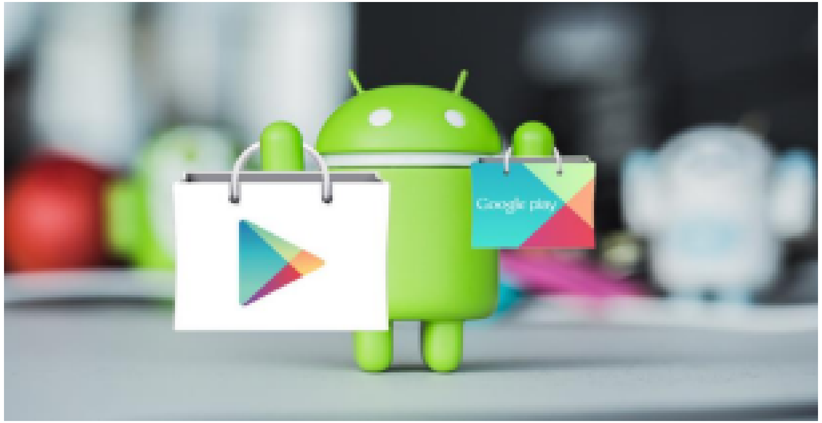
Capstone Project-1

(Exploratory Data Analysis)

Play Store App Review Analysis



GitHub https://github.com/Ashif-khan033/Play-Store-Apps-Review-Analysis-EDA.git

Team : DATA BATTALION

1. Ishan Sharma | ishansharma1132@gmail.com 2. Mohd Ashif Khan | khanashif033@gmail.com 3. Virender Chib | virenderchib9@gmail.com

4. Nitesh Pawar |pawarnitesh09@Gmail.com 5. Kismat Choudhary |kismatchoudhary002@gmail.com

Contents:

1. Introduction
2. Problem Statement
3. Dataset Description
    a) Play store App Data
    b) User Reviews Data
4. Data Processing Flowchart
5. Import Libraries
6. Data Loading
7. Null values Imputation / Data Cleaning
8. Data Segregation
9. Exploratory & Visualization Analysis
10. Conclusion

Introduction

- 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.

Problem Statement

New app maker company is trying to identify various factors to capture android market ,so they are trying to find out the following things from the app store data: |**Problem 1:**| How many Android Version Supported Apps Across the Whole Database?

|**Problem 2:**| What are the top most competing categories in play store ?

|**Problem 3:**| What is the Paid and Free apps ratio from all apps ?

|**Problem 4:**| What are the up gradation details of apps by year ? |**Problem 5:**| How the App pricing trend across popular categories? |**Problem 6:**| What are the Sentiment Data Across the All Reviews ? |**Problem 7:**| Is there any correlation between Sentiment polarity/Sentiment subjectivity with Installs/Rating/Reviews/Size/Price/Last Updated_Day/Last Updated_Month/Last Updated_Year ?

Dataset Description Column Wise of Play store Data.Csv

Two different datasets provided for analysis:

1.Play Store Data.csv

App : Categorical, the app name.

Category : Categorical, category the app belongs to.

Rating : Numerical, range from 0.0 to 5.0,Rating has received from the users. **Reviews** : Numerical, the number of reviews that the app received. **Size** : Numerical, the size of the app. The suffix M - megabytes, K - kilobytes. **Installs** : Numerical, describes the number of installs.

Type : Categorical, a label that indicates whether the app is free or paid. **Price** : Numerical, the price value for the paid apps.

Content Rating : Categorical, a categorical rating that indicates the age group for user.

Genre : Categorical, list of genres to which the app belongs. **Last Update** : Date Format, the date at which the app was last updated. **Current Version** : Version of the app as specified by the developers. **Android Version** : The Android OS the app is compatible with.

Dataset Description Column Wise of User Reviews.Csv

2.User Reviews.csv

App : the app name.
Translated_Review : the review text in English.
Sentiment : the sentiment of the review, positive, neutral, or negative.
Sentiment_Polarity : the sentiment in numerical form, ranging from -1.00 to 1.00.
Sentiment_Subjectivity : a measure of the expression of opinions, evaluations, feelings, and speculations.

Data Processing Flowchart

Import Libraries

library is a **collection of related modules**. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer.

Data Frame Loading.......

DataFrame Loading…….

1. The head () method returns a specified number of rows, string from the top.

2. The tail () method returns a specified number of rows, string from the bottom. Default o/p 5 bottom rows.

Drop Duplicate Entry

• Duplicate Enter data is repeated data in data frame,
• Either multiple columns have same data or multiple rows have same data. • Repeated
Entry add sums of number in final data and it miscalculates the true outcome. •
df.duplicated() method is used to identify duplicates from dataframe

1.Shape of data (Rows & Columns)2. Identify duplicate entries

3. keep the last instance of a duplicate row in dataframe

Drop Duplicate Entry

1.Print Duplicate Entry Data of play store_df

2.Remove Duplicate Entry Data from play store_df, print shape of data after removing duplicate entry.

Remove Visual impurities

Visual Impurities is like eyesight impurities in dataframe, without additional deep finding visible impurities

column have “+” sign

1. Check data frame and replace + and $ with blank so it works like delete.

Data after filter +

Find Unique, Null Count and Data type

For more cleaning of data we identified Unique data – Null Counts and Data type of each column ugh one that is perhaps not usable Data type -

Total value of App and Unique value is different, which means some App entries are repeated.

Information From Data Digging

Datatype are
not correct for

some columns
like price , size,
Installs and
Last Update

database are missing

Some columns have null values that need to be filled or filtered

Total count of each rows
are not same so

Data Filtering Column by Column

All Columns have some mistakes and unreliable things that need to be filtered

No	DataType	Mistakes and Unreliable things on specific Column
1	**App**	Repeated Entry in app column indicated some repeated apps are there
2	**Category**	1.9 entry in category is outliner, checked data and shifted the row
3	**Rating**	Null values present, filled it with mean() or median()

4	**Reviews**	Data type mistake, it’s a numerical data type
5	**Size**	Converted the size into one single unit
6	**Installs**	After removing +, corrected the data type to numerical
7	**Type**	1 null value present, filled it with proper data (After Crosscheck with play store fill with "Free")
8	**Price**	After removing $ sign, need to correct data type to numerical
9	**Content Rating**	Did not require any operation
10	**Genres**	Did not require any operation
11	**Last Updated**	Data type needed to replace to Date Format : datetime64[ns]
12	**Current Ver**	Did not require any operation
13	**Android Ver**	Did not require any operation

App Column Operation

1. Identifying the duplicate apps.

2. Remove duplicates and clean data and store in playstore_df3 data frame.

Category Column

Operation

1. 2.

1.outliner

2. Shift data

3. Replace with “LIFESTYLE”
(check with play store)

Rating Column Operation

1. Find the

Unique value.

2. Count the Null value.

3. Fill it with Median.

Review Column Operation

1. Convert the data type to “int”

Installs Column Operation

1. Convert the data type to “int”

Type Column Operation

Size Column Operation

1. Find Unique values

2. Identify Row Number

3. Replace with “Free” (check with play store.)

4. Check unique values after replace value.

,

2. Convert MB to KB and Remove K & M

3. Filter Data output

Price Column Operation

Content Rating Column Operation

1. Find value count of each type

2. Unrated apps ,also considerable for Everyone

Last Update Column Operation

1.Converted to datetime datatype

User Reviews Dataframe Operation

2. Separate Month, Day, year and Store in new column

1. Unreviewed Columns Dropped.

2. Describe Numerical Data

- Sentiment Polarity is Between -1 to 1
- Sentiment Subjectivity is Between 0 to 1

Segregate : Numerical, Categorical Variable Data

Segregate

Numerical and Categorical Variable is useful to do process fast and easily plotting

1. Dtype is not an object it's a

numerical data

3. Dtype is equal to object then it’s a categorical data

2. Output of numerical data

Data Visualisation

- **Correlations in Relationships**: Without data visualization, it is challenging to identify the correlations between the relationship of independent variables. By making sense of those independent variables, we can make better business decisions.
- **Trends Over Time**: While this seems like an obvious use of data visualization, it is also one of the most valuable applications. It’s impossible to make predictions without having the necessary information from

the past and present. Trends over time tell us where we were and where we can potentially go.

- **Frequency**: Closely related to trends over time is frequency. By examining the rate, or how often, customers purchase and when they buy gives us a better feel for how potential new customers might act and react to different marketing and customer acquisition strategies.
- **Examining the Market**: Data visualization takes the information from different markets to give you insights into which audiences to focus your attention on and which ones to stay away from. We get a clearer picture of the opportunities within those markets by displaying this data on various charts and graphs.
- **Risk and Reward**: Looking at value and risk metrics requires expertise because, without data visualization, we must interpret complicated spreadsheets and numbers. Once information is visualized, we can then pinpoint areas that may or may not require action.
- **Reacting to the Market:** The ability to obtain information quickly and easily with data displayed clearly on a functional dashboard allows businesses to act and respond to findings swiftly and helps to avoid making mistakes.

Problem 1: *How many Android Version Supported Apps Across the Whole Database ?*

Summary:

After identifying the total distribution percentage on data given details of more app supported Android OS versions .

Basically android 4.0 and above version supported app ratio is very higher and more than 60% app's support only on android 4.0 and above version.

Problem 2: What are the top most competing categories in play store ?

Problem 3: What is the Paid and Free apps ratio? from all apps ?

Summary:

A)None value in type column are 0.01%
B)Free apps in play store are 92.60%
C)Paid apps in play store are 7.39%

So, we can clearly see that only 7.39% are paid apps available on play store and the rest 92.60% are free.

Problem 4: What are the updation details of apps by year?

Summary:

Here, we are analyzing the added apps and up gradation of apps on play store. So, it's clearly visible that between time period 2017-2018, there are very wide range of app up gradation and addition occur.

Problem 5: How the App

pricing trend across popular categories ?

Summary:

1.From the above analyses as we find that there are different categories of apps demand different price ranges.
2.Like in simple & easy apps in category FAMILY LIFESTYLE, FINANCE and MEDICAL are high in price.
3. All Games apps are comparatively low in price. So, it could be the reason that it's have more downloads.

Problem-6: What are the

Sentiment Data Across the All Reviews ?

Summary:

1. From the above pie chart, we can say that the most of the reviews on the apps by the users on play store has received positive reviews i.e.(64.03%).
2. While, some of the apps have received negative reviews. i.e.(approx.15%).

Problem 7: Is there any correlation between Sentiment polarity/Sentiment subjectivity with Installs/Rating/Reviews/Size/Price/LastUpdated_Day/LastUpdated_Month/LastUpdated_Year ?

Summary:

We add Both the data frames that are related to each other. In these correlation heat map some values are negative and some are positive , for that purpose we merge the both data frames to obtain the most appropriate results and yes, we observed that:

1.Size and sentiment polarity are negatively correlated (-0.12) : There are lots of reasons of disliking those apps which have large size. First of all, it consumes more storage, takes more RAM and needs a high speed connection for its execution.

2.There is a positive correlation between reviews and number of installs (0.63) because as the reviews increases, people start noticing the app and install them.

3.There is a slightly positive correlation (0.27) between sentiment polarity and sentiment subjectivity that means if the users shares positive reviews then there is chance that users are sharing their personal opinion but not genuine information.

Conclusion

Through exploratory data analysis we have observed some trends and have made some assumptions that might lead to app success among the users in the play store.

• Android Version Supported Apps: - **60% apps support only android 4.0** and above version.

• Categories of app in play store :- most like or preferred by customers **family apps and gaming apps**

• Percentage of free apps :- **92.60% apps are free** available on the play store

• Percentage of Paid apps: - **7.39% apps are paid** apps. • Up gradation details of apps by year: - Latest **6000+ apps updated in 2018** which is the **maximum count**. • App pricing in popular categories: - we see **price upto 400 dollar** & **mostly medical and family apps are paid**. • Sentiments of reviews: - **64.03% customer gave positive feedbacks** while the **negative feedback is 21.28%**.