AAD Assignment 1 - Group 26

Omar, Eloise, Alina, Sue

2025-04-06

Content:

1 Descriptive analysis of the data set 1.1 Data loading and cleaning 1.2 Preliminary analysis on the data

2 Multiple linear regression 2.1 Initial MLR model using all appropriate predictors 2.2 Discussion of initial Model 2.3 Checking issues in the initial model 2.3.1 Outliers 2.3.2 High leverage points 2.3.3 Collinearity 2.4 Model improvements 2.5 Three most significant variables

3 Model Performance: Comparison between initial model and the improved one 3.1 Training MSE 3.1.1 Training MSE of the initial model 3.1.2 Training MSE of the improved model 3.2 Estimating testing error using the 80-20 split validation set approach 3.3 Estimating testing error using 5-fold cross validation 3.4 Estimating testing error using LOOCV 3.5 Conclusion: whether the final MLR model better than the initial

4 Prediction competition

1.1 Descriptive Analysis of the Data Set

```
# Load the data
data <- read.csv("Assignt1_data.csv", header = TRUE)

# Summary Statistics of the Original Data
str(data)
```

```
## 'data.frame':    18640 obs. of  10 variables:
##  $ id              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ longitude       : num  -122 -122 -122 -122 -122 ...
##  $ latitude        : num  37.9 37.9 37.9 37.9 37.9 ...
##  $ housingMedianAge: int  41 21 52 52 52 52 52 52 42 52 ...
##  $ aveRooms        : num  6.98 6.24 8.29 5.82 6.28 ...
##  $ aveBedrooms     : num  1.024 0.972 1.073 1.073 1.081 ...
##  $ population      : int  322 2401 496 558 565 413 1094 1157 1206 1551 ...
##  $ medianIncome    : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ medianHouseValue: num  452600 358500 352100 341300 342200 ...
##  $ oceanProximity  : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
```

```
head(data)
```

```
##   id longitude latitude housingMedianAge aveRooms aveBedrooms population
## 1  1   -122.23    37.88              41 6.984127   1.0238095        322
## 2  2   -122.22    37.86              21 6.238137   0.9718805       2401
## 3  3   -122.24    37.85              52 8.288136   1.0734463        496
## 4  4   -122.25    37.85              52 5.817352   1.0730594        558
## 5  5   -122.25    37.85              52 6.281853   1.0810811        565
## 6  6   -122.25    37.85              52 4.761658   1.1036269        413
##   medianIncome medianHouseValue oceanProximity
## 1       8.3252           452600       NEAR BAY
## 2       8.3014           358500       NEAR BAY
## 3       7.2574           352100       NEAR BAY
## 4       5.6431           341300       NEAR BAY
## 5       3.8462           342200       NEAR BAY
## 6       4.0368           269700       NEAR BAY
```

```
dim(data)
```

```
## [1] 18640    10
```

```
summary(data)
```

```
##        id          longitude         latitude     housingMedianAge
##  Min.   :    1   Min.   :-124.3   Min.   :32.55   Min.   : 1.00
##  1st Qu.: 5176   1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00
##  Median :10334   Median :-118.5   Median :34.26   Median :29.00
##  Mean   :10334   Mean   :-119.6   Mean   :35.63   Mean   :28.61
##  3rd Qu.:15502   3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00
##  Max.   :20640   Max.   :-114.3   Max.   :41.95   Max.   :52.00
##
##     aveRooms         aveBedrooms       population     medianIncome
##  Min.   :  0.8461   Min.   : 0.375   Min.   :    3   Min.   : 0.4999
##  1st Qu.:  4.4409   1st Qu.: 1.006   1st Qu.:  785   1st Qu.: 2.5668
##  Median :  5.2346   Median : 1.049   Median : 1167   Median : 3.5421
##  Mean   :  5.4370   Mean   : 1.097   Mean   : 1427   Mean   : 3.8798
##  3rd Qu.:  6.0599   3rd Qu.: 1.099   3rd Qu.: 1726   3rd Qu.: 4.7601
##  Max.   :141.9091   Max.   :34.067   Max.   :35682   Max.   :15.0001
##                     NA's   :190
##  medianHouseValue oceanProximity
##  Min.   : 14999   Length:18640
##  1st Qu.:119900   Class :character
##  Median :179900   Mode  :character
##  Mean   :207242
##  3rd Qu.:265600
##  Max.   :500001
##
```

```
# Clean the data
colSums(is.na(data))  # Count NAs in each column
```

```
##                id         longitude          latitude housingMedianAge
##                 0                 0                 0                 0
##        aveRooms      aveBedrooms        population      medianIncome
```

```
##                0              190                0                0
## medianHouseValue   oceanProximity
##                0                0
```

```
# 190 out of 18640 rows with na values in aveBedrooms
clean_data <- na.omit(data) # remove na rows

# Summary Stats of the Cleaned Data
str(clean_data)
```

```
## 'data.frame':    18450 obs. of  10 variables:
##  $ id              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ longitude       : num  -122 -122 -122 -122 -122 ...
##  $ latitude        : num  37.9 37.9 37.9 37.9 37.9 ...
##  $ housingMedianAge: int  41 21 52 52 52 52 52 52 42 52 ...
##  $ aveRooms        : num  6.98 6.24 8.29 5.82 6.28 ...
##  $ aveBedrooms     : num  1.024 0.972 1.073 1.073 1.081 ...
##  $ population      : int  322 2401 496 558 565 413 1094 1157 1206 1551 ...
##  $ medianIncome    : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ medianHouseValue: num  452600 358500 352100 341300 342200 ...
##  $ oceanProximity  : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
##  - attr(*, "na.action")= 'omit' Named int [1:190] 258 305 486 509 630 668 988 1214 1310 1447 ...
##   ..- attr(*, "names")= chr [1:190] "258" "305" "486" "509" ...
```

```
head(clean_data)
```

```
##   id longitude latitude housingMedianAge aveRooms aveBedrooms population
## 1  1   -122.23    37.88                41 6.984127   1.0238095        322
## 2  2   -122.22    37.86                21 6.238137   0.9718805       2401
## 3  3   -122.24    37.85                52 8.288136   1.0734463        496
## 4  4   -122.25    37.85                52 5.817352   1.0730594        558
## 5  5   -122.25    37.85                52 6.281853   1.0810811        565
## 6  6   -122.25    37.85                52 4.761658   1.1036269        413
##   medianIncome medianHouseValue oceanProximity
## 1       8.3252           452600       NEAR BAY
## 2       8.3014           358500       NEAR BAY
## 3       7.2574           352100       NEAR BAY
## 4       5.6431           341300       NEAR BAY
## 5       3.8462           342200       NEAR BAY
## 6       4.0368           269700       NEAR BAY
```

```
dim(clean_data)
```

```
## [1] 18450    10
```

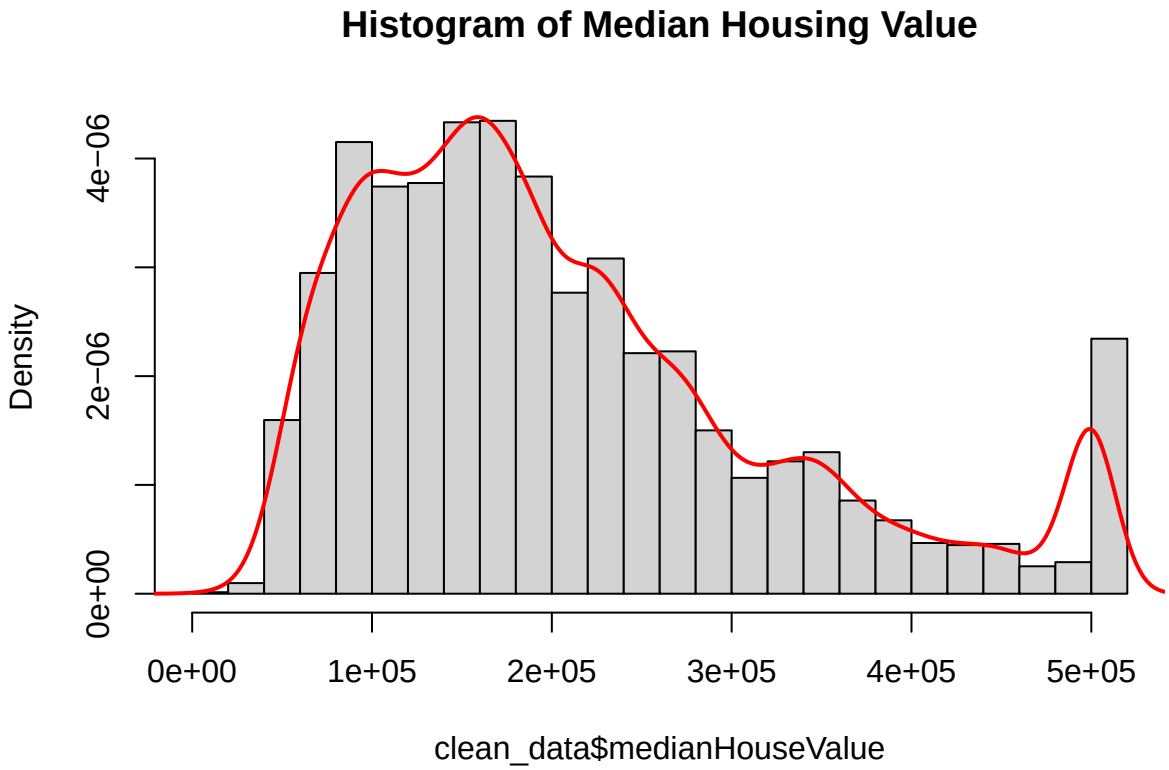
```
summary(clean_data)
```

```
##        id          longitude         latitude     housingMedianAge
##  Min.   :    1   Min.   :-124.3   Min.   :32.55   Min.   : 1.00
##  1st Qu.: 5180   1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00
##  Median :10332   Median :-118.5   Median :34.26   Median :29.00
```

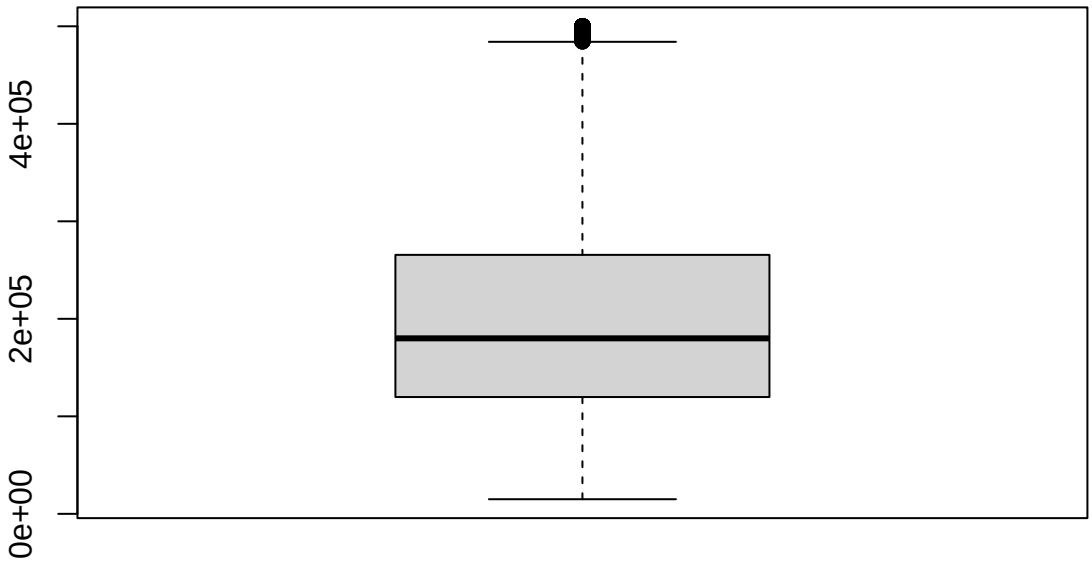
```
##  Mean   :10332   Mean   :-119.6   Mean   :35.63   Mean   :28.61
##  3rd Qu.:15498   3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00
##  Max.   :20640   Max.   :-114.3   Max.   :41.95   Max.   :52.00
##     aveRooms          aveBedrooms       population       medianIncome
##  Min.   :  0.8461   Min.   : 0.375   Min.   :    3.0   Min.   : 0.4999
##  1st Qu.:  4.4411   1st Qu.: 1.006   1st Qu.:  785.2   1st Qu.: 2.5669
##  Median :  5.2359   Median : 1.049   Median : 1167.0   Median : 3.5446
##  Mean   :  5.4392   Mean   : 1.097   Mean   : 1426.1   Mean   : 3.8803
##  3rd Qu.:  6.0599   3rd Qu.: 1.099   3rd Qu.: 1724.0   3rd Qu.: 4.7608
##  Max.   :141.9091   Max.   :34.067   Max.   :35682.0   Max.   :15.0001
##  medianHouseValue oceanProximity
##  Min.   : 14999   Length:18450
##  1st Qu.:119800   Class :character
##  Median :180000   Mode  :character
##  Mean   :207277
##  3rd Qu.:265600
##  Max.   :500001
```

Histogram of the Response Variable

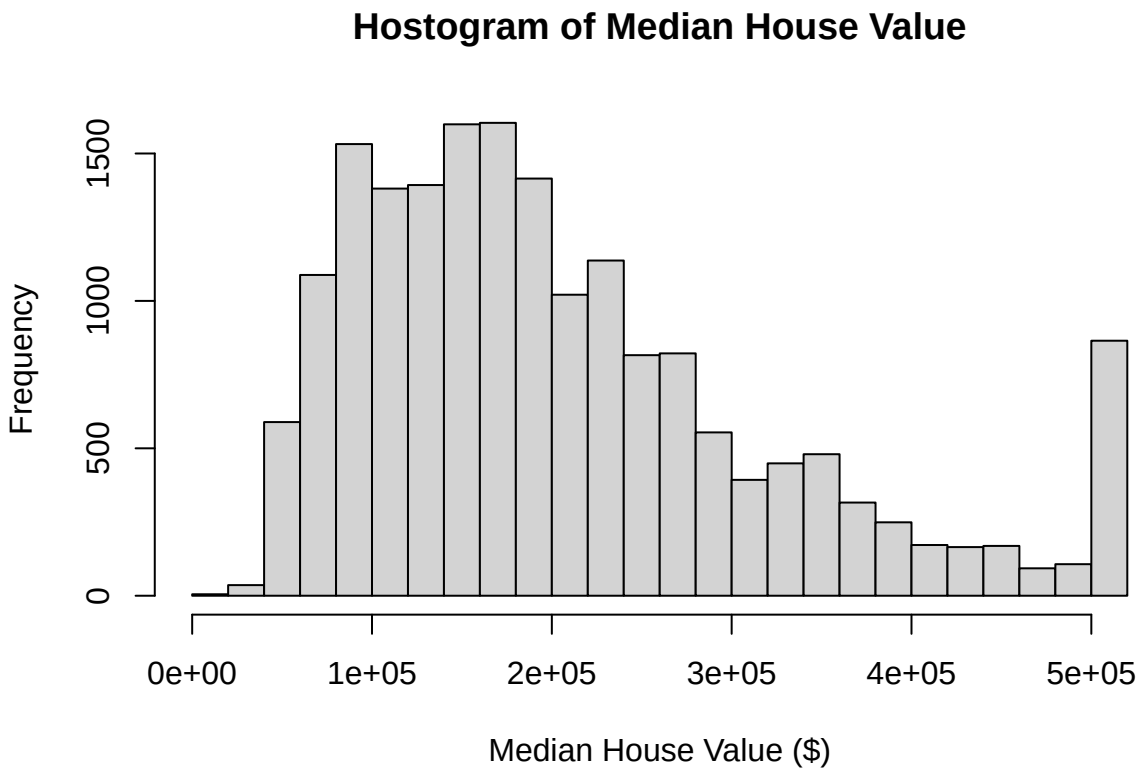
```
hist(clean_data$medianHouseValue,
     breaks = 30,
     freq = FALSE,
     main = "Histogram of Median Housing Value")
lines(density(clean_data$medianHouseValue), col = "red", lwd = 2)
```



```r
boxplot(clean_data$medianHouseValue)
```




```r
hist(clean_data$medianHouseValue,
     xlab = "Median House Value ($)",
     ylab = "Frequency",
     main = "Hostogram of Median House Value",
     breaks = 25)
```



The boxplot clearly indicates that the Median House Value is right-skewed, with a longer tail on the higher end. Additionally, there are several outliers present at the upper extreme of the distribution -

Correlations

From the correlation matrix plot we can see that among all the predictors, Median House Value is most correlated with Median Income. This is pretty intuitive, as we expect high-income households purchasing more expensive houses.

There is little correlation between Median Housing Value and Median Income.

Longitude and Latitude are highly negatively correlated, while Average Rooms and Average Bedrooms are highly positively correlated. This multicollinearity can lead to issues such as unstable coefficient estimates and inflated standard errors, as the model struggles to distinguish the individual effects of highly correlated predictors. Consequently, the interpretability of the regression results is compromised.To mitigate this, we can remove one of the correlated variables, which will be discussed later, in section 2.2.

- Note that ID is not a relevant predictor, so we don't care about the correlation between ID and other variables.

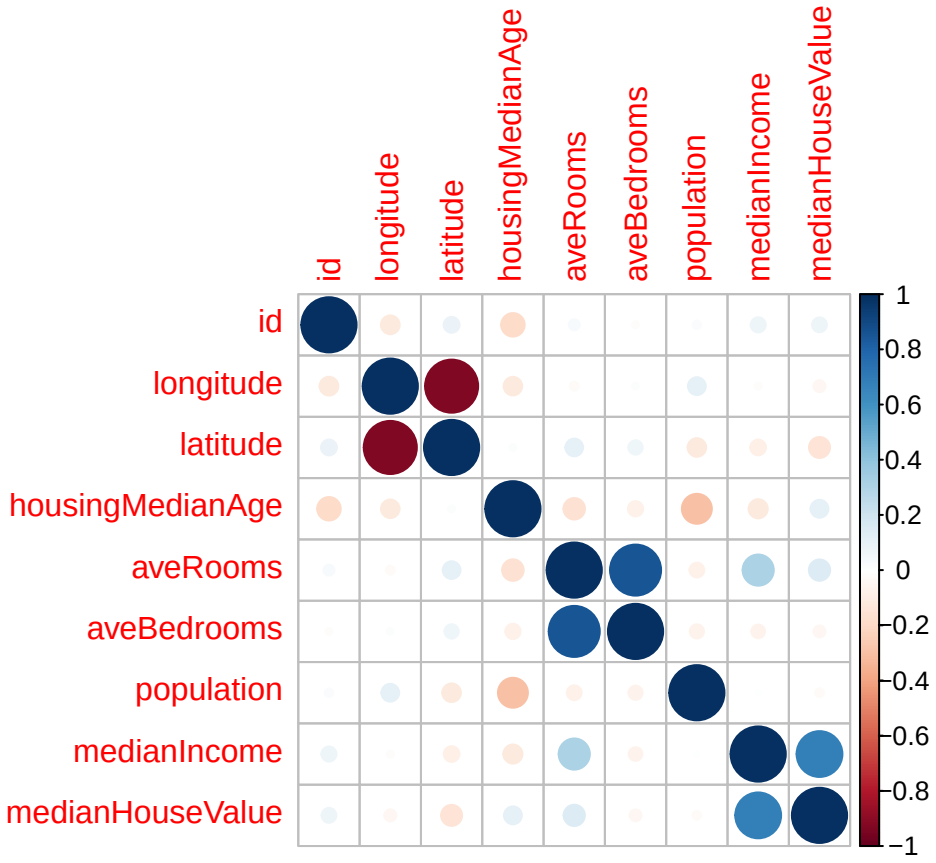
```
numeric.data <-  clean_data[sapply(clean_data, is.numeric)]
head(numeric.data)
```

```
##   id longitude latitude housingMedianAge aveRooms aveBedrooms population
## 1  1   -122.23    37.88              41 6.984127   1.0238095        322
## 2  2   -122.22    37.86              21 6.238137   0.9718805       2401
## 3  3   -122.24    37.85              52 8.288136   1.0734463        496
## 4  4   -122.25    37.85              52 5.817352   1.0730594        558
## 5  5   -122.25    37.85              52 6.281853   1.0810811        565
## 6  6   -122.25    37.85              52 4.761658   1.1036269        413
##   medianIncome medianHouseValue
## 1       8.3252           452600
## 2       8.3014           358500
## 3       7.2574           352100
## 4       5.6431           341300
## 5       3.8462           342200
## 6       4.0368           269700
```

```
cor_medianHouseValue <- cor(clean_data$medianHouseValue,
                            clean_data[, c("longitude",
                                           "latitude",
                                           "housingMedianAge",
                                           "aveRooms",
                                           "aveBedrooms",
                                           "population",
                                           "medianIncome")])
print(cor_medianHouseValue)
```

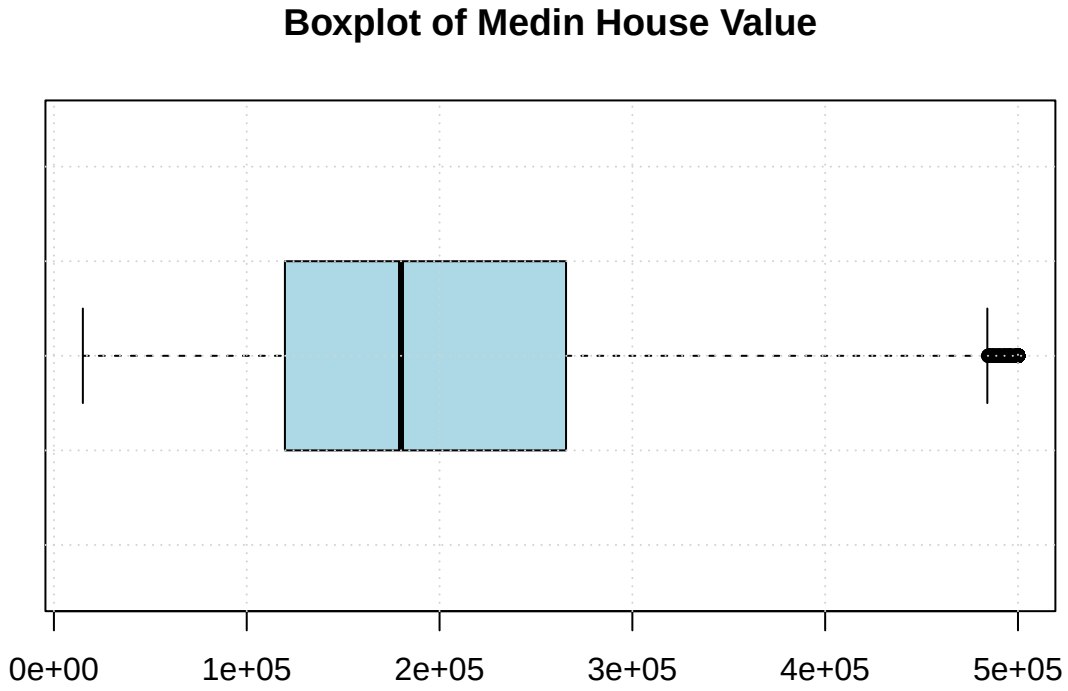
```
##        longitude   latitude housingMedianAge aveRooms aveBedrooms  population
## [1,] -0.04646254 -0.1439446        0.1052129 0.148625 -0.04545692 -0.02394206
##      medianIncome
## [1,]     0.689536
```

```
corrplot(cor(clean_data[, sapply(clean_data, is.numeric)]), method = "circle")
```



Outliers Detection

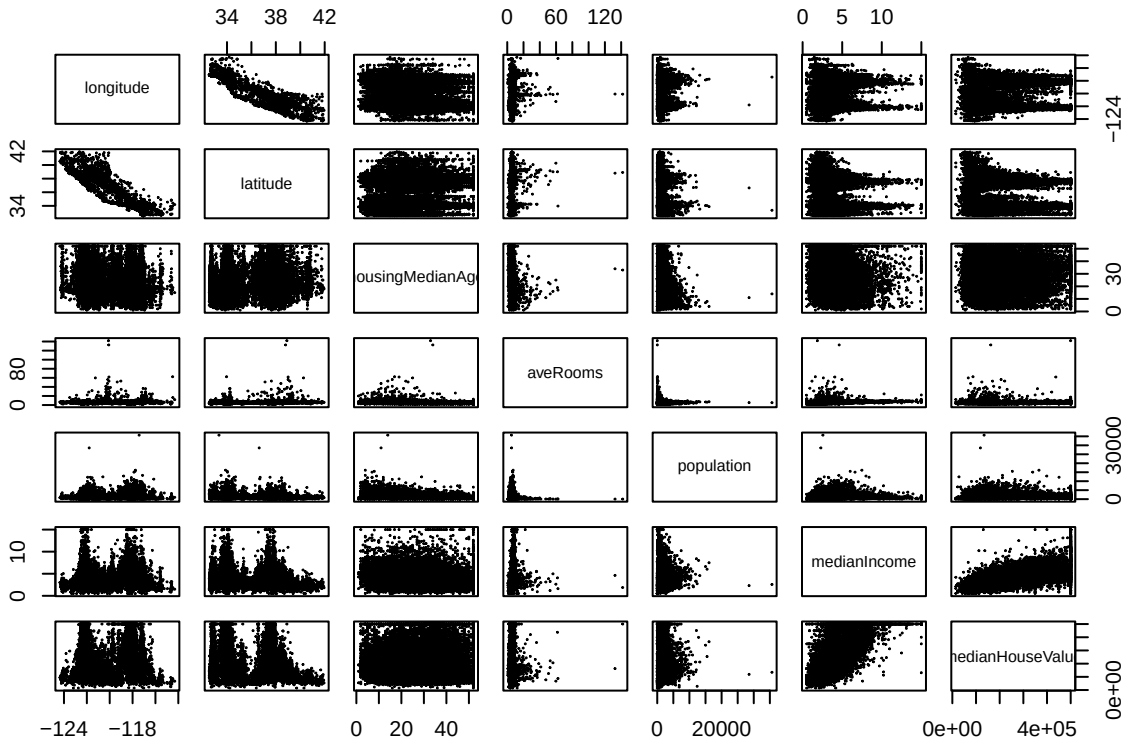
```
boxplot_medianHouseValue <- boxplot(clean_data$medianHouseValue,
        main = "Boxplot of Medin House Value",
        outline = TRUE,
        col = "lightblue",
        horizontal = TRUE,
        cex = 1,
        col.out = "red",
        pch = 16)
grid()
```




```
outliers <- boxplot_medianHouseValue$out
num_outliers <- length(outliers)
print(paste("Number of outliers:", num_outliers))
```

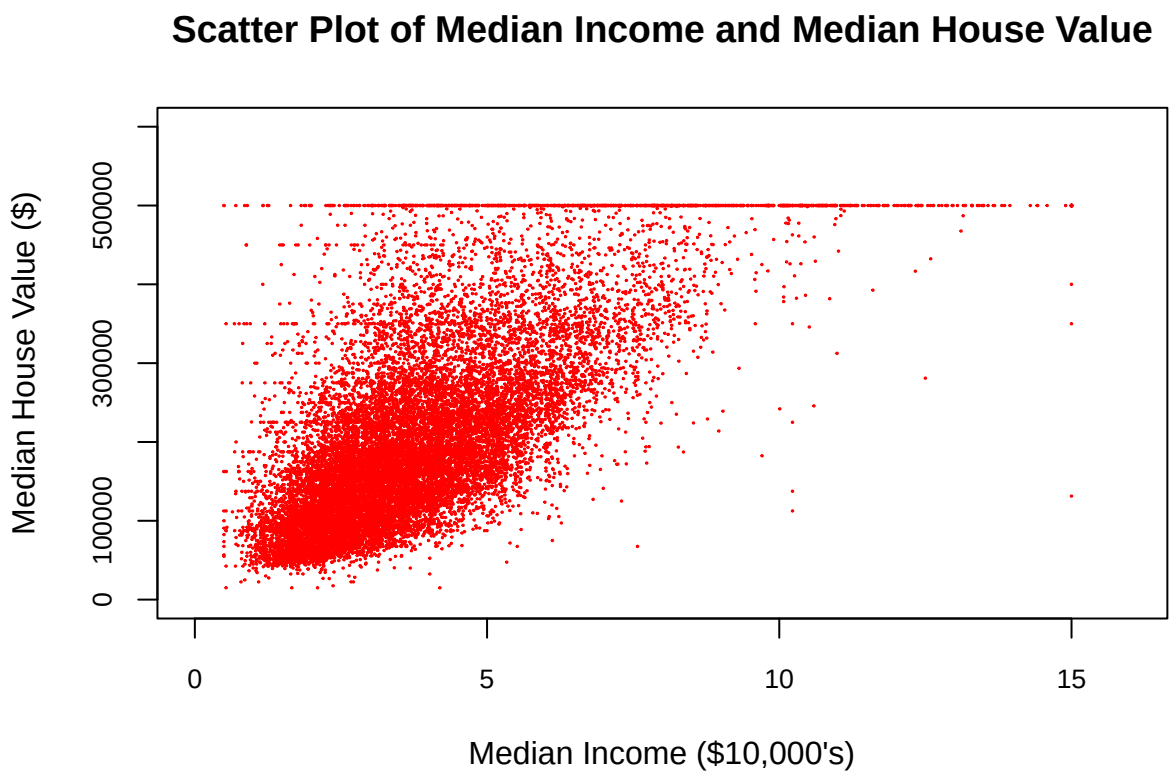
```
## [1] "Number of outliers: 952"
```

Pairwise Plots



More Plots: Median House Value with Median Income

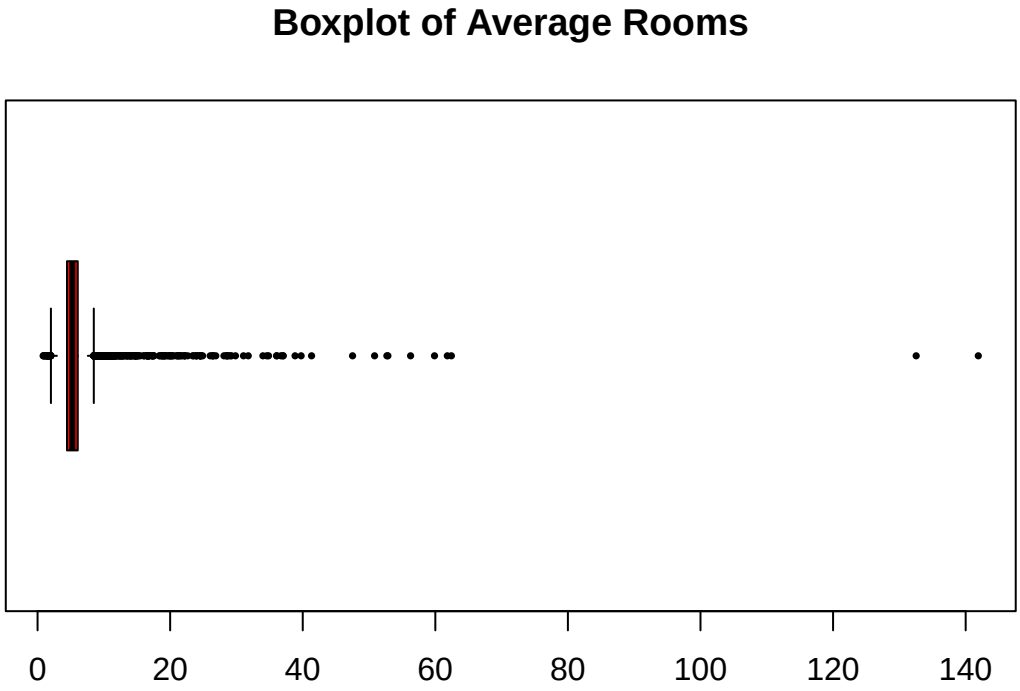
```
options(scipen = 999)
plot(clean_data$medianIncome, clean_data$medianHouseValue,
     xlab = "Median Income ($10,000's)", ylab = "Median House Value ($)",
     main = "Scatter Plot of Median Income and Median House Value",
     cex = 0.1,
     col = "red",
     xlim = c(0, 16),
     ylim = c(0, 600000),
     cex.axis = 0.8)
```



###

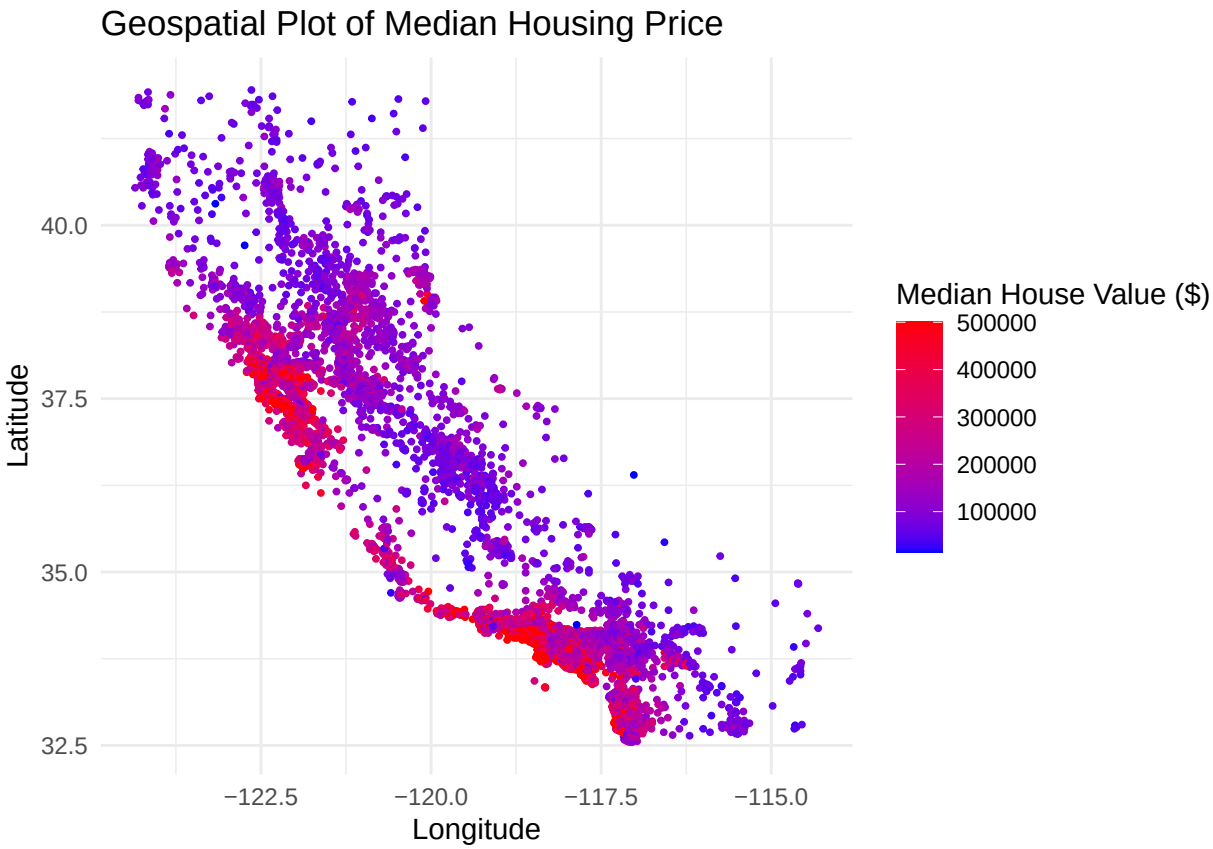
More Plots: Boxplot of Average Rooms Some extreme outliers here

```
boxplot(data$aveRooms,
        main = "Boxplot of Average Rooms",
        cex = 0.5,
        pch = 16,
        col = "red",
        horizontal = TRUE)
```



Geospacial Plot: Median House Value

```
ggplot(clean_data, aes(x = clean_data$longitude, y = clean_data$latitude)) +
  geom_point(aes(color = clean_data$medianHouseValue), size = 0.7) +
  scale_color_gradient(low = "blue", high = "red",
                       name = "Median House Value ($)") +
  theme_minimal() +
  ggtitle("Geospatial Plot of Median Housing Price") +
  labs(x = "Longitude", y = "Latitude")
```




```
# Longitude represents east-west position (-180: west, 180: east)
# Latitude represents north-south position  (-90: south, 90:north)
```

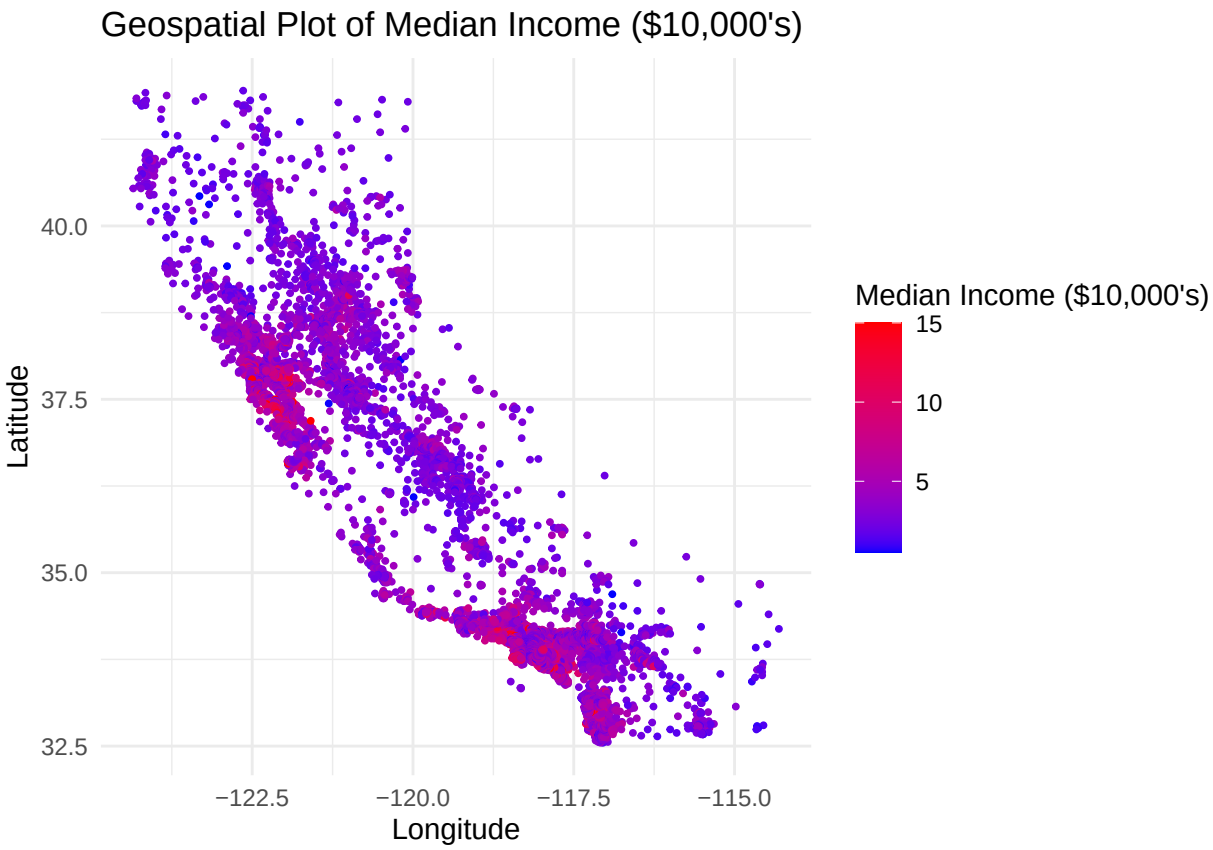
At a glance, median housing value is higher in the South than in the North.

There are some observable clusters. For example, the region around (-122.5, 37.5) and the region around (-118.75, 33.75) has highest density of high median housing values. These two regions has pretty high data point desity to begin with - thats why we also observe higher density of high value houses.

Geospacial Plot: Median Income

expect to see similar pattern as above

```
ggplot(clean_data, aes(x = clean_data$longitude, y = clean_data$latitude)) +
  geom_point(aes(color = clean_data$medianIncome), size = 0.7) +
  scale_color_gradient(low = "blue", high = "red",
                       name = "Median Income ($10,000's)") +
  theme_minimal() +
  ggtitle("Geospatial Plot of Median Income ($10,000's)") +
  labs(x = "Longitude", y = "Latitude")
```



2 Multiple Linear Regression

2.1 Initial MLR model using all appropriate predictors

We will be using the all the given predictors except for id, since it is irrelevant.

```
fit1 <- lm(clean_data$medianHouseValue ~. - id, data = clean_data)
summary(fit1)
```

```
##
## Call:
## lm(formula = clean_data$medianHouseValue ~ . - id, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -632258  -45300  -11782   30245  443819
##
## Coefficients:
##                               Estimate    Std. Error t value
## (Intercept)              -2271786.7538    97592.8298 -23.278
## longitude                  -26547.2849     1130.9763 -23.473
## latitude                   -24733.2245     1119.0102 -22.103
```

```
## housingMedianAge              820.5384       47.6087  17.235
## aveRooms                    -8784.2664      621.8502 -14.126
## aveBedrooms                 53381.3487     2972.9142  17.956
## population                     -0.6741        0.4942  -1.364
## medianIncome                42087.3726      447.0610  94.142
## oceanProximityINLAND       -38448.2476     1931.4427 -19.906
## oceanProximityISLAND       128990.5763    35857.5037   3.597
## oceanProximityNEAR BAY       4058.8488     2089.7062   1.942
## oceanProximityNEAR OCEAN     8919.4039     1713.9652   5.204
##                                    Pr(>|t|)
## (Intercept)              < 0.0000000000000002 ***
## longitude                < 0.0000000000000002 ***
## latitude                 < 0.0000000000000002 ***
## housingMedianAge         < 0.0000000000000002 ***
## aveRooms                 < 0.0000000000000002 ***
## aveBedrooms              < 0.0000000000000002 ***
## population                          0.172593
## medianIncome             < 0.0000000000000002 ***
## oceanProximityINLAND     < 0.0000000000000002 ***
## oceanProximityISLAND                0.000322 ***
## oceanProximityNEAR BAY              0.052115 .
## oceanProximityNEAR OCEAN           0.000000197 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71650 on 18438 degrees of freedom
## Multiple R-squared:  0.6168, Adjusted R-squared:  0.6166
## F-statistic:  2698 on 11 and 18438 DF,  p-value: < 0.00000000000000022
```

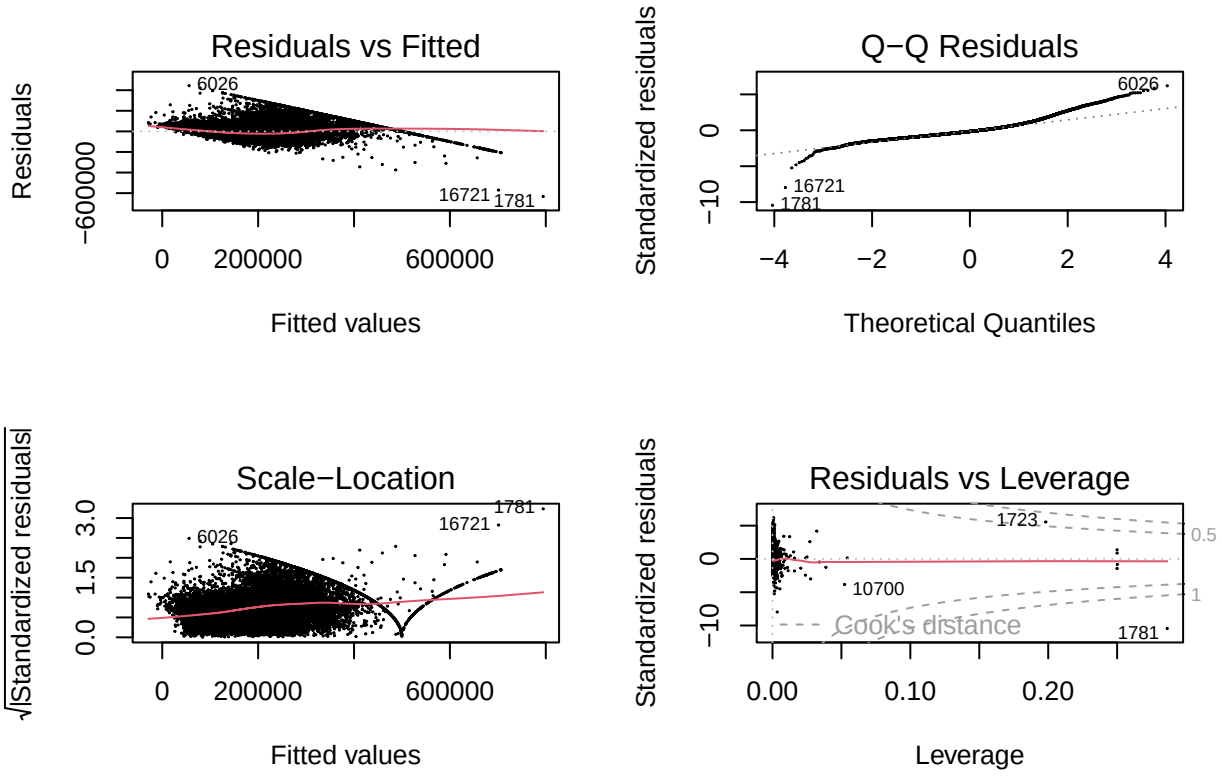
2.2 Discussion of the initial model

The p-value of the F-test is practically zero, therefore we have sufficient evidence to reject the null hypothesis that all regression coefficients are zero. This means that the model has overall significance, and that at least one of the predictors are useful in explaning the variation in the response variable (i.e. medianHouseValue).

Having concluded that this model (fit1_MHV) has overall significance, we can gauge the significance of individual predictors through the t-test p-values. It is clear that population is not useful, as we do not have sufficient evidence to reject the null hypothesis that its corresponding coefficient is non-zero. The dummy variable, "NEAR BAY", associated to the categorical predictor, "oceanProximity", is also shown to be insignificant. This might be an indication that comparing to the base case (<1H OCEAN), NEAR BAY does not lead to significant change in Median House Price.

2.3 Checking issues in the initial model

```
par(mfrow = c(2,2))
plot(fit1, cex = 0.1)
```



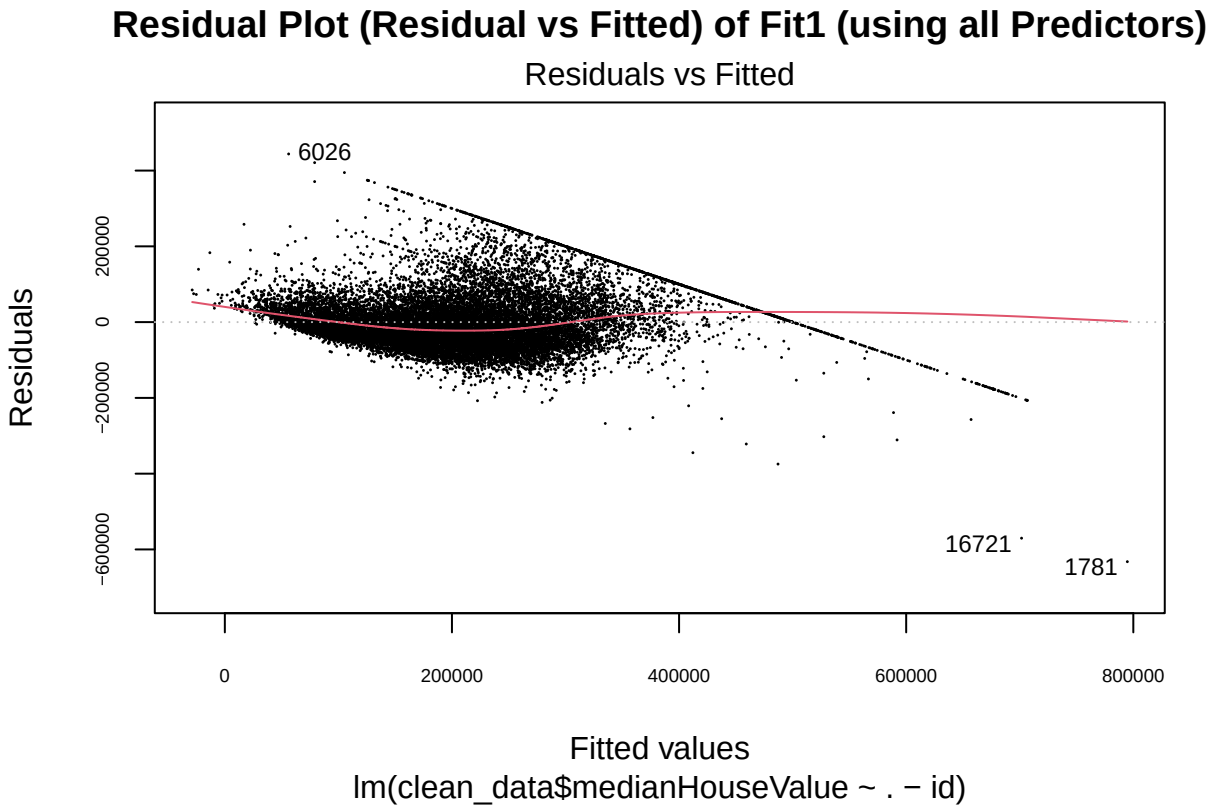
Residual Plot:

The residual plot shows that the points are not quite randomly scattered around zero. This means that the homoskedasiticity and linear assumption might be questionable. To address non-linearity, we can consider polynomial terms or interactions.

The funnel shape (the spread of residuals seems to increase as the fitted values increase) within the fitted value range (0, 400000) strengthens my doubt of heteroskedasticity. We may consider transforming some of our predictors later.

There are some outliers in the residuals that are far away from zero. These influential points may be high-leverage or outliers or both - should be investigated later.

```
par(mfrow = c(1, 1))
options(scipen = 999)
plot(fit1, which = 1, cex = 0.2, pch = 16, cex.axis = 0.6,
     main = "Residual Plot (Residual vs Fitted) of Fit1 (using all Predictors)")
```



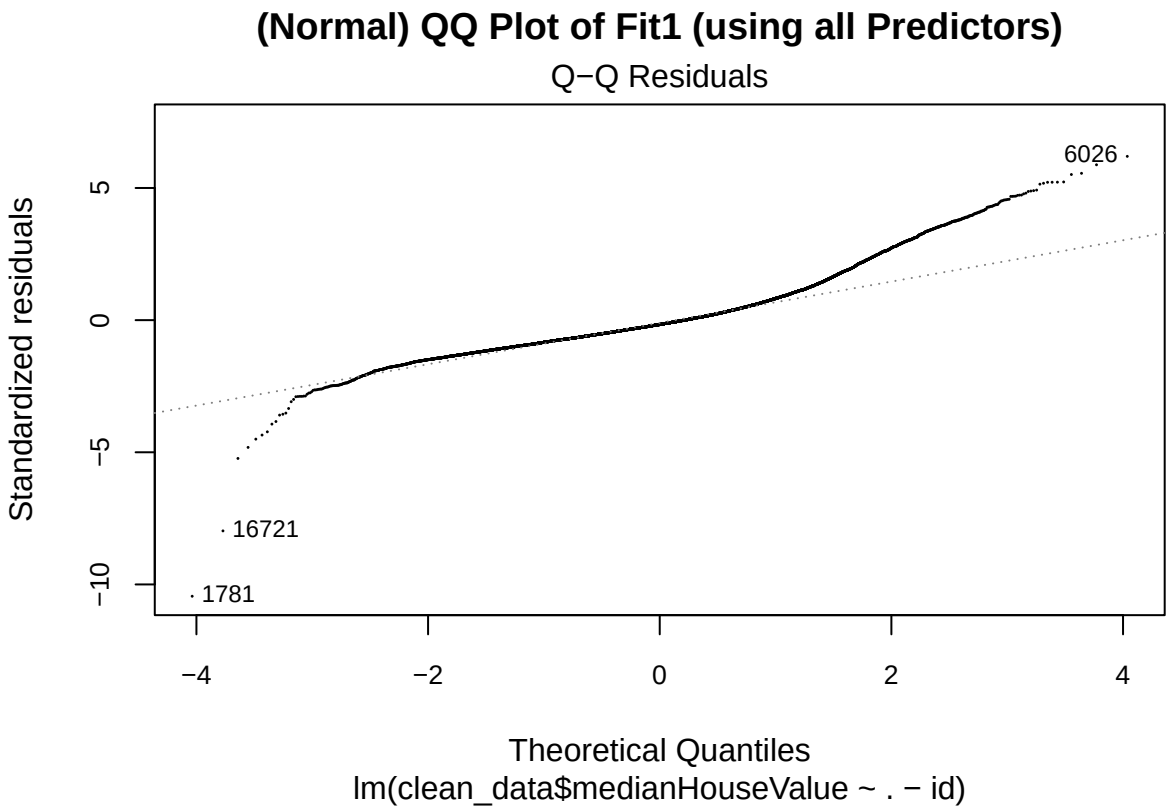
QQ Plot for Residuals:

A key assumption behind generalized linear model is that the error term is normally distributed. This is why t-statistics and F-statistics can be used in our previous testing.

T-stats are robust under some mild deviation from normality, but under extreme non-normality, these statistics become less reliable.

In our plot, the points deviates from the reference line (dashed line) for larger and smaller quantiles (roughly outside of this range: (-2, 1.5)), indicating non-normality (especially high skewness) and influence of outliers especially at the tails.

```
plot(fit1, which = 2, cex = 0.2, pch = 16, cex.axis = 0.8,
     main = "(Normal) QQ Plot of Fit1 (using all Predictors)")
```



Outliers

Outlier identification: as a rule of thumb, we consider those whose studentised residual has a magnitude greater than 3 as outliers here.

However we cannot just simply remove the outliers in this case. This is because these outliers coule be attributable to model specification or other problems. We will see what we can do with the model selection and then can come back to this later.

```
residuals_fit1 <- residuals(fit1)
stdresiduals_fit1 <- rstandard(fit1)
outlier_row_number <- which(abs(stdresiduals_fit1) > 3)

length(outlier_row_number) # gives how many outliers are there
```

```
## [1] 314
```

High Leverage Points

We will compute the leverage statistic hi and to see whether it is » (p + 1)/n. Where p is the number of predictors in the model and n is the sample size.

Having 3,317 high leverage points out of 18,450 data points means that about 18% of the data has high leverage. This indicates that our regression line can change dramatically with small changes in the predictors. One possible reason is that we are overfitting the data - and a reason to this is having too many predictor variables.

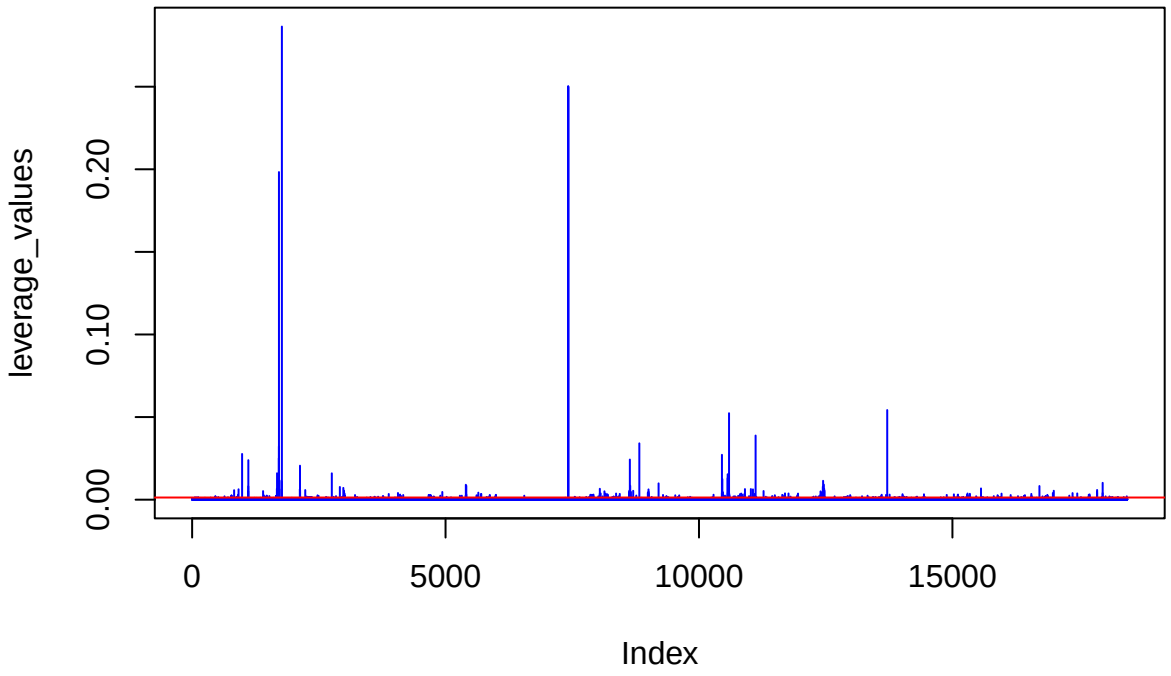
```r
leverage_values <- hatvalues(fit1)

p_lvg <- length(coef(fit1))
n_lvg <- nrow(clean_data)
threshold_lvg <- (p_lvg + 1) / n_lvg

highlvg_row_number <- which(leverage_values > threshold_lvg)
# highlvg_row_number
length(highlvg_row_number)
```

```
## [1] 3317
```

```r
plot(leverage_values, type = "h", col = "blue")
abline(h = 2 * mean(leverage_values), col = "red")
```



Collinearity

The arbitrary threshold of severe collinearity is VIF greater or equal to 5. Here, all the predictors are shown to have a non-severe VIF. However, Longitude and Latitude shows relatively high VIF comparing to other predictors - a cause of this is the high correlation between the two. We may consider using one of them instead of both in the model.

```r
vif(fit1)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## longitude        18.464342  1        4.297015
## latitude         20.529183  1        4.530914
## housingMedianAge  1.295871  1        1.138363
## aveRooms          8.994891  1        2.999148
```

```
## aveBedrooms       7.637752  1        2.763648
## population        1.134318  1        1.065044
## medianIncome      2.610282  1        1.615637
## oceanProximity    4.089926  4        1.192517
```

Making improvements

If we drop population:

```
fit2 <- lm(clean_data$medianHouseValue ~. - id - population, data = clean_data)
summary(fit2)
```

```
##
## Call:
## lm(formula = clean_data$medianHouseValue ~ . - id - population,
##     data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -634757  -45364  -11886   30308  444739
##
## Coefficients:
##                             Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)              -2267609.72    97547.05 -23.246 < 0.0000000000000002
## longitude                  -26463.72     1129.34 -23.433 < 0.0000000000000002
## latitude                   -24619.24     1115.91 -22.062 < 0.0000000000000002
## housingMedianAge              841.06       45.17  18.620 < 0.0000000000000002
## aveRooms                   -8755.75      621.51 -14.088 < 0.0000000000000002
## aveBedrooms                53362.29     2972.95  17.949 < 0.0000000000000002
## medianIncome               42101.44      446.95  94.197 < 0.0000000000000002
## oceanProximityINLAND      -38431.14     1931.45 -19.898 < 0.0000000000000002
## oceanProximityISLAND      129337.97    35857.44   3.607             0.000311
## oceanProximityNEAR BAY      4005.26     2089.39   1.917             0.055259
## oceanProximityNEAR OCEAN    9049.75     1711.34   5.288          0.000000125
##
## (Intercept)              ***
## longitude                ***
## latitude                 ***
## housingMedianAge         ***
## aveRooms                 ***
## aveBedrooms              ***
## medianIncome             ***
## oceanProximityINLAND     ***
## oceanProximityISLAND     ***
## oceanProximityNEAR BAY   .
## oceanProximityNEAR OCEAN ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71650 on 18439 degrees of freedom
## Multiple R-squared:  0.6168, Adjusted R-squared:  0.6166
## F-statistic:  2968 on 10 and 18439 DF,  p-value: < 0.00000000000000022
```

```
# no obvious changes here in terms of t and F test results, Std Error and Estimate
```

Collinearity

If we use Latitude instead of both Longtitude and Latitude:

```
fit3 <- lm(clean_data$medianHouseValue ~. - id - population - longitude,
           data = clean_data)
summary(fit3)
```

```
##
## Call:
## lm(formula = clean_data$medianHouseValue ~ . - id - population -
##     longitude, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -578634  -44947  -11828   29579  471249
##
## Coefficients:
##                            Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                 3958.64   11031.25   0.359                0.7197
## latitude                     545.02     307.83   1.771                0.0767 .
## housingMedianAge             939.04      45.64  20.576 < 0.0000000000000002 ***
## aveRooms                   -10031.37     628.26 -15.967 < 0.0000000000000002 ***
## aveBedrooms                 54353.93    3016.50  18.019 < 0.0000000000000002 ***
## medianIncome                43657.01     448.52  97.337 < 0.0000000000000002 ***
## oceanProximityINLAND       -66615.33    1533.54 -43.439 < 0.0000000000000002 ***
## oceanProximityISLAND       147482.36   36377.95   4.054        0.00005051941 ***
## oceanProximityNEAR BAY      12053.42    2091.37   5.763        0.00000000837 ***
## oceanProximityNEAR OCEAN    17822.39    1694.52  10.518 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72700 on 18440 degrees of freedom
## Multiple R-squared:  0.6054, Adjusted R-squared:  0.6052
## F-statistic:  3143 on 9 and 18440 DF,  p-value: < 0.00000000000000022
```

```
# Latitude become insiginificant this time - try interaction
```

If we include an interaction term, and keep both primary variables (Longitude and Latitude).

```
fit4 <- lm(clean_data$medianHouseValue ~. + longitude:latitude - id - population,
           data = clean_data)
summary(fit4)
```

```
##
## Call:
## lm(formula = clean_data$medianHouseValue ~ . + longitude:latitude -
##     id - population, data = clean_data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -610673  -45157  -12027   30261  448740
##
## Coefficients:
##                              Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)              -11692091.12    869881.34 -13.441 < 0.0000000000000002
## longitude                  -103658.32      7169.36 -14.459 < 0.0000000000000002
## latitude                    255254.94     25694.63   9.934 < 0.0000000000000002
## housingMedianAge               798.98        45.19  17.680 < 0.0000000000000002
## aveRooms                      -8482.54       620.04 -13.681 < 0.0000000000000002
## aveBedrooms                   51119.73      2970.62  17.208 < 0.0000000000000002
## medianIncome                  41788.71       446.45  93.602 < 0.0000000000000002
## oceanProximityINLAND         -45508.15      2031.78 -22.398 < 0.0000000000000002
## oceanProximityISLAND         134582.05     35746.61   3.765             0.000167
## oceanProximityNEAR BAY         7045.09      2101.32   3.353             0.000802
## oceanProximityNEAR OCEAN      15431.57      1803.53   8.556 < 0.0000000000000002
## longitude:latitude             2291.88       210.22  10.903 < 0.0000000000000002
##
## (Intercept)                 ***
## longitude                   ***
## latitude                    ***
## housingMedianAge            ***
## aveRooms                    ***
## aveBedrooms                 ***
## medianIncome                ***
## oceanProximityINLAND        ***
## oceanProximityISLAND        ***
## oceanProximityNEAR BAY      ***
## oceanProximityNEAR OCEAN    ***
## longitude:latitude          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71420 on 18438 degrees of freedom
## Multiple R-squared:  0.6192, Adjusted R-squared:  0.619
## F-statistic:  2726 on 11 and 18438 DF,  p-value: < 0.00000000000000022
```

On top of interaction, we consider using aveRooms instead of both aveRooms and aveBedrooms:

```
fit5 <- lm(clean_data$medianHouseValue ~. + longitude:latitude - id - population - aveBedrooms,
           data = clean_data)
summary(fit5)
```

```
##
## Call:
## lm(formula = clean_data$medianHouseValue ~ . + longitude:latitude -
##     id - population - aveBedrooms, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -533459  -46197  -12449   30995  451416
##
```

```
## Coefficients:
##                               Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)              -12706766.34    874798.59 -14.525 < 0.0000000000000002
## longitude                  -112370.04      7208.47 -15.589 < 0.0000000000000002
## latitude                    285187.47     25840.03  11.037 < 0.0000000000000002
## housingMedianAge               792.80        45.55  17.405 < 0.0000000000000002
## aveRooms                      1419.86       232.76   6.100        0.00000000108
## medianIncome                 36364.65       318.70 114.103 < 0.0000000000000002
## oceanProximityINLAND        -51836.07      2014.16 -25.736 < 0.0000000000000002
## oceanProximityISLAND        144225.58     36027.13   4.003        0.00006272281
## oceanProximityNEAR BAY        7767.50      2117.65   3.668              0.000245
## oceanProximityNEAR OCEAN     15575.25      1817.88   8.568 < 0.0000000000000002
## longitude:latitude            2542.36       211.38  12.027 < 0.0000000000000002
##
## (Intercept)                 ***
## longitude                   ***
## latitude                    ***
## housingMedianAge            ***
## aveRooms                    ***
## medianIncome                ***
## oceanProximityINLAND        ***
## oceanProximityISLAND        ***
## oceanProximityNEAR BAY      ***
## oceanProximityNEAR OCEAN    ***
## longitude:latitude          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71990 on 18439 degrees of freedom
## Multiple R-squared:  0.6131, Adjusted R-squared:  0.6129
## F-statistic:  2922 on 10 and 18439 DF,  p-value: < 0.00000000000000022
```

What if interaction?

```
fit6 <- lm(clean_data$medianHouseValue ~. + longitude:latitude + aveRooms:aveBedrooms - id - population
           data = clean_data)
summary(fit6)
```

```
##
## Call:
## lm(formula = clean_data$medianHouseValue ~ . + longitude:latitude +
##     aveRooms:aveBedrooms - id - population, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -571513  -45102  -11755   30125  576281
##
## Coefficients:
##                               Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)              -11628069.70    868968.81 -13.381 < 0.0000000000000002
## longitude                  -103307.69      7161.58 -14.425 < 0.0000000000000002
## latitude                    249931.40     25679.12   9.733 < 0.0000000000000002
## housingMedianAge               818.61        45.24  18.094 < 0.0000000000000002
```

```
## aveRooms                      -8112.88       621.97 -13.044 < 0.0000000000000002
## aveBedrooms                   60488.79      3299.80  18.331 < 0.0000000000000002
## medianIncome                  41749.78       446.00  93.610 < 0.0000000000000002
## oceanProximityINLAND         -45191.08      2030.11 -22.260 < 0.0000000000000002
## oceanProximityISLAND         129480.08     35715.47   3.625             0.000289
## oceanProximityNEAR BAY         6580.38      2100.20   3.133             0.001732
## oceanProximityNEAR OCEAN      14847.99      1803.76   8.232 < 0.0000000000000002
## longitude:latitude             2255.33       210.06  10.737 < 0.0000000000000002
## aveRooms:aveBedrooms           -137.57        21.20  -6.490      0.0000000000878
##
## (Intercept)              ***
## longitude                ***
## latitude                 ***
## housingMedianAge         ***
## aveRooms                 ***
## aveBedrooms              ***
## medianIncome             ***
## oceanProximityINLAND     ***
## oceanProximityISLAND     ***
## oceanProximityNEAR BAY   **
## oceanProximityNEAR OCEAN ***
## longitude:latitude       ***
## aveRooms:aveBedrooms     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71340 on 18437 degrees of freedom
## Multiple R-squared:  0.6201, Adjusted R-squared:  0.6198
## F-statistic:  2508 on 12 and 18437 DF,  p-value: < 0.00000000000000022
```

Heteroskedsiticity & Skewness (i.e. Non-normality)

All terms are

```
fit7 <- lm(log(medianHouseValue) ~ .
           + longitude:latitude + aveRooms:aveBedrooms - id - population,
           data = clean_data)
summary(fit7)
```

```
##
## Call:
## lm(formula = log(medianHouseValue) ~ . + longitude:latitude +
##     aveRooms:aveBedrooms - id - population, data = clean_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.40995 -0.21143 -0.01419  0.19978  1.75795
##
## Coefficients:
##                             Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)              -48.7829021   4.1638251 -11.716 < 0.0000000000000002
## longitude                 -0.5431950   0.0343160 -15.829 < 0.0000000000000002
## latitude                   1.2015608   0.1230463   9.765 < 0.0000000000000002
```

```
## housingMedianAge           0.0011021   0.0002168   5.084         0.0000003737
## aveRooms                  -0.0141665   0.0029803  -4.753         0.0000020149
## aveBedrooms                0.1725296   0.0158116  10.912 < 0.0000000000000002
## medianIncome               0.1711048   0.0021371  80.065 < 0.0000000000000002
## oceanProximityINLAND      -0.3461346   0.0097276 -35.583 < 0.0000000000000002
## oceanProximityISLAND       0.5362364   0.1711373   3.133              0.00173
## oceanProximityNEAR BAY     0.0136805   0.0100635   1.359              0.17403
## oceanProximityNEAR OCEAN   0.0152895   0.0086431   1.769              0.07691
## longitude:latitude         0.0111602   0.0010065  11.088 < 0.0000000000000002
## aveRooms:aveBedrooms      -0.0005657   0.0001016  -5.570         0.0000000258
##
## (Intercept)              ***
## longitude                ***
## latitude                 ***
## housingMedianAge         ***
## aveRooms                 ***
## aveBedrooms              ***
## medianIncome             ***
## oceanProximityINLAND     ***
## oceanProximityISLAND     **
## oceanProximityNEAR BAY
## oceanProximityNEAR OCEAN .
## longitude:latitude       ***
## aveRooms:aveBedrooms     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3418 on 18437 degrees of freedom
## Multiple R-squared:  0.6409, Adjusted R-squared:  0.6407
## F-statistic:  2743 on 12 and 18437 DF,  p-value: < 0.00000000000000022
```

2.4 Model improvements

There are a couple of issues we would like to address: 1. Collinearity: We will be doing a Ridge regression to address this problem. Although Lasso is also a relevant method, it only performs better if many predictors are useless. Given that many of the predictors in our initial model are useful (individially significant through the t-test), we want to keep all of them. 2.

2.5 Three most significant variables

```r
# this chunk need to be corrected
p_values_fit7 <- summary(fit7)$coefficients[, 4]
p_values_fit7 <- p_values_fit7[-1] # exlcude the intercept term
sorted_p_values_fit7 <- sort(p_values_fit7)
top_3_predictors_fit7 <- names(sorted_p_values_fit7)[1:3]
top_3_predictors_fit7
```

```
## [1] "medianIncome"         "oceanProximityINLAND" "longitude"
```

---

3 Model Performance: initial model vs improved model