# Test your knowledge of R

*The purpose of this set of exercises is to help build your familiarity with R. The goal here is not to make you the most proficient R programmer but rather focus on how to use R to help in analytics.*

1. Define: `x <- c(4,2,6)` and `y <- c(1,0,-1)`. Determine the result of the following using R.

```
x <- c(4,2,6)
y <- c(1,0,-1)
length(x)
```

```
## [1] 3
```

```
sum(x)
```

```
## [1] 12
```

```
sum(x^2)
```

```
## [1] 56
```

```
x+y
```

```
## [1] 5 2 5
```

```
x*y
```

```
## [1]  4  0 -6
```

```
x-2
```

```
## [1] 2 0 4
```

```
x^2
```

```
## [1] 16  4 36
```

```
x*y[1:2]
```

```
## Warning in x * y[1:2]: longer object length is not a multiple of shorter object
## length
```

```
## [1] 4 0 6
```

2. Decide what the following sequences are and use R to check your answers:

```
7:11
```

```
## [1]  7  8  9 10 11
```

```
seq(2,9)
```

```
## [1] 2 3 4 5 6 7 8 9
```

```
seq(4,10,by=2)
```

```
## [1]  4  6  8 10
```

```
seq(3,30,length=10)
```

```
## [1]  3  6  9 12 15 18 21 24 27 30
```

```r
seq(6,-4,by=-2)
```

```
## [1]  6  4  2  0 -2 -4
```

3. Determine what the result will be of the following R expressions, and then use R to check that you are right:

```r
rep(2,4)
```

```
## [1] 2 2 2 2
```

```r
rep(c(1,5),4)
```

```
## [1] 1 5 1 5 1 5 1 5
```

```r
rep(c(1,2),c(4,4))
```

```
## [1] 1 1 1 1 2 2 2 2
```

4. Define: `x <- c(5,9,2,3,4,6,7,0,8,12,2,9)`. Decide what each of the following is and use R to check your answers:

```r
x <- c(5,9,2,3,4,6,7,0,8,12,2,9)
x[2]
```

```
## [1] 9
```

```r
x[2:4]
```

```
## [1] 9 2 3
```

```r
x[c(2,3,6)]
```

```
## [1] 9 2 6
```

```r
x[c(1:5,10:12)]
```

```
## [1]  5  9  2  3  4 12  2  9
```

```r
x[-(10:12)]
```

```
## [1] 5 9 2 3 4 6 7 0 8
```

5. Create in R the matrices

$$x = \begin{bmatrix} 3 & 2 \\ -1 & -1 \end{bmatrix}, \text{ and, } y = \begin{bmatrix} 1 & 4 & 0 \\ 0 & 1 & -1 \end{bmatrix}.$$

Calculate the following and check your answers in R:

```r
x<-matrix(c(3,-1,2,-1),nrow=2, ncol=2)
y<-matrix(c(1,0,4,1,0,-1),nrow=2, ncol=3)
2*x
```

```
##      [,1] [,2]
## [1,]    6    4
## [2,]   -2   -2
```

```r
x*x
```

```
##      [,1] [,2]
## [1,]    9    4
## [2,]    1    1
```

```r
x%*%x
```

```
##      [,1] [,2]
## [1,]    7    4
## [2,]   -2   -1
```

```r
x%*%y
```

```
##      [,1] [,2] [,3]
## [1,]    3   14   -2
## [2,]   -1   -5    1
```

```r
t(y)
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    4    1
## [3,]    0   -1
```

```r
solve(x)
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]   -1   -3
```

With $x$ and $y$ as above, calculate the effect of the following subscript operations and check your answers in R.

```r
x[1,]
```

```
## [1] 3 2
```

```r
x[2,]
```

```
## [1] -1 -1
```

```r
x[,2]
```

```
## [1]  2 -1
```

```r
y[1,2]
```

```
## [1] 4
```

```r
y[,2:3]
```

```
##      [,1] [,2]
## [1,]    4    0
## [2,]    1   -1
```

6. Internet privacy has gained widespread attention in recent years. To measure the degree to which people are concerned about hot-button issues like Internet privacy, social scientists conduct polls in which they interview a large number of people about the topic. In this question, we will analyze data from a July 2013 Pew Internet and American Life Project poll on Internet anonymity and privacy, which involved interviews across the United States. The dataset `AnonymityPoll.csv` has the following fields (all Internet use-related fields were only collected from interviewees who either use the Internet or have a smartphone):

- Internet.Use: A binary variable indicating if the interviewee uses the Internet, at least occasionally (equals 1 if the interviewee uses the Internet, and equals 0 if the interviewee does not use the Internet).
- Smartphone: A binary variable indicating if the interviewee has a smartphone (equals 1 if they do have a smartphone, and equals 0 if they don't have a smartphone).

- Sex: Male or Female.
- Age: Age in years.
- State: State of residence of the interviewee.
- Region: Census region of the interviewee (Midwest, Northeast, South, or West).
- Conservativeness: Self-described level of conservativeness of interviewee, from 1 (very liberal) to 5 (very conservative).
- Info.On.Internet: Number of the following items this interviewee believes to be available on the Internet for others to see: (1) Their email address; (2) Their home address; (3) Their home phone number; (4) Their cell phone number; (5) The employer/company they work for; (6) Their political party or political affiliation; (7) Things they've written that have their name on it; (8) A photo of them; (9) A video of them; (10) Which groups or organizations they belong to; and (11) Their birth date.
- Worry.About.Info: A binary variable indicating if the interviewee worries about how much information is available about them on the Internet (equals 1 if they worry, and equals 0 if they don't worry).
- Privacy.Importance: A score from 0 (privacy is not too important) to 100 (privacy is very important), which combines the degree to which they find privacy important in the following: (1) The websites they browse; (2) Knowledge of the place they are located when they use the Internet; (3) The content and files they download; (4) The times of day they are online; (5) The applications or programs they use; (6) The searches they perform; (7) The content of their email; (8) The people they exchange email with; and (9) The content of their online chats or hangouts with others.
- Anonymity.Possible: A binary variable indicating if the interviewee thinks it's possible to use the Internet anonymously, meaning in such a way that online activities can't be traced back to them (equals 1 if he/she believes you can, and equals 0 if he/she believes you can't).
- Tried.Masking.Identity: A binary variable indicating if the interviewee has ever tried to mask his/her identity when using the Internet (equals 1 if he/she has tried to mask his/her identity, and equals 0 if he/she has not tried to mask his/her identity).
- Privacy.Laws.Effective: A binary variable indicating if the interviewee believes United States law provides reasonable privacy protection for Internet users (equals 1 if he/she believes it does, and equals 0 if he/she believes it doesn't).

(a) Using `read.csv()`, load the dataset from `AnonymityPoll.csv` into a data frame called poll and summarize it with the `summary()` and `str()` functions.

- How many people participated in the poll?
- *Ans: 1002 people participated in the polls (1002 observations with 13 variables).*

```
poll <- read.csv("AnonymityPoll.csv")
summary(poll)
```

```
##   Internet.Use      Smartphone        Sex                 Age
##  Min.   :0.0000   Min.   :0.0000   Length:1002        Min.   :18.00
##  1st Qu.:1.0000   1st Qu.:0.0000   Class :character   1st Qu.:37.00
##  Median :1.0000   Median :1.0000   Mode  :character   Median :55.00
##  Mean   :0.7742   Mean   :0.5078                      Mean   :52.37
##  3rd Qu.:1.0000   3rd Qu.:1.0000                      3rd Qu.:66.00
##  Max.   :1.0000   Max.   :1.0000                      Max.   :96.00
##  NA's   :1        NA's   :43                          NA's   :27
##     State              Region          Conservativeness Info.On.Internet
##  Length:1002        Length:1002        Min.   :1.000    Min.   : 0.000
##  Class :character   Class :character   1st Qu.:3.000    1st Qu.: 2.000
##  Mode  :character   Mode  :character   Median :3.000    Median : 4.000
##                                        Mean   :3.277    Mean   : 3.795
##                                        3rd Qu.:4.000    3rd Qu.: 6.000
##                                        Max.   :5.000    Max.   :11.000
##                                        NA's   :62       NA's   :210
##  Worry.About.Info Privacy.Importance Anonymity.Possible Tried.Masking.Identity
##  Min.   :0.0000   Min.   :  0.00     Min.   :0.0000     Min.   :0.0000
```

4

```
##  1st Qu.:0.0000    1st Qu.: 41.43    1st Qu.:0.0000    1st Qu.:0.0000
##  Median :0.0000    Median : 68.75    Median :0.0000    Median :0.0000
##  Mean   :0.4886    Mean   : 62.85    Mean   :0.3692    Mean   :0.1633
##  3rd Qu.:1.0000    3rd Qu.: 88.89    3rd Qu.:1.0000    3rd Qu.:0.0000
##  Max.   :1.0000    Max.   :100.00    Max.   :1.0000    Max.   :1.0000
##  NA's   :212       NA's   :215       NA's   :249       NA's   :218
##  Privacy.Laws.Effective
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2617
##  3rd Qu.:1.0000
##  Max.   :1.0000
##  NA's   :108
```

```r
str(poll)
```

```
## 'data.frame':    1002 obs. of  13 variables:
##  $ Internet.Use        : int  1 1 0 1 0 1 1 0 0 1 ...
##  $ Smartphone          : int  0 0 1 0 NA 1 0 0 NA 0 ...
##  $ Sex                 : chr  "Male" "Male" "Female" "Male" ...
##  $ Age                 : int  62 45 70 70 80 49 52 76 75 76 ...
##  $ State               : chr  "Massachusetts" "South Carolina" "New Jersey" "Georgia" ...
##  $ Region              : chr  "Northeast" "South" "Northeast" "South" ...
##  $ Conservativeness    : int  4 1 4 4 4 4 3 3 4 4 ...
##  $ Info.On.Internet    : int  0 1 0 3 NA 6 3 NA NA 0 ...
##  $ Worry.About.Info    : int  1 0 0 1 NA 0 1 NA NA 0 ...
##  $ Privacy.Importance  : num  100 0 NA 88.9 NA ...
##  $ Anonymity.Possible  : int  0 1 0 1 NA 1 0 NA NA 1 ...
##  $ Tried.Masking.Identity: int  0 0 0 0 NA 1 0 NA NA 0 ...
##  $ Privacy.Laws.Effective: int  0 1 NA 0 NA 0 1 NA 0 1 ...
```

(b) Look at the breakdown of the number of people with smartphones using the `table()` command on the `Smartphone` variable.

- How many interviewees responded that they use a smartphone? *Ans. 487 people have a smartphone.*
- How many interviewees responded that they don't use a smartphone? *Ans. 472 people do not use a smartphone.*
- How many interviewees did not respond to the question, resulting in a missing value, or `NA`, in the `summary()` output? *Ans. 43.*

```r
table(poll$Smartphone)
```

```
##
##   0   1
## 472 487
```

```r
summary(poll$Smartphone)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  1.0000  0.5078  1.0000  1.0000      43
```

(c) Look at the breakdown of the number of people with smartphones and Internet use using the `table()` command.

- How many interviewees reported not having used the Internet and not having used a smartphone *Ans. 186.*
- How many interviewees reported having used the Internet and having used a smartphone? *Ans. 470.*

5

- How many interviewees reported having used the Internet but not having used a smartphone? *Ans. 285.*
- How many interviewees reported having used a smartphone but not having used the Internet? *Ans. 17.*

```
table(poll$Internet.Use, poll$Smartphone)
```

```
##
##      0   1
##  0 186  17
##  1 285 470
```

(d) Many of the response variables (Info.On.Internet, Worry.About.Info, Privacy.Importance, Anonymity.Possible, and Tried.Masking.Identity) were not collected if an interviewee does not use the Internet or a smartphone, meaning the variables will have missing values for these interviewees.

- How many interviewees have a missing value for their Internet use? *Ans. 1.*
- How many interviewees have a missing value for their smartphone use? *Ans. 43.*

```
summary(poll)
```

```
##   Internet.Use      Smartphone         Sex                Age
##  Min.   :0.0000   Min.   :0.0000   Length:1002        Min.   :18.00
##  1st Qu.:1.0000   1st Qu.:0.0000   Class :character   1st Qu.:37.00
##  Median :1.0000   Median :1.0000   Mode  :character   Median :55.00
##  Mean   :0.7742   Mean   :0.5078                      Mean   :52.37
##  3rd Qu.:1.0000   3rd Qu.:1.0000                      3rd Qu.:66.00
##  Max.   :1.0000   Max.   :1.0000                      Max.   :96.00
##  NA's   :1        NA's   :43                           NA's   :27
##     State               Region         Conservativeness Info.On.Internet
##  Length:1002        Length:1002        Min.   :1.000    Min.   : 0.000
##  Class :character   Class :character   1st Qu.:3.000    1st Qu.: 2.000
##  Mode  :character   Mode  :character   Median :3.000    Median : 4.000
##                                        Mean   :3.277    Mean   : 3.795
##                                        3rd Qu.:4.000    3rd Qu.: 6.000
##                                        Max.   :5.000    Max.   :11.000
##                                        NA's   :62       NA's   :210
##  Worry.About.Info Privacy.Importance Anonymity.Possible Tried.Masking.Identity
##  Min.   :0.0000   Min.   :  0.00     Min.   :0.0000     Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 41.43     1st Qu.:0.0000     1st Qu.:0.0000
##  Median :0.0000   Median : 68.75     Median :0.0000     Median :0.0000
##  Mean   :0.4886   Mean   : 62.85     Mean   :0.3692     Mean   :0.1633
##  3rd Qu.:1.0000   3rd Qu.: 88.89     3rd Qu.:1.0000     3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :100.00     Max.   :1.0000     Max.   :1.0000
##  NA's   :212      NA's   :215        NA's   :249        NA's   :218
##  Privacy.Laws.Effective
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2617
##  3rd Qu.:1.0000
##  Max.   :1.0000
##  NA's   :108
```

```
# alternatively use
sum(is.na(poll$Internet.Use))
```

```
## [1] 1
```

```
sum(is.na(poll$Smartphone))
```

## [1] 43

(e) Use the subset function to obtain a data frame called `limited`, which is limited to interviewees who reported Internet use or who reported smartphone use.

- How many interviewees are in the new data frame? *Ans. 792*

```
limited <- subset(poll, poll$Internet.Use == 1|poll$Smartphone == 1)
str(limited)
```

```
## 'data.frame':    792 obs. of  13 variables:
##  $ Internet.Use        : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ Smartphone          : int  0 0 1 0 1 0 0 1 1 0 ...
##  $ Sex                 : chr  "Male" "Male" "Female" "Male" ...
##  $ Age                 : int  62 45 70 70 49 52 76 50 47 69 ...
##  $ State               : chr  "Massachusetts" "South Carolina" "New Jersey" "Georgia" ...
##  $ Region              : chr  "Northeast" "South" "Northeast" "South" ...
##  $ Conservativeness    : int  4 1 4 4 4 3 4 3 3 3 ...
##  $ Info.On.Internet    : int  0 1 0 3 6 3 0 1 0 9 ...
##  $ Worry.About.Info    : int  1 0 0 1 0 1 0 0 0 0 ...
##  $ Privacy.Importance  : num  100 0 NA 88.9 88.9 ...
##  $ Anonymity.Possible  : int  0 1 0 1 1 0 1 0 1 0 ...
##  $ Tried.Masking.Identity: int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Privacy.Laws.Effective: int  0 1 NA 0 0 1 1 0 0 0 ...
```

```
# alternatively use
nrow(limited)
```

## [1] 792

(f) For all the remaining questions use the `limited` data frame you have created.

- Which variables have missing values in the `limited` data frame?
  - Smartphone 20 `NAs`

  - Age 22 `NAs`

  - Conservativeness 45 `NAs`

  - Worry.About.Info 2 `NAs`

  - Privacy.Importance 5 `NAs`

  - Anonymity.Possible 39 `NAs`

  - Tried.Masking.Identity 8 `NAs`

  - Privacy.Laws.Effective 65 `NAs`

```
summary(limited)
```

```
##   Internet.Use      Smartphone          Sex                 Age
##  Min.   :0.0000   Min.   :0.0000   Length:792         Min.   :18.00
##  1st Qu.:1.0000   1st Qu.:0.0000   Class :character   1st Qu.:33.00
##  Median :1.0000   Median :1.0000   Mode  :character   Median :51.00
##  Mean   :0.9785   Mean   :0.6308                      Mean   :48.57
```

```
##  3rd Qu.:1.0000   3rd Qu.:1.0000                       3rd Qu.:62.00
##  Max.   :1.0000   Max.   :1.0000                       Max.   :93.00
##                   NA's   :20                            NA's   :22
##     State               Region         Conservativeness Info.On.Internet
##  Length:792         Length:792         Min.   :1.000    Min.   : 0.000
##  Class :character   Class :character   1st Qu.:3.000    1st Qu.: 2.000
##  Mode  :character   Mode  :character   Median :3.000    Median : 4.000
##                                        Mean   :3.237    Mean   : 3.795
##                                        3rd Qu.:4.000    3rd Qu.: 6.000
##                                        Max.   :5.000    Max.   :11.000
##                                        NA's   :45
##  Worry.About.Info Privacy.Importance Anonymity.Possible Tried.Masking.Identity
##  Min.   :0.0000   Min.   :  0.00     Min.   :0.0000     Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 41.43     1st Qu.:0.0000     1st Qu.:0.0000
##  Median :0.0000   Median : 68.75     Median :0.0000     Median :0.0000
##  Mean   :0.4886   Mean   : 62.85     Mean   :0.3692     Mean   :0.1633
##  3rd Qu.:1.0000   3rd Qu.: 88.89     3rd Qu.:1.0000     3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :100.00     Max.   :1.0000     Max.   :1.0000
##  NA's   :2        NA's   :5          NA's   :39         NA's   :8
##  Privacy.Laws.Effective
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2559
##  3rd Qu.:1.0000
##  Max.   :1.0000
##  NA's   :65
```

(g) What is the average number of pieces of personal information on the Internet, according to the `Info.On.Internet` variable?

- *Ans. 3.795*

```
mean(limited$Info.On.Internet)
```

```
## [1] 3.795455
```

(h) How many interviewees reported a value of 0 for `Info.On.Internet`? *Ans. 105.*

How many interviewees reported the maximum value of 11 for `Info.On.Internet`? *Ans. 8.*

```
sum(limited$Info.On.Internet==0)
```

```
## [1] 105
```

```
sum(limited$Info.On.Internet==11)
```

```
## [1] 8
```

```
#or you can also do
table(limited$Info.On.Internet)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11
## 105  84  95 101 104  94  67  63  40  18  13   8
```

(i) What proportion of interviewees who answered the `Worry.About.Info` question worry about how much information is available about them on the Internet?

- *Ans. 0.4886.*

```r
summary(limited$Worry.About.Info)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.4886  1.0000  1.0000       2
#or you can also do
mean(limited$Worry.About.Info, na.rm = TRUE)
```

```
## [1] 0.4886076
```

```r
table(limited$Worry.About.Info)
```

```
##
##   0   1
## 404 386
```

(j) What proportion of interviewees who answered the `Anonymity.Possible` question think it is possible to be completely anonymous on the Internet?

- *Ans. 0.3692*

```r
mean(limited$Anonymity.Possible, na.rm = TRUE)
```

```
## [1] 0.3691899
```

(k) Build a histogram of the age of interviewees. What is the best represented age group in the population - people aged around 20, people aged around 40, people aged around 60, people aged around 80?

- *Ans. From the histogram it seems people aged around 60 are best represented.*

```r
hist(limited$Age)
```

## Histogram of limited$Age
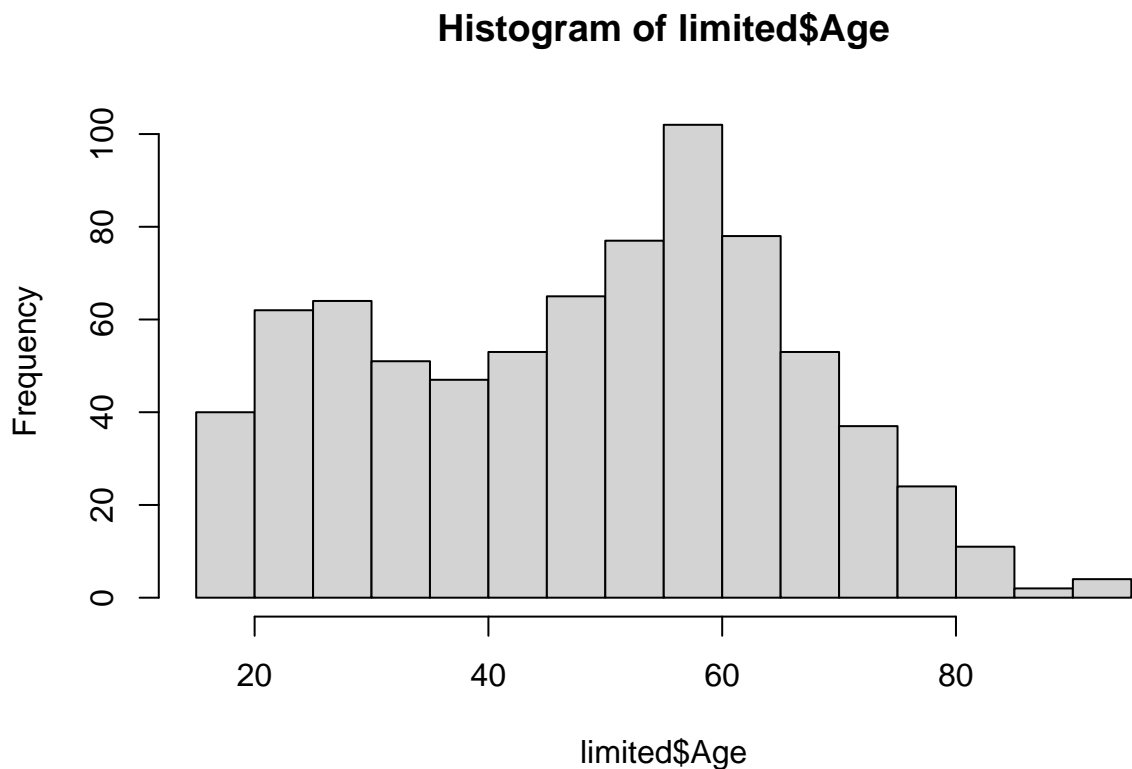


*We may also use* `ggplot` *to create the histogram as follows:*

```
library("ggplot2")
ggplot(limited)+geom_histogram(aes(Age),na.rm=T,binwidth=5,color="black",fill="lightblue")
```
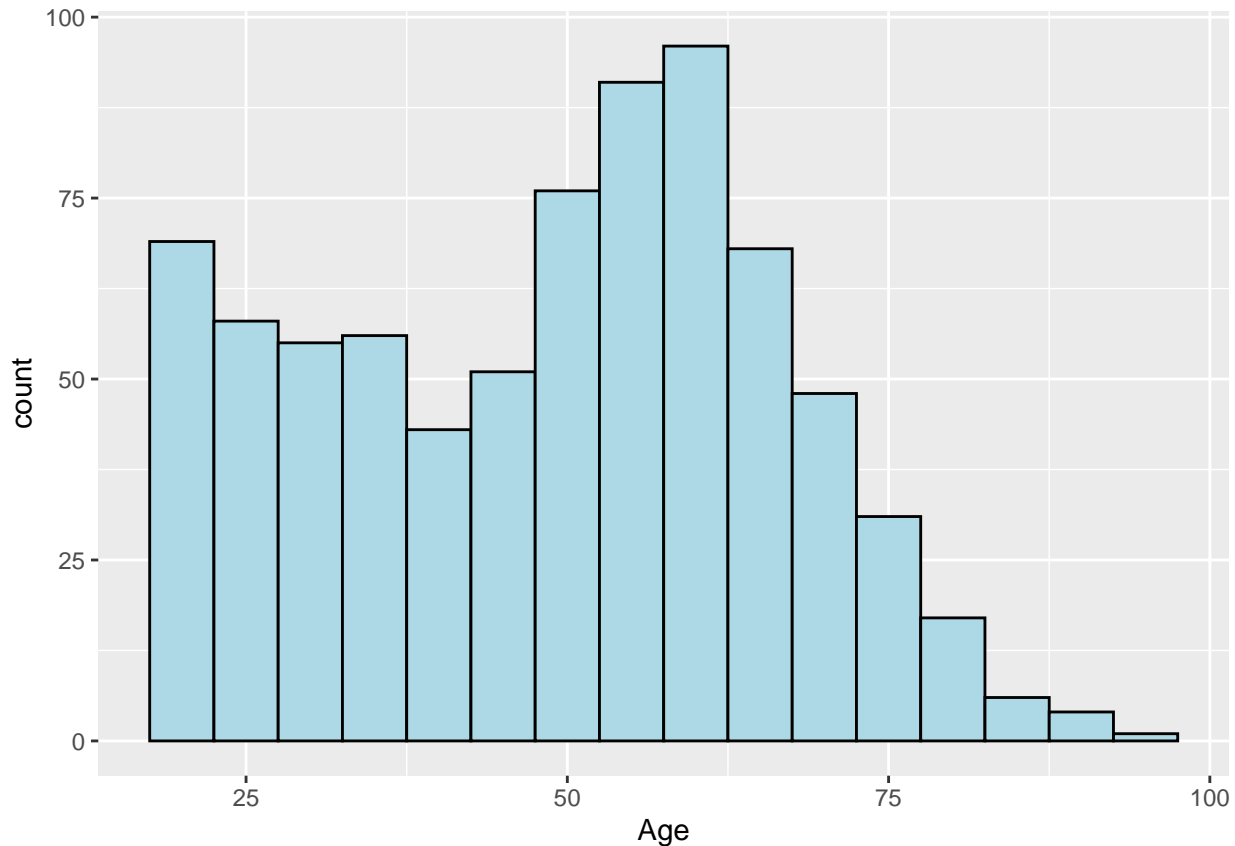


(l) Both `Age` and `Info.On.Internet` are variables that take on many values, so a good way to observe their relationship is through a graph. However, because `Info.On.Internet` takes on a small number of values, multiple points can be plotted in exactly the same location on this graph using the `plot()` function. What is the largest number of interviewees that have exactly the same value in their `Age` variable and the same value in their `Info.On.Internet` variable?

- *Ans. 6.*

```
max(table(limited$Info.On.Internet, limited$Age))
```

```
## [1] 6
```

(m) To avoid points covering each other up, we can use the `jitter()` function on the values we pass to the plot function. Experimenting with the command `jitter(c(1, 2, 3))`, what appears to be the functionality of the jitter command?

- *Ans. `jitter` adds some random zero mean noise to the vector.*

```
jitter(c(1,2,3))
```

```
## [1] 0.8275212 2.1584347 3.1707282
```

(n) Now, plot Age against Info.On.Internet with `plot(jitter(limited$Age), jitter(limited$Info.On.Internet))`. Comment on the relationship you observe between `Age` and `Info.On.Internet`?

- *Ans. From the jitter plot, older age is mildly associated with smalller values of the `Info.On.Internet`. However this does not seem to be partcularly strong.*

10

```
plot(jitter(limited$Age), jitter(limited$Info.On.Internet))
```
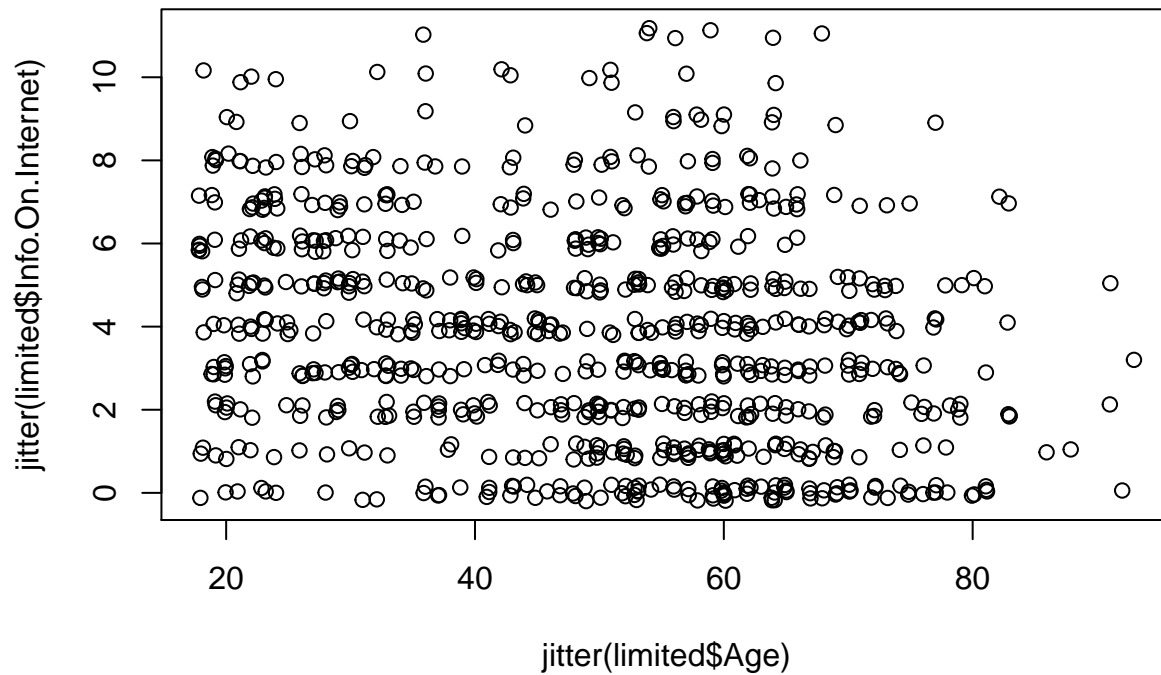


We may use `ggplot` to create a jitter plot.

```
ggplot(limited,aes(Age,Info.On.Internet)) +geom_point(position="jitter",na.rm=T)
```
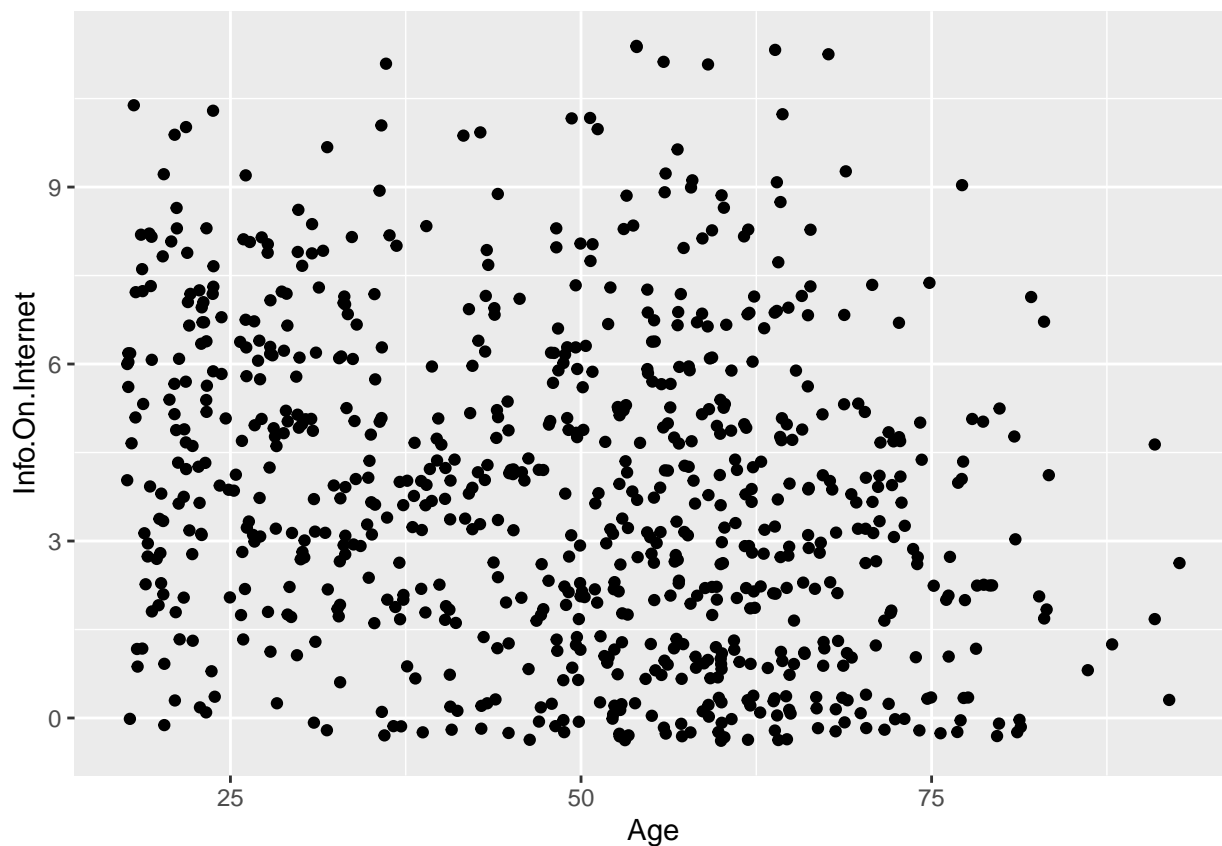


(o) Use the `tapply()` function to find the average of the Info.On.Internet value, depending on whether an

interviewee is a smartphone user or not? * *Ans. Average `Info.On.Internet` for smartphone users is 4.37 and for non-smartphone users it is 2.92.*

```r
tapply(limited$Info.On.Internet, limited$Smartphone, mean)
```

```
##        0        1
## 2.922807 4.367556
```

(p) Similarly use `tapply` to break down the Tried.Masking.Identity variable for smartphone and non-smartphone users.

- What proportion of smartphone users who answered the Tried.Masking.Identity question have tried masking their identity when using the Internet? *Ans. 0.1925.*
- What proportion of non-smartphone users who answered the `Tried.Masking.Identity` question have tried masking their identity when using the Internet? *Ans. 0.1174.*

```r
tapply(limited$Tried.Masking.Identity, limited$Smartphone, table)
```

```
## $`0`
##
##   0   1
## 248  33
##
## $`1`
##
##   0   1
## 390  93
```

```r
#or you can also do
tapply(limited$Tried.Masking.Identity, limited$Smartphone, mean, na.rm = TRUE)
```

```
##         0         1
## 0.1174377 0.1925466
```

7. In this question, we will investigate graphically the R internal dataset swiss using a different visualization tool. The data contains the variables:

- Fertility - common standardized fertility measure
- Catholic - % of catholics
- Agriculture - % of men working in agriculture environment
- Examination - % draftees receiving highest mark on army examination
- Education - % education beyond primary school for draftees
- Infant.Mortality - % of live births who live less than 1 year of 47 counties in the west of Switzerland dated at 1888. With ?swiss, you can get more information on the meaning of the variables.

(a) Read the help file of `stars()`. Make a star plot of all variables. What can you say about Sierre?

- *Ans. The county (canton) of Sierre has a high percentage of Catholics, high fertility rate and a large number of men working in agricluture (rural county). Education and examination percentages are low and so is the Infant mortality rate.*

```r
data(swiss)
head(swiss)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
```

```
## Porrentruy            76.1       35.3           9        7      90.57
##             Infant.Mortality
## Courtelary             22.2
## Delemont               22.2
## Franches-Mnt           20.2
## Moutier                20.3
## Neuveville             20.6
## Porrentruy             26.6
```
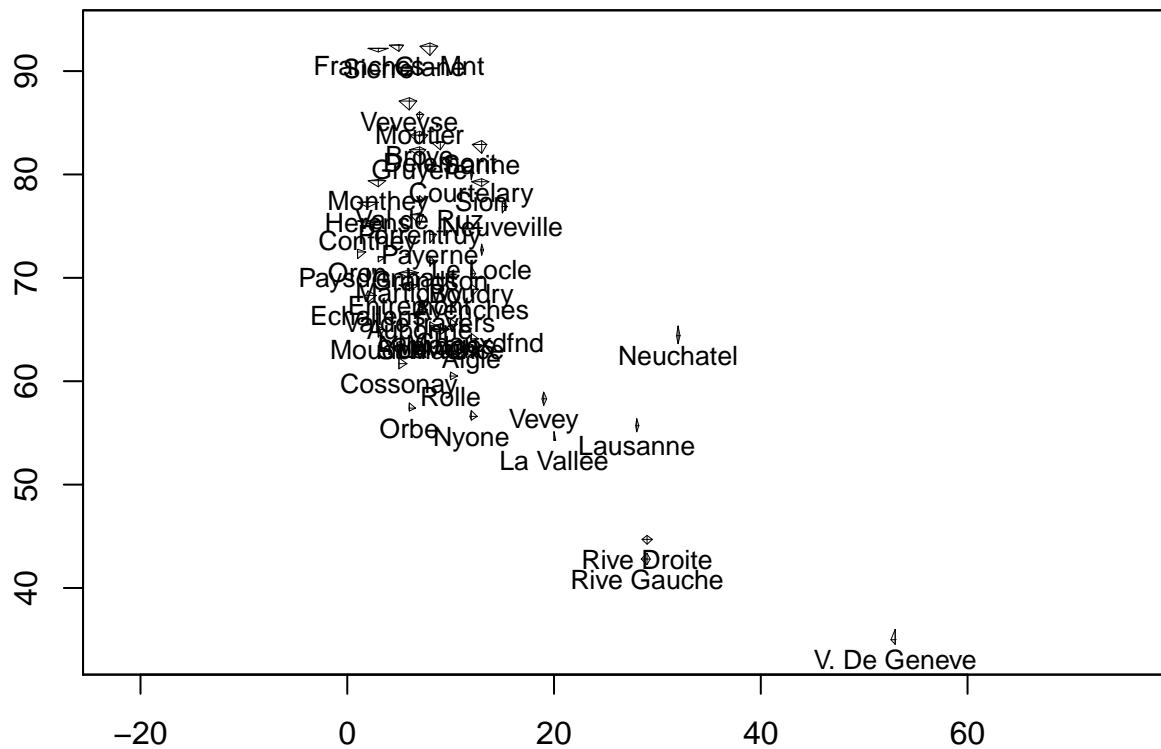
```
?stars
?swiss
stars(swiss, key.loc = c(18,2))
```
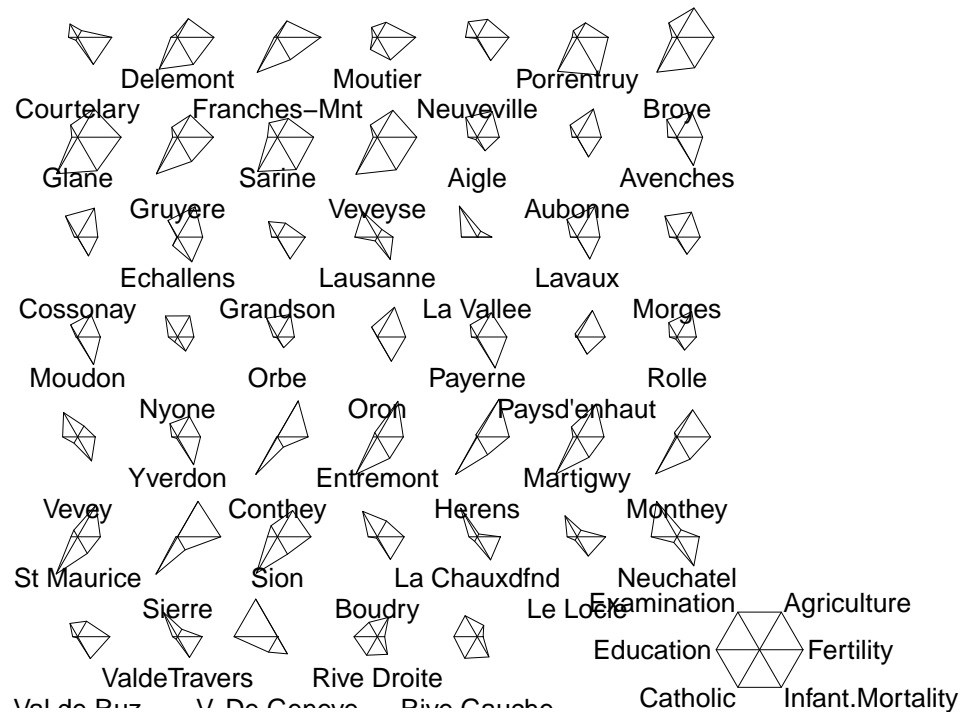


(b) We are interested in the relation between Fertility and Education. Therefore we would like to make a scatter-plot of Fertility against Education whose points are stars with the information of the other variables. In addition we need the argument location.

```
stars(as.matrix(swiss[,c(2,3,5,6)]), location = as.matrix(swiss[,c(4,1)]), axes = T)
```

```
stars(as.matrix(swiss[,c(2,3,5,6)]), location = as.matrix(swiss[,c(4,1)]), axes = T, labels = NULL, len
```

## Fertility against Education
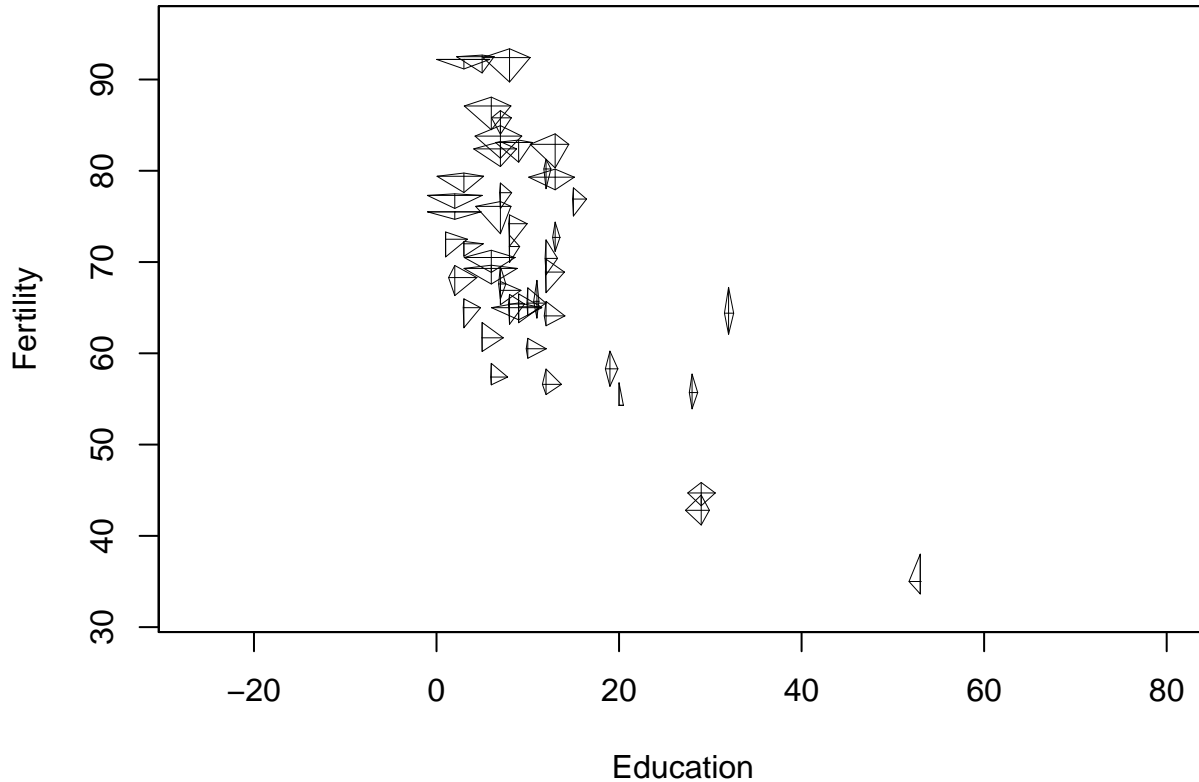


(c) Set the argument draw.segments to TRUE to get segments instead of stars. Place a legend with key.loc.

```
stars(as.matrix(swiss[,c(2,3,5,6)]), location = as.matrix(swiss[,c(4,1)]), axes = T, labels = NULL, len
```

**Fertility against Education**
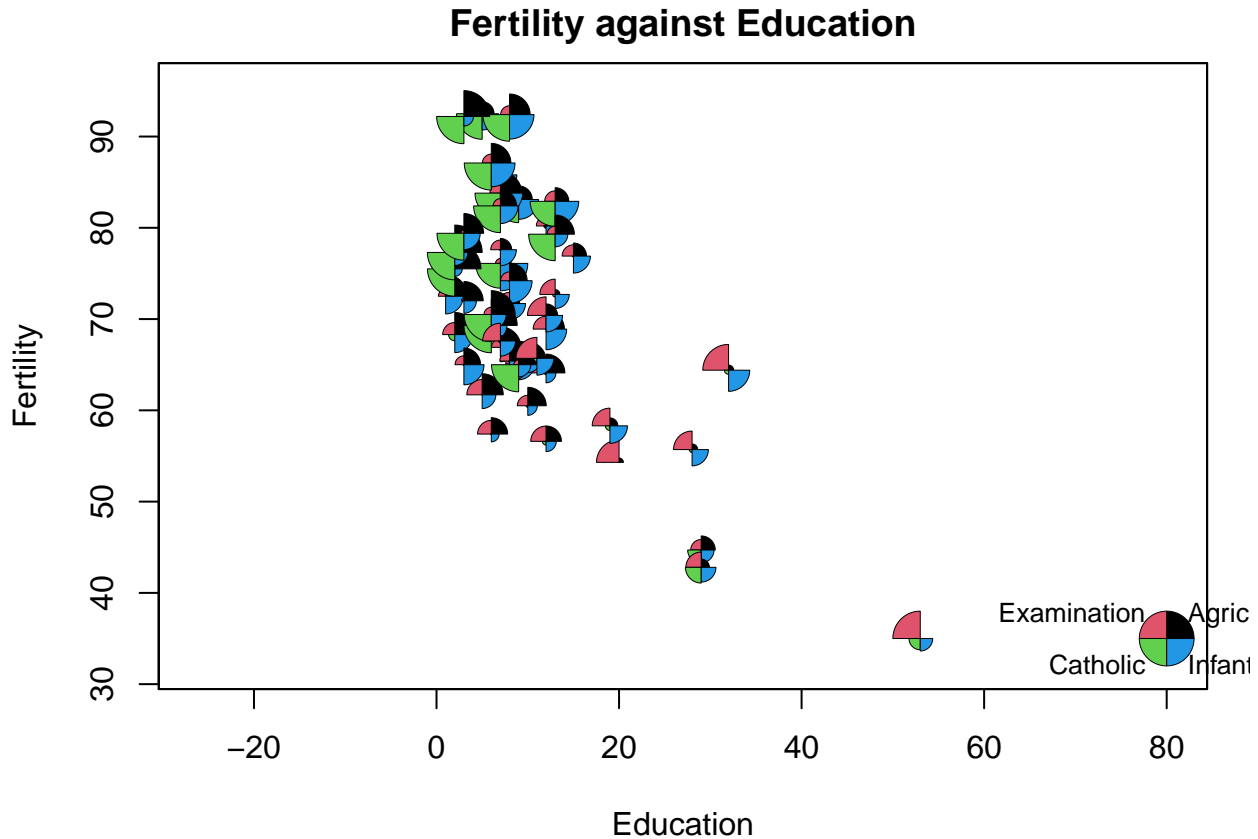


Education

(d) What relation do you get from the plots?

- *Ans. In counties with a high Catholic population and a large number of men working in agriculture, education levels are low and fertility is very high. There is no clear pattern fro infant mortality as such. Areas where there are more Protestants than Catholics, there seems to be higher levels of education*

8. In this question, we will visualize the attributes of parole violators from a dataset. In many criminal justice systems around the world, inmates deemed not to be a threat to society are released from prison under the parole system prior to completing their sentences. They are still considered to be serving their sentences while on parole and they can be returned to prison if they violate the terms of their parole. Parole boards use data on parole violators to better understand whether to approve or deny an application for parole. The dataset `Parole.csv` has the following fields:

- Male = 1 if the parolee is male, 0 if female
- Racewhite = 1 if the parolee is white, 0 otherwise
- Age = The parolee's age in years at the time of release from prison
- State = The parolee's state (Kentucky, Louisiana, Virginia, and Other). The first three states were selected due to having a high representation in the dataset.
- TimeServed = The number of months the parolee served in prison (limited by the inclusion criteria to not exceed 6 months).
- MaxSentence = The maximum sentence length for all charges, in months (limited by the inclusion criteria to not exceed 18 months).
- MultipleOffenses = 1 if the parolee was incarcerated for multiple offenses, 0 otherwise.
- Crime = The parolee's main crime leading to incarceration (Larceny, Drugs, Driving, and Other).
- Violator = 1 if the parolee violated their parole, and 0 if the parolee completed the parole without violation. In this question, we will visualize the attributes of parole violators using histograms with the ggplot2 package. We'll learn how to use histograms to show counts by one variable, and then how to

visualize 3 dimensions by creating multiple histograms.

(a) Read the data into a dataframe called `Parole`.

- What fraction of parole violators are female?
- *Ans. 14/(14+64)= 0.1794.*

```
Parole <- read.csv("Parole.csv")
str(Parole)
```

```
## 'data.frame':    675 obs. of  9 variables:
##  $ Male            : int  1 0 1 1 1 1 1 0 0 1 ...
##  $ RaceWhite       : int  1 1 0 1 0 0 1 1 1 0 ...
##  $ Age             : num  33.2 39.7 29.5 22.4 21.6 46.7 31 24.6 32.6 29.1 ...
##  $ State           : chr  "Other" "Other" "Other" "Other" ...
##  $ TimeServed      : num  5.5 5.4 5.6 5.7 5.4 6 6 4.8 4.5 4.7 ...
##  $ MaxSentence     : int  18 12 12 18 12 18 18 12 13 12 ...
##  $ MultipleOffenses: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Crime           : chr  "Driving" "Drugs" "Drugs" "Other" ...
##  $ Violator        : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
table(Parole$Violator,Parole$Male)
```

```
##
##        0    1
##   0  116  481
##   1   14   64
```

(b) In this dataset, which crime is the most common in Kentucky? *Ans. Drugs.*

```
table(Parole$State,Parole$Crime)
```

```
##
##             Driving Drugs Larceny Other
##   Kentucky        4    64      10    42
##   Louisiana       5    20      15    42
##   Other          34    34       9    66
##   Virginia       58    35      72   165
```

(c) In the **ggplot2** package, we need to specify a dataset, aesthetic, and geometry while creating visualizations. To create a histogram, the geometry will be geom histogram. Create a histogram to find out the distribution of the age of parolees, by typing the following command in your R console: `ggplot(data = Parole, aes(x = Age)) + geom histogram()`

```
ggplot(data = Parole, aes(x = Age)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```
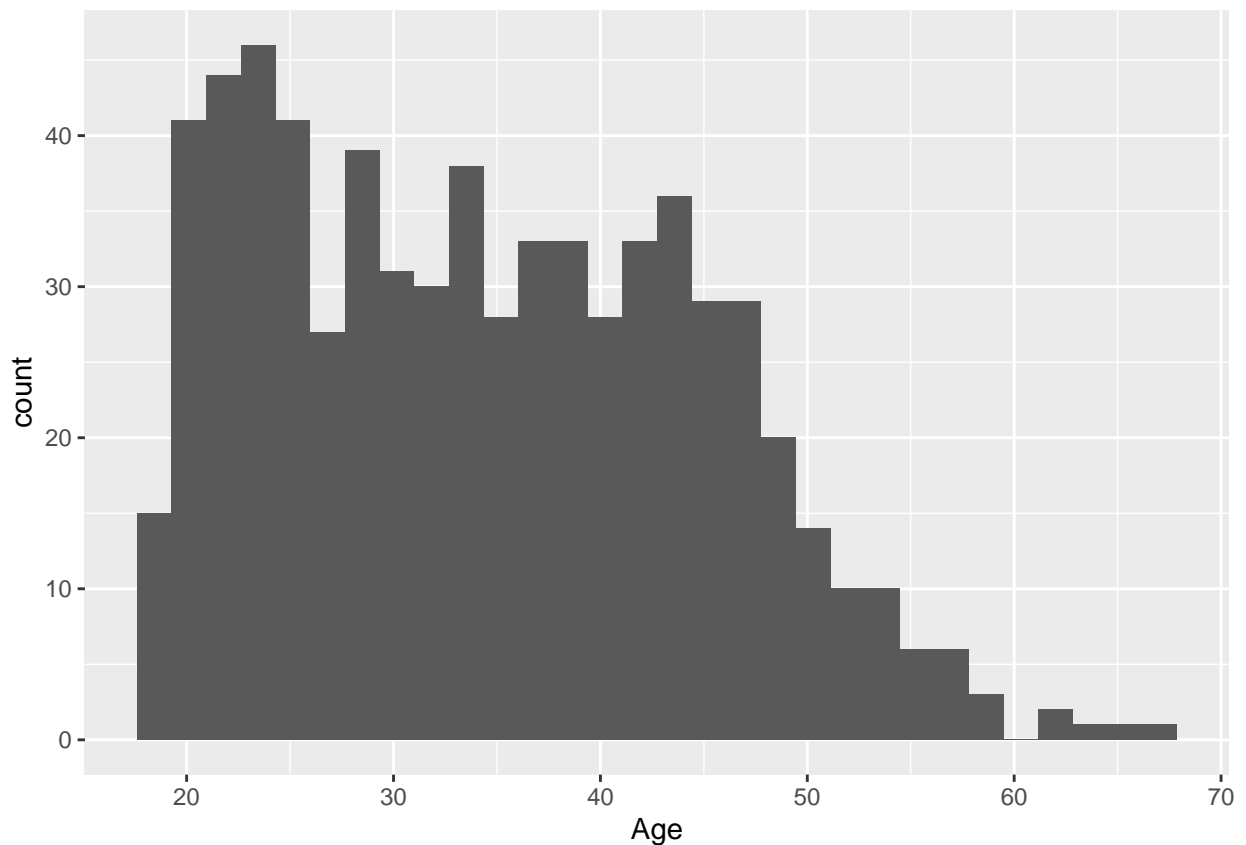
(d) By default, `geom_histogram` divides the data into 30 bins. Change the width of the bins to 5 years by adding the argument binwidth = 5. Also set the center of one of the bins to 17.5 by adding the argument center = 17.5. Also define the argument `closed = c(``left")` to indicate that left endpoint is included in the bin, but the right endpoint isn't. Which among these age brackets has the most parolees?

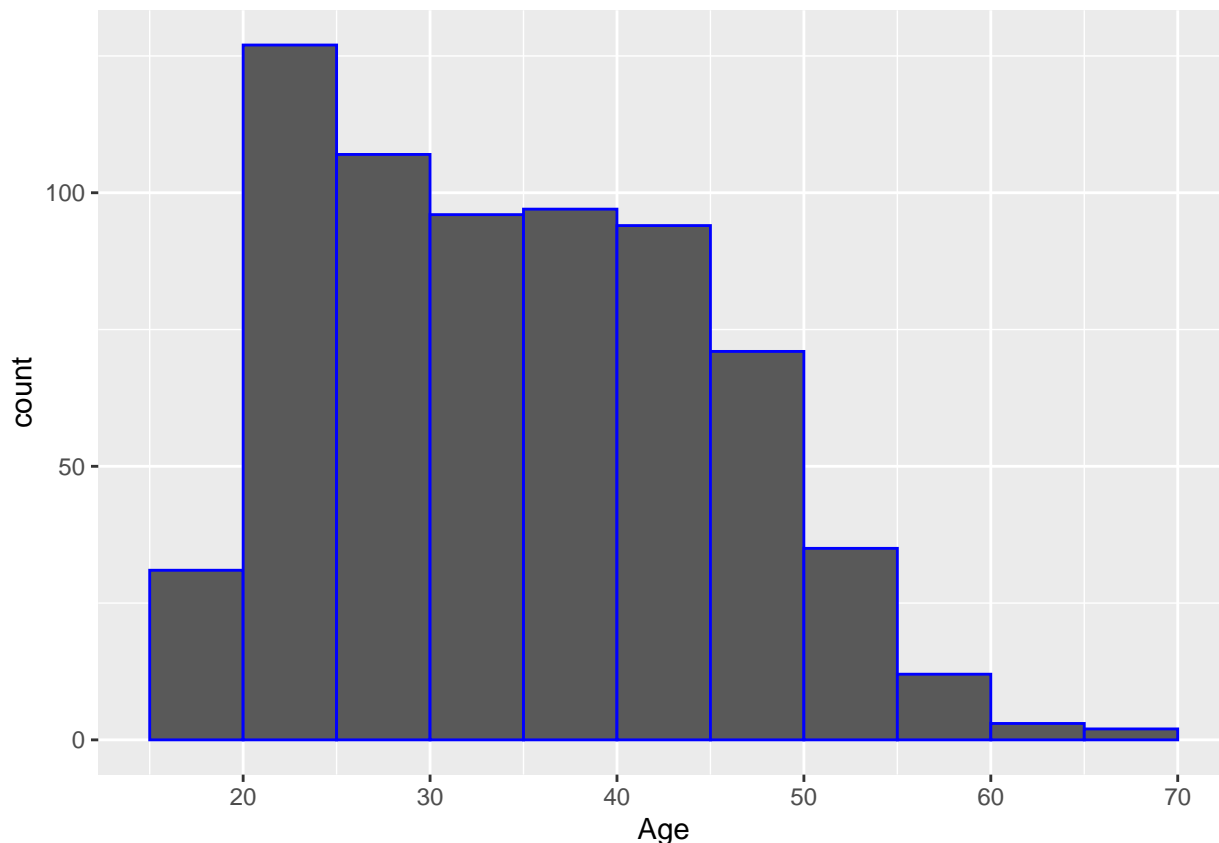- [20, 25) *(Correct answer)*
- [25, 30)
- [30, 35)
- [35, 40)

```
ggplot(data = Parole, aes(x = Age)) + geom_histogram(binwidth=5,closed=c("left"),center=17.5)
```
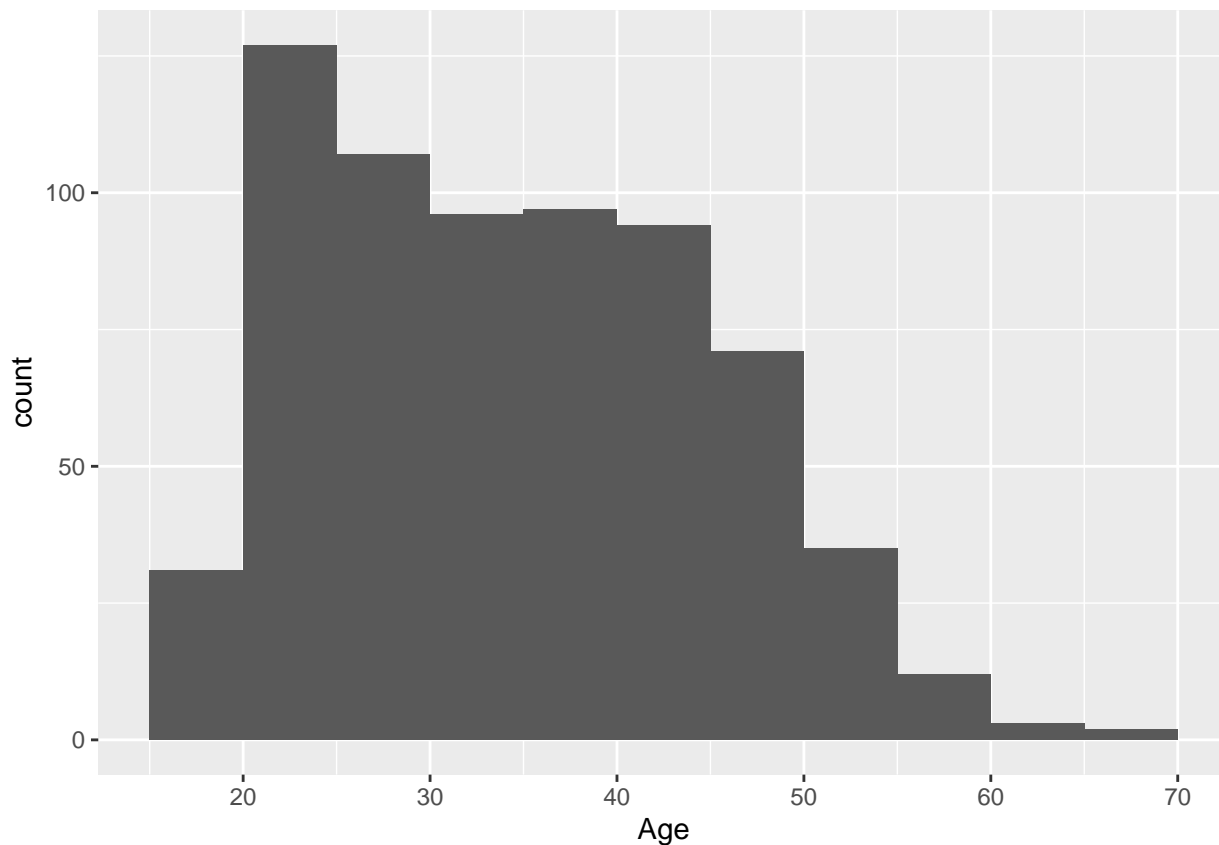
(e) Redo the histogram by adding the argument `color = c(``blue")` to `geom_histogram`. What does this argument do?

- Changes the fill color of the bars
- Changes the background color of the plot
- Changes the outline color of the bars *(Correct answer)*
- Changes the color of the axis labels

```
ggplot(data = Parole, aes(x = Age)) + geom_histogram(binwidth=5,closed=c("left"),center=17.5,color=c("b
```

(f) Now suppose we are interested in seeing how the age distribution of male parolees compares to the age distribution of female parolees. One option would be to create a heatmap with Age on one axis and Male (a binary variable in our data set) on the other axis. Another option would be to stick with histograms, but to create a separate histogram for each gender. ggplot has the ability to do this automatically using the facet grid command. To create separate histograms for male and female, type the following command into your R console: `ggplot(data = Parole, aes(x = Age)) + geom_histogram(binwidth=5,closed=c(``left"),center=17.5,color=c(``blue"))+facet grid(Male~.)` The histogram for female parolees is on the top and the male parolees is on the bottom. What is the age bracket with the most female parolees?

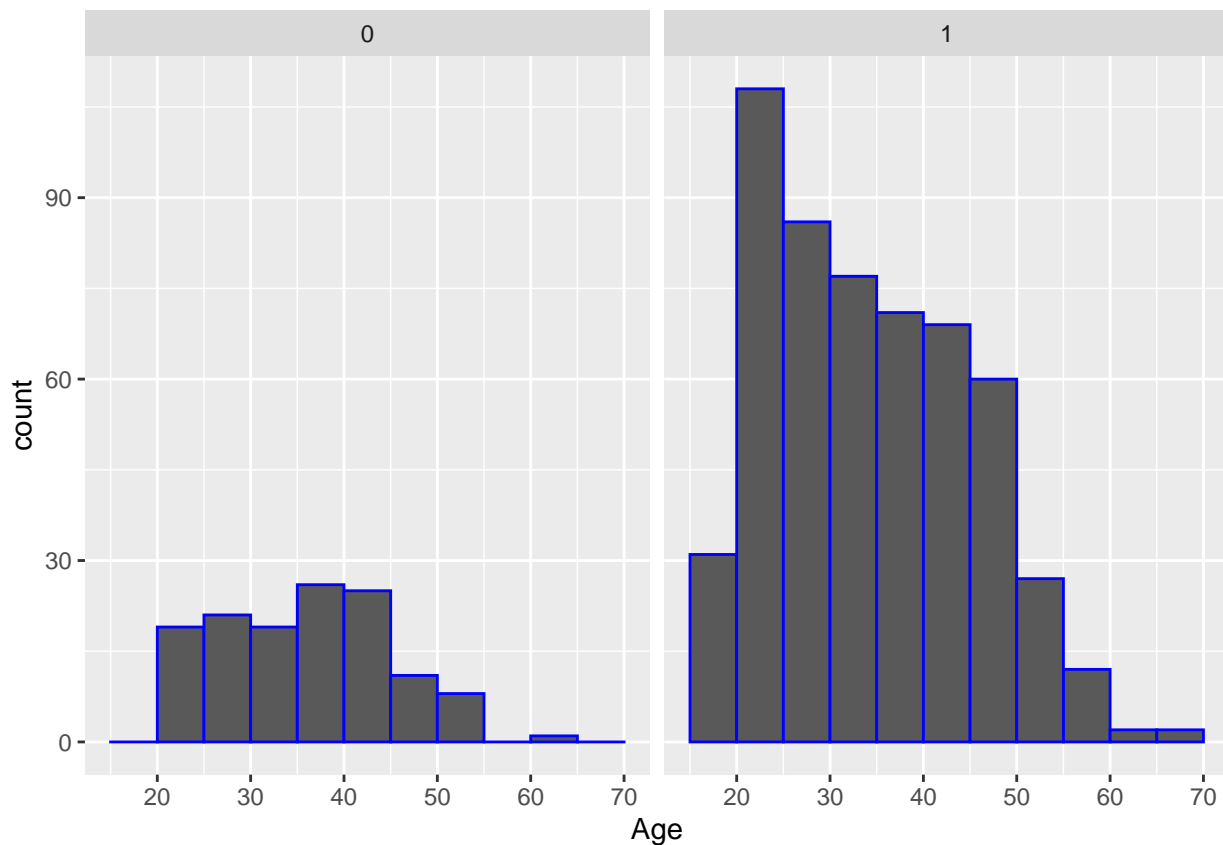- [20, 25)
- [25, 30)
- [30, 35)
- [35, 40) *(Correct answer)*

```
ggplot(data = Parole, aes(x = Age)) + geom_histogram(binwidth=5,closed=c("left"),center=17.5,color=c("bl
```
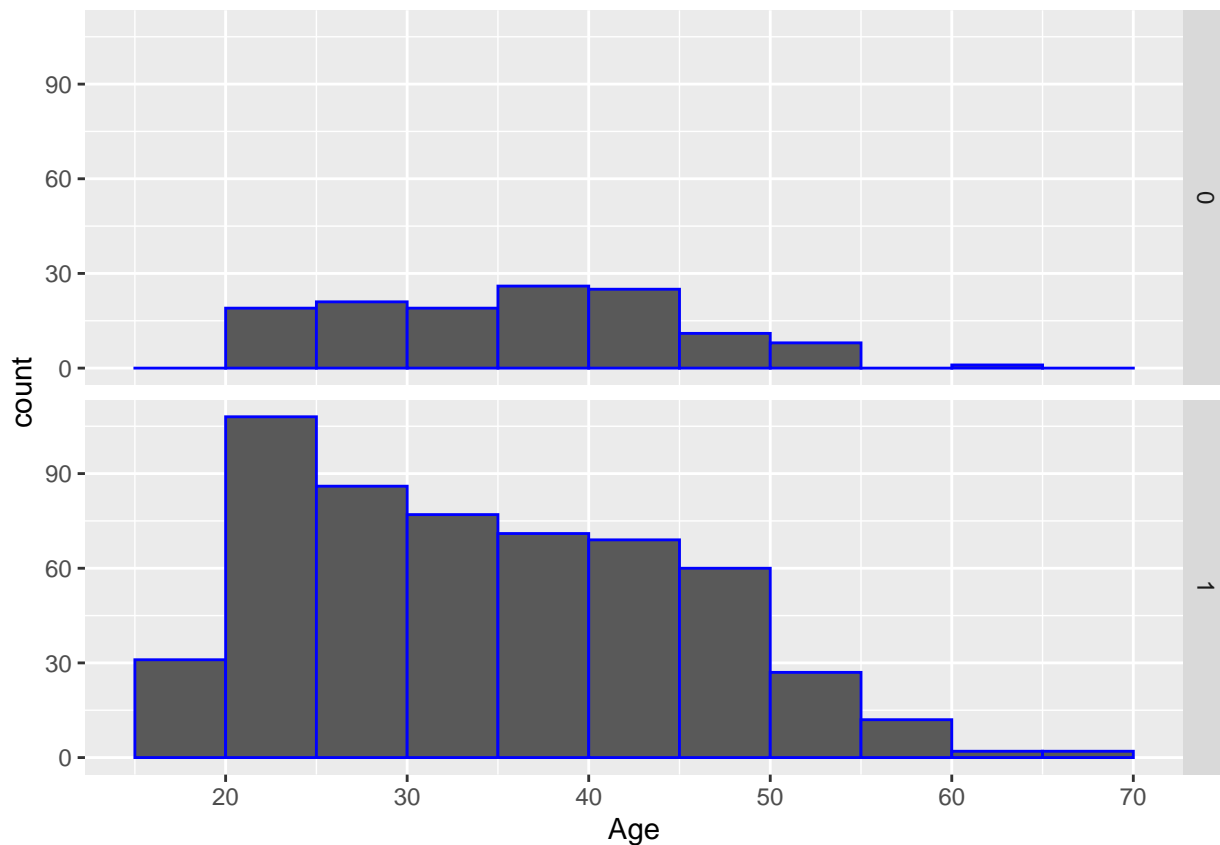
(g) Now change the facet grid argument to facet `grid(.~Male)`. What does this do?

- Creates histograms of the Male variable, sorted by the different values of age.
- Puts the histograms side-by-side instead of on top of each other. *(Correct answer)*
- Puts the histogram for male parolees on the top.
- This doesn't change anything - the plot looks exactly the same as it did before.

```
ggplot(data = Parole, aes(x = Age)) + geom_histogram(binwidth=5,closed=c("left"),center=17.5,color=c("bl
```

(h) An alternative choice to creating separate histograms is to color the groups differently. To do this, we need to tell ggplot that a property of the data (male or not male) should be translated to an aesthetic property of the histogram. We can do this with the fill parameter as follows: `ggplot(data = Parole, aes(x = Age,fill = as.factor(Male))) + geom histogram(binwidth=5,closed=``left",center=17.5,color=c(``blue")))` Here we need to specify the fill argument as a factor for the function to work. Create the new histogram.

```
ggplot(data = Parole, aes(x = Age,fill = as.factor(Male))) + geom_histogram(binwidth=5,closed=c("left")
```
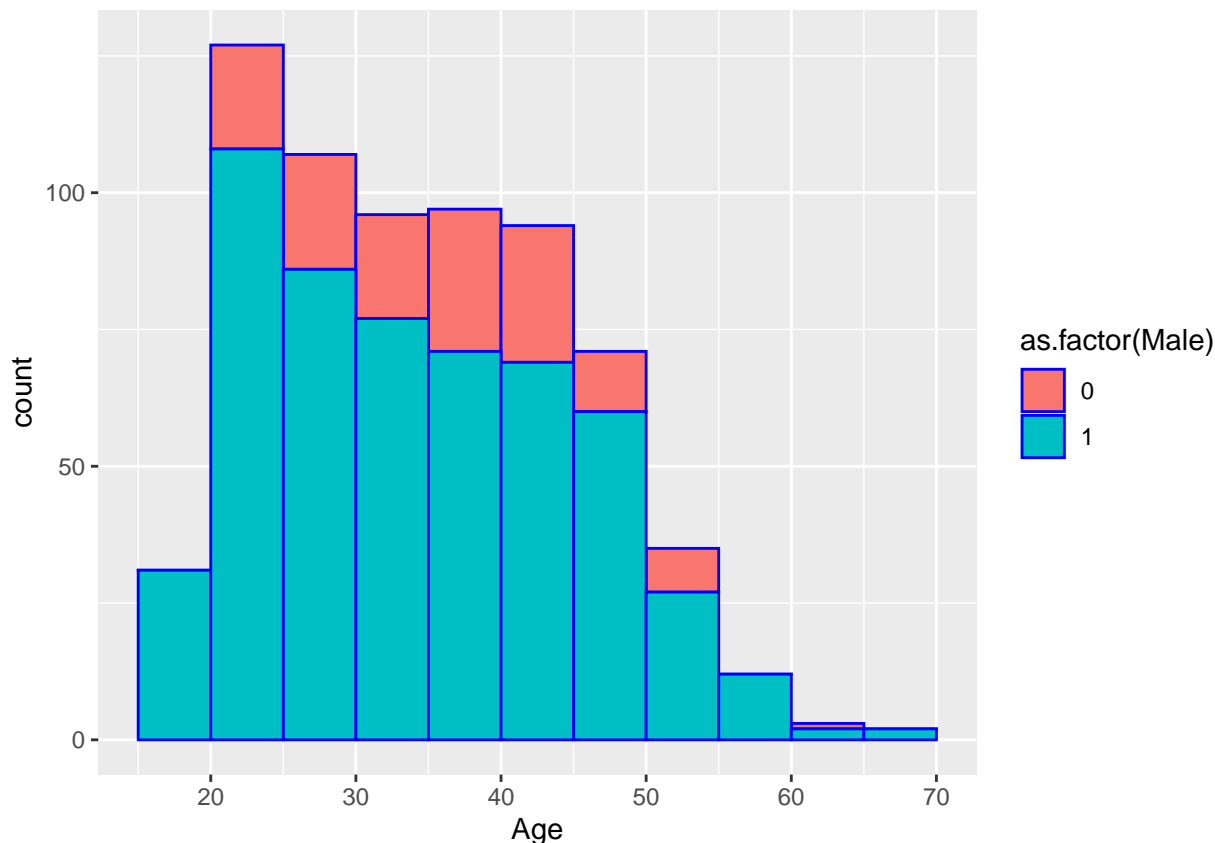
(i) Coloring the groups differently is a good way to see the breakdown of age by sex within the single, aggregated histogram. However, the bars here are stacked, meaning that the height of the bars in each age bin represents the total number of parolees in that age bin, not just the number of parolees in that group. An alternative to a single, stacked histogram is to create two histograms and overlay them on top of each other. This is a simple adjustment to our previous command. We just need to 1) Tell ggplot not to stack the histograms by adding the argument `position="identity"` to the `geom_histogram` function and 2) Make the bars semi-transparent so we can see both colors by adding the argument alpha=0.5 to the geom histogram function. The new arguments prevent the bars from being stacked and make them semi-transparent. Redo the plot, making both of these changes. Which of the following buckets contain no female paroles? Choose all that apply:

- [15, 20)
- [20, 25)
- [25, 30)
- [30, 35)
- [35, 40)
- [40, 45)
- [45, 50)
- [50, 55)
- [55, 60)
- [60, 65)
- [65, 70)

*– Ans. Age brackets with no female parolee: [15, 20), [55,60), [65,70).*

```
ggplot(data = Parole, aes(x = Age,fill = as.factor(Male))) + geom_histogram(binwidth=5,closed=c("left")
```
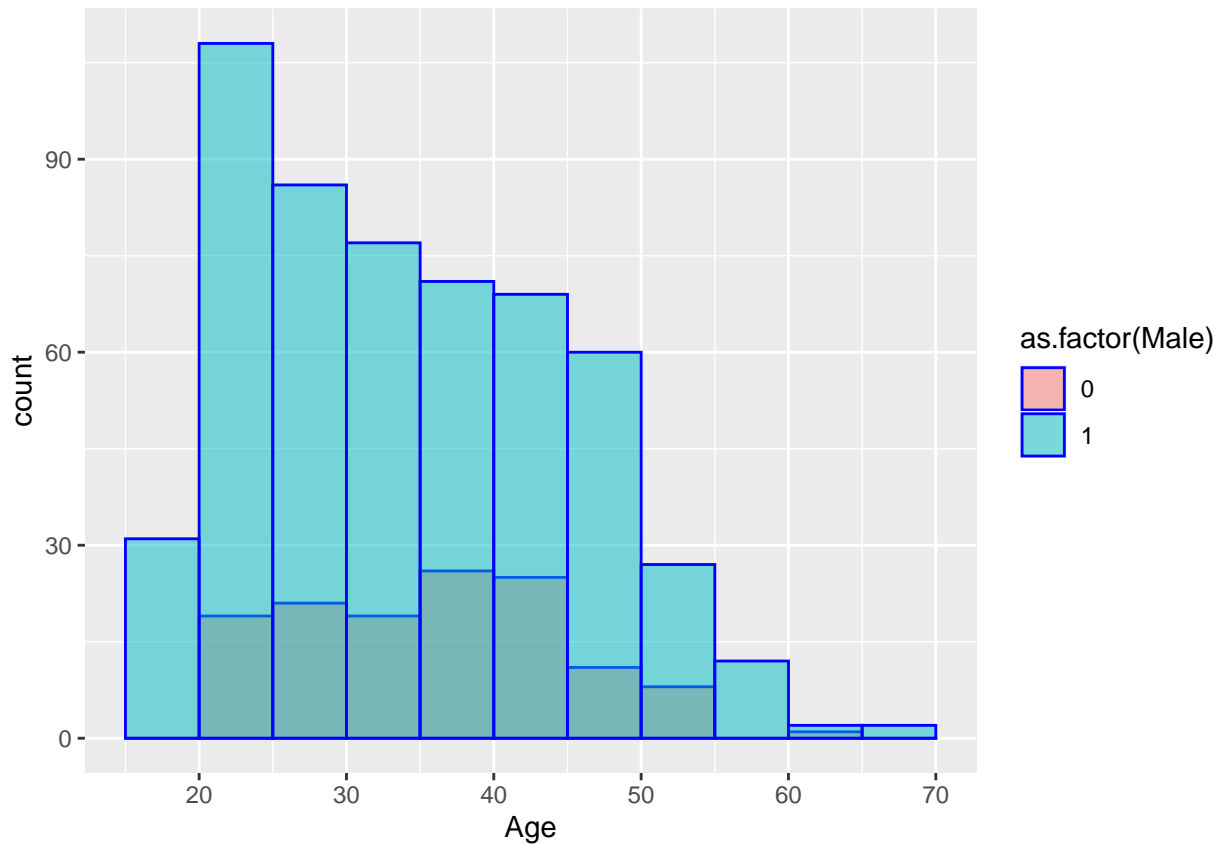


(j) Which of the histograms (faceting or overlaying) do you think better visualizes the data? Why?

- *Ans. Overlaying seems to be a better alternative to faceting, since it becomes easier to compare between factors.*

(k) Now let us explore the amount of time served by parolees. Create a basic histogram as in part (c) but with `TimeServed` on the x-axis. Set the binwidth to 1 month, center to 0.5 and closed to "right". What is the most common length of time served according to this histogram?

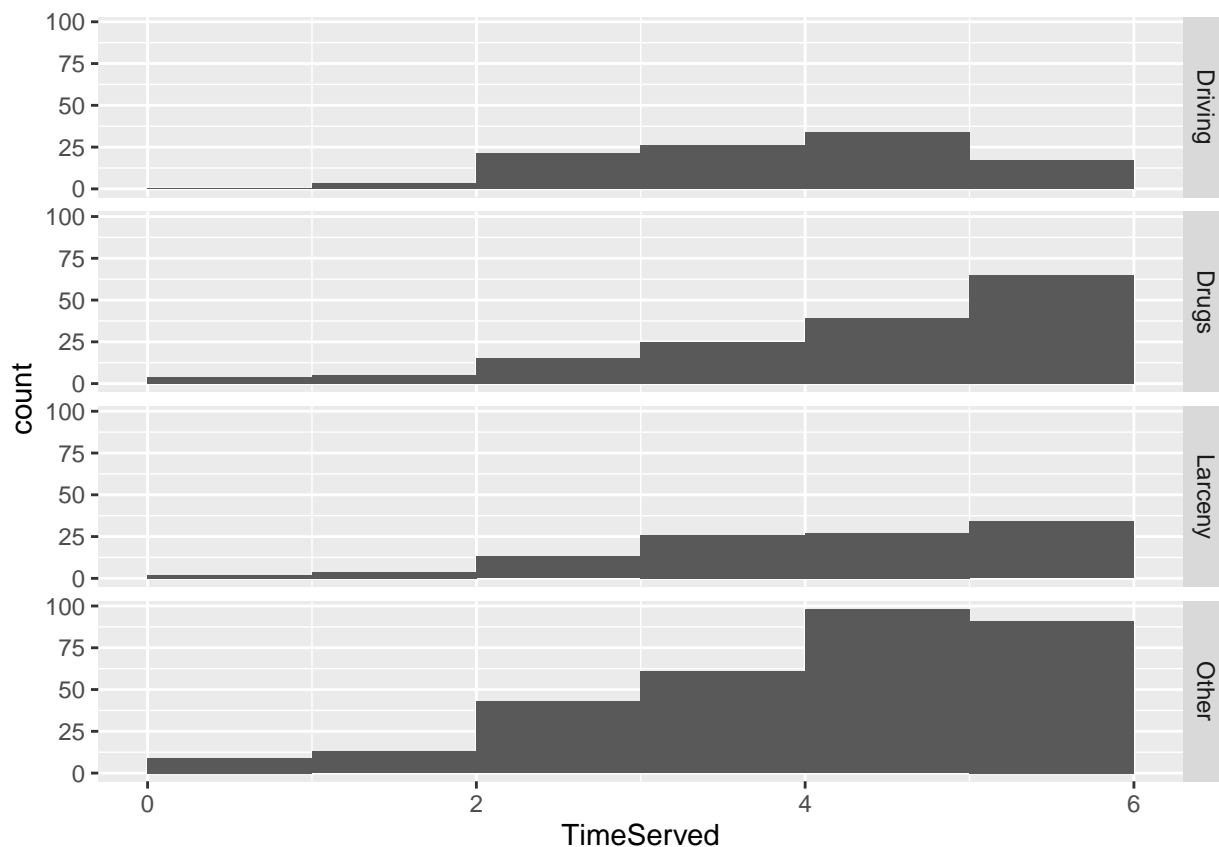- *Ans. In this case, the most common length is 5 to 6 months.*

```
ggplot(data = Parole, aes(x = TimeServed)) + geom_histogram(binwidth=1,center=0.5,closed="right")
```
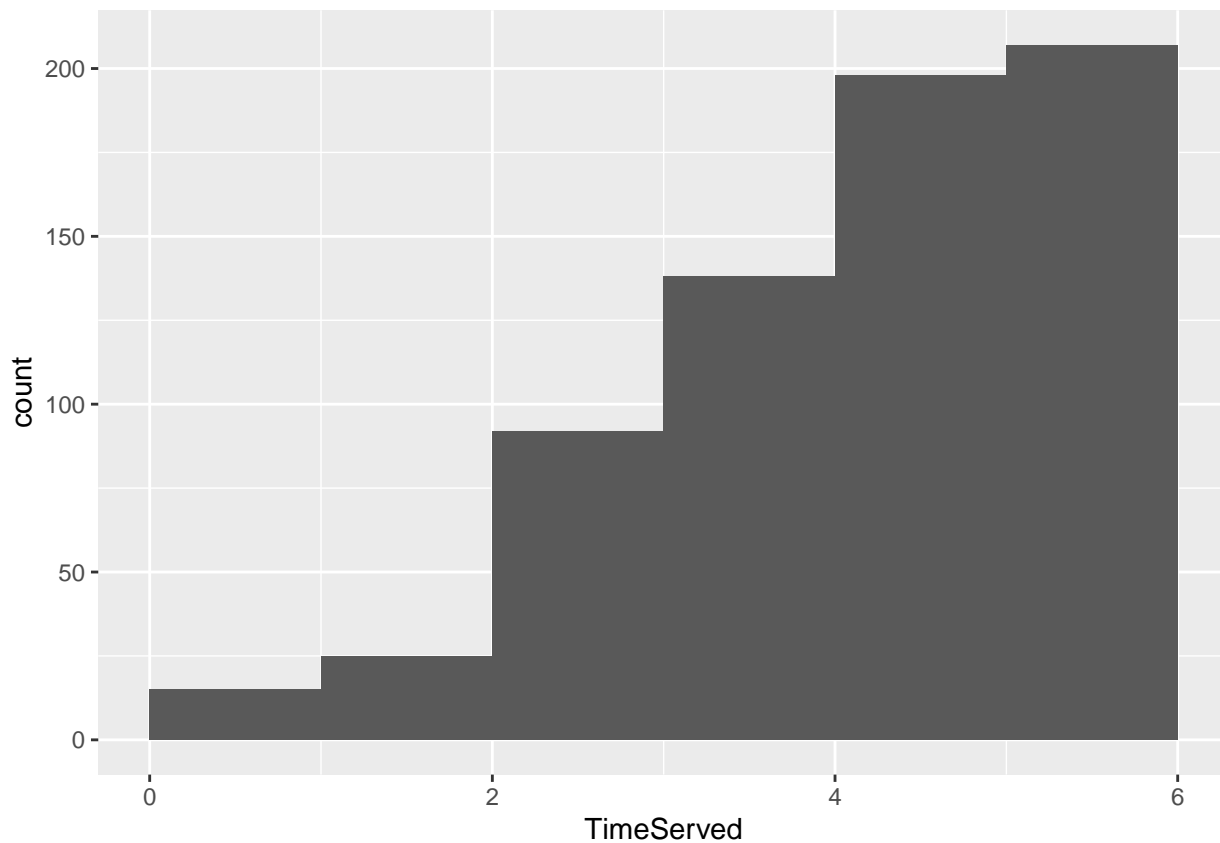
(l) Now, suppose we suspect that it is unlikely that each crime has the same distribution of time served. To visualize this change use facet grid to create a separate histogram of `TimeServed` for each value of the variable `Crime`. Which crime type has no observations where time served is less than one month?

- Drug
- Driving *(Correct answer)*
- Larceny
- Other

```
ggplot(data = Parole, aes(x = TimeServed)) + geom_histogram(binwidth=1,center=0.5,closed="right")+facet_
```

(m) Now instead of faceting the histogram, overlay them. Remember to set the position and alpha parameters so that histograms are not stacked. Also make sure to indicate the fill aesthetic is Crime. In this case, faceting seems like a better alternative. Why?

- With four different groups, it can be hard to tell them apart when they are overlayed, especially if they have similar values. *(Correct answer)*
- ggplot doesn't let us overlay plots with more than two groups.
- Overlaying the plots doesn't allow us to observe which crime type is the most common.

```
ggplot(data = Parole, aes(x = TimeServed,fill=Crime)) + geom_histogram(binwidth=1,center=0.5,closed="rig
```