Data Wrangling and Cleaning

The dataset, which includes missing and null values, was downloaded from data.medicare.gov. Several components of data wrangling were utilized to make the dataset ready to be used for analysis and prediction.

Step-1

The dataset was originally in the format of CSV and loaded to the python notebook.

Step-2

The dataset contains 6810 observations with 98 data fields with some missing/null values. Many of the names of the data fields are too long and some were shortened for convenience.

Step-3

The dataset was subset into fewer data fields/features depending on scalability and using feature selection tools in scikit-learn.

Step-4

Some of the features' name is too long to handle. Therefore, they are shortened to convenient names.

Step-5

The columns were regrouped to line up the predictor variables together and the mortality rate (the response variable) was moved to the last column.

Step-6

Some of the columns in dataset contain categorical variables, which are known to cover lots of useful information. In order not to miss valuable information, the categorical inputs of some data fields were converted to numerical nature.

Step-7

There are some null values in the response variable, Mortality rate, and the null/missing values were dropped and removed. Also, some of the features contain missing/null values, which will be handled at this stage or before model fitting. If the model classifiers could able to handle the null values, there is no need to clean it at this stage.

Step-8

Data fields, which contain values with % sign, were cleaned and converted into float.

After the dataset was passed through the above data wrangling steps, the dataset was left with 6140 observations and 50 data fields. It is now ready for the next step, exploratory data analysis.