Wrangle Report

Wrangle_act Project:

The wrangle and analyze data project for WeRateDogs twitter data is done so as to create interesting and trustworthy analyses and visualizations.

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention both for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter for breaking these aforementioned laws.

The project was constructed of 5 steps:

1. Gathering Data:
   The data was gathered from 3 sources:
   a. Twitter Archive Data which was available to be downloaded
      - This data contained basic tweet data for 5000+ tweets.
      - This data was extracted programmatically. The ratings weren't all correct. Which required more assessing and cleaning.

   b. Image Prediction File which was downloaded from a given url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
      - this data contained what is the breed of the dogs are as per each tweet id.
      - The data was archived through a neural network that could classify each breed of dogs.
      - The algorithms were divided into 3 levels (high prediction, medium prediction and low prediction).

   c. The real data from Twitter using the twitter API (Tweepy)
      - I first made a twitter developer account.
      - Then used the tweepy API to take out the needed data which included the retweeting counts and favorite counts as well.

2. Assessing Data:
   - Assessing was done visually using Microsoft Excel and programmatically using the Pandas functions.
   - The following data issues were observed:
     i. Missing data.
     ii. Inaccurate data.
     iii. Wrong data types.
     iv. Unneeded data.
     v. Duplicated records.
   - The following tidiness issues were observed:
     i. Data needed to be merged.
     ii. Columns were values not variables.

3. Cleaning Data:
   - All the cleaning was done programmatically.
   - First I made copies from the dataframes to keep the originals unmodified.
   - Each observed issue was defined, solution was coded and then test.

4. Storing Data:
   - The data was stored as requested in a file named twitter_acrhive_master.csv.

5. Analyzing Data:
   - 3 insights were taken from the data using pandas functions.
   - One visualization was made as per the insights taken.