MIDTERM PROJECT

Advances in Data Science/Architecture

Group 10
Jun, Wu
Vishwa, Shah
Avinita, Shetty

Table of Contents

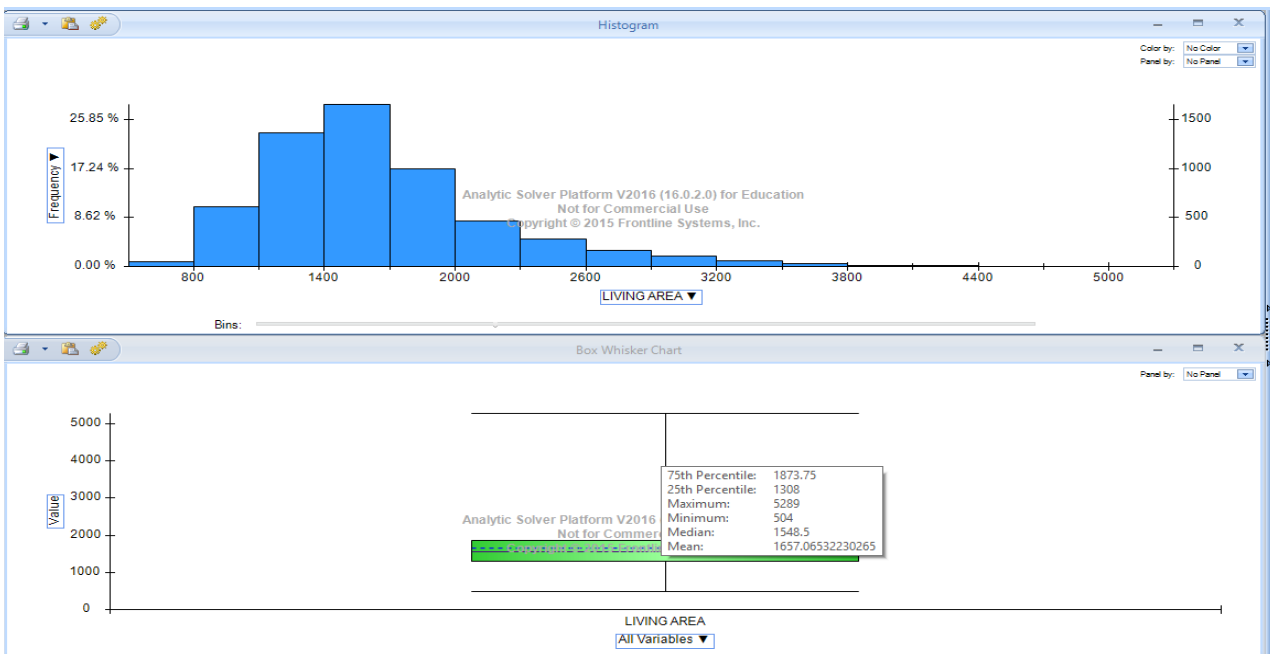
1. West Roxbury

Problem: West Roxbury is a dataset about the houses in the West Roxbury neighborhood and we need to compute the expected prices of a home in future from this given dataset.

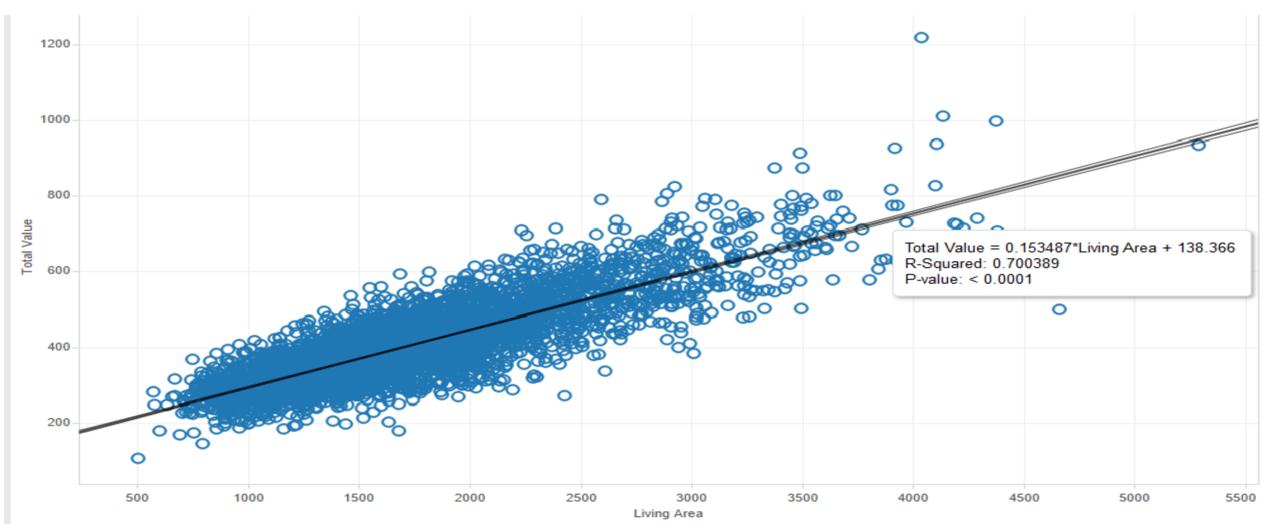
1.1. Exploratory data analysis using TABLEAU and XLMiner

Tool Used: Tableau and Excel

Living Area:

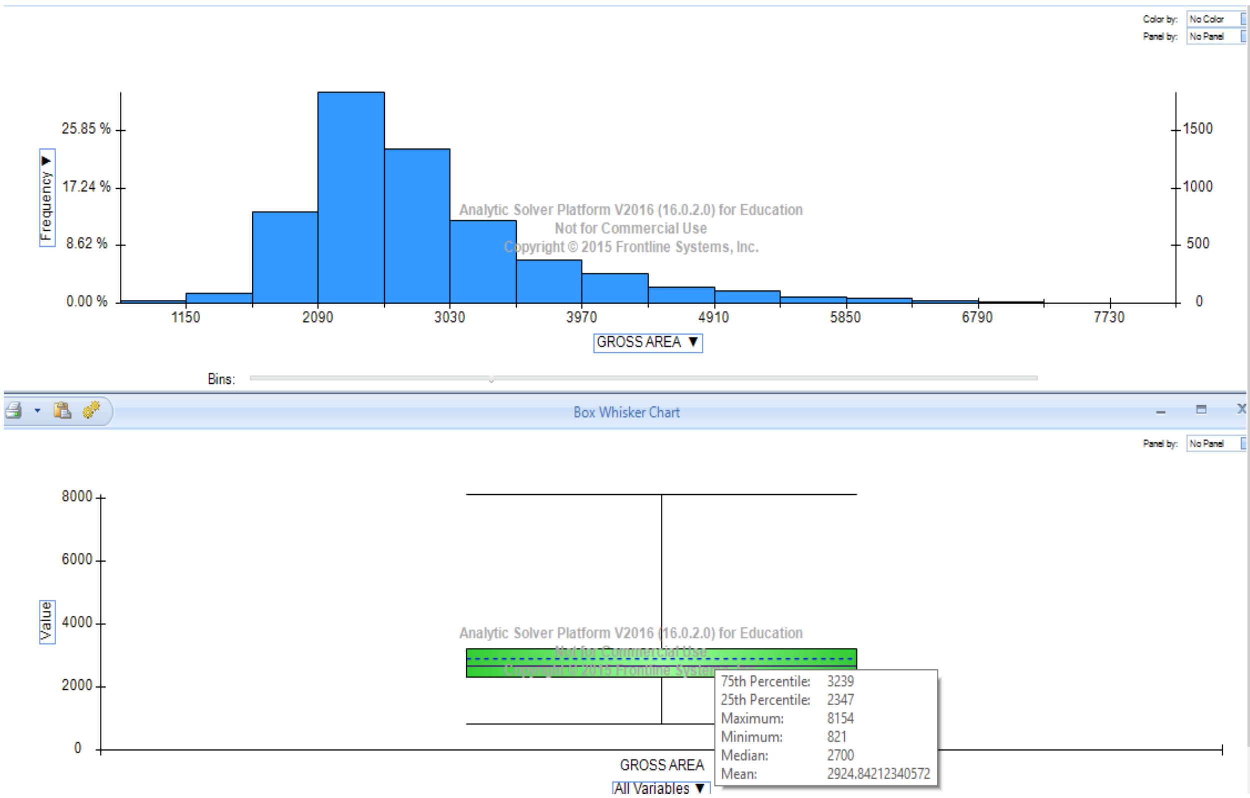


Scatter Plot in Tableau for Living Area

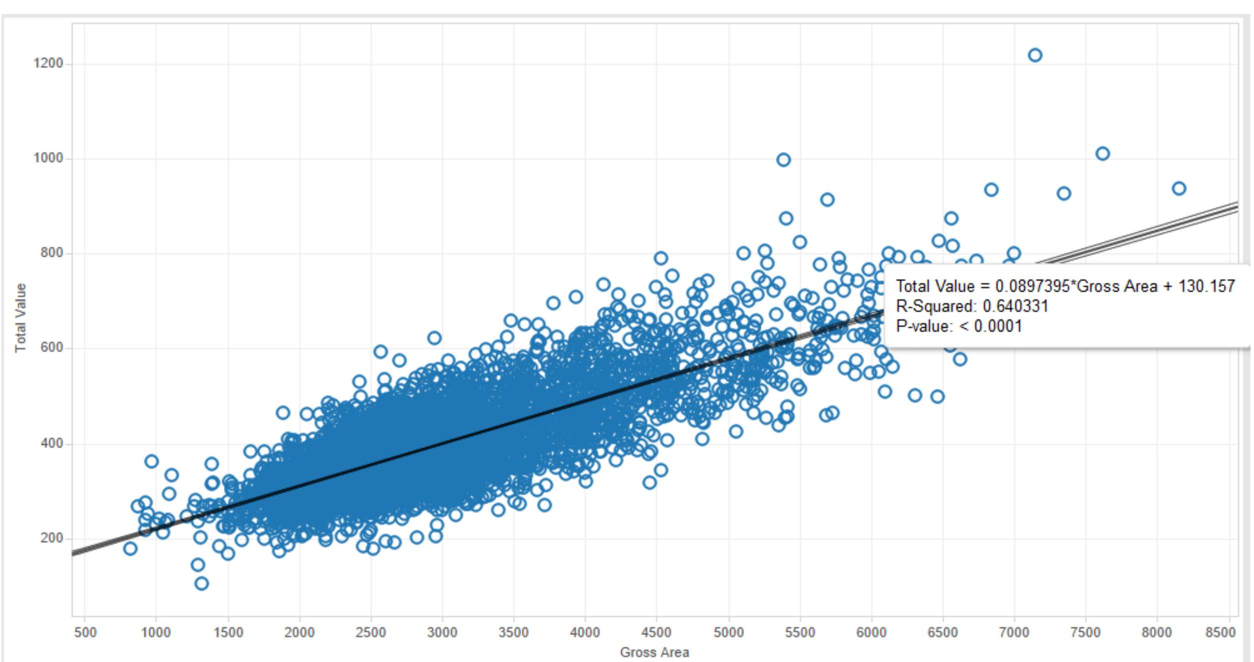


This Chart shows how the total price of the house depends on the Living area. The more the Living area of the house the price is higher.

Gross Area:

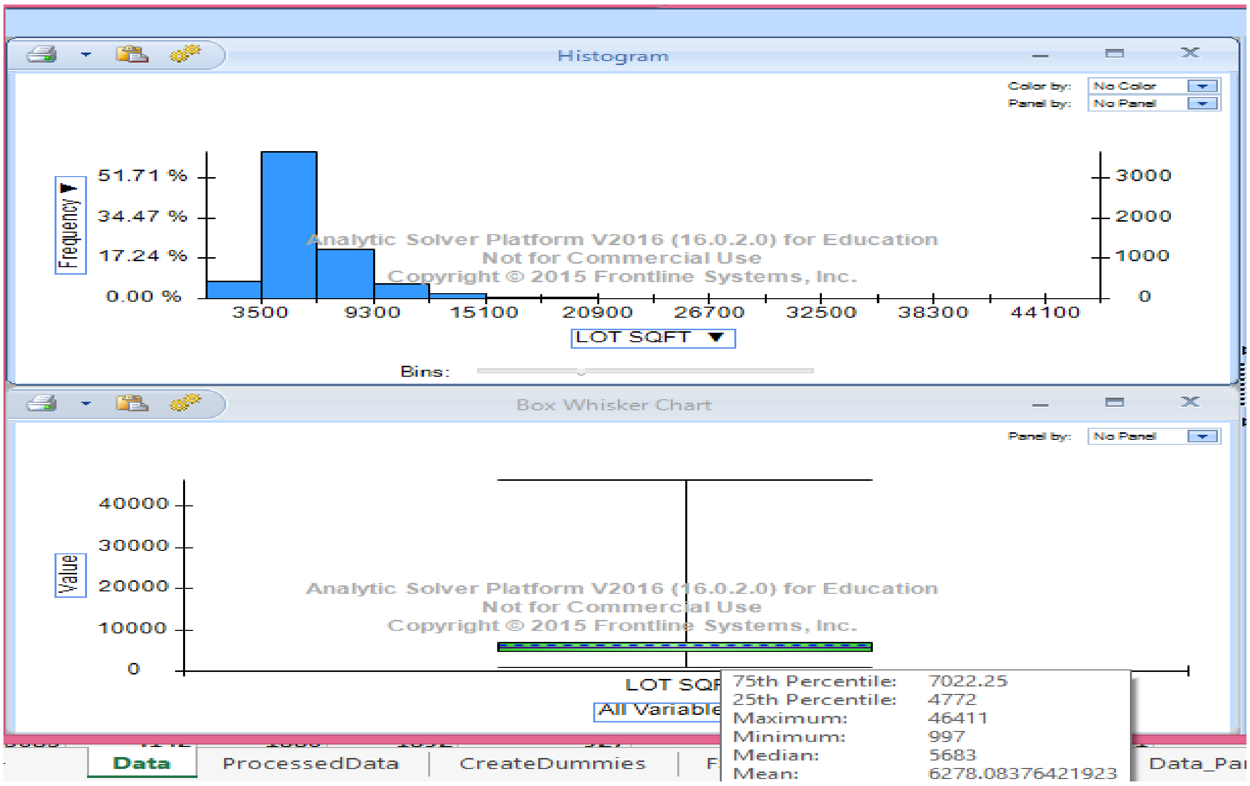


Scatter Plot in Tableau for Gross Area:

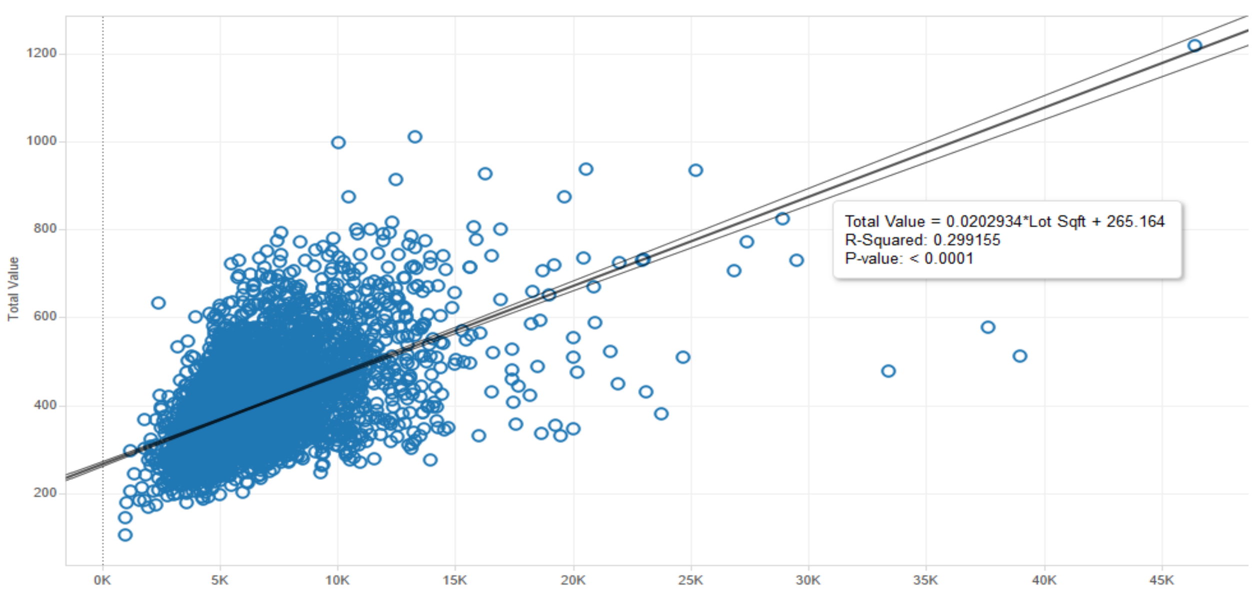


This Chart shows how the total price of the house depends on the Gross area. The more the gross area of the house the price is higher.

Lot Sqft:



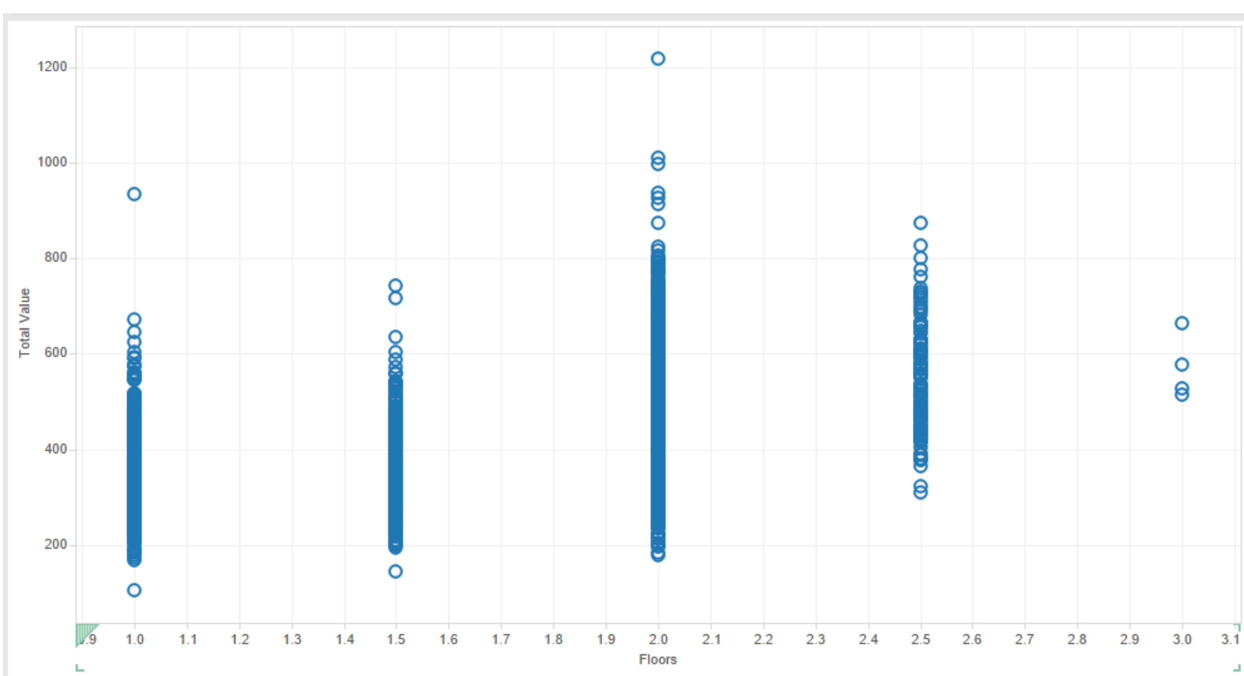
Scatter Plot in Tableau for Lot Sqft



This chart shows how the Lot Sqft effect the pricing of the house. There are a lot of houses of 1000-15K sqft and they are priced between 100$ to 800$. As the size of the Lot increases the price increases too.

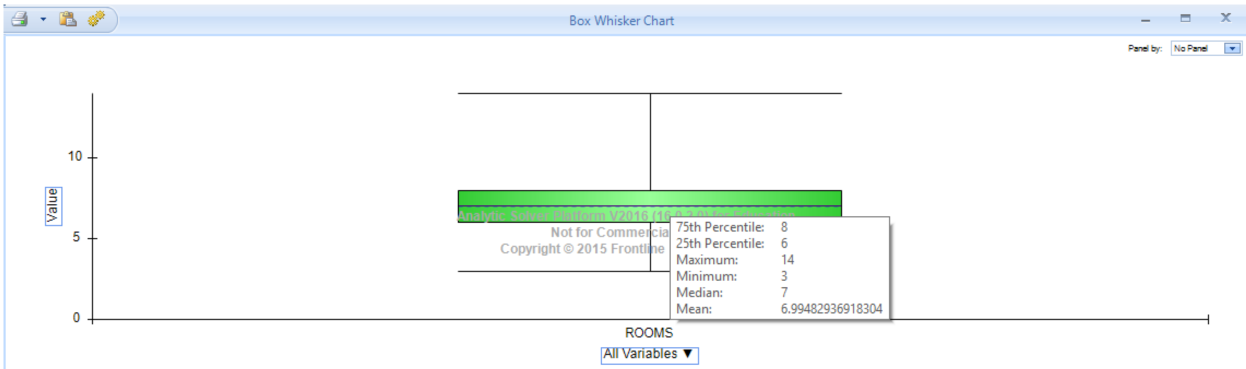
Floors:

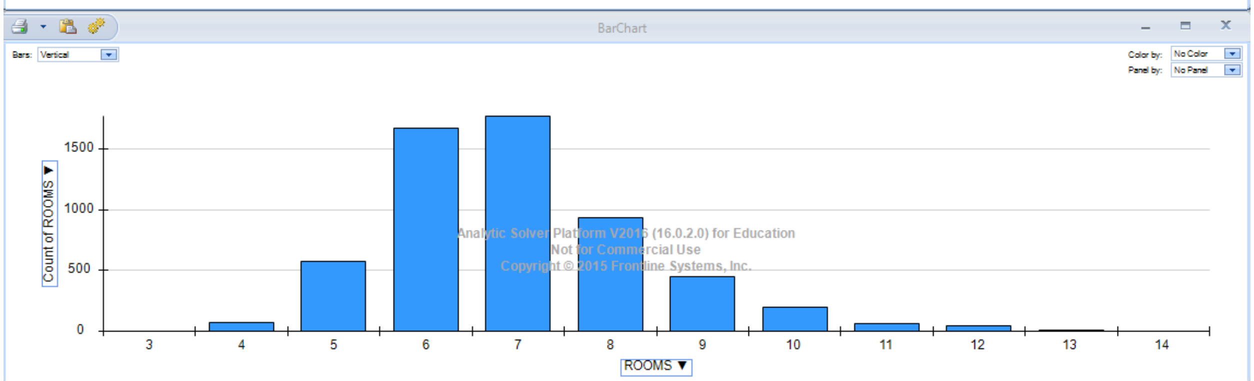
Scatter Plot on Tableau for Floor:



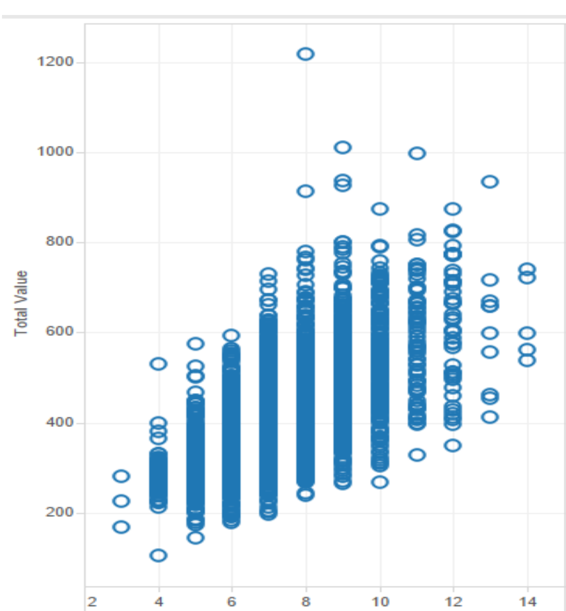
A very few houses in this dataset have 3 floors less than 1%. So this is an outlier.

Rooms:



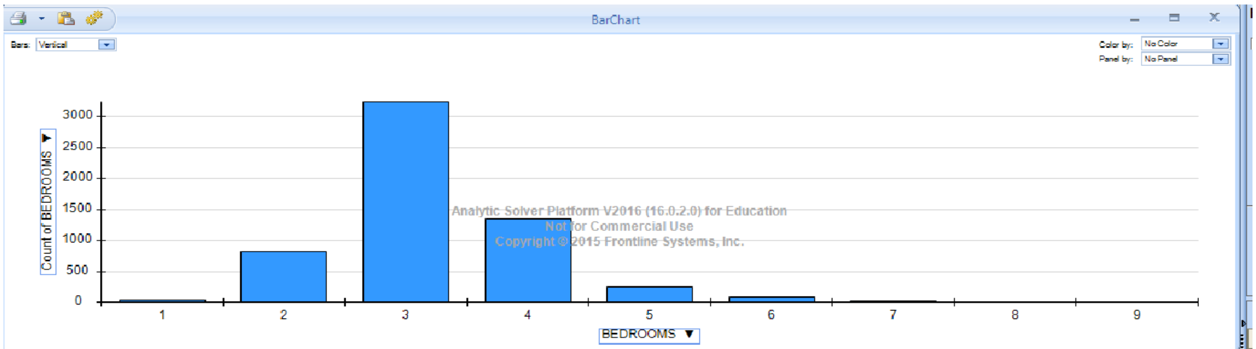


Scatter Plot on tableau for Room:

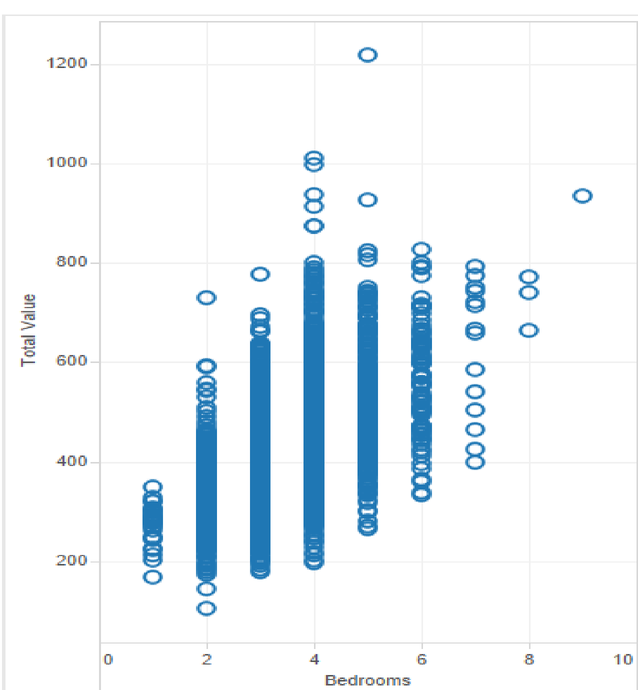


We can observe here that very few houses have 14 rooms. Most of the houses in this neighborhood has around 3-8 rooms in total.

Bedrooms:

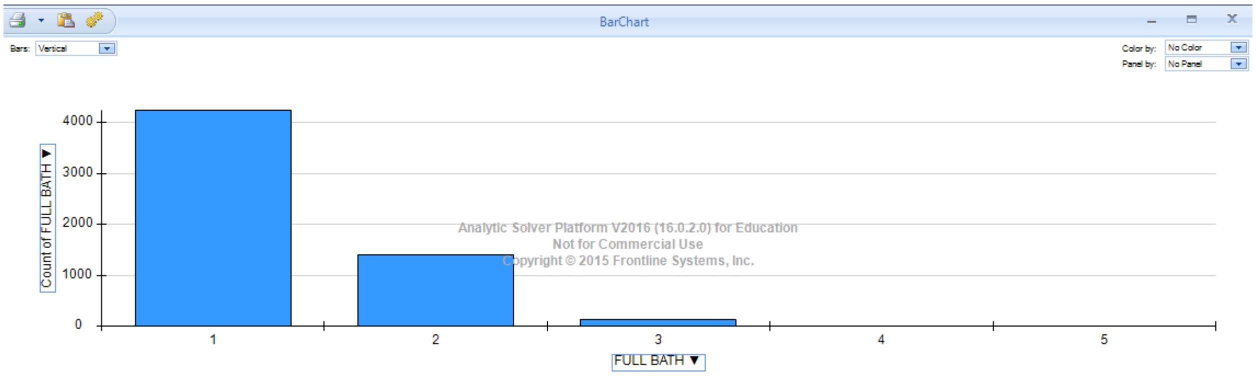


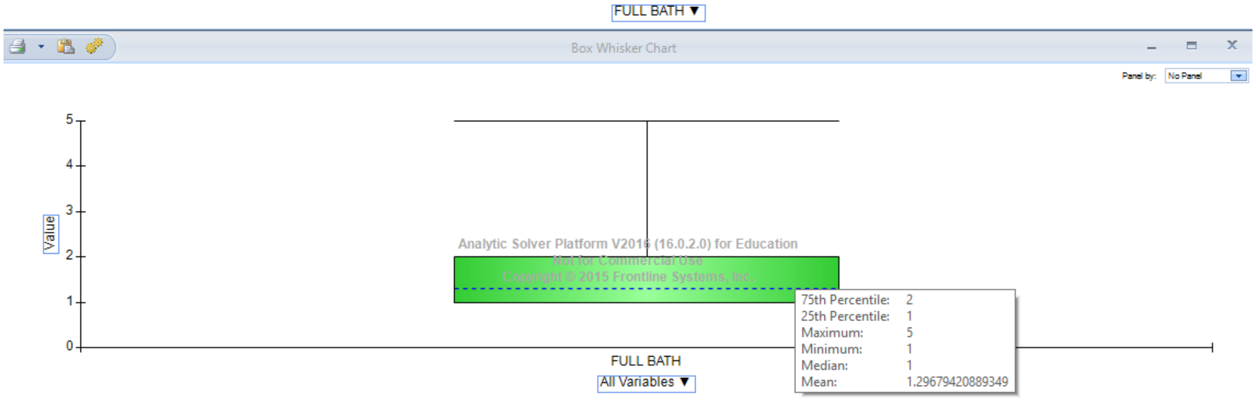
Scatter Plot in Tableau for Bedrooms:



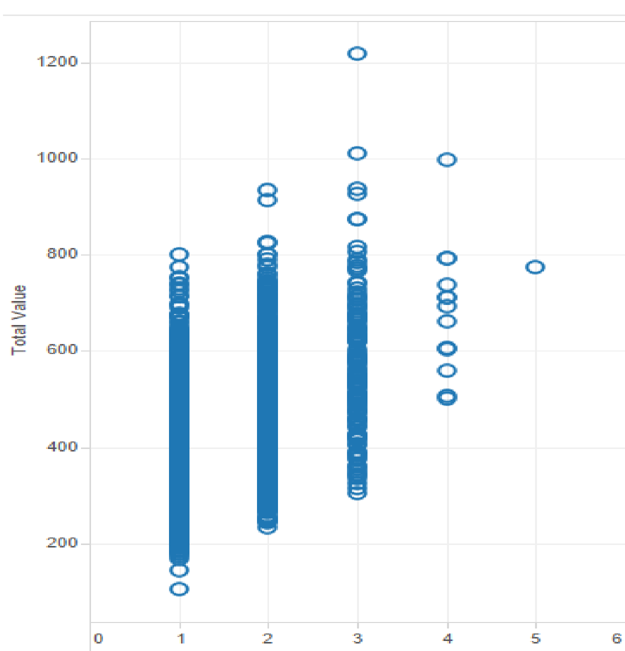
There is just one house with 9 bedrooms in this dataset. Also houses with 8 bedrooms are very less relatively.

Full Bath





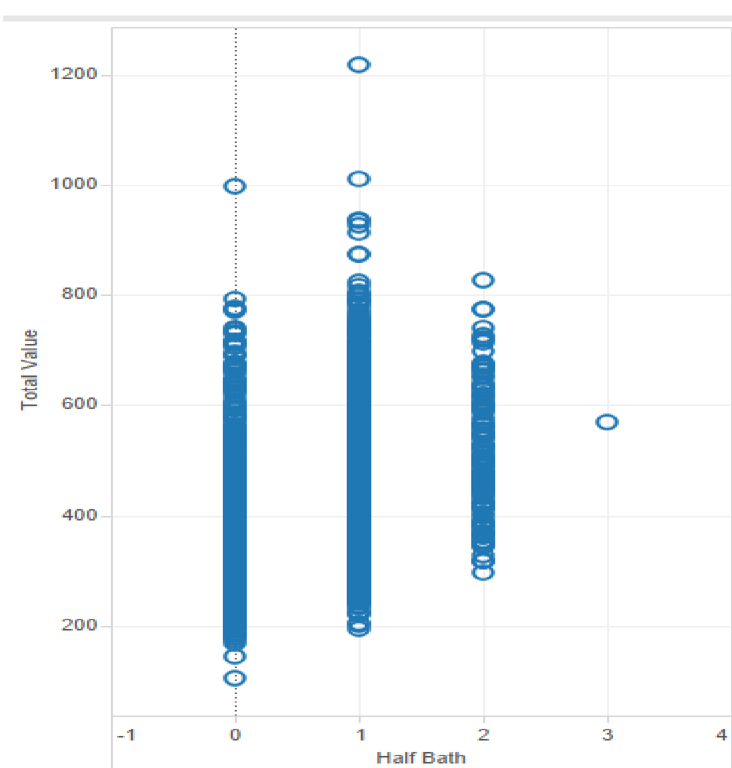
Scatter Plot in Tableau for Full Bath:



Here we can see very few houses have 4 Full Bath rooms and only 1 of them has 5 full baths. Also there are many houses with 1, 2 and 3 full baths. Houses with 3 full baths are priced a little higher.

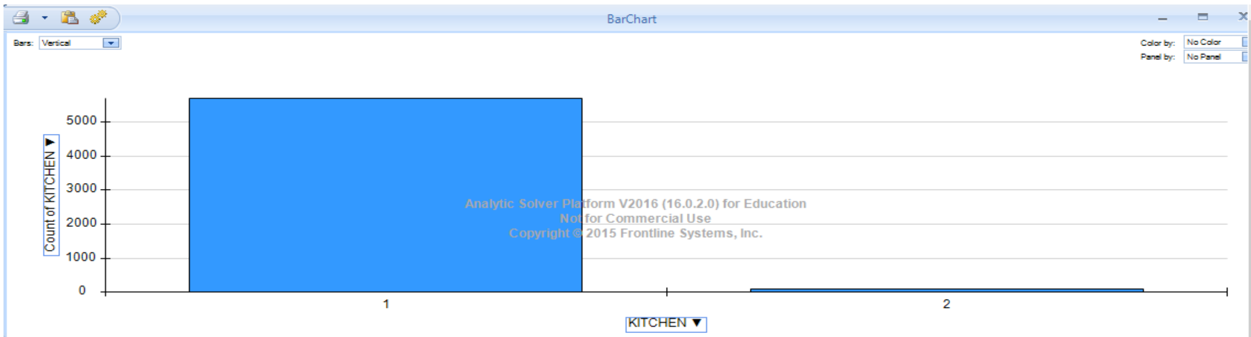
Half bath

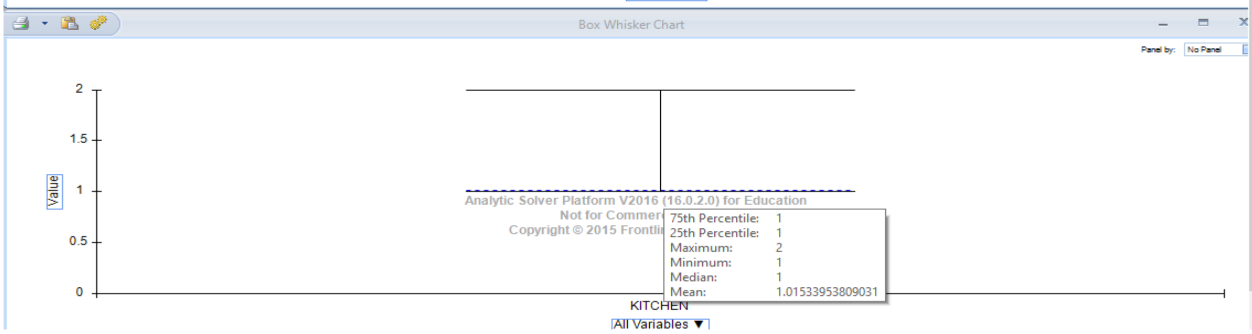
Scatter Plot in Tableau for Half bath



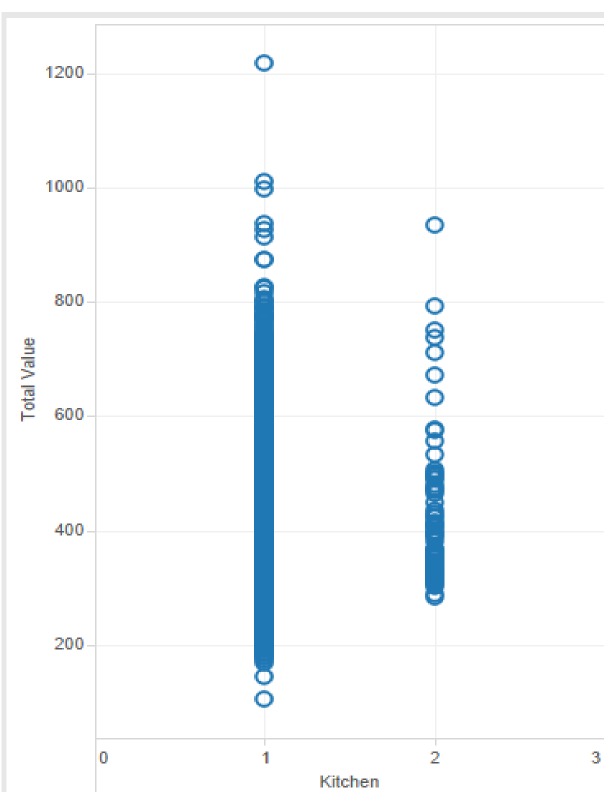
From looking at this we can find that there is only 1 data point which has 3 half baths, this may be an outlier.

Kitchen:



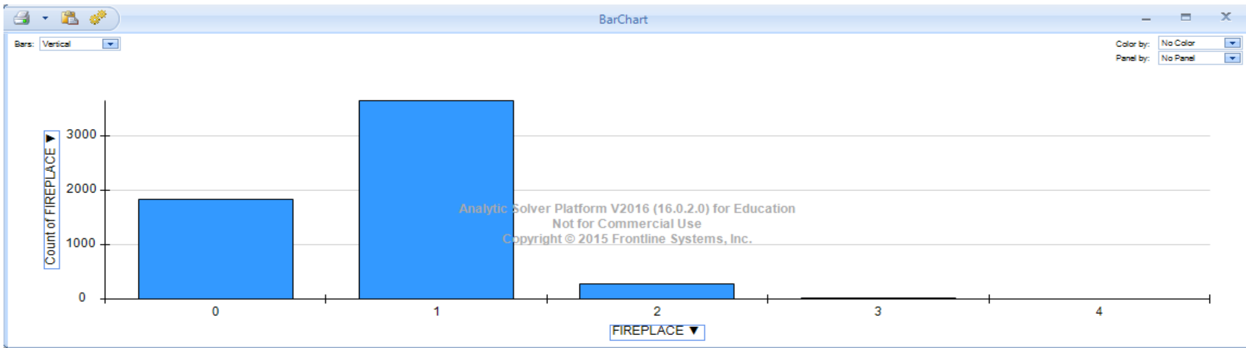


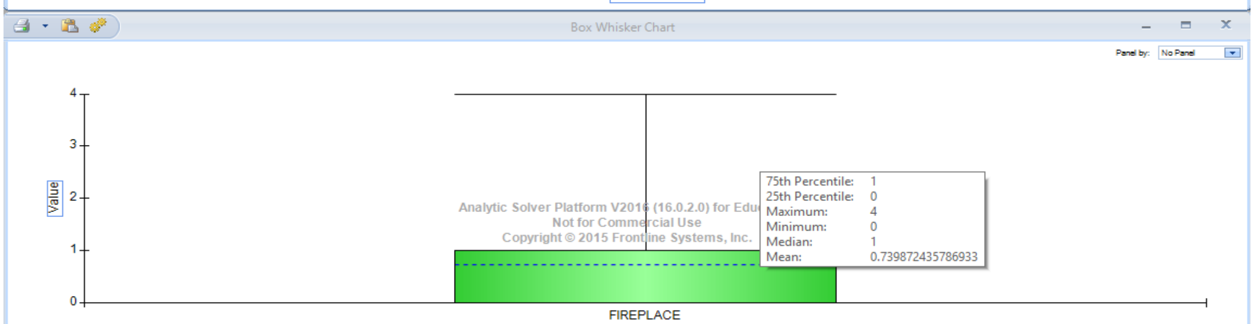
Scatter Plot in Tableau for Kitchen

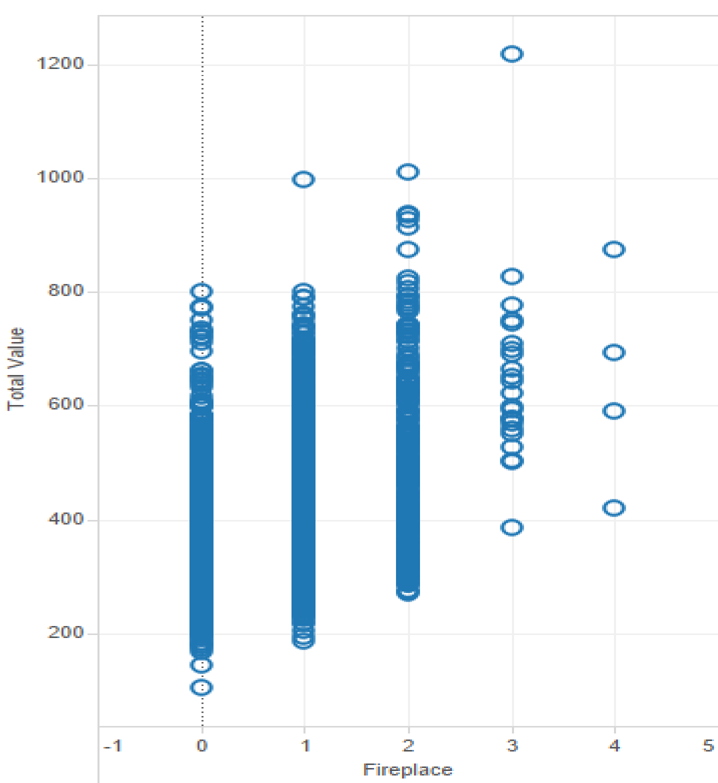


Here there are very few houses with 2 kitchens which can be seen using the Barchart. Only 90 observations have 2 kitchen out of 5890 total observation.

Fireplace:

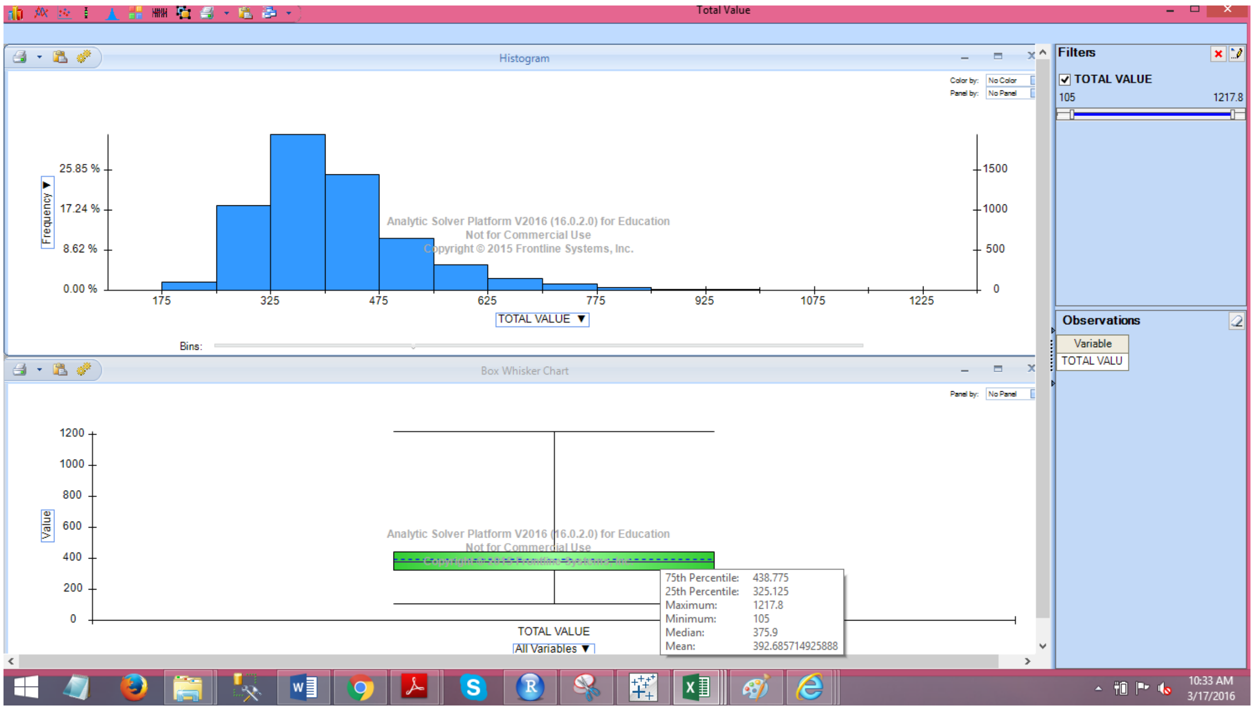


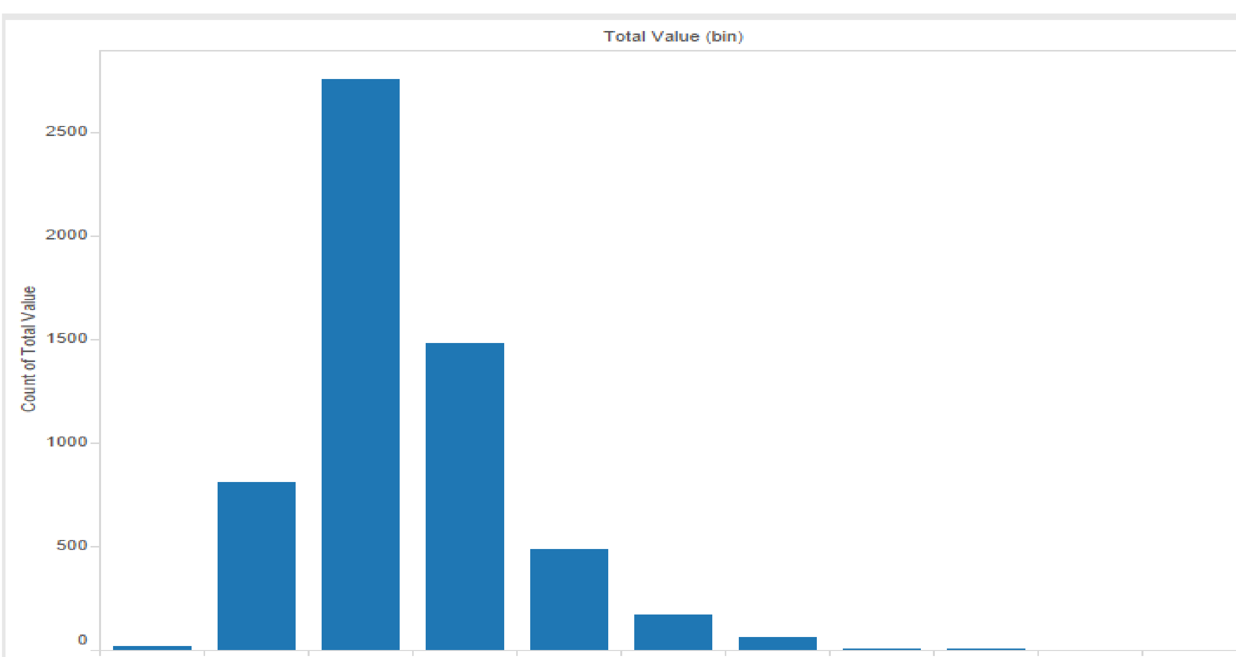




Here we can see there are very few (4) data points at fireplace=4 which means these are outliers.

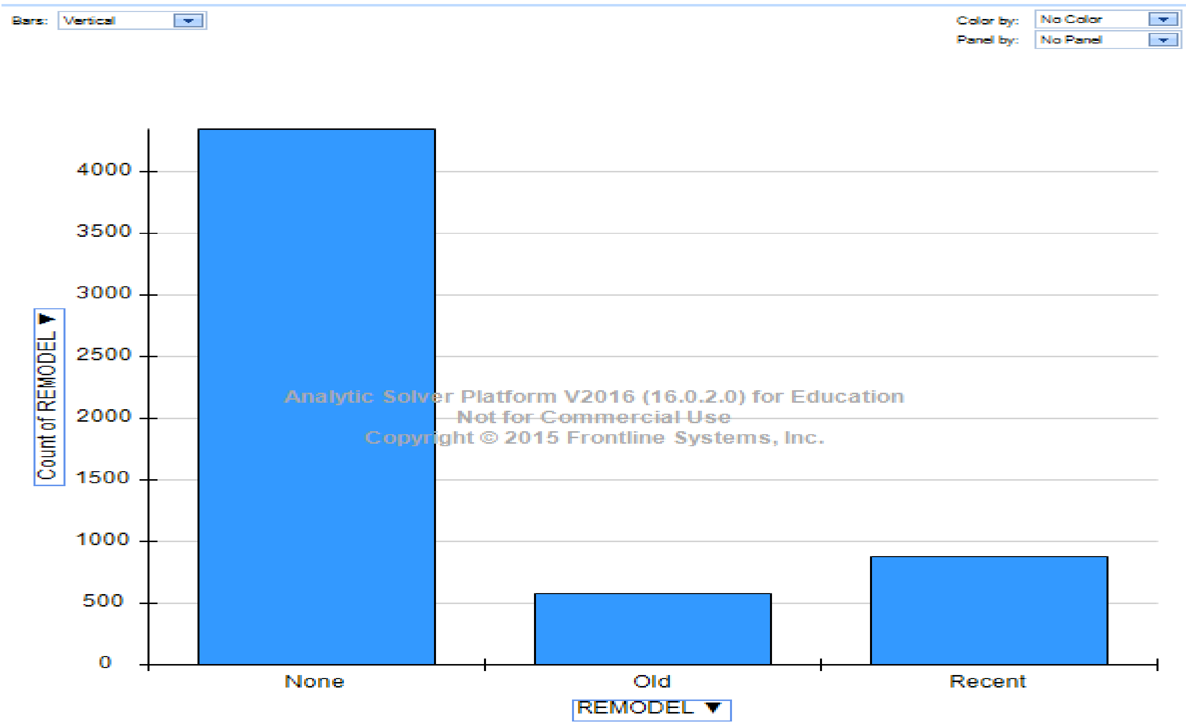
Summary of Total Value



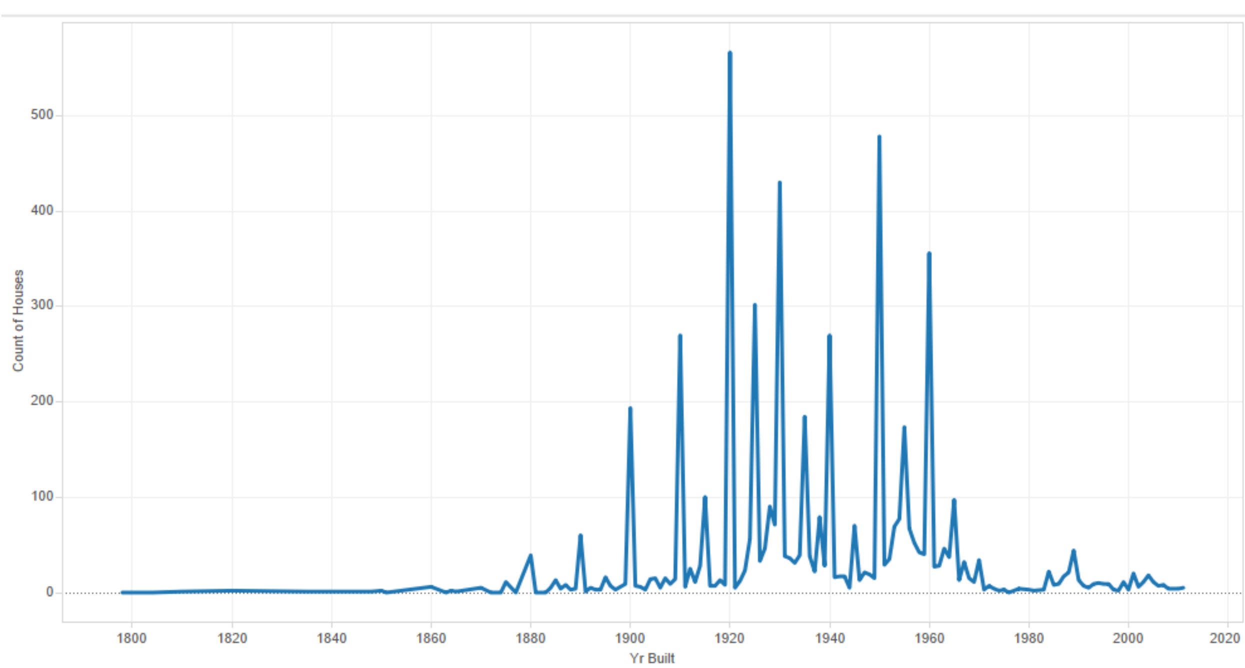


Most of the houses range from 200 to 600$ there are very few houses priced between 700-1000$ and only 1 house priced around 1200$.

Number of remodeled houses



Trend of Year in which houses were built

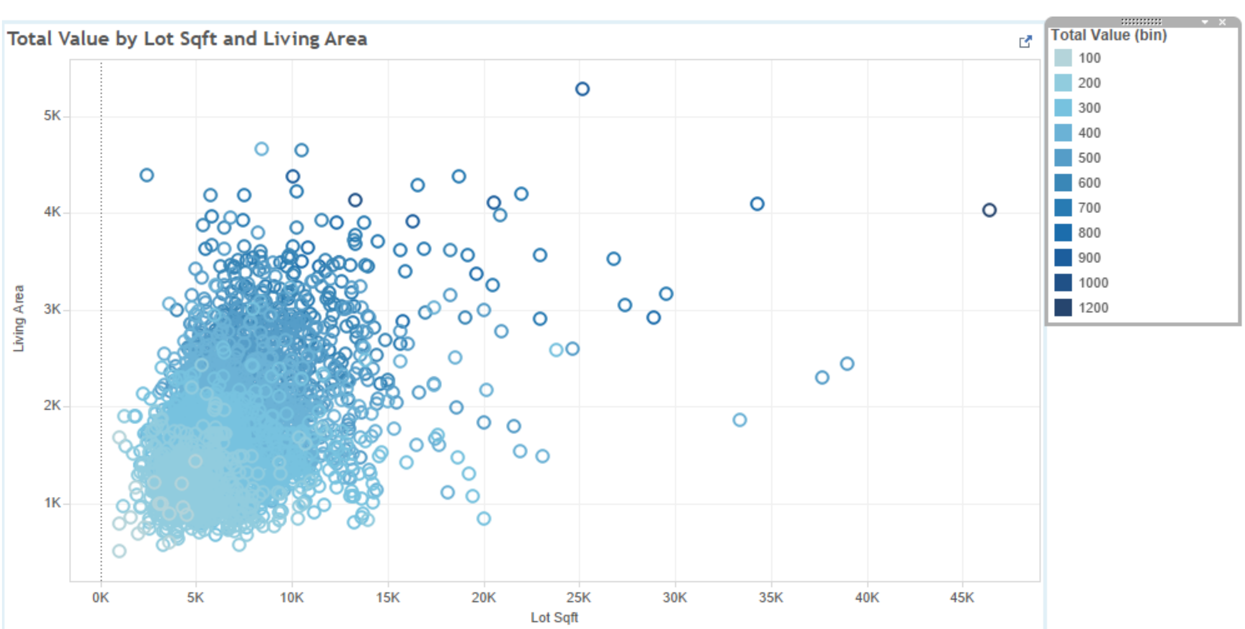


Below chart shows trend when most no. of houses were build. We can see here that there was increase in the number of houses build starting from 1880 and it increased a lot in and around the mid 90's and it went down again the early 2000's

Effect of Area on Total Value:

The chart shows when the Living Area and Gross area increases the total value increases too. The gradation in color shows the range of Total Value of the houses.

Effect of Lot Sqft and Living Area on Total Price



More Charts in the Problem1.twb show detailed Exploration of Data.

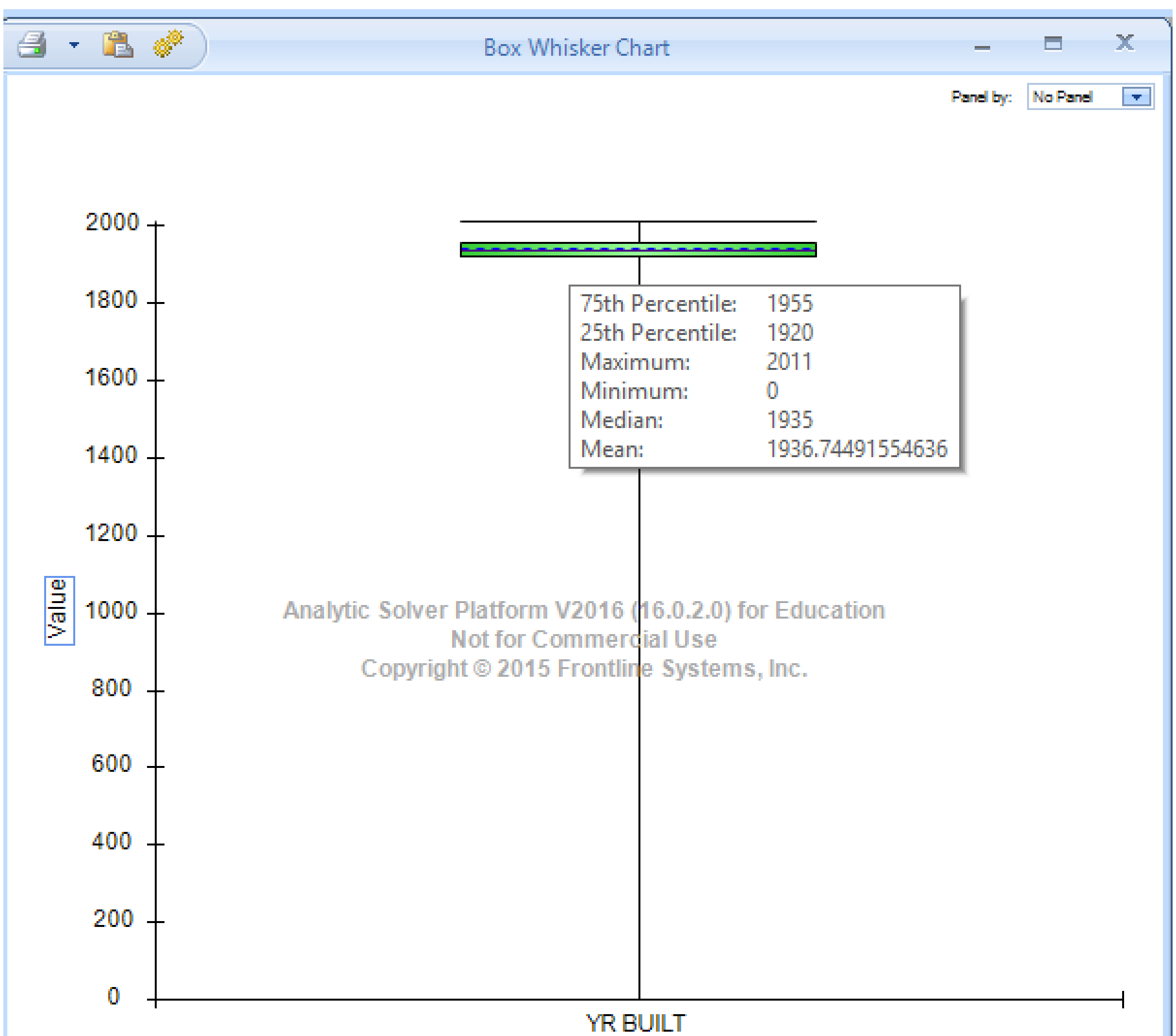
1.2. Building Prediction model using Multiple Linear Regression, CART and Random Forest

Tool: XLMiner

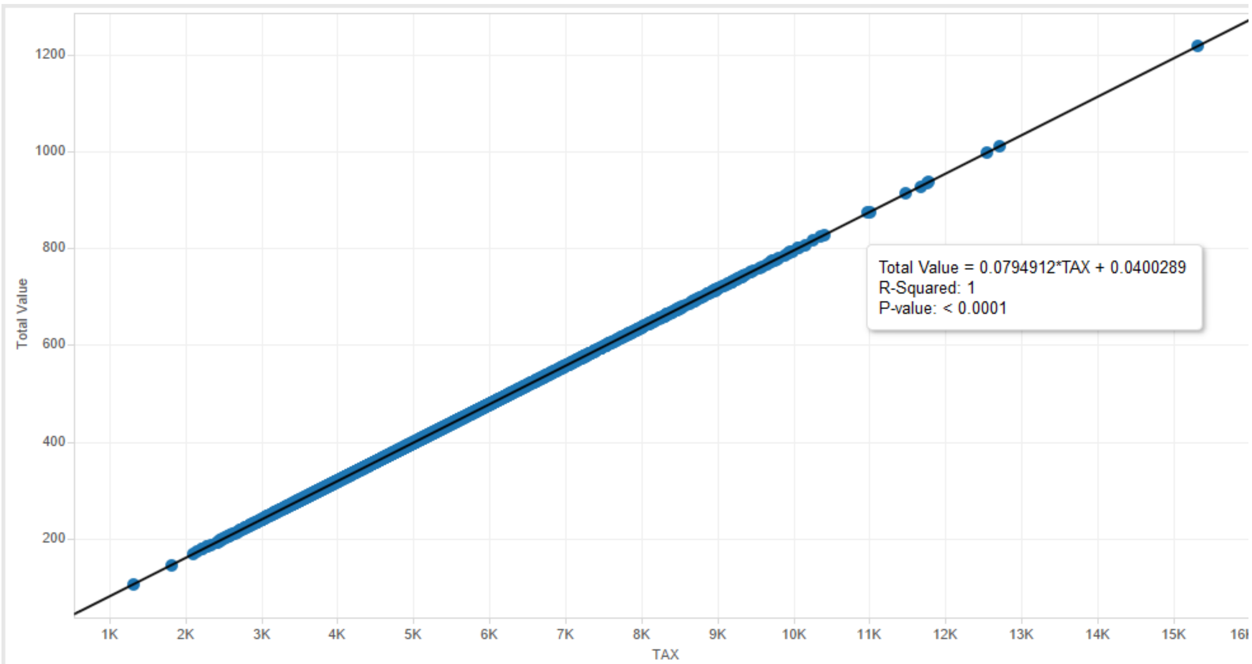
Steps:

1. **Data Pre-Processing:**

After carefully examining the dataset we find a row where the Year built is 0. Since we cannot randomly select a year and replace 0 with it, since many other predictors may be related to year. Also there is only one such row with year 0 so we can simply ignore this row and delete it.



Also the column Tax is calculated from the Total Value of the house. Tax is 12.5 times the Total Value of the house. Hence it is not an independent variable, it is directly dependent on Total Value. So we delete this column as we do not need it for the prediction.



The above chart shows the relation between Tax and Total amount.

2. **Creating Dummies for Categorical Variables:**

The Predictor Remodel (categorical feature) has 3 category which are NONE, OLD and RECENT. We create Dummies for this predictor since otherwise we cannot use them directly for Regression.

Similarly there are other variables like

ROOM (14 Categories)

BEDROOM (9 Categories)

KITCHEN (2 Categories)

FLOORS (5 Categories)

FULL BATH (5 Categories)

HALF BATH (4 Categories)

FIREPLACE (5 Categories)

Now we will have 49 variable in total now of which 8 of them would not be useful since if there are n dummy variables n-1 are only useful. So from these 41 variables we need to determine which of them matter how much.

We can determine this by Feature Selection feature of XLMiner. By using Feature selection we can see the Co-relation and the P value of the independent variables on Total Value. This Feature Selection uses Pearson, Spearman and Kendall Correlation for determining the Correlation between the variables. We will consider the values of Pearson Correlation for this problem. Higher Correlation and lower P values are considered to build a good model.

1.3. Building Model and its evaluation

2. Feature Selection

	Pearson Correlation		Spearman Correlation		Kendall Correlation	
Feature Identifier	Pearson Rho	Pearson: P-Value	Spearman Rho	Spearman: P-Value	Kendall Tau	Kendall: P-Value
LOT SQFT	0.5462	0	0.5208	0	0.3679	0
GROSS AREA	0.8004	0	0.7152	0	0.5294	0
LIVING AREA	0.837	0	0.7838	0	0.5974	0
FLOORS_2	0.3976	4.9709E-219	0.4815	0	0.3933	0
FLOORS_1	-0.3869	1.9121E-206	-0.4561	4.161E-296	-0.3725	0
FULL BATH_1	-0.3783	9.4494E-197	-0.3309	3.1261E-148	-0.2702	3.7168E-209
HALF BATH_0	-0.3216	1.1738E-139	-0.3624	1.3083E-179	-0.296	1.4946E-250
BEDROOMS_2	-0.3063	2.8863E-126	-0.3601	4.094E-177	-0.2941	2.6864E-247
ROOMS_5	-0.295	7.9246E-117	-0.3521	6.5872E-169	-0.2875	1.5061E-236
BEDROOMS_5	0.2846	1.7178E-108	0.2398	1.21472E-76	0.1958	8.6513E-111
FULL BATH_3	0.2818	2.4168E-106	0.194	2.50463E-50	0.1585	3.16508E-73
BEDROOMS_4	0.2803	3.2043E-105	0.2999	6.695E-121	0.245	2.9608E-172
ROOMS_9	0.2803	3.2964E-105	0.2692	6.63907E-97	0.2199	3.4048E-139
ROOMS_6	-0.2799	6.8811E-105	-0.2865	5.0914E-110	-0.234	2.4289E-157
FULL BATH_2	0.2736	4.2744E-100	0.2637	6.69306E-93	0.2154	1.2847E-133
FIREPLACE_0	-0.2673	1.74517E-95	-0.2904	3.8664E-113	-0.2372	1.2491E-161
ROOMS_10	0.2646	1.6714E-93	0.2167	1.29431E-62	0.177	7.34926E-91
HALF BATH_1	0.264	4.50932E-93	0.3114	1.312E-130	0.2543	1.875E-185
FIREPLACE_2	0.2465	5.21565E-81	0.1778	2.18338E-42	0.1452	9.65482E-62
REMODEL_Recent	0.2282	2.04587E-69	0.2379	2.00091E-75	0.1943	4.5409E-109
BEDROOMS_6	0.2263	2.86467E-68	0.1678	6.89942E-38	0.137	3.3354E-55
FLOORS_2.5	0.2229	3.37217E-66	0.1721	8.62176E-40	0.1405	5.68793E-58
REMODEL_None	-0.2173	5.94496E-63	-0.2267	1.58532E-68	-0.1852	2.7583E-99
BEDROOMS_3	-0.2025	9.44958E-55	-0.1388	2.40652E-26	-0.1134	2.47483E-38
ROOMS_11	0.2017	2.65436E-54	0.1443	2.25204E-28	0.1179	2.62002E-41
ROOMS_12	0.1988	8.33954E-53	0.1244	1.87048E-21	0.1016	3.81991E-31
HALF BATH_2	0.1787	8.07839E-43	0.1566	3.59906E-33	0.1279	2.5186E-48
ROOMS_8	0.1759	1.60406E-41	0.2192	4.57113E-64	0.179	6.21266E-93
FLOORS_1.5	-0.1679	6.35363E-38	-0.1794	3.5515E-43	-0.1466	6.93312E-63
FIREPLACE_3	0.1635	4.73192E-36	0.0965	1.72898E-13	0.0788	2.1626E-19
FULL BATH_4	0.1385	3.02265E-26	0.0762	6.08963E-09	0.0623	1.14473E-12
ROOMS_4	-0.1253	9.91787E-22	-0.1492	3.20488E-30	-0.1218	5.17205E-44
FIREPLACE_1	0.1244	1.94304E-21	0.1873	6.29122E-47	0.1529	2.56707E-68
YR BUILT	-0.1193	7.65107E-20	-0.1962	1.97381E-51	-0.1389	1.08799E-56
BEDROOMS_7	0.115	1.52817E-18	0.0706	7.20408E-08	0.0577	4.4164E-11
ROOMS_13	0.0919	2.28179E-12	0.0586	7.87751E-06	0.0479	4.5258E-08
BEDROOMS_1	-0.0864	4.27318E-11	-0.1015	9.19261E-15	-0.0829	2.82969E-21
BEDROOMS_8	0.0763	5.89576E-09	0.0386	0.003303847	0.0315	0.000321249
BEDROOMS_9	0.0718	4.3186E-08	0.0227	0.083744104	0.0185	0.034164498
ROOMS_14	0.0709	6.40534E-08	0.0467	0.000378241	0.0381	1.3482E-05
FIREPLACE_4	0.0666	3.82887E-07	0.0372	0.004634236	0.0304	0.000525972
FULL BATH_5	0.0504	0.000121523	0.0225	0.086012107	0.0184	0.035486274
FLOORS_3	0.047	0.000343286	0.0392	0.002858746	0.032	0.000260128
REMODEL_Old	0.0418	0.001437241	0.044	0.000808437	0.0359	4.10496E-05
ROOMS_3	-0.0386	0.003295764	-0.0373	0.004458915	-0.0305	0.000497249
HALF BATH_3	0.0233	0.076308527	0.02	0.126938464	0.0164	0.061560241
KITCHEN_1	-0.0183	0.163248566	-0.0018	0.888946458	-0.0015	0.864184526
KITCHEN_2	0.0183	0.163248566	0.0018	0.888946458	0.0015	0.864184526
ROOMS_7	-0.0064	0.627563643	0.0656	5.65872E-07	0.0536	9.28344E-10

3. Partition the Data

Next Step would be to partition the data

We create a random standard partition of 60% Training data and 40% for Validation

XLMiner: Data Partition Sheet

Date: 17-Mar-2016 12:44:45

Output Navigator		
Training Data	Validation Data	All Data

Elapsed Times in Milliseconds		
Partitioning Time	Report Time	Total
15	93	108

Data	
Data Source	$B$20:$AY$5821
Selected Variables	TOTAL VA, LOT SQFT, YR BUILT, GROSS AR, LIVING AR, FLOORS_1, FLOORS_1, FLOORS_2, FLOORS_2, FLOORS_3, ROOMS_1, ROOMS_1, ROOMS_
Partitioning Method	Randomly Chosen
Random Seed	12345
# Variables	50
# Training Rows	3481
# Validation Rows	2320
# Test Rows	0

4. Build Models

a. We build a model using Multiple Linear Regression, CART and Random Forest. For the Multiple Linear Regression we use the process of Backward Variable Selection to find out the best subsets of variables which would give us a good model.
The goal of the model is to have the RMSE value as low as possible. The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed.
So a good model would be the one with low RMSE value and high R-squared and adjusted R squared value. Also for a good model the RMSE of Training and Validation should be close to each other.

b. The most important factor in predicting the total value in any prediction problem is to select the variables correctly and find the best regression equation for prediction using these features.

a. Regression Model:

Using feature selection we can select the top variables and perform the regression. But how would we know how many variables will create best suitable model. For this we can use Variable Selection. We have used Backward Variable Selection as there are many variables to select from.

Variable Selection

Subset L	#Coeffs	RSS	Cp	R²	Adjusted	Probabili	Model 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Choose S	40	6E+06	40	0.8293	0.8274	1		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS	FLOORS	FLOORS	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROO
Choose S	39	6E+06	38	0.8293	0.8274	0.9925		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS	FLOORS	FLOORS	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROO
Choose S	38	6E+06	36.004	0.8293	0.8275	0.998		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS	FLOORS	FLOORS	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROO
Choose S	37	6E+06	34.019	0.8293	0.8275	0.9993		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5		FLOORS	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROOMS_	ROO
Choose S	36	6E+06	32.084	0.8293	0.8276	0.9991		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5		FLOORS	ROOMS_	ROOMS_	ROOMS_	ROOMS_13		ROOMS_	ROOMS_	ROOMS_	ROO
Choose S	35	6E+06	30.128	0.8293	0.8276	0.9997		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5		FLOORS	ROOMS_	ROOMS_	ROOMS_12			ROOMS_	ROOMS_	ROOMS_	ROO
Choose S	34	6E+06	28.857	0.8293	0.8276	0.9904		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5			ROOMS_	ROOMS_	ROOMS_12			ROOMS_	ROOMS_	ROOMS_	ROO
Choose S	33	6E+06	27.726	0.8292	0.8277	0.9734		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5			ROOMS_	ROOMS_11				ROOMS_	ROOMS_	ROOMS_	ROO
Choose S	32	6E+06	26.098	0.8292	0.8277	0.9778		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5			ROOMS_	ROOMS_11				ROOMS_	ROOMS_5		ROO
Choose S	31	6E+06	24.644	0.8292	0.8277	0.9767		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5				ROOMS_11				ROOMS_	ROOMS_5		ROO
Choose S	30	6E+06	23.337	0.8292	0.8277	0.9723		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5				ROOMS_11				ROOMS_	ROOMS_5		ROO
Choose S	29	6E+06	21.831	0.8291	0.8277	0.9745		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5				ROOMS_11				ROOMS_	ROOMS_5		ROO
Choose S	28	6E+06	20.132	0.8291	0.8278	0.9809		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5				ROOMS_11				ROOMS_	ROOMS_5		ROO
Choose S	27	6E+06	18.379	0.8291	0.8278	0.9864		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5				ROOMS_11				ROOMS_	ROOMS_5		ROO
Choose S	26	6E+06	17.378	0.8291	0.8278	0.9797		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5				ROOMS_11				ROOMS_	ROOMS_5		ROO
Choose S	25	6E+06	17.303	0.829	0.8278	0.9485		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5				ROOMS_11				ROOMS_	ROOMS_5		ROO
Choose S	24	6E+06	17.241	0.8289	0.8277	0.9029		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5				ROOMS_11				ROOMS_	ROOMS_5		ROO
Choose S	23	6E+06	17.358	0.8288	0.8277	0.837		LOT SQF	GROSS	LIVING A	FLOORS	FLOORS_1.5				ROOMS_11				ROOMS_	ROOMS_5		ROO

We select the Subset with 38 variables because the Cp value is close to no. of variables and also the R square and adjusted R squared values are pretty good and close to each other.

Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	341.8001	73.54523198	4.647481524	3.49E-06	197.6034	485.9968	540722411.1
LOT SQFT	0.008139	0.000302966	26.86404982	1.8E-144	0.007545	0.008733	9905709.829
GROSS ARE	0.034385	0.002009154	17.11406157	4.43E-63	0.030446	0.038324	13367220.69
LIVING ARE	0.045571	0.003790601	12.02204116	1.22E-32	0.038139	0.053003	2075462.9
FLOORS_1	-19.986	22.47510898	-0.88925131	0.37393	-64.0519	24.07988	538912.9785
FLOORS_1.5	-24.6147	22.49953276	-1.09400732	0.274028	-68.7284	19.49913	601546.4937
FLOORS_2	18.64762	22.35552313	0.834139351	0.40426	-25.1838	62.47905	363.3039184
FLOORS_2.5	20.86512	22.93701173	0.909670488	0.36306	-24.1064	65.83665	2333.612651
ROOMS_10	37.23013	26.60390217	1.399423666	0.161776	-14.9309	89.39116	3007.866886
ROOMS_11	50.7475	26.99506404	1.879880832	0.060209	-2.18046	103.6755	3831.566435
ROOMS_12	22.58288	26.92708692	0.838667789	0.401714	-30.2118	75.37757	2211.403098
ROOMS_13	37.06544	30.79765198	1.203515248	0.22886	-23.3181	97.44897	761.7997038
ROOMS_3	10.67489	50.61772169	0.210892276	0.832984	-88.5689	109.9187	2496.148914
ROOMS_4	46.90407	27.61164414	1.698706317	0.089465	-7.23279	101.0409	21.06347428
ROOMS_5	41.84147	26.80019043	1.561237867	0.11856	-10.7044	94.38736	582.0540858
ROOMS_6	35.61222	26.6467656	1.336455727	0.181489	-16.6329	87.8573	32197.84289
ROOMS_7	40.7545	26.59335696	1.532506651	0.125489	-11.3859	92.89485	6863.083696
ROOMS_8	42.30001	26.57558447	1.591686927	0.111547	-9.8055	94.40552	35.66348858
ROOMS_9	42.00837	26.52447446	1.583758982	0.11334	-9.99693	94.01368	1098.782126
BEDROOMS	-72.4235	46.90263921	-1.5441238	0.12265	-164.383	19.53634	3198.400417
BEDROOMS	-65.2438	45.74060159	-1.42638802	0.153847	-154.925	24.43762	377.5222916
BEDROOMS	-66.9018	45.638513	-1.4659061	0.142765	-156.383	22.57953	0.072342554
BEDROOMS	-66.4258	45.58970215	-1.45703418	0.145198	-155.811	22.95985	12.83633856
BEDROOMS	-70.0757	45.68452433	-1.53390561	0.125145	-159.647	19.49577	297.0250913

Residual DF	3442
R^2	0.824166
Adjusted R^2	0.822225
Std. Error Estimate	41.61421
RSS	5960658

BEDROOMS	-66.9018	45.638513	-1.4659061	0.142765	-156.383	22.57953	0.072342554
BEDROOMS	-66.4258	45.58970215	-1.45703418	0.145198	-155.811	22.95985	12.83633856
BEDROOMS	-70.0757	45.68452433	-1.53390561	0.125145	-159.647	19.49577	297.0250913
BEDROOMS	-67.4938	45.66736517	-1.47794304	0.139515	-157.032	22.04411	111.9945148
BEDROOMS	-69.7072	47.88499772	-1.45572124	0.145561	-163.593	24.17868	401.1030943
BEDROOMS	-72.2877	53.05489116	-1.36250848	0.173127	-176.31	31.73452	3395.575761
FULL BATH_	-146.139	42.80600426	-3.41398239	0.000648	-230.067	-62.2112	238960.0203
FULL BATH_	-122.735	42.74944941	-2.87104088	0.004116	-206.552	-38.9186	74417.98148
FULL BATH_	-94.0224	42.88187268	-2.19259106	0.028404	-178.099	-9.94592	3030.358989
FULL BATH_	-82.2477	45.72738779	-1.79865276	0.072161	-171.903	7.407868	1702.996949
HALF BATH	-37.9237	4.989213935	-7.60113758	3.76E-14	-47.7058	-28.1416	350298.4891
HALF BATH	-18.2175	4.844523563	-3.76042318	0.000172	-27.7159	-8.71903	27214.04238
KITCHEN_1	15.82654	6.051833828	2.615163952	0.008957	3.960989	27.69209	14935.77834
FIREPLACE_	-23.4342	1.574452486	-14.8840189	1.32E-48	-26.5211	-20.3472	395805.8339
FIREPLACE_	9.980857	3.484152448	2.864644237	0.0042	3.149642	16.81207	9625.659441
FIREPLACE_	11.56441	12.78354142	0.904632942	0.365723	-13.4997	36.62851	1457.179784
REMODEL_	4.840523	2.449209909	1.976361155	0.048194	0.038472	9.642575	9.186142784
REMODEL_	25.87654	2.076742851	12.4601556	6.89E-35	21.80477	29.94831	268862.5413

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
5960658	41.38044	-8.04699E-13

Training Data Scoring - Summary Report

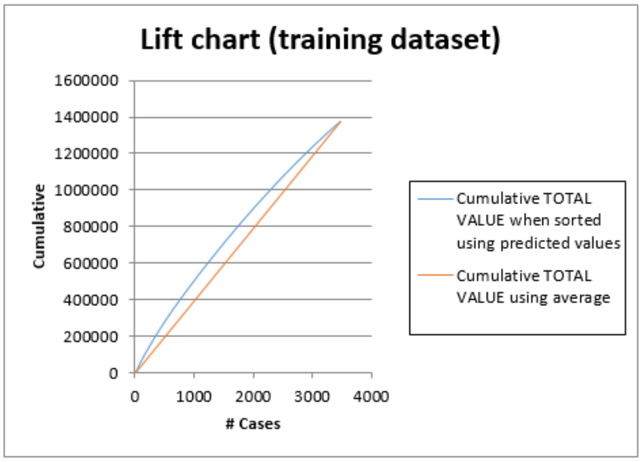
Total sum of squared errors	RMS Error	Average Error
5960658	41.38044	-8.04699E-13

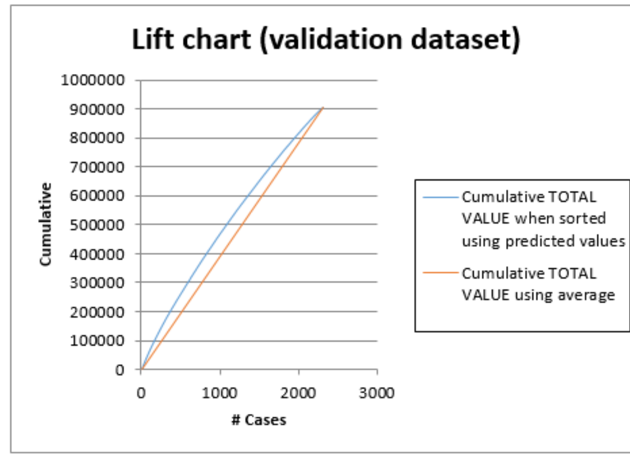
Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4189664	42.49576	-0.412080951

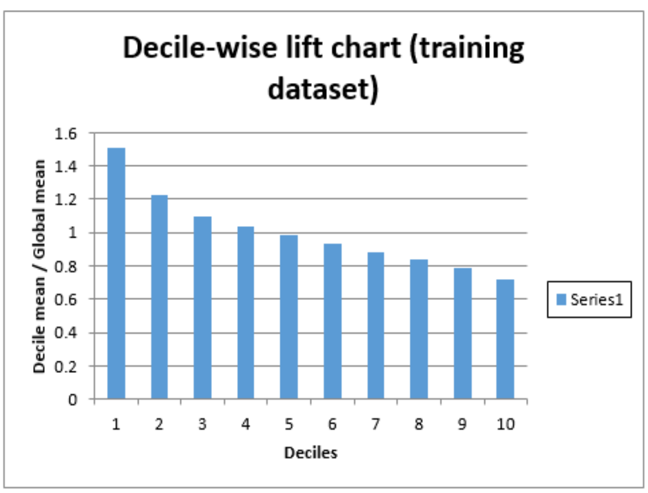
The RMSE we obtain for the Training is 41.38 and Validation dataset is 42.49 which are very much similar and also the average error is low. Also if we check the R-squared and adjusted R squared they are as high as 0.822 and are close to each other indicating this is a good model.

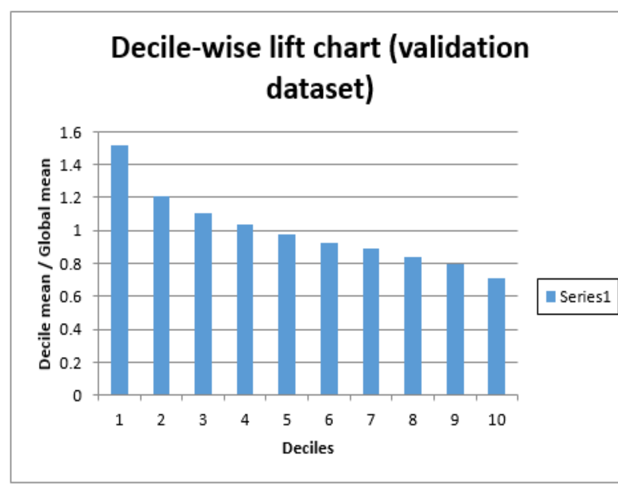
Lift Charts for Regression:



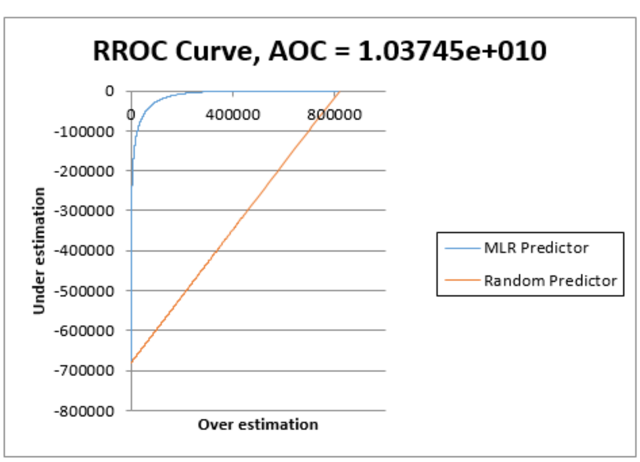


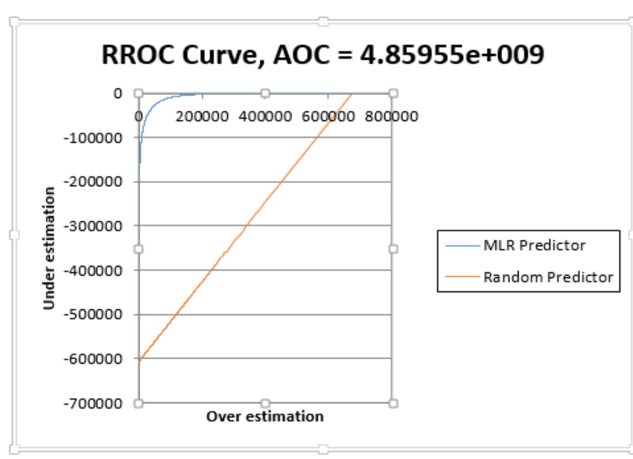
Decile wise lift chart for Regression





RROC Curve





The above charts show the Lift Chart, Decile Wise Lift Chart and the RROC charts for training and validation dataset of Multiple Linear Regression.

Lift Chart

Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.

The red line in the lift chart shows the baseline, which indicates the measure of effectiveness of any random model. The blue line shows the measure of effectiveness of our model which is above the baseline indicating our model is better than the baseline.

Decile Chart

After building a statistical model, a decile analysis is created to test the model's ability to predict the intended outcome. Each column in the decile analysis chart represents a collection of records that have been scored using the model. The height of each column represents the average of those records' actual behavior.

Ideal Situation: The Staircase Effect
When you're looking at a decile analysis, you want to see a staircase effect that is, you'll want the bars to descend in order from left to right.

Not-So-Ideal Situations

In contrast, if the bars seem to be out of order, the decile analysis is telling you that the model is not doing a very good job of predicting actual responses.

If the bars seem to be the same height, or the decile analysis looks "flat", the decile analysis is telling you that the model isn't performing any better than randomly binning people into deciles would. In both cases, your model should be improved before moving forward with it.

Our linear regression model follows the staircase effect to some extent which means it is a good model.

RROC Curve

RROC curves plot the performance of regressors by graphing over estimations (or predicted values that are too high) versus under estimations (or predicted values that are too low.) The closer the curve is to the top left corner of the graph (the smaller the area above the curve), the better the performance of the model. Area Over the Curve (AOC) is the space in the graph that appears above the ROC curve and is calculated using the formula: sigma2 * n2/2 where n is the number of records. The smaller the AOC, the better the performance of the model.

From above charts we can see that the AOC is very less which means that the model is a good model.

b. **CART Model**

Results for CART are as follows:

Training Data scoring - Summary Report (Using Full-Grown Tree)

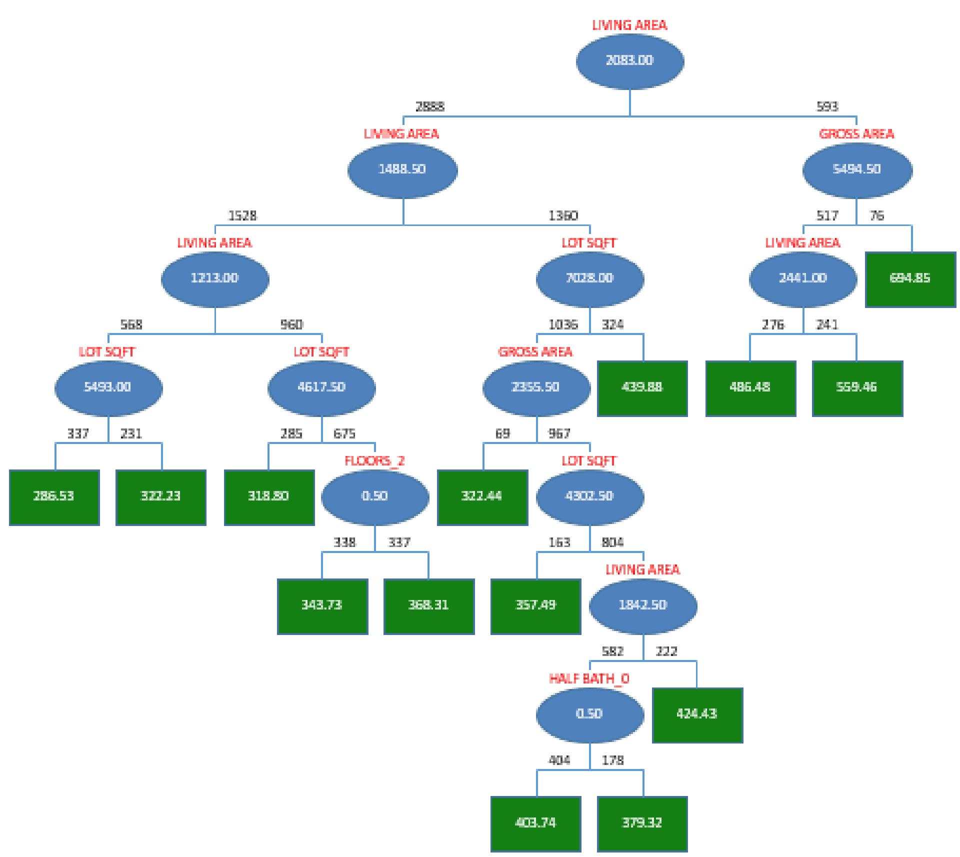
Total sum of squared errors	RMS Error	Average Error
8757306	50.15719	2.79563E-14

Validation Data scoring - Summary Report (Using Full-Grown Tree)

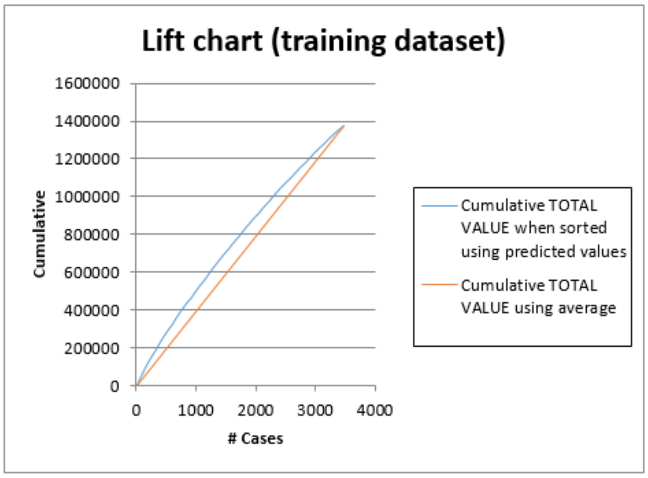
Total sum of squared errors	RMS Error	Average Error
7137641	55.46683	0.056711489

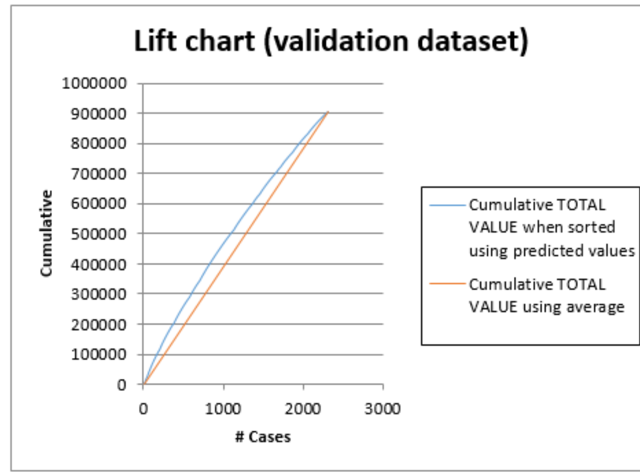
Here the RMSE is 50.15 for training dataset and 55.46 for Validation dataset. Also, only 5 variables are used by the Regression tree which can be seen below. The difference between RMSE for training and validation is greater in this model than in Multiple Linear Regression.

Full Grown Regression Tree

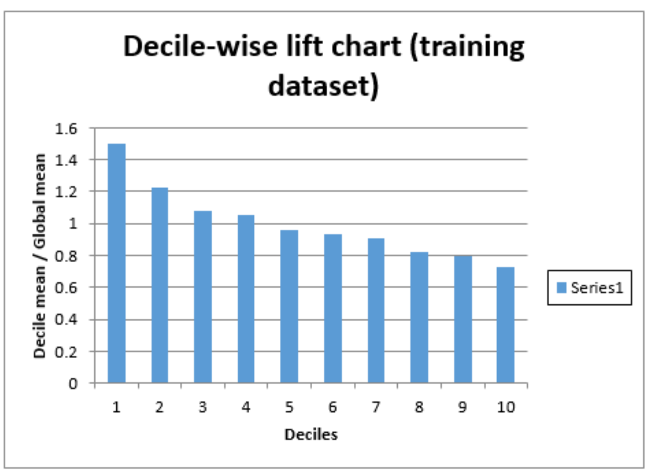


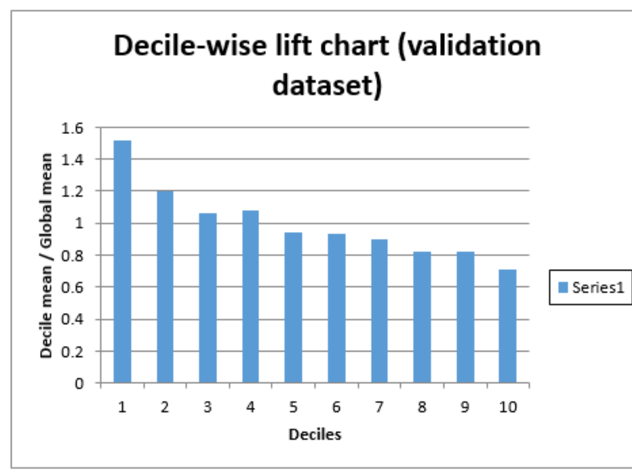
Lift Charts for CART



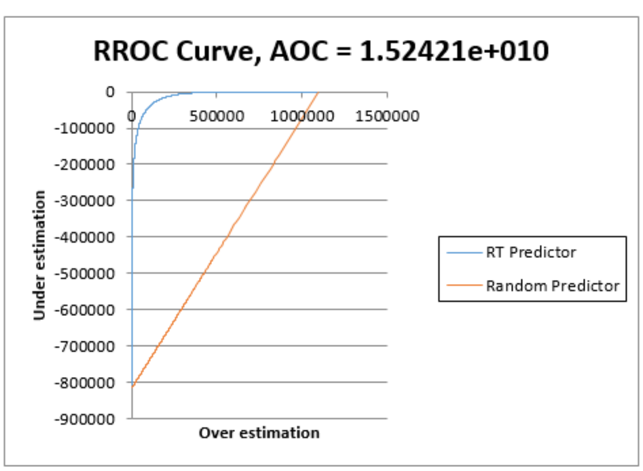


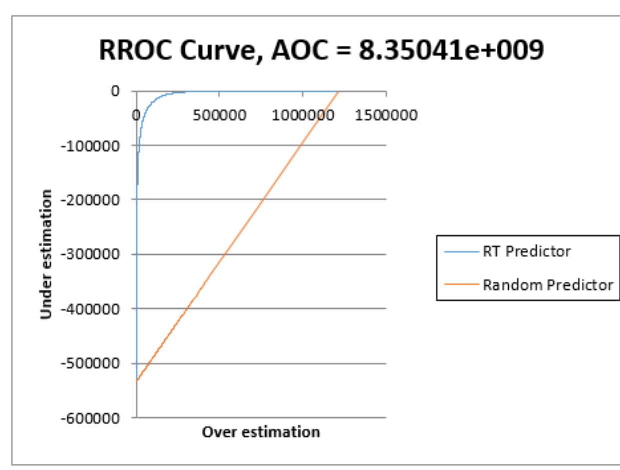
Decile wise Lift Chart for CART





RROC Curve





Lift Chart

The lift chart for Regression Tree is similar to the one of Multi Linear Regression so it is better than the baseline model. We cannot compare the Linear Regression with CART based on this lift charts as it is not giving us any significant difference.

Decile Chart

The decile chart for CART is little unstable which shows that the model is not very good.

RROC Curve

The value of the AOC is less in this model, but the AOC for multiple linear regression is the least amongst the 3 models.

c. Random Forest

Result for Random Forest are as follow:

Training Data scoring - Summary Report

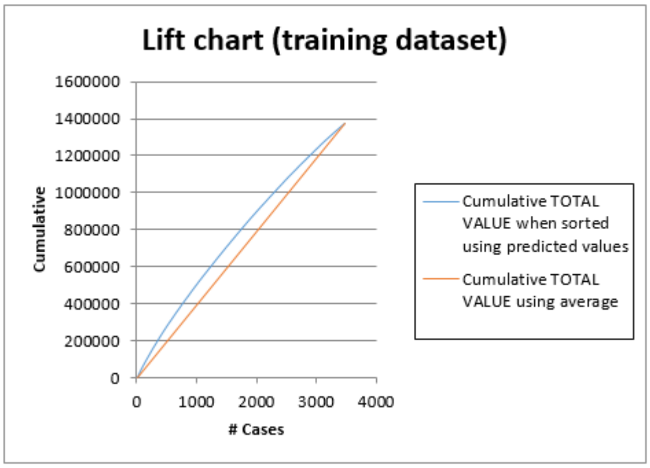
Total sum of squared errors	RMS Error	Average Error
7598308	46.72039	0.609269

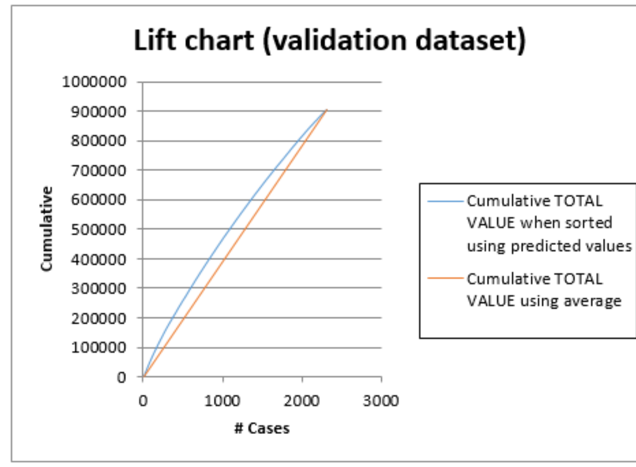
Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
5863802	50.27426	-0.9223

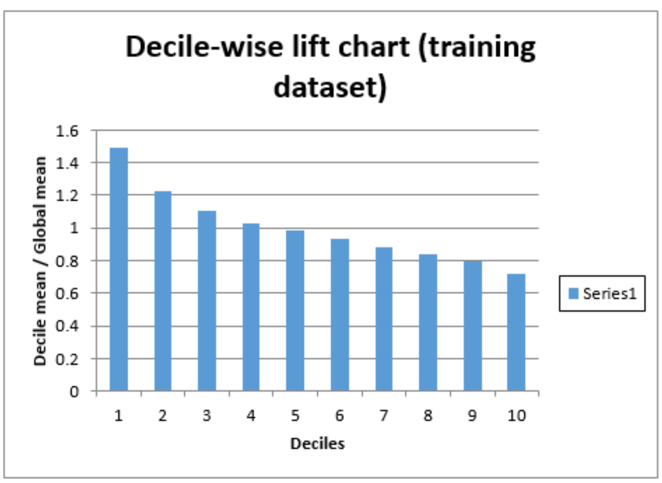
Here the RMSE for Training and Validation is better than CART since in CART only single tree is used and Random Forest creates many such trees and gives best output. The RMSE for Training and validation still differ more than that in Linear Regression. Also the RMSE for Multiple Linear Regression was lower than that in Random Forest.

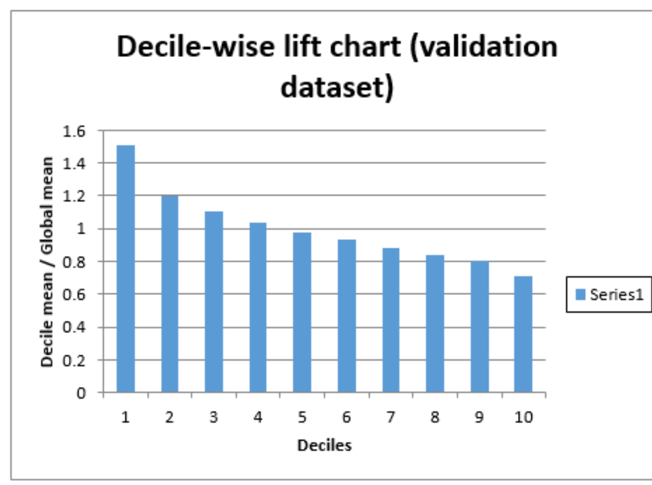
Lift Charts for Random Forest:



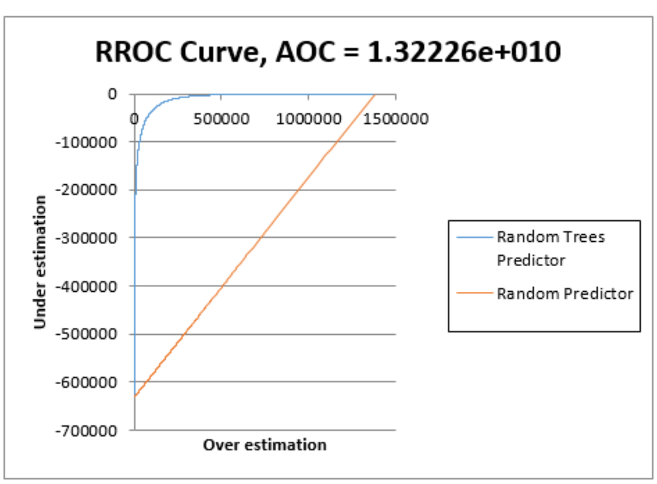


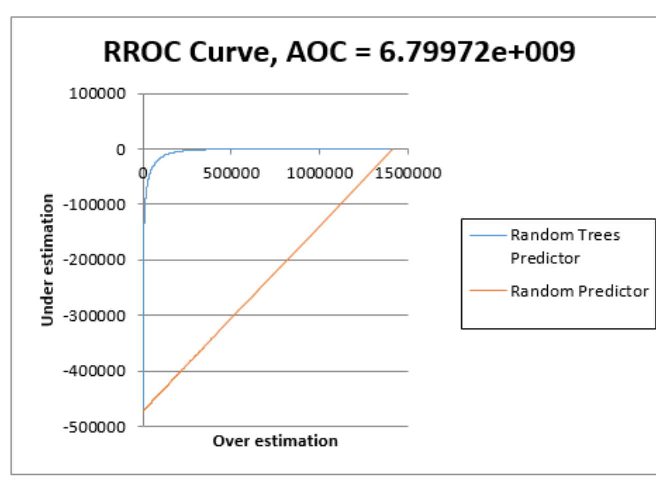
Decile Wise Lift Charts:





RROC Curve





Lift Chart

The lift chart for Random Forest is similar to the one of Multi Linear Regression so it is better than the baseline model. We cannot compare the Linear Regression with CART based on this lift charts as it is not giving us any significant difference.

Decile Chart

The decile chart for Random Forest is better in this case as compared to CART.

RROC Curve

The value of the AOC is less in this model. But the AOC for multiple linear regression is the least amongst the 3 models.

1.4. Model Recommendation

Multiple Linear Regression:

Residual DF	3442
R^2	0.824166
Adjusted R^2	0.822225
Std. Error Estimate	41.61421
RSS	5960658

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
5960658	41.38044	-8.047E-13

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4189664	42.49576	-0.41208095

Random Forest:

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
7598308	46.72039	0.609269

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
5863802	50.27426	-0.9223

CART Results:

Training Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
8757306	50.15719	2.79563E-14

Validation Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
7137641	55.46683	0.056711489

Comparing these three we can see the best (lowest) RMSE value of 41.38 for training and 42 for validation is derived by multiple linear Regression and the R-square as well as Adjusted R-squared is also around 0.82.

Also the Lift charts the Decile wise lift chart for the Multiple Linear regression are better in the case of Multiple Linear Regression.

It can be argued that since we have included a large no. of features for Multiple Linear Regression we are getting RMSE smaller than the other two model. But this is not the case. We have selected the variables by performing variable selection and feature selection. Also if we reduce the number of features that we input into the Multiple Linear Regression model by selecting only top 5-6 features even then the RMSE would be lesser than what we get by Random Forest and CART.

Proof that Multi Linear Model works better for this problem in comparison to CART and Random Forest even with lower number of features:

Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	152.795	4.477811698	34.12268708	2.4803E-220	144.0156	161.5744	540722411.1
LOT SQFT	0.007989	0.000320174	24.95114987	1.5652E-126	0.007361	0.008616	9905709.829
GROSS ARE	0.035737	0.002042122	17.49982131	8.8475E-66	0.031733	0.039741	13367220.69
LIVING ARE	0.055706	0.003668311	15.18568999	1.77873E-50	0.048514	0.062898	2075462.9
FLOORS_1	0.275506	2.557784457	0.10771257	0.914229922	-4.73941	5.290419	538912.9785
FLOORS_2	38.76852	2.330559037	16.6348604	7.61669E-60	34.19912	43.33793	497728.5435
BEDROOMS	3.526647	2.518317686	1.400398047	0.161483511	-1.41089	8.46418	2171.267664
FULL BATH_	-28.48256	1.985971916	-14.3418749	2.31113E-45	-32.3764	-24.5888	232212.8217
HALF BATH_	-23.13219	1.753401605	-13.1927503	8.17207E-39	-26.57	-19.6944	347520.3686

Residual DF	3472
R^2	0.795498
Adjusted R^2	0.795027
Std. Error Estimate	44.68428
RSS	6932491

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
6932491	44.62648	-6.98273E-13

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4863106	45.78391	-0.3612832

So the model derived by using Multi Linear regression would be the best model for this case of determining the prices of the houses in West Roxbury. The second best based on performance is Random Forest and CART is the last based on performance that we would recommend.

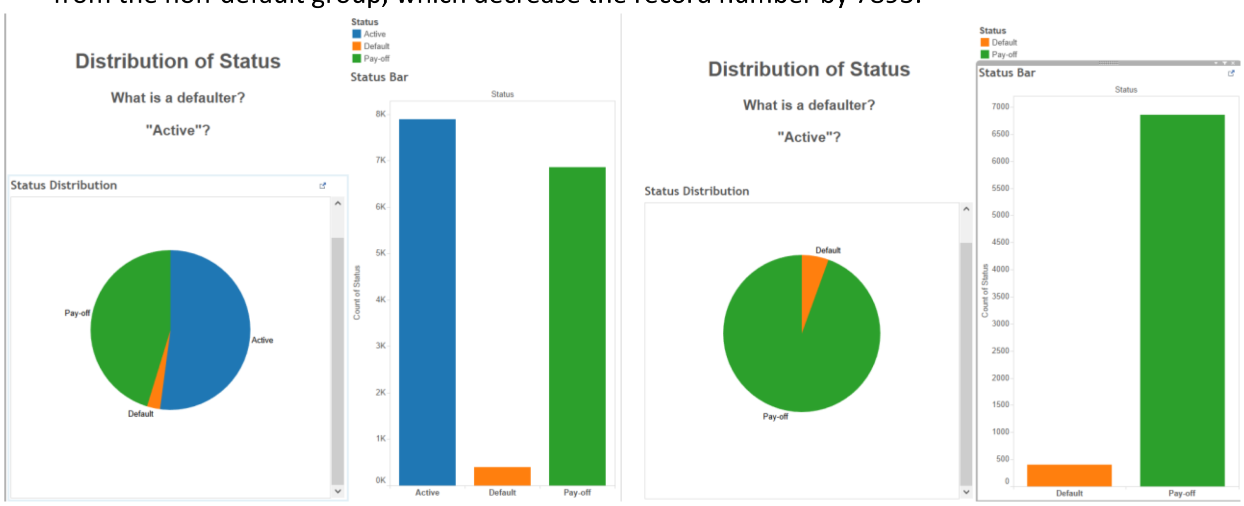
2 Mortgage Defaults

2.1 Exploratory data analysis

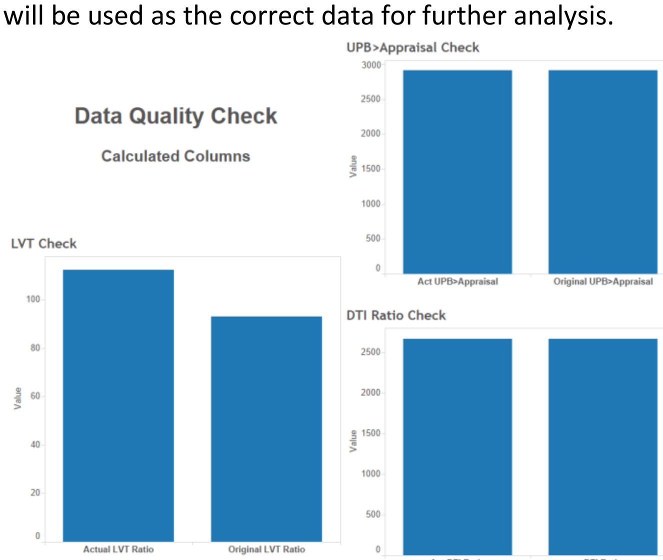
Tool: Tableau.

Steps:

1. According to the chart, noticed that the group of mortgage status "Active" has been classified as non-default. However, in the real life, these active mortgage may turn into a default as soon as the borrowers are no longer able to pay for that. Therefore, all the "active" records should be dropped from the non-default group, which decrease the record number by 7893.

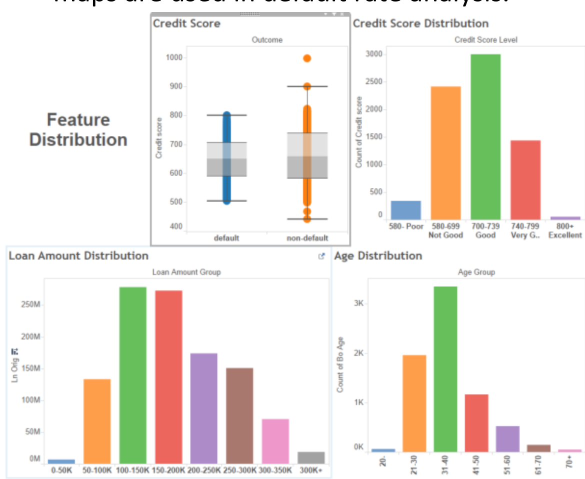


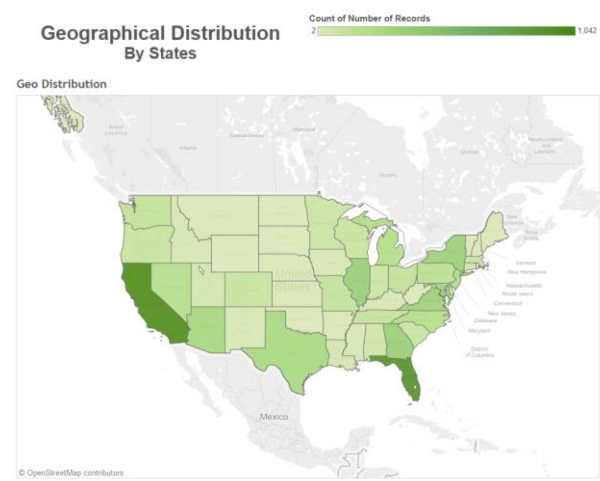
2. There are calculated columns in the dataset need to be checked by creating new calculated columns in Tableau. After the process, the data inaccuracy on LVT ratio have been discovered. The new calculated column will be used as the correct data for further analysis.

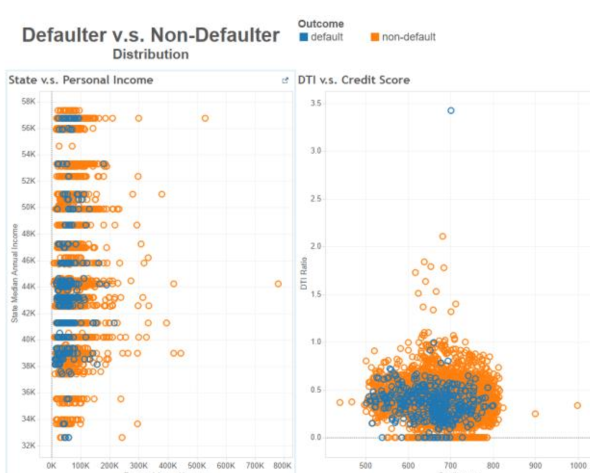


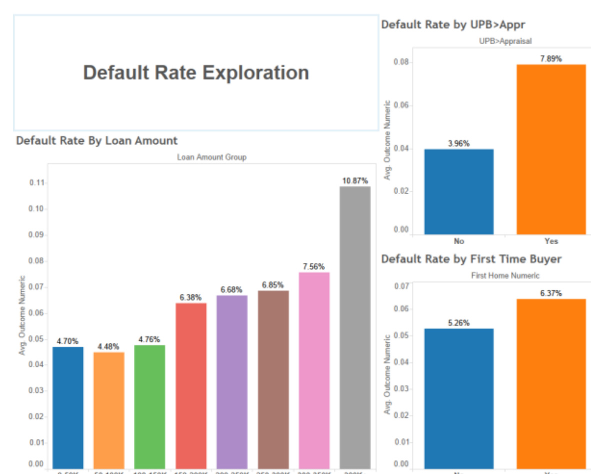
3. Further data exploration has been performed in multiple ways. First, mortgage distribution using different features has been visualized using bar chart, geo-map and box plot. Second, in order to

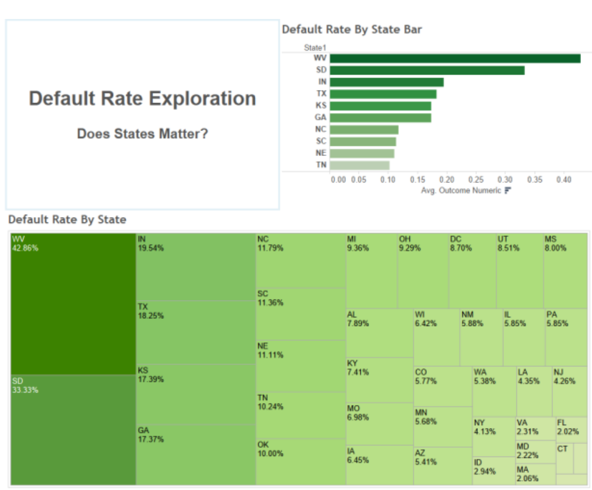
explore the relationships between default/non-default and different features, scatter plot charts have been used. Third, because of the significant sample size difference between default and non-default, the KPI default rate has been created to measure the impact of different features. Bar charts and heat maps are used in default rate analysis.











2.2 Build classification models

Tool: XLMiner

Steps:

1. Data screening and pre-processing:
    a. Filter out the rows having “active” in Status column.
    b. Create calculated column “Act_LTV_Ratio” to replace the “Orig_LTV_Ratio_Pct”. The formula used is Ln_Orig divided by the smaller between orig_apprd_val_amt and pur_prc_amt. If orig_apprd_val_amt = 0, then use pur_prc_amt[1].
    c. Drop unused column: Status, state.
    d. Convert binary categorical columns’ values into 1 and 0: First_home, OUTCOME.
2. Outlier identification and data cleansing: according to the result of visualization. The following outliers have been identified and need to be deleted from the dataset.
    a. One column has a credit score of 999, and with monthly income of $1,187, the borrower managed to pay-off a $228,000 loan. The house price is $50,000, which is far below the loan amount. This row shows obvious abnormal features of fraud data, and need to be deleted from the original dataset.
3. Partition data
    a. Data partition has been performed through standard partition option with a 60% training and 40% testing. Random seeds has been set as 12345. Same training and testing partition will be used in all the classification models.
4. Build models
    a. Set cut-off values. Cut-off value has been set to 0.05538 (5.538%), the generic percentage of default for the dataset.
    Reason: According to the scenario, the bank (lender)’s goal is to identify the potential defaulters so that they can purchase secondary insurance to prevent potential loss. Also, the cost of misidentifying a non-defaulter as a defaulter is much less than missing a real defaulter. Therefore, the model is expecting an unfair measurement of accuracy for a two-side classification. A cut-off value of 50% is not applicable in this case.
    b. Based on the cut-off value. Classification has been performed using the following machine learning algorithms: Logistic regression, CART (Classification and Regression Tree) and Random forest. Default settings are applied for CART & Random Forest.

[1] https://en.wikipedia.org/wiki/Loan-to-value_ratio

2.3 Performance evaluation

Measurement: confusion matrix. Lifting chart. Decile-wise lift chart.

1. **Logistic Regression:**
   a. In the confusion matrix of training dataset, the error rate for defaulter and non-defaulter are both close to 32%, while in the testing dataset, the error rate of defaulter increased to 36.6% and for non-defaulter it still remain lower than 32%. The overall error rate for both training and testing are similar (32%)

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.05538

Confusion Matrix

	Predicted Class	
Actual Class	1	0
1	164	77
0	1307	2807

Error Report

Class	# Cases	# Errors	% Error
1	241	77	31.95020747
0	4114	1307	31.76956733
Overall	4355	1384	31.77956372

Performance

Success Class	1
Precision	0.111488783
Recall (Sensitivity)	0.680497925
Specificity	0.682304327
F1-Score	0.191588785

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.05538

Confusion Matrix

	Predicted Class	
Actual Class	1	0
1	102	59
0	871	1872

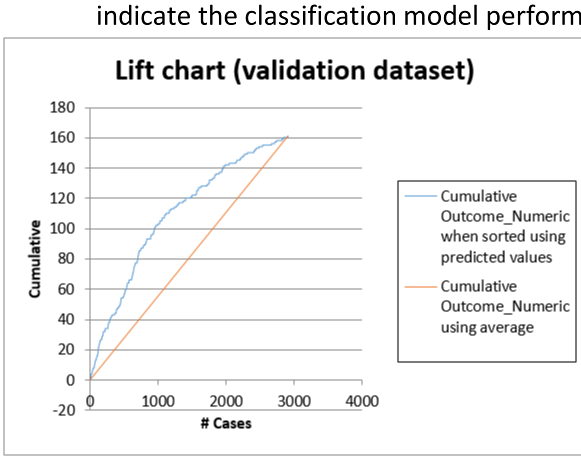
Error Report

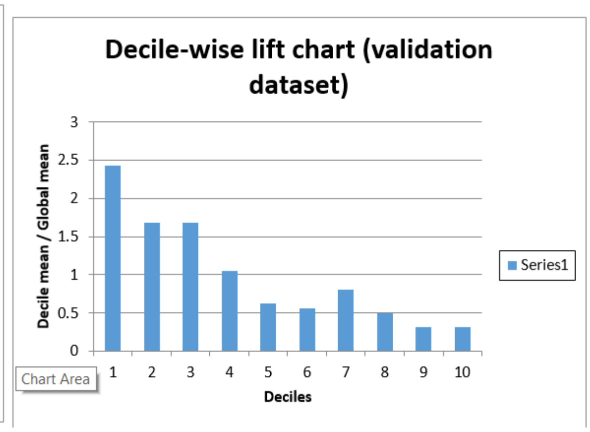
Class	# Cases	# Errors	% Error
1	161	59	36.64596273
0	2743	871	31.7535545
Overall	2904	930	32.02479339

Performance

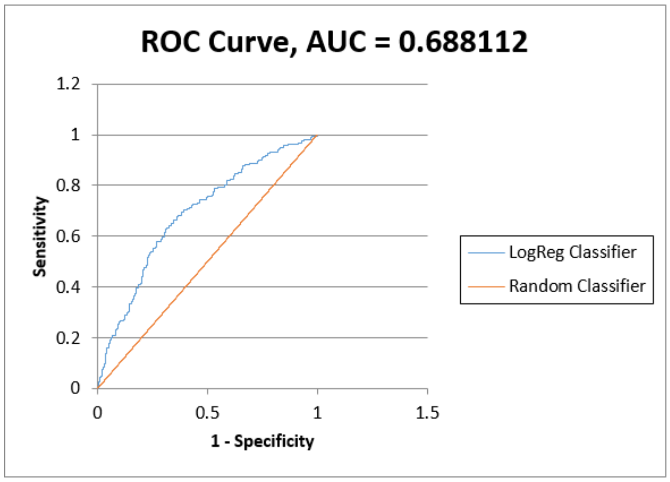
Success Class	1
Precision	0.104830421
Recall (Sensitivity)	0.633540373
Specificity	0.682464455
F1-Score	0.17989418

   b. In the lift chart for testing dataset, the curve is above the straight line which indicates random classification rules. The Decil-wise lift chart also has higher head and lower tail. Both lift chart indicate the classification model performs better than a random model.





   c. Roc Curve for testing dataset is above the straight line which indicates random classification rules. AUC = 0.688112, which also indicate the logReg Classifier performs better than a random classifier.



2. CART

a. In the confusion matrix of training dataset, the error rate for defaulter is 28% and for non-defaulter is 24%, while in the testing dataset, the error rate of defaulter increased to 100% and for non-defaulter dropped down to 0%. The overall error rate for training is 24% and for testing is 5%

Training Data scoring - Summary Report (Using Full

Cutoff probability value for success (UPDATABLE)	0.05538

Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	173	68
0	995	3119

Error Report			
Class	# Cases	# Errors	% Error
1	241	68	28.21577
0	4114	995	24.18571
Overall	4355	1063	24.40873

Performance	
Success Class	1
Precision	0.148116
Recall (Sensitivity)	0.717842
Specificity	0.758143
F1-Score	0.245564

Validation Data scoring - Summary Report (Using B

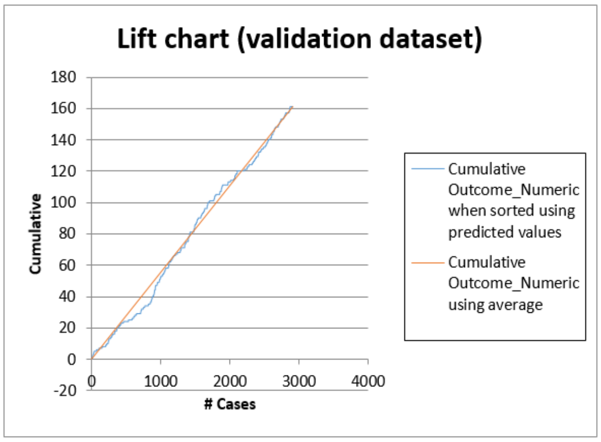
Cutoff probability value for success (UPDATABLE)	0.05538

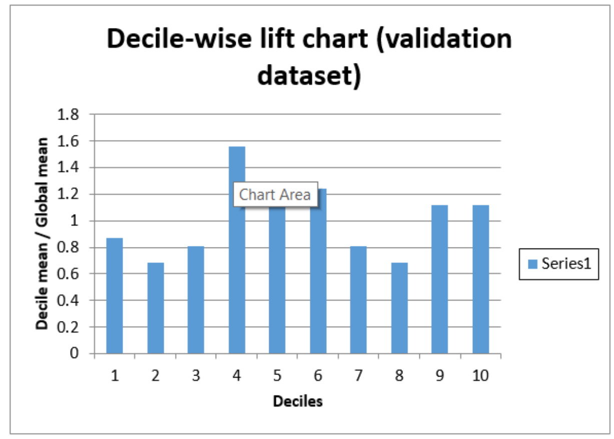
Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	0	161
0	0	2743

Error Report			
Class	# Cases	# Errors	% Error
1	161	161	100
0	2743	0	0
Overall	2904	161	5.544077

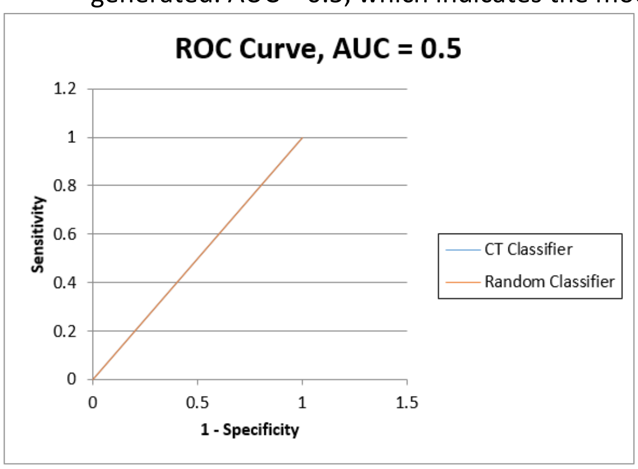
Performance	
Success Class	1
Precision	Undefined
Recall (Sensitivity)	0
Specificity	1
F1-Score	Undefined

b. In the lift chart for testing dataset, the curve is along with straight line which indicates randomly classify the outcome by using average. The Decil-wise lift chart also do not show higher head or lower tail. Both lift chart indicate the classification model performs no better than a random classification model.





c. Due to the 100% error rate on one side of classified outcome, ROC Curve cannot be generated. AUC = 0.5, which indicates the model is no better than a random classifier.



3. Random Forest

a. In the confusion matrix of training dataset, the error rate for defaulter is 1.24% and for non-defaulter is 76%, while in the testing dataset, the error rate of defaulter increased to 9.31% and for non-defaulter remains 76%. The overall error rate for training and testing are both around 72%

Training Data scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.05538

Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	238	3
0	3122	992

Error Report			
Class	# Cases	# Errors	% Error
1	241	3	1.244813
0	4114	3122	75.88721
Overall	4355	3125	71.7566

Performance	
Success Class	1
Precision	0.070833
Recall (Sensitivity)	0.987552
Specificity	0.241128
F1-Score	0.132186

Validation Data scoring - Summary Report

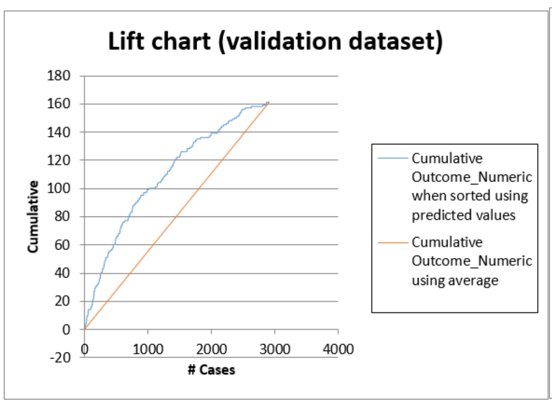
Cutoff probability value for success (UPDATABLE)	0.05538

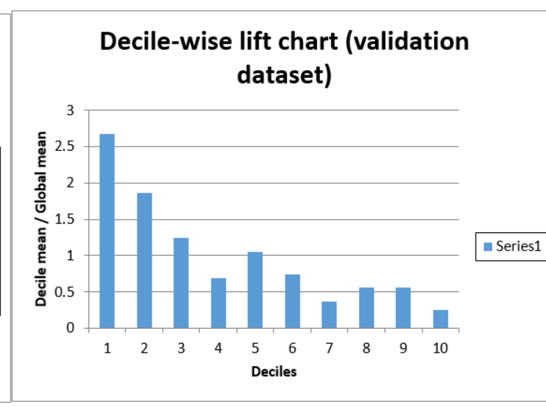
Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	146	15
0	2094	649

Error Report			
Class	# Cases	# Errors	% Error
1	161	15	9.31677
0	2743	2094	76.33977
Overall	2904	2109	72.62397

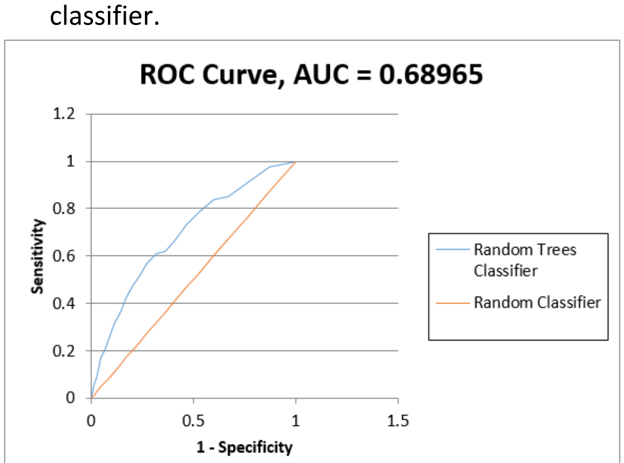
Performance	
Success Class	1
Precision	0.065179
Recall (Sensitivity)	0.906832
Specificity	0.236602
F1-Score	0.121616

b. In the lift chart for testing dataset, the curve is above the straight line which indicates a random classifier. The Decil-wise lift chart also has higher head and lower tail. Both lift chart indicate the classification model performs better than a random classifier.





c. Roc Curve for testing dataset is above the straight line which indicates random classification rules. AUC = 0.68965, which also indicate the logReg Classifier performs better than a random classifier.



4. Model Comparison:
    a. CART has the lowest overall error rate in the testing dataset. However, it failed to identify any defaulter, the performance is no better than a random classifier. Therefore, it is eliminated from the candidates of the classification models.
    b. Logistic regression has the lower overall error rate than random forest. However, the error rate of classifying a defaulter is much lower using random forest.
    c. AUC value and lift charts indicate random forest has a better performance than logistic regression.

2.4 Model selection and recommendation

1. Model selection:
   In order to perform model selection, the following assumption need to be made based on the operation of bank and financial institution (lender) in real life.
   Assumption 1: lender will purchase secondary insurance for any borrowers who has been classified as a potential defaulter in order to avoid their potential loss due to a mortgage default.
   Assumption 2: the premium payment of purchasing a secondary insurance is much lower than the cost of missing a defaulter in classification. Assume the annual premium rate for a Lenders mortgage insurance is in the range of 0.66% to 0.75%[2].
   Assumption 3: the average mortgage payment duration is around 10 years. Defaults normally happen during after the 5th year of mortgage. The assumption is made conservatively since economic and unemployment rate changes significantly in a 10-year period, according to Federal Reserve Bank, during economic crisis, the default rate may increase to over 10%[3].
2. Model Selection
   With the three assumptions above. Random Forest has been chosen as the recommended classification model for this scenario. The following reasons are presented for reference:
   a. Comparison has been made between Logistic Regression and Random Forest by calculating the expected costs based on the error rates of both models (testing + training).
      The expected cost for Logistic Regression is
      0.75%*(102/2+871+1307+164/2)*10+59+77-164-102 = 43.325
      The expected cost for Random Forest is
      0.75%*(146/2+2094+3115+238/2)*10+15+3-146-238 = 39.075
      As a result, the expected cost of using Random Forest is lower than Logistic regression with fairly conservative assumptions[4]
   b. Default rate may increases due to external factors like economy downhill and unemployment rate increases.

3. Recommendation
   a. Considering the current resources and available options, we recommend Random Forest as the best Classification models to identify mortgage defaults.
   b. In order to make comprehensive decisions and finding the best model. Other models such as boosting trees should be considered, tested and compared to the current Random Forest Model.
   c. Once new data and factors are collected, both model and expected cost calculation should be refined and updated for future use.

[2] https://en.wikipedia.org/wiki/Lenders_mortgage_insurance
[3] https://research.stlouisfed.org/fred2/series/DRSFRMACBS
[4] The highest premium rate has been applied.

3. Spambase

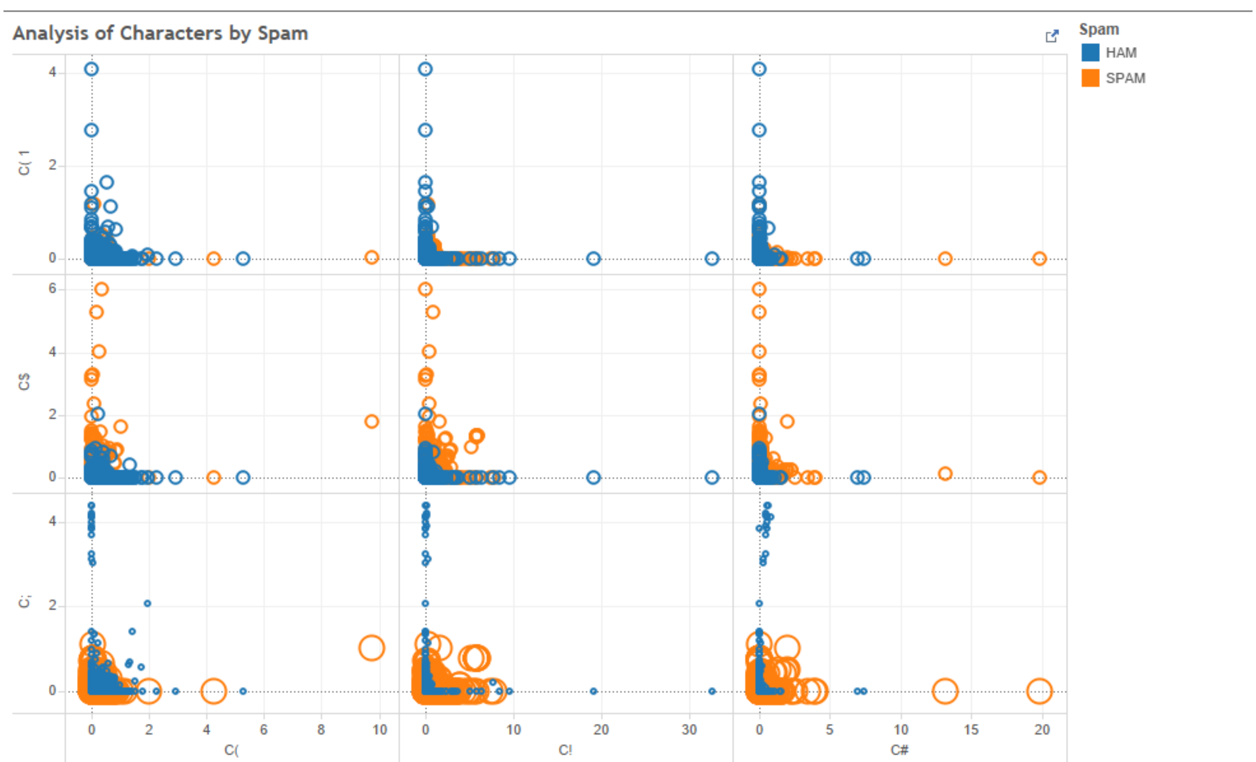
Analyzing the dataset

- There are total 48 attributes of type words
- Total 6 attributes of type characters like C
- 1 real attribute of type average of CAPITAL LETTERS
- 2 integer attributes of type total and longest of the length of CAPITAL LETTERS

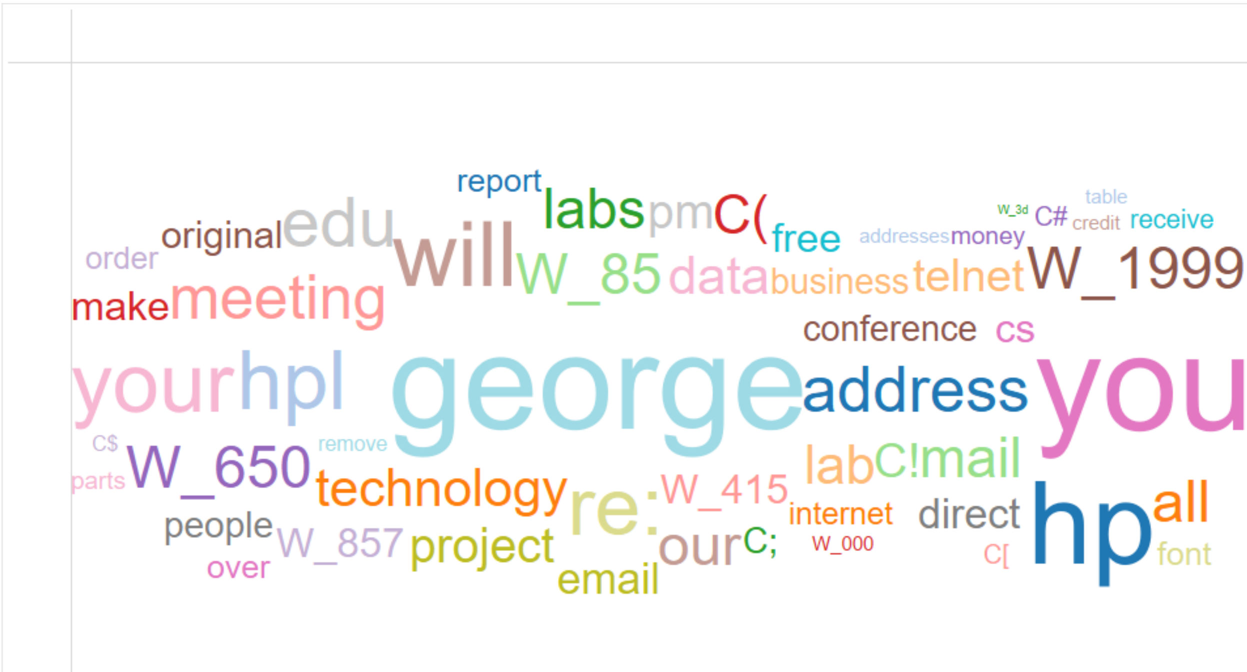
3.1 Exploratory Analysis:

Tool: Tableau

- Scatter Plot of Characters in the list by Spam



- Word Clouds showing Ham words by average of the percentage of the words



- Word Clouds showing Spam words by average of the percentage of the words

3.2 Building Models

Tool: XLMiner

Steps:

1. Data Screening and Pre-processing
    a. Performed Feature Selection: Feature selection is performed to select a subset of relevant features for use in model construction. Performing feature selection gives you the chi-squared P-Value which helps in determining whether a predictor accepts or reject 'Null Hypothesis'. There is no significant difference between observed and expected frequencies
2. Data Partition
    a. Data partition has been performed through standard partition option with a 60% training and 40% testing. Random seeds has been set as 12345. Same training and testing partition will be used in all the classification models.
3. Build Models
    a. Considering 1813 emails tagged as spam from 4601 emails. The initial cutoff probability of success is taken as 0.39. If the probability of success for an email is less than this value then the email would be a non-spam email and if it greater than this value then the email would be a spam email.
    b. Based on the cut-off value: Classification has been performed using the following machine learning algorithms: Logistic regression, CART (Classification and Regression Tree) and Random forest. Default settings are applied for CART & Random Forest.

3.3 Model Evaluation

Measurement: Confusion Matrix, Lift Chart, Decile-wise Lift Chart

1. Logistic Regression:
   For detecting spam messages, the error rate of spam messages from the confusion matrix of validation dataset 7.82 %

Validation Data Scoring - Summary Report

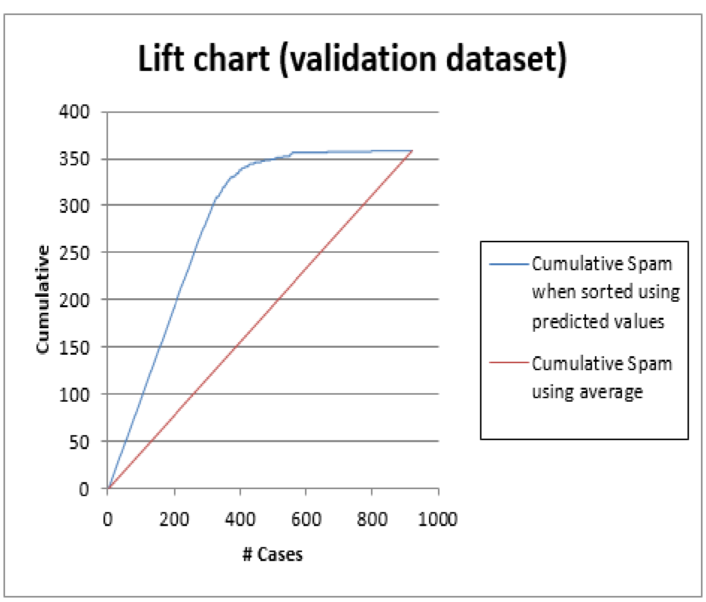
Cutoff probability value for success (UPDATABLE)	0.39	Updating the value here will NOT update value in detailed report

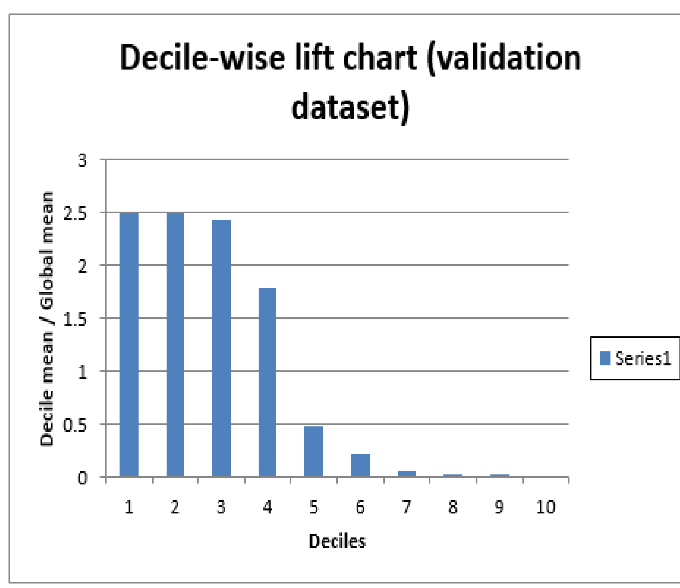
Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	330	28
0	45	517

Error Report			
Class	# Cases	# Errors	% Error
1	358	28	7.82122905
0	562	45	8.007117438
Overall	920	73	7.934782609

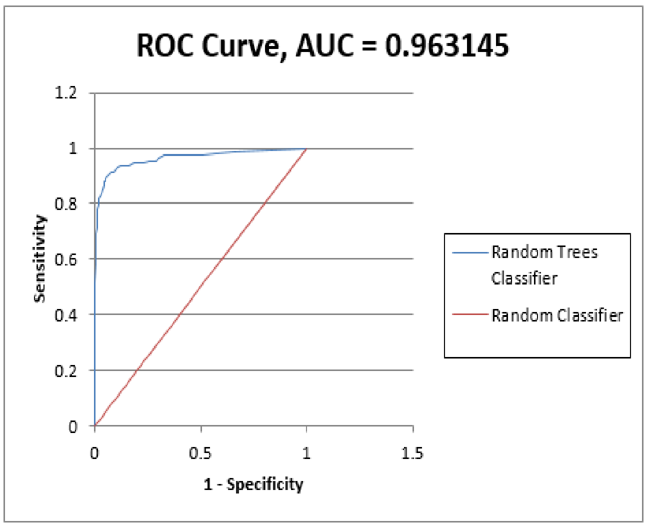
Performance	
Success Class	1
Precision	0.88
Recall (Sensitivity)	0.9217877
Specificity	0.9199288
F1-Score	0.9004093

The curve of the lift chart for validation dataset is above the straight line which is good. The decile-wise lift chart of the validation dataset has higher heads and lower tails. The top 2 decile contains 20% of the emails most likely to be spam emails. Whereas the bottom 2 decile contains 20% of the emails least likely to be the spam emails.





ROC Curve of the validation dataset is in the top left corner which indicates that the model performs better model than a random model



2. CART

a. The error rate of spam messages from the confusion matrix is about 19% which is high compared to the error rate of Logistic regression

Validation Data scoring - Summary Report (Using Best Pruned Tree)

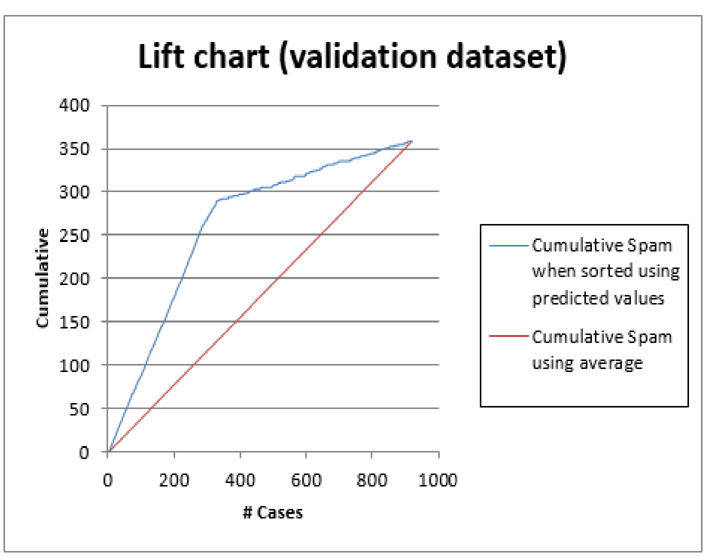
Cutoff probability value for success (UPDATABLE)	0.39	Updating the value here will NOT update value in detailed report

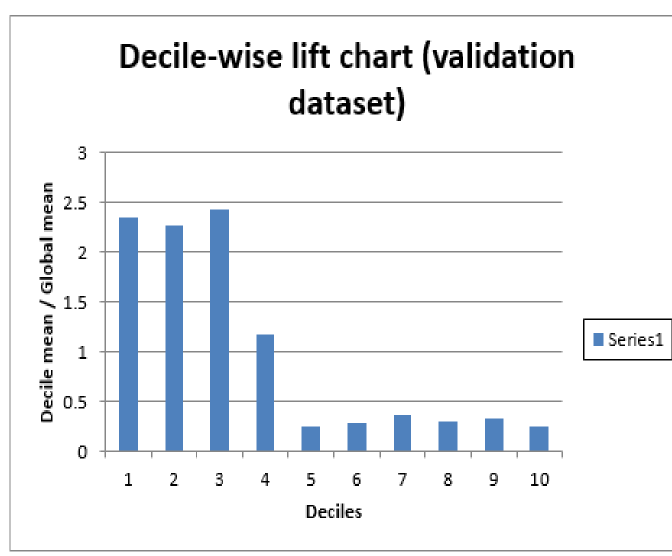
Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	290	68
0	43	519

Error Report			
Class	# Cases	# Errors	% Error
1	358	68	18.99441
0	562	43	7.651246
Overall	920	111	12.06522

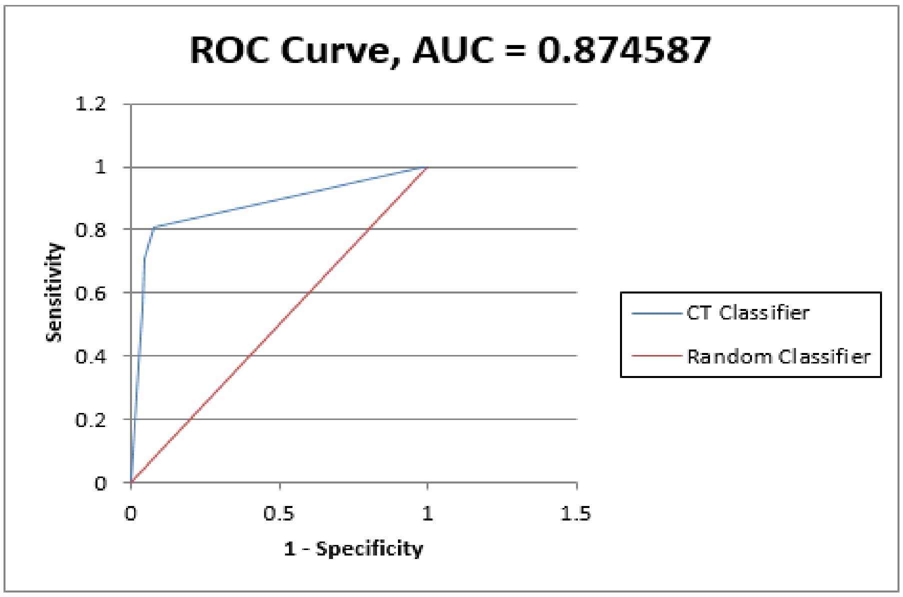
Performance	
Success Class	1
Precision	0.870871
Recall (Sensitivity)	0.810056
Specificity	0.923488
F1-Score	0.839363

b. The curve of the lift chart for Validation dataset of CART is above the straight line. Whereas in the decile chart the higher and the lower heads are out of order which indicates that the model is not good compared to a random classifier model which should be in good staircase order from left to right.





c.The ROC curve is away from top left corner which again indicates that the model is not good



3. Random Forest

a. The error rate of the spam messages from the confusion matrix of Random Forest using Random Forest is 6.14% which is low compared to Logistic Regression and CART.

Validation Data scoring - Summary Report

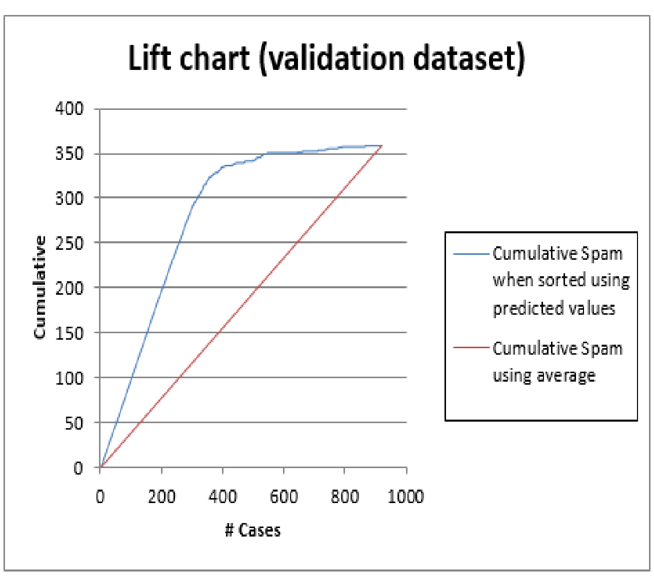
Cutoff probability value for success (UPDATABLE)	0.39	Updating the value here will NOT update value in detailed report

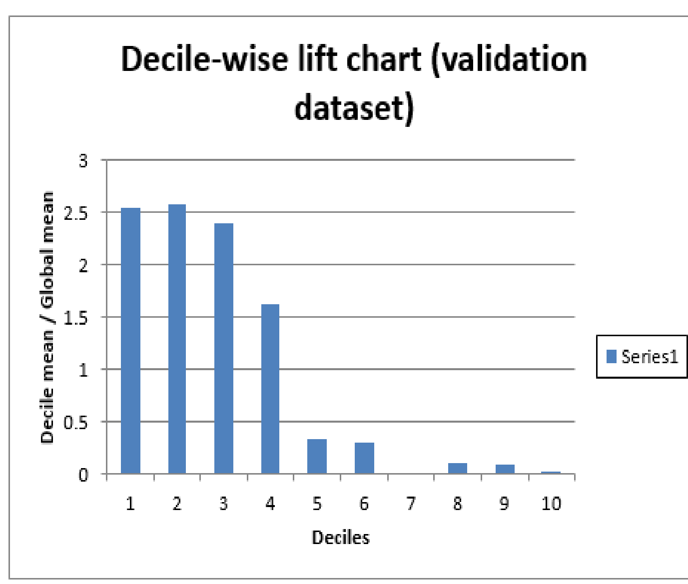
Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	336	22
0	73	489

Error Report			
Class	# Cases	# Errors	% Error
1	358	22	6.145251
0	562	73	12.98932
Overall	920	95	10.32609

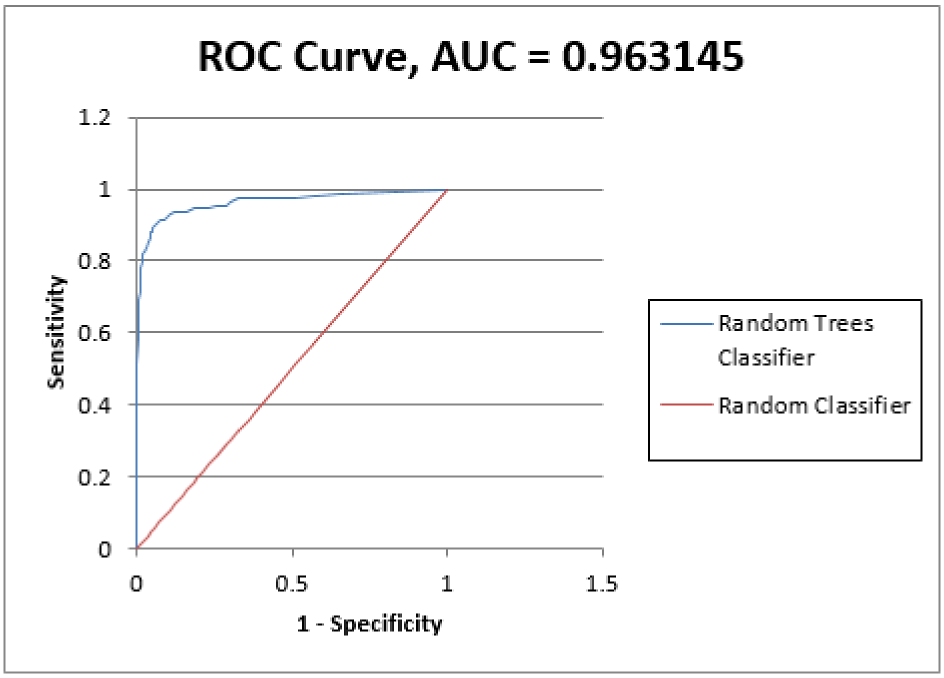
Performance	
Success Class	1
Precision	0.821516
Recall (Sensitivity)	0.938547
Specificity	0.870107
F1-Score	0.876141

b. The Lift chart of Random Forest is above the straight line which is good. The decile-wise lift chart shows that the higher heads and the lower tails with decreasing order from which it can be determined that this is a good model compared to a random classifier.





c. The ROC curve of the Random Forest is above the straight line and near the top left corner of the chart



3.3. <u>Model Comparison</u>:

a. The overall error rate of Logistic Regression is 7.9 % which is lower compared to Random Forest which is 10.32%. But for detecting the Spam emails from the validation dataset the error rate of Random Forest is lower 6.14% compared to Logistic Regression which is 7.82%.
b. CART has the highest overall error rate in the validation dataset. And, again from the decile-wise lift chart and ROC curve it can be seen that the model is not good for detecting the spam emails.
c. For detecting spam emails Random Forest provides better performance compared to Logistic Regression

3.4 **Model evaluation and Recommendation**

a. Assumption: According to the dataset, if we assume the initial cutoff probability of success to be 0.39 or 39% then Random Forest model is good taking into account the error rates which are:
   Logistic Regression: 7.9%

   CART: 18.99%

   Random Forest: 6.14%

   If we consider the initial cutoff probability of success to be 0.5 (default) or 50% for detecting spam emails then the error rates are:

   Logistic Regression: 12.29%

   CART: 18.99%

   Random Forest: 10.89%

   Hence, considering both the conditions Random Forest is a good model

4 Blog Feedback

4.1 Building Prediction Models

Tools: RStudio, Python

Steps:

1. Data Preprocessing:
    - Concatenated all the csv files having test data with the help of Python

      Python Code for concatenation:

      ```
      import glob
      import pandas as pd

      path ='C:/Users/vanwu/Desktop/INFO 7390 ADS/Midterm/BlogFeedback/test'
      allFiles = glob.glob(path + "/*.csv")
      frame = pd.DataFrame()
      list_ = []
      for file_ in allFiles:
        df = pd.read_csv(file_, header = None)
        list_.append(df)

      frame = pd.concat(list_)
      frame.to_csv('blog_data_test.csv', sep=',')
      ```

    - Looking into the data and referring the paper, we inferred that the significance of 200 bag-of-words is very weak in predicting the number of comments in next 24 hours. Also, the trend of the words may change which is not predictable. Hence, we deleted the 200 columns containing bag-of-words.

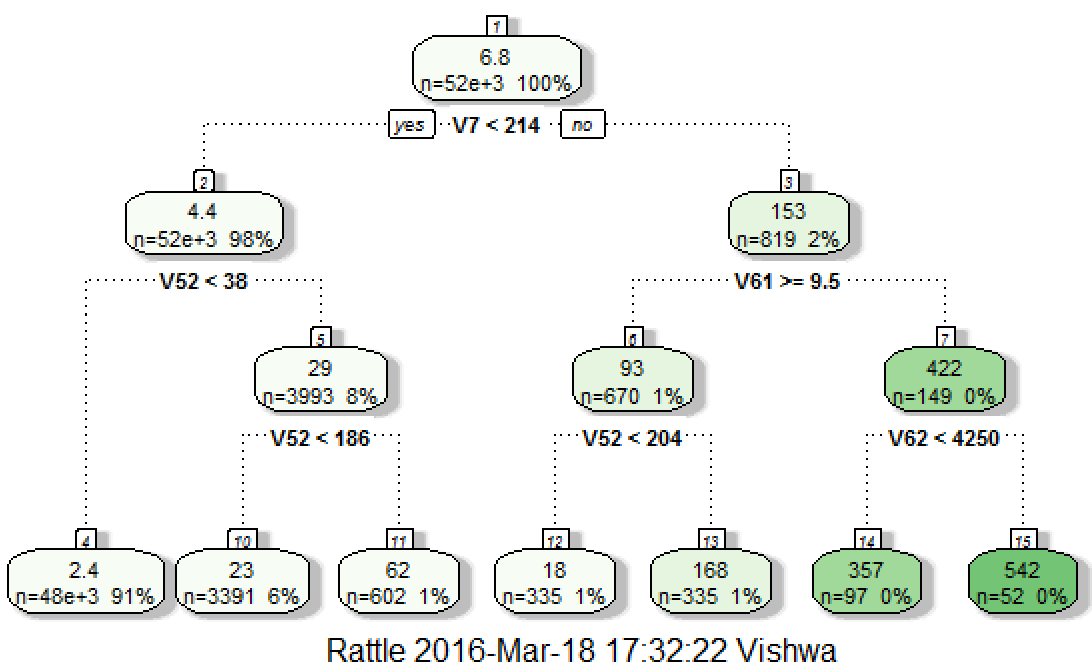
2. Build Models:
    - Prediction has been performed using the following machine learning algorithms: Multiple Linear Regression, CART(Classification and Regression Tree) and Random Forest.

4.2 Performance Evaluation

Tool: RStudio

Measurement: RMSE value

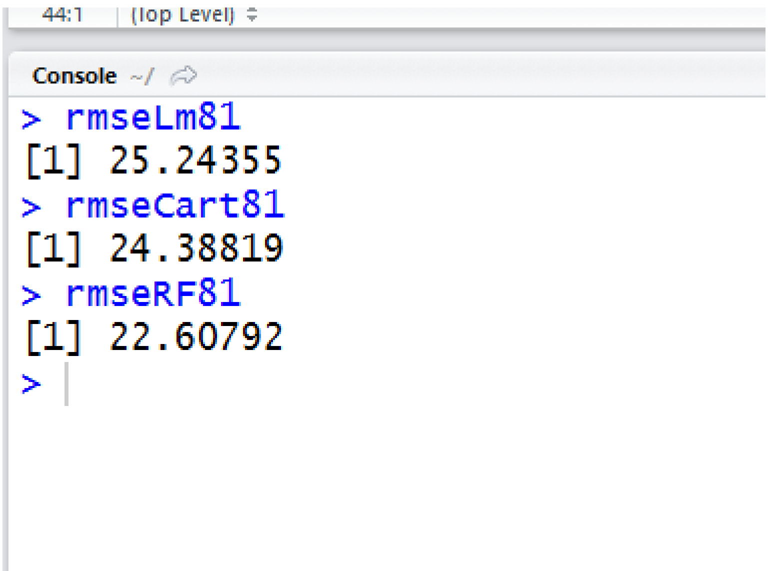
1. Linear Regression:
    - Applying the linear regression formula in training data and predicting the rmse value by applying the model to the testing data gives rmse value to be 25.24 and multiple R-squared value to be 0.36
2. CART:
    - The packages used for CART algorithm for calculating RMSE value and visualization of the tree are: rpart, rattle, dynamicGraph, rattle, rpart.plot, cartFit81
    - The RMSE value calculated with CART is 24.39 which is less compared to Regression model.
    - The regression tree that we got from CART algorithm is:



3. Random Forest:

- The package used for Random Forest in R is 'randomForest'
- The RMSE value calculated with Random Forest is 2.20 which is the lowest of all the models.

The list of RMSE values calculated using all the algorithms in R are:



4.3 <u>Model Selection and Recommendation</u>

1. Model Selection:
   In order to perform model selection, the following assumptions need to be made:
   Positive

   Assumption 1:

   More Total/Average/Max/Min (1,3-5, 6, 8-10, 11, 13-15) comments before basetime (51)/ within the 24 hours before basetime (52)/ within 48-24 hours before basetime, there may be more comments in next 24 hours (53).

   Assumption 2:

   More Total/Average/Max/Min (26, 28-30, 31, 33-35, 36, 38-40) trackbacks before basetime(56)/ within the 24 hours before basetime (57)/ within 48-24 hours before basetime(58), there may be more comments in next 24 hours.

   Assumption 3:

More Average/Max/Min (278 - 280) of comments the parent posts (277) received, there may be more comments in next 24 hours.

Negative

Assumption 1:

Longer the hours between publication & basetime (61), lesser comment will be in the next 24 hours.

Uncertain

Assumption 1:

Weekday of post publication (270-276) may affect the comments in the next 24 hours.

Assumption 2:

Weekday of basetime (263-269) Sat, Sun More

Assumption 3:

Length of post (62), too short or too long, less comments

Assumption 4:

Bag of words (63-262) not known, trends change.

Assumption 5:

Standard deviation (2,7,12,17,23,27,32,37,42,47) may affect consistency

Assumption 6:

24 hours after publication & before basetime:

1. Published at least 24 hours before basetime. More comments/links (54, 59), more comments in the next 24 hours. Positive
2. Published less than 24 hours before basetime. **---- to be determined.**

Assumption 7:

Difference of comments/Link (55,60) number between last 24 hours and last last 24 hours.

1. If >=0, attention grows/maintains, positive
2. If <0, attention drops, negative

At first we calculated the RMSE value of all the algorithms using all the parameters of the training and testing dataset. The results we gained using 281 parameters are:
Linear Regression: 21.96
CART: 27.06
Random Forest: 28.39
According to this, regression model would be a good model.
After close analysis, we found out that the significance of the bag-of-words on the model very less. Hence, we removed columns with bag-of-words from the csv files and use those files for all the algorithms. The result we received with 81 parameters are:
Linear Regression: 25.24
CART: 24.39
Random Forest: 22.61

2. Model Recommendation:

- If we take into account all the parameters, then Linear Regression is good, but if we remove bag-of-words columns and evaluate the model Random Forest is good.

http://www.cs. bme.hu/~buza/pdfs/gfkl_buza_social_media.pdf