THE FINAL REPORT OF DATA VISUALIZATION

Marvel vs DC Data Analysis

PREPARED BY…

1. **MAAZ SHAHNAWAZ**
2. **SUMAIYA AFZAL**
3. **MUHAMMAD USAMA**

Table of Contents

Introduction

There are many ways to compare comic's universes. Here I explain how

I created a visualization to compare Marvel and DC universes using data.

My goal was to define a simple visualization to show differences between Marvel and DC in a clear and fast way. I plotted a graph for each Company and showed them side-by-side.

Aim of Project

MCU vs DC. Which one is better? Which has more high-rated movies? Analysis of Marvel and DC movies based on gross value.

Marvel Cinematic vs DC Universe, it’s a never-ending debate, right? Fans got crazy when you oppose any of these cinematic universes. But in the article, we are going to do a fight over Marvel vs DC based on some data. Data always tells the truth. So, let’s start this data war, with a cup of coffee.

MCU vs DC

You can write the Python code in Jupyter Notebook**,** Google Colab, or any other preferred editor. I will recommend you Jupyter Notebook because I use it more often.

What is Data Visualization?

Data Visualization techniques are one of the key components of any analytics project. An end-to-end analytics use case involves ideation, requirement gathering, getting the raw data, analyzing the data, building a predictive model, deploying the model, and communicating the end result to the business.

Throughout this entire process, the analysis of data, and the communication of results to the business requires visualizing the raw data and understanding several inter-linked relations among the features. Python is the most preferred language which has several libraries and packages such as Pandas, NumPy, Matplotlib, Seaborn, and so on used to visualize the data.

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Import the libraries

- ## NumPy

  NumPy offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.

- ## Pandas

Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL database tables or queries, and Microsoft Excel.

- ## Matplotlib

  Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- ## Seaborn

  Seaborn is an open-source Python library built on top of Matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with data frames and the Pandas library.

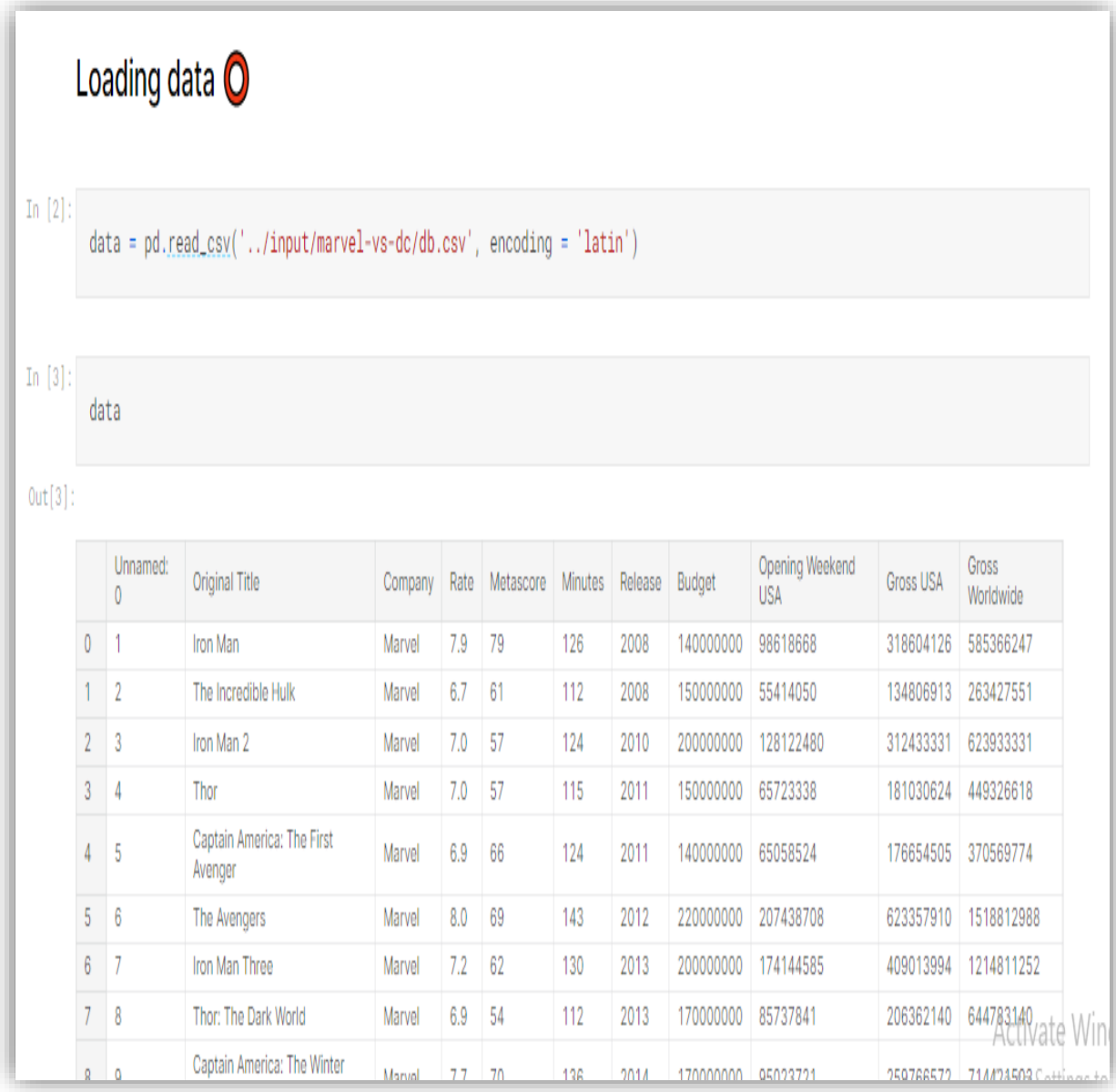
- ## Os

The OS module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules. This module provides a portable way of using operating system-dependent functionality.

In [1]:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version u
sing "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

- **Let’s load the data and take a sneak peek at the data.**



Loading data

In [2]:
```
data = pd.read_csv('../input/marvel-vs-dc/db.csv', encoding = 'latin')
```

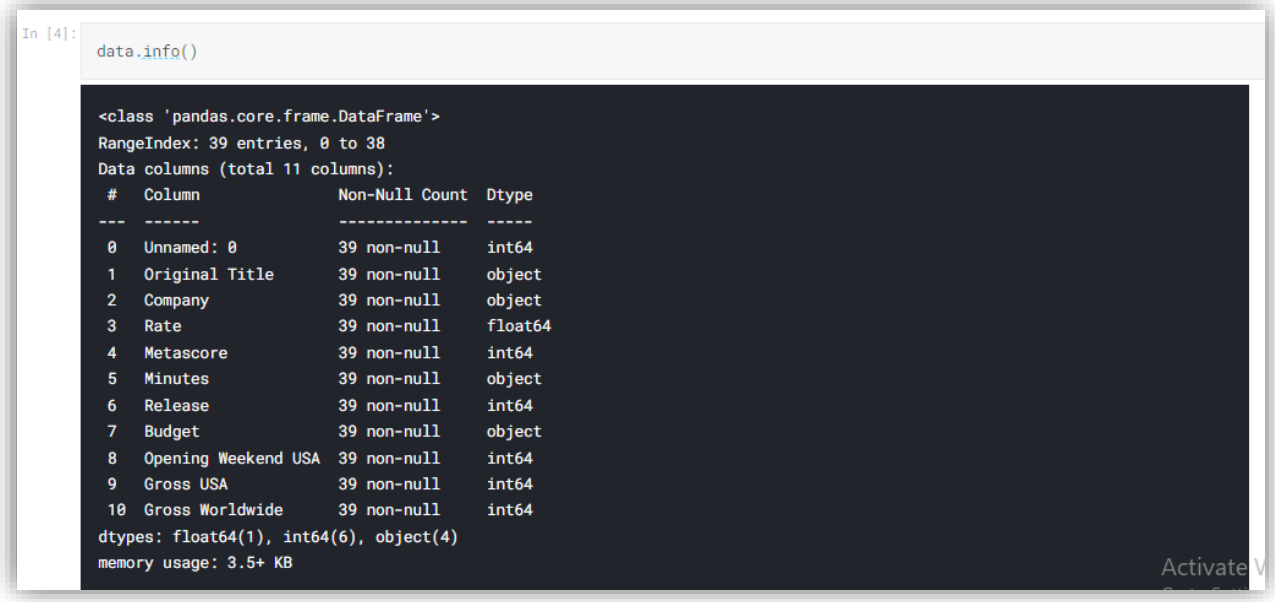
In [3]:
```
data
```

Out[3]:

	Unnamed: 0	Original Title	Company	Rate	Metascore	Minutes	Release	Budget	Opening Weekend USA	Gross USA	Gross Worldwide
0	1	Iron Man	Marvel	7.9	79	126	2008	140000000	98618668	318604126	585366247
1	2	The Incredible Hulk	Marvel	6.7	61	112	2008	150000000	55414050	134806913	263427551
2	3	Iron Man 2	Marvel	7.0	57	124	2010	200000000	128122480	312433331	623933331
3	4	Thor	Marvel	7.0	57	115	2011	150000000	65723338	181030624	449326618
4	5	Captain America: The First Avenger	Marvel	6.9	66	124	2011	140000000	65058524	176654505	370569774
5	6	The Avengers	Marvel	8.0	69	143	2012	220000000	207438708	623357910	1518812988
6	7	Iron Man Three	Marvel	7.2	62	130	2013	200000000	174144585	409013994	1214811252
7	8	Thor: The Dark World	Marvel	6.9	54	112	2013	170000000	85737841	206362140	644783140
8	9	Captain America: The Winter	Marvel	7.7	70	136	2014	170000000	95023721	259766572	714421503

We have names of movies, year of release, genre, IMDB rating, IMDB gross, entity, and so on.

- **Gather some more information of data.**

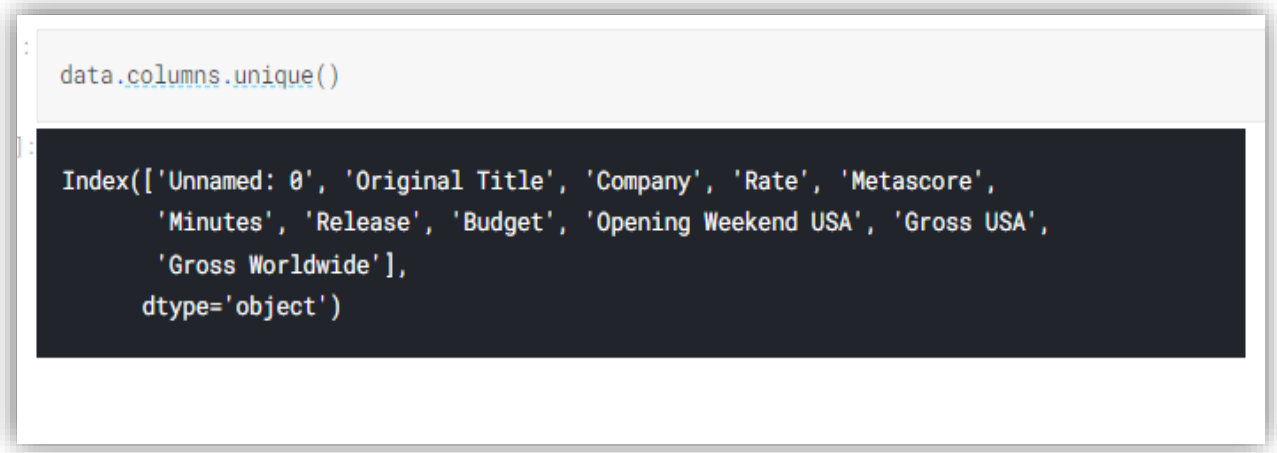


In [4]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39 entries, 0 to 38
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Unnamed: 0           39 non-null     int64
 1   Original Title       39 non-null     object
 2   Company              39 non-null     object
 3   Rate                 39 non-null     float64
 4   Metascore            39 non-null     int64
 5   Minutes              39 non-null     object
 6   Release              39 non-null     int64
 7   Budget               39 non-null     object
 8   Opening Weekend USA  39 non-null     int64
 9   Gross USA            39 non-null     int64
 10  Gross Worldwide      39 non-null     int64
dtypes: float64(1), int64(6), object(4)
memory usage: 3.5+ KB
```

Check out the null values in each column.
After that, get more information about our dataset with the type of each column attributes.



```
data.columns.unique()
```

```
Index(['Unnamed: 0', 'Original Title', 'Company', 'Rate', 'Metascore',
       'Minutes', 'Release', 'Budget', 'Opening Weekend USA', 'Gross USA',
       'Gross Worldwide'],
      dtype='object')
```

Checking the columns and using Unique to avoid repeated columns.

```
data = data.drop(['Unnamed: 0', 'Original Title'], axis = 1)
```

```
data
```

	Company	Rate	Metascore	Minutes	Release	Budget	Opening Weekend USA	Gross USA	Gross Worldwide
0	Marvel	7.9	79	126	2008	140000000	98618668	318604126	585366247
1	Marvel	6.7	61	112	2008	150000000	55414050	134806913	263427551
2	Marvel	7.0	57	124	2010	200000000	128122480	312433331	623933331
3	Marvel	7.0	57	115	2011	150000000	65723338	181030624	449326618
4	Marvel	6.9	66	124	2011	140000000	65058524	176654505	370569774
5	Marvel	8.0	69	143	2012	220000000	207438708	623357910	1518812988
6	Marvel	7.2	62	130	2013	200000000	174144585	409013994	1214811252

Dropping unnecessary columns.

All columns except 'Company' type should be converted to either float or int type.

```
convert_dtype = {
    'Minutes' : int,
    'Budget' : int
}

data = data.astype(convert_dtype)
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39 entries, 0 to 38
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Company             39 non-null     object
 1   Rate                39 non-null     float64
 2   Metascore           39 non-null     int64
 3   Minutes             39 non-null     int64
 4   Release             39 non-null     int64
 5   Budget              39 non-null     int64
 6   Opening Weekend USA 39 non-null     int64
```

Converted

Now find out the mean or average of each rating, metascore, length, budget etc.

```
data[data['Company'] == 'DC'].mean()
```

```
Rate                  6.806250e+00
Metascore             5.650000e+01
Minutes               1.341250e+02
Release               2.012500e+03
Budget                1.716250e+08
Opening Weekend USA   8.637872e+07
Gross USA             2.538687e+08
Gross Worldwide       6.056326e+08
dtype: float64
```

```
data[data['Company'] == 'Marvel'].mean()
```

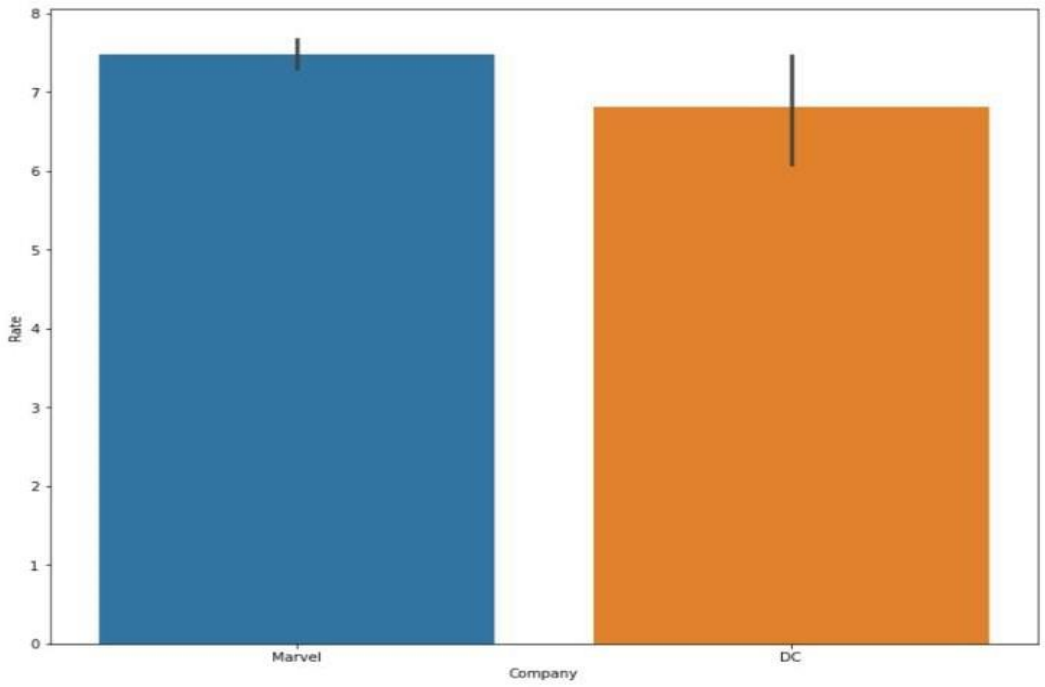
```
Rate                  7.478261e+00
Metascore             6.865217e+01
Minutes               1.302609e+02
Release               2.014696e+03
Budget                1.927826e+08
Opening Weekend USA   1.350966e+08
Gross USA             3.715423e+08
Gross Worldwide       9.819657e+08
dtype: float64
```

The average rating of DC movies is 6.886 and for Marvel movies, it's 7.47. DC has one of the highest-rated movies of all time.

Graphs showing correlation 📈

In [12]:

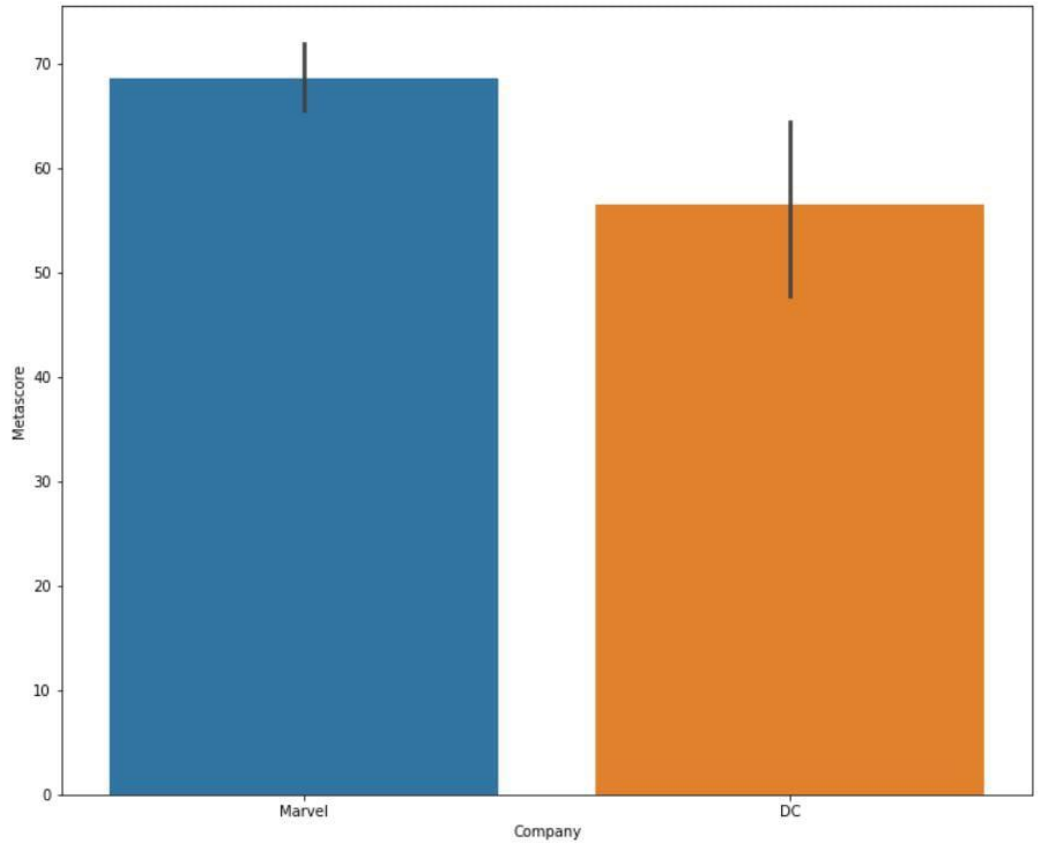
```
plt.figure(figsize = (12, 10))
sns.barplot(x = 'Company', y =
'Rate', data = data)
plt.show()
```



The above graph clearly shows the rating of Marvel movies is greater than DC. In the ratings game, Marvel wins by a large margin: 66% of Marvel films are certified fresh compared to 54% of DC films. Between the box office numbers and ratings, Marvel is still coming out on top.

In [13]:

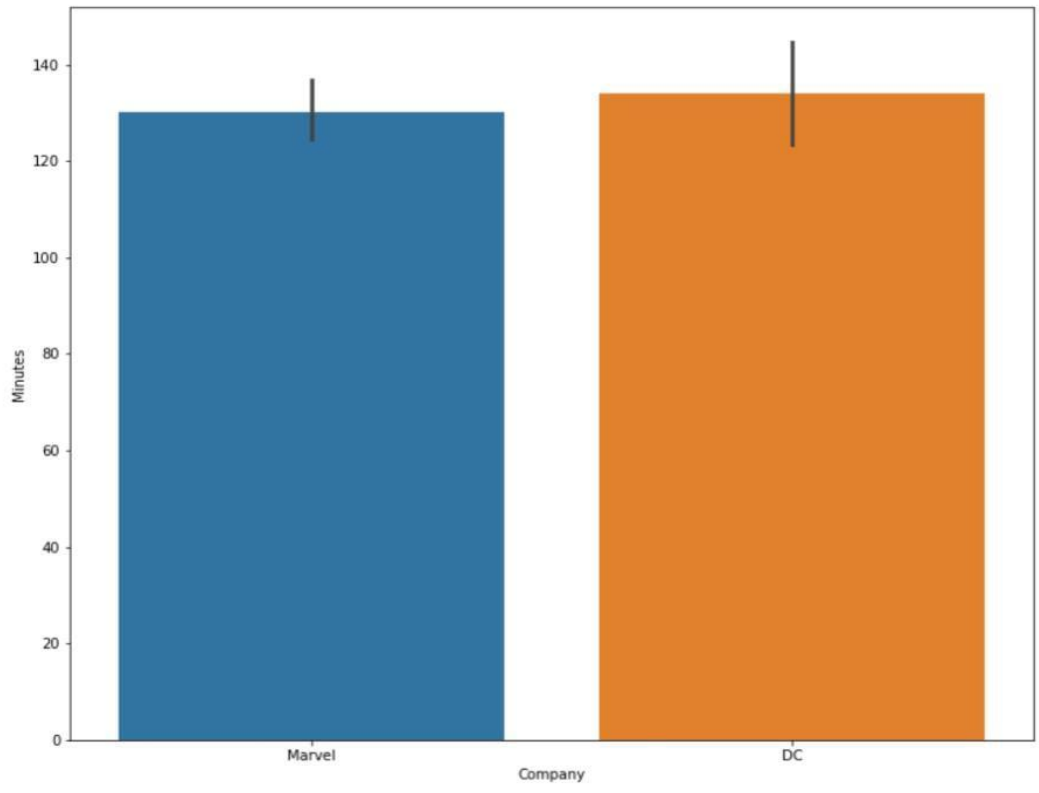
```
plt.figure(figsize = (12, 10))
sns.barplot(x = 'Company', y =
'Metascore', data = data)
plt.show()
```



This graph shows the Metascore of Marvel is also higher than DC. According to critics, Marvel is better than DC, but only by a hair's breadth, with an average Metacritique score of 58.73 and over DC's 56.71.

In [14]:

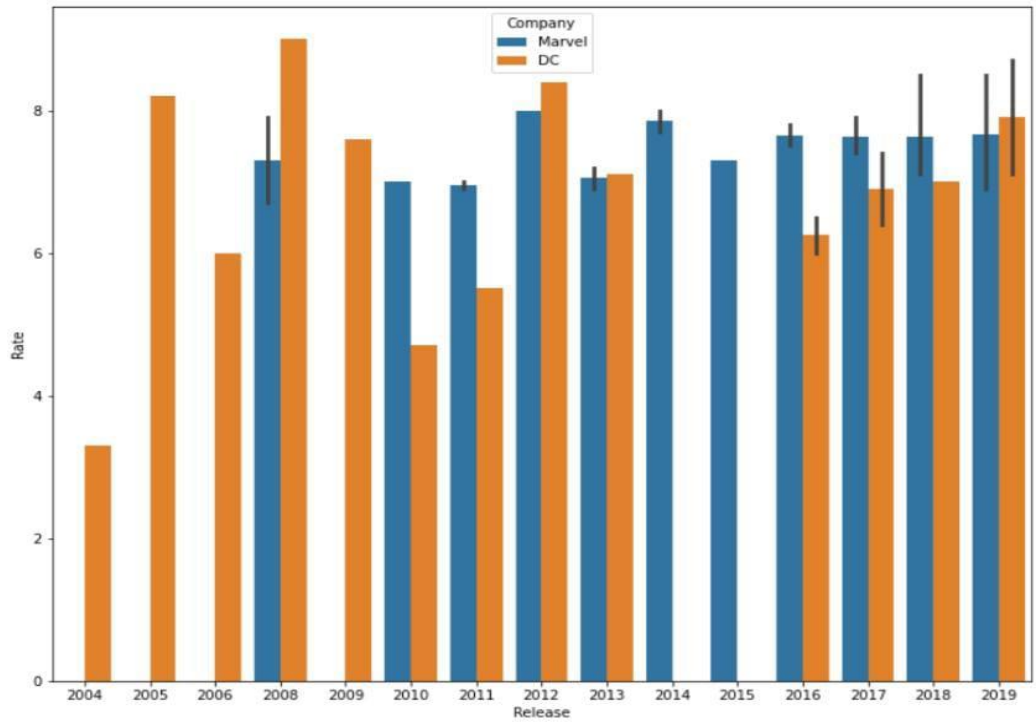
```
plt.figure(figsize = (12, 10))
sns.barplot(x = 'Company', y =
'Minutes', data = data)
plt.show()
```



This graph shows the length of the movies. Here the length of DC movies is greater than Marvel.

In [15]:

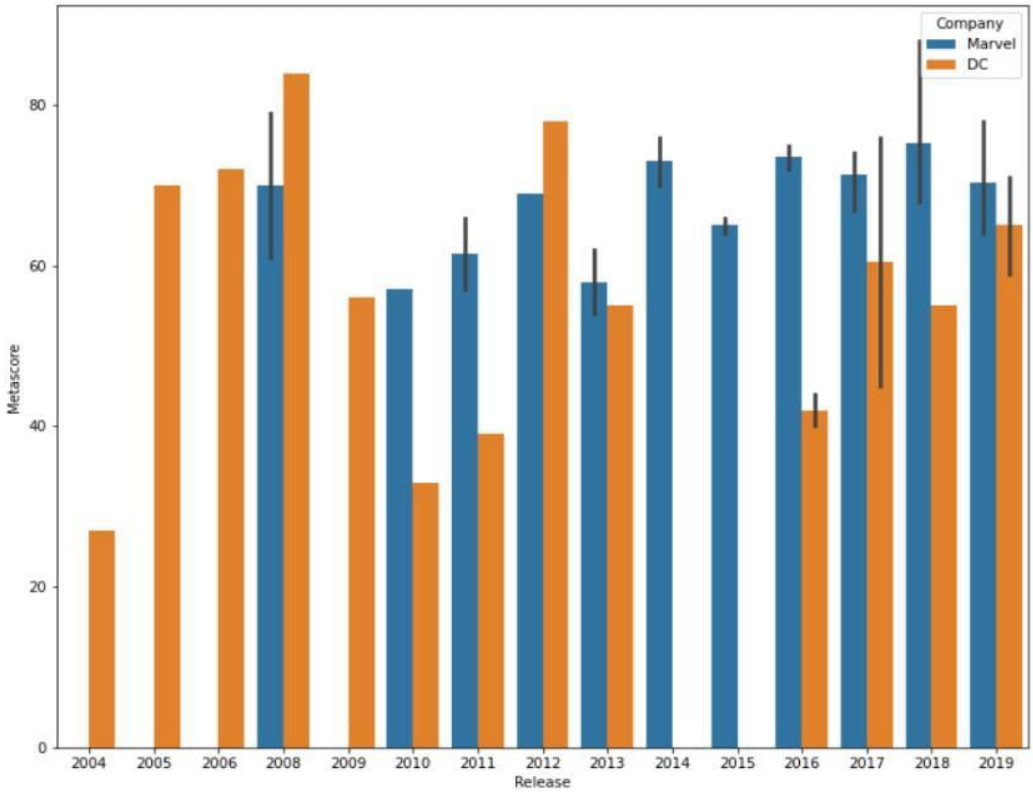
```
plt.figure(figsize = (12, 10))
sns.barplot(x = 'Release', y =
'Rate', hue = 'Company', data =
data)
plt.show()
```



This graph shows rating of each movies in the year 2004 to 2019.The Dark Knight is the Top-rated DC movie. It has an IMDB rating of 9. If you didn't watch it yet then do watch. You will witness the legendary act of Sir Heath Ledger. This movie shows that what DC Universe is capable of.

In [16]:

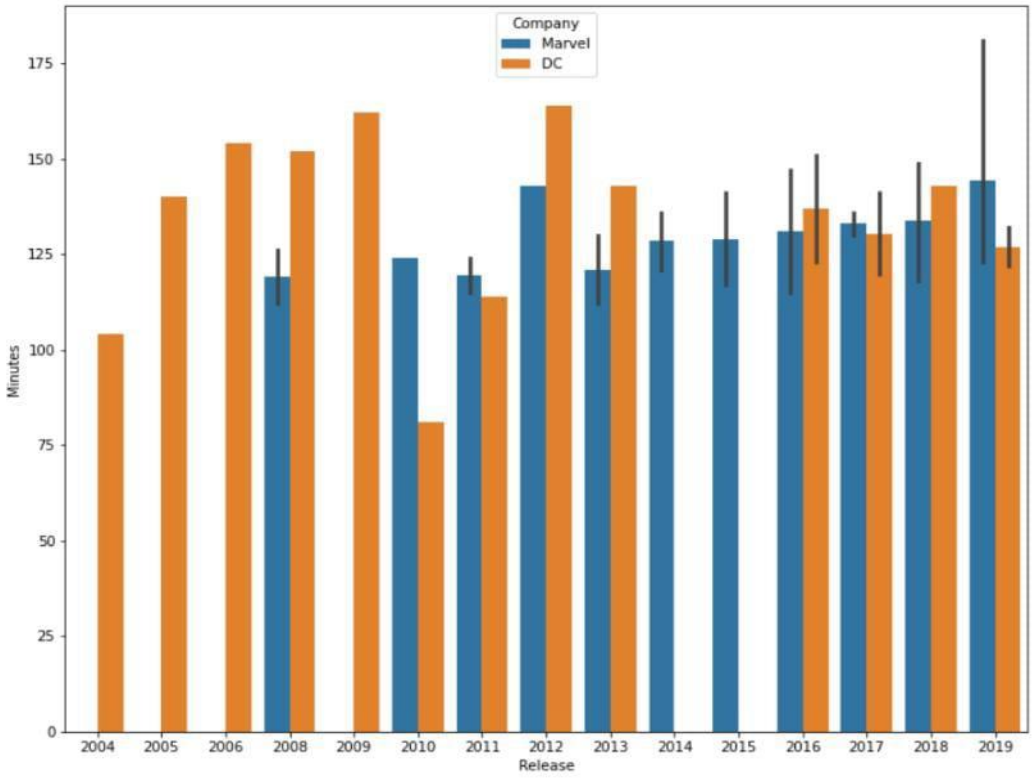
```
plt.figure(figsize = (12, 10))
sns.barplot(x = 'Release', y =
'Metascore', hue = 'Company', da
ta = data)
plt.show()
```



The critics also favor Marvel as the metascore of Marvel movies are also greater than DC.

In [17]:

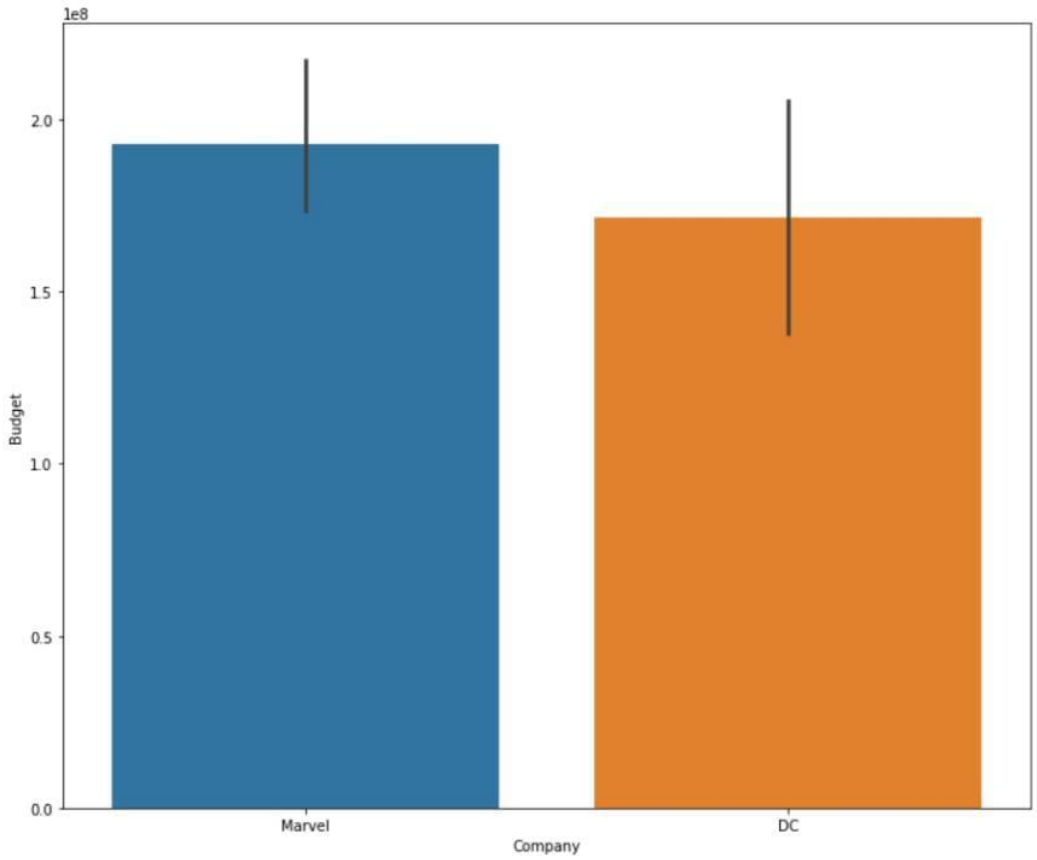
```
plt.figure(figsize = (12, 10))
sns.barplot(x = 'Release', y =
'Minutes', hue = 'Company', data
= data)
plt.show()
```



The average runtime of both the Marvel and DC movies is almost equal. But there is a huge difference there in highest runtime movies.

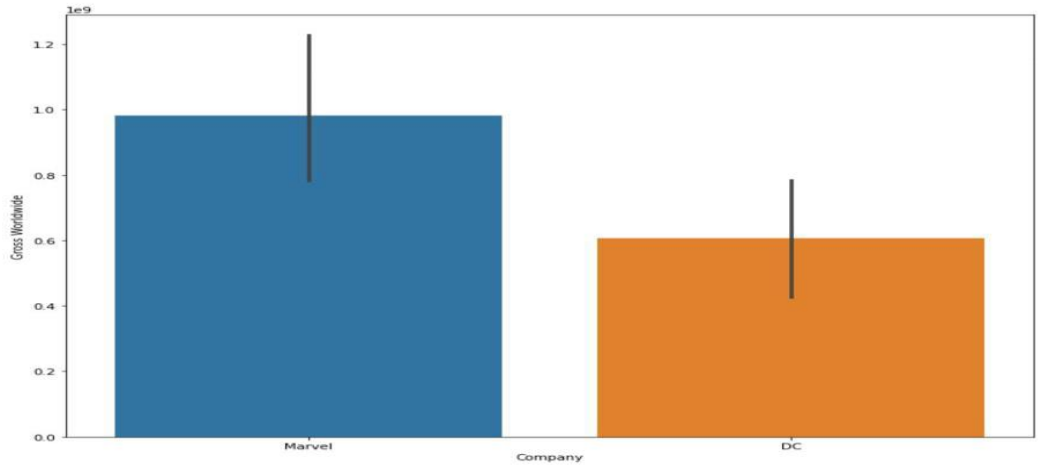
In [18]:

```
plt.figure(figsize = (12, 10))
sns.barplot(x = 'Company', y =
'Budget', data = data)
plt.show()
```



The Budget of Marvel is slightly higher than Dc.

```
plt.figure(figsize = (12, 10))
sns.barplot(x = 'Company', y =
'Gross Worldwide', data = data)
plt.show()
```



In terms of gross, few Marvel movies are far away from DC Movies.

Most of the Marvel movie has IMDB ratings lies between 6.7 to 8.2.

DC movies ratings are evenly distributed across the graph.

DC movies are performing well on IMDB gross but if you compare it with Marvel then they fall short.

We cannot compare both the Movie Making production house because who knows the future. In the future, DC may overshadow Marvel. But the best part is that both these productions houses are making good movies and entertained the audience for the past few decades.

Conclusion

- Avg. rating of Marvel is greater than DC.
- Avg. metascore of Marvel is greater than DC.
- Avg. duration of DC movies is greater than Marvel.
- DC movies were released earlier than birth of Marvel, but since 2010 people tended to like Marvel more over the years.
- Gradually duration of movies of Marvel increased and whereas of DC decreased.
- Moreover, we can see that the avg. budget for Marvel was more than DC, but the outcome was also good as the Gross World was also high. So, it was worth it!!!

Well, in my opinion as per stats Marvel won, but let's see how future places these two in the tough competition.

Well, that’s it for this article.

If this article sounds informative to you, make sure to follow and share it with your geek community.

References

https://www.kaggle.com/ikarosalpha/marvel-v-s-dc

https://medium.com/geekculture/marvel-vs-dc-data-analysis-in-python-e561cac72358

https://betterprogramming.pub/how-to-perform-exploratory-data-analysis-with-marvel-vs-dc-comics-data-ec75f457ac60