UNDERSTANDING AND PREDICTING THE SALE PRICE OF HOUSES IN AMES, IOWA

By:

TATHAGATA CHAKRABORTY

(MASTER OF SCIENCE, BUSINESS ANALYTICS)

STUDENT NUMBER: 40379173

MODULE CODE: MGT_7177

MODULE NAME: STATISTICS FOR BUSINESS

QUEEN'S UNIVERSITY BELFAST | MANAGEMENT SCHOOL

CONTENTS

Pg no.

INTRODUCTION

The most frequent factors influencing housing demand and supply in industrialised nations are fundamental home price variables. These include demographic indicators as well as economic and financial determinants. This report elaborates on how certain factors affect the sale price of houses in the city of Ames in Iowa, USA. Having a pre-emptive notion of which attributes influence the sale prices of a house can immensely aid real estate agents in designing the price structure of different houses as well as further analyse the probable prices of houses bearing certain characteristics.

BACKGROUND

The impact of housing on economic and social variables makes it crucial to analyse the elements that affect housing prices. First off, having a home improves inhabitants' health, residential mobility, and other social outcomes (Dietz et al., 2003). Second, housing is a good indicator that is intimately related to other markets and the overall state of the economy in the nation. The housing market is particularly linked to the financial sector because homes can be bought with mortgages as well as with residents' own money (Jureviciene et al. 2008). Finally, changes in house prices have an impact on the building industry as well as other economic factors like unemployment and inflation (Azbainis, 2014).

Historically, different hedonic-based techniques have been used to determine the correlation between house prices and housing attributes for many years (Adair et al., 1996, Selim, 2009). Meese and Wallace (2003) created hedonic-based regression techniques to analyse how market fundamentals affect the dynamics of home prices. Stevenson (2004) used data on the typical age of homes in Boston, Massachusetts to re-examine heteroscedasticity in hedonic housing price models. The acquired results confirmed the prior studies' indication of heteroscedasticity with reference to house age. Bin (2004) used a semi-parametric regression to estimate a hedonic pricing function and compared the performance of the price prediction to that of traditional parametric models. According to the findings, semi-parametric regression performed better in both in-sample and out-of-sample price predictions, and it may be used to predict prices (Bin 2004).

METHODOLOGY

However, hedonic-based methods have potential limitations relating to fundamental model assumptions and estimation(Park 2015). Pow et al. (2014) used an ensemble approach by combining KNN and Random Forest Technique to predict the price value of the property. They also used four regression techniques, namely Linear Regression, Support Vector Machine, K-Nearest Neighbours (KNN), and Random Forest Regression. Applying PCA (Principal Component Analysis) did not reduce the prediction error, while the ensemble technique predicted prices with the lowest error of 0.0985 (Manasa et al 2020). Thus, owing to the low error component of regression methodologies, for this scenario multi variate linear regression models have been used to estimate the relations between the attributes of the Ames house price dataset.

A set of five variables have been selected to form respective hypotheses and the bivariate analysis of these hypotheses have been conducted there on. Following the results of the hypotheses, linear regression models have been formulated using different combinations and the final model has been presented with the relevant inferential and predictive accuracy results.

Here under the five hypotheses are listed along with corroborating reasons as to why the said attribute has been considered for further testing.

[Note : H_i -> H – Hypothesis, i – index of the statement [1,5] as we are testing 5 hypotheses]

H_1 : Sale price of the house has a strong positive relationship with the living area of the house.

Thornes and McMillen (1998) and Lin and Evans (2000) found that size of living area is a controlling factor in determining residential property value. Furthermore, a larger house implies greater cost of infrastructural investment on the part of the construction company which naturally leads us to infer that considering moderate adjustments due to effect from additional factors, there must exist a positive relationship between the two.

H_2 : The sale prices of houses have a positive relationship with the quality of the house.

The quality of house directly impacts the quality of life in the dwelling of the occupants (Masri et al. 2018). Moreover, the better the quality of material and finish of the house, the greater is the initial investment and labour cost. This directly implies that the quality of the house must have a positive relationship with the sale price of the house.

H_3 : The sale price of houses has a strong positive relationship with total rooms in the house.

Fan(2006) suggests that property prices have the propensity to rise with the number of rooms in the property as it improves quality of life. Thus, it can be inferred that number of rooms may have a strong relation with the sale price of the house.

H_4 : There is a significant difference in mean Sale Price between houses that have central air conditioning versus the ones that do not.

Ames has a humidly continental climate (Mearns et al., 2012). It experiences hot summers and quite cold winters. Larsen et al. (2010) has also confirmed the positive effect that central air conditioning has on the house prices in USA. Considering the hot summers that Ames experiences, it is safe to assume that there shall be a significant variation between the prices of houses with and without central air conditioning.

H_5 : There is a strong positive relationship between the garage area and sale price of the house.

In 2020, the average home ownership in the city of Ames was 59.7% of the population where the employed population was estimated 66000 (DataUsa, 2020). The average car ownership in Ames and the whole of Iowa was 2 cars per household. The employed population in the state of Iowa was 1.52million with a 71.2% of home ownership status of the entire state population (DataUSA, 2020). Owing to these facts it is reasonable to believe that the area of garage space shall positively affect the price of the property.

ANALYSIS RESULTS AND DISCUSSION

A detailed exploratory analysis report of the hypotheses variables has been created using the **DataExplorer** package in R. (**See Appendix 2 – [DE**]**)

The dataset has been reduced to essential 16 attributes. The outliers and other data quality issues have been fixed. A table of descriptive statistics using the skim function followed by the summary function have been provided below.

```
> skim(data_new)
── Data Summary ────────────────────
                           Values
Name                       data_new
Number of rows             2883
Number of columns          16
_____________________
Column type frequency:
  character                5
  factor                   3
  numeric                  8
_____________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────
  skim_variable  n_missing complete_rate min max empty n_unique whitespace
1 Neighborhood           0             1   3   7     0       28          0
2 Bldg.Type              0             1   4   6     0        5          0
3 Central.Air            0             1   1   1     0        2          0
4 Heating.QC             0             1   2   2     0        5          0
5 Sale.Condition         0             1   6   7     0        6          0

── Variable type: factor ─────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 MS.Zoning             0             1 FALSE          7 RL: 2235, RM: 454, FV: 139, RH: 26
2 Overall.Qual          0             1 FALSE         10 5: 815, 6: 722, 7: 596, 8: 344
3 TotRms.AbvGrd         0             1 FALSE          8 6: 834, 7: 638, 5: 577, 8: 341

── Variable type: numeric ────────────────────────────────────────────────────────
  skim_variable  n_missing complete_rate      mean        sd    p0     p25    p50     p75    p100 hist
1 Lot.Area               0             1  10805.    40489.     1300   7440.   9434  11514.  215245 0
2 Gr.Liv.Area            0             1   1495.     496.      334   1126    1441   1741     5642 0
3 Bathrooms              0             1      2.21     0.803     1      1.5     2      2.5      7 0
4 Bedroom.AbvGr          0             1      3.85     0.822     1      3       4      4        7 0
5 Garage.Area            0             1    473.      214.       0    320     480    576     1488 0
6 Year.Remod.Add         0             1   1984.       20.9   1950   1965    1993   2004     2010 0
7 Mo.Sold                0             1      6.21     2.71      1      4       6      8       12 0
8 Sale.Price             0             1 215525.    91436.    15347. 155340  192000 255830.  699520. 0
```

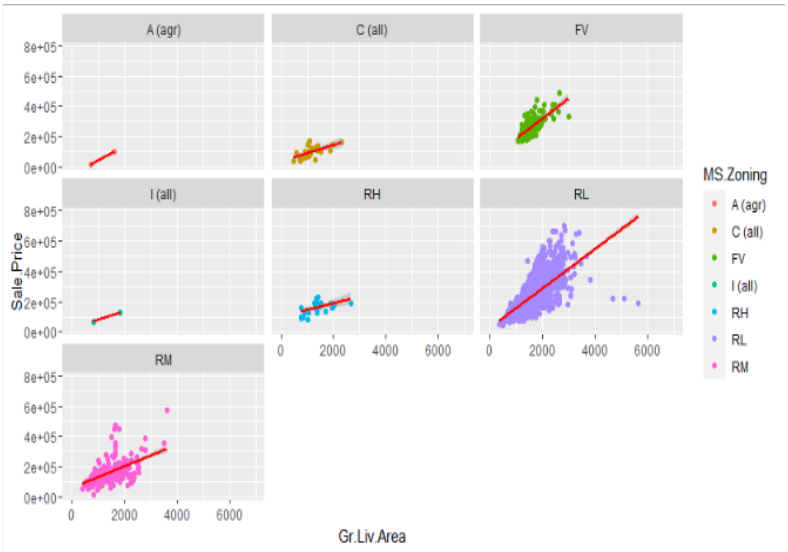
```
> summary(data_new)
   MS.Zoning     Neighborhood        Bldg.Type          Lot.Area       Gr.Liv.Area    Overall.Qual Central.Air        Heating.QC         Bathrooms
 A (agr):   2   Length:2883        Length:2883        Min.   :   1300   Min.   : 334   5      :815   Length:2883   Length:2883   Min.   :1.000
 C (all):  25   Class :character   Class :character   1st Qu.:   7440   1st Qu.:1126   6      :722   Class :character   Class :character   1st Qu.:2.000
 FV     : 139   Mode  :character   Mode  :character   Median :   9434   Median :1441   7      :596   Mode  :character   Mode  :character   Median :2.000
 I (all):   2                                         Mean   :  10805   Mean   :1495   8      :344                                            Mean   :2.212
 RH     :  26                                         3rd Qu.:  11514   3rd Qu.:1741   4      :220                                            3rd Qu.:2.000
 RL     :2235                                         Max.   :2152450   Max.   :5642   9      :103                                            Max.   :7.000
 RM     : 454                                                                          (Other): 83
 Bedroom.AbvGr   TotRms.AbvGrd  Garage.Area    Year.Remod.Add Sale.Condition       Mo.Sold        Sale.Price
 Min.   :1.000   6      :834   Min.   :   0   Min.   :1950   Length:2883        Min.   : 1.000   Min.   : 15347
 1st Qu.:3.000   7      :638   1st Qu.: 320   1st Qu.:1965   Class :character   1st Qu.: 4.000   1st Qu.:155340
 Median :4.000   5      :577   Median : 480   Median :1993   Mode  :character   Median : 6.000   Median :192000
 Mean   :3.852   8      :341   Mean   : 473   Mean   :1984                      Mean   : 6.214   Mean   :215525
 3rd Qu.:4.000   9      :265   3rd Qu.: 576   3rd Qu.:2004                      3rd Qu.: 8.000   3rd Qu.:255830
 Max.   :7.000   4      :203   Max.   :1488   Max.   :2010                      Max.   :12.000   Max.   :699520
                 (Other):  25
```

The results of the bivariate analysis of the hypotheses are categorically discussed below.

H_1 – The p-value was found to be **< 2.2e-16** which disproves the null hypotheses that there is no relationship between living area and sale price of the house.

(Correlation: **0.696**), Almost 70% of the variation in sale price can be explained by living area OR 1 unit increase in the living area will lead to an approximated 0.7 unit rise in the sale price.

```
> cor.test(data$Gr.Liv.Area,data$Sale.Price,method="pearson")

        Pearson's product-moment correlation

data:  data$Gr.Liv.Area and data$Sale.Price
t = 52.083, df = 2881, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6770943 0.7147226
sample estimates:
      cor
0.6963867
```



Also, if we facet by zoning, the strongest relationship between living area and Sale price is for:

- Floating Villages (**0.72**) and
- Residential Low density (**0.71**)

```
> d1 <- data %>%  filter(MS.Zoning == "RL")
> cor(d1$Gr.Liv.Area,d1$Sale.Price)
[1] 0.7126741
> d2<- data %>% filter(MS.Zoning == "FV")
> cor(d2$Gr.Liv.Area,d2$Sale.Price)
[1] 0.7174192
```

Thus, the alternate hypothesis is true with a strong positive relationship between living area and sale price.

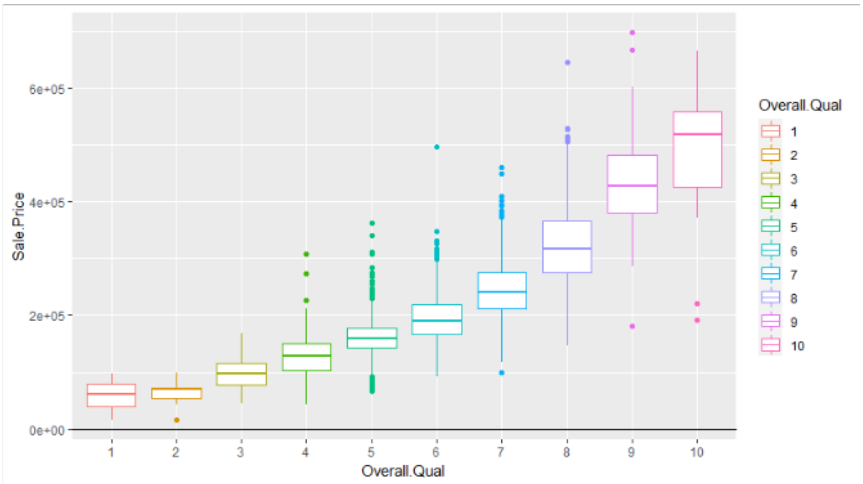
H_2 -- The p-value was found to be **< 2.2e-16** which disproves the null hypotheses that there is no relationship between living area and sale price of the house.

(Correlation: **0.806**), over 80% of the variation in sale price can be explained by the overall quality of the house OR 1 unit increase in the quality of house will lead to a 0.806 unit rise in the sale price.

```
> cor.test(as.numeric(data$Overall.Qual),data$Sale.Price,method = "spearman",exact = F)

        Spearman's rank correlation rho

data:  as.numeric(data$Overall.Qual) and data$Sale.Price
S = 774976959, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8059533
```

The trend suggests that with the increase in quality of the house, the median sale price also increases.

Thus, the alternate hypothesis is true with a very strong positive relationship between overall quality of the house and sale price.



H3 – The total number of rooms have been restricted to 9, in line with the analysis of Occupied Housing Units conducted by the US Census Bureau (2022) on the city of Ames, Iowa.

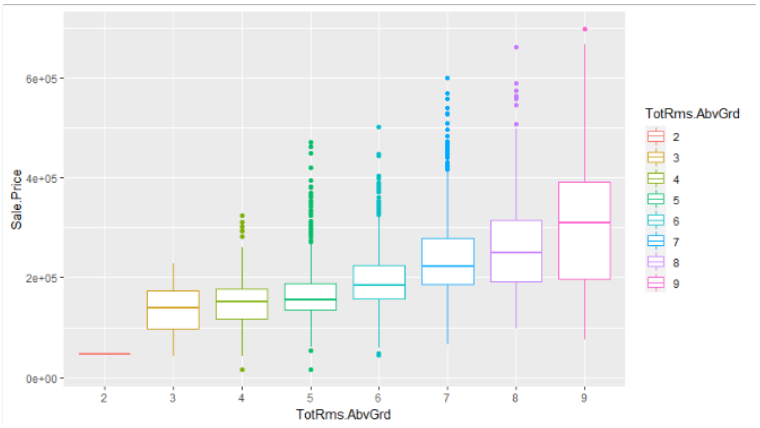
The p-value was found to be **< 2.2e-16** which disproves the null hypothesis that there is no relationship between total number of rooms and sale price of the house.

```
> cor.test(data$Sale.Price,data$TotRms.AbvGrd,method = "spearman",exact = F)

        Spearman's rank correlation rho

data:  data$Sale.Price and data$TotRms.AbvGrd
S = 2.014e+09, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4957052
```

(Correlation: **0.495**), almost 50% of the variation in sale price can be explained by the total number of rooms in the house OR 1 unit increase in the quality of house will lead to a 0.5 unit rise in the sale price.

The plot suggests an increasing trend between number of rooms and sale price.

Thus, the alternate hypothesis is true with a strong relationship between sale price and total number of rooms in the house.



H4 -- The p-value was found to be **< 2.2e-16** which disproves the null hypotheses that there is no difference in mean sale price between houses with and without central air conditioning.

(T-Test Statistic : -27.318) A significantly large t-test statistic shows that there is a significant difference in means between the two groups.

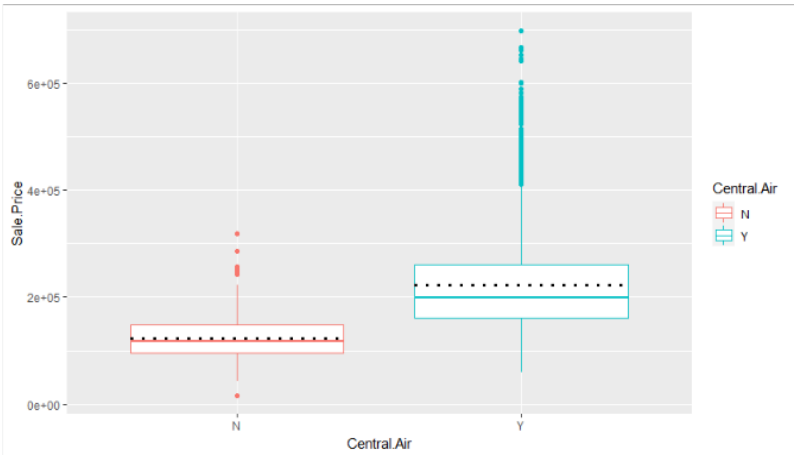
```
> t.test(Sale.Price~Central.Air,data=data)

        Welch Two Sample t-test

data:  Sale.Price by Central.Air
t = -27.318, df = 323.4, p-value < 2.2e-16
alternative hypothesis: true difference in means between group N and group Y is not equal to 0
95 percent confidence interval:
 -107264.27  -92852.85
sample estimates:
mean in group N mean in group Y
       122268.6        222327.1
```

Mean (With Central AC): $222327.1

Mean (Without Central AC): $122268.6

Thus, the alternative hypothesis is true

with a significant difference in mean sale price across the groups.

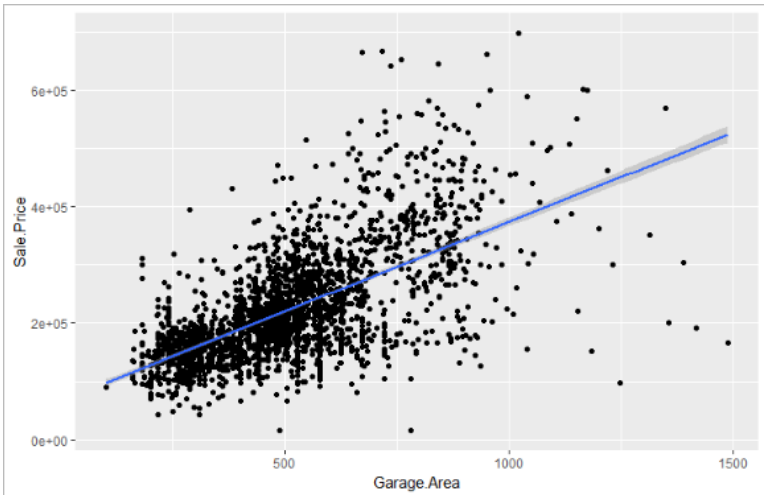


H5 -- The p-value was found to be **< 2.2e-16** which disproves the null hypotheses that there is no relationship between garage area and sale price of the house.

The analysis has excluded houses with no garages.

(Correlation: **0.636**), almost 64% of the variation in the house sale price can be explained by garage area.

```
> cor.test(data_new$Garage.Area[data_new$Garage.Area >0],data_new$Sale.Price)

        Pearson's product-moment correlation

data:  data_new$Garage.Area[data_new$Garage.Area > 0] and data_new$Sale.Price
t = 43.112, df = 2730, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6135817 0.6582292
sample estimates:
      cor
0.6364383
```



Thus, there is a strong positive relationship between garage area and sale price of the house.

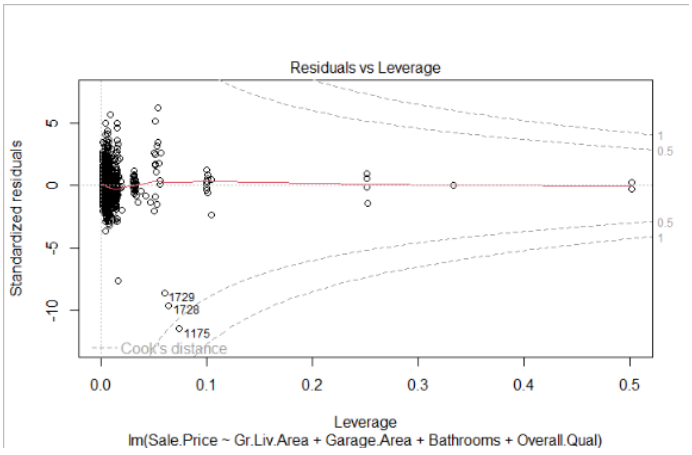
Regression modelling:

A tabulated result of the four models developed have been provided in Appendix 2 and the final model that has been selected for the analysis is discussed below.

model2<- lm(Sale.Price ~ Gr.Liv.Area + Garage.Area + Bathrooms + Heating.QC + Overall.Qual, data = train)

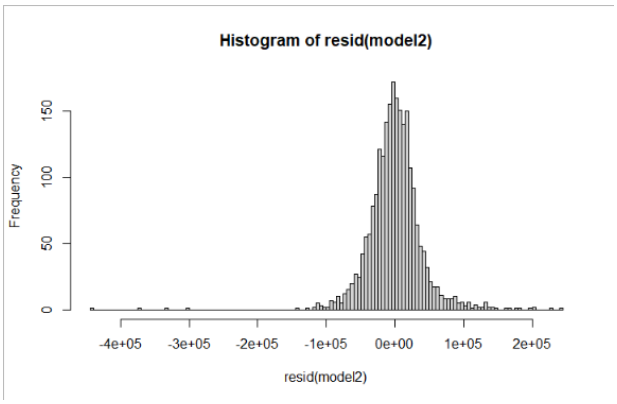
The final model is the best fit for its precise predictive accuracy on the sample dataset. 83.6% of variance in the sale price can be explained by this model. It has been confirmed that this model has no multicollinearity with an average VIF of 2.4.

This model also has no undue influencer as is clear from the graph showing the cook's distance as there is no element with cook's distance >1 (Field, 2012) .



There is also no homoscedasticity within the residuals and this model leads to an almost normally distributed residual plot.

hist(resid(model2))



Upon testing the predictive accuracy of this model, an R-squared of 0.849 was obtained, which can, to a good extent establish the reliability of this model. The root mean squared of this model was also calculated to be $36761, which can be considered for adjustment.

CONCLUSION

In conclusion, it is evident that living area and garage area bear a direct proportionality relation with the sale price of houses. In addition to that, the better the quality of a house gets, the sale price of the house rises steadily. It has also been noticed that residential zones of properties have a substantially large customer base than agricultural or industrial zones. This is rather understandable as generally people seem to avoid industrial or agricultural areas for residence, to remain closer to more socially active neighbourhoods with better proximity to medical and educational facilities. The model that has been designed can be used by the business to predict the house prices with relatively efficient accuracy.

From the summary statistics of the model, the coefficients can be used to determine the predicted sale price.

For instance, for a house with living area 1629, garage area 482, 2 bathrooms, good heating quality and an overall quality of 5, the sale price of the house can be predicted by:

- **nd <- data.frame(Gr.Liv.Area=c(1629),Bathrooms = c("2"),Garage.Area=c(482),Heating.QC=c("Gd"),Overall.Qual=c("5"))**
- **predict(model2, newdata = nd)**

The sale price is predicted as $188207 with an error margin of $36761.

The caveat with this model is that the residual error is quite high which means the predicted value of the house price will be always less than the actual by the error margin. Also, the p-value obtained in the Durbin-Wanger test is <0.05, which indicates that there exists a substantial positive autocorrelation between the successive residuals of this model. It is important to note that all linear models that have been tested (See Appendix), has returned with a p-value of < 0.05 for the Durbin-Wanger test. The autocorrelation can be adjusted in this model to improve the model performance.

REFLECTIVE SUMMARY

I had a very fundamental understanding of statistics before I started with this module. So far, I feel I have very steadily progressed with the concepts of bivariate analysis and linear regression. I have assessed that I need to work more on model improvement when assumptions are violated, as my understanding of error adjustment is still in a nascent stage and needs more refining. I am devoting more time to go through the concepts in detail from Descriptive Statistics using R by Andy Field and Statistics for Business and Economics by Cortinhas and Field book . I am really looking forward to learning how non-linear relationships between variables are handled using statistical tools. I am keen on advancing my career to the banking and finance industry as a risk analyst and wish to apply the different regression techniques for statistical forecasting of strategies.

REFERENCES

- Adair A., Berry J., McGreal W (1996). Hedonic modeling, housing submarkets and residential valuation. *Journal of Property Research*, 13 (1), pp. 67-83

- Azbainis, Vytautas & Rudzkienė, Vitalija. (2011). Pereinamojo laikotarpio ir ekonomikos krizės poveikio nekilnojamojo turto rinkai vertinimas. *Verslas: teorija ir praktika*. 12. 150-161.

- Bin O. (2004).A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics,* 13, pp. 68-84

- Deloitte, Massachusetts Institute of Technology (2020), Iowa, https://datausa.io/profile/geo/iowa , Available at: (https://datausa.io/profile/geo/iowa)

- Dietz, R. D., Haurin D. R. (2003). The social and private micro-level consequences of homeownership. *Journal of Urban Economics*, v. 54, n. 3, p. 401–450.

- Fan, G.Z., Ong, S.E. and Koh, H.C., 2006. Determinants of house price: A decision tree approach. *Urban Studies*, *43*(12), pp.2301-2315.

- Field, A., Miles, J. and Field, Z., (2012). Discovering statistics using R. *Great Britain: Sage Publications, Ltd*, *958*.

- Jureviciene D., Okuneviciute, Neverauskiene L. (2008). Būsto įsigijimo sąlygų įtaka jaunimui įsitvirtinti nacionalinėje darbo rinkoje. *Business: Theory & Practice*, v. 9, n. 2, p. 116–125.

- Larsen, J.E., (2010). The impact of buyer-type on house price: Some evidence from the USA. *International Journal of Housing Markets and Analysis*.

- Lin, T.C. and Evans, A.W. (2000). The relationship between the price of land and size of plot when plots are small. *Land economics*, pp.386-394.

- Manasa, J., Gupta, R. and Narahari, N.S. (2020). Machine learning based predicting house prices using regression techniques, *2nd International conference on innovative mechanisms for industry applications (ICIMIA) IEEE* , pp. 624-630.

- Masri, M. H. M., Nawawi, A. H., Mohd Safian, E. E., & Ahmad Saleh, A. F. (2018). House Qualities Characteristics Relationship on House Prices: Klang District. *Environment-Behaviour Proceedings Journal*, *3*(7), 295-305.

- Mearns L.O., Arritt R., Biner S., Bukovsky M.S., McGinnis S., Sain S., Caya D., Correia J., Flory D., Gutowski W. and Takle E.S. (2012). The North American regional climate change assessment program: overview of phase I results. *Bulletin of the American Meteorological Society*, *93*(9), pp.1337-1362.

- Meese R., Wallace N. (2003).House price dynamics and market fundamentals: The Parisian housing market*. Urban Studies*, vol 40, pp. 1027-1045

- Nissan Pow, Emil Janulewicz and L. Liu (2014), Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal, *McGill University*.

- Park B. and Bae, J.K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert systems with applications*, *42*(6), pp.2928-2934.

- Stevenson S. (2004). New empirical evidence on heteroscedasticity in hedonic housing models*. Journal of Housing Economics,* 13, pp. 136-153

- Thorsnes, P. and McMillen, D.P. (1998). Land value and parcel size: a semiparametric analysis. *The Journal of Real Estate Finance and Economics*, *17*(3), pp.233-244.

- US Census Bureau (2022), American Community Survey, https://www.census.gov/programs-surveys/acs/ . Available at: (https://www.iowadatacenter.org/index.php/data-by-source/american-community-survey/number-rooms-tenure)

APPENDIX 1

Model: Results

	Dependent variable:			
	Sale.Price			
Model no. :	(1)	(2)	(3)	(4)
Overall.Qual2	24,580.140	6,973.608	13,269.230	16,980.890
	(21,840.490)	(24,127.110)	(23,781.490)	(22,900.240)
Overall.Qual3	37,507.200*	17,293.570	24,490.440	25,603.760
	(19,521.860)	(21,724.810)	(21,295.660)	(20,530.580)
Overall.Qual4	51,775.860***	33,835.540	43,534.490**	38,958.800**
	(18,690.300)	(20,832.050)	(20,355.910)	(19,710.280)
Overall.Qual5	67,286.080***	55,664.460***	65,993.990***	57,766.530***
	(18,545.890)	(20,711.920)	(20,189.970)	(19,584.380)
Overall.Qual6	80,235.250***	69,175.290***	81,621.890***	72,563.480***
	(18,588.230)	(20,763.660)	(20,231.910)	(19,634.590)
Overall.Qual7	100,218.400***	93,278.380***	111,113.500***	101,093.100***
	(18,693.450)	(20,852.670)	(20,301.330)	(19,703.110)
Overall.Qual8	141,148.900***	147,082.800***	166,680.200***	158,437.500***
	(18,903.650)	(21,002.170)	(20,445.710)	(19,846.860)
Overall.Qual9	220,611.800***	236,154.700***	256,152.000***	248,318.500***
	(19,435.770)	(21,432.740)	(20,887.600)	(20,274.370)
Overall.Qual10	217,870.600***	216,315.800***	236,348.700***	228,705.000***
	(20,846.360)	(22,975.160)	(22,548.180)	(21,862.770)
Gr.Liv.Area	45.683***	45.134***	44.548***	43.656***
	(2.249)	(2.235)	(2.276)	(2.298)
Garage.Area	50.100***	64.713***	66.368***	53.722***
	(4.911)	(4.978)	(5.060)	(5.003)
Mo.Sold2			-2,484.263	-2,907.189

			(5,632.084)	(5,428.051)
Mo.Sold3			-408.185	-1,422.657
			(4,951.440)	(4,774.974)
Mo.Sold4			1,833.156	1,753.392
			(4,793.890)	(4,621.944)
Mo.Sold5			2,728.870	2,615.027
			(4,585.946)	(4,423.279)
Mo.Sold6			5,185.091	4,411.005
			(4,493.594)	(4,332.145)
Mo.Sold7			7,599.977*	6,939.990
			(4,534.236)	(4,371.087)
Mo.Sold8			110.326	-240.066
			(4,921.081)	(4,742.175)
Mo.Sold9			1,143.541	1,314.730
			(5,413.737)	(5,219.158)
Mo.Sold10			149.325	-886.343
			(5,289.026)	(5,101.028)
Mo.Sold11			1,225.863	2,034.662
			(5,499.522)	(5,298.606)
Mo.Sold12			-1,615.050	-837.682
			(5,978.713)	(5,766.908)
Bathrooms2	7,812.199***	12,641.000***	13,116.870***	18,160.860***
	(2,529.553)	(2,569.485)	(2,617.742)	(2,613.691)
Bathrooms3	24,180.350***	34,273.130***	35,746.480***	42,327.650***
	(3,446.611)	(3,513.988)	(3,577.653)	(3,560.808)
Bathrooms4	36,313.250***	46,437.650***	49,346.370***	55,752.230***
	(3,869.787)	(3,962.048)	(4,022.978)	(3,963.819)
Bathrooms5	10,156.610	-6,020.292	-818.744	26,917.610
	(26,343.510)	(28,084.100)	(28,594.050)	(27,909.810)
Bathrooms6	93,517.670**	98,719.640**	96,134.500**	120,423.300***
	(36,975.770)	(39,558.860)	(40,297.690)	(39,080.820)
Bathrooms7	36,480.590*	26,908.980	33,938.680	71,826.760***
	(22,017.740)	(23,237.740)	(23,864.470)	(23,470.850)
NeighborhoodBlueste	-42,638.920**			
	(17,247.280)			
NeighborhoodBrDale	-51,412.760***			
	(11,381.440)			
NeighborhoodBrkSide	-23,111.820**			

	(9,491.653)
NeighborhoodClearCr	17,844.280*
	(10,287.530)
NeighborhoodCollgCr	1,009.268
	(8,665.266)
NeighborhoodCrawfor	19,113.000**
	(9,348.406)
NeighborhoodEdwards	-30,219.130***
	(9,061.873)
NeighborhoodGilbert	-5,375.948
	(9,026.352)
NeighborhoodGreens	-17,836.360
	(17,355.680)
NeighborhoodGrnHill	63,782.910*
	(37,685.710)
NeighborhoodIDOTRR	-39,866.530***
	(9,721.311)
NeighborhoodLandmrk	-37,593.220
	(37,754.200)
NeighborhoodMeadowV	-38,345.980***
	(11,767.010)
Neighborhoodmes	-15,769.440*
	(8,751.187)
NeighborhoodMitchel	-7,189.210
	(9,364.115)
NeighborhoodNoRidge	36,210.710***
	(9,993.289)
NeighborhoodNPkVill	-33,782.320***
	(12,251.610)
NeighborhoodNridgHt	38,920.440***
	(9,248.407)
NeighborhoodNWAmes	-10,783.040
	(9,128.754)
NeighborhoodOldTown	-42,264.470***
	(8,964.294)
NeighborhoodSawyer	-15,774.840*
	(9,207.218)
NeighborhoodSawyerW	-11,095.920

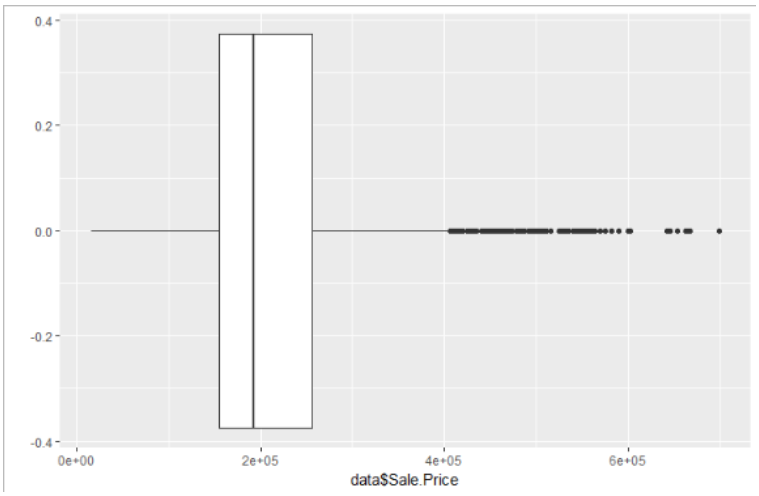
	(9,168.810)			
NeighborhoodSomerst	1,878.302			
	(8,834.666)			
NeighborhoodStoneBr	35,702.690***			
	(10,600.230)			
NeighborhoodSWISU	-39,309.640***			
	(10,307.120)			
NeighborhoodTimber	14,536.350			
	(9,779.718)			
NeighborhoodVeenker	15,217.700			
	(11,939.310)			
Heating.QCFa		-25,283.010***		
		(5,108.310)		
Heating.QCGd		-9,171.773***		
		(2,499.699)		
Heating.QCPo		-46,440.460*		
		(23,888.440)		
Heating.QCTA		-18,103.810***		
		(2,159.538)		
Central.AirY				21,089.560***
				(3,694.457)
as.factor(Bldg.Type)2fmCon				-23,810.940***
				(5,741.226)
as.factor(Bldg.Type)Duplex				-30,848.890***
				(4,547.075)
as.factor(Bldg.Type)Twnhs				-36,801.800***
				(4,645.389)
as.factor(Bldg.Type)TwnhsE				-15,595.610***
				(3,208.469)
Constant	29,395.290	24,358.730	-377.672	-4,594.605
	(20,455.030)	(20,834.260)	(20,427.930)	(19,687.540)
Observations	2,308	2,308	2,308	2,308
R^2	0.839	0.813	0.807	0.822
Adjusted R^2	0.836	0.811	0.805	0.819
Residual Std. Error	36,761.310 (df = 2263)	39,435.660 (df = 2286)	40,089.490 (df = 2279)	38,600.300 (df = 2274)
F Statistic	268.000*** (df = 44; 2263)	472.731*** (df = 21; 2286)	340.686*** (df = 28; 2279)	317.385*** (df = 33; 2274)

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

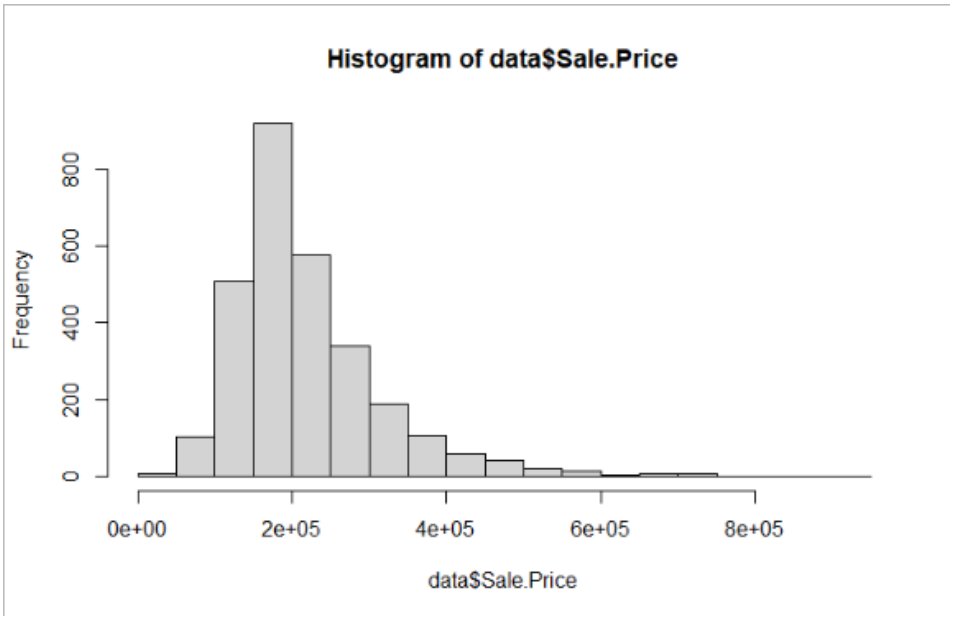
APPENDIX 2

```
setwd("C:/Users/Tathagata/OneDrive/Desktop/MGT_7177_ASgn1")
library(readxl)
data<-read_excel("ames.xlsx")    #reading in the data
library(tidyverse)
library(dplyr)
library(ggplot2)
library(skimr)
library(caret)

##exploratory analysis and fixing data quality issues

summary(data)
skim(data) #summarizing data using skim
View(data)

data<-data %>% mutate(Bathrooms = round(Bsmt.Full.Bath + Bsmt.Half.Bath * 0.5 +
Full.Bath + Half.Bath * 0.5,0)) #normalising the number of bathrooms for ease if calculation

    #Factorizing nominal and ordinal variables
data_new$Central.Air<-as.factor(data_new$Central.Air)
data$MS.Zoning<-as.factor(data$MS.Zoning)
data$Overall.Qual<-as.factor(data$Overall.Qual)
data$Mo.Sold <- as.factor(data$Mo.Sold)
data$Bathrooms <- as.factor(data$Bathrooms)
```

```
data$Heating.QC <- as.factor(data$Heating.QC)

data$TotRms.AbvGrd<-as.factor(data$TotRms.AbvGrd)



data<- data %>% filter(data$Sale.Price<700000)      #Removing outliers


data$Gr.Liv.Area[is.na(data$Gr.Liv.Area)] <- mean(data$Gr.Liv.Area,na.rm = T)     #Handling missing values

data$Garage.Area[is.na(data$Garage.Area)] <- mean(data$Garage.Area,na.rm=T)
#Replacing missing values with mean attribute


    #creating a subset of the dataset with selected variables of relevance (subjective).

data_new<- data %>%
select(MS.Zoning,Neighborhood,Bldg.Type,Lot.Area,Gr.Liv.Area,Overall.Qual,Central.Air,Heating.QC,Bathrooms,Bedroom.AbvGr,TotRms.AbvGrd,Garage.Area,Year.Remod.Add,Sale.Condition,Mo.Sold,Sale.Price)



data_new$Garage.Area[data_new$Garage.Area == 0] <- NA   #GarageArea 0 is interpreted as no garage

data_new <- data_new %>% filter(!is.na(data_new$Bathrooms))



## Exploratory Visualisation

qplot(data$Sale.Price,geom = "boxplot")  # identifying outliers in sale price
```



hist(data$Sale.Price) **#checking distribution of sale price**



#living area vs sale price trend by zones

ggplot(data=data_new, mapping = aes(x=Gr.Liv.Area, y=Sale.Price, color = MS.Zoning)) +

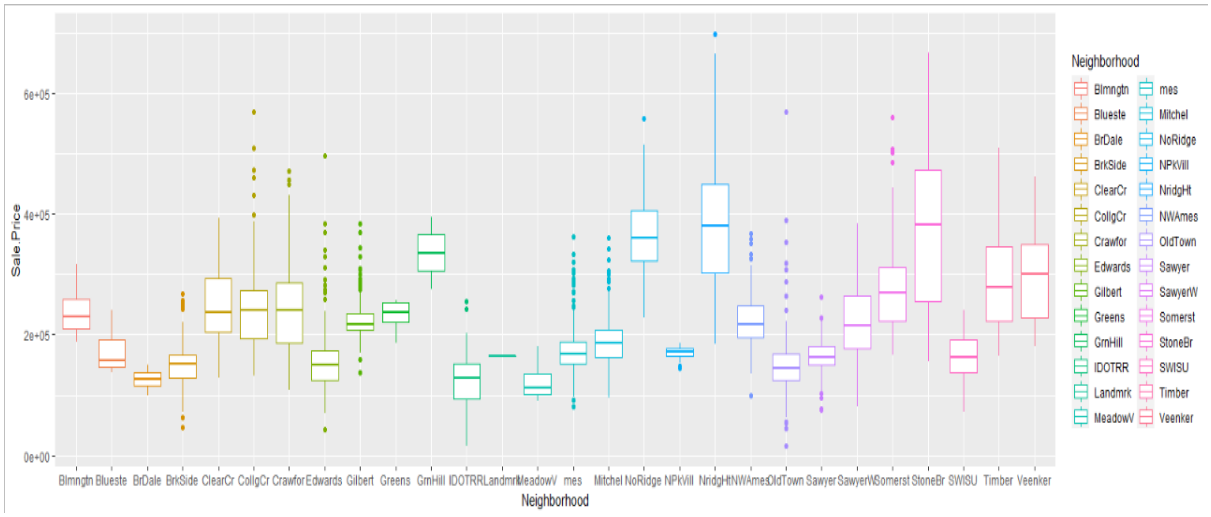
geom_point() + geom_smooth(method = lm,color="red") + facet_wrap(~MS.Zoning, nrow= 3)

#garage area vs sale price trend by zones

```
ggplot(data=data_new, mapping = aes(x=Garage.Area, y=Sale.Price, color = MS.Zoning)) +
  geom_point() + geom_smooth(method = lm,color="red") + facet_wrap(~MS.Zoning, nrow=
3)
```

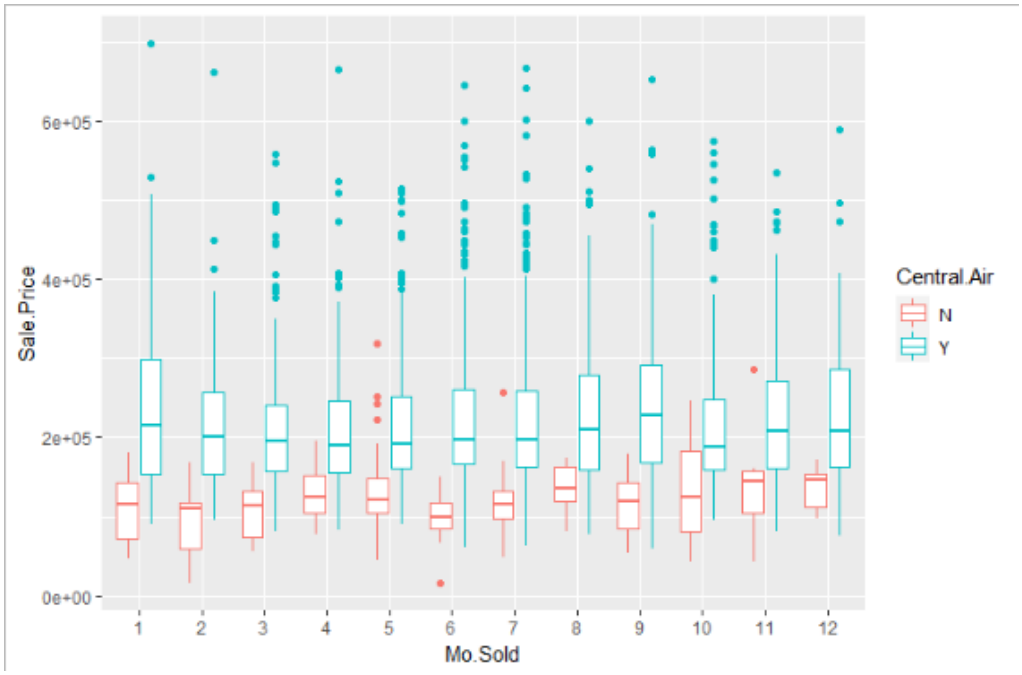
#sale price in respective neighborhood

```
ggplot(data=data_new, mapping = aes(x=Neighborhood, y=Sale.Price, color =
Neighborhood)) +
  geom_boxplot()
```



#Sale of houses each month with/without central air conditioning

```
ggplot(data=data_new, mapping = aes(x=Mo.Sold, y=Sale.Price, color = Central.Air)) +
  geom_boxplot()
```



#[DE] ##Summary Report of hypotheses variables using DataExplorer Package**

```
install.packages("DataExplorer")
```

```
library(DataExplorer)


dhyp<-
data_new[,c("Gr.Liv.Area","Garage.Area","Overall.Qual","Central.Air","TotRms.AbvGrd","Sal
e.Price")]

create_report(

  dhyp,

 output_file = "report.html",

 output_dir = getwd(),

 y = NULL,

 config = configure_report(),

 report_title = "Data Profiling Report"

)
```

#A detailed analysis summary report is generated as an html file in the current working directory of the R program. I intended to upload the static html file on Heroku and embed the link in my report for it to be accessible but since I am unaware of the IP policies of Queens, I refrained from doing so. Converting the html was also not an option as the image and text quality was being compromised.

#H1

```
ggplot(data_new,mapping=aes(x=Gr.Liv.Area,y=Sale.Price,color=MS.Zoning)) +
geom_point() + xlim(0,7000) + geom_smooth(method = lm,color="red") +
facet_wrap(~MS.Zoning,nrow=3)

cor.test(data_new$Gr.Liv.Area,data$Sale.Price,method="pearson")

d1 <- data_new %>%  filter(MS.Zoning == "RL")

cor(d1$Gr.Liv.Area,d1$Sale.Price)

d2<- data_new %>% filter(MS.Zoning == "FV")

cor(d2$Gr.Liv.Area,d2$Sale.Price)
```

#H2

```
plot(data_new$Overall.Qual,data$Sale.Price)


levels(data_new$Overall.Qual) #Contains house quality 11, No mention in data dictionary, delete records

data_new<- data_new[data_new$Overall.Qual != "11",]

table(data_new$Overall.Qual)

droplevels(data_new$Overall.Qual)

ggplot(data=data_new,mapping=aes(x=Overall.Qual,y=Sale.Price,color=Overall.Qual)) + geom_boxplot()

cor.test(as.numeric(data_new$Overall.Qual),data_new$Sale.Price,method = "spearman",exact = F)
```

#H3

```
class(data_new$TotRms.AbvGrd)

table(data_new$TotRms.AbvGrd)

data_new$TotRms.AbvGrd[data_new$TotRms.AbvGrd>9] <- 9

cor.test(data_new$Sale.Price,as.numeric(data_new$TotRms.AbvGrd),method = "spearman",exact = F)

ggplot(data = data_new, mapping = aes(x=TotRms.AbvGrd,y=Sale.Price, color = TotRms.AbvGrd)) + geom_boxplot()
```

#H4

```
summary(data_new$Central.Air)

table(data_new$Central.Air)

t.test(Sale.Price~Central.Air,data=data_new)

ggplot(data = data_new, mapping = aes(x=Central.Air,y=Sale.Price, color = Central.Air)) + geom_boxplot() + stat_summary(fun.y=mean,color="black",geom = "crossbar", fun = "mean", linetype = "dotted", width = .75)
```

```
cor.test(data_new$Garage.Area,data_new$Sale.Price)
```

#H5

```
summary(data_new$Garage.Area)

table(data_new$Garage.Area[data_new$Garage.Area==0])

data_new$Garage.Area[data_new$Garage.Area == 0] <- NA

ggplot(data=data_new, mapping=aes(x=Garage.Area,y=Sale.Price)) + geom_point() +
geom_smooth(method=lm)

cor.test(data_new$Garage.Area[data_new$Garage.Area >0],data_new$Sale.Price)
```

##Linear Regression analysis

```
set.seed(123456)

index <- createDataPartition(data_new$Sale.Price,p=0.8,list=F)

train <- data_new[index,]

test <- data_new[-index,]
```

#H1

```
fh1 <- Sale.Price ~ Gr.Liv.Area

modelh1 <- lm(fh1,data=train)

summary(modelh1)
```

#H2

```
OQ<-as.numeric(train$Overall.Qual)

fh2 <- Sale.Price ~ OQ

modelh2 <- lm(fh2,data=train)
```

```
summary(modelh2)
```

#H3

```
TR <- as.numeric(train$TotRms.AbvGrd)
fh3 <- Sale.Price ~ TR
modelh3 <- lm(fh3,data=train)
summary(modelh3)
```

#H4

```
CA <- as.numeric(train$Central.Air)
fh4 <- Sale.Price ~ CA
modelh4 <- lm(fh4,data = train)
summary(modelh4)
```

#H5

```
fh5 <- Sale.Price ~ Garage.Area
modelh5 <- lm(fh5, data = train)
summary(modelh5)
```

#MLR model test

#model1

```
model1 <- lm(Sale.Price ~ Overall.Qual + Gr.Liv.Area + Garage.Area + Bathrooms +
Neighborhood, data = train)
summary(model1)
```

```
install.packages("car")
```

```
library(car)



summary(model1)

vif(model1)

   #Overall.Qual and Neighboorhood has very high vif, needs to be dropped


##adjusting model

model2<- lm(Sale.Price ~ Gr.Liv.Area + Garage.Area + Bathrooms + Heating.QC +
Overall.Qual, data = train)


  #inferential testing. Testing Assumptions

summary(model2)

vif(model2)

plot(model2)

hist(resid(model2),breaks = 100)  #almost normally distributed residuals


install.packages("lmtest")

library(lmtest)

dwtest(model2)  #durban watson test value of 1.5938, within range


cook <- cooks.distance(model2)

sum(cook > 1)


mean(vif(model2))


#Checking Predictive accuracy

prediction <- predict(model2, newdata = test)

postResample(pred = prediction,obs = test$Sale.Price)
```

```
sqrt(mean((prediction - test$Sale.Price) ^ 2))   # $36761 sqrt mean squared error

head(resid(model2))

#eg prediction

View(test)

class(test$Bathrooms)

nd <- data.frame(Gr.Liv.Area=c(1629),Bathrooms =
c("2"),Garage.Area=c(482),Heating.QC=c("Gd"),Overall.Qual=c("5"))

predict(model2, newdata = nd)

##model3

model3 <- lm(Sale.Price ~ Overall.Qual + Lot.Area + Garage.Area + Bathrooms + Heating.QC,
data = train)

summary(model3)

dwtest(model3)

vif(model3)

plot(model3)

prediction <- predict(model3, newdata = test)

postResample(pred = prediction,obs = test$Sale.Price)

sqrt(mean((prediction - test$Sale.Price) ^ 2))   # $41000 sqrt mean squared error

head(resid(model2))

##model4
```

```
model4 <- lm(Sale.Price ~ Overall.Qual + Gr.Liv.Area + Garage.Area + Mo.Sold + Bathrooms + Central.Air + as.factor(Bldg.Type) , data = train)

summary(model4)

dwtest(model4)

vif(model4)


##Responses of all four models have been tabulated using the stargazer function

install.packages("stargazer")

library(stargazer)

stargazer(model1,model2,model3,model4,type="text",title="Model1: Results",align = T,out="model.html")
```