Wrangle Report

Goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

The data:

Enhanced Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. **Image predictions** a table full of image predictions alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction. **tweet_json** contains retweet count and favorite count.

This project was completed on the Udacity project workspace. However, the reports were created using Microsoft word.

The project was following The wrangling process :

1-Data Gathering

2-Assessing data

3-Cleaning data

Data Gathering :

Three data was gathered from different sources, first **Enhanced Twitter archive** was downloaded manually and uploaded to the workspace, then the **Image predictions** were downloaded programmatically using the requests library, and for **The Additional data from the Twitter API** due to some technical issue, the tweet_json file was downloaded programmatically using the requests library, and then was read line by line into a pandas DataFrame and later saved in a tweet_data.csv.

Assessing data:

The data was assesssd visually and programmtically. The following findings were concluded:

Quality issues

twitter_archive table

1.Retweet

2.Tweet_id is an integer instead of string

3.Erroneous datatypes(timestamp)

4.Rating other animals

5.Invalid names(None,a,an)

6.Uncorrect rating

7.Unnecessary columns (source,in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)

image_prediction table

1.Some names uppercase while other lowercase

tweet_df table

1.Missing rows (2354 instead of 2356)

Tidiness issues

1. Four variables columns in Twitter_archive table.

2.Image_prodiction and tweet_df should be part of the Twitter_archive table.

Cleaning data:

All mentioned above issues were cleaned in following ways :

1- Melt doggo,floofer,pupper and poppp columns to dogs_stage columns.then drop the four columns.
2- Merge all table.
3- Drop retweet and unnecessary columns.
4- Change tweet_id type to string and timestamp type to DateTime type.
5- Capitalize name,p1,p2 and p3 columns
6- Replace None,a and an with Null
7- Correct rating tweet manually and drop other animals rating