DATA WRANGLING PROJECT

Introduction

In this project, I will be wrangling the WeRateDogs Twitter archive containing that data is a Twitter account that rates people's dogs. At first, I should gather the data and I have three data tables with three different ways to gather it. second, I will also assess Data also I should assess data in different ways. The last thing, I will be cleaning data and Storing data to analyze and visualize data.

Gathering data

In this stage, I gathered three data tables in three different ways. The first file was (twitter_archive_enhanced.csv) and I downloaded the file manually, then I read it by pandas library in the variable named (Twitter_archive). the second file (image_predictions.tsv). These are images predicted from tweets according to using neural network and downloaded it programmatically used the Requests library from (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv) then, I open the file and I write all content from URL, in the last, I read the file by bandas in image_prediction variable. The third file (tweet-json.txt) I downloaded from Udacity and then I open the file and read the file by Jason's library and add all lines in a list then I create the DataFrame and add the list in it.

Assessing Data

After gathering all three pieces of data, I assessed them visually and programmatically for quality and tidiness issues and I detected and documented quality issues and tidiness issues.

In the **Twitter_archive** I documented five quality issues and one tidiness issues that was :

quality issues :

1. Some of the gathered tweets are replies and retweets
2. columns that won't be used for analysis
3. The timestamp has an incorrect datatype, should be DateTime.
4. the "source" is display as HTML.
5. Some values in rating_numerator and rating_denominator seem to be in error or outliers

Tidiness issues:

1. the columns doggo, floofer, pupper or puppo should be one column.

For **Image prediction table** I documented four quality issues and one tidiness issues that was :

Quality issues

1. Missing images there is only 2075 from 2356
2. unclear columns name
3. Dog breeds contain underscores
4. not all images predict dog

Tidiness issues:

1. Image predictions table should merge with Twitter_archive

For **Twitter API** the last table I documented two quality issues and one tidiness issues that was:

Quality issues:

1. Missing tweets
2. Erroneous datatype (tweet_id)

Tidiness issues:

1. Twitter API table should merge with Twitter_archive

Cleaning Data

In this section,I Cleaned all of the issues that were documented while assessing.

The first thing I made copy for all tablee and I cleaned the copy tables to save the original tables. Second,I cleaned the issues in **Twitter_archive** table:

- For some of the gathered tweets are replies and retweets, I removed all of them.
- There were some unnecessary columns ['in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id', 'retweeted_status_user_id','retweeted_status_timestamp', 'expanded_urls'], I droped all of them.
- The timestamp has an incorrect datatype, should be DateTime , I changed the type of timestamp from String to DateTime
- the "source" is display as HTML, I extracted all HTML values from source.
- Some values in rating_numerator and rating_denominator seem to be in error or outliers, for this I regenerated the values in culmns rating_numerator and rating_denominator
- add(doggo, floofer, pupper or puppo)in one column named (dog_stage)

thired, I cleaned the issues in **Image prediction table** table:

- I renamed all columns to clear names
- remove undog images
- I remove underscores from dog breeds then got the highest prediction confidence and its type of breed in all prediction confidence and add each one in a column

Fourth, I cleaned the issues in **Twitter API** table:

- change datatype (tweet_id) to int

The last thing I merged all datasets to gather and add to new dataset named master_archive

Storing Cleaned Data

I saved the master_archive table to twitter_archive_master.csv Then I started my analysis.