Project: Wrangling and Analyze Data

Reporting: Wrangle_report

Data Gathering

This project is intended to wrangle and analyse a tweet archive of twitter user @dogrates. Using jupyter notebook, I started the project by importing all the necessary packages that will be needed. I used three data sets in this project. I manually downloaded the *'twitter-archive-enhanced.csv'* data from Udacity classroom. I created a folder called *'image_predictions'* and downloaded image predictions data from Udacity server using request library. I then wrote the downloaded data into a file named *'image-predictions.tsv'* and stored it in the *'image_predictions'* folder. I dowloaded the last dataset (*tweet-json.txt*) manually from Udacity classroom and read it line by line into an empty list using json library, I then converted the list to a pandas DataFrame called *'twitter_df'*.

Assessing and Cleaning

I carried out visual and programmatic assessment on the three datasets and detected quality and tidiness issues. Below is a representation of the issues I detected and the steps I took to clean them.

Twitter archive data

Quality issues	Cleaning
Erroneous datatype(tweet_id)	Datatype of tweet_id will be converted to string.
Invalid rating_numerators and rating_denominators. Some rating numerators and denominators are less than 10.	Rating numerators and rating denominators will not be changed.
Unusual dog names in name such as *Stu','just', 'life', 'light', 'mad', 'my', 'not', 'old'*, etc. in name column.	Unusual dog names will be replaced with *'NaN'*. Nulls represented as 'None' will be changed to *'NaN'*
Nulls represented as 'None' in name column	Nulls represented as 'None' will be changed to *'No_name'*

timestamp and retweeted_status_timestamp should be datetime and not object.	timestamp and retweeted_status_timestamp will be converted to datetime
Nulls in doggo, floofer, pupper and puppo are represented by *'None'* instead of *'NaN'*.	Nulls in doggo, floofer, pupper and puppo represented as 'None' will be converted to *'NaN'*
Data types for retweeted_status_id should be string and not float.	Datatype will be converted to string
Missing data in these columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)	These columns will not be used in the analysis and therefore will be dropped.
Tidiness issues	**Cleaning**
Separate columns for dog stages	Dog stage columns will be combined to form one column

Twitter data

Quality issues	**Cleaning**
Count of tweets did not match that of twitter_archive_df. Therefore, there are missing tweets data.	Data will be represented as nulls
Tidiness issues	**Cleaning**
twitter_df table is separate from the main table which is twitter_archive_df	Merge tables

Image prediction

Quality issues	**Cleaning**
There are missing images data for the tweets.	Data will be represented as nulls

Erroneous datatype(tweet_id)	Datatype of tweet_id will be converted to string.
Tidiness issues	**Cleaning**
Image_predictions_df is separated from the main table twitter_archive_df	Merge tables

Storing data

I stored the cleaned data containing 22 columns and 2073 rows into a csv file name d ***'twitter_archive_master.csv'.***