Wrangle report

This report highlights the data wrangling efforts done on the WeRateDogs twitter dataset in the wrangle_act.ipynb notebook.

Gathering Data

The dataset was gathered through the following methods:

1. Downloaded file – twitter_archive_enhanced.csv
2. Udacity hosted file – image_predictions.tsv
3. Twitter API – tweet.json.txt

Downloaded file

The Pandas library was used to read the csv file into a dataframe

Udacity hosted file

The requests library was used to download the files from the Udacity servers. The downloaded files were then used to read the tsv file into a dataframe

Twitter API

The tweepy and json libraries were used then the tweets used to create a txt file. The text file is read to make favorite_count, retweet _count and tweet_id into a dataframe.

Data assessment

At this point the data has been gathered into separate dataframes, then the data is assessed programmatically and visually to look for tidiness and quality issues.

Programmatic methods:

1. .head()
2. .info()
3. .describe()
4. .sum()

5. .value_counts()
6. .duplicated()
7. .query()

The quality issues were in the following categories; validity, consistency and completeness. The tidiness issues were categorized by tidy data principles.

Quality issues

Validity

1. df_te: Retweets may capture the same dog twice with a different tweet_id
2. df_te: Replies do not have images
3. df_tsv: 324 predictions where the top 3 predictions are not dog breeds.

Consistency

1. df_te: Source displays url
2. df_te: Timestamp column is a string

Completeness

1. df_te: Missing and incorrect dog names
2. df_te: Benebop Cumberfloof not identified as floofer

Tidiness Issues

Each variable forms a column

1. df_tsv: Four columns for stages of dog (doggo, pupper, puppo, floofer) should be one category column

Each type of observational unit forms a table

1. df_tw: Retweet and favorite should be appended to df_te table
2. df_te: Obsevational unit is for image prediction, jpf_url shoule be part of the df_te table

Cleaning

This section explores the cleaning done on the dataset, the problems and improvements.

Problem 1: Incorrect and missing dog names

The beginning of the dog's name in each tweet in the dataset begin with "This is …".

The previous efforts to gather this data appeared to have taken note of the pattern, they were able to capture most of the names by capturing the dog name after the words "This is …".

In the event the tweet did not begin with "This is …" the default name was "None". This results to 745 records where the name of the dog is "None".

If you look further into the data, if the name of the dog was in lowercase, it was likely incorrectly labelled.

This method also explains why "a" is the second most dog name. In an example, if the tweet began with "This is a good boy …" then the function assigned "a" as the name of the dog.

In the cleaning process, a trial to correct the name of the dog by filtering the incorrectly labelled tweets and finding the name of the dog in the body of the text.

Due to practicality and time constraints, the notebook only includes correction for dog names labelled as "a".

More work can be don't to extract the name of dogs correctly from the tweets

Problem 3: Extracting nested dictionaries/lists from JSON creates messy data

The twitter JSON files are complex and include nested dictionaries and lists. In the attempt to convert these complex JSON files into a dataframe, problems came up as some nested dictionaries have the same key.

In the attempt to normalize the JSON files, it resulted in many empty columns and series of lists that proved difficult to work with.

A solution that was arrived upon was to only extract the columns of interest.

Additional insights may come up from careful handling of JSON files from the Twitter API

Problem 6: Top predictions are not dog breeds

For the number of most predictions where no dog predictions came up, most the images did not have a dog in the picture.

Furthermore, in some instances where a dog is in a busy photo and dog breed is not predicted

In an example, a dog photo taken from behind and the face of the dog is in the reflection of the computer monitor. The top predictions were for items on the desk.

More training on the model may provide more accurate breed predictions.