Data Wrangling Report

Project Overview

This project aims to extract and wrangle Twitter data from the WeRateDogs account and to produce interesting and insightful analyses related to the tweets from the account.

Data

This project utilizes the following 3 datasets:

- Enhance Twitter Archive - Contains basic tweet data for all the tweets from WeRateDogs
- Image Predictions File - Table containing dog classifications via a neural network
- Additional Data from Twitter API - Data such as retweet counts, favorite counts and URL

Gathering Data

Below contains the following steps taken to extract the 3 datasets explained above:

- Enhance Twitter Archive
    - This dataset was provided to all students
    - Dataset was uploaded directly to the environment
    - Loaded in as a comma separated table
- Image Predictions File
    - Dataset hosted on Udacity's servers
    - Extracted via the Requests library using the given URL
    - Loaded in as a tab separated table
- Additional Data from Twitter API
    - Dataset extracted via Twitter API
    - Data values were then extracted from their JSON format
    - Final data set loaded in as a pandas dataframe
    - **Note: I was unable to get the provided Twitter API code running and have manually uploaded the API data as a comma separated table**

Assessing Data

All 3 datasets were then assessed in terms of quality and tidiness both visually and programmatically.

- Enhance Twitter Archive
    - Quality Issues
        - Missing values for some columns (in_reply_to_status_id ,in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)
        - Invalid dog names like 'the', 'a', 'an' etc.
        - Data contains retweets
        - Timestamp column is not a date-time datatype
        - Retweeted_status_timestamp is not a date-time datatype
        - Contains retweet related data
    - Tidiness Issues
        - Dataset is apart of the other 2 datasets and should be combined with the others

        - Last four columns in the dataset are all related to the dog type and should be contained in a single column
- Image Predictions File
    - Quality Issues
        - P1, p2, p3 columns contain invalid dog types such as ‘web_site’, ‘nail’, ‘fire_engine’ etc.
        - P1, p2, p3 columns are inconsistent with their capitalization, some are capitalized and some are all lower case
        - P1, p2, p3 columns are inconsistent with their formatting, some have underscores and some don’t
    - Tidiness Issues
        - Dataset is apart of the other 2 datasets and should be combined with the others
- Additional Data from Twitter API
    - Quality Issues
        -
    - Tidiness Issues
        - Dataset is apart of the other 2 datasets and should be combined with the others

Cleaning Data

After assessing the data for quality and tidiness issues, the following steps were taken to clean the data:

1. Combining all 3 datasets into one dataset
2. Create 1 single column to store the 4 different dog types
3. Convert retweeted_status_timestamp to date-time datatype
4. Remove retweeted related columns and data
5. Remove columns with missing values
6. Convert timestamp to date-time datatype
7. Replace invalid dog names with ‘None’
8. Replace invalid dog types in p1, p2, p3 with ‘None’
9. Convert all names in p1, p2, p3 into lower case to keep consistency
10. Replace all underscores with space in p1, p2, p3 to keep consistency