What Types of Crimes are Principal Factors best at describing the rest of the crimes in USA (1973)

SDS348 Spring 2021

Name: Santhosh Saravanan

EID: sks3648

Introduction

The first dataset I chose was from the Vincentaelbuldock Github website concentrated on violent crime rates by states in the United States in 1973 titled USArrests. This dataset contains 50 rows by 4 columns, with the 50 rows containing the 50 states of the United States and the 4 columns are respectively Murder, Assault, Rape, and UrbanPop. These columns are ratios of murders, assaults, and rapes commited per 100,000 people in each state. The Urbanpop column features the percentage of the urban population in each state. These ratios were taken from the data in McNeil's monograph in 1975. The second dataset I chose to possible combine with this dataset is from the Bureau of Justice Statistics database and these observations are taken from "local, county, state, tribal, and federal law enforcement agencies" who want to contribute to the FBI's effort to modernize the reports of national crime data. The second dataset contains 2752 rows by 21 columns, with the 2752 rows containing all of the possible 50 states (including the United States as a whole nation), reporting all 21 different crime rates (per 100,000 people in the state/nation) from 1960 to 2012. The 21 columns are State ,Year, Data.Population , rates per 100,000 people in each state for all property, burglary, larceny, motor, assault, murder, and rape crimes for the various years.There are also 6-7 columns for total crimes related to property, burglary, larceny, motor, assault, murder, and rape crimes for the various years. There are also total columns adding up the total crimes per 100,000 for each criminal charge. From watching Forensic Files, Dateline, and other crime shows within the past year from quarantine, I'm interesting in see if there are any crimes that are correlated in 1973 (the year of the Vietnam war) and a curiosity as to whether anti-Vietnam tensions may have affected the crime rates, as correlation does not simply imply causation. From the national news on crime statistics, I have an underlying belief that burglary and larceny rates may be correlated, somewhat, as many riots and thefts were occuring during the United States during that time frame. Murder is also a ratio which strikes me as an interesting ratio to study as well.

Tidy

```r
#Import necessary packages
library(readxl)
library(tidyverse)
library(kableExtra)
USArrests6_1973 <- as.data.frame(read_csv("~/git/SDS348/Project/Datasets/USArrests6_1973.csv")) #read the Excel fi
le and save as a dataframe
glimpse(USArrests6_1973)
```

```
## Rows: 50
## Columns: 5
## $ X1       <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "California", "Co…
## $ Murder   <dbl> 13.2, 10.0, 8.1, 8.8, 9.0, 7.9, 3.3, 5.9, 15.4, 17.4, 5.3, 2.…
## $ Assault  <dbl> 236, 263, 294, 190, 276, 204, 110, 238, 335, 211, 46, 120, 24…
## $ UrbanPop <dbl> 58, 48, 80, 50, 91, 78, 77, 72, 80, 60, 83, 54, 83, 65, 57, 6…
## $ Rape     <dbl> 21.2, 44.5, 31.0, 19.5, 40.6, 38.7, 11.1, 15.8, 31.9, 25.8, 2…
```

```r
allArrests <- as.data.frame(read_csv("~/git/SDS348/Project/Datasets/state_crime_all.csv")) #read the Excel file an
d save as a dataframe
print(colnames(allArrests)) #print the colNames to get an understanding of the data we're dealing with
```

```
##  [1] "State"                      "Year"
##  [3] "Data.Population"            "Data.Rates.Property.All"
##  [5] "Data.Rates.Property.Burglary" "Data.Rates.Property.Larceny"
##  [7] "Data.Rates.Property.Motor"  "Data.Rates.Violent.All"
##  [9] "Data.Rates.Violent.Assault" "Data.Rates.Violent.Murder"
## [11] "Data.Rates.Violent.Rape"    "Data.Rates.Violent.Robbery"
## [13] "Data.Totals.Property.All"   "Data.Totals.Property.Burglary"
## [15] "Data.Totals.Property.Larceny" "Data.Totals.Property.Motor"
## [17] "Data.Totals.Violent.All"    "Data.Totals.Violent.Assault"
## [19] "Data.Totals.Violent.Murder" "Data.Totals.Violent.Rape"
## [21] "Data.Totals.Violent.Robbery"
```

```r
glimpse(allArrests) # look at the data previews to check if it's tidy.
```

```
## Rows: 2,751
## Columns: 21
## $ State                         <chr> "Alabama", "Alabama", "Alabama", "Alabam…
## $ Year                          <dbl> 1960, 1961, 1962, 1963, 1964, 1965, 1966…
## $ Data.Population               <dbl> 3266740, 3302000, 3358000, 3347000, 3407…
## $ Data.Rates.Property.All       <dbl> 1035.4, 985.5, 1067.0, 1150.9, 1358.7, 1…
## $ Data.Rates.Property.Burglary  <dbl> 355.9, 339.3, 349.1, 376.9, 466.6, 473.7…
## $ Data.Rates.Property.Larceny   <dbl> 592.1, 569.4, 634.5, 683.4, 784.1, 812.1…
## $ Data.Rates.Property.Motor     <dbl> 87.3, 76.8, 83.4, 90.6, 108.0, 106.9, 13…
## $ Data.Rates.Violent.All        <dbl> 186.6, 168.5, 157.3, 182.7, 213.1, 199.8…
## $ Data.Rates.Violent.Assault    <dbl> 138.1, 128.9, 119.0, 142.1, 163.0, 149.1…
## $ Data.Rates.Violent.Murder     <dbl> 12.4, 12.9, 9.4, 10.2, 9.3, 11.4, 10.9, …
## $ Data.Rates.Violent.Rape       <dbl> 8.6, 7.6, 6.5, 5.7, 11.7, 10.6, 9.7, 10.…
## $ Data.Rates.Violent.Robbery    <dbl> 27.5, 19.1, 22.5, 24.7, 29.1, 28.7, 32.0…
## $ Data.Totals.Property.All      <dbl> 33823, 32541, 35829, 38521, 46290, 48215…
## $ Data.Totals.Property.Burglary <dbl> 11626, 11205, 11722, 12614, 15898, 16398…
## $ Data.Totals.Property.Larceny  <dbl> 19344, 18801, 21306, 22874, 26713, 28115…
## $ Data.Totals.Property.Motor    <dbl> 2853, 2535, 2801, 3033, 3679, 3702, 4606…
## $ Data.Totals.Violent.All       <dbl> 6097, 5564, 5283, 6115, 7260, 6916, 8098…
## $ Data.Totals.Violent.Assault   <dbl> 4512, 4255, 3995, 4755, 5555, 5162, 6249…
## $ Data.Totals.Violent.Murder    <dbl> 406, 427, 316, 340, 316, 395, 384, 415, …
## $ Data.Totals.Violent.Rape      <dbl> 281, 252, 218, 192, 397, 367, 341, 371, …
## $ Data.Totals.Violent.Robbery   <dbl> 898, 630, 754, 828, 992, 992, 1124, 1167…
```

```r
#Each state form a row with information regarding the rates of crime based on the year. Other descriptions of the data are further divided into individual columns. Hence, this data is tidy.
colnames(USArrests6_1973)[1] <- "State"
#State wasn't capitalized in my 1973 dataset and there was a weird Unicode character when reading the dataset as an Excel spreadsheet. So that column was renamed.
allArrests1973 <- allArrests %>% filter(Year == 1973 & State != "United States")
#Save a dataframe where I filter the dataset by the year of interest and don't want to include the national estimates for crime rates, because I'm concerned only with the state crime rates.
```

Join/Merge

```r
USArrestscombined <- USArrests6_1973 %>% left_join(allArrests1973,by = c("State")) %>% filter(!is.na(State))
#All of the cases that didn't corresponding to the year 1973 were all discarded. Initial attempts to average the data from 1960 to 2012 was a possibility, but there were a myriad of NA values (45 to be exact) and I couldn't fill in all of those values with previous existing knowledge. The database did have some processing errors and couldn't accurately report the figures. To minimize bias, I decided to only select all 50 states and their crime figures for 1973 since there were no missing values. I added a filter just in case for good coding practice.
```

Summary Statistics

```r
#I want to add a categorical variable called Region, which is representative of which states belong in the NorthEast (NE) Region, Midwest (MW) Region, South (S) Region, and West (W) Region. The following vectors contains all of the states that are in that region.
NE<- c("Connecticut","Maine","Massachusetts","New Hampshire",
             "Rhode Island","Vermont","New Jersey","New York",
             "Pennsylvania")
MW<- c("Indiana","Illinois","Michigan","Ohio","Wisconsin",
             "Iowa","Kansas","Minnesota","Missouri","Nebraska",
             "North Dakota","South Dakota")
S<- c("Delaware","District of Columbia","Florida","Georgia",
            "Maryland","North Carolina","South Carolina","Virginia",
            "West Virginia","Alabama","Kentucky","Mississippi",
            "Tennessee","Arkansas","Louisiana","Oklahoma","Texas")
W<- c("Arizona","Colorado","Idaho","New Mexico","Montana",
            "Utah","Nevada","Wyoming","Alaska","California",
            "Hawaii","Oregon","Washington")
#Lines 66-71
# Step #1 : mutate a column called region where for every state in the State column, it's region value is either Midwest, West, NorthEast, or South depending on which vector that state is present in.
#Step #2 : arrange the dataframe in descending order starting from the first letter closest to
# Step #3 : don't select the rates.violent.murder, rates.violent.assault, and rate.violent.rape columns due to extraneous values and a lot of NA values. I could fill these in with the mean or median of the previous values, but I don't have any historical context and don't know how to predict these values without using ML procedures :(.
# Step #4: Filter the years to only include 1973 and arrange from ascending order by State (starting from state closest to the first letter of the alphabet.)
#Step #5: IMPORTANT ( Focus on Murder, Larceny, and Burglary for the analyses of numerical distributions of the estimators)
USArrestscombined <- USArrestscombined %>%
  mutate(Region = case_when(State %in% MW ~ "MidWest",
                        State %in% W  ~ "West",
                        State %in% NE  ~ "NorthEast",
                        State %in% S ~ "South")) %>% arrange(desc(Region))
USArrestscombinedImportant <- USArrestscombined %>% select(-c(Data.Rates.Violent.Murder, Data.Rates.Violent.Assault, Data.Rates.Violent.Rape)) %>% filter(Year == 1973) %>% arrange(State)

#summarize the mean rates for different crime rates without grouping by region and use the kbl and kable to convert it to a readable table.
USArrestscombinedImportant %>%
  summarise(mean_rape = mean(Rape, na.rm = TRUE),mean_assault = mean(Assault, na.rm = TRUE),mean_murder = mean(Murder, na.rm = TRUE), mean_burglary = mean(Data.Rates.Property.Burglary,na.rm = TRUE), mean_larceny = mean(Data.Rates.Property.Larceny,na.rm = TRUE), mean_motor = mean(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption = "Means from different crime rates in USA (1973) without grouping") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Means from different crime rates in USA (1973) without grouping

mean_rape	mean_assault	mean_murder	mean_burglary	mean_larceny	mean_motor
21.232	170.76	7.788	1096.822	2041.032	369.798

The distribution of the means is interesting to note in that several of the values seemed to be heavily impacted by outlers. For instance, the mean murder rates when considering all of the states as a whole is astonishgly low, while the rates for burglary and larceny are at an all-time high. To reiterate, the means are at an all-time high, so it's imperative that we check other numerical descriptions of our data.

```r
#summarize the minimum rates for different crime rates without grouping by region and use the kbl and kable to convert it to a readable table.
USArrestscombinedImportant %>%
  summarise(min_rape = min(Rape, na.rm = TRUE),min_assault = min(Assault, na.rm = TRUE),min_murder = min(Murder, na.rm = TRUE), min_burglary = min(Data.Rates.Property.Burglary,na.rm = TRUE), min_larceny = min(Data.Rates.Property.Larceny,na.rm = TRUE), min_motor = min(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption = "Minima from different crime rates in USA (1973) without grouping") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Minima from different crime rates in USA (1973) without grouping

min_rape	min_assault	min_murder	min_burglary	min_larceny	min_motor

min_rape	min_assault	min_murder	min_burglary	min_larceny	min_motor
7.3	45	0.8	383.4	824.9	107.1

The distribution of the minima give us an idea of the lowest possible crime rates for different crimes. Similar to the dsitributions of means, the minima for burglary and larceny is much higher than the minima of the other rates, leading into a hypothesis that these two variables may be crucial to other aspects of statistical analysis. It's suprising to note how low murder,rape, and assaults rates were in 1973; an inference could possible be the the mindset of the people in the United States could de directed towards national security and pride.

```
#summarize the maximum rates for different crime rates without grouping by region and use the kbl and kable to con
vert it to a readable table.
USArrestscombinedImportant %>%
  summarise(max_rape = max(Rape, na.rm = TRUE),max_assault = max(Assault, na.rm = TRUE),max_murder = max(Murder,
na.rm = TRUE), max_burglary = max(Data.Rates.Property.Burglary,na.rm = TRUE), max_larceny = max(Data.Rates.Proper
ty.Larceny,na.rm = TRUE), max_motor = max(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption = "Maxima fro
m different crime rates in USA (1973) without grouping") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Maxima from different crime rates in USA (1973) without grouping

max_rape	max_assault	max_murder	max_burglary	max_larceny	max_motor
46	337	17.4	2149.8	3720.1	1109.6

The distribution of the maxima give us an idea of the highest possible crime rates for different crimes. Similar to the distribution of means and minima, the maxima for burglary and larceny is much higher than the minima of the other rates, leading to a hypothesis that these two variables may be crucial to other aspects of statistical analysis.

```
#summarize the IQR values for different crime rates without grouping by region and use the kbl and kable to conver
t it to a readable table.
USArrestscombinedImportant %>%
summarise(IQR_rape = IQR(Rape, na.rm = TRUE),IQR_assault = IQR(Assault, na.rm = TRUE),IQR_murder = IQR(Murder, na
.rm = TRUE), IQR_burglary = IQR(Data.Rates.Property.Burglary,na.rm = TRUE), IQR_larceny = IQR(Data.Rates.Property
.Larceny,na.rm = TRUE), IQR_motor = IQR(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption = "IQRS from di
fferent crime rates in USA (1973) without grouping") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

IQRS from different crime rates in USA (1973) without grouping

IQR_rape	IQR_assault	IQR_murder	IQR_burglary	IQR_larceny	IQR_motor
11.1	140	7.175	460.725	1014.5	285.125

The distribution of the IQRs give us an idea of where the range between the 1st and 3rd quartiles lie for each crime rate. Similar to the previous distributions, the IQR values of burglary and larceny are exceptionally high, suggesting that it' smore like that rates of burglary and larceny have more variation compared to the rest of the crime rates. The murder and rape rates have the least amount of variation. These claims directly coincide with why we saw high means, minima, and maxima for burglary and larceny.

```
#summarize the Mean values for different crime rates grouping by region and use the kbl and kable to convert it to
a readable table.
USArrestscombinedImportant %>%
  group_by(Region) %>%
  summarise(mean_rape = mean(Rape, na.rm = TRUE),mean_assault = mean(Assault, na.rm = TRUE),mean_murder = mean(Mu
rder, na.rm = TRUE), mean_burglary = mean(Data.Rates.Property.Burglary,na.rm = TRUE), mean_larceny = mean(Data.Ra
tes.Property.Larceny,na.rm = TRUE), mean_motor = mean(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption =
"Means from different crime rates in USA (1973)") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Means from different crime rates in USA (1973)

Region	mean_rape	mean_assault	mean_murder	mean_burglary	mean_larceny	mean_motor
MidWest	18.44167	120.3333	5.700000	894.525	1972.400	317.7750
NorthEast	13.77778	126.6667	4.700000	1041.656	1677.667	508.7333
South	21.16250	220.0000	11.706250	1025.006	1657.412	293.0000
West	29.05385	187.2308	7.030769	1410.138	2828.092	416.1538

The distribution of the means grouped by region gives us a closer look as to how the individual means from each region contribute to a higher mean total. In the Northeast and South regions of the USA, there are two massive spikes for larceny and murder, and these two regions are the clear give-away as to why the means are extremely big when ungrouped. In contrast, the Northeast and Midwest have some of the lowest mean ratios for murder.

```r
#summarize the minima for different crime rates grouping by region and use the kbl and kable to convert it to a re
adable table.
USArrestscombinedImportant %>%
  group_by(Region) %>%
  summarise(min_rape = min(Rape, na.rm = TRUE),min_assault = min(Assault, na.rm = TRUE),min_murder = min(Murder,
na.rm = TRUE), min_burglary = min(Data.Rates.Property.Burglary,na.rm = TRUE), min_larceny = min(Data.Rates.Proper
ty.Larceny,na.rm = TRUE), min_motor = min(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption = " Minima fr
om different crime rates in USA (1973)") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Minima from different crime rates in USA (1973)

Region	min_rape	min_assault	min_murder	min_burglary	min_larceny	min_motor
MidWest	7.3	45	0.8	383.4	1406.0	131.4
NorthEast	7.8	48	2.1	685.0	1090.7	136.2
South	9.3	81	5.7	415.8	824.9	107.1
West	14.2	46	2.6	699.7	2237.8	207.0

The distribution of the minima grouped by region gives us a closer look as to where the largest of the relative minimum values are. In the Northeast and West regions of the USA, there are two massive spikes for minima for larceny and murder In contrast, the Northeast and Midwest have some of the lowest minimum ratios for murder.

```r
#summarize the maxima for different crime rates grouping by region and use the kbl and kable to convert it to a re
adable table.
USArrestscombinedImportant %>%
  group_by(Region) %>%
  summarise(max_rape = max(Rape, na.rm = TRUE),max_assault = max(Assault, na.rm = TRUE),max_murder = max(Murder,
na.rm = TRUE), max_burglary = max(Data.Rates.Property.Burglary,na.rm = TRUE), max_larceny = max(Data.Rates.Proper
ty.Larceny,na.rm = TRUE), max_motor = max(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption = "Maxima rat
e from different crime rates in USA (1973)") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Maxima rate from different crime rates in USA (1973)

Region	max_rape	max_assault	max_murder	max_burglary	max_larceny	max_motor
MidWest	35.1	255	12.1	1584.6	2771.3	548.3
NorthEast	26.1	254	11.1	1348.2	2312.3	1109.6
South	31.9	337	17.4	1857.2	3048.6	547.8
West	46.0	294	12.2	2149.8	3720.1	635.9

The distribution of the maxima grouped by region gives us a closer look as to where the largest of the relative maximum values are. In the South and West regions of the USA, there are two massive spikes for maxima for burglary and larceny.In contrast, the Northeast and Midwest have some of the lowest maximum ratios for murder.

```r
#summarize the standard deviations for different crime rates grouping by region and use the kbl and kable to conve
rt it to a readable table.
USArrestscombinedImportant %>%
  group_by(Region) %>%
  summarise(sd_rape = sd(Rape, na.rm = TRUE),sd_assault = sd(Assault, na.rm = TRUE),sd_murder = sd(Murder, na.rm
= TRUE), sd_burglary = sd(Data.Rates.Property.Burglary,na.rm = TRUE), sd_larceny = sd(Data.Rates.Property.Larceny
,na.rm = TRUE), sd_motor = sd(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption = "Standard Deviations fr
om different crime rates in USA (1973)") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Standard Deviations from different crime rates in USA (1973)

Region	sd_rape	sd_assault	sd_murder	sd_burglary	sd_larceny	sd_motor
MidWest	7.981736	71.53935	3.558345	337.1583	354.8975	139.8711
NorthEast	5.942806	64.85754	3.047950	250.4281	351.4370	334.4643
South	5.627536	74.20782	3.760934	332.2478	626.5774	129.6427
West	10.997774	80.32761	3.062511	480.6029	426.8271	152.4576

The distribution of the standard deviations grouped by region gives us a closer look as to the regions where this numerical value is the highest, as higher values of standard deviations lead to larger IQRs, in general. In the South and West regions of the USA, we see the highest standard deviation values in the larceny column. For the burglary column, we see the higher standard deviation values in the Midwest and West regions of the USA. In contrast, the Northeast and Midwest have some of the lowest minimum ratios for murder.

```
#summarize the variations for different crime rates grouping by region and use the kbl and kable to convert it to a readable table.
USArrestscombinedImportant %>%
  group_by(Region) %>%
  summarise(variation_rape = var(Rape, na.rm = TRUE),variation_assault = var(Assault, na.rm = TRUE),variation_murder = var(Murder, na.rm = TRUE), variation_burglary = var(Data.Rates.Property.Burglary,na.rm = TRUE), variation_larceny = var(Data.Rates.Property.Larceny,na.rm = TRUE), variation_motor = var(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption = "Variances from different crime rates in USA (1973)") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Variances from different crime rates in USA (1973)

Region	variation_rape	variation_assault	variation_murder	variation_burglary	variation_larceny	variation_motor
MidWest	63.70811	5117.879	12.661818	113675.70	125952.3	19563.93
NorthEast	35.31694	4206.500	9.290000	62714.25	123508.0	111866.37
South	31.66917	5506.800	14.144625	110388.59	392599.3	16807.23
West	120.95103	6452.526	9.378974	230979.14	182181.4	23243.33

The distribution of variance should look very similar to what was observed for the distribution of standard deviations. In the South and West regions of the USA, we see the highest standard deviation values in the larceny column. For the burglary column, we see the higher standard deviation values in the Midwest and West regions of the USA. In contrast, the Northeast and West were the regions with some of the lowest variance for murder.

```
#summarize the n_distinct values for different crime rates grouping by region and use the kbl and kable to convert it to a readable table.
USArrestscombinedImportant %>%
  group_by(Region) %>%
  summarise(Distinct_Values_rape = n_distinct(Rape, na.rm = TRUE),Distinct_Values_assault = n_distinct(Assault, na.rm = TRUE),Distinct_Values_murder = n_distinct(Murder, na.rm = TRUE), Distinct_Values_burglary = n_distinct(Data.Rates.Property.Burglary,na.rm = TRUE), Distinct_Values_larceny = n_distinct(Data.Rates.Property.Larceny,na.rm = TRUE), Distinct_Values_motor = n_distinct(Data.Rates.Property.Motor, na.rm = TRUE)) %>% kbl(caption = "Distinct Number of Samples from Different Crime Rates in USA (1973)") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Distinct Number of Samples from Different Crime Rates in USA (1973)

Region	Distinct_Values_rape	Distinct_Values_assault	Distinct_Values_murder	Distinct_Values_burglary	Distinct_Values_larceny	Distinc
MidWest	12	12	12	12	12	
NorthEast	9	9	8	9	9	
South	16	16	14	16	16	
West	13	12	13	13	13	

There seems to be a relatively symmetric distribution for the n_distinct features for the numeric ratios. This is because 2 of the ratios were NAs and I had to replace those rows with the median of the data. My data has several outliers, as seen with the burglary and larceny rates for each region, so it makes sense for me to fill in these rows with the median. I filled in these 2 values manully before processing the dataset.

```
#Need to clean our dataset for the correlation matrix algorithm to work.
#remove non-numeric columns and select columns that have rates(data of interest).
USArrestscombinedImportantCorrelation <- USArrestscombinedImportant %>%  select_if(is.numeric) %>% select(contains('Rates'))
#Copy over the Murder, Assault, and Rape columns to our new dataframe
USArrestscombinedImportantCorrelation$Murder <- USArrestscombinedImportant$Murder
USArrestscombinedImportantCorrelation$Assault <- USArrestscombinedImportant$Assault
USArrestscombinedImportantCorrelation$Rape <- USArrestscombinedImportant$Rape
#Create new data frame without the property rates, too many missing values and will create unnecessary errors when creating the heatmap.
USArrestscombinedImportantCorrelation <- USArrestscombinedImportantCorrelation%>% select(-c(Data.Rates.Property.All, Data.Rates.Violent.All)) %>% rename(
    Burglary = Data.Rates.Property.Burglary,
    Larceny = Data.Rates.Property.Larceny,
    Motor = Data.Rates.Property.Motor,
    Robbery = Data.Rates.Violent.Robbery,
    )
```

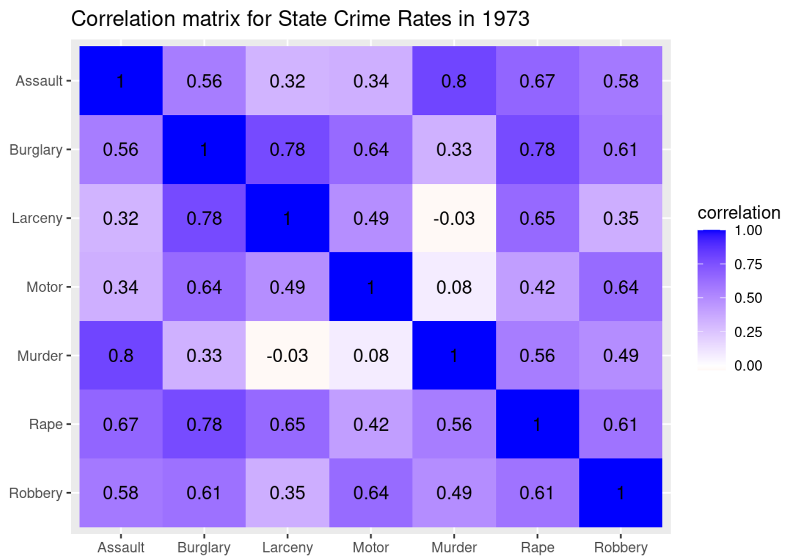
```
#summarize the rcorr values for different crime rates grouping by region and use the kbl and kable to convert it to a readable table.
library("Hmisc")
corMatrix <- rcorr(as.matrix(USArrestscombinedImportantCorrelation))
as.data.frame(corMatrix$r) %>% kbl(caption = "Correlation Matrix of Different State Crimes in 1973") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>% kable_material_dark()
```

Correlation Matrix of Different State Crimes in 1973

	Burglary	Larceny	Motor	Robbery	Murder	Assault	Rape
Burglary	1.0000000	0.7800244	0.6437819	0.6149926	0.3331508	0.5621404	0.7787774
Larceny	0.7800244	1.0000000	0.4897432	0.3459301	-0.0333869	0.3225882	0.6489864
Motor	0.6437819	0.4897432	1.0000000	0.6415893	0.0808595	0.3431231	0.4194696
Robbery	0.6149926	0.3459301	0.6415893	1.0000000	0.4921382	0.5804301	0.6056911
Murder	0.3331508	-0.0333869	0.0808595	0.4921382	1.0000000	0.8018733	0.5635788
Assault	0.5621404	0.3225882	0.3431231	0.5804301	0.8018733	1.0000000	0.6652412
Rape	0.7787774	0.6489864	0.4194696	0.6056911	0.5635788	0.6652412	1.0000000

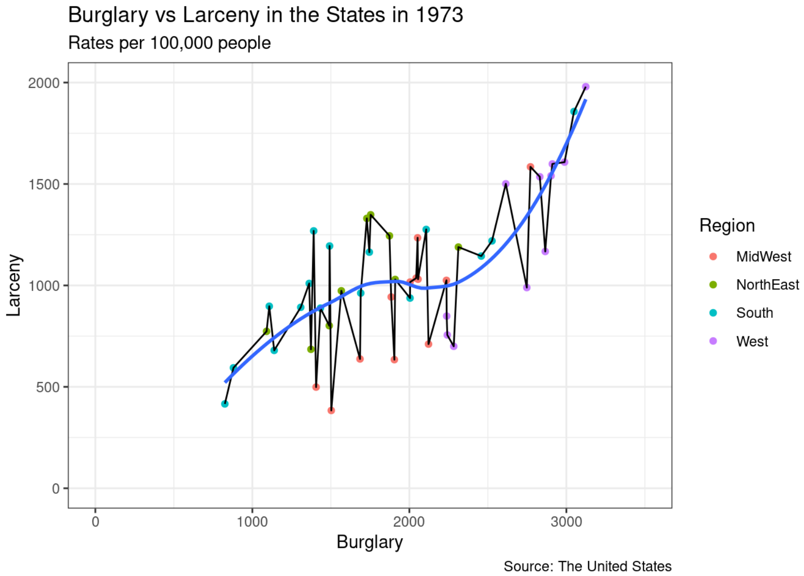
Burglary and Larceny have the highest correlation between each other. As noted from our previous analyses of the distribution of different numerical estimators, this shouldn't be a shocking revelation and matches my initial hypothesis.

```
#Cited from Dr.Guyot's Worksheet relating to pivoting on Canvas
cor(USArrestscombinedImportantCorrelation) %>%
  # Save as a data frame
  as.data.frame %>%
  # Convert row names to an explicit variable
  rownames_to_column %>%
  # Pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  # Specify variables are displayed alphabetically from top to bottom
  ggplot(aes(rowname, factor(other_var, levels = rev(levels(factor(other_var)))), fill=correlation)) +
  # Heatmap with geom_tile
  geom_tile() +
  # Change the scale to make the middle appear neutral
  scale_fill_gradient2(low="red",mid="white",high="blue") +
  # Overlay values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  # Give title and labels
  labs(title = "Correlation matrix for State Crime Rates in 1973", x = "", y = "")
```



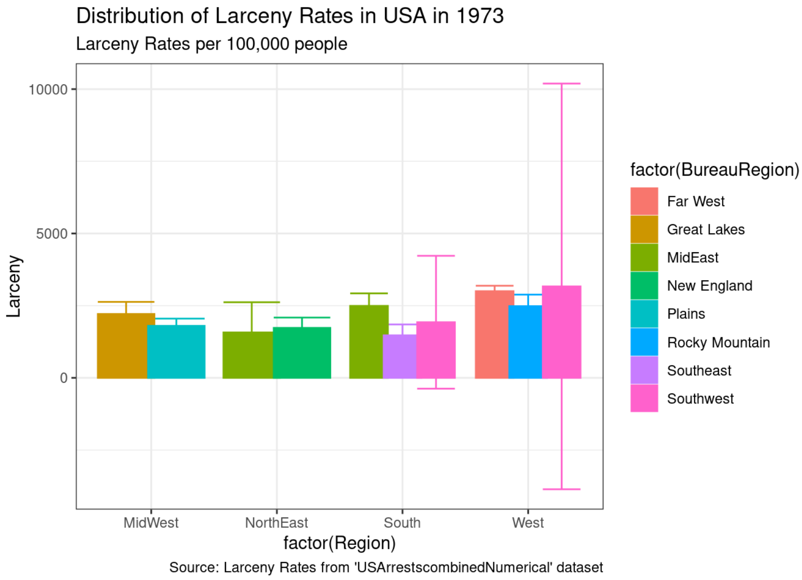
Visualizations

```
#add State and Region back to dataset to create third categorical variable (BureauRegion)
USArrestscombinedAnalysis <- USArrestscombinedImportantCorrelation
USArrestscombinedAnalysis$State <- USArrestscombinedImportant$State
USArrestscombinedAnalysis$Region <- USArrestscombinedImportant$Region
#Split states based on Bureau of Economic Analysis Regions
NewEngland <- c("Connecticut", "Maine", "Massachusetts", "New Hampshire", "Rhode Island","Vermont")
MidEast <- c("Delaware", "District of Columbia", "Maryland", "New Jersey", "New York" ,"Pennsylvania")
GreatLakes <-c("Illinois", "Indiana", "Michigan", "Ohio","Wisconsin")
Plains <- c("Iowa", "Kansas", "Minnesota", "Missouri", "Nebraska", "North Dakota", "South Dakota")
Southeast <- c( "Alabama", "Arkansas", "Florida", "Georgia", "Kentucky", "Louisiana", "Mississippi", "North Carol
ina", "South Carolina", "Tennessee", "Virginia","West Virginia")
Southwest <- c("Arizona", "New Mexico", "Oklahoma","Texas")
RockyMountain <- c("Colorado", "Idaho", "Montana", "Utah","Wyoming")
FarWest <- c("Alaska", "California", "Hawaii", "Nevada", "Oregon","Washington")
#Very similar logic structure to creating Region variable. Create BureauRegion column and possible values are Far
West, New England, Rocky Mountain, Southwest, Southeast, Plains, MidEast, and Great Lakes.
USArrestscombinedNumerical <- USArrestscombinedAnalysis %>%
  mutate(BureauRegion = case_when(State %in% NewEngland ~ "New England",
                        State %in% FarWest  ~ "Far West",
                        State %in% RockyMountain  ~ "Rocky Mountain",
                        State %in% Southwest ~ "Southwest",
                        State %in% Southeast ~ "Southeast",
                        State %in% Plains ~ "Plains",
                        State %in% MidEast ~ "MidEast",
                        State %in% GreatLakes ~ "Great Lakes",
                        )) %>% arrange(desc(BureauRegion))
#import ggplot2 package and set themes and scientific notation settings, if nay.
options(scipen=999)
library(ggplot2)
theme_set(theme_bw())
# Scatterplot
#Larceny and Burglary have the highest correlation,added visualization with difference from the medians

#Add in xlim and y limits for coordinate axes and titles
#graphing the relationship between Larceny and Burglary
gg <- ggplot(USArrestscombinedAnalysis, aes(x=Larceny, y=Burglary)) +
  geom_point(aes(col=Region)) +
  stat_summary(fun.y = median, geom='line') +
  geom_smooth(method="loess", se=F) +
  xlim(c(0, 3500)) +
  ylim(c(0, 2000)) +
  labs(subtitle="Rates per 100,000 people",
       y="Larceny",
       x="Burglary",
       title="Burglary vs Larceny in the States in 1973",
       caption = "Source: The United States")

plot(gg)
```



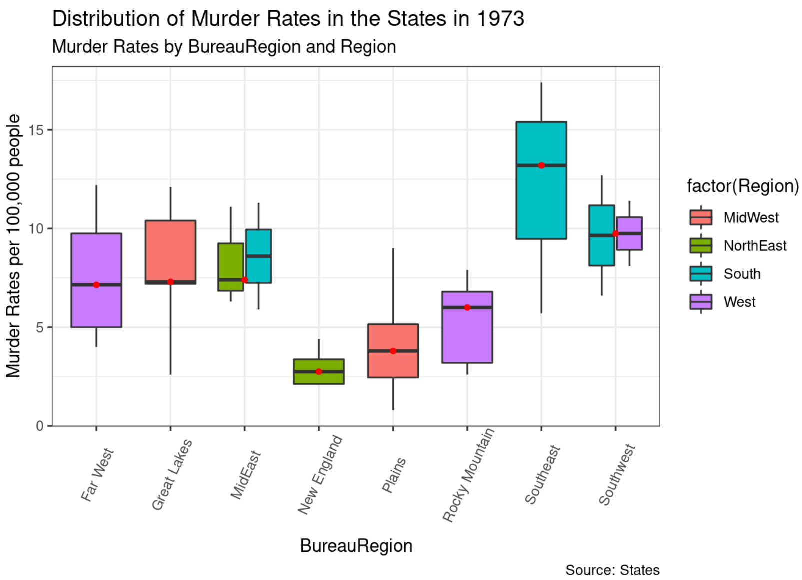
As noted by the correlation coefficient above a 0.5, the graph visually shows a rather strong and positive non-linear correlation between Larceny and Burglary. There are relatively small residuals between the means and actual data points for burglary and larceny. There are no outliers and the plot exhibits slight grouping by different regions.

```
#ggplot using Region as a factor and Larceny
#coloring points and plotting errors bars based on the mean and BureauRegion
#adding titles and theme to plot
options(scipen=999)
library(ggplot2)
ggplot(USArrestscombinedNumerical,aes(x=factor(Region),y=Larceny,colour=factor(BureauRegion),fill=factor(BureauReg
ion))) +
  stat_summary(fun.data=mean_cl_normal,position=position_dodge(0.8),geom="errorbar") +
  stat_summary(fun.y=mean,position=position_dodge(width=0.8),geom="bar")+
  labs(title="Distribution of Larceny Rates in USA in 1973",
       subtitle="Larceny Rates per 100,000 people ",
       caption="Source: Larceny Rates from 'USArrestscombinedNumerical' dataset")
```



This graph depicts a lot of variation for the larceny rates in the west region of the United States. Excluding the southwest subdivision of the western United States, most of the errors bars have shorter tails with top and bottom deviations kept at a minimum from the mean. This also entails that for the large amount of variation for larceny rates in these subregion of the United States.

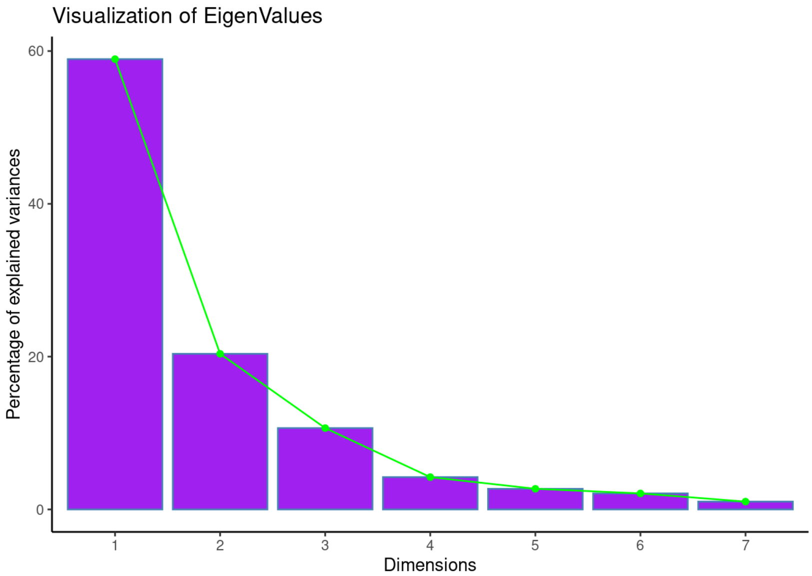
```
#ggplot using BureauRegion as a factor and Murder for my plotted variable
#coloring points and plotting errors bars based on the mean and  factor based on the region
#adding titles and theme to plot
library(ggplot2)
g <- ggplot(USArrestscombinedNumerical, aes(BureauRegion, Murder))
g + geom_boxplot(aes(fill=factor(Region)),outlier.colour = "red", outlier.shape = 1) + stat_summary(fun.y=median,
geom="point", shape=20, size=2, color="red", fill="blue") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Distribution of Murder Rates in the States in 1973",
       subtitle="Murder Rates by BureauRegion and Region",
       caption="Source: States",
       x="BureauRegion",
       y="Murder Rates per 100,000 people")
```



This boxplot shows that The SouthEast Region harbors a higher than median rate of murders, but overall the distribution seems relatively symmetric for all of the regions being described with little to no outliers.For the NewEngland sub-region, there seems to be lower than median rate of murders. The Plains, Rocky Mountains, and the New England sub-regions were relatively safe areas to settle in during the 1970's in the United States with respect to murder rates.

Dimensionality Reduction

```
library(factoextra)
library(cluster)
# Prepare data for PCA and run PCA Analysis
pca <- USArrestscombinedNumerical %>%  # Remove categorical variables
  select(-State,-Region,-BureauRegion) %>%
  # Scale to 0 mean and unit variance (standardize)
  scale() %>%
  prcomp()
percent <- 100*(pca$sdev^2/sum(pca$sdev^2))

fviz_screeplot(pca,linecolor="green",barfill = "purple") +labs(title="Visualization of EigenValues")  + theme_clas
sic()
```




```
get_eig(pca)
```

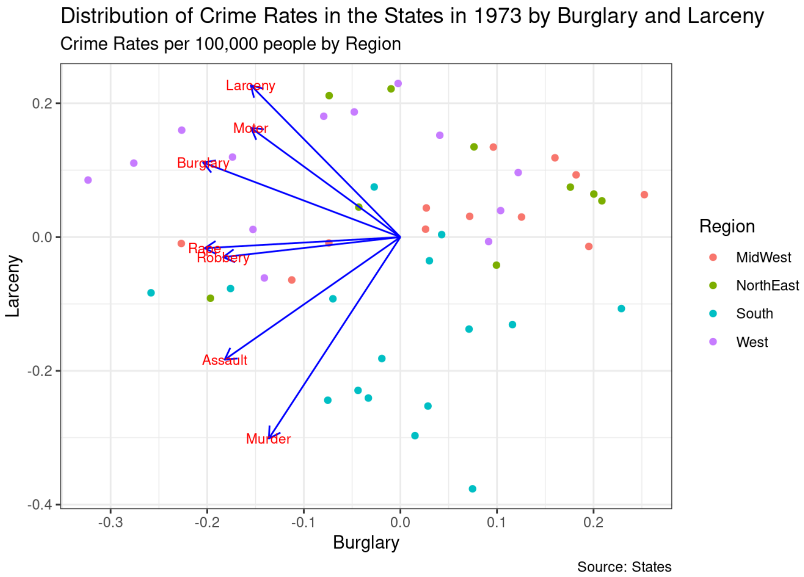
```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1 4.12605117        58.943588                    58.94359
## Dim.2 1.42593726        20.370532                    79.31412
## Dim.3 0.74537212        10.648173                    89.96229
## Dim.4 0.29582997         4.226142                    94.18844
## Dim.5 0.18888016         2.698288                    96.88672
## Dim.6 0.14648136         2.092591                    98.97931
## Dim.7 0.07144796         1.020685                   100.00000
```

```
# Visualize the rotated data
head(pca$rotation)
```

```
##                 PC1         PC2         PC3         PC4        PC5         PC6
## Burglary -0.4395793  0.23985913 -0.15106606  0.04452962 -0.2338576 -0.77362549
## Larceny  -0.3332765  0.48844419 -0.44590242  0.05661630  0.4027914  0.13040894
## Motor    -0.3328117  0.35123647  0.61158931  0.42656272 -0.3497310  0.26029218
## Robbery  -0.3949074 -0.06449804  0.52522939 -0.60102040  0.4454953 -0.06559029
## Murder   -0.2930194 -0.64788346 -0.04324300  0.03822569 -0.2669771 -0.15967935
## Assault  -0.3912314 -0.39411510 -0.08293912  0.54811233  0.4505737  0.14923559
##                  PC7
## Burglary -0.26692314
## Larceny   0.51871695
## Motor     0.14074197
## Robbery  -0.00533857
## Murder    0.62791816
## Assault  -0.39876986
```

2 principal components should be considered, as they both have eigenvalues greater than 1 and PC1 and PC2 form the crux of the elbow. Also, the cumulative proportion of variance is greater than 80%. Burglary and Larceny satisfy all three conditions of Kaiser's rule.

```
#Import necessary packages
library(ggplot2)
library(ggfortify)

# Prepare data for ggplot and include data with numerical + categorical
# variables from the PCA analysis and Specify by Region.
autoplot(pca, data = USArrestscombinedNumerical, colour = 'Region',
         loadings = TRUE, loadings.colour = 'blue',
         loadings.label = TRUE, loadings.label.size = 3) +   labs(title="Distribution of Crime Rates in the State
s in 1973 by Burglary and Larceny",
       subtitle="Crime Rates per 100,000 people by Region",
       caption="Source: States",
       x="Burglary",
       y="Larceny")
```



Burglary, Larceny, Motor, Robbery, Murder, and Assault contributed negatively to PC1.Burglary, Larceny, and Motor contributed positively to PC2, but Robbery, Murder, and Assault contributed negatively to PC2. Rape contributed negatively to both PC's.

Citations

Violent crime rates by US State. (n.d.). Retrieved March 22, 2021, from https://vincentarelbundock.github.io/Rdatasets/doc/datasets/USArrests.html
Statistics, B. (n.d.). Spreadsheets - crime & Justice electronic Data Abstracts at the Bureau of Justice Statistics. Retrieved March 22, 2021, from https://www.bjs.gov/content/dtdata.cfm#National