# Project story

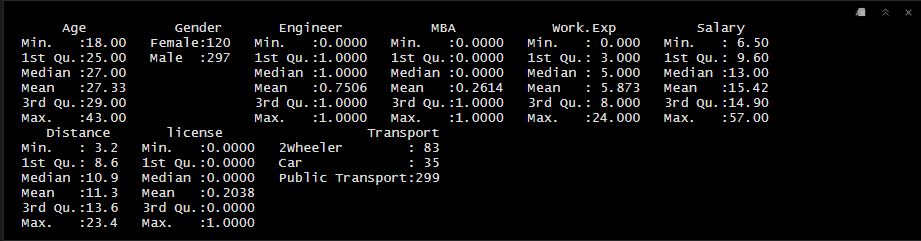
* This project requires you to understand what mode of transport employees prefers to commute to their office.
* The dataset "Cars-dataset" includes employee information about their mode of transport as well as their personal and professional details like age, salary, work exp.
* We need to predict whether or not an employee will use Car as a mode of transport. Also, which variables are a significant predictor behind this decision.

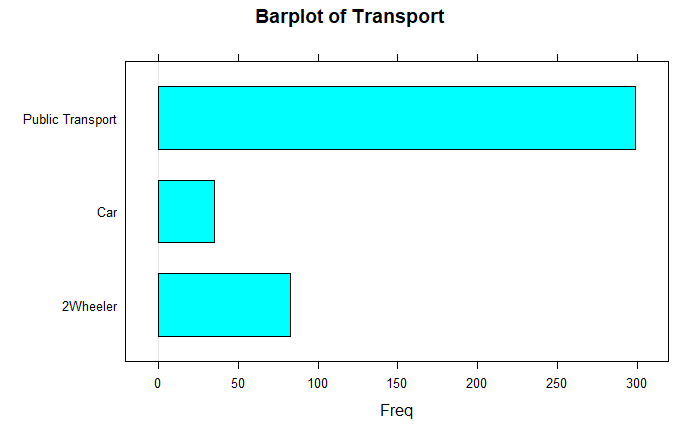
# Data Description

|  |  |
| --- | --- |
| AGE | Age of the employee |
| GENDER | Gender of employee |
| ENGINEER | Does employee have Engineering Degree. 1 indicates employee has engineering degree 0 indicates employee doesn’t |
| MBA | Does employee have MBA Degree. 1 indicates employee has MBA degree 0 indicates employee doesn’t |
| WORK EXP | Work experience in years |
| SALARY | Annual Salary of employee (in thousand) |
| DISTANCE | Distance from office (in KM) |
| LISCENSE | Does employee have license |
| TRANSPORT | Modes of transport chosen by employee |

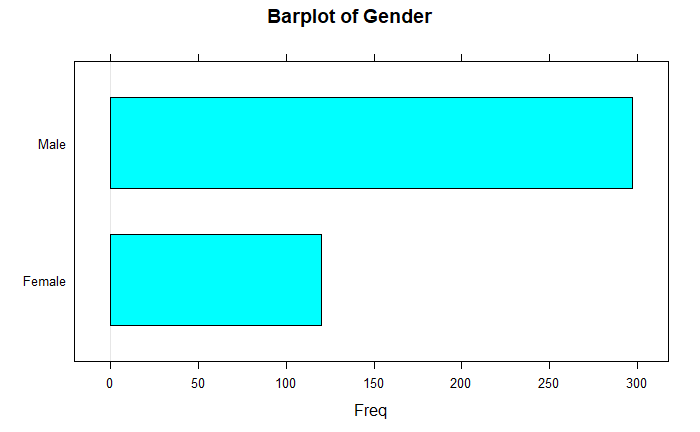
# EDA

* Number of Rows = 479
* Number of Columns = 9
* Grouped into age, salary, work experience, Engineer, Gender, MBA, Distance, License, Transport.



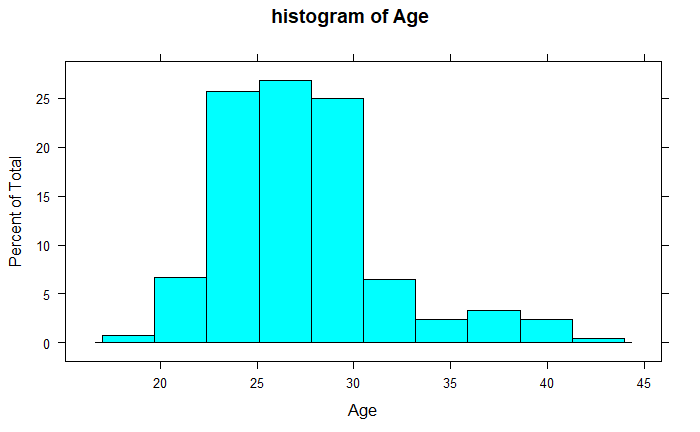


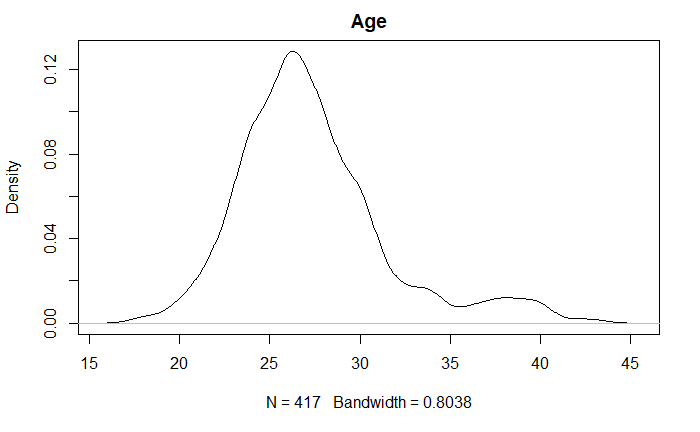
* Bar plot shows that … uses public transport
* Percentage people who uses each transport.
* Transport variable is the target variable, divided into 2wheeler – 83, public transport - 299 and cars – 35.
* After binning the variables 2wheeler and public transport into Public transport, public transport accounted for 91% of employees.

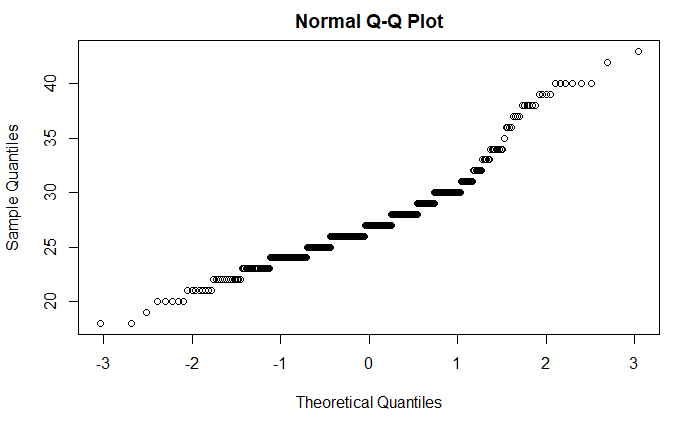


* More of male staff than female
* Female employee represents 29% of the staff population

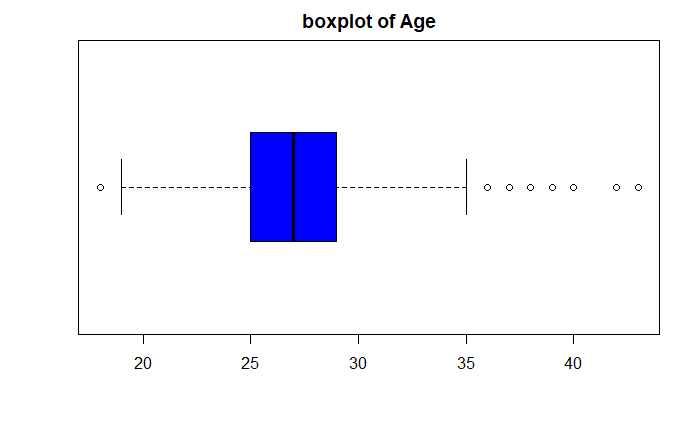
## AGE







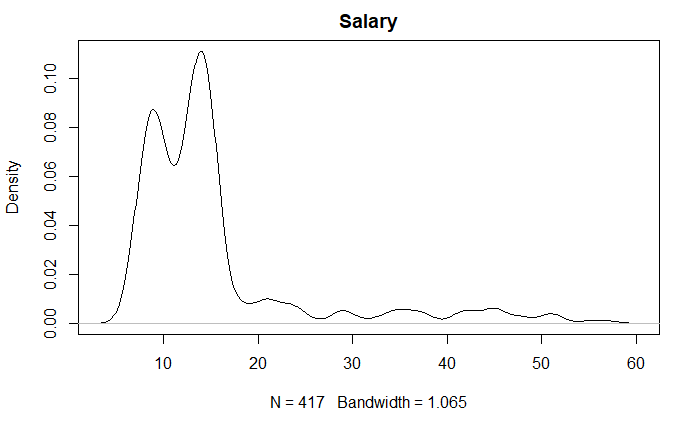
* Histogram of age is skewed to the right
* There’s no significant difference within the percentiles that indicates a big gap in outliers
* The minimum age of employee is 18 and the maximum is 42 years
* Average of employee is 27 years
* 50% of staff is 27 years

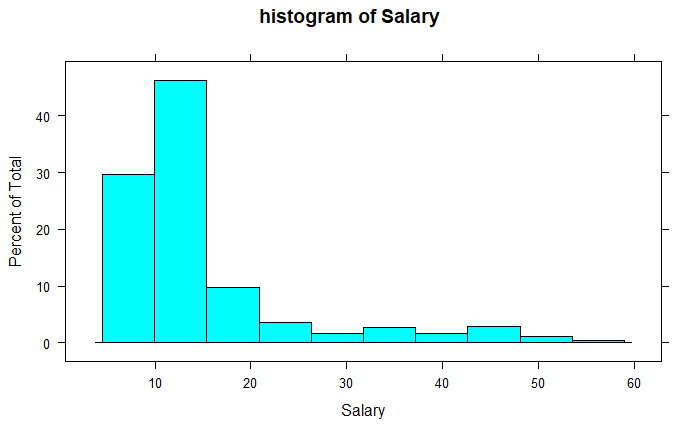


Shows the presence of extreme values in age among employee, pushing extreme values towards the right

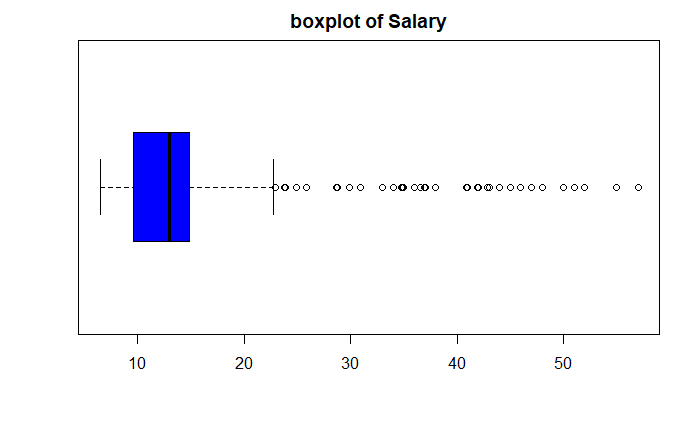
Maximum point falls under 35 years

## Salary



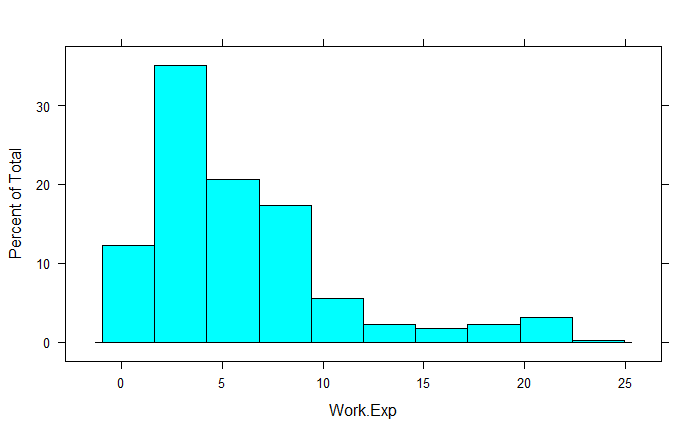


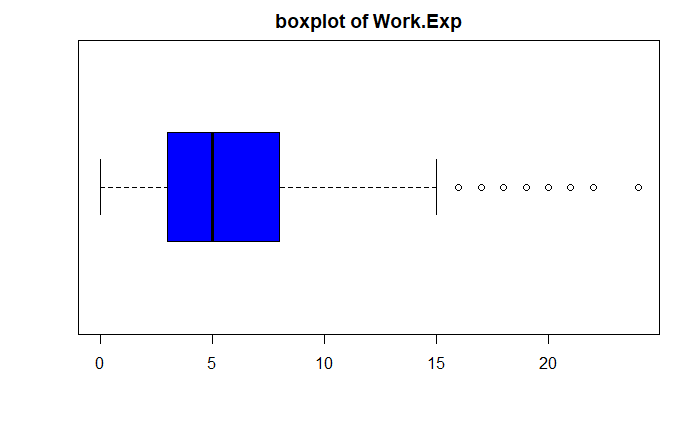
* Salary has extreme values making the histogram skewed to the right
* Extreme values as a result of variables like age and work experience has influenced it
* There is a significant difference in percentiles from the 90th percentile to 99th percentile.
* Average employee salary is $15,420
* 50% of staff earn $13,000
* Staff with extreme values earn around $57,000



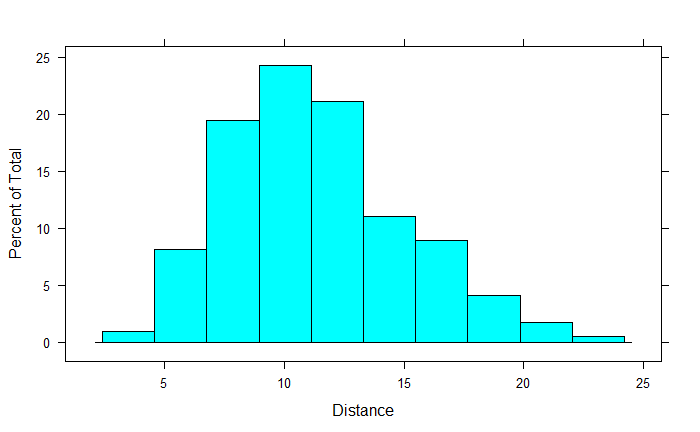
* There’s the presence of outliers pushing the extreme values extremely to the right
* 75% of staff earn $9,000
* And the minimum amount earned is $6000

## Work Exp

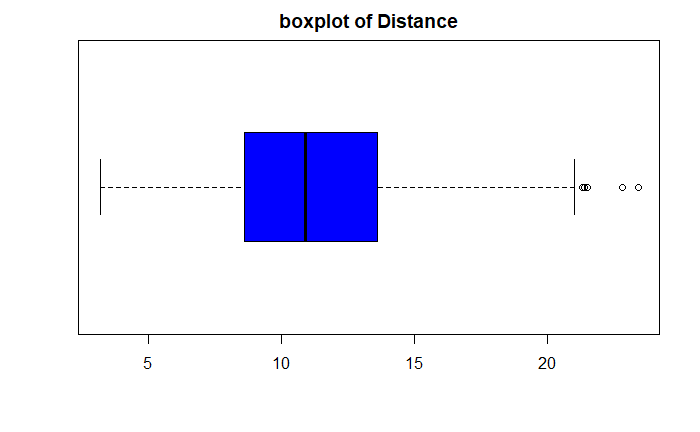




## Distance

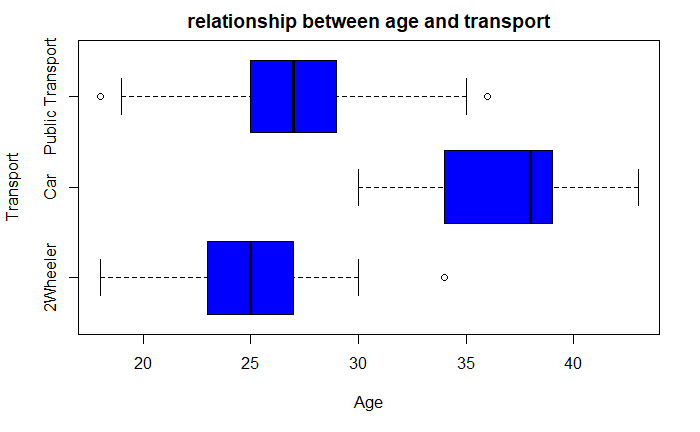


* Histogram of distance is normally distributed
* Average distance of employee is 11km
* 50% of employee live within a 10km range
* Maximum km is 23.4 km

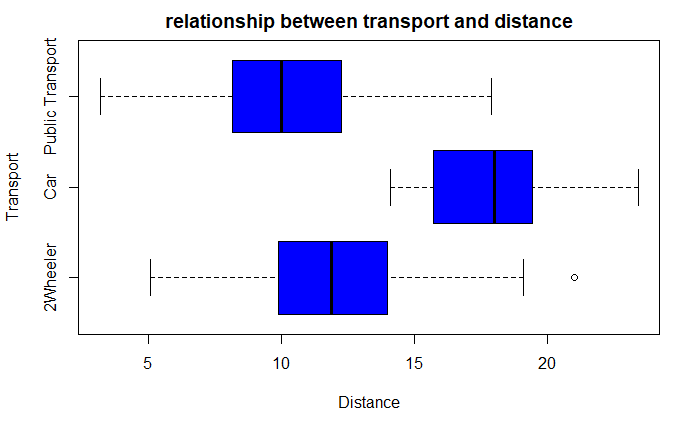


* The outlier is caused by the maximum distance of employee 23.4km
* Only a 25% population live as close as 8.6 km

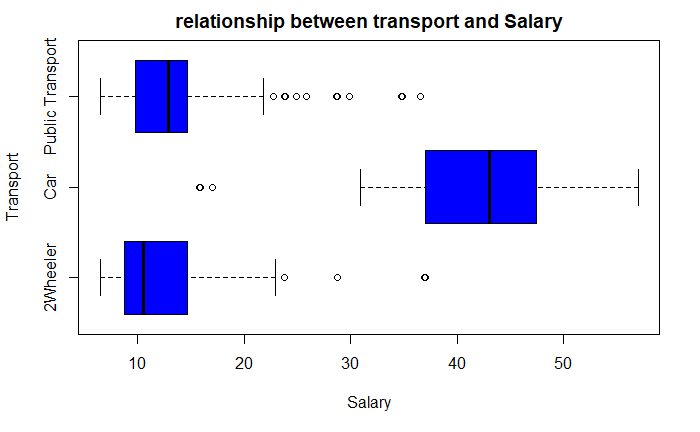
# Bivariate Analysis



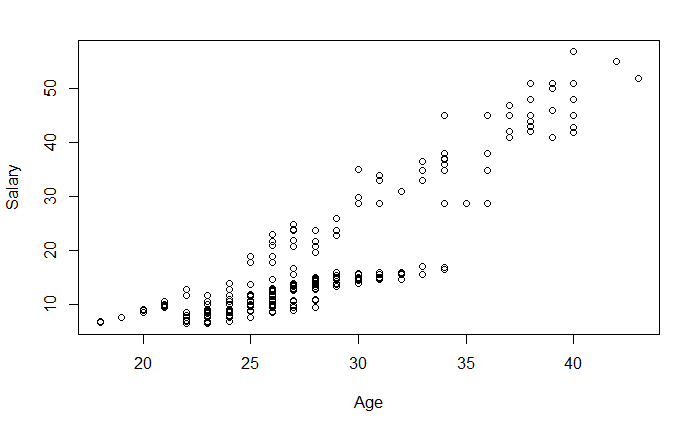
* The younger the employee, the more likely are they to use a 2-wheeler or public transport.
* There’s a linear relationship between the age of employee and transport.
* Older employees are more likely to use public transport or a car.
* The 8.3 % of employee that use car are aged between 30 - 43 years



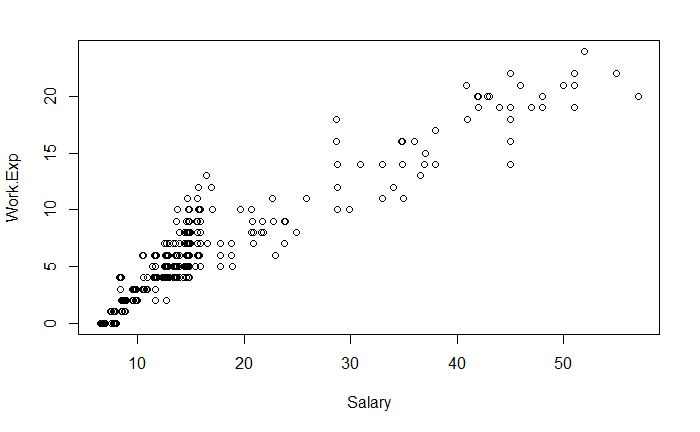
* Staff who use cars live far within
* Staff within a closer distance range uses public transport.
* Employee who uses car live 13km – 23,4
* People who live within 18km also uses public transport
* It is more likely that a person that lives within 8km range would use a public transport



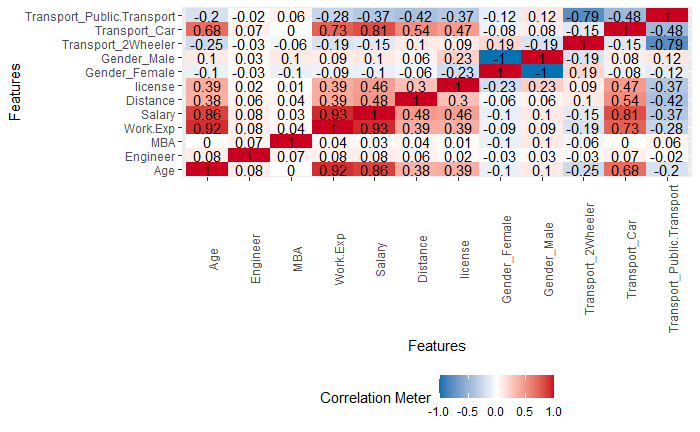
* Staff that earn $ 30 to $ 50 uses cars
* Staff who earn lesser than 23k shuffle between
* Employee that earn more can afford to use cars



* Age and salary share an increasing relationship. As age increases salary increases among employees. It is unlikely to find an employee age below 30 with a salary above $30.
* There’s a Linear relationship



* Work exp. Rises with salary, the more work exp. Of an employee the more amount of pay they receive. Which explains the increasing relationship between age and salary.



Correlation plot shows:

* A strong relationship between age and salary, work exp.
* A strong relationship between car transport and age, work exp., salary, distance
* A weak relationship with 2-wheeler and female employees and male employees with public transport.
* As the relationship with transport cars increases other mode of transport decreases.

# Data Partition

There is a class bias problem with the dataset, this condition is observed when a proportion of events is much smaller than the proportion of non-events. There’s a need to sample the observations in approximately equal proportions to get better models.

The dataset has been split into 30% Test and 70 % Train

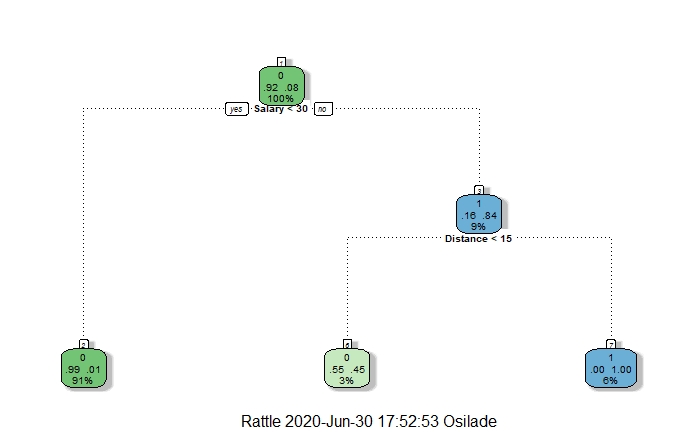
* Train data set – 292 observations and 9 variables, Target Variable – 268 (0), 23(1)
* Test data set - 126 observations and 9 variables, Target Variable – 114 (0), 12(1)

A high bias can cause an algorithm to miss the relevant relations between features and target outputs causing underfitting.

A High variance can cause an algorithm to model the random noise in the training data, rather than the intended outfits causing overfitting..

# Models

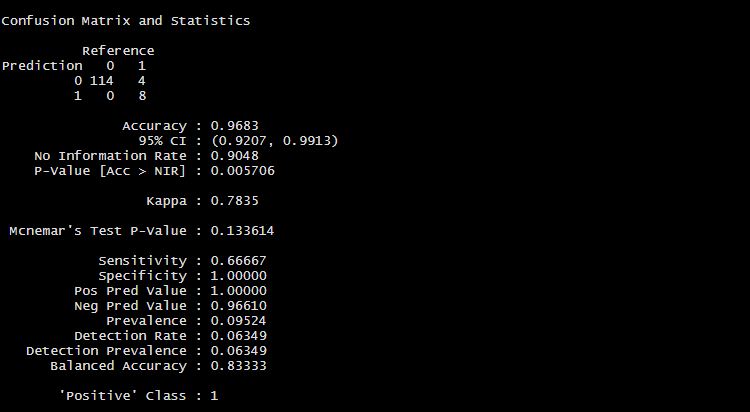
## Cart Model



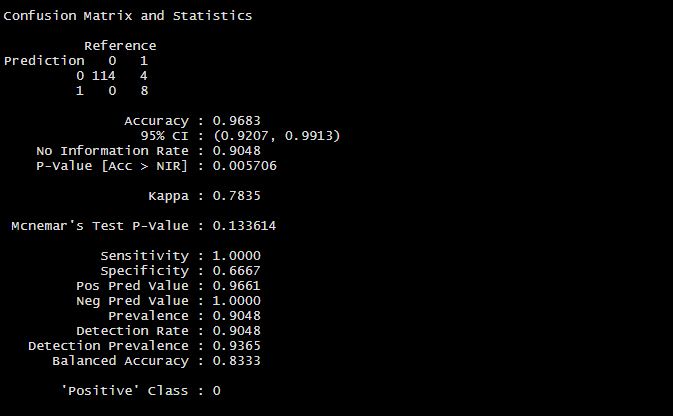
Using the decision trees model to decide who uses cars, based on the values of other variables in from the dataset.

* Only an 8% of employees uses cars
* Employees who earn salaries less than $30k have a 91% chance of not using a car.
* Employees who earn above $30k have a 9% chance of owning.
* Employees who earn above 30k and live more 15 kilometres away have a 6% chance of owning a car.
* Employees who do not earn above 30k but live closer than 15 kilometres have a 3% chance of owning a car.

Model Performance

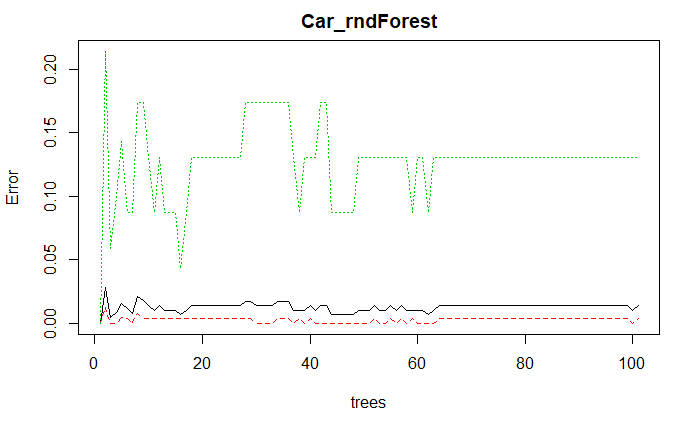


* On the train set we predicted 114 TPs.
* Accuracy is 0.93
* Sensitivity is 0.66
* Specificity is 1



* On the test set we predicted accuracy at 0.96, specificity at 1 and sensitivity at 0.66
* The above numbers are calculated on the validation sample that was not used for training the model.
* So, a truth detection rate of 97% on test data is good.

## Random Forest Model



* Out of bag error rate is 1.37%
* No. of variables tried at each split: 3 nodes
* We predicted 267 true positive and 1 False on the train set

### Plotting Variable importance

0 1 MeanDecreaseAccuracy MeanDecreaseGini

Age 3.031942 4.1567621 5.163106236 5.86928395

Gender 1.004988 1.4062626 1.672201105 0.04644393

Engineer -2.036517 0.8964697 -0.003450204 0.15758200

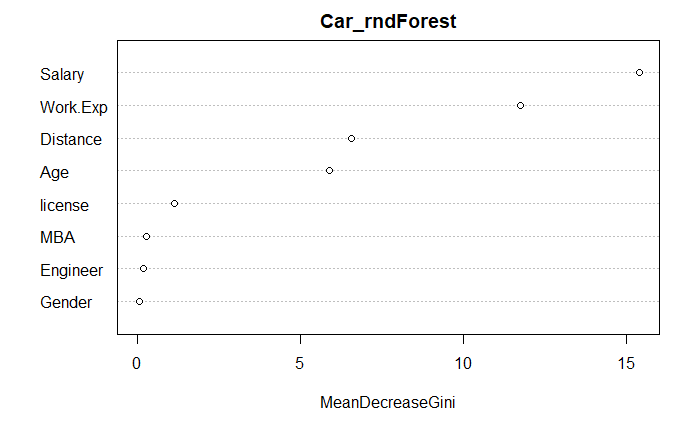
MBA 1.758212 0.2924816 1.269489090 0.26392711

Work.Exp 3.824456 7.1456998 7.169369109 11.72994640

Salary 6.695947 11.2649733 10.945053228 15.39786659

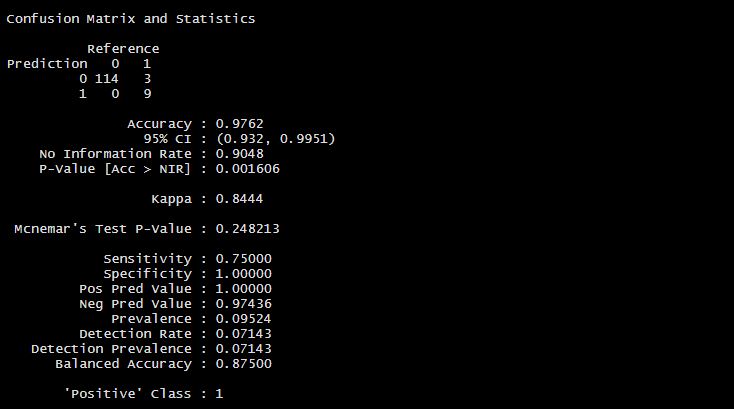
Distance 3.513830 8.3180871 8.778381467 6.56894560

license -1.017870 3.1212787 3.143002133 1.10938357

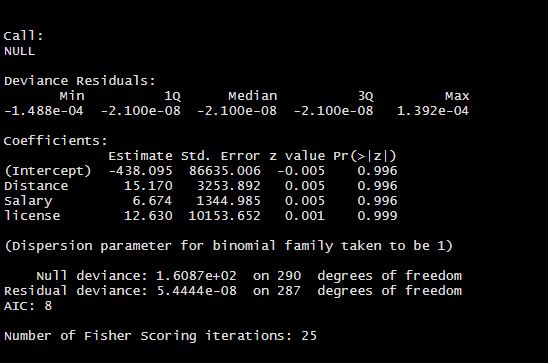


* Variables Salary and Work.Exp are considered high in importance.
* Gender is the least important, its not on/ does not contribute to whether employees will use car or not.

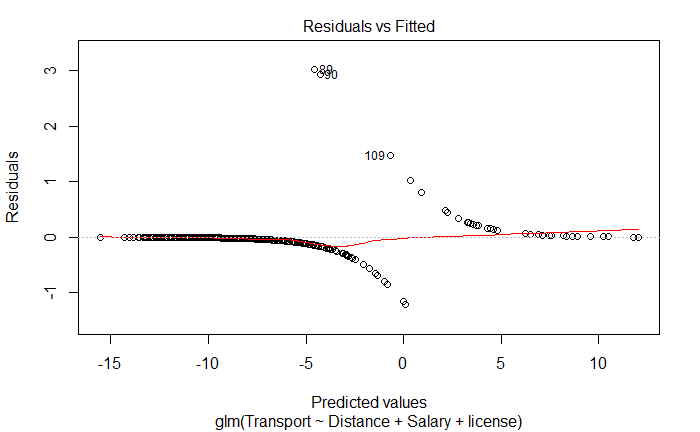
### Model Performance

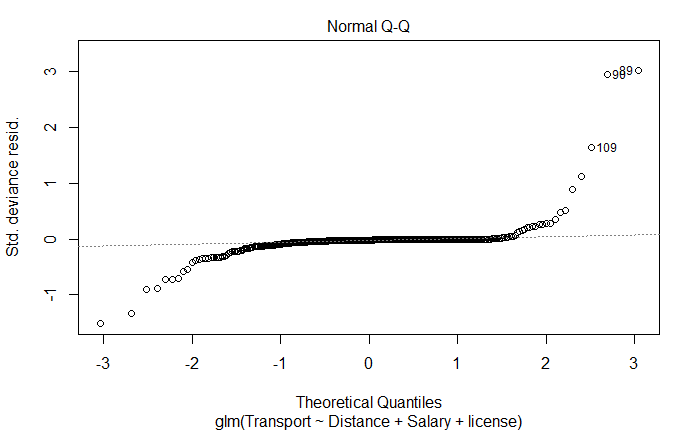


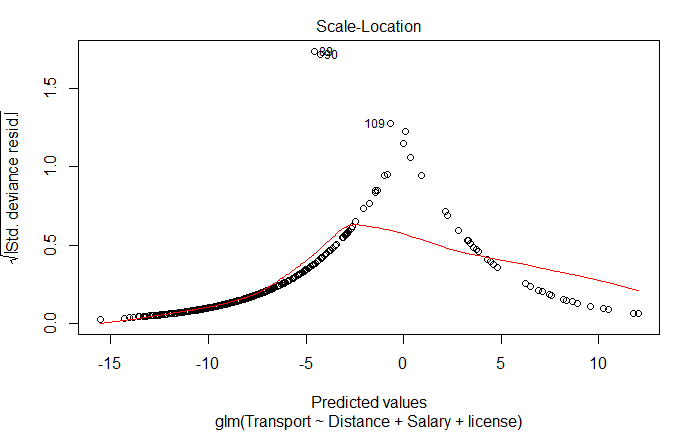
## Logistic Model

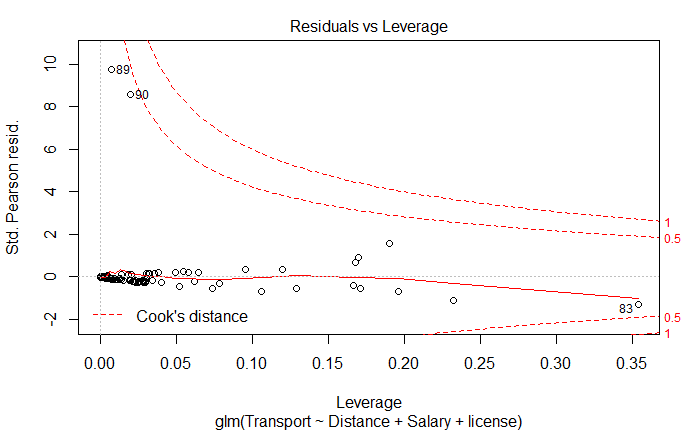


* Variables in the trainset don’t show any significance in the model.
* p Value turns out greater than significance level of 0.5 in determining transport
* Every 1-unit change in distance will increase the log odds of getting transport by 15.17

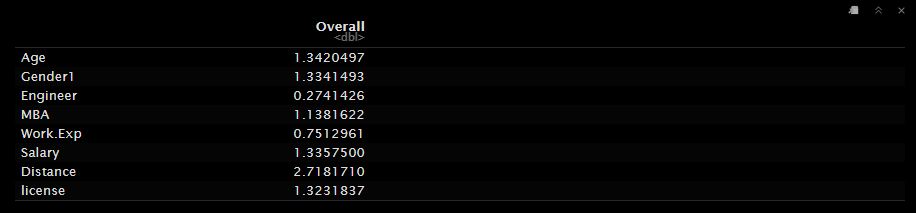








### Variable Importance in logistic regression



* Distance is considered high in importance but there’s no significant difference from other variables.
* Variables like Age, Gender, MBA, Salary, License are of least important but on the other hand does not does not contribute to whether employees will use car or not.
* Work exp. And Engineer have no business in predicting whether employees would use car or not.

### Variance inflation factor

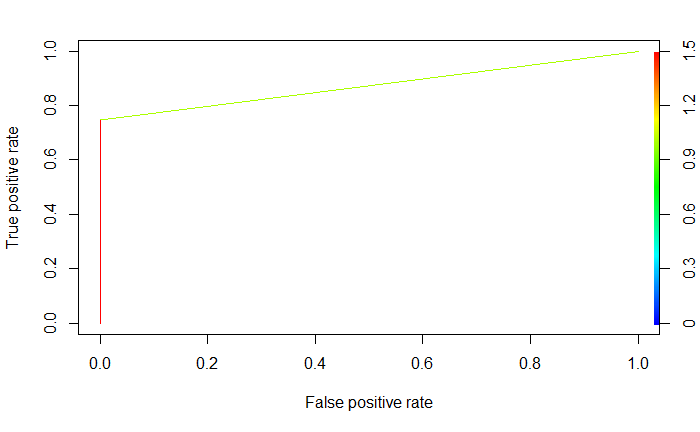
* Distance – 1.58
* Salary - 1.63
* License – 1. 31

As seen above, all the variables in the model have VIF well below 5.

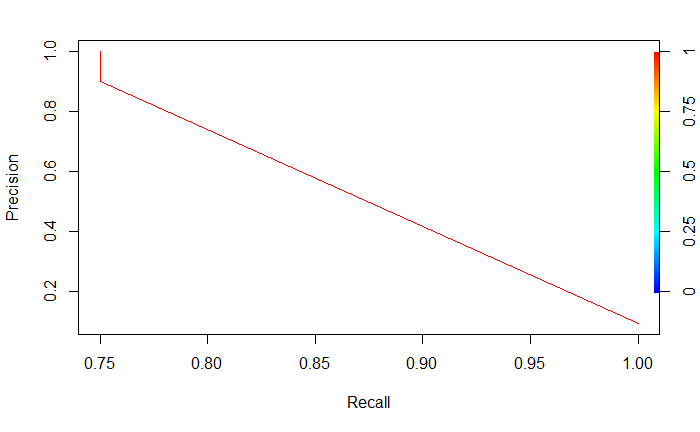
## Model performance

### ROC / AUC

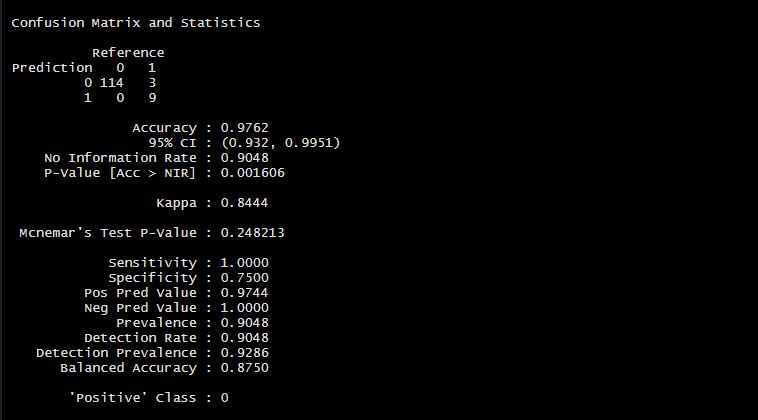
ROC shows the percentage of true positives accurately predicted by a given logit model as the prediction probability cut off is lowered from 1 to 0.



* AUC – 0.87



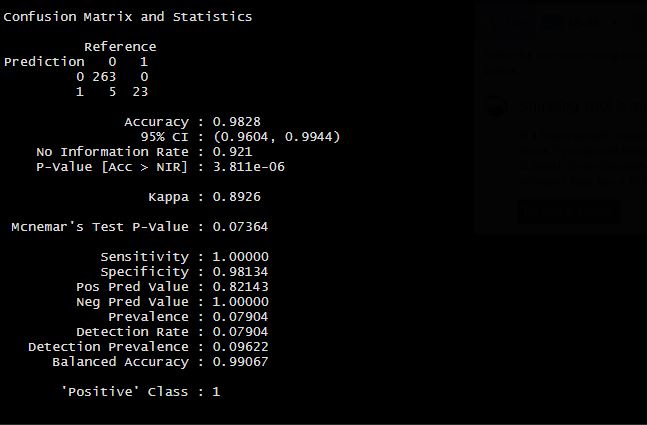
### Confusion Matrix



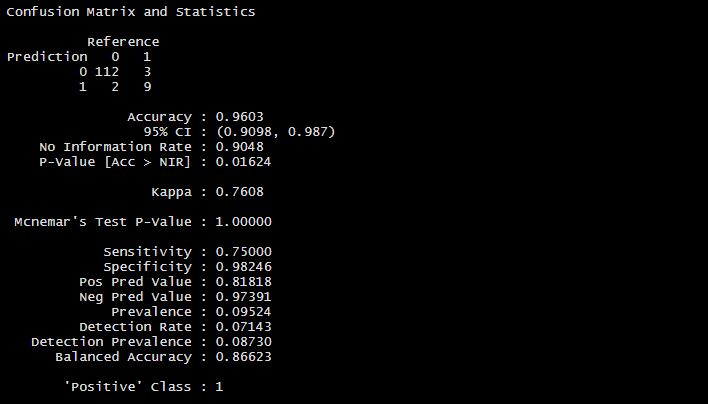
* The above numbers are calculated on the validation sample that was not used for training the model.
* Sensitivity is the percentage of 1’s correctly predicted by the model, while specificity is the percentage of 0’s correctly predicted.
* Sensitivity/ True positive is predicted at 1, Specificity is predicted at 0.75%
* So, a true detection rate of 1% on test data is good.

## Naïve Baiyes Model

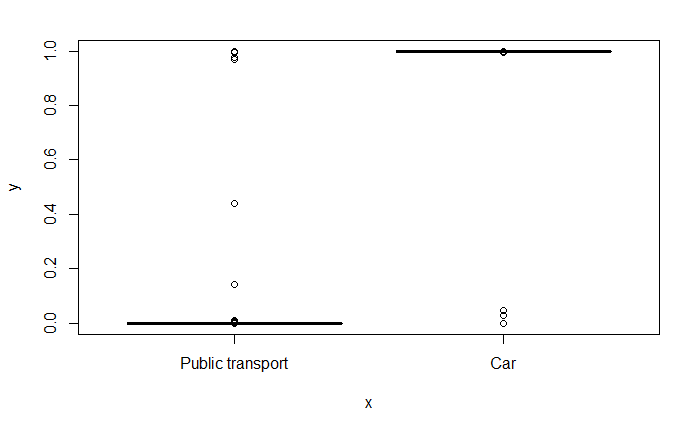
This model assumes all the variable to predict car transport is independent of each other and there is no correlation between them. The problem with the Naïve model is we need large dataset to have an estimate of all the different probabilities.



* Accuracy on the train set is train set is 98%
* Sensitivity is 1
* Specificity is 98%

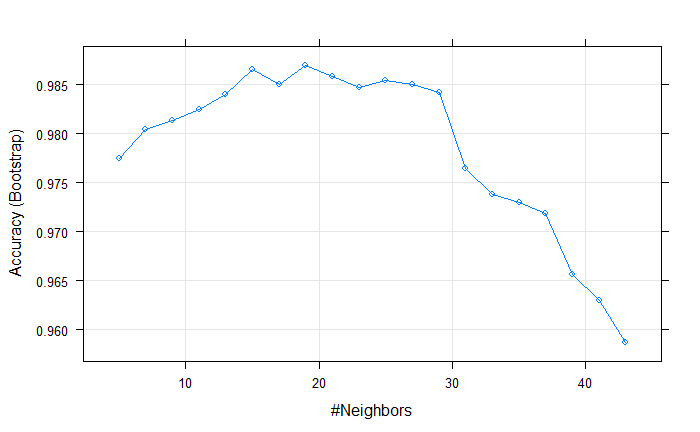


* Accuracy is 96%
* Sensitivity is 75%
* Specificity is 98%
* The accuracy on the training set is more than the accuracy on the test set and this explains that the model is too specific and not generalised.
* The model is overfitting



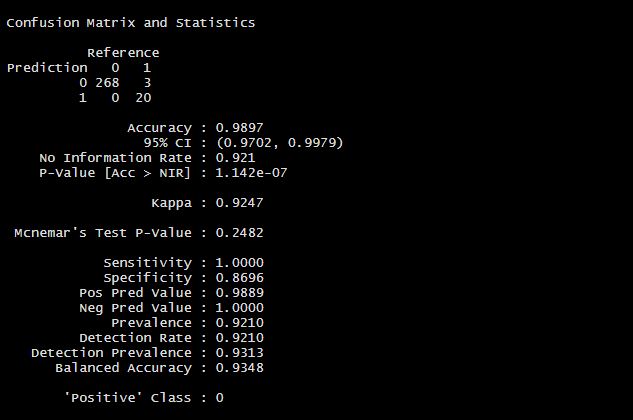
## KNN Model

We intend to find out the population that uses cars, It is necessary to scale the data, so that the variable can be on the same level, since we are finding nearest neighbours and KNN uses Euclidean distance. It tells us that that the population of people who use cars is dependent on the population around it.



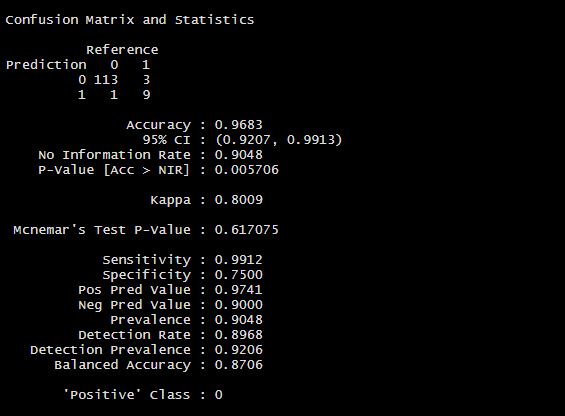
### Model performance

Train set



* We predicted 288 out of 291 and predicted 3 wrong on the train set
* Accuracy is 98%
* Sensitivity is 1
* Specificity is 86%

Test set



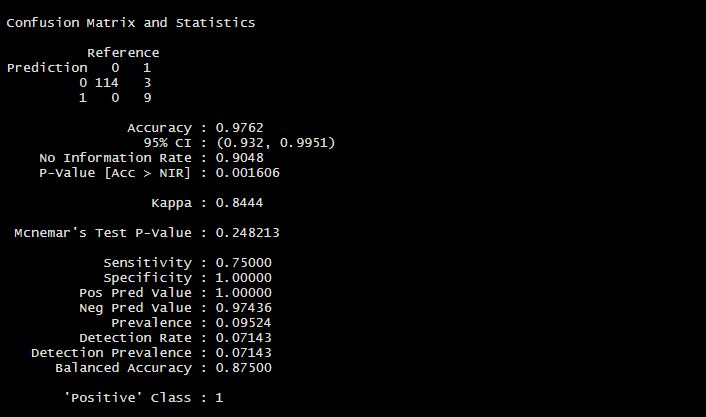
* The model seems to be overfitting
* Sensitivity is 99%
* We predicted 122 out of 126
* We predicted 4 value wrong
* Accuracy is 96%

## Bagging



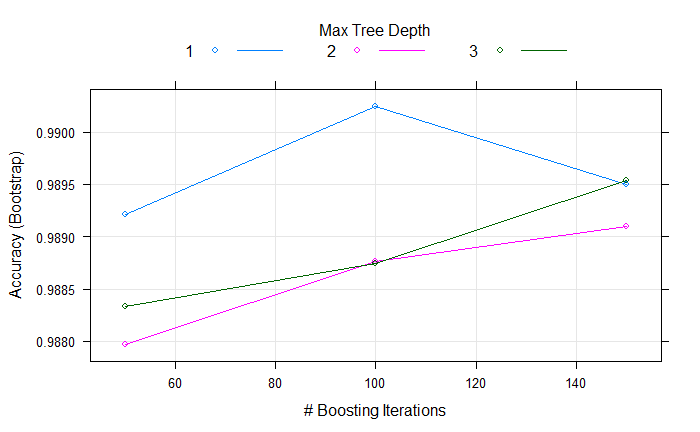
The important variables for bagging,

* Age -32.2, distance – 27.2 , salary - 40.9, workexp – 33.7 ., variable is of extreme importance
* Gender, Engineer and MBA as variables that add no significance

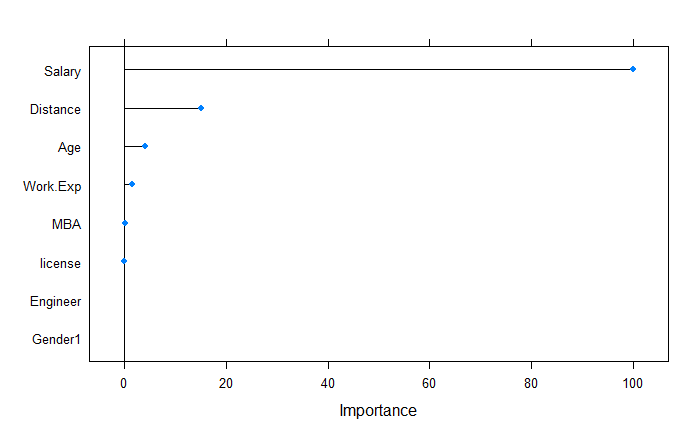


model performance has a sensitivity of 75%, we predicted 123 as true

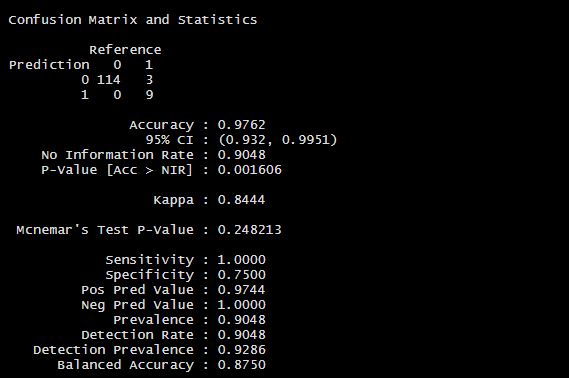
## Gradient Boosting



* Model has 3 trees



* The important variables are Salary, Distance
* License is a less important variable in gradient boosting, it doesn’t predict for who would use a car.



* The model predicted 123 out 126 to a 97% accuracy
* 3 wrong cases
* Sensitivity is at 1
* Specificity is at 75%

# Conclusion

Model Performed well in KNN model we predict more on the train set and less on the test set. A poorly chosen value for k may have resulted in a score with a high variance or a high bias Boosting model helps reduce both bias and variance in the data, we predicted 123 out of 126, with a high sensitivity, a good model in general, showed a stronger variable importance compared to other models. Bagging reduces the variance, but retains some of the bias, it showed more variable importance and more unimportant variables compared. The accuracy of the model KNN, Random forest, CART, Naïve bayes ranged from 96% - 98%, both bagging and boosting has an accuracy of 97% on the validation test, compared to KNN that has a 96% accuracy with a sensitivity of 98%