Bashair Altalhi

Wrangle Report

In this project, I used python and its libraries to gather data from different sources and formats. Then, assess its quality and tidiness, after that clean it, analyses, and visualizations.

Data Gathering:

We used three types of data.

- **Twitter archive**

It is a csv file that is provided in the project. I upload it to Jupter Notebook by using pandas.read_csv.

- **Image predictions**

It is a tsv file that is provided in the project. I read the file with open as file then upload it by using pandas.

- **Twitter json**

It is a txt file. I download the data from Project page and open it with pandas.

Assessing:

- **Visual assessment:**

I explored the data by looking at each column and row. This step can be done by display each data.

- **Programmatic Assessments:**

I investigated the data deeply by using different commands such as describe() that display a statistical summary. Also, the command .sample() which illustrate a sample of data. The command .duplicated() to check for duplicate in information. Also, the command .info() shows the number of non-null count and the type of each row.

Then, I documented some of the quality issues and tidiness issues of the three data.

Quality:

- **Twitter archive data**
  - Missing value in the following:
    - 'in_reply_to_status' the total of values are 78 and it is supposed to be 2356
    - 'in_reply_to_user_id' the total of values are 78 and it is supposed to be 2356.

- 'retweted_status_id' the total of values are 181 and it is supposed to be 2356.
- 'retweted_status_user_id' the total of values are 181 and it is supposed to be 2356.
- 'retweted_status_timestamp' the total of values are 181 and it is supposed to be 2356.
- 'expanded_urls' the total of values are 2297 and it is supposed to be 2356.

- 'timestamp' is object not datetime.
- 'retwetted_status_timestamp' is object not datetime (removed later no need to change).
- ' name' some names are enter as 'Nane' and 'a'.
- 'rating_denominator' has max value of 170 which can not be right because these ratings almost always have a denominator of 10.
- 'rating_numerator' type is int not float.

Twitter json

- 'id' rename.

Tidiness:

- 'puppo', 'doggo', 'floofer', 'pupper' are dog "stage" should be in one cloumn.
- Merge df_archive, df_image, and df_tweets into one by tweeter Id.

Cleaning:

In this section of the project I cleaned each issue by first define the issue then write the code to clean it finally test if the code worked.

Analyzing, and visualizing the wrangled data:

The last step of the project after storing the wrangled data, I asked three question and answered it by visulazining.