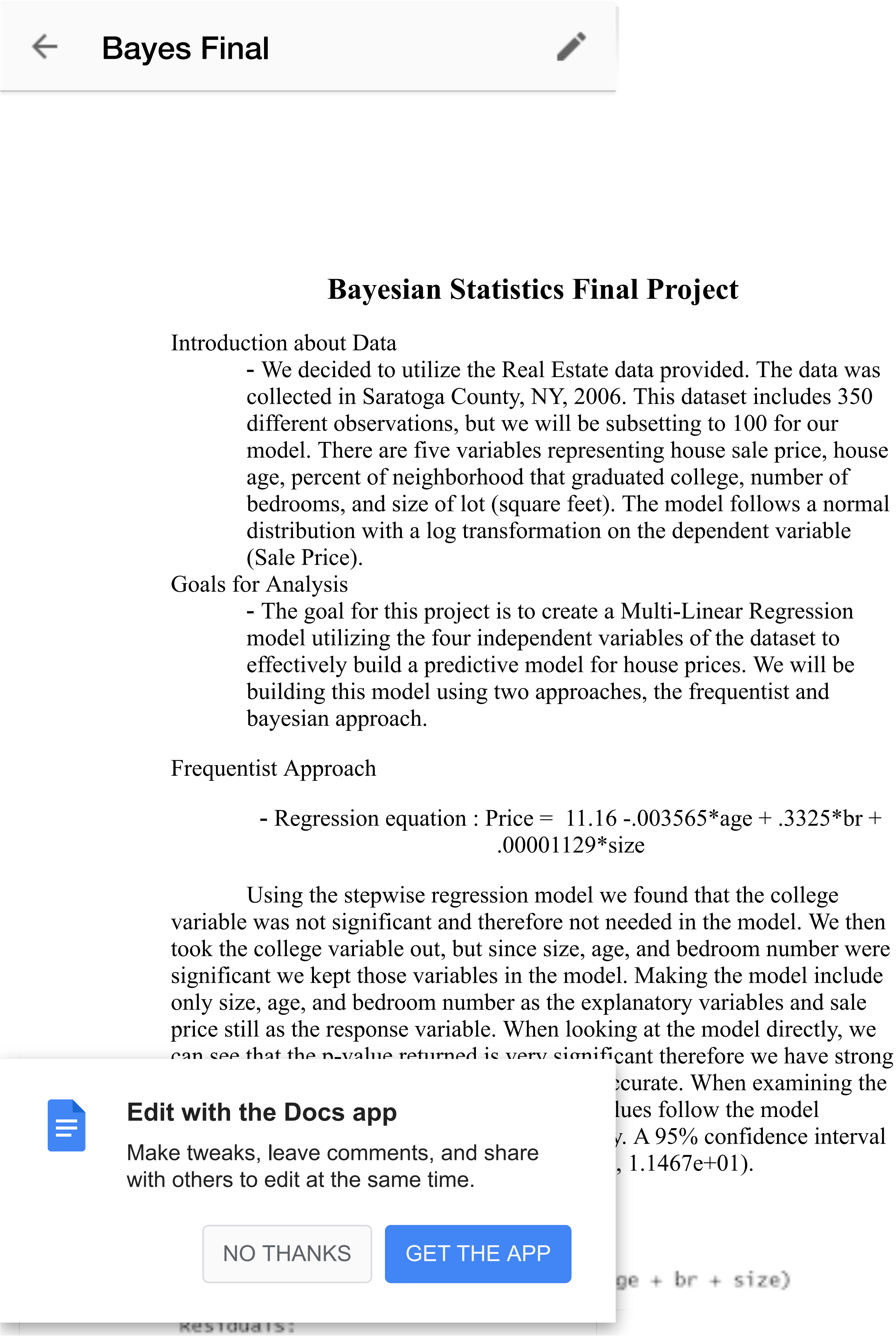


Bayesian Statistics Final Project

Introduction about Data

- We decided to utilize the Real Estate data provided. The data was collected in Saratoga County, NY, 2006. This dataset includes 350 different observations, but we will be subsetting to 100 for our model. There are five variables representing house sale price, house age, percent of neighborhood that graduated college, number of bedrooms, and size of lot (square feet). The model follows a normal distribution with a log transformation on the dependent variable (Sale Price).

Goals for Analysis

- The goal for this project is to create a Multi-Linear Regression model utilizing the four independent variables of the dataset to effectively build a predictive model for house prices. We will be building this model using two approaches, the frequentist and bayesian approach.

Frequentist Approach

- Regression equation : Price = 11.16 -.003565*age + .3325*br + .00001129*size

Using the stepwise regression model we found that the college variable was not significant and therefore not needed in the model. We then took the college variable out, but since size, age, and bedroom number were significant we kept those variables in the model. Making the model include only size, age, and bedroom number as the explanatory variables and sale price still as the response variable. When looking at the model directly, we can see that the p-value returned is very significant therefore we have strong

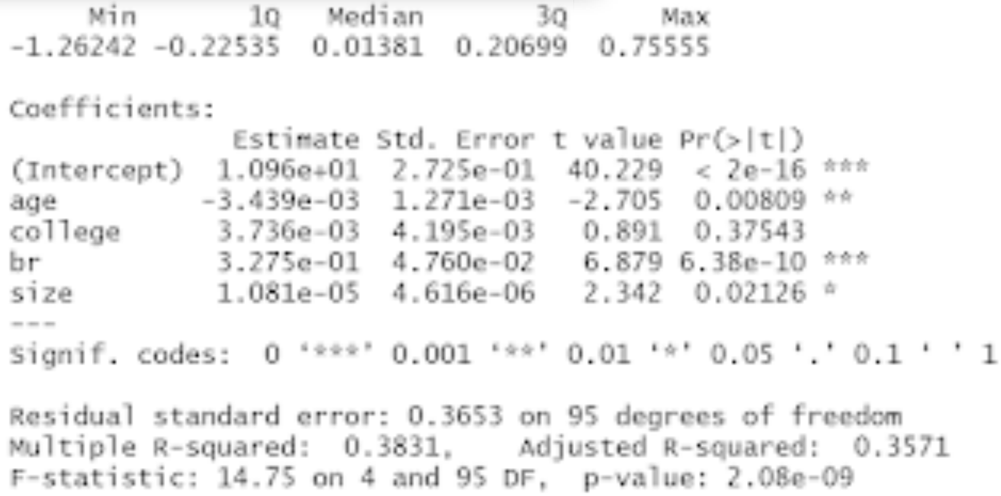
ccurate. When examining the

lues follow the model

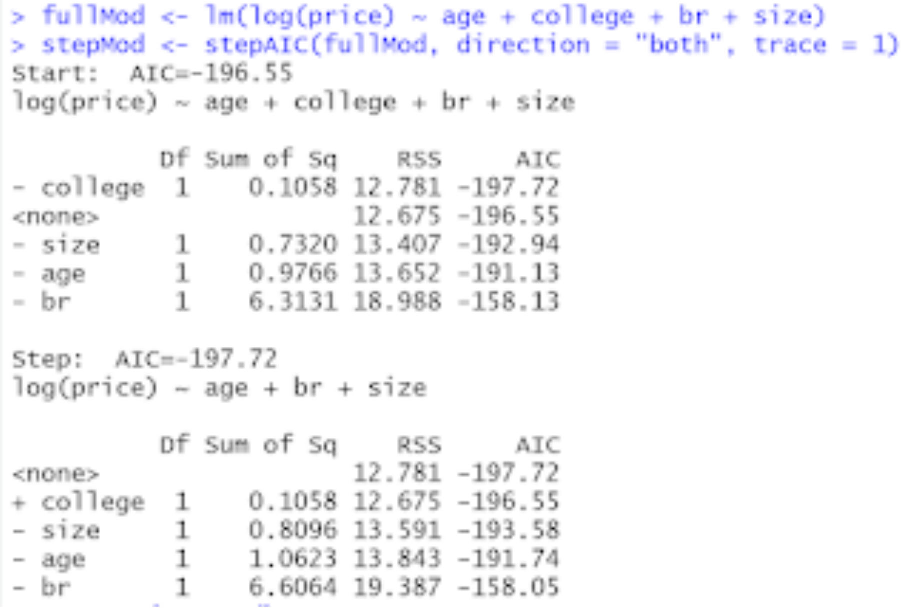
y. A 95% confidence interval

, 1.1467e+01).

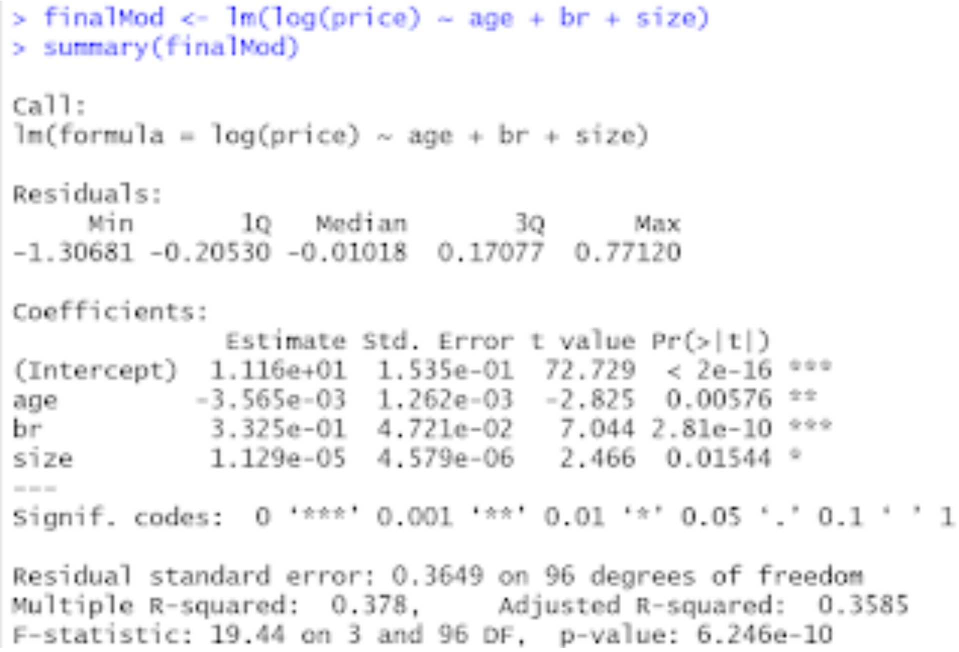
```
ge + br + size)

Residuals:
```



```
     Min       1Q   Median       3Q      Max 
-1.26242 -0.22535  0.01381  0.20699  0.75555 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.096e+01  2.725e-01  40.229  < 2e-16 ***
age         -3.439e-03  1.271e-03  -2.705  0.00809 ** 
college      3.736e-03  4.195e-03   0.891  0.37543    
br           3.275e-01  4.760e-02   6.879 6.38e-10 ***
size         1.081e-05  4.616e-06   2.342  0.02126 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3653 on 95 degrees of freedom
Multiple R-squared:  0.3831,	Adjusted R-squared:  0.3571 
F-statistic: 14.75 on 4 and 95 DF,  p-value: 2.08e-09
```

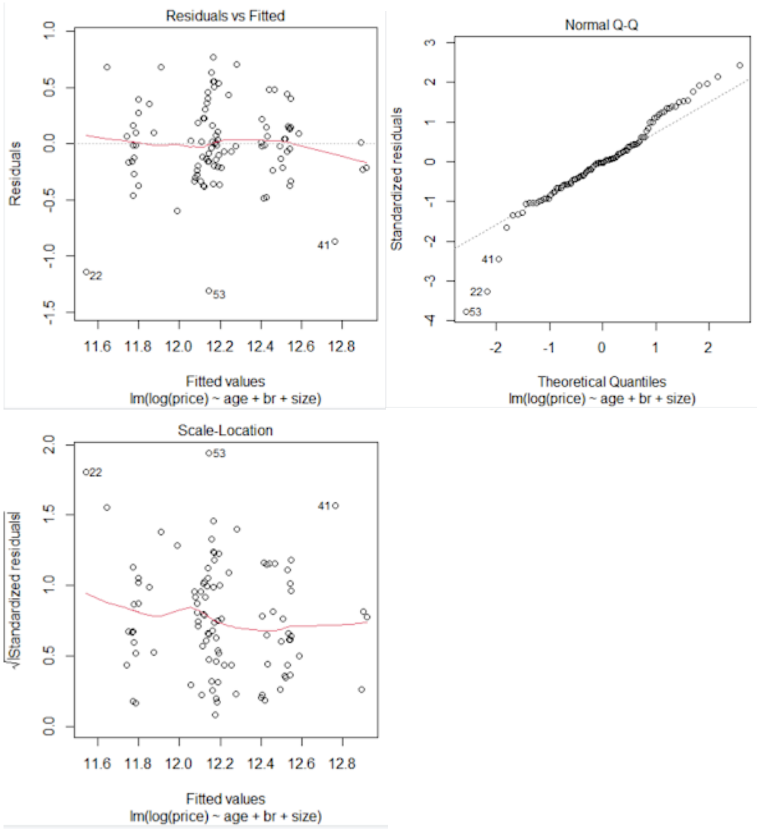


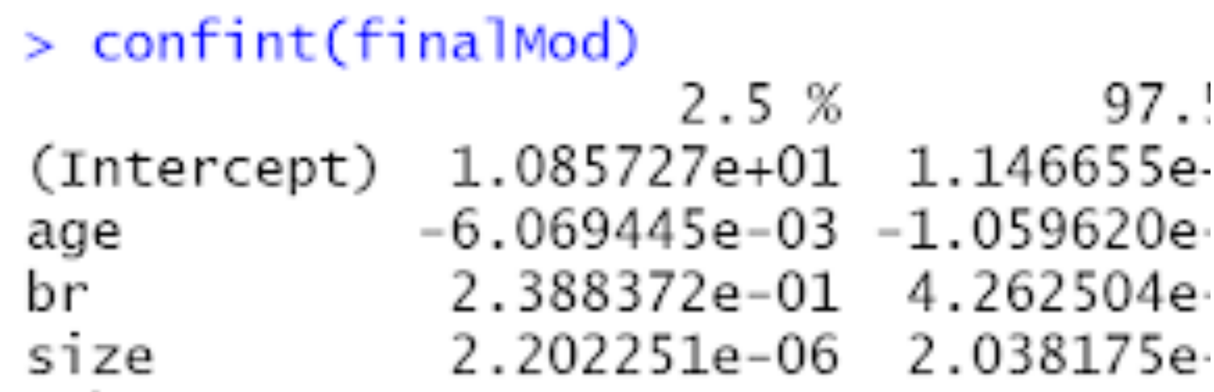
```
> fullMod <- lm(log(price) ~ age + college + br + size)
> stepMod <- stepAIC(fullMod, direction = "both", trace = 1)
Start:  AIC=-196.55
log(price) ~ age + college + br + size

          Df Sum of Sq    RSS     AIC
- college  1    0.1058 12.781 -197.72
<none>                 12.675 -196.55
- size     1    0.7320 13.407 -192.94
- age      1    0.9766 13.652 -191.13
- br       1    6.3131 18.988 -158.13

Step:  AIC=-197.72
log(price) ~ age + br + size

          Df Sum of Sq    RSS     AIC
<none>                 12.781 -197.72
+ college  1    0.1058 12.675 -196.55
- size     1    0.8096 13.591 -193.58
- age      1    1.0623 13.843 -191.74
- br       1    6.6064 19.387 -158.05
```



```
> finalMod <- lm(log(price) ~ age + br + size)
> summary(finalMod)

Call:
lm(formula = log(price) ~ age + br + size)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.30681 -0.20530 -0.01018  0.17077  0.77120 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.116e+01  1.535e-01  72.729  < 2e-16 ***
age         -3.565e-03  1.262e-03  -2.825  0.00576 ** 
br           3.325e-01  4.721e-02   7.044 2.81e-10 ***
size         1.129e-05  4.579e-06   2.466  0.01544 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3649 on 96 degrees of freedom
Multiple R-squared:  0.378,	Adjusted R-squared:  0.3585 
F-statistic: 19.44 on 3 and 96 DF,  p-value: 6.246e-10
```

Normality Checks on Model:



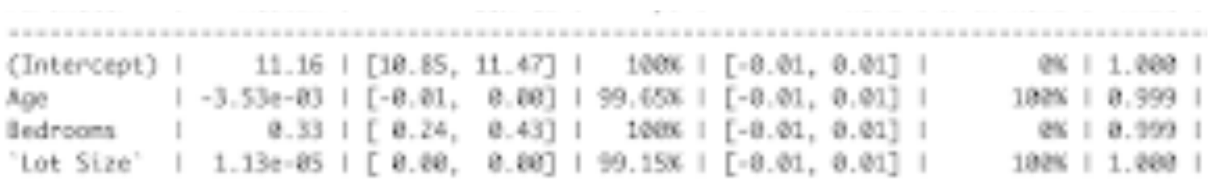


```
> confint(finalMod)
                      2.5 %          97.!
(Intercept)  1.085727e+01  1.146655e-
age         -6.069445e-03 -1.059620e-
br           2.388372e-01  4.262504e-
size         2.202251e-06  2.038175e-
```

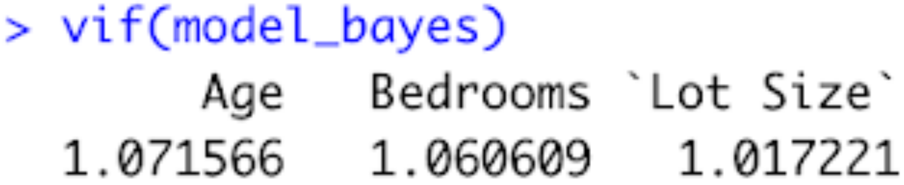
Bayesian Approach

When analyzing the output from the Bayesian methodology towards our model, we can infer that the housing price will change at a slope of 11.16 in correlation with the variables. Our output from the regression also provides similar results to the frequentists approach. When examining the variance inflation factors of the model, we can see that the values are all positively correlated to calculating the regression slope. That being said, we can determine that Age and Bedrooms are leading predictors for calculating the housing price.

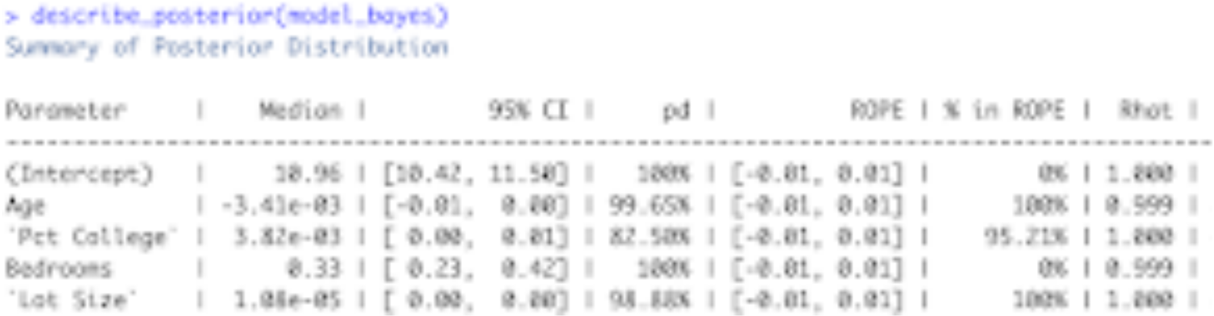
```
Parameter   |    Median |          95% CI |      pd |          ROPE | % in ROPE |  Rhat |
```



```
-------------------------------------------------------------------------------------------
(Intercept) |      11.16 | [10.85, 11.47] |   100% | [-0.01, 0.01] |          0% | 1.000 |
Age         | -3.53e-03 | [-0.01,  0.00] | 99.65% | [-0.01, 0.01] |        100% | 0.999 |
Bedrooms    |       0.33 | [ 0.24,  0.43] |   100% | [-0.01, 0.01] |          0% | 0.999 |
`Lot Size`  |  1.13e-05 | [ 0.00,  0.00] | 99.15% | [-0.01, 0.01] |        100% | 1.000 |
```



```
> vif(model_bayes)
       Age   Bedrooms `Lot Size`
  1.071566   1.060609   1.017221
```



```
> describe_posterior(model_bayes)
Summary of Posterior Distribution

Parameter    |    Median |         95% CI |     pd |          ROPE | % in ROPE |  Rhat |
-------------------------------------------------------------------------------------------
(Intercept)  |     10.96 | [10.42, 11.50] |   100% | [-0.01, 0.01] |        0% | 1.000 |
Age          | -3.41e-03 | [-0.01,  0.00] | 99.65% | [-0.01, 0.01] |      100% | 0.999 |
`Pct College`|  3.82e-03 | [ 0.00,  0.01] | 82.50% | [-0.01, 0.01] |    95.21% | 1.000 |
Bedrooms     |      0.33 | [ 0.23,  0.42] |   100% | [-0.01, 0.01] |        0% | 0.999 |
`Lot Size`   |  1.08e-05 | [ 0.00,  0.00] | 98.88% | [-0.01, 0.01] |      100% | 1.000 |
```

Comparing Initial Results

When analyzing the output from the Bayesian methodology towards our model, we can see that the slope of the regression model fits very well with what we had seen in our previous approach using the frequentist methodology. The 95% confidence interval returns the values of (1.085e+01, 1.147e+01), which also follows our previous confidence interval from the frequentists approach.

Conclusions

To conclude our study, we have been able to use dimensional reduction techniques to minimize our dataset into only significant predictor variables in our goal to create a predictive model for housing prices. We utilized both frequentist and bayesian approaches to our model, both generating ideal and comparable results. In conclusion, we found that the predictor variables of age, bedrooms, and lot size are the main determinants of housing price. Our model's fitted values followed the data very well, and our model p-values are significant therefore we have evidence to conclude that our model accurately predicts housing prices.