**Capstone Project-1 Submission**

**Project Title**

Play Store App Review Analysis



Project Type - EDA/Regression/Classification/Unsupervised

**Presented By : Bejjarapu Bharathwaj**

**Data science trainees,**

**AlmaBetter, Bangalore.**

Email id: [bharathbejjarapu123@gmail.com](mailto:bharathbejjarapu123@gmail.com)

Git Hub link: https://github.com/bharath977/Capstone-project--play-store-app-review-analysis

**Abstract:** The Google Play Store receives thousands of new applications from independent or group developers all over the world. The revenue model for most Play Store apps is unclear due to the prevalence of free apps with in-app purchases, advertisements, and subscriptions. Consequently, the success of an app is often determined by the number of installations and user ratings rather than generated revenue. a digital distribution service developed and operated by google, It is an official apps store that provides variety content such as apps, books, magazines, music, movies and television programs. It serves an as platform to allow users with 'Google certified' Android operating system devices to donwload applications developed and published on the platform either with a charge or free of cost. With the rapidly growth of Android devices and apps, it would be interesting to perform data analysis on the data to obtain valuable insights. However, app ratings can be biased due to inadequate or missing votes, and there are often discrepancies between numeric ratings and user reviews. This study aims to use python to predict the ratings of Google Play Store apps by analyzing and exploring the app dataset collected from alma better data set link. The study examines various attributes such as the app's pricing model (free or paid), user reviews, and app ratings to discover the correlation relationships between them.

Key words: Google Play Store Apps, Google certified, Installs Prediction, Exploratory Data Analysis, Sentimental analysis and correlation.

**1.Introduction:**

Google Play Store is a digital distribution service developed and operated by google, It is an official apps store that provides variety content such as apps, books, magazines, music, movies and television programs. It serves an as platform to allow users with 'Google certified' Android operating system devices to download applications developed and published on the platform either with a charge or free of cost. With the rapidly growth of Android devices and apps, it would be interesting to perform data analysis on the data to obtain valuable insights.

In this project we examine the different attributes present in the data set that affect the popularity of the application. We focused on to answer the questions like, what makes an app popular, what should be the price and size of the app. In our data set we have two csv files for data analysis: Play Store data User Reviews At first, we analysis the play store data and in the play store data we have 10841 rows and 13 columns & in the user review data we have 64295 rows and 5 columns of data. We have to take the maximum outcomes from the data which help us to analysis the which type of app is most preferable and comparisons between different insights. Our goal is to filter and make plots accordingly for a better EDA with respect to the final data.

The dataset with 10k Play Store applications is available to analyze the market of android. It can be examined to analysis the different Category App, Rating, Reviews, Size, Installs, Type, Price, Content, Rating, Genres, Last Updated, Current Ver, Android Ver.

**2.Problem Statement:**

Android is expanding as an operating system. It has captured around 74% of the total market which is a true indicator of the huge amount of population using android. Our goal is to help android developers to know what is the motivating factor for people to download an app. It will also help to find out the factors that affect someone’s decision to download an app. I would like to analyse category, reviews, price, ratings and installs for this purpose and find out how they are inter related.

1.Which app Rating has highest count?

2.How does the count of apps varies by Genres?

3.Which content Rating has highest count?

4.sentiment analysis of customer satisfaction?

5.which Category has highest installations?

6.Does the size of an app affect its rating?

7.Distribution of free app and paid apps?

8.Find the correlaction of the data set?

9.pairplot of the data set find visualize with one variable?

**3.Google Play Store Data Set:**

The dataset consists of Google play store application and is taken from Almabetter, which is the world’s largest community for data scientists to explore, analyze and share data.

This dataset is for Web scratched information of 10k Play Store applications to analyze the market of android. Here it is a downloaded dataset which a user can use to examine the Android market of different use of classifications music, camera etc. With the assistance of this, client can predict see whether any given application will get lower or higher rating level. This dataset can be moreover used for future references for the proposal of any application. Additionally, the disconnected dataset is picked so as to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset, I will examine various qualities like rating, free or paid and so forth utilizing Hive and after that I will likewise do forecast of various traits like client surveys, rating etc.

### The data set contains the following columns:

* **App:** This Column contains the name of the app
* **Category:** This contains the category to which the app belongs. The category column contains 33 unique values.
* **Rating:** This column contains the average value of the individual rating the app has received on the play store. Individual rating values can vary between 0 to 5.
* **Reviews:** This column contains the number of people that have given their feedback for the app.
* **Size:** This column contains the size of the app i.e. The memory space that the app occupies on the device after installation.
* **Installs:** This column indicates the number of time that the app has been downloaded from the play store, these are approximate values and not absolute values.
* **Type:** This column contains only two values- free and paid. They indicate whether the user must pay money to install the app on their device or not.
* **Price:** For paid apps this column contains the price of the app, for free apps it contains the value 0.
* **Content Rating:** It indicates the targeted audience of the app and their age group.
* **Genre:** This column contains to which genre the app belongs to, genre can be considered as a sub division of Category.
* **Last updated:** This column contains the info about the date on which the last update for the app was launched.
* **Current version:** Contains information about the current version of the app available on the play store.
* **Android version:** Contains information about the version of the android OS on which the app can be installed.

**4.User Review Dataset:**

* User reviews data frame has 64295 rows and 5 columns. The 5 columns are identified as follows:
* **App:** Contains the name of the app with a short description (optional).
* **Translated Review:** It contains the English translation of the review dropped by the user of the app.
* **Sentiment:** It gives the attitude/emotion of the writer. It can be ‘Positive’, ‘Negative’, or ‘Neutral’.
* **Sentiment Polarity:**It gives the polarity of the review. Its range is [-1,1], where 1 means ‘Positive statement’ and -1 means a ‘Negative statement’.
* **Sentiment Subjectivity:** This value gives how close a reviewer’s opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer’s opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

**5.Python**

Most of the info scientist use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is simplest programing language to select up compared to other language. That is the most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is straightforward to use. That is one among the most reasons to use python. Specifically, for data scientist the foremost popular data inbuilt open-source library is named panda. As we have seen earlier in our previous assignment once we got to plot scatterplot, heat maps, graphs, 3-dimensional data python built-in library comes very helpful.

**6.Data Cleaning and Preparation**

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**: We write a function play store head(), tail(),info (), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values, number of unique values in that column and percentage of null value in that columns in the play store dataset.
* **Step2**: finding out the duplicated values and made a graph
* **Step 3**: missing values for android, current and content rating columns is less (<0.05) we are using mode to replace the value 0
* **Step 4**: We can see that the ‘Rating’ column has 1474 null values. Due to low variations in the rating values As data comes in many shapes and forms, pandas aims to be flexible with regard to handling missing data. While NaN is the default missing value marker for reasons of computational speed and convenience, we need to be able to easily detect this value with data of different types: floating point, integer, boolean, and general object. In many cases, however, the Python None will arise and we wish to also consider that “missing” or “not available” or “NA”.
* **Step 5:** We can see that the ‘Reviews’ column despite being a numerical indicator is of the ‘object’ data type, we will convert this to ‘int’ data type using the as type(int) function.
* **Step 6:**  We can see that the size column, here we are applied replace function 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 1000 and ‘M’ with 1000000. Some values also have ‘+’ sign in them, which will be removed. Next, we will convert this column into ‘int’ datatype.
* **Step 7:** Here I have applied lambda functionThe ‘Installs’ column values contain the characters ‘+’ and ‘,’ which are going to prevent us from converting this column into a numeric datatype. We will get rid of these using the lambda() and replace() functions.
* **Step 8:** The values in the column ‘Price’ might have the ‘$’ sign in some values and the column is of the datatype ‘object’. We will first remove the ‘$’ sign using the lambda() function and then convert the column into ‘int’ datatype.
* **Step 9:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.
* **Step 10:** We write a function info(), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values ,number of unique values in that column and percentage of null value in that columns in the User review dataset.
* **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.

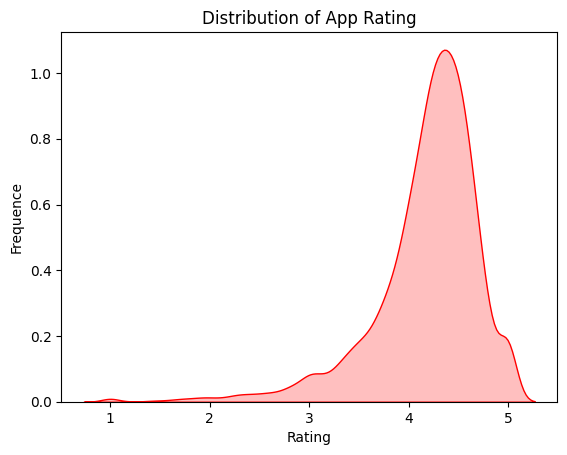
**7.EXPLORATORY DATA ANALYSIS**

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.

Fig:1

1.Which app Rating has highest count?

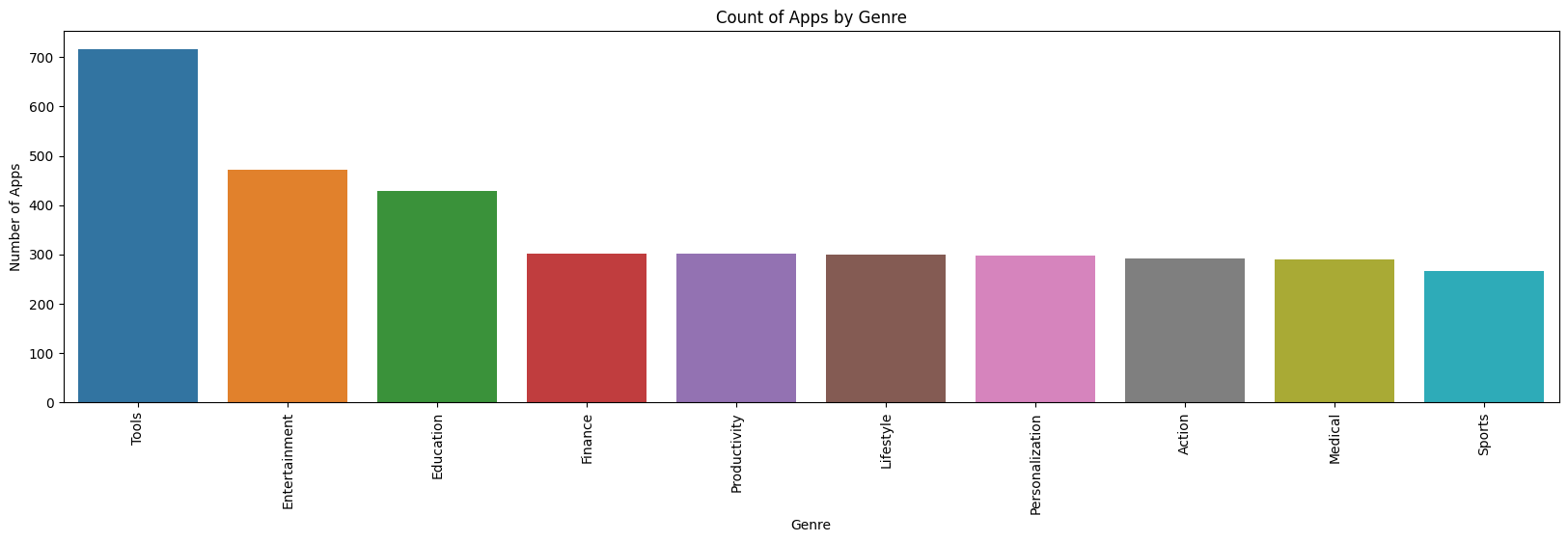


I have choosed the kdeplot its easy understandable by observing the graph the rating of the app distribution to plot the data against a single/univariate variable. It represents the probability distribution of the data values as the area under the plotted curve.

It is clear from the visualizations that the data in the Rating column is skewed towards the right.

Fig 2:

2.How does the count of apps varies by Genres?

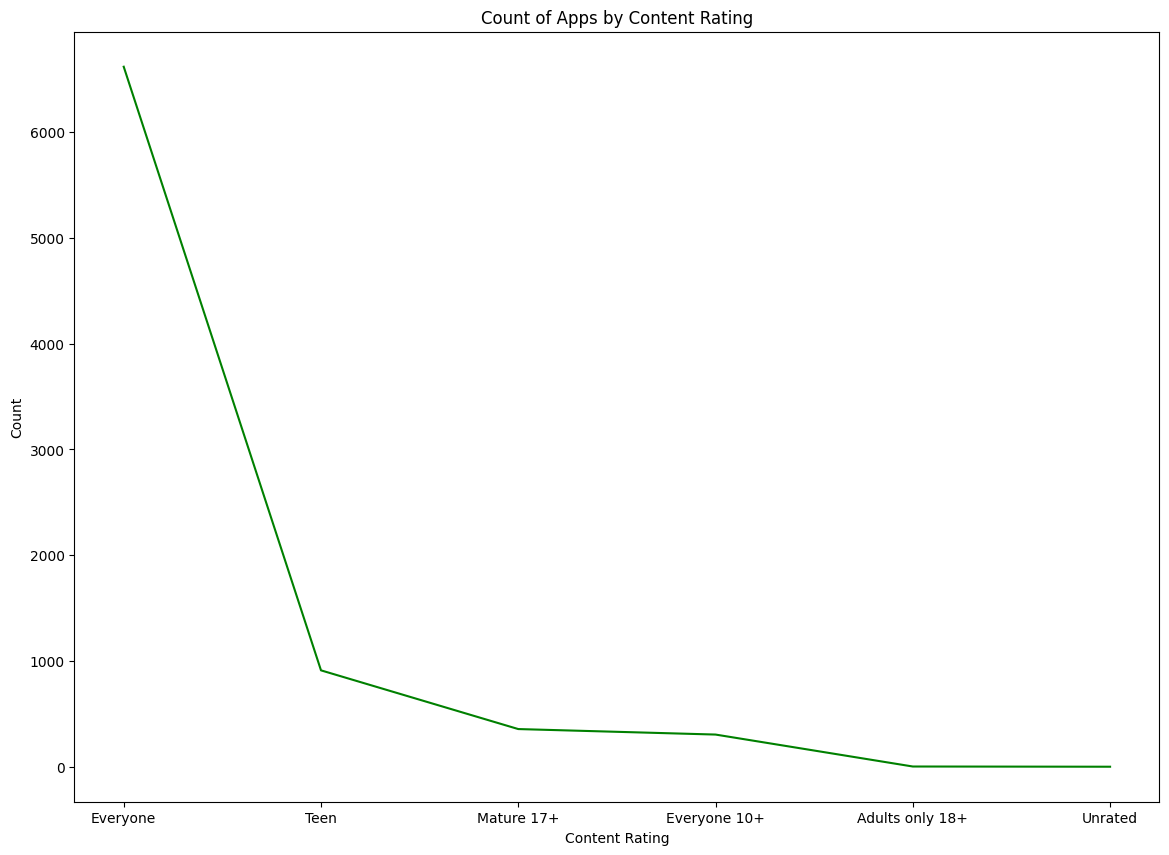


I have choose Bar graph the head(10) genres and counting the value to understand this selected bagplot because we are comparing the two items in x and y axis to know count of apps varies by Genres this chart is used to compare the data the insights.

The top genre is in number of apps are in Tools and Followed by Entertainment.

Fig 3:

3.Which content Rating has highest count?

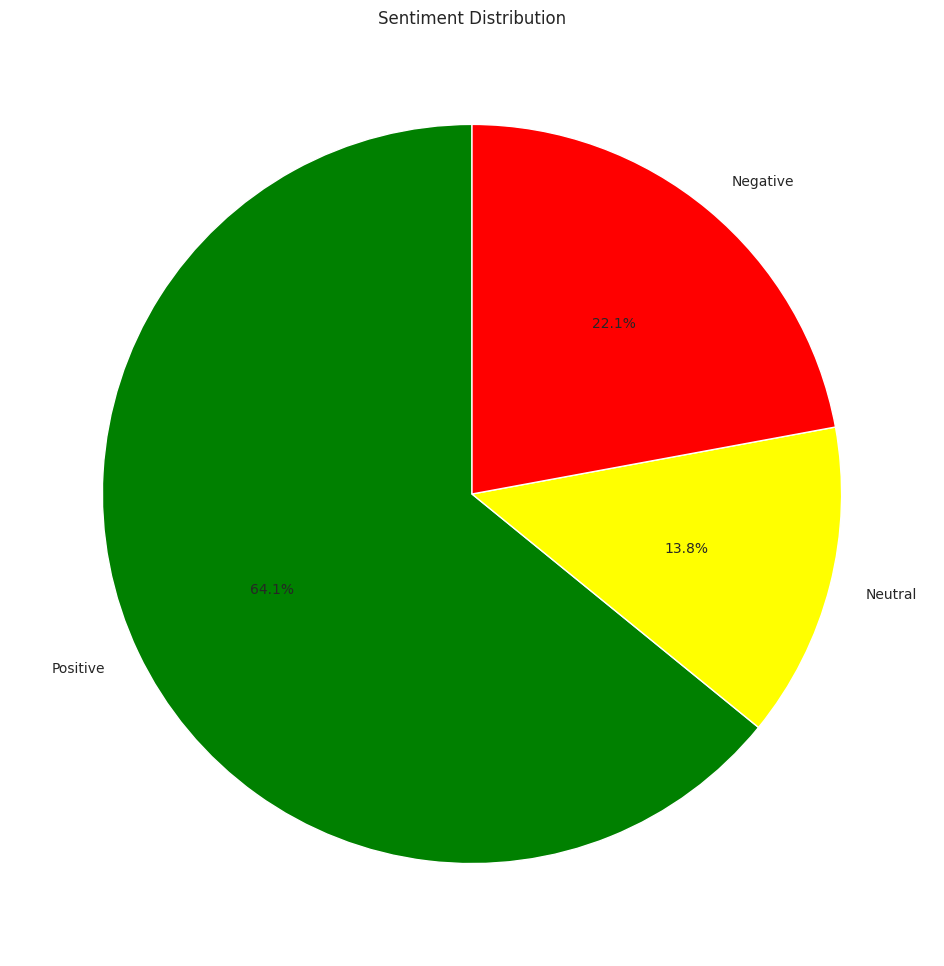


line chart use to understand easily by observing the line either it goes up or falling down this help to understand reasons. The relationship between x and y can be shown.

With good content rating count values will increase as app access to everyone there is no restriction as seen in the graph.

Fig 4:

4.sentiment analysis of customer satisfaction?



A pie chart is a common way to visualize the distribution of categorical data. In this case, we are visualizing the distribution of sentiment categories (positive, neutral, negative) in a set of reviews.

Pie charts are useful because they allow us to see the proportion of each category in the data at a glance. The size of each slice in the pie corresponds to the proportion of the data that falls into that category, making it easy to compare the relative sizes of the categories.

1.The majority of the reviews fall into the positive sentiment category, as indicated by the largest slice of the pie. This means that most customers had a positive experience with whatever product or service is being reviewed.

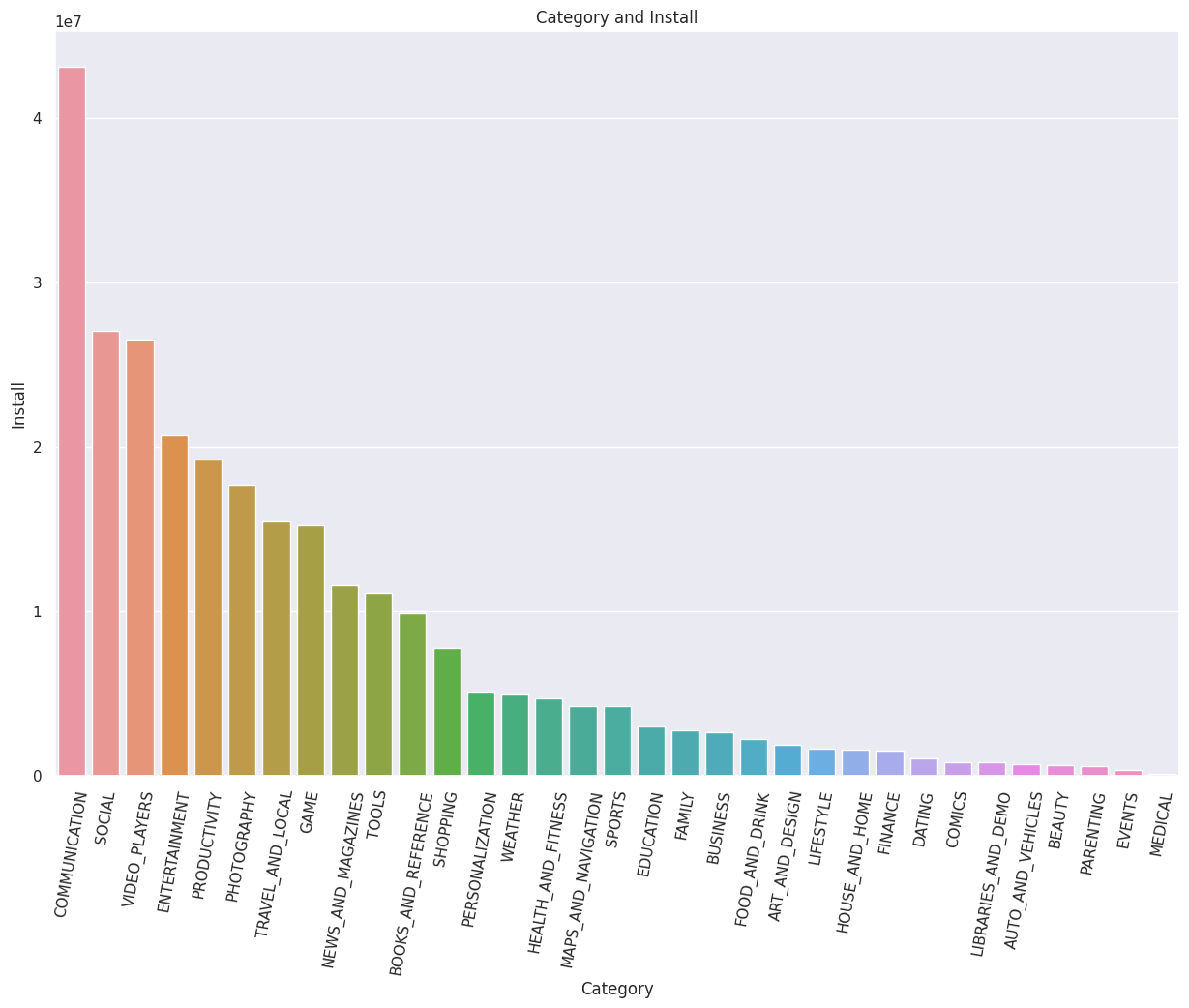
2.A smaller proportion of the reviews fall into the neutral category, as indicated by the middle slice of the pie. This means that some customers had a neutral experience, neither positive nor negative.

3.The smallest proportion of the reviews fall into the negative sentiment category, as indicated by the smallest slice of the pie. This means that relatively few customers had a negative experience with the product or service being reviewed.

Fig:4

Bivariate Analysis:

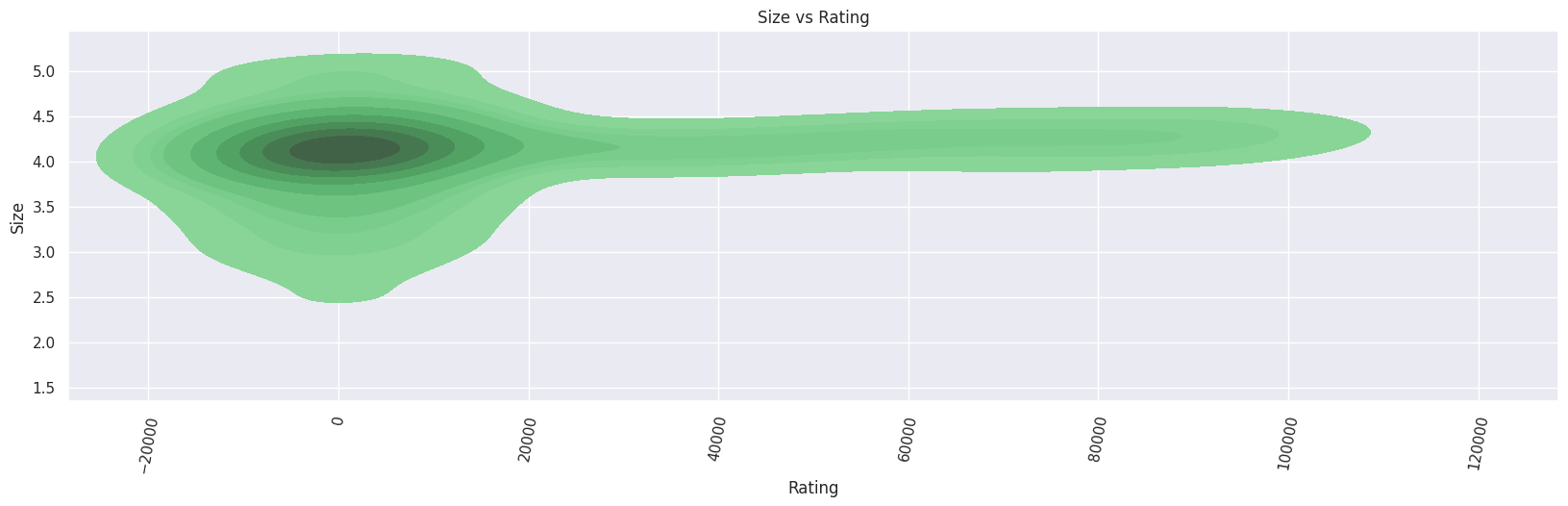
5.which Category has highest installations?



I used the barplot because we are comparing the two items in x:category and y:installs axis to know which category has highest installs this chart is used to compare the data the insights we find that the communication and followed by social category

Fig:6

6.Does the size of an app affect its rating?

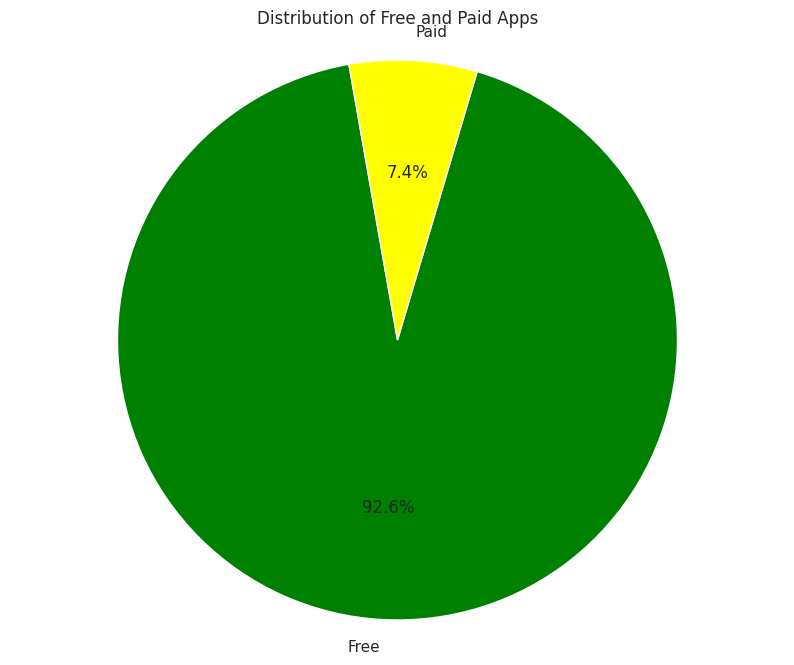


kdeplot is a good choice for visualizing large datasets in Seaborn. It is a type of density plot that estimates the probability density function of a continuous variable. KDE plots are useful for visualizing the distribution of a single variable, as well as comparing the distributions of multiple variables.

There seems to be a positive correlation between the size of the product and the average rating. Generally, this is less impact that size is relating to Rating.

Fig:7

7.Distribution of free app and paid apps?

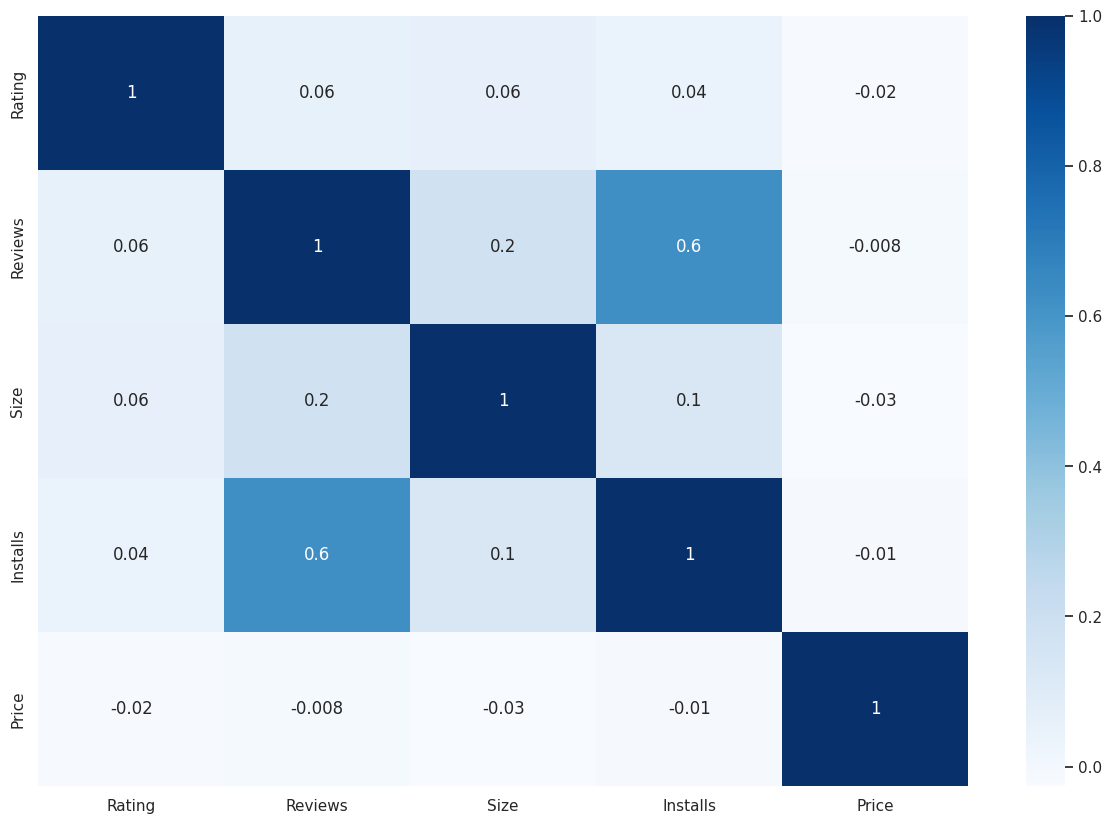


This chart because it is a simple and effective way to show the proportion of two categories. In this case, we want to compare the number of free apps to the number of paid apps in the Google Play Store. A pie chart allows us to quickly see the relative size of each category and compare them visually.

The insight we found that the distribution between the two free app and paid app the most number of customers are choosing the free app rather than paid once.by this understanding business should provide free installation of app.

Fig:8

8.Find the correlaction of the data set?



To visualise the correlations between all variables, not just the target variable, you can create a correlation matrix. This is essentially the same as the dataframe above, but with a row for each variable, and a neat colour coding scheme that allows you to see which values are most positively or negatively correlated based on the depth of their colour.

We then generate the correlation matrix using the .corr() method on the DataFrame df\_store, which calculates the Pearson correlation coefficient between each pair of columns in the DataFrame. The resulting matrix is assigned to the variable corr\_matrix.

Finally, we generate a heatmap of the correlation matrix using sns.heatmap(). We set annot=True to display the correlation coefficient values on the heatmap, fmt='.1g' to format the displayed values to one decimal place.

Positive correlation: A cell that is shaded in blue indicates a positive correlation between two variables. This means that the values of both variables tend to increase or decrease together.

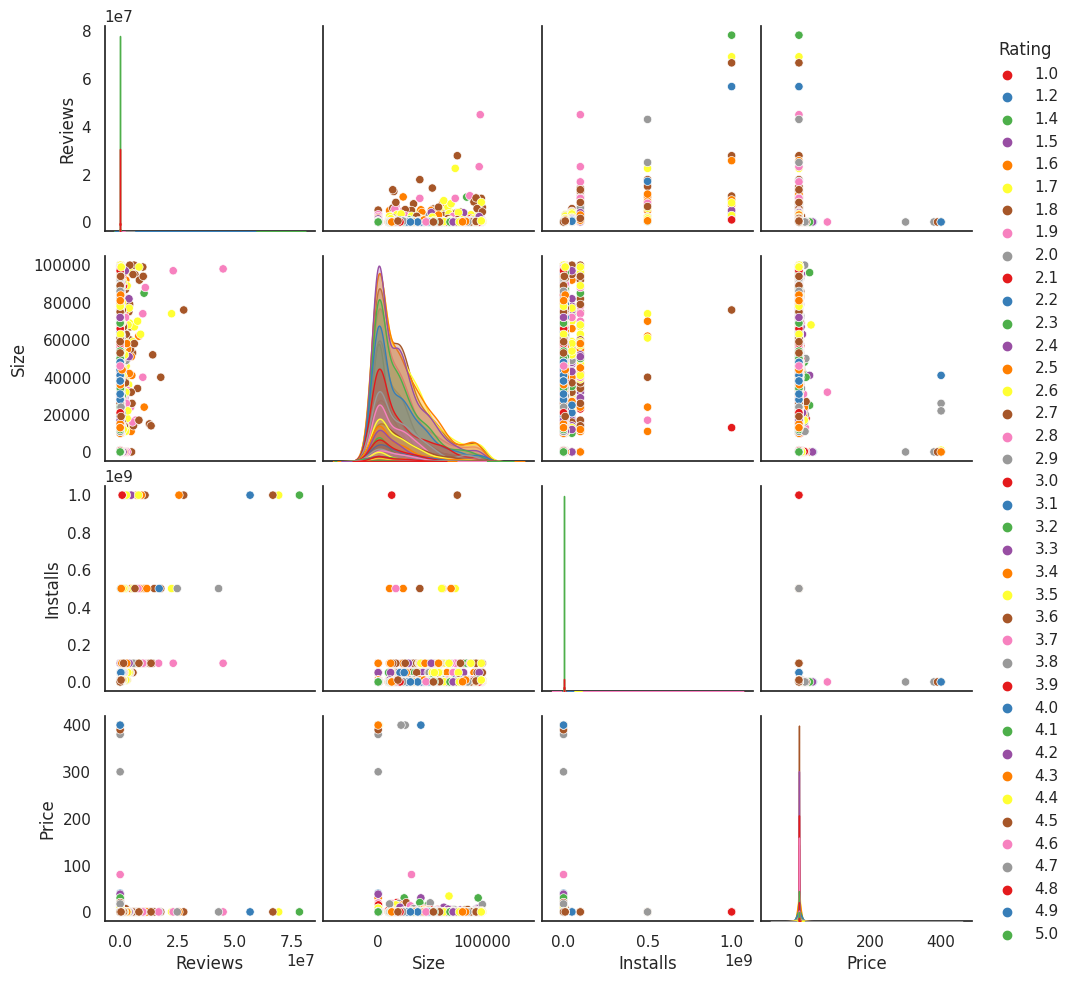
Negative correlation: A cell that is shaded in red indicates a negative correlation between two variables. This means that as one variable increases, the other variable tends to decrease.

Strength of correlation: The intensity of the color in each cell represents the strength of the correlation between the pair of variables. Darker shades of blue or red indicate a stronger correlation.

Overall, correlation maps can provide valuable insights into the relationships between variables in a dataset, which can inform decision-making and guide further analysis.

Fig:9

9.pairplot of the data set find visualize with one variable?



sns.pairplot shows the relationship between variables. By default, these plots are scatter plots, but you can change the plot type using the kind parameter. The diagonal plots show the distribution of each variable. By default, these plots are kernel density estimates (KDE), but you can change the plot type using the diag\_kind parameter.

The sns.pairplot function also supports color-coding the plots based on a categorical variable using the hue parameter. This can help to highlight differences between groups in the data.

Most of the customers have rated the store with 4 or 5 stars, as we can see a higher density of datapoints with these ratings.

The variables Price and Size appear to have a positive correlation, which means that larger products tend to be more expensive.

Overall, this sns.pairplot provides a useful visual summary of the pairwise relationships between variables in the df\_store dataset. The color-coding based on the Rating variable allows us to see how different ratings are distributed across the various scatter plots, and to identify potential correlations between the variables.

**The solution to the business objective**: of Play Store review analysis would depend on the specific objective of the company. Here are some general solutions for the common business objectives:

Focus Content Rating: Business should foucs on content rating of the apps because customer can understand well through the content this can bring impact on the business by increasing the customer rate

Improving customer satisfaction: To improve customer satisfaction, companies could use sentiment analysis to categorize reviews as positive, negative, or neutral. They could then identify common themes and issues in negative reviews and take steps to address them. Companies could also respond to negative reviews publicly and offer solutions to address customer concerns.

Brand reputation management: To manage brand reputation, companies could monitor Play Store reviews and respond to negative reviews promptly. They could also identify and address common issues raised in negative reviews to improve customer satisfaction.

Improving app ratings: To improve app ratings, companies could use sentiment analysis to categorize reviews as positive, negative, or neutral. They could then identify the factors that influence app ratings and take steps to improve them. For example, if users frequently complain about app crashes, the company could prioritize bug fixes to improve app stability.

with the positive correlation the business can improve highly by customers satisfaction with the negative correlation business should bring new implementation or make changes on existing plans.

**Conclusion**

Through our exploration and analysis of data, we discovered certain trends and assumptions that suggest our app could be successful among users in the Google Play Store. We believe these insights could lead to the triumph of our app among discerning users.

1.The app rating has highest count majorly stands around 4 to 5 ,From the distplot visualizations, it is clear that the ratings are left skewed.

2.The top 10 Genres are

Tools 717, Entertainment 471, Education 429, Finance 302, Productivity 301, Lifestyle 300, Personalization 298, Action 292, Medical 290, Sports 266.

3.A majority of the apps in the play store are can be used by everyone. The remaining apps have various age restrictions to use it.

4.(64.1) percent are postive user review sentiment analysis

5.communication and followed by social category are highest no.of. installed category

6.The size of a app is less impacted by rating with(0.06)of correlaction

7.Here we can see that 92.2% apps are free, and 7.80% apps are paid on Google Play Store, so we can say that Most of the apps are free on Google Play Store.

8.By observing heatmap of correlation There is a strong positive correlation between the Reviews and Installs column.

The Price is slightly negatively correlated with the Rating, Reviews, and Installs. This means that as the prices of the app increases, the average rating, total number of reviews and installs fall slightly.

The Rating is slightly positively correlated with the Installs and Reviews column. This indicates that as the average user rating increases, the app installs, and number of reviews also increase.

9.Most of the App are Free. Most of the Paid Apps have Rating around 4 As the number of installations increases the number of reviews of the particular app also increases. Most of the Apps are light-weighted.