wrangle_report

September 5, 2020

1 wrangle report

In this file, The Wrangling process undergone in this project is described which include: Gathering the Data, Accessing, Analyzing and Visualising it.

1.1 Data Gathering

Using what I learned from the lessons and other websites, the data was gathered from three different sources as follows:

Twitter Archive: This was given by udacity as downloadable csv format, I just call the functions to use it in the jupyter notebook. Image Prediction: The url of this file is also given by udacity, I downloaded it using the given url and upload it in the jupyter notebook for usage. Live Twitter Data: This data was to be generated from the twitter using tweepy but I chose to use it offline due to the fact that my twitter developer account was not approved and the time was tight.

1.2 Accessing Data

The data was assessed both visually and programmatically to identify potential data tidiness and quality issues that would be preventing any meaningful data insight analysis from happening. the issues identified are:

1.2.1 Tidiness Issues:

- doggo, floofer, pupper and puppo coluns should be combined because they are all states.
- tweet_data table, image prediction table and twitter achive table can be combined to form a single dataframe.

1.2.2 Quality Issues:

twitter_archive table

- columns such as in_reply_to_status_id, retweeted_status_id can be dropped.
- records with no image url can be dropped because the rating is invalid without an image.
- tweet_id is of type int, timestamp and source needs to be converted to string, time and category respectively
- source should be human readable
- names should be consistently capitalized and standardized, mis-spelt/incorrect dog names and should be tackled.

1.2.3 image_predictions table

- Convert tweet_id to type string
- columns such as p1, p2 and p3 should be renamed.
- capitalisation of the prediction dog breeds should be consistent.

1.3 Data Cleaning

The identified quality and tidiness issues identified were addressed and a test was issued to make sure the result was the desired one.

The 3 datasets was merged into one single master dataset and stored it to a csv file.

1.4 Data Analysis

The master dataset was copied and insight, analysis and visualization were carried out on itto produce a report documenting the wrangling effort and another, three insights were communicated which include:

- correlation between rating and favorites count
- correlation between rating and retweet counts
- correlation between favorite and retweet counts.