WRANGLE REPORT

1.0. INTRODUCTION

In this project, the wrangling effort can be highlighted under 3 main subheadings. These are:

1. Gather
2. Assess
3. Clean

2.0. GATHER

In the gathering stage of this project, data was to be gathered in three ways.

1. Directly download the WeRateDogs Twitter archive data
2. Programmatically download the tweet image predictions tsv file
3. Using the Tweepy library to query additional data via the Twitter API

I was able to successfully download the WeRateDogs twitter archive data and the image predictions tsv file using the python requests library. However, my application for a developer account to access the Twitter API is still ongoing. I proceeded with the tweet_json.txt file that was provided.

The json_file was read line by line into a python list and then loaded into a dataframe.

3.0. ASSESS

In the assess phase of this project, many functions and methods were used to understand the three separate datasets in order to find data quality and tidiness issues. Some of the methods and functions used are highlighted below:

1. Info
2. Shape
3. value_counts
4. groupby

The following quality issues were discovered

1. Presence of retweets int the Direct downloads dataset
2. The datetime format of timestamp in the direct downloads data set was not ideal
3. Missing and inaccurate data in the names column of the direct downloads dataset

4. Missing data in the expanded urls column of the direct downloads dataset
5. Missing data in the json dataset
6. Missing data in the doggo, floofer, pupper, and puppo in the direct downloads dataset
7. Presence of retweets int the json downloads dataset
8. Inaccurate and inconsistent ratings in the direct downloads dataset

The following tidiness issues were discovered

1. The stages of development in the direct downloads dataset should all be in one column
2. Many columns in the json dataset contained lists and dictionaries violating the tidiness rule

4.0. CLEAN and STORAGE

During the cleaning stage of this project, various methods and functions were used to clean the datasets in preparation for EDA and visualisation.

Once cleaning had occurred, the new datasets were combined into a new dataframe and stored in a csv file.

Some of the methods and functions employed include:

1. Isnull
2. String extraction
3. Datetime
4. Astype
5. Merge
6. Drop
7. Duplicated
8. Dtype
9. To_csv

5.0. ANALYSING AND VISUALISING

In this stage of the project, analysis was done on the clean data in order to perform EDA. Insights were drawn from the data and visualisations were used to elaborate on the insights. The matplotlib library was used for visualisations.

6.0. CHALLENGES

This project particularly had challenges. These have been highlighted below:

1. Inability to use Twitter API to query data from twitter.

2. Missing Data: Many of the datasets had missing data. This negatively impacted on the range and quality of analysis that could be done.

7.0. INSIGHTS AND CONCLUSIONS

The following insights were drawn from the analysis and visualisations of the dataset

1. The analysis and visualisations above suggested that the rating impacts the number of likes and retweets for each dog.
2. The top 10 breeds by the number of likes are the same top 10 breeds by the number of retweets, however, in a different order.
3. The analysis shows a strong positive correlation between the number of likes and retweets
4. Pembroke and Golden retriever are the top breeds across the variables (likes, retweets, and ratings).

More data and further analysis would be required to crystalize these observations and draw deeper insights.