Part A: Use the Alzheimer's Data (RM_xsect.csv) to complete the work below:

(1) Read data into R for analysis.

```
## first set working directory
setwd("~/Dropbox/AAU materials/datasets/Rosmap")

## double check this
getwd()

## load the data
RM <- read.csv("RM_xsect.csv")

## look at the data
head(RM)
```

(2) Generate descriptive statistics for "global_lv," including the minimum, first quartile, median, third quartile, and maximum.

```
> summary(RM$global_lv)
   Min.  1st Qu.   Median     Mean  3rd Qu.    Max.    NA's
-5.99024 -1.82601 -0.90609 -0.94956 -0.04854  2.99532       1
```

(3) Based on results in part (2), create a categorical variable for quartiles of "global_lv" as described below.

a. Create 1 categorical variable coded as:

i. 1: [min to Q1),
ii. 2: [Q1 to median),
iii. 3: [median to Q3),
iv. 4: [Q3 to max).

```
?cut
quantile(RM$global_lv, probs=c(0,0.25,0.5,0.75,1), na.rm=TRUE)
RM$global_lv_cat <- cut(RM$global_lv, c(-Inf,-1.83, -0.906, -0.049, Inf), right=FALSE,
include.lowest = TRUE)
table(RM$global_lv_cat)
```

b. Use the *factor* function to label the 4 levels.

```
?factor
RM$global_lv_cat <- factor(RM$global_lv_cat, labels = c("Q1","Q2","Q3","Q4"))
table(RM$global_lv_cat)
```

c. Check your work from part (3a) using the *tapply* function.

```
> quantile(RM$global_lv, probs=c(0,0.25,0.5,0.75,1), na.rm=TRUE)
```

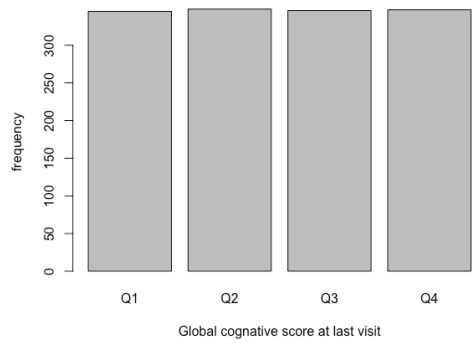
```
       0%        25%        50%        75%       100%
-5.99023809 -1.82600966 -0.90608958 -0.04853579  2.99531647
> tapply(RM$global_lv, RM$global_lv_cat, min, na.rm=TRUE)
       Q1        Q2        Q3        Q4
-5.99023809 -1.82617758 -0.90379142 -0.04776502
> tapply(RM$global_lv, RM$global_lv_cat, max, na.rm=TRUE)
       Q1        Q2        Q3        Q4
-1.83172937 -0.90838774 -0.05084808  2.99531647
```

d. Make a frequency table for the variable created in (3a). Be sure to include both frequency and relative frequency.

```
> tab_global <- table(RM$global_lv_cat)
> prop.table(tab_global)

     Q1       Q2       Q3       Q4
0.2489177 0.2510823 0.2496392 0.2503608
> cbind(tab_global, prop.table(tab_global))
  tab_global
Q1        345 0.2489177
Q2        348 0.2510823
Q3        346 0.2496392
Q4        347 0.2503608
```

Category	frequency	Relative frequency
Q1	0.2489	0.2489
Q2	0.2511	0.2489 + 0.2511 = 0.5
Q3	0.2496	0.2489 + 0.2511 + 0.2496 = 0.7496
Q4	0.2504	0.2489 + 0.2511 + 0.2496 + 0.2504 = 1

e. Create a bar graph for the variable you created in (3a). Be sure to label your axes and make your bar graph publishable.




```
barplot(tab_global, xlab="Global cognative score at last visit", ylab="frequency")
```

Part B: We will continue to use the Alzheimer's dataset (RM_xsect.csv) to complete the work below:

(1) Read data into R for analysis.

Same as part A (1).

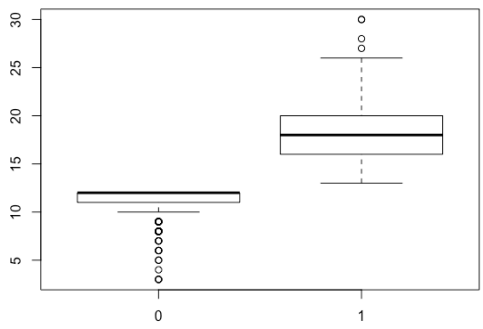
(2) Create a new binary variable for greater than high school education. The variable "educ" is the years of education. Your binary variable should have the following levels:
   a. 0 if education level is: High School (HS) or less education (education <= 12 years),
   b. 1 if education level is: University and more education (education > 12 years).

```
> RM$educ_cat <- as.numeric(RM$educ > 12)
> table(RM$educ_cat)

  0   1
302 1084
```

(3) Separately plot the distribution of years of education for each education category.
   a. Calculate the summary statistics for each group.

```
> tapply(RM$educ, RM$educ_cat, summary)
$`0`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00   11.00   12.00   11.17   12.00   12.00

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.00   16.00   18.00   17.55   20.00   30.00
```

   b. Create side-by-side boxplots of the distribution of years of education, stratified by education group.



boxplot(educ~educ_cat, data=RM)

(4) Test whether the proportion of those with university or more education is different than 75%.

```
> table(RM$educ_cat)

  0   1
302 1084
> prop.test(1084, 302+1084, p=0.25, correct=FALSE)

        1-sample proportions test without continuity correction

data:  1084 out of 302 + 1084, null probability 0.25
X-squared = 2093, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.25
95 percent confidence interval:
 0.7596100 0.8030441
sample estimates:
      p
0.7821068
```