

## CA682 Data Management and Visualisation

Name	Gautam Shanbhag
Student Number	18210455
Programme	MCM - DA
Module Code	CA682
Assignment Title	Data Visualisation
Submission date	17 <sup>th</sup> Dec 2018
Module coordinator	Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations set out in the module documentation. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found recommended in the assignment guidelines.

Name: Gautam Shanbhag

Date: 17<sup>th</sup> Dec 2018

# Marvel Studios

## 1. Introduction

My topic for Data Visualisation Assignment is Marvel Movies and its box office collection. The idea for this topic was my ever-growing craze for the marvel comics and movies since childhood. My purpose for visualisation was to identify which movie has had earned better domestic or in international market and comparing the collections with their reviews. I have tried to analyse the actor's career wise impact due to marvel and their appearances in marvel cinematics. I have collected data from various sources and manually extracted and merged it into excel. Dataset being short, I merged various other dataset so that interesting facts and charts could be plotted.

## 2. Dataset

- I have used the following link to fetch data  
<https://www.the-numbers.com/movies/franchise/Marvel-Cinematic-Universe#tab=summary>
- My next dataset is extracted manually opening each film and extracting the cast list for the particular movie
- Other data set is taken from  
<https://www.the-numbers.com/movies/franchise/Marvel-Cinematic-Universe#tab=acting>
- I have also taken IMDB rating from  
<https://www.imdb.com/imdbpicks/MCU-movies-ranked-by-imdb/ls038472133/mediaviewer/rm1589293056>
- Finally for the movie logo's, I individually googled it and cropped to the desired pixel size

### DATA CLEANING:





- Ms Excel
- Open Refine
- Python

Data cleaning was primarily done using MS Excel.

I have used Open Refine to have consistent values in my dataset. In addition, one of fields was comma separated and I have extracted that into columns using Open Refine.

Python dataframe was used to perform aggregate operations for plotting data.

### DATASET 1 SAMPLE:

	B	C	D	E	F	G	H	I
1	TITLE	IMDB RATING	PRODUCTION BUDGET	OPENING WEEKEND	DOMESTIC BOX OFFICE	WORLDWIDE BOX OFFICE	LOGO	IMAGES
17	Spider-Man: Homecoming	7.5	175000000	117027503	334201140	880070886		C:\Users\Gautam\Desktop\ma images\Spider-Man-Homecon
18	Thor: Ragnarok	7.9	180000000	122744989	315058289	846980024		C:\Users\Gautam\Desktop\ma images\Ragnarok.jpg
19	Black Panther	7.5	200000000	202003951	700059566	1347071259		C:\Users\Gautam\Desktop\ma images\Black-Panther.jpg
20	Avengers: Infinity War	8.7	300000000	257698183	678815482	2008425124		C:\Users\Gautam\Desktop\ma images\Avengers-Infinity-War

### DATASET 2 SAMPLE:

	A	B
1	TITLE	CAST
275	Iron Man 3	Rebecca Hall
276	Iron Man 3	Robert Downey Jr.
277	Iron Man 3	Stan Lee
278	Iron Man 3	Stephanie Szostak
279	Iron Man 3	Ty Broussard
280	Iron Man 3	William Sadler
281	Spider-Man: Far From Home	
282	Spider-Man: Homecoming	Bokeem Woodbine
283	Spider-Man: Homecoming	Chris Evans
284	Spider-Man: Homecoming	Donald Glover
285	Spider-Man: Homecoming	Gwyneth Paltrow

### DATASET 3 SAMPLE:

	A	B	C	D	E	F
1	Person	Nr. Of Movies	Role	Franchise Worldwide Box Office	Career Worldwide Box Office	Franchise/Career
2	Stan Lee	20	Himself	\$17,455,564,691	\$26,241,821,085	66.50%
3	Robert Downey Jr.	10	Tony Stark/Iron Man	\$9,636,820,961	\$12,260,462,709	78.60%
4	Samuel L. Jackson	10	Nick Fury	\$7,670,001,203	\$21,082,254,980	36.40%
5	Chris Evans	8	Steve Rogers / Captain America	\$8,034,498,559	\$9,411,837,504	85.40%
6	Gwyneth Paltrow	7	Pepper Potts	\$6,828,152,115	\$9,030,242,330	75.60%
7	Chris Hemsworth	6	Thor	\$6,870,284,142	\$8,601,823,064	79.90%
8	Scarlett Johansson	6	Natasha Romanoff/Black Widow	\$7,405,014,286	\$11,690,053,101	63.30%
9	Paul Bettany	6	Vision	\$7,906,004,669	\$10,554,732,068	74.90%
10	Don Cheadle	5	Lt. Col. James "Rhodey" Rhodes	\$6,388,068,772	\$9,432,121,645	67.70%
11	Sebastian Stan	5	Bucky Barnes/Winter Soldier	\$4,233,477,813	\$5,113,057,798	82.80%

### 3. Process

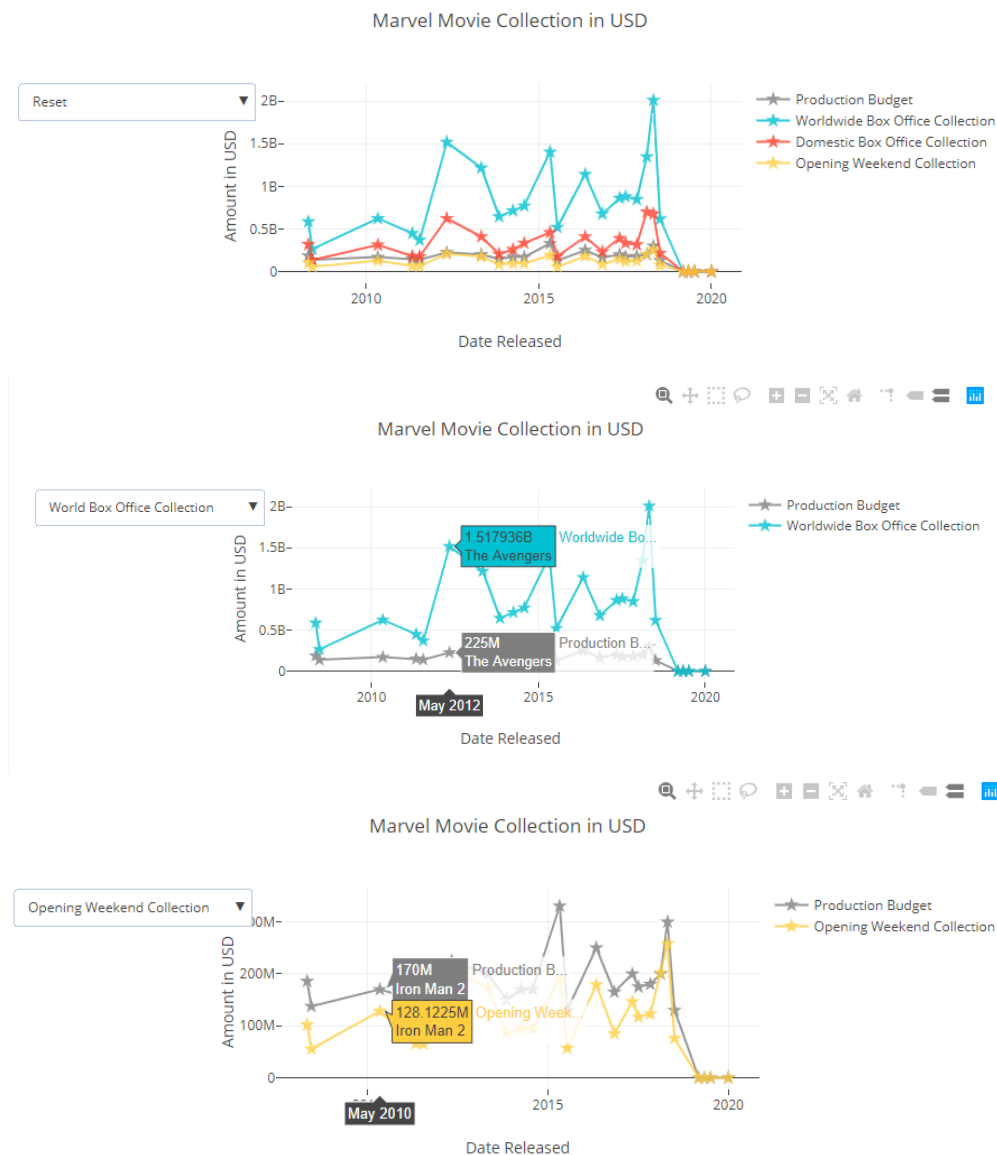
#### ❖ Interactive Line Chart with Dropdown options

My First graph is a line graph, which portrays Marvel Movie Collection in USD. X axis depicts the Date of Movie Release against Y axis depicting Amount in USD. The line graph projects 4 charts ie Production Budget, Worldwide Box Office Collection, Domestic Box Office Collection & Opening Box Office Collection, which is depicted by the legends. Dropdown options are provided which plots Worldwide Box Office Collection, Domestic Box Office Collection & Opening Box Office against Production Budget.

On Hover, the Movie Title with its respective Collection is visible. I have used datetime function to define X axis. Created 4 scatter plots and passed it into data. Then created a update menu bar which would toggle the 4 graphs depending on selection by basically hiding/showing the corresponding graphs. Used vibrant colors with Star markers to highlight points on each plot.

#### Visualization created using:

Plotly Using Python on Jupyter Notebook

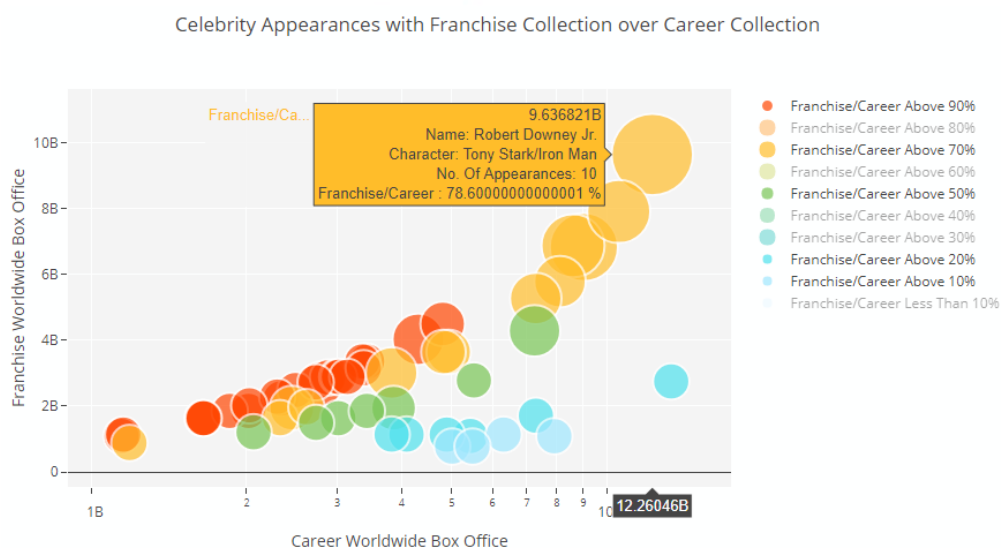
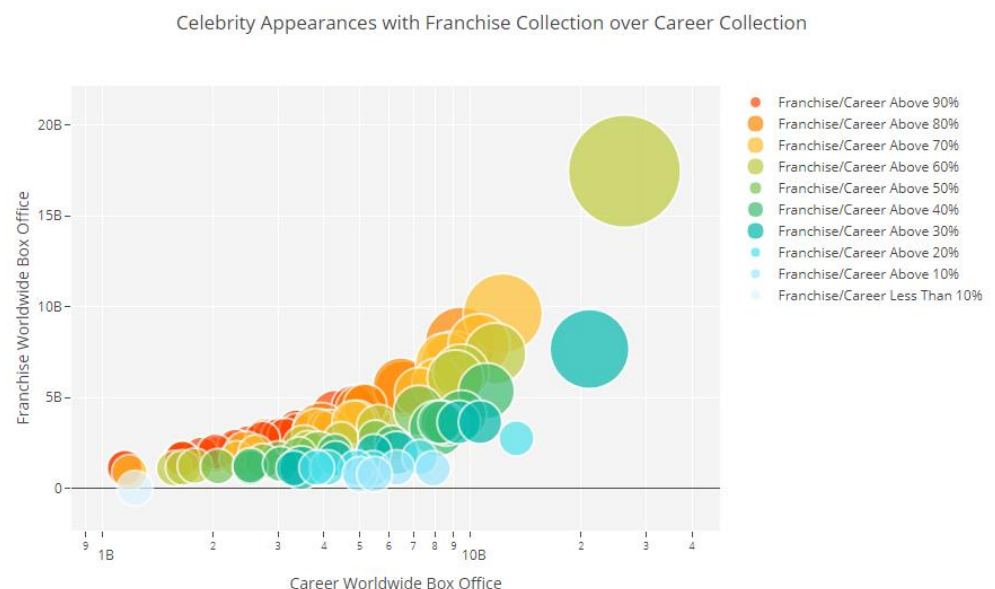


## ❖ Logarithmic Bubble Scatter Plot

Second Graph is a scatter plot with various attributes combined together. X-axis depicts celebrities' worldwide box office collection and Y-axis represents his/her Franchise worldwide box office collection. The size of the circle represents the no. of appearances in a marvel movies whereas the different colors symbolize the % of Franchise/Career in earning ie Marvel's impact on the celebrities total income. Due to big dataset for cast, I have used a logarithmic scale so that the points are distributed across the space. Every range is divided into traces with unique color and hover text info. Legends were not getting displayed properly due to plotly limitations. Color gradient is used with opacity increasing as we move from 90% to less than 10% of ratio of Franchise/Career. Manually assigned custom colors to each trace and calculated bubble size for plotting. Hover information displays the relevant information such as Celebrity Name ~ with his marvel character, No. of appearances, Franchise Collection vs Career Collection and the ratio between them.

### Visualization created using:

Plotly Using Python on Jupyter Notebook

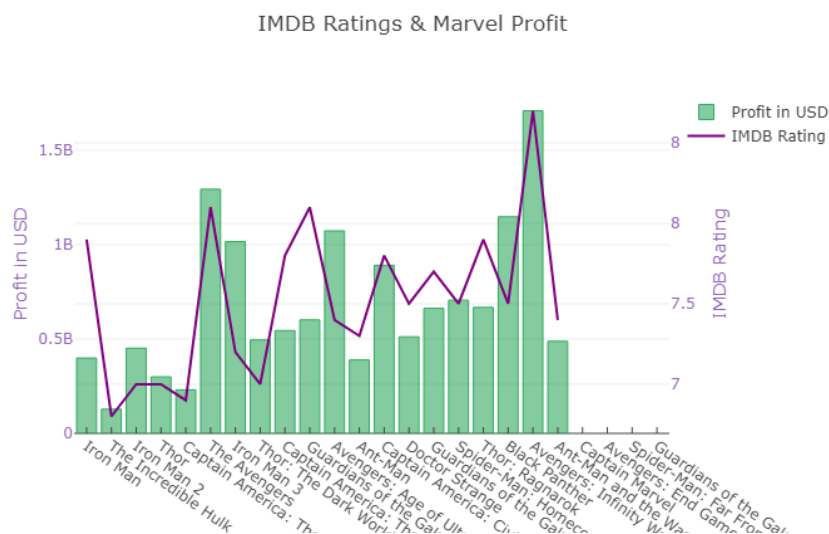


### ❖ Bar chart with Line Plot

Third Graph is a Double axis graph, where X represents Movie Titles and left sided Y-axis represents the movie profit earned in USD and the other Y-axis represents the IMDB movie ratings. A simple box plot works wonder with line graph overlapped which shows the audience popularity of a movie through earnings against critic reviews for the same. Simple color scheme has been used to represent bar graph with line graph. The Title names on the X axis have been slightly titled by an angle to fit the entire movie name.

**Visualization created using:**

Plotly using python on jupyter notebook



### ❖ WordCloud

Finally an extra graph just out of curiosity for trying out visualization ie a simple online-generated word cloud depicting the no. of appearances of celebrities in marvel movies. A small tribute to Stan Lee for his wonderful studio.

**Visualization created using:**

WordCloud using <https://www.jasondavies.com/wordcloud/>



## 4. Result

Following are that factors I considered while doing the visualisation:

- Use of multiple attributes in each graph to convey multiple information but cleanly and visually appealing with easy to understand.
- Use of colors to depict attributes.
- Proper labelling of Axes with Title info for every graph.
- Deciding the Scale for each axes.
- Providing dropdown where relevant.
- Legends to be displayed appropriately.

Following are challenges faced:

- Initially gathering the relevant dataset proved a challenge. I had to manually combine dataset searching online to create a dataset, which I wanted for myself.
- Tried my hands on various tools for visualization like Tableau, D3.js, Bokeh and finalized on plotly as it provided good visualization and had good support to fine tune various characteristics.
- Major issue which I faced while plotting scatter plot was to display legend markers. Tried various options but could not get to display the legend markers due to plotly limitations as the legend markers are proportionate to the area of circle plotted. Finally changed the slope size to make it visible but still is not proportionate.
- Second setback I faced was working with images. I had gathered movie logos to be used for visualization (present in dataset). Tried multiple libraries to fit the dataset. However, charts were not generated as planned with pixel issue or overlapping or size limitations. Failed to add images on hover or using images as bargraph plots.
- I created word cloud using wordcloud library in python but was not very satisfied with the outcome, hence dropped the plan.
- Also tried to create network-connected graph using D3.js and tableau for movie and cast list but faced technical difficulty in generating a meaningful graph, thus dropping those charts too.