Toronto Real Estate

#Load packages

```
realestatedata <- read.csv("real-estate-data.csv")
```

Descriptive statistics of the real estate dataset

```
summary(realestatedata)
```

```
##       id_             ward               beds           baths      
##  Min.   :100681   Length:3042        Min.   :0.000   Min.   :1.000  
##  1st Qu.:328133   Class :character   1st Qu.:1.000   1st Qu.:1.000  
##  Median :545680   Mode  :character   Median :1.000   Median :1.000  
##  Mean   :546242                      Mean   :1.445   Mean   :1.505  
##  3rd Qu.:766250                      3rd Qu.:2.000   3rd Qu.:2.000  
##  Max.   :998996                      Max.   :3.000   Max.   :3.000  
##                                      NA's   :54                     
##      DEN                size             parking            exposure        
##  Length:3042        Length:3042        Length:3042        Length:3042       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##      D_mkt          building_age        maint            price        
##  Min.   :  0.00   Min.   : 0.000   Min.   : 179.0   Min.   : 298000  
##  1st Qu.:  4.00   1st Qu.: 3.000   1st Qu.: 444.0   1st Qu.: 551000  
##  Median : 10.00   Median : 7.000   Median : 595.0   Median : 718000  
##  Mean   : 14.01   Mean   : 9.801   Mean   : 752.5   Mean   : 893422  
##  3rd Qu.: 19.00   3rd Qu.:14.000   3rd Qu.: 882.0   3rd Qu.:1008000  
##  Max.   :169.00   Max.   :75.000   Max.   :5395.0   Max.   :5688000  
##  NA's   :93                        NA's   :45       NA's   :61       
##        lt              lg        
##  Min.   :43.62   Min.   :-79.43  
##  1st Qu.:43.65   1st Qu.:-79.41  
##  Median :43.66   Median :-79.39  
##  Mean   :43.66   Mean   :-79.39  
##  3rd Qu.:43.67   3rd Qu.:-79.38  
##  Max.   :43.69   Max.   :-79.35  
##
```

```
class(realestatedata)
```

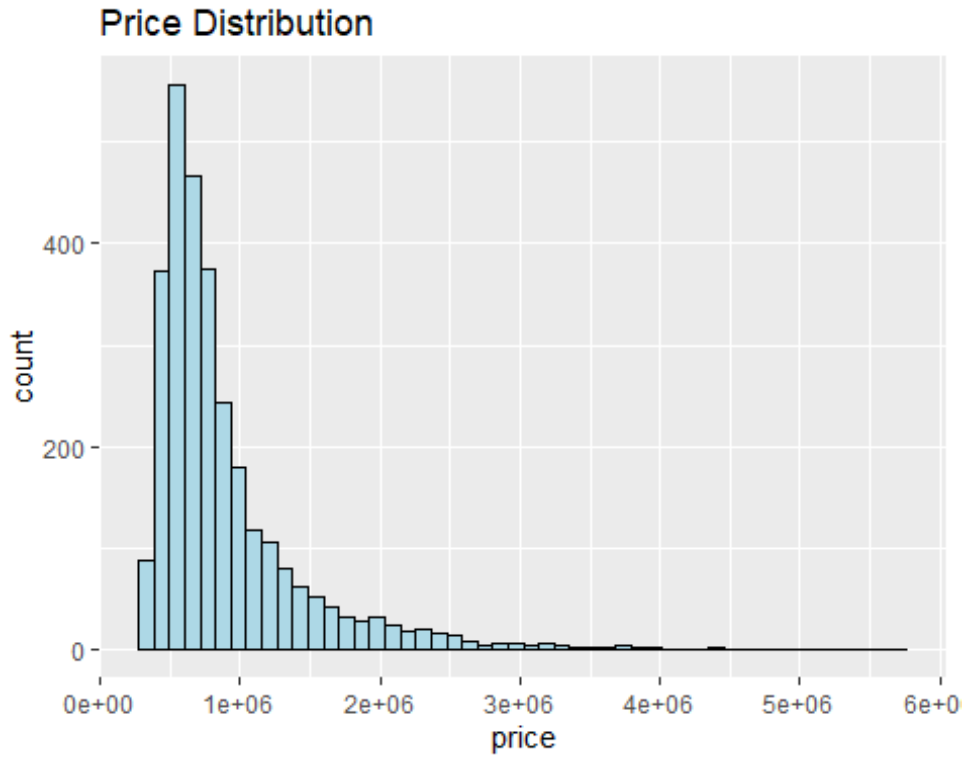
```
## [1] "data.frame"
```

```
table(realestatedata$size)
```

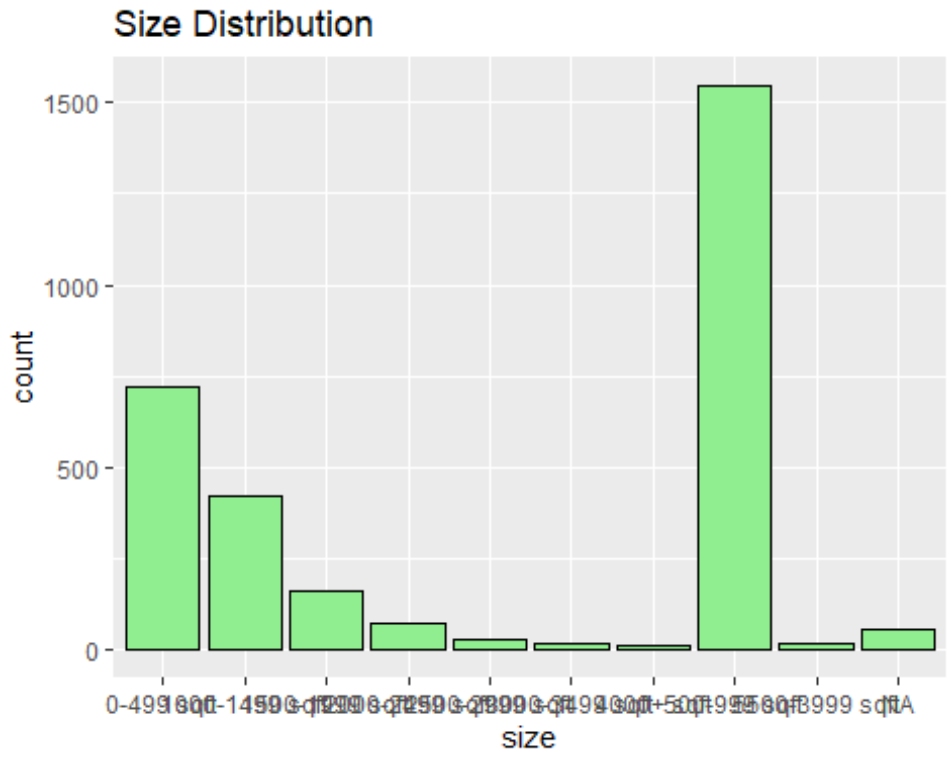
```
##
##     0-499 sqft 1000-1499 sqft 1500-1999 sqft 2000-2499 sqft 2500-2999 sqft
##            719            420            163             74             28
## 3000-3499 sqft     4000+ sqft   500-999 sqft 5500-3999 sqft
##             14             10           1546             15
```

Distributions of factors affecting Toronto housing prices

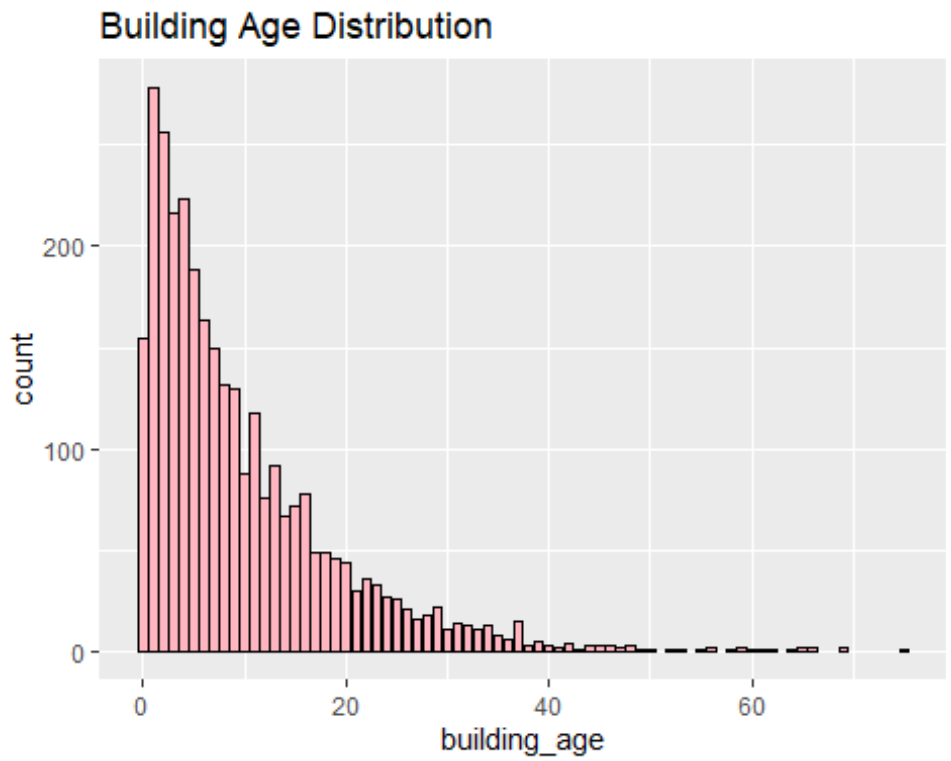
```
ggplot(data = realestatedata, aes(x = price))+geom_histogram(position =
"dodge", bins = 50, fill = "lightblue", color = "black") + labs(title =
"Price Distribution")

## Warning: Removed 61 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



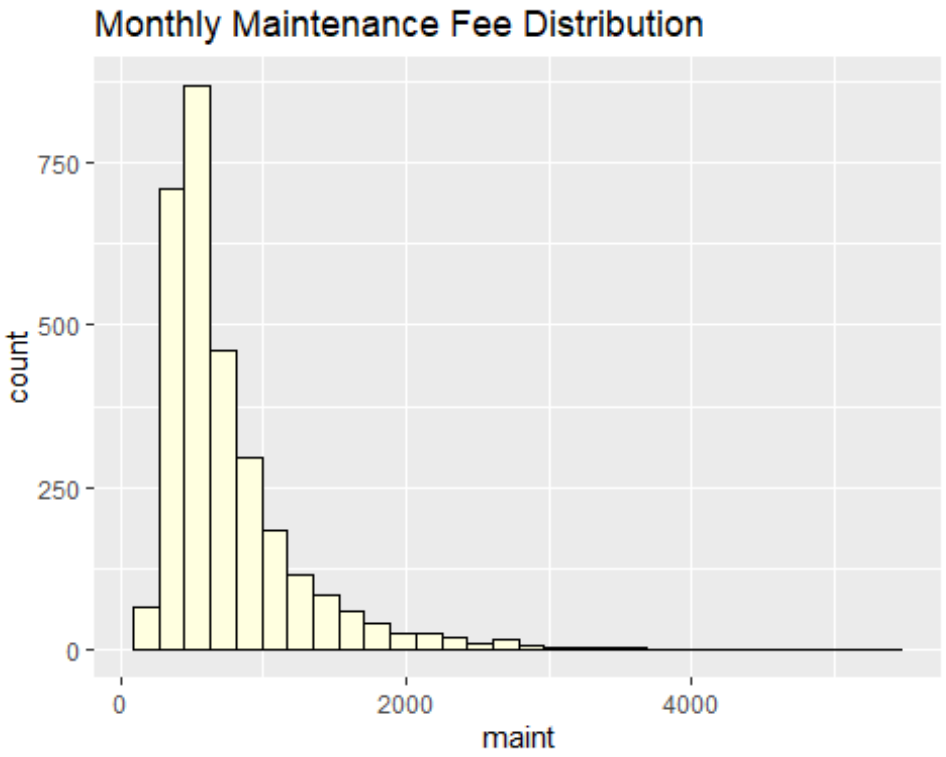

```
ggplot(data = realestatedata, aes(x = size, fill = ))+geom_bar(position =
"dodge", bins = 30, fill = "lightgreen", color = "black") + labs(title =
"Size Distribution")

## Warning in geom_bar(position = "dodge", bins = 30, fill = "lightgreen", :
## Ignoring unknown parameters: `bins`
```



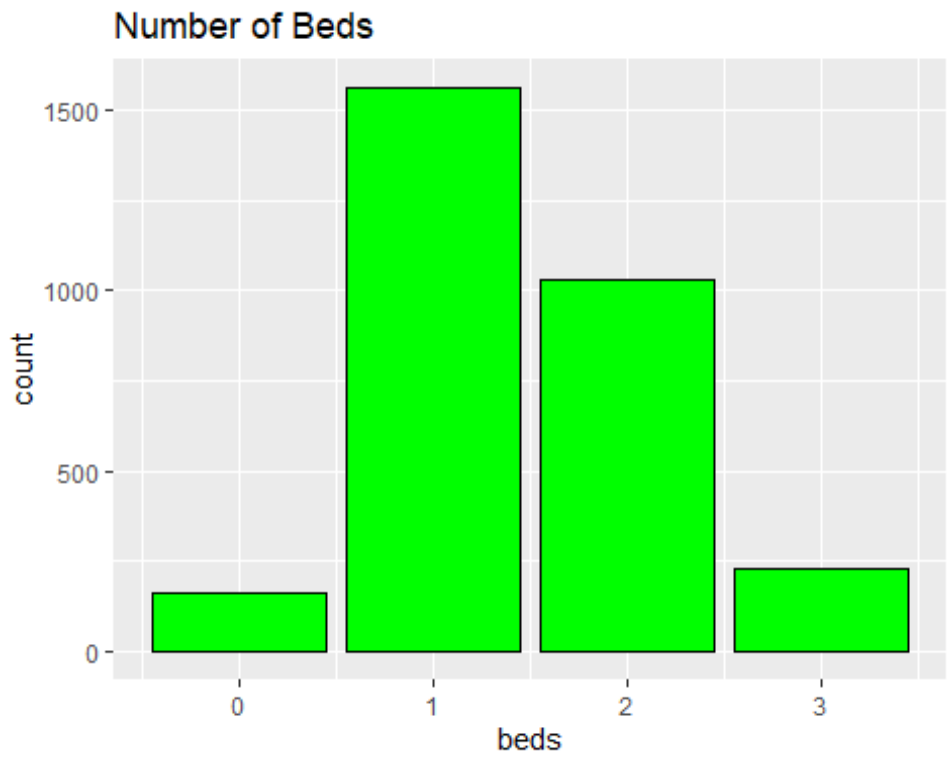

```
ggplot(data = realestatedata, aes(x = building_age, fill =
))+geom_bar(position = "dodge", bins = 30, fill = "lightpink", color =
"black") + labs(title = "Building Age Distribution")

## Warning in geom_bar(position = "dodge", bins = 30, fill = "lightpink",
color =
## "black"): Ignoring unknown parameters: `bins`
```




```
ggplot(data = realestatedata, aes(x = maint, fill =
))+geom_histogram(position = "dodge", bins = 30, fill = "lightyellow", color
= "black") + labs(title = "Monthly Maintenance Fee Distribution")

## Warning: Removed 45 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



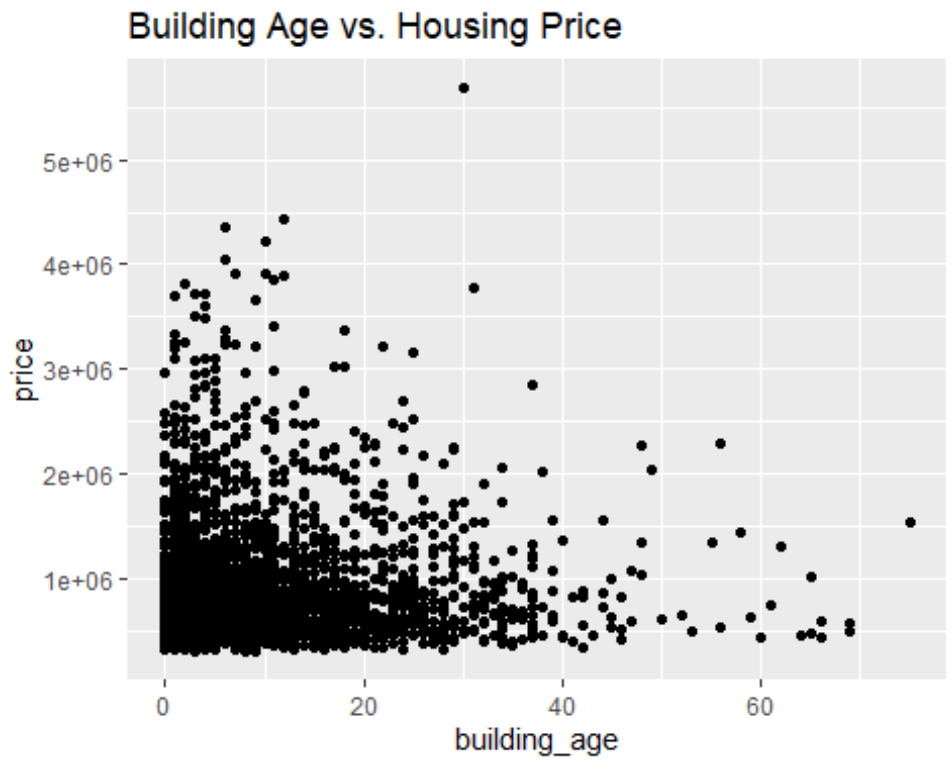

```
ggplot(data = realestatedata, aes(x = beds, fill = ))+geom_bar(position =
"dodge", bins = 30, fill = "green", color = "black") + labs(title = "Number
of Beds")

## Warning in geom_bar(position = "dodge", bins = 30, fill = "green", color =
## "black"): Ignoring unknown parameters: `bins`

## Warning: Removed 54 rows containing non-finite outside the scale range
## (`stat_count()`).
```



Data visualizations comparing Toronto housing price and related factors

```
ggplot(data = realestatedata, aes(x = size, y =
price))+geom_boxplot(fill="lightblue") + labs(title = "Size vs. Housing
Price")

## Warning: Removed 61 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

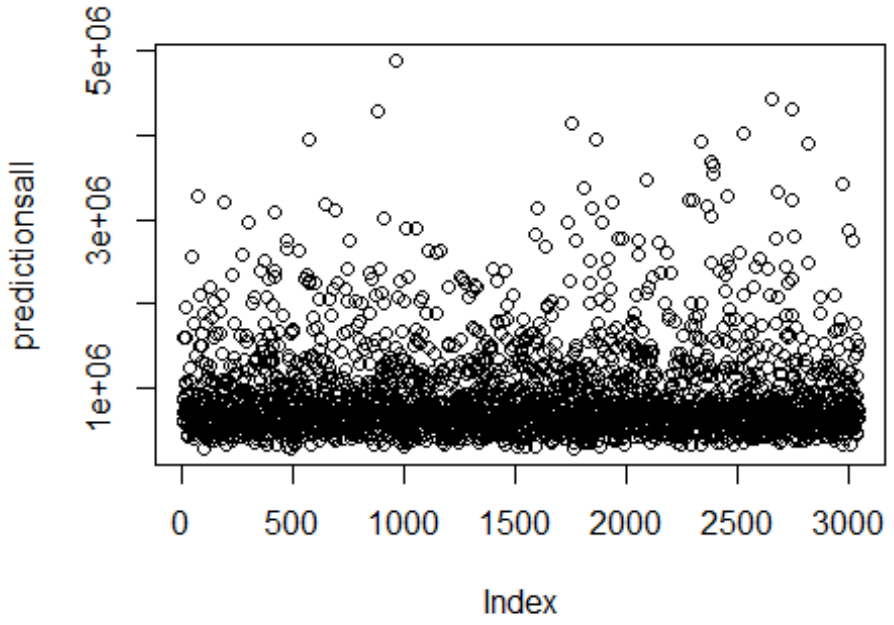

```
ggplot(data = realestatedata, aes(x = building_age, y = price))+geom_point()+
labs(title = "Building Age vs. Housing Price")

## Warning: Removed 61 rows containing missing values or values outside the
scale range
## (`geom_point()`).
```



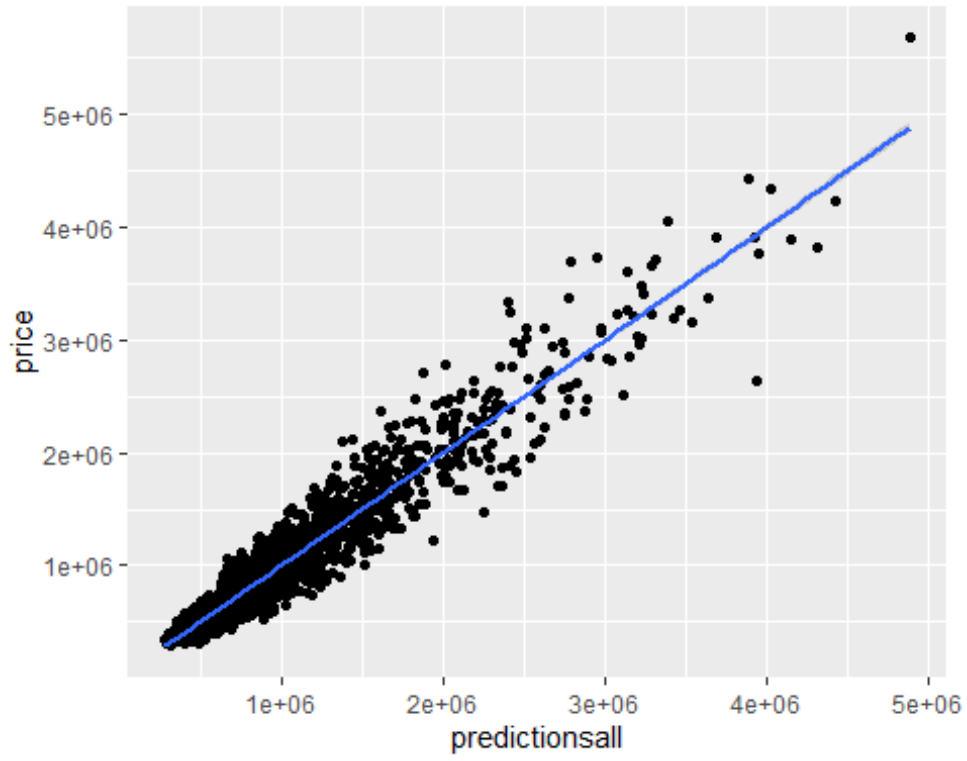
Exploring the correlation between the housing prices provided by the dataset and the prices generated by our multiple linear regression model

```
modelrem <- lm(formula = price ~ ward + beds + baths + DEN + size + parking +
D_mkt + building_age + exposure + maint + lt +lg, data = realestatedata)
summary(modelrem)

##
## Call:
## lm(formula = price ~ ward + beds + baths + DEN + size + parking +
##     D_mkt + building_age + exposure + maint + lt + lg, data =
realestatedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1305697   -78491    -5844    72814   928411
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.199e+07  3.031e+07   1.055  0.29135
## wardW11            -6.529e+03  1.194e+04  -0.547  0.58468
## wardW13            -2.337e+04  1.302e+04  -1.795  0.07278 .
## beds                7.309e+04  7.041e+03  10.382  < 2e-16 ***
```

```
## baths               -4.896e+03  7.338e+03  -0.667  0.50476
## DENYES               1.836e+04  6.494e+03   2.828  0.00473 **
## size1000-1499 sqft   3.172e+05  1.287e+04  24.646  < 2e-16 ***
## size1500-1999 sqft   6.028e+05  1.839e+04  32.776  < 2e-16 ***
## size2000-2499 sqft   8.711e+05  2.563e+04  33.990  < 2e-16 ***
## size2500-2999 sqft   9.987e+05  4.034e+04  24.756  < 2e-16 ***
## size3000-3499 sqft   1.190e+06  4.824e+04  24.667  < 2e-16 ***
## size4000+ sqft       1.759e+06  6.399e+04  27.494  < 2e-16 ***
## size500-999 sqft     9.659e+04  8.205e+03  11.772  < 2e-16 ***
## size5500-3999 sqft   1.632e+06  4.867e+04  33.539  < 2e-16 ***
## parkingYes           4.578e+03  6.137e+03   0.746  0.45577
## D_mkt                7.243e+01  2.201e+02   0.329  0.74215
## building_age        -8.383e+01  3.081e+02  -0.272  0.78559
## exposureNo          -2.909e+03  9.509e+03  -0.306  0.75966
## exposureS           -9.787e+03  9.249e+03  -1.058  0.29009
## exposureWe          -2.368e+04  1.089e+04  -2.174  0.02976 *
## maint                5.501e+02  1.139e+01  48.314  < 2e-16 ***
## lt                  -3.530e+05  3.985e+05  -0.886  0.37580
## lg                   2.062e+05  2.154e+05   0.957  0.33857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 158100 on 2726 degrees of freedom
##   (293 observations deleted due to missingness)
## Multiple R-squared:  0.9203, Adjusted R-squared:  0.9197
## F-statistic:  1431 on 22 and 2726 DF,  p-value: < 2.2e-16

predictionsall <- predict(modelrem, newdata = realestatedata)

realestatedata$Predicted_Price_All <- predictionsall

plot(predictionsall)
```



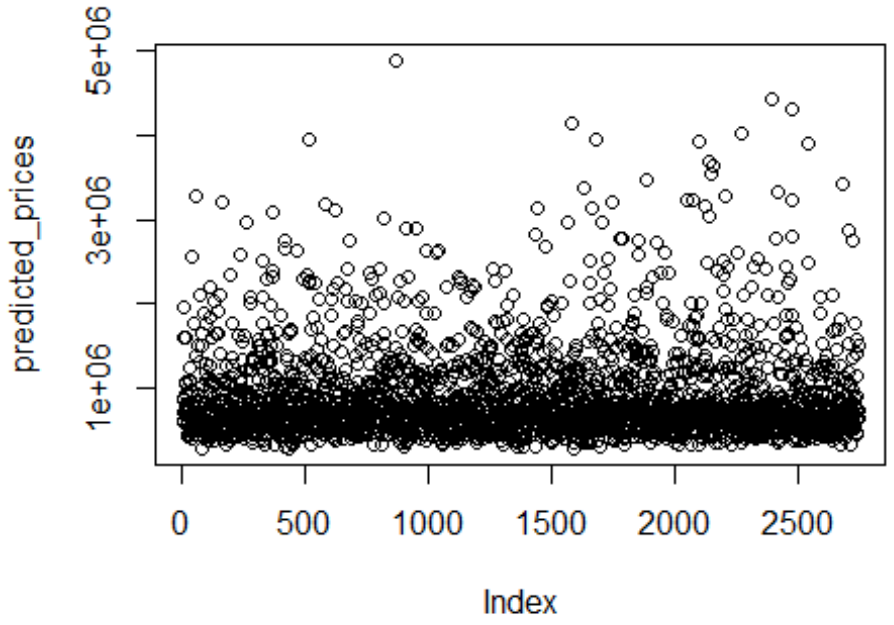

```
ggplot(data = realestatedata, aes(y = price, x =
predictionsall))+geom_point()+geom_smooth(method=lm)

## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 293 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 293 rows containing missing values or values outside the
scale range
## (`geom_point()`).
```



If the p-value of the predictor (intercept) was less than 0.05, it has a statistically significant relationship with housing prices. This means that there is a less than 5% likelihood that the observed values were found by chance. The R-squared value was computed to be 0.9203, which indicates that approximately 92.03% of the variation in Toronto housing prices can be represented by the predictors in this model.

Descriptive statistics for predicted housing prices

```
new <- data.frame(lt=c(-43.64), lg=c(-79.35))

predicted_prices <- predict(modelrem, realestatedata=new)

plot(predicted_prices)
```

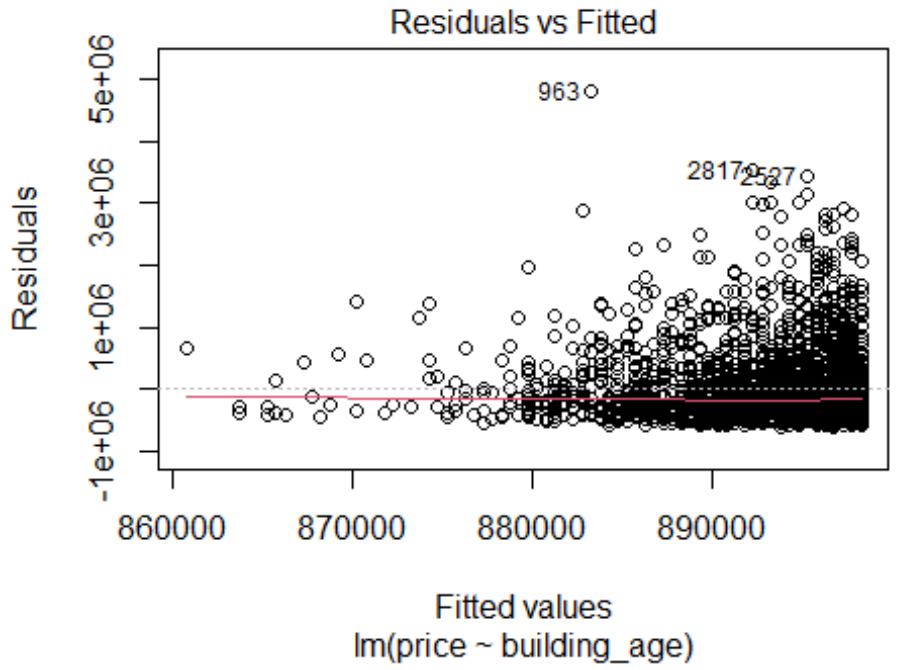



```
summary(predicted_prices)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  283824  581046  719644  894040 1019598 4885261
```

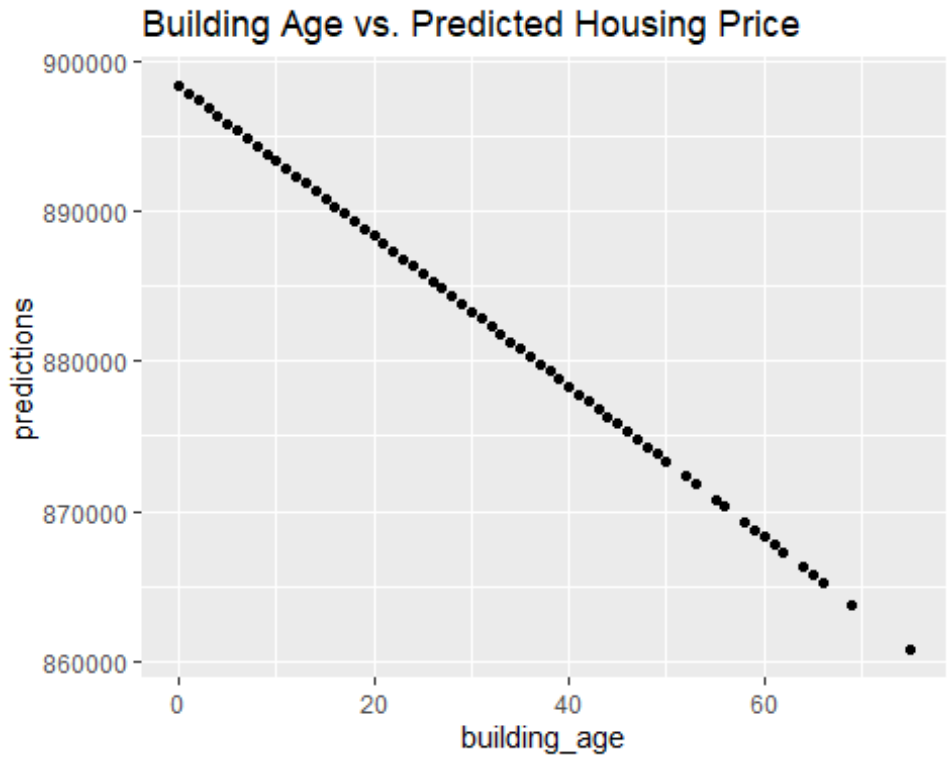
Predictor models

```
modelage <- lm(price ~ building_age , data = realestatedata)
summary(modelage)

##
## Call:
## lm(formula = price ~ building_age, data = realestatedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -595839 -341841 -175342  117163 4804675
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  898344.7    14458.1  62.134   <2e-16 ***
## building_age   -500.7     1039.9  -0.481     0.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 558100 on 2979 degrees of freedom
```

```
##   (61 observations deleted due to missingness)
## Multiple R-squared:  7.78e-05,   Adjusted R-squared:  -0.0002579
## F-statistic: 0.2318 on 1 and 2979 DF,  p-value: 0.6302

plot(modelage, which=1)
```



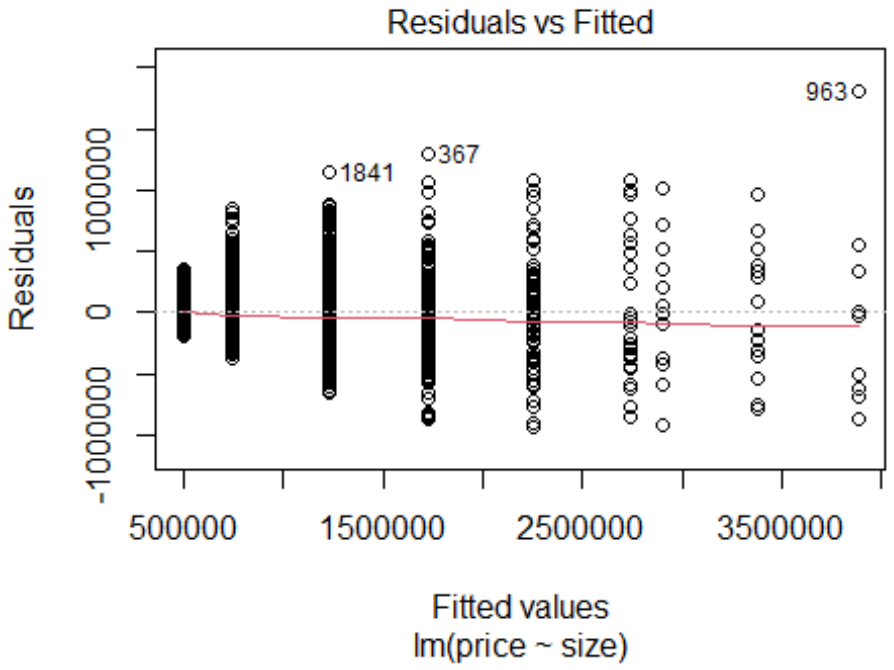

```
predictions <- predict(modelage, newdata = realestatedata)

realestatedata$Predicted_Price <- predictions

ggplot(data = realestatedata, aes(y = predictions, x =
building_age))+geom_point() + labs(title = "Building Age vs. Predicted
Housing Price")
```




```
modelsize <- lm(price ~ size , data = realestatedata)
summary(modelsize)

##
## Call:
## lm(formula = price ~ size, data = realestatedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -930699 -127629  -18629   95206 1804667
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           494377       8651   57.15   <2e-16 ***
## size1000-1499 sqft    737549      14180   52.01   <2e-16 ***
## size1500-1999 sqft   1232923      20115   61.29   <2e-16 ***
## size2000-2499 sqft   1762321      28242   62.40   <2e-16 ***
## size2500-2999 sqft   2242326      45045   49.78   <2e-16 ***
## size3000-3499 sqft   2404980      61997   38.79   <2e-16 ***
## size4000+ sqft       3388956      77054   43.98   <2e-16 ***
## size500-999 sqft      245252      10481   23.40   <2e-16 ***
## size5500-3999 sqft   2888556      59936   48.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229700 on 2919 degrees of freedom
##   (114 observations deleted due to missingness)
```

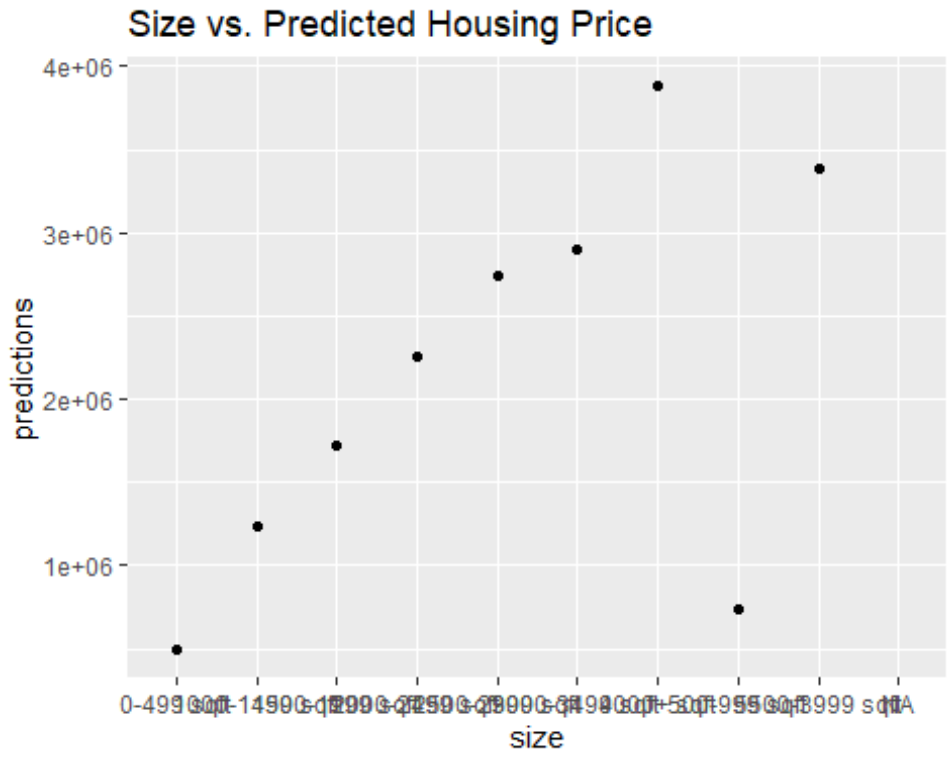
```
## Multiple R-squared:  0.8323, Adjusted R-squared:  0.8318
## F-statistic:  1811 on 8 and 2919 DF,  p-value: < 2.2e-16
```

```
plot(modelsize, which=1)
```

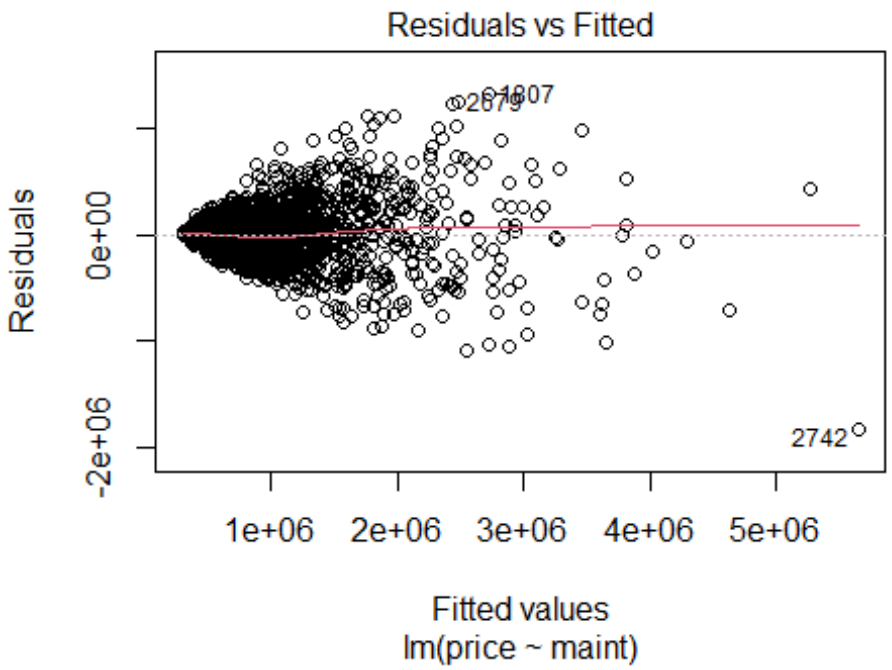



```
predictions <- predict(modelsize, newdata = realestatedata)

realestatedata$Predicted_Price <- predictions

ggplot(data = realestatedata, aes(y = predictions, x = size))+geom_point() +
labs(title = "Size vs. Predicted Housing Price")
```

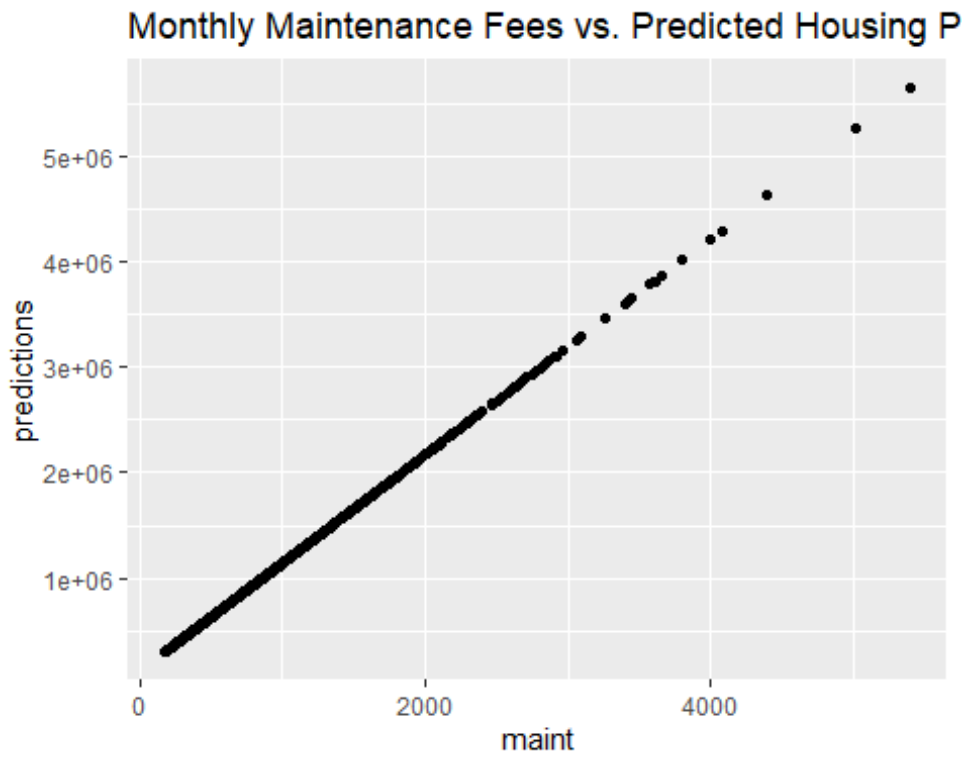
```
## Warning: Removed 53 rows containing missing values or values outside the
scale range
## (`geom_point()`).
```



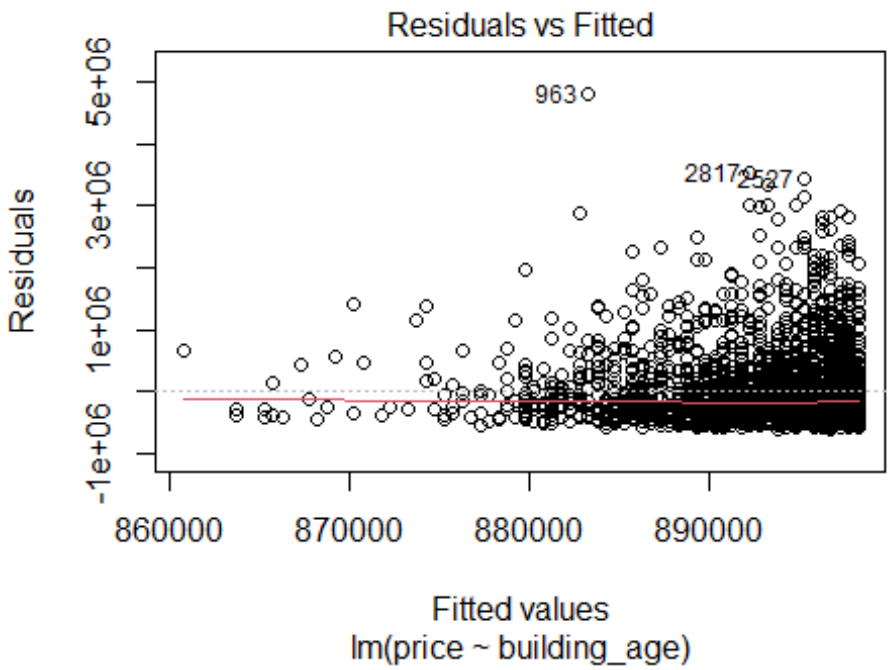

```
modelmaint <- lm(price ~ maint , data = realestatedata)
summary(modelmaint)

##
## Call:
## lm(formula = price ~ maint, data = realestatedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1825087   -97516    -2894    87176  1322432
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.223e+05  7.257e+03   16.86   <2e-16 ***
## maint       1.024e+03  8.013e+00  127.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 218300 on 2936 degrees of freedom
##   (104 observations deleted due to missingness)
## Multiple R-squared:  0.8475, Adjusted R-squared:  0.8475
## F-statistic: 1.632e+04 on 1 and 2936 DF,  p-value: < 2.2e-16

plot(modelmaint, which=1)
```



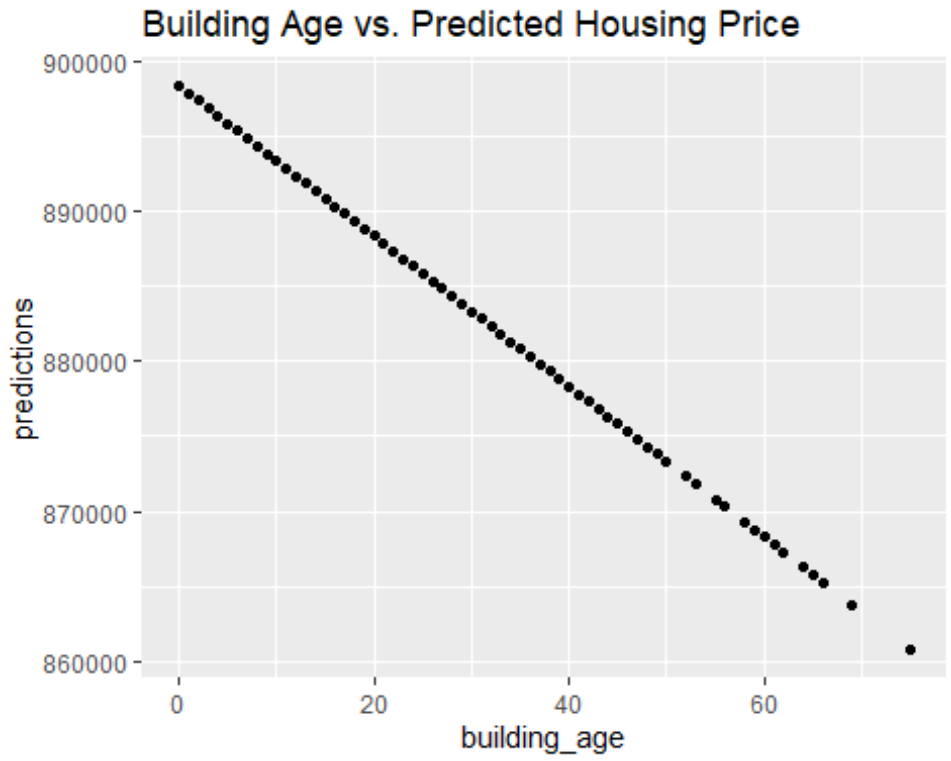

```
predictions <- predict(modelmaint, newdata = realestatedata)

realestatedata$Predicted_Price <- predictions

ggplot(data = realestatedata, aes(y = predictions, x = maint))+geom_point()+
labs(title = "Monthly Maintenance Fees vs. Predicted Housing Price")

## Warning: Removed 45 rows containing missing values or values outside the
scale range
## (`geom_point()`).
```



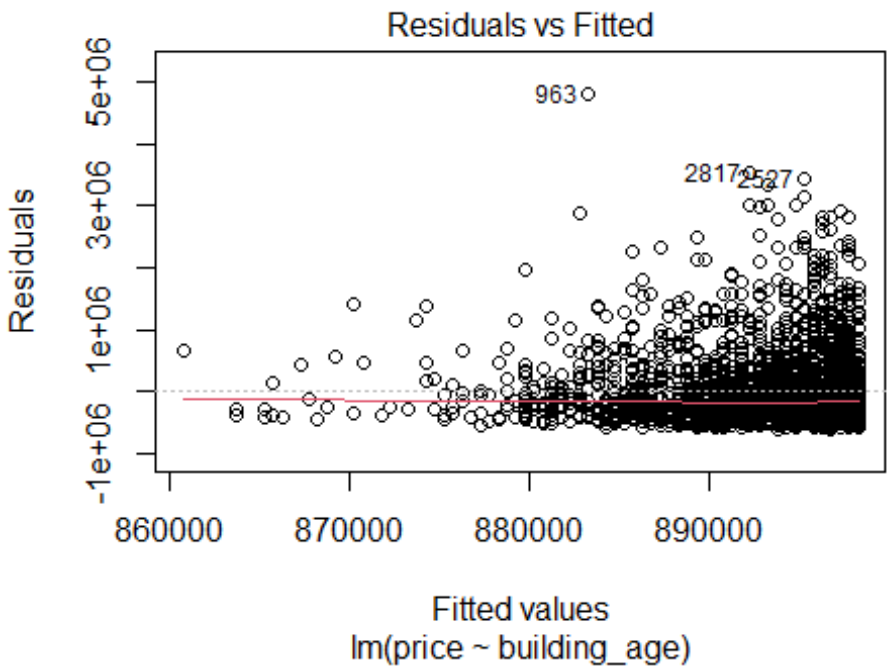

```
modelage <- lm(price ~ building_age , data = realestatedata)
summary(modelage)

##
## Call:
## lm(formula = price ~ building_age, data = realestatedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -595839 -341841 -175342  117163 4804675
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  898344.7    14458.1  62.134   <2e-16 ***
## building_age   -500.7     1039.9  -0.481     0.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 558100 on 2979 degrees of freedom
##   (61 observations deleted due to missingness)
## Multiple R-squared:  7.78e-05,   Adjusted R-squared:  -0.0002579
## F-statistic: 0.2318 on 1 and 2979 DF,  p-value: 0.6302

plot(modelage, which=1)
```



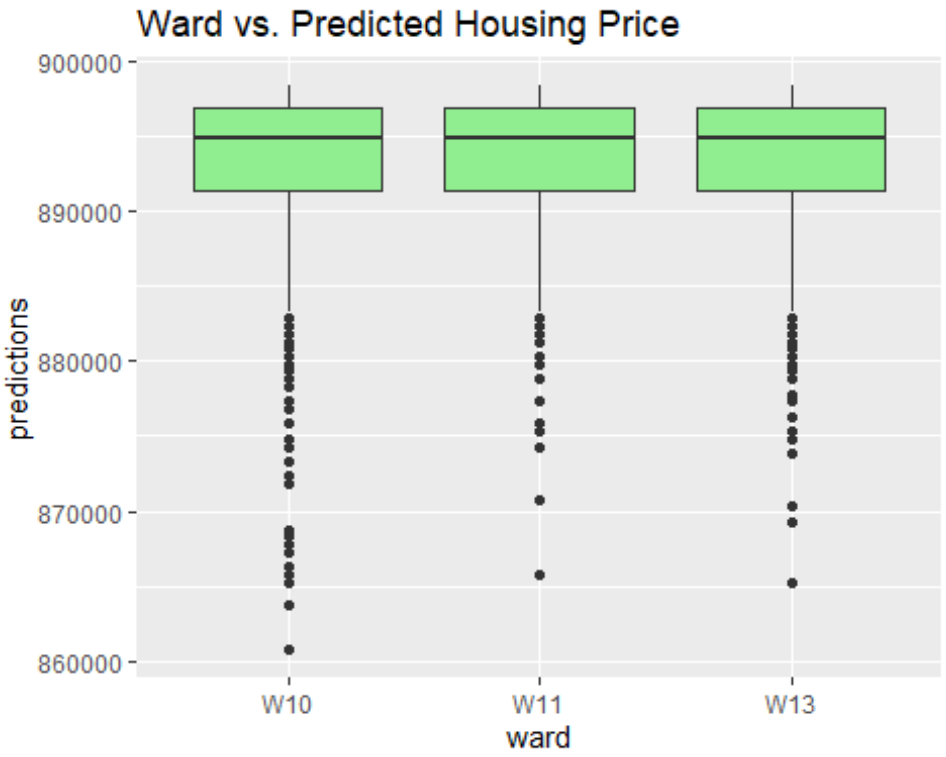

```
predictions <- predict(modelage, newdata = realestatedata)

realestatedata$Predicted_Price <- predictions

ggplot(data = realestatedata, aes(y = predictions, x =
building_age))+geom_point()+ labs(title = "Building Age vs. Predicted Housing
Price")
```




```
modelward <- lm(price ~ building_age , data = realestatedata)
summary(modelward)

##
## Call:
## lm(formula = price ~ building_age, data = realestatedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -595839 -341841 -175342  117163 4804675
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  898344.7    14458.1  62.134   <2e-16 ***
## building_age   -500.7     1039.9  -0.481     0.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 558100 on 2979 degrees of freedom
##   (61 observations deleted due to missingness)
## Multiple R-squared:  7.78e-05,   Adjusted R-squared:  -0.0002579
## F-statistic: 0.2318 on 1 and 2979 DF,  p-value: 0.6302

plot(modelward, which=1)
```




```
predictions <- predict(modelward, newdata = realestatedata)

realestatedata$Predicted_Price <- predictions

ggplot(data = realestatedata, aes(y = predictions, x =
ward))+geom_boxplot(fill="lightgreen") + labs(title = "Ward vs. Predicted
Housing Price")
```


Ward vs. Predicted Housing Price
900000
890000
880000
870000
860000
predictions
W10
W11
W13
ward