Northeastern University
College *of* Professional Studies

ALY 6040 Report – Finding patterns and performing EDA

Group Project

By –

Dishali Sonawane

Mrigank

Pranav Khanna

Swapnil Lokhande

Sakshi Dalicha

ALY6040 Data Mining Applications
SPRING 2019 CPS

Instructor: Justin Grosz

Due Date: June 2, 2019

Problem Statement

With the advancement in technology and cinematography, the film industry is growing rapidly and made huge business in last few years. According to reports, the American film industry generated $43.4 billion in revenue last year, increasing in each of the past five years at an annualized rate of just 2.2%. Thus, it is important to understand the factors behind the success of each movie; whether a huge star cast, a good director, the huge budget or is there any other factor behind the successful box-office story. In order to better understand the factors affecting the success of the movie, we intent to make predictive model to estimate the revenue of the movie released using different attributes.

Data Description

The data contain two different files:

Test raw data –	Objects: 4356, variables: 22
Train raw data -	**Objects: 3000, variables: 23**

Variables that are used in the Movie Dataset (raw)

Name	Definition	Variable
id	Unique movie ID	int
belongs_to_collection	The collections to which the movie	object

budget	The budget for the movie.	float
genres	The movie's genre.	object
homepage	The movie's homepage.	object
imdb_id	Movie’s IMDB ID	object
original_language	The original language of the movie	object
original_title	The original title of the movie	object
overview	A brief plot	object
popularity	An index of movie’s popularity	float
poster_path	The poster of the movie	object
production_companies	The producer companies	object
production_countries	Countries involved in the production	object
release_date	Release date of the movie	object
runtime	The length of the movie	float
spoken_languages	The language spoken in the movie	object
status	Released or rumored	object
tagline	Tagline of the movie	object
title	Title of the movie	object
Keywords	Keyword used	object
cast	Details of the cast involved	object
crew	Details of the crew involved	object

1. **Exploratory Data Analysis**

   We used descriptive statistics to understand and visualize the key insights as well as any anomalous behavior present in the data which can be rectified during Data Processing. Following are the key findings obtained upon analyzing the dataset:

- The train and test dataset have 3000 and 4398 number of entries respectively and have 24 and 23 attributes (columns) in the dataset.

In order to perform the exploratory analysis, we combined the test, train datasets together.

Missing Values:

- Upon analysis, it was observed that columns ‘belongs_to_collection’ and ‘homepage’ have more than 50% of missing value.
- The other columns which have missing values are keywords (9%), cast (0.35%), crew(0.51%), genres (0.31%), overview(0.29%), poster_path (0.02%), production_companies (5.5%), production_countries (2.12%), release_date (0.013%), runtime (0.08%), spoken_languages (0.83%), status (0.02%), tagline (19.73%) and title (0.04%).
- Other columns, 'id', 'imdb_id', 'original_title', 'overview', 'popularity', 'status', 'tagline', 'title' have been dropped since it has been assumed that these parameters would not have significance impact on the sales.

Columns Dropped:	Reason
belongs_to_collection	More than 50 percent missing values, similar attribute available in the data, does not add statistical significance to the analysis.

homepage	More than 50 percent missing values, does not add statistical significance to the analysis.
imdb_id	Does not add statistical significance to the analysis.
original_title	Does not add statistical significance to the analysis.
overview	Does not add statistical significance to the analysis.
popularity	Does not add statistical significance to the analysis.
poster_path	Does not add statistical significance to the analysis.
status	Does not add statistical significance to the analysis.
tagline	Does not add statistical significance to the analysis.

2. Data Pre-Processing

The raw column data consisted of dictionary objects. In order to extract the useful information which is stored in the form of key value pair we cleaned the data in R. The R environment is not capable of reading and processing data in form of dictionary, so we convert the objects to character vectors for the further cleaning of data.

Techniques used for data cleaning (feature engineering)

Since most of the attributes were in string form, we made use of gsub function to replace all the punctuations and the character they we do not need with black space and only extract the required part. We also used corpus for getting rid of non-alphanumeric characters.

The following attributes were cleaned using the above techniques

- **Production_companies**
- **Production_countries**
- **Genres**
- **Spoken languages**
- **Keyword_count**

Attribute Genres consisted of list of type of movie genres, we use DocumentTermMatrix function to find frequency of each genres in the data. Next, we use binded the columns to the data.

```
Genres =  "action"      +   "adventure" +
          "animation"   +   "comedy" +
          "crime"       +    "documentary" +
          "drama"       +     "family" +
          "fantasy"     +    "fiction" +
          "foreign"     +    "history" +
          "horror"      +   "movie" +
          "music"       +    "mystery" +
          "romance"     +    "science" +
          "thriller"    +    "war" +
          "western"
```

Similarly for the spoken_languages column is used, I distributed the spoken languages into different type of langauges unsing DocumentTermMatrix.

```
Spoken_languages = "deutsch"     +
                   "english"     +     "espaã±ol"    +
                   "español"     +      "franã§ais"   +
```

"français" + "italiano" +
"kiswali" + "latin" +
"magy" + "nernds" +
"norsk" + "polski" +
"portuguãªs" + "português" +
"romã¢näƒ" + "român" +
"shqip" + "ska" +
"slovina" + "somali" +
"srpski" + "suomi" +

The columns that contained date like **release_date** were also in the form of objects, so as to extract the date we used the as.Date function, it converts the objects representation into date like format. Then we split each column into respective day-weekday-month - year

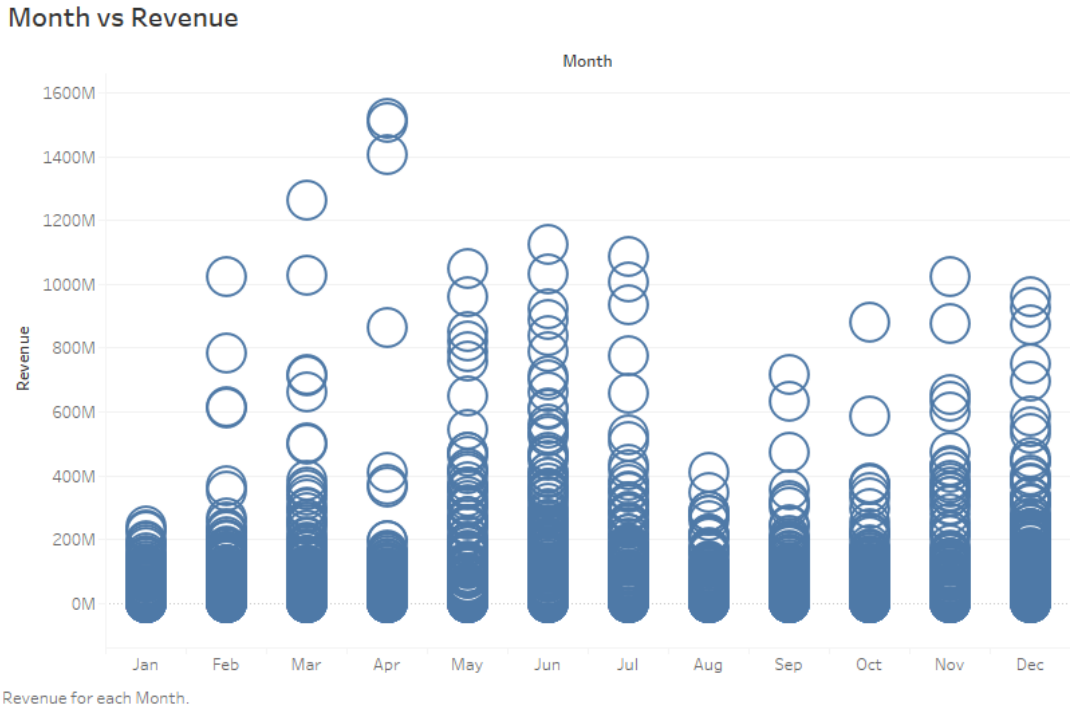
Release_date = "**month**" + "**year**" + "**day**" + "**weekday**"

When a movie is searched online, keywords make a greater impact on the revenue. We derived a column named **keyword_count**, that specifies the number of Key_words related to the movie.

Final Data:

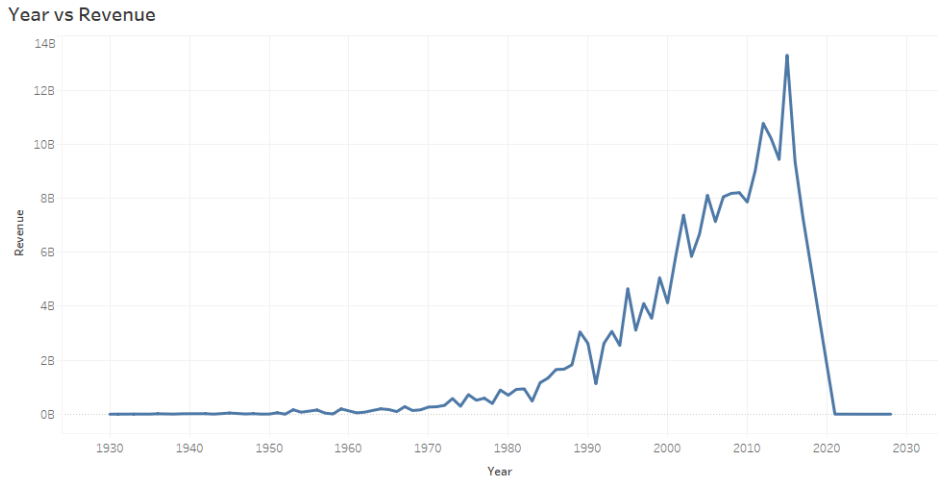
Total attributes: 57
Total data size: 7398

Data visualization



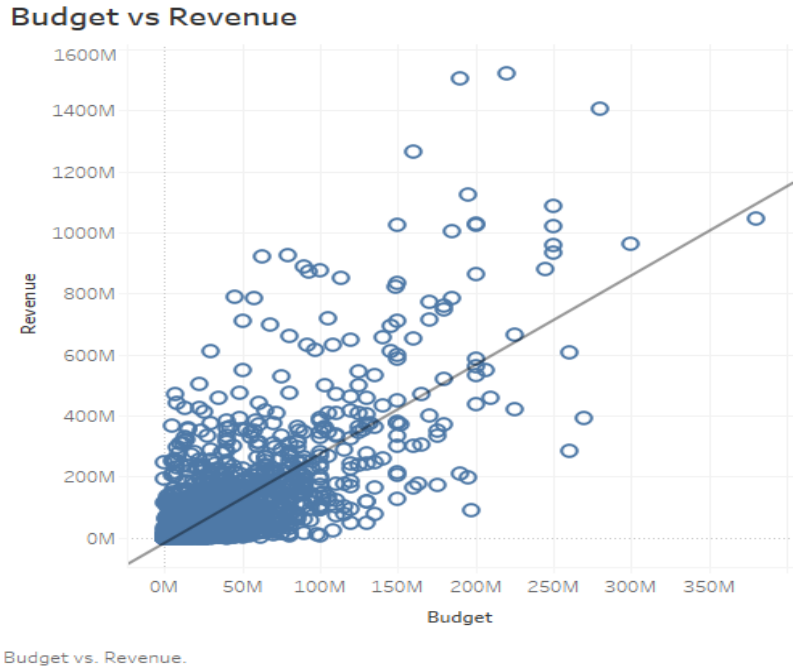
Inference:

This graph illustrates the monthly revenue collected we can observe that the highest revenue collected was in the month of April, the months of May, June, July, December have a significant increase in revenue collected as compared to other months. This might be because of the vacation periods most of the people are enjoying the movies.



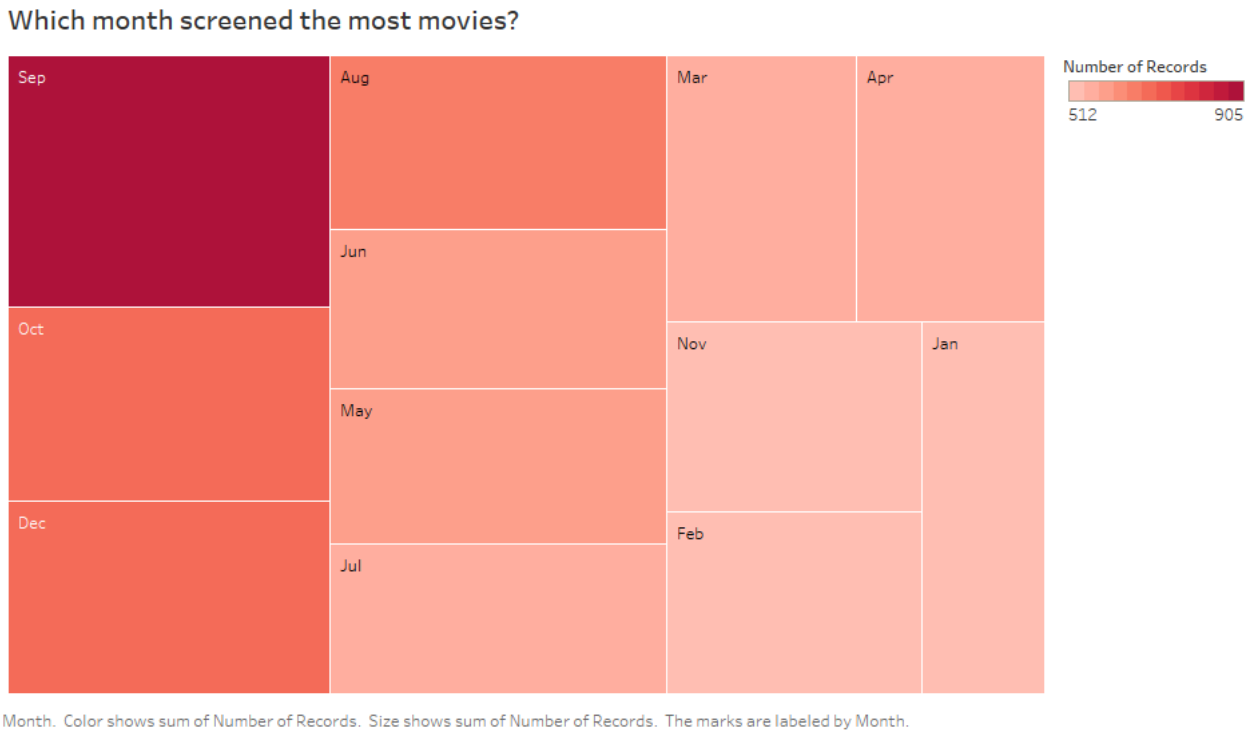
Inference:

As we can see the revenue is increasing every year as a greater number of movies are releasing.



Inference:

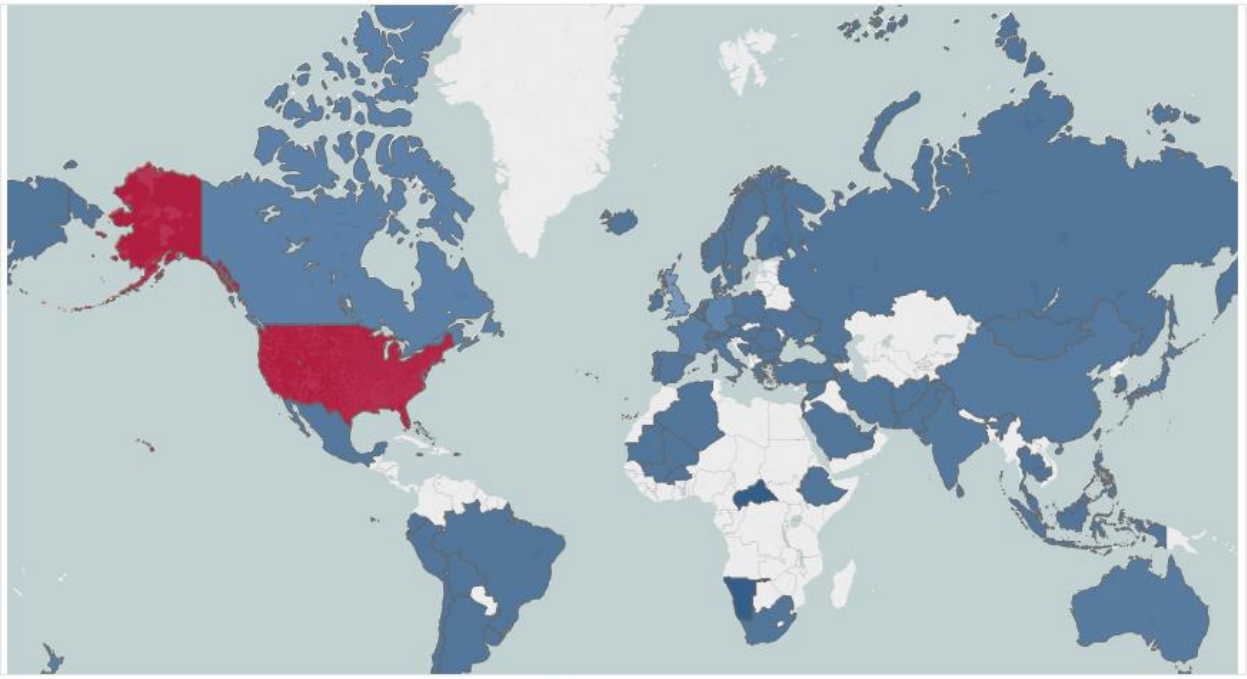
This graph describes the budget versus the revenue, both the attributes are highly correlated. The budget attribute is having a good impact in the revenue.



Inference:

The months September, October, December and August were the months when movie released generated the most revenue. This information will be vital in our further analysis.

Which countries generate the most revenue?



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Revenue. Size shows details about Production Countries. Map coloring shows 2018 Male/Female Ratio by Zip Code. The view is filtered on Longitude (generated) and Latitude (generated). The Longitude (generated) filter ranges from -111.2 to 172.4. The Latitude (generated) filter ranges from -42.6 to 66.2.

Inference:

We can clearly infer from the chart that movie made in USA generate the most revenue. There can be many reasons behind this, one of the most critical reason can be that most big production companies are present in the north America.

Conclusion

In this week assignment we were able to find important attribute for the analysis. We use different functions to find patterns in the data and extract meaningful information from the data. Next, we were able to create new dimensions from the existing columns, which will help us in our analyzing, modelling next coming weeks.

Also, we were able to further understand the relationship among distinct attributes like budget, Production countries, month of movie release, year of movie release with the target revenue attribute. This information will be valuable to us for generating the appropriate predictive model for the data.

References

Kaggle. TMDB Box office Prediction. Retrieved from https://www.kaggle.com/c/tmdb-box-office-prediction

Donges, N. (2018, Feb 22). The Random Forest Algorithm. Retrieved from https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Jha, V. (2017, June 18). Decision Tree Algorithm for a predictive model. Retrieved from https://www.techleer.com/articles/120-decision-tree-algorithm-for-a-predictive-model/

Microsoft Azure. (2017, Nov 03). The Team Data Science Process Lifecycle. Retrieved from https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

Statistics Solution. (2013). What is Linear Regression. Retrieved from https://www.statisticssolutions.com/what-is-linear-regression/