# Visualizing Superhero Attributes: An Analytical Comparison of Powers and Characteristics

Team Name: CESR CIPHER

Team Members:
Eric Li (eli1@college.harvard.edu)
Rohan Naidoo (rohannaidoo@college.harvard.edu)
Shahmir Aziz (shahmiraziz@college.harvard.edu)
Chris Canzano (ccanzano@college.harvard.edu)

## Background and Motivation

Through their extraordinary abilities and compelling stories, superheroes have captivated, and inspired awe and ambition in, audiences for generations. The impact they have had on modern art, as well as how the world views leadership and virtue, is immense.
Our motivation for this project hence stems from a shared interest in both data analytics and the superhero genre. By analyzing and visualizing superhero attributes, we aim to uncover patterns, correlations, and insights that highlight how different characters compare and contrast.
Moreover, we might also be able to match characters to users based on their inputs. We could also compare inter-fandom characters to see who are most similar to each other, and whether they represent similar ideologies.
This project allows us to apply our learnings in data visualization to a fun and engaging topic.

## Related Work

Our project is inspired by:

The Superhero Database: An online resource that catalogs detailed profiles of superheroes, showcasing the depth of available data. It can be found here: https://www.superherodb.com/

The Marvel and DC Wikis: Although these sites do not contain advanced visualizations, they contain a wealth of info related to superheroes from both franchises. These databases and articles are maintained by loving fans (Eric has made contributions before).
https://marvel.fandom.com/wiki/Marvel_Database

Class Discussions: Specifically on interactive data visualizations and the use of multidimensional data to tell a story.

Mostly though, we are inspired by our childhood love of superhero movies, comics, and animes. Eric in particular is a huge Spider-man fan, and is looking forward to comparing him to other superheroes.

**Data**

Using this dataset (https://www.kaggle.com/datasets/shreyasur965/super-heroes-dataset/data) and similar ones, we will sort and visualize hundreds of superheroes from various comic universes. The attributes can include things like:

- Superhero's alias or code name.
- Intelligence, strength, speed, durability, power, combat: Numerical representations of the character's abilities.
- Full-name, alter-egos, aliases: Real names and alternate identities.
- Place-of-birth, first-appearance, publisher, alignment: Biographical and affiliation details.
- Gender, race, height, weight, eye-color, hair-color: Physical characteristics.
- Occupation, base, group-affiliation, relatives: Professional and relational information.

We can group them by team affiliations, power, comic universe etc and compare each superhero against each other to see who excels in what category.

**Data Cleanup**

The planned data cleanup will involve:

- Handling Missing Values: We'll address incomplete entries by either filling in missing data or excluding certain records if needed.
- Standardizing Measurements: Converting all height and weight data to the same units (e.g., centimeters and kilograms) for consistency.
- Normalizing Numerical Attributes: Scaling scores to a uniform range, especially if they differ in scale.
- Consistent Formatting: Ensuring uniform text formatting (such as standardized capitalization and removal of special characters).
- Categorizing Variables: Simplifying analysis by grouping races and alignments into broader categories.
- Duplicate Removal: Detecting and removing duplicate entries to maintain data integrity.

Proposed Visualizations

We plan to create at least five distinct visualizations:

1. Superhero Abilities Radar Chart:
Type: Interactive Radar Chart
Description: Displays the ability scores (intelligence, strength, speed, durability, power, combat) for selected superheroes on a radial graph.
Purpose: To compare the strengths and weaknesses of different characters visually, allowing users to select and contrast multiple superheroes at once.

2. Abilities Distribution Histograms:
Type: Histogram
Description: Shows the distribution of each ability score across all superheroes. Users can toggle between different abilities to see how common certain levels are.
Purpose: To understand how abilities are distributed among superheroes, identifying if certain powers are rare or common.

3. Correlation Matrix Heatmap:
Type: Heatmap
Description: Illustrates the correlation coefficients between different numerical attributes, such as intelligence vs. combat skills.
Purpose: To reveal relationships between abilities, indicating whether superheroes who are strong are also likely to be durable, for example.

4. Alignment and Publisher Breakdown:
Type: Stacked Bar Chart
Description: Represents the count of superheroes by alignment (good, evil, neutral) within each publisher.
Purpose: To compare how different publishers distribute their characters across moral alignments, highlighting trends or biases.

5. Interactive Network Graph of Group Affiliations:
Type: Network Graph
Description: Nodes represent superheroes, and edges represent shared group affiliations. The graph is interactive, allowing users to explore connections between characters.
Purpose: To visualize the complex relationships and team dynamics within the superhero universes.

6. Geographical Map of Birthplaces:

Type: World Map Visualization
Description: Plots the places of birth of superheroes on an interactive map. Clicking on a location provides a list of characters from that area.
Purpose: To explore the geographic origins of superheroes, identifying hotspots of character birthplaces and any regional patterns.

7. Demographics Pie Charts:
Type: Pie Chart
Description: Displays the proportions of gender, race, and species among superheroes.
Purpose: To assess diversity within the superhero population, offering insights into representation in comic book characters.

---

**Week 9 | Final Project Map**

**Group Discussion:**

1. Audience Options:
    - Comic Enthusiasts and Fans: People who are passionate about comic book characters and narratives across various universes.
    - Data Science and Machine Learning Practitioners: Individuals interested in utilizing creative datasets for predictive modeling and classification, particularly those who might want to apply machine learning techniques to analyze comic book data.
    - Visual Analytics Educators and Students: Professors, teachers, and students interested in using this dataset as a unique example in data visualization or computer science courses, especially for exploring data preprocessing, exploratory data analysis, and visual storytelling techniques.

Chosen Audience: *Comic Enthusiasts*

2. Audience Profile:

Comic enthusiasts and fans are deeply invested in the lore, relationships, and power dynamics of characters across various comic universes. They are likely to have an in-depth knowledge of character backstories, affiliations, and iconic moments, but their familiarity with data analysis and visualization may vary widely. They might be comfortable with simple bar charts, pie charts, and character comparisons but could find complex statistical visualizations challenging. For this audience, information would be best presented at an accessible, story-driven level, using interactive elements and clear visual storytelling to draw connections between characters, explore character strengths and weaknesses, or highlight historical trends in character creation. Visualizations should be engaging, emphasizing comparisons or insights into attributes like power stats, alignments, and group affiliations rather than technical details.

3. Interesting Questions for the Audience:
    1. What are the primary characteristics that differentiate heroes from villains? Which power stats or physical attributes are significant indicators of alignment?
    2. How do intelligence, strength, and speed correlate with one another across characters? Are there trade-offs between these attributes?
    3. Are there any discernible patterns or trends in the physical attributes (height, weight) of characters with certain power levels?

4. Which publishers dominate certain character archetypes or power levels? Do some publishers tend to create characters with similar attributes?
5. What are the most common affiliations among characters? Are heroes or villains more likely to be affiliated with larger groups?
6. How do character attributes differ across comic universes? Are there any noticeable "signature" characteristics by universe?
7. What role does gender play in the power stats or physical characteristics of characters? Are there disparities in representation?
8. Are specific character attributes (e.g., intelligence, power) more prevalent among “neutral” characters than heroes or villains?
9. How do character traits correlate with their first appearance years? Are older characters more likely to have certain attributes? What has changed in heroes over the years?
10. How consistent are physical attributes (like height or weight) among characters with the same alignment? Do villains have a trend in stature compared to heroes?
11. Is there a relationship between combat skill and group affiliation? Are characters with high combat scores more or less likely to be part of a group?

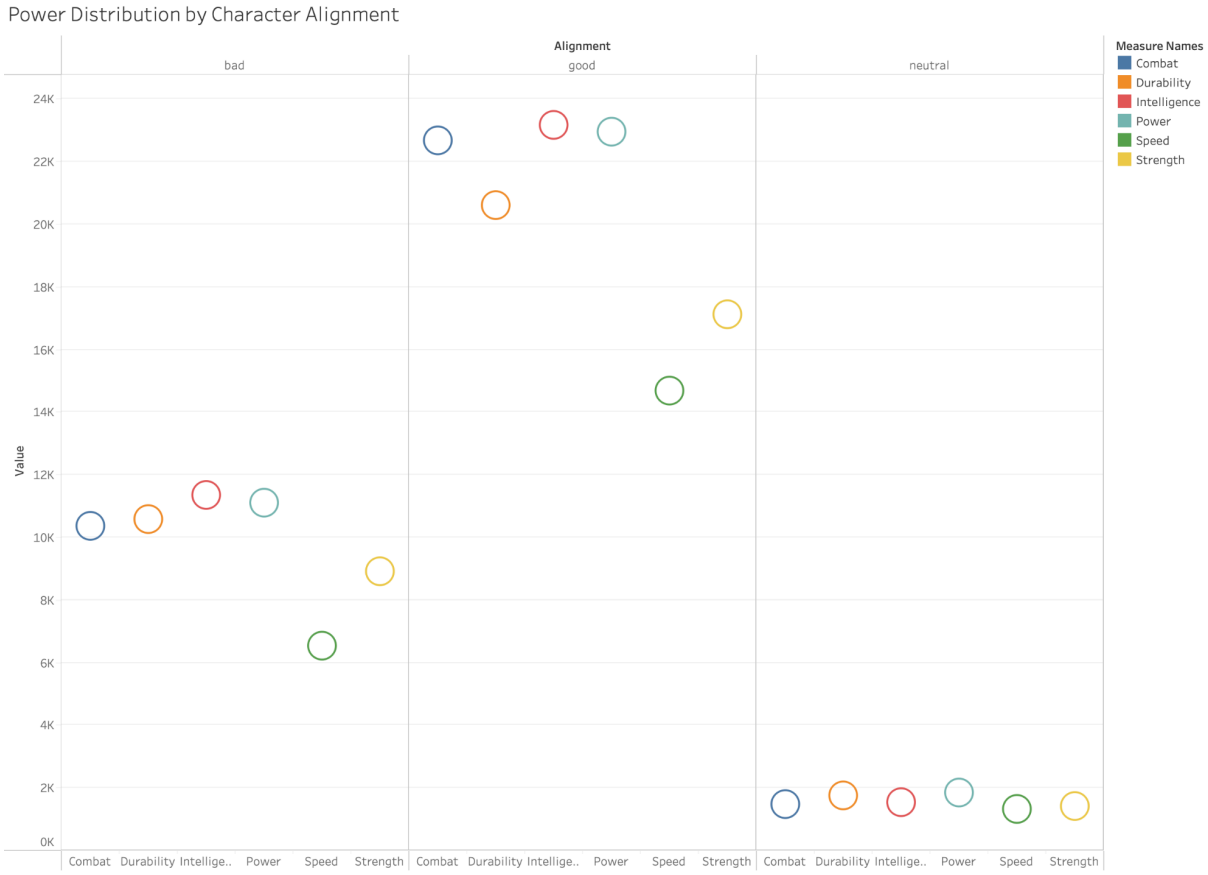
4. Dataset Overview:

- id: Unique identifier for each character (Quantitative).
- name: Name of the character (Categorical).
- intelligence, strength, speed, durability, power, combat: Core power stats represented as numerical values (Quantitative).
- full-name: The full name of the character, which may reveal additional biographical detail (Categorical).
- alter-egos: List of alter egos associated with the character (Categorical).
- aliases: Other names or aliases the character is known by (Categorical).
- place-of-birth: Location where the character originated or was first introduced (Categorical).
- first-appearance: The comic or year of the character’s debut (Categorical, potentially Ordinal if treated chronologically).
- publisher: The publishing company for the character (Categorical).
- alignment: Classification of the character as good, bad, or neutral (Categorical).
- gender: Gender of the character (Categorical).
- race: The race/species of the character, such as human or alien (Categorical).
- height and weight: Represented both as original strings and numeric values in centimeters and kilograms (Quantitative).
- eye-color and hair-color: Colors associated with the character’s appearance (Categorical).
- occupation: The character’s profession or role within their universe (Categorical).
- base: Character’s base or headquarters location (Categorical).
- group-affiliation: Affiliations or groups the character is associated with (Categorical).

- relatives: Known family or close relations, which could help explore connections across characters (Categorical).
- url: Link to the character's image (Categorical).

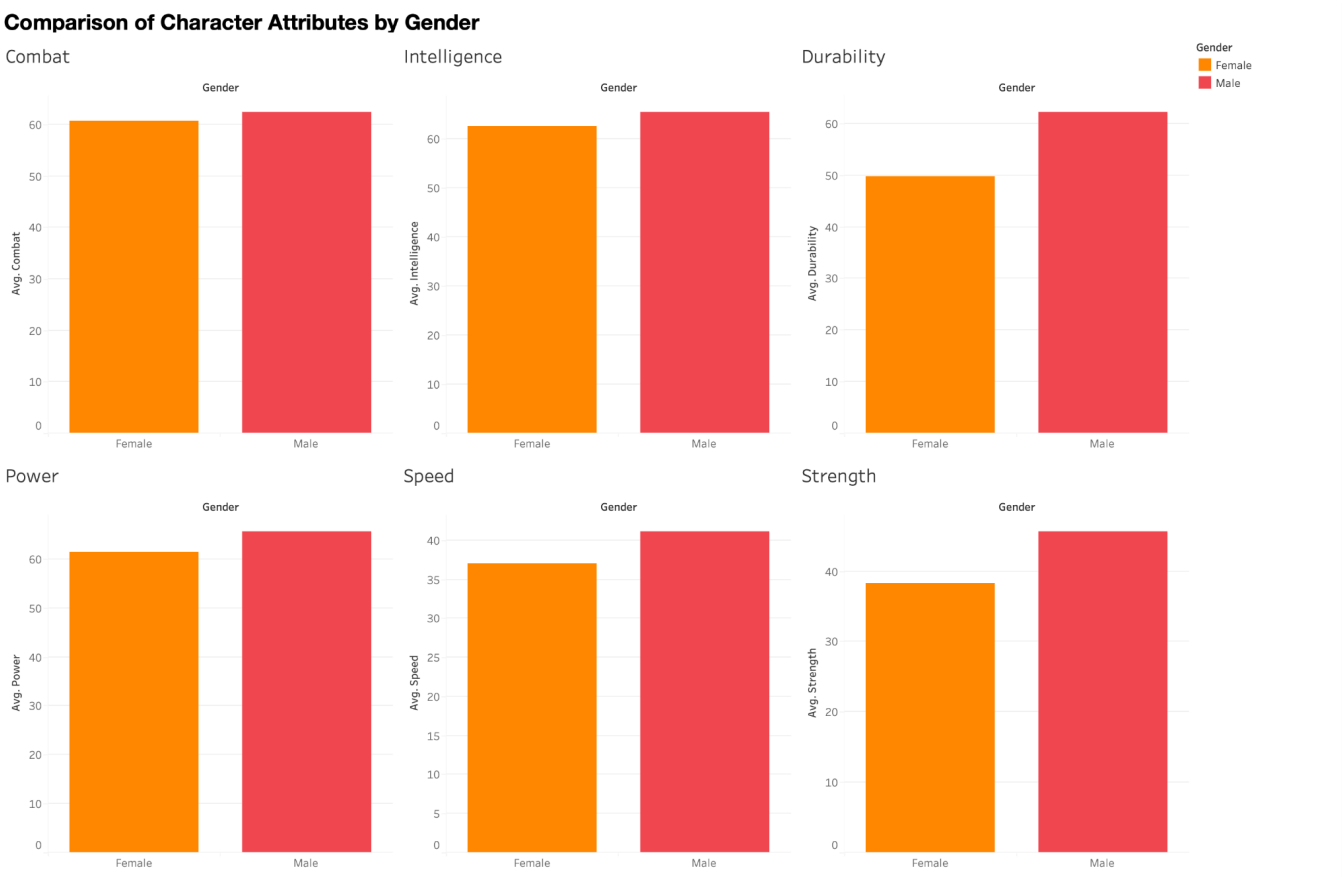
This dataset primarily consists of categorical and quantitative data, offering many possibilities for statistical and machine learning explorations, particularly around classification and predictive modeling tasks.

– Chris Canzano



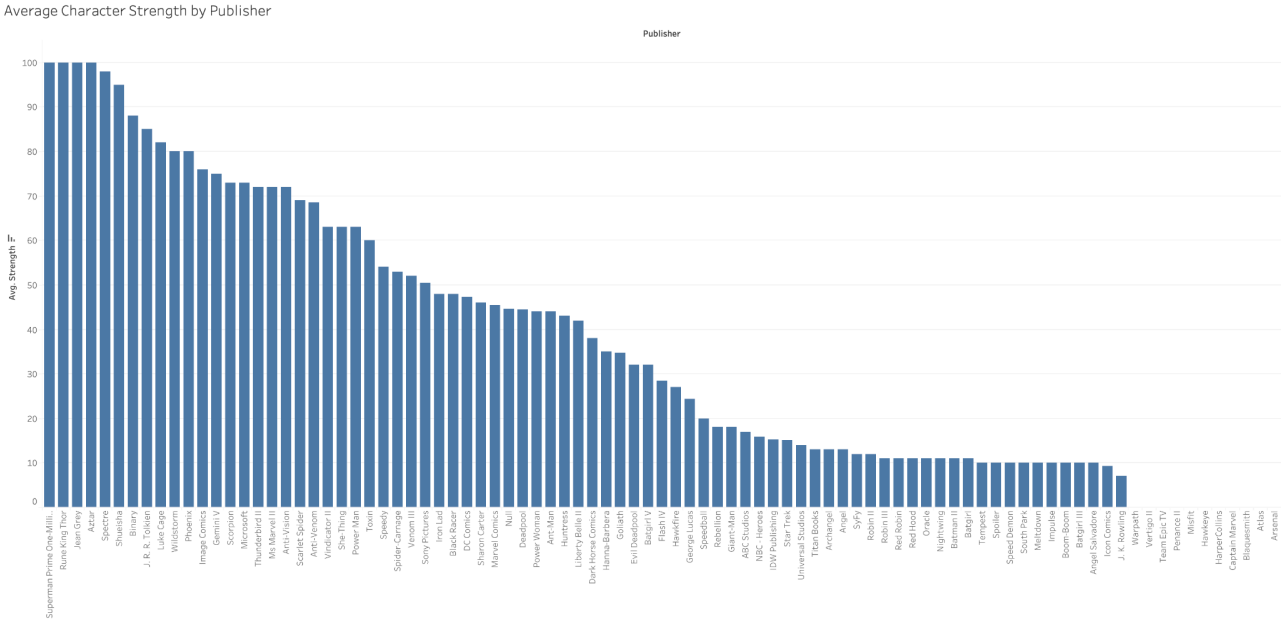

- The visualization provided above, titled "Power Distribution by Character Alignment," offers a high-level overview of how various power stats (such as combat, durability, intelligence, power, speed, and strength) are distributed across character alignments: "bad," "good," and "neutral." This approach aligns with some of the original questions, particularly those that explore the primary characteristics differentiating heroes from villains and the prevalence of specific attributes among neutrals, heroes, and villains. However, it diverges from questions focused on relationships or correlations between

attributes (such as intelligence, strength, and speed), physical traits (height and weight), and attributes across publishers or universes. The visualization's focus on alignment-based aggregation provides a simplified, accessible view that answers more categorical questions rather than complex, multivariable analyses. Although it offers insight into alignment-based distributions, it is limited for questions involving nuanced inter-attribute relationships, temporal patterns, or group affiliations, which would require additional cross-sectional or longitudinal data representations. This difference reflects a design decision in Tableau to address character alignment distributions rather than delve into the granular, multifaceted questions initially proposed, favoring clarity over depth.



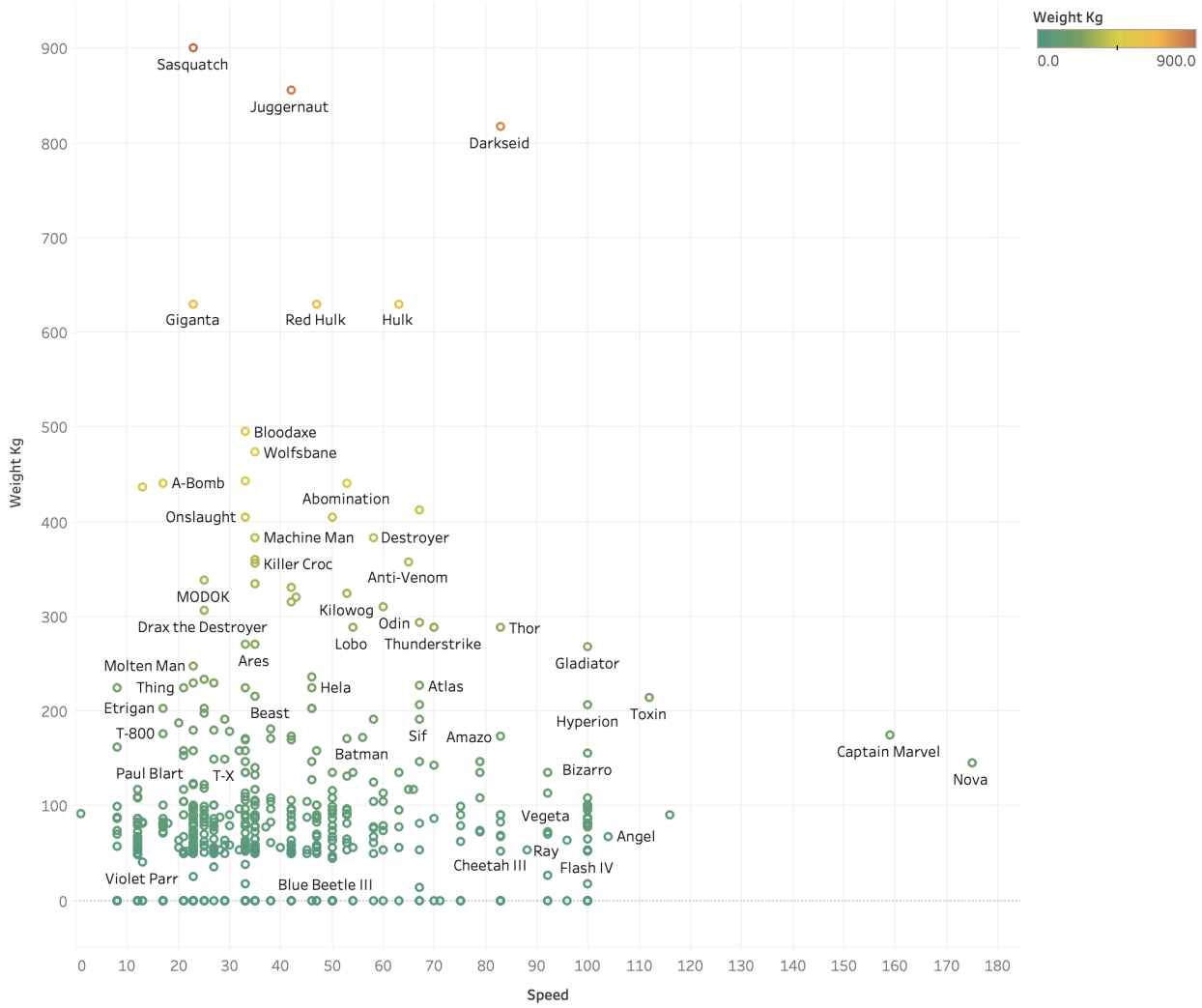

- The visualization above ("Comparison of Character Attributes by Gender") explores the average values of six key power stats (combat, intelligence, durability, power, speed, and strength) across genders. This aligns with some of our original questions, specifically those regarding gender disparities in power stats and physical characteristics. The visualization offers a straightforward gender-based comparison, revealing subtle but consistent trends where male characters generally have slightly higher averages across most attributes. However, it does not address more complex questions such as correlations between attributes, trends over time, or cross-universe comparisons. These

initial questions were designed to be exploratory and open-ended, covering a broader range of possible inquiries. In contrast, this visualization narrows the focus to a single demographic aspect, which is beneficial for examining gender disparities directly but lacks the depth required for multi-dimensional analyses. This approach was chosen likely for simplicity and clarity, aiming to answer one question well rather than diluting focus across multiple complex inquiries. This visualization is valuable for highlighting gender trends but could be complemented with other visuals to address our broader exploratory questions.



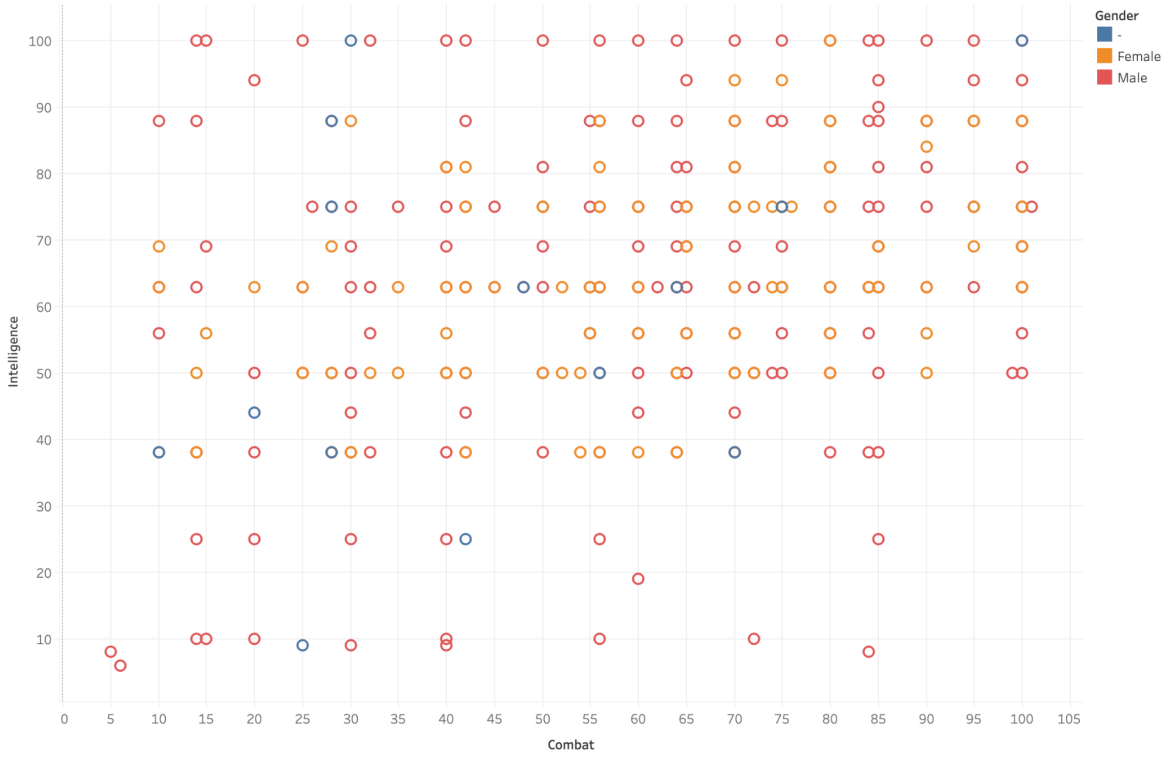

- The "Average Character Strength by Publisher" visualization presented here directly addresses one of our original questions regarding whether certain publishers create characters with specific attributes. By showcasing the average strength of characters by publisher, this visualization offers insight into which publishers emphasize strength in their character designs, possibly reflecting distinct creative philosophies or audience targeting strategies. However, it diverges from questions that seek to explore multi-dimensional relationships between attributes, such as correlations between intelligence, speed, and strength, or alignment-based differences. Focusing on a single attribute within publishers provides a straightforward view but lacks the depth of a multi-variable analysis. This choice suggests a strategic narrowing of scope, emphasizing clarity in presenting publisher-based trends over the complexity of inter-attribute relationships. While the original questions aimed for broad exploration, this visualization offers a concise answer to a specific facet, which may be more interpretable for audiences interested in publisher characteristics. Consequently, both approaches have merit; however, this targeted visualization simplifies storytelling and supports quick insights.
- Eric:

## Comparison by Weight (Kg) and Speed

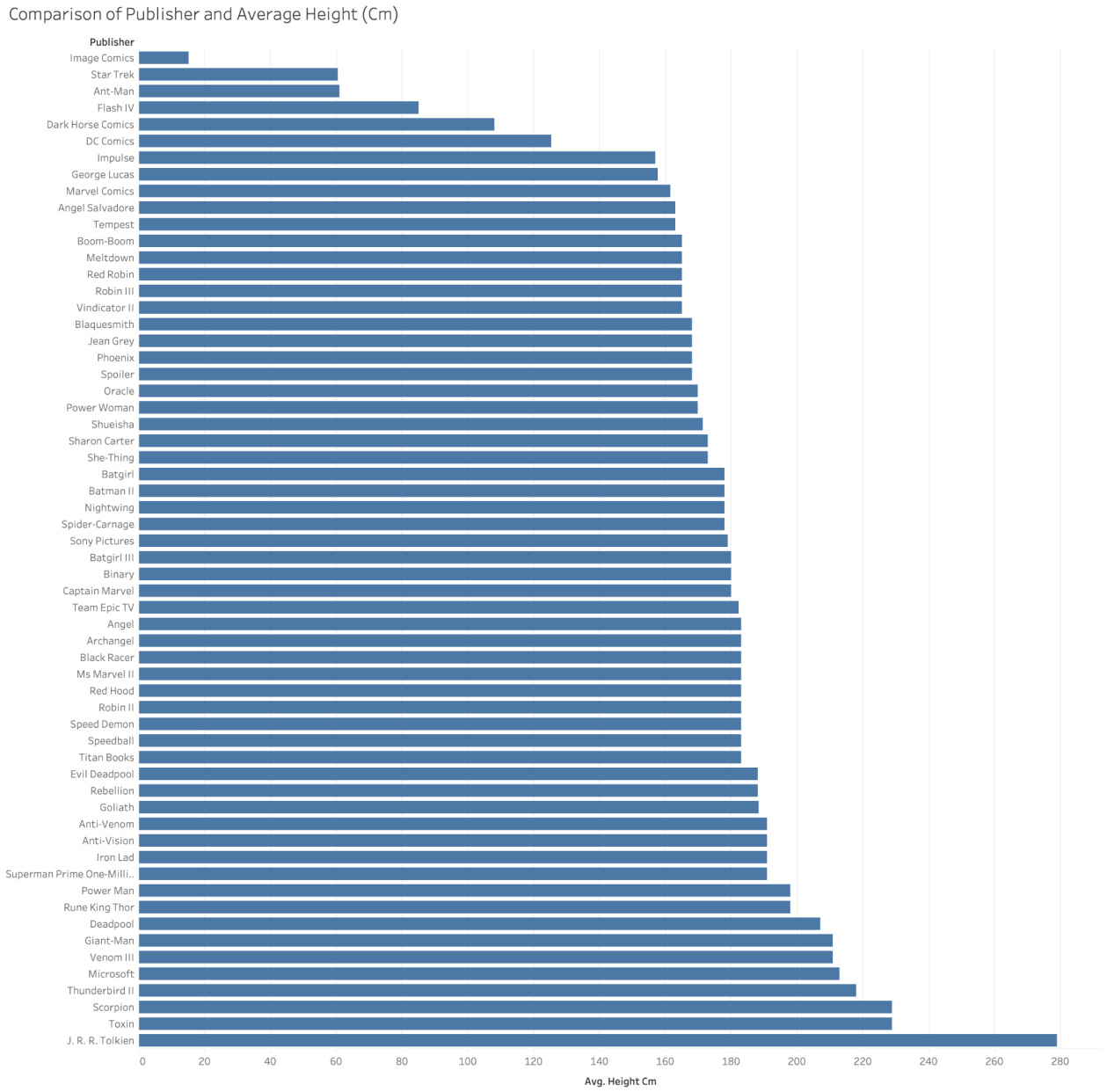



Sum of Speed vs. sum of Weight Kg. Color shows sum of Weight Kg. The marks are labeled by Name. The view is filtered on sum of Weight Kg, which ranges from 0.0 to 900.0.

## Combat vs Intelligence (Gender by Colour)



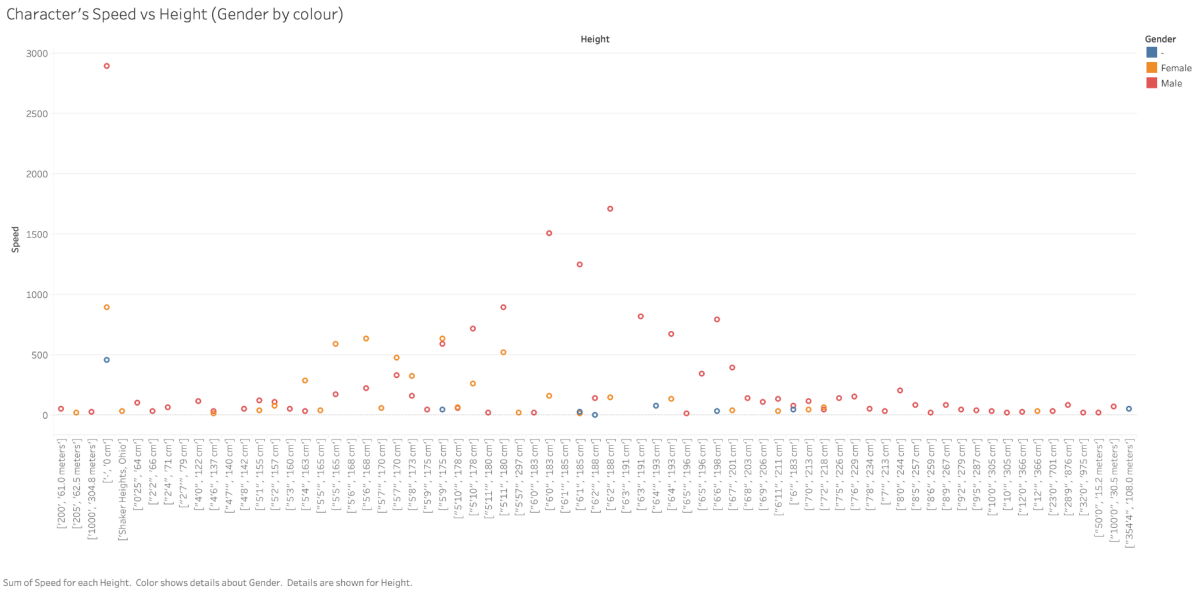

Combat vs. Intelligence. Color shows details about Gender.



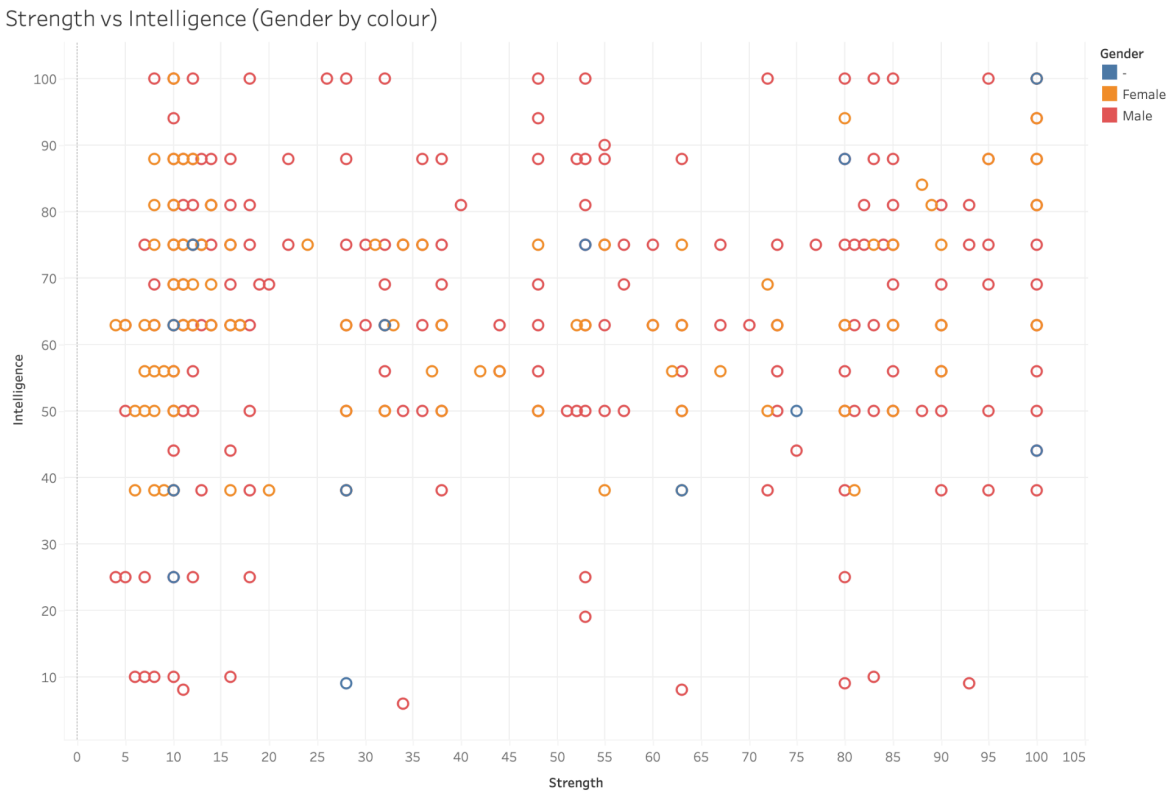

The three visualizations —comparing characters by weight and speed, analyzing combat versus intelligence by gender, and examining average height by publisher—align with several of our original questions but take a more targeted approach. Each visualization addresses a distinct dimension: the first explores the relationship between weight and speed, providing insight into physical attributes and their variation among characters; the second focuses on gender differences in combat and intelligence, partially addressing questions on gender-based disparities; and the third compares publishers based on average character height, revealing trends that may hint at "signature" physical attributes favored by certain publishers. While these visualizations answer narrower, attribute-specific questions, they diverge from more complex,

multifaceted inquiries like correlations across multiple attributes or trends over time. By sticking to focused visualizations, we gain clear, interpretable insights into specific questions, though at the cost of not fully addressing broader, more integrative queries. This selective approach is beneficial for clarity and audience engagement but may require additional, layered visualizations to holistically address the overarching questions initially posed.

- Rohan:

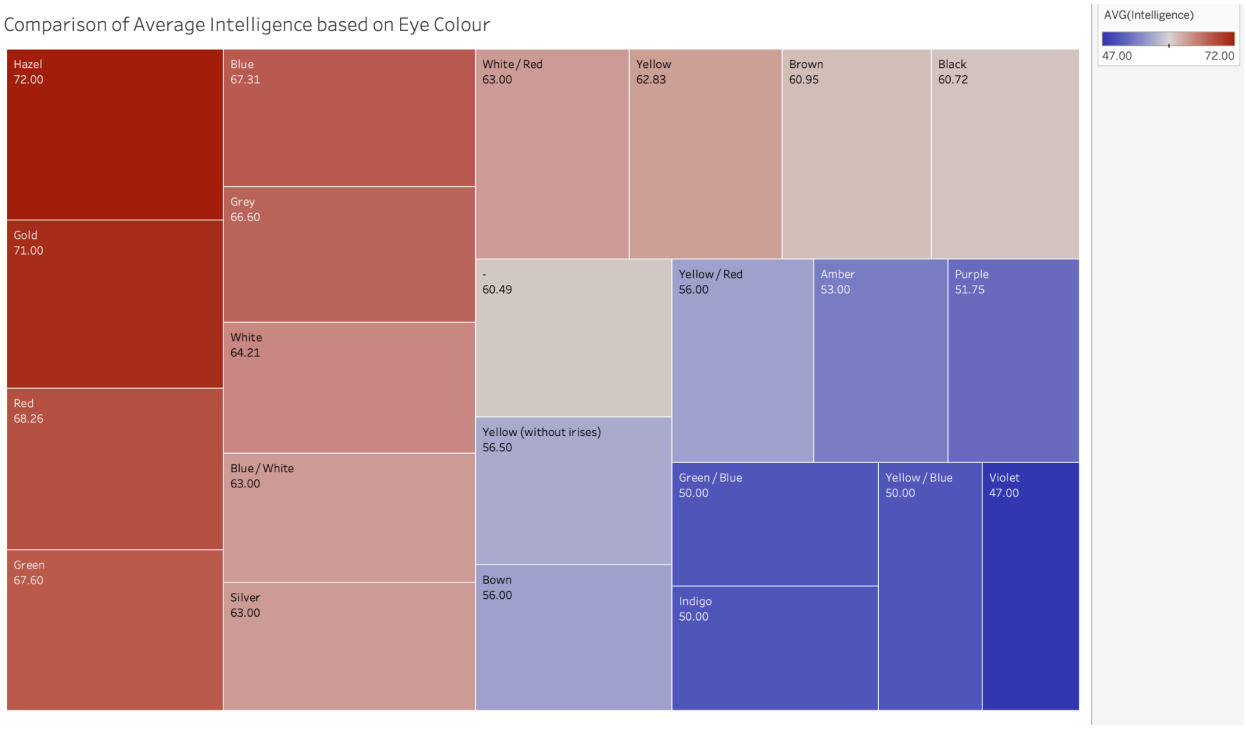



This visualization specifically answers the 3rd question we asked as a group: how do physical attributes affect certain physical abilities. In making this graph, I realized that we will need to do significant data cleaning, as quite a few null value errors popped up. The graph interestingly shows that height and speed seem to follow what looks quite similar to a normal distribution. While I wasn't initially expecting this, the more I think about it, the more it makes sense.




Strength vs. Intelligence. Color shows details about Gender.

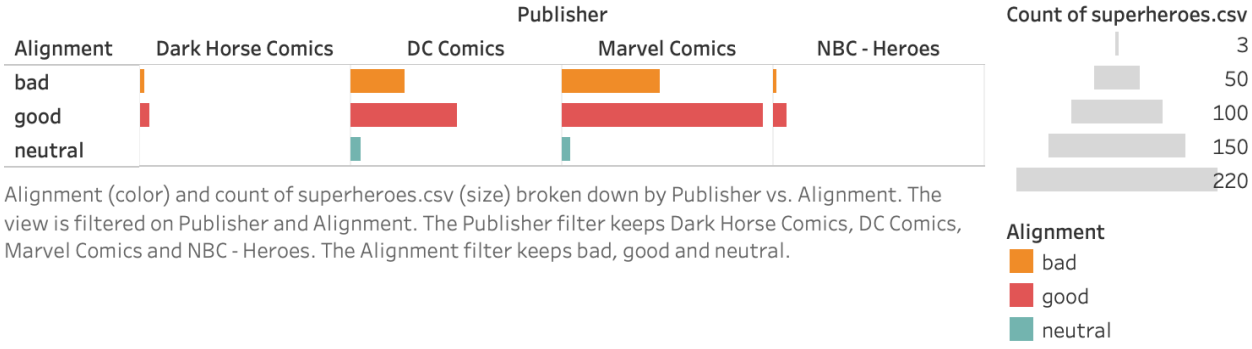
This visualization specifically answers the 2nd question we asked: are their tradeoffs between intelligence and physical abilities. Here, the results are actually quite surprising, as no clear pattern appeared. I was expecting an inverse relationship, but I guess these are superheroes after all! I was quite pleasantly surprised to see that there doesn't appear to be a major gender gap. I am curious whether a pattern may appear within specific universes/publishers.




This visualization doesn't specifically answer any of the questions we posed earlier, but I was curious to see if there may be any biases among common eye colors. Interestingly, characters with brown and black eye colors have low intelligence levels for such common eye colors. I think we should further explore where some of these biases may exist perhaps by race and gender.

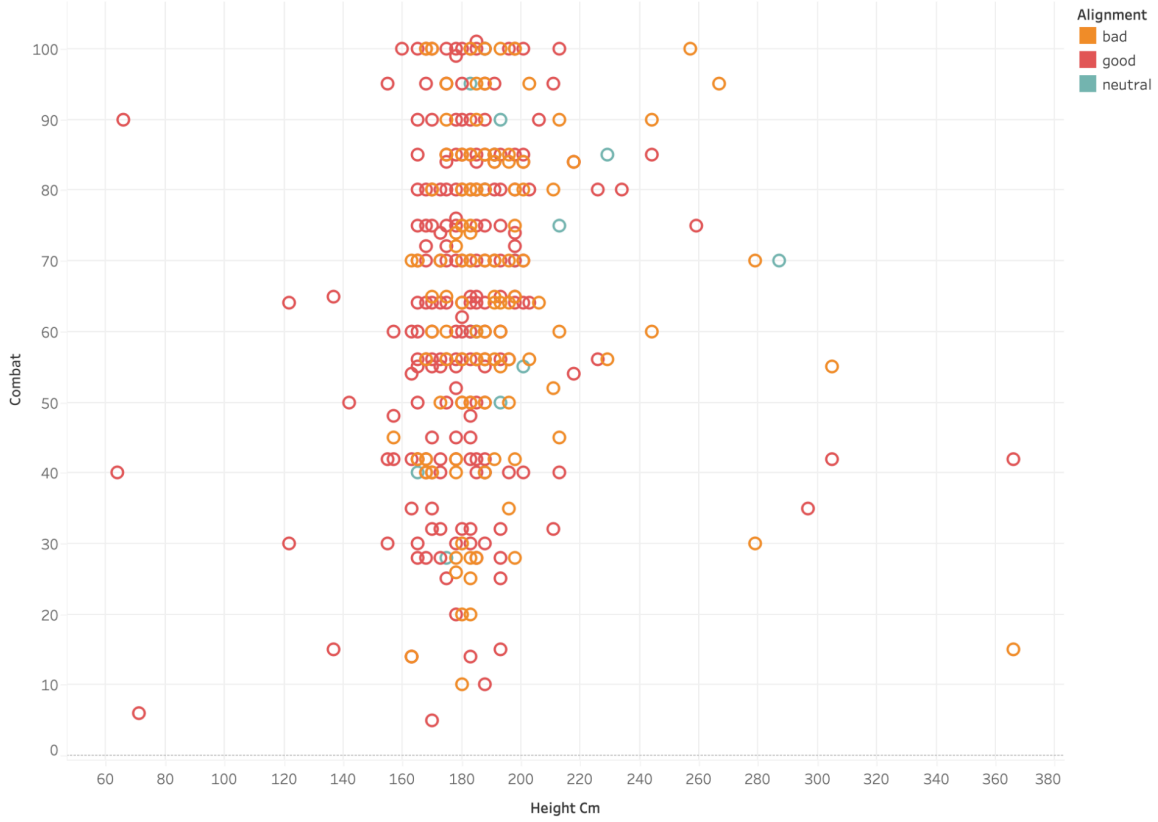
- Shahmir

## Distribution of Heroes vs Villians By Publisher



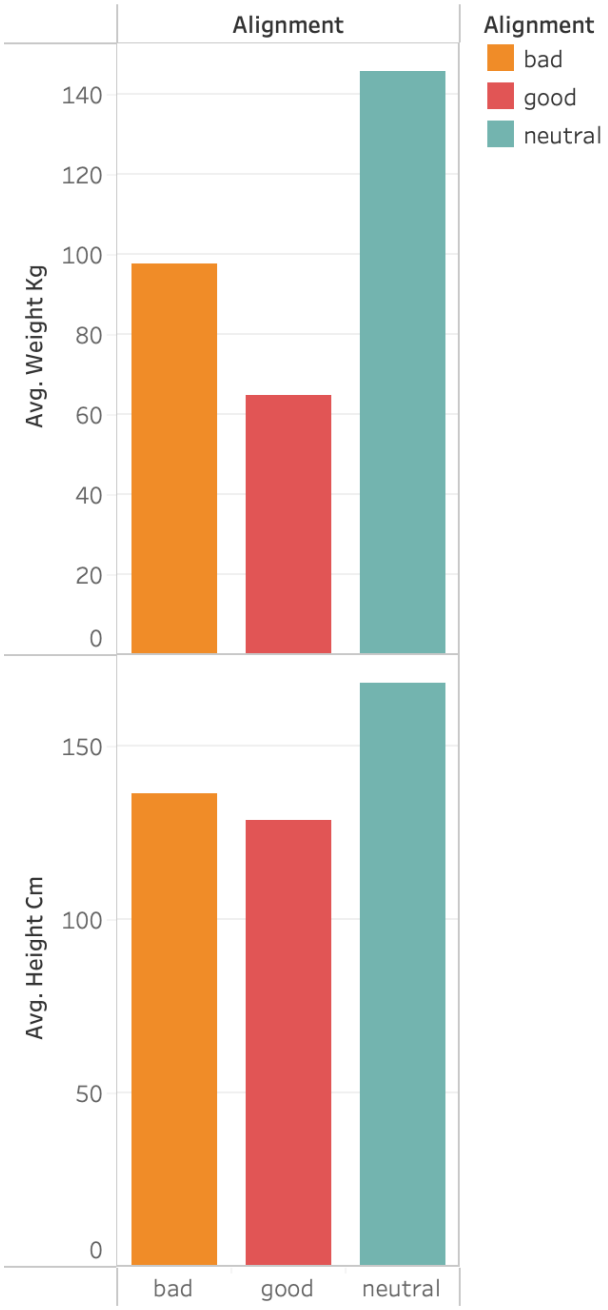

Alignment (color) and count of superheroes.csv (size) broken down by Publisher vs. Alignment. The view is filtered on Publisher and Alignment. The Publisher filter keeps Dark Horse Comics, DC Comics, Marvel Comics and NBC - Heroes. The Alignment filter keeps bad, good and neutral.

## Combat Skills vs Height




Height Cm vs. Combat. Color shows details about Alignment. The view is filtered on Height Cm and Alignment. The Height Cm filter ranges from 20.0 to 575.0. The Alignment filter keeps bad, good and neutral.

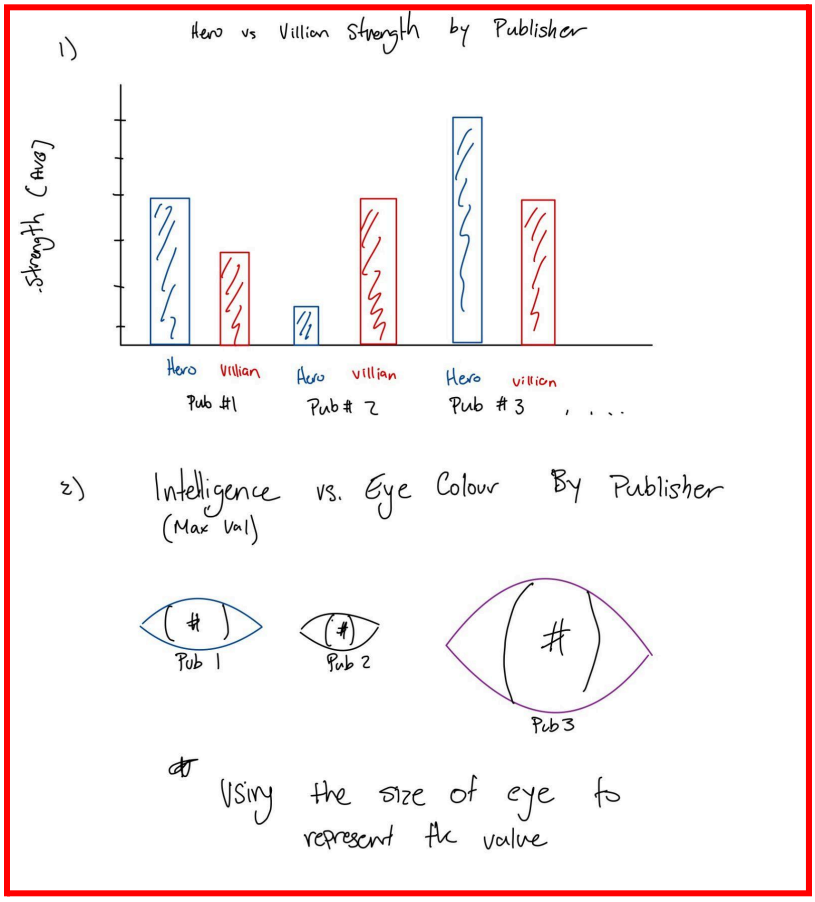
# Heroes vs Villians Height/Weight



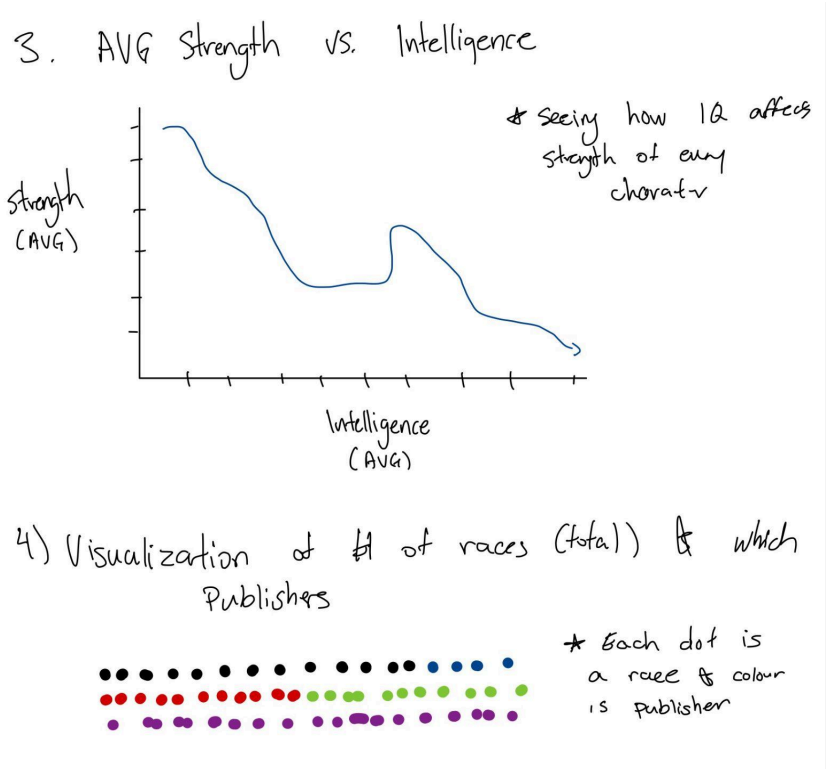

Average of Weight Kg and average of Height Cm for each Alignment. Color shows details about Alignment. The view is filtered on Alignment, which keeps bad, good and neutral.

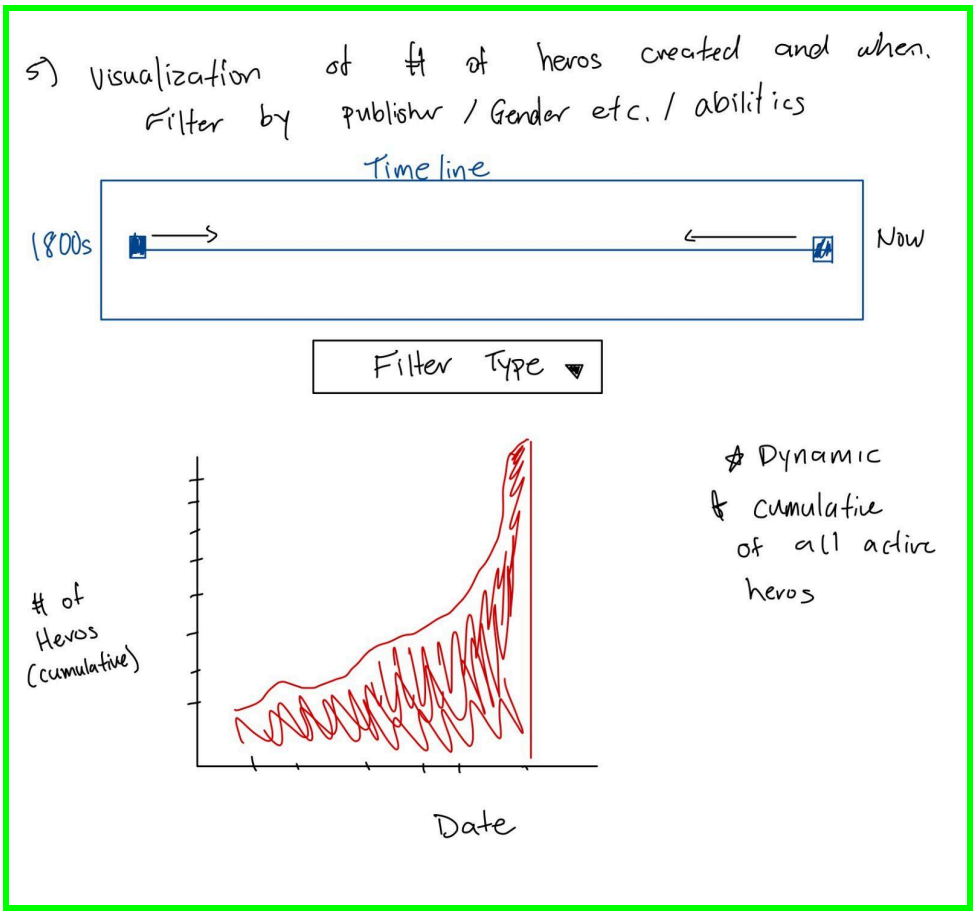
My first visualization tackles the distribution of character alignment by publisher, which is similar to question 4, but instead of just comparing quantitative attributes like physical or strength, I wanted to look into what the top publishers' proportions of heroes and villains are. This could give us insight into whether a particular publisher has historically preferred to overload its stories with positive role models and heroes, or maintained a balance between the number of heroes with villains and neutral characters as well. I think that this shows a broader trend in different publishers' attitudes towards storytelling, and the narratives that they tell.
The second visualization answers the question posed by question 3, wherein I sought to find how the physical attribute (height particularly) correlates with power levels (in terms of combat level). To my surprise, there is actually no clear trend, with the taller characters often having less than average combat level. I think that this was an interesting question, to challenge stereotypes of build vs physical ability, and I actually came upon this visualization without going through the questions particularly.
The last visualization is a different angle taken to answer question 1, which is to investigate how the characteristics of heroes differ from those of villains. I focused on physical attributes, namely height and weight, and found that villains tend to have higher levels of both, showing the physical dominance that heroes overcome in these stories so often.

- Eric 5 visualizations (sketched)



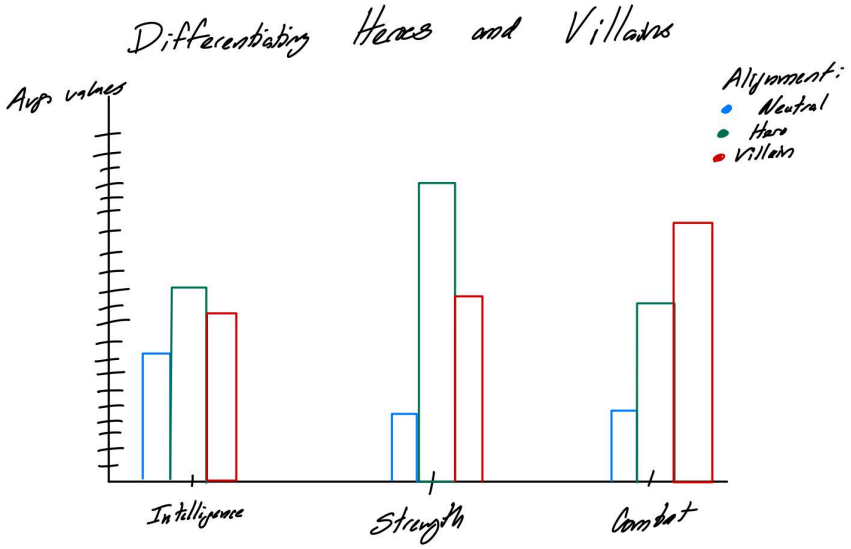

-



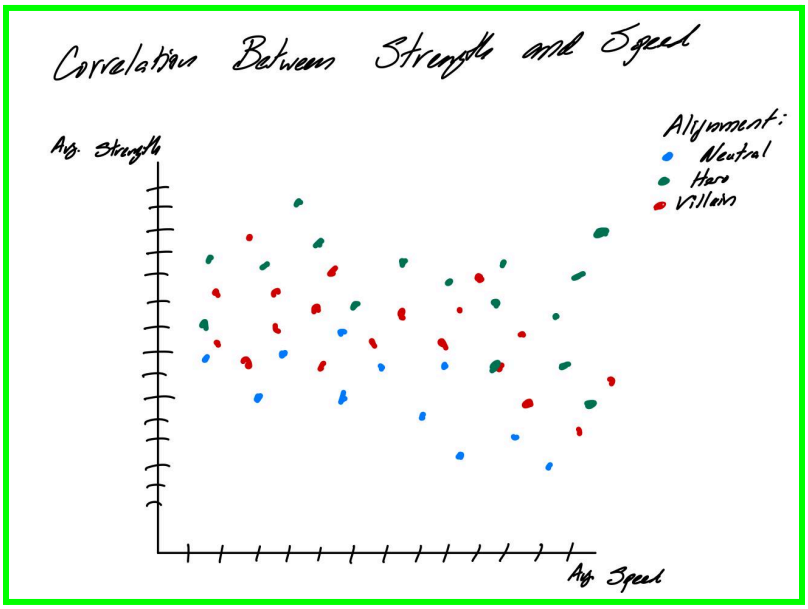



– Chris Canzano

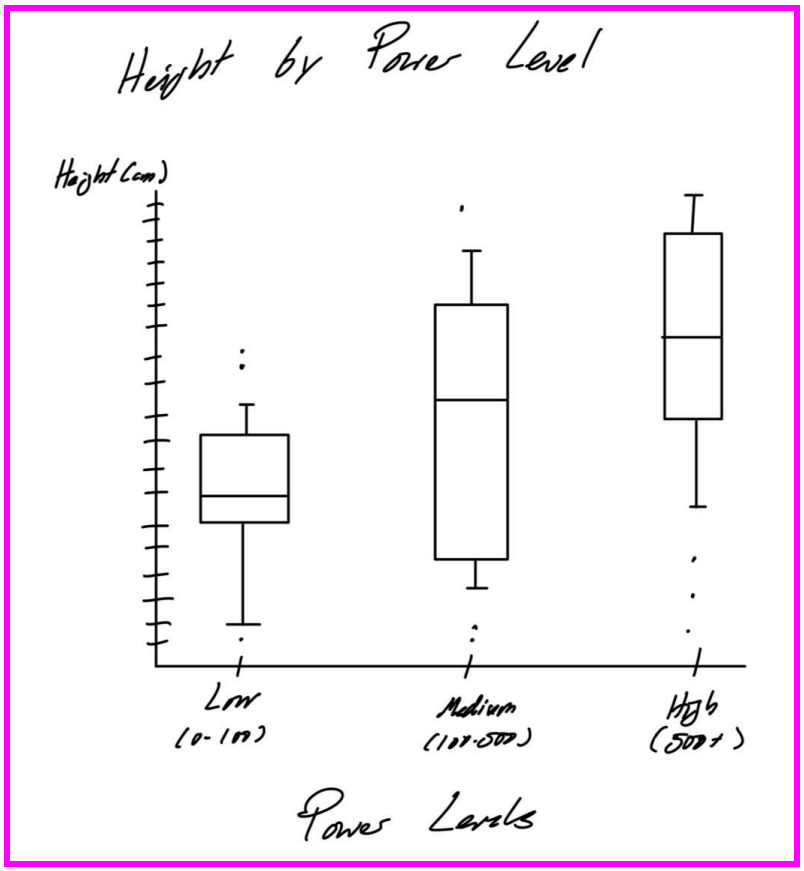
Question 1: What are the primary characteristics that differentiate heroes from villains? Which power stats or physical attributes are significant indicators of alignment?



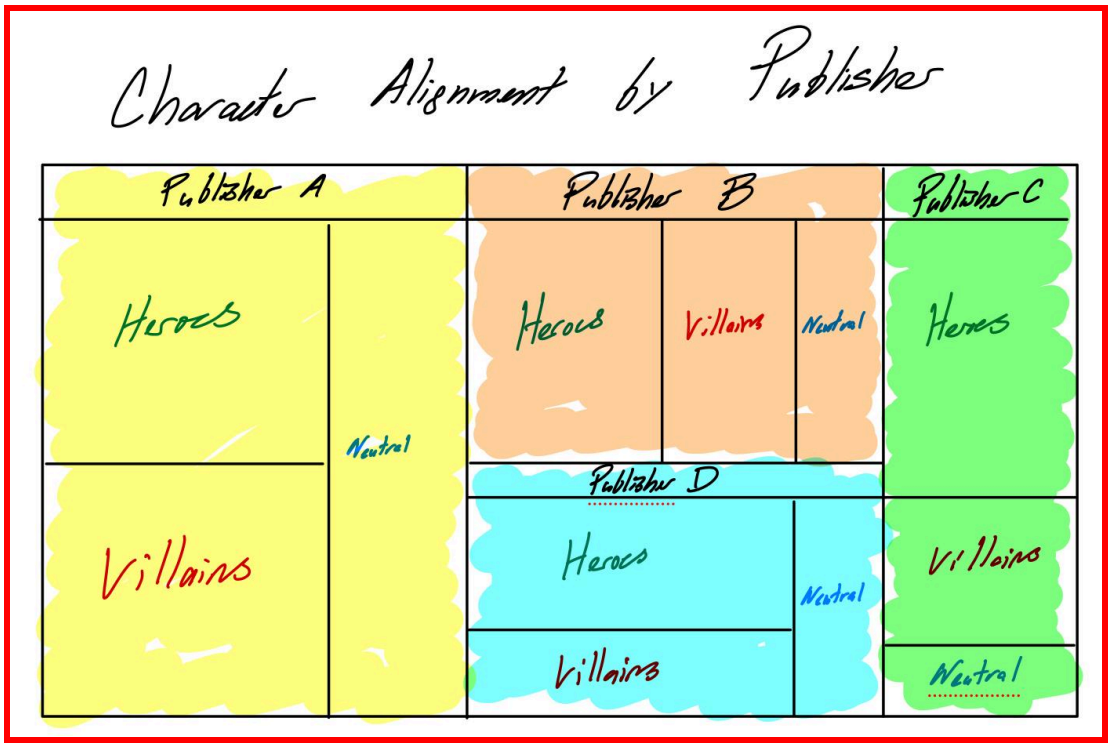
Question 2: How do intelligence, strength, and speed correlate with one another across characters? Are there trade-offs between these attributes?



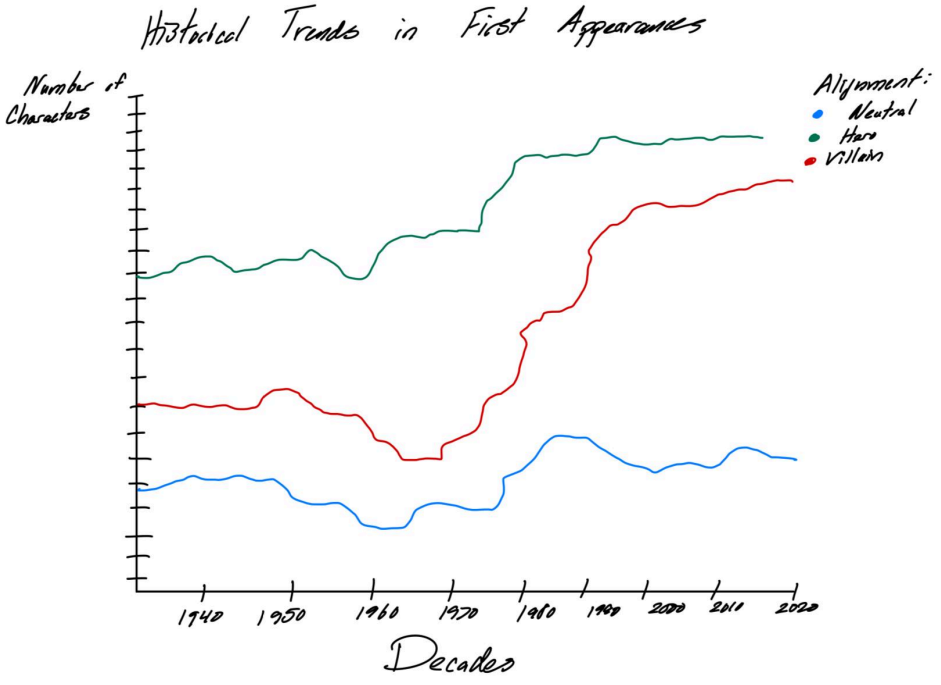

Question 3: Are there any discernible patterns or trends in the physical attributes (height, weight) of characters with certain power levels?



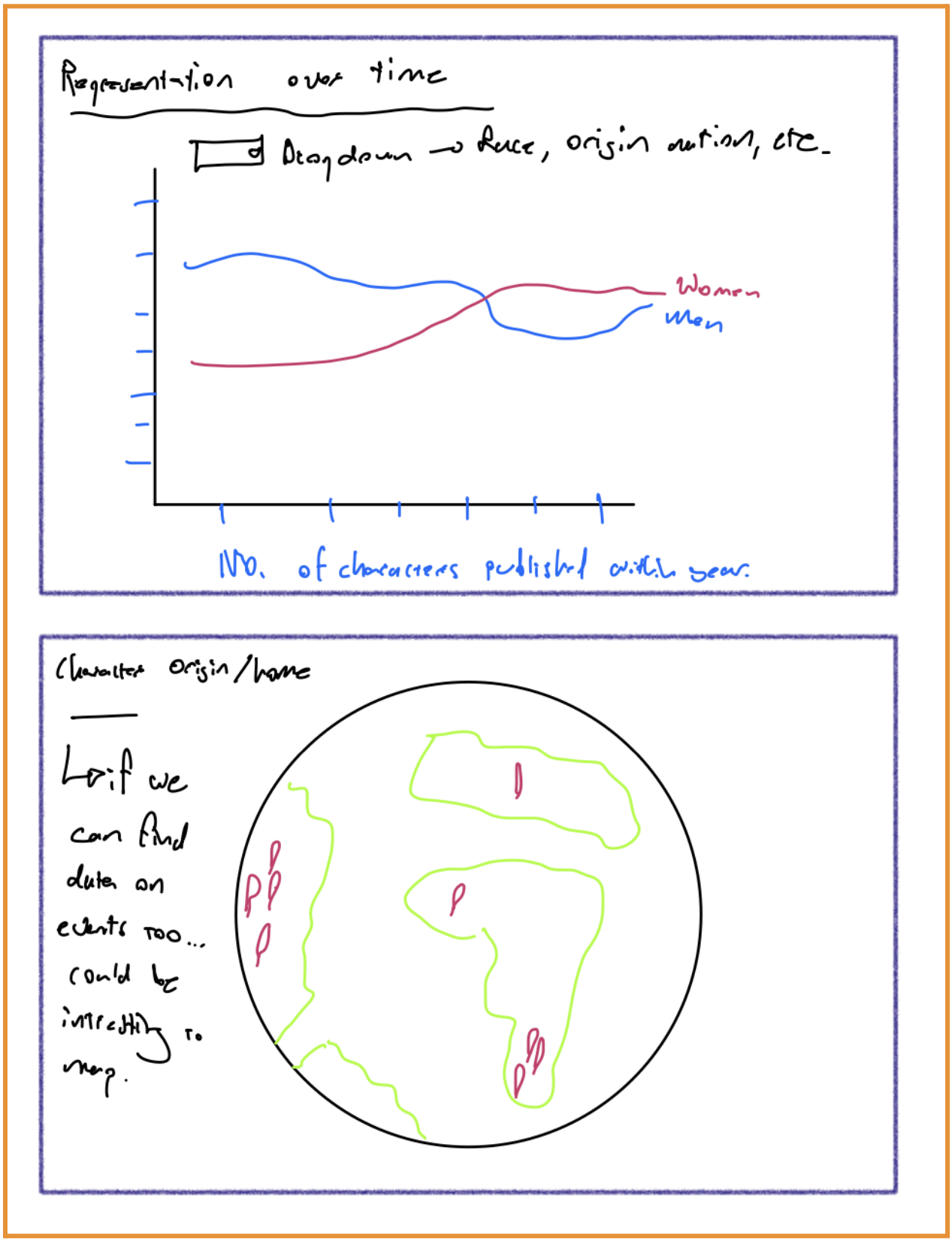

Question 4: Which publishers dominate certain character archetypes or power levels? Do some publishers tend to create characters with similar attributes?

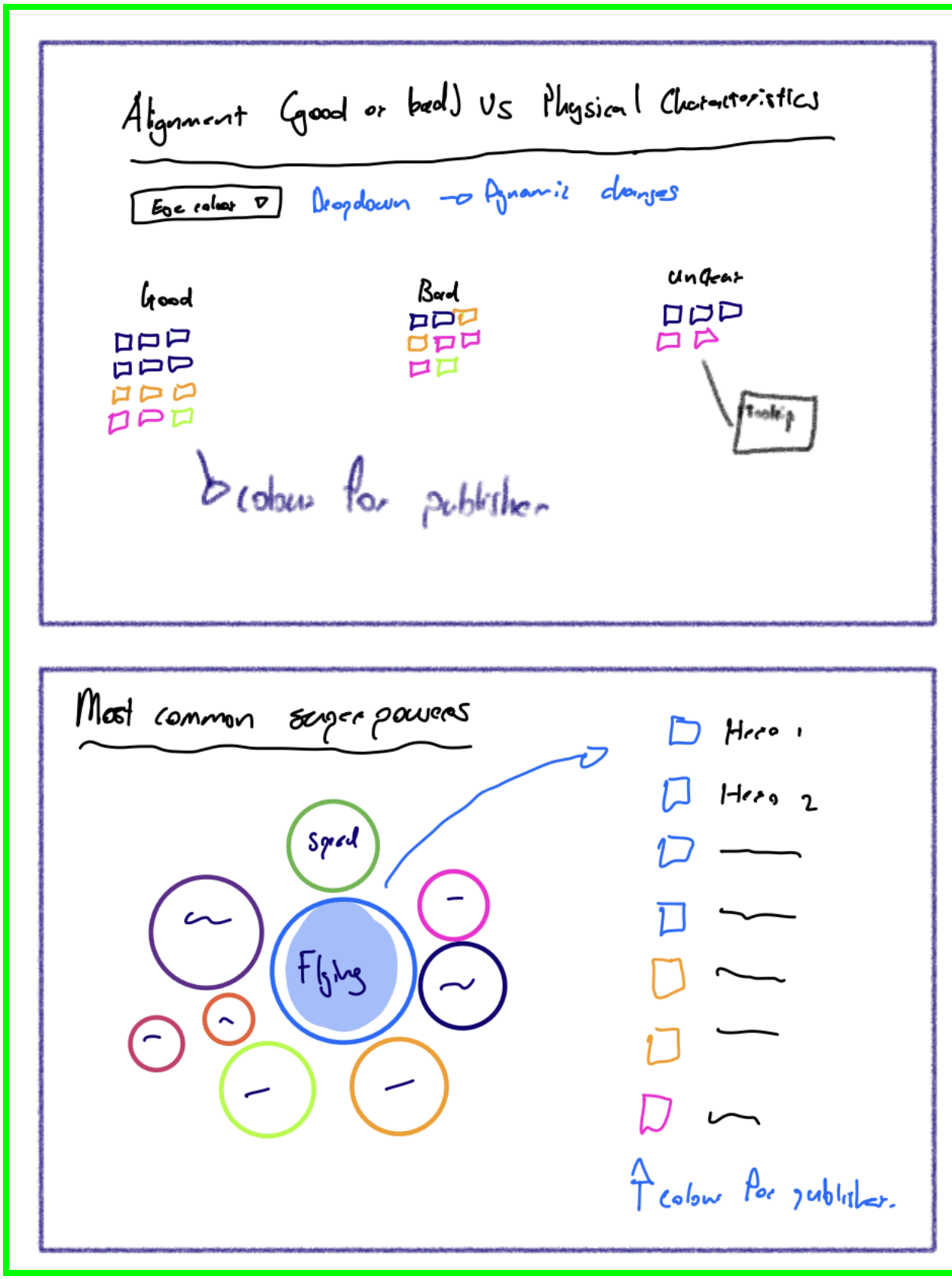


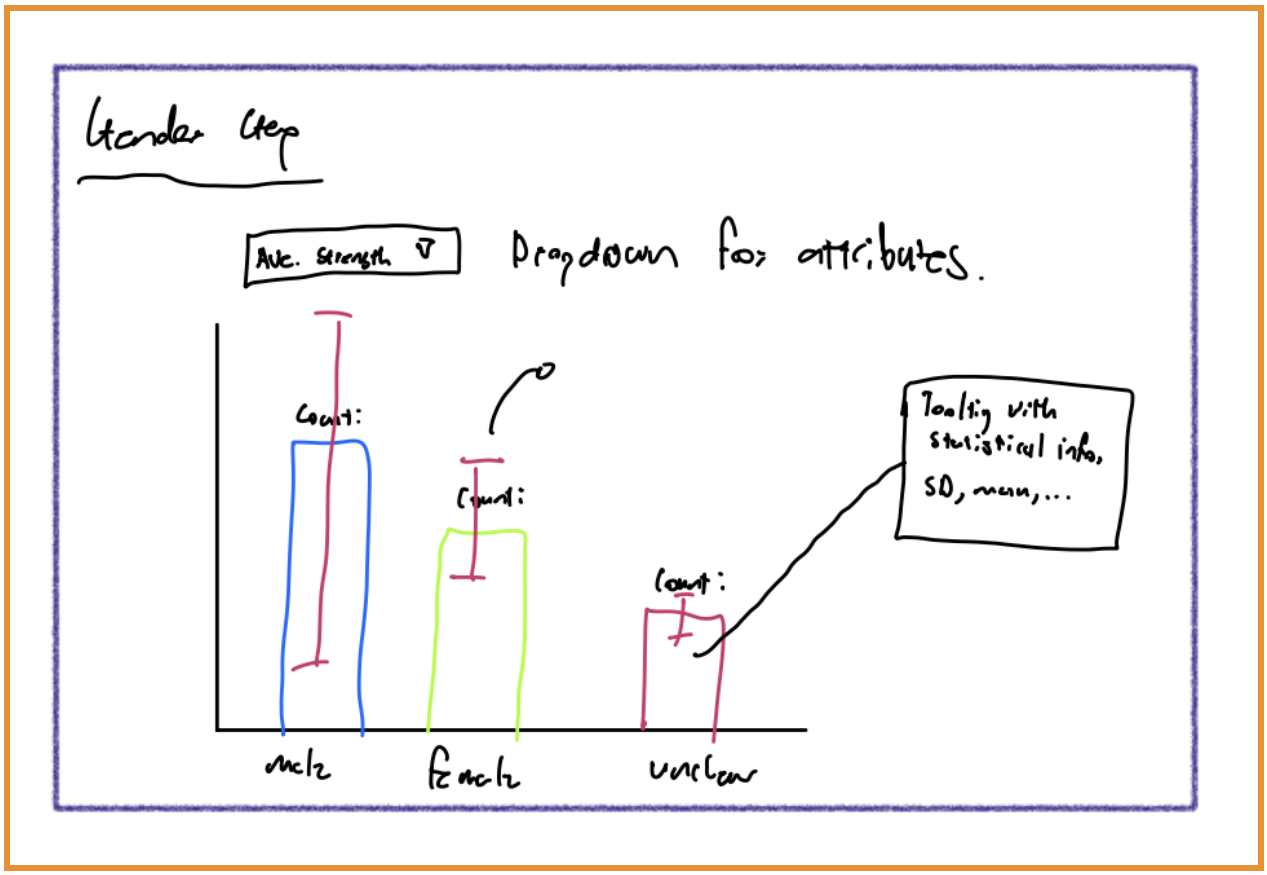

Question 5: What are the most common affiliations among characters? Are heroes or villains more likely to be affiliated with larger groups?



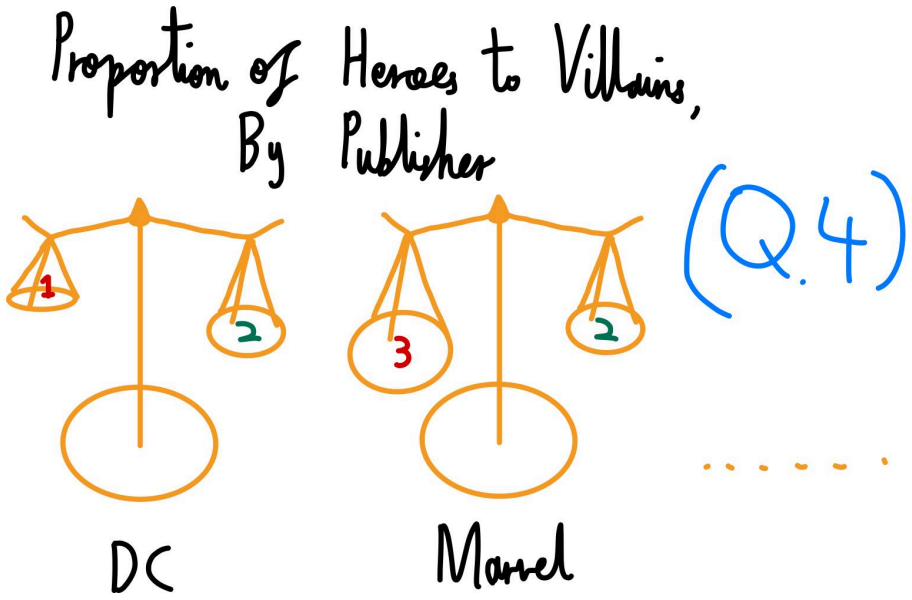
– Rohan Naidoo

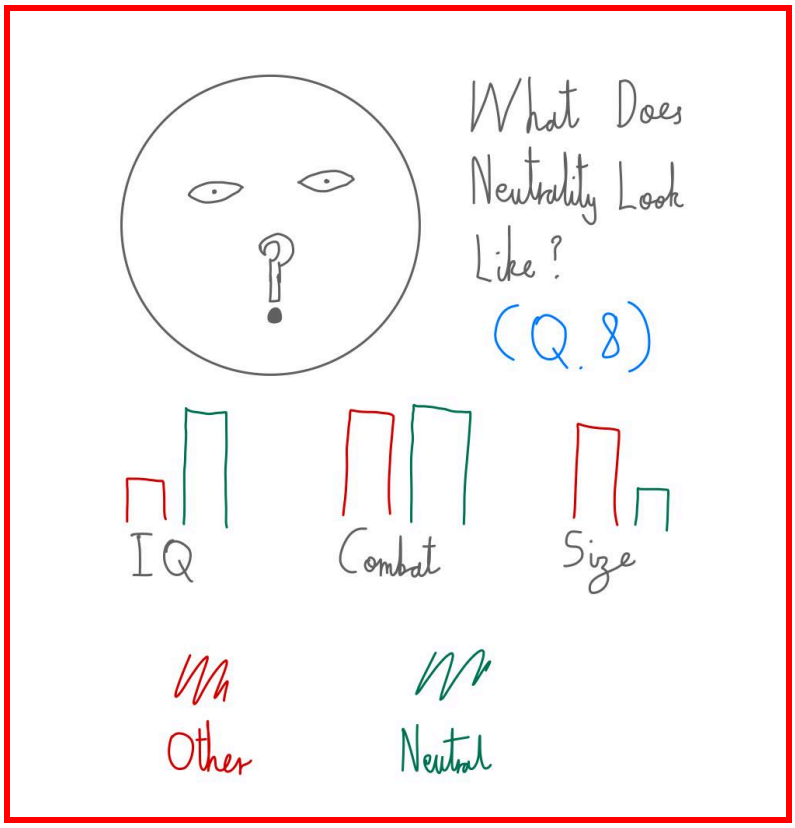


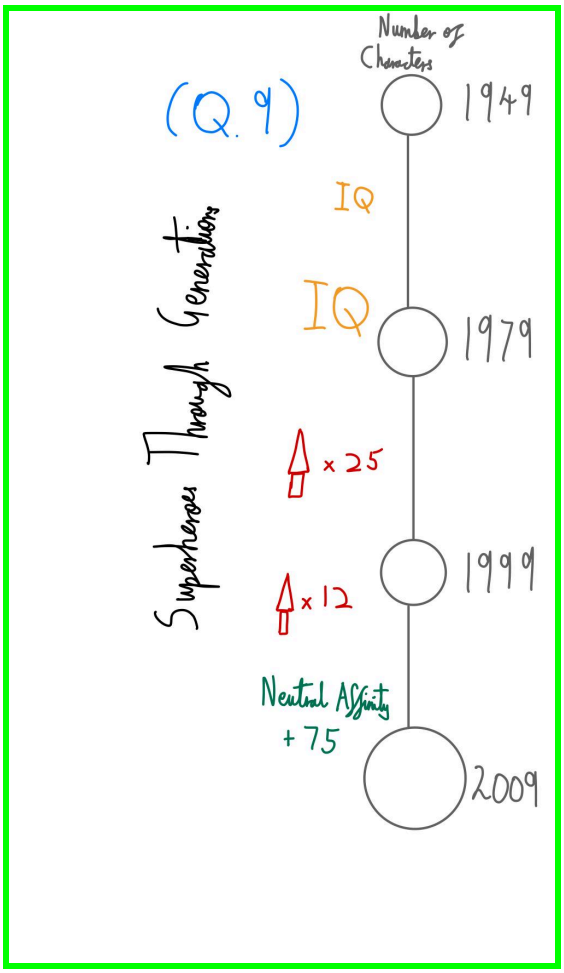

Alignment (good or bad) vs Physical Characteristics
Eye colour
Dropdown → dynamic changes
Good
Bad
Unclear
Tooltip
colour for publisher
Most common superpowers
Speed
Flying
Hero 1
Hero 2
colour for publisher.

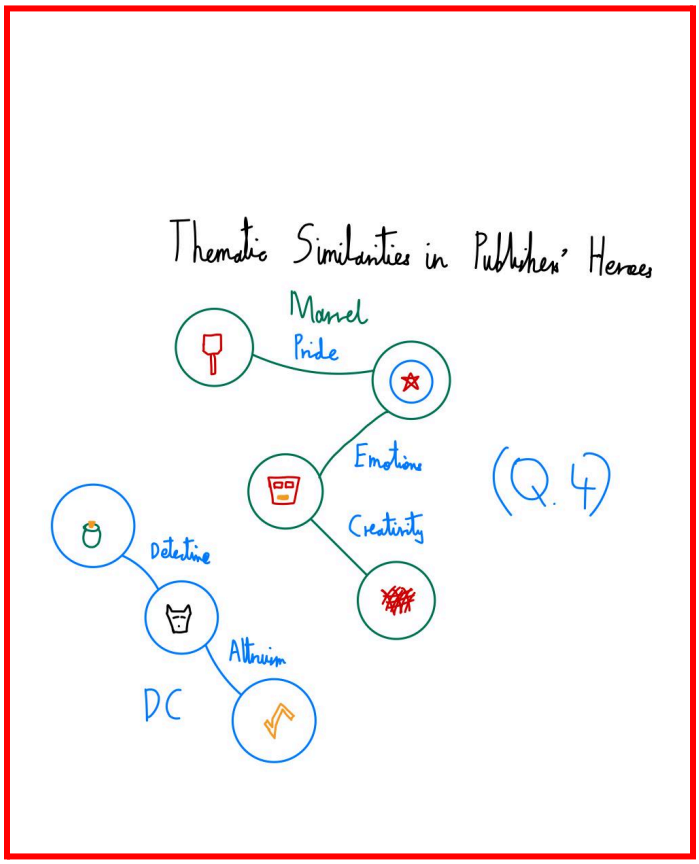


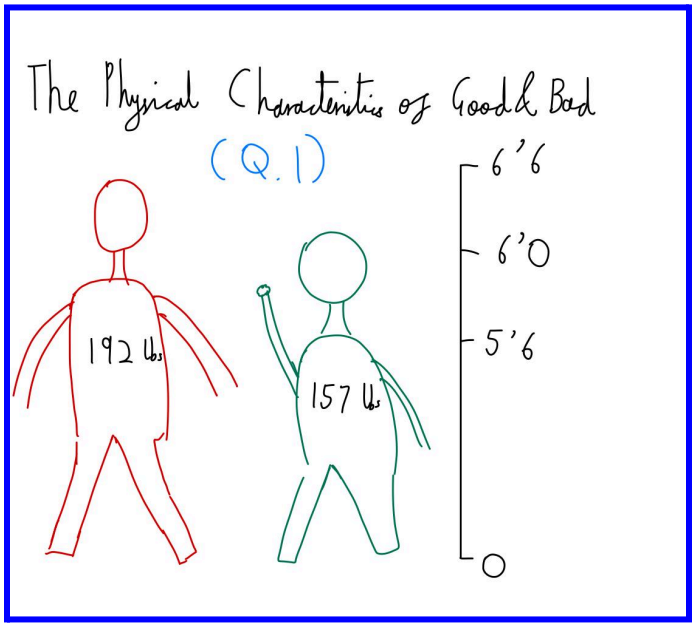

– Shahmir Aziz




What Does Neutrality Look Like?
(Q. 8)
IQ
Combat
Size
Other
Neutral


(Q. 9)
Superheroes Through Generations
Number of Characters
1949
IQ
IQ
1979
x 25
1999
x 12
Neutral Affinity
+ 75
2009


Thematic Similarities in Publishers' Heroes
Moral Pride
Emotions
Creativity
Detective
Altruism
DC
(Q. 4)


The Physical Characteristics of Good & Bad
(Q. 1)
192 lbs
157 lbs
6'6
6'0
5'6

---

**Insights**

– Chris Canzano

Insight 1: Power Distribution by Alignment

- Characters classified as "good" have higher aggregate values across most power categories (intelligence, speed, durability, etc.) compared to "bad" and "neutral" characters.
- "Neutral" characters tend to have lower power values overall, indicating that they might not rely on extraordinary abilities as much as aligned characters.

Insight 2: Comparison by Gender

- Male and female characters display similar average values across attributes like intelligence, combat, and power.
- However, males slightly outperform females in durability and strength, suggesting a gendered trend in character creation that emphasizes physical resilience for male characters.

Insight 3: Speed vs. Height

- Taller characters tend to exhibit higher speed, but there are notable outliers at extreme heights with unusually low or high speeds.
- Male characters dominate the distribution across all heights, while females are concentrated in mid-height ranges with moderate speeds.

– Shahmir Aziz

Insight 1: Good vs Evil Representation

- The representation of good vs evil varies greatly across publishers and times – heroes are generally shown as smarter, less powerful beings who overcome the odds to win the fight for good
- The proportion of heroes vs villains also ties into the takeaways that publishers are trying to convey

Insight 2: Generational Changes in Superheroes

- Over time, there has been significant changes in prioritization of different themes or characteristics seen as powerful, popular, and heroic.
- There has been a shift from more physical to more intellectual superiority and dominance through the characters.

Insight 3: Publisher Tendencies and Storytelling Intentions

- There tend to be thematic similarities between many characters from a particular publisher, which highlights the publisher's viewpoint of heroism and their intentions with their stories

– Eric Li

Insight 1: Intelligence and Eye Color

- Characters with hazel and gold eye colors have the highest average intelligence, while those with violet or blue/yellow combinations have the lowest.
- This might reflect creative decisions in character design linking intelligence to less common or "mysterious" traits.

Insight 2: Alignment vs. Publisher

- Marvel Comics has a greater proportion of "good" characters compared to DC Comics, which shows a balanced mix of "good" and "bad" characters.
- Neutral characters are rare across all publishers, but DC Comics has a slightly higher proportion of neutrals than Marvel or other publishers.

– Rohan Naidoo

Insight 1: Height/Weight by Alignment

- "Neutral" characters have significantly higher average height and weight than both heroes and villains, possibly representing characters designed as physically imposing but not strongly aligned with a moral side.
- Villains are shorter and lighter on average compared to heroes, which could reflect archetypes of cunning or less physically imposing antagonists.

Insight 2: Combat Skills vs. Height

- Combat skills peak for characters in mid-height ranges (around 170-200 cm), while very tall or short characters tend to have lower combat skills.
- Good-aligned characters have higher combat skills overall, while bad-aligned characters exhibit greater variability.

**Voting**

| Sketch ID | Question ID | Author |
|---|---|---|
| 1, 2, 9, 16, 19 | 4 | EL, CC, SA |
| 3, 7 | 2 | EL, CC |
| 4 | - | EL |
| 5, 18 | 9 | EL, SA |
| 6, 13, 20 | 1 | CC, RN, SA |
| 8 | 3 | CC |
| 10 | 5 | CC |
| 11, 15 | 7 | RN |
| 12 | - | RN |
| 14 | 6 | RN |
| 17 | 8 | SA |

**Main message**

Different superhero publishers represent the fight between good and evil through different characteristics and affinities, which highlights different storytelling philosophies.

- Side takeaway: The messaging of what it looks like to be a hero also changes across generations

Discussion:
The reason that our group chose this message is that it goes beyond just comparisons of fictional characters and their relative strengths, but takes these analytics to synthesize the messaging and purpose of superhero fiction through the lens of different large publishers as well as different generations. Drawing these visualizations will hopefully illuminate what the societal image of heroism or villainy is, how it's changed, and what these characteristics intend to show to the audience in terms of lessons to be learned.

**Click here to view:**
Storyboard

# V1 Prototype:

- Name of students that worked on prototype V1 submission.

Shahmir Aziz, Eric Li, Chris Canzano, Eric Li, Rohan Naidoo

- Data scraping and cleaning complete (using the real data sets)

We will be using the ‘Superhero Analytics’ dataset, taken from Kaggle. It contains information on 675 heroes and villains from various publishers and comic universes. (https://www.kaggle.com/datasets/shreyasur965/super-heroes-dataset/data)

The attributes include:

- Superhero's alias or code name.
- Intelligence, strength, speed, durability, power, combat: Numerical representations of the character's abilities.
- Full-name, alter-egos, aliases: Real names and alternate identities.
- Place-of-birth, first-appearance, publisher, alignment: Biographical and affiliation details.
- Gender, race, height, weight, eye-color, hair-color: Physical characteristics.
- Occupation, base, group-affiliation, relatives: Professional and relational information.

Concerns with the current state of the data:

We previously converted some of the columns for text to integers, but there was more work to do.

**Missing Values**:

- Some fields like full-name and height_cm have NaN values.
- The weight field has entries like '- lb' which need handling.

**Mixed Units**:

- Columns height and weight contain units like 'lb' and 'kg'. The numeric values are separated into height_cmand weight_kg, but we should confirm the consistency.

**Inconsistent Data**:

- Fields like group-affiliation and relatives have long, descriptive text, making it hard to analyze without parsing or categorization.
- The base column has generic entries like '-'.

**String Cleaning**:

- The occupation and relatives fields have inconsistent formatting and could benefit from standardization.
- The alias field has unnecessary brackets and quotations.

**Duplicates**:

- The id column should be unique, but this needs verification.

**Unnecessary Columns**:

- URL fields or repeated information might not be necessary for visualization.

```
import pandas as pd
import numpy as np
import os

# Step 0: Verify Current Directory
print("Current Directory:", os.getcwd())
print("Contents of Current Directory:", os.listdir())

# Load the dataset
file_path = 'superheroes.csv'
output_path = 'cleaned_superheroes.csv'

print("\nStep 1: Loading dataset...")
try:
    data = pd.read_csv(file_path)
    print("Dataset loaded successfully. Here's the first few rows:")
    print(data.head())  # Display a preview of the data
except FileNotFoundError:
    print(f"Error: File '{file_path}' not found. Ensure it's in the same directory as this script.")
    exit()

# Step 2: Handle Missing Values
print("\nStep 2: Handling missing values...")

# Replace `'-'` and `['-']` in text fields with "Unknown"
text_cols = data.select_dtypes(include=['object']).columns
for col in text_cols:
    data[col] = data[col].replace(['-', "['-']"], 'Unknown').fillna('Unknown')

# Clean up `aliases` column: Remove brackets and quotes
if 'aliases' in data.columns:
    data['aliases'] = data['aliases'].str.replace(r"[\'\"\[\]]", "", regex=True)

# Leave numeric missing values as NaN (default behavior)

# Verify missing values are handled
print("\nMissing values after handling:")
print(data.isnull().sum())

# Step 3: Parse Mixed Units in 'weight' and 'height'
print("\nStep 3: Parsing weight and height...")
if 'weight' in data.columns:
    data['weight_kg'] = data['weight'].str.extract('(\d+)').astype(float)
    # Replace invalid weight values (if any) with NaN
    data['weight_kg'] = data['weight_kg'].replace(0, np.nan)
    print("Weight parsing results:")
    print(data[['weight', 'weight_kg']].head())
else:
    print("Column 'weight' not found in dataset.")

# Step 4: Standardize Text Columns
print("\nStep 4: Standardizing text columns...")
if 'eye-color' in data.columns and 'hair-color' in data.columns:
    data['eye-color'] = data['eye-color'].str.lower().str.strip()
    data['hair-color'] = data['hair-color'].str.lower().str.strip()
else:
    print("One or both columns ('eye-color', 'hair-color') not found.")

# Step 5: Remove Unnecessary Columns
print("\nStep 5: Removing unnecessary columns...")
if 'url' in data.columns:
    data = data.drop(columns=['url'])
    print("Dropped 'url' column.")
else:
    print("'url' column not found, skipping.")

# Step 6: Deduplicate Data
print("\nStep 6: Removing duplicates...")
initial_rows = len(data)
data = data.drop_duplicates(subset='id')
final_rows = len(data)
print(f"Removed {initial_rows - final_rows} duplicate rows.")
```

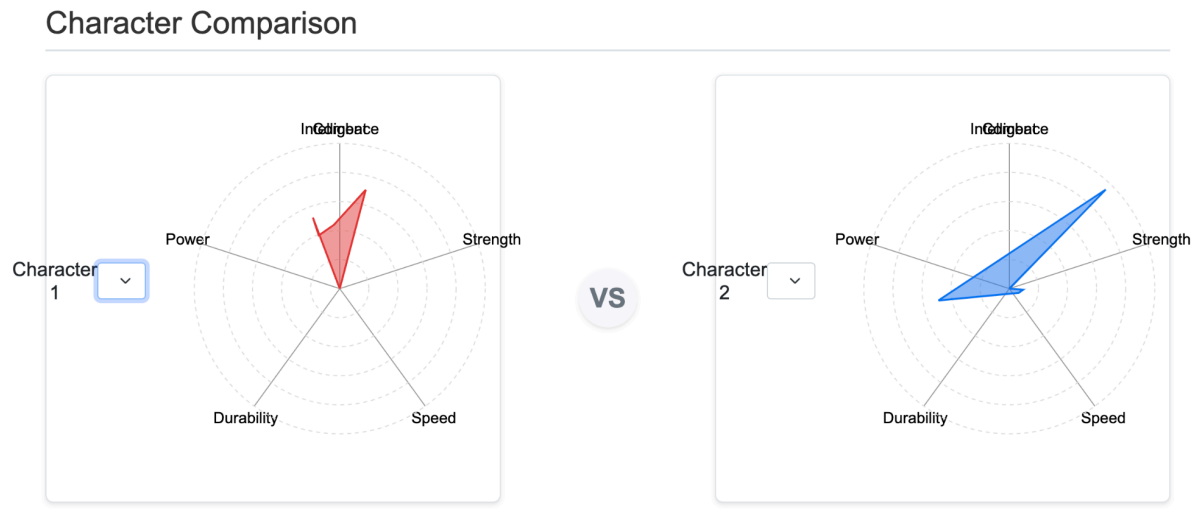
```python
# Step 7: Validate Numeric Ranges
print("\nStep 7: Validating numeric ranges...")
for col in ['intelligence', 'strength', 'speed', 'durability', 'power', 'combat']:
    if col in data.columns:
        data[col] = data[col].clip(lower=0, upper=100)  # Ensure values stay between 0 and 100
print("Validated numeric ranges for superhero attributes.")

# Step 8: Feature Engineering
print("\nStep 8: Adding derived features...")
if 'height_cm' in data.columns and 'weight_kg' in data.columns:
    # Replace invalid heights with NaN before BMI calculation
    data['height_cm'] = data['height_cm'].replace(0, np.nan)
    data['BMI'] = data['weight_kg'] / ((data['height_cm'] / 100) ** 2)
    data['BMI'] = data['BMI'].replace([np.inf, -np.inf], np.nan)  # Handle invalid BMI
    print("BMI column added.")
else:
    print("Missing columns for BMI calculation.")

# Step 9: Export Cleaned Data
print("\nStep 9: Saving cleaned data...")
try:
    data.to_csv(output_path, index=False)
    print(f"Cleaned data saved successfully to {output_path}")
except Exception as e:
    print(f"Error saving file: {e}")
    exit()

# Final Validation: Print Dataset Summary
print("\nFinal Dataset Info:")
print(data.info())
print("\nSample of Cleaned Data:")
print(data.head())
```

- **At least two D3 visualizations already partly implemented (including data loading and the basic vis, filtering does not have to work yet)**, and detailed drafts for 2-3 more visualizations




- Rough webpage design and structure has to be done and implemented (placeholders for visualizations, text, and images allowed)

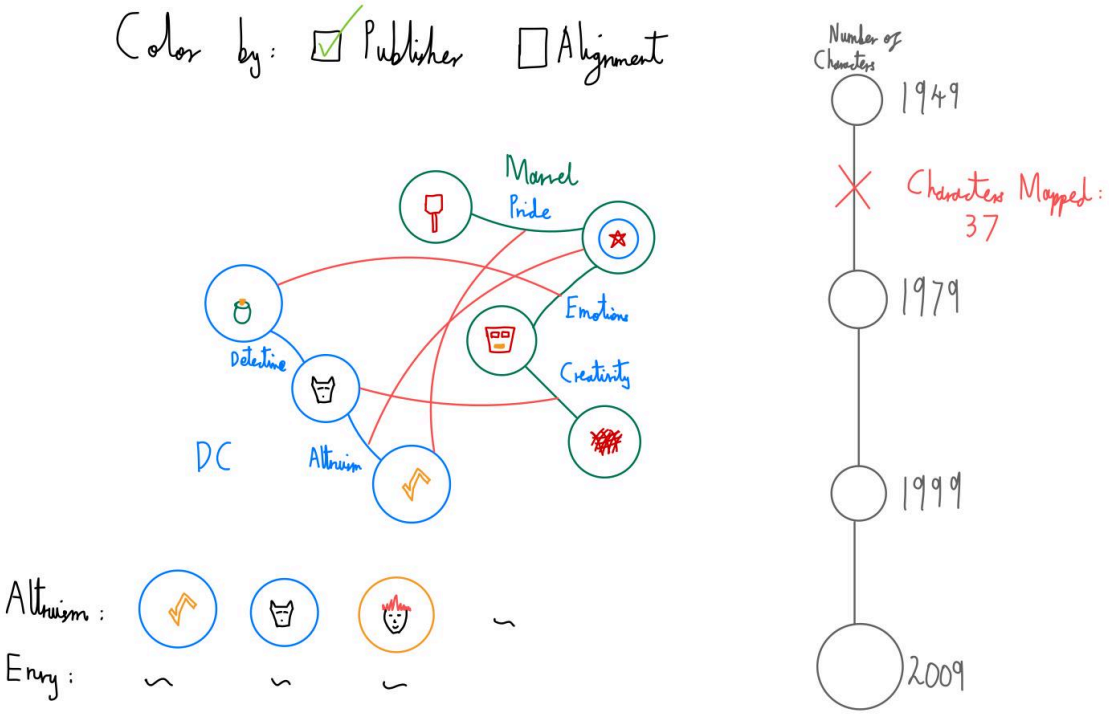
Code for historical plot for brushing exists, we just need to fix the data so that the historical dates are accessible
Code for the scatterplot above is also almost done, and needs to be made more ‘readable’ by the user.

- Storytelling is clear

We are trying to tell a story through our visualizations, and might rearrange our visualizations accordingly. The message we are trying to convey is how heroic messaging has evolved over

generations, and the different approaches different publishers have taken to display the fight between good and evil.

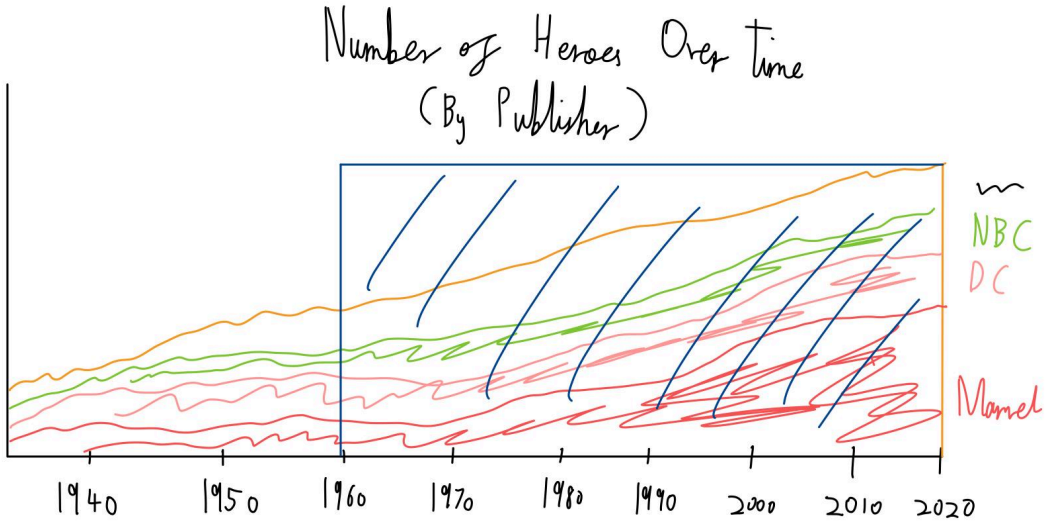
- **The first design of an innovative view**



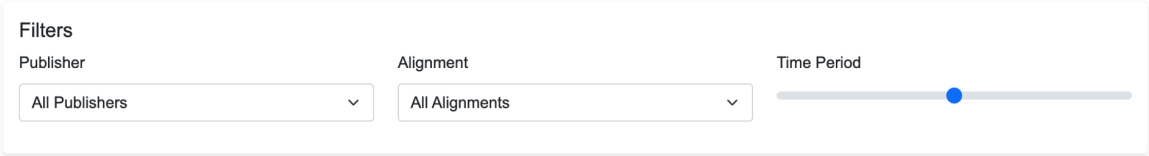

On the interactive vertical timeline, we can drag along a selector for a particular year. For that year, it will tell you how many characters are mapped in our ‘network graph’, in which different characters are plotted as circular logos containing their unique identifiers (which we will be drawing ourselves).
These logos will be connected to each other through central character traits, like shown above (empathy, pride etc.), to show how different traits are expressed across publishers or alignment types. We can choose how the logos are colored (above is by publisher, but by alignment is also possible). There is also a simple list at the bottom showing all the logos associated with a particular trait, but the network graph will show us connections between particular traits.

(Force directed layout for nodelink)
https://observablehq.com/@d3/force-directed-graph-component

- Interactions (e.g., filtering, brushing, etc.) have to be designed (at least in a textual description and some sketches)




The above chart will show the number of heroes created overtime, and there will be a brush element on this graph allowing us to select a particular range of years for which to show all the rest of the visualizations based on.
Accordingly, this chart will probably be on the top of the website, so users can interact with this before seeing all the other graphs.




At the top of the website, there will also be drop down menus for particular publishers to filter by, as well as particular alignments, to give further flexibility to users in addition to the time range for which to show all visualizations.

**PROTOTYPE V2 TASK LIST**

- Name of students that worked on prototype V1 submission.
- Data scraping and cleaning complete (using the real data sets)
- **At least two D3 visualizations already partly implemented (including data loading and the basic vis, filtering does not have to work yet)**, and detailed drafts for 2-3 more visualizations
- Rough webpage design and structure has to be done and implemented (placeholders for visualizations, text, and images allowed)
- Storytelling is clear
- **The first design of an innovative view**
- Interactions (e.g., filtering, brushing, etc.) have to be designed (at least in a textual description and some sketches)
- Up-to-date process book

Scrolling,
Interactivity

Predict bar charts

We can have yap slides

STORYTELLING

**Tasks**

1. Ideate the story we want to tell
2. Determine layout
    a. Go through homeworks and labs to see easily replicable code
    b. Emphasize interactivity
3. Implement
    a. Scrolling
    b. Innovative viz
    c. Fit into design

Which Hero are You Most Like?

---

**Intro**

- Marvel/DC style intro

**Slide 1:**

- You receive letter from Aziz's School for Gifted Youngsters
- Explain premise
- It has become aware

**Slide 2: [vis: heroes over time (with drag option for time?), bubbles for traits]**

- Showcase of different heroes/villains that embody the main characteristics: ALUMNI
- Each embodying one of the four traits
- Maybe like 2x2 set of profile images and when you hover you get a description and stat rundown

**Slide 3:**

- Are you ready for the exam?

**Slide 3: Speed [vis: outlier finder]**

- Lab we had

**Slide 4: Durability**

- Button clicker

**Slide 5: Intelligence**

- Riddles, etc

**Slide 6:  Strength**

- Questionnaire

**Slide 7: Results [vis: radar chart]**

- Radar chart

**Slide 8: [network graph]**

- Network chart

**Slide 9: [vis: bubbles for characters similar to you, can choose different top characteristic to show characters of those traits]**

- Show hero and villain most similar to you, description, backstory, etc (bubbles, can hover over any)

**Slide 10: [vis: geographic if possible]**

**Slide 11: [vis: histogram of heroes vs villains]**

- Show if villains or heroes tend to have your traits
- Give user choice: Nature v nurture
- Do you want to stick with the group you align with most based on characteristics or do you want to choose other slide?

**Slide 12:**

-

**References/Closing:**

**Think Aloud Study Results**

| | Notes (To be filled by project leads) |
|---|---|
| Tester Name | Peter Pich |
| Describe any usability issues or confusion the tester encountered while using the prototype. | It goes from storytelling and quizzes to straight comparison of various superheroes — seemed to not be a reason for the speed test and questionnaire (as it is still WIP so makes sense). |
| Was the tester able to understand the main message of the data story? (e.g., Yes/No + why/why not?) | Kind of — again, it is a WIP so some confusion was expected. They understood the story but confusion set in when it jumped form user input to immediate comparison of other superheroes without seeing their results in it. |
| What parts of the interface or visualization did the tester find most engaging or effective? | The network graph which was interactive seemed to be a hit. |
| What parts did the tester find confusing or less effective? | The bar chart seemed to be a little bit too simple and did not bring as much engagement as the other ones |
| Did the tester encounter any inconsistencies in design, data, or narrative? | The jump from the comic graphics to very basic visualizations was inconsistent with the webpage design |
| Were there any unexpected interactions or insights that emerged during the session? | Again, the jump from user input based interactions to just visualizations. |
| What specific improvements or changes did the tester suggest for the prototype? | To use the data gathers from the user and implement that — will be used in final implementation |
| Did the tester suggest any additional insights or visualizations to include? | Jazz up the visualizations a bit to match the comic book feel at the beginning |
| General observations or comments from the tester. | Off to a good start, but just needs polishing |

| | Notes (To be filled by project leads) |
|---|---|
| Tester Name | Andrew |
| Describe any usability issues or confusion the tester encountered while using the prototype. | Felt like there was inconsistent design (jump from questionnaire and interactive to comparisons of |

| | |
|---|---|
| | different superheroes). |
| Was the tester able to understand the main message of the data story? (e.g., Yes/No + why/why not?) | Yes they were — it was made clear that it was based on superhero stats from various universes. |
| What parts of the interface or visualization did the tester find most engaging or effective? | They liked the interactiveness of the speed test and the strength questionnaire. They also liked the network graph which was also interactive. |
| What parts did the tester find confusing or less effective? | Why there was questions/visualizations that gathered user info but was not used later on. This will be implemented later on in the final version. |
| Did the tester encounter any inconsistencies in design, data, or narrative? | They thought the abrupt change from comic book inspired designs to very basic visualizations like the bar chart on a white background was inconsistent. |
| Were there any unexpected interactions or insights that emerged during the session? | N/A |
| What specific improvements or changes did the tester suggest for the prototype? | To make design of the visualizations more consistent with the comic book design and keep the colour scheme the same. |
| Did the tester suggest any additional insights or visualizations to include? | They suggested to implement a user stat vs. various superhero visualization (which was already planned in the final version). |
| General observations or comments from the tester. | They enjoyed the interactive visualizations the most. |

| | Notes (To be filled by project leads) |
|---|---|
| Tester Name | Andrew Seybold |
| Describe any usability issues or confusion the tester encountered while using the prototype. | After the initial few slides, the assortment of visualizations were confusing and didn't seem to have structure. |
| Was the tester able to understand the main message of the data story? (e.g., Yes/No + why/why not?) | Yes- very clear that I was admitted to Xavier's school |

| | |
|---|---|
| What parts of the interface or visualization did the tester find most engaging or effective? | The beginning slides when Professor Xavier was explaining the premise were particularly effective. |
| What parts did the tester find confusing or less effective? | Mostly toward the end where Professor Xavier seemingly disappeared |
| Did the tester encounter any inconsistencies in design, data, or narrative? | Not any noted |
| Were there any unexpected interactions or insights that emerged during the session? | No |
| What specific improvements or changes did the tester suggest for the prototype? | Fixing the latter half of the website to continue the format of the beginning half of the story |
| Did the tester suggest any additional insights or visualizations to include? | Maybe some way to match up your skills with a superhero or villain of your choice. Or a surprise attack by villains later |
| General observations or comments from the tester. | Loved the idea, great scheme– just needs to be polished toward the end |

Based on the results of your 'think aloud' study, what would you improve in your data story?

I think the main thing we need to do is consistency in our story whether that is design or type of visualization. In the final implementation we will be collecting user data from interactive visualizations and using that data to compare it to existing superheroes in our visualizations.

Are there any additional insights and visualizations you would use? Would you amplify or change your message? Did your narrative work? Did the tester get your takeaways?

The testers did get the takeaways even if they weren't fully implemented. I think that we just didn't have time to implement everything we wanted or else the message would have been amplified and been much more clear.

Decide as a team which of these improvements you will implement and write down your decisions and why you made them in your process book as a numbered list.

The use of the user data from the initial visualizations will be first and foremost. For example, instead of doing character vs character comparison in the rader chart, we will start off with "Your stats" vs Superhero 1 so the user can see how they fare against other heroes and villains.

Post-Prototypes Work:

- Fix scrolling
- Add music to intro
- Add User info to radar and network charts
- Ensure story is cohesive