Wrange Report

Project: Wrangle and Analyze "WeRateDogs" Data

The purpose of this report is to summarize the tasks performed during the data wrangling process of this project. I utilized data from the Twitter account **WeRateDogs**, which rates people's dogs with humorous comments about them.

As we learned in the course, the data wrangling process involves three steps: gathering, assessing and cleaning the data. These steps can as well be iterative. In the following I describe how each stage of the data wrangling process was carried out for this project.

Data gathering

The data to be gathered consisted of three parts, as described below:

1) The **WeRateDogs** Twitter archive, in a **.csv** file provided by Udacity (**twitter_archive_enhanced.csv**). This file was included in the working directory and loaded into a Pandas data frame **df_twitter_archive**.

2) Tweet image predictions, i.e., the results of classification as to what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (**image_predictions.tsv**) is hosted on Udacity's servers and was downloaded programmatically using the requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. Then, the data was loaded into a Pandas data frame called **df_image_predictions.**

3) Each tweet's retweet count and favorite ("like") count. This is done by querying the Twitter API for each tweet's JSON data using Python's Tweepy library and storing this data in a file called **tweet-json.txt** file. In this case, because of mobile verification issues, the code was provided by Udactiy, together with the **tweet-json.txt** file containing the queried data. This was loaded into a third Pandas dataframe **df_twitter_api.**

Data assessing

The three data frames **df_twitter_archive**, **df_image_predictions** and **df_twitter_api** were assessed both visually and programmatically in order to identify at least two tidiness issues and eight quality issues, which were addressed in the data cleaning section.

For the assessment of each data frame I made use of Pandas built-in functions and dataframe attributes, such as s**ample(), info(), nunique(), value_counts(), dtypes, duplicated(), head(), tail(),** etc.

The following issues could be detected from the assessment:

- **Tidiness issues**

  df_twitter_archive

  - Unnecessary columns related to retweets and replies should be removed: **in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp**

  - Dog stage columns should be combined into a single column: (**doggo, floofer, pupper, puppo**)

  - The three data frames should be merged on **tweet_id**

- **Quality issues:**

  df_twitter_archive

  - Entries corresponding to retweets or replies should be removed
  - Data type of the **timestamp** column should be datetime, not string
  - Some dog names are not correct (eg. "a", "quite", "such", "very"), all of them appear in lowercase
  - Some dogs have more than one stage
  - Some dogs stages are NaN
  - A few rating denominator values are different from 10
  - A few rating numerator values are out of the range 1-10
  - The **expanded_urls** column has missing values

  df_image_predictions

  - Some prediction names appear in uppercase, while others in lowercase
  - In some prediction names there's an underscore instead of a whitespace to separate compound names
  - Some tweets don't have associated images

Data cleaning

The data cleaning process began by making a copy of each data frame, where the necessary modifications to the original data were made. Each issue mentioned in the data assessment section was addressed, following the steps learned in the course:

1) Define: action to take to solve the issue
2) Code: piece of program to implement the solution
3) Test: verify the piece of code works as expected

At the end of the process, a Pandas data frame **df_twitter_master** was obtained containing the clean data, which consists of 22 columns and 1948 entries. This data frame was stored in a **.csv** file named **twitter_archive_master.csv**, for further analysis.