Marvel Characters Clustering
unsupervised learning

Sonia Petrini

[1] University of Milan
[2] Data Science and Economics

Abstract. In this work, our aim is to employ unsupervised learning techniques to identify hidden patterns among Marvel characters. Superheroes are representative of a society's desires and aspirations, thus being an interesting topic for social research. In order to shed light on such aspirations and societal ideals, we will exploit both Clustering Analysis and Categorical Principal Component Analysis.

Keywords: Unsupervised Learning · Clustering Analysis · Categorical PCA · Marvel · Superheroes

1 Introduction

Superheroes have the possibility to embody anything a human could never be. They are thus created as an echo of the desires of a particular society in a given historical moment. The greatest example of this phenomenon is Superman (belonging to the DC Universe): the first comic was released on 10th June 1938, while Europe was distraught by totalitaristic regimes. In that context, he was created to epitomize the idea that strength and power could be used for good, and to give the population the hope that there was someone, despite fictional, that could fight the dictators. The research purpose of this work is thus to identify these concepts and ideas, peculiar to the time and society in which Superheroes are created, by analysing the *"Marvel Characters Data"* dataset available on Kaggle.

2 Data

2.1 Presentation

The original dataset collects information about 16376 characters, and for each of them the following features are reported:

- page_id: unique identifier (str)
- name: character's name and planet of origin (str)
- urlslug: URL Slug of the character (str)
- ID: type of identity (secret, public...) (fct)
- ALIGN: alignment (good, evil...) (fct)
- EYE: eye color (fct)
- HAIR: hair color (fct)
- SEX: sex of the character (fct)
- GSM: sexual orientation (fct)
- ALIVE: whether character is alive (fct)
- APPEARANCES: number of appearances(str)
- FIRST.APPEARANCE: day of first appearance (str)
- Year: year of creation (str)

Only alive characters have been considered, as the status of alive of deceased is not a relevant feature for the purposes of the analysis. The URL Slug and the unique identifier have also been excluded, while all the other variables have been processed. In particular, empty values in GSM have been imputed to "heterosexual" characters, and some misspecified values for both SEX and GSM have been verified and properly adjusted by referencing the original Marvel Database. Moreover, the original complete dates of first appearance have been re-encoded in a new factor variable, levelled according to the season of first appearance (Summer, Autumn, Winter, Spring).
Given the social scope of this project, it is important to consider minorities. For this reason, given the strong unbalance between heterosexual and non heterosexual characters, the analysis will be performed on a sub sample with reduced GSM imbalance.

2.2 Summary Statistics

In order to get an understanding of the data, we report here the summary statistics for numerical variables, and the absolute and relative frequencies of categorical variables.

Table 1: Numerical Variables Summary

APPEARANCES	Year
Min. : 1.0	Min. :1941
1st Qu.: 35.5	1st Qu.:1963
Median : 234.0	Median :1981
Mean : 739.9	Mean :1984
3rd Qu.:1235.2	3rd Qu.:2004
Max. :4043.0	Max. :2013

From **Table 1** we can observe that the range of variability of the number of appearances is quite large, spanning from 1 to 4043 appearances. This is reasonable, as the dataset includes both the most famous characters and the ones which only appear in one comic or movie.
Moreover, we are also able to identify the time span considered in the dataset, which goes from 1941 to 2013. In fact, Marvel was founded in 1939.

Fig. 1: Histogram of Year of first appearance

By plotting the histogram of the variable *Year* (**Figure 1**), we notice that the majority of the considered superheroes were presented around 1960. This is the year in which Stanley Lieber (in art: Stan Lee) became the director of Marvel Comics, inaugurating the real success of the company. Many characters have also been presented in recent years.

In **Table 2**, we can see the levels of each factor variable included in the analysis, with the connected frequencies. Overall, the majority of the characters are good and male, and there are very few occurrences of transvestites and pansexuals.

Table 2: Categorical Variables Summary

	Values	Freqs (% of Valid)
ID	Public	30 (38.5%)
	No Dual	15 (19.2%)
	Secret	33 (42.3%)
ALIGN	Good	58 (74.4%)
	Neutral	13 (16.7%)
	Bad	7 (9.0%)
SEX	Female	28 (35.9%)
	Male	48 (61.5%)
	Genderfluid	2 (2.6%)
GSM	heterosexual	30 (38.5%)
	bisexual	15 (19.2%)
	transvestites	1 (1.3%)
	homosexual	30 (38.5%)
	pansexual	2 (2.6%)
FIRST.APPEARANCE	summer	22 (28.2%)
	spring	22 (28.2%)
	autumn	21 (26.9%)
	winter	13 (16.7%)

3 Hierarchical Clustering

Among the different existing clustering techniques, this first section will concern the employment of hierarchical methods. In particular, we will deal with bottom-up methods, which start from the single observations and then proceed to agglomerate them according to their similarity.

3.1 Dissimilarity Measure

Clustering is intended to find the optimal splitting, the one which could guarantee the maximum variance between the groups, while minimizing the variance within them. For this reason, clusters formation depends upon the similarity (or dissimilarity) matrix of the observations, which in turn depends on the features' values.
In this particular case, the dataset contains a combination of numerical and categorical features; therefore, an appropriate distance measure, which allows to combine these two types of variables, is required. For this purpose we will use the **Gower Distance**, which is based on the following similarity score measures:

$$\text{qualitative variables: } \; s_{ijk} = 1 - |x_{ik} - x_{jk}|/R_k$$

$$\text{quantitative variables: } s_{ijk} = 1\{x_{ik} = x_{jk}\}$$

The final measure is then computed as follows:

$$\text{Gower Distance: } S_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}\delta_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

where δ_{ijk} is a term allowing to control for the comparability of the observations and of the features: it takes value 1 when both variables are non missing.
As we can see, the qualitative variables weight depends on the distance between its value in the two observations. Thus, this weight will be biased towards the variables presenting a higher scale, leading to an incorrect measure of dissimilarity. For this reason, on both *Year* and *APPEARANCES* standardization is performed, in order to grant commensurability.

3.2 Clustering Methods

In this section, we are going to implement various hierarchical clustering techniques to observe the different dendrogram configurations that can be obtained, and we will select the most appropriate.

Ward's method
First, let's consider the most commonly used method, namely *Ward's method*. At each step, the algorithm computes the square of the difference between each observation and the mean of each cluster; after trying all the possible combinations, it creates the new clusters following a variance minimization criterion. In **Figure 2** we can see the resulting dendrogram. The peculiarity of hierarchical clustering consists in the possibility of a user-specified number of clusters: in this case, we decide to cut the dendrogram at a height equal to 1.5. In this way, the 3 formed clusters look compact and fairly balanced.

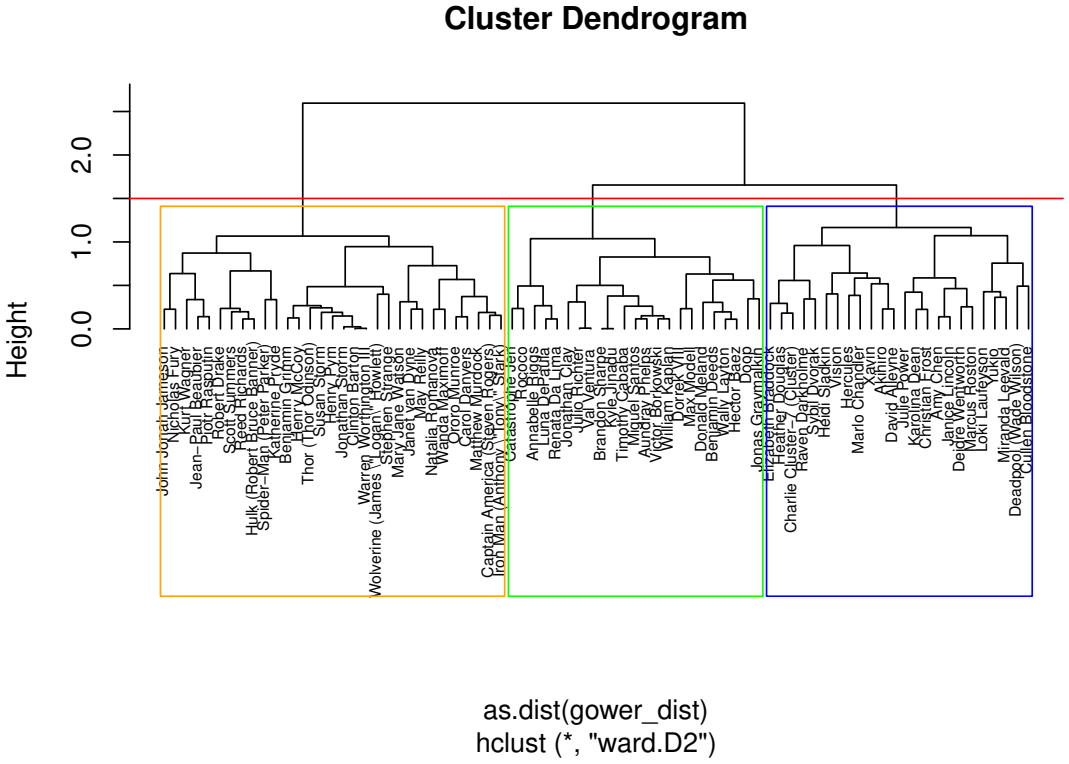

Fig. 2: Dendrogram obtained with Ward method

Complete-linkage method
The second explored technique is the *Complete-linkage method*. A number of different linkage-methods exist, each choosing a different aggregation system for all the similarities scores within the clusters. The "complete" method takes as intra-group dissimilarity the maximum one among those computed pairwise between each pair of clusters. In this case, we choose to cut the dendrogram at a height of 0.8, thus creating four different clusters.

The height variable on a dendrogram is of course the measure used to compute the dissimilarity among the clusters. Hence, a higher y value corresponds to a higher distance among the groups. Having this in mind, we notice that despite creating more compact groups, Ward's clusters are formed at a lower height comparing to those formed through the Complete method. This could mean that despite forming groups with a low within error sum of squares, Ward does not grant the creation of optimal clusters. We are going to verify whether this is true in the following sections.

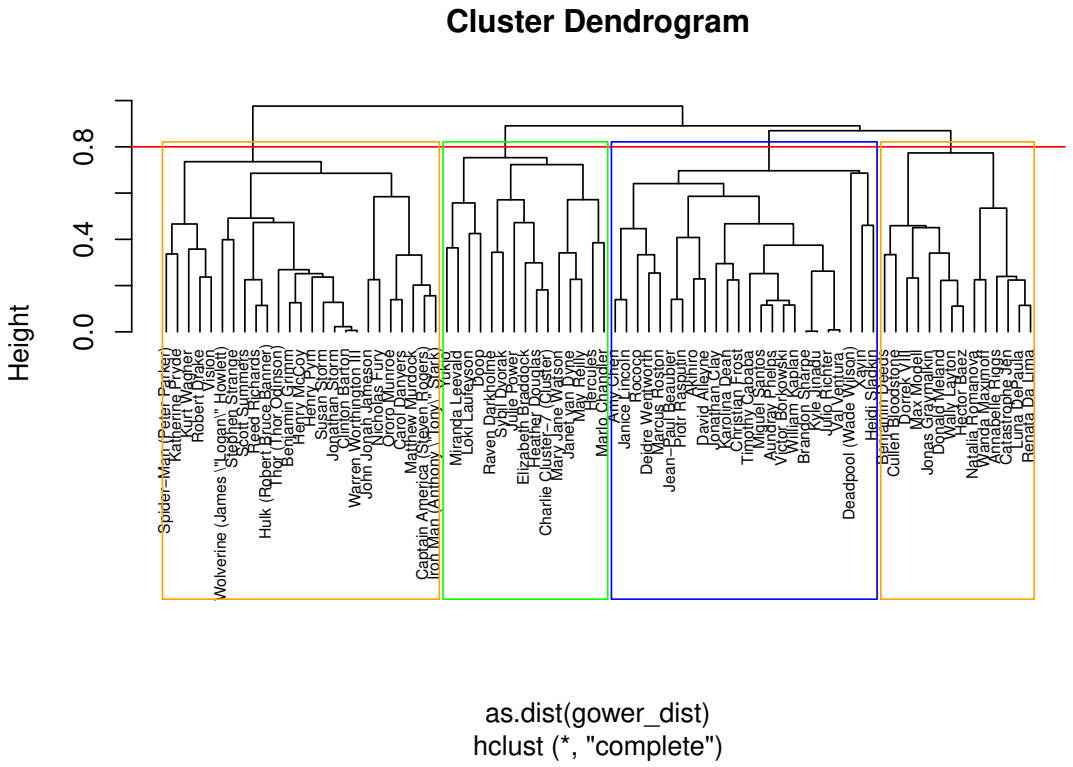

Fig. 3: Dendrogram obtained with Complete-linkage method

3.3 Goodness of Clustering

An important step of every analysis is to evaluate the obtained results, and in particular to compare them across different methods, to understand the pros and the drawbacks of each technique. Since we are dealing with unsupervised learning, and also we don't have any externally provided labels for the groups, we will rely on ***Internal Validation Measures***. We will consider two widely used measures: **Dunn's Index** and **Silhouette**. Both measures are based on the concepts of separation and compactness, in particular they are weighted ratios between these two properties, thus they are intended to be maximized.

Table 3: Internal Validation Measures

	Ward			Complete-linkage		
	Dunn	**Avg Sil**	**# Neg Sil**	**Dunn**	**Avg Sil**	**# Neg Sil**
3 clusters	0.152064	0.58	4	0.2059165	0.63	1
4 clusters	0.158867	0.59	1	0.2315025	0.58	3

In **Table 3** the resulting measures for the tested clusters configurations are displayed. We can notice the following:

- Ward's method yields the lowest scores for both configurations.
- The difference in performance between 3 and 4 clusters in Ward is very low compared to that in the Complete-linkage.
- While in Ward both measures are improving with an additional cluster, in Complete-linkage the measures go in opposite directions.

We are certainly going to opt for the Complete-linkage method, and among the two configurations, we decide to choose the one with 3 clusters. In fact, the decrease in Dunn's Index is offset by the increase in average silhouette, and the misplaced (negative silhouette value) character is only one, namely Heather Douglas.

In **Figure 4** we can see the Silhouette plot of the chosen configuration. Noticeably, the observation with negative silhouette belongs to group 2, and the third group is the only one with a silhouette below average.

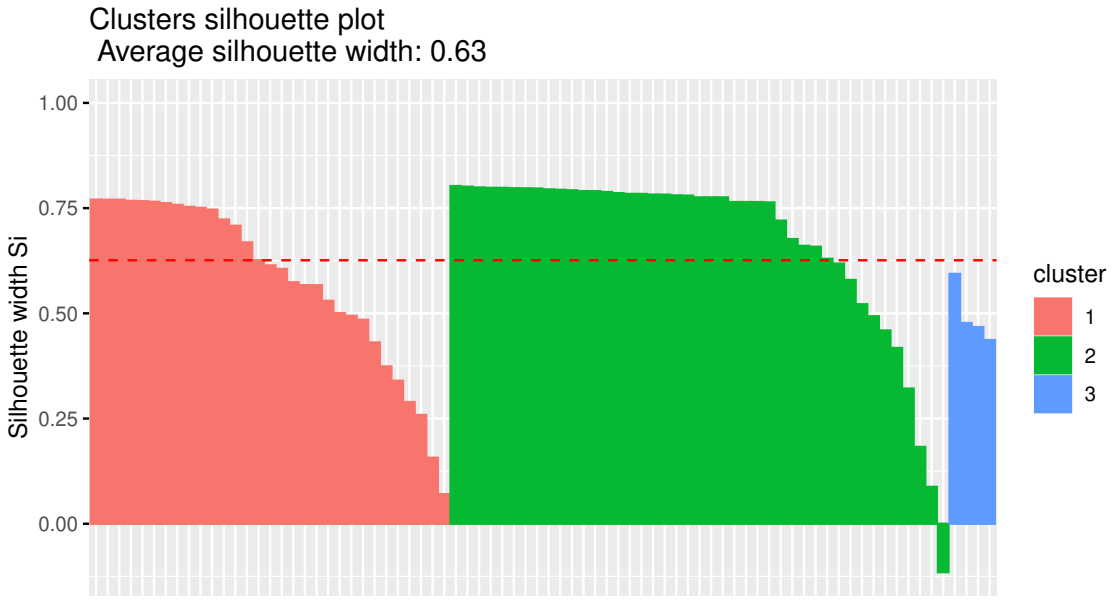

Fig. 4: Silhouette of Complete-linkage method with 3 clusters

4 Categorical Principal Component Analysis

This section will be devoted to the implementation of the Principal Component Analysis, with a dual purpose. First, the identification of the two main components will allow the visualization of the hierarchical clustering results. Second, it allows to rely on numerical variables, necessary for the implementation of non-hierarchical clustering.
Given the categorical nature of the majority of the features in the original dataset, we are going to implement a **Categorical PCA** through the *princals* function.

Table 4: Principal Components

	Component 1	**Component 2**	**Component 3**	**Component 4**
Eigenvalues	3.1943	1.6583	1.3604	1.1388
VAF	35.4921	18.4260	15.1158	12.6538
Cumulative VAF	35.4900	53.9200	69.0300	81.6900

In **Table 4** are reported the obtained components. As we can see, by retaining four components we can account for 81.7% of the original energy. Moreover, the eigenvalue linked to the fourth component is larger than 1, meaning that it is still able to explain more variance than the original variables to which it is related.

Now that we have chosen the number of components, we can assess their *loadings*, i.e. the correlations between the original variables and the unit-scaled components. This is useful to understand which variables are related to each other, and which underlying concepts are grasped by the components.

4.1 Components interpretation

The correlations are reported in **Table 5**, and can be better visualized in **Figure 5**. We notice the following:

- The number of appearances, year of creation, and sexual orientation are mainly spreading along the first component
- The season of first appearance, sex, hair and eye color are mainly spreading along the second component
- The third component is mainly related to alignment, but also to type of identity and eye color
- The fourth component is positively correlated with hair color and identity, and negatively with season of first appearance and eye color.

Table 5: Loadings

	Component 1	**Component 2**	**Component 3**	**Component 4**
GSM	0.936	-	-	-
APPEARANCES	-0.980	-	-	-
Year	0.947	0.132	-	-
HAIR	0.292	0.660	0.311	0.429
SEX	-0.157	0.717	0.351	-0.222
FIRST.APPEARANCE	0.458	-0.513	-0.106	-0.408
ALIGN	-	-0.381	0.818	-
ID	-0.271	-0.143	-0.449	0.662
EYE	-0.263	0.503	-0.483	-0.538

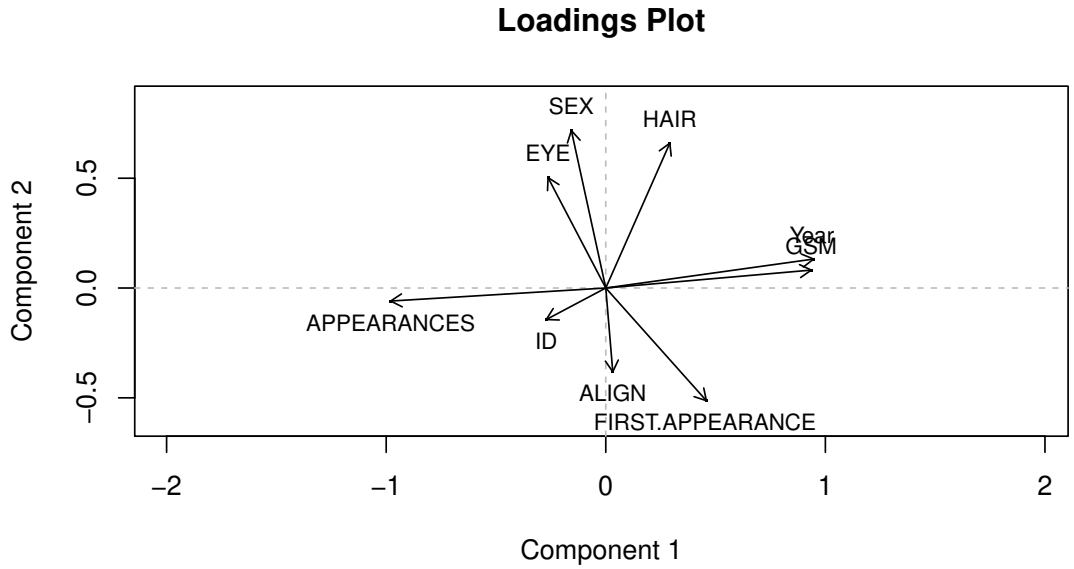

Fig. 5: Loading Plot of Categorical PCA

Among the above considerations, it is interesting to explore the concept related to the ***first component***. What we observe is a strong positive correlation between sexual orientation and year of appearance, and they both correlate negatively with the number of appearances. These relations are both depicted in **Figure 6**. First, the boxplots clearly show how the first characters to be created in the 60' were primarily heterosexuals. Through time, there was progressively more space for diversity among the heroes' sexual orientation. This reflects the development of the society, which has become more sensitive to sexual diversity in later years, and is becoming more and more inclusive as time goes by. Representation of minority groups is crucial from a social welfare point

of view, as it gives a voice to those individuals not belonging to the dominant classes.

Second, through the scatter plot we can easily visualize the negative correlation between the number of appearances and the year of first appearance. The ratio is probably dual: the first trivial explanation is that more recent heroes have had less time to appear in comics/movies. A second explanation may be related to the fact that the heroes presented around the 60' are the first, and thus most famous ones.

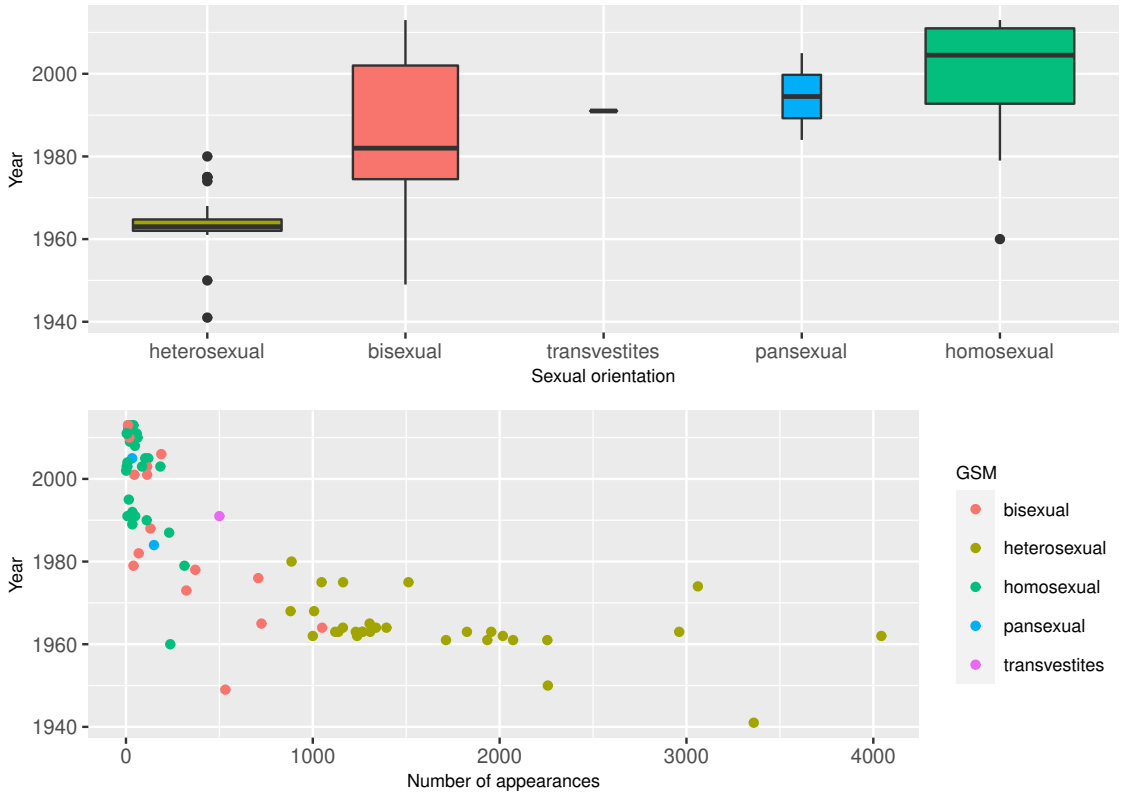

Fig. 6: Year boxplots by sexual orientation (upper), year and number of appearances colored by sexual orientation (lower).

4.2 Hierarchical Clustering Visualization

Now that we have identified the main directions of variability of the data, we can finally visualize the cluster configuration obtained in the previous sections. The model chosen was a 3-cluster partitioning with the Complete-linkage method. In **Figure 7** we can see the two dimensional representation over the first two principal components.

Noticeably, groups 1 and 3 are fairly separated along the first dimension, meaning that group 1 is on average the one including the most famous and "old" heroes, which are also probably heterosexual. Group 2 appears quite dispersed, but this is likely due to the fact that is is spreading along the other 2 principal directions. Yet, this latter group is the one displaying the lower values for component 2, meaning that it is probably including many females.

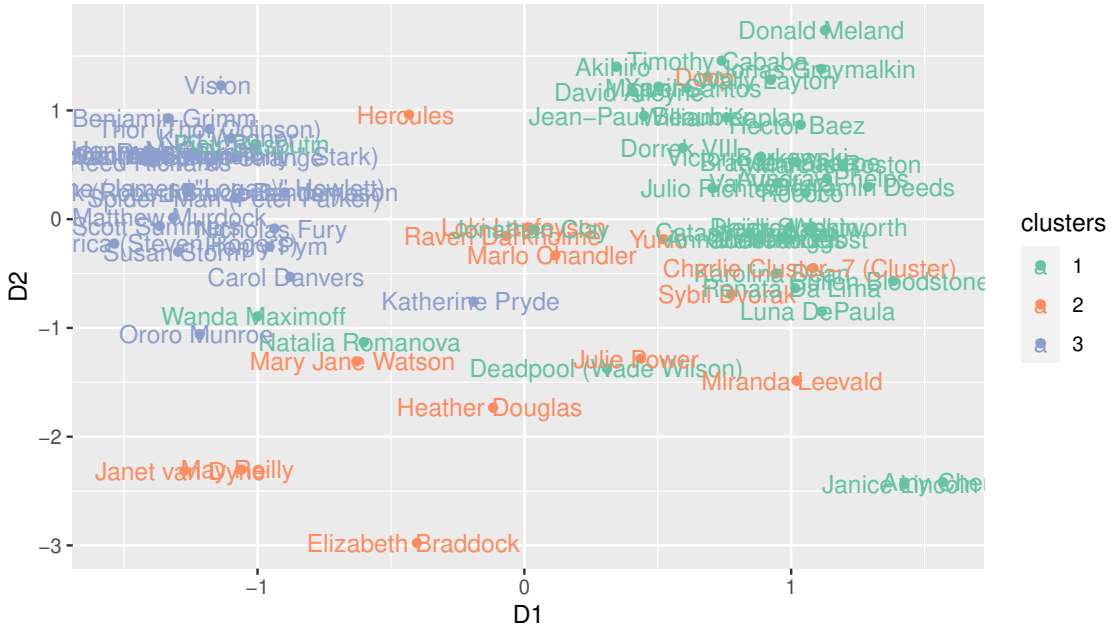

Fig. 7: Two dimensional representation of Complete-linkage clustering, over first two principal components.

5 K-Means Clustering

Thanks to Categorical PCA, we managed to obtain numerical components starting from categorical variables. With this new dataset, we are able to perform the **K-Means** algorithm, a technique for non-hierarchical clustering. We start by choosing 3 as the number of clusters to be created, and the obtained groups are shown in **Figure 8**, plotted over the three principal dimensions.

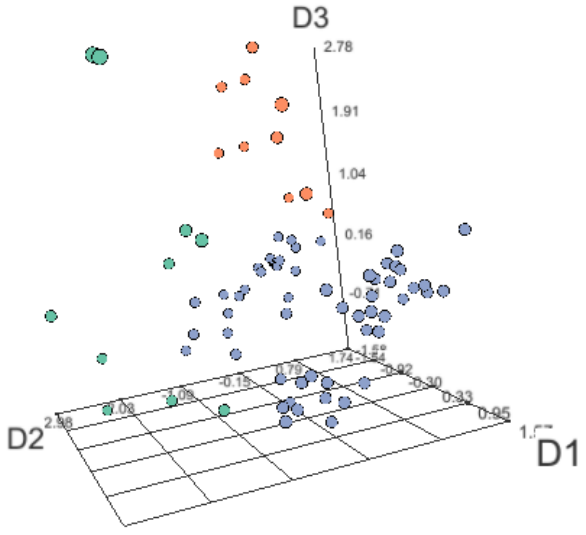

Fig. 8: Three dimensional representation of K-means clustering, over first three principal components.

The clusters are indeed well defined across the three dimensional space, as there is no overlapping. However, by applying the "elbow rule" to the within sum of squares against number of clusters plot we could choose both 3 and 4 as the optimal number of clusters, and we thus evaluate both

configurations with Dunn's Index and the Silhouette method. To allow the comparison with the previous findings, the 3-cluster Complete-linkage model is reported again.

Table 6: Internal Validation Measures

	K-means			Complete-linkage		
	Dunn	**Avg Sil**	**# Neg Sil**	**Dunn**	**Avg Sil**	**# Neg Sil**
3 clusters	0.1720901	0.26	10	0.2059165	0.63	1
4 clusters	0.2601682	0.3	8	0.2315025	0.58	3

From **Table 6**, we notice the following: the higher Dunn's Index (0.26) is obtained with 4-means clustering, but the average silhouette (0.3) is very low compared to the one obtained in the baseline model (0.63). Moreover, the misgrouped characters according to the silhouette methods are 8 in 4-means, compared to 1, as found in the benchmark. For these reasons, our final chosen model is the ***Complete-linkage Hierarchical Clustering with 3 clusters***.

6 Clusters Interpretation

To assess the identified clusters and interpret the results, we need to go back to the original unscaled data. However, since the majority of the examined variables is not numerical, we are mainly going to exploit percentage bar-charts as a visualization tool.

In **Figure 9** we can observe *identity* and *alignment* conditioned to the clusters. Noticeably, group 3 is almost entirely composed of "good" characters, is the only group without "bad" characters, and also the one with the highest percentage of public identities. Group 1 and 2 both show a certain degree of inside variability, with group 1 being the one with the highest number of characters with a secret identity.

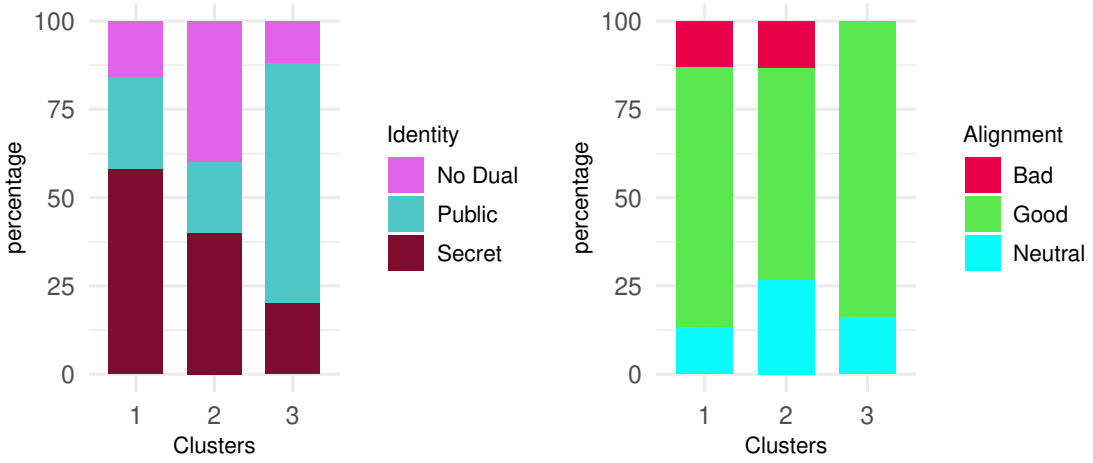

Fig. 9: Type of Identity and Alignment by cluster

Figure 10 depicts the characters' *sex* and *sexual orientation*. We can immediately notice how group 3 includes the majority of males and no genderfluid heroes, and is entirely composed of heterosexuals. On the other hand, group 2 is mainly populated by females and bisexuals, while group 1 includes all the homosexuals and the transvestites, being the more diverse. Pansexuals are included in both group 1 and 2.

The eye and hair colors are showed in **Figure 11**. Remarkably, cluster 3 is the one with the majority of the blue eyed characters, while group 2 has the wider range of hair colors.

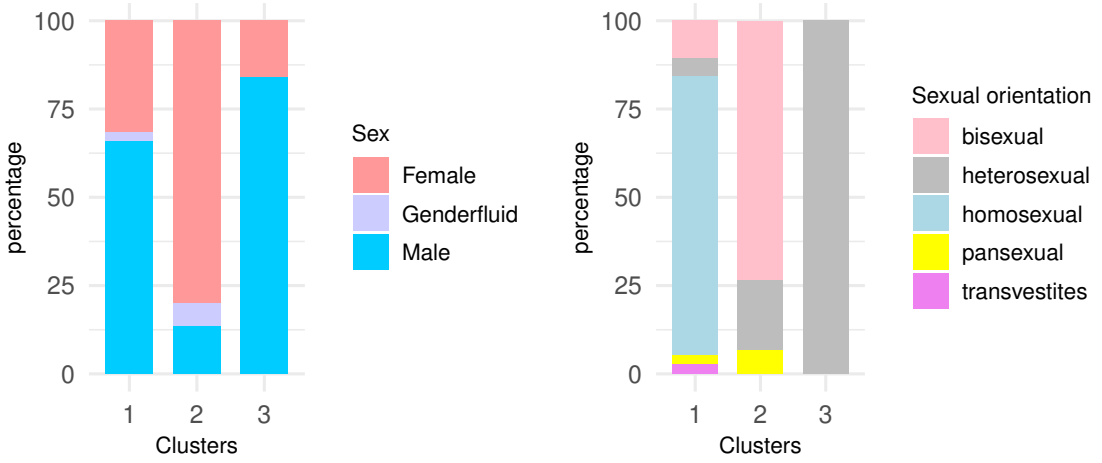

Fig. 10: Sex and Sexual orientation by cluster

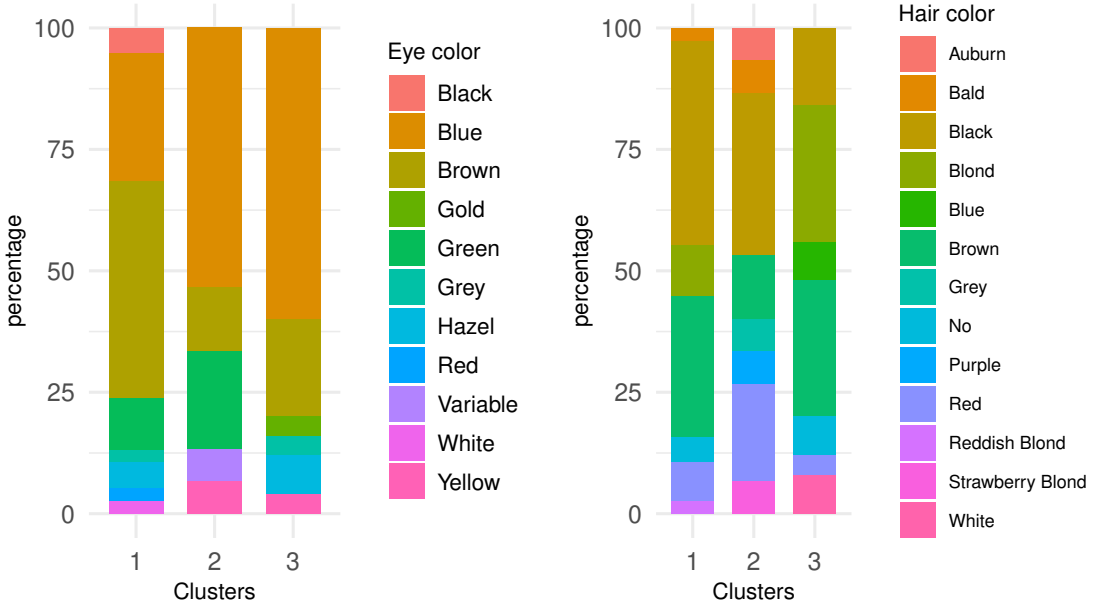

Fig. 11: Eye and Hair color by cluster

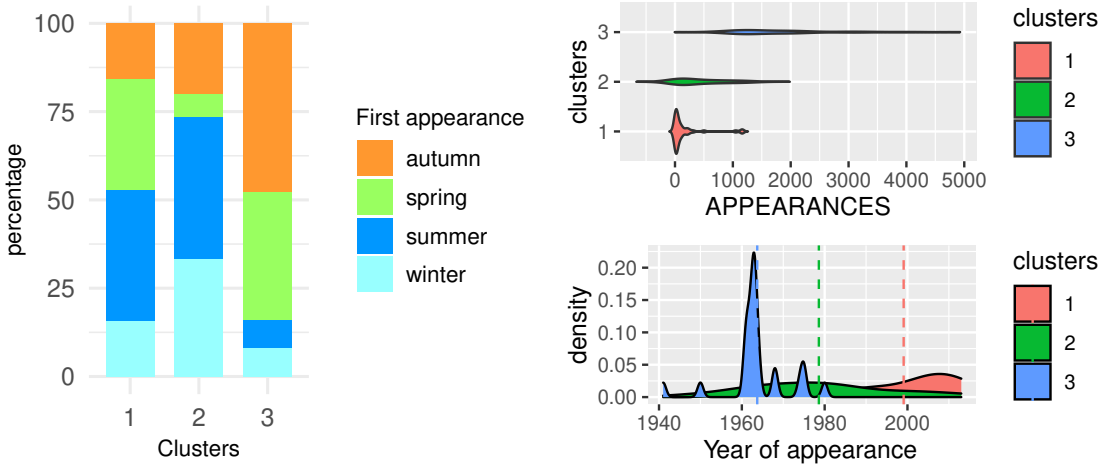

Fig. 12: Season of appearance, year of appearance, and number of appearances by cluster

Finally, **Figure 12** shows the variables related to the number and the moment of appearance. Despite not being particularly relevant for the research purpose of this work, the season of first appearance shows a great between group variability. More importantly, the two numerical variables accounting for the number and year of appearance are quite significant, and allow us to get the whole picture of the groups' meaning.

- **Group 1**: This is the group with the lowest number of appearances, all mainly occurring after the beginning of 2000th. As seen above, it is the most diverse group in terms of sexual orientation.
  Heroes belonging to this group are: *Xavin* (genderfluid, pansexual), *Deadpool* (male, transvestites), *Natalia Romanova* (female, bisexual).
- **Group 2**: Characters in this group don't have a specific time of appearance, and their number of appearances ranges from 0 to 2000. It is the group with the majority of the females and the bisexuals, and it also shows the greatest variability in hair color.
  Heroes belonging to this group are: *Loki Laufeyson* (genderfluid, bisexual), *Hercules* (male, bisexual).
- **Group 3**: This is the group of the famous and well-known heroes, introduced around the 60's and with the highest number of appearances. As seen above, it is entirely composed by heterosexuals, and it mainly comprehends "good" male characters with blue eyes.
  Heroes belonging to this group are: *Wolverine* (male, heterosexual), *Iron Man* (male, heterosexual), *Matthew Murdock* (male, heterosexual)

7 Conclusions

In this work various unsupervised learning techniques have been performed and compared in order to identify an appropriate cluster configuration. The aim was to find hidden patterns among characters belonging to the Marvel Universe, and indeed some socially relevant patterns have been identified. Through the Complete-linkage method we were able to identify 3 clusters.

Given the interpretation presented in the last section, we found the characters to represent different social classes. The first, most famous heroes are predominantly male, heterosexuals, and with blue eyes, all traits peculiar of the dominant class, which were considered to represent normality till recent decades. With the passing of time, new heroes with more diverse sexualities and traits have been introduced. This is a clear sign of an evolving society, which is becoming more inclusive and aware of the importance of minorities representation in the media.

8 Appendix

The R code used to perform the analysis is available on Github.