Capstone Project -1

Play Store App Review Analysis

Team Members

Saniya Pirjade
Ankur Sarkar
Vivek Kumar
Ratneshwar Kumar
Pravin Chawan

Content

1. Introduction
2. Problem statement
3. Data cleaning /null value implementation
4. Data processing
5. Data exploration
6. Basic observation
7. Insights from data
8. Conclusion
9. Challenges and future
10. Reference

Introduction

- The mobile industry is growing rapidly, and competition for apps has also grown significantly, so developers need to do enough research to make their apps successful.

- The Google Play Store has been discovered to be the world's largest app market. It has been observed that although it generates more than double the downloads than the Apple App Store, it makes only half the money compared to the App Store.

Why we analyze the play store?

Mobile App Market is set to grow 20% by 2023

Android Apps comprise 75% of the Market Share. 85% share in brazil,india,turkey.

What makes an App popular? Can we predict how popular it’s going to be?

What are some interesting patterns in user behavior related to app usage & feedback.

Problem Statement

- For this project, we analyzed play store data from 2017-18, Google Play Store is the most used app store worldwide and also the global leader in this segment.
- My main objective is to identify the key factors responsible for app success and user engagement.
- Thousands of new apps in various categories are added to the Play Store on a regular basis.
- We found the distribution of every app based on their size, installs, reviews, and much more.

Data cleaning

- We remove the row from the rating column that contained the outlier, and the histogram becomes perfect.
- We found the row number, which was 10472, and to replace all the null values so that the data frame looks nice and meaningful,
- In the rating column, there were the maximum null values, so replace all the null values with the median of all the values of a particular column(which contains numeric data type values) in a data frame. We clean up all null values in the rating column.
- We passed the rating column into that function as the variable. We also replaced the null values for the columns that have characteristic values with their modes. Then we found the price column and the install column.

Data Cleaning

- The Google Play store dataset has 10,841 observations of data with fields.
- data set 1) play store data 2) user reviews

List of fields:

❑ App
❑ Category
❑ Rating
❑ Reviews
❑ Size
❑ Installs
❑ Type
❑ Price
❑ Content rating
❑ Genres
❑ Last updated
❑ Current version
❑ Android version

Play store data

❑ App
❑ Translated review
❑ Sentiment
❑ Sentiment polarity
❑ Sentiment subjectivity

User reviews

Data Cleaning (Contd..)

Understand the structure of the dataset and clean data before analysis

- Finding Missing value in dataset.
- Correct data type(STRING, INT, FLOAT, DATE)
- Replace null value with aggregate function (mean, mode,median)
- Checking outliers.

Data Processing

- The dataset collected from the Play store is semi-structured or unstructured and contains significant superfluous data (defined as not contributing significant meaning). Some data types need to change in the required format, such as string, Int, float, Boolean and date time.

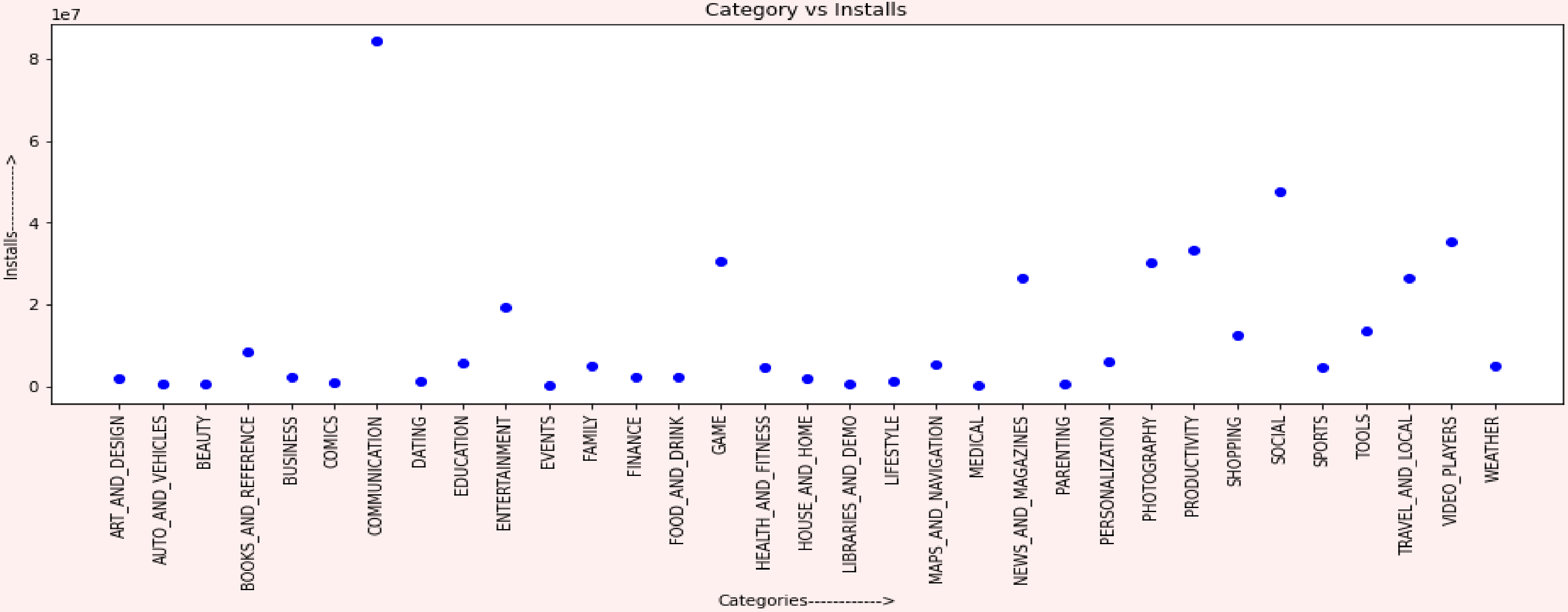
- The size of apps needs to be converted into one measurement, KB or MB. Preprocessing includes various tasks, including stemming, lowercase conversion, units, punctuation, and excluding terms.

Data Visualization

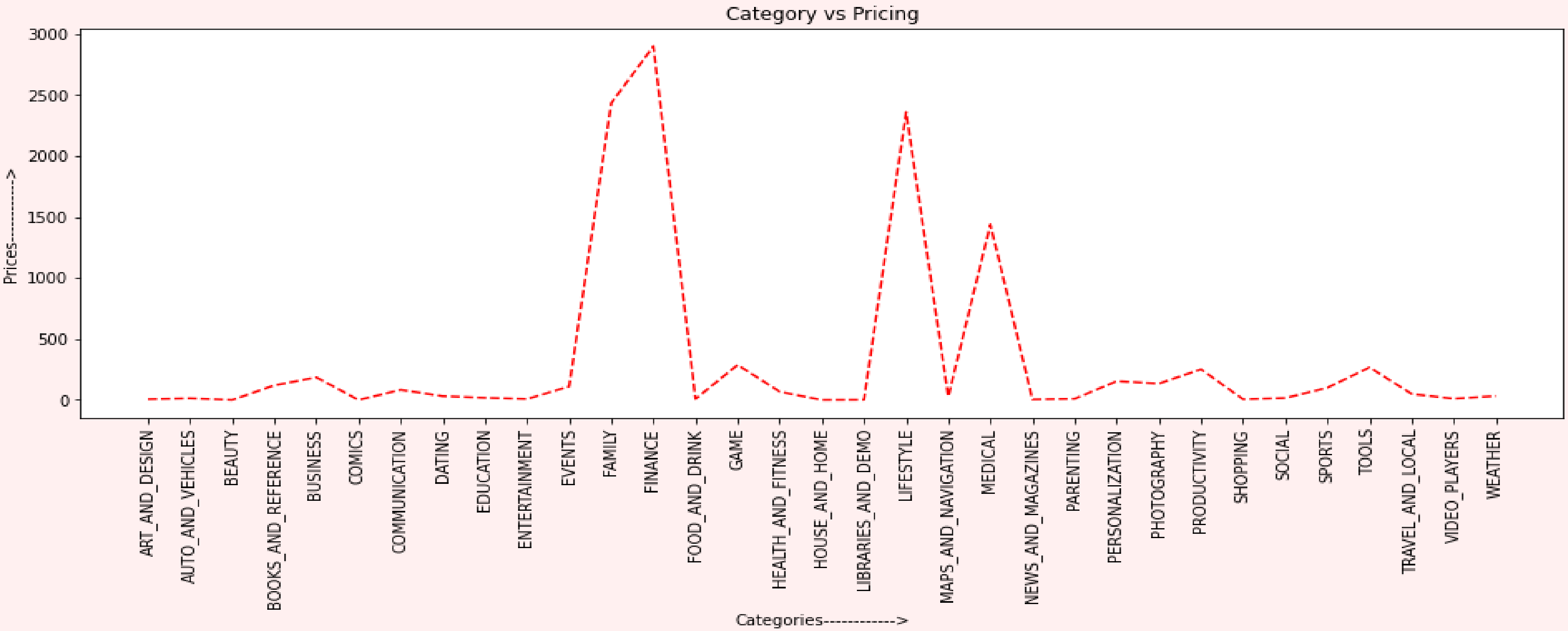
- We used a group by method to group the category column and made the objects for price, install, and review column as their respective sum, mean, and mean values.

- We plotted graphs("Category vs. Installs," "Category vs. Pricing," and "Category vs. Reviews") with the help of matplotlib and the Pyplot library.

Data Visualization



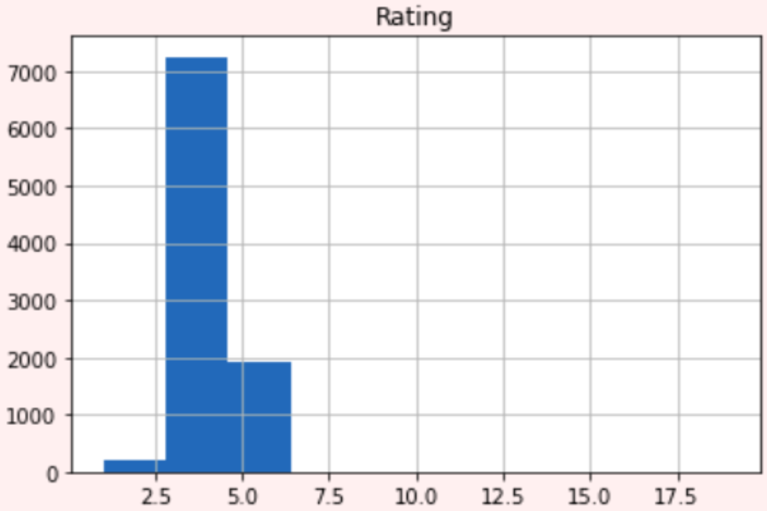
Fact


Category vs Pricing
Prices---------->
3000
2500
2000
1500
1000
500
0
ART_AND_DESIGN
AUTO_AND_VEHICLES
BEAUTY
BOOKS_AND_REFERENCE
BUSINESS
COMICS
COMMUNICATION
DATING
EDUCATION
ENTERTAINMENT
EVENTS
FAMILY
FINANCE
FOOD_AND_DRINK
GAME
HEALTH_AND_FITNESS
HOUSE_AND_HOME
LIBRARIES_AND_DEMO
LIFESTYLE
MAPS_AND_NAVIGATION
MEDICAL
NEWS_AND_MAGAZINES
PARENTING
PERSONALIZATION
PHOTOGRAPHY
PRODUCTIVITY
SHOPPING
SOCIAL
SPORTS
TOOLS
TRAVEL_AND_LOCAL
VIDEO_PLAYERS
WEATHER
Categories------------>

Rating of apps

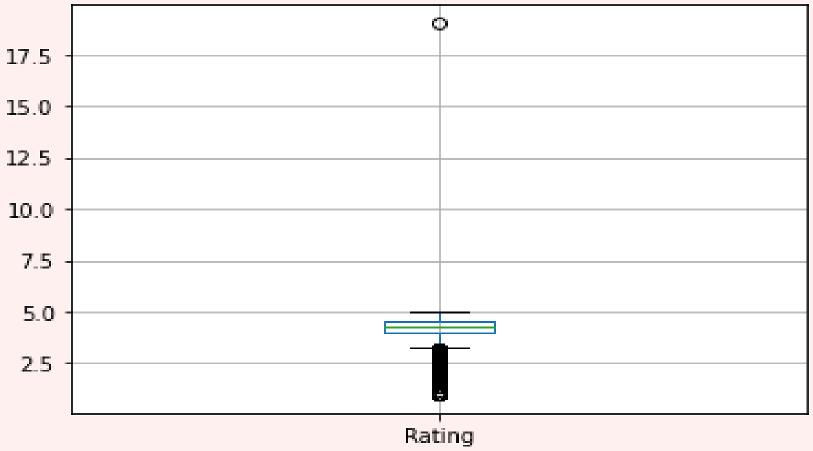
Plotted histogram and box plot for rating.

Rating Distribution of app



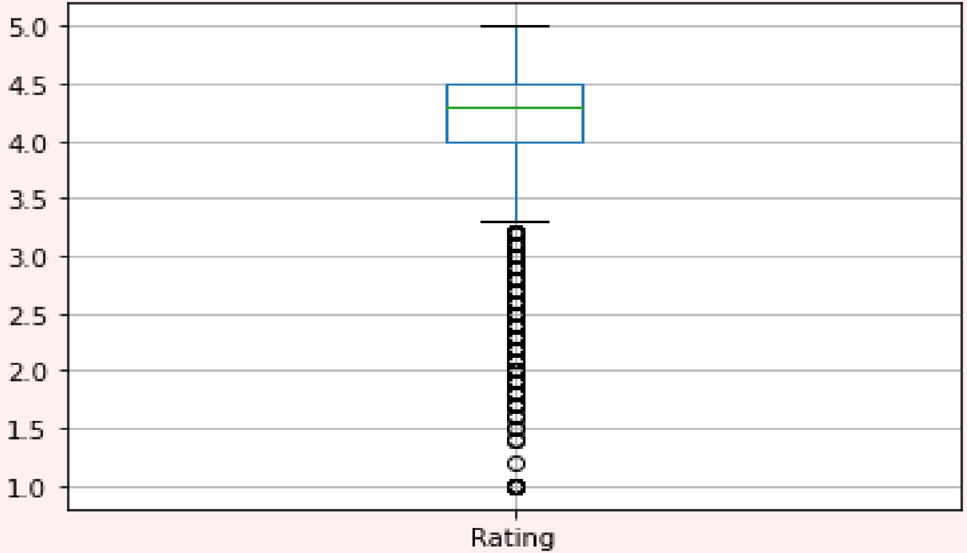
Average rating above 4.2

Content Rating of app



Most apps come under this everyone

Cont..



- The mean of the average ratings (excluding the NaN values) comes to be 4.2.
- The median of the entries (excluding the NaN values) in the 'Rating' column comes to be 4.3. From this, we can say that 50% of the apps have an average rating of above 4.3 and the rest are below 4.3.

Basic observation

Below are some observation by doing data wrangling.

❑ **Average app rating**	**4.30**
❑ Top five category highest average rating	1)Events 2)Education 3)Arts and design 4)Parenting 5)Personalization
❑ App with maximum reviews	Facebook
❑ Top 5 app having highest reviews	1)Face book 2) Subway surface 3)clash royal 4)Candy crush 5)UC-browser
❑ Most expensive app	I’m rich

Insights from Data

World Cloud

- Word Cloud is a data visualization technique used for representing text data in which
  the size of each word indicates its frequency or importance.

Sentiment Polarity

- The polarity of a sentiment measures how negative or positive the context is.

- In the data that we have, the polarity ranges from -1 (most negative) to +1 (most positive).

CONCLUSION

- **The average rating** we see of apps on the Google Play Store is **4.30.**

- Tools, entertainment, education, business, and medical are the top genres.

- Family, Games, and Tools are the top three categories, each having a 1906, 926, and 829 app count.

- **Family and games** apps are the most competitive category.

Cont..

- Users download a given app more if it has been reviewed by a larger number of people.
- Paid apps have a slightly higher number of favorable reviews than free apps.
- Free apps get more negative and neutral feedback, suggesting a wider range of opinions.
- **The Face book** app has the most reviews. While **family and fitness are the most** downloaded app.
- More than half of users rate **family, game, tools, entertainment and fitness apps** positively. Apps for games and social media get mixed reviews, with 50 percent of positive and 50 percent of negative responses.

Challenges

- The dataset contains NULL and NaN values.
- The main task is to clean the data, followed by data processing.
- Some data app names, etc., are in gibberish form and contain duplicates.

Future

- In this project, we perform EDA and discover relationships with specific features using the sentiment of users.

- Developers can use my work for their research purposes to make apps succeed.

Reference

- The data set consists of Google Play Store applications and is taken from Almabetter.

- Research paper based on play store analysis.

- https://learn.almabetter.com/courses/take/team-capstone-projects/presentations/25003924-sample-project-presentation

- https://github.com

Thank you