MATH 5557 Applied Regression Analysis

Group Project on Multiple Regression Analysis

Modelling the Median House Value Price in California

Instructor: Dr. Shu-Chuan Chen

Group: Shovan Chowdhury, Kiprop Kibet, Emmanuel Anokye Yeboah

---

December 6, 2020

Idaho State University

Introduction:

The demand for apartments in California is high and there is a lack of housing, especially apartments. In 1990 in California District census was and the data was published.

The dataset was updated 3 years ago. This data consists of 10 features. There are 207 unused rows. There is total 20640 data in the whole dataset. The data contains longitude, latitude, House media age, total rooms, total bedrooms, population, household, median income, median house value and ocean proximity.

In this data set we will use all the features mention above to predict the median house value. We will conduct a series of multiple regression analysis to narrow down the best model predictor of median house value.

Methodology:

Following flowchart represents the whole methodology for this project:

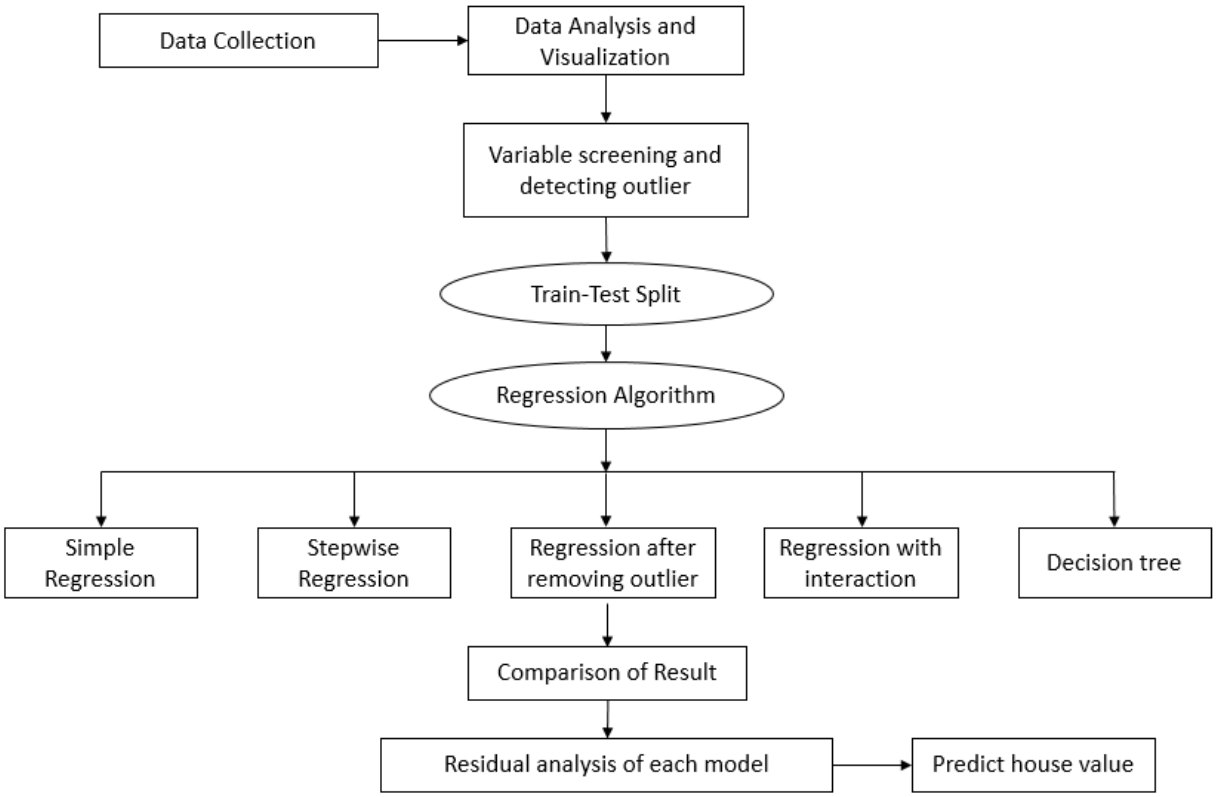
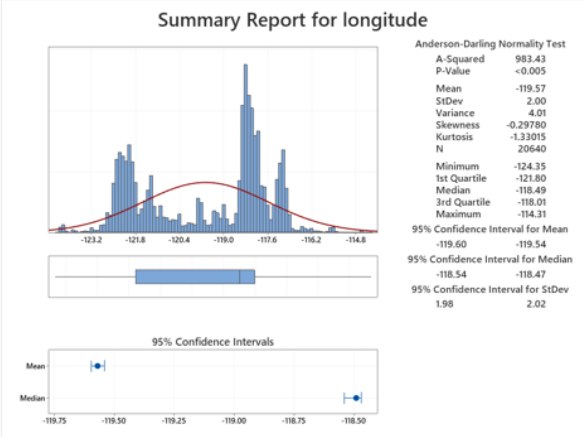
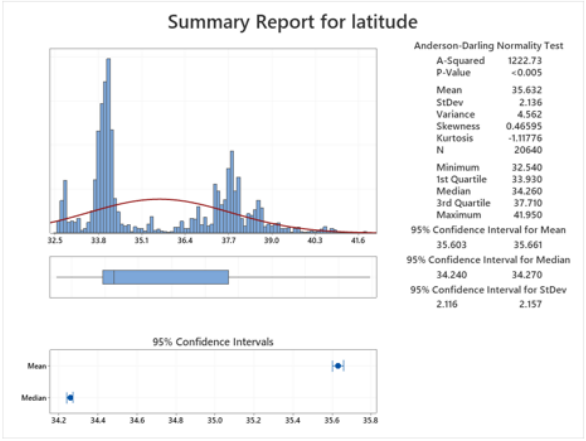


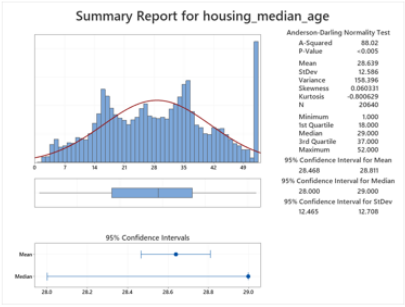
Figure: Flowchart of the project

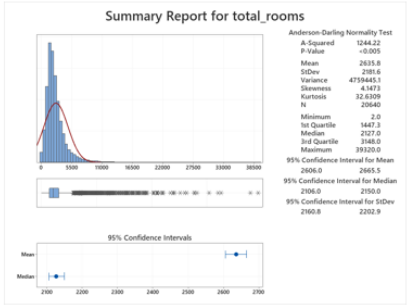
Data Visualization:

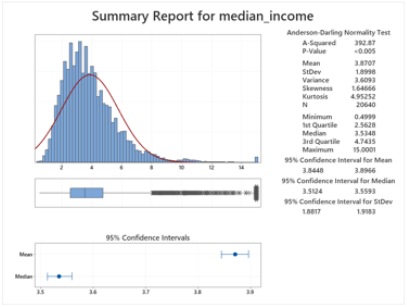
There are total 10 features in our analysis. From there, median house value is the dependent feature and other 9 features are independent features including one categorical feature. Graphical summary of each features are given below:

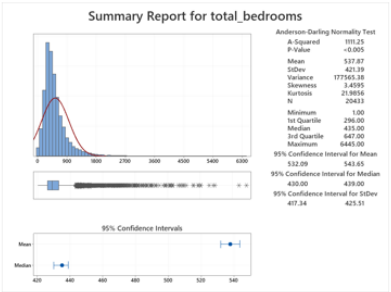


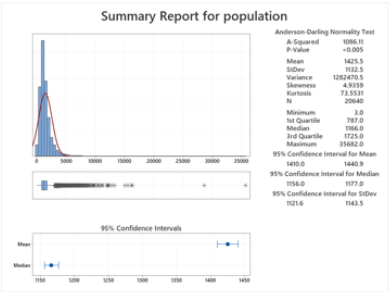


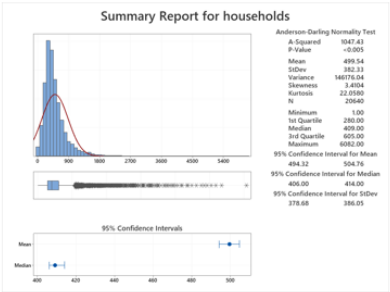


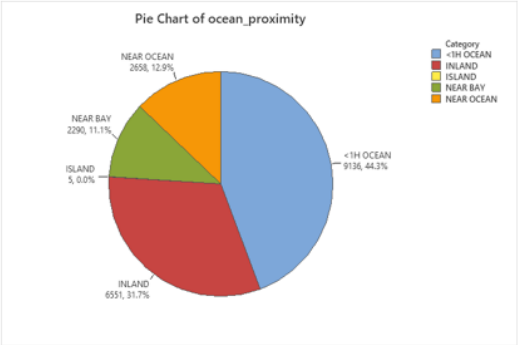


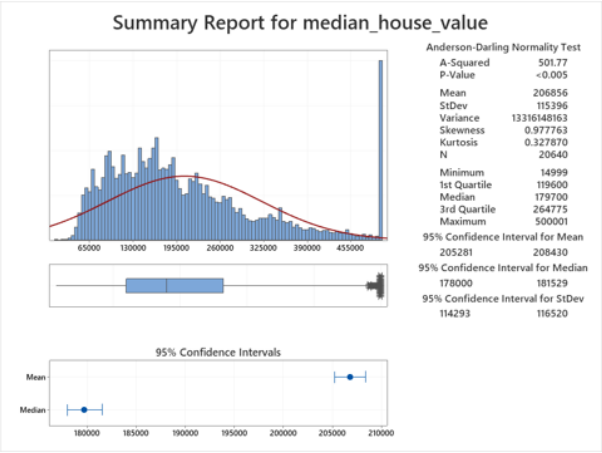












After observing the histogram of all features, it is observed that some of the features is highly skewed which can affect in our further analysis. In that case, we will consider transformation of the data. If we see our dependent feature median house value, we can see a huge peak in the histogram.

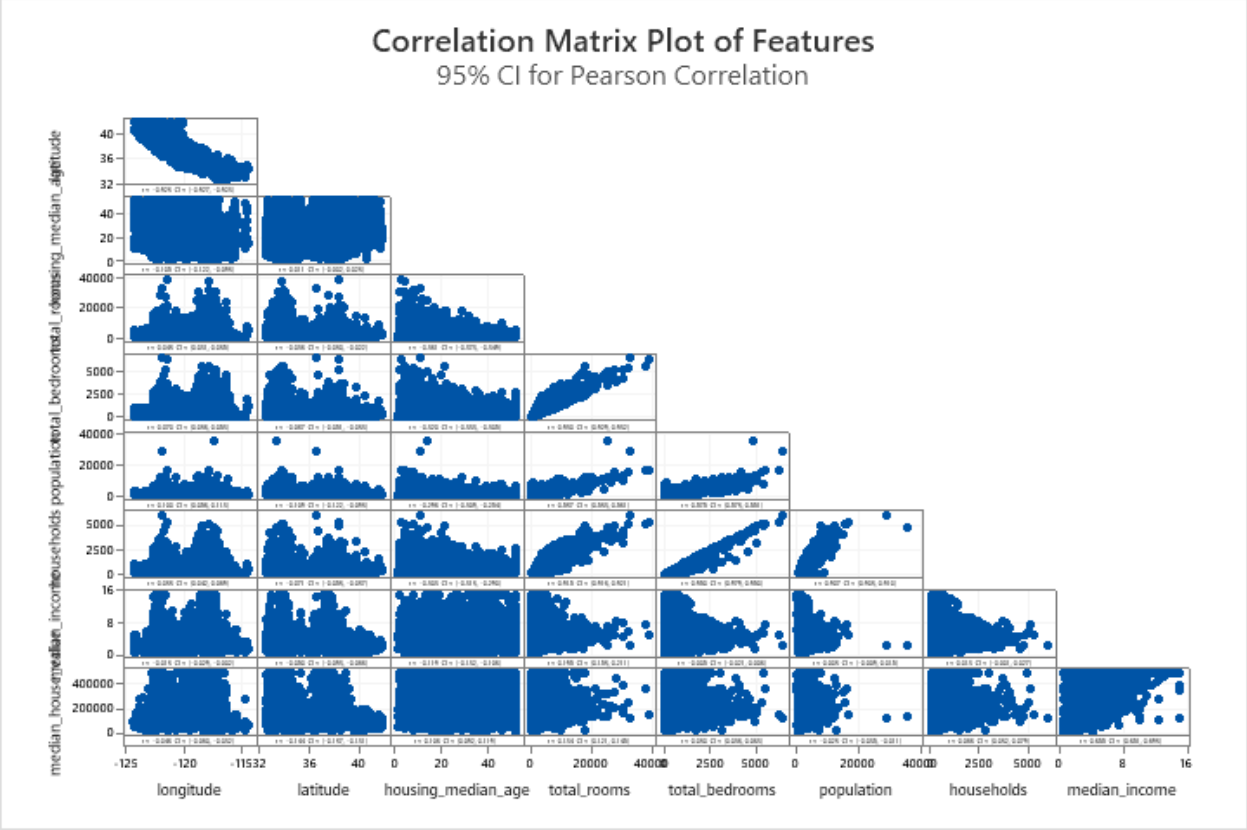
Recoding Categorical Features:

Since we have the qualitative data which is the ocean proximity, we recoded it and this is what we found.

Original Value	Recoded Value	Number of Rows
<1H OCEAN	1	9136
INLAND	2	6551
ISLAND	3	5
NEAR BAY	4	2290
NEAR OCEAN	5	2658

Correlation and multi collinearity analysis:

Before going for regression, we need to check out the correlation between each feature. Higher value of correlation indicates strong relationship between features. We need to find out which feature has strong correlation with median house value. Correlation matrix of features and the value is given below:



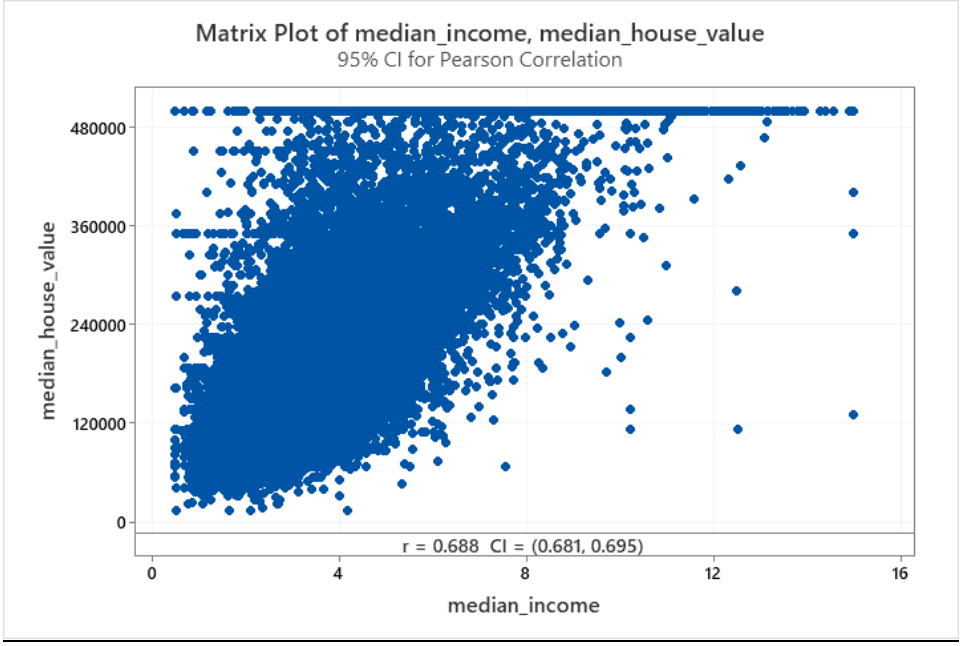
Correlations

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms
latitude	-0.925				
housing_median_age	-0.108	0.011			
total_rooms	0.045	-0.036	-0.361		
total_bedrooms	0.070	-0.067	-0.320	0.930	
population	0.100	-0.109	-0.296	0.857	0.878
households	0.055	-0.071	-0.303	0.918	0.980
median_income	-0.015	-0.080	-0.119	0.198	-0.008
median_house_value	-0.046	-0.144	0.106	0.134	0.050

	population	households	median_income
latitude			
housing_median_age			
total_rooms			
total_bedrooms			
population			
households	0.907		
median_income	0.005	0.013	
median_house_value	-0.025	0.066	0.688

We find out that some of the features has strong correlation between each other. Yellow highlighted values are the one associated with higher value. There is a chance of multicollinearity

as there is high correlation between some of the independent features. Median income has the highest correlation with median house value with a value of 0.688.



Simple Linear Regression:

First of all, we did simple linear regression and tried to see the basic result of regression. The multiple coefficients of determination R^2 (64.63%) is large enough. So, the data fit is good for the predictive model for Median House Value. From Minitab, we got following equation for the simple linear regression.

Regression Equation

Recoded ocean_proximity	
1	median_house_value = -2269954 - 26813 longitude - 25482 latitude + 1072.5 housing_median_age - 6.193 total_rooms + 100.56 total_bedrooms - 37.97 population + 49.62 households + 39260 median_income
2	median_house_value = -2309238 - 26813 longitude - 25482 latitude + 1072.5 housing_median_age - 6.193 total_rooms + 100.56 total_bedrooms - 37.97 population + 49.62 households + 39260 median_income
3	median_house_value = -2117052 - 26813 longitude - 25482 latitude + 1072.5 housing_median_age - 6.193 total_rooms + 100.56 total_bedrooms - 37.97 population + 49.62 households + 39260 median_income
4	median_house_value = -2273908 - 26813 longitude - 25482 latitude + 1072.5 housing_median_age - 6.193 total_rooms + 100.56 total_bedrooms - 37.97 population + 49.62 households + 39260 median_income
5	median_house_value = -2265676 - 26813 longitude - 25482 latitude + 1072.5 housing_median_age - 6.193 total_rooms + 100.56 total_bedrooms - 37.97 population + 49.62 households + 39260 median_income

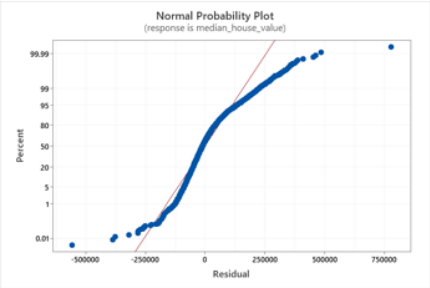
Model Summary

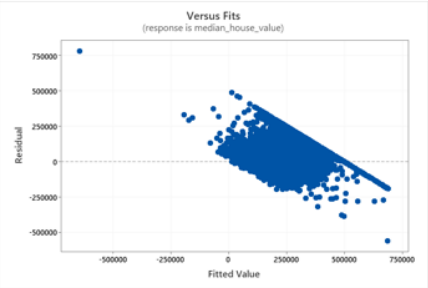
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
68657.0	64.65%	64.63%	9.66292E+13	64.51%	513121.00	513231.93

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-2269954	88014	(-2442468, -2097440)	-25.79	0.000	
longitude	-26813	1020	(-28812, -24814)	-26.30	0.000	18.09
latitude	-25482	1005	(-27451, -23513)	-25.36	0.000	19.97
housing_median_age	1072.5	43.9	(986.5, 1158.5)	24.44	0.000	1.32
total_rooms	-6.193	0.791	(-7.745, -4.642)	-7.83	0.000	12.97
total_bedrooms	100.56	6.87	(87.09, 114.02)	14.64	0.000	36.31
population	-37.97	1.08	(-40.08, -35.86)	-35.28	0.000	6.45
households	49.62	7.45	(35.01, 64.22)	6.66	0.000	35.17
median_income	39260	338	(38597, 39922)	116.15	0.000	1.79
Recoded ocean_proximity						
2	-39284	1744	(-42703, -35865)	-22.52	0.000	2.86
3	152902	30742	(92645, 213158)	4.97	0.000	1.00
4	-3954	1913	(-7704, -204)	-2.07	0.039	1.57
5	4278	1570	(1202, 7355)	2.73	0.006	1.20

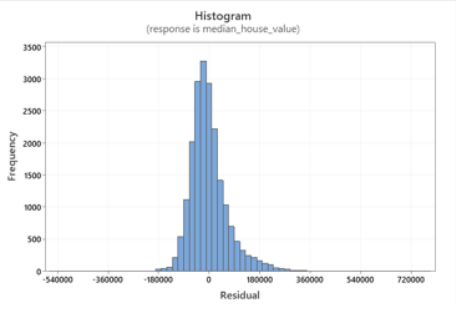
Residual analysis of this regression is given below:

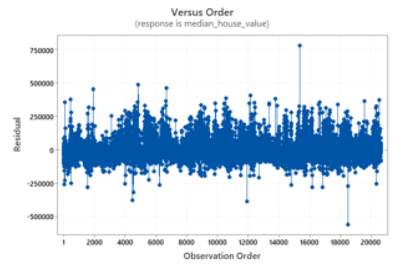




Durbin-Watson Statistic

Durbin-Watson Statistic = 0.977430





Stepwise Regression:

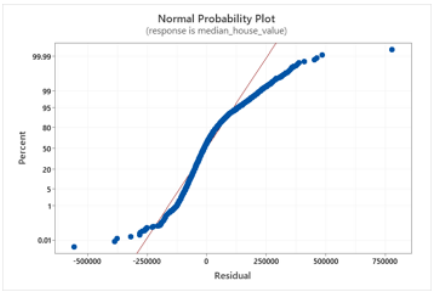
We undertook Stepwise Regression 9 times in Minitab to figure out a better predictive model for Median House Value with little change in R^2 . After the 9 separating Stepwise tests, we found that the variables had the largest insignificant p-value of 0.000, given by the first Stepwise test and C_p is 13.00.

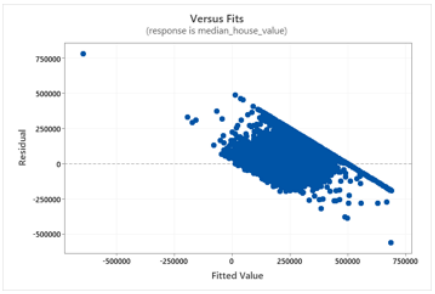
	------Step 7-----		------Step 8------		-------Step 9------	
	Coef	P	Coef	P	Coef	P
Constant	26110		-207095		-2269954	
median_income	40466	0.000	40483	0.000	39260	0.000
housing_median_age	1185.1	0.000	1183.1	0.000	1072.5	0.000
Recoded ocean_proximity	174152	0.000	174753	0.000	152902	0.000
total_bedrooms	81.31	0.000	86.64	0.000	100.56	0.000
population	-37.00	0.000	-36.34	0.000	-37.97	0.000
households	76.75	0.000	71.15	0.000	49.62	0.000
total_rooms	-7.509	0.000	-7.797	0.000	-6.193	0.000
longitude			-1960	0.000	-26813	0.000
latitude					-25482	0.000
S		69806.5		69728.3		68657.0
R-sq		63.45%		63.53%		64.65%
R-sq(adj)		63.43%		63.51%		64.63%
Mallows' Cp		700.60		654.28		13.00
AICc		513797.58		513752.76		513121.00
BIC		513892.66		513855.76		513231.93

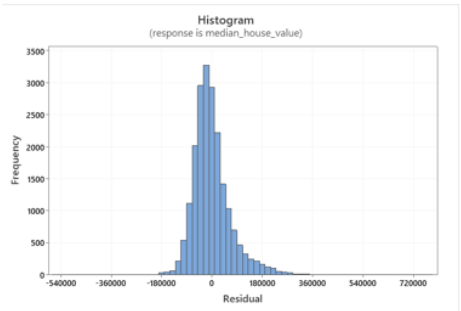
Model Summary

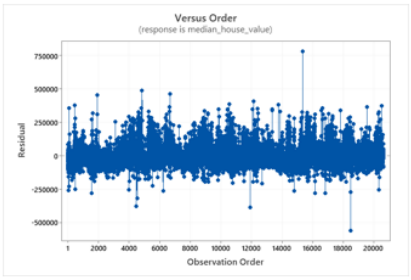
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
68657.0	64.65%	64.63%	9.66292E+13	64.51%	513121.00	513231.93

Residual analysis is given below:









Durbin-Watson Statistic

Durbin-Watson Statistic = 0.977430

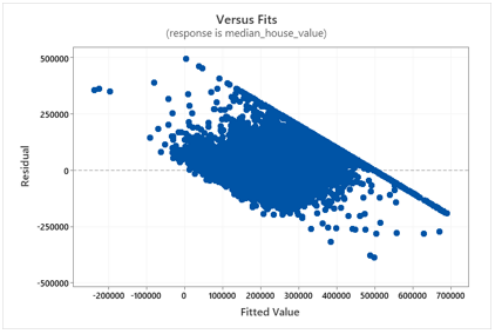
If we see the residual analysis, we can see that there is a huge outlier. There is deviation in normal probability plot which suggest us that the residual distribution is not normal. There is some pattern also observed in residual vs fits graph. We can delete the outlier and see how it impacts the results.

Detecting and removing Outlier:

There are two outliers detected in row no 15361 and 18502. We found out this data to analyze the reason of being outliers for this model. Following is the two data which is giving us outliers.

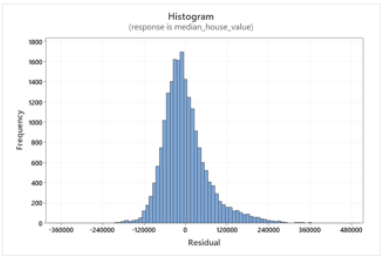
longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity	Recoded ocean_proximity	FITS	RESI
-117.42	33.35	14	25135	4819	35682	4769	2.5729	134400	<1H OCEAN	1	-644652	779052.3
-121.59	37.19	52	220	32	55	26	15.0001	131300	<1H OCEAN	1	688280.4	-556980

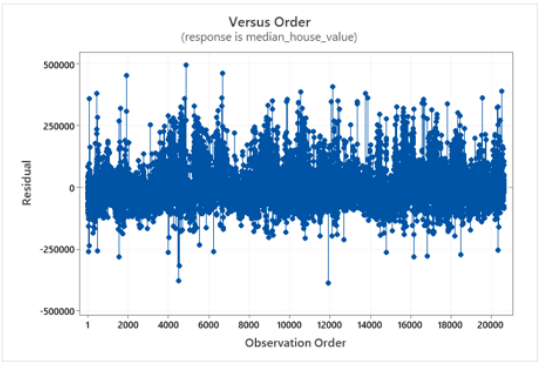
One outlier has high number of total rooms, total bedroom, population and household than others which actually suggests this data is from the place where population is very higher than others. So we consider deleting this unusual data. Another outlier has lower number of total rooms, total bedroom, population and household than others but it has higher median income. We deleted both of this outlier to improve the result. After deleting this two outliers, results are as follows:



Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
68300.8	65.01%	64.99%	9.54445E+13	64.94%	512858.24	512969.17

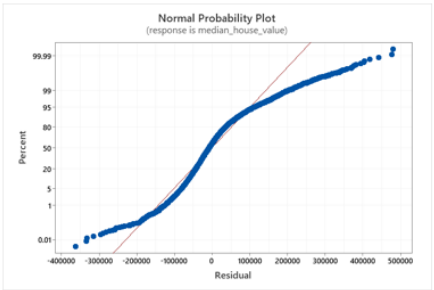




We can see that our accuracy is improved to 64.94% by deleting only two outliers.

Developing interaction and full model:

We developed full model with interaction of features and polynomial part. Results are improved to 71.48% with this interaction model.

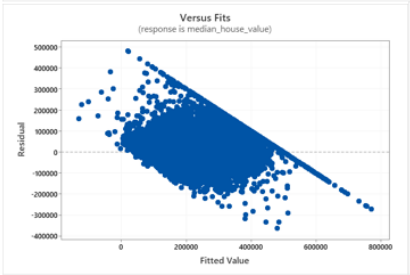


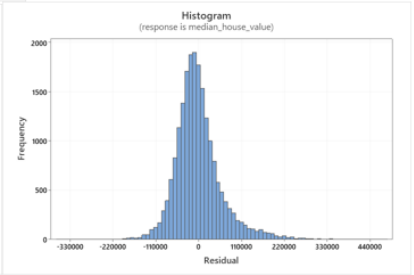
Model Summary

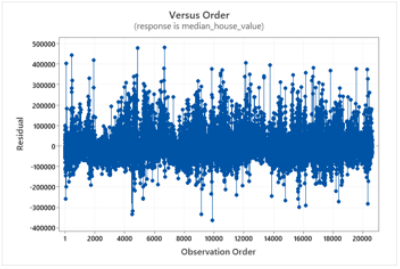
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
61652.1	71.57%	71.48%	*	*	508725.71	509248.31

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.16089



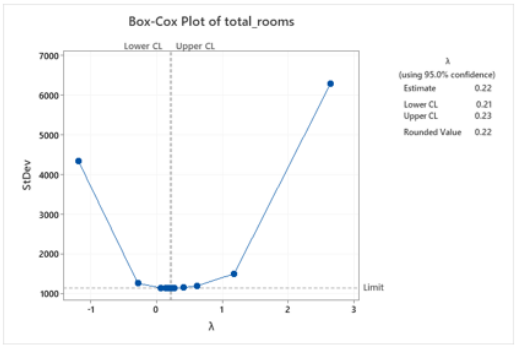


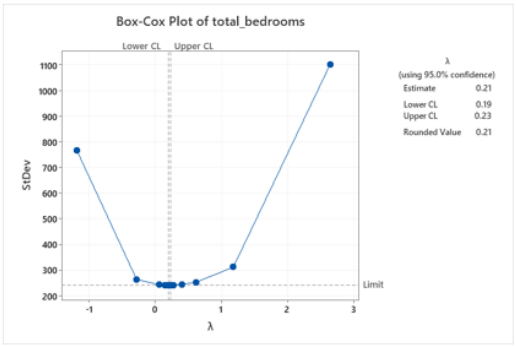


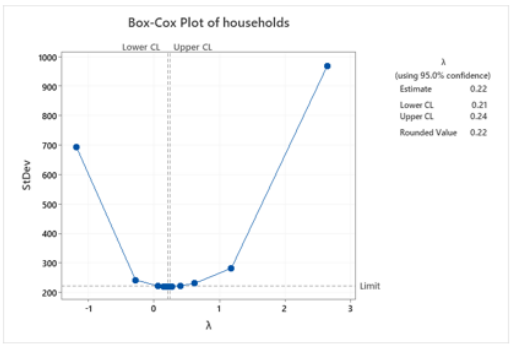
Durbin statistics 1.16089 represents that there is positive correlation among residuals. Developing full model increased the R^2 score but there is still some pattern observable in residual vs fits graph.

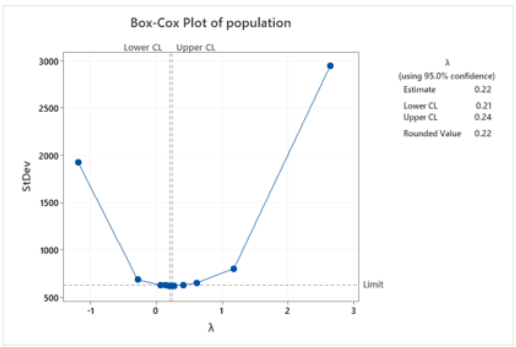
Transformation of data:

We transformed some of the features using box-cox transformation. After the transformation of data, we refit the model with transformed features.









Followings are the results using transformed features:

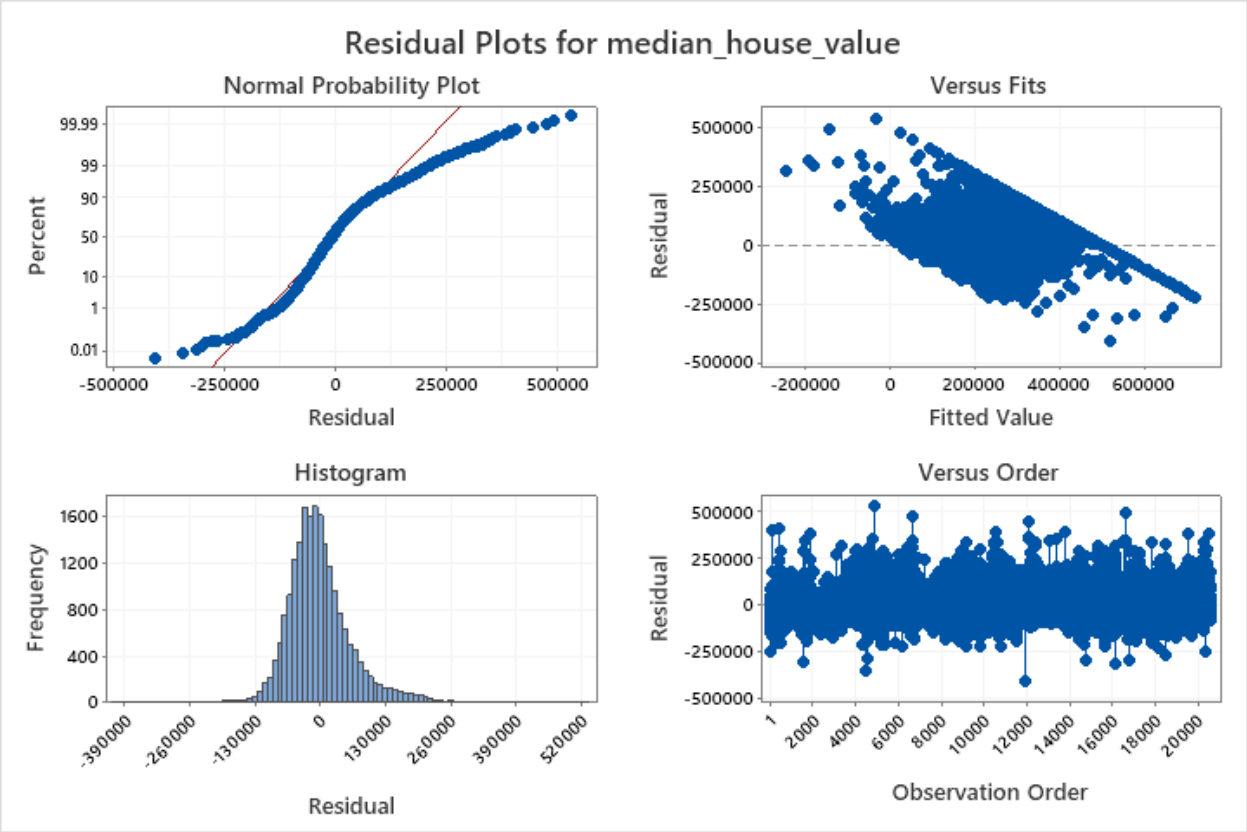
Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
66097.5	67.24%	67.22%	8.93824E+13	67.17%	511518.36	511629.29

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-2502034	85906	(-2670416, -2333652)	-29.13	0.000	
longitude	-29168	996	(-31119, -27216)	-29.29	0.000	18.61
latitude	-28350	985	(-30281, -26419)	-28.77	0.000	20.72
housing_median_age	1053.4	42.2	(970.8, 1136.1)	24.99	0.000	1.32
median_income	42171	396	(41394, 42947)	106.49	0.000	2.64
Transformed_total_room	-37498	2529	(-42456, -32540)	-14.82	0.000	21.72
Transformed_total_bedroom	127478	5472	(116753, 138203)	23.30	0.000	39.24
Transformed_population	-85161	1711	(-88514, -81808)	-49.78	0.000	7.78
Transformed_households	53001	4542	(44097, 61905)	11.67	0.000	34.94
Recoded ocean_proximity						
2	-33151	1699	(-36481, -29822)	-19.52	0.000	2.93
3	126032	29613	(67988, 184075)	4.26	0.000	1.00
4	-7969	1843	(-11582, -4356)	-4.32	0.000	1.57
5	-1221	1520	(-4200, 1757)	-0.80	0.422	1.21

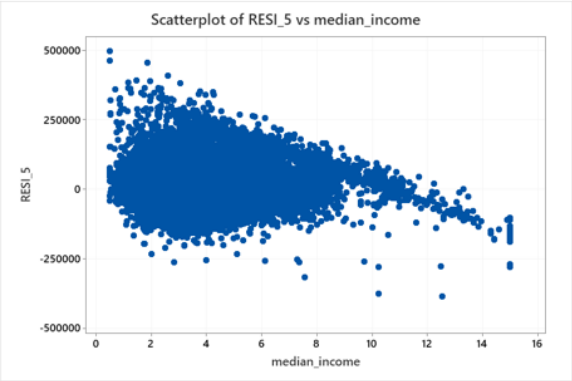
From the results, we can see that R^2 score is 67.17% which is greater than the one without transformation. So we can say that transformation of features improved our accuracy. Residual analysis of transformed data is given below:



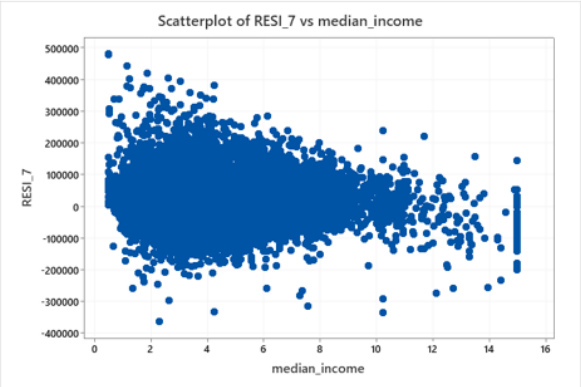
From the residual analysis, we can say that transformation could not able to diminish the pattern we observed in the residual vs fits graph.

Residual vs Predictor:

We plotted each of the features vs residual and tried to find out some special observation. Followings are the findings from residual vs predictor graph:



Simple regression model showing some pattern



Full model does not have this problem

Simple linear regression model is showing some pattern when residual plotted against median income. For the high median income, it is always giving the negative result. But for the full interaction model, we don't have this problem.

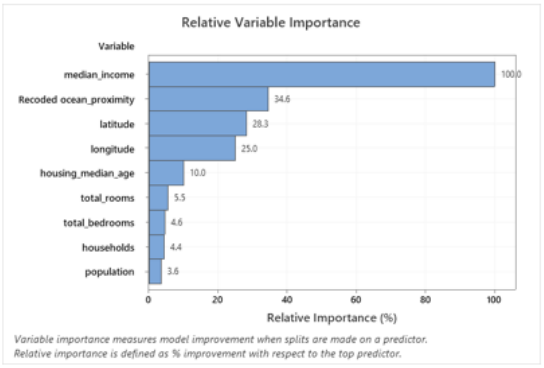
Decision Tree Model:

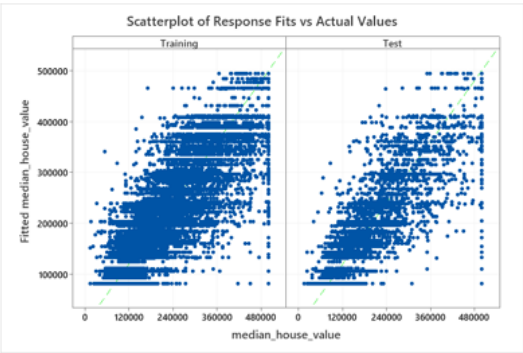
We build CART regression model if it outperforms the linear regression model. Observing the decision tree model, we can see that it is giving us R^2 score of 76.89% which is better than linear regression model.

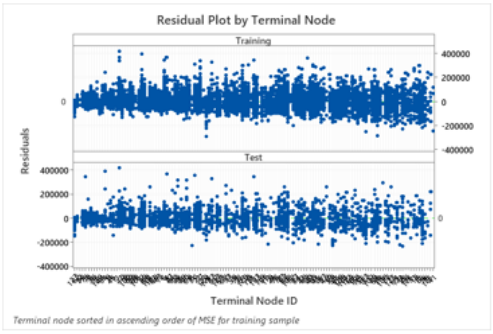
Model Summary

Total predictors	9
Important predictors	9
Number of terminal nodes	129
Minimum terminal node size	3

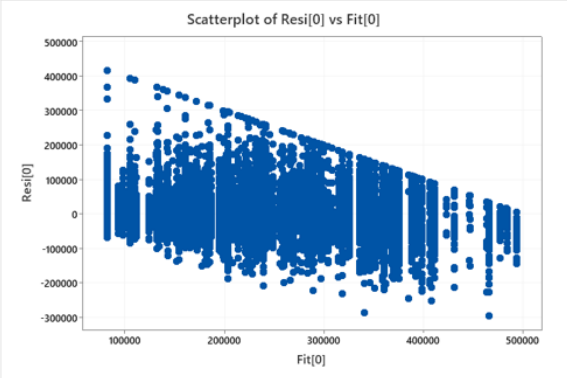
Statistics	Training	Test
R-squared	0.7689	0.7357
Root mean squared error (RMSE)	55359.0743	59817.5741
Mean squared error (MSE)	3.06463E+09	3.57814E+09
Mean absolute deviation (MAD)	39271.0872	41631.8983
Mean absolute percent error (MAPE)	0.2273	0.2388

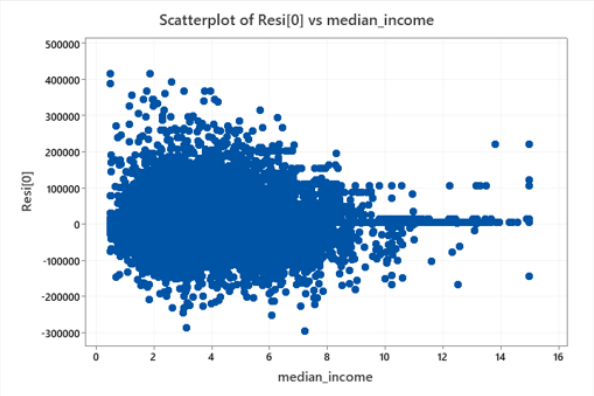






Residual plot of decision tree is given below for the CART model:





If we see the residual vs fits graph for the decision tree model, we can see that we are not getting any negative fits value in this model but still there is a pattern observable which is better than linear regression model. Residual vs median income graph is also equally distributed which satisfy the assumption of equal variance of residual.

Comparison of Results:

From all of the model, we got that decision tree model is giving us the best score. Among all of the regression model, full model is giving us best score.

	Simple regression model	Stepwise regression	Regression after removing outlier	Full model	Decision Tree
R square	64.65%	64.65%	65.01%	71.57%	76.89%

Conclusion:

From our analysis utilizing multiple methods of data processing technique, we have determined that the independent variables were important in determining the median house value. Finding out the correlation of each features is another important task before starting the regression. Residual analysis can give us very important information about outliers. Deleting outliers can improve the result greatly. Building interaction full model can improve the accuracy. CART regression is also useful regression technique which sometimes can outperform simple linear regression results.