Wrangling Report

By: Mohanad Ismail

Data Gathering

In this project, it was required to collect data from three different sources in three different formats.

First file was provided readily as a .csv file. It contained basic data about tweets of the twitter account "WeRateDogs". The file was read using the pandas read_csv function.

The second file was required to be downloaded using a link that was provided in the project description. To download the file, I used the requests library and saved the file in a .tsv file as required by the project. This file includes predictions made by a neural network to what breed of dogs are in the photos.

The third and last file is to be extracted from the Twitter API which contains extra data about the tweets. I used tweepy library to extract this file and saved it in a .txt file as required by the project.

Then I read each of these files into a separate data frame.

Data Assessing

The data was assessed in two ways: visually and programmatically. The problems found were classified into two types: Quality issues and Tidiness issues

Tidiness Issues:

1. Dog stages are distributed on four columns instead of only one
2. The 'text' column has the tweet text in addition to a shortened URL to the tweet.

Quality Issues:

1. Missing data in 'expanded_urls' column
2. Some rows are retweets and replies, which should be removed
3. Timestamps were of the wrong data type
4. The 'name' column has a lot wrong names (a, an, such, quite)
5. Some tweets have wrong ratings
6. Image predictions file has less tweets
7. Nulls are represented as 'None' throughout the dataframe
8. Column names need to be cleaned to be more clear
9. Columns that determine retweets and replies are of no use after deleting those tweets

Data Cleaning

Here I will write the approach on each issue. The code can be found in the jupyter notebook.

Tidiness

1. The melt method was used to combine the four columns into a single one. Then it was found that some rows have more than one dog so they were combined into one column
2. Split function was used to split the two columns

Quality

1. Rows were removed because that means they have no photos
2. Retweets and replies were also removed
3. Used the to_datetime function to change their datatype
4. All words in lowercase were replaced with null
5. Ratings were changed manually
6. Tweets with no predictions were removed because that means they have no photos
7. Used the replace method to change ‘None’ to NaN
8. Edited columns names and removed duplicated columns
9. Dropped those columns

Storing

The three datasets were combined into one master dataframe after cleaning and saved in a master .csv file