Dalhousie University

CSCI5408 Data Management and Warehousing Analytics

Project Report

PROJECT REPORT

Submitted By :
Deeksha Behara :B00784704
Tushar Gupta :B00782699

Supervisor:
Suhaib Qaiser

August 5, 2018

Abstract

Movies are the most sought-after source of entertainment. In a recent survey, it was observed that the entertainment sector is the most profitable and a majority of the investments are made in it. People make investments in this sector because of the possibility of the shared risks, personal satisfaction, the range of investment levels and possible long-term payoffs. Later in early 2000, there was a casual drift from watching the movies in the theaters to watching the movies online due to the upcoming of media service providers such as Netflix. People began binge-watching since these media service providers make user customized suggestions based on the history of user's interests. Our project performs descriptive and predictive analysis of the movie data but on a smaller scale.(GitHub link : https://github.com/bio33/dwh-project)

Contents

1 Introduction

A survey conducted by Adobe reveals that more people are watching TV online more than before. Though the trend has not completely shifted from watching movies offline, but the growth in the number of people who prefer to watch TV over watching movies using the online streaming service is increasing. This could be because of the surge of the media/streaming service providers such as Netflix. Researchers at Adobe have tracked 1.53 billion logins over a year and observed that TV viewing over the internet grew by 338 percent in the mid 2014 compared to the same time a year earlier. The number of unique viewers have more than doubled growing 146 percent in a year. All this is because of the services that these streaming service providers offer. These service providers know what you want to watch online before you actually do. Netflix, for example uses comprehensive recommendation system which performs extensive data analysis, uses various predictive algorithms and number crunching techniques to make suggestions based on the historic data. The historic data in this case is the users interest. (GitHub link : https://github.com/bio33/dwh-project)

2 Idea Summarization

Our project has similar functions to what Netflix does except to a smaller scale. We focused on performing descriptive and predictive analysis on the data retrieved from Kaggle data store. Our project focuses on solving the defined problem statement using two of these techniques.

2.1 Problem Statement

To emulate a recommendation system similar to that of Netlfix. With the taken dataset we would extract information that would help us to understand what factors precisely matter in affecting the rating of the movie. Our project is divided into two parts .

First making a Machine Learning model which would train based on the selected datasets. The model would be predicting if a movie would be a hit or not based on the attributes given.

Secondly finding features which could affect the rating of the movie. We have taken an hypothesis that movie genre and their release data combined have an affect on movie rating. We would be testing if this hypothesis is or not based on statistical and visual inference methods.

2.1.1 Challenges faced

Initially we faced quite a few challenges before we began with the project , some of which were:

1. The dataset was dispersed into 3 different CSV's each containing different set of features. Combining them based on keys and extracting them all in one dataset was the first issue we faced.

2. While the columns contained data in improper json format reading the data from each column was another issue which we solved using literal eval functionality in python To capture any pattern between movie release date and genre of period of time we had to experiment with lots of visualization before we could settle for one.

2.1.2 Steps to address

We have identified smaller problems in order to provide solution to the objectives of the project. The subset of problems are defined below.

1. Predictive Analysis

   (a) Pro-processing the data and preparing it for the model
   (b) Converting the data to their numerical equivalents for training the classifier
   (c) Splitting the dataset into training and Testing Dataset in 3:2 ratio
   (d) Using the training dataset to train the model
   (e) using the trained/intelligent model to predict the labels for the test data
   (f) Comparing Different models for their accuracies

(g) Deriving Confusion matrix

(h) Visualizing the results

2. Descriptive Analysis:

(a) Extracting features from the dataset.

(b) Generating word cloud from genre count

(c) Top 10 movie extractor based on movie profits

(d) Top 10 movies by a user given artist based on the revenue generated by their movies.

(e) Converting genre into features using count vectorizer .

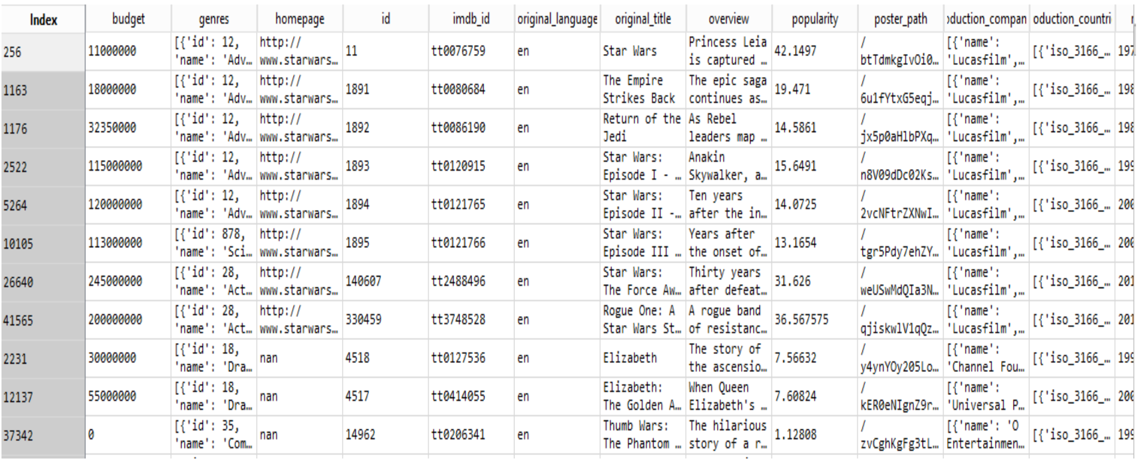
(f) Applying feature selection to select best features that contribute to a movie's success.

(g) Binning the movie ratings into smaller sections as well as the date of release into 4 quarters of the year to gain more accuracy on the classifier.

(h) Applying visualizations to find pattern for each genre.

2.2 Data Sources

The Dataset[1] was retrieved from Kaggle. The retrieved dataset consists of several features dispersed over several csvs. The data was present in a relational format and had features such as movie name, description, cast crew, budget, revenue, movie overview and several other important features. The dispersed data was brought to a common ground by merging the datasets using the similar movie_id. The features to be considered were detected by exploratory data analysis. The important attributes which contributed to the building the solution like movie review, cast ,crew and few others were retained in the dataset while the others which did not contribute were discarded.



Index	budget	genres	homepage	id	imdb_id	original_language	original_title	overview	popularity	poster_path	›duction_compan	oduction_countri	r
256	11000000	[{'id': 12, 'name': 'Adv…	http:// www.starwars…	11	tt0076759	en	Star Wars	Princess Leia is captured …	42.1497	/ btTdmkgIvOi0…	[{'name': 'Lucasfilm',…	[{'iso_3166_…	197
1163	18000000	[{'id': 12, 'name': 'Adv…	http:// www.starwars…	1891	tt0080684	en	The Empire Strikes Back	The epic saga continues as…	19.471	/ 6u1fYtxG5eqj…	[{'name': 'Lucasfilm',…	[{'iso_3166_…	198
1176	32350000	[{'id': 12, 'name': 'Adv…	http:// www.starwars…	1892	tt0086190	en	Return of the Jedi	As Rebel leaders map …	14.5861	/ jx5p0aH1bPXq…	[{'name': 'Lucasfilm',…	[{'iso_3166_…	198
2522	115000000	[{'id': 12, 'name': 'Adv…	http:// www.starwars…	1893	tt0120915	en	Star Wars: Episode I - …	Anakin Skywalker, a…	15.6491	/ n8V09dDc02Ks…	[{'name': 'Lucasfilm',…	[{'iso_3166_…	199
5264	120000000	[{'id': 12, 'name': 'Adv…	http:// www.starwars…	1894	tt0121765	en	Star Wars: Episode II -…	Ten years after the in…	14.0725	/ 2vcNFtrZXNwI…	[{'name': 'Lucasfilm',…	[{'iso_3166_…	200
10105	113000000	[{'id': 878, 'name': 'Sci…	http:// www.starwars…	1895	tt0121766	en	Star Wars: Episode III …	Years after the onset of…	13.1654	/ tgr5Pdy7ehZY…	[{'name': 'Lucasfilm',…	[{'iso_3166_…	200
26640	245000000	[{'id': 28, 'name': 'Act…	http:// www.starwars…	140607	tt2488496	en	Star Wars: The Force Aw…	Thirty years after defeat…	31.626	/ weUSwMdQIa3N…	[{'name': 'Lucasfilm',…	[{'iso_3166_…	201
41565	200000000	[{'id': 28, 'name': 'Act…	http:// www.starwars…	330459	tt3748528	en	Rogue One: A Star Wars St…	A rogue band of resistanc…	36.567575	/ qjiskwlV1qQz…	[{'name': 'Lucasfilm',…	[{'iso_3166_…	201
2231	30000000	[{'id': 18, 'name': 'Dra…	nan	4518	tt0127536	en	Elizabeth	The story of the ascensio…	7.56632	/ y4ynYOy205Lo…	[{'name': 'Channel Fou…	[{'iso_3166_…	199
12137	55000000	[{'id': 18, 'name': 'Dra…	nan	4517	tt0414055	en	Elizabeth: The Golden A…	When Queen Elizabeth's …	7.60824	/ kER0eNIgnZ9r…	[{'name': 'Universal P…	[{'iso_3166_…	200
37342	0	[{'id': 35, 'name': 'Com…	nan	14962	tt0206341	en	Thumb Wars: The Phantom …	The hilarious story of a r…	1.12808	/ zvCghKgFg3tL…	[{'name': 'O Entertainmen…	[{'iso_3166_…	199

Figure 1: screenshot of the dataset whoch contains the associated attributes.

The dataset consists of 26 columns with approximately forty five thousand sets of records.

2.3 Algorithms and Classifiers Used.

We started off with the data preprocessing and preparing the dataset for the machine learning. The data preprocessing followed the following steps:

1. Removing punctuations and white spaces
2. Removing nulls
3. Removing hyperlinks and emoticons from the text
4. Dropping columns which are not required.
5. Converting the required string values to floats.

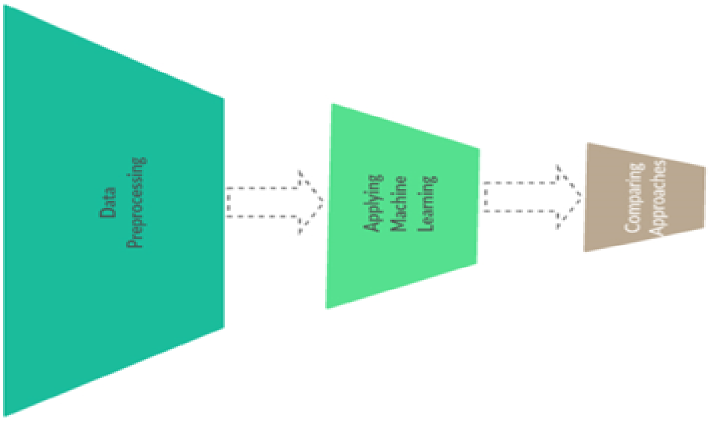


Figure 2: An over view of the approach followed by us.

The success of the movie was marked as either success or failure based on the following calculation.

Formula: (if ((revenue - budget) $>$ budget/3) ? success : failure)

Based on the formula mentioned above , the movie were classified either as a success or failure. After classifying the data and the data was converted to its numerical equivalents using label encoder. We were curious about the performance of the classifiers and trained several classifiers to determine which would be the best fit for the defined problem. We decided to choose the Random Forest , SVM , Logistic Regression , Naive Bayes and the decision tree classifier to review each of the classifier's performance.

1. Random Forest Random forest algorithm is an ensemble learning method for classification, regression and other tasks. It constructs multitude of decision trees at the training time and outputting the prediction of the desired classes.
2. SVM Support Vector Machine is a discriminative classifier which uses hyperplanes for
3. Logistic Regression This classifier helps predict the odds of an event happening based on the values of the independent variables.
4. Naive Bayes The Naive Bayes classifier uses the bayesian theorem and works using a simple rule.
5. Decision Tree This classifier follows a top down approach that partitions the data into subsets which consists of similar values.

2.3.1 Comparison

The accuracies for each of the classifiers were obtained and are depicted in the table below.

Classifier	Accuracy
Random Forest	0.9756
SVM	0.884
Logistic Regression	0.928
Naive Baye's	0.815
Decision Tree	0.9340

Table 1: Accuracies obtained using different classifiers are depicted in the table above.

Confusion matrix for Random Forest : [[2059 67][377 15713]]
Confusion matrix for Linear Regression : [[1235 891][404 15686]]
Confusion matrix for Multinomial Naive Bayes : [[1834 292] [3077 13013]]
Confusion matrix for Decision Tree : [[1861 265] [279 15811]]
Confusion matrix for Support Vector Machine : [[13 2113] [0 16090]]

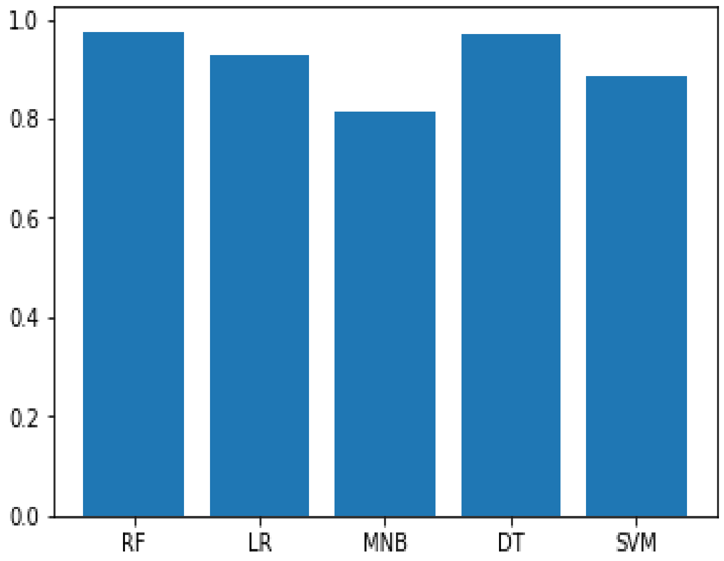


Figure 3: The graph above depicts the accuracies for the classifiers used

From the graph above it is observed that Random Forest classifier performs the best with 97.5 accuracy. The next best algorithm is decision tree with 97 percent accuracy and finally there were three classifiers which were Logistic Regression, Multinomial Naive Bayes and Support Vector Machine which were next in line. This concludes that the Random Forest Classifier performs the best

3 Visualization

There were multiple visualization used out of which the best and relevant were used to analyze patterns for our hypothesis.

After extracting data for each movie and their respective genre and release date grouped by months and year of release we plotted graphs for each genre to capture any patterns over period of time. The graphs we have plotted are timeline graphs with Years on the x-axis and movie rating on the y-axis, the color coding represents the month of the year the movie is being release in. Thus lighter shades are starting of the year while darker shader are towards the end of the year.

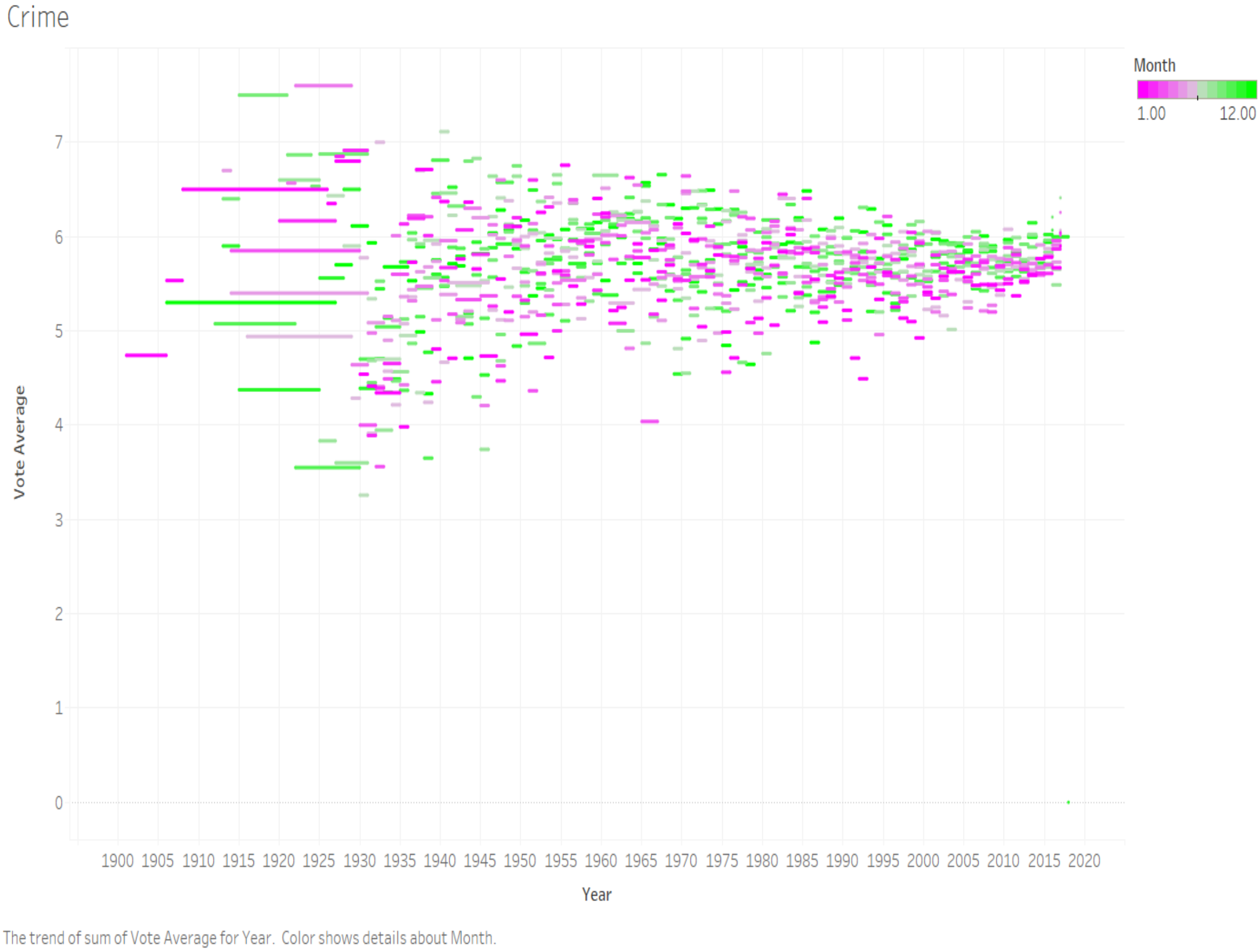


Figure 4: Crime genre movies performance over the years.

Inference : Crime movies have always performed better when released in the later part of the year than when released in early part of the year. Although the count of crime movies being release in the early part of the year has been increasing in last decade indicating time of release is not considered as a factor in respect to what kind of genre the movie has. Another inference We can draw from this visualization is that over the year variance in movie ratings has decreased and has shrunk down to the range of [6.2 - 5] for crime movies

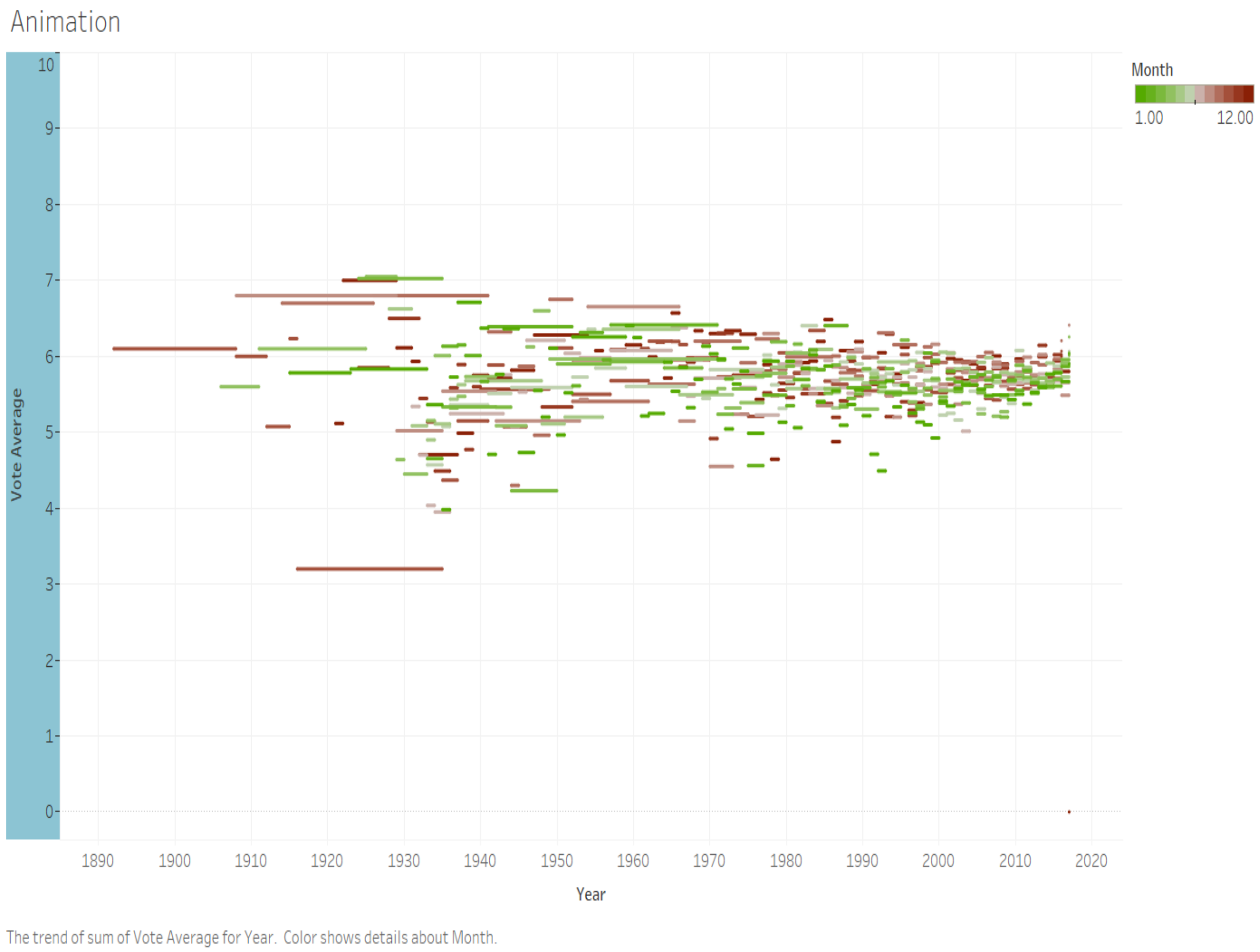


Figure 5: Animation genre movies performance over the years.

Inference : Animated movies have recently gained much poupularity due to production housed like Pixar which have revolutionized animation industry. We can see that from how dense the chart gets after 2000 since that is shortly after the launch of Pixar's first animated film Toy Story. It can also be observed on how movies are generally released in the first and second quarter of the year which is holiday time for children which is generally the target audience for such movies. In the last decade animated movie standards have been set really high hence low budget production houses are struggling to keep up , possibly the reason for the dip in the graph in the past years.The outliers are clearly visible which are mostly pixar movies release in the month of November like: Toy Story 1,2 , Coco , The Incredibles etc.

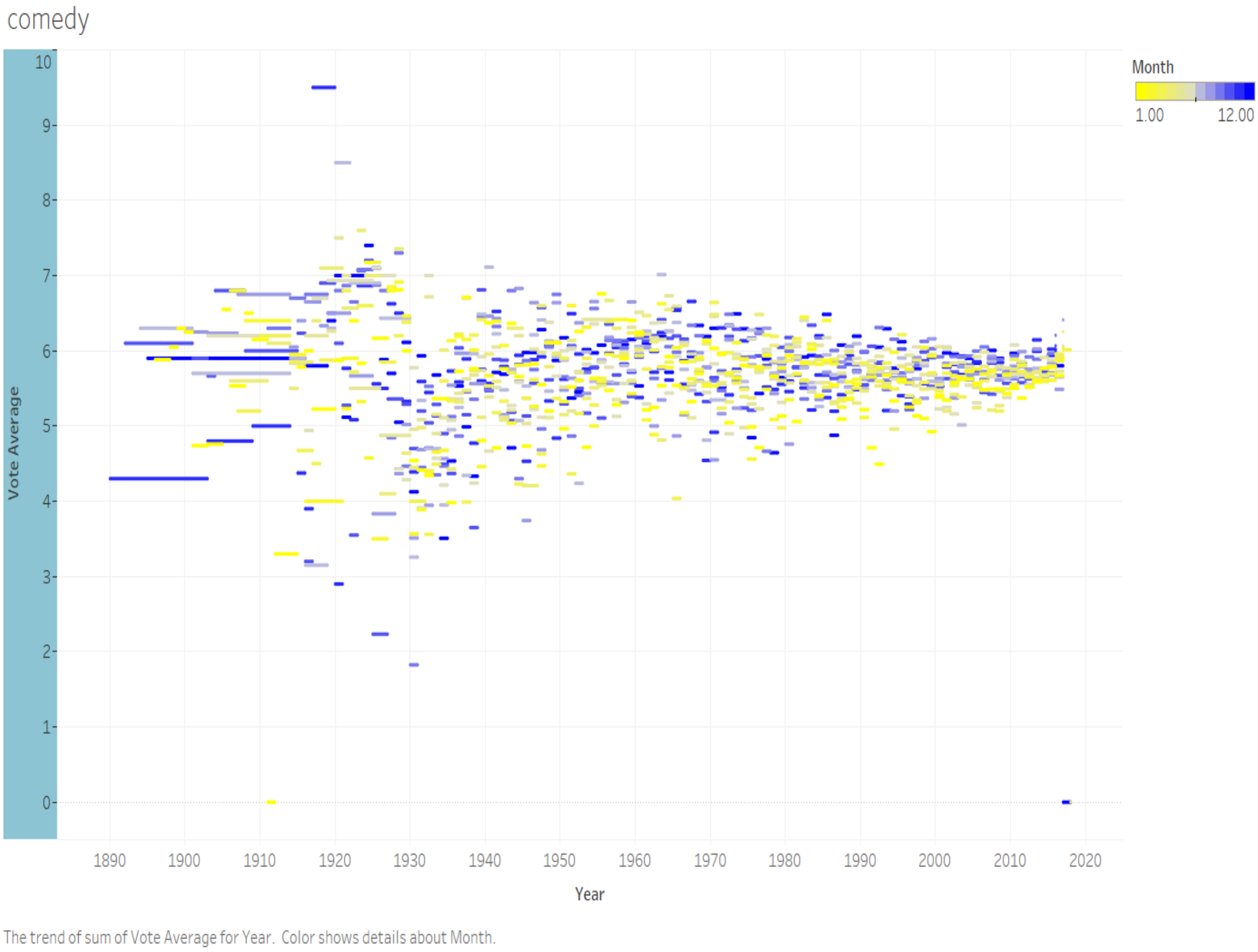
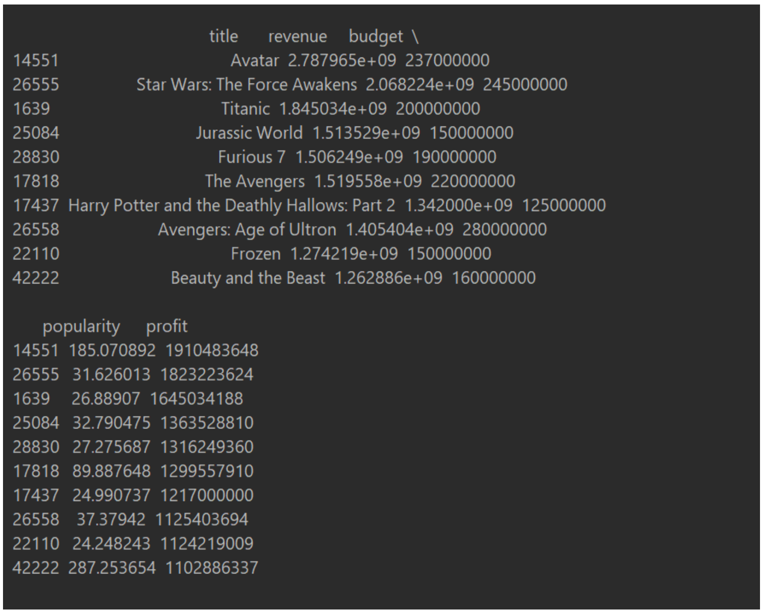


Figure 6: Comedy genre movies performance over the years.

Inference: Comedy movies have been made since the early 1800's with revolutionary artists like Charlie Chapplin setting the standards of comedy movies as high it can get. Which we can observe from the graph comedy movies in 1800's have averaged around 7 on movie rating bar. Comedy movies in released in the later part of the year although less in quantity have performed well than movies launched in early part of the year. Movie Ratings have strongly converged towards the value of 6 in recent years indicating saturation in content of comedy movies. Comedy movies are alone not doing great but action movies when combined with comedy have a better chance of doing well.

4 Sprint Reviews

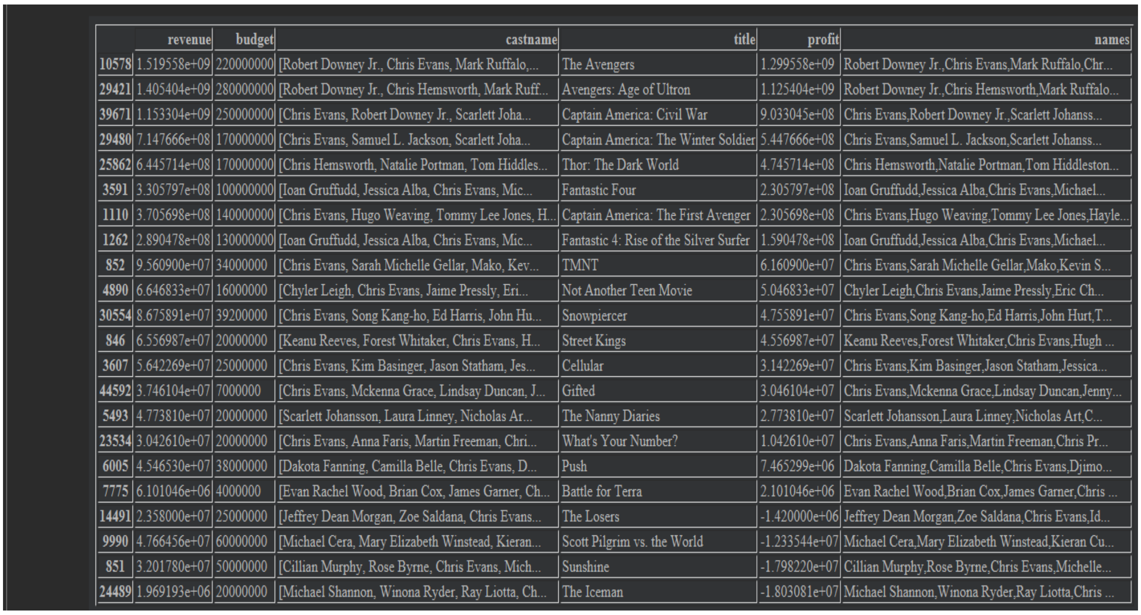
Our project was divided into 4 Sprints each spring having its own goals. First spring was dedicated in understanding the fields of dataset and what we can do with them, setting up the environment , and data-preprocessing. Second sprint was dedicated in applying Exploratory Data Analysis techniques (EDA) to understand the dataset mathematically like looking at count, variance , how to impede missing values, reason for any outliers, making small tools and visualization to get comfortable with dataset.For example Top 10 movies from the dataset :



```
                                   title       revenue     budget  \
14551                               Avatar  2.787965e+09  237000000
26555          Star Wars: The Force Awakens  2.068224e+09  245000000
1639                               Titanic  1.845034e+09  200000000
25084                       Jurassic World  1.513529e+09  150000000
28830                            Furious 7  1.506249e+09  190000000
17818                         The Avengers  1.519558e+09  220000000
17437  Harry Potter and the Deathly Hallows: Part 2  1.342000e+09  125000000
26558              Avengers: Age of Ultron  1.405404e+09  280000000
22110                               Frozen  1.274219e+09  150000000
42222                 Beauty and the Beast  1.262886e+09  160000000

       popularity      profit
14551  185.070892  1910483648
26555   31.626013  1823223624
1639     26.88907  1645034188
25084   32.790475  1363528810
28830   27.275687  1316249360
17818   89.887648  1299557910
17437   24.990737  1217000000
26558    37.37942  1125403694
22110   24.248243  1124219009
42222  287.253654  1102886337
```

Figure 7: Top 10 movies based on profits



	revenue	budget	castname	title	profit	names
10578	1.519558e+09	220000000	[Robert Downey Jr., Chris Evans, Mark Ruffalo,...	The Avengers	1.299558e+09	Robert Downey Jr.,Chris Evans,Mark Ruffalo,Chr...
29421	1.405404e+09	280000000	[Robert Downey Jr., Chris Hemsworth, Mark Ruff...	Avengers: Age of Ultron	1.125404e+09	Robert Downey Jr.,Chris Hemsworth,Mark Ruffalo...
39671	1.153304e+09	250000000	[Chris Evans, Robert Downey Jr., Scarlett Joha...	Captain America: Civil War	9.033045e+08	Chris Evans,Robert Downey Jr.,Scarlett Johanss...
29480	7.147666e+08	170000000	[Chris Evans, Samuel L. Jackson, Scarlett Joha...	Captain America: The Winter Soldier	5.447666e+08	Chris Evans,Samuel L. Jackson,Scarlett Johanss...
25862	6.445714e+08	170000000	[Chris Hemsworth, Natalie Portman, Tom Hiddles...	Thor: The Dark World	4.745714e+08	Chris Hemsworth,Natalie Portman,Tom Hiddleston...
3591	3.305797e+08	100000000	[Ioan Gruffudd, Jessica Alba, Chris Evans, Mic...	Fantastic Four	2.305797e+08	Ioan Gruffudd,Jessica Alba,Chris Evans,Michael...
1110	3.705698e+08	140000000	[Chris Evans, Hugo Weaving, Tommy Lee Jones, H...	Captain America: The First Avenger	2.305698e+08	Chris Evans,Hugo Weaving,Tommy Lee Jones,Hayle...
1262	2.890478e+08	130000000	[Ioan Gruffudd, Jessica Alba, Chris Evans, Mic...	Fantastic 4: Rise of the Silver Surfer	1.590478e+08	Ioan Gruffudd,Jessica Alba,Chris Evans,Michael...
852	9.560900e+07	34000000	[Chris Evans, Sarah Michelle Gellar, Mako, Kev...	TMNT	6.160900e+07	Chris Evans,Sarah Michelle Gellar,Mako,Kevin S...
4890	6.646833e+07	16000000	[Chyler Leigh, Chris Evans, Jaime Pressly, Eri...	Not Another Teen Movie	5.046833e+07	Chyler Leigh,Chris Evans,Jaime Pressly,Eric Ch...
30554	8.675891e+07	39200000	[Chris Evans, Song Kang-ho, Ed Harris, John Hu...	Snowpiercer	4.755891e+07	Chris Evans,Song Kang-ho,Ed Harris,John Hurt,T...
846	6.556987e+07	20000000	[Keanu Reeves, Forest Whitaker, Chris Evans, H...	Street Kings	4.556987e+07	Keanu Reeves,Forest Whitaker,Chris Evans,Hugh ...
3607	5.642269e+07	25000000	[Chris Evans, Kim Basinger, Jason Statham, Jes...	Cellular	3.142269e+07	Chris Evans,Kim Basinger,Jason Statham,Jessica...
44592	3.746104e+07	7000000	[Chris Evans, Mckenna Grace, Lindsay Duncan, J...	Gifted	3.046104e+07	Chris Evans,Mckenna Grace,Lindsay Duncan,Jenny...
5493	4.773810e+07	20000000	[Scarlett Johansson, Laura Linney, Nicholas Ar...	The Nanny Diaries	2.773810e+07	Scarlett Johansson,Laura Linney,Nicholas Art,C...
23534	3.042610e+07	20000000	[Chris Evans, Anna Faris, Martin Freeman, Chri...	What's Your Number?	1.042610e+07	Chris Evans,Anna Faris,Martin Freeman,Chris Pr...
6005	4.546530e+07	38000000	[Dakota Fanning, Camilla Belle, Chris Evans, D...	Push	7.465299e+06	Dakota Fanning,Camilla Belle,Chris Evans,Djimo...
7775	6.101046e+06	4000000	[Evan Rachel Wood, Brian Cox, James Garner, Ch...	Battle for Terra	2.101046e+06	Evan Rachel Wood,Brian Cox,James Garner,Chris ...
14491	2.358000e+07	25000000	[Jeffrey Dean Morgan, Zoe Saldana, Chris Evans...	The Losers	-1.420000e+06	Jeffrey Dean Morgan,Zoe Saldana,Chris Evans,Id...
9990	4.766456e+07	60000000	[Michael Cera, Mary Elizabeth Winstead, Kieran...	Scott Pilgrim vs. the World	-1.233544e+07	Michael Cera,Mary Elizabeth Winstead,Kieran Cu...
851	3.201780e+07	50000000	[Cillian Murphy, Rose Byrne, Chris Evans, Mich...	Sunshine	-1.798220e+07	Cillian Murphy,Rose Byrne,Chris Evans,Michelle...
24489	1.969193e+06	20000000	[Michael Shannon, Winona Ryder, Ray Liotta, Ch...	The Iceman	-1.803081e+07	Michael Shannon,Winona Ryder,Ray Liotta,Chris ...

Figure 8: Top 10 movies of Scarelett Johansson based on profits

The largest among all sprint was Sprint 3 we aimed to visualize our finding in this sprints, it took more time than expected since we weren't sure how to visualize and what we exactly wanted to visualize , we had to empirically to select our visualizations to get to the one which fits our findings. Spring 4 was to clean the code, test it for any errors and finalize and document our findings into one final report.

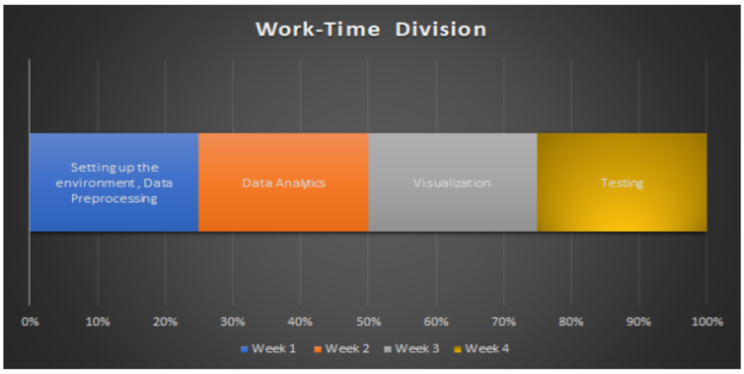
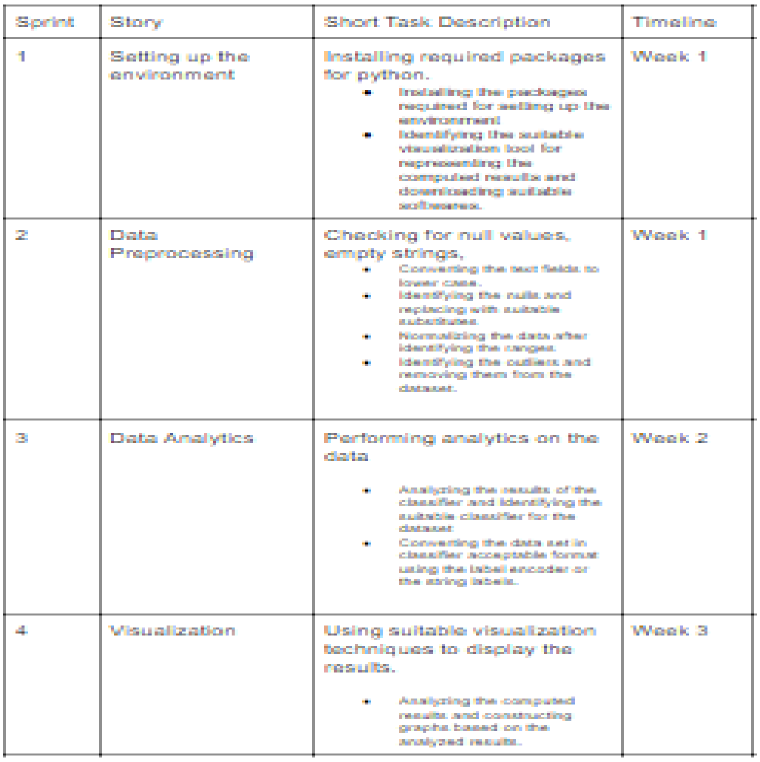


Figure 9: An overview of work breakdown for each Sprint



Sprint	Story	Short Task Description	Timeline
1	Setting up the environment	Installing required packages for python. • Installing the packages required for setting up the environment • Identifying the suitable visualization tool for representing the computed results and downloading suitable softwares.	Week 1
2	Data Preprocessing	Checking for null values, empty strings, • Converting the text fields to lower case. • Identifying the nulls and replacing with suitable substitutes. • Normalizing the data after identifying the ranges. • Identifying the outliers and removing them from the dataset.	Week 1
3	Data Analytics	Performing analytics on the data • Analyzing the results of the classifier and identifying the suitable classifier for the dataset • Converting the data set in classifier acceptable format using the label encoder or the string labels.	Week 2
4	Visualization	Using suitable visualization techniques to display the results. • Analyzing the computed results and constructing graphs based on the analyzed results.	Week 3

Figure 10: The work break down between the team mates is shown in the picture above.

5 Future Work

We aim to include more features in our model and test more hypothesis with data analytics to clear out any myths or to find any unnoticed feature which hasnt been used by anyone yet. With more datasets about actors , directors and movie related attributes like their country wise revenue, how much they earn in the first week of release and how well they do after that, how much is spent on movie advertisement , etc. With more dataset we hope to come close to a classifier which could help directors or movie production houses to analyze where they should spend or when they should release their movies for best possible profits.

6 Critical Review

The goal of the project is to perform two tasks. One is the descriptive and predictive analysis on the data retrieved from Kaggle. The problem was well defined for the project and the motivation and the approach to solve the problem was defined in the project plan document. The development process was aligned to the approaches described to the problem statement and results achieved so far was reasonable. However , there were few points which could have been approached in a different direction.

Though this project was well structured and well paced for the most part, it would have been better if some concrete results would have been found from the hypothesis taken. That being said the pattern observed for each genre are substantial enough to show that there is some relation between the movie genre and at what time of the year they are released to the rating of movie, maybe with more dataset and better statistical analysis more results can be generated.

7 Distribution Of Roles

The work for the project has been divided among the team mates as the following.

7.1 Team Role : Data Engineer

1. Understanding the dataset

   Exploring the attributes,data types, data size, columns, range and usefulness of the individual attributes to solve the problem statement.

2. Cleaning the dataset

   (a) *The cleaning of the dataset includes several steps such as filling missing values, converting the texts to lowercase etc.

   (b) Imputing Null Values : Assigning the null values to valid data in the data frames. In some cases , removing null values to make the data useful for the next processes.

   (c) Removing irrelevant columns: Identifying the columns which don't contribute to the problem statement and removing the irrelevant columns from the dataset.

   (d) Normalizing the dataset if required : After identifying the range of the attributes defined in the dataset, the values will be normalized to a specific range if required.

   (e) Binning: Combining two or more columns of the datasets to individual columns.

   (f) Cluterting : Grouping values in clusters and then detect and remove outliers (automatic or manual)

   (g) Aggregations: Performing regression by fitting the data into the regression functions.

7.2 Team Role: Data Scientist

1. Objective : Stating the objective of the problem, why are we solving it and how we are solving it. Having a clear understanding of how we are gonna be tackling the problem and what results we should get from solving the problems.

2. Importing Data : Reading the data into the IDE, having full accessibility to modify it and play with it.

3. Data Exploration : Learning about the columns, what all information is given in the dataset. What kind of analysis can be done on it , based on the columns it could be either a inference analysis gaining insights into the database from what has happened Or it could be a predictive analysis by learning what has happened what could happen in future given a set of attributes. If there is no target column then unsupervised learning algorithms like Kmeans, Dbscan algorithms would be used to find out clusters of dataset with similar properties/attributes , if there is a target column then depending on its values if its continuous or discrete supervised learning algorithms would be divided into regression and classification respectively. To gain more knowledge on the dataset EDA(Exploratory data analysis) techniques could be performed.

4. Baseline Modelling : Performing some basic machine learning techniques like linear regression , kmeans to get a baseline on which we could improve upon.

5. Evaluation of models : using evaluation metrics like confusion matrix or using ROC-AUC or MSE scores to test our analysis. Based on these results we would tweak our model burther.

6. Secondary Modelling : Performing complex algorithms like ANN, LSTM or Random Forest to increase accuracy and compare with our base models , and see if after performing the same tweaks on the the new model how do they perform against the base model.

7. Conclusion and Reporting : After the results are obtained we would be explaining the whole process and document it in the report.

8 Conclusion

Through Machine learning models and visual inferences we were able to create a model which can predict if a movie will be hit or not based on given attributes and we have statistically proven that our hypothesis is wrong there is no direct relation between the genre of the movie and the time of the year they are being release in. Maybe with more data we could find out a strong relation for our hypothesis to be valid but with the given data so such thing can be proven. Our project aimed to assist production houses with their movie promotions and where they should spend more so that they can make sure their movie earns the profits they deserve also we aimed to gain some useful insights from the hypothesis we tried to proove , if proven would have tremendously helped production houses to select their release dates and schedule their work accordingly for maximum profits.vc

References

[1] François Chollet et al. Kaggle. https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv, 2015.