DECI Project: Investigate a Dataset

Dataset chosen: **tmdb-movies.csv**

References: N/A

Data Wrangling Process:

Problems that needed cleaning were as follows:

1. NaN Values
    - Cleaned data as following:

```
- imdb_id: Drop Rows with NaN values.
- cast: Drop Rows with NaN values.
- homepage: Replace NaN values with "None".
- director: Drop Rows with NaN values.
- tagline: Replace NaN values with "None".
- keywords: Replace NaN values with "None".
- overview: Replace NaN values with "None".
- genres: Drop Rows with NaN values.
- production_companies: Replace NaN values with "Not Mentioned".
```

2. String Split Absence
    - Used ‘.str.split()’ function

3. Useless Data
    - Removed ‘imdb_id’, ‘id’ columns

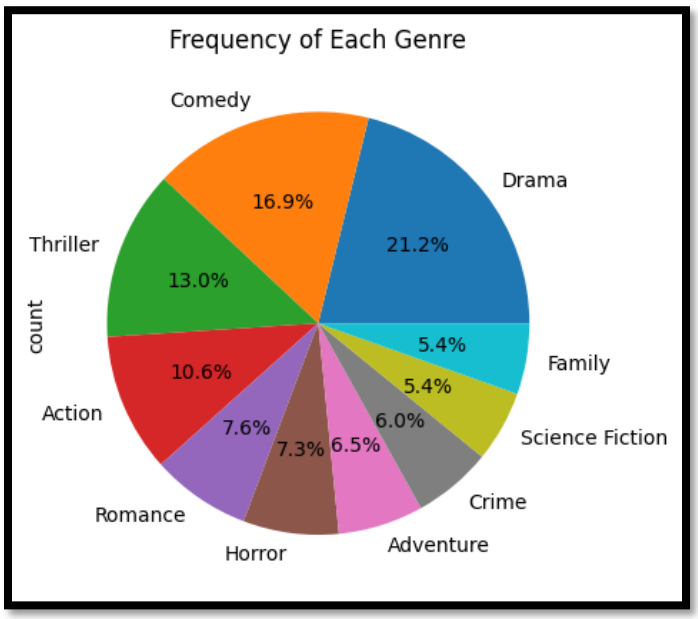
4. Unrealistic Runtime
    - Removed data with runtime > 300 using ‘.drop()’ function

Exploratory Data Analysis Process:

Posed Questions:

1. What is the most used genre?

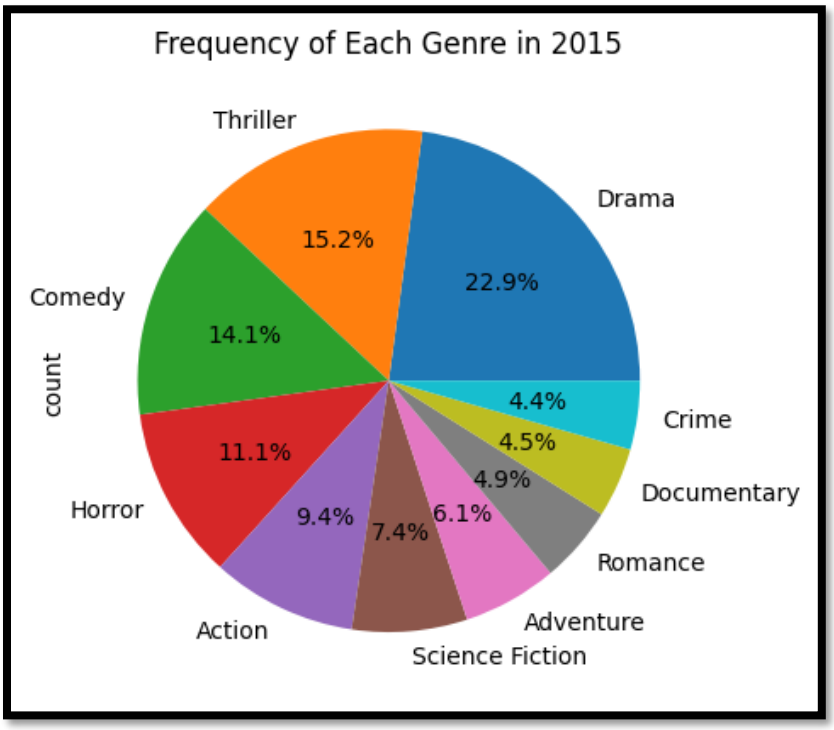
    - Dissected genres and formed new series
    - Used ‘value_counts()’ function to make a pie chart representing percentage as following:



❖ **Conclusion: Most used genre is *Drama* with 4754 movies.**

- Bonus Question: What is the most trending genre in 2015?

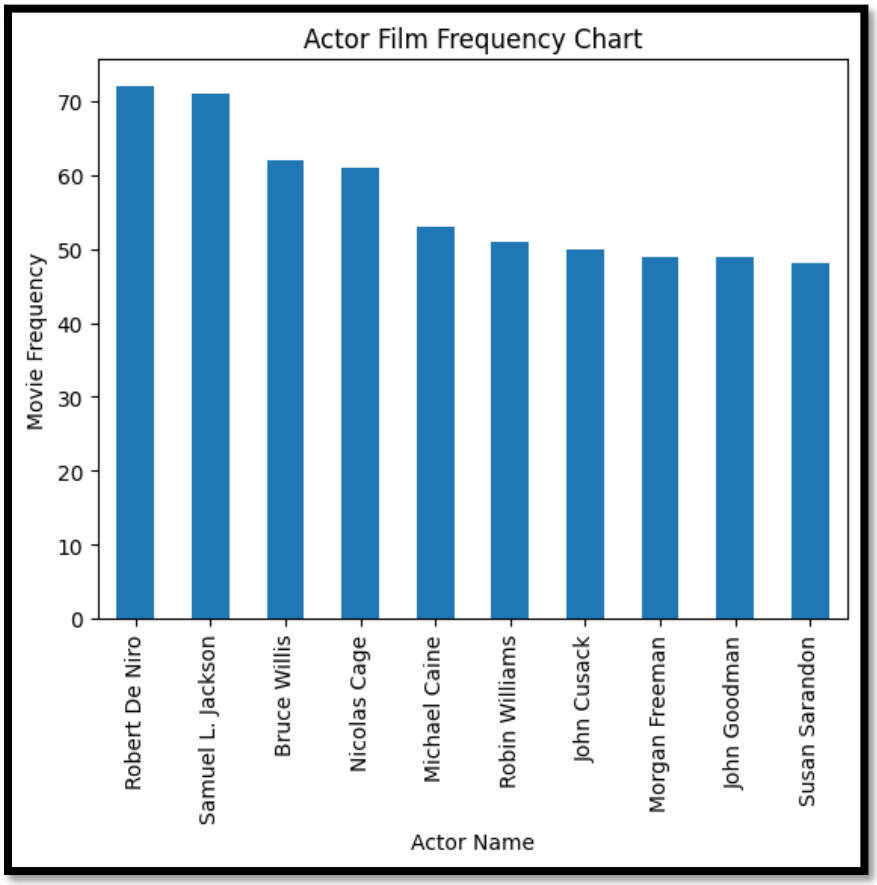
    - Sorting data frame by year and using '.head()' function to only include movies in 2015
    - Used 'value_counts()' again to represent data in pie chart as following:



❖ **Conclusion: Most used genre in 2015 is once again *Drama!***

2. Which actor played in the most movies?

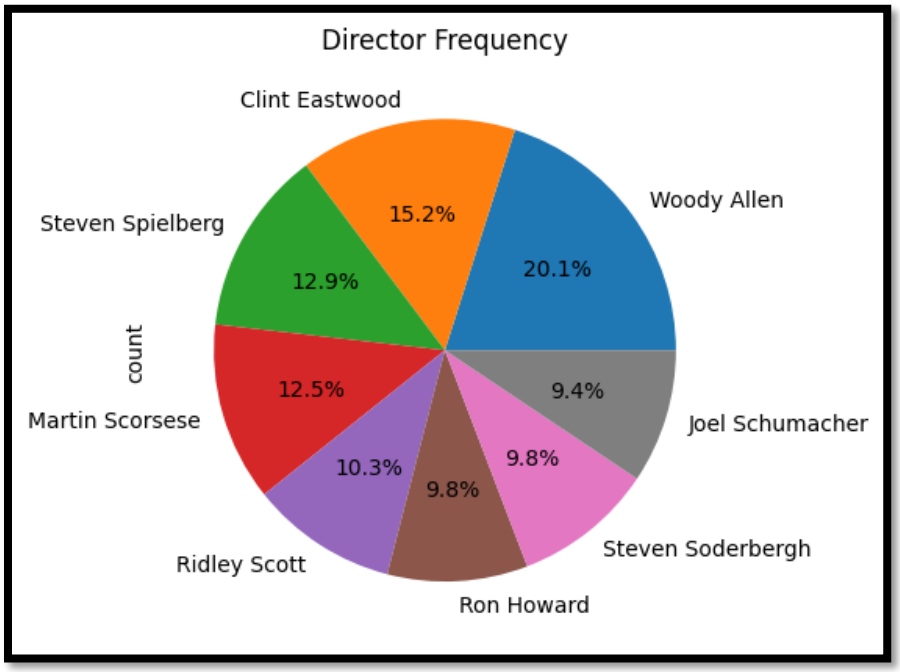
- Dissecting cast in data frame using '.explode()' function
- Using value count to represent the data in bar chart as following:



❖ Conclusion: **Robret De Niro played in the most films, starring in 72 movies!**

- Bonus Question: Which director made the most movies?

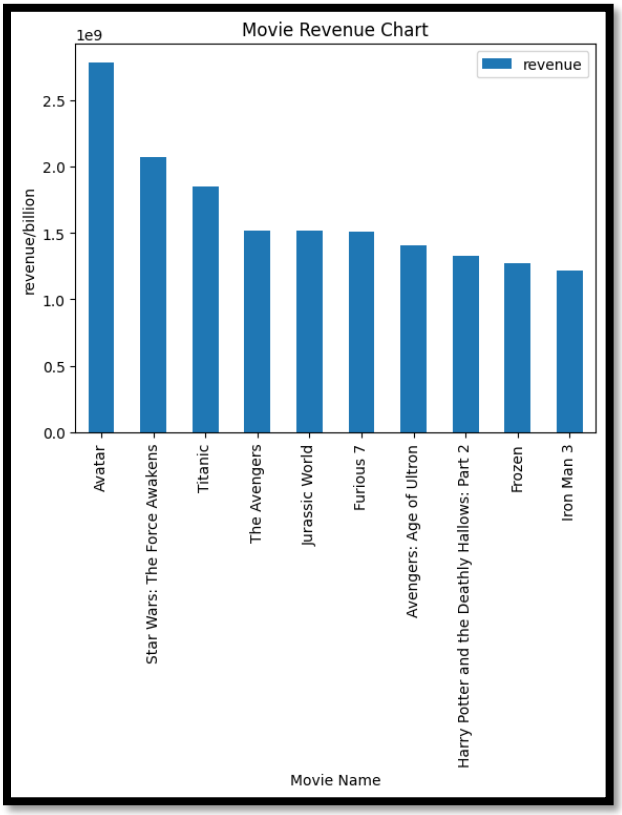
    - Using ‘value_count()’ function graph is represented in pie chart showing the top 8 directors by percentage as following:

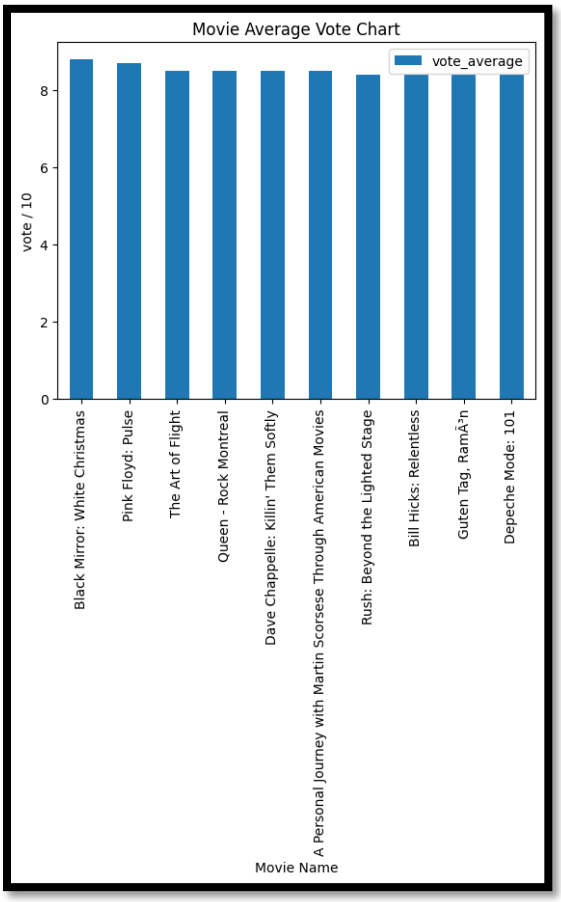


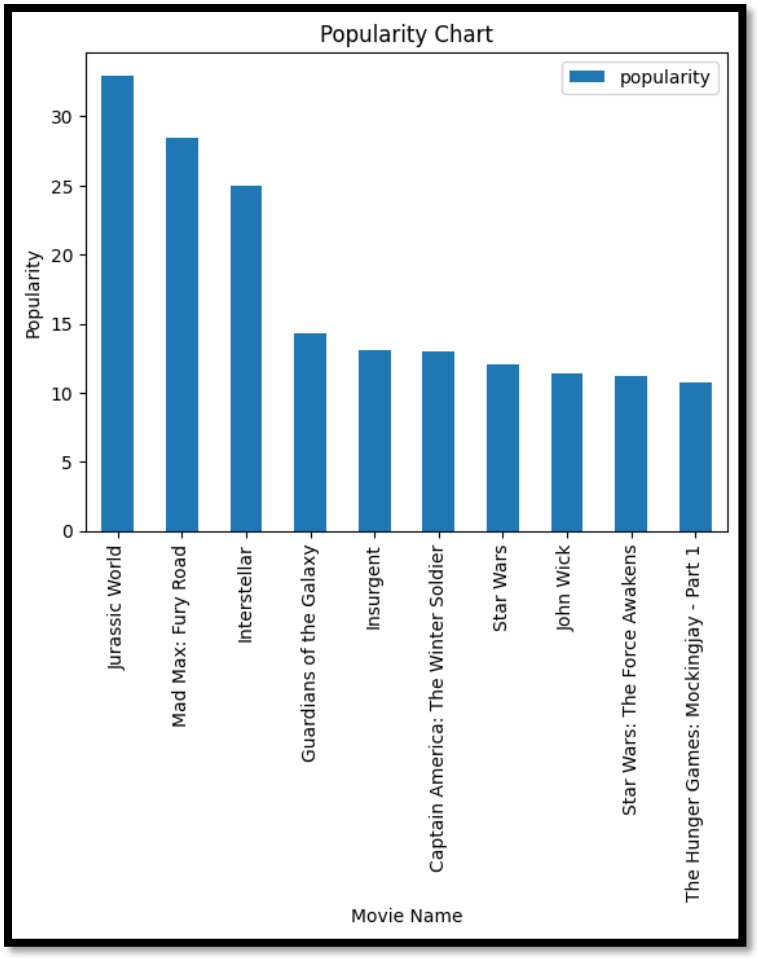
❖ **Conclusion: Woody Allen made the most movies (45 movies)**

3. What is the most successful film?

- Conclusion was based on Three aspects:
  1. Revenue
  2. Average Vote
  3. Popularity
- Using manually-created function sorting() each column was sorted in descending order in a copy of the data frame
- Copies of data frames were used to make following graphs:

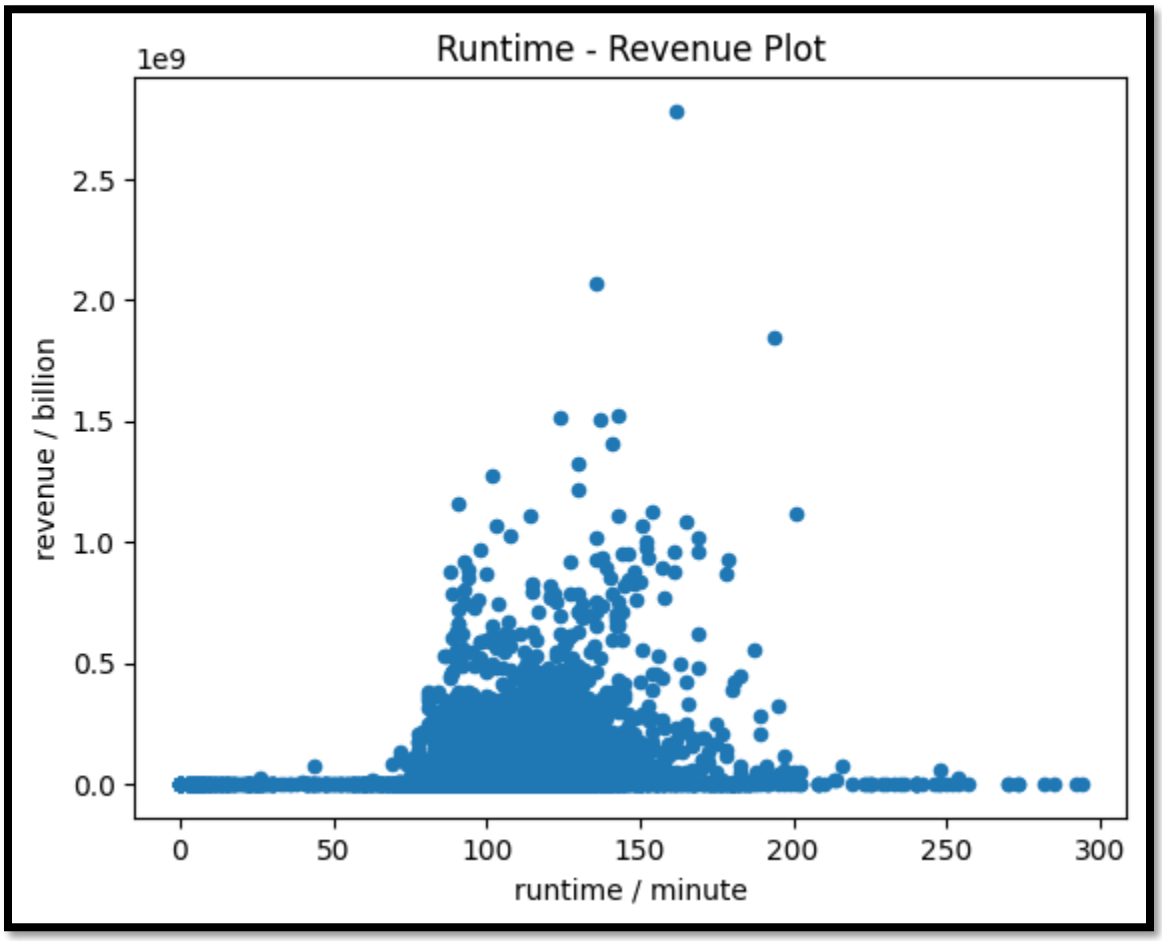






- **Conclusion: Jurassic World takes the cake for the most successful film, while it does lack in the average vote with a number of 6.5/10, it makes up for it in popularity where it ranks #1 and in revenue where it ranks #5, these statistics do certainly make it a promising contender!**

4. Is there any correlation between movie-length and revenue?
   - ❖ Using the '.plot.scatter()' function a plot was represented by the relation between revenue and runtime as following:



   - ❖ **Conclusion: The movies with the highest revenues are generally situated between 70-80 minutes long**