OFFICE COMMUTE

- An Analysis of CHOICE of Mode of Transport by Employees

Pradeep Panicker

4/3/20

MACHINE LEARNING

TABLE OF CONTENTS

PROJECT OBJECTIVE

DATA INTERPRETATION -

A DATA-SET CONTAINING 'EMPLOYEE's MODE OF COMMUTE TO OFFICE' & 'EMPLOYEE PROFILE' – FOR EMPLOYEES WORKING IN A FIRM.

THE DATA IS COMPILED FOR FOUR HUNDRED AND FORTY-FOUR EMPLOYEES.

PROJECT OBJECTIVE –

- TO PERFORM EXPLORATORY DATA ANALYSIS- WITH INSIGHTS.
- CHECK MULTICOLLINEARITY & PRESENT RESULTS GRAPHICALLY.
- PREPARE THE DATA WITH 'SMOTE'.
- MODEL BUILDING WITH – KNN; NAÏVE BAYES; LOGISTIC REGRESSION.
- APPLY BAGGING & BOOSTING MODELLING PROCEDURES – CREAT MODEL AND COMPARE MODEL PERFORMANCE PARAMETERS.
- GENERATE ACTIONALBE INSIGHTS WITH RECOMMENDATION.

WORKING ENVIORNMENT SETUP

I. SOFTWARE TOOL 'R STUDIO' IS ESTABLISHED.

II. WORKING DIRECTORY IS SET TO ACCESS WORKING FILE – 'Cars'.

   a. setwd ('C:/Users/PRADEEP PANICKER/Desktop/BABI/GREAT LAKES/MACHINE LEARNING/PROJECT')

III. Working file imported using [read.csv] command – to the Global Environment.

IV. The file is assigned as 'cars' in R Studio. Dimensions are verified -

   a. dim(bank) – gives a result of 444 Observations with 9 variables have been successfully imported.

V. The Structure of Data is checked for class of each variables.

   a. 9 VARIABLES are found in which –
      i. 05 Variables are integers
      ii. 02 Variables are Factor
      iii. 02 Variables are Numerical.
   b. Variables 'Engineer', 'MBA' & 'license' is being converted to FACTOR from INTEGER since these represent binary, discreet information.
   c. ALL VARIABLES ARE FOUND TO BE SIGNIFICANT.
   d. ONE Observation has Missing Value in the variable 'MBA'. The data will be cleaned by omitting this Observation. The new dimensions of the data set have become – 443 Observations with 09 Variables.

NOW THE DATA FILE - 'cars' - IS READY FOR DATA MANUPULATIONS AND STATISTICAL ANALYSIS.

UNIVARIATE ANALYSIS – SUMMARY.

SUMMARY of data –

The SUMMARY of Data is called for and verified for any missing values.

Command - [summary(cars)]

NOTE – FOR SUMMARY TABLE, PLEASE REFER TO ANNEXURE -I ; TABLE-I

As mentioned above, ONE MISSING VALUE was observed in the original dataset, which was treated and the new dataset ready for analysis has 443 Observations with 09 Rows.

The FIVE Basic Statistical elements of analysis namely – Minimum Value (Min.)of the observed 443 Values, First Inter Quartile Range (1st Qu.), Median of the 443 Observations (Median), Average / Mean of the 443 Observations (Mean), Third Inter Quartile Range (3rd Qu.) and Maximum Value of the observed 443 Values – for EACH OF THE FOUR CONTINEOUS VARIABLES, have been listed through the [summary] command..

For having better EDA of the VARIABLES Having Discreet data – 'Engineer' ; 'MBA' & 'license', the same were converted to FACTOR Variables.

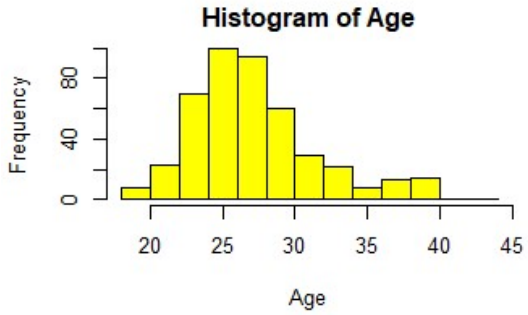
INSIGHTS -

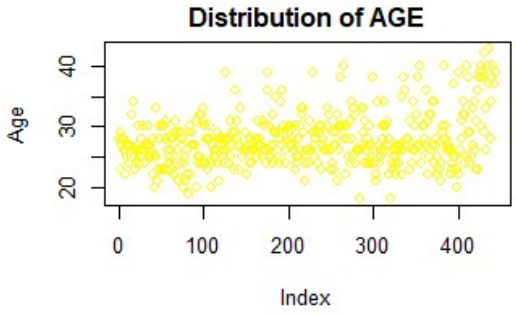
I. **All data points are complete in nature.**
II. **VARIABLES – All variables with continuous data seem to have outliers which can be treated if analysis demands so.**
III. **AGE – ranges from 18 to 43, found to be satisfactory as per driving regulations.**
IV. **TRANSPORT – Three types of transport are being used as mode of transport.**
V. **LICENCE – there are commuters who do not have a License.**

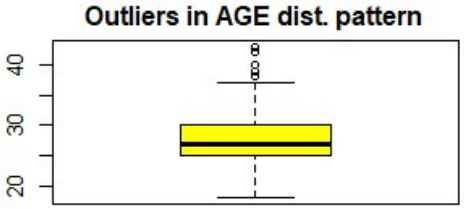
UNIVARIATE ANALYSIS – 'cars' Data File.

AGE of Employees [Age] –

Graph -1







INFERENCES -

Data Distribution – Histogram [Graph I] shows data is NORMALY distributed. Majority of Employees are in the age group 24 Years ~ 30 Years. A small group of employees do exist with age between 36 Years ~ 40 Years.

This also indicates that the Commute analysis concentrates of Method of Transport used by a YOUNG AGE Group of employees – which also indicates a NON-BIAS of usage of Transport due to age factor.

Scatter Plot indicates NO LINEAR Relationship.

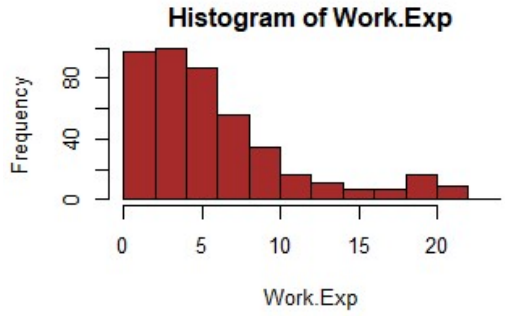
BOX PLOT Suggests no skewness in the data. OUTLIERS HAVE BEEN IDENTIFIED TOWARDS THE MAXIMUM DATA RANGE, i.e above 36Years of age.

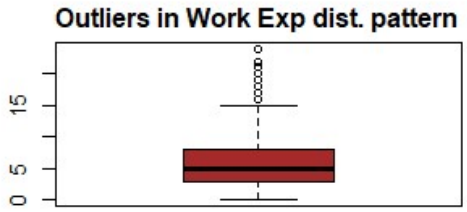
Above the 'Q3 + 1.5IQR' range there are 25 Employees which are categorized as OUTLIERS.

Outlier treatment will be done, during model building exercise, as the situation calls for.

WORK EXPERIENCE [Work. Exp] –

Graph-2





INFERENCES -

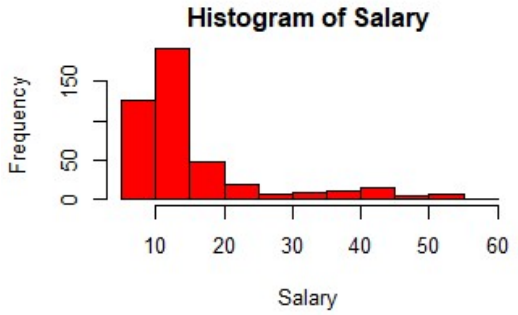
Data Distribution – Histogram [Graph 2] shows data is RIGHT SKEWED. MAJORITY OF EMPLOYEES FALL IN THE LESS EXPERIENCE CATEGORY.
Interestingly, we even have 29 Employees whose work experience has not reached ONE YEAR [employees having 'o' years' work experience]

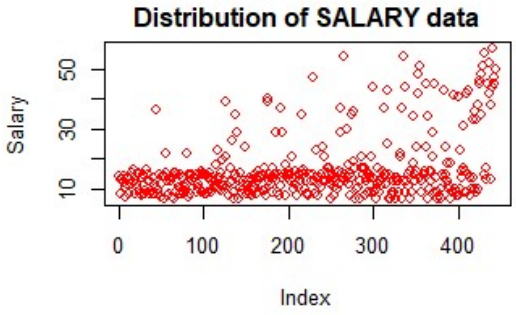
Scatter Plot - indicates NO LINEAR Relationship.

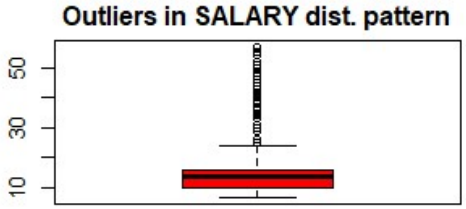
BOX PLOT - Suggests **OUTLIERS HAVE BEEN IDENTIFIED towards the MAXIMUM Data Range, above 15years of experience.**
Above the 'Q3 + 1.5IQR' range there are 38 Employees which are categorized as OUTLIERS.
Outlier treatment will be done, during model building exercise, as the situation calls for.

SALARY of Employees [Salary] –

Graph 3-







INFERENCES -

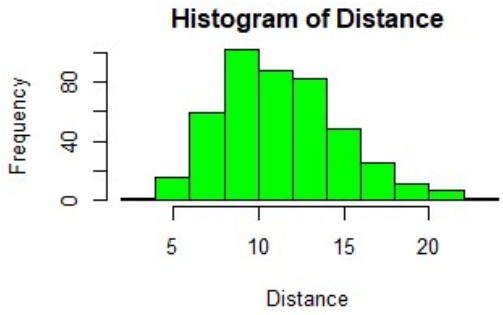
Data Distribution –Histogram [Graph 3] shows a RIGHT SKEWEDNESS of the data. **It is evident that high-salary range of employees is QUITE LESS in the data collected.**

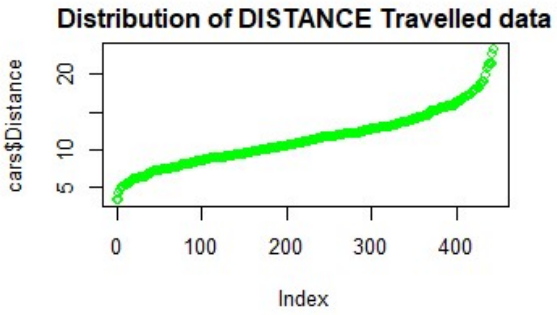
Scatter plot - suggests high density of employees present around the TEN LAKHS per Annum salary bracket.

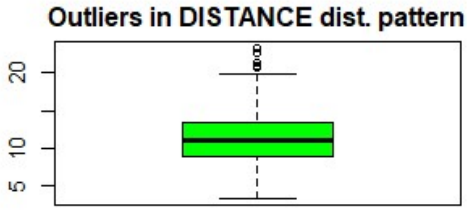
BOXPLOT - shows RIGHT SKEWNESS of DATA. **Outliers are present, highly significant at the 'Above the 1.5IQR + Q3 range'.**
Above the 'Q3 + 1.5IQR' range there are 59 Employees which are categorized as OUTLIERS.
Outlier treatment will be done, during model building exercise, as the situation calls for.

DISTANCE TRAVELLED[Distance] –

Graph 4 -







INFERENCES -

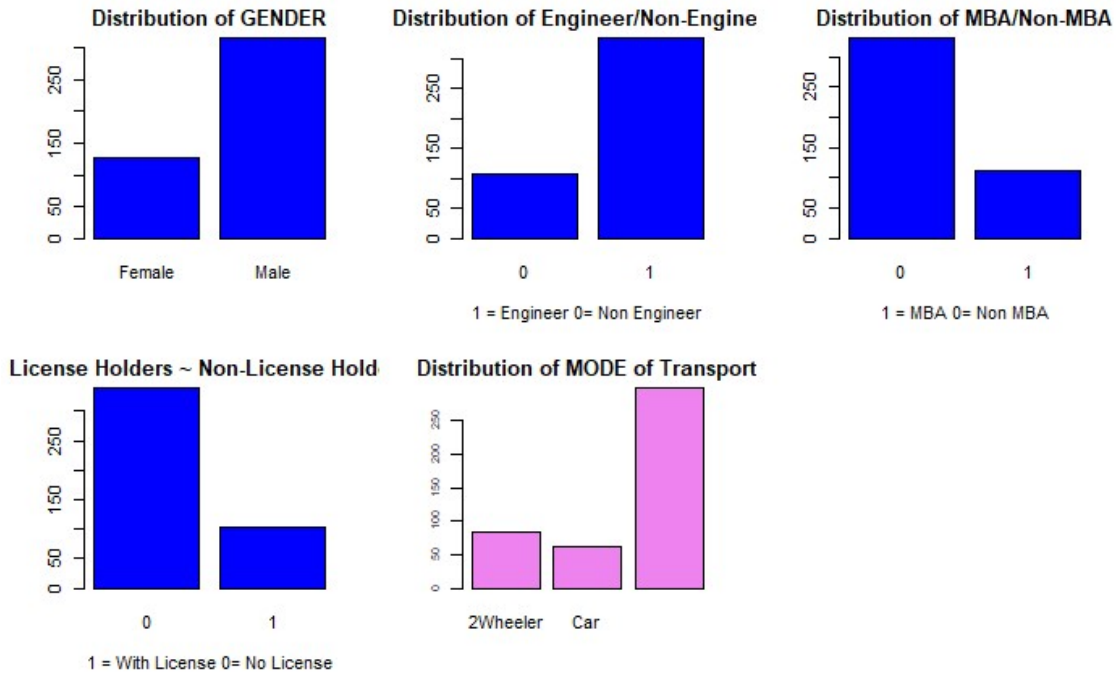
Data Distribution –Histogram [Graph 4] shows that data is QUITE NORMALLY DISTRIBUTED. Most of the employees cover a distance of 09 ~ 10 Units daily.

Scatter Plot - does not show any linear distribution pattern. But it also proves that EVERY EMPLOYEE HAS TRAVELLED SOME DISTANCE TO REACH TO THE WORKPLACE.

BOXPLOT indicates – OUTLIERS are Present at **'Above the 1.5IQR + Q3 range'. It is evaluated that 09 Employees fall in the outlier region – due to ABOVE AVERAGE distance travelled.**

ANALYSIS of 'Engineers/ Non-Engineers; MBA/ NON-MBAs; License Holders & GENDER' –

Graph 5 -



INFERENCES –

GENDER – From the graph above, it is evident that two-third of the employee's data base consists of MALE Employees.

ENGINEER / NON-Engineer – Data base evaluations shows majority of the employees to be engineering graduates.

MBA / NON- MBA – Data pattern shows only ONE -THIRD of the employee base to have a master's degree in business administration.

LICENSE / NO LICENSE - majority of the employees DON'T HAVE DRIVING LICENSE even when they know that they need to travel to reach their workplace.

MODE of TRANSPORT - The third BAR [majority one] represents the 'Public Transport'. Evident from this graphical pattern that most employees uses the Public Transport to reach their work destination.
CAR is the least preferred mode of transport.

MULTIVARIATE ANALYSIS – 'cars' Data File.

OVERALL DATA Scatter Plot Analysis ➔

Graph 9

PLOT ANALYSIS -

CORELATED VARIABLES ➔ Various plots when visually analyzed from GRAPH 9, the below listed MULTIVARIATE Analysis needs to be done for insights, since the same shows collinearity.

1) **'AGE' Vs 'WORK EXPERIENCE'.**
2) **'AGE' Vs 'SALARY'.**
3) **'WORK EXPERIENCE' Vs. 'SALARY'.**

'AGE' Vs 'WORK EXPERIENCE' –

A LOGICAL fact – As AGE INCREASES, WORK EXPERIENCE INCREASE – is proven here with the linear association shown above.

The Correlation factor is evaluated at 93.23%, which is an extremely high Positive collinearity.

This correlation also indicates that employees are encouraged to get motivated in their workplace such that their output remains positive as age increases.

'AGE' Vs 'SALARY' –

Correlation factor stands at 86.08% - VERY GOOD POSITIVE CORRELATION.

The data correlation as represented graphically – shows a positive linear correlation between the two variables. As AGE Increases, the SALARY of the employee also increases.

This IS A POSITIVE INDICATOR FOR THE EMPLOYEE TO INVEST ON ANY KIND OF ASSET PURCHASE – say an Automobile.

'WORK EXPERIENCE' Vs. 'SALARY'.

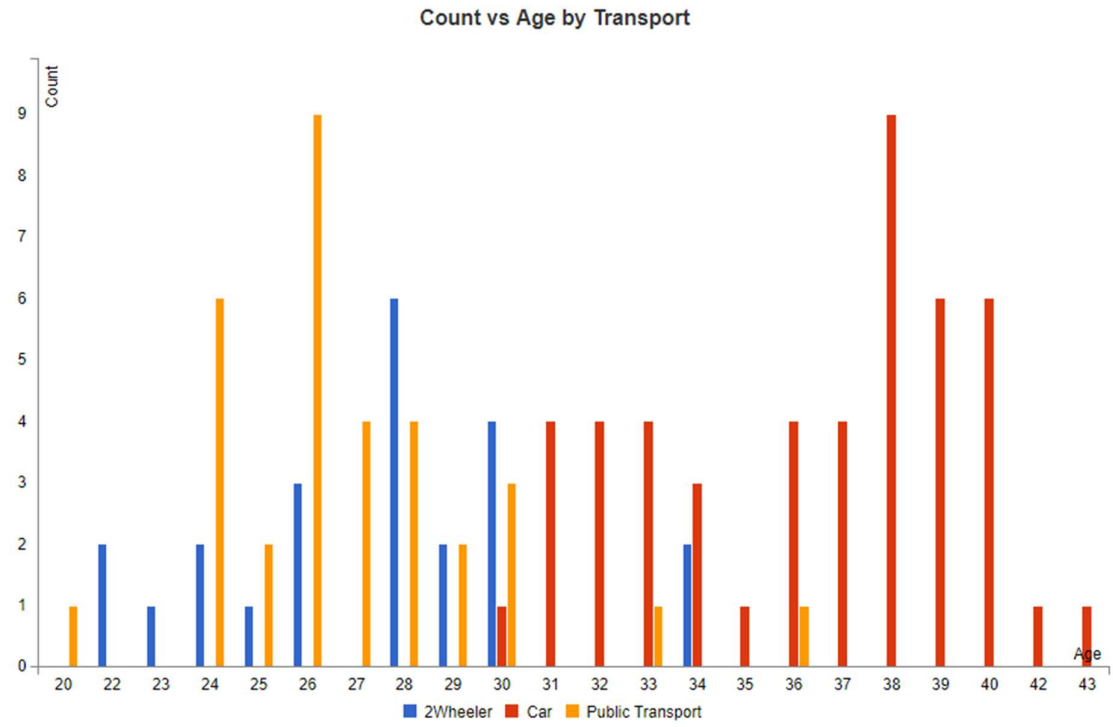
Correlation factor stands at 93.20% - EXTREMLY HIGH POSITIVE CORRELATION.

As the work experience increases in the company, the employee is benefited by a salary increase. The increase is in a positive linear equation. **This correlation motivates the employee to give better output to the company, such that his retainership is considered by the company.**

EDA ➔ DEEP-DIVE

AGE vs MODE of TRANSPORT Used, among License Holders –

Graph 10

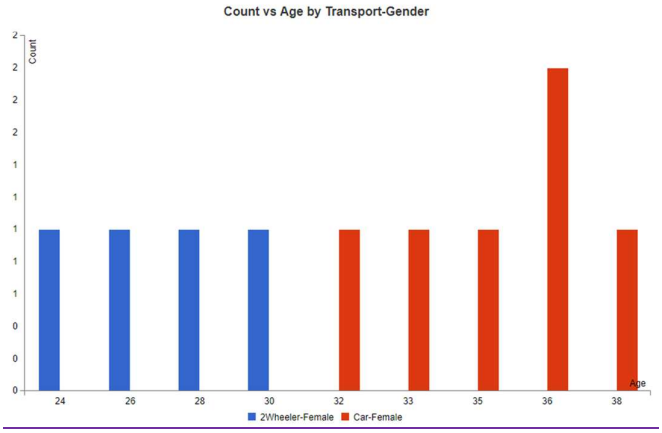


Inferences [from Graph 10] –

a) 2-Wheeler – is used by members between the age 22 ~ 30 ➔ The young age employees. These employees have less work experience and hence have invested on a Two-Wheeler [since less work experience calls for less salary].
b) Car – is used by members between the age 31 ~ 43, with the peak usage coming in at the age of 38. This indicates CAR being bought by employees who have higher income [since age has direct correlation with Salary].
c) Public Transport – Used generously by the Young and Middle aged [20 Years ~ 33 Years], even though they have kept a driving license as standby for usage.

GENDER vs. MODE OF TRANSPORT, among License Holders-

Graph 11- FEMALE EMPLOYEES

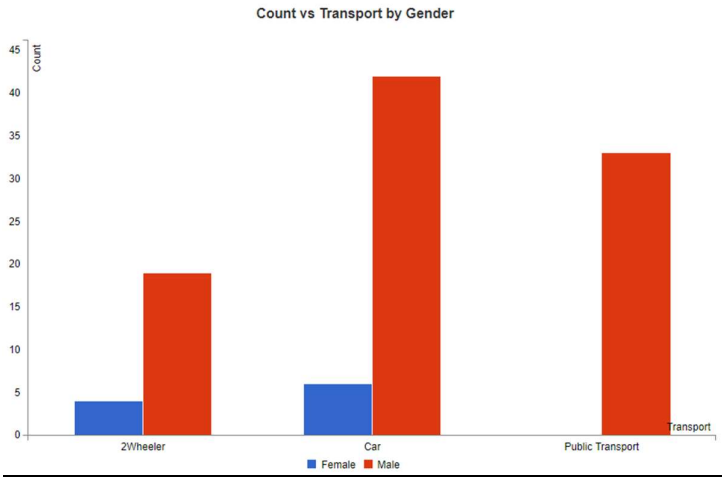


Inferences [from Graph 11] -

a) Interesting to note that ALL FEMALE LICENCE HOLDERS TRAVEL EITHER BY 2-WHEELER OR BY CAR – NOBODY USES 'PUBLIC TRANSPORT'. **IS THE PUBLIC TRANSPORT UNSAFE FOR FEMALE COMMUTE??**

d) While 2-Wheeler is preferred by the younger generation, CAR is preferred by the middle-aged employees.

Graph 12- MALE EMPLOYEES

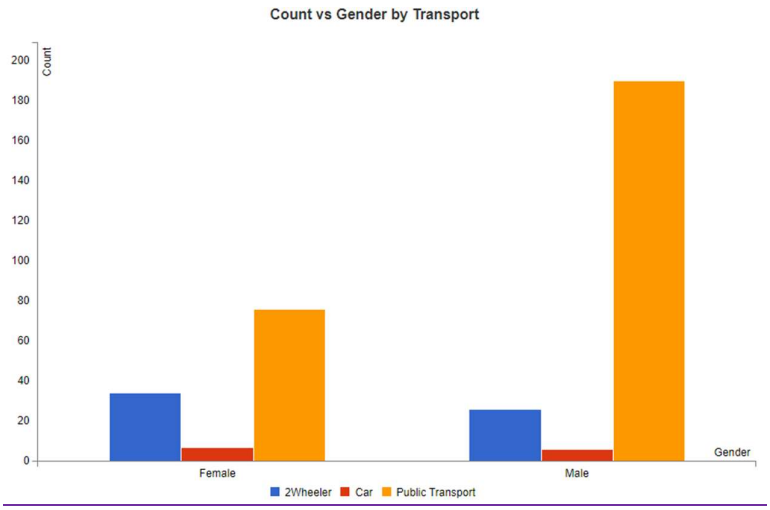


Inferences [from Graph 12] -

e) Public Transport is being used ONLY by the MALE Employees, even though they have a driving license.!!!

Mode of Transport vs. GENDER, among NON-LICENSE Holders -

Graph 13-



Inferences –

a) Female employees, even without license, prefer to travel in 2-Wheeler & Car rather than use the Public Transport.
b) Male employees have a less tendency to use vehicle, when without license, and prefer to use Public Transport.

Comparing inferences of Graph 11 & 13, it is evident that Female Employees prefer to travel in 2-Wheeler / Car rather than using Public Transport.

EDUCATION vs MODE OF TRANSPORT –

NINTY Employees are identified from the total 443 employees – to be having BOTH GRADUATION Degrees – i.e. ENGINEERING + MBA Qualifications.

Education level do have compelled NON-License holders to use Public Transport.

EDUCATION vs SALARY –

Comparing the data for EDUCATION & SALARY for the employees with singular qualification and DUAL qualification, we find –

In case of Singular qualification-

Minimum Age of Employee – 18 Years

Minimum Annual Salary offered – 6.50 Units

In case of DUAL Degree qualification –

Minimum Age of Employee – 21 Years

Minimum Annual Salary offered – 6.60 Units

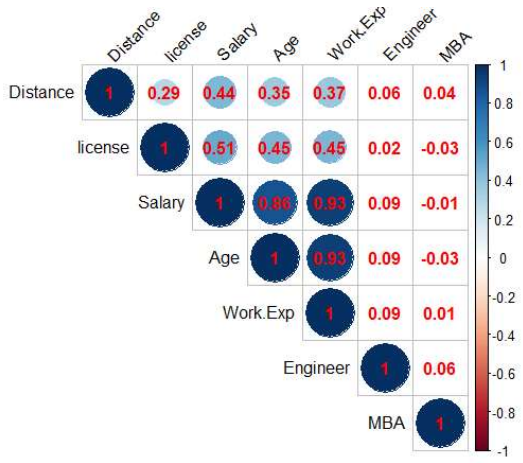
Inference – This sadly states that no advantage is given in renumeration of an employee with DUAL Degree over an Employee with Single Degree of qualification.

INSIGHTS – Based on EDA

I. One observation with a missing value identified – this observation was removed from the data set.
II. OUTLIERS are present in all the variables with continuous data.
III. AGE, WORK EXPERIENCE & SALARY are having a positive correlation – i.e. As AGE increases, the work experience as well as salary increases for the employee. This gives an intuition that the employee is motivated for being retained in the company.
IV. Out of the THREE types of mode of transport used-
    a. CARS are being used by those employees who are at the Higher end of Salary renumeration.
    b. FEMALE Employees prefer NOT-TO-USE Public Transport. They prefer to commute in a 2-Wheeler, self-driven OR as a Pillion-rider.
    c. PUBLIC Transport is used generously by the Young & the Middle-aged employees.
    d. PUBLIC Transport being avoided by FEMALE Employees – indicates either a Less-Socializing nature OR a Specific defined Lifestyle for the working woman class.
V. DISTANCE does not play a role in deciding the choice of MODE of Transport for office commute.
VI. Holding DUAL QUALIFICATION does not give an advantage to the employee w.r.t. renumeration nor experience.

MULTICOLLENIARITY

The collinearity between variables is well represented by the graph below –



The above graph indicates that NO NEGETIVE Collinearity exists, only POSITIVE CORRELATION EXISTS BETWEEN VARIABLES.

The high collinearity between Salary – Age – Work Experience indicates that Multicollinearity will exist. We will now proceed to check the Variation Inflation Factor through different modeling buildups.

VIF evaluation is done and below results were derived –

Applied Commands –

```
vif(lm(Age~Work.Exp+Engineer+MBA+Salary+Distance+license, data = carsmc))
vif(lm(Work.Exp~Age+Engineer+MBA+Salary+Distance+license, data = carsmc))
vif(lm(Engineer~Age+Work.Exp+Salary+Distance+license, data = carsmc))
vif(lm(MBA~Age+Work.Exp+Engineer+Salary+Distance+license, data = carsmc))
vif(lm(Salary~Age+Work.Exp+Engineer+MBA+Distance+license, data = carsmc))
vif(lm(Distance~Age+Work.Exp+Engineer+MBA+Salary+license, data = carsmc))
vif(lm(license~Age+Work.Exp+Engineer+MBA+Distance+Salary, data = carsmc))
```

Results

```
Work.Exp Engineer      MBA   Salary Distance  license
7.768963 1.013290 1.009150 8.782362 1.273160 1.364157
     Age Engineer      MBA   Salary Distance  license
3.893293 1.014872 1.008877 4.451166 1.260183 1.362346
      Age  Work.Exp    Salary  Distance   license
 7.775847 15.531796  8.838203  1.270051  1.379096
      Age  Work.Exp  Engineer    Salary  Distance   license
 7.787631 15.535795  1.010860  8.841972  1.270850  1.380942
     Age Work.Exp Engineer      MBA Distance  license
7.798734 7.887377 1.014811 1.017449 1.189702 1.299151
      Age  Work.Exp  Engineer       MBA    Salary   license
 7.870396 15.545085  1.014950  1.018025  8.282068  1.373517
      Age  Work.Exp  Engineer       MBA  Distance    Salary
 7.780589 15.505349  1.014140  1.020643  1.267268  8.344393
```

From the above analysis, it is found that the variable 'Work Experience' has multi-collinearity with other variables and hence will influence the MODEL Building extensively.

The VIF for 'Work Experience' stand at 15.50

A model was built by removing the variable 'Work Experience', VIF checked. Results were found as follows –

vif(lm(Age~Engineer+MBA+Salary+Distance+license, data = carsmc))
vif(lm(Salary~Age+Engineer+MBA+Distance+license, data = carsmc))

```
Engineer      MBA   Salary Distance  license
1.013110 1.007009 1.555083 1.255225 1.360148
     Age Engineer      MBA Distance  license
1.360182 1.014740 1.008173 1.174901 1.290794
```

MULTI-COLLINEARITY TREATMENT

It is proved, from the above treatment , that removing 'Work Experience' as an independent variable from model building exercises, will ensure a precise estimation & prediction.

SMOTE ANALYSIS

The problem statement of this project is to – Predict whether an Employee will choose CAR as a mode of transport for commute to Office.

The original data shows 13.8% of employees use CAR as a mode of transport to commute to office.

To have a precise prediction, we will apply SMOTE Analysis to the data.

1) The values of the variable 'Transport' are revalued such that 'Car' is represented by '1' and 'Public Transport' & '2Wheeler' is represented by '0'.
2) By this conversion, we understand that the majority class 'NON-CAR Users' are 86.23% of the data base and 'Car Users' are 13.77% of the data base.
3) To create a balanced data, we will now increase the presence of the minority class to 20% of the data base.

The dataset now has the following proportion –

prop.table(table(carssmote.train$car))

```
        0         1
0.8612903 0.1387097
```

Over various iterations to choose the desired 'perc.over' & 'perc.under' combinations to achieve a dataset with 20% of minority class presence, the following dataset was generated by the command –

balanced.carstrain= SMOTE(car~.,carssmote.train, perc.over =50, perc.under = 1300, k=5)

The dimensions of the balanced TRAIN dataset created for Model building is as –

```
table(balanced.carstrain$car)
  0   1
273  64
```

```
prop.table(table(balanced.carstrain$car))
        0         1
0.810089 0.189911
```

MODEL BUILDING – KNN; NAÏVE BAYES & LOGISTIC REGRESSION

KNN – K Nearest Neighbour method of Model building and analysis.

Dataset is prepared with scaling all variables and converting the independent variables to numeric format. The target variable 'car' is sustained as a Factor variable. Scaling is important to normalize the data.

Model is built using the command –

##Version 1 k = 3

carsfit= knn(train = carsknnscaled.train[, 1:7], test = carsknnscaled.test[, 1:7], cl= carsknnscaled.train[,8], k= 3, prob = TRUE)

table(carsknnscaled.test[,8],carsfit)

Also, a model is built with the SMOTE enabled data, model is built using the command –

carsfitsm= knn(train = carsknnbalscaled[, 1:7], test = carsknnscaled.test[, 1:7], cl= carsknnbalscaled[,8], k= 3, prob = TRUE)

table(carsknnscaled.test[,8],carsfitsm)

Various iterations are done to the model with 'K' value ranging from 2 to 17, and values tabled for the model performance on test data -

KNN

	Option 1 'k' = 3		**Option 2 'k' = 17**		**Option 3 'k' = 13**		**Option 4 'k' = 9**		**Option 5 'k' = 5**		**Option 6 'k' = 2**	
	TEST DATASET		TEST DATASET		TEST DATASET		TEST DATASET		TEST DATASET		TEST DATASET	
	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)
0	114	1	115	0	115	0	115	0	114	1	111	4
1	6	12	8	10	8	10	8	10	7	11	3	15
	With SMOTE -		**Option 2 'k' = 17**		**Option 3 'k' = 13**		**Option 4 'k' = 9**		**Option 5 'k' = 5**		**Option 6 'k' = 2**	
	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)
0	110	5	113	2	113	2	114	1	112	3	108	7
1	3	15	4	14	3	15	3	15	3	15	3	15

From the above data table, we find that the model built with 'K' value at 9, is giving the best results of prediction.

With SMOTE application to data set – to boost the accuracy of minority class prediction – the least errors of predictions are found with 'K' value @ 9. This value of 'K' balances the tradeoff between Specificity & Sensitivity.

NAÏVE BAYES – method of Model Building and analysis

The basic conditions to be followed for application of NAÏVE BAYES is –

a) The data Set should be too large enough
b) All the independent variables to be in character format.

Even though our dataset is not too large, we will attempt to build a model using NAÏVE BAYES . We will convert the independent variables to numerical format, leaving alone the target variable in factor format.

The dataset is also scaled for normalization.

The model is built with the original proportion dataset as well as the SMOTE applied dataset.

Model built command with Original proportion dataset –

nb= naiveBayes(x= carsnbscaled.train[,1:7], y= carsnbscaled.train[,8])
pred.nb= predict(nb, newdata = carsnbscaled.test[,1:7])
table(carsnbscaled.test[,8],pred.nb)

Model built command with SMOTE applied dataset –

nbsm= naiveBayes(x=carsnbscaledsm.train[,1:7],y=carsnbscaledsm.train[,8])
pred.nbsm= predict(nbsm, newdata = carsnbscaled.test[,1:7])
table(carsnbscaled.test[,8],pred.nbsm)
The values captured for the predicted Vs the actual for the model when performed on the test dataset is –

NAÏVE BAYES

	TEST DATASET	
	False (0)	True(1)
0	111	4
1	3	15
	With SMOTE -	
	False (0)	True(1)
0	110	5
1	2	16

Application of SMOTE technique has improved the prediction of the minority class of data set.

LOGISTIC REGRESSION – method of Model Building and analysis

With Logistic Regression, we will attempt to build a model for prediction as well as derive actionable insights.

Basic Rule validation of dataset –

1) Target variable in binary values ➔ Verified as OK.
2) LARGE Data Set – Ratio of Rows to columns is 54 Rows per column. Status is WEAK.
3) Multicollinearity – Variable causing multi-collinearity is identified as ' Work. Experience'. This variable will be removed from dataset while model building.

The dataset is built as 'carlogittrain' & 'carlogittest'.

Model is generated using the command –
modellogit= glm(car~Age+Engineer+MBA+Salary+Distance+license,data= carslogittrain, family=binomial(link = 'logit'))

Summary of model is generated, and AIC Value is checked, and significant variables are identified.
The next iteration of model is done on the train dataset , after removing the Non-Significant variables – to reach the best AIC Value [LEAST].

While the initial model command gave an AIC value of 65.878, after removing the in-significant variables , we could re-build a model with AIC Value of 63.322. The Multiple R2 also increased by 4.50 % showing more command on the independent variables on the dependent variables.

To check the impact of the SMOTE dataset , we now build a model using the same.

Initial model is built as per –

modelsmlogit= glm(car~Age+Engineer+MBA+Salary+Distance+license, data= carssmlogittrain, family= binomial(link = 'logit'))

In-significant variables are removed after reviewing the summary of 'modelsmlogit' and further model is developed with lesser AIC value [Initial AIC Value 73.418 ; Final AIC Value 72.043] .

The predicted values and compared to the actual ones – for the target variable . The same is tabulated as below –

Logistic Regression

	TRAIN DATASET			TEST DATASET	
	False (0)	True(1)		False (0)	True(1)
0	264	3		114	1
1	7	36		5	13
	With SMOTE -				
	False (0)	True(1)		False (0)	True(1)
0	269	4		113	2
1	7	57		6	12

Astonishingly , the prediction with the SMOTE applied dataset is weak in prediction of the minority class than the original dataset.

Comparing the Model performance of each method of model building – ON THE TEST DATA SET, the below tabulated data will give insights as to choose the best model that performed accurate predictions –

	Logistic Regression		**KNN with 'k' = 9**		**NAÏVE BAYES**	
	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)
0	114	1	115	0	111	4
1	5	13	8	10	3	15
	With SMOTE Applied Dataset -					
	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)
0	113	2	114	1	110	5
1	6	12	3	15	2	16

From the above data and understanding the sensitivity of the project –

1) To classify and understand which employee will use a car as mode of transport – WE CAN SURELY CHOOSE THE LOGISTIC RGESSION MODEL generated with THE NORMAL ORIGINAL DATA PROPORTION.
2) With respect to the prediction accuracy, 'Employees without CAR' count was 99.13% accurately predicted & 'Employees with Car' count was 83.3% accurately predicted – THEREBY 'KNN' MODEL WITH SMOTE APPLIED ON TRAIN DATA can be chosen as the best model.

APPLYING 'ENSEMBLING METHODS' – for BETTERMENT of Prediction.

Two methods are applied over dataset for ensembling – Bagging & Boosting.

1) Bagging – primarily reduces the variance on the predictive values by generating additional data for training, during model building, from the original dataset itself.
2) Boosting – is a technique where, if an observation is predicted wrongly, it tries to overweigh the same over the others during the next iteration and thereby brings in more accurate prediction results.

BAGGING -

Separate dataset was created for the BAGGING exercise.

A primary model was generated with the command –
carsbag= bagging(car~.,data =carsbag.train, control=rpart.control(minsplit = 4, maxdepth = 5),nbag=25)

Multiple iterations were made to the above model with changing the values of the parameters 'minsplit(ms)' / 'maxdepth(md)' & 'nbag(nb)'.

	ms/md/nb- 4/5/25		ms/md/nb- 6/5/20		ms/md/nb- 4/10/30		ms/md/nb- 10/3/25	
	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)
0	112	3	113	2	114	1	112	3
1	5	13	6	12	3	15	3	15

The best model was generated with *minsplit = 4, maxdepth = 10 & nbag=30*.

BOOSTING -

Adaptive Boosting – Separate dataset was created , ensuring that the target variable is in a numeric format with binary values limited to '0' & '1'.

Basic model was generated with the command –

carsgbm= gbm(formula = car~., distribution = 'bernoulli', data = carsgbm.train, n.trees = 5000, interaction.depth = 1,shrinkage = 0.001,n.cores = NULL, verbose = FALSE, cv.folds = 5)

Iteration to the above model was done with changing values of 'ntrees'/ 'shrinkage' & 'cv.folds'.

Values were recorded as follows –

ADAPTIVE BOOSTING

	n.tree=5000		n.tree=10000		n.tree=15000	
	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)
0	113	2	114	1	114	1
1	7	11	5	13	6	12

The best model was generated with 'ntrees' = 10000.

XG BOOSTING –

Model creation requires dataset to be in MATRIX Format – with target variable separately identified as a separate matrix from the train dataset.

```
cardata= as.matrix(carsxg.train[,1:7])
carlabel= as.matrix(carsxg.train[,8])
carstest= as.matrix(carsxg.test[,1:7])
```

Basic model was generated with the command –

carsxgfit= xgboost(data = cardata, label = carlabel,eta= 0.001, max_depth=3,min_child_weight =3,nrounds = 10000, nfold= 5, objective= 'binary:logistic', verbose=0, early_stopping_rounds = 10)

The generate the parameters for a best fit model, iterations were made with a range of values being assigned for 'eta' / 'max_depth'/ 'nrounds'.

The values were recorded as under –

XG BOOSTING

	Normal		BEST FIT - eta =0.001		BEST FIT - eta =0.7	
	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)
0	113	2	113	2	113	2
1	7	11	7	11	4	14

The best fit model was assigned to the command (values highlighted i above)–

carsxgfit= xgboost(data = cardata,label = carlabel,eta= 0.7,max_depth=9, min_child_weight =3, nrounds = 5000,nfold= 5, objective= 'binary:logistic', verbose=0,early_stopping_rounds = 10)

COMPARING ACCURACY of MODEL Performances –

On the test data, the Predicted values to the Actual values TABULATED RESULTS of the best models from Logistics Regression , Bagging, Adaptive Boosting & XG Boosting is given below –

	KNN Dataset with SMOTE Application		Bagging		Adaptive Boosting		XG BOOSTING	
	TEST DATASET		**TEST DATASET**		**TEST DATASET**		**TEST DATASET**	
	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)	False (0)	True(1)
0	114	1	114	1	114	1	113	2
1	5	13	3	15	5	13	4	14
Accuracy =>	**0.9549**		**0.9699**		**0.9549**		**0.9549**	

From the above , BAGGING Method gives the best accuracy of prediction .

ACTIONABLE INSIGHTS using LOGISTICS REGRESSION

Below is an out derived from the Logistic Regression model created –

	Intercept[a]	Exponential Coeff.[b]	P/1-P [c]= 1-[b]	Probability of Y(No Car) , if X=1
AGE as X	0	2.58	1.58	0.3876
MBA Holder as X	0	0.061	0.939	0.5157
DISTANCE as X	0	1.55	0.55	0.6452

The above variables were identified as the most significant ones which contribute to the decision factors as to whether an Employee uses a CAR or not.

AGE – The influence of AGE on whether to buy a CAR or not in the future, for an employee of the company WHO DOES NOT HAVE A CAR YET , IS ONLY 38.76 %.

MBA – For an MBA Graduate degree Employee and having NO CAR at present, 51.57% probability is there that he / she will buy a CAR in future.

DISTANCE – For an employee who is commuting to office daily and HAS NO CAR at Present, 64,52% chances are there that the employee will buy a CAR due to the distance factor.