Wrangle and Analyze Data Report

I completed the “Wrangle and Analyze Project” as part of the Udacity’s Data Analyst Nanodegree.

About the dataset (based on Udacity’s description):

The dataset I was wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Gathering data:

Data was gathered from three different sources:

1. The WeRateDogs Twitter archive (twitter_archive file) was provided by Udacity and downloaded manually.
2. The tweet image predictions (image_prediction file) were downloaded programmatically using the Requests library from Udacity’s server.
3. Favorite and retweet count was gathered by using the Twitter API

Assessing data:

Following methods were used in order to assess the data:

-.head()
-.sample()
-.info()
-.value_counts()

Tidiness issues that were cleaned:

twitter_archive

- Missing values were in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
- Last 4 columns were merged (doggo, floofer, pupper, puppo) and column was called "stage"

image_prediction

- Was merged with twitter_archive, as data was talking about the same tweets

status_df

- Was merged with twitter_archive, as data was talking about the same tweets

Quality issues that were cleaned:

twitter_archive

- Some column names were not very specific to what they stand for
- tweet_id was an integer
- timestamp and retweeted_status_timestamp were type 'object'
- source was with a and \a tags surrounding the text
- contained retweets and therefore duplicates
- text column contained untruncated text instead of displayable text
- the way the rating was displayed in general was not standardized and therefore difficult to analyze
- dog names were starting with lowercase characters (e.g. a, an, actually, by)

image_prediction

- Some column names were not very specific to what they stand for

status_df

- Some column names were not very specific to what they stand for
- tweet_id was an integer

Clean data:

Following methods were used on order to clean the data:

merge()
reduce()
extract()
slice()
drop()
isnan()
astype()
to_datetime()
islower()
replace()

rename()
set_option()
loc()
value_counts()
info()
head()
Loops