Analysis of Popular Comic Characters

Shaunak Sen

ss8n18@soton.ac.uk

# Data Story Summary

The topic I have chosen for the Data Story is **Analysis of Popular Comic Characters**.

Few questions I pose in my data story are:

1. How is the popularity of comic characters related to their Character Alignment and Sex?
2. How is the popularity of comic characters impacted by Character Alignment, Sex and Alive field (whether the character is alive or not) - Take a deeper look?
3. What are the factors which affect Popularity - Strength, Durability, Intelligence, and Speed?
4. Who are the most powerful comic characters? - Which of them are super villains?
5. How many of the most powerful comic characters are female?
6. How popular are the most powerful characters?
7. Which are some most common superpowers? - Which are some of the rarest superpowers?
8. Which characters apart from Spider Man can sense danger?

While posing these questions I have used narrative design patterns like **Flow, Engagement** and **Argumentation**. Questions are posed such that there is a **Gradual Reveal** of information. We start with a basic overview of a topic and explore the results until we reach our conclusions. **Exploration** is promoted by the use of visualizations, with which users can easily interact and derive their own insights and conclusions (e.g. Bubble Chart, Heat Map and Word Cloud used in the story all support user interaction). At many stages we **Compare** different characters based on various attributes.

# Dataset Summary

Here you should briefly summarise the data sources you used, describe where you retrieved them from, describe how you linked together multiple datasets (if applicable), and describe any processing you needed to do to the data.

## Data Sources

The Data Sources used are as follows:

### <https://github.com/fivethirtyeight/data/blob/master/comic-characters/marvel-wikia-data.csv>

### <https://www.kaggle.com/claudiodavi/superhero-set>

Additionally, I scraped information about each character in the first dataset from the website: <http://marvel.wikia.com/>. For each character, the superhero name and attribute scores like strength, speed, fighting skills, intelligence and energy projection were scraped.

## Data Preprocessing

In the first dataset, most of the data extracted was text based, and required a lot of pre processing, like removal of leading and trailing whitespaces, converting them to appropriate case, removing special characters etc. After cleaning, the dataset was filtered to get the top hundred most popular superheroes to be used for the analysis. The number of appearances of the character in comic books has been considered as the indicator to popularity of that character. Also, few of the most popular characters had missing values which I looked up from the website and filled in.

The second dataset did not have any missing values or garbage text data. However, the data was in a wide format. Each superpower had its own columns. The data was converted from this wide format to a long format, where a new column called “Super power possessed” and populated.

This process was done to help create the Word Cloud as well as the Heat Map visualizations.

# Visualisations

## Bar Chart to get an overview of how Character Alignment and Sex impact Popularity

[Insert screenshot]

### Description

The Bar Charts provide an overview of the popularity of the characters based on their Character Alignment and Sex. It plots each character on the X axis and the Appearances on the Y axis. Sex and Character Alignment are represented by colors. For each visualization, Tableau’s color-blind palette has been used.

### Justification

Discuss why the chosen visualisation is suitable for your narrative, and the data. Describe how the visualisation makes the data readable by the intended audience, avoids bias or confusion, and emphasises the key data. You should justify why (and how) you are using each visualisation with regard to concepts covered in the course; make sure you cite any relevant papers.

This visualization introduces the user to the Data Story and provides an overview of the data. The chart is also supported with some basic conclusions that emphasize key findings from the chart.

### Narrative Design Patterns

This visualization uses the **Gradual Reveal** design pattern in the sense that it provides a basic overview to the user, before going deep and exploring. It also makes use of the **Compare** narrative design pattern as it compares popularity of characters across different attributes.

### Strengths and Weaknesses

This chart displays a lot of data points. While necessary to provide an overview of the data, sometimes it might be a bit confusing for the user. This is why the text under the plot highlights the important conclusions to help the audience follow along with the story.

### Improvements

There is only one genderfluid character – Loki. If additional data about other genderfluid characters are found, the comparisons could be better.

## Stacked Bar Chart to explore how Character Alignment, Sex and Alive fields impact Popularity

### Description

Stacked Bar Charts show how the popularity of characters is impacted by character alignment and sex of the character. We plot average number of appearances on the Y axis and Character Alignment on the X axis. The colors represent sex and alive (whether the character is deceased or living) fields respectively.

### Justification

The first chart provided an overview to the user. This one explores the topic and also looks at another dimension – alive field to understand the data. The key data, i.e. there are no popular female villains is highlighted to the user.

### Narrative Design Patterns

This visualization uses the **Gradual Reveal** design pattern in the sense that it picks up from where the previous chart left off and explores the question. It also makes use of the **Compare** narrative design pattern as it compares popularity of characters across different attributes. Additionally, it promotes **Exploration** by providing various filters by which users can interact with the chart.

### Strengths and Weaknesses

The key point is highlighted in the chart. Also conclusions are mentioned as text below the chart to aid the user’s understanding

### Improvements

More data about genderfluid characters would result in a better analysis.

## Scatter Plots to Identify factors that impact Popularity – Visualizations 3.1 and 3.2

### Description

The Scatter Plots plot the number of appearances on the Y axis and different factors like Strength, Intelligence, Fighting Skills and Energy Projection scores on the X axis. Also it makes use of the size of each circle to indicate a third variable like Durability and Speed. Trend lines are fitted to each Scatter plot, which not only help the user see the overall trend, but provide additional statistical measures like R squared value and p value to test the significance of the relationship.

### Justification

Scatter Plots are necessary to address our third question, which tries to find out the factors that affect popularity. Key points are highlighted on the plots, calling out user’s attention.

### Narrative Design Patterns

This visualization supports the **Humans-behind-the-dots** Narrative Design Pattern. Users can hover over the circles (dots) and see which character it belongs to.

### Strengths and Weaknesses

Scatter plots tend to be very good to understand if there is relationship between two variables. Trend lines further support that. However, when the X axis is discrete, overlapping of data points occurs [2]. Fortunately, this problem is not so severe for our data, and in our context, the trend of the data is more important than the individual data points. To fix the problem, the fill of the mark has been removed [2] and some overlapping data points have been manually removed.

### Improvements

The highlighted points in the plot helps draw user’s attention to key features and data points. As user clicks on the text, a feature that would highlight the relevant data points would be very user friendly.

## Bubble Chart to Visualize the Most Powerful Heros and Villains and the Most Powerful Male and Female Characters

### Description

Strength is plotted on the X axis and Speed is plotted on the Y axis. In the first plot, the color of the circles represent Character Alignment, while in the second plots the color represents Sex of the characters The size of the circles represents Intelligence Score. To prevent overlapping and to observe the clusters, a random noise (jitter) has been added to the strength and speed scores. However, the tooltip that is generated when a user hovers over a point represents the correct value for the Strength and Speed. Also, users can use the filter to see where their favorite superheroes fall on the plot.

### Justification

Bubble Charts are very useful to visualize a cluster of data points and explore the relationship between them. Towards the right-upper part of the plots, we can see a cluster of data points which are large in area. These are the characters who are the most powerful, i.e. have high Strength, Speed and Intelligence. This helps us understand the questions 4 and 5 posed earlier.

### Narrative Design Patterns

This visualization supports the **Humans-behind-the-dots** Narrative Design Pattern. Users can hover over the circles (dots) and see which character it belongs to. Additionally, it promotes **Exploration** by providing various filters by which users can interact with the chart. I have also used the **Familiarisation** Design Pattern. Wherever possible I have used examples of popular characters that have appeared in movies like Thanos, Thor, Rogue and All-Father Odin, so that the user can relate to them.

### Strengths and Weaknesses

As discussed for Scatter Plots, Bubble Charts also suffer from the problem of overplotting. However, with the use of jitter, removing the fill of our data points and reducing the opacity solves this problem to a great deal. One of the strengths of this plot is that it not only helps us answer posed questions, but also helps in discovering interesting facts. For e.g. an interesting observation from this plot is there are very few female characters who are highly intelligent. Thus the comic world may be gender biased and we kind of intuitively know that the most popular intelligent characters like Tony Stark, Dr. Bruce Banners, Hank Pym are all males.

### Improvements

The highlighted points in the plot helps draw user’s attention to key features and data points. As user clicks on the text, a feature that would highlight the relevant data points would be very user friendly.

## Packed Bubble Chart to explore how Popular the most Powerful characters are

### Description

We already found who the most powerful (based on Strength, Speed and Intelligence) are. We use the packed bubble chart to compare the popularities of these most powerful characters. The size represents the popularity (number of appearances), while the color represents the sex of the character.

### Justification

This chart helps us easily answer the sixth question and also highlights an interesting finding: Rogue is very popular and is the only female character among the most powerful ones.

### Narrative Design Patterns

This visualization uses the **Gradual Reveal** design pattern in the sense that it picks up from where the previous chart left off and explores the question. It also makes use of the **Compare** narrative design pattern as it compares popularity among the most powerful characters

### Strengths and Weaknesses

The human eye does not discern size(area) very well. The sizes of the data points of Nova, Thanos, The Watcher and Galactus all seem similar. However this problem is addressed by the use of labels in the plot. Also, since we are using only six data points, the plot is appropriate. But this plot will not scale well if we have a larger number of data points

### Improvements

If we have a lot of data points, to improve scalability other alternatives like bar chart or a stacked row chart might be more appropriate.

## Heat Map

### Description

The Heat Map helps to visualize the attribute scores of the top fifty most popular comic characters. The data is sorted by the sum of their scores (descending)

### Justification

In the previous plots, while answering who the most powerful characters were, we chose strength, speed and intelligence. In this plot, we consider fighting skills, durability and energy projection scores as well. The most powerful Marvel character is a female character (Rogue). This is a very interesting finding.

### Narrative Design Patterns

This visualization uses the **Gradual Reveal** design pattern in the sense that it picks up from where the previous chart left off and explores the question further. It also makes use of the **Compare** narrative design pattern as it compares attribute scores among the most popular characters. Additionally, it promotes **Exploration** by providing various filters by which users can interact with the chart.

### Strengths and Weaknesses

The Heat Map can compare characters across all the attribute scores, which was a limitation of the Bubble Chart. However, for the Heat Map, data is required in a long format in Tableau. This causes data duplication and additional data processing overheads.

### Improvements

This plot selects the top 50 most popular characters and compares the attribute scores among them. The option of the number of characters to select (instead of 50) can be given to the user to promote better user interaction.

## Word Cloud – Visualize the most common and rare superpowers and the charcaters that possess them

### Description

The Word Cloud uses the second dataset and visualizes the most frequently occurring superpowers with greater size. On clicking on a particular power, the characters possessing that power will be populated below

### Justification

This plot promotes user engagement and helps us answer our last two questions posed above.

### Narrative Design Patterns

This visual promotes **Exploration** as users can search for a power to highlight in the cloud. On clicking on a particular power, the characters possessing that power will be populated below. This also promotes the **Users-find-themselves** narrative as the users can explore this visual and find their own insights

### Strengths and Weaknesses

Word Clouds are often criticized because the visibility changes if the length of the word varies. However, it is quite easy to create and engaging.

### Improvements

To solve the problem of different word lengths, the smaller length word has been padded with spaces and the longer words have been abbreviated, wherever possible. This leads to better visualization.

[Insert image]

# Conclusion

The Data Story answers all the questions addressed above. It also finds some interesting observations along the way. Additionally, users have the scope of interacting with many of the plots and deriving their own insights.

##### References

1. B. Bach, M. Stefaner, J. Boy, S. Drucker, L. Bartram, J. Wood, P. Ciuccarelli, Y. Engehardt, U. Köppen, and B Tversky. “Narrative design patterns for data-driven storytelling.” In Data-Driven Storytelling, N. H. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale, Eds. CRC Press, USA, 2018, ch. 5, pp. 107–134.

[2] https://visual.ly/blog/the-pros-and-cons-of-scatterplots/

[3] https://www.keatext.ai/en/blog/artificial-intelligence/3-strengths-and-3-weaknesses-of-word-clouds/