Data Wrangling Report

By Ghada Maher Mohamed

Septemper 2020

The purpose of this project is wrangleing WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.This Report shows main steps involved in data wrangling of Twitter account “ WeRateDogs” .

Data Gathering

There were three main sources for the data to deal with :

1-the WeRateDogs Twitter archive dataset, I dowloaded it manually from the given instructions of the project as twitter_archive_enhanced.csv and then imported into working environment using Pandas function “pd.read_csv”.

2-The tweet image predictions dataset, I dowloaded it programmatically from the given instructions of the project as a url using Requests library get function and pd.read_csv pandas function .

3-The tweets dataframe that is consist of each tweet's retweets and likes.the dataset was gathered from twitter REST API by python's Tweepy library to extract information and store each tweet using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data in a file called tweet_json.txt.

Data Assessment

I assessed the data using two methods, visually and programmatically for quality and tidiness issues. I started to detect and document each issue in the wrangle_act.ipynb jupyter notebook. I identified 16 quality issues and two tidiness issues .The issues are:

Quality issues

From the Twitter archive dataset

- **data types(consistency issues): all timestamps are object type**
- **all tweet_ids are integers**
- **missing entries in expanded_urls**
- **Reformat source column to display clear text.**
- **there are retweets in the dataset**
- **Unecessary Colums: remove columns(retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp) because it is unuseful.**
- **erroneous names like the letter a and an**
- **inconsistent repersentation of null values as None strings in the (name ,doggo, floofer,pupper,puppo) columns.**
- **incorrect values of the rating_numerator and rating_numerator that extracted wrong.**
  - **change values that extracted wrong**
  - **Set tweet (516) ratings to Null**
  - **drop tweet (979) which have unvalid rating**

From the Image predictions dataset

- **Erroneous datatype for tweet_id.**
- **There are 66 duplicated url**
- **Columns need to reshaping**

From the Api Tweet dataset

- **Erroneous datatype for tweet_id.**

Tidiness issues:

- **Melt the 4 colums (doggo,floofer,pupper,puppo) to one colum for each row with the name Dog_stage.**
- **Merging all the datasets in one dataframe.**

Data Cleaning

I started the cleaning process by making a copy of each dataset. Then, I went for each issue to clean it in 3 steps, which are define, code and test.Several pandas methods were used (i.e. info(), .drop(), .astype(), .loc(), etc.) and many loops and functions .

After cleaning all of the issues listed before, I combined all the datasets in one named twitter_archive_master.csv .