Relatório

Análise dos dados da Google Play Store

Danilo Morales Teixeira
09/05/2019

O objetivo deste projeto era analisar os dados da Google Play Store, analisando desde quantas vezes um aplicativo foi instalado, a faixa etária dominante de usuários de cada aplicativo assim como os sentimentos das opiniões de cada usuário que utilizou um determinado aplicativo.

Na primeira etapa analisou-se os dados do arquivo googleplaystore.csv verificando que este dado possui 13 colunas e 10841 linhas e que as colunas contendo informações foram nomeadas como: 'App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver' e 'Android Ver'.

Antes de realizar uma análise detalhada deste arquivo, verificou-se quantas linhas apresentam dados ausentes do tipo Not A Number (NANs). Os resultados encontrados para cada coluna foram:

Categoria	Total	Total Percentual
Rating	1474	0.135965
Current Version	8	0.000738
Android Version	3	0.000277
Content Rating	1	0.000092
Type	1	0.000092
Last Updated	0	0
Genres	0	0
Price	0	0
Installs	0	0
Size	0	0
Reviews	0	0
Category	0	0
App	0	0

Observou-se que apenas as colunas Rating, Current Version, Android Version, Content Rating e Type possuem valores ausentes, porém numa taxa que podemos considerar desprezível em relação ao total de dados presentes nesta amostra. Desta forma decidiu-se remover todas as linhas que possuíssem valores ausentes. Desta forma, passamos a analisar uma amostra com 13 colunas e 9360 linhas.

Uma informação muito importante em relação aos aplicativos é a nota média de cada um, ou seja, a nota que cada usuário que utiliza este aplicativo deu para o mesmo. Verificou-se que a nota média, considerando todos os aplicativos, foi de 4.19 com um desvio padrão de 0.51 onde observou-se que alguns usuários deram notas máximas e

mínimas para os aplicativos. A distribuição de notas de todos os aplicativos é apresentada no gráfico da figura 1. A partir deste gráfico observamos, conforme mencionado anteriormente, que a maior concentração de notas está na região da nota 4.0

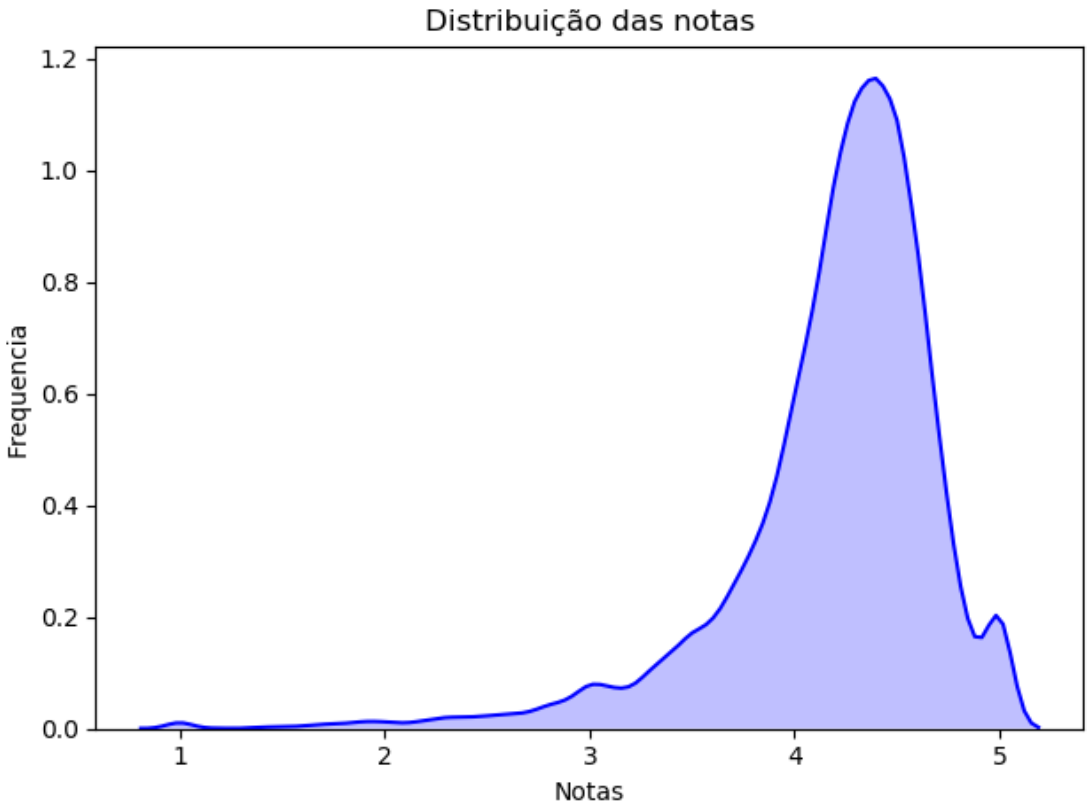

Figura 1: distribuição das notas dos aplicativos

Em seguida, listamos as categorias existentes onde cada aplicativo pode ser encontrado. Estas categorias são: ART_AND_DESIGN, AUTO_AND_VEHICLES, BEAUTY, BOOKS_AND_REFERENCE, BUSINESS, COMICS, COMMUNICATION, DATING, EDUCATION, ENTERTAINMENT, EVENTS, FINANCE, FOOD_AND_DRINK, HEALTH_AND_FITNESS, HOUSE_AND_HOME, LIBRARIES_AND_DEMO, LIFESTYLE, GAME, FAMILY, MEDICAL, SOCIAL, SHOPPING, PHOTOGRAPHY, SPORTS, TRAVEL_AND_LOCAL, TOOLS, PERSONALIZATION, PRODUCTIVITY, PARENTING, WEATHER, VIDEO_PLAYERS, NEWS_AND_MAGAZINES e MAPS_AND_NAVIGATION.

Realizamos a contagem de aplicativos de cada categoria e os resultados estão presentes no gráfico de barras apresentados na figura 2. Observa-se que as categorias Game e Família possuem maior número de aplicativos.

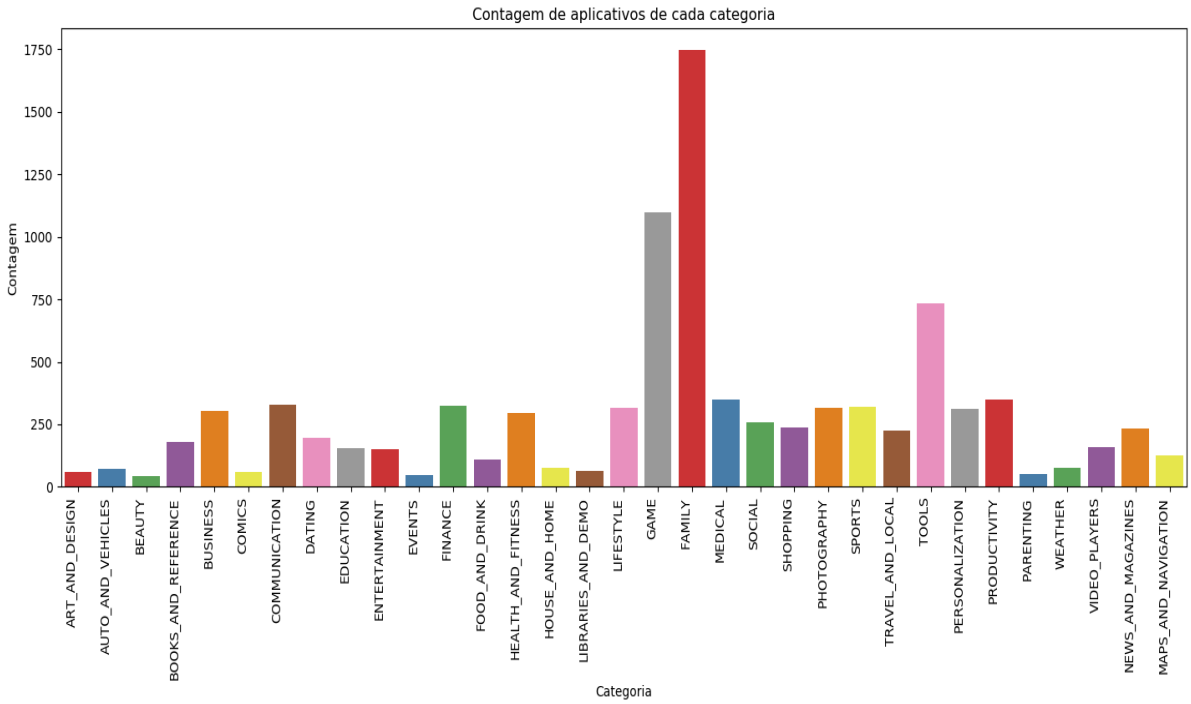

Figura 2: contagem de aplicativos de cada categoria

A próxima etapa do processo de análise foi gerar um gráfico do tipo boxplot das notas de cada categoria. Este resultado está presente na figura 3, onde observa-se que as categorias apresentam uma distribuição de notas muito parecidas considerando-se as barras de erro.

Analisou-se a relação entre o número de vezes que um aplicativo foi instalado em função da sua nota média. Antes, porém, verificou-se os tipos de dados presentes na coluna Installs onde notou-se que se trata de um objeto e que os valores possuem os caracteres + e a vírgula em seus valores. Para que os mesmos pudessem ser analisados, removeu-se o sinal de + e a vírgula, categorizamos o mesmo como uma variável do tipo inteiro e ordenamos os valores iniciando a partir dos aplicativos que foram instalados mais vezes. O gráfico das notas em função do número de instalações é apresentado na figura 4. A partir do gráfico notou-se que existe uma tendência de que um aplicativo seja mais instalado se a sua nota for maior.

Sabemos que nem todos os aplicativos são gratuitos e por este motivo fizemos um gráfico do tipo CountPlot mostrando a porcentagem de aplicativos gratuitos e pagos que é apresentado na figura 5. O gráfico nos revela que 93.1% dos aplicativos são gratuitos.

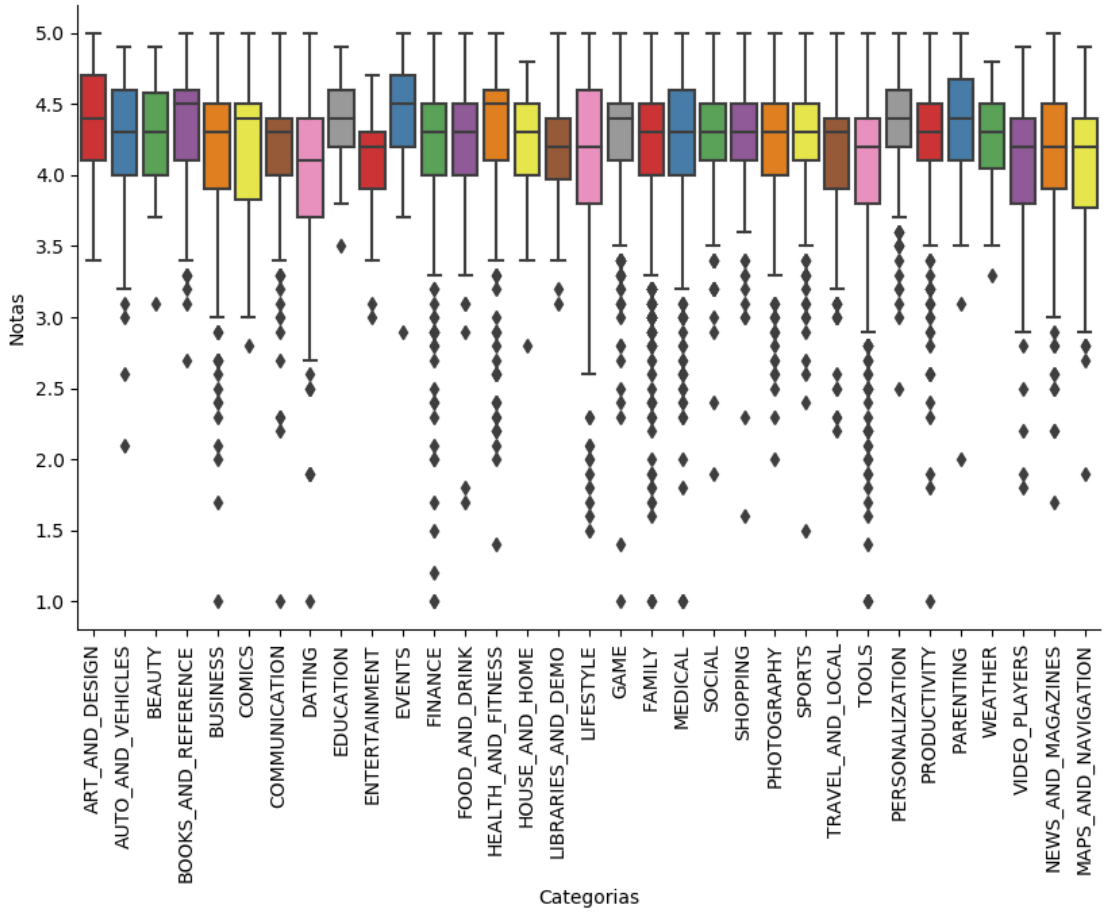

Figura 3: Boxplot das notas de cada categoria

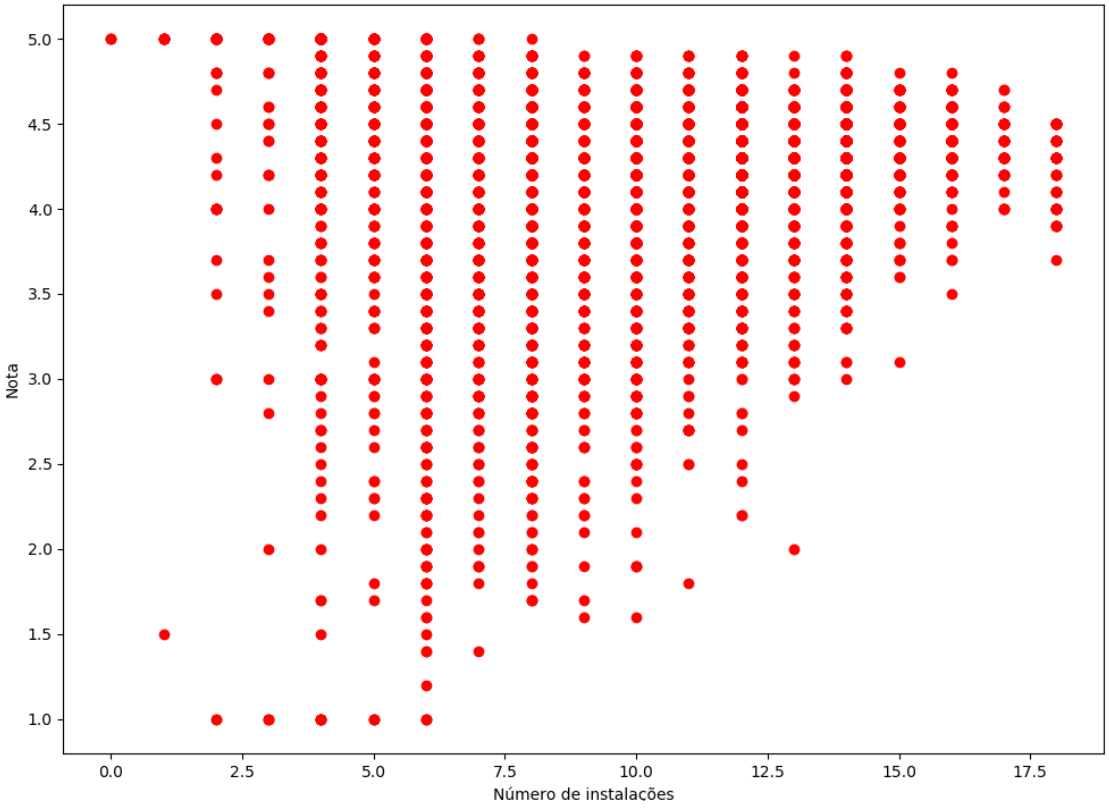

Figura 4: Notas em função do número de instalações

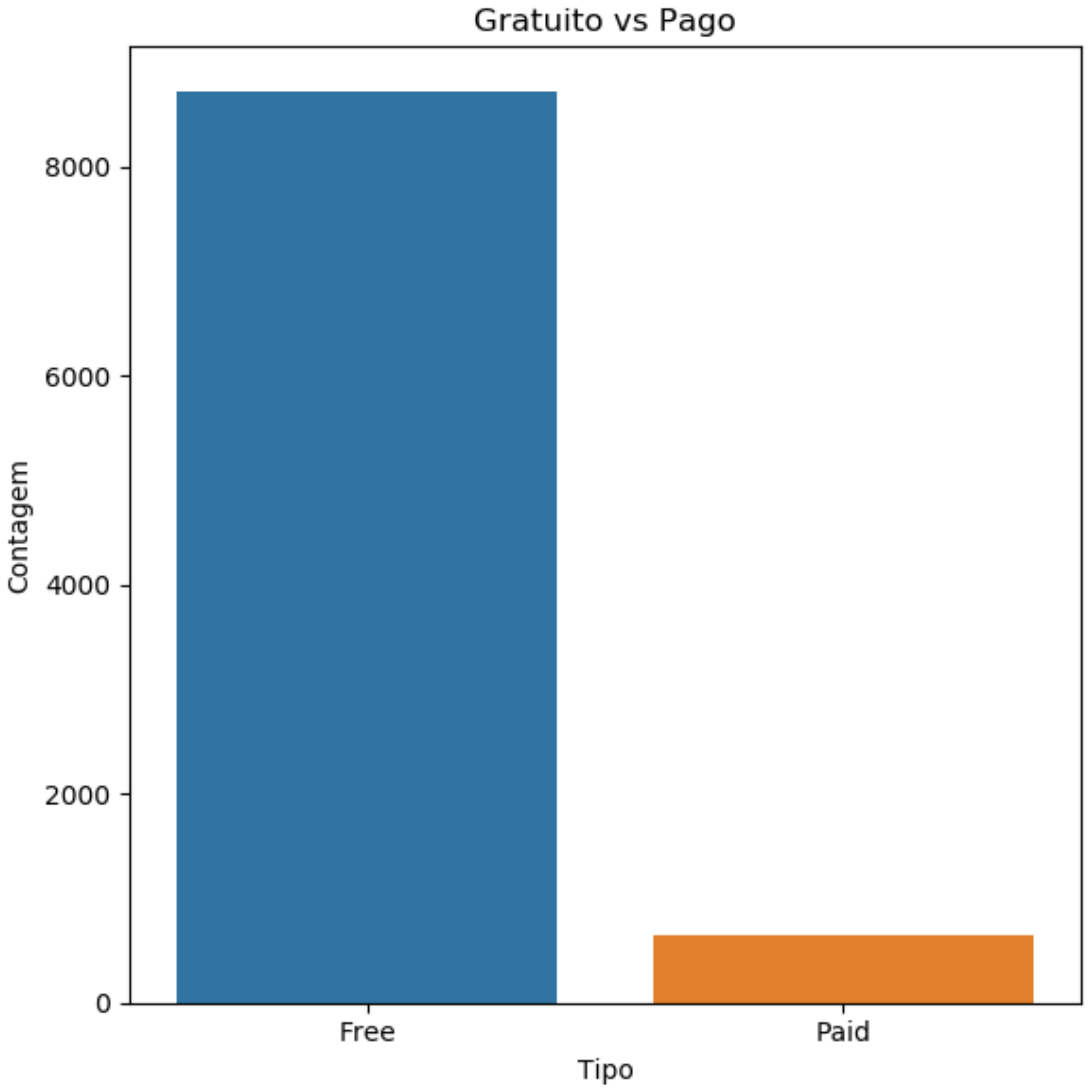

Figura 5: contagem de aplicativos gratuitos e pagos

Verificamos a correlação entre o número de instalações com a nota média. Este resultado apresentado na figura 6 nos diz que existe uma baixa correlação entre estas duas variáveis.

Na figura 7 apresentamos um gráfico Boxplot das notas médias em função da faixa etária, onde através deste gráfico notamos que a faixa etária afeta pouco as notas. No entanto para aplicações para usuários maiores de 17 anos existe um baixo índice.

Analisou-se também como a faixa etária afeta o numero de instalações de um aplicativo. Este resultado apresentado na figura 8 revela que as categorias Mature e Teen fazem mais instalações de aplicativos.

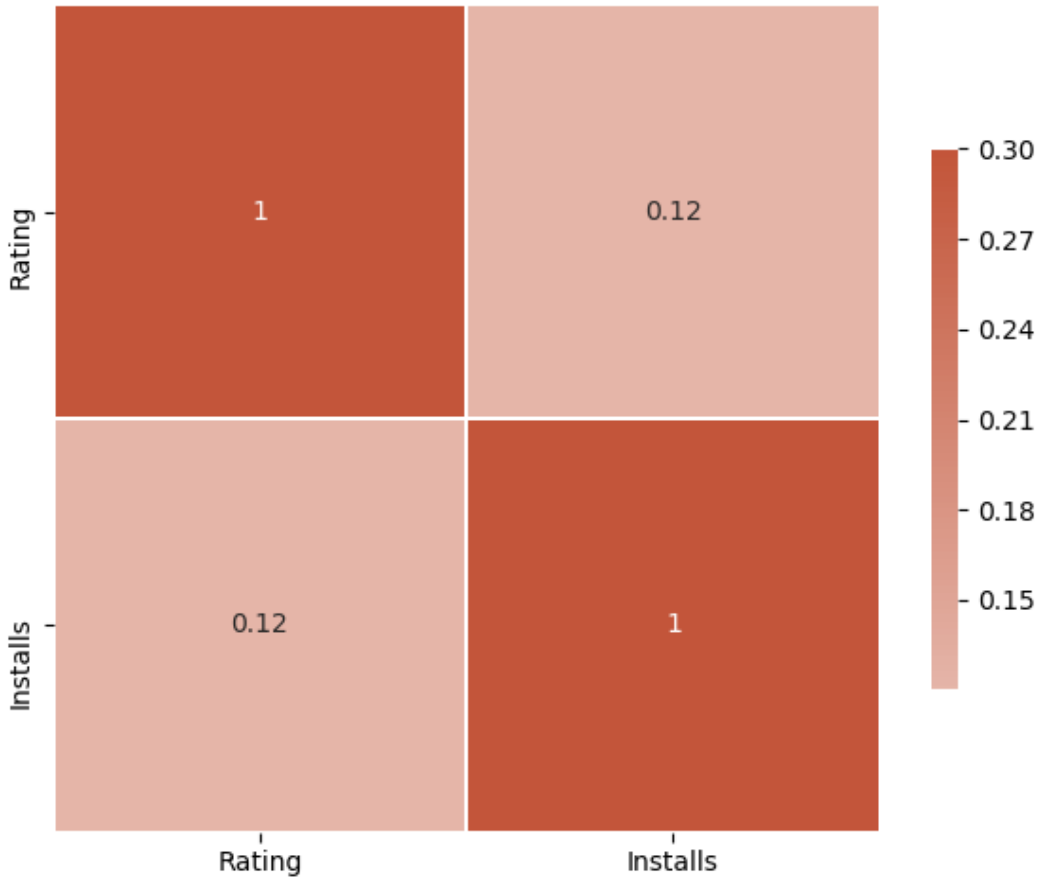

Figura 6: correlação entre notas e número de instalações

Analisamos as diferentes categorias e notou-se que existem um total de 33 categorias que são: ART_AND_DESIGN, AUTO_AND_VEHICLES, BEAUTY, BOOKS_AND_REFERENCE, BUSINESS, COMICS, COMMUNICATION, DATING, EDUCATION, ENTERTAINMENT, EVENTS, FINANCE, FOOD_AND_DRINK, HEALTH_AND_FITNESS, HOUSE_AND_HOME, LIBRARIES_AND_DEMO, LIFESTYLE, GAME, FAMILY, MEDICAL, SOCIAL, SHOPPING, PHOTOGRAPHY, SPORTS, TRAVEL_AND_LOCAL, TOOLS, PERSONALIZATION, PRODUCTIVITY, PARENTING, WEATHER, VIDEO_PLAYERS, NEWS_AND_MAGAZINES, MAPS_AND_NAVIGATION.

A partir das informações sobre as categorias gerou-se um gráfico de barras a nota média de cada uma delas. Este resultado apresentado na figura 9 nos mostra que a distribuição de notas é muito semelhante (considerando o desvio padrão) e que as categorias Events e Dating apresentam a maior e menor nota respectivamente.

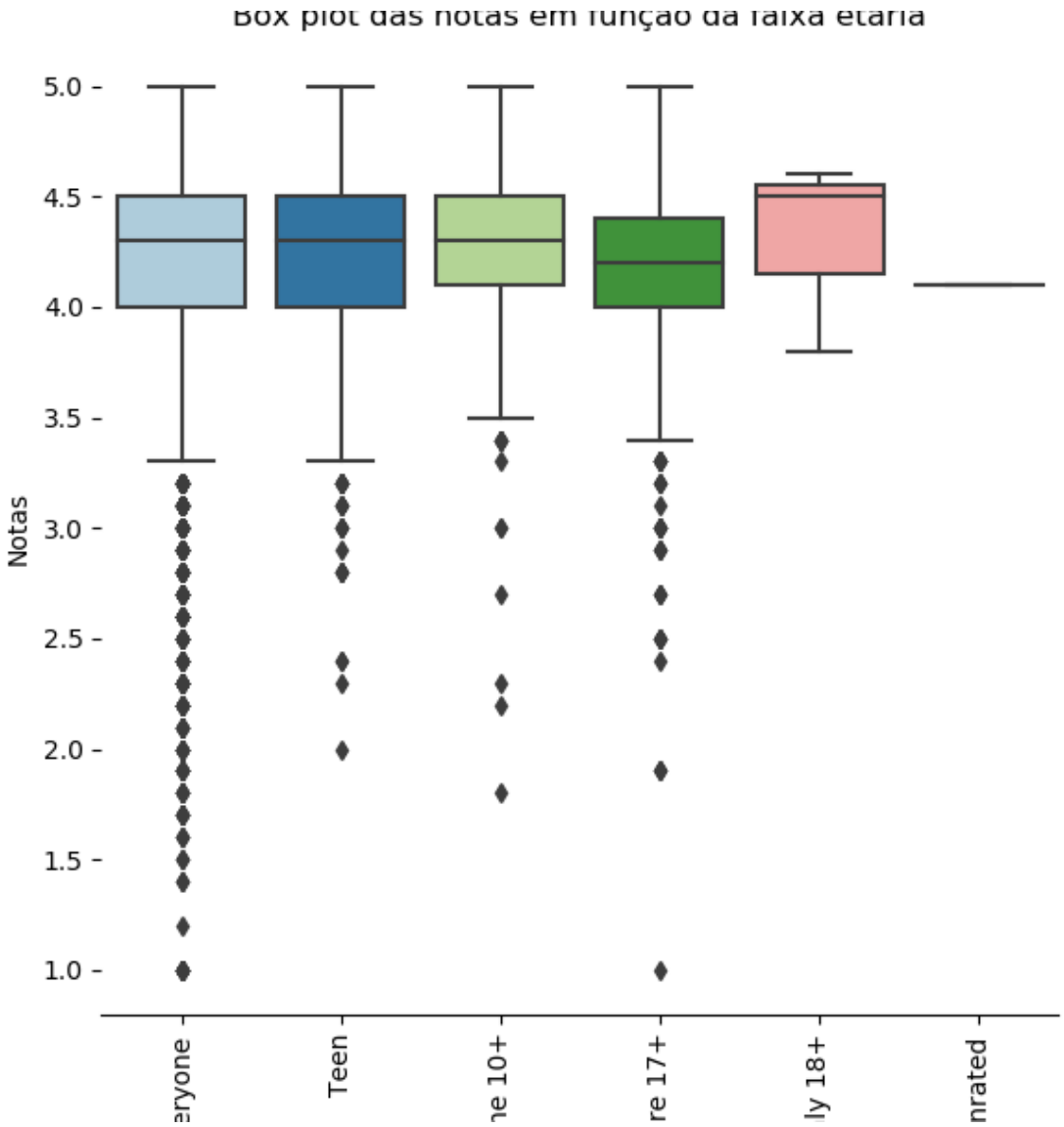

Figura 7: BoxPlot das notas em função da faixa etária

Uma informação importante que devemos levar em consideração é o número de revisões em função das categorias. Através de um gráfico de espalhamento, apresentado na figura 10, notamos que as categorias que possuem maior número de revisões são House and Home, Communication, Productivity e Finance. As demais têm um comportamento parecido.

Analisamos o numero de aplicativos por gênero e a partir do resultado apresentado no gráfico de barras da figura 11, notou-se que as categorias Ferramentas, Entretenimento e Educação apresentam maior contagem, enquanto que as demais apresentam um número parecido. Desta forma, analisou-se o porcentual de aplicativos gratuitos e pagos destas três categorias, apresentado na figura 12, e a categoria ferramenta é a que possui maior número de aplicativos seguida pelas categorias educação e entretenimento.

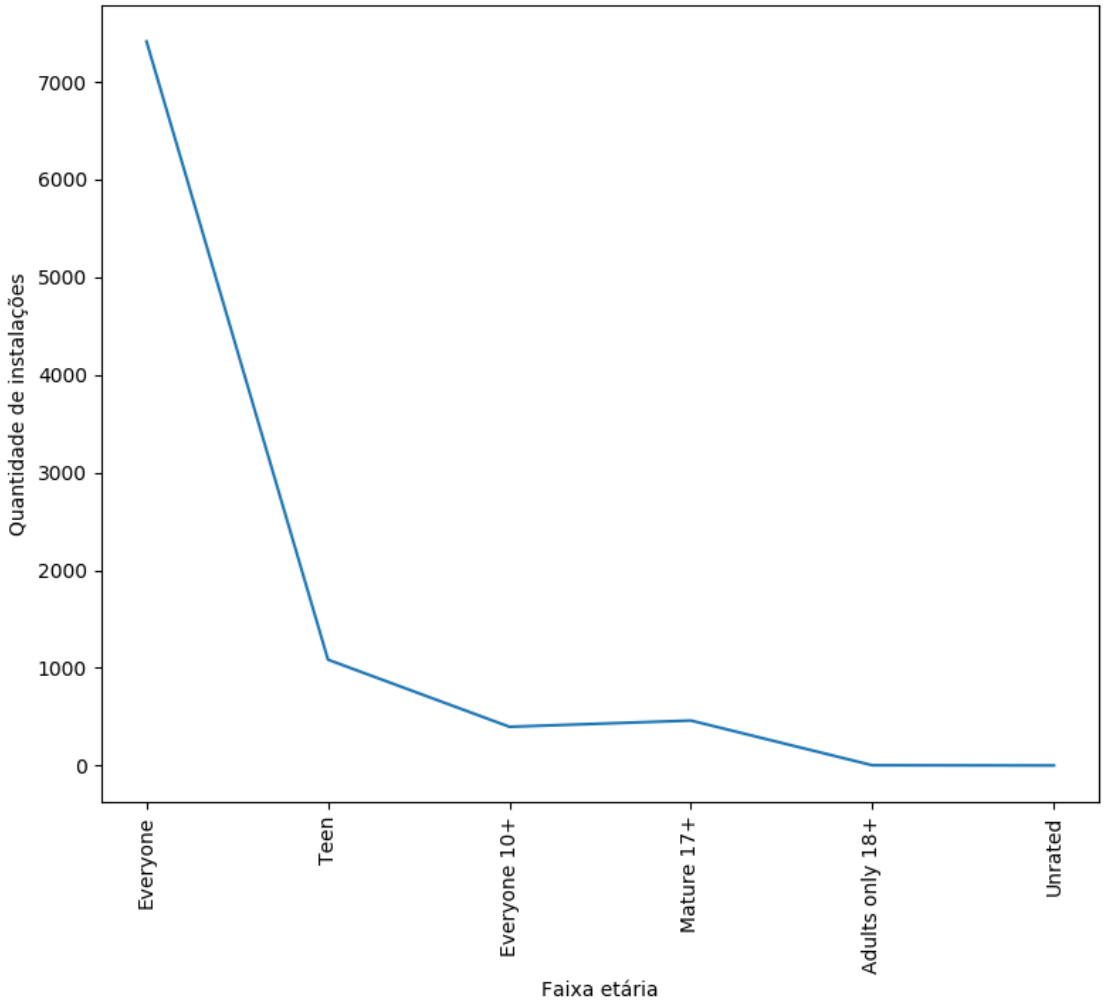

Figura 8: número de instalações em função da faixa etária

Um ponto importante da nossa análise é determinar se os comentários dos usuários são positivos, negativos ou neutros. Este resultado é apresentado na figura 13 onde notamos que a grande maioria dos comentários são positivos e que o número de comentários negativos e neutros são parecidos.

Por fim, aplicamos um processo de lematização aos comentários removendo palavras desnecessárias para detectarmos as palavras mais utilizadas e mais fortes com um objetivo de identificar os sentimos mais importantes dos usuários. Este resultado é apresentado na figura 14. As palavras acct, achieve, actual, aknowledge, advertisse, activate, adjust, addon e admin estão entre as mais utilziadas pelos usuários. Dentre estas palavras podemos destacar advertisse que pode ser considerada negativa uma vez que a grande maioria dos aplicativos gratuitos forçam os usuários a assistirem anúncios de patrocinadores como forma de arrecadar dinheiro.

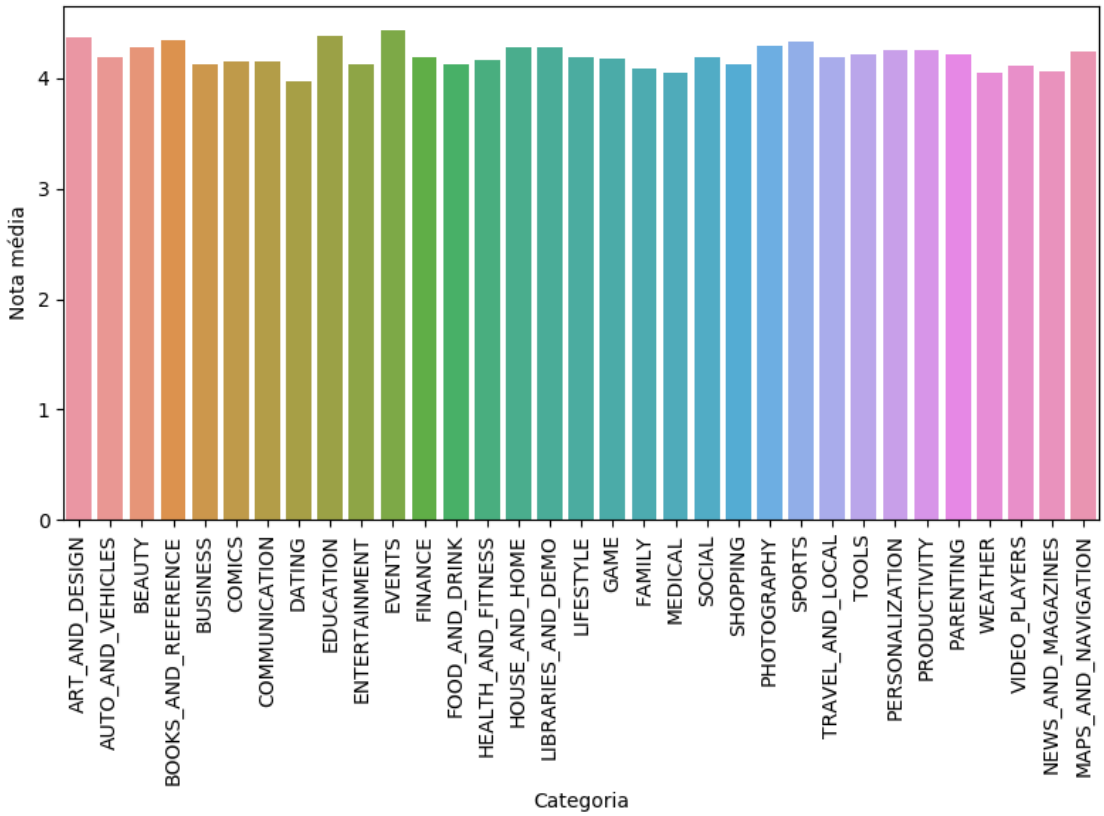

Figura 9: número de aplicativos de cada categoria

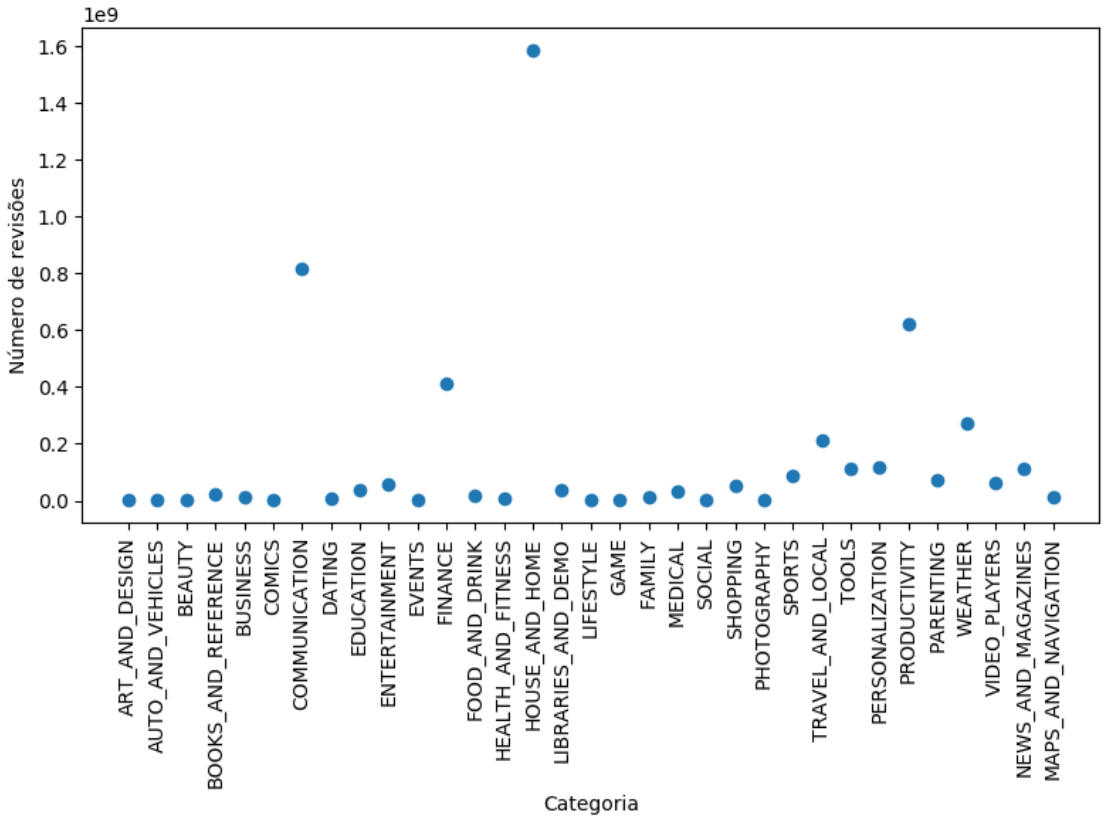

Figura 10: número de revisões de cada categoria

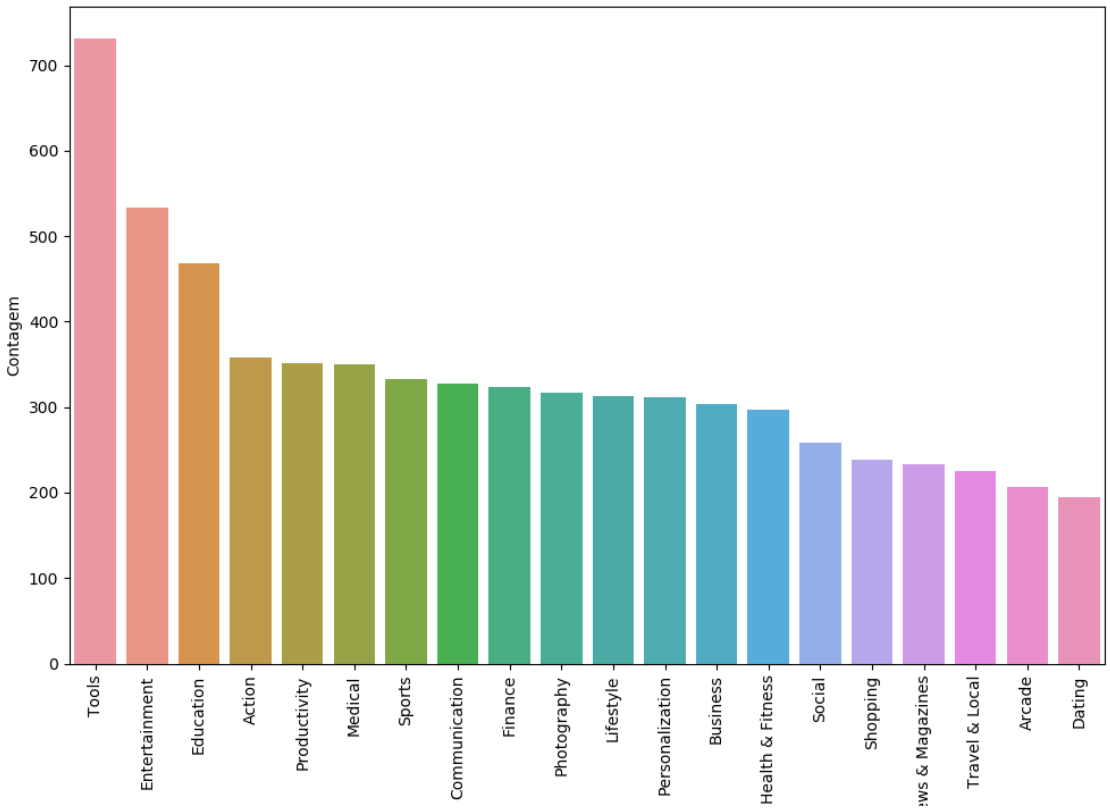

Figura 11: contagem de aplicativos das diferentes categorias

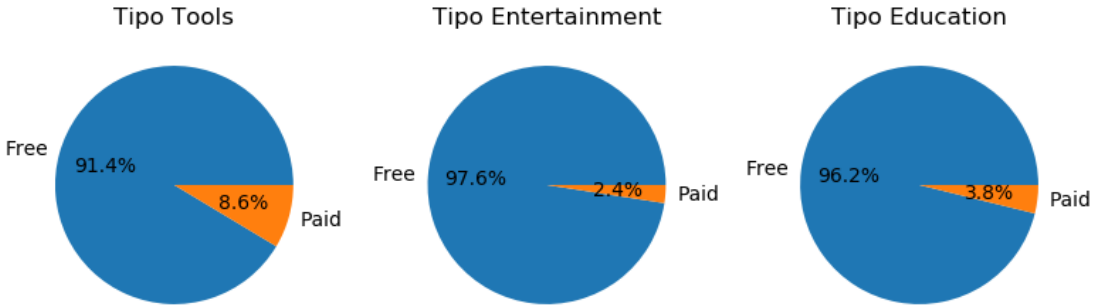

Figura 12: Porcentual de aplicativos pagos das categorias com maiores números de aplicativos

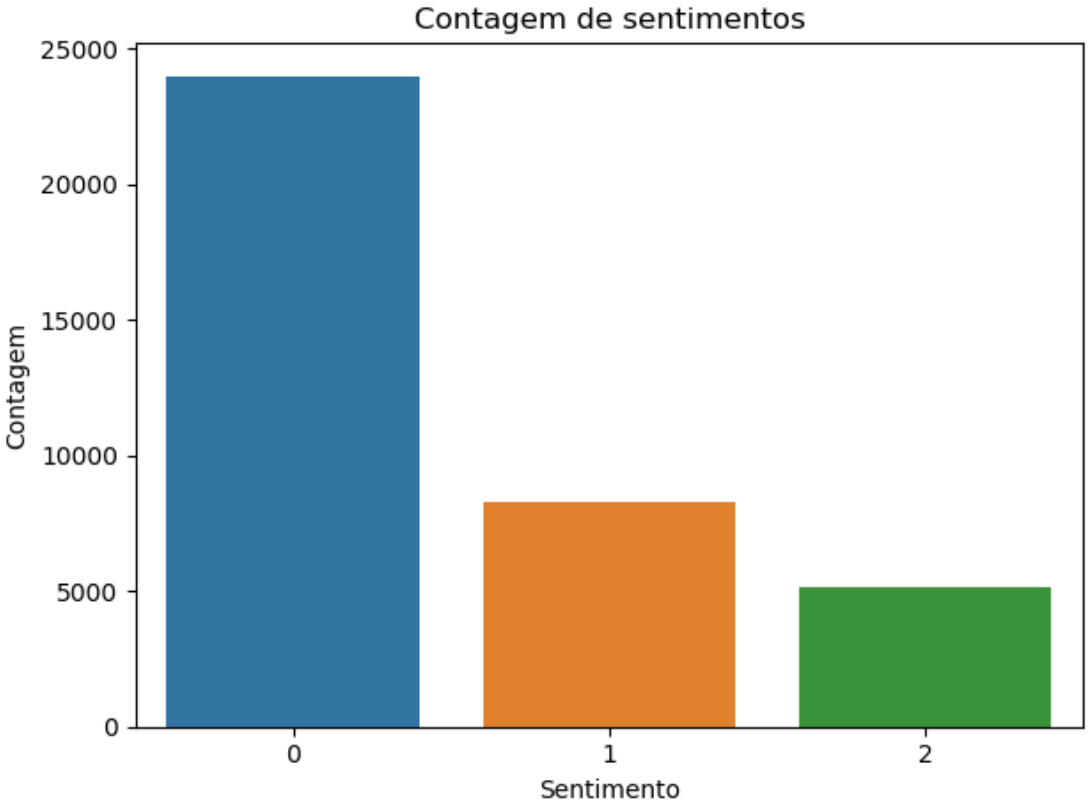

Figura 13: Contagem de sentimentos. 0 se refere aos comentários positivos, 1 aos comentários negativos e 2 aos comentáios neutros.

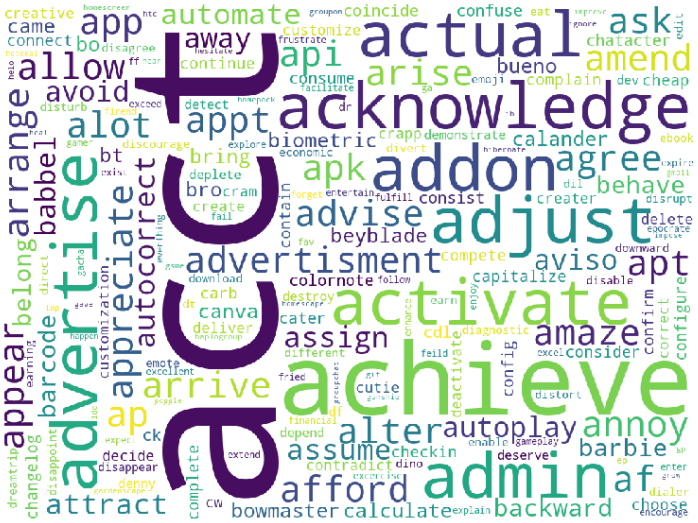

Figura 14: palavras mais utilizadas nos comentários.