Report on Data Wrangling: WeRateDog Twitter-Datasets

Data wrangling is the process that involves data cleaning, data remediation, or data munging which refers to a variety of processes designed to transform raw data into more readily used formats. Data can be represented in various formats and the role of data wrangling is also merging, grouping, concatenating and getting ready for analyzing or building models on the given datasets. Python has built-in features to apply these wrangling methods to various data sets to achieve the analytical goal. In this data wrangling process basically, data gathering, data Assessing and data cleaning are performed.

1.1 Data Gathering

Data gathering is one of the activities during data wrangling process. In this project three different datasets are collected. One of the three datasets was given and the two datasets were gathered programmatically using web scrapping libraries(requests) and API (tweepy) for the remaining twitter dataset. Additionally, the data become stored and represented in suitable format which contains all the necessary information and would able to read into data frame. So, the data formats are represented in comma separated value and tab separated value which are suitable to manipulate for farther analysis. Furthermore, the datasets are get read to the Jupyter notebook editor and visualized in pandas data frame. The head and info methods are used to observe the datasets.

1.2 Data Assessment

The objective data assessment is to identify incorrect or dirty data and untidy or messy data to implement corrective action. During the assessment process on the WeRateDogs twitter data, numerous amounts of problems are identified. The assessment was conducted visually and programmatically with help of pandas library and exploration data analysis tools (EDA). As a result of the assessment, the following quality and tidiness issues are identified.

- Null variable or columns
    - NaN values
- Insignificant columns, columns which contains similar value
- Incorrect datatypes such as
    - The timestamp needs to be datetime instead of string
    - The tweet id needs to be string instead of numeric
- Messy or unstructured data that needs collapsing and merging

    - The columns that describe about dog stages
    - Mering and concatenating the datasets based on their id
- Dirty and duplicated records
    - Extracting the text and sources
    - Removing special characters
- Incorrect and invalid data such as,
    - Names with ‘a’
    - Rating numerators with the fraction value instead of the whole decimal value
- Inconsistency issues
    - Some starts with capital and the other with small letter
    - Column name divergence
    - Values some of them None the other NaN
- Some non-descriptive columns with their name
- The retweet rows duplicate duplication issue are identified

Lastly, the result of the assessment is summarized and documented.

1.3 Data Cleaning

Data cleaning is the process of correcting inaccurate records from the datasets and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or deleting the data. So, based the result of the assessment, more than quality related issues and two tidiness issue are get cleaned and corrected. During the cleaning phase three steps are employed. These are:

- **Define** phase: in this phase the problem is described and the solution to correct it is specified ideally.
- **Code** phase: in this phase, it specifies the methods, libraries to be used and writes the actual code to fix the problem.
- **Test**: in this phase, testing is performed to assure the problem is get fixed. The pandas methods are helpful in this process.

Lastly, during the cleaning process the following activities are conducted.

- Inaccurate records are replaced
- Insignificant columns to analysis and null records are removed
- Duplicate records are managed

- Divergent column name gets managed
- Non descriptive column names are replaced with appropriate names
- Datatype issues get corrected
- Tidiness issues get resolved
- Invalid issues are managed
- Standardizations has been made regarding capitalization, values and names.
- The retweet rows which are the duplicates of the actual tweets and may cause skewness on analysis are removed

Finally, the preprocessed datasets are saved and stored to a master file for further analysis purpose.