Analysis of the Factors Affecting Price of Real Estates

MBA652A – Statistical Modelling for Business Analytics

Project Report

Submitted by:

Group:11

Ashwani Prabhakar (18114006)

Danish Nawaz(18114008)

Dhruv Patel(18114010)

Rohit Gupta(18114020)

Guided By:

Prof. Devlina Chatterjee

Contents

Acknowledgement:

We are highly indebted to Prof. Devlina Chatterjee, for her guidance and continuous support in completing this project. It is because of the knowledge and skills acquired during the course work, along with her comprehensive style of teaching, that we are able to understand the subject in a better way and are able to complete this modelling project successfully.

Declaration:

This is to certify that the project report entitled 'Analysis of the Factors Affecting Prices of Real Estates' is based on our original research work. Our indebtedness to other works, studies and publication's have been duly acknowledge at the relevant places.

Dhruv Patel

(IME MTech)

Danish Nawaz

(IME MTech)

Ashwani Prabhakar

(IME MTech)

Rohit Gupta

(IME MTech)

Introduction:

The Real Estate valuation is influenced by various factors. It is complex function of various economic factors, geographical factors, neighborhood, Rea Estate condition etc. The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan.

The data consists of deal prices of various houses under certain conditions characterized by attributes such as location, distance from metro station, house age, deal time and number of convenient stored available in neighborhood. We have analyzed the effect of each of these factors on the prices of houses and have tried to construct various regression models based on these factors and included models that were found significant.

Objective:

The objective of the project is to study the various factors affecting the **Price of Real Estate per square feet** using various Regression Techniques and to formulate models depicting the effects of these factors. We will try to evaluate all the possible combination of variables that explains the variation in price of Real Estate and try to conclude the best possible combination.

Data Source:

https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set#

The dataset we have chosen is historical dataset of houses at Sindian Dist., New Taipei City, Taiwan from year 2012-2014

Variables:

1)Dependent variable(Y):

- **Sales price of house:** It is a continuous variable whose change in value is to be analyzed using regressors.

2)Independent variables(X):

Following factors affect the prices of house:

- Deal Time:

Time has major role in determining real estate prices. With change in time various policies related to real estate, inflation rate, neighborhood area may change leading to variation in prices at the time of sell. We have included variable **Time of Deal(td)** to capture these details.

- Condition of House:

Condition of house may change with time as house may deteriorate with time. Also, as house becomes old, design/style of house may become outdated with time, leading to decrease in house price ideally. We have included variable **House Age(ha)**to incorporate these details. The intuitive relation may not be true always for example, house may be renovated in mean time and hence effect of age on price becomes difficult to estimate.

- Location of House and neighborhood:

Location has prime effect on prices for any real estate. Location in a posh area or having a popular place nearby may drastically lift the prices of houses. To incorporate these details, we have used four variables. **Latitude(lat), Longitude(long)**, **distance from MRT station(dtmrt)** and **Number of convenient stores nearby(nos).** As we will see in Exploratory Data Analysis, some of these variables are correlated with each other; hence we may have to carry out a multicollinearity test and remove some of them before finalizing our model.

So dependent variables are:

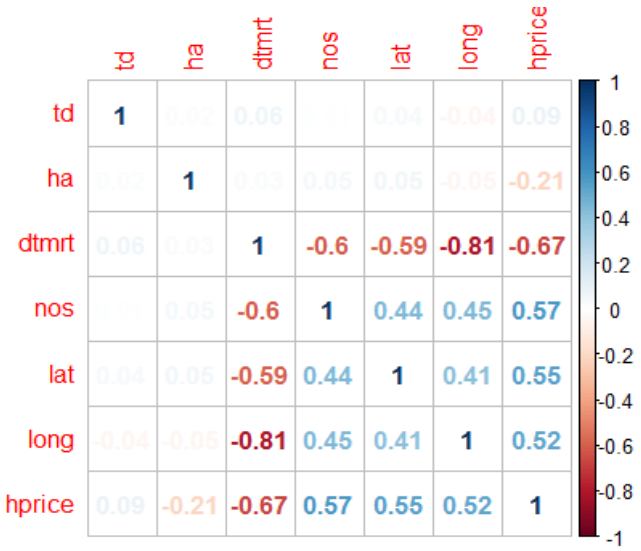
- **Transaction date(td**)-continuous
- **House age(ha)-**continuous
- **Distance to nearest metro station(dmrt)-** continuous
- **Number of convenience stores in locality(noc)-** ordinal
- **Latitude & Longitude(lat& long)-** continuous

3)**Omitted variables:**

- **Renovations in lifetime of house(**May affect condition of house**)**
- **Type of House(**It can capture details like facilities in house, type of foundation, Build Quality of house etc. **)**
- **Size of house (**although the house prices given are per square feet but not always house prices are proportional to area**)**

Exploratory Data Analysis:

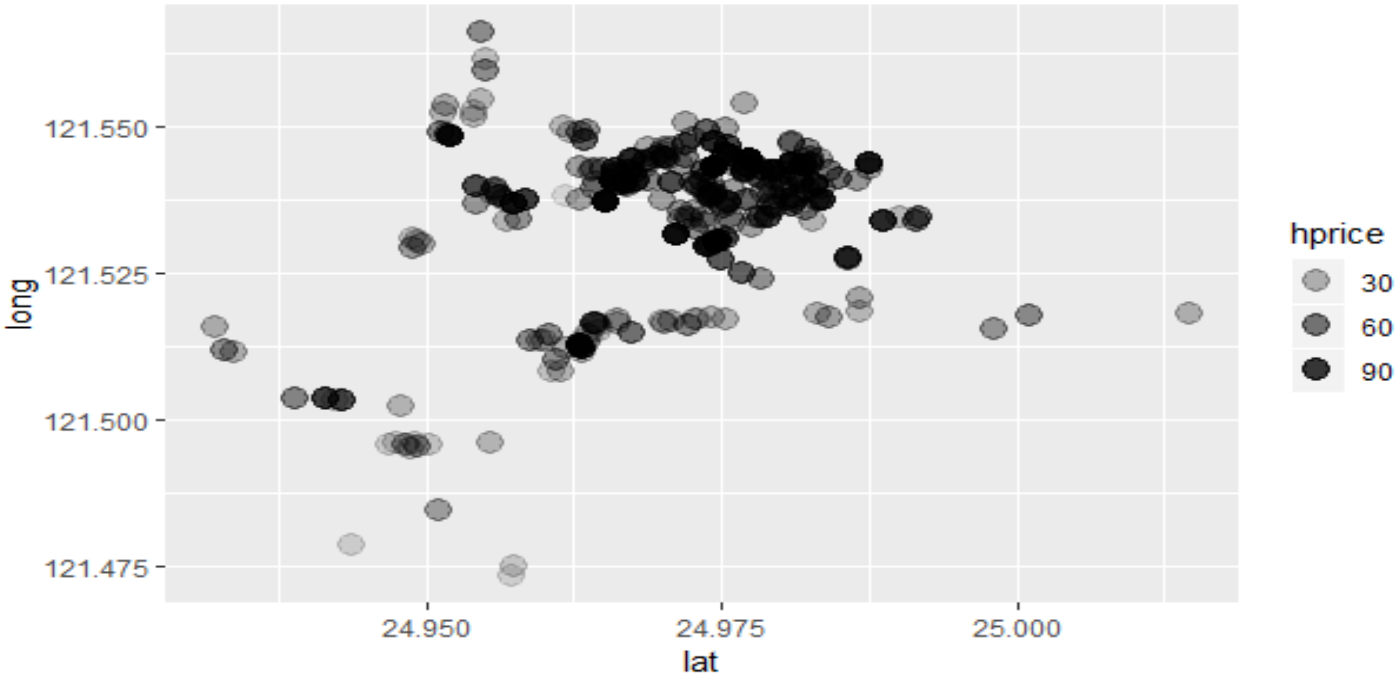
- **Correlation plot between variables:**



-From this plot this can be concluded that there is no perfect multicollinearity but there is imperfect multi collinearity between longitude and distance from metro station as it can be thought intuitively also that as longitudinal co-ordinate changes then accordingly distance from metro station would also change.

-All 3 attributes longitude, latitude and Distance from nearest railway station gives insight about the same thing i.e. the effect of location on prices. So instead using all 3 variables in regression it would be better to try only one or maximum two of them in the final equation.

- **Longitude vs Latitude with house price as decision variable for darkness of color.**



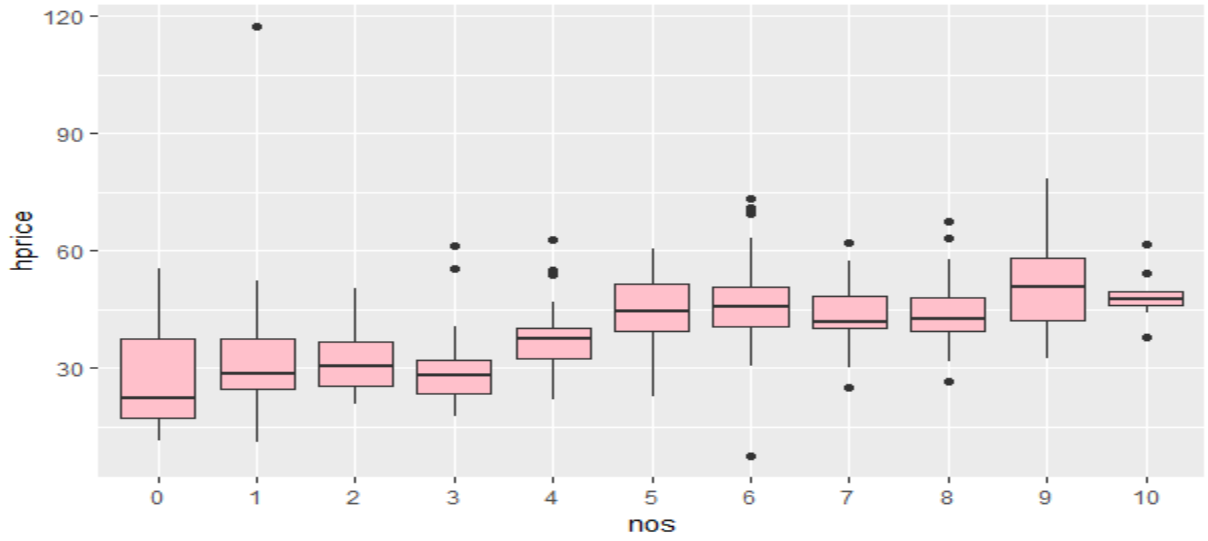
X-axis=latitude

Y-axis=longitude

Decision variable: house price

From the plot, it is clearly seen that there is a cluster of pointes where price of houses are high why?

- **Boxplot of house price vs no. of stores.**

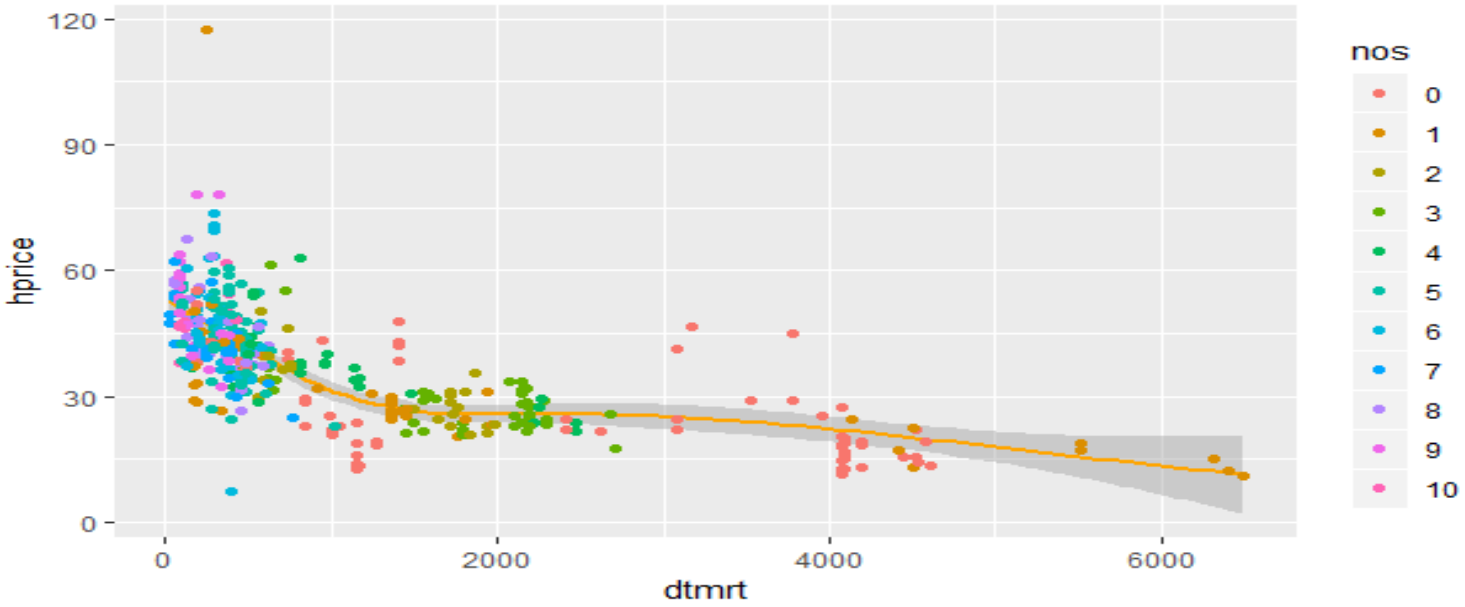


X-axis=No. of convenient stores in a locality.

Y-axis=House price.

This graph explains about variability of house price I accordance with no of stores in locality. It can be observed from graph that the houses with less no of stores near them have relatively lower prices.

- **Scatter plot of House price with distance from market decision variable being no of stores.**

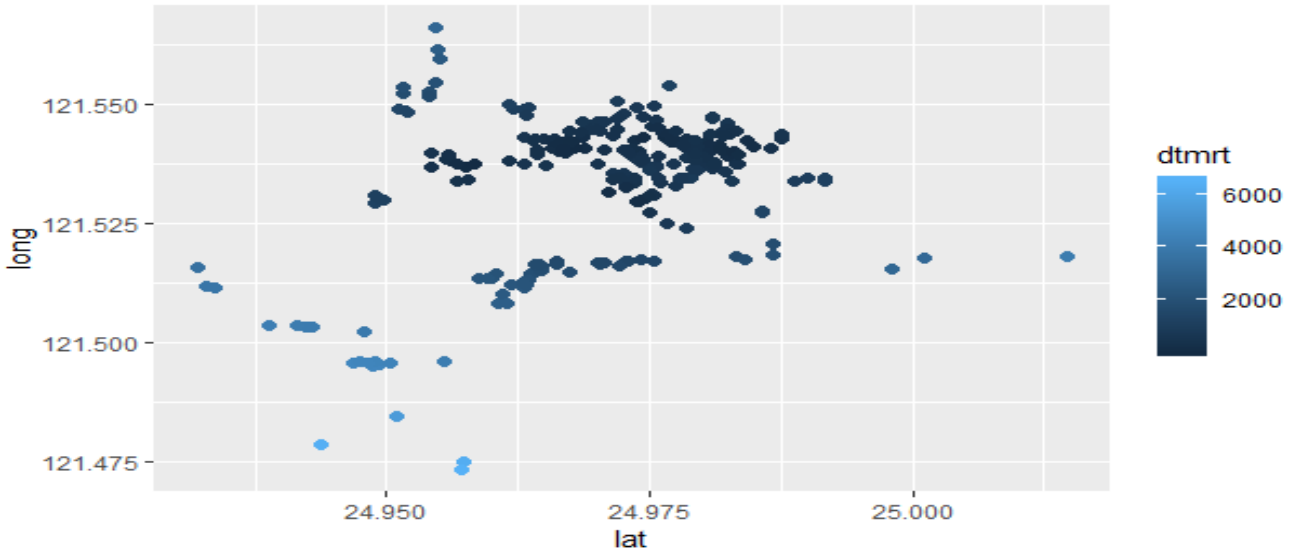


X-axis=Distance from mart.

Y-axis=House price.

It can be drawn from this graph that the houses that are nearer to metro have higher sales price and also have higher no of stores which can be seen from graph.

- **Scatter plot of longitude vs latitude decision variable being distance from mart:**

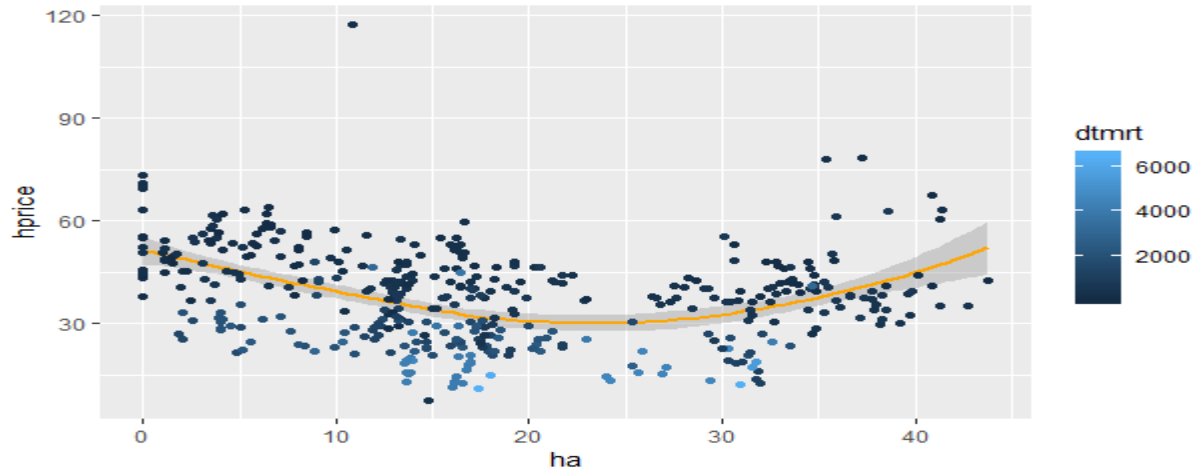


X-axis=latitude

Y-axis=Longitude

The darker region is the region closer to the metro station, that explains our earlier observed pattern that in that particular cluster, prices are high because of proximity to station.

- **Scatter plot of house price with age of house decision parameter being distance from metro station:**

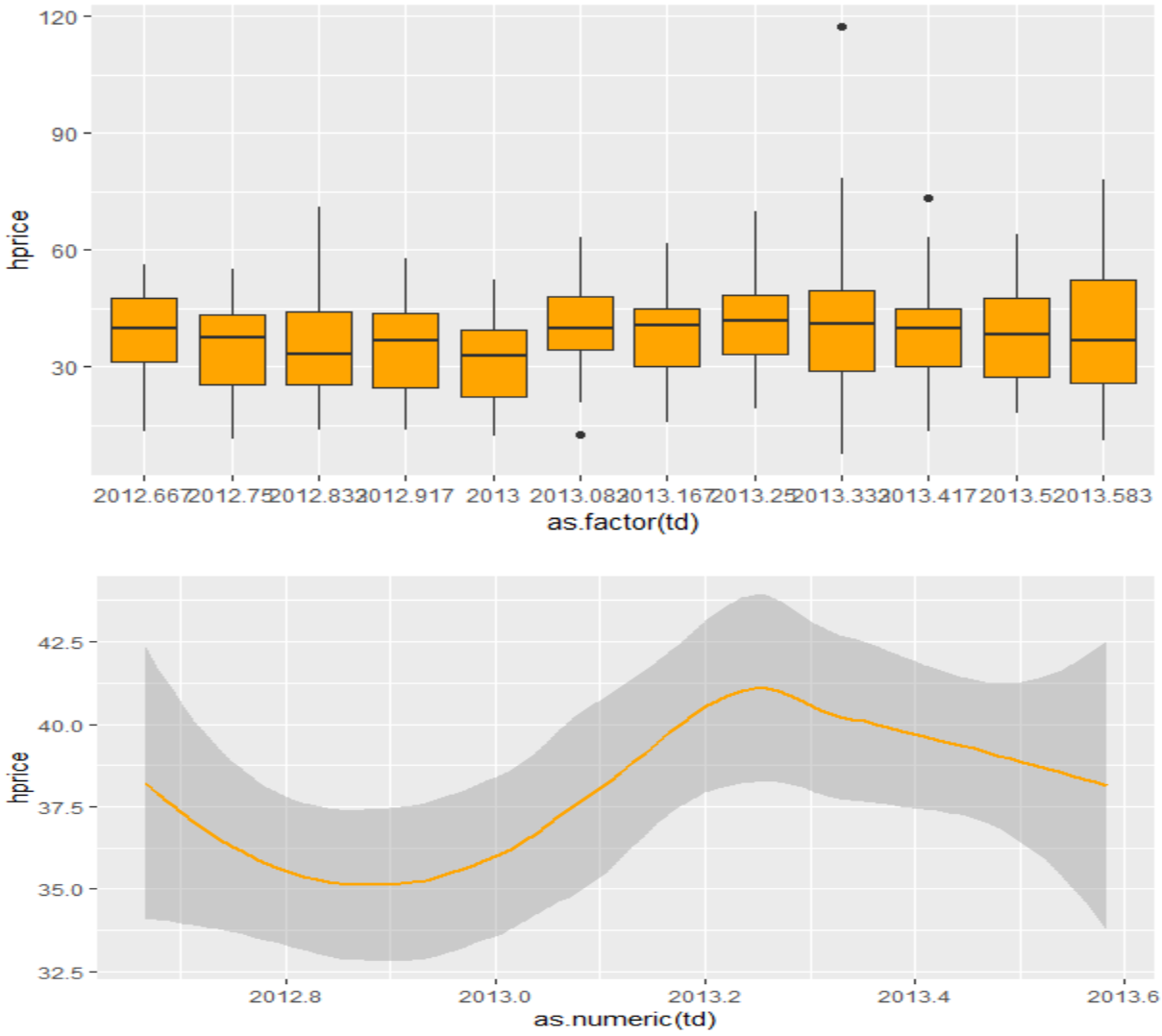


X-axis=Age of house.

Y-axis=House price.

Interesting thing as we might expect intuitively house price increases with decreasing age, that is not the case here that explains role of other variables in analysis.

- **Plot of House price vs transaction year:**



X-axis=transaction date

Y-axis=House price.

Plot shows that there is a non-linear variation in prices of houses with time which explains role of various economic factors affecting price of houses.

X-axis =Transaction in numerical format

Y-axis=House price

Linear Models:

Linear regression is a method in which we fit regression line between a dependent variable and one or more independent variables. We formulated 12 different models to understand their relationship with house sales price of a locality. Following is the list of models formulated in this study:

(Regression Models)

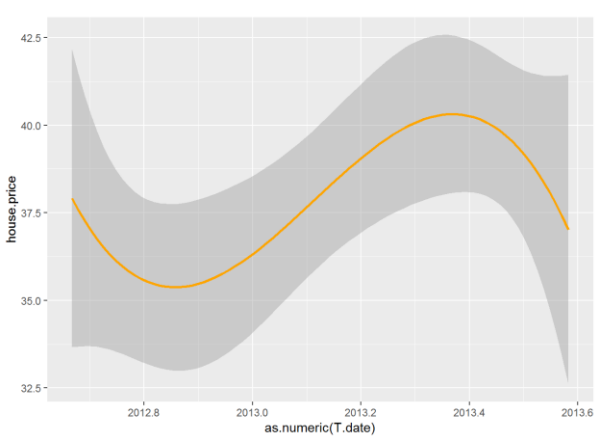
Model 1:

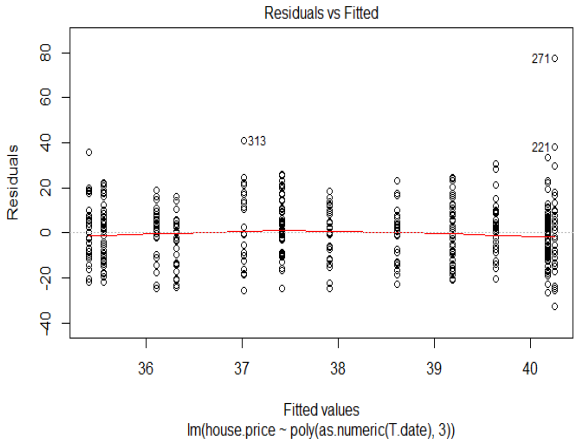
Sales price=37.980+24.1926*HouseTransctionDate-1.41*HouseTransctionDate^2-25.37*HouseTransctionDate^3

(.6627) (13.5458) (13.5458) (13.5458)

Residual standard error: 13.55 on 410 degrees of freedom

Multiple R-squared: 0.0161, Adjusted R-squared: 0.0089, p-value=0.0086





From the scatter plot (drawn in gg2plot in R) we saw there is a non-linear relationship. so we built a non-linear model. With a beta's value of 24.1926,-1.41and 25.37. The numeric value of beta is large, the standard error is significantly small. Hence we get a p-value close to zero and all beta's are statistically significant. Thus, we rejected the null hypothesis that sales price has no effect on House transaction date at 5% level of significance as the p-value is very low.

The R-Square value of this model is 0. 0089. It means that we can explain 0.89% of the variations in Sales price by the variable house transaction date. Our R-square value is still too much low such that there seems possibility of omitted variable bias.

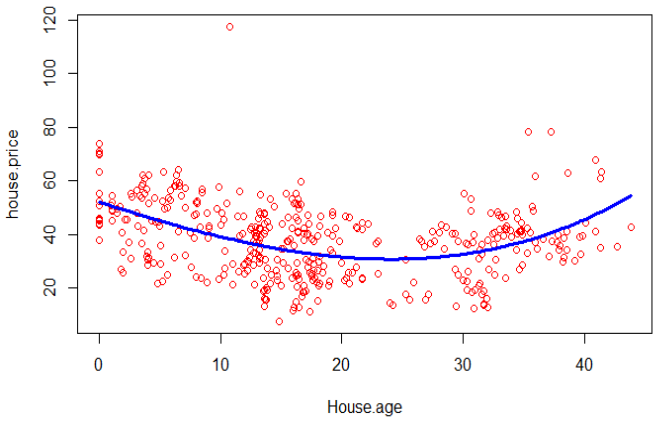
Model 2:

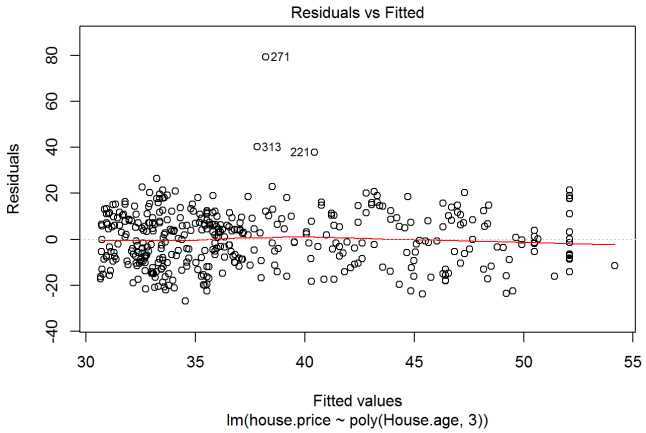
Sales price=37.980-58.2253*HouseAge+109.6353*HouseAge^2+14.9804*HouseAge^3

(.5986) (12.1802) (12.1802) (12.1802)

Residual standard error: 12.81 on 410 degrees of freedom

Multiple R-squared: 0.2045, Adjusted R-squared: 0.1987, p-value: 2.2*e-16





From this model we can say that there is a polynomial relationship between sales price and House age with beta value of -58.2253,109.653 and-14.9804. All beta's have t-statistic value of greater than 1.96. Hence it is statistically significant. We reject the null hypothesis that sales price has no effect on House transaction date as the p-value is very low. The R-Square value of this model is 0. 1987. It means that we can explain 19.87% of the variations in Sales price by the variable house age. Our R-square value is still too much low such that there seems possibility of omitted variable bias.

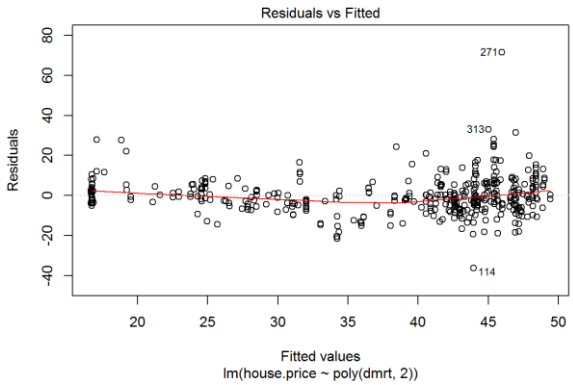
From residual plot this can be concluded that there is no heteroskedasticity as the graph's residue value does not changes with change in fitted value.

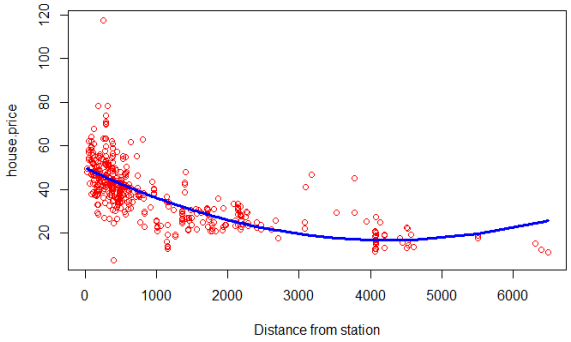
Model 3:

Sales price=37.980-186.265*DistanceFromMetro+74.824* DistanceFromMetro ^2		
(.461)	**(9.381)**	**(9.381)**

Residual standard error: 9.381 on 411 degrees of freedom

Multiple R-squared: 0.527, Adjusted R-squared: .5247,p-value:2.2*e-16





From this model we can say that there is a polynomial relationship between sales price and distance to metro station with beta value of -186.265, 74.828. The t-statistic value of beta's are large, the standard error is significantly small. Hence it is statistically significant. We reject the null hypothesis that sales price has no effect on House transaction date as the p-value is low at 5% level of significance**.** The R-Square value of this model is 0. 527. It means that we can explain

52.7% of the variations in Sales price by the variable Distance from metro station. Our R-square value is still too much low such that there seems possibility of omitted variable bias.

From residual it can be concluded that the data is not heteroskedastic as there is not too much variation of residue as the fitted value changes.

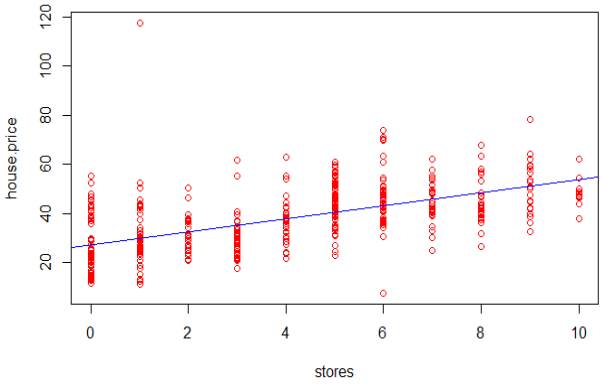
Model 4:

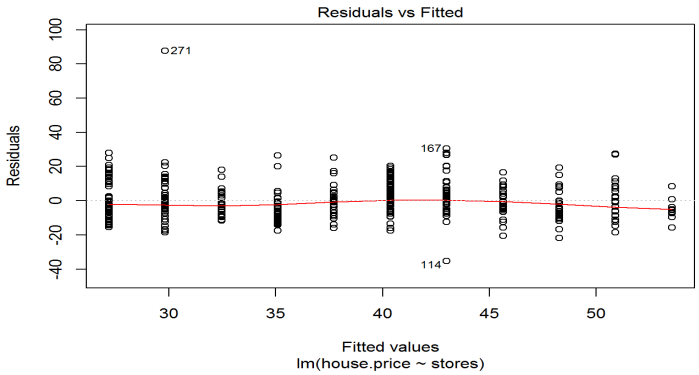
Sales price=27.1811+2.6377*No.of.Stores

(.9419) (.1868)

Residual standard error: 11.18 on 412 degrees of freedom

Multiple R-squared: 0.326, Adjusted R-squared: 0.3244,p-value:2.2*e-16





From this model we can say that there is a linear relation between sales price and no. of stores with beta value of 2.6377 which mean that for unit change in no. of store there will be 2.6377 increases in sales price of house. Even if the numeric value of beta is not large, the standard error is significantly small. Hence it is statistically significant. We reject the null hypothesis that sales price has no effect on House transaction date as the p-value is less**.** The R-Square value of this model is 0. 324. It means that we can explain 32.4% of the variations in Sales price by the variable no. of stores in locality. Our R-square value is still too much low such that there seems possibility of omitted variable bias.

From residual plot it can be concluded that the data is not heteroskedastic as there is not too much variation of residue as the fitted value changes.

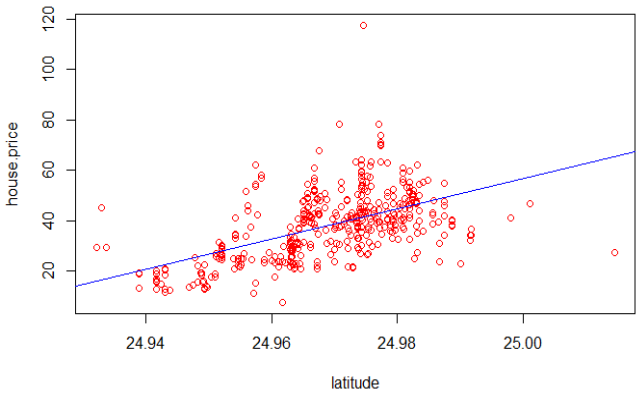
Model 5:

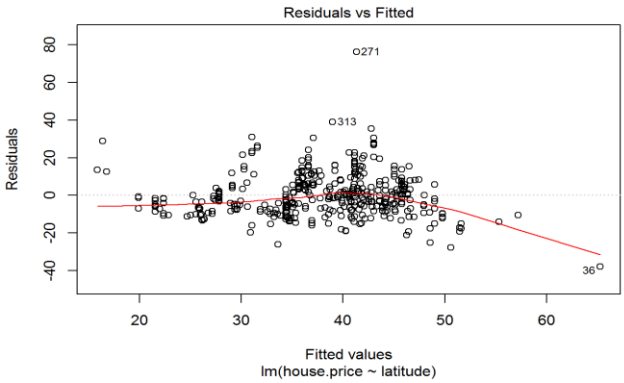
Sales price=-14917.68+598.97*latitude

(1129.66) (45.24)

Residual standard error: 11.41 on 412 degrees of freedom

Multiple R-squared: 0.2985, Adjusted R-squared: 0.2967,p-value:2.2*e-16





From this model we can say that there is a linear relation between sales price and latitude with beta value of 598.97 which mean that for unit change in latitude there will be 598.97 increase in sales price of house. Even if the numeric value of beta is large, the standard error is significantly small. Hence it is statistically significant. We reject the null hypothesis that sales price has no effect on latitude as the p-value is very less. The R-Square value of this model is 0.2985. It means that we can explain 29.85% of the variations in Sales price by the variable latitude. Our R-square value is still too much low such that there seems possibility of omitted variable bias.

From residual plot the red line diverges from the zero line so we concluded that there is a polynomial relationship.We further improves this model in model 5a.

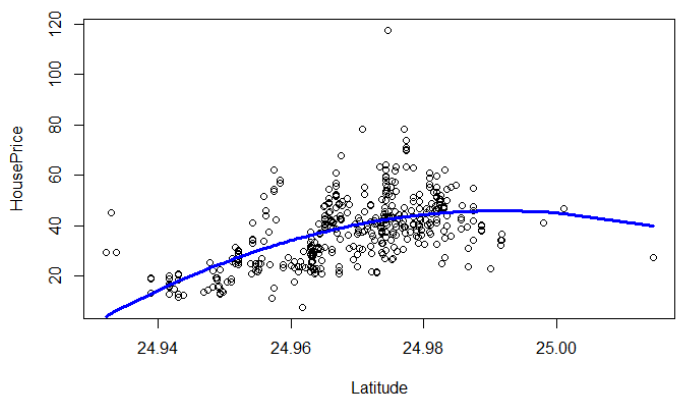
Model 5a:

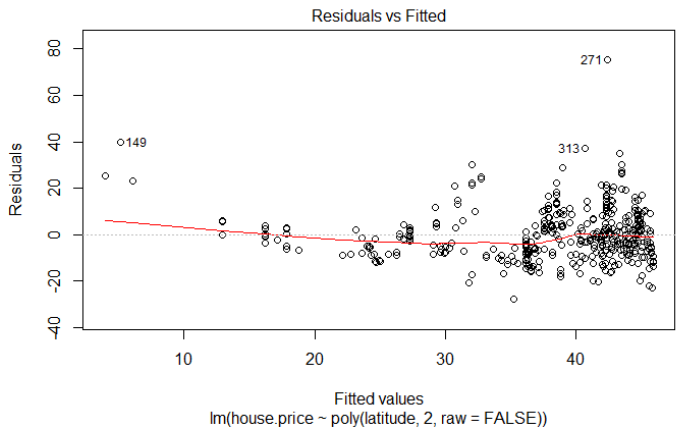
Sales price=-37.9802+151.0428*latitude-52.750latitude^2

(.5467) (11.124) (11.124)

Residual standard error: 11.12 on 411 degrees of freedom

Multiple R-squared: 0.3348, Adjusted R-squared: 0.3316,p-value:2.2*e-16





As from previous graph it was seen that residue was changing for fitted values so to remove heteroskedasticity polynomial of 2 degree is applied as shown in the graph. The beta value of 151.0428and -52.75 is large enough and, the standard error is significantly small. Hence it is statistically significant. We reject the null hypothesis that sales price has no effect on latitude as the p-value is very less. The R-Square value of this model is 0. 2985. It means that we can explain 29.85% of the variations in Sales price by the variable latitude. Our R-square value is still too much low such that there seems possibility of omitted variable bias.

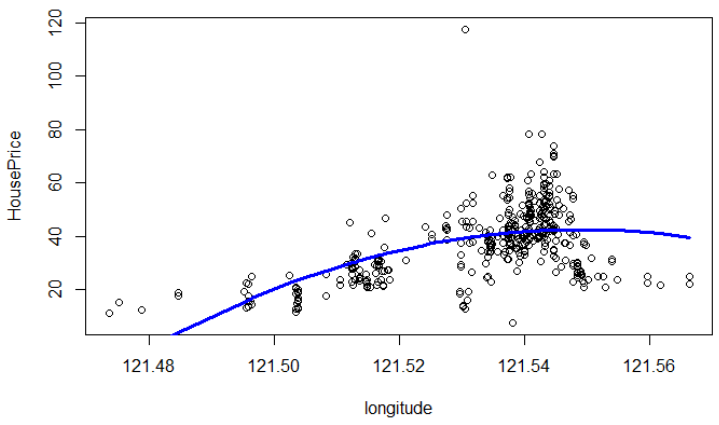
Model 6:

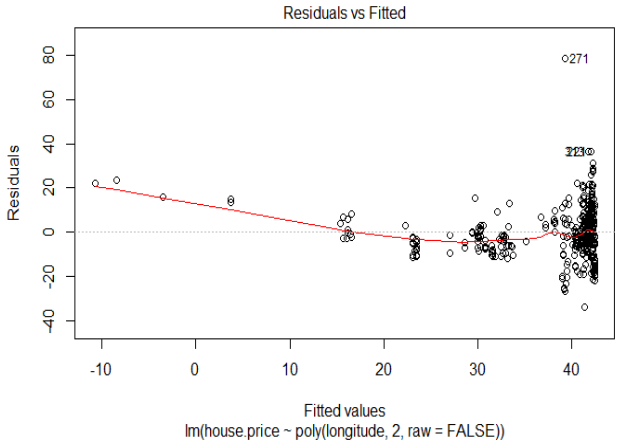
Sales price=-37.9802+144.69*longitude-58.460longitude^2

(.5534) (11.2596) (11.2596)

Residual standard error: 11.26 on 411 degrees of freedom

Multiple R-squared: 0.3185, Adjusted R-squared:0 .3152,p-value:2.2*e-16





From this model we can say that there is a non-linear relationship between sales price and latitude with beta value of 37.9802, 144.69 and -58.460. With t-statistic greater than 1.96 for all variable all beta's are statistically significant. We reject the null hypothesis that sales price has no effect on longitude as the p-value is very less. The R-Square value of this model is 0.3185. It means that we can explain 29.85% of the variations in Sales price by the variable longitude. Our R-square value is still too much low such that there seems possibility of omitted variable bias and also that longitude is strong predictor which should be used.

Observations:

- Model 5a is the best fit as it has the highest adjusted R square value of 0.332. Both the intercept and estimates for this model is statistically significant. The standard error of regression is also low for this model. This model explains nearly 33.2% of the variance in house price which is still quite low.
- Model 4 has also nearly same adjusted R squared value as model 5a but the estimate is not statistically significant. So we fail to reject the null hypothesis that beta value is zero at 5% level of significance.
- Model 3 has intercepts and estimates which are statistically significant but has low adjusted R squared value
- Model 1 and model 2 has very low adjusted R squared value. So they cannot explains much variance in house price.

Table for regression models.

		Model 1 (H.price)	Model 2 (H.price)	Model 3 (H.price)	Model 4 (H.price)	Model 5a (H.price)	Model 6 (H.price)
Intercept	$(X)^0$ (SD)	37.980*** (0.666)	37.980*** (0.599)	37.980*** (0.461)	27.181*** (0.942)	37.980*** (0.547)	37.98*** (0.553)
Date	$poly(date)^1$ (SD)	24.193* (13.546)					
	$poly(date)^2$ (SD)	-1.414 (13.546)					
	$poly(date)^3$ (SD)	-25.374* (13.546)					
House age	$poly(house\ age)^1$ (SD)		-58.225*** (12.180)				
	$poly(house\ age)^2$ (SD)		109.635*** (12.180)				
	$poly(house\ age)^3$ (SD)		14.980 (12.180)				
Distance	$poly(distance)^1$ (SD)			-186.265** (9.381)			
	$poly(distance)^2$ (SD)			74.824*** (9.381)			
Stores	(store) (SD)				2.638*** (0.187)		
Latitude	$poly(latitude)^1$ (SD)					151.063*** (11.124)	
	$poly(latitude)^2$ (SD)					-52.751*** (11.124)	
Longitude	$poly(longitude)^1$ (SD)						144.697*** (11.260)
	$poly(longitude)^2$ (SD)						-58.466*** (11.260)
	Summary						
	SER	13.546	12.180	9.381	11.184	11.124	1.260
	R Squared	0.016	0.204	0.527	0.326	0.335	0.319
	Adjusted R Squared	0.009	0.0003424	0.199	0.324	0.332	0.315

Significance level: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple Regressors:

Model 7:

Sales price=-1.444e+04 +(5.146e+00 *transaction.Date) –(2.697e-01 *HouseAge)

(6.776e+03) (1.557e+00) (3.853e-02)

-(4.488e-03 *distanceFromMetro)+ (1.133e+00* Stores)+(2.255e+02 *Latitude) – (1.242e+01* longitude)

(7.180e-04) (1.882e-01) (4.457e+01) (4.858e+01)

Residual standard error: 8.858 on 407 degrees of freedom

Multiple R-squared: 0.5824, Adjusted R-squared:0 .5762, F-statistic:94.59

All the variables have t-statistic value of greater than 1.96 and are statistically significant except for longitude the t-statistic is -0.256 which shows longitude is not statistically significant variable. The F-statistic is 94.59 so we **reject** the null hypothesis that all the beta values are zeros.

Model 8:

Sales price=-1.596e+04+(5.135e+00*transaction.Date) –(2.694e-01*HouseAge)

(6.776e+03) (1.557e+00) (3.853e-02)

-(4.353e-03 *distanceFromMetro)+ (1.136e+00 *Stores)+(2.269e+02*Latitude)

(7.180e-04) (1.882e-01) (4.457e+01)

Residual standard error: 8.848 on 408 degrees of freedom

Multiple R-squared: 0.5823, Adjusted R-squared: 0.5772, F-statistic:113.80

Model 8 is constructed by removing longitude which was a statistically non significant variable. Here F-statistic is 113.80 so we so we **reject** the null hypothesis that all the beta values are zeros. All remaining variables are statistically significant with very low p-values.

Model 9(Log-linear model):

Log(Sales price)=-5.117e+02+(1.355e-01*transaction.Date) –(6.967e-03*HouseAge)

(1.695e+02) (3.896e-02) (9.641e-04)

-(1.455e-04*distanceFromMetro)+ (2.775e-02*Stores)+(7.925e+00*Latitude) -(3.688e-01* longitude)

(1.797e-05) (4.708e-03) (1.115e+00) (1.216e+00)

Residual standard error: .2216 on 407 degrees of freedom

Multiple R-squared: 0.6857, Adjusted R-squared: 0.6811,F-statistic:148

From this model it can be interpreted that adjusted R squared is increasing by applying log over sales price. when we plotted the graphs individually we saw that some variables have there was non- linear relation with sales. All beta's values are close to zero and much significant. We also reject that null hypothesis with F-statistic value of 148.

Model 10(Linear-log model):

Sales price=-1.337e+05+(1.318e+04*log(transaction.Date)) –(2.304e-01*log(HouseAge))

(2.916e+04) (2.876e+03) (3.538e-02)

-(6.675e+00*log(distanceFromMetro))+ (3.646e-01*log(Stores))+(7.048e+03*log(Latitude)) +(2.256e+03* log(longitude))

(5.873e-01) (1.914e-01) (9.370e+02) (4.206e+03)

Residual standard error: .2216 on 407 degrees of freedom

Multiple R-squared: 0.6857, Adjusted R-squared: 0 .6811,F-statistic:127.4

In this model we applied log over all independent variables. The result is that there is increase in adjusted R squared value. All beta's are significant with very low p-value's. F-statistic is 127.4, so we reject the null hypothesis that all beta's are zero.

Model 11(Log-log model):

Log(Sales price)=-3.412e+03+(3.370e+02*log(transaction.Date)) –(6.068e-03*log(HouseAge))

(5.674e+02) (7.462e+01) (9.180e-04)

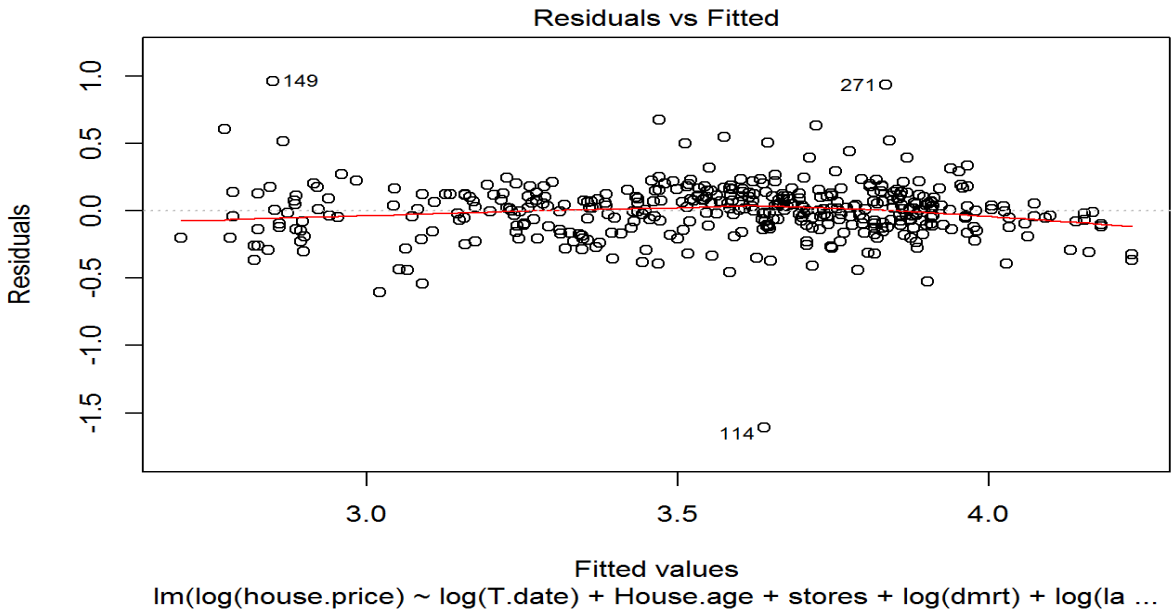
-(1.944e-01*log(distanceFromMetro))+ (1.027e-02*log(Stores))+(2.652e+02*log(Latitude))

(1.333e-02) (4.966e-03) (2.395e+01)

Residual standard error: .2097 on 408 degrees of freedom

Multiple R-squared: 0.7181, Adjusted R-squared: 0.7146,F-statistic:207.8

In this model we applied log over both independent and dependent. Here F-statistic is 207.80 so we so we **reject** the null hypothesis that all the beta values are zeros. The adjusted R-squares value further increases and all the beta's value is significant with very low p-value .



From residual plot it can be concluded that the data is not heteroskedastic as there is not too much variation of residue with changes in fitted values. But zero line is still not perfectly flat. So we decided to build a model taking the best out of best models we found so far.

Model 12(Picking Best Individual model):

Log(Sales price)=-6.533e+02+(8.980e-01* (transaction.Date) +2.020e-01* (transaction.Date)^2) -(1.665e+00* (HouseAge)

(8.319e+01) (2.089e-01) (2.089e-01) (2.087e-01)

-1.166e+00* (HouseAge)^2)) -(3.819e+00* (distanceFromMetro) -1.107e+00* (distanceFromMetro)^2)

(2.223e-01) (3.117e-01) (2.256e-01)

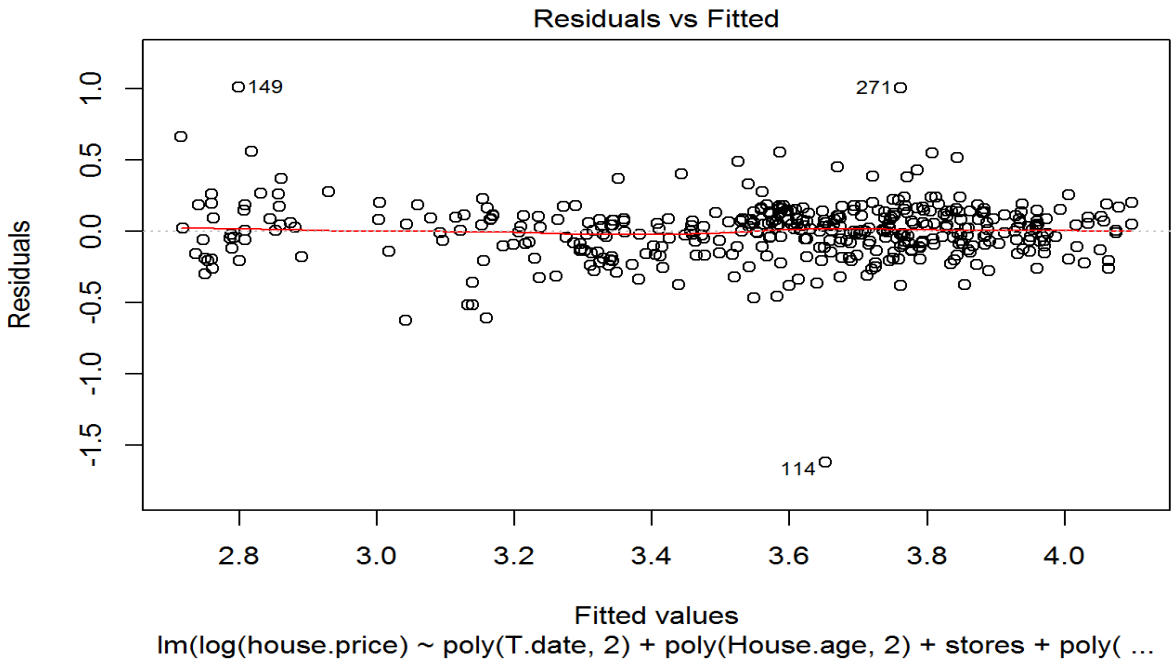
+ (1.706e-02* (Stores)) + (2.041e+02*log(Latitude))

(4.724e-03) (2.586e+01)

Residual standard error: 0.22069 on 405 degrees of freedom

Multiple R-squared: 0.7274, Adjusted R-squared: 0.7221, F-statistic:135.10

In this model we applied those function to variables which best describes each individual model from (model1-model11) . All the beta's value is significant with very low p-value .Here F-statistic is 135.10 so we so we reject the null hypothesis that all the beta values are zeros the results were in favor of our analysis so far. We get the best R squared value and the perfectly flat residual plot (as shown below)



From the residual fit it can be observed that the data is homoscedastic and our model explains sales price very well.

Observations:

- Model 12 gives the best adjusted R Squared value of 0.722 with a very low standard error of regression 0.207. All the estimates and intercepts are statistically significant except poly(date)^2 . we reject the null hypothesis that all beta's are zero as F-statistic value is 135.10 with 405 degree of freedom.
- Model 11 also has high adjusted R squared value close to model 12 but stores variable are non significant. This model is over fitting the data. This is an example of low bias and high variance.
- Model 9 has lower adjusted R squared value and has high standard error of regression. Moreover the longitude variable is not statistically significant.
- Model 7 and model 8 has low adjusted R squared value compared to all combined models below. The standard error of regression is also very high for these two model.

Table for multiple regression model:

		Model 7 (H.price)	Model 8 (H.price)	Model 9 log(H.price)	Model 10 (H.price)	Model 11 log(H.price)	Model 12 log(H.price)
Intercept	$(X)^0$ (SD)	-14437.100*** (6775.671)	-15959.260*** (3233.450)	-511.676*** (169.534)	-133670.100*** (29152.850)	-3412.312*** (567.417)	-653.327*** (83.191)
Date	(date) (SD)	5.146*** (1.557)	5.135*** (1.555)	0.135*** (0.039)			
	poly(date)^1 (SD)						0.898*** (0.209)
	poly(date)^2 (SD)						0.202 (0.209)
	log(date) (SD)				13177.120*** (2875.734)	337.014*** (74.619)	
House age	(house age) (SD)	-0.270*** (0.039)	-0.269*** (0.038)	-0.007*** (0.001)	-0.230*** (0.035)	-0.006*** (0.001)	
	poly(house age)^1 (SD)						-1.665*** (0.209)
	poly(house age)^2 (SD)						1.166*** (0.222)
Distance	(distance) (SD)	-0.004*** (0.001)	-0.004*** (0.005)	-0.0001*** (0.00002)			
	poly(distance)^1 (SD)						-3.819*** (0.312)
	poly(distance)^2 (SD)						1.107*** (0.226)
	log(distance) (SD)				-6.675*** (0.587)	-0.194*** (0.013)	
Stores	(store) (SD)	1.133*** (0.188)	1.136*** (0.188)	0.028*** (0.005)	0.365* (0.191)	0.010** (0.005)	0.017*** (0.005)
Latitude	(latitude) (SD)	225.473*** (44.567)	226.882*** (44.174)	7.925*** (1.115)			
	log(latitude) (SD)				7048.356*** (936.961)	265.215*** (23.946)	204.132*** (25.85)
Longitude	(longitude) (SD)	-12.424 (48.582)		0.369 (1.216)			
	log(longitude) (SD)				2256.146 (4206.019)		
	Summary						
	SER	8.858	8.848	0.222	8.079	0.210	0.207
	R Squared	0.582	0.582	0.686	0.653	0.718	0.727
	Adjusted R Squared	0.576	0.577	0.681	0.715	0.715	**0.722**

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Test for Multi Collinearity using Variable Inflation factor:

1)Check for all independent variables:

modelA=lm(hprice~., data=train)- all variables included,

vif(modelA):

	VIF
Transaction date	1.016603
House age	1.015215
dmrt	4.332612
stores	1.617776
latitude	1.610762
longitude	2.931430

2) Since dmrt has high vif value, we tried removing all location variables one by one and one with best linear fit was found to be:

modelB=lm(hprice ~ .-long, data=train)- we excluded longitude variable.

vif(modelB):

	VIF
Transaction date	1.015868
House age	1.014089
dmrt	2.018013
stores	1.611872
latitude	1.585901

3)Checking on our best R squared model(including non-linearities and log):

model12<-lm(log(house.price)~poly(T.date,2)+poly(House.age,2)+stores+poly(dmrt,2)+log(latitude))

vif(model12):

	GVIF	Df	GVIF^(1/(2*Df))
poly(Transaction date,2)	1.038331	2	1.009448
poly(House age,2)	1.174104	2	1.040942
stores	1.868147	1	1.366802
poly(dmrt, 2)	2.560972	2	1.265031
log(latitude)	1.593353	1	1.262281

-so, we can conclude that since there is no higher value of GVIF in final model, model has no significant multicollinearity.

CONCLUSION:

- From the above results we conclude that since our data was non-linear, the log and non-linear models were able to explain the variance better.
- Distance from metro station found to be most important attribute deciding the prices of houses which explains role of location in deciding prices.
- It seemed surprising that transaction date has very little effect in model. This can be due to very small range of dates covered in given observations and this could have affected R squared as well due to bias in data.
- Our model is able to explain 72.7 of the variance, rest 28.3% explains the role of omitted variables discussed earlier.

Codes:

Data Cleaning

```
library(knitr)

#read excel file

data<-read.csv("C:/Users/DANISH/Desktop/project_sem2/SMBA_project1/project_data.csv")

#checking for any missing values

sum(is.na(data))

Data summary

summary(data)

str(data)

#correlation check

correlation Matrix

round(cor(data),2)
```

Data Visulaization

```
library(corrplot)

M <- cor(data)

corrplot(M, method = "number")

plot(data)

library("ggplot2", lib.loc="~/R/win-library/3.5")

ggplot(data=data,mapping=aes(x=latitude, y=longitude, alpha= house.price))+ geom_point(size=4)

#theres is a dense region where prices are high? why?

ggplot(data=data,mapping=aes(x=stores, y=house.price))+ geom_boxplot(fill="pink")

ggplot(data=data,mapping=aes(x=dmrt, y=house.price))+ geom_smooth(color="orange") + geom_point(aes(color=
stores))
```

```
#no of stores effect

ggplot(data=data,mapping=aes(x=latitude, y=longitude, color= dmrt))+ geom_point(size=2)



# we can see the high price points near mrt station in graph.

ggplot(data=data,mapping=aes(x=House.age, y=house.price))+ geom_smooth(color="orange") + geom_point(aes(color=
dmrt))


# interesting thing as we might expect intuitively house price increases with decreasing age, that is not the case here
that explains role of other variables in analysis.

summary(data$T.date)

ggplot(data=data,mapping=aes(x=as.factor(T.date), y=house.price))+ geom_boxplot(fill="orange")

ggplot(data=data,mapping=aes(x=as.numeric(T.date), y=house.price))+ geom_smooth(color="orange")


#effect of time of deal

cor(as.numeric(data$T.date),data$house.price)

Multicollinearity check

plot(data$dmrt,data$longitude, xlab="Distance from station", ylab="Longitude", col="dark green")

 #There is no perfect multicollinearity


checking for outlier

#checking for outlier

boxplot(data$T.date , main="Transaction date",

        col = "brown")

boxplot(data$House.age ,main="House age",

        col = "blue")

boxplot(data$dmrt ,main="Distance from stores",

        col = "red") #large potential outliers exist

boxplot(data$latitude , main="Latitude",

        col = "green")

boxplot(data$stores , main="Stores",

        col = "grey")

boxplot(data$longitude , main="Longtitude",
```

```
        col = "dark green")
boxplot(data$T.date , main="Transaction date",
        col = "brown", outline=FALSE)
boxplot(data$House.age ,main="House age",
        col = "blue",outline=FALSE)
boxplot(data$dmrt ,main="Distance from stores",
        col = "red",outline=FALSE) #large potential outliers exist
boxplot(data$latitude , main="Latitude",
        col = "green",outline=FALSE)
boxplot(data$stores , main="Stores",
        col = "grey",outline=FALSE)
boxplot(data$longitude , main="Longtitude",
        col = "dark green",outline=FALSE)
```

Model Fitting

single Regressor

Model 1

```
model1=lm(data=data,house.price~poly(as.numeric(T.date),3))
summary(model1)
ggplot(data=data,mapping=aes(x=as.numeric(T.date), y=house.price))+ geom_smooth(color="orange",
method="lm",formula=y~poly(x,3))
```

 # Model 2

```
model2=lm(data=data, house.price~ poly(House.age,3))
summary(model2)
#plot(House.age,house.price, col="red")
#lines(smooth.spline(House.age,predict(model2)),col="blue",
    #lwd=3)
plot(model2)
```

Model 3

```
model3=lm(data=data, house.price~ poly(dmrt,2))
summary(model3)
```

```
#plot(dmrt,house.price, col="red" ,xlab="Distance from station")
#lines(smooth.spline(dmrt,predict(model3)),col="blue",
    #lwd=3)
plot(model3)
```

Model 4

```
Model 4 (house.price=27.1811+2.6377*stores)
attach(data)
plot(as.factor(stores),house.price,
   col="blue",
xlab="stores", ylab="HousePrice")
plot(stores,house.price, type="p", col="red")

model4<-lm(house.price~stores)
summary(model4)
abline(model4, col="Blue" )
plot(model4)
```

Model 5

```
plot(latitude,house.price, type="p", col="red")
model5<-lm(house.price~latitude)
abline(model5, col="Blue" )
summary(model5)
plot(model5)
```

Model 5

```
plot(latitude,house.price,
xlab="Latitude", ylab="HousePrice")
model5<-lm(house.price~poly(latitude,2, raw=FALSE))
summary(model5)
lines(smooth.spline(latitude,predict(model5)),col="blue",
lwd=3)
```

Model 6

```
plot(longitude,house.price,
xlab="longitude", ylab="HousePrice")
model6<-lm(house.price~poly(longitude,2, raw=FALSE))
summary(model6)
lines(smooth.spline(longitude,predict(model6)),col="blue",
lwd=3)
```

Multiple Regressor

Model 7

```
model7<-lm(house.price~T.date+House.age+dmrt+ stores+latitude+longitude)
summary(model7)
```

Model 8

```
model8<-lm(house.price~T.date+House.age+dmrt+ stores+latitude)
summary(model8)
```

Model 9 LOg-linear model

```
model9<-lm(log(house.price)~T.date+House.age+dmrt+ stores+latitude+longitude)
summary(model9)
```

Model 10 Linear-log model

```
model10<-lm(house.price~log(T.date)+House.age+stores+log(dmrt)+log(longitude)+log(latitude))
summary(model10)
```

Model 11 Log-Log Model

```
model11<-lm(log(house.price)~log(T.date)+House.age+stores+log(dmrt)+log(latitude))
summary(model11)
plot(model11)
```

Model 12 (Picking Best Individual model)

```
model12<-lm(log(house.price)~poly(T.date,2)+poly(House.age,2)+stores+poly(dmrt,2)+log(latitude))
summary(model12)
plot(model12)
```