Wrangling Report

By: Amr Seleem

1. Background

This report briefly presents my wrangling efforts. The aim of this project is to apply what we have learned in data wrangling, assessing, and cleaning from the Udacity Professional Data Analysis Nanodegree program. The dataset is about a Twitter account @dog_rates, also known as WeRateDogs. It is a Twitter account that rates people's dogs with funny comments. These ratings almost always have a denominator of 10. The project is comprised of the following parts:

- Gathering data
- Assessing data
- Cleaning data

Gathering Data

The data are gathered from three difference sources as follows:

- First, the tweet archive of Twitter account @dog_rates, it is given by Udacity to download. twitter_archive_enhanced.csv
- Second, tweet image predictions, which is a file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using ag given URL
- Third, Twitter API & JSON: by using the tweet IDs in the Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data

Assessing data

Once the three tables were obtained, I assessed the data as following:

- First visually, by seeing the three entire dataframes separate in Jupyter Notebook and checking the csv files in Excel.
- Second programmatically, by applying different pandas functions (e.g. info, value_counts, sample, duplicated, describe, etc). Then I collected the issues encountered in quality and tidiness issues.

From the assessment, the following quality and tidinesss issues came out.

Tidiness issues:

1. all of the three data sources have tweet_id,so this can help combining the three together in one dataset
2. drop all unncessary columns that are not relevant.

Quality issues:

3. remove the retweets

4. fix columns datatype: tweet_id and timestamp
5. remove the tweets that have no image
6. fix the numerators and denominators ratings
7. fix the outliers and high ratings
8. combine the numberator and denominator
9. combine the dog names to one column
10. fix missing values of NONE in column name

Cleaning Data

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described in the assess section. First I created a copy of the three original dataframes.

I used python and searching over the internet i.e. stackoverflow for references and possible guidance to resolve the aforementioned issues.

Overall, I learned a lot about how to use library commands efficiently to clean data and store it. The analysis results are presented in Act_report.pdf