WeRateDogs

Data Wrangling and analysis Project

Data Gathering

- Gathered Data from from 3 different sources:

1. Image prediction - used the request liberary to get the required tsv file.
2. Twitter archive - downloaded the csv file and loaded using pandas.
3. Json-data - I supposed to get this data using twitter Api but the permission to the developer account granted late so I downloaded the file as suggested within the project details then used a while loop to go through each line using readlines() to store the data from txt to columns.

Assessing Data

- I did use Microsoft-Excel to asses the 3 files visually used filters and some functions like count.
- Used some pandas function and methods to assess data on notebook eg”df.describe(), df.info(), df.value_counts().sum(), df.columns(), df.sample(), df.iloc[index].
- Used comment with each cell to describe the purpose of the cell.

Assessment findings and cleaning

- Image prediction.
- I found that columns names are not descriptive, and the a tideness issue that the last 9 columns needs to be merged into only 3 columns using the wide_to_long general function of pandas then to remove the duplicates.
- Twitter Enhanced archive.

- This file has some issues as follows:

Assessment findings and cleaning 2nd

1. timestamp column has " +0000" will be removed using .split().
2. 'timestamp" column will be converted using to_datetime().
3. "name" column has some names that are lowercased and not a valid names replacing with "None“ using a function lower_names with IF statement then replaced the resulted list with the column values.
4. Dropped 181 retweets.
5. Dropped 78 replies.
6. source column can be simplified using .extract() and a regex.
7. dropping the tweets with rating denomirator bigger than 10 as "not valid or many dogs“.
8. dog category columns will be merged in 1 column "dog_category using .replace() and “+” operator between columns.
9. Dropped empty columns with retweets and replies using twt_enh_c.drop(columns=[“columns names”], inplace= True)

Assessment and cleaning 3rd

- Json-data

- For this file I found that it should be a part of the twitter archive so I merged both using pd.merge() on tweet_id and a left join to keep all columns from twt_enh_c and add the 2 new columns from the json-data DataFrame.