Wrangling Report

Data gathering:

There were three different sources of data & each source needed different method to be loaded into a dataframe.

The data sources are as follows:

1. **Enhanced Twitter Archive**
   - Type: CSV.
   - Downloaded using: manually from Udacity.
   - Content: basic tweet data for all 5000+ of their tweets, but not everything.
2. **Image Predictions File**
   - Type: TSV.
   - Downloaded using: programmatically using the Requests library from web link
   - Content: images & what stage of dog is present in each tweet according to predictions & confidence level.
3. **Tweet Json file**
   - Type: TXT.
   - Downloaded using: tweet IDs in the Twitter archive are queried the Twitter API for each tweet's JSON data using Tweepy library which are stored in a file called tweet_json.txt
   - Read using: reading line by line to get tweet ID, retweet count, and favorite count and stored in csv.
   - Content after read: tweet ID, retweet count, and favorite count

Assessing Data

Once the three data frames were gathered, I start assessing using two methods

- **Visual assessment**: Done using excel sheets & notes

- **Programmatical assessment**: Done using different functions such as (info, duplicated, value_counts,etc)

I found the main three issues encountered to be missing data, quality and tidiness issues.

One interesting fact: Although I was done with assessing, after I start cleaning, I found more and more issues has become clearer such as images that are not for dogs.

Cleaning Data

I started by copying the three main data frames (good practice). This really helped a lot since while cleaning, I wanted to test some codes, made some mistakes such as wrong enters or wrong codes. So, each time I went back and load the copies.

Then, using the common way I divided each clean of issued points in assessing part into three main parts: Define, Code and Test.

There were some interesting cleaning codes. For instance, merging four columns respectfully from image prediction file to create dog stages and confidence level columns.

Conclusion

I like data wrangling since it made me feel that I have done somethings to improve the data for future. I am really looking for more data wrangling personal projects.

I think anyone interested in data analysis should master data wrangling first.