Wrangling and analysis data

Introduction:

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled 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

Data Gathering

I gathered data from 3 sources, stored in separate files:
1. WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
2. The image predictions file, programmatically downloaded from the Udacity servers.
3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy
library. The favourite_count and retweet_count were extracted programmatically from this file.

Assessing data

Once the three tables were obtained I assessed the data as following:
1)Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel.
2) Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc).

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 and very helpful step was to create a copy of the three original dataframes. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original. Could create another copy of the dataframes and continue working on the cleaning part

There were a couple of cleaning steps that were very challenging. One of them was twitter archive try to clean the table remove the retweet in the text and try to get the value from the 4 columns (doggo,popper,pupper,floofer) into one table and remove unwanted data

And also to two different methods in clean data manually which was hard and the anther one was programmatically

One more thing was making the data tidy data by changing types of columns to be able to complain them

Conclusion

Data wrangling is a core skill that whoever handles data should be familiar with.
I have used Python programming language and some of its packages. There are several advantages of this tool (as compared to e.g. Excel)

One can re-run analysis automatically every period. Thus, we could actually re-run the dog analysis every month with much less effort now because I have set it up once.
Handling, assessing, cleaning and visualizing of data is possible programmatically using code.

At the end funny pic.