

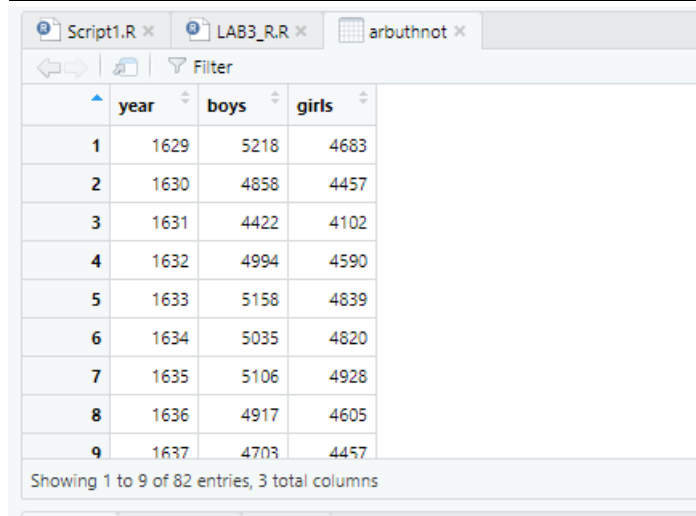
Compte Rendu

Lab3 : Introduction to Data

Part 1: Introduction to R:

Birth rates -boys vs girls

```
> source("http://www.openintro.org/stat/data/arbuthnot.R")
>
```



	year	boys	girls
1	1629	5218	4683
2	1630	4858	4457
3	1631	4422	4102
4	1632	4994	4590
5	1633	5158	4839
6	1634	5035	4820
7	1635	5106	4928
8	1636	4917	4605
9	1637	4703	4457

Showing 1 to 9 of 82 entries, 3 total columns

The Data: Dr. Arbuthnot Baptism Records

-We can take a look on this data also by typing its name:

```
> arbuthnot
  year boys girls
1 1629 5218 4683
2 1630 4858 4457
3 1631 4422 4102
4 1632 4994 4590
5 1633 5158 4839
6 1634 5035 4820
7 1635 5106 4928
8 1636 4917 4605
9 1637 4703 4457
10 1638 5359 4952
11 1639 5366 4784
12 1640 5518 5332
13 1641 5470 5200
14 1642 5460 4910
15 1643 4793 4617
16 1644 4107 3997
17 1645 4047 3919
18 1646 3768 3395
19 1647 3796 3536
20 1648 3363 3181
21 1649 3079 2746
22 1650 2890 2722
23 1651 3231 2840
24 1652 3220 2908
25 1653 3196 2959
26 1654 3441 3179
27 1655 3655 3349
28 1656 3668 3382
29 1657 3386 3280
```

-We can see the dimensions of this data frame by typing:

```
> dim(arbuthnot)
[1] 82  3
> |
```

This data containing 82 and 3 columns

-To see what the data frame contains we type this command:

```
> names(arbuthnot)
[1] "year" "boys" "girls"
> |
```

some Exploration:

-This command will only show the number of boys baptized each year

```
> arbuthnot$boys
[1] 5218 4858 4422 4994 5158 5035 5106 4917 4703 5359 5366 5518 5470 5460 4793 4107 4047 3768 3796 3363 3079 2890 3231 3220
[45] 6073 6113 6058 6552 6423 6568 6247 6548 6822 6909 7577 7575 7484 7575 7737 7487 7604 7909 7662 7602 7676 6985 7263 7632
> |
```

Script1.R x LAB3_R.R x arbuthnot\$boys x

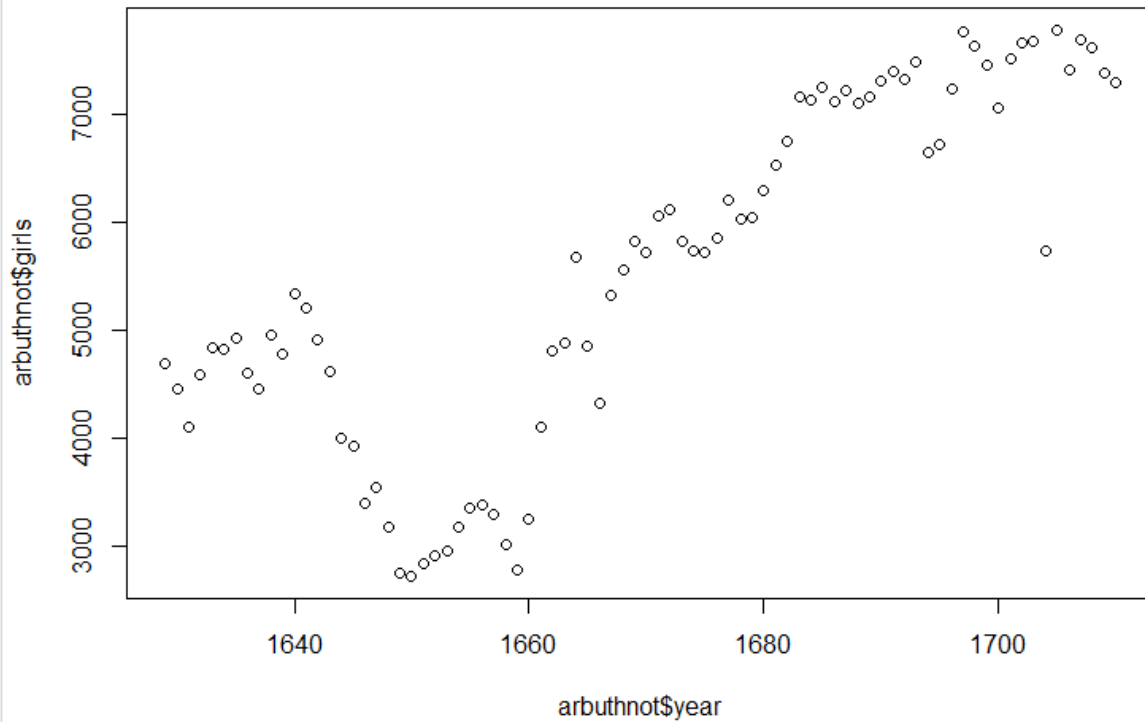
Filter

	V1
1	5218
2	4858
3	4422
4	4994
5	5158
6	5035
7	5106
8	4917
9	4703
10	5359
11	5366
12	5518
13	5470
14	5460
15	4793

Q1:

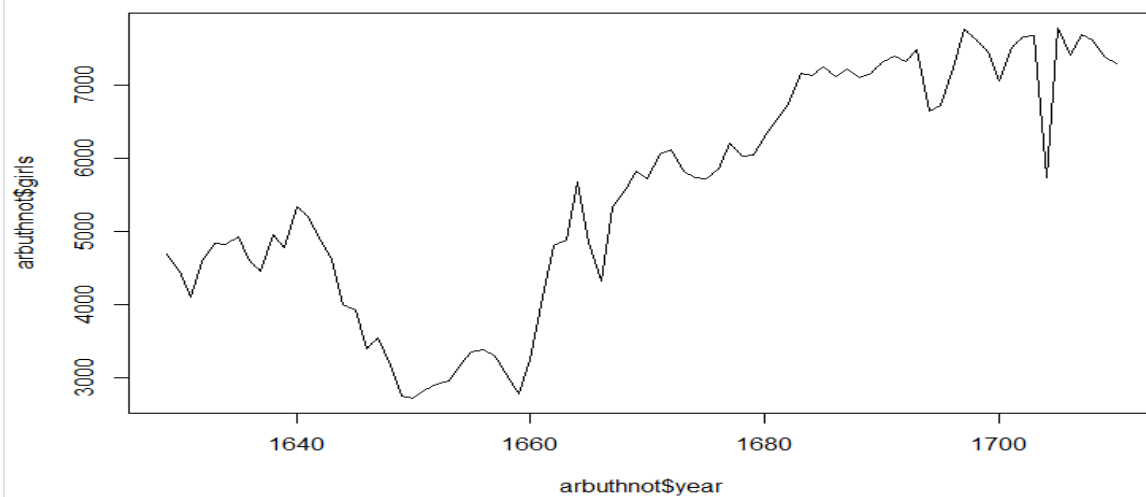
-create a simple plot of the number of girls or boys baptized per year with:

```
> plot(x=arbuthnot$year, y=arbuthnot$boys)
> plot(x=arbuthnot$year, y=arbuthnot$girls)
>
```



-If we want to connect the data points with line, we type this command:

```
19 plot(x=arbuthnot$year, y=arbuthnot$girls, type="l") #type=l pour relier entre les points dans le graphe
20
>
> plot(x=arbuthnot$year, y=arbuthnot$girls, type="l")
> |
```



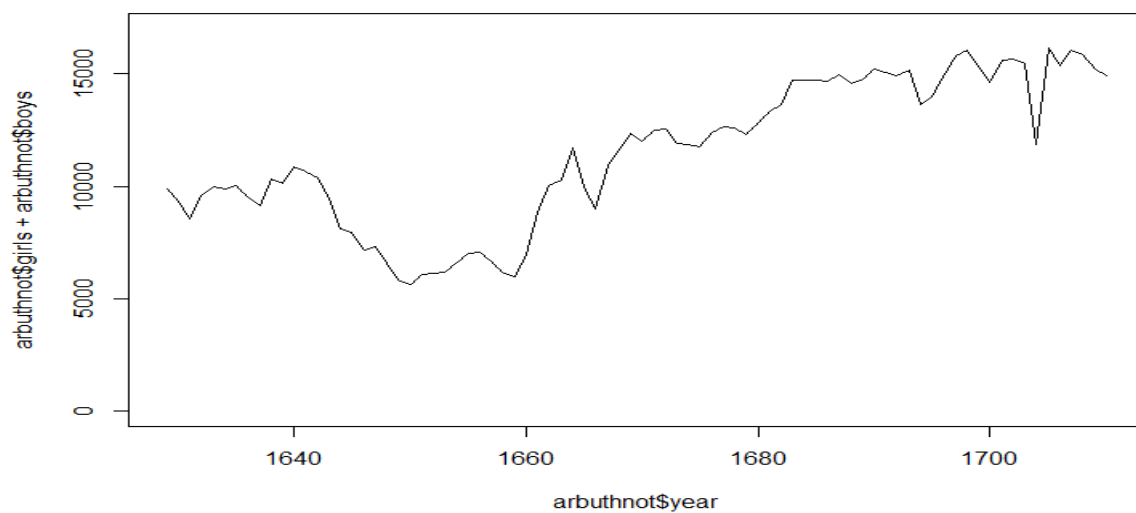
Q2:

-To compute the total number of baptisms we can type this mathematical expression

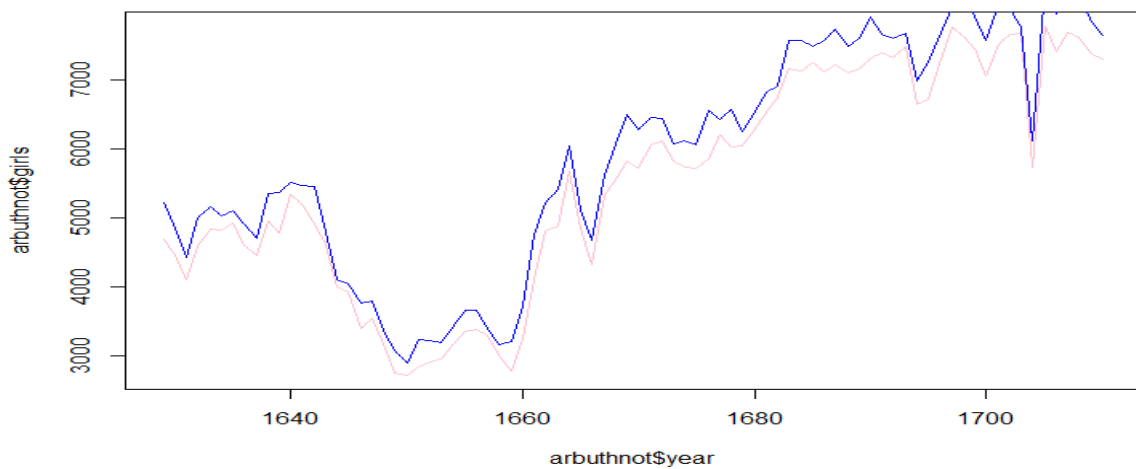
```
> 5218/4683
[1] 1.114243
>
>
> arbutnot$boys+arbutnot$girls
[1] 9901 9315 8524 9584 9997 9855 10034 9522 9160 10311 10150 10850 10670 10370 9410 8104 7966 7163 7332 6544 5825
[22] 5612 6071 6128 6155 6620 7004 7050 6685 6170 5990 6971 8855 10019 10292 11722 9972 8997 10938 11633 12335 11997
[43] 12510 12563 11895 11851 11775 12399 12626 12601 12288 12847 13355 13653 14735 14702 14730 14694 14951 14588 14771 15211 15054
[64] 14918 15159 13632 13976 14861 15829 16052 15363 14639 15616 15687 15448 11851 16145 15369 16066 15862 15220 14928
```

-we can a plot of the total number of baptisms per year with this command

```
>
>
> plot(x=arbutnot$year,y=arbutnot$girls+arbutnot$boys,type="l",ylim=c(0,17000))
>
```



```
>
> plot(x=arbutnot$year,y=arbutnot$girls+arbutnot$boys,type="l",ylim=c(0,17000))
> plot(x=arbutnot$year, y=arbutnot$girls, type="l") #type=l pour relier entre les points dans le graphe
>
> lines(arbutnot$year,arbutnot$girls,col="pink")
> lines(arbutnot$year,arbutnot$boys,col="blue")
>
```



-we can see here that the birth rate of the boys per year more than the birth rate of the girls.

Q3: make a plot of proportion of boys over time, what we see?

-When we ask if boys outnumber girls in each year, we can type this mathematical expression:

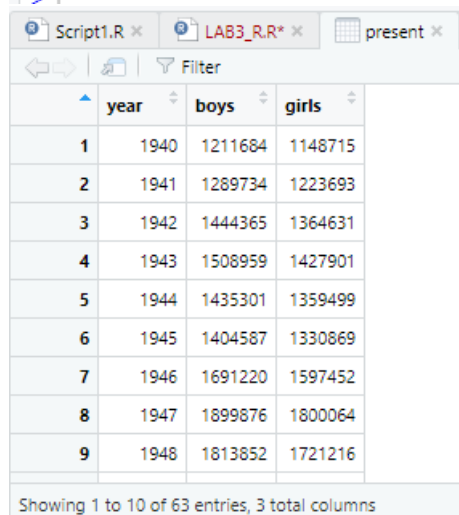
```
>
> arbutnot$boys>arbutnot$girls
[1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[27] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[53] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[79] TRUE TRUE TRUE TRUE
> |
```

*We notice that this command returns 82 values are TRUE, it means all these years had more boys than girls.

Exercise1:

Q1:

```
>
> source("http://www.openintro.org/stat/data/present.R")
> view(present)
> |
```



	year	boys	girls
1	1940	1211684	1148715
2	1941	1289734	1223693
3	1942	1444365	1364631
4	1943	1508959	1427901
5	1944	1435301	1359499
6	1945	1404587	1330869
7	1946	1691220	1597452
8	1947	1899876	1800064
9	1948	1813852	1721216

Q2:

-To see all the years include in this data set we type this command: **present\$year**

```
> #Q2
> present$year
[1] 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965
[27] 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991
[53] 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002
> |
```

-To see the dimension of the data frame we type this command: **dim(present)**

```
> dim(present)
[1] 63 3
> |
```

-To see the column names of this data we type this command: **names(present)**

```
> names(present)
[1] "year" "boys" "girls"
> |
```

Q3:

```
> range(present$year)
[1] 1940 2002
> range(arbuthnot$year)
[1] 1629 1710
> |

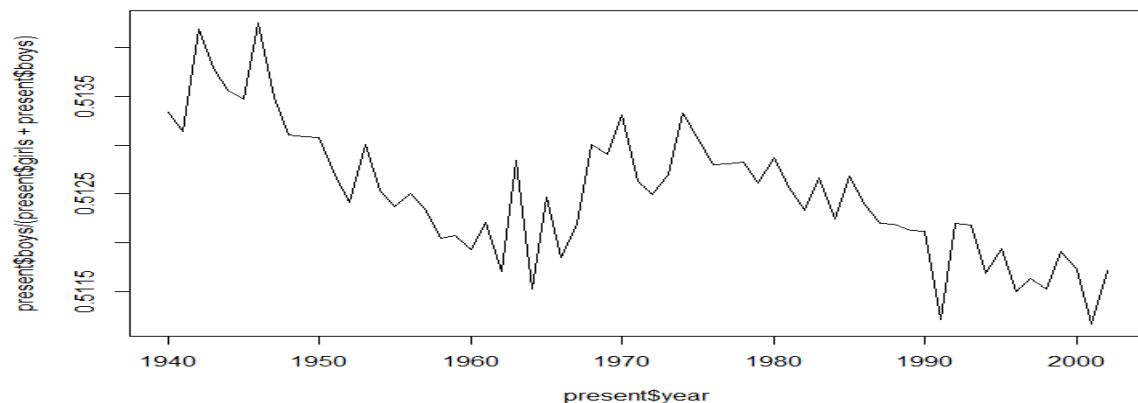
> head(present)
  year   boys  girls
1 1940 1211684 1148715
2 1941 1289734 1223693
3 1942 1444365 1364631
4 1943 1508959 1427901
5 1944 1435301 1359499
6 1945 1404587 1330869
> head(arbuthnot)
  year  boys  girls
1 1629  5218  4683
2 1630  4858  4457
3 1631  4422  4102
4 1632  4994  4590
5 1633  5158  4839
6 1634  5035  4820

> tail(present)
  year   boys  girls
58 1997 1985596 1895298
59 1998 2016205 1925348
60 1999 2026854 1932563
61 2000 2076969 1981845
62 2001 2057922 1968011
63 2002 2057979 1963747
> tail(arbuthnot)
  year  boys  girls
77 1705  8366  7779
78 1706  7952  7417
79 1707  8379  7687
80 1708  8239  7623
81 1709  7840  7380
82 1710  7640  7288
> |
```

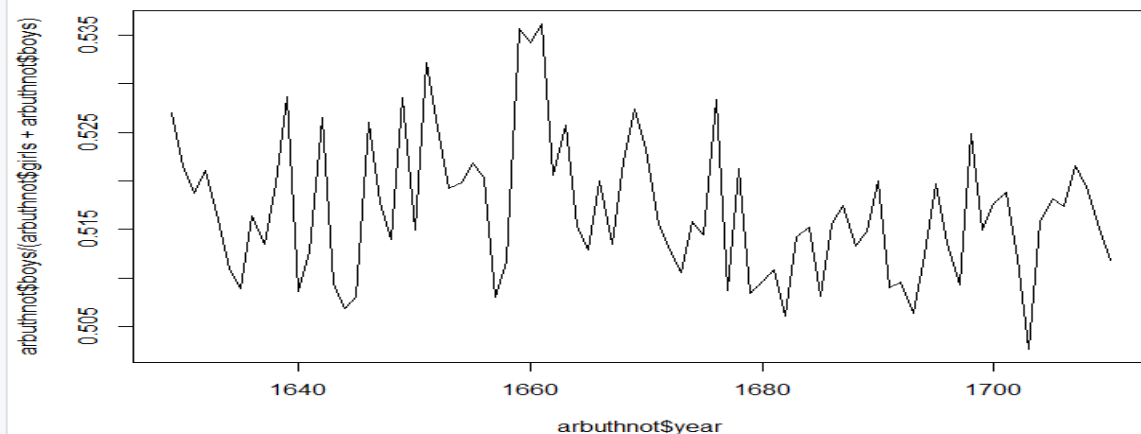
-We notice that both of those data frame aren't on similar scale because they are not on the same dimension and range.

Q4:

```
>
> plot(present$year, present$boys/(present$girls+present$boys), type="l")
> |
```



```
>
> plot(arbuthnot$year, arbuthnot$boys/(arbuthnot$girls+arbuthnot$boys), type="l")
> |
```



Q5:

```
>
> subset(present$year, present$boys + present$girls == max(present$boys + present$girls))
[1] 1961
>
```

Part 2: Introduction to data:

```
> #PART2
> source("http://www.openintro.org/stat/data/cdc.R")
> view(cdc)
```

	genhlth	exerany	hlthplan	smoke100	height	weight	wt Desire	age	gender
1	good	0	1	0	70	175	175	77	m
2	good	0	1	1	64	125	115	33	f
3	good	1	1	1	60	105	105	49	f
4	good	1	1	0	66	132	124	42	f
5	very good	0	1	0	61	150	130	55	f
6	very good	1	1	0	64	114	114	55	f
7	very good	1	1	0	71	194	185	31	m
8	very good	0	1	0	67	170	160	45	m
9	good	0	1	1	65	150	130	27	f
10	good	1	1	0	70	180	170	44	m

Showing 1 to 11 of 20,000 entries, 9 total columns

-to view the names of the variables:

```
>
> names(cdc)
[1] "genhlth" "exerany" "hlthplan" "smoke100" "height" "weight" "wt Desire" "age" "gender"
>
```

Q1: -To see how many cases in this data set and variables:

```
>
> head(cdc)
  genhlth exerany hlthplan smoke100 height weight wt Desire age gender
1    good      0        1         0    70   175    175   77      m
2    good      0        1         1    64   125    115   33      f
3    good      1        1         1    60   105    105   49      f
4    good      1        1         0    66   132    124   42      f
5 very good    0        1         0    61   150    130   55      f
6 very good    1        1         0    64   114    114   55      f
> tail(cdc)
  genhlth exerany hlthplan smoke100 height weight wt Desire age gender
19995    good      0        1         1    69   224    224   73      m
19996    good      1        1         0    66   215    140   23      f
19997 excellent    0        1         0    73   200    185   35      m
19998    poor      0        1         0    65   216    150   57      f
19999    good      1        1         0    67   165    165   81      f
20000    good      1        1         1    69   170    165   83      m
>
```

-Summaries and tables

```
>
> summary(cdc$weight)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  68.0  140.0  165.0  169.7  190.0  500.0
>
```

-if we want to compute the interquartile range for the respondents' weight:

```
> 190-140
[1] 50
>
```

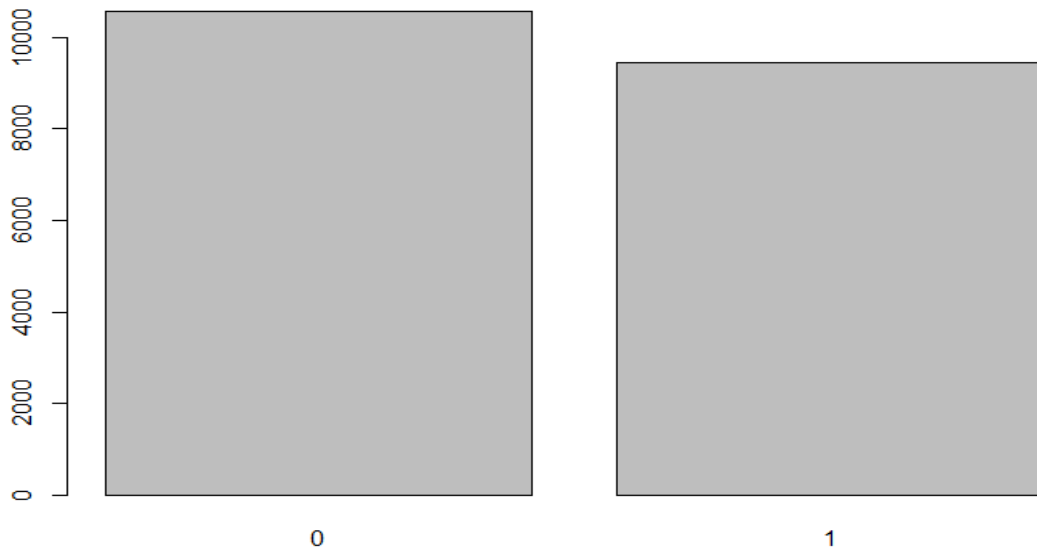
-To compute summary statistics one by one:

```
>
> mean(cdc$weight)
[1] 169.683
> var(cdc$weight)
[1] 1606.484
> median(cdc$weight)
[1] 165
> |
[1] 165
> table(cdc$smoke100)

      0      1
10559  9441
> table(cdc$smoke100)/20000 #afficher la probabilité résultat sous forme de pourcentage

      0      1
0.52795 0.47205
> |

> barplot(table(cdc$smoke100)) #tracer un graphe
>
> smoke<- table(cdc$smoke100)
> barplot(smoke)
> |
```



Q2:

```
> #Q2
> table(cdc$gender,cdc$smoke100)

      0      1
m 4547 5022
f 6012 4419
> mosaicplot(table(cdc$gender,cdc$smoke100))
>
> summary(cdc$height)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 48.00  64.00   67.00  67.18  70.00   93.00
> summary(cdc$weight)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 48.00  64.00   67.00  67.18  70.00   93.00
> IQR(cdc$weight)
[1] 50
>
> cdc[1:10,5]
[1] 70 64 60 66 61 64 71 67 65 70
>
> cdc[567,6]
[1] 160
>
```



```
table(cdc$gender, cdc$smoke100)
```



```
> #subsetting
> mdata=subset(cdc, cdc$gender=="m")
>
> data_30=subset(cdc,cdc$age >30)
>
> m_and_over30=subset(cdc,cdc$gender=="m" & cdc$age >30)
>
> view(m_and_over30)
>
```

Script1.R x LAB3_R.R x m_and_over30 x									
Filter									
	genhlth	exerany	hlthplan	smoke100	height	weight	wtdesire	age	gender
1	good	0	1	0	70	175	175	77	m
7	very good	1	1	0	71	194	185	31	m
8	very good	0	1	0	67	170	160	45	m
10	good	1	1	0	70	180	170	44	m
11	excellent	1	1	1	69	186	175	46	m
12	fair	1	1	1	69	168	148	62	m
14	excellent	1	1	1	70	170	170	69	m
16	good	1	1	1	73	185	175	79	m
17	good	0	0	1	67	156	150	47	m
18	fair	0	1	1	71	185	185	76	m
19	good	1	1	1	75	200	190	43	m
23	very good	0	1	1	73	160	160	43	m
30	excellent	1	1	1	74	185	175	63	m
31	very good	1	1	0	67	166	160	74	m
32	excellent	1	1	0	71	180	175	41	m
33	very good	1	1	0	71	182	182	36	m
34	good	1	1	1	68	185	160	67	m
38	excellent	1	1	1	69	190	180	65	m

Showing 1 to 19 of 7,244 entries, 9 total columns

```

> cdc$gender=="m"
[1] TRUE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE
[22] FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE
[43] FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE TRUE
[64] FALSE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE TRUE TRUE TRUE TRUE FALSE TRUE FALSE FALSE TRUE TRUE TRUE TRUE
[85] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE FALSE TRUE FALSE FALSE TRUE
[106] TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[127] TRUE FALSE TRUE TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[148] FALSE TRUE FALSE TRUE FALSE FALSE TRUE FALSE FALSE TRUE TRUE TRUE FALSE FALSE FALSE TRUE TRUE FALSE TRUE TRUE TRUE
[169] FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE
[190] TRUE FALSE FALSE TRUE FALSE TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE
[211] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE
[232] TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE
[253] TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE TRUE FALSE TRUE TRUE TRUE TRUE
[274] TRUE TRUE TRUE FALSE FALSE TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE FALSE TRUE
[295] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE FALSE TRUE FALSE TRUE TRUE FALSE FALSE FALSE
[316] FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE TRUE FALSE TRUE TRUE FALSE FALSE
[337] FALSE TRUE TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE
[358] TRUE FALSE FALSE FALSE FALSE TRUE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE TRUE FALSE
[379] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE TRUE TRUE TRUE
[400] FALSE FALSE FALSE TRUE FALSE TRUE TRUE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE TRUE
[421] FALSE TRUE TRUE FALSE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE FALSE FALSE TRUE
[442] FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE FALSE
[463] TRUE TRUE TRUE TRUE FALSE TRUE FALSE FALSE TRUE FALSE TRUE TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE TRUE TRUE
[484] TRUE TRUE FALSE TRUE FALSE TRUE TRUE FALSE TRUE FALSE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE FALSE
[505] FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE

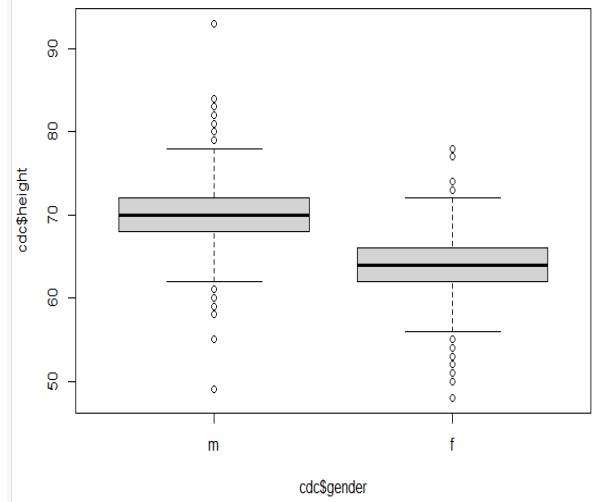
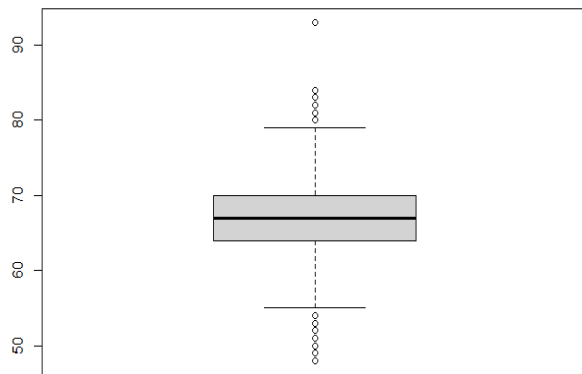
```

Q3:

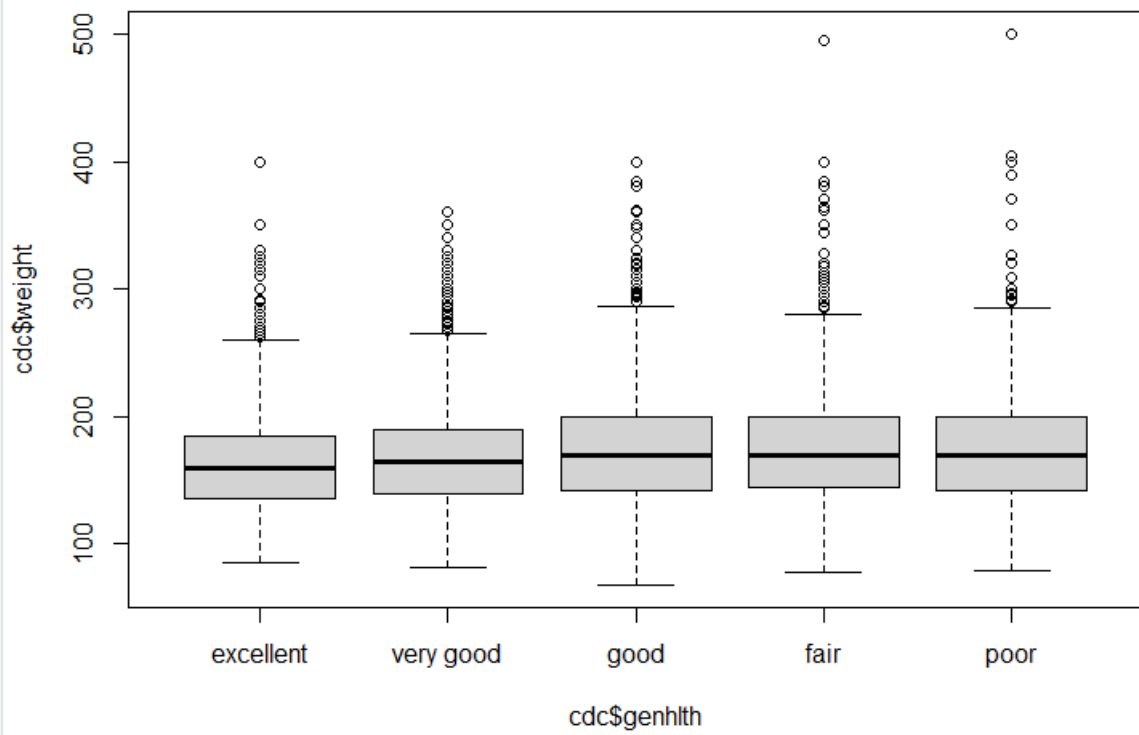
```

> #Q3
> under_23_andsmoke_100=subset(cdc, cdc$age <23 & cdc$smoke100==1)
>
> #Quantitative data : moyenne, max, min
> boxplot(cdc$height)
> boxplot(cdc$height~cdc$gender)
>
> boxplot(cdc$weight~cdc$genhlth)
>
> bmi=(cdc$weight/(cdc$height^2))*703
> boxplot(bmi~cdc$weight)
>
> hist(cdc$age)
> hist(cdc$age, breaks=50 )
~

```



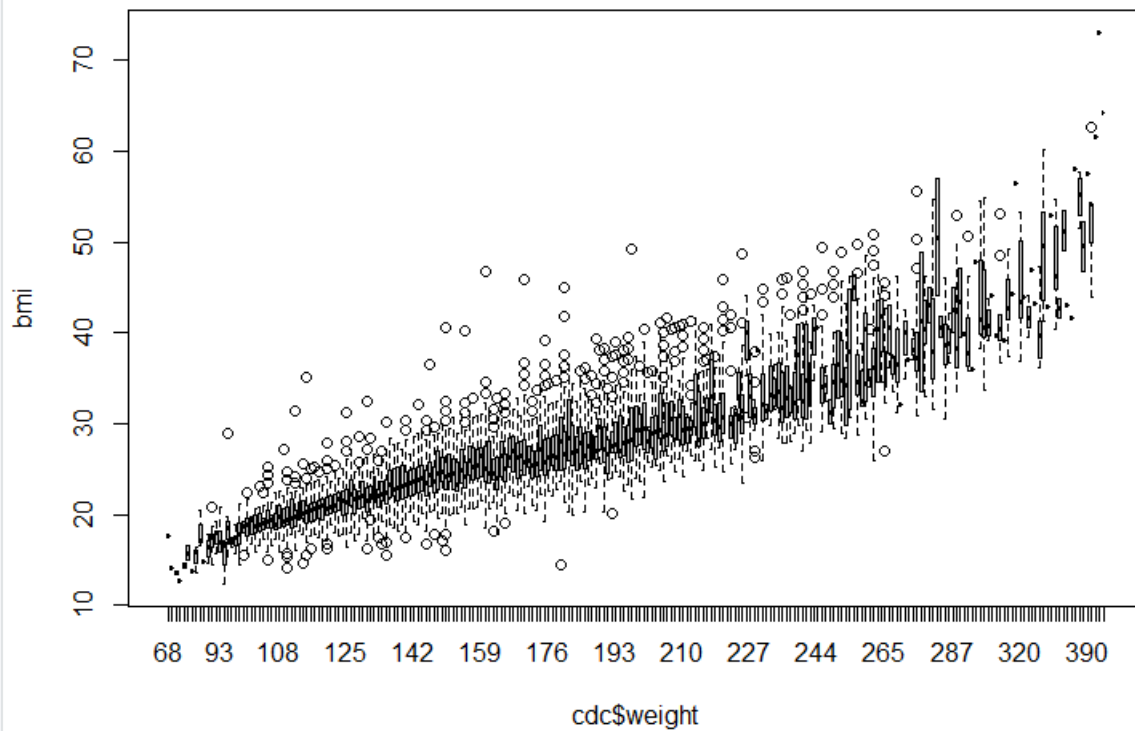
```
133
134 boxplot(cdc$weight~cdc$genhlth)
135
```



```

35
36 bmi=(cdc$weight/(cdc$height^2))*703
37 boxplot(bmi~cdc$weight)
38

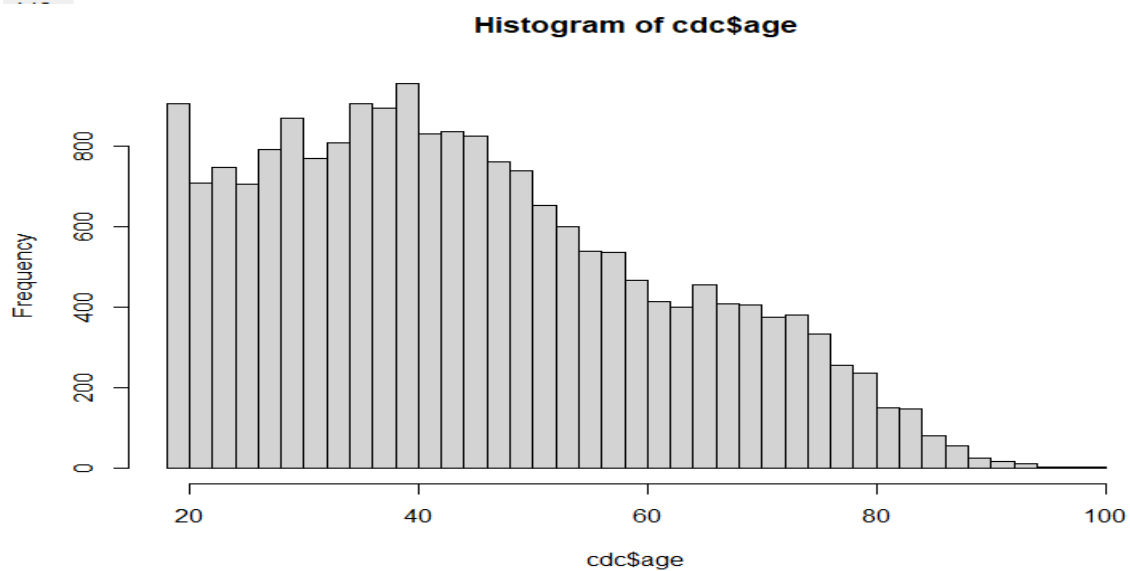
```



```

139 hist(cdc$age)
140 hist(cdc$age, breaks=50 )
141

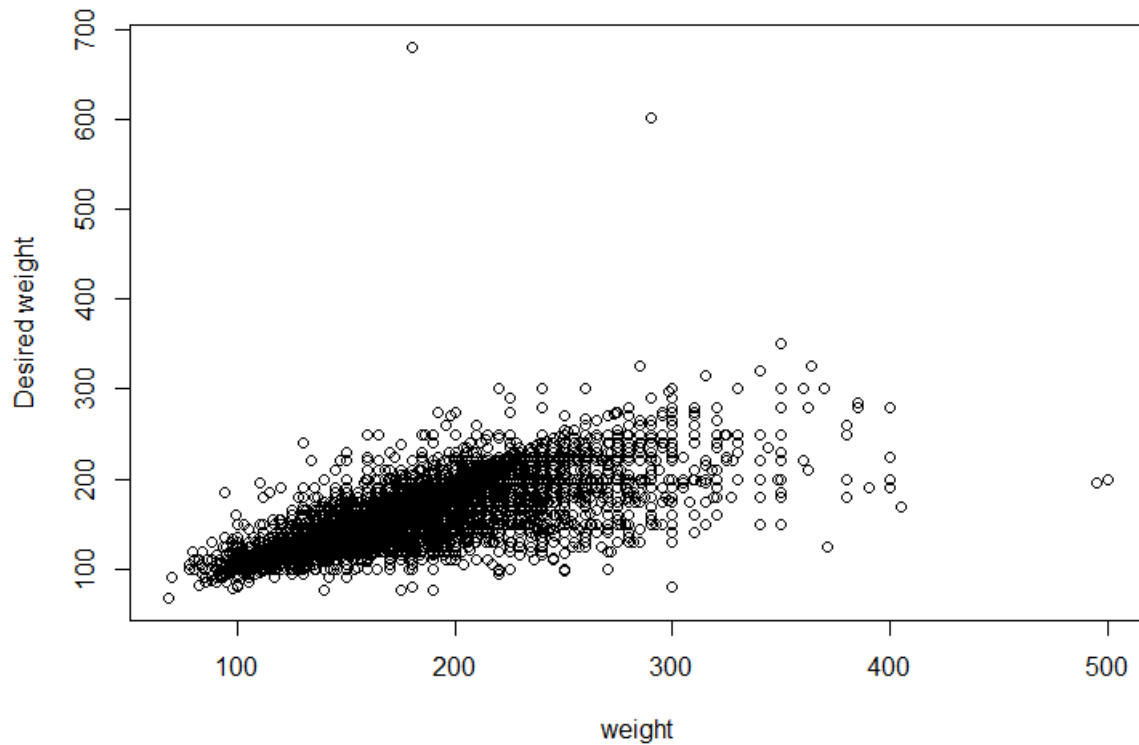
```



Exercise2:

Q1:

```
3 #Q1
4 plot(cdc$weight,cdc$wtdesired,xlab="weight",ylab="Desired weight")
5
6 |
```



-we notice that the weight increases, also the desired increases.

Q2:

```
145
146 wdiff <-(cdc$weight-cdc$wtdesired)
147 wdiff
148 |
```

```
> wdiff <- (cdc$weight-cdc$wtdesired)
> wdiff
[1] 0 10 0 8 20 0 9 10 20 10 11 20 -35 0 0 10 6 0 10 5 50 20 0 7 -15 40 25 10 10 10 6 5
[33] 0 25 60 7 20 10 0 15 6 13 0 40 40 35 10 0 -3 1 29 7 10 2 0 0 10 48 10 40 11 24 50 20
[65] 0 15 -10 5 10 0 25 0 10 20 -10 30 -12 6 22 10 -20 0 25 10 28 8 30 0 23 20 100 16 15 0 0 -25
[97] 20 10 0 0 0 2 0 0 0 30 4 28 11 20 0 0 0 13 0 12 10 10 45 -5 80 0 5 44 16 0 5 70
[129] 5 5 20 20 4 0 20 60 20 10 65 0 30 -7 25 40 -5 30 10 0 0 0 17 0 0 -5 20 28 15 0 0 -3
[161] 5 40 20 12 0 30 10 0 30 3 0 30 30 0 0 3 50 20 60 50 20 0 -5 25 9 15 20 0 18 12 10 36
[193] -15 40 0 0 0 42 25 0 30 15 5 25 10 75 0 0 -5 -15 7 6 5 0 20 10 24 4 5 15 0 6 5 0 0
[225] 5 15 0 40 0 0 0 15 0 10 7 35 37 5 0 10 40 5 23 30 5 0 5 0 40 10 25 -10 27 0 0 30
[257] 45 5 40 6 20 0 0 0 0 38 12 5 0 0 0 50 0 0 -9 35 30 80 20 30 25 5 10 20 10 0 0
[289] 15 6 0 0 0 60 55 15 5 -12 45 35 -10 17 5 0 0 0 0 30 5 3 35 100 12 20 5 0 0 26 0 50 0
[321] 10 30 10 45 0 0 6 0 7 0 50 40 5 20 20 0 20 10 -5 4 35 -10 30 0 16 27 10 45 5 108 0 0
[353] 8 0 15 0 15 65 10 0 20 0 0 0 0 0 0 40 0 20 80 5 25 25 -25 10 0 15 54 15 120 145 40 12 0

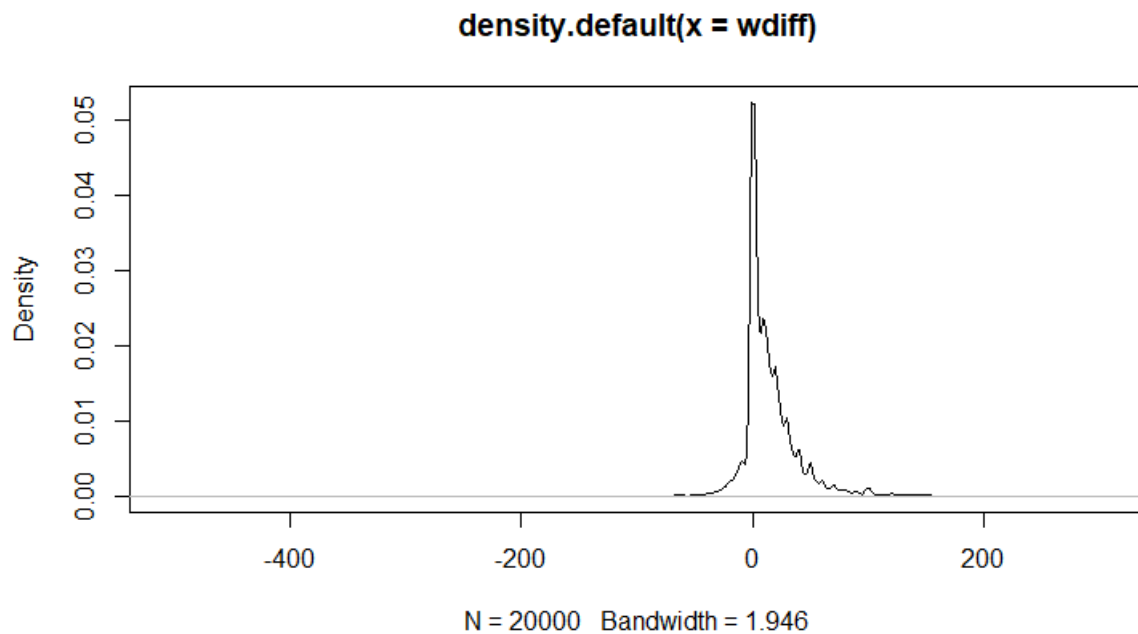
wdiff int [1:20000] 0 10 0 8 20 0 9 10 20 10 ...
```

Q3:

```
> typeof(wdiff)
[1] "integer"
> |
```

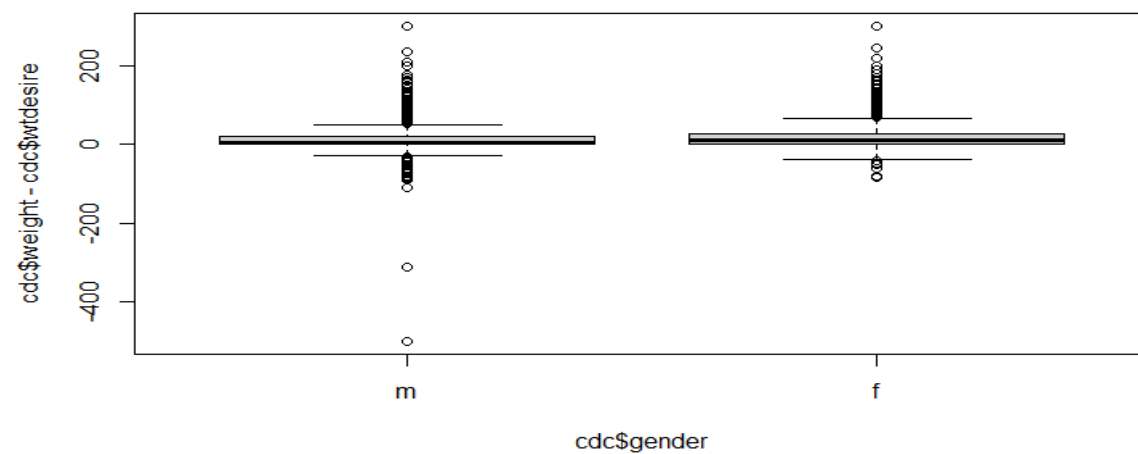
Q4:

```
> mean(wdiff)
[1] 14.5891
> sd(wdiff)
[1] 24.04586
> plot(density(wdiff))
> |
```



Q5:

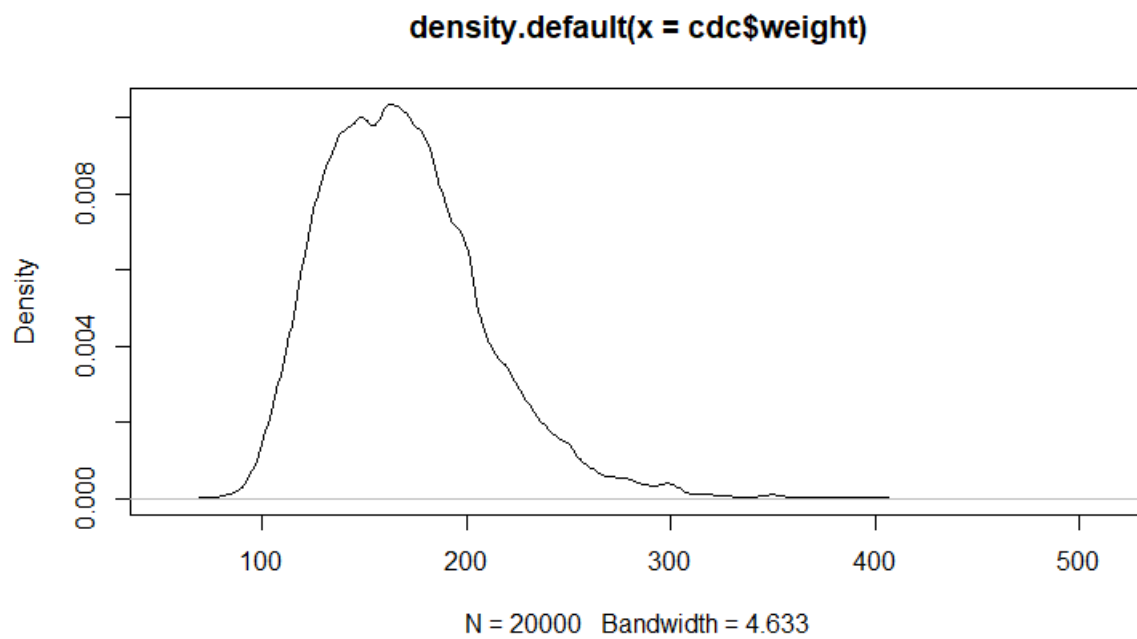
```
> #Q5
> boxplot(cdc$weight-cdc$wtdesired ~ cdc$gender)
> |
```



-we can notice that there is no difference.

Q6:

```
>  
> #Q6  
> mean(cdc$weight)  
[1] 169.683  
> sd(cdc$weight)  
[1] 40.08097  
> plot(density(cdc$weight))  
>
```



-It's a normal distribution.