Wrangle Report

WeRateDogs Twitter Data , **Kareem Fouad Hamoda**

Objective

This short report describes the wrangling efforts involved in completing the “WeRateDogs” project as part of FWD Data Analysis Professional Track.

The Data Wrangling process consists of:

- Gathering the data
- Assessing the data
- Cleaning the data

Gathering

The data for this project consist on three different dataset:

1. **Twitter archive file:** the (twitter_archive_enhanced-2.csv) was provided by Udacity and downloaded manually
2. **The tweet image predictions File:** This file (image_predictions.tsv) was provided by Udacity and downloaded manually Or on Udacity servers and was downloaded programmatically using the Requests library and URL information
3. **Twitter JSON File:** (tweet-json.txt)was provided by Udacity and downloaded manually

Assessing

I used pandas to open and load the csv files and txt file into jupyter notebook and check the csv files in Excel

I used some Programmatically methods to explore the data well like(head(), describe(),corr(),sample(),info(),value_counts(),duplicated(), groupby(),) and from those I somehow be able to understand the three given datasets well and I made some EDA to explore the data and find if there is a relationship between any columns.

I Tested them visually and programmatically for accuracy and tidiness problems after gathering each of the above pieces of evidence. In your wrangle.act.ipynb Jupyter Notebook, detect and report at least eight (8) consistency problems and two (2) tidiness problems. The concerns that satisfy the project motivation (see the Main Points header on the previous page) must be analyzed in order to fulfill the requirements.

Cleaning

Cleaning the data for accuracy and tidiness problems is the final step in the wrangling process and this section of the data wrangling was split into three parts: describe, code and test the code. On each of the problems mentioned in the evaluation section, these three measures were discussed.

Creating a copy of the three original data frames was the first and very beneficial step. To manipulate the copies, I wrote the codes. If a mistake happened, I should have made a new copy of the original.
I was able to make another copy of the data frames if I made a mistake and begin working on the cleaning part.

Most of the cleaning was conducted using programmatic methods, such as def functions or built-in pandas (concat, filt, etc.), or buy filtering some columns from the three data frames together to get some useful insight but some manual cleaning was also done to fix ratings and dog type row errors.
After that I stored the data in new CSV file called ('twitter_archive_master.csv').

Conclusion:

Data wrangling offers a tidy data frame for future analysis and visualization, in our case we finished with 'twitter archive master.csv.' This file can even be shared with others without the need to wrangle the details.