Wrangle report

Introduction.

The purpose of this project is to strengthen data wrangling skills during the Data Analyst Nanodegree programm. The dataset represented is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

Project steps.

Gather

There three parts to work with:

- Twitter archive file: the twitter_archive_enhanced.csv was available for downloading
- Images with predictions of some neural network. This file (image_predictions.tsv) was on Udacity serveres and was downloaded with help of the `requests` library.
- Twitter API & JSON: Using Twitter API and python libraries get the data and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Assess

Personally, I used programmatic way of assessing data – pandas functions and methods.

After getting acquainted with the data, I divided process into two subparts: quality and tidiness.

Clean

In this part of data wrangling the main idea for me was to identify issues (duplicates, inconsistency in the data etc.) and work on it. The most interesting part was uniting all twitter massages into entire text document and cleaning it - deleting links, numbers, characters, using regular expressions.

Visualize

After all of above was done, the data was ready to be visualized. Apart from plotting bars, I decided to make a cloud of words using dog picture as a background, it helped to understand the most frequently used words and reveal some funny words and phrases.

Conclusion.

Obviously, data wrangling is one of the core skills for anyone who works with data. Technologies has grown up, so now you can get data with just a few movements, using different techniques and programming languages. To sum up, I enhanced my knowledge in following:

- Working with API, understanding API
- Working with JSON file format
- Creating nice visuals with wordcount
- In general, working on this data wrangling project is similar to ETL process, except for the last procedure (LOAD) as we didn't need to load our data to any database system.