Wrangle Report

by: Hassan Almaghrabi

Introduction

In this project, I will go through the wrangling process of the data. The dataset that is used in this project is the Twitter tweets archive of the user WeRateDogs @dog_rates, which is a Twitter account rating people’s dogs with a humorous comment about the dog. This wrangling report will contain the three main steps of data wrangling those are: gathering data, assessing data, and cleaning data.

Gathering the data

In this project, I should gather data from three different sources using several techniques

the data I have gathered for this project are as follow:

1. the WeRateDogs Twitter archive, which is come as CSV file and it is provided to download directly from Udacity.
2. The tweet image predictions, It is a plain text file of type TSV, and I download it programmatically using the requests library in Python.
3. Twitter API JSON file, it is a file that is containing the retweet count and favorite count of the tweets in the WeRateDogs Twitter archive, I manage to get that information by utilizing Twitter app API.

Assessing the data

After I have done gathering the data, I come to assess the data either programmatically or visually to look for the quality and tidiness issues in the data and I documented the issues that I have notice so I can deal with the issues in the cleaning step

The issues in the data:

Quality Issues

1. in twitter_archive the HTML tag should be removed from the source column.
2. in twitter_archive some columns’ data types need to be changed.
3. in twitter_archive drop any reply tweets, we can find those tweets where the in_reply_to_status_id is not null, then drop the columns related to in_reply_to since there are a lot of missing values in those columns.
4. in twitter_archive the retweeted tweets should be removed, along with the columns related to them.
5. in twitter_archive the rating_denominator column has other values than 10, for seek of consistency we need to drop those entries.
6. in twitter_archive the "expanded_urls" column has some rows with missing values.
7. in image_predictions there are some duplicated jpg_url, this issue occurs due to the retweeted tweets so we need to drop them.
8. in twitter_api the retweet and favorite count data type shall be converted to integer.

Tideness Issues

1. in twitter_archive the last 4 columns shall be combined to be one column called "dog_stage".
2. in image_predictions we need to get the most likely dog breed into a new column.
3. merege the three dataframes into one dataframe called "df_master".

Cleaning the data

In the last step of data wrangling which is the cleaning step I have to solve the issues that I discover it in the data and come up with a trustful data to do the analyzing and visualizing later on , So I made a copy of each dataframe that I gathered earlyer and I have use the Define, Code, Test methodology to approach the issues in the data. After dealing with all of the issues, I store the clean data into a csv file to do further cleaning or analysis if needed.