The University of British Columbia

Irving K. Barber Faculty of Science

DATA 101

Assignment 5

Please submit your assignment as an R script file named with your last name, student number, assignment number and with the suffix R. For example, if Joe Smith, student number 87654321 hands in Assignment 4, he would name the file `Smith87654321A2.R`.

Within your answer file, include answers with your R code preceded by the `#` sign. For example, to answer the 5th question on an assignment which is "Perform the calculation $2 + 2$", you would type

```
# Question 5
2 + 2 #coding
# 4 (your answer here )
```

Due Date: Dec 3, 2020

1. We have made a function called `hornerpoly()` in question 3 lab 8. You can use this function to write another function that can evaluate $\sin(x)$ by using the formula below:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

or

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}.$$

Of course, you cannot evaluate the infinite sum, so you will need to evaluate only the first m terms of the sum, say, with $m = 9$ or $m = 10$. Remember that $n!$ can be calculated with the function in R `factorial()`. Call your function `sintaylor()`. (5 points)

Solution:

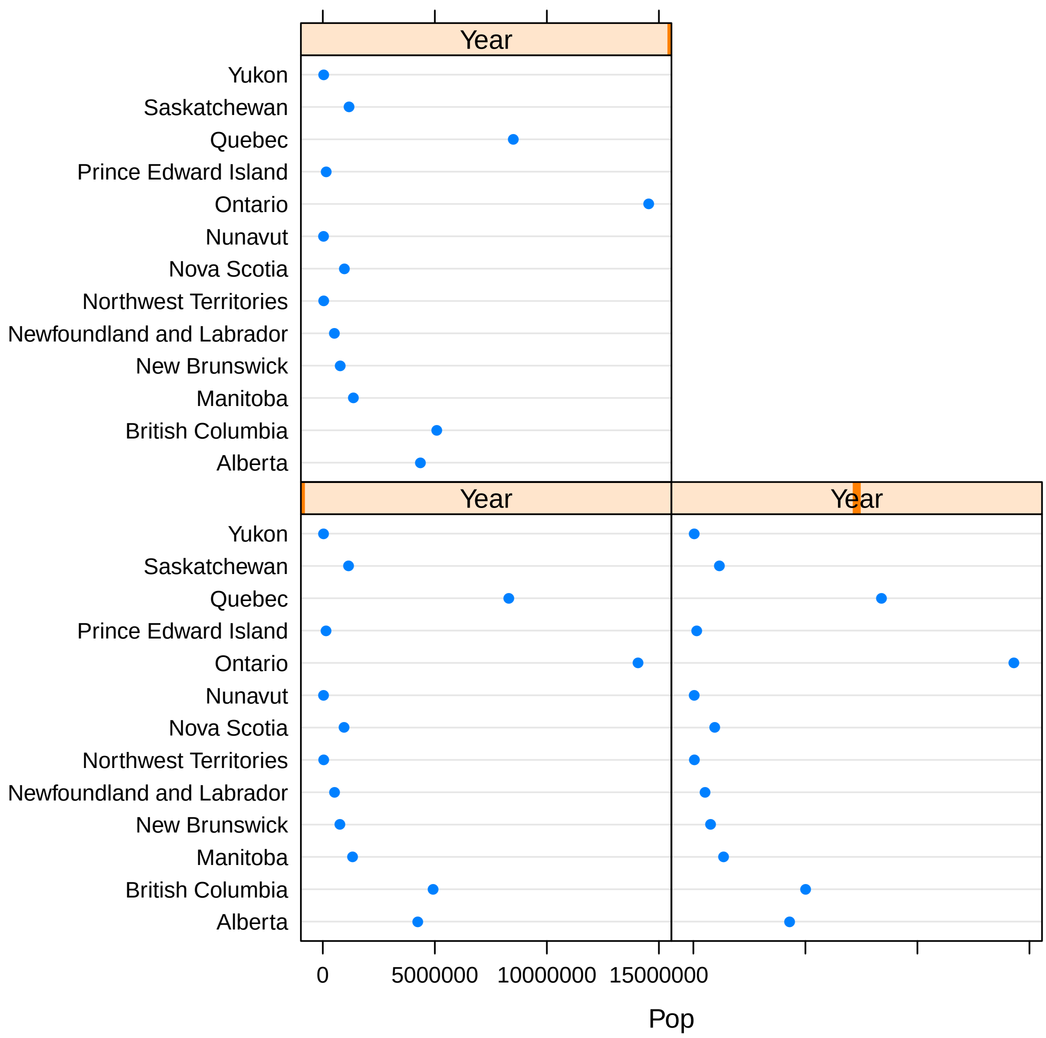
```r
sintaylor <- function(x) {
     a <- numeric(12)
     a[seq(2, 12, 2)] <- (-1)^n/factorial(2*n+1)
     hornerpoly(x, a)
}
```

2. Population figures for all Canadian provinces and territories for 2017 through 2019 can be found in the 13th column of the file *population.csv*. The first two columns of that file give the year and the region. Read the data into R and create a new data frame object called `CanPop` that contains only those three columns, with labels `Year`, `Province` and `Pop`. In addition, create a *lattice* dotplot that has three panels showing the populations for each region, one panel for each of the years. (4 points)

Solution:

```
CanPop <- read.csv("population.csv")
CanPop <- CanPop[, c(1, 2, 13)]
names(CanPop) <- c("Year", "Province", "Pop")
```

```
library(lattice)
dotplot(Province~ Pop|Year, data=CanPop)
```



3. Numbers of homicides for all Canadian provinces and territories for 2017 through 2019 can be found in the 11th column of the file *murders.csv*. The first two columns of that file give the year and the region.

   (a) Read the data into R and create a new data frame object called `CanMurders` that contains only those three columns (i.e. 1, 2 and 11), with labels `Year`, `Province` and `Murders`. (2 points)

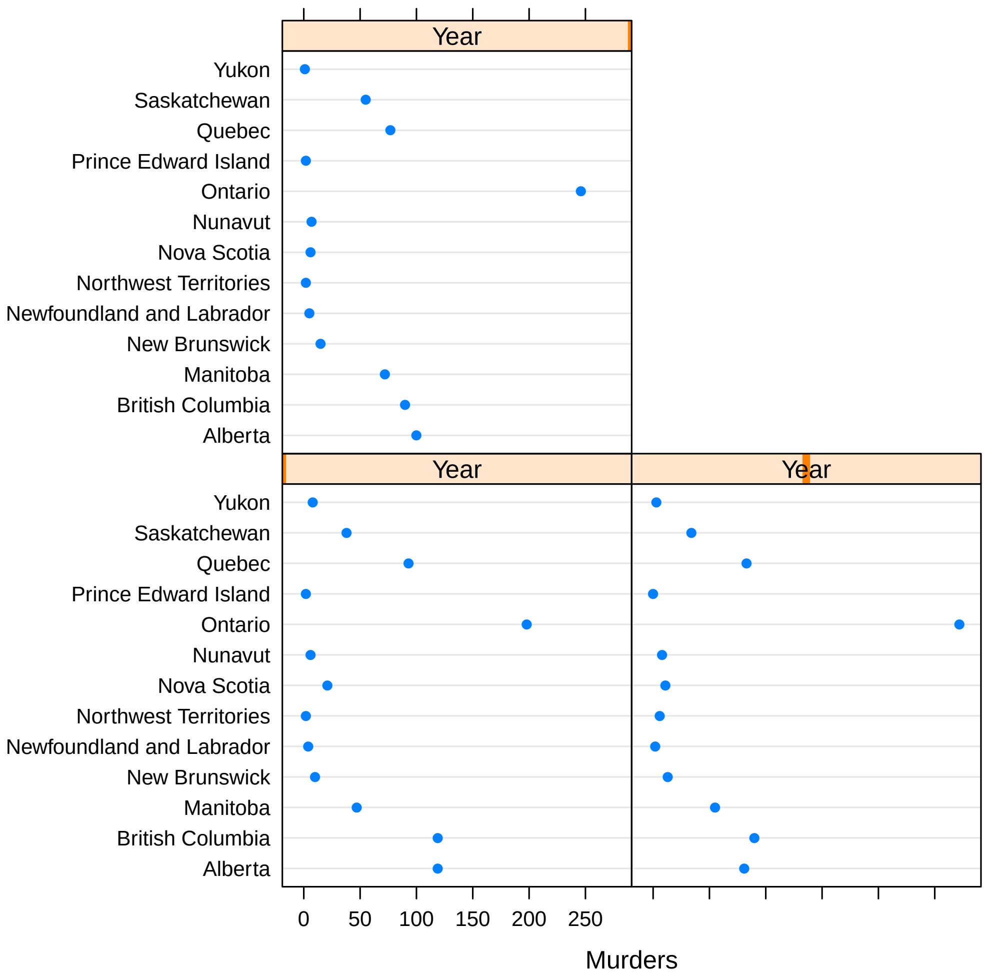
Solution:

```
CanMurders <- read.csv("murders.csv")
CanMurders <- CanMurders[, c(1, 2, 11)]
names(CanMurders) <- c("Year", "Province", "Murders")
```

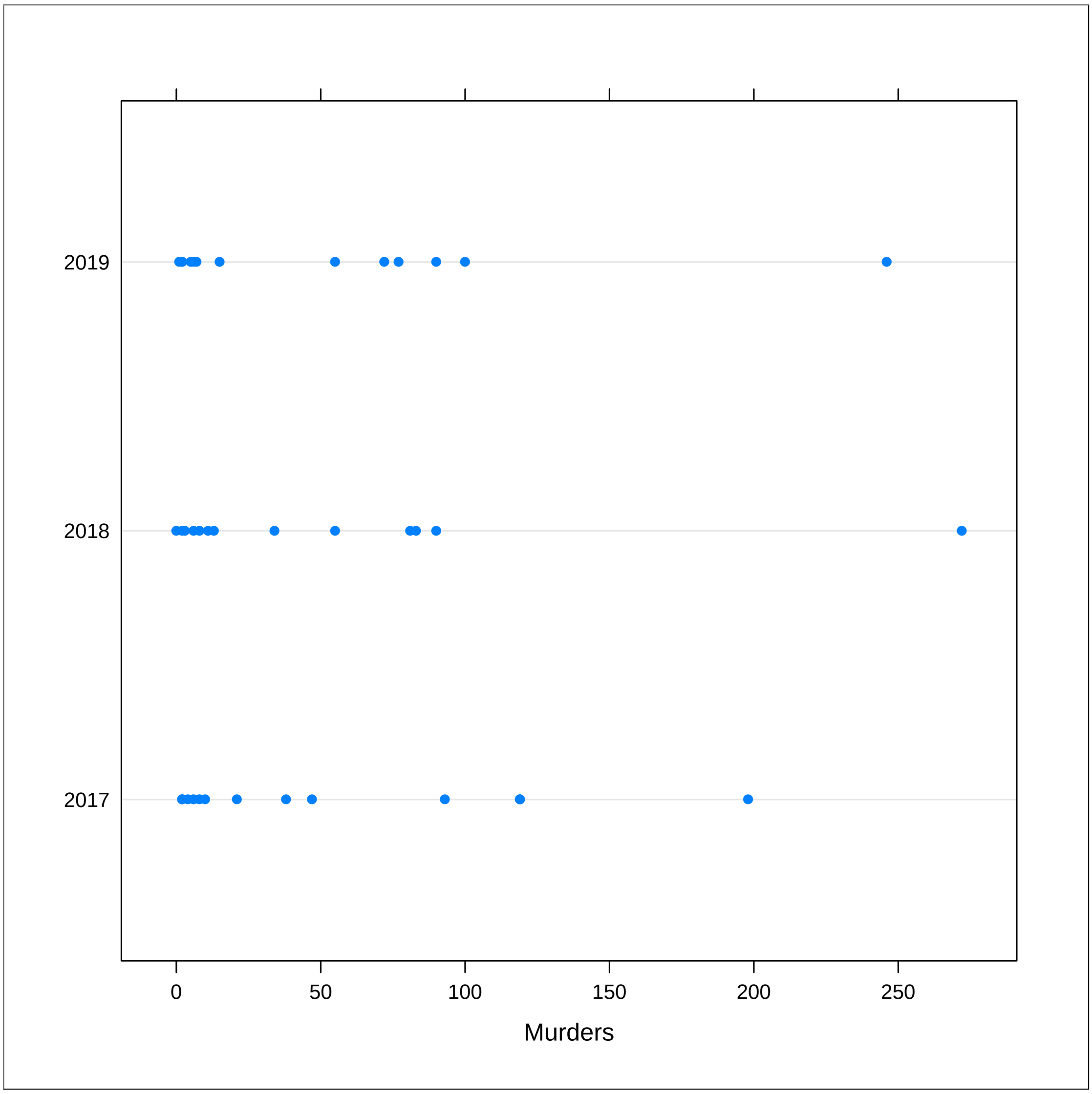
(b) Create *lattice* dotplots showing the numbers of murders in each region, by year and a single dotplot showing the numbers of murders each year. (You will need to convert the `Year` column to character data using the `as.character()` function beforehand.) (3 points)

Solution:

```
dotplot(Province~ Murders|Year, data=CanMurders)
```



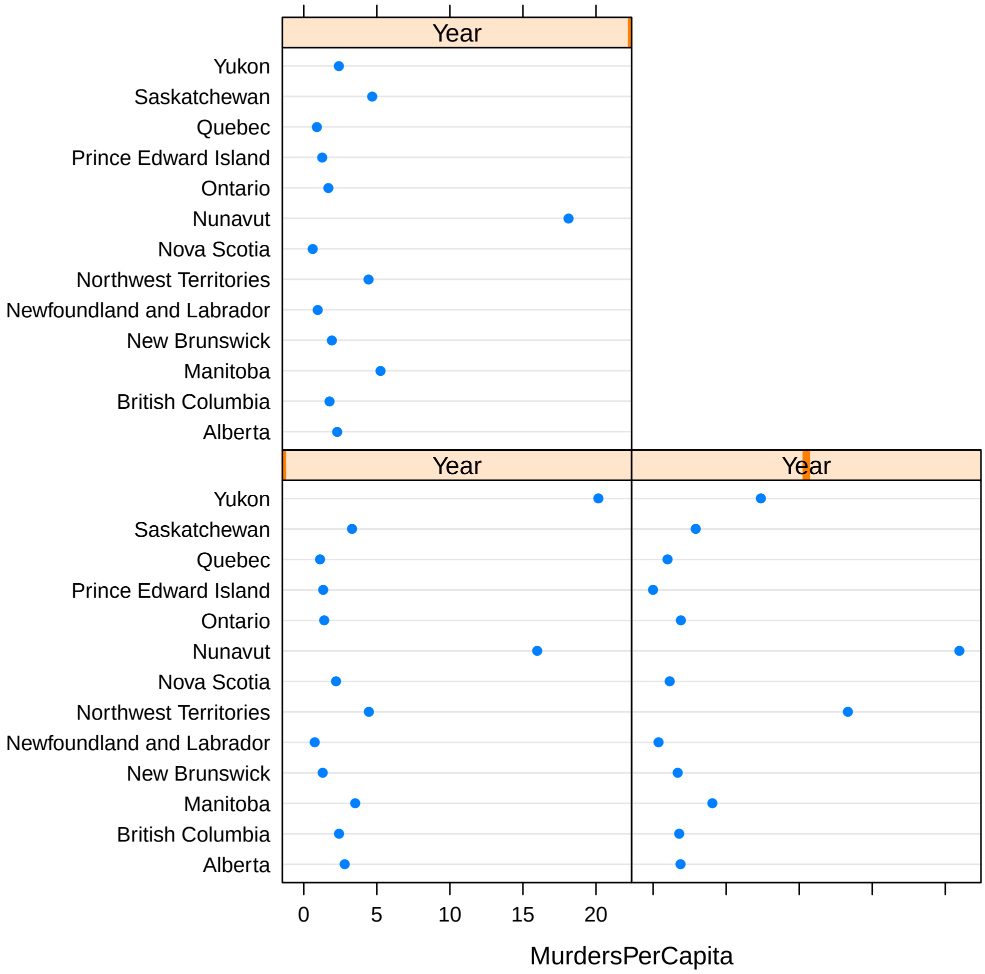

```
dotplot(as.character(Year)~ Murders, data=CanMurders)
```



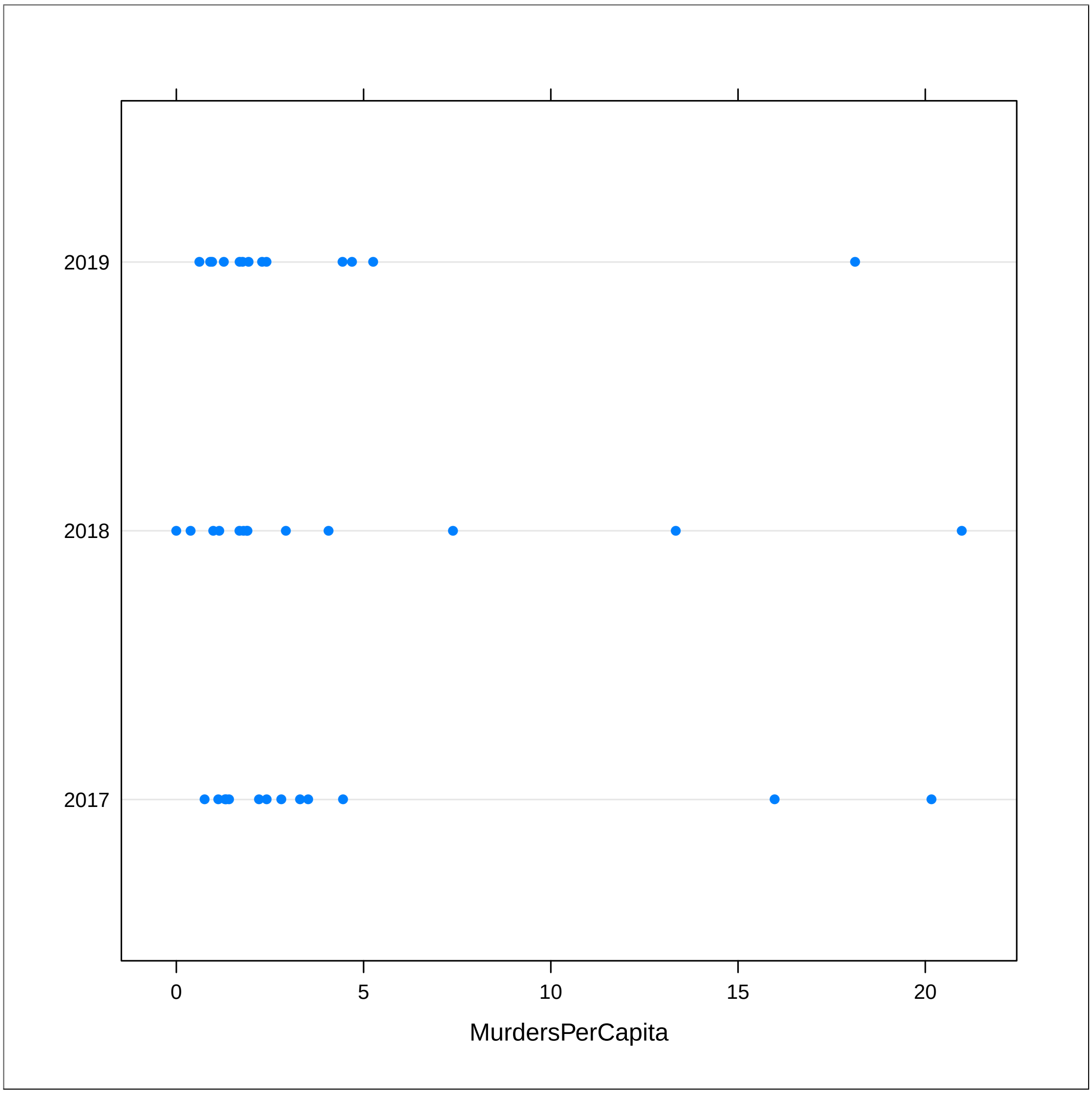
(c) Create a new column in the data frame `CanMurders`, called `MurdersPerCapita` which contains the number of murders divided by the provincial or territorial population multiplied by 100000. This is the number of murders per 100000 population and is a more useful basis for regional comparisons. (The population for the provinces and territories is in the data frame discussed in the previous question.) Then create dotplots showing murders per capita in each region by year and a dotplot showing numbers of murders in each year. (2 points)

Solution:

```
CanMurders$MurdersPerCapita <- CanMurders$Murders/CanPop$Pop*100000
dotplot(Province~ MurdersPerCapita|Year, data=CanMurders)
```




```r
dotplot(as.character(Year)~ MurdersPerCapita, data=CanMurders)
```



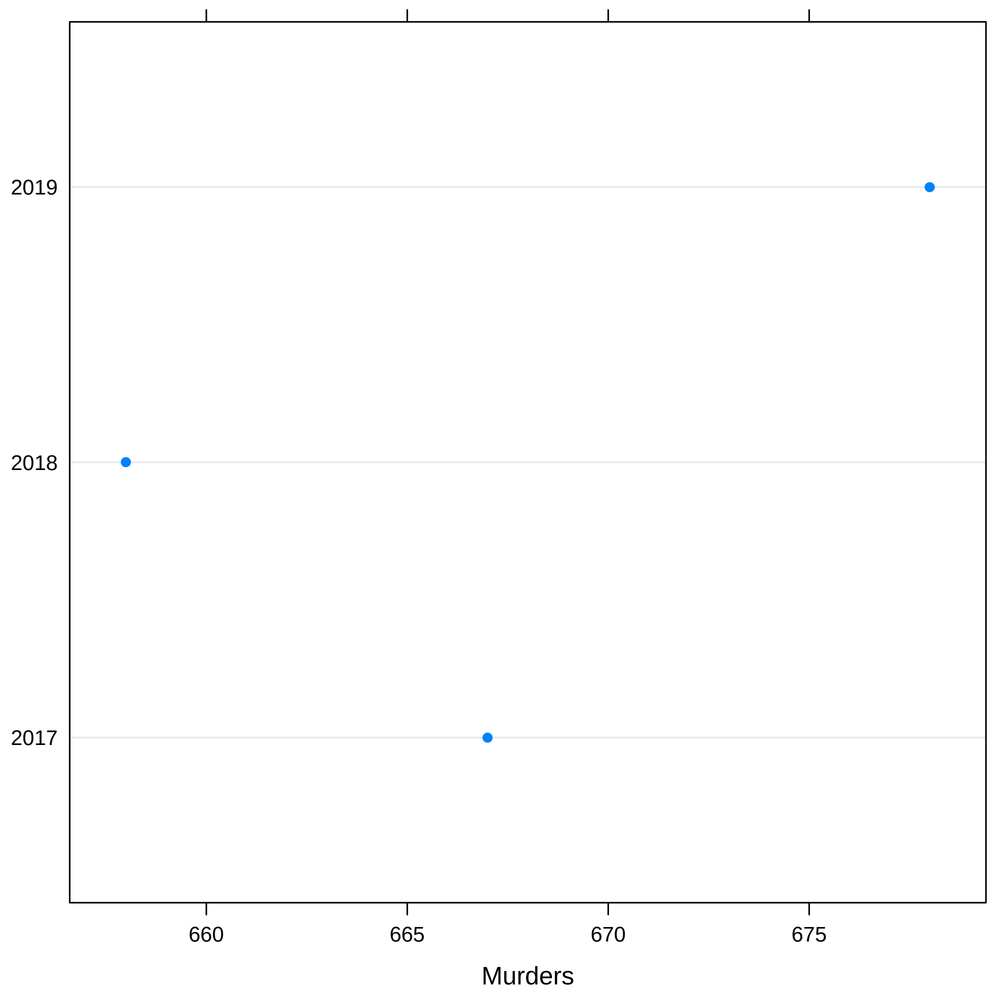
(d) Using the data frame `CanMurders`, create a new data frame called `SumMurders` that contains three columns and three rows. The first column name is `Year` and will contain the year data. The second column name is `Murders` that shows the total number of murders for all of Canada in that year. And the third column names MurdersPerCapita that shows the total number of murders per capita in that year for all of Canada. (You may use the data frame `CanPop` to help construct this column.) And creat the dot plot showing the total numbers of murders and the numbers of murders per capita for all nation each year.(10 points)

Solution:

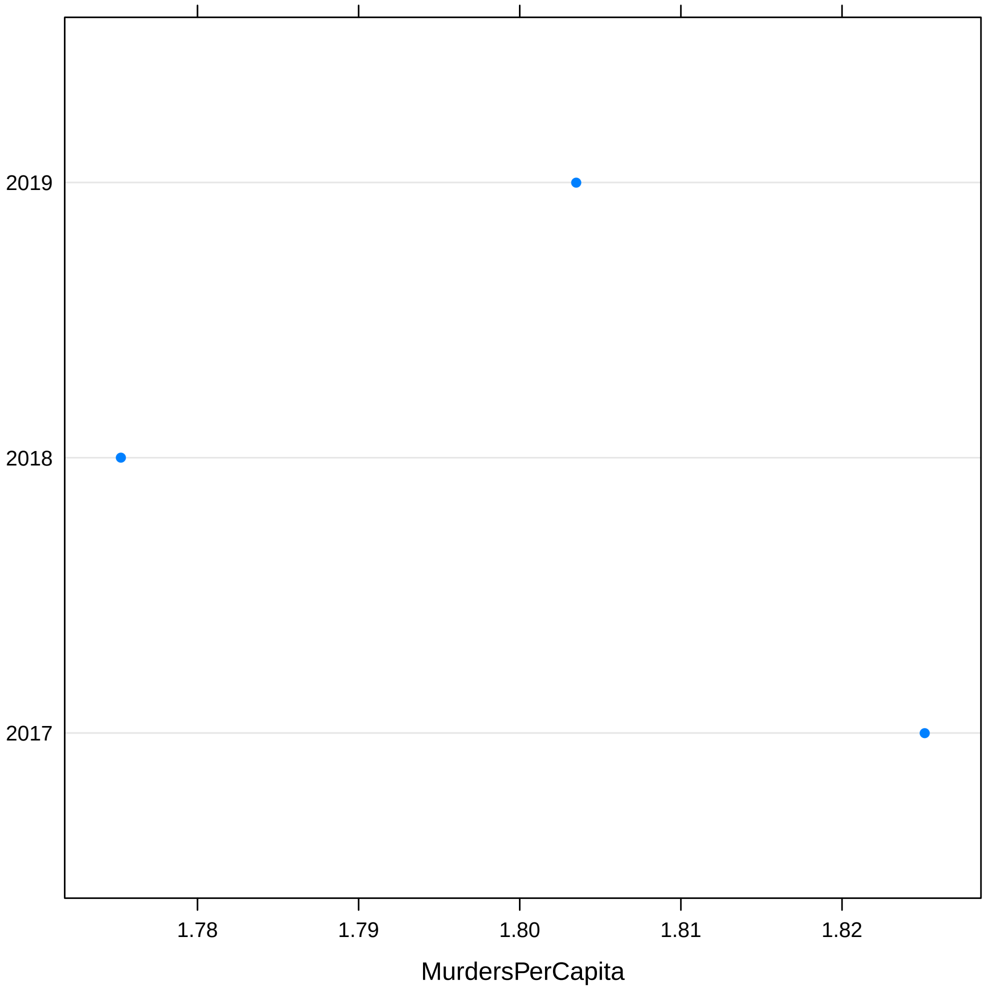
```r
year <- 2017:2019
SumMurders <- CanMurders[3, c(1,3,4)]
for (i in 1:length(year)){
      subyear <- subset(CanMurders, Year==year[i])
      subpop <- subset(CanPop, Year==year[i])
      SumMurders[i, ]<- c(year[i], sum(subyear$Murders),
          sum(subyear$Murders)/sum(subpop$Pop)*100000)
}
row.names(SumMurders)<- NULL
SumMurders
```

```
##   Year Murders MurdersPerCapita
## 1 2017     667         1.825132
## 2 2018     658         1.775251
## 3 2019     678         1.803509
```

```r
dotplot(as.character(Year)~ Murders, data=SumMurders)
```



```
dotplot(as.character(Year)~ MurdersPerCapita, data=SumMurders)
```

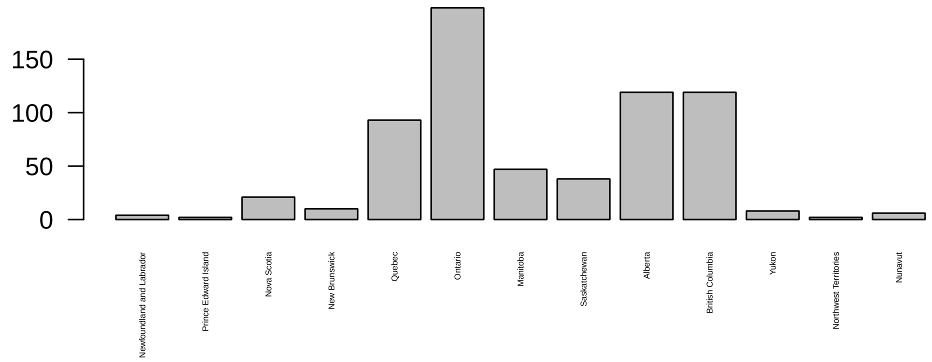


(e) Create a bar plot of 2017 provincial and territorial homicide counts and compare it with the corresponding bar plot of 2017 provincial and territorial per capita homicide rates. Include labels for the bars using the `names.arg` argument containing the entries of the `Province` column and using `cex.name = .35` and `las=2` for readability. (4 points)

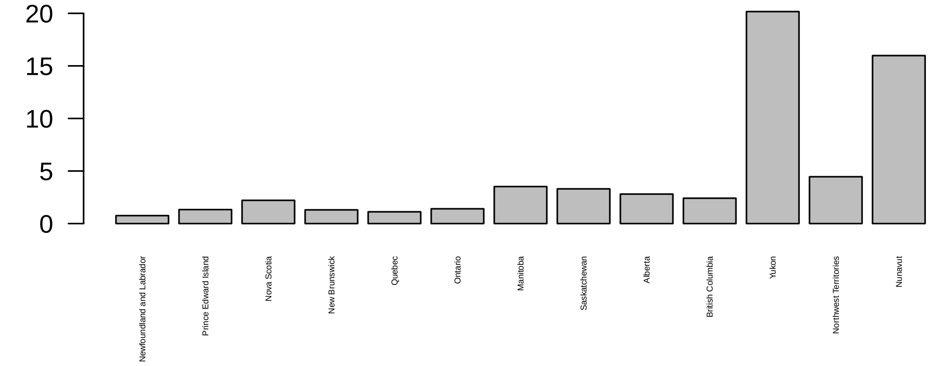
Solution:

```
CanMurders17 <- subset(CanMurders, Year==2017)
CanMurders18 <- subset(CanMurders, Year==2018)
CanMurders19 <- subset(CanMurders, Year==2019)
```

```
barplot(CanMurders17$Murders, names.arg=CanMurders17$Province,
    las=2, cex.names=.35)
```



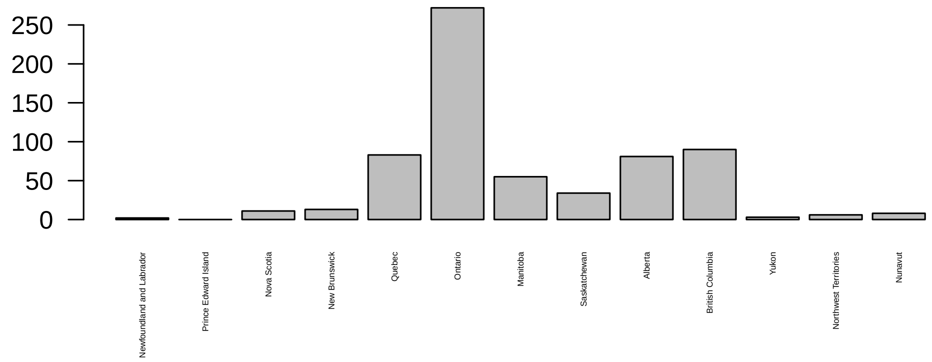

```r
barplot(CanMurders17$MurdersPerCapita, names.arg=CanMurders17$Province,
      las=2, cex.names=.35)
```



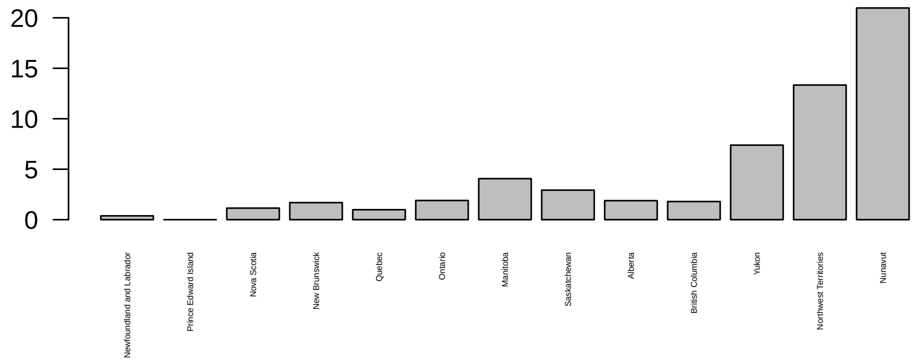
(f) Repeat the previous exercise for 2018 and 2019. (4 points)

Solution:

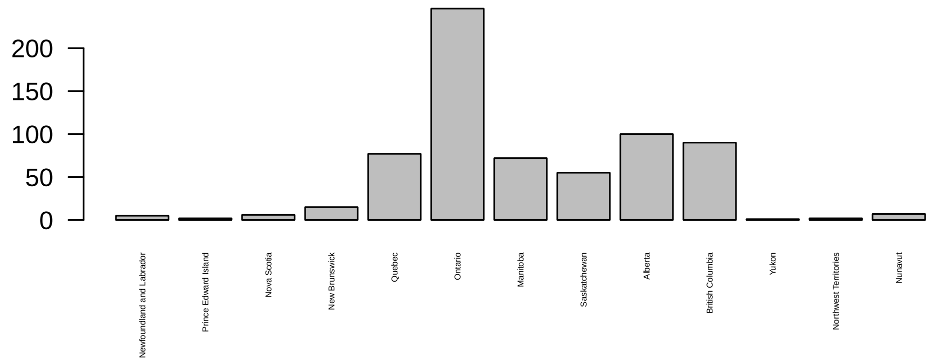
```r
barplot(CanMurders18$Murders, names.arg=CanMurders17$Province,
     las=2, cex.names=.35)
```



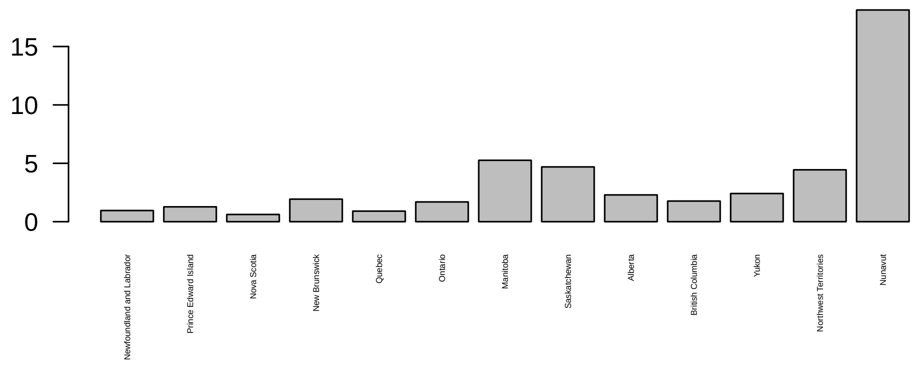

```
barplot(CanMurders18$MurdersPerCapita, names.arg=CanMurders17$Province,
     las=2, cex.names=.35)
```




```
barplot(CanMurders19$Murders, names.arg=CanMurders17$Province,
    las=2, cex.names=.35)
```



```
barplot(CanMurders19$MurdersPerCapita, names.arg=CanMurders17$Province,
     las=2, cex.names=.35)
```

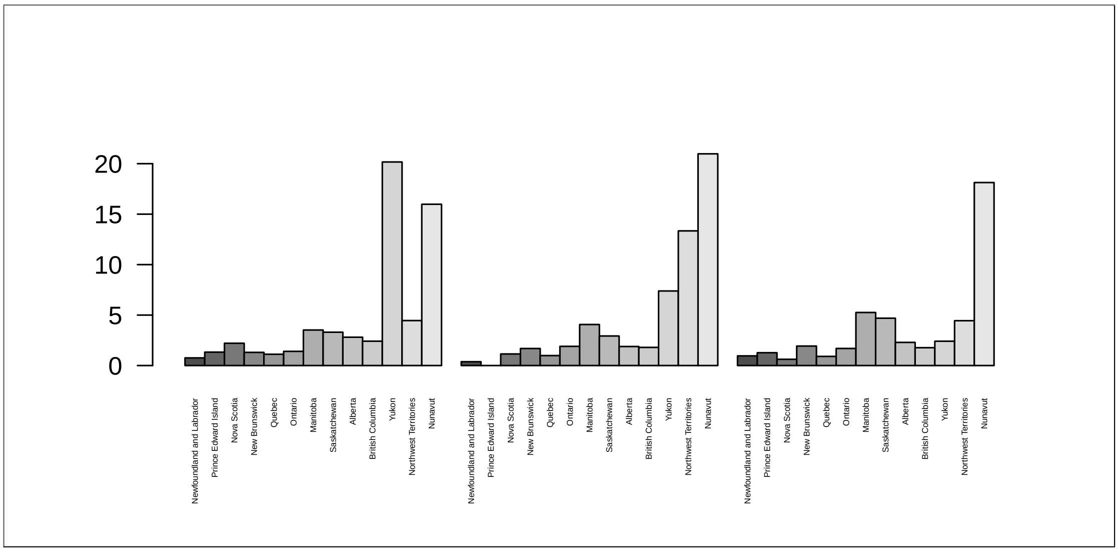


(g) Create a 3 column matrix called `CanHomicides` consisting of the per capita murder rates for the provinces and territories for the three years, using the `cbind` function. Name the columns of the matrix `2017`, `2018` and `2019`.

Also, create side-by-side bar plots of the per capita homicide rates for the provinces and territories, one for each of the three years. Use the `rep` function and the `Province` column of `CanMurders17` in the `names.arg` argument to print bar labels below the horizontal axis. (5 points)

Solution:

```
CanHomicides <- cbind(CanMurders17$MurdersPerCapita,
    CanMurders18$MurdersPerCapita,
    CanMurders19$MurdersPerCapita)
colnames(CanHomicides) <- c("2017", "2018", "2019")
rownames(CanHomicides) <- CanMurders17$Province
barplot(CanHomicides, beside=TRUE, names.arg=rep(CanMurders17$Province, 3),
    las=2, cex.names=.35)
```



4. Consider the data in *radon.R*.

(a) Fit a regression model relating measurement to diameter. Compute the PRESS statistic using the `PRESS` function in the *MPV* package. (3 points)

Solution:

```
source("radon.R")
summary(radon)

##   measurement       diameter          time     
##  Min.   :60.00   Min.   :0.370   Min.   :1.00  
##  1st Qu.:66.75   1st Qu.:0.510   1st Qu.:1.75  
##  Median :74.00   Median :0.865   Median :2.50  
##  Mean   :72.38   Mean   :1.000   Mean   :2.50  
##  3rd Qu.:77.50   3rd Qu.:1.400   3rd Qu.:3.25  
##  Max.   :85.00   Max.   :1.990   Max.   :4.00  

radon1.lm <- lm(measurement ~ diameter, data = radon)
library(MPV) # contains PRESS()

## Loading required package:  KernSmooth
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009

PRESS(radon1.lm)

## [1] 266.6634
```

(b) Fit a regression model relating measurement to both diameter and time. The time is the order for recording measurment for each diameter such as `time<- rep(1:4, 6)` Compute the PRESS statistic. Which of your two models is better? (4 points)

Solution:

```
time <- rep(1:4, 6)
radon$time <- time
radon2.lm <- lm(measurement ~ diameter + time, data = radon)
PRESS(radon2.lm)

## [1] 151.6716
```

The PRESS is lower for the second model, so we would prefer the second model.

(c) Predict the first measurement for a diameter of 0.9 (i.e. assume `time` is 1). Compare with what you would predict for the fourth measurement for a diameter of 0.9 (i.e. assume `time` is 4). What are the 95% prediction intervals in each case? (2 points)

Solution:

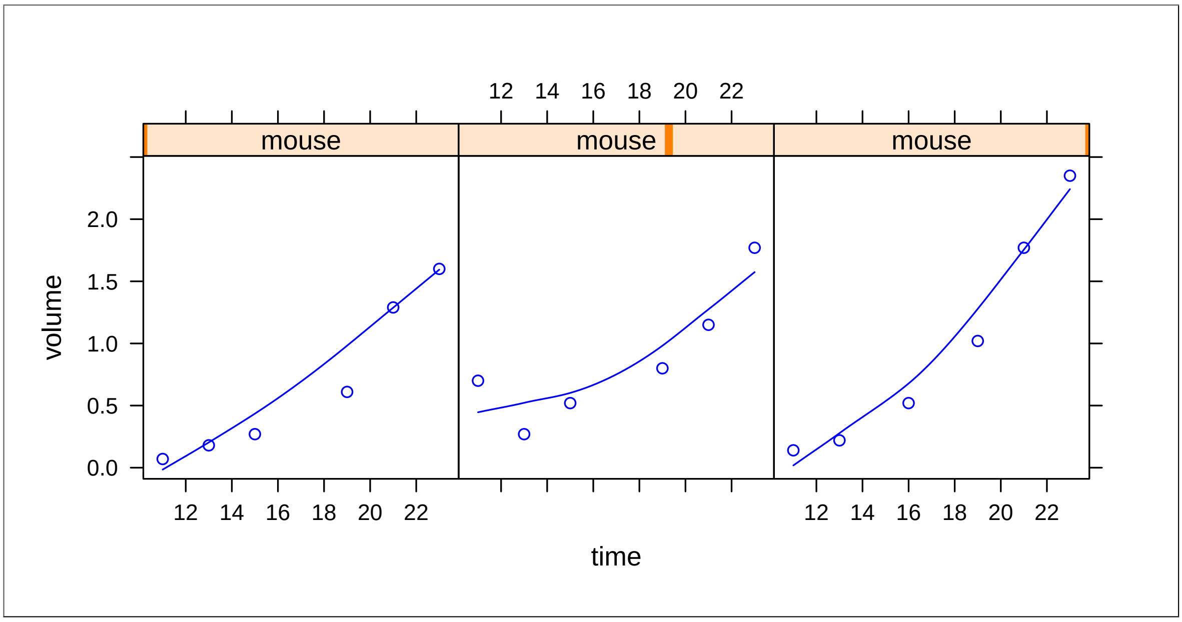
```
predict(radon2.lm, newdata = data.frame(diameter = c(.9, .9),
                time = c(1, 4)), interval="prediction")

##        fit      lwr      upr
## 1 70.73473 65.61045 75.85902
## 2 76.38473 71.26045 81.50902
```

5. The mouse tumor data `mousetumours.R` were collected at University Hospital in London, Ontario. They concern measurements of tumours growing in three different mice at different times following injection of a known carcinogen. Use `xyplot()` to construct a graph of volume against time, for each mouse. Overlay a smooth curve in each panel. (4 points)

Solution:

```
source("mousetumours.R")
```

```
xyplot(volume ~ time|mouse, data=mousetumours,
    type=c("p", "smooth"), span=1.2, col=4)
```



6. The `table.b4` data frame in the *MPV* package has 24 observations on property valuation. There are 10 columns in the data frame. Use `help()` function to check what the variables are. Try out the models below and identify the one that you think is best. (Using `PRESS()`)

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \varepsilon && (1)\\
y &= \beta_0 + \beta_1 x_1 + \varepsilon && (2)\\
y &= \beta_0 + \beta_1 x_1 + \beta_8 x_8 + \varepsilon && (3)\\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon && (4)\\
y &= \beta_0 + \beta_1 x_1 + \beta_5 x_5 + \varepsilon && (5)\\
y &= \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_8 x_8 + \varepsilon && (6)\\
y &= \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \varepsilon && (7)\\
y &= \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_9 x_9 + \varepsilon && (8)
\end{aligned}
$$

If we don't have any information on taxes, which aspects of the house are most useful for us to look at in order to predict house price? Write down the predictive model. (5 points)

Solution:

```
b4.lm <- lm(y~., data=table.b4)
PRESS(b4.lm)

## [1] 393.492

b4_1.lm <- lm(y~x1, data=table.b4)
PRESS(b4_1.lm)

## [1] 224.6111

b4_2.lm <- lm(y~x2+x3+x4+x5+x6+x7+x8+x9, data=table.b4)
PRESS(b4_2.lm)

## [1] 424.1779

b4_3.lm <- lm(y~x1+x8, data=table.b4)
PRESS(b4_3.lm)

## [1] 253.9058

b4_4.lm <- lm(y~x2+x3+x5+x8, data=table.b4)
PRESS(b4_4.lm)

## [1] 331.0418

 b4_5.lm <- lm(y~x1+x2, data=table.b4)
PRESS(b4_5.lm)

## [1] 224.441

 b4_6.lm <- lm(y~x1+x5, data=table.b4)
PRESS(b4_6.lm)

## [1] 236.814

b4_7.lm <- lm(y~x2+x3+x5, data=table.b4)
PRESS(b4_7.lm)

## [1] 374.1862

 b4_8.lm <- lm(y~x2+x3+x5+x9, data=table.b4)
PRESS(b4_8.lm)

## [1] 422.3961
```

The best models is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ because the PRESS is 224.441.
If we don't have information of tax, the best model is

```
b4_4.lm <- lm(y~x2+x3+x5+x8, data=table.b4)
PRESS(b4_4.lm)

## [1] 331.0418

summary(b4_4.lm)

##
## Call:
## lm(formula = y ~ x2 + x3 + x5 + x8, data = table.b4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6728 -1.8525 -0.0162  0.9609  7.7641
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.58367    4.17021   3.497  0.00241 **
## x2          12.51268    3.01261   4.153  0.00054 ***
## x3           0.92173    0.39149   2.354  0.02946 *
## x5           2.78020    1.11662   2.490  0.02221 *
## x8          -0.10088    0.04975  -2.028  0.05682 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.126 on 19 degrees of freedom
## Multiple R-squared:  0.776,Adjusted R-squared:  0.7289
## F-statistic: 16.46 on 4 and 19 DF,  p-value: 5.618e-06
```

$$\hat{y} = 15.58 + 12.51x_2 + 0.92x_3 + 2.78x_5 - 0.1x_8$$

Data References

1. Statistics Canada. Table 35-10-0068-01 Number, rate and percentage changes in rates of homicide victims DOI: https://doi.org/10.25318/3510006801-eng

2. Statistics Canada. Table 17-10-0060-01 Estimates of population as of July 1st, by marital status or legal marital status, age and sex DOI: https://doi.org/10.25318/1710006001-eng