Lab 2 Solution

```r
library(car)
library(ggplot2)
```

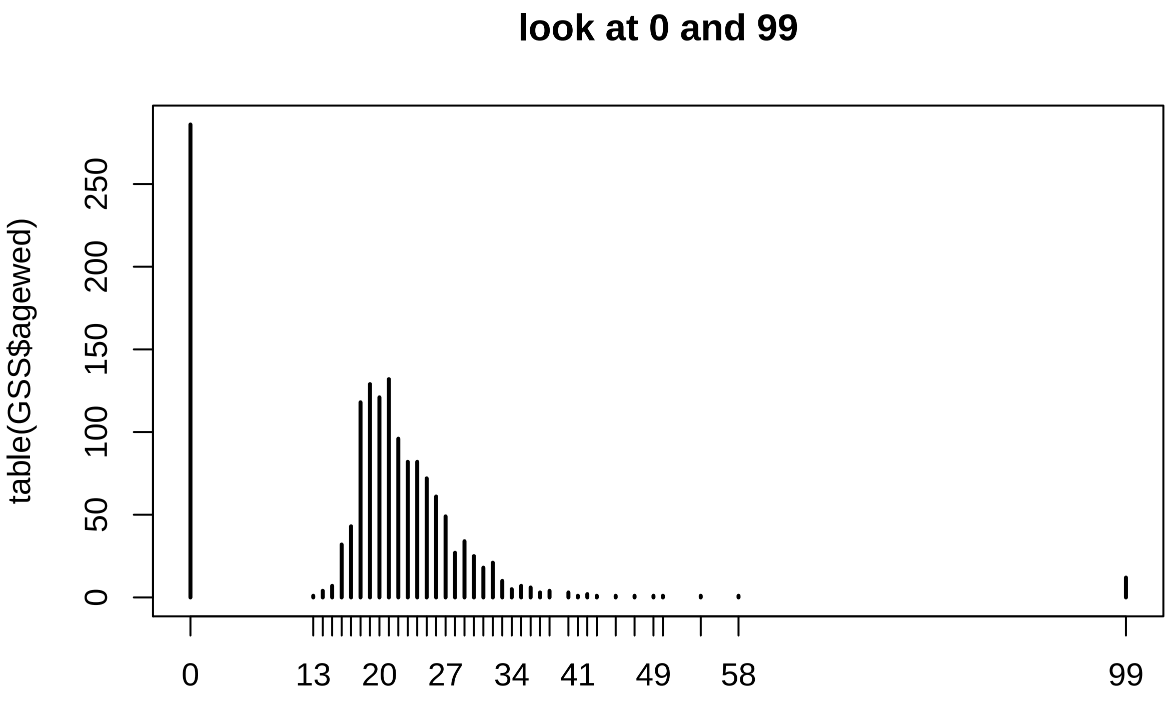
```r
load('GSS.Rdata')
ls()
```

```
## [1] "GSS"
```

```r
### see the data
table(GSS$agewed,useNA='always')
```

```
## 
##     0   13   14   15   16   17   18   19   20   21   22   23   24   25   26 
##   286    1    4    7   32   43  118  129  121  132   96   82   82   72   61 
##    27   28   29   30   31   32   33   34   35   36   37   38   40   41   42 
##    49   27   34   25   18   21   10    5    7    6    3    4    3    1    2 
##    43   45   47   49   50   54   58   99 <NA> 
##     1    1    1    1    1    1    1   12    0 
```

```r
# yes: you can plot a table.
plot(table(GSS$agewed), main='look at 0 and 99')
```



Maybe also looking at agewed>age

```
# BAD = subset(GSS,agewed > age)
# table(BAD$agewed)
```

```
#maybe this helps too:
table(GSS$agewed, GSS$marital)
```

```
##
##       married widowed divorced separated never married   NA
##   0         0       0        0         0           286    0
##   13        1       0        0         0             0    0
##   14        1       2        1         0             0    0
##   15        0       1        5         1             0    0
##   16       12       9        9         2             0    0
##   17       22       8       11         2             0    0
##   18       62      24       27         5             0    0
##   19       77      19       28         5             0    0
##   20       85      12       22         2             0    0
##   21       97      13       19         3             0    0
##   22       68       9       17         2             0    0
##   23       60      11       10         1             0    0
##   24       58      12        8         4             0    0
##   25       49      10       11         2             0    0
##   26       45       5        9         2             0    0
##   27       34       6        8         1             0    0
##   28       22       2        3         0             0    0
##   29       26       1        7         0             0    0
##   30       18       3        3         1             0    0
##   31       11       1        3         3             0    0
##   32       14       5        1         1             0    0
##   33        8       0        1         1             0    0
##   34        3       2        0         0             0    0
##   35        3       3        1         0             0    0
##   36        6       0        0         0             0    0
##   37        2       0        1         0             0    0
##   38        2       1        1         0             0    0
##   40        3       0        0         0             0    0
##   41        1       0        0         0             0    0
##   42        0       2        0         0             0    0
##   43        0       1        0         0             0    0
##   45        0       0        0         1             0    0
##   47        1       0        0         0             0    0
##   49        0       1        0         0             0    0
##   50        0       0        1         0             0    0
##   54        1       0        0         0             0    0
##   58        1       0        0         0             0    0
##   99        2       2        6         1             0    1
```

```
GSS$agewed = recode(GSS$agewed, recodes="0=NA;99=NA")
# see the table again
table(GSS$agewed,useNA='always')
```

```
##
```

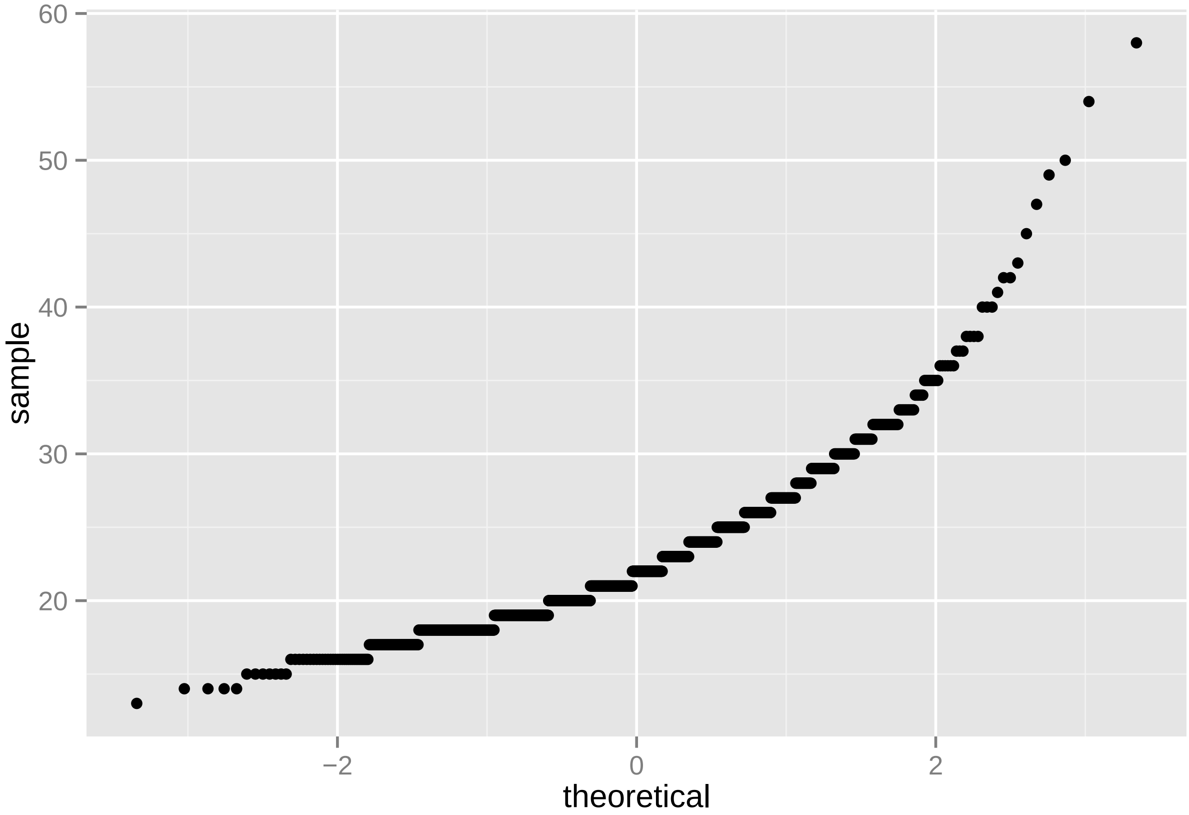
```
##   13   14   15   16   17   18   19   20   21   22   23   24   25   26   27
##    1    4    7   32   43  118  129  121  132   96   82   82   72   61   49
##   28   29   30   31   32   33   34   35   36   37   38   40   41   42   43
##   27   34   25   18   21   10    5    7    6    3    4    3    1    2    1
##   45   47   49   50   54   58 <NA>
##    1    1    1    1    1    1  298
```

```
# What is the mean?
mean(GSS$agewed, na.rm=TRUE)
```

```
## [1] 22.79201
```

```
# qq-plot: far from normal
qplot(sample = GSS$agewed,stat='qq')
```

```
## Warning: Removed 298 rows containing missing values (stat_qq).
```



```
# Shapiro-Wilk test: The null of normality is rejected
shapiro.test(GSS$agewed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  GSS$agewed
## W = 0.8896, p-value < 2.2e-16
```

```
# now, looking at variances.
# report "mean" and "variance" by gender
by(GSS$agewed,GSS$sex, var,na.rm=TRUE  )
```

```
## GSS$sex: Male
## [1] 23.6843
## --------------------------------------------------------
## GSS$sex: Female
## [1] 24.29948
```

```
# We can't reject the null of homoskedasticity
leveneTest(GSS$agewed,group=GSS$sex)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  0.9609 0.3272
##       1200
```

Null: Mean = 23

```
# z test: null's mean = 23, given population sd = 5
n.obs = sum( !is.na(GSS$agewed))
zscore = (mean(GSS$agewed,na.rm=T) - 23)/ 5 * sqrt(n.obs)
print(zscore)
```

```
## [1] -1.442174
```

```
# calculate a two-tailed pvalue.
# I take absolute value to always calculate my pvalue based on positive zscore
pvalue = 2 * pnorm( -abs(zscore) )
cat(' p-value = ',pvalue)
```

```
##  p-value =  0.1492532
```