Wrangle and Analyze Data

By Dana alqahtani

WeRateDogs

Introduction:

This project focuses on wrangling and visualizing a dataset that has been gathered from different sources. In order to explore a twitter account called @WeRateDogs where the account post a tweet of a dog and letting the audience rate.

Gathering data:

The data used in this project was collected from three different sources:

- twitter_archive.csv given from Udacity.
- Image_predictions.tsv extracted from Udacity sever.
- tweet.json scraped from Twitter api.

Assessing data:

This step was divided into two assessments:

- Visual assessment

This assessment was done by viewing different records in the data (the first 20 records, a sample of 20 records, and the last 20 records for each dataset gathered previously.

- Programmatic assessment

This assessment was done by writing python code to view the basic info of each dataset, checking null values, and duplicated records.

Summary of findings:

Quality issues

Twitter_archive .csv

- The tweet_id as well as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, and retweeted_status_user_id data type needs to be string.
- Timestamp and retweeted_status_timestamp data type needs to be timestamp.
- expanded_urls has only 2297 entries, so we don't have image url for 59 entries, delete those records.
- name column has irrelevant entries (letter a instead of names)
- Drop unnecessary columns such as source, in_reply_to_status_id, in_reply_to_user_id, expanded_urls, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.

image_prediction.tsv

- **Remove false predictions for dogs.**
- **Extract breed from the p1, p2, and p3.**
- **Convert the values of p1,p2, and p3 to lowercase letters.**
- **Rename column names to more readable names.**

Tidiness

- **Create a categorical attribute of dog_stage which contains doggo, floofer, pupper, puppo.**
- **Merge the three datasets in one dataframe.**

Cleaning data:

This step is manly fixing what was identified previously in the gathering stage.

- **Changing data types for tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp , and timestamp.**
- **Delete expanded urls where the url doesn't contain any image.**
- **Delete any irrelevant entries from the name column.**
- **Delete useless column from the dataset such as source, in_reply_to_status_id, in_reply_to_user_id, expanded_urls, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.**
- **Changing the headers name to more readable names.**
- **Remove false predictions for dogs and only keep the true prediction.**
- **Create a new column where the breed has been extracted from the p1, p2, and p3.**
- **Convert the values of p1,p2, and p3 to lowercase letters.**
- **Create a categorical attribute of dog_stage which contains doggo, floofer, pupper, puppo.**
- **Merge the three datasets in one dataframe.**