This reports briefly explains and describes my wrangling effort and findings in this project

This dataset is about movies specifications and it consists of 10866 row (which means that we are going to analyze the data of 10866 movie) and 21 column and such as the popularity and averge rating for each movie and it's title and release date and other attributes

our aim is to analyze this data and get answers to some questions in the dataset and clean and asses the dataset and make visualizations that will help us in answering those questions

The wrangling process done is divided into 3 three steps:
1-Gathering data
2- Assessing data
3-Cleaning data

Each step will be briefly explained below.
First: Gathering data

The dataset was downloaded from Kaggle and its shape is 10866 row and 21 column which are :

- homepage
- id
- original_title
- overview
- popularity
- production_companies
- production_countries
- release_date
- spoken_languages
- status
- tagline
- vote_average

Second:Assessing data

After the gathering process of the data the data need to be assessed and assessing is done visually and programmatically for the quality and tidiness issues

- we need to remove some columns such as id and imdb_id

- There are null values in the dataset

- There are duplicated rows

- we need to make some changes to release_date column so that we can use the date for the analysis

- there are incorrect runtimes

- there are movies have zero budgets and revenues in the dataset

- there are a correlation between some columns such as popularity and vote_count

Third: Cleaning

The last step in wrangling the data in data cleaning and in this step we have to solve every issue we found in the assessing process to result a high quality and tidy data

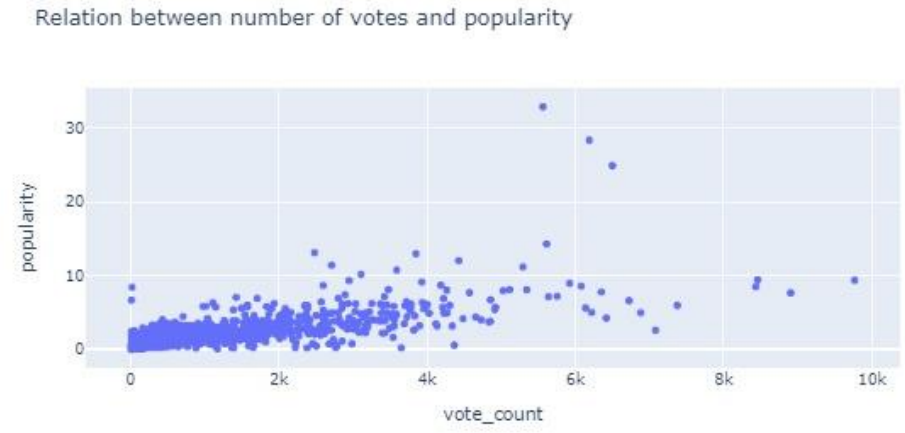
Now in this part of the report we will communicate our findings from this dataset

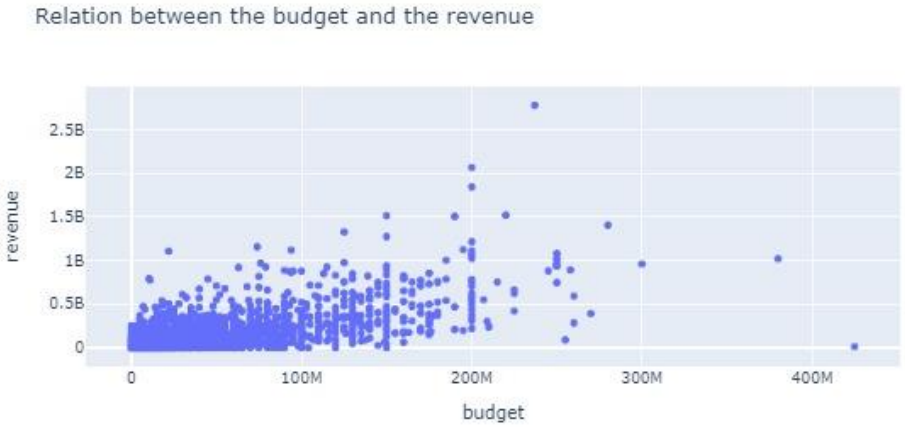
Questions we want to answer it in the analysis:

- What is the relationships between the dataset attributes?
- What is the most popular genre from 1960 to 2015?
- Who are the top 10 directors ?
- Who are the top 10 actors ?
- What is the average revenue in each year?
- What is the average movies budget in each year?
- What is the most famous production company?

First: What is the relationships between the dataset attributes?

We made a scatter plot to see the correlation between the attributes

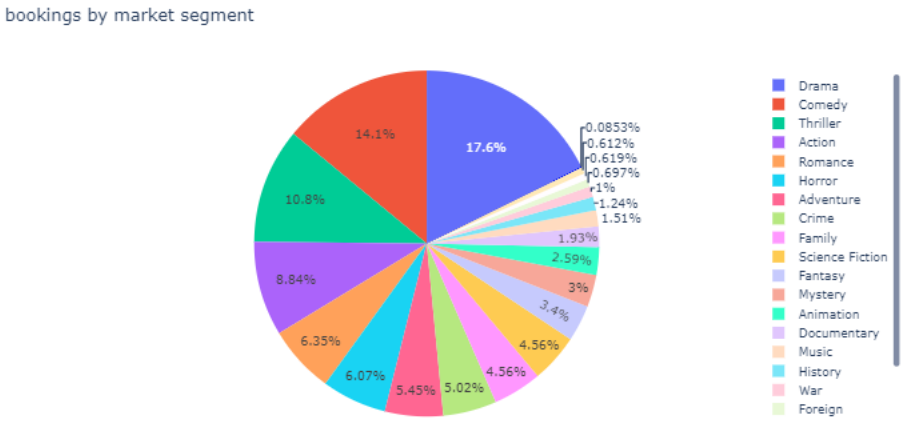




From those plots we can conclude that there is a strong positive correlation between the popularity of the movies and the number of votes and between the budget and the revenue which means most of the movies with high budgets gets high revenue

Second: What is the most popular genre from 1960 to 2015?

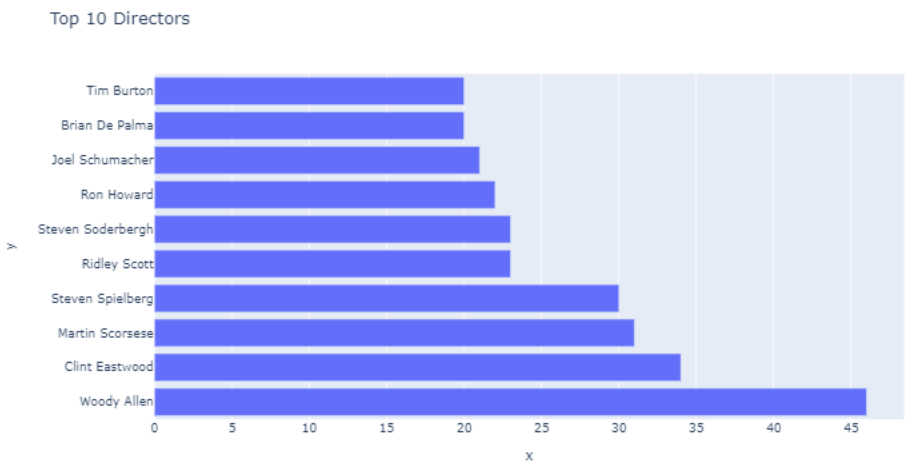
Here we made a list contains all the genres mentioned in the dataset and number of times they appeared then I made a pie chart to find what is the most popular genres



we can conclude here that the most popular movie genre from 1960 to 2015 is Drama the Comedy in the second place and Thriller in the third place

Third: Who are the top 10 directors ?

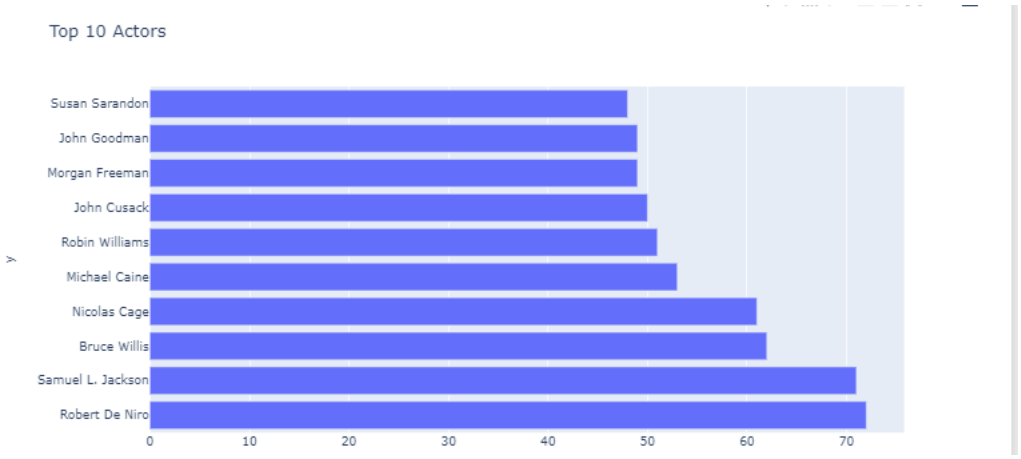
Here we made a list contains all the directors names mentioned in the dataset and also the number of times they appeared in the dataset and then I made a bar plot



we conclude here that the most famous director is Woody Allen with more than 45 movie then comed after him Clint Eastwood with 34 movie and so on..

Fourth: Who are the top 10 actors ?

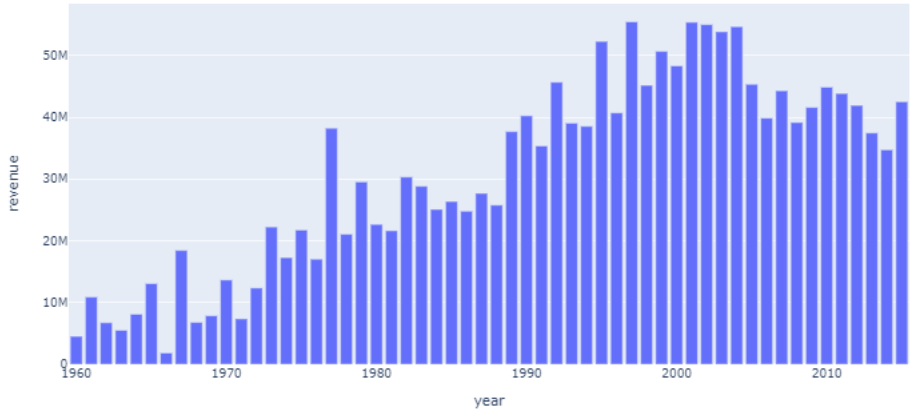
As I done before I made a list contains all the actors name and number of times they appeared in the dataset and I made a bar plot to display the results



Here we can see that the most famous actor is Robert De Niro as he made about 72 movie then in the second place Samuel L.Jackson with about 71 movie

Fifth: What is the average revenue in each year?

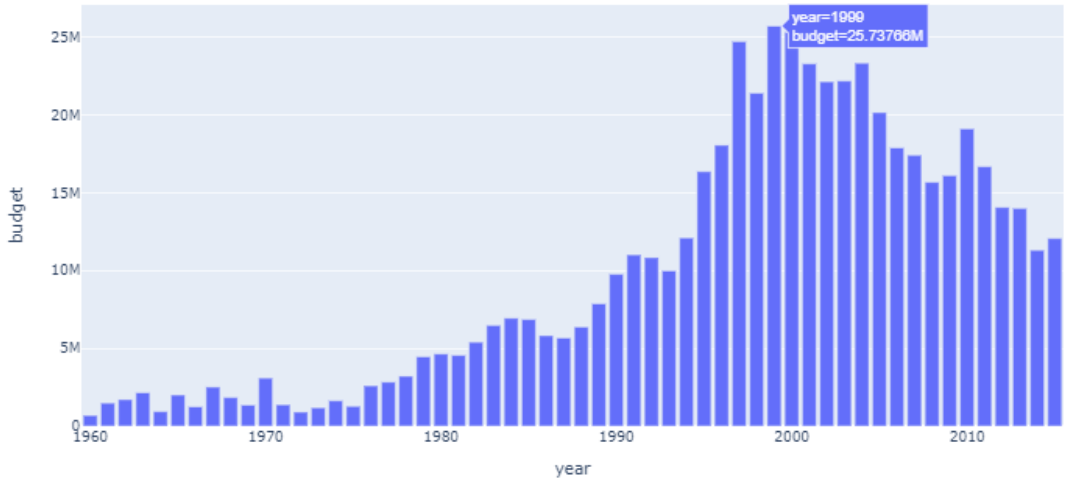
Here I made a new dataframe and I grouped each year with the average revenues in that year and the resulted dataframe contains 2 columns one for year and the other for the average revenues and then uses this data I made a bar plot to display the results



It's obvious here that the biggest average was in 1997 then in 2001

Sixth: What is the average movies budget in each year?

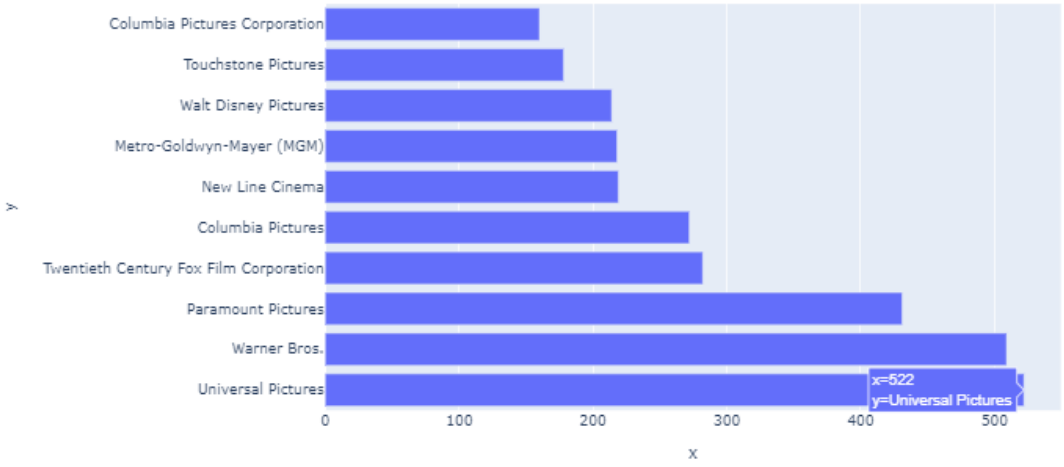
As same as before I made a new dataframe and I grouped each year with the average budgets in that year and the resulted dataframe contains 2 columns one for year and the other for the average budgets and then uses this data I made a bar plot to display the results



It's obvious here that the biggest average budget was in 1999 then in 2000

Seventh: What is the most famous production company?

I made a list contains all the production companies names and number of times they appeared in the dataset and I made a bar plot to display the results



Universal Pictures is the most famous production company then Warner Bros. in the second place