**Tor Thogersen**

**DA 460 – Fall 2017**

**Lab 1 - Handout 0 R and Handout 0 SAS**

## **Part 2 – R Handout**

**Exercise 1**

* How many cases are there in this data set / How many variables

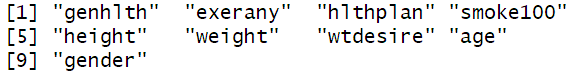
> dim(cdc)

[1] 20000 9

(total sample would have 200+ different variables, our dataset will be reduced for our class)

* For each variable (identify its data type)

> names(cdc)



**Exercise 2**

* Create a numeric summary for height and age, and compute the interquartile range for each.

> summary(cdc$height)

Min. 1st Qu. Median Mean 3rd Qu. Max.

48.00 64.00 67.00 67.18 70.00 93.00

> quantile(cdc$height,.25)

25%

64

> quantile(cdc$height,.75)

75%

70

> IQR(cdc$height)

[1] 6

> summary(cdc$age)

Min. 1st Qu. Median Mean 3rd Qu. Max.

18.00 31.00 43.00 45.07 57.00 99.00

> quantile(cdc$age,.25)

25%

31

> quantile(cdc$age,.75)

75%

57

> IQR(cdc$age)

[1] 26

* Compute the relative frequency distribution for gender and exerany.

> table(cdc$gender)/20000

m f

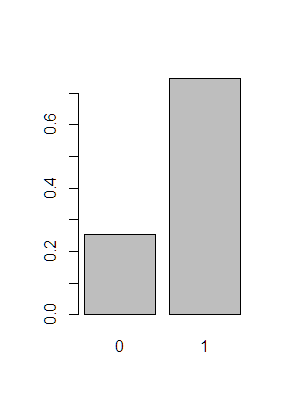
0.47845 0.52155

> table(cdc$exerany)/20000

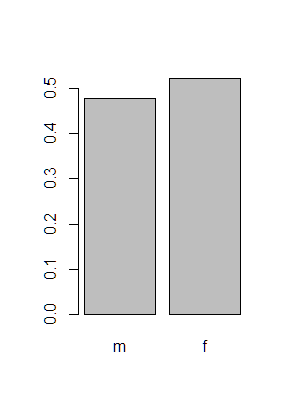
0 1

0.2543 0.7457

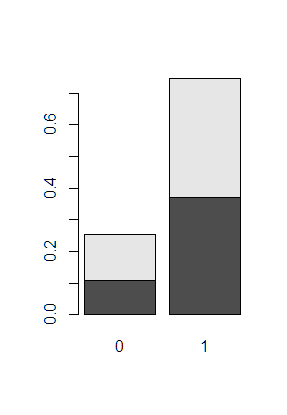
> barplot(table(cdc$exerany)/20000)



> barplot(table(cdc$gender)/20000)

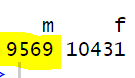


> barplot(table(cdc$gender, cdc$exerany)/20000)



* How many Males are in the sample.

> table(cdc$gender)



* What proportion of the sample reports being in excellent health.

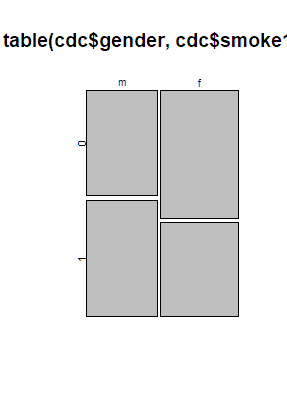
> table(cdc$genhlth)/20000



**Exercise 3**

* What does the mosaic plot reveal about smoking habits and gender.

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



The sample shows males are more likely to have smoked 100 cigarettes versus females in the sample

**Exercise 4**

* Create a new object called “under23\_and\_smoke” that contains all observations of respondents under the age 23 that have smoked 100 cigarettes in their lifetime.

> under23\_and\_smoke <- subset(cdc, age > 23 & smoke100)

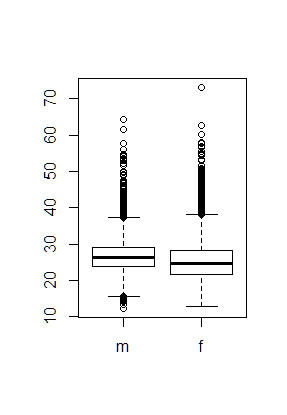
**Exercise 5**

* What does this box plot show?

The Sample data shows that smaller BMI range for respondents that reported excellent health, while Fair / Poor respondents have the highest BMI range.

* Pick another categorical variable from the data set and see how it relates to BMI.

> bmi <- (cdc$weight / cdc$height^2)\*703; boxplot(bmi ~ cdc$gender)



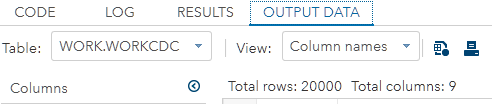
* List the variable you chose.
  + Gender
* Why you might think it would have a relationship to BMI and indicate what the figure seems to suggest.
  + Due to the different makeup between Male and Female, I would hypothesis the sample would show the average BMI range would be great for females versus males.
  + Based on the data, it shows males had a smaller BMI range vs Females but females had Lower Min and Max BMI overall.

**Part 2 – SAS Handout**

**Exercise 1**

* How many cases are there in this data set?

proc print data=work.cdc (obs=10);  
run;



* How many variables? 9
* For each variable, identify its data type (for example, categorical, numeric).

proc contents data=work.cdc short;

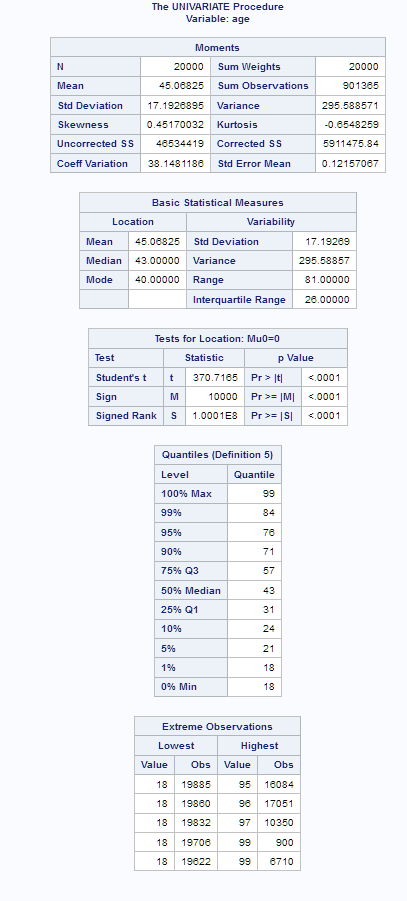
run;



**Exercise 2**

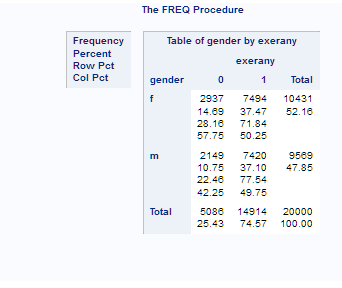
* Create a numerical summary for **height** and **age**, and compute the interquartile range for each.

proc univariate data=work.cdc;  
 var height;  
 var age;  
run;

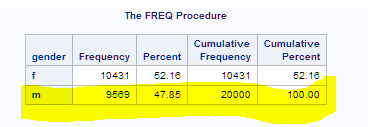
* Compute the relative frequency distribution for **gender** and **exerany**.

proc freq data=work.cdc;  
 tables gender\*exerany;  
run;



* How many males are in the sample?

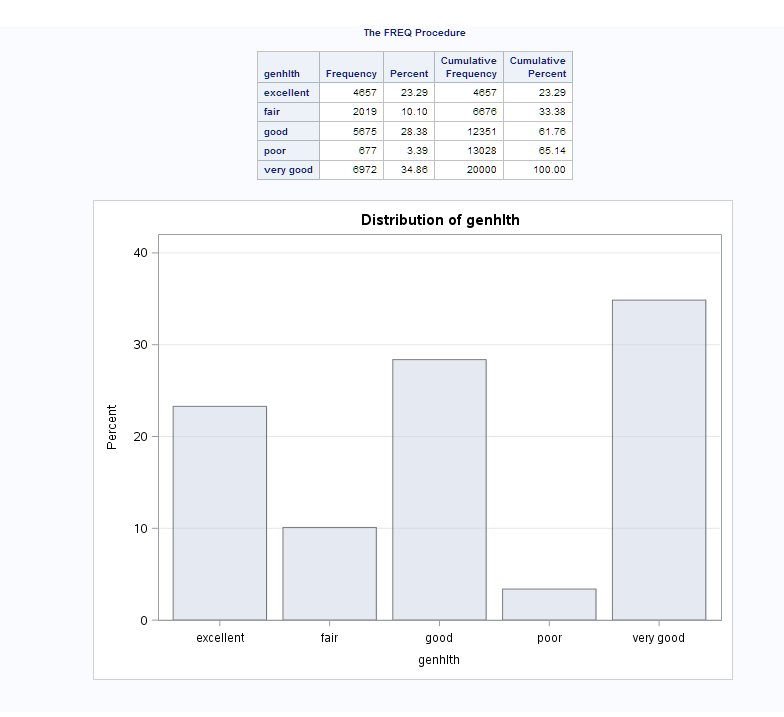
proc freq data=work.cdc; tables gender;  
run;



* What proportion of the sample reports being in excellent health?

23.29 percent

proc freq data=work.cdc;  
 tables genhlth / plots=freq;  
run;



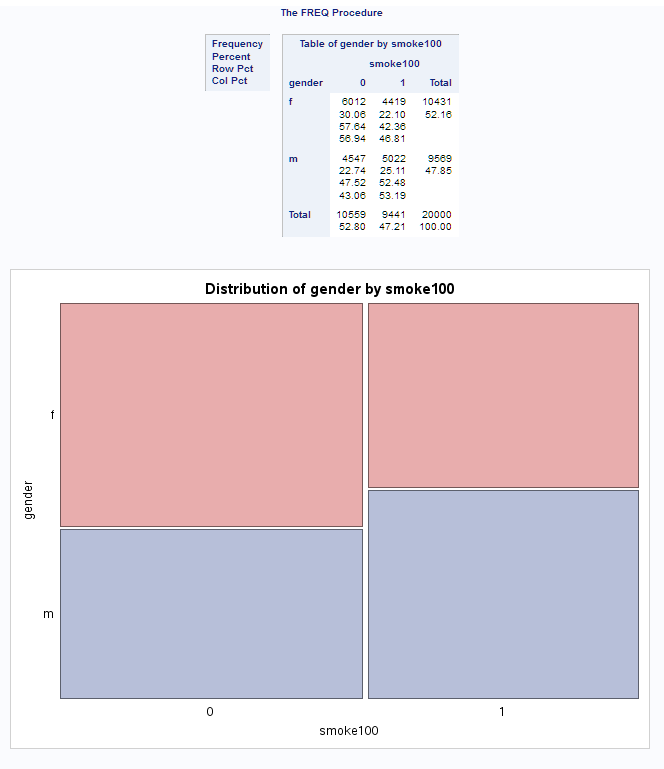
**Exercise 3**

* What does the mosaic plot reveal about smoking habits and gender?

Sample data shows a greater number of males vs females have smoked 100 or more cigarettes within their lifetime.

Additionally the data shows greater number of female in the sample did not smoke or smoked less < 100 cigarettes while the men data shows greater number of men in the sample did smoke versus sample of men that did not smoke or smoked less < 100 cigarettes.

proc freq data=work.cdc;  
 tables gender\*smoke100 / plots=mosaicplot;   
run;

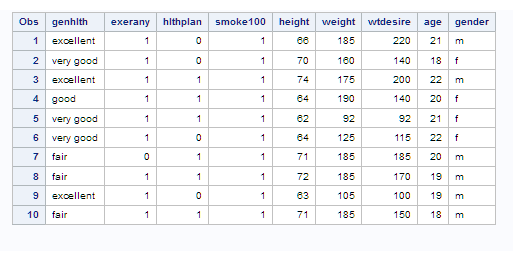


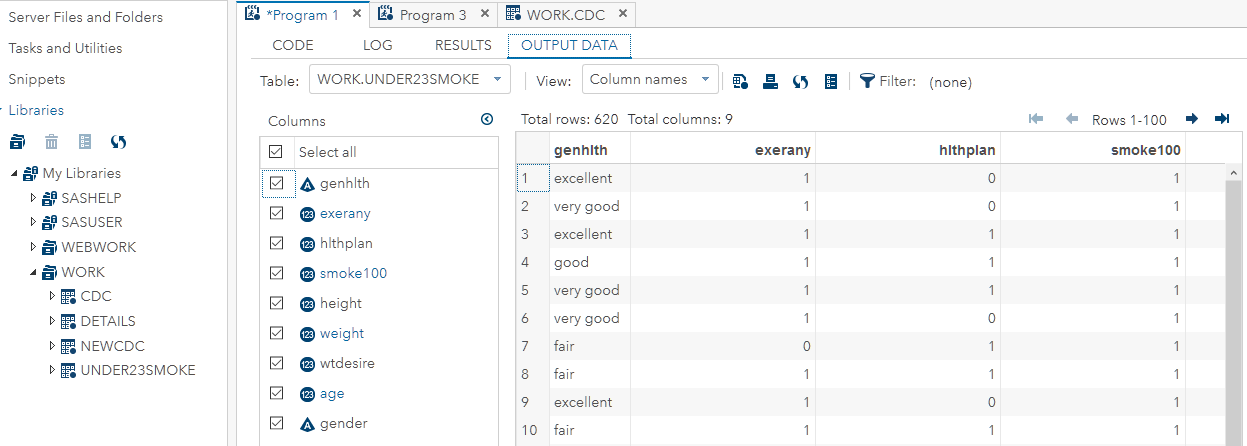
**Exercise 4**

Create a new data set named **under23smoke** that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime.

Write the programming statements you used to create the new data set as the answer to this exercise.

data work.under23smoke;   
 set work.cdc;  
 where smoke100 and age < 23;  
run;  
  
proc print data=work.newcdc (obs=10);   
run;

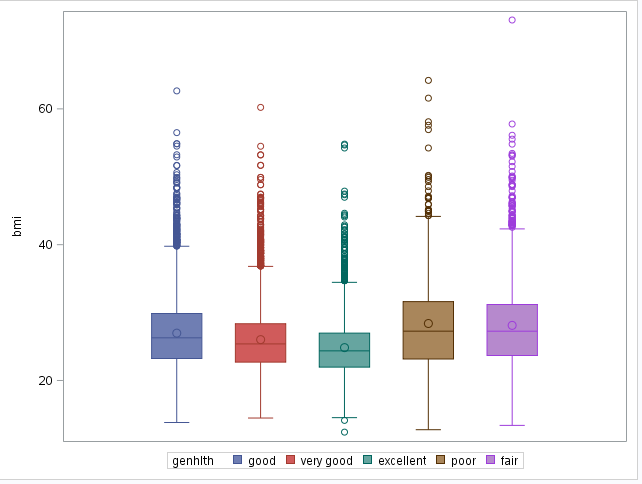




**Exercise 5**

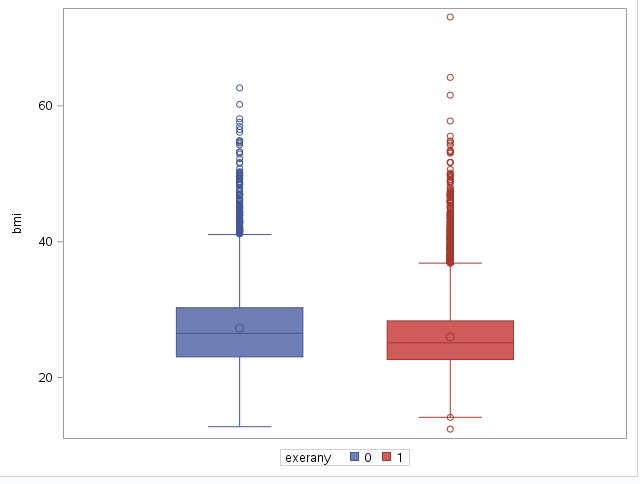
* What does this box plot show?

shows the BMI min, max, 1st / 3rd Q, Med sample data with general health responses



* Pick another categorical variable from the data set and see how it relates to BMI.

proc sgplot data=work.cdcbmi;   
 vbox bmi / group=exerany;  
run;



* List the variable you chose, why you might think it would have a relationship to BMI, and indicate what the figure seems to suggest.
  + Exerany
  + my hypothesis suggests that if you exercise your BMI on average should be less than someone who does not exercise.
  + Based on the sample data, the BMI average is smaller for people who exercise.