The three data sources I am going to use for this project are related to Marvel comic characters. Most of the data provided by the sources are of descriptive nature. The collections of information on comic characters are valuable sources of data for fans and people interested in the details about the characters from Marvel comics.

The first data source – the flat csv file (marvel-wikia-data.csv) was downloaded from <https://www.kaggle.com/fivethirtyeight/fivethirtyeight-comic-characters-dataset>. According to the creator of the data set, it was scraped from <https://marvel.fandom.com/wiki/Marvel_Database>.

The file contains 16376 observations. The number of variables is 12.

Description of the variables:

|  |  |
| --- | --- |
| variable | description |
| page\_id | id of character’s page |
| name | character’s name |
| urlslug | url related to the character |
| ID | identity status |
| ALIGN | good, bad neutral character |
| EYE | eye color |
| HAIR | hair color |
| SEX | gender |
| GSM | gender or sex minority |
| ALIVE | status: alive or deceased |
| APPEARANCES | the number of appearances in the comic books |
| FIRST APPEARNCE | the month and year of first appearance |
| YEAR | the year of first appearance |

The second data source - website <https://www.marvel.com/characters> includes the information about 2587 Marvel characters. Each character page contains information on the following: **name, image, height, weight, eyes, hair, universe, aliases, education, place of origin, relatives.** These categories represent the biography of the characters. This data is visible in the source page of the character page, and I assume will be scraped in the future. I am not sure yet how it could be done.

The third data source that I found for this project is a Comic Vine API (<https://comicvine.gamespot.com/api/documentation>). It contains 17 variables about a great number of characters by different publishers.

Description of the variables:

|  |  |
| --- | --- |
| variables | description |
| aliases | aliases |
| api\_detail\_url | API for the character details resource |
| count\_of\_issue\_appearances | total number of issues it appears in |
| date\_added | date of being added to Comic Vine |
| date\_last\_updated | data of being updated on Comic Vine |
| deck | summary about the character |
| description | description of the character |
| first\_appeared\_in\_issue | Issue of first appearance |
| id | character’s id |
| image | character’s image(url) |
| birth | date of birth |
| gender | male, female, other |
| name | character’s name |
| origin | the origin of the character: human, alien and etc. |
| publisher | primary publisher |
| real\_name | character’s real name |
| site\_detail\_url | url with character’s information on GiantBomb |

I was able to send the request and retrieve the data. It is possible to filter by a lot of variables but unfortunately retrieving resources about characters and filtering characters by publisher is not available yet (I tried it but no luck either with the publisher’s name or ID , and went through the forum, which confirmed this feature still needs to be added). The total number of characters is around 137 300. I was able to loop through 13730 pages and get json with information on 100 comic characters limit for each page, upload all of it into a final data frame and filter the data frame by the publisher post query, remove the duplicates. The total of 18769 observations of comic characters was retrieved and added into a data frame for Marvel publisher. It was a time-consuming procedure.

I think these datasets could be merged on the name of characters (the relationship). Character’s name is present in the data from all of the data sources and truly represents the point which the rest information about the characters is associated with. Before I do that, I believe I will have to clean the data by checking the headers, identifying the outliers, duplicates, verify and validate problematic values, make sure the data is in the right format, standardize the data, visualize it, compute summary statistics, subset the data. I believe that the final dataset should have the main physical and personal traits characteristics of the comic character and its presence in the comics world (name, eyes, hair, gender, good/bad/neutral character, its origin, summary, universe, aliases, and information on its appearance in comics).