Effect of Box-Office on Unemployment

Amogh A Kori
Illinois Institute of Technology Chicago, IL U.S.A. akori@hawk.iit.edu

Mounika Gampa
Illinois Institute of Technology
Chicago, IL U.S.A.
mgampa@hawk.iit.edu

Pragati Khekale
Illinois Institute of Technology
Chicago, IL U.S.A.
pkhekale@hawk.iit.edu

Rajesh Patel
Illinois Institute of Technology
Chicago, IL U.S.A
rpatel166@hawk.iit.edu

Abstract

This project aims to measure how reviews and ratings of movies released in theatres affect their sales at the box office and to analyze how the Unemployment rate in US has an effect on box office revenue. To start with, we have taken the movie dataset as the primary dataset which was retrieved from The Movie Database (TMDB) and GroupLens online sites. And contextual dataset as Unemployment dataset from the U.S. Bureau of Labor Statistics website was taken as a contextual dataset. We have performed data preprocessing and cleaning both the datasets and we have combined the two tables to form a final dataset on which exploratory data analysis was performed to extract insights. Later, have applied predictive modeling algorithms to identify meaningful patterns from the available data.

Keywords—movie, recommendation, ratings, AI, revenue, entertainment

I. MOTIVATION

This project was carried out to determine whether the risk of investing in feature films for theatrical releases could be reduced via the analysis of movie ratings by movie enthusiasts.

II. PROBLEM STATEMENT

Since the beginning of the 20th century, the motion pictures industry invests increasingly large amounts of money into its cinematic productions, and movie studios have tried several different approaches to maximize ROI on theatrical releases. Even so, pleasing the public with motion pictures continues to be a risky endeavor, and box office revenue predictions remainas elusive as ever.

One approach commonly used to predict revenue is the eliciting of reviews in pre-release private screenings. In these events, movies are exhibited before their final editing or release to the public, and the general sentiments of a hand-picked audience towards their potential are “captured”. Nonetheless, this seems to have little effect on box office results. Mega-productions, known as blockbusters, often fail to meet box office expectations, while low-budget movies sometimes perform beyond the wildest hopes and dreams.

Addressing how the general sentiments of the public towards specific theatrical releases relate to box office results could bring substantial benefits to the movie industry. Thus, this project attempts to determine whether the general sentiments of the public towards theatrical releases, expressed in the form of movie ratings, can be used to predict how a movie will fair at the box office.

III. DATA ACQUISITION

To answer this analytical question in a manner that can be universally applicable, machine learning algorithms were utilized to automate the analysis and build suitable predictive models.

Information initially assumed to be relevant for the purpose of the defined use case included movie release dates, budget, revenue, genre, and the associated ratings by motion picture enthusiasts. In addition to this relevant information, scoured datasets should be sizable representative population samples so the scope of the resulting predictive model expands beyond datasets scrutinized for this project.

Online sources were combed for suitable data, and a total of four different datasets, described in the following sections and further detailed in Appendix A, were curated for both manual and automated analysis.

A. Primary Dataset: The Movies Dataset

This is the primary dataset and consists of a collection of data from The Movie Database (TMDb) and GroupLens online sites. TMDb is a user-editable online database with data on millions of movies and TV shows, while GroupLens is a movie recommender website with hundreds of thousands of registered users. The dataset is distributed in 7 separate files, containing metadata for 45,000 movies released on or before July 2017. The included files are described as follows by the creator:

1. ***movies_metadata.csv:*** The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries, and companies.
2. ***keywords.csv:*** This contains the movie plot keywords for our MovieLens movies. Available in the form of a stringifiedJSON Object.
3. **credits.csv:** Consists of Cast and Crew Information for all our movies. Available in the form of a stringified JSON Object.

4. ***links.csv:*** The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset.
5. ***links_small.csv***: Contains the TMDB and IMDB IDs of a small subset of 9,000 movies of the Full Dataset.
6. ***ratings_small.csv***: The subset of 100,000 ratings from 700 users on 9,000 movies.
7. ***ratings.csv:*** 26 million and 750,000 tag applications applied to 58,000 movies by 280,000 users.

The dataset was deemed reliable because it contained data mined from not one but two reliable sources, and the preliminary analysis (visualization of the .csv files in Microsoft Excel) revealed that it did not contain any anomalous values. The relevant content for the purpose of the project is contained in files number 7 (user ratings) and file number 1 (movies metadata), whose contents are described in Appendix A.

B. Backup dataset: IMDb movies extensive dataset

IMDb is one of the most reputable and most comprehensive online databases of entertainment-related information, including movies and associated reviews by both fans and movie critics. The database has been owned by Amazon since 1998 and is well known for its exhaustiveness.

This dataset contains data automatically scrapped with Python scripts from IMDb, includes movies with 100 reviews or more, was last updated in September 2020, and is found split into the following 5 separate files:

1. ***IMDb movies.csv:*** Contains information on 85,855 movies and associated general attributes such as release date, genre, duration.
2. ***IMDb names.csv:*** Contains 20 attributes pertaining to each of 297,705 cast member entries.
3. ***IMDb ratings.csv:*** Contains aggregate rating information for each of the 85,855 movies scrapped from IMDb.
4. ***IMDb title_principals.csv:*** Contains role information for the main cast members of each movie.

This dataset was captured to serve as a backup to the primary dataset, to replace it should that dataset present any significant issues. Additionally, this dataset would also be used to validate the resulting model if time permits.

C. Contextual dataset: US Unemployment Dataset (2010-2020)

This dataset contains a collection of unemployment rates in the US, dating from January 2010 until 2020, grouped by categories (education level, race, gender, etc.). It was put together with U-3 rate data collected from the U.S. Bureau of Labor Statistics (BLS) and it is contained in a single file.

The unemployment rate is a macroeconomic indicator that rises and falls in the wake of changing economic conditions and which can help judge the overall health of an economy. This provides a window into specific circumstances of the labor market for given demographic groups at the time when motion pictures were released in movie theaters.

The relatively short timeframe of this dataset (a 10 year period) is suitable for the temporal scope of the project (between 2010 and 2017). An analysis of box office revenue including older movies would have to account not only for currency inflation but also for lack of online reviews in the past and for the way box office reporting has evolved over time.

This dataset offers unemployment rates for given categories but does not present the overall unemployment rate. Thus, it needs to be combined with the report published by the U.S. Bureau of Labor Statistics (BLS) containing seasonally adjusted unemployment rates for each month of the year.

The contextual dataset was combined with the primary dataset by correlating the year and month in which a movie was released with the corresponding year and month on the unemployment rate report. Although the unemployment rate is a lagging macroeconomic indicator, this dataset was selected to provide context because it can serve as a predictor of future economic performance and can provide insights into whether the general state of the economy is related to movie theater affluence.

D. Contextual Dataset: Bureau of Labor Statistics (BLS) Unemployment Rates

The Bureau of Labor Statistics publishes the unemployment rate on the first Friday of every month (with few exceptions) accounting for the employment situation of the preceding month. This report considers unemployed all those individuals with 16 years of age or more, who are willing and available to work, and who have actively sought work within the past four weeks. This overall unemployment data is retrieved from the website of the Bureau of labor Statistics at https://data.bls.gov/cgi-bin/surveymost?bls.

The data can be extracted in one of several formats with the elected option for the project being a normalized table format.

IV. Data Preparation

To prepare the data in the four datasets for analysis and use in the modeling process, the project employed an 8-step data wrangling process to produce an Analytic Base Table (ABT) data structure. The wrangling process consisted of the following steps: discovery, structuring, cleaning, validating, enrichment, aggregation, integration, and publishing. Note that some of the steps are not separated by hard definition lines, such as between cleaning and validating.

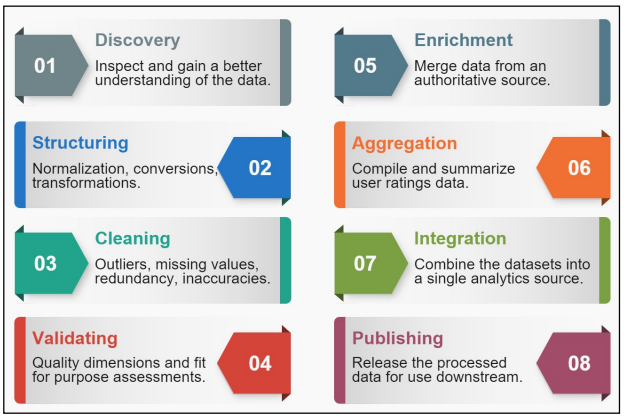
One lesson learned from this process is that whenever feasible, as much as possible of the entire dataset should be prepared for analysis, with extraneous variables being removed at the end of the data preparation process. This provides the benefit of having additional variables ready for the analysis process if a decision is alter made to expand the scope of the analysis.

The steps of the data wrangling process used to prepare the data are further detailed in the subsequent sections.

a) the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert

figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

Fig. 1. Steps of the employed data wrangling process



B. *Discovery*

During this user-oriented process, the datasets were visually explored, in their raw form, to better understand the available data and to gain insights into the use case and more clearly define the associated key entities (i.e., movies and ratings). With better insights into the nature of the data, better questions can be asked of it for business purposes.

As a result, a movie was defined as a motion picture that had its first public release carried out in a movie theater. This excluded movies never released to the public or released initially on video. Additionally, notice was taken of the different ranges used for the rating scales of the different online movie recommender sites. While TMDb uses an integer scale ranging from 1 to10, GroupLens employs a scale from 1 to 5 with 0.5 increments.

C. *Structuring*

Data structuring techniques were used to normalize the data and reduce complexity. For example, it was ensured that each variable contained an atomic value and datatype conversions were applied where variables were not assigned the correct datatypes. Some schema conversions were also performed to transform some of the JSON formatted strings into tabular format.

D. *Cleaning*

The data cleaning process had outliers stripped from the datasets, such as movies with movie the value of revenue set at 0. Additionally, movies released straight to video or with any status other than *RELEASED* were also dropped from analysis consideration. Entries with missing values for required entries, such as the release date were also excluded.

As part of the data cleaning process, our primary dataset was reduced from 45,466 individual movie entries to 7,406 entries.

E. *Validating*

This step consists of validating the results of the data cleansing process to ensure that the resulting dataset is still fit for purpose. This included verifying that categorical values contained only acceptable values, that date variables had the correct datatypes, etc.

F. *Enrichment*

The enrichment step consists of merging a dataset with third-party data from a reputable data source. For this project, initial enrichment was performed by merging unemployment statistics with unemployment rates published by the Bureau of Labor Statistics, the authoritative source on the matter.

G. *Aggregation*

For the aggregation step, user rating information by numerous users was summarized through the computing of aggregates grouped by individual movies.

H. *Integration*

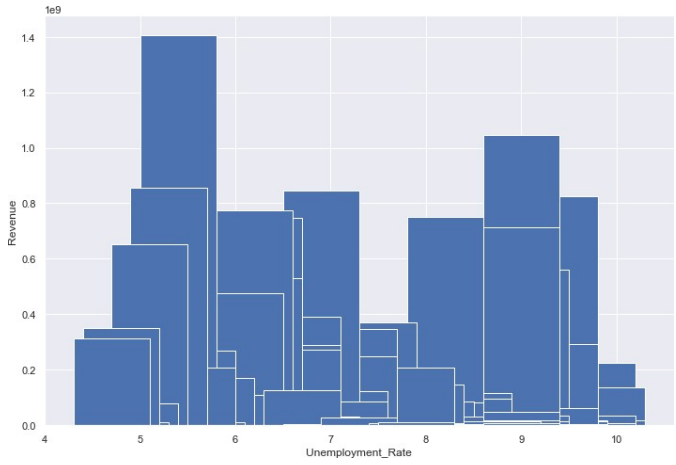
Once all datasets were cleaned and validated, they were combined into a single unified view, using common fields (i.e., movie identifier to associate movies with ratings, and both the year and the month of the reporting period to associate the datasets containing unemployment data and correlate them with movies based on the movie release data). The outcome was an Analytical Base Table to be used for analysis.

I. *Publishing*

The outcome of the data wrangling process was a combined dataset, containing clean and validated data, that was formally made available for data analysis and the model building stages of the project.

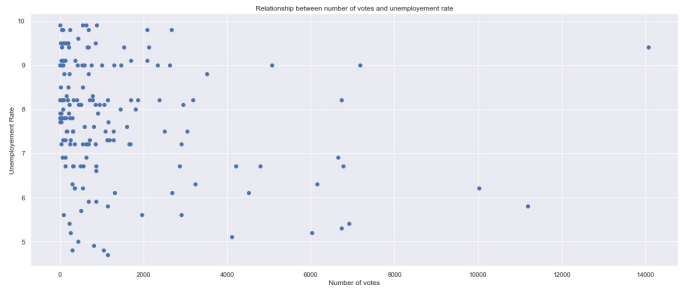
V. Exploratory Data Analysis (EDA)

A. Unemployment Rate VS Revenue

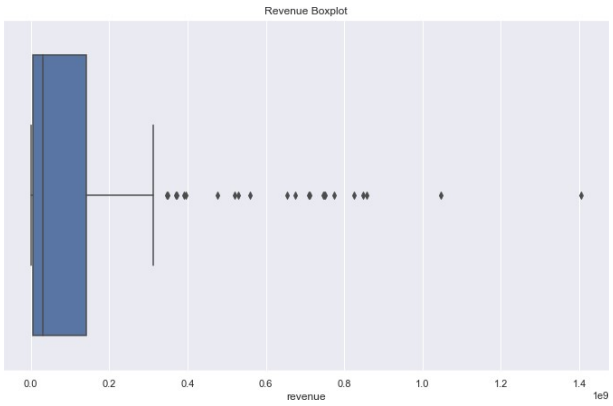


Movies which earned higher revenue tend to have lower employment rate.

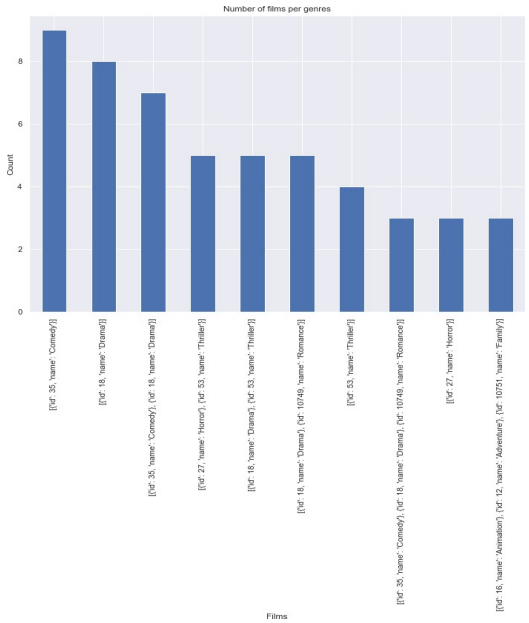
B. Unemployment Rate VS Number of Votes



C. Revenue Boxplot

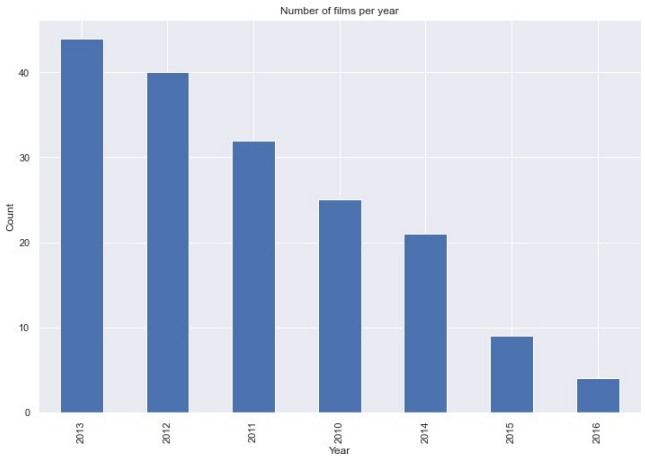


D. Number of films per genres



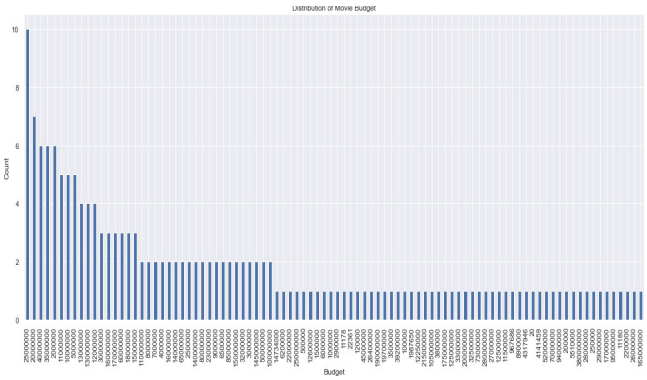
Most movies are of comedy genre followed by thriller.

E. Number of films per year



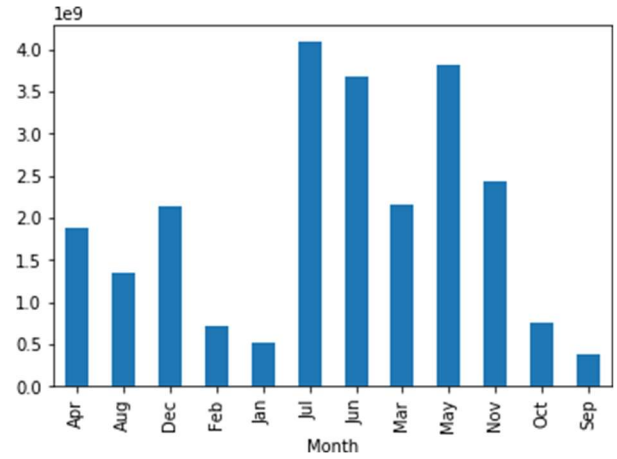
Most movies were released in year 2013 in comparison to various year (these statistics were framed from the data we had)

F. Distribution of Movie Budget



There was good amount of high budget movies on box office

G. Movie released in a given month



Movies which were released in summer tend to create more money whereas the ones which were during the peak wintertime tend to make less revenue in comparison.

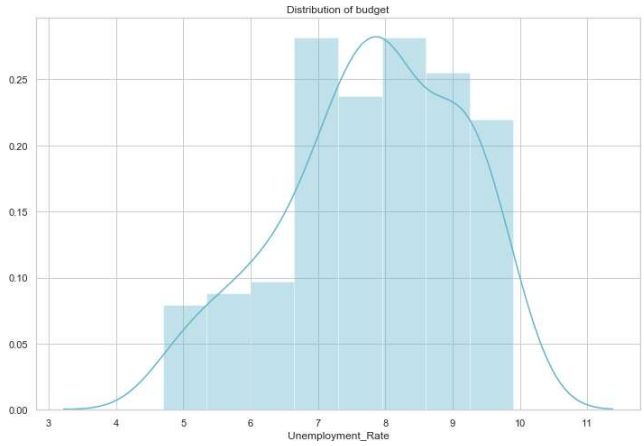
H. Movies with highest Average Ratings

Movie Title	Average Rating
127 Hours	3.30
13 Sins	3.125
17 Girls	2.875
30 Minutes or Less	3.50
5 Days of War	3.428

I. Movies with highest Ratings

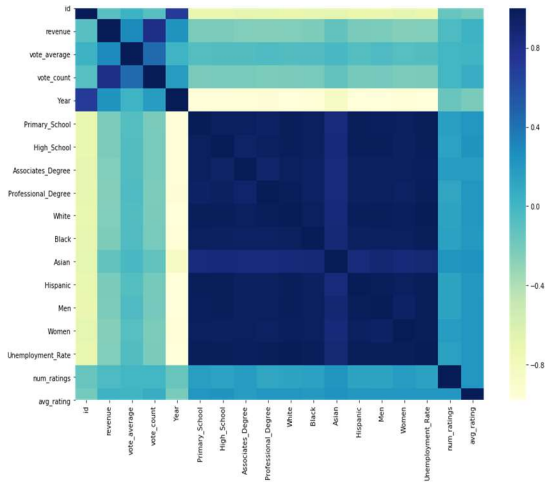
Movie Title	Ratings
Labor Day	4.5
Resident Evil: The Final Chapter	4.0
The Call	4.0
The Rover	4.0
Guardians of Galaxy	4.0

J. Distribution of budget



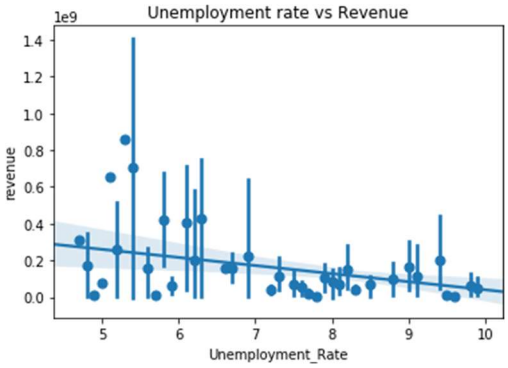
Movies with good amount of budget tend to show median unemployment rate

K. Correlation matrix



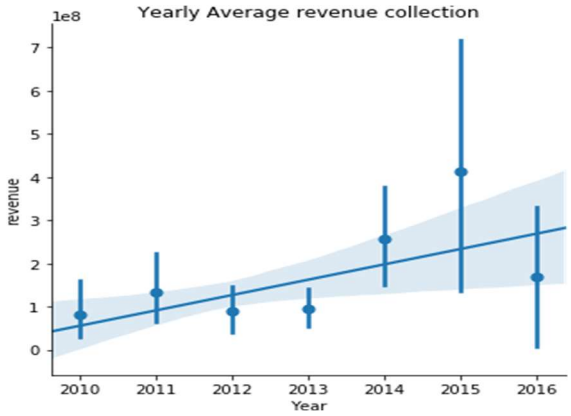
This shows how various variables are correlated to each other and how effective is each feature on every other feature.

L. Unemployment Rate VS Revenue



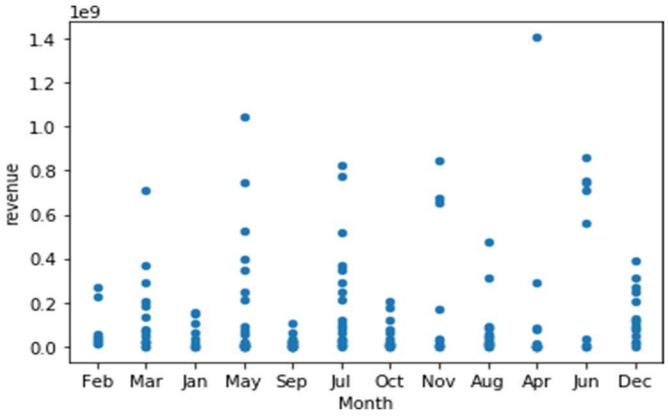
Unemployment rate increases when the movie fails to make good amount of profit or is unable create good amount.

M. Average Revenue collection based on Year



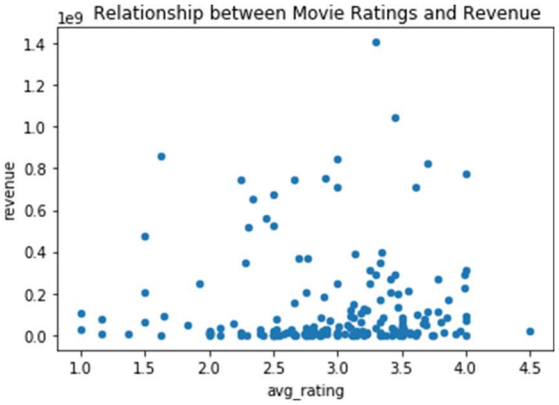
In year 2015 the most average revenue was collected, most movies performed the best during that duration.

N. Revenue Scatter plot



We can find a movie in April created a good profit shown as an outlier in the plot above, but overall there were was good quantity of movies released in July

O. Relationship between Movie Ratings and Revenue



This shows that even when the movie rating were the movie did make a good revenue on box office and few movies which had good movie reviews failed terribly at box office.

REFERENCES

[1] T. Gebru, J. Morgenstern, B. Vecchione, J. Vaughan, H. Wallach, H. Daume III, and K. Crawford, "Datasheets for datasets," arXiv:1803.09010v7 [cs.DB], Mar 2020.

APENDIX A

Datasheets for Datasets

This appendix contains datasheets for each of the datasets curated for the project.

I. THE MOVIES DATASET

A. Motivation

This dataset was assembled by Rounak Banik as part of a project (the Capstone Project) for Springboard's Data Science Career Track. It was built as a narrative for the history of cinema and to serve as support for movie recommender systems.

The creator of this self-contained dataset anticipated that the dataset can also be used to predict box office revenue based on any one of a set of variables, to determine characteristics of movies with more engagement from users and to support movie recommender systems.

B. Composition

Each instance in the movies metadata file of this dataset represents a motion picture that has been rumored about, planned for production, currently being produced, canceled, in post-production, or released to the public. The dataset contains 45,466 unique movie instances as registered by The Movie Database (TMDb) site, but does not contain all possible instances as there are more exhaustive datasets available.

The ratings file contains 26,024,289 movie ratings submitted by registered users of the GroupLens online movie recommender site, yet without any associated personally identifiable information.

These two instances are related via a unique movie identifier and the composition of each instance is detailed below. There is no recommended data split for either instance.

TABLE I. GROUPLENS RATINGS DATASET VARIABLES

Variable	Type	Description
userId	Integer	The unique identifier of the GroupLens user who rated the movie.
movieId	Integer	The unique identifier of the rated movie.
rating	Float	The rating value assigned by the GroupLens user to the movie (on a scale of 0 to 5, in 0.5 increments).
timestamp	Integer	The date and time when the Grouplens user rated the movie.

TABLE II. TMDB MOVIES DATASET VARIABLES

Variable	Type	Description
adult	Text [TRUE, FALSE]	An indication of whether the movie is intended for an adult audience.
belongs_to_collection	Text (JSON)	Collection details if the movie is part of a fil series.
budget	Integer	The amount of money (in dollars) spent on the entire movie project.
genres	Text (JSON)	The category (or set of categories) that define a movie based on its narrative elements. These have changed and evolved over time.
homepage	Text (URI)	A link to the official website of the film.
id	Integer	The unique identifier of the movie.
imdb_id	Text	The unique identifier of the movie in the IMDB database.
original_language	Text	The two- character ISO 639-2 language code of the original language of the movie.
original_title	Text	The original title of the movie.
overview	Text	A synopsis of the movie describing the context and plot of the movie.
popularity		
poster_path	Text	A relative path to a .jpg image of the movie poster.
production_companies	Text (JSON)	The production company (or companies) that produced the movie.

Variable	Type	Description
production_countries	Text (JSON)	The country (or countries) where the movie was filmed on location.
release_date	Date	The date when the movie was first released through movie theaters for the public (the movie premiere).
revenue	Integer	The amount of money generated by the movie through theater movie ticket sales.
runtime	Integer	The elapsed time (in minutes) from the start of the movie until the end of the credits scene.
spoken_languages	Text (JSON)	The language (or languages) spoken during the movie.
status	Text (category)	The current stage of the movie production [CANCELED, IN PRODUCTION, PLANNED, POST PRODUCTION, RELEASED, RUMORED].
tagline	Text	A phrase used to market and advertise the movie (advertising slogan).
title	Text	The title of the movie.
video	Text [TRUE, FALSE]	Whether the movie had a theatrical release before being released on video.
vote_average	Integer	The average rating by TMDb users (on a scale of 0 to 10).
vote_count	Integer	The total number of TMDb user ratings.

Some released movies were found not to have any associated revenue, which was deemed an error. Additionally, status information was not provided for some of the instances either and although the dataset description page specified that it included movies released on or before July 2017, the data contained movies released as late as August 2018.

C. Collection Process

As per the creator, movie entities were collected by registering and requesting access to the publicly available TMDB Open API (https://www.themoviedb.org/documentation/api), while user ratings were extracted from datasets made available to the public by GroupLens on its website (https://grouplens.org/datasets/movielens/latest/). The dataset was created on October 24, 2017 and last updated on November 10, 2017, and was downloaded from Kaggle at the following URI: https://www.kaggle.com/rounakbanik/the-movies-dataset/metadata.

D. Preprocessing

The following steps were taken to process this dataset:

1. **Relevant data:** Removed from the dataset all the non-relevant variables to address the identified business need.
2. **Movie revenue:** Identified and removed instances for which no revenue was recorded (i.e., a value of $0).
3. **Feature films:** Removed from the dataset all films made for and premiered in television an ensured only feature films were part of the dataset.
4. **Dated releases:** Removed from the dataset all the movies that were not released to the public, or for which the release date was not specified.
5. **Rated movies:** Removed from the dataset all movies for which there were no associated user ratings.
6. **Date conversion:** Converted the review time data from a timestamp to a date and time value.
7. **Existing movies:** Removed from the dataset all the ratings without a corresponding movie in the processed movies metadata dataset.

After this process, the initial dataset was reduced to 232 movies meeting the required criteria and 61,387 corresponding user ratings. The unprocessed dataset is saved and the software used to process the data, Jupyter Notebook, is freely available.

E. Lifecycle

The dataset was used uniquely in the scope of this project and will not be distributed, published, or maintained outside of its scope.

II. IMDB Movies Extensive Dataset

A. Motivation

This dataset was created by Stefano Leone, who scrapped the data from the IMDb website (https://www.imdb.com/). The inspiration behind this effort was to inquire which aspects make movies successful from a business viewpoint and from the perspective of movie watchers alike.

This dataset can also be used to answer a myriad of movie-related questions concerning revenue predictions, genres, plot trends, etc.

B. Composition

The self-contained dataset contains 85,855 movie instances, representing past, present, and future motion picture releases with attributes that include title, description, genre, etc. Not all instances are represented though, as the IMDb database contains over 500,000 featured films. Additionally, the dataset offers 85,855 movie ratings by registered IMDb users. Other instances, with less relevance for the project include names of cast members (297,705) and cast member roles (835,513).

All instances are related via unique movie identifiers and there are no recommended data splits.

C. Data Collection

The dataset creator scrapped data from the publicly available website https://www.imdb.com, and included only movies with more than 100 votes. The dataset was created on November 25, 2019 and last updated on September 14, 2020. For the project, the dataset was downloaded from Kaggle at https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset/metadata.

D. Preprocessing

No preprocessing was performed on this dataset as it was set aside to use as a backup in case of need dur to issues with the primary dataset.

E. Lifecycle

The dataset was used uniquely in the scope of this project and will not be distributed, published, or maintained outside of its scope.

III. US Unemployment Dataset (2010-2020)

A. Motivation

Aniruddha Shirahatti created this dataset primarily for combination with COVID-19 datasets to provide a greater understanding of how the unemployment rate in the U.S. is being impacted. The dataset can also be combined with other datasets to provide context on the potential impact of the unemployment rate.

B. Composition

Each entity in this dataset represents the unemployment rate each year and month for a specified demographic group. The range of dates covered spans from January of 2010 to December of 2020, for a total of 132 instances. The dataset contains all instances in the stipulated period.

Although self-contained, this dataset is dependent on external data to correlate the unemployment rate for the included demographic groups with the overall unemployment rate.

The contents of the US Unemployment Dataset (2010-2020) are described below.

TABLE III. US Unemployment Dataset Variables

Variable	Type	Description
Year	Integer	The reporting year.
Month	Float	The reporting month.
Primary_school	Text	Unemployment rate among individuals with a primary school level of education.
Date	Date	The month and year of reporting.
High_School	Float	Unemployment rate among individuals with a high school level of education.
Associates_Degree	Float	Unemployment rate among individuals with an associate degree.
Professional_Degree	Float	Unemployment rate among individuals with a professional degree.
White	Float	Unemployment rate among individuals of white ethnicity.
Black	Float	Unemployment rate among individuals of black ethnicity.
Asian	Float	Unemployment rate among individuals of Asian ethnicity.
Hispanic	Float	Unemployment rate among individuals of Hispanic ethnicity.
Men	Float	Unemployment rate among male individuals.
Women	Float	Unemployment rate among female individuals.

The instances for the dates ranging from April 2020 to December 2020, although part of the dataset, contained no unemployment rate data and were thus discarded as erroneous, reducing the effective number of instances from 132 to 123.

The dataset also contains statewide unemployment statistics, which were disregarded for this project due to time constraints.

C. Data Collection

The dataset creator downloaded and transformed data from the site of the Bureau of Labor Statistics and created this dataset on April 18, 2020. The dataset was last updated on April 30, 2020 and was downloaded from Kaggle at the following URI: https://www.kaggle.com/aniruddhasshirahatti/us-unemployment-dataset-2010-2020/metadata.

D. Preprocessing

The following steps were taken to process this dataset:

1. **Missing values:** Looked for missing values in the dataset and deleted entries for which there was no corresponding unemployment rate information.

The unprocessed dataset is saved, and the software used toprocess the data, Jupyter Notebook, is freely available.

E. Lifecycle

The dataset was used uniquely in the scope of this project and will not be distributed, published, or maintained outside of its scope.

IV. Bureau of Labor Statistics Unemployment Rates

A. Motivation

The Bureau of Labor Statistics (BLS) publishes the unemployment rate monthly as part of its responsibility to report on the current employment situation in the nation.

B. Composition

Each instance of this dataset represents the overall unemployment rate in a given year and month. The dataset is self-contained but does not offer all possible instances because the desired period was specified as a filter during its extraction. As a result, 1232 instances, ranging from January 2010 to December 2020 can be found in the dataset. None of the data is confidential as it was sourced from a publicly accessible site.

The data can be extracted in one of several formats with the elected option for the project being a table with data for the specified reporting periods and in the following format:

TABLE IV. BLS Unemployment Rates Dataset Variables

Column	Type	Description
Series id	Text	The identifier of the associated BLS report.
Year	Integer	The year of the reporting period.
Period	Text	A string representation of the month of the reporting period [M01-M12].
Value	Float	The unemployment rate in the reporting period.

C. Data Collection

This dataset was created specifically for this project by exporting data from the publicly accessible site of the Bureau of Labor Statistics at https://data.bls.gov/cgi-bin/surveymost?bls.

D. Preprocessing

The following steps were taken to process this dataset:

1. **Data augmentation:** Inserted a composite column to specify the month of the reporting period to match the representation in the unemployment dataset.

The unprocessed dataset is saved and the software used to process the data, Jupyter Notebook, is freely available.

E. Lifecycle

The dataset was used uniquely in the scope of this project and will not be distributed, published, or maintained outside of its scope.

Appendix B

Model cards for models developed

This appendix contains model cards for all models developed.

MODELLING:

One of the important parts in Data Science project is modelling. Based on the data retrieved, we can build a model to predict the future events or to recognize the patterns that helps in making decisions on our business effectively. Models uses statistical approach to predict the output. Models are trained using set of data as input to an algorithm that helps model in learning or recognizing the patterns. Once the models learn from the data, we can use it over a new set of data and make predictions or identify the patterns. There are two types of models, supervised and unsupervised models. In this project we have used both.

I. Supervised Model:

In supervised learning models, the training dataset contains labelled outputs. It will have a knowledge of what the output should be. These models tend to be accurate, but they require human intervention to label the data appropriately. Supervised model can be divided into two types – Regression and Classification.

*A. **Regression***:

To understand the relation between independent and dependent variables we use regression method. These methods will be useful when output is of continuous real values. Linear regression, polynomial regression, random forest regression, and logistic regression are some of the popular regression methods to find trends in the data.

***B. Classification*:**

Classification models are used to classify the test data into different labelled categories that was learnt from the training set of the data. For example, an algorithm for classification model can identify an email is spam or not spam. Some of the classification models are decision tree, support vector machine (SVM), k-nearest neighbors (KNN) etc.

II. Unsupervised Model:

Unsupervised learning is training a dataset which is not classified or not labelled. The task of the algorithm is to group the data as per their similarities, differences, and patterns without any prior knowledge of the data. The dataset given are unlabeled and model has to cluster and analyze the datasets. Unsupervised learning models are classified into two algorithms – Clustering and Association.

A. Clustering:

It is one of the techniques in data mining to group the data based on its similarities and dis-similarities. As defined in Wikipedia (https://en.wikipedia.org/wiki/Cluster_analysis), "Clustering is a task of grouping a set of objects in such a way that objects in same group are more similar to each other than those in other groups." This is a general task that is performed and can be accomplished by various algorithms based on density of the dataset, distance between data points and on statistical distributions. The most popular clustering algorithms are K-means and Hierarchical clustering.

B. Association:

Association is a type of unsupervised learning models used to find the relation/rules between the variables. It is frequently used in recommendation systems and in market basket analysis. In this project, we are determined to find the relationship between unemployment rate and various factors of movie dataset like revenue, average rating etc. Since the values in the dataset are continuous real values, we have decided to use Linear Regression model. Following is the brief description of the model and its implementation on this project.

Linear Regression:

The main idea for linear regression model is to find the best fit line for the data. With linear regression model we can predict the value of dependent variable using one more independent variable. When we use single independent variable to predict, it is known as simple linear regression model. Similarly, when we use one or more independent variables to predict then it is known as multiple linear regression. Let us consider an independent variable x, using which we need to predict the output y.

For simple linear regression, the model would be

$$Y=B(X)+C$$

Where B=coefficient/weight of X and C= bias coefficient.

And for multiple linear regression, the model would be:

$$Y= B0(X1) + B1(X2) +\ldots+ Bn(Xn) + C$$

Where X1, X2,…,Xn are input values, B0, B1,..,Bn are its corresponding coefficients/weights and C is bias coefficient. Below mentioned is an example graph indicating simple linear regression.

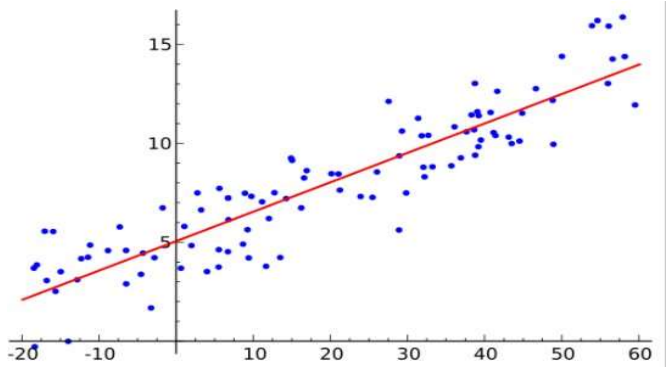


Fig 2: Simple linear regression, Source: https://en.wikipedia.org/wiki/Linear_regression#Simple_and_multiple_linear_regression

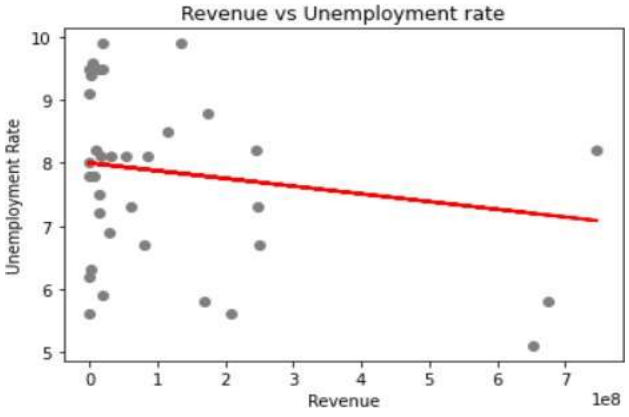
IMPLEMENTATION:
In the implementation of modelling, the main aim to find the relationship between unemployment rate and movie revenue and to predict movie revenue based on unemployment rate. Since the values in these attributes are continuous, we have applied linear regression model.

A. Unemployment Rate & Revenue:

The dataset we are using has been pre-processed and cleaned as explained in previous sections. The final dataset which is the combination of primary (Movie dataset) and contextual (Unemployment dataset) was used. Firstly, we took values of revenue in variable x and unemployment rate values in y. 80% of these values are divided into training and remaining into test sets. There are packages exists in python sklearn library to implement linear regression. We have performed the sklearn's LinearRegression model on the training dataset. Later we have took the values of coefficient which is a slope of the line, intercept value and score from the model. Below are the values.

Intercept: [8.00097999]
Coefficient: [[-1.22633911e-09]]
Score: 0.053275521476430776

Here the coefficient value has negative sign, this indicates that there is negative correlation between Revenue and Unemployment rates i.e., when unemployment rate decreases revenue increases and vice versa. And the score indicates R^2 score. R^2 score value is used to evaluate performance of the model. The best score is 1, it occurs when predicted values are same as true values. 0 is baseline score of the model and negative values occurs when the model's performance is worse. This score indicates the relationship between two values. Since, we have received a score of 0.05, there is not any strong relation between Unemployment rate and Movie revenue. Now we have predicted the values of test data and have plotted a graph for the model and calculated error rates with which the model has predicted.



Mean Absolute Error: 1.0933204020686056
Mean Squared Error: 1.7074806756987433
Root Mean Squared Error: 1.3067060402778978

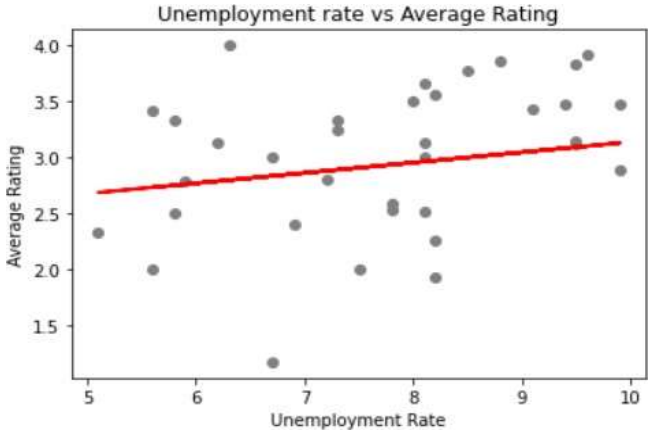
It can be inferred from the scatter plot points that most points are at low revenue when the unemployment rate is high. With this we can say that there is somewhat a negative linear relationship between these two variables instead of being completely independent. This can be confirmed by the score value of 0.0532.
Root mean squared error (RMSE) for this model is approximately 1.3. This error calculates how far the points spread out from the linear line. This will tell how best fit the line is for the data. Since the mean of unemployment rate is 7.8, RMSE value of 1.3 is below 10% of the average value hence it is a good value. Hence finally we can conclude from this implementation that unemployment rate has meaningful relationship.
Now, we wanted to explore more on how unemployment rate effects other factor of the movie like average rating. Since these attributes also include continuous values, we will be applying linear regression model.

B. Unemployment Rate & Average Rating:

For this prediction model, we have taken Unemployment rate values on variable x and Average rating values on variable y. This dataset has been divided two parts training set and test set. Training set consists of 80% of the data. On applying linear regression on the training set with the help of sklearn's library, below are the values of intercept, coefficient, and score.

Intercept: [2.21187059]
Coefficient: [[0.09239123]]
Score: 0.031089008230032694

And on predicting the values for test data and plotting it on the graph, below is the output:



Mean Absolute Error: 0.5111230051487996
Mean Squared Error: 0.38697618736451633
Root Mean Squared Error: 0.6220741011845102

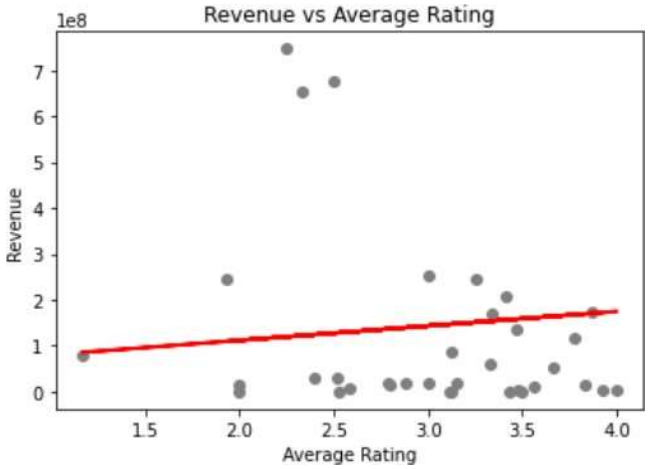
From the above results, it can infer that Unemployment rate and average rating of the movie have positive but small linear correlation coefficient of 0.09. And the R^2 score is almost 0. This indicates that these attributes are independent of each

other. RMSE value was calculated as 0.62, and the mean of average rating is 2.94. Since RMSE was more than 10% of the mean of average rating, the variables are almost independent of each other, there exists no relationship.

C. Revenue & Average Rating:

To explore more on primary dataset, we have decided to check what effect average rating has on revenue. Like previous models, we have applied linear regression model for this as well because of continuous values for the attributes selected. We have taken Revenue values in X and average rating values in Y. Divided 80% of data into training set and rest of 20% to test set. On applying linear regression model, below are the results.

Intercept: [48576631.11682867]
Coefficient: [[31547118.2490375]]
Score: 0.007713784162576176

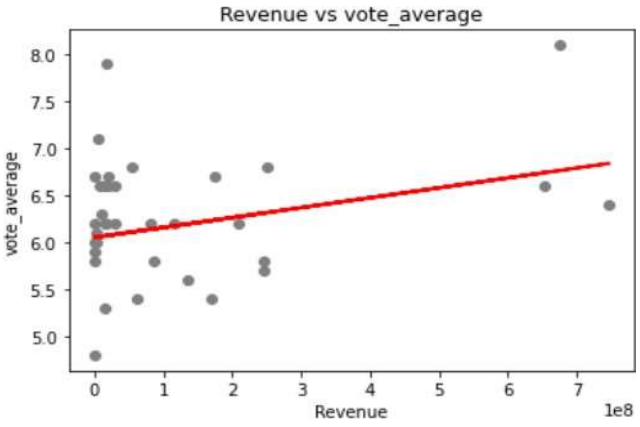


Mean Absolute Error: 146804559.82723203
Mean Squared Error: 4.042542276327888e+16
Root Mean Squared Error: 201060743.96380532

On analyzing these results, the error and the coefficients values are very large. This indicates this model is not good fit for the prediction of revenue based on average rating. Or maybe we need to standardize the values so that the error will be normalized. The outliers in the attribute might be skewing the results. So, let's check on predicting the revenue based on another factor i.e., vote average and check if we receive this high error.

D. Revenue & Vote Average:

Similar to previous model, we wanted to find the vote average on revenue of the movie. On applying linear regression model on revenue and vote average, below are the results.
Intercept: [6.05245955]
Coefficient: [[1.05524221e-09]]
Score: 0.08524368948085714



Mean Absolute Error: 0.483715371630039
Mean Squared Error: 0.3999048823754229
Root Mean Squared Error: 0.632380330477967

As per the results shown, we can say that both attributes have small positive correlation coefficient. The mean of the vote average is 6.2 and RMSE value is 0.63 which is 10% of the mean value. This indicates the variables are independent of each other and do not have strong linear relationship. And the error rate was not high, so the high error received in previous model might not be because of high values in revenue attribute, it might be because of issue with the model itself.

E. Model for profit production:

Based on the attributes available in the primary dataset, we could compute the profit. We have created a variable in the dataset called Profit, which is computed by taking difference between revenue and budget values and subset of original dataset. Now the new dataset for this model would include Budget, Revenue, Vote Average, Vote count, Year and Profit as variables. All the variables except for label i.e., Profit were assigned to data frame X and the Profit values are assigned to an array called Y. Now X and Y are split into training and testing sets. As these values are continuous and other classification models doesn't apply to this dataset, we again are going to perform Linear Regression model on training set to predict Profits on the movie. Below are the results.

Mean Absolute Error: 2.416283158319337e-08
Mean Squared Error: 2.277767914913387e-15
Root Mean Squared Error 4.772596688295992e-08

As shown above the RMSE value is very high. We happened to discuss if this error value will be same even if use any other models and does any other model perform better than Linear Regression for this dataset? Since we have continuous real values in the dataset, we must use regression model. Random forest regressor is one such model.

Random Forest Regressor:
This algorithm uses ensemble learning method. In this method predictions are combined from multiple predictive models to make more accurate predictions.

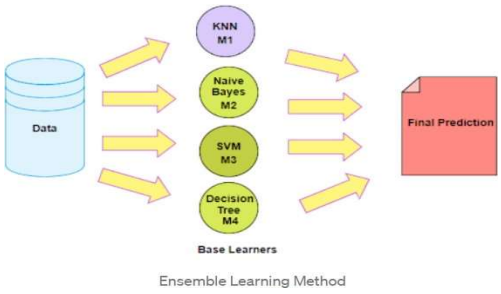


Figure 2: Ensemble Learning Method, Source: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f
On implementing this model on profit dataset on training set and predicting on testing set, below are the results.

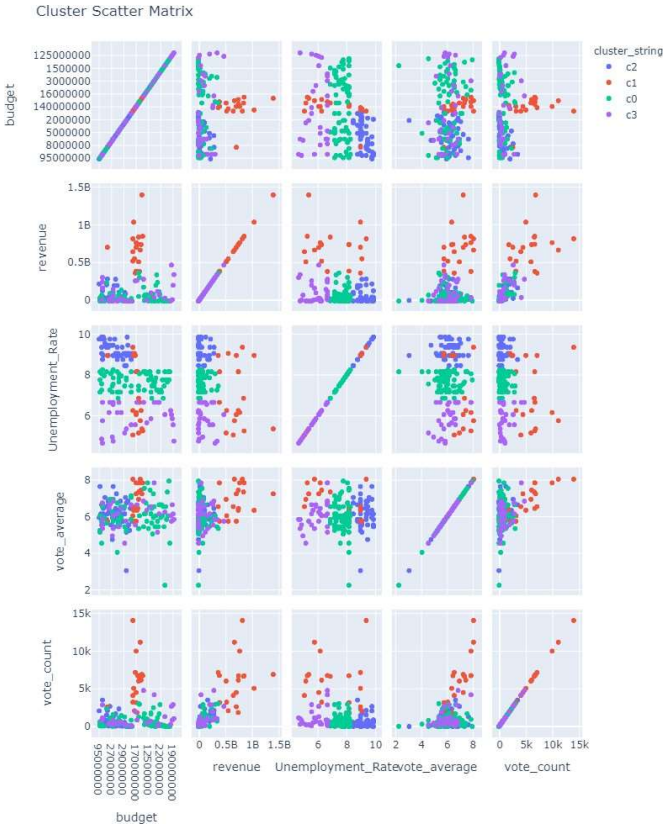
Mean Absolute Error: 14929153.74885714
Mean Squared Error: 967662608989788.2
Root Mean Squared Error: 31107275.820775244

Compared to linear regression, this model has performed well on predicting movie profits.

Next Steps:
If there is more time for this project, we would have explored and implemented more models on the dataset to get insights on the data. We would also further analyze the tested models to better understand their results and that we would also run them on different datasets. We would have also explored other datasets as contextual and perform analysis combining all three datasets. And would have built recommendation system on the primary dataset.

K Mean Clustering:

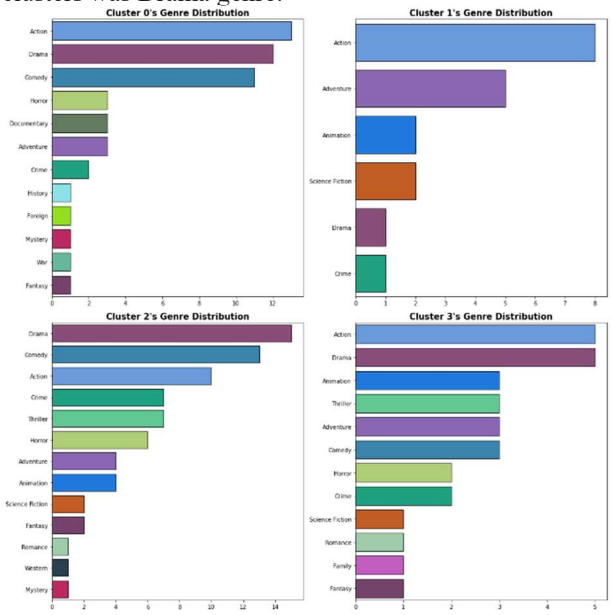
By applying K Mean Clustering model we get some comparisons between the variables in the cluster Scatter matrix. From the scatter matrix we can say that the revenue is mostly low when the unemployment rate is high. And the vote count is also low when unemployment rate is high. The budget does not have any major effect on the unemployment rate. We can also find that the revenue is high when the budget is high. The cluster scatter matrix has the comparisons among budget, revenue, unemployment rate, average votes and vote count. Since the comparison cannot be done with itself, we get a diagonal scatter plot matrix as show in the figure. The various colors dots represent they belong to different clusters. With the help of this matrix, we can conclude that there is a certain relationship on cluster co and c3 with unemployment rate. They show some pattern and similarities when compared to any other cluster. The revenue and the vote count is very low when the unemployment rate is low. On a close watch we can also find that there is a similarity with vote count and the revenue with respect to unemployment rate. They show similar scatter plot, so the unemployment rate shows a similar effect on movie revenue and the number of votes.



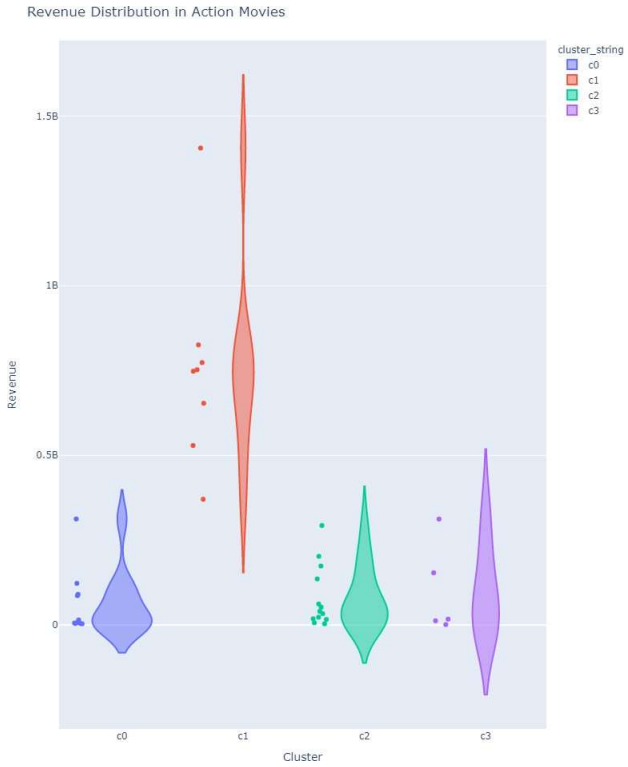
Cluster genres distribution

Apart from contextual data, using the K mean clustering we have plotted the bar graph with respect to the type of genres. The count of genres is calculated in the bar graph in a descending order. We have clustered the movie genres and highest genre count in each cluster Almost all the clusters turned out to have the highest number of Action genres except cluster 2 where Drama genres has the highest count. Since we have Action as the highest, we now plot a revenue distribution among all the clusters. The cluster 0 suggests that the Action genre has the overall highest and the Mystery genre records the lowest. Cluster 1 suggests that the Action genre has the overall highest and the Drama genre records the lowest. Cluster 2 suggests that the Drama genre has the overall highest and the Western genre records the lowest. Cluster 3 suggests that the Action genre has the overall highest and the Science Fiction genre records the lowest. So, there is a significant difference between the overall lowest among all the clusters but the highest is Action for many of the clusters. So, we conclude that the action genre movies we high in number during the unemployment. Each distribution chart did not have all the genres because some of the clusters did not have all the genre of movies. So, since the action movies were the highest, we

have done furthermore analysis with respect to the action movies. The revenue distribution plot was created for the two movies genres which were on the top of the list i.e., which has top and the second highest count of the movies among all the clusters was Drama genre.



We have plotted a revenue distribution graph for action movies. The action movies revenue was highest on cluster 1 where there were a smaller number of movies released which compared to the other clusters. Even though cluster 0 have a greater number of movies the revenue seems to be comparatively low.



Conclusion

From this project we learned that the feature revenue has shown positive effect on unemployment rate i.e., when movies perform well on box office, they have low unemployment rate.

Also, in many cases movies with higher ratings and high budget fails to make big bucks at box office. There were many movies which failed to impress the critics but made good profit on box office

References:

https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
https://en.wikipedia.org/wiki/Linear_regression#Simple_and_multiple_linear_regression
https://en.wikipedia.org/wiki/Association_rule_learning
https://en.wikipedia.org/wiki/Cluster_analysis
https://towardsdatascience.com/machine-learning-basics-random-forest-regression-be3e1e3bb91a
https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/#:~:text=As%20the%20square%20root%20of,of%20RMSE%20indicate%20better%20fit.
https://github.com/jayantsingh123/Movies-Data-Analysis/blob/master/movie_analysis.md

Appendix C

Individual Report

A. Detailed description of the work done on this project

- We have worked on the Exploratory Data Analysis(EDA) for this project.
- The EDA explores the movies and unemployment rate dataset through visualization and graphs using python libraries, matplotlib, and seaborn.
- Some of the key features in EDA are identifying the features, calculating the number of observations, checking for null or missing values in the data, identifying the data types for features, checking for empty cells or columns in data, and identifying them and exploring the data.
- With the work done on EDA we can investigate the critical decisions regarding which features are important and which are not worth following up on, with the help of EDA we can build a hypothesis using the relationship among variables.

B. Challenges faced

- EDA doesn't always yield expected results immediately.
- We cannot assess the risks related to the results.
- The data predicted from models have some amount of error.
- To find the most suitable model we need greater techniques.
- The effectiveness can be further improved by using advanced data analytics.

C. Learning

- This was a great project for understanding EDA and using various coding skills.
- With the help of EDA, we were able to solve the questions on this project.
- We were able to identify the missing values and how to deal with those values
- Training different kinds of data.
- Adding, changing, and removing the features.

D. Group Performance

- My group performed very well during the whole project
- We had assigned roles for everyone to help with collaboration
- We all had strong teamwork skills
- We also had designated time for group meetings
- I would grade all my team members 8/10 for their enthusiasm, hard work and dedication shown in the project.
- I am proud of my team members for their skills and knowledge.