wrangle_report

July 31, 2022

1 Wrangle Report

1.1 Introduction

We Rate Dogs is a Twitter Account that provided us with an archive containing data from 2017 of Tweets of dogs, with ratings. The data set comes messy and untidy and I am tasked to clean the dataset to be ready for analysis. The process involves incorporating 3 data sources - Twitter Enhanced dataframe, Twitter API & Image prediction file.

1.2 1) Gathering the Data

There are 3 sources of which we utilised to gather data on the We Rate Dogs Tweets.

1) Twitter_archive_enhanced.csv was provided by Udacity. The CSV was loaded into a dataframe called "archive"

2) Image_predictions.tsv was provided by Udacity. The TSV file was loaded into a dataframe called "predictions"

3) Tweepy API which was an additional datasource which we exract te favourite counts and retweet counts. This loaded into a dataframe called tweet_stats

1.3 2) Assess the data

Using visualising and programmtic assessing to find issues in the data. The following issues where identified:

Quality Issues: 1) Archive dataframe - tweet_id is an integar and needs to be a string data type. Timestamp is a string and needs to change to date data type.

2/3) Archive dataframe - The dataframe is containing information that is not needed in the data - retweets & replies. (2) remove retweets & (3) replies. Drop the fields that relate to reply & retweet data

4) Archive dataframe - All denominators need to be 10 - change denominators that are not 10 to 10

5) Archive dataframe - 'name' column - there are names that are not nessecarily names in the values. Names starting with "a" or "an", "the" and "my" which should rather be "None"

6) Predictions dataframe - There is some tweets that is predicted that it is not a dog, this data needs to be removed. Remove none dogs from the dataframe (column - p1_dog, p2_dog or p3_dog). Combine into one column called "breed" and with its associated confidence

7) Tweet_stats dataframe - ID is a integar data type where it needs to be a string. For joining purposes, it is best that 'id' be renamed to 'tweet_id' and so that there is no confusion in understanding that those columns contain the same information. Convert ID to string data type & rename to tweet_ID

8) Predictions dataframe - tweet_id a integar and needs to be a string. Change tweet_id to string data_type

Tidiness Issues: 9) Archive dataframe - doggo, floofer, pupper & puppo are in 4 seperate columns - should be in one column & categorical

10) To compare dog scores relativey to one another it is best practise to create a score of percentages to compare. Create column with rating in archive daraframe using the rating_numerator & rating_denominator

1.4 3) Clean the Data

Before cleaning the data, a copy of the dataframe needs to be made: 1. Archive_clean 2. Predictions_clean 3. Tweet_stats_clean

The following needs to be done: 1) Archive_clean = tweet_id needs to be a string data type. Timestamp needs to change to date data type. 2) Remove retweets from archive_clean dataframe 3) Remove replies from archive_clean dataframe 4) Archive_clean dataframe - change denominators that are not 10 to 10 5) Archive_clean dataframe - name column - make names starting with "a" or "an", "the" and "my" which should rather be "None" 6) Predictions dataframe - remove none dogs from the dataframe (column - p1_dog, p2_dog or p3_dog). Combine into one column called "breed" and with its associated confidence 7) Tweet_stats dataframe - convert ID to string data type & rename to tweet_ID 8) Predictions dataframe - change tweet_id to string data_type

9) Archive dataframe - doggo, floofer, pupper & puppo should be in one column (breed) & make the column categorical. Drop unnessecary colums once breed column is created.
10) Create column with rating in archive daraframe using the rating_numerator & rating_denominator

1.5 4) Store Data

Create one master CSV with all the data_frames combined - joined in the tweet_id primary key. Call the CSV: twitter_archive_master.csv