Wrangle_report

December 18, 2018

1 Wrangle_report

1.0.1 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 taking a break from wrangling. Anyhow, as I noted – understanding the project requirements is crucial. This involved setting up the working environment. Importing the necessary libraries like the pandas, Numpy, requests, tweepy, matplotlib and the json library. And ensuring that the Jupyter workspace is correctly setup.

1.0.2 Gathering data:

Three pieces of data from different sources were collected: The sources were data from a CSV, a website and from the twitter api. Tweepy python library was used to access the twitter api and the tweets data gathered was written to a json text file. This was a good exercise indeed.

1.0.3 Assess data:

Data was assessed both visually and programmatically for quality and tidiness issues like typos, dtype errors, missing rows and data; and tidiness issues like over spread data or timestamps with more than one variable in a column. Pandas methods like head(), shape(), info() etc were used to provide an indepth view of the dataset. About 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.

1.0.4 Cleaning:

Following assessment, the issues identified in the assessment stage were cleaned programatically with code 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 said to be iterative!. This process goes on into the analysis stage. After cleaning the data was then stored and then analysed and visualisation. The insights are documented in the act_report.