Capstone-Project--Play-Store-App-Review-Analysis I create this EDA based capstone project for take a deep look into google play store app review for collecting,cleaning and organise its data to use for a suitable purpose.

Play Store App Review Analysis

**📋 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 catergory, 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. 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. The objective of this experiment is to deliver insights to understand customer demands better and thus help developers to popularize the product. We have tried to discover the relationships among various attributes such as which application is free or paid, what are the user reviews, rating of the application.

-----------------------------------------------------

**💾 Project Files Description** This Project includes 1 individual colab notebook,1 team notebook, 2 datasets csv files along with technical documentation:

Executable Files: Play Store App Review Analysis - Includes all functions required for clustering operations. Output: Google Colab - All the outputs are visible in the provided colab notebook. Input Files: Play Store Data.csv - It contains the basic details of the app like number of user reviews, ratings, etc. User Reviews.csv - It contains the user reviews and its sentiment score for the respective app. Data Source:

**Dataset taken from Almabetter**

**📖 Introduction:** In today’s scenario we can see that mobile apps playing an important role in any individual’s life. With enormous challenge from everywhere throughout the globe, it is important for a designer to realize that he/she is continuing in the right way or not. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The dataset with 10k Play Store applications is available to analyze the market of android. It can be examined to analysis the different category such as family, communication,entertainment, tools, music, camera etc. 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, is there some trends in user sentiments. 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. We need to explore and analyze the data to discover key factors responsible for app engagement and success.

The contents of Play Store Data are: App: It contains the name of the app with a short description (optional). Category: This section gives the category to which an app belongs. In this dataset, the apps are divided among 33 categories. Size: The disk space required to install the respective app. Rating: The average rating given by the users for the respective app. It can be in between 1 and 5. Reviews: The number of users that have dropped a review for the respective app. Installs: The approximate number of times the respective app was installed. Type: It states whether an app is free to use or paid. Price: It gives the price payable to install the app. For free type apps, the price is zero. Content rating: It states which age group is suitable to consume the content of the respective app. Genres: It gives the genre(s) to which the respective app belongs. Last updated: It gives the day in which the latest update for the respective app was released. Current Ver: It gives the current version of the respective app. Android Ver: It gives the android version of the respective app. The contents of User Reviews are: App: It 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.**

**📋Problem Statements**

* Which category has highest share of app in market?
* What is distribution of Apps rating?
* Which content Rating has highest count?
* Which Genres have most number of apps?
* What are the sentiments of people towards apps? Is there any Correlation between App Size and ratings?
* How is price trending across some categories?
* Is subjectivity and polarity proportional to each other?
* What are top 10 downloaded apps based on reviews and what's difference between their rating & reviews?
* Is there any corelation between Rating,Size,Installs,Reviews and Price?
* Is there any corelation between all quantitative value in both datasets?

**📔 What is Exploratory Data Analysis?**

* Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets for patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset and summarize their main characteristics, often employing data visualization methods. It is an important step in any Data Analysis or Data Science project. It helps determine how best to manipulate data sources to get the answers you need.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better and make it more attractive and appealing.

The following are the various steps involved in the EDA process:

Problem Statement - We shall brainstorm and understand the given data set. We shall study the attributes present in it and try to do a philosophical analysis about their meaning and importance for this problem. Hypothesis - Upon studying the attributes present in the data base, we shall develop some basic hypothesis on which we can work and play with the data to look for the varied results which we can get out of it. Univariate Analysis - It is the simplest form of analyzing the data. In this we would initially pick up a single attribute and study it in and out. It doesn't deal with any sort of co-relation and it's major purpose is to describe. It takes data, summarizes that data and finds patterns in the data. Bivariate Analysis - This analysis is related to cause and the relationship between the two attributes. We will try to understand the dependency of attributes on each other. Multivariate Analysis - This is done when more than two variables have to be analyzed simultaneously. Data Cleaning - We shall clean the dataset and handle the missing data, outliers and categorical variables.

**Testing Hypothesis - We shall check if our data meets the assumptions required by most of the multivariate techniques.**

📖 Steps Involved After loading the dataset, we can start the exploration but before that, we need to check and see that the dataset is ready for performing several exploration operations or not, so let’s first have a look at the structure and the manner in which the data is organized.

Data Cleaning Our data set contains a large number of null values in the rating column, so we drop them. Some of the columns have a smaller number of null values, so we replace the null values in these columns with the mode value of that particular column. Our data set also contain the duplicate rows for a single application. We also drop the duplicate rows because the rows contain the identical data. Also drop the rows, which have rating greater than 5.

Data Transforming From the information of data frame, we can see that all the columns except rating have the object data type but some of the columns like, reviews, size, installs and price have the numerical value. So, we have to transform them in proper data type and also remove the unwanted values from the numerical columns like ‘+’ and ‘,’ from installs and ‘$’ from price. In the size column we have some values in KB and some values in MB, so we transform all the values in MB.

Exploratory Data Analysis After establishing a good sense of each feature, we proceeded with plotting a pairwise plot between all the quantitative variables to look for any evident patterns or relationships between the features. There is a high variance in the number of installs and in number of reviews. To overcome this problem, we add two new columns to the data frame named: log\_installs and log\_review, which contain the logarithmic values of installs and review columns, respectively.

**📋 Conclusion:**

Our main objective here was to handle all the missing value and duplicates , providing app developers and stakeholders a insight to make their application according to marketing strategies follow according to analysis. Some points as we already discussed that what we have found regarding this dataset so for more generalised things that a company could focus are give here as-

Overall rating distribution: The analysis can provide information on the overall distribution of ratings for the apps in the Play Store. This can help understand the popularity of different rating levels, and also identify the areas where the app developers can improve to attract more users.

Ratings by category: The analysis can also provide insights on the ratings distribution by category, such as entertainment, productivity, education, etc. This can help identify the categories that are most popular among users, and also understand the expectations of users from apps in different categories.

Reviews sentiment analysis: The analysis can also involve sentiment analysis of the app reviews, which can help understand the sentiment of users towards different apps. This can be useful in identifying the areas where the app developers need to improve, as well as understanding the strengths of popular apps.

App size and ratings correlation: Another possible insight is to identify if there is any correlation between app size and ratings. This can help understand if users prefer smaller apps, and whether developers should focus on optimizing the app size to improve ratings.

📚 References

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

📜 Feedback If you have any feedback, please reach out to us at [bharathbejjarapu123@gmail.com](mailto:bharathbejjarapu123@gmail.com)