Project Report

Wrangling and Analyzing Data Project

After reading the project overview I gained some knowledge about the data I have.

Gathering Data :

I used jupyter notebook to gather the data.
First I read the csv file provided called ‘twitter-archive-enhanced.csv’ using pandas library.
Then I programmatically downloaded the image prediction dataset (image-predictions.tsv) using requests , and then read it.
Then I was supposed to use the twitter API but because I don’t have the access I just normally read the text file provided by udacity. It was a JSON file so I used read_json() method provided by pandas library.

Assessing Data :

First I started by doing visual assessments using head() method and sample().
I immediately noticed that in the name column the missing values appead as a string value ‘none’ , some columns names had to be changed like ‘p1’ , and the dog stages had to be in one column instead of four.
After that I started assessing it programmatically using methods like info() , and dublicated(). And I noticed a lot of issues :

- timestamps data type are objects not datetime.
- tweet_id and other id values data type are int or float , and not object.
- drop the retweeted tweets because we don't want their ratings.
- dublicated urls in the column jpg_url.
- some id columns such as 'in_reply_to_status_id' and 'in_reply_to_status_id_str' are duplicated. since we don't need them .
- we have columns that only has null values in tweet_json dataset.

And also there is another tidenes issue which is that all datasets should be in one dataset.

Cleaning Data:

- I took a copy of the datasets I have so I don’t make mistakes in the original data I have.
- Then I started by changing the timestamp column to datetime data type using pandas method to_datatime().
- After that I changed the id columns to string using astype() method.

- Then I droped the retweets by dropping the rows that doesn't have the 'retweeted_status_user_id' column as null .
- Then I replaced the string value 'None' to NaN in the name column.
- Then I renamed the column that needed to be renamed like 'p1'.
- Then I dropped the dublicates in the image predictions dataset using pandas method drop_dublicates().
- After that I dropped the columns that only has null values in the tweet_json dataset.
- Then I dropped the columns that had a lot of missing values and we also won't use them.
- Then I pandas 'loc[]' method to fix the tidiness issue and make the four columns one column , and then I dropped the 4 columns.
- Then I simply merged the datasets into one dataset , and the merge was on the 'tweet_id' and 'id' in the tweet_json dataset.

After that I simply stored the data set into a csv file using pandas method to_csv() and called it 'twitter_archive_master.csv' .