**DATA SCIENCE PROJECT**

**Title: DATA ANALYSIS ON TMDB MOVIES DATASET**

**Team Members**:

|  |  |  |  |
| --- | --- | --- | --- |
| SLNO | Name | Section | SRN |
| 1 | S Ritesh Kumar | G | PES1UG19CS403 |
| 2 | Shreyas Bhaktaram | G | PES1UG19CS464 |
| 3 | Sahana Ramesh | G | PES1UG19CS413 |
| 4 | Sachin Shenoy | G | PES1UG19CS408 |

Dataset: TMDB 5000 movies

Source:<https://www.kaggle.com/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv>

**Abstract:**

The entertainment industry is a rapidly growing industry. The industry is also very unpredictable and volatile. The production houses which produce movies tend to suffer huge losses if the movie isn’t successful. Production houses spend a lot of money on producing a movie with a long runtime and the movies might not perform well. Our project aims to analyse the performance of movies in the box office. We aim to find out the correlation between different factors such as revenue of the movie and the budget or the vote count with the popularity, etc. We also try to hypothesize the ideal budget of a movie for a movie to be well received by the audience.

**Introduction:**

We found this dataset on kaggle to be relevant to our project as the records matched our needs for the project. In present times, it is important to understand whether a movie’s success can be determined by its budget and it’s vote count. It is worth noting that predicting the revenue before the theatrical release must leverage data that are available only before the movie is released. The accurate estimation of the movie box office revenue, mainly before the movie is released in the theater, is a more challenging problem for the movie industry. A movie with a higher vote count, has a chance of higher viewership and thus earning a higher revenue. The dataset is having a diverse range of movies (five thousand data entries) with many genres,languages,runtime and varying vote counts.

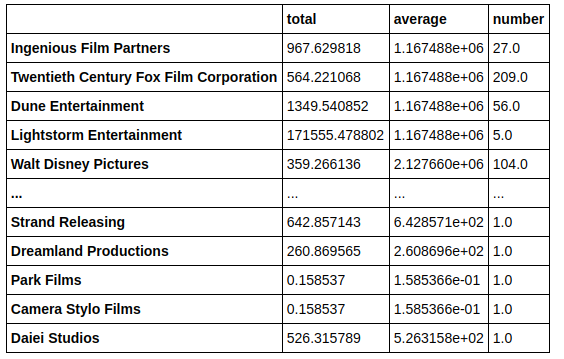
So our main motive behind choosing this dataset was to find the correlation between the budget, revenue, vote count and average rating, so that we can come to a conclusion whether the movies with higher budgets will always have a higher popularity and a revenue or there is a shift in trend with this dataset.

**Data Cleaning:**

1)**Removing redundant rows with missing data or NA data**

Condition: If a particular row has 3 or more missing or NA columns, that row is completely dropped. Rows with no “runtime”, “production\_country”, “spoken\_languages” are deleted. Movies with NaN revenue are replaced by 0 value.

2)**Selective removal of movies which are important parameters**

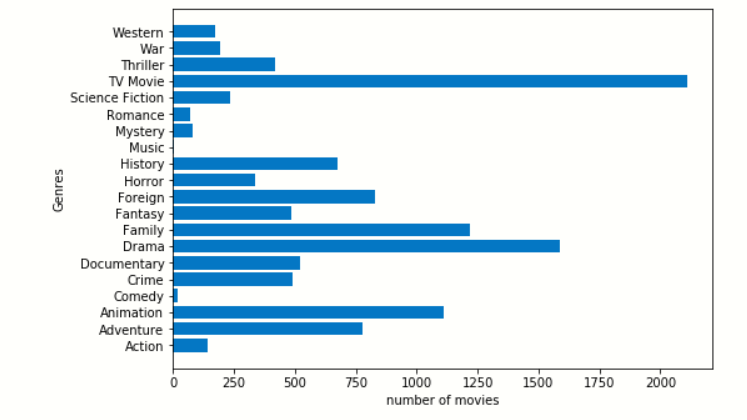
The rows with No “production\_companies”, no “runtime”, no “production\_countries” and no “spoken\_languages” are dropped.

**DataFrame Creation**

A data frame has been created which contains the data of every production company. This will be used to fill missing data.

**3)Handling the missing budget data**

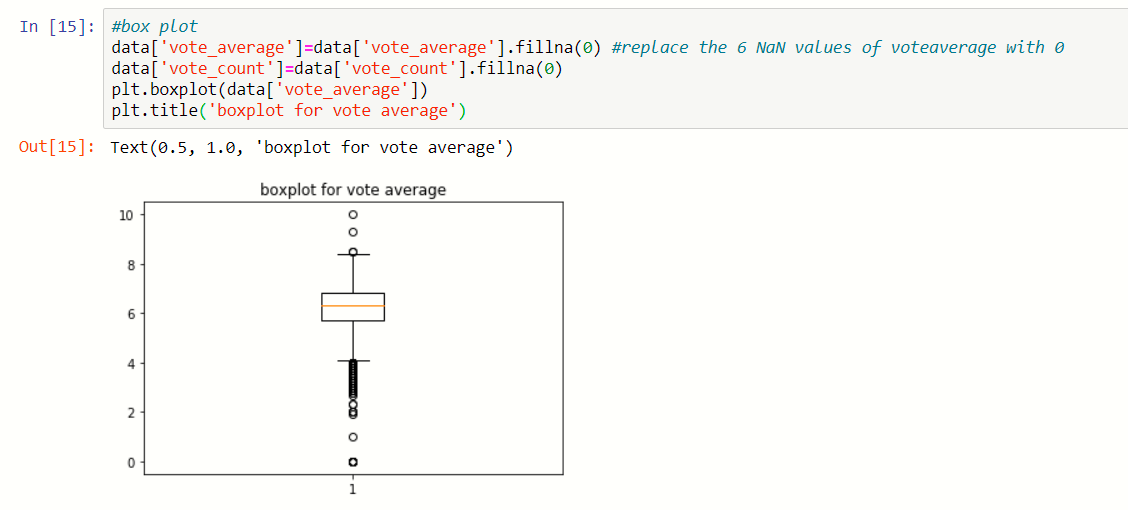
The movies with null data for the budget column have been filled with the average budget of the production companies(obtained from the data-frame) they have been produced by.



**Graph visualisation**

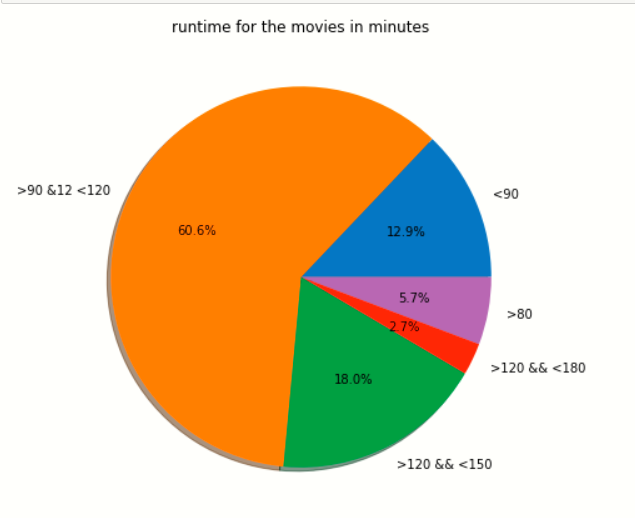
1)**Bar chart for number of movies in different genres** :

A horizontal bar graph was used to visualize the different number of movies in each genre.



2)**Box plot for vote average:**

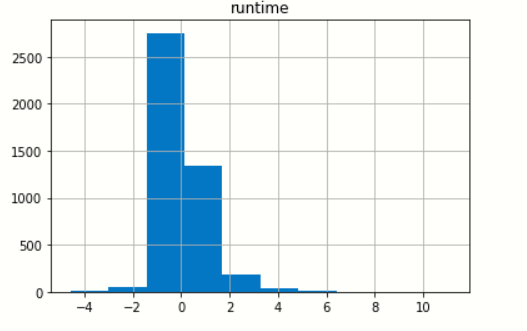
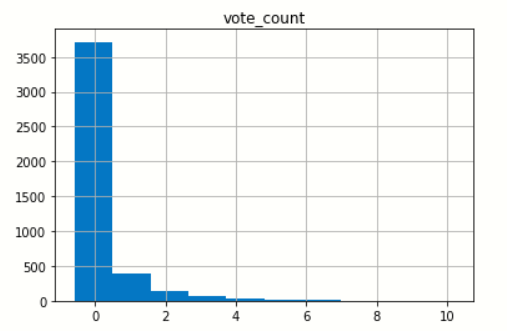
A box plot was used to visualize the mean and interquartile range.

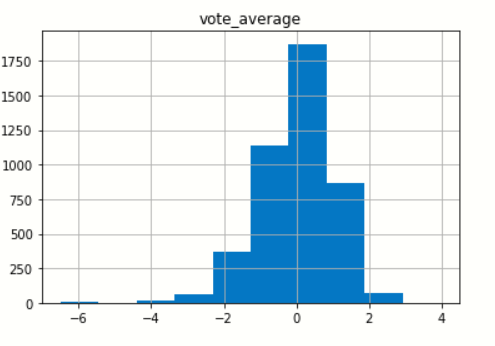
3)**Pie Chart :**

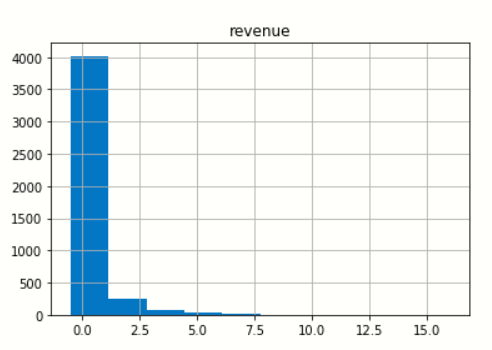
A pie chart was used to plot the ratio of movies with different runtimes and it was observed that almost 60% of movies had a runtime between 90 to 120 minutes.

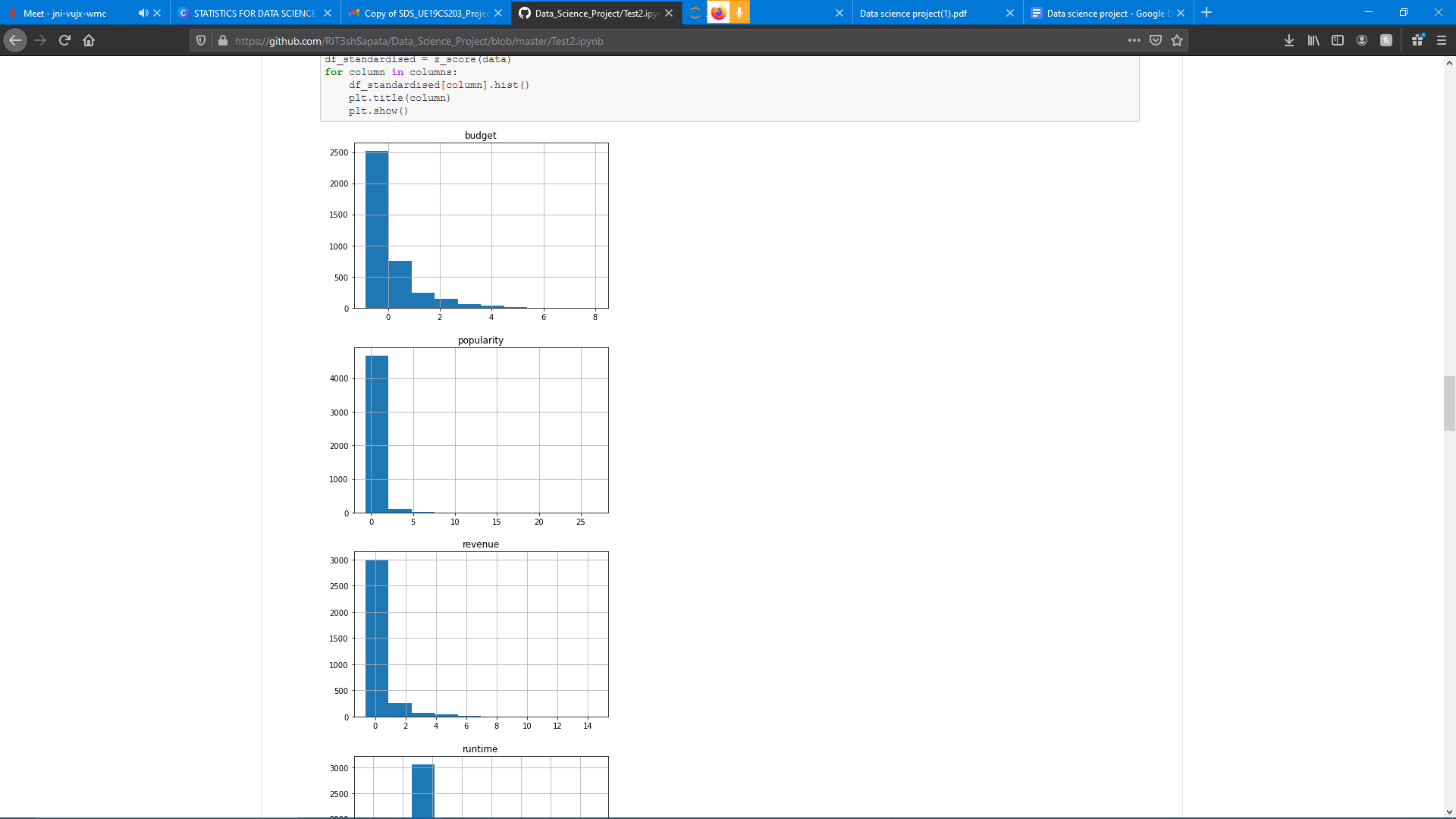
**Normalisation**

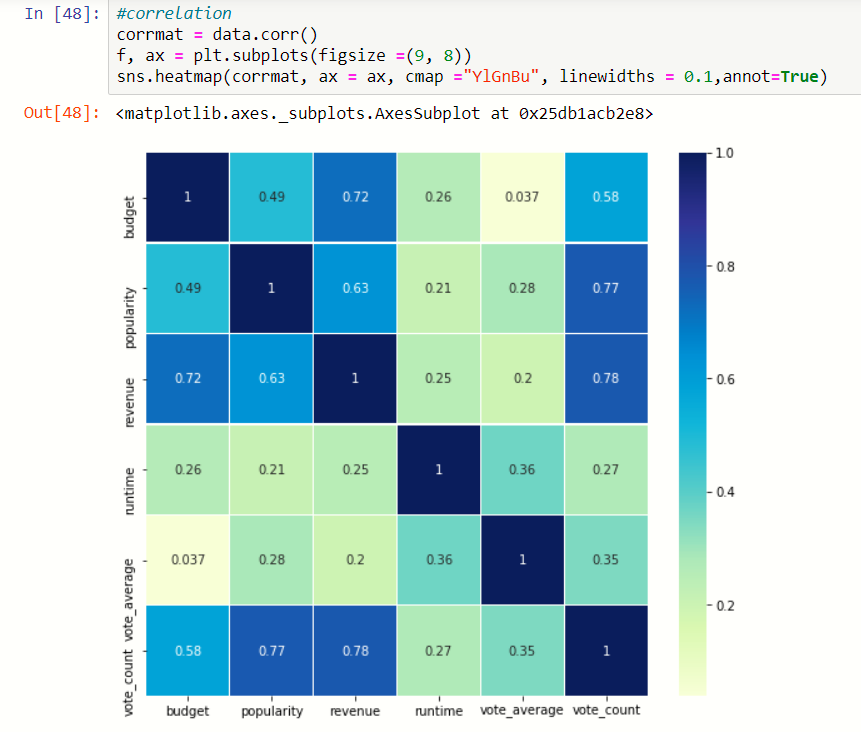
Normalisation refers to rescaling by the minimum and range of the vector, bringing all the values of numeric columns in the dataset to a common scale.Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data. Due to the huge range of data in our dataset, we have normalised the data for easier computation and scaling of the data for easier operations on data. We used standard normalisation( z-score) to normalise the data. Below are the graphs of the normalized numerical columns.









**Correlation** 

Correlation matrix ( with the correlation coefficient) using heatmap

Inference : All the variables are positively related to each other .

.**corr()** is used to find the pairwise correlation of all columns in the dataframe. Any na values are automatically excluded. For any non-numeric data type columns in the dataframe it is ignored. Note: The correlation of a variable with itself is 1.

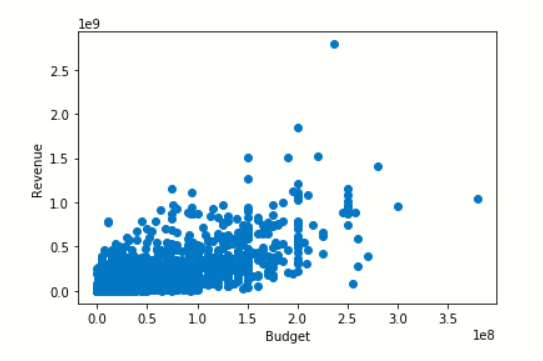
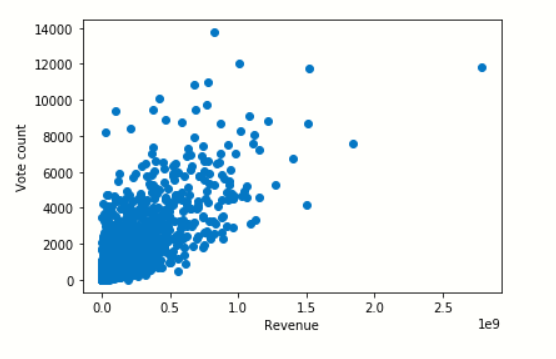
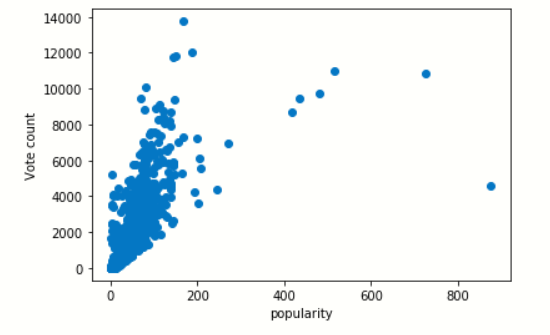
**Scatter plot to show relation between variables in the data**

Conclusions : Vote count increases with popularity,

Revenue of the movie increases with increase in votes of the audience.

Revenue increases with increase in budget in most cases.

Movies with higher budgets usually have more popularity and a wider audience, hence the vote counts are high too. Since the vote counts are high, a lot of viewers have watched the movie and the revenue will be high.



**HYPOTHESIS TESTING**

**Case 1:** Budget > 40 million and runtime

For this case, we have considered the budget and the runtime of the movie. The mean runtime is 120mins and along with a budget constraint of over 40 million, the null hypothesis is rejected, which means that all movies having a budget over 40million will have runtime over 120mins.

**Case 2:** Budget > 40 million and vote average

For this case, we have considered movies with a budget of over 40 million which is the approximate mean budget and have compared it to the vote average score with a mean of 6.2. Here, the null hypothesis is plausible which means that if a movie has a budget of over 40 million, it doesn’t mean that the movie will have a rating greater than the mean rating.

**CONCLUSION**

From the above hypothesis and graphs we have come to a conclusion that even if the Production company has a budget over most of the movies, the rating will be not more than the movies having a lower budget. The case 1 hypothesis shows the runtime of the movies which tells us that the company will have more content and more of a storyline to show in the movie if they have a higher budget which will lead to a better experience for the end user. The bar graph which displays the genres gives an idea to what is more popular in the film industry and what category the company must produce the next movie on.

**DISCUSSION**

This dataset primarily focuses on the way production companies must approach the market before spending a ton of money and later being not such a successful movie by either just breaking even or making a loss. The market is very diverse and is evolving at a very quick rate, hence it is important to know the current trends which can lead to a blockbuster.