#WeRateDogs Archive

This Analysis Done By

Zahra Abdelkareem Al-Hourani

Data Analyst Nanodegree Program

Udacity, One Million Arab Coder

July, 2021

In this report, I will discuss the Wrangling processes that I have to apply on the #WeRateDdogs Archive. These processes include the following:

1- Gathering
2- Assessing
3 –Cleaning

Let me explain this in detail:

1- Gathering

I extracted the data needed for this project from several sources, which are:

1- The WeRateDogs Twitter archive, This file is available as a link on the Udacity website. Then I upload it as 'csv' file in the Workspace.
2- Image Prediction dataset, I had download it programmatically from the Udacity server.
3- Twitter dataset, I have created a developer account, and I have obtained this data from Twitter API using tweepy.

2- Assessing

After gathering the data, I have to Assess it visually and programmatically, and I also noticed some Quality and Tidiness Issues with this data:

Quality Issues

Enhanced Archiv Tables

- The name column contains 'None' Value.
- The name column contains lowercase.
- The name column contains words that are not names like 'a', 'this', 'that', 'an'.
- The timestamp column is stored as a string.
- The Tweet_id is stored as int.
- Missing data in 'in_reply_to_status_id' , 'in_reply_to_user_id' 'retweeted_status_id' , 'retweeted_status_user_id' , 'retweeted_status_timestamp' , 'expanded_urls'.
- 'source' column contains HTML symbols.
- 'source' should be categorical data type.
-

Image Prediction Tables

- 'p1' , 'p2' , 'p3' columns have Lowercase.
- drop 'p1_dog' , 'p2_dog' , 'p3_dog' rows when all of columns containes. 'False' value
- Tweet_id stored as int.

Tweets Tables

- Tweet_id stored as int

Tidiness Issues

Enhanced Archiv Table

- Too many columns for the same information like 'doggo' , 'floofer' , 'pupper' , 'puppo'.

All Tables

- All Tables should be combined as one dataset

4-Cleaning

That is the final step of wrangling the data. Here I will discuss with you the steps,I follow when I clean the data.

Quality Issues

Enhanced Archiv Tables

- The name column contains 'None' Value.
- The name column contains lowercase.
- The name column contains words that are not names like 'a', 'this', 'that', 'an'

For these columns there is a lot of 'None' value about 745 records. Also, the 'name' column contains 'a', 'an', 'this' etc... this is not a valid values and will affect the validity of our data. So I decide to convert all lowercase names to 'None' and delete them.

- The timestamp column is stored as a string.
- The Tweet_id is stored as int.

'timestamp' column stored as a string. So I will convert it to date.
'Tweet_id' column stored as int. So I will convert it to string.

- Missing data in 'in_reply_to_status_id' , 'in_reply_to_user_id' 'retweeted_status_id' , 'retweeted_status_user_id' , 'retweeted_status_timestamp' , 'expanded_urls'.

These columns I don't need in my analysis anymore. So I decide to drop them all because they have a lot of missing data.

- 'source' column contains HTML symbols.
- 'source' should be categorical data type.

'source' column stored as a string. So I will convert it to categorical data type. The 'Source' column contains HTML symbols. That will affect the consistency of the data. So I decide to remove these symbols using replace function.

Image Prediction Tables

- 'p1' , 'p2' , 'p3' columns have Lowercase.

  'p1' , 'p2' , 'p3' columns have Lowercase I will make it uppercase for all columns

  using[.str.title()]

- drop 'p1_dog' , 'p2_dog' , 'p3_dog' rows when all of columns containes. 'False' value.

  'p1_dog' , 'p2_dog' , 'p3_dog' when these rows contain False. That means that the dog is not one of the expected types. So I have drop all of these values.

- Tweet_id stored as int.

  Tweet_id stored as int. So I will convert it to string

Tidiness Issues

Enhanced Archiv Table

- Too many columns for the same information like 'doggo' , 'floofer' , 'pupper' , 'puppo'.

I have combine all of these column in one column and drop the 4 columns.Then I noticed that there is a lot Of NaN. So I decide to replace it with None untile I solve this problem.

All Tables

- All Tables should be combined as one dataset

All the existing data belong to each other and belong to the same topic, so I decided to combine the three tables.

Conclusion

Data Assessing and cleaning is a process that can be repeated over and over.