GOOGLE PLAY STORE APP'S RATINGS

EXPLORATORY DATA ANALYSIS

HRISAV BHOWMICK
A20011
STS-I PROJECT
PRAXIS BUSINESS SCHOOL

CONTENTS

INTRODUCTION

Google Play is a digital distribution service operated and developed by Google Inc. It serves as the official app store for the Android operating system, allowing users to browse and download. Applications are available through Google Play either free of charge or at a cost. They can be downloaded directly on an Android device through the mobile app or by deploying the application to a device from the Google Play website.

The aim of the analysis is to provide insights about android applications and their categories. Deeper dive in data will help to find out the factors of influences on an application. An analysis on category, price, ratings and installs would be carried out for this purpose. It would help to find out how they are inter related and also will help android developers to analyze and understand the factors behind the ratings of the apps.

DATASET

The dataset is extracted from Kaggle. It contains details of Android applications from Google Play. There are 6 includes that depict each application and an aggregate of 1796 applications. Below table consists of the column headers of the dataset and its brief description.

COLUMN HEADERS	DESCRIPTION
App	Name of the App
Category	Category of the app
Rating	Over all user rating of the app out of 5
Install	Number of user downloads for the app
Price	Cost of the App
Content Rating	Age group the app is targeted at

DESCRIPTIVE STATISTICS

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way. Some measures that are commonly used to describe a data set are measures of central tendency and measures of dispersion. Measures of central tendency include the mean, median and mode, while measures of dispersion include the standard deviation (or variance), the minimum and maximum values of the variables, quartile deviation.

```
df.describe()
```

	Rating	Installs	Prioe
oount	1708.000000	1.708000e+03	1708.000000
mean	4.210363	2.337989e+07	0.263759
std	0.473817	1.073932e+08	1.682448
min	1.000000	1.000000e+02	0.000000
25%	4.000000	5.000000e+04	0.000000
50%	4.300000	1.000000e+06	0.000000
75%	4.500000	1.000000e+07	0.000000
max	5.000000	1.000000e+09	39.990000

The *describe()* function shows some basic statistical details like percentile, mean, standard deviation, etc. of the data frame.

```
#finding mean
df.mean()
```

```
Rating      4.210363e+00
Installs    2.337989e+07
Price       2.637588e-01
dtype: float64
```

```
#finding median
df.median()
```

```
Rating           4.3
Installs    1000000.0
Price             0.0
dtype: float64
```

```
#finding mode
df.mode()
```

	App	Category	Rating	Installs	Prioe	Content Rating
0	Cardiao diagnosis (heart rate, arrhythmia)	FAMILY	4.4	1000000.0	0.0	Everyone
1	ROBLOX	NaN	NaN	NaN	NaN	NaN

The mean value for “Rating” is 4.21, “Installs” is 20 million, “Price” is $ 0.263. The mean value for “Rating” is 4.3, “Installs” is 1 million, “Price” is $ 0. The mode value for “Rating” is 4.4 and “Installs” is 1 million.

```
#finding range of Installs
df['Installs'].max()-df['Installs'].min()
```

```
999999900
```

```
#finding range of Price
df['Price'].max()-df['Price'].min()
```

```
39.99
```

```
#finding range of Rating
df['Rating'].max()-df['Rating'].min()
```

```
4.0
```

The range value for “Rating” is 4.0 and “Price” is $ 39.99.

```
#finding standard deviation
df.std()
```

```
Rating      4.738174e-01
Installs    1.073932e+08
Price       1.682448e+00
dtype: float64
```

```
#finding variance
df.var()
```

```
Rating      2.245030e-01
Installs    1.153330e+16
Price       2.830632e+00
dtype: float64
```

The standard deviation for "Rating" is 0.473 and "Price" is 1.68. Standard deviation tells how spread out the data is from the mean. The variance for "Rating" is 0.224 and "Price" is $ 2.83.

```
# First quartile (Q1) of Installs
Q1 = np.percentile(df['Installs'], 25, interpolation = 'midpoint')

# Third quartile (Q3) of Installs
Q3 = np.percentile(df['Installs'], 75, interpolation = 'midpoint')

# Interquartile range (IQR) of Installs
IQR = Q3 - Q1

IQR
```

```
9950000.0
```

```
# First quartile (Q1) of Price
Q1 = np.percentile(df['Price'], 25, interpolation = 'midpoint')

# Third quartile (Q3) of Price
Q3 = np.percentile(df['Price'], 75, interpolation = 'midpoint')

# Interquartile range (IQR) of Price
IQR = Q3 - Q1

IQR
```

```
0.0
```

```
# First quartile (Q1) of Rating
Q1 = np.percentile(df['Rating'], 25, interpolation = 'midpoint')

# Third quartile (Q3) of Rating
Q3 = np.percentile(df['Rating'], 75, interpolation = 'midpoint')

# Interquartile range (IQR) of Rating
IQR = Q3 - Q1

IQR
```

```
0.5
```

The Inter Quartile Range for "Installs" is 9.95 million and "Rating" is 0.5. The interquartile range, which tells us how far apart the first and third quartile are. It indicates how spread out the middle 50% of our set of data is.

```
#finding kurtosis
df.kurtosis()
```

```
Rating        6.413741
Installs     61.424869
Price       230.136725
dtype: float64
```

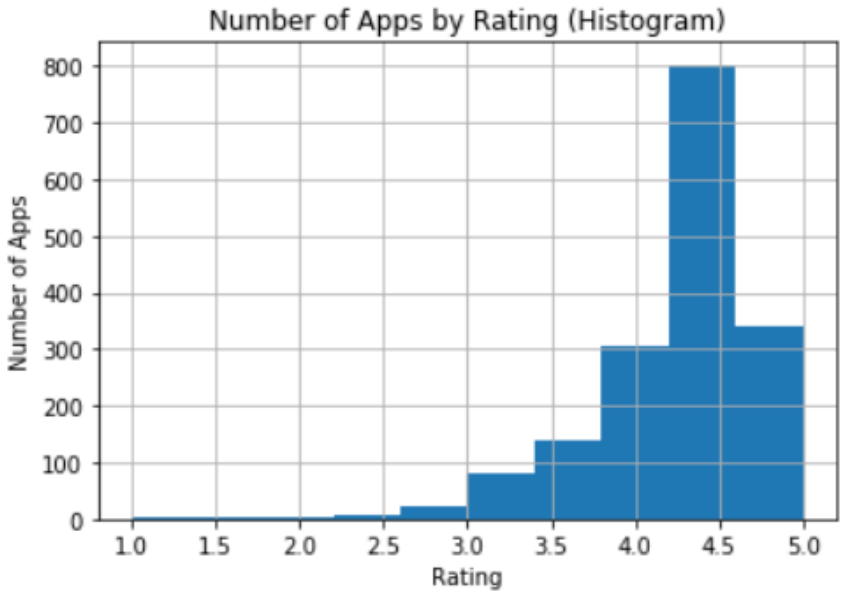
```
#finding skewness on columns
df.skew(axis = 0)
```

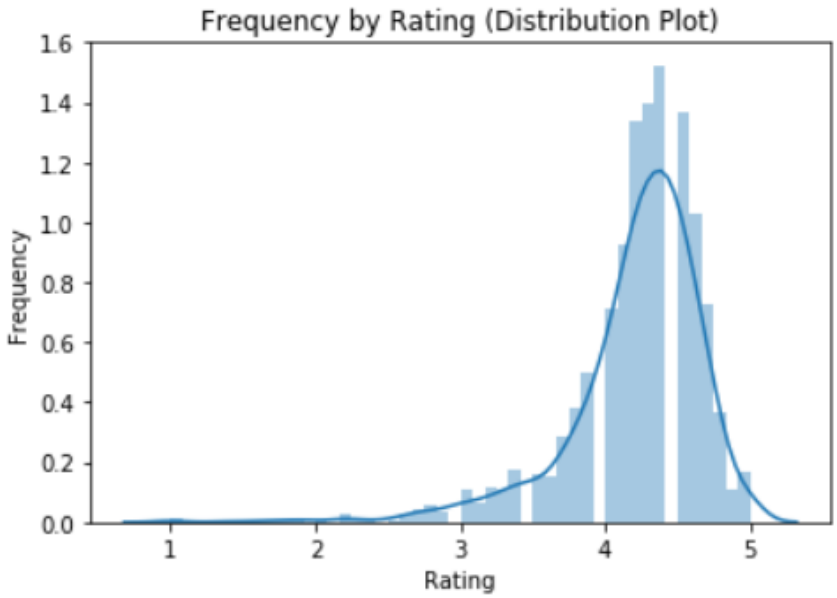
```
Rating      -1.906738
Installs     7.552619
Price       12.711100
dtype: float64
```

'Rating' is Left Skewed or Negatively Skewed. Kurtosis is around 6.4 for 'Rating' that means data is above normal distribution, which also means it has large outliers.

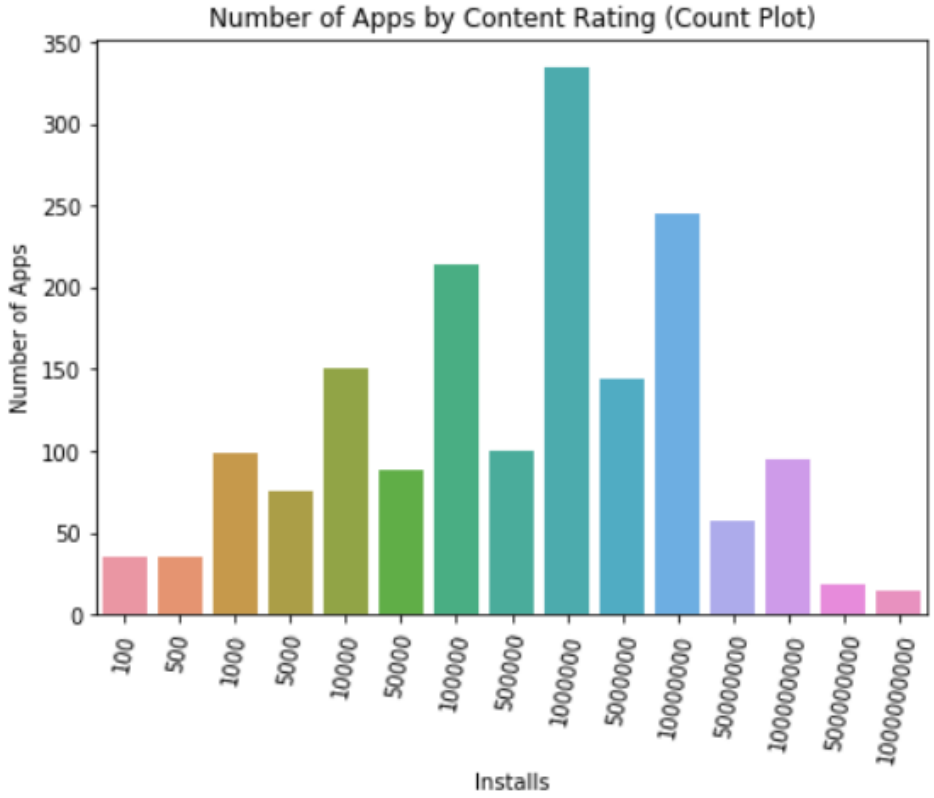
'Installs' is Positive Skewed with Kurtosis around 61.4. 'Price' is also Positive Skewed with Kurtosis around 230.

GRAPHICAL REPRESENTATIONS

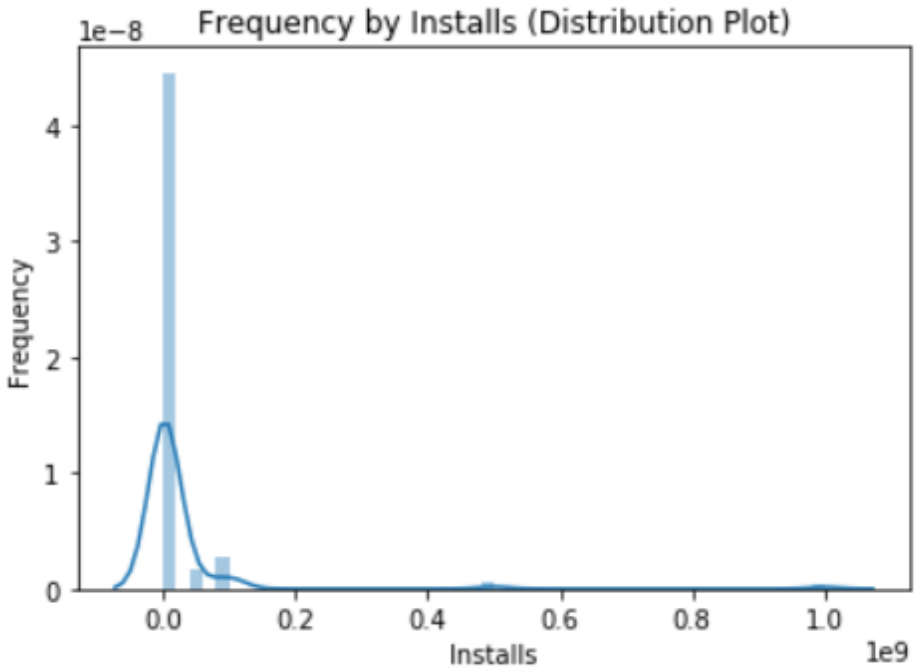




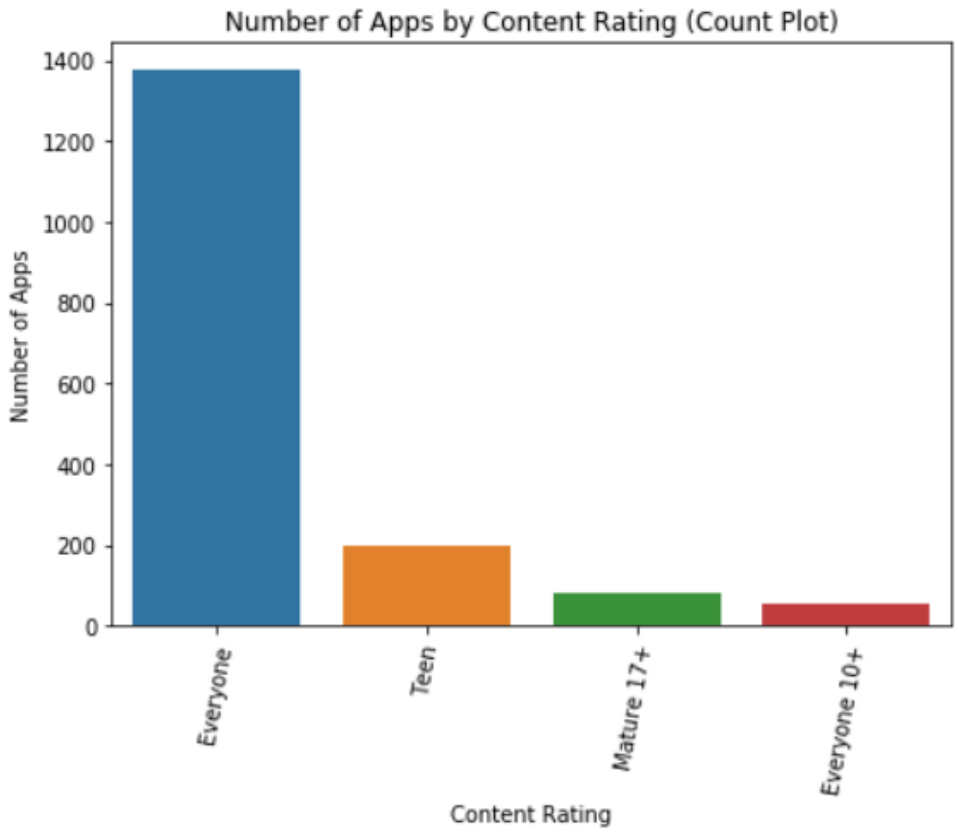
Rating has negatively skewed data with mode value of 4.3.

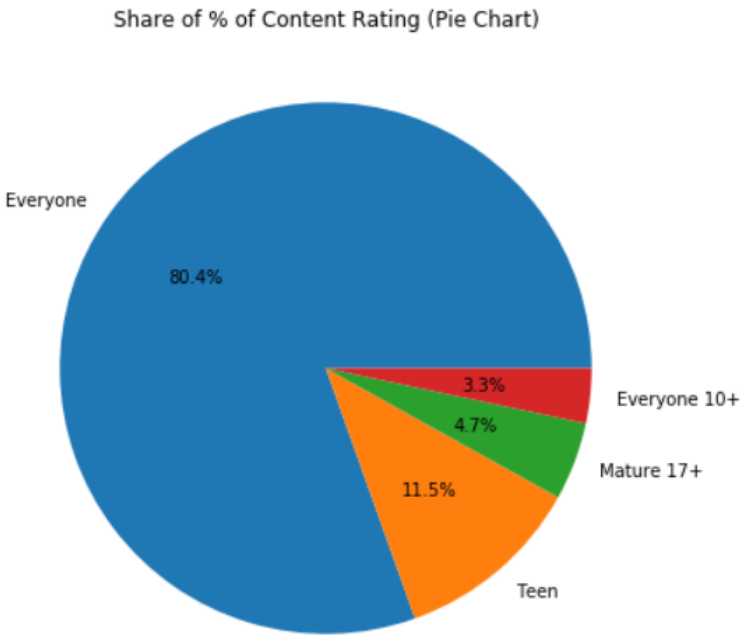


The Installs category of 1,000,000+ has the highest number of apps, whereas, the Installs category of 1,000,000,000+ has the lowest number of apps.

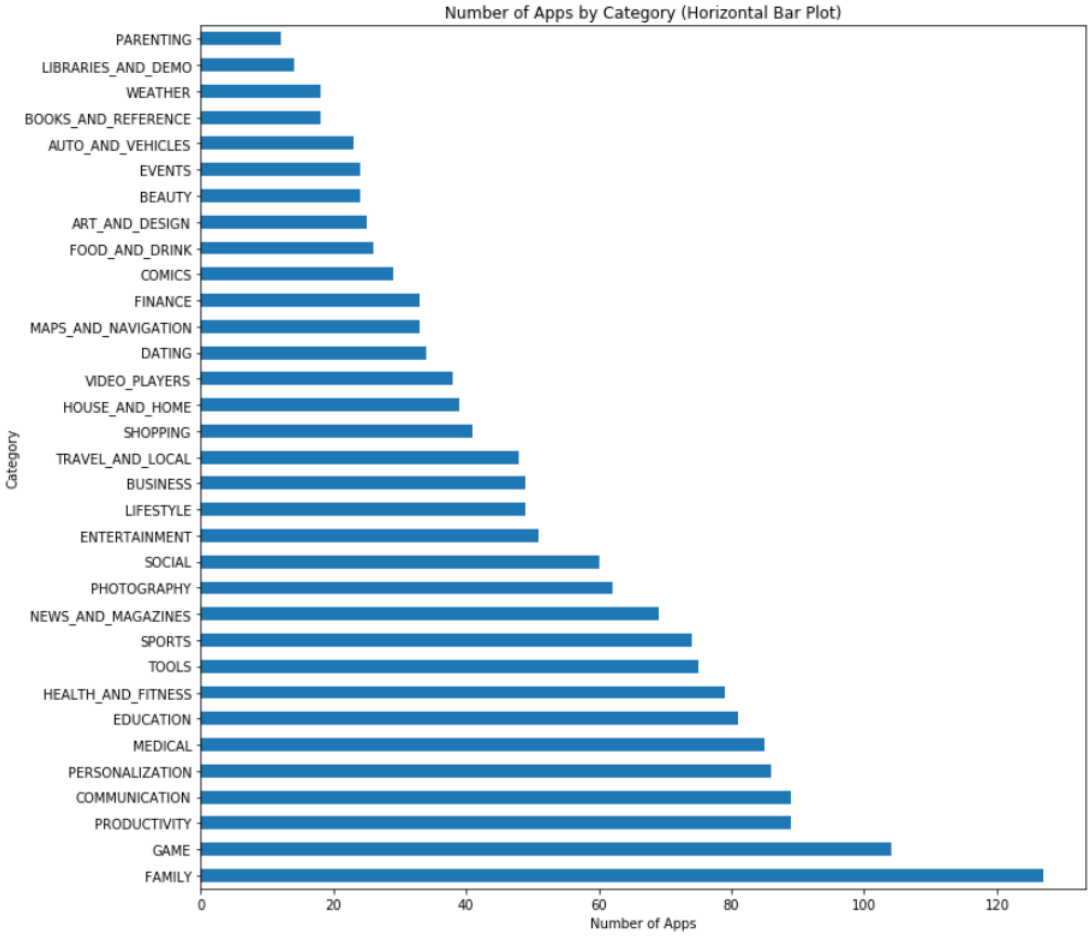


Installs has positively skewed data.

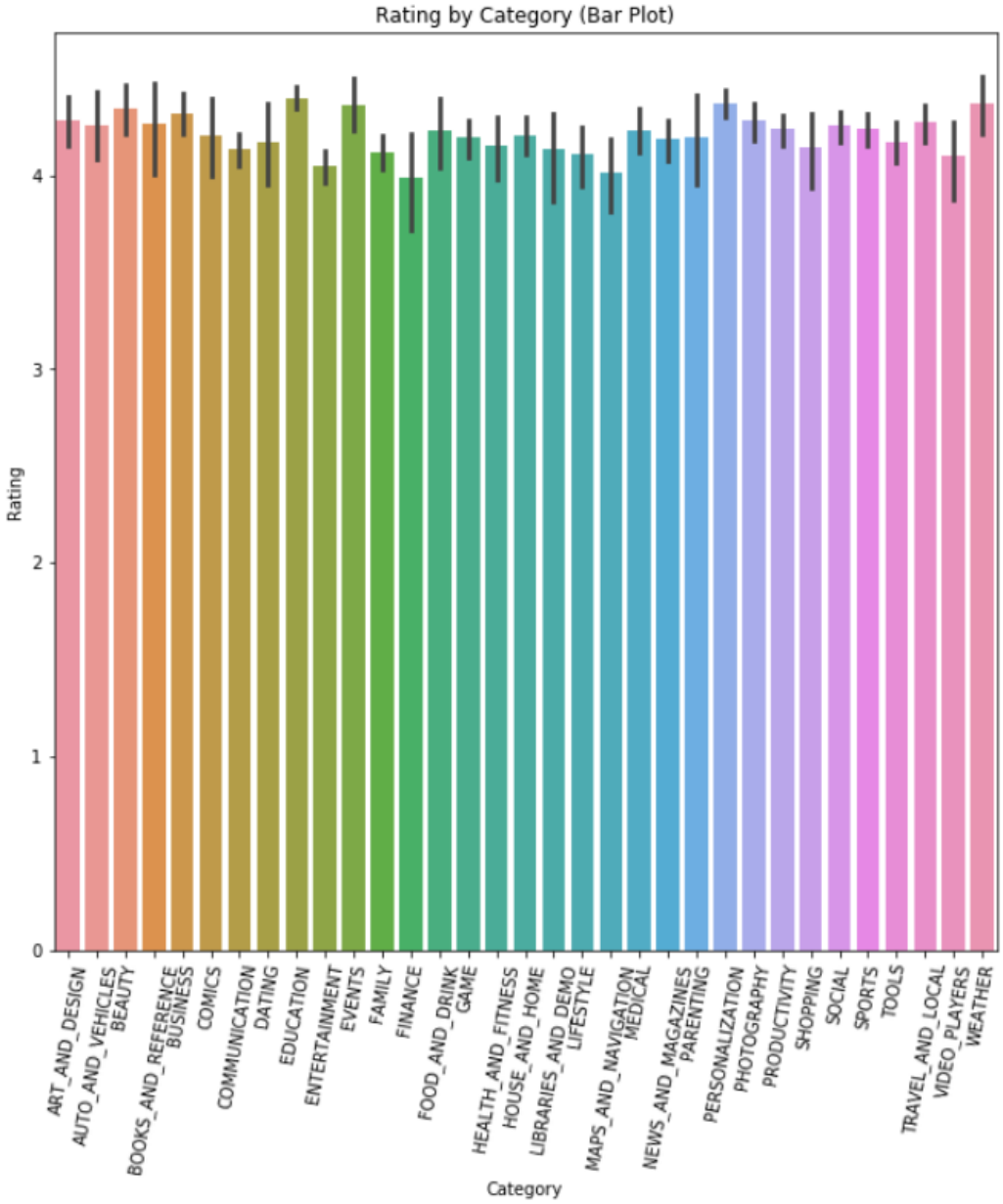




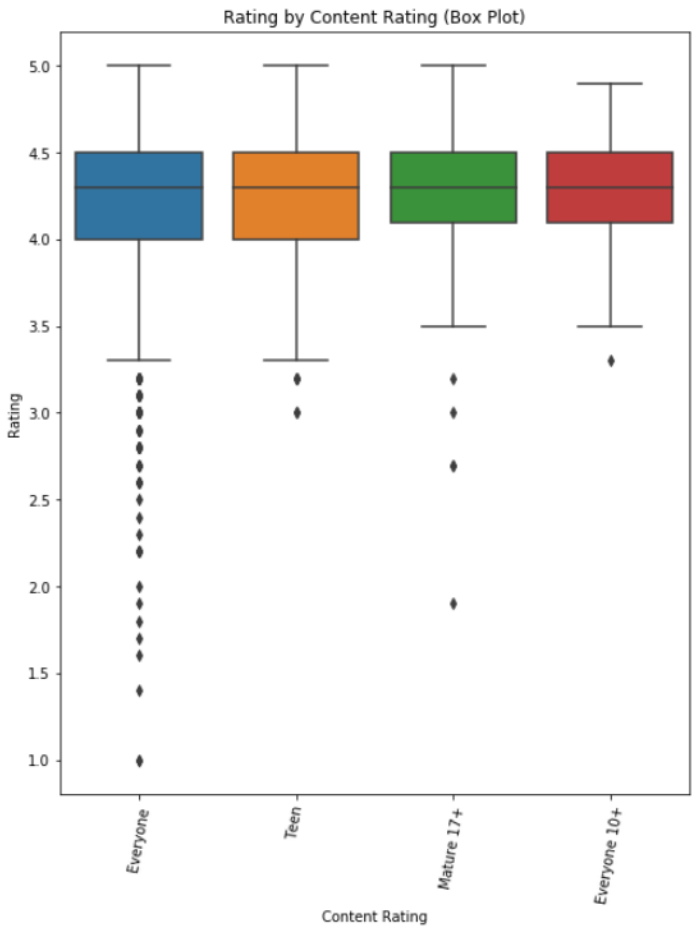
Content Rating category 'EVERYONE' has the maximum share percentage of around 80.4%, whereas, Content Rating category 'EVERYONE 10+' has the least share percentage of 3.3%. Number of apps under 'EVERYONE' is 1374.



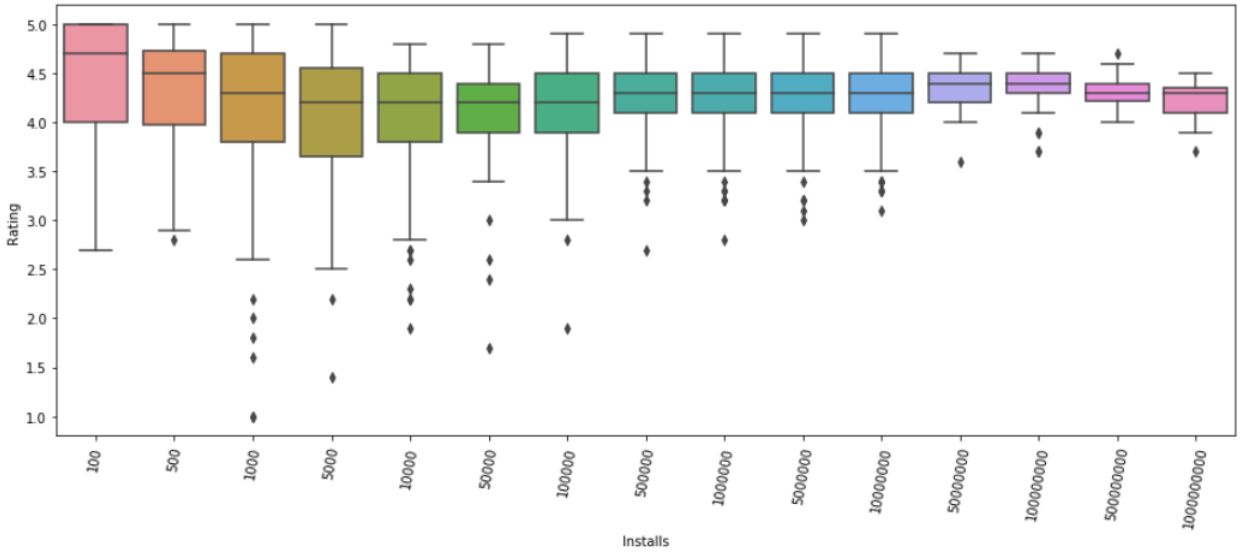
The horizontal bar plot shows that the maximum number of apps belong to the Family Category, followed by Game category. The least number of apps belong to the Parenting category.



'EDUCATION' category has the highest mean rating followed by 'WEATHER' and 'PERSONALIZATION' categories. Meanwhile, 'FINANCE' category has the lowest mean rating.



Inter quartile range for Everyone and Teen categories is greater than Mature 17+ and Everyone 10+ categories. It means Everyone and Teen has more dispersed data.



The apps in the category of 100+ installs has the highest rating median value.

CONCLUSION

The project was started from scratch where the dataset which was taken was totally raw. A lot of cleaning was done to put it out in a presentable form. Missing values were also removed during this process.

The motive during the project was to analyze the data and provide insights regarding them, which would in turn help the android developers to understand the factors behind the ratings of any app.

The costliest app was found out to be “LTC AS Legal”. Majority of the apps were found to be lying in the region of 1 million plus installs. Content rating category “Everyone” has the maximum share percentage.

The most number of apps belong to the Family category, while Education category has the highest mean rating.