# Project Investigate a Data Set

## Introduction

In this project, I chose to analyze the **IMDB movie data**. The questions that this document attempts to answer are the following:

**Question 1**: Which is the most popular genre for each year?

**Question 2**: What properties are associated with movies that have high revenues?

In order to enable the user to answer the above questions, some data processing was required which I performed in the code provided. In the section below, I will attempt to explain the steps that I went through to answer the above questions. As a first step I load the .csv file provided into a pandas dataframe. This is done by the 1st function in the code, the ***read\_data()*** function.

# Answer to Question 1:

In order to explore and answer the first question, data is required from 3 columns. By noticing the “key” words in question 1, these columns are: **‘grenres’**, **‘release\_year’** and **‘popularity’.**

In the 2nd function of the code, ***miss\_values(df),*** I check if there are any missing values (NaN’s) in any of these 3 columns. Since there are rows that do not contain info on the genre of the film, still inside this function, I drop these rows.

In the previous analysis, I notice that the values in the **‘grenres’** column are multiple for each row and separated by (|). In practice, this means that some films can belong to many film genres and not just one.

For the purpose of our analysis, and in order to be fair and accurate when we answer the question which is the most popular genre, it is essential to split this column in more rows, one for every genre a film belongs to. This way, I will take into account all information to answer to my main question. For this reason, I created the function ***final\_dataframe(df).*** Inside this function, I first split the elements in each row in the **‘grenres’** column and create a list out of it (line 33). Thereafter, using the *.explode* method, I create a new row for every element in the **‘grenres’** column.

Finally, I create the final dataframe which I will use to create figure 1 you see in the next page. Still in the ***final\_dataframe(df)*** function and in line 39, I groupby both the **‘release\_year’** and the **‘grenres’** column to get the average popularity of each film genre for each year.Now, to get the film genres that have the highest average popularity each year, I sort the **‘popularity’** values by descending order and at the same time drop the rows with the same release year since I am only interested in the most popular genre for each year (line 43).

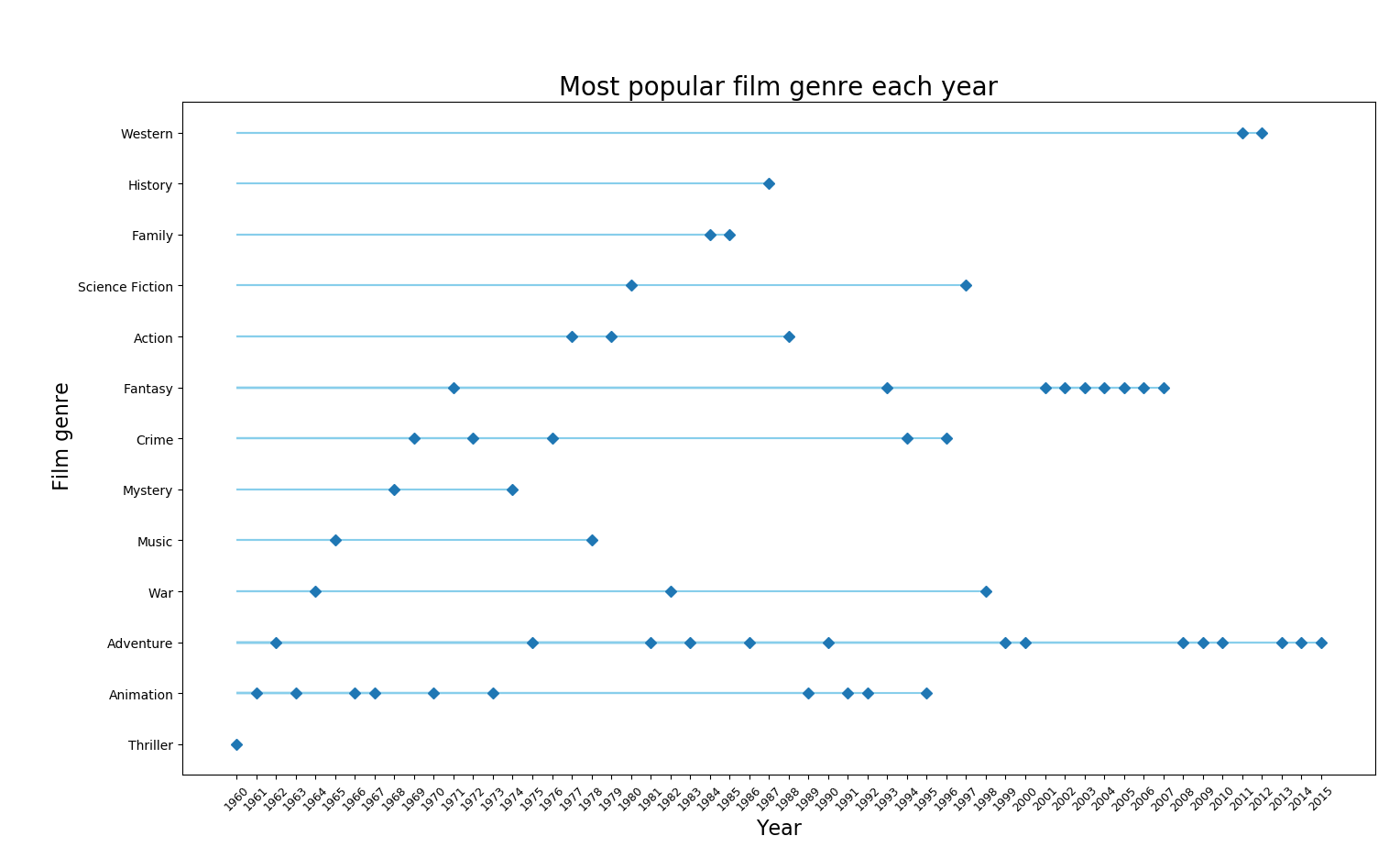
Now I have created the ***df\_year\_genre*** dataframe that consists of 3 columns, the release year of the film, the most popular genre in this year and the score of its popularity. Therefore, I am ready to create the graph that will visualize my result in the figure 1 in the next page.

The 1st part of the ***visualize(df\_year\_genre,df)*** function is used to produce figure 1. The code that does this action is the following:

# Visualize which the year and most popular corresponding genre  
plt.hlines(y=df\_year\_genre['genres'], xmin=1960, xmax=df\_year\_genre['release\_year'], color='skyblue')  
plt.plot(df\_year\_genre['release\_year'], df\_year\_genre['genres'], "D")  
plt.xticks(df\_year\_genre['release\_year'], fontsize=9)  
plt.xticks(rotation=45)  
plt.title('Most popular film genre each year', fontsize=20)  
plt.xlabel('Year', fontsize=16)  
plt.ylabel(' Film genre', fontsize=16);

**Conclusions Question 1:**

Figure 1 that is displayed in the next page shows the most popular film genre for each year. By observing the graph, it’s clearly seen that the most popular genre is changing year by year. However, it seems that the last 15 years the two most popular genres are ‘fantasy’ and ‘adventure’ with the latter having ranked the most popular genre for the most years in the entire history.

fdfgdfyjfjfjvcnv **Figure 1: Most popular film genre**

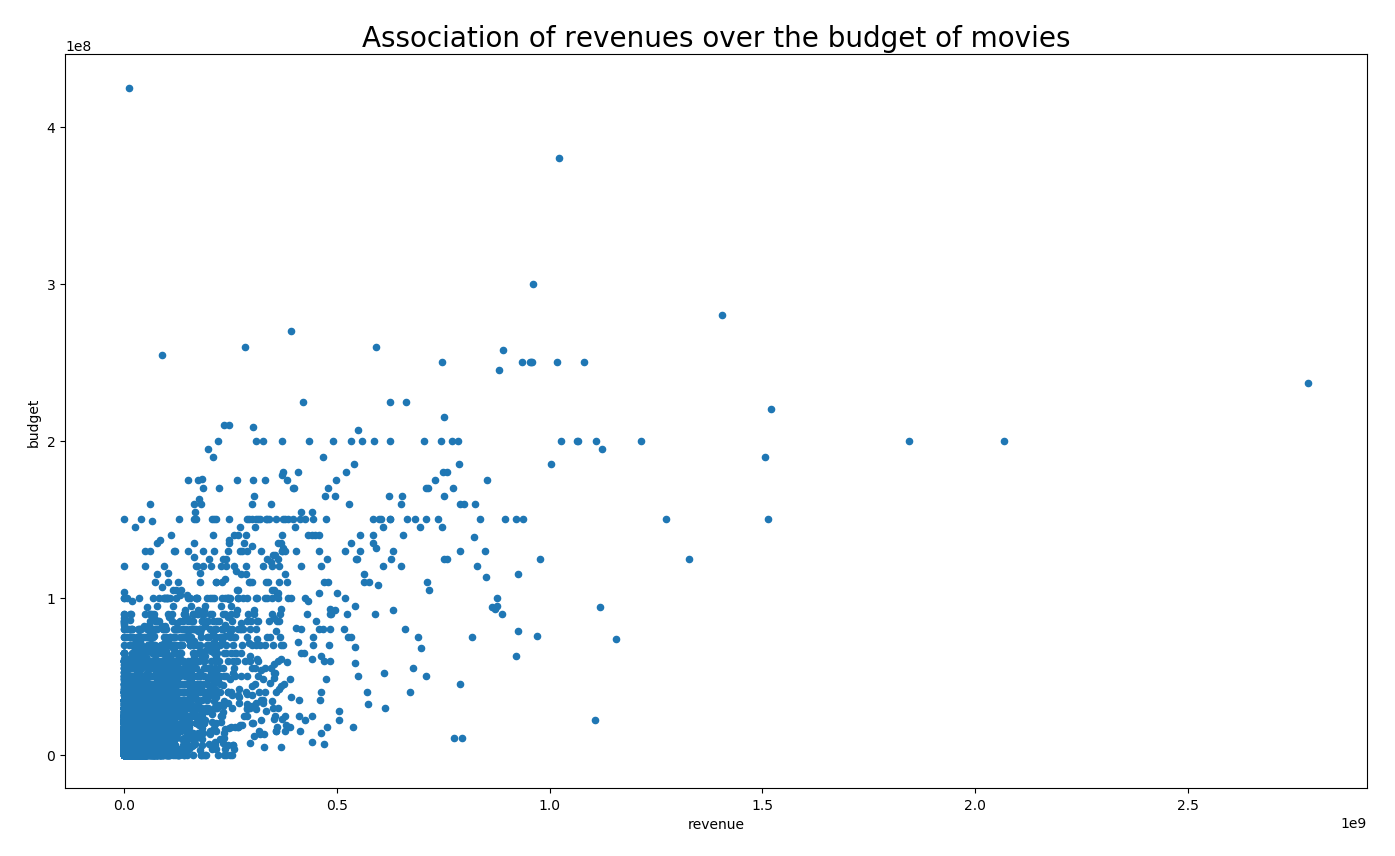
# Answer to Question 2:

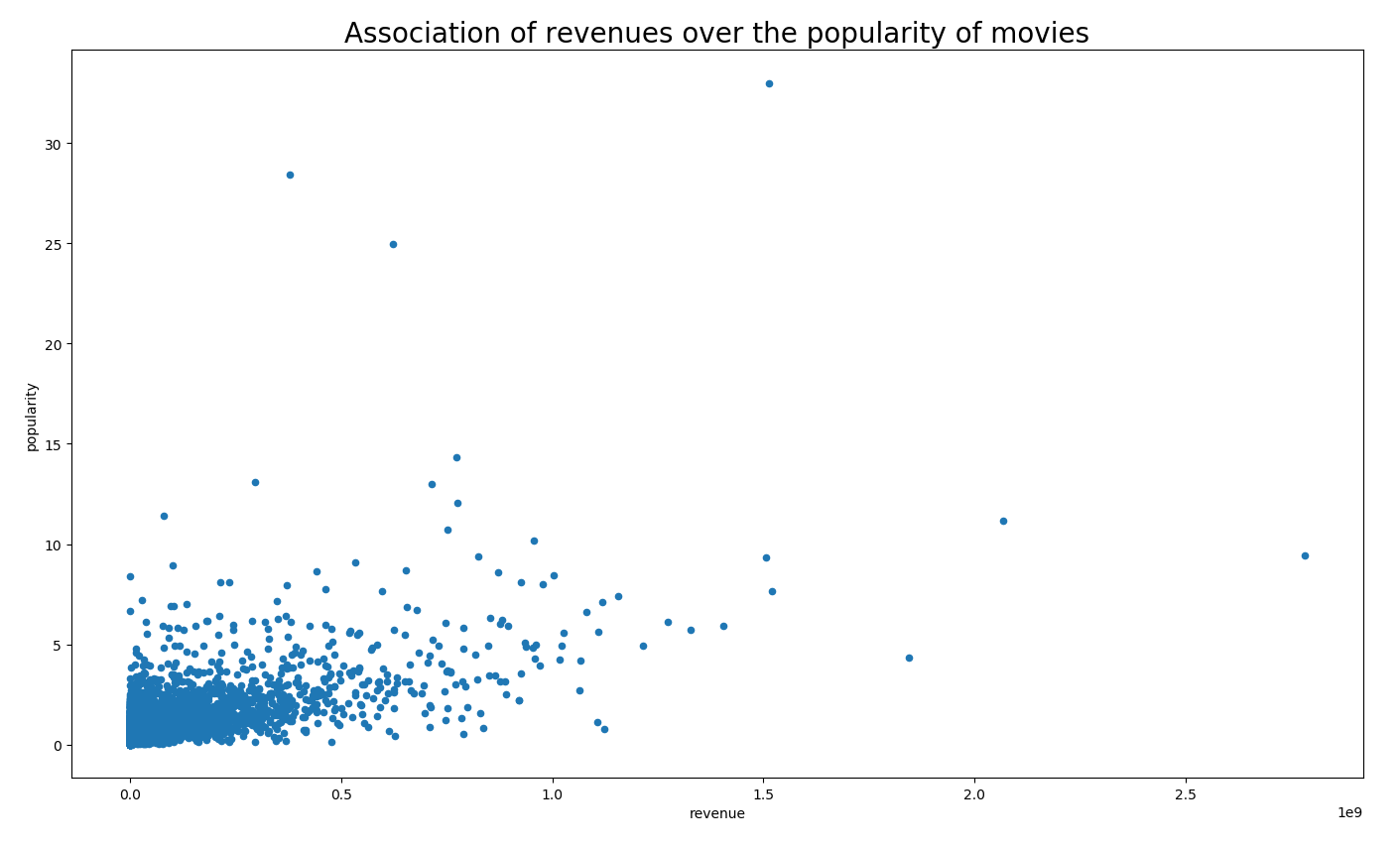
To answer the 2nd question that was posed, it’s necessary to identify the correlation that the **‘revenue’** column has with the rest of the properties described in the IMDB database. In the ***visualize(df\_year\_genre,df)*** function and in line 60 I print the table that shows how is the **‘revenue’** column correlated with the rest of the data. According to theory, the closer to 1 a value is, the highest the relation between the two columns. From the table produced, I note that only 3 columns seem to be associated with the **‘revenue’** column. These are the **‘popularity’**, **‘budget’** and **‘vote\_count’**. I distinguish these 3 columns as they are the only ones that have correlation higher than 0.5 with **‘revenue’** column. What’s left is to illustrate this correlation. That is being done through the code in the ***visualize(df\_year\_genre,df)*** in lines 61 – 66 where the 3 plots showing the relation with the revenues are produced. All three figures are shown in the pages below.

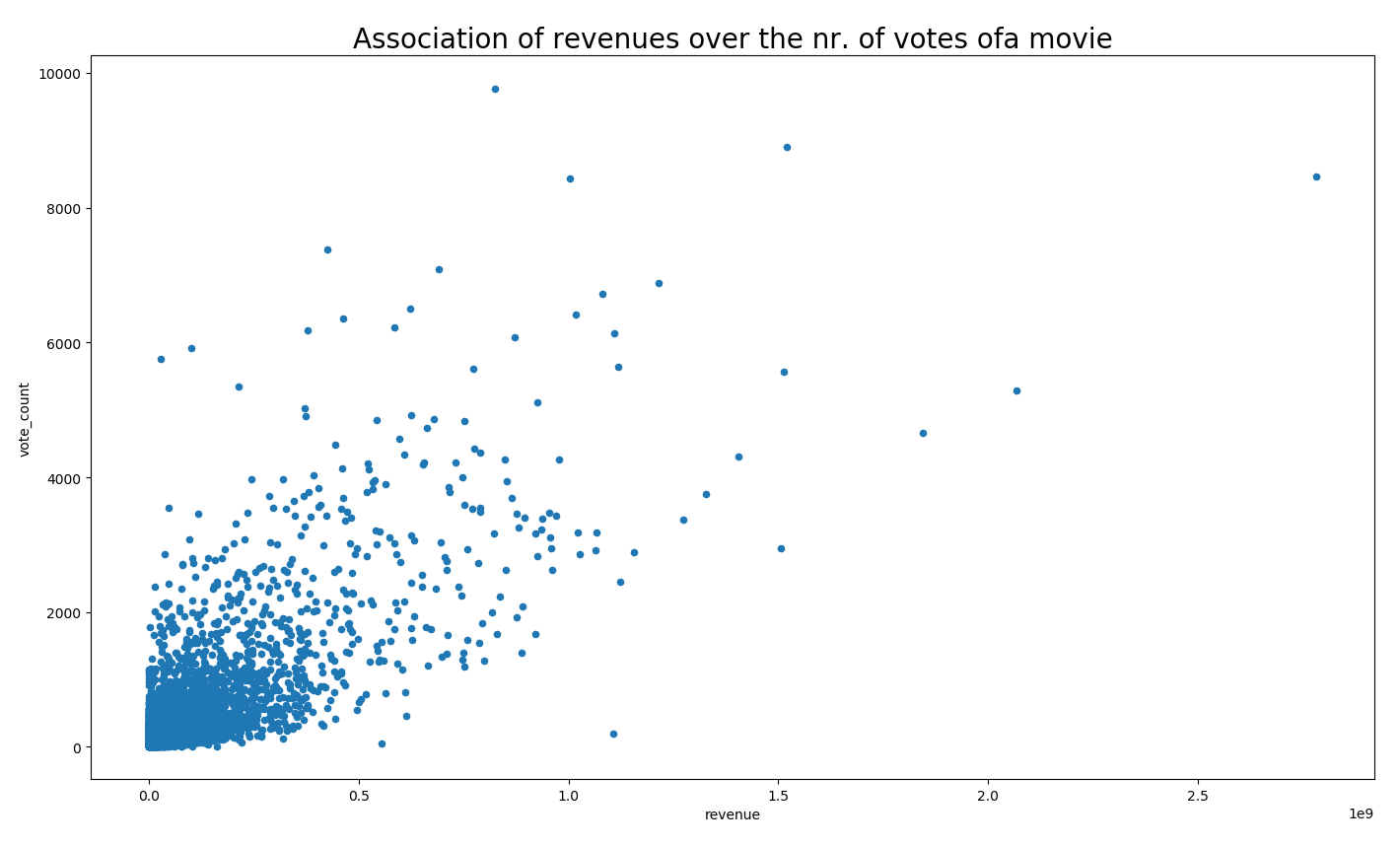
**Conclusions Question 2:**

From Figure 2, Figure 3 and Figure 4 in the next page, we can see that all 3 properties have a positive correlation with the revenues of the film. Namely, the higher the budget that is invested in the film the higher the possibility that will yield high revenues. The same applies for the popularity of the film as well as the number of votes that it receives.

From all the three properties that are associated with high revenues, it’s interesting to see that the number of votes a film receives is the safest indicator of high revenues as their correlation is the strongest. The latter is confirmed also from the value that this correlation has (0.79) which is the highest of the three. However, although all these 3 properties seem to be associated with high revenues, their correlation is not very strong to safely conclude that a high value of a property would definitely result in a high revenue movie.

 **Figure 2: Correlation of revenues and budget of films**

 **Figure 3: Correlation of revenues and popularity of films**

 **Figure 4: Correlation of revenues and number of votes of films**