Stephanie Irvine

Wrangle Report

The wrangling for this project was broken down into the following steps:

- Gathering
- Assessing
- Cleaning

Gathering:

Udacity provided the WeRateDogs enhanced Twitter archive which contained basic tweet data for 2356 tweets. This was a csv file and was downloaded manually from the Udacity classroom. This was read into a pandas DataFrame 'df1'.

The next file to gather was the image predictions text file which contained the predictions of dog breeds present in the tweet according to a neural network. This was downloaded programmatically using the Requests library. This was then read into a pandas DataFrame, 'df2'.

The final file to gather for this analysis was using a Twitter API to retrieve tweet_id, retweets, and favourites. I also retrieved timestamp but later dropped this. After signing up for a Twitter Developer account I attempted to query Twitter's API however only successfully created a DataFrame 'df_api' containing 2331 rows. Since there were only 2331 rows it seemed like some were missing therefore for the purposes of Udacity submission I manually downloaded the Twitter API file from the classroom and created a pandas DataFrame 'df3'.

Assessing:

The next stage in the wrangling process was assessing the DataFrames. These were assessed both visually and programmatically in order to identify both quality and tidiness issues. Assessing the data, I followed two steps; detecting the issues and then documenting these in the Jupyter notebook. Throughout the Data Quality Dimensions of Completeness, Validity, Accuracy, and Consistency were considered. Visual assessment was achieved by using Excel,

as well as displaying the DataFrame in the Jupyter notebook. Programmatic assessment was then conducted for each of the DataFrames by using the following Pandas functions:

- df.head()
- df.tail()
- df.sample(5)
- df.info()
- df.describe()
- value_counts()

The following issues were noted from the Assess stage:

Quality

`df1` **- Twitter Archive Table**

- Considerable missing data in 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp' columns with less than 200 non-null values.
- Unclear column headers and unhelpful for analysis: 'in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp'.
- In 'expanded_urls' column there are some links that are given twice in the same row
- In the 'name' column there are 745 named 'None'. These should instead be null values
- In the 'doggo', 'floofer', 'pupper', and 'puppo' columns 'None' is also written
- In the 'name' column there are incorrect names such as 'a' which appears to be a mistake.
- Erroneous datatypes ('timestamp' is an object and should be a datetime data type, 'tweet id' is an integer and should be a string since no calculations will be performed, rating_numerator and rating_denominator should be floats instead of integers as a ratings sometimes include a decimal)
- Numerator_rating and Denominator_rating have min value of 0 and numerator_rating has a max value of 1776.

`df2` **- Image Predictions Table**

- Erroneous datatypes ('tweet id' is an integer and should be a string since no calculations will be performed)
- Consistency issue: Underscores in the dog breeds and sometimes lower case and sometimes upper case
- Predictions that are not dog types for example fruits/paper towels
- There is missing data since there are only 2075 rows compared to the 2356 from `df1`
- Image Number column not accurate or meaningful

`df3` **- Twitter API Data Table**

- Erroneous datatypes (as before)

Tidiness

`df1` **- Twitter Archive**

- 4 separate columns for each 'doggo', 'floofer', 'pupper', and 'puppo' i.e. dog stage is in 4 separate columns and could be 1

`df2` **- Image Predictions**

- Several different columns to indicate dog breed

`df3` **- Twitter API Data**

- Can merge all of these DataFrames together as they are all extremely closely related and can be merged on tweet_id

Cleaning

This data was then cleaned as documented in ‘wrangle_act.ipynb’.