Wrangle Report

13 OCT 2020

Wrangle and Analyze Data

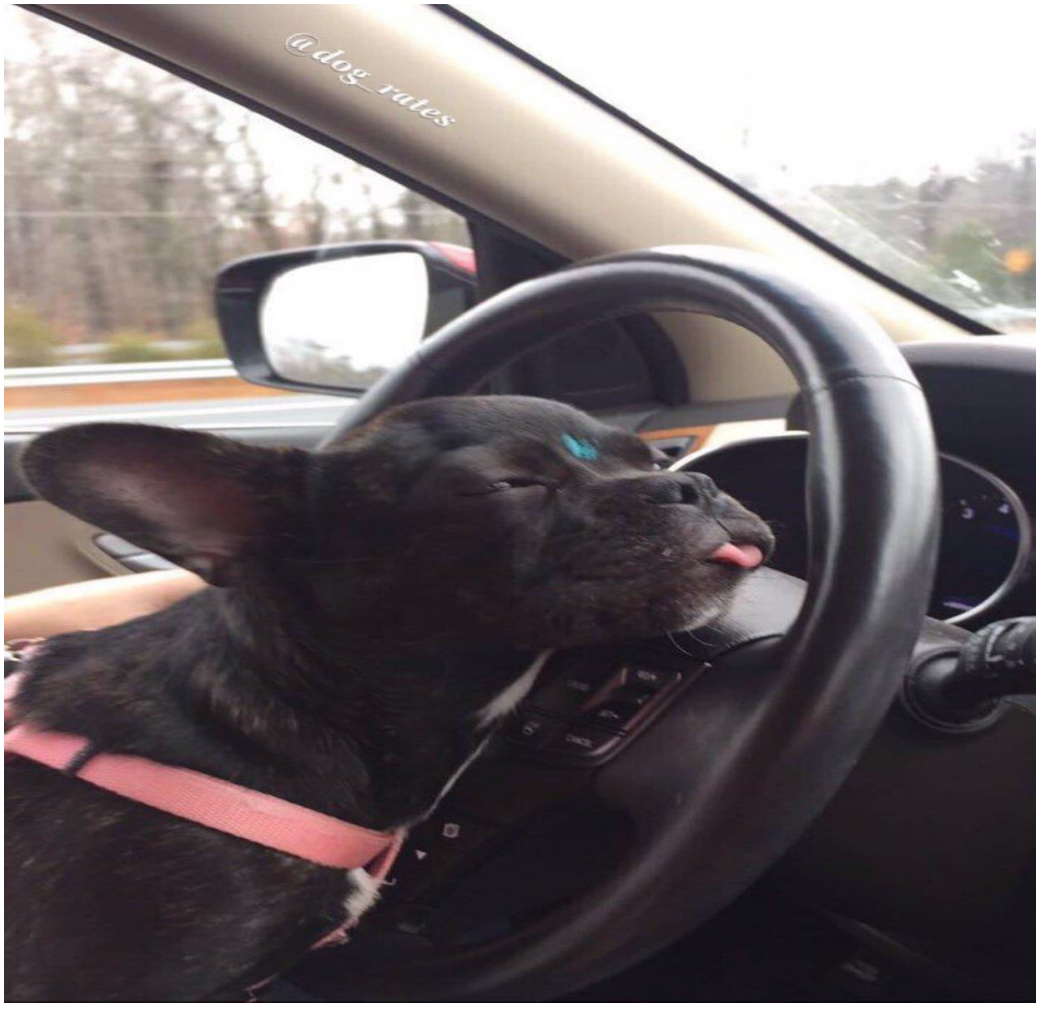
Prepared by

Mohamed fathy

This report is part of data wrangling Udacity nanodegree program projects

Heading 1

WeRateDogs is a twitter account that rates people's dogs with humorous comment about the dog.



WeRateDogs

The dataset wrangle in the project is the tweet archive of Twitter user, known as WeRateDogs. WeRateDogs is a twitter account that rates people's dogs with humorous comment about the dog.

Report goal

The WeRateDogs Twitter project goals included:

1- Wrangling the twitter data through the following processes:
2- ▪ Gathering Data
3- ▪ Assessing Data
4- ▪ Cleaning Data
5- Storing, analyzing and visualizing your wrangled data
6- Reporting on the data wrangling efforts and data analyse and visualization

1-Gathering

Wrangling efforts for the WeRateDogs Twitter project included gathering data from the following sources:

a. The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets
b. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students.
c. Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite count at minimum, and much more, I should request it by api , but I couldn't due to some error with api developer account and shortage time . so I used the one provided by Udacity class room resources.

2- Assessing

Quality issues:

"twitter _archive_enhanced.csv"

Completeness:

- Missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls.
- Tweet_id is an int (all tables)

Validity:

1. Dog names: some dogs have 'None' as a name, or 'a', or 'an'.
2. This dataset includes retweets, which means there is duplicated data (as a result, these columns will be empty: retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp).

Accuracy:

1- retweeted_status_timestamp is also an object (the other retweeted statuses are floats).
2- Timestamp is an object

Consistency:

1. The Source column still has the HTML tags
2. ▪ rating_denominator should be a standard 10, but there are a multitude of other values

'image_predictions.tsv':

Validity:

1) p1, p2 and p3 columns have invalid data...why would the algorithm labelled a dog photo as a starfish, boathouse, or mailbox.

Consistency:

1) p1, p2 and p3 columns aren't consistent when it comes to capitalization: sometimes the dog breed listed is all lowercase, sometimes it is written in Sentence Case.
2) ▪ In p1, p2 and p3 columns there is an underscore for multi-word dog breeds.

'tweet_json':

Completeness:

- Missing Some Data

Tidiness Issue:

twitter-archive-enhanced.csv:

- The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo).

‘image_predictions.tsv’:

- This data set is part of the same observational unit as the data in the **‘**twitter-archive-enhanced-2.csv’ - one table with all basic information about the dog ratings.

‘tweet_json’:

- This data set is also part of the same observational unit - one table with all basic information about the dog ratings.

Cleaning Data:

Define , Code and Test

- Merge the clean versions of archive, images, and twitter_counts_df data frames Correct the dog types.
- Create one column for the various dog types: doggo, floofer, pupper, puppo Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
- Delete retweets.
- Remove columns no longer needed.
- Change tweet_id from an integer to a string.
- Change the timestamp to correct datetime format.
- Correct naming issues and Standardize dog ratings.
- Creating a new dog_breed column using the image prediction data.