Data Wrangling Report
Udacity WeRateDogs project

Purpose of the project:

Is to gather, assess, clean, store and analysis data from many sources, and find the present insights from the collected data using what I learned in Udacity Data Analysis program.

Project Details

The tasks of this project are as follows:

- Gathering Data
- Assessing data
- Cleaning data
- Storing, Analyzing, and Visualizing Data

• Gathering Data

For this part, data was divided into 3 parts,

1- Twitter archive file: was provided in the resources to be downloaded manually
2- Image prediction: was downloaded programmatically from udacity server
3- Twitter data: should have been downloaded by twitter API tweepy, but my account was not approved till this moment for the developer account, so I downloaded it manually

• Assessing data

After downloading all the data, assessing data is the next step using 2 methods:

1- Visually: I do that by previewing the data frame itself and search for points that grasp my attention
2- Programmatically: by using methods I learned in the Nano Degree, like (info, head, tail, describe, etc) those methods help to get overview on the data without reviewing them line by line

After assessing the data using previous methods, all problems were listed to be solved in the next step

• Cleaning data

Cleaning data consist of 3 parts, define, code, and test, define part already done in assessing data section, so next step is to start to solve the problems one by one.

Solving problems shouldn't be in the same order you spotted them, some actions should be done before others for data consistency

First thing to do is to have a copy of the data, not to ruin the original data, if any problem occurred you can just have another copy

One of the problem that faced me was the outliers in numerators rates, I wasn't sure if to remove them or to make them normalized to the mean or median of data, yet I choose to remove them, as they were minimal, and changing the rates would make the results inconsistent

Another good step was to merge all 4 dog stages into one column

- Storing, Analyzing, and Visualizing Data

First storing the data into new file
Then the Analyzing and Visualizing part was simple after all the previous efforts, using what I learned in visualizing the data, it gave a very clear insights regarding the data.

Finally

It was a great experience to start working on data from scratch, till I reach the final conclusion, I've learned a lot, and had the chance to apply it in a real life example.