THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Bayesian Analysis on UK Used Car Prices 2008-2020

By Abdulaziz Gebril, Jake Lieberfarb, and Divya Parmar
Professor: Dr. James Huang

DATS 6311_10 Bayesian Computing

December 8th, 2021

Introduction:

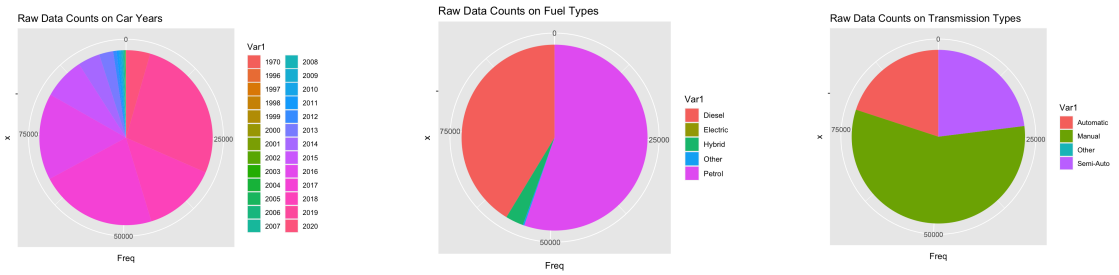
For this project, the dataset of 100,000 used car sales in the United Kingdom analyzed. Originally, there were 9 total variables with 108,540 rows. All further analysis could be found on the researchers' Github. The variable of 'model' was dropped from the analysis as there were too many classes to build a model around it. The multiple csv files from Kaggle were compiled and identified with a new variable 'Car.Make'. The dependent variable was 'price'. The four categorical independent variables were 'year', 'transmission', 'fuelType', and 'Car.Make'. the four quantitative independent variables were 'mileage', 'tax', 'mpg' and 'engineSize'. One-hot encoding was applied to the categorical data and the final dataframe had a total of 32 columns. Outliers were kept in the analysis as the researcher wanted to avoid class imbalance for the Car.Make categories. The goal of this project was to construct Bayesian based Hierarchical, Logistic, and Linear models and compare them to frequentist versions to see if more accurate evaluations could be made about the price of used cars in the United Kingdom.

Variable	**Description**	**Variable**	**Description**
year_2008	'0' no or '1' yes	price	Value of car in British Pounds
year_2009	'0' no or '1' yes	transmission_Automatic	'0' no or '1' yes
year_2010	'0' no or '1' yes	transmission_Manual	'0' no or '1' yes
year_2011	'0' no or '1' yes	transmission_Semi.Auto	'0' no or '1' yes
year_2012	'0' no or '1' yes	mileage	Total miles on car
year_2013	'0' no or '1' yes	fuelType_Diesel	'0' no or '1' yes
year_2014	'0' no or '1' yes	fuelType_Hybrid	'0' no or '1' yes
year_2015	'0' no or '1' yes	fuelType_Petrol	'0' no or '1' yes
year_2016	'0' no or '1' yes	tax	Tax on sale of car
year_2017	'0' no or '1' yes	mpg	Miles per gallon
year_2018	'0' no or '1' yes	engineSize	Engine size
year_2019	'0' no or '1' yes	Car.Make_Audi	'0' no or '1' yes
year_2020	'0' no or '1' yes	Car.Make_Skoda	'0' no or '1' yes
Car.Make_BMW	'0' no or '1' yes	Car.Make_Toyota	'0' no or '1' yes

Car.Make_Ford	‘0’ no or ‘1’ yes	Car.Make_Vauxhall	‘0’ no or ‘1’ yes
Car.Make_Hyundai	‘0’ no or ‘1’ yes	Car.Make_Volkswagen	‘0’ no or ‘1’ yes
Car.Make_Mercedes	‘0’ no or ‘1’ yes		

(Figure 1: Description of Variables)

The 33 variables utilized in this analysis were listed out and described **(Figure 1)**. Once the variables for the dataset were constructed, cleaning the data followed. The following steps were taken as part of data cleaning. All NA values were removed. Next no years were included from before 2008 and after 2020. 'Electric' and 'Other' were removed from the fuel type variable. 'Other' was removed from the transmission type. These classes were removed as they had no or very few data points of interest. The final dataframe removed 9.31% of the original data. The final dataframe had 98,439 rows. The alpha of 0.05 was utilized for all subsequent statistical analysis.



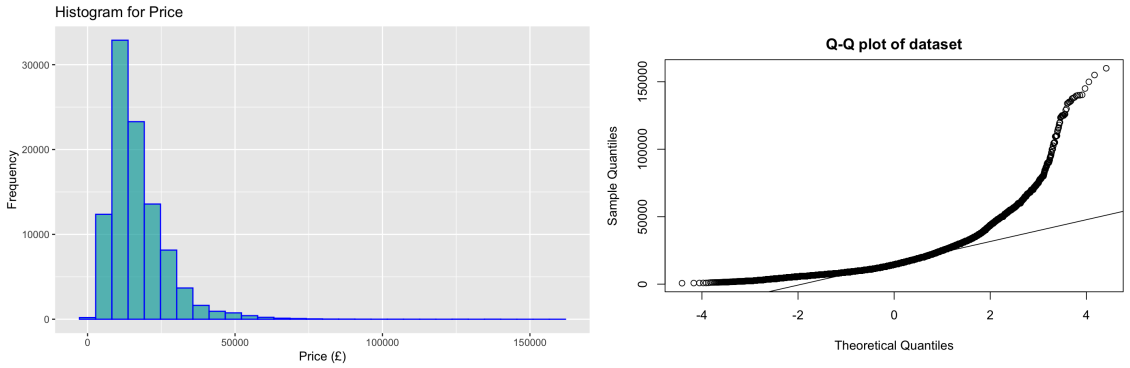
(Figure 2: pie Charts for Car Year, Fuel Types, and Transmission Types)

Pie charts were constructed for the Car years, Fuel Type and Transmission Type to see if any groups had low or no values **(Figure 2)**. The car years before 2008 were removed from the data. The Fuel Types of electric and other were removed from the data. Finally the Transmission type of Other was removed.

Price							
Min	1st Quantile	Median	Mean	3rd Quantile	Max	Standard Deviation	Variance
694	10000	14498	16868	20900	159999	9850	97021765.64

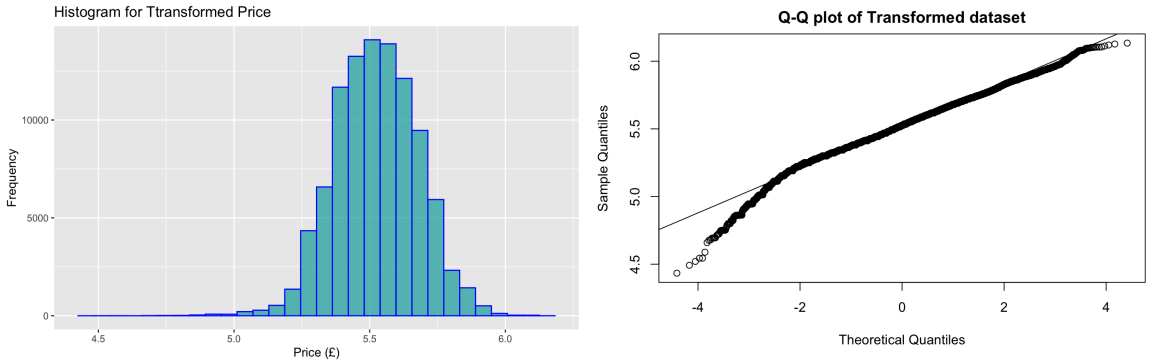
(Figure 3: Descriptive Statistics for Price)

The summary Statistics indicate a right skew in the data as the median is less than the mean **(Figure 3)**. Normality should be assessed on the price variable and a proper transformation should be applied.



(Figure 4: Histogram and Q-Q plot of Raw Price data)

The histogram and Q-Q plot were constructed for the price data to assess the normality of the data **(Figure 4)**. The price data appeared very skewed to the right from the histogram. The Q-Q plot confirmed this assessment as by the 2nd and 4th theoretical quantile, the data appeared to drastically increase which indicated not normally distributed data. The Box-Cox procedure was implemented on the data.

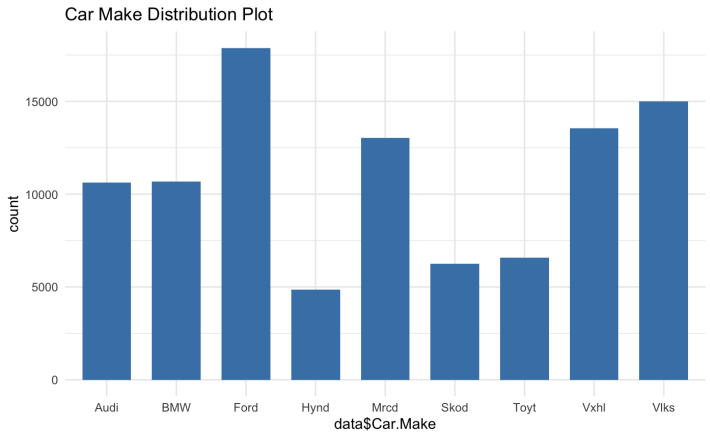


(Figure 5: Histogram and Q-Q plot of Transformed Price data)

The Box-Cox procedure identified a lambda of -.128. The transformation was applied to the pricing data. The histogram for the transformed data appeared normally distributed **(Figure 5)**. The Q-Q plot for the transformed data appeared to have tails moving away from the data at around the -4th to -2 theoretical quantile and the 4th theoretical quantile. The Kolmogorov-Smirnov was applied to the raw data and it yielded a very low p-value of <2e-16. With this information the researchers concluded that the data was not normally distributed. The Kolmogorov-Smirnov was further applied to the transformed data and it too yielded a low

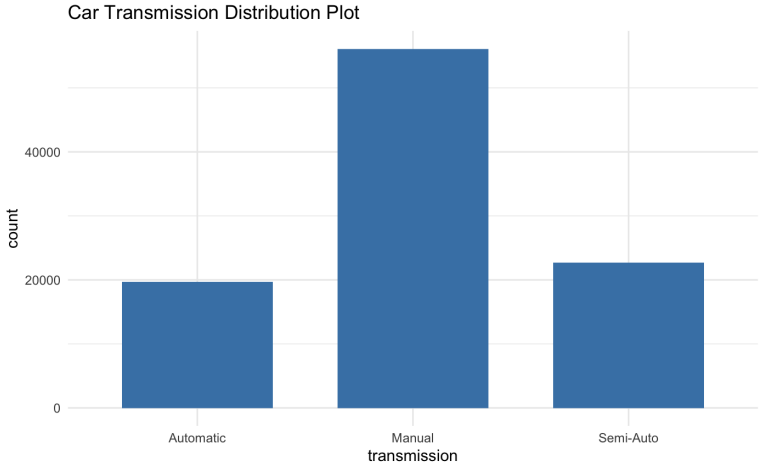
p-value of <2e-16. Since the transformed data was found to mathematically normalize the data, the original raw data was used for further analysis. Additionally analysis on the car data was assessed.

Exploratory Data Analysis:



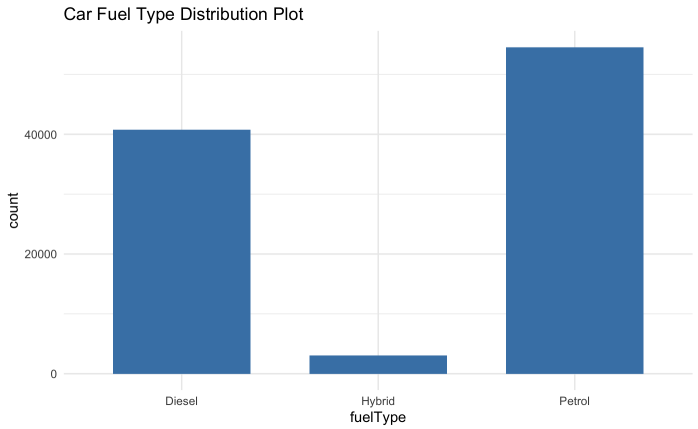
(Figure 6: Bar Chart for Count of Car Makes)

A bar chart of Car Make distributions was constructed to see the types of cars that were being sold in England **(Figure 6).** The most popular car markers were Ford. Vlks, and Vxhl. The least popular Car Makers were Hynd, Skod, and Toyota. The dataset appeared to have a majority of European based models.



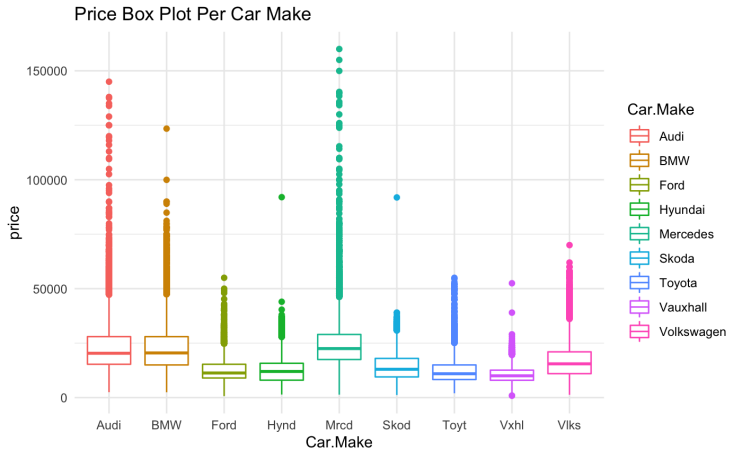
(Figure 7: Bar Chart for Transmission Count)

A graph for the count of transmission types was constructed as well **(Figure 7)**. Manual appeared to be the most common. This was then followed by Semi-Auto adn the Automatic.



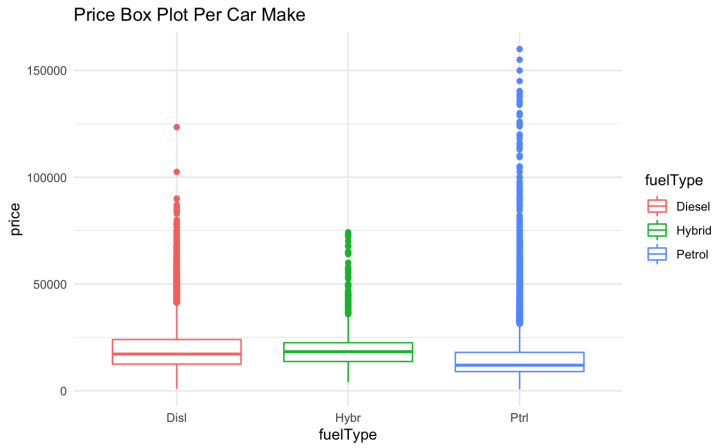
(Figure 8: Bar Chart for Fuel Type Count)

A bar chart was constructed for the count of fuel types **(Figure 8).** As Hybrid cars have only recently been introduced into the market, it made sense to use it in the lowest of fuelTypes for this dataset. Petrol and Diesel based cars have a clear majority in the used car market.



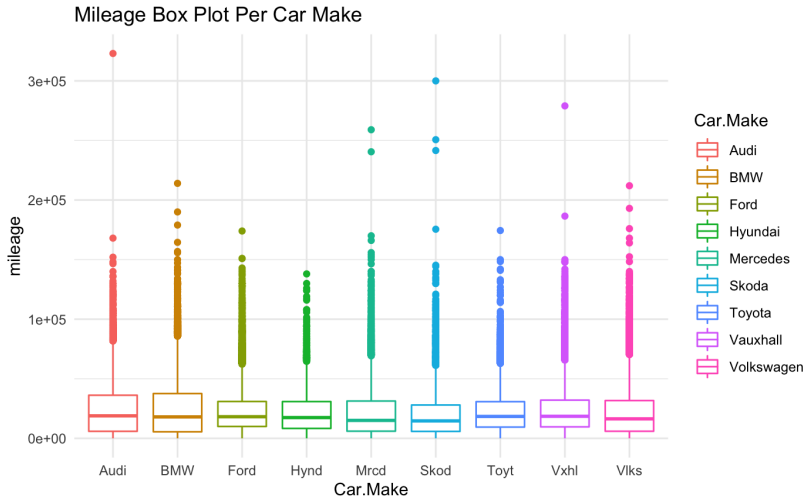
(Figure 9:Box Plot for Price and Car Make)

It appeared that Mercedes had the highest price values for their vehicles followed by Audi and then BMW **(Figure 9)**. the prices for carmakes tended to range around the 25000 pound range. Each box appeared skewed to the right as there were noticeable outliers within each group. Further testing was employed to see if there was any statistical difference between the Car Makes.An ANOVA test for price of different Car Makes had a p-value of <2e-16 which revealed that there was a statistical difference between the groups. Pairwise comparison was introduced to see which groups were different from each other. The Tukey pairwise comparison showed that all Car Make prices were statistically different from each other except for the three groups of BMW and Audi, Toyota and Ford, and Toyota and Hyundai. This analysis revealed that the only Car Makes that have statistically similar prices are Toyota-Hyundai and BMW-Audi. It is important to note that Hyundai and Ford would be statistically different at a 95% confidence but the original alpha was set to .05.



(Figure 10: Box Plot for price and Fuel Type)

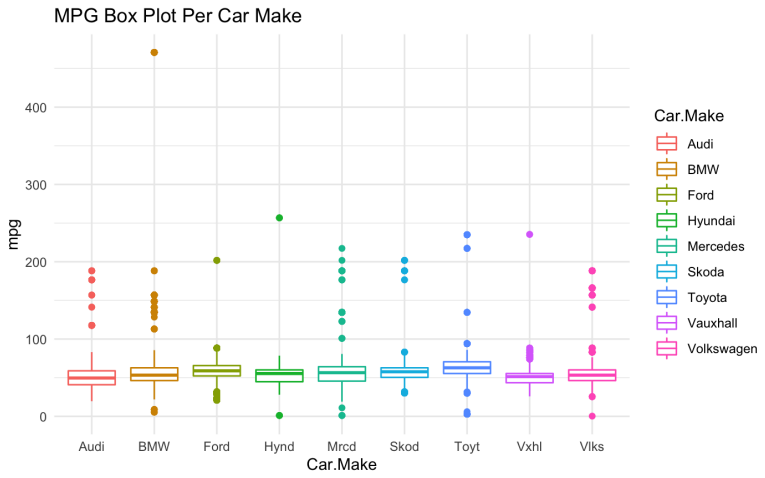
The price of the cars were compared to their respective fuel types **(Figure 10)**. The box plot showed a strong right skew in the data but there was slightly overlap of the three boxes. The ANOVA test on price of cars by fuel type revealed a p-value of <2e-16 which indicated there was a statistical difference between groups of fuel types by prices. The Tukey pairwise comparison revealed Petrol to be statistically different from Diesel and Hybrid.



(Figure 11: Box Plot for Mileage and Car Make)

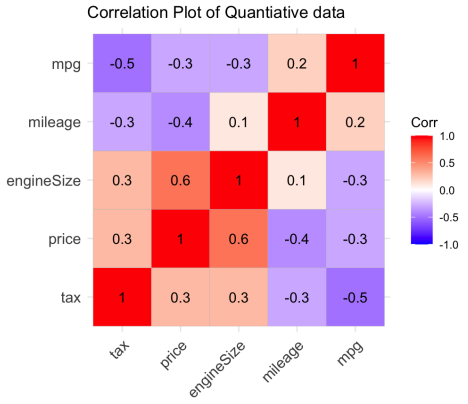
The mileage of the cars were compared to their respective Car Makes **(Figure 11)**. The box plot showed a strong right skew in the data but there was slightly overlap of the three boxes. It appeared that many of the Car Makes had similar mileage numbers as they tended to overlap a similar range of data. The ANOVA test on mileage of Car Makes revealed a p-value of <2e-16 which indicated there was a statistical difference between groups of fuel types by prices. The pairwise comparison revealed the following groups to not be statistically similar to each other were BMW-Audi, Toyota-Ford, Vauxhall-Ford, Mercedes-Hyundai,

Toyota-Hyundai, Volkswagen-Hyundai, Toyota-Mercedes, Volkswagen-Mercedes, Vauxhall-Toyota, and Volkswagen-Toyota.



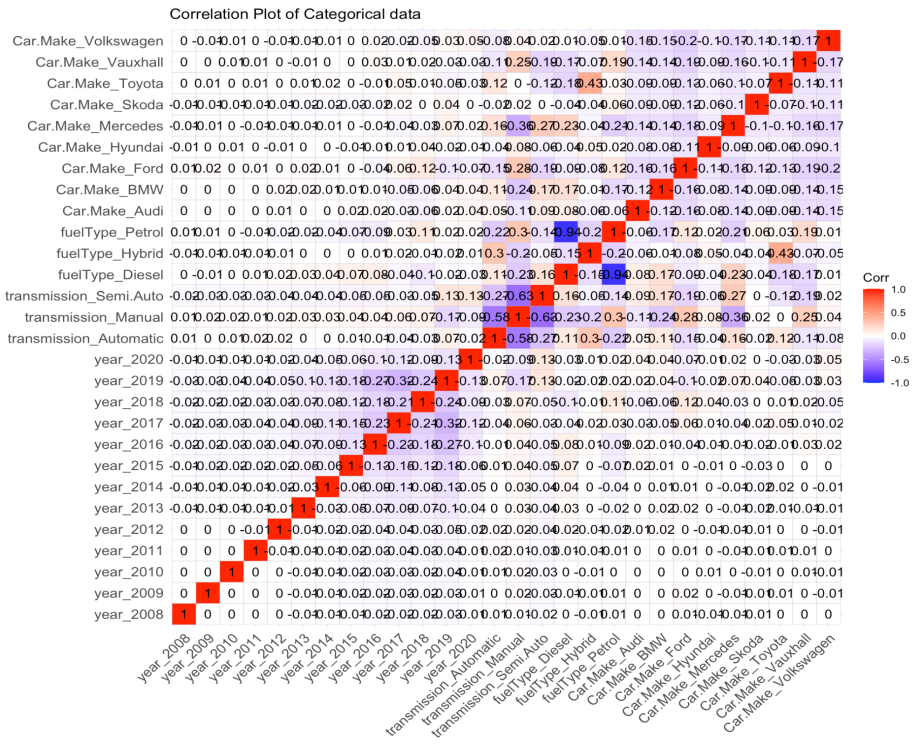
(Figure 12: Box Plot for mpg and Car Make)

A box plot chart was created for mpg and Car Make **(Figure 12)**. Each of the groups showed outliers in the upper quartile range of the data. Most of the means of mpg for each of the Car Makes was around 50. The ANOVA test did reveal there was a statistical difference between the groups of Car Makes by mpg as the p-value was <2e-16. The Tukey test revealed the mpg of Car Makes for the following pairwise comparisons to be statistically similar: Skoda-BMW and Volkswagen-Hyndai. The other pairwise comparisons were all statistically different.



(Figure 13:Correlation Matrix for Quantitative Data)

Correlation matrices were constructed to compare the categorical data and the quantitative data **(Figure 13)**. This step was implemented to check for multicollinearity between the variables. For the quantitative correlation comparisons, Pearson's correlation coefficient was utilized and it revealed engine size to have the strongest correlation with price at 0.6. This was followed by mileage at -.4, tax at 0.3 and mpg at -.3. Although these correlations to price are not very strong, They do have a moderate correlation with the target variable.



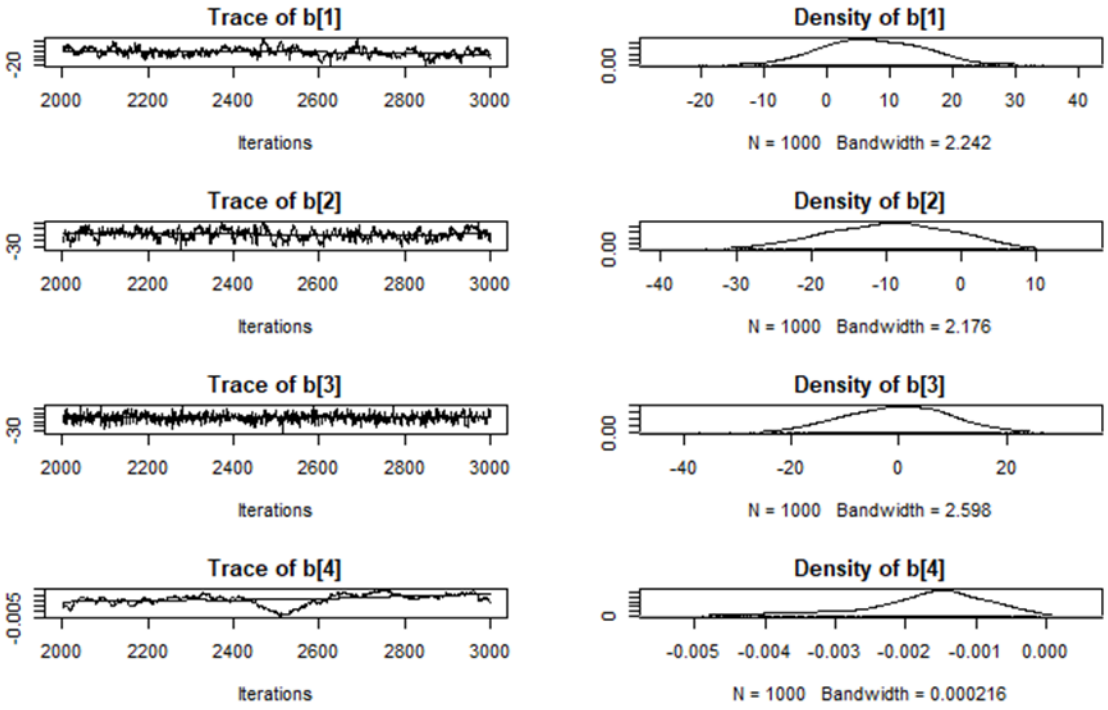
(Figure 14:Correlation Matrix for Categorical Data)

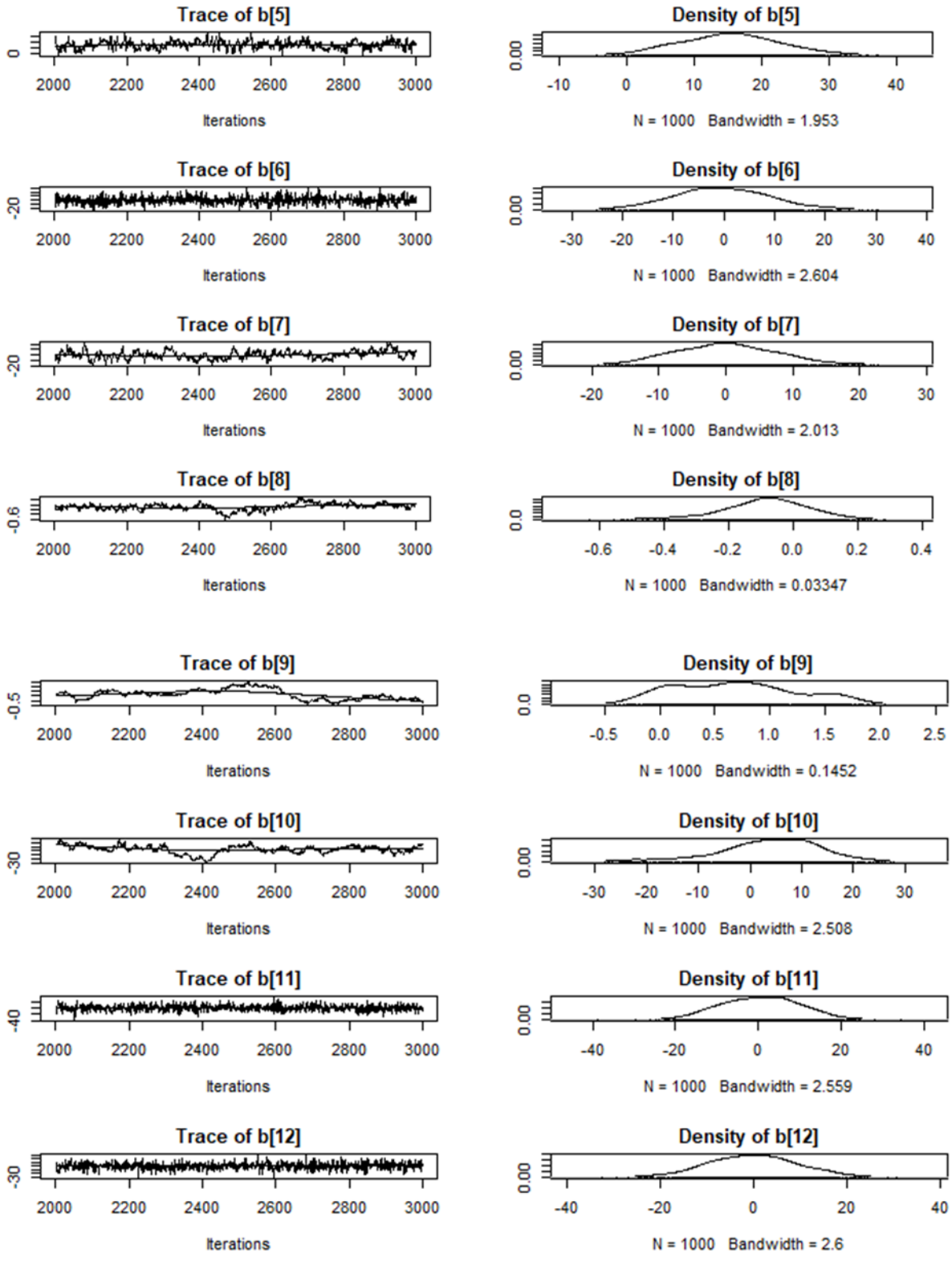
The categorical data was compared using Spearman's correlation coefficient **(Figure 14)**. Only the variables of fuelType_Petrol and fuelType_Diesel had a strong correlation coefficient of -0.94. The final correlation calculation was the Point-Biserial Correlation. This correlation calculation was utilized as the researchers were interested in seeing the relationship between the target value of price and how it related to the different categorical data values. The categorical values that had the strongest correlation with price were transmission_Manual at 0.55, transmission_Semi.Auto at -.412. The other variables had a much lower correlation with the price data. With exploratory data analysis completed, models were subsequently constructed.

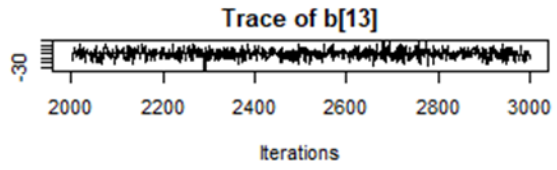
Logistic regression model:

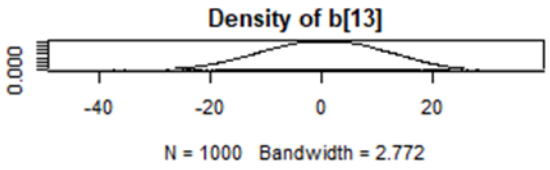
We investigated whether the dependent variable “Price” can be engineered in a different way to be able to apply the Logistic Regression model. A new dependent variable “Price_Range” was defined. This variable consists of binary response. The responses are non-luxury and luxury cars which have values of 0 and 1 respectively. To have balanced classes, the mean value of car price was used as a cutoff to classify both classes, where the values below car price mean represented the non-luxury cars and vice versa.

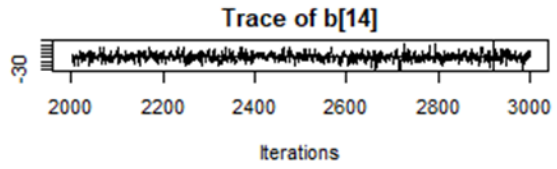
The logistic Regression model was conducted using Jags in R. The dependent variable prior was set to follow Bernoulli distribution and a non-informative priors for all coefficients was set to have a normal distribution with mean of 0 and standard deviation of 0.02. The following plots show the posterior plots for all estimated coefficients and trace plots.

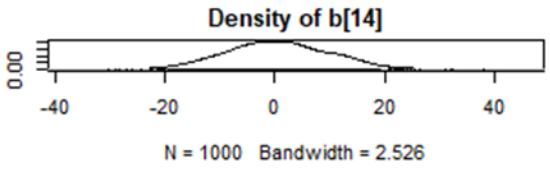


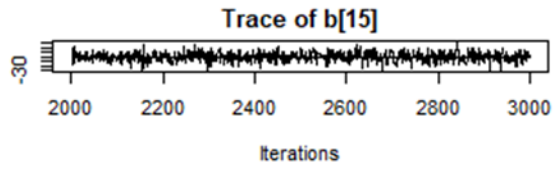

Trace of b[5]
Density of b[5]
N = 1000 Bandwidth = 1.953
Trace of b[6]
Density of b[6]
N = 1000 Bandwidth = 2.604
Trace of b[7]
Density of b[7]
N = 1000 Bandwidth = 2.013
Trace of b[8]
Density of b[8]
N = 1000 Bandwidth = 0.03347
Trace of b[9]
Density of b[9]
N = 1000 Bandwidth = 0.1452
Trace of b[10]
Density of b[10]
N = 1000 Bandwidth = 2.508
Trace of b[11]
Density of b[11]
N = 1000 Bandwidth = 2.559
Trace of b[12]
Density of b[12]
N = 1000 Bandwidth = 2.6
Iterations

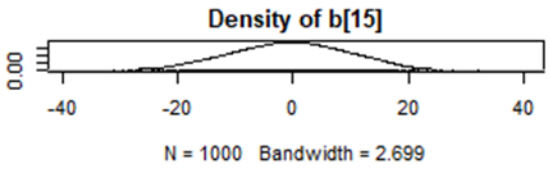

Trace of b[13]
-30
2000
2200
2400
2600
2800
3000
Iterations

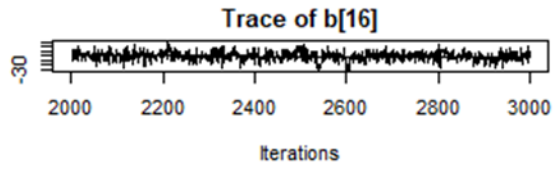

Density of b[13]
0.000
-40
-20
0
20
N = 1000 Bandwidth = 2.772

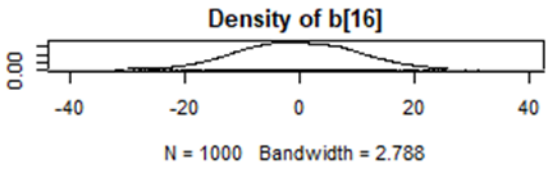

Trace of b[14]
-30
2000
2200
2400
2600
2800
3000
Iterations

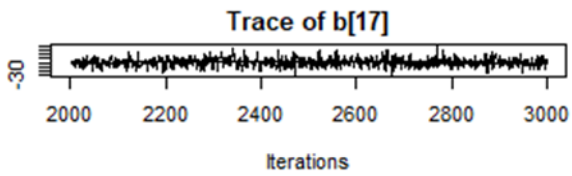

Density of b[14]
0.00
-40
-20
0
20
40
N = 1000 Bandwidth = 2.526

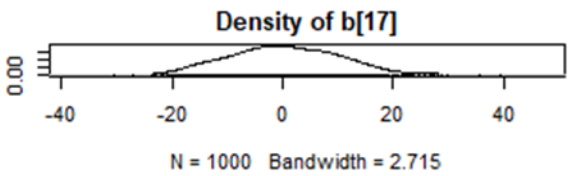

Trace of b[15]
-30
2000
2200
2400
2600
2800
3000
Iterations

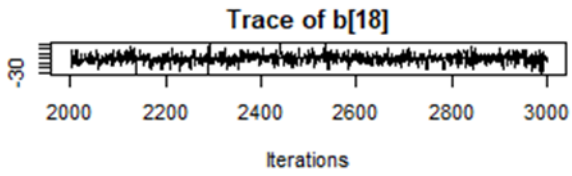

Density of b[15]
0.00
-40
-20
0
20
40
N = 1000 Bandwidth = 2.699

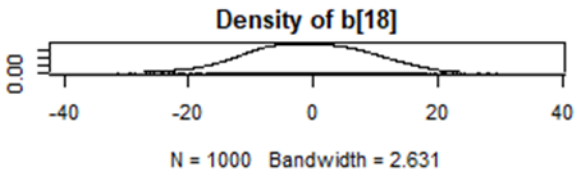

Trace of b[16]
-30
2000
2200
2400
2600
2800
3000
Iterations

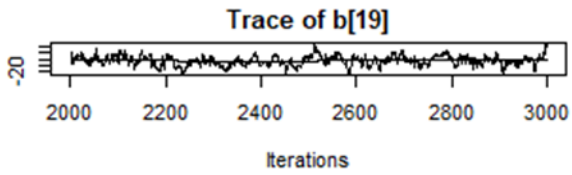

Density of b[16]
0.00
-40
-20
0
20
40
N = 1000 Bandwidth = 2.788

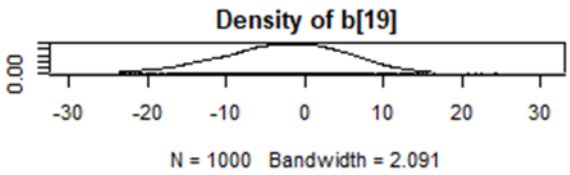

Trace of b[17]
-30
2000
2200
2400
2600
2800
3000
Iterations

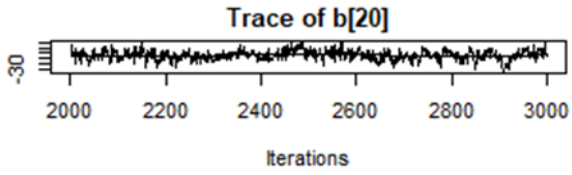

Density of b[17]
0.00
-40
-20
0
20
40
N = 1000 Bandwidth = 2.715

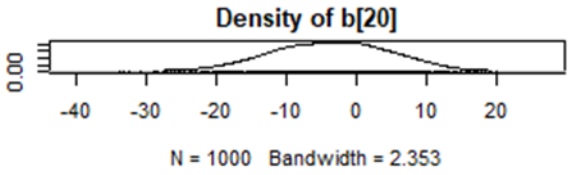

Trace of b[18]
-30
2000
2200
2400
2600
2800
3000
Iterations

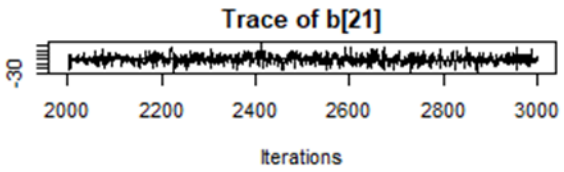

Density of b[18]
0.00
-40
-20
0
20
40
N = 1000 Bandwidth = 2.631

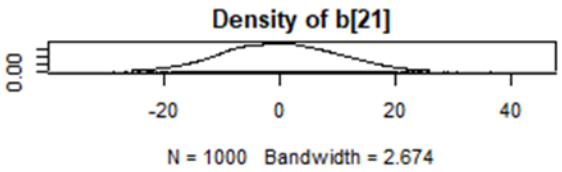

Trace of b[19]
-20
2000
2200
2400
2600
2800
3000
Iterations

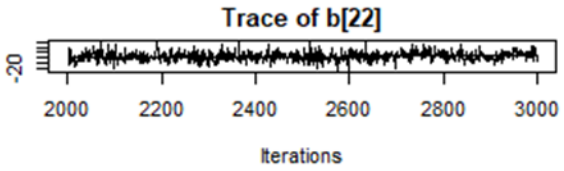

Density of b[19]
0.00
-30
-20
-10
0
10
20
30
N = 1000 Bandwidth = 2.091

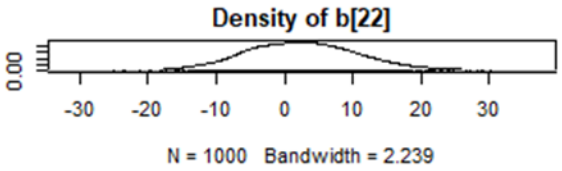

Trace of b[20]
-30
2000
2200
2400
2600
2800
3000
Iterations

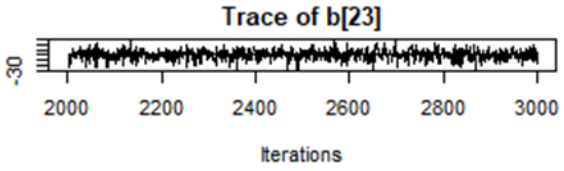

Density of b[20]
0.00
-40
-30
-20
-10
0
10
20
N = 1000 Bandwidth = 2.353

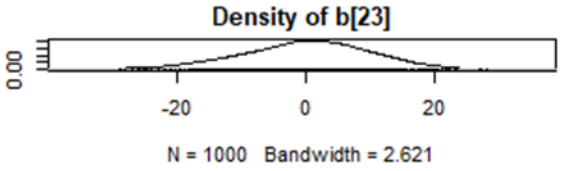

Trace of b[21]
-30
2000
2200
2400
2600
2800
3000
Iterations

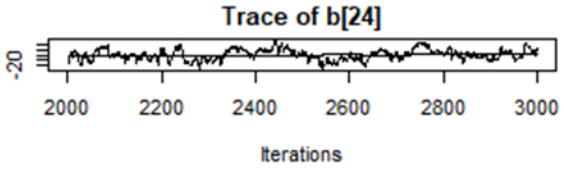

Density of b[21]
0.00
-20
0
20
40
N = 1000 Bandwidth = 2.674

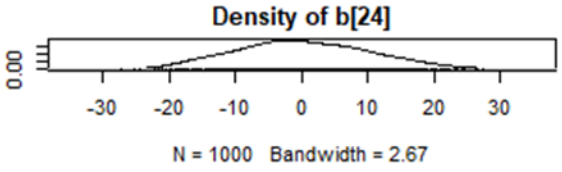

Trace of b[22]
-20
2000
2200
2400
2600
2800
3000
Iterations

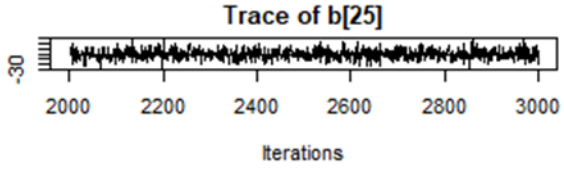

Density of b[22]
0.00
-30
-20
-10
0
10
20
30
N = 1000 Bandwidth = 2.239

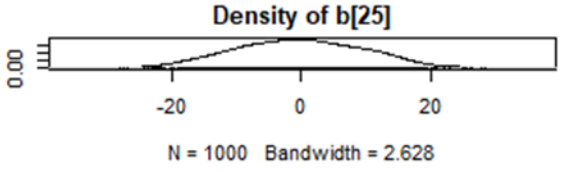

Trace of b[23]
-30
2000
2200
2400
2600
2800
3000
Iterations

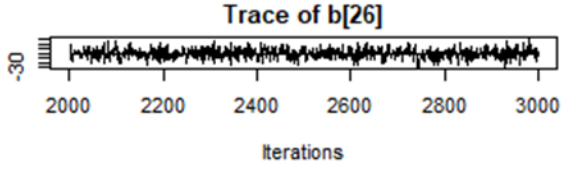

Density of b[23]
0.00
-20
0
20
N = 1000 Bandwidth = 2.621

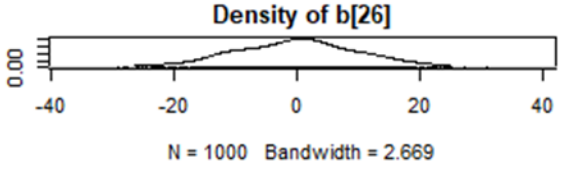

Trace of b[24]
-20
2000
2200
2400
2600
2800
3000
Iterations

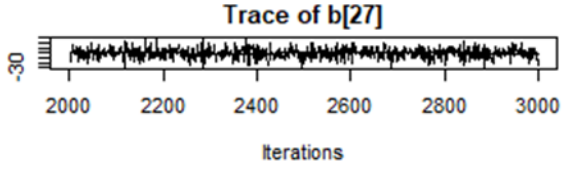

Density of b[24]
0.00
-30
-20
-10
0
10
20
30
N = 1000 Bandwidth = 2.67

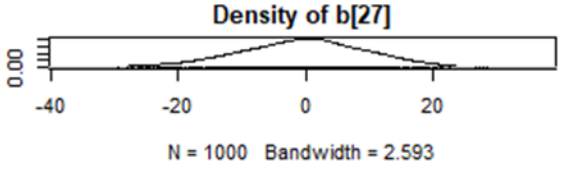

Trace of b[25]
-30
2000
2200
2400
2600
2800
3000
Iterations

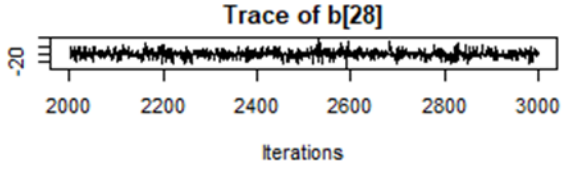

Density of b[25]
0.00
-20
0
20
N = 1000 Bandwidth = 2.628

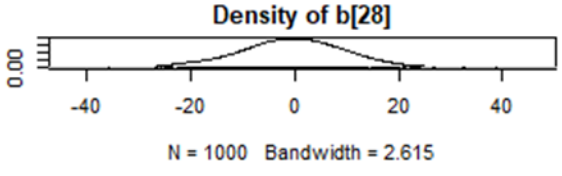

Trace of b[26]
-30
2000
2200
2400
2600
2800
3000
Iterations

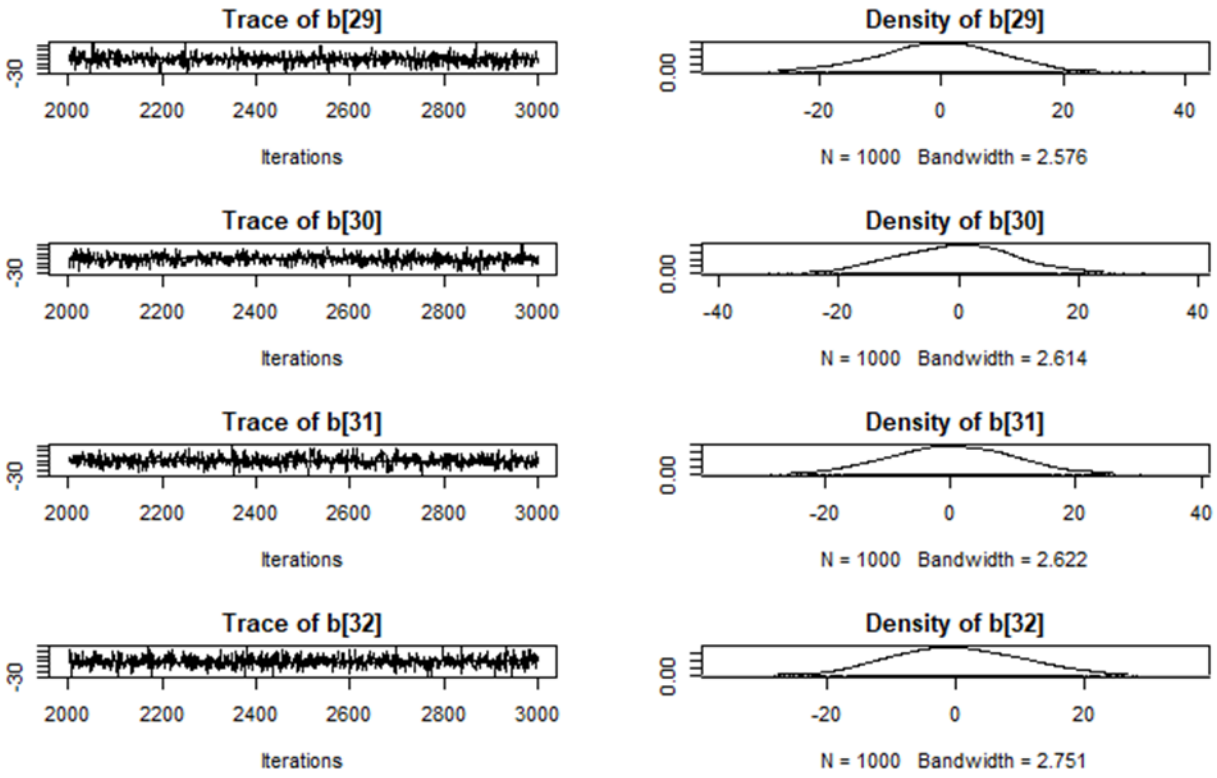

Density of b[26]
0.00
-40
-20
0
20
40
N = 1000 Bandwidth = 2.669


Trace of b[27]
-30
2000
2200
2400
2600
2800
3000
Iterations


Density of b[27]
0.00
-40
-20
0
20
N = 1000 Bandwidth = 2.593


Trace of b[28]
-20
2000
2200
2400
2600
2800
3000
Iterations


Density of b[28]
0.00
-40
-20
0
20
40
N = 1000 Bandwidth = 2.615



Calculating the probabilities and applying the model for prediction resulted in a model accuracy of 75%. The following table;

	Predicted (Non-Luxury)	**Predicted (Luxury)**
Actual (Non-Luxury)	**42,282**	**17,236**
Actual (Luxury)	**6,893**	**32,028**

Hierarchical model:

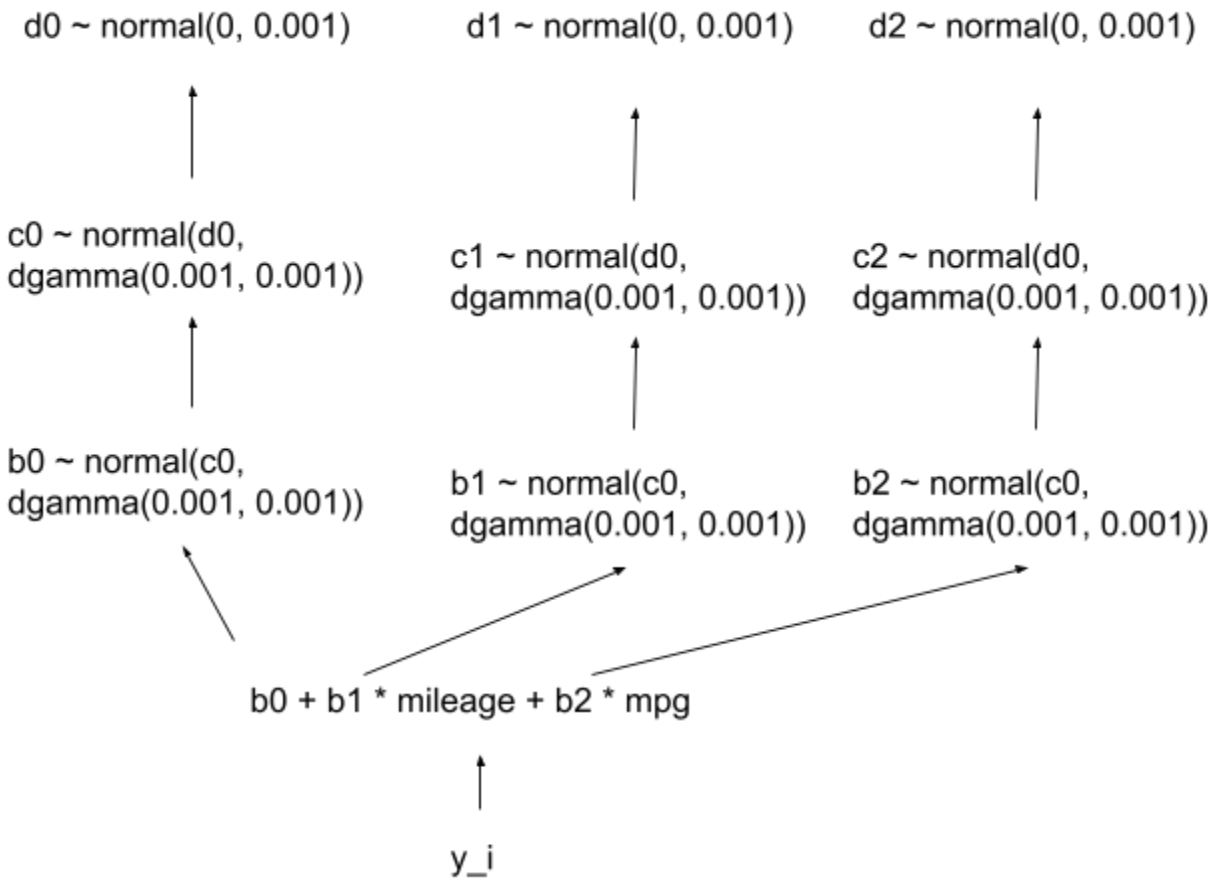
To apply a concept learned in class, a hierarchical model was applied. The concept came from the idea of coins and coin factories. In this application, a car maker makes cars, and the car has a price.

The model is as follows:

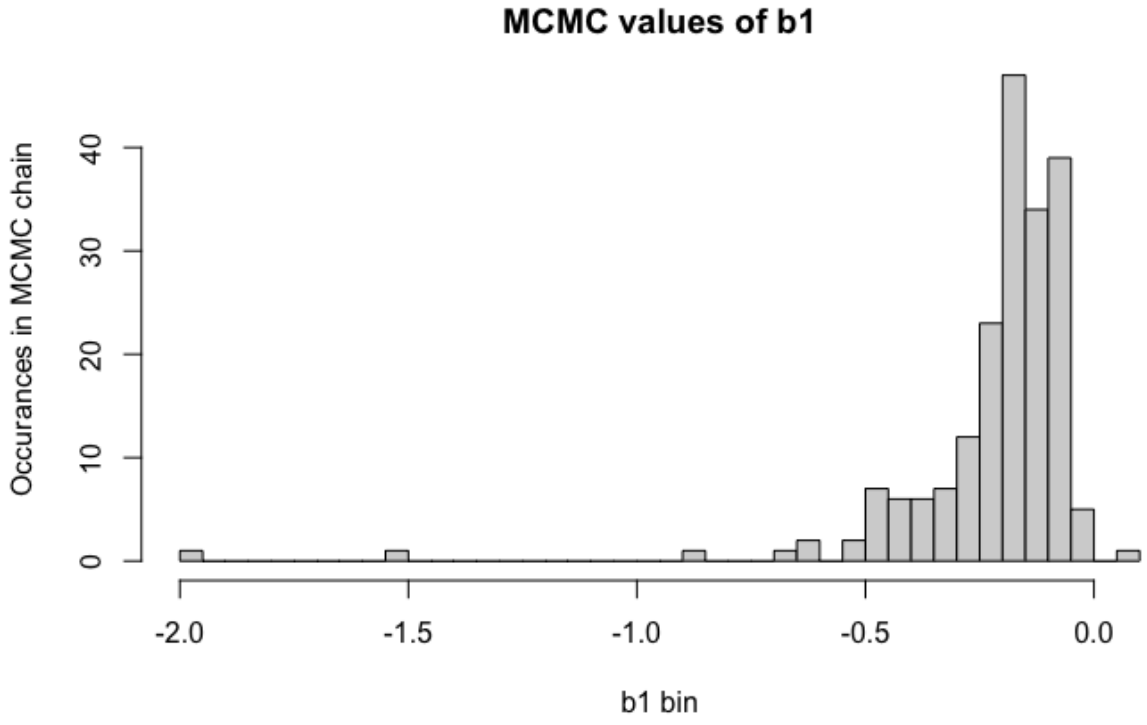
$$y \sim b0 \ + \ b1 \ * \ mileage \ + \ b2 \ * \ mpg$$

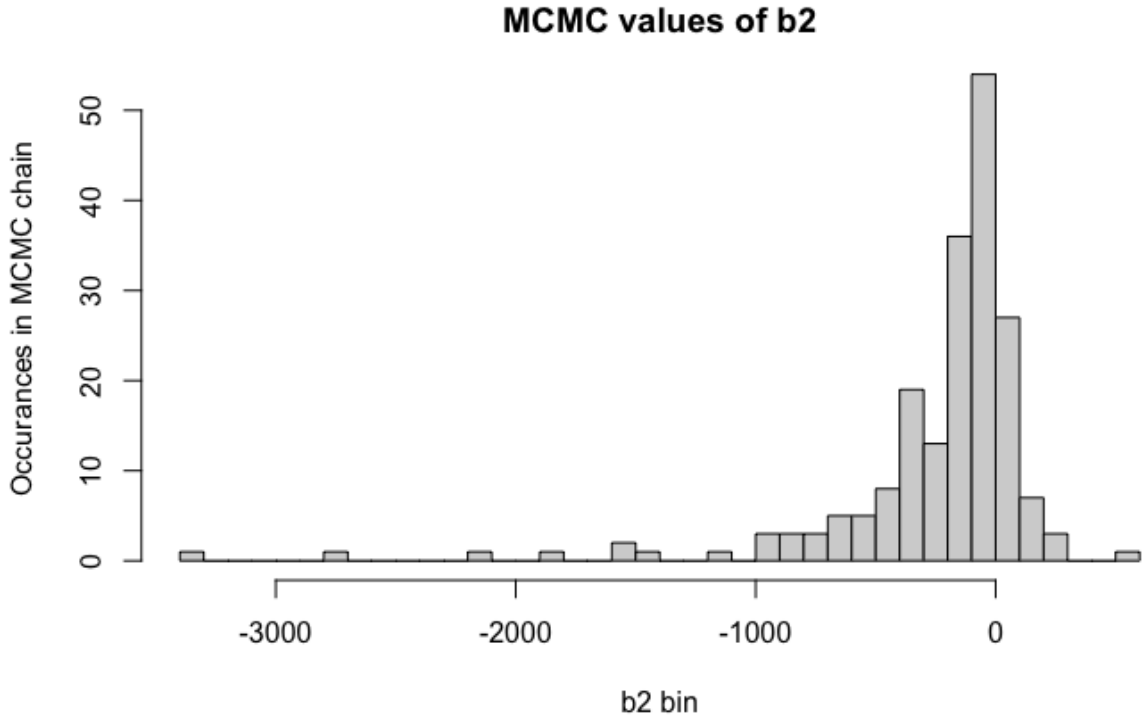
To incorporate the hierarchical aspect, there are two new levels, one for car models and one for carmakers.

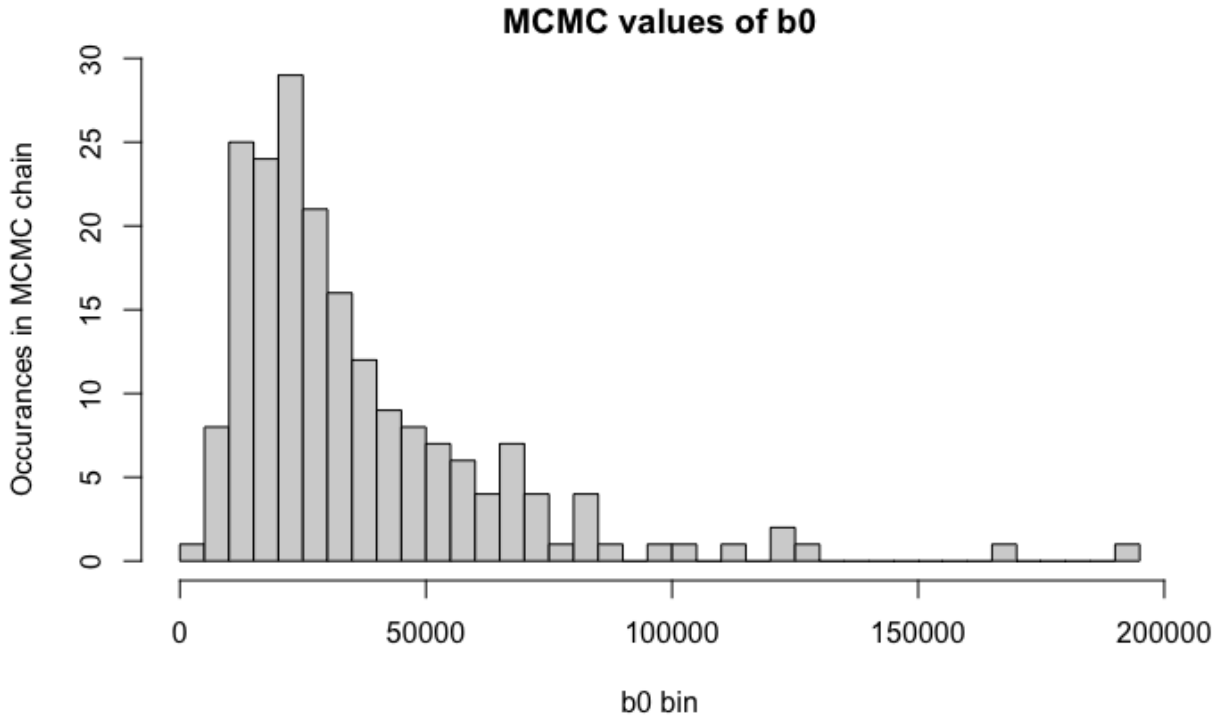
For carmakers, d0/d1/d2 represent priors for the parameters in the linear regression model. Those priors go into c0/c1/c2, and those in turn go into b0/b1/b2.



The priors were initialized with normal distributions with gamma variance. Then MCMC was run with 1000 iterations. Below are the mean values of b0/b1/b2 from the 1000 iterations.


MCMC values of b1
Occurances in MCMC chain
0
10
20
30
40
-2.0
-1.5
-1.0
-0.5
0.0
b1 bin


MCMC values of b2
Occurances in MCMC chain
0
10
20
30
40
50
-3000
-2000
-1000
0
b2 bin

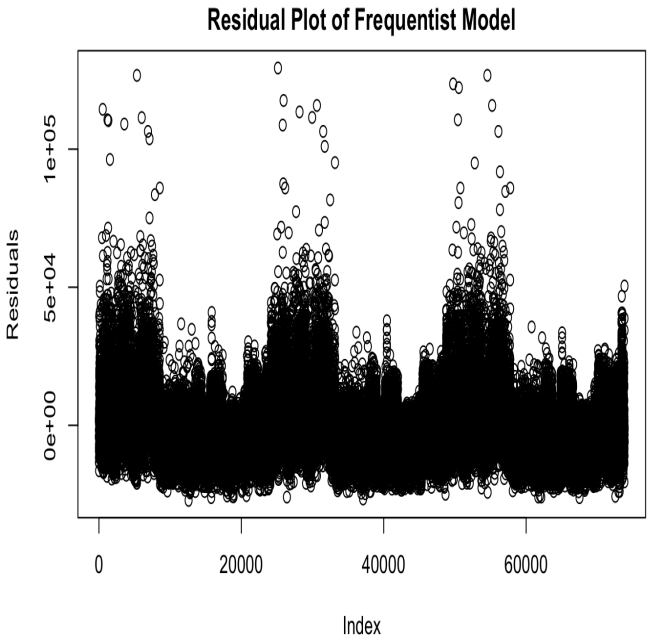


The median value of b0, b1, and b2 were used with the x1 and x2 variables to predict y. The MSE was 76088524, meaning our average prediction was off by around $8722. There was a wide variance in the values that MCMC produced, and that was likely due to a weak, uninformed prior. In the future, a stronger prior could be used.

Linear regression model:

A linear model was constructed to predict the price of used cars. The variables utilized in this regression model were the 12 years. Issues were encountered in deploying the full model so the variables that did work were selected for this analysis. A frequentist model and Bayesian based model were constructed and compared by looking at their respective MSE on the training data which was 75% (73,829) of the cleaned data. The uninformed prior of N(0,0.001) was assigned to each of the 12 predictors for the Bayesian model. The frequentist model was constructed first

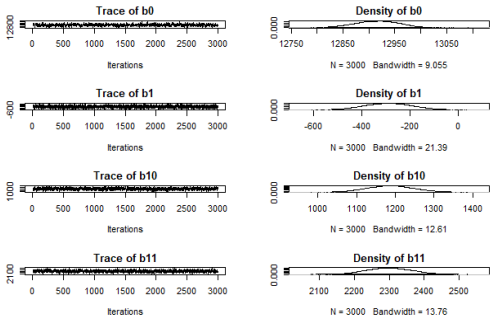
```
Call:
lm(formula = price ~ year_2008 + year_2009 + year_2010 + year_2011 +
    year_2012 + year_2013 + year_2014 + year_2015 + year_2016 +
    year_2017 + year_2018 + year_2019 + year_2020, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-19494  -4679  -1487   3367 143232

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    28484        143   199.6   <2e-16 ***
year_20081    -24573        683   -36.0   <2e-16 ***
year_20091    -24051        591   -40.7   <2e-16 ***
year_20101    -23046        520   -44.4   <2e-16 ***
year_20111    -21768        480   -45.3   <2e-16 ***
year_20121    -20935        412   -50.8   <2e-16 ***
year_20131    -19852        234   -84.8   <2e-16 ***
year_20141    -18500        206   -89.8   <2e-16 ***
year_20151    -16855        180   -93.8   <2e-16 ***
year_20161    -15236        162   -94.3   <2e-16 ***
year_20171    -14107        157   -90.0   <2e-16 ***
year_20181    -12195        164   -74.2   <2e-16 ***
year_20191     -4903        154   -31.8   <2e-16 ***
year_20201        NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8260 on 73816 degrees of freedom
Multiple R-squared:  0.305,     Adjusted R-squared:  0.305
F-statistic: 2.7e+03 on 12 and 73816 DF,  p-value: <2e-16
```

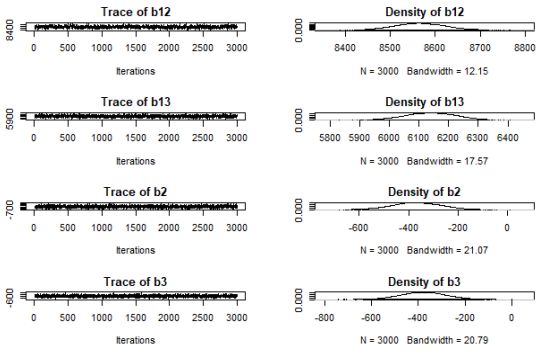


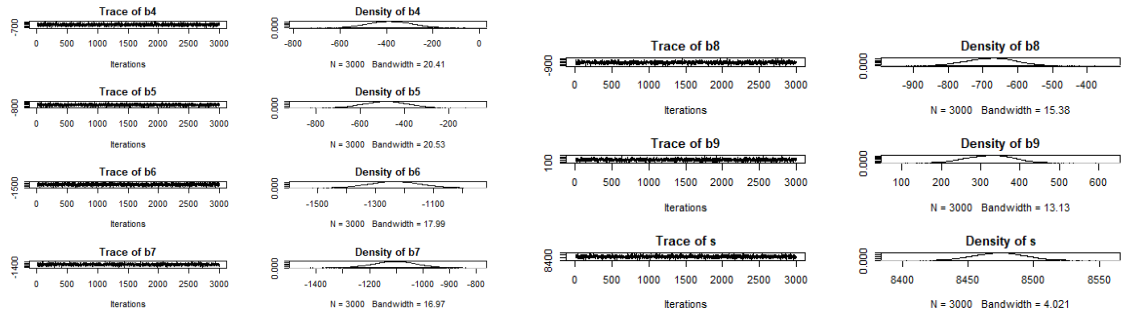
(Frequentist Linear Regression Model and Residual Plot)

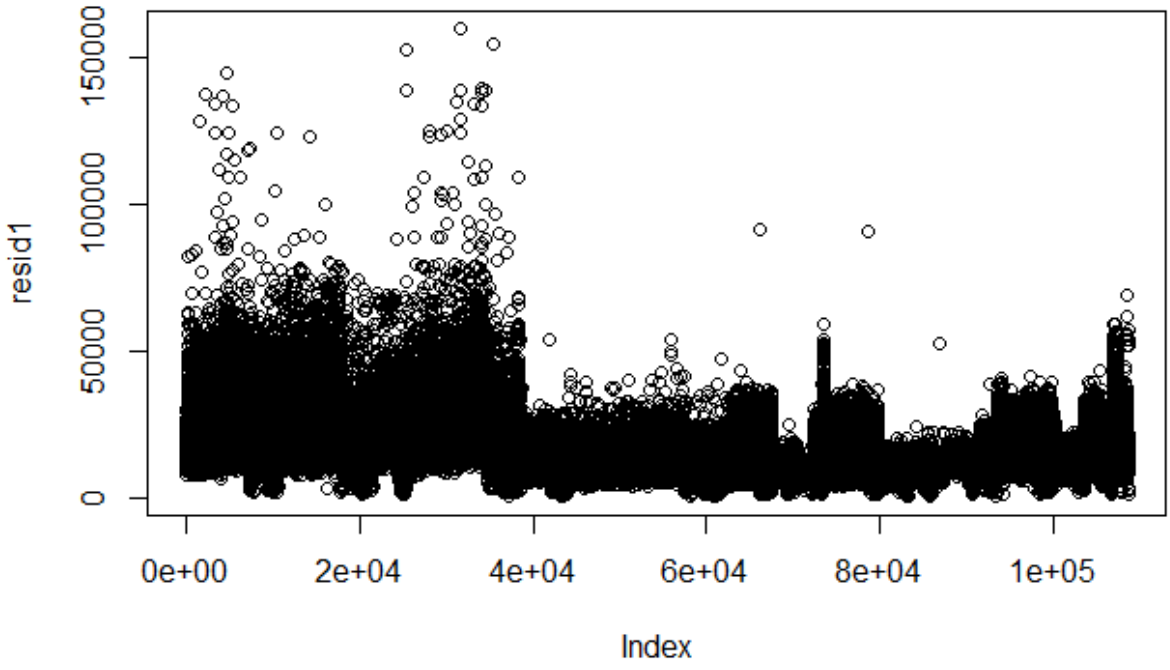
The results of the frequentist model were recorded. The model had a low adjusted r squared value of 0.305. This value indicated that the model was not explaining most of the variation in the data. Regardless, all of the coefficient values are statistically significant. The MSE of the frequentist model was - 9,124,511,777,148.26.

Next, the Bayesian model was constructed. The trace plots and coefficient distributions were constructed to see if there were any abnormalities. For trace plots, this would mean that there was a pattern in the data. For the density plots the coefficient values would not form a normal distribution. 1000 samples were made for the Bayesian model.









(Bayesian Trace, Density, and Residual Plots)

The Density and Trace plots for the 12 coefficients appeared to be normally distributed. THe residual plot did show a slight trend in the data with large increases in outliers from the first to index 4000. The MSE of the Baysian model was found to be 3,253,350,353,165,.63. This value is three times closer to zero than the MSE of the frequentist model.

Conclusion:

The prior cleaning steps for the data **(Figures 1-3)** were very useful in adjusting the data to fit the target value of price. The Bayesian logistic regression predicted 75% of the data correctly. The Bayesian linear model predicted price three times as well as the frequentist model. Some further steps that can be taken in this analysis would be to remove outliers from the data

model. The frequentist logistic regression model should be constructed as well to compare it to the Baysian based model. Finally, further research should be explored in trying to construct a full linear Bayesian model.

The hierarchical linear model suffered from a weak, uninformed prior. Doing more research on the priors and feature selection would have improved MSE and made the model more useful. We could have set b1 as slightly negative with a wide variance (as more mileage tends to lower price) and b2 as slightly positive with a wide variance (as more mpg tends to increase price). We could have used a t-distribution in the hierarchy between the linear equation and y_i (as opposed to simply a normally distributed y). This would better account for outliers in the data and set up a robust GLM model as we discussed in class. The t-distribution has three parameters, and one of those can extend the tails to capture outliers.