

Wrangle report

Wrangling the data was challenging and exciting. To accomplish this task I was required to gather, assess and clean data. And each of these steps was not so easy if you compare it with the previous Udacity projects.

Data collection could be not a trivial task at all. In this project, I gathered data from many sources. I loaded data from a CSV file that I downloaded manually from Udacity website and it was quite easy. More challenging was to download a file from the internet using requests library. It was a new way for me and it took me some time to read the documentation about the library. The hardest part was to use Twitter's API to get retweets and favorites counts. I had to create a developer account and create new API keys. Using tweepy library and the API keys, I was able to collect json data for almost all tweets from the CSV file data and store it to file. Then I used pandas to load json data to DataFrame and filter out all required information.

Assessing and cleaning data took me a lot of time and effort. All the data had so many issues so that I even didn't know where to start. Following the guides from Udacity, I started to collect issues and clean them at the same time. While I was cleaning I noticed new issues and repeated this process numerous times. Several steps of cleaning required much time and energy to find an appropriate solution. The hardest steps were cleaning names and ratings where I had to use a lot of code and review data many times.

Overall, it was a great journey to data wrangling. It showed me how difficult data analyst work could be. I learned a bunch of new things and skills and I'm ready to go further.