WeRateDogs Wrangling Report

By: Asmaa Ahmed Kamal

April 2021

As a part of the project "Data wrangle and Analyse" in the Udacity Data Analyst - Professional Nanodegree; this report illustrates the main steps done in the data wrangling of the Twitter account "WeRateDogs"

- Data Wrangling includes three main steps:
  1- Data Gathering
  2- Data Assessment
  3- Data Cleaning

Data Gathering:

In this step, data of twitter archive is collected from 3 main resources:
1- Twitter_archive_enhanced.csv file, this file was downloaded directly from the workspace of the course, and was imported to the working environment of the project using pandas function "pd.read_csv".

2- Image_prediction.tsv, is the second source, and it has been downloaded from its URL using requests library get function, then it is read using pandas function 'pd.read_csv'. The file contains image predictions for the dog breeds in the archive, obtained through a neural network.

3- The final dataset was gathered in a jason-file.txt, the file has been read line by line to extract data like retweets count and favourite count. I couldn't gather it through twitter API, as the approval process was taking longer than it should to proceed the project.

Data Assessment:

In this step, I assessed data after its gathering visually and programmatically to look for any quality and tidiness issues.

- Visual and programmatic assessments were done on Jupyter notebook, using functions as info, sample and value_counts, duplicates.
- Then assessment was separated into quality issues and tidiness issues.

- **Quality Issues:**

  archive_df:

  - Timestamp column data type is an object not datetime
  - Name column contains null values under "None" not Nan, or have weird names as a, an
  - Data types of columns: (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id) is float, could be string since we aren't doing any actions on them. (could be dropped too)
  - Some rows don't have pet's classifications
  - Some tweet ids have collective ratings for packs of dogs (not rated individually)
  - Inaccurate numerator ratings because data type of ratings columns are int, not float
  - Source column can hardly be read

image_pred_df:

- Number of data entries in this table is only 2075 while in archive_df is 2356 (#tweet_ids without image)
- Non descriptive column names (p1_conf, p1_dog, p2_conf, p2_dog, p3_conf, p3_dog)
- Some tweets are retweets and replies
- Tweet ids with the same jpg_url (Duplicate data)
- Columns of predictions and configurations (first predictions are statically are always higher), so we might exclude the other two predictions.

- **Tidiness Issues:**

archive_df

- Column headers (doggo, floofer, pupper, puppo) are values, not variable names.

Merging Dataframes

- Merge image_pred_df and api_df into archive_df

Cleaning Data:

This process was divided into three main steps for each issue; Define, Code and Test.

First, I created a copy of each data frame so if there is an error made during coding, the original data frames remain the same.

Then I merged the 3 data frames into 1 data frame on tweet_id, and by that; any duplicate in the jpg_url were cleaned automatically and any tweet_id without jpd_url are dropped. After that I dropped tweet_id with retweets and ids and accordingly dropped their unnecessary columns.

Then trying to restructure the columns of (doggo, floofer,..), cleaning rating of numerator and denominators, dropping unnecessary columns of the other predications and cleaning every other remaining quality issues.

Then storing the final version of the cleaned data frame in a csv file (twitter_archive_master.csv)

On a side note, I believe that there is no final version for a cleaned dataset, as the knowledge of the analyst increases, his skills in cleaning gets better in finding more and more data quality issues. And so, this dataset might be satisfyingly clean at this point of my learning but I believe it can be cleaner.