Assignments

Niko Amber Assignments!

Assignment 1

Collaborators: Eliza Epstein.

Problem 1

Install the datasets package on the console below using `install.packages("datasets")`. Now load the library.

Answer: I've loaded the library!

Load the USArrests dataset and rename it `dat`. Note that this dataset comes with R, in the package datasets, so there's no need to load data from your computer. Why is it useful to rename the dataset?

```
dat <- USArrests
```

Answer: It is beneficial to rename the data set, so we can replicate analyses without disturbing the original data set. Additionally, it is nice to rename your data set to know exactly it is called

Problem 2

First I am making states lowercarse, to be used as variables later.

```
dat$state <- tolower(rownames(USArrests))
```

This dataset has the state names as row names, so we just want to make them into a new variable. We also make them all lower case, because that will help us draw a map later - the map function requires the states to be lower case.

List the variables contained in the dataset `USArrests`.

The variables contained in the dataset 'USArrests' are Murder, Assault, and Rape. Additionally, the data set shows us what percentage of people live in urban areas.

```
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
## Oklahoma          6.6     151       68 20.0
## Oregon            4.9     159       67 29.3
## Pennsylvania      6.3     106       72 14.9
## Rhode Island      3.4     174       87  8.3
## South Carolina   14.4     279       48 22.5
## South Dakota      3.8      86       45 12.8
## Tennessee        13.2     188       59 26.9
## Texas            12.7     201       80 25.5
## Utah              3.2     120       80 22.9
## 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
```

Problem 3

What type of variable (from the DVB chapter) is `Murder`?

Answer: Murder is a categorical variable, it is one of many catagories that crime falls into. It is not ordinal, there is no ordering of crime.

What R Type of variable is it?

Answer: 'Murder' is a character variable, it contains information that isn't numeric.

Problem 4

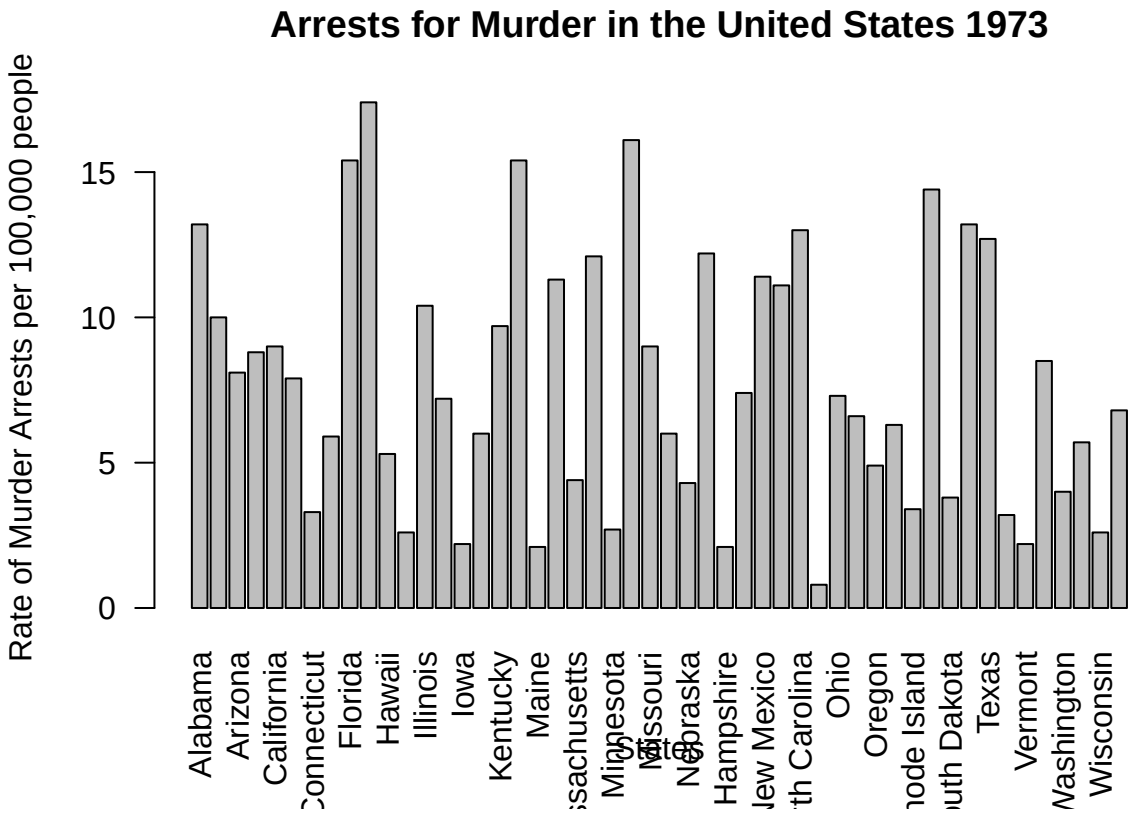
What information is contained in this dataset, in general? What do the numbers mean?

Answer: The Data set USArrests contains data about the rate of arrests for murder, rape and assault per 100,000 residents in each US state in 1973. The data set also includes the percent of the population living in urban cities in each state. The collums represent each type or crime (and urban population percentage) The numbers in each row are the rate of arrests per state (per 100,000).

Problem 5

Draw a (histogram) bar graph of `Murder` with proper labels and title. I used a bar graph instead of a histogram because I feel that is a better way to represent this data.

```
barplot(USArrests$Murder , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Murder Arrest
        main = "Arrests for Murder in the United States 1973")
```



Problem 6

Please summarize `Murder` quantitatively. What are its mean and median? What is the difference between mean and median? What is a quartile, and why do you think R gives you the 1st Qu. and 3rd Qu.?

I used the summary function to gather this information

```
summary(USArrests$Murder)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

The mean for murder is 7.788 meaning this is the average rate of people (per 100,000) murdered in each US state in 1973. The median for murder is 7.250 meaning in the United States in 1973 half of the states had a rate per 100,000 more that 7.250 and half had fewer. Median is the middle of numbers in a data set, while mean is the average of all numbers in said set. If the data is evenly distributed the median will equal the mean.

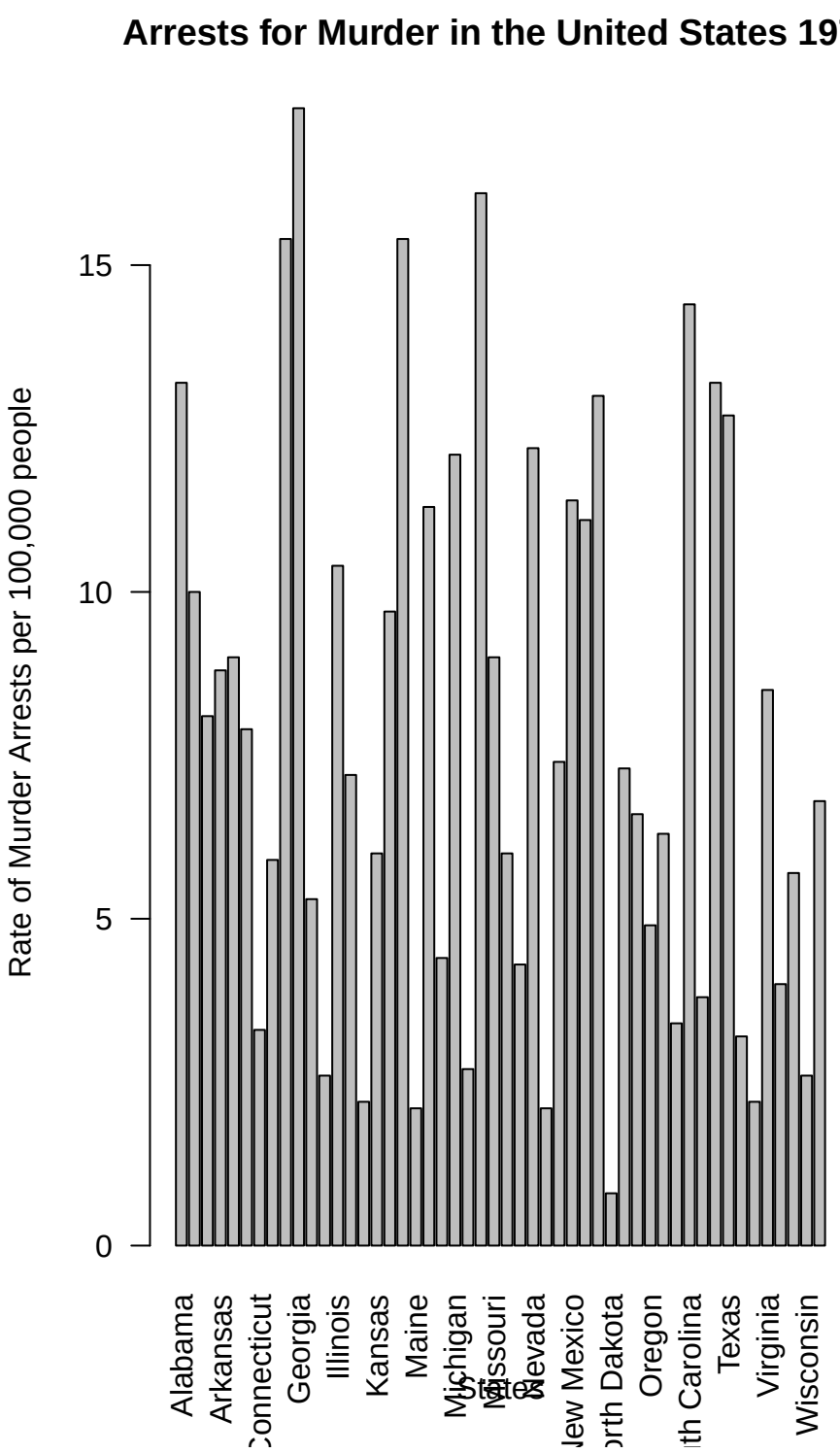
A quartile is when the data is divided into four equal parts: the 1st, 2nd, 3rd, and 4th quartile. R gives the 1st and 3rd quartile because it represents the middle of the data. 1st quartile is lowest 25% and 3rd quartile is highest 25%.

Problem 7

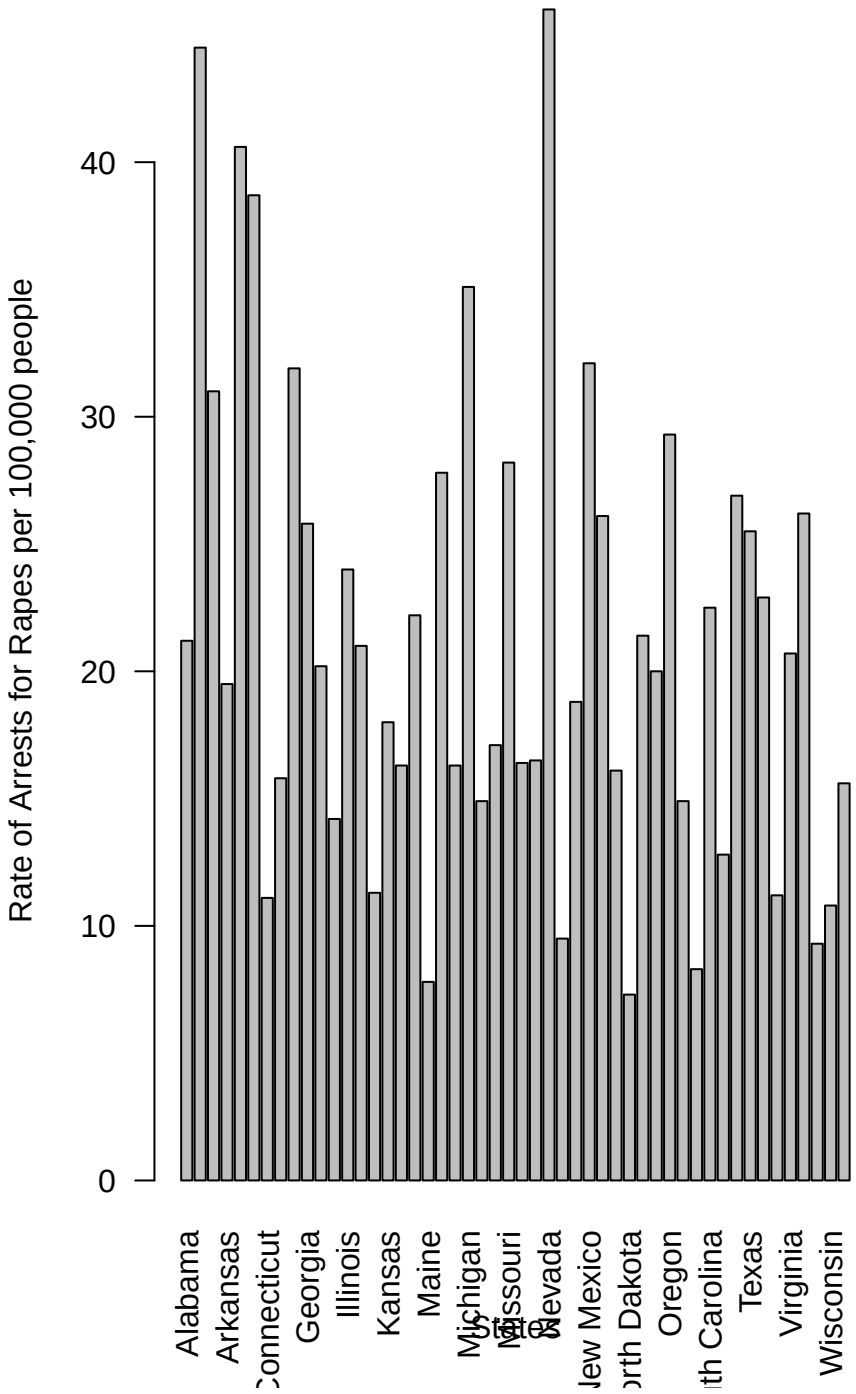
Repeat the same steps you followed for `Murder`, for the variables `Assault` and `Rape`. Now plot all three histograms together. You can do this by using the command `par(mfrow=c(3,1))` and then plotting each of the three.

Note: I used bar graphs

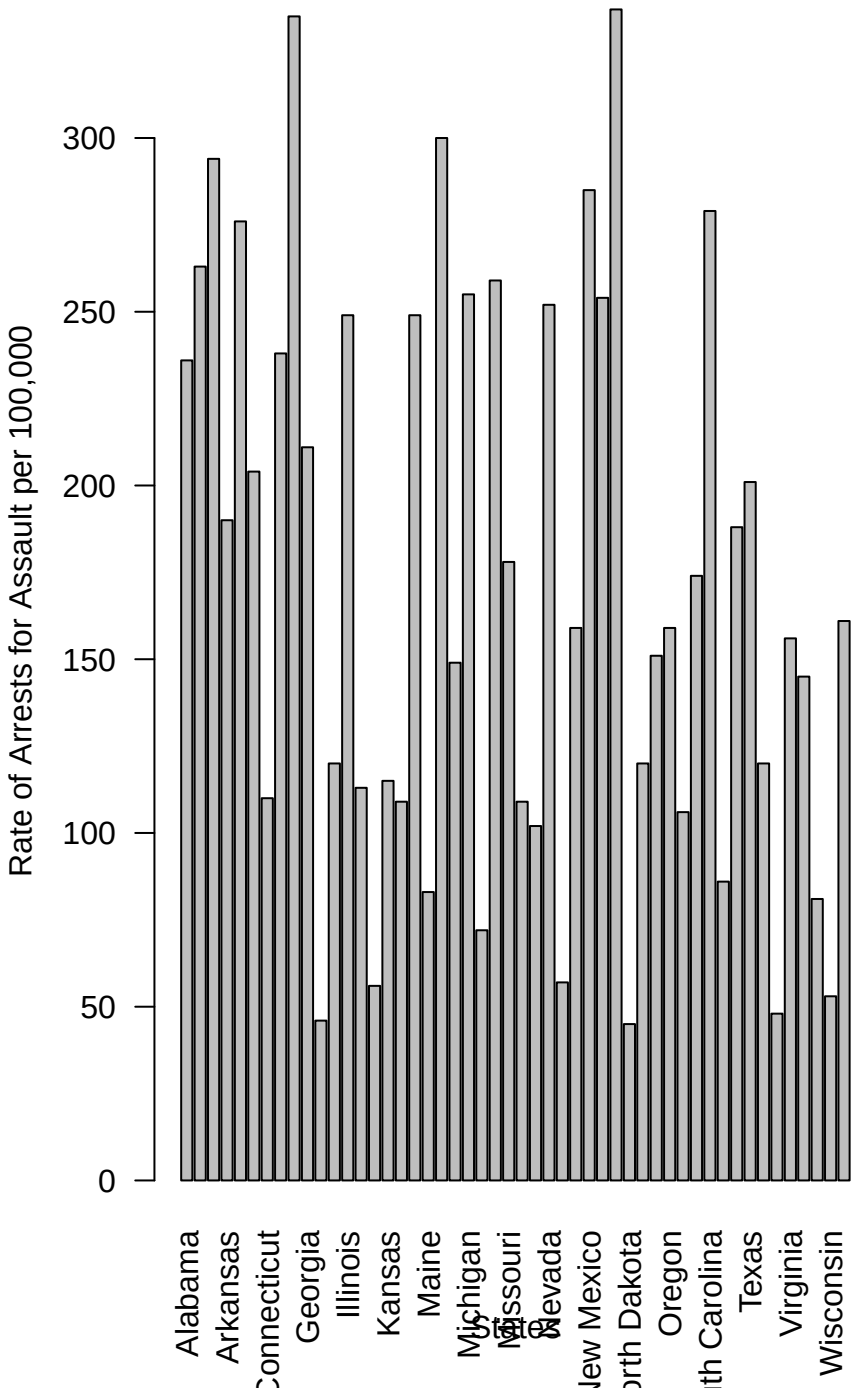
```
barplot(USArrests$Murder , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Murder Arrest
        main = "Arrests for Murder in the United States 1973")
```




```
#Bar plot for rape arrests
barplot(USArrests$Rape , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Arrests for Rape
        main = "Rate of Arrests for Rape in the United States in 1973")
```




```r
#Bar plot for assault arrests
barplot(USArrests$Assault , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Arrests for A
        main = "Rate of Arrests for Assaults in the United States")
```

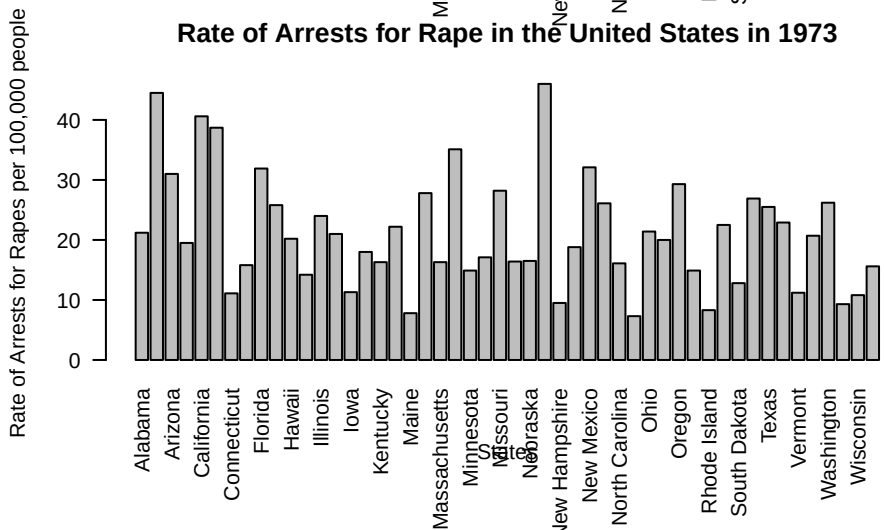


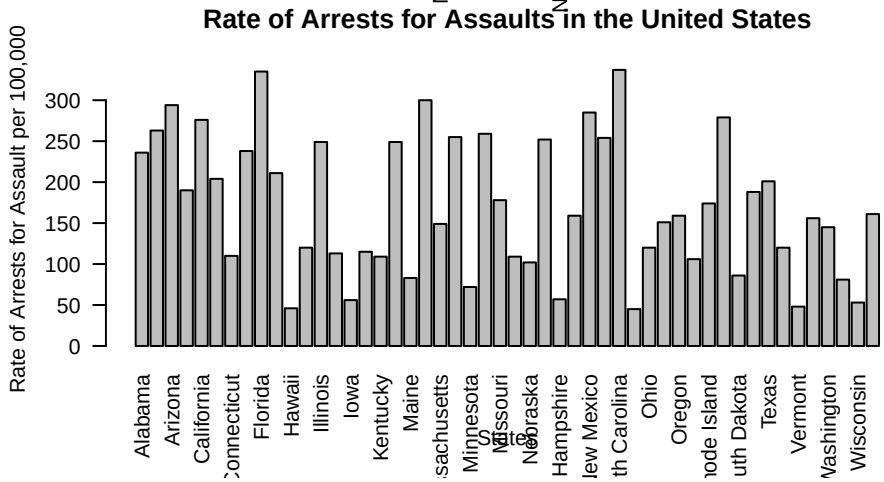

```
#plotting the graphs together

par(mfrow=c(3,1))
barplot(USArrests$Murder , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Murder Arrests
        main = "Arrests for Murder in the United States 1973")
```

```r
#Bar plot for rape arrests
barplot(USArrests$Rape , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Arrests for Rape
        main = "Rate of Arrests for Rape in the United States in 1973")

#Bar plot for assault arrests
barplot(USArrests$Assault , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Arrests for A
        main = "Rate of Arrests for Assaults in the United States")
```






```
par
```

```
## function (..., no.readonly = FALSE)
## {
##     .Pars.readonly <- c("cin", "cra", "csi", "cxy", "din", "page")
```

```
##     single <- FALSE
##     args <- list(...)
##     if (!length(args))
##         args <- as.list(if (no.readonly)
##             .Pars[-match(.Pars.readonly, .Pars)]
##         else .Pars)
##     else {
##         if (all(unlist(lapply(args, is.character))))
##             args <- as.list(unlist(args))
##         if (length(args) == 1) {
##             if (is.list(args[[1L]]) || is.null(args[[1L]]))
##                 args <- args[[1L]]
##             else if (is.null(names(args)))
##                 single <- TRUE
##         }
##     }
##     value <- .External2(C_par, args)
##     if (single)
##         value <- value[[1L]]
##     if (!is.null(names(args)))
##         invisible(value)
##     else value
## }
## <bytecode: 0x7faeb98ae760>
## <environment: namespace:graphics>
```

What does the command par do, in your own words (you can look this up by asking R `?par`)?

Answer: This command allows R to set paramaters, this way multiple data sets can be graphed together.

What can you learn from bar graphs the histograms together?

Answer: When we plot these bar graphs together we can compare each state's arrest rates for different crimes. For example, when looking at the bar graphs it is easy to see that North Carolina's arrest rate for assault is much higher than the arrest rate for rape. This can lead researchers to ask questions: why were there more arrests for assaults? Were there possible sexually based assault that should have been charged as rape?

Problem 8

In the console below (not in text), type `install.packages("maps")` and press Enter, and then type `install.packages("ggplot2")` and press Enter. This will install the packages so you can load the libraries.

Run this code:

```
library('maps')
library('ggplot2')

#this code creates a map
ggplot(dat, aes(map_id=state, fill=Murder)) +
  geom_map(map=map_data("state")) +
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat)
```

What does this code do? Explain what each line is doing.

Answer: This code creates a colored map that uses our data set to show murder rates. The lighter the blue, the higher the murder rate.

The first line of code creates a map that is divided by states. TYhe first line also applies the data set to the map, so the color of the state will be lighter if murder rates are higher. The second line of code designs the map by breaking it up by state. The last line of code expands the limits of the map

Assignment 2

(Coming soon)