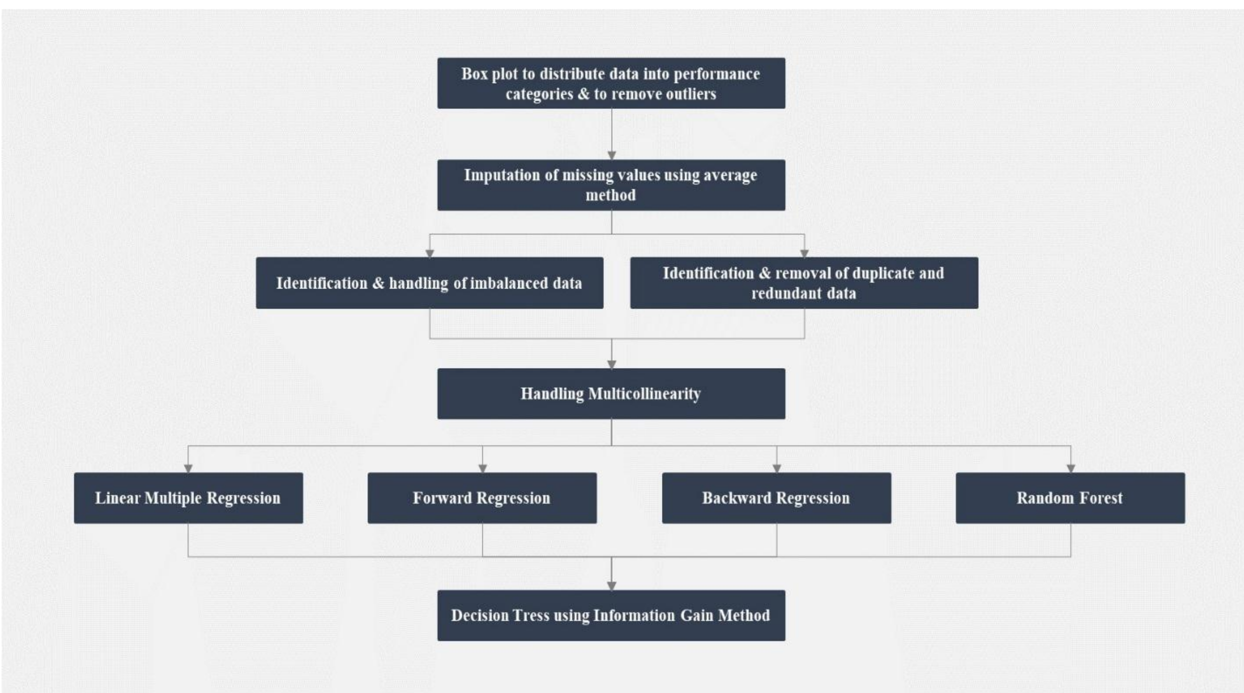
ASHUTOSH AGARWAL

Approach & Result Document

1. **Opening of new stores or relocating stores:** Estimate sales that would be generated by a new location given the characteristics of the new store and location.
2. **Identify high performance stores:** Identify stores that are exceeding expectations so that their success formula can be applied to other store **3. Identify low performance stores:** Identify stores that aren't performing as well as expected and take appropriate decisions including closing them down

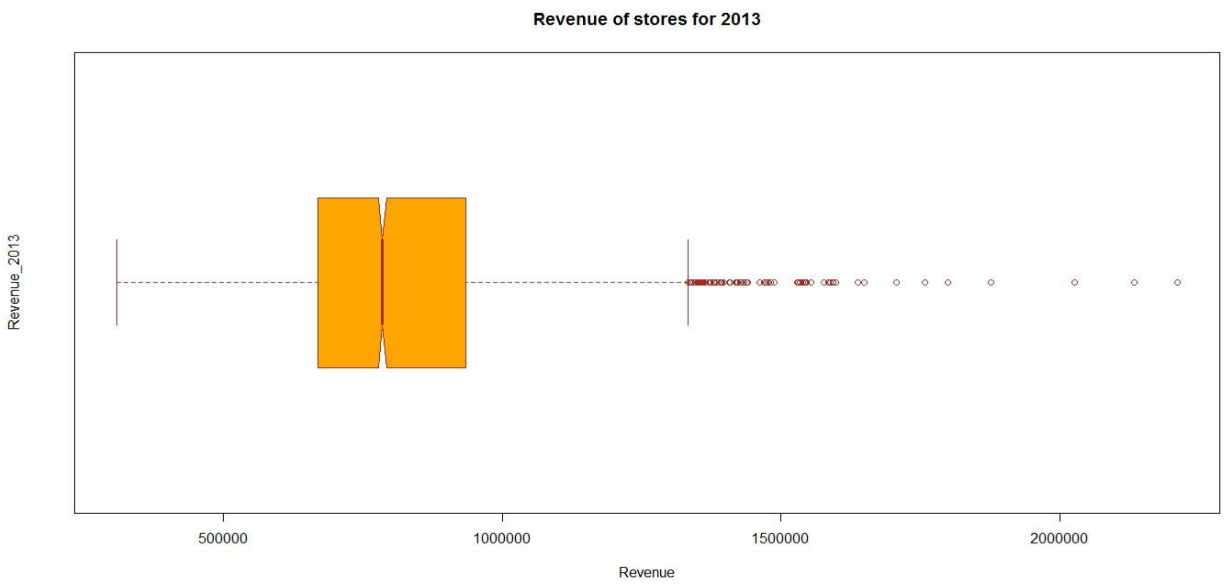
Below flowchart summarizes the approach used for solving the above problems.

Various statistical techniques were used to first clean the data for outliers, imbalanced data, duplicate data, redundant data and to prepare it for regression modelling by eliminating multicollinearity
Box plot was used to distribute data by revenue into three performance categories, namely Excellent, Good & Bad.
These categories were used as target variables in a decision tree to identify decision rules on significant stores characteristics (identified using regression techniques), which can be used to identify a high performance or a low performance store
Regression model was also built to estimate sales for new stores



II.1. Box Plot

Below Box plot was created to distribute store data by revenue into three performance categories namely Excellent, Good & Bad.



Below is the output of the above Box plot.

Results	Value (Revenue)
Min Value	309,408
First quartile	669,492.5
Second Quartile	785,471
Third Quartile	934,952.5
Fourth Quartile	1,332,707
Max Value	2,211,249

Data points outside the fourth quartile (revenue greater than 1,332,707) were treated as outliers and were deleted from the data set. After that, based on the identified quartiles, below conditions were used to categorize store data into Excellent, Good & Bad categories as shown below.

Performance Categories	Condition	Count of Data Points
Excellent	Revenue>= 934,952	640
Good	669,493< Revenue <=934952	1411
Bad	Revenue<=669493	706

Additionally, as it can be inferred from the above table, data is free from any imbalance anomalies.

II.2. Data Cleaning

Below table summarizes the results of data cleaning activates performed.

Activity Performed	Row/Column Impacted	Rationale
Column deleted	PERC_CONVERTED_TO_AGREEMENT	Considered column **number of agreements**. Since number of agreements gave a better explanation. **PERC_CONVERTED_TO_AGREEMENT was** deleted.
Column deleted	CYB02V001	Duplicate columns (CYB07VBASE, CYB02V001). Only column CYB07VBASE was kept.
Column deleted	CENSUS_DIVISION CENSUS_REGION	Considered only the **state** column.
	U_CITY	
Column deleted	PERC_CYEA07V007	Values were almost negligible
Column deleted	SINGLE_TENANT_IND PAD_IN_SHOP_CENTER_IND COMP_PRESENCE_IND PAYLESS_IND WALMART_IND TARGET_IND AUTOZONE_IND	More than 80% of the rows had same values. Therefore, these column were deleted as these wouldn't have made any variation in the model.
	NUM_PARKING_SPACES	
Rows deleted	FRONTAGE_ROAD	Rows deleted where value was "**Unable to determine"** or "**Yes No"**
Values imputed	TOT_ATTRITION_2012 TOT_ATTRITION_2013 NUM_ASSISTANT_MANAGERS NUM_CUST_ACC_REPS NUM_STORE_MANAGERS NUM_EMP_PAY_TYPE_H AVG_PAY_RATE_PAY_TYPE_S AVG_PAY_RATE_PAY_TYPE_H	Missing values were imputed for these columns. Average value for the column was used for filling the missing data

II.3. Handling Multicollinearity

Multicollinearity is a problem because it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret. To eliminate multicollinearity in the regression model, correlated continuous & categorical variables were removed as shown below.

II.3.1. Chi-square test for categorical variables

Chi-square was used to identify correlated categorical variables. The test was used on the variables: **FRONTAGE_ROAD", "STRIP_SHOP_CENTER_IND".**

Null hypotheses: FRONTAGE_ROAD, STRIP_SHOP_CENTER_IND are independent
Below is the result for chi square test.

Results	Values
Chi-square test statistic (X^2)	9.4608
Degrees of freedom (df)	1
P-value	0.002099

Since p-Value is less than the significance level of 0.05, null hypothesis was rejected and it was concluded that the two variables are in fact dependent. Therefore, the variable "**STRIP_SHOP_CENTER_IND**" was deleted.

II.3.2. Correlation for continuous variables

Before applying correlation, continuous variables were normalized by using **Z-Score** methodology. This was to ensure that the variables are at the same scale to facilitate to accurate application of correlation.

Variables which were highly correlated that is with correlation coefficient greater than or equal to 0.9 were deleted. Below are the results after running correlation test between all the continuous variables in the data.

Highlighted cells in the below correlation matrix shows highly correlated variable pairs. For instance, **NAT_CURR_BURGLARY** is correlated to **NAT_PAST_BURGLARY**. Therefore, one of the correlated variable was deleted for each pair. Deleted variables were: **NAT_PAST_BURGLARY, NAT_PAST_MOT_VEH_THEFT & NAT_PAST_ROBBERY.**

Correlation Matrix	Y	RY	H_THEFT
NAT_CURR_BURGLARY	0.490253909	1	0.522482183
NAT_PAST_BURGLARY	0.462288233	0.952152028	0.464386853
NAT_CURR_MOT_VEH_THEFT	0.805064151	0.522482183	1
NAT_PAST_MOT_VEH_THEFT	0.77499171	0.520196538	0.92290718
NAT_CURR_ROBBERY	1	0.490253909	0.805064151
NAT_PAST_ROBBERY	0.971109376	0.483519924	0.740677959

Similarly, variable **"PERC_CYB11V006"** was deleted for the below correlation matrix.

Correlation Matrix	PERC_CYB11V006	PERC_CYB11V007
PERC_CYB11V006	1	0.947791664
PERC_CYB11V007	0.947791664	1

Only variable **CYA01V001** was kept and all the other variables were deleted for the below correlation matrix.

Correlation Matrix	CYA01V001	CYA12V003
CYA12V001	0.9687948	0.8345908
CYA12V002	0.96853677	0.99108425
CYA12V003	0.9359856	1
CYA12V007	0.9338927	0.8742704
CYA12V008	0.9382616	0.8852796
CYB07VBASE	0.98560684	0.92023831
Total_White_Population	0.825679237	0.910799698

II.4. Regression

After performing all data cleaning activities & removing correlated variables, **Linear multiple regression** model was built.

Dependent Variable: "revenue_2013"

Independent Variables:

"U_STATE", "SQUARE_FEET", "TOT_ATTRITION_2012", "TOT_ATTRITION_2013", "NUM_ASSISTANT_MANAGERS", "NUM_CUST_ACC_REPS", "NUM_STORE_MANAGERS", "NUM_EMP_PAY_TYPE_H", "AVG_PAY_RATE_PAY_TYPE_S", "AVG_PAY_RATE_PAY_TYPE_H", "NAT_CURR_ROBBERY", "NAT_CURR_BURGLARY", "NAT_CURR_MOT_VEH_THEFT ", "FRONTAGE_ROAD", "MARKETING_EXP_2013", "MARKETING_EXP_2012", "TOT_NUM_LEADS", "NUM_CONVERTED_TO_AGREEMENT", "CYA01V001", "CYA12V001", "CYA21V001", "XCX03V069" "PERC_CYB11V007", "PERC_CYC13VV01", "Total_Black_African_American_Population", "Total_Asian_Population**"**

Below are the summary screen shots (portioned into three for sake of clarity) of the results of the regression model

Summary (1/3):

```
Residuals:
   Min      1Q  Median      3Q     Max
-347744  -64750   -4702   65708  473116

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
(Intercept)                                -2.867e+05  7.288e+04  -3.935 8.61e-05 ***
U_STATEAL                                   6.057e+03  5.610e+04   0.108 0.914031
U_STATEAR                                   3.081e+03  5.742e+04   0.054 0.957215
U_STATEAZ                                  -1.409e+05  5.527e+04  -2.549 0.010864 *
U_STATECA                                   6.517e+04  5.781e+04   1.127 0.259786
U_STATECO                                  -1.061e+05  5.599e+04  -1.894 0.058307 .
U_STATECT                                   2.047e+04  5.759e+04   0.355 0.722353
U_STATEDC                                  -5.291e+04  8.223e+04  -0.643 0.519994
U_STATEDE                                   1.590e+04  6.069e+04   0.262 0.793310
U_STATEFL                                  -6.269e+04  5.501e+04  -1.140 0.254593
U_STATEGA                                  -1.060e+05  5.571e+04  -1.903 0.057190 .
U_STATEHI                                   2.691e+04  6.784e+04   0.397 0.691603
U_STATEIA                                   4.831e+04  5.781e+04   0.836 0.403454
U_STATEID                                  -9.205e+04  6.637e+04  -1.387 0.165599
U_STATEIL                                  -5.810e+04  5.503e+04  -1.056 0.291176
U_STATEIN                                  -8.240e+04  5.509e+04  -1.496 0.134880
U_STATEKS                                   2.811e+03  5.672e+04   0.050 0.960481
U_STATEKY                                  -6.337e+04  5.596e+04  -1.132 0.257609
U_STATELA                                  -1.271e+04  5.661e+04  -0.225 0.822382
U_STATEMA                                   4.649e+04  5.545e+04   0.838 0.401905
U_STATEMD                                   5.243e+03  5.619e+04   0.093 0.925679
U_STATEME                                   2.883e+04  5.804e+04   0.497 0.619445
U_STATEMI                                   6.365e+03  5.514e+04   0.115 0.908120
U_STATEMO                                  -7.471e+04  5.600e+04  -1.334 0.182322
U_STATEMS                                   1.362e+04  5.699e+04   0.239 0.811092
U_STATEMT                                  -7.709e+04  7.122e+04  -1.082 0.279206
U_STATENC                                  -4.580e+04  5.500e+04  -0.833 0.405094
U_STATEND                                   2.327e+05  1.173e+05   1.984 0.047424 *
U_STATENE                                  -6.008e+04  6.508e+04  -0.923 0.356015
U_STATENH                                   6.819e+04  6.125e+04   1.113 0.265727
```

Summary (2/3):

U_STATEND	2.327e+05	1.175e+05	1.984	0.047424	
U_STATENE	-6.008e+04	6.508e+04	-0.923	0.356015	
U_STATENH	6.819e+04	6.125e+04	1.113	0.265727	
U_STATENJ	6.511e+04	5.688e+04	1.145	0.252470	
U_STATENM	-8.052e+04	5.787e+04	-1.392	0.164197	
U_STATENV	-2.955e+04	5.972e+04	-0.495	0.620759	
U_STATENY	5.565e+04	5.422e+04	1.026	0.304904	
U_STATEOH	-2.474e+04	5.483e+04	-0.451	0.651946	
U_STATEOK	2.782e+04	5.741e+04	0.485	0.628078	
U_STATEOR	-1.809e+05	5.719e+04	-3.162	0.001588	**
U_STATEPA	8.170e+03	5.446e+04	0.150	0.880762	
U_STATERI	-5.163e+04	5.992e+04	-0.862	0.388942	
U_STATESC	-5.767e+04	5.607e+04	-1.028	0.303879	
U_STATESD	-7.207e+04	7.115e+04	-1.013	0.311241	
U_STATETN	-5.931e+04	5.568e+04	-1.065	0.286977	
U_STATETX	-1.635e+04	5.417e+04	-0.302	0.762885	
U_STATEUT	-9.971e+04	5.967e+04	-1.671	0.094893	.
U_STATEVA	-1.315e+04	5.580e+04	-0.236	0.813667	
U_STATEVT	3.491e+03	6.845e+04	0.051	0.959327	
U_STATEWA	-1.431e+05	5.622e+04	-2.545	0.010989	*
U_STATEWV	1.261e+05	5.848e+04	2.156	0.031160	*
U_STATEWY	-6.010e+04	7.432e+04	-0.809	0.418780	
SQUARE_FEET	6.935e+00	1.785e+00	3.886	0.000105	***
TOT_ATTRITION_2012	4.622e+02	1.547e+03	0.299	0.765103	
TOT_ATTRITION_2013	5.009e+03	1.549e+03	3.234	0.001238	**
NUM_ASSISTANT_MANAGERS	2.023e+03	1.291e+03	1.567	0.117269	
NUM_CUST_ACC_REPS	3.197e+03	9.277e+02	3.447	0.000579	***
NUM_EMP_PAY_TYPE_H	4.972e+04	3.575e+03	13.905	< 2e-16	***
AVG_PAY_RATE_PAY_TYPE_S	4.105e+00	3.610e-01	11.372	< 2e-16	***
AVG_PAY_RATE_PAY_TYPE_H	1.806e+04	3.193e+03	5.656	1.76e-08	***
NAT_CURR_ROBBERY	-2.099e+00	2.724e+01	-0.077	0.938593	
NAT_CURR_BURGLARY	-3.660e+01	1.654e+01	-2.213	0.027011	*
NAT_CURR_MOT_VEH_THEFT	6.850e+01	2.313e+01	2.962	0.003092	**
FRONTAGE_ROADYes	7.645e+03	4.836e+03	1.581	0.114057	
MARKETING_EXP_2013	6.010e+00	4.219e+00	1.424	0.154455	
MARKETING_EXP_2012	-4.729e+00	2.363e+00	-2.002	0.045452	*

Summary (3/3):

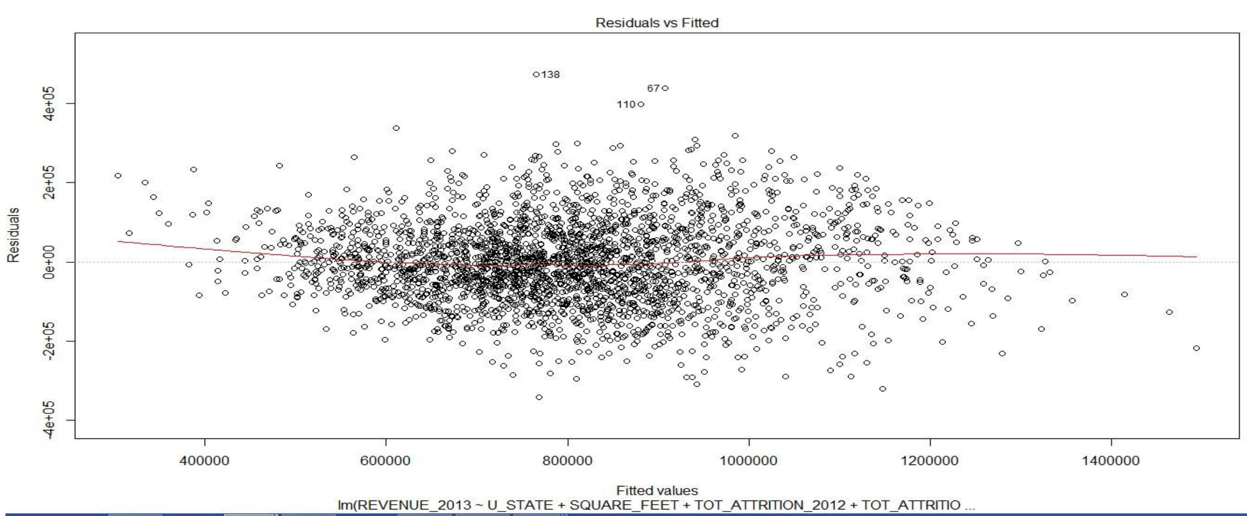
```
NAT_CURR_ROBBERY                               -2.112e+00  2.728e+01  -0.077 0.938299
NAT_CURR_BURGLARY                              -3.622e+01  1.656e+01  -2.187 0.028826 *
NAT_CURR_MOT_VEH_THEFT                          6.832e+01  2.315e+01   2.951 0.003207 **
FRONTAGE_ROADYes                                7.448e+03  4.846e+03   1.537 0.124453
MARKETING_EXP_2013                              6.100e+00  4.225e+00   1.444 0.148969
MARKETING_EXP_2012                             -4.737e+00  2.367e+00  -2.001 0.045476 *
TOT_NUM_LEADS                                   1.906e+01  1.232e+00  15.474  < 2e-16 ***
NUM_CONVERTED_TO_AGREEMENT                      5.198e+02  2.993e+01  17.370  < 2e-16 ***
CYA01V001                                       3.273e-01  1.320e-01   2.480 0.013211 *
CYA12V001                                      -8.082e-01  3.545e-01  -2.279 0.022740 *
CYA21V001                                       6.631e+01  2.556e+01   2.594 0.009546 **
XCX03V069                                       3.837e+02  2.449e+02   1.567 0.117319
PERC_CYB11V007                                 -2.093e+04  2.194e+05  -0.095 0.924011
PERC_CYC13VV01                                 -1.800e+04  7.971e+03  -2.258 0.024047 *
Total_Black_African_American_Population         2.260e-02  1.193e-01   0.189 0.849739
SIGNAGE_VISIBILITY_INDUnable to determine      -3.995e+03  1.507e+04  -0.265 0.790908
SIGNAGE_VISIBILITY_INDYes                      -3.161e+03  6.181e+03  -0.511 0.609119
SIGNAGE_VISIBILITY_INDYes No                    4.216e+04  6.128e+04   0.688 0.491578
Total_Asian_Population                          2.364e-01  3.315e-01   0.713 0.475876
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 104300 on 2102 degrees of freedom
Multiple R-squared:  0.7183,    Adjusted R-squared:  0.7082
F-statistic: 70.54 on 76 and 2102 DF,  p-value: < 2.2e-16
```

As evident from the above summary, **R square value is moderately high at 71.83%** and **adjusted R square at 70.82%** is close to R square. This implies that the model explains the variability of the response data to a good extent.

II.4.1. Testing Regression Model

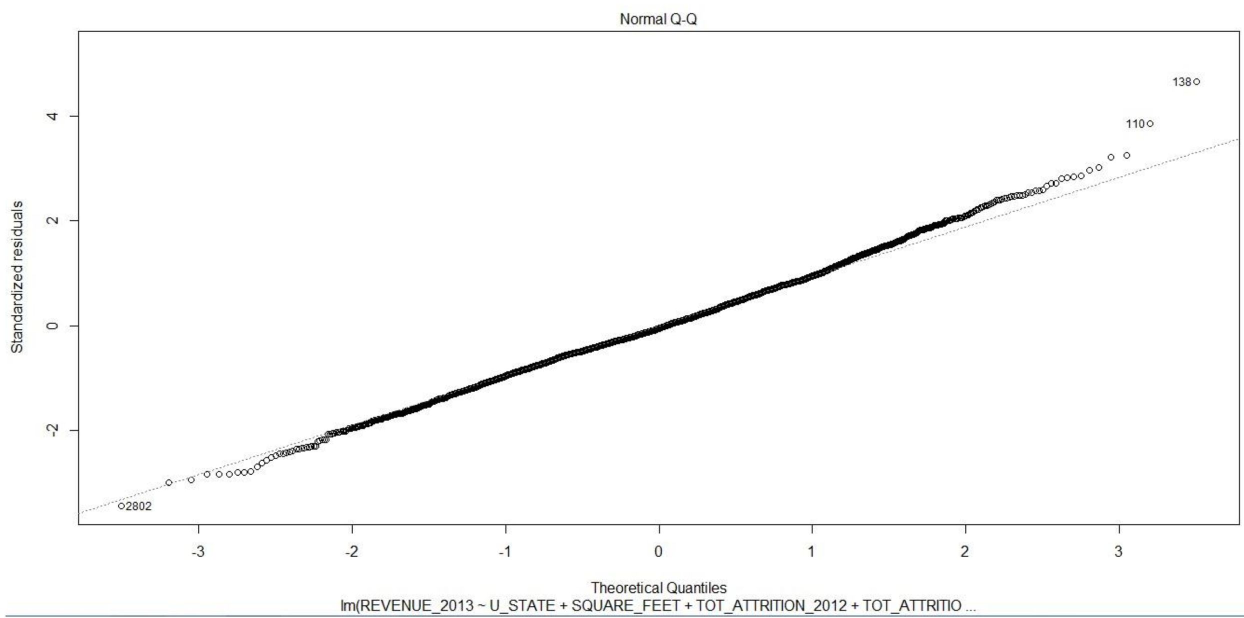
In addition to looking at R-square, other tests were also conducted to validate the model as below.

Homoscedasticity Test



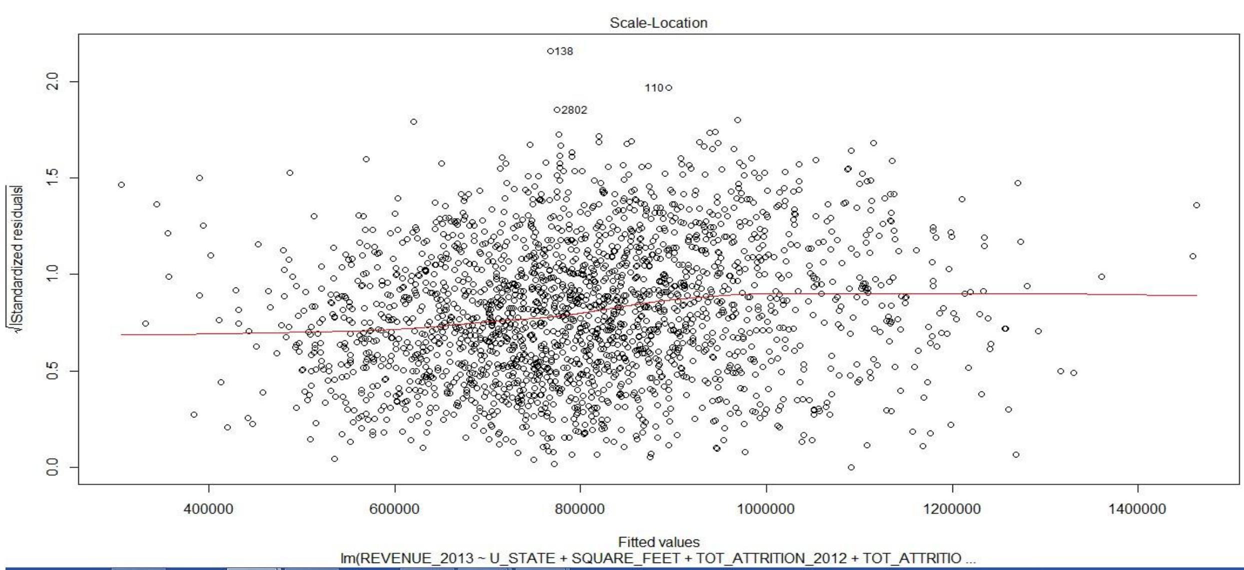
Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. From the above graph, it was inferred that there is Homoscedasticity in the model which means there are not outliers in the model & the true variance and covariance are not underestimated.

QQ plot



The QQ plot was a straight line which indicates the errors are normally distributed.

Standardized Residual Plot



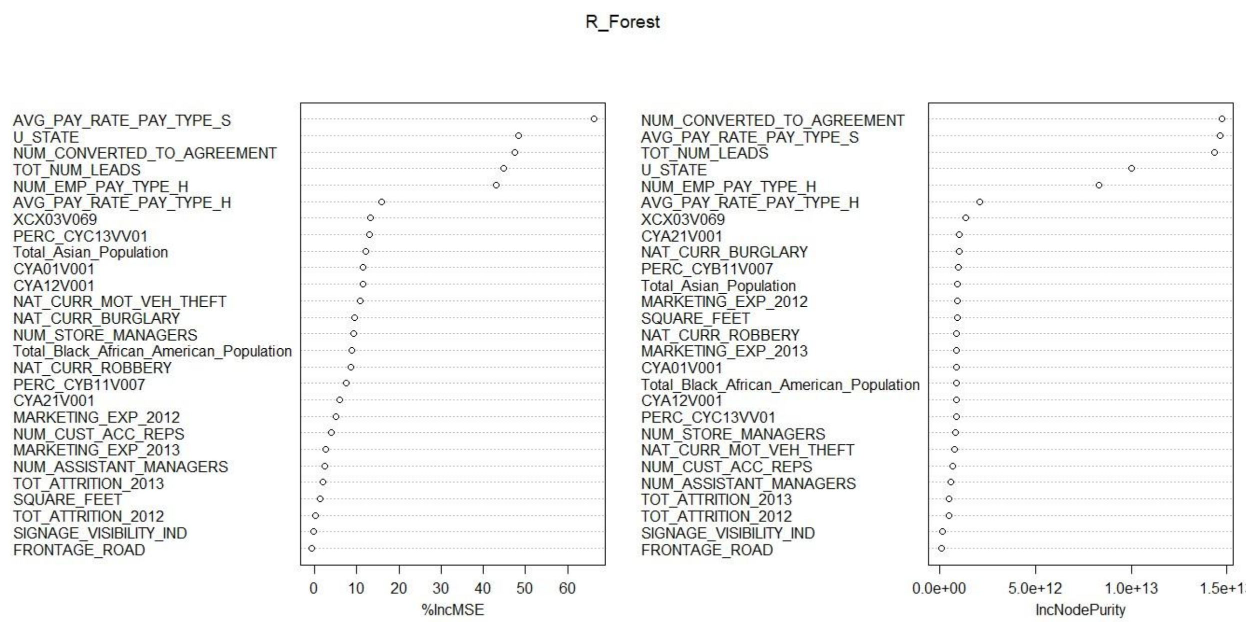
The Standardized Residual plot was homogeneously distributed, and no patterns were observed. This indicates that error terms

<u>Thus, all the above tests helped to validate the assumptions taken in the model.</u>

Additionally, stepwise regression models (both backward and frontward) were also run and gave the same results.

II.4.2. Random Forest

Random forest was used along with the Linear multiple regression to select a list of common significant variables. This was to done to validate the results of the liner regression. Dependent & Independent variables similar to linear multiple regression model were used. Below is the variable importance plot of the random forest.



Using the results of both the **Linear Multiple Regression model** & the **Random forest**, common top significant variables were identified. Using these significant variables, regression equation was built.

<u>Regression Equation</u>

Sales= (-2.960e+05)
+ (2.942e+00*AVG_PAY_RATE_PAY_TYPE_S)
+ (5.653e+02*NUM_CONVERTED_TO_AGREEMENT)
+ (6.022e+04*NUM_EMP_PAY_TYPE_H)
+ (1.534e+01*TOT_NUM_LEADS)
+ (2.150e+04*AVG_PAY_RATE_PAY_TYPE_H)
+ (3.837e+02*XCX03V06)

+ (-1.426e+05*U_STATEAZ)
+ (2.326e+05*U_STATEND)
+ (-1.813e+05*U_STATEOR)
+ (-1.441e+05*U_STATEWA)
+ (1.255e+05*U_STATEWV)

II.5. Decision Tree

Decision tree was built to come up with decision rules that can be used to evaluate the performance of a store. It was build using the **Information Gain Methodology.**

Dependent Variables

Performance categories based on revenue, as identified earlier using the box plot, were used as dependent variables.

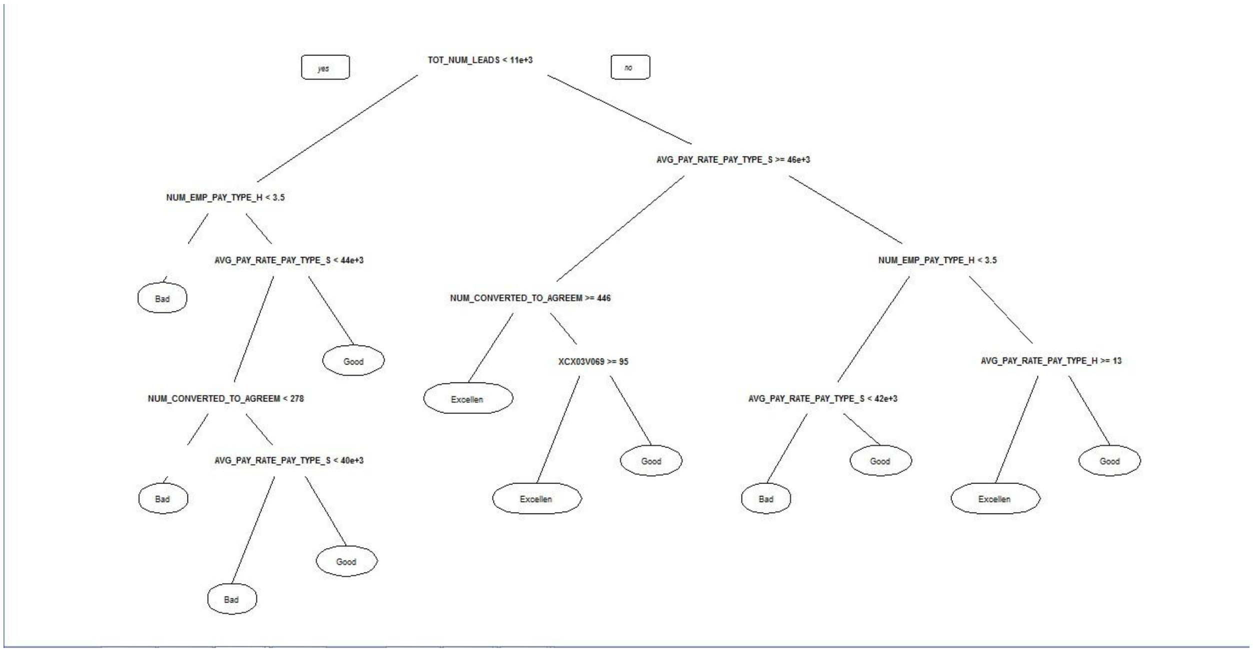
Independent Variables

Below significant variables identified using the regression were used as independent variables in the decision tree.

U_STATE_AZ
U_STATE_ND
U_STATE_OR
U_STATE_WA
U_STATE_WV
AVG_PAY_RATE_PAY_TYPE_S
NUM_CONVERTED_TO_AGREEMENT
NUM_EMP_PAY_TYPE_H
TOT_NUM_LEADS+AVG_PAY_RATE_PAY_TYPE_H
XCX03V06

Please note that the new variables were created for the states based on output of the regression equation. These are U_STATE_AZ, U_STATE_ND, U_STATE_OR, U_STATE_WA & U_STATE_WV. These variables could take tow values either 0 (which implies that store is not in specified state) or 1 (which implies that store is in specified state).

Below is the decision tree built using the above inputs.



We have assumed that the performance catego ry " Excellent" equates to the high perfo rm ing sto res & the cate go ry " Bad" equate s to lo w perfo rm ing sto res. **Final summarized results of the decision tress are given in the next section.**

III. Results

Below are the results for the Problem 2.

1. Linear multiple regression model was used to come up with regression equation to estimate the sales at a new store given characteristics of the new store & location. Additionally, forward & backward regression was also done which gave exactly the same significant variables as linear multiple regression. Random forest was used along with the Linear multiple regression to select a list of common significant variables. This was to done to validate the results of the liner regression.

   Below is the regression equation to estimate sales of a new store.

   Sales= (-2.960e+05)
   + (2.942e+00*AVG_PAY_RATE_PAY_TYPE_S)
   + (5.653e+02*NUM_CONVERTED_TO_AGREEMENT)
   + (6.022e+04*NUM_EMP_PAY_TYPE_H)
   + (1.534e+01*TOT_NUM_LEADS)
   + (2.150e+04*AVG_PAY_RATE_PAY_TYPE_H)
   + (3.837e+02*XCX03V06)
   + (-1.426e+05*U_STATEAZ)
   + (2.326e+05*U_STATEND)
   + (-1.813e+05*U_STATEOR)
   + (-1.441e+05*U_STATEWA)

+ (1.255e+05*U_STATEWV)

As evident fr om the above equation, it's not advisable to open a store in the states: AZ, OR & WA since these have negative impact on sales because of negative regression coefficients. On the other hand, states: ND & WV are favourable locations to open new stores.

2. Decision tree was used to come up with the conditions that can be applied to the stores characteristics to evaluate if a store is high performing or low performing. Below table summarizes the various decision rules identified. There are three decision rules to identify a high performing store & four decision rules to evaluate a low performing store.

   Please note that for each decision rules, all the conditions on store characteristics should be satisfy.

Performance	Decision Rules
High Performance Stores	Tot_Num_Leads>=11,000 Avg_Pay_Rate_Pay_Type_S>=46,000 Num_Converted_To_Agreem>=446
	Tot_Num_Leads>=11,000 Avg_Pay_Rate_Pay_Type_S>=46,000 Num_Converted_To_Agreem<446 XCX03V069>=95
	Tot_Num_Leads>=11,000 Avg_Pay_Rate_Pay_Type_S<46,000 Num_Emp_Pay_Type_H>=3.5 Avg_Pay_Rate_Pay_Type_H>=13
Low Performance Stores	Tot_Num_Leads<11,000 Num_Emp_Pay_Type_H<3.5
	Tot_Num_Leads<11,000 Num_Emp_Pay_Type_H>=3.5 Avg_Pay_Rate_Pay_Type_S<44000 Num_Converted_To_Agreem<278
	Tot_Num_Leads<11,000 Num_Emp_Pay_Type_H>=3.5 Avg_Pay_Rate_Pay_Type_S<44000 Num_Converted_To_Agreem>=278 Avg_Pay_Rate_Pay_Type_S<40,000
	Tot_Num_Leads>=11,000 Avg_Pay_Rate_Pay_Type_S<46,000 Num_Emp_Pay_Type_H<3.5 Avg_Pay_Rate_Pay_Type_S<42,00