Project 2 : Wrangle Report
on the Wrangling and analyzing data project

Udacity Data Analyst Nanodegree

Mame Diarra NIANE

Introduction

This is a part of the project Wrangle and analyzes data .In this project ,we use the twitter archive dataset from the twitter account WeRatesDogs that rates people's dog with a humorous comment about the dog.
These are the steps that I have used for the wrangling part of the project.

1. GATHERING DATA

We need to gather data from differents sources of data:

- The first one is a file on hand that contains the WeRatesDogs Twitter Archive which has basic tweet data for all 5000 + of their tweets. The archive Data contains a column text that is the actual tweets,the rating(numerator and denominator),the dog name , and the dog stage (doggo, fluffer,pupper,puppo). there is 2356 entries in the archive data which corresponds to the tweets with ratings

- In second hand we have the Tweet Image Prediction File that is a flat file that needed to be downloaded programmatically from the Udacity's servers using the request library.This file is a table full of image prediction (the Top 3 prediction only) alongside each tweet with the image url ,the image number that correspond to the most confident image prediction ,the breed of dog that correspond to each prediction and the level of confidence of this prediction

- Lastly we needed to gather data via the Twitter API to get the retweet count and favorite count of each tweet missing in the tweet archive file using Python's Tweepy library .

2. ASSESSING DATA

In this part of the project ,we already have our 3 files that contain the needed information for our assessing and cleaning.
First I did a visual assessment in the Jupyter notebook and also in Excel and Sublime Text for the JSon File, I also did a programmatic assessment and both of these assessments show some quality and tidiness issues.

The Quality issues was :

Tweet archive Table

- remove the tweets that are retweets ,we only need the tweets with actual rating
- in_reply_to_status_id ,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id are float instead of int
- rename source info for more visibility`
- tweet_id should be an string(object) instead of int
- TimeStamp should be Datetime type and remove the timezone in the timestamp column

- there is others data types errors in theses columns (retweeted_status_timestamp for example)
- Rating_numerator should be a float instead of int
- There are numerators (960,182 ,1776 etc...) that are not common compared to what the WeRatedogs account usually puts as a rating. We can check the actual posts from the urls and delete them if necessary
- Same for the denominator that are commonly 10 but might change for multiple reasons that we'll find out and remove these tweets if necessary
- Replace the 'None' by NaN Values

tweet image prediction table

- tweet_id should be an string(object) instead of int
- Some prediction are not dog race (ex : hamster ,llama,guinea pig) these should be removed
- display the most confident_level of prediction and the breed that it predict

tweet_json table

- there are 2325 entries out of the 2356 (twit_archive table) possible. This shows that there's missing tweets compared to the twit_archive table where we take our tweet_id .
- tweet_id should be an string(object) instead of int in the tweet_json table

The Tidiness issues was :

- doggo,floofer,pupper and puppo in the twit_archive table are stages of dog and should be in one column
- the columns in twit_image_prediction table and tweet_json table should be added in the twit_archive table

3. CLEANING DATA

After the assessment part , I clean all the issues documented in that part .Even if there's for sure more quality issues but unfortunately that will be very long to clean .I'll continue on my own to make a better clean version of this data.
I make a copy of the 3 original data.For a very simple pattern,I use the Define-Code-Test framework . This cleaning results to a final dataframe twitter_archive_master that is cleaned and tidy for visualization.