Wrangle Report

Introduction

The project included a deep analysis of the data related to the twitter account @WeRateDogs which is quite popular when it comes to dog lovers. Gathering the data and identifying the sources of the data was the most crucial step specifically for this data analysis. Getting a hands on experience with the Data Wrangling process from various sources was the biggest takeaway from the project.

Project details

The tasks of this project are as follows:

- Gathering data
- Assessing data
- Cleaning data

Gathering Data

Gathering the data for this project was quite a learning experience. Gathering the data from many sources and gradually creating a uniform dataset that is relatable is quite remarkable for this project. The 3 important sources of data were :

1] In Hand File : The In Hand File provided formed the basis of the whole analysis process. Although the File has some missing columns which were added at the later stage in this process, The In Hand File gave an idea of the project and helped in the further of data.

2] Programmatically Downloading A File : This was one of the most significant takeaway which can be very handy when we need to download some files. Programmatically downloading a file and then storing it was quite an experience. Making the use of requests package and then downloading the file programmatically from a server seemed like magic.

3] API Interaction : This data source helped me in filling up the missing columns as mentioned in the first data source. This source had an API interaction (twitter API) using the tweepy library to download the required data.

Assessing Data

Assessment of the obtained data was quite revealing step and It helped me to understand the data better and finding the problems that were to be rectified by the cleaning process.

- The tools that I used to assess the data were the Jupyter Notebool that I was working on and MS Excel.
- There are many things that one can misinterpret when assessing the data in Jupyter notebook but opening it in MS Excel and getting a glance over it can really help in getting the bigger picture of the data.

- Assessment of data includes two phases :
  1] Visual Assessment : As mentioned earlier this step can be carried out in the Jupyter Notebook and MS Excel.

  2] Programmatic Assessment : Getting information about the data, getting the distribution of this data points, the distinct values in a particular columns are the most significant insights that form a valuable assessment phase.

- Identifying the columns that are to be cleaned and noting them down in the Assessment Summary is a traditional process that one can follow.

Cleaning Data

Cleaning the data is something that one implements the actions specified in the Assessment Summary. Cleaning the data may include various actions like :

- Filling up the missing values.
- Dropping certain records.
- Dropping certain columns.
- Correcting specific records.
- Merge or join two dataframes.
- Dealing with the null values.
- Dealing with certain typos.

There are 2 types of issues :

- Tidiness Issues : There were 4 tidiness issues that I encountered while assessing the data. The issues are generally related to the structure of the dataset that we are analyzing. Generally, It is suggested that the tidiness issues must be solved first and then one must solve the quality issues.
- Quality Issues : There were many quality issues that I encountered. These issues are generally related to the data i.e the content of the data. All the issues related to the data are quality issues.

Conclusion

- Data Wrangling is a very important skill that one need to be good at in order to be a data analyst.
- A careless Data Wrangling can lead to a false insight which ultimately would reflect in a wrong decision for an organization.
- Data Wrangling for this particular project was fun and the Gathering process required a lot of new skills.
- Data Wrangling must be carried out in a systematic manner and each step must be clearly defined and executed.
- Cleaning the data makes the further analysis phase simpler and ready to extract insights.