Wrangle_report

December 14, 2018

1 Wrangle_report

1.0.1 Olive MGK

1.0.2 Procedure:

As in any problem-solving scenario, 'understanding' what is expected is the initial challenge. Wikipedia and the WeRateDogs site provided a good overview of the nature of business. 'WeRateDogs' – is a great site – for laughs when wrangling efforts hit a low. Anyhow, as I noted – understanding the project requirements is crucial. This involved setting up my environment. I imported the necessary libraries like the pandas, Numpy, requests, tweepy, matplotlib and the json library. And ensured that my Jupyter workspace was in order.

1.0.3 Gathering data:

I gathered three pieces of data from different sources -ie CSV, website and from the twitter api. I accessed the api using tweepy python library and the tweets data gathered was written to a json text file.

1.0.4 Assess data:

The data was assessed both visually and programmatically for quality issues like typos, dtype error, missing details and tidiness issues like over spread data. Pandas methods like head(), shape(), info() etc were used to provide an indepth view of the dataset. This gives detailed summary about the size, dtypes , all variables in the data and the storage size of the dataset. describe() gives a good summary of the quantitative aspects of the dataset. 8 quality problems and 2 tidy were identified and documented to be used in the next stage of the dwrangling process. Such included missing data, NaNs, wrong datatypes, duplicates, empty or useless columns, incomplete data and yes for tidiness, some columns needed to be gathered which some needed melting like the timestamp.

1.0.5 Cleaning:

Following assessment, the issues identified in the assessment stage were cleaned programatically to create a clean and tidy dataset fit for analysis and visualisation. Cleaning is a never ending process of defining, cleaning with code and testing that's why its iterative. This process also continued on into the analysis stage. After cleaning the data was then stored in a file ready for the next process of analysis and visualisation to derive insights and predictions