Report

The Movie Database (TMDB)

The Movie Database (TMDB)

A PDF file containing my analysis and final results.

1- A note specifying which dataset you analyzed:

I have analyzed The Movie Database (TMDB).

2- A statement of the question(s) you posed:

a- Which genres are most popular from year to year?
b- What kinds of properties are associated with movies that have high revenues?

3- A description of what you did to investigate those questions:

1- I downloaded the data.

2- Then I explored the data and knew the dimensions of my dataframe and the number of missing values and the number of duplicate rows.

3- I did data wrangling in all its stages such as data cleaning to delete the above-mentioned problems and find solutions to them.

4- Then I started using the split () function in order to get rid of the special character ' | ' in the rows (cast , director , genres).

5- Then I answered the first question by drawing a histogram for each of the years from 1960 to 2015, where this figure shows the most prevalent types of films from year to year.

6- Then I started answering the second question and I chose the following classes (runtime, budget, cast, director, genres release_date, production campaigns) to clarify whether these classes directly or indirectly affect the revenues of the film.

7- I started using different forms and different statistical methods to estimate the amount of impact.

8- Finally, I got amazing results. I will review them later and will explain them.

4- Documentation of any data wrangling you did:

1- Gathering the data:

```
1- df = pd.read_csv("tmdb-movies.csv")
```

2- Discovery:

```
1- df.shape
2- df.duplicated().sum()
3- df.info()
4- df.isnull().sum()
```

3- Data Cleaning:

```
1- df.drop_duplicates(inplace= True)
2- df.drop([..................})
3- df.dropna(subset=[..............],inplace=True)
4- df.fillna(method ='pad',inplace=True)
5- x.split(str_split)
```

4- Validating:

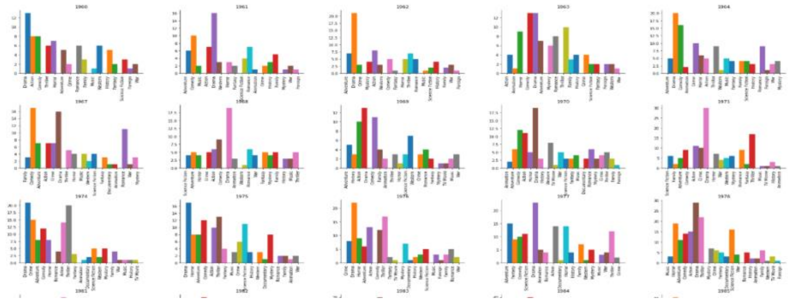
verifying that my data is both consistent and of a high enough quality by printing it.

5- Summary statistics and plots communicating your final results:

1- Question 1: What are the most popular types from year to year?

Tape charts show how the most watched and most popular movie genres changed from year to year from 1960 to 2015.

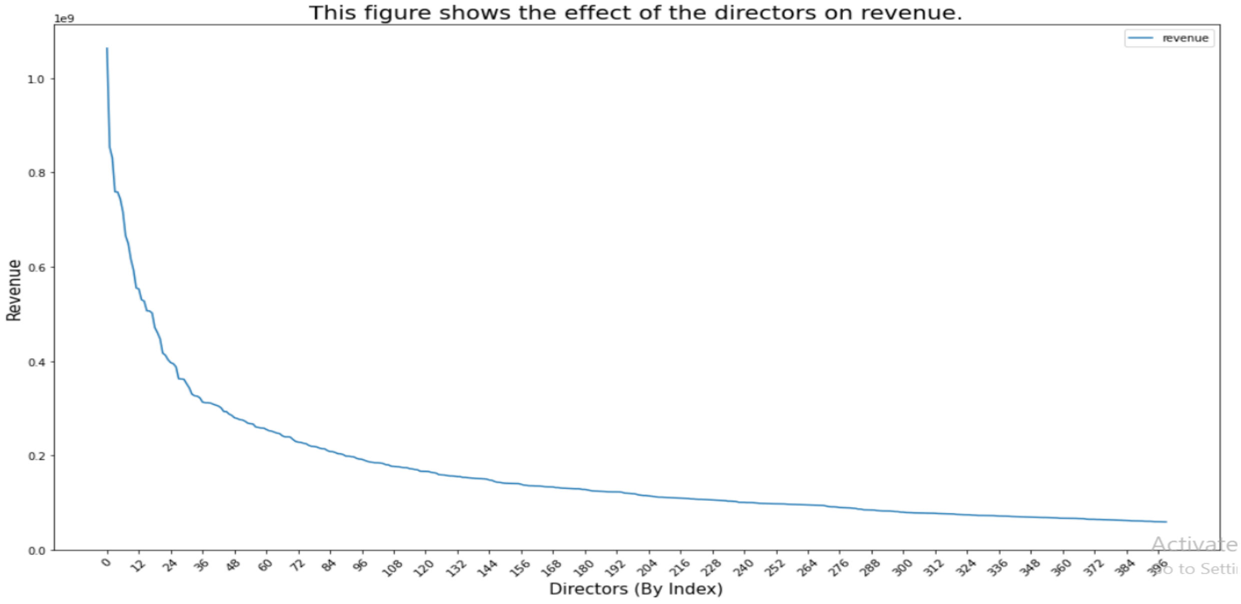
We note that no film has been reformed during these years. Instead, all types of movies change proportionally from year to year, as a result of several factors that we will mention later, such as the cast, the director, the running time…etc.



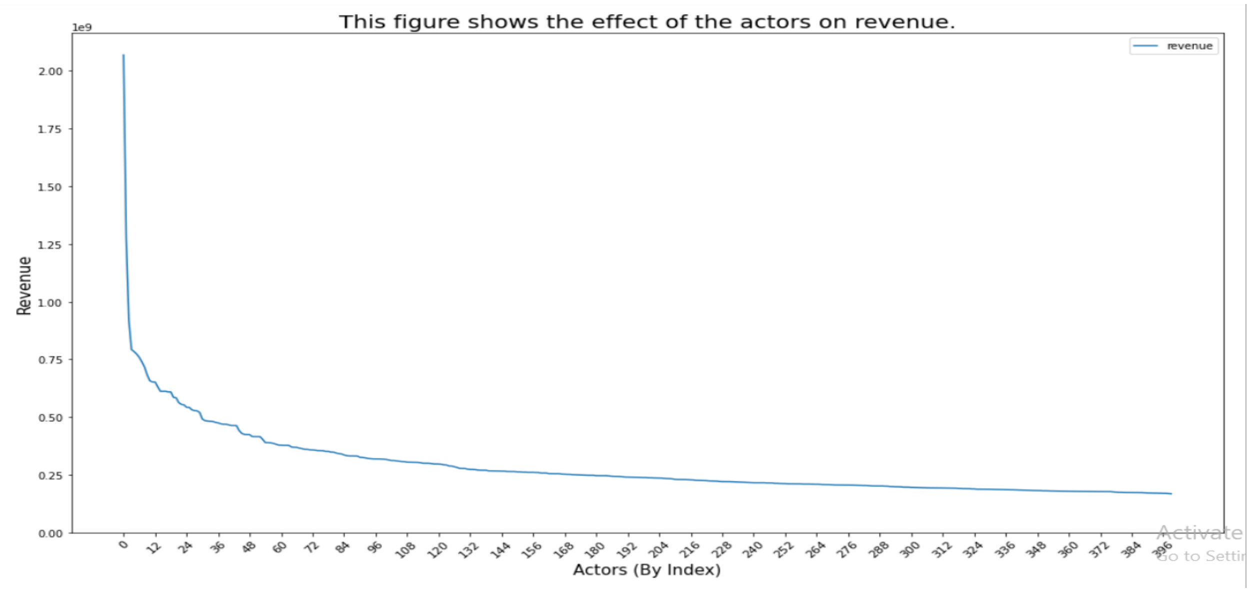
2- The second question: What are the types of real estate associated with films with high returns?

I chose 7 pillars that I think might affect the movie's revenue:

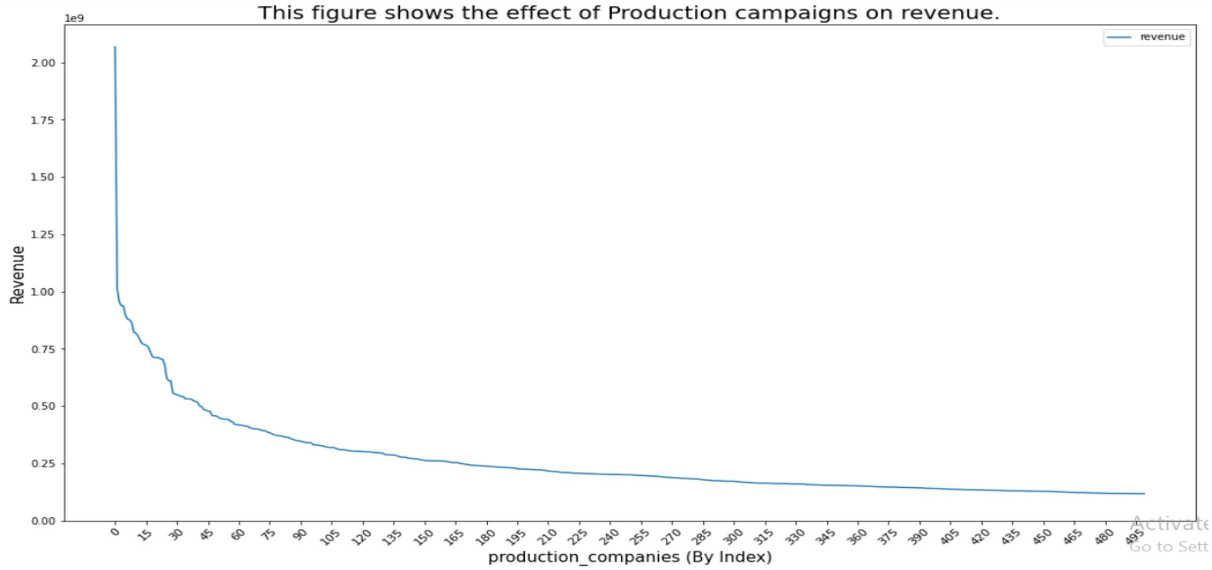
A - The first figure is the line chart, which represents how the directors of films affect the revenues of the film, I examined only 400 samples, but if you want to be higher or lower, just add the number inside the code. And if you want to achieve high revenues from the film, you should choose about the first 120 directors from my list, but if you choose a director after the 120th order, the curve of the level of revenues will go down, and therefore it is possible not to achieve high revenues.


This figure shows the effect of the directors on revenue.

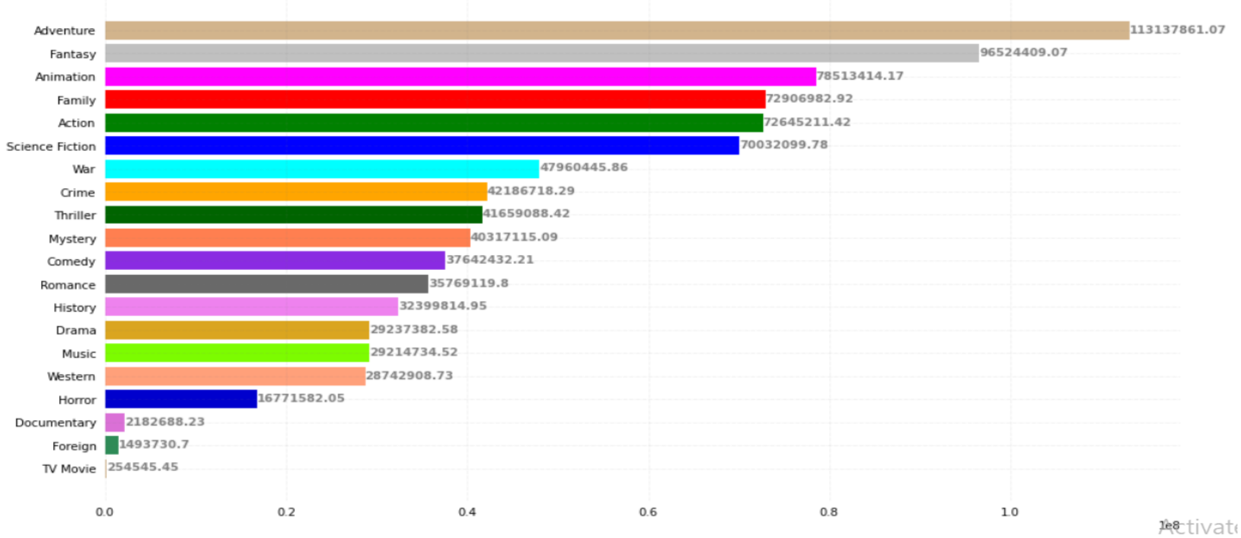
b - In the second line chart, which shows the effect of casting on revenues, as is the case with the first line chart, but here if you want to achieve good revenues, you should choose the casting team to be most of the first 200 actors in my list, they will guarantee you a large percentage of achievement High revenue.


This figure shows the effect of the actors on revenue.

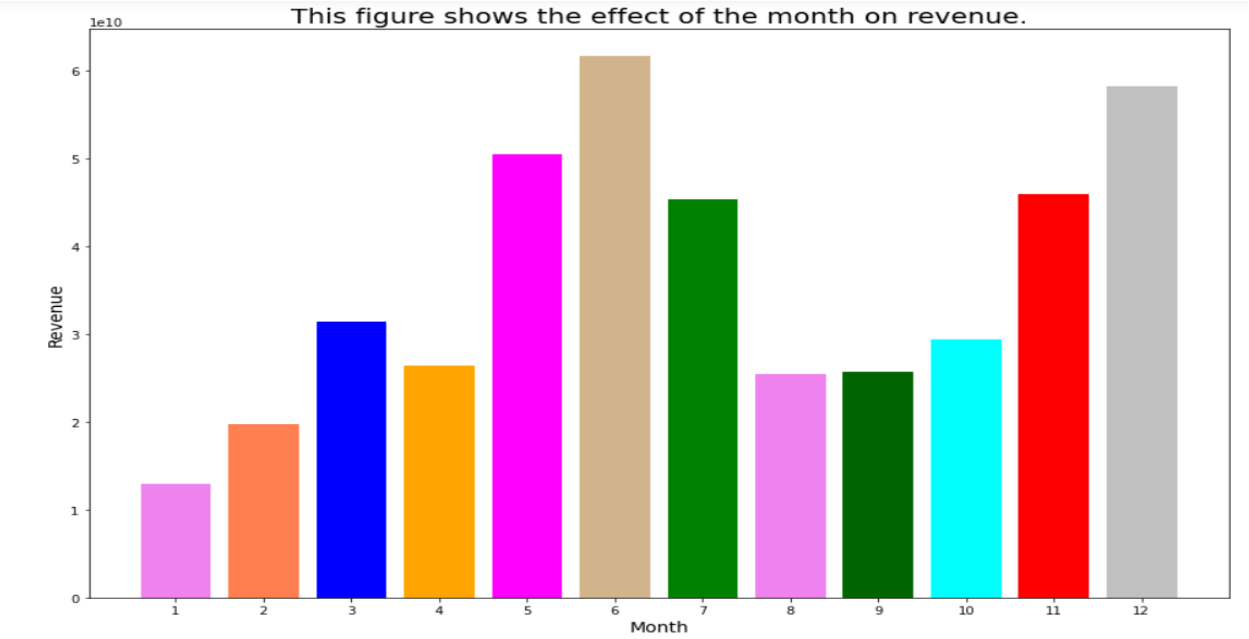
C- The same thing is what happens with the third line chart, which shows the effect of production companies on the revenues of films, and we also note that there is a great influence in choosing the appropriate production companies to produce your films and achieve high revenues, and also to ensure high revenues you have to choose production companies Only from the first 180 companies from my ranked list, and if companies are selected after the 180th rank, then you will risk the success of the film and you will risk achieving high revenues or not.


This figure shows the effect of Production campaigns on revenue.

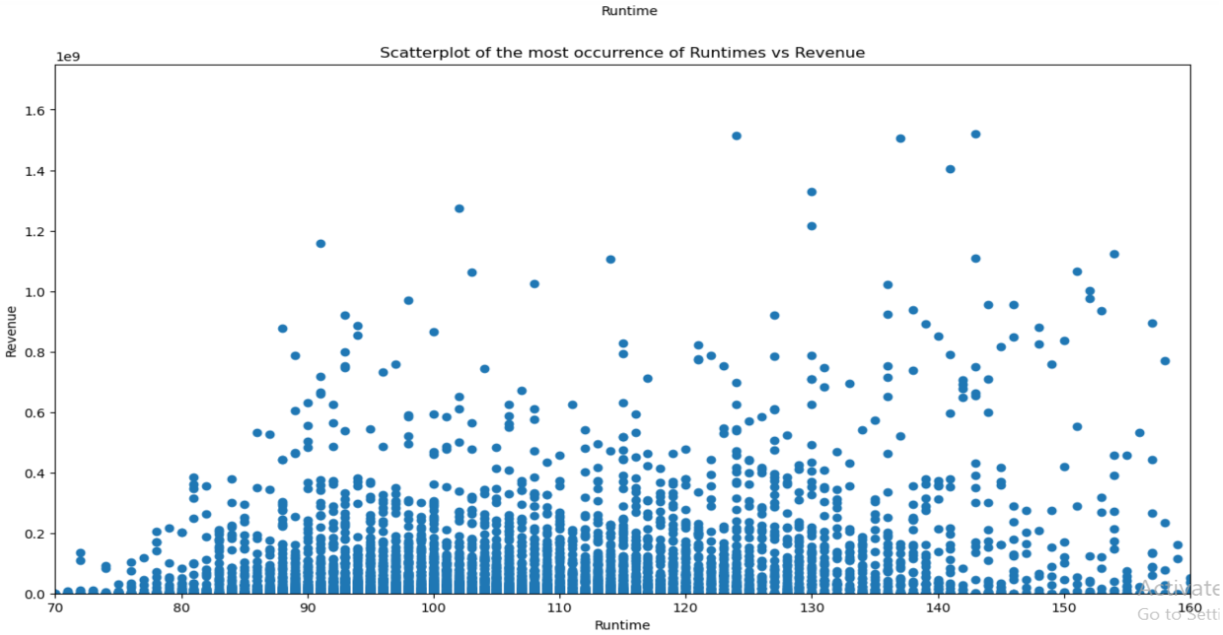
d- The fourth figure represents the horizontal bar chart, which represents the most widespread types of films that achieve the highest revenues. As we can see, the Adventures films top the list and TV Movies remove the list, and according to this figure, it is preferable that the type of film you produce be one of the following types (Adventures, fantasy, animation, family, action, Sience fiction) because the seventh film in the ranking, which is the war movies, is far from the sixth type in terms of revenue generation, and it is also clear from the graph that the genre of the film directly affects the volume of sales and revenue because of the preference for People are more interested in certain types of films than others.



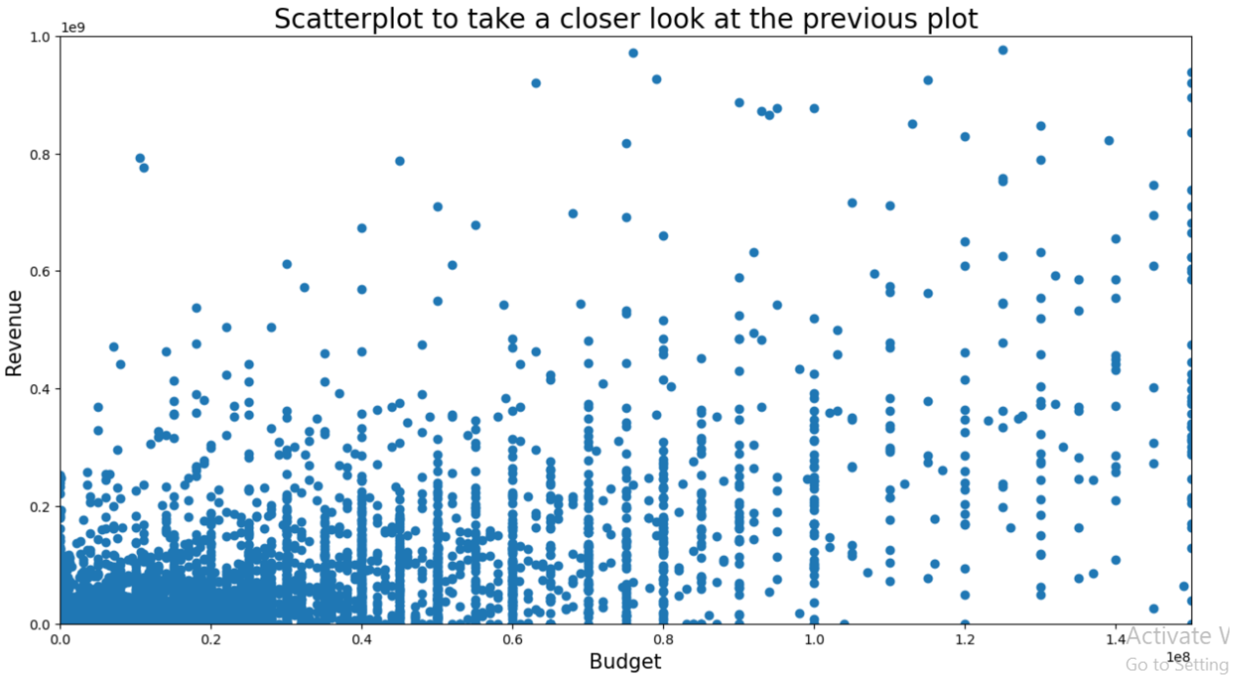
e- The fifth figure represents a bar chart showing the effect of the month in which the movie is released on the amount of revenue that the movie earns, As we can see here that the highest two months in terms of revenue are the months of June and December, and this certainly has many reasons. It is possible that these two months are the months of official holidays in many countries, or that these two months have important occasions or the date of the beginning of the spring and autumn seasons, and therefore the weather is mild. Which encourages people to get out of their homes and go to the cinema.



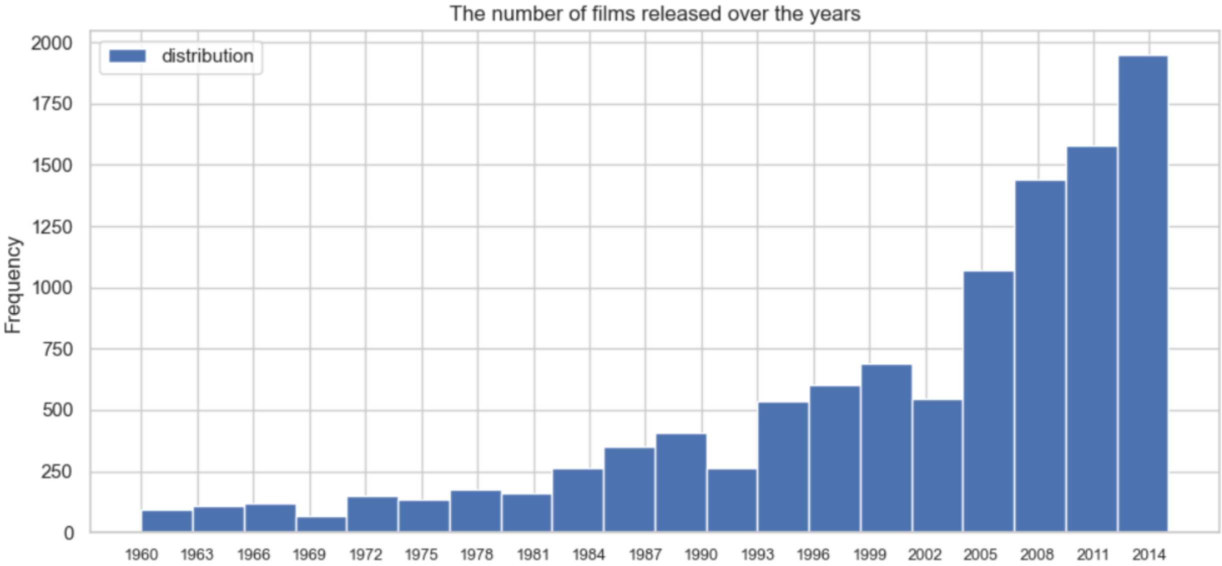
f- In the sixth figure, a scatterplot was drawn, and it shows the appropriate film duration that people prefer, which positively affects the revenues, It is preferable that the duration of the film ranges from 90 to 140 minutes, because this period is the one that brings more revenue and therefore people prefer it.



g- In the seventh figure, which also represents a scatter plot to show the effect of the budget on the revenues of the movie, we find that the correlation between them is very little and that the budget does not strongly affect the revenues.



h- It is also clear from this figure that the rate of films released during the year is increasing annually from year to year, and we see that the rate is incremental, and this indicates progress in the artistic field and people's interest in films over the years, and thus the revenues generated by films increase from year to year.



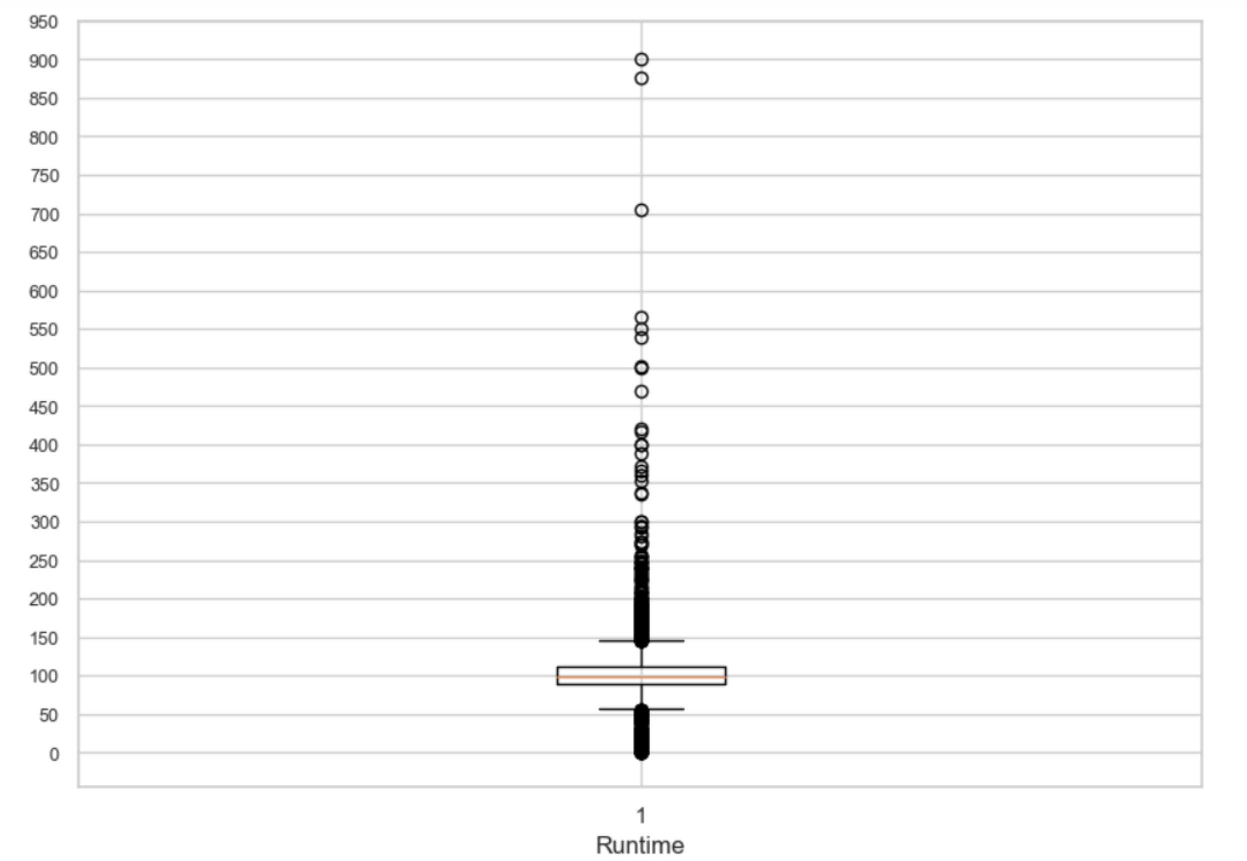
i- The box-plot gives us an overall idea of how spread the distribution is in the case of the runtime of the movies. It also shows the outliers here.

By looking at both of the plots and calculations, we can conclude that:

25% of the movies have a runtime of fewer than 90 minutes

50% (median) of movies have a runtime of fewer than 99 minutes.

75% of movies have a runtime of fewer than 112 minutes.


Runtime