

Predicting the Sale Price of Houses in Ames using Machine Learning

Chukwuekwu Musei

1.0 Introduction and Background

In the history of development economics, housing has been thought of as a key factor in the development of any nation's economy and it is one of the basic needs of man. As such, the demand for residential houses have continued to be on the rise as house owners continue to evaluate the worth of their properties in order to place the right values on them. The purpose of this paper is to develop a model for predicting sale price of houses using multiple regression. then followed by the task procedure. It will then discuss results of findings and make recommendation.

A considerable amount of literature has been published on house prices prediction using different regression techniques. According to Du and Cross (2007) the grey model (GM) was used to predict house prices in the ancient days which depicts a rising and falling process that yielded poor metrics for predicting house prices. Varol et al. (2009) study show how the support vector machine (SVM) can be used to predict house prices because it helps to solve the issue of overfitting that reduces expected errors in analyses when a model tries to learn from data. Despite the merits of SVM over GM in predicting house prices, it has its limitations particularly as it relates to how the value of the training data is picked in the right way (Gu *et al.*, 2011), leading to the emergence of a better approach in which generic algorithm optimises SVG to yield G-SVM. Gu *et al.* (2011) claim that the G-SVM approach for house price prediction has performed better when applied to predict the national selling price of houses in China within a 10-year period.

Manasa *et al.* (2020) and Viktorovich *et al.* (2018) use the ordinary least square method in multiple regression to predict the sale price of houses alongside multiple independent variables claim to have effect on the sale price of houses. It can be claimed that this regression technique provides more certainty as the datasets are split into train and test, where the latter will be used to check the accuracy of the trained data.

However, the hypotheses of this paper will be built on certain features in houses believed to have a significant positive effect on sale price, such as, Lot Area, Bedrooms, Ground Living Area, Overall Quality and Basement Finishing that are in the ames dataset. The decision to choose these variables were based on Civil Engineering experience as well as the reviewed extant literatures. Five hypotheses will be tested in this task, and they are as follows:

H1 Lot.Area is positively related to Sale Price

H2 Bedroom is positively related to Sale Price

H3 Ground living area is positively related to Sale Price

H4 Overall quality is positively related to Sale Price

H5 BsmtFin.Type.1 is positively related to Sale Price

2.0 Methodology

The CRISP-DM framework for data mining was used to carry out the data analyses of this task. The ames dataset was imported into R studio after the working directory had been set. This was followed by assigning the ames dataset to the object *musei* in preparation for the exploratory data analyses (EDA). After importing and assigning the ames dataset, a summary statistics was taken. It was observed that a total of 2893 observations and 80 variables were in the ames dataset. During the EDA, such as looking out for data quality issues, checking the minimum and maximum values for possible outliers as well as missing values, it was observed that 11 variables have a combined total of 466 missing data out of the 2893 observations in the dataset.

The histogram and boxplot in ggplot2 were used to visualise the variables to ascertain where to subset the outliers in order not to compromise the quality of the analyses (see Appendix 3). In the end, the variables with data quality issues were either fixed by removing errors coding them as NAs and replacing them by the mean value. These were done because the removed observations were not significant enough to affect the analyses.

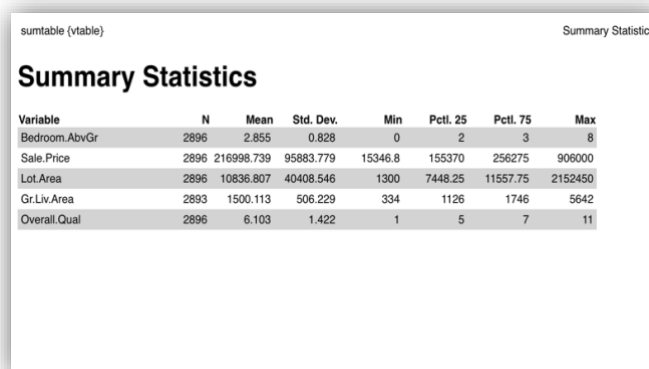
In addition, the multiple regression technique was used to predict the sale price of houses in ames. First, the data was split into two, train and test. The train data consist of 80% of the data while the test data that will be used to check the accuracy of the trained data contained 20% of the data. Also, assumption checks such as homoscedasticity, cook distance, Durban-Waton and multicollinearity were done on the multiple regression model.

3.0 Results and Discussions

3.1.1 Descriptive Statistics

The descriptive statistics in tables 1a and 1b show the ames dataset summaries from the selected numeric variables that formed part of the hypotheses before and after fixing data quality issues. It was observed that the sale price, which is the target variable recorded a sharp increase in its maximum value. After removing the outlier, the maximum value now reduced slightly. Similarly, there was an error in *Overall.Qual* levels contrary to the description in the data dictionary. These errors were fixed appropriately and documented.

Table 1a: Initial summary statistics of selected numeric variables before fixing data quality issues



The image shows a screenshot of the 'Summary Statistics' window in RStudio. The window title is 'sumtable (vtable)' and 'Summary Statistics'. The table displays the following data:

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Bedroom.AbvGr	2896	2.855	0.828	0	2	3	8
Sale.Price	2896	216998.739	95883.779	15346.8	155370	256275	906000
Lot.Area	2896	10836.807	40408.546	1300	7448.25	11557.75	2152450
Gr.Liv.Area	2893	1500.113	506.229	334	1126	1746	5642
Overall.Qual	2896	6.103	1.422	1	5	7	11

Table 1b: Final descriptive statistics of selected numeric variables after fixing data quality issues

sumtable {vtable}

Summary Statistics

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Bedroom.AbvGr	2892	2.855	0.829	0	2	3	8
Sale.Price	2892	216596.128	94224.774	15346.8	155400	256200	750000
Lot.Area	2892	10097.112	6932.586	1300	7448.25	11556	164660
Gr.Liv.Area	2892	1500.513	506.196	334	1126.75	1746.75	5642
Overall.Qual	2892	6.096	1.411	1	5	7	10

3.1.2 Table 2a: Showing final regression model.

```
> summary(model6)

Call:
lm(formula = formula, data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-595398 -18708      42    17439   245269

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.468e+04  1.832e+04   0.801  0.4230
Lot.Area      1.205e+00  1.293e-01   9.319 < 2e-16 ***
Gr.Liv.Area   6.420e+01  2.832e+00  22.669 < 2e-16 ***
Overall.Qual  2.189e+04  1.100e+03  19.897 < 2e-16 ***
Bedroom.AbvGr -5.908e+03  1.351e+03  -4.375 1.27e-05 ***
BsmtFin.Type.1Unf -2.249e+04  9.404e+03  -2.391 0.0169 *
BsmtFin.Type.1LwQ -1.543e+04  9.988e+03  -1.545 0.1225
BsmtFin.Type.1Rec -1.311e+04  9.697e+03  -1.352 0.1766
BsmtFin.Type.1BLQ -7.972e+03  9.722e+03  -0.820 0.4123
BsmtFin.Type.1ALQ -3.130e+03  9.582e+03  -0.327 0.7440
BsmtFin.Type.1GLQ 1.511e+03  9.604e+03  0.157 0.8750
FoundationCBlock 7.379e+03  3.140e+03  2.350 0.0188 *
FoundationPConc 2.162e+04  3.460e+03  6.248 4.94e-10 ***
FoundationSlab 1.186e+04  1.149e+04  1.032 0.3020
FoundationStone -4.863e+03  1.411e+04  -0.345 0.7303
FoundationWood -2.737e+03  1.878e+04  -0.146 0.8841
Condition.1Feedr 1.191e+04  5.851e+03  2.035 0.0420 *
Condition.1Norm 2.044e+04  4.834e+03  4.229 2.44e-05 ***
Condition.1PosA 4.721e+04  1.211e+04  3.897 0.0001 ***
Condition.1PosN 1.628e+04  9.157e+03  1.778 0.0755 .
Condition.1RR Ae 4.233e+03  9.601e+03  0.441 0.6593
Condition.1RRAn -1.767e+04  8.405e+03  -2.103 0.0356 *
Condition.1RRNe -1.752e+03  2.166e+04  -0.081 0.9355
Condition.1RRNn 5.084e+03  1.636e+04  0.311 0.7560
Total.Bsmt.SF 3.492e+01  2.873e+00  12.151 < 2e-16 ***
Exter.QualFa -9.048e+04  1.168e+04  -7.744 1.44e-14 ***
Exter.QualGd -7.423e+04  5.094e+03  -14.572 < 2e-16 ***
Exter.QualGood -8.806e+04  1.754e+04  -5.020 5.57e-07 ***
Exter.QualTA -9.024e+04  5.973e+03  -15.108 < 2e-16 ***
Exter.CondFa -1.874e+04  1.473e+04  -1.273 0.2033
Exter.CondGd 7.375e+03  1.349e+04  0.547 0.5847
Exter.CondGood -1.128e+02  2.081e+04  -0.005 0.9957
Exter.CondPo -2.133e+04  2.770e+04  -0.770 0.4413
Exter.CondTA -1.636e+03  1.327e+04  -0.123 0.9019
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41310 on 2281 degrees of freedom
Multiple R-squared:  0.8179,    Adjusted R-squared:  0.8153
F-statistic: 310.5 on 33 and 2281 DF,  p-value: < 2.2e-16
```

In tab. 2a above, it can be seen that the Multiple R Squared accounts for 81.79% of the *Sale.Price* variation, indicating a good model.

The first model consisted of the 5 variables in the hypotheses, after which, individual variable were added to output the second to the seventh model (see Appendix 1). It was observed that at the 7th model showed no significant increase in the Adjusted R-Squared, thus, the reason for dropping the model and using model 6 as the final model.

In addition, the F-statistics value of the model is significant because a p-value < 0.05 is outputted. This means that *Sale.Price* as the target variable has a significant relationship with all the 10 independent variables in model 6 (see tab. 2a). The estimates of 5 variables are explained below:

Lot.Area: The b-value for the *Lot.Area* also known as the slope of the regression is given as 1. 205. This indicates that for one square foot increase in *Lot.Area* will result in an increase in *Sale.Price* by \$1.205. It can be seen that the relationship between *Sale.Price* and *Lot.Shape* are significant at 0.001 level, having in mind that a p-value < 0.05 is significant. Therefore it can be concluded there is a relationship between *Sale.Price* and *Lot.Area*, though not a very strong one.

Gr.Liv.Area: The b-value = 64.20 for the *Gr.Liv.Area*. This indicates that for one square foot increase in *Gr.Liv.Area* will result in an increase in *Sale.Price* by \$64.20. It can be seen that the relationship between *Sale.Price* and *Gr.Liv.Area* are significant at 0.001 level, having in mind that a p-value < 0.05 is significant. Therefore it can be concluded there is a relationship between *Sale.Price* and *Gr.Liv.Area*.

Overall.Qual: The b-value =21890 for the *Overall.Qual*. This indicates that for one square foot increase in *Overall.Qual* will result in an increase in *Sale.Price* by \$21,890. It can be seen that the relationship between *Sale.Price* and *Overall.Qual* are significant at 0.001 level, having in mind that a p-value < 0.05 is significant. Therefore it can be concluded there is a strong positive relationship between *Sale.Price* and *Overall.Qual*.

Foundation: The *Foundation* can be interpreted as 5 binary variables. Compared to a *FoundationCBlock*, houses with *FoundationPConc* are more expensive. This is significant at 0.001 level. Therefore it can be predicted that a house with *FoundationPConc* is worth expensive by an extra of \$21,620 compared to that of *FoundationCBlock* valued at \$7379 with 0.05 significant level. It is also crucial to note that independent variables that are not significant should also be used for prediction, but it cannot be used to conclude there is a genuine relationship in the population. For instance, houses with *FoundationSlab* are valued at \$11,860 more when prediction is made, but it cannot be used to generalise on the wider population.

Total.Bsmt.SF: The b-value =3492 for the *Total.Bsmt.SF*. This indicates that for one square foot increase in *Total.Bsmt.SF*. will result in an increase in *Sale.Price* by \$3492. It can be seen that the relationship between *Sale.Price* and *Total.Bsmt.SF*. are significant at 0.001 level, having in mind that a p-value < 0.05 is significant. Therefore it can be concluded there is a genuine relationship between *Sale.Price* and *Total.Bsmt.SF*.

However, the measures of accuracy of the multiple regression (model6) is interpreted below:

Table 2b: Measures of accuracy

```
> ### check model6 accuracy of the test data (20%)
> predictions <- predict(model6, test)
> postResample(predictions, test$Sale.Price)
          RMSE          Rsquared          MAE
4.364751e+04 7.582546e-01 2.409724e+04
```

The average error in the prediction of *Sale.Price* of the house is at \$43,647, which can be seen in the RMSE value in tab. 2b. This value is a bit on the high side which will not be appreciated by the owner of the property to loose such amount due to error. Comparing the model accuracy for each model, it can be observed there was improvement on the average error i.e. RSME from models 1- 5 before a slight increase in model 6. Therefore, model 6 can be investigated to know if the model will drop one step further to model 5, having previously dropped model7 because there was no significant increase in the adjusted mean square.

Table 2c: Model accuracies

MODEL ACCURACY (\$)						
	Model1	Model2	Model3	Model4	Model5	Model6
RSME	45578	45444	44335	44966	43517	43,647

3.1.3 Assumption Checks

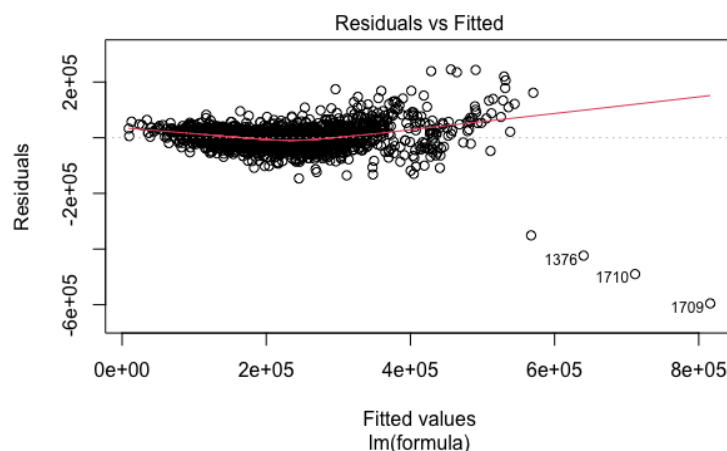
3.1.3.1 Multicollinearity $VIF > 3$

```
> vif(model6)
```

	GVIF	Df	GVIF^(1/(2*Df))
Lot.Area	1.173774	1	1.083409
Gr.Liv.Area	2.726268	1	1.651141
Overall.Qual	3.256447	1	1.804563
Bedroom.AbvGr	1.708770	1	1.307199
BsmtFin.Type.1	5.652346	6	1.155276
Foundation	6.222689	5	1.200599
Condition.1	1.252524	8	1.014172
Total.Bsmt.SF	2.089000	1	1.445337
Exter.Qual	3.843265	4	1.183280
Exter.Cond	1.415923	5	1.035390

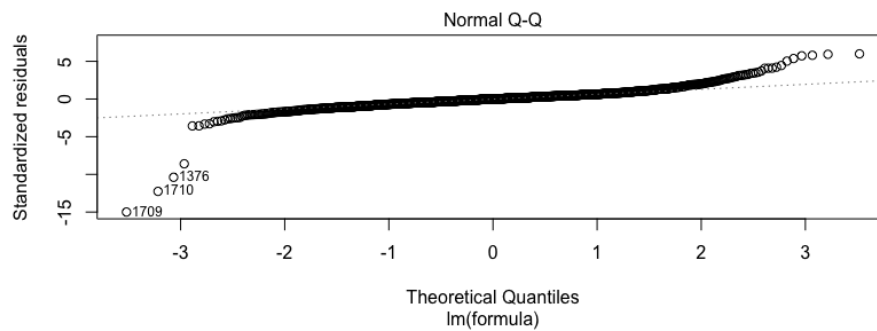
The vif result above indicate that independent variables of *Overall.Qual*, *BsmtFin.Type.1*, *Foundation* and *Exter.Quality* have a slight VIF value > 3 . These variables have failed the multicollinearity check. The VIF for *BsmtFin.Type.1* and *Foundation* is high. This may be because both variables are highly correlated.

3.1.3.2 Checking for Homoscedasticity

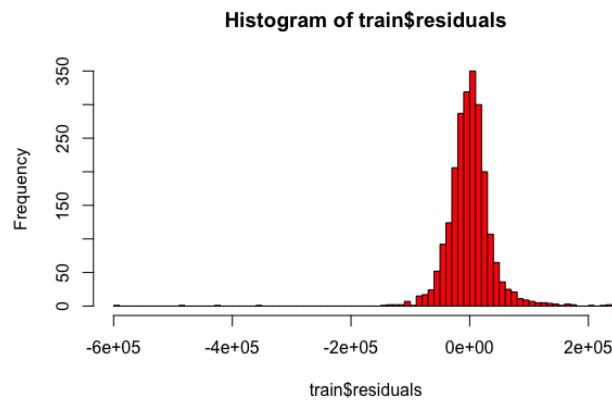


The residuals against the fitted plot show a curved like shape. This may indicate a violation of the assumption of linearity.

3.1.3.3 Check for Normal Q-Q residual plot

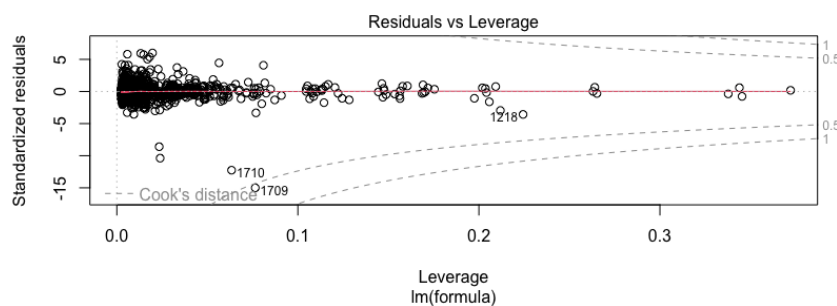


Points from the extremes from the line looks like a deviation from the line. This may indicate that the normality assumption may have been violated. To check further, a histogram of residuals is produced.



The histogram of residuals shows normal distribution, and the Q-Q plot can be seen as a diagonal line. This indicate that the model is good.

3.1.3.4 Cook distance test



The cook distance plot of > 1 is shown above. It can be seen that no observation is found around the 1 mark at the top of the graph. This indicates that model6 passes the cook distance test.

3.1.3.5 Durbin-Watson test

```
> dwtest(model6)

Durbin-Watson test

data:  model6
DW = 1.4589, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The Durbin-Watson assumption test conducted on independent residuals show a dw= 1.45 and p-value < 0.05 indicate the model6 did not meet the assumption test.

3.1.4 Measures of association results

For the measures of association, the Pearson's and spearman's correlation methods were used for the analyses of the selected variables in the ames dataset. It was observed that the correlation between *Sale.Price* and *Lot.Area*, *Bedrooms*, *BsmtFin.Type.1*, *Condition.1* and *Exter.Cond* have a weak positive correlation at 0.2769, 0.1980, 0.3915, 0.1888, 0.1348 and 0.0493 respectively.

However, there was a high positive correlation between *Sale.Price* and *Gr.Liv.Area*, *Overall.Qual*, *Foundation* and *Total.Bsmt.SF* of the buildings at 0.6896, 0.7986, 0.5164 and 0.6265 respectively. In addition, there was a strong negative correlation of *Sale.Price* and *Exter.Qual* of the house at -0.6241, including a weak negative relationship with *Lot.Shape* at -0.3404. It can be argued that the measure of association is not a valid measure of the relationships between variables, thus, a regression model is appropriate.

4.0 Conclusions

This essay was set out to predict sale price of Ames houses using multiple regression model. Five hypotheses were set and one of which is that there is a positive relationship with lot area of a house and sale price. It was interesting to know that despite the expected relationship of sale price and lot area, findings show that there was a slight positive relationship. It could be argued that such findings may be because the property may not be in a location that may interest buyers, thus giving out a large chunk for it expecting to gain huge financial reward may not always follow such trajectory. In practice, the developed model will be useful to accurately predict the value of houses by effectively deploying findings in the regression model. For instance, the issue of multicollinearity will aid in making sure that variables with similar characteristics would not be paired in a multiple regression model as it will result to wrong prediction

5.0 Reflective Commentary

Learning R programming has been rewarding and challenging. It has given me a bigger horizon of how to solve business problems with the use of data. Though most statistics concept were knew to me, but it was exciting knowing the concepts behind certain theories, particularly the concept of nominal, ordinal and numeric values, which builds on the foundation of regression models. I will rate my R skills 5 on a scale of 10, which require that I keep practicing with more datasets to gaining more confidence in the use of the programming language. Another plus in this module is learning the CRISP-DM framework in solving business problems, which will greatly improve my chances at securing a job after the master's programme.

References

- Du, J. and Cross, S. A. (2007) 'Cold in-place recycling pavement rutting prediction model using grey modeling method', *Construction and Building Materials*, 21(5), pp. 921-927
<https://doi.org/10.1016/j.conbuildmat.2006.06.001>.
- Gu, J., Zhu, M. and Jiang, L. (2011) 'Housing price forecasting based on genetic algorithm and support vector machine', *Expert Systems with Applications*, 38(4), pp. 3383-3386.
Available at: <https://doi.org/10.1016/j.eswa.2010.08.123>.
- Manasa, J., Gupta, R. and Narahari, N.S. (2020) 'Machine Learning based Predicting House Prices using Regression Techniques', *International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pp. 624-630. Available at: <https://doi.org/10.1109/ICIMIA48430.2020.9074952>.
- Varol, Y., Oztop, H. F., Koca, A. and Avci, E. (2009) 'Forecasting of entropy production due to buoyant convection using support vector machines (SVM) in a partially cooled square cross-sectional room', *Expert Systems with Applications*, 36(3), pp. 5813-5821.
Available at: <https://doi.org/10.1016/j.eswa.2008.07.016>.
- Viktorovich, P. A., Aleksandrovich, P. V., Leopoldovich, K. I. and Vasilevna, P. I. (2018) 'Predicting sales prices of the houses using regression methods of machine learning', *3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)*, pp. 1-5.
Available at: <https://doi.org/10.1109/RPC.2018.8482191>.

Appendix 1: R Code

Appendix 1: R Code

```
#### Set working directory ##
```

```
setwd("/Volumes/GoogleDrive/My Drive/Documents/MGT_7177/ST-Assignment 1/dataset")
```

```
#### load in libraries
```

```
library(tidyverse)
```

```
library(readxl)
```

```
library(psych)
```

```
library(gridExtra)
```

```
library(factoextra)
```

```
library(dplyr)
```

```
#### load in data
```

```
musei <- read_excel("ames.xlsx")
```

UNDERSTANDING AMES DATASET

```
#### Understanding data, looking for outliers, missing values
```

```
glimpse(musei)
```

```
#### looking at the first 10 rows of the dataset as well as the last 7 rows of the observations
```

```
head(musei, 10)
```

```
tail(musei, 7)
```

```
names(musei)
```

```

### summarise data
summary(musei)

## carrying out descriptive statistics on selected variables that are likely to predict Sale Price
for building hypotheses

SH <- c('Bedroom.AbvGr','Sale.Price',
'Lot.Area','Gr.Liv.Area','Overall.Qual','BsmtFin.Type.1','Foundation',
'Condition.1','Total.Bsmt.SF','Exter.Qual','Exter.Cond')

summary(musei[SH])

### DATA QUALITY ISSUES ##

### Checking for outliers in Sale.Price using ggplot ###

mean(musei$Sale.Price)
summary(musei$Sale.Price)

histo1<- ggplot(musei) +
  geom_histogram(aes(Sale.Price), bins = 100, colour= "red") +
  labs(title = "Sale Price Outlier Fig. 1 - Histogram")

boxplt1 <- ggplot(musei) +
  geom_boxplot(aes(Sale.Price), colour= "red") +
  labs(title = "Sale Price Outlier Fig. 2 - boxplot")

### combining the above visualisation histo1 and boxplt1
grid.arrange(histo1, boxplt1)

### subset to remove outliers in Sale.Price ###
histo2 <- ggplot(musei[musei$Sale.Price < 750000,]) +
  geom_histogram(aes(Sale.Price), bins = 100, colour="Green") +
  labs(title = "Sale Price Clean Fig. 3 - Histogram")

```

```

boxplt2 <- ggplot(musei[musei$Sale.Price < 750000,]) +
  geom_boxplot(aes(Sale.Price), colour="Green") +
  labs(title = "Sale Price Clean Fig. 4 - Boxplot")
### combining the above visualisation histo1 and boxplt1
grid.arrange(histo2, boxplt2)
### Assigning Sale.Price outlier as NA
musei$Sale.Price[musei$Sale.Price > 750000] <- NA

### Remove 2 NAs from Sale.Price outlier and replace with mean value
musei$Sale.Price[is.na(musei$Sale.Price)] <- mean(musei$Sale.Price, na.rm=TRUE)
summary(musei$Sale.Price)

### Checking for outliers in Lot.Area using ggplot ###
histo3 <- ggplot(musei) +
  geom_histogram(aes(Lot.Area), bins = 100, colour= "Blue")+
  labs(title = "Lot Area Outlier Fig. 5 - Histogram")

boxplt3 <- ggplot(musei) +
  geom_boxplot(aes(Lot.Area), colour= "Blue") +
  labs(title = "Lot Area Outlier Fig. 6 - Boxplot")

### ggplot combining Lot Area outliers and Lot Area clean data
grid.arrange(histo3, boxplt3)

### subset to remove outliers in Lot.Area ###
histo4 <- ggplot(musei[musei$Lot.Area < 750000,]) +
  geom_histogram(aes(Lot.Area), bins = 100, colour="Green")+
  labs(title = "Lot Area Clean Fig. 7 - Histogram")

```

```
boxplt4 <- ggplot(musei[musei$Lot.Area < 750000,]) +
  geom_boxplot(aes(Lot.Area), colour="Green") +
  labs(title = "Lot Area Clean Fig. 8 - Boxplot")
```

```
### ggplot combining Lot Area outliers and Lot Area clean data
grid.arrange(histo4, boxplt4)
```

```
### Assigning Lot.Shape outlier as NA
```

```
musei$Lot.Area[musei$Lot.Area > 750000] <- NA
```

```
### Remove 1 NA from Lot.Area outlier and replace with mean value
```

```
musei$Lot.Area[is.na(musei$Lot.Area)] <- mean(musei$Lot.Area, na.rm=TRUE)
```

```
summary(musei$Lot.Area)
```

```
### Checking for data issues in Overall Quality using ggplot ###
```

```
boxplot5 <- ggplot(musei,aes(x=as.factor(Overall.Qual), y= Sale.Price))+
  geom_boxplot()+
  geom_point(colour='red')+
  labs(title = "Overall Quality Dirty Fig. 9 - Boxplot")
```

```
### Fixing data quality issue in Overall.Qual. It is supposed to have 10 levels not 11 levels
according to the ames data dictionary
```

```
musei<-musei %>%
```

```
  filter(Overall.Qual<=10)
```

```
boxplot6 <- ggplot(musei,aes(x=as.factor(Overall.Qual), y= Sale.Price))+
  geom_boxplot()+
  geom_point(colour='green')+
  labs(title = "Overall Quality Clean Fig. 10 - Boxplot")
```

```
### ggplot combining Initial error and clean Overall.Quality variable  
grid.arrange(boxplot5, boxplot6)
```

```
### Checking for data issues in Ground Living Area using ggplot ###
```

```
histo6<- ggplot(musei) +  
  geom_histogram(aes(Gr.Liv.Area), bins = 100, colour= "red") +  
  labs(title = "Ground Living Area Fig. 11 - Histogram")
```

```
boxplt7 <- ggplot(musei) +  
  geom_boxplot(aes(Gr.Liv.Area), colour= "red") +  
  labs(title = " Ground Living Area Fig. 12 - boxplot")
```

```
### combining the two visualisations above using ggplot2  
grid.arrange(histo6, boxplt7)
```

```
### Summarising Gr.Liv.Area  
summary(musei$Gr.Liv.Area)
```

```
### Removing 3 NAs and assigning the mean value to them  
musei$Gr.Liv.Area[is.na(musei$Gr.Liv.Area)] <- mean(musei$Gr.Liv.Area, na.rm=TRUE)  
summary(musei$Gr.Liv.Area)
```

```
### Checking for data issues in Total.Bsmt.SF using ggplot ###  
histo9 <- ggplot(musei) +  
  geom_histogram(aes(Total.Bsmt.SF), bins = 100, colour= "Blue")+  
  labs(title = "Total.Bsmt.SF Outlier Fig. 13 - Histogram")
```



```

boxplt9 <- ggplot(musei) +
  geom_boxplot(aes(Total.Bsmt.SF), colour= "Blue") +
  labs(title = "Total.Bsmt.SF Outlier Fig. 14 - Boxplot")

### ggplot combined plot of histogram and boxplot showing NA in Total.Bsmt.SF
grid.arrange(histo9, boxplt9)

### summarising Total.Bsmt.SF
summary(musei$Total.Bsmt.SF)

### Removing 1 NA and assigning the mean value to them
musei$Total.Bsmt.SF[is.na(musei$Total.Bsmt.SF)] <- mean(musei$Total.Bsmt.SF,
na.rm=TRUE)
summary(musei$Total.Bsmt.SF)

### Converting character variables to factor in preparation for correlation test

### Setting musei$BsmtFin.Type.1 to factor
musei$BsmtFin.Type.1 <- factor(replace_na(musei$BsmtFin.Type.1, '0'), levels = c('0', 'Unf',
'LwQ', 'Rec', 'BLQ', 'ALQ', 'GLQ'))

### Setting musei$BsmtFin.Type.2 to factor
musei$BsmtFin.Type.2 <- factor(replace_na(musei$BsmtFin.Type.2, '0'), levels = c('0', 'Unf',
'LwQ', 'Rec', 'BLQ', 'ALQ', 'GLQ'))

### Setting External condition to factor
musei$Exter.Cond <- factor(musei$Exter.Cond)

```

```
### Setting Roof Style to factor
musei$Roof.Style <- as.factor(musei$Roof.Style)
levels(musei$Roof.Style)
```

```
### Setting Lot.shape to factor
musei$Lot.Shape <- factor(musei$Lot.Shape)
```

HYPOTHESES

```
## h1 lot area is positively related to Sale Price
## h2 bedroom is positively related to Sale Price
## h3 Ground living area is positively related to Sale Price
## h4 overall quality is positively related to Sale Price
## h5 BsmtFin.Type.1 is positively related to Sale Price
```

VISUALISATION

```
## sale price by lot area showing scatter plot and lm
ggplot(data=musei) +
  geom_point(mapping=aes(x=Lot.Area, y=Sale.Price, colour="black")) +
  geom_smooth(method = "lm", formula = y~x, mapping = aes(x=Lot.Area, y= Sale.Price)) +
  labs(title="Sale Price by Lot Area", x="Lot.Area(sqft)", y="Sale.Price($)")
```

```
## sale price by Bedrooms
ggplot(data=musei) +
  geom_boxplot(mapping=aes(x=as.factor(Bedroom.AbvGr), y=Sale.Price)) +
  labs(title="Sale Price by Bedroom", x="No of Bedrooms (Features)", y="Sale.Price($)")
```

```
## sale price by Ground Living area showing scatter plot and lm
ggplot(data=musei) +
```

```

geom_point(aes(x=Gr.Liv.Area, y=Sale.Price, colour= 'green')) +
geom_smooth(method = "lm", formula = y~x, mapping = aes(x=Gr.Liv.Area, y= Sale.Price))
+
labs(title="Sale Price by Ground Living Area",x="Gr.Liv.Area (sqft)", y="Sale.Price($)")

```

sale price by lot Overall Quality

```

ggplot(data=musei) +
geom_boxplot(mapping=aes(x=as.factor(Overall.Qual), y=Sale.Price)) +
geom_point(aes(x=Overall.Qual, y=Sale.Price, colour= 'green')) +
labs(title="Sale Price by Overall Quality", x="Overall Quality", y="Sale.Price($)")

```

sale price by Basement Finish Type 1

```

ggplot(data=musei) +
geom_boxplot(mapping=aes(x=as.factor(BsmtFin.Type.1), y=Sale.Price)) +
geom_point(aes(x=BsmtFin.Type.1, y=Sale.Price, colour= 'red')) +
labs(title="Sale Price by Basement Finish Type 1", x="Basement Finish 1 (Features)",
y="Sale.Price($)")

```

sale price by lot Foundation

```

ggplot(data=musei) +
geom_boxplot(mapping=aes(x=Foundation, y=Sale.Price)) +
geom_point(aes(x=Foundation, y=Sale.Price, colour= 'green')) +
labs(title="Sale Price by Foundation", x="Foundation (sqft)", y="Sale.Price($)")

```

sale price by Building Type

```

ggplot(data=musei) +
geom_boxplot(mapping=aes(x=as.factor(Bldg.Type), y=Sale.Price)) +
geom_point(aes(x=Bldg.Type, y=Sale.Price)) +
labs(title="Sale Price by Building Type", x="Building Type (Features)", y="Sale.Price($)")

```

```
## sale price by Basement Quality
ggplot(data=musei) +
  geom_boxplot(mapping=aes(x=as.factor(Bsmt.Qual), y=Sale.Price)) +
  geom_point(aes(x=Bsmt.Qual, y=Sale.Price, colour= 'green')) +
  labs(title="Sale Price by Basement Quality", x="Basement Quality (Features)",
y="Sale.Price($)")
```

```
## sale price by Basement Finish Type 2
ggplot(data=musei) +
  geom_boxplot(mapping=aes(x=as.factor(BsmtFin.Type.2), y=Sale.Price)) +
  geom_point(aes(x=BsmtFin.Type.1, y=Sale.Price)) +
  labs(title="Sale Price by Basement Finish Type 2", x="Basement Finish 2 (Features)",
y="Sale.Price($)")
```

```
## sale price by External Condition
ggplot(data=musei) +
  geom_boxplot(mapping=aes(x=as.factor(Exter.Cond), y=Sale.Price)) +
  geom_point(aes(x=Exter.Cond, y=Sale.Price)) +
  labs(title="Sale Price by External Condition", x="External Condition (Features)",
y="Sale.Price($)")
```

CORRELLATION

```
## correlation 1
cor(musei$Lot.Area, musei$Sale.Price)
cor.test(musei$Lot.Area, musei$Sale.Price)
```

```
## correlation 2
cor(musei$Bedroom.AbvGr, musei$Sale.Price, method = "spearman")
```

```

## correlation 3
cor(musei$Gr.Liv.Area, musei$Sale.Price)

## correlation 4
cor(as.numeric(musei$Overall.Qual), musei$Sale.Price)

## correlation 5
bsmt_type <- as.numeric(musei$BsmtFin.Type.1)
cor(bsmt_type, musei$Sale.Price, method = "spearman")

## correlation 6
musei$Foundation <- as.factor(musei$Foundation)
Fdn <- as.numeric(musei$Foundation)
cor(Fdn, musei$Sale.Price, method="spearman")

## correlation 7
musei$Condition.1 <- as.factor(musei$Condition.1)
Cd1 <- as.numeric(musei$Condition.1)
cor(Cd1, musei$Sale.Price, method = "spearman")

## correlation 8
cor(musei$Total.Bsmt.SF, musei$Sale.Price)

## correlation 9
musei$Exter.Qual <- as.factor(musei$Exter.Qual) ## negative correlation
Extd <- as.numeric(musei$Exter.Qual)
cor(Extd, musei$Sale.Price, method = "spearman")

## correlation 10

```

```
musei$Exter.Cond <- as.factor(musei$Exter.Cond)
```

```
Ext <- as.numeric(musei$Exter.Cond)
```

```
cor(Ext, musei$Sale.Price, method = "spearman")
```

```
## correlation 11
```

```
lahPE <- as.numeric(musei$Lot.Shape)      ## negative correlation
```

```
cor(lahPE, musei$Sale.Price, method="spearman")
```

```
## correlation 12
```

```
bsmt_type2 <- as.numeric(musei$BsmtFin.Type.2)
```

```
cor(bsmt_type2, musei$Sale.Price, method = "spearman")
```

MULTIPLE REGRESSION

```
### Load in caret library in preparation for regression
```

```
library(caret)
```

```
### Set seed to keep values of regression constant
```

```
set.seed(1845)
```

```
index <- createDataPartition(musei$Sale.Price, times =1, p =0.8, list= FALSE)
```

```
train <- musei[index,]
```

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

```
### build the model on train data
```

```
### model 1
```

```
formula <- Sale.Price ~ Lot.Area + Gr.Liv.Area + Overall.Qual + Bedroom.AbvGr +  
BsmtFin.Type.1
```

```
model1 <- lm(formula, train)
```

```
summary(model1)
```

```
### model 2
```

```
formula <- Sale.Price ~ Lot.Area + Gr.Liv.Area + Overall.Qual + Bedroom.AbvGr +  
BsmtFin.Type.1 + Foundation
```

```
model2 <- lm(formula, train)
```

```
summary(model2)
```

```
### model 3
```

```
formula <- Sale.Price ~ Lot.Area + Gr.Liv.Area + Overall.Qual + Bedroom.AbvGr +  
BsmtFin.Type.1 + Foundation + Condition.1
```

```
model3 <- lm(formula, train)
```

```
summary(model3)
```

```
### model 4
```

```
formula <- Sale.Price ~ Lot.Area + Gr.Liv.Area + Overall.Qual + Bedroom.AbvGr +  
BsmtFin.Type.1 + Foundation + Condition.1 + Total.Bsmt.SF
```

```
model4 <- lm(formula, train)
```

```
summary(model4)
```

```
### model 5
```

```
formula <- Sale.Price ~ Lot.Area + Gr.Liv.Area + Overall.Qual + Bedroom.AbvGr +  
BsmtFin.Type.1 + Foundation + Condition.1 + Total.Bsmt.SF + Exter.Qual
```

```
model5 <- lm(formula, train)
```

```
summary(model5)
```

```
### model 6
```

```
formula <- Sale.Price ~ Lot.Area + Gr.Liv.Area + Overall.Qual + Bedroom.AbvGr +  
BsmtFin.Type.1 + Foundation + Condition.1 + Total.Bsmt.SF + Exter.Qual + Exter.Cond
```

```
model6 <- lm(formula, train)
```

```
summary(model6)
```

```
### model 7
```

```
formula <- Sale.Price ~ Lot.Area + Gr.Liv.Area + Overall.Qual + Bedroom.AbvGr +  
BsmtFin.Type.1 + Foundation + Condition.1 + Total.Bsmt.SF + Exter.Qual+ Exter.Cond +  
Lot.Shape
```

```
model7 <- lm(formula, train)
```

```
summary(model7) ### No significant increase in the model looking at the Adjusted R squared  
result reason for dropping the model
```

```
### check model1 accuracy of the test data (20%)
```

```
predictions <- predict(model1, test)
```

```
postResample(predictions, test$Sale.Price)
```

```
### check model2 accuracy of the test data (20%)
```

```
predictions <- predict(model2, test)
```

```
postResample(predictions, test$Sale.Price)
```

```
### check model3 accuracy of the test data (20%)
```

```
predictions <- predict(model3, test)
```

```
postResample(predictions, test$Sale.Price)
```

```
### check model4 accuracy of the test data (20%)
```

```
predictions <- predict(model4, test)
```

```
postResample(predictions, test$Sale.Price)
```



```
### check model5 accuracy of the test data (20%)  
predictions <- predict(model5, test)  
postResample(predictions, test$Sale.Price)
```

```
### check model6 accuracy of the test data (20%)  
predictions <- predict(model6, test)  
postResample(predictions, test$Sale.Price)  
test$pred <- predictions
```

CHECKING FOR ASSUMPTIONS

```
### Checking for Multicollinearity to see if VIF > 3  
library(car)  
vif(model6)
```

```
### Checking for Homoscedasticity  
plot(model6) ### 1st plot showing Residuals Vs Fitted values
```

```
### Normally distributed residuals- Check Normal Q-Q residual plot  
plot(model6) ### 2nd plot showing Residuals Vs Normal distribution
```

```
### Influential cases: Check cook distance > 1  
cooks <- cooks.distance(model6)  
sum(cooks > 1)
```

Independent residuals : run dwtest . value between 1.5 and 2.5 is good

```
library(lmtest)
```

```
dwtest(model6)
```

```
train$residuals <- resid(model6)
```

```
train$predictions <- fitted(model6)
```

```
hist(train$residuals, breaks= 100, col = "red") ### plot histogram to see if it is normally distributed
```

```
plot(train$residuals, train$Sale.Price)
```

```
summary(model6)
```