Exploratory Data Analysis on Google Play Store Apps Dataset

Deepkumar Patel
Data Visualization
DATA-53000-002
Lewis University
Romeoville, USA
deepkumarpatel@lewis.edu

***Abstract*— The goal of this project is to discover patterns, test different hypothesis or assumptions using visualizations. Further, we investigate the dataset to summarize main characteristics of the play store dataset. Once we thoroughly analyze the dataset, we will be able to draw insights which can help us make business decisions as beginner app developers.**

I. Introduction

The dataset we will examine and work with contains a set of features from Google Play store or App store [1]. The dataset itself contains individual apps and their associated properties such as category, reviews, installs, price and many more. This dataset is publicly available on Kaggle [2] and the author, or the user has used web scraping to gather all its content. The author has provided the data under creative commons license, which is free to use, share and adapt.

Since we are exploring the data and testing our hypothesis, it is very important that we process the dataset before we do any type of analysis, especially when the data is acquired using web scraping because its prone to errors. Therefore, before visualizing any data, I have applied measure transforms, removed nulls, replaced any inconsistent formatting, and renamed columns for our better understanding. The transformation and cleaning helps improve the data quality; leaving us with only the high-quality information.

The future section of this project includes the following information: Section II describes the dataset we will use for visualization. Section III describes a set of questions and test hypotheses regarding the dataset. Section IV includes the visuals and its corresponding drawn conclusions. Section V is the conclusion of this project.

II. Data Description

After the preprocessing step, we are left with the following attributes shown in Table I. There are also other attributes which are not included in Table 1, for instance app name, type (free or paid – this is derived by the price column), last updated, and a few other attributes. The main reason we don't include the above attributes is simply because we want to avoid unnecessary columns which don't give us meaningful information.

Table I – Google Play Store Dataset

Attribute	***Type***	***Example Value***	***Description***
Category	Nominal (string)	Game	Category this app belongs to
Rating	Numeric (float)	4.3	Number of ratings on the app
Reviews	Numeric (integer)	100	Number of reviews on the app
Number of Installs	Numeric (integer)	10,000	Number of installs of the app
Price	Numeric (float)	$2.99	Price for the app ($0 is free)
Content Rating	Nominal (string)	Mature 17+	Minimum maturity level of content in app
Genres	Nominal (string)	Action	Genre this app belongs to
Android Version	Numeric (float)	4.1	Current targeted Android version

The original dataset contained 10,842 instances and 13 attributes. However, after the preprocessing step, we are left with 9,465 instances and 8 attributes. Using our new transformed dataset, we will be able to test our hypothesis and look for patterns.

III. Questions and Hypotheses

Our main objective of this section, is to derive a set of hypotheses and questions which later can either be proved or disproved. Sine we know nothing about this dataset, I have come up with the following questions and or hypotheses. The following set is by no means concrete, which means that exploring data is solely based on the end goal we are trying to achieve. As beginner Android developers, we want to explore the Google play market as our starting point with following questions.

1) *Is there any relationship between the number of installs and Android version?*
2) *Which category have the most installs and what's the most targeted Android version in those categories?*
3) *Does content rating have an impact on the number of installs?*
4) *Does having more ratings mean more installs?*

5) *Which types of app are mostly paid and how are the ratings for those apps?*
6) *Which type of audience spent the most money on apps?*
7) *Which type of genres get the highest ratings and what are their category distribution like?*
8) *Which categories have the highest reviews?*

As mentioned earlier, we can keep adding new questions as per our needs. However, for the purpose of this report I have selected the above set because I think it is a good starting point for any app developer as it will give us some vital information before we start developing an app.

IV. DATA EXPLORATION

1) Is there any relationship between the number of installs and Android version?

One may wonder why even ask this question; the reason is very simple. We want to reach as many audiences as we could, so if we can figure out which Android version is most installed, we can target that specific version. Since Android is the world's most popular OS, we want to make sure we strategically target widely used version. As developers, we could support all versions but due to the dependency issues, apps might not work properly in older versions. There are pros and cons to both sides, however, for our purpose we will look at the highly installed version.

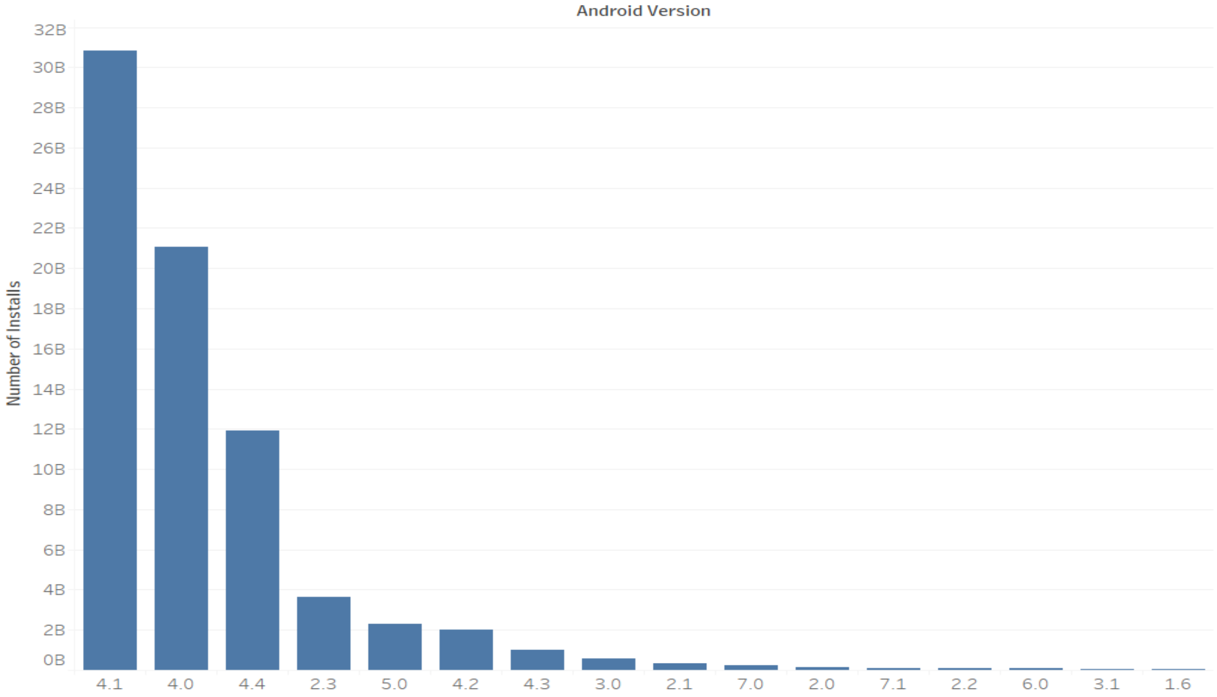

Fig 1. Number of installs by Android version

In the figure above, we can see that 4.1 is a widely used Android version. It is interesting to see how the newer versions like 5.0 and 7.1 are still not as popular as 4.0 and 4.1. This could also mean that a lot of people still own older phones, because the newer phones doesn't support old versions.

2) Which category have the most installs and what's the most targeted Android version in those categories?

If the above claim is true, that is, if 4.0 or 4.1 is the most installed version, then this question should prove current hypothesis in terms of categories. This will also give us an idea on how categories are distributed among the number of installs. Answering this question will help us get a sense of what people enjoy the most.

Below, fig.2 shows that "Game" category has the highest number of installs, followed by communication, family and photography which are among the most installed categories. One interesting observation we can see is that "Social" category is no where near the most installed apps. This is surprising, because these days social media apps are very popular such as Twitter, Facebook, Instagram and many more.

In terms of the version, we can see that 4.0 is indeed among the most popular android version. Now, in fig. 1, we saw that 4.1 was the most popular version, so how come it is not displayed in fig.2? The answer is, it only shows the major version and not the minor versions. I did this to remove the graph clutter and to keep things simple, otherwise, it would get difficult to distinguish versions in the individual bars.

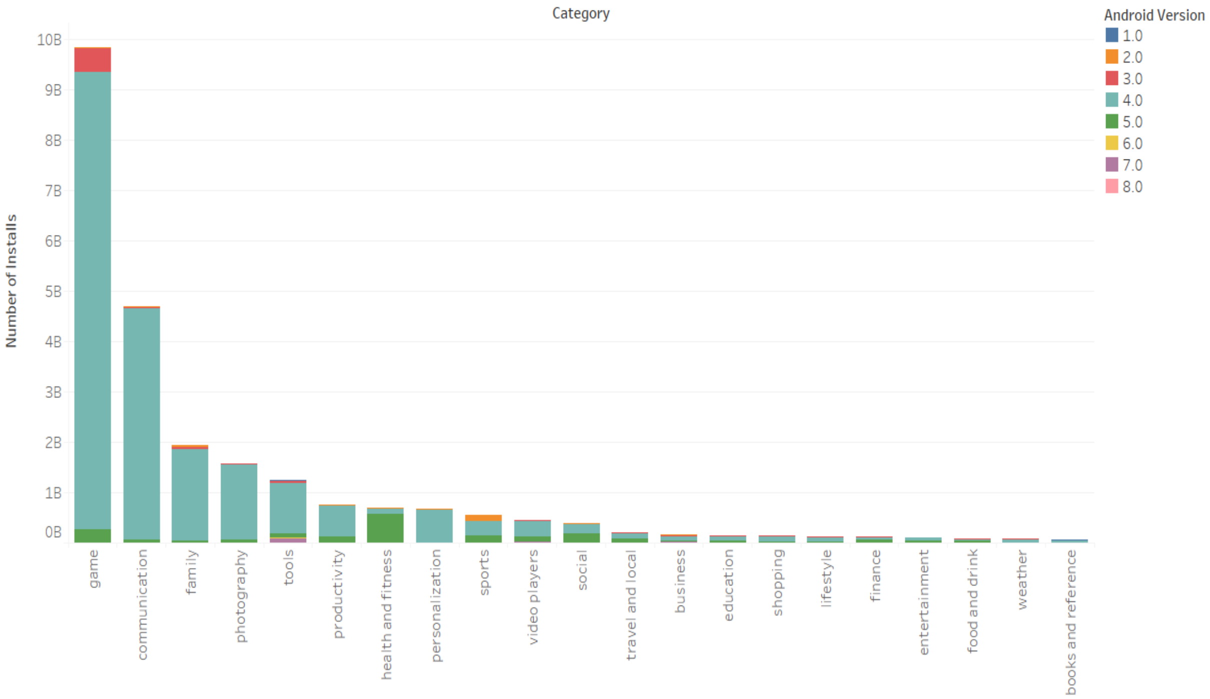

Fig 2. Number of installs per category and version distribution

3) Does content rating have an impact on the number of installs?

Content rating is nothing but the minimum maturity level of a content in an app. We want to know if restricted content would have more or less installs.

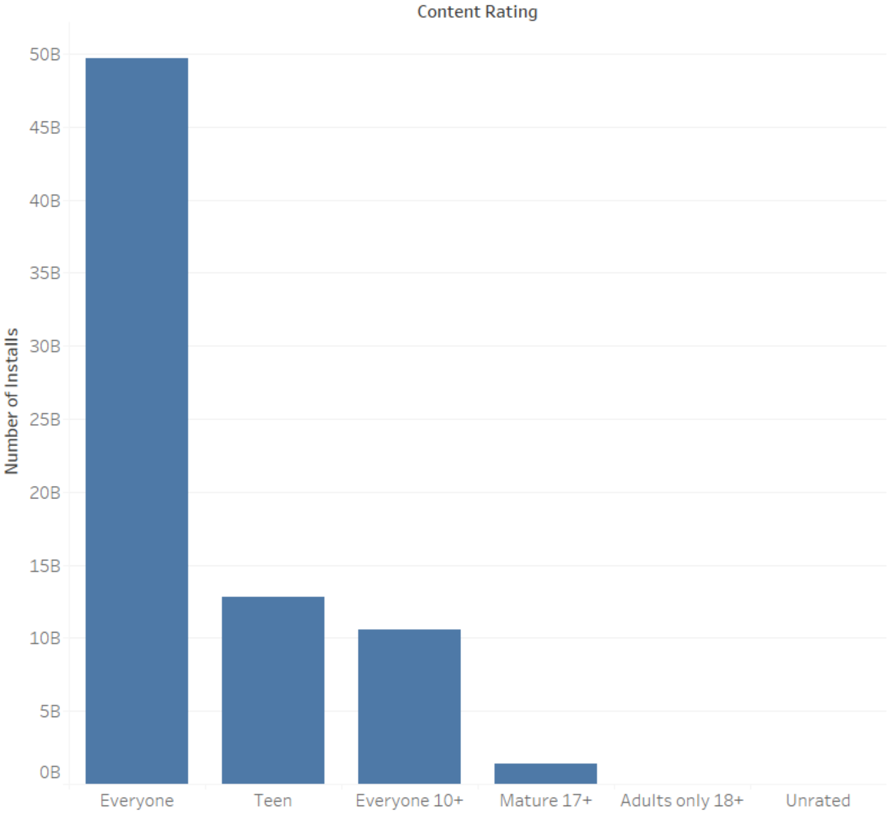

Fig 3. Number of installs per content rating

The above trend somewhat make sense because as the content gets less restrictive, more people download the app. We also learn that "Teen" content has among the second highest installs. It could be because most apps target the younger generation. Again, this may explain why “Game” category have the highest number of installs in fig.2 because young people love games.

4) Does having more ratings mean more installs?

This is an interesting question because we want to see if ratings attract more downloads. The questions we want to ask ourselves are, is there any positive or negative correlation or is there some kind of trend or pattern its following.

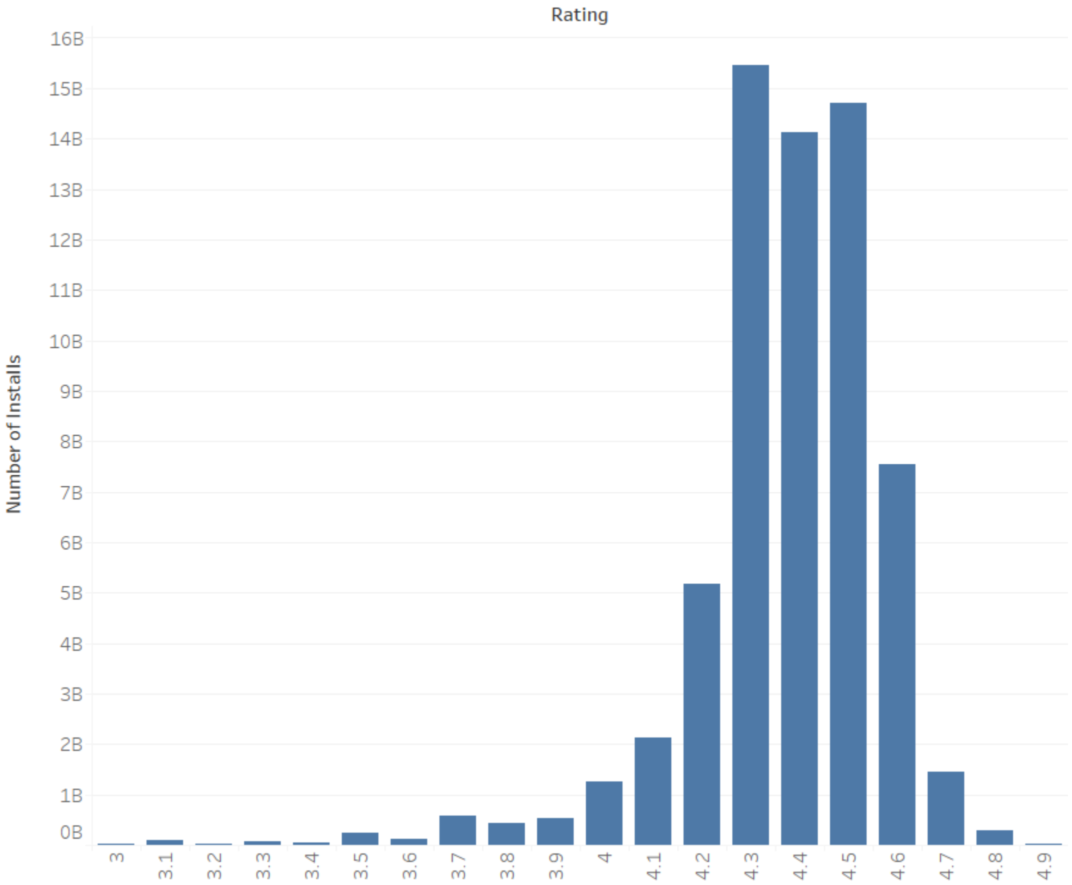

Fig. 4 Number of installs per rating

The above figure shows that there is no positive or negative correlation between ratings and the number of installs. However, we do see a trend, that is apps with ratings between 4.3 to 4.5 tend to get the most downloads. It is also interesting to see that more ratings don’t necessarily mean more downloads. For instance, app with 4.9 rating has way less downloads than the previous ratings. A general conclusion we can make is that if an app has a rating between 4 and 4.7 (inclusive), it will get more downloads.

5) Which types of app are mostly paid and how are the ratings for those apps?

If we want to generate revenue from apps, then it is necessary to see what type of genre are usually paid and how are the ratings for those apps. Since we are dealing with ratings, in the previous question we saw that ratings of 4 will get the most downloads, at least that’s what the trend showed us. It will be interesting to capture the same trend in this visualization.

Below fig.5 shows that finance, entertainment, medical and education are among the most expensive apps in the market. It is shocking to see that some of the app’s charge $400. Now, this could be because users are on monthly subscription and the price shows as a cumulative sum. In terms of ratings, we can see that the top 3 expensive apps tend to have a rating of 4, which further supports our previous idea that these apps should have the greatest number of installs.

There are always pros and cons for making the app free and paid. Generally, most free apps tend to get the most downloads, however, it is difficult to make revenues with free apps, unless we somehow integrate ads in our apps. Therefore, it all depends on our business model. Looking at fig.5 we could conclude that the entertainment genre would be a viable option because despite having the high price, it has high ratings and high downloads. The reason we pick entertainment is because it is something that targets almost every age group, and to further support this idea fig.3 also showed that if the content rating is for “Everyone” then it will get the most downloads.

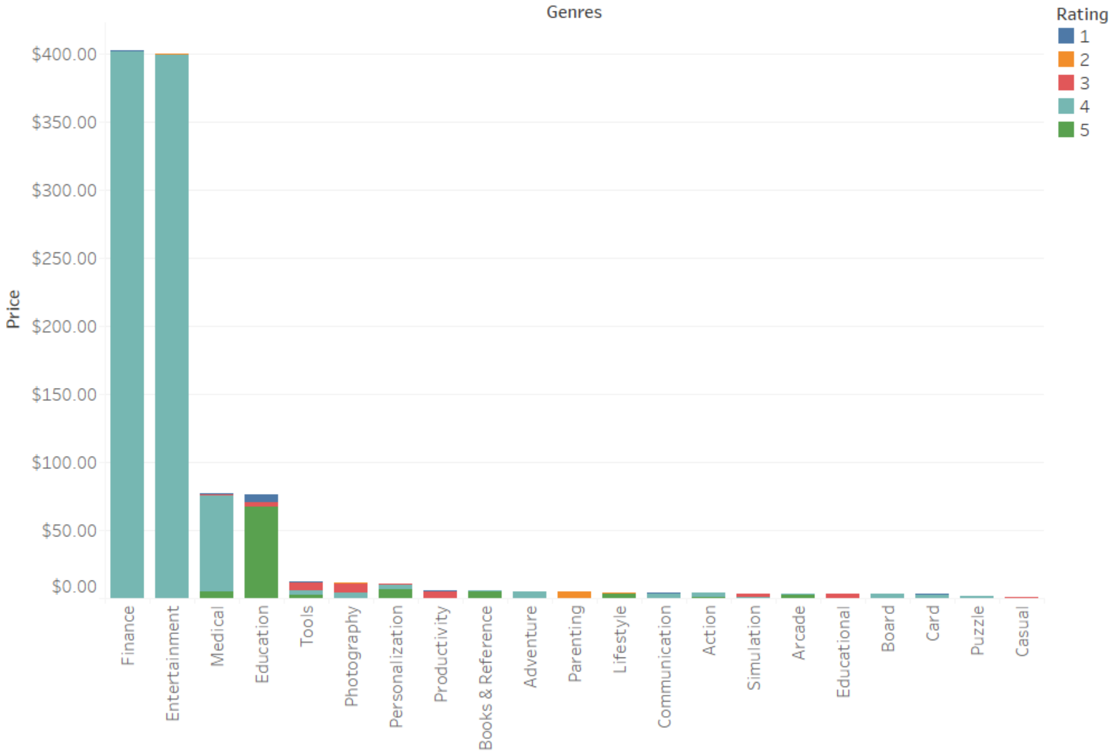

Fig. 5 Price distribution among genres and their ratings

6) Which type of audience spent the most money on apps?

Since we know which type of genres are mostly paid, we can now determine the type of audience that spends the most money. This is important for cases where we want to target a specific audience or a certain age group.

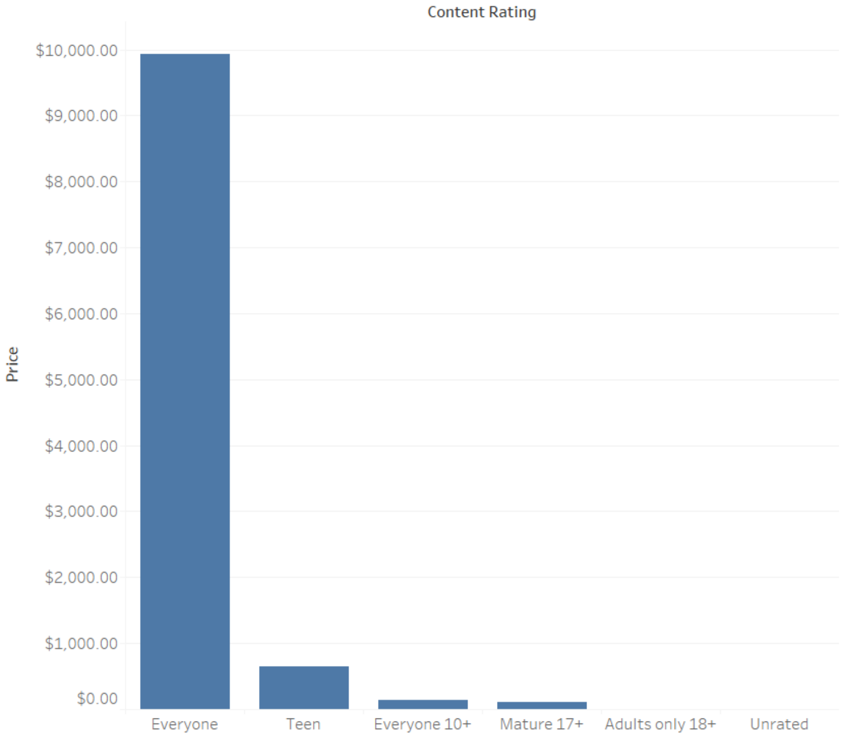

Fig. 6 Price distribution among content ratings

Above figure shows that apps with no age restrictions tend to make up the majority of apps that are paid. In addition, we also know that apps with content rating “Everyone” tend to get the highest installs, which we know this by using fig. 3. Using this visual, and in conjunction with fig 3, we can almost predict the number of installs our app might get.

7) Which type of genres get the highest ratings and what are their category distribution like?

Since there are so many genres, it is interesting to see which genre gets the highest ratings, and it is also important to look at the category distribution of each genre because a genre can belong to many categories.

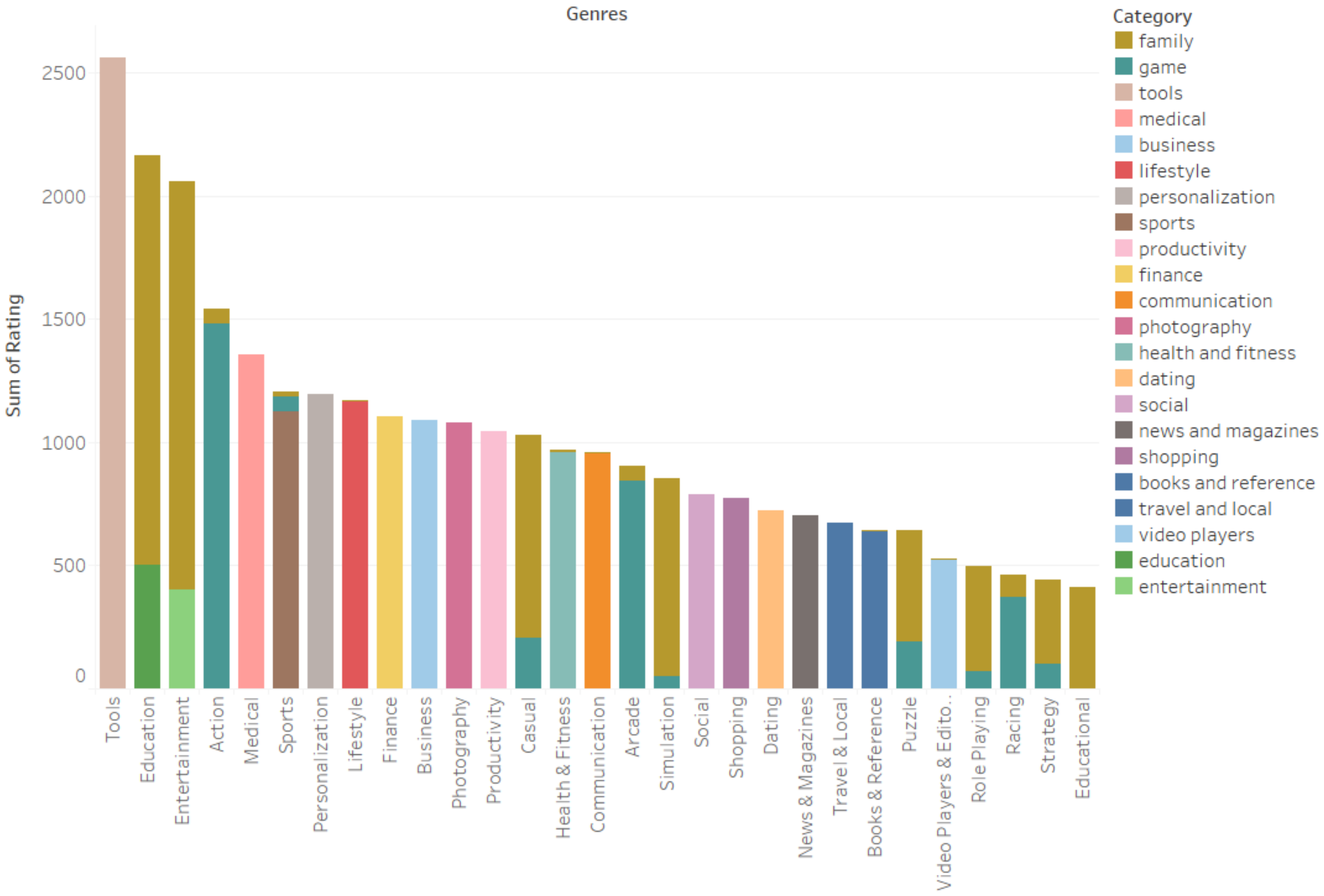

Fig 7. Ratings per genre and its category distribution

The above figure shows that tools, education and entertainment are among the highest rated apps in the market. It is interesting to see how when we compared genres and the number of installs in fig.2, “tools” was the fifth highest installed app. Therefore, looking at the above visual, we can conclude that apps such as tools and education tend to get high ratings but low number of installs. The entertainment genre still has a fair amount of ratings and installs (using fig. 2). As mentioned earlier, a genre could belong to multiple categories. If we look at the education genre, it belongs to both the family and education category. Obviously, if we want to get high ratings, then we would target a category that has a high coverage.

8) Which categories have the highest reviews?

Most of the time people first read the reviews before buying a product. This simple idea applies to apps as well. If the reviews are bad for the apps, then no one would download the app. While we visualize the graph, we will also able to tell if there is any kind of correlation between the number of installs and the number of reviews. Due to the limitations of this dataset, we do not know if the reviews given in this dataset are positive or negative, so for the purpose of this project we will assume it is a mix of both.

In the below fig.8, apps that belongs to game and family are among the most installed and reviewed apps in the market. We do see a positive correlation between the number of installs and the number of reviews. However, there are some categories which have a constant number of reviews despite the number of installs. We could perhaps pick a category which targets both game and family for high reviews and installs.

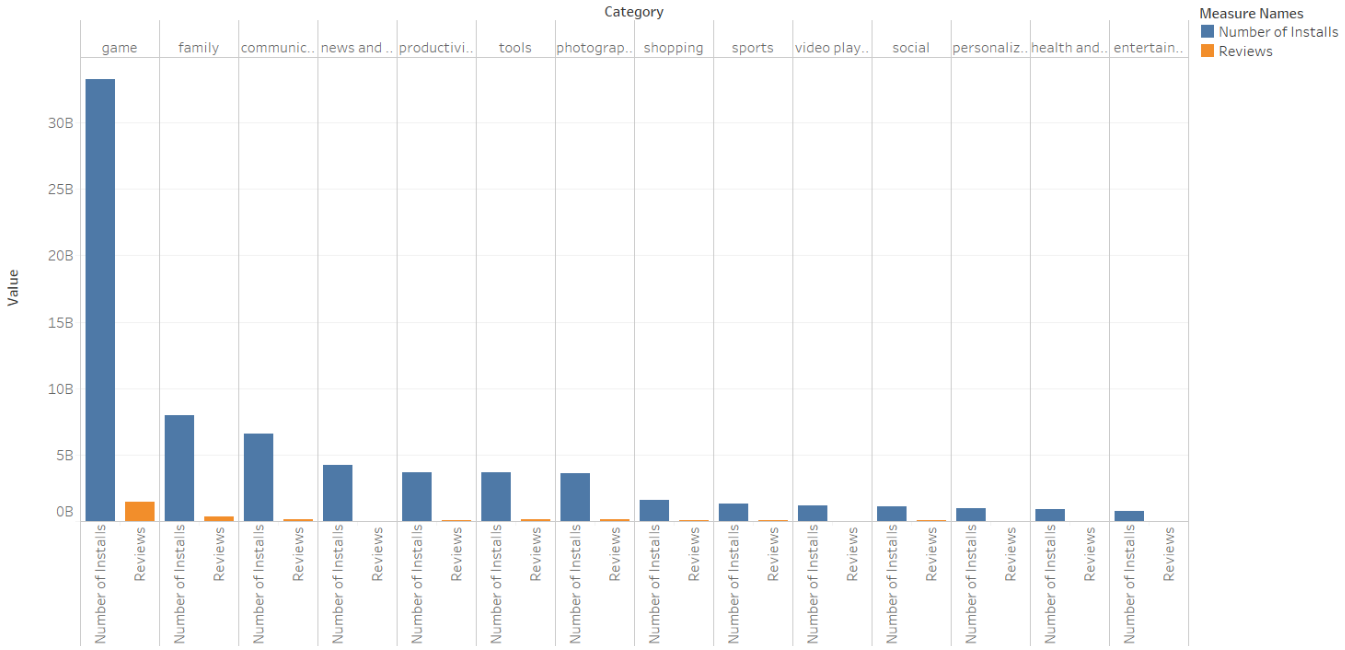

Fig. 8. Reviews and number of installs per category

After careful examination, I have created all the visuals in a simple and easy to understand manner. I have applied filters to remove clutters from the graphs, applied different colors to show distribution within each bin of histograms. In addition, I have removed any inconsistent formatting with different attributes and applied units where it seemed necessary. I also added sorting to help user easily draw conclusions about each visual.

V. Conclusion

Before we started using any of the models, we applied some pre-processing techniques to our dataset. Sometimes, we might not even need extra pre-processing, but most of the time, real-world data is not always clean. Therefore, it is very important to clean and transform the dataset in a way we can interpret them and possibly draw conclusions. When we try to do exploratory data analysis, we always want to start with questions, essentially, it is giving us a direction which will lead us to draw different conclusions about the dataset. We started out with 8 questions, but by no means we are confined to this number. I thought these set of questions were a great way to know about the app market. If the goal of this project is different, then we will have a different set of questions. Our goal was to get a general feel about the android app market, and the way we did this is by asking and testing different hypotheses.

We drew many conclusions from this dataset. For instance, we should try to target android 4.0 or 4.1 with a category of either game or family because they have the highest number of installs. We also learned that if we keep the content rating to everyone, then it will not only get us more installs but high reviews as well. In terms of paid apps, we should try to explore the entertainment industry because according to the data, this genre not only has high installs, but revenue is also high. If its possible we should try to target the family category because it has the highest coverage in terms of rating. Having all the above information, we now at least have an idea of the type of app we could potentially make. There are also many other factors that play a vital role on the success of an app, such as the functionality, availability in terms of country, language and many more, and since we don't have those datapoints, we can't draw conclusions. However, for an android app developer this report is a good starting point, because it covers the general topics of the Google play market.

References

[1] "Google Play" www.play.google.com/store/apps?hl=en_US&gl=US

[2] L. Gupta. "Google Play Store Apps" Kaggle, 3 Feburary 2019, www.kaggle.com/lava18/google-play-store-apps