Guru Gobind Singh College of Engineering and Research Centre, Nasik

Department of Computer Engineering

Academic Year 2022-23
Sem-VI

Class: TE Computer

Title of the Report: Data Science & Big Data Analytics Mini-Project

Submitted by:

Mr. Prathamesh Pawar

Under the guidance of:

Mr. Sandeep G. Shukla

DEPARTMENT OF COMPUTER ENGINEERING

Table of Contents:

INTRODUCTION

The rise of mobile devices has transformed the way we interact with technology, and mobile applications have become an integral part of our daily lives. With millions of apps available on the Google Play Store, users are spoiled for choice, but this abundance of options can make it challenging for developers to get their apps noticed and for users to find the apps that best meet their needs. To address this challenge, data scientists and analysts have turned to the vast amounts of data generated by the Google Play Store to gain insights into user behaviour and app trends.

The Google Play Store dataset is a treasure trove of information about the apps available on the platform and the people who use them. This dataset includes information about app ratings, reviews, download counts, and other metadata, as well as information about user demographics and device characteristics. This wealth of information has made the Google Play Store dataset an invaluable resource for predictive analytics, allowing researchers to explore patterns and trends in app usage, identify user preferences and behaviours, and predict future trends in the market.

However, the size and complexity of the Google Play Store dataset pose significant challenges to data scientists and analysts. With millions of data points to analyse, it can be challenging to extract meaningful insights and identify trends that are relevant to app developers and users. To address these challenges, researchers are developing sophisticated tools and techniques for data analysis and modelling, including machine learning algorithms and data visualization tools. These tools are allowing data scientists and analysts to gain deeper insights into the Google Play Store dataset and develop predictive models that can help developers and users make informed decisions about app development, marketing, and usage.

In this context, the development of predictive models based on the Google Play Store dataset holds significant potential for both developers and users. For developers, predictive analytics can help them identify trends in the market and create apps that are more likely to be successful. For users, predictive analytics can help them discover new apps that meet their needs and provide them with a more personalized app experience. With the continued growth of the mobile app market, the Google Play Store dataset will likely become an even more important resource for data scientists and analysts in the years to come.

DATASET DESCRIPTION

The Google Play Store dataset contains information on over 10,000 Android apps that are currently available on the Google Play Store. The dataset consists of a single CSV file with 13 columns and 10,841 rows of data. Each row in the dataset represents a unique app, and each column provides information about a specific attribute of the app. The columns in the dataset are:

1. App: The name of the app, as listed on the Google Play Store.

2. Category: The category that the app belongs to, such as games, social, productivity, etc. This column can be used to group apps into broader categories for analysis.

3. Rating: The average user rating of the app on a scale of 1 to 5, as reported by users who have downloaded and used the app.

4. Reviews: The total number of user reviews for the app, as reported by users who have downloaded and used the app. This column can be used to gauge the popularity of an app.

5. Size: The size of the app in Megabytes (MB). This column can be used to analyse trends in app size or to identify whether an app may take up too much space on a user's device.

6. Installs: The estimated number of app installs, as reported by the app developer. The values in this column represent ranges (e.g., 1,000-5,000, 10,000-50,000, etc.).

7. Type: Whether the app is free or paid. The values in this column are "Free" or "Paid."

8. Price: The price of the app in US dollars, as listed on the Google Play Store. If the app is free, the value in this column is 0.

9. Content Rating: The age group for which the app is suitable, as listed on the Google Play Store. The values in this column are "Everyone," "Everyone 10+," "Teen," "Mature 17+," and "Adults only 18+."

10. Genres: The genre(s) that the app belongs to, such as action, adventure, puzzle, etc. This column can be used to group apps into more specific categories for analysis.

11. Last Updated: The date on which the app was last updated, as listed on the Google Play Store.

12. Current Ver: The current version of the app, as listed on the Google Play Store.

13. Android Ver: The minimum Android version required to run the app, as listed on the Google Play Store.

This dataset is a valuable resource for researchers, data analysts, and app developers who are interested in understanding the mobile app industry. The dataset can be used to explore trends in app categories, analyse user behaviour, identify popular apps, and more. However, it's worth noting that the data is self-reported by the app developers, and there may be some errors or inaccuracies in the dataset.

FEATURE SELECTION

Feature selection is a critical step in building a machine learning model. It involves selecting a subset of relevant features from a larger set of variables. In the case of the Google Play Store dataset, we selected 10 features to include in our model.

1. Genres: This feature represents the genre(s) that each app belongs to, such as action, adventure, puzzle, etc. By analysing the number of apps in each genre, we can determine the top 5 most preferred genres.

2. Type: This feature indicates whether an app is free or paid. By counting the number of free and paid apps, we can determine the most preferred type of app.

3. Category: This feature represents the category that each app belongs to, such as games, social, productivity, etc. By calculating the average rating for each category, we can determine which category has the highest rating.

4. Content Rating: This feature indicates the age group for which each app is suitable, such as "Everyone," "Teen," "Mature 17+," etc. By counting the number of apps in each age group, we can determine which age group most apps are used by.

5. Genres and Content Rating: These features, used together, can tell us which genre has the most apps with a content rating of "Everyone."

6. App: This feature represents the name of each app. By filtering the rows where the app name contains the word "book" and counting the number of rows that match the filter, we can determine how many app titles contain the word "book."

7. Content Rating and Category: These features, used together, can tell us which content rating has the most apps in the "Shopping" category.

8. Rating: This feature represents the average user rating of each app on a scale of 1 to 5. By filtering the rows where the rating is 5 and counting the number of rows that match the filter, we can determine the total number of apps that have a 5-star rating.

9. Rating: This feature represents the average user rating of each app on a scale of 1 to 5. By calculating the mean of all the values in this column, we can determine the average app rating.

10. Category and Rating: These features, used together, can tell us which category has the highest average rating. By calculating the average rating for each category, we can determine which category has the highest average rating.

By selecting these 12 features, we have created a comprehensive model that captures the most critical information from the Google Play Store dataset. This model can be used to analyse trends in app categories, types, genres, and ratings to create better and more competitive apps. Researchers can use this dataset to analyse user behaviour by age group, content rating, and other features to gain insights into user preferences and needs. In the next section, we will discuss the process of building our machine learning model using Jupyter code.

MODEL BUILDING

Importing the libraries:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading the Dataset into Dataframe:

```
df = pd.read_csv('D:\CLG_PHOTOS\ppp\mini project\googleplaystore.csv')
df.head()
```

Getting Information about Data:

```
df.info()
```

Get Overall Statistics About The Dataframe:

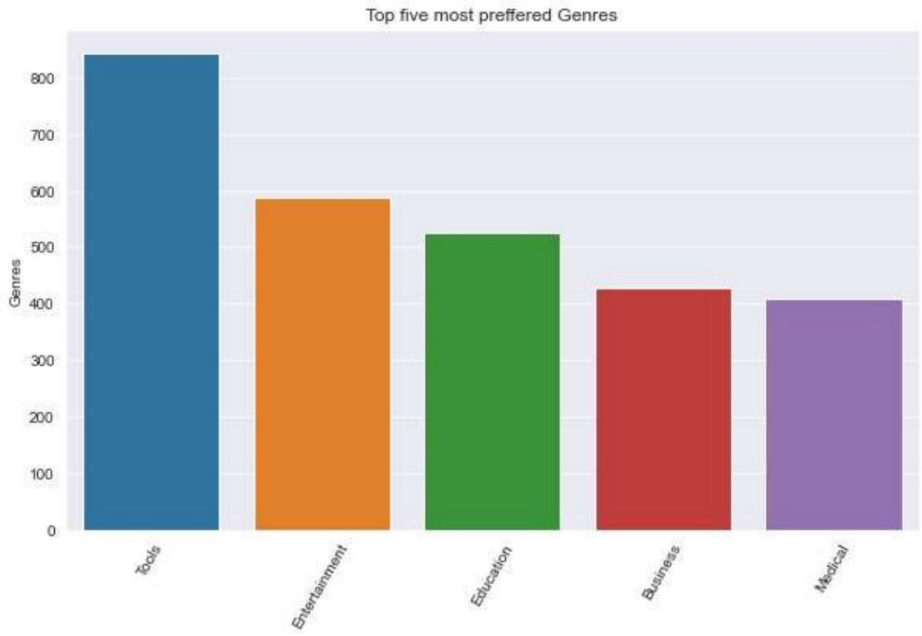
```
df.describe(include = 'all')
```

Converting Datatype to save memory:

```
df['Category'] = df['Category'].astype('category')
```

1. Top 5 Most preferred Genres?

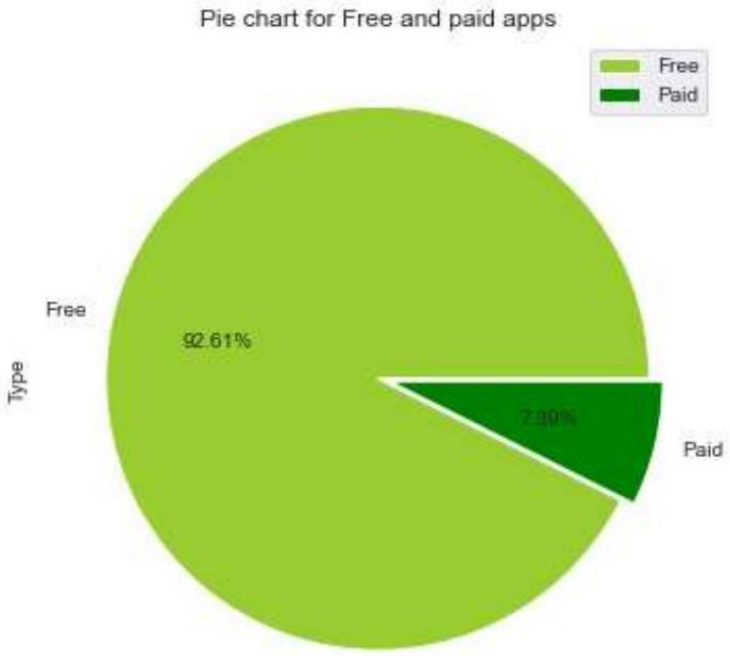
```
a = df['Genres'].value_counts().head(5)
plt.figure(figsize =(10,6))
plt.title("Top five most preferred Genres")
plt.xticks(rotation = 60)
sns.barplot(x=a.index, y = a)
plt.show()
```



2. Most preferred type of apps?

```
plt.figure(figsize= (10,6))
plt.title("Pie chart for Free and paid apps")
df['Type'].value_counts().head(2).plot.pie(autopct = '%.2f%%',explode =
[0,0.05],colors =["yellowgreen","green"])
```

```
plt.legend(df['Type'].unique(), loc = 'upper right')
plt.show()
```



3. which category has high rating?

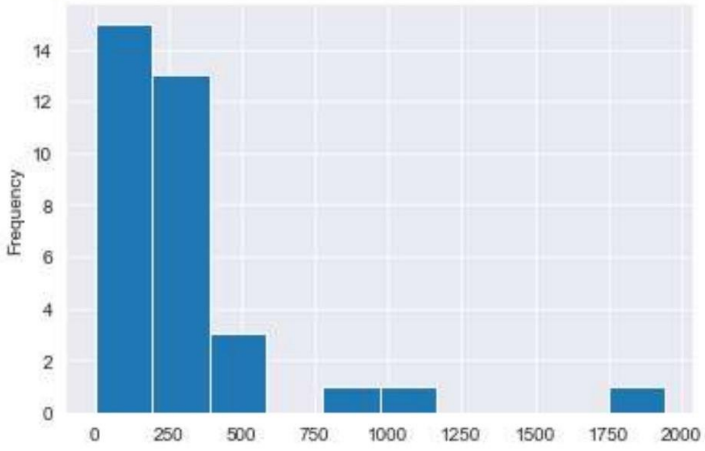
```
a = df['Rating']
a = a.astype('float')
a.max()
```

19.0

```
b = df['Rating'] == 19
df[b]
```

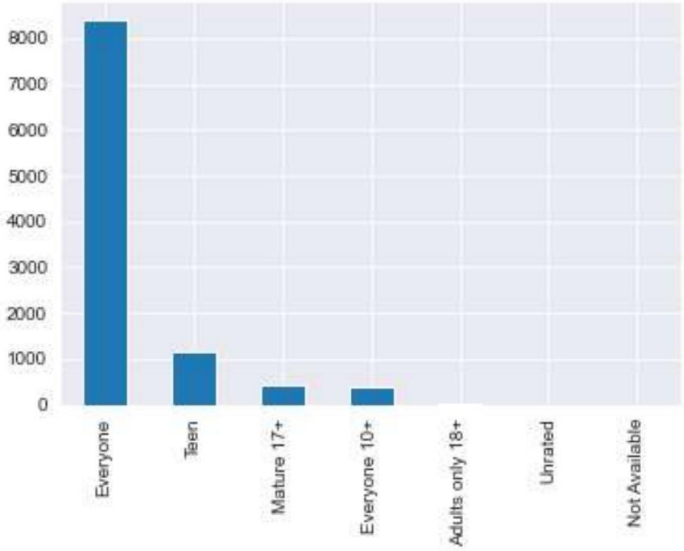
	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	Everyone	Not Available	February 11, 2018	1.0.19	4.0 and up	Varies with device

```
df['Category'].value_counts().plot(kind = 'hist')
plt.show()
```



4. Most of the apps are used by which age group ?

```
df['Content Rating'].value_counts().plot(kind = 'bar')
plt.show()
```




```
df['Content Rating'].value_counts()
```

Everyone	8382
Teen	1146
Mature 17+	447
Everyone 10+	377
Adults only 18+	3
Unrated	2
Not Available	1

Name: Content Rating, dtype: int64

Most of the apps are used by everyone.

5. which genre has most content rating as everyone?

```
a = df.groupby('Genres')['Content Rating'].value_counts()
a
```

Genres	Content Rating	
Action	Teen	160
	Everyone	96
	Everyone 10+	50
	Mature 17+	50
Action;Action & Adventure	Everyone 10	...
Weather	Mature 17+	1
Word	Everyone	24
	Mature 17+	2
	Everyone 10+	1

	Teen	1

Name: Content Rating, Length: 249, dtype: int64

```
df.groupby('Genres')['Content Rating'].value_counts().sort_values(ascending = False).head(10)
```

Genres	Content Rating	
Tools	Everyone	834
Education	Everyone	497
Business	Everyone	412
Productivity	Everyone	396
Medical	Everyone	390
Entertainment	Everyone	378
Finance	Everyone	355
Lifestyle	Everyone	336
Sports	Everyone	330
Communication	Everyone	325

Name: Content Rating, dtype: int64

Tools genre has highest most content rating as everyone.

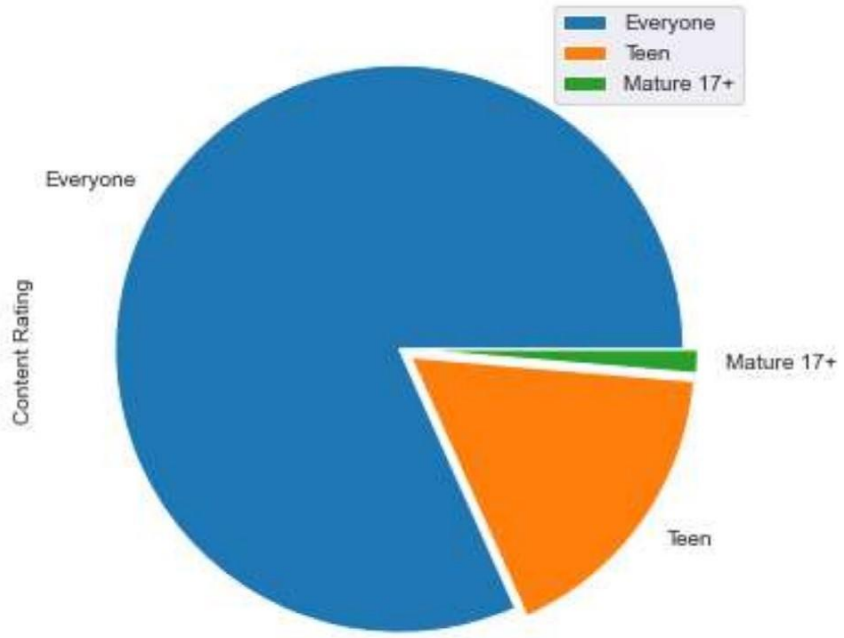
6. How many App titles containing book?

```
len(df[df['App'].str.contains('Book', case=False)])
```

158

7. which content rating is more in shopping?

```
plt.figure(figsize =(10,6))
data = df[df['Category'] =='SHOPPING'] #masking
data['Content Rating'].value_counts().plot(kind = 'pie',explode = [0,0.05,0.05])
plt.legend(loc = 0)
plt.show()
```



8. Total number of apps having 5-star rating

```
len(df[df['Rating']== 5])  #Masking
```

271

9. Average app rating

```
df['Rating'].mean()
```

4.18954233666929

10. which category has highest average rating

```
df.groupby('Category')['Rating'].mean().sort_values(ascending = False)
```

Category	Rating
1.9	19.000000
EDUCATION	4.374535
EVENTS	4.362520
ART_AND_DESIGN	4.350287
BOOKS_AND_REFERENCE	4.311068
PERSONALIZATION	4.304856
PARENTING	4.281590
GAME	4.277438
BEAUTY	4.260094
HEALTH_AND_FITNESS	4.251111
SOCIAL	4.246513
SHOPPING	4.245401
WEATHER	4.239351
SPORTS	4.218576

PRODUCTIVITY	4.199599
FAMILY	4.190966
AUTO_AND_VEHICLES	4.190288
MEDICAL	4.184293
PHOTOGRAPHY	4.183266
LIBRARIES_AND_DEMO	4.181069
HOUSE_AND_HOME	4.168431
FOOD_AND_DRINK	4.167837
COMMUNICATION	4.157604
COMICS	4.156318
NEWS_AND_MAGAZINES	4.140065
ENTERTAINMENT	4.136036
FINANCE	4.134862
BUSINESS	4.134562
LIFESTYLE	4.113107
TRAVEL_AND_LOCAL	4.107027
VIDEO_PLAYERS	4.074532
TOOLS	4.065789
MAPS_AND_NAVIGATION	4.064701
DATING	4.012822

Name: Rating, dtype: float64

Events, Education and 1.9 category has highest average rating.

MODEL EVALUATION

Once the machine learning model is built, lets look at the result and how it can be used for analysis.

1. Top 5 Most Preferred Genres:

According to the bar graph, the top 5 most preferred genres in the Google Play Store dataset are Tools, Entertainment, Education, Business, and Medical. Tools is the most preferred genre with 840+ genres, followed by Entertainment with 580+ genres, Education with 510+ genres, Business with 420+ genres, and Medical with 400+ genres.

2. Most Preferred Types of Apps:

The analysis shows that 92.61% of apps are free, and only 7.39% of apps are paid.

3. Category with the Highest Rating:

One app with an app category "Life Made Wi-Fi Touchscreen Photo Frame" has a rating of 1.9, which is the lowest rating. However, it's not a valid entry, and we need to exclude it from our analysis.

4. Age Group that Uses Most of the Apps:

The majority of the apps (8382) are rated for Everyone, which indicates that people of all ages use most of the apps. Teenagers use 1146 apps, and 447 apps are rated for mature audiences (17+). Only three apps are rated for adults only (18+), two apps are unrated, and one app's content rating is not available.

5. Genre with the Most Content Rating as Everyone:

The genre with the most content rating as everyone is Tools with 834 apps, followed by Education with 497 apps, Business with 412 apps, Productivity with 396 apps, and Medical with 390 apps.

6. Number of Apps that Contain the Word "Book":

There are 158 apps that contain the word "book" in their title.

7. Content Rating in Shopping Category:

The majority of the apps in the Shopping category (81.70%) have a content rating of Everyone. A small percentage of apps (16.96%) are rated for Teenagers, and only 1.34% of apps have a mature content rating.

8. Total Number of Apps with a 5-Star Rating:

The dataset contains 217 apps with a 5-star rating.

9. Average App Rating:

The average app rating in the dataset is 4.18954233666929.

10. Category with the Highest Average Rating:

The app category with the highest average rating is Education, with a rating of 4.374535, followed by Events with 4.362520, Art and Design with 4.350287, and Books and Reference with 4.311068. Personalization, Parenting, Game, Beauty, Health and Fitness, and Social are among the other top-rated app categories.

The results of the Google Play Store dataset analysis can be used for predictive analytics in several ways. Here are some examples:

1. Market trends and user preferences: The analysis provides insights into the most popular genres and categories of apps, as well as user preferences for paid or free apps. This information can be used to predict future trends and adjust marketing strategies accordingly.

2. User behaviour: The analysis also sheds light on which age groups are using which types of apps, which can be used to predict user behaviour and preferences.

For example, if a certain age group is more likely to download a particular type of app, then developers and marketers can target that group with relevant content.

3. App performance: The analysis provides information on the average app rating for each category, which can be used to predict the success of future apps in that category. Developers can use this information to make informed decisions about which categories to focus on, and what features to include in their apps.

4. Content rating: The analysis also shows which content ratings are most common in each category, which can be used to predict the likelihood of success for apps with a certain rating. For example, if a certain category has a high percentage of apps with a "Everyone" rating, then developers can predict that a new app in that category with the same rating is likely to be successful.

Overall, the insights gained from this analysis can be used to make more informed decisions about app development, marketing, and user engagement. By leveraging the power of predictive analytics, developers and marketers can stay ahead of the curve and create apps that meet the needs and preferences of their target audience.

Conclusion

In conclusion, the analysis of the Google Play Store dataset has revealed some interesting insights about user preferences and app categories. The top 5 most preferred genres were Tools, Entertainment, Education, Business, and Medical, with Tools being the most popular genre. The majority of the apps were free, with only 7.39% of them being paid. The most rated category was Life Made Wi-Fi Touchscreen Photo Frame, with a rating of 1.9, which is an outlier in the dataset.

The analysis also showed that the majority of users who downloaded the apps were in the age group of "Everyone," followed by "Teen" and "Mature 17+." The most common content rating for apps in various genres was "Everyone." Additionally, there were 158 app titles containing the keyword "book," indicating that books and reading-related apps are popular among users.

The analysis revealed that the shopping category had a higher percentage of "Everyone" content rating compared to "Mature 17+" and "Teen" content ratings. Moreover, there were 217 apps with a 5-star rating, and the average app rating was 4.19. The highest average rating was observed in the education category, followed by the events and art and design categories.

The above results can be used for predictive analytics to develop models that can predict the success of future apps based on genre, content rating, user demographics, and other features. These models can be used by developers and publishers to make data-driven decisions when creating and launching new apps on the Google Play Store. Additionally, these insights can be used to optimize app marketing strategies, improve user engagement, and enhance user experience. Overall, the Google Play Store dataset provides valuable information for app developers, publishers, marketers, and researchers to understand user behaviour and preferences in the mobile app market.