**Play Store App Review Analysis**

**Iqbal Babwane,**

**Sameer Ansari,**

**Lukman Haider**

**Data Science Trainees,**

**AlmaBetter, Mumbai**

**Abstract:**

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, size, and more. Another dataset contains customer reviews of the android apps.

Explore and analyze the data to discover key factors responsible for app engagement and success.

**1.Introduction**

**Google Play**, also branded as the **Google Play Store** and formerly **Android Market**, is a [digital distribution](https://en.wikipedia.org/wiki/Digital_distribution) service operated and developed by [Google](https://en.wikipedia.org/wiki/Google). It serves as the official app store for certified devices running on the [Android operating system](https://en.wikipedia.org/wiki/Android_(operating_system)) and [its derivatives](https://en.wikipedia.org/wiki/Google_Operating_System) as well as [ChromeOS](https://en.wikipedia.org/wiki/ChromeOS" \o "ChromeOS), allowing users to browse and download applications developed with the [Android software development kit](https://en.wikipedia.org/wiki/Android_software_development) (SDK) and published through Google. Google Play has also served as a [digital media](https://en.wikipedia.org/wiki/Digital_media) store, offering music, books, movies, and television programs. Content that has been purchased on [Google Play Movies & TV](https://en.wikipedia.org/wiki/Google_TV_(service)) and [Google Play Books](https://en.wikipedia.org/wiki/Google_Play_Books) can be accessed on a [web browser](https://en.wikipedia.org/wiki/Web_browser), and through the [Android](https://en.wikipedia.org/wiki/Android_(operating_system)) and [iOS](https://en.wikipedia.org/wiki/IOS) apps.

Applications are available through Google Play either free of charge or at a cost. They can be downloaded directly on an Android device through the [exclusive](https://en.wikipedia.org/wiki/Proprietary_software) Google Play Store [mobile app](https://en.wikipedia.org/wiki/Mobile_app) or by [deploying](https://en.wikipedia.org/wiki/Software_deployment) the application to a device from the Google Play website. Applications operating hardware capabilities of a device can be targeted to users of devices with specific hardware components, such as a motion sensor (for motion-dependent games) or a front-facing camera (for online video calling). The Google Play Store had over 82 billion app downloads in 2016 and reached over 3.5 million apps published in 2017, while after a purge of apps is back to over 3 million. It has been the subject of multiple issues concerning safety, in which [malicious software](https://en.wikipedia.org/wiki/Malware) has been approved and uploaded to the store and downloaded by users, with varying degrees of severity.

Google Play was launched on March 6, 2012, bringing together Android Market, Google Music, Google Movies and the Google eBookstore under one brand, marking a shift in Google's digital distribution strategy. Following their re-branding, Google has expanded the geographical support for each of the services. Since 2018, Google has gradually sunsetted the Play brand: [Play Newsstand](https://en.wikipedia.org/wiki/Google_Play_Newsstand) was rebranded as [Google News](https://en.wikipedia.org/wiki/Google_News) in 2018; [Play Music](https://en.wikipedia.org/wiki/Google_Play_Music) was withdrawn in favor of [YouTube Music](https://en.wikipedia.org/wiki/YouTube_Music) in 2020; and Play Movies & TV was rebranded as Google TV in 2021. In 2022, [Play Games](https://en.wikipedia.org/wiki/Google_Play_Games) is expected to shut down its mobile app in favor of an Android emulator for Windows with the same name. The remaining standalone mobile app will be Play Books.

**2. Description of Dataset**

### Dataset consist of two csv file.

1. Playstrore.csv: In this file total number of columns is 13 and number of rows is 10839.
2. User\_reviews.csv: In this file total number of columns is 5 and total number of rows is 64294

Playstrore.csv:

|  |  |  |
| --- | --- | --- |
| Sr. No | Column | Description |
| 1 | App | Name of the App and Game. |
| 2 | Category | Category of App |
| 3 | Rating | Rating of App |
| 4 | Reviews | Total Number of reviews |
| 5 | Size | Size of app in MB |
| 6 | Install | Number of installs |
| 7 | Type | Free type or paid type |
| 8 | Price | App price |
| 9 | Content Rating | Audience of app |
| 10 | Genres | Different genres under app falls |
| 11 | Last Updated | Last update date |
| 12 | Current Version | Current version of app |
| 13 | Android Version | Android version required |

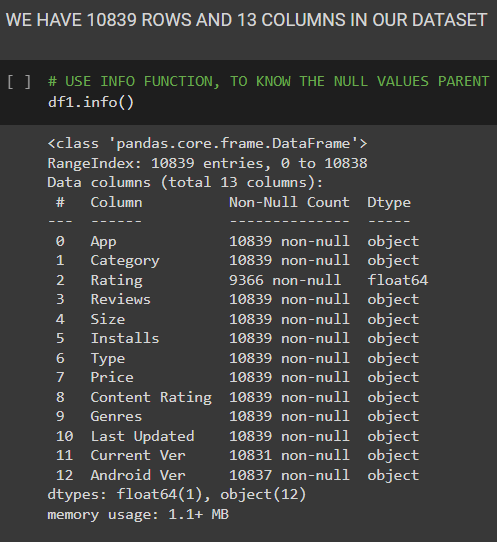
User\_reviews.csv

|  |  |  |
| --- | --- | --- |
| Sr. No | Column | Description |
| 1 | App | Name of the App |
| 2 | Translated reviews | User comments on app |
| 3 | Sentiment | User sentiment |
| 4 | Sentiment Polarity | Polarity refers to the strength of an opinion. |
| 5 | Sentiment Subjectivity | Subjectivity refers to the degree to which a person is personally involved in an object. |

## **3. Filling Null Values**

In this we observe all the values of dataset, if there is any missing value, null value, empty cell or blank cell.

After observing the first file which is Playstore.csv we found the column rating, current version and android version have null values. As shown in fig



## **4. Data Cleaning**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Data cleaning means fixing bad data in your data set.

Bad data could be:

* Empty cells
* Data in wrong format
* Wrong data
* Duplicates

Now we discuss the cleaning of our Dataset.

We start cleaning data from Reviews column. The datatype of this column is object and we want it to convert it to integer. We do some coding and changed it to integer successfully. Now we take size column from dataset. In this column there is outlier(‘M’) is present and datatype is object. We want to remove the outlier(‘M’) from it and change the datatype as float. We do some coding and store the data of size in new column called as size\_in\_mb. Now we take Installs column. In this column, outliers (‘+’) are present and datatype is object. We want datatype as integer. So, we did some coding and remove the outlier (‘+’) and change datatype as integer. Now we take price column. ‘$’ outlier is present in price column and datatype is object. We want to remove the outlier and change the datatype as float. We did some coding and remove the outlier (‘$’) and change datatype as float. After we take Android ver column. The datatype is object and we convert it as float and store in new column named it as min\_andriod\_ver. After that we take last updated column and using this column, we make new column and named it as year. In this we store year of each application. Now we done with dataset and clear all duplicates, empty cell and wrong data. Now our data is clean and using this data we create new Dataframe named it as df\_new.

# **5. Data Visualization**

Data visualization is the discipline of trying to understand data by placing it in a visual context so that patterns, trends, and correlations that might not otherwise be detected can be exposed.

Python offers multiple great graphing libraries packed with lots of different features. Whether you want to create interactive or highly customized plots, Python has an excellent library for you.

To get a little overview, here are a few popular plotting libraries:

* Matplotlib: low level, provides lots of freedom
* Pandas Visualization: easy to use interface, built on Matplotlib
* Seaborn: high-level interface, great default styles
* plotnine: based on R’s ggplot2, uses Grammar of Graphics

1. Plotly: can create interactive plots

Using these libraries, we do some visualization.

1. Observation-1

In this observation, we take category, minimum size of app and type columns and do visualization on it.

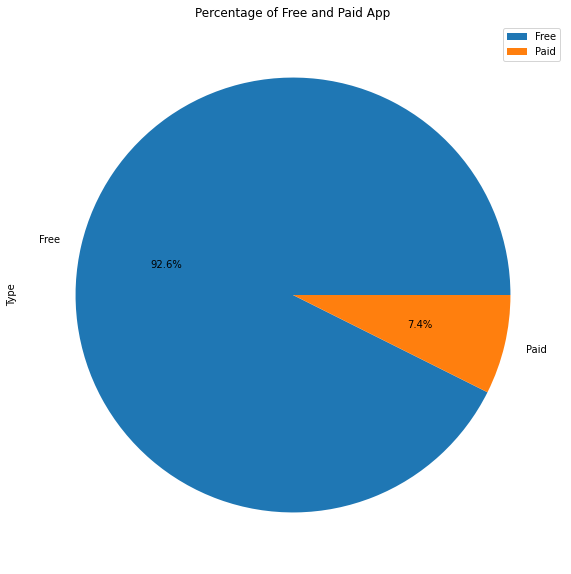
# 

In this graph, we take category on x-axis and average size of app on y-axis. One more thing, we distribute category by its type (free and paid). Now we observed that, highest average size of free apps in category is ‘Game’ and the highest average size of paid apps in category is ‘Education’.

Now we observed that, the lowest average size of free apps in category is ‘Tools’ and the lowest average size of paid apps in category is ‘Beauty, Comics and House and Home’.

1. Observation – 2

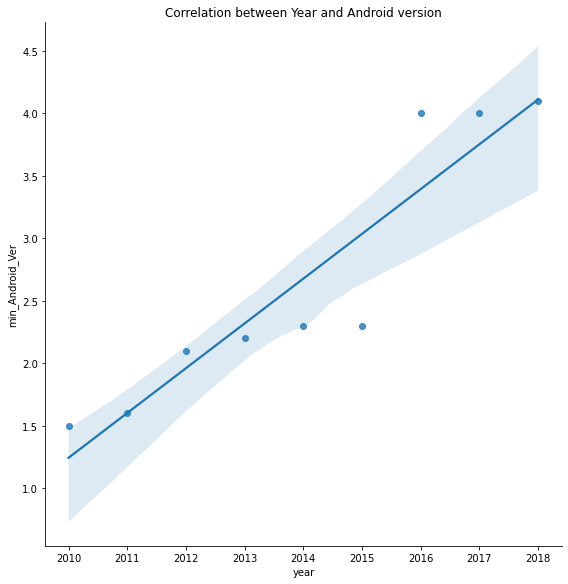
In this observation, we take Type column and do visualization on it.



In this graph, Type has two values Free and Paid. We use these values and make a pie chart of it. We easily observed that free apps are ruling the market by huge percentage that is 92.6% and percentage of paid apps present in Play store is 7.4% which is very low as compare to free apps present.

1. Observation – 3

In this observation, we take year and minimum android version columns and do visualization on it.

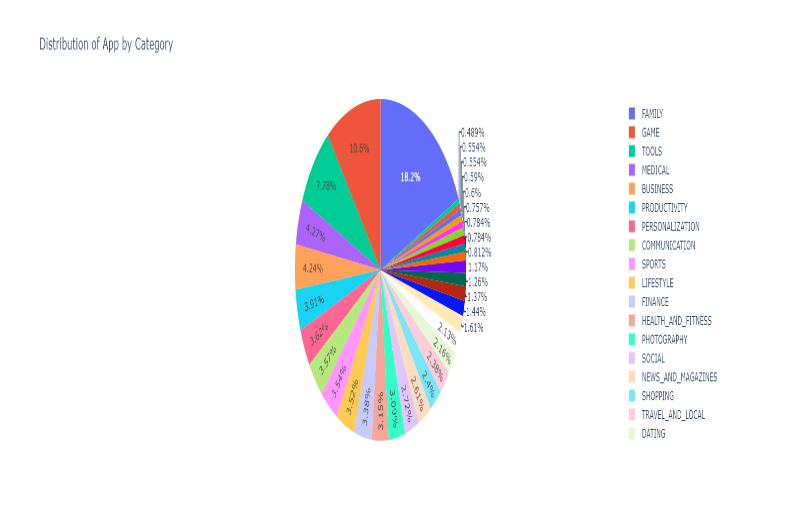


In this graph, we take Year on x-axis and Minimum Android Version Required on y-axis. Now we observed that the positive correlation of year and minimum android version required.

The minimum android version required in year 2010 is 1.5 and the minimum android version required in year 2018 is 4.1.

1. Observation – 4

In this observation, we take Category and App columns and do visualization on it.



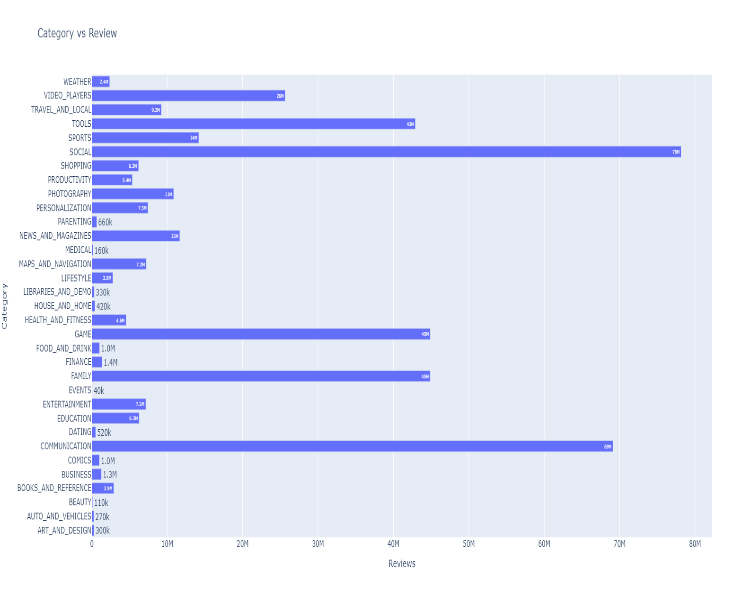
In this graph, we distribute number apps present in category wise and make a pie chart of it. Now we figure out easily, the ‘Family’ has highest number of apps in play store and the number is 1971.

The category which has second highest number of apps present in play store is ‘Game’ and number is 1144.

The category which has lowest number of apps present in play store is ‘Beauty’ and the number is 53 apps.

1. Observation – 5

In this observation, we take Category and Reviews columns and do visualization on it.



In this graph, we take Category on x-axis and Reviews on y-axis.

Now we observed that, which category has highest and lowest reviews.

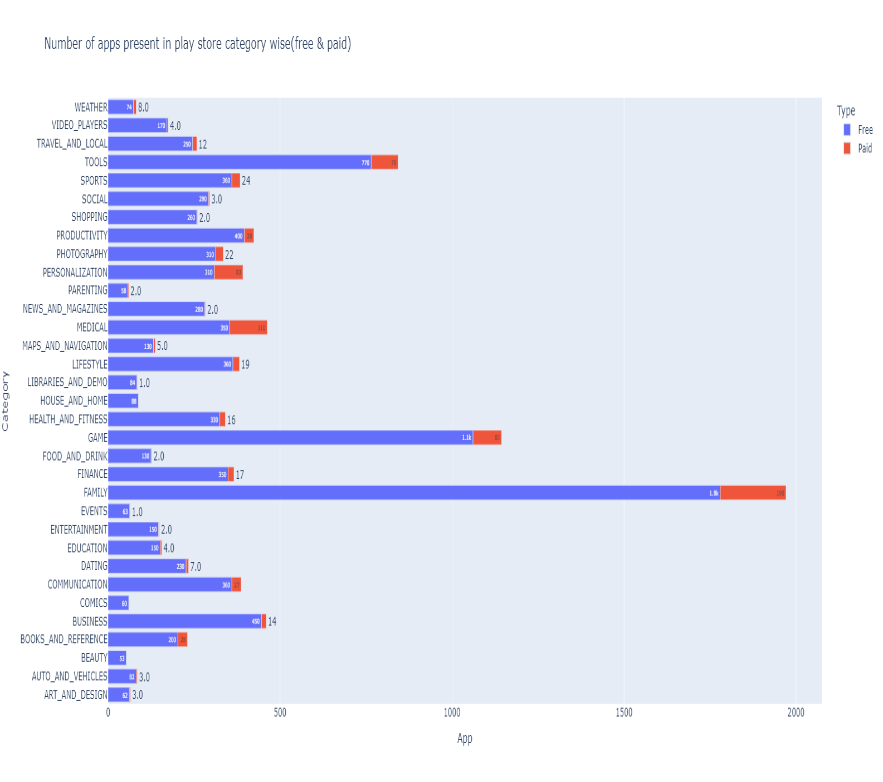
The highest review of category is ‘Social’ and number of reviews is 78 million

The second highest review of category is ‘Communication’ and number of reviews is 69 million

The lowest review of category is ‘Events’ and number of reviews is 40 thousand.

1. Observation – 6

In this observation, we take Category, App and Type columns and do visualization on it.



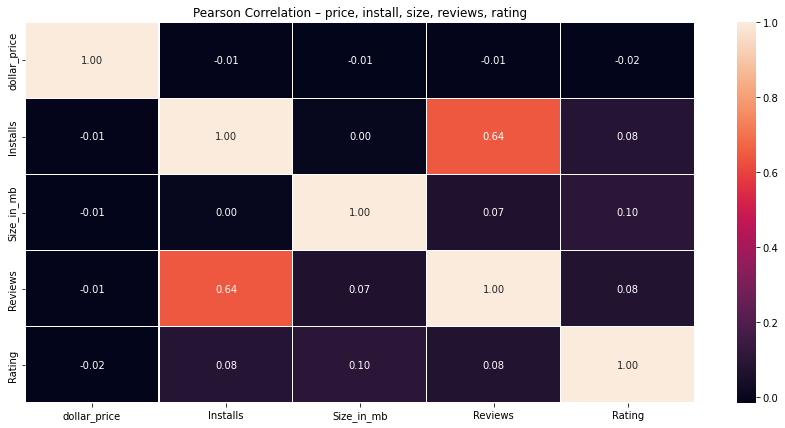
In this plot, category wise we observe number of free and paid apps present in play store.

The Family category takes the lead in both the section (free and paid).

In comics and beauty category, there is no any paid apps.

1. Observation – 7

In this observation, we take dollar price, Installs, Size\_in\_mb, Reviews and Rating columns and do visualization on it.

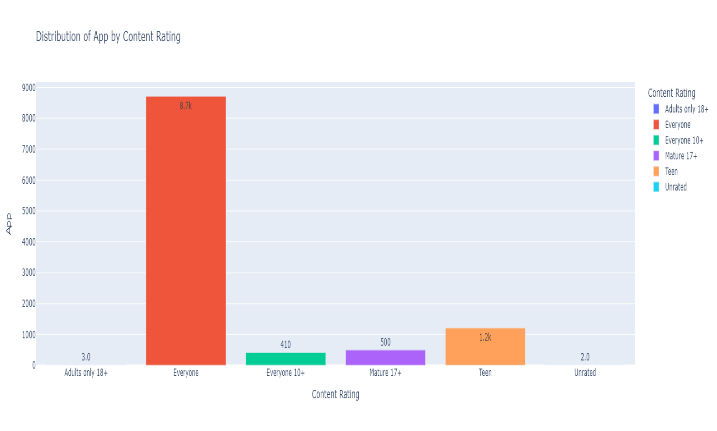


In this Pearson correlation plot, we easily observe the relation of price, installs, size, reviews and rating.

The correlation between Reviews and Installs is 0.64 and it is very strong correlated with each other.

1. Observation – 8

In this observation, we take Content rating and App columns and do visualization on it.

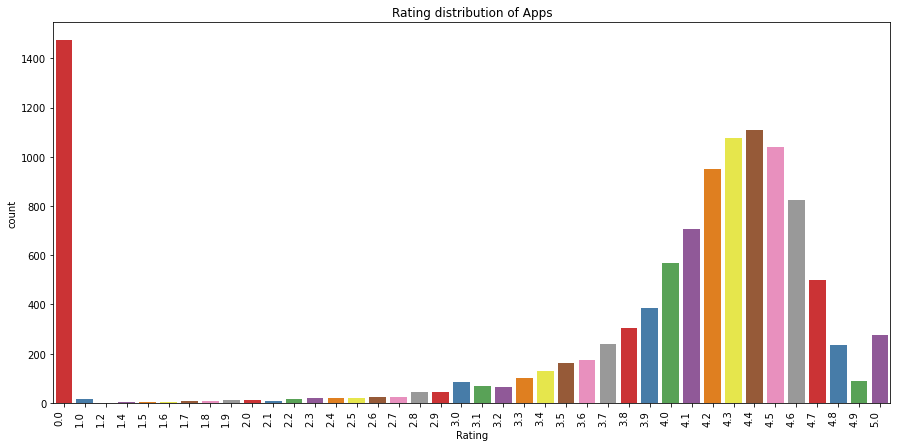


In this plot we observe, distribute of application by content rating.

We see ‘Everyone’ content rating has highest number of apps, which is 8.7 thousand and the lowest apps is present in ‘Adult only 18+’ and ‘unrated’.

1. Observation – 9

In this observation, we take Size, reviews and Installs columns and do visualization on it.

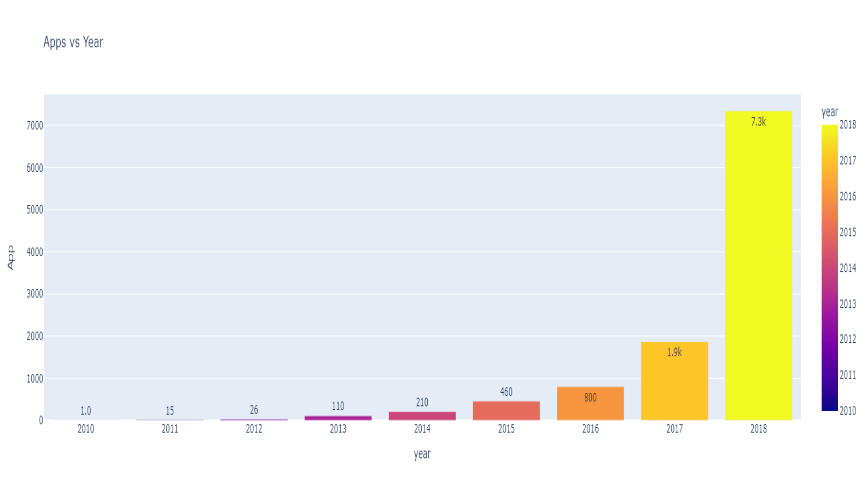


In this plot we observe, distribution of application by rating.

We clearly see, the average rating is between 4.0 to 4.7.

1. Observation – 10

In this observation, we take Year and App columns and do visualization on it.



In this plot we see, distribution of application by year.

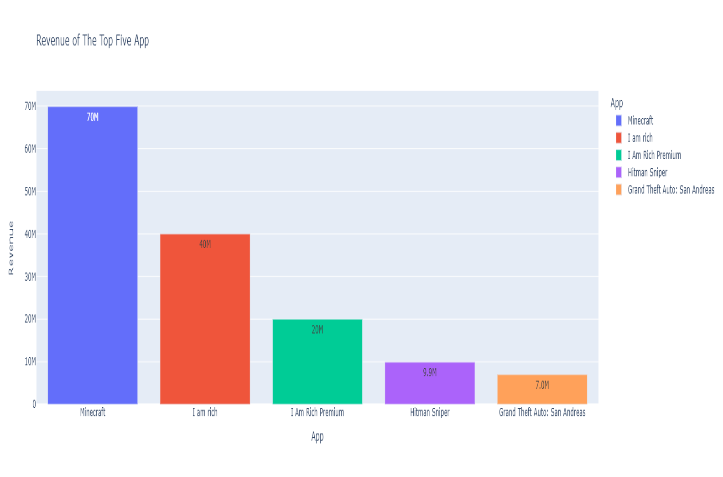
We observe, number of applications is increasing year by year.

Since 2017 to 2018 number of applications is increase by 4 times.

We easily predict the number will be increase in future also.

1. Observation – 11

In this observation, we take Revenue and App columns and do visualization on it.

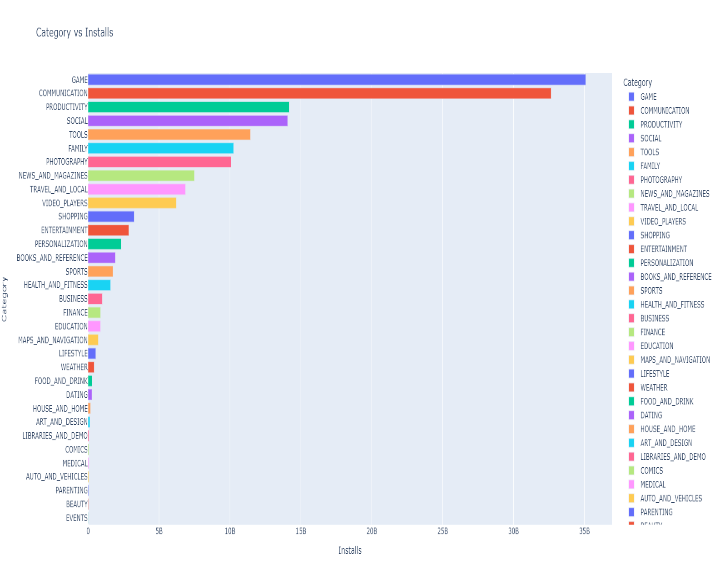


In this plot we see, the top five revenue applications.

The highest revenue application is ‘Minecraft’ and category is ‘Game’

1. Observation – 12

In this observation, we take Category and Installs columns and do visualization on it.



In this plot we see, distribution of category with respect to install.

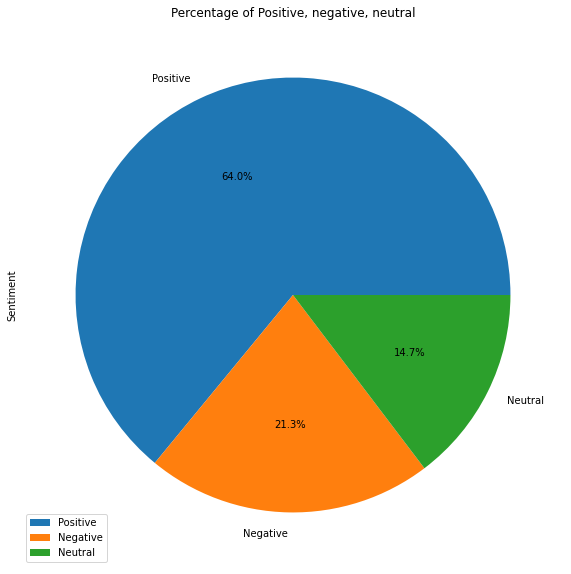
The ‘Game’ category has highest number of installs, which has approximate value is 35 billion installs.

The second highest category is ‘Communication’ which has approximate value is 32.6 billion installs

The lowest installs category is ‘Event’, which has approximate value is 16 million installs.

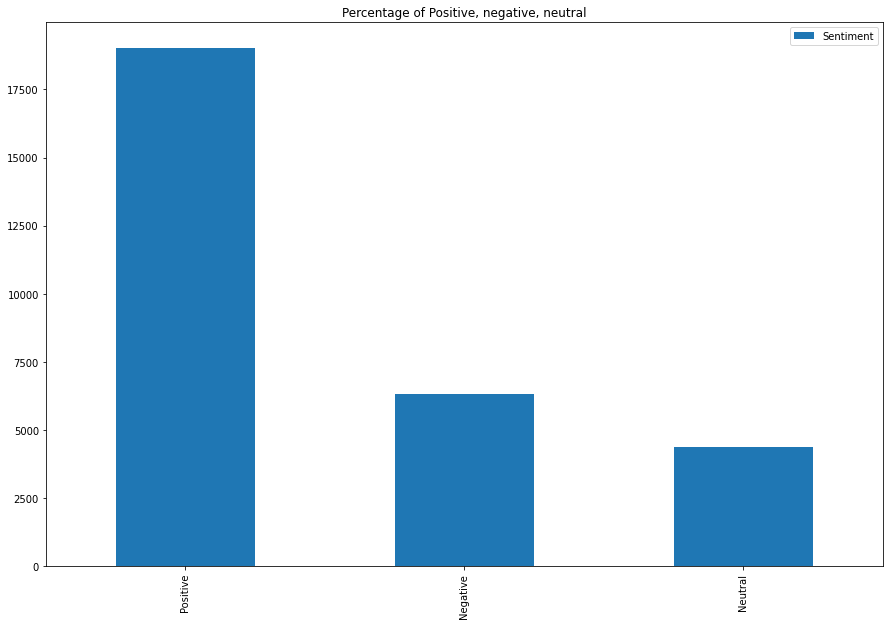
1. Observation – 13

In this observation, we take Sentiment column and do visualization on it.



In this plot we observe, distribution of application with respect to sentiment.

The positive sentiment is more, which is 64 percent. Negative sentiment is 21.3 percent and the neutral sentiment is 14.7 percent.



This is bar plot for distribution of application with respect to sentiment.

1. Observation – 14

In this observation, we take Sentiment, Sentiment\_polarity and Sentiment\_subjectivity columns and do visualization on it.

****

In this plot we see, user sentiment polarity and user sentiment subjevtivity toward an application.

By analyse this plot we observe three cases:

Case – 1 :

When sentiment is positive, the value of setiment polarity is positive and the value of snetimet subjectivity also positive.

Case – 2 :

When sentiment is negative, the value of setiment polarity is negative and the value of snetimet subjectivity positive.

Case – 3 :

When sentiment is neutral, the value of setiment polarity is zero and the value of snetimet subjectivity positive.

**6. Conclusion:**

After the completion of project. I have learned and got exposure to different tools and techniques in data analysis. I was able to complete the project successfully with the help of tools like Python, Pandas, Matplotlib, NumPy, Seaborn, and Plotly. Also, I learned different techniques like Data Cleaning, Data Preparation, Data Exploration and visualization, and Data Interpretation.

**References:**

1. Stackoverflow
2. GeeksforGeeks
3. w3school