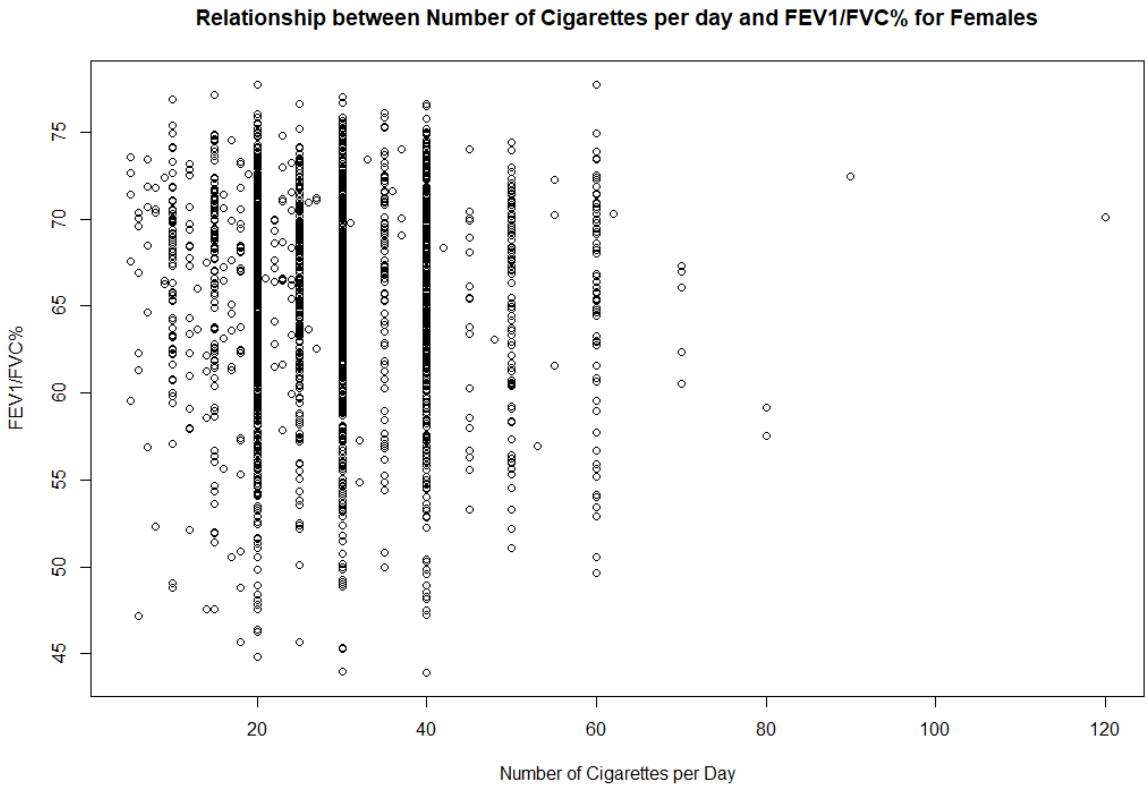
R Assignment 2

Code segments provided for each question, full R script for the assignment at the bottom

Subsetting Data

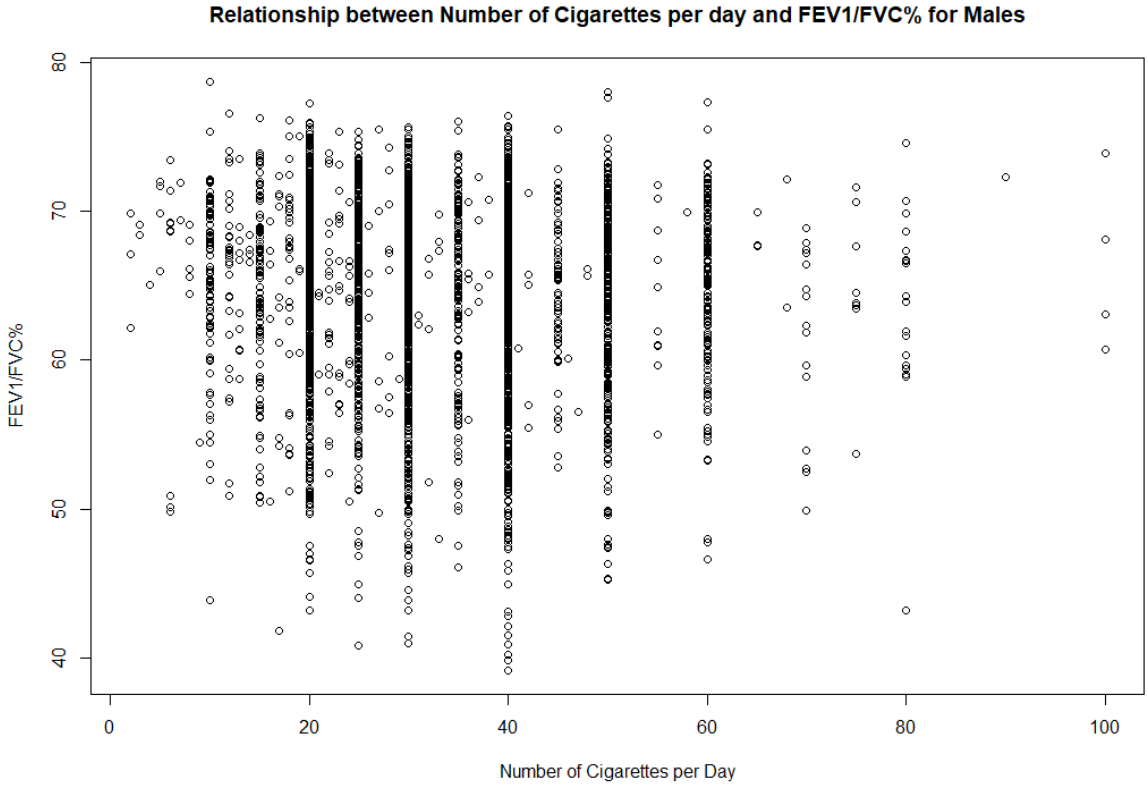
> Investigate the relationship between the FEV1/FVC% at baseline (FEVFVC02; Y) and the number of cigarettes smoked per day at baseline (f10cigs; X) for the data set that contains only females (lhs_f) by creating a scatterplot.

```
# subset data to only choose rows where AGENDER == "F"
lhs.f1<-lhs[lhs$AGENDER=="F", ]

# scatterplot for relationship between FEV1/FEC% at baseline and
# cigarettes smoked per day for new data set with females only
plot(lhs_m$f10cigs, lhs_m$FEVFVC02,
     xlab = "Number of Cigarettes per Day",
     ylab = "FEV1/FVC%",
     main = "Relationship between Number of Cigarettes per day and FEV1/FVC% for
Males")
```



> Create a data set that contains only males (lhs_m). Investigate the relationship between the FEV1/FVC% at baseline (FEVFVC02; Y) and the number of cigarettes smoked per day at baseline (f10cigs; X) for the data set that contains only males (lhs_m) by creating a scatter plot.

```
lhs_m <- lhs[lhs$AGENDER=="M", ]
plot(lhs.f1$f10cigs, lhs.f1$FEVFVC02,
     xlab = "Number of Cigarettes per Day",
     ylab = "FEV1/FVC%",
     main = "Relationship between Number of Cigarettes per day and FEV1/FVC% for
            Females")
```



> Describe what you see in each plot using terms from class (associated or independent, what is the direction and strength of the association, do you see outliers etc). Compare your summaries of males to those of females.

Relationship between Number of Cigarettes per day and FEV1/FVC% for Females:
A weak, positive linear association seems to be between the Number of Cigarettes per day and FEV1/FVC% for females. The two points on the top right corner can be considered outliers, as they are the only two point grater than 80 on the x-axis while being much higher on the y-axis than the other points with a high value for Numbers of Cigarettes per Day. The distribution of all points is right-skewed.

Relationship between Number of Cigarettes per day and FEV1/FVC% for Males:
There is no clear association between Number of Cigarettes per day and FEV1/FVC% for males, which makes these two variables independent variables for this population. The distribution is slightly right-skewed with the four points greater than 80 on the x-axis being possible outliers.

The plot for males is much less right-skewed that the plot for females, with more but less obvious possible outliers. Additionally, while there seems to be a weak, positive linear associations for females, no association is visible for males.

Summarizing Data

> How many of the participants in the LHS study were located at Clinic G (see ACLINIC variable)?

```
table(lhs$ACLINIC)

  A   B   C   D   E   F   G   H   I   J
618 607 613 594 510 530 673 519 615 608
```

673 participants were located at Clinic G.

> What is the summary of number of cigarettes per day at baseline (see f10cigs variable)? What is the mean number of cigarettes per day at baseline?

```
summary(lhs$f10cigs)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   20.00   30.00   31.27   40.00  120.00
```

The complete summary of number of cigarettes per day is listed above as a result from `summary(lhs$f10cigs)`, the mean number of cigarettes per day is 31.27.

> Which of the following variables have missing values: number of years of education at baseline (yearedu), weight change (kg) from baseline to 2nd annual visit (wgtchg02), salivary cotinine at 4th annual visit (COT4)?

```
na_yearedu <- sum(is.na(lhs$yearedu))
na_wgtchg02 <- sum(is.na(lhs$wgtchg02))
na_COT4 <- sum(is.na(lhs$COT4))

na_yearedu
[1] 0
na_wgtchg02
[1] 550
na_COT4
[1] 3452
```

The variable number of years of education at baseline does not have any missing values, the variables weight change (kg) from baseline to 2nd annual visit and salivary cotinine at 4th annual visit both have missing values (550, 3452).

> Does it make sense to interpret the summaries computed for the race variable? Why or why not?

It does not make sense to interpret the summaries computed for the race variable. The variable is a categorical nominal variable, as there is no order to race, but as the races are coded as numbers, R interprets it as numerical values. It never makes sense to interpret mathematical results such as the mean for categorical variables.

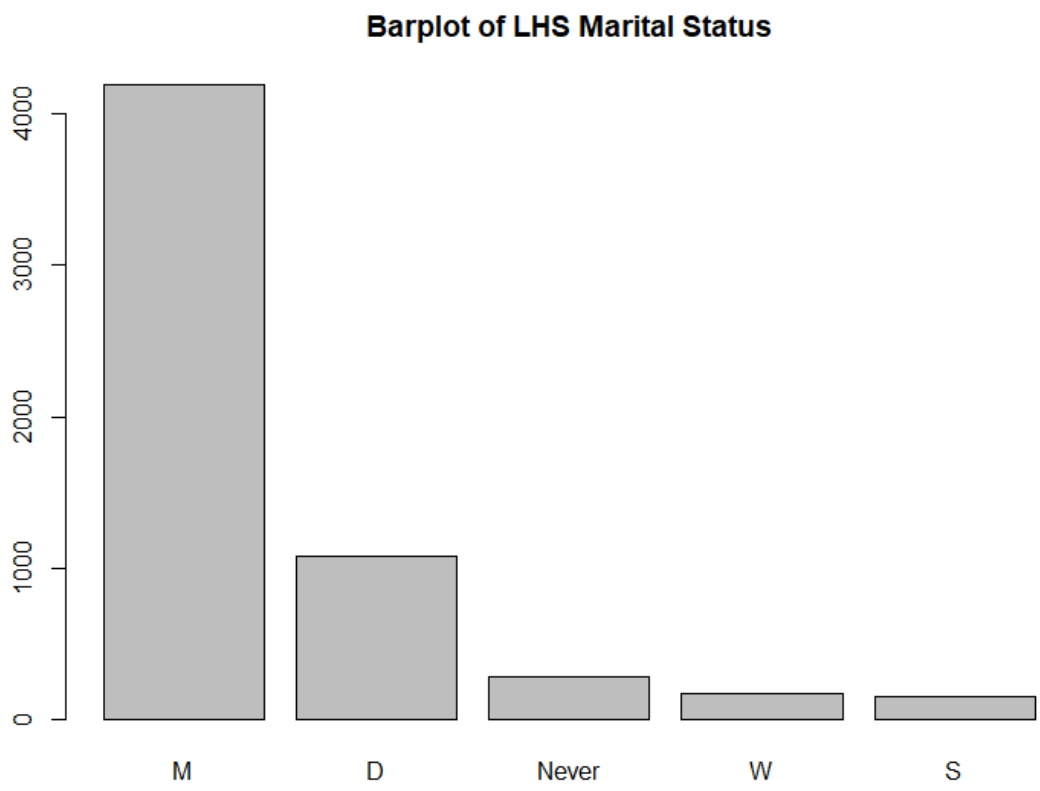
> For the marital status variable (marital0), what does the category of 1 represent? How many 1's are there in the data set?

```
table(lhs$marital0)

   1    2    3    4    5
 288 4188  177  152 1082
```

Category 1 for the marital status represents *Never Married* according to the documentation. There are 288 1's in the data set.

> Using the bar plot, write down the marital status groups in descending order (from largest to smallest).

```
table_marital0 <- table(lhs$marital0)
names(table_marital0) <- c("Never","M","W","S","D")
table_marital0 <- sort(table_marital0, decreasing = TRUE)
barplot(table_marital0, main = "Barplot of LHS Marital Status")
```



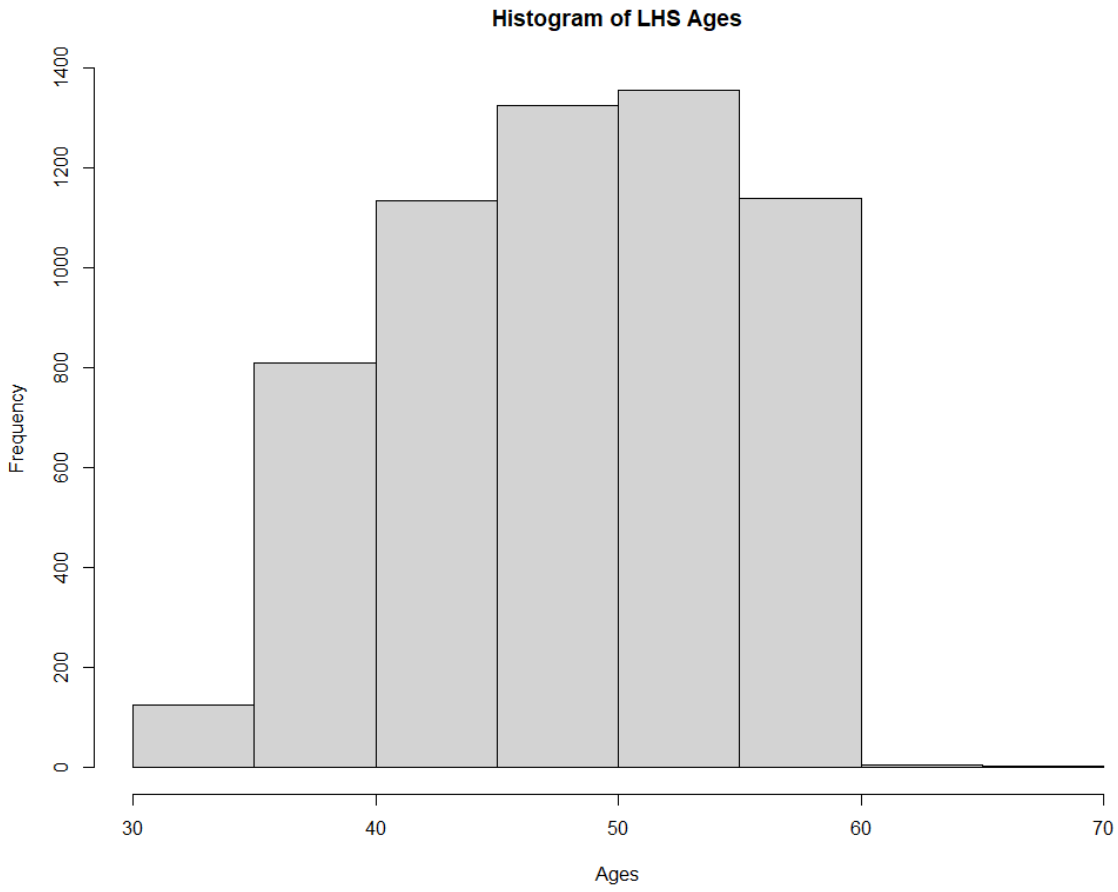
> What is the mean age of the participants in the LHS study? What is the median age? The standard deviation of age? The minimum and maximum age?

```
summary(lhs$age)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  34.00   43.00   49.00   48.47   54.00   67.00 

sd(lhs$age)
[1] 6.82513
```

The mean age for participants in the LHS study is 48.47 years, the median 49.00 and the standard deviation of age is 6.82513. The minimum age is 34.00 and the maximum 67.00.

> Using the histogram, approximately how many participants were between 30 and 35 years old at the beginning of the LHS study?

```
hist(lhs$age, main="Histogram of LHS Ages", xlab="Ages", breaks = 10)
```



According to the histogram, approximately 130 people were between 30 and 35 years old at the beginning of the LHS study.

> Using the counts object, figure out how to create a table of proportions instead of counts for the marital status variable. [Hint: Use prop.table() function.]

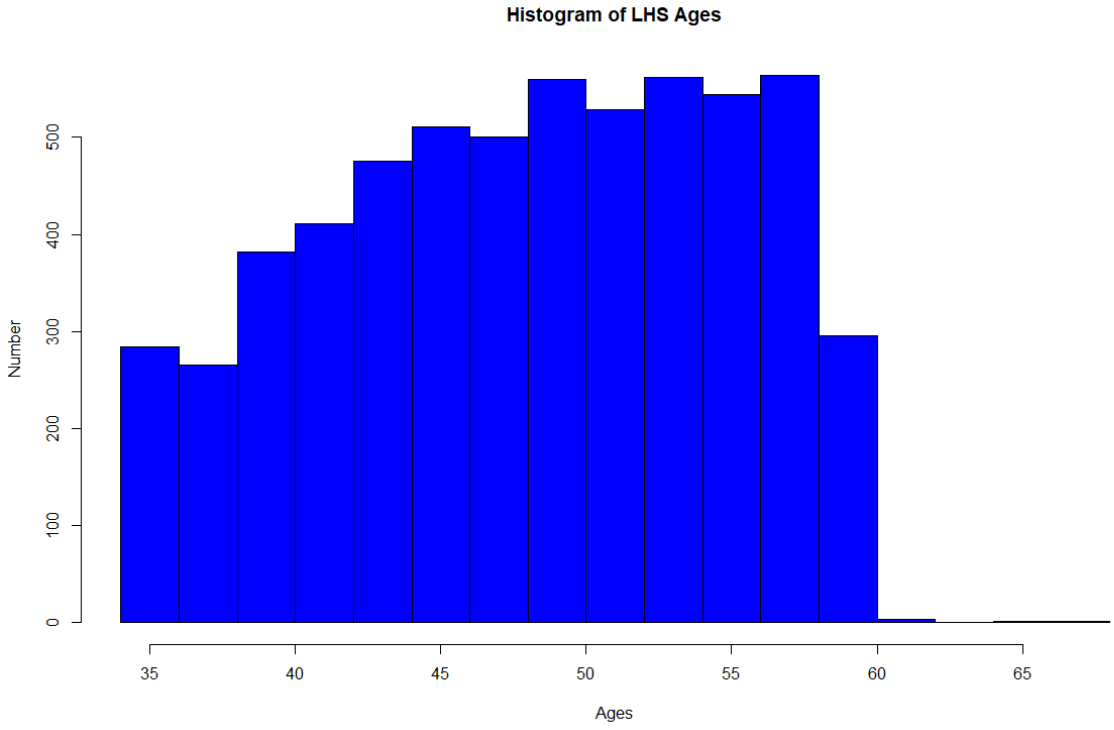
```
counts<-table(lhs$marital0)
names(counts) <- c("Never","M","W","S","D")
prop.table(counts)

    Never         M         W         S         D
0.04892135 0.71139800 0.03006625 0.02581960 0.18379480
```

> Figure out how to make the color of bars in the bar plot for marital status, marital0, to be blue, and how to change the y-axis title to be “Number”.

```
hist(lhs$age, main = "Histogram of LHS Ages",
     xlab = "Ages", ylab = "Number",
     col = "blue")

barplot(lhs$age, main = "Barplot of LHS Ages",
     xlab = "Ages", ylab = "Number",
     col = "blue")
```

Changing the color for a histogram or bar plot is done by setting the `col` property to a different value. The value can either be a string with a predefined color or a customized color specified using the `rgb()` function.

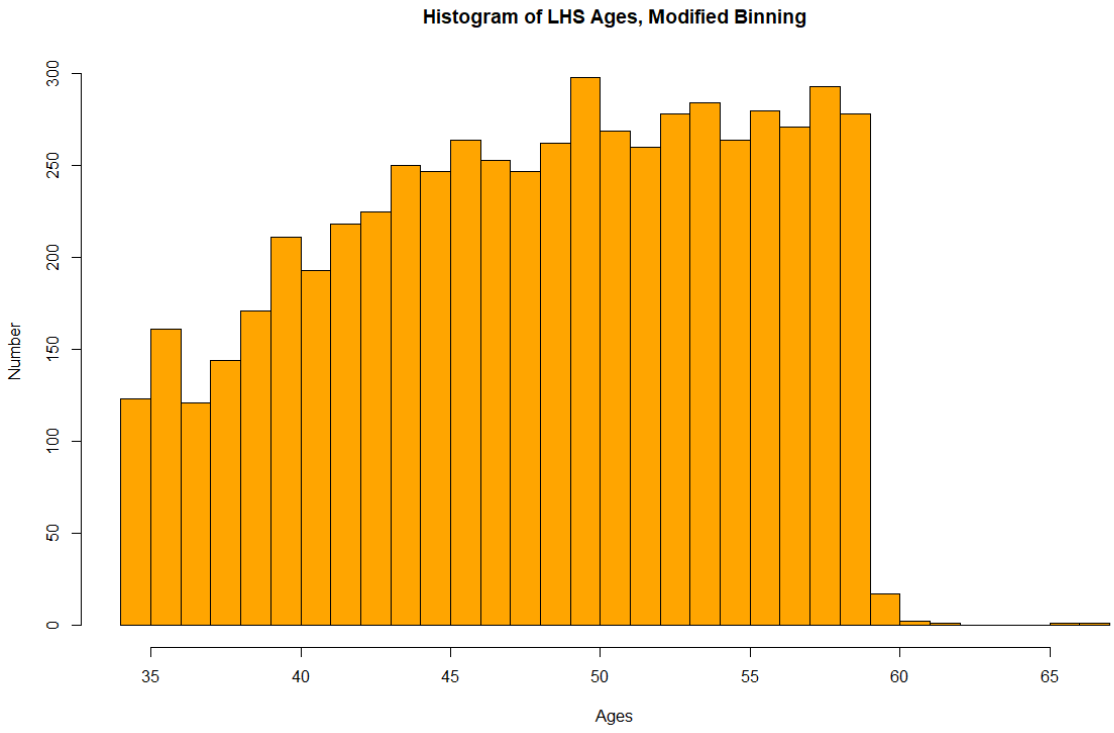
The title for the x-axis and y-axis can be changed by setting the `xlab` and `ylab` properties respectively. Each property takes a string value that will be displayed as the title for the axis.



> Figure out how to make the color of the bars in the histogram for age, age, to be orange, and how to modify the binning of the histogram to your liking.

```
hist(lhs$age, main = "Histogram of LHS Ages, Modified Binning",
     xlab = "Ages", ylab = "Number",
     col = "orange",
     breaks = 25)
```

The color will be changed using the `col` property as before, this time using `"oragne"` as a value. The binning can be modified with the `breaks` property. That property sets the number of bins used for in the histogram, here `25`.



Complete Script for R Assignment 2

```
install.packages("psych")

# Load data, choose lhs.csv
lhs <- read.csv(file = file.choose(), header = TRUE)

# subset data to only choose rows where AGENDER == "F"
lhs.f1<-lhs[lhs$AGENDER=="F", ]
# Alternative
# lhs.f2<-subset(lhs, lhs$AGENDER=="F")

# scatterplot for relationship between FEV1/FEC% at baseline and
# cigarettes smoked per day for new data set with females only
plot(lhs.f1$f10cigs, lhs.f1$FEVFVC02,
     xlab = "Number of Cigarettes per Day",
     ylab = "FEV1/FVC%",
     main = "Relationship between Number of Cigarettes per day and FEV1/FVC% for
Females")


lhs_m <- lhs[lhs$AGENDER=="M", ]
plot(lhs_m$f10cigs, lhs_m$FEVFVC02,
```

```
     xlab = "Number of Cigarettes per Day",
     ylab = "FEV1/FVC%",
     main = "Relationship between Number of Cigarettes per day and FEV1/FVC% for
Males")

# summaries for all variables in data set depending on the data type
summary(lhs)

table(lhs$ACLINIC)
summary(lhs$f10cigs)

na_yeareduc <- sum(is.na(lhs$yeareduc))
na_wgtchg02 <- sum(is.na(lhs$wgtchg02))
na_COT4 <- sum(is.na(lhs$COT4))

table_marital0 <- table(lhs$marital0)
names(table_marital0) <- c("Never","M","W","S","D")
table_marital0 <- sort(table_marital0, decreasing = TRUE)
barplot(table_marital0, main = "Barplot of LHS Marital Status")

summary(lhs$age)
sd(lhs$age)

# Alternatively:
###Access the package using the library() function
library(psych)
###Calculate multiple summary statistics using the describe() function
describe(lhs$age, IQR=TRUE, quant=c(.25,.75))

hist(lhs$age, main="Histogram of LHS Ages", xlab="Ages", breaks = 10)

counts<-table(lhs$marital0)
names(counts) <- c("Never","M","W","S","D")
prop.table(counts)

barplot(lhs$age, main = "Barplot of LHS Ages",
     xlab = "Ages", ylab = "Number",
     col = "blue")

hist(lhs$age, main = "Histogram of LHS Ages, Modified Binning",
     xlab = "Ages", ylab = "Number",
     col = "orange",
     breaks = 25)
```