Midterm Sol

1. (16 pts) The Behavioral Risk Factor Surveillance Survey: is conducted biannually. We have the data from 1970 and 1990. The following measurements were made on each individual surveyed:

   - `weight`: measured in pounds, with a value of 999 for missing
   - `height`: measured in inches
   - age: measured in years
   - `daysSick`: number of days in the past 30 days that health was not good. Values include 1 to 30. None is coded as 88.
   - `lastCheckup`: time since last routine checkup. Values are 1 for within a year, 2 for more than 1 and less than 2 years, 3 for more than 2 and less than 5 years, 4 for more than 5 years, and NA for don't know.
   - phone: the respondent was contacted by land line (1) or cell phone (2).

   The data are in a data frame called `BRFSS`. To prepare the data for analysis, write code to perform each of the following operations.

   (a) Recode the number of days sick so that 88 is 0.

```
BRFSS$daysSick[BRFSS$daysSick==88] = 0
or  BRFSS[BRFSS$daysSick==88, 'daysSick'] = 0
```

   (b) Convert a weight of 999 to NA.

```
BRFSS$weight[BRFSS$weight==999] = NA
```

   (c) Turn `lastCheckup` into a factor with appropriate labels for the levels.

Note levels = c(1:4, NA) is an error

```
BRFSS$lastCheckup = factor(BRFSS$lastCheckup,
    labels = c("< 1yr", "1 to 2 yrs", "2 to 5yrs", "> 5yrs"))
```

   (d) Drop all records from the data frame that have a value of NA for the time of the last check up. (Assign the smaller data frame to `BRFSS2`).

```
BRFSS2 = BRFSS[!is.na(BRFSS$lastCheckup), ]
```

2. (20 pts) Write a function called qcd(), short for quartile coefficient of dispersion. This function takes two arguments: the required x, which is a numeric vector; and the optional `na.rm`, which indicates whether NAs should be removed x. The default value for this argument is FALSE. The function returns the single numeric value that is the ratio of the interquartile range to the sum of the lower and upper quartiles. In addition, check that the input for x is numeric and if not terminate execution and provide an informative error message.

```
qcd = function(x, na.rm = FALSE){
  if (!is.numeric(x)) {
      stop("x must be numeric")
  }
  qs = quantile(x, probs = c(0.25, 0.75),
                na.rm = na.rm)

  return((qs[2]-qs[1])/sum(qs))
}
```

3. (16 pts) What is the value printed to the console for each of the following expressions.

(a)
```
> x = c(-1, -1, -1, -1, 1, -1)
> y = cumsum(x*3)
> y
```

w/out * 3

-3 -6 -9 -12 -9 -12	-1 -2 -3 -4 -3 -4

(b)
```
> z = which(y < -10)
> z
```

4 6	integer(0)

(c)
```
> z[length(z)]
```

length(z) is 0

6	integer(0)

(d)
```
> x[z]
```

-1 -1	numeric(0)

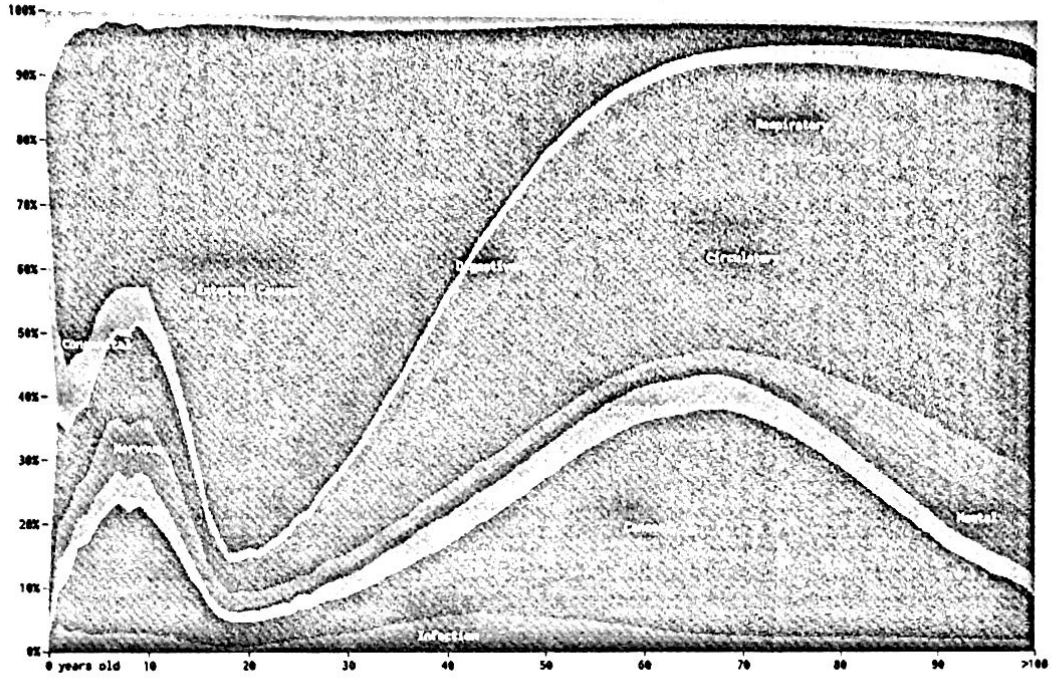
4. (12 pts) For the following function call show the steps in the computations that R carries out.

```
afunc(1:4, 10:13)
```

afunc = function(x, y) {	1	2	3	4	5	6	7	8	9	10
if (length(x) != length(y)) {	X									
stop("x and y must be the same length")										
}										
if (any(x < 0 | y < 0)) {		X								
bump = -(min(x) + min(y))										
} else {										
bump = min(1, min(x), min(y))			X							
}										
return(x + y + bump)				X						
}										

ok to put X in } else { row and at } is

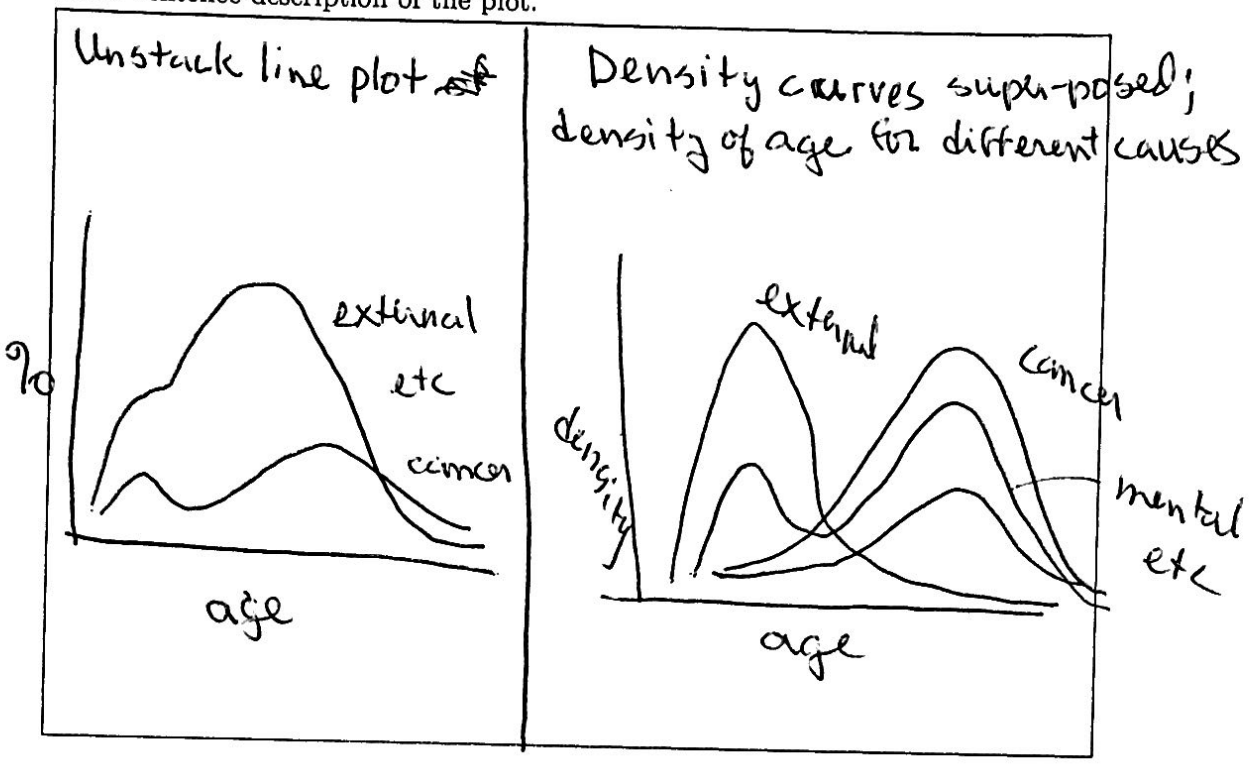
5. (20 pts) Consider the plot below. It exams the causes of death among males at various ages. These data are from the Centers for Disease Control and Prevention data base. In answering the questions be sure to use the graphics terminology introduced in the course.

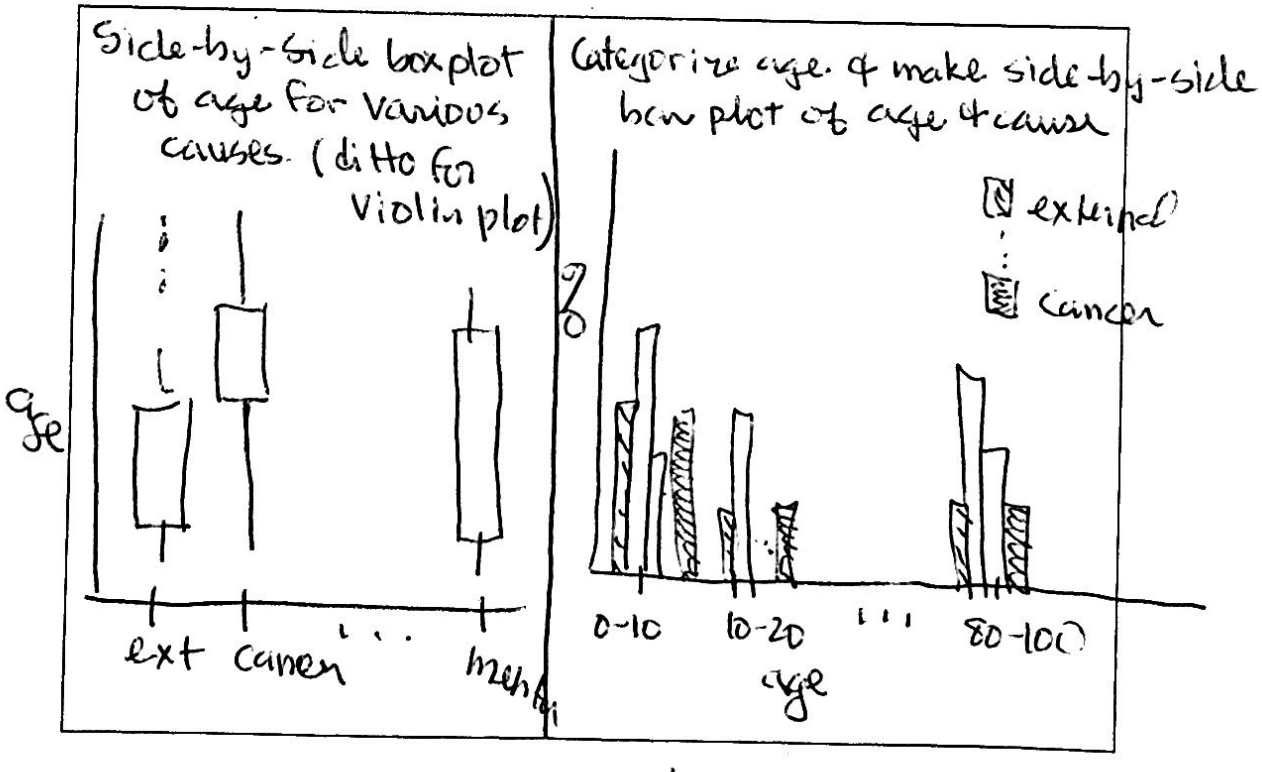


(a) What are the aesthetics in this plot? For each aesthetic describe the variable that it maps to and the data type of the variable.

x - age quantitative (numeric)
y - % quantitative (numeric)
fill - cause categorical (factor)

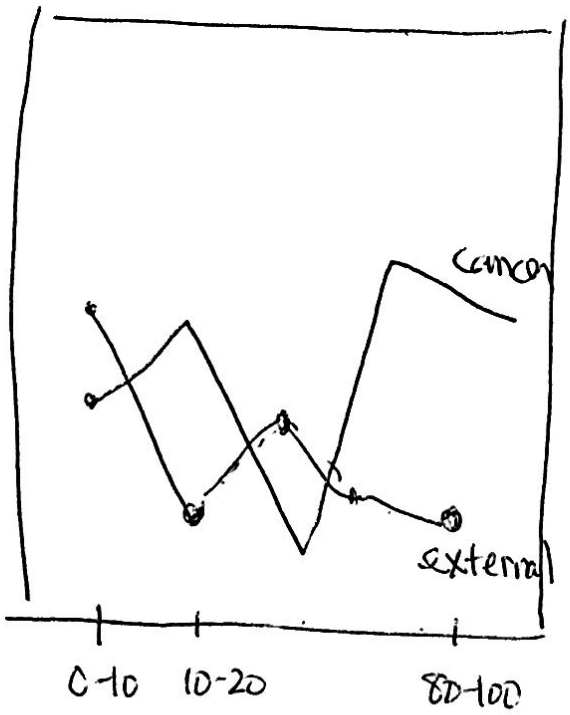
(b) Sketch two alternative plots that avoid stacking. There is no need to provide exact details; simply provide enough information that it is clear what you are plotting. Provide a one sentence description of the plot.





Categorizing age is less desirable

line plot for
categorical age &
cause



%

No Credit
For anything stacked
scatterplot
piechart
second plot nearly
identical to first

6. (16 pts) Consider the following list, aList:

```
aList
$x
[1] "a" "b" "c" "d" "e"

$mat
     [,1] [,2]
[1,]    8    5
[2,]    7    4
[3,]    6    3

$zz
$zz$x
[1] 1 2 3

$zz$y
[1]  7  8  9 10 11

$zz$z
[1]  TRUE  TRUE FALSE FALSE  TRUE

$one
[1] 100
```

Write down what will appear at the console when R evaluates each of the following expressions (note that some expressions may result in an error message):

(a) `length(aList)`

4 for x, mat, zz, $one

(b) `aList$mat + aList$one`

108 105
107 104
106 103

essentially matrix + 100

(c) `length(aList[["zz"]][1])`

1 return value from aList[["zz"]][1] is a list with one element (x)

(d) `sapply(aList$zz, min)`

1 7 0 logical vector coerced to numeric