Data Wrangling Report

This project involved the wrangling of data using python libraries. The data is the we rate dogs data and the processes involved were in three phases which includes;

Data Gathering

Accessing Data

Cleaning Data

Data Gathering:

The data was gathered from three sources

1. The we rate dogs twitter archive that was provided in the udacity classroom, I manually downloaded this and read it using pandas and named it tweet_arch
2. The image prediction data which I had to download programmatically using the provided url in the classroom, after which I read it into my workspace with the alias img_pred
3. The additional data which I got by querying the twitter API using the tweepy library, I rad it into my workspace and called it tweet_df.

Data Assessment:

In this phase of the data wrangling, I went through all three dataframes in search of quality and tidiness issues using both visual assessment and Programmatic assessment. During the visual assessment, I noticed there were a lot of NAN, I couldn't easily understand the source columns too. Outlined below are the issues I came across;

Quality Issues:

- ✓ Some retweets and replies were included in the datasets, these constitutes most percentages of the null values
- ✓ Erroneous datatypes
- ✓ Innacurate values in the ratings_denominator and ratings_numerator columns
- ✓ Some names of dogs in the tweet archive table are not given, instead they are replaced with the words 'None' and 'a'
- ✓ Some names of dogs in the tweet archive table are not given, instead they are replaced with the words 'None' and 'a'
- ✓ In the expanded url column in tweet archive, the urls are repeated on one row making them irregular
- ✓ The source column is unfiltered, making the data not understandable
- ✓ Some values were gotten from vine and not twitter as seen in the source column
- ✓ non descriptive names in the img_pred columns

Tidiness Issues:

- ✓ The stages of dog format should be in one column

- ✓ Irregularities in the casing of the predicted names of dogs in the image prediction table, some and in lower case and others in title case
- ✓ the ratings should be in a column
- ✓ all dataframes should be in one table

Data Cleaning

- ✓ I dropped rows with retweets, Replies and other missing values as I want to work with only tweets
- ✓ I changed the timestamp datatype to Datetime
- ✓ I corrected some innacurately entered values in the ratings_denominator and ratings_numerator columns and then dropped outrageous figures from both columns
- ✓ I Identified dogs that were misnamed and renamed them
- ✓ I dropped Duplicates and NA values in the expanded URL columns
- ✓ I identified values in the expanded urls columns that are repeated on the same rows and adjusted them
- ✓ I extracted relevant source of the tweet from the html tag using the split method and replace them with the values on the source column
- ✓ I dropped values gotten from vine
- ✓ I renamed columns to more descriptive name
- ✓ I melted the doggo, floofer, pupper and puppo co;umns into one column(dog_stages)
- ✓ I converted all predicted names into lower cases and remove the underscore between each words in the names.
- ✓ I created a ratings column to house both the ratings_numerator and the ratings_denominator
- ✓ I merge all three dataframes into one