WRANGLE REPORT

Abstract

This is the report of dog dataset. It covers the gathering stage, assessing and cleaning stage.

Nkemakolam Onyemachi

Executive Summary

This project was about finding insights from three datasets about dogs extracted from twitter user name @weratedogs. The first dataset was archived dataset from the twitter username mentioned, the second dataset was image predictions using neural network about the dog specie and the third was me extracting counts of retweets and favorites using the twitter id from the first and second datasets. They were gathered and assessed and then cleaned to understand more about the data. In this report, gathering, assessing and cleaning the data will be discussed.

1.0 Gathering Data

The datasets were gathered from twitter. In this project, three datasets were used; twitter-archive-enhanced.csv, image_prediction.tsv and tweet_json.txt. These were all gotten from the username @weratedogs. The first two datasets were already given to us by Udacity, while the third, I got access from twitter via their API to gather the data.

2.0 Assessing Data

The three datasets which were in csv, tsv and txt format were converted to dataframes to enable wrangling. The first dataset had 2356 rows and 17 columns; it was named dog_archive. The second dataset had 2075 rows and 12 columns; it was named dog-_images. The third dataframe from the had 2327 rows and three columns, it was name apiData.

After assessing the dataframes, these were the issues that were noted.

Quality issues

Dog archive dataset

- Timestamp is of object datatype
- Name column has values that are 'None' instead of 'Nan' to show that they are null values.Some names with lowercase do not look like dog names.
- 'None' is used here instead of 'Nan' in the doggo, floofer, pupper, and puppo columns.
- Some dog name has less than two letters as names
- Most of retweeted_status_id, retweeeted_status_userId, retweeted_status_timestamp,in_reply_to_status_id and in_reply_to_user_id are null values

Twiter Prediction Image Dataset

- Duplicate images
- Missing rows in images dataset; There are 2075 rows instead of 2356
- Names in p1, p2 and p3 contain underscores instead of spaces

Twitter API data

- Missing tweets compared to the archive

Tidiness Issues

- The four columns doggo, floofer, pupper, and puppo seems vague. It can be as a single column as a category.
- Dataframe can be merged into one.

3.0 Cleaning Data

Cleaning the data meant taking care of the quality issues and the tidiness issues that were noted above.

i. Creating copies of the dataframes in order to work with the copies.
ii. Merging the three dataframes into one.
iii. Removing duplicate images from the dataframe
iv. Replacing the underscore issue from p1, p2 and p3 columns to space.
v. Drop columns such as retweeted_status_id, retweeeted_status_userId, retweeted_status_timestamp,in_reply_to_status_id and in_reply_to_user_id. This is due to the fact that they had more null columns and that made them not necessary to draw insights from them.
vi. Converting the Datetime column which was in string to datetime format.
vii. Removing names that just had one letter as a name and names that were in lowercase. The lowercase names investigating it looked more like verbs and not nouns not to talk of names.
viii. Fixed issues with the rating column