Wrangle Report

Gathering:

Following the project instruction from Udacity, I've gathered data from three different sources.

First, Twitter_Archive_Enhanced.csv Which was provided on hand. I've downloaded it from the classroom then I imported the CSV file using pandas.

Second File image_prediction.tsv I've extracted the data programmatically using Python request library.

The third data supposed to be extracted from Twitter API using Python tweepy library but unfortunately, I did not get the authorization from twitter however I downloaded The file tweet_json.txt from classroom then I've read this file line by line to extract id tweet, favorites, and retweet counts for each tweet.

Finally, I've made a copy of those three data types before starting data cleaning.

Assessing:

I've asses the data frames Visually and Programmatically to combine quality issues and tidiness issues. here is what I found:

- Quality Issues:

- Twitter df clean:

1- Delete retweeted tweets rows because they are not an original tweet.

2- remove columns that are not needed for analysis.

3- Change timestamp from string to date time.

4- rating denominator column has numbers like (170,150,130) I think they are typo mistakes because the denominator has to be 10 and Rating numerator that having more than 10 will be decreased to 10 and values less than 10 will not be changed.

5- Extract texts from the source column.

- images clean:

1- Capitalize the first letter of p1, p2, and p3 because they have inconsistent capital words.

2- drop duplicate jpg URL.

3- Replace _ and - with white space in p1,p2, and p3 column.

- Tweet clean:

1- rename id to tweet_id to merge column.

2- changing tweet_id column from number to string.

3-remove all columns except ('id' , 'retweet_count' , 'favorite_count')

- Tidiness Issues:

1- combining dog stages to one column.

2- create a Date and time column to change from an object(string) to date-time format.

3- perform inner merge join between three data frames as they all have data for the same tweet.

Cleaning:

I've used basic python function to clean and fix all the qualities and tidiness issues like the drop, info, and value count. Using the drop function: - Retweeted tweets were removed because they are not original tweet and - columns that are not needed for analysis were removed

- changed timestamp from string to date time using pd.to_datetime.
- rating denominator was assign to 10
- numerator that having more than 10 decreased to 10
- Extract text from source column using .str.extract()
- Capitalize the first letter of p1,p2 and p3 using .str.capitalize()
- Drop duplicate using .drop_duplicates()
- Replace _ and - with white space in p1,p2 and p3 column using .str.replace()
- rename id to tweet_id to merge column
- changed tweet_id from number to string.
- removed all columns except ('id' , 'retweet_count' , 'favorite_count')
- combining dog stages to one column
- create Date and time column to change from object(string) to date time format
- perform inner merge join between three data frames as they all have data for the same tweet