Marvel Database

Sytze Van Herck, Cara Pelsmaekers, Carole Urlings

2015-2016

Motivation

For this project we chose to use a database from the publishing company *Marvel Comics.* The reasons for this choice are twofold. The first reason is somewhat personal, because we're all Marvel fans we figured we could take a subject that interests us all and motivates us to keep working. At the same time this ensures that anyone else reading or listening to our findings is at least mildly entertained. The second reason is somewhat more serious. While still keeping things fun, we're still trying to reveal important societal trends as they are displayed through popular culture, for example developments in sexual orientation and gender balance.

Dataset

As we started on the assignment, we had a lot of difficulties with the first dataset we chose. For starters, it was difficult to get all the information from the Marvel Developers portal. First of all we were limited to 3000 results per day, also having to query only 100 results at a time. This returned JSON files, which we converted to CSV files in order for us to use the results in Tableau. The data was also quite limited and chaotic when looking into the series, comics and events in which characters occurred.

We then moved to a dataset found on Github containing more detailed information on the characters. Since our genderize API did not perform well in determining the gender of creators, we decided to drop the creators dataset all together and focused on visualizing the dataset of characters as much as possible. Finally, we also came across a dataset created by the website FiveThirtyEight.com for their comparison between Marvel and DC characters. This dataset also described sexual orientation and provided two more types of gender, namely agender and genderfluid.

Visualization design and implementation

	Marvel wiki	Marvel vs. DC
Bubble chart	gender & popularity, status, weight	gender & popularity
Bar chart	gender & popularity, status, height	gender & popularity, sexual orientation
Boxplot	weight, height	
Word Cloud	powers & abilities, occupation, categories	
Timeline		sexuality
Line chart		status
Tree map		eyes, hair

All our visualizations were created using Tableau. We created bubble charts to show gender and popularity, by means of counting in how many comics a character appeared and color coding gender. Another bubble chart for the marvel wiki dataset shows the status of characters meaning alive, deceased or revived where the size of the bubbles still indicates popularity. We showed the status for the other dataset in a line chart to really show the proportion of dead and alive characters over time. This visualization also gave some historical insights. For both weight and height of characters, we created a boxplot to show the distribution per gender. We also created a bubble chart showing weight per character, while a bar chart indicates height. The word clouds concerning powers and abilities, occupation and categories of characters were generated in voyant-tools.org and helped us gain insight into the descriptions. Furthermore we found a bar chart visualized the proportions in sexual orientations best, although we also showed sexuality in a time line to see the evolution over time. Finally, we mapped eye and hair color per gender in tree maps to show the most popular colors.

Since the Tableau dashboards didn't provide the overview we were truly after, we also implemented all our visualizations in a website. We started from the bootstrap template called Greyscale, which is a one page theme. Because we had two different datasets we decided on creating two pages that were easily accessible using a switch button in our navigation bar. Other than implementing our visualizations by embedding the code Tableau provided, we also wrote a small about section describing our team as well as a section explaining and providing links to our datasets. In the end, we decided not to use Gephi for creating social networks based on creators that worked together or characters that appeared in the same comics, because the data simply didn't lend itself to create such visualizations.

Because in data visualisation it is of course important to keep aesthetics of your visualisations in mind, we tried to integrate this into our website as well, keeping the colors uniform over the datasets and including some fun graphics throughout too. We bore the five design aesthetics in mind as discussed in the course of data visualisations.
Furthermore we applied *Guidance for the encoding of quantitative, ordinal and categorical data* by Mackinlay (1986). To visualise the popularity and status of characters, which are both quantitative and categorical data, we created a color coded bubble chart, including volume and color hue as suggested by the guidelines. The overview of our visualisations is obtained by the navigation bar at the top of our webpage, but we decided to show one visualisation at a time with details on demand for each one. .

Reflection
We did not aim to create a scientific visualisation, but instead visualized information with this dataset. Initially we just wanted to visualize the characters included in the Marvel dataset, based on things like popularity, powers, status and gender. However, as we were making these graphs, we realised the power of such visualizations and

truly mapped out interesting societal information. This way, we found some historical events truly influenced the content of the comics and even the creation of new characters. These findings would've been invisible, if not for the visualizations. In other words, there is real power in visualizing data and you have to be cautious with such graphs as wellIf we were to do things differently, we might start off by looking more critically at the data we wanted to use. Finding the right dataset from the start would have saved us a lot of time. Although we had some ideas on what type of visualisations to create from the start, we should have narrowed it down earlier on. We wasted a lot of time trying to add gender to almost 6000 creators as well as finding all the comics per character and creator in order to come up with our social network in Gephi.

Effort

Though the visualizations were relatively straightforward and easily adapted to fit our needs, the formatting of the dataset itself was pretty hard. Scraping information from Wikipedia and matching it with the data from the Marvel API took a lot of time. On top of that, transforming the JSONs to CSV also resulted in some errors, which we had to manually sort out. Though Tableau is fun and easy tool, it's feature of embedding within a website is not transparent and badly documented. Especially the Public version does not allow complete customization, as for example the menu bar always remains at the bottom of the graph. The size - or better dimensions - were also difficult to adapt to our wishes. While you can easily determine the border of the graph to be a certain number of pixels, the content would still be oversized and scroll bars would appear. So we spent a lot of time - and effort - trying out different dimensions and settings.

We distributed the creation of visualisations throughout the group based on who knew more of what tool and had the time to work on it. However, the content of what needed to be visualized was discussed together beforehand. Presentations were always created together and the website was made by Cara and Sytze. We all worked on this report.