

Capstone Project -1

Play Store App Review Analysis



Team Members

Saniya Pirjade

Ankur Sarkar

Vivek Kumar

Pravin Chawan

Ratneshwar Kumar

Points of Discussion

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 application 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 25%
by 2026



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 found the row number, which was 10841, and we replaced all the null values so that the data frame looks nice and meaningful.
- To make the histogram perfect, we remove the row from the rating column that contains the outlier.
- 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 Play store dataset has 10,841 observations of data with fields.
- data set 1) play store data 2) user reviews

List of fields:

<ul style="list-style-type: none"><input type="checkbox"/> App<input type="checkbox"/> Category<input type="checkbox"/> Rating<input type="checkbox"/> Reviews<input type="checkbox"/> Size<input type="checkbox"/> Installs<input type="checkbox"/> Type<input type="checkbox"/> Price<input type="checkbox"/> Content rating<input type="checkbox"/> Genres<input type="checkbox"/> Last updated<input type="checkbox"/> Current version<input type="checkbox"/> Android version	<p>Play store data</p>	<ul style="list-style-type: none"><input type="checkbox"/> App<input type="checkbox"/> Translated review<input type="checkbox"/> Sentiment<input type="checkbox"/> Sentiment polarity<input type="checkbox"/> Sentiment subjectivity	<p>User reviews</p>
--	---------------------------------------	--	--------------------------------

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.

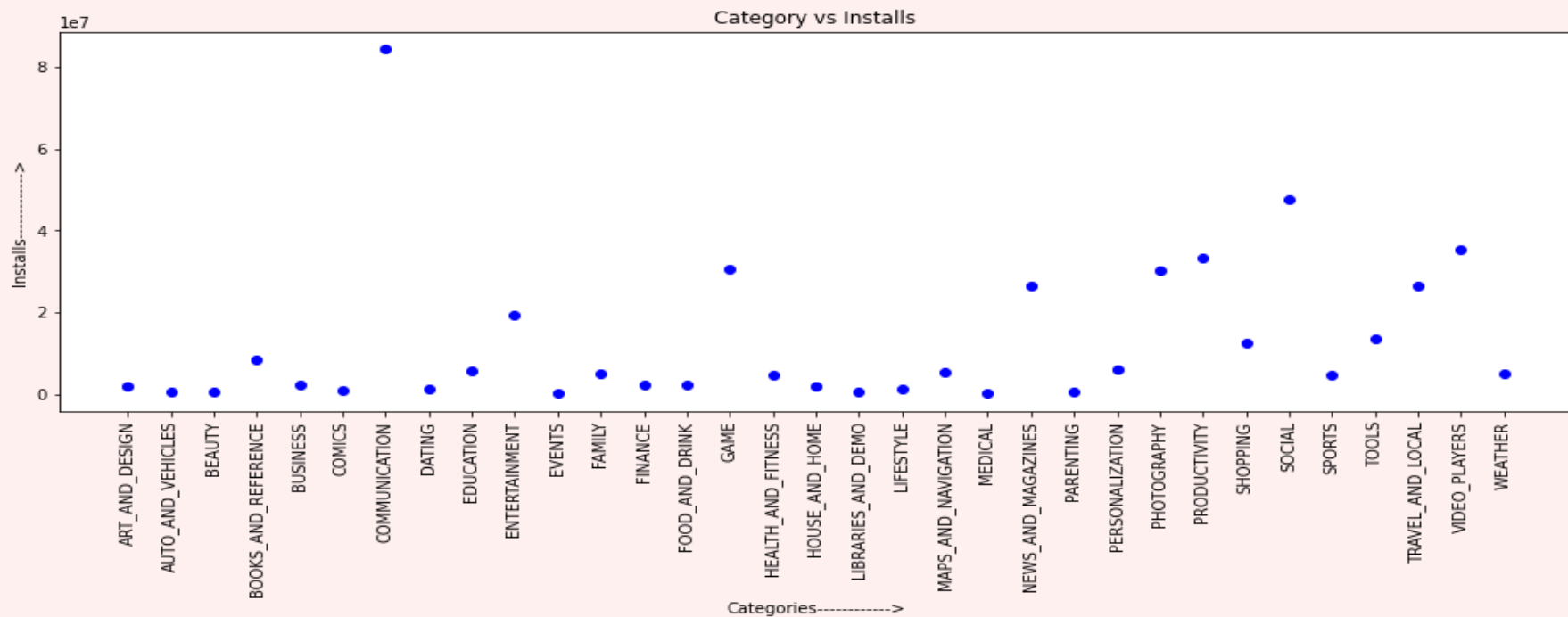


- The size of apps needs to be converted into one measurement, KB or MB. Preprocessing includes various tasks, including stemming, lowercase conversion, units, punctuation, and excluding terms.

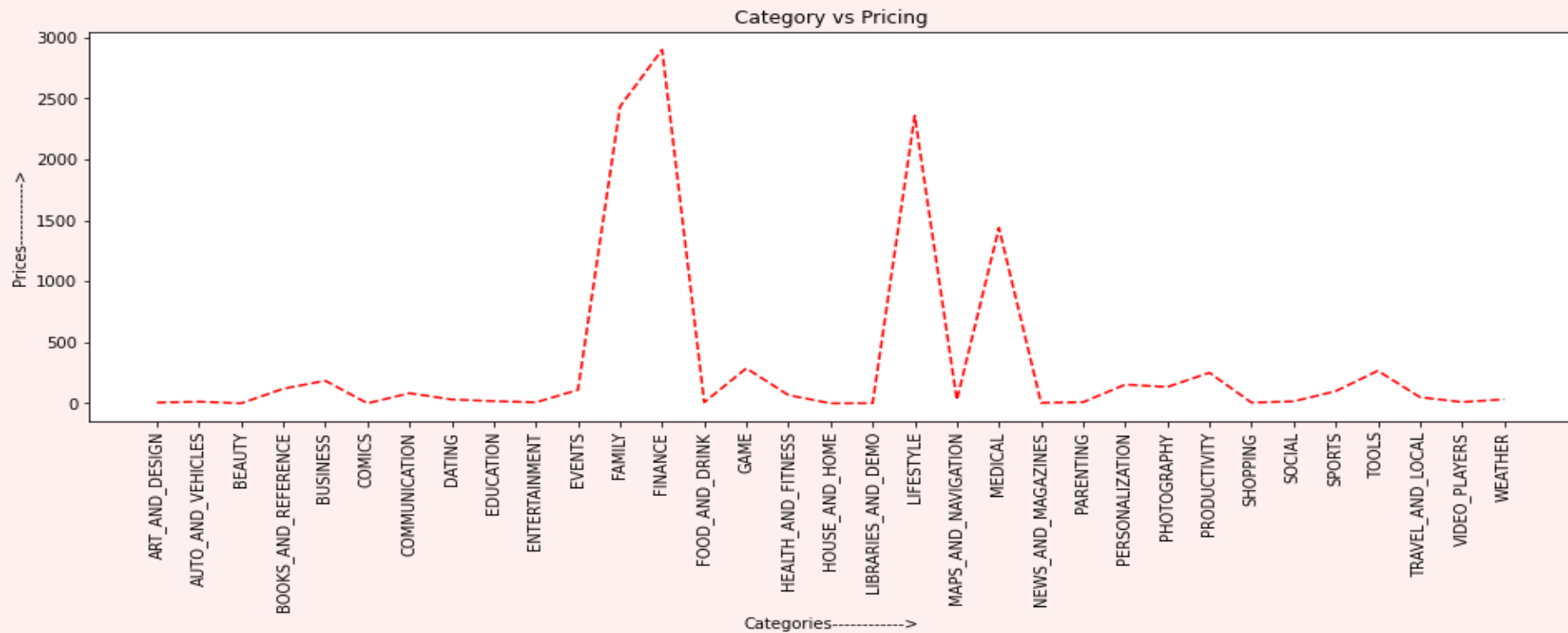
Data Visualization

- 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



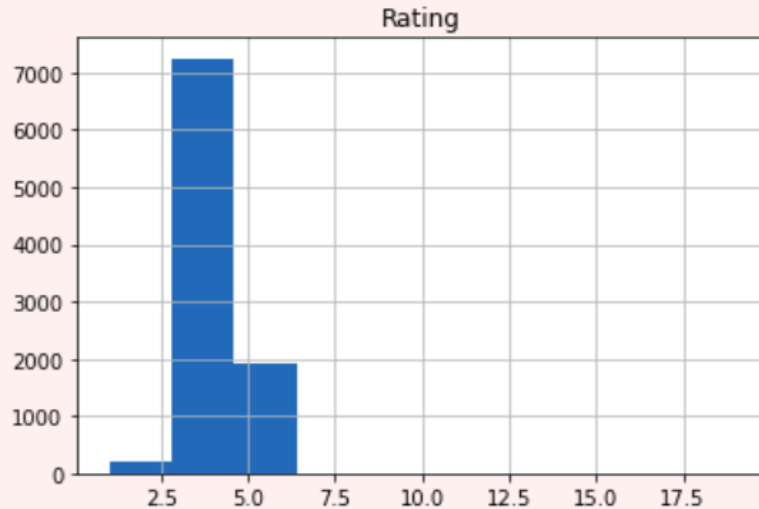
Fact



Rating of apps

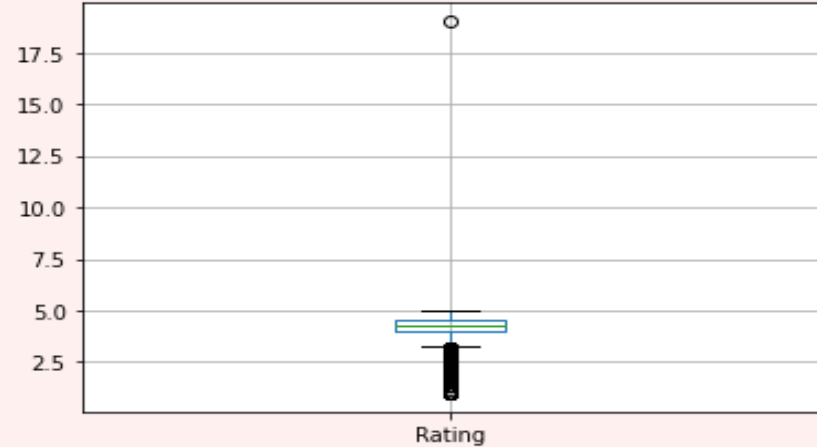
Plotted histogram and box plot for rating.

Rating Distribution of app



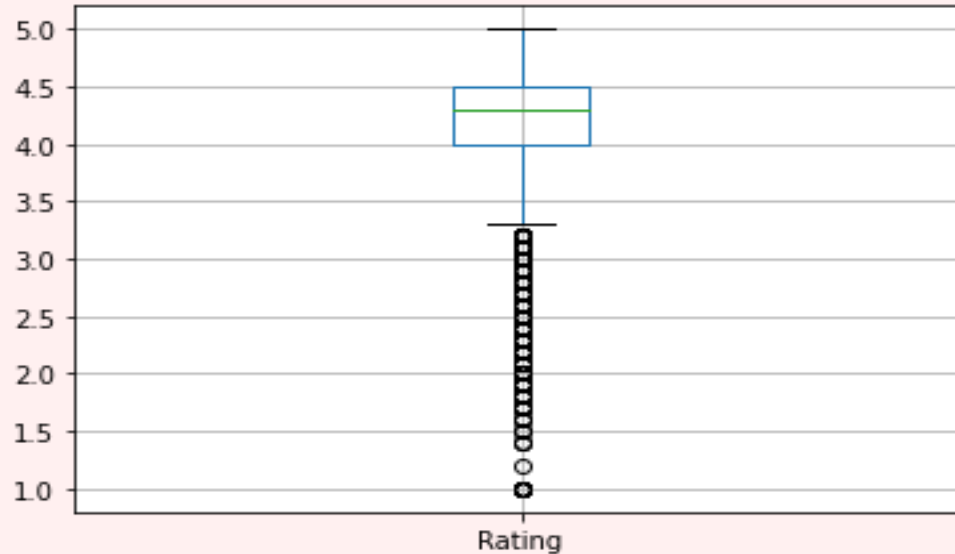
Average rating above 4.2

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.

<input type="checkbox"/> Average app rating	4.30
<input type="checkbox"/> Top five category highest average rating	1)Events 2)Education 3)Arts and design 4)Parenting 5)Personalization
<input type="checkbox"/> App with maximum reviews	Facebook
<input type="checkbox"/> Top 5 app having highest reviews	1)Face book 2)Subway surface 3)clash royal 4)Candy crush 5)UC-browser
<input type="checkbox"/> Most expensive app	I'm rich

Insights from Data

- 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 ratings than free apps.
- Free apps get more negative and neutral feedback, suggesting a wider range of opinions.
- 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 5 (most positive).
- More than half users rate **Family, Sports and Health & Fitness apps** positively. Apps for games and social media get mixed reviews, with 50 percent positive and 50 percent negative responses.

CONCLUSION

- **The average rating** we see of apps on the Play Store is **4.30**.
- Tools, entertainment, education, business, and medical are the top genres.
- **The Face book** app has the most reviews. While **family and fitness are the most** downloaded app.
- 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.

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*

