WeRateDogs

Data wrangling project report

This project’s cleaning was divided to four parts:

- Gathering data
- Assessing data
- Cleaning data
- Compiling and storing all data

Gathering data

Data was given in three different sources:

- A CSV file “twitter-archive-enhanced” which was given in the classroom I just downloaded it.
- A TSV file “image-predictions” that I downloaded programmatically from this link given in the classroom. Link: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Last one I got from twitter using twitter API, I queried data using the tweet IDs given in the “twitter-archive-enhanced” CSV file after creating a twitter developer account and using API keys, secrets, and tokens. I saved the data in a TXT file called “tweet_json”

Assessing data

I used these function on all tables:

- .head(), to have a look at the rows and make sure data is good
- .tail(), to have more view of the rows in each data frame
- .info(), to know the types of each columns and the number of entries
- .columns, to make sure the columns are named right and to note the columns I don’t need and simply copy their names to use it in anther function to delete the ones I will not need.
- I noticed that some dog names in “twitter-archive-enhanced” was not actually dog names and the start with small letter so I used a function to get all name starting with small letters to delete them later

Cleaning data

I got the parts that needed to be cleaned and I divided them into two parts

- Quality issues
- Tidiness issues

Quality issues

twitter-archive-enhanced

1. Time stamp column in "twitter-archive-enhanced" is an object.

2. Source column has HTML labels in it '< a>'.

3. Column names are not in capital letters.

4. Dog stages columns has "none" in empty cells and are in 4 columns.

5. There are nonrealistic names in the column "name" like a, such, my,..etc.

image-predictions

6. Renaming dogs with no breed to Nan.

7. Dog breeds has "_" in it.

tweet_json

8. "id" column should be renamed to "Tweet_ID" to match other df.

Tidiness issues

twitter-archive-enhanced

1. Time stamp column should be split to "Day" and "Month" columns

image-predictions

2. Compiling 6 columns of dog breeds and confidence to 2 columns

3. Deleting useless columns

tweet_json

4. Removing all the columns I will not use

Compiling and storing all data

Finally, all sheets were merged using the tweet IDs in a CSV file called “twitter_archive_master”