Housing_project-DANA-4810

Khushank Sethi (100423378)

2024-10-25

```
housing <- read.csv("House_listing.csv")
attach(housing)
cat("total number of records:", nrow(housing))
```

```
## total number of records: 35768
```

```
for(col in unique(names(housing))) {
  cat("Missing values of", col, ":", length(housing[is.na(housing[[col]])]), "\n")
}
```

```
## Missing values of City : 0
## Missing values of Price : 0
## Missing values of Address : 0
## Missing values of Number_Beds : 0
## Missing values of Number_Baths : 0
## Missing values of Province : 0
## Missing values of Population : 0
## Missing values of Latitude : 0
## Missing values of Longitude : 0
## Missing values of Median_Family_Income : 0
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
cat("Out of range values of Latitude", length(housing$Latitude[!between(Latitude, 42, 83)]), "\n")
```

```
## Out of range values of Latitude 0
```

```r
cat("Out of range values of Longitude", length(housing$Longitude[!between(housing$Longitude, -141, -52)]
```

```
## Out of range values of Longitude 91
```

```r
housing$Longitude[housing$Longitude == 63.1005] <- -63.1005
cat("Out of range values of Latitude after cleaning", length(housing$Longitude[!between(housing$Longitu
```

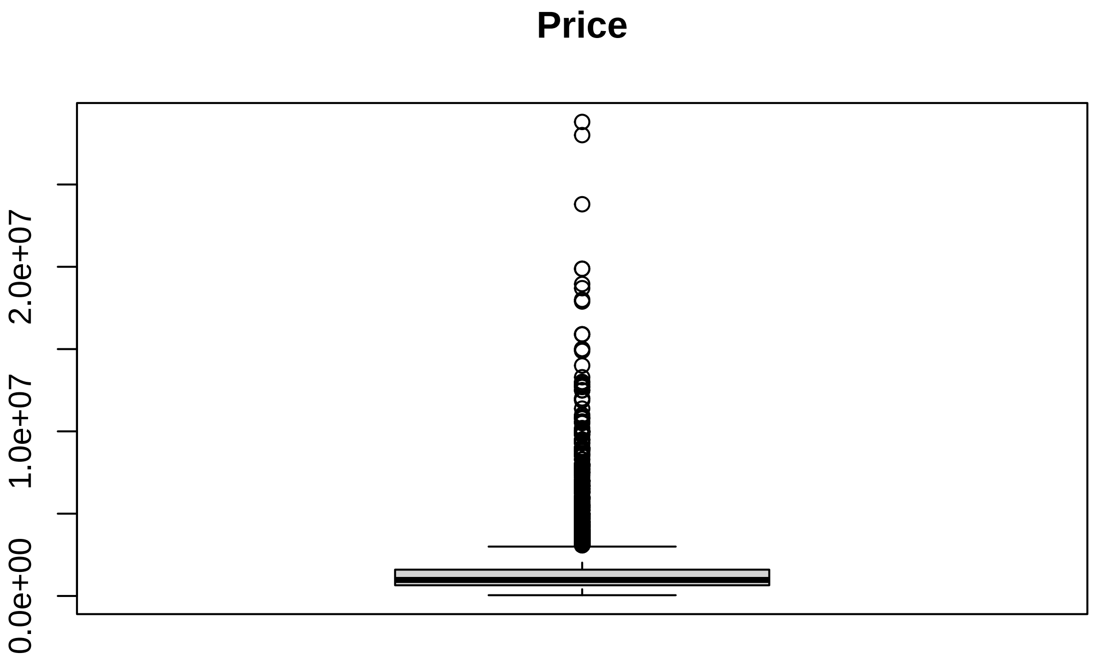
```
## Out of range values of Latitude after cleaning 0
```

```r
housing$Price <- as.numeric(gsub(",", "", Price))
summary(housing)
```

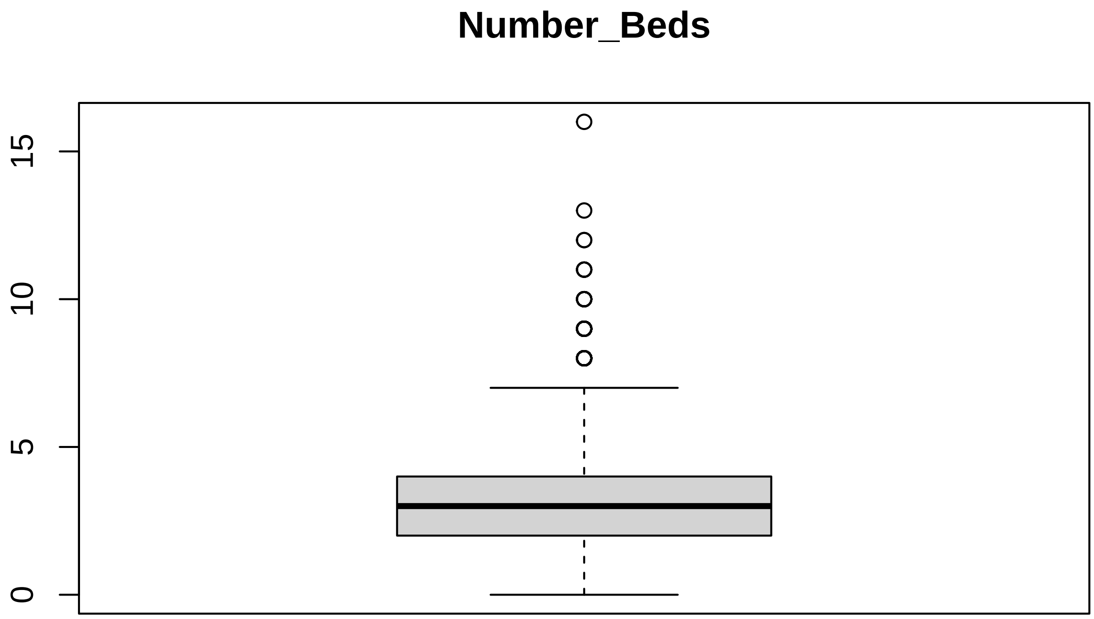
```
##      City               Price             Address           Number_Beds
##  Length:35768       Min.   :   21500   Length:35768       Min.   :  0.000
##  Class :character   1st Qu.:  459900   Class :character   1st Qu.:  2.000
##  Mode  :character   Median :  699000   Mode  :character   Median :  3.000
##                     Mean   :  943296                      Mean   :  3.284
##                     3rd Qu.: 1095000                      3rd Qu.:  4.000
##                     Max.   :37000000                      Max.   :109.000
##   Number_Baths      Province           Population         Latitude
##  Min.   : 0.000   Length:35768       Min.   :  63382   Min.   :42.28
##  1st Qu.: 2.000   Class :character   1st Qu.: 109167   1st Qu.:43.87
##  Median : 2.000   Mode  :character   Median : 242460   Median :49.02
##  Mean   : 2.532                      Mean   : 636015   Mean   :47.45
##  3rd Qu.: 3.000                      3rd Qu.: 522888   3rd Qu.:49.89
##  Max.   :59.000                      Max.   :5647656   Max.   :53.92
##    Longitude        Median_Family_Income
##  Min.   :-123.94   Min.   : 62400
##  1st Qu.:-122.32   1st Qu.: 82000
##  Median :-104.61   Median : 89000
##  Mean   : -98.74   Mean   : 89643
##  3rd Qu.: -79.87   3rd Qu.: 97000
##  Max.   : -52.80   Max.   :133000
```

```r
write.csv(housing, file = "cleaned_housing.csv")
```

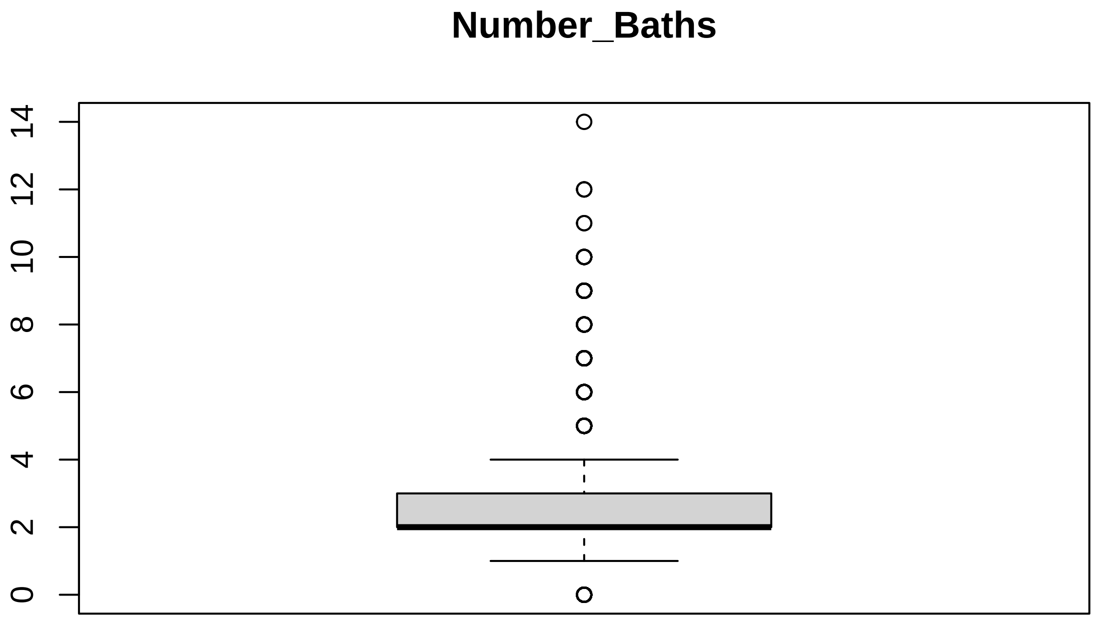
```r
bcHousing <- housing[housing$Province == "British Columbia",]
bcHousing_numCols <-  bcHousing[sapply(bcHousing, is.numeric)]
for(col in names(bcHousing_numCols)) {
  boxplot(bcHousing_numCols[[col]], main=col)
  cat(col, "outliers:")
  print(length(boxplot.stats(bcHousing_numCols[[col]])$out))
}
```



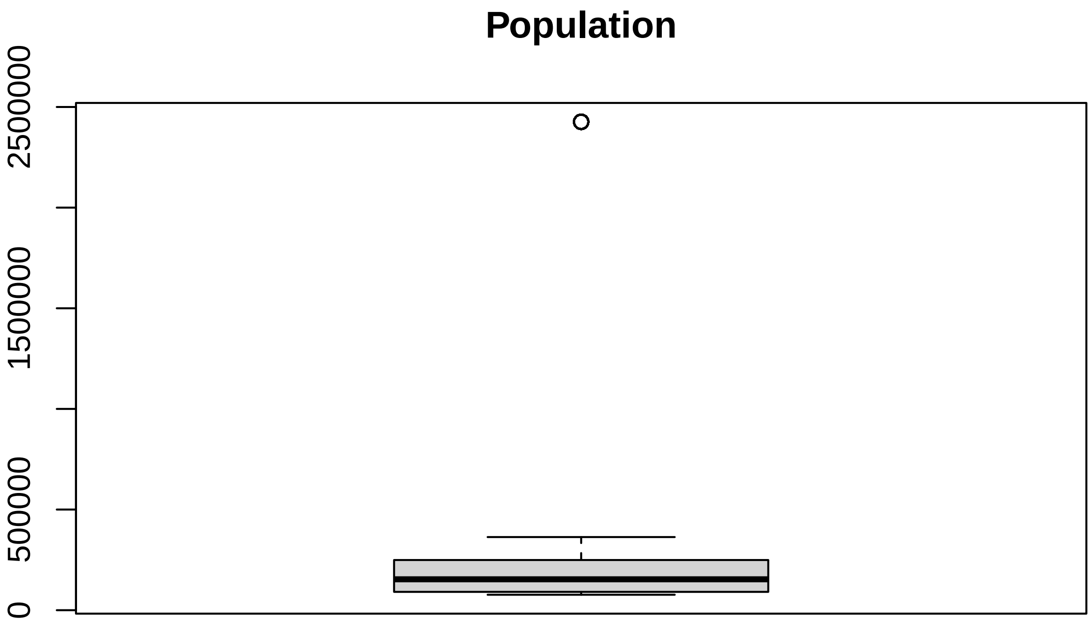
```
## Price outliers:[1] 683
```



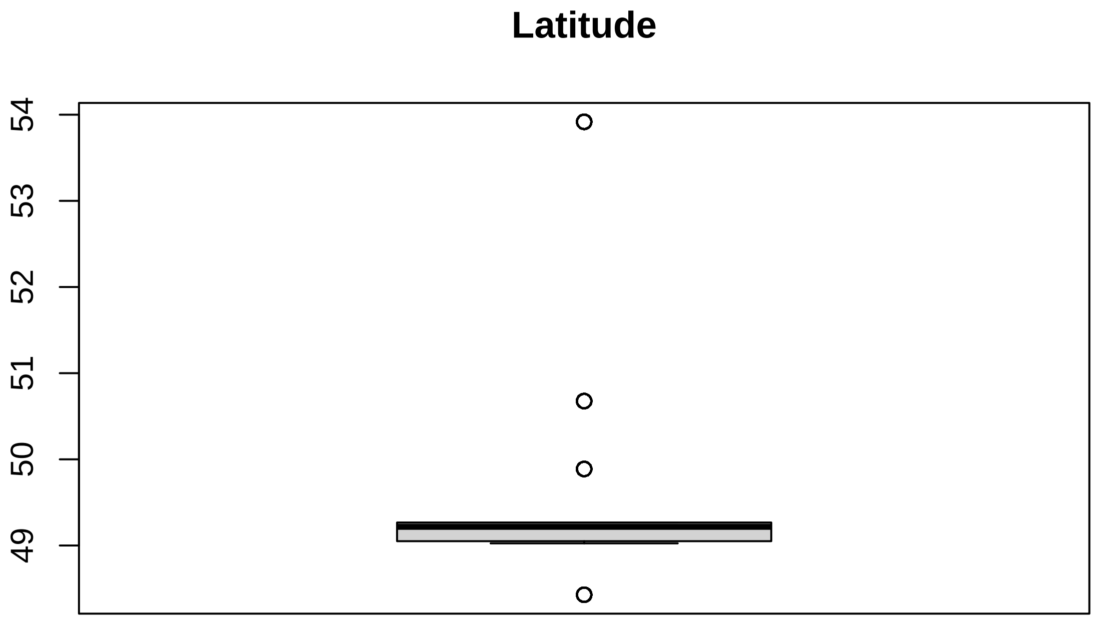
```
## Number_Beds outliers:[1] 212
```



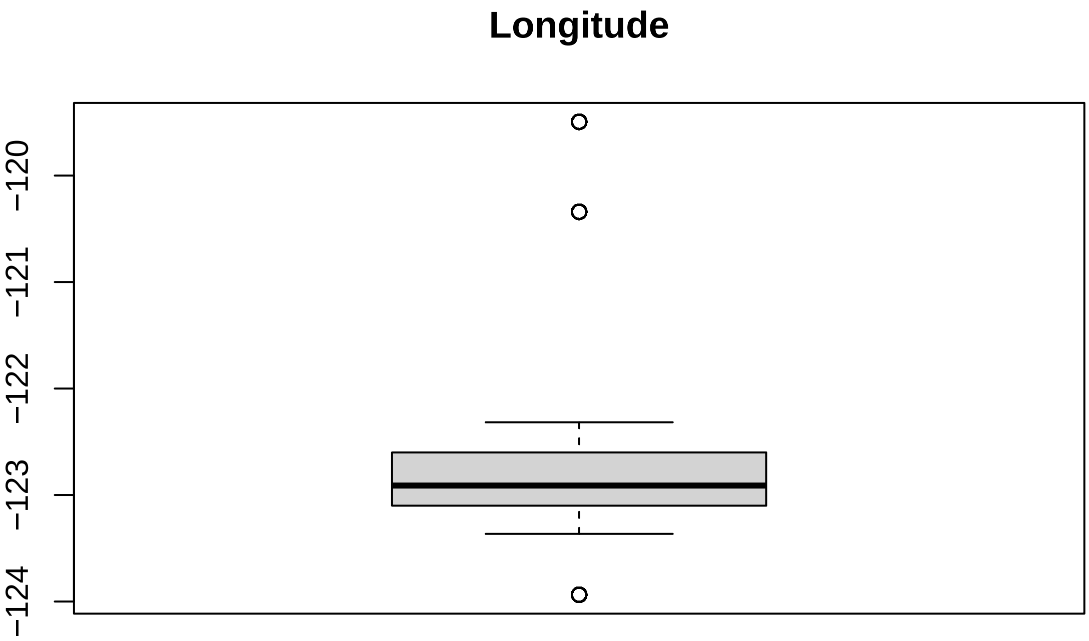
```
## Number_Baths outliers:[1] 1153
```



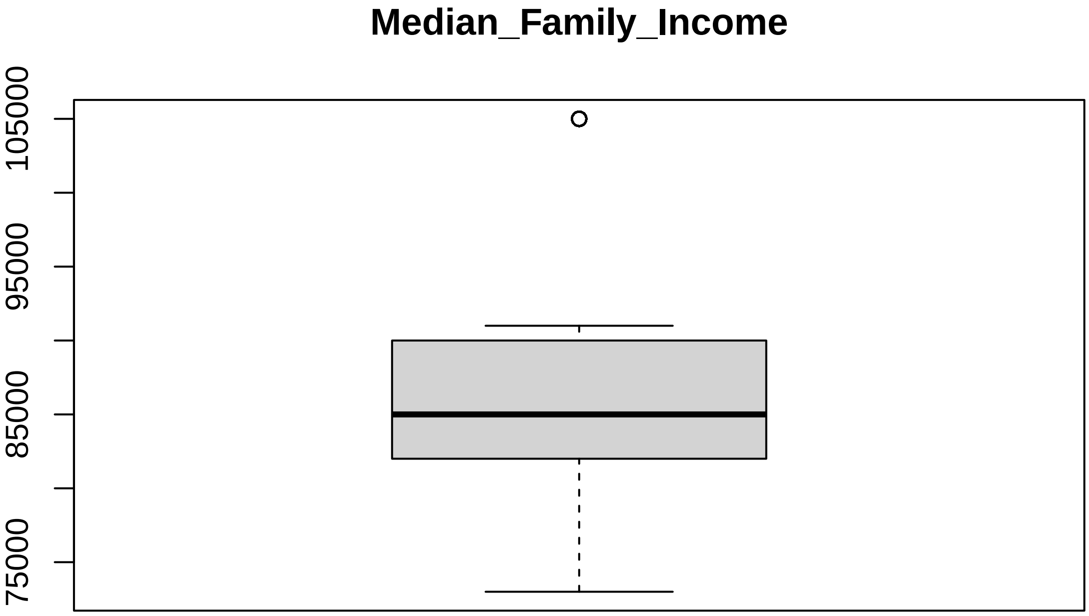
```
## Population outliers:[1] 1328
```



```
## Latitude outliers:[1] 3354
```



```
## Longitude outliers:[1] 2401
```



```
## Median_Family_Income outliers:[1] 1282
```

- method to remove outliers until it is 0 *remove_outliers <- function(data, col){
- len_outlier <- length(boxplot.stats(data[[col]])$out)
- if(len_outlier > 0){
- data <- data %>% filter(!data[[col]] %in% boxplot.stats(data[[col]])$out)
- remove_outliers(data, col)
- }
- else{
- return(data)
- } *}

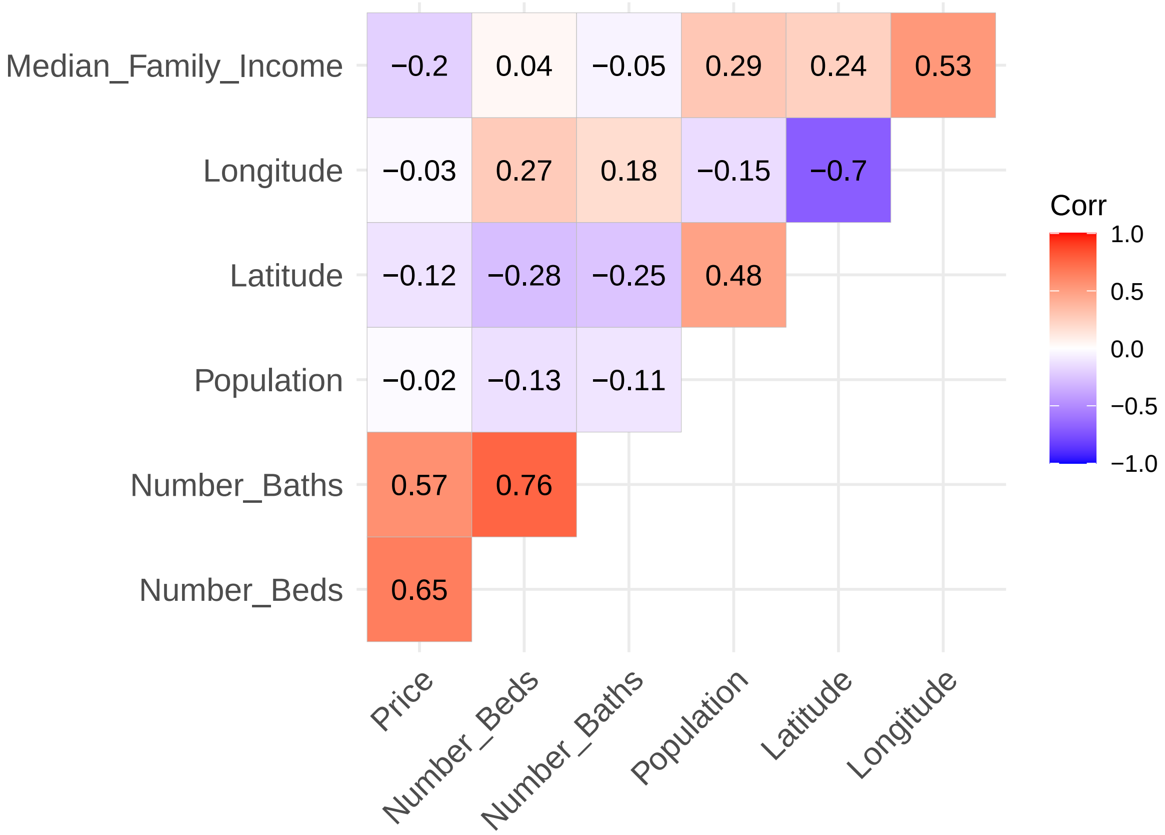
```
for(col in unique(names(bcHousing_numCols))){
  bcHousing_numCols <- bcHousing_numCols %>% filter(!bcHousing_numCols[[col]] %in% boxplot.stats(bcHous
    #remove_outliers(bcHousing_numCols, col)
  cat(col, "outliers:", length(boxplot.stats(bcHousing_numCols[[col]])$out), "\n")
}
```

```
## Price outliers: 204
## Number_Beds outliers: 0
## Number_Baths outliers: 0
## Population outliers: 0
## Latitude outliers: 0
## Longitude outliers: 0
## Median_Family_Income outliers: 1484
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.4.1

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.4.1
```

```
cor_matrix <- cor(bcHousing_numCols)
ggcorrplot(cor_matrix, lab = TRUE, type = "upper")
```



```
full_model <- lm(Price ~ ., data = bcHousing_numCols)
summary(full_model)
```

```
##
## Call:
## lm(formula = Price ~ ., data = bcHousing_numCols)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1415257  -233753   -70850   145551  2477183
##
## Coefficients: (1 not defined because of singularities)
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -4.818e+07  4.053e+06 -11.886   <2e-16 ***
## Number_Beds           2.359e+05  7.104e+03  33.208   <2e-16 ***
## Number_Baths          9.004e+04  1.048e+04   8.591   <2e-16 ***
## Population            1.073e+00  1.102e-01   9.739   <2e-16 ***
## Latitude             -1.394e+06  1.076e+05 -12.958   <2e-16 ***
## Longitude            -9.513e+05  4.280e+04 -22.225   <2e-16 ***
## Median_Family_Income         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 408800 on 3906 degrees of freedom
## Multiple R-squared:  0.507,  Adjusted R-squared:  0.5064
## F-statistic: 803.5 on 5 and 3906 DF,  p-value: < 2.2e-16
```

```r
simple_model <- lm(Price ~ 1, data = bcHousing_numCols)
summary(simple_model)
```

```
##
## Call:
## lm(formula = Price ~ 1, data = bcHousing_numCols)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -981511 -462411 -179961  386589 1888589
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1111411       9304   119.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 581900 on 3911 degrees of freedom
```

```r
both_select <- step(full_model, direction = "both")
```

```
## Start:  AIC=101100.4
## Price ~ Number_Beds + Number_Baths + Population + Latitude +
##     Longitude + Median_Family_Income
##
##
## Step:  AIC=101100.4
## Price ~ Number_Beds + Number_Baths + Population + Latitude +
##     Longitude
##
##                Df  Sum of Sq        RSS    AIC
## <none>                        6.5286e+14 101100
## - Number_Baths  1 1.2337e+13 6.6520e+14 101172
## - Population    1 1.5852e+13 6.6872e+14 101192
## - Latitude      1 2.8064e+13 6.8093e+14 101263
## - Longitude     1 8.2563e+13 7.3543e+14 101564
## - Number_Beds   1 1.8432e+14 8.3718e+14 102071
```

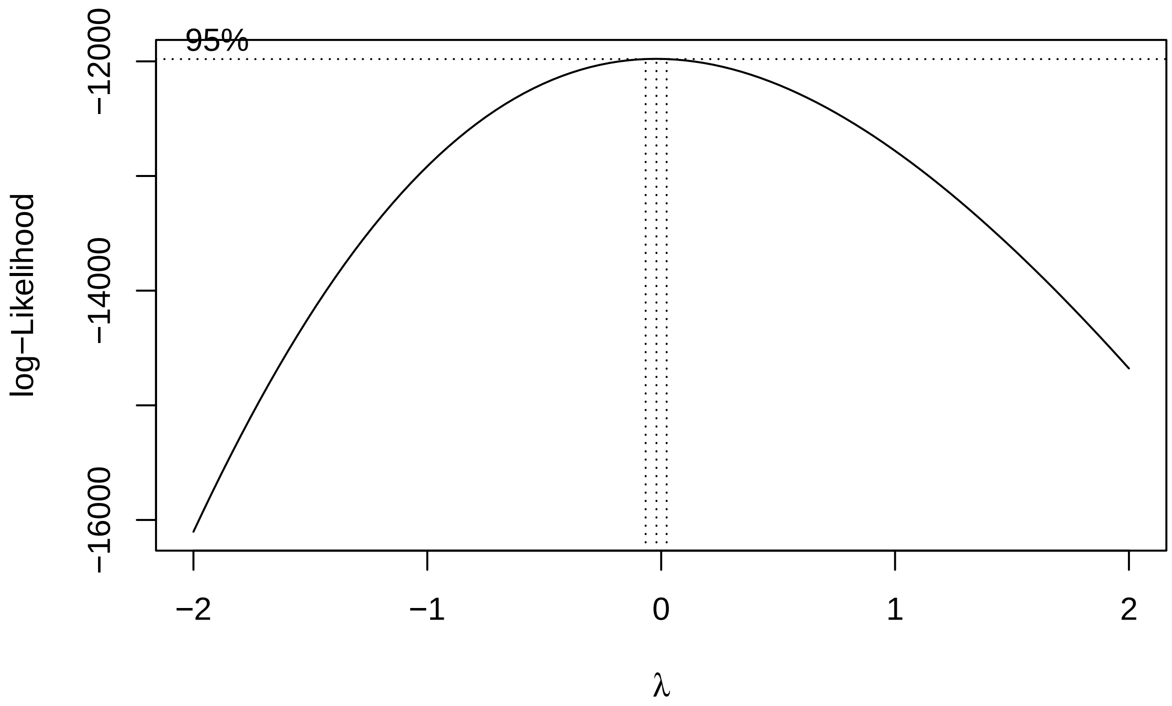
```
summary(both_select)
```

```
##
## Call:
## lm(formula = Price ~ Number_Beds + Number_Baths + Population +
##     Latitude + Longitude, data = bcHousing_numCols)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1415257  -233753   -70850   145551  2477183
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.818e+07  4.053e+06 -11.886   <2e-16 ***
## Number_Beds   2.359e+05  7.104e+03  33.208   <2e-16 ***
## Number_Baths  9.004e+04  1.048e+04   8.591   <2e-16 ***
## Population    1.073e+00  1.102e-01   9.739   <2e-16 ***
## Latitude     -1.394e+06  1.076e+05 -12.958   <2e-16 ***
## Longitude    -9.513e+05  4.280e+04 -22.225   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 408800 on 3906 degrees of freedom
## Multiple R-squared:  0.507,  Adjusted R-squared:  0.5064
## F-statistic: 803.5 on 5 and 3906 DF,  p-value: < 2.2e-16
```

```
model <- lm(formula = Price ~ Number_Beds + Number_Baths + Population +
    Latitude + Longitude, data = bcHousing_numCols)

library(MASS)
```

```
##
## Attaching package: ’MASS’

## The following object is masked _by_ ’.GlobalEnv’:
##
##     housing

## The following object is masked from ’package:dplyr’:
##
##     select
```

```
boxncox <- boxcox(model)
```



```
optimal <- boxncox$x[which.max(boxncox$y)]
optimal
```

```
## [1] -0.02020202
```

```
bcHousing_numCols$new_Price <- (bcHousing_numCols$Price^optimal)/optimal
new_model <- lm(new_Price ~ Number_Baths + Longitude + Latitude  + Population +
   Number_Beds, data = bcHousing_numCols)
summary(new_model)
```

```
##
## Call:
## lm(formula = new_Price ~ Number_Baths + Longitude + Latitude +
##     Population + Number_Beds, data = bcHousing_numCols)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43893 -0.16603 -0.02299  0.13763  1.27706
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.629e+01  2.566e+00  -29.74   <2e-16 ***
## Number_Baths  8.171e-02  6.634e-03   12.32   <2e-16 ***
## Longitude    -6.384e-01  2.709e-02  -23.56   <2e-16 ***
## Latitude     -8.196e-01  6.810e-02  -12.04   <2e-16 ***
## Population    7.234e-07  6.973e-08   10.37   <2e-16 ***
## Number_Beds   1.582e-01  4.496e-03   35.19   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2588 on 3906 degrees of freedom
## Multiple R-squared:  0.562,  Adjusted R-squared:  0.5614
## F-statistic:  1002 on 5 and 3906 DF,  p-value: < 2.2e-16
```