WRANGLING ACT REPORT

By: Naseer Faheem – Sep 2018

INTRODUCTION:

In this project I am going to gather, asses, clean and analyze real world data using multiple sources on WeRateDogs Twitter archive. WeRateDogs which is a popular Twitter page with more than 7.28 million followers as of Sep/2018. WeRateDogs rate people's dogs with a humorous comment about the dog. Their rating is also very humorous as they use a denominator of 10.

WeRageDogs downloaded their Twitter archive and sent it to Udacity via email which about 5000 entries. Udacity has done some preliminary sorting and gave me a parsed dataset with about 2356 tweets. I use multiple analysis skills to gather extra information and clean this dataset to make it into one big master file.

This project takes advantage of the following packages:

- Pandas
- Numpy
- Requests
- Tweepy
- Json
- Os
- Io
- Matplotlib

This is by far one of my top favorite projects where I had to do everything myself and showcase my data analysis abilities. I have also included a list of references that I used during the project at the end of the code file.

This project is divided into 4 main sections;

1. Gathering
2. Assessing
3. Cleaning
4. Analysis and Visualization

GATHERING

There are a total of 3 different sources used in this project:

1. Enhanced Twitter Archive: This is the complied file shared by Udacity. I had to manually download this and read it in python. There are 2356 rows in this dataset that will be assessed and cleaned for the analysis purposes.

2. Image Predictions File: This file contains 2075 rows which is produced by Udacity. This file includes a result from a neural network analysis that was used to classify each dog's breed from its jpg_url. I used python's request library to download this file programmatically and save it on my local machine.

3. Additional Data via Twitter API ('Twitter_Json): Both of the above datasets lacked vital information such as retweet_count and favorite_count for each tweet. I used the Twitter API and Tweepy to download those myself and put it a new dataframe.

ASSESSING

After gathering the three datasets, I starting assessing each table visually to see if I could find any quality or tidiness issues. Then, I programmatically assessed each table and listed all the assessments in at the bottom of the assessment section of the python file. I divided my assessment into two subsections; quality and tidiness. I tried to follow Hadley Wikcham's paper on tidy data and labeled my data accordingly.

I was able to find most of the tidiness issues by my visual assessment of tables individually, however, the majority of quality issues were revealed programmatically.

CLEANING

After having a full list of assessments, I used a Define, Code and Test strategy to address each of the assessments individually. First, I started cleaning the twitter_archive table since it was the main dataset that I had, then I cleaned image predictions and finally I cleaned the data that I downloaded through Twitter API. I finally merged all three tables using Panda's merge function. I finished the cleaning task by removing all null values created after merging the datasets.

After the dataset was merged, I saved it in a CSV file named twitter_archive_master.csv.

ANALYSIS AND VISUALIZATION

Now that I had everything cleaned and organized, I tried to do a quick explanatory data analysis to see the trend within the dataset. Finally, I visualized 3 features of the dataset using Panda's plotting functions.