Wrangle Report

Introduction

In this report, we will describe the things I have done for the Udacity class Data Wrangling. For this project, we looked at the Twitter/X archive data for the WeRateDogs account.

The project had the following components.

- Gathering of the data
- Access of the data
- Cleaning of the data

1.) Gathering of the data

The needed information for this project came from multiple sources. Such as the original Twitter archive data, which we downloaded from the Udacity site and was then uploaded to my Jupiter notebook, the predictions data, and the Twitter data.

2.) Accessing the data

After gathering all the needed data, we had to assess the data both visually and programmatically. I found that there were some tidiness issues that needed to be addressed.

The final step for this project was to clean the data. The following issues needed to be addressed.

Twitter Archive

-- The twitter_id should be of the string typing instead of the int that it currently is.

-- The timestamp column is set to string and it should be a date time object.

-- Some values for both the rating numerator and denominator appear to not always be correct.

-- There are quite a few NaN values in the Twitter archive.

-- We can drop the unneeded column

Image prediction

-- The names in the P column should be standardized with an upper-case character instead of a mixture of the upper and lower case.

-- Once again, the tweet_id should be string and not int.

Tweet data

-- The tweet_id should be string and not int once again.

3.) Cleaning the data

Once I figured out all the issues I wished to address, I set myself to clean the data issues that were addressed in the previous stage. This took some work, but it was most definitely a learning experience.

4.) Conclusion

In conclusion, there was a lot of raw data in this project that needed to be cleaned, I felt like it was a quality representation of how the real world would work however, because there is so much data out there and the quality is often going to be broken and missing. It was nice to learn a little about how to gather needed data and clean it up so we can get a fuller picture.