Capstone Project

Play Store App Review Analysis

Team

SAQUIB NEYAZ
RAJNI SHUKLA
PRANALI DONGRE
RAKESH BANGAR
MOOL JANGID

CONTENTS

PROBLEM STATEMENT

Google Play, formerly known as the Google Play Store and Android Market, is a digital distribution service run and developed by Google.
Android is growing as an operating system. It has taken up about 74% of the total market which is a true indicator of the large number of people using android. Our goal is to help android developers know what motivates people to download the app. It will also help to identify factors that affect a person's decision to download an app. we would like to analyze the category, reviews, price, ratings and insertions for this purpose and find out how they relate.

APPROACH

In this Project Our analysis approach is divided into two phases: Apps Analysis and Sentiment Analysis.

In the first step, we did, cleaning and analyzed various features of the Apps dataset through data exploration and visualizations.

In the second step, we loaded and cleaned the User Reviews dataset. The two datasets are then merged to visualize the composition of the total reviews and the sentiment polarity distribution.

INTRODUCTION

The google play store is one of the largest and most popular Android app stores. It has an enormous amount of data that can be used to make an optimal model. We have used a raw data set from the Google Play Store from the Kaggle website.

Mobile applications are one of the fastest-growing segments of the downloadable software application market. Out of all of the markets we choose Google to play due to its increased popularity and recent past growth. One of the main reasons is the fact that about 81% of the app are free of cost.

Hence proposing to analyze data to the developer that what customer is likely to download, which category got the maximum downloads this all plays a crucial role in app development. Generally, customers download apps depending on the number of downloads, positive reviews, negative reviews, ratings, and comments. So, in this project, we are going to help the users by categorizing positive, negative, and neutral reviews and comments on the particular.

WHY ANALYZE THE PLAY Store?

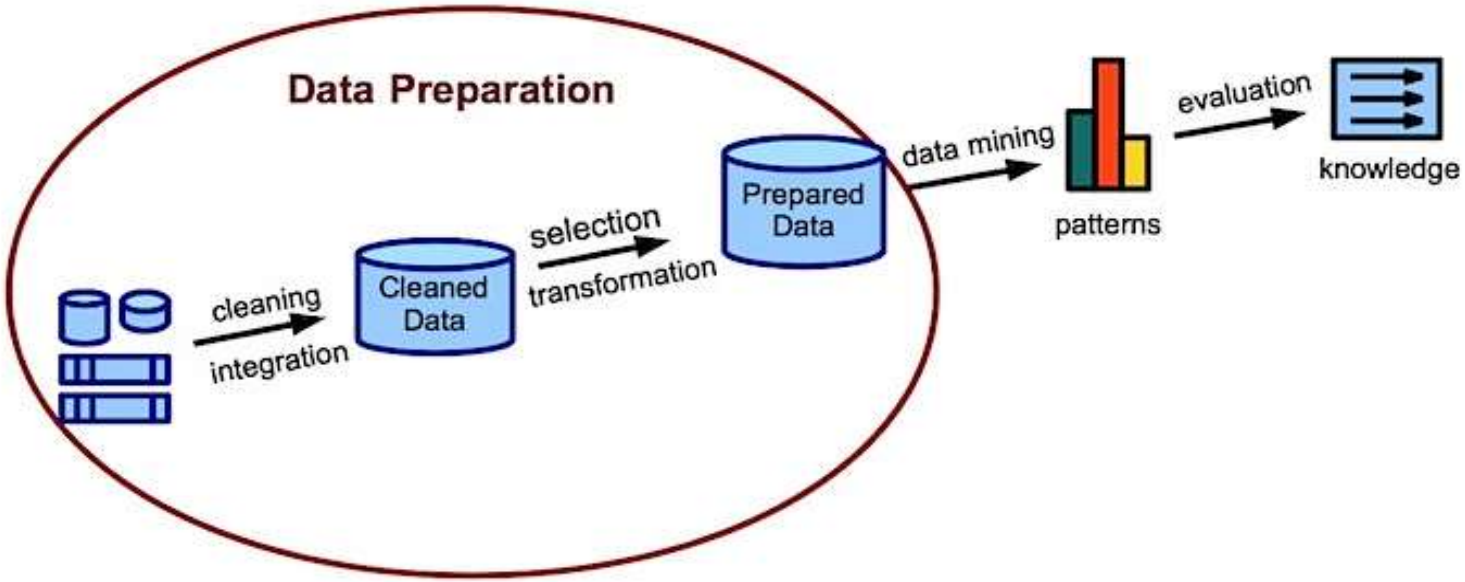
- **Android Apps comprise 75% of the Market Share.85% share in brazil,india,turkey**

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

- **What makes an App popular? Can we predict how popular it's going to be?**

  => Only way an app remains popular is by satisfying all the basic user needs an delighting them with great user experience,full of surprises and excitement.

Data Preparation and Cleaning



Data Cleaning:

Data cleaning is the process of detecting and correcting(or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate,or irrelevant parts of the data and then replacing,modifying,or deleting the dirty or coarse data.

```
#Drop NaNs and duplicates in the dataframe
apps = df.dropna()
apps = apps.drop_duplicates()
print(len(apps))

# Remove unwanted characters
chars = ['$', '+', ',']
cols = ['Installs','Price']

for col in cols:
  for char in chars:
    apps[col] = apps[col].astype(str).str.replace(char,'')

# Convert columns to numeric data type
  apps[col] = pd.to_numeric(apps[col])

# Change the size of Apps from KB to MB
apps['Size'] = apps['Size'].astype(str).str.replace('M','')
apps['Size'] = apps['Size'].astype(str).str.replace('k','e-3')

# Change the size 'Varies with device' to average app size as
reported by Google
apps['Size'] = apps['Size'].astype(str).str.replace('Varies with
 device','11.5')

apps['Size'] = pd.to_numeric(apps['Size'])
```

```
#Convert the 'Last Updated' column to Datetime object
apps['Last Updated'] = pd.to_datetime(apps['Last Updated'])
apps.sample(10)
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
4202	H Band	HEALTH_AND_FITNESS	4.0	214	8.9	10000	Free	0.00	Everyone	Health & Fitness	2018-06-13	3.0.2	4.4 and up
3610	Johny Johny Yes Papa Nursery Rhyme - offline V...	PARENTING	4.5	806	11.0	500000	Free	0.00	Everyone	Parenting	2018-08-04	4.2	4.0.3 and up
2284	Recognise Foot	MEDICAL	4.2	9	95.0	1000	Paid	7.49	Everyone	Medical	2017-09-14	1.0.3	4.1 and up
7647	Hashtags For Likes.co	SOCIAL	4.3	420	18.0	50000	Free	0.00	Everyone	Social	2016-12-19	1.1	4.0.3 and up
832	Learn languages, grammar & vocabulary with Mem...	EDUCATION	4.7	1107948	11.5	10000000	Free	0.00	Everyone	Education	2018-08-02	Varies with device	Varies with device
10145	Ez iCam	PHOTOGRAPHY	3.1	7300	46.0	1000000	Free	0.00	Teen	Photography	2018-01-12	V4.1.0	5.0 and up
1948	Zombie Hunter King	GAME	4.3	10538	50.0	1000000	Free	0.00	Mature 17+	Action	2018-08-01	1.0.8	2.3 and up
8880	Wheelie Challenge	GAME	4.6	137338	51.0	5000000	Free	0.00	Everyone	Racing	2018-07-13	1.44	4.1 and up
9114	My Ooredoo Algeria	TOOLS	4.2	3606	17.0	100000	Free	0.00	Everyone	Tools	2018-06-24	1.22.1	4.4 and up
3988	C Programming	FAMILY	4.3	22248	1.8	1000000	Free	0.00	Everyone	Education	2015-06-28	3.0	2.3 and up

DATA PREPARATION:

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data, and the combining of data sets to enrich data.

Description of data:

- showing first 6 values- df.head(6)
- ps_data = df.copy()
- finding data type of each column- ps_data.info()
- list of columns-list(ps_data.columns)
- Find out the size of play store data -ps_data.shape
- ps_data.describe()

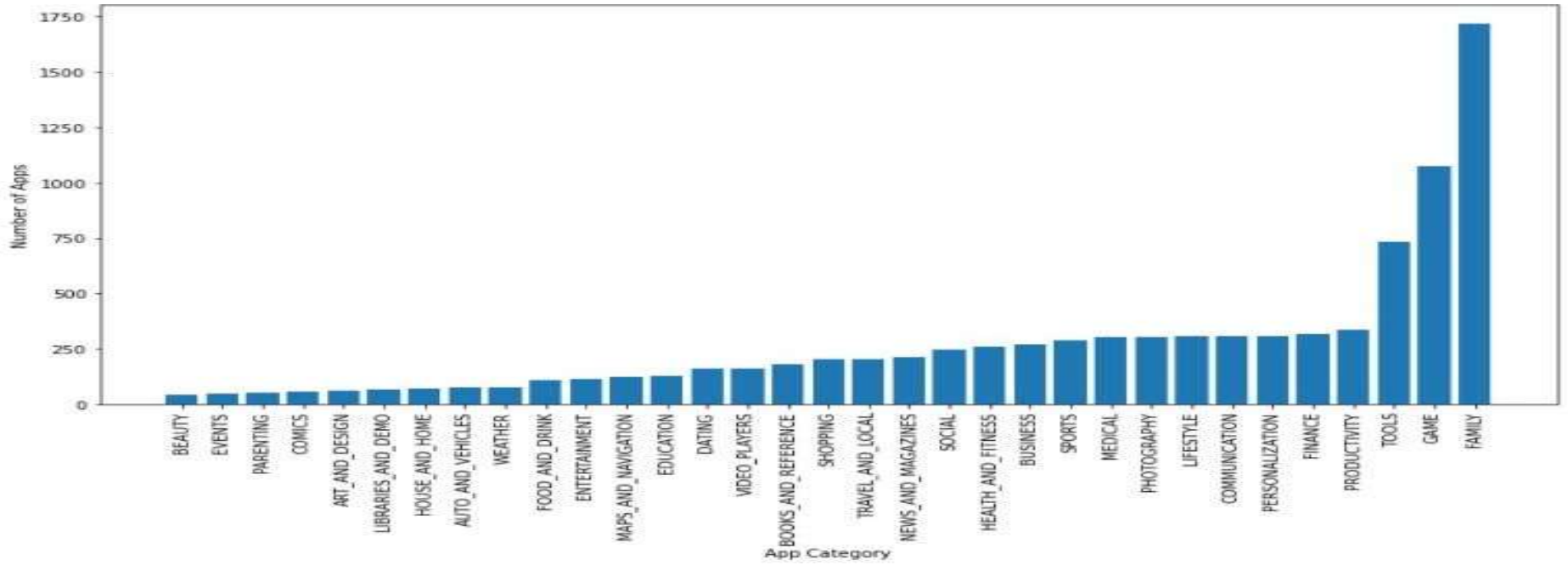
Exploratory Analysis and Visualization

- **In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.**

- **Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images.**

DATA VISUALIZATIONS

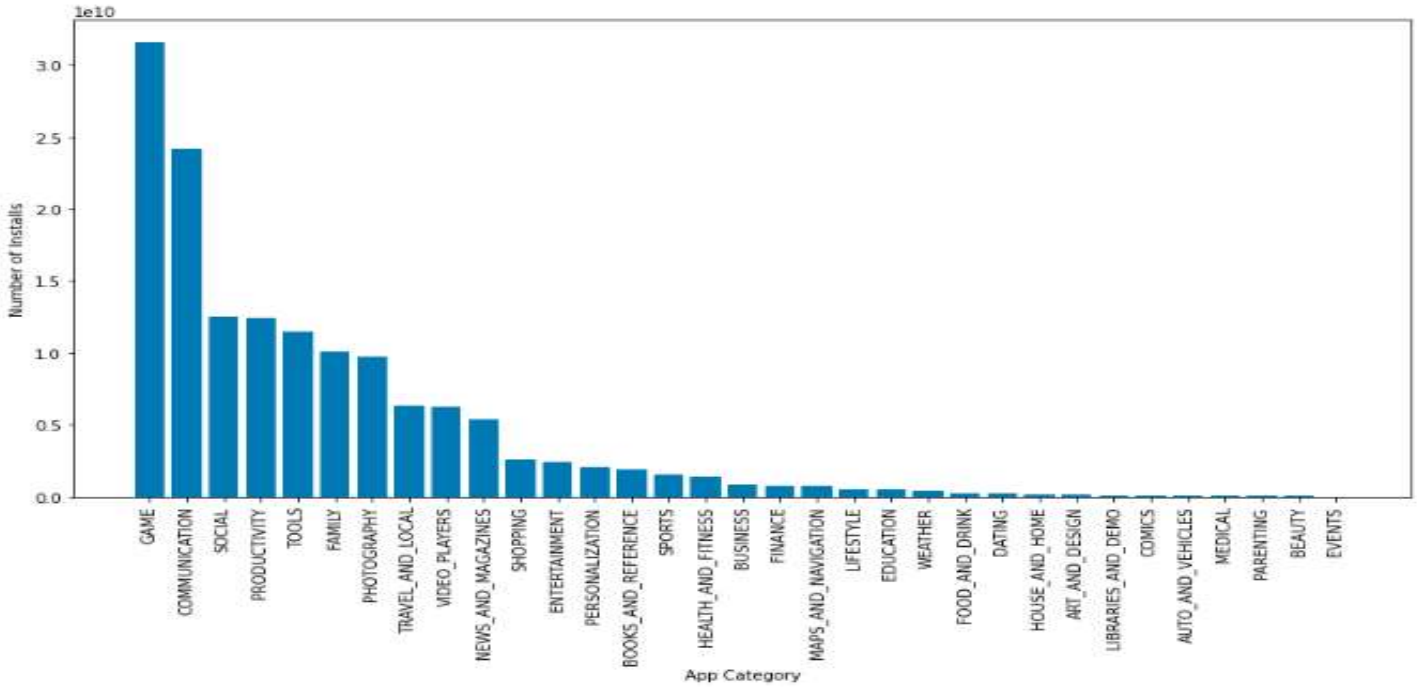
Number of Apps vs. App Category :



App Category vs Installed :

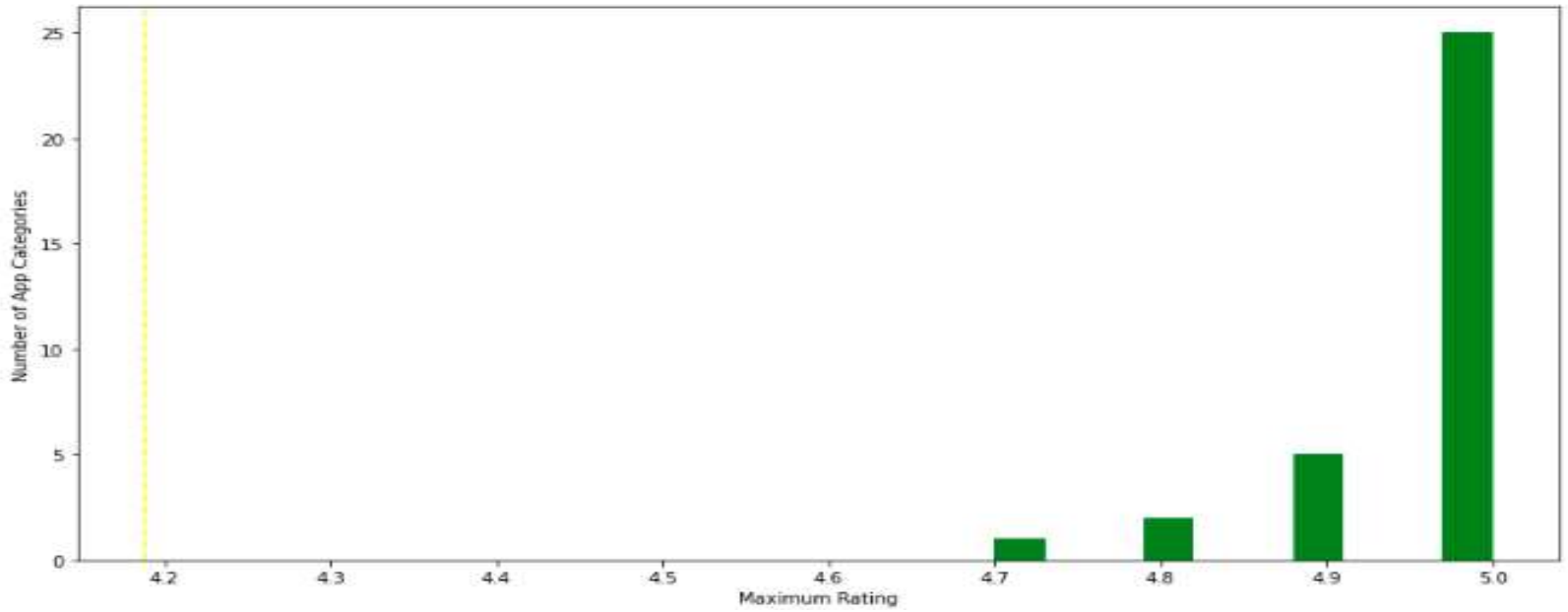
•There are all total of 33 categories in the dataset from the above output we can come to the conclusion that in the play store most of the apps are Game, Family, Communication, News & Magazines, & Tools.

•Game's and Communication apps is the most installed application and Less number of the installed application is beauty and event etc.

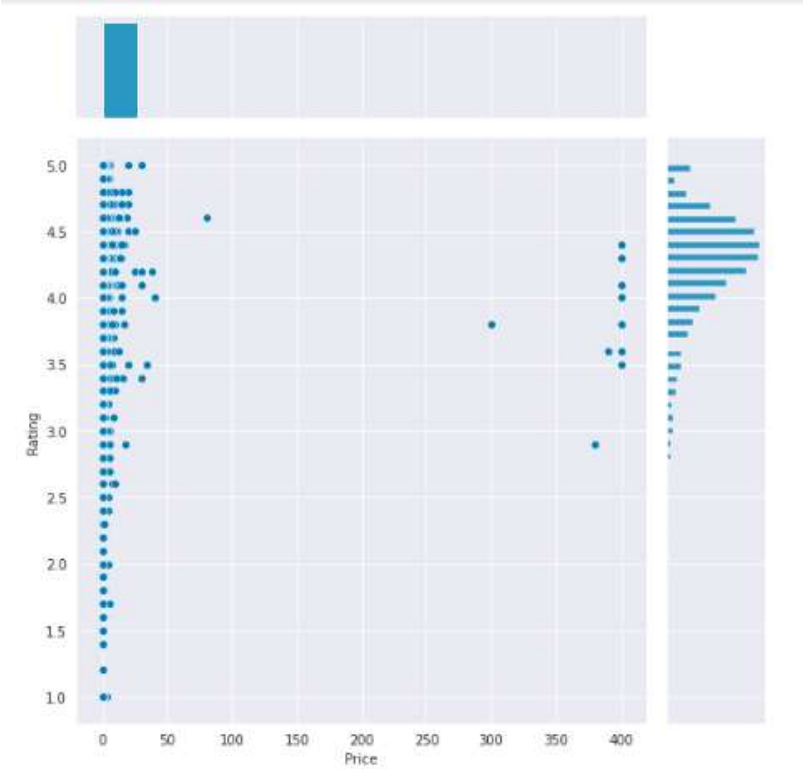


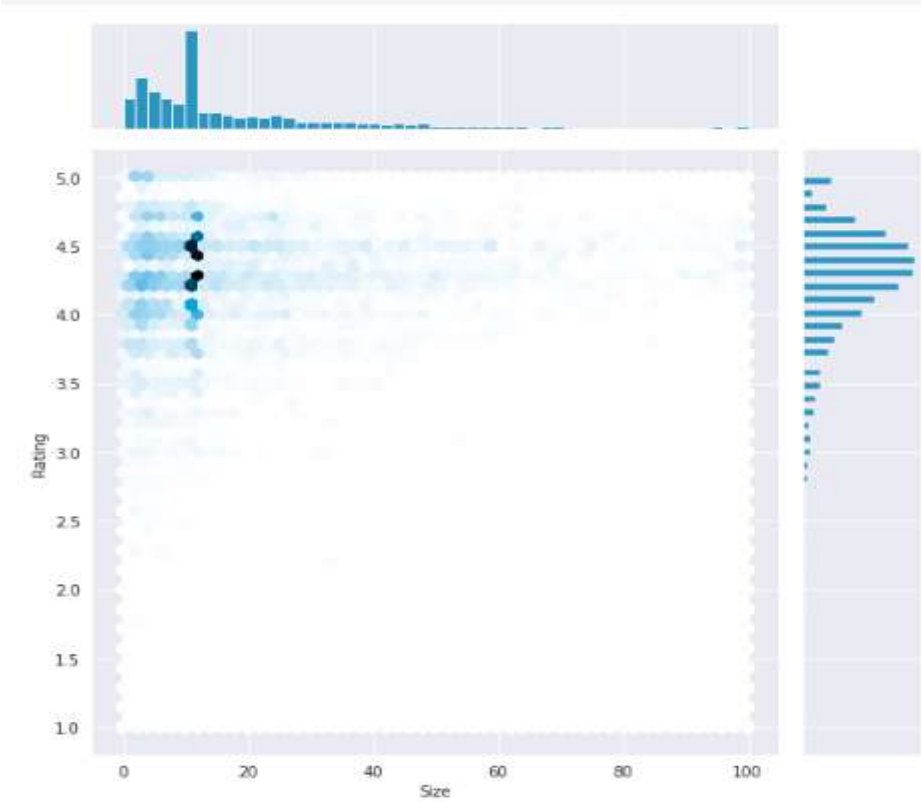
The Top Content Rating :

From the above graph, we can come to the conclusion that most of the apps in the google play store are rated between 4.9 to 5

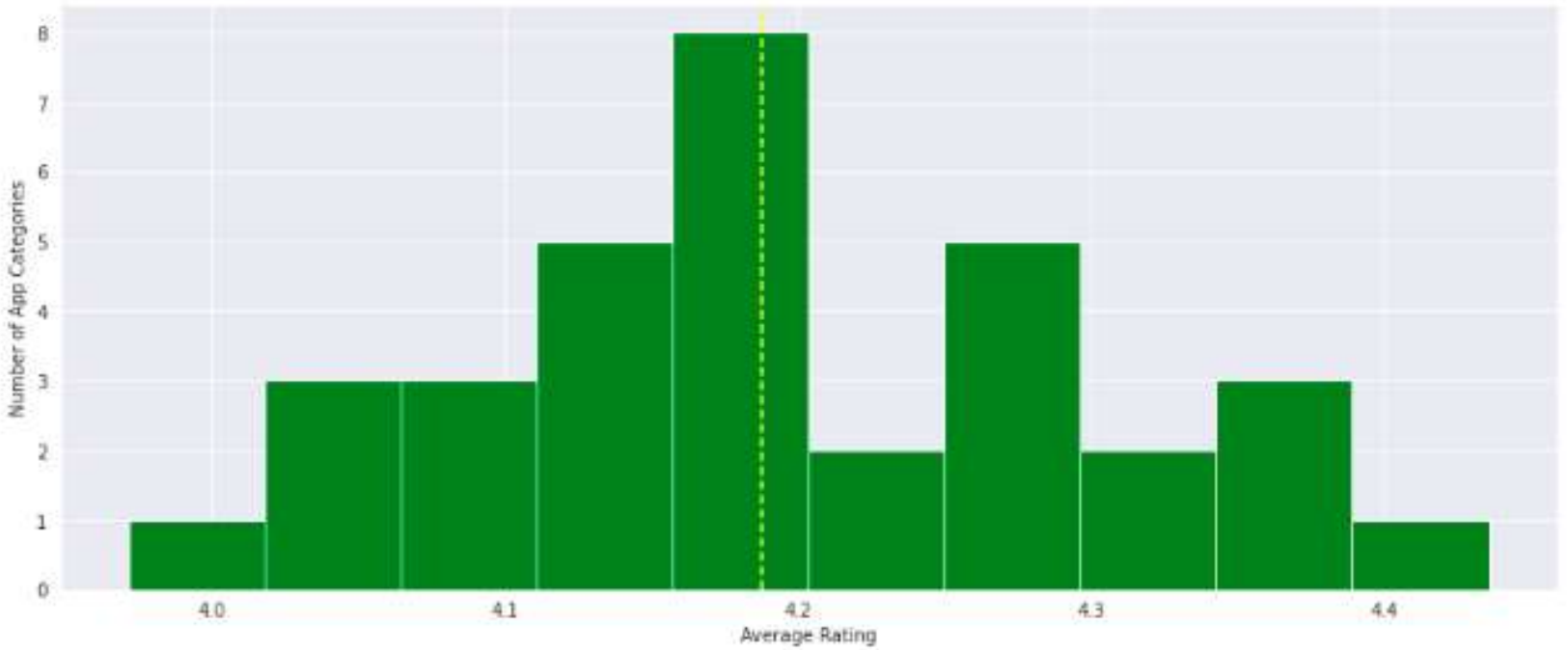


Effect of Price and Size Vs. Rating :





Number of App categories vs. Average Rating :



From the above scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of cases, it shows a proportional behavior when variance is too high or low.

Basic Observation :

Below are some observation by doing data wrangling.

Average app rating	**4.18**
Category of Apps having most number of installs	**i. Game** **ii. Communication** **iii. Social**
Top Content rating apps	**i. Art and Design** **ii. Auto and Vehicles** **iii. Beauty**
Popularity of Paid Apps vs Free Apps	**i. Free Apps - 93.12%** **ii. Paid Apps - 6.88%**
Composition Of Positive, Negative and Neutral reviews of app	**i. Positive - 63.98%** **ii. Negative - 23.88%** **iii. Neutral – 12.14%**

CONCLUSIONS

1. **Hence, from the above observations and visualizations, we can draw the following conclusions:**
2. **The dataset contains possibilities to deliver insights to understand customer demands better and thus help developers to popularize the product.**
3. **The most popular App Category is "Game".**
4. **A large number of Apps fall into "Family" Category i.e., this is the category with highest number of subsequent apps.**
5. **The total average rating of Play Store Apps is [4.18].**
6. **The App Categories with least and highest average ratings are "Dating" and "Events" respectively.**
7. **We deal with missing data and outliers, we tested some of the fundamental statistical assumptions and we even transformed categorial variables into dummy variables. That's a lot of work that python helped us make easier.**
8. **Free apps are highly popular when compared to Paid apps.**

THANK YOU