Data Wrangler process

This document contains information about the process of gathering, assessing, cleaning and storing data as part of the process of data wrangling. The data source used is WeRateDogs, which is a Twitter user rates for dogs.

Step 1: gathering data

The three data sources used in the project are listed below

- Enhanced Twitter Archive(twitter_file)
- Additional Data via the Twitter API(tweet_json_fil)
- Image Predictions File(prediction_image)

Step 2: assesing data

In this step the gathered data is discovered and data issues like: missing data issue, quality issue and tidiness issues are detected. Two ways of assessing data is used: visualisation method and programming method.

For the twitter_file, tweet_json_file and prediction_image, I checked the data types, information in each column, Null values, information no related to dogs, inconstancies and inaccuracies. The quality and tideness iusses for each file is stated below.

Summary of quality and tidiness data issues

Data quality
twitter_file

- O1: Error in data type for column: timestamp. Current data type object, it should be datetime
- O2: Values in rating denominator are not accurate. There are values of 170, 150, 130. The common range is 10
- O3: Duplications are found in numerous urls and there are two different Urls in the same line
- O4: Remove retweet no relevant for our analysis
- O5: Columns with no images
- O6: The tweet_id columns must be string not integer
- O7: Source information is no easy to read. There are numerous characters no readily identifiable
- O8: Original images should be included which are relevant for our analysis. Remove columns related to retweets and replies
- O9: Only columns relevant for our analysis should include. Remove no relevant columns

prediction_image

- O11: Duplicated image in urls
- O12: Remove the column img_num is not relevant for analysis
- O13: Change the information in each column separated by underscore to space
- O14: Capitalisation in columns p1,p2 and p3 are not consistent

- O15: Inaccurate entries. Some of the observations are not related to dogs but objects such as mailbox, etc
- O16: Predictions p2 an p3 contains errors. Keep prediction 1
- O17: Change name P1 column to ease analysis
- O18: Column with most reliable image prediction by dog categories is considered to support the analysis

tweet_json_file

- O20: Observations no related to dogs

Tidiness

twitter_file

- O10: Categories of doggo, floofer, pupper, puppo are all stages of dog. This should be merged in one column

prediction_image

- O19: Datasets require merging with twitter_file

tweet_json_file

- O21: Datasets *tweet_json_file* require merging with twitter_file

Step 3: cleaning data

I addressed the issues that impact my analysis. I use the programmatic data cleaning approach that comprises three steps: define, code and test. The initial work is to copy the original datasets to avoid overriding the data. Each issue noticed in the assessing stage is documented the particular [problem is defined, a code is generated to fix the issues and then the code is tested to check the expected results. This process it iterative. The tools used to fix these issues are python libraires.

Step 4: cleaning data

The master file with cleaned and merged files is stored in Jupiter notes as a csv file.