sta141a_project

2023-12-10

```r
# All the packages and Libraries
library(GGally)
```

```
## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(ggplot2)
library(ISLR)
library("tibble")
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
```

```r
library("tidyr")
```

Data Processing

```r
data <- read.table("../sta141a_Project/California_Houses.csv", sep = ",", header = T) # uploading all t
data <- na.omit(data) # removing all the NA values

fullData = data %>% filter(data$Median_House_Value < 500000) # removing the outliers for the house valu
head(data)
```

```
##   Median_House_Value Median_Income Median_Age Tot_Rooms Tot_Bedrooms Population
## 1             452600        8.3252         41       880          129        322
## 2             358500        8.3014         21      7099         1106       2401
## 3             352100        7.2574         52      1467          190        496
## 4             341300        5.6431         52      1274          235        558
## 5             342200        3.8462         52      1627          280        565
```

```
## 6             269700        4.0368         52       919           213        413
##   Households Latitude Longitude Distance_to_coast Distance_to_LA
## 1        126    37.88   -122.23          9263.041       556529.2
## 2       1138    37.86   -122.22         10225.733       554279.9
## 3        177    37.85   -122.24          8259.085       554610.7
## 4        219    37.85   -122.25          7768.087       555194.3
## 5        259    37.85   -122.25          7768.087       555194.3
## 6        193    37.85   -122.25          7768.087       555194.3
##   Distance_to_SanDiego Distance_to_SanJose Distance_to_SanFrancisco
## 1             735501.8            67432.52                 21250.21
## 2             733236.9            65049.91                 20880.60
## 3             733525.7            64867.29                 18811.49
## 4             734095.3            65287.14                 18031.05
## 5             734095.3            65287.14                 18031.05
## 6             734095.3            65287.14                 18031.05
```

```r
summary(data)
```

```
##  Median_House_Value Median_Income      Median_Age      Tot_Rooms
##  Min.   : 14999     Min.   : 0.4999   Min.   : 1.00   Min.   :    2
##  1st Qu.:119600     1st Qu.: 2.5634   1st Qu.:18.00   1st Qu.: 1448
##  Median :179700     Median : 3.5348   Median :29.00   Median : 2127
##  Mean   :206856     Mean   : 3.8707   Mean   :28.64   Mean   : 2636
##  3rd Qu.:264725     3rd Qu.: 4.7432   3rd Qu.:37.00   3rd Qu.: 3148
##  Max.   :500001     Max.   :15.0001   Max.   :52.00   Max.   :39320
##   Tot_Bedrooms      Population      Households        Latitude
##  Min.   :   1.0   Min.   :    3   Min.   :   1.0   Min.   :32.54
##  1st Qu.: 295.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.:33.93
##  Median : 435.0   Median : 1166   Median : 409.0   Median :34.26
##  Mean   : 537.9   Mean   : 1425   Mean   : 499.5   Mean   :35.63
##  3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.:37.71
##  Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :41.95
##    Longitude       Distance_to_coast  Distance_to_LA      Distance_to_SanDiego
##  Min.   :-124.3   Min.   :   120.7   Min.   :    420.6   Min.   :    484.9
##  1st Qu.:-121.8   1st Qu.:  9079.8   1st Qu.:  32111.3   1st Qu.: 159426.4
##  Median :-118.5   Median : 20522.0   Median : 173667.5   Median : 214739.8
##  Mean   :-119.6   Mean   : 40509.3   Mean   : 269422.0   Mean   : 398164.9
##  3rd Qu.:-118.0   3rd Qu.: 49830.4   3rd Qu.: 527156.2   3rd Qu.: 705795.4
##  Max.   :-114.3   Max.   :333804.7   Max.   :1018260.1   Max.   :1196919.3
##  Distance_to_SanJose Distance_to_SanFrancisco
##  Min.   :   569.4    Min.   :   456.1
##  1st Qu.:113119.9    1st Qu.:117395.5
##  Median :459758.9    Median :526546.7
##  Mean   :349187.6    Mean   :386688.4
##  3rd Qu.:516946.5    3rd Qu.:584552.0
##  Max.   :836762.7    Max.   :903627.7
```

```r
incomeAndHousingValData = subset(fullData, select = c(Median_House_Value, Median_Income))

# Converting the values
incomeAndHousingValData$Median_Income <- 10000*incomeAndHousingValData$Median_Income # need to convert
incomeAndHousingValData$Median_House_Value <- 1*incomeAndHousingValData$Median_House_Value
head(incomeAndHousingValData$Median_Income)
```

```
## [1] 83252 83014 72574 56431 38462 40368
```

```r
# distanceData for linear regression of the house prices and the distance from different Metropolitan a
distanceData <- subset(fullData, select = c(Median_House_Value, Distance_to_coast,
                                           Distance_to_LA, Distance_to_SanDiego, Distance_to_SanJose,Distanc
#ggpairs(subData)
#plot(subData)
```

Summaries

```r
summary(incomeAndHousingValData) # housing and income
```

```
##  Median_House_Value Median_Income
##  Min.   : 14999     Min.   :  4999
##  1st Qu.:116475     1st Qu.: 25263
##  Median :173600     Median : 34490
##  Mean   :192055     Mean   : 36764
##  3rd Qu.:247900     3rd Qu.: 45825
##  Max.   :499100     Max.   :150001
```

```r
summary(distanceData) # housing and distances to coast, LA, San Diego, San Jose, SF
```

```
##  Median_House_Value Distance_to_coast  Distance_to_LA      Distance_to_SanDiego
##  Min.   : 14999     Min.   :   120.7   Min.   :    420.6   Min.   :    484.9
##  1st Qu.:116475     1st Qu.:  9851.1   1st Qu.:  33053.6   1st Qu.: 158890.6
##  Median :173600     Median : 21327.1   Median : 177200.7   Median : 223113.6
##  Mean   :192055     Mean   : 41967.7   Mean   : 271815.9   Mean   : 399800.7
##  3rd Qu.:247900     3rd Qu.: 53250.1   3rd Qu.: 528960.0   3rd Qu.: 707671.8
##  Max.   :499100     Max.   :333804.7   Max.   :1018260.1   Max.   :1196919.3
##  Distance_to_SanJose Distance_to_SanFrancisco
##  Min.   :   569.4    Min.   :   456.1
##  1st Qu.:117875.9    1st Qu.:120402.5
##  Median :458025.4    Median :524571.7
##  Mean   :349802.7    Mean   :387223.1
##  3rd Qu.:517889.8    3rd Qu.:585659.7
##  Max.   :836762.7    Max.   :903627.7
```

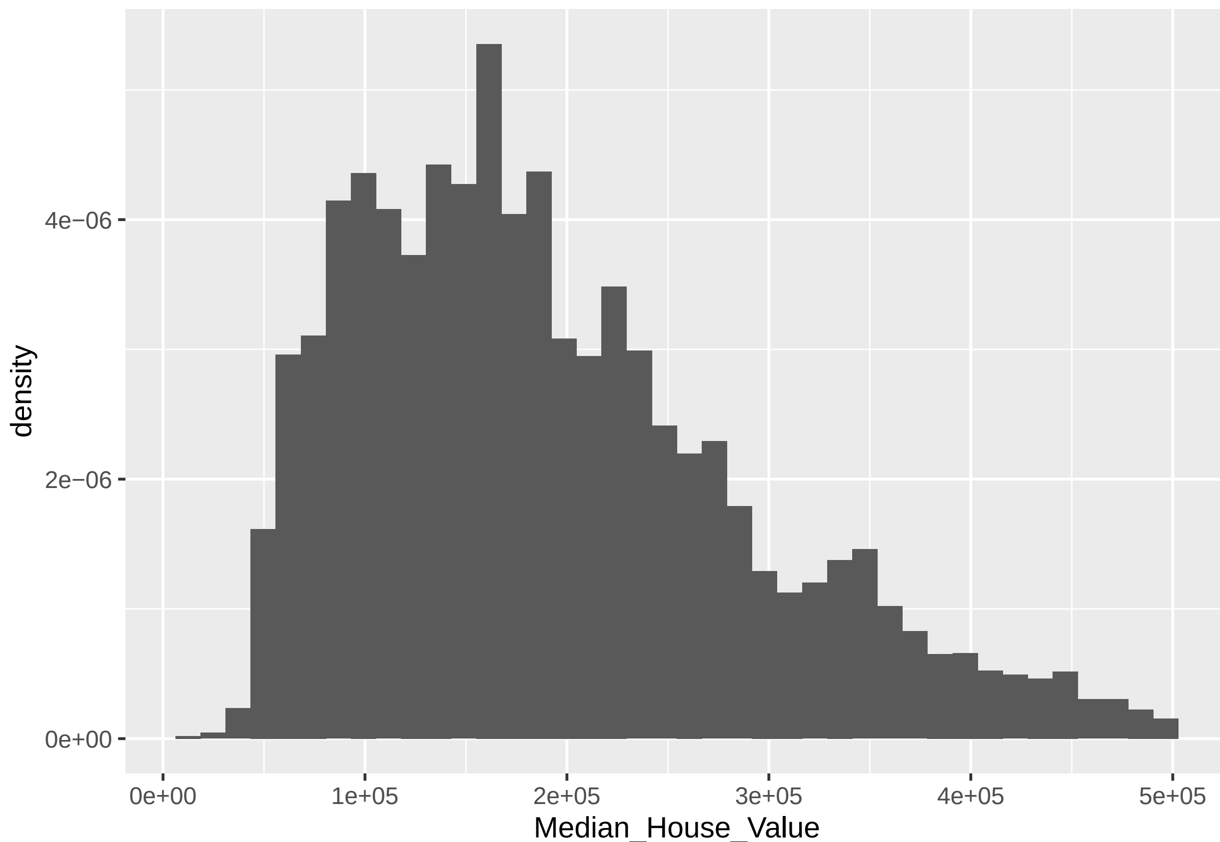
Visualization and Methodology

Linear Regression: Starting off with Median housing prices vs Median income

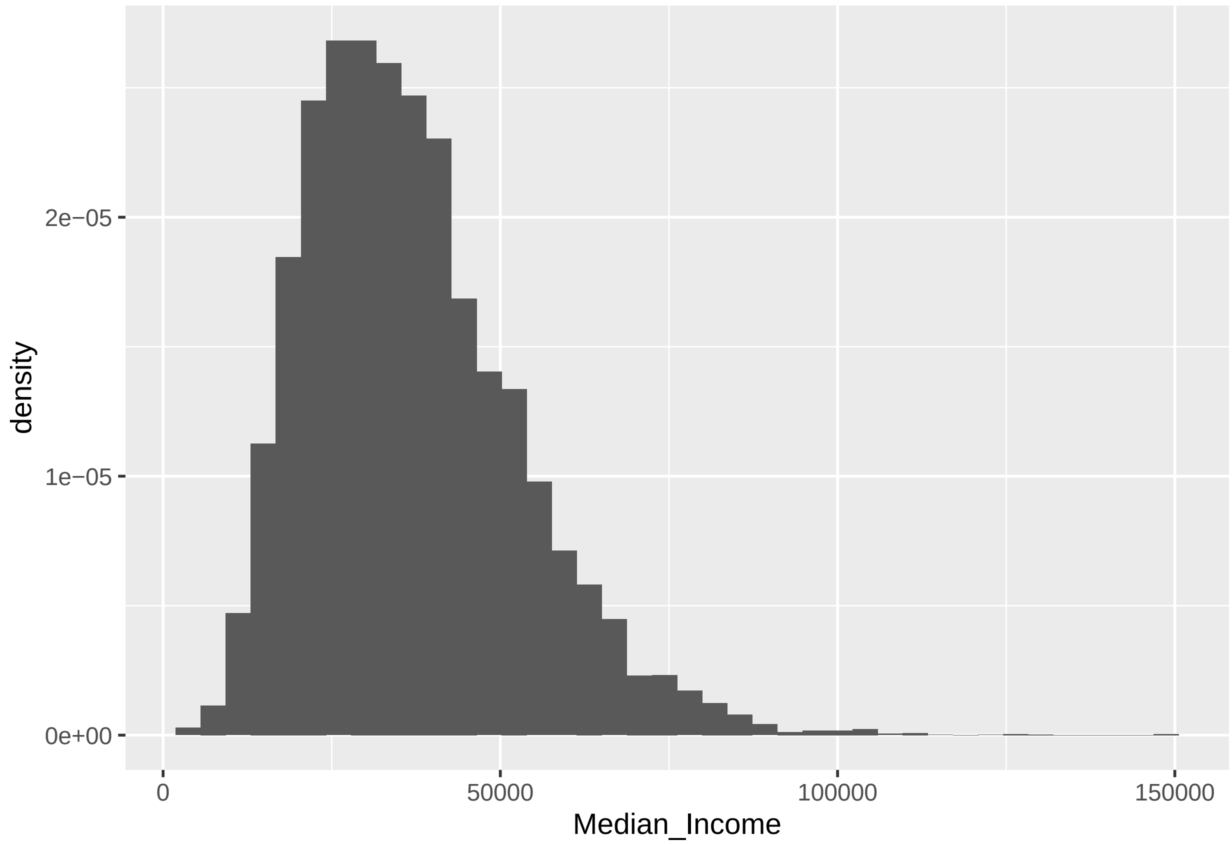
```r
head(incomeAndHousingValData)
```

```
##   Median_House_Value Median_Income
## 1             452600         83252
## 2             358500         83014
## 3             352100         72574
## 4             341300         56431
## 5             342200         38462
## 6             269700         40368
```

```r
#histograms for median house value and median income
ggplot(incomeAndHousingValData, aes(x=Median_House_Value))+
  geom_histogram(aes(y = after_stat(density)), bins = 40)
```




```r
ggplot(incomeAndHousingValData, aes(x=Median_Income))+
  geom_histogram(aes(y = after_stat(density)), bins = 40)
```



Correlation

```
cor(incomeAndHousingValData$Median_Income,incomeAndHousingValData$Median_House_Value)
```

```
## [1] 0.6467194
```

Linear regression between median Housing value and median income

```
 # median income is def gamma distribution whereas house value is either gamma or normal or both
# Gamma would work because all values are great than zero
# Gamma is the distribution we used because our data is right skewed with continuous positive values
# poisson will not be a good idea for this type of data because it's not really discrete
# for the link after experimenting for a but it seems that identity would not be a good idea because th
# line would have been y = 49926.1 + 38514.6(x) which doesn't really make sense for our data

#verySimpleModel = glm(fullData$Median_House_Value ~ fullData$Median_Income, data = fullData,family = G
verySimpleModel = glm(incomeAndHousingValData$Median_House_Value ~ incomeAndHousingValData$Median_Income
summary(verySimpleModel)
```

```
##
## Call:
## glm(formula = incomeAndHousingValData$Median_House_Value ~ incomeAndHousingValData$Median_Income,
##     family = Gamma(link = log), data = incomeAndHousingValData)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -1.83225  -0.32616  -0.08016   0.18103   1.92546
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1.138e+01  7.785e-03  1461.5   <2e-16 ***
## incomeAndHousingValData$Median_Income 2.008e-05  1.947e-07   103.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1837816)
##
##     Null deviance: 5227.8  on 19647  degrees of freedom
## Residual deviance: 3284.9  on 19646  degrees of freedom
## AIC: 493998
##
## Number of Fisher Scoring iterations: 4
```

```r
head(incomeAndHousingValData)
```

```
##   Median_House_Value Median_Income
## 1             452600         83252
## 2             358500         83014
## 3             352100         72574
## 4             341300         56431
## 5             342200         38462
## 6             269700         40368
```

The effect of Median Income on Median Housing Value is that if Median Income increases by one then log of Median Housing Value increases by 0.00002008. Keep in mind that our median income is converted to the 10k.

Test: H0 = B1 = 0 and Ha = B1 != 0

The effect of Median Income is significant at alpha 0.01 therefore we can reject the null hypothesis that H0 = B1 = 0.

```r
verySimpleModelBad = glm(incomeAndHousingValData$Median_House_Value ~ 1, data = incomeAndHousingValData
#summary(verySimpleModelBad)
AIC(verySimpleModel)
```

```
## [1] 493997.7
```

```r
BIC(verySimpleModel)
```

```
## [1] 494021.3
```

```r
AIC(verySimpleModelBad)
```

```
## [1] 503447.3
```

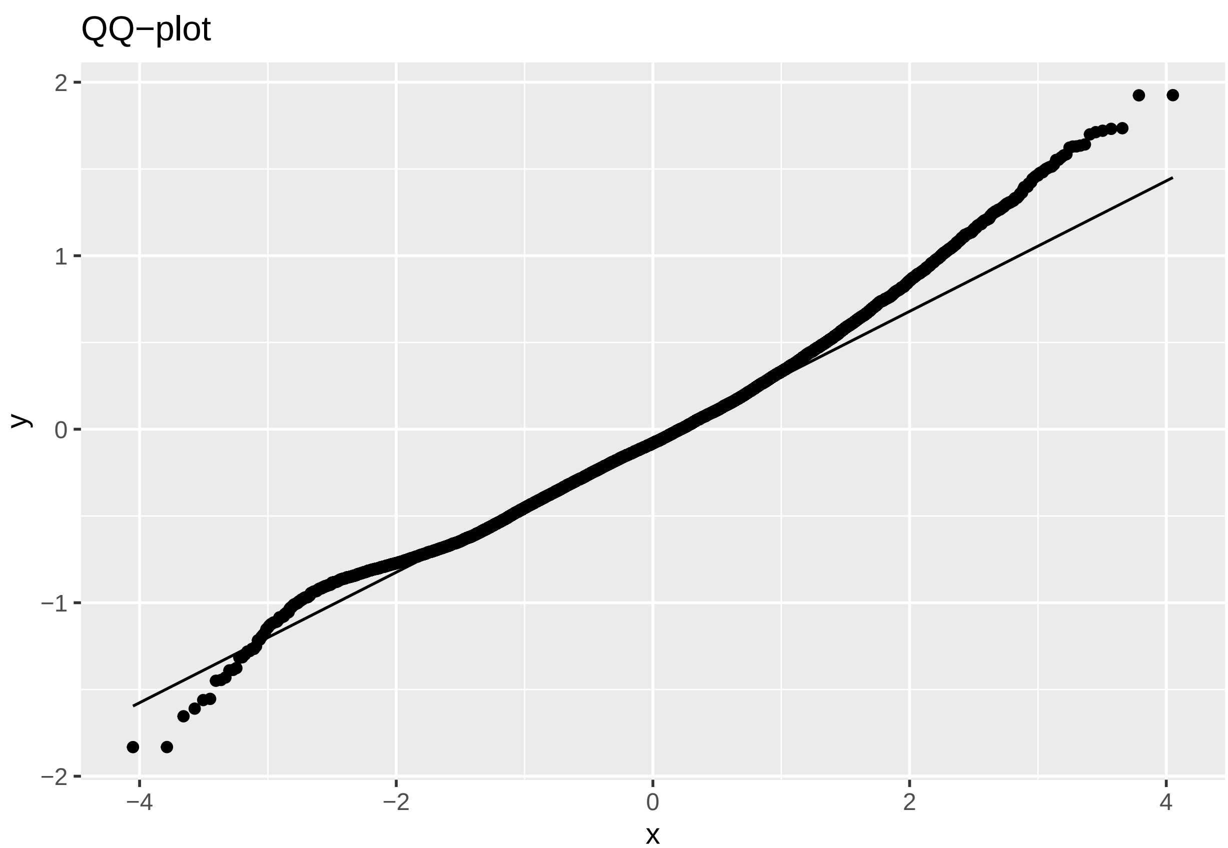
```r
BIC(verySimpleModelBad)
```

```
## [1] 503463.1
```

We can see that the model with Median Income is a better model compared to the model with just intercept, because both the AIC and BIC is lower for the model with Median Income. ??

```r
residualForVerySimpleModel <- resid(verySimpleModel)
fittedForVerySimpleModel <- fitted(verySimpleModel)

# QQ-plot for the residual of the very simple model
 ggplot(fullData, aes(sample = residualForVerySimpleModel)) +
  stat_qq() + stat_qq_line() + labs(title = "QQ-plot")
```




```r
# COMEBACK: maybe do a resudual vs fitted plot
```

Multiple Linear Regression We will be looking at: Does being part of a metropolitan area play a part in a higher average cost of living? Might need to change this later

```r
fullModel = glm(Median_House_Value~(.), distanceData,family =  Gamma(link = log))
fullModel
```

```
##
## Call:  glm(formula = Median_House_Value ~ (.), family = Gamma(link = log),
##     data = distanceData)
##
## Coefficients:
##             (Intercept)         Distance_to_coast            Distance_to_LA
##               1.292e+01                -5.280e-06                -6.690e-07
##     Distance_to_SanDiego       Distance_to_SanJose  Distance_to_SanFrancisco
##              -1.777e-07                 2.497e-07                -1.076e-06
##
## Degrees of Freedom: 19647 Total (i.e. Null);  19642 Residual
## Null Deviance:       5228
## Residual Deviance: 3463  AIC: 495100
```

```r
badModel = glm(distanceData$Median_House_Value~1, distanceData ,family =  Gamma(link = log))
summary(fullModel)
```

```
##
## Call:
## glm(formula = Median_House_Value ~ (.), family = Gamma(link = log),
##     data = distanceData)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.79271  -0.33919  -0.08473   0.21289   2.64683
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.292e+01  3.161e-02 408.811  < 2e-16 ***
## Distance_to_coast        -5.280e-06  7.394e-08 -71.416  < 2e-16 ***
## Distance_to_LA           -6.690e-07  4.379e-08 -15.275  < 2e-16 ***
## Distance_to_SanDiego     -1.777e-07  5.671e-08  -3.134  0.00172 **
## Distance_to_SanJose       2.497e-07  1.396e-07   1.789  0.07360 .
## Distance_to_SanFrancisco -1.076e-06  1.522e-07  -7.068 1.62e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1837657)
##
##     Null deviance: 5227.8  on 19647  degrees of freedom
## Residual deviance: 3462.8  on 19642  degrees of freedom
## AIC: 495071
##
## Number of Fisher Scoring iterations: 6
```

```r
summary(badModel)
```

```
##
## Call:
```

```
## glm(formula = distanceData$Median_House_Value ~ 1, family = Gamma(link = log),
##     data = distanceData)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.80438  -0.46168  -0.09936   0.26658   1.13464
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.165539   0.003607    3372   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.2556718)
##
##     Null deviance: 5227.8  on 19647  degrees of freedom
## Residual deviance: 5227.8  on 19647  degrees of freedom
## AIC: 503447
##
## Number of Fisher Scoring iterations: 4
```

```r
# comeback for residuals

# ANVOA AND ALSO F-TEST
```

```r
anova( fullModel, test = 'LRT')
```

```
## Analysis of Deviance Table
##
## Model: Gamma, link: log
##
## Response: Median_House_Value
##
## Terms added sequentially (first to last)
##
##
##                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                    19647     5227.8
## Distance_to_coast        1  1595.52     19646     3632.2 < 2.2e-16 ***
## Distance_to_LA           1     7.01     19645     3625.2 6.485e-10 ***
## Distance_to_SanDiego     1    84.82     19644     3540.4 < 2.2e-16 ***
## Distance_to_SanJose      1    68.13     19643     3472.3 < 2.2e-16 ***
## Distance_to_SanFrancisco 1     9.49     19642     3462.8 6.694e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-test Ho = B1 = B2 = B3 = B4 = B5 = 0, Ha: B1 != 0 OR B2 != 0 OR B3 != 0 OR B4 != 0 OR B5 != 0 At 0.001 we can reject the null.

```r
smallerModel = glm(Median_House_Value~Distance_to_coast + Distance_to_LA + Distance_to_SanFrancisco, di
smallerModel
```

```
##
```

```
## Call:  glm(formula = Median_House_Value ~ Distance_to_coast + Distance_to_LA +
##     Distance_to_SanFrancisco, family = Gamma(link = log), data = distanceData)
##
## Coefficients:
##              (Intercept)         Distance_to_coast            Distance_to_LA
##                1.283e+01                -5.408e-06                -7.387e-07
## Distance_to_SanFrancisco
##               -7.299e-07
##
## Degrees of Freedom: 19647 Total (i.e. Null);  19644 Residual
## Null Deviance:       5228
## Residual Deviance: 3465  AIC: 495100
```

```r
summary(smallerModel)
```

```
##
## Call:
## glm(formula = Median_House_Value ~ Distance_to_coast + Distance_to_LA +
##     Distance_to_SanFrancisco, family = Gamma(link = log), data = distanceData)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.79619  -0.33899  -0.08297   0.21118   2.62994
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.283e+01  1.488e-02  862.43   <2e-16 ***
## Distance_to_coast        -5.408e-06  6.372e-08  -84.88   <2e-16 ***
## Distance_to_LA           -7.387e-07  2.384e-08  -30.98   <2e-16 ***
## Distance_to_SanFrancisco -7.299e-07  2.329e-08  -31.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1839911)
##
##     Null deviance: 5227.8  on 19647  degrees of freedom
## Residual deviance: 3464.7  on 19644  degrees of freedom
## AIC: 495078
##
## Number of Fisher Scoring iterations: 6
```

F-test Ho = B1 = B2 = B3 = B4 = B5 = 0, Ha: B1 != 0 OR B2 != 0 OR B3 != 0 OR B4 != 0 OR B5 != 0

ANOVA

```r
anova(smallerModel, fullModel, test = 'LRT')
```

```
## Analysis of Deviance Table
##
## Model 1: Median_House_Value ~ Distance_to_coast + Distance_to_LA + Distance_to_SanFrancisco
## Model 2: Median_House_Value ~ (Distance_to_coast + Distance_to_LA + Distance_to_SanDiego +
##     Distance_to_SanJose + Distance_to_SanFrancisco)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
```

```
## 1     19644     3464.7
## 2     19642     3462.8  2   1.8846 0.005929 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0 = B1 = B2 = B5 = 0, HA B1 != 0 OR B2 != 0 OR B5 != 0 We cannot reject the null hypothesis at alpha 0.001 therefore we choose the small model and get rid of the full model.

```r
AIC(badModel)
```

```
## [1] 503447.3
```

```r
BIC(badModel)
```

```
## [1] 503463.1
```

```r
AIC(fullModel)
```

```
## [1] 495071.4
```

```r
BIC(fullModel)
```

```
## [1] 495126.6
```

```r
AIC(smallerModel)
```

```
## [1] 495078.4
```

```r
BIC(smallerModel)
```

```
## [1] 495117.9
```

We can see that the badmodel is the worst model by looking at the aic and bic since it has the largest value. I would say that smaller model would be the best as it has the lowest BIC despite the AIC being greater than the AIC for full model. Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.