Wrangle Report

Background

The purpose of this project was to put in practice what we have learned about data wrangling part of the Data Analyst Nanodegree. The dataset comes from the tweets of the user WeRateDogs, a twitter account that rates the photos of dogs. The ratings are usually above 10/10.

The main aspects of data wrangling are

- Gathering data
- Assessing data
- Cleaning data

Gathering data

For this project data was gathered from three different loctions;

- Twitter file: this twitter_archieve_enhanced.csv file contains an archieve of several 1000s Tweets from WeRateDogs
- Dog prediction file: this file contains the predicion of what breed the dog is based on a neuel network aligotherm. The predictions were based on the three best dog breeds and their associated confidence level. This file was downloaded programmaitically using requests library.
- Twitter API – Using the tweet_id of the files above, the Twitter API was queried and stored each tweets JSON data.

All of these were loaded into pandas dataframes.

Assessig data

After the initial gathering stage, next was the assessment of data which was done in two stages; visually and programmatically. Whilst Jupyter Notebooks is good for manipuliting data, an initial assessment of data was completed in Excel Spreadsheets. Following this, the data was then assessed programmatically using different pandas functions such as .info, .value_counts, .describle, .sampe, .head, groupby, .duplicated, etc.

After this notes were then written up with issues separated based on quality and tidyiness data. The tiydiness format comes from the concept of tidy data by Hadley Wickham. I read one of his papers to understand tidy data more for this project.

Cleaning data

This was the hardest part of the project. For each part to be cleaned there are three parts;

- Define : this is where we state the problem to be fixed and how we can fix it.
- Code: this is the code that can be impemented to fix problem.
- Test: this is an operation/code to submit to make sure or code to fix the problem truly worked.

Before starting this step a copy of the dataframe was made as we don't want to edit on raw data.

When I made a mistake with the cleaning data, I went back to the origioanl data and made another copy and started again. Some of the cleaning operations took several iterations in oder to get right. For example there are 3 predictions of dogs and I wanted to keep the highest confidence value prediction as the main prediction. In order to tidy this dataset the origonal stages of dog (such as doggo and pupper), which was four columns was combiend into one column.

Conclusions

- I found the initial visual assessment easier to see how the data looked and had a point of reference of where to go programmeritally. I think I would find it hard to do this without a visual assessment first as viewing a dataframe in Jupyter Notebooks is clunky.
- I learned a lot about how APIs can be used in place of web scraping.
- By runnng each cleaning operation into the steps of define-code-test, traceabiliy can be improved and too make sure the Notebook does not get too busy.