Exercise 1 : Introduction to Data

Shikhar Dhwaj (A17030)

24 September 2017

Reading the data file

```
source("http://www.openintro.org/stat/data/cdc.R")
names(cdc)
```

```
## [1] "genhlth"  "exerany"  "hlthplan" "smoke100" "height"   "weight"
## [7] "wtdesire" "age"      "gender"
```

Q-1) How many cases are there in this data set? How many variables? For each variable, identify its data type (e.g. categorical, discrete).

```
dim(cdc)
```

```
## [1] 20000     9
```

Number of cases are 20000 Number of variables are 9

```
head(cdc)
```

```
##     genhlth exerany hlthplan smoke100 height weight wtdesire 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
```

Categorical Variables

- genhlth
- exerrany
- hlthplan
- smoke100
- gender

Continuous Variables (or Pseudo Discrete)

- height
- weight
- wtdesire
- age

Hence , Number of continuous variable are 4, Number of categorial variables are 5

Q-2) Create a numerical summary for height and age, and compute the interquartile range for each. Compute the relative frequency distribution for gender and exerany. How many males are in the sample? What proportion of the sample reports being in excellent health?

```
sh <- summary(cdc$height)
sh
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.00   64.00   67.00   67.18   70.00   93.00
```

```
sa <- summary(cdc$age)
sa
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.00   43.00   45.07   57.00   99.00
```

Interquartile Range

Height

```
sh[5] - sh[2]
```

```
## 3rd Qu.
##       6
```

Age

```
sa[5] - sa[2]
```

```
## 3rd Qu.
##      26
```

Relative Frequency Distribution for Gender and Exerany

```
table(cdc$gender,cdc$exerany) / sum(table(cdc$gender,cdc$exerany))
```

```
##
##           0       1
##   m 0.10745 0.37100
##   f 0.14685 0.37470
```

Number of Males in the sample

```
table(cdc$gender)
```

```
##
##     m     f
##  9569 10431
```

Number of Males are 9569

Proportion of sample report in excellent health

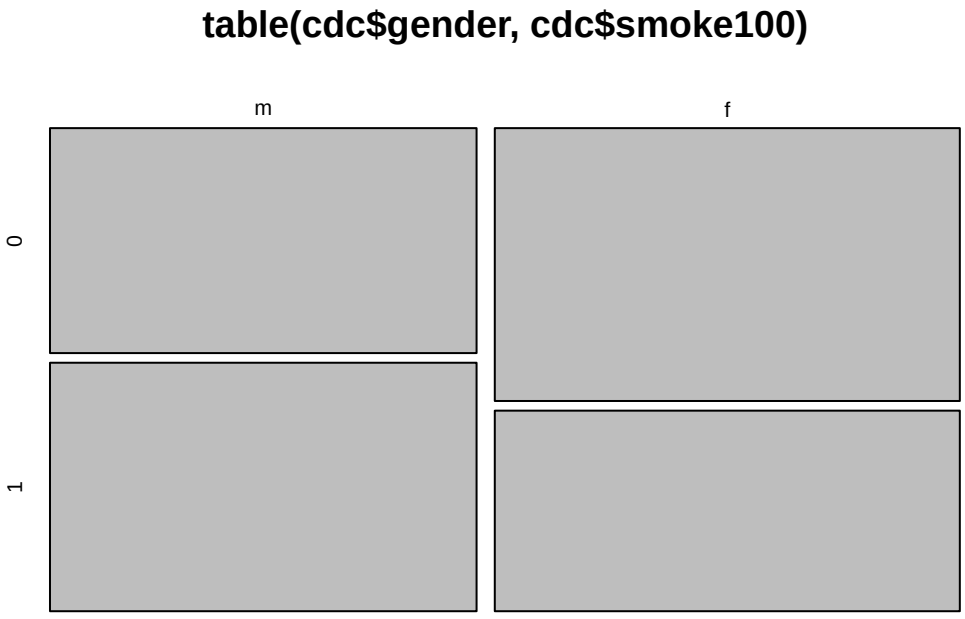
```
table(cdc$genhlth)
```

```
##
## excellent very good      good      fair      poor
##      4657      6972      5675      2019       677
```

4657 cases in excellent health out of 20000

Q-3) What does the mosaic plot reveal about smoking habits and gender?

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



Mosaic Plot tells us that number of males who smoke are greater than number of females who smoke. Hence, it may be possible that Males are more likely to smoke than females

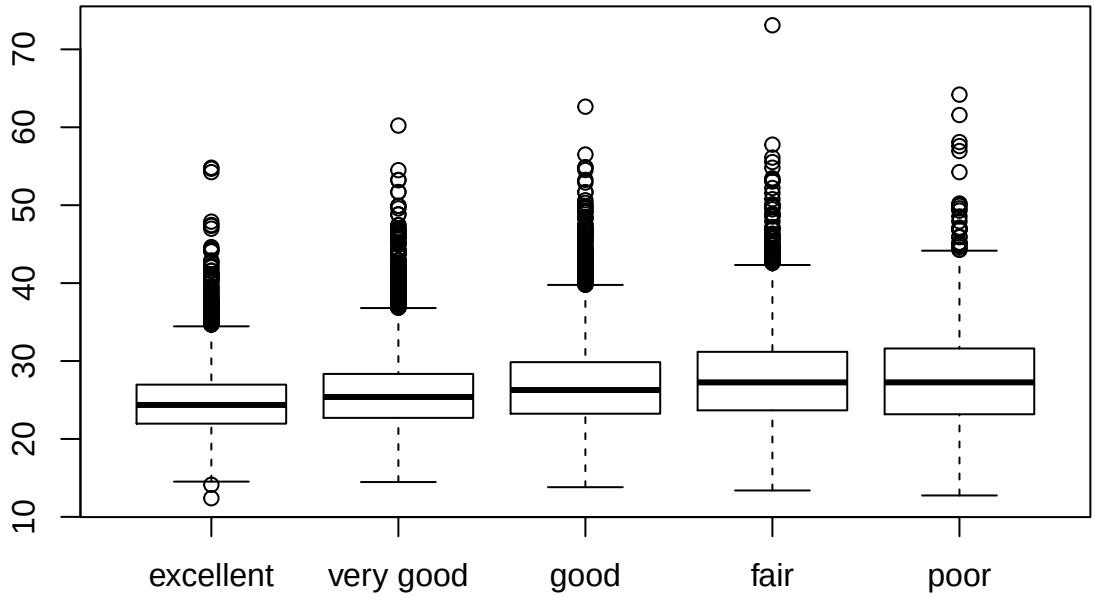
Q-4) Create a new object called under23_and_smoke that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

```
under23_and_smoke <- cdc[cdc$age==23 & cdc$smoke100==1,]
```

Q-5) What does this box plot show? Pick another categorical variable from the data set and see how it relates to BMI?

Variable chosen is general health

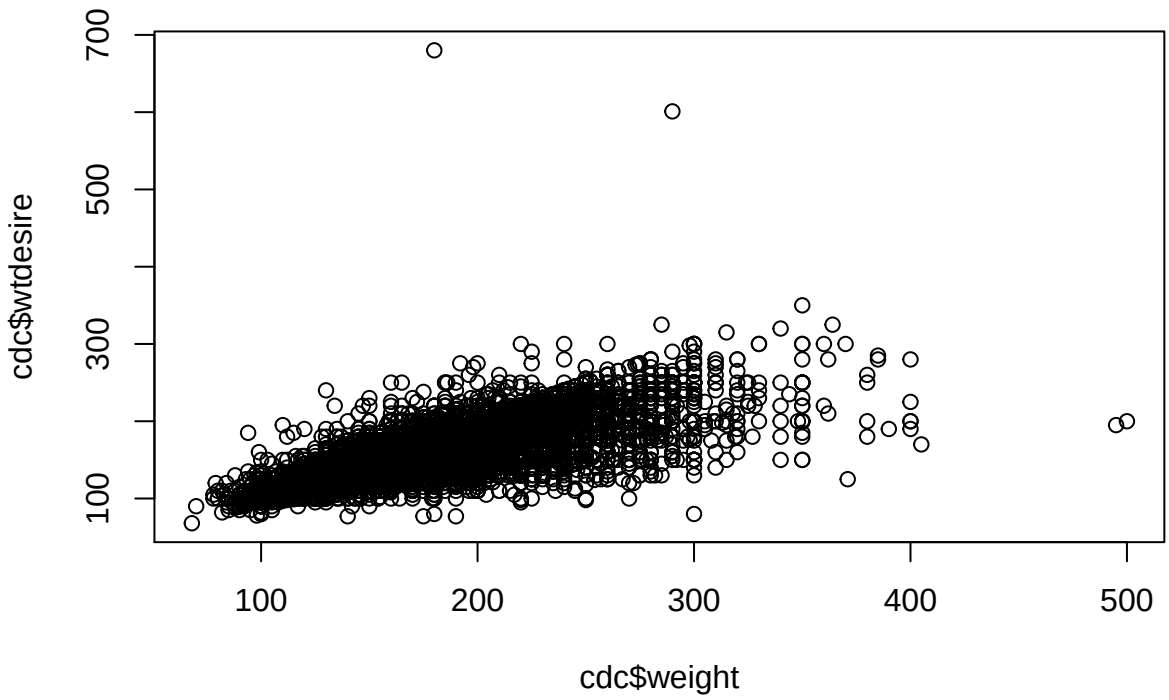
```
cdc$BMI <- cdc$weight / cdc$height^2 * 703
boxplot(cdc$BMI ~ cdc$genhlth)
```



The distributions of BMIs within each health status group is left skewed

Q-6) Make a scatterplot of weight versus desired weight. Describe the relationship between these two variables ?

```
plot(cdc$weight,cdc$wtdesire)
```



Relatioship appears to be linear with few outliers

Q-7) Let's consider a new variable: the difference between desired weight (wtdesire) and current weight (weight). Create this new variable by subtracting the two columns in the data frame and assigning them to a new object called wdiff.

```
wdiff <- cdc$wtdesire - cdc$weight
```

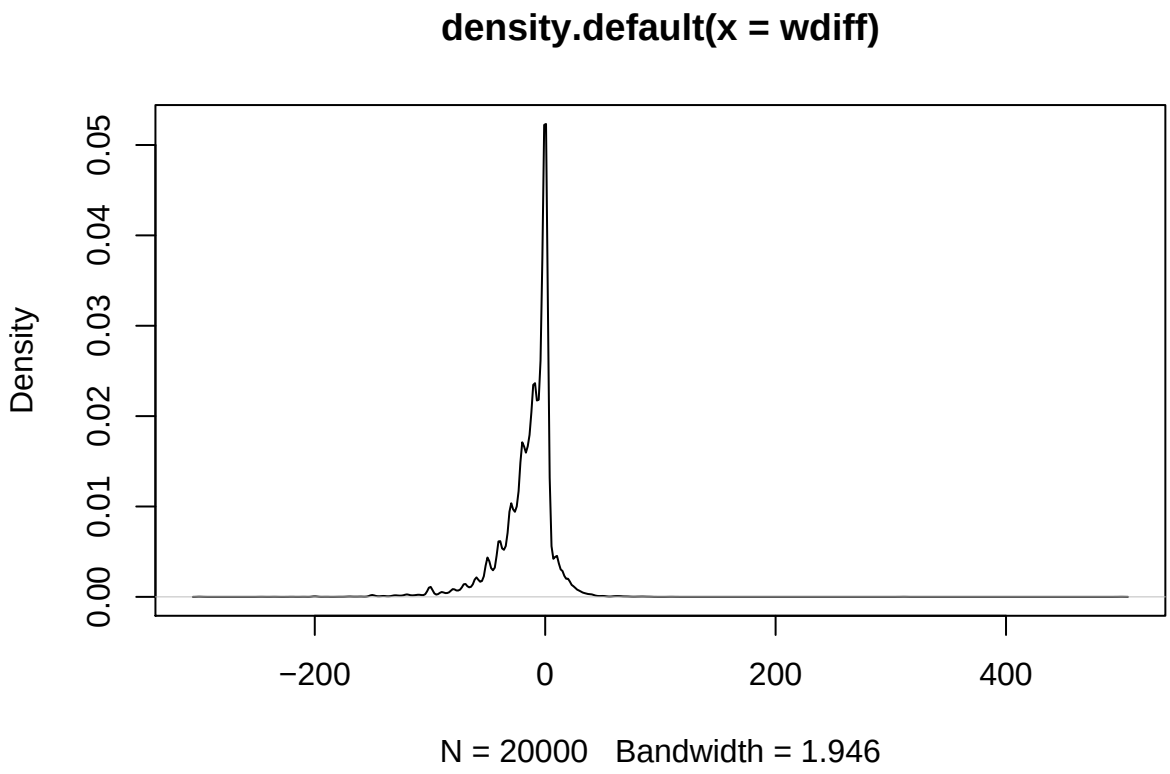
Q-8) What type of data is wdiff? If an observation wdiff is 0, what does this mean about the person's weight and desired weight. What if wdiff is positive or negative?

```
summary(wdiff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -300.00  -21.00  -10.00  -14.59    0.00  500.00
```

It is Continuous Variable If wdiff is 0 then desired weight is current weight If wdiff is positive then current weight is less than desired weight If wdiff is negative then current weight is more than desired weight ### Q-9)Describe the distribution of wdiff in terms of its center, shape, and spread, including any plots you use. What does this tell us about how people feel about their current weight?

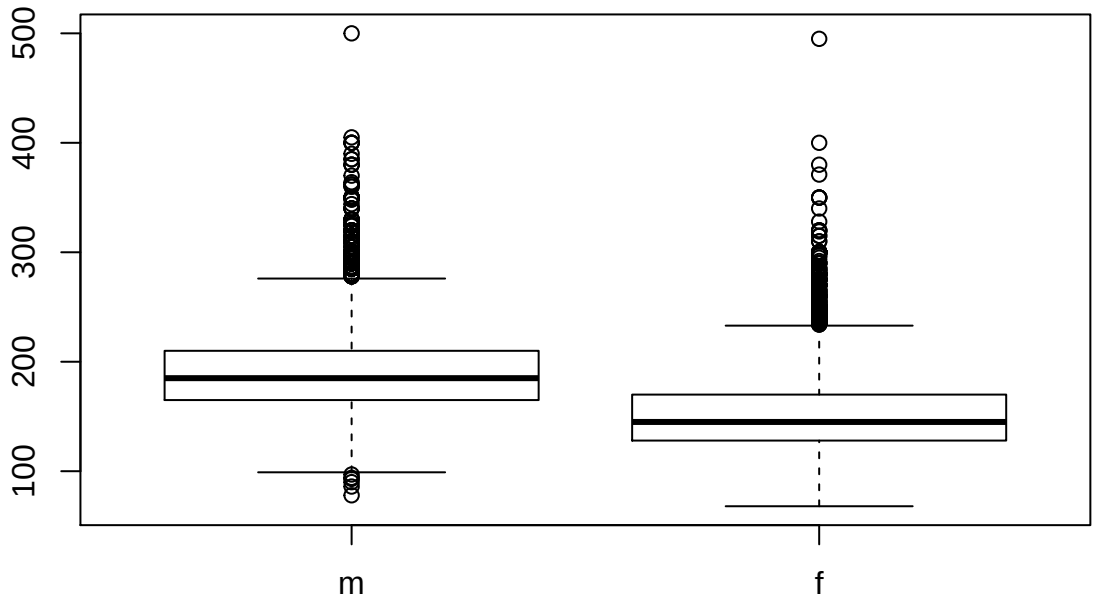
```
plot(density(wdiff))
```



Plot is centered at 0 Plot is a nearly normal It is spread approximatly from -175 to 125 For most of the cases current weight is equal to desired weight

Q-10)Using numerical summaries and a side-by-side box plot, determine if men tend to view their weight differently than women.

```
boxplot(cdc$weight~cdc$gender)
```



Yes,men tend to view their weight differently as compared to women

Q-11)Now it's time to get creative. Find the mean and standard deviation of weight and determine what proportion of the weights are within one standard deviation of the mean.

```
m<-mean(cdc$weight)
m
```

```
## [1] 169.683
```

```
s<-sd(cdc$weight)
s
```

```
## [1] 40.08097
```

```
table(cdc$weight>m-s & cdc$weight<m+s)
```

```
##
## FALSE  TRUE
##  5848 14152
```