Toronto

Regression Diagnostics

Soni Manan[a]

[a] *College of Professional Studies, Master of Professional Studies in Analytics.*

***Subject**: ALY6015 **NUID**: 002982645*

Under the guidance of

Dr. Prof. Alex Maizlish

Introduction

This is the data about housing prices and its information of Ames. It consist of 2930 samples and 82 variables. The name of data set is Ames Iowa. Initially, dataset was uncleaned. Data has samples in the range of 2006 to 2010.

What do we get by this dataset?

This dataset is about characteristic of houses of Ames and its sales. Therefore, we can serve analysis to gain such benefits. From the first look, the most important variable is the sales price by which we can have best line for sectors like Real estate and banking. For real estate, variables are much useful as data has a lot of characteristics about houses from build and banks can increase their dept as mortgage.

Business Analysis

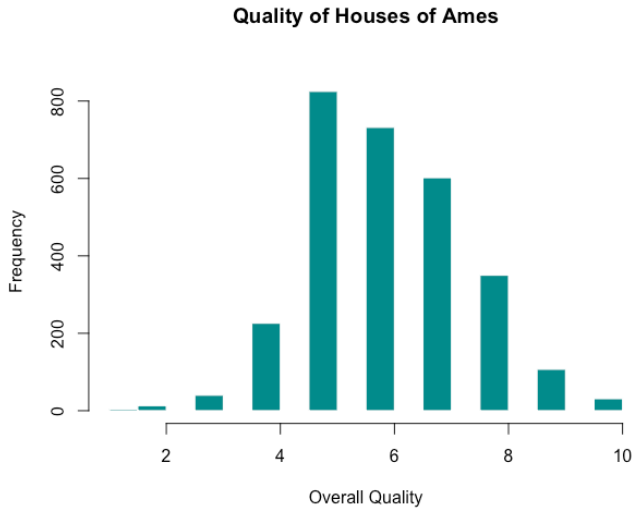
- **Initial**

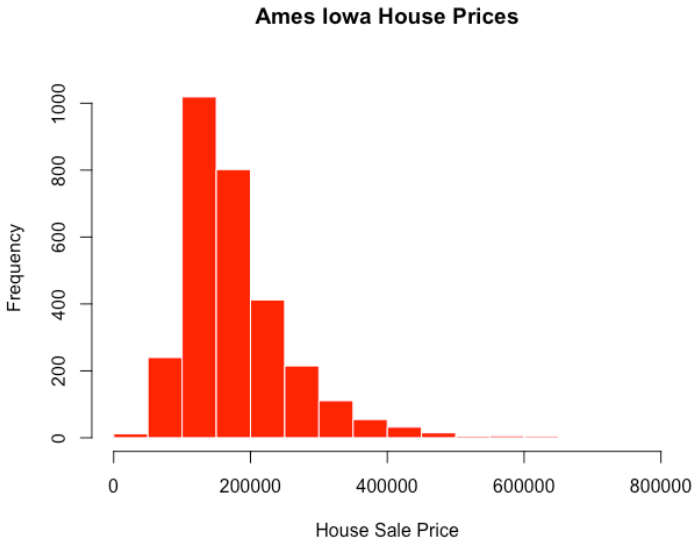
  There are 82 variables in total where there are categorical, discrete and continuous. It's not clean and data has only 2930 samples, instead of removing placed mean of respective variables.

- **Descriptive Analysis**
  - Sale price is varying from 12789 to 755000 with mean value of 180796. Looking at min value it's not possible so it will have outliers.
  - Garage cars, garage area and overall quality has minor skewness and almost normal distributions.
  - Sale price has 1.74 and 1st floor sq.ft has 1.47 which are positive skewed.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
SalePrice	1	2930	180796.06	79886.69	160000	170429.15	54856.20	12789	755000	742211	1.74	5.10	1475.84
X1st.Flr.SF	2	2930	1159.56	391.89	1084	1127.17	349.89	334	5095	4761	1.47	6.95	7.24
Garage.Cars	3	2930	1.77	0.76	2	1.77	0.00	0	5	5	-0.22	0.24	0.01
Garage.Area	4	2930	472.82	215.01	480	468.32	182.36	0	1488	1488	0.24	0.95	3.97
Gr.Liv.Area	5	2930	1499.69	505.51	1442	1452.25	461.09	334	5642	5308	1.27	4.12	9.34
Overall.Qual	6	2930	6.09	1.41	6	6.08	1.48	1	10	9	0.19	0.05	0.03
Overall.Cond	7	2930	5.56	1.11	5	5.47	0.00	1	9	8	0.57	1.48	0.02

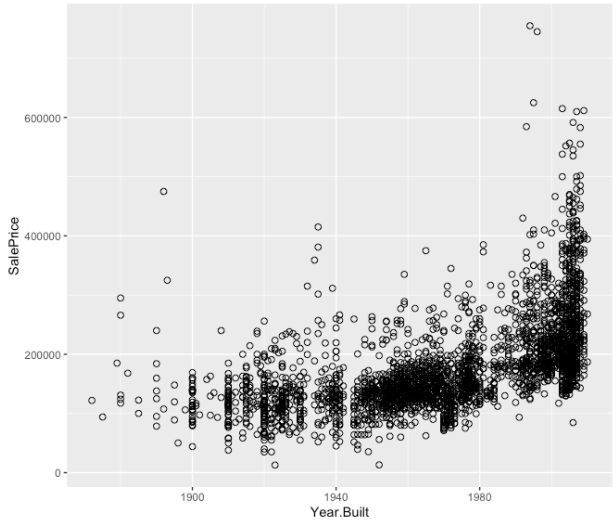
- **Exploratory Analysis**
  - The average price is 180000 and so the most of the samples lies between 0 to 200000. And looking at histogram of house qualities are 5 or above and there are very few like less than 50 for quality of 1,2 and 10.
  - Data cleaning is important so I replaced value with mean and took numeric variables.

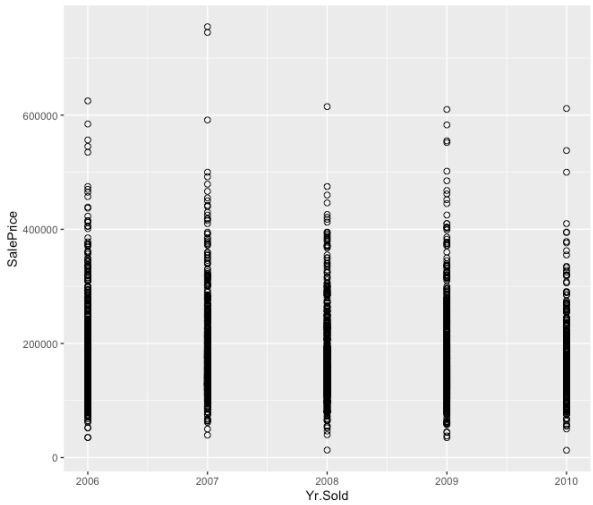


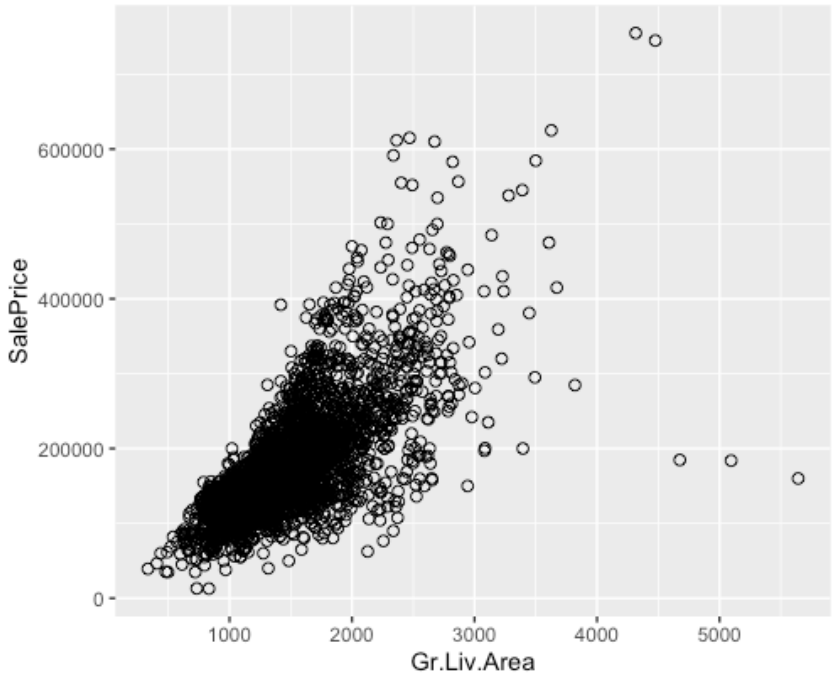


.

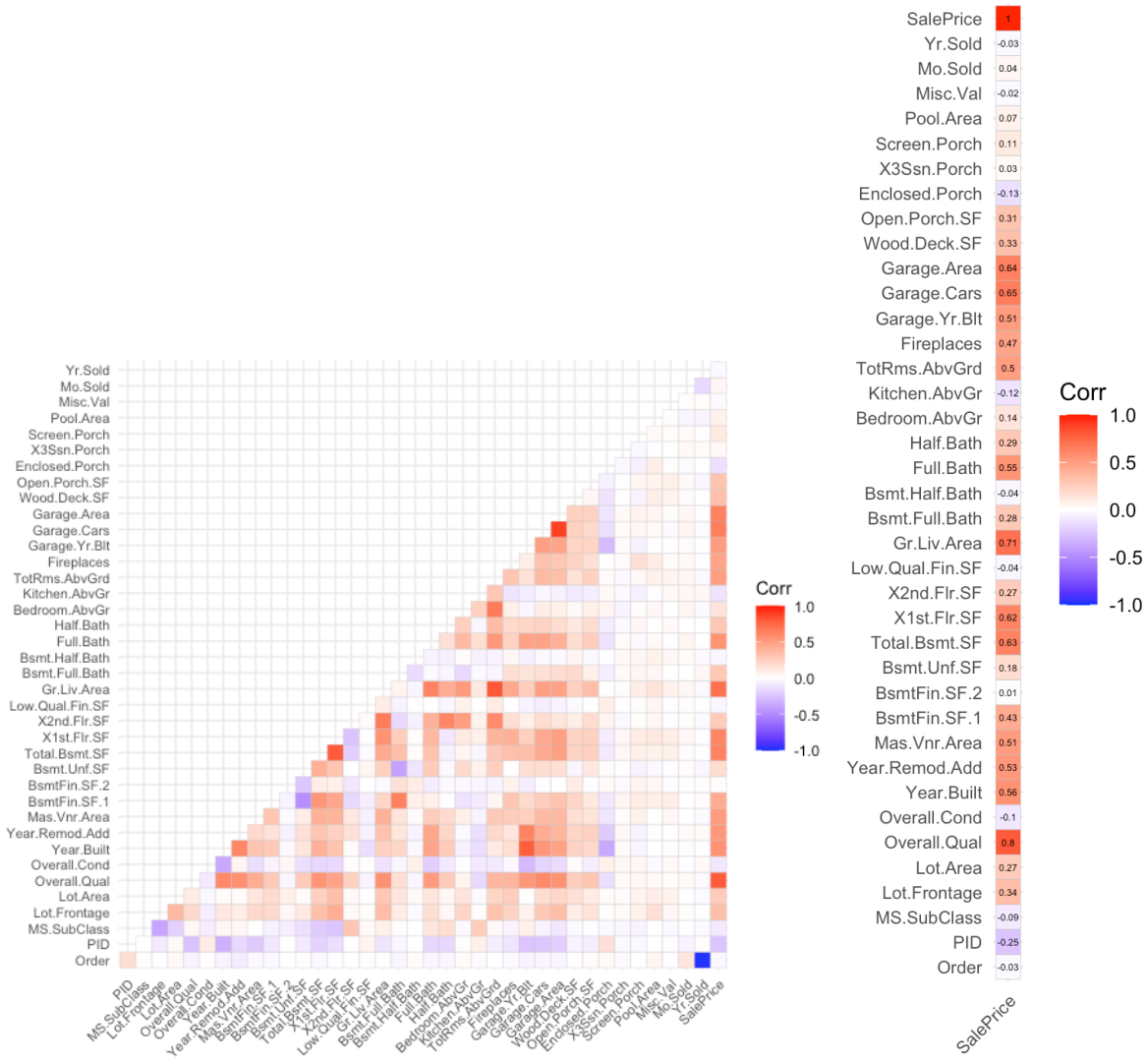
- **Linearity and Correlation**
  - Linearity is the important check for regression diagnostics. Variables have some kind of relation that grows in linear fashion.
  - As we can be seen there is no relation between house and its sold year.
  - Other to have relation but strong and weak like area above ground level looks like linear pattern where year build has weak linear.
  - To identify, Correlation matrix is useful.



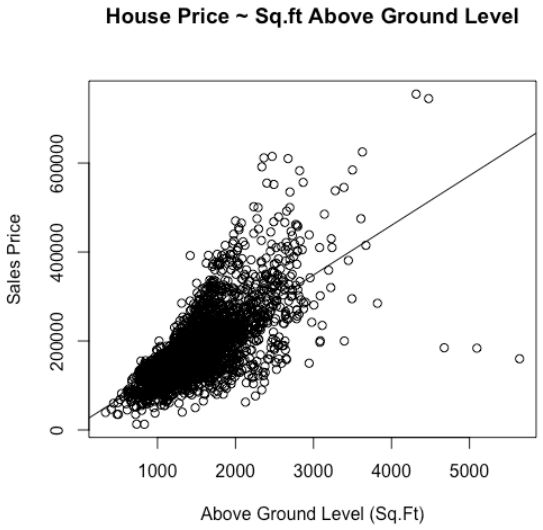


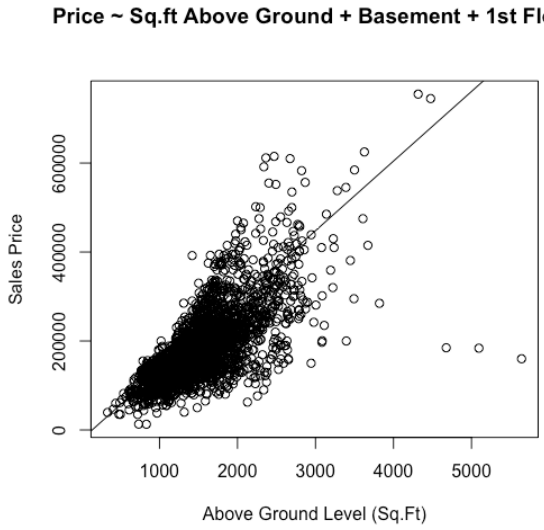


  - Color closes to red means value of correlation is near to 1 is positive linear and to -1 is negative.


Yr.Sold
Mo.Sold
Misc.Val
Pool.Area
Screen.Porch
X3Ssn.Porch
Enclosed.Porch
Open.Porch.SF
Wood.Deck.SF
Garage.Area
Garage.Cars
Garage.Yr.Blt
Fireplaces
TotRms.AbvGrd
Kitchen.AbvGr
Bedroom.AbvGr
Half.Bath
Full.Bath
Bsmt.Half.Bath
Bsmt.Full.Bath
Gr.Liv.Area
Low.Qual.Fin.SF
X2nd.Flr.SF
X1st.Flr.SF
Total.Bsmt.SF
Bsmt.Unf.SF
BsmtFin.SF.2
BsmtFin.SF.1
Mas.Vnr.Area
Year.Remod.Add
Year.Built
Overall.Cond
Overall.Qual
Lot.Area
Lot.Frontage
MS.SubClass
PID
Order
Corr
1.0
0.5
0.0
-0.5
-1.0
SalePrice
1
-0.03
0.04
-0.02
0.07
0.11
0.03
-0.13
0.31
0.33
0.64
0.65
0.51
0.47
0.5
-0.12
0.14
0.29
0.55
-0.04
0.28
0.71
-0.04
0.27
0.62
0.63
0.18
0.01
0.43
0.51
0.53
0.56
-0.1
0.8
0.27
0.34
-0.09
-0.25
-0.03
SalePrice
Corr
1.0
0.5
0.0
-0.5
-1.0

- Values near to 0 are independent. Value > 0.5 is good for regression diagnostics.
- Overall Quality, Sq.ft greater than ground level, Garage cars and Garage Area are correlated to sale prices.
- As we can note in the diagrams line of regression has inclined more and according to trend which demonstrates that accuracy of this model has been improved and errors are minimized to some extend.

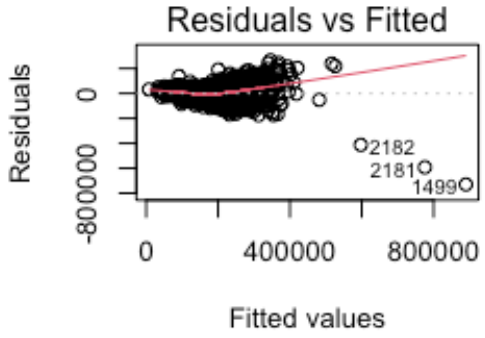


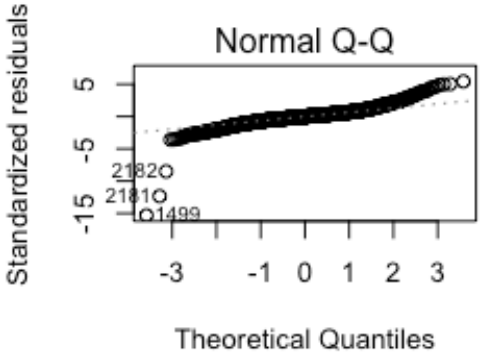


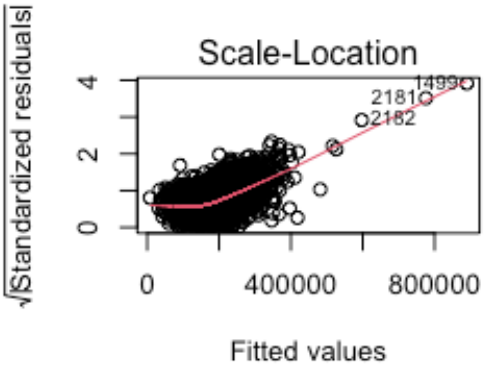
- As we can see that R-square value is 49.95% at first. But later it increased to 62.52%
- But we need to check for more if we can improve.
- All of the value of P is almost 0 or less than 0.05 which shows that all of the variable are significant for model.
- Adding variable who are correlating can help to getting line of the best fits.

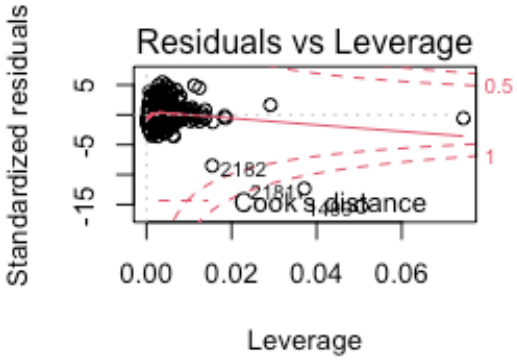
```
                   Estimate Std. Error t value             Pr(>|t|)
(Intercept)       13289.634   3269.703   4.064            0.0000494 ***
nData$Gr.Liv.Area   111.694      2.066  54.061 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56520 on 2928 degrees of freedom
Multiple R-squared:  0.4995,    Adjusted R-squared:  0.4994
F-statistic:  2923 on 1 and 2928 DF,  p-value: < 0.00000000000000022
```

```
                     Estimate Std. Error t value             Pr(>|t|)
(Intercept)         -21415.703   3161.525  -6.774      0.0000000000151 ***
nData$Gr.Liv.Area       83.043      2.162  38.417 < 0.0000000000000002 ***
nData$X1st.Flr.SF        4.096      4.169   0.982                0.326
nData$Total.Bsmt.SF     69.345      3.424  20.253 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48910 on 2926 degrees of freedom
Multiple R-squared:  0.6256,    Adjusted R-squared:  0.6252
F-statistic:  1630 on 3 and 2926 DF,  p-value: < 0.00000000000000022
```









- There are a lot things which can lead to ambiguity and some samples and outliers can do that.

- Adding variable who are correlating can help to getting line of the best fits.
- Accuracy 0.5 to 0.9 is good greater than that can be overfitting as there can't be perfect sample.

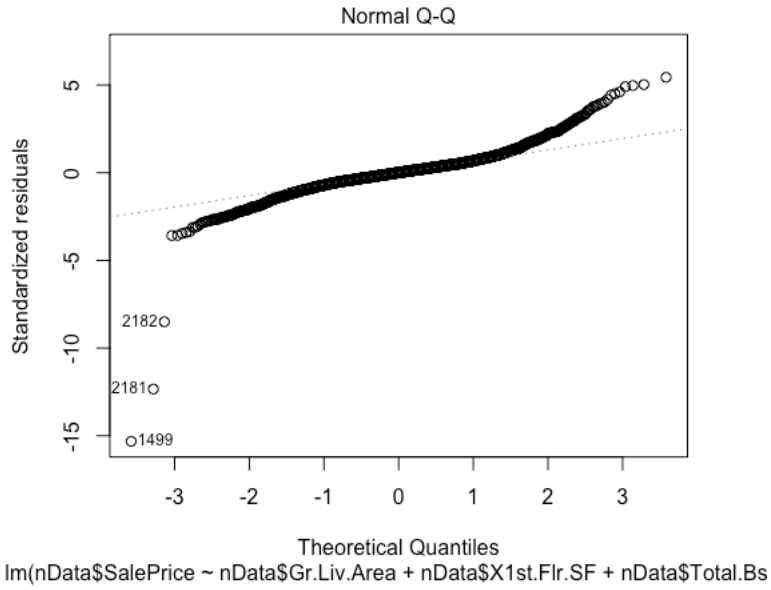
Multi-collinearity

- This exists when an independent variable corelates with multiple variables. So simple regression becomes multi linear regression. As above demonstrated, The accuracy had been increased with adding variables like basement sq.ft and first floor sq.ft.
- A variance inflation factor (VIF) is tool to identify the degree of collinearity. In statistical term, a multi linear regression is difficult to guess which variable affects more to the model.
- The values of vif illustrates relation between variables.

```
nData$Gr.Liv.Area   nData$X1st.Flr.SF nData$Total.Bsmt.SF
         1.462236            3.268570            2.786283
```

Outliers who effects the performance

- Those who are around -5 to -15 are outliers. Probably removing them can cause positive effects on model.



- Removing the outlier has advantages as show in the fig R-squared value has increased to 68%.

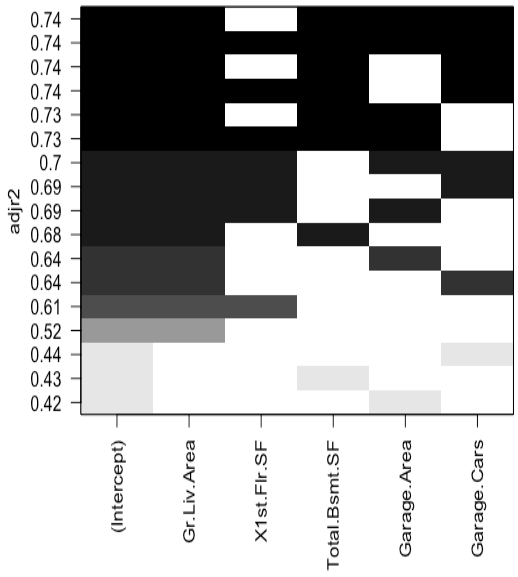
```
Coefficients:
                            Estimate Std. Error t value            Pr(>|t|)
(Intercept)               -29314.955   2880.508  -10.18 <0.0000000000000002 ***
cleanedData$Gr.Liv.Area       84.351      1.892   44.59 <0.0000000000000002 ***
cleanedData$X1st.Flr.SF        2.670      3.708    0.72                0.471
cleanedData$Total.Bsmt.SF     76.067      3.075   24.74 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41070 on 2890 degrees of freedom
Multiple R-squared:   0.68,     Adjusted R-squared:  0.6797
F-statistic:  2047 on 3 and 2890 DF,  p-value: < 0.00000000000000022
```

The Formula:

SalePrice = Intercept + coeff1 * (Gr.Liv.Area) + coeff2 * (X1st.Flr.SF) + coeff3 * (total.Bsmt.SF)

All Subset Regression

- This is the method to identify the value of R-square. i.e. How we can achieve or with what variables we can achieve.
- As you can see black portion is the variable parts by which we can get 0.74 accuracy.



References

RDocumentation, *2022. Regsubs Source:* https://www.rdocumentation.org/packages/leaps/versions/3.1/topics/regsubsets

Databizpyr, 2022. *Pchs in R Source*: https://datavizpyr.com/pch-in-r-built-in-shapes-in-r/

RDocumentation, *2022. ggcorrplot: Visualization of a correlation matrix using ggplot2 Source:* https://www.rdocumentation.org/packages/leaps/versions/3.1/topics/regsubsets

Sthda, 2022, ggplot2 scatterplot Source: http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization

Appendix

```
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

#1
mData = read.csv('./AmesHousing.csv')
str(mData)
# 2
# getting rows and cols count
length(mData) # columns
nrow(mData) # rows

library(dplyr)
nData = mData %>% select(where(is.numeric))

for(i in 1:ncol(nData)){
  nData[is.na(nData[,i]), i] <- mean(nData[,i], na.rm = TRUE)
}

psych::describe(nData[c('SalePrice','X1st.Flr.SF','Garage.Cars','Garage.Area','Gr.Liv.Area','Overall
.Qual','Overall.Cond')])
options(scipen=999)
hist(nData$SalePrice,xlab='House Sale Price',main='Ames Iowa House
Prices',col='red',border='white')
hist(nData$Overall.Qual,xlab='Overall Quality',main='Quality of Houses of
Ames',col='cyan4',border='white')

cor(nData['SalePrice'],nData[1:38])

library(ggcorrplot)
ggcorrplot(cor(nData),  insig = "blank", lab_size=2,type='lower',tl.cex = 8)
ggcorrplot(cor(nData['SalePrice'],nData[1:39]), lab=TRUE,lab_size=1.5, insig = "blank", tl.cex=8)

library(ggplot2)
options(scipen=999)
# highest
```

```
ggplot(nData, aes(x=Gr.Liv.Area, y=SalePrice)) +
  geom_point(size=2, shape=1)
# lowest
ggplot(nData, aes(x=Yr.Sold, y=SalePrice)) +
  geom_point(size=2, shape=1)
# 0.5
ggplot(nData, aes(x=Year.Built, y=SalePrice)) +
  geom_point(size=2, shape=1)

# 7
par(mfrow=c(1,2))
# a
SL_lmodel = lm(nData$SalePrice ~ nData$Gr.Liv.Area)
plot(nData$SalePrice ~ nData$Gr.Liv.Area,xlab="Above Ground Level (Sq.Ft)",ylab="Sales Price",main='House Price ~ Sq.ft Above Ground Level')
abline(a= SL_lmodel$coefficients[1],b=SL_lmodel$coefficients[2])
summary(SL_lmodel)

# b
ML_lmodel = lm(nData$SalePrice ~ nData$Gr.Liv.Area + nData$X1st.Flr.SF + nData$Total.Bsmt.SF)
plot(nData$SalePrice ~ nData$Gr.Liv.Area,xlab="Above Ground Level (Sq.Ft)",ylab="Sales Price",main='Price ~ Sq.ft Above Ground + Basement + 1st Floor')
abline(a= ML_lmodel$coefficients[1], b=ML_lmodel$coefficients[2]+ML_lmodel$coefficients[3]+ML_lmodel$coefficients[4])
summary(ML_lmodel)

par(mfrow=c(1,1))
plot(SL_lmodel)
plot(ML_lmodel)


library(car)
vif(ML_lmodel)

# 11
outlierTest(ML_lmodel)
```

```
par(mfrow=c(1,1))
hat.plot <- function(ML_lmodel) {
  p <- length(coefficients(ML_lmodel))
  n <- length(fitted(ML_lmodel))
  plot(hatvalues(ML_lmodel), main="Index Plot of Hat Values")
  abline(h=c(2,3)*p/n, col="red", lty=2)
  identify(1:n, hatvalues(ML_lmodel), names(hatvalues(ML_lmodel)))
}
hat.plot(ML_lmodel)

# 12
cooksValue = cooks.distance(ML_lmodel)
influential_points = cooksValue[(cooksValue>3*mean(cooksValue,na.rm=T))]
names = names(influential_points)
outliers = nData[names,]
cleanedData = nData%>%anti_join(outliers)

#qqnorm()
par(mfrow=c(1,1))
WO_ML_lmodel<-
lm(cleanedData$SalePrice~cleanedData$Gr.Liv.Area+cleanedData$X1st.Flr.SF
+cleanedData$Total.Bsmt.SF)
plot(cleanedData$SalePrice ~ cleanedData$Gr.Liv.Area)
abline(a=WO_ML_lmodel$coefficients[1],b=WO_ML_lmodel$coefficients[1]+WO_ML_lmodel$c
oefficients[2])
summary(WO_ML_lmodel)
plot(WO_ML_lmodel)

#13
library(MASS)
stepAIC(WO_ML_lmodel,direction = "both")
par(mfrow=c(1,1))
library(leaps)
leaps <- regsubsets(SalePrice~Gr.Liv.Area+X1st.Flr.SF +Total.Bsmt.SF + Garage.Area +
Garage.Cars,
                    data=withoutOutliers,nbest=4)
plot(leaps,scale="adjr2",ylab='Adjusted R-Square Value')
```