SUMMARY

ON

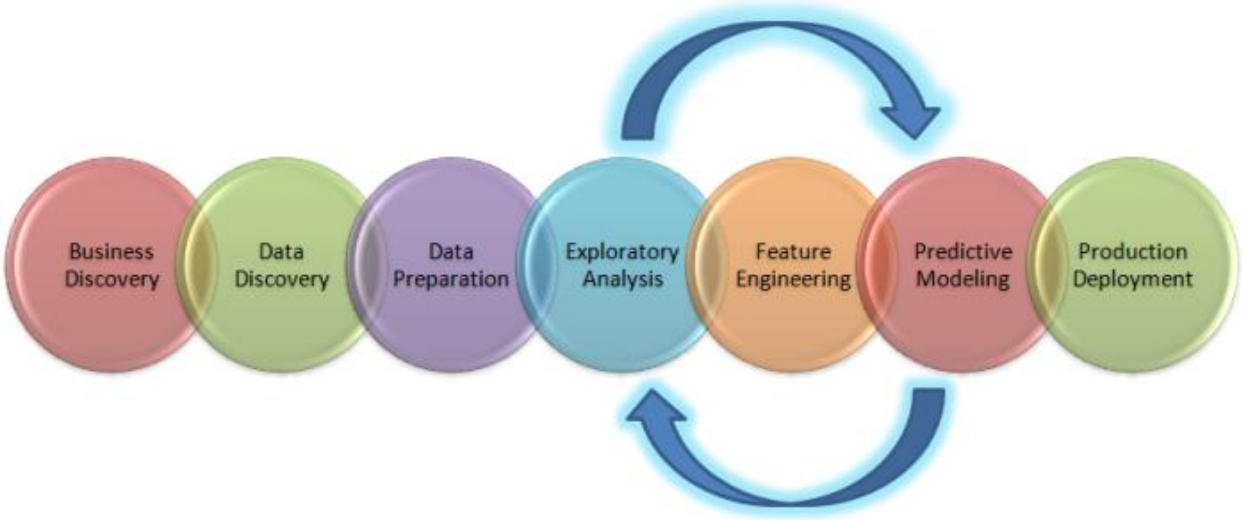
PLAY STORE APP REVIEW ANALYSIS(EDA)

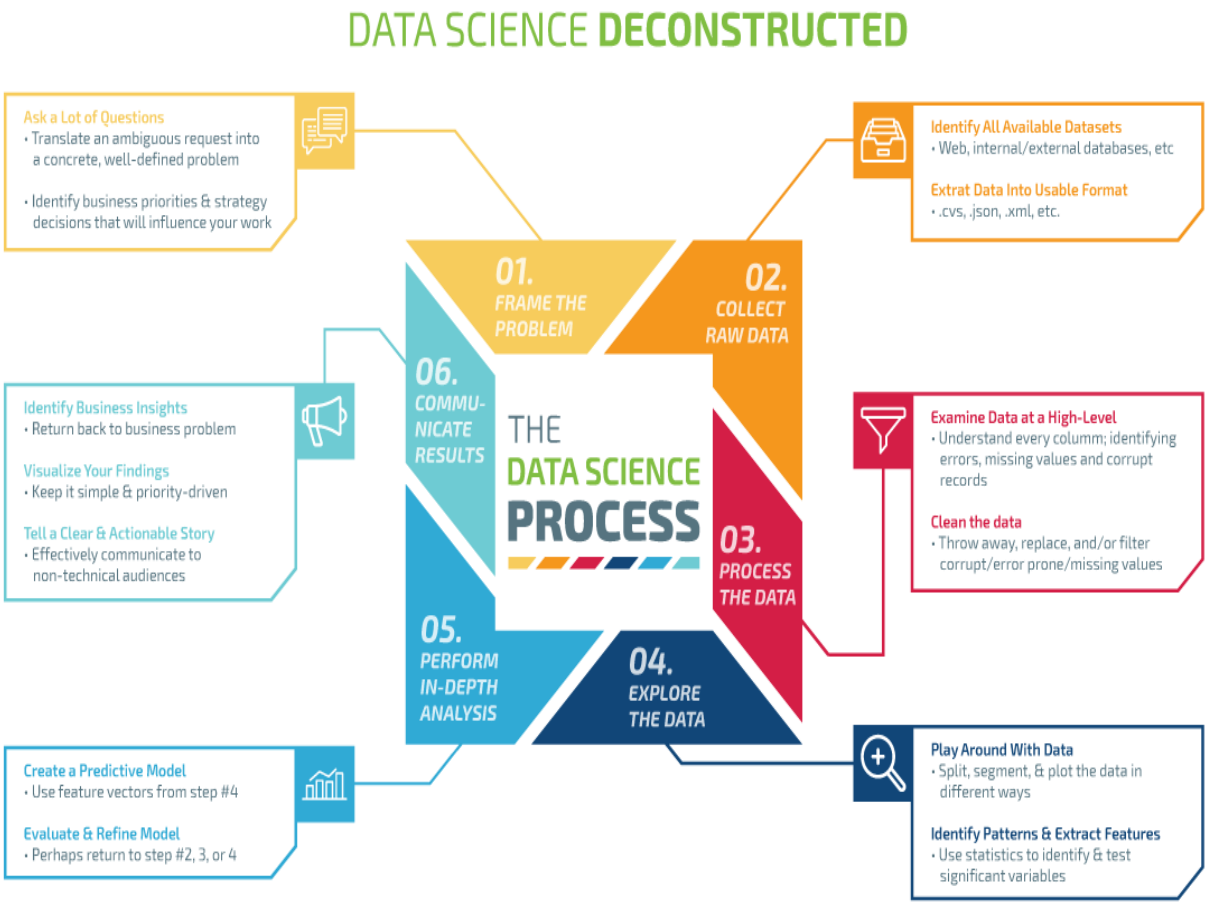
ALMABETTER CAPSTONE PROJECT

-By Yamini Peddireddi

1: Introduction

1.1. Data science problem solving





2: Methodology/Approach in solving

2.1. Problem Statement

- The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.
- Each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps.
- Explore and analyze the data to discover key factors responsible for app engagement and success.

2.2. Steps to problem solving

According to the problem statement, it is clear that we need find the success factors which is making app business grow. Here there are two datasets given for making analysis and finding insights from them.

The steps in solving this problem or the steps involved in analysis are the following:

- Understanding problem statement
- Viewing the data
- Data understanding and summary
- Missing value analysis
- Data Cleaning and transformation
- Feature engineering
- Outliers' detection
- Correlation plots
- Feature important variables
- Data Visualization
- Some case studies

2.3. Challenges

The challenges in creating the model are

- Data collection
- Data limitation
- Privacy and security
- Data manipulation
- Missing values or vague data

- Chance of misinterpretations if data is modified a lot

3: Data understanding and analysis

3.1. Description of the data

```
Store_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10841 non-null  object 
 1   Category        10841 non-null  object 
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object 
 4   Size            10841 non-null  object 
 5   Installs        10841 non-null  object 
 6   Type            10840 non-null  object 
 7   Price           10841 non-null  object 
 8   Content Rating  10840 non-null  object 
 9   Genres          10841 non-null  object 
 10  Last Updated    10841 non-null  object 
 11  Current Ver     10833 non-null  object 
 12  Android Ver     10838 non-null  object 
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

```
Store_df.describe()
```

	Rating
count	9367.000000
mean	4.193338
std	0.537431
min	1.000000
25%	4.000000
50%	4.300000
75%	4.500000
max	19.000000

```
Users_df.describe()
```

	Sentiment_Polarity	Sentiment_Subjectivity
count	37427.000000	37427.000000
mean	0.182171	0.492770
std	0.351318	0.259904
min	-1.000000	0.000000
25%	0.000000	0.357143
50%	0.150000	0.514286
75%	0.400000	0.650000
max	1.000000	1.000000

3.2. Features in the data

- There are two datasets given for analysis in which play store data tells about the apps and its related success factors with their updated details whereas user reviews dataset only tells about the users sentiment with their reviews on different apps.
- There are 10841 rows,13 columns in Play store data and there are 64295 rows, 5 columns in the User reviews data.
- When I check for datatypes in play store data it says 11 object variables and 1 float variable which is not how it has to be because there are some variables which are categorized as object instead it has to be in float and vice-versa.
- In User reviews dataset there are 3 object and 2 float variables.

3.3. Data cleaning

- There is a lot of data cleaning and pre-processing work to be done before making any analysis on play store data as the raw data is not so good for analysis.
- Firstly, when I checked for datatypes they do not seem to be the correct datatypes for the variables. So I have changed them after converted the values in them to proper format.
- For example, the size and price columns cannot be object variables. So converted them to numeric by making required changes so that the data is not distorted. In price column there is dollar sign attached to the values so I have removed punctuations(,) and currency symbol or notation. I have renamed the column as Price($) and size contains

'MB','KB' so I have created new column with units(MB,KB) and size values as another column.

- Now all the features can be used for further analysis.

3.5. Feature engineering and selection

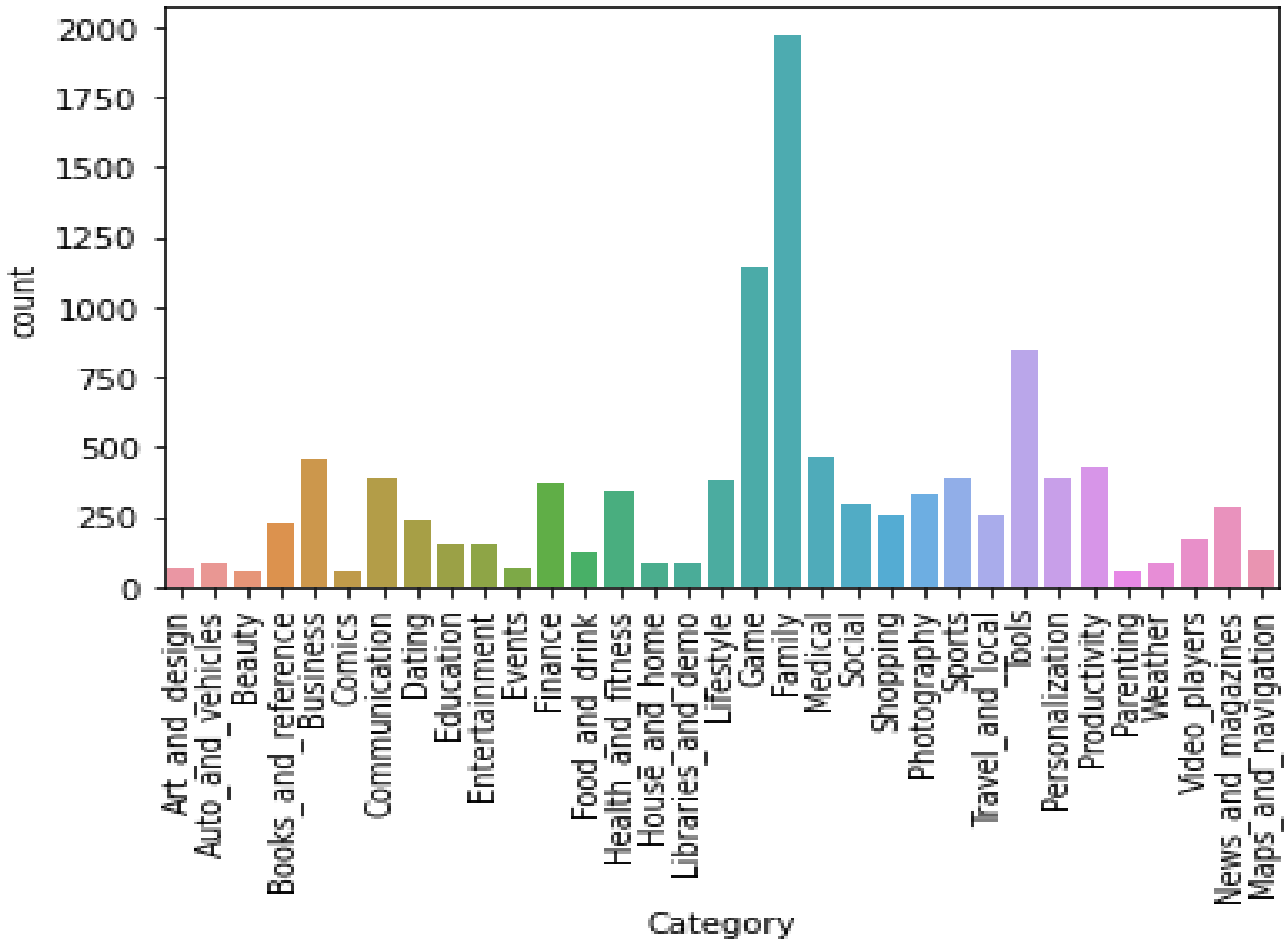
- New features like Size (MB, KB), Size values (actual given), Size (all in KB unit), Last updated month, last updated year, last updated quarter have been introduced for better analysis of Play store data.
- App category data values are capitalized for better representation and Installs column is also renamed as Installs (+) after removing the punctuations from it.
- In Size column there are some rows which says that the app size varies with device so that rows have replaced with 11.5M which is the average size of the app in the android device as play store is only contained in android devices.
- There is not much changes done in User reviews dataset as it contains the reviews and sentiment scores. I have not applied NLP techniques here because as it only analysis we can read the reviews this way better than systems way of removing stop words and punctuations. So, the User reviews data is ready to use for analysis and finding insights from them.
- The below is the description of play store data after data cleaning, transformation and feature engineering.

```
Store_df.describe()
```

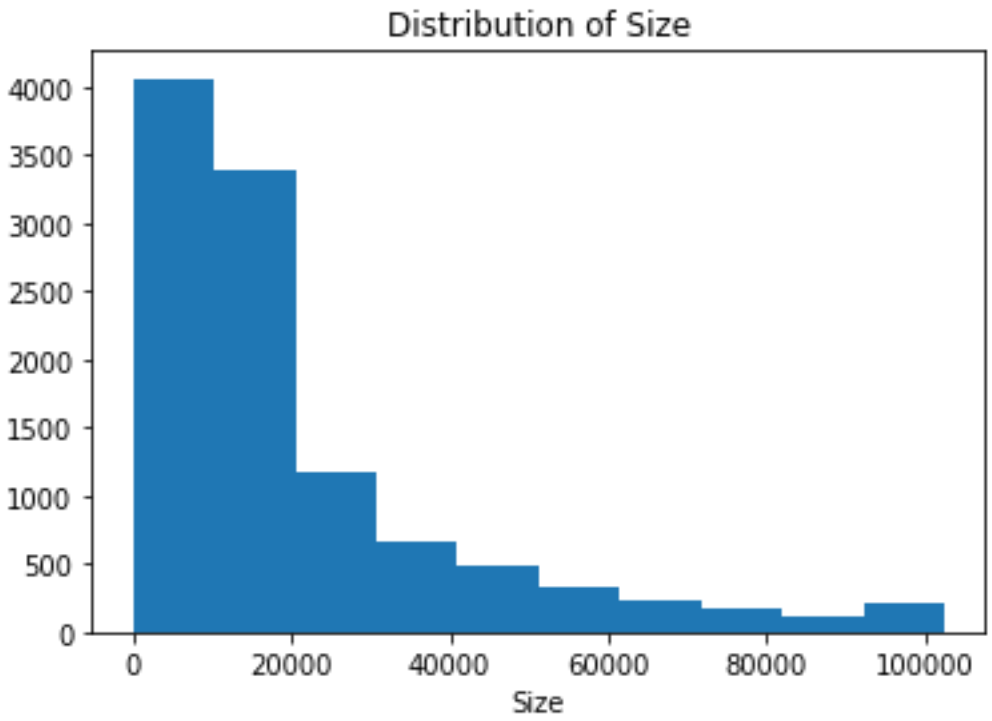
	Rating	Reviews	Size	Installs(+)	Price($)	Size(Values as given)	Last_updated_day
count	10841.000000	1.084100e+04	10841.000000	1.084100e+04	10841.000000	10841.000000	10841.000000
mean	4.206485	4.441136e+05	20426.899908	1.546291e+07	1.027273	33.039627	15.609353
std	0.480321	2.927628e+06	21569.790208	8.502557e+07	15.948971	91.283080	9.561235
min	1.000000	0.000000e+00	1.000000	0.000000e+00	0.000000	1.000000	1.000000
25%	4.100000	3.800000e+01	6041.600000	1.000000e+03	0.000000	6.800000	6.000000
50%	4.300000	2.094000e+03	11776.000000	1.000000e+05	0.000000	11.500000	16.000000
75%	4.500000	5.476800e+04	26624.000000	5.000000e+06	0.000000	28.000000	24.000000
max	5.000000	7.815831e+07	102400.000000	1.000000e+09	400.000000	1020.000000	31.000000

4: Data visualization

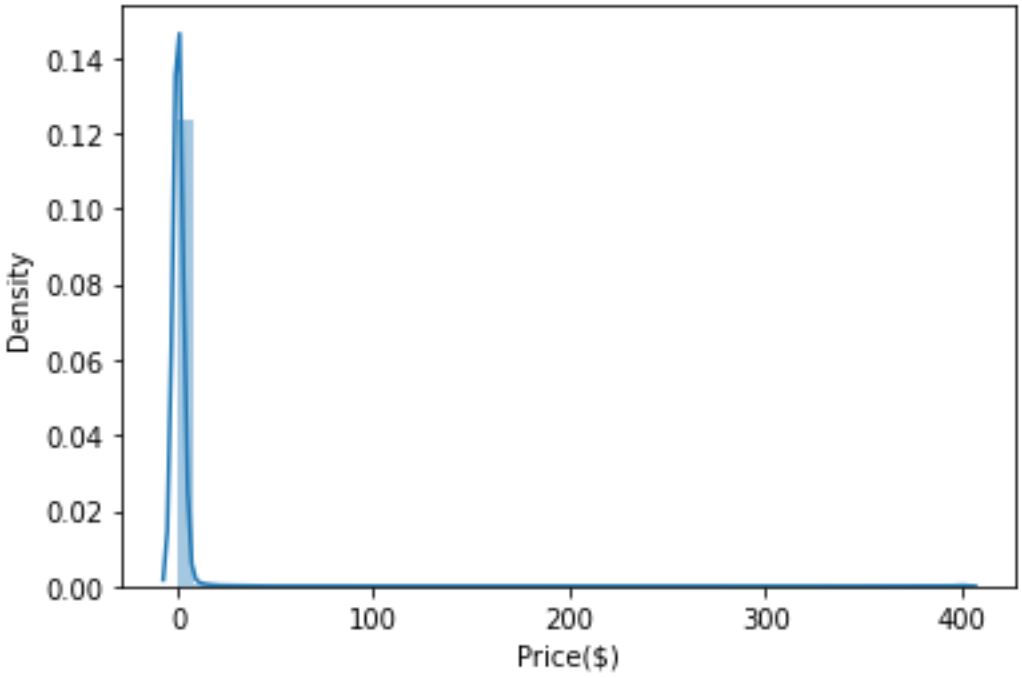
4.1. Univariate analysis



- There are more apps of Category 'Family' then 'Game' and next 'Tools'. These are the three category apps which are more in the Play Store data and occupy more space in Play Store.
- The Family apps are app. 2000, Game apps app.1200 and Tools category apps are app.850. All the other apps of each category are less than 500.



- Most of the app size is in between 0-20000 KB. There are even apps with greater size than normal. They require up to 1,00,000 KB which takes more space in a system.



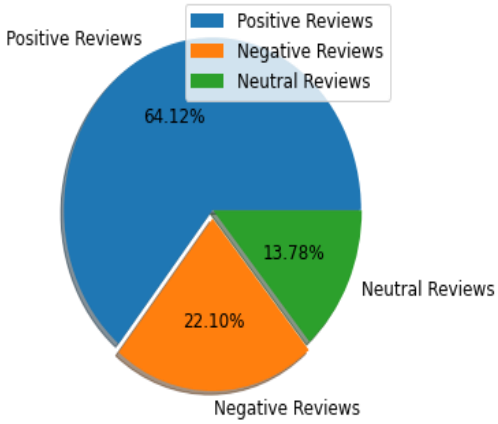
```
#skewness and kurtosis
print("Skewness: %f" % Store_df['Price($)'].skew())
print("Kurtosis: %f" % Store_df['Price($)'].kurt())
```

```
Skewness: 23.708490
Kurtosis: 578.196704
```

- From the above dist plot it can be said that there are more free apps which require no money to purchase or install them. They are free to use and maybe there are some clauses like after using for some period it has to be purchased or something.
- There are even the costly apps which are more than 400$. Very costly!!!
- The skewness indicates it is right-tail distributed because of some costly apps present in the play store. The Kurtosis is much higher than the normal because of the peakedness or consistency of many free apps.

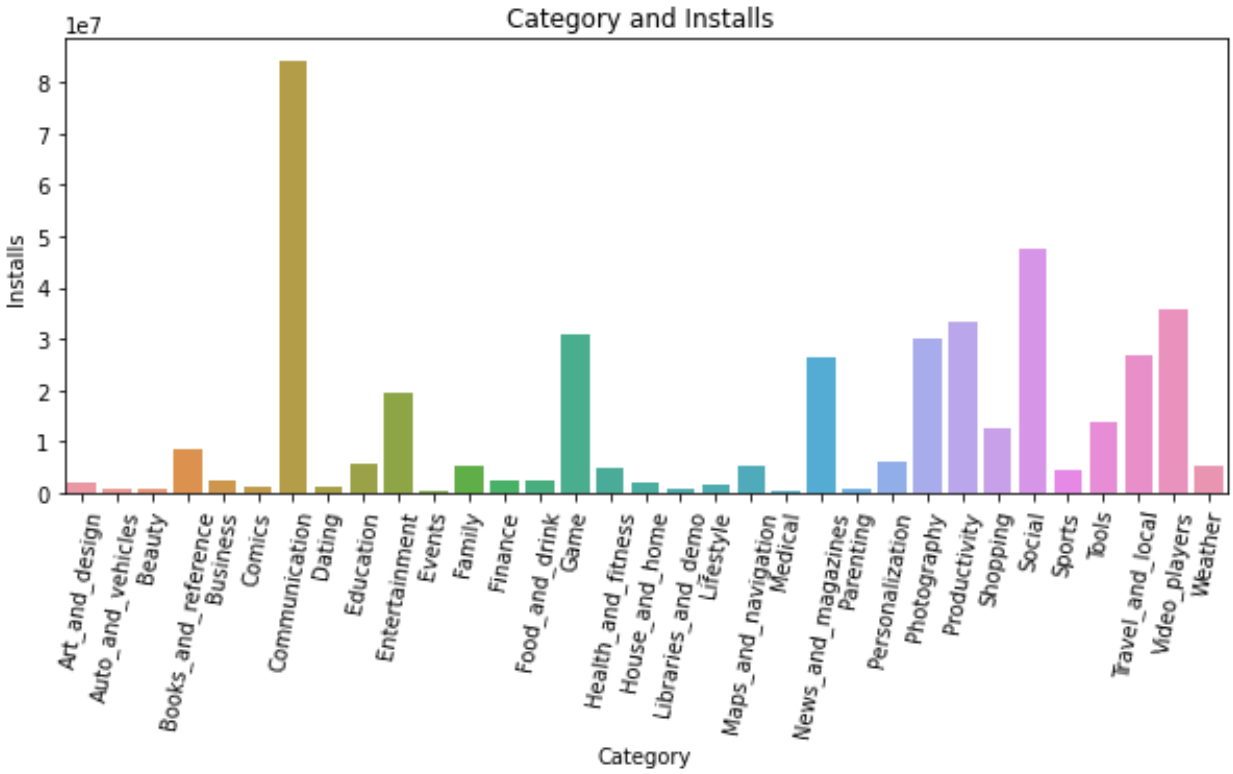
- From the above Word Cloud, it can be seen that game size is bigger than any other, next comes good, app and so on. It can also be said that there are positive words can be seen which indicates there are lot of apps which are given positive reviews by the Users.
- The most commonly used words from Users Translated reviews says that they have used more positive words to say about their app experience and which says it has good app engagement by its users. So, this definitely talks about good user experience and good word of mouth increases the publicity for the app then it impacts the success rate of the respective apps.

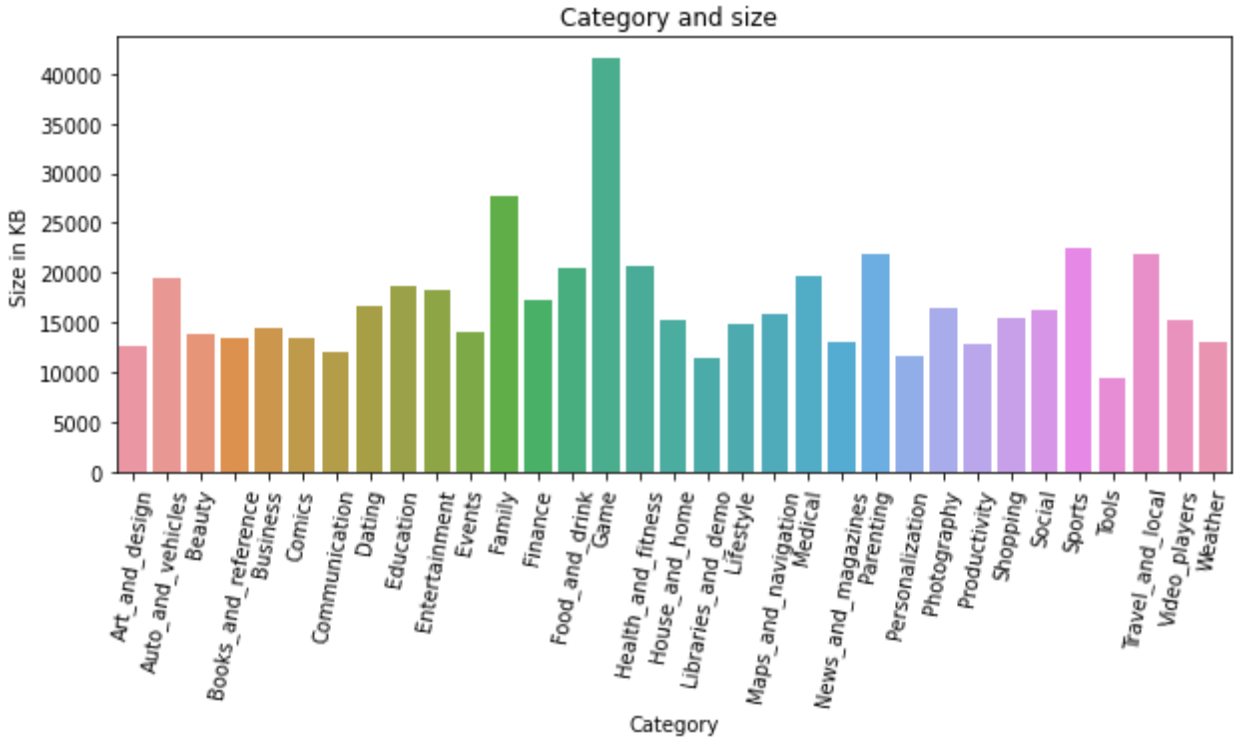
A Pie Chart Representing Percentage of Sentiments for all the apps in Users Reviews dataset



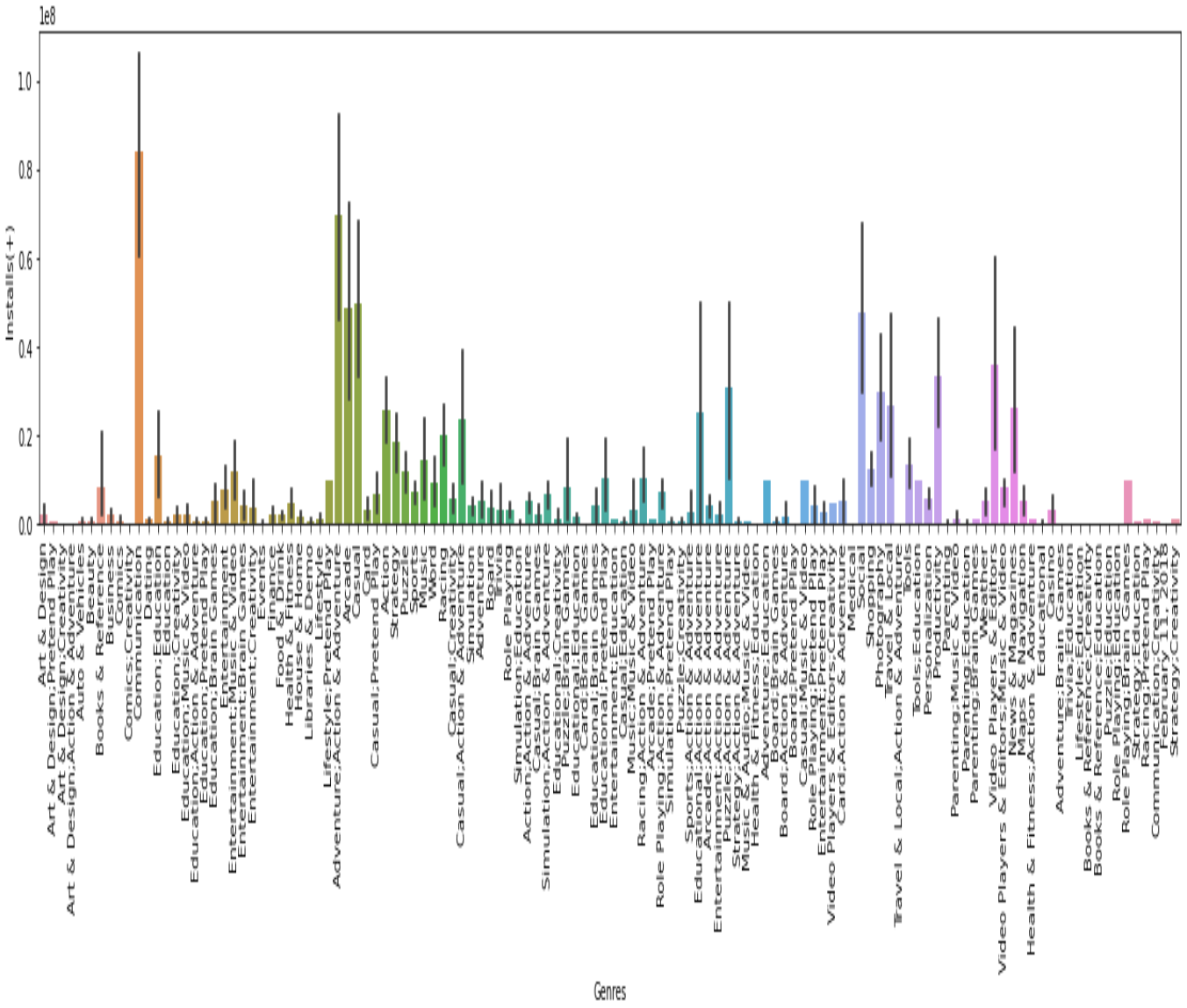
- The Users sentiment dataset tells the same that there are more positive reviews for the apps than the negative and neutral reviews for the play store data.
- This sentiment tells that there is a good user interaction with apps and good app engagement with the apps for the users. So, it has positive impact for the success of the apps in the play store.

4.2. Bivariate analysis

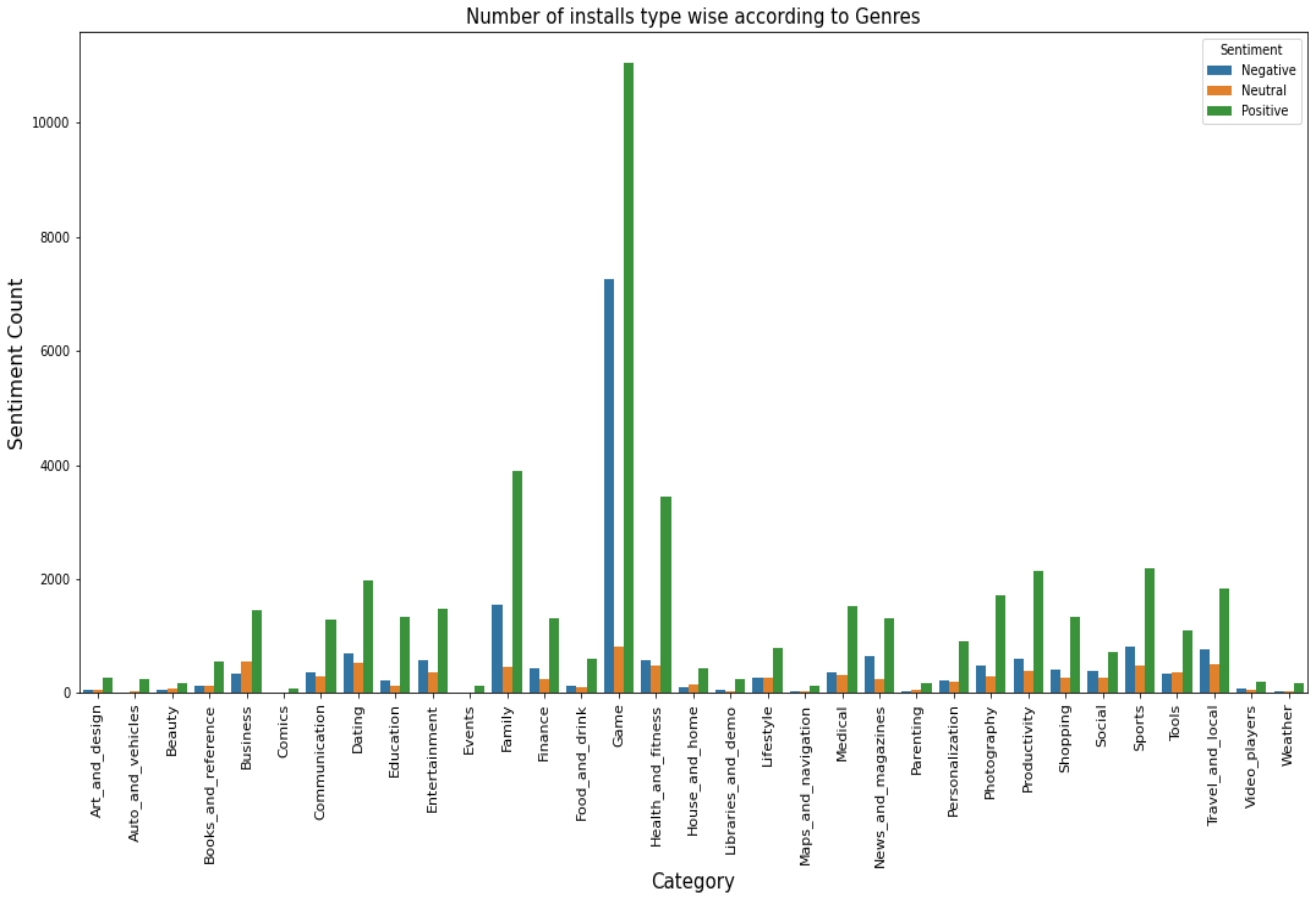




- It can be seen that average number of installations are high for Communication apps then Social followed by Game, Video players, photography (editing and good filter apps "Common"),News and magazines and, Travel etc.
- According the installation there is high rate of app engagement with most commonly downloaded apps like communication apps. There is a high chance of success rate for these apps if there no other constraints to stop using it.
- The Game category has high average size app. >40000 KB. So, this requires good space in the mobile to download. All other apps size is very much less than this.
- Even though game category requires good memory space, users are installing it which can be seen in the first plot. So, from this it can be said that there is good app engagement and also there is a high chance for other apps also be used by the people because of its small size. So, by this it can be said that success rate of the apps are good even though it is larger in size.

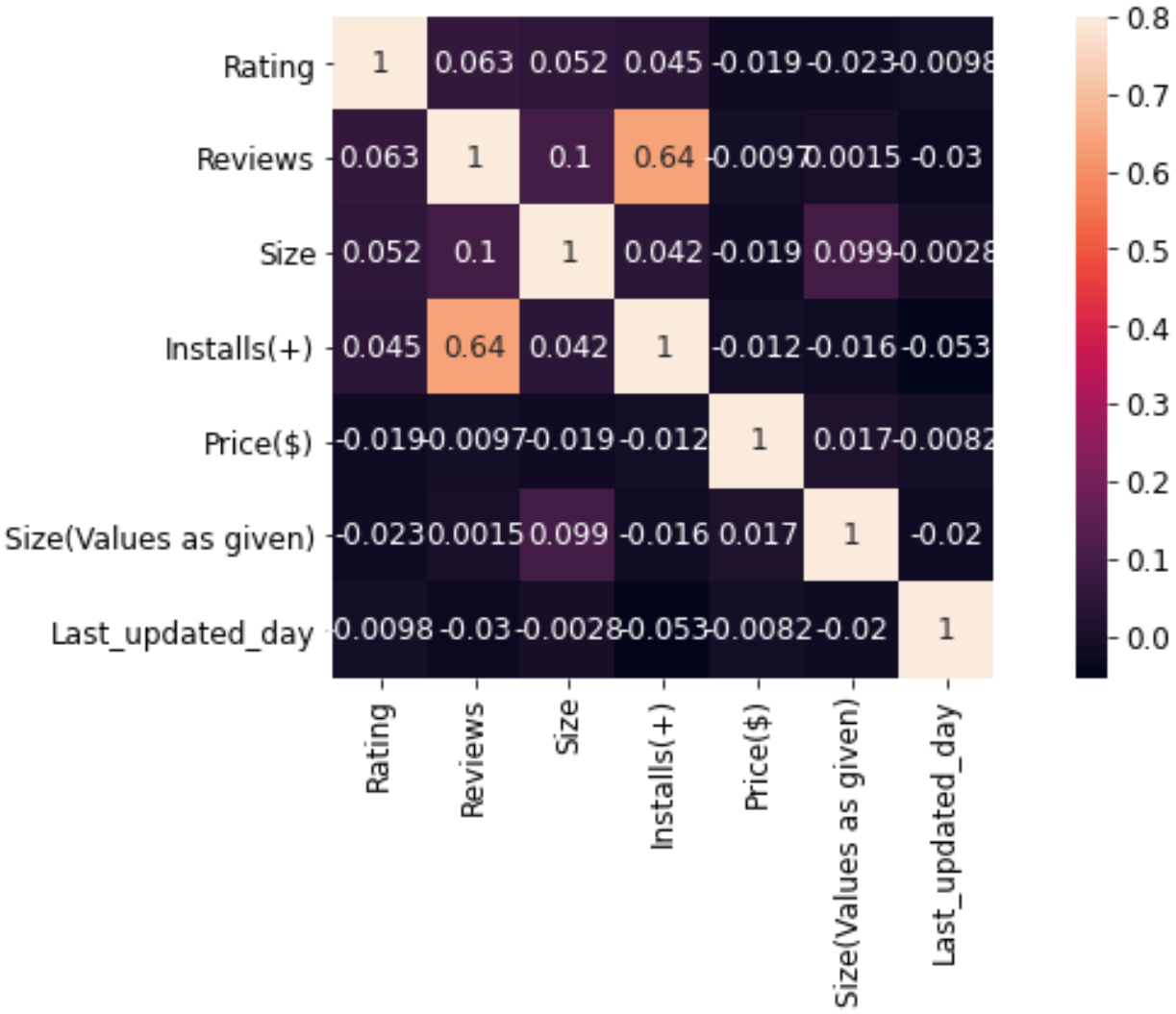


- From the genres and the number of downloads made, it can seen that the app that are downloaded are more from Communication genres, next Adventure and action genre followed by Arcade, Casual and Social genre apps.

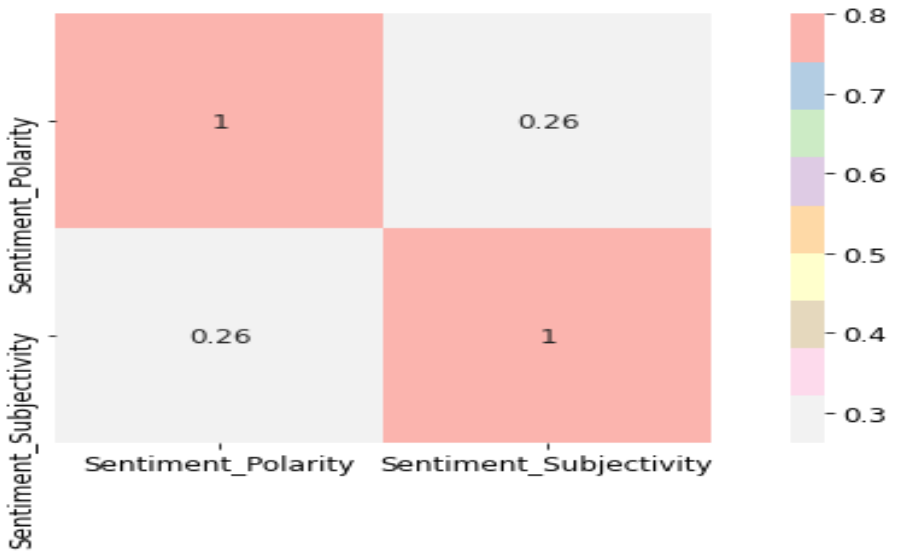


- There are a total of 33 category apps which are merged to Play store data from User reviews dataset.
- Here the count for all the apps has been obtained for each level of sentiment categories (i,,e, positive, negative and neutral sentiment). It can be seen from the above plot that there are more positive and highest number of positive reviews in Game category then the negative and least number of neutral reviews among all the apps.
- The same trend has been continued for the all the apps. Most of the apps have more positive reviews than negative and neutral reviews which says the apps businesses (Play store apps) are good and the success rate for the apps is also much higher even though there are some negative reviews in User reviews dataset with good app engagement features in the apps.

4.3. Multivariate analysis



- This is the correlation plot on Play Store data.
- The correlation between Reviews and Installs is higher and they both are strongly correlated variables. Hence multicollinearity exists between these two features.
- Size and Reviews are next correlated variables with correlation of 0.1.
- All the other variables are not much correlated with each other. Based on the requirement or predictor variable only uncorrelated variables are chosen and whichever is have good correlation with target are preferred.



- Sentiment polarity and subjectivity have little correlation between them and it doesn't affect much as it doesn't have high multicollinearity.

5: Case studies

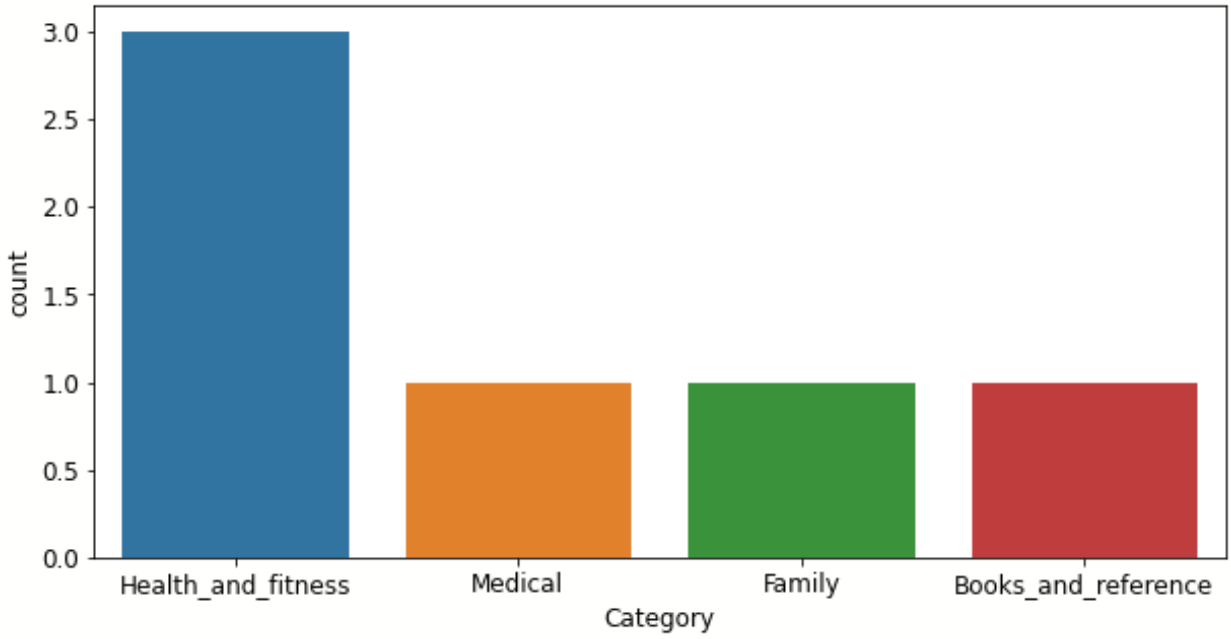
1.Find out the apps which are rated more than 4.8, which come under paid category, with 10000+ installs and also reviews are given for it.

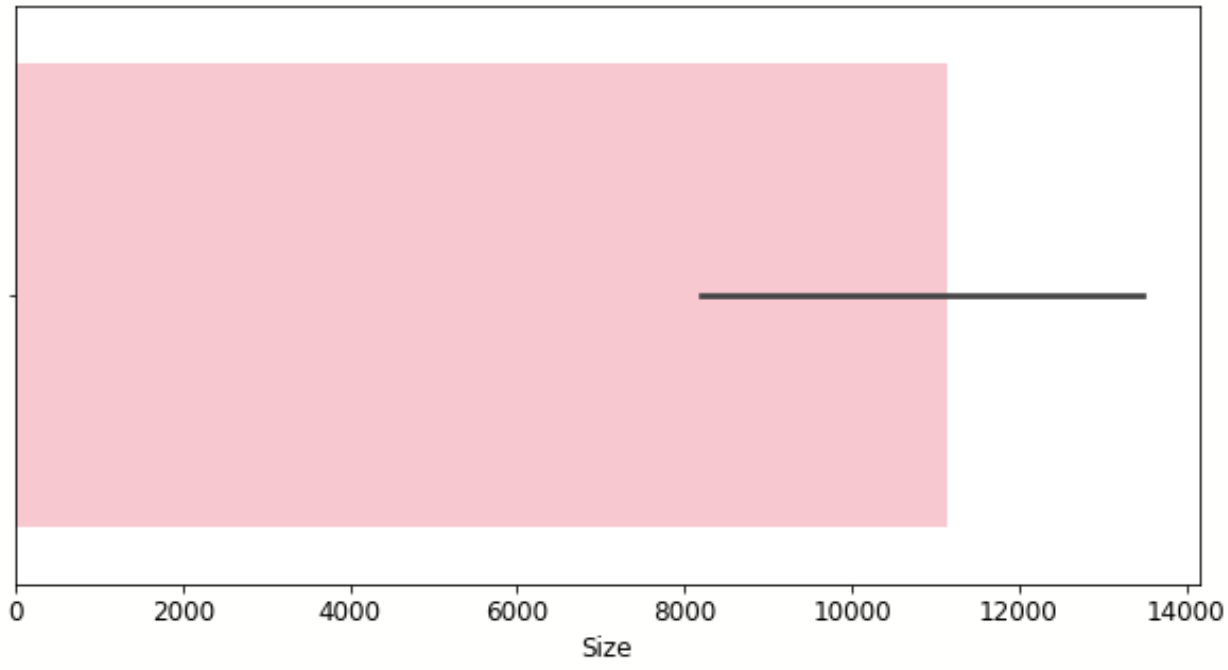
	App	Category	Rating	Reviews	Size	Installs(+)	Type	Price($)	Content Rating	Genres	Current Ver	Android Ver	Size(MB,KB)	Size(Values as given)	Last_updated_
1833	The Room: Old Sins	Game	4.9	21119	49152.0	100000	Paid	4.99	Everyone	Puzzle	1.0.1	4.4 and up	MB	48.0	

Observations:

- There is only one such row in the whole dataset.
- There is one gaming which is rated high(4.9) which costs 4.99$ with good number of installation (>10000+ installs).
- The genre of this app is Puzzles and size you require to download this app is 48.0 MB
- The new features or updated version of this app was in the second quarter of 2018 that is on 18th Feb'18.

2. **Find out the app with number of installations with more than 10,00,000 and with ratings greater than 4.8, reviews>50000 and size< 25000 KB**





Observations:

- There are 6 such apps which satisfy the given conditions and they are of different categories and genres.
- These are all free apps
- These apps are from Health and fitness, Medical, Family and Books and reference categories.
- The size of all these apps is up to 14000 KB

6: Conclusion

- The features in Play store data that mostly helps in **predicting the success rate of an app are the Rating, reviews, Installs and type of an app.**
- The features in Users review dataset that would help in the success rate or by which it can be said **that the app is successful or not. They are firstly Sentiment is useful then Sentiment polarity and subjectivity.**
- From the analysis it is seen that there are good number of apps with positive reviews than negative and neutral reviews.
- There are many categories of apps present in the play store and **the apps that are high in particular category is Communication and Social apps. It says these apps are more successful and have high app engagement.** "Yes, everyone uses communication apps for talking with others than calls and Social apps for sharing the things and also knowing different kind of things happening around us."
- There are more **apps in the Play store than that are given reviews by the users. There are more ratings given to an app than the reviews.** "Yes, I do that and how many does it? Many! It's a choice."
- There are more updates happening in third quarter then second quarter and few of them in first and fourth quarter. The last updated quarter for android apps in the play store happened more in third quarter with 5071 apps and next second with 2867 apps, first and fourth are 1692,1211 respectively.
- Mostly the updates happened in the Month of July, August and June in the descending order of the updates happened for apps with 3163, 1594 and 1273 respectively. **So may be most of the updated versions are released in these months and quarters which increases the success rate of an app with slight improvements provided to it.**
- **There is an increasing trend that can be observed from 2010 to 2018, the users who have rated the apps have increased. So, the success rate of the apps has also grown.** "Definitely, even now the people who are using the android apps is also high in number because digitalization and technological developments".