Analyzing effect of Box Office on Unemployment

Pragati Khekale, A20471024
Mounika Gampa, A2048807
Amogh A Kori, A20491465
Rajesh Patel, A20503050

Problem Statements

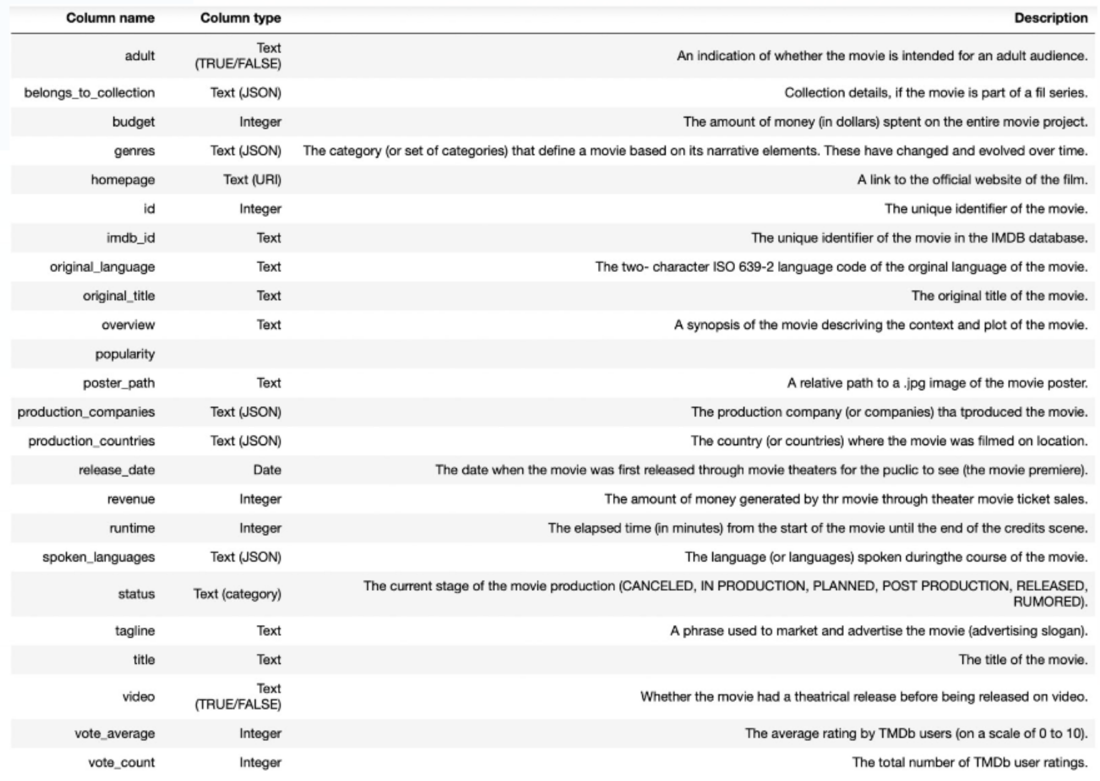
- Analyze how the unemployment rate in US has effect on box office revenue.
- Measure how reviews and ratings of movies released in theaters to their sales at the box office.
- Another objective of the project is to visualize the analysis.

PROJECT OVERVIEW

- Perform data collection of all 4 mentioned distinct datasets (These datasets are raw but obtained from authoritative source which maintains the data quality and authenticity)
- Perform Data preparation steps such as: Data observation, cleaning, validating, integration, etc
- After we obtain a processed data as a result of the previous steps, now we combine all the processed datasets
- While combining keep the relevant fields from the datasets and alter the fields to make the final resulting dataset uniform
- Perform Exploratory data visualization and analysis

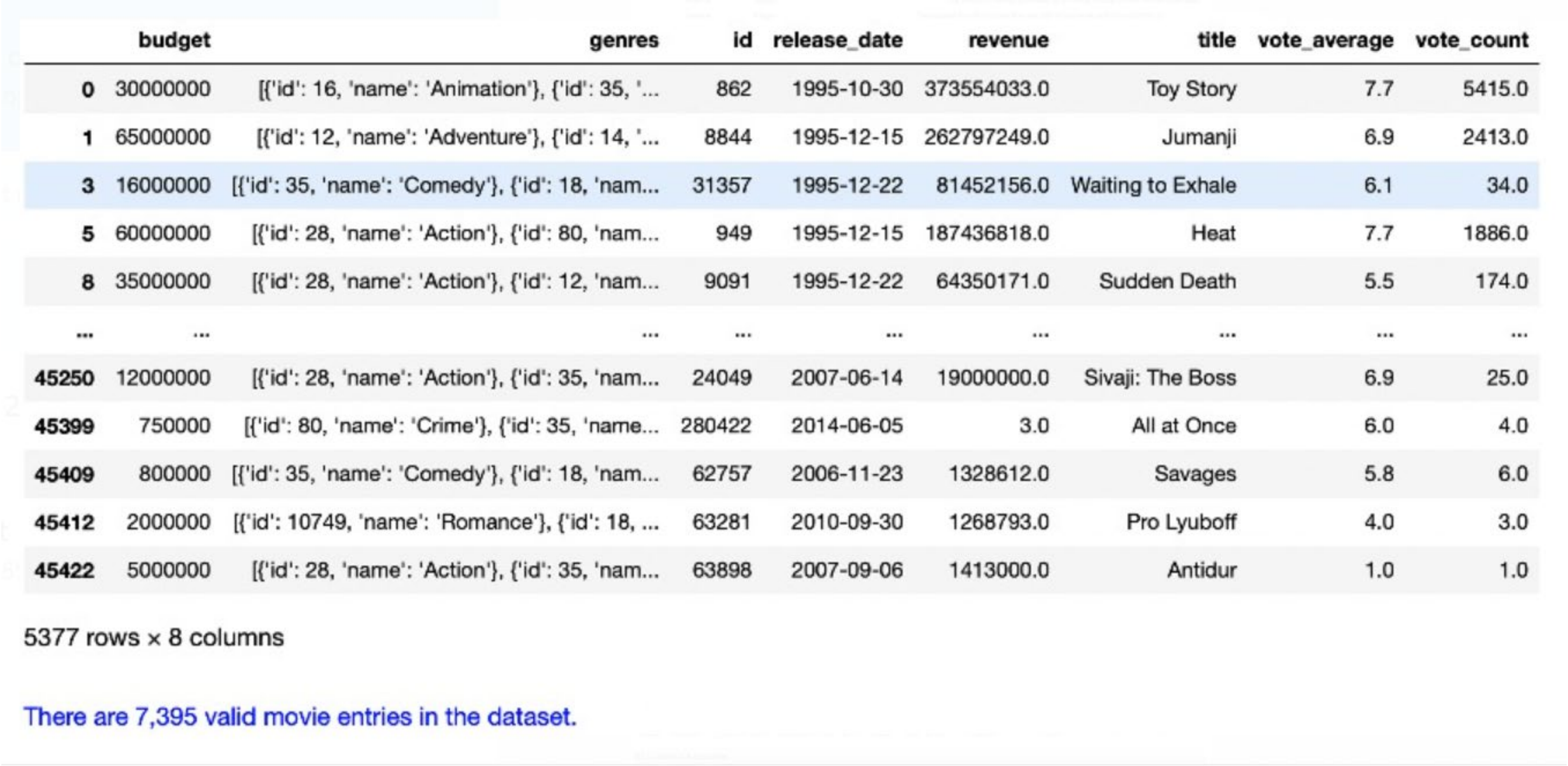
DATA

“Primary Data: The Movies Dataset” - The contents of the movies_metadata.csv are described below.



Column name	Column 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) sptent 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 orginal language of the movie.
original_title	Text	The original title of the movie.
overview	Text	A synopsis of the movie descriving 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) tha tproduced the movie.
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 puclic to see (the movie premiere).
revenue	Integer	The amount of money generated by thr 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 duringthe course of 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.

"Secondary Data: IMDB movies extensive dataset" – The contents of the IMDB movies.csv dataset are described below



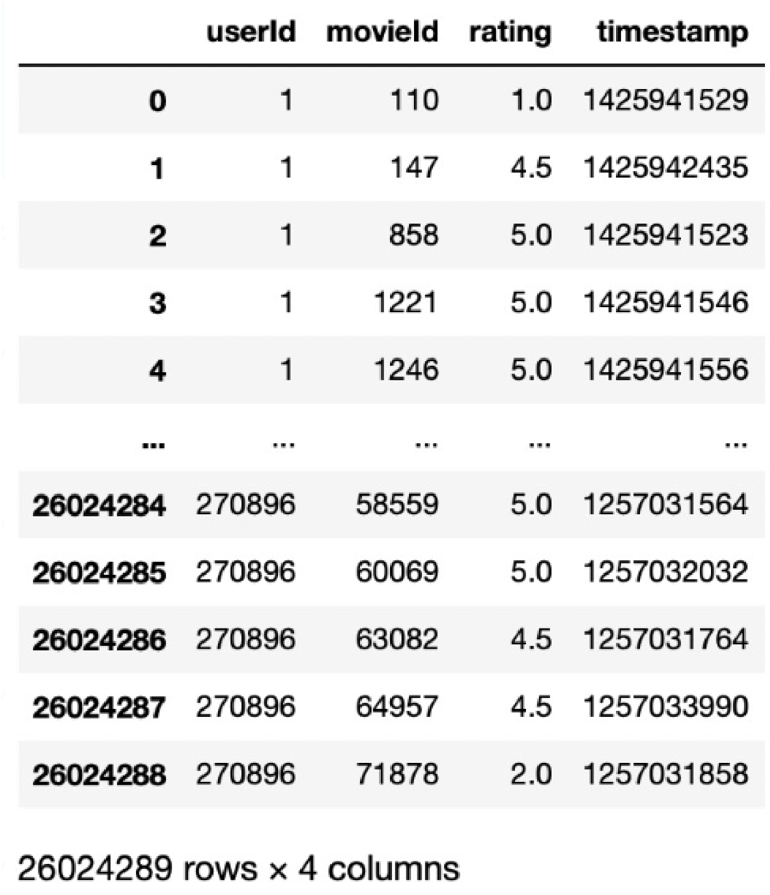
	budget	genres	id	release_date	revenue	title	vote_average	vote_count
0	30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	862	1995-10-30	373554033.0	Toy Story	7.7	5415.0
1	65000000	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	8844	1995-12-15	262797249.0	Jumanji	6.9	2413.0
3	16000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	31357	1995-12-22	81452156.0	Waiting to Exhale	6.1	34.0
5	60000000	[{'id': 28, 'name': 'Action'}, {'id': 80, 'nam...	949	1995-12-15	187436818.0	Heat	7.7	1886.0
8	35000000	[{'id': 28, 'name': 'Action'}, {'id': 12, 'nam...	9091	1995-12-22	64350171.0	Sudden Death	5.5	174.0
...	...	...	...	...	...	...	...	...
45250	12000000	[{'id': 28, 'name': 'Action'}, {'id': 35, 'nam...	24049	2007-06-14	19000000.0	Sivaji: The Boss	6.9	25.0
45399	750000	[{'id': 80, 'name': 'Crime'}, {'id': 35, 'name...	280422	2014-06-05	3.0	All at Once	6.0	4.0
45409	800000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	62757	2006-11-23	1328612.0	Savages	5.8	6.0
45412	2000000	[{'id': 10749, 'name': 'Romance'}, {'id': 18, ...	63281	2010-09-30	1268793.0	Pro Lyuboff	4.0	3.0
45422	5000000	[{'id': 28, 'name': 'Action'}, {'id': 35, 'nam...	63898	2007-09-06	1413000.0	Antidur	1.0	1.0

5377 rows × 8 columns

There are 7,395 valid movie entries in the dataset.

Continued…

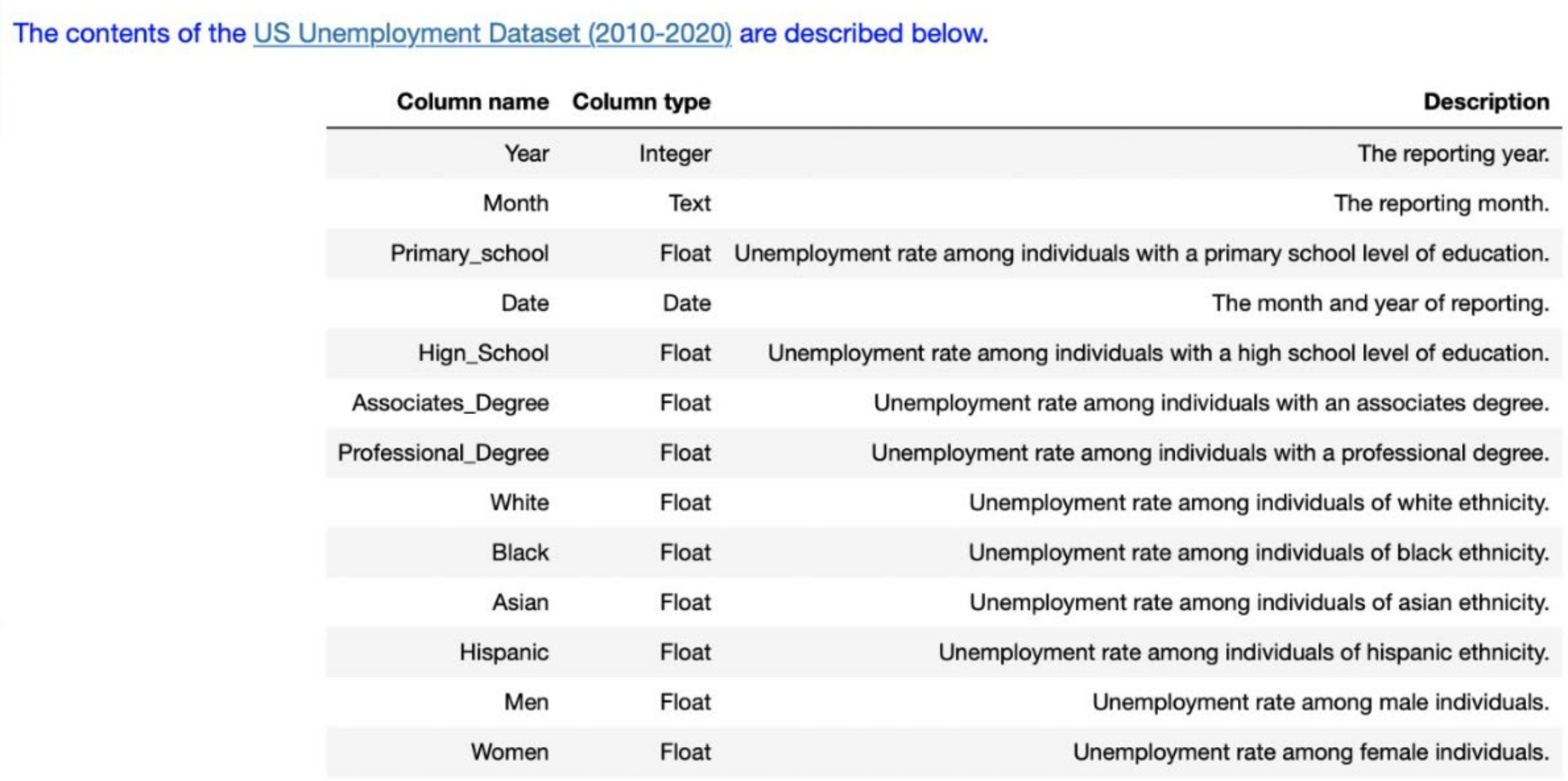
The contents of the IMDB ratings.csv dataset are described below



	userId	movieId	rating	timestamp
0	1	110	1.0	1425941529
1	1	147	4.5	1425942435
2	1	858	5.0	1425941523
3	1	1221	5.0	1425941546
4	1	1246	5.0	1425941556
...	...	...	...	...
26024284	270896	58559	5.0	1257031564
26024285	270896	60069	5.0	1257032032
26024286	270896	63082	4.5	1257031764
26024287	270896	64957	4.5	1257033990
26024288	270896	71878	2.0	1257031858

26024289 rows × 4 columns

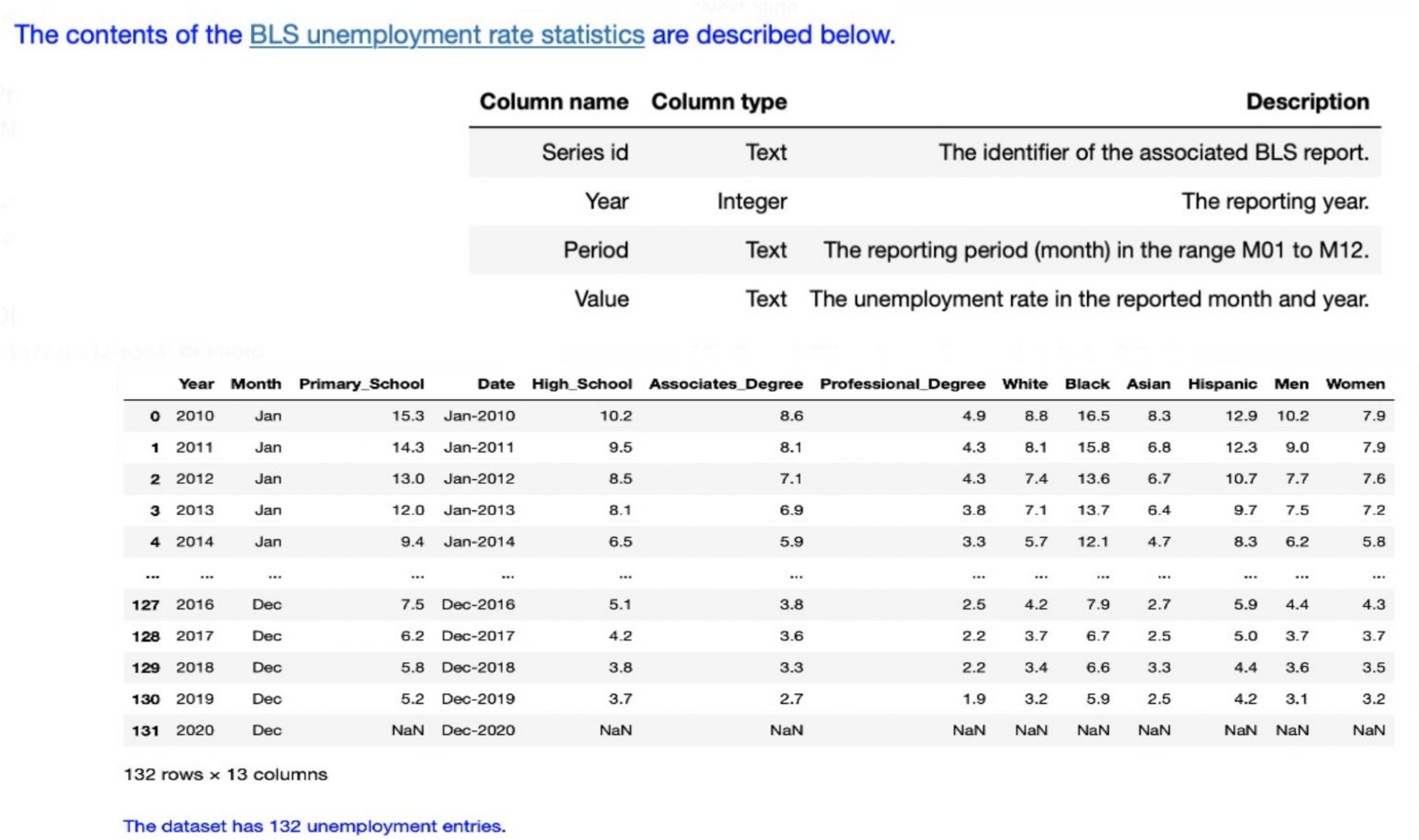
“Secondary Data: (B) Contextual dataset: US Unemployment Dataset (2010- 2020)”



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

Column name	Column type	Description
Year	Integer	The reporting year.
Month	Text	The reporting month.
Primary_school	Float	Unemployment rate among individuals with a primary school level of education.
Date	Date	The month and year of reporting.
Hign_School	Float	Unemployment rate among individuals with a high school level of education.
Associates_Degree	Float	Unemployment rate among individuals with an associates 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.

Secondary Data: (C) Bureau of Labor Statistics (BLS) Unemployment Rates (2010- 2020)



The contents of the BLS unemployment rate statistics are described below.

Column name	Column type	Description
Series id	Text	The identifier of the associated BLS report.
Year	Integer	The reporting year.
Period	Text	The reporting period (month) in the range M01 to M12.
Value	Text	The unemployment rate in the reported month and year.

	Year	Month	Primary_School	Date	High_School	Associates_Degree	Professional_Degree	White	Black	Asian	Hispanic	Men	Women
0	2010	Jan	15.3	Jan-2010	10.2	8.6	4.9	8.8	16.5	8.3	12.9	10.2	7.9
1	2011	Jan	14.3	Jan-2011	9.5	8.1	4.3	8.1	15.8	6.8	12.3	9.0	7.9
2	2012	Jan	13.0	Jan-2012	8.5	7.1	4.3	7.4	13.6	6.7	10.7	7.7	7.6
3	2013	Jan	12.0	Jan-2013	8.1	6.9	3.8	7.1	13.7	6.4	9.7	7.5	7.2
4	2014	Jan	9.4	Jan-2014	6.5	5.9	3.3	5.7	12.1	4.7	8.3	6.2	5.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...
127	2016	Dec	7.5	Dec-2016	5.1	3.8	2.5	4.2	7.9	2.7	5.9	4.4	4.3
128	2017	Dec	6.2	Dec-2017	4.2	3.6	2.2	3.7	6.7	2.5	5.0	3.7	3.7
129	2018	Dec	5.8	Dec-2018	3.8	3.3	2.2	3.4	6.6	3.3	4.4	3.6	3.5
130	2019	Dec	5.2	Dec-2019	3.7	2.7	1.9	3.2	5.9	2.5	4.2	3.1	3.2
131	2020	Dec	NaN	Dec-2020	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

132 rows × 13 columns

The dataset has 132 unemployment entries.

DATA PREPARATION STEPS

- To produce an Analytic Base Table (ABT) data structure.
- Data wrangling process: It consisted of the following steps:
  - Discovery
  - Structuring
  - Cleaning
  - Validating
  - Enrichment
  - Aggregation
  - Integration
  - publishing.

Data Preparation Steps: Analytic Base Table (ABT) data structure

- 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.
- Note: 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.

Data Wrangling Steps: (A) Discovery

- Datasets were explored initially in its raw form, to better understand the dataset
- 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.

Structuring

Data structuring techniques were used to:

1) normalize the data
2) To 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.

Cleaning

- Data cleansing or data cleaning is the process of detecting and correcting corrupt or inaccurate records from a given dataset
- The data cleaning process had outliers stripped from the datasets
- For instance, in the given data movie with movie revenue value as 0 were considered as outlier and stripped off
- Also movies released straight to video or with any status other than RELEASED were also dropped from the analysis
- Entries with missing values for required entries, such as the release date were also excluded.
- After data cleansing our primary dataset was reduced from 45,466 individual movie entries to 7,406 entries.

Validating

- Data validation refers to the process of ensuring the accuracy and quality of data
- This step is 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.

Enrichment

- Data Enrichment allows companies to make their raw data useful
- It also allows businesses to add additional as well as missing data to the original data set to make it more useful.
- Here in this step we performed merging a dataset with third-party data from a reputable data source
- Initial enrichment was performed by merging unemployment statistics with unemployment rates published by the Bureau of Labor Statistics.

Aggregation

- Data aggregation is any process whereby data is gathered and expressed in a summary form.
- When data is aggregated, atomic data rows typically gathered from multiple sources are replaced with totals or summary statistics.
- For the aggregation step, user rating information by numerous users was summarized through the computing of aggregates grouped by individual movies.

Integration

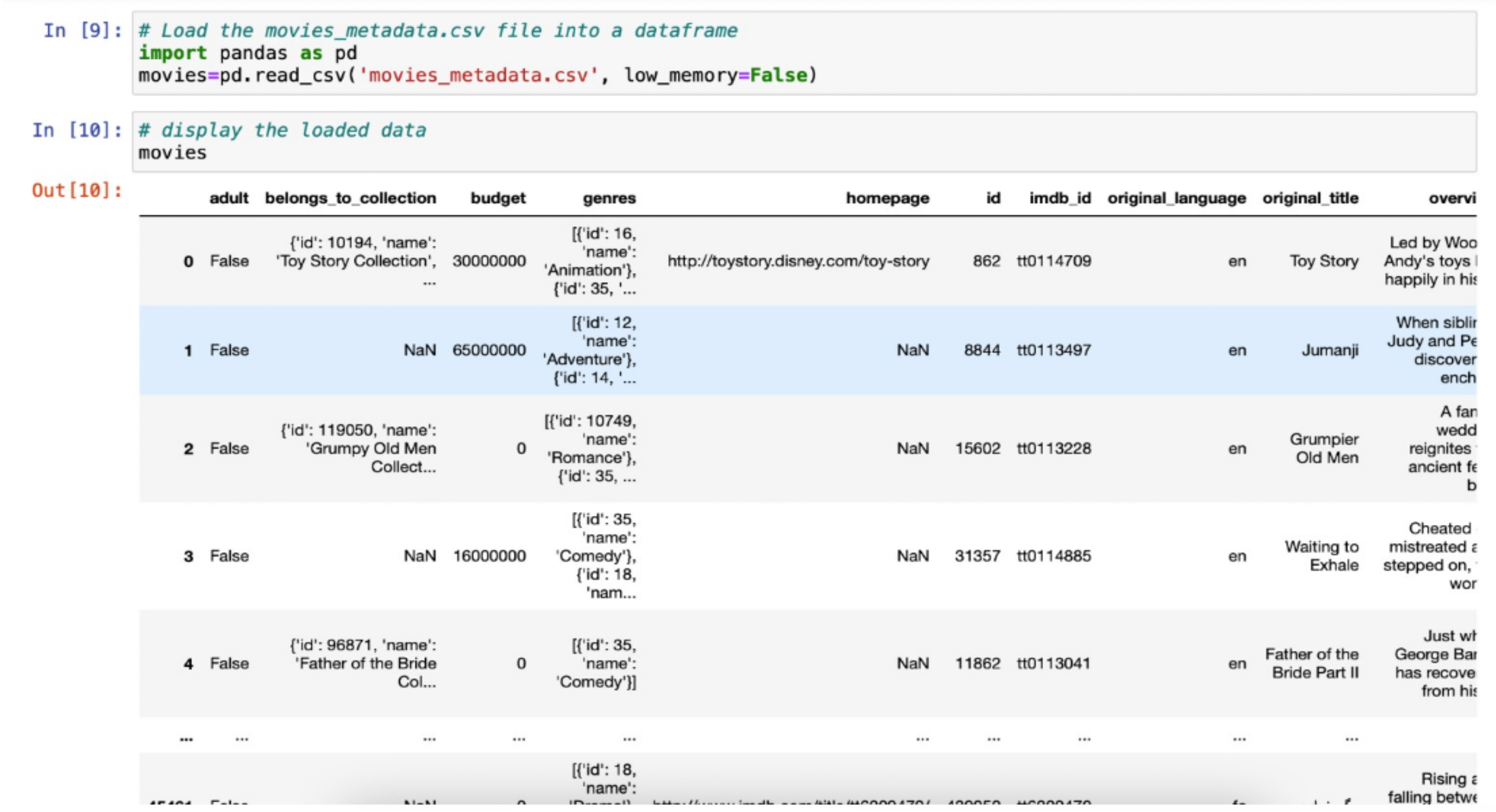
- Integration of dataset is performed to combine the required and useful information from the dataset
- Once all datasets were cleaned and validated, they were combined into a single unified view, using common fields
- For instance: movie identifier to associate movies with ratings, and both the year 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.

Publishing

- After performing all the above the steps, we need to display the resulting data
- 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.

Implementation

Loading the Dataset

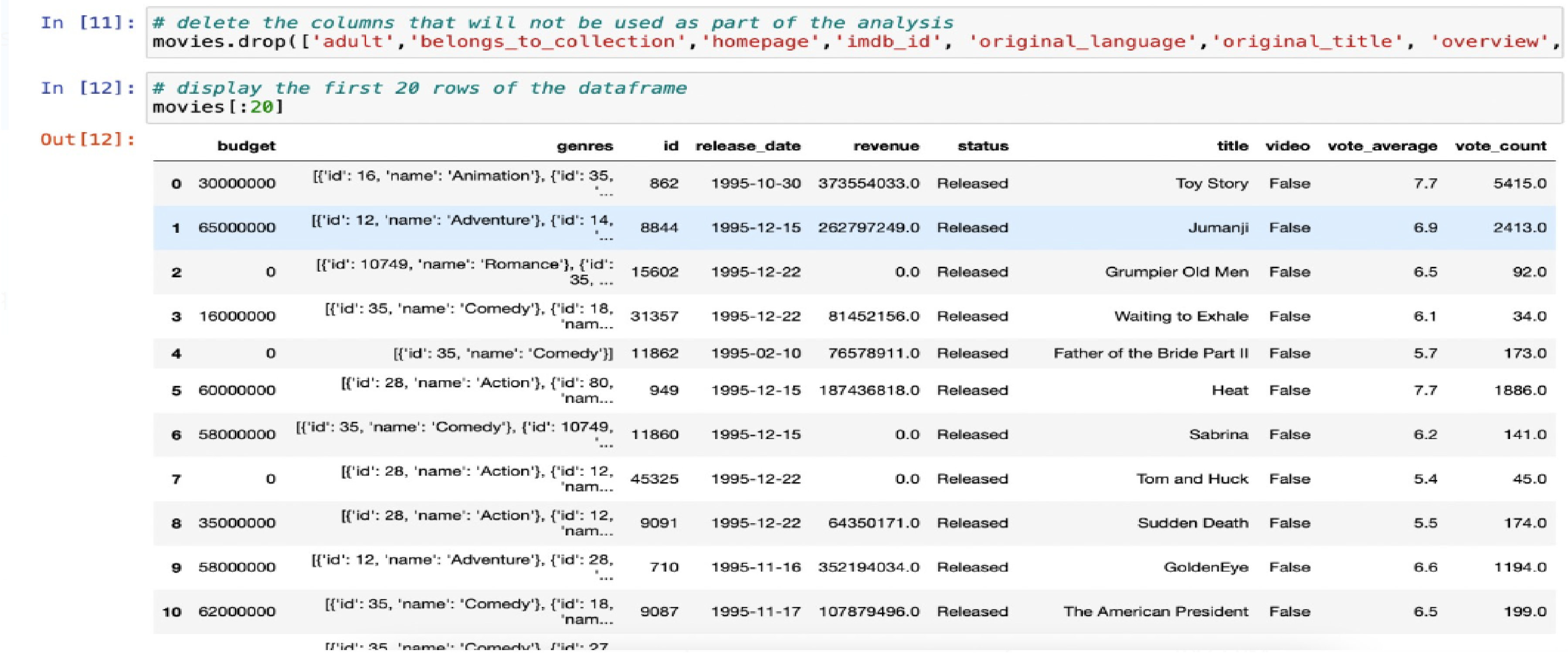


```
In [9]:  # Load the movies_metadata.csv file into a dataframe
         import pandas as pd
         movies=pd.read_csv('movies_metadata.csv', low_memory=False)

In [10]: # display the loaded data
         movies

Out[10]:
```

	adult	belongs_to_collection	budget	genres	homepage	id	imdb_id	original_language	original_title	overvi
0	False	{'id': 10194, 'name': 'Toy Story Collection', ...	30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	http://toystory.disney.com/toy-story	862	tt0114709	en	Toy Story	Led by Woo Andy's toys happily in his
1	False	NaN	65000000	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	NaN	8844	tt0113497	en	Jumanji	When siblir Judy and Pe discover ench
2	False	{'id': 119050, 'name': 'Grumpy Old Men Collect...	0	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...	NaN	15602	tt0113228	en	Grumpier Old Men	A fan wedd reignites ancient fe b
3	False	NaN	16000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	NaN	31357	tt0114885	en	Waiting to Exhale	Cheated mistreated a stepped on, wor
4	False	{'id': 96871, 'name': 'Father of the Bride Col...	0	[{'id': 35, 'name': 'Comedy'}]	NaN	11862	tt0113041	en	Father of the Bride Part II	Just wh George Bar has recove from his
...	...	...	...	...	...	...	...	...	...	
45461	False	NaN	0	[{'id': 18, 'name': 'Drama'},	http://www.imdb.com/title/tt6209470/	439050	tt6209470	fa		Rising a falling betwe

Continued..

- Here we perform Data Cleansing



```
In [11]: # delete the columns that will not be used as part of the analysis
         movies.drop(['adult','belongs_to_collection','homepage','imdb_id', 'original_language','original_title', 'overview',
```

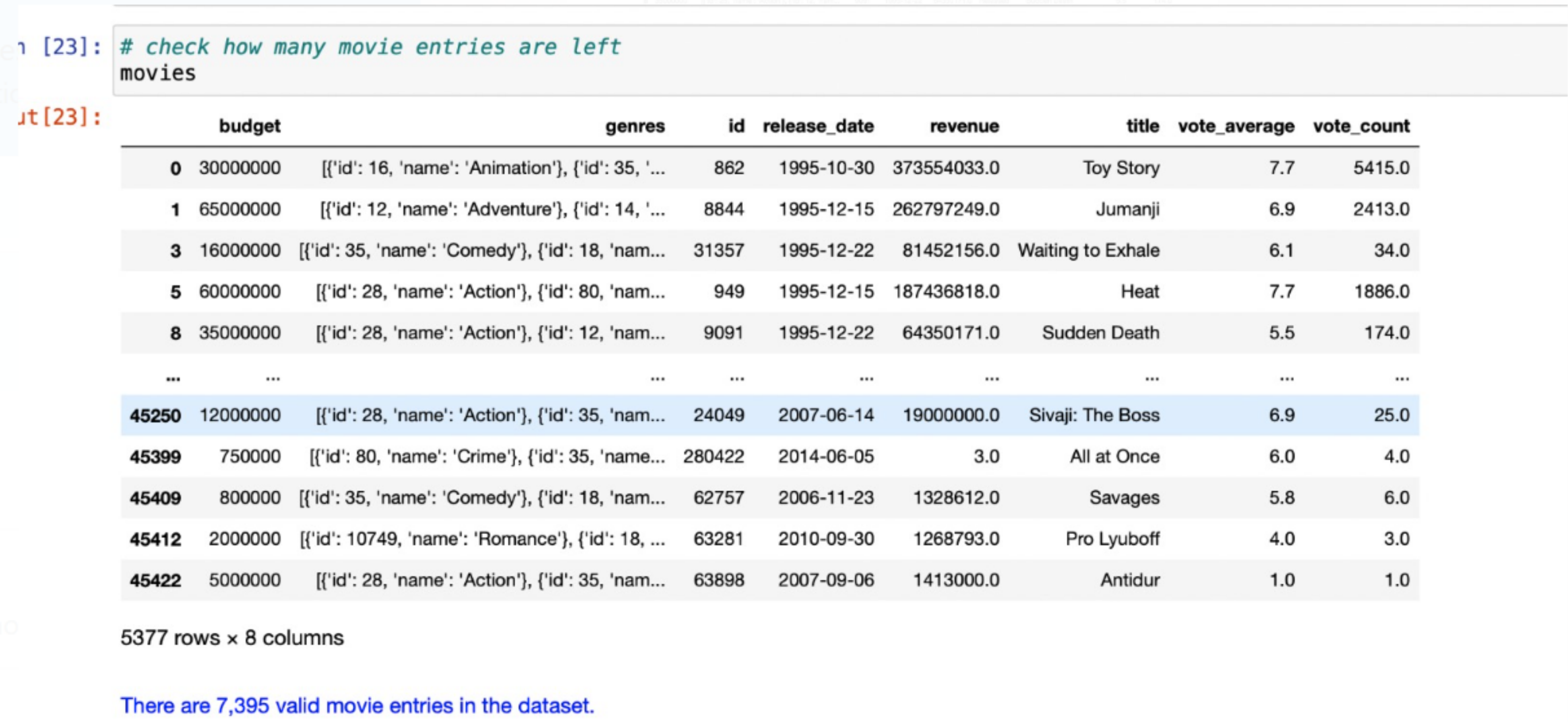
```
In [12]: # display the first 20 rows of the dataframe
         movies[:20]
```

Out[12]:

	budget	genres	id	release_date	revenue	status	title	video	vote_average	vote_count
0	30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	862	1995-10-30	373554033.0	Released	Toy Story	False	7.7	5415.0
1	65000000	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	8844	1995-12-15	262797249.0	Released	Jumanji	False	6.9	2413.0
2	0	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...	15602	1995-12-22	0.0	Released	Grumpier Old Men	False	6.5	92.0
3	16000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	31357	1995-12-22	81452156.0	Released	Waiting to Exhale	False	6.1	34.0
4	0	[{'id': 35, 'name': 'Comedy'}]	11862	1995-02-10	76578911.0	Released	Father of the Bride Part II	False	5.7	173.0
5	60000000	[{'id': 28, 'name': 'Action'}, {'id': 80, 'nam...	949	1995-12-15	187436818.0	Released	Heat	False	7.7	1886.0
6	58000000	[{'id': 35, 'name': 'Comedy'}, {'id': 10749, '...	11860	1995-12-15	0.0	Released	Sabrina	False	6.2	141.0
7	0	[{'id': 28, 'name': 'Action'}, {'id': 12, 'nam...	45325	1995-12-22	0.0	Released	Tom and Huck	False	5.4	45.0
8	35000000	[{'id': 28, 'name': 'Action'}, {'id': 12, 'nam...	9091	1995-12-22	64350171.0	Released	Sudden Death	False	5.5	174.0
9	58000000	[{'id': 12, 'name': 'Adventure'}, {'id': 28, '...	710	1995-11-16	352194034.0	Released	GoldenEye	False	6.6	1194.0
10	62000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	9087	1995-11-17	107879496.0	Released	The American President	False	6.5	199.0
		[{'id': 35, 'name': 'Comedy'}, {'id': 27,								

- Like this we need to perform required data preparation steps for the given datasets

Continued..



```
n [23]: # check how many movie entries are left
        movies
```

ut[23]:

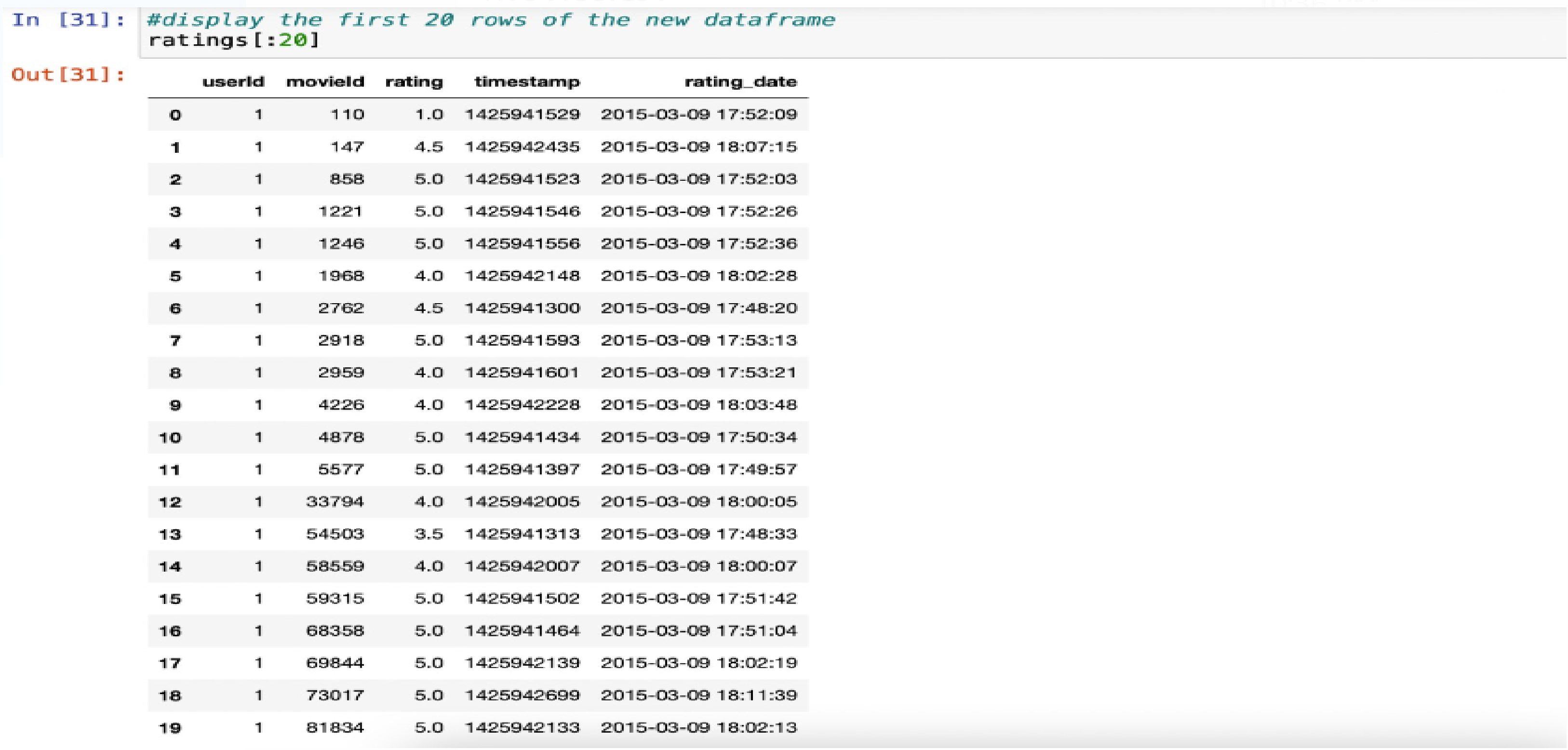
	budget	genres	id	release_date	revenue	title	vote_average	vote_count
0	30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	862	1995-10-30	373554033.0	Toy Story	7.7	5415.0
1	65000000	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	8844	1995-12-15	262797249.0	Jumanji	6.9	2413.0
3	16000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	31357	1995-12-22	81452156.0	Waiting to Exhale	6.1	34.0
5	60000000	[{'id': 28, 'name': 'Action'}, {'id': 80, 'nam...	949	1995-12-15	187436818.0	Heat	7.7	1886.0
8	35000000	[{'id': 28, 'name': 'Action'}, {'id': 12, 'nam...	9091	1995-12-22	64350171.0	Sudden Death	5.5	174.0
...	...	...	...	...	...	...	...	...
45250	12000000	[{'id': 28, 'name': 'Action'}, {'id': 35, 'nam...	24049	2007-06-14	19000000.0	Sivaji: The Boss	6.9	25.0
45399	750000	[{'id': 80, 'name': 'Crime'}, {'id': 35, 'name...	280422	2014-06-05	3.0	All at Once	6.0	4.0
45409	800000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	62757	2006-11-23	1328612.0	Savages	5.8	6.0
45412	2000000	[{'id': 10749, 'name': 'Romance'}, {'id': 18, ...	63281	2010-09-30	1268793.0	Pro Lyuboff	4.0	3.0
45422	5000000	[{'id': 28, 'name': 'Action'}, {'id': 35, 'nam...	63898	2007-09-06	1413000.0	Antidur	1.0	1.0

5377 rows × 8 columns

There are 7,395 valid movie entries in the dataset.

Movies.csv

Continued..



In [31]:

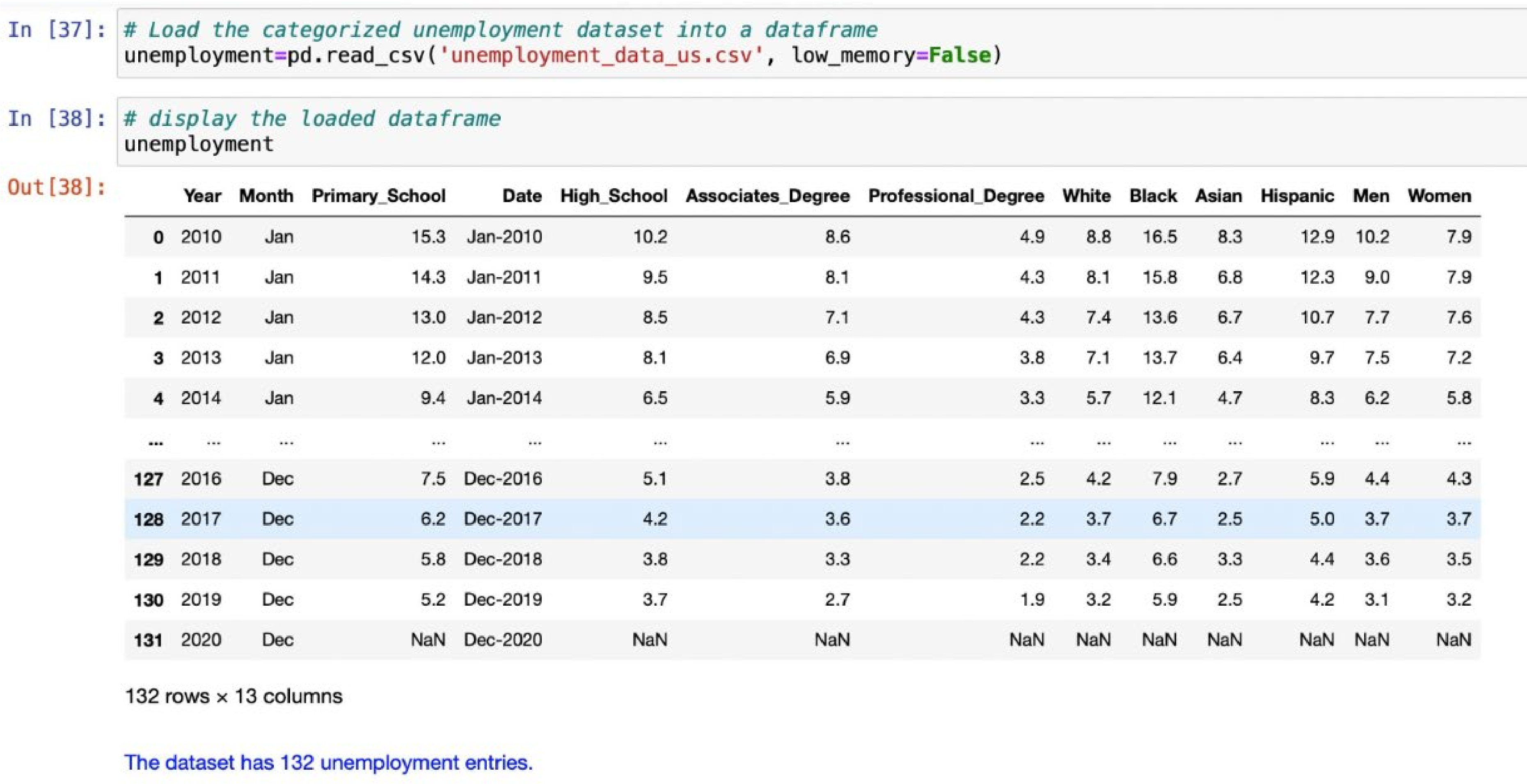
```
#display the first 20 rows of the new dataframe
ratings[:20]
```

Out[31]:

	userId	movieId	rating	timestamp	rating_date
0	1	110	1.0	1425941529	2015-03-09 17:52:09
1	1	147	4.5	1425942435	2015-03-09 18:07:15
2	1	858	5.0	1425941523	2015-03-09 17:52:03
3	1	1221	5.0	1425941546	2015-03-09 17:52:26
4	1	1246	5.0	1425941556	2015-03-09 17:52:36
5	1	1968	4.0	1425942148	2015-03-09 18:02:28
6	1	2762	4.5	1425941300	2015-03-09 17:48:20
7	1	2918	5.0	1425941593	2015-03-09 17:53:13
8	1	2959	4.0	1425941601	2015-03-09 17:53:21
9	1	4226	4.0	1425942228	2015-03-09 18:03:48
10	1	4878	5.0	1425941434	2015-03-09 17:50:34
11	1	5577	5.0	1425941397	2015-03-09 17:49:57
12	1	33794	4.0	1425942005	2015-03-09 18:00:05
13	1	54503	3.5	1425941313	2015-03-09 17:48:33
14	1	58559	4.0	1425942007	2015-03-09 18:00:07
15	1	59315	5.0	1425941502	2015-03-09 17:51:42
16	1	68358	5.0	1425941464	2015-03-09 17:51:04
17	1	69844	5.0	1425942139	2015-03-09 18:02:19
18	1	73017	5.0	1425942699	2015-03-09 18:11:39
19	1	81834	5.0	1425942133	2015-03-09 18:02:13

ratings.csv

Continued..



In [37]:
```
# Load the categorized unemployment dataset into a dataframe
unemployment=pd.read_csv('unemployment_data_us.csv', low_memory=False)
```

In [38]:
```
# display the loaded dataframe
unemployment
```

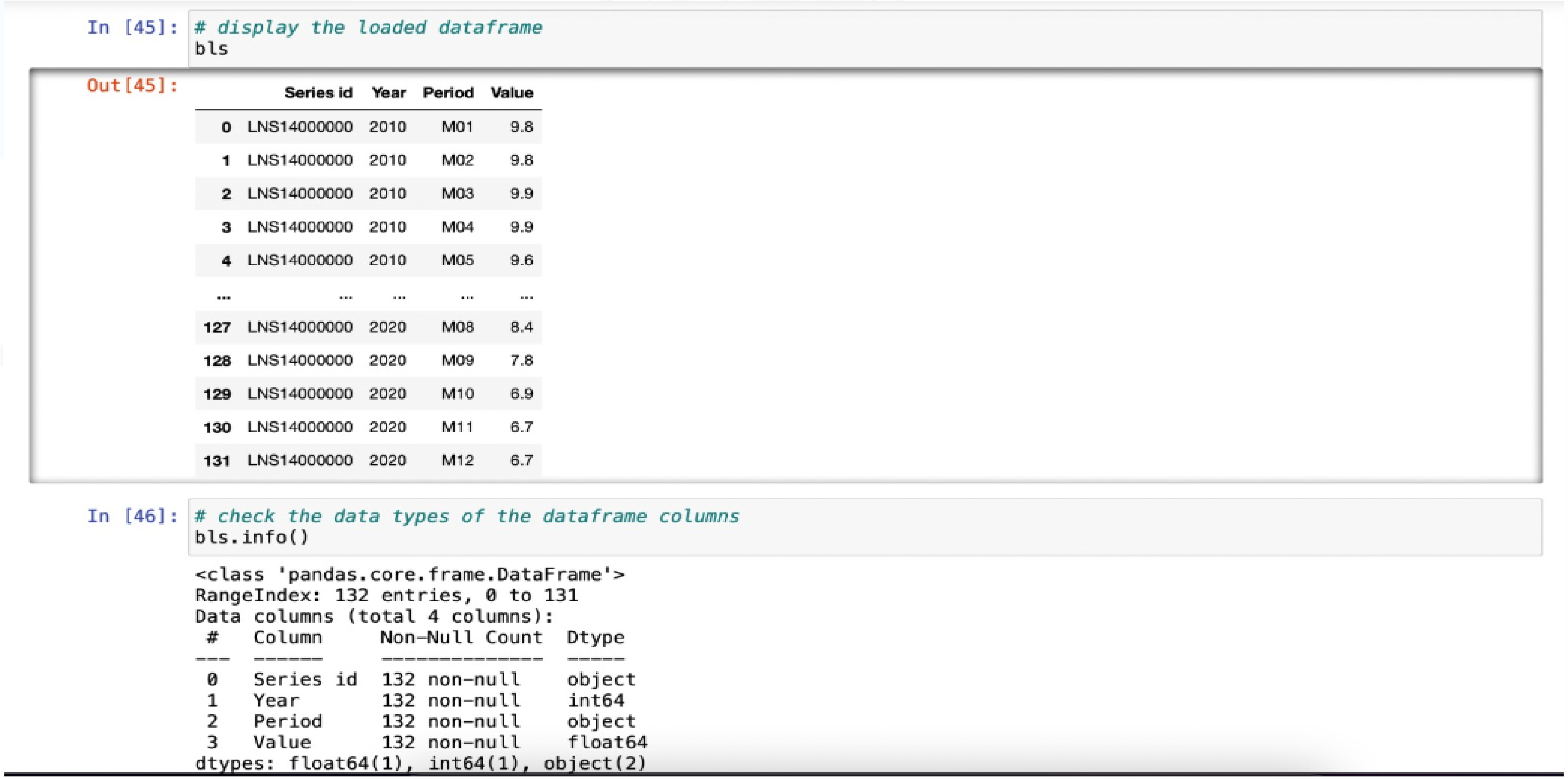
Out[38]:

	Year	Month	Primary_School	Date	High_School	Associates_Degree	Professional_Degree	White	Black	Asian	Hispanic	Men	Women
0	2010	Jan	15.3	Jan-2010	10.2	8.6	4.9	8.8	16.5	8.3	12.9	10.2	7.9
1	2011	Jan	14.3	Jan-2011	9.5	8.1	4.3	8.1	15.8	6.8	12.3	9.0	7.9
2	2012	Jan	13.0	Jan-2012	8.5	7.1	4.3	7.4	13.6	6.7	10.7	7.7	7.6
3	2013	Jan	12.0	Jan-2013	8.1	6.9	3.8	7.1	13.7	6.4	9.7	7.5	7.2
4	2014	Jan	9.4	Jan-2014	6.5	5.9	3.3	5.7	12.1	4.7	8.3	6.2	5.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...
127	2016	Dec	7.5	Dec-2016	5.1	3.8	2.5	4.2	7.9	2.7	5.9	4.4	4.3
128	2017	Dec	6.2	Dec-2017	4.2	3.6	2.2	3.7	6.7	2.5	5.0	3.7	3.7
129	2018	Dec	5.8	Dec-2018	3.8	3.3	2.2	3.4	6.6	3.3	4.4	3.6	3.5
130	2019	Dec	5.2	Dec-2019	3.7	2.7	1.9	3.2	5.9	2.5	4.2	3.1	3.2
131	2020	Dec	NaN	Dec-2020	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

132 rows × 13 columns

The dataset has 132 unemployment entries.

Displaying processed data for bls statistics dataset



```
In [45]: # display the loaded dataframe
         bls
```

Out[45]:

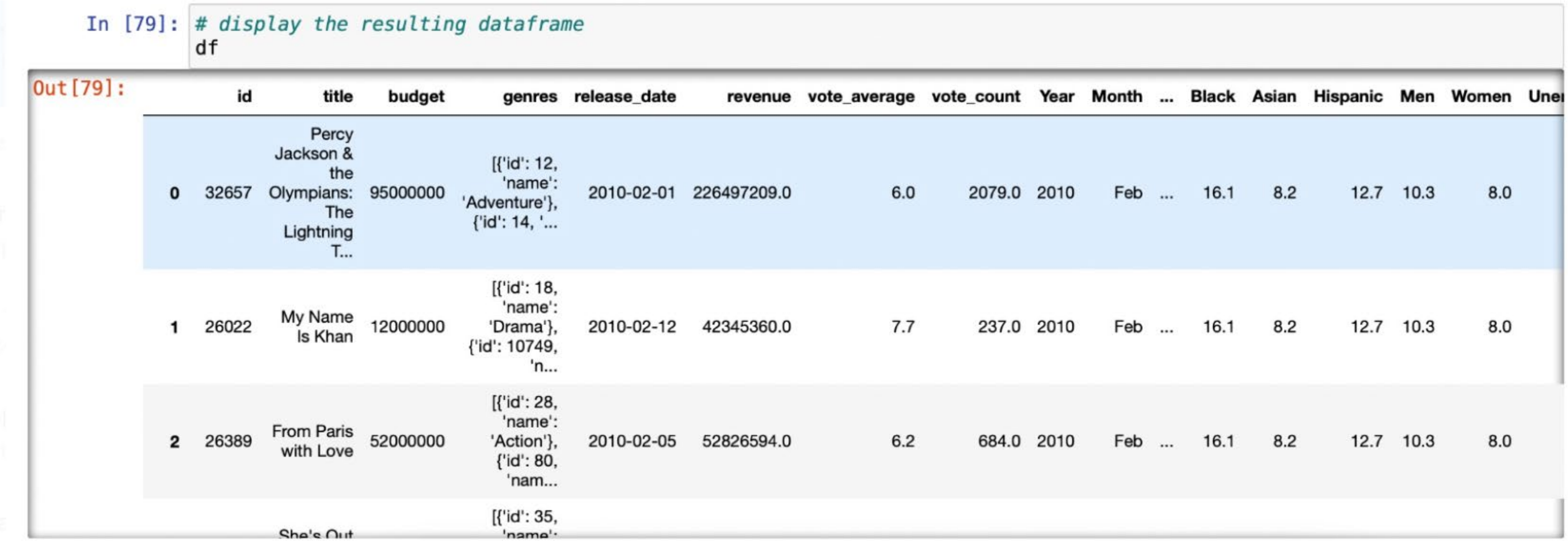
	Series id	Year	Period	Value
0	LNS14000000	2010	M01	9.8
1	LNS14000000	2010	M02	9.8
2	LNS14000000	2010	M03	9.9
3	LNS14000000	2010	M04	9.9
4	LNS14000000	2010	M05	9.6
...	...	...	...	...
127	LNS14000000	2020	M08	8.4
128	LNS14000000	2020	M09	7.8
129	LNS14000000	2020	M10	6.9
130	LNS14000000	2020	M11	6.7
131	LNS14000000	2020	M12	6.7

```
In [46]: # check the data types of the dataframe columns
         bls.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 132 entries, 0 to 131
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Series id  132 non-null    object
 1   Year       132 non-null    int64
 2   Period     132 non-null    object
 3   Value      132 non-null    float64
dtypes: float64(1), int64(1), object(2)
```

Displaying processed data for bls statistics dataset

Combining all the datasets with carefully setting the combining constraints



In [79]:
```
# display the resulting dataframe
df
```

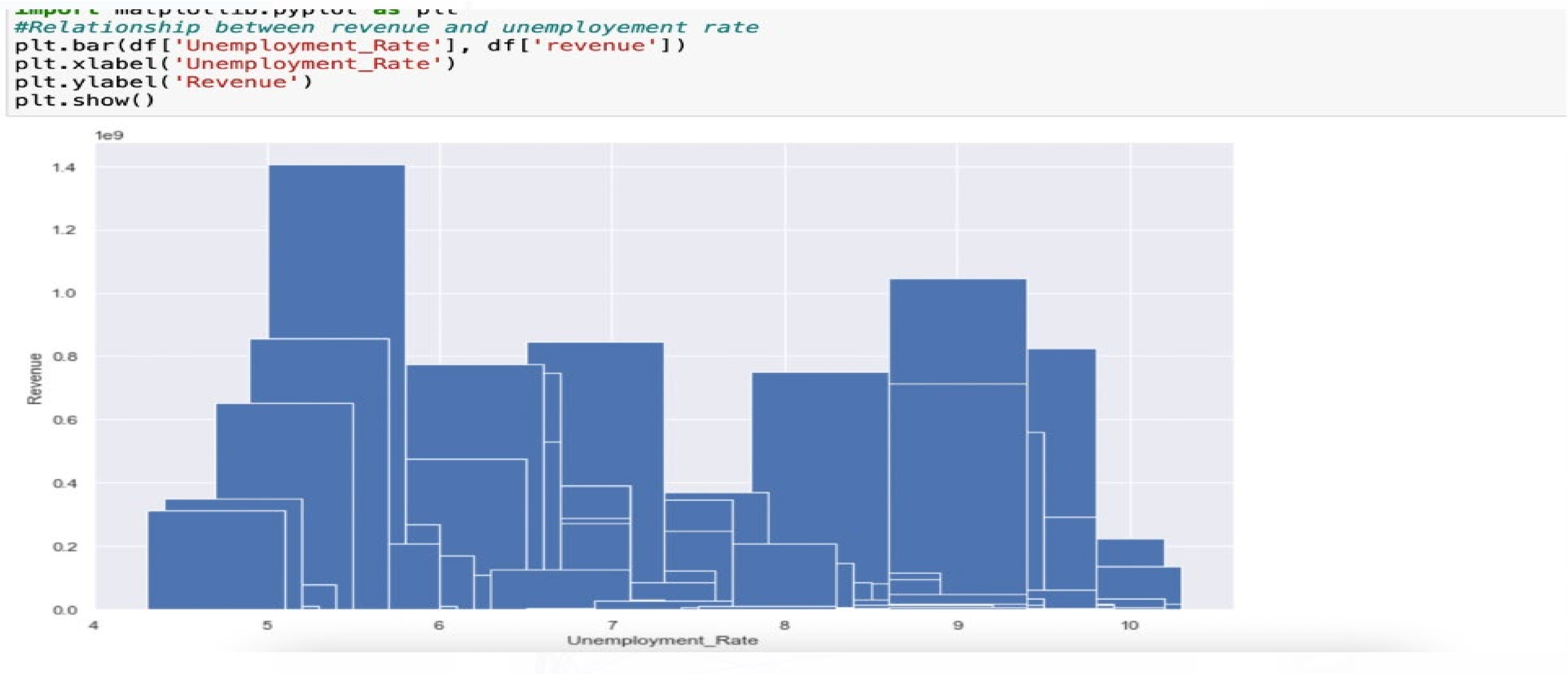
Out[79]:

	id	title	budget	genres	release_date	revenue	vote_average	vote_count	Year	Month	...	Black	Asian	Hispanic	Men	Women	Une
0	32657	Percy Jackson & the Olympians: The Lightning T...	95000000	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	2010-02-01	226497209.0	6.0	2079.0	2010	Feb	...	16.1	8.2	12.7	10.3	8.0	
1	26022	My Name Is Khan	12000000	[{'id': 18, 'name': 'Drama'}, {'id': 10749, 'n...	2010-02-12	42345360.0	7.7	237.0	2010	Feb	...	16.1	8.2	12.7	10.3	8.0	
2	26389	From Paris with Love	52000000	[{'id': 28, 'name': 'Action'}, {'id': 80, 'nam...	2010-02-05	52826594.0	6.2	684.0	2010	Feb	...	16.1	8.2	12.7	10.3	8.0	
		She's Out		[{'id': 35, 'name':													

Combining all the datasets with carefully setting the combining constraints

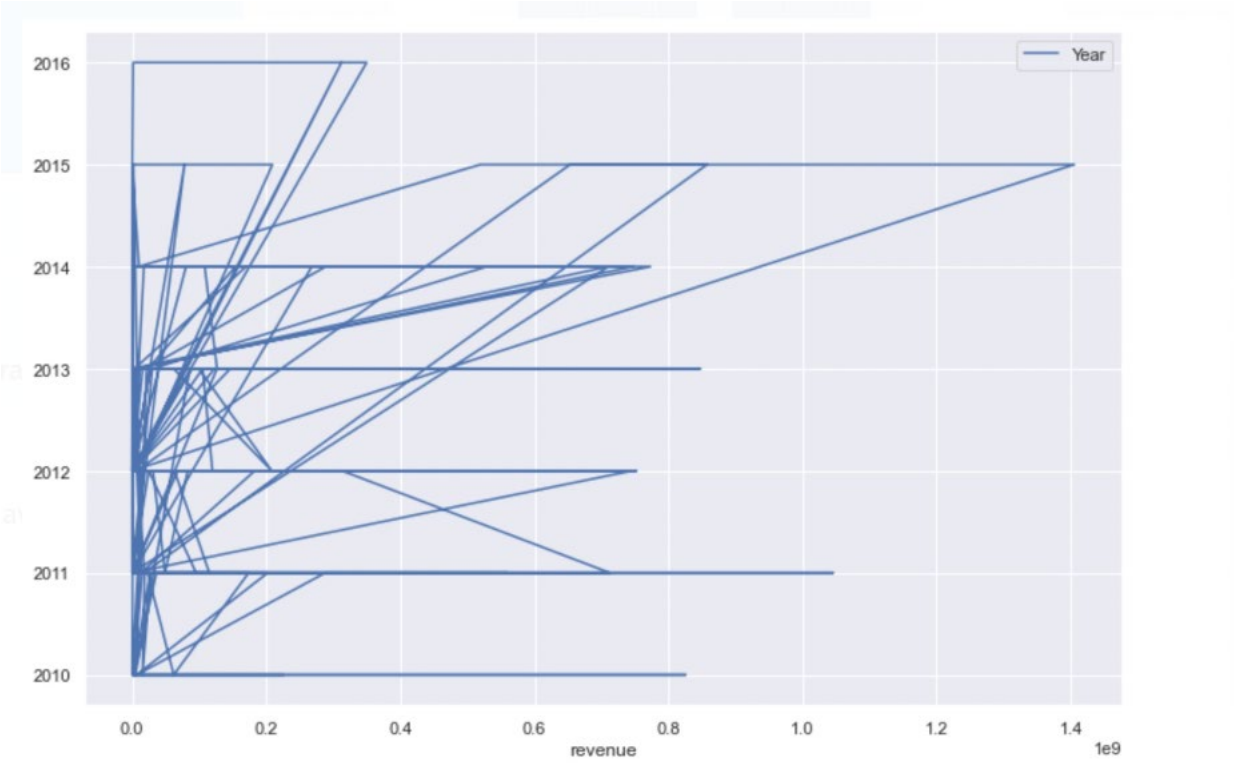
Results

- VIZUALIZATION AND ANALYSIS



```
import matplotlib.pyplot as plt
#Relationship between revenue and unemployement rate
plt.bar(df['Unemployment_Rate'], df['revenue'])
plt.xlabel('Unemployment_Rate')
plt.ylabel('Revenue')
plt.show()
```

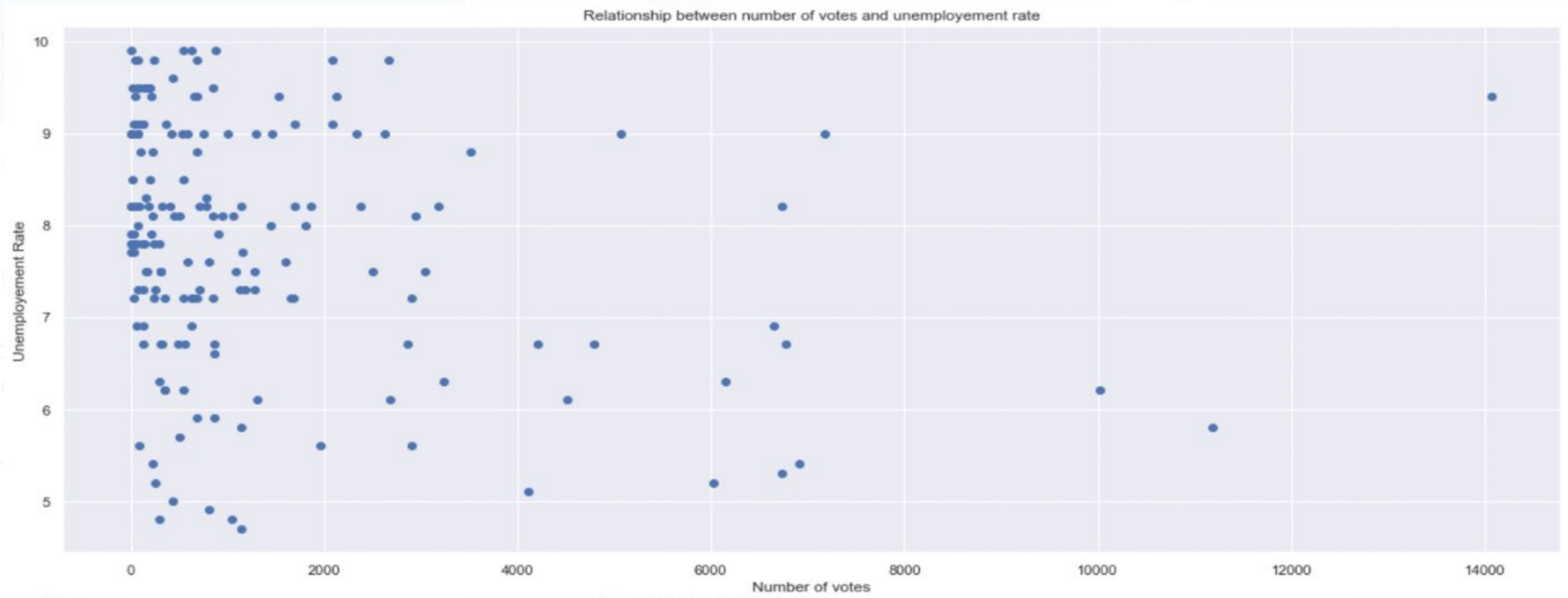
Analyzing the revenue with respect to each year



Analyzing the revenue with respect to each year

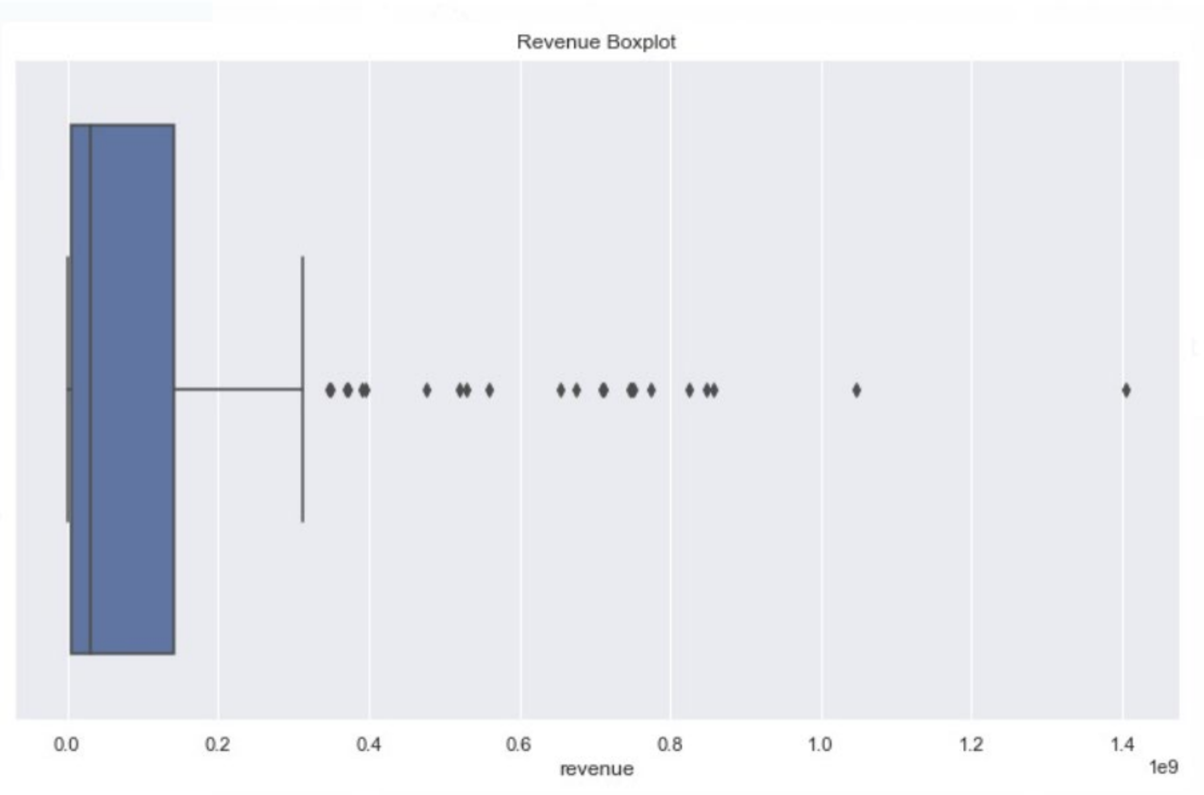
Visualizing the relation between Number of votes and unemployment rate

- This is result obtained from combining rating and unemployment datasets



Vizualizing the relation between Number of votes and unemployement rate

Revenue Boxplot

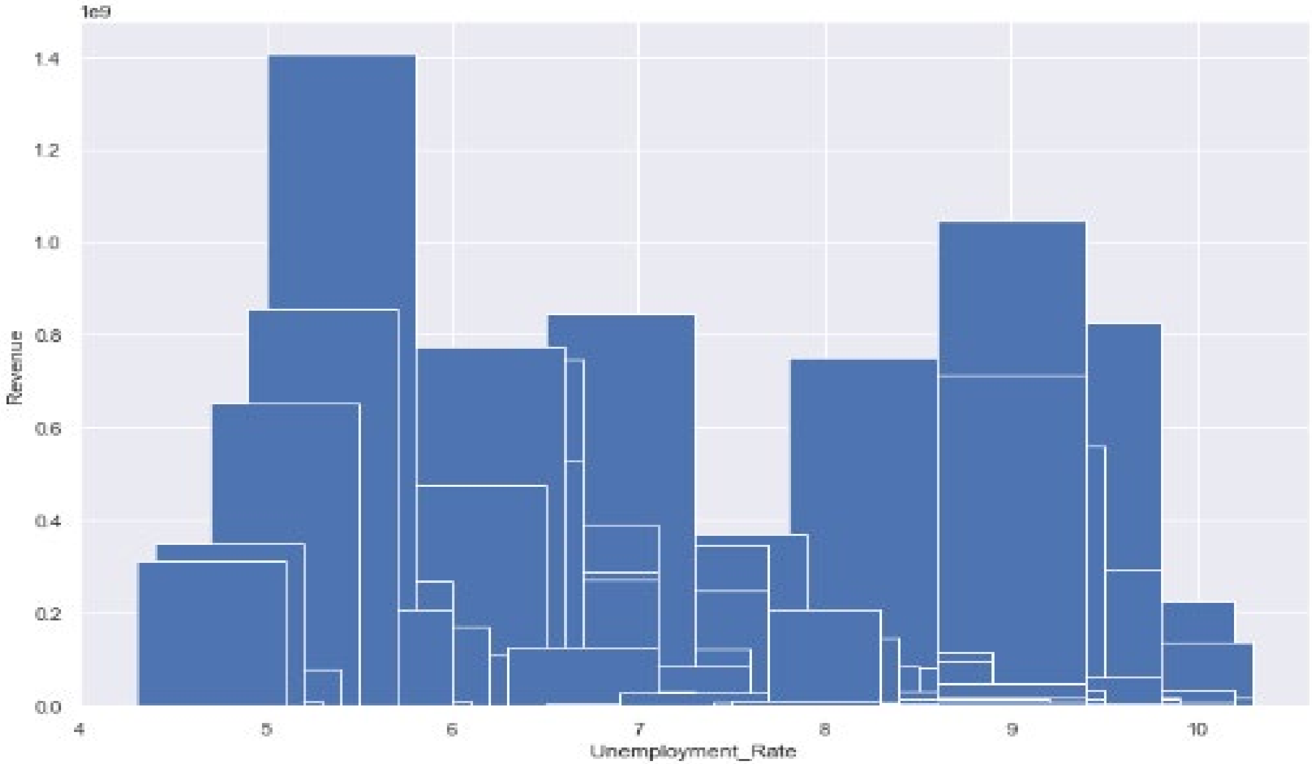


Revenue Boxplot

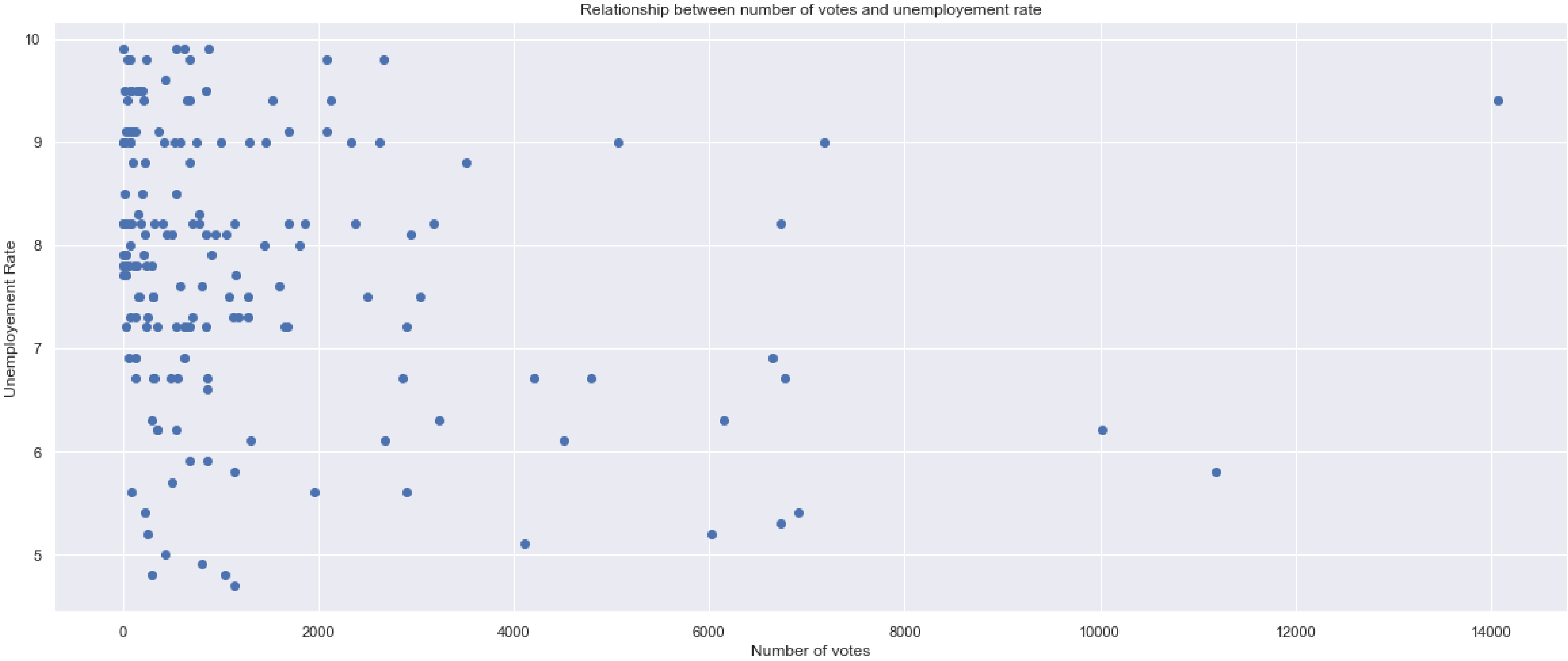
EXPLORATORY DATA ANALYSIS (EDA)

- Unemployment Rate VS Revenue

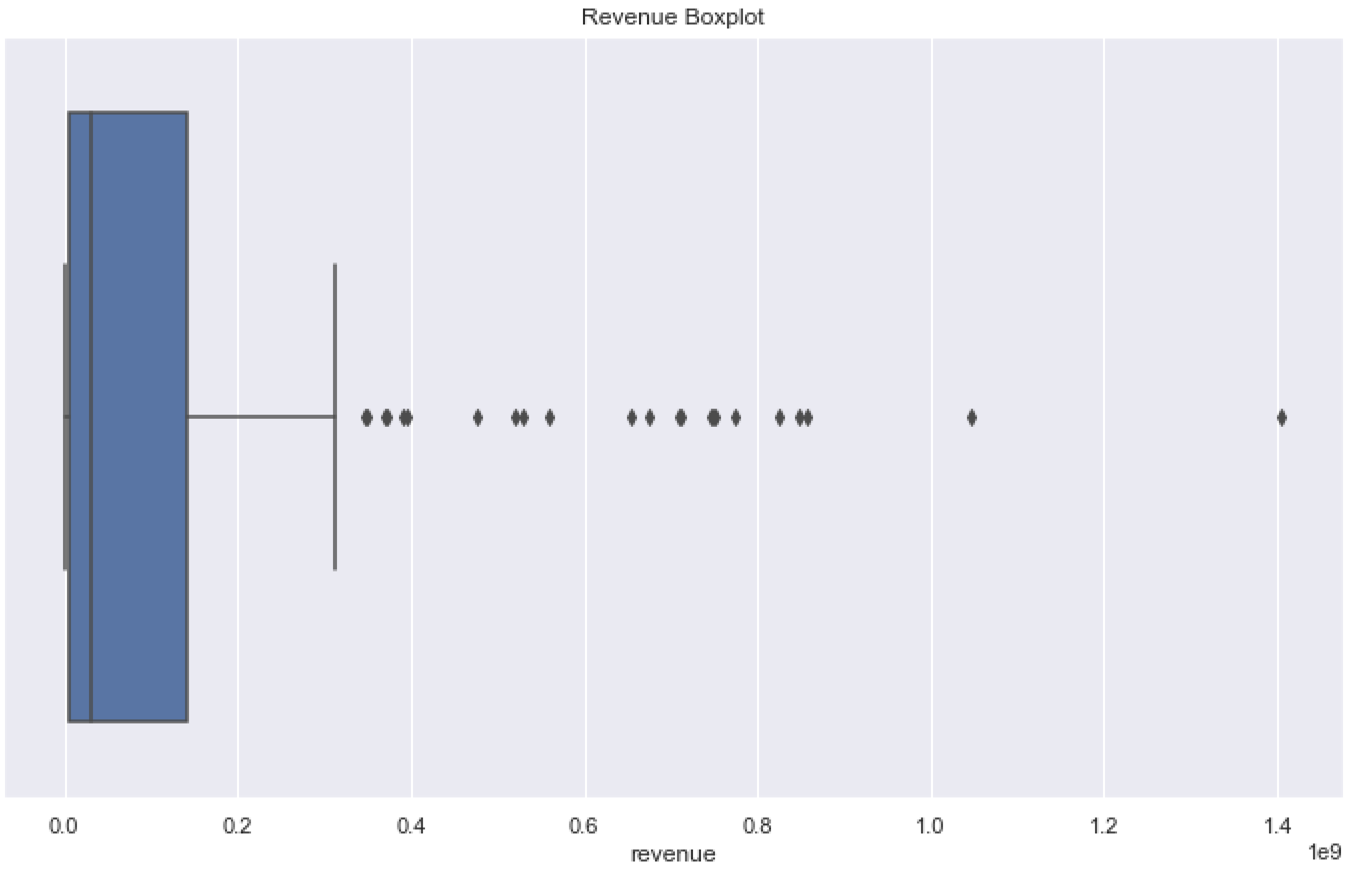
Movies that earned higher revenue tend to have lower employment



Unemployment Rate VS Number of Votes

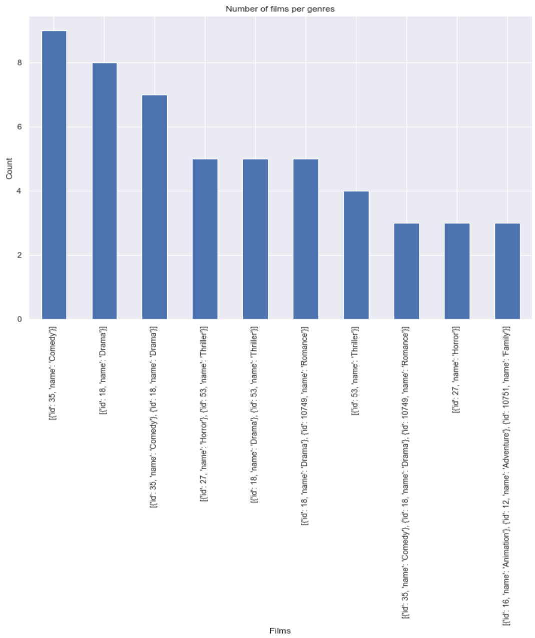


Revenue Boxplot



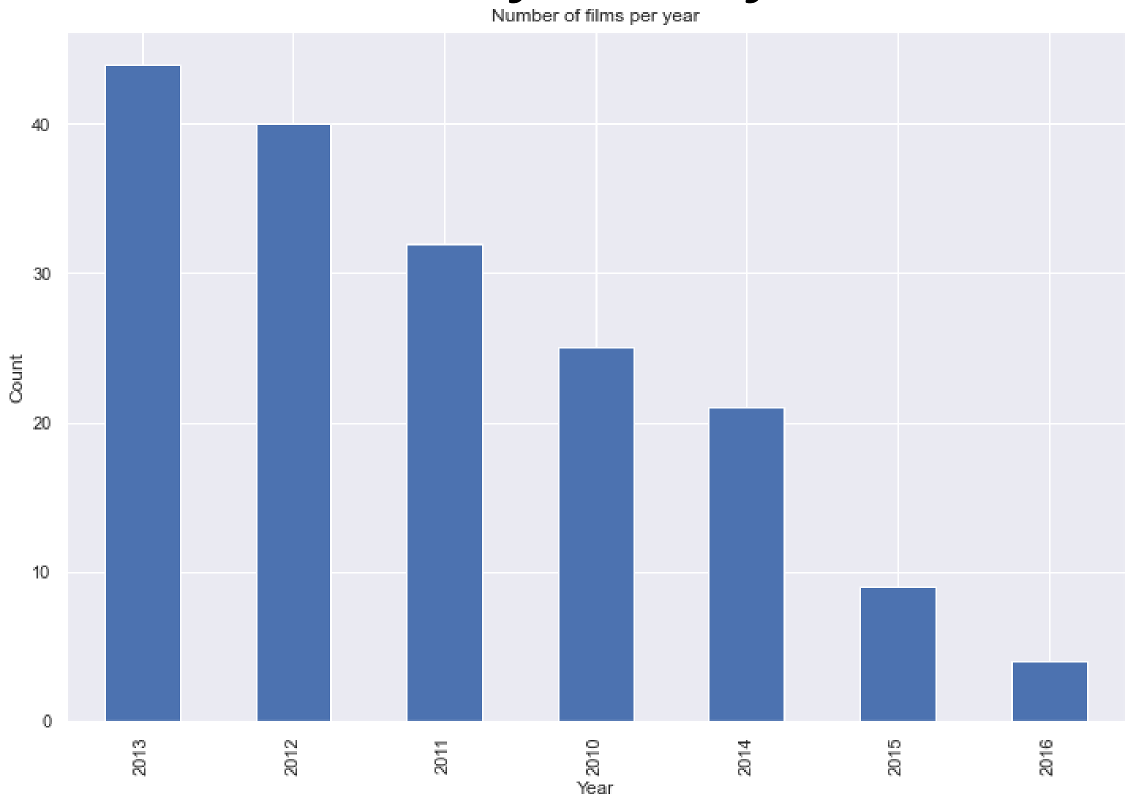
Number of films per genre

- Most movies are of the comedy genre followed by thriller



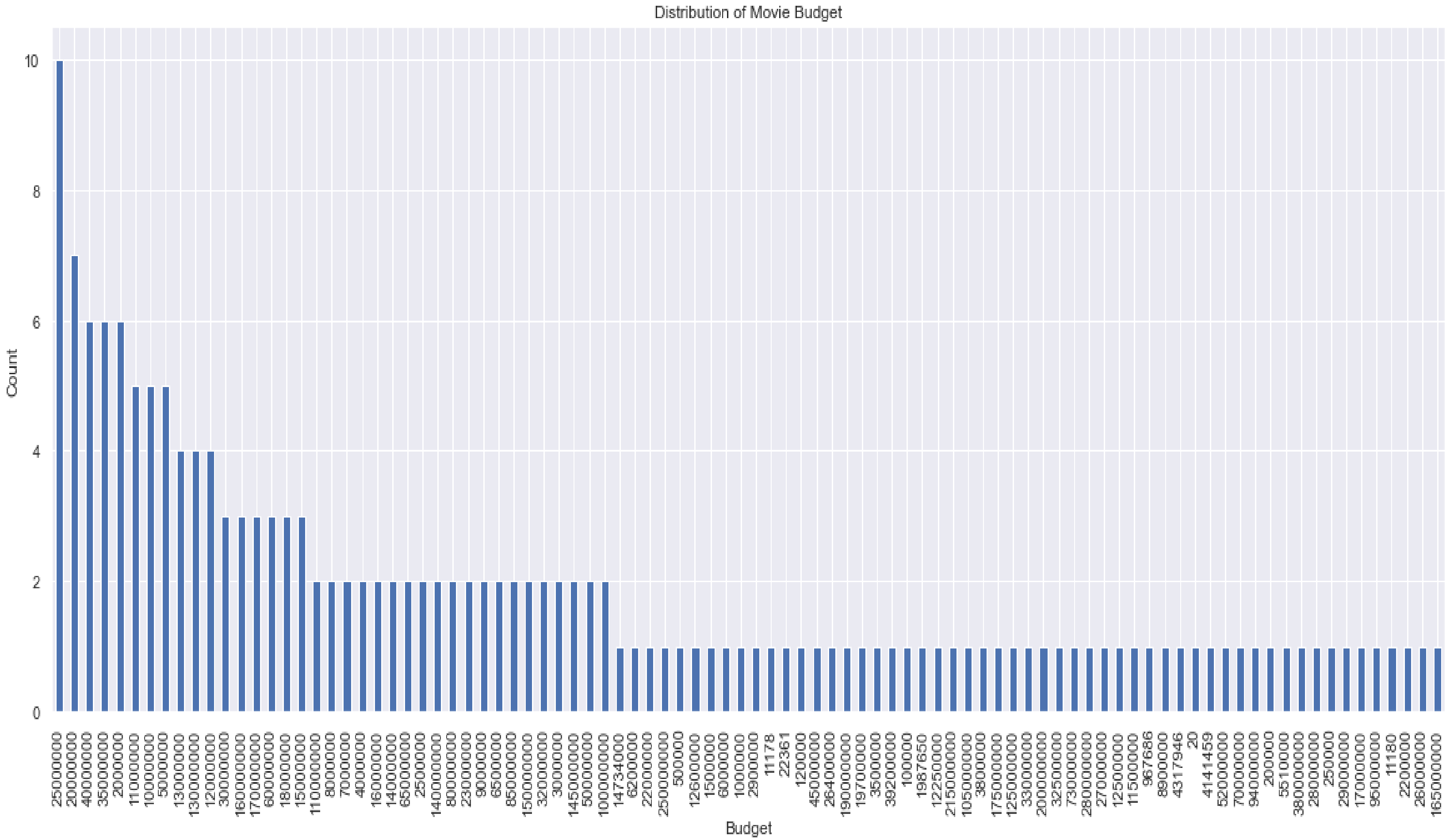
Number of films per year

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



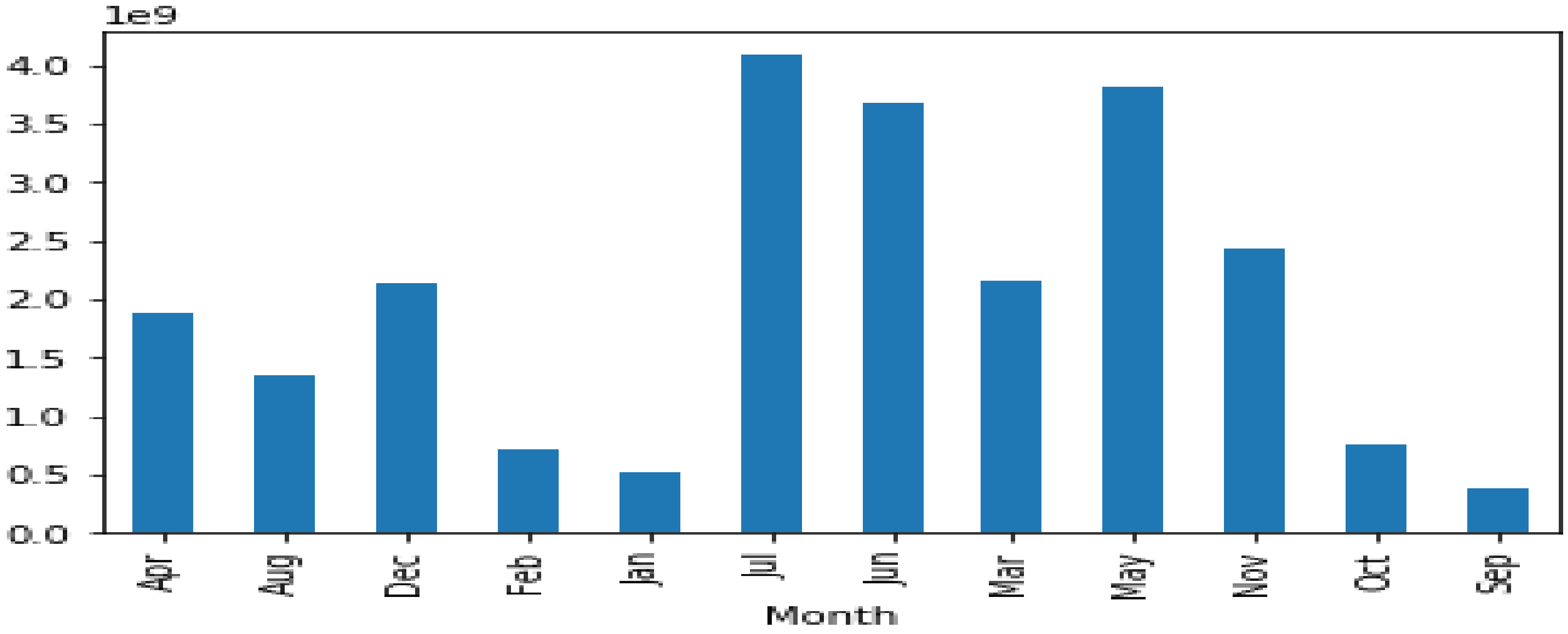
Distribution of Movie Budget

- *There was a good amount of high-budget movies at the box office*

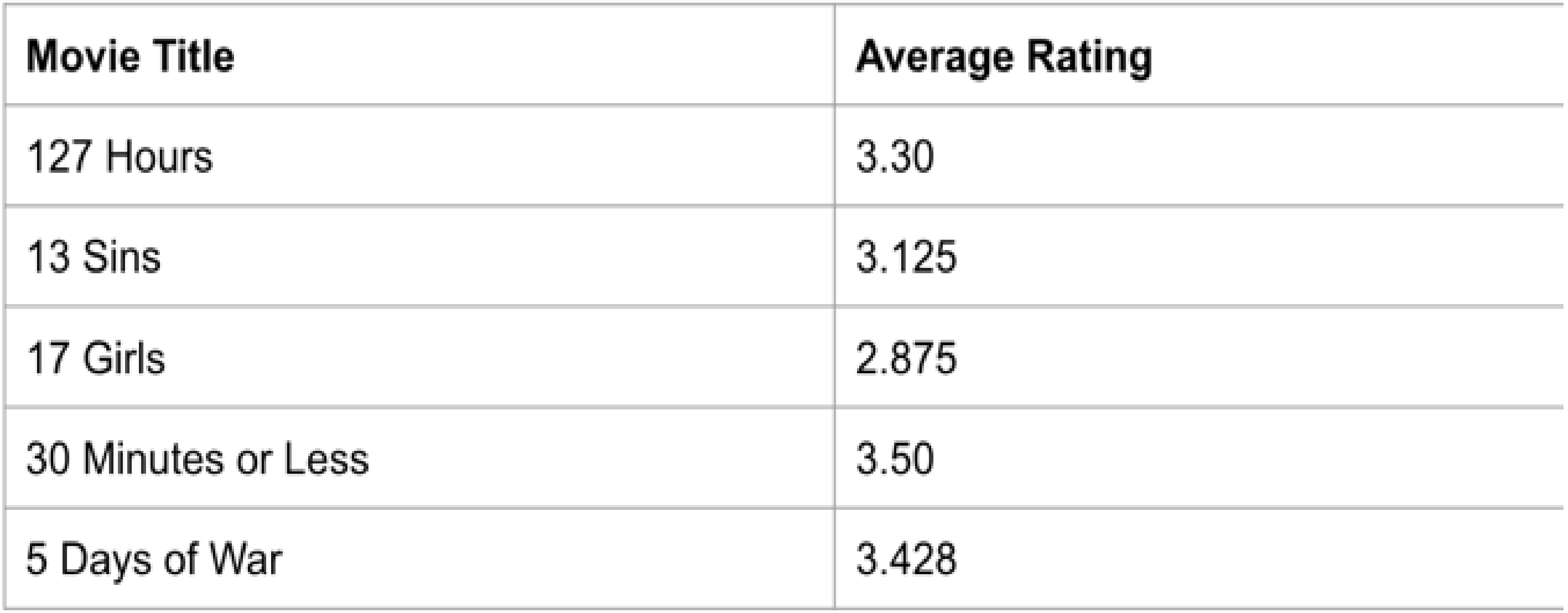


Movie released in a given month

- *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.*

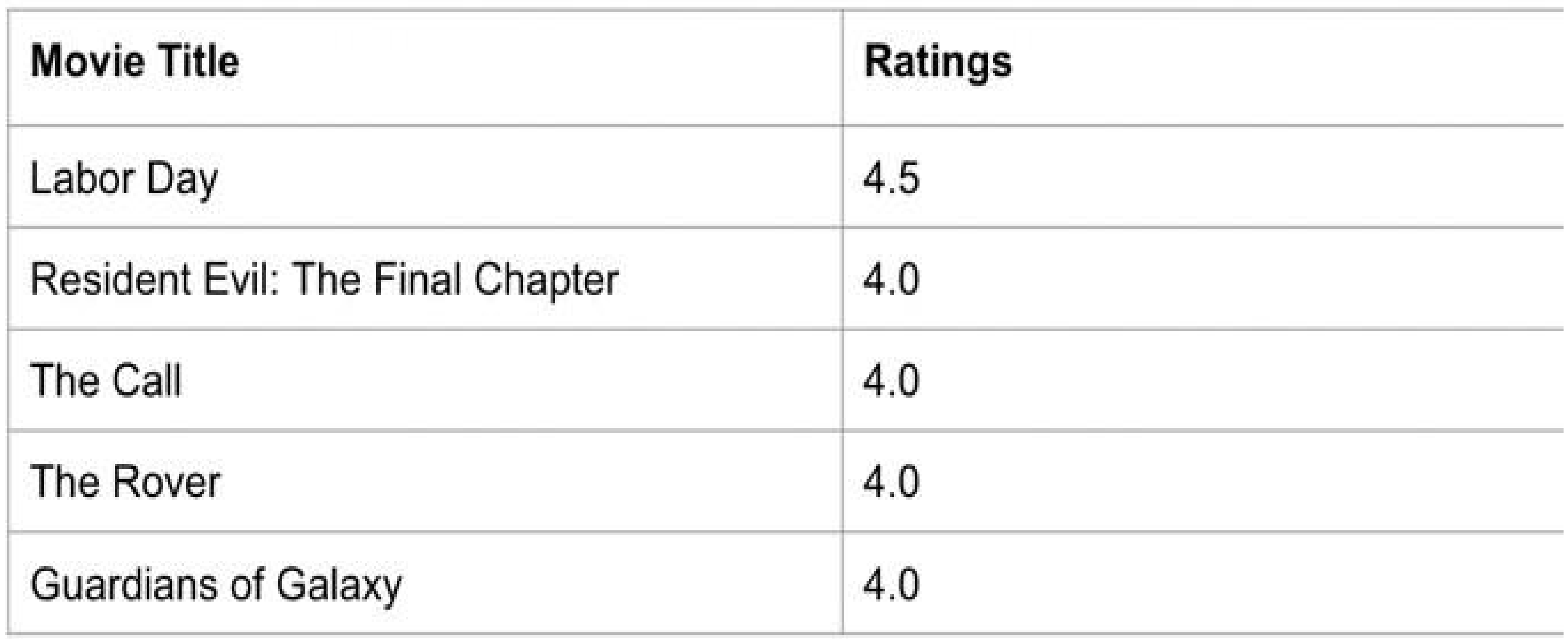


Movies with highest Average Rating



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

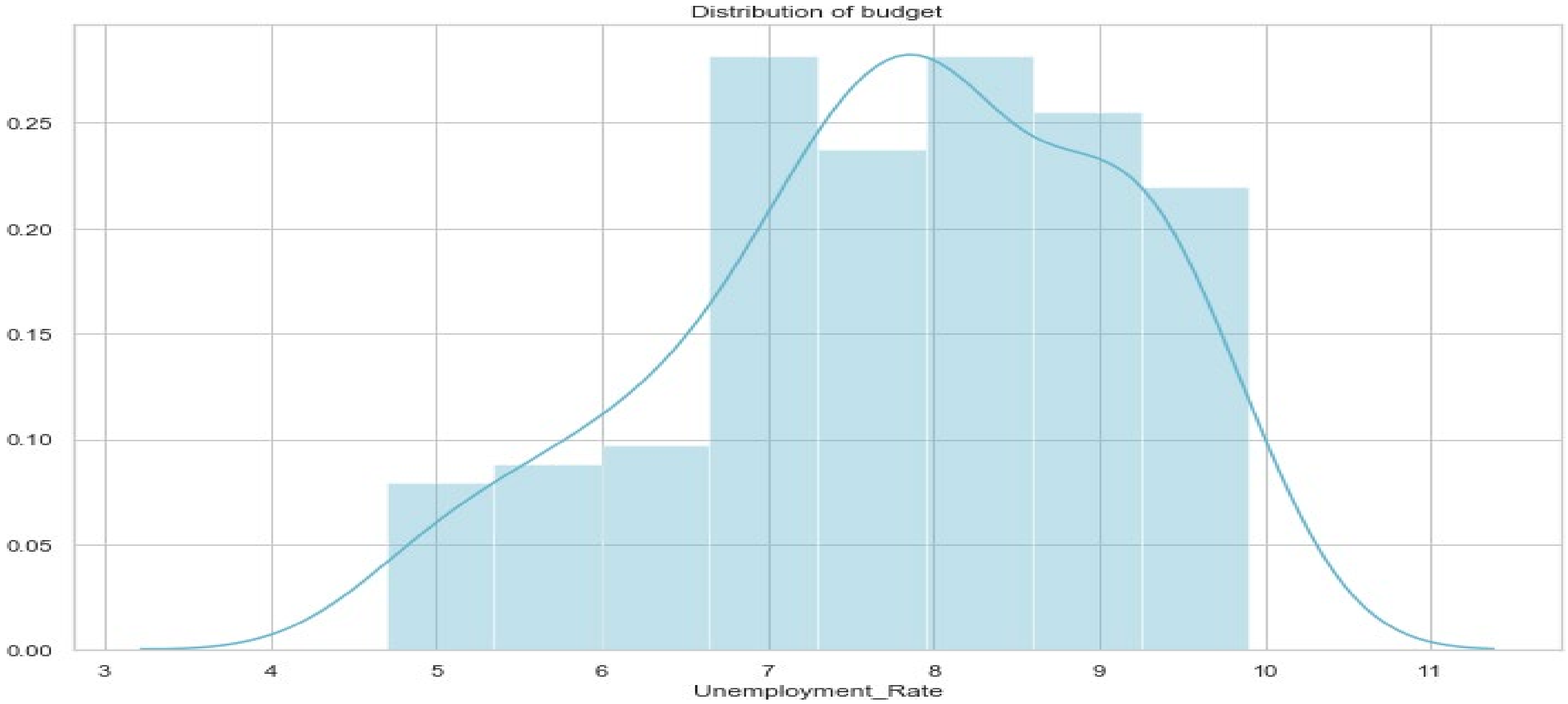
Movies with highest Ratings



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

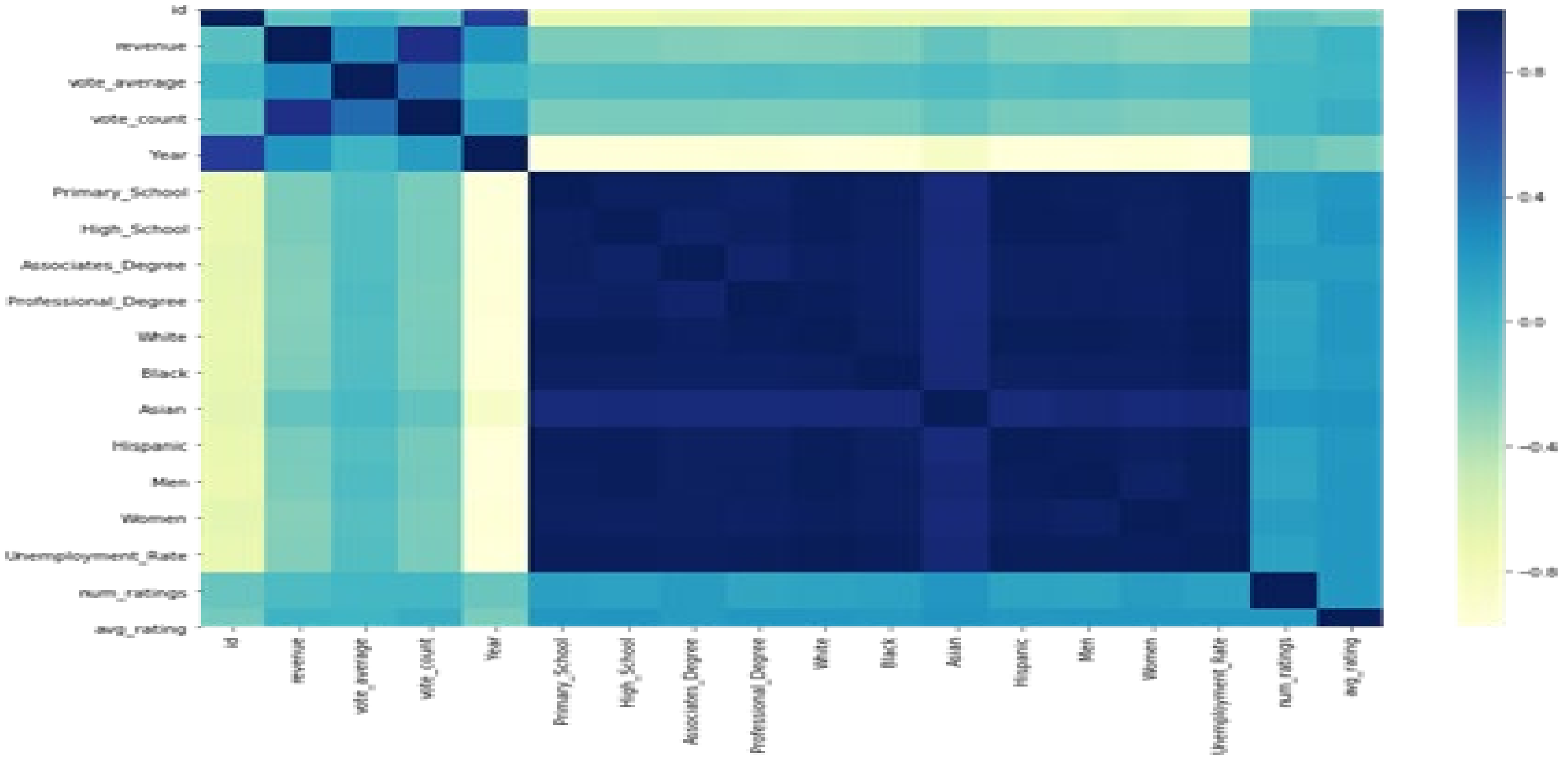
Distribution of budget

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



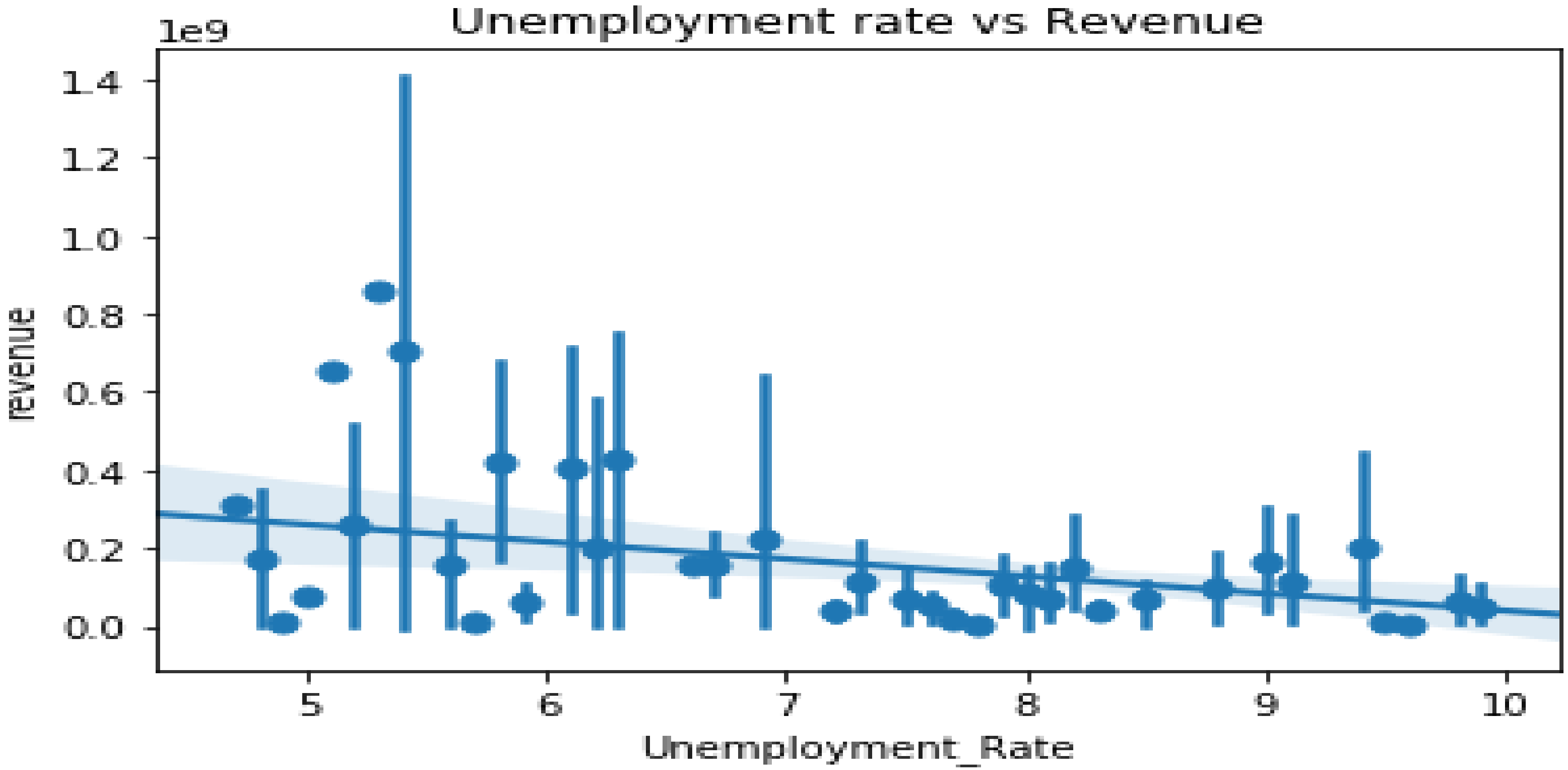
Correlation matrix

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



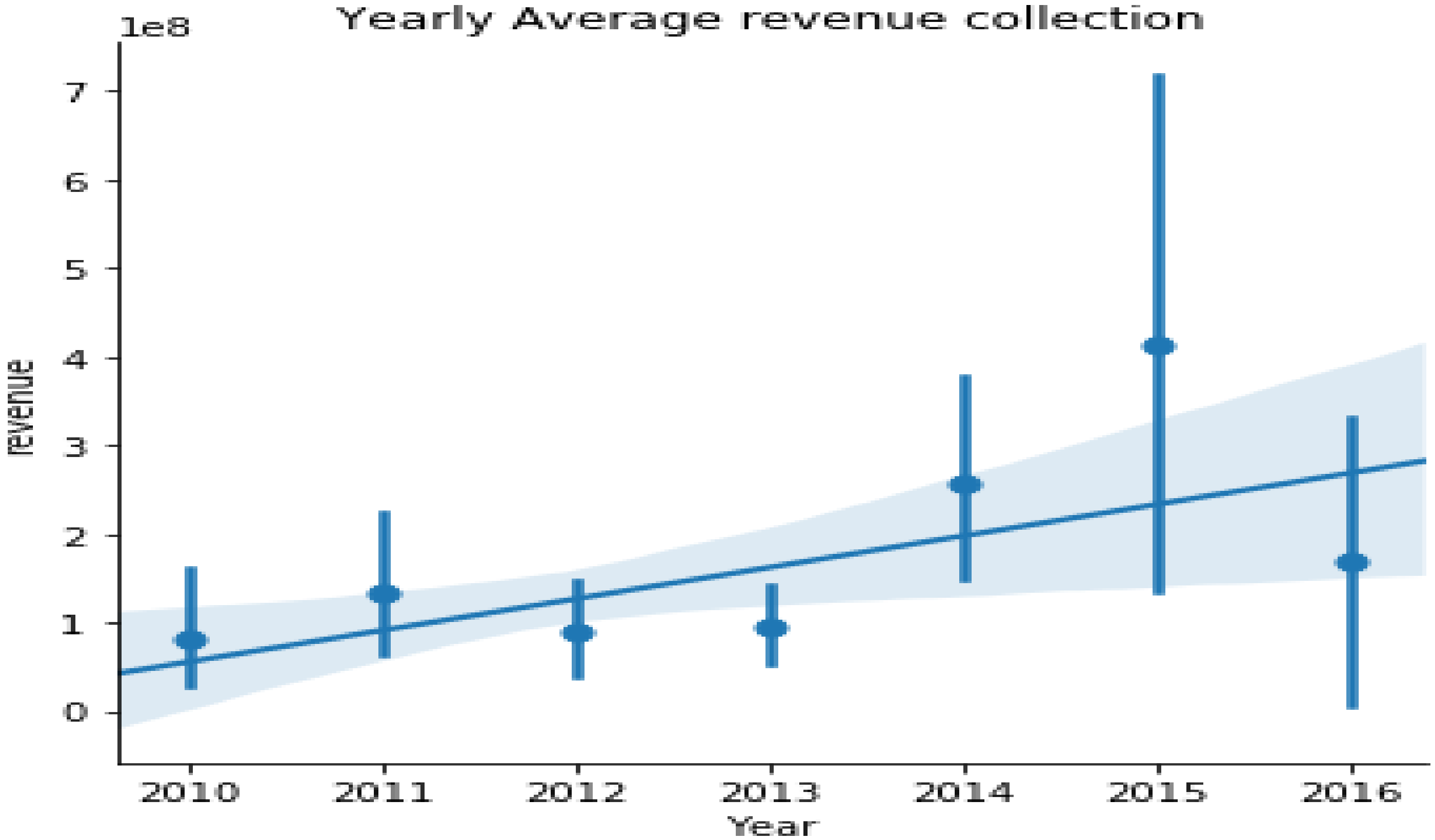
Unemployment Rate VS Revenue

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



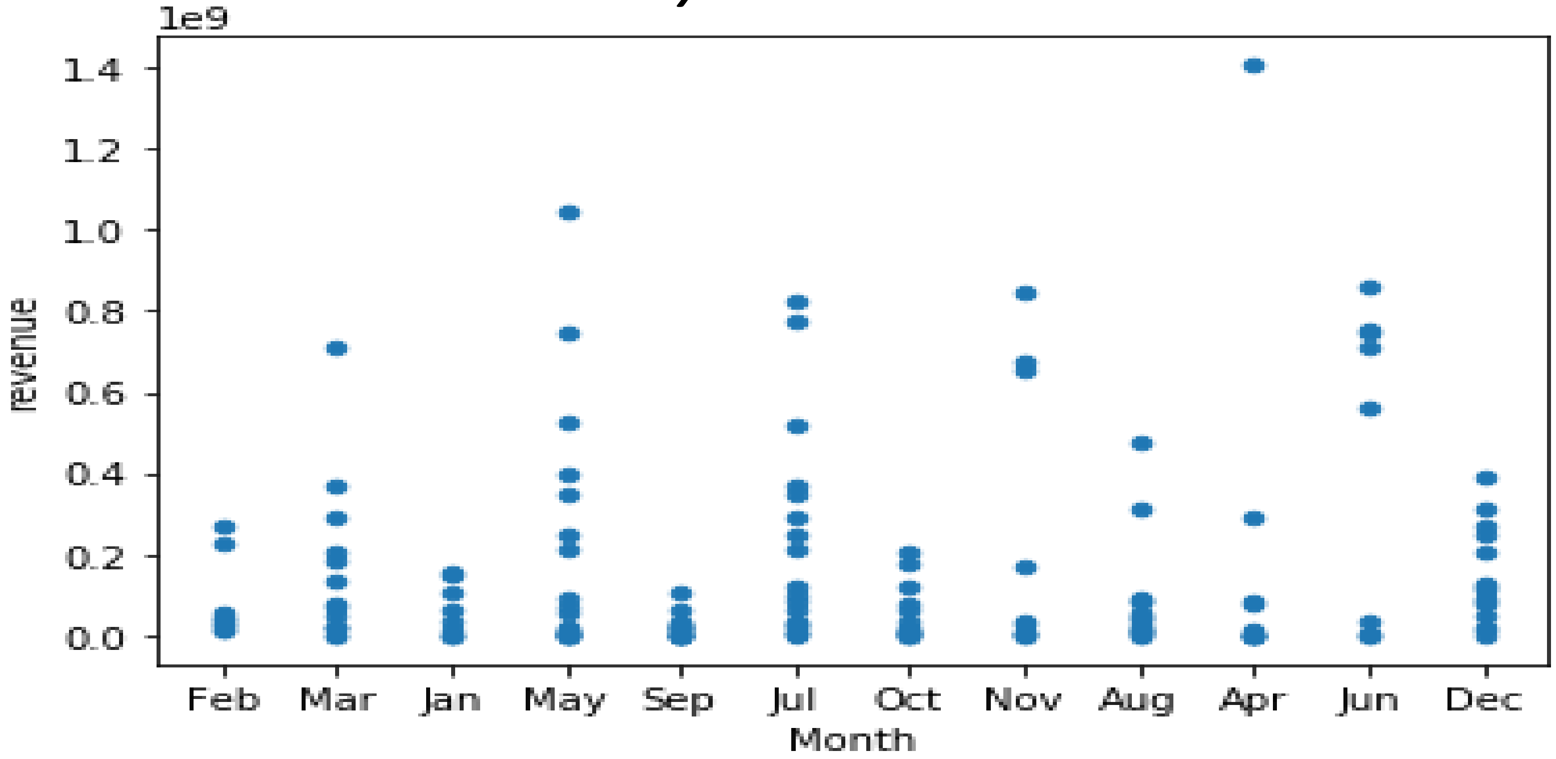
Average Revenue collection based on Year

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



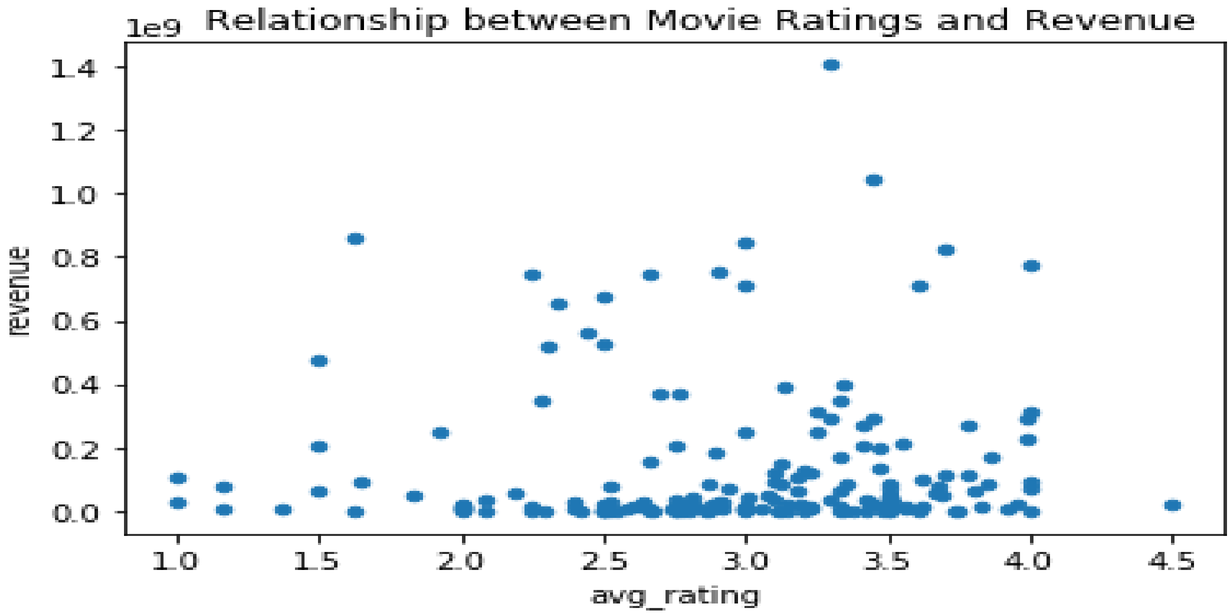
Revenue Scatter plot

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



Relationship between Movie Ratings and Revenue

- *This shows that even when the movie rating was the movie did make a good revenue on box office and few movies which had good movie reviews failed terribly at the 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.
- [2] https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
- [3] https://www.kaggle.com/aniruddhasshirahatti/us- unemployment-dataset-2010-2020/metadata.
- [4] https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset/metadata.
- [5] The United States BLS Statistics data: https://data.bls.gov/cgi-bin/surveymost?bls.