Google Play store EDA

Project Report

Course Name: Data Science Lab (R4EC3012P)

Date: January-May 2023

Team Members:

Yash Deshpande

Email: yadeshpande_b20@el.vjti.ac.in

Chirag Patil

Email: cupatil_b20@el.vjti.ac.in

Atharva Bendre

Email: aubendre_b20@el.vjti.ac.in

Shreyas Bhatlawande:

Email: sgbhatlawande_b20@el.vjti.ac.in

TABLE OF CONTENTS:

1. PROJECT OVERVIEW

Description of Use Case & Project

The objective of this experiment is to deliver insights to understand customer demands better and thus help application developers to popularize their products. 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,

1. Which category has the greatest number of installations?

2. How many free apps does the Play Store have?

3. Which is the most common category of apps on Play Store?

4. Which is the most expensive category?

5. Which category has the highest number of reviews on Play Store?

What is Exploratory Data Analysis:

Exploratory data analysis (EDA) is used by data scientists to analyse 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.

Types of EDA

1.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 correlation and its major purpose is to describe. It takes data, summarizes that data and finds patterns in the data.

2.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.

3.Multivariate Analysis - This is done when more than two variables have to be analysed simultaneously.

2. INTRODUCTION:

Brief Idea

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. A few thousands of new applications are regularly uploaded on Google play store. A huge number of designers working freely on designing the apps and making them successful. With the enormous challenge from everywhere throughout the globe, it is important for a developer to know whether he/she is continuing the correct way or not. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, in-application 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. We aim to provide these insights to developers by performing EDA on Google Play store dataset.

Basic Project Domains:

- Exploratory Data Analysis
- Data Visualization
- Data Science

3. METHODS & STAGES OF PROGRESS

Theory & Approach:

Steps involved in EDA

1.**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.

2.**Data Collection -** Every business knows the importance of using data beneficially by properly analysing it. However, this depends on collecting the required data from various sources through surveys, social media, and customer reviews, to name a few.

3.**Data Cleaning -** We shall clean the dataset and handle the missing data, outliers and categorical variables.

4.**Identify correlated variables -** Finding a correlation between variables helps to know how a particular variable is related to another.

5.**Visualizing and Analysing Results**-Once the analysis is over, the findings are to be observed cautiously and carefully so that proper interpretation can be made. The trends in the spread of data and correlation between variables give good insights for making suitable changes in the data parameters.

Contents of the Play Store Dataset:

- 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.

Libraries Used:

1. NumPy:

NumPy is a powerful numerical computing library for Python, providing support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays.

2. Seaborn:

Seaborn is a statistical data visualization library based on Matplotlib. It simplifies the creation of informative and attractive statistical graphics through a high-level interface, with built-in themes and color palettes.

3. Pandas:

Pandas is a data manipulation and analysis library for Python, offering data structures like Data Frames and Series for efficient handling and manipulation of structured data, including tools for cleaning, exploring, and transforming datasets.

4. Matplotlib:

Matplotlib is a versatile 2D plotting library for Python. It enables the creation of a wide variety of static, animated, and interactive visualizations, providing fine-grained control over plot elements and customization options.

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.

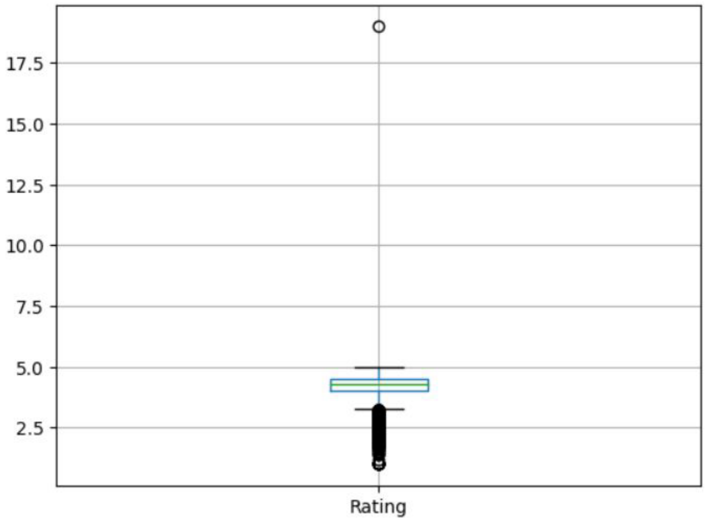
1. **Data Cleaning**

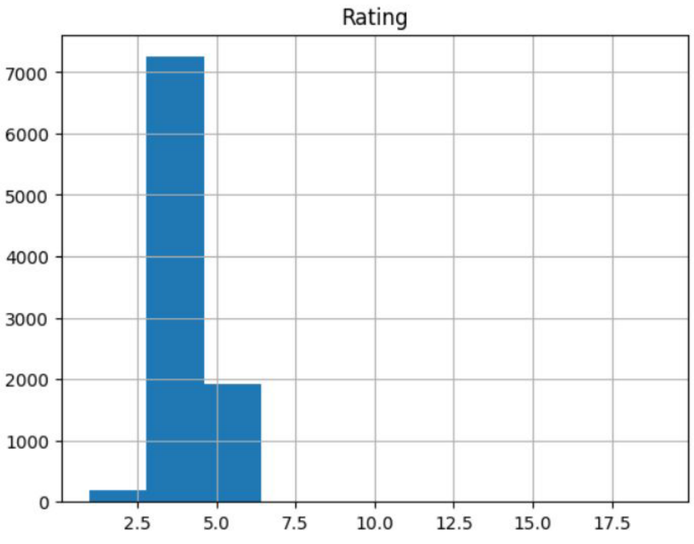
Our data set contains a large number of null values in the rating column, so we drop them. Some of the columns have a smaller number of null values, so we replace the null values in these columns with the mode value of that particular column. Our data set also contains duplicate rows for a single application.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

2. **Checking how many Outliers are there and removing them**

Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different manner. They can be caused by measurement or execution errors. The analysis of outlier data is referred to as outlier analysis or outlier mining.

Box plot and Histogram when outliers present





We find Point Outlier in our dataset by giving the condition of Ratings greater than 5

```
df[df.Rating>5]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	Everyone	NaN	February 11, 2018	1.0.19	4.0 and up

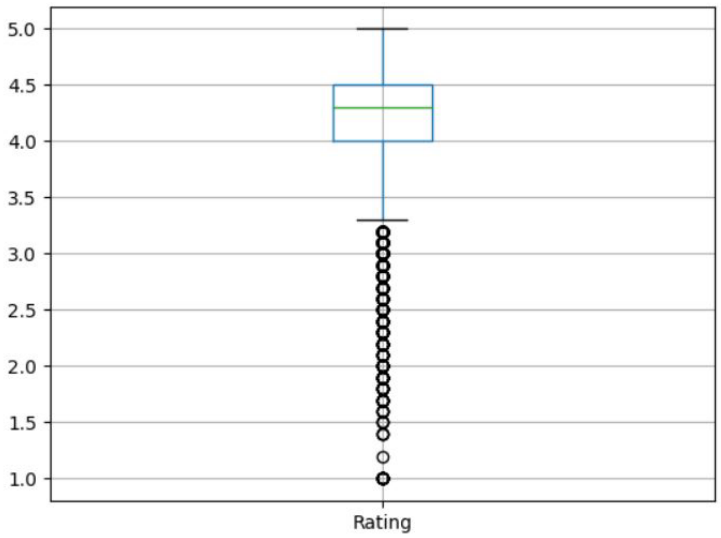
3. Removing Outliers

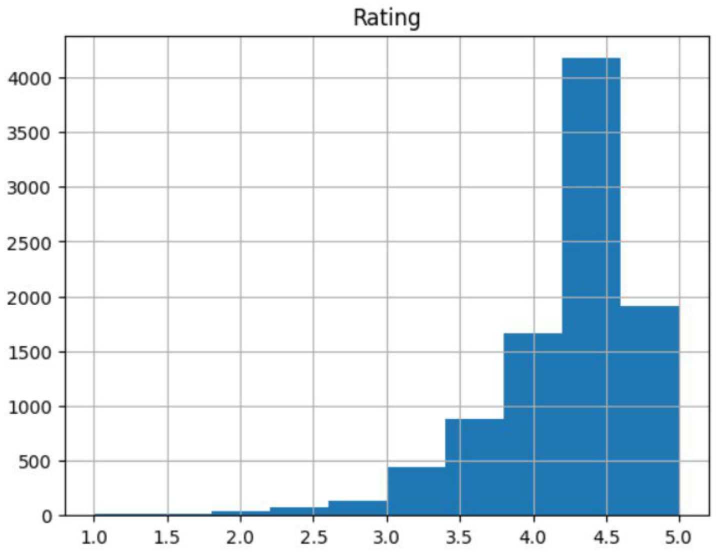
```
df.drop([10472], inplace=True)
```

```
df[10470:10475]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	L Upda
10470	Jazz Wi-Fi	COMMUNICATION	3.4	49	4.0M	10,000+	Free	0	Everyone	Communication	Febru 10, 2
10471	Xposed Wi-Fi-Pwd	PERSONALIZATION	3.5	1042	404k	100,000+	Free	0	Everyone	Personalization	Augus 2
10473	osmino Wi-Fi: free WiFi	TOOLS	4.2	134203	4.1M	10,000,000+	Free	0	Everyone	Tools	Augus 2
10474	Sat-Fi Voice	COMMUNICATION	3.4	37	14M	1,000+	Free	0	Everyone	Communication	Novem 21, 2
	Wi-Fi										May

Box plot and Histogram when outliers removed



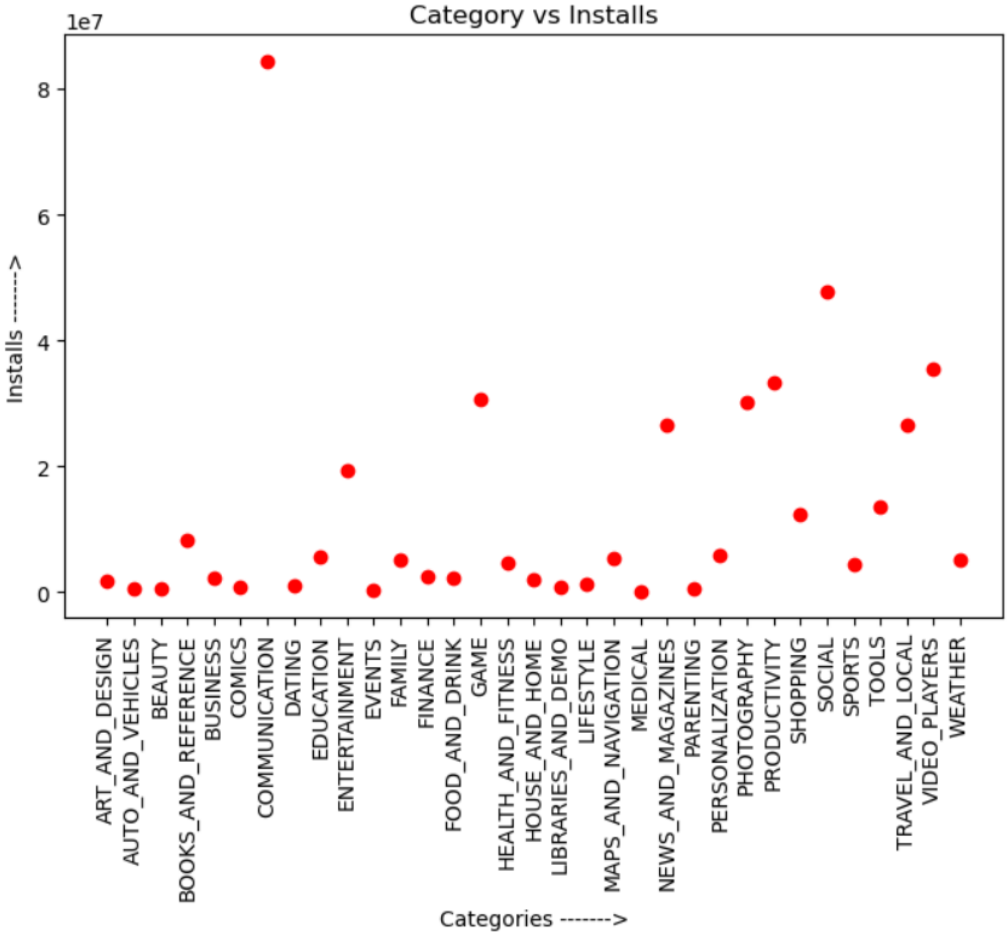


4. Data Visualization

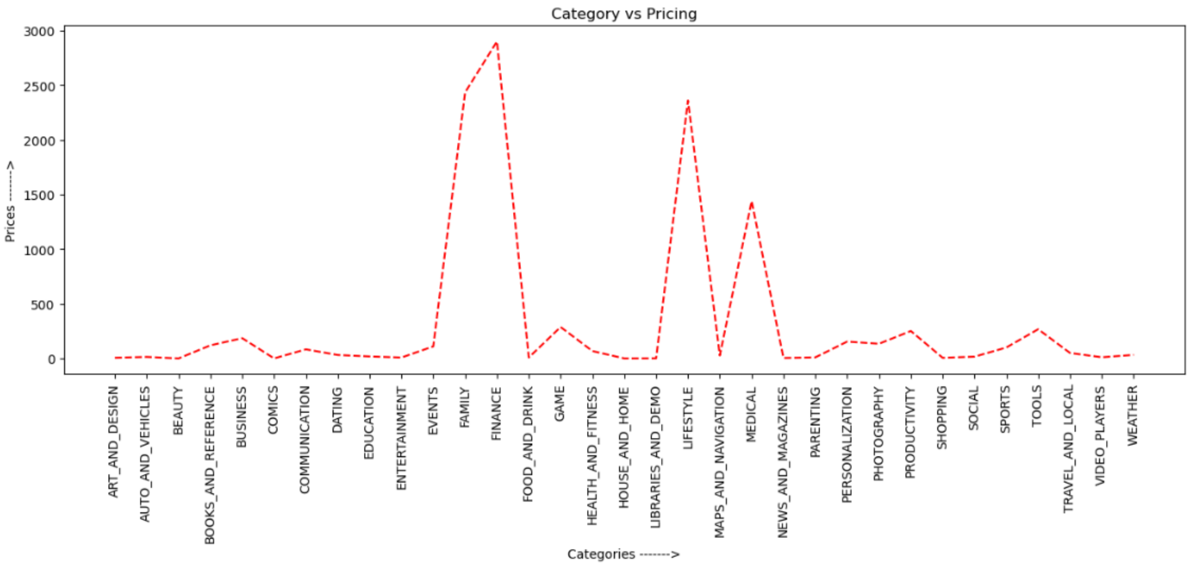
For this project of exploratory data analysis on the Google Play store dataset, we performed data visualization because visual representations play a crucial role in unravelling meaningful insights from the vast and complex dataset. Through the use of charts, graphs, and plots, we aimed to distil intricate patterns, trends, and relationships existing within the data. Visualization enabled us to gain a comprehensive understanding of user engagement metrics, such as app installations, ratings, and reviews, across different categories. This approach not only facilitated the identification of popular app genres but also allowed us to assess the impact of various factors on app performance. Additionally, visualizing market trends, outliers, and distribution of key variables aided in making informed decisions regarding potential strategies for app development, marketing, and user experience enhancement. Ultimately, data visualization served as a powerful tool for effectively communicating our findings to stakeholders, providing a visually compelling narrative that supported the conclusions drawn from the exploratory data analysis of the Google Play store dataset.

In this case study we generate three graphical representations-

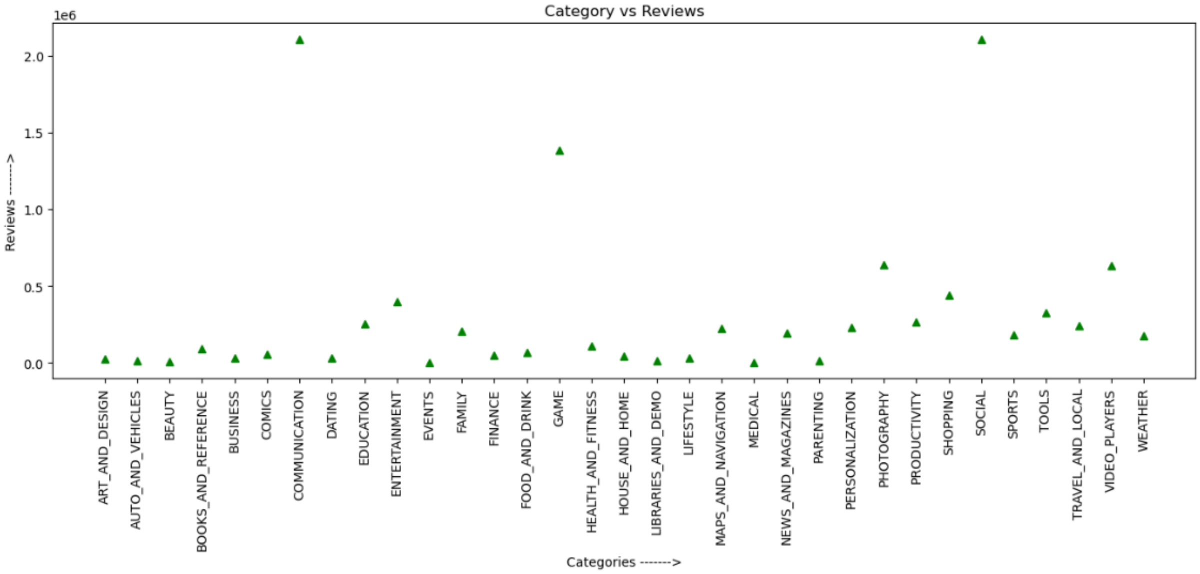
1.Category VS Install



2.Category VS Pricing



3.Category VS Reviews



5. CONCLUSION & FUTURE WORK

Most of the apps are free so developers should focus on creating free apps to have a huge customer base. More Apps should be in the category like Events, Beauty, Parenting as they have not been explored much but still quite popular with huge installations. In order to retain the customer base apps should be updated regularly Developers should develop apps such that their content is available for everyone.

1. Most common category of apps on the Play Store is: **Family**

2. Percentage of free apps on the Play Store is: **~92%**

3. Category with the greatest number of app installs on the Play Store is: **Communication**

4. Category with the greatest number of reviews is: **Communication**

5. Category with the most expensive apps on the Play Store is: **Finance**

Future Scope

1. Advanced Predictive Modelling:

Future exploration could involve implementing advanced machine learning models to predict app success metrics, such as user ratings or download counts, based on historical data patterns from the Google Play store dataset.

2. User Behaviour Analysis in Real-Time:

Integrating real-time data streams can enable continuous monitoring and analysis of user behaviour, allowing for dynamic adjustments in app strategies and features to meet evolving user preferences.

3. Sentiment Analysis and User Feedback Mining

Extending the analysis to include sentiment analysis and mining user reviews can provide valuable insights into user satisfaction, helping developers understand specific strengths and weaknesses of apps in the Google Play store ecosystem.

4. Market Trend Forecasting

Exploring time-series analysis and forecasting techniques can contribute to predicting future market trends, allowing app developers and marketers to stay ahead by adapting strategies to changing preferences and market dynamics.

5. Cross-Platform Analysis

Considering the increasing diversity of platforms, expanding the analysis to include data from other app distribution platforms can provide a holistic view of the mobile app landscape, aiding in strategic decision-making beyond the Google Play store.

6. REFERENCES

- Blog Reference:

  Play Store EDA on Kaggle

- Data Set from Kaggle:

  Kaggle Link