Wrangle_report

March 29, 2022

1 Conclusion Wrangle and Analyze Project

In order to draw conclusion from the data sets, I first had to gather the data, then assess it and then clean it.

1.1 Gathering the Data

For this project, three different sources of data were used. - archive .csv file - Image prediction data from the web - twitter_data from twitter

1.1.1 Archive File

The archive file was given in the scope of the Udacity course

1.1.2 Image Prediction Data

The requests library was used to get the data about the image predicions from the web. The file is hosted on a server and has .tsv file format.

1.1.3 Tweet Json File

This file was gathered from the twitter API. To access that api a twitter account with developer access was created. From there the different keys and tokens were saved to access the data later. The get the data, the tweepy library was used and the data saved in a json format called tweet_json_1.txt.

1.2 Assessign the Data

Before I cleaned the data, I first looked thorugh the data to find out what does not comply with the standarts. To do this, I used visual assessing in google docs and programmatic assessing with Python. These are all the issues I hace found and cleand in the data sets: ### Quality issues

1. timestamp is not of data type datetime `(archive)`
2. Some dogs' names are weired. Like 'one' or 'such'. They always begin wich a lower case letter`(archive)`
3. Standarize nominator values of rating column `(archive)`
4. rating_denominator column is not necessary. `(archive)`
5. Only original ratings. No retweets nor replies `(archive)`
6. Not needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls, `(archive)`

7. Dog names does not need '_' and will be replace with a space. Also capitalize the first letter of p1, p2, p3 `(image_predicitons)`
8. We can see that not all ratings are on dogs. p1_dog is whether or not the #1 prediction is a breed of dog is True. OI will remoce the elements when they are not predicted to be a dog. `(image_predictions)`)

1.2.1 Tidiness issues

1. The rating uses two columns but can be done in one. `(archive)`
2. There are 4 columns for the category of the dog when it's only one variable (1 variable. Category -> doggo floofer pupper puppo) `(archive)`
3. Joining the dataframes

1.2.2 Archive File

To assess the data in the archive file, methods and functions like this .info() and .isna().sum() were used. Addicionally, I used google docs to look through the data to find issues. Qulaitiy issues found contain wrong data types, useless columns and that some rows contain data about retweets or replies which are not of concern for our analysis.

1.2.3 Image Prediction Data

When looking through this file, it became evident that not all data was about dogs. Also, the dogs' names were not standarized containing character like '_'.

1.2.4 Tweet Json File

This file was not heavily used. But from programmatic assessing, it became evident, that none of the needed values were to change.

1.3 Cleaing the Data

After gathering and assesing I was ready to start cleaning the data.

1.3.1 Archive File

Things like data types, column names and wrong dog names' were changed or deleted as part of the quality issues. Tidiness issues were solved by combining columns that only contain one variable for instance.

1.3.2 Image Prediction Data

Here the dogs' names were change to capital letters and by removing the '_'. Also, the no columns that were not needed were deleted.

1.3.3 Tweet Json Data

This file was not cleaned. The needed columns were already in good shape.

1.4 Joining the tables

After the cleaning I joined these tables. The tweet_id was used as the key. That created a well stuctured and high quality data set.