Wrangling Report

Introduction

In this report I will briefly describes my wrangling efforts made in the Wrangle and Analyze Data project.

Gathering

There are 3 datasets was gathered for this project.

1. The “**twitter-archive-enhanced.csv**” data was provided by Udacity. I manually downloaded from this link, and uploaded to the Jupyter Notebook Workspace.
2. The “**image_predictions.tsv**” data was hosted in this link. The file was downloaded programmatically by using Request library.
3. The “**tweet_json.txt**” was populated by querying Twitter API via Tweepy library.

Assessing

I used visual and programatice assessments to identify quality and tidiness issues. Below is the issues I found from the data.

Quality Issues

twitter_archive

1. Some tweets are retweets.
2. Some tweets have no image.
3. Some rating denominator values range from 0 to 170.
4. Some tweets have rating numerator larger than rating denominator.
5. Both column (rating_numerator and rating_denominator) could be merged into a single column.
6. The datatype of 'tweet_id' should be string.
7. Values of 'a', 'an', 'the' and 'None' found in the 'name' column.
8. English word found in 'name' column such as 'very', 'getting', 'mad' and etc.

Image_prediction

1. The Datatype of 'tweet_id' should be string.
2. Columns (p1, p2, p3) has inconsistent capitalization.

Tidiness Issues

1. twitter_archive, image_predictions, and tweet_json can be merged into a single dataframe by joining on ‘tweet_id’..

2. 'doggo', 'floofer', 'pupper', and 'puppo' columns could be merged into a single column.
3. Some columns are not useful or unnecessary.

Cleaning

Final step in data wrangling is data cleaning. I try to clean on every issue I found from assessment step. In this project, I used programmatic method to clean the data. Before cleaning, the original pieces of data was copied. And then I followed below 3 steps to clean data programmatically:

1. Define: Describe the issue and explain how to clean.
2. Code: Convert the idea into program/code.
3. Test: Verify and validate the new data set.

Conclusion

After wrangling the data by above 3 steps Gathering, Assessing and Cleaning. We have a datasets that much easier to understand and ready to be analyzed.