Wrangle and Analyze Data Report

Name: Gilbert Woasieworm Adjei

This report includes insights gained from wrangling and analyzing data from the WeRataDogs twitter page.

The wrangling process follows the following steps:

Gathering the data

Assessing the data

Cleaning the data

Storing the data

Drawing insights from visualization

Python and the jupyter notebook were used throughout the process of wrangling, cleaning, storing and visualization of the clean data.

The gathering and assessing process:

The required libraries were imported into the jupyter notebook, these were then used to read the data from WeRataDogs twitter archive into the jupyter notebook.

Retweet and favorite data were extracted using twitter API to query for each tweet's JSON data and then written into their own line. This was then read line by line into pandas DataFrame.

The last set of data was gathered using the Requests library and the url given in the notes.

After gathering the data, they were all assessed and merged together for cleaning and storing.

After gathering and assessing the data, there were some quality issues which includes the following,

1. Missing data
2. Time data recognize as object
3. Unwanted data
4. Different names for tweet IDs

Also, there were some tidiness issues such as,

1. Different Data Frames
2. DataFrames has different shapes

A copy of the data was made and all the quality and tidiness issues were addressed.
After cleaning, the cleaned data was then stored.

The data cleaning stage was followed by the visualization stage where the bar charts and regression plots were used to gain insights from the data and relationships between some of the data. This was made possible with the help of matplotlib, pyplot and the seaborn libraries.

The visualizations display the top 10 dog breeds. It shows that the Golden retriever, the Labrador retriever and the Pembroke are the top three dogs breeds to be rated. It also shows the top 10 dog names.
The final visualization that there is a strong positive correlation between the retweet counts and favorite counts from the data.