**Project 4**

With the renewed popularity and success of the superhero genre lately, particularly in movies, an abundance of previously lesser known superheroes have become mainstream. This explosion of names and characteristics can become confusing, especially to someone just dipping their toes into the genre. Therefore, the ability to quickly group similar characters into groups of similar characteristics could potentially assist in quickly understanding an otherwise complex genre.

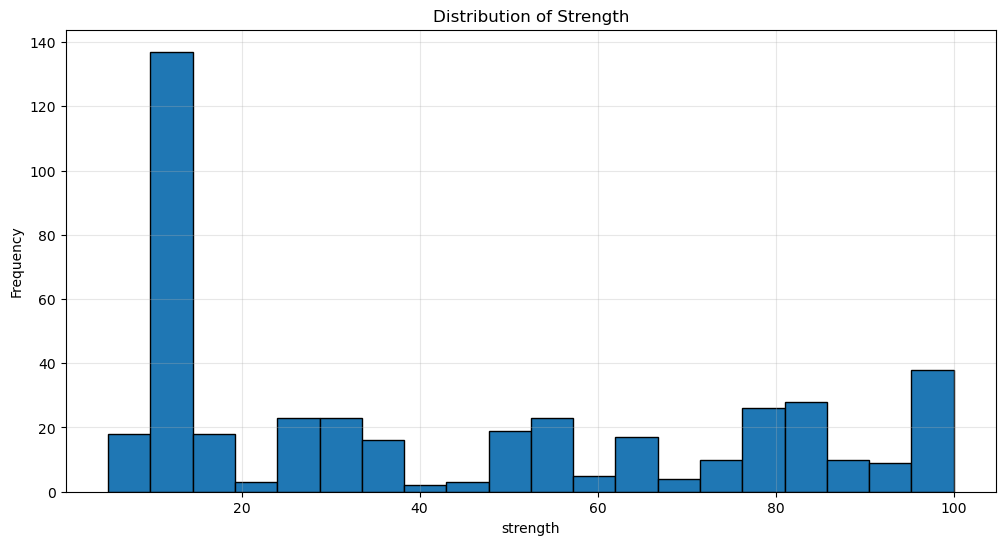
The way this will be done is with a process known as clustering. Clustering takes a large set of data and attempts to find trends within it, so that the dataset can be broken into smaller groups of similar data points. This allows for data that is otherwise potentially overwhelming to be much more understandable. More specifically, the clustering algorithm k-means will be used. K-means creates *k* random data points, where the value of *k* is chosen by the modeler. These data points are inserted randomly into the data space, and called the centroids. At this point the distance between each data point and the nearest centroid is calculated. Then the centroids are recalculated, using the average location of the data points associated with it, with the hope being that it becomes more similar to the data points it represents. This process of calculation and reposition is repeated until the centroids can no longer be further optimized. At this point the best possible k-values have been found and the modeling is complete.

The data for this project was sourced from Kaggle, and has a shape of 731 rows and 26 columns (https://www.kaggle.com/datasets/shreyasur965/super-heroes-dataset). The columns in the dataset were “id” (a numeric id with no predictive value), “name” (their superhero moniker), “intelligence”, “strength”, “speed”, “durability”, “power”, “combat”, “full-name” (their full, human name), “alter-egos” (alternate, lesser used superhero names), “aliases” (any nicknames, titles, or alter-egos), “place-of-birth”, “first-appearance” (first comic book appearance of the character), “publisher” (the publisher of the character’s comics), “alignment” (good or evil), “gender” (male or female), “race” (not race in the traditional sense, but whether the character is a human or some other non-human species), “height” (in both inches and centimeters), “weight” (in both pounds and kilograms), “eye-color”, “hair-color”, “occupation”, “base”, “group-affiliation” (what superhero groups the character has been a part of or been associated with), “relatives”, and “url” (a link to an image of the character). The columns “intelligence”, “strength”, “speed”, “durability”, “power”, and “combat” all had values ranging from one to one-hundred, with high values indicating a particular aptitude in that category. In total the dataset consisted of six float64 columns, one int64 column, and nineteen object columns.

In order to get a quick view of the dataset a function was created which made a histogram of every numeric column. These histograms give a quick view of each column and show the distribution. One notable graph in particular was the distribution of the strength column, which can be seen in Figure 1. This histogram showed that the vast plurality of superheroes have a strength level between ten and fifteen. This might make clustering difficult, as many of the data points are going to be extremely similar.

**Figure 1**

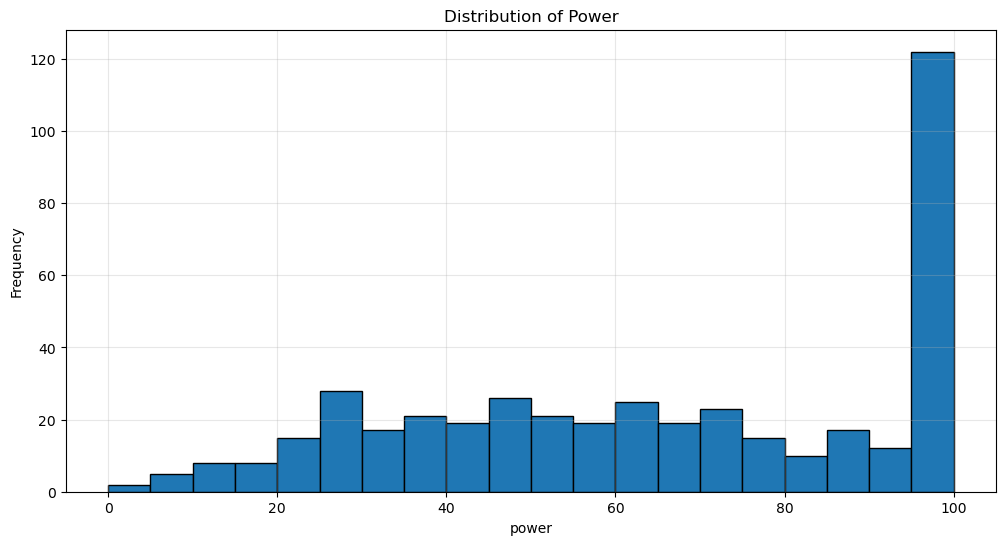
*Histogram for Strength*



Similarly, the power histogram showed that a very substantial amount of the dataset had a power level greater than 95. Similar to the strength attribute, this may cause difficulty for the clustering algorithm, as a large portion of the dataset is going to want to go toward the same centroid.

**Figure 2**

*Histogram for Power*



The other attribute columns had much more typical curves, though some, such as speed, skewed right, and some, such as intelligence, skewed right. The other notable finding from these histograms was that height and weight both had almost no variance, with the vast majority all falling within the same bin. This makes sense, as most humans have similar physical proportions, though there were a few extreme outliers.

Though the original dataset was generally clean, some preprocessing was still necessary. The first thing which was done was that all columns which obviously had no predicting power were removed. The removed columns were “name”, “id”, “url', 'full-name”, “alter-egos”, “place-of-birth”, “aliases”, “first-appearance”, “occupation”, “base”, “group-affiliation” and “relatives”. All of these columns were object types which contained information which would be extremely difficult for a clustering model to understand.

The next step taken was to check for null values. The most null-filled column was “race”, with three-hundred and two null values. The columns “intelligence”, “speed”, “durability”, “power” and “combat” all had one-hundred and sixty-five null values. “strength” had one-hundred and two null values. Finally, “publisher” had fifteen null values. After the null values and useless columns had been removed the new shape of the dataset was thirteen columns and five-hundred and fifty-five instances. This amount of data loss was suboptimal, but unfortunately unavoidable.

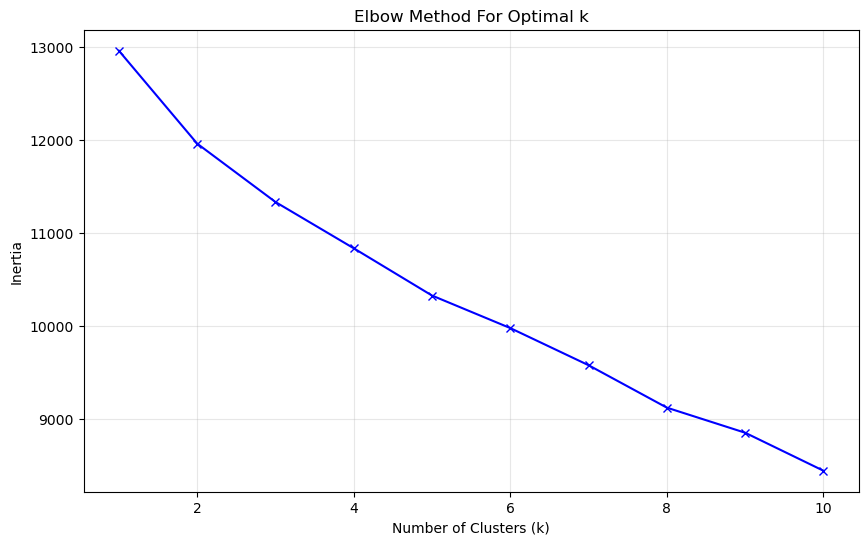
Another preprocessing step which was taken was to simplify the “height” and “weight” columns. Previously their outputs had been in the format “["6'8", '203 cm']” and “['980 lb', '441 kg']”. Of course, this object type format would be incredibly difficult for a clustering algorithm to understand, so the height in centimeters and weight in pounds were stripped from the original values, with the datatype being switched to float64. At this point I also noticed that the weight and height for some characters was being listed as zero, seemingly in place of null values. When these nulls were removed four-hundred and fifty-one instances remained.

The final preprocessing step was to split and one-hot encode the columns “publisher”, “alignment”, “gender’, “eye-color”, and “hair-color”. Once this process had taken place the former object types were changed to boolean types, where every different value became its own column. Any of these newly created columns with less than 10 positive instances were dropped from the dataset. After this step was taken a total of thirty columns were present in the dataset.

The first modeling step which was taken was to create an elbow curve, in order to test how many centroids the model should use. In order to do this a copy of the dataset was standardized so that each feature had a mean of zero and a standard deviation of one, ensuring that outliers would not skew this test. The k-means algorithm was then run ten times, using every k-value from one and ten. The inertia, which measures how similar the data points around a centroid are, was then measured for each k-value. The elbow curve was then plotted, visualizing the efficacy of each k-value in this problem, which is shown in Figure 3.

**Figure 3**

*Elbow Curve*



Looking at the elbow curve, a few candidate k-values were noted. The best k-values to choose are those which have a notable flatlining of the data between them and the subsequent datapoint, as that means that the value of an additional centroid is reduced, compared to the previous ones. According to this graph, the highest drop-off between centroids was between the second and third centroid. However, only using two centroids would likely not give very useful information. The next biggest drop-off seemed to be between k-values five and six. Therefore, a k-value of five was chosen for this project.

With the necessary preprocessing and data understanding complete, the final model was created. Using a k-value of five and a random state of forty-two the clustering was calculated. Print statements were used to show the average value of every categorical variable for each cluster, in order to give an effective snapshot of each group’s makeup. The same was done for the boolean values, so that those features ranged from zero to one, with higher values indicating a larger number of characters with that characteristic.

With five clusters created their contents could be viewed. When looking at the first cluster, the immediate standout features were the second highest strength, durability, and power stats, along with the highest speed. However, this cluster had the second lowest intelligence. Additionally, cluster one had the highest average height, and by far the highest average weight. With an average weight of two-thousand five-hundred and seventy-nine pounds cluster one was almost double the weight of the second heaviest group. These immediate metrics paint a picture of extremely large, tough characters whose lack of intelligence is made up for by their physical abilities. Going beyond the physical abilities of the characters, group one was overwhelmingly male, at eighty-eight percent, as well as being almost exclusively created by either DC or Marvel Comics. Interestingly, this cluster was also composed of exclusively neutrally aligned characters, in fact being every neutral character in the entire dataset. Clearly, this was a huge factor in creating this cluster. In total cluster one had seventeen heroes in it.

Cluster two had very average physical attributes, being neither the highest or lowest scoring group for any score, with the exception of intelligence, where it was the highest scorer. In general, the typical group two member seems to be a somewhat large, but still humanly possible, person with somewhat high levels of strength, combat, and power abilities. Within the boolean values, group two was ninety-eight percent male, largely from Marvel and DC Comics, ninety-seven percent heroic, seventy-two percent blue-eyed, with an even smattering of hair-colors. Overall, group two seemed to be composed of intelligent humans, who were not as physically imposing as group one, but much more heroic. There were one-hundred and fifteen characters in group two.

Group three was the dumbest, shortest, and lightest of the groups. Its other physical attributes were middle of the road, generally being the fourth highest group in any given stat. Group three’s other metrics showed that it was composed of ninety-eight percent female characters, majority Marvel, and eighty-three percent heroic. Physically, group three was plurality blue-eyed with a mixture of hair colors. Clearly, this cluster was largely defined by gender, which also likely explains the lower weight and height measurements. There were one-hundred and eighteen characters in total in this group.

Group four was the weakest group, with a strength more than half of the two strongest groups. Additionally, group four was the slowest, least durable, least powerful, and the worst at combat. In fact, the only physical characteristics which group four was not the lowest at were intelligence, height, and weight. Otherwise, group four was seventy-one percent Marvel characters, almost evenly split between heroic and villainous, and ninety-six percent male. The characters were largely brown-eyed and black or brown-haired. The extremely unimpressive physical display clearly shows how this group was likely chosen, with it largely composing the characters whose physical scores were the weakest.

The final group was the second most intelligent, fastest, and heaviest group, as well as the strongest, most durable, and most powerful group. These are very impressive physical traits. The group was also ninety-four percent male, majority from Marvel comics, and majority villainous. This group was plurality red-eyed, which was very interesting as that is an uncommon eye-color. Also strangely, the group was plurality bald, at forty-seven percent. There were seventy in group 5 in total. This group was dramatically different from many of the others, clearly being composed of strong, villainous characters who also looked the part.

Concluding this project, while the total impact is likely limited it is nevertheless not nonexistent. By grouping characters into archetypes the tropes and patterns of those characters can more easily be made familiar with, hopefully allowing new fans to quickly feel like they at least have a general grasp on the characters and their attributes. A limitation of this study is that it can really only teach so much, and grouping characters by their attributes will not help explain the storylines, character interactions, or more nuanced parts of characters that really set them apart.

Link to code:

<https://github.com/bigbadraj/Project-4---Superhero-Clustering>

Sources:

<https://www.kaggle.com/datasets/shreyasur965/super-heroes-dataset>

ChatGPT was used in some sections to help build code and debug, particularly when building the model.