Data Visualization and Exploration - Final Project

DC versus Marvel - A Data Visualization and Machine Learning Approach

Submitted By: Tanvi Sahay

Abstract:

In this project, I attempt to visualize differences and similarities between the two comic book franschisees DC and Marvel. The data has been obtained from FiveThirtyEight's GitHub page - Comic Characters and was inspired by their blog post - Comic Books Are Still Made By Men, For Men And About Men, although all analysis has been done independently. I also apply the clustring algorithm K-Means in order to cluster similar characters of both DC and Marvel together. The final visualizations can be found at his url: http://www-edlab.cs.umass.edu/~tsahay/. The purpose of this project is to present a story to the user about how the two biggest comic book giants compare with each other on different levels. All codes for running this webpage have been included in the zip file attached with this submission. Simply run Index.html for the final webpage.

About Data

The data chosen is an open source dataset provided by FiveThirtyEight, about comic book characters from two different universes: DC and Marvel. Each universe has a separate table in the dataset and both tables have the same attributes. The DC data has a total size of 1.1 MB, with 6897 data instances and 13 attributes. The Marvel data set has a total size of 2.4 MB and consists of 16,377 data instances and the same 13 attributes as DC. These common attributes are:

- *Page id:* The page id on which the character was first introduced. Type: numeric
- *Name:* Name of the character. Type: text
- *urlslug:* The url of the character. Type:text
- *ID:* Identity of the character. Type: Nominal - secret or public
- *Align:* Alignment of the character. Type: Nominal - good or bad
- *Eye:* Eye color of the character. Type: text
- *Hair:* Hair color of the character. Type: text
- *Sex:* Sex of the character. Type: Nominal - male or female
- *GSM:* Sexuality of the character. Type: Nominal - Bisxeual, homosexual, none
- *Alive:* Whether the character is alive or not. Type: Nominal - alive or deceased
- *Appearances:* Total number of appearances a character has had. Type: numeric
- *First Appearance:* The date and month of first appearance of the character. Type: dd-mm
- *Year:* The year of first appearance of the character. Type: year

Data Cleaning

The data had several missing values for both nominal and numeric data types which needed to be filled before the data could be used for visualization or analysis.
A small section of the tables being used has been provided in Table 1 and Table 2.

While numeric values could have been imputed by using the mean or median, doing the same for ordinal values might have drastically changed the trends and hence, the analysis. For this purpose, any tuple with a missing value was removed. This reduced the dataset to a total of 6525 rows, with 2123 rows for DC and 4402 rows for Marvel. For each of the two files, two extra columns were also added. One was the name of comic franchisee('dc' or 'marvel') and the other was Appearances of a character relative to all others in that franchisee : Relative Appearance = Absolute Appearance / Total Appearances by all characters. After addition of these attributes, the final data, both DC and Marvel combined was a total of 6562 x 15. This data was used for all further visualizations.

Table 1: Subset of Data for DC

page_id	name	align	eye	hair	sex
1422	Batman(Bruce Wayne)	Good Characters	Grey Eyes	Black Hair	Male Character
1380	Flash(Barry Allen)	Good Characters			Male Character
1448	Wonder Woman(Diana)	Good Character		Blond Hair	Female Characters
1451	Timothy Drake(New Earth)	Good Characters			Male Characters

Table 2: Subset of Data for Marvel

page_id	name	align	eye	hair	sex
1836	Human Torch	Good Characters	Blue Eyes	Blond Hair	Male Characters
1071	Thomas Halloway	Good Characters	Brown Eyes	White Hair	Male Character
1463	Dorma	Neutral Character	Blue Eyes	Auburn Hair	Female Characters
11798	Phineas Horton	Good Characters	Blue Eyes	Brown Hair	Male Characters

Interest in Data

The debate between which of the two comic book giants, DC and Marvel is better than the other has been going on for several decades, since Marvel's introduction of the Human Torch in 1939. In this project I try to take a statistical look at the data available to see how the two universes compare with each other. I feel it would be interesting to take an objective look at the differences between the two juggernauts, see how they have grown with time and how the characters have changed. The questions that I try to answer are:

1. How have the two franchises grown over time in terms of the number of characters added in each?
2. How do the two franchises compare in terms of their story compositionality and writing?
    - compositionality - How many good versus bad characters do each of the two franchises have?
    - writing - Do the good/bad characters occur more often towards the beginning of a comic or towards the end?
3. How do the two comic books differ in their character construction?
    - How has the sex ratio changed in comic books over time?
    - Have the comics seen an addition of more LGBT characters with changing time?
    - Which comic has more radical characters(varied hair color/eye color)?
    - Which comic has more characters with a secret identity?

Analytics

In this project, I have applied K-Means Clustering to cluster characters having similar attributes across DC and Marvel together into a single group. Basic analysis for the same has been added into the webpage. Since the clustering visualization is dynamic and K-Means clusters depende upon the initialiation of centroid, analysis may differ across multiple runs of the algorithm. Thus, only some commonly observed analysis have been added. The visualization has been adapted from - Visualizing K-Means algorithm with D3.js. A basic summary of K-Means Clustering can be given as follows:

K-means is an iterative optimization algorithm for clustering that alternates between two steps. It randomly initializes k cluster centroids in a vector space D, with N data cases. From there on it repeats the following two steps until convergence: a) For each data point x i in X, find the nearest centroid (cluster) to it (using Euclidean distance). Assign this data point to that cluster. b) For each of the clusters formed in step a), find the new centroids for each them, by finding out the mean in each of the clusters. Repeat step a)again.

Visualizations

In this project, I have prepared several visualiations, which answer the questions mentioned above either individually or when composed with another. The visualizations can be deided into two separate sections:

descriptive and exploratory.

- Descriptive -
  1. **Table** - This is a standard table showing the top 4 most frequently appearing marvel and DC characters sorted according to their number of appearances. The table is static.
  2. **Timeline** - This is a timeline showing the top 8 characters from both DC and Marvel, sorted according to their year of introduction. The characters included are the presently most famous ones from each franchisee. The timeline allows probing of each point and upon clicking, it takes the user to the url for that character, as spcified in the dataset and shown in the table above.
- Exploratory - The first set of visualizations is fully interactive, with a filter applied in each plot reflecting on every other plot. It contains these plots:
  1. **Pie Chart** - Showing the division of total number of characters between DC and Marvel.
  2. **Bar Charts** - One bar chart for DC and Marvel each showing what the per year addition of characters has been like.
  3. **Row Chart** - Showing the number of good, bad and neutral characters
  4. **Group Chart** - Showing the trend of additions of good, bad and neutral characters for each page range of 50,000 pages.

  The second visualization is a **series chart** between year and number of characters added, with each series colored on the basis of the gender of the character. This chart can be refined on the basis of which franchisee to choose. Viewers can either observe trends for both DC and Marvel, or for only DC or for only Marvel. This aspect has been provided by adding a drop down menu to select the data for visualization.

  The third set of visualizations are four **heatmaps**, each showing the trend of number of appearances for both DC and Marvel based on the character's eye color, hair color, sexual orientation and identity. Radio Buttons have been added to switch between absolute number of appearances and relative number of appearances.

Analysis for all the visualizations has been added in the webpage.

Design Layout

All designing has been done according to Chrome. However, the same has also been tested with Mozilla Firefox and Safari as well.