|  |
| --- |
| Machine Learning - Group Assignment |
| **Concepts covered – MLR Package** |

|  |  |
| --- | --- |
| **Group No:** | 3 |
| **Team Members** | 1. Divyalakshme Arangasubramanian 2. Meenu Sethuraman 3. Priya Srinivasan 4. Ravichandran Narayanan 5. Shiyam Sundar T K 6. Subashree Bhaskaran |

***Question:***

The attached data has records of 444 employees in a firm. The variables are described below

Age: Age of the employee

Gender: Gender of the employee

Engineer: Whether the employee is an Engineer

MBA: Whether the employee is a MBA

Work Exp (work experience in completed years)

Salary: Salary in Rs Lakhs

Distance (in Km): The distance between the employee’s residence and office

License: Whether the employee has a license

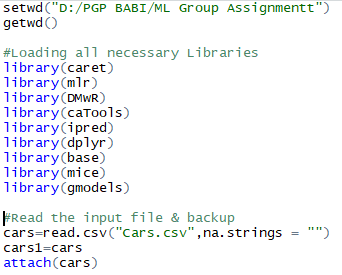
Transport: Main mode of transport taken by the employee

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data (first few records) Age** | **Gender** | **Engineer** | **MBA** | **Work Exp** | **Salary** | **Distance** | **license** | **Transport** |
| 28 | Male | 0 | 0 | 4 | 14.3 | 3.2 | 0 | Public Transport |
| 23 | Female | 1 | 0 | 4 | 8.3 | 3.3 | 0 | Public Transport |
| 29 | Male | 1 | 0 | 7 | 13.4 | 4.1 | 0 | Public Transport |
| 28 | Female | 1 | 1 | 5 | 13.4 | 4.5 | 0 | Public Transport |
| 27 | Male | 1 | 0 | 4 | 13.4 | 4.6 | 0 | Public Transport |
| 26 | Male | 1 | 0 | 4 | 12.3 | 4.8 | 1 | Public Transport |
| 28 | Male | 1 | 0 | 5 | 14.4 | 5.1 | 0 | 2Wheeler |

**Solution:**

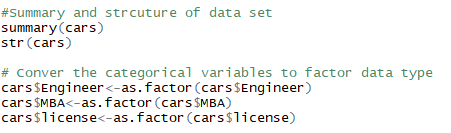
**Step1: Check the given problem statement read the CSV file into R.**

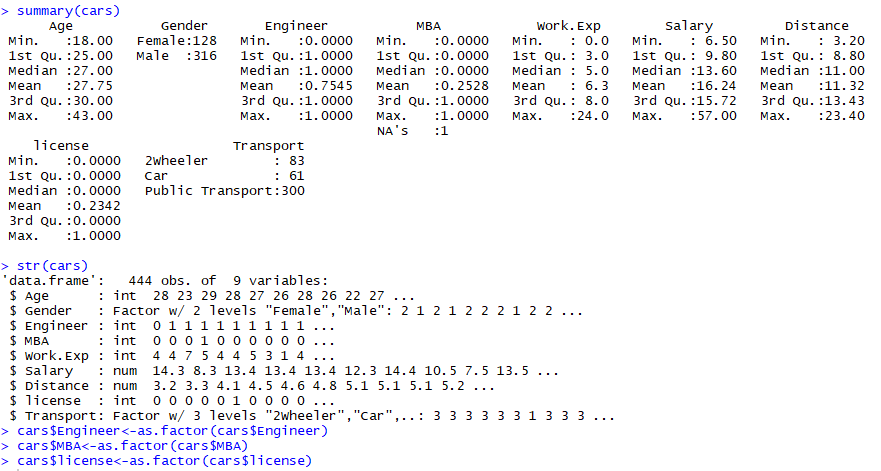
The working directory is set and the csv file is loaded in to R and the necessary packages are loaded following the file upload.



**Step2: Run a summary of the csv file and check for missing values and variable’s data type:**

Run a summary of the csv file to understand the distribution of the variables. But this step should be associated with str function to understand the data type of each variable. The variables “Engineer, MBA & license” which are categorical in nature but in different format in R, are converted into factor variables. It can be seen that there is on e

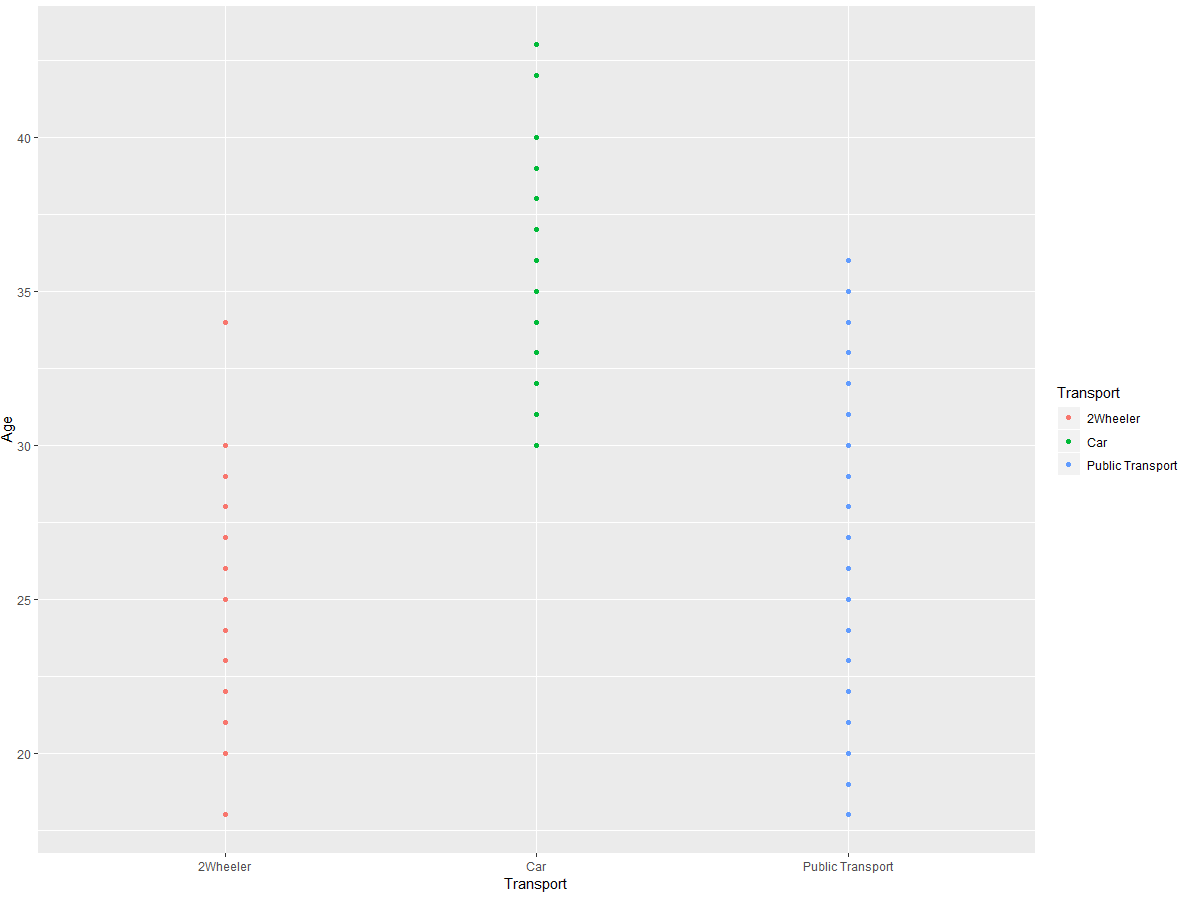




**Step3: Exploratory Data Analysis (Univariate and Bivariate):**

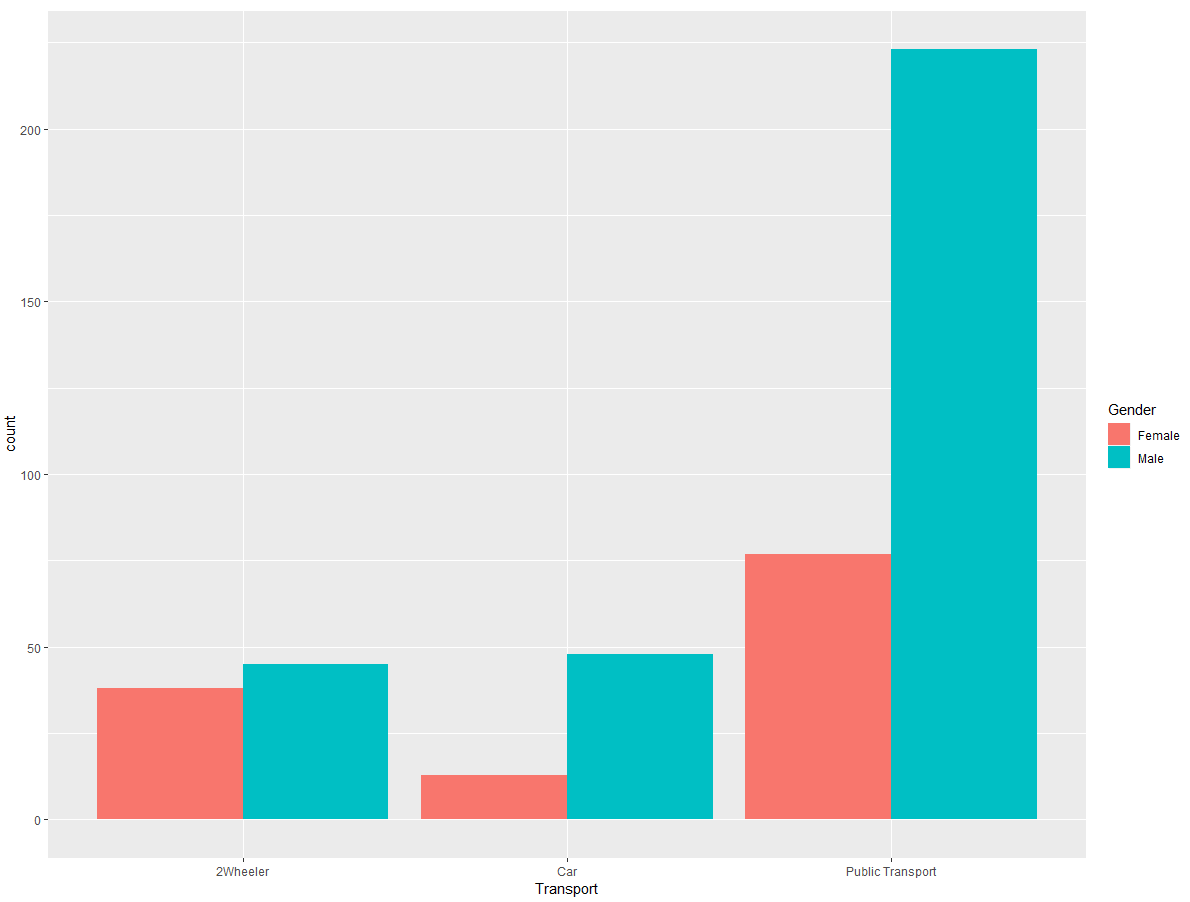
Using ggplot package, try different combination of variables to understand the inter-relationship. 

***Plot 1: Age vs Transport***



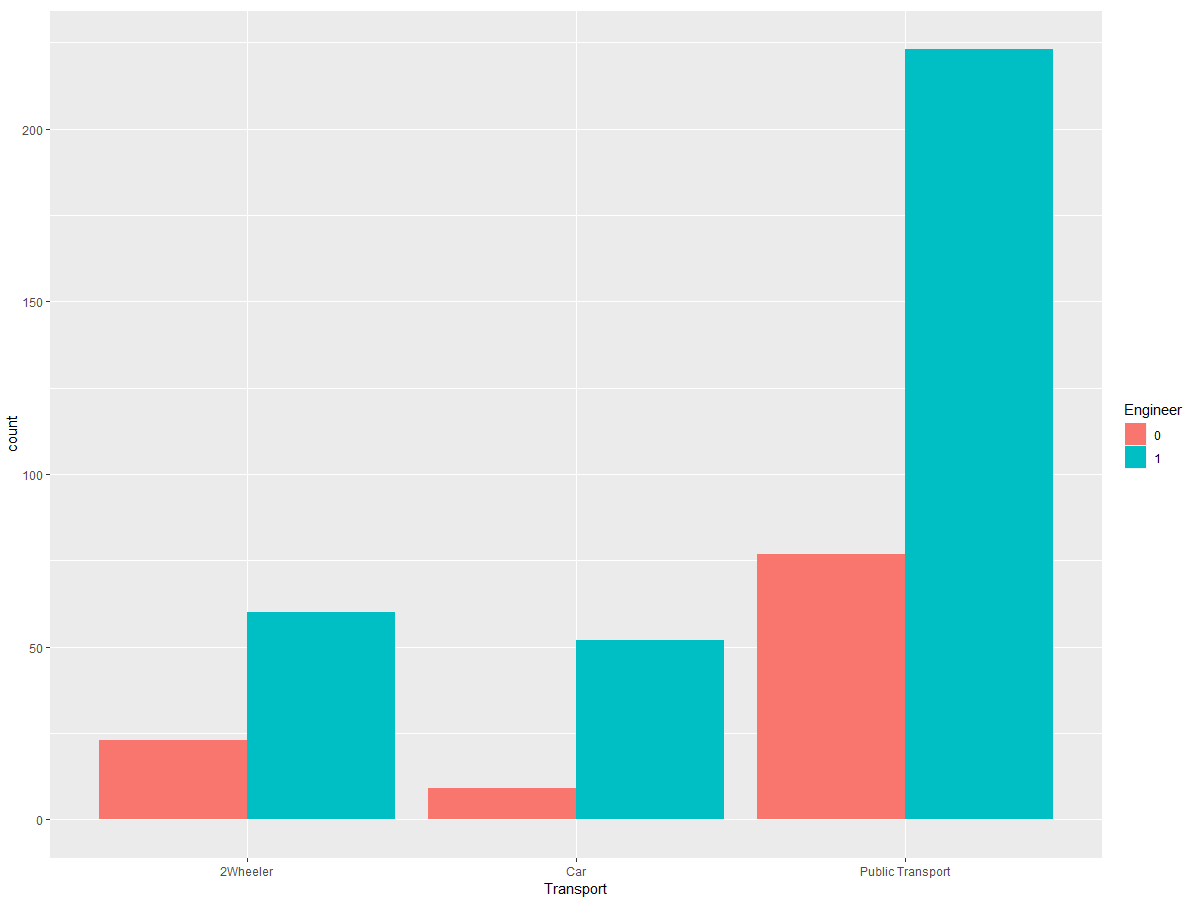
It can be seen from the above plot that car user’s age is specifically ranged high (30 – 50) compared to other two mode of transports.

***Plot 2: Transport type (count) vs Gender***



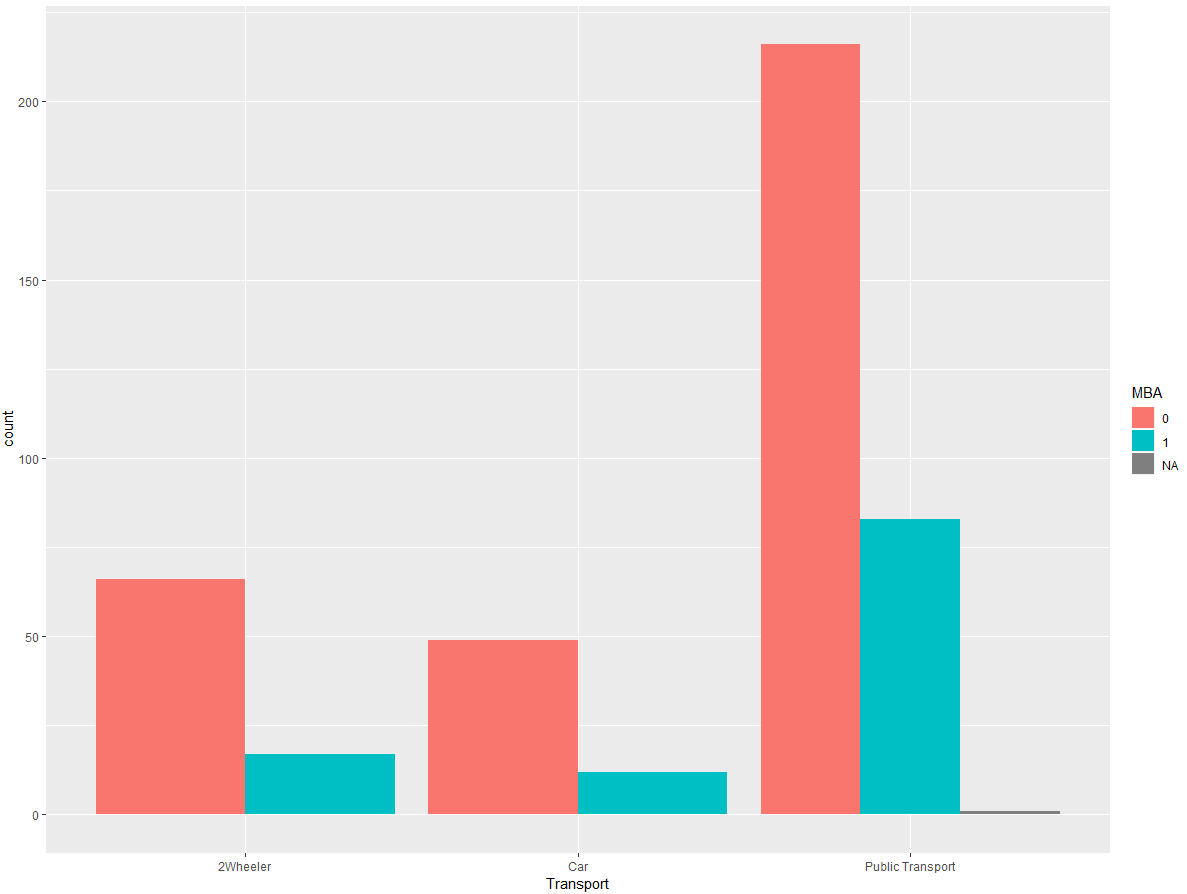
It very evident that Male users show more affinity towards public transport and of all transport modes, males count more in terms of gender vs transportation mode type proportion.

***Plot 3: Transport type (count) vs Engineer status***



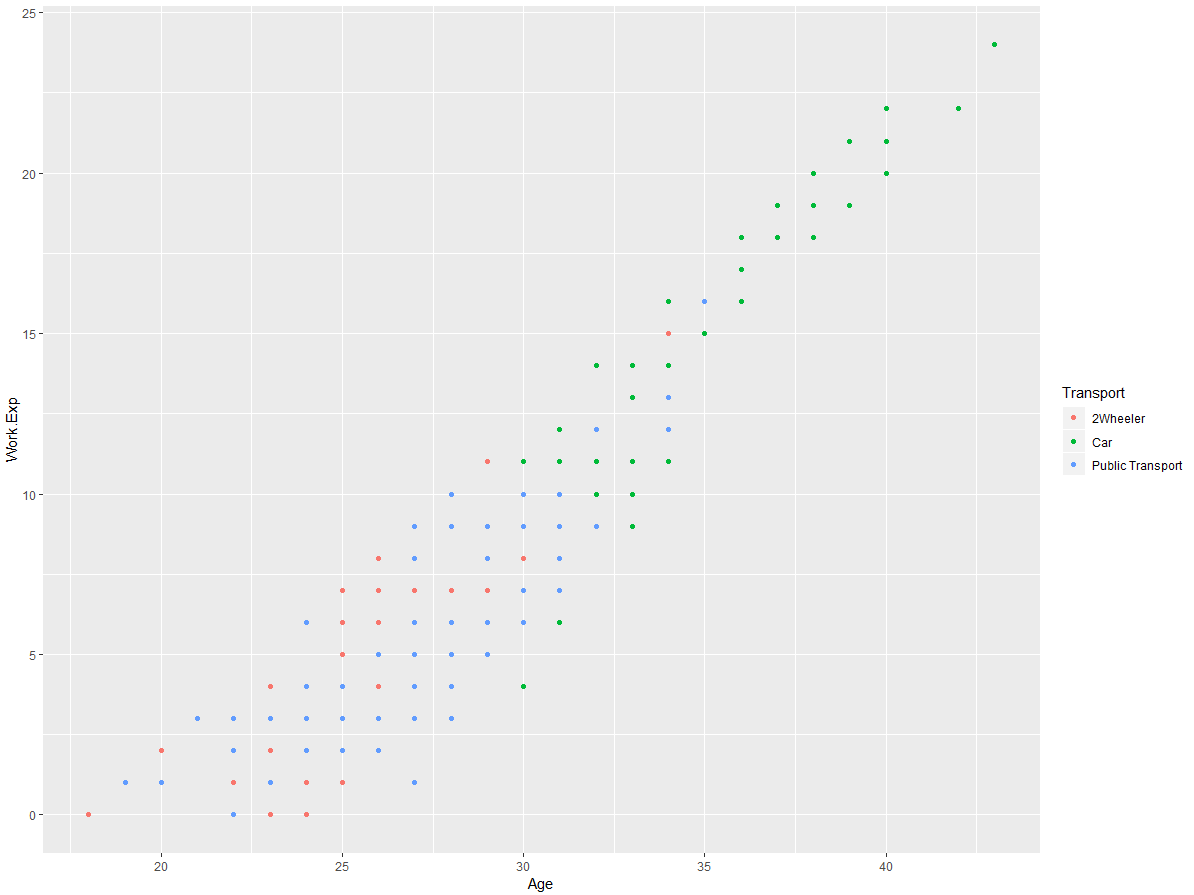
It can be seen that Engineers prefer public transport more compared to other modes of transport.

***Plot 4: Transport type (count) vs MBA status***



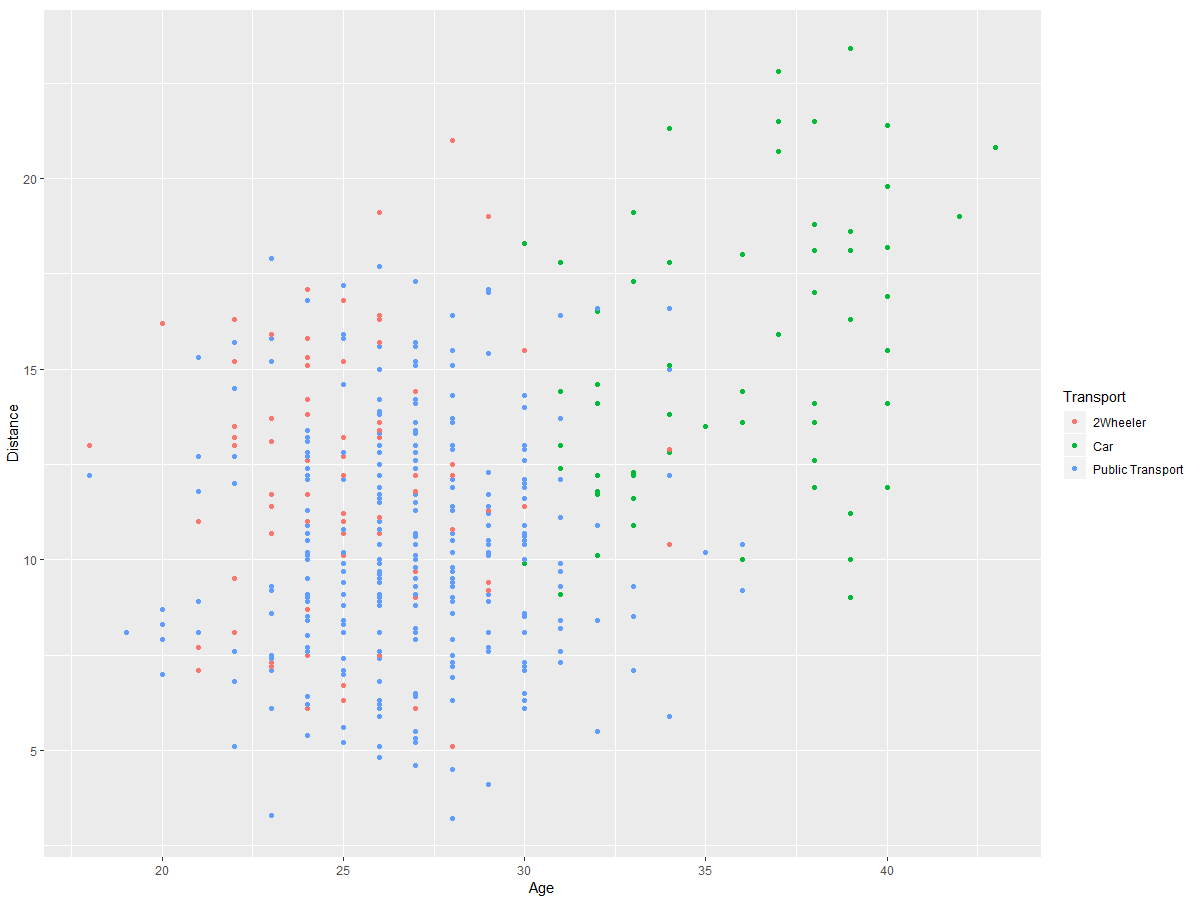
Unlike Engineers, in MBA grads segment, Non-MBA’s prefer public transport more when compared to non-MBA grads. This plot also shows that, the number of Non-MBA’s are more compared to number of MBA grads.

***Plot 5: Work Experience vs Age on Transport Mode***



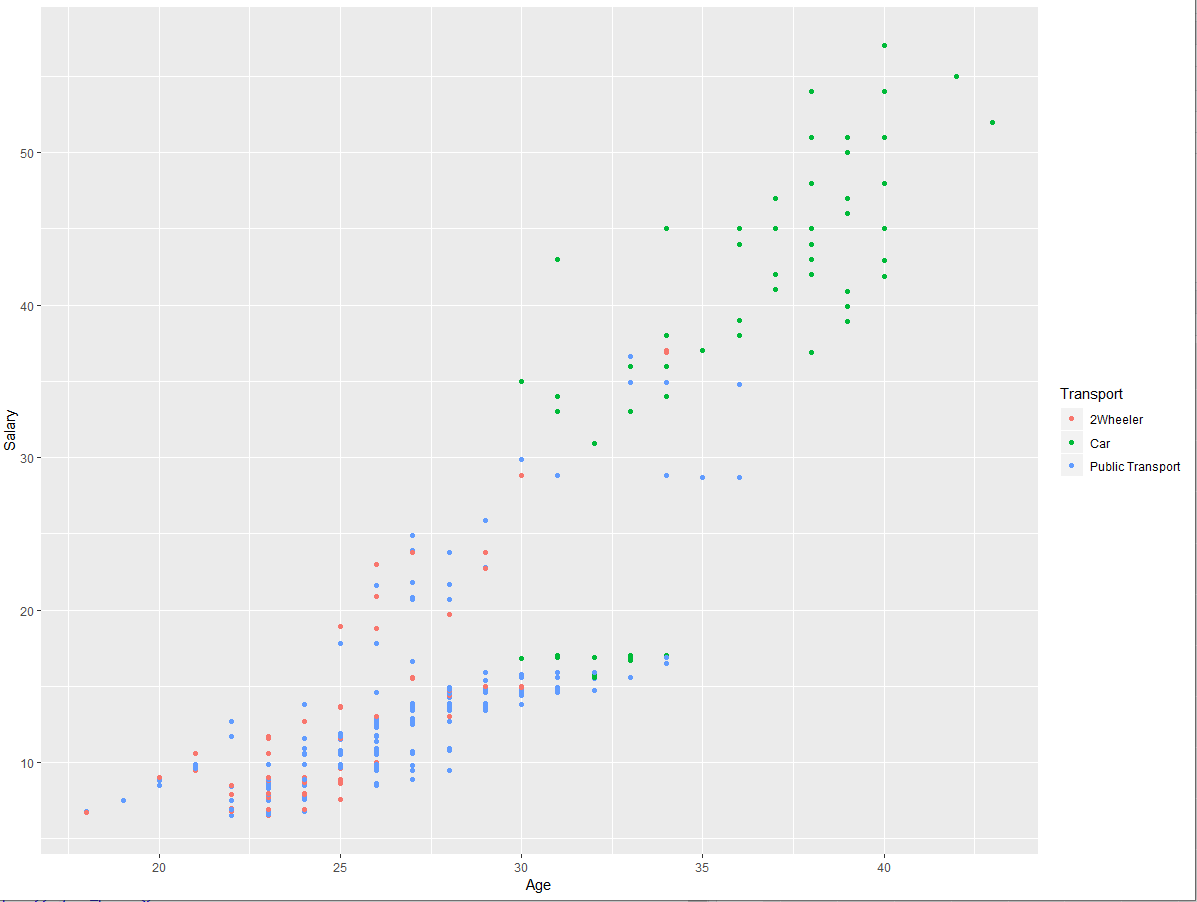
It can roughly inferred that, from age 30 and above, with a experience level ranging from 10 yrs and above, people mostly prefer car compared to people below this segmentation.

***Plot 6: Work Experience vs Age on Transport Mode***



People with age 30 and above, falling under a distance bracket of 10 units and more, prefer car as mode of transport.

***Plot 7: Salary vs Age on Transport Mode***



People aged more than 30 and having income level more than 30 Lakhs prefer car over other modes of transport.

***Plot 8: Salary vs Age on Transport Mode***



It can noticed, that Non-MBA’s who are aged more prefer car. This signifies Age is am important factor in determining the mode of transport.

***Plot 9: Work Experience vs Age on Transport Mode***



More experienced more people prefer car. Except for some outliers, people with experience of more than 10 years prefer car. This confirms in addition to age, work experience too is an important factor in determining the transport mode preference.

***Plot 10: Gender vs Age on Transport Mode***



The given data set is more composed of male when compared to female and car users are more of male and in terms of 2wheeler, people with age<30 are equally contributing.

***Plot 11: Engineer vs Age on Transport Mode***

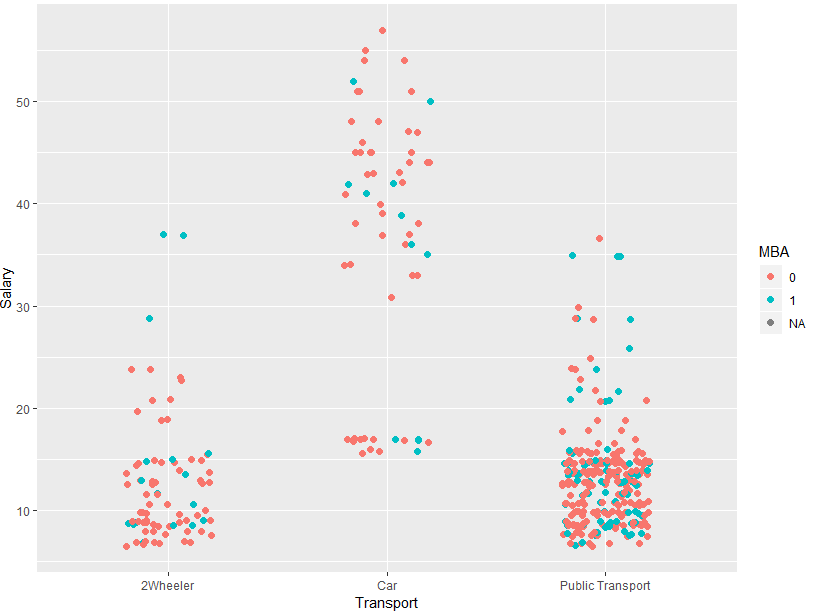


***Plot 12: Engineer vs Salary on Transport Mode***



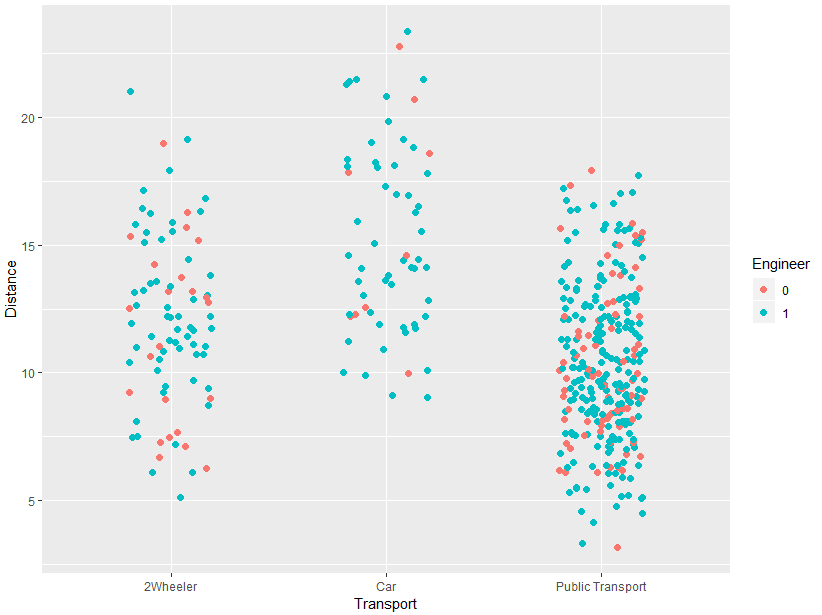
Inference: High salaried engineers prefer car and low salaried engineers prefer public transport more when compared to 2wheeler.

***Plot 13: MBA vs Salary on Transport Mode***



Inference: Non-MBA’s with high salary prefer car when compared to MBA’s with high salary.

***Plot 14: Engineer vs Distance on Transport Mode***



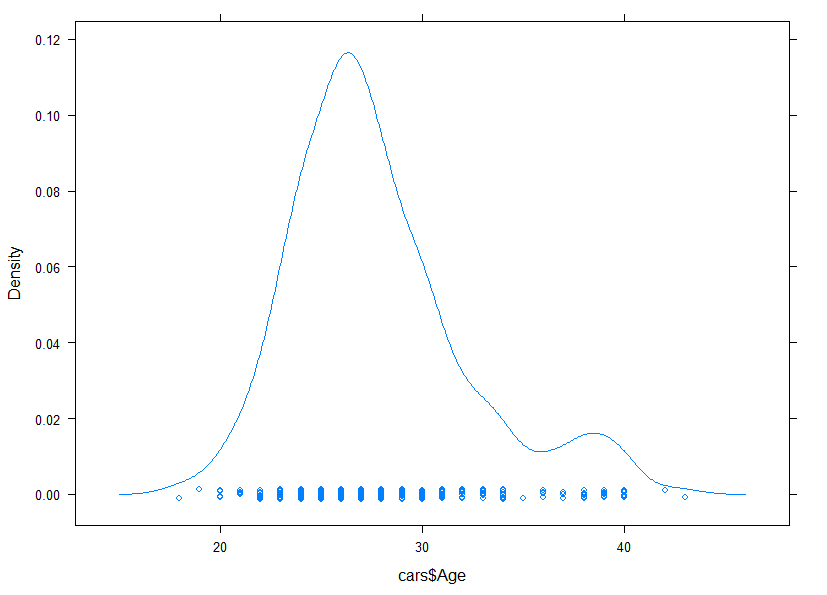
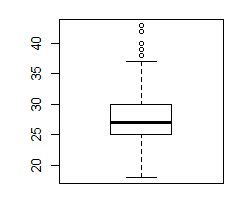
Inference: Non-Engineers mostly prefer public transport.

***Plot 14: Engineer vs Distance on Transport Mode***

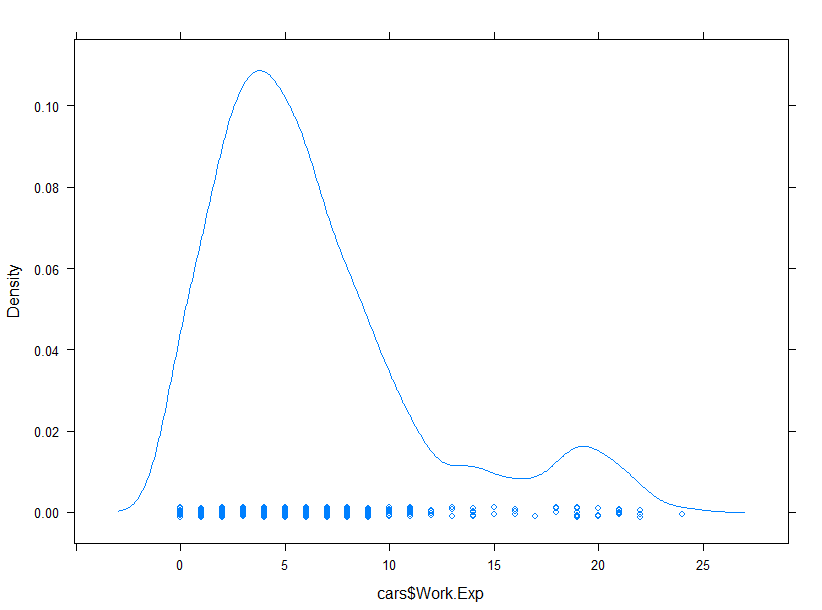
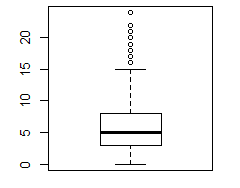


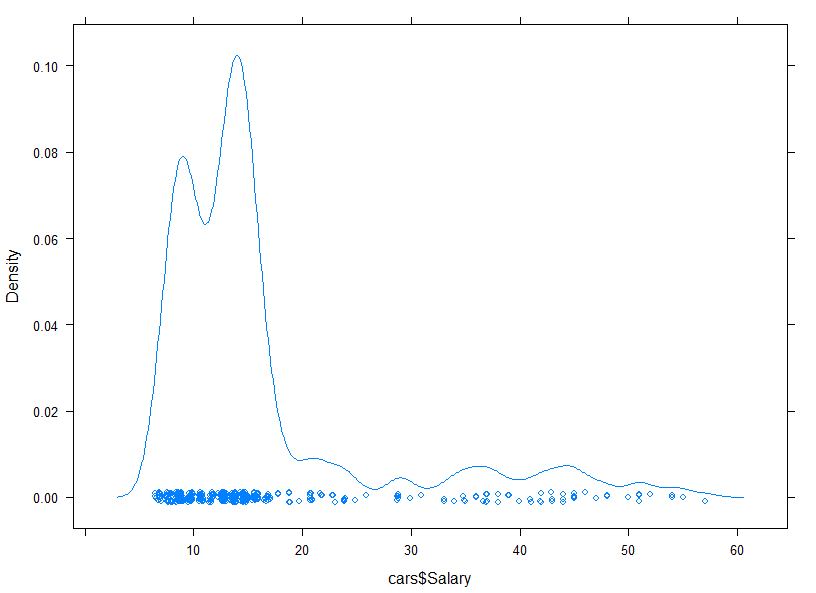
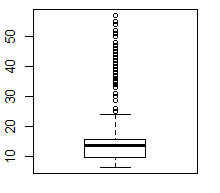
Inference: No clear conclusion could be derived except for a rough estimate depicting non-MBA’s prefer car when the distance is more than 10 Kms and this bucket could be high earning people.

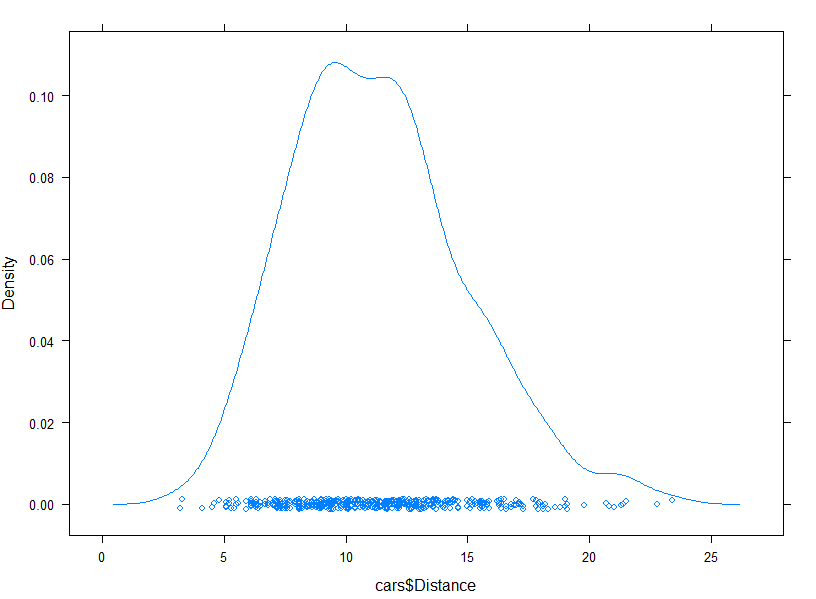
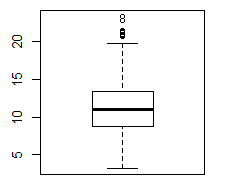
***Plot 15: Age Distribution***

***Plot 16: Work Experience Distribution***

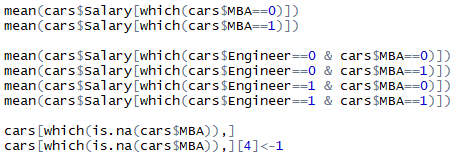
 ***Plot 17: Salary Distribution***

 ***Plot 18: Distance Distribution***

General inference on Univariate EDA: Even the box plot/s shows outliers, those outliers are the essential contributors for the car preferring population (This can be observed from the multivariate plots above). If we consider them as outliers and treat them, the dataset will become imbalanced. Hence, the outliers are to be retained as such.

**Step4: Missing value imputation:**

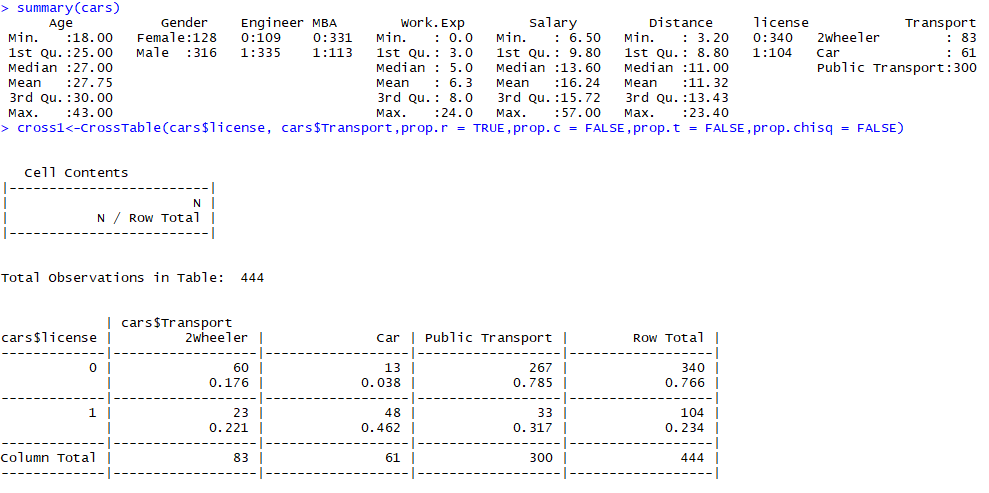


The one missing value exists in the MBA column. It can be replaced with mean (or) mode, since one imputation won’t show any big effects in the data set. We chose to calculate the conditional mean of existing MBA=0 & 1 and compared the salary variable trait corresponding to the missing itemset and made replacement accordingly as MBA=1.

**Step5: Recheck for missing value and do a cross table for license variable:**

Recheck using the summary function for missing values. Since license was not covered in the EDA portion, do a cross table to under its effect.

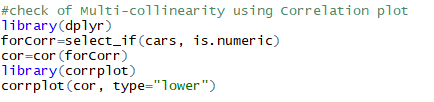


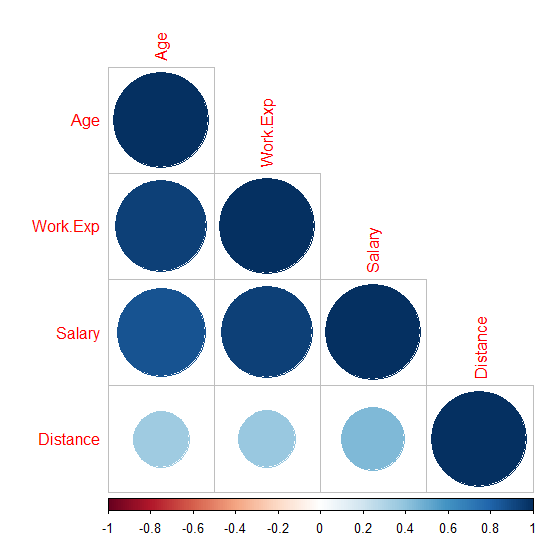


It can be seen that, no missing values exists and people with license tend to prefer their own four wheeler transport (Car).

**Step6: Check for multicollinearity:**

Using the correlation matrix, check if any multicollinearity exists. Looking at the data and EDA’s above, it can roughly estimated that with Age, people will become more experienced professionally and their salaries could be linear associated with age and work experience. Let us check if this is reflected in the correlation plot.

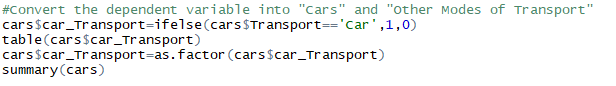


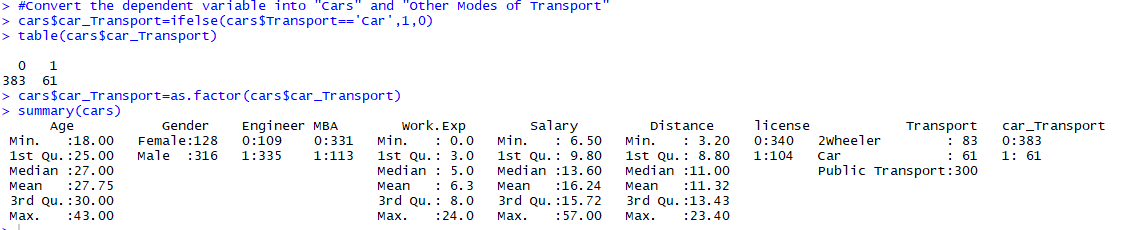


It looks like the initial assumption is indeed true. Hence, the issue of multicollinearity has to be addressed if we build models which banks on no-multicollinearity assumptions.

**Step7: Convert the dependent variable into “Cars” and “Other Modes of Transport”:**

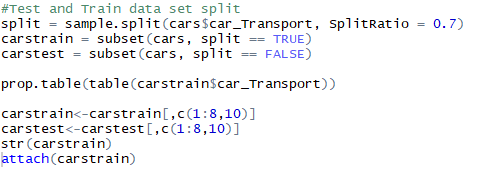
Since, the problem statement is to build model which best explains the employee’s decision for opting car, we have converted the dependent variable into “0” and “1” under a new variable name “car\_Transport”, where “0” stands for other modes of transport and “1” stands for car. Other reason being, logistic regression cannot work on dependent variables having more than two factors, this methodology has been carried out.

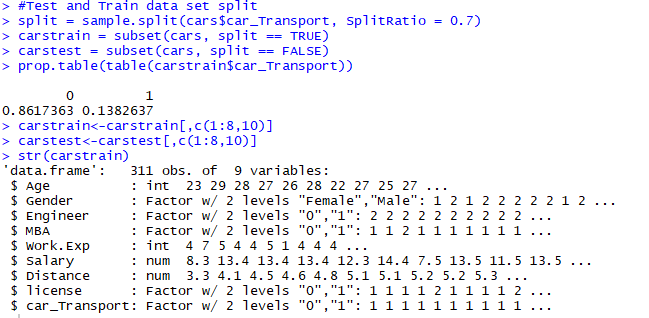




**Step8: Split the data set into train and test sets:**

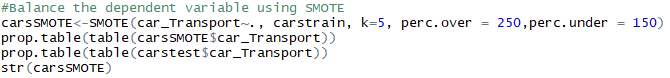
Split the data set into train and test at 70:30 ratio. The initial Transport variable is to be removed from both Train and Test data sets. Finally, the data types of the other variables are checked to ensure nil discrepancy.

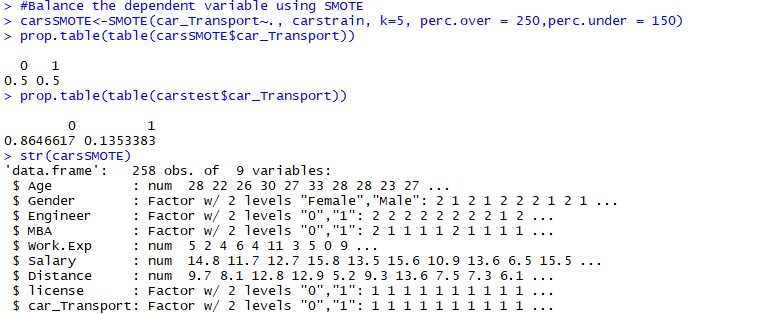




**Step9: Use SMOTE to balance the data set:**

Since the proportion of 0’s and 1’s in target columns is 86% and 14%, use smote function to create synthetic data points to balance the data set.





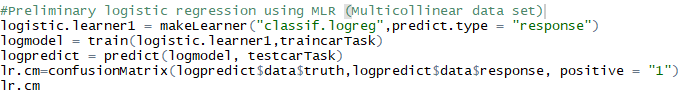
**Step10: Create Train and Test Task:**

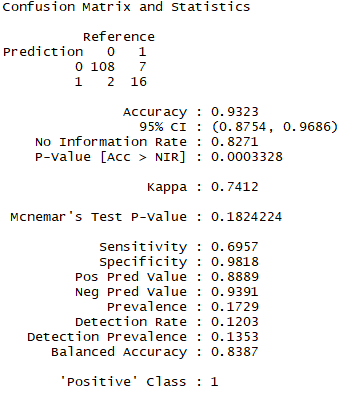
Create train task using the smote data set and test task using the test data set with positive value indicator as “1”.



**Step11: Build Preliminary Logistic Regression Model using MLR:**

Build a preliminary logistic regression using MLR with prediction type as response. Then predict the model on the test task and build a confusion matrix. One point to note is that, this model is based on the data set which has multicollinearity problem. This needs to be rectified.



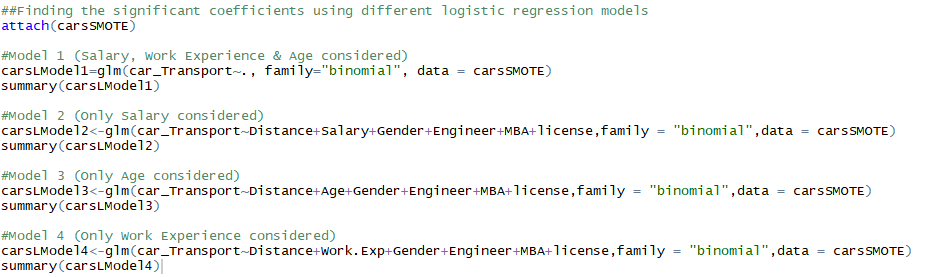


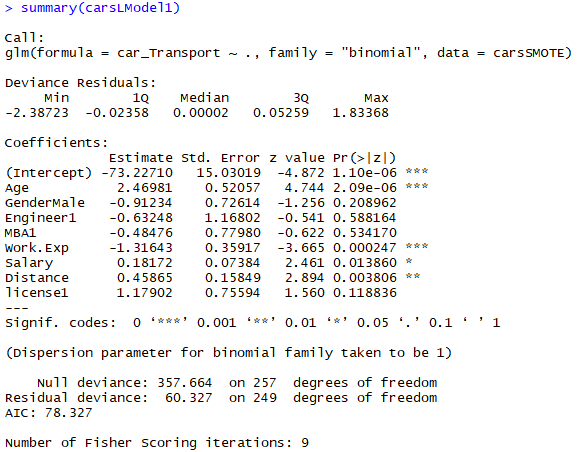
This model shows an accuracy of 93% with multicollinearity problem. Next step is to rectify this multicollinearity issue by building different logistic regression models with different variables.

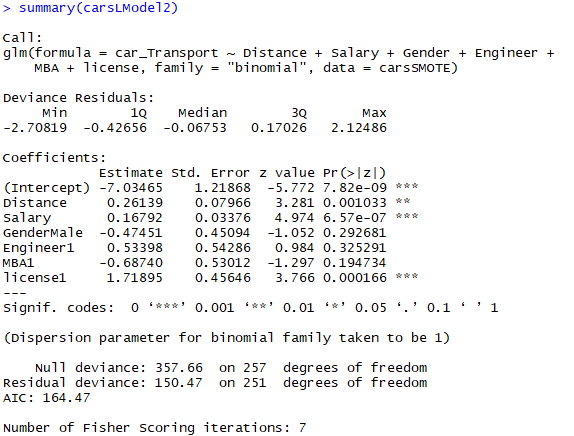
**Step12: Build Preliminary Logistic Regression Model using MLR:**

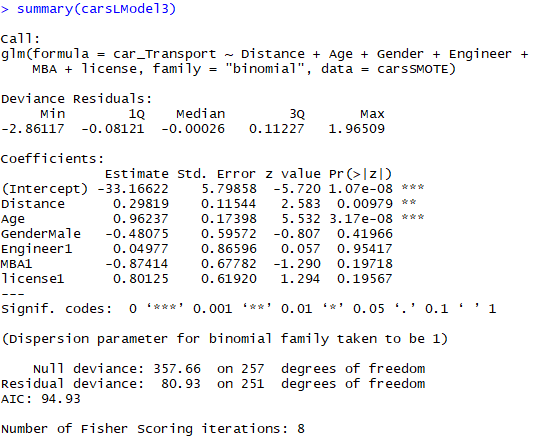
In addition to the preliminary logistic regression model, build three more models to identify which among the three multi-collinear variables (age, work experience & salary) is significant in model building.

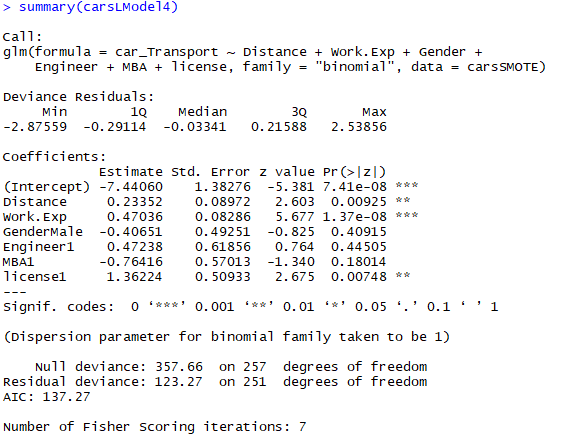
This can be identified by looking at the Z statistics shown in the summary of the logistic regression. Higher the Z value, more important the variable is. In that regard, variable “Age” and “Work Experience” seems to exhibit higher Z value. But looking at the ROC and AUC Plots, model with “Age” variable shows higher AUC. Hence, “Age” variable is deemed important and the train / test train sets are to be amended accordingly.





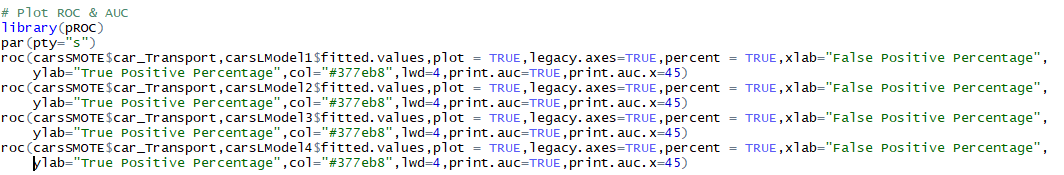




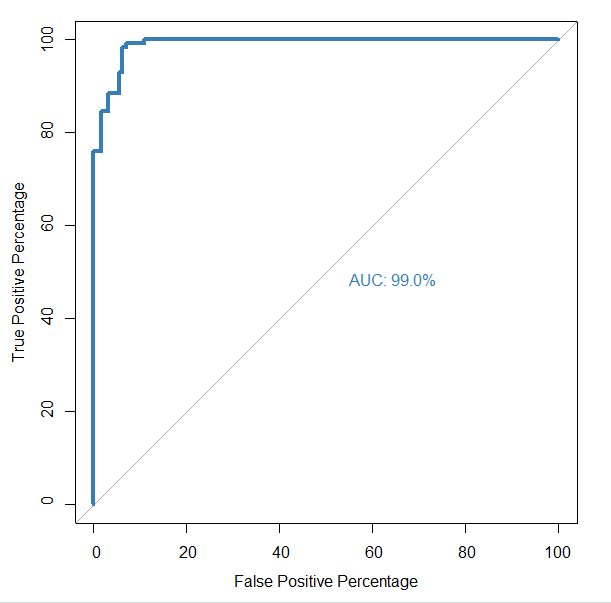


**Step13: Build ROC curve and measure the AUC:**

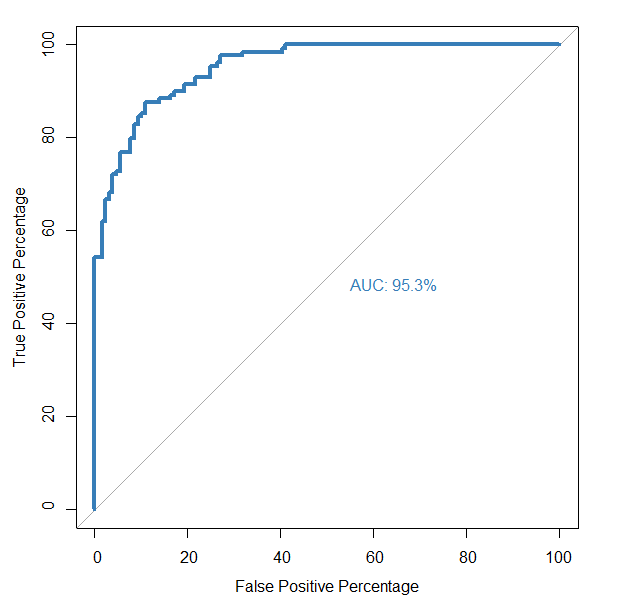
For the different logistic models stated above, the ROC and AUC graphs are shown below:



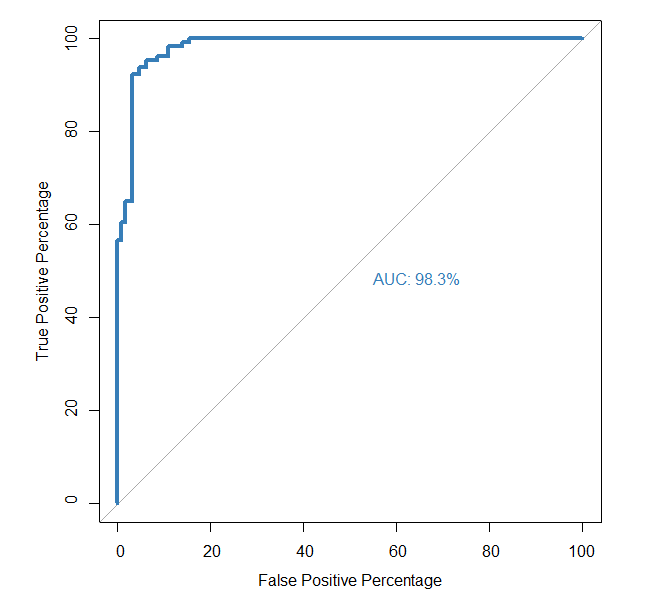
1. ROC and AUC for full model (Model 1):



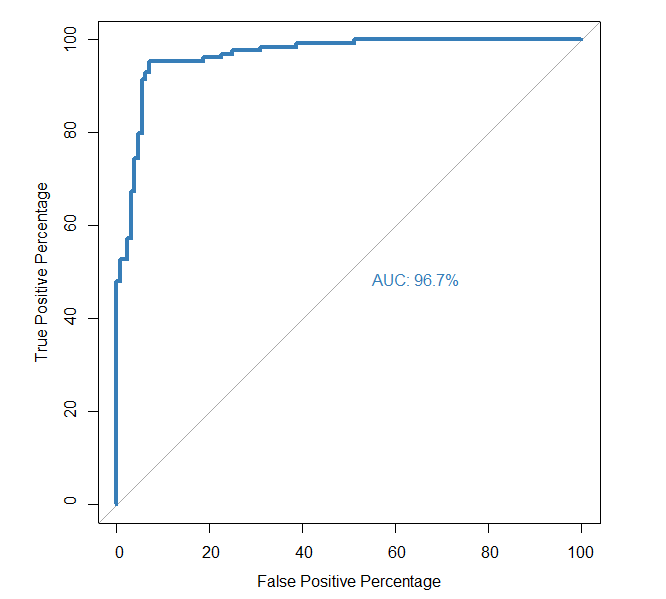
1. ROC and AUC for model with “Salary” (Model 2):



1. ROC and AUC for model with “Age” (Model 3):



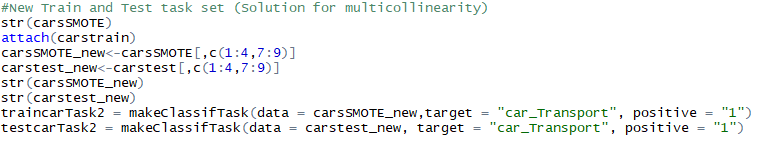
1. ROC and AUC for model with “Work Experience” (Model 4):



Since model 3 shows higher AUC, its source data set will be used as source for remaining models.

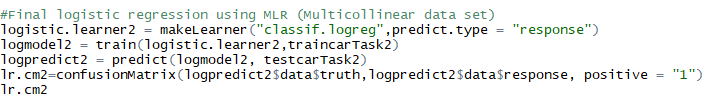
**Step13: Update Train and Test tasks:**

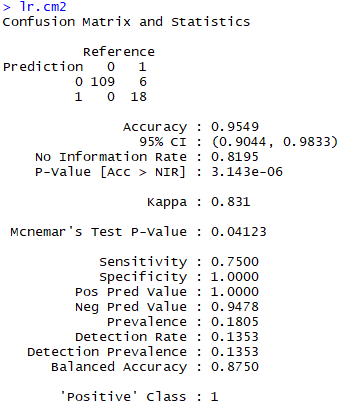
Inline with model 3, update the train and test tasks. Among the multi-collinear variables, “Work Experience and salary” have been removed leaving behind only variable “Age”.



**Step14: Final Logistic Regression Model:**

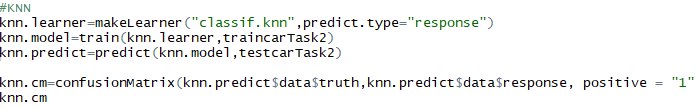
Using the updated train and test task, build the final logistic regression model. The confusion matrix and accuracy measures are shown below:

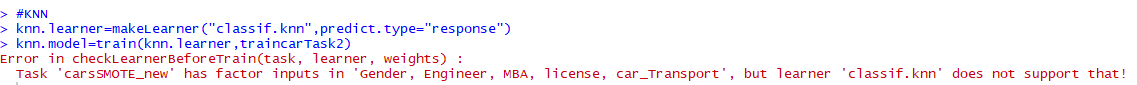




**Step14: K-Nearest Neighbour Model:**

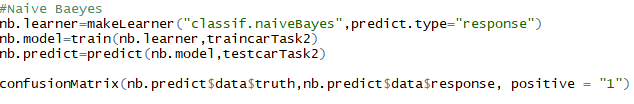
When trying to train the model with KNN learner, an error pops up saying KNN model doesn’t work with factor inputs. This can be rectified by converting the factor variables into dummy variables if needed. In this case, we can work with other models instead.

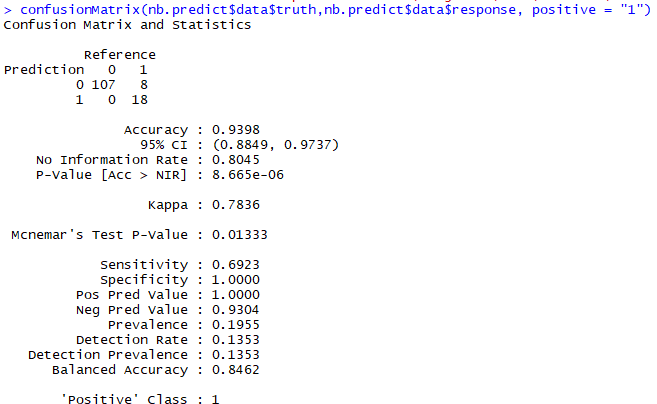




**Step15: Naïve Baeyes Model using MLR:**

Confusion matrix of NB Model built using MLR is shown below for reference:

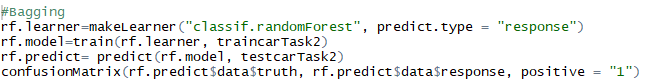


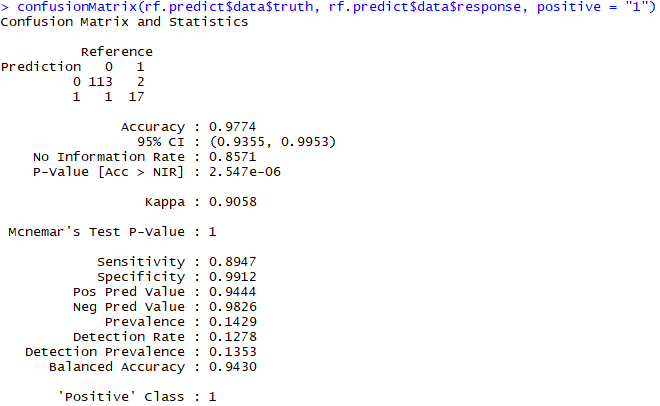


Since Naïve Baeyes is the least accurate model of all (owing to the assumption NB model employs – concept of independence between predictor variables), logistic regression shows accuracy of 95% which is higher than NB’ accuracy of 93.98%. In this case, we can go with logistic regression for future batch predictions.

**Step16: Bagging Model using MLR:**

Ensemble techniques are generally employed to boost the model accuracy. In this case, we will build a random forest (Bagging Model) to see if the model accuracy increases compared to logistic regression’s final model.

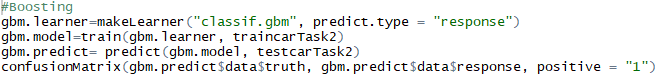


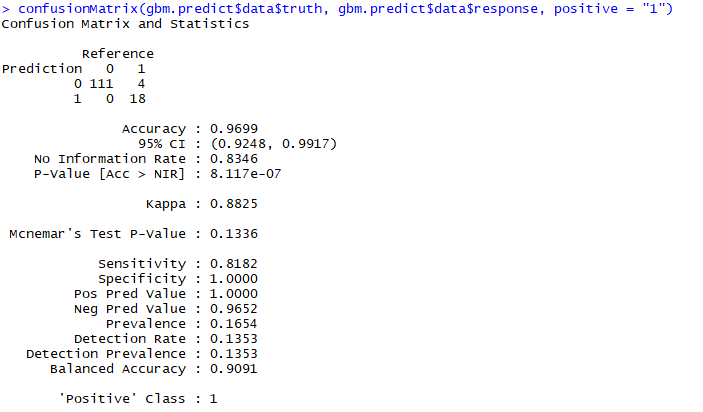


The random forest model shows a model accuracy of 97.74%. This is high compared to the logistic regression’s accuracy of 95%.

**Step16: Boosting Model using MLR:**

Since the data set is small, we will limit to only Gradient Boosting. XGBoost might result in overfitting the model.





Conclusion: Compared to random forest, boosting model scores a little low on accuracy measure. Hence, if model accuracy is the target, we can opt for Random Forest for future batch predictions. If the requirement is for a parametric model, then we can opt of Logistic regression for future batch predicitions.

**FINAL QUESTION:**

-What would your predictions be regarding their choice of transport be for the following two employees?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Gender | Engineer | MBA | Work Exp | Salary | Distance | license |
| 25 | Male | 0 | 0 | 2 | 10 | 5 | 1 |
| 25 | Female | 1 | 0 | 2 | 10 | 5 | 0 |

**Answer:** Based on the EDA/s done, it can inferred that, since the age of employees is only 25 and being less experienced with low salary and less distance from home to office, they are more likely to prefer either 2wheeler (or) public transportation with strong possibilities for public transportation.