DATA WRANGLING

Introduction

The aim of this project is to put into practice data wrangling techniques and used what I have learned in the data wrangle part of the Udacity data Analyst NANO DEGREE, the dataset that was wrangled is the tweet archive of twitter account known as WeRateDogs, an account that rates people dogs with ridiculous comments about the dog.

Project overview

The flow of steps in this project:

- Gathering or collecting the data
- Assessing the data
- Cleaning the data

Gathering the data

- Twitter archive file: A file that I downloaded it from the Udacity data analyst course twitter_arvhive_enhanced.csv file.
- Tweet image prediction: for this dataset I collected it using the request library with a given URL from Udacity data analyst. https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

- Twitter API & JSON: using the tweepy library that can return information about a tweet by its ID, I retrieved the favorite and retweet count of tweets that are in the twitter archive file ‘using the tweet ID in the file’ and then append the return information in a Pandas data frame, then store it in a tweet_json.txt and tweet_json.csv.

Assessing the data

I performed tow types of assessments Visual Assessment and programming assessments, and I discovered 12 issues not limited:

- **Quality issues**

1. There are 59 missing extended_URLs.
2. Names must be correct some are None and others are not true names such as 'a' and 'an'.
3. The source column in the df_twitter contain useless information, so it needs to be clear.
4. Change the column timestamp type to Date Time.
5. There are 66 duplicated jpg_URLs.
6. Columns that is unnecessary for the analysis need to be dropped.
7. Data of the retweets will be dropped since they are not the original data.
8. Change the data type of rating_numerator and rating_denominator to float instead of int.
9. putting the rating in a single column.
10. Combining the predictions of dog type.

- **Tidiness issues**

1. In the df_twitter there are four columns of dog stages, that can be represented in a single column.

2. All three datasets need to be merged.

Cleaning the data

In this section of the project, I solved the issues that I discovered in last phase, First I made a copy of each data frame I collected at phase one, then I start solving the issues, then I stored the clean data in twitter_archive _master.csv.

Finally, after data was clean and arranged, I made some visualizations and found insights you can find them at act_report.