Wrangle Report

This is a brief description of my efforts to wrangle the WeRateDogs database acquired from the Twitter account carrying the same name. I start by importing the libraries that I will use during the project.

Then I start with loading the three pieces of data from different sources. First from a CSV file of the twitter archive of the account activity using pandas read_csv. Then download the second file of Algorithm image prediction using the requests library. Finally, I imported the third piece from a JSON file. For simplicity, I used the file prepared by the Udacity team and didn't download the data myself from Twitter API. After that, I tried to assess the data to identify errors. Both visually and programmatically. For visual assessment, it can be done in Microsoft excel and through pandas functions like df.head(), df.tail() and df.sample(). By programming, it can be done through functions like df.info() and df.describe(). I discovered several issues in both quality and tidiness summaries in the below bullet points.

Quality Issues

Twitter Archive table

- 1- timestamp in string not in datetime format
- 2- +0000 in timestamp column entries
- 5- letters in name column
- 6- unnecessary rows with retweeted data
- 7- unnecessary rows with reply data

Image Prediction table

- 8- p1, p2 and p3 names are misleading
- (3-4) 2356 rows in Twitter archive, 2075 in image prediction and 2354 in tweets data. (missing data)

Tweets Data

- 9- id column name should be tweet_id

Tidiness Issues

Archive table

- 1- Unnecessary columns i.e in_reply_to_status_id, in_reply_to_user_id empty and retweeted_status_id retweeted_status_user_id and retweeted_status_timestamp.

Tweet Data Table

- 2- Unnecessary columns like 'created at' which is similar to timestamp.

After that, I tried to clean most of these issues. I changed the string into datetime format and removed the hour notation. Then removed the unnecessary columns in each table. I filtered the data for original tweets and removed every retweet. And I changed the column names on several

occasions to make them easier to use and understand. Finally, I reported several insights supported with graphical visualizations for example most popular dog name and how favorite count and retweet count changed over time.