Wrangle Report

By Mohamed Rashwan

FEB 2021

In this short report, We will have a walk through the wrangling process for the project (data-wrangling) of twitter account "WeRateDogs", as part of Udacity's Professional Data Analysis Nanodegree.

1. Data Gathering:

We will start our journey by gathering the required data using three different data sources that almost cover the most common ways for gathering data. These data sources are:

a. **"Importing Enhanced Twitter Archive"**: This file was downloaded manually and then imported into our working environment using Pandas library.
b. **"Gather data via Requests library":** We will gather some additional data using requests library. We send a request to download "image-predictions.tsv" file using provided "url".This way will help us to apply the concept of reproducibility of code.
c. **Gather Twitter Data Using Twitter API:** For this Particular Step, I had some issues with getting twitter developer account so I decided to use the given "tweet_json.txt" file.

2. Data Assessment:

In this stage, we are going to explore the data and see what's wrong with it so we can fix these issues later before start analyzing the data.

There are two assessment parameters we can look for which are:

1. **Quality**: Issues related to content (Low quality data is also known as dirty data).
2. **Tidiness**: Issues related to structure of data that affect the ease of analysis (Untidy data is also known as messy data).

In order to explore the data we will use two types of assessment:

1. **Visual assessment**: scrolling through the data.
2. **Programmatic assessment**: using code to view specific portions and summaries of the data (pandas' head, tail, and info methods, for example).

3. Data Cleaning:

The final step in the wrangling process is cleaning the data for quality and tidiness issues.

First, we need to take a copies for our datasets to protect the original data frames that we get from Gathering step.

Second, we are going to take every single issue in Assessment summary and solve it through applying [Define, Code, Test] strategy.

Most of the cleaning was performed programmatically, such as defining functions, developing regular expression to capture right records or using pandas built-in functions (merge, melt, extract, etc.)
Also, some manual cleaning was performed to correct ratings error values.

4. Conclusion:

The final product from our wrangling process was a master database with our date cleaned and ready to use in the analysis and visualization.

Also, we exported SQL database file. *"master_df.db".*