Data Wrangling Report

By Hani Jubail

Gathering Data for this Project

The First Step in the Data Gathering process is to download twitter_archive_enhanced.csv From The Udacity Classroom .

Then upload it to my workspace online so i can work with it later.

The Second step is to programmatically download image_predictions.tsv file using using the Requests library and then save it to my workspace .

The third step is to get additional data from twitter using Python's Tweepy library such as “retweets count and Likes “ , then i saved in a file called tweet_json.txt.

Assessing The Data

After gathering all the data from the different sources i saved each file to a dataframe .

- twitter_archive_enhanced.csv to twitter_archive
- Image_predictions.tsv toimg_predictions
- tweet_json.txt to tweet_json_df

Then i looked at each data frame using dataframe.info() , dataframe[‘column_name’].value_counts(),etc.

I uploaded the original FIles to Google Sheets so i can make a visual assessment.

During this process i was taking notes on the problems that i found and what needs to be done in the next steps.

Quality Issues

`twitter_archive` table:

- tweet_id column needs to be a string not int.
- timestamp column needs to be a datetime object and we remove the 0000 part from it .
- rating_denominator column need to be equal to 10 in all rows .
- rating_numerator column contains some outliers that we should take them into consideration while analyzing the data .
- name column have missing names and some names are wrong like "Such","a","old,"an","very".
- we need to extract the part between "><" in the source to get the actual source

`img_predictions` table:

- the img_prediction table contains only 2075 so there is missing data .
- tweet be a string.
- we will capi_id needs totalize and remove "_" from the dogs breeds in these columns P1,P2,P3 .

`tweet_json_df` table:

- id_str column needs to a string.

Tidiness Issues

`twitter_archive` table:

- retweeted_status_id retweeted_status_user_id and retweeted_status_timestamp columns will be dropped from the dataframe.
- doggo,floofer,pupper and puppo columns will be converted to one column named "DoggoLingo"

All tables will be merged into one final table for analysis.

Cleaning The Data

The first step of cleaning is to take a copy from the dataframes .

Then i defined each Issue and then wrote the code then tested it .

The hardest issue is Combining doggo,floofer,pupper and puppo columns to one column named "DoggoLingo", because it took me a lot of time and search to find the right steps to get the final column Cleaned.

After Cleaning the data and merging all the datasets i saved the final dataframe to a CSV file named “twitter_archive_master.csv”.

Data Analysis

After cleaning and merging the data frames i started the data analysis process .

1st step visualising some variables .

2nd step visualising two variables to look for correlations between them .

3rd step making queries and groups to get the correct plots from the data .

4th step is writing each observation and each plot to explain the results .