WeRateDogs Analysis Documentation

By

Umar Fauwzziyyah Ozohu

June,2022

INTRODUCTION

In this report, we are going to document the process used during the project. The steps used in this project are:

-Data Gathering

-Assessing Data

-Cleaning Data.

DATA GATHERING

The data used in this project come from the sources listed below:

- twitter-archieve-enhanced.csv: The WeRateDogs Twitter Archive data was downloaded directly from Udacity web page
- image_prediction.tsv: which is gotten by using the request library to download it
- tweet_json.txt: Provides the tweet JSON data, using the JSON library to extract data like retweet_count and favourite_count

ASSESSING DATA

Data was assessed both visually and programmatically on all the three datasets. For the Visual Assessment the data was viewed through the jupyter notebook and Microsoft Excel. For the programmatic assessment functions such as describe(), info(), duplicate(), isnull() and samples() have been used. Here is the information gotten for the three dataset when info() was used:

Twitter Archive Dataset

```
#a summary of the data frame
twitter_a.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   tweet_id                    2356 non-null   int64  
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object 
 4   source                      2356 non-null   object 
 5   text                        2356 non-null   object 
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object 
 9   expanded_urls               2297 non-null   object 
 10  rating_numerator            2356 non-null   int64  
 11  rating_denominator          2356 non-null   int64  
 12  name                        2356 non-null   object 
 13  doggo                       2356 non-null   object 
 14  floofer                     2356 non-null   object 
 15  pupper                      2356 non-null   object 
 16  puppo                       2356 non-null   object 
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

Image Prediction Dataset

```
#Summary of the image Prediction data frame
image_P.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   tweet_id  2075 non-null   int64  
 1   jpg_url   2075 non-null   object 
 2   img_num   2075 non-null   int64  
 3   p1        2075 non-null   object 
 4   p1_conf   2075 non-null   float64
 5   p1_dog    2075 non-null   bool   
 6   p2        2075 non-null   object 
 7   p2_conf   2075 non-null   float64
 8   p2_dog    2075 non-null   bool   
 9   p3        2075 non-null   object 
 10  p3_conf   2075 non-null   float64
 11  p3_dog    2075 non-null   bool   
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

Tweet-Json Dataset

```
#Summary of the json tweet data frame
json_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 3 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   tweet_id        2354 non-null   int64
 1   favorite_count  2354 non-null   int64
 2   retweet_count   2354 non-null   int64
dtypes: int64(3)
memory usage: 55.3 KB
```

After assessing the data both visual and programmatically, this are the issues found. It was classified into quality and tidiness issues

QUALITY ISSUES

1. Missing Value (reply status id, userid, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)
2. Name column in twitter archive not clearly defined
3. Sources are defined with hyperlink (instead of URLs)
4. Incorrect datatype for column Timestamp (Datetime instead of object)
5. p1, p2 and p3 column name in Image Prediction not clear (should be renamed)
6. Filter out data above 08-02-2017
7. name= none is equivalent to a null value
8. In the expand URL column some rows have more than one URL as value

TIDINESS ISSUES

1 Necessary data frame Should mergered with twitter archive

2 doggo, floofer, pupper and poppo should be categorized

3 Column not needed should be removed

CLEANING DATA

A copy of all the dataset was made before the cleaning process started, copy was made incase we will need to refer to the original anytime. The ***Define-Code-Test*** Framework was used in the

cleaning stage. Define was use to explain an issue, Code is used to fix it and the Test was use to show the result of the code used for change.

SUMMARY

Python Libraries allowed us to have access to various data source, formats, methods and libraries. We used pandas to access and manipulate our data. Matplotlib was used for visualization, it also helps us understand our data better and draw some interesting conclusion.

Every step we took has improved the dataset quality and also tidy our dataset. We now have one dataset that is the combination of the 3 datasets after quality and tidiness check.