ASSIGNMENT 5

- Anirudh Narayanan(an9425)

Q1.
The data in the json file contains information regarding the currencyLabel, title, imdb_id, cost, distributor_label and box_office revenue for the imdb movies.
We can match these values with our existing information on the basis of the imdb_id. The data also consists of other information regarding the datatype for each data.

Q2.
In order to pre-process that data I read in the whole json file into a pandas dataframe, the only stored the value in the "value" field in the json data, then cleaned up the data and loaded it on mongodb. I handled the null values while pre-processing the data.

I have updated 47598 records in the movies collection using the new data provided. I have kept the revenue field for only documents where the currency is USD. This will help me filter out revenue based on currency as required.
My program updates the movie collection where the IMdb_Id matches the on the _id field in the movies collection.

```
/Users/anirudhramesh/Desktop/INTRO_TO_BIGDATA/ASSN4/venv/bin/python /Users/anirudhramesh/Desktop/INTRO_T
47590 documents updated successfully based on IMDB ID.
```

Q3.
In order to update the movies collection without the imdb id i used the primary title and the Original_title field in the movies collection and matching it with the titleLabel in the new_Data collection. Then I am performing an 'updatemany' operation where multiple documents can get updated if the 'primarytitle' or the 'original_title' gets matched. We can see this by the number of documents getting updated.
We can see that 445536 documents got updated which is a lot more than when they got matched with the imdb_id.
I tried running my update query for a while , but it was taking too long, so I added indexes to the primarytitle and original_title fields in the movies collection on mongo.

```
/Users/anirudhramesh/Desktop/INTRO_TO_BIGDATA/ASSN4/venv/bin/python /Users/anirudhramesh/Desktop/INTRO_TO_BIGDATA/AS
Time Taken: 90.38seconds
Updated 445536 documents in movies collection.
```

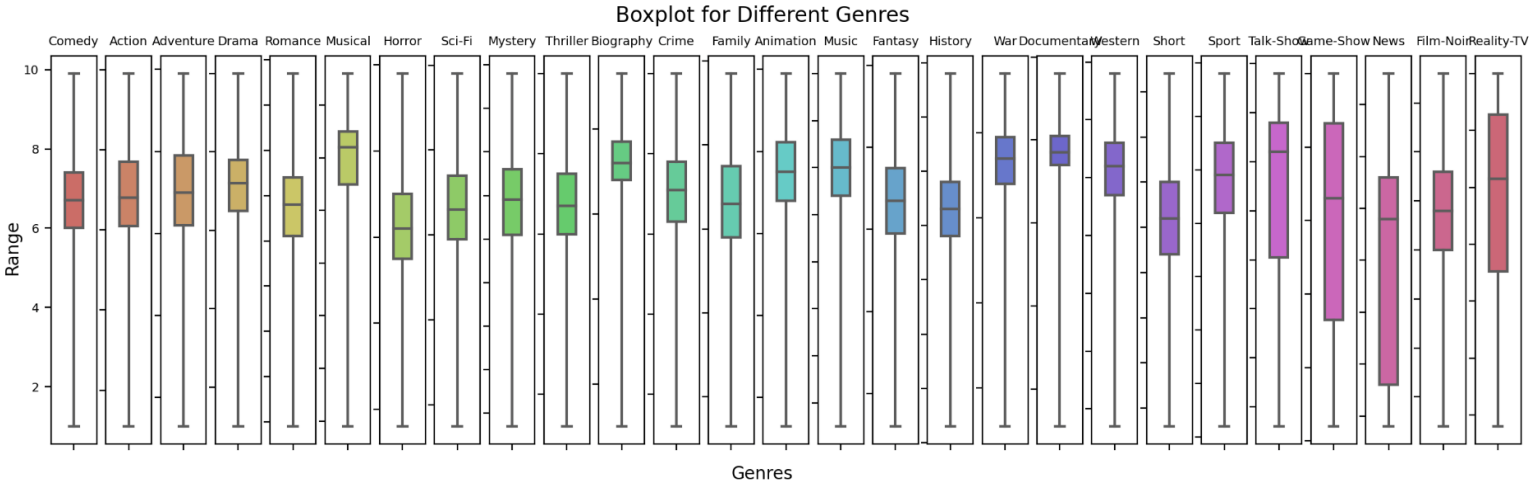
Now, I added a dictionary to keep track of the titles already updated and then ignoring the titles already modified.

```
/Users/anirudhramesh/Desktop/INTRO_TO_BIGDATA/ASSN4/venv/bin/python /Users/anirudhramesh/Desktop/INTRO_T
Time Taken: 82.02seconds
Updated 106211 documents in movies collection.
```

Now, I have updated 106211 documents, which shows that there are a lot of titles that have duplicate titleLabels and causing so many updates.

Q4

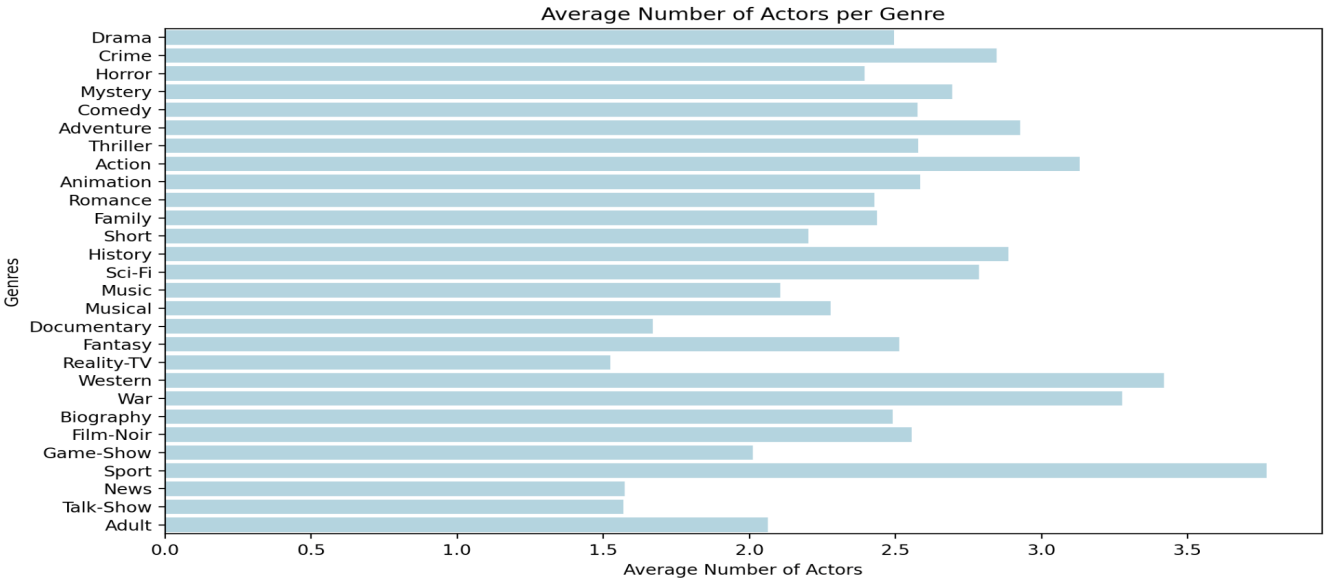
1. This is my boxplot for 4.1 where I have found the minimum value , value at 25th, 50th, 75th percentile and the maximum rating. I have read in the whole collection and populated a dictionary in python where the conditions exist and the value is met.



Time Taken:

```
/Users/anirudhramesh/Desktop/INTRO_TO_BIGDATA/ASSN4/venv/bin/python
 /Users/anirudhramesh/Desktop/INTRO_TO_BIGDATA/ASSN4/venv/q4.1.py
Time Taken: 80.37seconds
```

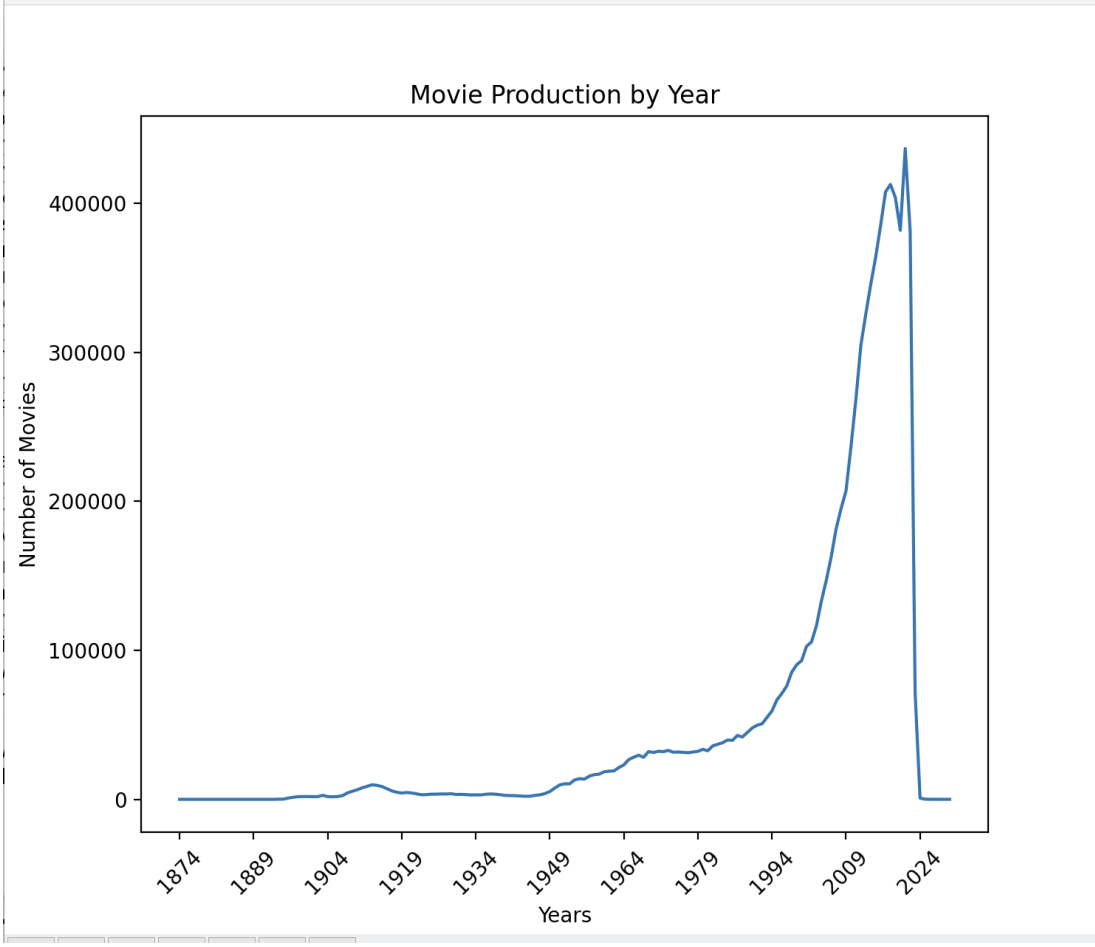
2. This is my barplot for the average number of actors in all movies grouped by genres.I read in the movies collection from the mongodb database and grouped the number of actors in each movie on the basis of genres. Then I found the average number of actors and plotted a barplot.



Time Taken:

```
/Users/anirudhramesh/Desktop/INTRO_TO_BIGDATA/ASSN4/venv/bin/
Time Taken: 85.50seconds
```

3. This is my time series plot for the number of movies released by year. I read in the movies collection from the mongodb database and grouped the number of movies for every staryear. Then I plotted a time series plot using after sorting the data based on movie start year.



Time Taken:

```
/Users/anirudhramesh/Desktop/INTRO_TO_BIGDATA/ASSN4/venv/bin/python /Users/anir
Time Taken for Plot: 98.26seconds
```