Project 3

Group 32

2022-11-20

I: Panel Data Models

1. Intro

Hello, we are Project Group #32 with Aaron Chien, Shaune Le, Yoojin Min, and Doris Wu. With our dataset and model, we are attempting to answer the economic/business/finance question about how sales or emphasis on the prices of cigars are affected by the state's statistics. Our cigar dataset, from the Ecdat package, was collected from 1963 to 1992, and it is a panel data of 50 observations with the data collected from 50 states and Washington DC. The column names or variables are state, year, price (price per pack of cigarettes in dollars), pop (population in millions), pop16 (population above the age of 16 in millions), cpi (consumer price index), ndi (per capita disposable income), sales, and pimin (minimum price in adjoining states per pack of cigarettes in dollars). The dimension of our panel data is 1380 x 9, or 1380 rows and 9 columns. After examining our index of state, we can see that they each have the same value of observations (30), making our data a balanced panel. Our index is 50 and time is 30, so we have relatively wide and short panel data.

```
# reading in initial data
data(Cigar)
head(Cigar)
```

```
##   state year price  pop  pop16  cpi      ndi sales pimin
## 1     1   63  28.6 3383 2236.5 30.6 1558.305  93.9  26.1
## 2     1   64  29.8 3431 2276.7 31.0 1684.073  95.4  27.5
## 3     1   65  29.8 3486 2327.5 31.5 1809.842  98.5  28.9
## 4     1   66  31.5 3524 2369.7 32.4 1915.160  96.4  29.5
## 5     1   67  31.6 3533 2393.7 33.4 2023.546  95.5  29.6
## 6     1   68  35.6 3522 2405.2 34.8 2202.486  88.4  32.0
```

```
dim(Cigar)
```

```
## [1] 1380    9
```

Doing some data cleaning, we are first trying to see if there are any missing values.

```
lapply(Cigar, function(x) {which(is.na(x))})
```

```
## $state
## integer(0)
```

```
##
## $year
## integer(0)
##
## $price
## integer(0)
##
## $pop
## integer(0)
##
## $pop16
## integer(0)
##
## $cpi
## integer(0)
##
## $ndi
## integer(0)
##
## $sales
## integer(0)
##
## $pimin
## integer(0)
```

Since there are no NA values, we can safely continue.

```
table(Cigar$state)
```

```
##
##  1  3  4  5  7  8  9 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
## 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30
## 30 31 32 33 35 36 37 39 40 41 42 43 44 45 46 47 48 49 50 51
## 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30
```

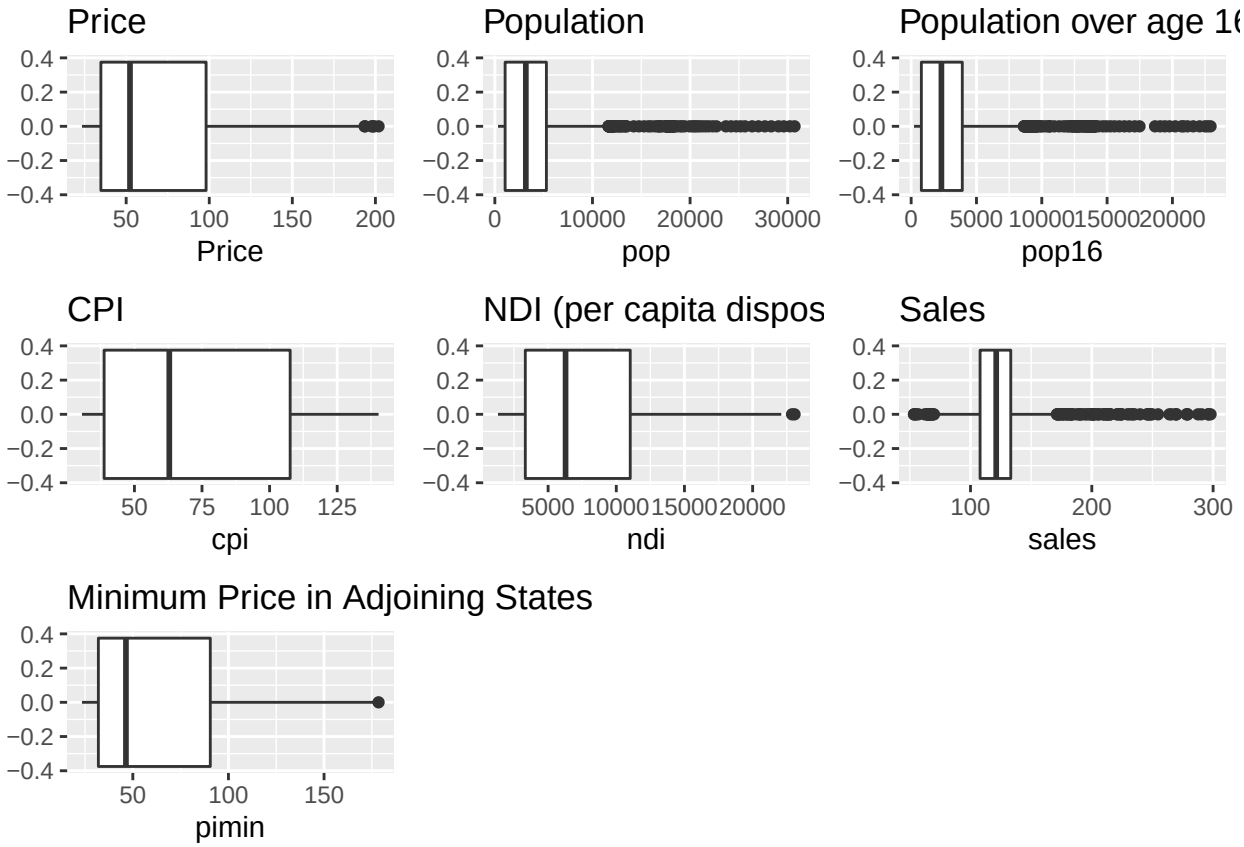
Examining our index of state, we see that they each have the same value of observations, making this a balanced panel.

2

2.1 : Boxplots of Continous Variables

```
cigar_price <- Cigar %>%
  ggplot(mapping = aes(x = price)) +
  geom_boxplot() +
  ggtitle("Price") +
  labs(x = "Price")

cigar_pop <- Cigar %>%
  ggplot(mapping = aes(x = pop)) +
  geom_boxplot() +
```

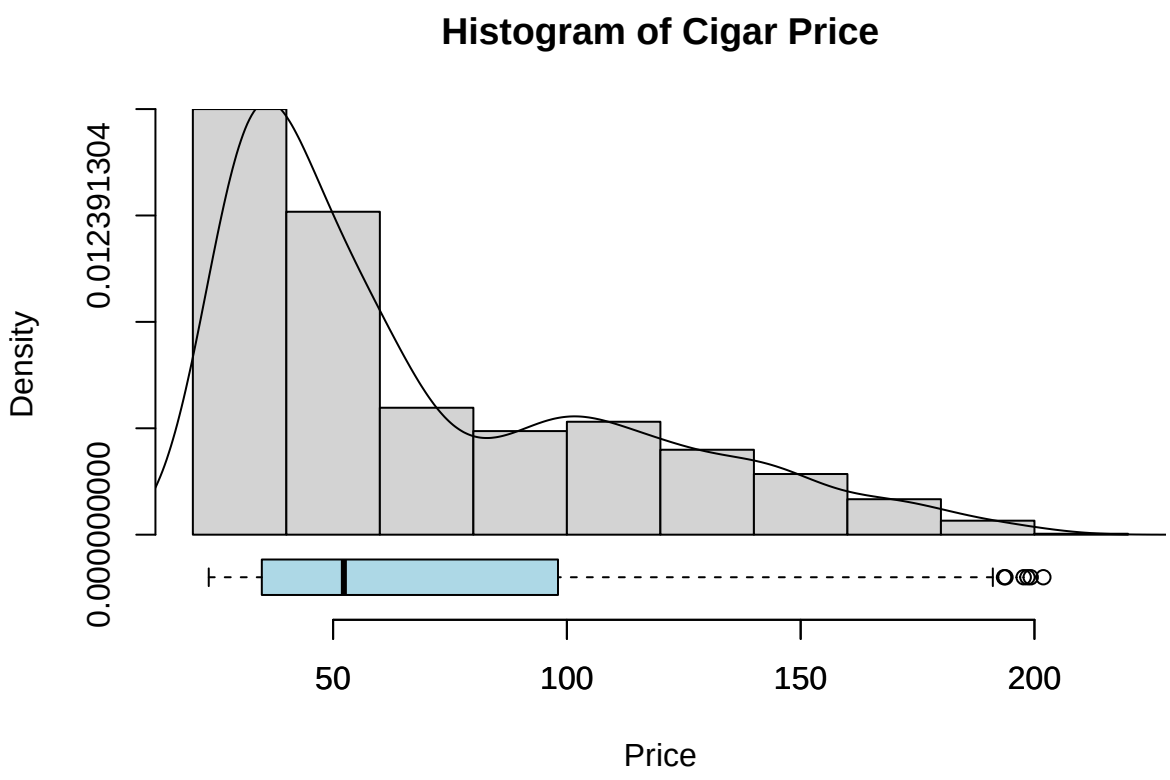
```
  ggtitle("Population") +
  labs(x = "pop")

cigar_pop16 <- Cigar %>%
  ggplot(mapping = aes(x = pop16)) +
  geom_boxplot() +
  ggtitle("Population over age 16") +
  labs(x = "pop16")

cigar_cpi <- Cigar %>%
  ggplot(mapping = aes(x = cpi)) +
  geom_boxplot() +
  ggtitle("CPI") +
  labs(x = "cpi")

cigar_ndi <- Cigar %>%
  ggplot(mapping = aes(x = ndi)) +
  geom_boxplot() +
  ggtitle("NDI (per capita disposable income)") +
  labs(x = "ndi")

cigar_sales <- Cigar %>%
  ggplot(mapping = aes(x = sales)) +
  geom_boxplot() +
  ggtitle("Sales") +
  labs(x = "sales")

cigar_pimin <- Cigar %>%
  ggplot(mapping = aes(x = pimin)) +
  geom_boxplot() +
  ggtitle("Minimum Price in Adjoining States") +
  labs(x = "pimin")

plot_grid(cigar_price, cigar_pop, cigar_pop16, cigar_cpi, cigar_ndi, cigar_sales, cigar_pimin, ncol = 3
```



- The boxplot of price indicates that the data for price is heavily skewed to the right as the median is closer to the bottom of the box, and the whisker is shorter on the lower end of the box. There are some high outliers present.
- The boxplot of population indicates that the data for population is skewed to the right. Although the median is close to the mean, the whisker is much longer on the higher end of the box, and there are a lot of high outliers present. Moreover, since there are a lot of outliers in this data, population may not be a good use in predicting sales.
- The boxplot of population above the age of 16 indicates that the data is skewed to the right. Although the median is close to the mean, the whisker is much longer on the higher end of the box, and there are a lot of high outliers present.
- The boxplot of CPI indicates that the data is slightly skewed to the right because median is slightly closer to the bottom of the box, and the whisker is longer on the higher end of the box. But there are no outliers present.
- The boxplot of NDI indicates that the data for NDI is skewed to the right as the median is slightly closer to the bottom of the box, and the whisker is shorter on the lower end of the box. There are a few high outliers present.
- The boxplot of sales indicates that the data for sales is skewed to the right. Although the median is close to the mean, the whisker is longer on the higher end of the box. Also, there are a lot more high outliers to the right rather than lower outliers. Moreover, there are a lot of outliers in sales rather than price, so it may be better to predict price rather than sales.
- The boxplot of the minimum price in adjoining states indicates that the data is skewed to the right because median is closer to the bottom of the box, and the whisker is longer on the higher end of the box. There seems to be 1 high outlier present.

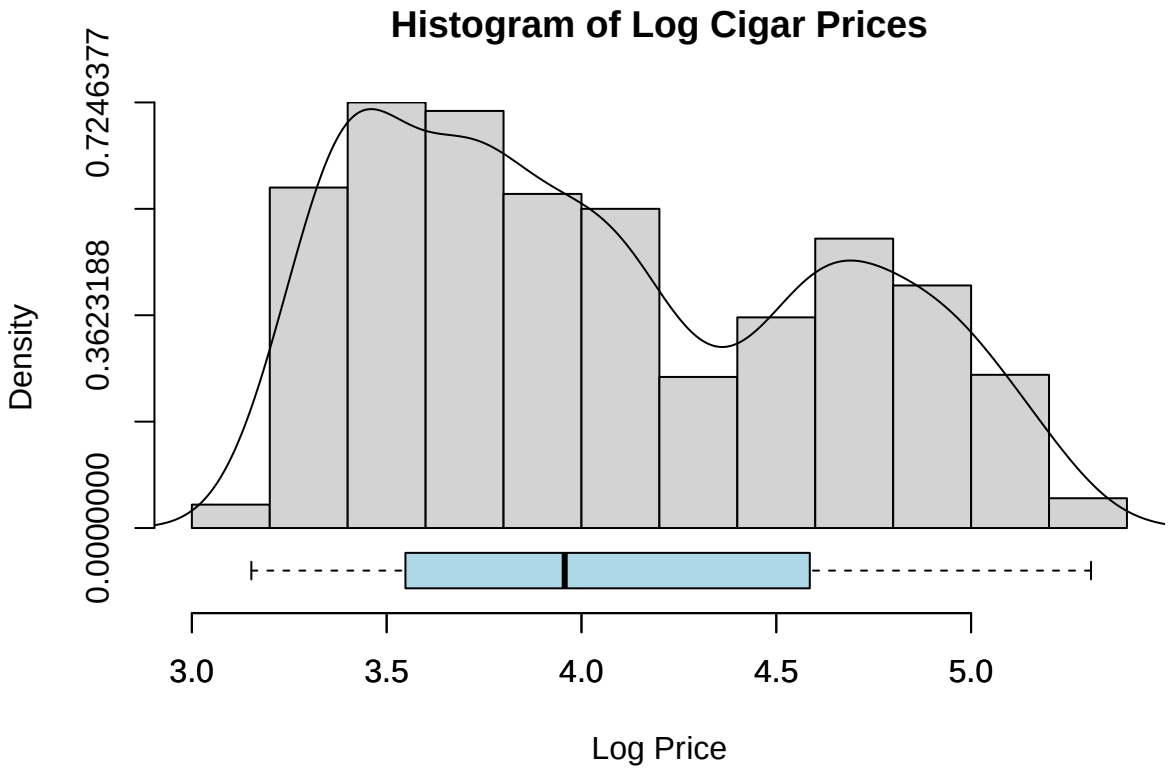
2.3: Histogram and Density Line of Variables

```
hist_boxplot(Cigar$price, freq = FALSE,
             main = "Histogram of Cigar Price",
             xlab = "Price")
lines(density(Cigar$price))
```



Examining the histogram of Cigar Price, we see that it is skewed to the right, so the median is higher than the median. Since this histogram is skewed, we should use the log of the price in order to make it more normally distributed.

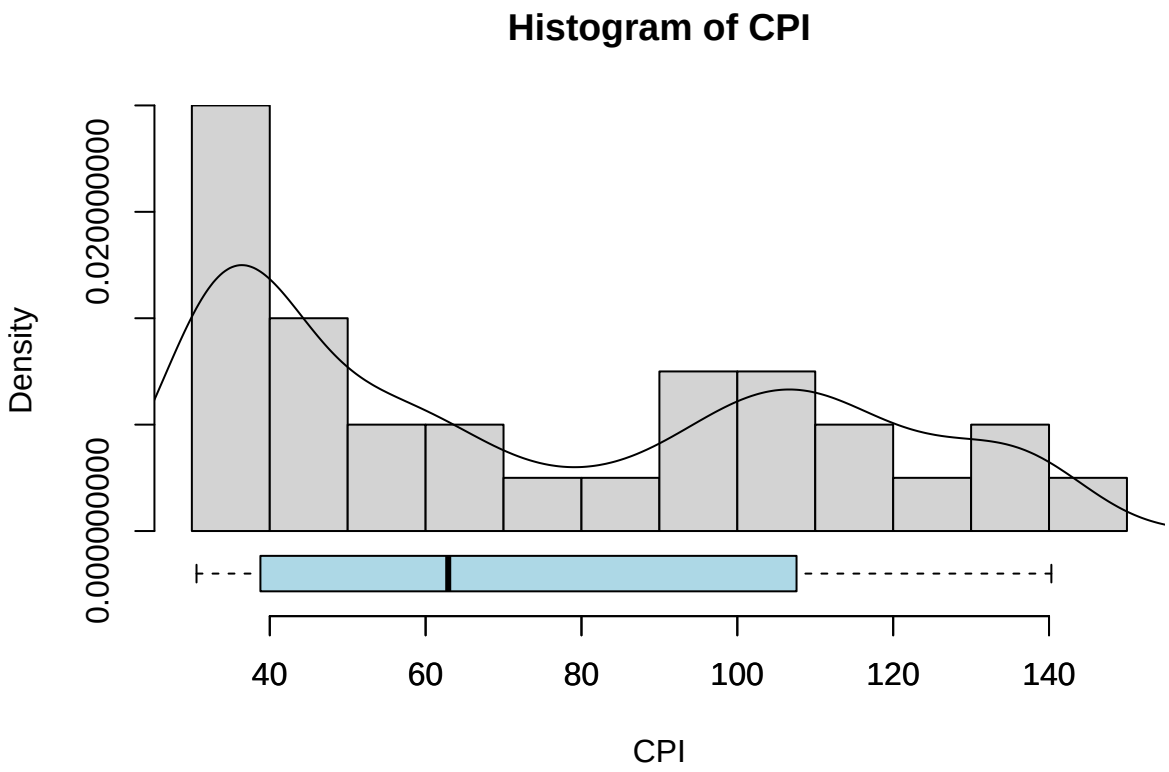
```
hist_boxplot(log(Cigar$price), freq = FALSE,
             main = "Histogram of Log Cigar Prices",
             xlab = "Log Price")
lines(density(log(Cigar$price)))
```



Examining the log price of the cigars, we see that the skewness is a bit fixed, but it is now multimodal, however this is better than the heavily skewed histogram, so we shall continue using log price as our predictor.

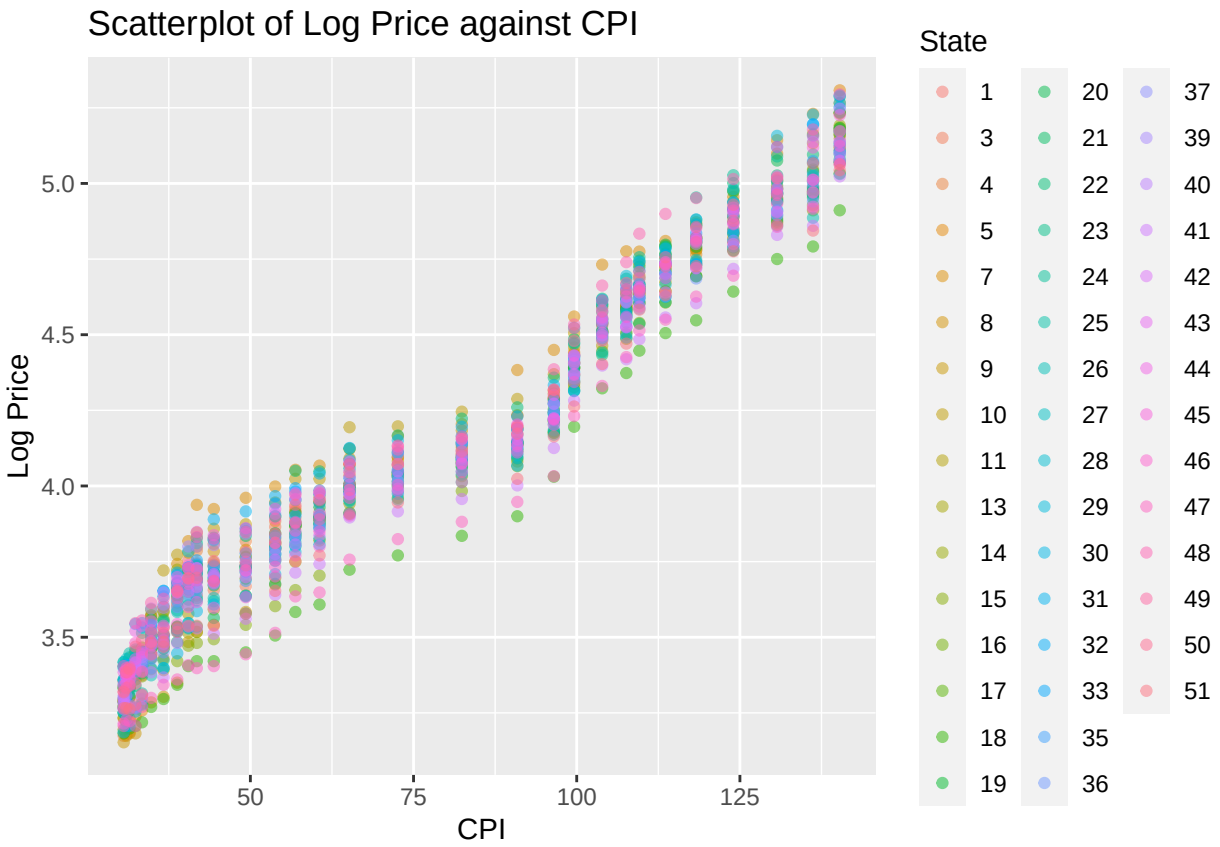
```
Cigar$lprice <- log(Cigar$price)
```

```
hist_boxplot(Cigar$cpi, freq = FALSE, main = "Histogram of CPI",
             xlab = "CPI")
lines(density(Cigar$cpi))
```



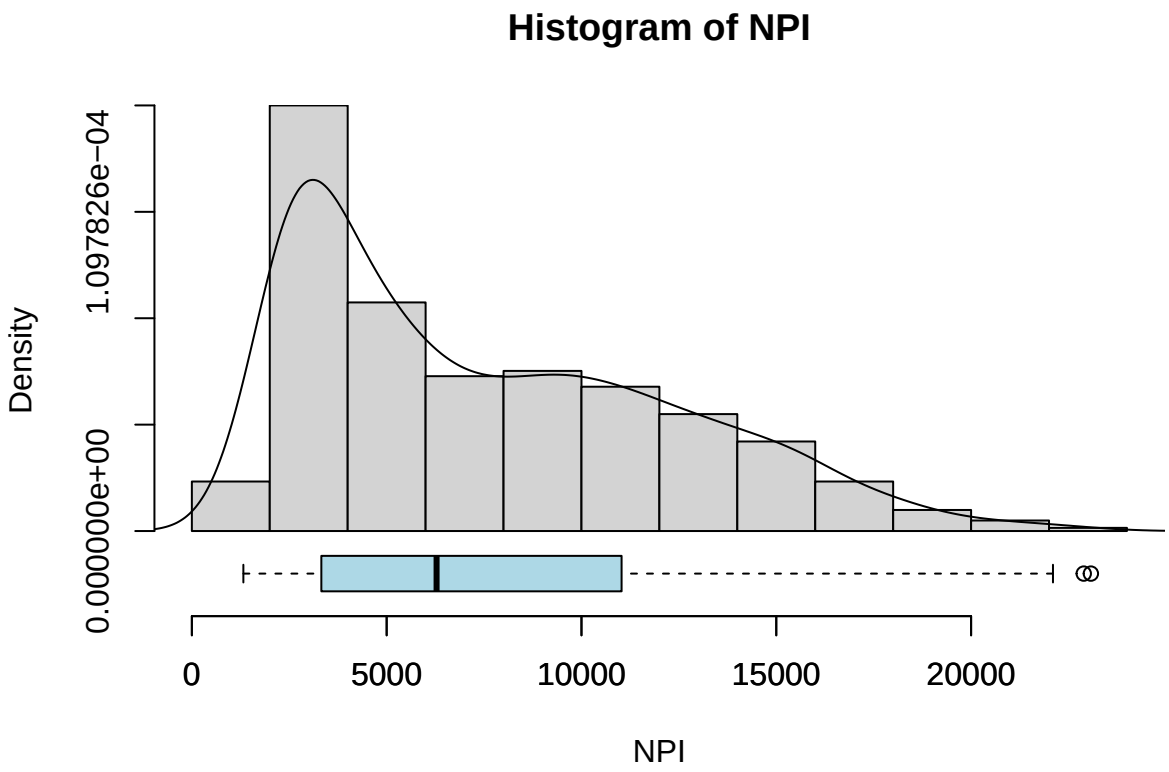
Examining the CPI variable, we see that this is right-skewed so the median is larger than the mean and the boxplot also shows that there are no outliers.

```
Cigar %>%
ggplot(aes(x = cpi, y = lprice, colour = factor(state))) +
  geom_point(alpha = 0.5) +
  xlab("CPI") +
  ylab("Log Price") +
  guides(color = guide_legend(title = "State")) +
  ggtitle("Scatterplot of Log Price against CPI")
```



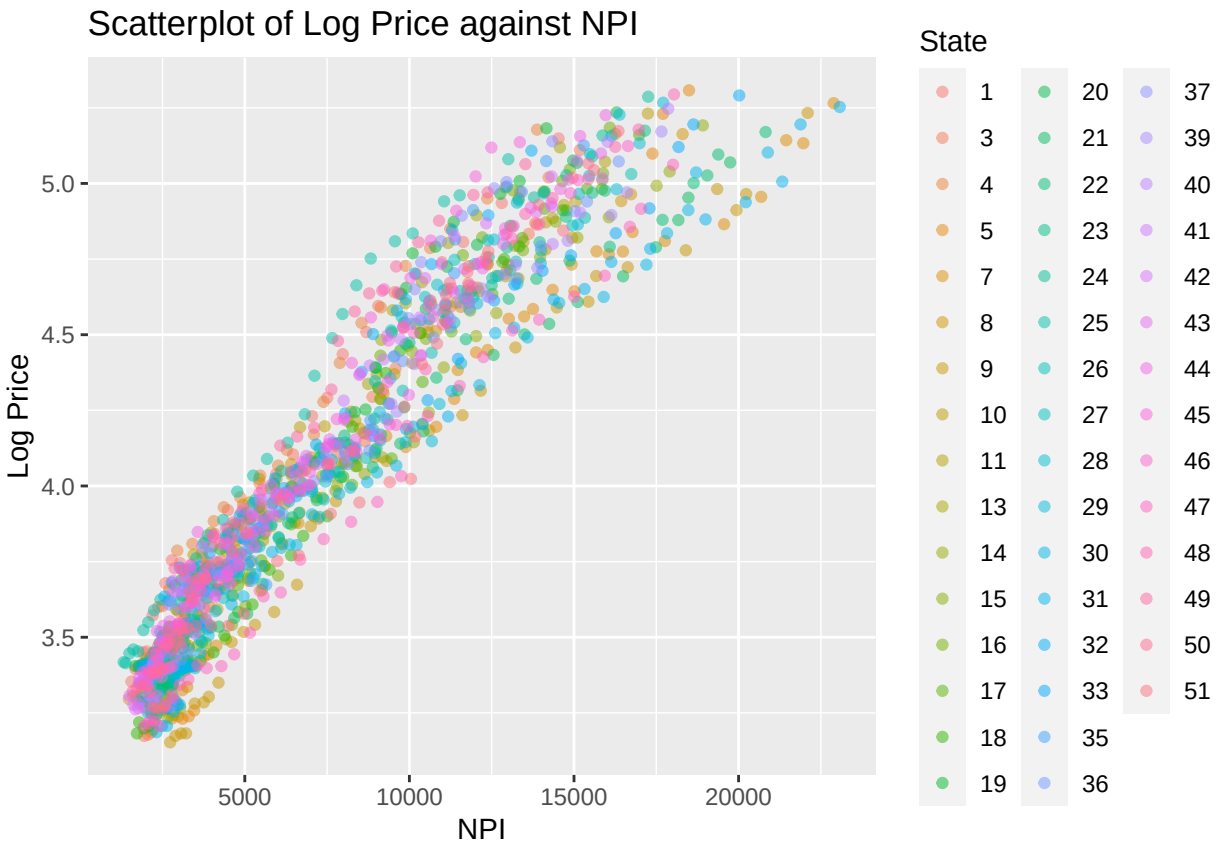
We created a scatterplot to visualize the log of the price per pack of cigarettes against the consumer price index (CPI) by state. Looking at the scatter plot, all the states display a similar relationship between the log of the price per pack of cigarettes and CPI. More specifically, the scatterplot shows a strong, positive, linear association between the log of the price per pack of cigarettes and CPI for all 50 states and Washington DC. This suggests that the higher the CPI, the higher the log of the price per pack of cigarettes are for all 50 states and Washington DC.

```
hist_boxplot(Cigar$ndi, freq = FALSE, main = "Histogram of NPI", xlab = "NPI")
lines(density(Cigar$ndi))
```



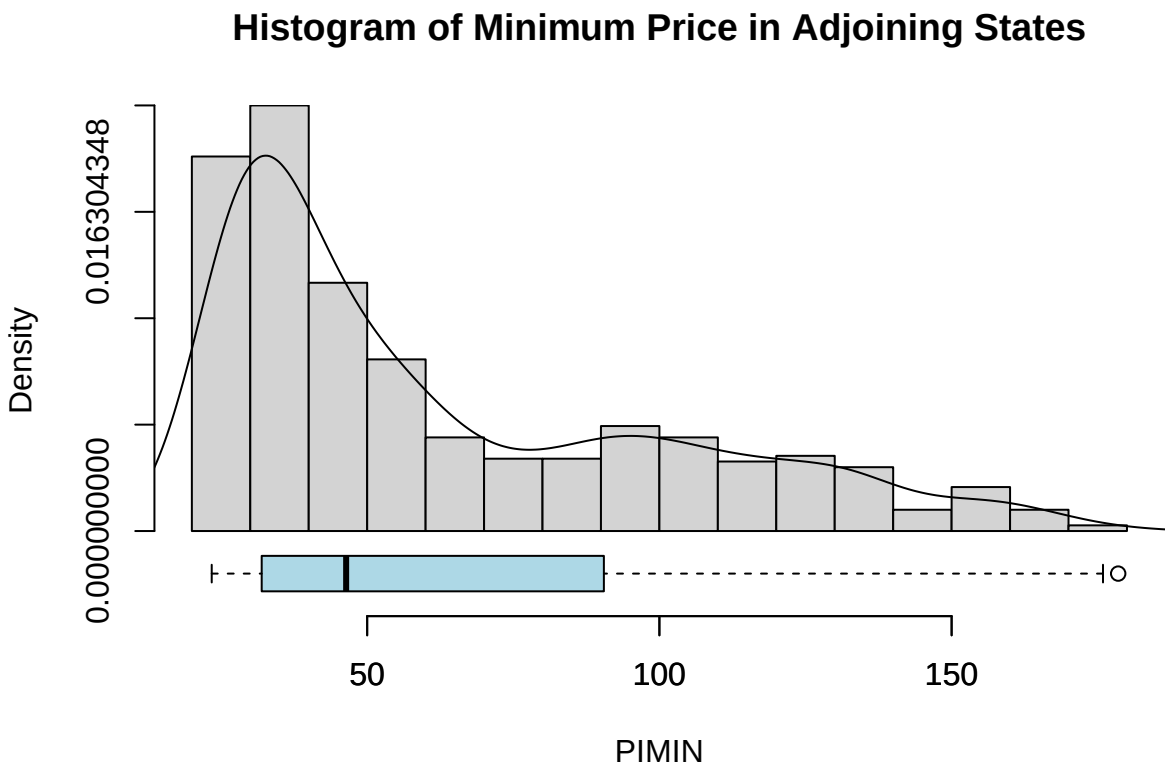
The NPI variable, we see that it is right-skewed and contains a bit of outliers. Because it is right-skewed, the median is also higher than the mean. Also, since it is skewed, using a log would be better in order to make it more normally distributed, but since our predictor is already logged, we do not need to make it more normally distributed.

```
Cigar %>%
ggplot(aes(x = ndi, y = lprice, colour = factor(state))) +
  geom_point(alpha = 0.5) +
  xlab("NPI") +
  ylab("Log Price") +
  guides(color = guide_legend(title = "State")) +
  ggtitle("Scatterplot of Log Price against NPI")
```



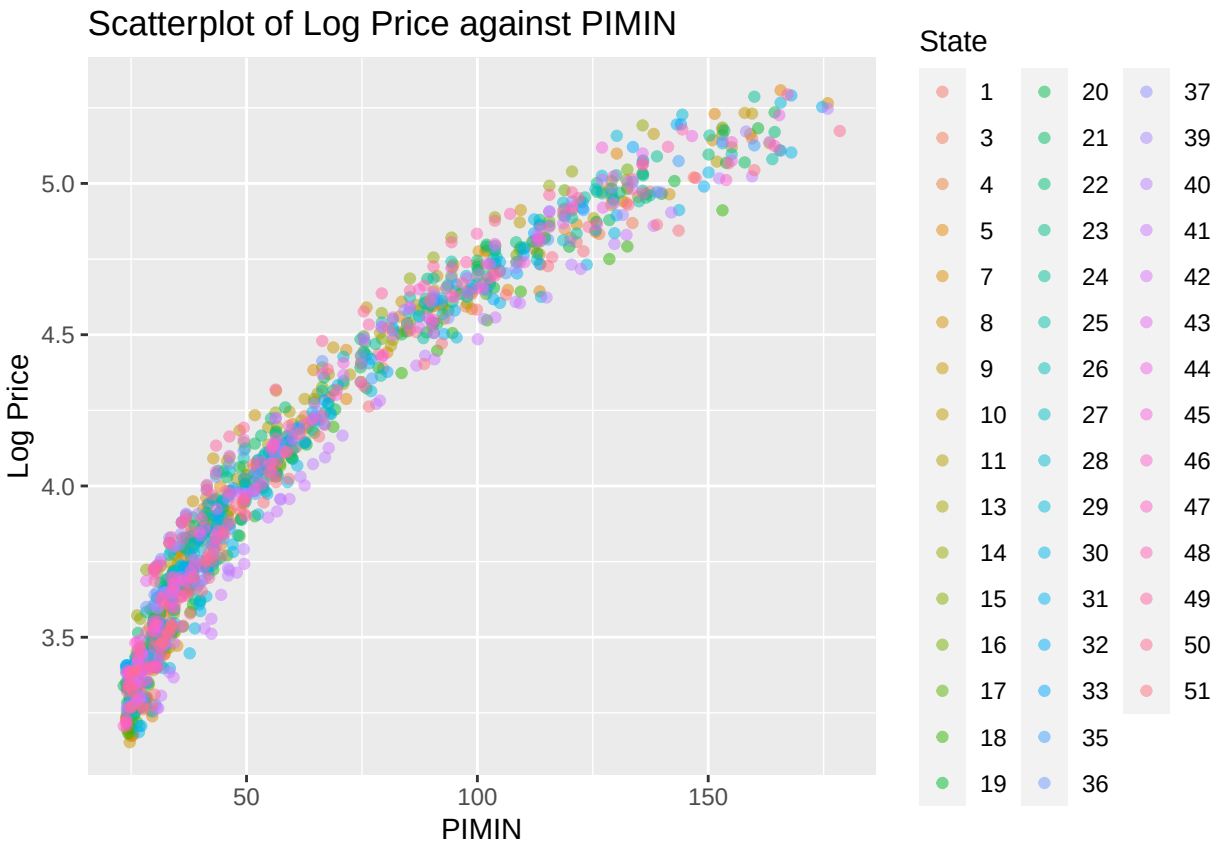
There seems to be a strong positive correlation between the log price and NPI, which means that the higher the NPI, the higher the price of the cigars are.

```
hist_boxplot(Cigar$pimin, freq = FALSE, main = "Histogram of Minimum Price in Adjoining States",
             xlab = "PIMIN")
lines(density(Cigar$pimin))
```



Examining the histogram of PIMIN, we also see that it is right-skewed, meaning that the median is higher than the mean and there only appears to be one outlier in the data. Since our predictor variable is already logged, we do not need to attempt to make this more normally distributed.

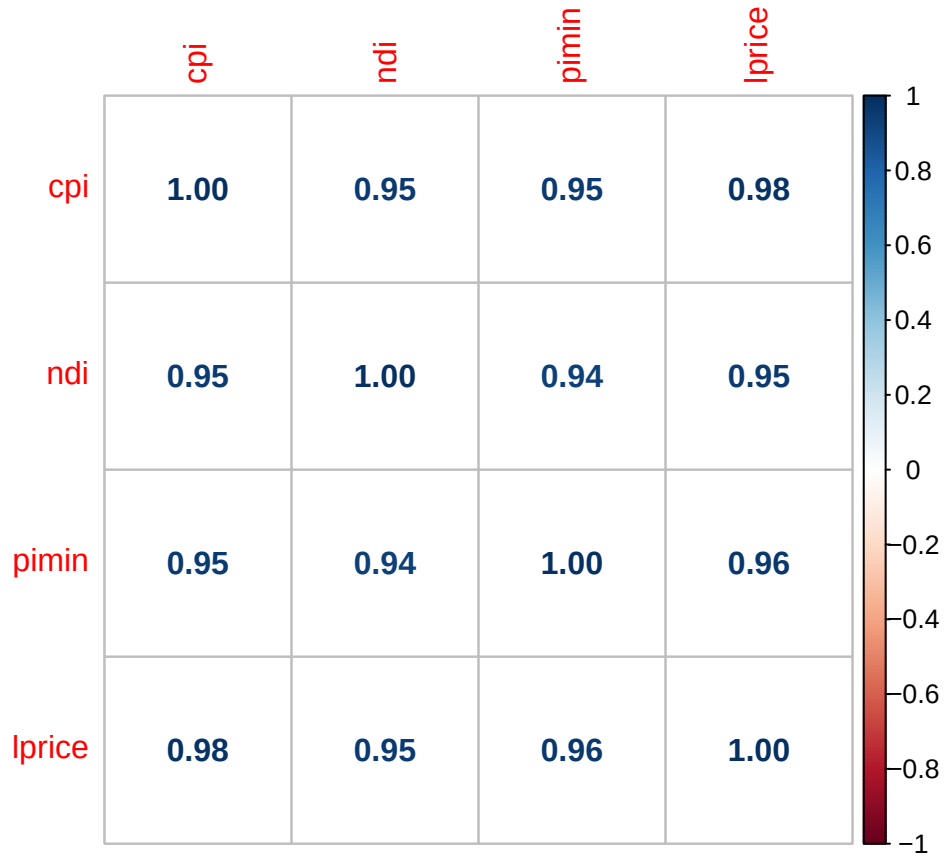
```
Cigar %>%
ggplot(aes(x = pimin, y = lprice, colour = factor(state))) +
  geom_point(alpha = 0.5) +
  xlab("PIMIN") +
  ylab("Log Price") +
  guides(color = guide_legend(title = "State")) +
  ggtitle("Scatterplot of Log Price against PIMIN")
```



We created a scatterplot to visualize the log of the price per pack of cigarettes against the minimum price in adjoining states per pack of cigarettes (PIMIN) by state. Looking at the scatter plot, all the states display a similar relationship between the log of the price per pack of cigarettes and PIMIN. More specifically, the scatterplot shows a strong, positive, nonlinear association between the log of the price per pack of cigarettes and PIMIN for all the states.

2.4: Correlation Plot

```
corr_cig <- Cigar[, (names(Cigar) %in%
                       c('lprice', 'cpi', 'ndi', 'pimin'))]
res <- cor(corr_cig)

corrplot(res, method = 'number')
```



Based on the correlation heat map, all of the variables are extremely strongly positively correlated, however, variables that have the strongest correlation between them are log price and CPI with a 0.98 correlation coefficient. PIMIN and log Price have the second strongest correlation with a 0.96 correlation coefficient. NDI and log Price have the third strongest correlation with a correlation coefficient of 0.95.

2.5: Summary Statistics

```
summary(Cigar$lprice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.153   3.549   3.957   4.060   4.586   5.308
```

The five number summary for the log of price per pack of cigarettes reveals that the log price per pack of cigarette has a minimum of `$3.153`, a 1st quartile value of `$3.549`, a median of `$3.957`, a mean of `$4.060`, a 3rd quartile value of `$4.586`, and a maximum of `$5.308`. Since the **median is slightly lower than the mean, our data for log price would be slightly skewed to the right**.

```
summary(Cigar$cpi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    30.6    38.8    62.9    73.6   107.6   140.3
```

The five number summary for the consumer price index (CPI) reveals that the CPI has a minimum of 30.6, a 1st quartile value of 38.8, a median of 62.9, a mean of 73.6, a 3rd quartile value of 107.6, and a maximum of 140.3. Since the **median is slightly lower than the mean, our data for CPI would be slightly skewed to the right**.

```
summary(Cigar$ndi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1323    3328    6281    7525   11024   23074
```

The five number summary for the per capita disposable income (NDI) reveals that the NDI has a minimum of `$1,323`, a 1st quartile value of `$3,328`, a median of `$6,281`, a mean of `$7,525`, a 3rd quartile value of `$11,024`, and a maximum of `$23,074`. Since the **median is lower than the mean, our data for NDI would be skewed to the right**.

```
summary(Cigar$pimin)
```

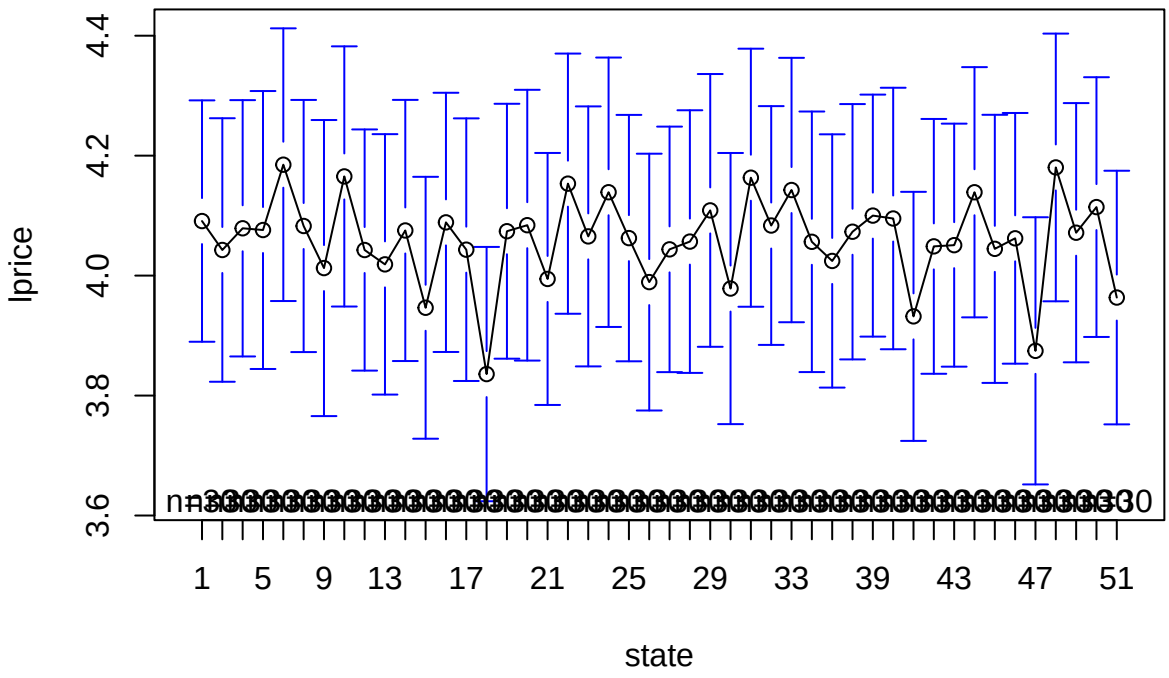
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.40   31.98   46.40   62.90   90.50  178.50
```

The five number summary for the minimum price in adjoining states per pack of cigarettes (PIMIN) reveals that the PIMIN has a minimum of `$23.40`, a 1st quartile value of `$31.98`, a median of `$46.40`, a mean of `$62.90`, a 3rd quartile value of `$90.50`, and a maximum of `$178.50`. Since the **median is lower than the mean, our data for PIMIN would be skewed to the right**.

3

3.1:

```
plotmeans(lprice ~ state, data = Cigar)
```



We plotted the means of the logged price separated by state and we can see that the log price of cigars in each state does differ for most states, so there does seem to exists heterogeneity in the model.

```
Cigar_pd <- pdata.frame(Cigar, index = c("state", "year"))

cigar_pool <-
  plm(lprice ~ cpi + ndi + pimin,
    data = Cigar_pd, model = "pooling")
summary(cigar_pool)
```

```
## Pooling Model
##
## Call:
## plm(formula = lprice ~ cpi + ndi + pimin, data = Cigar_pd, model = "pooling")
##
## Balanced Panel: n = 46, T = 30, N = 1380
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.2864004 -0.0643610  0.0020411  0.0694512  0.3063612
##
## Coefficients:
##               Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) 2.9927e+00 6.8697e-03 435.644 < 2.2e-16 ***
## cpi         9.7076e-03 2.9169e-04  33.281 < 2.2e-16 ***
## ndi         1.5840e-05 1.9635e-06   8.067 1.555e-15 ***
```

```
## pimin       3.7125e-03 2.4191e-04  15.347 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    452.97
## Residual Sum of Squares: 13.483
## R-Squared:      0.97023
## Adj. R-Squared: 0.97017
## F-statistic: 14950.2 on 3 and 1376 DF, p-value: < 2.22e-16
```

Our pooled regression mode is: $ln(PRICE) = 2.9927 + 0.0097076cpi + 1.584 \times 10^{-5}ndi + 0.0037125pimin$

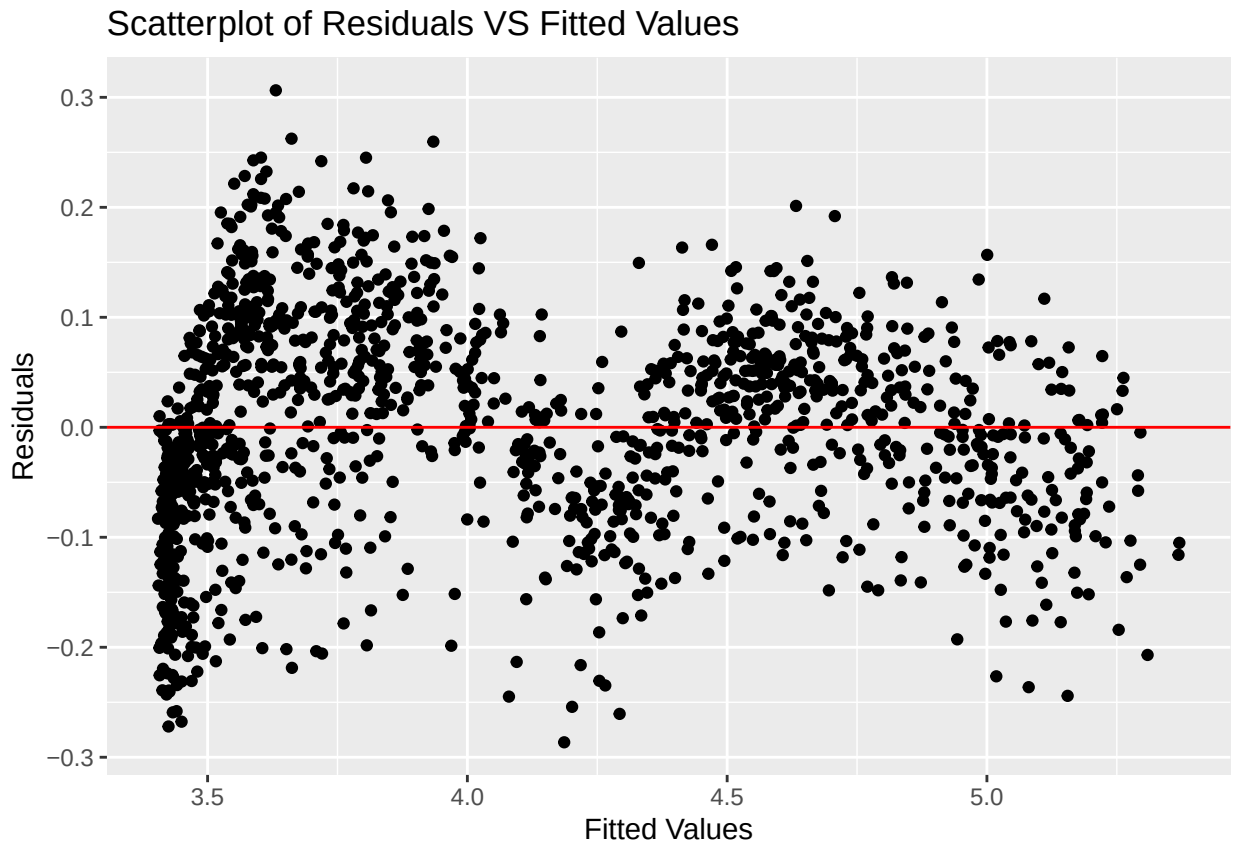
In our pooled regression model, the intercept along with all the estimators are statistically significant. The intercept is approximately 2.99287, which suggests that the average change in price per pack of cigarettes is approximately 2.99% when CPI, NDI, and PIMIN equal to 0. The coefficient for CPI is 0.0097076, which indicates that a unit increase in the overall change in consumer prices leads to an approximately 0.971% increase in the price per pack of cigarettes, all else equal. The coefficient for NDI is 0.00001584, which indicates that a dollar increase in per capita disposable income leads to an approximately 0.00158% increase in price per pack of cigarettes, all else equal. Lastly, The coefficient for PIMIN is 0.0037125, which indicates that a dollar increase in the minimum price in adjoining states per pack of cigarettes leads to an approximately 0.371% increase in price per pack of cigarettes. The R-squared value is 0.97023, which suggests that 97.023% of the variation in the change in price per pack of cigarettes is explained by our pooled model.

```
bptest(cigar_pool)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  cigar_pool
## BP = 65.073, df = 3, p-value = 4.84e-14
```

Base on the p value of our bptest, our pooled model does show signs of heterogeneity as we reject the null that the variance of our errors is non-dependent on all of the explanatory variables.

```
Cigar_pd$Pool_fitt <- fitted(cigar_pool)
Cigar_pd$Pool_resid <- resid(cigar_pool)

Cigar_pd %>%
  ggplot(aes(y = Pool_resid, x = Pool_fitt)) +
  geom_point() +
  xlab("Fitted Values") +
  ylab("Residuals") +
  ggtitle("Scatterplot of Residuals VS Fitted Values") +
  geom_hline(yintercept = 0, col = "red")
```



Examining the residuals vs fitted values plot, we do see a slight trend and clustering at certain points in which as the values get larger, the residuals go more downward, indicating the presence of hetereoskedascity.

```
coeftest(cigar_pool, vcov = vcovHC(cigar_pool, type = "HC0",
         cluster = "group", adjust = T))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 2.9927e+00 1.4163e-02 211.3038 < 2.2e-16 ***
## cpi         9.7076e-03 6.2392e-04  15.5589 < 2.2e-16 ***
## ndi         1.5840e-05 4.6237e-06   3.4258  0.000631 ***
## pimin       3.7125e-03 4.2536e-04   8.7280 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Correcting for the hetereoskedascity in our pooled mode, we see that the standard errors for the intercept become smaller, but the standard errors for our coefficients become larger. However, our coefficients are still significant, so correcting for hetereoskedascity did not invalidate any of the variables chosen.

```
cigar_fem <-
 plm(lprice ~ cpi + ndi + pimin,
    data = Cigar_pd, model = "within")

summary(cigar_fem)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lprice ~ cpi + ndi + pimin, data = Cigar_pd, model = "within")
##
## Balanced Panel: n = 46, T = 30, N = 1380
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.237822 -0.047814  0.002606  0.056215  0.258103
##
## Coefficients:
##         Estimate Std. Error t-value  Pr(>|t|)
## cpi   1.0304e-02 3.0329e-04 33.9733 < 2.2e-16 ***
## ndi   1.8180e-05 2.6391e-06  6.8889  8.65e-12 ***
## pimin 2.8259e-03 2.2451e-04 12.5873 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    445.8
## Residual Sum of Squares: 8.2091
## R-Squared:      0.98159
## Adj. R-Squared: 0.98092
## F-statistic: 23650.1 on 3 and 1331 DF, p-value: < 2.22e-16
```

Our fixed effects model is: $ln(PRICE) = Intercept_i + 0.010304cpi + 1.8180 \times 10^{-5}ndi + 0.0028259pimin$ where i represents state.

```
fixef(cigar_fem)
```

```
##      1      3      4      5      7      8      9     10     11     13     14
## 3.0376 2.9697 3.0332 2.9642 3.0414 3.0035 2.9108 3.0801 2.9900 2.9646 3.0046
##     15     16     17     18     19     20     21     22     23     24     25
## 2.9019 3.0179 2.9633 2.7992 3.0108 3.0165 2.9152 3.0489 2.9919 3.0525 3.0209
##     26     27     28     29     30     31     32     33     35     36     37
## 2.9401 2.9866 2.9897 3.0111 2.8692 3.0276 3.0252 3.0355 2.9869 2.9744 3.0143
##     39     40     41     42     43     44     45     46     47     48     49
## 3.0215 2.9851 2.8760 2.9944 3.0172 3.0571 2.9977 2.9894 2.8175 3.0847 3.0504
##     50     51
## 3.0247 2.8870
```

These are the values for the intercept where the header is i and the value below it is the value of the intercept at that i (state).

In our Fixed Effects regression model, all the estimators are statistically significant. The coefficient for CPI is 0.010304, which indicates that a unit increase in the overall change in consumer prices leads to an approximately 1.03% increase in the price per pack of cigarettes, all else equal. The coefficient for NDI is 0.00001818, which indicates that a dollar increase in per capita disposable income leads to an approximately 0.00181% increase in price per pack of cigarettes, all else equal. Lastly, the coefficient for PIMIN is 0.0028259, which indicates that a dollar increase in the minimum price in adjoining states per pack of cigarettes leads to an approximately 0.283% increase in price per pack of cigarettes. The R-squared value is 0.98159, which suggests that 98.159% of the variation in the change in price per pack of cigarettes is explained by our Fixed Effects model.

```r
# FEM vs pool
pFtest(cigar_fem, cigar_pool)
```

```
##
##  F test for individual effects
##
## data:  lprice ~ cpi + ndi + pimin
## F = 19.003, df1 = 45, df2 = 1331, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Performing a f-test across individuals between the pooled and fixed effects model, our p-value is well below the usual acceptable limits of significance, so we would reject the null that the coefficients are the same, i.e. fixed effects are preferred over the pooled model.

```r
cigar_rem <-
  plm(lprice ~ cpi + ndi + pimin,
    data = Cigar_pd, model = "random")

summary(cigar_rem)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = lprice ~ cpi + ndi + pimin, data = Cigar_pd, model = "random")
##
## Balanced Panel: n = 46, T = 30, N = 1380
##
## Effects:
##                     var  std.dev share
## idiosyncratic 0.006168 0.078534 0.644
## individual    0.003404 0.058346 0.356
## theta: 0.7614
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -0.22496350 -0.05246825  0.00005921  0.05681738  0.27050536
##
## Coefficients:
##               Estimate Std. Error  z-value  Pr(>|z|)
## (Intercept) 2.9872e+00 1.0491e-02 284.7445 < 2.2e-16 ***
## cpi         1.0280e-02 2.9517e-04  34.8261 < 2.2e-16 ***
## ndi         1.7834e-05 2.5132e-06   7.0958 1.286e-12 ***
## pimin       2.8930e-03 2.2154e-04  13.0586 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    446.21
## Residual Sum of Squares: 8.5164
## R-Squared:      0.98091
## Adj. R-Squared: 0.98087
## Chisq: 70718.8 on 3 DF, p-value: < 2.22e-16
```

Our random effects model is: $ln(PRICE) = 2.9872 + 0.010280cpi + 1.7834 \times 10^{-5}ndi + 0.0028930pimin$

The variance of the random error in our model is 0.003404 for the individual error, 0.006168 for the idiosyncratic error, and the total variance is 0.009572.

In our Random Effects regression model, the intercept along with all the estimators are statistically significant. The intercept is approximately 2.9872, which suggests that the average change in price per pack of cigarettes is approximately 2.99% when CPI, NDI, and PIMIN equal to 0. The coefficient for CPI is 0.01028, which indicates that a unit increase in the overall change in consumer prices leads to an approximately 1.03% increase in the price per pack of cigarettes, all else equal. The coefficient for NDI is 0.000017834, which indicates that a dollar increase in per capita disposable income leads to an approximately 0.0018% increase in price per pack of cigarettes, all else equal. Lastly, The coefficient for PIMIN is 0.002893, which indicates that a dollar increase in the minimum price in adjoining states per pack of cigarettes leads to an approximately 0.29% increase in price per pack of cigarettes. The R-squared value is 0.98091, which suggests that 98.09% of the variation in the change in price per pack of cigarettes is explained by our Random Effects model.

```
# REM vs Pool
plmtest(cigar_rem, type = c("bp"))
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan)
##
## data:  lprice ~ cpi + ndi + pimin
## chisq = 2625.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

In our BP test of the pooled model against the random effects model, the p-value is well below the usual acceptable limits of significance, so we would reject the null hypothesis that the variance of the error term is equal to 0. In other words, there is statistical evidence that there exists individual heterogeneity, which validates the use of the Random Effects model.

```
# FEM vs REM
phtest(cigar_fem, cigar_rem)
```

```
##
##  Hausman Test
##
## data:  lprice ~ cpi + ndi + pimin
## chisq = 9.9491, df = 3, p-value = 0.019
## alternative hypothesis: one model is inconsistent
```

In our Hausman test of the fixed effects model vs the random effects model, our p-value of 0.019 is below the 5% significance level, so we would reject the null that the Random Effects model is the preferred model. This implies that there is endogeneity present (i.e. the error term is correlated with the explanatory variables), therefore we should use the **Fixed Effects model**.

$ln(PRICE) = Intercept_i + 0.010304cpi + 1.8180 \times 10^{-5}ndi + 0.0028259pimin$

II: Binary Models

For the qualitative dependent variable model, we used "blackpoliticians" dataset in order to answer the economic/finance/business question with respect to our regression model: What variables causes legislators

to respond to emails? In other words, what characteristics, whether it be the characteristics of the legislator itself or the characteristic of the person sending the email, would influence the legislator responding to an email. Our dataset was collected in 2013 from Broockman on a field experiment where the author sent fictional emails purportedly sent by Black people to legislators in the US. It is a data frame with 5,593 rows and 14 variables. In our model, our dependent variable is "responded", which means that the legislator responded to email, and it is an indicator variable. Out of 13 independent variables, there are 6 indicator variables: leg_black (legislator receiving email is Black), treat_out (email is from out-of-district), nonblacknonwhite (legislator receiving email is neither Black nor White), leg_senator (legislator receiving email is a senator), leg_democrat (legislator receiving email is in the Democratic party), and south (legislator receiving email is in the Southern United States). The remaining 7 independent variables are continuous variables: totalpop (district population), medianhhincom (district median household income), black_medianhh (district median household income among Black people), white_medianhh (District median household income among White people), blackpercent (percentage of district that is Black), and urbanpercent (percentage of district that is urban). We disregarded the variable "statessquireindex", which is state's squire index, because there is a lack of information about this variable from google search and the dataset description itself.

```
data("black_politicians")

head(black_politicians)
```

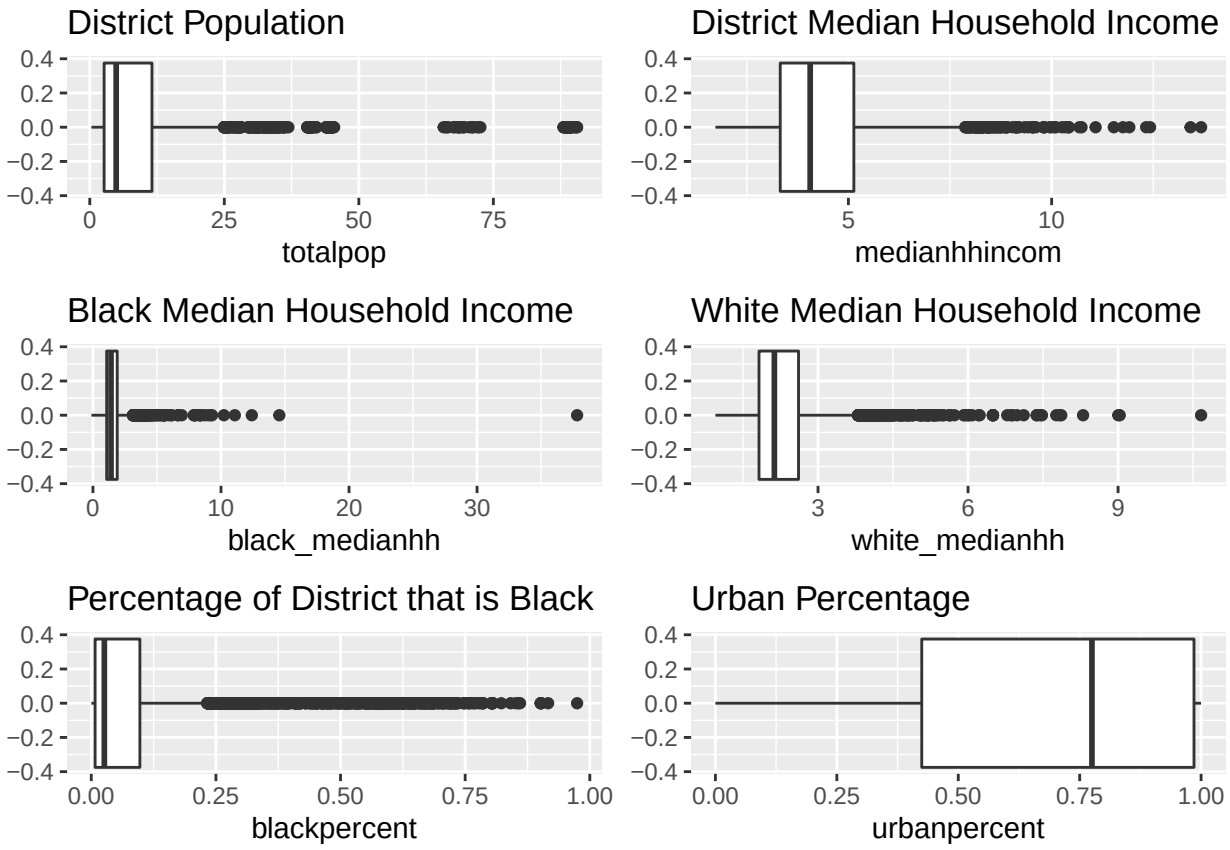
```
## # A tibble: 6 x 14
##   leg_black treat_out responded totalpop media~1 black~2 white~3 black~4 state~5
##       <dbl>     <dbl>     <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1         0         0         0     1.59    5.06    2.68    2.66 0.00712   0.227
## 2         0         0         1     1.62    4.97    2.71    2.66 0.00580   0.227
## 3         0         0         1     1.67    6.96    2.31    3.00 0.0120    0.227
## 4         0         0         1     1.61    4.18    2.47    2.49 0.00428   0.227
## 5         0         1         1     1.56    3.12    2.15    2.06 0.00826   0.227
## 6         0         1         0     1.62    6.07    1.99    2.70 0.0118    0.227
## # ... with 5 more variables: nonblacknonwhite <dbl>, urbanpercent <dbl>,
## #   leg_senator <dbl>, leg_democrat <dbl>, south <dbl>, and abbreviated
## #   variable names 1: medianhhincom, 2: black_medianhh, 3: white_medianhh,
## #   4: blackpercent, 5: statessquireindex
```

2

2.1

```
bp_totalpop <- black_politicians %>%
  ggplot(mapping = aes(x = totalpop)) +
  geom_boxplot() +
  ggtitle("District Population") +
  labs(x = "totalpop")

bp_median <- black_politicians %>%
  ggplot(mapping = aes(x = medianhhincom)) +
  geom_boxplot() +
  ggtitle("District Median Household Income") +
  labs(x = "medianhhincom")
```

```
bp_bmedian <- black_politicians %>%
  ggplot(mapping = aes(x = black_medianhh)) +
  geom_boxplot() +
  ggtitle("Black Median Household Income") +
  labs(x = "black_medianhh")

bp_wmedian <- black_politicians %>%
  ggplot(mapping = aes(x = white_medianhh)) +
  geom_boxplot() +
  ggtitle("White Median Household Income") +
  labs(x = "white_medianhh")

bp_blackp <- black_politicians %>%
  ggplot(mapping = aes(x = blackpercent)) +
  geom_boxplot() +
  ggtitle("Percentage of District that is Black") +
  labs(x = "blackpercent")

bp_urbanp <- black_politicians %>%
  ggplot(mapping = aes(x = urbanpercent)) +
  geom_boxplot() +
  ggtitle("Urban Percentage") +
  labs(x = "urbanpercent")

plot_grid(bp_totalpop, bp_median, bp_bmedian, bp_wmedian, bp_blackp, bp_urbanp, ncol = 2, nrow = 3)
```



The boxplot of district population, district median house income, black median house income, white median house income, percentage of district that is black are all extremely skewed to the right with multiple outliers, whereas, the percentage of the district that is urban is skewed to the left. We also see that there exists many outliers and we would have to remove them in order to solve the issue of skewness.

```
outliers <-
c(which(black_politicians$totalpop %in%
          boxplot.stats(black_politicians$totalpop)$out),
  which(black_politicians$medianhhincom %in%
          boxplot.stats(black_politicians$medianhhincom)$out),
  which(black_politicians$black_medianhh %in%
          boxplot.stats(black_politicians$black_medianhh)$out),
  which(black_politicians$white_medianhh %in%
          boxplot.stats(black_politicians$white_medianhh)),
  which(black_politicians$blackpercent %in%
          boxplot.stats(black_politicians$blackpercent)))

black_politicians_c <- black_politicians[-outliers, ]

outliers1 <-
c(which(black_politicians_c$totalpop %in%
          boxplot.stats(black_politicians_c$totalpop)$out),
  which(black_politicians_c$medianhhincom %in%
          boxplot.stats(black_politicians_c$medianhhincom)$out),
  which(black_politicians_c$black_medianhh %in%
          boxplot.stats(black_politicians_c$black_medianhh)$out),
  which(black_politicians_c$white_medianhh %in%
          boxplot.stats(black_politicians_c$white_medianhh)),
  which(black_politicians_c$blackpercent %in%
          boxplot.stats(black_politicians_c$blackpercent)))

black_politicians_c1 <- black_politicians_c[-outliers1, ]

outliers2 <-
c(which(black_politicians_c1$totalpop %in%
          boxplot.stats(black_politicians_c1$totalpop)$out),
  which(black_politicians_c1$medianhhincom %in%
          boxplot.stats(black_politicians_c1$medianhhincom)$out),
  which(black_politicians_c1$black_medianhh %in%
          boxplot.stats(black_politicians_c1$black_medianhh)$out),
  which(black_politicians_c1$white_medianhh %in%
          boxplot.stats(black_politicians_c1$white_medianhh)),
  which(black_politicians_c1$blackpercent %in%
          boxplot.stats(black_politicians_c1$blackpercent)))

black_politicians_c2 <- black_politicians_c1[-outliers2, ]

outliers3 <-
c(which(black_politicians_c2$totalpop %in%
          boxplot.stats(black_politicians_c2$totalpop)$out),
  which(black_politicians_c2$medianhhincom %in%
          boxplot.stats(black_politicians_c2$medianhhincom)$out),
  which(black_politicians_c2$black_medianhh %in%
          boxplot.stats(black_politicians_c2$black_medianhh)$out),
```

```
  which(black_politicians_c2$white_medianhh %in%
          boxplot.stats(black_politicians_c2$white_medianhh)),
  which(black_politicians_c2$blackpercent %in%
          boxplot.stats(black_politicians_c2$blackpercent)))

black_politicians_c3 <- black_politicians_c2[-outliers3, ]

outliers4 <-
c(which(black_politicians_c3$totalpop %in%
          boxplot.stats(black_politicians_c3$totalpop)$out),
  which(black_politicians_c3$medianhhincom %in%
          boxplot.stats(black_politicians_c3$medianhhincom)$out),
  which(black_politicians_c3$black_medianhh %in%
          boxplot.stats(black_politicians_c3$black_medianhh)$out),
  which(black_politicians_c3$white_medianhh %in%
          boxplot.stats(black_politicians_c3$white_medianhh)),
  which(black_politicians_c3$blackpercent %in%
          boxplot.stats(black_politicians_c3$blackpercent)))

black_politicians_c4 <- black_politicians_c3[-outliers4, ]

outliers5 <-
c(which(black_politicians_c4$totalpop %in%
          boxplot.stats(black_politicians_c4$totalpop)$out),
  which(black_politicians_c4$medianhhincom %in%
          boxplot.stats(black_politicians_c4$medianhhincom)$out),
  which(black_politicians_c4$black_medianhh %in%
          boxplot.stats(black_politicians_c4$black_medianhh)$out),
  which(black_politicians_c4$white_medianhh %in%
          boxplot.stats(black_politicians_c4$white_medianhh)),
  which(black_politicians_c4$blackpercent %in%
          boxplot.stats(black_politicians_c4$blackpercent)))

black_politicians_c5 <- black_politicians_c4[-outliers5, ]
nrow(black_politicians_c5)
```

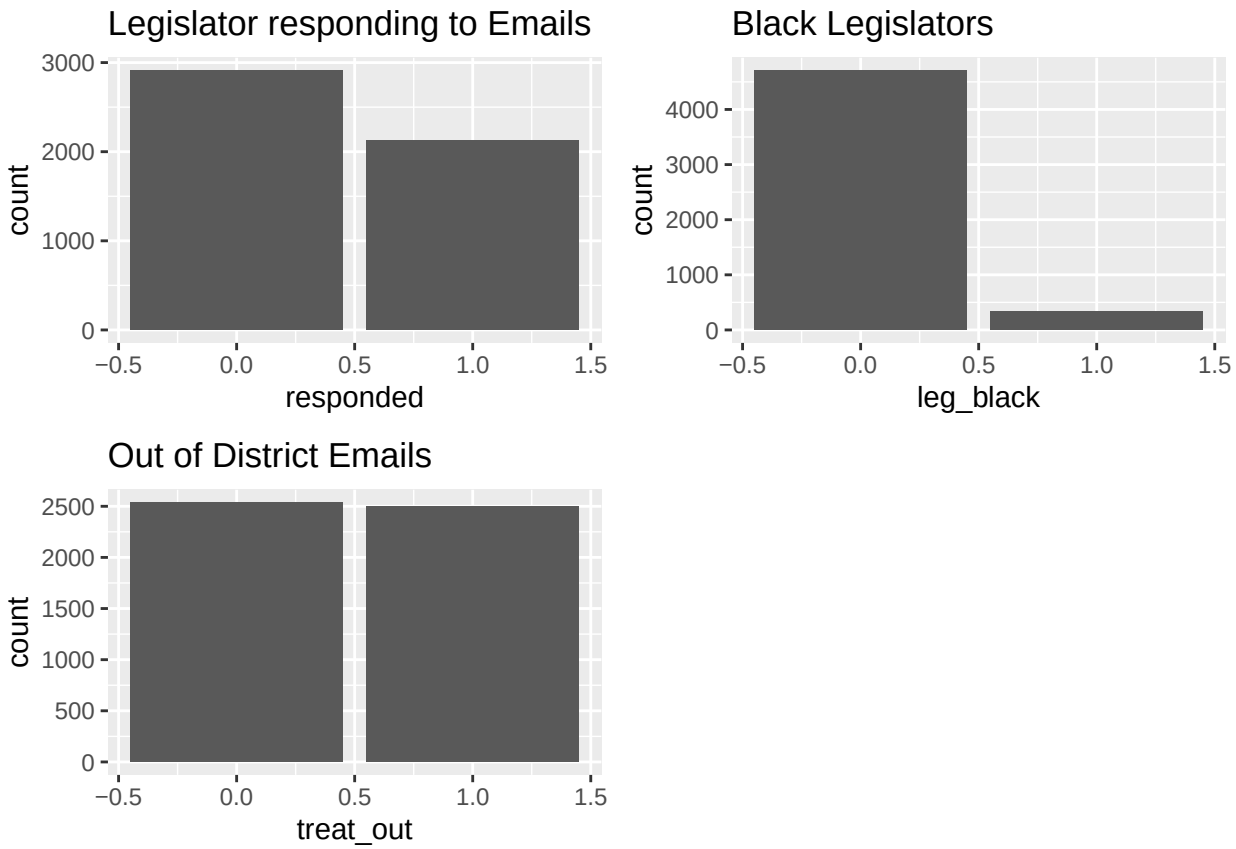
```
## [1] 4943
```

After a few rounds of removing outliers, we still have 4,943 observations to work with and we may continue.

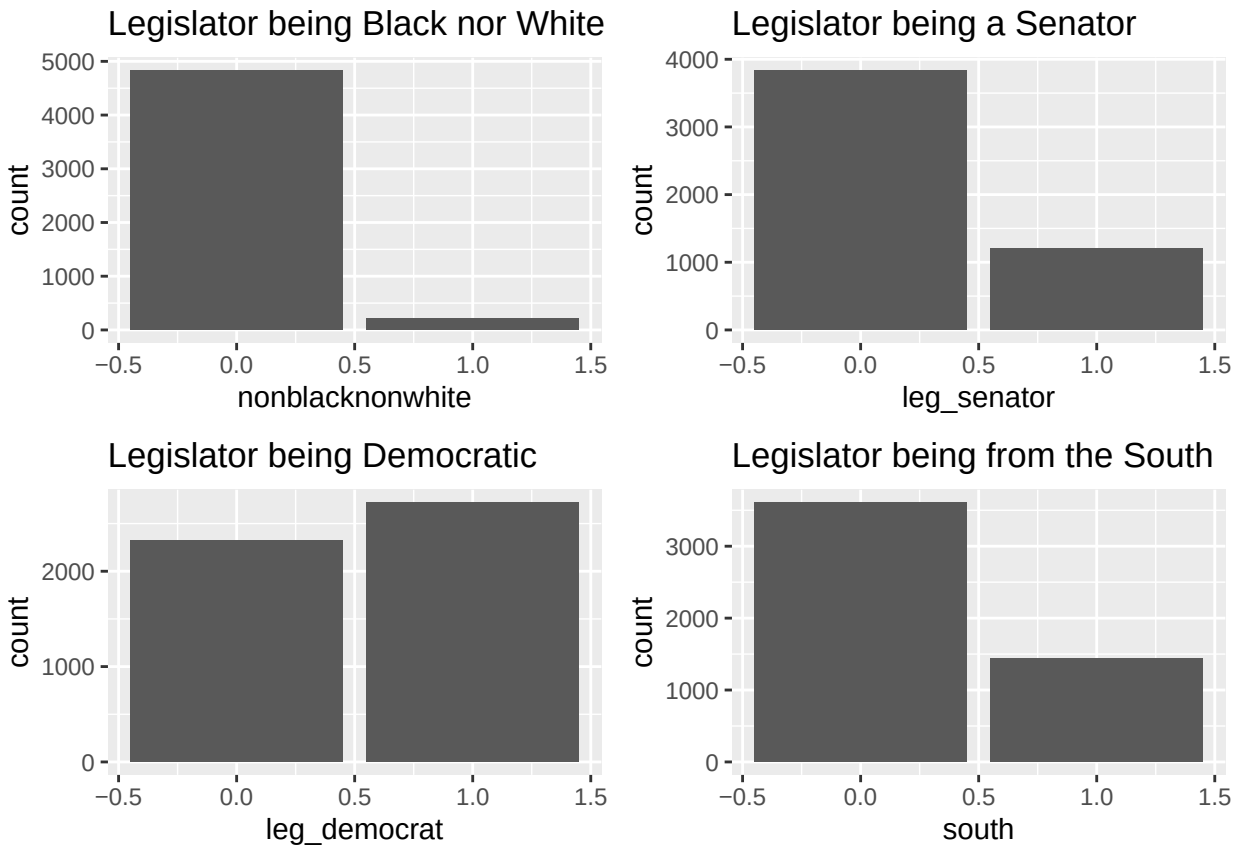
2.2

```
response <-
black_politicians_c %>%
  ggplot(mapping = aes(x = responded)) +
  geom_bar() +
  ggtitle("Legislator responding to Emails") +
  labs(x = "responded")

legblack <-
black_politicians_c %>%
```

```
  ggplot(mapping = aes(x = leg_black)) +
  geom_bar() +
  ggtitle("Black Legislators") +
  labs(x = "leg_black")

outod <-
black_politicians_c %>%
  ggplot(mapping = aes(x = treat_out)) +
  geom_bar() +
  ggtitle("Out of District Emails") +
  labs(x = "treat_out")

nbnw <-
black_politicians_c %>%
  ggplot(mapping = aes(x = nonblacknonwhite)) +
  geom_bar() +
  ggtitle("Legislator being Black nor White") +
  labs(x = "nonblacknonwhite")

legsen <-
black_politicians_c %>%
  ggplot(mapping = aes(x = leg_senator)) +
  geom_bar() +
  ggtitle("Legislator being a Senator") +
  labs(x = "leg_senator")

legdem <-
black_politicians_c %>%
  ggplot(mapping = aes(x = leg_democrat)) +
  geom_bar() +
  ggtitle("Legislator being Democratic") +
  labs(x = "leg_democrat")

south <-
black_politicians_c %>%
  ggplot(mapping = aes(x = south)) +
  geom_bar() +
  ggtitle("Legislator being from the South") +
  labs(x = "south")

plot_grid(response, legblack, outod, nrow = 2, ncol = 2)
```




```
plot_grid(nbnw, legsen, legdem, south, nrow = 2, ncol = 2)
```



Using bar charts to examine our binary variables, we see that the variables with roughly around the same number of observations are treat_out, the email being out of district and leg_democrat, the legislator being democratic. The other variables do show a large amount of different observations, including leg_black, nonblacknonwhite, leg_senator, and south.

2.3

```
library(Boruta)
bor_res <- Boruta(responded ~., data = black_politicians_c5[, -c(1, 2, 9, 10, 11, 12, 13, 14)], doTrace
```

```
## After 9 iterations, +44 secs:

##  confirmed 1 attribute: totalpop;

##  still have 4 attributes left.

## After 13 iterations, +1.1 mins:

##  confirmed 3 attributes: black_medianhh, blackpercent, medianhhincom;

##  still have 1 attribute left.

## After 28 iterations, +2.3 mins:
```

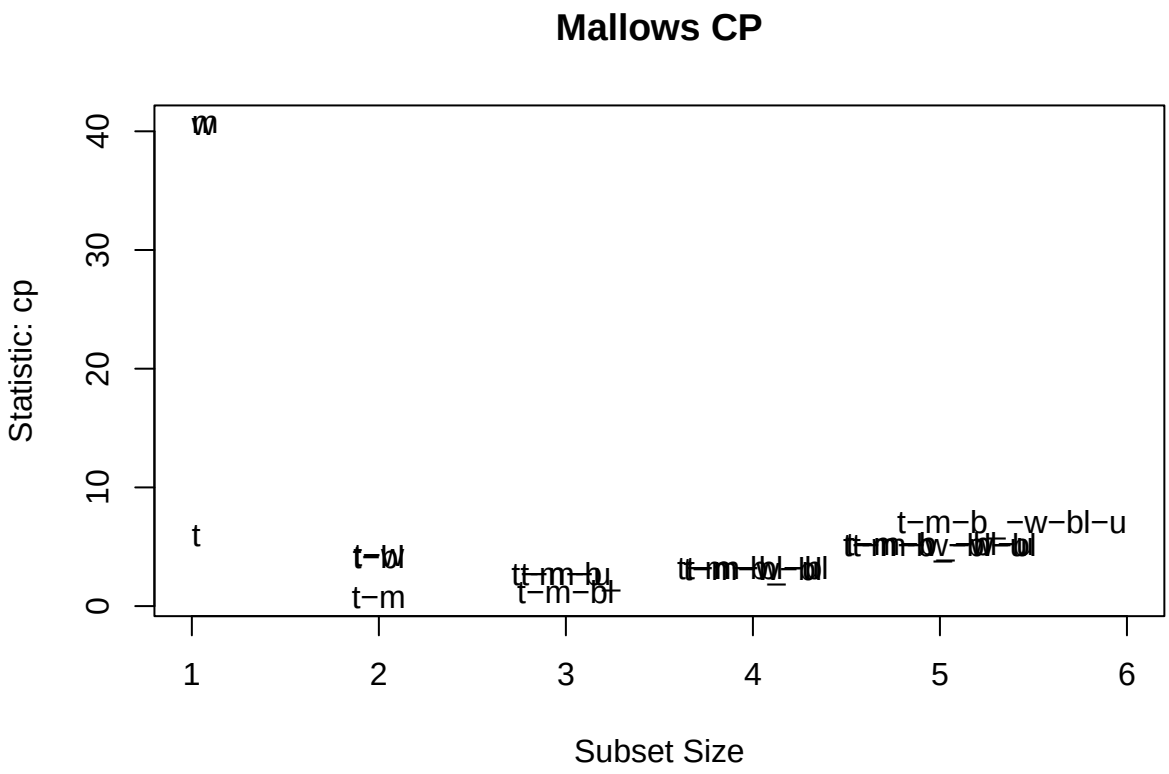
```
##  confirmed 1 attribute: white_medianhh;

##  no more attributes left.
```

```
attStats(bor_res)
```

```
##                  meanImp medianImp      minImp    maxImp  normHits  decision
## totalpop       17.050781 16.987650 13.77383624 20.348081 1.0000000 Confirmed
## medianhhincom   5.357964  5.174152  3.38377293  7.964129 0.9642857 Confirmed
## black_medianhh  3.945491  4.000386  0.57970630  7.405637 0.9642857 Confirmed
## white_medianhh  2.484451  2.645994 -0.03009006  4.518268 0.7857143 Confirmed
## blackpercent    6.044611  6.412760  2.54259632  7.787520 0.9642857 Confirmed
```

We used the Boruta algorithm to perform feature selection and found that the variables for total population, median household income, black median household income, and the percentage of district that is black were confirmed; these attributes were deemed as important. On the other hand, the Boruta algorithm outputted white median household income as a tentative attribute, meaning that this variable may or may not be the best input depending on the model and other attributes.

```
library(leaps)
ss <- regsubsets(responded ~ totalpop + medianhhincom + black_medianhh +
                 white_medianhh + blackpercent + urbanpercent,
                 method = c("exhaustive"), nbest = 3,
                 data = black_politicians_c5)

library(AER)
subsets(ss, statistic = "cp", legend = F, main = "Mallows CP")
```

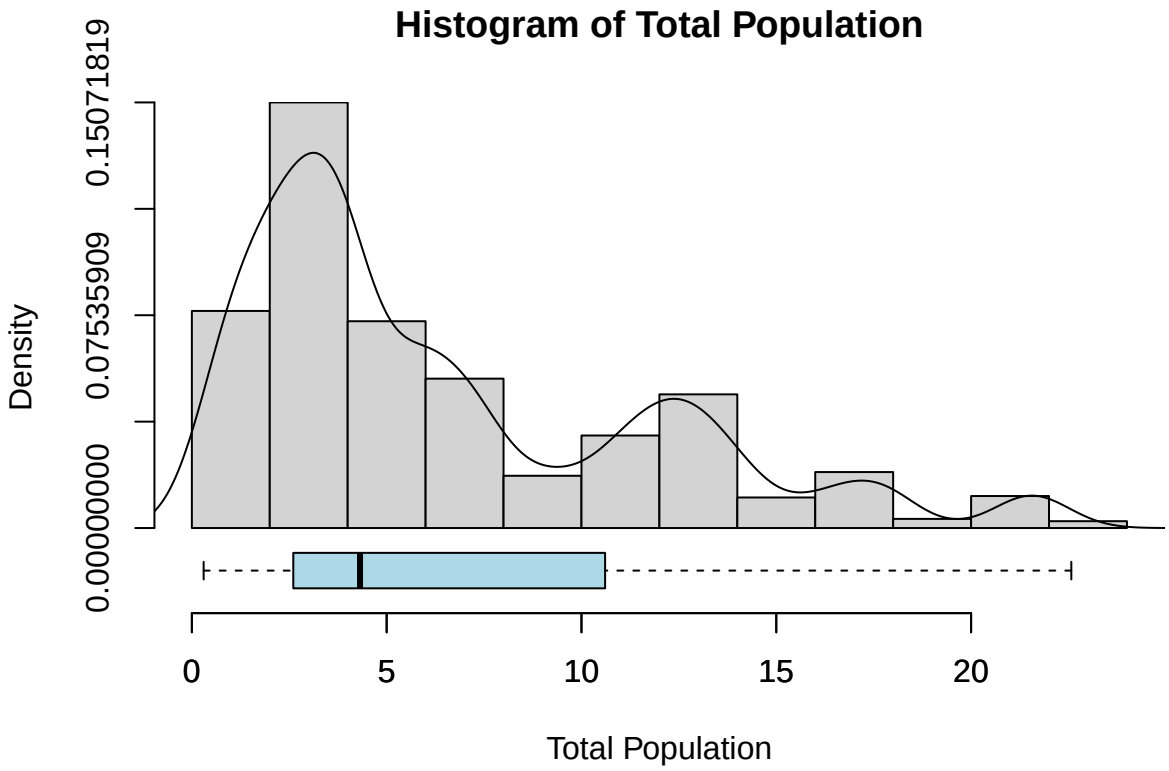



```
##                Abbreviation
## totalpop                  t
## medianhhincom             m
## black_medianhh           b_
## white_medianhh            w
## blackpercent             bl
## urbanpercent              u
```

```
# gives a group of variables that is needed for the model, the lower score is the best
```

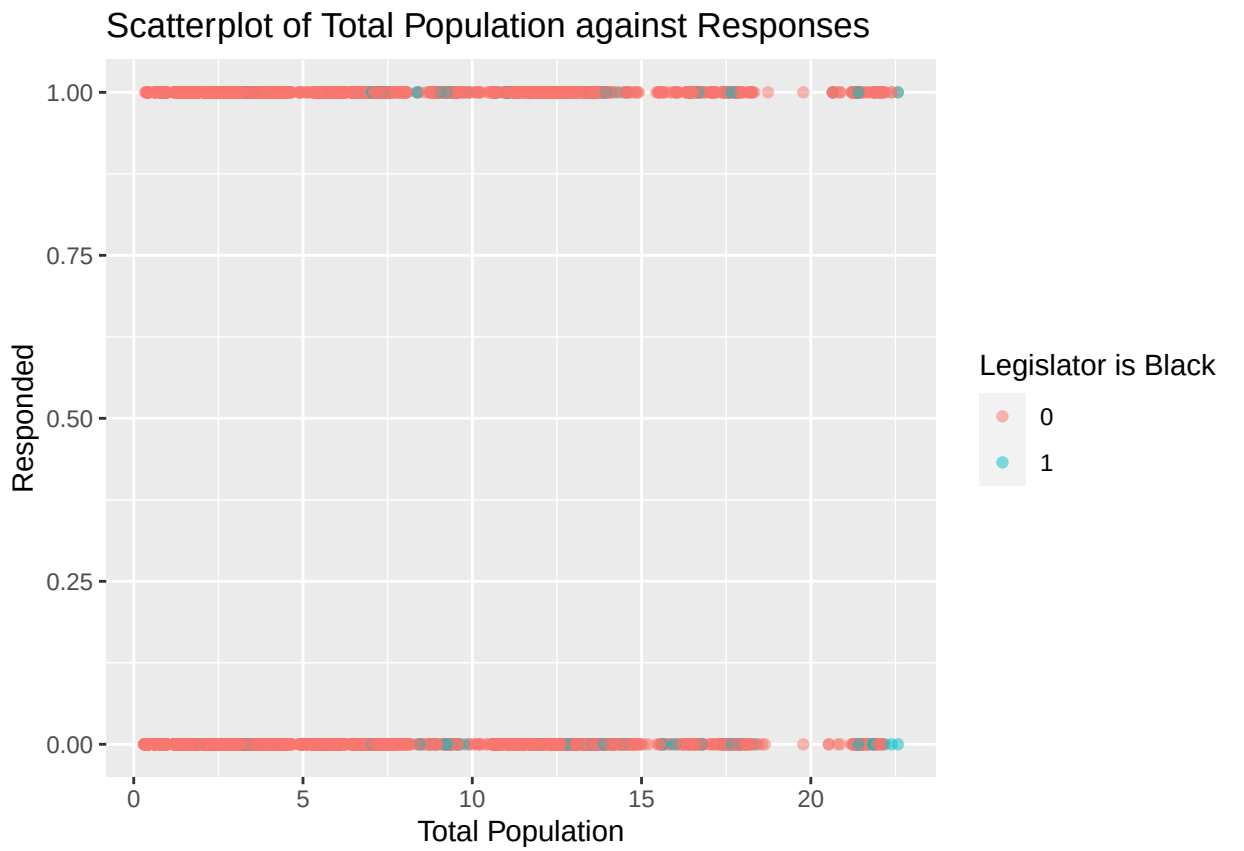
According to the Mallows CP model, model 2 (t-m) has the smallest Mallows' Cp value. Therefore, we would keep predictor variables total population and median household income. Taking both inputs from the Boruta Algorithm and Mallows CP, we decided to run with totalpop (total population) and medianhhincom (district median household income) as our continuous variables.

```
hist_boxplot(black_politicians_c5$totalpop, freq = FALSE,
             main = "Histogram of Total Population",
             xlab = "Total Population")
lines(density(black_politicians_c5$totalpop))
```



The histogram and boxplot of total population is skewed to the right with the majority of the data concentrated at around 4, indicating most states have a population of around 4 million.

```
black_politicians_c5 %>%
ggplot(aes(x = totalpop, y = responded, colour = factor(leg_black))) +
  geom_point(alpha = 0.5) +
  xlab("Total Population") +
  ylab("Responded") +
  guides(color = guide_legend(title = "Legislator is Black")) +
  ggtitle("Scatterplot of Total Population against Responses")
```



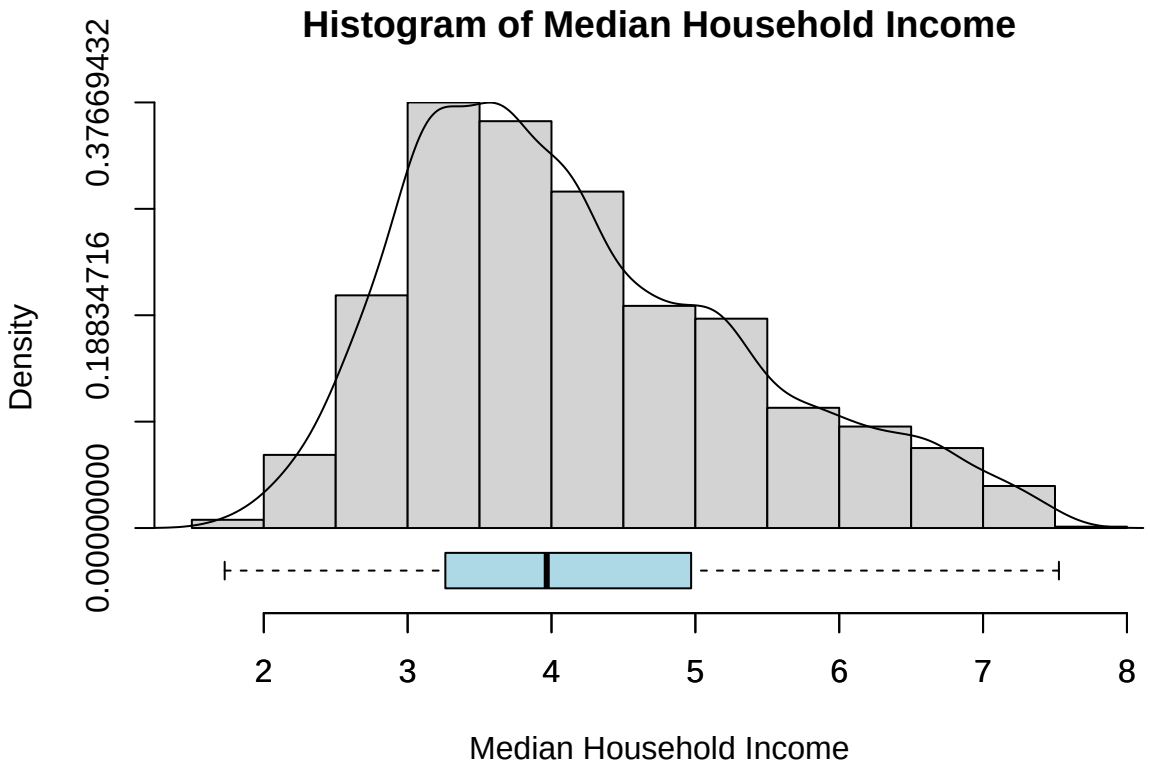
Examining the scatter plot of the legislator responding vs the total population separated by if the legistlator is black, we see that there isn't a clear relationship between the two in the scatter plot, but we do see a lot of red, indicating that the legistlator is black, on both sides.

```
summary(black_politicians_c5$totalpop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3026  2.6034  4.3171  6.5533 10.6066 22.5718
```

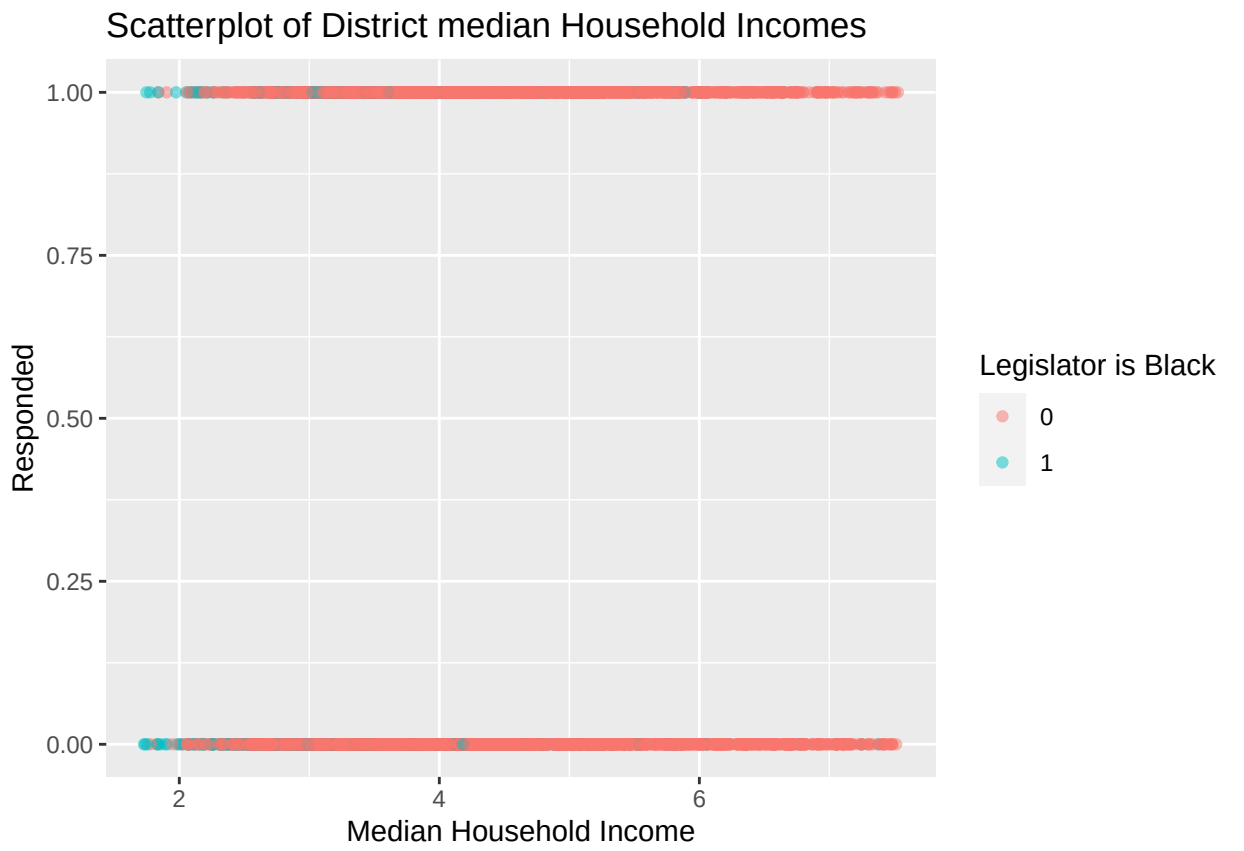
Since our population values are in the millions, the lowest population value in this data set is 302,600, the median population value is around 4 million and the maximum value is 22 million. With 25% of the data being below the value 2.6 million and 75% of the data being below 6.55 million.

```
hist_boxplot(black_politicians_c5$medianhhincom, freq = FALSE,
             main = "Histogram of Median Household Income",
             xlab = "Median Household Income")
lines(density(black_politicians_c5$medianhhincom))
```



The histogram is unimodal, slightly skewed to the right with majority of the data concentrated around median household incomes of $35,000.

```
black_politicians_c5 %>%
ggplot(aes(x = medianhhincom, y = responded, colour = factor(leg_black))) +
  geom_point(alpha = 0.5) +
  xlab("Median Household Income") +
  ylab("Responded") +
  guides(color = guide_legend(title = "Legislator is Black")) +
  ggtitle("Scatterplot of District median Household Incomes")
```




```
summary(black_politicians_c5$medianhhincom)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.728   3.262   3.967   4.185   4.971   7.527
```

Since our median household income value is in the 10,000s, the lowest median house income value in this data set is `$17,280`, the median household income value is `$39,670` and the maximum value is `$75,270`. With 25% of the data being below the value `$32,620` and 75% of the data being below `$49,710`.

3

3.1 Linear Probability Model

```
LPM <- lm(responded ~ totalpop + medianhhincom + leg_black + treat_out +
            nonblacknonwhite + leg_senator + leg_democrat + south,
          data = black_politicians_c5)

summary(LPM)
```

```
##
## Call:
## lm(formula = responded ~ totalpop + medianhhincom + leg_black +
```

```
##     treat_out + nonblacknonwhite + leg_senator + leg_democrat +
##     south, data = black_politicians_c5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7801 -0.3764 -0.2304  0.4595  0.8382
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.500886   0.030618  16.359  < 2e-16 ***
## totalpop           0.008909   0.001412   6.310 3.03e-10 ***
## medianhhincom      0.008110   0.006166   1.315 0.188511
## leg_black         -0.013602   0.028595  -0.476 0.634312
## treat_out         -0.271754   0.013406 -20.270  < 2e-16 ***
## nonblacknonwhite -0.040045   0.033989  -1.178 0.238783
## leg_senator        0.030827   0.016827   1.832 0.067021 .
## leg_democrat      -0.045017   0.014137  -3.184 0.001460 **
## south             -0.055488   0.015865  -3.498 0.000474 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4711 on 4934 degrees of freedom
## Multiple R-squared:  0.09218,    Adjusted R-squared:  0.09071
## F-statistic: 62.63 on 8 and 4934 DF,  p-value: < 2.2e-16
```

The results for our Linear Probability model show that the estimators median household income, the legislator being black, and non black nor white are highly insignificant. The estimator for the legislator being a senator was found to be insignificant at the 5% level. On the other hand, the intercept along with the remaining estimators are all significant. The coefficient for total population is 0.008909, which suggests that an additional person in the district population increases the probability of the legislator responding to emails by 0.891%, all else equal. The negative sign in front of the coefficient for treat_out indicates that the email being out-of-district decreases the probability of the legislator responding to emails by 27.18%, all else equal. The negative sign in front of the coefficient for leg_democrat indicates that the probability of the legislator responding to emails decreases by 4.5% if they are in the Democratic party, all else equal. The negative sign in front of the coefficient for south indicates the probability of the legislator responding to emails by decreases 5.55% if they are from southern United States, all else equal. Lastly, the R-squared value is 0.09218, which means that 9.22% of the variation in whether or not legislators respond to emails is explained by the Linear Probability model.

```
LPM2 <- lm(responded ~ totalpop + medianhhincom + treat_out + leg_democrat
          + south,  data = black_politicians_c5)
summary(LPM2)
```

```
##
## Call:
## lm(formula = responded ~ totalpop + medianhhincom + treat_out +
##     leg_democrat + south, data = black_politicians_c5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7672 -0.3771 -0.2320  0.4589  0.8343
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.504973   0.030120  16.765  < 2e-16 ***
## totalpop       0.009610   0.001312   7.324 2.79e-13 ***
## medianhhincom  0.007937   0.006039   1.314 0.188819
## treat_out     -0.271844   0.013407 -20.277  < 2e-16 ***
## leg_democrat  -0.049134   0.013684  -3.591 0.000333 ***
## south         -0.058775   0.015617  -3.764 0.000169 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4712 on 4937 degrees of freedom
## Multiple R-squared:  0.09126,    Adjusted R-squared:  0.09034
## F-statistic: 99.16 on 5 and 4937 DF,  p-value: < 2.2e-16
```

We re-estimated our Linear Probability Model after noticing that the estimates for black legislators, legislators that are senators, and legislators that were neither white nor black were insignificant. The estimate for median household income remains insignificant, so we are removing this variable from our model. The remaining estimates are still statistically significant and do not appear to have changed significantly in value from our original Linear Probability Model. Lastly, the R-squared value is now 0.09126, which means that 9.13% of the variation in whether or not legislators respond to emails is explained by the Linear Probability model. This is lower than the R-squared value of our previous model, 9.22%, possibly because the number of explanatory variables decreased after our re-estimation.

```
LPM3 <- lm(responded ~ totalpop + treat_out + leg_democrat
           + south,  data = black_politicians_c5)

LPM3R <- coeftest(LPM3, vcov = hccm(LPM3, type = "hc1"))
LPM3R
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)   0.5393389  0.0153729  35.0836 < 2.2e-16 ***
## totalpop      0.0099454  0.0013087   7.5992 3.544e-14 ***
## treat_out    -0.2716125  0.0133933 -20.2797 < 2.2e-16 ***
## leg_democrat -0.0522577  0.0134637  -3.8814 0.0001052 ***
## south        -0.0649098  0.0147542  -4.3994 1.108e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(LPM3)
```

```
## [1] 6596.693
```

In our final re-estimation of the Linear Probability Model, we regress the variable "responded" (legislators responded to emails) on "totalpop" (district population), "treat_out" (email is from out-of-district), "leg_democrat" (legislators in the Democratic Party), and "south" (legislators from southern United States). All of these variables are statically significant as their p-values are extremely small. Lastly, the R-squared value is now 0.09094, which means that 9.10% of the variation in whether or not legislators respond to emails is explained by the Linear Probability model. This is slightly lower than the R-squared value of our previous model, 9.13%, possibly because the number of explanatory variables decreased after our re-estimation. We also perform an AIC calculation on the model and received a score of 6,596. We will be using this score to compare with the probit and logit models.

```
linearHypothesis(LPM3, c("treat_out = 0",
                         "leg_democrat = 0",
                         "south = 0"),
                 vcov. = vcovHC(LPM3, type = "HC1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## treat_out = 0
## leg_democrat = 0
## south = 0
##
## Model 1: restricted model
## Model 2: responded ~ totalpop + treat_out + leg_democrat + south
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   4941
## 2   4938  3 151.79 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Examining further if our binary variables are significant, we run a linear hypothesis code to see if all three variables equal 0. Since we reject the null that the variables "treat_out," "leg_democrat," and "south" equal to zero in the linear hypothesis test, we determined that these variables are statistically significant and should be included in our model. Our final linear probability model is: $Responded = 0.539 + 0.00994totalpop + -0.27161treatout + -0.05226legdemocrat + -0.06491south$

```
LPM_predict <-
round(predict(LPM3, newdata = data.frame(
  totalpop = black_politicians_c5$totalpop,
  treat_out = black_politicians_c5$treat_out,
  leg_democrat = black_politicians_c5$leg_democrat,
  south = black_politicians_c5$south)))
```

```
LPM_results <- (table(LPM_predict == 1, black_politicians_c5$responded))

LPM_results
```

```
##
##            0    1
##   FALSE 1988  889
##   TRUE   865 1201
```

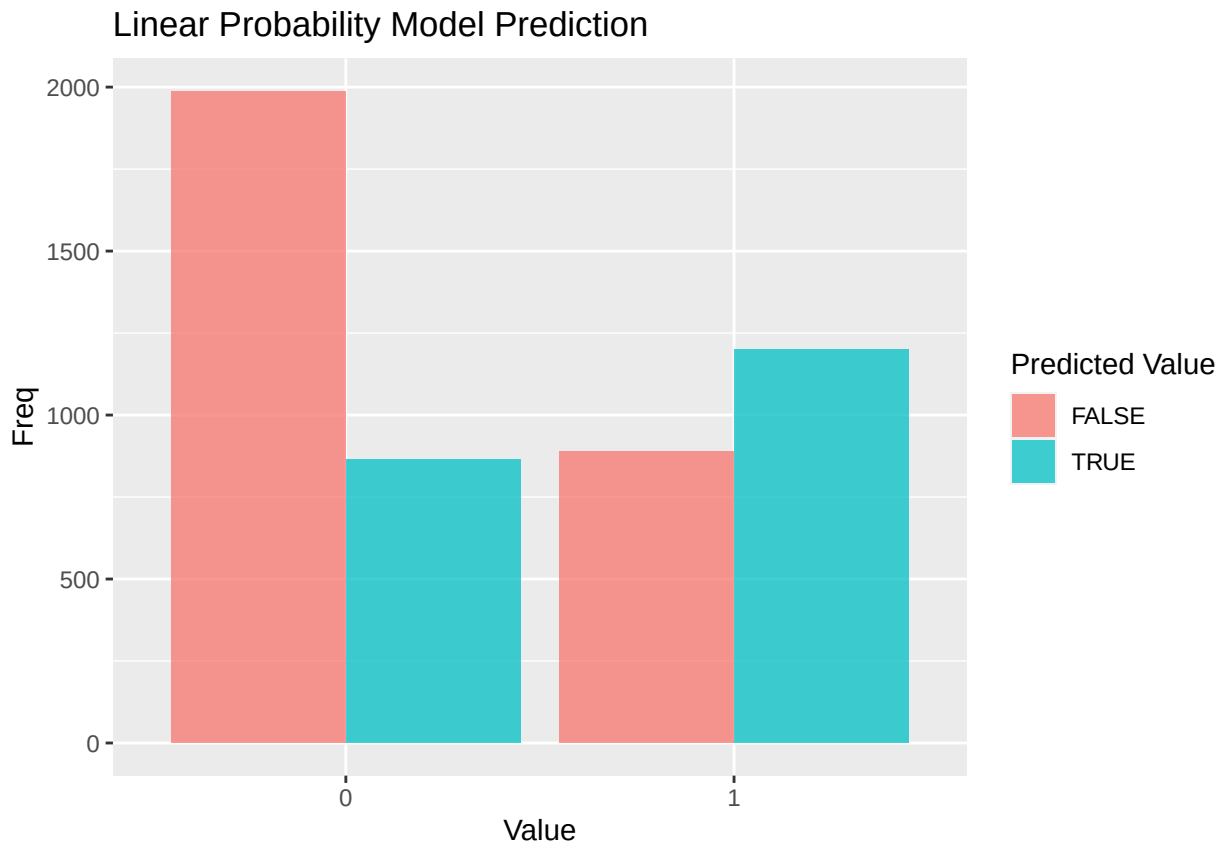
```
c(LPM_results[1, 1] / sum(LPM_results[, 1]),
  LPM_results[2, 2] / sum(LPM_results[, 2]))
```

```
## [1] 0.6968104 0.5746411
```

Using the "predict" function, we are able to run our linear probability model. We decided to use a threshold of 50%, meaning that if it above or equal to 50%, 0.50, we consider that 1 and the legislator responded,

otherwise it is 0 and the legislator did not respond. The accuracy of the model is that if it is FALSE and 0 in the matrix or TRUE and 1, the model predicted it accurately. We see that our final linear probability model predicted 0 correctly 1,988 times and 1 correctly 1,201 times. Examining the percentage it got correct, it was right 69.68% of the time predicting 0s, and 57.46% of the time predicting 1s.

```
Results_LPM <- data.frame(LPM_results)

Results_LPM %>%
  ggplot(aes(x = Var2, y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", position = position_dodge(), alpha = 0.75) +
  labs(x = "Value") +
  guides(fill = guide_legend(title = "Predicted Value")) +
  ggtitle("Linear Probability Model Prediction")
```



This bar chart demonstrates the accuracy of our Linear Probability Model, to prevent confusion, the predicated values are noted by FALSE or TRUE, where FALSE corresponds to predicting 0 and TRUE corresponds to predicting 1. Here, we see that out linear probability model better predicts 0 values than predicting 1 values.

```
margins(LPM3)
```

```
## Average marginal effects

## lm(formula = responded ~ totalpop + treat_out + leg_democrat +     south, data = black_politicians_c
```

```
##  totalpop treat_out leg_democrat    south
##  0.009945   -0.2716     -0.05226 -0.06491
```

For our linear probability model, the average marginal effect of total population, totalpop, is 0.009945402 or 0.99%, the average marginal effect of an email being sent outside the district is, treat_out, is -0.2718125 or -27.18%, the average marginal effect of the legislator being democratic, leg_democrat, is -0.05225774 or -5.22%, and the average marginal effect of the legislator being from the south, south, is -0.0649 or -6.49%.

3.2 Probit Model

```
PM <- glm(responded ~ totalpop + treat_out + leg_democrat + south,
          data = black_politicians_c5, family = binomial(link = "probit"))

summary(PM)
```

```
##
## Call:
## glm(formula = responded ~ totalpop + treat_out + leg_democrat +
##     south, family = binomial(link = "probit"), data = black_politicians_c5)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6799  -0.9678  -0.7354   1.1074   1.8354
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.099426   0.040543   2.452   0.0142 *
## totalpop      0.027137   0.003543   7.659 1.87e-14 ***
## treat_out    -0.720465   0.037066 -19.437  < 2e-16 ***
## leg_democrat -0.145292   0.037147  -3.911 9.18e-05 ***
## south        -0.179749   0.041308  -4.351 1.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6734.2  on 4942  degrees of freedom
## Residual deviance: 6272.3  on 4938  degrees of freedom
## AIC: 6282.3
##
## Number of Fisher Scoring iterations: 4
```

In this probit model, we regressed the variable "responded" (legislators responded to emails) on "totalpop" (district population), "treat_out" (email is from out-of-district), "leg_democrat" (legislators in the Democratic Party), and "south" (legislators from southern United States). All estimates are statistically significant as their p-values are extremely small. The AIC score for our model is 6,282, which is lower than our linear probability model, which indicates our probit model is a better model. Our probit model is $responded = \Phi(0.09943 + 0.02714 totalpop + -0.72047 treat\_out + -0.14529 leg\_democrat + -0.17975 south)$

```
linearHypothesis(PM, c("treat_out = 0",
                       "leg_democrat = 0",
                       "south = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## treat_out = 0
## leg_democrat = 0
## south = 0
##
## Model 1: restricted model
## Model 2: responded ~ totalpop + treat_out + leg_democrat + south
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   4941
## 2   4938  3 403.17  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Further testing to see if our binary variables (treat_out, leg_democrat, south) are all insignificant, we see that our p-value is extremely small, indicating that our binary variables are significant enough to be in the model.

```
PM_predict <-
round(pnorm(predict(PM, newdata = data.frame(
  totalpop = black_politicians_c5$totalpop,
  treat_out = black_politicians_c5$treat_out,
  leg_democrat = black_politicians_c5$leg_democrat,
  south = black_politicians_c5$south))))
```

```
PM_results <-
table(PM_predict == 1, black_politicians_c5$responded)

PM_results
```

```
##
##            0    1
##   FALSE 2019  914
##   TRUE   834 1176
```

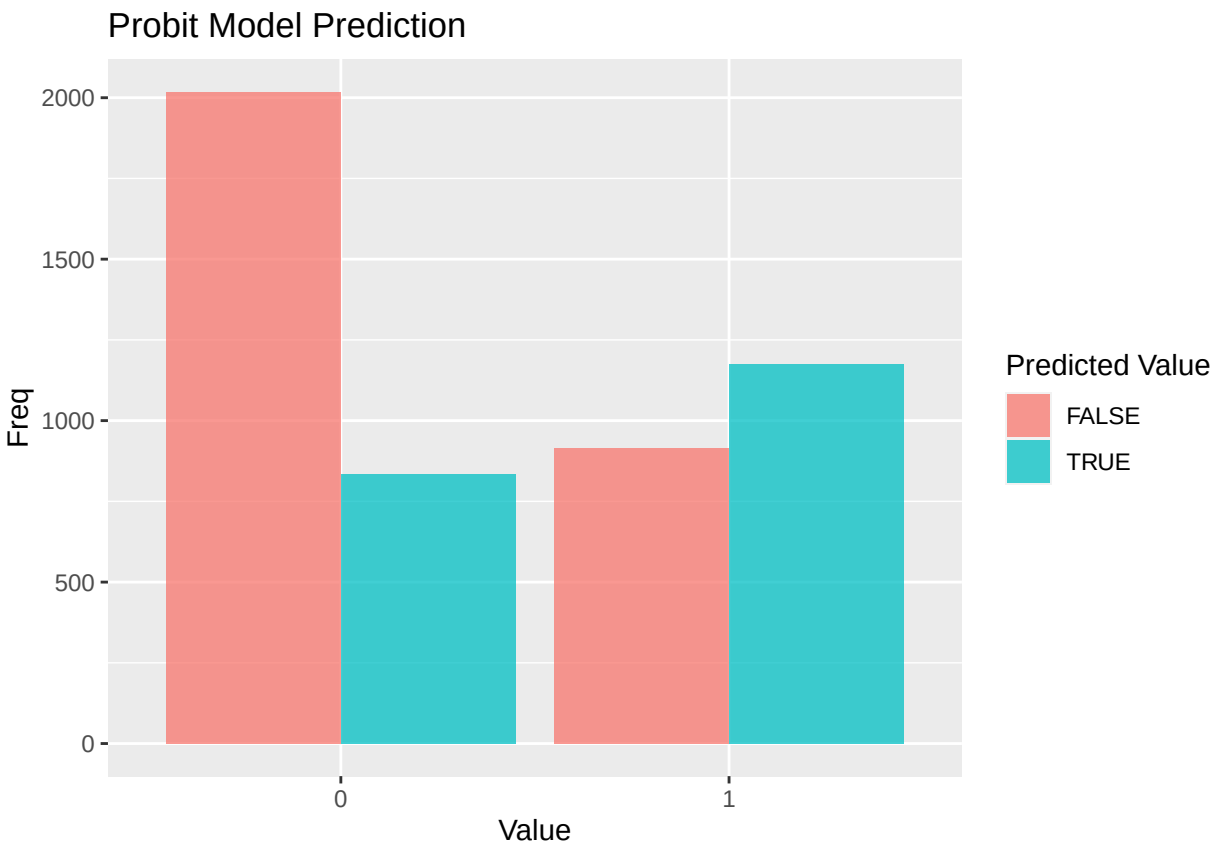
```
c(PM_results[1, 1] / sum(PM_results[, 1]),
  PM_results[2, 2] / sum(PM_results[, 2]))
```

```
## [1] 0.7076761 0.5626794
```

Using the predict and pnorm function, since the probit model follows a normally distributed CDF, we predicted the results and compared them with the actual results into a matrix. The FALSE and 0 section indicates that our probit model predicted 0 values correctly using a 50% threshhold, and the TRUE and 1 section indicates that the probit model predicted 1 values correctly using a 50% threshhold. We also formatted into percentage form and see that our probit model predicted 0 values with 70.77% accuracy and 1 values with 56.27% accuracy.

```
Results_PM <- data.frame(PM_results)

Results_PM %>%
```

```
  ggplot(aes(x = Var2, y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", position = position_dodge(), alpha = 0.75) +
  labs(x = "Value") +
  guides(fill = guide_legend(title = "Predicted Value")) +
  ggtitle("Probit Model Prediction")
```



Like before, our probit model predicts FALSE if it thinks the value is 0, and TRUE if it thinks the value is 1. In the bar chart, we see that our probit model better predicts values of 0 rather than values of 1.

```
margins(PM)
```

```
## Average marginal effects

## glm(formula = responded ~ totalpop + treat_out + leg_democrat +     south, family = binomial(link = '

##  totalpop treat_out leg_democrat    south
##  0.009835   -0.2611     -0.05266 -0.06515
```

The average marginal effect for total population, totalpop, is 0.0098, or 0.98% increase in the chances of having a legislator responding. The average marginal effect at the binary variable if the email was sent out of district, treat_out, is -0.2611, or -26.11% change in the chances of having a legislator responding. The average marginal effect if the legislator is democratic, leg_democrat, is -0.0526 or -5.26% change in the chances of having a legislator responding. The average marginal effect if the legislator is from the south, south, is -0.0651, or -6.51% change in the chances of having a legislator responding.

3.3 Logit Model

```
LM <- glm(responded ~ totalpop + treat_out + leg_democrat + south,
          data = black_politicians_c5, family = binomial(link = "logit"))

summary(LM)
```

```
##
## Call:
## glm(formula = responded ~ totalpop + treat_out + leg_democrat +
##     south, family = binomial(link = "logit"), data = black_politicians_c5)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6780  -0.9661  -0.7377   1.1069   1.8237
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.158627   0.065609   2.418 0.015616 *
## totalpop      0.044217   0.005794   7.631 2.33e-14 ***
## treat_out    -1.167091   0.060894 -19.166  < 2e-16 ***
## leg_democrat -0.235551   0.060743  -3.878 0.000105 ***
## south        -0.293006   0.067724  -4.327 1.51e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6734.2  on 4942  degrees of freedom
## Residual deviance: 6272.7  on 4938  degrees of freedom
## AIC: 6282.7
##
## Number of Fisher Scoring iterations: 4
```

In the last model we are testing, we are using the logit model, in order to predict the results. In our logit model, the equation is $responded = \Lambda(0.15863 + 0.04422totalpop + -1.16709treat\_out + -0.23555leg\_democrat + -0.29301south)$. Our AIC score is 6,282.7, which is only a little bit higher than our probit model, so these two models are extremely close in terms of how good the model is.

```
linearHypothesis(LM, c("treat_out = 0",
                       "leg_democrat = 0",
                       "south = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## treat_out = 0
## leg_democrat = 0
## south = 0
##
## Model 1: restricted model
## Model 2: responded ~ totalpop + treat_out + leg_democrat + south
```

```
##
##   Res.Df Df Chisq Pr(>Chisq)
## 1   4941
## 2   4938  3   389  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Going further, we run a linear hypothesis on the three binary variables (treat_out, leg_democrat, south) to see if they are all equal to 0 and since the p value is extremely low, we reject that they are insignificant.

```
LM_predict <-
round(plogis(predict(LM, newdata = data.frame(
  totalpop = black_politicians_c5$totalpop,
  treat_out = black_politicians_c5$treat_out,
  leg_democrat = black_politicians_c5$leg_democrat,
  south = black_politicians_c5$south))))
```

```
LM_results <-
table(LM_predict == 1, black_politicians_c5$responded)

LM_results
```

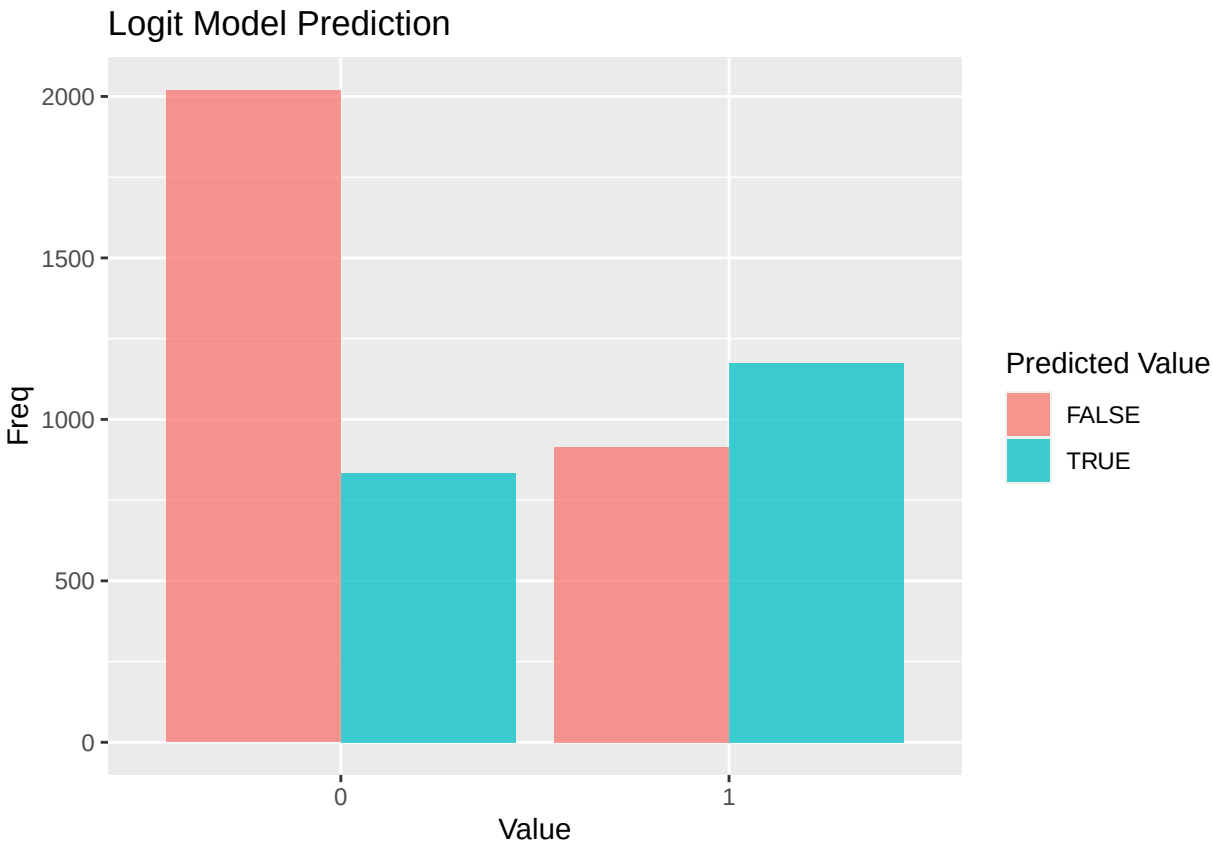
```
##
##            0    1
##   FALSE 2019  915
##   TRUE   834 1175
```

```
c(LM_results[1, 1] / sum(LM_results[, 1]),
  LM_results[2, 2] / sum(LM_results[, 2]))
```

```
## [1] 0.7076761 0.5622010
```

Using a threshold of 50% and the plogis function since the logit model follows a logistic function, indicating that it has fatter tails than the normal distribution in the probit model, we see how accurate our logit model predicted the output. The FALSE and 0 section predicts the 0 values correctly and the TRUE and 1 section predicts the 1 values correctly. We see that the logit model predicted 0 correctly 2,019 times and 1 correctly 1,175 times. The model predicted 0 with 70.77% accuracy and 1 with 56.22% accuracy.

```
Results_LM <- data.frame(LM_results)

Results_LM %>%
  ggplot(aes(x = Var2, y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", position = position_dodge(), alpha = 0.75) +
  labs(x = "Value") +
  guides(fill = guide_legend(title = "Predicted Value")) +
  ggtitle("Logit Model Prediction")
```



Like before, our logit model predicts FALSE if it thinks the value is 0, and TRUE if it thinks the value is 1. As we can see in the bar chart, our logit model predicts relatively the same as our probit model since it better predicts values of 0.

```
margins(LM)
```

```
## Average marginal effects

## glm(formula = responded ~ totalpop + treat_out + leg_democrat +     south, family = binomial(link =

##  totalpop treat_out leg_democrat    south
##  0.009808   -0.2589     -0.05225 -0.06499
```

The average marginal effect of total population, totalpop, is 0.0098, or 0.98% increase in the chance of getting a response in the email for every 1 unit increase in totalpop, the average marginal effect of emails received from out of district, treat_out, is -0.2589, or -25.89% change in the chance of getting a response to the email if the email was from out of district, the average marginal effect of the legislator being democratic, leg_democrat, is -0.0522, or -5.22% change in the chance of getting a response to the email if the legislator is democratic, and the average marginal effect of the legislator being from the south is -0.0650, or -6.50% change in the chance of getting a response to the email if the legislator is from the south.

4

Finally comparing our models, we see that our probit model, $responded = \Phi(0.09943 + 0.02714totalpop + -0.72047treat\_out + -0.14529leg\_democrat + -0.17975south)$, is the best as it had the lowest AIC at

6282.3. This was the tie-breaker between choosing the probit or the logit model as both models predicted the same outcomes relatively closely.