Project 1

Wong Jing Hei (Koby)

Contents

Choosing Descriptive Variables

The dataset train.csv contains 79 explanatory variables. The data description and csv file can be downloaded directly from kaggle. Your task, as suggested on the kaggle website, is to build a model to predict final home prices. Note, this is part of a kaggle competition which you might consider participating in later on. Before you start the parts below, identify any 10 variables of your choice and write a brief paragraph of why you selected them. These are the predictors you will use for solving the problem.

```
data<-read.csv("train.csv")
library(car)
library(MASS)
library(AER)
library(broom)
library(leaps)
library(caret)
library(rpart)
library(corrplot)
```

Var1. Above Grade (Ground) Living Area in Square Feet (GrLivArea):
The size of housing has always been a key factor that affects its price. While other measures of size in square feet are provided in the dataset, it is expected that most buyers would care more about the size of the living area of the house as they would spend the most of their time there instead of other areas in the property such as the garage or the basement. Therefore, it is chosen as the key measure of size for as one of my variables.

Var2 Year Sold (YrSold) (2006-2010):
It is important to consider the year the house was sold as inflation plays a key role in affecting house prices. Exogenous variables may also affect housing prices, such as the financial crisis in 2008. By holding the year sold constant, the effects of other variables would become more apparent.

Var3 Overall Material and Finish of the House (OverallQual):
Higher quality of material would tend to lead to higher prices, and is hence included in the regression.

Var4 Overall Condition of the House (OverallCond):
Houses in poor condition irrespective of the quality of material may still cause prices to decrease, and hence must be considered as a key factor in affecting house prices.

Var5 Heating quality and condition (HeatingQC):
The survey was conducted in Iowa, an area that gets very low temperatures especially during the latter months of the year, hence heating quality will become a key factor many buyers consider when buying houses. This would be expected to heavily affect house prices and hence is included as one of my variables.

Var6 Kitchen Quality (KitchenQual):
All homes have kitchens, and its quality would be a important consideration for buyers when they purchase a property, affecting housing prices. Hence it is included as one of my variables of analysis.

Var7 Basement Height (BsmtQual):
Summary statistics of the dataset indicates that most houses in this dataset have a basement. While homeowners may use the basement in various ways (Storage/ potential living area), a basement with a higher height is beneficial regardless of it's intended purposes (higher height means more storage or a more comfortable living area) and would affect house prices. Hence it is chosen as one of my variables.

Var8 Garage Capacity (GarageCars):
Summary statistics of the dataset also indicates that most houses in this dataset have a garage. While measures of garage size in square feet is available, I think the most important consideration for home buyers is it's total car capacity. Increased capacity of car storage is expected to increase house prices, and is therefore chosen as one of my variables.

Var9 Remodel Date (YearRemodAdd):
Newer houses are expected to be more expensive, while older or less refurbished houses are expected to be cheaper. While this variable may potentially be correlated with the houses' condition, preliminary summary statistics indicate a low correlation between the two factors, and hence is included as one of my variables.

Var10 Full Bathrooms Above Grade:
Bathrooms are a necessity for a home. The more bathrooms a home has, the higher functionality and convenience for the house and the house owner. Hence It is included as one of my variables.

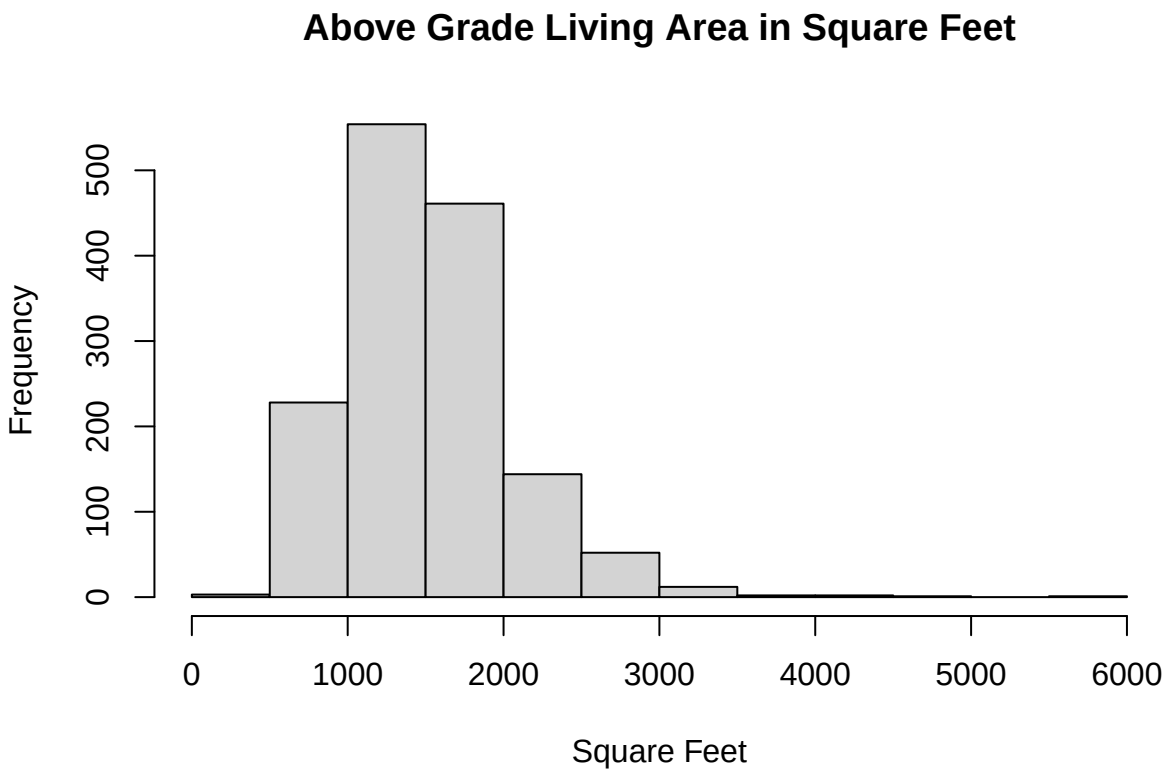
Variables that were considered but not included:
Many variables in this dataset were initially considered but not included as key variables due to the disporportional distribution of these variables in the data set. For example, while access to utilities would be a key factor that affects housing prices, all but one house in the entire data set does has access to all public utilities, and hence this variable is omitted as there is not a sufficiently large enough sample size for meaningful analysis to be made. Other variables that are omitted for similar reasons include proximity to nearby locations, sale condition, and zoning classification of the sale.

Descriptive Analysis

(a) Provide a descriptive analysis of your variables. This should include histograms and fitted distributions, quantile plots, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.

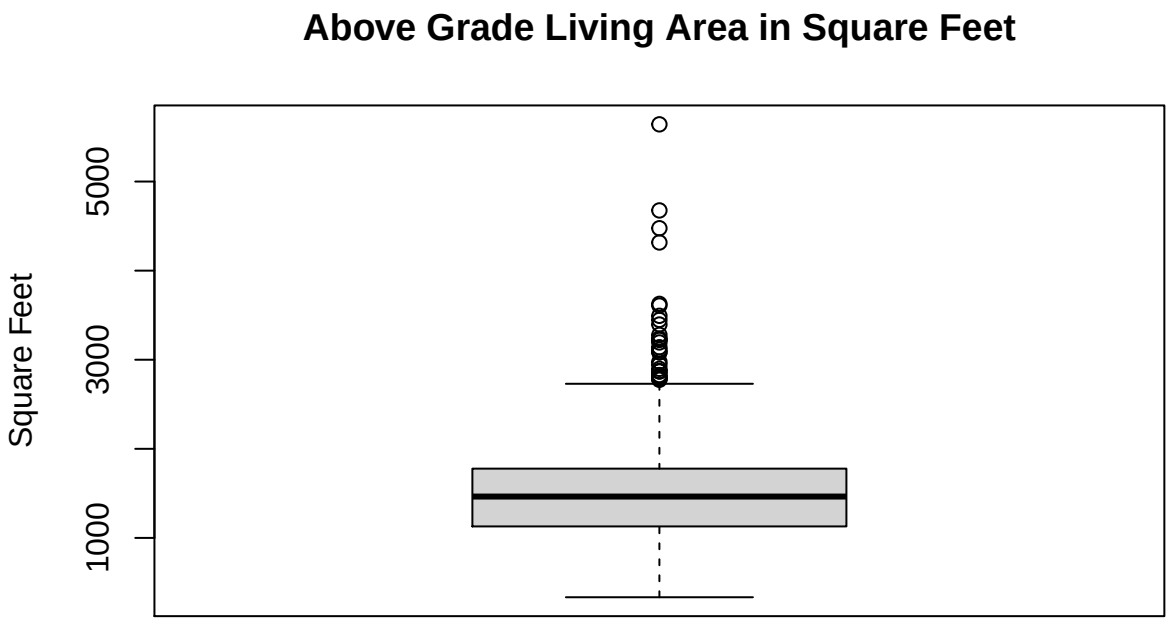
Var1. Above Grade (Ground) Living Area in Square Feet (GrLivArea)

```
hist(data$GrLivArea, main="Above Grade Living Area in Square Feet",
     xlab="Square Feet",ylab="Frequency")
```



A relatively normal distribution of above grade living area in square feet can be observed, with obvious outliers at >5000 square feet.

```
boxplot(data$GrLivArea, main="Above Grade Living Area in Square Feet",ylab="Square Feet")
```




```
summary<-c(mean(data$GrLivArea),sd(data$GrLivArea),min(data$GrLivArea),
           quantile(data$GrLivArea,.25),median(data$GrLivArea),
           quantile(data$GrLivArea,.75),max(data$GrLivArea),
           quantile(data$GrLivArea,.75)-quantile(data$GrLivArea,.25))
tablea<-matrix(summary,ncol=8)
colnames(tablea)<-c("Mean","Standard Deviation","Minimum", "25th Percentile",
                    "Median","75th Percentile","Maximum","IQR")
tablea
```

```
##          Mean Standard Deviation Minimum 25th Percentile Median 75th Percentile
## [1,] 1515.464           525.4804     334          1129.5   1464         1776.75
##      Maximum    IQR
## [1,]    5642 647.25
```

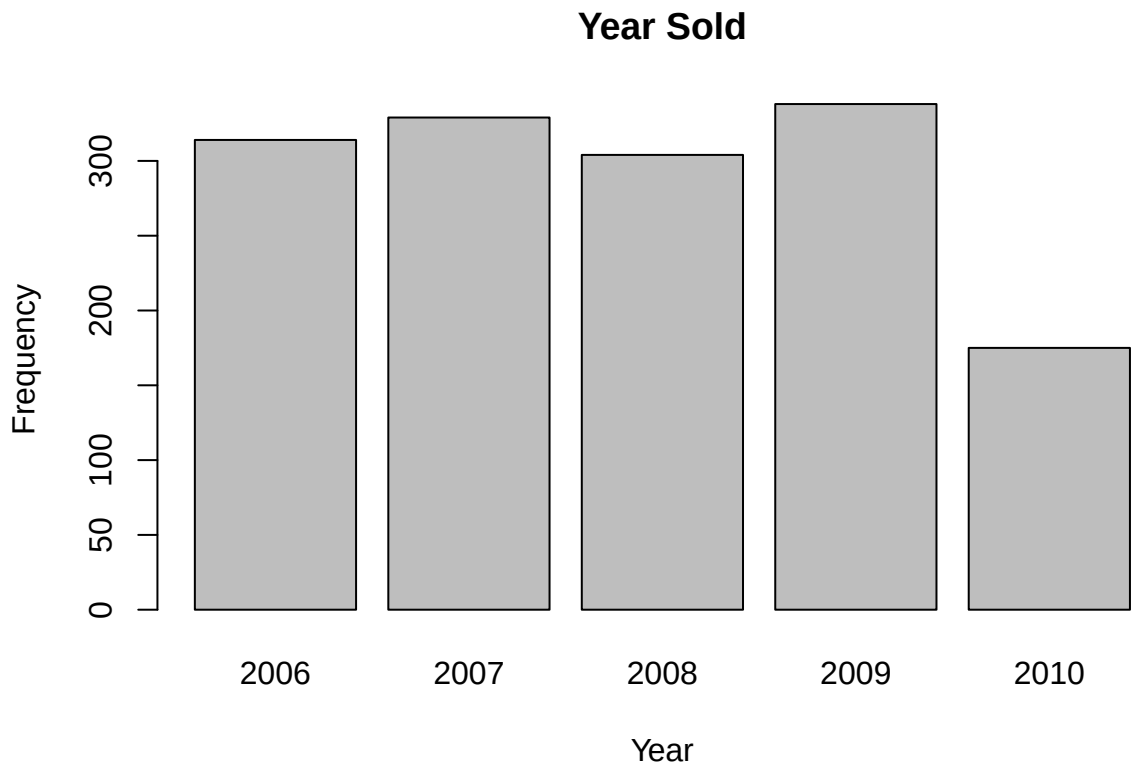
```
quantile(data$GrLivArea,.75)+1.5*(quantile(data$GrLivArea,.75)-
                                  quantile(data$GrLivArea,.25))
```

```
##      75%
## 2747.625
```

The median is relatively similar to the mean. In the boxplot, outliers above 2747.625 square feet can be more clearly observed and can potentially be removed.

Var2 Year Sold (YrSold) (2006-2010)

```
year<-table(data$YrSold)
barplot(year, main="Year Sold", xlab="Year", ylab="Frequency")
```



The distribution of sale year is relatively even in this given data set. The earliest sale year in this data set is 2006, and the latest sale year is 2010, as proven by the following function:

```
min(data$YrSold)
```

```
## [1] 2006
```

```
max(data$YrSold)
```

```
## [1] 2010
```

```
year
```

```
##
## 2006 2007 2008 2009 2010
##  314  329  304  338  175
```

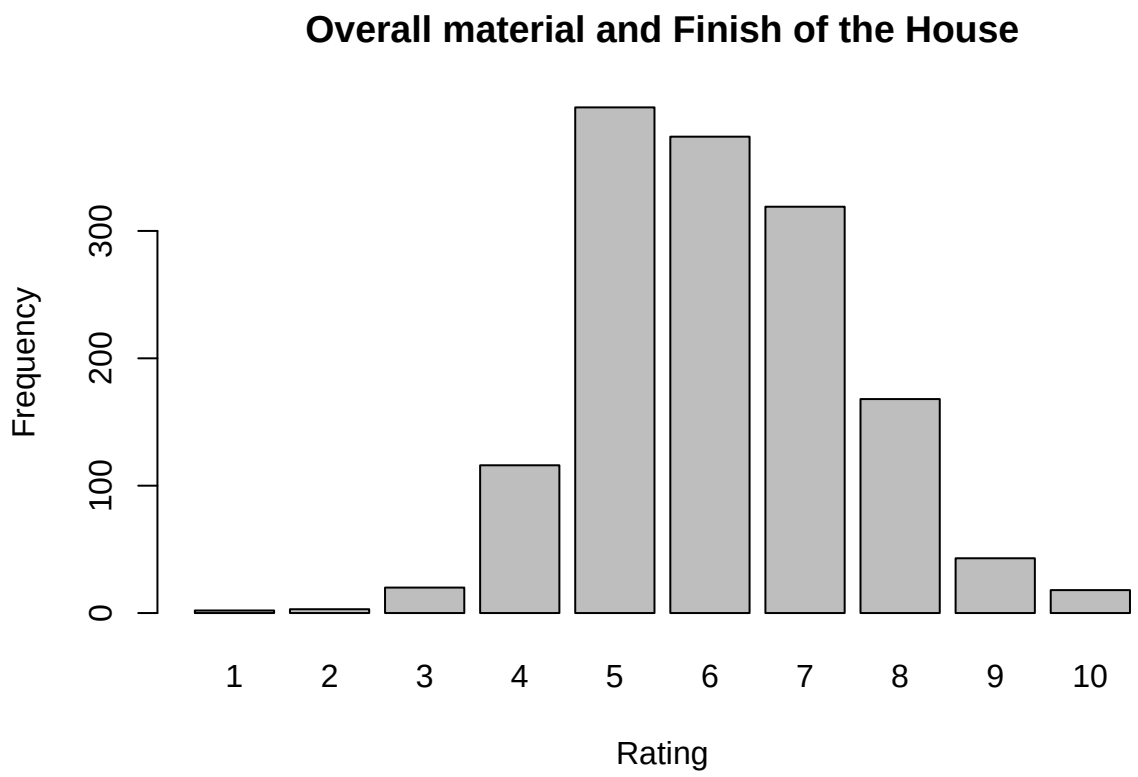
The actual numbers of the distribution of sale years are reported in the table above.

```r
data$y2006<-ifelse(data$YrSold==2006,1,0)
data$y2007<-ifelse(data$YrSold==2007,1,0)
data$y2008<-ifelse(data$YrSold==2008,1,0)
data$y2009<-ifelse(data$YrSold==2009,1,0)
data$y2010<-ifelse(data$YrSold==2010,1,0)
```

Since each year may have unique effects, e.g. exogeneous shocks, inflation etc., dummy variables are constructed for each year to be used in the regression, omitting one to avoid multicollinearity.

Var3 Overall Material and Finish of the House (OverallQual)

```
qual<-table(data$OverallQual)
barplot(qual,main="Overall material and Finish of the House",xlab="Rating",ylab="Frequency")
```



A somewhat normal distribution of the quality of the overall material and finish of the house can be observed in the histogram above.

```
qual
```

```
##
##   1   2   3   4   5   6   7   8   9  10
##   2   3  20 116 397 374 319 168  43  18
```

The actual numbers of frequency of each corresponding rating is reported in the table above. There is a disproportionately low number of low frequency ratings in the 1 and 2 range, which could potentially be a cause for concern due to the small sample size.

```
data$highqual<-ifelse(data$OverallQual>5,1,0)
sum(data$highqual)
```

```
## [1] 922
```

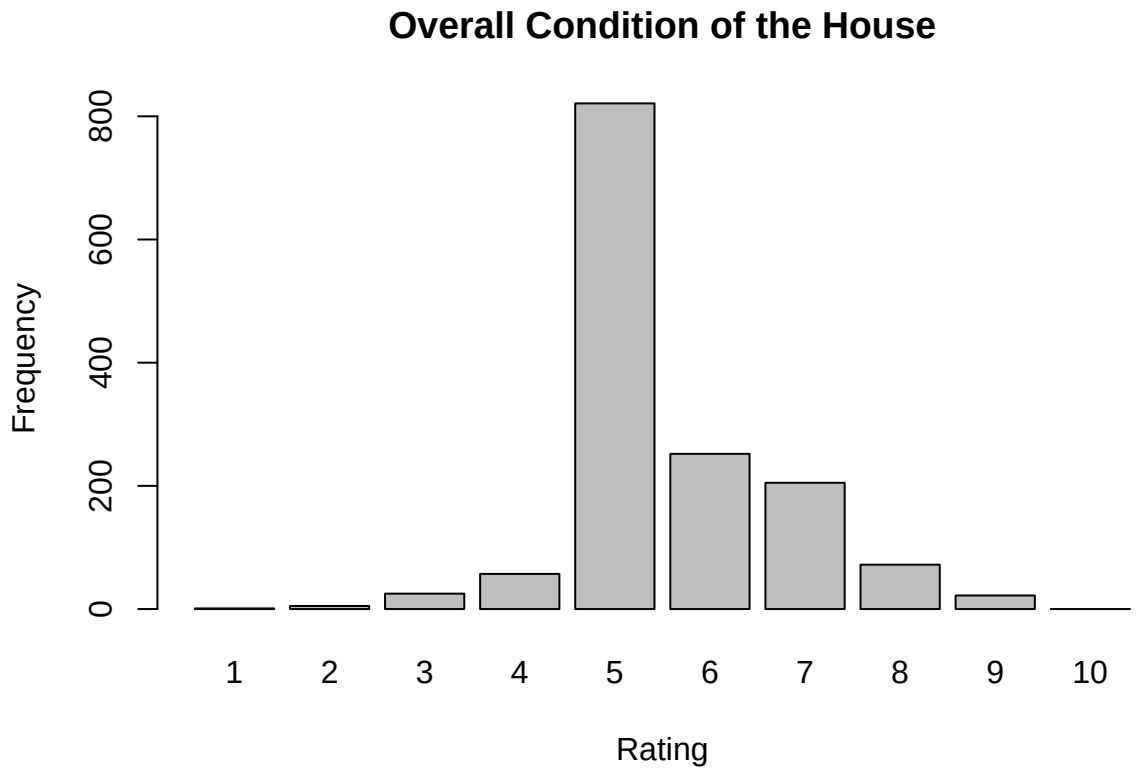
```r
1460-sum(data$highqual)
```

```
## [1] 538
```

To mitigate the problem of a small sample set, a new dummy variable is constructed that equals to 1 if ratings of overall material quality of the house is above 5, and zero otherwise. This differentiates between above average ratings of quality and below average ratings while maintaining a sufficiently size sample set between the two outcomes for meaningful analysis.

Var4 Overall Condition of the House (OverallCond):

```
cond<-table(factor(data$OverallCond, levels = 1:10))
barplot(cond,main="Overall Condition of the House",xlab="Rating",ylab="Frequency")
```



The distribution of the overall condition of the house is skewed towards the middle at a rating of around 4 to 5.

```
cond
```

```
##
##   1   2   3   4   5   6   7   8   9  10
##   1   5  25  57 821 252 205  72  22   0
```

The frequency of each corresponding rating of the overall condition of the house is reported in the table above. There is no data for the rating = 10 in the dataset, and only a very small number of houses that have a rating <5 in terms of condition, which again is a cause for concern.

```
data$highcond<-ifelse(data$OverallCond>5,1,0)
sum(data$highcond)
```

```
## [1] 551
```

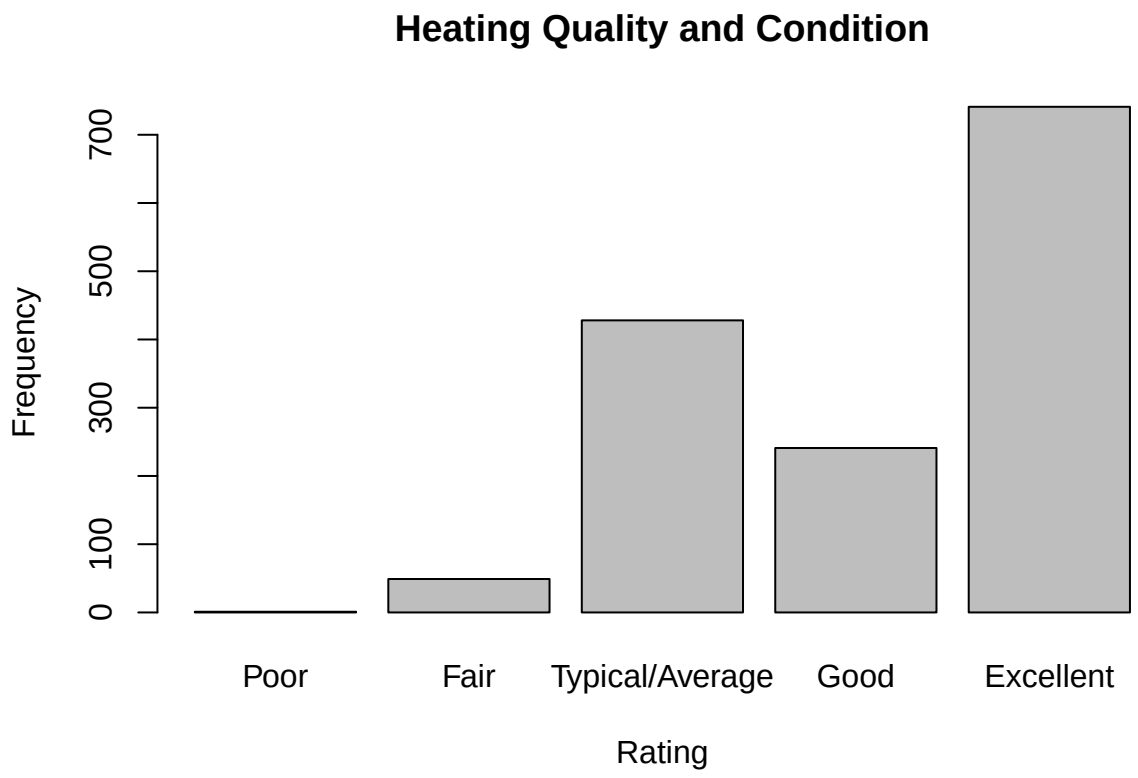
```r
1460-sum(data$highcond)
```

```
## [1] 909
```

To mitigate this problem, similar to the quality variable, a dummy variable for high ratings of overall condition is set that equals to 1 for ratings above 5, and 0 for ratings below or equal to 5. This differentiates between above average ratings of house condtion and below average ratings while maintaining a sufficiently size sample set between the two outcomes for meaningful analysis.

Var5 Heating quality and condition (HeatingQC):

```
heat<-table(data$HeatingQC)
heat2<-heat[c(4,2,5,3,1)]
barplot(heat2, names.arg=c("Poor","Fair","Typical/Average","Good","Excellent"),
        main="Heating Quality and Condition",xlab="Rating",ylab="Frequency")
```



The above histogram shows the distribution of reported heating quality in ascending order. As seen, the distribution is skewed towards the right of the diagram, meaning that more people report excellent heating quality. There is also a very small amount of houses that report poor heating quality.

```
heat2
```

```
##
##  Po  Fa  TA  Gd  Ex
##   1  49 428 241 741
```

Given the relatively smaller scale of only 5 variables and how the distribution is heavily skewed towards the higher ratings, a dummy variable separating between bottom three and upper two ratings can be considered that represents the divide between below average and above average heating quality.

```
data$hheat<-ifelse(data$HeatingQC=="Ex"|data$HeatingQC=="Gd",1,0)
sum(data$hheat)
```

```
## [1] 982
```

```r
1460-sum(data$hheat)
```

```
## [1] 478
```

Var6 Kitchen Quality (KitchenQual)

```
data$KEx<-ifelse(data$KitchenQual=="Ex",1,0)
data$KGd<-ifelse(data$KitchenQual=="Gd",1,0)
data$KTA<-ifelse(data$KitchenQual=="TA",1,0)
data$KFa<-ifelse(data$KitchenQual=="Fa",1,0)
data$KPo<-ifelse(data$KitchenQual=="Po",1,0)
kitchen<-data.frame(sum(data$KEx),sum(data$KGd),sum(data$KTA),sum(data$KFa),sum(data$KPo))
data$tablekit<-ifelse(data$KEx==1,5,ifelse(data$KGd==1,4,ifelse(data$KTA==1,3,
                          ifelse(data$KFa==1,2,ifelse(data$KPo==1,1,0)))))
ktable<-table(factor(data$tablekit,levels=1:5))
barplot(ktable,names.arg=c("Poor","Fair","Average/Typical","Good","Excellent"),
        main="Kitchen Quality",xlab="Rating",ylab="Frequency")
```

The above histogram shows the distribution of reported kitchen quality, with numbers 1 to 5 corresponding to ratings of Poor, Fair, Typical/Average, Good and Excellent respectively. As seen, no house in the dataset has reported a kitchen quality rating of 1, with the dataset also skewing towards the middle of the graph.

```
kitchen
```

```
##   sum.data.KEx. sum.data.KGd. sum.data.KTA. sum.data.KFa. sum.data.KPo.
## 1           100           586           735            39             0
```

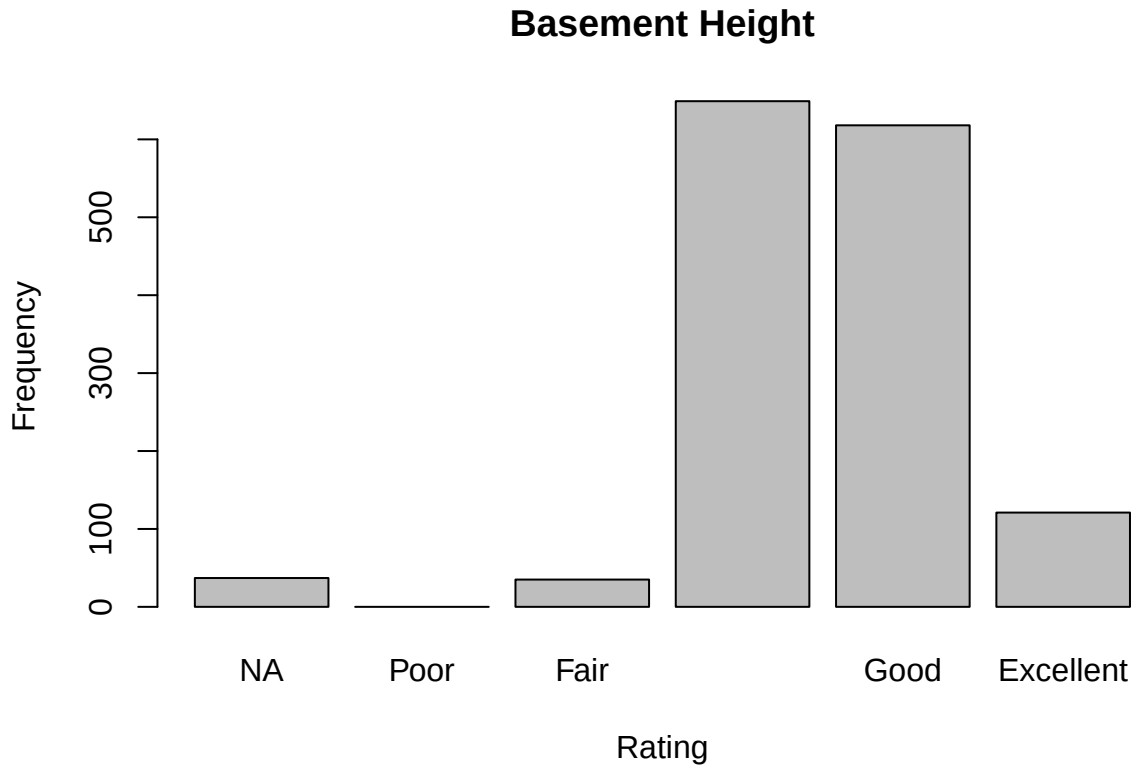
The above table reports the actual values of the distribution of kitchen quality in the dataset. Again, given the uneven distribution and the size of skew, a dummy variable seperating between below average and above average kitchen ratings can be explored as a potential variable for consideration.

```r
data$hkitchen<-ifelse(data$KitchenQual=="Ex"|data$KitchenQual=="Gd",1,0)
sum(data$hkitchen)
```

```
## [1] 686
```

Var7 Basement Height (BsmtQual)

```
data$BsmtQual[is.na(data$BsmtQual)] <- "NA"
data$BEx<-ifelse(data$BsmtQual=="Ex",1,0)
data$BGd<-ifelse(data$BsmtQual=="Gd",1,0)
data$BTA<-ifelse(data$BsmtQual=="TA",1,0)
data$BFa<-ifelse(data$BsmtQual=="Fa",1,0)
data$BPo<-ifelse(data$BsmtQual=="Po",1,0)
data$BNA<-ifelse(data$BsmtQual=="NA",1,0)
bqual<-ifelse(data$BEx==1,5,ifelse(data$BGd==1,4,ifelse(data$BTA==1,3,
      ifelse(data$BFa==1,2,ifelse(data$BPo==1,1,ifelse(data$BNA==1,0,0))))))
basement<-table(factor(bqual,levels=0:5))
barplot(basement,names.arg=c("NA","Poor","Fair","Average/Typical","Good","Excellent"),
        main="Basement Height",xlab="Rating",ylab="Frequency")
```



The data is skewed towards the middle of the distribution. However, there are no recorded data of houses with poor basement height in the dataset.3 dummy variables separating houses with above average quality, below average quality, and no basement can be used to mitigate this problem.

```
basement
```

```
##
##   0   1   2   3   4   5
##  37   0  35 649 618 121
```

```r
data$lbasement<-ifelse(data$BsmtQual=="Po"|data$BsmtQual=="Fa"|data$BsmtQual=="TA",1,0)
data$hbasement<-ifelse(data$BsmtQual=="Gd"|data$BsmtQual=="Ex",1,0)
sum(data$lbasement)
```

```
## [1] 684
```

```r
sum(data$hbasement)
```

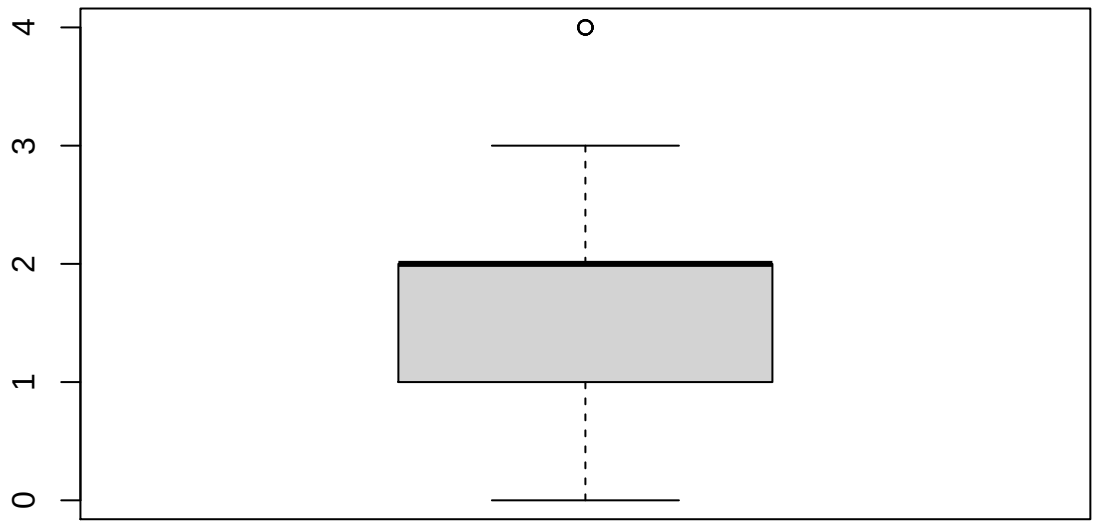
```
## [1] 739
```

```r
sum(data$BNA)
```

```
## [1] 37
```

Var8 Garage Capacity (GarageCars)

```
boxplot(data$GarageCars)
```



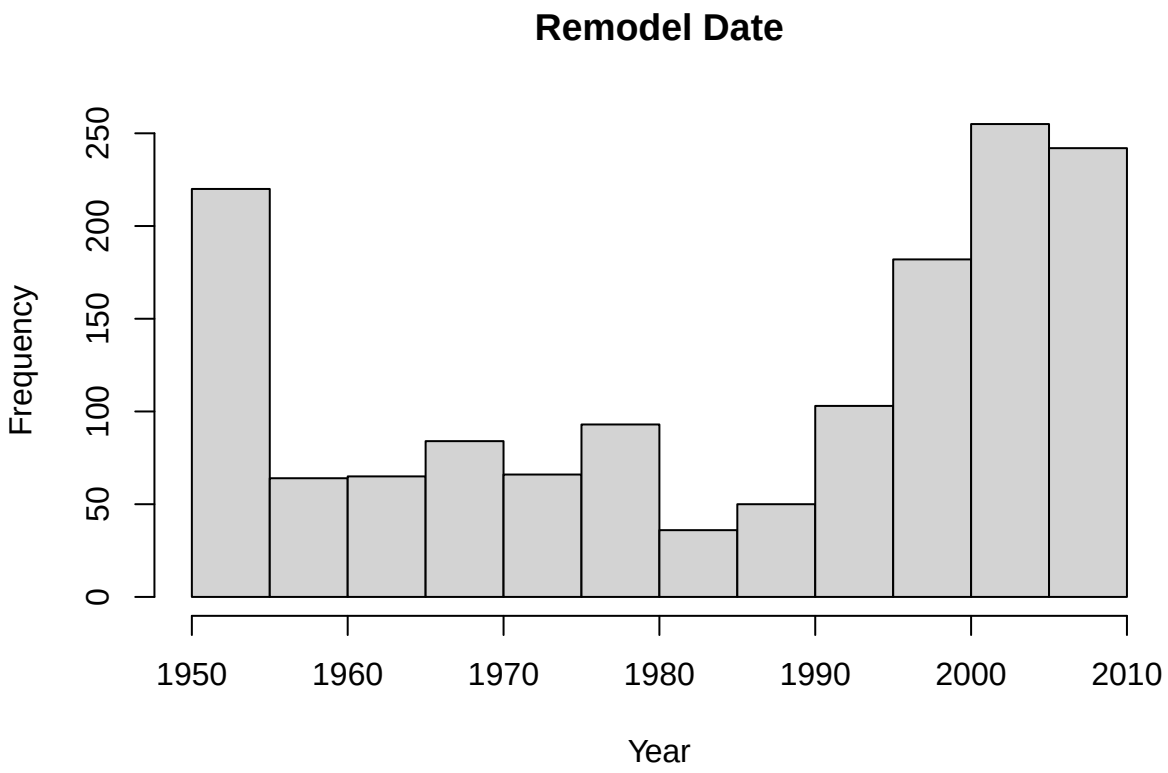
The median value of garage capacity in houses in the dataset is 2, with only one recorded outlier in the dataset of an individual that has a garage capacity of 4 cars as indicated in the boxplot.

```
summary<-c(mean(data$GarageCars),sd(data$GarageCars),min(data$GarageCars),
           quantile(data$GarageCars,.25),median(data$GarageCars),
           quantile(data$GarageCars,.75),max(data$GarageCars),
           quantile(data$GarageCars,.75)-quantile(data$GarageCars,.25))
table<-matrix(summary,ncol=8)
colnames(table)<-c("Mean","Standard Deviation","Minimum",
                   "25th Percentile","Median","75th Percentile","Maximum","IQR")
table
```

```
##          Mean Standard Deviation Minimum 25th Percentile Median 75th Percentile
## [1,] 1.767123           0.747315       0               1      2               2
##      Maximum IQR
## [1,]       4   1
```

Var9 Remodel Date (YearRemodAdd)

```
hist(data$YearRemodAdd,main="Remodel Date",xlab="Year")
```



There is a wide range of remodel dates in this dataset, ranging from 1950 to 2010.

```
min(data$YearRemodAdd)
```

```
## [1] 1950
```

```
max(data$YearRemodAdd)
```

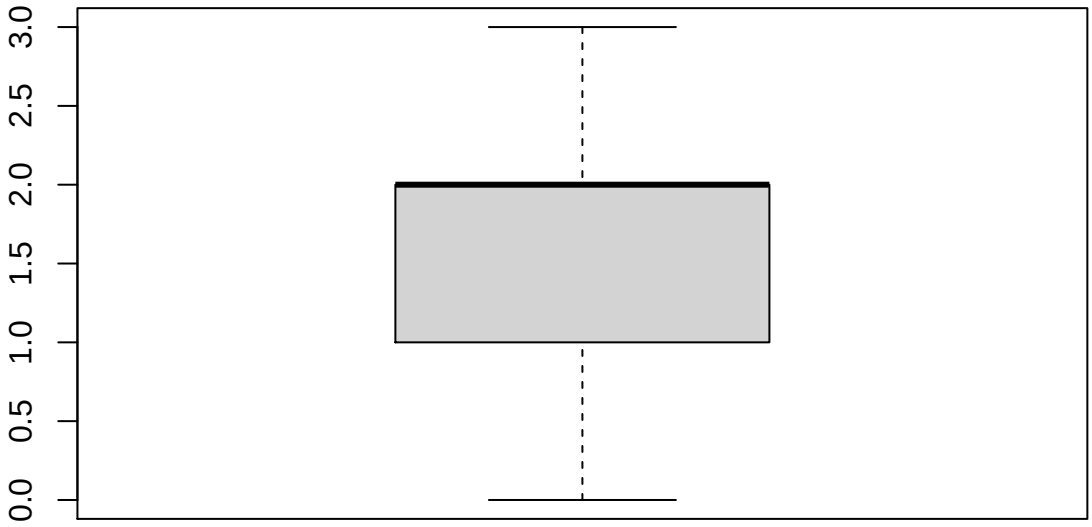
```
## [1] 2010
```

```
data$fittedyremod<-data$YearRemodAdd-1950
```

We can normalize the remodel date as seen in the equation above, where an increase in this new variable represents the difference between the house's remodel date and the earliest remodel date in the dataset - 1950. e.g. This variable equals 0 if the house is remodeled at 1950, 1 if the house is remodeled at 1951 etc.

Var10 Full Bathrooms Above Grade

```
boxplot(data$FullBath)
```



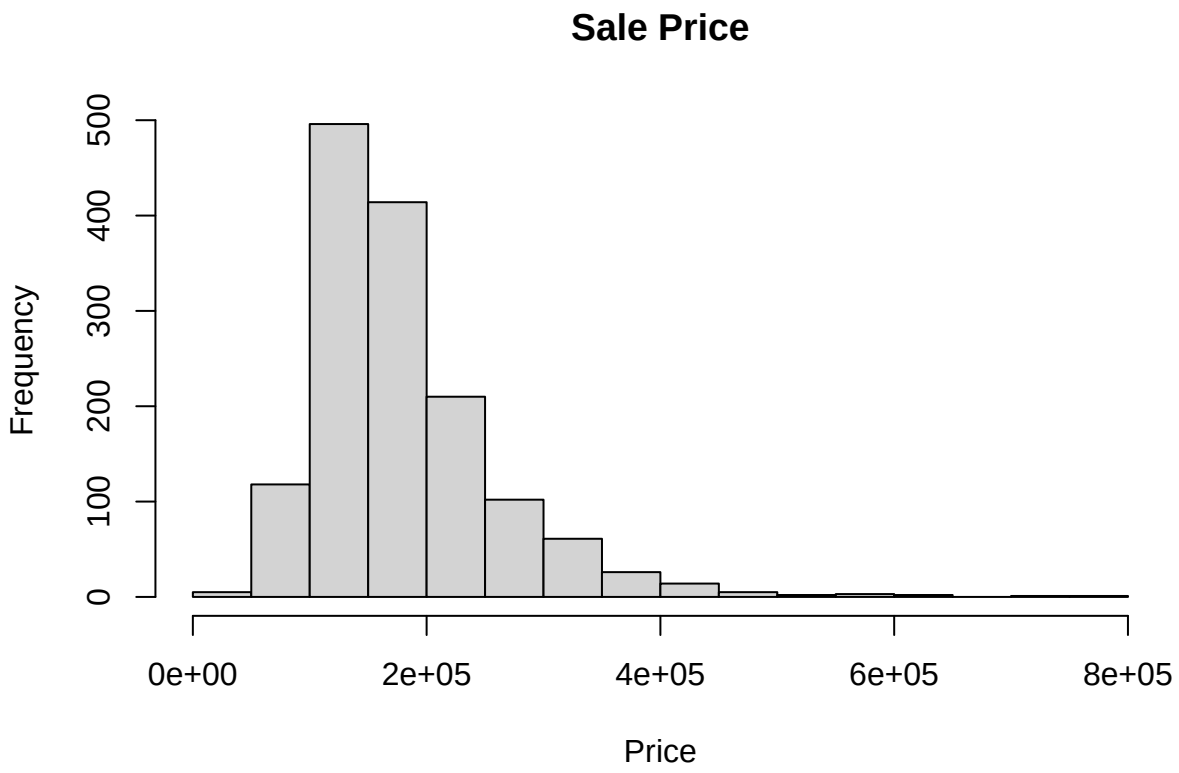
The median number of above ground full bathrooms in the dataset is 2 bathrooms, with minimum and maximum values of 0 and 3 respectively.

```
summary<-c(mean(data$FullBath),sd(data$FullBath),min(data$FullBath),
           quantile(data$FullBath,.25),median(data$FullBath),
           quantile(data$FullBath,.75),max(data$FullBath),
           quantile(data$FullBath,.75)-quantile(data$FullBath,.25))
table<-matrix(summary,ncol=8)
colnames(table)<-c("Mean","Standard Deviation","Minimum",
                   "25th Percentile","Median","75th Percentile","Maximum","IQR")
table
```

```
##          Mean Standard Deviation Minimum 25th Percentile Median 75th Percentile
## [1,] 1.565068          0.5509158       0               1      2               2
##      Maximum IQR
## [1,]       3   1
```

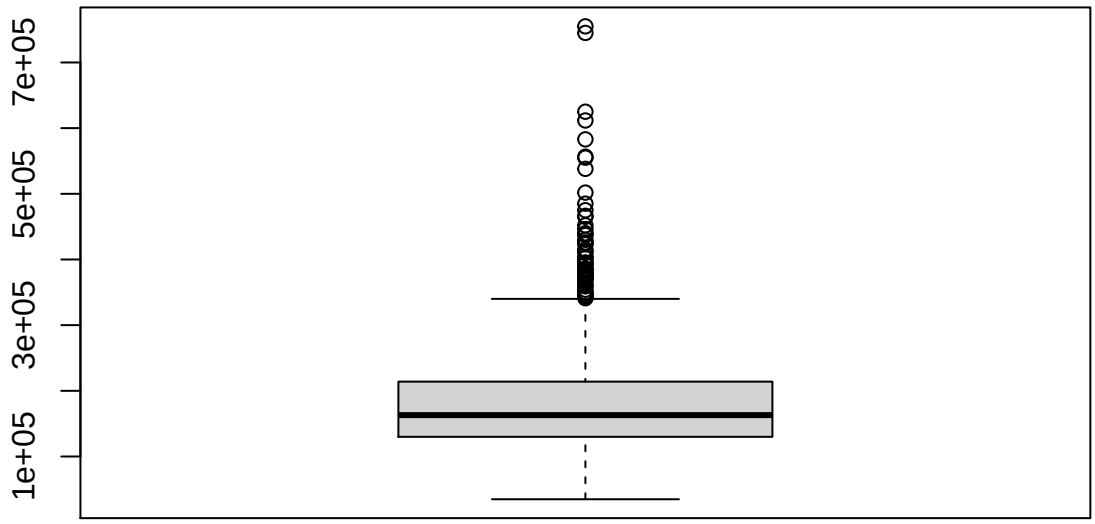
Y Variable Sale Price

```
hist(data$SalePrice,main="Sale Price",xlab="Price")
```



The histogram of the sale price indicates a relatively normal distribution of sale price. However, there seem to be outliers in the dataset towards the far right of the histogram.

```
boxplot(data$SalePrice)
```




```
quantile(data$SalePrice,.75)+1.5*(quantile(data$SalePrice,.75)-
                                  quantile(data$SalePrice,.25))
```

```
##      75%
## 340037.5
```

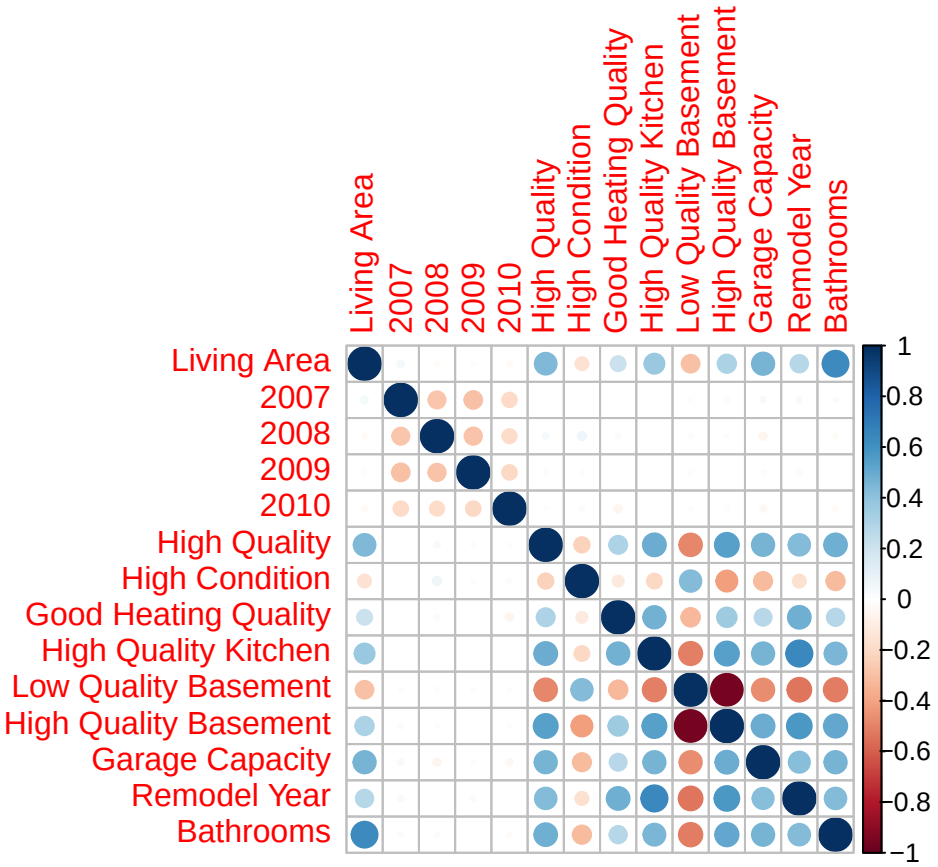
The outliers can be seen more clearly in the boxplot above, with outliers above $340037.5.

```
summary<-c(mean(data$SalePrice),sd(data$SalePrice),min(data$SalePrice),
           quantile(data$SalePrice,.25),median(data$SalePrice),
           quantile(data$SalePrice,.75),max(data$SalePrice),
           quantile(data$SalePrice,.75)-quantile(data$SalePrice,.25))
tabley<-matrix(summary,ncol=8)
colnames(tabley)<-c("Mean","Standard Deviation","Minimum", "25th Percentile",
                    "Median","75th Percentile","Maximum","IQR")
tabley
```

```
##          Mean Standard Deviation Minimum 25th Percentile Median 75th Percentile
## [1,] 180921.2            79442.5   34900          129975 163000          214000
##      Maximum   IQR
## [1,]  755000 84025
```

Correlation Plot

```
datac<-matrix(c(data$GrLivArea,data$y2007,data$y2008,data$y2009,data$y2010,data$highqual,
      data$highcond,data$hheat,data$hkitchen,data$lbasement,data$hbasement,data$GarageCars,
      data$fittedyremod,data$FullBath),nrow=1460)
colnames(datac)<-c("Living Area","2007","2008","2009","2010","High Quality",
                   "High Condition","Good Heating Quality","High Quality Kitchen",
                   "Low Quality Basement","High Quality Basement","Garage Capacity",
                   "Remodel Year","Bathrooms")
corrplot(cor(datac))
```



The correlation between most variables are not a cause for concern. However, the absolute value of the correlation between low quality and high quality basements seem to be particularly high. This may be due to the small number of houses that reported to not have a basement. A potential solution for this would be to omit houses with no basements from the dataset and effectively treating them as outliers, then removing the dummy variable for low quality basements from the regression to avoid multicollinearity. But for now, I would keep the variables as it is just to see how the regression would work out if I don't omit them. (I will omit them after part d)

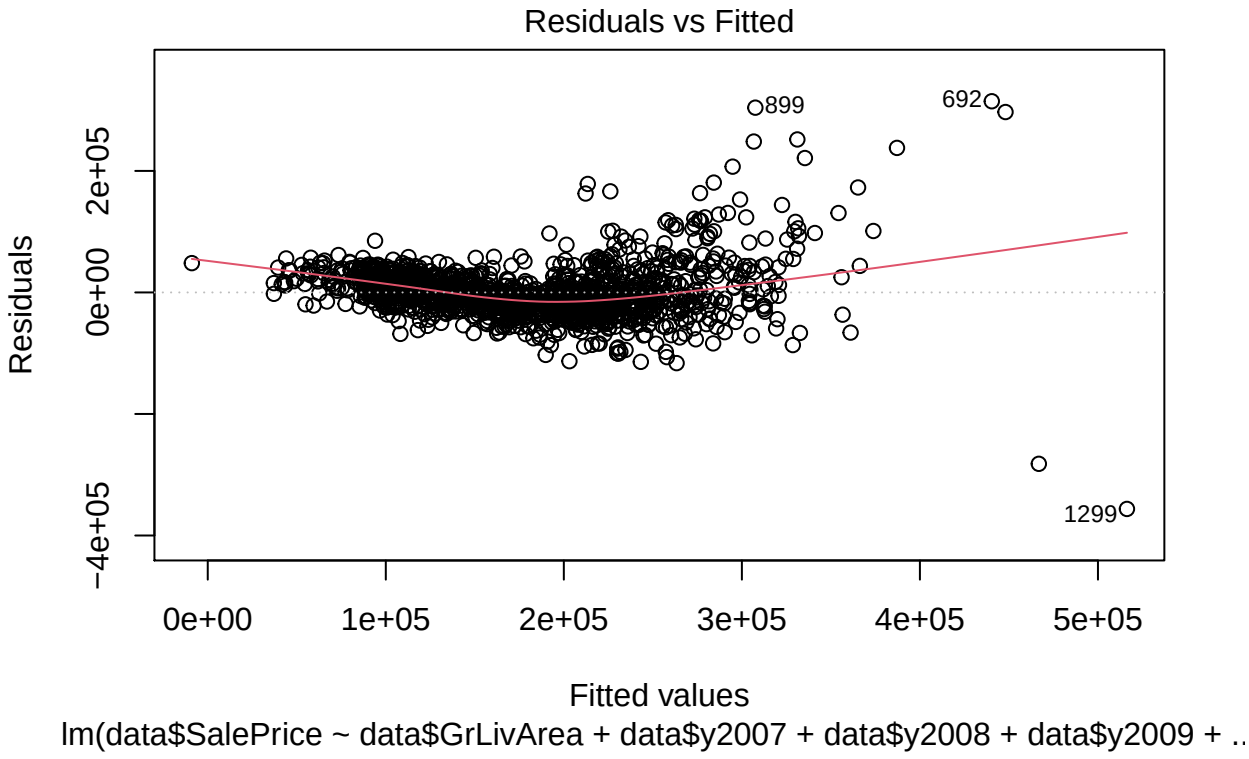
Transformation to Linearity

1b. For each variable (except indicator ones), test if a transformation to linearity is appropriate, and if so, apply the respective transformation, and comment on the transformed predictor(s).

Sale Price

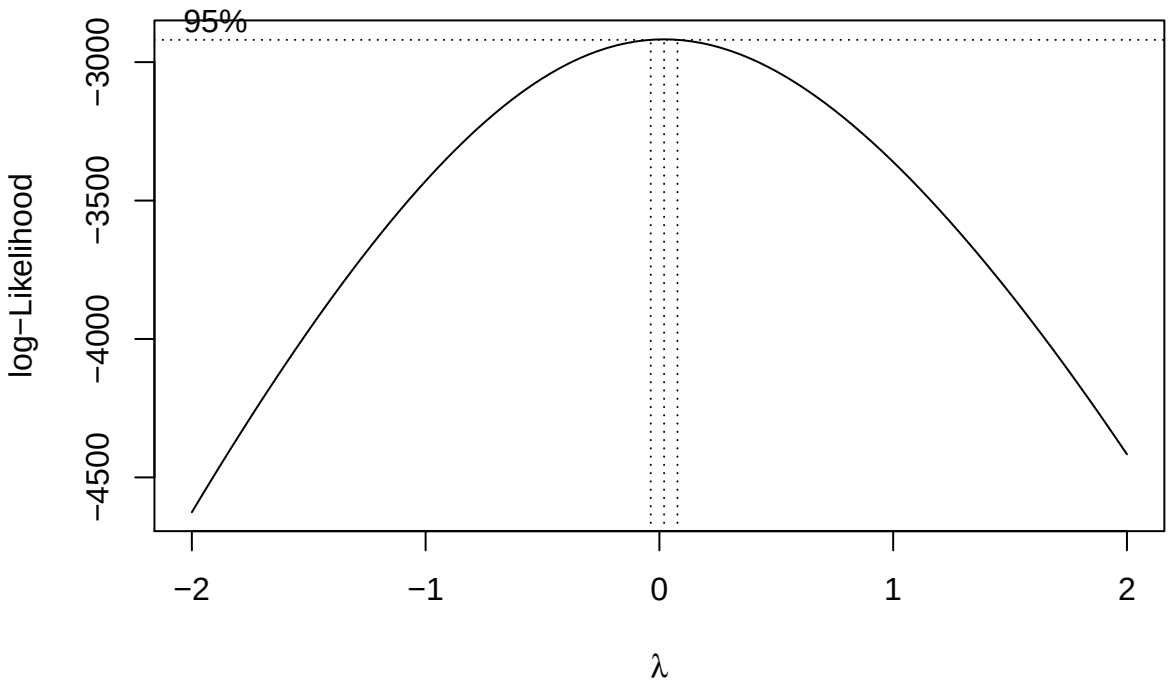
First let us consider transformations to the y variable. Consider a regression that includes all variables with no transformations.

```
reg1<-lm(data$SalePrice~data$GrLivArea+data$y2007+data$y2008+data$y2009+
           data$y2010+data$highqual+data$highcond+data$hheat+data$hkitchen+
           data$lbasement+data$hbasement+data$GarageCars+data$fittedyremod+
           data$FullBath)
plot(reg1,1)
```



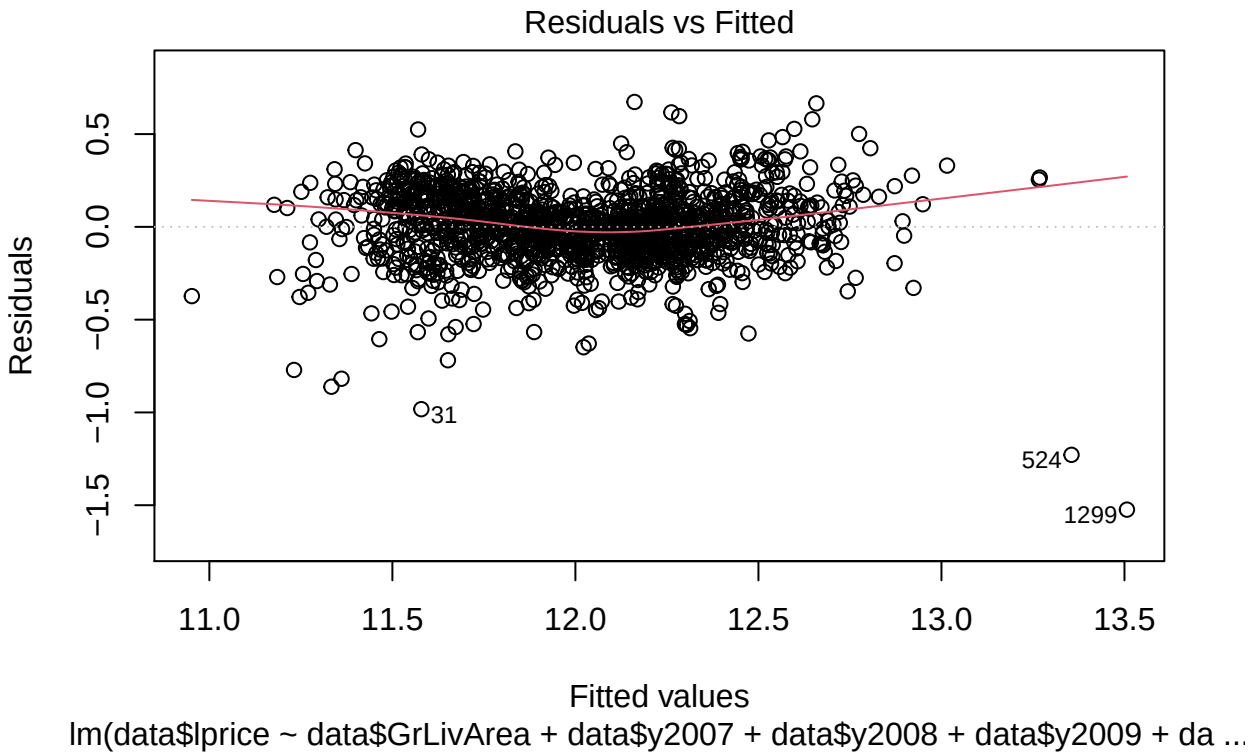
In this plot, we can see a large deviation in the residuals from zero.

```
boxcox(reg1)
```



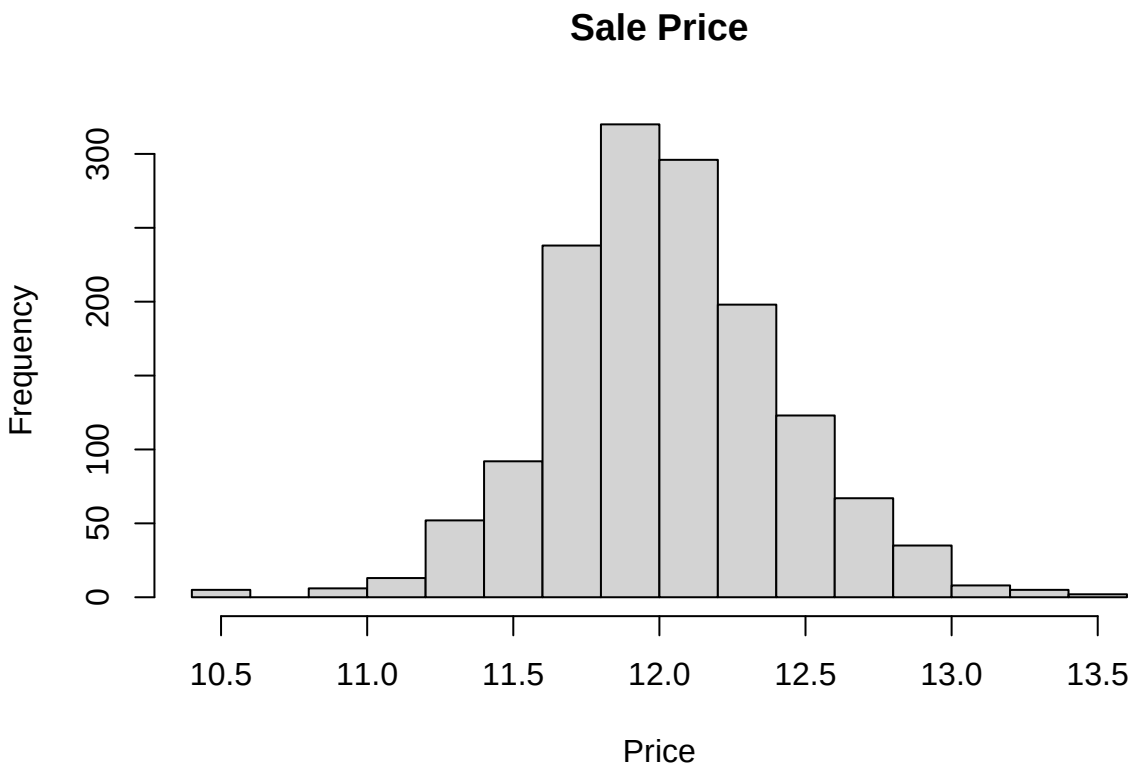
The Box-Cox plot of this regression predicts a landa value around 0, which suggests that the y varaible has a logarithmic relationship with the x variables in the regression. Therefore, we transform the y variable logarithmically.

```
data$lprice<-log(data$SalePrice)
reglp<-lm(data$lprice~data$GrLivArea+data$y2007+data$y2008+data$y2009+data$y2010+
            data$highqual+data$highcond+data$hheat+data$hkitchen+data$lbasement+
            data$hbasement+data$GarageCars+data$fittedyremod+data$FullBath)
plot(reglp,1)
```



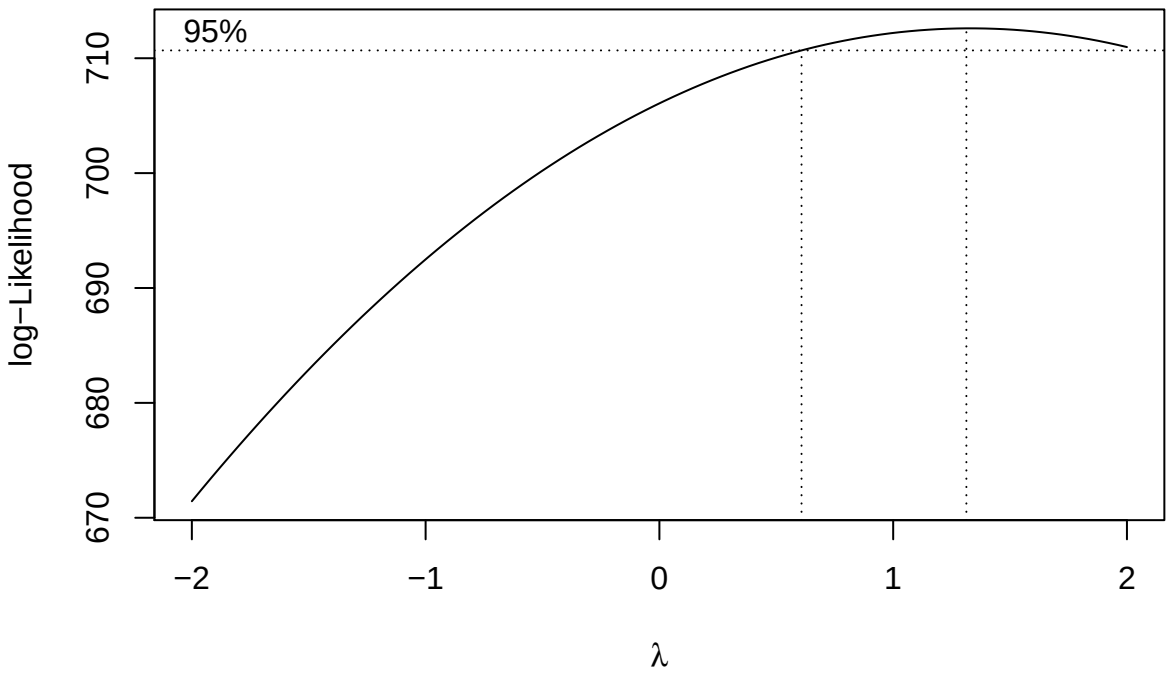
Here, the size of the residuals are much smaller and deviates much more closely around 0. The pattern of observations in the second diagram is also a lot more randomly distributed, suggesting that the relationship between the x and y variables are more linear. Hence a logarithmic transformation of the y variable is chosen.

```
hist(data$lprice,main="Sale Price",xlab="Price")
```



The histogram of log(saleprice) is also much more normally distributed, again justifying a logarithmic transformation.

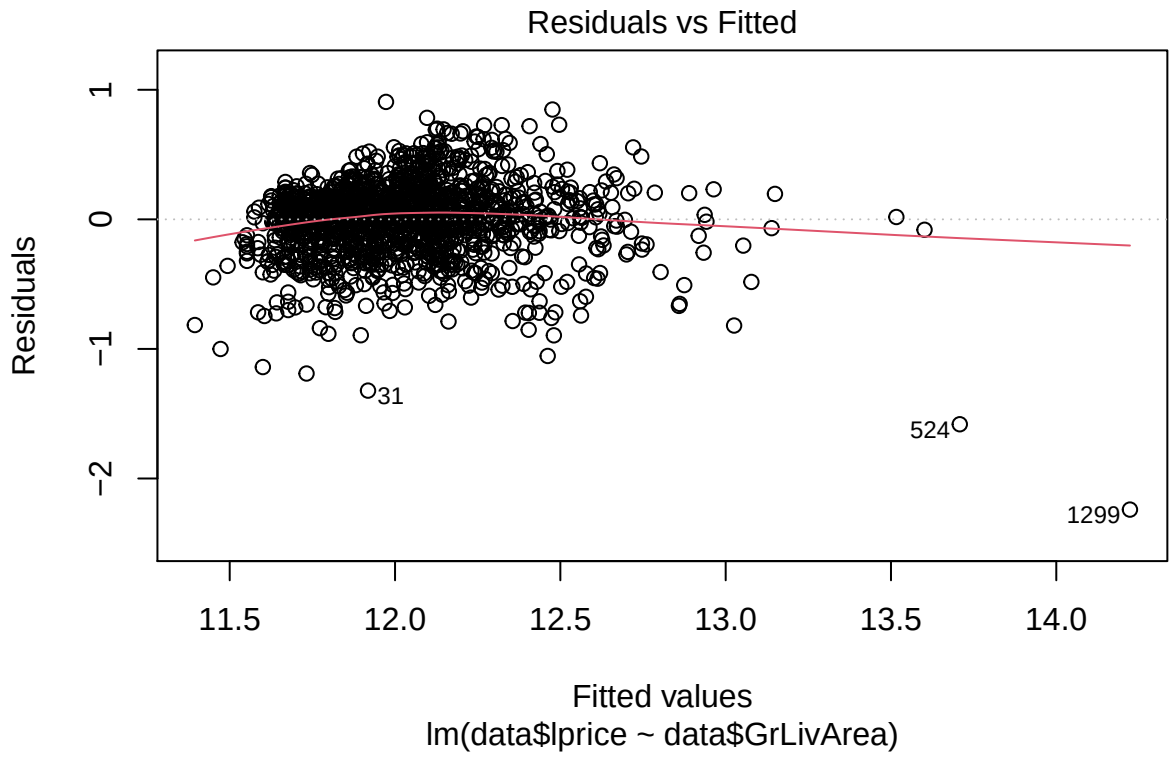
```
boxcox(reglp)
```



The Box Cox plot of this regression also has a landa value around the value of 1 after the y variable is logarithmically transformed, suggesting that linearity is improved.

Above Ground Living Area

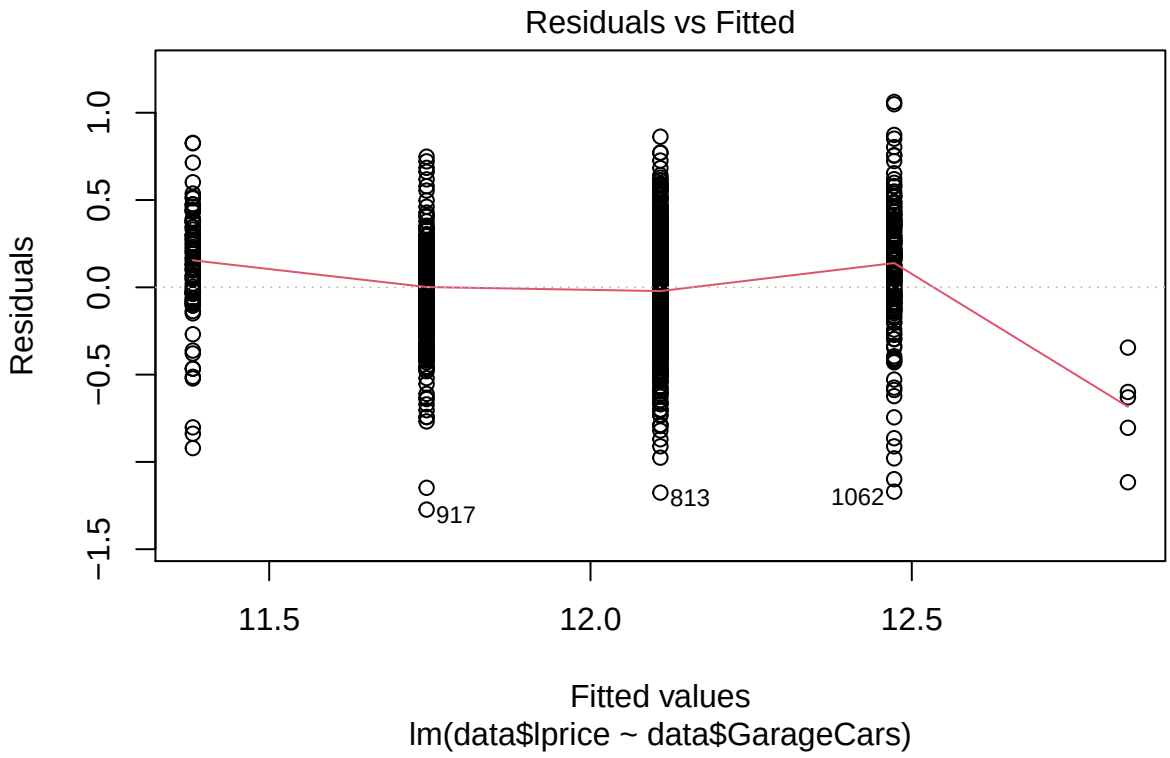
```
regla<-lm(data$lprice~data$GrLivArea)
plot(regla,1)
```



The above residual vs fitted plot shows a distribution of residuals that deviates close enough to 0. Different linear transformations of this variable did not result in a sizeable observable increase in deviation likelihood around 0. Both factors combined led me to conclude that no linear transformation is needed for this variable.

Garage Capacity

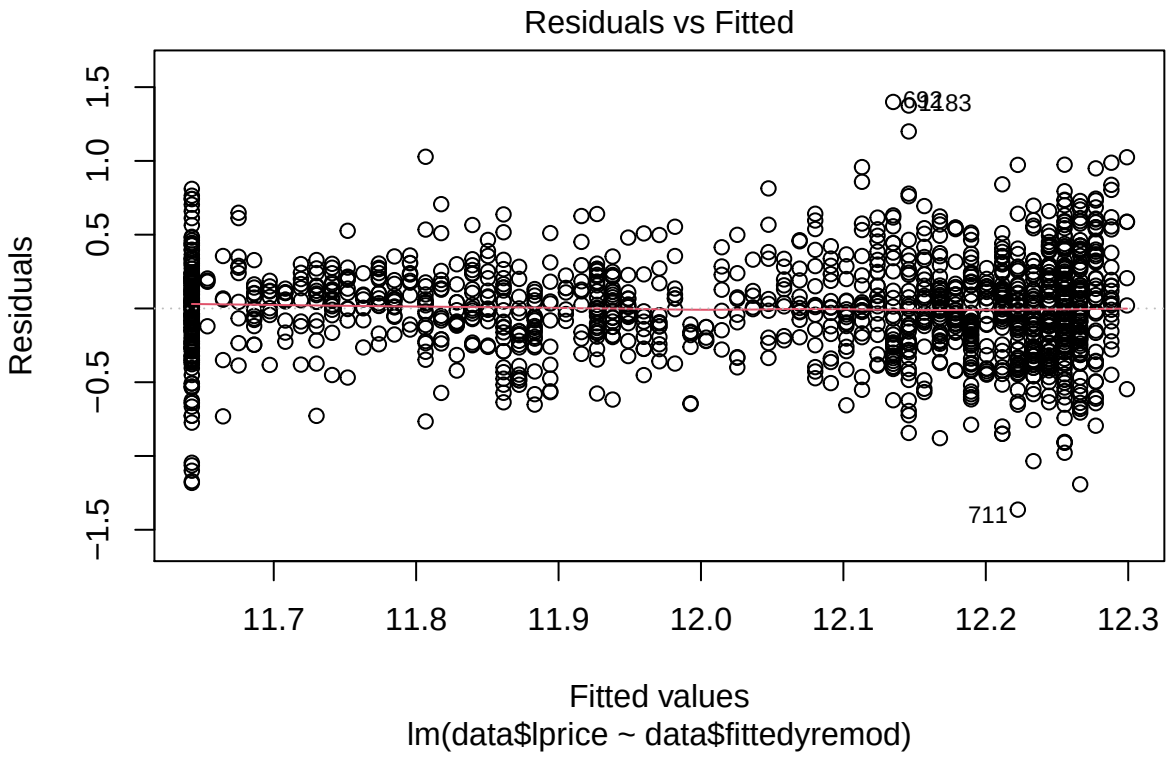
```
regcar<-lm(data$lprice~data$GarageCars)
plot(regcar,1)
```



Residuals are distributed randomly with no observable trend and deviates around 0. Hence no linear transformation is needed.

Remodel Year

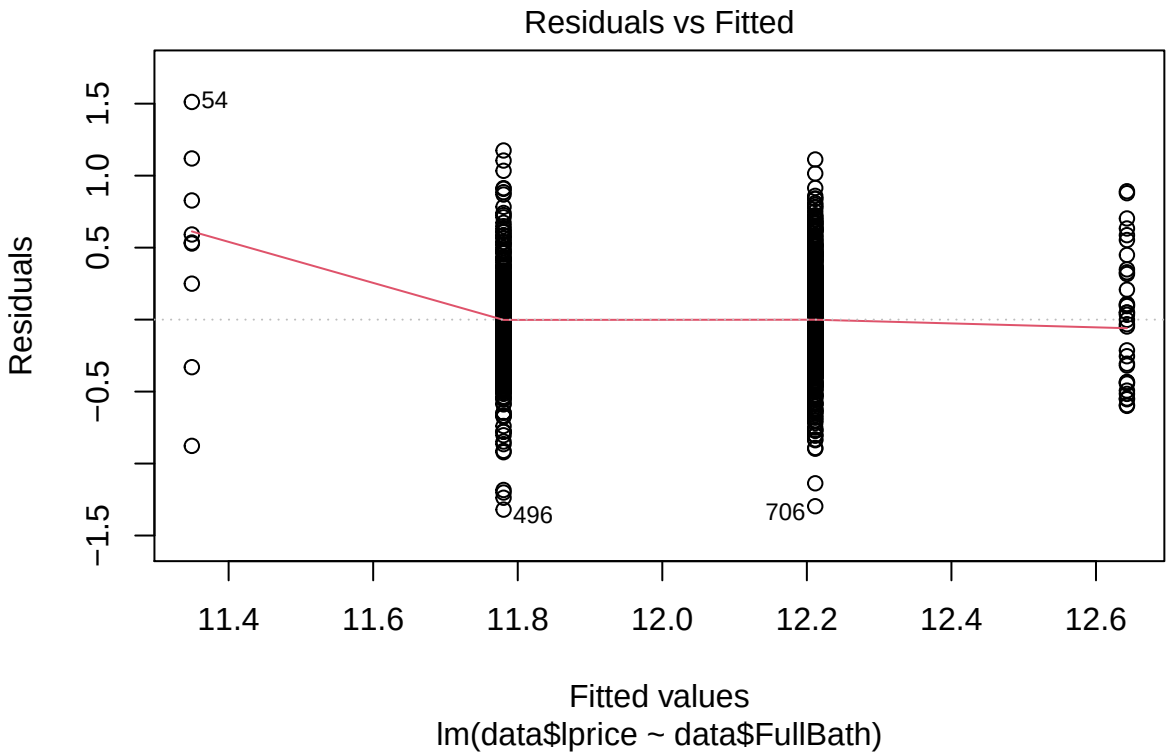
```
regyremod<-lm(data$lprice~data$fittedyremod)
plot(regyremod,1)
```



Residuals are distributed randomly with no observable trend and deviates around 0. Hence no linear transformation is needed.

Number of Bathrooms

```
regbath<-lm(data$lprice~data$FullBath)
plot(regbath,1)
```



Residuals are distributed randomly with no observable trend and deviates mostly around 0. The deviation above 0 on the left of the diagram may be due to a small sample size of houses in the dataset that has reported 0 above ground full bathrooms in the house, and cannot be resolved through linear transformations.

Regression Analysis

1c. Estimate a multiple linear regression model that includes all the main effects only (i.e., no interactions nor higher order terms). We will use this model as a baseline. Comment on the statistical and economic significance of your estimates. Also, make sure to provide an interpretation of your estimates.

Baseline

```
regb<-lm(data$lprice~data$GrLivArea+data$y2007+data$y2008+data$y2009+data$y2010+
           data$highqual+data$highcond+data$hheat+data$hkitchen+data$lbasement+
           data$hbasement+data$GarageCars+data$fittedyremod+data$FullBath)
summary(regb)
```

```
##
## Call:
## lm(formula = data$lprice ~ data$GrLivArea + data$y2007 + data$y2008 +
##     data$y2009 + data$y2010 + data$highqual + data$highcond +
##     data$hheat + data$hkitchen + data$lbasement + data$hbasement +
##     data$GarageCars + data$fittedyremod + data$FullBath)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52384 -0.10892  0.00311  0.11879  0.67317
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.086e+01  3.961e-02 274.222  < 2e-16 ***
## data$GrLivArea      3.066e-04  1.328e-05  23.086  < 2e-16 ***
## data$y2007         -9.280e-03  1.539e-02  -0.603  0.54657
## data$y2008         -7.860e-03  1.573e-02  -0.500  0.61747
## data$y2009         -2.345e-02  1.528e-02  -1.535  0.12511
## data$y2010         -5.300e-03  1.842e-02  -0.288  0.77355
## data$highqual       8.659e-02  1.402e-02   6.177 8.46e-10 ***
## data$highcond       5.376e-02  1.198e-02   4.486 7.83e-06 ***
## data$hheat          3.852e-02  1.284e-02   3.000  0.00275 **
## data$hkitchen       8.644e-02  1.483e-02   5.827 6.93e-09 ***
## data$lbasement      1.750e-01  3.387e-02   5.167 2.72e-07 ***
## data$hbasement      2.853e-01  3.536e-02   8.068 1.48e-15 ***
## data$GarageCars     1.492e-01  8.835e-03  16.885  < 2e-16 ***
## data$fittedyremod   2.308e-03  3.620e-04   6.376 2.44e-10 ***
## data$FullBath      -3.121e-03  1.372e-02  -0.228  0.82005
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1941 on 1445 degrees of freedom
## Multiple R-squared:  0.7662, Adjusted R-squared:  0.7639
## F-statistic: 338.2 on 14 and 1445 DF,  p-value: < 2.2e-16
```

Almost all variables chosen to be included in the baseline regression have shown a high level of significance, except for sale years and the number of above ground full bathrooms in the house. It is also important to note that the regression also has an R-squared value at 0.7639, which although is not perfect is decent.

Above Ground Living Area

This is a significant variable as shown in the baseline regression model at the 0.1% significant level. A unit

change in above ground living area in square feet of the house on average leads to an approximately 0.03066% increase in house sale price, holding other factors constant.

Sale Year
None of the sale year dummy variables are shown to be significant variables under any standard significance levels. Each coefficient of each respective sale year dummy represents the proportional difference in sale price compared to sale prices in 2006. For example, the coefficient in the year 2007 dummy variable indicates that sale prices in 2007 are approximately -0.928% cheaper than sale prices in 2006 on average, holding other factors constant. However as mentioned, none of these sale year variables are significant. This could potentially suggest the robustness of house prices to exogenous economic effects, such as the financial crisis in 2008. However, this interpretation is not conclusive, and more evidence is needed to solidify this claim.

Overall Quality of Material and Finish of House
This is a significant variable as shown in the baseline regression model at the 0.1% significant level. Houses that have reported an above average quality in material and finish are 8.659% more expensive on average than houses that have reported a below average quality in material and finish, holding other factors constant.

Overall Condition of the House
This is a significant variable as shown in the baseline regression model at the 0.1% significant level. Houses in above average conditions are approximately 5.376% more expensive than houses in below average conditions, holding other factors constant.

Heating Quality
This is a significant variable as shown in the baseline regression model at the 1% significant level. Houses that have above average heating quality are approximately 3.852% more expensive than houses that have below average heating quality, holding other factors constant.

Kitchen Quality
This is a significant variable as shown in the baseline regression model at the 0.1% significant level. Houses that have above average kitchen quality are approximately 8.644% more expensive than houses that have below average kitchen quality, holding other factors constant.

Basement Quality
Both basement dummy variables are significant at the 0.1% significant level. Houses that have a below average quality of basement in terms of height are approximately 17.5% more expensive than houses that don't have a basement, holding other factors constant. Houses that have a above average quality of basement in terms of height are approximately 28.53% more expensive than houses that don't have a basement, holding other factors constant.

Garage Capacity
This is a significant variable as shown in the baseline regression model at the 0.1% significant level. An increase in garage capacity per car will lead to a 14.92% increase in house price on average, holding other factors constant.

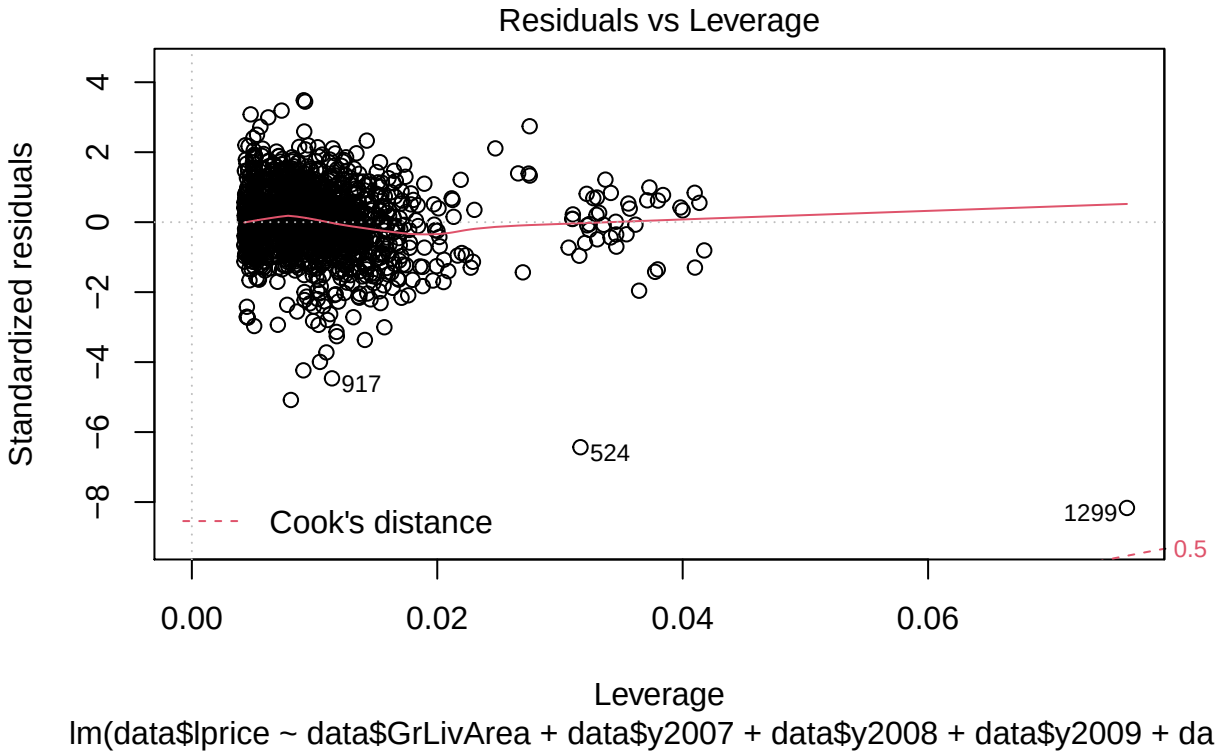
Remodelling Year
This is a significant variable as shown in the baseline regression model at the 0.1% significant level. An additional year in remodelling year of the house since 1950 will lead to an 0.2308% increase in house price on average per year, holding other factors constant.

Number of Bathrooms
This is not a significant variable as shown in the baseline model under any standard significance level. The estimated coefficient indicates that an additional above ground full bathroom in the house would lead to a 0.3121% decrease in house price on average, holding over factors constant. However, as mentioned, this coefficient is not significant.

Outlier Analysis

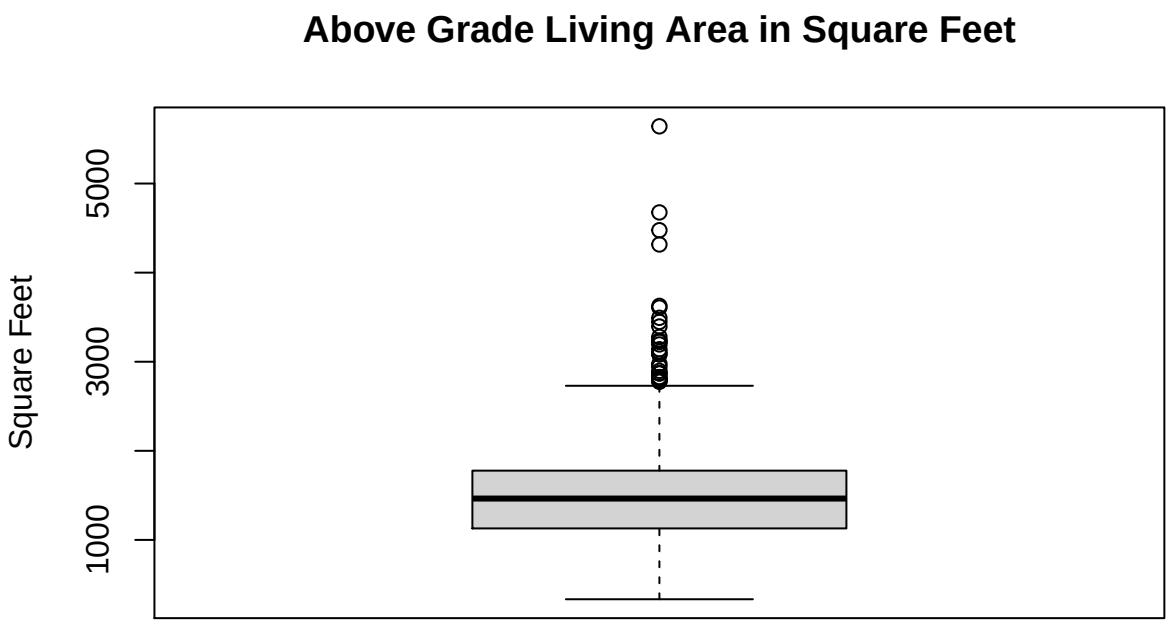
1d. In your model from part (c), identify if there are any outliers worth removing. If so, remove them but justify your reason for doing so and re-estimate your model from part (c)

```
plot(regb,5)
```



Consider the residuals vs leverage plot of the baseline regression. Here, there are no data points outside Cook's distance, suggesting that there are no overly influential outliers within the dataset. However, as mentioned in part a, since there is only a small sample set of houses that do not have a basement, the dummy variables for houses with a high quality basement and low quality basement may have issues of multicollinearity. To avoid this problem, we should treat houses that do not have basements as outlier variables and omit the dummy variable for houses with low quality basements in the new regression to avoid multicollinearity

```
boxplot(data$GrLivArea, main="Above Grade Living Area in Square Feet",
        ylab="Square Feet")
```




```
tablea
```

```
##          Mean Standard Deviation Minimum 25th Percentile Median 75th Percentile
## [1,] 1515.464           525.4804     334          1129.5   1464         1776.75
##      Maximum    IQR
## [1,]    5642 647.25
```

```
data$SalePrice[data$GrLivArea>4000]
```

```
## [1] 184750 755000 745000 160000
```

```
data$SalePrice[data$GrLivArea==4676]
```

```
## [1] 184750
```

```
data$SalePrice[data$GrLivArea==5642]
```

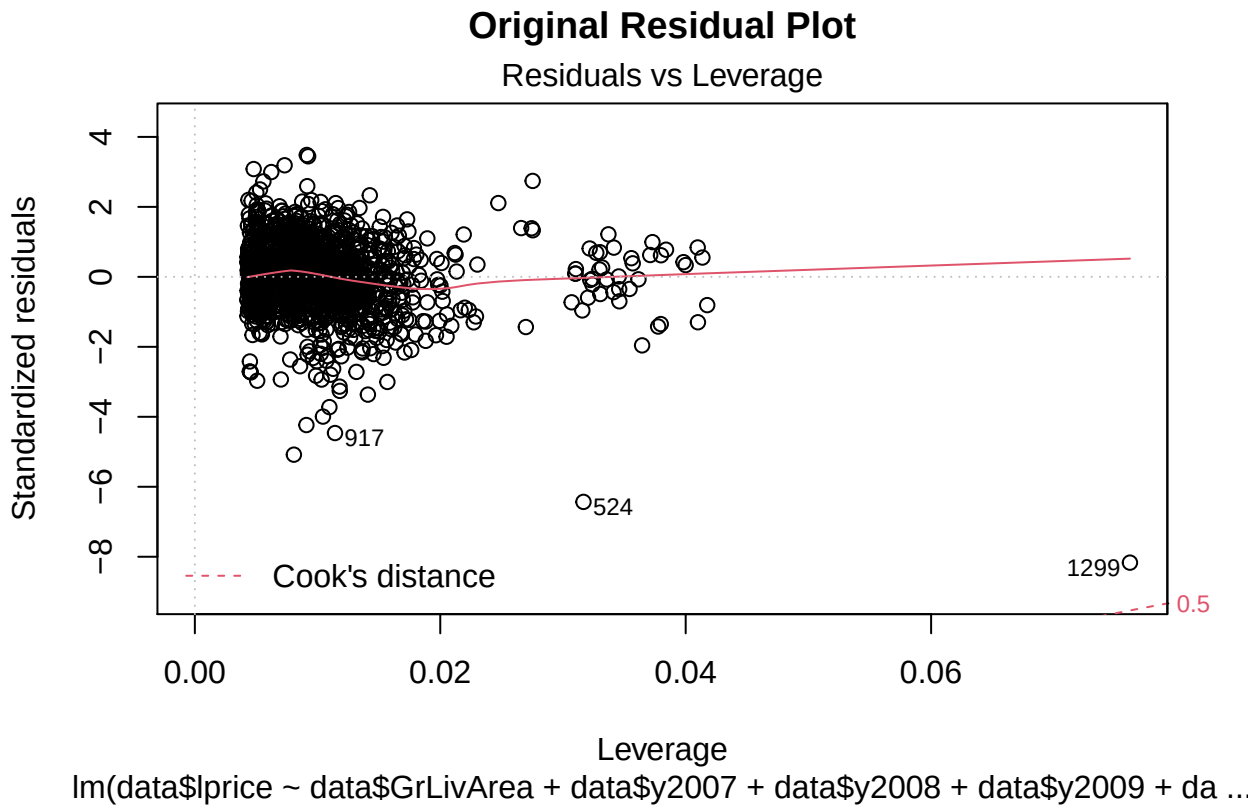
```
## [1] 160000
```

```
mean(data$SalePrice)
```

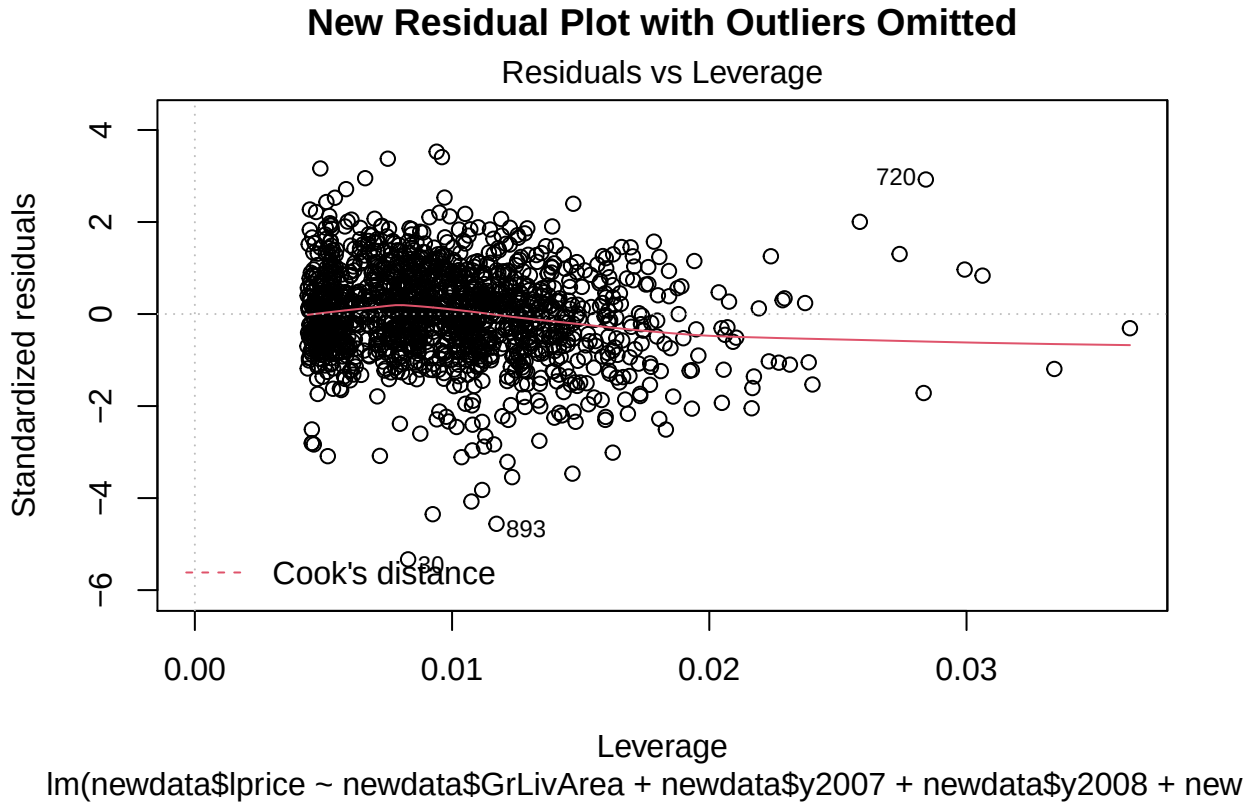
```
## [1] 180921.2
```

Furthermore, there are four clear outliers shown in the boxplot of above grade living area in square feet. When examining the corresponding house prices of the outliers, the first and fourth house with the biggest above grade living area in square feet has a sale price of just $160000 and $184750 respectively, only one of which is barely above the mean of sale price at 180921.2, the opposite trend illustrated in the baseline regression. Therefore these two outliers will also be dropped.

```
newdata<-subset(data, data$BNA==0&data$GrLivArea!=4676&data$GrLivArea!=5642)
newdata$y2006<-ifelse(newdata$YrSold==2006,1,0)
newdata$y2007<-ifelse(newdata$YrSold==2007,1,0)
newdata$y2008<-ifelse(newdata$YrSold==2008,1,0)
newdata$y2009<-ifelse(newdata$YrSold==2009,1,0)
newdata$y2010<-ifelse(newdata$YrSold==2010,1,0)
newdata$highqual<-ifelse(newdata$OverallQual>5,1,0)
newdata$highcond<-ifelse(newdata$OverallCond>5,1,0)
newdata$hheat<-ifelse(newdata$HeatingQC=="Ex"|newdata$HeatingQC=="Gd",1,0)
newdata$hkitchen<-ifelse(newdata$KitchenQual=="Ex"|newdata$KitchenQual=="Gd",1,0)
newdata$lbasement<-ifelse(newdata$BsmtQual=="Po"|newdata$BsmtQual=="Fa"|
                          newdata$BsmtQual=="TA",1,0)
newdata$hbasement<-ifelse(newdata$BsmtQual=="Gd"|newdata$BsmtQual=="Ex",1,0)
newdata$fittedyremod<-newdata$YearRemodAdd-1950
newdata$lprice<-log(newdata$SalePrice)
newregb<-lm(newdata$lprice~newdata$GrLivArea+newdata$y2007+newdata$y2008+
              newdata$y2009+newdata$y2010+newdata$highqual+newdata$highcond+
              newdata$hheat+newdata$hkitchen+newdata$hbasement+
              newdata$GarageCars+newdata$fittedyremod+newdata$FullBath)
plot(regb,5,main="Original Residual Plot")
```



```r
plot(newregb,5, main="New Residual Plot with Outliers Omitted")
```



Once the living area outliers are removed, the two residuals closest to Cook's distance shown in the original Residuals vs Fitted diagram is removed, hence justifying the omission of those two outliers in particular.

```r
summary(newregb)
```

```
##
## Call:
## lm(formula = newdata$lprice ~ newdata$GrLivArea + newdata$y2007 +
##     newdata$y2008 + newdata$y2009 + newdata$y2010 + newdata$highqual +
##     newdata$highcond + newdata$hheat + newdata$hkitchen + newdata$hbasement +
##     newdata$GarageCars + newdata$fittedyremod + newdata$FullBath)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99554 -0.10758  0.00187  0.11810  0.65835
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.101e+01  2.281e-02 482.598  < 2e-16 ***
## newdata$GrLivArea     3.505e-04  1.365e-05  25.681  < 2e-16 ***
## newdata$y2007        -5.349e-03  1.502e-02  -0.356  0.72174
## newdata$y2008        -1.825e-03  1.539e-02  -0.119  0.90558
## newdata$y2009        -2.144e-02  1.500e-02  -1.429  0.15311
```

```
## newdata$y2010        -4.896e-04  1.802e-02  -0.027  0.97833
## newdata$highqual      7.845e-02  1.362e-02   5.758 1.04e-08 ***
## newdata$highcond      4.748e-02  1.170e-02   4.058 5.21e-05 ***
## newdata$hheat         3.530e-02  1.263e-02   2.796  0.00525 **
## newdata$hkitchen      8.489e-02  1.444e-02   5.877 5.21e-09 ***
## newdata$hbasement     1.139e-01  1.477e-02   7.714 2.29e-14 ***
## newdata$GarageCars    1.431e-01  8.731e-03  16.389  < 2e-16 ***
## newdata$fittedyremod  2.492e-03  3.573e-04   6.975 4.68e-12 ***
## newdata$FullBath     -2.115e-02  1.361e-02  -1.554  0.12046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1875 on 1407 degrees of freedom
## Multiple R-squared:  0.7757, Adjusted R-squared:  0.7736
## F-statistic: 374.3 on 13 and 1407 DF,  p-value: < 2.2e-16
```

As seen, the signs and significance of all values remain the same in this new model as the original baseline model. Most of the values of the estimated coefficients of the new regression remain the same as well compared to the original baseline model, with the obvious exception of the high basement quality dummy variable which now has a lower coefficient value since the comparison is between houses with a low quality basement instead of a house with no basement at all.

Mallows CP and Multicollinearity

1e. Use Mallows Cp for identifying which terms you will keep from the model in part (d) and also test for multicollinearity. Based on your findings estimate a new model.

```
ss=regsubsets(newdata$lprice~newdata$GrLivArea+newdata$y2007+newdata$y2008+
                newdata$y2009+newdata$y2010+newdata$highqual+newdata$highcond+
                newdata$hheat+newdata$hkitchen+newdata$hbasement+
                newdata$GarageCars+newdata$fittedyremod+newdata$FullBath,
              data=newdata,nbest=1)
s=summary(ss)
coef(ss, which.min(s$cp))
```

```
##           (Intercept)    newdata$GrLivArea     newdata$highqual
##         1.099424e+01         3.398967e-04         7.671881e-02
##     newdata$highcond        newdata$hheat     newdata$hkitchen
##         4.968603e-02         3.474034e-02         8.517370e-02
##    newdata$hbasement   newdata$GarageCars newdata$fittedyremod
##         1.095116e-01         1.425703e-01         2.391518e-03
```

Mallow's cp indicate that a regression without variables on the number of bathrooms and sale year (only 8 variables) would be the ideal regression model. This matches with the summary of the baseline model which suggests that sale year and the number of bathrooms are insignificant explanatory factors that affect sale price.

```
newregcp<-lm(newdata$lprice~newdata$GrLivArea+newdata$highqual+
               newdata$highcond+newdata$hheat+newdata$hkitchen+newdata$hbasement
             +newdata$GarageCars+newdata$fittedyremod)
vif(newregcp)
```

```
##    newdata$GrLivArea     newdata$highqual     newdata$highcond
##             1.437824             1.689837             1.274633
##        newdata$hheat     newdata$hkitchen    newdata$hbasement
##             1.393338             2.098810             2.120465
##   newdata$GarageCars newdata$fittedyremod
##             1.698221             2.105273
```

```
mean(vif(newregcp))
```

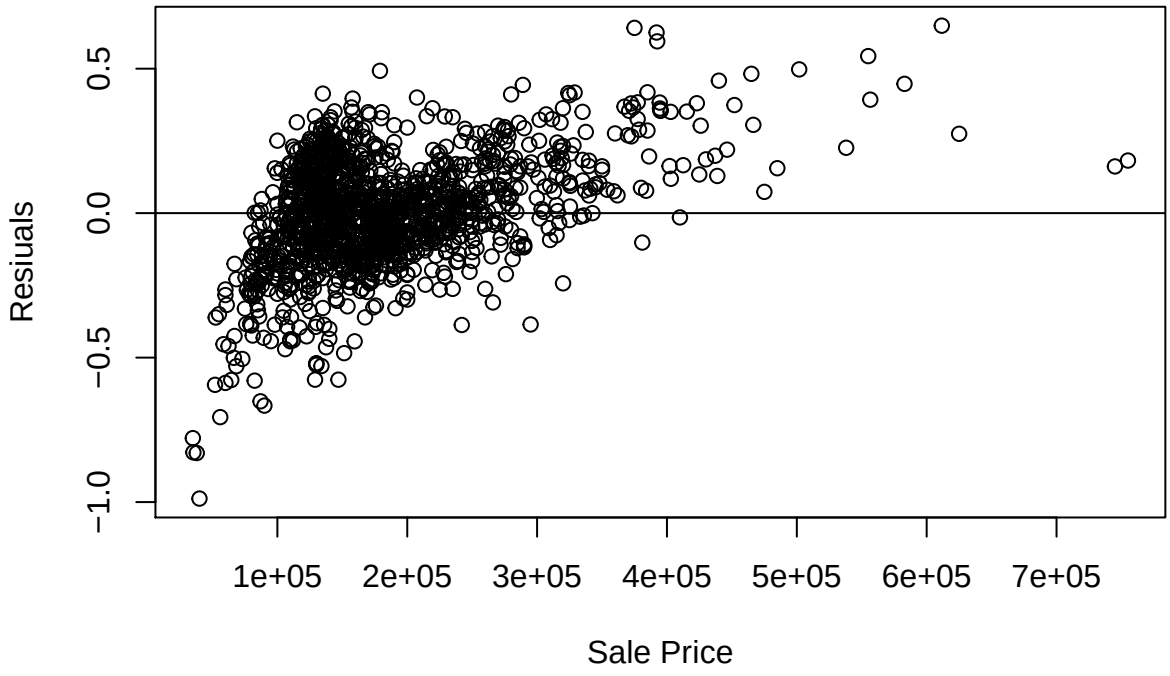
```
## [1] 1.7273
```

A vif test is run on the new Mallow's CP estimated model to test if multicollinearity exists. All of the vif values and mean vif values are relatively low below 10, meaning multicollinearity is not a problem in this regression.

Residual Analysis

1f. For your model in part (e) plot the respective residuals vs. y, and comment on your results.

```
resid_newreg<-resid(newregcp)
plot(newdata$SalePrice,resid_newreg,ylab="Resiuals",xlab="Sale Price",abline(0,0))
```



The residuals are more or less distributed around 0 at random. However, there is a slight trend upwards in residuals as sale prices increase.

AIC and BIC Analysis

1g. Using AIC and BIC for model comparison, identify which model is better, (c) or (e). Why?

```
AIC(regb) #AIC for Regression in c
```

```
## [1] -626.7781
```

```
AIC(newregcp) #AIC for Regression in e
```

```
## [1] -713.6509
```

The AIC for the regression in e is lower than the AIC for the regression in c, therefore model e is better.

```
BIC(regb) #BIC for Regression in c
```

```
## [1] -542.199
```

```
BIC(newregcp) #BIC for Regression in e
```

```
## [1] -661.0598
```

The BIC for the regression in e is lower than the AIC for the regression in c, therefore model e is better.

Interaction Terms

1h. Estimate a model based on (g) that includes interaction terms and if needed, any higher power terms. Comment on the performance of this model compared to your other two models.

```
newdata$qualcond<-(newdata$highqual*newdata$highcond)
regint<-lm(newdata$lprice~newdata$GrLivArea+newdata$highqual+newdata$highcond+
             newdata$qualcond+newdata$hheat+newdata$hkitchen+newdata$hbasement+
             newdata$GarageCars+newdata$fittedyremod)
summary(regint)
```

```
##
## Call:
## lm(formula = newdata$lprice ~ newdata$GrLivArea + newdata$highqual +
##     newdata$highcond + newdata$qualcond + newdata$hheat + newdata$hkitchen +
##     newdata$hbasement + newdata$GarageCars + newdata$fittedyremod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97526 -0.11160  0.00226  0.11639  0.65094
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.098e+01  2.070e-02 530.373  < 2e-16 ***
## newdata$GrLivArea      3.431e-04  1.178e-05  29.138  < 2e-16 ***
## newdata$highqual       1.109e-01  1.849e-02   5.996 2.56e-09 ***
## newdata$highcond       8.495e-02  1.742e-02   4.877 1.20e-06 ***
## newdata$qualcond      -6.305e-02  2.334e-02  -2.701  0.00699 **
## newdata$hheat          3.101e-02  1.264e-02   2.453  0.01427 *
## newdata$hkitchen       8.129e-02  1.447e-02   5.620 2.30e-08 ***
## newdata$hbasement      1.041e-01  1.460e-02   7.126 1.64e-12 ***
## newdata$GarageCars     1.420e-01  8.703e-03  16.317  < 2e-16 ***
## newdata$fittedyremod   2.282e-03  3.546e-04   6.433 1.71e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1871 on 1411 degrees of freedom
## Multiple R-squared:  0.776,  Adjusted R-squared:  0.7746
## F-statistic: 543.2 on 9 and 1411 DF,  p-value: < 2.2e-16
```

A number of interaction terms were used and tested for significance, and the only interaction term that yielded any significance is an interaction between the dummies for high quality and high condition.

```
AIC(newregcp)
```

```
## [1] -713.6509
```

```
AIC(regint)
```

```
## [1] -718.9796
```

```
BIC(newregcp)
```

```
## [1] -661.0598
```

```
BIC(regint)
```

```
## [1] -661.1293
```

The AIC and BIC values of the new regression with the added interaction term is slightly lower than the model specified in model e, hence it is slightly better.

```
newdata$areasq<-(newdata$GrLivArea)^2
summary(lm(newdata$lprice~newdata$GrLivArea+newdata$highqual+newdata$highcond+
             newdata$qualcond+newdata$hheat+newdata$hkitchen+newdata$hbasement+
             newdata$GarageCars+newdata$fittedyremod))
```

```
##
## Call:
## lm(formula = newdata$lprice ~ newdata$GrLivArea + newdata$highqual +
##     newdata$highcond + newdata$qualcond + newdata$hheat + newdata$hkitchen +
##     newdata$hbasement + newdata$GarageCars + newdata$fittedyremod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97526 -0.11160  0.00226  0.11639  0.65094
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.098e+01  2.070e-02 530.373  < 2e-16 ***
## newdata$GrLivArea     3.431e-04  1.178e-05  29.138  < 2e-16 ***
## newdata$highqual      1.109e-01  1.849e-02   5.996 2.56e-09 ***
## newdata$highcond      8.495e-02  1.742e-02   4.877 1.20e-06 ***
## newdata$qualcond     -6.305e-02  2.334e-02  -2.701  0.00699 **
## newdata$hheat         3.101e-02  1.264e-02   2.453  0.01427 *
## newdata$hkitchen      8.129e-02  1.447e-02   5.620 2.30e-08 ***
## newdata$hbasement     1.041e-01  1.460e-02   7.126 1.64e-12 ***
## newdata$GarageCars    1.420e-01  8.703e-03  16.317  < 2e-16 ***
## newdata$fittedyremod  2.282e-03  3.546e-04   6.433 1.71e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1871 on 1411 degrees of freedom
## Multiple R-squared:  0.776,  Adjusted R-squared:  0.7746
## F-statistic: 543.2 on 9 and 1411 DF,  p-value: < 2.2e-16
```

The only term where a square term would be considered would be the living area of the house. However the coefficient of the square term is insignificant. Hence it will not be included.

Five Fold Cross Validation

1i. Lastly, choose you favorite model from all the ones estimated and perform a five-fold cross validation test on it. Then use the test.csv dataset to evaluate how well your model predicts home prices for out of sample data, and comment on your overall findings.

```
train_control<- trainControl(method="cv", number=5, savePredictions = TRUE,
                             returnResamp = 'all')
model <- train(x=newdata[,c(47,87,88,89,96,104,62,105)],y=newdata[,106],
               method="lm", trControl=train_control)
model
```

```
## Linear Regression
##
## 1421 samples
##    8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1137, 1138, 1136, 1136, 1137
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1882477  0.7744743  0.1434117
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

I chose the model used in part e. The resampled dataset has a has R squared and Residual standard error values very similar to the estimated variables in the regression in part e.

```
summary(newregcp)
```

```
##
## Call:
## lm(formula = newdata$lprice ~ newdata$GrLivArea + newdata$highqual +
##     newdata$highcond + newdata$hheat + newdata$hkitchen + newdata$hbasement +
##     newdata$GarageCars + newdata$fittedyremod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98782 -0.10803  0.00166  0.11714  0.64882
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.099e+01  1.985e-02 553.786  < 2e-16 ***
## newdata$GrLivArea    3.399e-04  1.174e-05  28.949  < 2e-16 ***
## newdata$highqual     7.672e-02  1.352e-02   5.674 1.69e-08 ***
## newdata$highcond     4.969e-02  1.156e-02   4.299 1.83e-05 ***
## newdata$hheat        3.474e-02  1.259e-02   2.759  0.00588 **
## newdata$hkitchen     8.517e-02  1.443e-02   5.904 4.43e-09 ***
## newdata$hbasement    1.095e-01  1.450e-02   7.554 7.55e-14 ***
## newdata$GarageCars   1.426e-01  8.719e-03  16.351  < 2e-16 ***
```

```
## newdata$fittedyremod 2.392e-03  3.531e-04   6.773 1.84e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1875 on 1412 degrees of freedom
## Multiple R-squared:  0.7749, Adjusted R-squared:  0.7736
## F-statistic: 607.5 on 8 and 1412 DF,  p-value: < 2.2e-16
```

```
testdata<-read.csv("test.csv")
saleprice<-read.csv("sample_submission.csv")
testdata$saleprice<-saleprice[,2]

testdata$highqual<-ifelse(testdata$OverallQual>5,1,0)
testdata$highcond<-ifelse(testdata$OverallCond>5,1,0)
testdata$hheat<-ifelse(testdata$HeatingQC=="Ex"|testdata$HeatingQC=="Gd",1,0)
testdata$hkitchen<-ifelse(testdata$KitchenQual=="Ex"|testdata$KitchenQual=="Gd",1,0)


testdata$BsmtQual[is.na(testdata$BsmtQual)] <- "NA"
testdata$BEx<-ifelse(testdata$BsmtQual=="Ex",1,0)
testdata$BGd<-ifelse(testdata$BsmtQual=="Gd",1,0)
testdata$BTA<-ifelse(testdata$BsmtQual=="TA",1,0)
testdata$BFa<-ifelse(testdata$BsmtQual=="Fa",1,0)
testdata$BPo<-ifelse(testdata$BsmtQual=="Po",1,0)
testdata$BNA<-ifelse(testdata$BsmtQual=="NA",1,0)
sum(testdata$BNA)
```

```
## [1] 44
```

Again, since our model does not work for houses with no basements and there is only a small sample set of houses without basements, we treat them as outliers and omit them.

```
testdata<-subset(testdata, testdata$BNA==0)
testdata$hbasement<-ifelse(testdata$BsmtQual=="Gd"|testdata$BsmtQual=="Ex",1,0)

testdata$hkitchen<-ifelse(testdata$KitchenQual=="Ex"|testdata$KitchenQual=="Gd",1,0)
testdata$KitchenQual[is.na(testdata$KitchenQual)] <- "NA"
testdata$KNA<-ifelse(testdata$KitchenQual=="NA",1,0)
sum(testdata$KNA)
```

```
## [1] 1
```

There is 1 house that report not having a kitchen, which I would omit as an outlier as it is incompatible with this model.

```
testdata<-subset(testdata, testdata$BNA==0&testdata$KNA==0)

testdata$GarageCars[is.na(testdata$GarageCars)] <- 0

min(testdata$YearRemodAdd)
```

```
## [1] 1950
```

Minimum year in this dataset is also 1950, so the same variable can be used.

```
testdata$fittedyremod<-testdata$YearRemodAdd-1950

testdata$p_price<-exp(as.numeric(newregcp$coefficients[1])+
                    as.numeric(newregcp$coefficients[2])*testdata$GrLivArea+
                    as.numeric(newregcp$coefficients[3])*testdata$highqual+
                    as.numeric(newregcp$coefficients[4])*testdata$highcond+
                    as.numeric(newregcp$coefficients[5])*testdata$hheat+
                    as.numeric(newregcp$coefficients[6])*testdata$hkitchen+
                    as.numeric(newregcp$coefficients[7])*testdata$hbasement+
                    as.numeric(newregcp$coefficients[8])*testdata$GarageCars+
                    as.numeric(newregcp$coefficients[9])*testdata$fittedyremod)

actualvalues<-read.csv("sample_submission.csv")

testdata$difference<-testdata$p_price-testdata$saleprice
mean(testdata$difference)
```

```
## [1] -1857.839
```

The model chosen on average underestimates the actual sale price of houses. This could potentially be due to a significant variable that was not chosen as one of my 10 variables for analysis, giving a downward bias for the estimates.