wrangle_report

October 29, 2019

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. MY WRANGLING EFFORTS

Wrangling of the data all started by undertanding the context of the data which is about dog ratings based on an twitter dog group. The Major issues for me was during the data gathering phase of the project which was querying the twitter Api in order to extract the data for retweet count and favorite count. It was a length and tedious process especially as it needed some keys and access password which I hadn't at the time I began wrangling. I spent a full day trying to access the data before I found some keys on udacity knowledge hub page. Even after getting the access pass to query the API, twitter had a rate limit which meant that I had to wait for almost 7 intervals of 20 to 30 minute each for my data to successfully download.

After gathering my data, I began accessing it for data quality and tidiness issues both manually using microsoft excel and programmatically using Jupyter notebook. After drafting out all the problems that needed cleaning, I began to clean and tidy my data. The melting of the data was a little problematic given the values of 'none' that existed in the variable. I had to use Lamda which I really did not have a good working knowledge of at the time.

Just as the instruction in the course goes, I gave priority to tidiness issues first before cleaning the missing data and quality issues. This way, it was easy for me to clean up the data and thus, at the end, conduct my analysis and visualization. The tidiness isses involved:

Melting the twitter archive dataframe variable of doggo, puppo, pupper and friggo into the dog stages

Merging all the three data frames of twitter api, twitter archive and image predictions into one master data frame and then storing it to my local drive

The quality issues in details were enomous, more than the required 8. I spotted and cleaned close to 12 data quality issues ranging from:

Deleting rows with missing values

Changing the data types for timestamp from string to datetime

chaning the data types for the Id from integer to string

Eliminating duplicate data

Renaming and replacing values in different variables

In [ ]: