P8130 Final Project

Abstract

Introduction (brief context and background of the problem)

Methods (data description and statistical methods)

Results

Conclusions/Discussion

```
library(tidyverse)
library(ggplot2)
library(GGally)
library(PerformanceAnalytics)
library(performance)
```

Read in dataset

```
cdi = read_csv("./cdi.csv") %>%
  janitor::clean_names()
```

```
## no missing value
cdi %>%
  select(everything()) %>%
  summarise_all(funs(sum(is.na(.)))) %>%
  knitr::kable()
```

id	cty	state	area	pop	pop18	pop65	docs	beds	crimes	hsgrad	bagrad	poverty	unemp	pcincome	totalinc	region
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Data cleaning

```
# some normalization for better comparison
cdi =
  cdi %>%
  mutate(crm_1000 = crimes/pop*1000,  # as indicated by the project prompt
         docs_1000 = docs/pop*1000,  # every 1000 people how many doctors
         beds_1000 = beds/pop*1000,  # ratio of beds per doctor
         pop_density = pop/area,  # how many people per square miles
         region = factor(region)) %>%
  dplyr::select(-id, -crimes,-area, -docs, -beds, -totalinc)
```

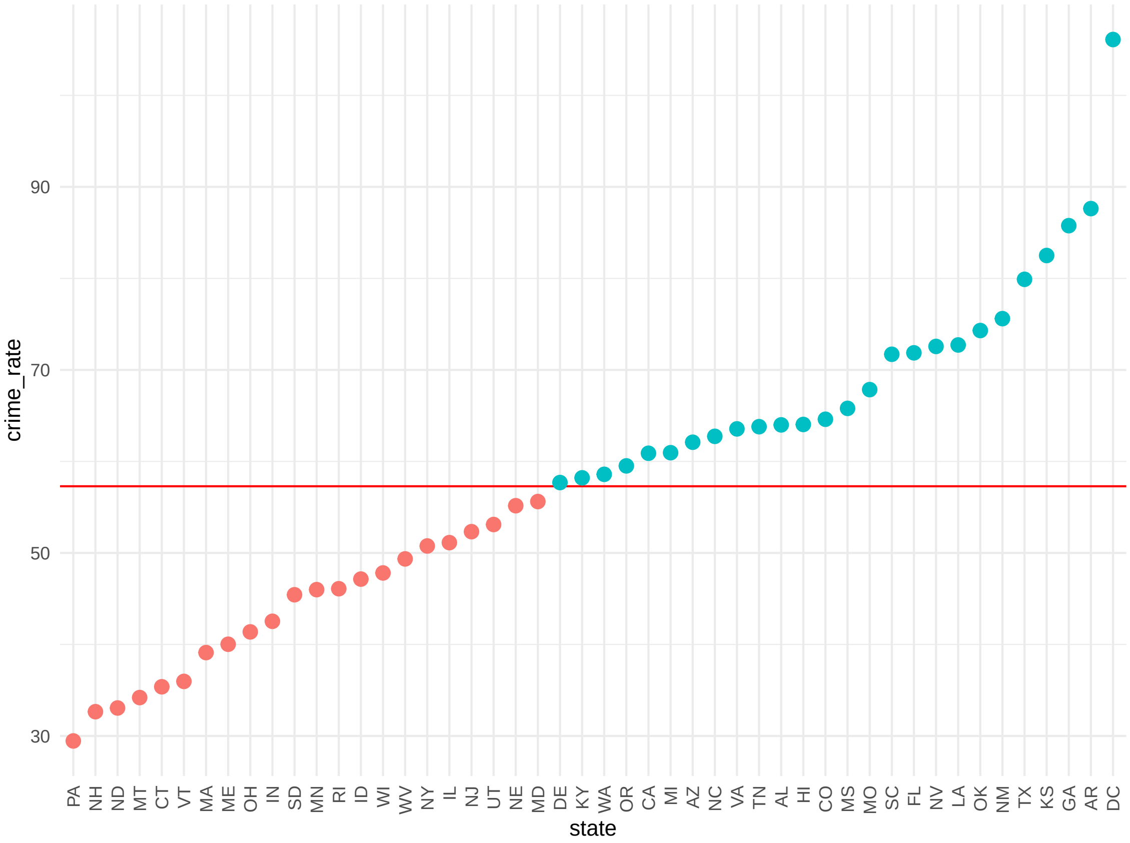
Data Exploration

```
## summary statistics, tentative, NOT FINAL
sum_cdi =
  cdi %>%
  dplyr::select(crm_1000, docs_1000, pop_density, pop, pop18, pop65, hsgrad, bagrad, poverty,unemp, pcin
summary(sum_cdi)
```

```
##     crm_1000         docs_1000        pop_density            pop
##  Min.   :  4.601   Min.   : 0.3559   Min.   :   13.26   Min.   : 100043
##  1st Qu.: 38.102   1st Qu.: 1.2127   1st Qu.:  192.34   1st Qu.: 139027
##  Median : 52.429   Median : 1.7509   Median :  335.91   Median : 217280
##  Mean   : 57.286   Mean   : 2.1230   Mean   :  888.44   Mean   : 393011
##  3rd Qu.: 72.597   3rd Qu.: 2.4915   3rd Qu.:  756.55   3rd Qu.: 436064
##  Max.   :295.987   Max.   :17.0377   Max.   :32403.72   Max.   :8863164
##      pop18           pop65            hsgrad          bagrad
##  Min.   :16.40   Min.   : 3.000   Min.   :46.60   Min.   : 8.10
##  1st Qu.:26.20   1st Qu.: 9.875   1st Qu.:73.88   1st Qu.:15.28
##  Median :28.10   Median :11.750   Median :77.70   Median :19.70
##  Mean   :28.57   Mean   :12.170   Mean   :77.56   Mean   :21.08
##  3rd Qu.:30.02   3rd Qu.:13.625   3rd Qu.:82.40   3rd Qu.:25.32
##  Max.   :49.70   Max.   :33.800   Max.   :92.90   Max.   :52.30
##     poverty          unemp            pcincome       beds_1000
##  Min.   : 1.400   Min.   : 2.200   Min.   : 8899   Min.   : 0.1649
##  1st Qu.: 5.300   1st Qu.: 5.100   1st Qu.:16118   1st Qu.: 2.1972
##  Median : 7.900   Median : 6.200   Median :17759   Median : 3.3287
##  Mean   : 8.721   Mean   : 6.597   Mean   :18561   Mean   : 3.6493
##  3rd Qu.:10.900   3rd Qu.: 7.500   3rd Qu.:20270   3rd Qu.: 4.5649
##  Max.   :36.300   Max.   :21.300   Max.   :37541   Max.   :19.6982
```

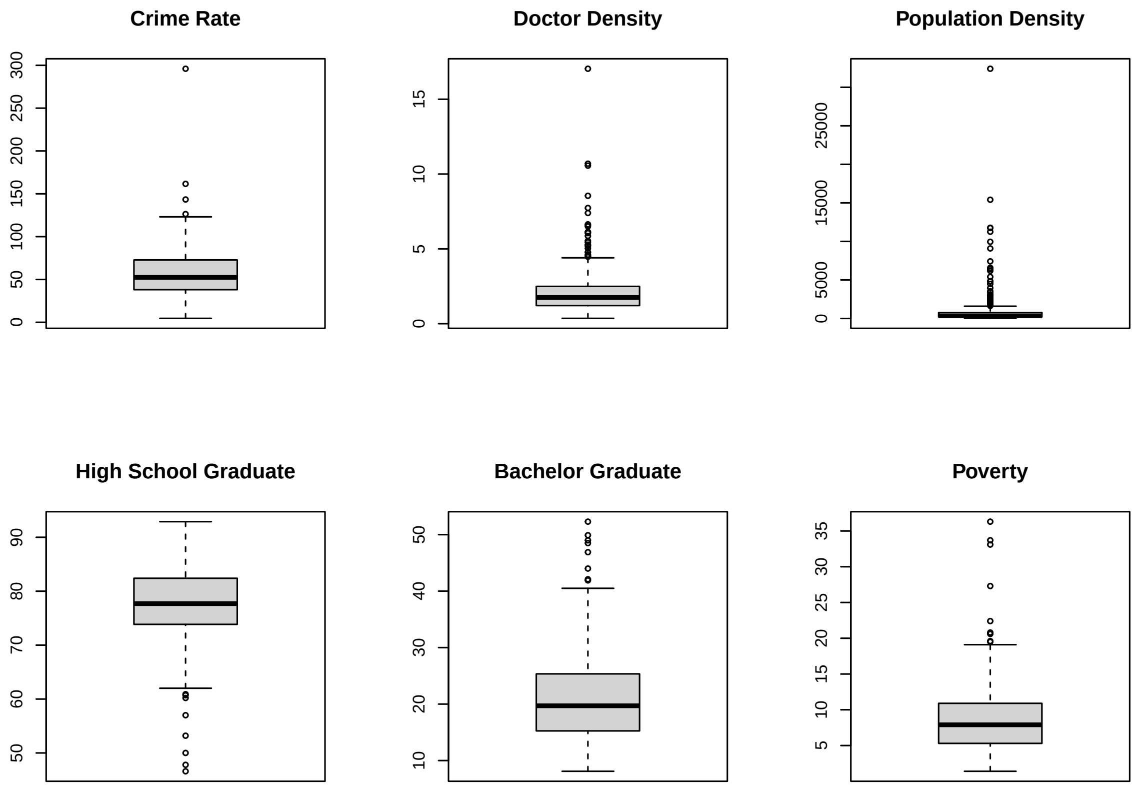
```
mean_crm = mean(sum_cdi$crm_1000)
cdi_state = cdi %>%
  group_by(state) %>%
  summarize(crime_rate = mean(crm_1000)) %>%
  mutate(low_high = ifelse(crime_rate>mean_crm, TRUE,FALSE))


cdi_state %>%
  mutate(state = fct_reorder(state, crime_rate)) %>%
  ggplot(aes(x = state, y = crime_rate))+
  geom_hline(yintercept = mean_crm, color = "red")+
  geom_point(aes(color = low_high),size = 3)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust= 1),
        legend.position = "none")
```

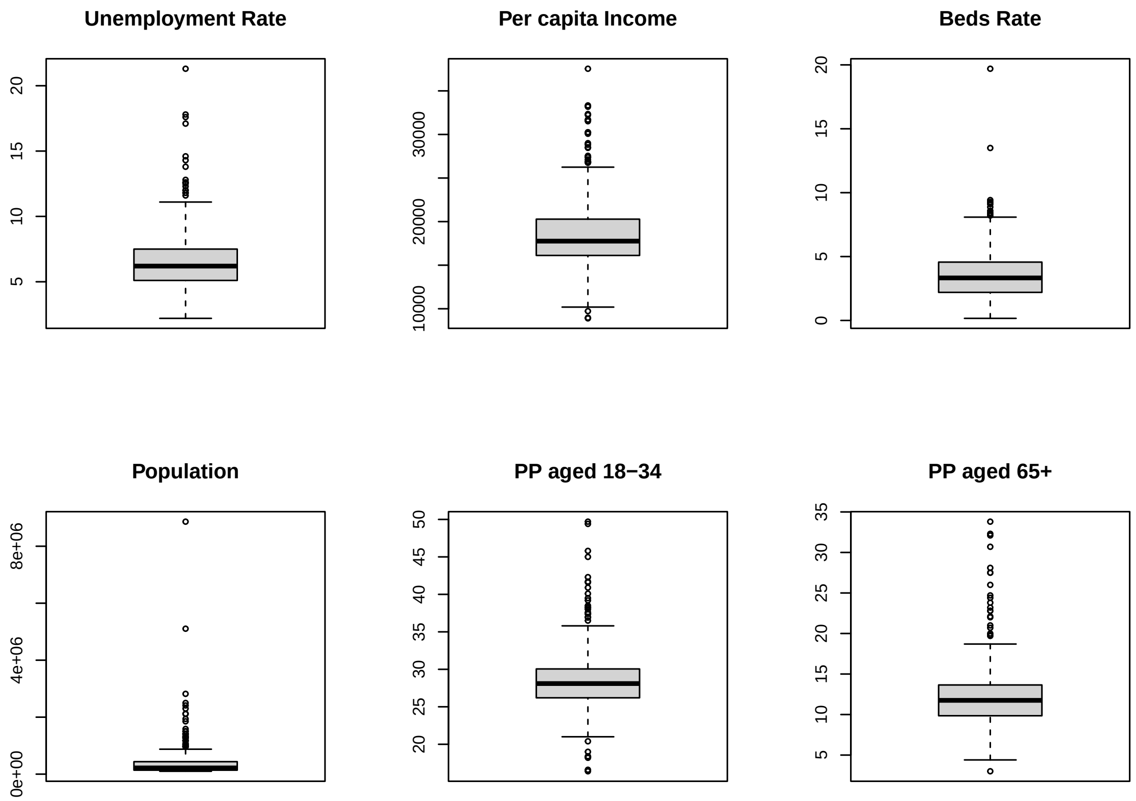


boxplot for each variable

```
par(mfrow=c(2,3))
boxplot(sum_cdi$crm_1000, main='Crime Rate')
boxplot(sum_cdi$docs_1000, main='Doctor Density')
boxplot(sum_cdi$pop_density,main='Population Density' )
boxplot(sum_cdi$hsgrad, main='High School Graduate')
boxplot(sum_cdi$bagrad, main='Bachelor Graduate')
boxplot(sum_cdi$poverty, main='Poverty')
```

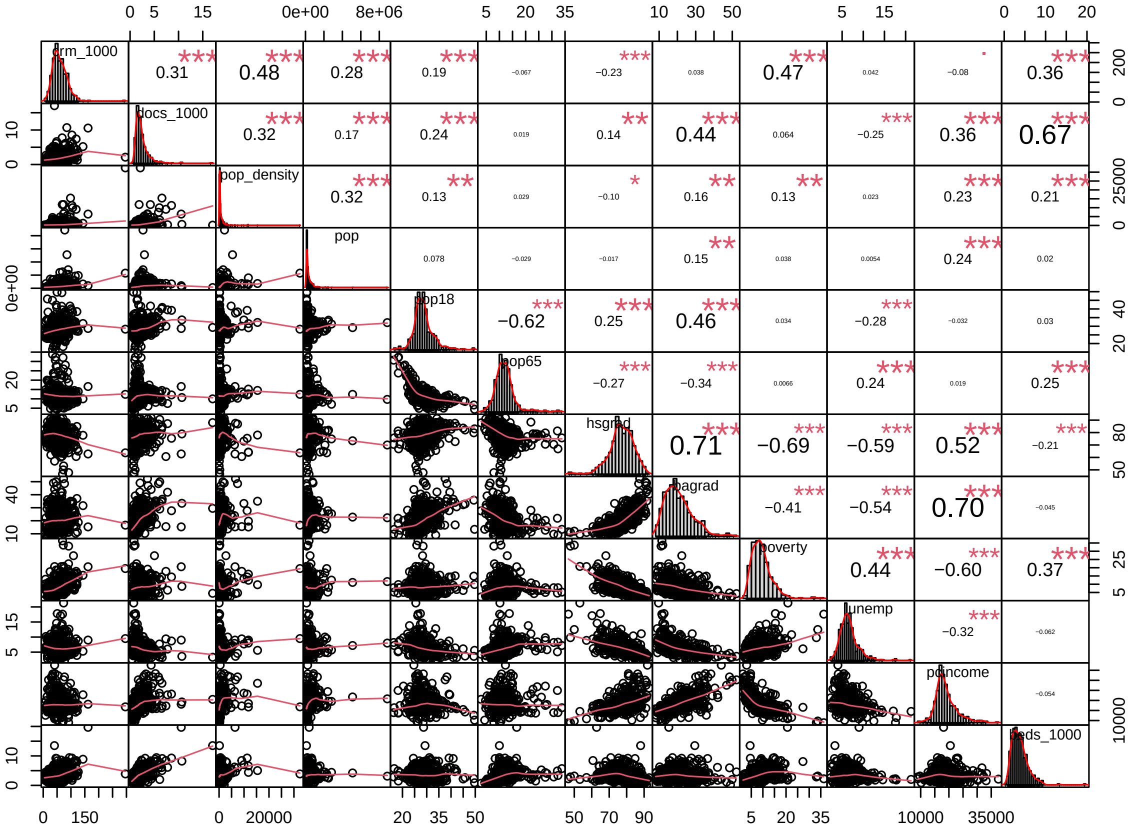


```
par(mfrow=c(2,3))
boxplot(sum_cdi$unemp, main='Unemployment Rate')
boxplot(sum_cdi$pcincome, main='Per capita Income')
boxplot(sum_cdi$beds_1000, main='Beds Rate')
boxplot(sum_cdi$pop, main='Population')
boxplot(sum_cdi$pop18, main='PP aged 18-34')
boxplot(sum_cdi$pop65, main='PP aged 65+')
```



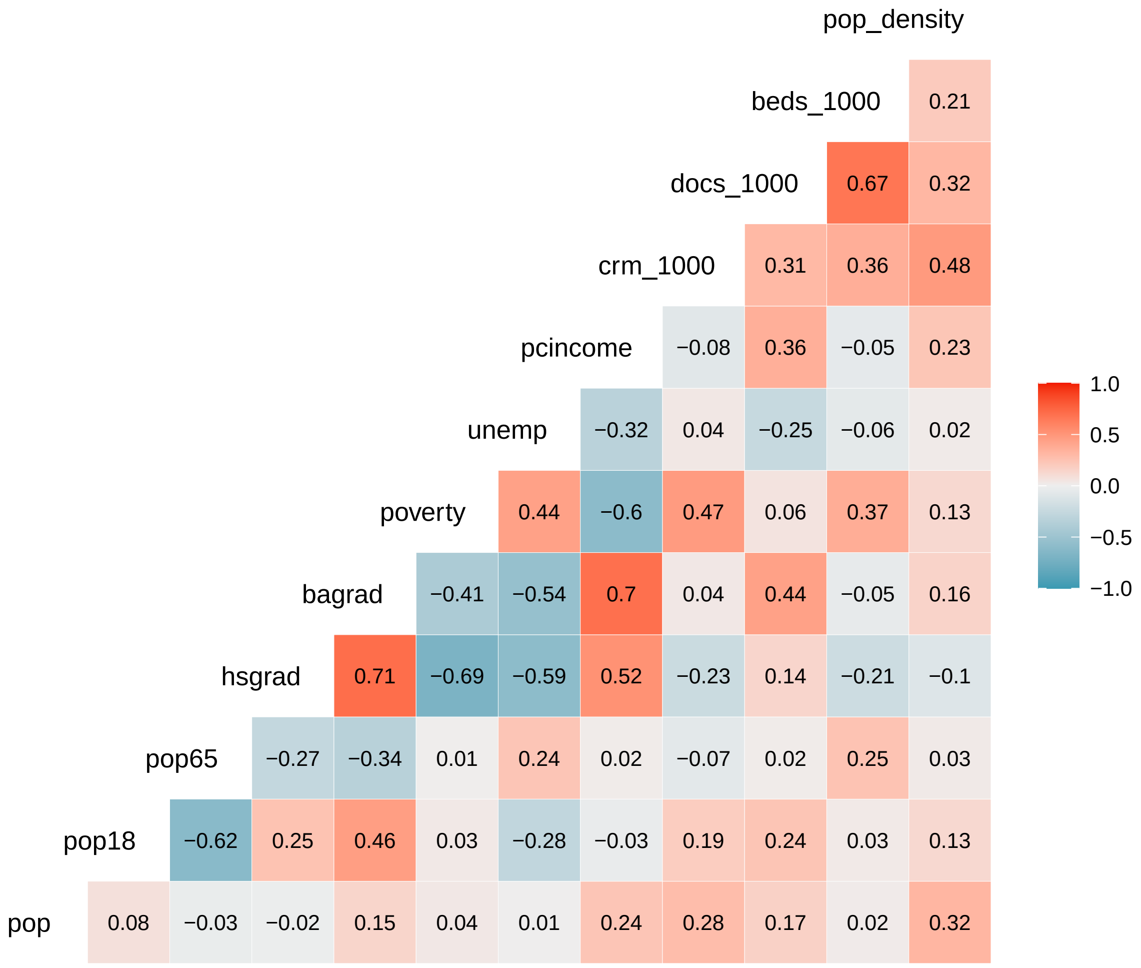
Marginal Correlation and Correlation martix

```
corr_matrix =
  #cdi %>%
  #dplyr::select(-state, -region, -cty) %>%
  sum_cdi %>%
  chart.Correlation(histogram = TRUE, method = "pearson")
```



Correlation Heatmap

```
cdi %>%
  dplyr::select(-state, -region, -cty) %>%
 # sum_cdi %>%
  ggcorr(label=TRUE, hjust = 0.9, layout.exp = 2, label_size = 3, label_round = 2)
```



											pop_density
										beds_1000	0.21
									docs_1000	0.67	0.32
								crm_1000	0.31	0.36	0.48
							pcincome	−0.08	0.36	−0.05	0.23
						unemp	−0.32	0.04	−0.25	−0.06	0.02
					poverty	0.44	−0.6	0.47	0.06	0.37	0.13
				bagrad	−0.41	−0.54	0.7	0.04	0.44	−0.05	0.16
			hsgrad	0.71	−0.69	−0.59	0.52	−0.23	0.14	−0.21	−0.1
		pop65	−0.27	−0.34	0.01	0.24	0.02	−0.07	0.02	0.25	0.03
	pop18	−0.62	0.25	0.46	0.03	−0.28	−0.03	0.19	0.24	0.03	0.13
pop	0.08	−0.03	−0.02	0.15	0.04	0.01	0.24	0.28	0.17	0.02	0.32

```
#corrplot(cor(cdi_1), type = "upper", diag = FALSE)
```

Build Model

Backward Elimination

```
mult_fit = lm(crm_1000 ~ ., data = sum_cdi)
summary(mult_fit)
```

```
##
## Call:
## lm(formula = crm_1000 ~ ., data = sum_cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.873 -12.099  -1.752  12.515  68.501
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.212e+01  2.979e+01  -0.407 0.684403
## docs_1000   -1.671e-01  1.128e+00  -0.148 0.882350
## pop_density  4.278e-03  5.095e-04   8.397 6.76e-16 ***
## pop          6.141e-06  1.756e-06   3.496 0.000521 ***
## pop18        3.287e-01  3.686e-01   0.892 0.373062
## pop65       -2.195e-01  3.388e-01  -0.648 0.517367
## hsgrad       3.306e-01  2.761e-01   1.198 0.231744
## bagrad       2.536e-02  3.247e-01   0.078 0.937773
## poverty      2.930e+00  4.170e-01   7.026 8.40e-12 ***
## unemp       -1.043e+00  5.688e-01  -1.833 0.067424 .
## pcincome     2.881e-04  5.297e-04   0.544 0.586836
## beds_1000    1.816e+00  8.431e-01   2.154 0.031778 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.15 on 428 degrees of freedom
## Multiple R-squared:  0.4698, Adjusted R-squared:  0.4561
## F-statistic: 34.47 on 11 and 428 DF,  p-value: < 2.2e-16
```

```
multi_back = step(mult_fit, direction='backward')
```

```
## Start:  AIC=2654.79
## crm_1000 ~ docs_1000 + pop_density + pop + pop18 + pop65 + hsgrad +
##     bagrad + poverty + unemp + pcincome + beds_1000
##
##               Df Sum of Sq    RSS    AIC
## - bagrad       1       2.5 173835 2652.8
## - docs_1000    1       8.9 173841 2652.8
## - pcincome     1     120.1 173953 2653.1
## - pop65        1     170.5 174003 2653.2
## - pop18        1     322.9 174155 2653.6
## - hsgrad       1     582.5 174415 2654.3
## <none>                     173832 2654.8
## - unemp        1    1365.3 175198 2656.2
## - beds_1000    1    1884.9 175717 2657.5
## - pop          1    4964.9 178797 2665.2
## - poverty      1   20052.0 193885 2700.8
## - pop_density  1   28640.6 202473 2719.9
##
## Step:  AIC=2652.8
## crm_1000 ~ docs_1000 + pop_density + pop + pop18 + pop65 + hsgrad +
##     poverty + unemp + pcincome + beds_1000
##
##               Df Sum of Sq    RSS    AIC
## - docs_1000    1       7.1 173842 2650.8
## - pop65        1     169.0 174004 2651.2
## - pcincome     1     271.2 174106 2651.5
## - pop18        1     457.8 174293 2652.0
## <none>                     173835 2652.8
## - hsgrad       1     877.1 174712 2653.0
## - unemp        1    1509.9 175345 2654.6
## - beds_1000    1    2020.8 175856 2655.9
## - pop          1    4988.1 178823 2663.2
## - poverty      1   25605.9 199441 2711.3
## - pop_density  1   28692.4 202527 2718.0
##
## Step:  AIC=2650.82
## crm_1000 ~ pop_density + pop + pop18 + pop65 + hsgrad + poverty +
##     unemp + pcincome + beds_1000
```

```
##
##               Df Sum of Sq    RSS    AIC
## - pop65        1     168.8 174011 2649.2
## - pcincome     1     286.7 174129 2649.5
## - pop18        1     455.4 174297 2650.0
## <none>                     173842 2650.8
## - hsgrad       1     870.6 174713 2651.0
## - unemp        1    1504.9 175347 2652.6
## - beds_1000    1    3329.7 177172 2657.2
## - pop          1    4981.0 178823 2661.2
## - poverty      1   25960.5 199803 2710.1
## - pop_density  1   28823.8 202666 2716.3
##
## Step:  AIC=2649.24
## crm_1000 ~ pop_density + pop + pop18 + hsgrad + poverty + unemp +
##     pcincome + beds_1000
##
##               Df Sum of Sq    RSS    AIC
## - pcincome     1     312.0 174323 2648.0
## <none>                     174011 2649.2
## - hsgrad       1     997.6 175009 2649.8
## - pop18        1    1206.5 175217 2650.3
## - unemp        1    1663.8 175675 2651.4
## - beds_1000    1    3248.7 177260 2655.4
## - pop          1    4957.7 178969 2659.6
## - pop_density  1   28656.0 202667 2714.3
## - poverty      1   28689.5 202700 2714.4
##
## Step:  AIC=2648.03
## crm_1000 ~ pop_density + pop + pop18 + hsgrad + poverty + unemp +
##     beds_1000
##
##               Df Sum of Sq    RSS    AIC
## <none>                     174323 2648.0
## - pop18        1      1042 175365 2648.7
## - hsgrad       1      1383 175706 2649.5
## - unemp        1      1608 175931 2650.1
## - beds_1000    1      3763 178086 2655.4
## - pop          1      5976 180299 2660.9
## - poverty      1     32317 206640 2720.9
## - pop_density  1     34197 208520 2724.8
```

Forward Selection

```
multi_forward = step(mult_fit, direction = 'forward')
```

```
## Start:  AIC=2654.79
## crm_1000 ~ docs_1000 + pop_density + pop + pop18 + pop65 + hsgrad +
##     bagrad + poverty + unemp + pcincome + beds_1000
```

Both direction

```
multi_both = step(mult_fit, direction = "both")
```

```
## Start:  AIC=2654.79
## crm_1000 ~ docs_1000 + pop_density + pop + pop18 + pop65 + hsgrad +
##     bagrad + poverty + unemp + pcincome + beds_1000
##
##               Df Sum of Sq    RSS    AIC
## - bagrad       1       2.5 173835 2652.8
## - docs_1000    1       8.9 173841 2652.8
## - pcincome     1     120.1 173953 2653.1
## - pop65        1     170.5 174003 2653.2
## - pop18        1     322.9 174155 2653.6
## - hsgrad       1     582.5 174415 2654.3
## <none>                     173832 2654.8
## - unemp        1    1365.3 175198 2656.2
## - beds_1000    1    1884.9 175717 2657.5
## - pop          1    4964.9 178797 2665.2
## - poverty      1   20052.0 193885 2700.8
## - pop_density  1   28640.6 202473 2719.9
##
## Step:  AIC=2652.8
## crm_1000 ~ docs_1000 + pop_density + pop + pop18 + pop65 + hsgrad +
##     poverty + unemp + pcincome + beds_1000
##
##               Df Sum of Sq    RSS    AIC
## - docs_1000    1       7.1 173842 2650.8
## - pop65        1     169.0 174004 2651.2
## - pcincome     1     271.2 174106 2651.5
## - pop18        1     457.8 174293 2652.0
## <none>                     173835 2652.8
## - hsgrad       1     877.1 174712 2653.0
## - unemp        1    1509.9 175345 2654.6
## + bagrad       1       2.5 173832 2654.8
## - beds_1000    1    2020.8 175856 2655.9
## - pop          1    4988.1 178823 2663.2
## - poverty      1   25605.9 199441 2711.3
## - pop_density  1   28692.4 202527 2718.0
##
## Step:  AIC=2650.82
## crm_1000 ~ pop_density + pop + pop18 + pop65 + hsgrad + poverty +
##     unemp + pcincome + beds_1000
##
##               Df Sum of Sq    RSS    AIC
## - pop65        1     168.8 174011 2649.2
## - pcincome     1     286.7 174129 2649.5
## - pop18        1     455.4 174297 2650.0
## <none>                     173842 2650.8
## - hsgrad       1     870.6 174713 2651.0
## - unemp        1    1504.9 175347 2652.6
## + docs_1000    1       7.1 173835 2652.8
## + bagrad       1       0.7 173841 2652.8
## - beds_1000    1    3329.7 177172 2657.2
## - pop          1    4981.0 178823 2661.2
## - poverty      1   25960.5 199803 2710.1
## - pop_density  1   28823.8 202666 2716.3
##
```

```
## Step:  AIC=2649.24
## crm_1000 ~ pop_density + pop + pop18 + hsgrad + poverty + unemp +
##     pcincome + beds_1000
##
##               Df Sum of Sq    RSS    AIC
## - pcincome     1     312.0 174323 2648.0
## <none>                     174011 2649.2
## - hsgrad       1     997.6 175009 2649.8
## - pop18        1    1206.5 175217 2650.3
## + pop65        1     168.8 173842 2650.8
## + docs_1000    1       6.9 174004 2651.2
## + bagrad       1       0.1 174011 2651.2
## - unemp        1    1663.8 175675 2651.4
## - beds_1000    1    3248.7 177260 2655.4
## - pop          1    4957.7 178969 2659.6
## - pop_density  1   28656.0 202667 2714.3
## - poverty      1   28689.5 202700 2714.4
##
## Step:  AIC=2648.03
## crm_1000 ~ pop_density + pop + pop18 + hsgrad + poverty + unemp +
##     beds_1000
##
##               Df Sum of Sq    RSS    AIC
## <none>                     174323 2648.0
## - pop18        1      1042 175365 2648.7
## + pcincome     1       312 174011 2649.2
## - hsgrad       1      1383 175706 2649.5
## + pop65        1       194 174129 2649.5
## + bagrad       1       177 174146 2649.6
## + docs_1000    1        26 174297 2650.0
## - unemp        1      1608 175931 2650.1
## - beds_1000    1      3763 178086 2655.4
## - pop          1      5976 180299 2660.9
## - poverty      1     32317 206640 2720.9
## - pop_density  1     34197 208520 2724.8
```

Residuals vs. Fitted && QQ Plots

Check Multicollinearity

```
check_collinearity(multi_forward)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##         Term  VIF Increased SE Tolerance
##    docs_1000 2.63         1.62      0.38
##  pop_density 1.00         1.00      1.00
##          pop 1.00         1.00      1.00
##        pop18 2.52         1.59      0.40
##        pop65 1.97         1.40      0.51
##       hsgrad 2.71         1.65      0.37
##       bagrad 3.47         1.86      0.29
##      poverty 2.26         1.50      0.44
##        unemp 1.69         1.30      0.59
##     pcincome 1.02         1.01      0.98
##    beds_1000 2.91         1.71      0.34
```

```
check_collinearity(multi_back)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##         Term  VIF Increased SE Tolerance
##  pop_density 1.00         1.00      1.00
##          pop 1.00         1.00      1.00
##        pop18 1.23         1.11      0.81
##       hsgrad 2.60         1.61      0.38
##      poverty 2.38         1.54      0.42
##        unemp 1.74         1.32      0.57
##    beds_1000 1.25         1.12      0.80
```

```
check_collinearity(multi_both)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##         Term  VIF Increased SE Tolerance
##  pop_density 1.00         1.00      1.00
##          pop 1.00         1.00      1.00
##        pop18 1.23         1.11      0.81
##       hsgrad 2.60         1.61      0.38
##      poverty 2.38         1.54      0.42
##        unemp 1.74         1.32      0.57
##    beds_1000 1.25         1.12      0.80
```