Wrangle Report

WeRateDogs Data Wrangling Project

1.1 Introduction

The relevance of data in our world today continues to grow, it is arguably everywhere and it plays an important role in our daily activities. In its simplest definition, data is the resource that every data professional (scientist, analyst, engineers, architect) works with and there is much data in the world today.

"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days."
~ Eric Schmidt, Executive Chairman at Google

One crucial thing to understand is that **data is not information**. Before a data can be considered as an information which is communicated and easily understood by every individual, it has to pass through an interesting process, mostly known as "Data Wrangling"

1.2 Data Wrangling

Data Wrangling is the process of gathering data, assessing its quality as well as structure and cleaning it before further activities such as, analysis, visualization or building predictive models using ML can be done with it. Real world data rarely comes clean and in actual fact, Gartner research estimates that poor quality data costs organizations an average of **$12.9M** every year.

Wrangling is a core skill that every data scientist must have, as it is used to varying degree on every data set, they work with.

1.3. WeRateDogs Project

in this project, the tweet archive of Twitter user @dog_rates, also known as WeRateDogs is being wrangled. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The ratings almost have a denominator of 10. The numerators caries values that are almost greater than 10 ($\frac{11}{10}, \frac{12}{10}, \frac{13}{10}$) etc. This is because *"They're good dogs Brent."*

1.4. Wrangling Process

Data Gathering
The datasets used for this project was gathered from three (3) different sources which involves

- A CSV file at hand
  This CSV file ("twitter-archive-enhanced.csv") was read into the DataFrame using the Pandas function.
- A TSV file downloaded programmatically

This file ("image-predictions.tsv") was downloaded from the internet programmatically using the "Requests" library as it is best for scalability and reproducibility. Afterwards, the file was read into a DataFrame using the Pandas function.

- Querying the Twitter API
  This process involves specifying the required tweet IDs for the project and querying Twitter's APIs for each tweet IDs JSON data.

Data Assessing

This wrangling process involves two (2) steps.

- Visual Assessments
  Each file was visually assessed and it means systematically looking through each table of data in the *Jupyter Notebook* using Pandas and scrolling through the data, looking for interesting and relevant issues such as (completeness, validity, accuracy and consistency)
- Programmatic Assessments
  All files went through this process and it involves using functions and methods to reveal information on issues about each dataset's quality and tidiness.

Some of the issues identified in the data assessing process of each datasets include the following:

Quality

1. Unusual characters in the *name* column.
2. Erroneous datatypes.
3. Non-descriptive column names.
4. Columns with non-useful information.
5. Problems of outliers and inaccurate values.

Tidiness

1. A new variable to form a column.
2. Merging the 3 datasets to for form a table.

Data Cleaning

This process involves cleaning the issues that was revealed during the assessment of the datasets. It follows three (3) steps

- Define the issue.
- Writing the cleaning code.
- Test the code for successful execution.

1.5. Conclusion.

The data wrangling process was tedious but effective enough to help in cleaning our dirty and messy data. This provides the opportunity to go ahead with successfully analyzing and visualizing our data in order to communicate our findings.