WRANGLE REPORT

INTRODUCTION

Data Wrangling is a process of Gathering, Assesing and Cleaning Data
In this project we are dealing with 3 different datasets

GATHERING DATA

1.Twitter Archive Data: The WeRateDogs Twitter archive. This has been given as a CSV file Using pandas, the dataset has been successfully imported

2.Image Predictions Data: The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. Had to pull a request to store the data in a .tsv file

3.Twitter API Data: Each tweet's retweet count and favorite count at minimum and Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count

ASSESSING DATA

After assesssing the data the following issues were found

Quality Issues
1.Timestamp has string values. Need to change to correct format
2.Rating numerators have values above and below 10. I we would assume the ratings to be on a scale of 10, then the rest shouldn't be considered valid. Remove rating denominators anything other than 10
3.Name field has articles in it like an,a,the etc, which we humans would know is not a name, but any software would consider it valid
4.Remove tweets that do not have a rating associated with it
5.Make sure rating numerator has valid entries
6.Convert tweetID from tweetjson to int
7.Remove P1_dog and p2_dog = False, keep only true values
8.Remove Expanded URL's
9.Remove retweet related columns
10.Delete rating denominator column

Tidiness Issues
1.columns pupper,puppo,floofer and doggo will be grouped to single column called DogStage
2.img_pred and tweetjson have the same informational attributes as twitter archive so the tables can be merged.

CLEANING DATA

Cleaning data happens in 3 stages

Define, Code and Test. Followed the same for addressing all the issues

1.Timestamp has been changed using pd.to_datetime() function

2.Rating denominators other than value 10 have been removed from the dataset

3.lower case words in the name field have been removed as they do not associate with dog names

4.Tweets with no ratings have been removed(found 2, dropped 2)

5.Rating numerator value above 10 have been removed, as we have considered ratings to be on a scale of 10

6.TweetID of Twitter API dataset has been changed to int to accommodate easy merging

7.P1 and P2 dogs corresponding to true values have been kept and only P1 and P2 are used for our analysis

8.Expanded URLs are removed

9.Retweet related columns are removed

10.Rating denominator column is deleted as it serves no purpose

11.DogStage column has been created

12.All the 3 datasets have been merged using inner join

STORING DATA

After all the cleaning and merging has been done, the final dataset has been written and stored into a csv file called twitter_master_archive