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

Pedro Henrique Ventura

Introdução

**Situação atual no Brasil:**

- População brasileira: 209,3 milhões de habitantes em 2017;
- 230 milhões de smartphones em uso (abril de 2019)

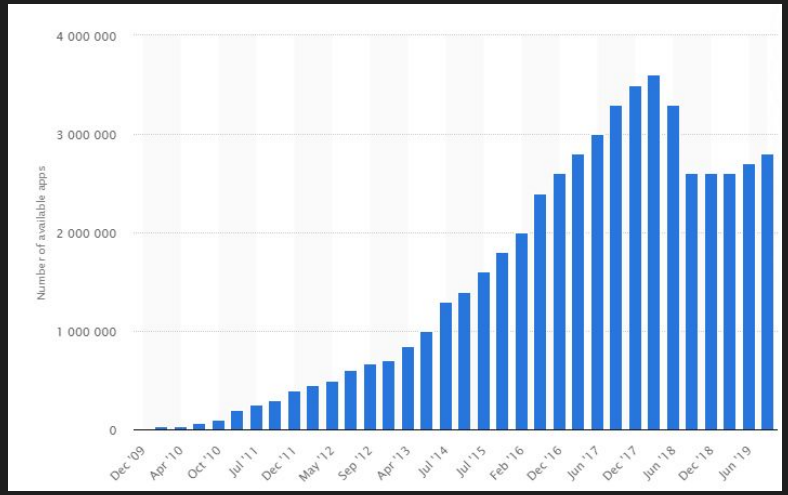



Fig1. Total de aplicativos na Google Play Store

GET IT ON Google Play

# Motivação

Tudo tem um porquê...

## Motivação

Empresas de aplicações mobile como uber, que possuía uma receita de 3,81 bilhões de dólares em novembro de 2019. Como todo bom empreendimento, análise de dados do mercado alvo é importantíssimo para tomadas de atitudes. Mas como e quais dados obter deste nicho?

## Objetivo

Este trabalho busca através de dados de aplicativos da “Google Play Store”, 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.

# Metodologias e Métodos

# Etapas de desenvolvimento

01

## FIRST

Obtenção da base, visualização de sua estrutura e componentes

02

## SECOND

Visualização inicial dos dados e tratamento

03

## THIRD

Avaliando a distribuição

04

## FOURTH

Análise dos eixos

# 01

Obtenção da base, visualização de sua estrutura e componentes

## Base - Dataset

Base: Google Play Store Apps

Gerada por: Web scraping

Link: https://www.kaggle.com/lava18/google-play-store-apps

Registros: 10841

Atributos: 13

# Base - Dataset

App: Application name

Category: Category the app belongs to

Rating: Overall user rating of the app (as when scraped)

Reviews: Number of user reviews for the app (as when scraped)

Size: Size of the app (as when scraped)

Installs: Number of user downloads/installs for the app (as when scraped)

Type: Paid or Free

Price: Price of the app (as when scraped)

Content: Rating: Age group the app is targeted at - Children / Mature 21+ / Adult

Genres: An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.

Last Updated: Date when the app was last updated on Play Store (as when scraped)

Current Ver: Current version of the app available on Play Store (as when scraped)

Android Ver: Min required Android version (as when scraped)

02

# Visualização inicial dos dados e tratamento

# Tratamento

## Remoção

Remoção de valores faltantes.

Redução de 13.66%

Novo total de 9360

## Formato

Foram transformados em valores reais: Reviews, Size, Installs e Price

## Outliers

Optou-se pela não remoção, por possuir uma amostra relativamente pequena pelo todo.

0,036%

# Busca de correlação

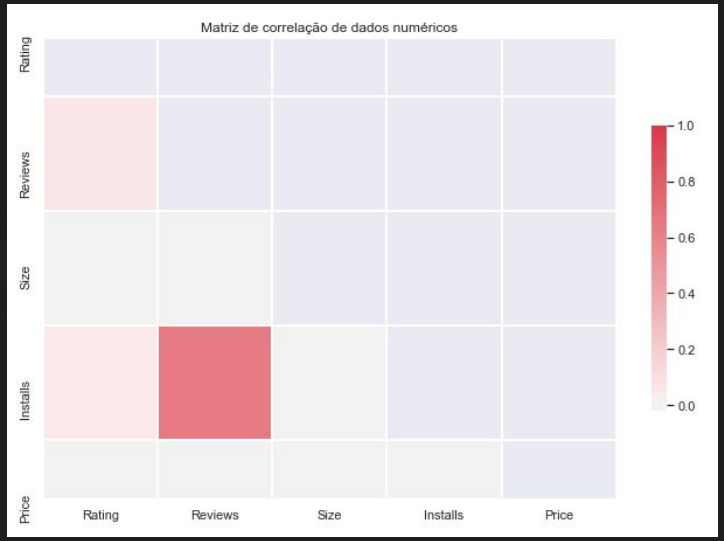



Fig2. Matriz de Correlação

| 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 |

Tabela 1. Matriz de Correlação

# Análise das variáveis

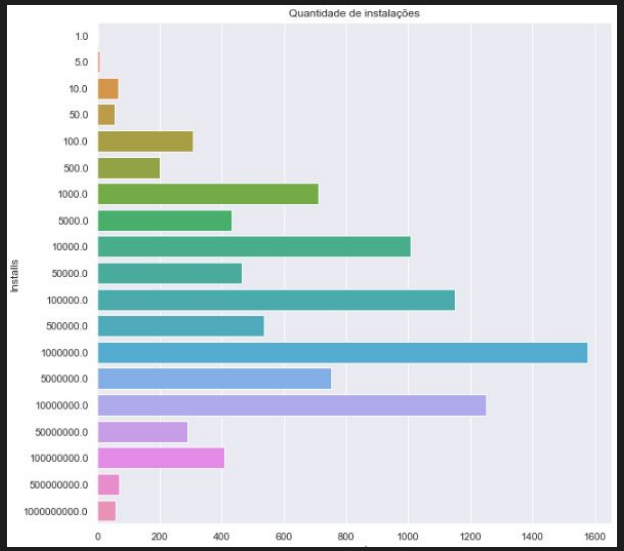
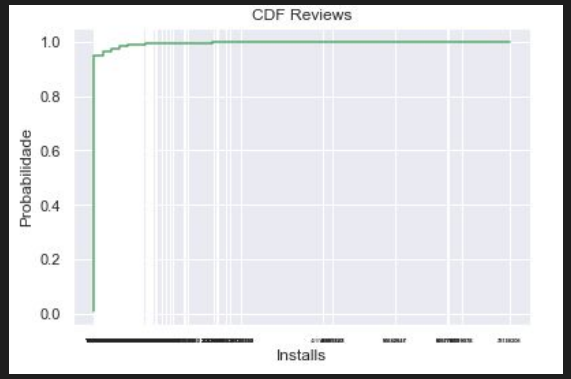



Fig3. Histograma de quantidade de instalações (Installs



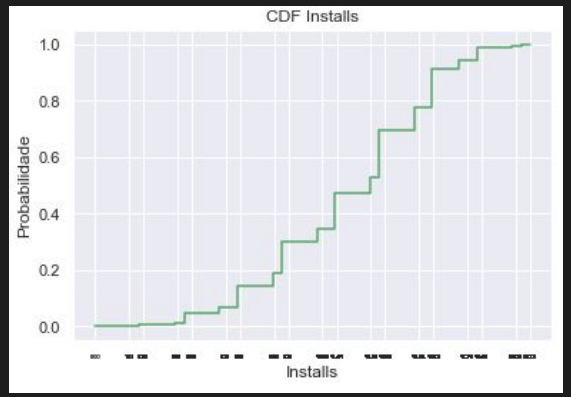




Fig 4 e 5. CDF Reviews e Installs


# 03
# Avaliando a distribuição

# Parâmetros para Regressão

## Reta de Regressão

Interceptação no eixo Y da reta: b0, 118421.30
Coeficiente de inclinação da reta: b1, 0.022

## Confianças dos parâmetros de regressão

$R^2$: 0.41165731
SSR: 3.81077229
SSE: 5.44637491
SST: 9.2571472

## Qualidade do Modelo

Desvio padrão
b0: 2.52
b1: 0.02

## Intervalo de Confiança 90%

b0: -514.36, -506.01
b1: -1.12, -1.047

## Intervalo de Confiança 95%

b0: -515.17, -505.21
b1: -1.13, -1.04

## Intervalo de Confiança 99%

b0: -516.11, -504.27
b1: -1.13, -1.03

# Análise Visual

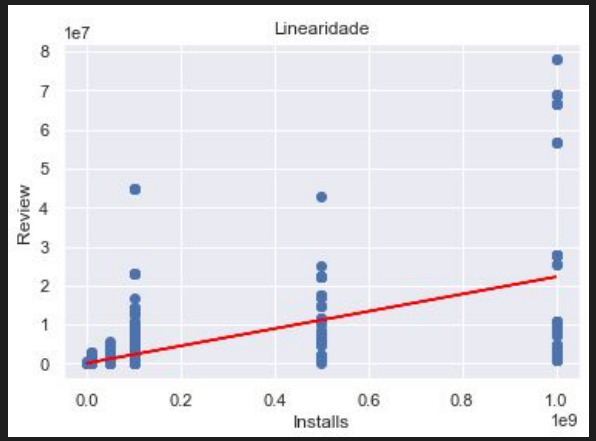



Figura 6: Teste Visual de Linearidade

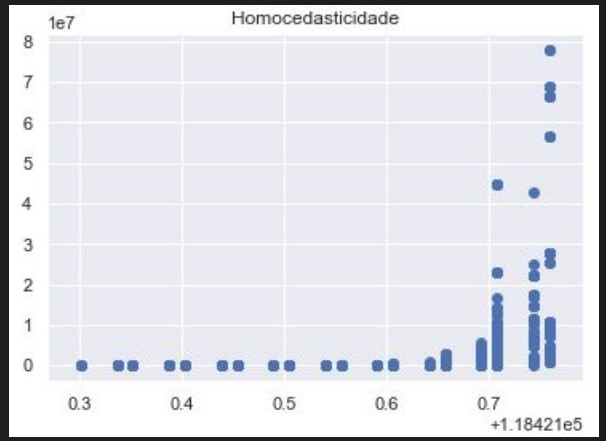



Figura 7: Grafico de Homocedasticidade

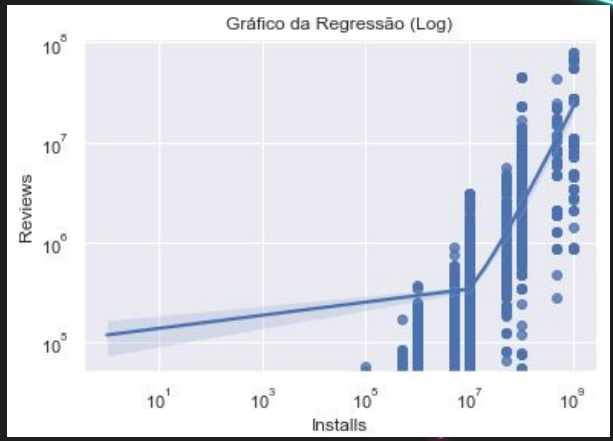



Figura 8: Gráfico da regressão gerada em log

# 04

# Análise dos eixos

# Gráfico de Probabilidade


Figura 9: Gráfico de probabilidade da variável Reviews

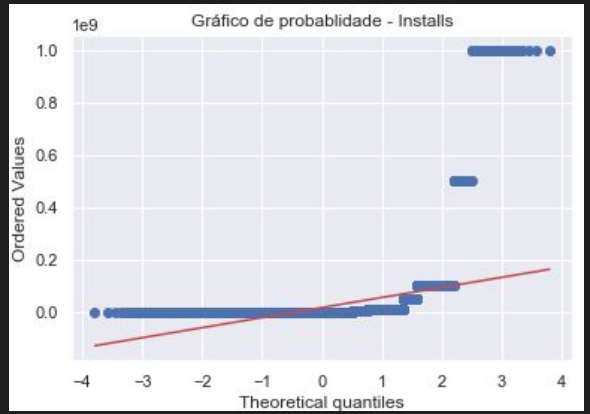
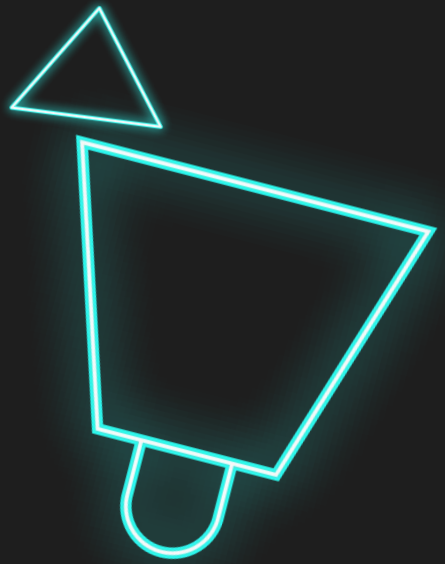



Figura 10: Gráfico de probabilidade da variável Installs



# Conclusões e Pesquisas Futuras

... e também um final.

## Conclusões

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.

## Pressuposições

1. Amostra muito pequena em relação à população (0,036%);
2. Grande impacto de outliers;
3. Poucas variáveis, principalmente variáveis com valores numéricos;
4. Diferença de comportamento por nicho de clientes.

## Sugestões para pesquisas futuras

1. “Web scraping” mais significativo;
2. Utilização de métodos que levem em considerações as variáveis nominais ou não lineares.;
3. Reprodução do experimento, agrupando por categoria, uma vez que usuários de gêneros específicos de aplicações, possam possuir características distintas.

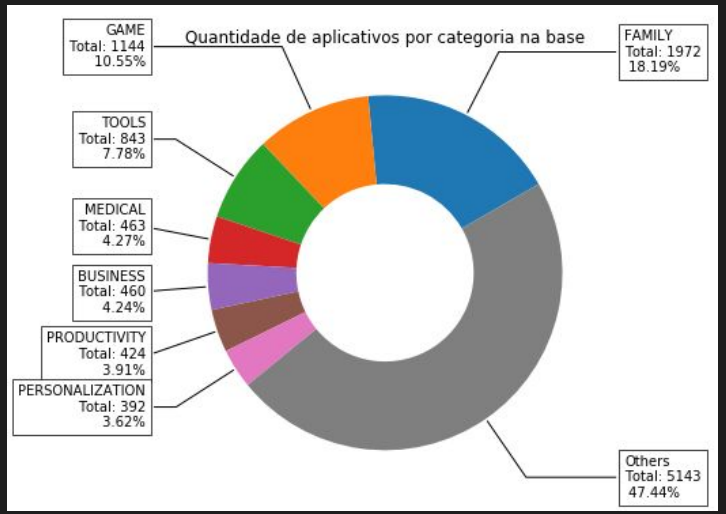



Figura 11: Quantidade de aplicativos por categoria na base

# THANKS!


Do ~~you~~ not have any questions ~~?~~

pedro.ventura@ice.ufjf.br
github.com/ArtFicer/PGCC-MQ
/tree/master/MQFinal