Data wrangle project

By: Abeer Alharbi

Introduction

This project focused on wrangling data from the WeRateDogs Twitter account using Python, documented in a Jupyter Notebook (wrangle_act.ipynb). This Twitter account rates dogs with humorous commentary. The rating denominator is usually 10, however, the numerators are usually greater than 10. This aspect was not cleaned as it is part of the humor and popularity of WeRateDogs.

For this project I used three data sets that we combined together later in one dataframe.

1- twitter archive data
2- image predation data
3- json data

The goal of the project was to wrangle the dataset and create interesting and trustworthy analyses and visualizations. In order to achieve that we had to request the Twitter API and collect additional data for the provided WeRateDogs tweets. Furthermore, another dataset which contained image predictions of a trained neural network of the aforementioned tweets needed to be downloaded programmatically from the Udacity website and cleaned as required.

The tasks for this project were:

- Data wrangling, which consisted of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing the wrangled data
- Reporting

Gathering

The first data set was from WeRateDogs archive provided. Second, the image predictions data which needed to be downloaded programmatically from the Udacity website. Finally, the tweets data that has been taken from text file provided.

Assessing

In this part we assess the three dataframes in order to identify data quality issues and tidiness issues that needed to be solved. Eight quality issues were found and two tidiness.

Quality summary from twitter_archive table:

- Retweets present in the dataset (we only want original tweets with images)
- Columns relevant to retweets and replies needs to be dropped
- Erroneous datatype for tweet_id
- Timestamp have to changed to a uniform date type
- 23 rating denominator not on 10
- 440 numerator below to 10
- doggo, floofer, pupper and puppo have None string we have to convert it to NaN
- Incorrect dog names
- Missing values in 'name'

Tidiness summary:

- Timestamp should be split into date and time
- doggo, floofer, pupper, puppo columns in twitter_archive_enhanced.csv should be combined into a single column as this is one variable that identify stage of dog.
- Information about one type of observational unit (tweets) is spread across three different files/dataframes. So these three dataframes should be merged as they are part of the same observational unit.

Quality summary from twitter_images table:

- Erroneous datatype for tweet_id.
- Lowercase dog breed names
- Non-descriptive column names for several variables.

Quality summary from df_tweet_json table:

- tweet_id has to be a string
- Missing columns: favorite_count and retweet_count

Cleaning

We started this step by making a copy of all three dataframes. So, at this part we cleaning the issues found in the second part. By using several pandas methods ,loops and functions. After that we creat a master DataFrame that contained all the cleaned variables from our the three dataframes and we saved it as csv file.