KARNAVATI UNIVERSITY

USCI

R PROGRAMMING PROJECT REPORT

ACADEMIC YEAR 2023-24

Project Title:

US Arrests

Students:

Sr. No.	Student Name	Enrolment No	Sem / Course
1	Rishiraj Patel	20220701047	03/B.sc

GitHub Project Link:

rishiimortal/R_project_USArrests (github.com)

Faculty: Deepti Ameeta

Dean: Dr. Raju Shanmugam

UNITEDWORLD SCHOOL OF COMPUTATIONAL INTELLIGENCE

KARNAVATI UNIVERSITY

Index

Introduction of the project

Data Analysis and Visualization of US Arrests

In this R programming project, we delve into the inclusive analysis of the "US Arrests" dataset, a dataset that summarises various crime-related statistics across different states in the United States. The dataset includes vital information such as murder rates, assault rates, urban population percentages, and rape rates for each state.

Objective:
The primary focus of this project is to conduct an investigative data analysis using the R programming language within the "RStudio" compiler. Through statistical analysis, data visualization, we aim to gain insights into relationships present in the US Arrests dataset.

Scope of Analysis:
1. Descriptive Statistics: Utilize R's statistical functions to generate descriptive statistics, including measures of central tendency and dispersion, providing a broad overview of the dataset.

2. Data Cleaning and Pre-processing: Address missing values, outliers, or any inconsistencies in the dataset to ensure the reliability of subsequent analyses.

3. Data Visualization: Leverage R's powerful visualization libraries, such as ggplot2, to create insightful graphs, charts, and maps. Visualizations will aid in understanding the geographical distribution of crime rates and potential correlations between different crime categories.

4. Inferential Statistics: Apply inferential statistical techniques to draw conclusions about the population from the dataset.

5. Clustering Analysis: Explore the possibility of grouping states based on crime profiles using analytical tools, providing a deeper understanding of insights within the dataset.

Tools and Technologies:
- R Programming Language
- RStudio Integrated Development Environment (IDE)
- Tidyverse (for data manipulation and visualization)
- ggplot2 (for advanced data visualization)
- Statistical packages in R (for inferential statistics)

Benefits:
This project serves as an excellent opportunity to enhance our proficiency in R programming, statistical analysis, and data visualization. The insights gained from this

analysis could potentially contribute to a better understanding of crime trends across different states in the United States.

By the end of this project, we aim to produce a comprehensive report that communicates our findings, methodologies, and visualizations effectively.

Aim of the project:

The project aims to enhance our proficiency in R programming, statistical analysis, and data visualization while contributing meaningful insights into the complex landscape of crime rates in the United States.

Intended outcomes of the project:

The expected outcomes of this project are multifaceted, ranging from a deeper understanding of the US Arrests dataset.

Dataset description:

In R, the inbuilt dataset that corresponds to US Arrests is also known as the "USArrests" built-in dataset. This dataset is built into the base R package and does not require additional installations. It provides information on crime rates in different states of the United States. The dataset is often used for introductory data analysis and statistical modelling exercises.

Here's a brief description of the variables in the "USArrests" dataset:

1. State:
   - Description: The name of the state.
   - Data Type: Character/String.

2. Murder:
   - Description: The murder rate per 100,000 population.
   - Data Type: Numeric (Continuous).

3. Assault:
   - Description: The rate of assaults per 100,000 population.
   - Data Type: Numeric (Continuous).

4. UrbanPop:
   - Description: The percentage of the state's population living in urban areas.
   - Data Type: Numeric (Continuous).

5. Rape:
   - Description: The rate of reported rapes per 100,000 population.
   - Data Type: Numeric (Continuous).

Usage:
You can access the "USArrests" dataset directly in R using the following command:
R
data(USArrests)

These commands provide a glimpse of the data, summary statistics, and a scatterplot matrix for exploring relationships between variables.

Proposed method

Input: head(USArrests)

```
1  head(USArrests)
```

Output:

```
> head(USArrests)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
>
```

Input: tail (USArrests)

```
1  tail(USArrests)
```

Output:

```
> tail(USArrests)
              Murder Assault UrbanPop Rape
Vermont          2.2      48       32 11.2
Virginia         8.5     156       63 20.7
Washington       4.0     145       73 26.2
West Virginia    5.7      81       39  9.3
Wisconsin        2.6      53       66 10.8
Wyoming          6.8     161       60 15.6
>
```

Input: print(USArrests)

```
1  print(USArrests)
2
3
```

Output

```
> print(USArrests)
               Murder Assault UrbanPop Rape
Alabama          13.2     236       58 21.2
Alaska           10.0     263       48 44.5
Arizona           8.1     294       80 31.0
Arkansas          8.8     190       50 19.5
California        9.0     276       91 40.6
Colorado          7.9     204       78 38.7
Connecticut       3.3     110       77 11.1
Delaware          5.9     238       72 15.8
Florida          15.4     335       80 31.9
Georgia          17.4     211       60 25.8
Hawaii            5.3      46       83 20.2
Idaho             2.6     120       54 14.2
Illinois         10.4     249       83 24.0
Indiana           7.2     113       65 21.0
Iowa              2.2      56       57 11.3
Kansas            6.0     115       66 18.0
Kentucky          9.7     109       52 16.3
Louisiana        15.4     249       66 22.2
Maine             2.1      83       51  7.8
Maryland         11.3     300       67 27.8
Massachusetts     4.4     149       85 16.3
Michigan         12.1     255       74 35.1
Minnesota         2.7      72       66 14.9
Mississippi      16.1     259       44 17.1
Missouri          9.0     178       70 28.2
Montana           6.0     109       53 16.4
Nebraska          4.3     102       62 16.5
Nevada           12.2     252       81 46.0
New Hampshire     2.1      57       56  9.5
New Jersey        7.4     159       89 18.8
New Mexico       11.4     285       70 32.1
New York         11.1     254       86 26.1
North Carolina   13.0     337       45 16.1
North Dakota      0.8      45       44  7.3
Ohio              7.3     120       75 21.4
```

Input: rownames(USArrests)

```
1  rownames(USArrests)
```

Output:

```
> rownames(USArrests)
 [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"       "California"
 [6] "Colorado"       "Connecticut"    "Delaware"       "Florida"        "Georgia"
[11] "Hawaii"         "Idaho"          "Illinois"       "Indiana"        "Iowa"
[16] "Kansas"         "Kentucky"       "Louisiana"      "Maine"          "Maryland"
[21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"    "Missouri"
[26] "Montana"        "Nebraska"       "Nevada"         "New Hampshire"  "New Jersey"
[31] "New Mexico"     "New York"       "North Carolina" "North Dakota"   "Ohio"
[36] "Oklahoma"       "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
[41] "South Dakota"   "Tennessee"      "Texas"          "Utah"           "Vermont"
[46] "Virginia"       "Washington"     "West Virginia"  "Wisconsin"      "Wyoming"
>
```

Input: ncol(USArrests)

```
1  ncol(USArrests)
```

Output:

```
> ncol(USArrests)
[1] 4
>
```

Input: dim(USArrests)

```
dim(USArrests)
```

Output:

```
> dim(USArrests)
[1] 50  4
>
```

Input: nrow(USArrests)

```
nrow(USArrests)
```

Output:

```
> nrow(USArrests)
[1] 50
>
```

Statistical analysis

Input:

mean(USArrests$Murder)

mean(USArrests$Assault)

mean(USArrests$UrbanPop)

mean(USArrests$Rape)

```
mean(USArrests$Murder)
mean(USArrests$Assault)
mean(USArrests$UrbanPop)
mean(USArrests$Rape)
```

```
> mean(USArrests$Murder)
[1] 7.788
> mean(USArrests$Assault)
[1] 170.76
> mean(USArrests$UrbanPop)
[1] 65.54
> mean(USArrests$Rape)
[1] 21.232
>
```

Input:

median(USArrests$Murder)

median(USArrests$Assault)

median(USArrests$UrbanPop)

median(USArrests$Rape)

```
median(USArrests$Murder)
median(USArrests$Assault)
median(USArrests$UrbanPop)
median(USArrests$Rape)
```

Output:

```
> median(USArrests$Murder)
[1] 7.25
> median(USArrests$Assault)
[1] 159
> median(USArrests$UrbanPop)
[1] 66
> median(USArrests$Rape)
[1] 20.1
>
```

Input:

min(USArrests$Murder)

min(USArrests$Assault)

min(USArrests$UrbanPop)

min(USArrests$Rape)

```
min(USArrests$Murder)
min(USArrests$Assault)
min(USArrests$UrbanPop)
min(USArrests$Rape)
```

Output

```
> min(USArrests$Murder)
[1] 0.8
> min(USArrests$Assault)
[1] 45
> min(USArrests$UrbanPop)
[1] 32
> min(USArrests$Rape)
[1] 7.3
```

Input :

var(USArrests$Murder)

var(USArrests$Assault)

var(USArrests$UrbanPop)

var(USArrests$Rape)

```
var(USArrests$Murder)
var(USArrests$Assault)
var(USArrests$UrbanPop)
var(USArrests$Rape)
```

Output:

```
> var(USArrests$Murder)
[1] 18.97047
> var(USArrests$Assault)
[1] 6945.166
> var(USArrests$UrbanPop)
[1] 209.5188
> var(USArrests$Rape)
[1] 87.72916
>
```

Input:

sd(USArrests$Murder)

sd(USArrests$Assault)

sd(USArrests$UrbanPop)

sd(USArrests$Rape)

```
sd(USArrests$Murder)
sd(USArrests$Assault)
sd(USArrests$UrbanPop)
sd(USArrests$Rape)
```

Output:

```
> sd(USArrests$Murder)
[1] 4.35551
> sd(USArrests$Assault)
[1] 83.33766
> sd(USArrests$UrbanPop)
[1] 14.47476
> sd(USArrests$Rape)
[1] 9.366385
>
```

Data visualization

Input:

summary(USArrests)

require(graphics)

pairs(USArrests, panel = panel.smooth, main = "USArrests data")

USArrests["Maryland", "UrbanPop"]

UA.C <- USArrests

UA.C["Maryland", "UrbanPop"] <- 76.6

s5u <- c("Colorado", "Florida", "Mississippi", "Wyoming")

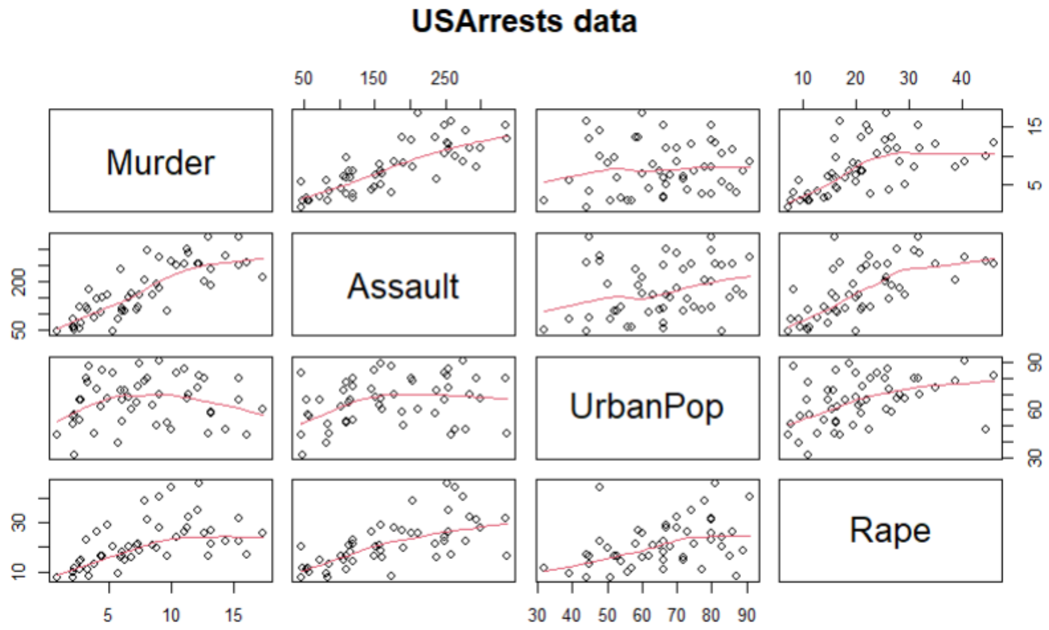
s5d <- c("Nebraska", "Pennsylvania")

UA.C[s5u, "UrbanPop"] <- UA.C[s5u, "UrbanPop"] + 0.5

UA.C[s5d, "UrbanPop"] <- UA.C[s5d, "UrbanPop"] - 0.5

```
summary(USArrests)

require(graphics)
pairs(USArrests, panel = panel.smooth, main = "USArrests data")


USArrests["Maryland", "UrbanPop"]
UA.C <- USArrests
UA.C["Maryland", "UrbanPop"] <- 76.6

s5u <- c("Colorado", "Florida", "Mississippi", "Wyoming")
s5d <- c("Nebraska", "Pennsylvania")
UA.C[s5u, "UrbanPop"] <- UA.C[s5u, "UrbanPop"] + 0.5
UA.C[s5d, "UrbanPop"] <- UA.C[s5d, "UrbanPop"] - 0.5

```

Output:



Input:

par(mfrow=c(1, 1))

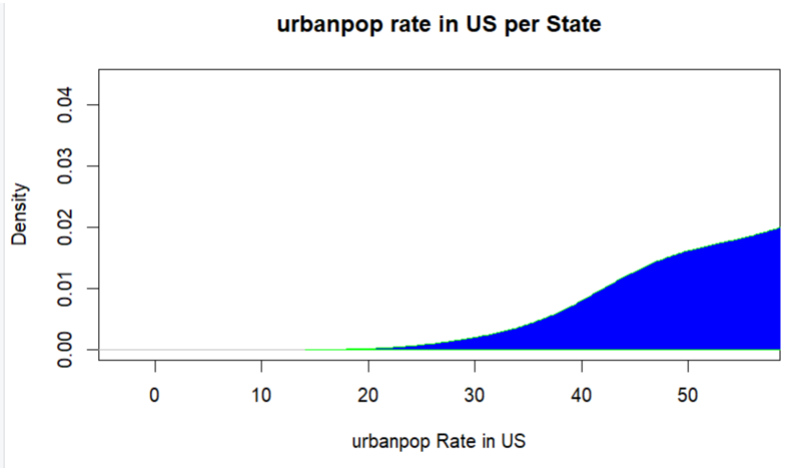
c<-density(USArrests$UrbanPop)

plot(d,type="n",main="urbanpop rate in US per State",xlab="urbanpop Rate in US")

polygon(c, col="blue", border="green")

```
par(mfrow=c(1, 1))
c<-density(USArrests$UrbanPop)
plot(d,type="n",main="urbanpop rate in US per State",xlab="urbanpop Rate in US")
polygon(c, col="blue", border="green")
```

Output:



Input:

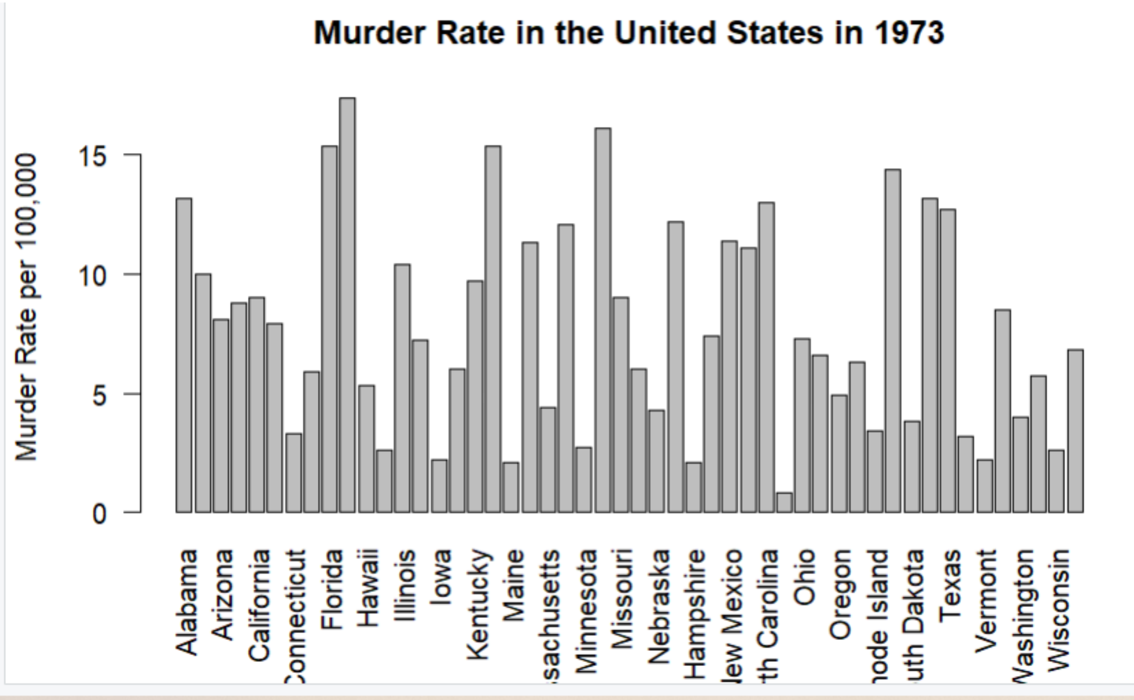
state.names = row.names(USArrests)

barplot(USArrests$Murder, names.arg = state.names, las = 2, ylab = "Murder Rate per 100,000",

main = "Murder Rate in the United States in 1973")

```
state.names = row.names(USArrests)
barplot(USArrests$Murder, names.arg = state.names, las = 2, ylab = "Murder Rate per 100,000",
        main = "Murder Rate in the United States in 1973")
```

Output :

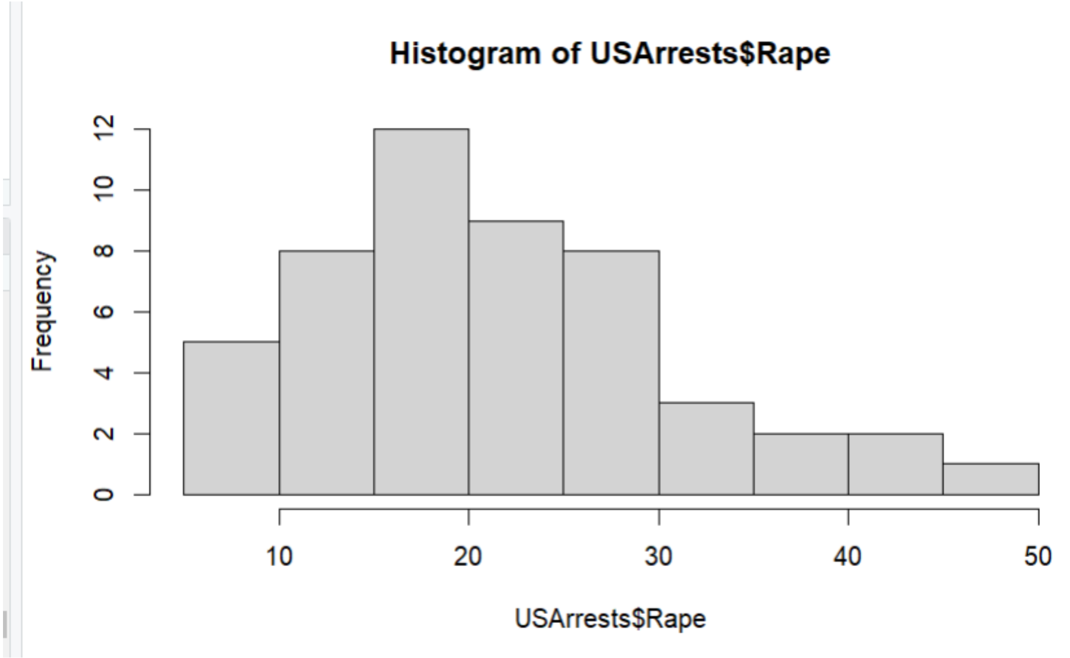


Input:

Hist(USArrests$Rape)

```
hist(USArrests$Rape)

```

Output :


Histogram of USArrests$Rape
Frequency
0
2
4
6
8
10
12
10
20
30
40
50
USArrests$Rape

Functions:

Input:

```
USArrests_data <- data.frame(
  State = c("Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado",
"Connecticut", "Delaware", "Florida", "Georgia",
          "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa", "Kansas", "Kentucky", "Louisiana",
"Maine", "Maryland",
          "Massachusetts", "Michigan", "Minnesota", "Mississippi", "Missouri", "Nebraska",
"Nevada", "New Hampshire", "New Jersey",
          "New Mexico", "New York", "North Carolina", "North Dakota", "Ohio", "Oklahoma",
"Oregon", "Pennsylvania",
          "Rhode Island", "South Carolina", "South Dakota", "Tennessee", "Texas", "Utah",
"Vermont", "Virginia",
          "Washington", "West Virginia", "Wisconsin", "Wyoming"),
  Murder = c(13.2, 10.0, 8.1, 8.8, 9.0, 7.9, 3.3, 5.9, 15.4, 17.4, 5.3, 2.6, 10.4, 7.2, 2.2, 6.0, 9.7,
15.4, 2.1,
           11.3, 4.4, 12.1, 2.7, 16.1, 9.0, 4.3, 12.2, 2.1, 7.4, 11.4, 11.1, 13.0, 0.8, 7.3, 6.6, 4.9,
6.3,
           3.4, 14.4, 3.8, 13.2, 12.7, 3.2, 2.2, 8.5, 4.0, 5.7, 2.6, 6.8),
  Assault = c(236, 263, 294, 190, 276, 204, 110, 238, 335, 211, 46, 120, 249, 113, 56, 115,
109, 249, 83, 300,
            149, 255, 72, 259, 178, 102, 252, 57, 159, 285, 254, 337, 45, 120, 151, 159, 106,
174, 279, 86,
            188, 201, 120, 48, 156, 145, 81, 53, 161),
  UrbanPop = c(58, 48, 80, 50, 91, 78, 77, 72, 80, 60, 83, 54, 83, 65, 57, 66, 52, 66, 51, 67,
85, 74, 66, 44,
             70, 62, 81, 56, 89, 70, 86, 45, 44, 75, 68, 67, 72, 87, 48, 45, 59, 80, 80, 32, 63, 73,
39, 66, 60),
  Rape = c(21.2, 44.5, 31.0, 19.5, 40.6, 38.7, 11.1, 15.8, 31.9, 25.8, 20.2, 14.2, 24.0, 21.0,
11.3, 18.0, 16.3,
         22.2, 7.8, 27.8, 16.3, 35.1, 14.9, 17.1, 28.2, 16.5, 46.0, 9.5, 18.8, 32.1, 26.1, 16.1, 7.3,
21.4, 20.0,
         29.3, 14.9, 8.3, 22.5, 12.8, 26.9, 25.5, 22.9, 11.2, 20.7, 26.2, 9.3, 10.8, 15.6)
)
USArrests_data$CrimeCategory <- ifelse(USArrests_data$Murder > 10, "High Crime", "Low
Crime")

print(USArrests_data)
```

```r

USArrests_data <- data.frame(
  State = c("Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado", "Connecticut", "Delaware", "Florida", "Georgia",
            "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa", "Kansas", "Kentucky", "Louisiana", "Maine", "Maryland",
            "Massachusetts", "Michigan", "Minnesota", "Mississippi", "Missouri", "Nebraska", "Nevada", "New Hampshire", "New Jersey",
            "New Mexico", "New York", "North Carolina", "North Dakota", "Ohio", "Oklahoma", "Oregon", "Pennsylvania",
            "Rhode Island", "South Carolina", "South Dakota", "Tennessee", "Texas", "Utah", "Vermont", "Virginia",
            "Washington", "West Virginia", "Wisconsin", "Wyoming"),
  Murder = c(13.2, 10.0, 8.1, 8.8, 9.0, 7.9, 3.3, 5.9, 15.4, 17.4, 5.3, 2.6, 10.4, 7.2, 2.2, 6.0, 9.7, 15.4, 2.1,
             11.3, 4.4, 12.1, 2.7, 16.1, 9.0, 4.3, 12.2, 2.1, 7.4, 11.4, 11.1, 13.0, 0.8, 7.3, 6.6, 4.9, 6.3,
             3.4, 14.4, 3.8, 13.2, 12.7, 3.2, 2.2, 8.5, 4.0, 5.7, 2.6, 6.8),
  Assault = c(236, 263, 294, 190, 276, 204, 110, 238, 335, 211, 46, 120, 249, 113, 56, 115, 109, 249, 83, 300,
              149, 255, 72, 259, 178, 102, 252, 57, 159, 285, 254, 337, 45, 120, 151, 159, 106, 174, 279, 86,
              188, 201, 120, 48, 156, 145, 81, 53, 161),
  UrbanPop = c(58, 48, 80, 50, 91, 78, 77, 72, 80, 60, 83, 54, 83, 65, 57, 66, 52, 66, 51, 67, 85, 74, 66, 44,
               70, 62, 81, 56, 89, 70, 86, 45, 44, 75, 68, 67, 72, 87, 48, 45, 59, 80, 80, 32, 63, 73, 39, 66, 60),
  Rape = c(21.2, 44.5, 31.0, 19.5, 40.6, 38.7, 11.1, 15.8, 31.9, 25.8, 20.2, 14.2, 24.0, 21.0, 11.3, 18.0, 16.3,
           22.2, 7.8, 27.8, 16.3, 35.1, 14.9, 17.1, 28.2, 16.5, 46.0, 9.5, 18.8, 32.1, 26.1, 16.1, 7.3, 21.4, 20.0,
           29.3, 14.9, 8.3, 22.5, 12.8, 26.9, 25.5, 22.9, 11.2, 20.7, 26.2, 9.3, 10.8, 15.6)
)


USArrests_data$CrimeCategory <- ifelse(USArrests_data$Murder > 10, "High Crime", "Low Crime")


print(USArrests_data)

```

Outcome:

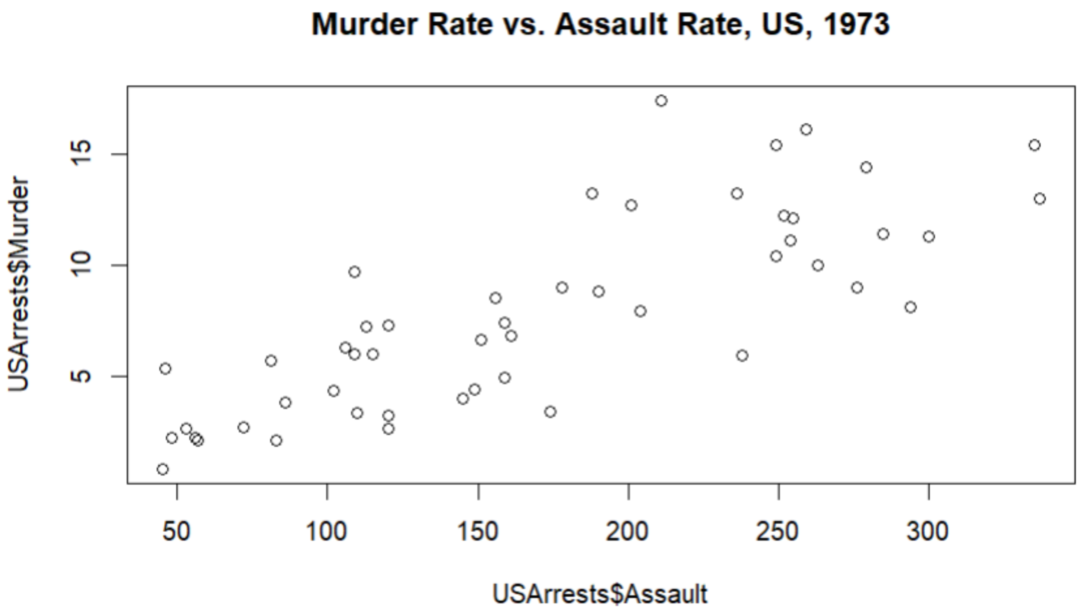
```
> print(USArrests_data)
            State Murder Assault UrbanPop Rape CrimeCategory
1         Alabama   13.2     236       58 21.2    High Crime
2          Alaska   10.0     263       48 44.5     Low Crime
3         Arizona    8.1     294       80 31.0     Low Crime
4        Arkansas    8.8     190       50 19.5     Low Crime
5      California    9.0     276       91 40.6     Low Crime
6        Colorado    7.9     204       78 38.7     Low Crime
7     Connecticut    3.3     110       77 11.1     Low Crime
8        Delaware    5.9     238       72 15.8     Low Crime
9         Florida   15.4     335       80 31.9    High Crime
10        Georgia   17.4     211       60 25.8    High Crime
11         Hawaii    5.3      46       83 20.2     Low Crime
12          Idaho    2.6     120       54 14.2     Low Crime
13       Illinois   10.4     249       83 24.0    High Crime
14        Indiana    7.2     113       65 21.0     Low Crime
15           Iowa    2.2      56       57 11.3     Low Crime
16         Kansas    6.0     115       66 18.0     Low Crime
17       Kentucky    9.7     109       52 16.3     Low Crime
18      Louisiana   15.4     249       66 22.2    High Crime
19          Maine    2.1      83       51  7.8     Low Crime
20       Maryland   11.3     300       67 27.8    High Crime
21  Massachusetts    4.4     149       85 16.3     Low Crime
22       Michigan   12.1     255       74 35.1    High Crime
23      Minnesota    2.7      72       66 14.9     Low Crime
24    Mississippi   16.1     259       44 17.1    High Crime
25       Missouri    9.0     178       70 28.2     Low Crime
26       Nebraska    4.3     102       62 16.5     Low Crime
27         Nevada   12.2     252       81 46.0    High Crime
28  New Hampshire    2.1      57       56  9.5     Low Crime
29     New Jersey    7.4     159       89 18.8     Low Crime
30     New Mexico   11.4     285       70 32.1    High Crime
31       New York   11.1     254       86 26.1    High Crime
32 North Carolina   13.0     337       45 16.1    High Crime
33   North Dakota    0.8      45       44  7.3     Low Crime
34           Ohio    7.3     120       75 21.4     Low Crime
35       Oklahoma    6.6     151       68 20.0     Low Crime
36         Oregon    4.9     159       67 29.3     Low Crime
37   Pennsylvania    6.3     106       72 14.9     Low Crime
38   Rhode Island    3.4     174       87  8.3     Low Crime
39 South Carolina   14.4     279       48 22.5    High Crime
```

Input:
plot(y = USArrests$Murder, x = USArrests$Assault, main = "Murder Rate vs. Assault Rate, US, 1973")

Source on Save

```r
plot(y = USArrests$Murder, x = USArrests$Assault, main = "Murder Rate vs. Assault Rate, US, 1973")
```

Output:



Conclusion:

In conclusion, the examination of the "USArrests" dataset has offered valued insights into the patterns of crime across different states in the United States. Through examining data, we have identified notable variations in arrest rates and crime types, shedding light on potential factors influencing these disparities.

Additionally, the visualization of the dataset has provided a compelling way to comprehend geographical patterns and outliers. Identifying states with unusually high or low arrest rates prompts a deeper examination of the unique circumstances contributing to these deviations.

While this analysis has offered valuable insights, it is crucial to acknowledge the limitations of the dataset.