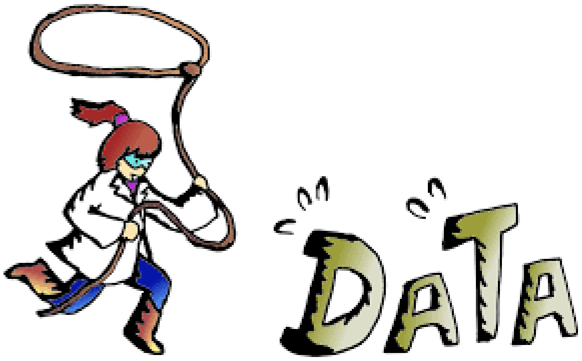
Project
Wrangle And Analyze Data



Wrangle_Report

Udacity Data Analyst Nanodegree
21.04.2021
By Yalçın FİLİZ

1.Introduction

Goal of this project is to get the “*WeRateDogs*” twitter account’s data and wrangle for analysis.

The Data wrangling process consists of 3 steps.I took the similar approach in this project.

1. Gather Data
2. Assess Data
3. Clean Data

2. Gather Data

There are 3 different sources for WeRateDogs data.

1. WeRateDogs Twitter archive

   Twitter_archive file was given in csv format on Udacity project page. I uploaded this file to the jupyter notebook in the project workspace. And then I made it into a dataframe using the *read_csv* method.

   There are tweet datas in this file such as id, timestamp, text and ratings.

2. Tweet image predictions

   I downloaded the image predictions file in tsv format from the link provided on the project page. I used pyhton request library for this. Later, I used the *read_csv* method to make this file a dataframe.

   There are dog image predictions and jpg ulr’s in that file.

3. Twitter API (tweet_json.txt)

   I needed a twitter developer account to get extra data with tweepy library (Twitter API). But I had some verification issues. That’s why I downloaded tweet_json.txt file from udacity web page as mentioned in the project's Twitter API section. I've copied the given code to the jupyter notebook as mentioned in the project's Twitter API instructions. I've studied how that code works and I got it.

   I had to convert the json file to a pandas dataframe. I used the *json.load ()* method for this.

   When we examine the file, we see that the numbers of retweets and favorites are for each tweet.

3. Assess Data

I used visual and programmatic assessment methods to find quality & tidiness issues.

I used methods like *head()* and *sample()* for visual assessment. Thanks to this, I had the chance to see the writing-related inconsistencies in values. I used the info*()* method to check data types & null value counts for each column. And also I checked if there are missing records.

For Programmatic assessment, the Duplicated(), counts(), sum() etc.. methods were used for check if there are duplicate records.

With value_counts() method, I noticed outliers and possible incorrect values for investigation.

While checking Quality issues, I also looked for tidiness issues in dataframes. And I determined development areas related to columns in dataframes.

4. Assess Data - Findings

Quality Items

Twitter Archive

- Tweet_id data type must be string
- timestamp columns data type must be datetime
- Source column are not easy to read
- There are 181 retweets
- There are some incorrect Rating (nominator & denominator) values. for example some denominators are greater than 10.
- There are so many 'a' in the name column.
- Name column has 'None' values instead of NaN
- **doggo,floofer,pupper & puppo colums data type to category**
- Missing record in expanded_urls (2297 < 2356)

Image predictions

- Tweet_id data type must be string
- **Missing records in image.There are more records in archive dataframe rather than image dataframe**
- writing is not consistent in p1, p2, p3
- **There are some non-dog images**

Df_tweets (tweet_json)

- Tweet_id data type must be string

Tidiness

- Type of dogs should be one column as category
- We can merge all data frames by tweet_id
- Remove unnecessary columns
- Create a new column (numerator / denominator) to compare ratings easily and for better analysis

5- Clean Data

First I copied all of the dataframes to a new dataframe before cleaning.

I followed the below steps for each issue.

- Define
- Code
- Clean

I started the cleaning process by clearing out incorrect data types and writing gaps.

I waited to merge 3 data frame into a single dataframe to fix some quality issues. However, I corrected many values such as rating and name column which were previously incorrect. Incorrect I mostly used the data in the text column to correct the values. And of course, I deleted the columns that I thought would not be useful during the analysis phase and combined some columns.

I usually used pandas methods to do this cleaning work. I manually replaced some incorrect data (ratings, name etc ..) with the correct data as much as I could, using the data in the text column.

And finally I stored cleaned data into a new csv file as ‘’twitter_archive_master.csv ‘’.

6 - References

https://jingwen-z.github.io/how-to-play-with-regular-expression-via-python/

https://stackoverflow.com/questions/25351968/how-to-display-full-non-truncated-dataframe-information-in-html-when-convertin

https://stackoverflow.com/questions/45416684/python-pandas-replace-multiple-columns-zero-to-nan

https://www.geeksforgeeks.org/python-pandas-series-str-extract/#:~:text=extract()%20function%20is%20used,match%20of%20regular%20expression%20pat.

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html

https://www.geeksforgeeks.org/python-pandas-series-str-contains/#:~:text=contains()%20function%20is%20used,of%20a%20Series%20or%20Index.