Uma Introdução à Análise de Dados

An Introduction to Data Analysis

Luiz Guilherme Teixeira Bim[1], **Fernando José Antônio**[2]

RESUMO

Nesta pesquisa tem-se como objetivo analisar os dados obtidos das lojas de aplicativos da *Apple*, a *App Store*, e da *Google*, a *Play Store,* em busca de identificar e quantificar padrões. Os dados foram analisados utilizando a linguagem de programação *Python*, também foi utilizado o método de *bootstrapping*, que tem como objetivo a obtenção de intervalos de confiança para as estimativas dos parâmetros de interesse. Através da pesquisa foram obtidas as categorias com mais aplicativos, as categorias com menos aplicativos, a média de avaliação por categoria e relação entre a quantidade de aplicativos em uma dada categoria e a sua média de avaliações. Com isso pode-se concluir que os usuários da *App Store* estão mais satisfeitos com os aplicativos apresentados, o que pode ser entendido como um forte indicativo de um melhor controle de qualidade sobre os aplicativos que são aceitos pela plataforma da *App Store* e que a linguagem *Python* se mostrou uma eficiente ferramenta para tal análise.

Palavras-chave: Análise de Dados, Loja de Aplicativos, Avaliação de Usuários

ABSTRACT

This research aims to analyze data obtained from Apple's app stores, the App Store, and from Google, the Play Store, in order to identify and quantify patterns. The data were analyzed using the Python programming language, the bootstrapping method was also used, which aims to obtain confidence intervals for the estimates of the parameters of interest. Through the research, the categories with the most apps, the categories with the fewest apps, the average rating per category and the relationship between the number of applications in a given category and its average ratings were obtained. Therefore, it can be concluded that App Store users are more satisfied with the applications presented, which can be understood as a strong indicator of better quality control over the applications that are accepted by the App Store platform and that the language Python proved to be an efficient tool for such analysis.

Keywords: Data Analysis, App Store, User Rating

1 INTRODUÇÃO

O uso de conceitos e métodos de física estatística abre portas para novas formas de identificar e quantificar padrões em dados nas mais diversas áreas do conhecimento. Sistemas da natureza costumam ser formados por muitas partes interagentes, a partir das quais podem emergir comportamentos coletivos presentes em diversas escalas de magnitude. Sistemas que apresentam tais características são usualmente

[1] Engenharia de Controle e Automação, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil; lbim@alunos.utfpr.edu.br
[2] Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil; fjantonio@utfpr.edu.br

denominados de sistemas complexos Tal classificação é bastante abrangente, e inclui um grande número de problemas envolvendo desde sistemas físicos a sistemas biológicos ou mesmo sociais.

Classicamente, temos métodos quantitativos exatos para descrever sistemas de poucas partículas interagentes, tipicamente, não mais do que três. O caso oposto também pode ser bem entendido por métodos de mecânica estatística, na qual considera-se um número muito grande de partículas, tipicamente da ordem do número de Avogadro. Nesse cenário, não se obtém o comportamento individual de cada partícula, mas sim, o comportamento médio. Entretanto, é bastante difícil descrever sistemas cujo número de partículas é grande demais para ser resolvido exatamente, mas não é grande o suficiente para a observação apenas de comportamentos médios em casos limites. Apesar disto, para um número razoavelmente grande de partículas, entender o comportamento médio é suficiente para várias aplicações.

Para esta pesquisa, teve-se como objetivo a análise de dados obtidos das plataformas da *App Store* da *Apple* e da *Play Store* da *Google*, esses dados foram obtidos na plataforma *Kaggle* (763K iOS Apps, 2021; Google Play Store Apps, 2021), uma plataforma online que disponibiliza diversos bancos de dados gratuitamente. Tais dados incluem a categoria dos aplicativos, avaliação média, número de avaliações, preço em dólar e tamanho do aplicativo. O banco de dados da *App Store* contém um total de 763.731 aplicativos, enquanto que o banco de dados da *Play Store* contém mais de 2,3 milhões de aplicativos. Realizou-se uma filtragem nos dados de ambas as plataformas, descartando-se aplicativos duplicados e aplicativos sem nota, resultando em um total de 165.224 aplicativos para o banco de dados da *App Store* e 1.118.136 aplicativos para o banco de dados da *Play Store*. A última data de atualização do banco de dados da *Play Store* foi em junho de 2021, já o banco de dados da *App Store* foi atualizado em outubro de 2019.

2 MÉTODOS

Para a análise dos dados, utilizou-se a linguagem de programação gratuita *Python* (PYTHON, 2020) junto a plataforma de data science anaconda (ANACONDA, 2020). Inicialmente, cada variável foi investigada isoladamente, sempre confrontando os dois bancos de dados disponíveis. A seguir, passou-se a investigar a presença de correlação entre pares de variáveis de um mesmo banco de dados.

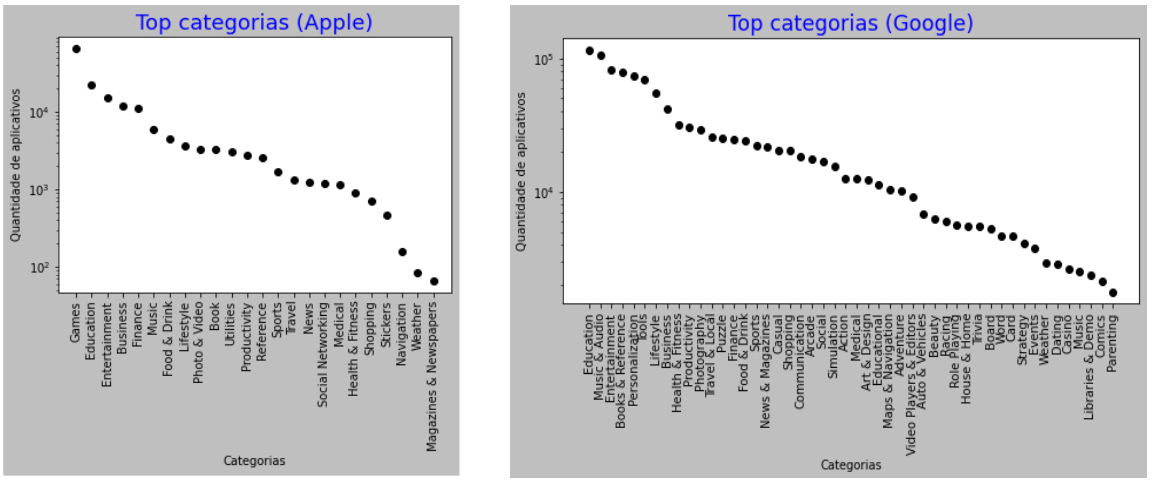
Outra ferramenta utilizada foi o método de bootstrapping, termo que se refere a um método de simulação que possui como função objetivar a obtenção de intervalos de confiança para as estimativas dos parâmetros de interesse, fazendo uso de reamostragem do conjunto de dados original. Dessa forma, essa técnica estima a distribuição de amostragem, coletando uma quantidade significativa de amostras com reposição de uma amostra aleatória única, a qual é chamada de reamostra (EFRON e TIBSHIRANI, 1993).

3 RESULTADOS

Para caracterizar os dados, estudamos a distribuição de *rank* dos aplicativos para cada banco de dados. Nesse caso, é feito um histograma com o número de aplicativos dentro de cada categoria (como jogos, produtividade e viagens), sendo que o eixo horizontal foi ordenado do maior para o menor número. Os resultados obtidos estão ilustrados na Figura 1.

Figura 1 – Distribuição de *ranking* da quantidade de aplicativos em cada categoria presentes (a) no banco de dados da *App Store*, (b) no banco de dados da *Play Store*.

(a) (b)

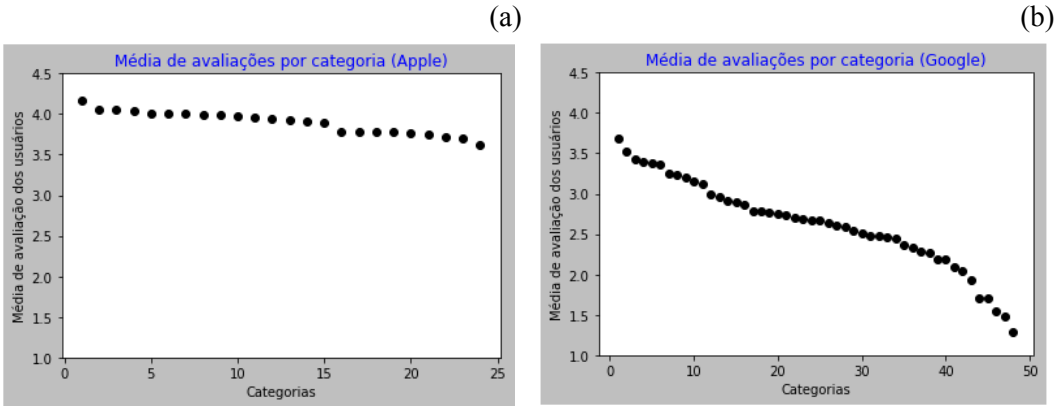


Fonte: Autoria própria (2021)

Percebe-se da Figura 1 que a distribuição de *rank* decresce mais rapidamente do que uma função exponencial negativa, indo a zero quase que abruptamente no caso dos dados da *App Store*. Em particular, as 14 categorias mais populosas seguem um padrão diferente das demais, indicando uma transição de fase. Também pode-se observar que os dados retirados da *Play Store* possuem uma variedade maior de categorias. Nesse caso, as 8 categorias mais populosas seguem um decaimento quase linear, ao passo que a tendência de decrescimento a partir de então se aproxima bastante de uma exponencial negativa.

Outra variável analisada foi a avaliação média dos aplicativos dentro de cada categoria para cada um dos bancos de dados de acordo com as categorias dos aplicativos. Os gráficos obtidos estão apresentados na Figura 2 e o resultado fornece um parâmetro mensurável da qualidade e/ou efetividade do aplicativo na dada categoria.

Figura 2 – Avaliação média dos aplicativos dentro de cada categoria para os aplicativos em cada categoria presentes (a) no banco de dados da *App Store*, (b) no banco de dados da *Play Store*.



Fonte: Autoria própria (2021)

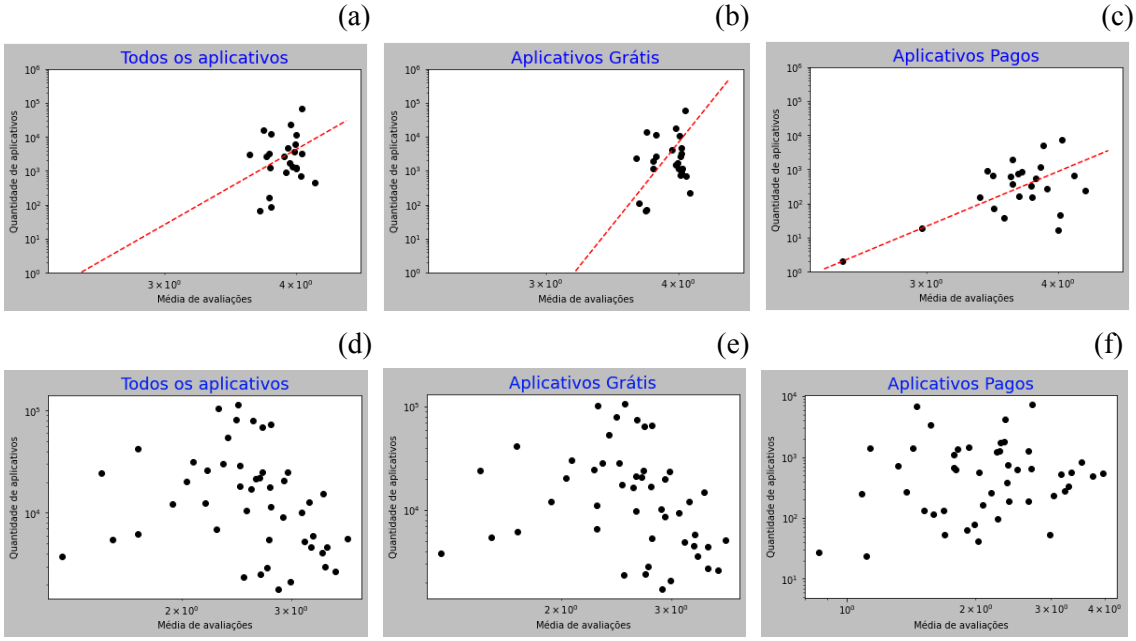
As categorias com maior avaliação média no banco de dados da *App Store* foram respectivamente "*stickers*", "*games*" e "*book*". Já para o banco de dados da *Play Store* foram "*role playing*", "*casino*" e "*cards*", respectivamente. Dessa forma, identificamos que essas são as categorias nas quais os aplicativos são mais bem desenvolvidos, mantidos e/ou produtivos. Por outro lado, as três categorias com menor avaliação média no banco de dados da *App Store* foram "*entertainment*", "*magazines & newspapers*" e "*utilities*". Já para o banco de dados da *Play Store* as três categorias com a menor avaliação foram "*house and home*", "*food and drink*" e "*events*". Dessa forma foi identificado que tais categorias possuem os aplicativos que mais apresentam problemas e geram insatisfação no consumidor final.

Também pode-se perceber que, no geral, as avaliações médias recebidas pela *App Store* são maiores do que as avaliações médias recebidas da *Play Store*. De fato, a categoria com menor média da *App Store* ("*utilities*") teve avaliação média de 3,61. Esse valor é razoavelmente próximo da categoria de maior avaliação média apresentada pela *Play Store*, a saber, trata-se da categoria "*Role Playing*" cuja avaliação média foi 3,96. Esse resultado pode ser um reflexo de um maior controle de qualidade por parte da Apple sobre quais aplicativos são aprovados para a sua *App Store*. Outro indicativo disso também é evidenciado pela quantidade reduzida de aplicativos presentes na *App Store*, com menos de 764 mil aplicativos no total, em relação a da *Play Store*, que apresentava mais de 2,3 milhões de aplicativos no momento da coleta dos dados.

O próximo passo da análise consistiu em investigar a relação entre a quantidade de aplicativos e a média de avaliações dentro de uma dada categoria. Além disso, os dados foram rearranjados em 3 grupos: (i) todos

os aplicativos de uma loja, (ii) apenas os aplicativos grátis dessa loja e (iii) apenas os aplicativos pagos dessa loja. Os gráficos obtidos estão dispostos na Figura 3.

Figura 3 – Relação entre a quantidade de aplicativos em uma dada categoria e a média de avaliações. (a) Para todos os aplicativos da *App Store*, uma regressão linear sugere uma reta com inclinação de 17,70. (b) Para os aplicativos grátis da *App Store*, uma regressão linear sugere uma reta com inclinação 39,01. (c) Para os aplicativos pagos da *App Store*, uma regressão linear sugere uma reta com inclinação 12,86. (d) Para todos os aplicativos da *Play Store*. (e) Para os aplicativos grátis da *Play Store*. (f) Para os aplicativos pagos da *Play Store*.



Fonte: Autoria própria (2021)

Da figura 3, percebe-se que através da regressão linear dos dados da *App store* foram encontradas três comportamentos distintos. De (a), temos uma reta com inclinação 17,70 quando todos os aplicativos foram considerados, ao passo que a inclinação aumentou para 39,01 quando apenas os aplicativos gratuitos foram considerados (b) e reduziu-se para 12,86 quando apenas os aplicativos pagos foram considerados. Esse resultado mostra que quando a avaliação média dos aplicativos é mais alta, temos uma maior quantidade de aplicativos nas categorias. Além disso, foi notável que esse padrão é mais evidente quando se leva em conta apenas os aplicativos pagos. Para o caso da loja *Play Store*, os aplicativos pagos seguem um padrão diferente dos aplicativos gratuitos. Entretanto, não surge uma relação alométrica. Por outro lado, surge um comportamento do tipo "v" voltado para baixo. Esses dados sugerem uma transição de fase nos dados que merece ser explorada futuramente.

4 CONCLUSÃO

Através da análise proposta pode-se concluir que a média de avaliações por categoria da *Play Store*, é significativamente menor que a da *App Store*, o que mostra que os usuários do iOS estão, no geral, mais satisfeitos com o que lhes é oferecido em sua loja de aplicativos. Isso pode se dar pelo fato de que a *App Store* possui um melhor controle de qualidade sobre quais aplicativos serão aceitos em sua plataforma. Ainda sobre a plataforma da *App Store*, pode-se perceber que quanto mais aplicativos uma determinada categoria possui, mais bem avaliada ela tende a ser. Ademais, notou-se que a plataforma *Python* que foi utilizada ao longo deste estudo se mostrou muito eficiente, fornecendo as ferramentas necessárias para analisar os dados fornecidos de forma satisfatória e gratuita.

REFERÊNCIAS

763K iOS Apps. Kaggle, 2021. Disponível em: <https://www.kaggle.com/cmqub19/763k-ios-app-info>. Acesso em: 08 de jun. de 2020.

ANACONDA. The World's Most Popular Data Science Platform, 2021. Disponível em <https://www.anaconda.com>. Acesso em: 28 de jun. de 2020.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. New York: John Wiley & Sons, 1993. p. 642.

Google Play Store Apps. Kaggle, 2021. Disponível em: <https://www.kaggle.com/gauthamp10/google-playstore-apps>. Acesso em: 25 de jun. de 2021.

PYTHON. Python Language, 2020. Disponível em: <https://www.python.org>. Acesso em: 10 de nov. de 2020.