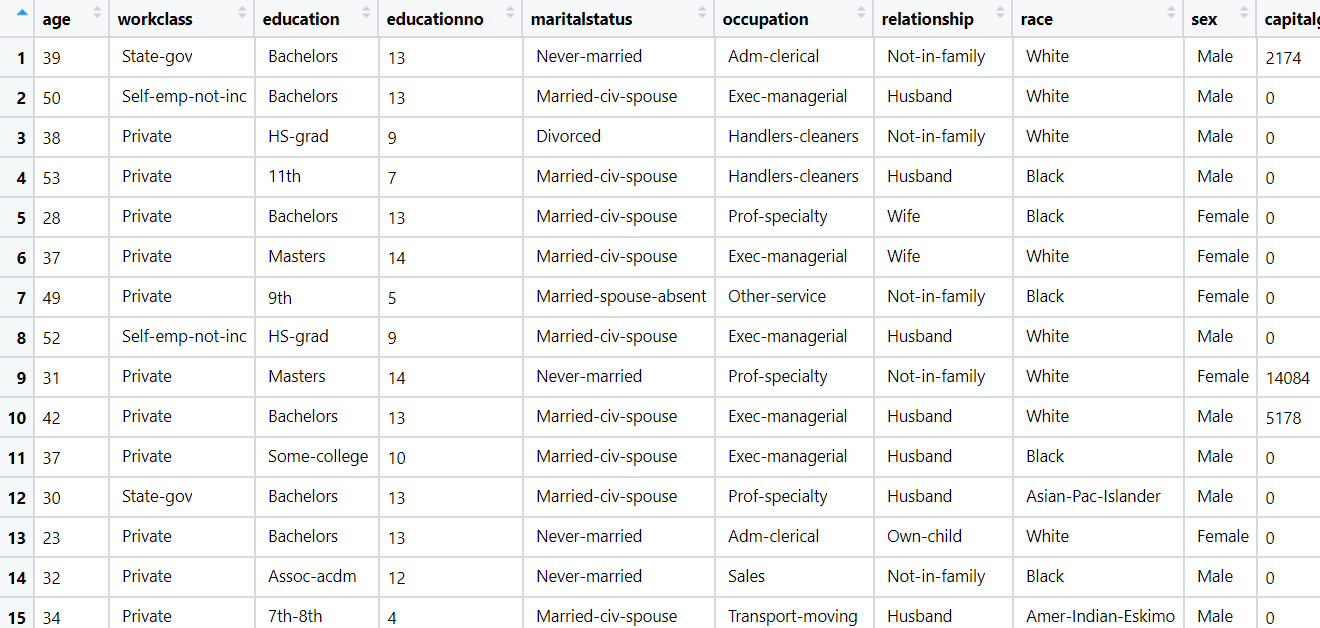
# Topic: Support Vector Machine (SVM)

1) Prepare a classification model using SVM for salary data



**Ans:**

**Analyzing the input and output variables:**

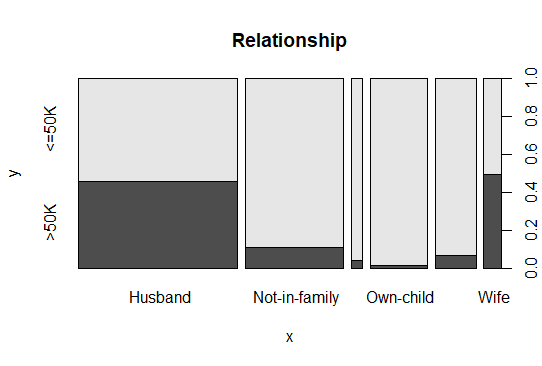
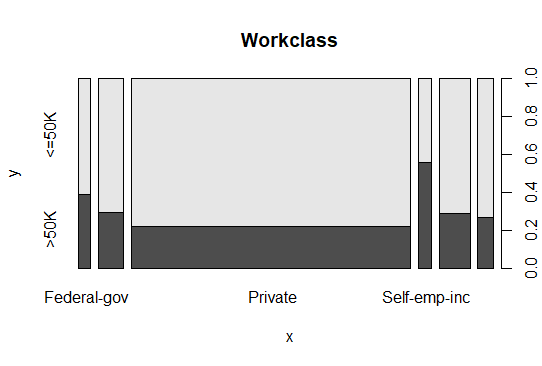
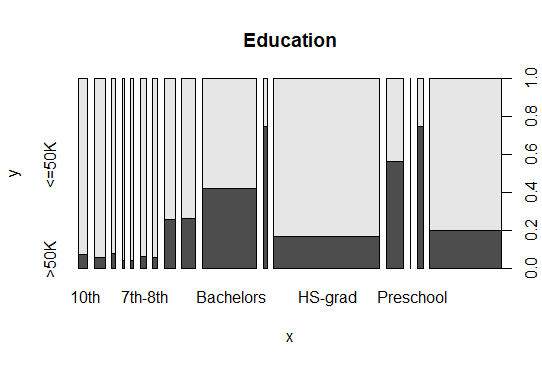
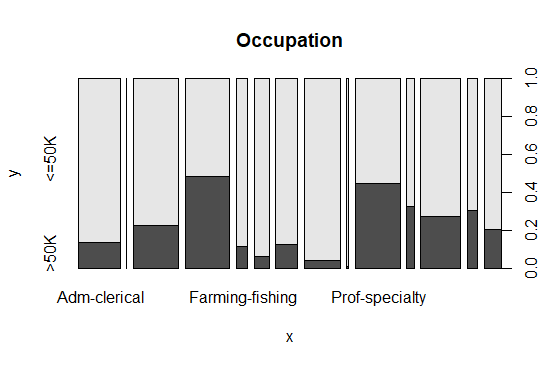
* Output Variable(y) - Salary
* Input Variable (x) - Other Factors

**Importing data and Data Preprocessing:**

* Need to import Dataset using Text(base) converting strings into factors.
* As there are no missing values, So imputation of data is not required.
* Import both dataset Salary train & Salary test using the text base converting dataset from strings to factors.

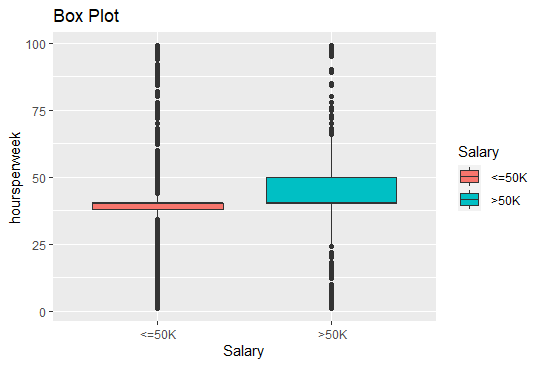
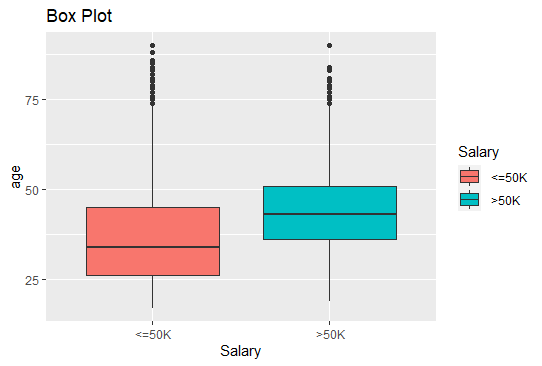
**Exploratory Data Analysis:**

**Plot Representation:**

****

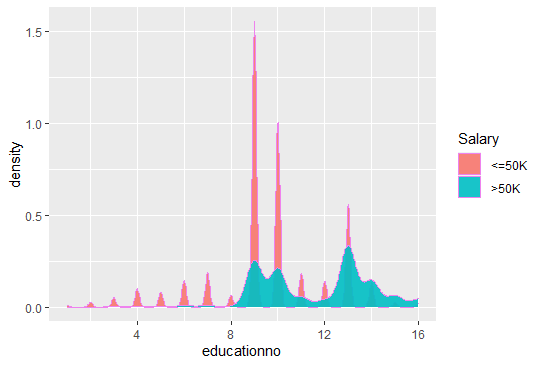
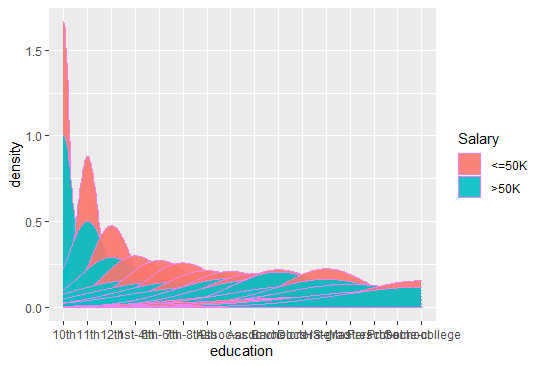
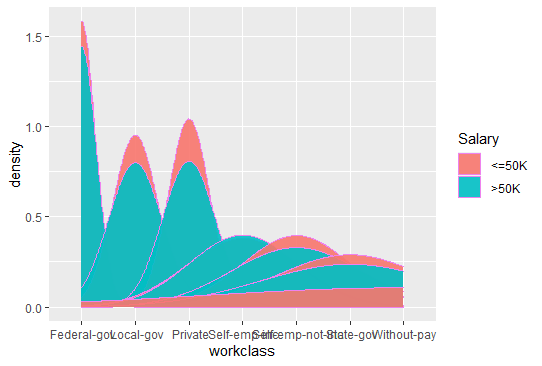
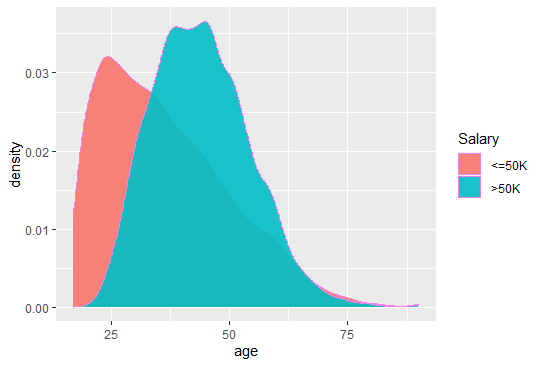
* From the above graphical representation, we can clearly see how the data is spread in each individual variable.

**Box Plot Representation:**

****

* The box plots represent that there exist outliers in the input and output variables.

**Density Plots:**

****

* The density plot observes the distribution of data of a variable.

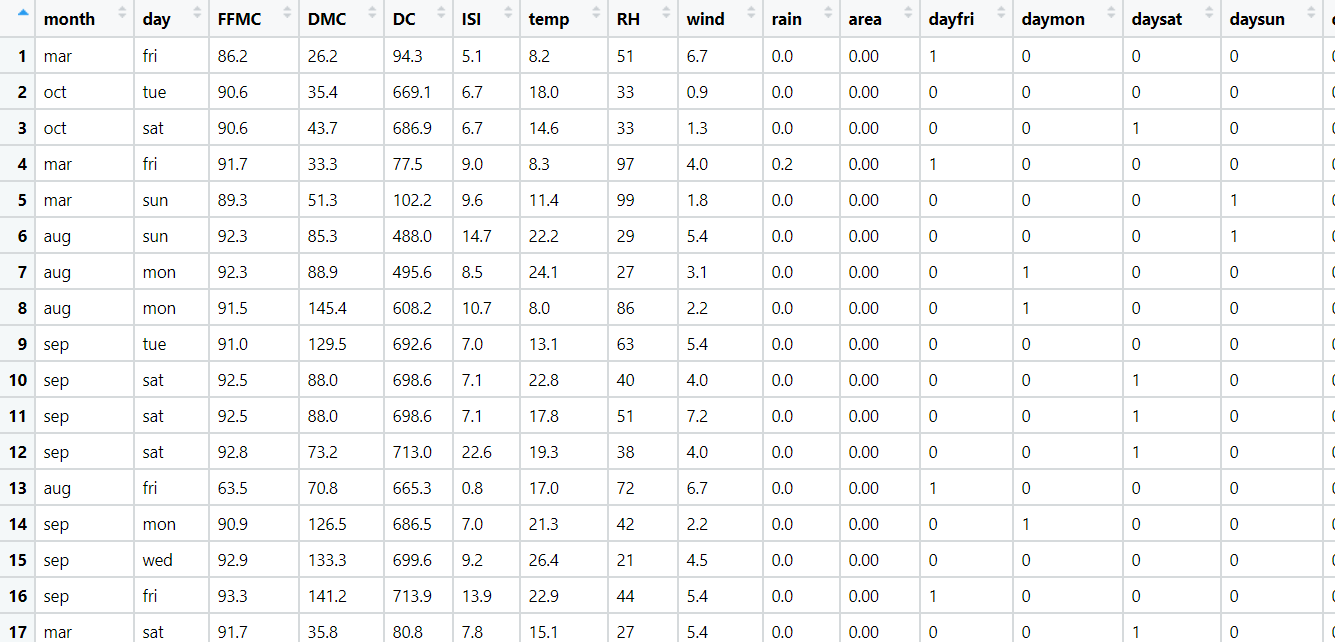
**Model Building:**

* Building a model using a kernel parameter as “**vanilla dot**” and predicting the model on the test data and the **accuracy is 0.85**
* Now used different kernel parameter to “**rbfdot”** to check the improvisation of the model performance and the **accuracy is 0.85**

**Conclusion**:

**So, after comparing above 2 model, both models give the same accuracy and are the best fit models.**

2.) Prepare svm model for classifying the area under fire for forest fires data



**Ans:**

**Analyzing the input and output variables:**

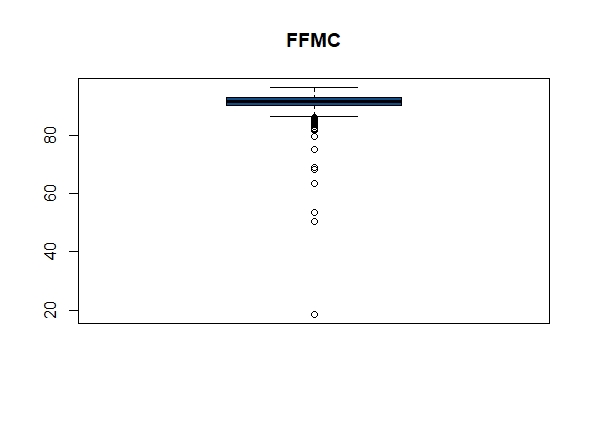
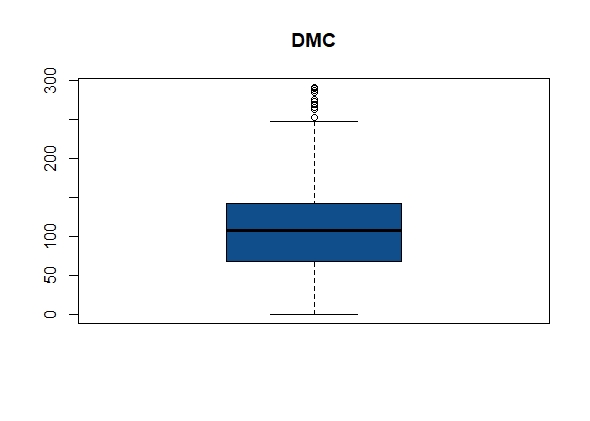
* Input Variables (x) = Other Factors influencing the output variable
* Output Variable(y) = area

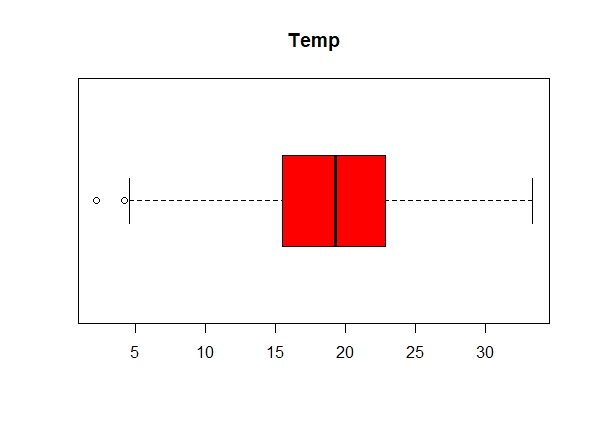
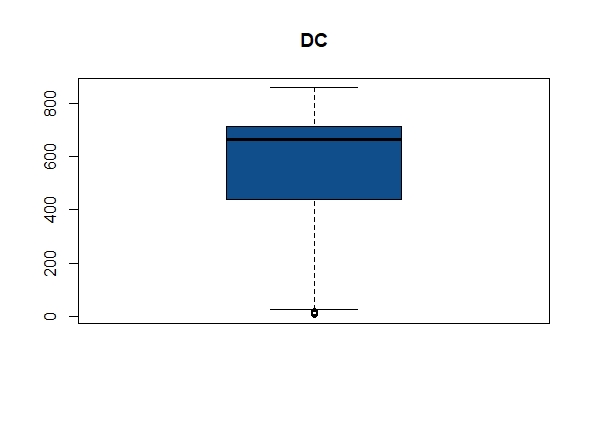
**Data Preprocessing:**

* Columns are rearranged to have the easy access of the variables.
* In R , dummy variable are created automatically when object(x) is created as model using Matrix command which has an inbuild feature of converting dummy variables and we can also factorize the variables. whereas, in python we need to write label encoder () code for creation of dummy variables.
* Removing unnecessary columns and checking for the NA value

**Exploratory Data Analysis:**

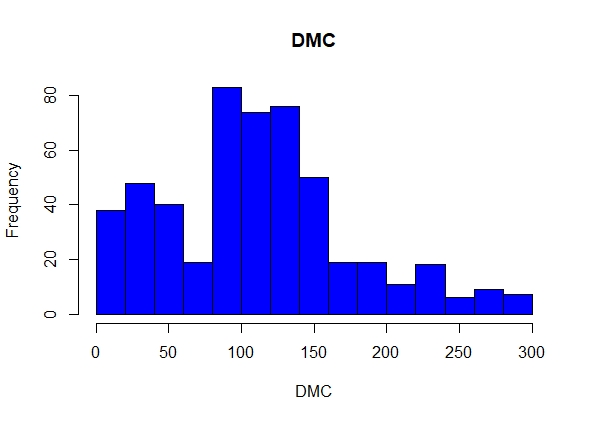
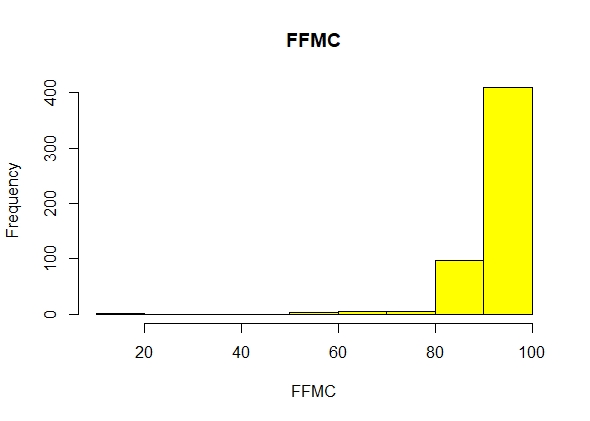
**Boxplot Representation:**

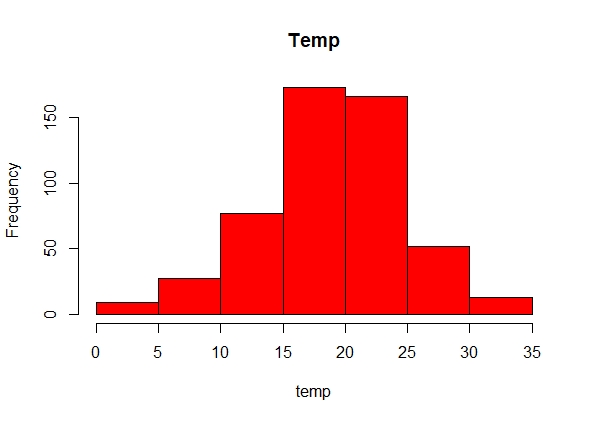
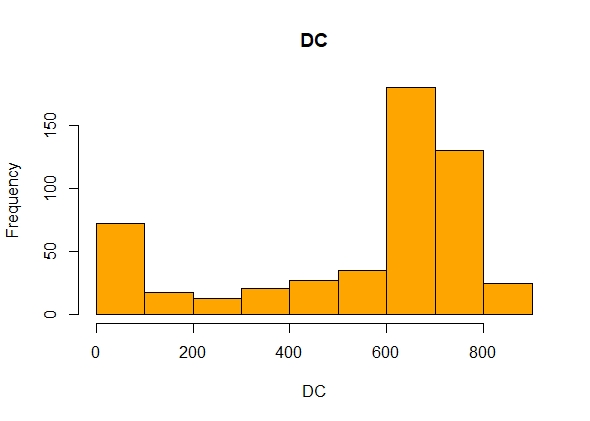
****

****

* The above boxplot represents outlier exist in the individual variables

**Histogram Representation:**

****

****

* From the above graphs data is normally distributed in case of Temp and DMC but incase of FFMG and DC the data is right skewed.

**Normalizing the data:**

* Using the normalization function, the data is normalized and the data will come under same level as shown in below formula.

**normalise <- function(x) {**

**return((x - min(x)) / (max(x) - min(x)))**

**}**

**Converting output variable into categorical:**

* The output variable “**area**” is converting into 2 level categorical variable and the variable data is factorized.

**Model Building:**

1. Now the model is built on train with different kernel methods to check the training error.
2. Building a model using a kernel parameter as “**linear**” and predicting the training error which gives as 0.02
3. Building a model using a kernel parameter as “**polydot**” and predicting the training error which gives as 0.23
4. Building a model using a kernel parameter as “**rbfdot**” and predicting the training error which gives as 0.25
5. Building a model using a kernel parameter as “**sigmoid**” and predicting the training error which gives as 0.25
6. By comparing above 4 models, the model with “**linear”** kernel parameter shows less training error.
7. Now with “**linear**” kernel parameter, the model is tested on test data to find the accuracy.

**The accuracy rate of the test data is 0.98**

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy Ex: Feature Engineering, Hyper Parameter tuning etc.
3. All the codes (executable programs) are running without errors
4. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here