WeRateDogs Wrangle Report

Following tasks are performed in this project:

- Data wrangling, which consists of:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on
    - 1) Data wrangling efforts and
    - 2) Data analyses and visualizations

Gathering Data

We get data for this project from three different sources:

1. The WeRateDogs Twitter archive is provided by Udacity.
2. 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. This file (image_predictions.tsv) is hosted on Udacity's servers and is downloaded programmatically using the Requests library and the following URL:

   https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Using Python's Tweepy library Using the tweet IDs in the WeRateDogs Twitter archive,we query the Twitter API for each tweet's JSON data. Each tweet's JSON data is 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

I followed process as mentioned below:

1. I assessed the data by exposing them through them through dataframes in pandas, and directly looking at them through editors.
2. I also used several functions in pandas for analysing the data for any discrepancy.

Cleaning Data

1. As suggested in entire course, the wrangling was done in three parts: defining, coding and testing.
2. Create a copy of all three dataframes, and analyse and clean the three copies.
3. Some of the interesting steps were:
   a. Combine dog stages in 1 column, rather than 4 in the archived dataset
   b. Updation of numerals that were decimals

Conclusion

It was an awesome project, and correctly emphasises the importance of data wrangling. Python is the language which was useful throughout the course.