wrangle_report

February 15, 2019

1 Wrangling steps:

1.1 Gathering Data

1- Read "twitter-archive-enhanced.csv" file and save it as df1

2- get "image_predictions.tsv" file using requestes package, then read it and save it as df2

3- Query Twitter's API for JSON data for each tweet ID in the Twitter archive using tweepy package, then save json file as df3

1.2 Assessing Data

Quality issues

1- Timestamp column should be date & time datatype.

2- the Denominator should be 10, yet the min value is 6.7 and max value is 170. These rows must be identified and cleaned.

3- since Denominator is not always 10, new coulmn "rating" should be added for fair comparison

4- some dogs are not associated with any stage and some do not have names

5- some dogs are associated with more than one stage

6- new col satge should be category

7- p1, p2, p3 columns values should be categories

8- p1, p2, p3 columns values should be in lower case

Tidiness issues

1- dog stage in the first dataframe (doggo, floofer, pupper, puppo) should be one column since the dog can only be at one stage.

2- There are some tweets that contain rating for two dogs. Each row should represents a tweet that includes one dog rating

3- all the three dataframes represent tweets, therefore they should be combined in one dataframe after eliminating unwanted columns

1.3 Cleaning Data

1- change Timestamp column to date & time datatype.

2- fix numerator & denominator values where rating_denominator does not equal 10

3- add new column to calculate rating, which solves the issue with different denominator

4- delete rows where dogs are not associated with any stage (not accomplished)

5- delete unwanted columns from the first dataframe

6- fix rows where a dog is associated with many stages

7- add new column stage and delete 4 columns (doggo, floofer, pupper, puppo)

8- merge the first and second dataframes

9- delete unwanted columns in the third dataframes

10- merge df3 with df4. The result is one dataframe that contains all the needed columns from the three files. Each row represents a tweet for one dog rating

11- change stage, p1,p2,p3 columns to categories

12- change p1,p2,p3 values to lower case because some start with upper case and some do not, which would create issue when we group by breed

In [ ]: