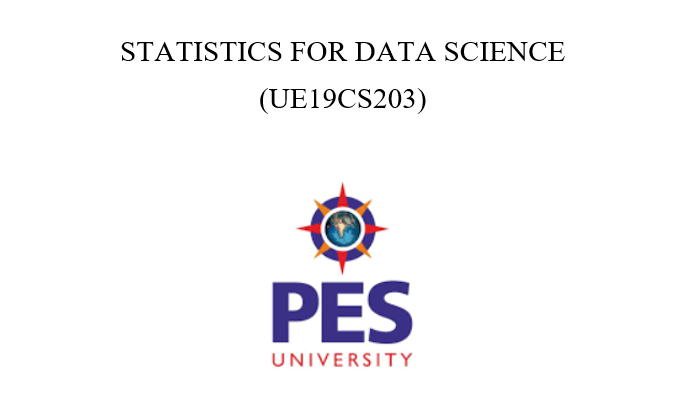
****

**TOPIC: GOOGLE PLAY STORE APPS**

**TEAM MEMBERS:**

**KUNAL MUJOO – PES2UG19CS196 KIRAN BS – PES2UG19CS186**

**GOWTHAM S R – PES2UG19CS133 KOUSHIK P L – PES2UG19CS193**

**ABSTRACT**

Our dataset is called GOOGLE PLAY STORE APPS. We performed several operations as mentioned below to perform efficient analysis.

* DATA CLEANING
* DATA VISUALIZATION
* NORMALISATION AND STANDARDISATION
* CORRELATION
* HYPOTHEIS TESTING

**INTRODUCTION**

The problem statement we chose to work on is as follows:

***Is the quality of paid apps on Google Play store significantly different than free apps?***

Makes one wonder doesn’t it? Is there really such a difference between some apps that we must pay to use them. Are they more fun to use?

Such are some questions we try to answer is this project through several techniques.

**ABOUT THE DATASET**

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!

The source of this dataset was Kaggle.com. **Kaggle** allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

It consists of many features including App name, Category, Reviews, Ratings, Size of apps, Number of installations, Type, Price is the app is paid else 0, Content Rating and more. It consists of about 11000 rows and 13 columns and gives sufficient amount of data to do a descriptive and non-biased data analysis.

**DATA CLEANING**

We used various techniques in order to clean our dataset. They are as follows:

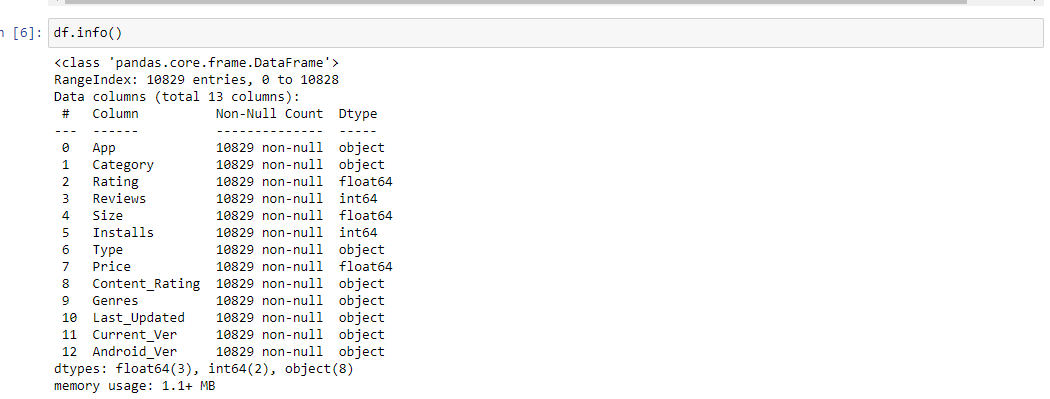
* Get rid of empty spaces and NaN values by removing when the row was not too important and otherwise by taking column mean.
* Converting several string columns that contained important numeric data, to numeric types.
* Remove duplicates.
* Change text from Lower to Upper and vice versa to avoid repetitions in columns.
* Spell check.
* Remove unnecessary formatting.

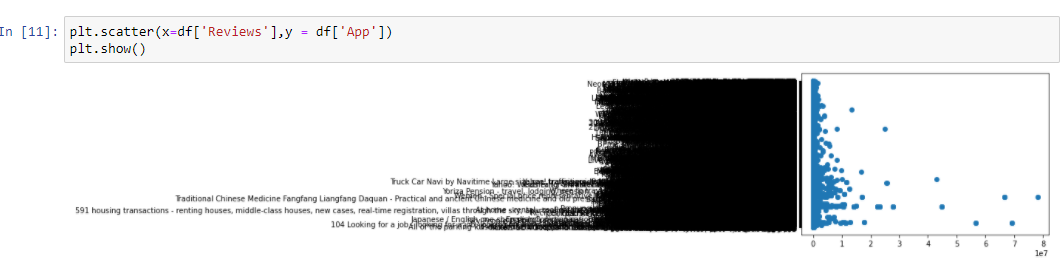
This dataset had several rows with NaN values. There were also many important numeric columns in string format. Hence it was essential to clean this dataset to perform proper data analysis.

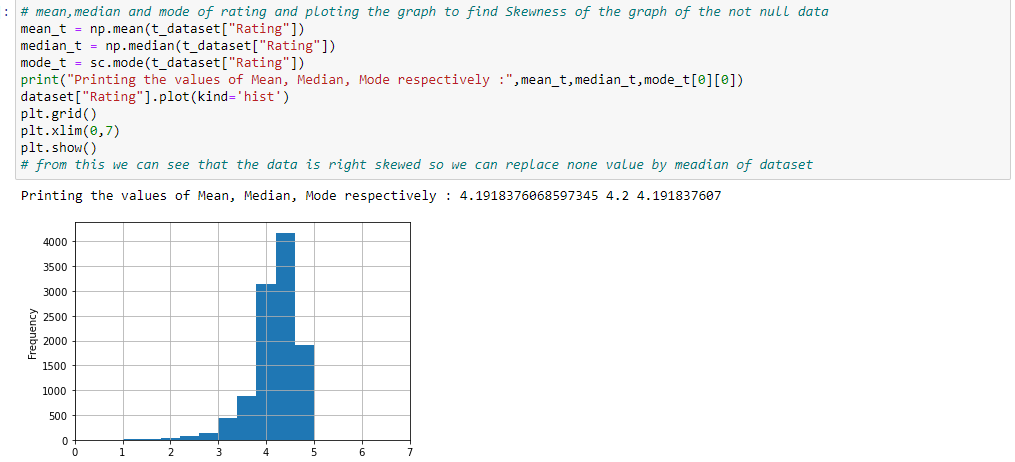
**EXPLORATORY DATA ANALYSIS**

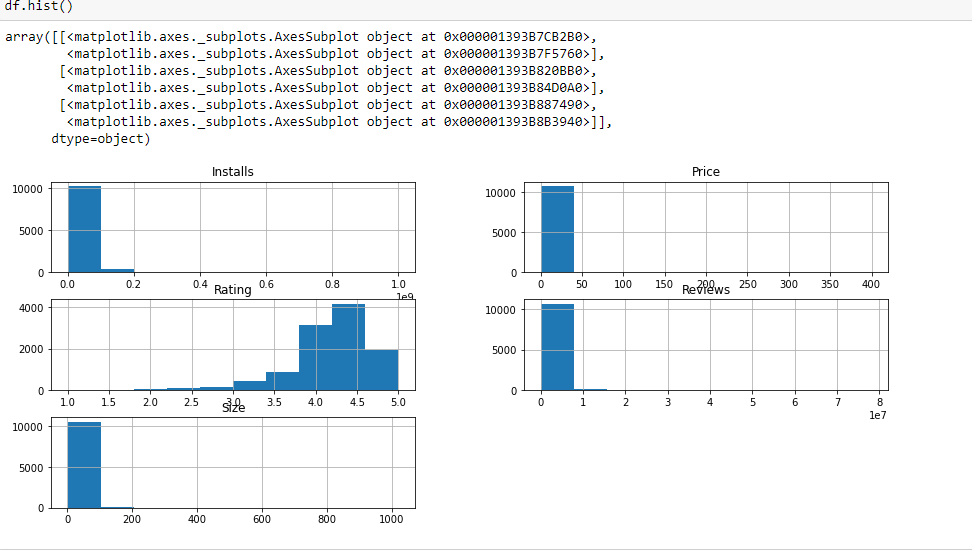
Firstly it is important to remove outliers from our dataset because they cause serious issues to statistical analysis like, skew the data, significant impact on the mean and standard deviation.

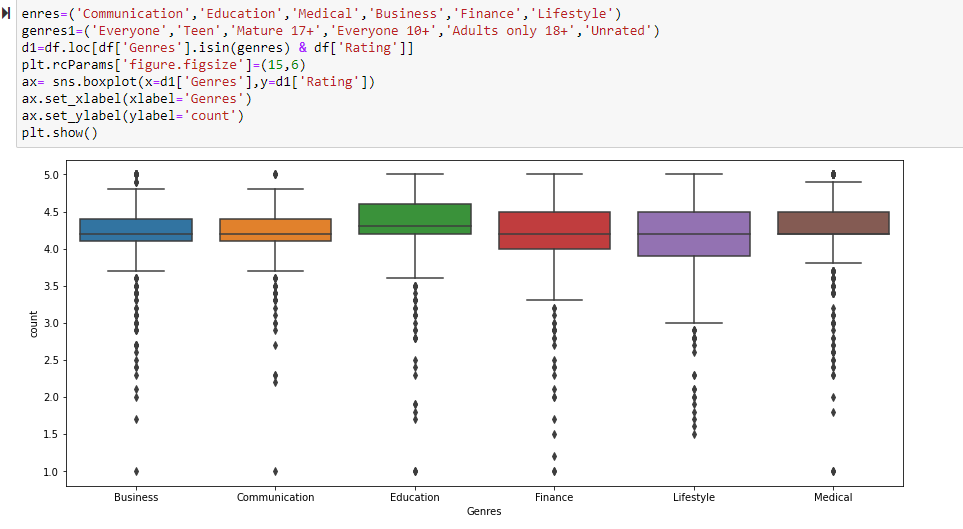
*Using the box plot, scatter plot and the histogram we see that there are many outliers in our dataset. (Skewed graphs, large values in scatter plot)*





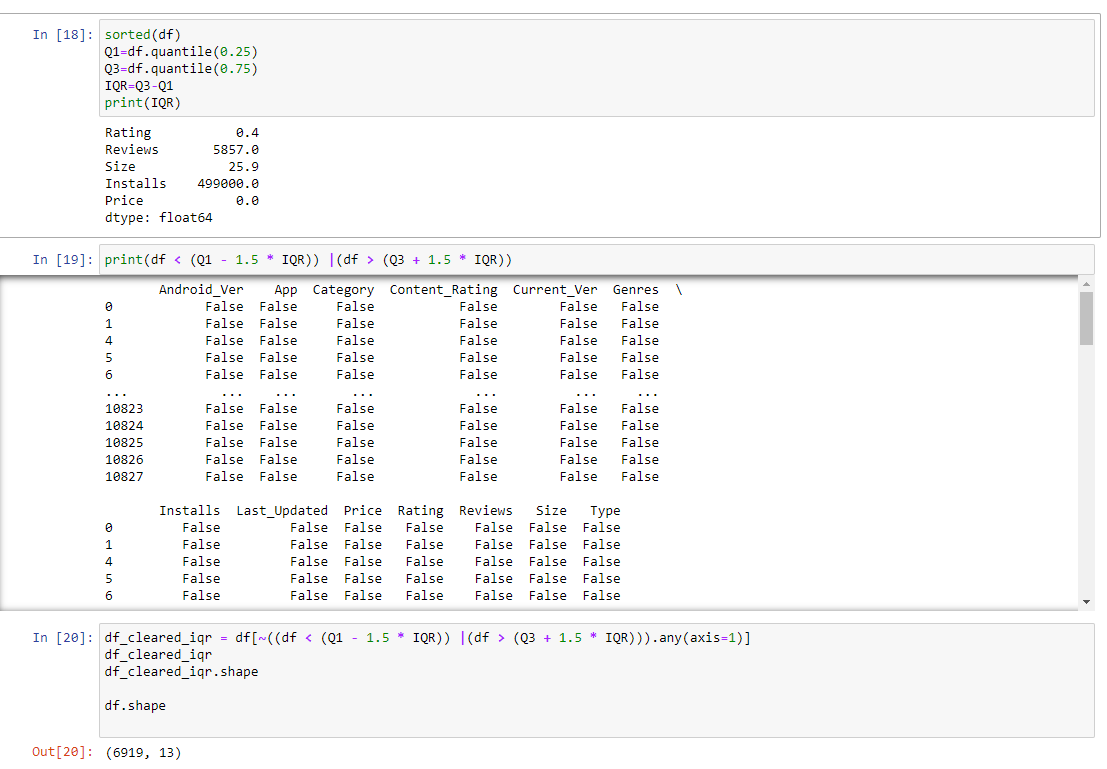




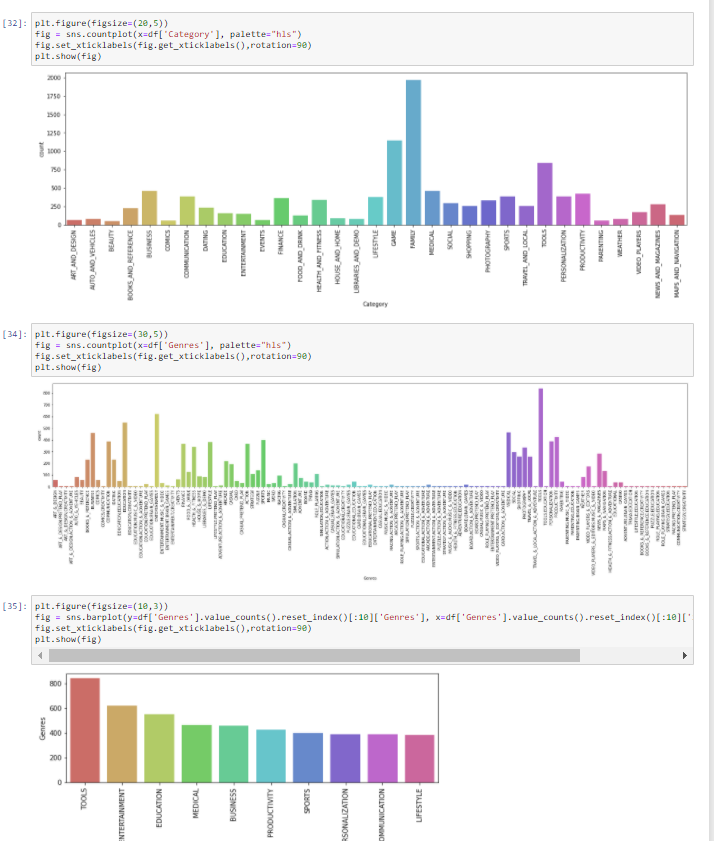


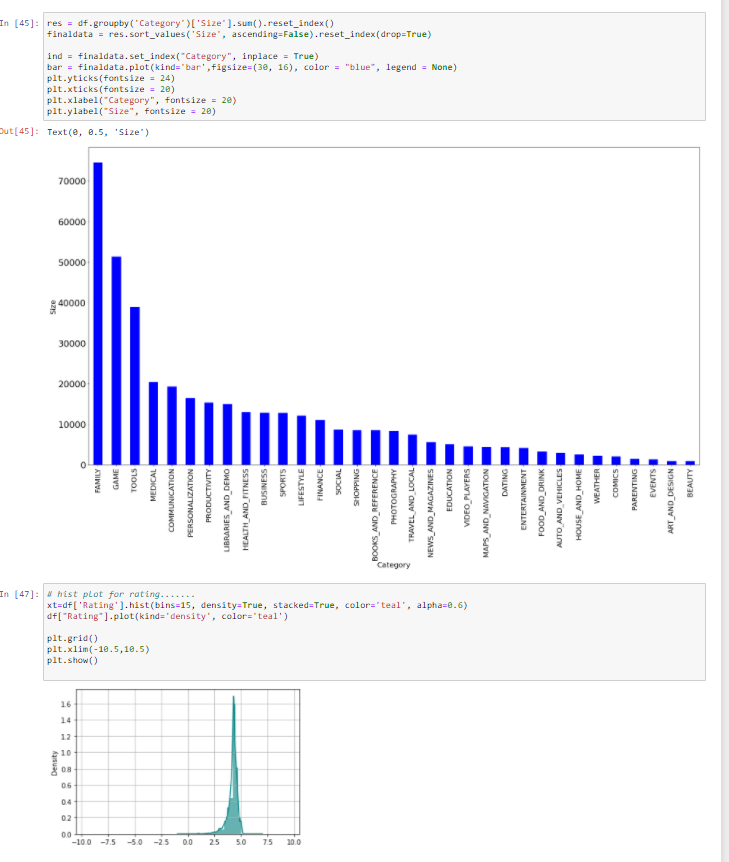
**OUTLIER REMOVAL:**

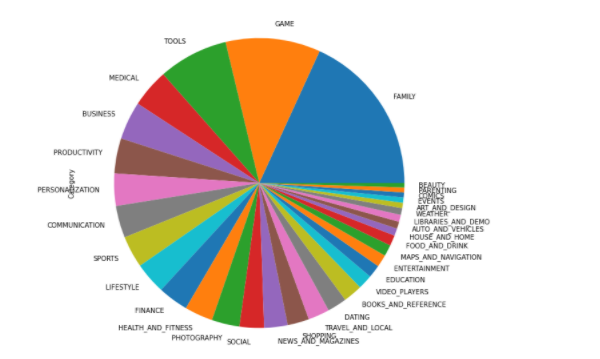
*Hence we remove these outliers using IQR formula. The values beyond 75th percentile and below 25th percentile are removed. This gives us a more accurate dataset to work on.*

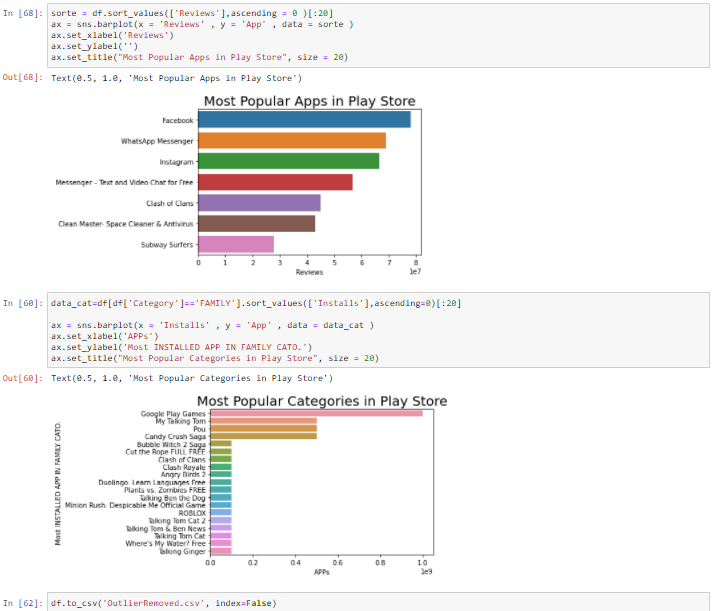


**GRAPH VISUALIZATION**

****

****

****

****

**The following are among several insights drawn from the graphs:**

* The number of free apps(92.5%) is much greater than the number of paid apps(7.5%).
* Most apps are for everyone.
* Category “Family” has the greater number of apps.
* Category “Beauty” has the lowest number of apps.
* Facebook has the highest number of reviews followed why Whatsapp, Instagram, etc etc.
* App “Word Search” has greatest size.
* 58 apps have highest number of installs 1000000000.

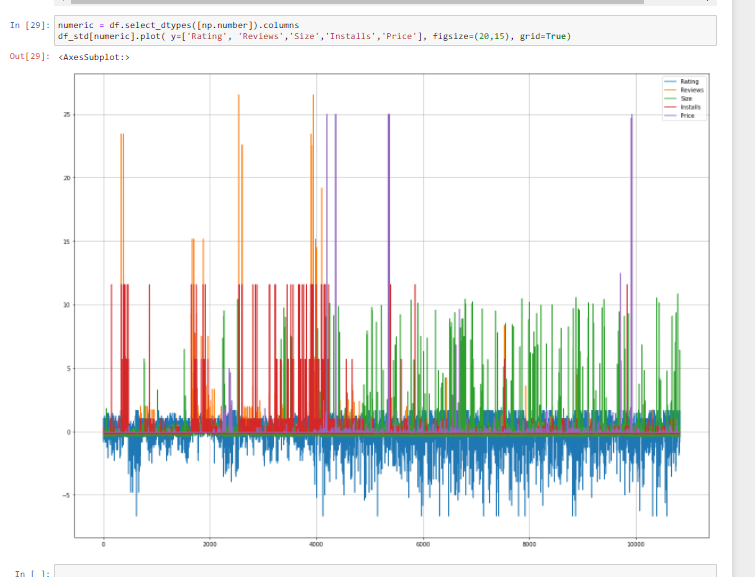
**NORMALISATION AND STANDARDISATION**

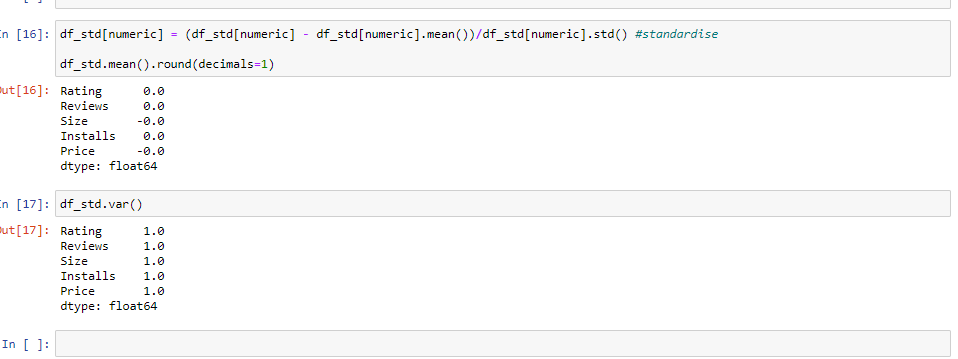
The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges.

**BEFORE NORMALISING:**

****

**AFTER NORMALISING:**

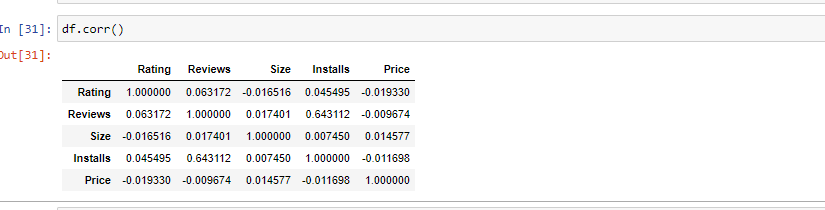
****

****

**CORRELATION**

**Measure of the strength of linear relationship between 2 quantitative variables.**

Positive correlation is when both variables move in same direction, and negative correlation is when both variables move in opposite directions.



From the table, ratings and reviews has the highest positive correlation, and ratings and price has the highest negative correlation.



**HYPOTHEIS TESTING**

***Is the quality of paid apps on Google Play store significantly different than free apps?***

The null hypothesis is that quality of paid apps on the play store is the same as the quality of free apps on the play store.

The alternate hypothesis is that the quality of paid apps is different from the quality of free apps on the play store.

(Difference between means method)

Let the level of significance α = 0.05.

H0 : υx -υy = DELTA0

H1 : ux – uy != DELTA0



The null hypothesis is rejected. The ratings of paid apps are not significantly different from the ratings of free apps.

**RESULTS AND DISCUSSION**

After extensive visual and numerical analysis, and through testing of our Hypothesis , we come to a conclusion that the quality of paid apps is not significantly different from the quality of free apps.

We see that the ratings of Free and Paid apps are very similar, and gives insight into where developers should invest their time and money.