Wrangle Report for WeRateDogs Twitter Data

1.Collecting Data

The dataset that I will be investigating is the archive of tweets from Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a funny comment about the dog. The WeRateDogs Twitter account has over 4 million subscribers and received international media coverage. This means that there is a large amount of data generated by this Twitter account. There are three sets of data that make up the WeRateDogs twitter account data and they include : twitter_archive data, image_predictions data and twitter_api data.The efforts to extract the data from these different sources is documented as follows

- **Collecting the Twitter archive data**
  I downloaded the twitter_archive_enhanced.csv file from the udacity website and read the data using pandas read_csv function. I then imported it into the df_archive dataframe
- **Collecting the image predictions data**
  I downloaded the image_predictions.tsv data set from the udacity website using python's requests library and imported the data to the img_pred dataframe
- **Collecting the twitter api data**
  I downloaded the twitter_api data from the udacity website and using the pandas read_json function. I then imported the data to the df_api dataframe

2.Assessing Data

Through programmatic and visual assessment, I was able to identify 8 quality and 2 tidiness issues from the data sets

Programmatic assessment(Quality Issues)

I employed the use of python functions: info(), value_counts(), dtypes, shape, duplicated() and isnull() to identify various issues in the dataframes. These helpful functions helped in identifying the 8 quality issues as listed below:

- There were some missing values.
- Timestamp was in the wrong data fromat.
- Tweet_id was of the wrong data type. It should be a string.
- There were some unnecessary columns.
- Some column names were very vague in description.
- The first letter of some names were in lowercase.
- The api dataframe had 2354 entries while the archive data set had 2356 entries(Which meant that there were some missing values in the retweet and favourite count).

- There were some rating denominators that were more than the standard limit of 10.

Visual assessment(Tidiness issues)

By opening the twitter_archive_enhanced.csv and image_predictions.tsv in google sheets, I was able to visually identify some flaws in the data sets. The Tidiness issues observed were:

- The 3 dataframes were scattered and a merge wass necessary.
- The doggo, floofer, pupper and puppo classifications werre unnecessary and should be in one column.

3.Cleaning Data

First, I ,merged the 3 dataframes into one dataframe 'raw-df'. I then saved a copy of the dataframe and named it 'new_df'. It was this dataframe(new_df) that I performed the cleaning operations on. I used the Define, Code , Test method to effectively cleanse the data of the quality and Tidiness issues.

4.Data Storage

After cleaning the data, I saved it all in a csv file named 'twitter_archive_master.csv'.