Data Wrangling Report

This report will include an explanation of the wrangling process for WeGetRate data.

Let’s introduce the datasets we worked on:

1- WeGetRate is a Twitter account that provides rates of dogs with a funny comment, we got an archive of this account ‘tweets.

2- Also from twitter, it considers an additional dataset to archive dataset.

3- The result of classification model dataset from Udacity. This model predicts whether the image in tweets is it dog and from which type (breeds).

We gathered each dataset in various ways upon on the file type. For example:

Archive's tweets is CSV files so just need to load and read. However, Image prediction and tweet's additional JSON file need to request from the Udacity server then load and read it.

After gathering data is assessing the datasets programmatically and visually. Using different Panda and NumPy functions such as:

- Info, sample to view the data
- Isnull , duplicate, and describe to identify the null, duplicates, unique value

We determined many quality and tidiness issues as follow:

- Note: In this project we focused on some of them that serve as in the analysis -

Quality issues:

- *Missing data:*

• Many tweet_id of archive table are missing

• There are a lot of missing values or unnecessary columns such as in_reply_to_status_id

- *Incorrect datatype:*

• Some of the column’s datatypes are incorrect format such as (timestamp)

- *Inconsistent data:*

• The source column in the archive table is messy

• The P1, P2, P2 in the image predication table doesn’t start with capital

• The name column in archive table starts with lowercase

Invalid data:

• There is duplicated tweets in the archive table, which is the retweeted

• There is a rating system (they're good dogs Brent) so the rating should be out of 10 but we have in the archive table more in denominator

• For further analysis and to make insight from the two rating columns in the archive table we create a separate column to calculate the rating

• There are names value in the archive table that doesn’t give a meaning.

• There are many tweets with predicted False value in all Ps tries in the image prediction table, assuming it doesn't contain dogs such as origin prediction or others

• There are many columns related to the predicted try of the model in the image prediction table, so we need to choose the best / most accurate try of model that identify the dog breeds

• There is a missing dog stage in the archive table

Tidiness issues:

• There are three stages of dogs in the archive tables represented in separate columns we should combine them into one

• There is a spread dataset with the same shared IDs so we should join them in one dataframe

After finding the issues in the all tables, we worked on cleaning the data. To help streamline the process, I prefer to clean each table individually.

There are many functions and methods in python used at this stage such as :

- For loop & Regex to extract Source

- Apply & lambda to deal with invalid names.
- Capitalize to deal with inconsistent data.
- Query & condition filters & drop to deal with null, missing, and unnecessary columns as well as the duplicates in tweets
- Append & function & if condition to deal with selecting the best P tries and blending three age/stages of dogs in one column
- Merge to deal with merging all three tables on the tweet_id and left.

Ultimately, we merged all tables into one master then store them in a new CSV file.