WRANGLE REPORT

INTRODUCTION

This project involved gathering, assessing, cleaning and analyzing data from tweets of “WeRateDogs” twitter account

Python packages used includes numpy, pandas, requests, os, matplotlib

DATA GATHERING

Data was gathered from three sources and read into dataframes for further assessment and cleaning

- The file named ’twitter_archive_enhanced.csv’ contained the archived data of tweets and was read into ‘twitter_archive’ dataframe
- The ‘image_predictions.tsv’ file contained data on image predictions of the dogs from the tweets. It was extracted from the given URL into a folder and then read into ‘image_pred’ dataframe
- The given ‘tweet-json.txt’ file contained additional columns on retweet counts and favourite counts. It was read into a ‘tweet’ dataframe

ASSESSING DATA

Visual and programmatic assessments of the 3 dataframes were carried out to identify quality and tidiness issues. The following issues were identified:

Quality Issues:

- Datatype issues: tweet_id column is stored as integer in 'twitter_archive', 'image_prediction' and 'tweet' dataframes also timestamp column in 'twitter_archive' is stored as string datatype
- Wrong dog names, the wrong dog names are noted to all be in lowercase
- Rows with replies and rows with retweets
- Irrelevant columns: 'in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id' , 'retweeted_status_user_id' and 'retweeted_status_timestamp' contain replies and retweet and are not relevant to analysis

- Some rating_denominators are not 10
- Unclear column names description in image_pred dataframe
- Some dog rating_numerators were wrongly extracted
- From image_pred dataframe, the three image predictions (p1_dog, p2_dog and p3_dog) of some tweets were not dogs

Tidiness Issues:

- Dog stage variable in 'twitter_archive' stored in four columns
- All the data should be on one table

CLEANING DATA

- To clean the dataframes, original copies were first made to keep the data intact
- Identified issues were then cleaned using the 'Define-Code-Test' Sequence

SAVING DATA

The cleaned data was saved as a master dataset. This was then saved to a CSV file named 'twitter_archive_master.csv'

ANALYSIS AND VISUALISATIONS

Analysis and visualisations of the data was then done using pandas and matplotlib libraries

INSIGHTS

Insights gathered:

- The highest value of rating is 14
- Only 336 of the 2356 tweets had dog_stages
- Of the tweets that had dog stages, pupper was the most, consisting about two thirds of the total

- Of the tweets with dog names, Obie's tweet seemed to have the highest number of retweets and favorites count
- For the image predictions, about three-fourths of the first predictions of images turned out to be dogs
- Out of the correct first predictions, 'golden_retriever' is the most predicted dog breed