Wrangle Report

Introduction:

In this report I will document the steps and actions I took while wrangling the data of the tweets from twitter account WeRateDogs @dog_rates WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog, the tweets desired are tweets with ratings and pictures mostly.

Gathering:

The data gathered were from three different sources, with the first source being exclusive data given to udacity from WeRateDogs twitter archive, the process of importing the data was done through downloading a csv file manually and importing directly into the workspace by reading it normally using typical python packages.

The second source was an image prediction file, which is a table generated through a neural network that run through the images and classify dog breeds. The process of importing the data was done programmatically using the provided url, it was in tsv form which required some different form of reading.

The third source was downloading data that gives information about retweet and favorite count through twitter api. As I didn’t have access to a twitter developer account, I read the json file using appropriate parameters to match the properties of the file.

Assessing:

While assessing the data for possible quality(8) and tidiness(3) issues, I pinpointed the issues I wanted fixed in a clear list with a description of each issue, here are the following issues I documented to be fixed In the cleaning phase of the data wrangling process.
Issues listed were found programmatically and some were found visually by inspecting the data.

Tidiness Issues:

1- Column names on the image predication are not easily understandable.
2- The three data frames all describe the same thing (A unique tweet). They should be merged together to make analysis easier.
3-What we call a "Dog Stage" is described over 4 different columns(doggo, floofer, pupper, puppo).

Quality Issues:

1-"timestamp" column should be of type datetime.
2-There are some clear outliers in the rating_numerator column.
3-The denominators are not all consistent.
4-The predicted values are not standardized (Some are capital other are small letters).
5-All tweet ids are of integer type and not Object type(We don't need arithmetic operations on ids).
6-Some tweets are retweets.
7-Some expanding urls are missing.
8-retweeted_status_id and retweeted_status_user_id are of type float.

Cleaning:

Before I started cleaning I saved original copies of the data frames in order to preserve the original data.

In the cleaning section, I started solving the issues I found in my data one by one, with the Define, Code, Test approach.

I started with tidiness issues followed by the quality issues. Most of the cleaning and testing was done programmatically, but I also used visual assessment in some issues for validation.

Conclusion:

Going through the wrangling process will help me and future analystes on analyzing the data by gathering the data and assessing it, even though not all the issues are dealt with and some parts of the data are still messy, I've ensured at least that I can analyze and extract insights on the aspects cleaned in the cleaning phase.