WRANGLING REPORT

Introduction

The dataset that was provided for this project is a twitter archive data for a particular user @dog_rates, a page which is used to rate pictures and videos of dogs out of ten.

My task was to go through a systematic Wrangling processes and provide three insights and provided one visualization.

The Wrangling processes are:

1. Gathering data,,
2. Accessing data and
3. Cleaning data.

Gathering Data

I made worked on the twitter archive data to gather data from three different sources:

1. The WeRateDogs Twitter Archive. The data was in Comma separated version (CSV) and named **twitter-archive-enhanced.csv**. I downloaded it from Udacity student portal as provided by the instructors. The data contains 2356 rows and 17 columns of raw data. I read the data and saved as df.
2. The tweet image predictions was downloaded programmatically from the Udacity student portal and this came in the form of tab separated version (TSV). I load the data and saved as image_predictions_df. This dataset has 2075 rows and 12 columns.
3. Also, the third dataset that was used tweet_json was gathered by downloading the Twitter's Json data. The process of getting the data was a bit daunting as I had to use Tweepy library. I created a function to extract columns. The dataset has 4 columns and 2354 rows.

ASSESSING DATA

After the process of gathering the data. I started assessing the data for quality and tidiness issues and came up with the listed issues below for different tables.

Quality Issues/solutions

Twitter Archive Table

1. Convert timestamp to datetime and remove +0000. I converted the timestamp datatype to datatime and removed the extra +0000 from al the columns.
2. Wrong datatype used in the some columns. I converted all the columns to the appropriate datatype that will give a more sensible exploration.
3. Some dog names are mistyped (as a, an the, and such). Converted the error names to none.
4. Some columns (timestamp to tweet_timestamp, text to tweet_text, rating_numerator to dog_ratings", name" to dog_names) will be renamed so it can make more sense. I had to rename some columns so it ca be more presentable.

5. Remove columns that wont' be used for data anaysis and vusialisation. Removed columns that are not going to be used for analysis.
6. Source column is not a string datatype but in HTML-format. I changed the source column to string datatype as it was in an HTML-string format.

Twitter API Table

7. Wrong datatype for tweet_id column. Tweet_id was converted to string datatype.
8. Some Missing tweets. I actually did not touch this.

Image Prediction Table

9. Some values in p1, p2, and p3 columns have their first letter in capital letter which makes our date to be inconsistent. I changed all the first letters in p1,p2, and p3 columns to be capital letters

Tidiness issues

Twitter Archive Table

1. doggo, floofer, pupper and puppo columns in twitter-archive table should be in one column and probably named as **dog_stage.** I concatenated doggo, floofer, pupper, and puppo columns and named as dog_stage.

Image Prediction

2. The image prediction table should be joined with the twitter archive table. I joined image prediction with the twitter archive table.

Twitter API Table

3. The twitter Api with the following columns(retweet_count, favorite_count, followers_count) should be joined with twitter archive table. I joined the retweet_count, favorite_count, followers_count with the twitter archive table.

Storing The Cleaned Data To CSV

After completing the three steps of wrangling. I then saved the cleaned data into

df_clean.to_csv('twitter_archive_master.csv')

image_predictions_clean.to_csv('image_prediction_master.csv')

twitter_api_clean.to_csv('tweet_json_master.csv')