Data Wrangling Report

INTRODUCTION

This project is a data wrangling project that involves the tweet archive of Twitter user @dog_rates, also known as WeRateDogs, which is a Twitter account that rates people's dogs with a humorous comment about the dog. The project mainly focuses on fixing that quality and tidiness issues of the data using python.

DATA GATHERING

1. The WeRateDogs Twitter archive, which was provided by Udacity. I read the file ('twitter_archive_enhanced.csv) into the notebook using pd. read_csv ()
2. The tweet image predictions, 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 and was downloaded programmatically using the **Requests** library and the following URL: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**
3. Tweet data, which was provided by Udacity in the form of two files: twitter_api.py and tweet_json.txt. Using the tweet IDs in the Twitter archive, 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. I created a data frame from this json including tweet_id, retweet_count, favorite_count, display_text_range data.

DATA ASSESSING

This involves inspecting the dataset for two things: Data quality issues and Tidiness issues.

Tidiness Issues

These are issues that affect the structural integrity of data.

1. One variable in 4 columns (doggo, floofer, pupper, puppo)

2. The tweet_data and twitter_archive tables should be merged.

Quality Issues

<twitter_archive>

1. Missing values in the following columns: in_reply_to_status_id, in_reply_user_id, retweeted_status_id, retweeted_status_user_id, expanded_urls, timestamp
2. Erroneous data types - in_reply_to_status_id, in_reply_user_id, retweeted_status_id, retweeted_status_user_id, tweet_id columns(should be str), timestamp and retweeted_status_timestamp columns(need to be split into data and time), dog status column (should be categorical)
3. Some exceptional high values in rating_numerator column, leading to high rating which might be inaccurate
4. Rating_denominator other than the standard value of 10
5. Some names in the name column are invalid data such as: quite, a, an, such, etc.
6. There are six incorrectly entered rating_numerator values.

<image_predictions>

1. Column names are not descriptive enough
2. Use of _ instead of space in p1, p2, p3 column values.
3. Values are sometimes uppercase and other times lowercase in p1, p2, p3 columns
4. Image duplicate predictions present for jpg_url with different tweet ids and other data the same.

<tweet_data>

Tweet id title is different in different tables. id here, tweet_id in others.

DATA CLEANING

- Tidiness issue 1: Combine the doggo, floofer, pupper and puppo columns into one column - dog_status and drop the unnecessary columns after the combination.
- Tidiness issue 2: Merge the tables twitter_archive_clean and tweet_data_clean on tweet_id column

- Quality issue 1: Remove unnecessary columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_timestamp. Fill in the missing values of expanded_urls column in twitter_archive_clean table using tweet_id
- Quality issue 2: Change the datatype of timestamp to datetime and remove the observations where tweet_id matches retweeted_status_id
- Quality issue 3: Correct the observations with incorrect denominators and change the numerator rating accordingly
- Quality issue 4: Replace all the instances of name column having invalid data with NaN
- Quality issue 5: Check for the correct words in the text column to correctly interpret the dog_status and then reform the dog_status using these words
- Quality issue 6: Replace the values in the source column to human-readable text
- Quality issue 7: Rename the columns to descriptive names
- Quality issue 8: Replace the '_' with ' ' and change the values to uppercase in p1, p2, p3.
- Quality issue 9: Drop all the observations where image_url is duplicated