Gathering and Transforming Data about Marvel Comic Characters

The focus of this project was data wrangling. The data used came from three sources: a flat csv file (downloaded from kaggle.com), API (ComicVine API), and a website (marvel.com). After the three sources have been cleaned and prepared, they were merged and loaded into in SQLite database.

What I learned working on this project:

* Sourcing data from a website and API.
  1. Originally, I had to scrape all of the URL-addresses with the information about Marvel characters from marvel.com sitemap, and then scrape the data by looping through the URL-addresses. When I realized that all of the data wasn’t in a table, or located on one page, writing out what I needed to accomplish helped me come up with the code. It was time-consuming, but also rewarding.
  2. When pulling the data from the API I came across two problems: 1) pagination (only 100 observations were allowed per request), and 2) the fact that filtering by publisher wasn’t possible when sending a request (ComicVine API has data on comic characters by various publishers but no way to filter by them). I had to find the way to extract all of the 137000 characters and filter them by publisher after the data was received in the JSON format before I created a final data set. That was a lot of learning, and in some way my first complex coding mostly from scratch where I had to come up with ideas similar to that of scraping from the website where the solution wasn’t that obvious.
* Working with regular expressions and using lambdas were other important skills that I learned and were of great use when preprocessing the data.
* Since the data that was identified as a relationship among the three datasets was categorical, I had to perform fuzzy matching, which helped with removing the duplicates.
* I have never worked with databases before and loading the data into one and creating visualizations from the data stored in the data base was a great way to learn about SQLite, SQL, and visualizations with Python.

With the help of this project, I was able to merge three data sets from different sources, apply data wrangling techniques, explore and practice the concepts learned during this term. As a result, I created a data frame with 12 variables, cleaned and loaded it into SQLite.