An Exploration into Superhero Traits

Ethan Olpin, Huy Tran, Jeff Gay, John Duffy

University of Utah

ABSTRACT

With the aim of revealing how comic character design evolves over time, we apply clustering, collision counting, and regression to a dataset pertaining to comic characters from both Marvel and DC Universes. From our results we have identified different trends in superhero characteristics such as, the diversity of characters increasing over time, repeat appearances, or particular powers and traits overlapping more than others. Following a brief introduction to our dataset we describe our methods and findings in observing the trends of these characteristics.

Introduction To the Dataset

Our dataset was gathered from the following sources: 538's Comic Characters Dataset[1] and Danniel R's Marvel Superheros Dataset[2]. The data originates from the Marvel and DC Wikias and specifies the attributes of a wide array of comic book characters. These attributes include things such as their name, identity, moral alignment, gender, and aspects of their appearance (eye and hair color). Many characters appear less frequently than others in their respective comics, for this reason the data on them is incomplete, resulting in some unknown values, all such datapoints were filtered or accounted for in our figures.

Collision Counter (Marvel)

The backing logic for the Collision Counter is the Birthday Paradox/ Pigeonhole Principle. The label for these collisions are created with the parameters of 'eyes', 'hair', and 'sex' (biogoically). The 'unknown' values that do come up in the labels mentioned in the introduction does create 'outliers' or non-specific subsets.

The Marvel dataset was chosen over DC for this due smaller amount of 'unknown' values

With this section the figures show the differences between four time different periods. Among the finding for male and female there has complete trend difference. On the Female subsets (Figures 1 and 2),the trend from before 1990 and after it there was change in female physical attributes. The Male subsets (Figures 3 and 4) ended up in similar manner.

A conclusion that can be drawn is that some of the ideal physically attributes for male and female have changed between the four different time periods.

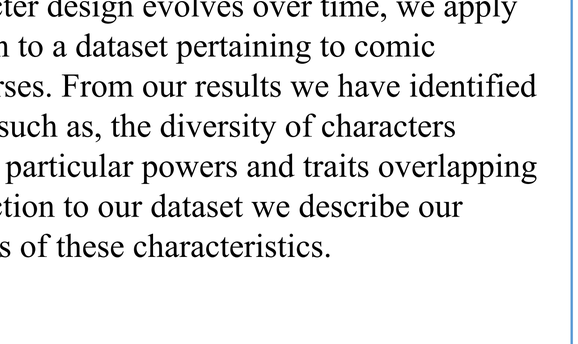

Fig. 1:Female Characters 1975-1990 vs before 1975

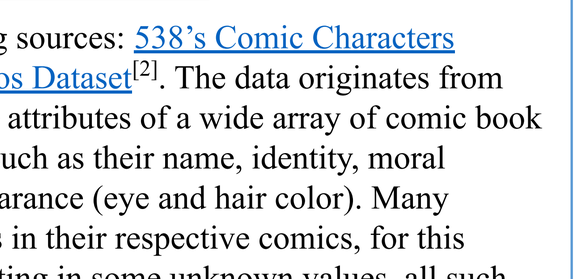

Fig. 2

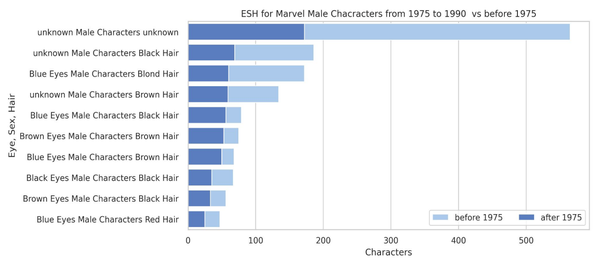

Fig. 3: Male Characters 1975-1990 vs before 1975

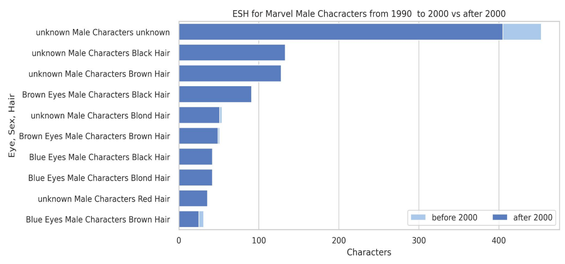

Fig. 4

Regression

We found that the datasets for characters' attributes also have the number of appearances, which indicates how many times one character has been on comic books (as of Sep 14'). Thus, we decided to perform some Machine Learning algorithms such as regressions, decision trees and boosting to predict the number of appearances of characters. As Table 1 and Table 2 show results when we apply K-fold Cross Validation on various methods using only the visual appearances we have, our following tables will show how using different columns, whether including or excluding will make our models better or worse.

Table 1: **DC** - Regression using ['eye', 'hair','sex']

Regressor	KFold Test MSE
Linear Reg.	7447.156
Ridge Reg.	7447.156
Logistic Reg.	8219.782
Lasso Reg.	7447.157
Random Forest	7323.726
Gradient Boost	7308.014
K-Neighbors Reg.	8921.632

Table 2: **Marvel** - Regression using ['eye', 'hair','sex']

Regressor	KFold Test MSE
Linear Reg.	9206.066
Ridge Reg.	9206.066
Logistic Reg.	9257.364
Lasso Reg.	9206.005
Random Forest	9171.763
Gradient Boost	9201.730
K-Neighbors Reg.	10307.132

Table 3: **DC**- Regression using ['id','align']

Regressor	KFold Test MSE
Linear Reg.	7617.247
Ridge Reg.	7617.247
Logistic Reg.	8219.772
Lasso Reg.	7617.247
Random Forest	7396.171
Gradient Boost	7384.140
K-Neighbors Reg.	14764.620

Table 4: **Marvel** - Regression using ['id','align']

Regressor	KFold Test MSE
Linear Reg.	9245.027
Ridge Reg.	9253.201
Logistic Reg.	9257.364
Lasso Reg.	9397.039
Random Forest	9245.027
Gradient Boost	9257.364
K-Neighbors Reg.	14764.620

Table 5: **DC**- Regression using ['id','align','eye','hair','sex','alive','year']

Regressor	KFold Test MSE
Linear Reg.	7018.653
Ridge Reg.	7018.653
Logistic Reg.	7848.797
Lasso Reg.	7018.652
Random Forest	6791.093
Gradient Boost	7114.129
K-Neighbors Reg.	7542.965

Table 6: **Marvel** - Regression using ['id','align','eye','hair','sex','alive','year']

Regressor	KFold Test MSE
Linear Reg.	9062.278
Ridge Reg.	9062.278
Logistic Reg.	9712.405
Lasso Reg.	9062.277
Random Forest	8297.963
Gradient Boost	7766.886
K-Neighbors Reg.	8581.421

Clustering

Our clusters are based on power combinations that appeared for various characters. It is a three-dimensional plot viewed from the top, as the larger a circle is, the "higher up" it is. We chose to focus on a few key examples that were particularly interesting, either because of greater separation between clusters or more homogeneity. Consider Figure 5, for k=2 clusters over Intelligence, Strength, Power the computed k-means cost was 34.767, and the k-center cost was 76.917, whereas for Figure 6. Speed, Durability, and Power, the k-means was 35.420 and the k-center cost was 67.241. This suggests that the average squared distance (k-means) for each set of clusters is roughly the same, but that the speed, durability, power clusters tended to be closer overall than the intelligence, strength, power clusters (k-centers). This could indicate that particular power combinations for heroes and villains tend toward specific groups, like Strong characters tending to be more durable or fast over intelligent, or Intelligent characters tending toward speed over power.

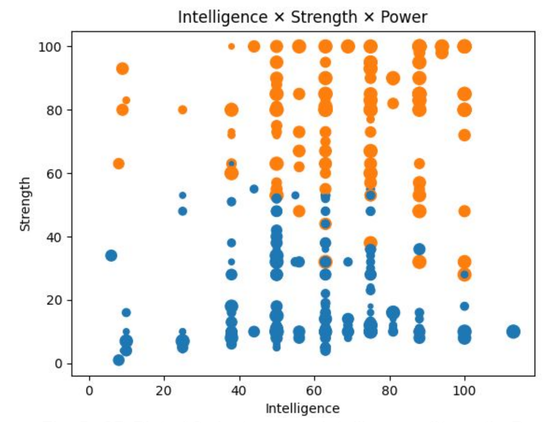

Fig. 5: 3D Plot of 2 clusters over Intelligence, Strength, Power

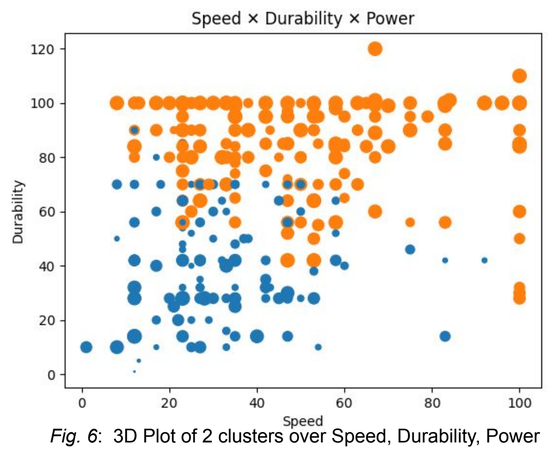

Fig. 6: 3D Plot of 2 clusters over Speed, Durability, Power

References

[1] FiveThirtyEight - *FiveThirtyEight Comic Characters Dataset*, 2019
Retreived From: https://www.kaggle.com/fivethirtyeight/fivethirtyeight-comic-characters-dataset

[2] Daniel R. - *Marvel Superheroes,* 2018
Retreived From: https://www.kaggle.com/dannielr/marvel-superheroes

[3] Pedregosa, F, Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., V, erplas, J., Passos, A., Cournapeau, D., Brucher, M., and Perrot, M., Duchesnay, E. *Scikit-learn: Machine Learning in Python* - vol. 12 Journal of Machine Learning Research, 2011.
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
https://scikit-learn.org/stable/modules/ensemble.html

GitHub Repo: https://github.com/baohuy251210/ComicCharactersMining