WeRateDogs Wrangle Report

July 5, 2022

1 WeRateDogs Wrangle Report

1.0.1 Introduction

In the WeRateDogs project, I worked with three datasets: the enhance twitter archive dataset, image prediction file and additional data from twitter API. The enhanced twitter archive dataset contains basic tweets data for over 5000+ tweet from the WeRateDogs Twitter account. The image prediction file contains classified dog breeds alongside each tweet id, image URL and the image number. The data gathered from twitter API has basic retweet counts and favorite counts.

I performed data wrangling process on the dataset and in this report I will highlight the whole process. The steps are gathering, accessing and cleaning the data.

1.0.2 Gathering Data

- I directly downloaded the WeRateDogs twitter archive csv file and uploded the file to my notebook workspace and read the data into a pandas DataFrame.
- I also downloaded the image prediction tsv file from the URL that was provided by Udacity using the Request library.
- Encountered a challenge wile trying to access the twitter data therefore I opted to use the tweet json file that was already provided by Udacity. I proceeded to read the tweet json txt file line by line into a pandas DataFrame with tweet ID, retweet count and favorite count.

1.0.3 Accessing Data

> There are two type of assessing data which I used. Visual assessment and the programmatic assessment. Each piece I gathered, I displayed on my jupyter notebook for visual assessment and at first glance all I could some columns with missing values. I also opened the twitter archived enhanced csv file on excel worksheet and I noticed the many ’none’ values on the doggo, floofer, pupper and puppo which are the dog stages.
>
> With the programmatic assessment I used some pandas functions and methods to access the data. Some of these fuctions are .head() to access first 5 rows, .info() to get basic information, and .describe() to get statistical infomation. After all the accessment I documented the major issues I found on the datasets. Then proceeded to cleaning.

1.0.4 Cleaning Data

I made a copy of the original data before cleaning. Some of the issues I detected involved missing values, erroneous data types, invalid and inconsistence data and undescriptive columns names. For instance: the image prediction file had columns names like p1, p2 and p3 which are not clear variable names.

To clean the data, I droped missing values and replaced some, corrected erroneous data type, renamed undescriptive columns names and I used the define-code-test framework and documented it.After a thorough cleaning process, I merged the datasets and formed a master pandas DataFrame which I saves as twitter_archive master csv file.

1.0.5 Conclusion

The wrangling process is a rigorous task but it's very important in data analysis that will help present accurate and meaning insights and visualizations.