Wrangle Report

Introduction

This project put in practice what I've learned in 'wrangle and analyze data' section from Udacity Data Analyst Nanodegree program. The dataset that used in this project is the archive data of Twitter user WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. This report will describe how I wrangle the data in this project.

Gathering data

- **Twitter archive file**
  I downloaded the twitter_archive_enhanced.csv from Udacity website as in-hand data.
- **Twitter image prediction file**
  The image_predictions.tsv is on Udacity's servers and I downloaded this dataset programmatically by using Requests library and pass in the URL.
- **Twitter JSON file**
  I downloaded the tweet-json.txt from Udacity website manually due to the fact that I cannot set up a Twitter developer account. Then I read the tweet-json.txt line by line into a pandas dataframe with four columns: tweet ID, favorite count, retweet count and retweeted.

Assessing data

I assess the three dataframes both manually and programmatically by using different methods like info, value_counts, duplicated, isnull and so on. Then I summarize the quality and tidiness problems within these three dataframes as follows:

Quality problem
data_archive
Issue 1: Remove lines for no image
Issue 2: Correct lines' rating
Issue 3: Delete the 181 retweets rows
Issue 4: Delete columns with too many missing values or not useful for analysis
Issue 5: Split the 'timestamp' column into year, month and day three columns
data_prediction
Issue 6: Delete columns that are not used for analysis
Issue 7: Drop the 66 repeated rows of 'jpg_url' column
Issue 8: Create two columns for dog type prediction and confidence interval since some predictions are not dog
Tidiness problem
Issue 1: Melt the four columns of dog type into one column in archive data
Issue 2: Merge the three tables

Cleaning data

The data cleaning part was based on the former issues. For each dataset, firstly, I define the issue that I want to fix. Secondly, I create a copy of the dateset to run the data cleaning code. Thirdly, I test my cleaning effort by using methods like info, head and so on.

There are three challenging parts of data cleaning in this project as following:

- **data_archive**: Correct lines' rating
  I used iloc method to pinpoint the specific lines and correct the lines' ratings manually, this takes a lot of time and I wonder there might be a more efficient way to do this.
- **data_archive:** Melt the four columns of dog type into one column in archive data
  I used melt method to melt the four clolumns into 'dog' column and 'dog_type' column, then I dropped the 'dog' column and left the 'dog_type' column.
- **data_prediction**: Create two columns for dog type prediction and confidence interval
  I created a nested if-elseif-else loop to draw the correct information of dog type prediction and confidence interval.

Conclusion

Data wrangling is an important skill for data analyst. During this project, I became more familiar with many packages in Python like numpy, pandas, requests and so on. Furthermore, I was exposed to API and JSON file, I learned how to read, load and extract information form JSON file.

Limitation

There could more efficient ways to clean the data programmatically instead of fill in the correct data manually, I will keep enhancing my skills in this aspect.