STAT1003 Introduction to Data Science

Solution: S1 2019 Test 1

Instructions

1. Save this file to `I:\STAT1003` and rename it `StudentID_S1_2019_STAT1003_Test_1.Rmd`;
2. Download and save the file `S1_2019_STAT1003_Test_1.RData` into the same directory; it contains two data frames, `USCrime` and `USGeogData`;
3. Carry out all of your work in this R Markdown file;
4. **Save your work frequently!**
5. When you are finished, upload this Rmarkdown file and the resulting 'knitted' Word file to the Assessments section from which you downloaded the files. **Please close the Word document before you upload it to BB.**
6. Total number of marks: **67**

Question 1 Short Answer Questions

a. What are some of the characteristics of contemporary data science that distinguish it from applied statistics? (5 marks)

There are lots of characteristics that distinguish contemporary data science from 'conventional' applied statistics:

1. it sits at the intersection of different domains, e.g., statistics, computer science, visualization and communication;
2. defining features include very large datasets; 1 x 5
3. data that is passively measured or observed;
4. a focus on predictive rather than generative models;
5. dynamic analysis and visualization of data, etc.

b. The article by Cukier and Mayer-Schoenberger (2013) "The Rise of Big Data: How It's Changing the Way We Think About the World" introduces the term "datafication". Briefly describe what "datafication" means, and list three examples that you can think of yourself (not from the article itself!) and why they constitute datafication. (5 marks) 1

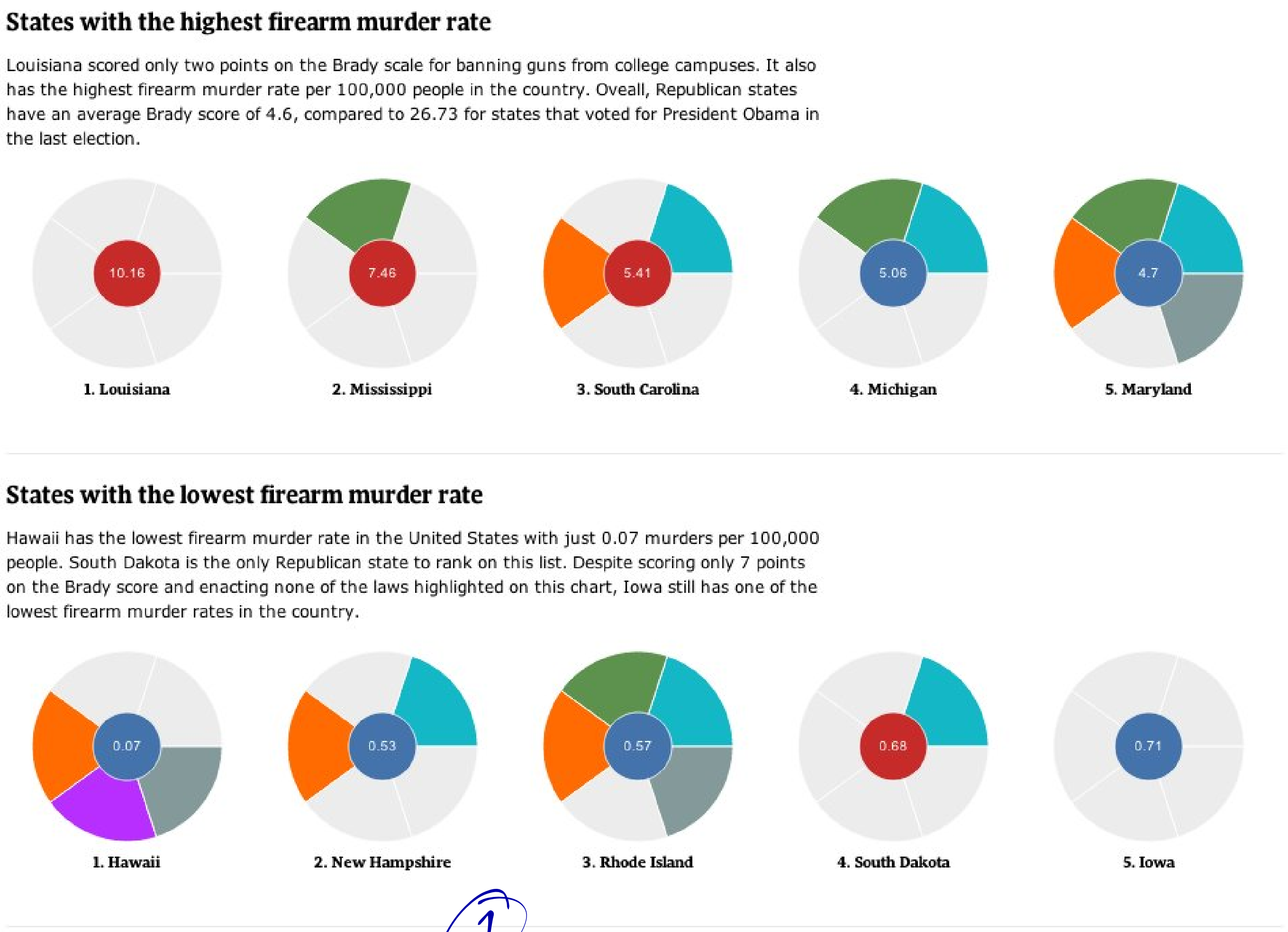
1 "Datafication" can be defined as the increasing trend towards taking many aspects of our lives and turning them into data that can be used by individuals to their own benefit, or by external agencies to the detriment of the individual. There are lots of examples of "datafication", and I was looking for sensible examples 1/2 with some brief explanation 1/2 of why they constitute datafication, e.g.,

1. Using BB accesses to predict student 1/2 performance (frequency and timing of BB accesses is data); 1/2
2. Using wifi logins and signals to estimate room usage at Curtin (logins and signal strength as data for tracking individuals); 1/2
3. Now that everyone is using cards or other forms of electronic payment for even the smallest transactions, banks are using their customers' retail transaction history to determine whether they should be, for example, given loans (automatic collection of financial data that is being interpreted in a way that an individual may not have intended). 1/2

(anything sensible is acceptable)

Question 2 Questionable graphics

Briefly outline why the two graphics below might be either confusing, misleading or just plain uninformative. (5 marks)



States with the highest firearm murder rate

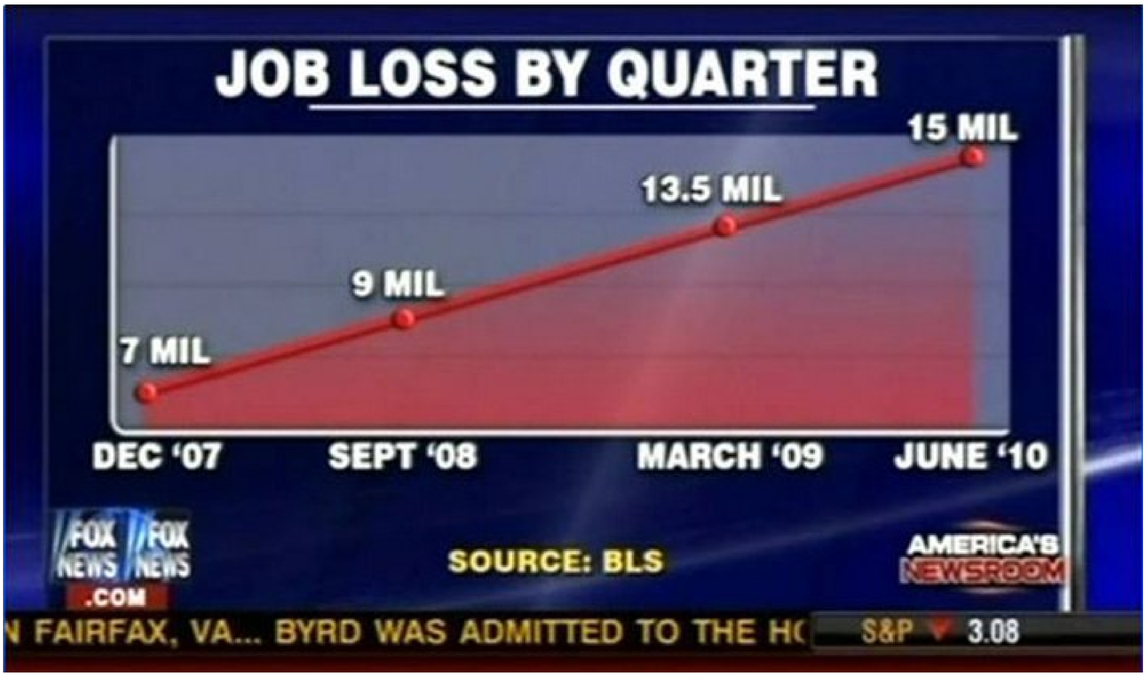
Louisiana scored only two points on the Brady scale for banning guns from college campuses. It also has the highest firearm murder rate per 100,000 people in the country. Oveall, Republican states have an average Brady score of 4.6, compared to 26.73 for states that voted for President Obama in the last election.

States with the lowest firearm murder rate

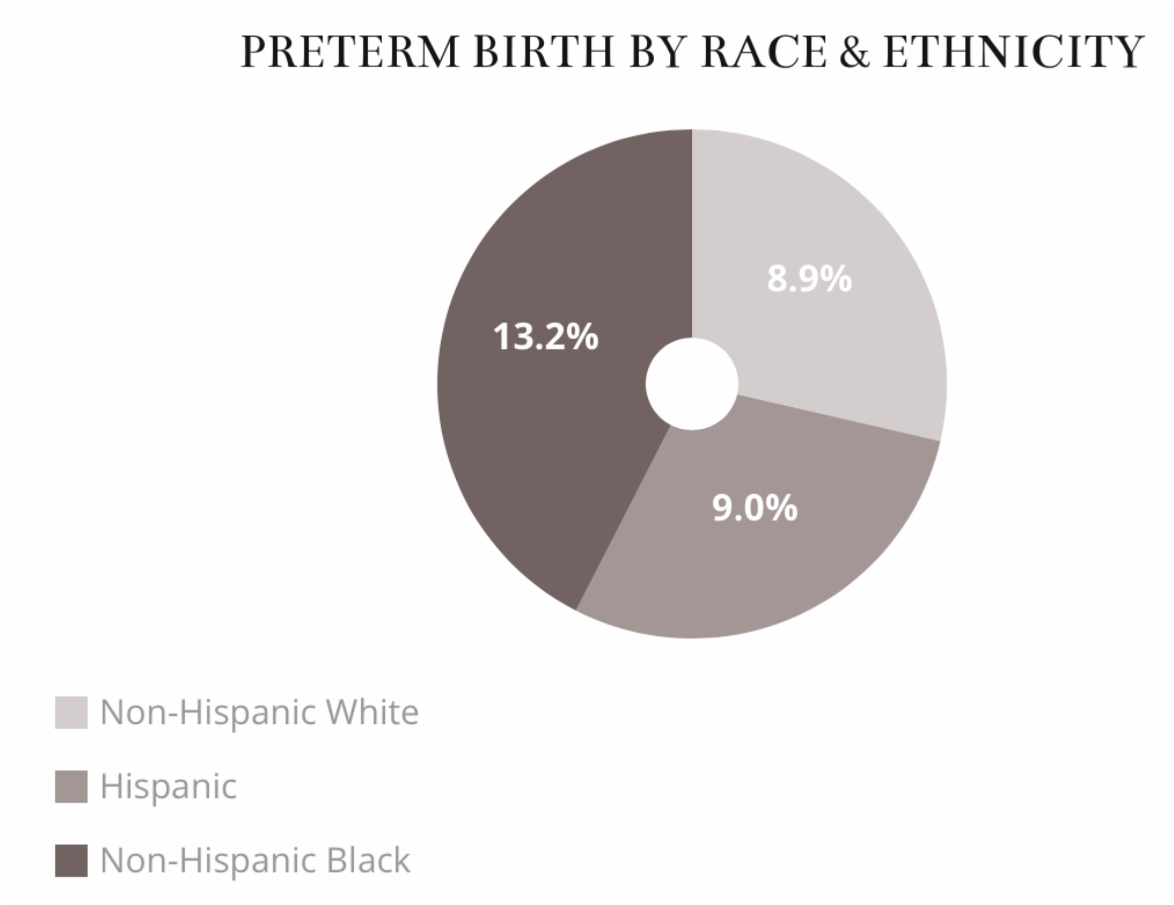
Hawaii has the lowest firearm murder rate in the United States with just 0.07 murders per 100,000 people. South Dakota is the only Republican state to rank on this list. Despite scoring only 7 points on the Brady score and enacting none of the laws highlighted on this chart, Iowa still has one of the lowest firearm murder rates in the country.

1

a. Confusing because these are not pie charts!! They're simply showing the firearm murder rates in the central circle. Also, it's not clear why the central circle is either red or blue. 1



b. The horizontal scale is not linear, so there is not a linear increase in jobless rate! 1 1



c. We might expect the numbers in the pie chart to add up to 100%, but they don't. 1

Question 3 Crime in the US

The US Federal Bureau of Investigation (FBI) publishes estimates of crime statistics for each US state and the District of Columbia, and they are available from 1960 onwards. The data frame `USCrime` contains crime statistics in each state—except the District of Columbia—for the period 1990–2014 and includes both violent crime and property crime. The variables in `USCrime` include the following

- `Abbr` : a two-letter abbreviation of the name of the state;
- `Div` : the name of the geographical division (e.g., New England, Middle Atlantic, etc.) that the state has been assigned;
- `State` : the full name of the state;
- `Year` : the year for which the data have been collected;
- `Population` : the estimated population of the state;
- `TotVCrime` : the total number of violent crimes recorded—includes `Murder` , either `LRape1` (or, where available, `RRape`), `Robbery` , and `AggAssault` ;
- `Murder` : the number of murders and nonnegligent homicides;
- `LRape` : the number of recorded rapes according to a 'legacy' definition that included only female victims;
- `Rrape` : (after 2013 only) the number of recorded rapes according to a new definition that included both males and females and new categories;
- `Robbery` : the number of robberies, defined as "the taking or attempting to take anything of value … by force …";
- `AggAssault` : the number of aggravated assaults;
- `TotPCrime` : the total number of property crimes—includes `Burglary` , `LarcenyTheft` , and `MVTheft` ;
- `Burglary` : the number of burglaries, defined as "the unlawful entry of a structure to commit a felony or theft";
- `LarcenyTheft` : the number of crimes under the rubric 'Larceny-Theft', which can be briefly described as robbery without

violence; includes embezzlement, forgery, etc.;

- `MVTheft`: the number of motor vehicle thefts or attempted thefts;

In addition to `USCrime` and additional data frame, `USGeogData`, is also available. It contains the variables `Abbr`, which contains the same two-letter state abbreviations as in `USCrime`, `Area`, the area in square miles of the state; and `Lat` and `Long`, which provide the latitude and longitude values of the approximate geographical center of each state. Longitude is negative, because the US is west of the prime meridian at Greenwich.

Make sure that all plots are correctly labelled.

```
# Load the data here by uncommenting the statement below:

print(load("S1_2019_STAT1003_Test_1.RData"))
```

```
[1] "USCrime"    "USGeogData"
```

a. Use an appropriate *R* command to determine the number of rows there are in `USCrime`? What does each row represent? (2 marks)

```
nrow(USCrime)
```

1 or dim ( )

```
[1] 1250
```

1

There are 1250 rows in `USCrime`. Each row represents the crime statistics for a US state in the years 1990 to 2014.

b. Which of the variables in `USCrime` are categorical, and which are quantitative? You could go through the variables one-by-one and decide which are categorical and which are quantitative, but a single *R* command will tell you that information. (2 marks)

The function `str` provides information about the class of each of the variables in the data frame:

```
str(USCrime)
```

1

```
'data.frame':   1250 obs. of  15 variables:
 $ Abbr        : Factor w/ 50 levels "AK","AL","AR",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Div         : Factor w/ 9 levels "New England",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ State       : Factor w/ 50 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Year        : int  1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 ...
 $ Population  : int  4040587 4089000 4136000 4187000 4219000 4253000 4273000 4319000 4352000 436
9862 ...
 $ TotVCrime   : int  28630 34518 36052 32676 28844 26894 24159 24379 22286 21421 ...
 $ Murder      : int  467 469 455 484 501 475 444 426 354 345 ...
 $ Lrape       : int  1319 1455 1704 1471 1487 1350 1397 1396 1443 1513 ...
 $ Rrape       : int  NA NA NA NA NA NA NA NA NA NA ...
 $ Robbery     : int  5805 6246 6819 6677 7223 7900 7124 6931 5698 5297 ...
 $ AggAssault  : int  21039 26348 27074 24044 19633 17169 15194 15626 14791 14266 ...
 $ TotPCrime   : int  169974 184882 181837 171598 178015 179294 181803 186809 177779 171398 ...
 $ Burglary    : int  44585 51873 49053 45578 44064 43586 42821 43786 41965 38648 ...
 $ LarcenyTheft: int  111336 118151 117801 111878 119951 120967 123350 127616 120943 119616 ...
 $ MVTheft     : int  14053 14858 14983 14142 14000 14741 15632 15407 14871 13134 ...
```

We can see that the variables `Abbr`, `Div`, and `State` are categorical, and the remainder are quantitative. 1

c. How many rows are there in `USGeogData`? What does each row represent? (2 marks)

```
nrow(USGeogData)
```

1 or dim ( )

```
[1] 50
```

1

There are 50 rows in `USGeogData`. Each row represents geographical data for each of the 50 US states.

d. Use R commands to identify which state has the largest area. What is its area? (2 marks)

The function `which.max()` can be used to identify the row of `USGeogData` containing the state that has the largest area, as follows:

```
USGeogData[which.max(USGeogData$Area), ]
```

1

```
  Abbr   Area    Long   Lat
2   AK 589757 -127.25 49.25
```

The state that has the largest area is Alaska; its area is 589757 square miles. 1

e. Using the function `subset`, create a new dataset (call it `California`) containing only the crime data for the state of California. Check to make sure that it has the appropriate number of rows and columns. (2 marks)

The function `subset` is easy to use, but I was also hoping you would check the number of rows and columns using the function `dim`, or another equivalent. Typing it out is not useful—what if it were *really* large?

```
California <- subset(USCrime, subset = (Abbr == "CA"))
dim(California)
```

1

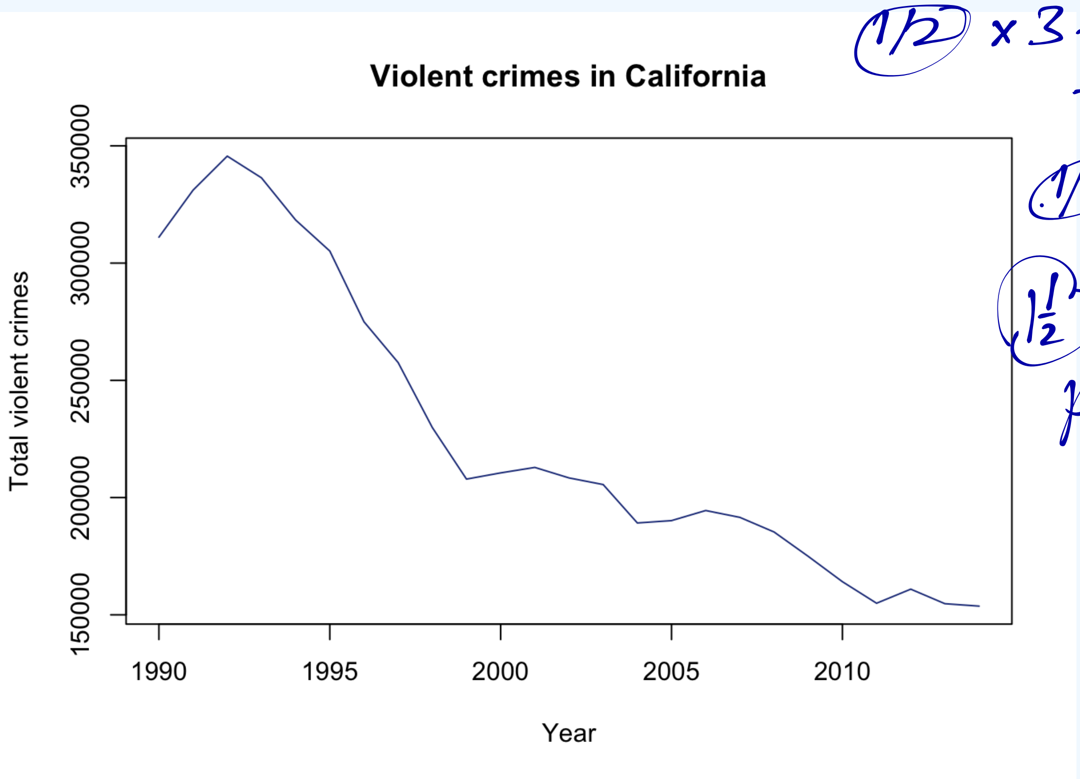
1

```
[1] 25 15
```

0 → data from 1990-2014

f. For California, construct separate plots of the total number of violent and property crimes. Label them appropriately, include a main title, and construct a line plot in a colour different from the default black. Comment on any trend you notice. (8 marks)

```
plot(TotVCrime ~ Year, data = California, xlab = "Year", ylab = "Total violent crimes",
    main = "Violent crimes in California", type = "l", col = "royalblue4")
```

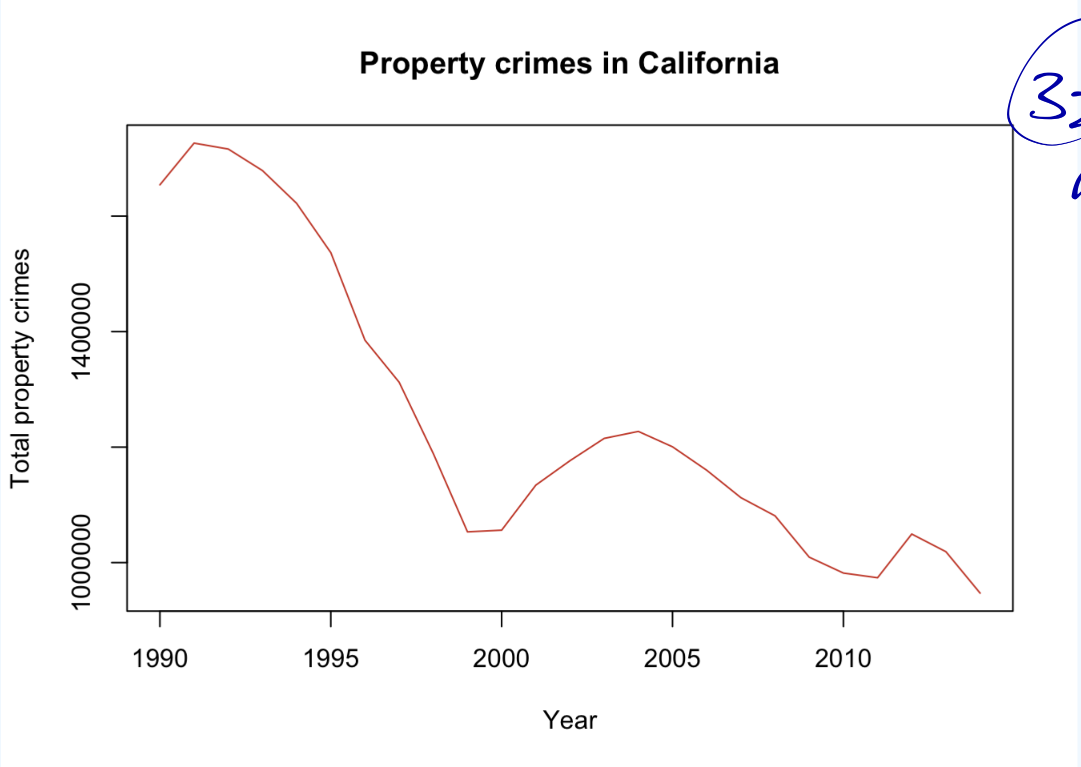


1/2 x 3 for labels & title

1/2 for colour

$1\frac{1}{2}$ for correct plot

```
plot(TotPCrime ~ Year, data = California, xlab = "Year", ylab = "Total property crimes",
    main = "Property crimes in California", type = "l", col = "tomato3")
```



3½ - breakdown as above

Clearly the number of violent and property crimes has declined, but this decline is sharpest throughout the 1990s. 1

g. In wanting to compare the crime rate across states, it makes sense to normalize crime statistics by the state population. Why? (1 mark)

We would expect states with larger populations to have a greater number of crimes, so normalizing by population provides us with the crime *rate*. 1

h. Construct an appropriately labelled plot that compares the (total) violent crime rate over time (total number of violent crimes per 100000 residents) between California and Alabama. Note that you will probably need several steps to construct such a plot, for example,

   a. extract data for Alabama;
   b. calculate total violent crime rate per 100000 people for the two states;
   c. plot them on the same plot against time.

Show those steps so that you get part marks even if you don't succeed in producing the required plot. A legend is *not* required, but a main title is. Choose a red line for Alabama and a blue one for California. Comment on any differences you see. (9 marks)

The first few steps are easy:

```
Alabama <- subset(USCrime, subset = (State == "Alabama"))
dim(Alabama)
```

1

```
[1] 25 15
```

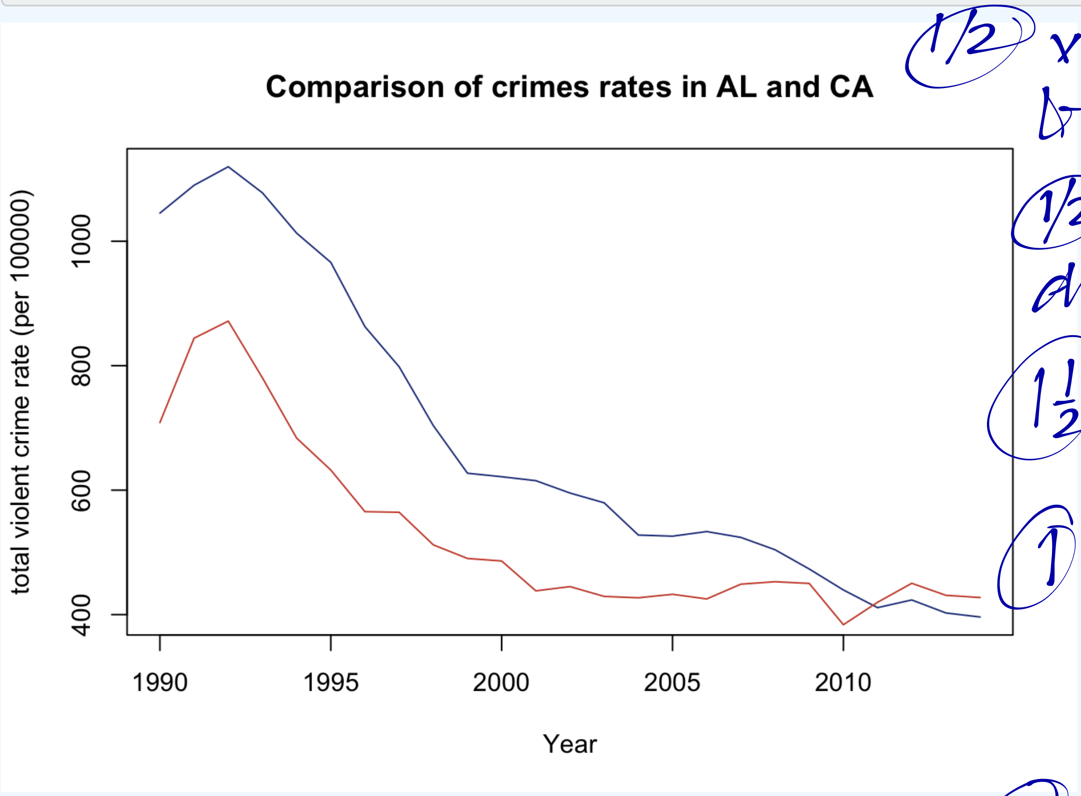
½ ½

```
CACrimeRate <- 1e+05 * California$TotVCrime/California$Population
ALCrimeRate <- 1e+05 * Alabama$TotVCrime/Alabama$Population
```

½ ½

There are several ways to plot two or more lines on a plot: first, by creating a plot and then adding another line to it using the function `lines`, or second, by combining the two sets of crime rates into a matrix or data frame, and then using the function `matplot`. Here, I show the former:

```
plot(CACrimeRate ~ Year, data = California, col = "royalblue4", type = "l",
    ylab = "total violent crime rate (per 100000)", main = "Comparison of crimes rates in AL and
CA")
lines(1990:2014, ALCrimeRate, col = "tomato3")
```



1/2 x 3 for labels & title

1/2 x 2 for different colours

$1\frac{1}{2}$ for the state

1 for second state

We can see that after about 2010, the crime rate in Alabama is higher than that of California. 1

i. Construct a dataset that consists of the crime data for all states for the year 2014. Call it `USCrime2014`. (2 marks)

As before, we can use the `subset()` command:

```
USCrime2014 <- subset(USCrime, subset = Year == 2014)
```

2

j. For 2014, calculate the total violent crime rate per 100000 people in each state, and store it in a vector `USCrimeRate2014`. (3 marks)

1 1 1

```
USCrimeRate2014 <- 1e+05 * USCrime2014$TotVCrime/USCrime2014$Population
```

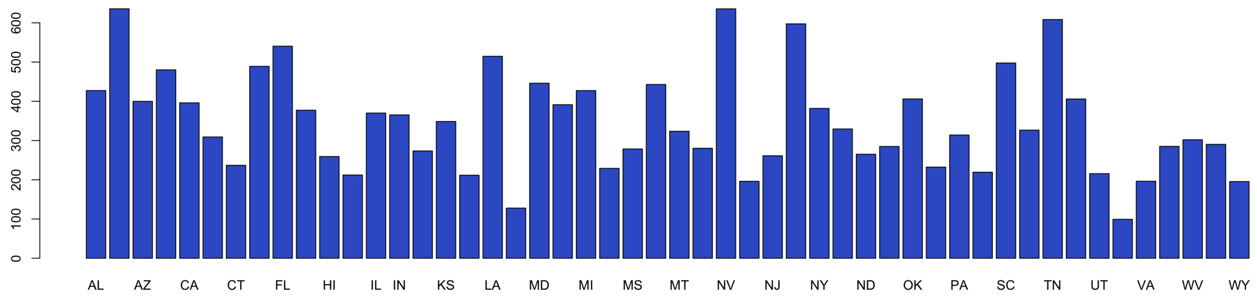
k. Produce an appropriately labelled barplot of the 2014 crime rate in each state. Bonus marks if you can show it in increasing or decreasing order. (4 marks)

If you look at the help file for the function `barplot`, you'll see that the argument `names.arg` can be used to plot labels for each bar, and we can use the variable `Abbr` for this purpose. To get all the labels, we could plot them at a 45-degree angle, but that's getting too fancy. Let's just widen the plot instead to get at least some of the lables.

```
barplot(USCrimeRate2014, names.arg = USCrime2014$Abbr, col = "royalblue3")
```

1

3



Plotting the bars in increasing or decreasing order is not much more difficult; we simply `sort` the crime rate, but in order to get the correct ordering of state labels, we have to remember to order the labels according to the crime rate, using the function `order`. We'll also make the plot a bit fancier by using a colour gradient using the function `brewer.pubu` in the library `pals`. It constructs a gradient from light purple to dark blue.

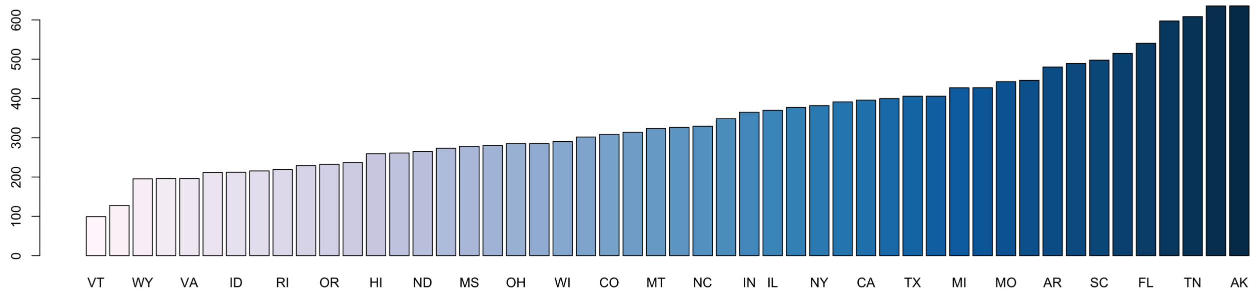
To run the chunk below, install the library `pals`, and then change `eval=FALSE` to `eval=TRUE`.

Bonus +3

Bonus +2.

```
library(pals)
barplot(sort(USCrimeRate2014), names.arg = USCrime2014$Abbr[order(USCrimeRate2014)],
    col = brewer.pubu(50))
```

not necessary



l. Which state has the highest rate in 2014, and which has the lowest rate? What are they (the rates)? (2 marks)

From the plot above, it looks as if Vermont (VT) and Alaska (AK) are at the two ends of the spectrum, but we can confirm it as follows. First, the minimum and maximum crime rates are straightforward to extract:

```
range(USCrimeRate2014)
```

```
[1]  99.2719 635.7807
```

1 or min/max

The elements of the vector `USCrimeRate2014` do not give us the state abbreviations (or names), but we can do so using the `function` and remembering that the ordering of the elements of `USCrimeRate2014` is alphabetically according to the states. Thus, we can do the following:

```
names(USCrimeRate2014) <- USCrime2014$State  # assign state abbreviations to each element
sort(USCrimeRate2014)[c(1, 50)]  # sort the 2014 crime rate, and then extract and display the fir
st and 50th elements
```

```
 Vermont   Alaska 
 99.2719 635.7807 
```

1/2 x 2

We can use similar syntax to find out the three lowest and the three highest:

```
sort(USCrimeRate2014)[c(1:3, 48:50)]
```

```
  Vermont      Maine    Wyoming Tennessee     Nevada     Alaska
  99.2719   127.8110   195.4967  608.4266   635.5890   635.7807
```

m. In Workshop 3, you produced a plot of the Perth region, and then superimposed circles at the geographical center of each suburb, where the size of the circle was proportional to the number of students in that suburb who were enrolled in a particular unit. In this question, you will be doing something similar, except that the map will be of the continental US, and the size of the circles will be proportional to the total violent crime rate in 2014.

i. First, load the appropriate libraries into your session, and then modify the code from Workshop 3 to produce **only** a map of the continental US (without Alaska and Hawaii). The continental US is (roughly) between longitude -125 and -66, and latitude 25 and 50. **Do not add a scale nor add cities!** (4 marks)

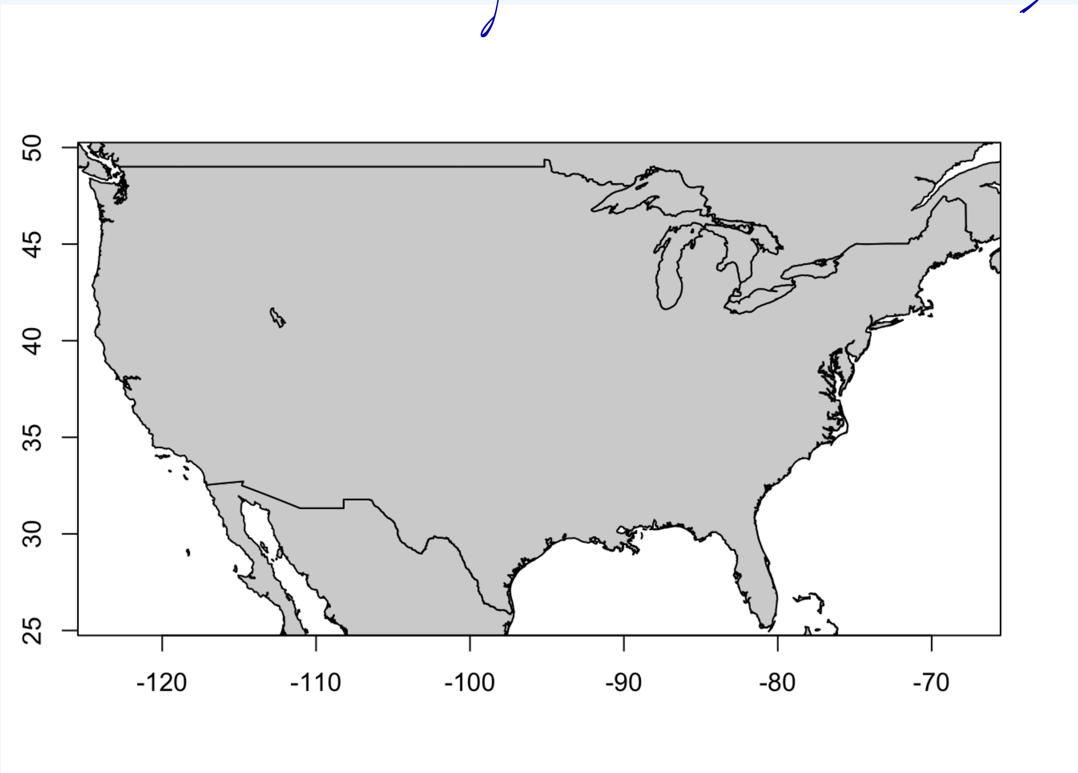
Plotting this should be straightforward: all you need to do is to remember that latitude is plotted on the y-axis and longitude on the x-axis.

```
library(maps)
library(mapdata)
```

1 1

```
map("worldHires", ylim = c(25, 50), xlim = c(-125, -66), fill = TRUE, col = "lightgrey",
    mar = c(4.5, 4, 1, 1))
map.axes()
```

2 for map(" . . . " ,) command

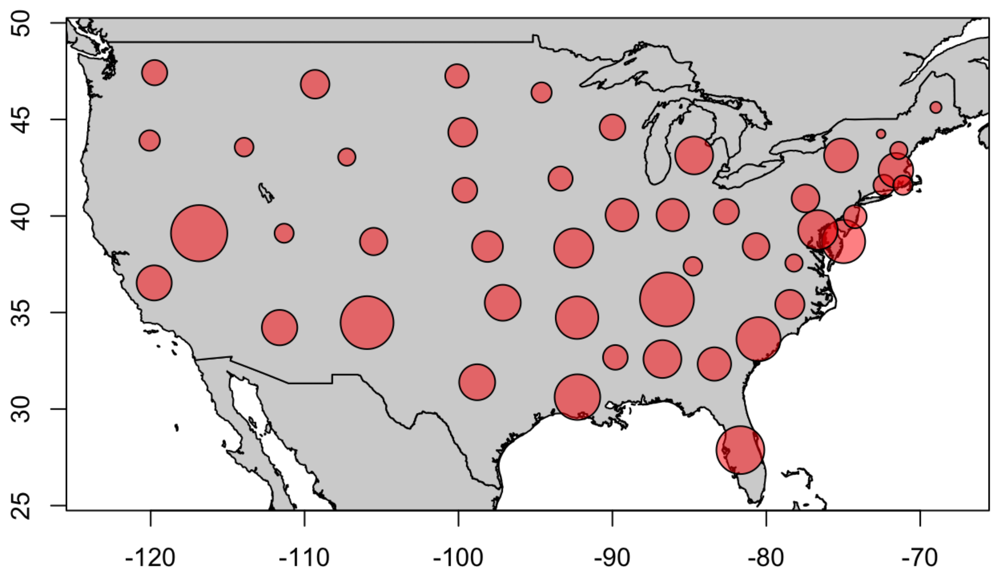


ii. Once you have been able to produce a map, then plot circles whose diameter is proportional to the total crime rate. The divisor for the `cex` argument may have to be large, somewhere between 50 and 300! Use the same colours and symbols as in Workshop 3. Don't forget that the centers of each state are given in the data frame `USGeogData`. (3 marks)

This should have been straightforward: the key again was to remember that in plotting the points on the map, longitude is on the x-axis and latitude is on the y-axis. Getting the size of the circles just right requires adjusting the divisor of `USCrimeRate2014` in the argument `cex` just right. I gave you a hint about its rough magnitude in the question.

```
map("worldHires", ylim = c(25, 50), xlim = c(-125, -66), fill = TRUE, col = "lightgrey",
    mar = c(4.5, 4, 1, 1))
map.axes()
points(USGeogData$Long, USGeogData$Lat, cex = USCrimeRate2014/130, pch = 21,
    col = "black", bg = rgb(1, 0, 0, 0.5))
```

1 1 1



Note that because this is a map of the continental US, Alaska doesn't appear, and the two big circles that we see correspond to Nevada in the west and Tennessee in the south.

A more common way of presenting such data is to use *choropleth* maps, which are maps where geographic or administrative regions are coloured according to some metric; in our case, it would be nice to colour each state with, for example, a more intense colour corresponding to a higher 2014 crime rate.

To run the chunk below, make sure you have installed the libraries `maps`, `pals`, and `SDMTools`, and then change `eval=FALSE` to `eval=TRUE`.

```
require(pals)
require(SDMTools)

names(USCrimeRate2014) <- tolower(names(USCrimeRate2014))
SCols <- brewer.pubu(length(USCrimeRate2014))
USCrimeRate2014 <- sort(USCrimeRate2014)
CrimeRateCols <- data.frame(State = names(USCrimeRate2014), CrimeRate = USCrimeRate2014,
    Cols = SCols)

State <- maps::map("state", xlim = c(-125, -66), ylim = c(25, 50), mar = c(0,
    0, 0, 0), plot = FALSE)$names
State <- unlist(lapply(strsplit(State, ":"), function(x) x[[1]]))
State <- data.frame(State = State)
StateCols <- merge(State, CrimeRateCols, all.x = TRUE)

map("state", xlim = c(-130, -66), ylim = c(25, 52), fill = TRUE, mar = c(0,
    0, 0, 0), col = as.character(StateCols$Cols))
par(oma = c(0, 0, 4, 0))
title(main = "2014 US Total Violent Crime Rate (per 100,000 people)", outer = TRUE,
    line = 3)
x = c(-20, -18, -18, -20)
y = c(25, 53, 53, 25)
legend.gradient(cbind(x = x - 110, y = y), cols = as.character(CrimeRateCols$Cols),
    limits = round(range(CrimeRateCols$CrimeRate)), title = "")
```

2014 US Total Violent Crime Rate (per 100,000 people)