QUALITY AND CLEANING ASSIGNMENT

OJAAS HAMPIHOLI.

OpenRefine is great data manipulation tool which can be used when we have a large data set with multiple features which hold multiple types of categorical as well as numeric columns. This tool can be used very easily when we want to change the features by a bit while still maintaining the integrity of the data set as well as keeping a hevy speed and efficiency of data cleaning. OpenRefine works better than spreadsheets in terms of accuracy of data manipulation and cleaning as well as the speed of the same when we work both of those on very large data sets. This tool will come in handy when the speed as well as the security of data is important, as the tool sets up a local server on any computer that it is installed in, so that the data never leaves local computer when it is being processed.

A tidy dataset is a dataset where each column is a feature and each row is an observation and each table is a whole dataset in itself, such that there can be a standard procedure to extract the required feature from one or more columns as per the requirements. This dataset has various features that describe the ships capacity, its jorney and current status of the fleet. This dataset is not a tidy dataset, because it has a lot of features that are named very similar to each each other and fall in the same group, but fall under different categories due to the wrong spelling, wrong capital letters as well as lack/excess of spaces. This dataset needs a lot of refinement and manipulation for it to be made worthy of data wrangling and visualisation in order to make sense of the raw data. The most basic changes would be to convert all categorical columns to lower case and to remove trailing/leading spaces and clustering the various categorical data features to get very few features. This cleans the data to a very high extent such that we can easily visualize the data in order to look for conclusions.

The MISC INFO column contains details about whether the ship completed the trip and was reused, or faced a breakdown and reused, faced a breakdown and was scraped or whther the ship sank on the journey. This data can be used to study the area where maximum accidents happen, as well as the locations with the least accidents, what most of the ship wrecks have in common and the most common fate met by most of the shipwreck as well as the lives lost and the investments lost (ship value and cargo).