Data Wrangling Report

1. Gathering Data

I will be wrangling (fixing the data quality and tidiness issues) the tweet archive dataset, also known as WeRateDogs. This dataset has 2356 basic tweet data from November 2015 to August 2017. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Gather Twitter archive CSV file

I downloaded the WeRateDogs Twitter archive manually as twitter_archive_enhanced.csv through the followin link provided by Udacity (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archiveenhanced/twitter-archive-enhanced.csv).

Gather tweet image predictions

Using Python's Requests library I downloaded the tweet image predictions file hosted on Udacity's servers programmatically and saved it to df_predictions.tsv file locally. Then, I imported this file into a Python Pandas dataframe.

Gather data from Twitter API

First, I contacted tweeter in order to have access for tweeter API. After I received the keys, I accessed the entire data for every tweet from Twitter API and stored every tweet's entire set of JSON data in a file called tweet_json.txt file. Created a dataframe tweet_df from this JSON including only tweet_id, retweet_count, favorite_count and display_text_range data.

2. Assessing Data

After gathering each of the above pieces of data, we will assess them in two different ways, which are visually, and programmatically for quality and tidiness issues. Exploring data set for two things: data quality issues and lack of tidiness

• Quality Issues related to content such as missing, duplicate, or incorrect data

• Untidy Data has specific structural issues

I run some invistinatons on all the tree tablels by looking using methods like (info, sample, head, …)

3. Cleaning Data

Most of the quality and tidiness issues were related to df table and df_ predections, I created a copy of the 3 tables.

I created copy of the the three tables and named them with same names adding (.copy). For each of the following quality or tidiness issue, I performed the programmatic data cleaning process in three stages, which are (Define, Code, and Test).

Cleaning of Quality issues

- remove duplicate retweets
- delete missing tweet_id(s) in (image predictions)
- dropp df table without any duplicates (retweets)
- convert feilds that contain both date and time parts such as timestamp to datetime data type
- remove the 59 records that have no information
- get rid of unnecessary html tags in source column in place of utility name e.g. <a href=""http://twitter.com/download/iphone""rel=""nofollow"">Twitter for iPhone bout images ('expanded_urls' is NaN)
- delete short displayable text
- convert rating_numerator column which has values less than 10
- convert rating_denominator column that has values other than 10
- edit some dog names starting with lowercase characters to uppercase
- change the frequent incorrect dog name to None

Tidiness

- merge doggo, floofer, pupper and puppo columns in df table into one column named "stage"
- Create a new variable – ‘stage’ to show the four dog stages, drop the four columns, and fill the empty with NaN

df_predictions table

- Columns 'names such as p1, p2 are not understandable.
- Dog breeds prediction involve both uppercase and lowercase for the first letter.
- jpg_url, img_num columns from df_predictions table are not needed for me since I am not going to use them in the analysis.

4. Data Storing

After the completion of the cleaning process, I stored the df_clean DataFrame   to CSV. Then I reread them in a reverse to make sure they have saved to CSV properly.

5. Analyzing and Visualizing Data

I analyzed the df dataframe and then generated two visualizations

- Analyze and Visualize the Distribution of Dog Stage
- Analyze and Visualize The Distribution of Source