wrangle_report

August 10, 2022

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

This project consisted of performing data wrangling process on data from a twitter page called WeRateDogs. The gathering of the data was performed in three ways: - titter_archive_enhanced_csv file was provided to us on the Udacity platform and imported into a dataframe. - image_predictions.tsv file was imported via Udacity's servers and downloaded programmatically using the Requests library. - for the tweet_json.txt we were requested to query the Twitter API for each tweet's JSON data using Python's Tweepy library. I was however not able to connect to twitter and performed this step manually by reading the tweet_json.txt file line by line into a pandas DataFrame

After gathering the data we were required to Assess the data both visuallay and programatically. This part was quite difficult for me as beginner as there were a lot of issues to choose from but I was able to identify 8 quality issues and 2 tidiness issues by using methods such as info(), describe(), value_counts()and also scrolling through the data in excel as well as on the jupyter notebook. I also had at this point an idea of what type of insights and analysis I wanted to perform so I focused more on the dataframes and columns and rows containing this data to find issues.

For cleaning the data, it was important to first make copies of the documents before cleaning. For each quality issue I used the Define, Code and Test process in order to ensure that the steps were being followed: - twitter_clean dataframe: focused more on converting dtypes, removing null values, non-valid names, removing retweets and replies. - images_clean dataframe: removing false predictions - I did not perform any data cleaning on tweet_json.txt as I was not going to use this data in my analysis and wanted to focus more on the other dataframes.