Report Data Wrangling and analyzing
WeRateDogs Twitter

By

Shahad Alotaibi

Project Overview

The objective of this project is to verify three basic operations in dealing with data, which are: Gathering data , Assessing data and Cleaning data. This is to reach more accurate data and give realistic results and solutions that can be used. In this project, we discussed obtaining Twitter user @dog_rates, also known as WeRateDogs.
Here in this project, I used the data that Udacity presented in the project file, because of the delay of Twitter in responding to the request for data, the rooms and the goal are the same in both cases

Project Details

- Data wrangling, which consists of:
  - Gathering data (downloadable file)
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting the final result

What is Data Wrangling?

Data wrangling is the process of cleaning, structuring and enriching raw data into a desired format for better decision making in less time.

Data wrangling is increasingly ubiquitous at today’s top firms. Data has become more diverse and unstructured, demanding increased time spent culling, cleaning, and organizing data ahead of broader analysis.

At the same time, with data informing just about every business decision, business users have less time to wait on technical resources for prepared data.

1-What is Gathering Data ?

Gathering Data is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

How to Gathering Data ?

We import this data

- Twitter archive file: download this file manually by clicking the following link: twitter_archive_enhanced.csv
- The tweet image predictions, i.e.This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Twitter JSON

We have used different methods to open these files

This is because the first file is an csv file, the second file is a link, and the third file is a Json file.

- How to gathering Twitter archive file using

```
twitter = pd.read_csv('twitter-archive-enhanced-2.csv')
```

- How to gathering tweet image predictions

```
response = requests.get(url)
```

- How to gathering JSON file

```
Json = pd.read_json('tweet-json copy', lines=True)
```

2- What is Assessing Data ?

Data Quality Assessment is a distinct phase within the data quality life-cycle that is used to verify the source, quantity and impact of any data items that breach pre-defined data quality rules.
The Data Quality Assessment is a task typically executed by dedicated Data Quality Software

We have analyzed the files that need to be analyzed and extracted the quality of the data and the problems inside the files to find out how to clean them and extract the problems they contain, and here we have limited a set of problems that the files dealt with twitter_archive_enhanced.csv and image_predictions.tsv

Quality And Tidiness Issues:

Quality Issues:

- Delete cells with a lot of missing values
- Drop values that contain repeated information
- Drop duplicates in jpg_url column
- Remove html tags in source column
- The columns (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) have only 181 values and 2175 missing values (Null values), and retweeted_status_timestamp Column has invalid format.
- Create a backup to preserve the original files
- The Column timestamp has invalid format, it should be a datetime type.
- The Column tweet_id Change from an integer to string .

Tidiness Issues:

- Combine “'doggo', 'floofer', 'pupper', ‘puppo' ”columns into a single column called “ Dog_Stage “
- Information about one type of observational unit (tweets) is spread across three different files/dataframes.
- Drop columns with missing values: 'in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id', ’retweeted_status_user_id’,‘retweeted_status_timestamp','expanded_urls
- The file has 2354 rows, the same of tweeter archive file, no issu Merge the columns (p1, p1_dog, p2, p2_dog, p3, p3_dog) in one column called: image prediction , and Merge the columns(p1_conf, p2_conf, p3_conf) in one column called confidence level.
- Merge all three dataset and remove repetitive columns. Take both the twitter_clean and Json_clean tables and image_clean merge into one table using the join() method on the columns tweet_id.

3- What is Cleaning data ?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled

In the cleaning process, we used the method of defining the problem, writing the required code to solve this problem, and finally testing the validity of the code to verify the results in executing the code.

Step cleaning process

Define : Defining the problem and what is the purpose of this solution to this problem.

Code : Actual code to solve the defined problem.

Test : Test the validity of the executed code and did it show the results expected to be resolved from this code

After checking the existing problems and what are the ways to solve them, and the definition of each problem and the way to solve it, a backup copy is now created of all the files used with their original information without change in order to preserve the original copies of the data

4- Merge Dataset

After making a backup copy of all the used files and doing the cleaning process on the backup files to preserve the original information, the process of merging all three files is now being done to reach related information and have a goal, by knowing the tweets that have a lot of likes and knowing what images are used for these tweets This is through a photo file and a data file for these tweets

We used the file name “twitter_archive_master-2.csv” to Merage All Dataset into one file , and also take a copy form the Marge file

At the end of the project, we discussed the analyzing and visualizing of this data

- Find out what are the most common animals among the names of these dogs
- Find out the percentage of repeated tweets and favorite tweets
- Get the most common types of animals by comparing 3 different sides
- Comparing a type of dog compared to other dogs to find out the proportion of this type compared to others