Project: Wrangle and Analyze Data

Wrangle Report

Introduction

This report is a summary of the "Wrangle and Analyze Data" project completed for the Udacity Data Wrangling course, which is part of their Data Analysis nanodegree program. The project directions are to gather, manipulate, and analyze data from three separate sources:

- The WeRateDogs Twitter archive
- Tweet image predictions from the Udacity servers
- Twitter API for JSON data

Gathering Process

WeRateDogs Twitter archive: This data was available in CSV format through the course page and was a simple download. Once downloaded, the data was read into a pandas dataframe.

Tweet image predictions: This data was downloaded programmatically from a specified URL as a CSV file, then read into a pandas dataframe.

Twitter API: This step was the most difficult. First, it was necessary to obtain access keys from the Twitter developer portal. Armed with these keys, I was then able to query their database using Python's Tweepy library and store each tweet's entire set of JSON data in a text file. The desired data from the JSON file were then combined into a list object that was eventually converted into a pandas dataframe.

Assessment Process

Each pandas dataframe was then inspected both visually and programmatically. The visual inspection led to details which were then inspected further programmatically.

WeRateDogs Twitter archive: The assessments illuminated the following issues:

Quality Issues

- The dog "type" columns were contained in 4 different columns (doggo, floofer, pupper, and puppo), which have either the column name or "None" as values. I felt that it would be more helpful if these were Boolean datatypes
- The value in the denominator column was mostly 10, but there were a few off occurrences where the value was different. I felt that it was best to drop the rows where it was not 10 to maintain consistency.
- The datatype for the timestamp was a string type, and I feel that datetime format would be better.
- Some values in the name column were "a", "an", and "None" which are not descriptive or helpful

Tidiness Issues

- We were told not to include retweets in our analysis, and there were 181 occurrences.
- We were told not to include replies in our analysis, and there were 78 occurrences.

Tweet image predictions:

Quality Issues

- Some of the column names were not meaningful
- The breed names are mostly capitalized, but some are lowercase
- There are columns that I do not plan to use for analysis

Twitter API:

Quality Issues

- The datatype for the timestamp was a string type, and I feel that datetime format would be better.

Cleaning Process

The following methods were used to clean the data to address the issues listed above

WeRateDogs Twitter archive:

- Used .replace function to change the value of the four dog type columns to boolean true/false values
- Used .drop method to remove rows where the denominator does not equal 10 for consistency
- Converted string datatype for timestamp column to datetime using to_datetime
- Used . replace method to replace "a", "an", and "None" with null data
- Since null values are used when messages are retweeted, I droppp all rows where the retweeted status ID is not a null using .drop method
- I used the same method as above to remove all rows where the in_reply_to_status_id is null

Tweet image predictions:

- Created an array mapping the old names to the new names, then use the.rename function to rename the columns using the array
- Used str.capitalize method to capitalize the name in the prediction column. (Note: Will only capitalize prediction 1 since the others are being dropped in the next step)

Project: Wrangle and Analyze Data

- Deleted unused columns using .drop method

Twitter API:

- Converted string datatype for timestamp column to datetime using to_datetime