Wrangle report

Gathring :

In the gathring stage, there were three files csv file, tsv file, Json file , and each file had different readers, for example a csv file It was read like this df=pd.read_csv('/content/twitter-archive-'enhanced.csv ')

And the tsv file, it was read in this way df_image=pd.read_csv("/content/image-predictions.tsv", sep='\t'), and the Json file , it was completely read in this way df_tweets=pd.read_json('/content/tweet-json .txt',lines=True)

Assessing

In the evaluation stage, it was a stage for me to discover quality problems and Tidiness problems , we discovered all files to search for any problem such as missing values, incorrect values, etc .

I will review the quality problems and Tidiness problems that I found :

Quality issues:

twitter-archive-enhanced:

- full text column was not fully displayed
- Column is hard to read source so I'll make it easy to read
- Change data type for some column from floot to integer
- Deal with a missing value in expanded_urls column
- Fix name column there is some name unacceptable
- Make all name(value) in name column start with Capital letter
- Change the data type of:timestamp ,retweeted_status_timestamp columns to datatime data type
- Drup that columns that i don’t needed in my analysis.
- Extract year column from timestamp column.

Image-predictions:

- Rename some columns : img_num, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog.

- Make all name(value) of prediction column start with Capital letter

Tweet-json:

- Drup that columns that i don’t needed in my analysis
- Rename some columns to Unify column names with other data-frame

Tidiness:

- Merging the three data-frame into one data-frame
- Combine the four columns: doggo،floofer،pupper،puppo columns into one column and make this column category data type and fixe an none value.

- We must merge two columns rating_numerator،rating_denominator into one column and separate them with a (/) sign to clarify the numerator and denominator

Cleaning

In the cleaning stage fix or solve all the quality problems and Tidiness problems that were previously identified , This part of the data divided in three parts: Define,was wrangling code and , test, Where I created another copy of my data frames

```
{
df_clean = df.copy ()
df_image_clean = df_image.copy ()
df_tweets_js_clean = df_tweets.copy ()
}
```

to continues to work perfectly on the cleaning part ,Then I took the clean merged file and ('saved it in ('twitter_archive_master.csv) .