Predictive Modelling

Cars Case Study

Project 5

Kudakwashe Nyikadzino

1.) Exploratory Data Analysis

1.1) Data and Variable Analysis

1.1.1) Dataset structure

The dataset has 418 observations and 9 variables (Chart1.1). The target variable is Transport which has 3 levels namely 2Wheeler, Car and Public Transport. These are the modes of transport that employees use which we will predict.

The remaining 8 predictor variables include gender which is the only categorical variable with two levels i.e. Female and Male. Three variables have been recorded as integer type when they should be considered factor variables. These are Engineer, MBA and license which determine whether someone is an Engineer, has an MBA or a license. These will be converted to factor variables. Age, Work Experience, Salary and Distance are the remaining variables which are integers or numeric variables.

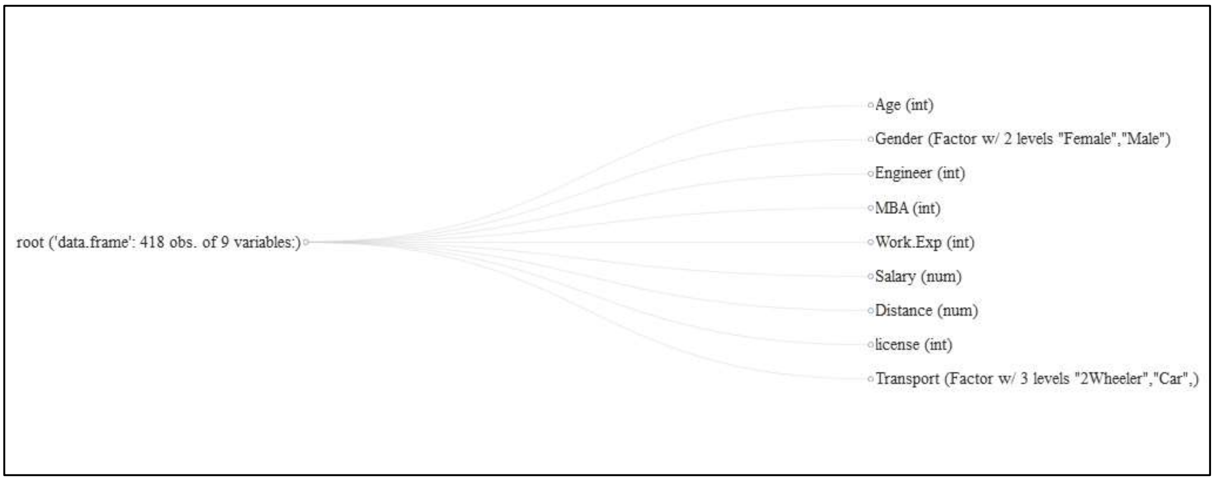

root ('data.frame': 418 obs. of 9 variables:)
Age (int)
Gender (Factor w/ 2 levels "Female","Male")
Engineer (int)
MBA (int)
Work.Exp (int)
Salary (num)
Distance (num)
license (int)
Transport (Factor w/ 3 levels "2Wheeler","Car",)

Chart1.1

1.1) Missing Variable Identification

The original dataset shows 78% of the columns are continuous and 22% are discrete. There are no missing columns but 99.8% of the rows are complete and no observations are missing.

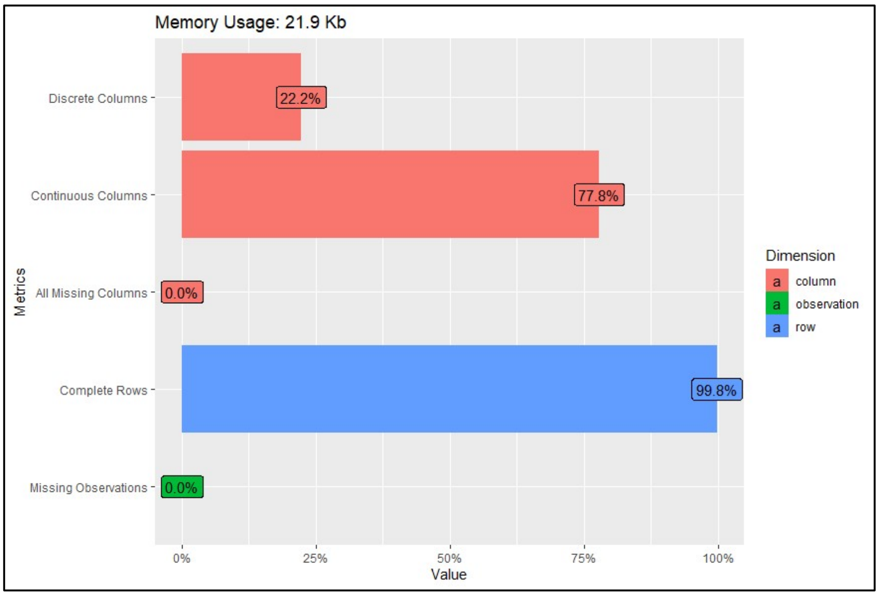


Chart1.2a

There is one missing value, and this is in the MBA variable. The missing value observations will be imputed instead of being removed. This ensures we maintain as much information from the dataset as possible. In addition to keeping the missing value, there is little risk of introducing significant bias as this is one value that is being imputed. Details of the imputation will be included as part of the data preparation process later.

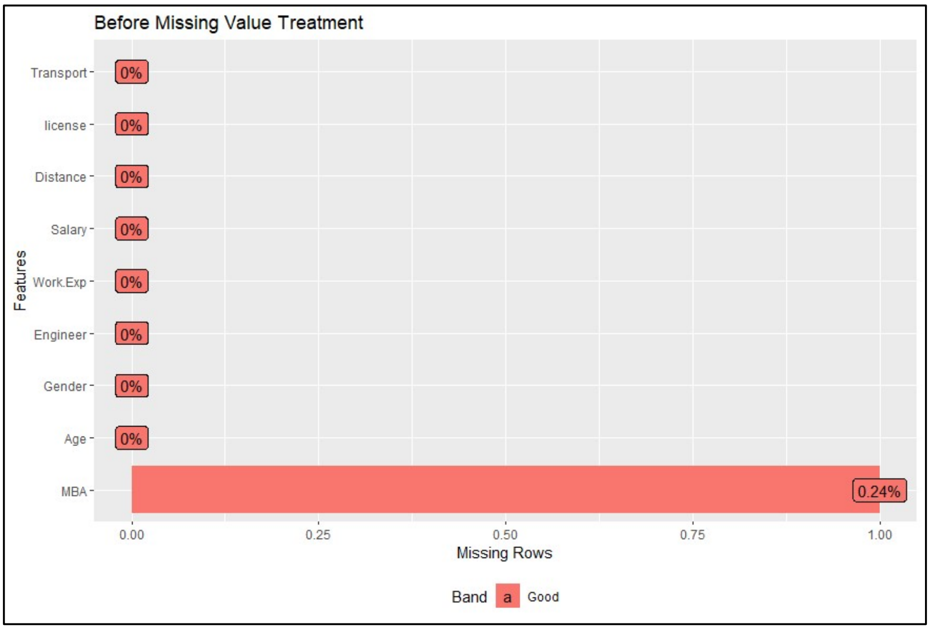


Chart1.2b

1.2) Univariate Analysis

1.2.1) Categorical Variables

1.2.1.1) Transport

Most employees use public transport which makes up 71.8% of the different transport modes. Those using 2 wheels as a mode of transport amount to about 19.9% and the least used mode of transport is car at 8.4%. Transport is our target variable and public transport is the dominant mode of transport.

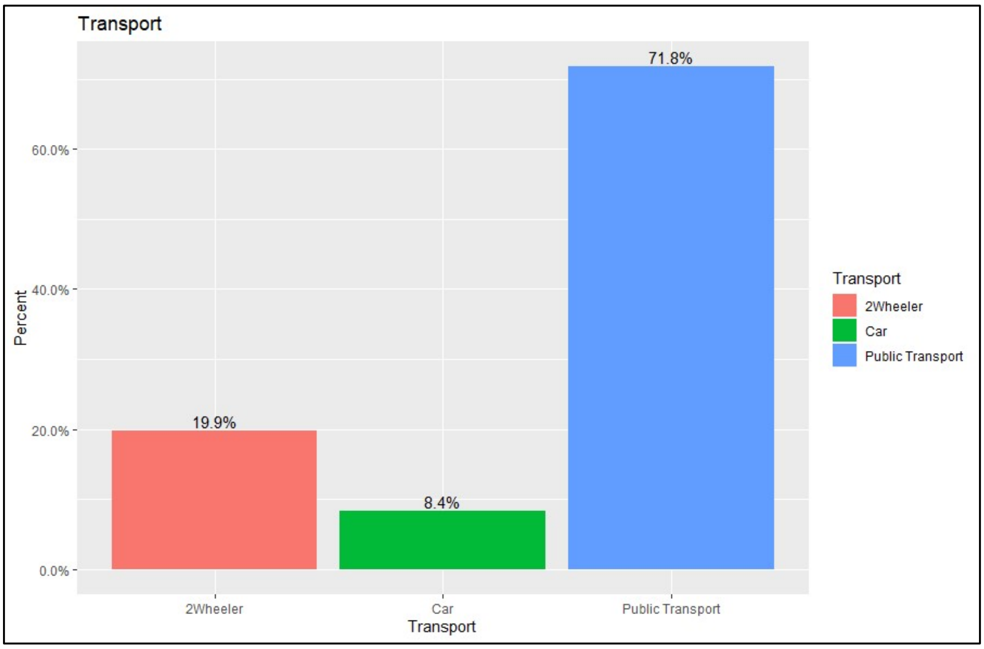


Chart1.2.1.1

1.2.1.2) Gender

Male employees make up 71.1% of all employees with female employees making up the remaining 28.9%. Males outnumber females by close to 3:1.

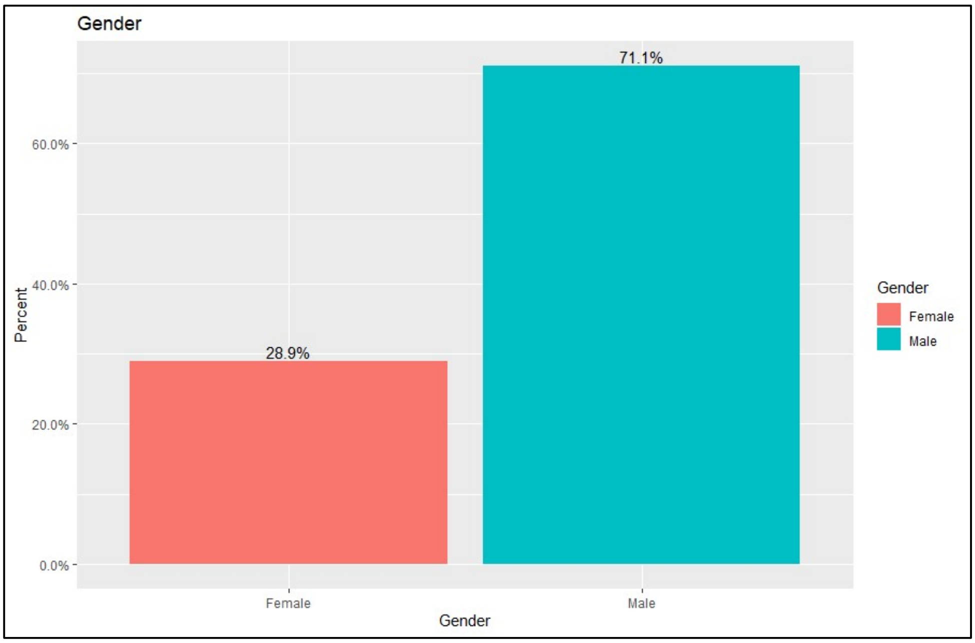


Chart1.2.1.2

1.2.1.3) Engineer

About three quarters of the employees are engineers (74.9%). Anecdotally, the engineering field is male dominated. This seems to be the same pattern being shown with males dominating the number of employees where most of the employees are engineers.

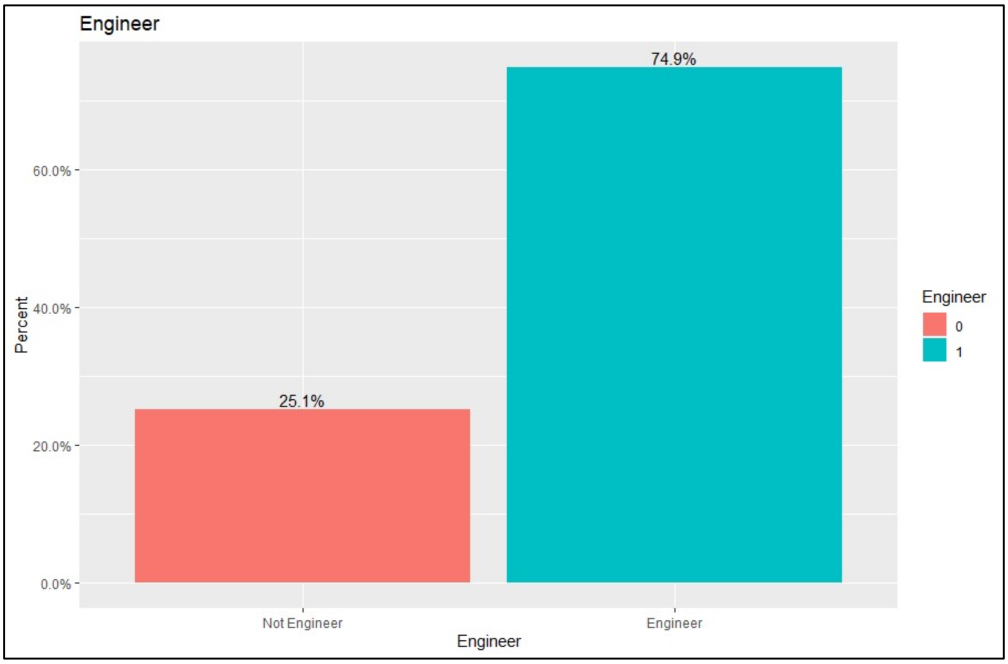


Chart1.2.1.3

1.2.1.4) MBA

Most employees do not hold an MBA at 73.9% and those that hold an MBA amounting to 26.1%. With engineers dominating the employee numbers, it could be expected that fewer would have a business qualification and potentially those in leadership having an MBA. This is what we are potentially seeing with a significant number not having an MBA.

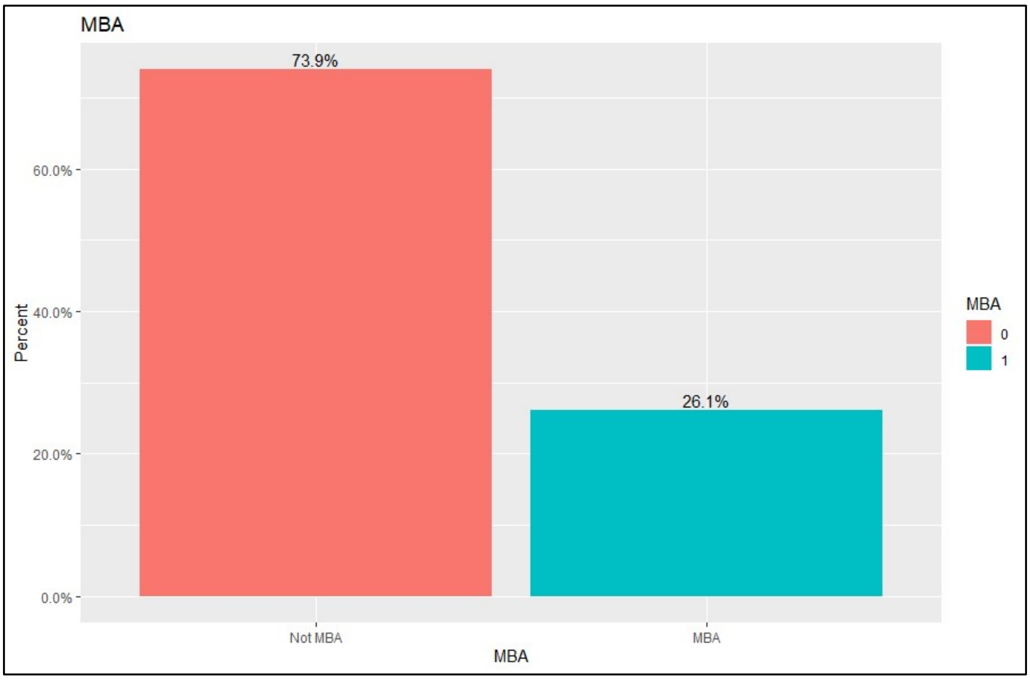


Chart1.2.1.4

1.2.1.5) License

Almost 80% (79.7%) of the employees do not have a license and only 20.3% have licenses. Considering the large number of employees using public transport, this could be a significant factor driving the mode of transport used. This will be part of the analysis to determine the important variables.

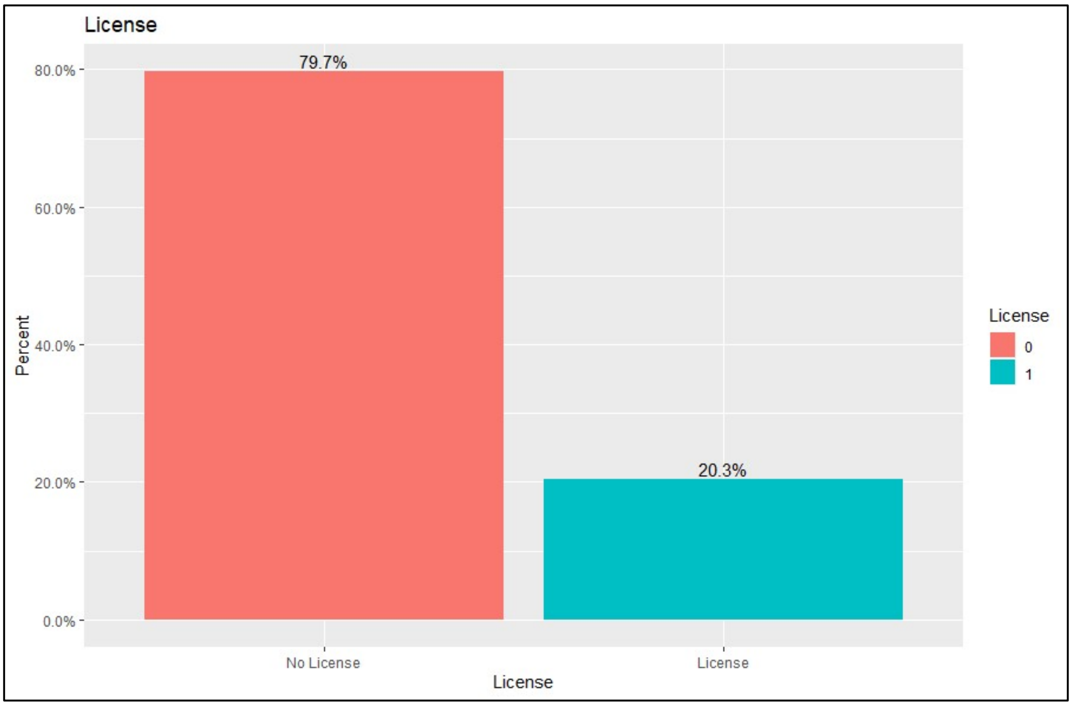


Chart1.2.1.5

1.2.2) Numeric Variables

1.2.2.1) Age

The distribution of age is right skewed with most outliers on the right whisker of the box plot (Chart 1.2.2a). There is one outlier on the left whisker which is the minimum age of all employees at 18 years (Table1.2.2b). The oldest employee in the dataset is 43 years old making the range 25 years. The average employee age is about 27 years. The standard deviation of Age is about 4 years. The QQ-Plot for Age also shows the presence of skewness, the outliers and the deviation from normality (Chart1.2.2c). It's the least dispersed variable around the mean as indicated by its coefficient of variation of 0.15.

1.2.2.2) Work Experience

About 57% of all employees have a work experience of 5 or less years with 29 just starting their careers with zero years of experience. This seems to be a young workforce and considering their average age is about 27 years. Considering that most employees are engineers, typical engineering studies are 4 to 5 years long. Most would finish their studies in their early or mid-twenties. Their average work experience of about 5.9 years and their average age around 27 years are closely related. The most experienced employee has 24 years which makes the range 24 years as well. The median age is 5 years which the histogram indicates as more than half of the employees have 5 years or less years of experience.

The distribution of work experience is not normal. There is a heavy right skew which is influenced by the presence of some outliers above the right whisker of the box plot. The right whisker is also longer than the left whisker with most of the distribution congested towards the minimum years of experience. The QQ-Plot also confirms the presence of outliers and deviation from the mean as points are distributed away from the diagonal on either end. Its standard deviation is 4.8 years. Its more disperse around its mean compare to Age with a coefficient of variation of 0.82 compared to 0.15.

1.2.2.3) Salary

The Salary earned by employees seems to be the most skewed and the furthest from a normal distribution. The presence of 52 outliers or 12.4% of employees on the right side of the distribution contributed to the heavy right skew. The QQ-Plot shows a significant deviation from normal compared to all the numeric variables. The average salary for an employee is about 15418 (we assume this is a monthly income) with a median salary of 13000. Considering the maximum income is 57500 given the average and median income, the existence of the skew is more apparent. The lowest salary is 6500 giving a wide range of 50500. Though its standard deviation of about 9.7 seems higher than other variables, its coefficient of variation at 0.627 is lower than that of Work Experience.

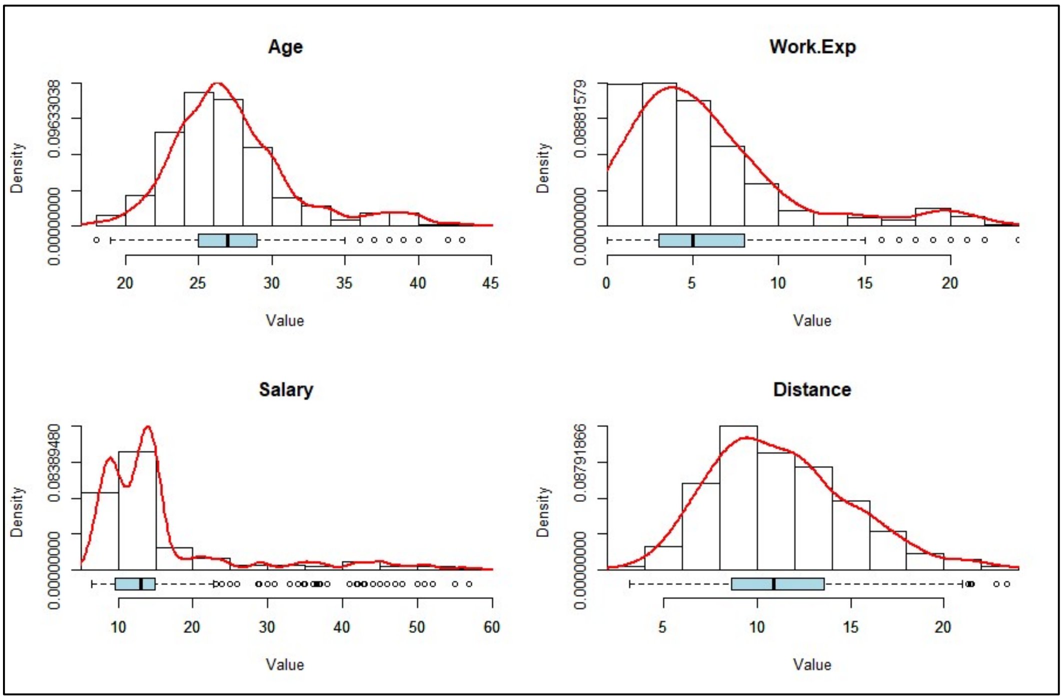
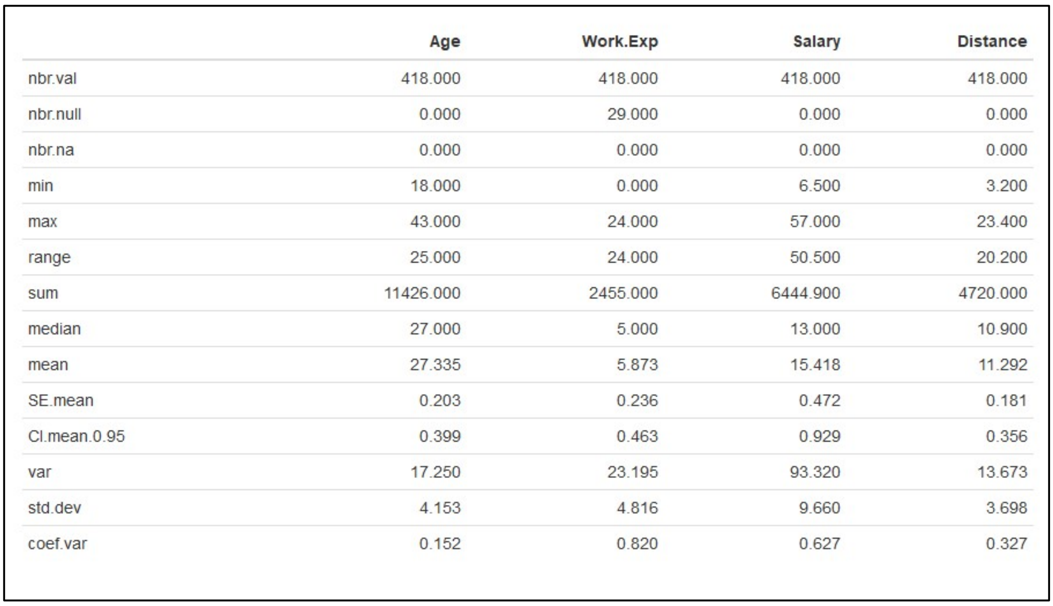


Chart1.2.2.a



	Age	Work.Exp	Salary	Distance
nbr.val	418.000	418.000	418.000	418.000
nbr.null	0.000	29.000	0.000	0.000
nbr.na	0.000	0.000	0.000	0.000
min	18.000	0.000	6.500	3.200
max	43.000	24.000	57.000	23.400
range	25.000	24.000	50.500	20.200
sum	11426.000	2455.000	6444.900	4720.000
median	27.000	5.000	13.000	10.900
mean	27.335	5.873	15.418	11.292
SE.mean	0.203	0.236	0.472	0.181
CI.mean.0.95	0.399	0.463	0.929	0.356
var	17.250	23.195	93.320	13.673
std.dev	4.153	4.816	9.660	3.698
coef.var	0.152	0.820	0.627	0.327

Table1.2.2.b

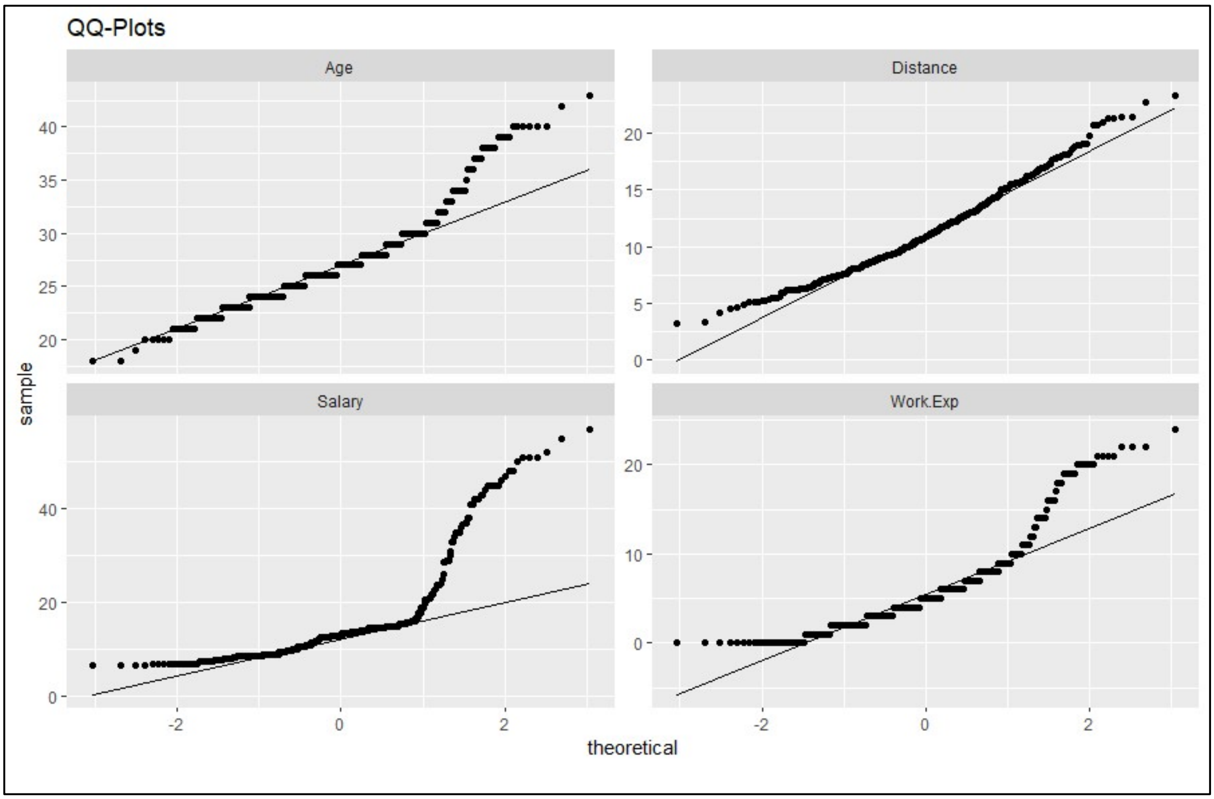


Chart1.2.2.c

1.2.2.4 Distance

The distance travelled by employees seems to be the closed to a normal distribution and contains the least outliers at 6. The QQ-Plot also shows most of the values falling on or close to the diagonal indicating a distribution close to normal. Employees on average travel 11.3 km to work with the least travelled distance work being 3.2 km and the most 23.4 km. the median distance travelled is less than the mean at 10.9 km. the right whisker of the box plot is longer than the left which indicates the right skew. The few outliers above the right whisker do not give as much of a skew compared to the other variables. The standard deviation is the lowest at about 3.7 km though its coefficient of variation is the second lowest at 0.327 after Age.

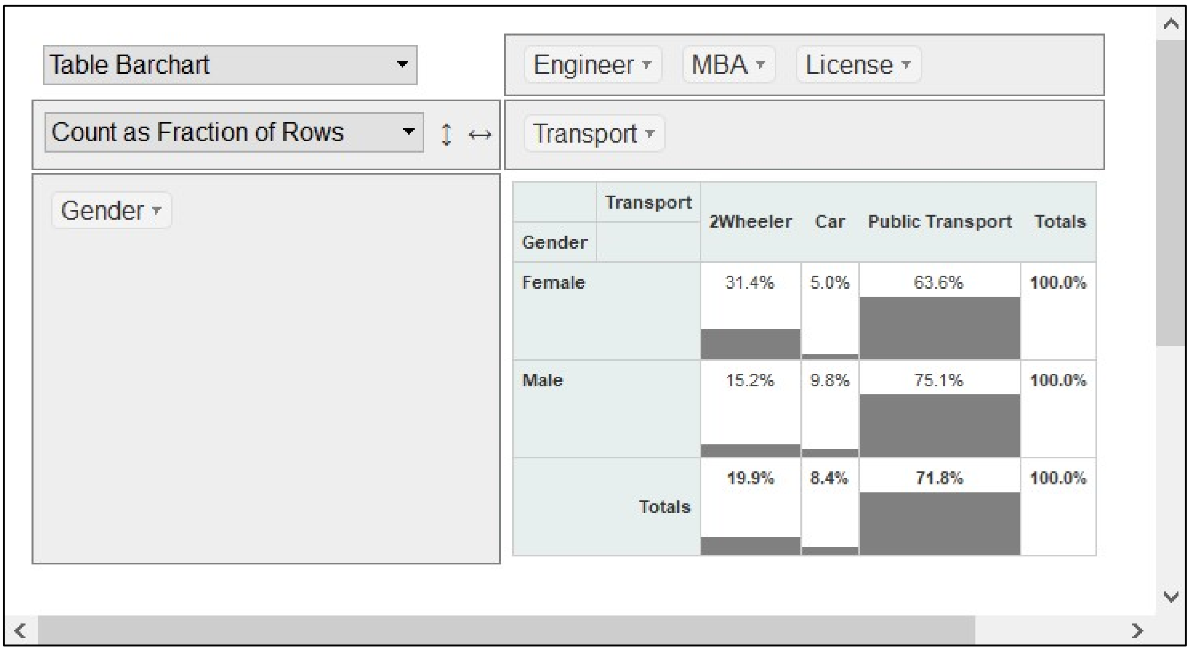
1.3) Bivariate Analysis

The bivariate analysis is based on clients that accepted the personal loan offer and other features that could help explain why they accepted the offer.

1.3.1) Bivariate Analysis – Categorical Variables

1.3.1.1) Transport and Gender

The relationship between the genders is not the same compared to Transport. Larger proportions of female employees use 2Wheeler as a mode of transport compared to male employees. A total of 31.4% of all female employees use 2Wheelers compared to 15.2% of male employees (Chart1.3.1.1a). More male employees use cars at 9.8% compared to 5.0% of female employees. The most popular mode of transport for male and female employees is public transport at 75.1% and 63.6% respectively. Since male employees make up 71.1% of all employees, this influences the overall proportions of the different modes of transport. For example, female employees using 2Wheeler transport make up 31.4% of all females but the overall average of all employees using 2Wheeler transport is only 19.9%. The differences in distributions of modes of transport across Gender is potentially useful in the prediction of the mode of transport compared to a situation where they are similar.

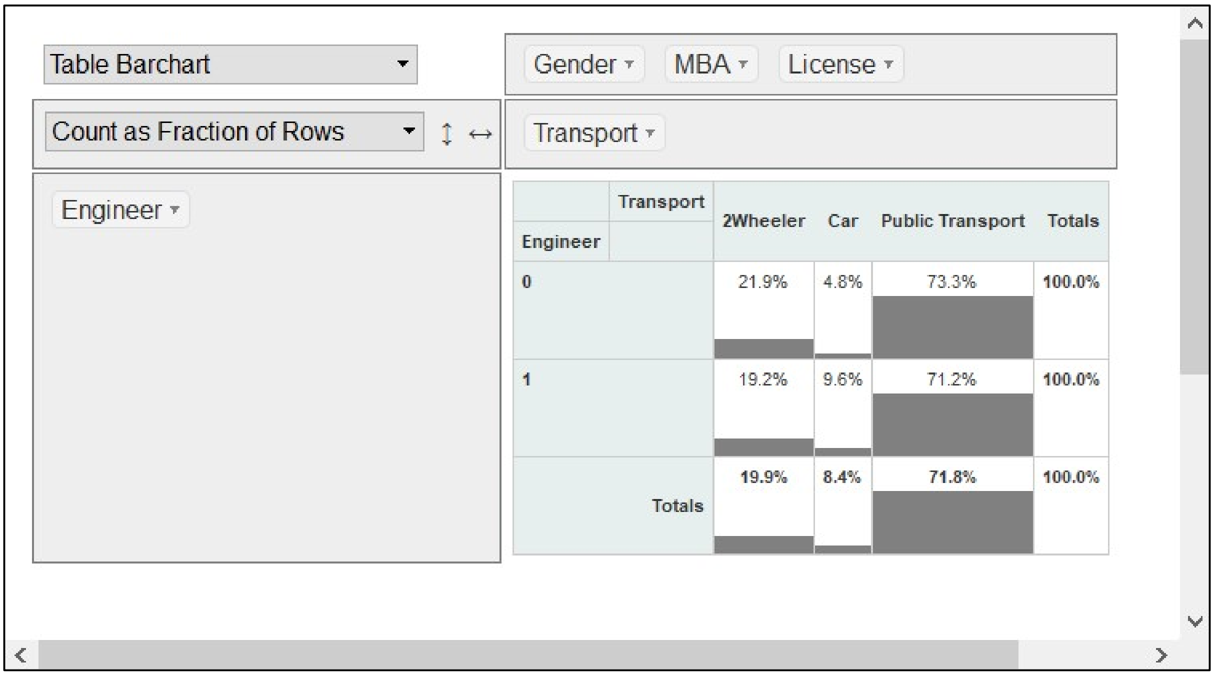


Gender \ Transport	2Wheeler	Car	Public Transport	Totals
Female	31.4%	5.0%	63.6%	100.0%
Male	15.2%	9.8%	75.1%	100.0%
Totals	19.9%	8.4%	71.8%	100.0%

Chart1.3.1.1

1.3.1.2) Transport vs Engineer

The pattern of modes of transport used by engineers and non-engineers is similar. The order of the most used mode of transport is public transport, 2Wheeler and cars across the two groups. Slightly more non-engineers use public transport at 73.3% compared 71.2% engineers. The difference between car usage is the largest across the different modes of transport with 4.8% of non-engineers using cars compared to 9.6% engineers. Since engineers dominate the workforce, the overall proportions of the different transport modes are like those of engineers. This factor on its own may mean it would be difficult to distinguish the mode of transport an engineer or non-engineer is likely to use.

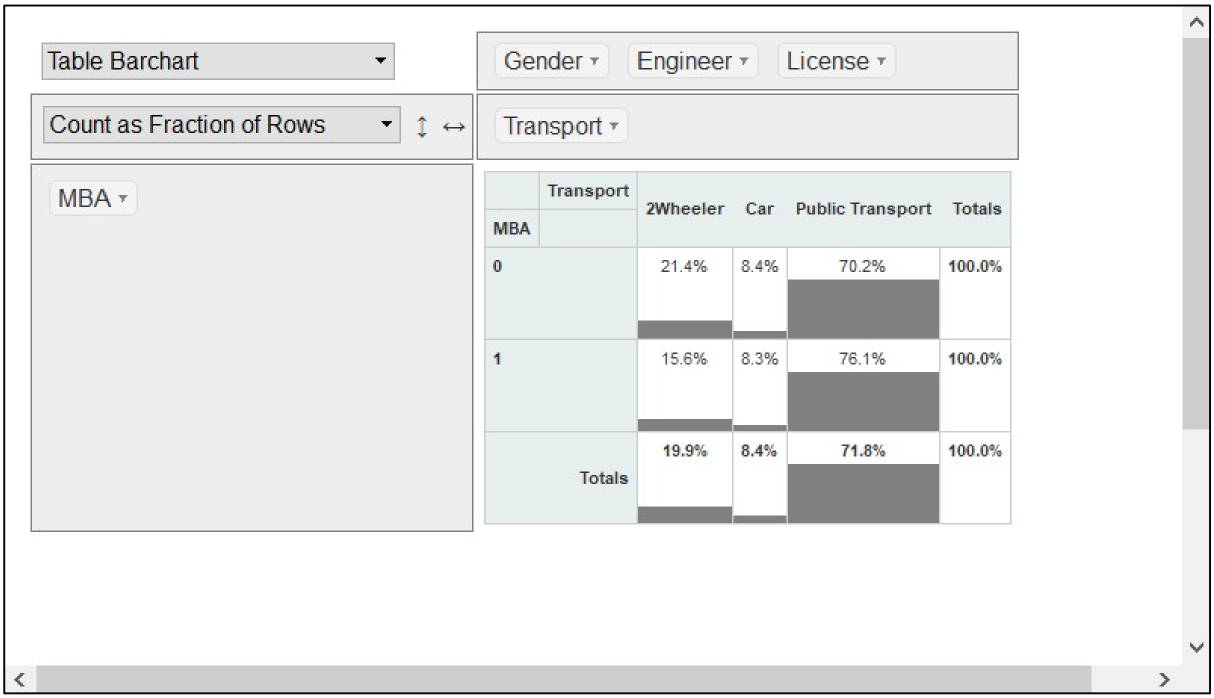


Engineer \ Transport	2Wheeler	Car	Public Transport	Totals
0	21.9%	4.8%	73.3%	100.0%
1	19.2%	9.6%	71.2%	100.0%
Totals	19.9%	8.4%	71.8%	100.0%

Chart1.3.1.2

1.3.1.3) Transport vs MBA

Employees using cars are the same regardless of whether they have an MBA or not at 8.3% and 8.4% respectively. The most popular mode of transport is public transport with 70.1% for those without and 76.1% for those with an MBA. More employees without an MBA use 2Wheelers relative to those with an MBA at 21.4% and 15.6% respectively. Most employees do not hold an MBA and their influence is evident on the overall proportions.



Transport / MBA	2Wheeler	Car	Public Transport	Totals
0	21.4%	8.4%	70.2%	100.0%
1	15.6%	8.3%	76.1%	100.0%
Totals	19.9%	8.4%	71.8%	100.0%

Chart1.3.1.3

1.3.2) Bivariate Analysis – Numeric Variables

Visualising the relationship between our predictor and outcome variable using box plots and density plots can bring to the surface variables that are potentially important in prediction. Visual inspections alone may not be the basis of confirming the relationships between variables.

1.3.2.1) Transport vs Salary

The average salaries of employees using public transport and 2Wheelers is close with that of those using cars significantly distant from the two (Chart1.3.2.1). There are outliers on the right side of both the 2Wheeler and public transport distribution in contrast to cars where the outliers are lower in the distribution. The density plot also shows a clear break in distribution for cars and a different pattern compared to the other two with similar distributions both in kurtosis and skewness. The difference between cars and the other two distributions is an indicator that salary can help in predicting the mode of transport (Chart1.3.2.2). However, the similarity in the distribution of 2Wheeler and public transport may make it difficult to distinguish between the two in determining the importance of Salary as a predictor of Transport.

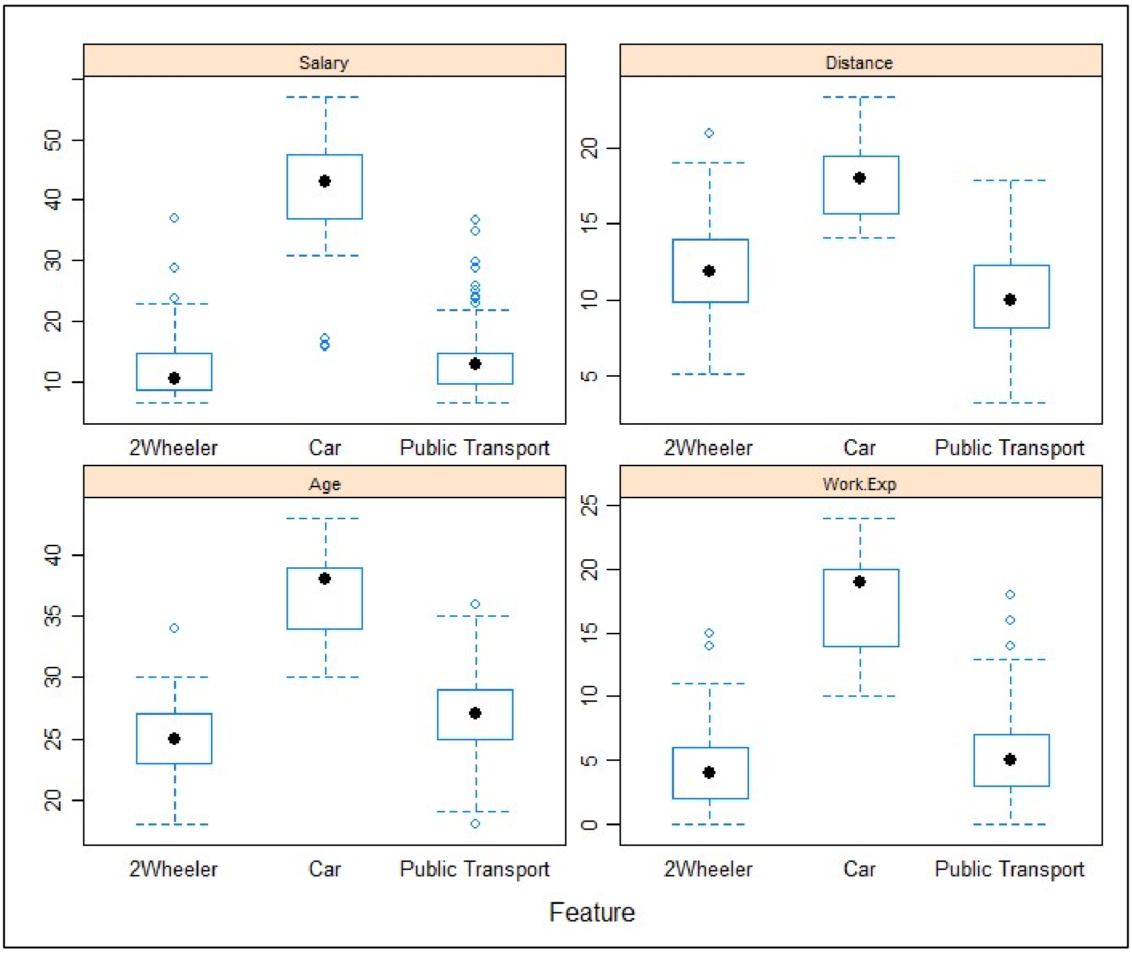


Chart1.3.2.1

1.3.2.2) Transport vs Distance

Distance seems to show differences between the three modes of transport, though these differences may not extreme. Employees using cars seems to be further from work travelling longer distance on average than 2Wheeler and public transport users. Those using public transport on average stay nearest to work though their distribution is the widest of the three. The patterns of distribution of the three modes of transport have a similar kurtosis though their placement is different. Bar the outlier in the 2Wheeler distribution, the upper ends of the 2Wheeler and public transport users is close to the average distance travelled by those using cars. The maximum distance travelled by employees using public transport is 17.9 km, the average distance for those using cars is also 17.9 km and the maximum for 2Wheeler excluding the outlier is 19.1 km. there is potentially a distance that becomes a decision boundary where using a certain type of transport becomes feasible. In addition, the differences in placement of the distributions make Distance a strong candidate in determining the mode of transport used.

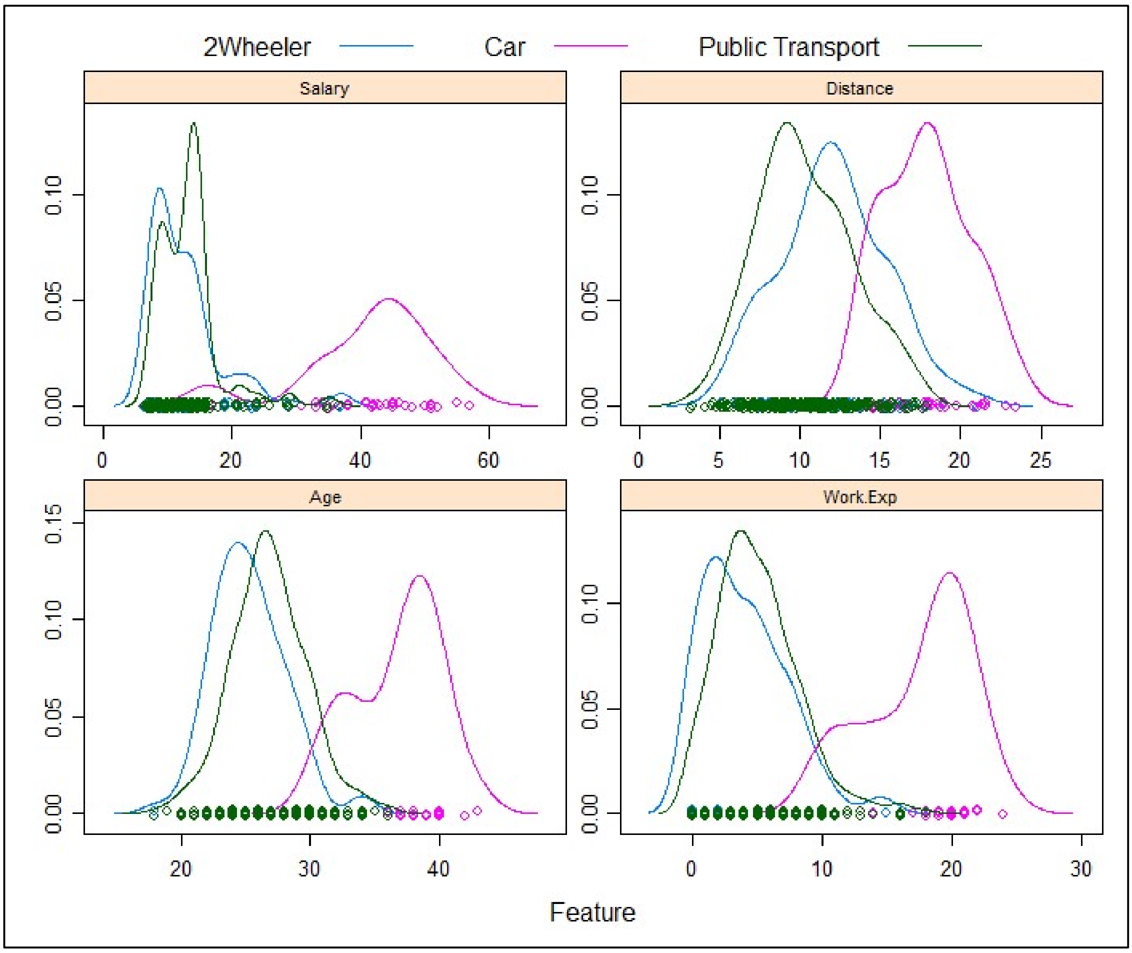


Chart1.3.2.2

1.3.2.3) Transport vs Age

Older employees seem to use cars more than the other modes of transport. The high correlation between Age, Work experience and Salary possibly indicates that the higher income enables them to afford cars and the expense associated with running them. Potentially the older employees may no longer be able to use modes of transport like bicycles and motorbikes that require more exertion. The distributions between 2Wheeler and public transport are similar though different in their placement with average ages of 25.3 and 26.8 years respectively. Those using cars have a distribution different from the other two in both placement and kurtosis with an average age of 36.7 years. Some employees using public transport or 2Wheelers are outlier and those using cars having no outliers. The difference between cars and the two variable given age is helpful in possibly predicting the mode of transport. If we look at some of the similarities between public transport and 2Wheelers, it makes a bit more difficult to distinguish the two from each other. This may decrease the overall strength of age being significant predictor of the mode of transport used.

1.3.2.4) Transport vs Work Experience

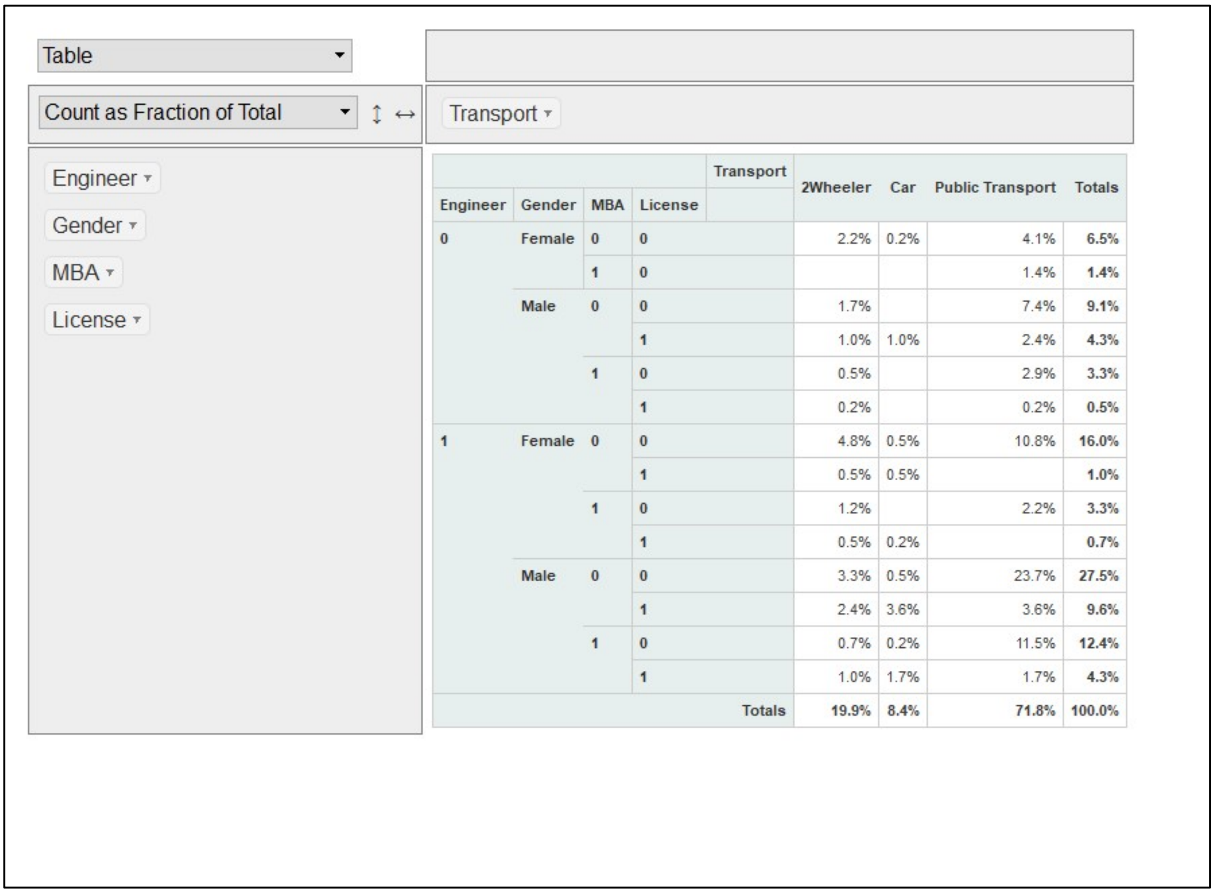
Work experience and Age have similar patterns and distributions. Like age, work experience shows those using 2Wheeler and public transport having similar density plots with a light offset between the two. Those using cars have a very different from the other two just like Age with average work experience of 17.5 years. There are outliers for both 2Wheeler and public transport with average experience of 4 and 5 years respectively.

The differences years for 2Wheeler and public transport is 1.5 years in respect of Age and 1 year in respect of Work experience. The difference between public transport and cars for

Age is about 10 years and about 12 years for work experience. This highlights the similarity between these two variables and shown by their high correlation of 92%. Meaning using the two in the same model may not be helpful in prediction particularly in situations where independence is assumed. Similarly, using work experience maybe helpful in predicting the mode of transport between cars and the other two modes. However, the similarity between 2Wheeler and public transport may diminish this predictive power.

1.3.3) Overall view across all variables

The table below shows a view across all categorical variables as a proportion of the total number.



				Transport	2Wheeler	Car	Public Transport	Totals
Engineer	Gender	MBA	License					
0	Female	0	0		2.2%	0.2%	4.1%	**6.5%**
		1	0				1.4%	**1.4%**
	Male	0	0		1.7%		7.4%	**9.1%**
			1		1.0%	1.0%	2.4%	**4.3%**
		1	0		0.5%		2.9%	**3.3%**
			1		0.2%		0.2%	**0.5%**
1	Female	0	0		4.8%	0.5%	10.8%	**16.0%**
			1		0.5%	0.5%		**1.0%**
		1	0		1.2%		2.2%	**3.3%**
			1		0.5%	0.2%		**0.7%**
	Male	0	0		3.3%	0.5%	23.7%	**27.5%**
			1		2.4%	3.6%	3.6%	**9.6%**
		1	0		0.7%	0.2%	11.5%	**12.4%**
			1		1.0%	1.7%	1.7%	**4.3%**
				Totals	**19.9%**	**8.4%**	**71.8%**	**100.0%**

Chart1.3.3

2.) Insights from EDA

- Most employees are male engineers without an MBA or license using public transport.
- No female employee with a license uses public transport.
- The numeric variables all have outliers and will require treatment to build robust models.
- From the observations, knowing whether someone is an engineer or not may not be useful in predicting the mode of transport they are likely to use.
- Gender seems to be the most important categorical variable in determining the mode of transport used by employees.
- Employees with higher salaries tend to use cars.
- Older employees tend to use cars.
- Employees with higher work experience also tend to use cars.
- Age, Work experience and Salary have high pairwise correlations and as a result having all three variables in a model may provide redundant information.
- Age and weight have almost indistinguishable distributions and with their high correlations, information from one is very similar to the other.
- Distance seems to be the most important numeric variable in determining the mode of transport used.

3.) Most challenging aspect of the problem

3.1) Multicollinearity

Three of the four numeric variables are highly correlated. These variables are Age, Work experience and Salary (Chart3.1). Distance has the lowest correlation with all the other three variables. This presence a problem of multicollinearity just from inspection of these variables. High correlations make it difficult to determine the true impact of an independent variable on a dependent variable. This is a major issue in situations where there is an assumption of independence between the variables for example regression.

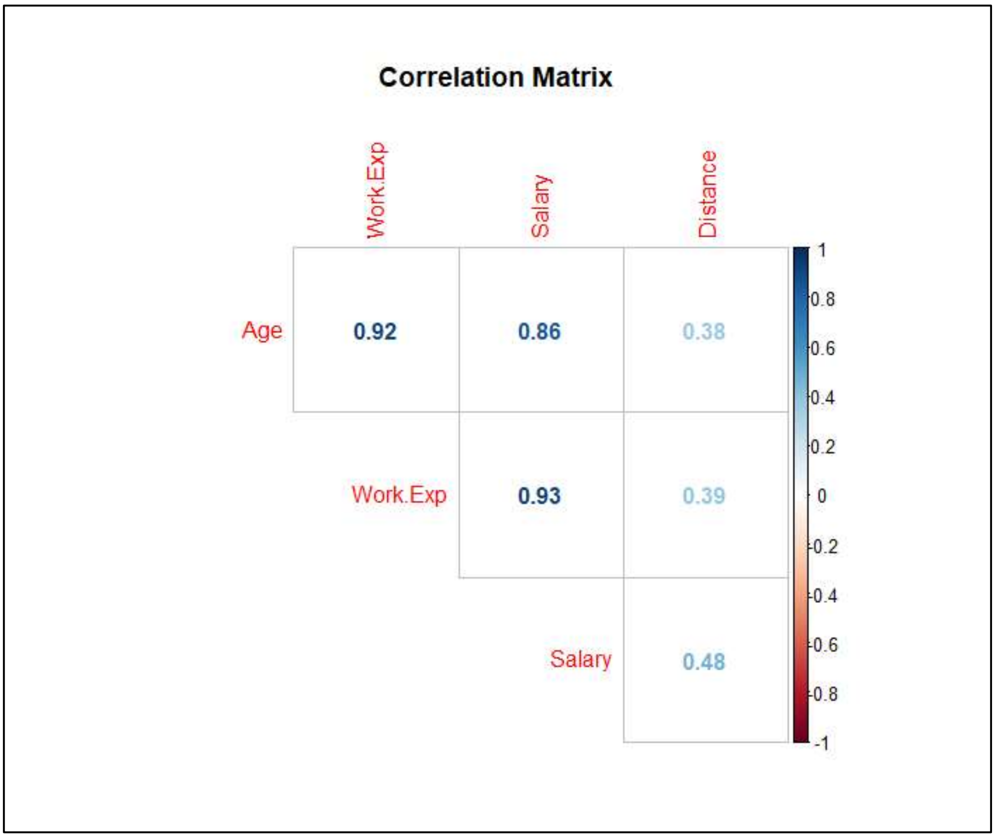
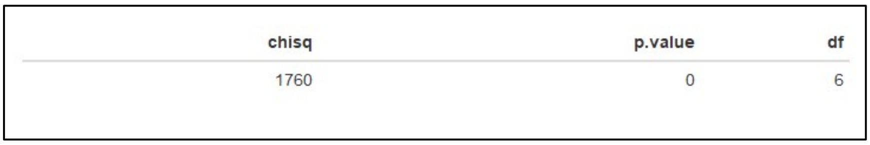


Chart3.1

3.2) How to deal with the problem

3.2.1) Bartlett's Test for Sphericity

The test compares a correlation matrix with an identity matrix to see if independent variable correlations are significantly different from zero. The null hypothesis for the test that independent variables are not correlated. Where we reject the null hypothesis, the test is looking to see if any of the numeric variables can be combined to fewer factors due the redundancy that results from correlations. The results of the test show a p-value that practically equal to zero (Table3.2.1).



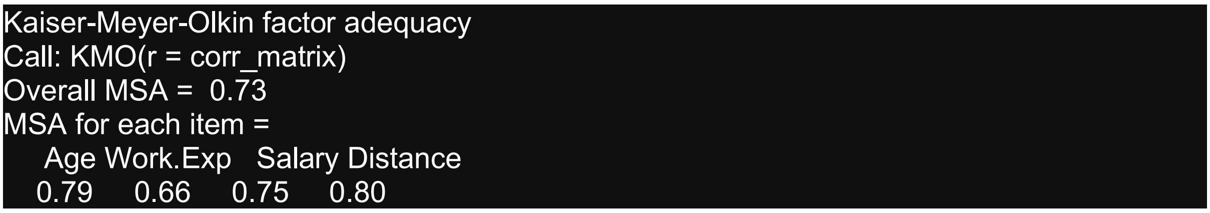
chisq	p.value	df
1760	0	6

Table3.2.1

We can reject the null hypothesis that the independent variables are not correlated. This means at a level of significance of 5%, we can conclude the variables are correlated. This makes the variables candidates for dimensionality reduction in our case with factor analysis.

3.2.2) KMO – Test: Test for sampling adequacy

The KMO Test measures the proportion of variance among the variables that may be common variance. This measure is on individual variables as well a group of variables and varies from 0 - 1. This allows us to see whether we can proceed with dimensionality reduction like factor analysis. A general guide is if the value is less than 0.5, then it may not be useful to proceed with factor analysis, but some judgement is needed here. The higher the value the better. The results of the test below show that the variables individually and as group at 0.73 are candidates for factor analysis (Table3.2.2).



```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = corr_matrix)
Overall MSA =  0.73
MSA for each item = 
     Age Work.Exp   Salary Distance 
    0.79     0.66     0.75     0.80 
```

Table3.2.2

4.) Data Preparation

4.1) Missing value treatment

There is one value missing in the entire dataset which was for the MBA variable. The decision was made to keep as much information as possible in the data, so the value was imputed. The imputation was made using the MICE package to complete the dataset.

4.2) Outlier treatment

Outliers are prevalent in the dataset with all numeric variables containing outliers. The method used uses imputation and the MICE package. All outliers are replaced by NAs and then the missing values are completed by the random forest method in the MICE package. Each column with missing values is used as the outcome variable with the other values as predictors. We start with the column with the least number of outliers till we get to the column with the most to build a much more robust dataset. The order is as follows Distance, Age, Work Experience and Salary. After the outlier treatment, only two variables i.e. Work experience and Salary have outliers. The number of outliers for Work experience have come down from 29 to 21 and the represent about 5% of values (Chart 4.2). Those from Salary have decreased from 52 to 4 which is about 1% of the values. The treatment of outliers is satisfactory, and we can proceed with the dataset. The decision to use this approach has an advantage of ensuring a complete dataset before proceeding to the next imputation.

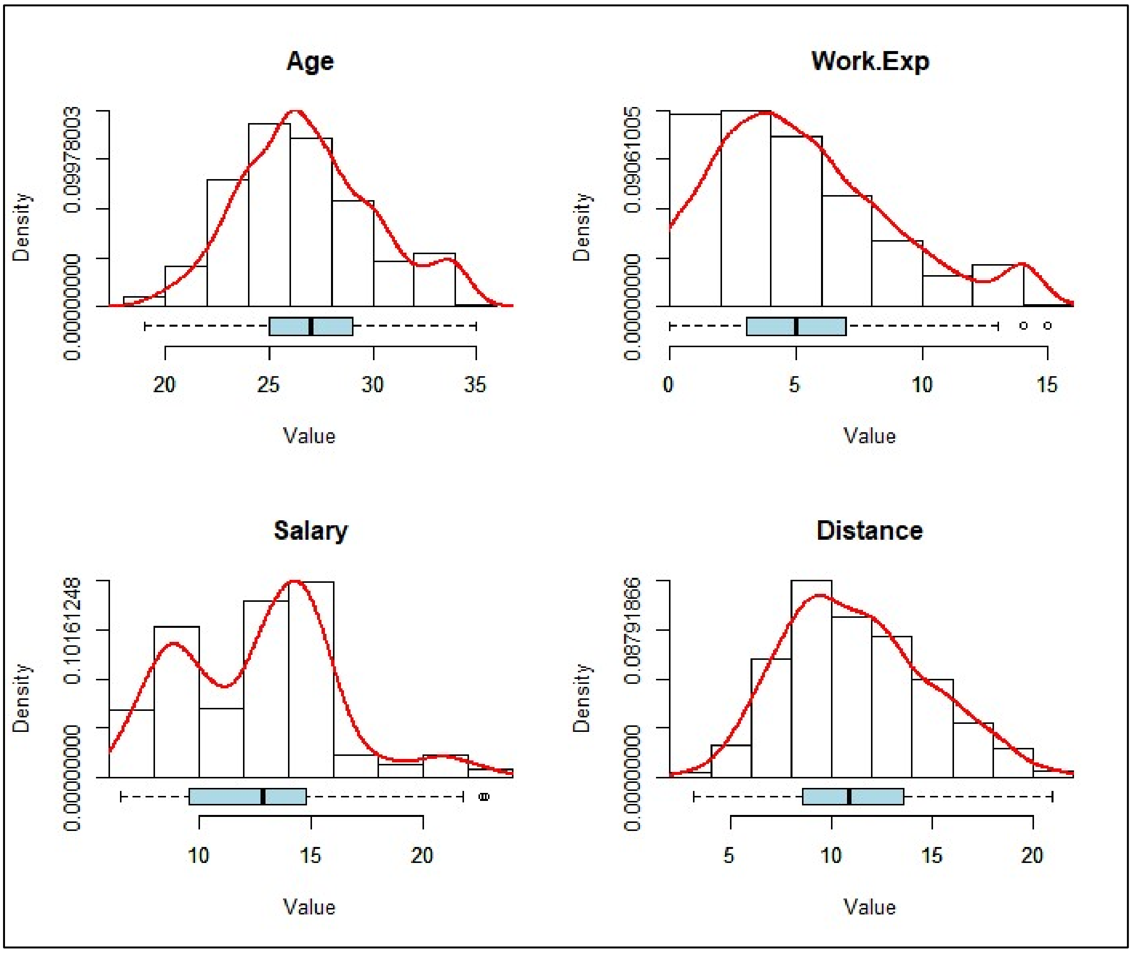


Chart4.2

4.3) Scaling

The units of measurement are not the same so scaling has been applied. This ensures that values for all numeric variables are between 0 and 1 shown below.

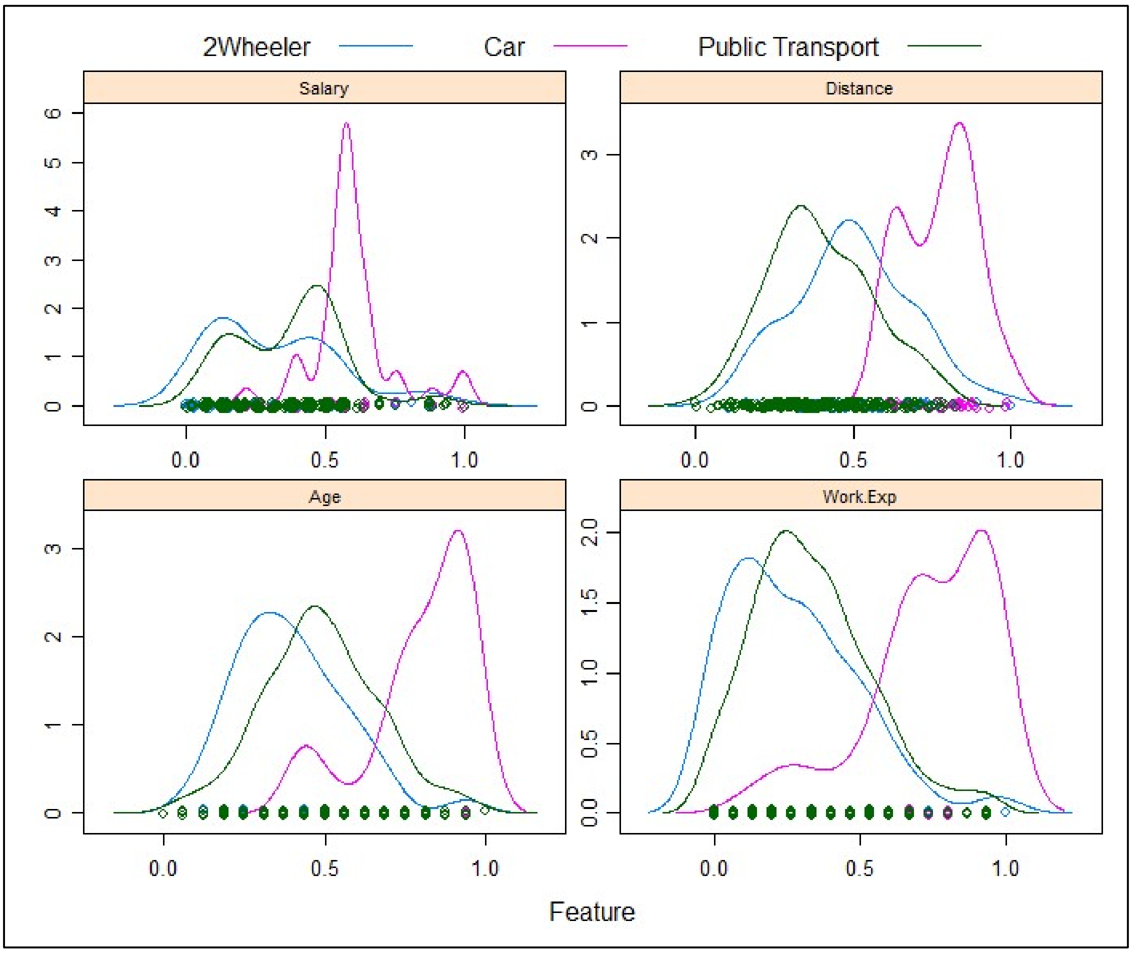


Chart4.3

4.4) Factor Analysis

We are going to use factor analysis to deal with the problem of multicollinearity identified earlier. This allows us to reduce the number of columns into common factors helping to resolve the multicollinearity issue. The Scree Plot and the Eigen values help in identifying the number of factors that will be used in the analysis. Based on the Scree plot if values above one is used then only one factor will be used. We can see the second factor is close to one with an eigen value of 0.927. The decision in this case will be to use two factors.

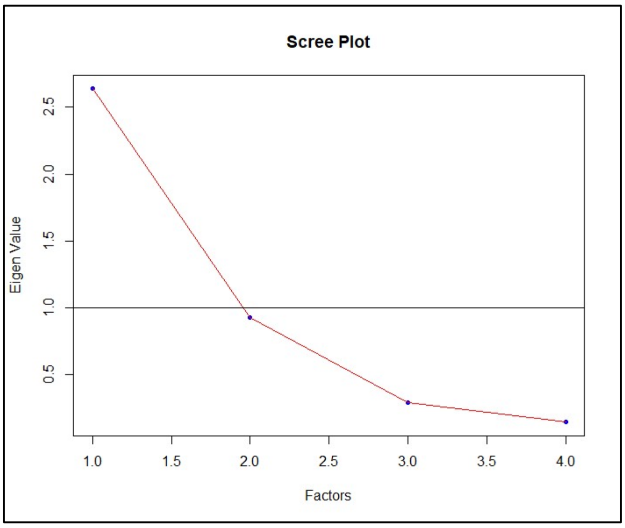


Chart4.4a

Rotation is used and the oblimin method is used because we assume correlation exists between some of the variables. The diagram below shows that Age, Work experience and Salary are combined into one factor as we saw earlier because of their high correlations. Distance is considered a separate factor on its own. The plot of the factor analysis below.

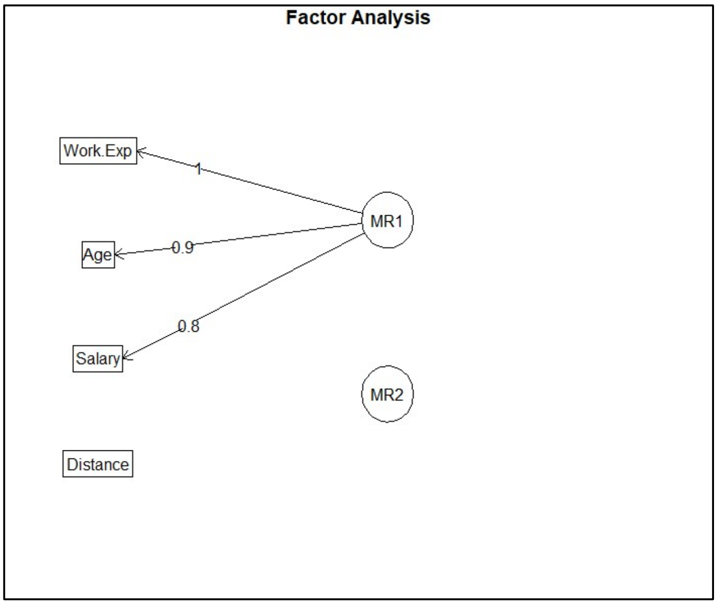


Chart4.4b

We proceed to remove the three variables and replace them with the factor scores of the first factor while keeping Distance.

4.5) Splitting data into train and test

The data has been split into training and test into 75% and 25% respectively. The split preserves the proportions of the Transport variables in the two datasets. Splitting the data enables us to evaluate how well the model performs.

4.6) One-hot encoding dummy variables

This helps us convert categorical variables into binary variables to allow us to use all variables with different machine learning algorithms.

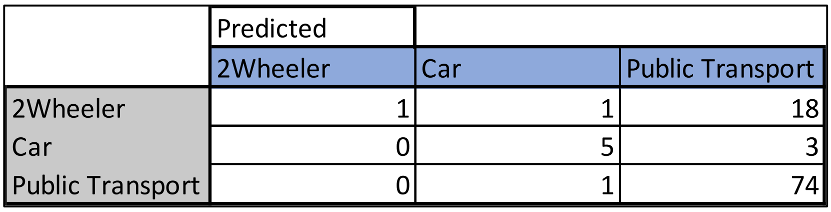
5.) Models and Performance

5.1) KNN Model

The model that was used has a k = 17 and an accuracy of 78.7%.

```
k-Nearest Neighbors

315 samples
 10 predictor
  3 classes: '2Wheeler', 'Car', 'Public Transport'

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 210, 210, 210
Resampling results across tuning parameters:

  k   Accuracy  Kappa
   5  0.740     0.327
   7  0.733     0.319
   9  0.740     0.301
  11  0.768     0.349
  13  0.781     0.375
  15  0.781     0.359
  17  0.787     0.378
  19  0.775     0.329
  21  0.771     0.317
  23  0.768     0.304

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 17.
```

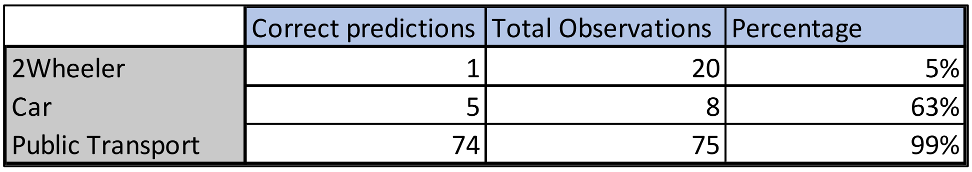
When this model is applied to the test data set, the output shows an accuracy of 77.7% which is (1+5+74)/103. The performance is about 1% lower than that of the training dataset which shows when applied to potentially new data it may do well.



	Predicted		
	2Wheeler	Car	Public Transport
2Wheeler	1	1	18
Car	0	5	3
Public Transport	0	1	74

KNN Model Evaluation

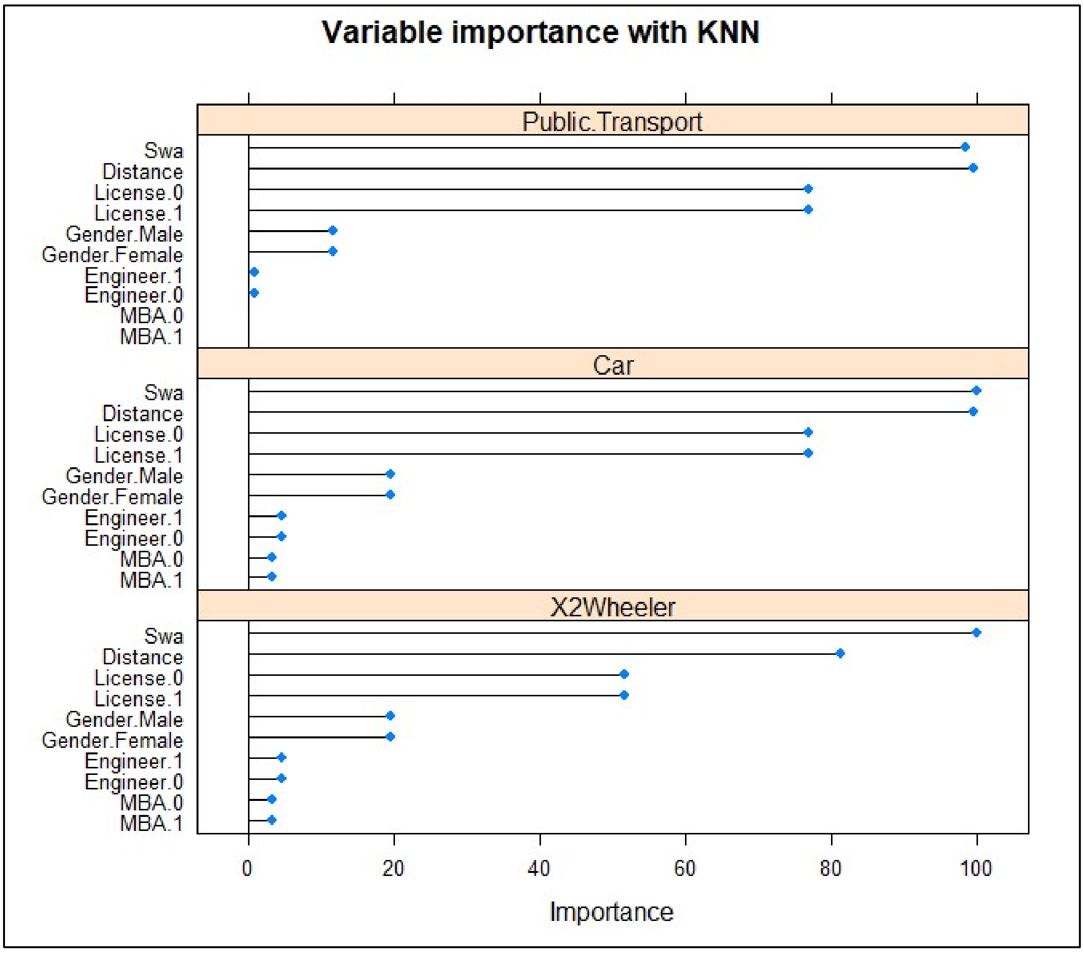
The model performs very well when predicting if an employee uses public transport. It accurately predicted 99% of all employees that use public transport. Its does poorly when predicting employees that use 2Wheelers. It inaccurately predicted that 18 of the 20 employees using 2Wheelers used public transport. It only accurately predicted 5% which is very low. The performance of the car predictions was not very high at 63%.



	Correct predictions	Total Observations	Percentage
2Wheeler	1	20	5%
Car	5	8	63%
Public Transport	74	75	99%

The overall performance at 77.7% accuracy on the test dataset may be improved using other algorithms which we will use.

From KNN, the most important variable to predict whether an employee will use a car is based on the factor comprised of Salary, Age and Work Experience (Swa). Distance is the second most important variable.



5.2) Naïve Bayes

The Naïve Bayes model correct predicted 71.8% of the different transport modes used by employees.

	Predicted		
	2Wheeler	Car	Public Transport
2Wheeler	4	1	15
Car	1	6	1
Public Transport	10	1	64

Naïve Bayes Model Evaluation

The model correctly predicted 2Wheelers 20% of the time and this low although it is an improvement from the KNN model. There was also an improvement in the accuracy when predicting employees that use cars to 75% from 63% in the KNN model. There was a decrease in accuracy for the public transport users from 99% to 85% though it is still a high level of accuracy. The Naïve Bayes seems to be generalising better than the KNN model.

	Correct predictions	Total Observations	Percentage
2Wheeler	4	20	20%
Car	6	8	75%
Public Transport	64	75	85%

Appropriateness of Naïve Bayes

The algorithm assumes independence between variables holds. In this case, for the numerical variables we have multicollinearity between the variables through factor analysis. This has helped us with variables that are highly dependent on each other like Age and Work experience where an increase in work experience cannot occur without an increase in age.

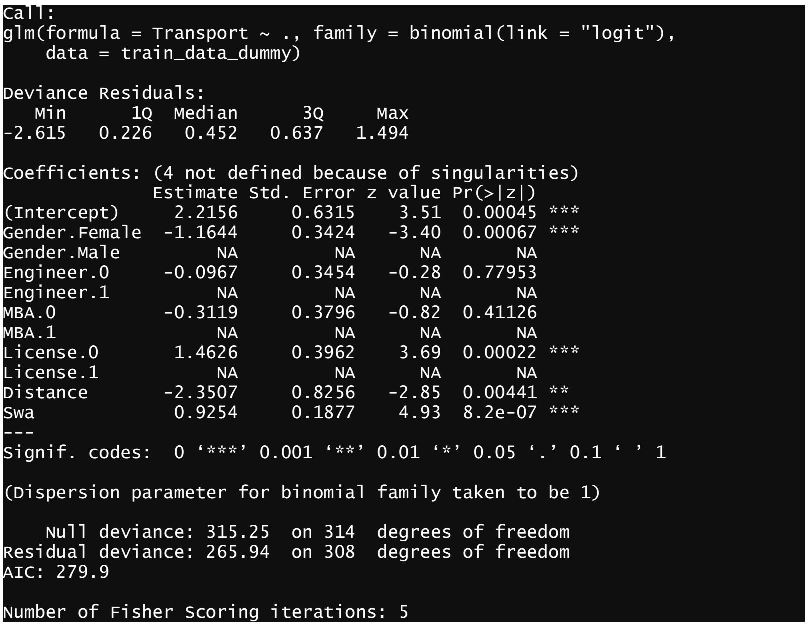
Naïve Bayes performs better for categorical variables compared to numeric variables. In this case we have used a dataset that contains numeric variables and as a result may not produce the best performance. In addition, the numeric variables are assumed to follow a normal distribution which not the case with some of our variables based on observations made in exploratory data analysis.

5.3) Logistic regression model

We have used probabilities in determining whether an observation is classified as true or false for each mode of transport. When the probability is greater than 0.5, then it is classified as true.

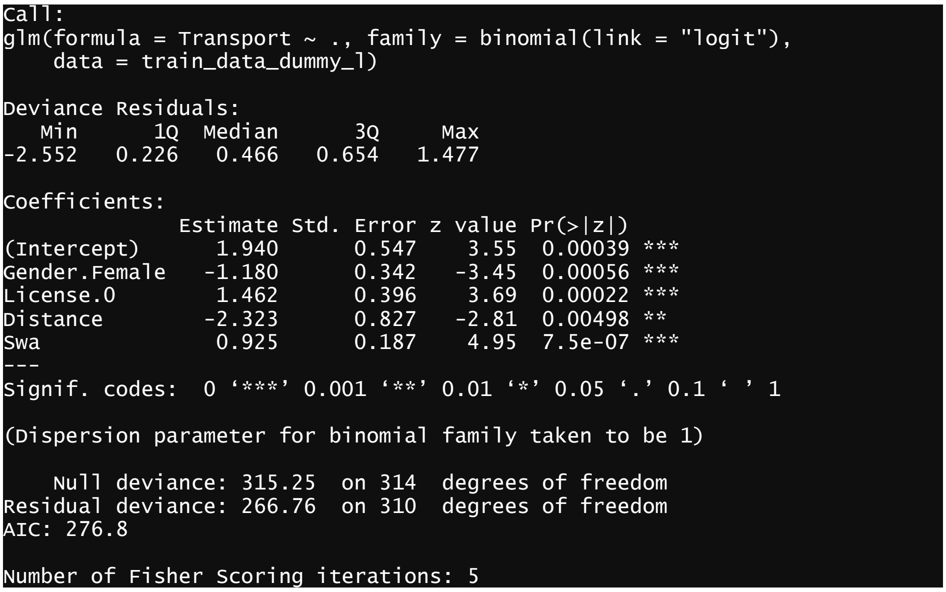
We determine which variables are significant and remove the non-significant variables in the model and run the model again. To refine the model, the following variables have been removed:

- Gender.Male
- Engineer.0
- Engineer.1
- MBA.0
- MBA.1
- License.1



```
Call:
glm(formula = Transport ~ ., family = binomial(link = "logit"), 
    data = train_data_dummy)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
-2.615   0.226   0.452   0.637   1.494  

Coefficients: (4 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)     2.2156     0.6315    3.51  0.00045 ***
Gender.Female  -1.1644     0.3424   -3.40  0.00067 ***
Gender.Male         NA         NA      NA       NA    
Engineer.0     -0.0967     0.3454   -0.28  0.77953    
Engineer.1          NA         NA      NA       NA    
MBA.0          -0.3119     0.3796   -0.82  0.41126    
MBA.1               NA         NA      NA       NA    
License.0       1.4626     0.3962    3.69  0.00022 ***
License.1           NA         NA      NA       NA    
Distance       -2.3507     0.8256   -2.85  0.00441 ** 
Swa             0.9254     0.1877    4.93  8.2e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 315.25  on 314  degrees of freedom
Residual deviance: 265.94  on 308  degrees of freedom
AIC: 279.9

Number of Fisher Scoring iterations: 5
```

After the second iteration of the model with fewer variables is below. In this instance, Salary, Age and Work experience as the factor Swa is the most important variable with the smallest Pr(>|z|).



```
Call:
glm(formula = Transport ~ ., family = binomial(link = "logit"), 
    data = train_data_dummy_l)

Deviance Residuals: 
   Min       1Q   Median       3Q      Max  
-2.552    0.226    0.466    0.654    1.477  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)      1.940      0.547    3.55  0.00039 ***
Gender.Female   -1.180      0.342   -3.45  0.00056 ***
License.0        1.462      0.396    3.69  0.00022 ***
Distance        -2.323      0.827   -2.81  0.00498 ** 
Swa              0.925      0.187    4.95  7.5e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 315.25  on 314  degrees of freedom
Residual deviance: 266.76  on 310  degrees of freedom
AIC: 276.8

Number of Fisher Scoring iterations: 5
```

Logistic regression model evaluation

The model has an accuracy of 93.2% ((17+5+74)/103) which is the best accuracy we have seen compared to KNN and Naïve Bayes. It also predicted employees using 2Wheelers better than the other models at 80% accuracy and has matched KNN for predicting public transport and car usage.

	Predicted			
	FALSE	TRUE	Total	Percentage
2Wheeler	4	16	20	80%
Car	3	5	8	63%
Public Transport	1	74	75	99%

5.4) Bagging

The overall accuracy using bagging is 79.6% with the model still performing well in predicting public transport users at 95%.

	Predicted		
	2Wheeler	Car	Public Transport
2Wheeler	5	0	15
Car	0	6	2
Public Transport	4	0	71

There is a general improvement in predicting car usage in comparison to most models when bagging is used.

	Correct predictions	Total Observations	Percentage
2Wheeler	5	20	25%
Car	6	8	75%
Public Transport	71	75	95%

5.5) Boosting

For the boosting algorithm, I decided to reduce the output variables to two levels that is those using a car and those that are not using a car to make the implementation of the algorithm work.

	Predicted			
	FALSE	TRUE	Total	Percentage
Car	0	8	8	100%
Other	94	1	95	99%

The output shows that the model accurately predicted all instances when an employee used a car with an overall accuracy of 99%. It also accurately predicted when an employee did not use a car 99% of the time. Using SMOTE did not improve the results.

6.) Actionable Insights and Recommendations

- The important factors in predicting whether an employee will use a car are, Salary, Age and Work experience. With the higher earning, older and experienced employees likely to use a car. If the company wants to help employees improve their health from use of other modes of transport like bicycles, it can start a wellness program for the older employees that could provide discounts for bicycle purchases for example.

- Distance is also an important factor when determining the use of a car with employee above 19 kms from work likely to use a car. The company can offer work from home options for employees that are not required to be in the office and use cars because of the distance from the office.

- Knowing whether someone is an engineer or not or has an MBA or not is unlikely to assist in predicting whether someone uses a car. The use of a car is related to potential increase in wealth which older employee are likely to have because of their experience.

- Cars are the least used mode of transport with employees favouring public transport. The company can encourage those using public transport to continue doing so because of the lower impact on the environment and the resulting congestion. An awareness campaign on the benefits of using public transport can be launched to ensure those that use it continue doing so and potentially move other employees from other modes.

- If the company wants to help employees improve their health from use of other modes of transport like bicycles, it can start a program for the older employees that could provide discounts for bicycle purchases.