Data Wrangling Report

Project Description

The project is part of the Udacity Data Analysis Nanodegree in the Data Wrangling lesson. To have a meaningful dataset Udacity provided a tweet archive file with information for ratings of dogs from the twitter user @dog_rates « WeRateDogs ».

The report shows the efforts and experiences made in the wrangling process.

Gathering Data

- **Twitter Archive File**
  Udacity provided all students with a twitter_archive_enhanced.csv for manual download from: `twitter_archive_enhanced.csv`
- **Image Predictions**
  Udacity also provided us with the image_predictions.tsv file containing the output of a machine learning / AI algorithm to find out the dog breed of a dog on a picture on twitter. The file was manually downloaded from: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- **Twitter JSON API**
  To get additional data and learn using APIs programmatically in python we created a twitter developer account to get the credentials for our API calls. With the tweet_id contained in the twitter archive file I queried the API to get the entire stored JSON data for those tweets. This process takes about 30 minutes. I stored the results in the file: tweet_json.txt.

Assessing Data

The assessing part of the project was done visually first, just visualizing the files in the jupyter notebook. There might be better tools for this since jupyter always cuts lines if there are too many rows, but for a first quick glance it was enough.

In a further step I assessed the three data files also programmatically invoking methods like info, sample, groupby, value_counts, duplicated, head etc.

I summarized all my findings in an Assessment Summary directly in the jupyter notebook file. For a better readability I assigned IDs to the findings which I later addressed in the cleaning part of the project.

Cleaning Data

As described above I assigned IDs to my findings of the assessment step. In the cleaning step I rementioned those IDs for each section and always splitted the work in three cleaning steps: define, code, test. Before I started I copied over all dataframes to have a possibility to revert my changes in case of any issues.

Conclusion

The project was very helpful to further strengthen my data wrangling skills acquired in the lesson and all of the exercises. It again showed me that the major part of the work always comes during the wrangling process. Running the analyses and documenting the insights was much less time-intensive. However any small changes tot he whole process always require to rerun all the steps again and check each output and statement.