Data exploration is the first step of performing any meaningful data analysis.

The data analysts can use both visual exploration tools and descriptive statistics to understand what are the characteristics of the data. Characteristics that are interested include:

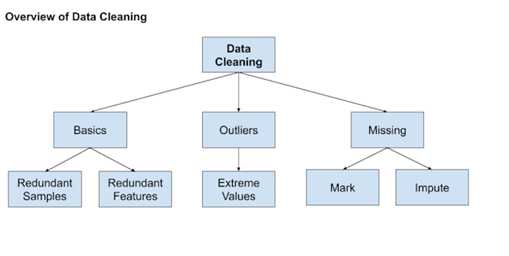
1. **the data size (both number of columns and number of rows);**

**Dataset**:

All four files of the dataset

Describe the input variables and output variables

* Cleveland Dataset (shape and head())
* Hungarian Dataset (shape and head())
* Switzerland Dataset (shape and head())
* VA dataset (shape and head())



1. **Data problems such as Basics, Outliers,missingness;**

* **Basics:** Redundant samples( zero rows)

Redundant Features (column with only one variable)

* **Missing values:**

1. **Mark:**

* Missing values( with ?/NA/ blank)---mark them as NAN
* Draw the histogram matrix
* Describe the data( once it is marked NAN then we can calculate)
* **Outliers:** plot boxplot for all input variables

1. **Imputation:**

KNN:(Cl, cl+sw, cl+hu, cl+sw+hu)--print the accuracy with std and k

Iterative: (Cl, cl+sw, cl+hu, cl+sw+hu)--print accuracy and std

**Decision: Final dataset for our project (cl+sw+hu with accuracy 0.857407 and std= 0.024038, k=3)**

**EDA of new dataset:**

* shape(), head(0, describe(),
* Outliers: extreme values

**(3) data distribution such as normalness, skewness, and kurtosis;**

* Histogram matrix (with color coded if possible)-- if skew then we can try transformation---in classification we need transformation on input only, am I right??
* class distribution to check the balance in the output variable

**(4) relationship among variables such as linear correlation, monotone correlation, and association between categorical variables.**

* Scatterplot matrix
* Correlation matrix to confirm the relation as seen in histogram

--------------------------------------------------------------------------------------------------------------------

**Feature selection: backward, forward, with some algorithms**

I think we did it in our lab session too..(also, homework assignment 1)

**Building a model:**

* Logistic regression
* Knn classifier
* SVM
* Generalized additive models
* Random forest
* Boosting