Wrangle Report

With no doubt, knowing how to query APIs, clean the extracted data and analyse it, is certainly a skill that is highly useful for a data analyst. As such, this project was designed to provide the relevant tools and knowledge to the Data Analyst Udacity students.

More specifically, the WeRateDogs downloaded their Twitter archive and sent it to Udacity. This archive contained basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017 and were given to Udacity students. The goal of the project was to wrangle the dataset and create interesting and trustworthy analyses and visualizations. In order to achieve that we had to requery the Twitter API and collect additional data for the provided WeRateDogs tweets. Furthermore, another dataset which contained image predictions of a trained neural network of the aforementioned tweets needed to be downloaded programmatically from the Udacity website and cleaned as such.

The wrangling part of the project was splitted in 3 parts. In the first part we had to gather all three datasets and import each one of them in a pandas DataFrame. We used Tweepy library for quering the Twitter API and the requests library for downloading the image predictions dataset from Udacity.

In the second part a visual and programmatic assessment of the three dataframes were done in order to identify data quality issues and tidiness issues that needed to be solved. Eight quality issues were found and five tidiness issues i.e. retweets present in the DataFrame (we only wanted original tweets with images), erroneous datatypes, erroneous dog ratings, four variables that should be merged in one column etc.

The third part of the project was all about cleaning the issues found in the second part. Several pandas methods were used (i.e. isnin(), info(), .drop(), .astype(), .loc(), .sample() .str.title(), etc.) and many loops and functions were created to succesfully address the quality and tidiness issues. The final step of the cleaning process was the creation of a master DataFrame that contained all the cleaned variables from the three initial dataframes. This Dataframe was then saved in a .csv format.