Data-Wrangling Report

By Huda Ahmed Abd El-Majeed

February 2021

As an assignment for Udacity Data Analyst Nanodegree; This Report illustrate the main steps involved in the data-wrangling of Twitter account "**WeRateDogs**".

Data Gathering

In this step, collection data takes place. For this project, there were three main sources for the data to deal with:

** Twitter_archive_enhanced.csv file, this file was delivered by email and downloaded manually to our working directory and then imported into our working environment using Pandas function "pd.read_csv".

** Image_predections.tsv file, this file has been hosted on a webpage and downloaded from its relevant URL using the Requests library get function and "pd.read_csv" Pandas' function. . This file encompassed image predictions for the dogs' breeds obtained through a neural network on most of the tweets in the archive file.

** Final dataset was gathered from twitter API via the tweepy library by querying the API to obtain extra information pertinent to the tweets' ids in the first file, e.g. retweets count and favorite count aspects.

Data Assessing

After gathering the data , I started to assess the data **Visually** and **programmatically**.

Visually : It's done on spreadsheet application "Excel"

Programmatically : It's done in Jupyter Notebook by using Pandas' functions and methods that are used to assess data .

Assessing process also done on **Quality** and **Tidiness** issues

Archive Table :

Quality Aspects

- 'name' column has alot of unmeaningful and invalid values (a, an, etc...).
- 'timestamp' and 'retweeted_status_timestamp' columns are str instead of datetime.
- 'source' column's values are formatted as html tags "< a > href=url < / a >".
- Retweets and replies on the original tweets are presented in the data.
- columns that include ID variables are sometimes integers or floats instead of strings.
- 'rating_denominator' column has values more or less than 10.
- 'rating_numerator' column should be float type so it can accounting for decimals correctly.

Tidiness Aspects

- Many columns for the same measurement unit ('doggo', 'floofer', 'pupper', 'puppo'), i.e. Dog stage

Image predictions Table:

Quality Aspects

- Some columns name are unmeaningful and non-descriptive.
- Missing photos for some IDs as the total number of records (2075 instead of 2356)
- Inconsistent capitalization for some columns values (p1, p2, p3)

Tidiness Aspects

- Many columns for the same measurement unit ('p1', 'p2', 'p3') and they all present breed predictions

API Table:

Quality Aspects

- Missing retweet or favorite counts for some IDs as the total number of records (2331 instead of 2356)

Tidiness Aspects

- Not observitional unit to have its own table

Cleaning Data

I cleaned the data after the assessment through 3 steps

** **Define** the issue **Write a **Code** to solve it ** **Test** my code to make sure it's solved

_Convert timestamp columns to datetime

_Extract tweet source from source column

_Remove retweets and replies

_Remove values that more or less than 10 in 'rating_denominator' column

_Convert 'rating_numerator' column to float type so it can accounting for decimals correctly

_Create new column 'dog_stage' instead of the four columns 'doggo', 'floofer', 'pupper' and 'puppo

_Replace 'None' with np.nan in 'name' column and remove any rows with invalid names.

_Remove rows with NaNs for 'expanded_urls' column i.e. missing photos

_Capitalize the first letter of names of 'p1', 'p2', 'p3' columns

_Create breed and confidence columns with highest confidence predictions and drop other columns

_Rename 'id_str' column to 'tweet_id' so we can merge the data into database

_Convert 'tweet_id' columns to str for the three data sets

Finally : Merge the three datasets