Wrangling Data Report

Gathering Data

From the instructions on the Udacity project detail page I started by downloading the Twitter archive file ‘twitter_archive_enhanced.csv’ which was provided for download, secondly, the file ‘image_predictions.tsv’ was downloaded programmatically from Udacity server using request library and third file which caused me a lot of trouble was tweet_json.txt file using twitter API for @WeRateDogs I had applied to twitter for more than a week now but to no avail, they kept asking me all sort of questions. Finally, I handed the issue to my mentor and she directed me to use the predefined tweet_text.json file so I read this file line by line to create data frame for at least three columns id(tweet) , favorite and retweet counts. After having three dataframe(df) I make copies of them so if I make any changes it won’t affect the original ones. I called copied df_clean, image_clean and tweet_clean.

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. Each variable forms a column. 2. Each observation forms a row. 3. Each type of observational unit forms a table.

Tidiness issues 1. change tweet_id from number to string. 2. Newly created Date and time column needed to change from object(string) to date time format. 3. perform inner join between three data frames as they all have data for the same tweet.

Cleaning

I used basic python functions like duplicates , drop, sort , value_count ,describe , info and others to comply with the above mentioned point. I struggled with few issues and had to spend a lot of time to get my understanding. As little help was provided its first time I used so many websites for checking syntax and possible solutions.

Assessing Data

Asses df visually and programmatically and found lots of Quality and Tidiness issues but done minimum requirement for project 8 quality and 2 tidiness issues as below for cleaning Quality Issues. It mainly includes issues like completeness, validity, accuracy and consistency df_clean 1. remove tweet that has been retweet as it's not original. 2. combining dog stages to one column 3. remove columns that are not needed for analysis. 4. Change timestamp from string to date time and make separate columns for date and time. image_clean 5. p1,p2 and p3 have inconsistent capital words 6. drop duplicate jpg_url. 7. p1,p2 and p3 have unnecessary underscores instead of space. tweet_clean 8. rename id to tweet_id so can merge later.

Conclusion

I think only after learning thoroughly from Udacity platform I have started to grasp coding mindset but as I am changing career so still a lot more to learn.