Data Wrangling

To get the data into a good state before analysing it there are some steps we need to go through . The following process is a basic outline of what needs to happen before data is ready for modelling and analysis.

- **Data Discovery**: This is an all-encompassing term that describes understanding what your data is all about. In this first step, you get familiar with your data and ensure it is capable of being used to help work on your problem.
- **Data Structuring**: When you collect raw data, it initially is in all shapes and sizes, and has no definite structure. Such data needs to be restructured to suit the analytical model that your enterprise plans to deploy
- **Data Cleaning**: Raw data comes with some errors that need to be fixed before data is passed on to the next stage. Cleaning involves the tackling of outliers, making corrections, or deleting bad data completely
- **Data Enriching**: By this stage, you have kind of become familiar with the data in hand. Now is the time to ask yourself this question – do you need to embellish the raw data? Do you want to augment it with other data?
- **Data Validating**: This activity surfaces data quality issues, and they have to be addressed with the necessary transformations. The rules of validation rules require repetitive programming steps to check the authenticity and the quality of your data
- **Data Publishing**: Once all the above steps are completed, the final output of the wrangling can be published as part of a data pipeline and is ready for analysis.

Data wrangling is a core iterative process that throws up the cleanest, most useful data possible before you start your actual analysis. It is THE most important stage in building credible models as everything else depends on the quality of the data.

Examples of the steps

Data Discovery

Think about the Titanic data set – what can be said about it – what cannot be said; it gives us information on survivors / deaths / gender / ticket class etc but not on weight of passengers (does it influence survival chance?). Also it gives us no physical state information (what was the average drop height of lifeboats into the sea for example).

Data Structuring

If you are collecting raw data you will need to put the data into a standardized format, you might need to convert integers to text or text to numerical data. For example if you

are scraping twitter feeds you might need to parse dates and put them in a standardized YYYY-mm-dd format – adjust American dates etc. Structuring the data so it can be used in a 2D Pandas dataframe for example with columns and a row for each column.

Data Cleaning

This is what most people find is the core activity in data preparation – as it involves ensuring the data is consistent and credible. Activities include IMPUTING values – if there are missing values then either replacing them using a standard method such as taking an average of previous and subsequent data points or removing the data completely etc. If there are missing values then decisions need to be made and recorded as part of the documentation so that the changes are transparent.

Data Enriching:

This step should be taken carefully as it involves transforming original data in new ways or adding data that wasn't part of the original data set. An example could be similar to doing a database join to connect data from 2 separate sources to create a new one (for example name and address information combined with geographic data and some demographic data to see if age and location is correlated with name choices (this is what German health bodies had to do to try and 'guess' the age of people to call in for their covid vaccination!). It could be that like the Danish CPR system there is date of birth information inside another piece of data – so you can embellish your data using further data embedded inside other data items.

Data Validating

You need to make sure all the data is of the data type you are expecting, that it falls within an expected range and that it is consistent. This step is concerned with making sure there is high confidence that data is not being missed because it is outside a range where the range is incorrectly set, where data has accidentally been stored as an integer rather than a floating point number etc. It is a double check that your data structuring and data cleaning have all been done properly (for example, if you are supposed to have unique values in a column that they are indeed unique).

Data Publishing

This is the final step and is a strong candidate for being done inside a version control tool like git. Often for big data projects such as in marketing this will sit inside a distributed datastore (can be relational (Postgresql) , key-value (Google BigTable / HBase) or other types of NoSQL database. In Machine learning then tools like Cassandra and Couchbase or Postgresql are popular.