UNIVERSIDADE FEDERAL DE GOIÁS – UFG
FACULDADE DE ADMINISTRAÇÃO, CIÊNCIAS CONTÁBEIS E CIÊNCIAS ECONÔMICAS - FACE
DEPARTAMENTO DE CONTABILIDADE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS CONTÁBEIS – PPGCONT

IVAN RIBEIRO MELLO

O IMPACTO DO BIG DATA NA PERFORMANCE DAS FIRMAS DE CAPITAL ABERTO NO BRASIL

GOIÂNIA – GO
2023

UNIVERSIDADE FEDERAL DE GOIÁS
FACULDADE DE ADMINISTRAÇÃO, CIÊNCIAS CONTÁBEIS E CIÊNCIAS ECONÔMICAS

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Ivan Ribeiro Mello

3. Título do trabalho

O Impacto do Big Data na performance das firmas de capital aberto no Brasil.

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Aletheia Ferreira Da Cruz**, **Professora do Magistério Superior**, em 02/10/2023, às 18:37, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ivan Ribeiro Mello**, **Discente**, em 03/10/2023, às 14:14, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4092097** e o código CRC **69E4132E**.

IVAN RIBEIRO MELLO

O IMPACTO DO BIG DATA NA PERFORMANCE DAS FIRMAS DE CAPITAL ABERTO NO BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Contábeis – PPGCONT, da Faculdade de Administração, Ciências Contábeis e Ciências Econômicas (FACE) da Universidade Federal de Goiás (UFG) como requisito para a obtenção de título de Mestre em Ciências Contábeis.

Área de Concentração: Ciências Contábeis

Orientadora: Profa. Dra. Alethéia Ferreira da Cruz

GOIÂNIA – GO

2023

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Mello, Ivan Ribeiro
O Impacto do Big Data na performance das firmas de capital aberto no Brasil [manuscrito] / Ivan Ribeiro Mello. - 2023.
71 f.

Orientador: Profa. Dra. Alethéia Ferreira da Cruz.
Dissertação (Mestrado) - Universidade Federal de Goiás, Faculdade de Administração, Ciências Contábeis e Ciências Econômicas (FACE), Programa de Pós-Graduação em Ciências Contábeis, Goiânia, 2023.
Bibliografia. Apêndice.
Inclui gráfico, tabelas, lista de figuras, lista de tabelas.

1. Big Data. 2. Performance da Firma. 3. Visão Baseada em Recursos (RBV). 4. Python. 5. Robotic Process Automation (RPA). I. Cruz, Alethéia Ferreira da, orient. II. Título.

CDU 657

UNIVERSIDADE FEDERAL DE GOIÁS

FACULDADE DE ADMINISTRAÇÃO, CIÊNCIAS CONTÁBEIS E CIÊNCIAS ECONÔMICAS

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº 09 da sessão de Defesa de Dissertação de Ivan Ribeiro Mello, que confere o título de Mestre em Ciências Contábeis na área de concentração em Ciências Contábeis.

Aos quatro dias do mês de julho do ano de dois mil e vinte e três, a partir das 14 horas e 00 minutos, na transmissão em videoconferência, pela Faculdade de Administração, Ciências Contábeis e Ciências Econômicas, realizou-se a sessão pública de Defesa de Dissertação intitulada **"O Impacto do Big Data na performance das firmas de capital aberto no Brasil".** Os trabalhos foram instalados pela Orientadora, Professora Doutora Alethéia Ferreira da Cruz (PPGCONT/UFG), com a participação dos demais membros da Banca Examinadora: Professor Doutor Pedro Henrique Melo Albuquerque, da Universidade de Brasília (PPGA/UnB), membro titular externo; cuja participação também ocorreu através de videoconferência, e Professor Doutor Moisés Ferreira da Cunha (PPGCONT/UFG), membro titular interno. Durante a arguição os membros da banca **não fizeram** sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pela Professora Doutora Alethéia Ferreira da Cruz, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos quatro dias do mês de julho do ano de dois mil e vinte e três.

TÍTULO SUGERIDO PELA BANCA

"O Impacto do Big Data na performance das firmas de capital aberto no Brasil".

Documento assinado eletronicamente por **Pedro Henrique Melo Albuquerque**, **Usuário Externo**, em 04/07/2023, às 15:45, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Moisés Ferreira Da Cunha**, **Professor do Magistério Superior**, em 04/07/2023, às 15:46, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alethéia Ferreira Da Cruz**, **Professora do Magistério Superior**, em 04/07/2023, às 15:52, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3865636** e o código CRC **8113E3BB**.

Referência: Processo nº 23070.036012/2023-49 SEI nº 3865636

Reitora

Profa. Dra. Angelita Pereira de Lima

Vice-Reitor

Prof. Dr. Jesiel Freitas Carvalho

Diretora da Faculdade de Administração Ciências Contábeis e Ciências Econômicas

Prof. Dra. Andréa Freire de Lucena

Coordenadora do Programa de Pós-Graduação em Ciências Contábeis

Profa. Dra. Michele Rílany Rodrigues Machado

UNIVERSIDADE FEDERAL DE GOIÁS – UFG
FACULDADE DE ADMINISTRAÇÃO, CIÊNCIAS CONTÁBEIS E CIÊNCIAS ECONÔMICAS - FACE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS CONTÁBEIS – PPGCONT

IVAN RIBEIRO MELLO

O IMPACTO DO BIG DATA NA PERFORMANCE DAS FIRMAS DE CAPITAL ABERTO NO BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Contábeis – PPGCONT, da Faculdade de Administração, Ciências Contábeis e Ciências Econômicas (FACE) da Universidade Federal de Goiás (UFG) como requisito para a obtenção de título de Mestre em Ciências Contábeis.

Orientadora: Profa. Dra. Alethéia Ferreira da Cruz.

Goiânia-GO, 04 de julho de 2023.

BANCA EXAMINADORA

_____

Profa. Dr. Pedro Henrique Melo Albuquerque
Universidade Federal de Brasília
Membro Externo

_____

Profa. Dr. Moises Ferreira Cunha
Universidade Federal de Goiás
Membro Interno

_____

Profa. Dra. Alethéia Ferreira da Cruz
Universidade Federal de Goiás
Orientadora

AGRADECIMENTOS

São tantas etapas superadas para iniciar a escrita desse parágrafo, que não há como escrevê-lo sem que os olhos estejam marejados desde o momento em que a primeira letra surgiu.

Inicialmente gostaria de agradecer à minha mãe, Beatriz Ribeiro de Oliveira. Você é a pessoa que fez impossíveis se tornarem realidade para me dar oportunidades que muitas vezes você nunca teve. Eu serei eternamente grato por você ter feito alguns muitos sonhos teus esperarem para que eu pudesse chegar mais próximo de alcançar os meus. Quanto mais o tempo passa, mais conhecimento e experiência eu adquiro, mais te admiro. Isso, certamente, é privilégio para poucos. Que honra poder escrever essas linhas te agradecendo e saber que você está aqui comigo, vendo essa vitória, que é nossa!

À minha avó, Eneida Ramos Ribeiro (*in memorian*), agradeço por ser minha maior inspiração acadêmica e ter me feito descobrir que "livro" e "liberdade" têm a mesma origem etimológica, "liber". Quanto mais conhecimento tivermos, mais livres seremos. Eternizei esse ensinamento para todo sempre em meu coração, interna e externamente, justamente para que quando você partisse, eu pudesse olhar para ti todos os dias. Infelizmente não consegui finalizar esse programa a tempo de poder lhe dar a alegria de me ver mestre em vida, mas saiba que sinto sua falta todos os dias, e ainda me pego te ligando sem querer quando o cérebro entra no automático. Eu prefiro acreditar que você foi acompanhar aí de cima, para levar boas novas para tanta gente especial que estava com saudade de você, em especial meu pai, Armando Silveira Mello Filho (*in memorian*). Tenho certeza de que ele está igualmente orgulhoso ao seu lado.

À minha companheira, Ana Cláudia Marchi, que foi fundamental desde o primeiro dia do meu ingresso no programa. Viver ao seu lado me dá a certeza de que o mundo ainda pode ser doce, seguro e que eu posso retirar minha armadura e deixar de estar o tempo inteiro preparado para a guerra. Você, mais do que ninguém, viu e participou ativamente de todo o esforço dessa grande jornada. Das madrugadas em claro aos pedidos, muitas vezes negados, para que eu desligasse o computador e fosse dormir... Só a gente sabe como foi duro e de todas as abdicações que você fez para que eu pudesse ter mais horas de estudo e concentração. Obrigado por ser exatamente quem você é. Tenho muita sorte e honra de estar ao seu lado, meu amor.

À minha irmã, Alice Ribeiro Mello, que sempre esteve ao meu lado em todos os momentos de incertezas, inseguranças e desabafos, durante toda a minha vida. Poder contar e

aprender com você me torna uma pessoa melhor todos os dias. Obrigado por sempre acreditar em mim, até quando nem eu mesmo acreditava mais. Essa conquista é nossa!

À professora Dra. Alethéia Ferreira Cruz, minha orientadora, agradeço por ter me direcionado no cumprimento dos objetivos acadêmicos, sempre me oferecendo oportunidades de pesquisa, extensão e aprofundamento do meu conhecimento, exercendo o máximo de compreensão a respeito dos meus compromissos profissionais e me dando forças para continuar em frente.

Ao professor Dr. Juliano de Lima Soares agradeço por ter me ensinado a importância de se fazer Ciência da forma mais correta possível. Com você aprendi muito pelo exemplo, tenho certeza de que um dia realizarei meu sonho de me tornar professor, e você terá contribuído muito para isso.

À minha colega de turma e parceira de trincheiras, Juliana Ferreira de Carvalho, obrigado por ter compartilhado tanto conhecimento e de maneira tão humilde. Além disso, obrigado por ter representado nossa turma nas reuniões do programa de maneira exemplar e ter me apresentado a UFG no momento do Estágio Docência. A sua parceria também foi fundamental para o meu sucesso!

À turma de 2021, o meu muito obrigado! Nós somos a prova de que a diversidade entrega muito valor e todos podemos ensinar e aprender, independente de quaisquer outros fatores. O maior trunfo é a união e o compartilhamento do conhecimento, sempre! Muitos vivas ao *Open Science*!

EPÍGRAFE

Impressiona-me a urgência por fazer. Saber não é suficiente; deve-se praticar. Querer não é o suficiente; deve-se agir.

Leonardo da Vinci

O mundo é formado não apenas pelo que já existe, mas pelo que pode efetivamente existir.

Milton Santos

RESUMO

Existe uma convicção comum de que as empresas devem se engajar ativamente em estratégias de *Big Data* para se manterem competitivas. No entanto, a preocupação constante das empresas está relacionada à estimativa do valor dos ganhos e dos gastos envolvidos na aquisição ou desenvolvimento dessas soluções. Nesse contexto, este trabalho busca responder a seguinte pergunta de pesquisa: qual o impacto do uso de *Big Data* na performance das companhias de capital aberto no Brasil? Assim sendo, seu objetivo é mensurar o impacto do uso de *Big Data* na performance das companhias de capital aberto no Brasil no período de 2010 a 2022. Dessa forma, a presente pesquisa replica o modelo de Cappa et al. (2020) no ambiente brasileiro, utilizando dados de aplicativos móveis disponibilizados pelas empresas na *Google Play Store* como *proxy* objetiva para o *Big Data* disponível para cada uma delas; tendo como plataforma teórica basilar, a Visão Baseada em Recursos (RBV). Os dados foram coletados de três fontes diferentes: *Refinitiv Eikon*©, Comissão de Valores Mobiliários (CVM) e *Google Play Store*, com o uso de *Robotic Process Automation* (RPA), através da biblioteca *PyAutoGUI* do *Python*. Os resultados encontrados demonstram que o impacto do *Big Data* em 2022 é menos significante do que observado em pesquisas anteriores e que, no período de 2010 a 2022, a crescente adoção da estratégia por diversas firmas em quase todos os setores avaliados, indica que a aplicação de *Big Data* parece ter gerado vantagem competitiva posicional e não sustentável. Nesse sentido, as principais contribuições deste trabalho estão relacionadas à desmistificação do *Big Data* como um conceito que ainda representa uma inovação vanguardista. Os achados indicam que o investimento em *Big Data* ainda faz sentido, mas sem a crença de que seja algo capaz de resolver quaisquer problemas pelo simples fato de ser investimento em tecnologia.

Palavras-chave: *Big Data*; Performance da Firma; Visão Baseada em Recursos (RBV); *Python*; *Robotic Process Automation (RPA)*

ABSTRACT

There is a common belief that companies should actively engage in Big Data strategies to remain competitive. However, companies' ongoing concern is related to estimating the value of gains and expenses involved in acquiring or developing these solutions. In this context, this study seeks to answer the following research question: What is the impact of using Big Data on the performance of publicly traded companies in Brazil? Therefore, its objective is to measure the impact of using Big Data on the performance of publicly traded companies in Brazil from 2010 to 2022. Thus, this research replicates the model of Cappa et al. (2020) in the Brazilian context, using mobile application data made available by companies on the Google Play Store as an objective proxy for the available Big Data for each of them; with the Resource-Based View (RBV) as the fundamental theoretical framework. Data was collected from three different sources: Refinitiv Eikon©, the Brazilian Securities and Exchange Commission (CVM), and the Google Play Store, using Robotic Process Automation (RPA) through the PyAutoGUI library in Python. The results show that the impact of Big Data in 2022 is less significant than observed in previous research, and that from 2010 to 2022, the increasing adoption of the strategy by various firms in almost all evaluated sectors indicates that the application of Big Data seems to have generated positional and unsustainable competitive advantage. In this sense, the main contributions of this study are related to demystifying Big Data as a concept that still represents a cutting-edge innovation. The findings suggest that investing in Big Data still makes sense, but without the belief that it is something capable of solving any problems simply because it is a technology investment.

Keywords: Big data; Firm performance; Resource-based View (RBV); Python; Robotic Process Automation (RPA)

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

A capacidade tecnológica de uma determinada sociedade tem se demonstrado cada vez mais fundamental para a geração de riqueza (Hilbert; López, 2011). Nesse contexto, o termo *Big Data* se tornou praticamente onipresente tanto no âmbito de mercado quanto acadêmico (Diebold, 2020). Essa grande quantidade de dados é considerada uma nova força-motriz capaz de aumentar o nível de competitividade das firmas e sustentar altos índices de crescimento por meio de três principais práticas: aprendizado constante, através de fluxos de dados, a respeito das estratégias que atingem os melhores resultados; utilização de força de trabalho especializada, como, por exemplo, Cientistas e Analistas de Dados e descentralização das tarefas relacionadas à Inteligência Analítica (*Analytics*) para outras áreas das firmas que não somente a de Tecnologia da Informação (Akoka, Comyn-Wattiau; Laoufi, 2017).

Firmas que aplicam estratégias bem-sucedidas de *Big Data Analytics* (BDA), ou seja, analisam o mar de dados por elas coletados, criam um componente crucial capaz de melhorar substancialmente o processo de tomada de decisões (Hagel, 2015). Além disso, resultados de pesquisas indicam que a adoção de BDA tem relação com uma postura ativa na execução de estratégias, fazendo com que as firmas sejam capazes de projetar o futuro com menos incerteza do que suas concorrentes, reduzindo o custo de aquisição de clientes em 47% e aumentando a receita em até 8% (Liu, 2014).

As firmas que desenvolvem seus produtos em torno de uma estratégia adequada de uso do *Big Data* alcançam crescimento exponencial principalmente por melhorias em três mecanismos fundamentais: redução de custos, acessibilidade e estrutura inovadora do modelo de negócios (Johnson et al., 2017). Um *case* de sucesso quanto à redução de custos a ser citado é o da Premier Healthcare Alliance, que usou o compartilhamento de dados e inteligência analítica para reduzir o custo de operação em 2.85 bilhões de dólares (IBM, 2012). Outros vários exemplos de estratégias bem-sucedidas de implementação de BDA são citados por Wamba et al. (2015) em sua pesquisa que objetivou desenvolver um *framework* a respeito do que seria, de fato, *Big Data*.

Recente revisão bibliométrica realizada por Nobanee (2021) considerando as palavras-chave *Big Data* e *Finance* apontou um crescimento exponencial na produção acadêmica relacionando os dois temas, corroborando com as impressões de Diebold (2020) de que esse é um tema central para todas as ciências.

De 2011 a 2020, as palavras-chave mais mencionadas pelos artigos envolvendo *Big Data* e *Finance* foram: mineração de dados (*Data Mining*) - com 97 ocorrências; aprendizado de máquina (*Machine Learning*) - com 90 ocorrências; mercados de capital (*Financial Markets*) - com 71 ocorrências; inteligência artificial (*Artificial Intelligence*) - com 67 ocorrências; investimentos (*Investments*) - com 61 ocorrências; *Big Data Analytics* (BDA) - com 59 ocorrências; economia (*Economics*) - com 52 ocorrências e tomada de decisão (*Decision Making*) - com 50 ocorrências (Nobanee, 2021).

Não são apenas aspectos positivos que surgem a partir da utilização do *Big Data* pelas companhias, sendo a proteção dos dados contra vazamentos e a utilização ética da informação para geração de valor os principais desafios enfrentados pelas companhias (Cappa et al., 2020; Freund et al., 2019). Com relação ao custo de processamento, armazenamento e transmissão de dados, no período de 1992 a 2016, foi possível observar uma intensa redução através da democratização do acesso à *Cloud Computing*, ou Computação em Nuvem (Silva, Bonacelli e Pacheco, 2020).

Para este trabalho, utiliza-se como base o esforço de De Mauro, Greco e Grimaldi (2015) na elaboração de uma definição consensual sobre *Big Data* de uma maneira concisa e mais consensual possível: "*Big Data* representa os ativos informacionais caracterizados por um alto volume, velocidade e variedade que requerem Tecnologia e Métodos Analíticos específicos para sua transformação em valor" (De Mauro et al., 2015, p.103).

Pesquisas anteriores encontraram relação explicativa entre *Big Data* ou BDA e a performance das companhias. Wamba et al. (2016) pesquisaram os efeitos das Capacidades Dinâmicas (CD) desenvolvidas com o uso de BDA como mediador através de *surveys* enviadas para mais de 297 Gerentes de Tecnologia e Analistas de Dados com experiências em *Big Data* na China. A análise encontrou que 65% da variação da performance da companhia, tanto financeira quanto de mercado, foi explicada pelos dois conceitos.

Côrte-Real et al. (2016) pesquisaram o valor do BDA usando também uma *survey* para a qual receberam resposta de 175 Gerentes de Tecnologia e de Negócios por toda a Europa. O resultado encontrado por meio do modelo proposto com base nas teorias das CD e Conhecimento (*Knowledge-based view*) demonstrou que 77.8% da variação na vantagem competitiva foi explicada pelo uso de BDA.

Johnson et al. (2017) pesquisaram como o *Big Data* era crucial para a fase de Desenvolvimento de Novos Produtos (DNP). O método utilizado também foi um *survey* enviado para 261 Gerentes de Produto nos Estados Unidos e identificaram que quando uma

firma adota uma abordagem exploratória, BDA é fundamental para geração de valor através de novos produtos.

Cappa et al. (2020) pesquisaram como o *Big Data* impacta a geração de valor para as companhias, sendo o valor medido através do Q de Tobin, mas dessa vez utilizando os dados dos Aplicativos Móveis disponíveis na *Play Store* do *Google* como *proxy* para quantificar o *Big Data* disponível para cada uma das firmas integrantes do índice S&P 500 dos Estados Unidos. Com essa pesquisa, os autores encontraram que o volume, representado pela quantidade de *downloads* do aplicativo, tem um impacto negativo na performance; que uma maior variedade, representada pela quantidade de permissões de coleta de dados que cada um dos aplicativos (GPS, Acesso à Câmera, etc...) solicita, moderava positivamente o efeito negativo do volume na performance; que a veracidade, representada pelo percentual da força de trabalho especializada em dados dentro de cada uma das firmas, afetava positivamente a geração de valor.

Entre todos os trabalhos anteriormente citados que investigaram o aspecto explicativo entre *Big Data* e a performance das firmas, um ponto em comum é naturalmente percebido: o uso da Visão Baseada em Recursos (*Resource-based view* – RBV) como plataforma teórica suplantando as hipóteses e desenhos de pesquisa.

Nesse sentido, foi realizada uma busca por trabalhos cujos objetivos fossem a revisão sistemática da literatura de *Big Data* e performance das firmas. Foi constatado, novamente, que a plataforma teórica mais utilizada é a da Visão Baseada em Recursos (Maroufkhani et al. 2019; Arunachalam et al., 2018). Tendo em vista tal convergência, dentre os trabalhos que buscam entender a relação entre o uso de *Big Data* e a performance das firmas, no uso da Visão Baseada em Recursos (*Resource-based view* – RBV) como arcabouço teórico basilar, essa é a mesma escolha realizada para a condução da presente pesquisa.

Essa teoria indica que a diferença de performance entre firmas similares é gerada pelos recursos internos valiosos, raros, imperfeitamente imitáveis e organizados para capturar valor (Barney, 1986; 1991). Os recursos internos que possuem as características citadas anteriormente, são fundamentais na geração da vantagem competitiva sustentável (Kretzer; Menezes, 2006).

1.2 Problema de pesquisa

Para Davenport (2014) já é possível entender que o impacto do *Big Data* na geração de diversas oportunidades é grande, porém ainda não é possível entender os detalhes de como as

companhias e as indústrias são afetadas. Isto quer dizer, também, que ainda não se tem clareza suficiente se as grandes companhias são as mais beneficiadas ou não e como o mercado responde de forma geral ao tema. Assim, o autor provoca: "Há pouca dúvida de que a análise de *Big Data* pode transformar as organizações. As firmas que reconhecerem toda a extensão dessas oportunidades aproveitarão o maior valor" (Davenport, 2014b, p.50).

Nesse contexto, este trabalho busca responder a seguinte pergunta de pesquisa: Qual o impacto do uso de *Big Data* na performance das companhias de capital aberto no Brasil?

1.3 Objetivos

Dessa forma, o objetivo desse trabalho é mensurar o impacto do uso de *Big Data* na performance das companhias de capital aberto no Brasil no período de 2010 a 2022.

1.4 Justificativas

Os resultados dessa pesquisa visam contribuir ao debate, no âmbito das empresas de capital aberto do Brasil, entre investimento em *Big Data* e a sua materialização em resultados positivos para a firma que o faz, já que o mercado global de *Big Data* e *Business Analytics* foi avaliado em 168.8 bilhões de dólares em 2018 (Statista, 2022).

A pergunta da pesquisa é relevante uma vez que busca não só responder às provocações realizadas por Davenport (2014), mas também replicar o modelo de Cappa et al. (2020) atendendo ao pedido dos autores da realização da pesquisa em outros mercados que não fossem os dos Estados Unidos, Europa ou Ásia.Outro ponto importante que essa pesquisa enseja responder é o quanto há realmente de inovação no uso do *Big Data* enquanto recurso capaz de gerar vantagem competitiva sustentável pelas empresas de capital aberto no Brasil. Na grande maioria das vezes em que o tema e o termo são mencionados muito avanço é discutido, mas nem é sempre que é possível comprová-lo por métodos mais objetivos.

Nesse sentido, recente revisão sistemática da literatura envolvendo *Big Data* e performance da firma demonstrou que a maioria dos artigos publicados utilizam *surveys* como metodologia principal (Maroufkhani et al., 2019). Outrossim, a utilização de medidas objetivas, como *proxy* de *Big Data*, permite a replicação da pesquisa em outros contextos.

Adicionalmente, o uso de medidas objetivas para mensurar as dimensões do *Big Data* por meio dos dados de aplicativos móveis disponíveis para *smartphones* busca diminuir a

subjetividade que está presente em maior grau quando as informações são coletadas através de *surveys* (Cappa et al., 2020).

1.5 Contribuições

Esta pesquisa contribui no desenvolvimento dos estudos a respeito do impacto do *Big Data* na performance das firmas em alguns sentidos. Primeiro, entende-se que a comparação dos resultados dessa pesquisa com a de Cappa et al. (2020) poderá agregar de forma relevante para a compreensão de como fatores exógenos de mercado podem afetar a geração de valor a partir do uso de *Big Data* quando se compara o mercado dos Estados Unidos com o de países emergentes, como Brasil.

Segundo, acrescenta-se o aspecto da velocidade do *Big Data* ao considerar o quão atualizados os aplicativos disponibilizados pelas firmas estavam no momento da coleta de dados, já que foi possível obter a data de última atualização de cada um deles. O acréscimo dessa dimensão do *Big Data* ao presente trabalho objetiva capturar a heterogeneidade que pode existir entre os *players*, no que tange a capacidade de cada firma em atualizar eventuais erros, incluir novas funcionalidades e absorver as sugestões de seus clientes, com velocidades distintas, em cada nova versão disponibilizada em produção.

Adicionalmente, este trabalho pode auxiliar gestores a definirem uma estratégia de *Big Data* mais acurada visando alcançar melhorias na performance de suas companhias e não apenas seguir a moda (Davenport, 2014).

Por fim, o presente trabalho busca formar uma ponte entre a pesquisa acadêmica e o uso de seus resultados na tomada de decisões dos gestores e do mercado.

Além desta introdução, o capítulo 2 traz uma revisão da literatura que alicerça os argumentos teóricos desta pesquisa, a formulação das hipóteses e o modelo empírico construído. O capítulo 3 trata dos procedimentos metodológicos utilizados para realização da pesquisa tais como: caracterização do presente trabalho, descrição dos processos de coleta e limpeza dos dados e abordagem estatística; para posterior avaliação das hipóteses. O capítulo 4 apresenta os resultados obtidos na avaliação das hipóteses. Por fim, o capítulo 5 traz as conclusões finais e o relacionamento com pesquisas anteriores e com a plataforma teórica escolhida.

2 REVISÃO DA LITERATURA

2.1 Teoria da visão baseada em recursos (RBV)

Segundo Wernerfelt (1984), uma determinada firma possui recursos que são utilizados para a elaboração de um ou mais de seus produtos que, por sua vez, são comercializados em diferentes mercados na busca da realização do lucro. Nesse sentido, a definição de um recurso, para Wernerfelt (1984), diz respeito aos ativos tangíveis ou intangíveis controlados por uma firma em um determinado período. Tais recursos, disponíveis de forma heterogênea entre os *players* de um mesmo mercado, são utilizados pelas firmas na busca da geração de valor.

Quando se utiliza a análise baseada em recursos, diferentemente da perspectiva tradicional, a empresa se apoia na dinâmica de oportunidades e ameaças ao produto que um determinado ambiente oferece à medida em que enfatiza as condições internas de cada uma das companhias como propulsoras para o diferencial competitivo e, portanto, melhora na performance (Wernerfelt, 1984). Os recursos, segundo a RBV, são fruto dessas condições internas específicas de cada empresa.

Os principais recursos de uma firma, de acordo com Barney (1986, 1991), são capacidades, processos organizacionais, atributos, informação e conhecimento que são classificados em 3 categorias: i) recursos de capital físico – ex.: maquinário; ii) recursos de capital humano – ex.: time de gestores ou engenheiros de *software*; iii) recursos de capital organizacional – ex.: cultura e processos internos.

Em um ambiente de competição, sabe-se que a disponibilidade dos recursos não é linear entre todos os concorrentes de um determinado mercado. Essa heterogeneidade entre os recursos disponíveis para cada um dos competidores gera implementações estratégicas distintas dos mesmos e é isso que origina os diferentes resultados alcançados pelas companhias segundo a RBV (Barney, 1986, 1991, 2001; Peteraf, 1993).

Uma estratégia oriunda da aplicação de um recurso que não está sendo implementada por nenhum concorrente atual ou potencial e que gera valor pode ser classificada como vantagem competitiva posicional. Quando os benefícios dessa mesma estratégia não conseguem ser replicados pelos concorrentes, ainda que eles se utilizem de outros meios ou substituições, origina-se a vantagem competitiva sustentável (Barney, 1986, 1991; Kretzer; Menezes, 2006).

Dessa forma, os recursos geradores de vantagens competitivas sustentáveis são: i) valiosos – ao permitirem a implantação de estratégias que tornam a firma mais eficaz e eficiente;

ii) raros – não disponíveis com facilidade para todos os *players* de um determinado mercado; iii) imperfeitamente imitáveis – as firmas que não possuem esses recursos não conseguem obtê-los a não ser que os construam já que eles são frutos de condições históricas únicas, casualmente ambíguos e socialmente complexos e iv) organizados para capturar valor (Barney, 1986, 1991, 2001).

Uma metáfora que pode ser utilizada para explicar o funcionamento da RBV dentro de uma determinada organização é a de uma banheira. Investimentos em pesquisa e desenvolvimento, por exemplo, são como o fluxo de água que vai, aos poucos, aumentando o volume de líquido dentro de uma banheira – o estoque de ativos intangíveis estratégicos. Esse estoque vai se tornando cada vez mais específico e qualificado de acordo com as dinâmicas internas da companhia – sua organização para capturar valor ou, no caso da metáfora, o formato específico de cada banheira - e é acumulado através da escolha apropriada da quantidade e frequência de investimento dentro de um determinado período (Dierickx; Cool, 1989).

Assim, não é possível atingir o mesmo estoque de ativos estratégicos se o fluxo de investimentos não foi dosado oportunamente durante um período determinado. É como dobrar o fluxo de água por apenas 2 minutos e esperar que o volume de líquido na banheira seja igual ao de uma outra que foi preenchida, *ceteris paribus*.

Por esse motivo, se o ponto de partida no estoque de ativos estratégicos entre duas companhias semelhantes é distinto, igualmente distinta será a acumulação deste ponto do tempo em diante, ainda que ajustes no fluxo de investimento sejam realizados no meio do caminho (Dierickx; Cool, 1989).

Isso quer dizer que heterogeneidade implica que as firmas com recursos medianos só podem esperar atingir o ponto de equilíbrio, enquanto as que acumularam um estoque de ativos estratégicos maior e melhor receberão rendimentos adicionais (Peteraf, 1993).

Outro ponto a destacar diz respeito ao conceito de que, muitas vezes, um recurso de vantagem competitiva sustentável é classificado como casualmente ambíguo. A analogia, dessa vez, é a de uma máquina caça-níquel em um cassino.

Uma pessoa pode ganhar todas as moedas disponíveis na máquina com apenas uma tentativa e pouco dinheiro investido, ainda que a probabilidade de outra pessoa que pode tentar mais vezes, por ter mais recursos, seja maior. Da mesma forma, uma firma pode desenvolver um recurso extremamente valioso, raro, inimitável com pouco investimento, ainda que seja mais improvável (Dierickx; Cool, 1989).

Dentre os recursos disponíveis, o *Big Data* tem se apresentado como um ativo valioso na aplicação de estratégias corporativas. Sob a perspectiva da RBV, investimentos dessa

natureza podem culminar na geração de vantagem competitiva sustentável (Fosso Wamba et al., 2015; Côrte-Real et al., 2016; Johnson et al., 2017; Cappa et al., 2020).

A utilização da RBV para explicar o fenômeno do *Big Data* enquanto recurso interno capaz de gerar valor pelo fato de desencadear uma vantagem competitiva sustentável segue trabalhos de revisão sistemática da literatura envolvendo o *Big Data* e a performance das firmas (Maroufkhani et al. 2019; Arunachalam et al., 2018).

2.2 Big Data

Para Diebold (2020) *Big Data* não é tão somente um termo, mas um fenômeno e uma disciplina emergente. Enquanto fenômeno o autor entende que este é um tema científico chave de nossos tempos para todas as ciências. Enquanto disciplina, no campo acadêmico, há bastante discussão quanto à redundância no ensino, uma vez que para os céticos disciplinas tradicionais como ciência da computação e estatística seriam capazes de assimilar o fenômeno de forma satisfatória. Por outro lado, no mundo dos negócios, é cada vez mais comum encontrar executivos em cargos estratégicos mencionando o termo (Diebold, 2020; Francisco et al., 2020).

Apresentar as origens do termo *Big Data* não é tarefa fácil, conforme afirma Diebold (2020). De toda forma, a primeira vez que o termo foi caracterizado de maneira mais completa e enriquecida foi no relatório de pesquisa corporativa não publicado elaborado por Laney (2001) na empresa META Group, hoje pertencente à GARTNER. Este documento destacou, de modo inédito, as primeiras dimensões do *Big Data*: volume, variedade e velocidade. Essas dimensões, por sua vez, vieram a ficar comumente conhecidas como os 3 V's do *Big Data*.

No relatório de Laney (2001), o comércio eletrônico (*e-commerce*) é apontado como o propulsor do *Big Data*. À medida em que esses canais de venda conseguem atingir muito mais consumidores com um menor custo a profundidade e a largura dos dados coletados em uma única transação chega a ser 10 vezes maior do que em um modelo tradicional. Isso representa o volume, que significa o aumento exponencial na quantidade de informação armazenada pelas companhias em bancos de dados locais, como *mainframes*, ou *online* utilizando soluções de *Cloud Computing*.

O autor também apontou que a troca de informações cada vez mais rápida a cada interação com os consumidores forçava as companhias a fazerem uma série de melhorias nas suas capacidades tecnológicas de processamento - velocidade. Além disso, um dos grandes desafios mencionados ainda naquela época era a expectativa de aumento da variedade de formatos de dados incompatíveis entre si - variedade.

Outro esforço de definição do escopo relacionado ao termo *Big Data* foi realizado por Gantz e Reinsel (2011). Para os autores uma nova geração de tecnologias e arquiteturas de sistemas desenhadas para extrair valor econômico a partir de grandes volumes de dados variados capturados e analisados em alta velocidade forma o conceito dinâmico da atividade que intersecciona diversas áreas da Tecnologia da Informação.

De Mauro, Greco e Grimaldi (2015), a partir de uma extensa revisão sistemática da literatura, observaram que os trabalhos científicos acerca do *Big Data* se concentram em 4 grandes assuntos: informação - o combustível moderno para a existência e permanência do fenômeno; tecnologia - os equipamentos necessários para armazenar e processar a informação; métodos - as técnicas utilizadas para converter *Big Data* em valor e impacto - como a existência dessas enormes quantidades de informação nas mãos das firmas e governos afetam, principalmente, a privacidade dos consumidores e sociedade em geral.

Outra contribuição do trabalho de De Mauro, Greco e Grimaldi (2015) foi a elaboração de uma definição consensual sobre *Big Data* de forma concisa. Para os autores, "*Big Data* representa os ativos informacionais caracterizados por um alto volume, velocidade e variedade que requerem Tecnologia e Métodos Analíticos específicos para sua transformação em valor." (De Mauro et al., 2015, p.103)

Com o objetivo de sintetizar o conceito de *Big Data* no contexto da presente pesquisa, a Tabela 1 apresenta a definição consensual geral e as definições das suas 5 dimensões, bem como a definição do ato de utilizar o *Big Data* para gerar valor, denominado na literatura de *Big Data Analytics* (BDA).

Tabela 1 - Definições de Big Data

Atributo	Definição	Fonte
Definição consensual	"*Big Data* representa os ativos informacionais caracterizados por um alto volume, velocidade e variedade que requerem Tecnologia e Métodos Analíticos específicos para sua transformação em valor."	(De Mauro et al., 2015 p. 103)
Volume	"O grande volume de dados que é ou consumido ou armazenado através de um número grande de registros."	(Fosso Wamba et al., 2015 p.236)
Variedade	"Dados gerados atráves de uma grande variedade de fontes e formatos contendo campos multidimensionais."	(Fosso Wamba et al., 2015 p.236)
Velocidade	"Frequência da geração ou entrega dos dados."	(Fosso Wamba et al., 2015 p.236)
Veracidade	"A imprevisibilidade de alguns conjuntos de dados demanda análise especializada do *Big Data* para gerar confiança nas previsões."	(Fosso Wamba et al., 2015 p.236)

Atributo	Definição	Fonte
Valor	"Até que ponto o *Big Data* gera insights e ou benefícios economicamente dignos por meio de extração e transformação."	(Fosso Wamba et al., 2015 p.236)
Big Data Analytics (BDA)	"Tecnologias, por exemplo bancos de dados e ferramentas de mineração de dados, e técnicas, por exemplo métodos analíticos, que uma determinada companhia pode empregar para analisar grandes volumes de dados complexos de variadas aplicações com o objetivo de aumentar a sua performance em diversas dimensões"	(Mikaelef et al., 2017 p.556)

Fonte: Adaptado de Fosso Wamba et al. (2015) e Cappa et al. (2020)

Tendo em vista o tamanho do mercado de *Big Data* e *Business Analytics*, estimado em 168.8 bilhões de dólares em 2018 (Statista, 2022), relacionar os volumosos investimentos que as firmas fazem nessas tecnologias com a performance obtida é de fundamental importância.

2.3 Big data vs. Performance

Firmas que performam bem geram lucro e boas perspectivas para seus *stakeholders* no longo prazo. Isso quer dizer que, a partir do crescimento da firma, novas oportunidades de emprego são criadas melhorando a capacidade de geração de riqueza da sociedade na qual a firma se encontra inserida (Taouab e Issor, 2019).

A partir da definição de De Mauro, Greco e Grimaldi (2015), na Tabela 1, *Big Data* pode ser considerado um ativo informacional. Esse ativo informacional, por sua vez, quando existente dentro de uma determinada firma é considerado um ativo intangível, já que os dados não estão exatamente registrados nas suas demonstrações contábeis, mas agregam valor relevante ao permitirem um processo decisório mais robusto por meio de uma cultura *data-driven*. Isto é, quando decisões são tomadas a partir da análise de dados que representam os fatos e fenômenos que estão sendo discutidos (Zhang et al., 2020). Além disso, ativos intangíveis são responsáveis por direcionar de forma significativa como as firmas criam valor na economia moderna (Haji e Ghazali, 2018).

Nesse contexto, para a presente pesquisa, *Big Data* é classificado como um ativo informacional, intangível, que, se utilizado como recurso estratégico interno, pode gerar vantagem competitiva sustentável conforme os preceitos da RBV.

Tendo em vista a classificação do *Big Data* como ativo intangível não registrado nas demonstrações contábeis, torna-se relevante investigar o impacto que o seu uso pode trazer em termos de performance de mercado, isto é, a maneira como o ambiente externo enxerga a firma utilizadora.

É evidente que o desenvolvimento contínuo de tecnologia ao qual as empresas se submetem para dar manutenção em um ativo intangível como o *Big Data* demanda investimentos robustos por meio de times especializados e que, muitas vezes, acabam retirando recursos das atividades principais das companhias (Francisco et al., 2020). Logo, a pressão por resultados advindos desse tipo de estratégia é uma realidade nas firmas que optam por implementá-las. Essa dualidade entre uma expectativa de valor positiva e uma expectativa de custo negativa foi confirmada no trabalho de Maçada et al. (2019) por meio de *survey* aplicada em 99 empresas brasileiras.

O *Big Data*, enquanto fenômeno que possui diversas dimensões, demanda uma análise específica para cada uma delas, para que seja possível entender as suas relações na geração de valor (Cappa et al., 2020). E, dessa forma, auxiliar na tomada de decisões se existe ou não uma das dimensões que possa gerar mais valor do que as outras, ou até mesmo prejudicar a geração de valor da empresa através da aplicação da referida estratégia.

A dimensão de volume do *Big Data* está diretamente relacionada a uma das principais características do termo, que é a grandiosidade em termos de tamanho. Por outro lado, quanto maior o tamanho dos *datasets* – conjuntos de dados - em poder das firmas aumentam, igualmente, os riscos de vazamento de informações, bem como a possibilidade de ocorrência do fenômeno da "obesidade de informação" (Cappa et al., 2020).

Esse fenômeno indica que a simples existência de uma quantidade imensa de informação pode mais confundir do que auxiliar as firmas na elaboração de suas estratégias de geração de valor que culminam em melhoria na performance. Dessa forma é apresentada a seguinte hipótese:

H1 Quanto maior o volume, menor o impacto do *Big Data* na performance da firma.

A dimensão de variedade do *Big Data* funciona como um moderador positivo do efeito negativo do volume. Isto é, um *dataset* que possua um grande volume e uma grande variedade tem um potencial muito maior na geração de valor impactando positivamente a performance da firma.

Tal moderação positiva advém do fato de que uma maior variedade de dados permite às firmas estabelecerem estratégias mais robustas entendendo de forma mais assertiva as influências de diversas variáveis no comportamento de seus clientes, por exemplo. Sendo assim, é apresentada a seguinte hipótese:

H2 A variedade do *Big Data* disponível impacta positivamente a performance da firma.

A dimensão de veracidade diz respeito à capacidade das firmas de extrair informação confiável e replicável do imenso mar de dados disponível. É essa dimensão do *Big Data* que versa a respeito do processo de análise dos dados propriamente dito. Assim, as firmas que são capazes de conferir maior veracidade aos *datasets* que estão em seu poder estão mais preparadas para conceber estratégias de sucesso que culminarão em melhora nas suas performances (Francisco et al., 2020; Maçada et al., 2019). Dessa forma, é apresentada a seguinte hipótese:

H3 A veracidade do *Big Data* impacta positivamente a performance da firma.

Cappa et al. (2020) assumiram que a dimensão da velocidade seria uma variável constante para todas as firmas tendo em vista que na economia moderna o uso de aplicativos móveis está ligado, principalmente, à existência de trocas de dados em tempo real. Nesta pesquisa, porém, propõem-se a utilização dessa dimensão do *Big Data* baseando-se na velocidade com que as firmas disponibilizam atualizações de seus aplicativos móveis para seus clientes, uma vez que segundo os preceitos da RBV essa é uma dimensão do *Big Data* que pode gerar heterogeneidade entre os *players.*

Em outras palavras: é esperado que o movimento de atualização e, portanto, melhor manutenção do ativo intangível do *Big Data*, seja heterogêneo entre as firmas e que aquelas que mais disponibilizam atualizações sejam as mesmas que mais se preocupam em atingir as necessidades de seus clientes. Além disso, coletam dados com maior rapidez o que, consequentemente, afetam positivamente a sua performance. A partir dessa contextualização, é apresentada a hipótese abaixo:

H4 A velocidade de atualização do *Big Data* impacta positivamente a performance da firma.

Tendo em vista os preceitos da RBV, principalmente aqueles desenvolvidos por Dierickx e Cool (1989), entende-se que as firmas que lançaram seus aplicativos móveis antes que seus concorrentes, passaram a coletar dados e a utilizar *Big Data* primeiro. Assim, conforme explicam os autores, a posição inicial favorável de um ativo estratégico tem relação positiva com a capacidade da firma de gerar mais vantagem competitiva sustentável do que seus concorrentes. Nesse sentido a seguinte hipótese é apresentada:

H5 Firmas que implementaram o *Big Data* primeiro obtiveram melhor performance.

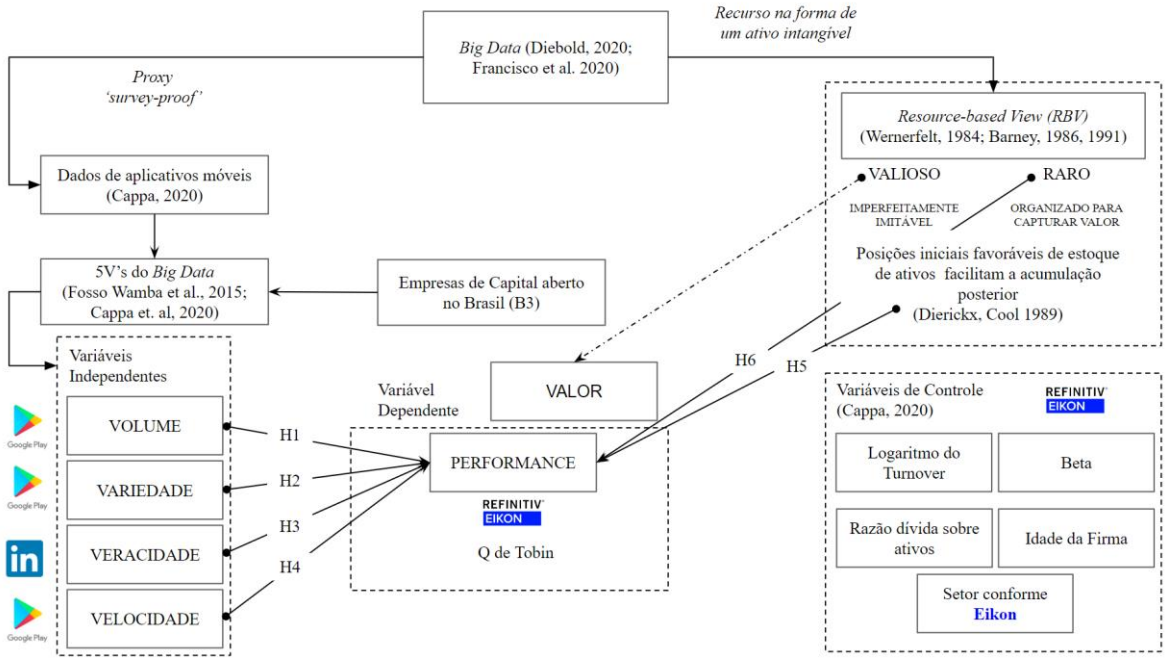
À luz da RBV um recurso que gera vantagem competitiva sustentável está disponível de forma heterogênea em um determinado mercado (Peteraf, 1993). Firmas que possuem

recursos não raros conseguem apenas atingir o ponto de equilíbrio, uma vez que não existe vantagem competitiva sustentável frente aos seus concorrentes. Nesse sentido, torna-se importante avaliar o impacto da raridade da aplicação do *Big Data* na performance das firmas, como gerador de vantagem competitiva. Assim, a seguinte hipótese é apresentada:

H6 Nos setores onde o uso de *Big Data* é maior, o impacto na performance é menor.

Com o objetivo de resumir a relação da RBV, o *Big Data* e as hipóteses formuladas para essa pesquisa, a Figura 1 foi construída.

Figura 1 – Modelo Empírico da Pesquisa



Fonte: Elaboração própria do autor.

O conceito empírico basilar da pesquisa é o *Big Data*. Partindo do topo do desenho para a esquerda, essa pesquisa segue estudos anteriores ao utilizar os Dados de Aplicativos Móveis como *proxy* mais objetiva para medir as dimensões independentes do *Big Data* seguindo o modelo de 5Vs. O intuito de seguir essa linha de raciocínio, evitando abordagens de *survey*, visa facilitar a replicação dessa pesquisa em situações futuras e ofertar como resultado, uma validação empírica mais aderente à realidade de mercado das firmas de capital aberto no Brasil.

Além disso, o desempacotamento do *Big Data* em dimensões distintas ajuda a capturar individualmente os efeitos de cada uma delas na geração de valor, e não um único efeito exclusivo somado (Johnson et al., 2017).

Por esse motivo, no lado esquerdo da Figura 2, é possível identificar a conexão das dimensões independentes do *Big Data* por meio das hipóteses H1 a H4 com a dimensão dependente do modelo de 5Vs, que é o valor entregue pelo *Big Data* medido pelo Q de Tobin, conforme Cappa et al. (2020) no presente modelo empírico.

O conceito teórico basilar da pesquisa é a RBV. Segundo essa lente teórica. um recurso que gera vantagem competitiva sustentável tem de ser: valioso, raro, imperfeitamente imitável e organizado para capturar valor. Além disso, empresas que são precursoras – *early adopters* - em uma determinada estratégia, na presente pesquisa, implementação do *Big Data*, tendem a gerar cada mais valor por meio dos estoques cada vez maiores desse mesmo ativo e por desenvolverem uma capacidade maior de fazer a sua manutenção (Dierickx; Cool 1989).

Para refletir essa linha teórica no modelo empírico, a parte superior direita da Figura 2 foi construída. Logo, partindo do topo do desenho para a direita, essa pesquisa propõe duas hipóteses, H5 e H6, visando testar mais profundamente a aderência do *Big Data* enquanto recurso - ativo intangível – capaz de gerar vantagem competitiva sustentável.

Assim, busca-se avaliar de forma objetiva o que Francisco et al. (2020) provocam em seu trabalho: a crença generalizada de que as empresas deveriam, obrigatoriamente, se envolver ativamente na adoção de estratégias de *Big Data* para se manterem competitivas. Porém, segundo os preceitos da RBV já discutidos, quanto mais um recurso estratégico é disseminado em um determinado mercado, menos tende a gerar vantagens competitivas sustentáveis, já que os preceitos de raridade e não imitabilidade são violados.

A principal diferença entre as hipóteses H5 e H6 diz respeito ao nível de análise que está sendo empregado. Em H5 são as empresas propriamente ditas, em H6 o setor, mas ambas almejam verificar se a diminuição da raridade e do aumento da replicação da mesma estratégia de *Big Data* por outros *players* do mercado que não os *early adopters* diminuiu a geração de valor percebido e performance de mercado pelas empresas precursoras ao longo do tempo.

3 PROCEDIMENTOS METODOLÓGICOS

3.1 Caracterização da pesquisa

Este estudo é classificado como descritivo porque objetiva descrever a realidade do impacto do uso de *Big Data* na performance das companhias de capital aberto no Brasil, de abordagem quantitativa, uma vez que intenciona responder ao problema de pesquisa por meio de medidas objetivas através dos dados coletados dos *apps* das empresas disponíveis na *Google Play Store*. Outrossim, a abordagem quantitativa busca auxiliar o processo de tomada de decisões por meio de uma abordagem baseada em evidências empíricas com a aplicação de métodos estatísticos de análise.

A amostra é composta pelas empresas de capital aberto listadas na B3. O início da janela temporal inicia em 2010, por ser o ano em que as normas internacionais de contabilidade são aplicadas de forma obrigatória no Brasil, e se encerra no ano de 2022, por ser o último ano de divulgação de dados anterior à execução dessa pesquisa.

Conforme discutido anteriormente nas seções 1 e 2 do presente trabalho, as dimensões independentes do *Big Data* são características que influenciam de forma heterogênea a dimensão dependente que é a geração de valor. Na presente pesquisa, a geração de valor será medida através de *proxies* que reflitam a performance de mercado das firmas.

3.2 Variáveis de interesse – *big data*

Encontrar uma *proxy* que reflita o *Big Data* disponível para cada uma das firmas não é tarefa fácil, porém, no contexto dessa pesquisa, mantém-se o direcionamento de Cappa et al. (2020) na utilização dos dados de aplicativos móveis disponíveis na *Play Store do Google*. Essa *proxy*, ainda que possua limitações, permite mensurar de forma objetiva as dimensões do *Big Data* e facilita a replicação da pesquisa.

Assim sendo, para capturar cada uma das dimensões independentes do *Big Data*, é apresentada a Tabela 2 a seguir com o objetivo de explicar como cada uma delas será contemplada na presente pesquisa.

Tabela 2 – Dimensões do Big Data e suas proxies correspondentes

Dimensão e Hipótese de Pesquisa	Definição da *Proxy*
Volume (H1)	Número de *downloads* dos aplicativos, para cada firma, na Google Play Store
Variedade (H2)	Número de tipos diferentes de dados coletados pelas firmas em cada aplicativo na Google Play Store (ex.: GPS, Permissão para acessar as fotos, Permissão para acessar a câmera, entre outros)
Veracidade (H3)	Percentual da força de trabalho especializada em BDA em relação ao total de empregados de cada uma das firmas conforme dados coletados da rede social profissional *Linkedin*
Velocidade (H4)	A diferença de dias entre a data da coleta e a última atualização do aplicativo disponibilizada pelas firmas na Google Play Store. Quanto menor a diferença de dias entre as datas, maior velocidade de atualização
Implementação do *Big Data* (H5)	Data de lançamento do aplicativo de cada firma na Google Play Store, organizada ordinalmente da mais antiga para a mais nova, por setor e empresa
Maturidade do uso de Big Data (H6)	Razão da Quantidade de empresas com *app* lançado na *Play Store* pela Quantidade Total de Empresas, por setor, por ano fiscal ao qual as demais variáveis do modelo se referem

Fonte: Adaptado de Cappa et al. (2020) para as hipóteses H1, H2 e H3. Elaboração própria para as hipóteses H4, H5 e H6.

3.3 Variável dependente – *TOBIN'S Q*

A fórmula de cálculo utilizada para o Q de Tobin, *proxy* da performance nessa pesquisa, é a mesma seguida por Gompers, Ishii e Metrick (2003) e Kaplan e Zingales (1997), conforme equação abaixo:

$$TobinQ_{i,t} = \frac{Ativos\ Totais_{i,t} + Valor\ de\ Mercado_{i,t} - PL\ Contábil_{i,t}}{Ativos\ Totais_{i,t}} \quad (1)$$

O Q de Tobin, também conhecido como *Q Ratio*, representa a razão entre o valor de mercado de uma determinada empresa e o custo de reposição de seus ativos (Chung; Pruitt, 1994). Chung e Pruitt (1994) concluem em seu trabalho, que uma das virtudes do Q de Tobin como *proxy* para performance das firmas reside no fato dessa medida ser uma razão e, portanto, ser uma medida de desempenho com um nível maior de padronização entre as diversas empresas.

A interpretação da *Q Ratio* pode ser a seguinte: se o resultado é maior do que 1, significa que a performance da firma está sendo premiada pelos seus investidores, pelo aumento do valor de mercado, em relação ao custo de substituição de seus ativos. Por outro lado, se o resultado é

menor do que 1, significa que o valor de mercado da firma não é suficiente para cobrir a substituição de seus ativos e os investidores não estão premiando a companhia. Logo, em linhas gerais, vemos que o Q de Tobin mensura performance através da indicação do quão bem a firma está usando seus ativos para gerar valor aos seus acionistas.

Assim sendo, a presente pesquisa utilizará como *proxy* para a performance o Q de Tobin (Johnson et al., 2017; Cappa et al., 2020). Assim, no contexto desse trabalho, performance da firma quer dizer: performance de mercado, medida pelo Q de Tobin.

3.4 COLETA DE DADOS

3.4.1 Variáveis de Controle e Variável Dependente

Os dados referentes às variáveis de controle foram coletados na base *Refinitiv Eikon*©, anualmente, no período de 2010 a 2022, compreendendo um total de 13 exercícios, na data de 08/03/2023.

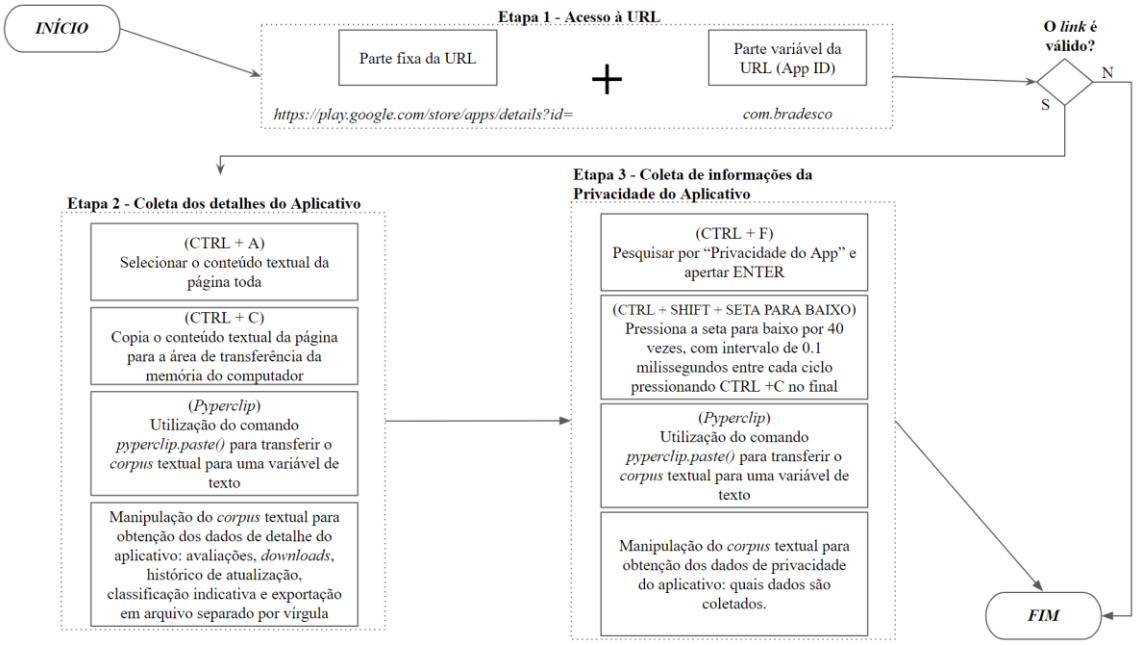
Os dados referentes às informações cadastrais das empresas de capital aberto listadas na B3 no período de 2010 a 2022 foram coletados do Portal de Dados Abertos da CVM, especificamente do *dataset* de Formulário Cadastral (FCA), que consiste em um documento eletrônico, de encaminhamento periódico e eventual, previsto no artigo 21, inciso I, da Instrução CVM nº 480/09 (CVM – FCA, 2022).

Esses dados foram utilizados como base de consulta auxiliar, principalmente na obtenção de identificadores como os *tickers* das empresas em comparação aos obtidos na base *Refinitiv Eikon*© e a vinculação com o Cadastro Nacional da Pessoa Jurídica (CNPJ).

3.4.2 Variáveis Independentes

Os dados referentes aos aplicativos disponíveis na *Google Play Store* foram encontrados na plataforma *Kaggle* e foram extraídos, a primeira vez, em junho de 2021 (Prakash e Koshy, 2021). Baseado nesse *dataset* inicial, uma nova rodada de atualização da extração foi realizada utilizando as bibliotecas do *Python*: *PyAutoGui*, *Pyperclip* e *Pandas*. A atualização automática das informações dos aplicativos foi realizada em março de 2023.

Figura 2 – Etapas automatizadas com o uso da biblioteca PyAutoGui de Robotic Process Automation (RPA)



Fonte: Elaboração própria do autor.

A Figura 2 demonstra detalhadamente cada uma das etapas de coleta das variáveis independentes utilizadas como *proxies* para o *Big Data* disponível para cada uma das empresas. Essa forma de coleta de dados, ao diminuir a necessidade de ações manuais do pesquisador, reduz a possibilidade de coletar dados de forma equivocada.

Vale ressaltar que muitas empresas possuíam mais de um aplicativo disponível na *Google Play Store*. Tendo em vista que o objetivo de coletar esses dados é a utilização como *proxy* do *Big Data*, não fazia sentido considerar mais de um aplicativo por companhia, já que dificilmente seria possível definir quantos *downloads* ou aplicativos ativos da mesma companhia um determinado usuário poderia ter (Cappa et al., 2020).

Assim, após a coleta de todos os aplicativos possíveis por empresa, a seguinte ordenação foi definida como critério: aplicativo com a data de lançamento mais antiga; em caso de empate no primeiro critério, para a mesma companhia, foi utilizado o aplicativo com a maior quantidade de *downloads.* Dessa forma, foi possível identificar o principal aplicativo de cada firma.

Para que fosse possível relacionar o aplicativo disponível na *Google Play Store* com as empresas de capital aberto da B3, foram utilizados como chave o domínio corporativo do e-mail do desenvolvedor do aplicativo na *Google Play Store* em conexão com o domínio

corporativo do e-mail disponível no setor de Relação com Investidores de cada uma das empresas. Ressalta-se que esse relacionamento não é perfeitamente exato, considerando que algumas companhias podem ter terceirizado o desenvolvimento de seus aplicativos e, como consequência, o domínio do e-mail do desenvolvedor do aplicativo pode ser divergente do domínio do setor de Relação com Investidores.

É importante ressaltar que após uma primeira execução do *script* detalhado na Figura 2, foram identificadas empresas cuja existência de aplicativos era de conhecimento prévio do pesquisador, mas que os dados não foram corretamente retornados. Procedeu-se, então, uma revisão sistemática e minuciosa do *app id* na *Google Play Store* para cada uma dessas empresas. Uma nova lista incremental de *app id* foi confeccionada, e o *script* foi executado uma segunda vez para coletar esses dados faltantes.

Para os dados coletados da rede social profissional *Linkedin*, foram realizadas duas etapas. A primeira consistiu em coletar o *link* da página corporativa de cada uma das empresas constantes na listagem obtida através da base *Refinitiv Eikon*©. A segunda etapa compreendeu em acessar cada uma das páginas corporativas, clicar no botão que demonstra o total de colaboradores da empresa, filtrar os colaboradores por palavras-chave "dados" ou "*data*" e, por último, utilizar o *software Linked Booster*© para extrair todas as páginas de colaboradores disponíveis em um arquivo separado por vírgulas (*.csv*). Essas etapas foram repetidas para cada uma das empresas.

3.5 Caracterização e tratamento dos dados

A série temporal de 2010 a 2022 coletada da base de dados Refinitiv Eikon© referente às empresas de capital aberto no Brasil consistiu em um total de 3.365 observações para 268 empresas únicas pelos respectivos *tickers*.

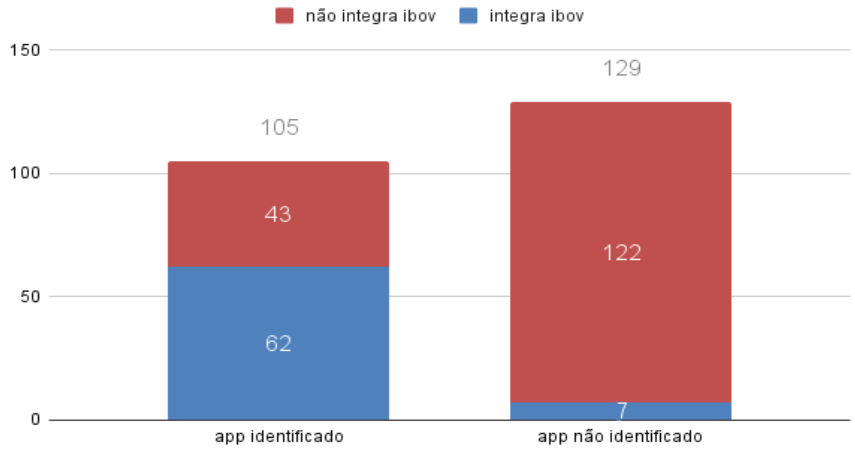
As próximas subseções apresentam as análises descritivas dos *datasets* específicos para a realização dos procedimentos estatísticos de teste para cada uma das hipóteses construídas na seção 2.3 e ilustradas na Figura 2.

Posteriormente, eventuais tratamentos necessários para adequação dos procedimentos estatísticos aos respectivos *datasets* são abordados.

3.5.1 Caracterização do *dataset* utilizado para testar as hipóteses H1, H2, H3 e H4

Para os testes envolvendo as hipóteses H1, H2, H3 e H4, foi selecionado apenas o ano de 2022, uma vez que os dados das dimensões do *Big Data*, conforme Tabela 2, dos aplicativos disponíveis na *Google Play Store*, só puderam ser coletados com alto grau de confiança uma única vez, em março de 2023. Não foi possível coletar os dados no encerramento do exercício de 2022 pois os *scripts* de coleta demonstrados, cujo fluxo foi demonstrado anteriormente na Figura 2, ainda não estavam totalmente finalizados.

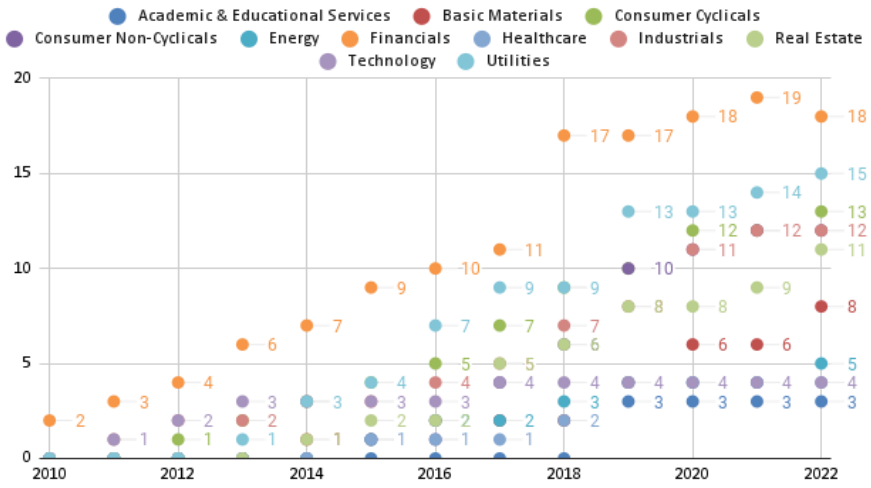
Figura 3 – Empresas brasileiras de capital aberto vs. Apps na Google Play Store 2022



Fonte: Elaboração própria do autor.

Após o filtro por ano de 2022, foram identificadas 234 empresas com 1 observação cada, excluindo fundos de investimento. A Figura 3 demonstra graficamente a divisão entre empresas integrantes do IBOV e a existência ou não de aplicativo na *Google Play Store* no *cross-section* de 2022. Foram identificados aplicativos para 44,87% das empresas das 234 integrantes dessa seleção.

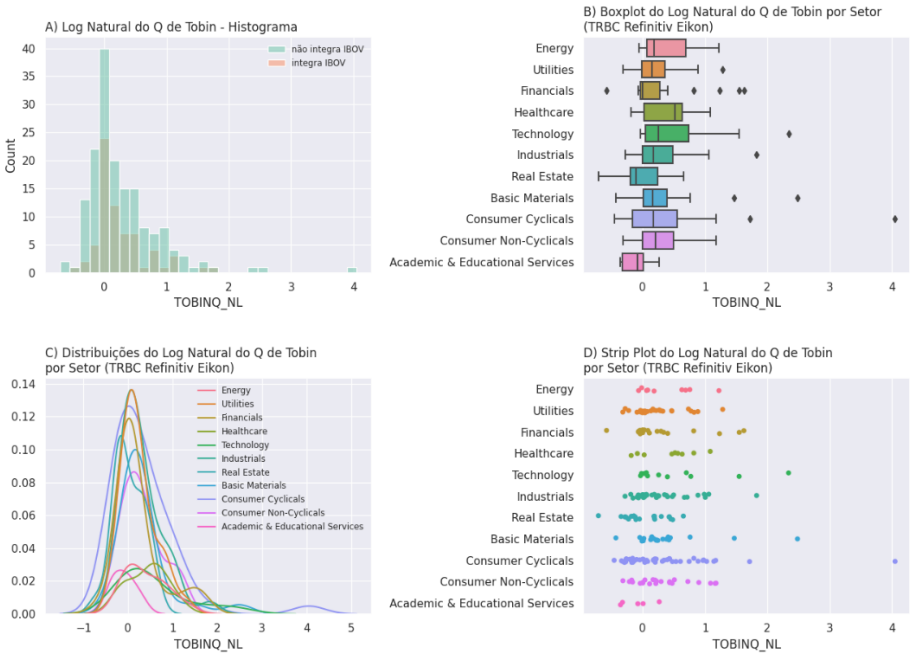
Figura 4 – Quantidade de apps por setor econômico da Refinitiv Eikon© - 2010 a 2022



Fonte: Elaboração própria do autor.

A partir da Figura 4, é possível identificar a evolução na adoção do *Big Data*, medido pela quantidade de *apps* disponíveis na *Google Play Store*. Os setores – conforme o campo categórico coletado da *Refinitiv Eikon© "The Refinitiv Business Classification – Economic Sector"* (*TRBC – Economic Sector*) – que mais se destacam na adoção de *Big Data* como estratégia de geração de valor são: *Financials*, *Utilities* e *Consumer Cyclicals*.

Figura 5 –Análise Exploratória da Variável Dependente Q de Tobin



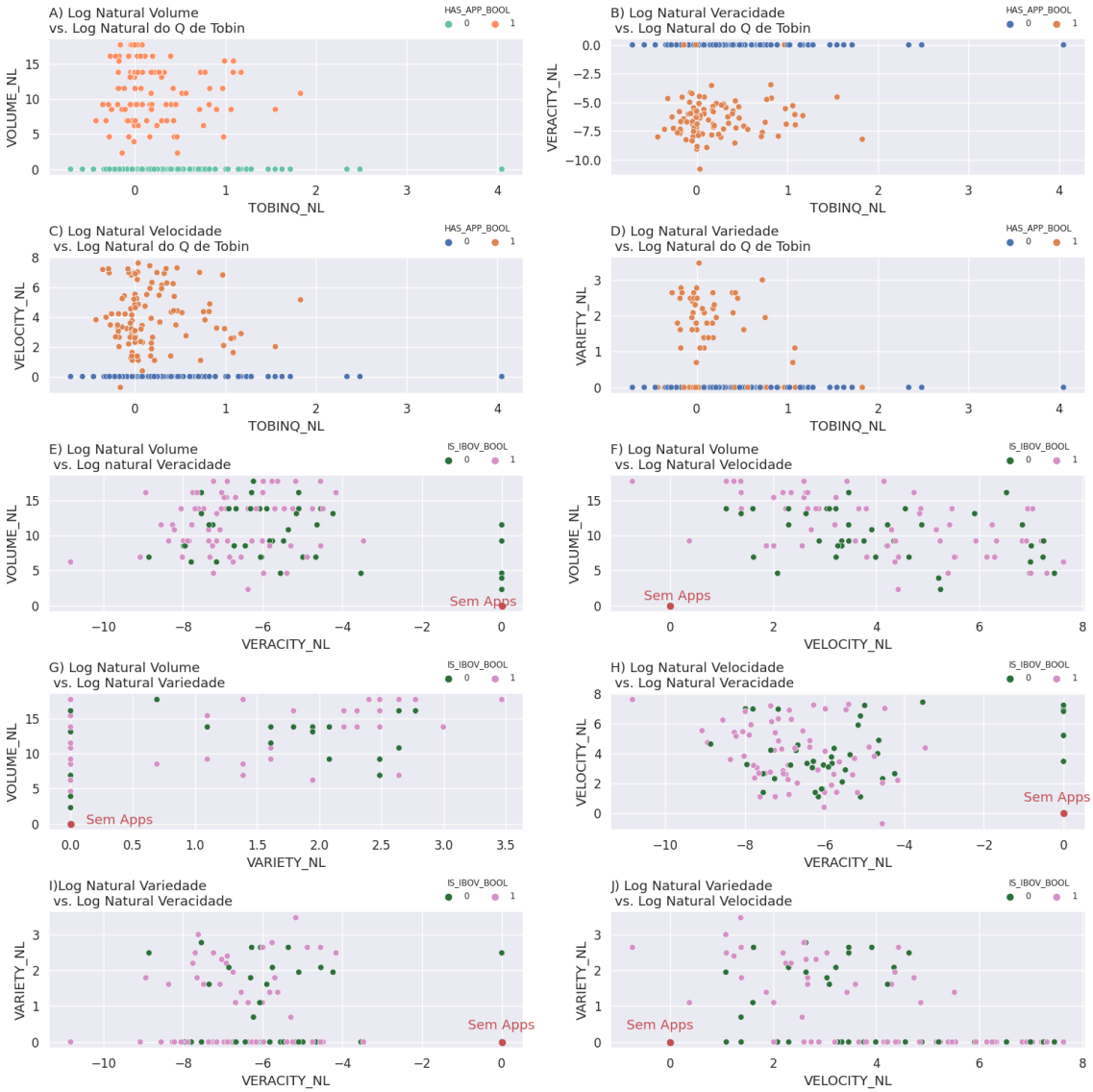
Fonte: Elaboração própria do autor.

A Figura 5 demonstra que a variável dependente do modelo empírico proposto através

da Figura 1, Q de Tobin, calculada por meio da aplicação da equação (1), possui comportamento distinto entre os setores coletados da *Refinitiv Eikon*©. Esse comportamento se difere, através do *subplot* A, quanto à frequência de valores, pelo fato das empresas integrarem ou não o índice Bovespa. Além disso, através do *subplot* B, é possível identificar a presença de valores extremos, *outliers*, principalmente no setor *Financials*. Adicionalmente, através da análise do *subplots* C e D, é possível identificar que as distribuições dos valores também variam bastante por setor.

Logo, para obtenção de um modelo estatístico adequado, algumas etapas de limpeza e tratamento da variável dependente Q de Tobin deverão ser realizadas, principalmente no que tange aos valores extremos, os denominados *outliers* inferiores e superiores.

Figura 6 – Variável Dependente Q de Tobin vs. Proxies do Big Data Variáveis Independentes



Fonte: Elaboração própria do autor.

A Figura 6 descreve graficamente as relações entre as *proxies* do *Big Data*, calculadas conforme descrições da Tabela 2, e a Variável Dependente Q de Tobin. O *subplot* A demonstra que o Volume é uma variável quantitativa discreta. Além disso, outro ponto que vale ressaltar, é que a *Google Play Store* não exibe em sua página de *app* o número exato de *downloads* de um determinado *app* e sim um número arredondado. Dessa forma, essa é uma das limitações da fase de coleta de dados descrita pela Figura 1, já que as informações coletadas por meio de *web scraping* refletem exatamente o que aparece na página.

O *subplot* B demonstra que a Veracidade é uma variável que nem sempre está presente para as firmas que possuem aplicativos. Isto é: a firma pode ter um *app* disponível na *Google Play Store* e não ter sido possível identificar colaboradores alocados em funções de Dados através da rede social *Linkedin*. Esse comportamento não era esperado e fica evidenciado por meio dessa visualização.

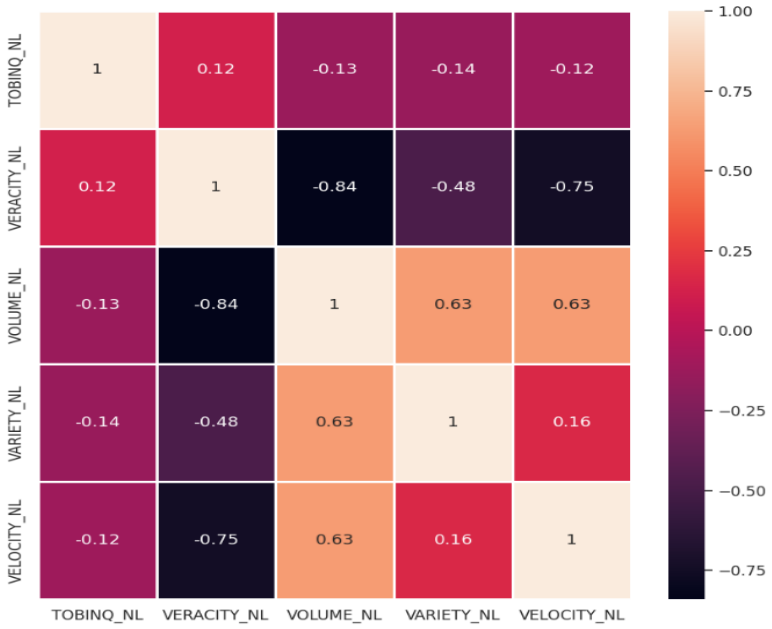
O *subplot* C demonstra que a Velocidade também possui uma enorme variância entre as empresas que possuem aplicativo lançado na *Google Play Store*. Isso quer dizer que nem todas as companhias demonstram uma preocupação de atualização dos seus aplicativos, sendo que, para essa variável, quanto menor a diferença em dias da data de coleta dos dados para a última atualização do aplicativo, melhor.

O *subplot* D demonstra que a Variedade não é um atributo que existe para todas as empresas que possuem aplicativo lançado na *Google Play Store*. Esse fato se deve principalmente pela etapa de coleta descrita pela Figura 1. Existiram dois cenários principais ocasionadores de ausência de dados para Variedade, são eles: i) algumas empresas não informam quais tipos de dados estão coletando de seus usuários, e isso não aparenta ser algo considerado negativo pela loja de aplicativos, pois apenas exibe um aviso de que o desenvolvedor do *app* não forneceu detalhes sobre os dados que coleta; ii) as empresas realmente informam que não coletam nem armazenam nenhum tipo de dados dos seus usuários, movimento que pode ser atribuído ao aumento das exigências de governança de dados no âmbito da Lei Geral de Proteção de Dados (LGPD). Tal movimento está em linha com a literatura anterior, conforme já explicitado sobre os riscos inerentes a utilização de *Big Data* pelas empresas (Cappa et al., 2020; Freund et al., 2019).

Os *subplots* de E a J demonstram as *proxies* do *Big Data* contra elas mesmas. A intenção é demonstrar, principalmente, que o Volume é a única variável presente para todas as empresas que possuem *apps* lançados na *Google Play Store*. Todas as outras dimensões do *Big Data* têm

algum nível de ausência por empresa. Além disso, em cada um dos gráficos há uma marcação demonstrando se a empresa integra ou não o índice Bovespa, separação essa que demonstrou que o fato das empresas participarem ou não desse índice não aparenta ter uma relação direta com o investimento ou não em *Big Data*. Foi marcado o ponto de todas as empresas que não possuem *apps* em destaque vermelho para que o leitor pudesse identificar a relação dos *subplots* de A a D com os de E a J.

Figura 7 – Mapa de Calor das Correlações de Pearson entre Q de Tobin e as proxies do Big Data



Fonte: Elaboração própria do autor.

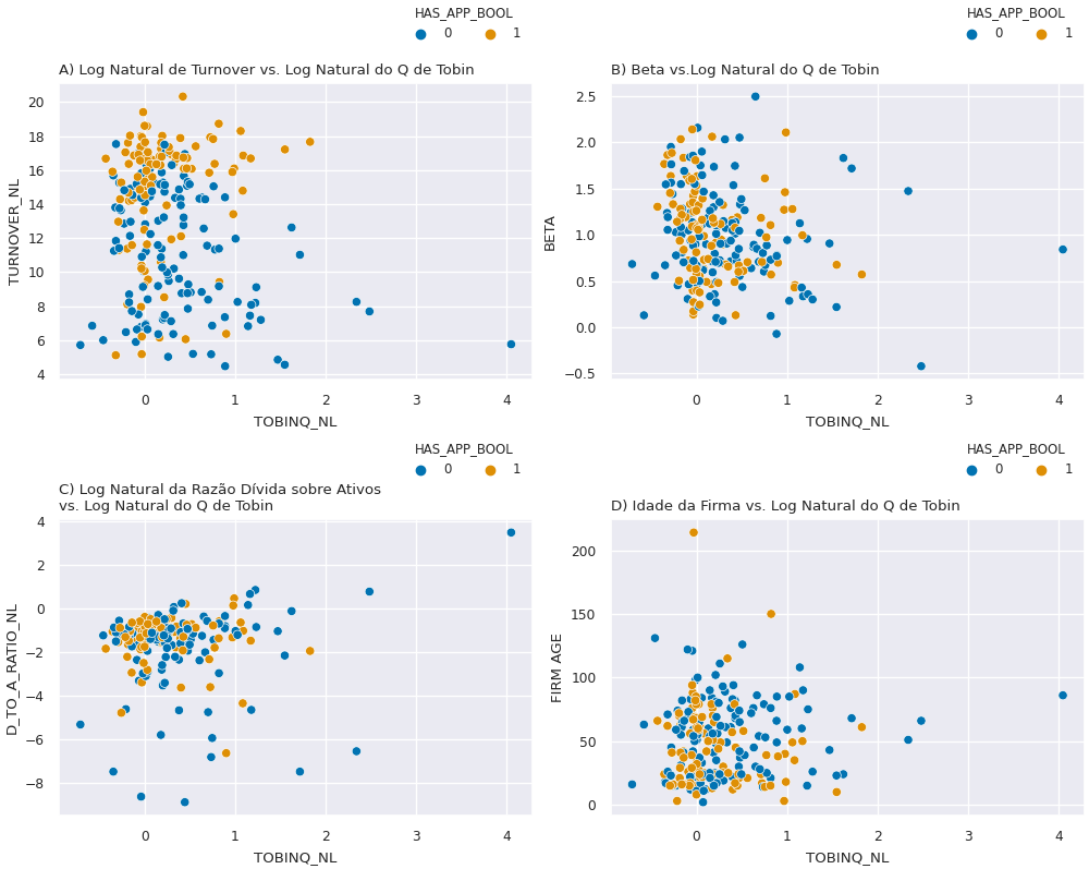
A Figura 7 demonstra visualmente as correlações, calculadas pelo método de Pearson, existentes entre as Variáveis Independentes coletadas como *proxies* para a representação do *Big Data* e a Variável Dependente Q de Tobin. É possível perceber que não existe nenhuma correlação alta, apenas valores muito tímidos (<0.50), na relação entre Q de Tobin com as dimensões do *Big Data*.

Quando se analisam as correlações de Pearson das dimensões do *Big Data* entre si, é possível perceber que Volume possui alta correlação positiva com Velocidade e Variedade. Isto quer dizer que, no *dataset* coletado nesta pesquisa, quando Volume aumenta, também

aumentam a Velocidade e a Variedade do *Big Data*, ou seja, quanto maior o número de *downloads* (Volume) do *app*, maior parece ser a diferença de dias entre a última versão do *app* disponível da loja do *Google* e o dia da coleta dos dados. E igualmente maior parece ser a quantidade de tipos de dados diferentes coletados pelo *app* (Variedade).

Outro destaque cabe à dimensão da Veracidade. Quanto maior o percentual de colaboradores atuando especificamente com dados em relação à força de trabalho total das empresas de capital aberto da B3 (veracidade), menor a Velocidade. Isto é, os *apps* se tornam mais atualizados, já que a diferença de dias entre a data da última atualização e a data da coleta das informações, execução do *script* da Figura 1, é menor. Uma correlação alta e de sentido negativo que chama atenção, é da Veracidade com o Volume. Os dados coletados nesse *dataset*, conforme *script* mencionado na Figura 1, demonstram que quanto maior o percentual de colaboradores atuando especificamente com dados em relação à força de trabalho (Veracidade), menor o Volume de dados coletados pelos *apps*.

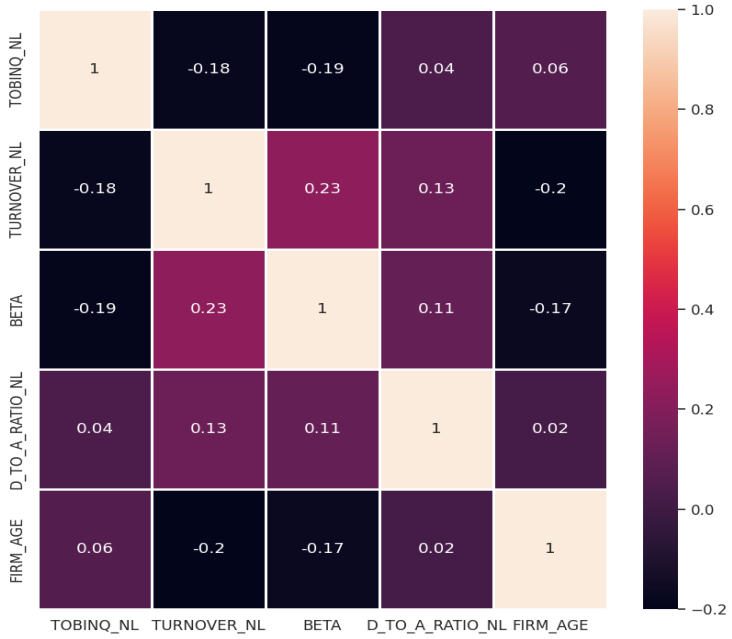
Figura 8 – Variável Dependente Q de Tobin vs. Variáveis de Controle



Fonte: Elaboração própria do autor.

A Figura 8 demonstra a relação gráfica entre o Q de Tobin e as Variáveis de Controle obtidas a partir da pesquisa de Cappa et al. (2020). É possível identificar que para todos os *subplots*, assim como na Variável Dependente e *proxies* do *Big Data* analisadas graficamente anteriormente, existe a presença de *outliers*.

Figura 9 – Mapa de Calor das Correlações de Pearson entre Q de Tobin vs. Variáveis de Controle

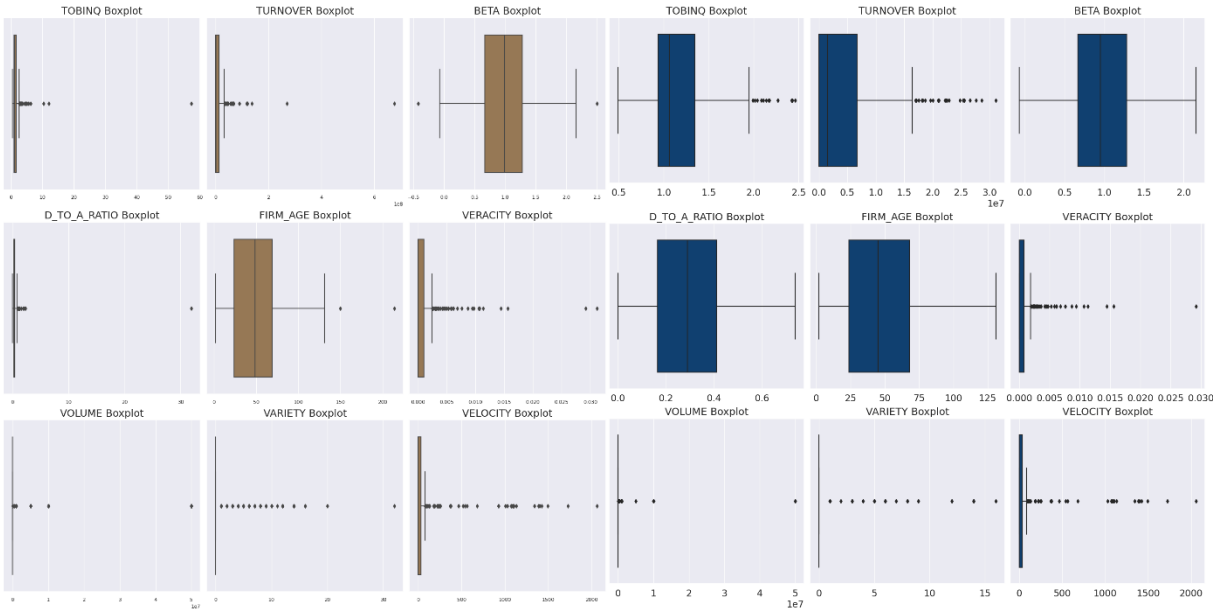


Fonte: Elaboração própria do autor.

A Figura 9 demonstra que a Variável Dependente Q de Tobin não possui uma alta correlação com nenhuma das Variáveis de Controle.

A primeira etapa, realizada na fase de pré-processamento, foi a remoção dos *outliers* da Variável Dependente Q de Tobin e das Variáveis de Controle. A Figura 10 a seguir demonstra os efeitos desse tratamento, com a cor marrom sendo antes e a cor azul sendo o estado após a sua conclusão. Essa primeira etapa gerou uma diminuição de 50 firmas, saindo de 234 para 184 observações.

Figura 10 – Boxplots Modelo 1 antes e depois da remoção dos outliers da Variável Dependente e das Variáveis de Controle



Fonte: Elaboração própria do autor

A Tabela 3 a seguir demonstra que, depois da remoção dos *outliers*, além do aspecto visual evidenciado na Figura 10 através dos *Boxplots*, 41.30% das empresas que permaneceram após o tratamento possuíam *apps* na *Google Play Store* representando 76 ocorrências num total de 184 observações.

Tabela 3 Composição do Dataset de teste das hipóteses H1 a H4 após remoção dos outliers da Variável Dependente e das Variáveis de Controle

Setor (*TRBC Economic Sector*)	Não Possui *Apps*	Possui *Apps*	Total	Não Possui *Apps* %	Possui *Apps* %	Total %
Consumer Cyclicals	25	6	31	13,59	3,26	16,85

Industrials	19	9	28	10,33	4,89	15,22
Utilities	12	13	25	6,52	7,07	13,59
Basic Materials	11	6	17	5,98	3,26	9,24
Real Estate	11	10	21	5,98	5,43	11,41
Consumer Non-Cyclicals	9	9	18	4,89	4,89	9,78
Financials	6	13	19	3,26	7,07	10,33
Healthcare	5	1	6	2,72	0,54	3,26
Technology	5	4	9	2,72	2,17	4,89
Energy	3	2	5	1,63	1,09	2,72
Academic & Educational Services	2	3	5	1,09	1,63	2,72
Total	108	76	184	58,70	41,30	100,00

Fonte: Elaboração própria do autor

Na sequência, foi realizada a transformação da Variável Dependente Q de Tobin com a aplicação do Log Natural. Esse procedimento foi realizado para tratar os *outliers* remanescentes, conforme pode ser observado no lado direito da Figura 10, no primeiro *Boxplot* azul, que ainda é notada a existência de pontos de dados discrepantes.

Ao analisar a Figura 10, também fica clara a existência de diversos *outliers* para a Variável Independente *proxy* da dimensão da Variedade do *Big Data*. Em uma investigação mais aprofundada, sobretudo com a utilização de casos de uso e reprodução do *script* evidenciado na Figura 1, conforme já comentado no momento de análise da Figura 6, a quantidade de empresas que possuía *app* e para as quais foi possível coletar a quantidade de tipos de dados diferentes coletados dos usuários – Variedade - é muito baixa.

Em seguida, todas as dimensões do *Big Data* (Volume, Variedade, Veracidade e Velocidade) passaram pelo processo de transformação em Log Natural.

Vale ressaltar, novamente, que no caso das empresas brasileiras, não foi possível coletar as despesas com Pesquisa e Desenvolvimento como foi na pesquisa internacional de Cappa et al. (2020). A adição dessa variável ao modelo poderia ajudar na redução da variância do erro sem induzir multicolinearidade.

Após a realização de todos esses procedimentos, foram estimados os seguintes modelos: Modelo Base – contendo como variáveis explicativas apenas as variáveis de controle; Modelo H1 – contendo como variáveis explicativas as variáveis de controle e adicionando apenas a dimensão do Volume (quantidade de *downloads*) do *Big Data*; Modelo H2 – contendo como variáveis explicativas as variáveis de controle e adicionando apenas a dimensão da Variedade (quantidade de tipos de dados coletados pelo *app* GPS, câmera, fotos, etc..) do *Big Data*; Modelo H3 – contendo como variáveis explicativas as variáveis de controle e adicionando apenas a dimensão da Veracidade (percentual da força de trabalho vinculada a dados em relação ao total de colaboradores da firma) do *Big Data*; Modelo H4 – contendo como variáveis

explicativas as variáveis de controle e adicionando apenas a dimensão da Velocidade (diferença em dias entre a data de coleta dos dados e a data da última atualização) do *Big Data*; Modelo Interativo – contendo como variáveis explicativas as variáveis de controle e adicionando a interação entre os 4 Vs do *Big Data* Volume, Variedade, Veracidade e Velocidade.

Todos esses modelos explicitados anteriormente foram estimados usando tanto a Regressão Linear Múltipla pelo método dos Mínimos Quadrados Ordinários, quanto utilizando a Regressão Robusta, com o estimador *M* como *Trimmed Mean* para lidar melhor com os *outliers* que foram demonstrados anteriormente. Assim sendo, há 6 modelos em cada cenário, totalizando 12 modelos.

Finalizada a etapa da caracterização do *dataset* utilizado para realizar o teste das hipóteses H1, H2, H3 e H4, a abordagem estatística utilizada é detalhada na seção 3.6.1.

3.5.2 Caracterização do *dataset* utilizado para testar a hipótese H5

A série temporal de 2010 a 2022 coletada da base de dados Refinitiv Eikon© referente às empresas de capital aberto no Brasil consistiu em um total de 3.365 observações para 268 empresas únicas pelos respectivos *tickers*.

A partir desta base, foi acrescentada a informação de existência do *app* ao decorrer da série temporal. Para um exemplo prático, é apresentada a Tabela 4 a seguir.

Tabela 4 – Exemplo da atribuição do Ano de Lançamento do App na Google Play Store na base extraída da Refinitiv Eikon© - TOTVS

Ticker	**Ano**	**Total de Ativos em USD (apenas para ilustração)**	**Ano da Data de Lançamento do App na Google Play Store**	**Quantidade de Apps**
TOTS3.SA	2010	778.548.785,63	*null*	*null*
TOTS3.SA	2011	718.230.525,58	*null*	*null*
TOTS3.SA	2012	693.831.990,23	*null*	*null*
TOTS3.SA	2013	782.784.317,05	*null*	*null*
TOTS3.SA	2014	806.874.670,68	*null*	*null*
TOTS3.SA	2015	672.360.013,13	*null*	*null*
TOTS3.SA	2016	751.233.247,26	*null*	*null*
TOTS3.SA	2017	752.883.367,05	2017	1
TOTS3.SA	2018	616.244.974,74	2017	1
TOTS3.SA	2019	879.802.687,24	2017	1
TOTS3.SA	2020	990.893.967,69	2017	1
TOTS3.SA	2021	1.785.366.138,27	2017	1
TOTS3.SA	2022	2.008.360.385,93	2017	1

Fonte: Elaboração própria do autor, relacionando dados da *Refinitiv Eikon*©

A data de lançamento do principal aplicativo de cada uma das companhias foi coletada utilizando o *script* que executa, de maneira automatizada, as etapas descritas na Figura 2. No

caso ilustrado na Tabela 4, para a empresa TOTVS, a data de lançamento obtida foi de 14/08/2017.

Dessa forma, a coluna Ano da Data de Lançamento foi preenchida como "2017", a partir deste ano em diante no *dataset*. Adicionalmente, foi inserido na coluna Quantidade de *Apps* o número 1, indicando que a companhia passou a ter um aplicativo lançado e, portanto, conforme os critérios definidos na presente pesquisa, passou a utilizar o *Big Data*.

Para as outras linhas anteriores a 2017, não foi inserido nenhum tipo de marcação indicando que o *Big Data* ainda não havia sido adotado. A mesma lógica foi realizada para todas as companhias.

Tabela 5 –Primeiros players a lançarem aplicativos na *Google Play Store*

Ticker	Nome da Empresa	Setor (*TRBC Economic Sector*)	Data de Lançamento do App
ITUB4.SA	Itau Unibanco Holding SA	Financials	18/12/2009
CIEL3.SA	Cielo SA	Industrials	12/06/2011
VIVT3.SA	Telefonica Brasil SA	Technology	27/10/2011
LREN3.SA	Lojas Renner SA	Consumer Cyclicals	09/10/2012
ENGI4.SA	Energisa SA	Utilities	10/01/2013
IGTI3.SA	Iguatemi SA	Real Estate	02/10/2014
DXCO3.SA	Dexco SA	Basic Materials	15/11/2014
ODPV3.SA	Odontoprev SA	Healthcare	31/07/2015
NTCO3.SA	Natura & Co Holding SA	Consumer Non-Cyclicals	02/12/2015
UGPA3.SA	Ultrapar Participacoes SA	Energy	09/12/2015
ANIM3.SA	Anima Holding SA	Academic & Educational Services	08/02/2019

Fonte: Elaboração própria do autor, através dos dados coletados conforme *script* descrito na Figura 1

A Tabela 5 demonstra as primeiras firmas a lançarem seus aplicativos na *Google Play Store*, por setor econômico (TRBC Economic Sector - *Refinitiv Eikon*©). O Itaú Unibanco é a firma que lançou o primeiro aplicativo e, seguindo os critérios dessa pesquisa, foi a primeira companhia a adotar uma estratégia de *Big Data*. Vale ressaltar que a *Google Play Store* em seu lançamento, em outubro/2008, se chamava *Android Market*. Logo, é realmente um destaque que, 1 ano e 2 meses depois, o Itaú já tenha sido capaz de lançar a primeira versão de seu aplicativo. Outro destaque a ser obtido da Tabela 5 é que o setor de *Academic & Educational Services* foi o último setor a ter a primeira firma, Anima Holding SA, com um aplicativo na mesma loja.

Em linhas gerais a Tabela 5 demonstra que, embora *Big Data* seja usualmente considerado um tema atual, as firmas precursoras na adoção desse tipo de estratégia já o fazem há um tempo considerável, sendo a primeira delas há exatos 13 anos e 2 meses - Itaú Unibanco – a partir do mês de coleta das informações, março de 2023.

Em sequência, as firmas listadas na Tabela 5, foram separadas em um grupo específico

de primeiras adotantes da estratégia de *Big Data (early adopters)* em seus respectivos setores. Todas as outras firmas foram marcadas como grupo de controle. Assim, o grupo das primeiras adotantes possui 10 integrantes, 1 firma por setor e o de outras companhias, com 184 empresas no espaço temporal de 2010 a 2022.

Finalizada a etapa da caracterização do *dataset* utilizado para realizar o teste da hipótese H5, a abordagem estatística do teste de média por meio de *ztest* é discutida na seção 3.6.2.

3.5.3 Caracterização do *dataset* utilizado para testar a hipótese H6

Para testar H6, que demanda informações por setor econômico e não mais por empresa, foi necessário remover 4 das 184 empresas sem *apps* na *Google Play Store*, pois elas não apareceriam em todos os anos do período. Portanto, sobraram 180 empresas sem aplicativos somados as 10 empresas da Tabela 4, que foram as precursoras na adoção do *Big Data* em seus setores, totalizando 190 empresas. A base final após remoção de eventuais linhas sem dados totalizou 1.520 observações, isto é, cada uma das 190 empresas aparecendo por 8 anos fiscais, de 2015 a 2023.

A partir dessa base de 1.520 observações, foi construído um *dataset* por setor econômico contendo os 11 setores, no período de 8 anos de 2015 a 2022, totalizando 88 observações. A agregação das linhas por empresa em linhas por setor foi realizada obtendo a média, por ano, da Variável Dependente Q de Tobin, das Variáveis de Controle e somando a quantidade de *apps* lançados de todas as empresas que compunham o setor no ano.

A hipótese H6 demanda uma agregação por setor. Assim sendo, o *dataset* produzido para a estimação deste modelo contém 88 observações totais, sendo *t* os anos no período de 2015 a 2022, e *i* os 11 setores econômicos coletados da base Refinitiv Eikon©.

Foi mantido o período de 2015 a 2022 porque o *dataset* passou por um processo de remoção das observações que estavam nulas para a Variável Dependente Q de Tobin e para as Variáveis de Controle, conforme relatado anteriormente na seção 3.5.2 do presente trabalho. Em sequência, o mesmo *dataset* passou por um processo de balanceamento.

Foi criada uma coluna que apresenta a razão do total de firmas com *apps* na *Google Play Store* em relação ao total de firmas por setor e por ano. Essa é a Variável Independente do Modelo 2 da Regressão de Linear Múltipla com dados em painel com o objetivo de avaliar a hipótese H6. A Variável Dependente é a média do Log Natural do Q de Tobin, por setor e

por ano. As Variáveis de Controle são: Log Natural do *Turnover*, Beta, Razão Dívida sobre Ativos e Idade da Firma.

Vale ressaltar mais uma vez, que a Variável de Controle de Despesas com Pesquisa e Desenvolvimento não é divulgada em número adequado pelas empresas brasileiras, o que permitiu uma replicação imperfeita do modelo de Cappa et al. (2020).

Um primeiro modelo foi estimado e apresentou *Variance Inflation Factor* (VIF) alto para a Variável de Controle Log Natural do *Turnover*, acima de 10. De modo que foi estimado um modelo removendo essa Variável de Controle para que a multicolinearidade fosse melhor controlada. Os detalhes dos valores podem ser observados no Apêndice.

Assim, foi finalizada a etapa da caracterização do *dataset* utilizado para realizar o teste da hipótese H6, cuja abordagem estatística utilizada é detalhada na seção 3.6.3.

3.6 ABORDAGEM ESTATÍSTICA

A modelagem estatística dessa pesquisa foi realizada utilizando em sua extensa maioria a biblioteca *Statsmodels* na linguagem de programação *Python* conforme o trabalho de Seabold e Perktold (2010).

As obras de Wooldrige (2002) e Fávero e Belfiore (2017) foram utilizadas como base teórica, e a de Heiss e Brunner (2020), como base técnica para aplicar as modelagens e testes estatísticos especificamente em *Python*.

3.6.1 Modelo 1 – Regressão Linear Robusta vs. Regressão Linear MQO (H1, H2, H3 e H4)

O modelo geral de Regressão Linear Múltipla utilizado para testar as hipóteses H1 até H4 é o mesmo tratado por Wooldrige (2002) e Heiss e Brunner (2020) em suas obras, sendo a equação abaixo a base:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u \qquad (2)$$

Na equação 2 acima temos que:

- y é a Variável Dependente, ou seja, a qual objetiva-se explicar. Para a finalidade dessa pesquisa, que tem como objetivo explicar a performance de mercado das companhias brasileiras de capital aberto da B3, é o Q de Tobin calculado

conforme equação 1 da seção 3.3;

- β_0 é o intercepto da equação, em outras palavras, o valor que a Variável Dependente – Q de Tobin, para essa pesquisa – assumiria caso todas as outras Variáveis Independentes fossem 0;
- $\beta_1 x_1$, indica dois termos, o primeiro, β_1, é o parâmetro associado ao efeito estimado da Variável Independente x_1 sobre a Variável Independente y. A mesma lógica se repete para cada uma das outras Variáveis Independentes do modelo, x_2 e x_3;
- $\beta_k x_k$ é apenas uma abstração de que poderia ser utilizada uma quantidade "K" de Variáveis Independentes para explicar a Variável Dependente y já que o modelo é o de regressão linear **múltipla**;
- u representa o erro ou o resíduo do modelo, isto é, a parte "sobressalente" da Variável Dependente que não é explicada pelas "K" Variáveis Independentes.

Em outras palavras, representa os fatores não observados e a variação aleatória.

Conforme visto anteriormente na seção 3.5, principalmente nas Figuras 5, 6 e 8, os dados coletados para essa pesquisa apresentam uma quantidade representativa de *outliers*, isto é, pontos de dados que influenciam fortemente o cálculo de medidas estatísticas resumidoras sobre o todo, como por exemplo a média.

Nesse sentido, para trazer robustez aos resultados dessa pesquisa e lidar melhor com os *outliers*, serão apresentados para a validação das hipóteses H1, H2, H3 e H4 o modelo de regressão linear múltipla pelo método dos mínimos quadrados ordinários conforme equação 2 anteriormente demonstrada como "Cenário A"; e o modelo de regressão linear robusta utilizando o *M-estimator* como "média aparada" (*Trimmed Mean*), que por sua vez atribuirá menor peso aos *outliers*, reduzindo suas influências na estimação dos coeficientes da regressão (β da equação 2) como "Cenário B".

Mantendo a busca por robustez nos resultados encontrados o "Cenário C" foi calculado utilizando o estimador de Newey-West de forma que os erros-padrão do modelo sejam mais consistentes em relação a heterocedasticidade e autocorrelação (Andrews, 1991).

Os gráficos de ajuste dos modelos e o cálculo do *Variance Inflation Factor* (VIF) são apresentados no Apêndice deste trabalho.

3.6.2 Teste de Média (H5)

O Teste de Média no presente estudo é o mesmo aplicado conforme as bases teóricas explicitadas na obra de Fávero e Belfiore (2017, p.214) sobre o Teste Z. Nesse sentido, abaixo são descritos os desdobramentos necessários para realização do Teste Z, sejam eles:

- **H5 Firmas que implementaram o Big Data primeiro obtiveram melhor performance**
- **População:** firmas brasileiras de capital aberto na B3 – Bovespa
- **Parâmetro de interesse:** média do Log Natural do Q de Tobin entre as empresas precursoras na adoção do *Big Data* - *early adopters* μ_1 - conforme Tabela 4, e as demais empresas - μ_2
- **H0 (hipótese nula)**: não existe diferença significante entre os dois grupos, logo $\mu_1 = \mu_2$
- **Há (hipótese alternativa)**: existe diferença significante entre os dois grupos, logo $\mu_1 \neq \mu_2$

3.6.3 Modelo 2 – Regressão Linear Múltipla MQO em dados em Painel (H6)

Para a estimação do Modelo 2 de Regressão Linear Múltipla pelo método dos mínimos quadrados ordinários (MQO) em dados em painel, serão utilizadas as definições de Heiss e Brunner (2020): "Um conjunto de dados em painel inclui várias observações em diferentes pontos no tempo *t* para o mesmo conjunto de unidades transversais *i*" (Heiss e Brunner, 2020, p. 233).

Assim apresenta-se a equação básica de um modelo de Regressão Linear Múltipla MQO no formato de dados em painel é a seguinte:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + \beta_3 x_{it3} + \cdots + \beta_k x_{itk} + u_{it;}$$
$$t = 1,\ldots,T;$$
$$i = 1,\ldots,n \qquad (4)$$

O paralelo da equação básica 4, para Regressão Linear Múltipla com dados em painel, com a equação básica 2, para Regressão Linear Múltipla, é natural. As únicas alterações que existem se referem a inclusão dos termos *t*, representando o tempo, e *i*, representando o indivíduo.

Os gráficos de ajuste dos modelos e o cálculo do *Variance Inflation Factor* (VIF) são apresentados no Apêndice deste trabalho.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Primeiramente, apresentam-se os resultados de cada uma das hipóteses de pesquisa obtidos através das etapas metodológicas aplicadas e, em seguida, uma avaliação em conjunto de todas as hipóteses, bem como o relacionamento dos achados com trabalhos anteriores e com a base teórica escolhida da Visão Baseada em Recursos (RBV).

4.1 Hipóteses H1, H2, H3 e H4

Conforme disposto na seção de procedimentos metodológicos, para a avaliação das hipóteses de H1 a H4, foram implementados 6 modelos em dois cenários diferentes: Cenário A – por meio de Regressão Linear Múltipla pelo método dos mínimos quadrados ordinários (MQO); Cenário B – por meio de Regressão Robusta para melhor tratamento dos *outliers* identificados na etapa de caracterização do conjunto de dados coletados, por meio do estimador M *Trimmed Mean*, que diminui os efeitos dos *outliers* nos parâmetros e na reta resultante da regressão; Cenário C – por meio de Regressão Linear Múltipla pelo método dos mínimos quadrados ordinários (MQO) utilizando o estimador de Newey-West com o objetivo de trazer consistência com relação a heterocedasticidade e autocorrelação.

A partir da análise das Tabelas e Figuras inseridas no Apêndice B do presente trabalho, é possível identificar que os modelos estimados para testar as hipóteses H1 até H4 estão relativamente bem ajustados e a multicolinearidade está controlada, já que em todos os casos o resultado do *Variance Inflation Factor* (VIF) de cada uma das variáveis utilizadas foi inferior a 10 e a média inferior a 5.

A Tabela 6 apresenta o resultado dos modelos estimados para analisar as hipóteses H1 a H4. O Quadro A demonstra os resultados obtidos através de Regressão Linear Múltipla pelo método dos mínimos quadrados ordinários. O Quadro B demonstra os resultados obtidos através de Regressão Robusta. O Quadro C demonstra os resultados obtidos através de Regressão Linear Múltipla pelo método dos mínimos quadrados ordinários utilizando o estimador de Newey-West com o objetivo de trazer consistência com relação a heterocedasticidade e autocorrelação. Vale ressaltar que, ao controlar o setor das companhias conforme o *The Refinitiv Eikon© Business Classification* (TRBC Economic Sector), o setor *Academic & Educational Services* apresenta seu impacto no intercepto das regressões e encontra-se omitido.

O Modelo Base foi estimado em todos os Quadros da Tabela 6, utilizando exclusivamente as variáveis de controle. Vale destacar, novamente, a limitação da realidade de mercado no Brasil no que tange a divulgação dos gastos com Pesquisa e Desenvolvimento por parte das companhias de capital aberto. Devido à escassez de dados sobre essa variável de controle, ela não foi incluída nos modelos.

O Modelo H1 avalia se o volume do *Big Data* interfere negativamente na performance das firmas de capital aberto no Brasil em 2022. Em todos os cenários de estimação, Quadros A, B e C da Tabela 6, existiu significância estatística a 10% e 5%, respectivamente, para essa variável de interesse. Esses resultados corroboram a pesquisa de Cappa et al. (2020), a qual também observou que o Volume interferiu negativamente na performance das firmas na realidade das empresas de capital aberto nos Estados Unidos integrantes do índice S&P500 para o recorte do ano de 2018.

Chama atenção o fato de que na presente pesquisa, para a realidade das companhias de capital aberto no Brasil em 2022, o coeficiente possuir significância estatística, mas o seu valor ser muito menor do que o obtido por Cappa et al. (2020). Outro ponto que vale ressaltar, é o aumento do poder explicativo entre o Modelo Base, sem a dimensão do volume do *Big Data*, e o Modelo H1 incluindo essa variável medida através do *Rsquared*, comportamento também alinhado à pesquisa de Cappa et al. (2020).

O Modelo H2 avalia se a variedade do *Big Data* interfere positivamente na performance das firmas de capital aberto no Brasil em 2022. Em todos os cenários de estimação, Quadros A, B e C da Tabela 6, não existiu significância estatística a 10% e 5%, respectivamente, para essa variável de interesse. Esses resultados divergem da pesquisa de Cappa et al. (2020), a qual observou que a Variedade interferiu positivamente na performance das firmas na realidade das empresas de capital aberto nos Estados Unidos integrantes do índice S&P500 para o recorte do ano de 2018.

O Modelo H3 avalia se a veracidade do *Big Data* interfere positivamente na performance das firmas de capital aberto no Brasil em 2022. Quando os *outliers* não estão sendo controlados, Quadro A da Tabela 6, não existiu significância estatística para essa variável de interesse. Quando os *outliers* são controlados, Quadro B da Tabela 6, existiu significância estatística a 10% para essa variável de interesse. Quando o modelo foi calculado utilizando o estimador de Newey-West, usado como teste de robustez para a presente pesquisa, novamente não foi identificada significância estatística. A diferença da significância estatística quando há o controle dos eventuais *outliers* pode ser oriunda do fato de que as firmas possuem estruturas muito diferentes no que tange ao investimento em tecnologia. Mesmo sem ser

possível controlar os investimentos em Pesquisa e Desenvolvimento, durante o processo de coleta de dados foi possível observar que poucas firmas detinham muitos colaboradores alocados em funções de dados. Adicionalmente, o fato de não ser observada significância estatística quando há utilização do estimador de Newey-West, identifica que a veracidade do *Big Data* pode ter um efeito instável ao longo da série temporal utilizada na pesquisa. Isto é, possivelmente no Quadro A da Tabela 6 o coeficiente está inflado por autocorrelação ou heterocedasticidade.

Esses resultados divergem da pesquisa de Cappa et al. (2020), a qual observou que a Veracidade interferiu positivamente na performance das firmas na realidade das empresas de capital aberto nos Estados Unidos integrantes do índice S&P500 para o recorte do ano de 2018.

O Modelo H4 avalia se a Velocidade do *Big Data* interfere positivamente na performance das firmas de capital aberto no Brasil em 2022. Em todos os cenários de estimação, Quadros A, B e C da Tabela 6, não existiu significância estatística a 10% e 5%, respectivamente, para essa variável de interesse. Essa dimensão do *Big Data* foi proposta em adição à pesquisa de Cappa et al. (2020), base deste trabalho e, portanto, representa um novo resultado para os quais não há base comparativa anterior.

Por fim, o Modelo Interativo avalia se a interação entre as 4 dimensões independentes do *Big Data* – Volume, Variedade, Veracidade e Velocidade - em conjunto, afetam a performance das firmas de capital aberto no Brasil em 2022. Nos cenários de estimação, Quadro A e Quadro B da Tabela 6, não existiu significância estatística a 10% e 5%, respectivamente, para essa variável de interesse. Porém no Quadro C da Tabela 6, ao utilizar o estimador de Newey-West no cálculo do modelo foi identificada a significância estatística para a interação dos 4 V's do *Big Data* na performance de mercado das companhias de capital aberto no Brasil. Esse fato indica que a heterocedasticidade e a autocorrelação presentes nos modelos estimados nos Quadros A e B reduziram o efeito da variável. De qualquer forma, vale notar que apesar da significância estatística, o coeficiente apresenta valor muito baixo.

Em linhas gerais, ao comparar os resultados obtidos na presente pesquisa e os resultados obtidos por Cappa et al. (2020), as dimensões do *Big Data* aparentam ter reduzido a sua significância estatística em termos de impacto na performance das firmas. Para além das variáveis exógenas que diferenciam o mercado brasileiro do mercado americano, há também a variável de tempo. Os dados incluídos no modelo da pesquisa de Cappa et al. (2020) são de 2018 enquanto os dessa pesquisa são de 2022. Em um período de 4 anos, muito pode ser feito e implementado pelas firmas.

Logo, à luz da RBV, na medida em que a curva de adoção da estratégia de *Big Data* foi aumentando ao longo do tempo pelas empresas de capital aberto do Brasil, a raridade desse recurso foi diminuindo e, por consequência, a heterogeneidade entre os *players* também. Assim sendo, os resultados das hipóteses podem ter sido afetados por esse movimento.

Nesse sentido, apesar das hipóteses H2 até H4 terem sido rejeitadas para o cenário das companhias brasileiras de capital aberto no Brasil, parte da explicação para esses resultados terem sido observados advém da lente teórica. Tal explicação teórica pode ser traduzida em dados através da análise da Tabela 3, a qual demonstra que 41,30% das empresas na B3 já estavam fazendo uso da estratégia de *Big Data* em 2022, logo esse recurso não tem mais uma alta raridade.

Outrossim, vale ressaltar que, no período de 2018 a 2022, muitas mudanças estruturais aconteceram do ponto de vista regulatório que podem ter afetado a capacidade de geração de valor das firmas através da adoção de estratégias de *Big Data*.

No mercado europeu, foi lançada a *General Data Protection Law* (GDPR) em maio de 2018. No mercado dos Estados Unidos, foi lançada a *California Consumer Privacy Act* (CCPA) em junho de 2018. No mercado brasileiro, foi lançada a Lei Geral de Proteção de Dados Pessoais (LGPD), em agosto de 2018. Essas regulações, em linhas gerais, traçam um limite até onde as firmas poderão chegar no quesito obtenção de dados pessoais de seus clientes.

Logo, essas regulamentações podem ter afetado a eficiência da estratégia de *Big Data* para além do efeito gerado pela maior possibilidade de replicação pelos concorrentes. Podendo gerar hipóteses para pesquisas futuras no que tange ao aprofundamento das divergências que foram observadas entre a presente pesquisa e o trabalho de Cappa et al. (2020), já que os recortes temporais são 2022 e 2018, respectivamente.

Fato que apresenta uma direção nesse sentido, é que o *script* adotado na Figura 2 ao ser direcionado para a coleta de tipos de dados aos quais as companhias pediam acesso em seus *apps* – dimensão Variedade do *Big Data* - identificou muitos casos nos quais as empresas afirmam que não coletavam mais dados pessoais de seus consumidores, indicando um impacto dessas mudanças regulatórias em 2022.

Tabela 6 –Resultados dos Modelos Estimados para avaliação de H1 até H4

Quadro A - Resultados da Regressão Linear Múltipla pelo Método dos Mínimos Quadrados Ordinários

	Modelo Base *(variáveis de controle)*		**Modelo H1** *(variáveis de controle + volume)*		**Modelo H2** *(variáveis de controle + variedade)*		**Modelo H3** *(variáveis de controle + veracidade)*		**Modelo H4** *(variáveis de controle + velocidade)*		**Modelo Interativo** *(variáveis de controle + 4v's)*	
Variáveis	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**
Intercept	0,1400	0,4463	0,1172	0,5214	0,1331	0,4718	0,1335	0,4673	0,1455	0,4289	0,1239	0,5022
TURNOVER_NL	-0,0001	0,9916	0,0036	0,6116	0,0003	0,9661	0,0024	0,7334	0,0018	0,7995	0,0009	0,8993
BETA	-0,1419	0,0118[ab]	-0,1377	0,0139[ab]	-0,1400	0,0135[ab]	-0,1443	0,0105[ab]	-0,1462	0,0098[ab]	-0,1389	0,0139[ab]
D_TO_A_RATIO	-0,0772	0,5724	-0,0386	0,7783	-0,0681	0,6236	-0,0469	0,7355	-0,0547	0,6936	-0,0575	0,6781
FIRM_AGE	0,0000	0,9775	-0,0001	0,9383	0,0000	0,9724	-0,0001	0,9268	-0,0001	0,8969	0,0000	0,9788
VOLUME_NL			-0,0083	0,0593[b]								
VARIETY_NL					-0,0144	0,6678						
VERACITY_NL							0,0092	0,2352				
VELOCITY_NL									-0,0096	0,3435		
VOLUME_NL_i_VARIETY_NL_i_VERACITY_NL_i_VELOCITY_NL											0,0001	0,3490
TRBC_ECONOMIC_SECTOR[T.Basic Materials]	0,1573	0,3222	0,1381	0,3821	0,1565	0,3259	0,1487	0,3491	0,1484	0,3514	0,1544	0,3315
TRBC_ECONOMIC_SECTOR[T.Consumer Cyclicals]	0,1363	0,3613	0,1273	0,3906	0,1385	0,3549	0,1208	0,4194	0,1219	0,4168	0,1409	0,3459
TRBC_ECONOMIC_SECTOR[T.Consumer Non-Cyclicals]	0,1613	0,3080	0,1627	0,3002	0,1650	0,2989	0,1661	0,2934	0,1533	0,3334	0,1731	0,2758
TRBC_ECONOMIC_SECTOR[T.Energy]	0,2993	0,1225	0,3005	0,1182	0,3063	0,1162	0,2927	0,1306	0,2752	0,1592	0,3082	0,1124
TRBC_ECONOMIC_SECTOR[T.Financials]	0,0228	0,8843	0,0630	0,6875	0,0340	0,8307	0,0340	0,8278	0,0187	0,9048	0,0372	0,8128
TRBC_ECONOMIC_SECTOR[T.Healthcare]	0,3764	0,0456[ab]	0,3530	0,0592[b]	0,3782	0,0451[ab]	0,3550	0,0600[b]	0,3526	0,0633[b]	0,3811	0,0431[ab]
TRBC_ECONOMIC_SECTOR[T.Industrials]	0,2263	0,1267	0,2120	0,1498	0,2245	0,1309	0,2161	0,1448	0,2163	0,1453	0,2218	0,1347
TRBC_ECONOMIC_SECTOR[T.Real Estate]	0,0335	0,8252	0,0284	0,8504	0,0371	0,8076	0,0292	0,8471	0,0274	0,8566	0,0394	0,7951
TRBC_ECONOMIC_SECTOR[T.Technology]	0,2842	0,0973[b]	0,3079	0,0712[b]	0,2847	0,0975[b]	0,2780	0,1045	0,2763	0,1075	0,2862	0,0951[b]
TRBC_ECONOMIC_SECTOR[T.Utilities]	0,1546	0,3101	0,1618	0,2849	0,1558	0,3077	0,1516	0,3191	0,1451	0,3420	0,1548	0,3096
No. Observations	184		184		184		184		184		184	
Rsquared	0,1404		0,1585		0,1414		0,1476		0,1450		0,1449	
Breusch-Pagan	0,0000810		0,0000601		0,0000953		0,0000910		0,0001317		0,0000861	
Shapiro-Wilk p-value	0,0151416		0,0481450		0,0191203		0,0350259		0,0245254		0,0294460	

Continua

Quadro B - Resultados da Regressão Robusta (Trimmed Mean)

	Modelo Base *(variáveis de controle)*		**Modelo H1** *(variáveis de controle + volume)*		**Modelo H2** *(variáveis de controle + variedade)*		**Modelo H3** *(variáveis de controle + veracidade)*		**Modelo H4** *(variáveis de controle + velocidade)*		**Modelo Interativo** *(variáveis de controle + 4v's)*	
Variáveis	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**
Intercept	0,0552	0,7209	-0,0064	0,9669	0,0228	0,8857	0,0053	0,9727	0,0252	0,8717	-0,0203	0,8957
TURNOVER_NL	0,0017	0,7680	0,0074	0,2081	0,0047	0,4208	0,0070	0,2430	0,0055	0,3612	0,0067	0,2461
BETA	-0,1205	0,0103[ab]	-0,1198	0,0103[ab]	-0,1224	0,0111[ab]	-0,1272	0,0069[ab]	-0,1278	0,0071[ab]	-0,1250	0,0079[ab]
D_TO_A_RATIO	-0,0473	0,6810	0,0351	0,7611	-0,0414	0,7282	0,0389	0,7402	0,0171	0,8842	-0,0048	0,9668
FIRM_AGE	0,0006	0,4379	0,0005	0,4744	0,0004	0,6036	0,0005	0,4976	0,0005	0,5400	0,0007	0,3595
Variáveis	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**
VOLUME_NL			-0,0090	0,0141[ab]								
VARIETY_NL					-0,0167	0,5616						
VERACITY_NL							0,0126	0,0533[b]				
VELOCITY_NL									-0,0108	0,2067		
VOLUME_NL_i_VARIETY_NL_i_VERACITY_NL_i_VELOCITY_NL											0,0001	0,2825
TRBC_ECONOMIC_SECTOR[T.Basic Materials]	0,1123	0,4001	0,0981	0,4604	0,1167	0,3928	0,1086	0,4171	0,1098	0,4152	0,1075	0,4203
TRBC_ECONOMIC_SECTOR[T.Consumer Cyclicals]	0,0896	0,4751	0,0893	0,4738	0,1009	0,4321	0,0780	0,5364	0,0829	0,5145	0,0815	0,5158
TRBC_ECONOMIC_SECTOR[T.Consumer Non-Cyclicals]	0,1233	0,3532	0,1213	0,3582	0,1283	0,3461	0,1262	0,3431	0,1143	0,3941	0,1289	0,3333
TRBC_ECONOMIC_SECTOR[T.Energy]	0,2975	0,0669[b]	0,3019	0,0615[b]	0,3091	0,0639[b]	0,2921	0,0728[b]	0,2741	0,0971[b]	0,3040	0,0614[b]
TRBC_ECONOMIC_SECTOR[T.Financials]	0,0686	0,6022	0,1237	0,3485	0,0887	0,5156	0,0953	0,4704	0,0739	0,5776	0,0888	0,5015
TRBC_ECONOMIC_SECTOR[T.Healthcare]	0,4891	0,0019[ab]	0,3616	0,0211[ab]	0,3911	0,0152[ab]	0,3584	0,0236[ab]	0,3613	0,0241[ab]	0,3888	0,0135[ab]
TRBC_ECONOMIC_SECTOR[T.Industrials]	0,1760	0,1563	0,1663	0,1784	0,2020	0,1122	0,1676	0,1787	0,1702	0,1753	0,1764	0,1555
TRBC_ECONOMIC_SECTOR[T.Real Estate]	0,0793	0,5342	0,0814	0,5209	0,0848	0,5163	0,0823	0,5198	0,0806	0,5313	0,0908	0,4767
TRBC_ECONOMIC_SECTOR[T.Technology]	0,2949	0,0399[ab]	0,3298	0,0211[ab]	0,2966	0,0435[ab]	0,2966	0,0392[ab]	0,2951	0,0417[ab]	0,3025	0,0350[ab]
TRBC_ECONOMIC_SECTOR[T.Utilities]	0,0816	0,5233	0,0940	0,4600	0,0842	0,5200	0,0819	0,5229	0,0748	0,5630	0,0792	0,5356
No. Observations	184		184		184		184		184		184	
Breusch-Pagan	0,0000382		0,0000371		0,0000431		0,0000657		0,0000791		0,0000433	
Shapiro-Wilk p-value	0,0045282		0,0099696		0,0043077		0,0088514		0,0034802		0,0032970	

Continua

Quadro C - Resultados da Regressão Linear Múltipla pelo Método dos Mínimos Quadrados Ordinários (com estimador Newey-West)

	Modelo Base *(variáveis de controle)*		**Modelo H1** *(variáveis de controle + volume)*		**Modelo H2** *(variáveis de controle + variedade)*		**Modelo H3** *(variáveis de controle + veracidade)*		**Modelo H4** *(variáveis de controle + velocidade)*		**Modelo Interativo** *(variáveis de controle + 4v's)*	
Variáveis	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**
Intercept	0,1400	0,3800	0,1172	0,4615	0,1331	0,4027	0,1335	0,4036	0,1455	0,3623	0,1239	0,4373
TURNOVER_NL	-0,0001	0,9921	0,0036	0,6348	0,0003	0,9676	0,0024	0,7570	0,0018	0,8132	0,0009	0,9038
BETA	-0,1419	0,0131[ab]	-0,1377	0,0116[ab]	-0,1400	0,0140[ab]	-0,1443	0,0106[ab]	-0,1462	0,0102[ab]	-0,1389	0,0138[ab]
D_TO_A_RATIO	-0,0772	0,5870	-0,0386	0,7876	-0,0681	0,6389	-0,0469	0,7544	-0,0547	0,7123	-0,0575	0,6910
FIRM_AGE	0,0000	0,9766	-0,0001	0,9365	0,0000	0,9713	-0,0001	0,9237	-0,0001	0,8921	0,0000	0,9781
VOLUME_NL			-0,0083	0,0179[ab]								
VARIETY_NL					-0,0144	0,5232						
VERACITY_NL							0,0092	0,1928				
VELOCITY_NL									-0,0096	0,2901		
VOLUME_NL_i_VARIETY_NL_i_VERACITY_NL_i_VELOCITY_NL											0,0001	0,0848[b]
Variáveis	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**	**Coeficientes**	**P-valores**
TRBC_ECONOMIC_SECTOR[T.Basic Materials]	0,1573	0,1600	0,1381	0,2113	0,1565	0,1640	0,1487	0,1749	0,1484	0,1771	0,1544	0,1717
TRBC_ECONOMIC_SECTOR[T.Consumer Cyclicals]	0,1363	0,2079	0,1273	0,2402	0,1385	0,2016	0,1208	0,2689	0,1219	0,2663	0,1409	0,1964
TRBC_ECONOMIC_SECTOR[T.Consumer Non-Cyclicals]	0,1613	0,2338	0,1627	0,2259	0,1650	0,2221	0,1661	0,2143	0,1533	0,2592	0,1731	0,2049
TRBC_ECONOMIC_SECTOR[T.Energy]	0,2993	0,0806[b]	0,3005	0,0667[b]	0,3063	0,0717[b]	0,2927	0,0781	0,2752	0,1112	0,3082	0,0678
TRBC_ECONOMIC_SECTOR[T.Financials]	0,0228	0,8326	0,0630	0,5521	0,0340	0,7541	0,0340	0,7466	0,0187	0,8617	0,0372	0,7301
TRBC_ECONOMIC_SECTOR[T.Healthcare]	0,3764	0,0346[ab]	0,3530	0,0474[ab]	0,3782	0,0338[ab]	0,3550	0,0474[ab]	0,3526	0,0502[b]	0,3811	0,0331[ab]
TRBC_ECONOMIC_SECTOR[T.Industrials]	0,2263	0,0242[ab]	0,2120	0,0339[ab]	0,2245	0,0264[ab]	0,2161	0,0300[ab]	0,2163	0,0303[ab]	0,2218	0,0289[ab]
TRBC_ECONOMIC_SECTOR[T.Real Estate]	0,0335	0,7579	0,0284	0,7923	0,0371	0,7341	0,0292	0,7851	0,0274	0,7971	0,0394	0,7205
TRBC_ECONOMIC_SECTOR[T.Technology]	0,2842	0,0514[b]	0,3079	0,0286[ab]	0,2847	0,0510[b]	0,2780	0,0547[b]	0,2763	0,0571[b]	0,2862	0,0500[ab]
TRBC_ECONOMIC_SECTOR[T.Utilities]	0,1546	0,1523	0,1618	0,1324	0,1558	0,1497	0,1516	0,1544	0,1451	0,1786	0,1548	0,1533
No. Observations	184		184		184		184		184		184	
Rsquared	0,14040000		0,15850000		0,14140000		0,14760000		0,14500000		0,14490000	
Breusch-Pagan	0,00008101		0,00006007		0,00009535		0,00009103		0,00013172		0,00008608	
Shapiro-Wilk p-value	0,01514165		0,04814498		0,01912032		0,03502591		0,02452537		0,02944596	

a: p valor <=0.05; **b**: p valor <= 0.10

Fonte: Elaboração própria do autor.

Uma vez que foi possível coletar, através do *script* descrito na Figura 2, a data de lançamento de cada um dos aplicativos considerados para as empresas, conforme detalhado nas Tabelas 4 e 5, se torna interessante analisar o *Big Data* enquanto recurso ao longo do tempo e a sua aderência aos preceitos da Teoria da Visão Baseada em Recursos (RBV), conforme discussões anteriores na seção 2.1.

Assim, as hipóteses H5 e H6 a seguir buscam realizar testes com *datasets* no formato de Série Temporal e de Dados em Painel, e não mais *cross-section* focando apenas no período mais atual.

4.2 Hipótese H5

A Tabela 7 a seguir demonstra os resultados dos cálculos do Teste Z e os p-valores ao nível de significância de 5% e 10%, respectivamente. Os resultados foram obtidos usando o método *stats.ztest* do pacote Statsmodels do *Python*.

Tabela 7 – Resultados do Teste Z de Média por ano (H5)

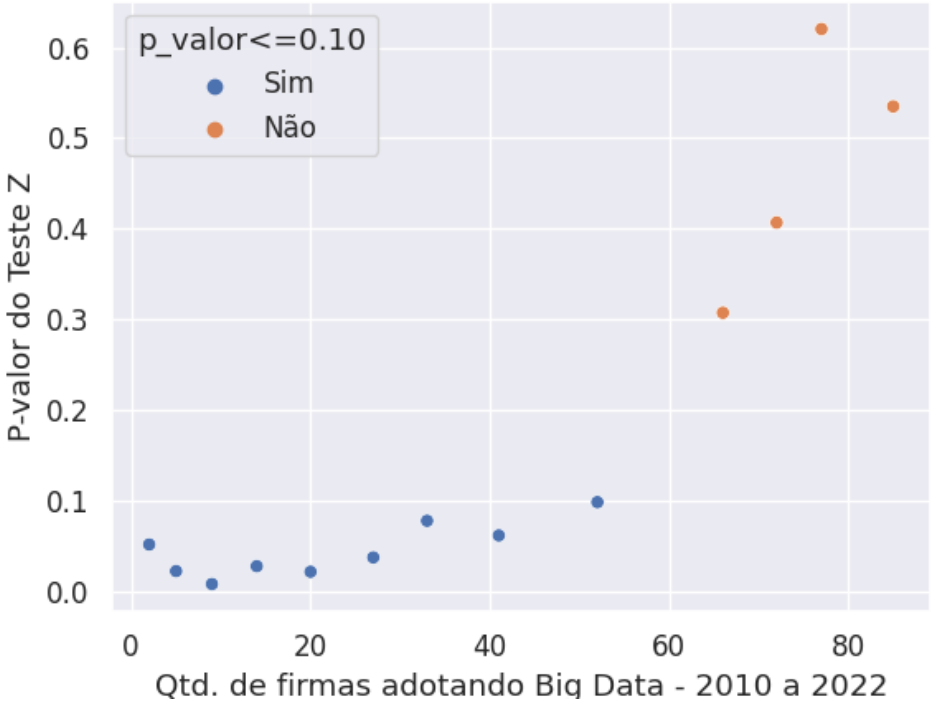
Ano	Quantidade de empresas usando *Big Data* (*app* existente na *Google Play Store*)	Média do Log Natural do Q de Tobin - Grupo 1 *(early adopters)*	Média do Log Natural do Q de Tobin - Grupo 2 *(non-early adopters)*	Estatística do Teste Z	P-valor
2010	2	0,8569	0,4510	-1,9479	0,0514[b]
2011	5	0,7106	0,2550	-2,2896	0,0220[ab]
2012	9	0,8034	0,2764	-2,6617	0,0078[ab]
2013	14	0,6948	0,2252	-2,2048	0,0275[ab]
2014	20	0,6047	0,1447	-2,3025	0,0213[ab]
2015	27	0,4514	0,0044	-2,0842	0,0371[ab]
2016	33	0,5657	0,1986	-1,7647	0,0776[ab]
2017	41	0,6129	0,2682	-1,8705	0,0614[b]
2018	52	0,5098	0,2265	-1,6538	0,0982[b]
2019	66	0,6318	0,4502	-1,0210	0,3073
2020	72	0,5121	0,3606	-0,8295	0,4068
2021	77	0,2602	0,3503	0,4951	0,6205
2022	85	0,1760	0,2863	0,6203	0,5350

a: p valor <=0.05; **b**: p valor <= 0.10
Fonte: Elaboração própria do autor.

É possível notar uma certa relação de linearidade entre a quantidade de *players* que passaram adotar o *Big Data* (para essa pesquisa, lançamento do *app* na *Google Play Store*) e a queda da significância estatística do p-valor referente ao Teste Z, tanto ao nível de 95% de confiança, quanto ao nível de 90% de confiança. Isso faz bastante sentido com os preceitos da Teoria da Visão Baseada em Recursos (RBV), uma vez que na medida em que um recurso capaz de gerar vantagem competitiva sustentável passa a ser menos raro e os concorrentes conseguem replicar os seus benefícios, os chamados *early adopters* passam a obter cada vez menos diferencial na execução da mesma estratégia.

A partir da Tabela 7 foi construído o gráfico da Figura 11 a seguir para resumir visualmente o comportamento da relação entre o p-valor observado ao realizar o Teste Z de Média para avaliar a hipótese H5, e a quantidade crescente de empresas adotando *Big Data*, isto é, lançando aplicativos na *Google Play Store*.

Figura 11 – P-valores do Teste Z de Média vs. Adoção do Big Data (H5)



Fonte: Elaboração própria do autor.

Relacionando a Figura 11 com a Teoria da Visão Baseada em Recursos (RBV) para o período de 2010 a 2022, é possível confirmar aspectos mencionados por Peteraf (1993) no que tange a importância da manutenção da baixa oferta dos recursos superiores, pois a vantagem

competitiva sustentável existe quando tais recursos superiores não podem ser expandidos livremente ou imitados por outras empresas.

A partir do momento em que a condição de heterogeneidade do uso do *Big Data* vai diminuindo pela replicação da mesma estratégia por empresas concorrentes que não foram as precursoras *(early adopters)*, a diferença de média do parâmetro de interesse Log Natural do Q de Tobin vai se reduzindo até não ser mais estatisticamente significante de 2019 em diante, inclusive ao nível de 90% de confiança, conforme demonstrado na Tabela 7.

À vista disso, os dados obtidos durante a avaliação da hipótese H5 permitem uma relação natural com o trecho teórico a respeito da RBV: "Independentemente da natureza dos aluguéis, a vantagem competitiva sustentável exige que a condição de heterogeneidade seja preservada. Se a heterogeneidade é um fenômeno de curta duração, os aluguéis também serão passageiros" (Peteraf, 1993, p.182).

Outro ponto interessante é perceber a relação entre os modelos estimados para avaliação das hipóteses de H1 a H4, os resultados dos testes de média, ano a ano, utilizados para avaliar a hipótese H5 e o recorte temporal da pesquisa de Cappa et al. (2020).

Das dimensões independentes do *Big Data*, apenas o Volume se demonstrou estatisticamente significante na realidade das companhias de capital aberto no Brasil em 2022. Porém, a Tabela 7 demonstra que 2018 foi o último ano que existiu diferença estatisticamente significante entre a performance das firmas que adotaram o *Big Data* primeiro em seus setores e as outras. Esse é justamente o ano do recorte temporal utilizado por Cappa et al. (2020).

Nesse sentido, temos que os resultados de H5 corroboram os resultados de H1 a H4 e demonstram aderência a RBV, plataforma teórica deste trabalho: de 2018 em diante, a estratégia de geração de valor através de *Big Data* parece ter se disseminado em níveis tão grandes que a vantagem competitiva obtida pelos *early adopters* parece ter minguado até que, em 2022, apenas a dimensão do Volume do *Big Data* manteve significância estatística.

4.3 Hipótese H6

A partir da análise das Tabelas e Figuras inseridas no Apêndice C, é possível identificar que o modelo utilizado para testar a hipótese H6 está relativamente bem ajustado, exibe multicolinearidade controlada, após a remoção da variável Turnover, que apresentou VIF acima de 21. Esse modelo estimado apresentou significância estatística para a razão do total de firmas com *apps* na *Google Play Store* em relação ao total de firmas por setor e por ano

(TA_TO_TC_RATIO) afetando positivamente a performance medida pelo Log Natural do Q de Tobin. Os resultados obtidos estão descritos na Tabela 8.

Tabela 8 – Modelo 2 resultados Regressão Linear Múltipla com dados em painel (H6)

Variável	**Coeficiente**	**P-valor**
Intercepto	0.6852	0.0118
BETA	-0.4590	0.0048**
D_TO_A_RATIO	0.2391	0.0610*
FIRM_AGE	-0.0039	0.4268
TA_TO_TC_RATIO	0.3589	0.0829*
Variável Dependente	TOBINQ_NL	
Nro. De Indivíduos	11	
Períodos de Tempo	8	
Tipo do Modelo	*PanelOLS*	
Breusch-Pagan	0.035192	
Shapiro-Wilk	0.616620	

*p-valor < 0.10; **p-valor<0.05
Fonte: Elaboração própria do autor.

Ceteris-paribus os resultados do modelo estimado para avaliação de H6 indicam que um incremento na adoção do *Big Data* por setor, medido nessa pesquisa por um novo lançamento de *app* na *Google Play Store* por uma empresa que ainda não tenha adotado essa estratégia, tem um impacto positivo de 0.3589 no Log Natural do Q de Tobin.

Em outras palavras, o aumento do uso do *Big Data*, por setor, afeta positivamente a performance deste setor e não negativamente como previsto na hipótese H6, baseando-se na lente teórica da RBV de que um recurso capaz de gerar vantagem competitiva sustentável precisa ser raro e estar disponível de forma heterogênea entre os competidores (Wernerfelt, 1984; Barney, 1986, 1991; Dierickx; Cool, 1989).

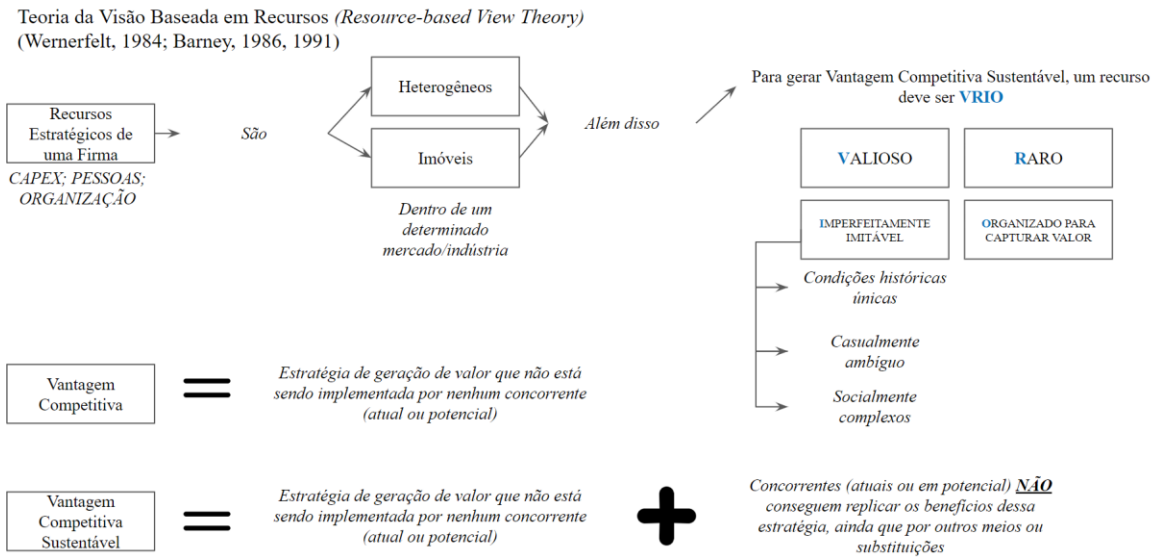
A hipótese H6 indica que, quando se altera a análise saindo da firma para a agregação dos dados por setor, setores que adotaram em maior intensidade estratégias de *Big Data* o impacto na performance média desse setor foi positivo.

A próxima seção de resultados busca relacionar os resultados das hipóteses entre si, além de relacioná-los com a RBV, plataforma teórica utilizada nesta pesquisa.

4.4 Análise e discussão geral dos resultados

Para a realização da discussão dos resultados nesta seção, foi construído um esquema visual, representado na Figura 14 a seguir, com o objetivo de retomar os pontos teóricos a respeito da Visão Baseada em Recursos (RBV).

Figura 12 – Resumo Visual da Teoria da Visão Baseada em Recursos (RBV)



Fonte: Elaboração própria do autor.

A Figura 12 objetiva explicar os pontos principais da RBV. Essa plataforma teórica indica que os resultados alcançados pelas firmas, em diferentes mercados e ambientes de competição, podem ser analisados por meio dos recursos que estão disponíveis internamente em cada uma delas.

Saindo do topo do desenho à esquerda, os recursos de uma firma podem assumir os mais variados formatos: podem ser ativos físicos (CAPEX), a força de trabalho qualificada (PESSOAS), ou a própria cultura organizacional da companhia (ORGANIZAÇÃO). Para que um determinado recurso seja considerado estratégico, ele precisa ser heterogêneo e imóvel, ou seja, não estar disponível com facilidade para outras firmas que competem em um mesmo mercado.

Todas essas características de um recurso estratégico, quando confirmadas segundo a RBV, o levam a gerar vantagem competitiva. Em outras palavras, partindo do lado direito do desenho, um recurso que tem a capacidade de gerar vantagem competitiva sustentável é: valioso, raro, imperfeitamente imitável e organizado para capturar valor.

Quando alguma dessas características é violada, a vantagem competitiva é apenas posicional e não sustentável. É isso que é retratado da metade para o final da Figura 12.

Relacionando a Figura 12 com o *Big Data*, tem-se a crença geral de considerar este ativo intangível um Recurso Estratégico de uma firma capaz de gerar vantagem competitiva sustentável (Francisco et al., 2020). Porém, à luz da RBV, conforme demonstrado na Figura

12, um recurso que gera vantagem competitiva sustentável precisa ser: valioso, raro, imperfeitamente imitável e organizado para capturar valor.

Ao decorrer da seção 4 e da avaliação das hipóteses da pesquisa, ficou evidenciado que as empresas brasileiras de capital aberto na B3 passaram por um grande movimento de lançamento de *apps* na *Google Play Store*, *proxy* objetiva do *Big Data* utilizada nessa pesquisa. Os dados coletados a partir da execução do *script* detalhado na Figura 1 e organizados de forma resumida na Tabela 3 demonstram que, no recorte de março de 2023, 76 empresas, ou 41,30%, já adotam essa estratégia.

Logo, ao relacionarmos essas informações com o resumo da RBV da Figura 12, fica evidente que *Big Data*, embora seja comumente tratado como uma inovação de vanguarda, tem uma relação muito mais forte com o passado das firmas. Isto é, não é mais uma estratégia que aparenta gerar vantagem competitiva sustentável atualmente, justamente porque não obedece boa parte dos preceitos teóricos da RBV: não tem sido mais tão raro e, aparentemente, firmas que não faziam parte do grupo de *early adopters* têm conseguido replicar os benefícios da estratégia.

Fato este que é reforçado ao verificar-se significância estatística apenas para a hipótese H1 e quando o modelo utiliza dados *cross-section* de 2022. Além disso, no Quadro C da Tabela 6, quando o modelo utiliza o estimador de Newey-West, observou-se significância estatística da interação os 4V's do *Big Data* ao nível de 10%, entretanto o coeficiente de apenas 0,0001 indica que, tudo o mais constante, o Q de Tobin sofre uma variação muito pequena quando existe um aumento de 1 unidade na utilização de *Big Data*.

Além disso, juntam-se os argumentos gerados pela avaliação de H5 revelando que o grupo de *early adopters*, detalhado na Tabela 5, experimentou ganho de performance diferencial enquanto o número de empresas com aplicativos lançados na *Google Play Store* era reduzido, até 2018, para ser mais preciso, conforme demonstrado na Tabela 7 e Figura 11.

Por outro lado, quando a análise foi agregada por setor e não mais no nível individual de cada firma, a avaliação de H6 sugere que, quanto maior o percentual de adoção do *Big Data* por um determinado setor econômico, um efeito positivo na performance de mercado medida pelo Q de Tobin foi estatisticamente significante.

Tais resultados demonstram que o *Big Data* aparenta ter deixado a posição de ativo intangível crítico e estratégico justamente pelo fato de que muitos concorrentes lograram êxito, ao longo de 2010 a 2022, ao imitar a mesma estratégia das empresas precursoras e obtiveram benefícios similares (Dierickx ; Cool 1989).

Ao relacionar os resultados da presente pesquisa com os obtidos em âmbito nacional, usando a base Scielo, não foi possível identificar trabalhos que abordaram os temas de *Big Data* e Performance das firmas brasileiras de forma objetiva, apenas via *surveys*.

O trabalho nacional encontrado, o de Medeiros et al. (2021), concluiu que a analítica de dados de negócios, tanto isoladamente quanto em conjunto, pode transmitir o efeito do *Big Data* para a Gestão do Desempenho Corporativo. A *survey* foi realizada com 312 gestores que utilizavam *big data analytics* (BDA) em organizações brasileiras. Esses resultados se relacionam com a presente pesquisa uma vez que o *Big Data* apesar de permanecer relevante para a Gestão do Desempenho Corporativo, aqui nessa pesquisa sendo medido pelo Q de Tobin, quando enquadrado à luz da RBV, demonstrou ser um recurso cada vez mais disponível para as firmas de capital aberto no Brasil e, portanto, não ser exatamente uma fonte de vantagem competitiva sustentável, principalmente de 2018 em diante.

Sendo assim, vale ressaltar mais uma vez que: "Independentemente da natureza dos aluguéis, a vantagem competitiva sustentável exige que a condição de heterogeneidade seja preservada. Se a heterogeneidade é um fenômeno de curta duração, os aluguéis também serão passageiros" (Peteraf, 1993, p.182).

Relacionando os resultados da presente pesquisa com os obtidos por Cappa et al. (2020) é possível observar que o poder explicativo obtido no presente estudo (R2) é inferior, possivelmente devido a ausência da variável de Pesquisa e Desenvolvimento para o mercado brasileiro.

Ademais, mesmo quando existiu significância estatística do *Big Data* neste estudo, como foi o caso do Modelo Interativo utilizando o estimador de Newey-West para produzir robustez em relação aos erros padrão consistentes com heteroscedasticidade e autocorrelação (HAC), o coeficiente obtido foi de valor muito baixo (0,0001), contrariando o que foi observado na pesquisa internacional (0,0036).

Adicionalmente, os resultados obtidos na presente pesquisa, para a realidade das firmas de capital aberto no Brasil, de 2010 a 2022, permitem sugerir que a democratização e a redução de custo para implementação de iniciativas de *Big Data* que ocorreram no mesmo período, conforme observado por Silva et al. (2020), podem ter contribuído para a adoção mais abrangente da estratégia. E, novamente à luz da RBV, se o recurso estratégico não é mais heterogêneo em um determinado mercado, a vantagem competitiva obtida é posicional e não sustentável.

5 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo mensurar o impacto do uso de *Big Data* na performance das companhias de capital aberto no Brasil, utilizando a Teoria da Visão Baseada em Recursos (RBV) como base.

Ao abordar os objetivos, foi utilizado o Q de Tobin como *proxy* para a performance das empresas e os dados dos *apps* lançados pelas empresas na *Google Play Store* como *proxy* objetiva para mensurar o *Big Data* disponível. Essas escolhas foram baseadas em pesquisas anteriores, como o trabalho de Cappa et al. (2020), visando tornar a replicação da pesquisa mais factível.

As hipóteses de pesquisa foram formuladas e testadas utilizando diferentes métodos estatísticos. As hipóteses H1 e H5 apresentaram significância estatística, indicando que quanto maior o volume de *Big Data*, menor o impacto na performance sugerindo efeito negativo da obesidade de informações sem moderação. E quanto mais cedo as empresas adotaram a estratégia de *Big Data*, maior foi o impacto na performance. Além disso, quando utilizado o estimador de Newey-West no modelo interativo dos 4V's do *Big Data* observou-se significância estatística, mas com um coeficiente de valor muito baixo (0,0001). Para todas as outras hipóteses, não foram observadas significâncias estatísticas, sugerindo que a vantagem competitiva para as firmas de capital aberto no Brasil no período de 2010 a 2022 gerada pelo uso do *Big Data* parece ter sido posicional, e não sustentável.

Esses resultados se relacionam com a Teoria da Visão Baseada em Recursos (RBV), fornecendo evidências de que a utilização do *Big Data* como recurso estratégico não é mais uma novidade. Os resultados apontam para um nível de maturidade no uso do *Big Data*, uma vez que muitas empresas brasileiras de capital aberto adotaram essa estratégia nos últimos anos.

As principais contribuições deste trabalho estão relacionadas à desmistificação do *Big Data* como um conceito inovador de vanguarda. Os resultados indicam que o investimento em *Big Data* ainda faz sentido, mas sem a crença de que seja capaz de resolver quaisquer problemas pelo simples fato de ser investimento em tecnologia.

Do ponto de vista metodológico, a presente pesquisa tem relevância na medida em que prezou sempre pela possibilidade de replicação de seus resultados, optando por alternativas que produzissem resultados determinísticos com a aplicação de *scripts* e automações, principalmente com a utilização do *Python*.

Foram enfrentadas algumas limitações na realização do presente estudo. Um dos objetivos iniciais era coletar dados de *apps* também da *App Store* da *Apple*, mas variáveis fundamentais para a construção e a execução dos testes das hipóteses como, por exemplo, data de lançamento do aplicativo, não são disponibilizados pela *Apple* na página da web de cada um dos aplicativos. Isto é, o nível de transparência dos dados dos aplicativos na *App Store* é bem menor do que na *Google Play Store*.

Outro aspecto limitador da pesquisa, foi a impossibilidade de acesso à Rais Identificada, que permitiria validar com maior exatidão a quantidade de empregados em cargos relacionados com Dados e melhorar a mensuração da dimensão da Veracidade do *Big Data*. A alternativa disponível foi coletar os dados da rede social *Linkedin*, mas essa rede não tem mecanismos para garantir que determinada pessoa realmente tenha vínculo empregatício vigente com determinada empresa.

A ausência de divulgação de dados relacionados a despesas com Pesquisa e Desenvolvimento por parte das empresas brasileiras de capital aberto, impediu que essa variável fosse adicionada nos modelos estatísticos, o que pode ter aumentado o erro de ajuste dos mesmos e reduzido o seu poder explicativo acerca dos fenômenos modelados, por exemplo.

Nesse sentido, o principal desafio foi a ausência de dados e a ausência de políticas que auxiliam na colaboração entre departamentos para obtenção de bases de dados já liberadas para a universidade, ou até mesmo do Ministério do Trabalho com as universidades, como o caso da Rais Identificada.

Do ponto de vista teórico, a pesquisa é relevante ao enquadrar o *Big Data*, à luz da Teoria da Visão Baseada em Recursos (RBV), como recurso capaz de gerar vantagem competitiva posicional e não sustentável, pois como foi identificado, no contexto das companhias brasileiras de capital aberto, os concorrentes aparentam ter conseguido replicar, posteriormente, os benefícios obtidos pelos *early adopters*.

Como sugestão para pesquisas futuras, indica-se a investigação do uso da Inteligência Artificial enquanto recurso estratégico capaz de gerar vantagem competitiva sustentável à luz da Teoria da Visão Baseada em Recursos (RBV). Além disso, a inclusão de variáveis que possam atuar como proxies de Pesquisa e Desenvolvimento para o mercado brasileiro, pode trazer mais robustez ao modelo proposto por essa pesquisa com inspiração em Cappa et al. (2020).

Do ponto de vista das áreas de Finanças e Ciências Contábeis, este estudo levanta a questão da evolução das estratégias de Big Data e sua capacidade contínua de gerar vantagem competitiva. Considerando que essas estratégias podem não ter o mesmo impacto no presente

que tiveram no passado, surge a indagação sobre se os profissionais e pesquisadores estão preparados para assimilar os avanços tecnológicos futuros, como a Inteligência Artificial.

Portanto, é relevante questionar o nível de conhecimento tecnológico incorporado na formação de futuros pesquisadores e profissionais contábeis, bem como considerar a necessidade de uma atualização profunda do currículo acadêmico nestas áreas.

REFERÊNCIAS

AGUIAR, G. de A. et al. Análise da influência dos ativos intangíveis no desempenho das empresas brasileiras. **Revista de Administração da UFSM**, v. 14, p. 907-931, 2021.

AKOKA, J.; COMYN-WATTIAU, I.; LAOUFI, N. Research on Big Data–A systematic mapping study. **Computer Standards & Interfaces**, v. 54, p. 105-115, 2017.

ANDREWS, D. WK. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica: **Journal of the Econometric Society**, p. 817-858, 1991.

ARUNACHALAM, D.; KUMAR, N.; KAWALEK, J.P. Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. **Transportation Research Part E: Logistics and Transportation Review**, v. 114, p. 416-436, 2018.

BARNEY, J. Firm resources and sustained competitive advantage. **Journal of management**, v. 17, n. 1, p. 99-120, 1991.

BARNEY, J. B. Resource-based theories of competitive advantage: A ten-year retrospective on the resource-based view. **Journal of management**, v. 27, n. 6, p. 643-650, 2001.

BUALLAY, A. Is sustainability reporting (ESG) associated with performance? Evidence from the European banking sector. **Management of Environmental Quality: An International Journal**, v. 30, n. 1, p. 98-115, 2019.

CAPPA, F. et al. Big data for creating and capturing value in the digitalized environment: unpacking the effects of volume, variety, and veracity on firm performance. **Journal of Product Innovation Management**, v. 38, n. 1, p. 49-67, 2021.

CHUNG, K. H.; PRUITT, S. W. A simple approximation of Tobin's q. **Financial management**, p. 70-74, 1994.

CÔRTE-REAL, N.; OLIVEIRA, T.; RUIVO, P. Assessing business value of Big Data Analytics in European firms. **Journal of Business Research**, v. 70, p. 379-390, 2017.

CVM - DFP. Cias Abertas: **Documentos: Formulário de Demonstrações Financeiras Padronizadas (DFP).** Portal Dados Abertos CVM: 2022. Disponível em: https://dados.cvm.gov.br/dataset/cia_aberta-doc-dfp. Acesso em 15.09.2023

CVM - FCA. (2022). Cias Abertas: **Documentos: Formulário Cadastral (FCA)**. Portal Dados Abertos CVM: Disponível em: https://dados.cvm.gov.br/dataset/cia_aberta-doc-fca. Acesso em 15.09.2023

DAVENPORT, T. Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press, 2014. Disponível em: https://books.google.com.br/books?id=apjBAgAAQBAJ. Acesso em 15.09.2023

DAVENPORT, T. How strategists use “big data” to support internal business decisions, discovery and production. **Strategy & leadership**, v. 42, n. 4, p. 45-50, 2014b.

DE MAURO, A.; GRECO, M.; GRIMALDI, M. What is big data? A consensual definition and a review of key research topics. In: AIP conference proceedings. **American Institute of Physics**, 2015. p. 97-104.

DIEBOLD, F. X. On the origin (s) of the term "Big Data". **arXiv preprint arXiv**:**2008.05835**, 2020.

DIERICKX, I.; COOL, K. Asset stock accumulation and sustainability of competitive advantage. **Management science**, v. 35, n. 12, p. 1504-1511, 1989.

FÁVERO, L.P.; BELFIORE, P. **Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®**. Elsevier Brasil, 2017.

FRANCISCO, E. R. et al. Beyond technology: Management challenges in the Big Data era. **Revista de Administração de Empresas**, v. 59, p. 375-378, 2020.

FREUND, G. P. et al. Mecanismos tecnológicos de segurança da informação no tratamento da veracidade dos dados em ambientes Big Data. **Perspectivas em Ciência da Informação**, v. 24, p. 124-142, 2019.

GANTZ, J. et al. Extracting value from chaos. **IDC iview**, v. 1142, n. 2011, p. 1-12, 2011.

GOMPERS, Paul; ISHII, Joy; METRICK, Andrew. **Corporate governance and equity prices. The quarterly journal of economics**, v. 118, n. 1, p. 107-156, 2003.

HAGEL, J. Bringing analytics to life. **Journal of Accountancy**, v. 2019, p.24-25, 2015

HAJI, A. A.; MOHD GHAZALI, N. A. The role of intangible assets and liabilities in firm performance: empirical evidence. **Journal of Applied Accounting Research**, v. 19, n. 1, p. 42-59, 2018.

HEISS, F.; BRUNNER, D. Using Python for introductory econometrics. New York: **Independently published**, 2020.

HILBERT, M.; LÓPEZ, P. The world's technological capacity to store, communicate, and compute information. **science**, v. 332, n. 6025, p. 60-65, 2011. doi:10.1126/science.1200970

IBM. Premier Healthcare Alliance trusts IBM to deliver comprehensive healthcare solution. Media Center IBM. 2012. Disponível em: https://mediacenter.ibm.com/media/Premier+Healthcare+Alliance+trusts+IBM+to+deliver+comprehensive+healthcare+solution/1_p3weduvd. Acesso em 15.09.2023

JAY, B. Strategic Factor Markets: Expectations, Luck, and Business Strategy. Management Science, pp. 1231-1241, 1986. DOI: 10.1057/978-1-349-94848-2_519-1.

JOHNSON, J.S.; FRIEND, S. B.; LEE, H. S. Big data facilitation, utilization, and monetization: Exploring the 3Vs in a new product development process. **Journal of Product Innovation Management**, v. 34, n. 5, p. 640-658, 2017.

KAUFMANN, M. Big data management canvas: a reference model for value creation from data. **Big Data and Cognitive Computing**, v. 3, n. 1, p. 19, 2019.

KITCHIN, R.; MCARDLE, G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. **Big Data & Society**, v. 3, n. 1, p. 2053951716631130, 2016. doi:10.1177/2053951716631130

KRETZER, J.; MENEZES, E. A. A importância da visão baseada em recursos na explicação da vantagem competitiva. **Revista de economia mackenzie**, v. 4, n. 4, 2006.

LANEY, D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Tech. rep., META Group. 2001. Disponível em: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Acesso em 15.09.2023

LIU, Y. Big data and predictive business analytics. **The Journal of Business Forecasting**, v. 33, n. 4, p. 40, 2014.

MAÇADA, A C. G.; BRINKHUES, R. A.; FREITAS, JOSÉ CARLOS DA SILVA. Capacidade de gestão da informação e implementação de estratégia de big data. **Revista de Administração de Empresas**, v. 59, p. 379-388, 2020.

MAROUFKHANI, P. et al. Big data analytics and firm performance: **A systematic review. Information**, v. 10, n. 7, p. 226, 2019.

MEDEIROS, M. M.; MAÇADA, A. C.G; HOPPEN, N. O papel da administração e análise de Big data como habilitadoras da gestão do desempenho corporativo. RAM. **Revista de Administração Mackenzie**, v. 22, 2021.

MIKALEF, P. et al. Big data analytics capabilities: a systematic literature review and research agenda. **Information systems and e-business management**, v. 16, p. 547-578, 2018.

NOBANEE, H. A bibliometric review of big data in finance. **Big Data**, v. 9, n. 2, p. 73-78, 2021.

PETERAF, M. A. The cornerstones of competitive advantage: a resource-based view. **Strategic management journal**, v. 14, n. 3, p. 179-191, 1993.

PRAKASH, G. Google Play Store Apps. Kaggle.com. Disponível em: https://www.kaggle.com/datasets/gauthamp10/google-playstore-apps?resource=download. Acesso em: 15 set. 2023.

ROSS, S. A., e al. Corporate Finance: Core Principals & Applications. **New York: McGraw-Hill/Irwin**. 2007.

SAGIROGLU, S.; SINANC, D. Big data: A review. In: 2013 international conference on collaboration technologies and systems (CTS). **IEEE**, 2013. p. 42-47.

SEABOLD, S.; PERKTOLD, J. Statsmodels: Econometric and statistical modeling with python. **In**: Proceedings of the 9th Python in Science Conference. 2010. p. 10-25080.

SILVA NETO, V. J.; BONACELLI, M. B. M.; PACHECO, C. A. Digital Technology System: artificial intelligence, cloud computing and Big Data. **Revista Brasileira de Inovação**, v. 19, 2021.

STATISTA. Revenue from big data and business analytics worldwide from 2015 to 2022. (2022). Disponível: Statista: https://www.statista.com/statistics/551501/worldwide-big-data-business-analytics-revenue/. Acesso 15.09.2023

KAPLAN, S. N.; ZINGALES, L. Do investment-cash flow sensitivities provide useful measures of financing constraints? **The quarterly journal of economics**, v. 112, n. 1, p. 169-215, 1997.

TAOUAB, O.; ISSOR, Z. Firm performance: Definition and measurement models. **European Scientific Journal**, v. 15, n. 1, p. 93-106, 2019.

WAMBA, S. Fosso et al. How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. **International journal of production economics**, v. 165, p. 234-246, 2015.

WERNERFELT, B. A resource-based view of the firm. **Strategic management journal**, v. 5, n. 2, p. 171-180, 1984.

WOOLDRIGE, J. M. Introductory econometrics: A modern approach. Atlanta, GA: AMAC **Accessibility Solutions**, 2002.

ZHANG, C. et al. Linking big data analytical intelligence to customer relationship management performance. **Industrial Marketing Management**, v. 91, p. 483-494, 2020.

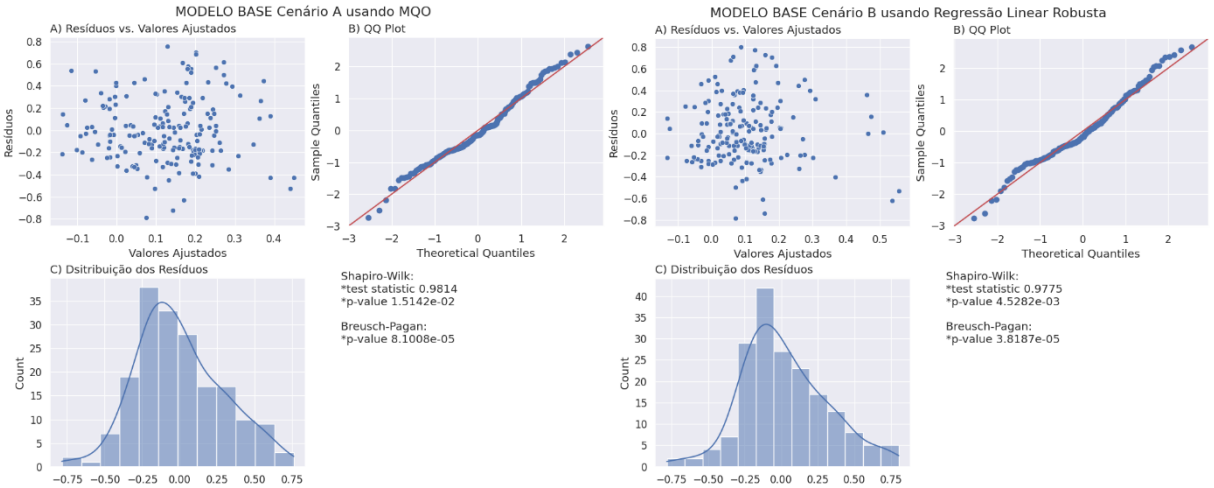
Apêndice A - Repositório *online* de *scripts*, dados coletados e utilizados na pesquisa

Com o objetivo de demonstrar a execução de todas as etapas da pesquisa e proporcionar ao leitor uma melhor compreensão dos pacotes na linguagem *Python* que foram utilizados no presente estudo, bem como fornecer os dados coletados no formato mais bruto até a sua versão mais tratada, foi criado um repositório *online* na plataforma *Github* de forma pública.

Para acessar o repositório, basta digitar em qualquer navegador de Internet o seguinte *Uniform Resource Locator* (URL): https://github.com/ivanrmello/big_data_research.

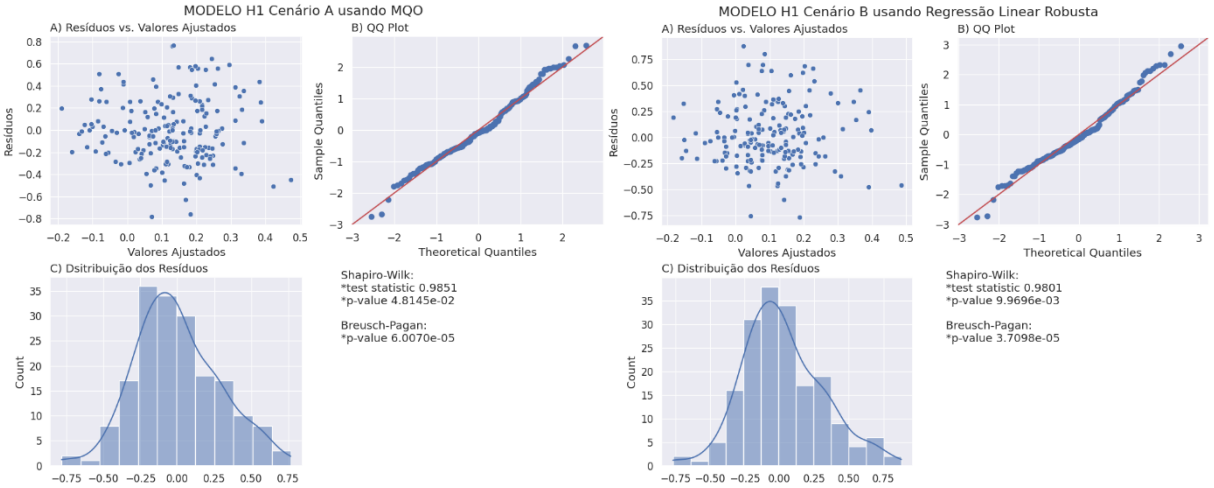
Apêndice B - Hipóteses H1, H2, H3 e H4 – Gráficos de Ajuste dos Modelos e cálculo do *Variance Inflation Factor* (VIF)

Gráficos de Ajuste do Modelo Base



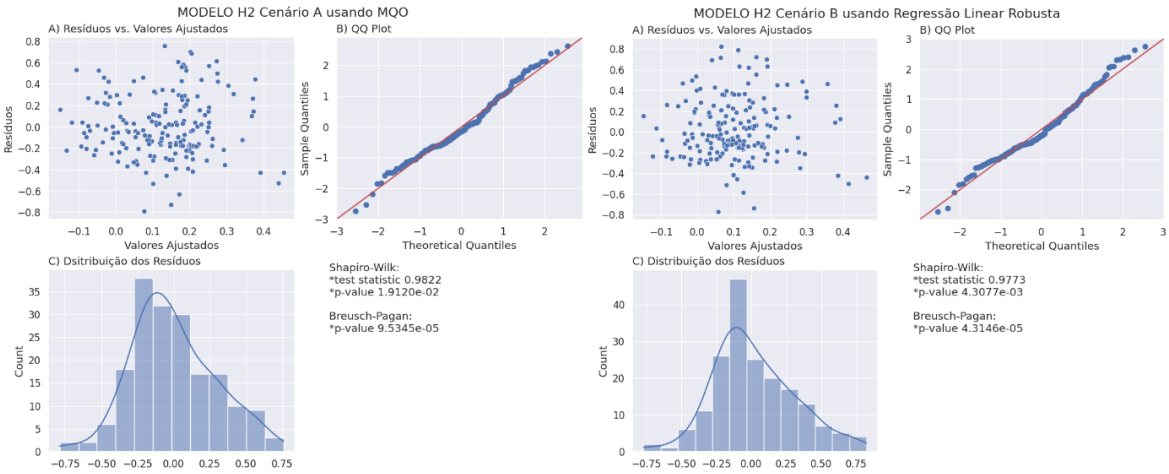
Fonte: Elaboração própria do autor.

Gráficos de Ajuste do Modelo H1



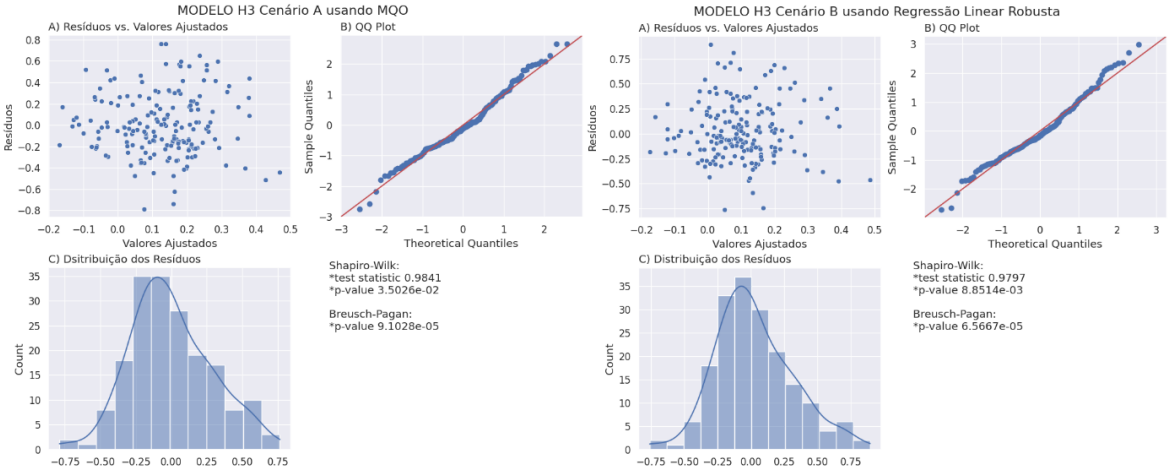
Fonte: Elaboração própria do autor.

Gráficos de Ajuste do Modelo H2



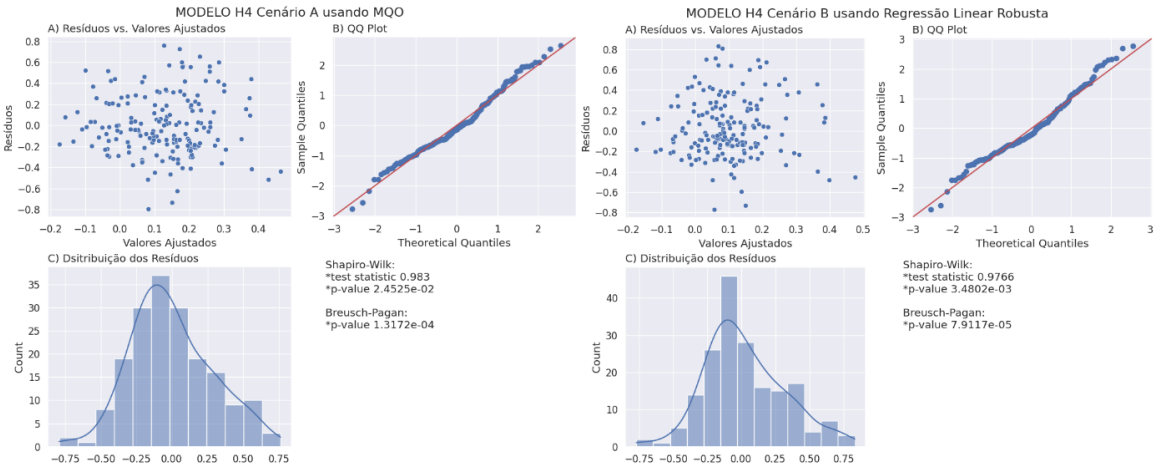
Fonte: Elaboração própria do autor.

Gráficos de Ajuste do Modelo H3



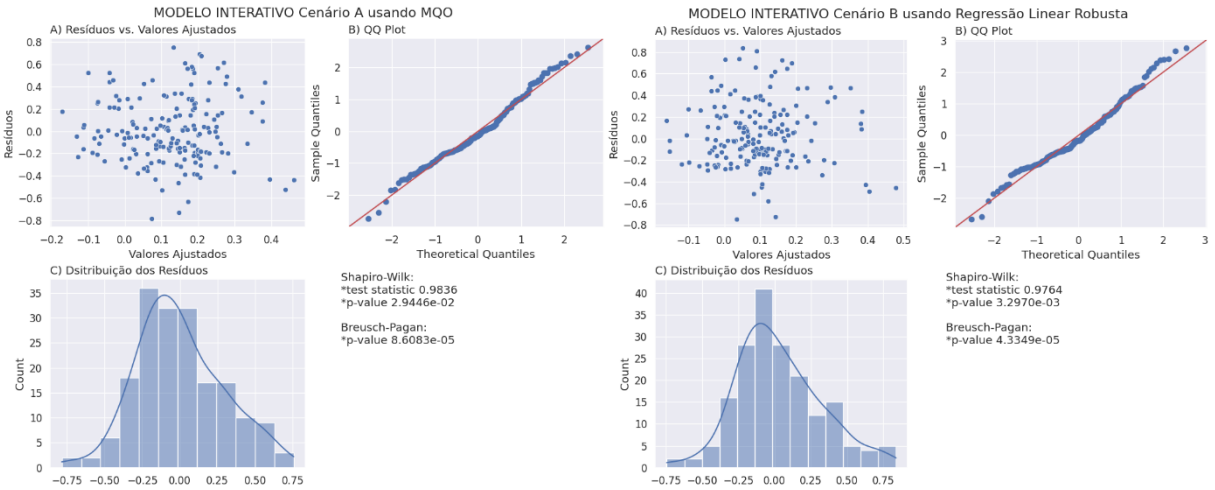
Fonte: Elaboração própria do autor.

Gráficos de Ajuste do Modelo H4



Fonte: Elaboração própria do autor.

Gráficos de Ajuste do Modelo Interativo



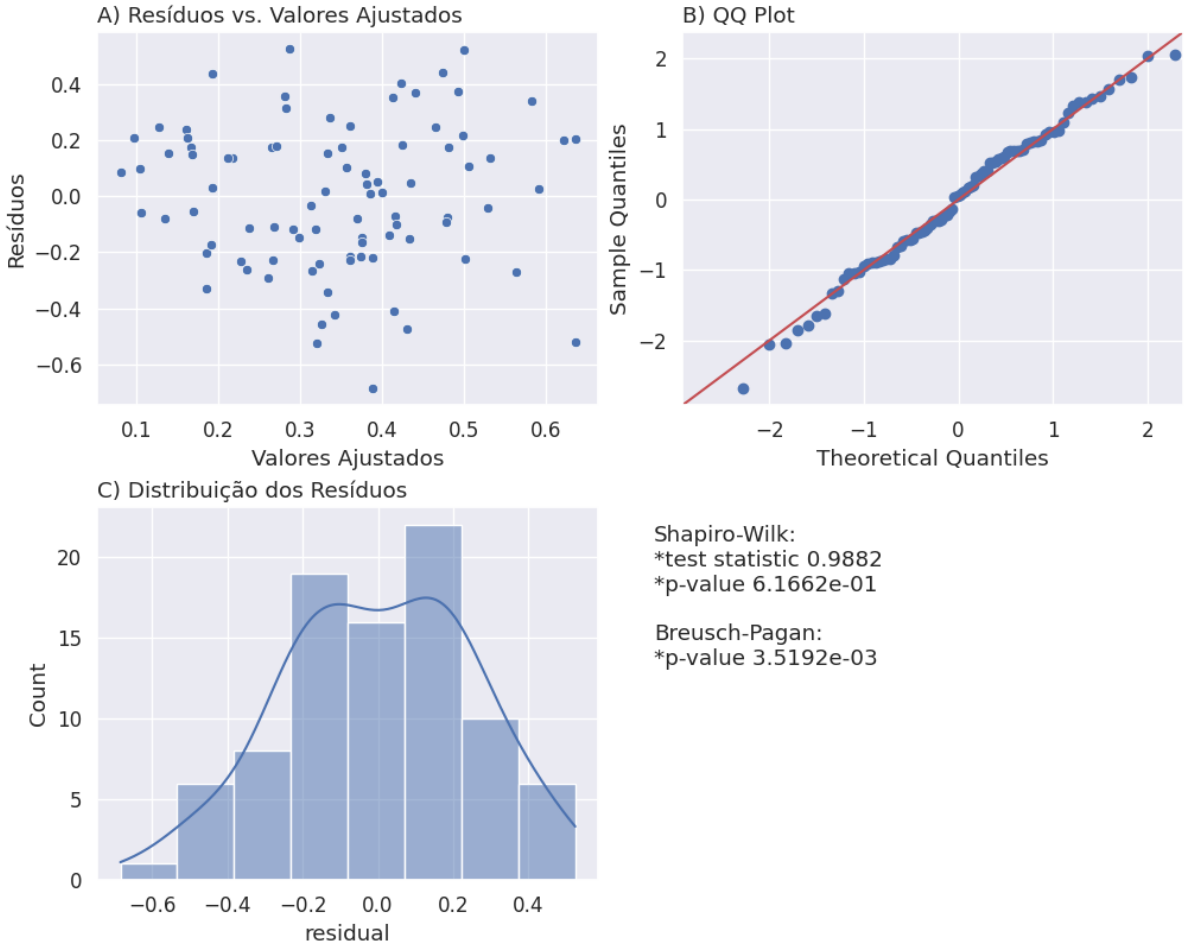
Fonte: Elaboração própria do autor.

Cálculo do Variance Inflation Factor (VIF) - H1 a H4

Variáveis	**Modelo Base** *(variáveis de controle)* **VIF**	**Modelo H1** *(variáveis de controle + volume)* **VIF**	**Modelo H2** *(variáveis de controle + variedade)* **VIF**	**Modelo H3** *(variáveis de controle + veracidade)* **VIF**	**Modelo H4** *(variáveis de controle + velocidade)* **VIF**	**Model Interativo** *(variáveis de controle + 4v's)* **VIF**
BETA	5,1882	5,1938	5,1904	5,2679	5,2304	5,1892
D_TO_A_RATIO	3,7241	3,7581	3,7650	3,8151	3,8128	3,7693
FIRM_AGE	2,7322	2,7782	2,7473	2,8027	2,8497	2,7500
TURNOVER_NL	8,4371	9,3186	8,5482	9,7021	9,5979	8,5764
VOLUME_NL		1,7641				
VARIETY_NL			1,2194			
VERACITY_NL				1,9067		
VELOCITY_NL					1,8387	
VOLUME_NL_i_VARIETY_NL_i_VERACITY_NL_i_VELOCITY_NL						1,1949
MÉDIA GERAL	**5,0204**	**4,5625**	**4,2941**	**4,6989**	**4,6659**	**4,2960**

Apêndice C - Hipótese H6 - Gráficos de Ajuste dos Modelos e cálculo do Variance Inflation Factor (VIF)

Gráficos de Ajuste do Modelo – após a remoção da variável de Controle Turnover devido a VIF acima de 10



Fonte: Elaboração própria do autor.

Cálculo do *Variance Inflation Factor* (VIF) - H6

Variáveis	**VIF**
BETA	8,7918
D_TO_A_RATIO	3,3972
FIRM_AGE	9,1575
TA_TO_TC_RATIO	4,3647
MÉDIA GERAL	**5,0204**

Fonte: Elaboração própria do autor.