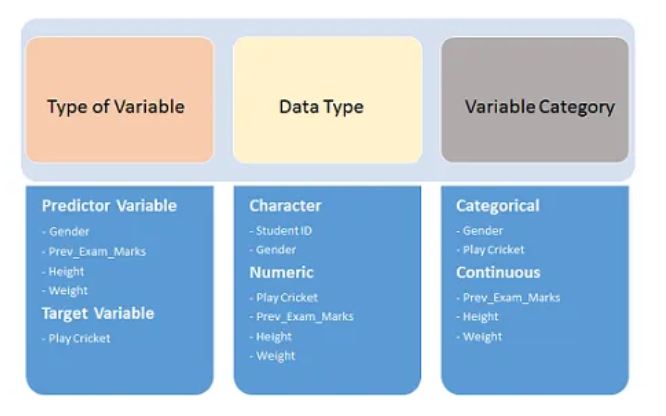
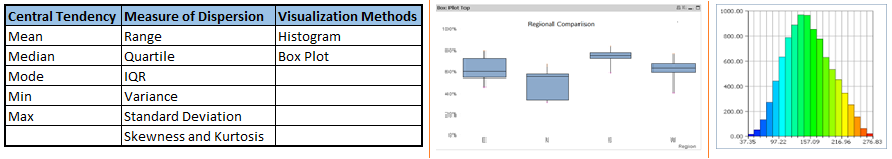
**PREPROCESSING STEPS**

**Variable Identification**: Check the types of data (nominal/ordinal/ratio/interval) and Object, Integers, Float. Separation of numerical and categorical variables.



**Univariate Analysis:**

**For Continuous Variables:** Central tendency, Measures of Dispersion and Visualization Methods - ggplot in R (or) seaborn and matplotlib in Python - **KDE Plot, Pairplot**

**For Categorical Variables**: Use Frequency Table to understand the distribution. Count and Count Percentage against each category. Bar Chart can be used to visualize the data.

**Bivariate Analysis:** Finding the relationship between two variables.The combination can be Categorical & Categorical, Categorical & Continuous and Continuous & Continuous.

Continuous & Continuous - Correlation Matrix

Categorical & Categorical - Two Way Table, Stacked Column Chart and Chi Square Test

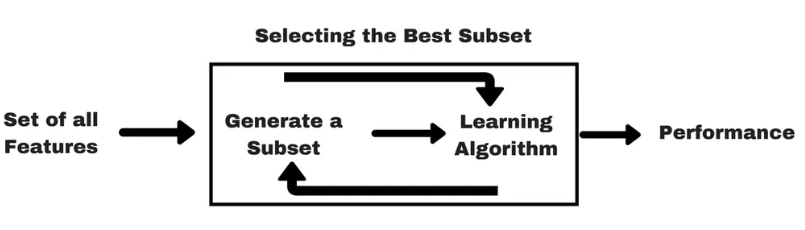
Categorical & Continuous - Z Test/T-Test and ANOVA

Feature Selection and Feature Elimination

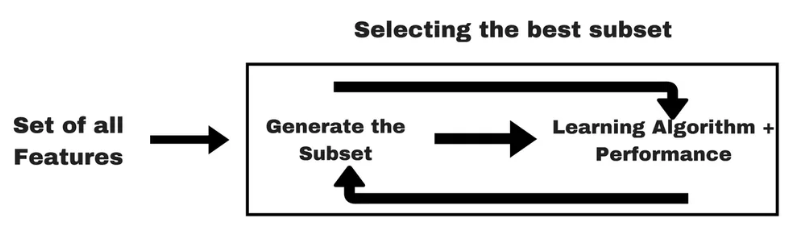
Filter Methods:



* Linear Discriminant Analysis -It is used to find a linear combination of features that characterizes or separates two or more classes (or levels) of a categorical variable.
* Pearson’s Correlation- consider the important variables which has strong relationship with the target.
* ANOVA-Provides the statistical test provides Means of several groups equal or not.
* Chi-Square-Provides the statistical test applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution.

Wrapper Method: RFE

* **Recursive Feature Elimination**: It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

Embedded Method: It combines Filter and Wrapper Methods.The best example is L1 regularization LASSO and L2 regularization RIDGE regression which has inbuilt penalty functions to reduce overfitting.

**Note**: By keeping the most important features, we can improve the accuracy. We can improve the interpretability, reduce the complexity and reduce the training time of the model.

**Missing values in dataset:** The reason for the occurence of these missing values.

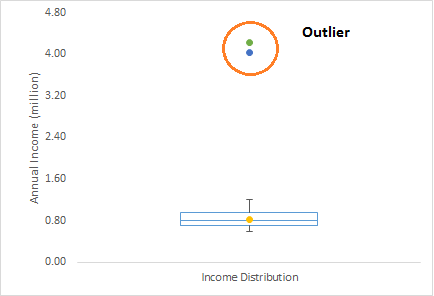
1. Data Extraction
2. Data Collection: Missing completely at random, Missing depending on unobserved predictors. **For Example: In Medical Domain -** if a particular diagnostic causes discomfort, then there is a higher chance of dropping out from the study. This missing value is not at random unless we have included “discomfort” as an input variable for all patients.

Missing that depends on missing value itself: Probability of missing values is directly correlated with missing values itself. **For example**: People with higher or lower income are likely to provide non-response to their earnings.

**Methods to treat missing values:**

1. Deletion: List wise and Pair Wise
2. Mean/Median (Quantitative attributes) /Mode (Qualitative Attribute) Imputation
3. KNN Imputation: The missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function.

**Techniques of Outlier Treatment:**

****

**Outlier -** An observation which appears far away and diverges from overall pattern.

**Outlier are 2 Types**: Univariate (Box Plot) and Multivariate (Scatter Plot to look how the data is distributed in multi-dimensions)

**Removing Outliers**: Deleting Observations, Transforming Variables (Log of Values), Imputing by use of mean/median/mode methods.

**Feature Engineering: Variable Transformation**

* Symmetric Distribution is preferred over Skewed Distribution - For right skewed distribution, we take square / cube root or logarithm of variable and for left skewed, we take square / cube or exponential of variables.

**Common Methods of Variable Transformation** - Log, Square/Cube Root Transformation. In Medical Domain or any other applications, Log Transformation is mostly used.

**Feature Extraction**

* + Principle Component Analysis
  + Factor Analysis
  + Singular Value Decomposition

**Feature Scaling**

Feature scaling is the method to limit the range of variables. It is performed on continuous variables. Plot the distribution of all the continuous variables in the dataset.

**For Example:** we have worked in Credit Card Loan dataset where we have features like applicant income, co-applicant income, credit history, loan amount and loan amount term.

if we try to apply distance based methods such as kNN on these features, feature with the largest range will dominate the outcome results and we’ll obtain less accurate predictions. We can overcome this trouble using feature scaling practically in R/Python.

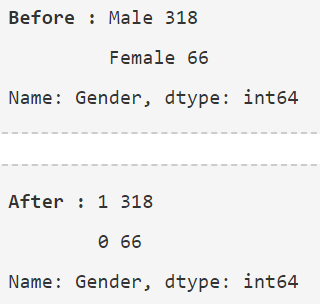
Perform Standard Scaling and Min Max Scaling

**Standard Scaling (Standardization):** It removes the mean and scales the data to unit variance. It cannot be assured there is a balance scaling in presence of outliers.

**Minmax Scaling (Normalization):** Scales the data within the range of 0 to 1. It is sensitive to outliers. In Python, sklearn provides a tool Minmax scalar which will scale down those loan features between 0 and 1.

Perform Label encoding and one hot encoding.

**Label Encoder:** Sklearn provides a very efficient tool for encoding the levels of a categorical features into numeric values. Label Encoder encode labels with value between 0 and 1.

**Note**: Before fit the logistic/KNN/SVM/Tree based methods, encode the features which are discrete ones because sklearn needs only numeric arrays.

One Hot Encoding: It transforms each category feature with n possible values into n binary features.

Important Note: If your categories are ordinal, like low, medium, high. Then it makes sense to use a LabelEncoder with a MinMaxScaler. But if you have non-ordinal categorical values, Mouse, Dog, Cat then it would be better to use a OneHotEncoder.

Prediction: After cleaning the data and performing the exploratory data analysis. Predict the model using test data. Arrive Cross Validation Scores along with other metrics for linear regression. It will test the effectiveness of the trained model.

