Análise exploratória de dados da “Google Play Store” e estudos iniciais sobre correlação de atributos quantitativos para regressão.

Pedro Henrique Ventura Rodrigues de Alemida[1]

[1]Departamento de Mecânica Aplicada e Computacional
Departamento de ciência da computação
Universidade Federal de Juiz de Fora (UFJF)

pedro.ventura@ice.ufjf.br

Abstract. *Every day more and more apps are being released and made available in app stores like the “Play Store” as a result of the huge increase in smartphone accessibility (increasingly present in the lives of people around the world and at low cost) and from the ease of obtaining knowledge for mobile application development, adjunct of course, in the interest of the market. With this, this work seeks to study public data from the Google Play Store app store, in order to understand the current state of the platform and its applications from a quantitative point of view, and to look for factors of "success" of the apps, besides seeking correlation and a possible prediction of some data.*

Resumo. *A cada dia mais e mais aplicativos são lançados e disponibilizados em lojas de aplicativo como a “Play Store”, em consequência do grande aumento da acessibilidade a smartphones (cada vez mais presente na vida das pessoas ao redor do mundo e com baixo custo) e da facilidade de obtenção de conhecimento para desenvolvimento de aplicativos mobile, adjunto é claro, do interesse do mercado. Com isso, este trabalho busca estudar dados públicos da loja virtual de aplicativos “Google Play Store”, com o intuito de compreender o estado atual da plataforma e seus aplicativos em um ponto de vista quantitativo, e buscar fatores de “sucesso” dos aplicativos, além de buscar correlação e uma possível previsão de alguns dados.*

1. Introdução

Dentro da realidade brasileira, é visível no dia a dia o aumento de usuários de smartphones, onde em um país de aproximadamente 209,3 milhões de habitantes em 2017 [1], já possui atualmente 230 milhões de smartphones em uso (abril de 2019) [2].

Com preços cada vez mais acessíveis, os smartphones inundam os mercados físicos e virtuais. Incontáveis negócios e lojas abrem constantemente, todas voltadas para os aparelhos, vendendo além deles (ou nem os vendem), acessórios como capinhas, carregadores, personalização, películas, manutenção entre muitos outros produtos e serviços, como os de planos de dados para mobile.

O impacto disso pode ser visto além do mundo físico, no mundo digital. Segundo “statista.com” [3], no final de 2009, no “início” dos smartphones, haviam em torno de 16.000 (dezesseis mil) aplicativos para platadormas Android na plataforma “Google Play Store”, chegando a ter em março de 2018 aproximadamente 3,6 milhões de aplicativos. Atualmente a marca é de

aproximadamente 2,8 milhões, com aplicativos nos mais diversos segmentos.

Empresas de aplicações mobile como uber, segundo o Google Finanças [4], possuía uma receita de 3,81 bilhões de dólares em novembro de 2019. Logo, é visível o interesse de empresas e desenvolvedores individuais de se adentrar neste mercado. Contudo, como todo bom empreendimento, análise de dados do mercado alvo é importantíssimo para tomadas de atitudes, mas como e quais dados obter deste nicho é uma questão de interesse neste trabalho.

2. Objetivo

É difícil mensurar e identificar motivos que levam a popularização de um aplicativo. Contudo, este trabalho busca através de dados extraídos da "Google Play Store" sobre aplicativos contidos na loja, compreender os dados contínuos e/ou discretos disponíveis com métodos e ferramentas estatísticas, buscando correlações e descobrir alguma característica que influencie na popularização de um aplicativo.

3. Metodologias e Métodos

Para realização do estudo, utilizou-se da linguagem "Python 3", com um conjunto de ferramentas (bibliotecas) disponíveis na linguagem para manipulação de dados, análises estatísticas e geração de gráficos. O ambiente de desenvolvimento utilizado foi o Jupyter notebook com o gerenciador de pacotes do anaconda. Todo o código feito para este pode ser encontrado em [6], sendo possível a reprodução completa do experimento, havendo necessidade de instalações de bibliotecas definidas no escopo do código.

3.1. Obtenção da base, visualização de sua estrutura e componentes

Para o estudo, foi e selecionada uma base obtida em [5]. Trata-se de uma base criada por meio de "web scraping", ou seja, extração de dados de páginas web, no caso, da "Google Play Store". A mesma possuía no momento da obtenção 10841 registros e 13 atributos, sendo eles *"App, Catergory, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Updated, Current Ver e Android Ver"*, todas em formatos object, exceto "Rating", que se encontra como float. Trata-se então de uma amostra da loja de aplicativos da Google.

Como o objetivo deste não é a princípio a análise de variáveis categóricas, somente um conjunto de varieis possível de transformação para valores discretos ou contínuas foram selecionadas;

Rating, Reviews, Size, Installs e Price.

3.2. Visualização inicial dos dados e tratamento

Inicialmente foi feita uma remoção de instâncias com valores faltantes, havendo uma redução de 13.66% (novo total de 9360), que se julga não ser uma perda significativa devido à quantidade restante de amostras. Posteriormente foi analisado o formato dos dados, que exceto pelo "Rating", os demais necessitaram de transformações para valores reais, passando assim por um tratamento inicial de conversão.

Após o tratamento, foi realizado um teste de correlação, com intuito de buscar dados correlatos para análise. Em sua maioria não houve sinal de correlação, exceto por "Review" e "Intalls", que teve aproximadamente 64,16%, correlação moderada segundo Pearson[7]. Apesar de se induzir ao fato de quanto mais instalações ("Installs") um aplicativo tem (por ter correlação positiva), em consequência poderá haver mais pessoas escrevendo "resenhas" na plataforma ("Review"), aqui optou-se por seguir a linha de raciocínio inversa, que busca ver se aplicativos com mais resenhas, tendem a incentivar novos usuários a experimentar a aplicação.

Figura 1: Mariz de Correlação dos dados numéricos

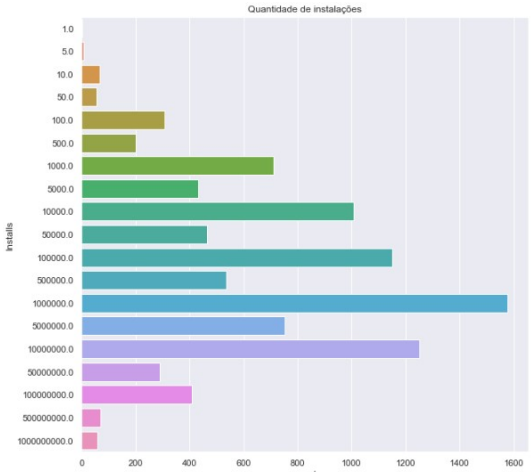

Figura 2: Histograma de quantidade de instalações (Installs)

Tabela 1: Tabela de Correlação dos atributos numéricos da base

Correlação	Rating	Reviews	Size	Installs	Price
Rating	1	-	-	-	-
Reviews	6,81%	1	-	-	-
Size	-1,89%	3,65%	1	-	-
Installs	5,13%	64,16%	-0,98%	1	-
Price	-2,19%	1,65%	1,84%	-0,11%	1

Como com base no resultado de correlação não se pode afirmar uma real correlação, fez-se um estudo com o intuito de validar os dados e buscar uma certeza . Em um primeiro momento, foram gerados gráficos de Histograma baseado na quantidade de cada instância, para compreender melhor a distribuição dos dados e verificar se é de uma distribuição normal.

Contudo, como pode ser visto pela figura 2, apesar de "apresentar" um formato de sino aparente, ainda não é possível tirar conclusões sobre o todo. O histograma sobre "Reviews" não foi inserido neste por uma questão de proporção e qualidade, uma vez que possui milhares de valores distintos. O mesmo pode ser visto em [6], assim como alguns outros experimentos e gráficos como CDFs e boxplots. Contudo, visualmente não aparenta seguir uma distribuição normal.

3.3 Avaliando a distribuição

Como os dados aparentemente não seguem uma distribuição normal, alguns testes foram feitos para verificar a validade dos mesmos, e se serviriam para uma regressão linear.

Foi então criado uma base bivariável, x → Installs e y → Reviews. Dela foi calculado valores relativos ao ponto de interceptação no eixo Y da reta de probabilidade de XY (b0, 118421.30) e o coeficiente de inclinação da reta (b1, 0.022).

Posteriormente foi calculado o SSE, SST e SSR para poder calcular o R^2, ou seja, estatística para verificar a confianças dos parâmetros de regressão. Obteve-se como resultado para R^2 aproximadamente 41,16%, sendo então o modelo não muito propício para uma regressão.

R^2: 0.41165731

SSR: 3.81077229

SSE: 5.44637491

SST: 9.2571472

Com isto, restou verificar apenas para confirmação da má qualidade da regressão a qualidade dos parâmetros b0 e b1 em intervalos de confiança de

90%,95% e 99%.

Tabela 2: Qualidade do Modelo

	Desvio padrão	90%	95%	99%
b0	2.52	-514.36, -506.01	-515.17, -505.21	-516.11, -504.27
b1	0.02	-1.12, -1.047	-1.13, -1.04	-1.13, -1.03

Baseado nas leituras dos dados da tabela 2, todos os valores de b1 encontram-se próximos à 0 e b0 muito elevados e maiores que b1, podendo concluir que os parâmetros encontrados não descrevem muito bem os dados (resultado da qualidade da regressão), não sendo relevantes para o modelo de regressão.

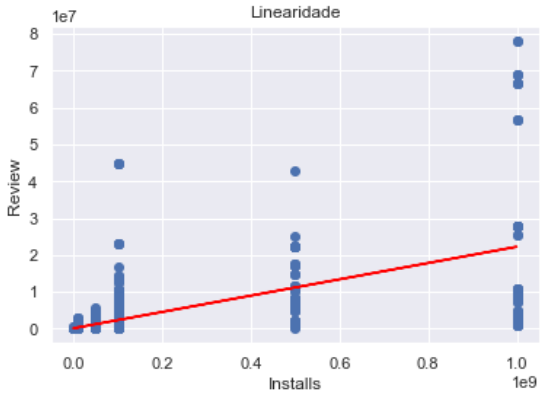

Figura 3: Teste Visual de Linearidade

Tal fato também se torna visível no teste visual de linearidade (figura 3), onde os pontos possuem padrões de comportamento de crescimento em sobre o eixo Y, não seguindo a reta de regressão gerada.

Padrões também podem ser encontradas na figura 4, no teste de Homoscedasticidade, sugerindo que para este modelo, talvez um teste de regressão não-linear fosse mais interessante.

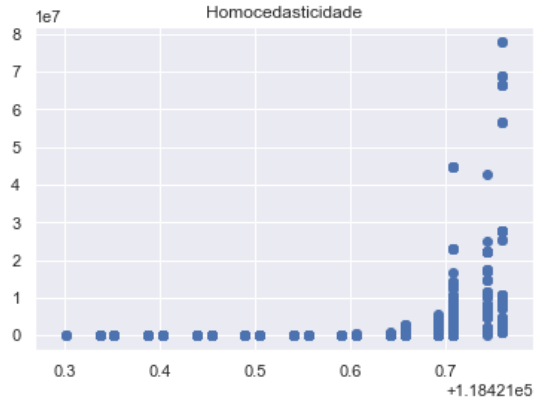

Figura 4: Gráfico de Homoscedasticidade log

3.4 Análise dos eixos

Uma vez comprovado a má qualidade dos eixos para a regressão, fez-se uma análise dos mesmos, com intuito de verificar o efeito de outliers no processo e comportamento do eixo.

Figura 5: Gráfico de probabilidade da variável Reviews

Como pode ser visto na figura 6, o comportamento de “Reviews” tende a manter-se sobre a linha de regressão, contudo, em seu final, apresenta comportamento exponencial sobre o eixo y, mostrando uma mudança drástica de comportamento com peso de possíveis outliers.

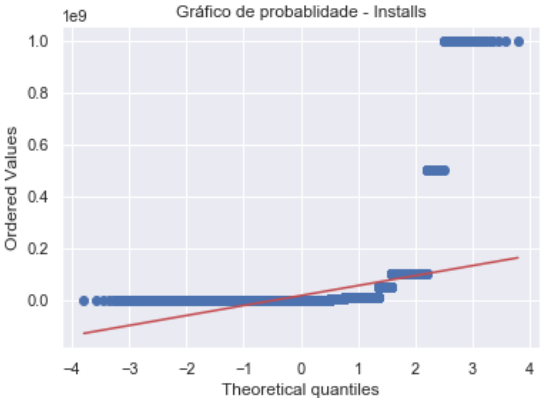

Figura 6: Gráfico de probabilidade da variável Installs

Já na figura 7, o comportamento de “Installs” demonstra não seguir a reta, com presença fortes de outliers.

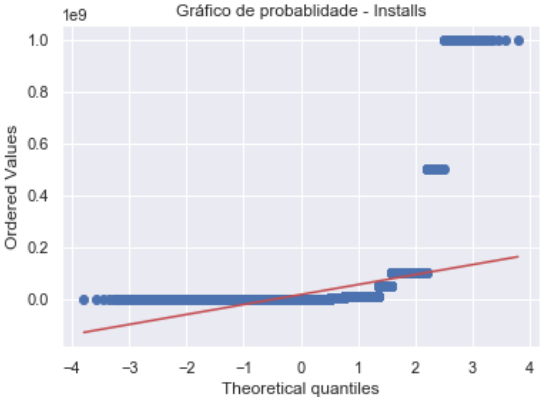

Figura 7: Gráfico de probabilidade da variável Installs

4. Conclusões e Pesquisas futuras

Baseado nas análises feitas, constatou-se a impossibilidade de com a base atual, se obter correlações reais e geração de regressões através dos métodos utilizados neste trabalho.

Pressupõe-se que tal fato ocorra por diversos fatores, entre eles: o fato de termos uma amostra muito pequena em relação a população (aproximadamente 0,036%, grande impacto de outliers, poucas variáveis, principalmente variáveis com valores numéricos e diferença de comportamento por nicho de clientes.

Sugere-se para trabalhos futuros um “web scraping” mais significativo, obtendo uma amostra correspondente a uma fração maior da população. Outro fator sugerido é a utilização de métodos que levem em considerações as variáveis nominais ou não lineares. Por fim, seria válido uma reprodução do experimento, agrupando por categoria, uma vez que usuários de gêneros específicos de aplicações, possam possuir características distintas.

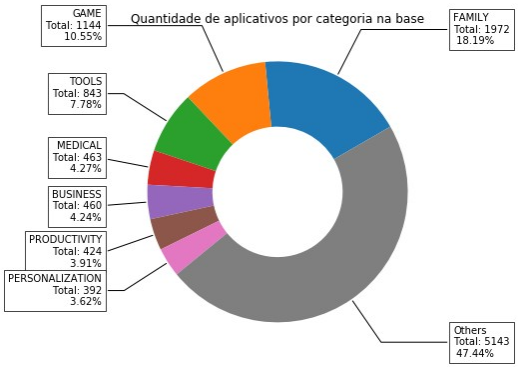

Figura 8: Distribuição dos dados por categoria

Referências

Google Data Explorer; (2019), *“Indicadores do desenvolvimento Mundial – População (Brasil)”*, https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&met_y=sp_pop_totl&idim=country:BRA:ARG&hl=pt&dl=pt, Novembro. [1]

Globo.com. (2019), *“Brasil tem 230 milhões de smartphones em uso”*, https://epocanegocios.globo.com/Tecnologia/noticia/2019/04/brasil-tem-230-milhoes-de-smartphones-em-uso.html, Novembro 2019. [2]

Statista, (2019),*”Number of available applications in the Google Play Store from December 2009 to September 2019”*, https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/. Novembro 2019. [3]

Google Finance. (2019), *“Market Sumary, Uber Technologies Inc”*, https://www.google.com/search?q=uber&tbm=fin#scso=_PRjgXbHPDJq85OUP_K2o8A49:0&wptab=OVERVIEW. Novembro 2019. [4]

Gupta, Lavanya (2019). *“Google Play Store Apps. Web scraped data of 10k Play Store apps for analysing the Android market.”*, https://www.kaggle.com/lava18/google-play-store-apps. Novembro 2019 [5]

de Almeida, Pedro. (2019). *“Trabalho Final de Métodos Quantitativos. Análise exploratória de dados da “Google Play Store” e estudos inicias sobre correlação de atributos quantitativos para predição”*.https://github.com/ArtFicer/PGCC-MQ/tree/master/MQFinal. Dezembro 2019. [6]

Mukaka, M.M. (2012),*”Statistics Corner: A guide to appropriate uso of Correlation coefficient in medical research”*. Malawai Medical Journal. PMC 3576830. Novembro 2019. [7]