Data Wrangling Report

Introduction

The dataset I used in this project to wrangling, 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. The project will include two main task which is data wrangling, which consists of: gathering data, assessing data and cleaning data. Second task is storing, analyzing, and visualizing wrangled data.

Gathering Data

1. Download the WeRateDogs Twitter archive by using twitter_archive_enhanced.csv

2. Download the tweet image predictions from image_predictions.tsv (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv). The mage_predictions.tsv is hosted on Udacity's servers.
3. Use the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Accessing Data

1. Assess the data visually and programmatically
2. Detect and document at least eight quality issues and two tidiness issues:

Quality Issues

'archive' table:

- Tweet_id is an int
- The dataset included retweets, which means it included duplicated data like retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
- timestamp is an object
- Missing one column for the rating instead of rating_numerator and rating_denominator
- There are missing data in the columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
- Dogs' name has a, the, this or an
- Lowercase given Dogs' name
- The source column’s content are not clear

'Prediction' table

- tweet_id is an int
- The column jpg_url has 66 duplicated data
- The first letter of the column p1, p2, p3 some lowercase, some uppercase, and the symbol connect word are nor consistent

Tidiness Issues

'archive' table

- The variable for dog stage is in four different columns (dogoo, pupper, puppo)

'prediction' table

- This dataset has part of the same observation in the archive table

'tweet_json' table

- This dataset has part of the same observation in the archive table

Cleaning Data

1. Clean each of the issues that documented in assessing part

The steps of cleaning data is define, code, test. In the cleaning section, I cleaned the data by:

1. Convert tweet_id column's data type from an int to a string using astype
2. Drop jpg_url duplicated data
3. Merge archive table, prediction table and tweet_json table, joining on tweet_id
4. Remove Retweets data
5. Convert timestamp to datetime data type
6. Remove the doggo, floofer, pupper and puppo from archive table. These are dog stages
7. Calculate rating by using rating_numerator/rating_denominator
8. Remove useless columns like in_reply_to_status_id, in_reply_to_user_id, rating_numerator , rating_denominator
9. Replace Dogs' name with a, the, this or an to None
10. Get main content of source columns, drop useless information
11. Replace lowercase letter to uppercase letter for name
12. Replace lowercase letter to uppercase letter for p1,p2,p3, replace'-' to '_' to make consistent

Storing, Analyzing, and Visualizing Data

1. Store the clean DataFrame in a CSV file called twitter_archive_master.csv

   The witter_archive_master.csv file contains 1994 rows and 21 columns.

2. The following questions will be solved:
   1. What's the common dog stage?
   2. Which dog stage has highest rating?
   3. What’s the top 10 dog's name?
   4. What’s the relationship between tweet count and favorite count?
   5. Which month has most retweet and like?