wrangle_report

October 23, 2018

This project started with gathering data from three different sources described in project details section: Downloading file manually by clicking the link provided (twitter_archive_enhanced.csv), downloading programmatically using the Requests library (image_predictions.tsv) and using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file (tweet_json.txt).

Next step was to access data visually (in excel, google sheets etc.) and programatically (using info, head, describe etc. functions). Having a visual look on the data is kind of overviewing data by scrolling down and checking the data (like few numbers are float, few are integars etc.) and programatically assessment involves going deeper into the details like having a clear idea of what are the issues associated with our dataset and making a clear note of the different issues (quality and tidiness) associated with data at the end of the access section.

Next step was cleaning data which contains three steps: 1. Define 2. Code 3. Test

First step in cleaning preocess is the define statement which clearly states the issue associated with the data and how we are going to solve it (explain which functions/methods etc we are going to use) and next is the code to resolve that issue and last was the test which confirms that the issue has been resolved or not. If we get the correct answer it means we can preceed to the next step, otherwise we can go back and check what is wrong/missing. Not only after cleaning, at any step of the data wrangling process if more data is needed to analyse or any other issue comes up, we can go back to the previous steps and check upon the details required.

Next step was merging the necessary dataframes to a single dataframe and saving it as a master dataframe in csv format.

Nest step involved exploring the master dataframe and getting insights and visualizations from the final dataframe. In the EDA process we define and explain various insights and make plots wherever needed and explain those plots in detail.