Wrangle Report

1. Introduction

Data preparation is always the hardest part of a data analyst's work flow, in this project, we will use the data wrangling skills to pull real-world data from Twitter, clean it, and do some analysis. We will get the original Twitter data from Twitter user @dog_rates, along with a image prediction dataset, to build our analysis.

WeRateDogs is a popular Twitter hash tag, as the name tells, people rate dogs with a denominator of 10 and the numerator is usually higher than 10 to show how lovely the dog is.

2. Gathering data

The data was gathered from three sources:

- **Enhanced Twitter archive**: This part is kind of on-hand data, stored in the twitter_archive_enhanced.csv
- **Image prediction**: We get the image prediction data from web scraping
- **Twitter API**: This dataset is archived from Twitter's API and parsed from JSON to csv

3. Assessing Data

After assessing the datasets, I summarized several quality and tidiness problems of them:

Quality problems

df_twitter_archive:

1-remove tweets that has been replay as its not original.

2-remove tweets that has been retweet as its not original.

3-doggo, floofer, pupper, and puppy have values (1976) that are the string "None" instead of NaN

4-Correct denominators not equal 10.

5-remove tweets that has been numerator bigger than 14.

df_image_predictions:

1- drop duplicate jpg_url.

2-p1,p2 and p3 have inconsistent capital words.

3-p1, p2, and p3 contain underscores instead of spaces in the labels

Tidiness:

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types. In tidy data:

1-Merge three data frames.

2-Drop unneeded columns.

3-replace the url from value of source.

4. Cleaning data

The data quality and tidiness problems mentioned above was cleaned by multiple methods including pandas join, regular expression, combining multiple columns, pandas sub setting, removing missing values and so on.

In the end of this part, I stored the cleaned version of the data into a csv file for future usage.

5. Conclusion

This data wrangling project is a good practice of what I learned in the course, I've struggled to find subtle data problems, and cleaned those hard, time-consuming data quality problems. The hardest part of this project is the dealing with strings by using regular expressions.

In the end, I'm able to build a well-structured wrangling processing notebook with details in every part: exhaustive steps in how I gathered, assessed, and cleaned the data with every problem and its solving process illustrated.

The analyzing and visualizing part will be in another file called act_report.html