ITM 891 Project Write-Up

IMDB Action Movie Revenue Predictor

Project Summary:

Given the parameters before the release of the film we attempt to predict its Revenue in USA. Data was fetched, cleaned, explored and visualized using Python programming language. A lot of concepts studied over the semester were applied. Finally designed a PCA applied Linear model to get the whole a prediction tool and is tested for an unreleased movie: Wonder Woman 1984.

What question / purpose did you identify?

I strove to find out whether, knowing only things I could know before a film was released, what the revenue of the film would be? What parameters best predict a good or top grossing film?

How did you go about answering that question / purpose numerically?

I trained a model on randomized 80% of the data and then tested it on the remaining 20%. To predict the revenue, one of the most correlated variables was its **Budget**.

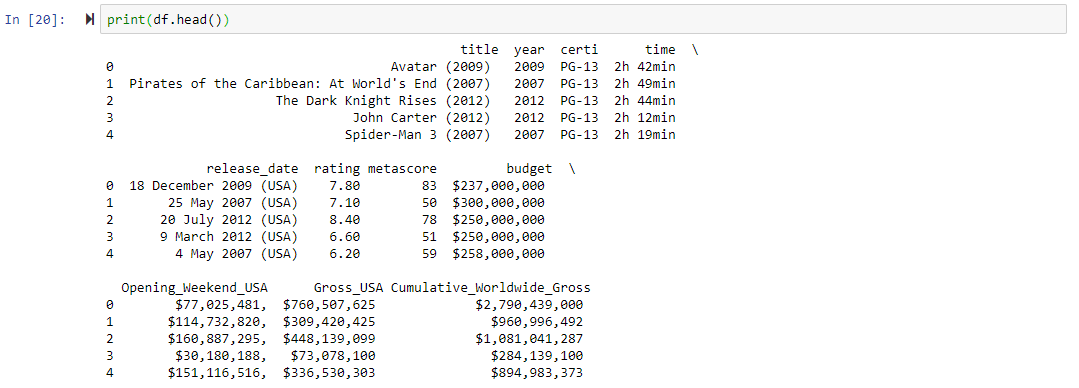
After cleaning the data, I designed few **Linear regression** models and got an **average = 0.54**, with “1” being the perfect model. **The predictor variable selected is Gross\_USA.** We will be predicting just for Action movies and their revenues in USA.

**How did you gather your data?**

Using the method of **Web Scrapping**, data was fetched From **IMDB** using the **beautiful Soup Package**. Following are the attributes fetched:

1. Title
2. Release Data
3. Certifications
4. Runtime
5. Rating
6. Metascore
7. Budget
8. Opening weekend revenue
9. USA revenue
10. Worldwide revenue

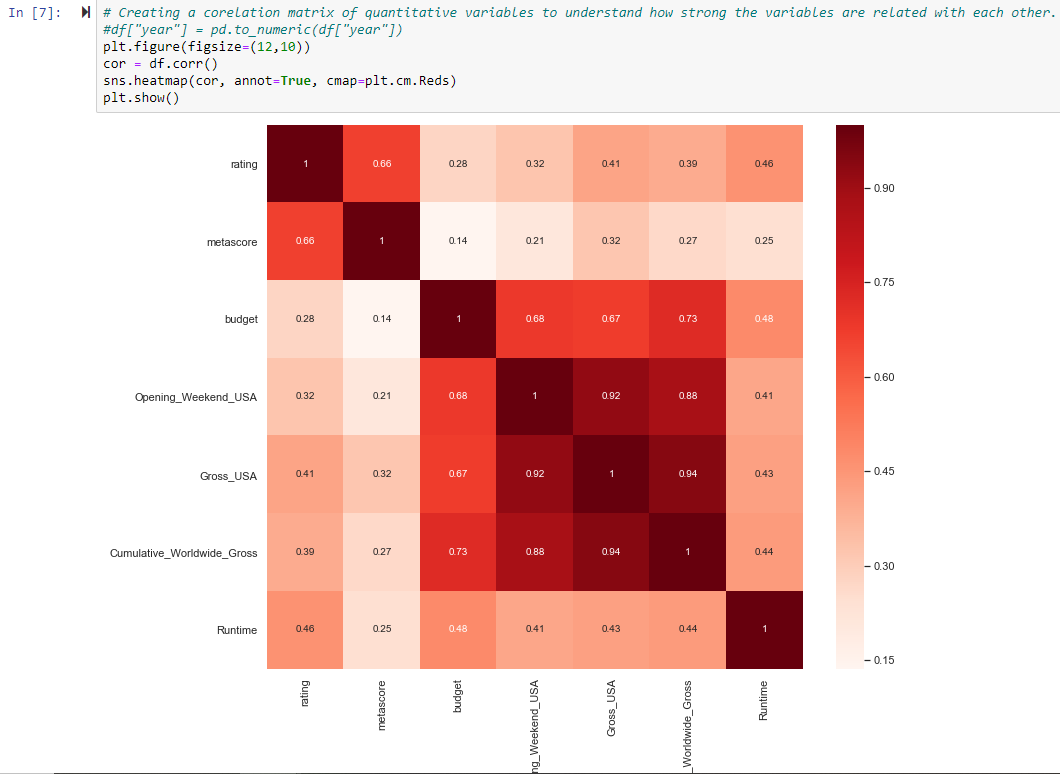
Before scraping, a csv was generated with around 1500 action movie URL. After processing each url, the Web scrapping tool scanned and fetched the required data from the HTML web pages and converted it to a data-frame. It took around **30 mins** to perform this task. The output data frame was exported and saved into local drive- **Name: Web\_Scrape\_data**



How did you explore your data?

Looking at the info and head of the data-frame we came to know about its dimension, datatype and an overview. The package **Seaborn** was used to explore what the data means and how each parameter is relate with each other. Few of the results in data exploration were used to clean the data set further.

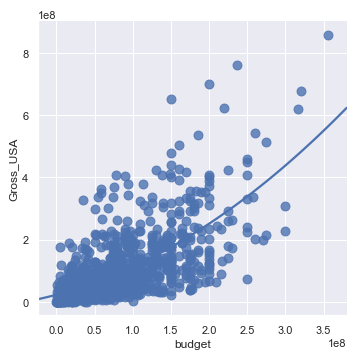
A correlation matrix was made and plotted to get a better pitcher of how the variables are related numerically.



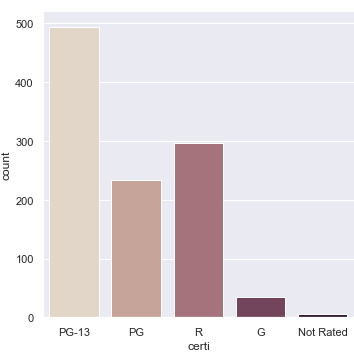
From here we were able to conclude that Budget has the highest correlation with Gross\_USA.

We can see that Opening\_Weekend\_USA and cumulative\_Worldwide\_gross have exceptionally high corelation value. This might create the effect of multicolinearity in the model. With help of some business understanding it is clear to us that Opening\_Weekend\_USA is part of Gross\_USA and Gross\_USA is part of cumulative\_Worldwide\_gross. Thus, these are essentially categories of outcome variable and hence can be neglected.

Next, we look at the relation of Budget and Gross\_USA. Below is the graph plotting the relationship. It can be observed that the relationship average line is not a straight line due to which we will be exploring polynomial relationships as well



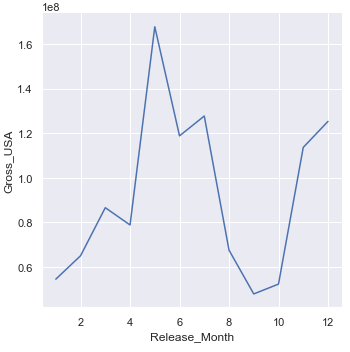
As Proportion of G and Not rated movies is low, we will be excluding that from the data set.:



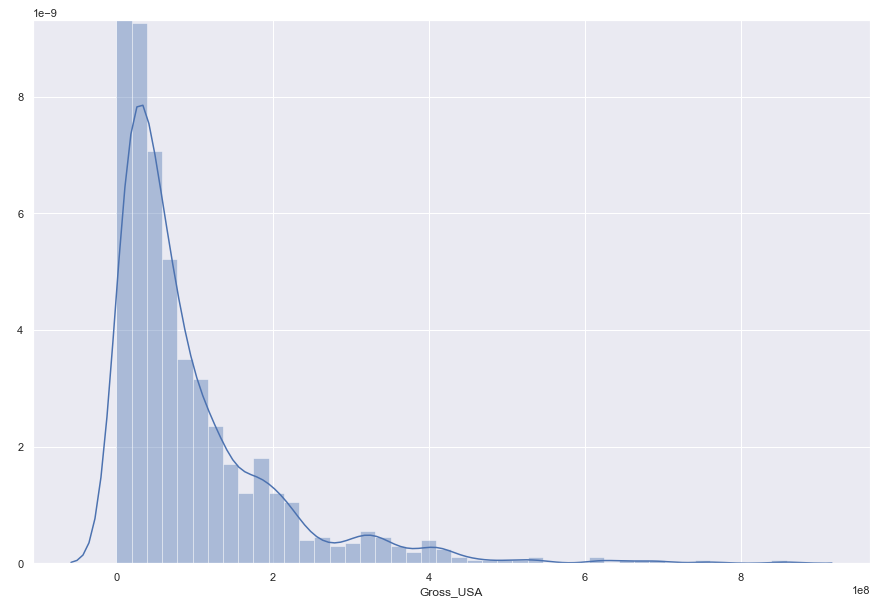
Yearly budget trend certification wise:



Monthly trend of Gross\_USA :



Distribution of Gross\_USA:



**How did you transform your data for the tools you used?**

After scraping the data from the website, there were lots of missing values, garbage values, wrong data types and there was requirement to transforming data. Also, we had details of movies which were very old and did not provide good relationship financially.

Thus, we did the following data cleaning steps:

1.Select movies released on and after the year 2000, this will help us reduce the effect of inflation.

#Remove observations with "-" in title, Gross\_USA and budget. Selecting USA movies

2. Remove "," and "$" characters from the Multiple columns.

3. Converting variables to required datatypes.

4. Converting the duration of movies from hr,min fromat to mins and then placing them as numeric.

# eg: "1h 48min" --> 108

5. Converting the dates of movies from to months and then placing them as numeric form.

# eg: "01 January 2011" --> 1

6. we can see that from the above chart that certifications "G" and "Not rated" amount for 40 count out of total observations. Thus, we can exclude it as it forms just 4% of the data.

Data was also scaled (especially the financial data) and checked the model performance but there was no substantial effect in model performance, thus a decision was made to use to original data

**How did you use your tools?**

After cleaning and exploring the data as discussed above, I started by creating a **Base Linear model** using **SKLearn** package. This is the package used to design all machine learning package though SciPy and stats models’ packages were also explored.

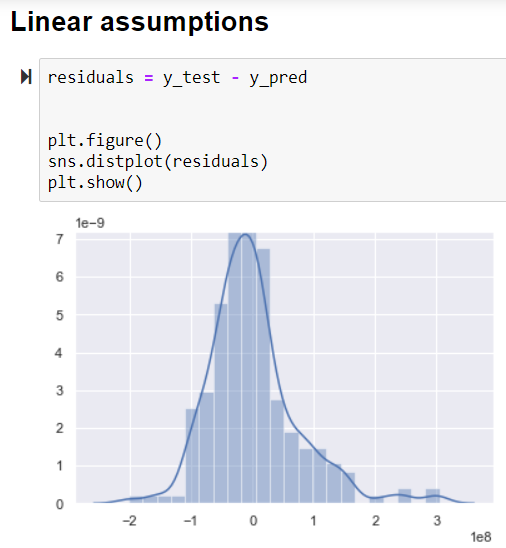
All categorical variables were converted to dummies before designing the models resulting in 17 predictor variables in total. Using train-test(validation) method on base model I got  **= 0.5423 and MAPE = 2360.151.**

Checking for Assumptions of linear regression:

1. The residual terms, {ε}, have a mean value of 0.

2. The residual terms are homoscedastic; they have constant variance, this means the standard deviation of residuals associated with one value of X equals the standard deviation of the residuals for any other value of X.

5. For any value of X, the residuals associated with that value of X follow the Normal Distribution with mean 0 and standard deviation



Furthering the Base LM model, methods of **Ridge and PCA** were applied to regularize and reduce variables, respectively where the **MAPE achieved was around 2000 and 1476**. The option of **polynomial LM** was also explore but did not provide better insights in the output ( **= 0.54599 and MAPE = 2859.491)**.

**How are you going to share these results with us?**

Results are going to be shared in the following manner:

1. Discussing the Business Question and giving insights into data.
2. Overview of data cleaning, web scrapping and analysis.
3. Visualization.
4. Explaining Output with test cases.

Prediction is going to be made of few movies which have released and yet to release. We can check how are model is doing as well as predict for future release.

We will also make recommendations on few strategies to boost revenues.

$$ How much will Wonder Woman 1984 Earn in the US $$

Budget: 175000000

Runtime: 110

Certification: PG-13

Release date: 14 August 2020

Gross USA Revenue: ???????????