**Capstone Project-1 Submission**

**Play Store App Review Analysis**

**Aanchal Kankrecha**

**Sonika Baheti**

**Data science trainees**



**AlmaBetter, Bangalore**

**Cohort: Chicago**

--------------------------------------------------------------------------\*\*\*--------------------------------------------------------------------

**Aanchal Kankrecha**- [kankrechaaanchal11@gmail.com](mailto:kankrechaaanchal11@gmail.com)

**Sonika Baheti** – [bahetisonika50@gmail.com](mailto:bahetisonika50@gmail.com)

**GitHub Link~**

**Aanchal Kankrecha:** - <https://github.com/AanchalKankrecha/Playstore-Data-Analysis>

**Sonika Baheti:** - <https://github.com/sonika-07/Play_Store_App_Review_Analysis>

**Abstract** –Play store dataset contains information regarding applications with their genres, category, ratings and reviews. The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market! Each app (row) has values for category, rating, reviews, size, installs, price, rated, last updated, and version. We have tried to discover the relationships among various attributes present in my dataset such as which application is free or paid, about the user reviews, rating of the application. In this notebook, we will do a comprehensive analysis of the Android app market by comparing over ten thousand apps in Google Play across different categories. We'll look for insights in the data to devise strategies to drive growth and retention.

***Key Words*:** Google Play Store Apps, Data Visualization, Exploratory Data Analysis.

# OBJECTIVE

This project focuses on the analysis of the Play Store data set. Main objective of this project is to extract useful insights from the data and utilize it in making business decisions and determining most popular and profitable applications. We tried to **let our data tell the story** using Exploratory Data Analysis by unfolding it through various graphs, charts and tables. Our objective will be to structure the data, clean it and present certain trends that we observe that can help us draw very preliminary conclusions

# INTRODUCTION

Mobile apps are everywhere. They are easy to create and can be money making. Because of these two factors, more and more apps are being developed. In this notebook, we will do a comprehensive analysis of the Android app market by comparing over ten thousand apps in Google Play across different categories. We'll look for insights in the data to devise strategies to drive growth and retention. Google play store is engulfed with a few thousands of new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make them successful, with the enormous challenge from everywhere throughout the globe. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created.

# Google Play store and User Review Analysis

Play store App review Dataset contains all the details of the applications on Google Play. There are 13 features that describe a given app. User review dataset contains 100 reviews for each app, most helpful first. The text in each review has been pre-processed and attributed with three new features: Sentiment (Positive, Negative or Neutral), Sentiment Polarity and Sentiment Subjectivity.

We develop Android apps & release on Play Store. As a Developer or say Business Perspective it’s very important to know whether users are enjoying the app or facing any issues. To know this, Play Store has a Ratings & reviews section for each app released on play store. Users can submit the ratings and has a freedom to write a review for a particular app. This approach is quite a lengthy to rate & review app i.e., navigate to Play store to submit feedback or redirect leaving a current app workflow to open Play Store App link using URI. We never wanted our customers to leave our application, but with this flow, we are forced to redirect the control to Play store app.

# Google Play store Dataset

This file contains all the details of the apps on Google Play. 13 features describe a given app.

• App: Name of the app

• Category: Category of the app. Some examples are ART\_AND\_DESIGN, FINANCE, COMICS, BEAUTY etc.

• Rating: The current average rating (out of 5) of the app on Google Play

• Reviews: Number of user reviews given on the app

• Size: Size of the app in MB (megabytes)

• Installs: Number of times the app was downloaded from Google Play

• Type: Whether the app is paid or free

• Price: Price of the app in US$

• Content Rating: A content rating (also known as maturity rating) rates the suitability of TV broadcasts,

movies, comic books, or video games to its audience. To show which age group is suitable to view media and

entertainment.

• Genres: A category of artistic, musical, or literary composition characterized by a particular style, form, or

content

• Last Updated: Date on which the app was last updated on Google Play

• Current Ver: Current Version means a version of the software that is currently being supported by its

publisher.

• Android Ver: Android versions (codenames) are used to describe the various updates for the open-source

Android mobile operating system.

# User Review Dataset

This file contains a random sample of 100 most helpful first user reviews for each app. The distribution of positive

and negative reviews in each category has been pre-processed and passed through a sentiment analyzer.

• App: Name of the app on which the user review was provided. Matches the App column of the

play\_store\_data.csv file

• Translated Review: The pre-processed user review text.

• Sentiment: Sentiment category of the user review - Positive, Negative or Neutral.

• Sentiment Polarity: Sentiment score of the user review. It lies between [-1,1]. A higher score denotes a

more positive sentiment.

# STEPS INVOLVED

•**Data Overview**

As a first step we take the overview of data, where we specially made our focus on understanding what each column means. So that we can be clear from what perspective we have to analyze our data. After understanding different column, we marked few important columns. These columns involve category, rating, reviews, size, installs, price, genres.

•**Data Preprocessing and Cleaning**

Preprocessing is important into transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data are dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e., there can be some missing values.

Now we started cleaning our data. So, we first identified the null values and we replaced this null value according to their data type. We dropped the duplicate values and fake records and removed the special characters from price and installs and convert app size in Kbs. Replace android and current version ‘vary with device’ value by **mode.** Changed data types and made the data ready for EDA. Finally, our data is ready to jump in the next step **EDA**.

• **Exploratory Data Analysis**

After establishing a good sense of each feature, we have proceeded firstly with correlation matrix to get the extent of relationship among variables. EDA is the process of trying to understand data in the ways possible in order to derive insights from it. Use exploratory data analysis to understand important factors or characteristics such as Avg, Std Mean Deviations, and also check for missing or null values and outliers. Exploratory data analysis is the process of looking at available data sets to identify patterns and anomalies, test hypotheses, and validate assumptions using statistical means. Using Python in exploratory data analysis processes and visual comparisons between variables is easy to understand and insightful.

• Set of Questions:

1. Is different variables related?

2. How many reviews are there for apps having 1 billion installs?

3. How many installs are made for each application?

4. How many apps are uninstalled?

5. Which category has highest installs?

6. What are the top 3 gaming Apps?

7. What are the top 3 communication Apps?

8. What are the top 3 tools Apps?

9. How many apps are there for each content rating?

10. How many installs are made for each type of content rating?

11. What is the revenue generated by paid apps?

12. How apps are distributed on the basis of category?

13. Percentage of free vs paid apps?

14. How many apps are free and paid for each category?

15. Apps with their type having rating 5?

16. Distribution of Rating?

17. Distribution of size of apps?

18. Average rating of each category?

19. No. of apps in each genre?

20. Rating vs price?

21. Rating vs installs?

22. When apps were last updated?

23. No. of translated reviews of different apps?

24. Which app has maximum translated reviews?

25. Distribution plot of sentiment polarity?

26. Distribution plot of sentiment subjectivity?

27. Percentage of sentiment count?

Throughout the analysis, we tried to answer questions that help us understand the factors determining the data trends and answers to the above mentioned questions are as follows:

**DATA PREPROCESSING**

Preprocessing is important into transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data are dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e., there can be some missing values.

The available data is raw and unusable for Exploratory data analysis, so before we do anything with the data we will have to explore and clean it to prepare it for data analysis.

* **Step1**: Checked and removed duplicates in App column.
* **Step2**: Dropped record where Category=1.9 as it corresponds to rating=19 which is not possible, install is free, price is everyone, which is not possible
* **Step 3**: Null value of Rating column is filled with ‘0’.
* **Step 4**:Changd the datatype of Review Column to float.
* **Step 5:** Removed special characters (“+”, “,”) and changed datatype to integer in install column.
* **Step 6:**  We can see that the size column, which should be numeric, is of the data type ‘object’, it also has characters ‘k’ and ‘M’ in the values which stand for kilobytes and Megabytes, we will replace the ‘k’ with “” and ‘M’ with 1000. Some values also have ‘+’ sign in them, which will be removed. Next, we will convert this column into ‘float’ datatype.
* **Step 7:** Filled null values of Type column with ‘free’.
* **Step 8:** The values in the column ‘Price’ might have the ‘$’&’,’ sign in some values and the column is of the datatype ‘object’. We removed the ‘$’ sign and then convert the column into ‘int’ datatype.
* **Step 9:**Replaced Current Version varies with device value with NaN and then with its mode.
* **Step 10:** Replaced Android Version varies with device value with NaN and then with its mode.
* **Step11:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total 26863 NaN value are present so we drop them using dropna() function.

# EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use **Python** language (**Pandas** library) for this purpose.

## HEATMAP

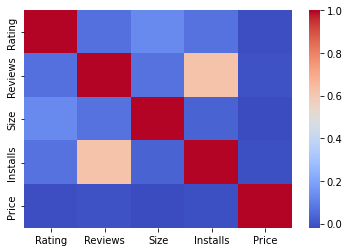
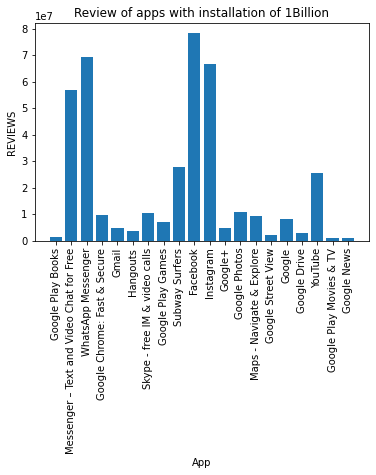


Fig -1: Heatmap

This shows that there exist positive correlation between Installs and Reviews and very less correlation among other Variables. Maximum no. of reviews are **8871** from the range 0 to 20,00,000 having maximum no. of installs as **7666**. Thus validating the conclusion of heatmap that there exist **positive correlation** between Reviews and Installs

**no. of reviews of apps having 1 Billion installs**



**Fig -2**: To check no. of reviews of apps having 1 Billion installs

This graph suggests that **Facebook** has the maximum no. of reviews when installs is 1B.

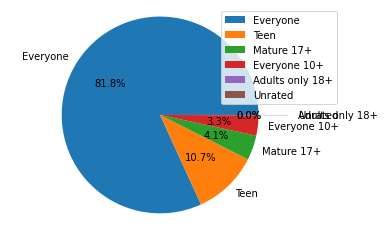
## category v/s Installs

## 

**Fig -3**: Category v/s Installs

Games are the most downloaded category followed by communication and tools.

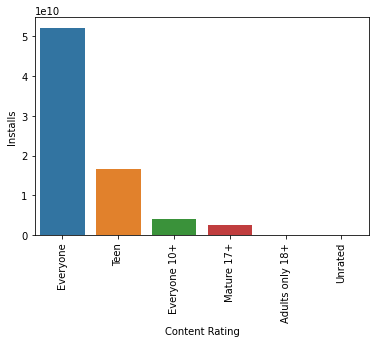
**Number of apps belonging to each Content**



**Fig -4: Content rating**

Everyone content rating has 81.8% of the total apps.

**Installs v/s content Rating**



**Fig -5: Installs v/s Content rating**

Everyone Content Rating has the maximum installs.

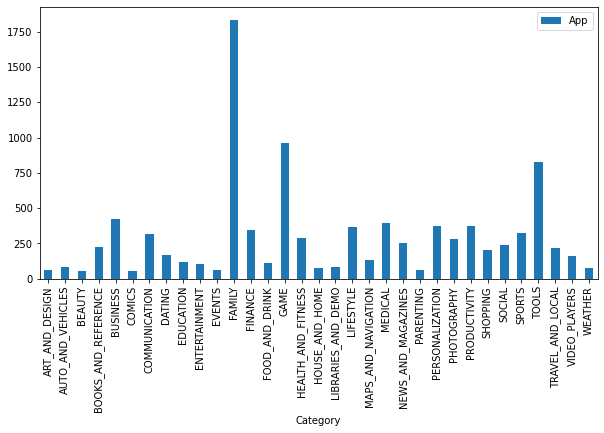
## revenue of Top paid apps

|  |  |
| --- | --- |
| **App** | **Revenue** |
| Minecraft | 69900000.000000 |
| I am rich | 39999000.000000 |
| I Am Rich Premium | 19999500.000000 |
| Hitman Sniper | 9900000.000000 |
| Grand Theft Auto: San Andreas | 6990000.000000 |

**Fig -6: Table of Revenue of top paid apps**

Minecraft generates the maximum revenue

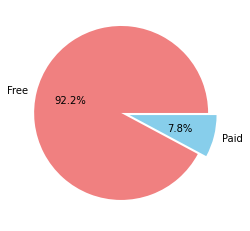
**Category wise app distribution**



**Fig -7 Category v/s No of apps**

play store has most apps for family followed by games and tools

**Free v/s Paid Apps**

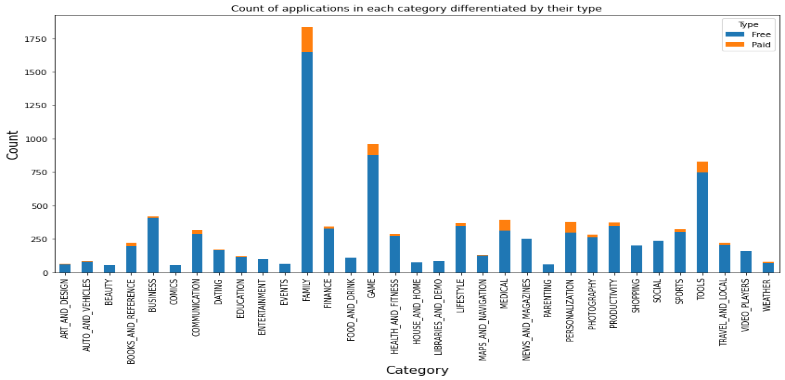


**Fig -:8 Free v/s Paid Apps**

More free apps are available on play store.

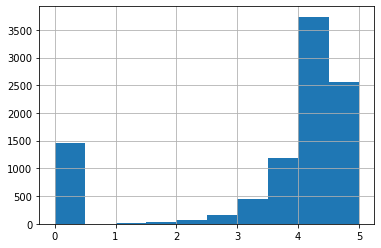
Approximately 92% of the apps on play store are free while 8% are paid apps.

**No. of free and paid apps for each category**



**Fig -9: type v/s Category**

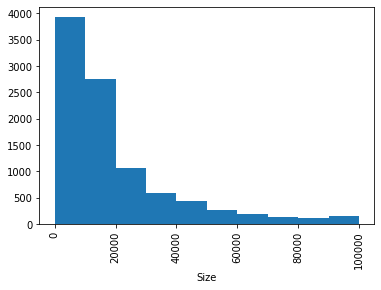
**Histogram of rating**

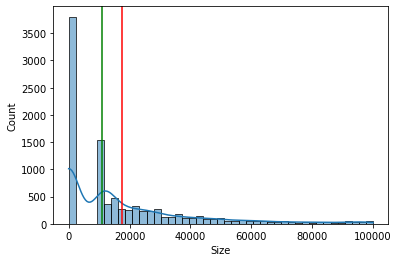


**Fig -10: Rating histogram**

Most of the apps has rating between 3 and 5. Average rating of app is 4.2 on play store.

## Distribution of Size



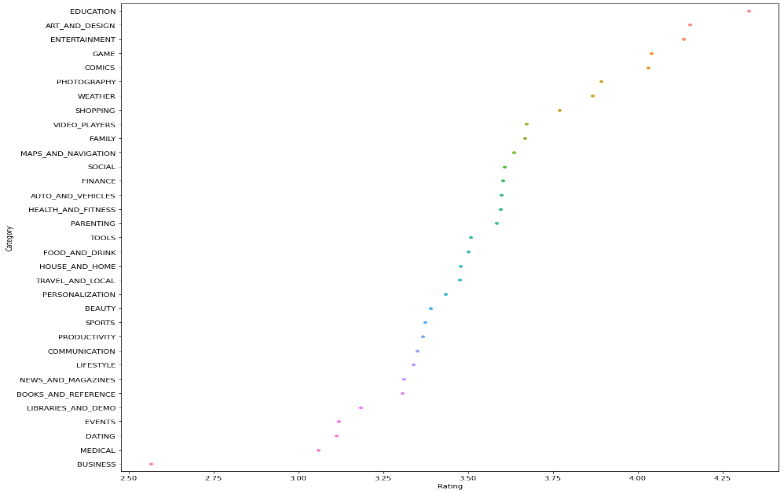


**Fig-11: Distribution of Size**

From the above plot we can infer that data of size is right tailed because mean>median

Most of the apps are of size between 1kb to 20 mb.Mean Size of app is : 17435.69264934258 kb Median Size of app is : 11000.0 kb

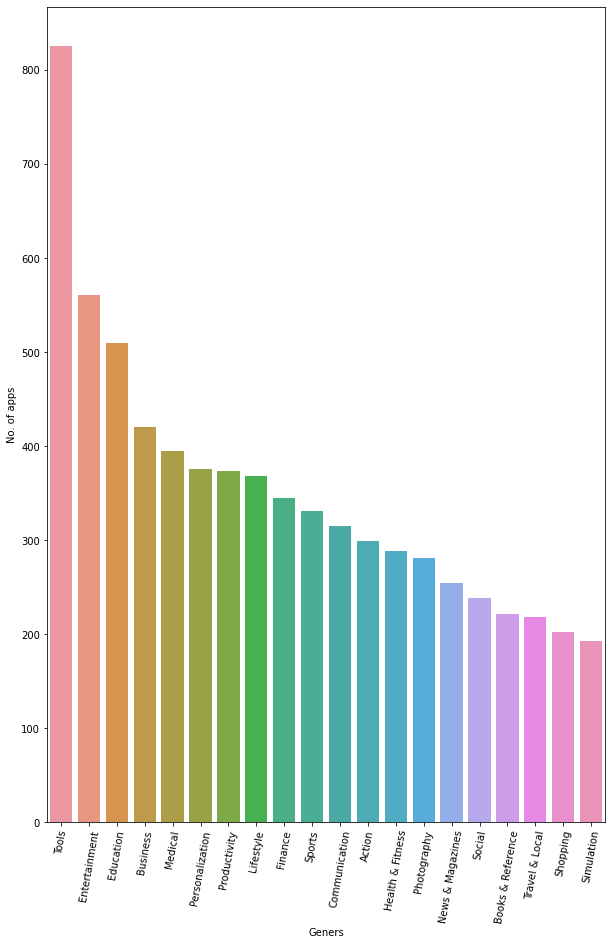
**Average rating of each category**

****

**Fig -12 Average rating of Category**

Education has the maximum average ratings of 4.3 followed by ART\_AND\_DESIGN 4.153125 & ENTERTAINMENT 4.135294

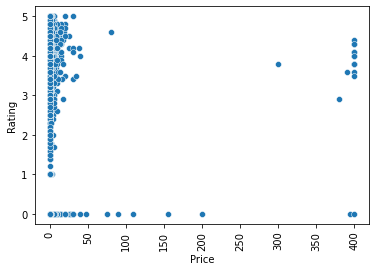
**Genres with Number of Apps**

****

**Fig 13: Genres with Number of Apps**

**Tools are the Most trending Genres among all genres, followed by Entertainment and Education.**

**Rating v/s Price**

****

**Fig -14: Rating v/s Price**

Different Ratings are given to free apps while high paid apps are given much good ratings.

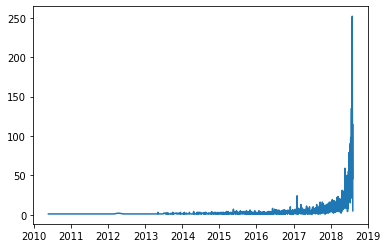
**Rating and Installs**



**Fig -15: Rating and Installs**

Higher rated apps have more installs

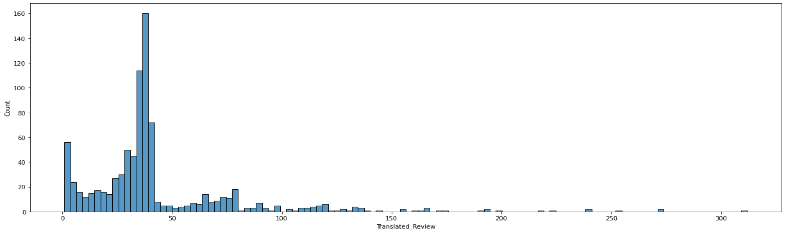
**Last Updated of A****pps**



**Fig -16: Last Updated of Apps**

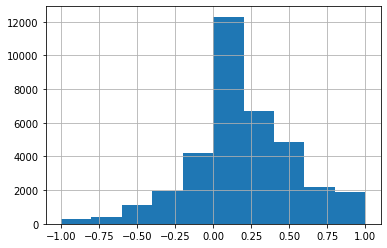
Mostly Apps are updated in the year 2017,2018,2019

**Distribution plot of translated reviews**

****

**Fig -17: Distribution plot of translated reviews**

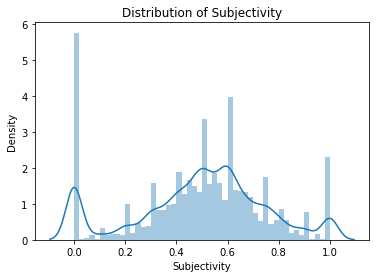
**Distribution of sentiment polarity**



**Fig -18: Sentiment Polarity**

Sentiment polarity is almost normally distributed.

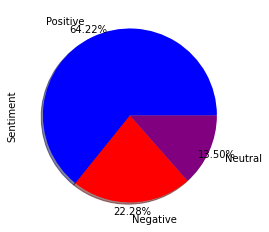
**Distribution of sentiment subjectivity**

****

**Fig -19: Sentiment subjectivity**

From the above plot sentiment subjectivity mostly lies between 0.4 to 0.8

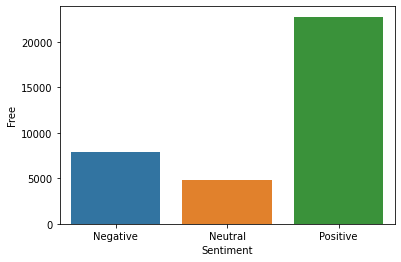
**Sentiment Counts**

****

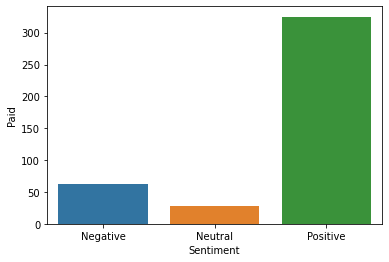
**Fig -20: Sentiment Count**

Here we can see that there are almost 64% positive Reviews given to apps ,i.e., most of the apps have positive effect on users.

**Sentimental analysis for free and paid apps**

****

**Fig -21: Sentiment of free**

****

**Fig -22: Sentiment of Paid**

Free and paid apps mostly have positive sentiments.

Paid apps have low negative and neutral sentiments compare to free.

**Conclusion~**

* Facebook has the maximum number of reviews.
* There are total 15 apps which are uninstalled.
* There are total 20 apps having 1 B installs.
* Games are the most downloaded category followed by communication and tools.
* Most of the apps are from content rating as everyone equivalent to 81.8%.
* Content rating as Everyone has the maximum no. of installs.
* Minecraft application generates the maximum revenue.
* Most of the apps are for family followed by games and tools.
* 92% of the apps on play store are free while 8% are paid apps.
* Average rating of apps on play store is 4.2
* Most of the apps are of size between 1Kb to 20Mb.
* Education category has maximum average rating as 4.3
* Tools are the most famous genres having 826 apps on play store.
* Free apps has varying ratings while paid apps generally have high ratings.
* Mostly apps are updated in the year 2017,2018,2019.
* Bowmaster app has the maximum translated reviews as 312.
* Sentiment subjectivity generally lies b/w 0.2 to 1
* There are no apps having Sentiment subjectivity as -1.
* There are almost 64% positive Reviews given to apps, i.e., most of the apps have positive effect on users.

**Future Work~**

Any particular user generally installs an app after going through its Review. Hence the developer must be aware of creating best user-interface with best contents, so that it can maximize installation of app and hence maximize the profit.

**Challenges Faced~**

* There were null values in the columns: Rating, Type, Content Rating, Last Updated, Current Version and Android Version. We have to fill the null values with its appropriate replacement.
* Handling 10841 records is quite tedious, so in dealing with such large data, we must be careful so that none of the information is lost.
* Duplicacy was one of the challenge. Hence removing it should not result in losing useful information.
* Choosing the correct visual for the analysis is one of the challenging task.

**References~**

* GeeksforGeeks
* Analytics Vidhya
* Stackoverflow
* Towards data science
* Python libraries documentation

Thank You