WERATEDOGS®

SAT/SUN
May 1/2
2021
/Easter (Orthodox)

This is George. He doesn't chew socks. He just likes to hold them. They make him happy.

14/10 a self-care king

PHOTO COURTESY TRACY KASIK

Data Wrangling Report

By Ahmed Osama Mohamed

February 2021

As an assignment for the Udacity Data Analyst Professional Track. This report illustrates the main steps involved in Data Wrangling process of Twitter Account: “WeRateDogs”.

Data Gathering:

This is the step for Collecting data. For this project, there were 3 main sources to gather data from

1- The WeRateDogs Twitter archive(twitter-archive-enhanced.csv): This file is given to us, so I Downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv . Then import it to the working environment by using pandas read method (pandas.read_csv()).

2- image-predictions.tsv is the second file and it was hosted on a webpage. So, I downloaded it using the Python Requests Library then imported it into the working environment by using pandas read method (pandas.read_csv()).

3- tweet_json.txt is the third file and it was gathered through Twitter API. I had to apply to twitter developers account at first, then after my application was approved, I Started to gather this dataset using ‘Tweepy’ python library.

Data Assessment:

This is the step for finding Quality and Tidiness issues whether manually or programmatically.

1- I did some visual assessment in Jupyter notebook by showing the whole dataset inside the notebook. Then I did programmatic assessment.
2- Quality issues was addressed first, then the Tidiness ones.
3- For the Quality issues I have addressed some points like:
    - Missing values
    - Wrong datatypes
    - Inaccurate values
    - Wrong Extracted values
    - Lower Case Names
    - Records with replies, re-tweets and no-image. Which isn’t in our criteria.
4- For the tidiness issues, I have addressed some points like:
    - Not every variable is represented in one column issue.
    - Different tables that should be in one table.

Data Cleaning:

This step is for Cleaning the data from the issues we have addressed above. And I have done it through three main steps:

- Define: I defined how would I deal with the issue
- Code: writing the code
- Test: test the output of this code, if it would match the needed or not.

So, I started the cleaning step with copying the 3 datasets. Then I did some mixed steps. I didn't follow an order in cleaning the data.

- Dropping the unneeded records
- Solving missing values issues
- Correcting wrong data types
- Upper casing the lower-case issues
- Correctly extracting the wrong extracted values
- Correcting the in-accurate values issue
- Solving the tidiness issues and converting every variable into one column
- Merging the tables

Data Storing:

This step is for storing the data after finishing the cleaning step. It's done by using the pandas method

"To_csv () " and saving the final data set into a csv file called " twitter_archive_master.csv "