**DATA ANALYTICS PRACHI SINGLA DATE: 08-09-2019**

1. **INTRODUCTION**

In this project I will analyse movies data set containing information about 4803 movies between time span of 1969 and 2068.

* 1. **Problem: MOVIES DATA ANALYTICS**

Data contains20 variables for 4803 movies, spanning between 1969-2068 in 88 countries. I have to analyse data to deduce some insights which would be useful in future in multiple ways.

* 1. **Interest:**
     1. The results can help film companies to understand the secret of generating a commercial success movie.
     2. The insights can be useful for various service providers whether they are offline or online service providers in recommending movie.
     3. Analysis of data can help people to watch movies with high rating which is already giving idea to people in advance about movie.

1. **Data Acquisition and Cleaning**
   1. **Data sources:** The dataset is provided to me in csv format. It contains 20 variables for 4803 movies, spanning between 1969-2068 in 88 countries. This section describes the various data cleaning and data wrangling methods applied on the Movie datasets to make it more suitable for further analysis.
   2. **Data cleaning: It includes multiple steps for implementing:**
      1. **Removing Unnecessary Features** Some features such as the original\_title, homepage, ID, keyword, tagline were unnecessary attributes and were dropped to reduce the dimensions of the dataset.Then we have made **statistics** of data to see its null values, its mean, max values, information about types of features, no. & name of given columns, removing unwanted columns, replacing null values with nan or median according to requirement. The data frame was exploded wherever the analysis demanded it.
      2. **Feature selection**: In feature selection, I have created a new features which I think will give better idea for movies success i.e. returns, which will be the ratio of revenue to budget. Other than this I will take every feature one by one to explore, analyse, visualize and then correlate.

### Exploratory Data Analysis:

### In this section, the various insights produced through descriptive statistics and data visualisation is presented

### Release dates: I have extracted year from release dates and made a data frame against title of movies. Then I plotted it which is showing no. of movies against year. The plot is showing that maximum no. of movies are in mid between of our time span of years

### 3.2 Title: Most popular Title of movie can be observed if we know about most frequent words in multiple titles of movies from which we can get to know that which type of movies are coming .So I have made word cloud which will represent frequent words in large and bold and less frequent words in small letters.

### 

### We can see that some words like man, girl, love, dead, good are very frequent in titles of movies according to this word cloud for titles of movies. . I think this summarizes the idea of the presence of romance, love, friendship in movies pretty well.

### 3.3 Overview: Most popular movie overviews can be observed if we know about most frequent words in multiple overviews of movies from which we can get to know that which type of movies are coming .So I have made word cloud which will represent frequent words in large and bold and less frequent words in small letters.

### 

### Overview corpus is showing some frequent words for movies that are world, family, friend, life, find. So summarizing these two features above us can get good idea about movie titles, overviews & their themes.

### 3.4 Original Language: There are 37 unique languages in all 4803 movies. So we have created a data frame of all 37 unique languages against their total numbers followed by its visualization in a bar chart.

### 

### 

### Here we have visualize language distribution for all languages accept English which is quite obvious and clear from above data frame which we made for numbers along with unique language.

### Other from English mostly movies are in French, Spanish and then Dutch and so on as seen above in bar chart.

### 3.5 Popularity: Firstly I have made popularity feature as float numeric no. and then applied some statistics to know more about data. I have created a distribution plot to see popular movies on a scale of its scores.

### 

### We can see that all movies score is around 100 which is more than 75 percentile of distribution plot (i.e. 28.31) but it is somewhere right skewed distribution in which mean is less than standard deviation.

### I have sorted values according to popularity & title so that we can find most popular movies. From analysis I found that Minions is most popular movie amongst all.

### 3.6 Vote\_Count: Same thing that I did in popularity, I have repeated same for vote count also and then created a distribution plot.

### 

### Plot is right skewed and which means movies vote count are more.

### 3.7 Vote\_average: Same thing that I did in popularity, I have repeated same for vote average also and then created a distribution plot

### 

### If we see vote average rather than vote count, then we can say that this is normally distributed, which means movies votes are averagely distributed.

### Now, Question that comes in mind now, is popularity has any relationship like there is between vote\_avg. and vote\_count. To see this we have to make scatter plot and if there would be any relation between vote\_count/vote\_avg. and popularity, we will join popularity also to above data frame.

### 

### 

### In both above cases Pearson coefficient is between 0-1 which implies that there is some relation between popularity and these two features but is not strong enough that we could deduce any insight from this.

### 3.8 Status: Although movie status is not really data analysis which is going to be useful for us, because almost every movie status is released. But in case of recommendation of movies or in case of intelligent movie chatbots, it can be useful.

### After data exploration, Status of 4795 movies are released status out of 4803 movies.

### 3.9 Runtime: Now we will see runtime effect on movies data. We will extract some interesting insight from length of movie

### From statistics of data for runtime of movie which I have extracted, maximum runtime of movie in data is 338 minutes i.e. 5.6 hours. I have created a distribution plot for runtime of all movies to see scenario.

### 

### This distribution is somewhere right skewed and which shows that most of movies with runtime in between 1.6 & 2 hours, which are relatively more liked by people and are more popular.

### Let’s see the relationship between movie run time and profitability or success of movie (returns of movies)

### 

### Here we got pearson coefficient much less than zero which is showing negligible relation between movie returns/success and it's runtime, as its distribution is wide spread and which is showing no particular trend or relation.

### We can also deduce shortest and longest movies ranking by further exploring data. After exploratory analysis I have found that Vessel is shortest movie and Carlos is longest movie

### 3.10 Spoken Languages: Let's talk about spoken languages but before exploring and analysis we need to convert this into numeric feature. After conversion, I have explored data and extracted no. of spoken languages per movie in form of data frame as below:

### 

### We can see that mostly movies are made with one language only but few movies are present which have made with multiple languages and in our dataset 9 is maximum no. of languages which are used within same movie.

### But if we see movie popularity with respect to no. of spoken languages used within movie then we can say that these two features are not directly proportional. This will be clearer by visualizing plot as below:

### 

### From above visualization it has been more cleared that spoken language has no effect on popularity of movie and i am sure this has no strong relation with other features also because popularity has good relations with other important features like budget, revenue, return, vote count etc. but spoken languages don’t have.

### 3.11 Genres: After exploration of data we can say that there are 20 unique genres for all movies present in our dataset. Let's look at no. of movies per genre.

### 

### Let's visualize genres against no. of movies for more clarity.

### 

### We mostly have drama movies, then comedy movies then thriller and so on in ranking.

### 3.12 Production Countries: This feature does not really affect movie popularity, its success or anything important. But for some knowledge based insights it can be done

### 

### US is most popular country for movie shooting as shown across all 88 countries. So we can visualize other also other than US for their existence in movie shooting countries.

### 

### From above bar plot, it has been shown that other than US, Germany, France, Canada are next at high score

### 3.13 Budget: Since BUDGET is most crucial part for movies data so we will analyse this data in respect of revenue, return, success of movie to extract some useful insights

### We have applied some stats on budget data & we noticed that maximum value is very far from mean so we can say that there are outliers in dataset which are influencing budget data.

### We made a distribution plot for budget

### 

### Plot is right skewed and decaying which is showing most of movies have budget are less than mean.

### We can find out most expensive and least expensive movies as we did for longest and shortest movies. After exploring data I found that “Pirates of the Caribbean: On Stranger Tides” is most expensive movies with highest budget in 2011.

### But if we see return of movie/ success of movie it has lowest budget. From this data analysis as above we can say that most successful movie is “Modern Times”

### We can say that budget and return has no direct proportional relationship

### Let's visualize these features in a plot

### 

### Here pearson coefficient is 0.71 which is showing close and strong relationship between these two features i.e. revenue and budget.

### 3.14 Revenue: As we did for budget we will do that same for revenue also. We will extract statistics for revenue data followed by exploratory data analysis. Since mean is very less than max value therefore there are so many outliers in data.

### We already have seen relationship with budget, returns & revenue which is not so positive but has some relation.

### I have created distribution plot for showing revenue of movies over time span

### 

### This is showing same results as for budget and plot is right skewed plot and showing that most of movies have less revenue and less movies have high revenue

### Now we will see for high revenue movies as we did for high return and high budget movies

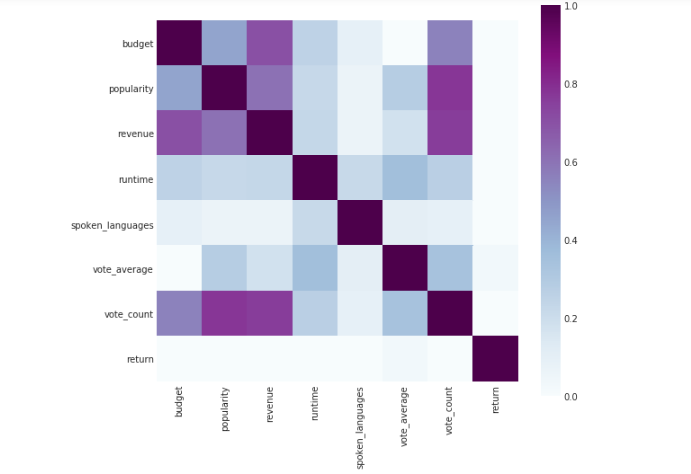
### ‘Avatar’ is at highest revenue movie rank but not with highest return and highest budget so this is showing that revenue has no direct proportional relationship with either budget and returns of movies

### 3.15 Return: Since return has no directly proportional relationship with budget and revenue so I have made a data frame in sorted way of returns where we can see clearly the values of return, revenue, budget against title of movies.

### 

1. **Correlation Analysis:** It's time to create correlation matrix for all features with each other

For this I have removed some column which are not useful & then I have made a heat map matrix to see the relationship between multiple features with each other.



A heat map matrix as shown above is a graphical representation of data with multiple features where the individual feature contained in a **matrix** are represented as colours.

The vertical line with different intensity colours and values shows the extent of correlation between different features which we can examine by this matrix. So taking account all features based on strong and weak correlations we can make a table for more clear and straight picture of analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Strong Relation** | **Fair Relation** | **Weak Relation** | **No Relation** |
| Budget | Revenue | Popularity, Vote count | Run time | Vote Avg, Return |
| Popularity | Vote Count , Revenue | Budget | Vote Avg, Run Time | Return |
| Revenue | Budget, vote count | Popularity | Run time, vote avg | Return |
| Run Time |  | Budget, vote avg | Popularity, revenue, vote count | Return |
| Vote Avg |  | Popularity, runtime, vote count | Revenue, budget | Return |
| Vote Count | Popularity, revenue | Budget, vote avg | Runtime | Return |
| Return |  |  | Vote avg | Popularity, revenue, budget, runtime, vote count |

1. **Conclusion**: I have concluded all correlations with their relation intensity that to how extent they are related or not related. Other than this in whole analysis I have explore and analyse data followed by their visualizations in different forms like bar plots, distribution plots or scatter plots.
2. **Discussion:** Since this is open ended analysis without any specified target so I just have analysed data by picking each feature and some relationship with each other. Otherwise many more things can be done with this dataset according to any target if specified like as follows:
   1. If we want to use machine learning model on this problem statement to find revenue generated for all movies regression techniques would be applied.
   2. The Classification model can be used with the same Feature Engineering steps as those followed by the Regression Model if we want to predict movie success.
   3. Alternatively, based on rating, popularity or may be other factors also, movie recommendation system can be generated of any type out of three (i.e. content based recommendation system, collaborative filtering based, hybrid recommendation system)

This report highlighted the processes of data wrangling, inferential statistics, data visualization, feature engineering performed on the Movies Dataset. All the results and insights gained as part of these processes were also highlighted.

1. **Future Insights:**
   1. The results can help film movie companies to understand the secret of generating a commercial success movie.
   2. The insights can be useful for various service providers whether they are offline or online service providers in recommending movie.
   3. Analysis of data can help people to watch movies with high rating or high popularity.