REPORT ON DATA WRANGLING

By Priyanka Krishnan

Introduction:

For the Wrangle and Analyze Data project, I have wrangled (and analyzed and visualized) the tweet archive dataset 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.

As part of this project, I have had to do the below steps:

- Data wrangling
    - Gathering data
    - Assessing data - quality & tidiness
    - Cleaning data
- Storing, analyzing, and visualizing wrangled data
- Reporting on
    - data wrangling efforts and
    - data analyses and visualizations

This report pertains to my data wrangling efforts.

Data Wrangling:

a. Gathering data:

For this project, I had to gather data from 3 different sources:

- Downloaded the WeRateDogs Twitter archive data file manually by clicking the following link provided by Udacity: twitter_archive_enhanced.csv. Upload this file into the jupyter notebook using the upload button. Imported this dataset into a dataframe using 'pd.read_csv'.
- The tweet image predictions dataset, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers URL. Used requests library to programmatically download the tsv file. Once the file had been downloaded, imported this dataset into a dataframe using 'pd.read_csv'.
- Got additional information like retweet count and favorite count from each tweet. Using tweet IDs in the WeRateDogs Twitter archive dataset, queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called 'tweet_json.txt' file. Each tweet's JSON data is written to its own line. Read this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

b. Assessing data:

The gathered data was assessed visually and programmatically for both quality and tidiness issues.

My assessment summary is stated below:

- **Quality Issues:**
  - archive data:
    i. 'in_reply_to_status_id' and 'in_reply_to_user_id' have missing data. Only 78 records present out of 2356. 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp' have missing data. Only 181 records present out of 2356. As per project instructions we do not need these columns, only original tweet information is required.
    ii. 'expanded_urls' has some missing data. Only 2297 records present out of 2356.
    iii. Erroneous datatypes (for 'timestamp' and 'tweet_id')
    iv. Column 'floofer' should be 'floof'.
    v. Columns 'rating_numerator' and 'rating_denominator' have some outliers. Numerators above 20 and less than 1 will be considered invalid/outlier. The denominator above 10 or below 10 will be considered invalid/outlier.
    vi. Duplicated 'expanded_urls' present. Not necessary to clean this for my analysis.
    vii. Missing names under 'None' and incorrect names like 'a', 'an', 'the' and so on. This will not be cleaned as part of my analysis.
    viii. Missing dog stage information under columns 'floof', 'doggo','puppo' and 'pupper'. This will not be cleaned as part of my analysis
    ix. 'source' column has no significant information.
  - image prediction data:
    i. 'p1', 'p2' and 'p3' has some names in upper case whereas some are in lower case.
    ii. Erroneous datatype for 'tweet_id'
    iii. 'img_num' makes no logical sense to me. Can't interpret this column. This will not be cleaned as part of my analysis.
    iv. If 'p1_dog', 'p2_dog' and 'p3_dog' are all false, chances are the entry is not for a dog. This will not be cleaned as part of my analysis.
  - additional tweet details data:
    i. id_str needs to change to tweet_id to help merge the different datasets.

- **Tidiness Issues:**
  - The dog stage is one variable and hence should form a single column
  - All three tables can be merged on 'tweet_id'

c. Cleaning data:

After assessment, the data was cleaned using Define, Code and Test method.

As part of cleaning, the following steps were processed:

- Created copies of all 3 datasets.
  - twt_arch_clean
  - img_pred_clean
  - twt_add_dets_clean
- Deleted the rows that have values in retweet columns and replies columns. Once the rows are deleted, deleted the retweet and replies columns too from twt_arch_clean.
- Deleted the rows that have missing values in 'expanded_urls' columns using dropna method.
- Changed the dtype of column timestamp from object to datetime using to_datetime()
- Changed the dtype of column tweet_id from int64 to object using the astype()
- Updated column name from 'floofer' to 'floof' using rename method.
- Deleted any record that has a numerator above 20 and less than 1. Also deleted any record that has a denominator less than 10 or greater than 10. Created a column called rating, the value in it was populated by divding rating_numerator by rating_denominator.
- Converted all the names under p1, p2 and p3 using str.lower()
- Updated column name from 'id_str' to 'tweet_id' using rename method.
- Columns doggo, floofer, pupper and puppo were combined into one column called stage. Then deleted the 4 parent columns.
- Merged twt_arch_clean and twt_add_dets_clean on the column tweet_id into one table using the pd.merge() method
- Merged merge_twt_data and img_pred_clean on the column tweet_id into one single table using the pd.merge() method