# Week 1 Tutorial: Sampling from a population, summarizing data, and the normal distribution (EXAMPLE SOLUTIONS)

Introduction to Statistical Thinking and Data Analysis

*MSc in Epidemiology and MSc in Health Data Analytics, Imperial College London*

*7 October 2019*

1. The dataset `perulung_ems.csv` contains data from a study of lung function among a sample of 636 children aged 7 to 10 years living in a deprived suburb of Lima, Peru, introduced on page 27 of Kirkwood and Sterne. FEV1 is the *forced expiratory volume* in 1 second, the maximum amount of air which children could breath out in 1 second measured using a spirometer.

| Variable | Description |
|---|---|
| id | Participant ID number |
| fev1 | Forced Expiratory Volume in 1 second |
| age | Age in years |
| height | Height in centimeters |
| sex | Sex (0 = female, 1 = male) |
| respsymptoms | Presence of respiratory symptoms (0 = no symptoms; 1 = symptoms) |

   a. What type of variable is each variable in the dataset?

```
str(data1)
```

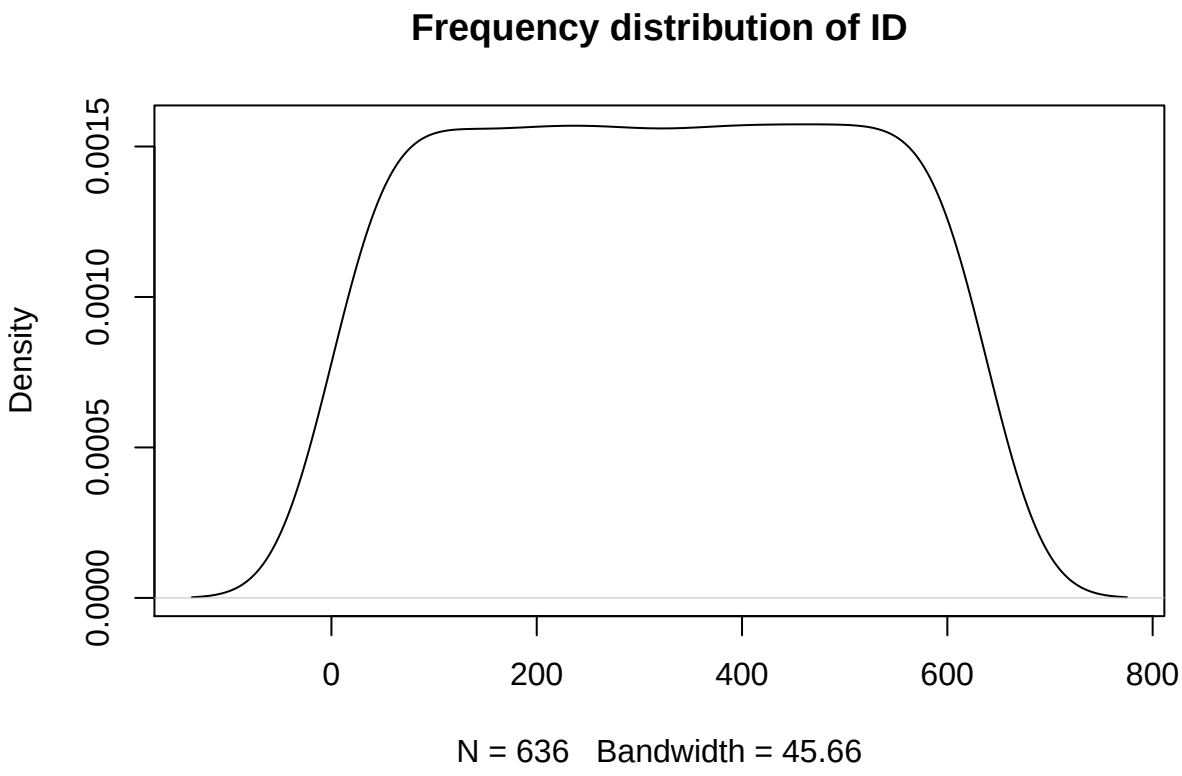
```
## 'data.frame':    636 obs. of  6 variables:
##  $ ï..id       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fev1        : num  1.56 1.18 1.87 1.49 1.62 2.11 1.73 1.47 1.83 1.41 ...
##  $ age         : num  9.59 7.49 9.86 8.59 8.97 ...
##  $ height      : num  125 111 136 119 121 ...
##  $ sex         : int  0 1 0 0 1 0 1 0 1 0 ...
##  $ respsymptoms: int  0 0 0 0 0 1 0 1 0 0 ...
```

- sex and respsymptoms are binary variables although encoded in the dataframe as integers.

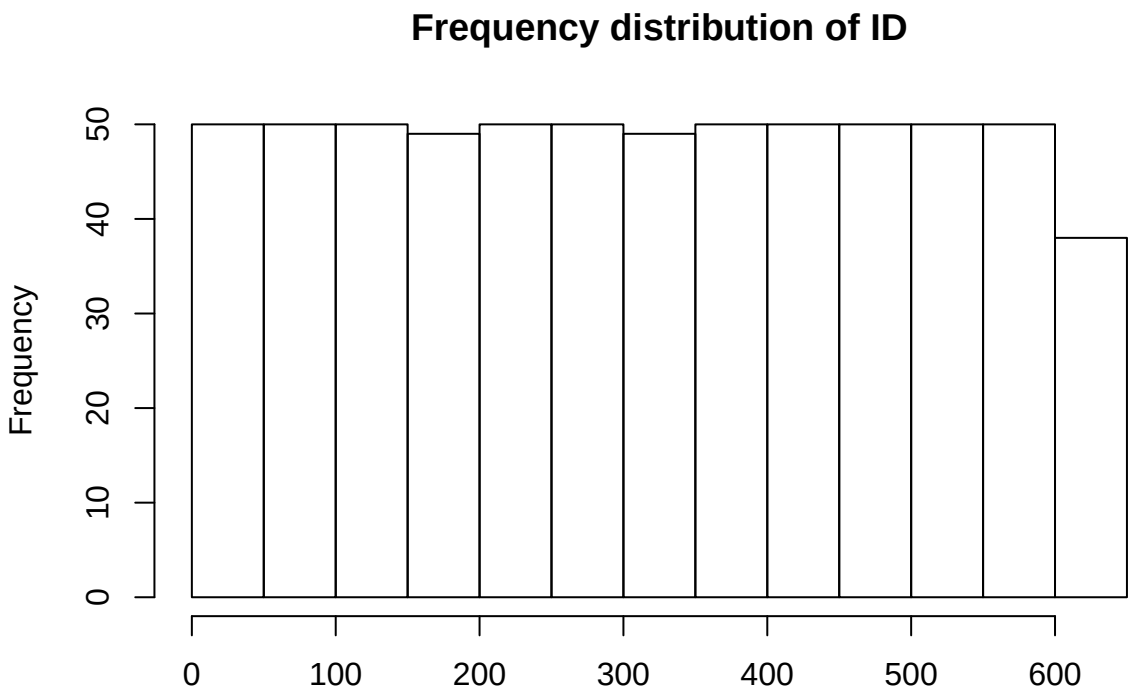
   b. What shape is the frequency distribution of each of the variables?

- id: each appears just once so the data are uniformly distributed.

```
plot(density(data1$ï..id),main="Frequency distribution of ID")
```

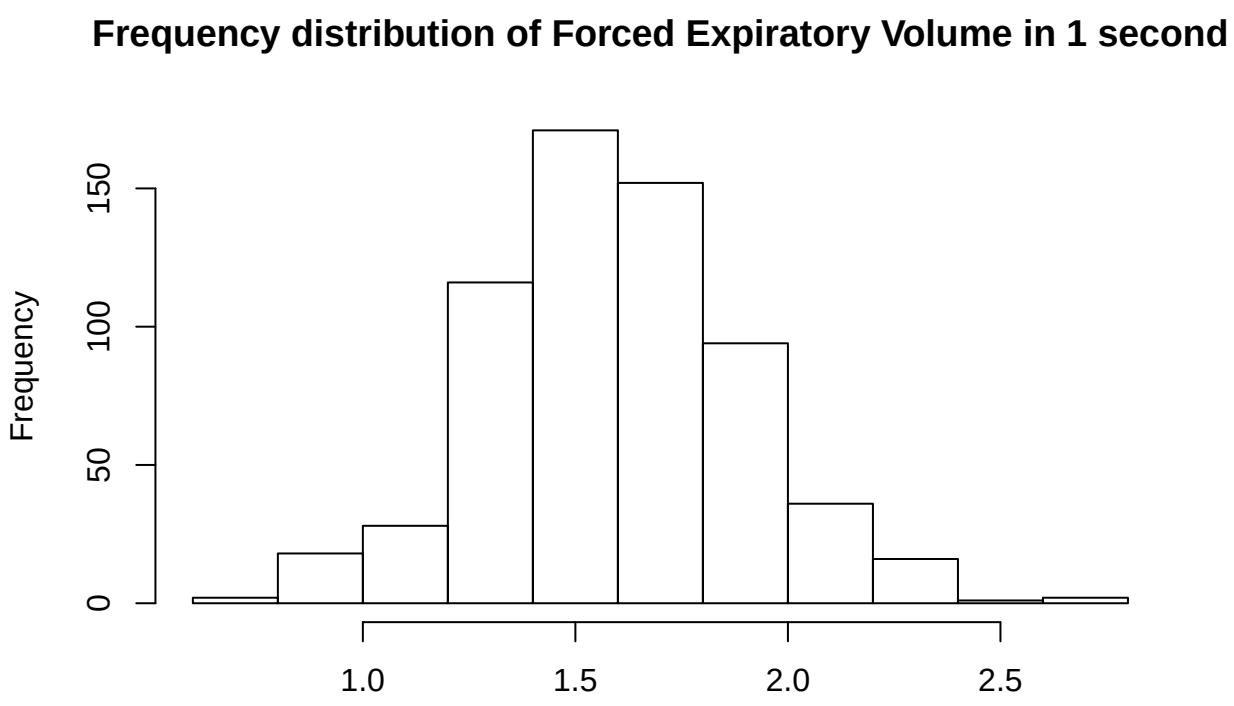



```r
hist(data1$ï..id,main="Frequency distribution of ID",xlab="")
```



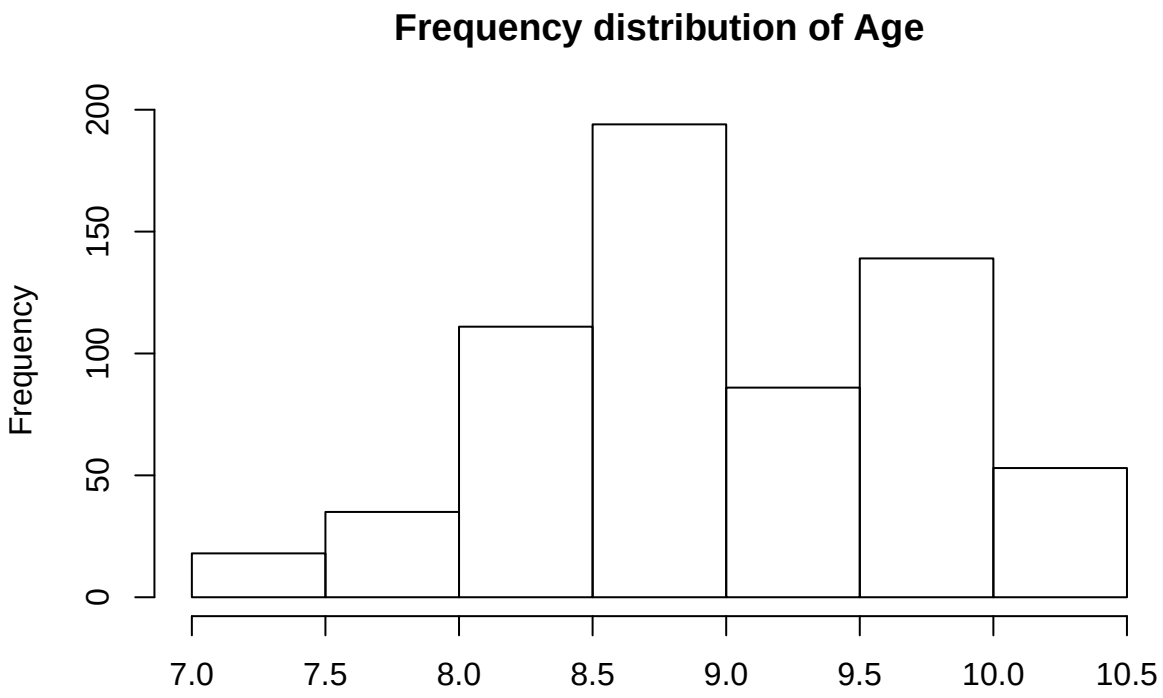

- fev1: normally distributed

```
#plot(density(data1$fev1),main="Frequency distribution of Forced Expiratory Volume in 1 second")
hist(data1$fev1,main="Frequency distribution of Forced Expiratory Volume in 1 second",xlab="")
```



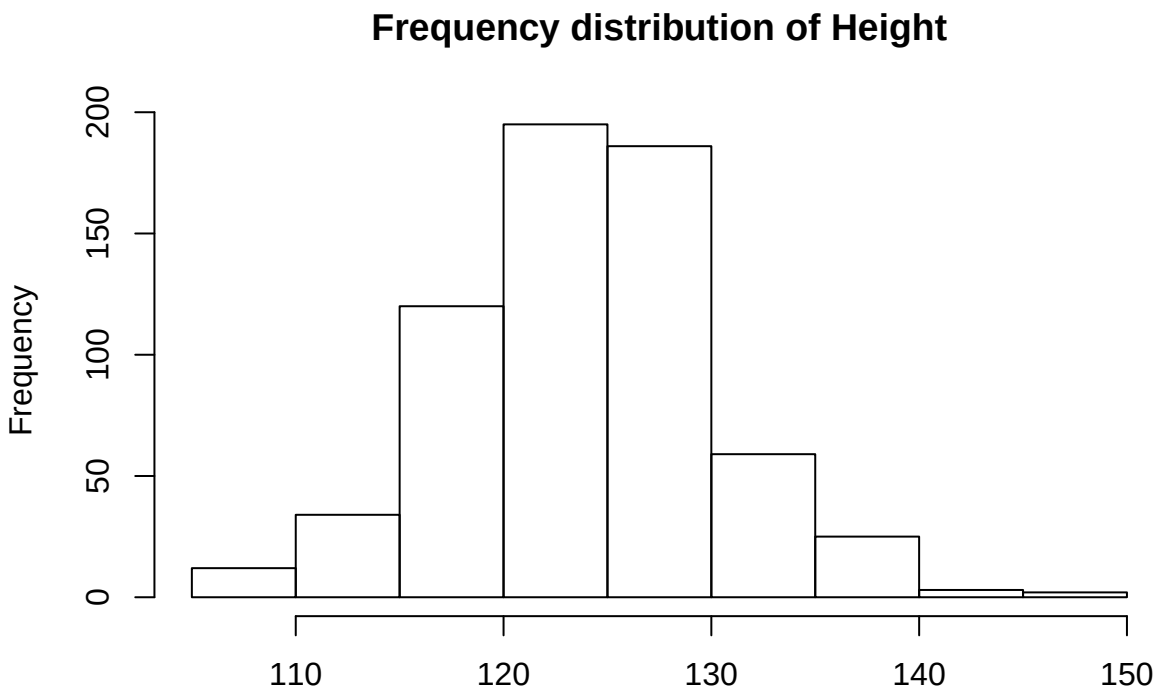

- age: bimodal

```
#plot(density(data1$age),main="Frequency distribution of Age")
hist(data1$age,main="Frequency distribution of Age",xlab="")
```



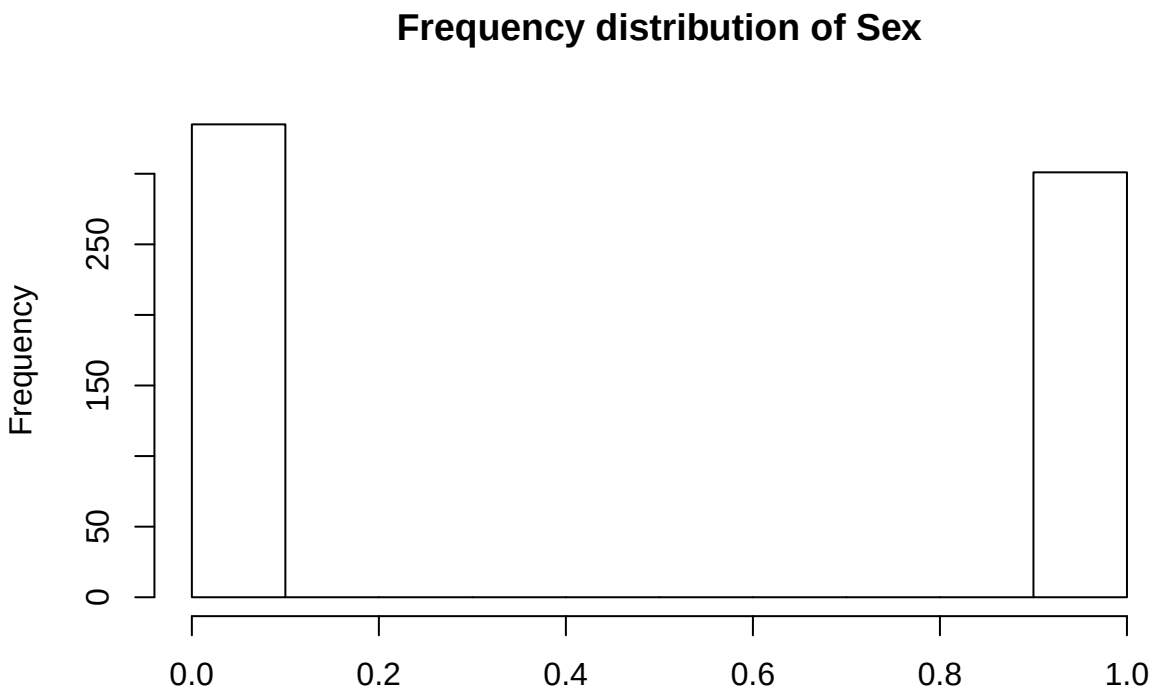

- height: normally distributed

```
#plot(density(data1$height),main="Frequency distribution of Height")
hist(data1$height,main="Frequency distribution of Height",xlab="")
```



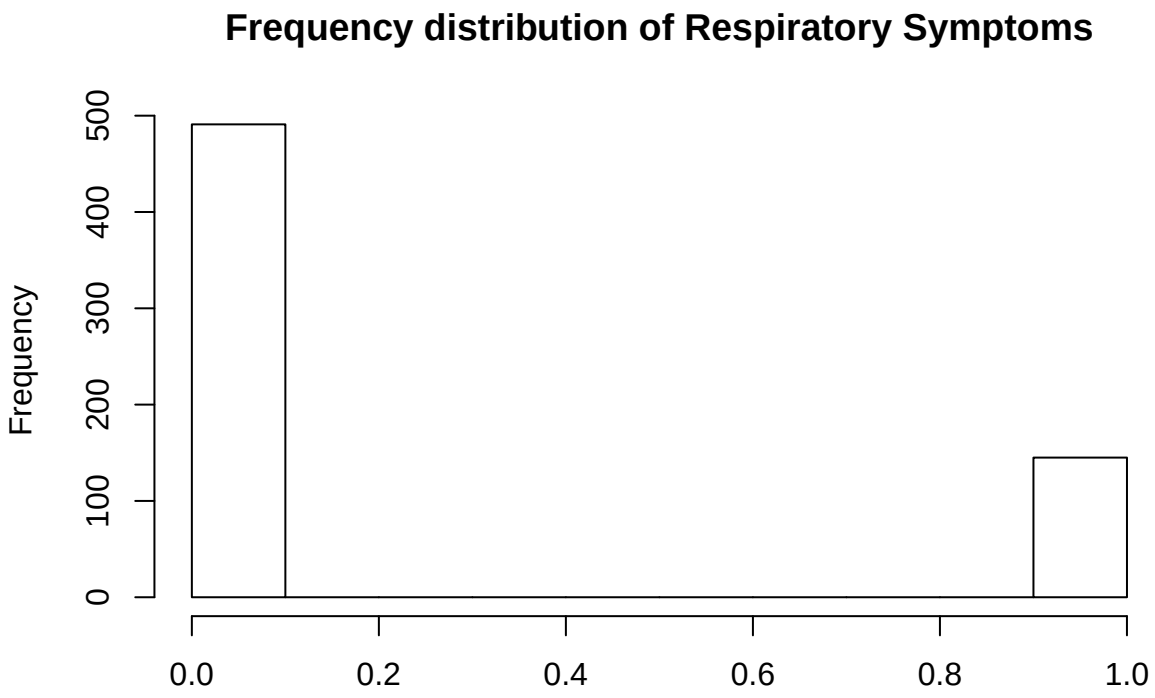

- sex: bimodal (response fairly well balanced)

```
#plot(density(data1$sex),main="Frequency distribution of Sex")
hist(data1$sex,main="Frequency distribution of Sex",xlab="")
```




- respsymptoms: bimodal (response imbalanced - more without symptoms)

```
#plot(density(data1$respsymptoms),main="Frequency distribution of Respiratory Symptoms")
hist(data1$respsymptoms,main="Frequency distribution of Respiratory Symptoms",xlab="")
```



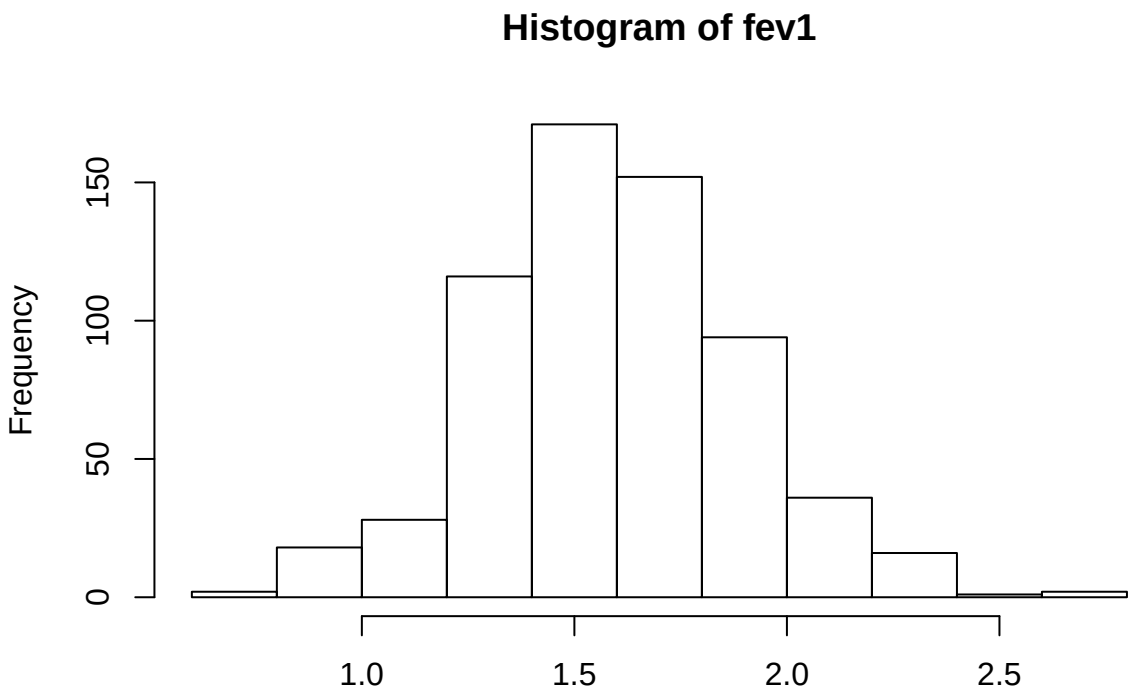

c. What are some research questions which these data could have been collected to address?

- The effects of age and gender in determining whether a child has respiratory symptoms
- The relationship between respiratory symptoms and forced expiratory volumes in 1 second
- The relationship between height and age (are the children under-developed for their age) and how this relates to forced expiratory volumes
- Based on these relationships, what groups of children are most at risk of respiratory illness

d. Use R to create appropriate univariate graphical summaries of each of the variables.

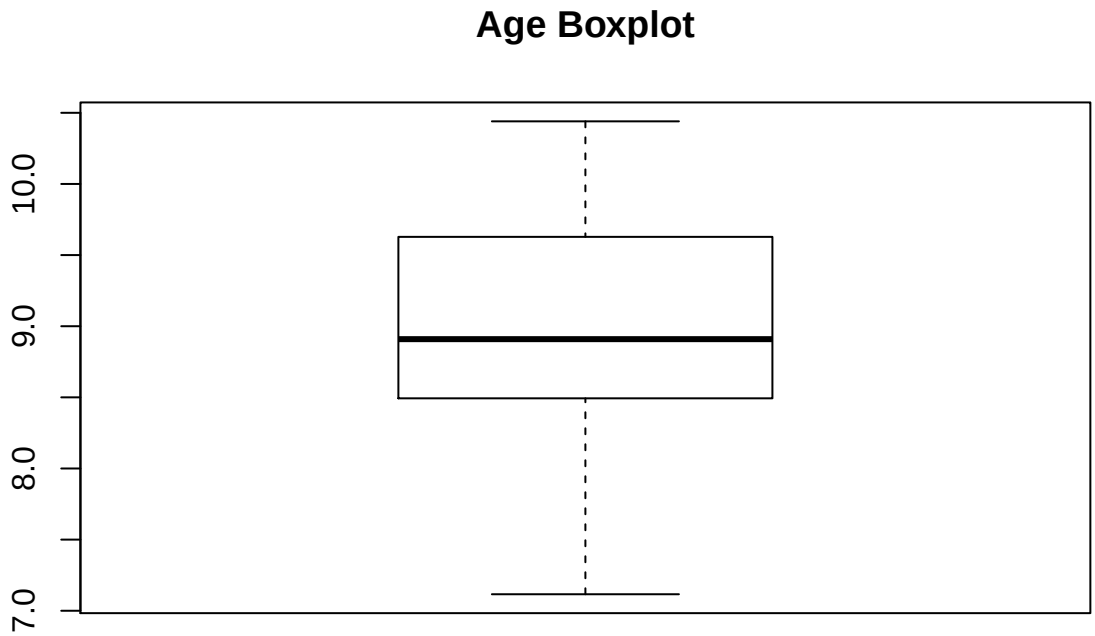
- ID: n/a
- fev1: histogram

```
hist(data1$fev1,main="Histogram of fev1",xlab="")
```



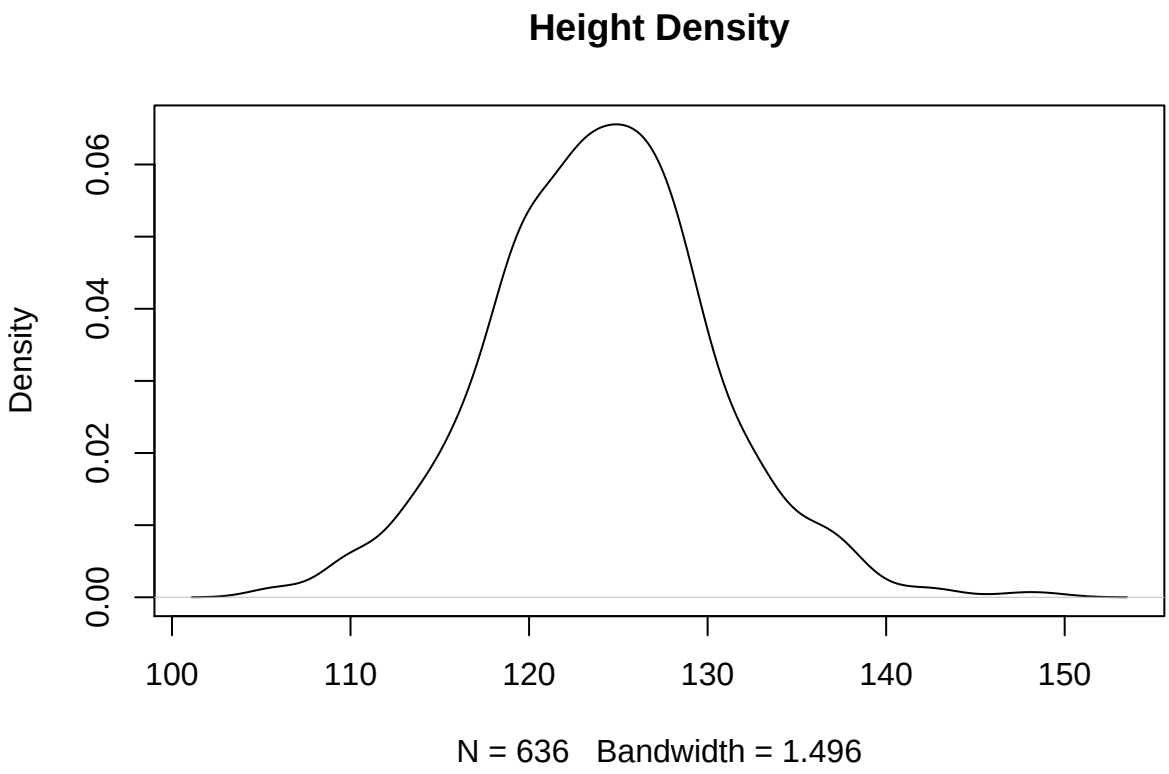

- age: boxplot

```
boxplot(data1$age,main="Age Boxplot")
```



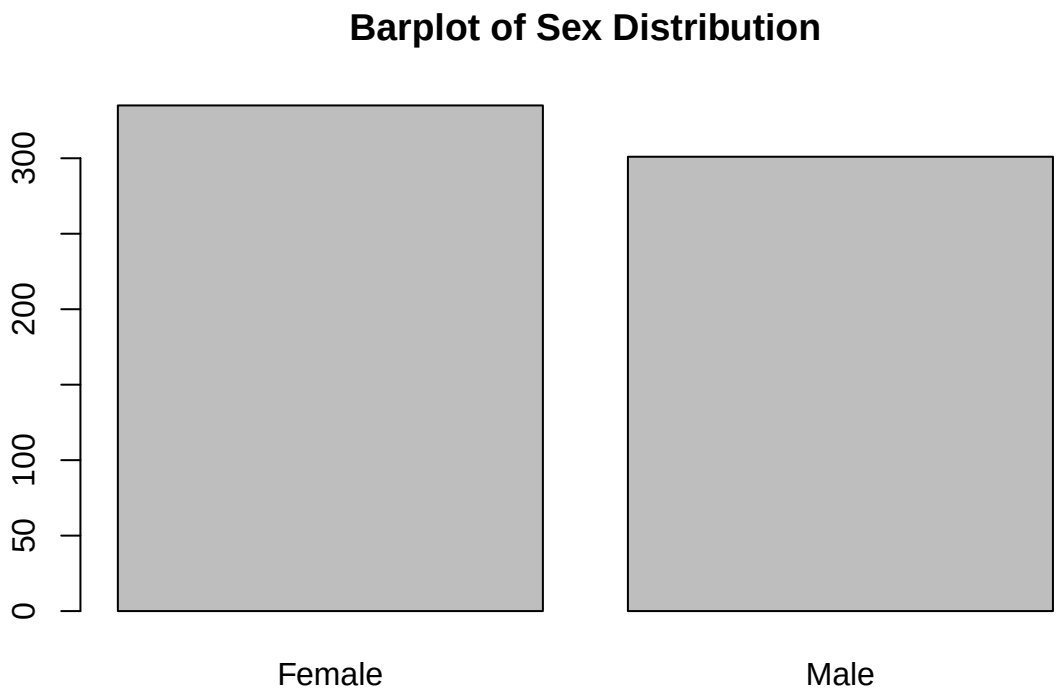

- height: histogram

```
plot(density(data1$height),main="Height Density")
```



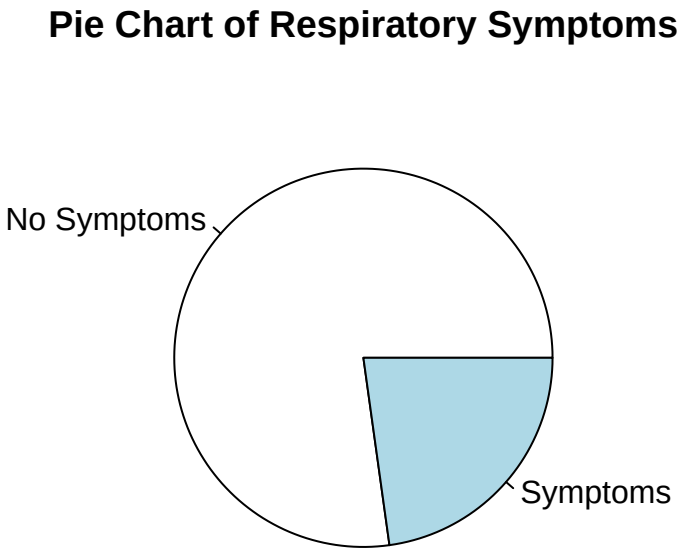

- sex: barplot

```
sex_counts <- table(data1$sex)
barplot(sex_counts,main="Barplot of Sex Distribution",names.arg=c("Female","Male"))
```




- respsymptoms: piechart

```
resp_counts <- table(data1$respsymptoms); labels <- c("No Symptoms","Symptoms")
pie(resp_counts,labels=labels,main="Pie Chart of Respiratory Symptoms")
```




e. Create a single table summarizing key characteristics of the sample—an appropriate 'Table 1' for a medical or epidemiologic paper. (It is probably possible to construct a full table with R commands, but you might find it easier to do calculations of summary statistics with R and copy the R output into a separate table in MS Word, Excel, or similar.)

- continuous variables: fev1, age, height

```
quantile(data1$fev1) #can also do min(data1$fev1); median(data1$fev1); max(data1$fev1)
```

```
##     0%    25%    50%    75%   100%
## 0.6400 1.3975 1.5800 1.7900 2.6900
```

```
mean(data1$fev1)
```

```
## [1] 1.594654
```

```
summary(data1$fev1) #alternatively use this function to get all of this in one line of code
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.640   1.397   1.580   1.595   1.790   2.690
```

```
summary(data1$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.116   8.493   8.909   8.984   9.627  10.440
```

```
summary(data1$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   105.6   119.9   124.0   124.1   128.0   149.0
```

- binary variables: sex, respsymptoms

```
table(data1$sex)
```

```
##
##   0   1
## 335 301
```

```
table(data1$respsymptoms)
```

```
##
##   0   1
## 491 145
```

f. In this sample of 636 children, does there appear to be an association between:

(i) sex and height,
(ii) age and height,
(iii) sex and lung function,
(iv) sex and presence of respiratory symptoms,
(v) respiratory symptoms and lung function.

Support your answers with graphical or numerical evidence.

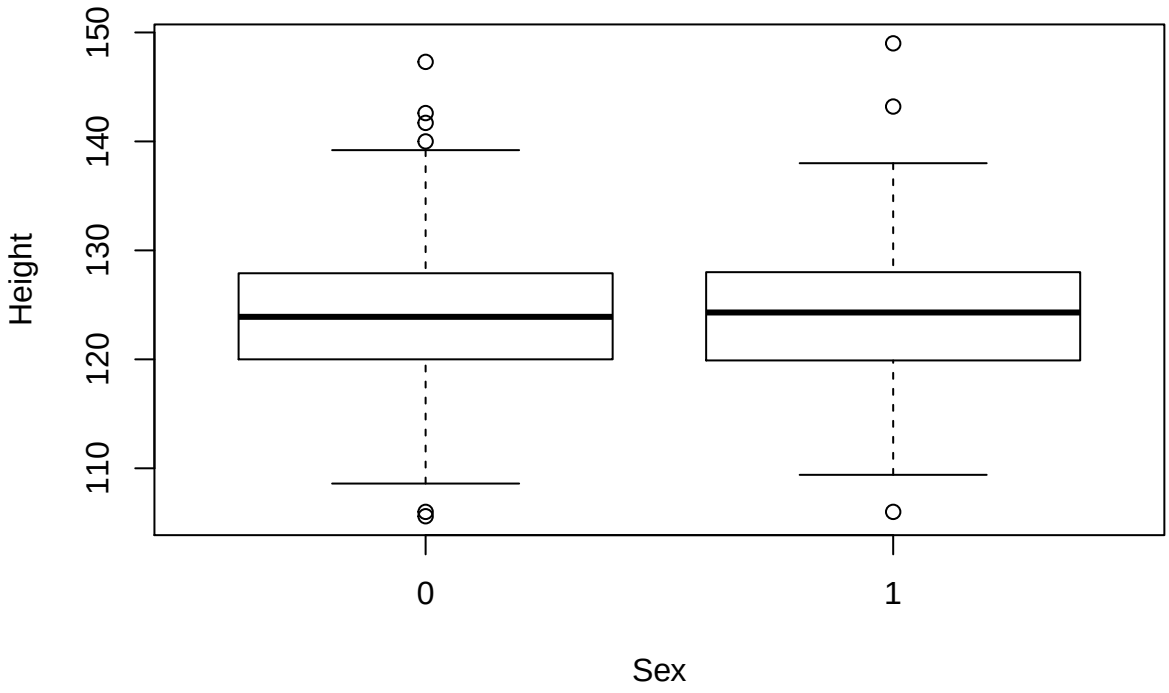
- Use scatterplots or boxplots for graphical evidence and correlations as numerical evidence.

```
cor(data1)
```

```
##                      ï..id        fev1         age       height
## ï..id          1.000000000 -0.01850828 -0.03826454 -0.005488525
## fev1          -0.018508277  1.00000000  0.51616575  0.637609596
## age           -0.038264537  0.51616575  1.00000000  0.594601497
## height        -0.005488525  0.63760960  0.59460150  1.000000000
## sex            0.027554089  0.19512354 -0.03329329  0.006846386
## respsymptoms   0.012394614 -0.20629114 -0.17401164 -0.100283622
##                        sex respsymptoms
## ï..id          0.027554089   0.01239461
## fev1           0.195123537  -0.20629114
## age           -0.033293288  -0.17401164
## height         0.006846386  -0.10028362
## sex            1.000000000  -0.03471080
## respsymptoms -0.034710800   1.00000000
```

(i) sex and height

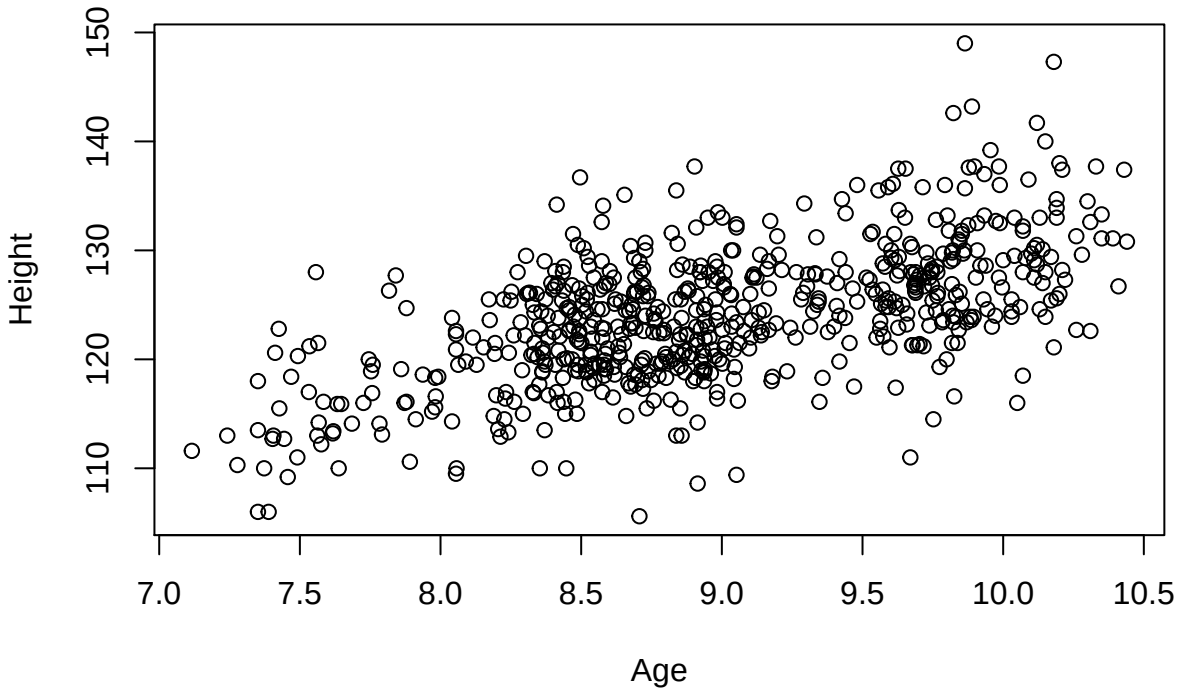
```
boxplot(data1$height~data1$sex,xlab="Sex",ylab="Height")
```




The boxplots are very similar, indicating that there is not an association between the two. This is confirmed by the correlation which is close to 0 (0.007).

(ii) age and height

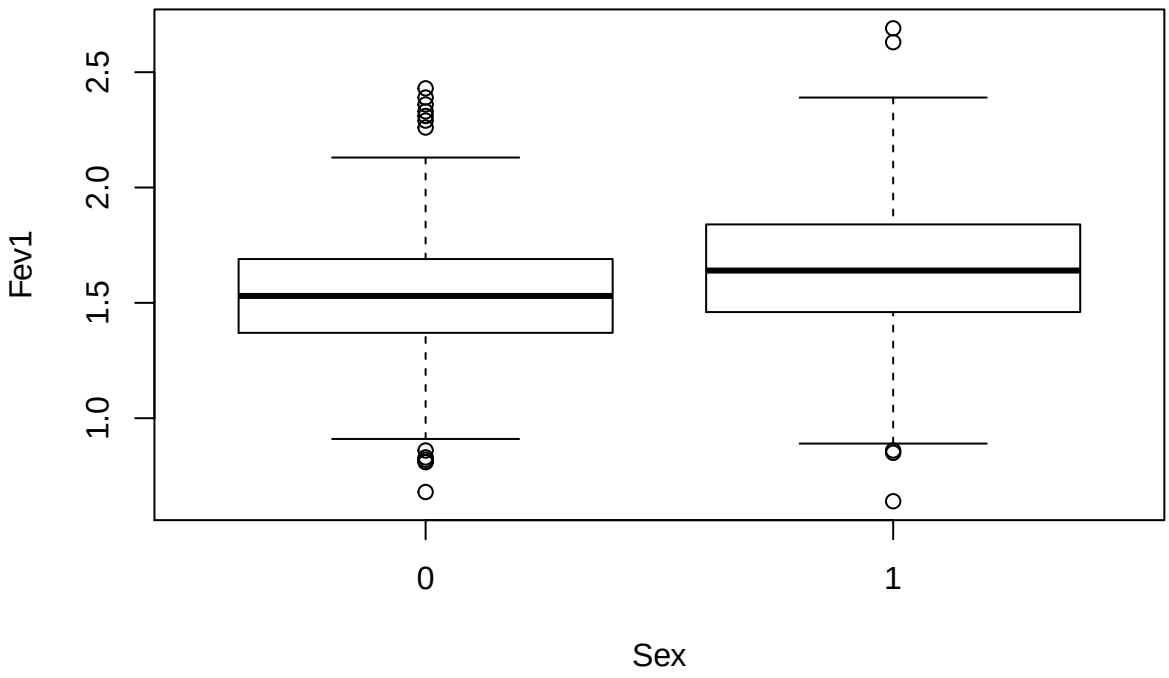
```
plot(data1$height~data1$age,xlab="Age",ylab="Height")
```




The scatterplot indicates a positive association between the two. This is confirmed by the correlation which is 0.59.

(iii) sex and lung function

```
boxplot(data1$fev1~data1$sex,xlab="Sex",ylab="Fev1")
```



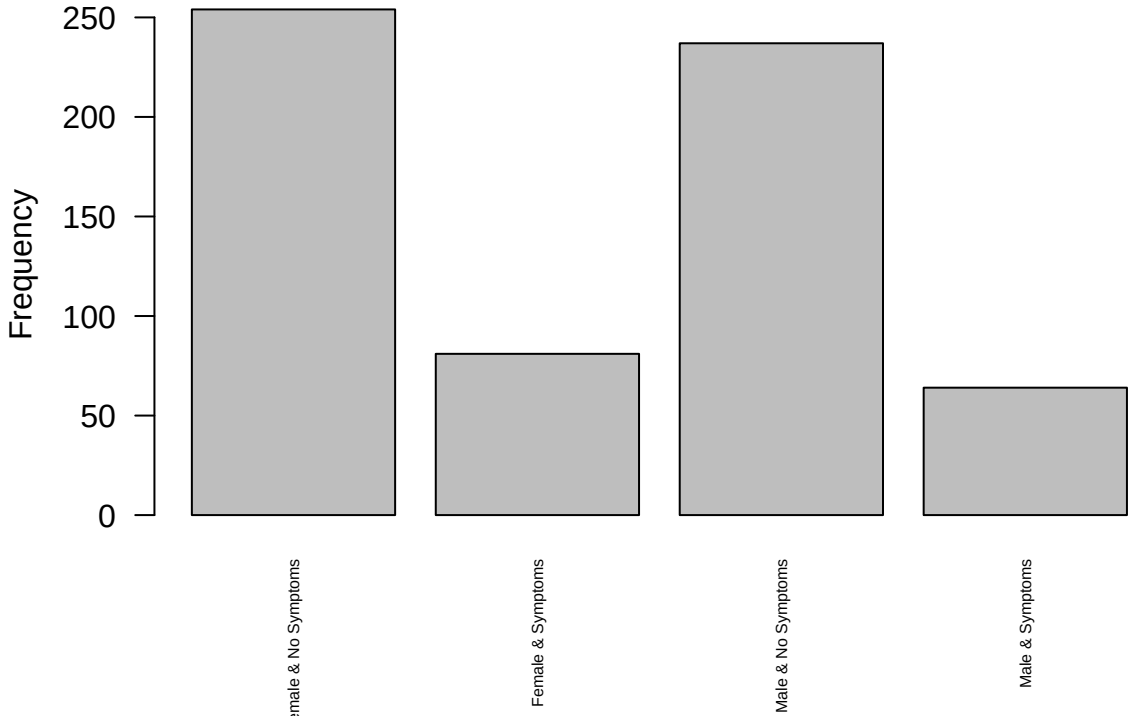

Again the boxplots are very similar. Correlation is 0.2, indicating they are weakly associated.

(iv) sex and presence of respiratory symptoms

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.5.3
```

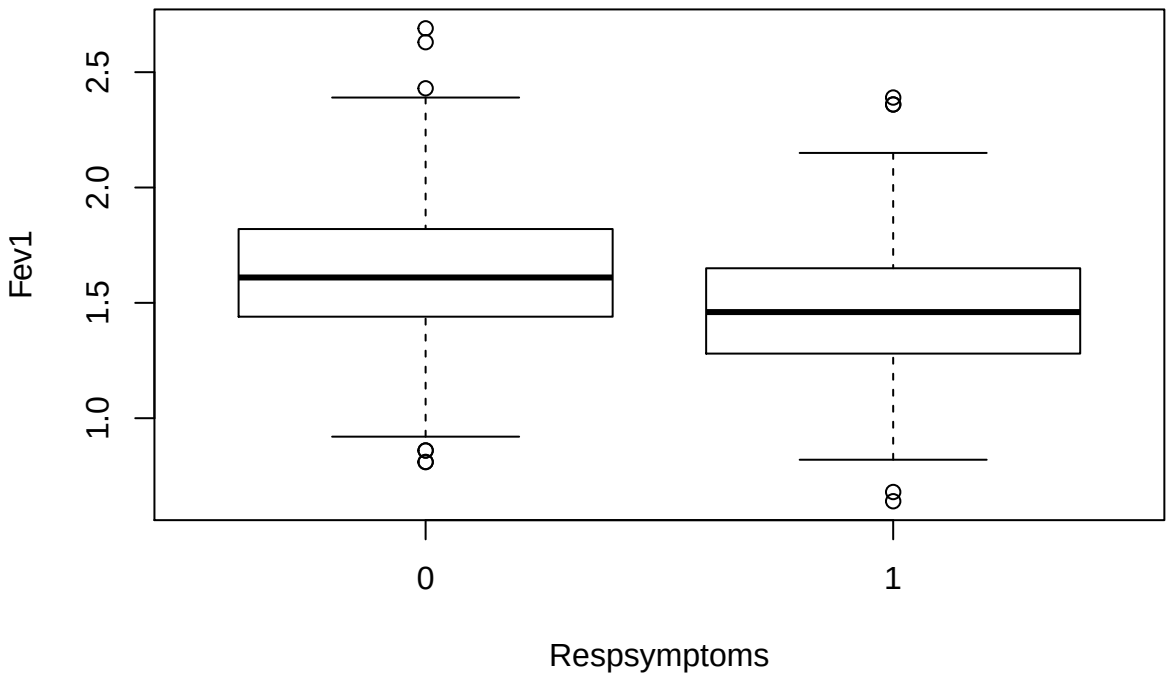
```
sex_resps_count <- count(data1,vars=c("sex","respsymptoms"))
barplot(sex_resps_count$freq,xlab="",ylab="Frequency",names.arg=c("Female & No Symptoms","Female & Symp
```




Similar pattern in respiratory symptoms across male and female. Very weak negative association detected by correlation (-0.035) implying that males are slightly more associated with having respiratory symptoms.

(v) respiratory symptoms and lung function

```
boxplot(data1$fev1~data1$respsymptoms,xlab="Respsymptoms",ylab="Fev1")
```




Those with respiratory symptoms have lower fev1 than those who do not. Correlation confirms this slight (negative) association with a value of -0.21.

g. What is the target population to which your conclusions about these questions might generalize?

- Children from deprived areas in suburban Peru.

2. The *National Health and Nutrition Examination Survey* (NHANES) is a nationally representative survey to assess the health and nutrition of adults and children in the United States. The survey was first conducted in the 1960s and has been conducted continuously since 1999 with around 5000 respondents sampled and interviewed in their homes every year. The survey consists of a combination of questionnaire responses and physical and biomarker measurements. Further information about the survey and datasets can be found here: https://www.cdc.gov/nchs/nhanes/index.htm.

The R package `NHANES` contains an extract of 75 variables about 10,000 respondents to NHANES between 2009 and 2012 abstracted for educational purposes. The actual NHANES survey datasets include sampling weights to account for non-equal sampling probability of certain population groups to increase the statistical efficiency of the survey, which is not covered in this course. The dataset of 10,000 respondents in the `NHANES` R package has been constructed such that the dataset can analysed as if it were a simple random sample from the American population. See the package documentation for information and links about accessing and analysing the actual NHANES data for research purposes; there are other R packages available on CRAN to assist with accessing and processing the actual NHANES survey data.

The objective of this exercise is to practice loading large datasets into R, understanding the structure and variables in a dataset, and conducting exploratory analysis.

a. Open and explore the NHANES dataset in R through the following steps:

- Install the `NHANES` R package: `install.packages("NHANES")`
- Load the R package: `library(NHANES)`
- Load the NHANES dataset into your workspace: `data(NHANES)`
- Use the command `?` to access the help page for the dataset: `?NHANES`

```
## Warning: package 'NHANES' was built under R version 3.5.3
```

Review the dataset edocumentation, particularly looking

(i) What was the purpose for collecting the data?

- To monitor the health and nutrition of children and adults in the United States.

(ii) When and how were the data in the dataset collected?

- Collected from 2009-2012. Data is collected through interviews in the respondent's home and a health examination conducted in a mobile examination centre.

(iii) What is the target population of the sample?

- The non-institutionalised civilian resident population of the United States.

(iv) What is the sample size? Who was eligible to be included in the dataset? Are there different eligibility or inclusion criteria for certain variables?

- Sample size: raw data has 20293 obs, the NHANES data itself has been resampled to have 10000 rows (to account for sampling biases).
- Eligibility: think it is everyone in US (from NHANES website)
- Different eligibility criteria for certain variables: educational level and marital status for participants aged 20 or over only; length only for children under 3 etc.

(v) What are the areas of information available in the dataset?

- Examples include age, race, gender, education and marital status, indicators related to poverty, blood pressure (lots more).

In the documentation, note that several of the variables are only collected for respondents of a certain age range or in one of the survey rounds but not the other. This is important to take note because it may affect what questions can be addressed by the data, or result in errant conclusions and incorrect interpretation if eligibility and inclusion criteria are not appropriately considered during analysis. For the remainder of the tutorial, we will only consider the subset of the sample who are adults aged 20 years and older.

Construct this dataset with the R command: `nhanes20pl <- NHANES[NHANES$Age >= 20, ]`

(vi) Confirm that your new dataset has 7235 respondents remaining. Use at least one different R command to acheive construction of the same subsetted dataset. Confirm that your alternate command has the same number of rows and columns.

```
nrow(nhanes20pl)
```

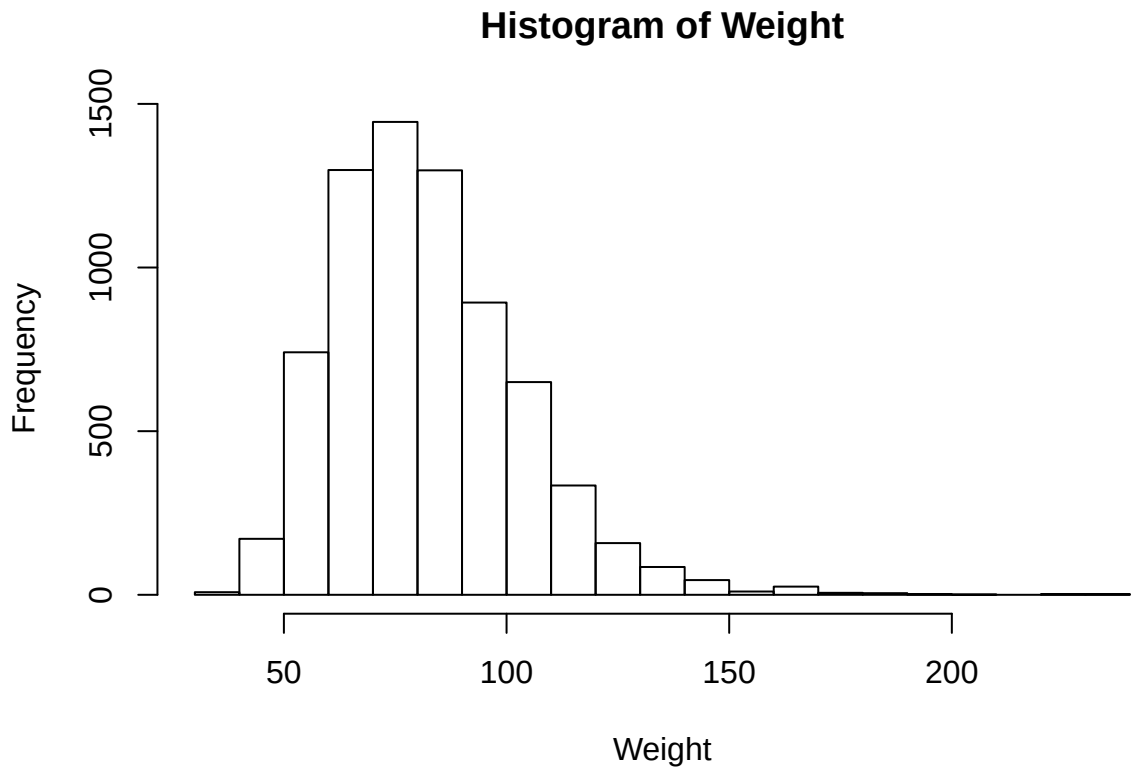
```
## [1] 7235
```

b. *Types of variables.* Using the subsetted `nhanes20pl` dataset, identify at least one variable of each of the types of variables: continuous, discrete numeric, binary, categorical, and ordered categorical. For an identified variable of each type, create an appropriate summary of the frequency distribution and calculate an appropriate measure of central tendency and variation.

```
str(nhanes20pl)
```

- Continuous example: Weight

```
hist(nhanes20pl$Weight,xlab="Weight",main="Histogram of Weight")
```




```
mean(nhanes20pl$Weight,na.rm = TRUE)
```

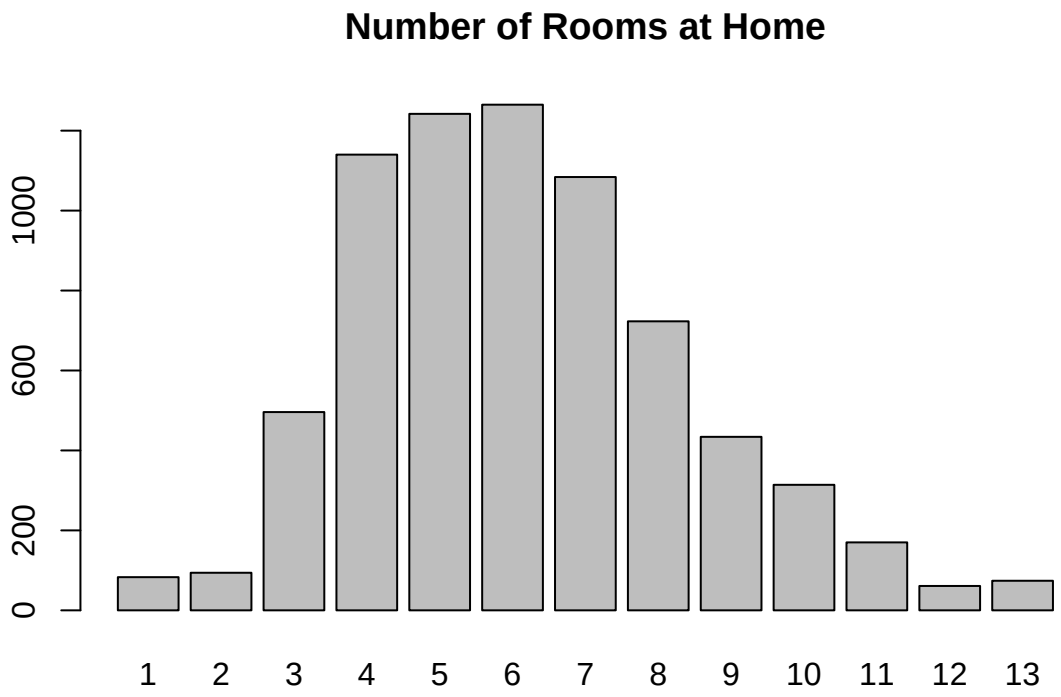
```
## [1] 82.21705
```

```
sd(nhanes20pl$Weight,na.rm=TRUE)
```

```
## [1] 21.22707
```

- Discrete numeric example: HomeRooms

```
room_count <- table(nhanes20pl$HomeRooms)
barplot(room_count,main="Number of Rooms at Home")
```




```
mean(nhanes20pl$HomeRooms,na.rm = TRUE)
```

```
## [1] 6.142201
```

```
sd(nhanes20pl$HomeRooms,na.rm=TRUE)
```

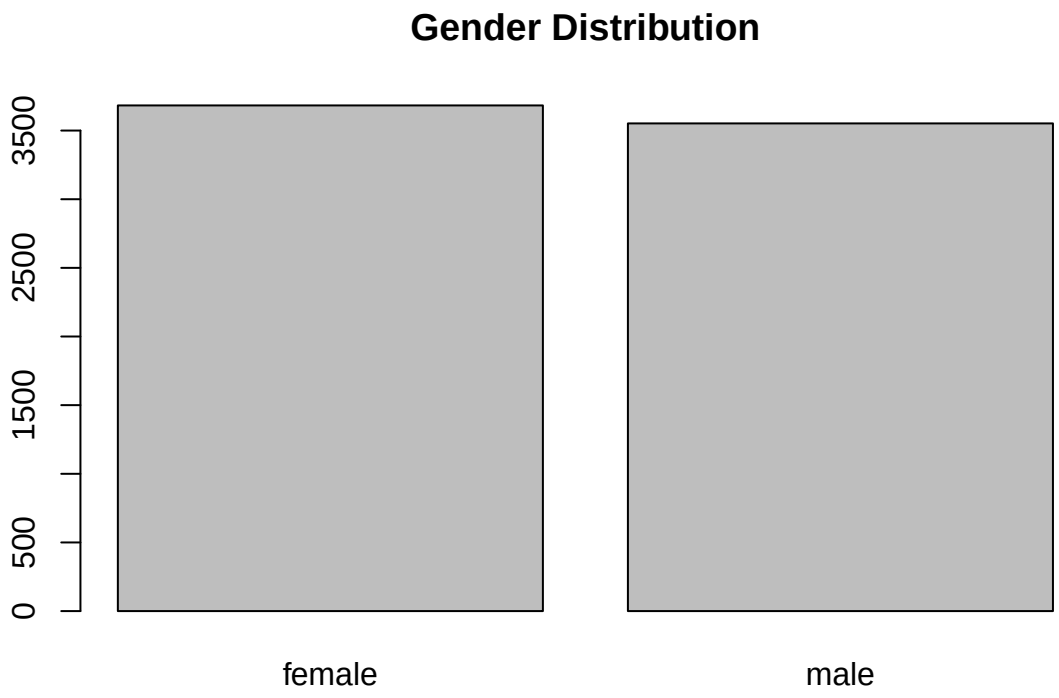
```
## [1] 2.260319
```

- Binary example: Gender

```
gender_count <- table(nhanes20pl$Gender)
gender_count
```

```
##
## female   male
##   3683   3552
```

```
barplot(gender_count,main="Gender Distribution")
```

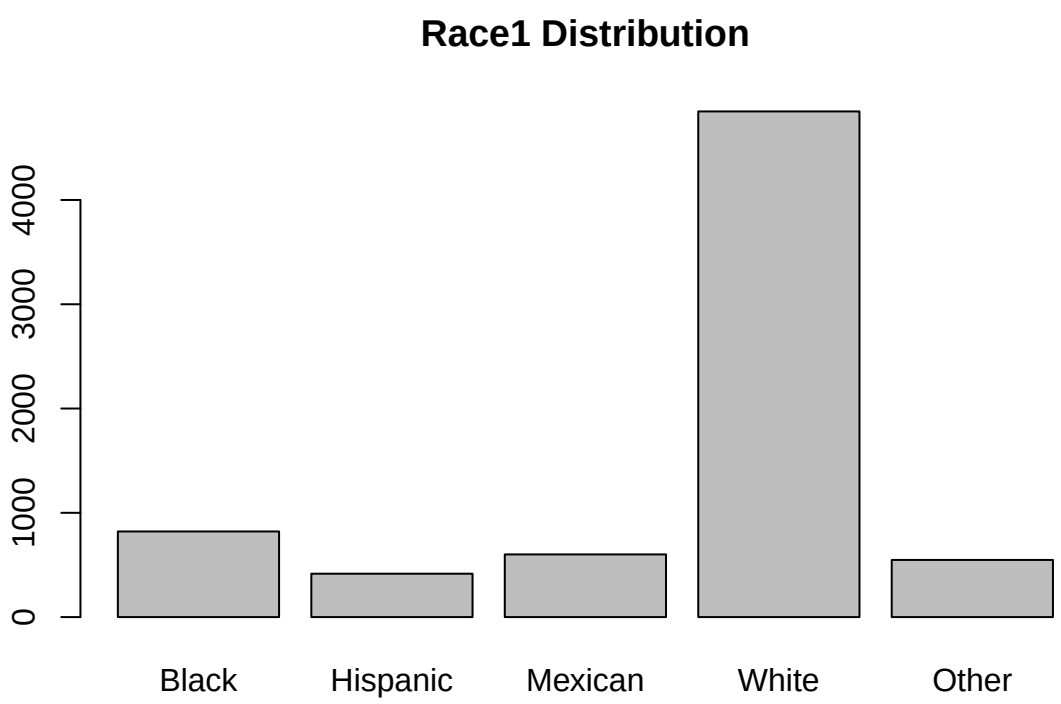



- Categorical example: Race1

```
race_count <- table(nhanes20pl$Race1)
race_count
```

```
##
##    Black Hispanic  Mexican    White    Other
##      821      416      601     4849      548
```

```
barplot(race_count,main="Race1 Distribution")
```

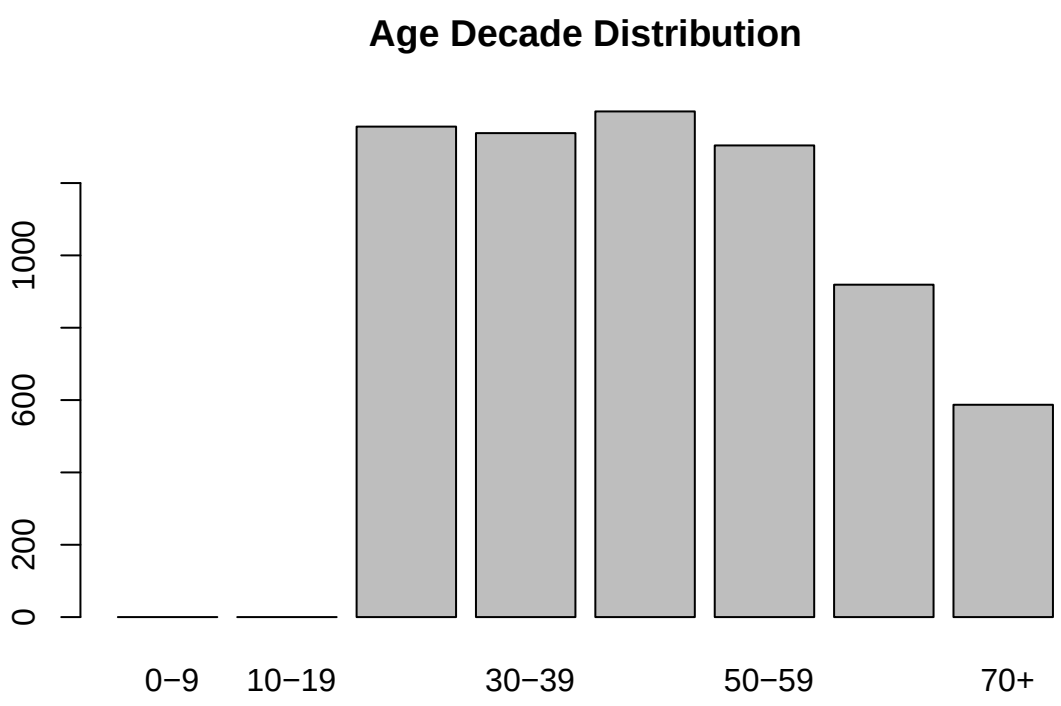



- Ordered categorical example: AgeDecade

```
agedec_count <- table(nhanes20pl$AgeDecade)
agedec_count
```

```
##
##    0-9  10-19  20-29  30-39  40-49  50-59  60-69    70+
##      0      0   1356   1338   1398   1304    919    587
```

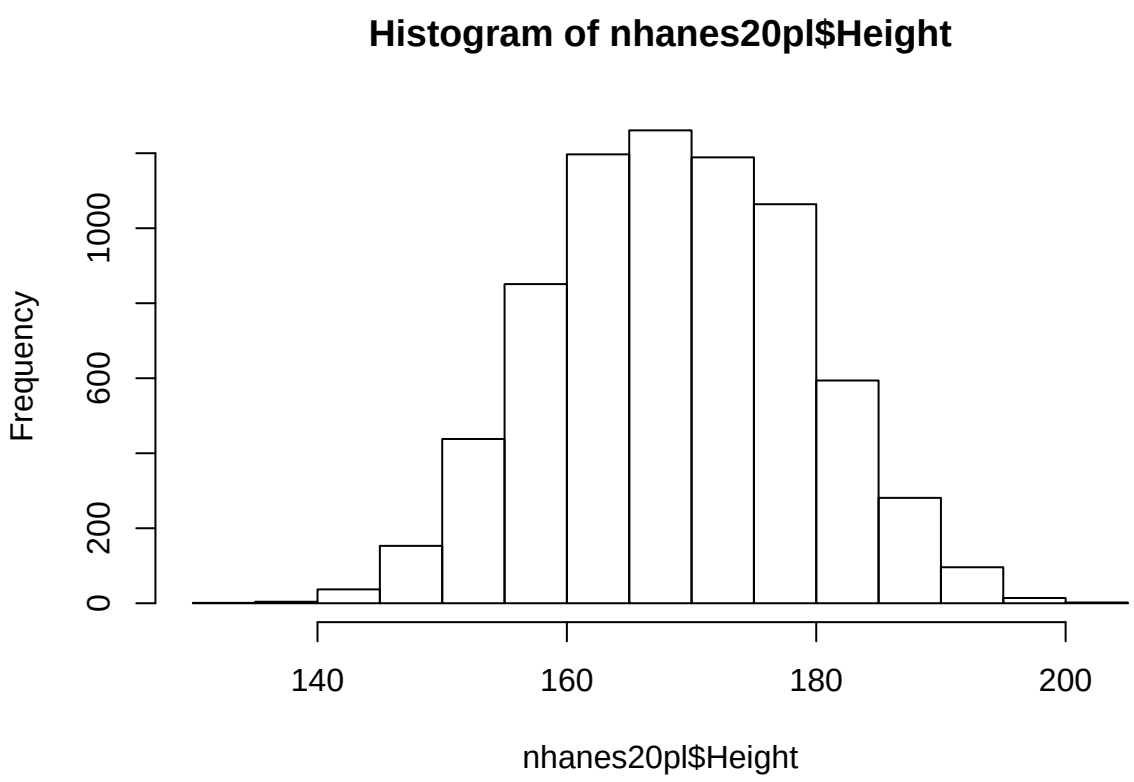
```
barplot(agedec_count,main="Age Decade Distribution")
```




c. *Frequency distributions.* Identify at least one variable in the dataset that has a frequency distribution matching each of the shapes described in Kirkman and Sterne Figures 3.5 and 3.6 (pages 20-21). For each of the identified variables, calculated the mean, median, mode, variance, standard devation, range, and interquartile range.

- Symmetric: Height

```
hist(nhanes20pl$Height)
```

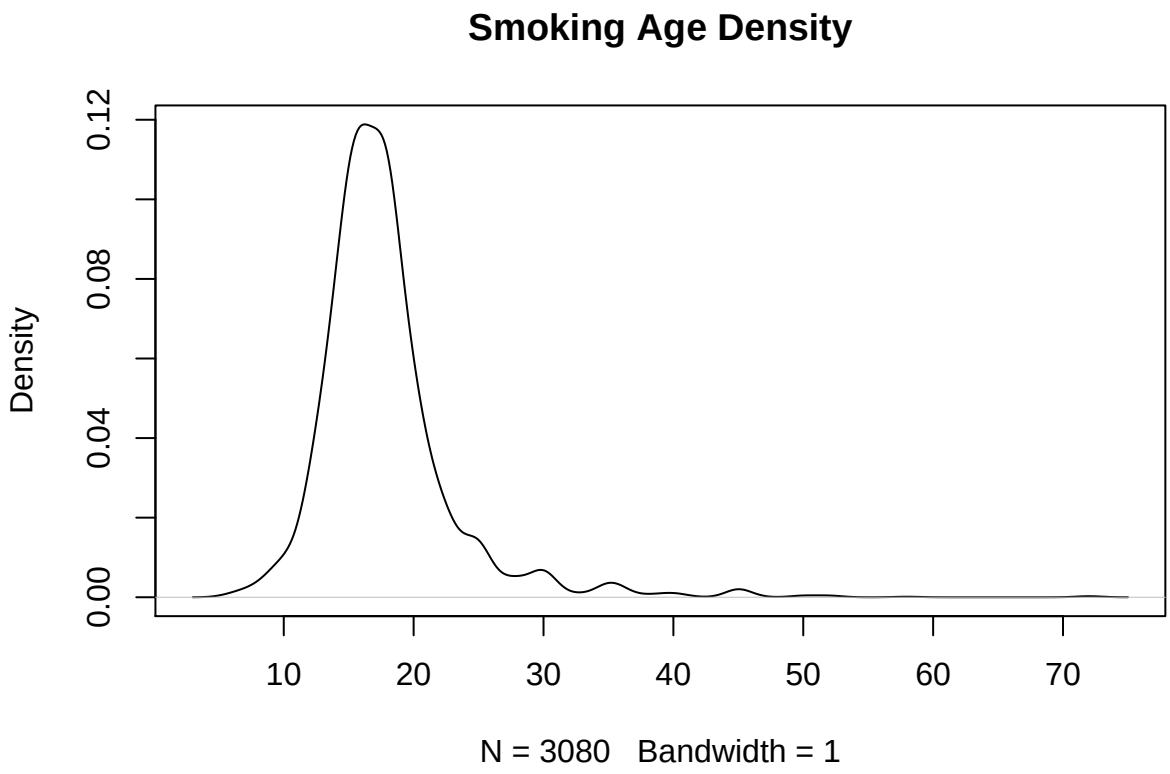



```
summary(nhanes20pl$Height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   134.5   161.4   168.7   168.8   176.0   200.4      53
```

- Positively skewed: SmokeAge

```
plot(density(nhanes20pl$SmokeAge,na.rm=TRUE,bw=1),main="Smoking Age Density")
```

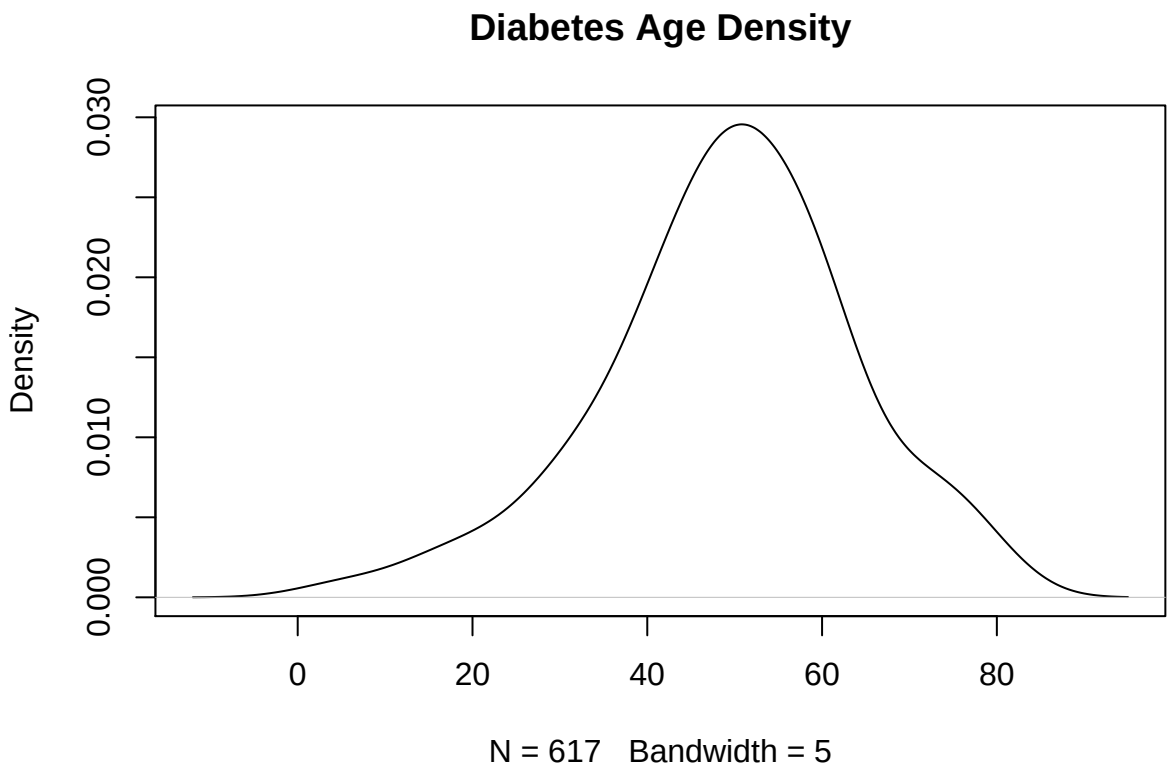



```
summary(nhanes20pl$SmokeAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    6.00   15.00   17.00   17.83   19.00   72.00    4155
```

- Negatively skewed: PhysActiveDays

```
plot(density(nhanes20pl$DiabetesAge,na.rm = TRUE,bw=5),main="Diabetes Age Density")
```

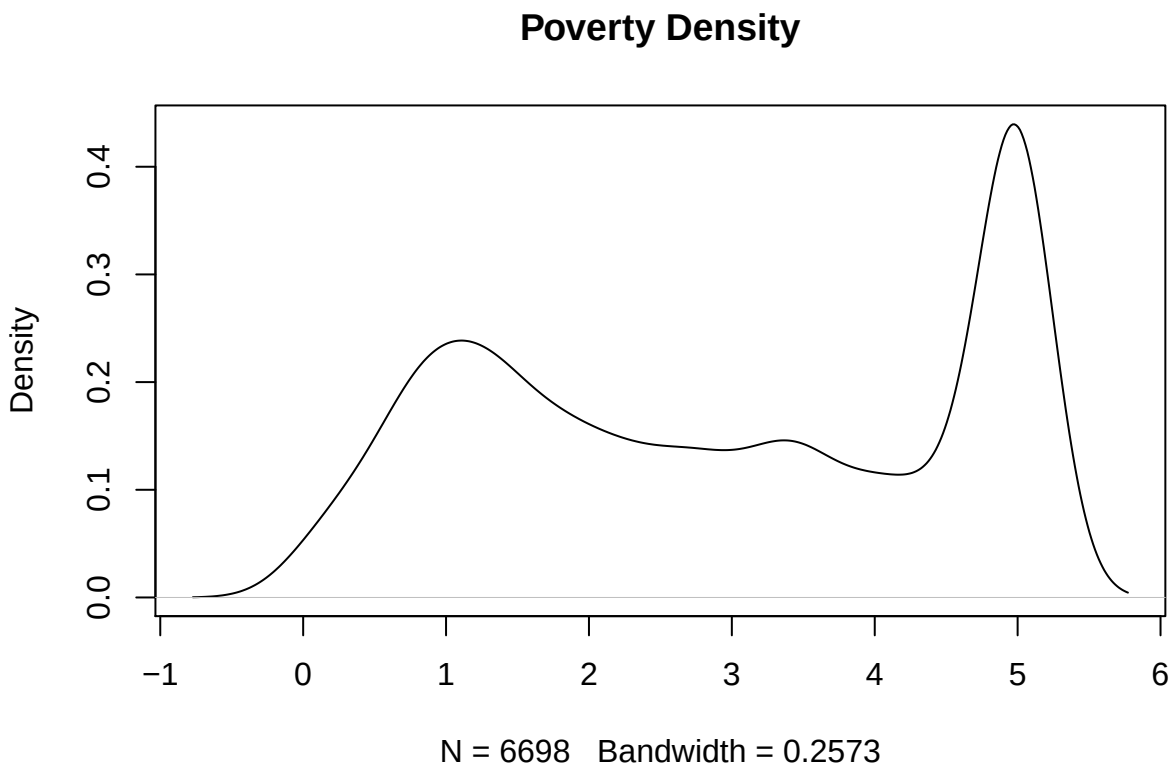



```
summary(nhanes20pl$DiabetesAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.00   41.00   50.00   49.27   59.00   80.00    6618
```

- Bimodal: Poverty

```
plot(density(nhanes20pl$Poverty,na.rm=TRUE),main="Poverty Density")
```

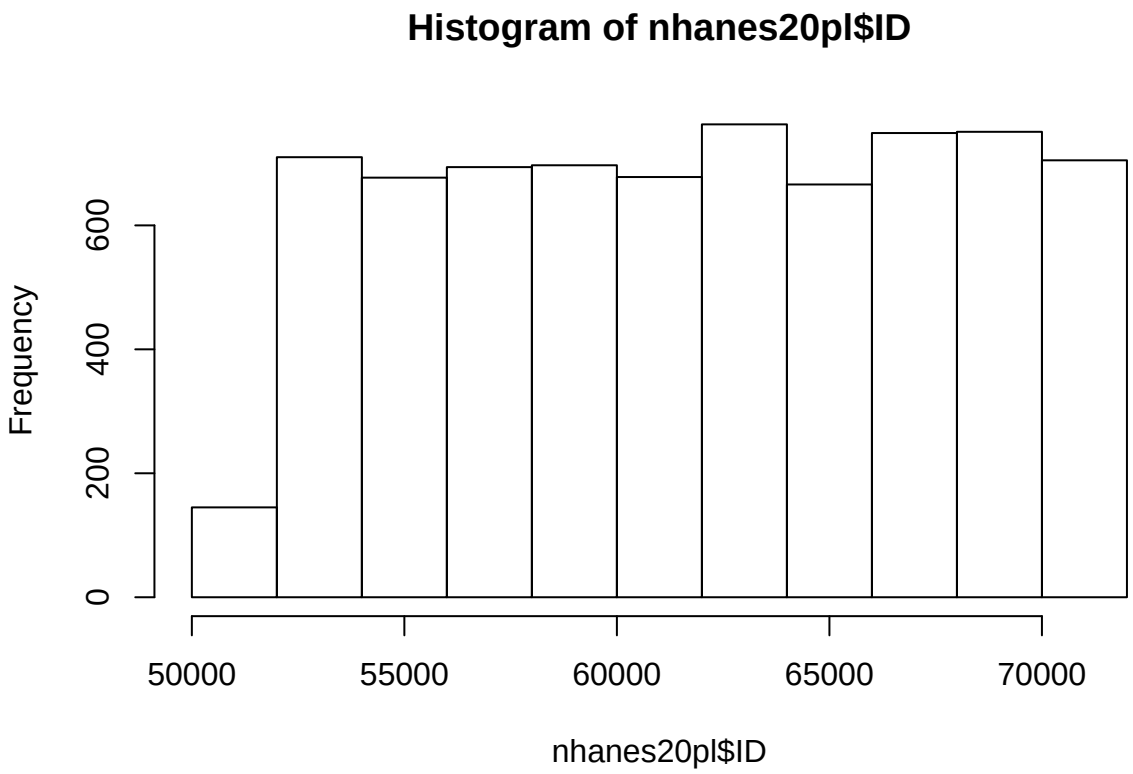



```
summary(nhanes20pl$Poverty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.360   2.910   2.945   5.000   5.000     537
```

- Reverse J-shaped: not found
- Uniform: ID

```
hist(nhanes20pl$ID)
```




```
summary(nhanes20pl$ID)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51624   56902   62056   61904   67003   71915
```

d. *Missing data.* Using the `nhanes20pl` dataset, review the amount of missing data for each of the variables pertaining to *Demographic Variables* and *Physical Measurements*.

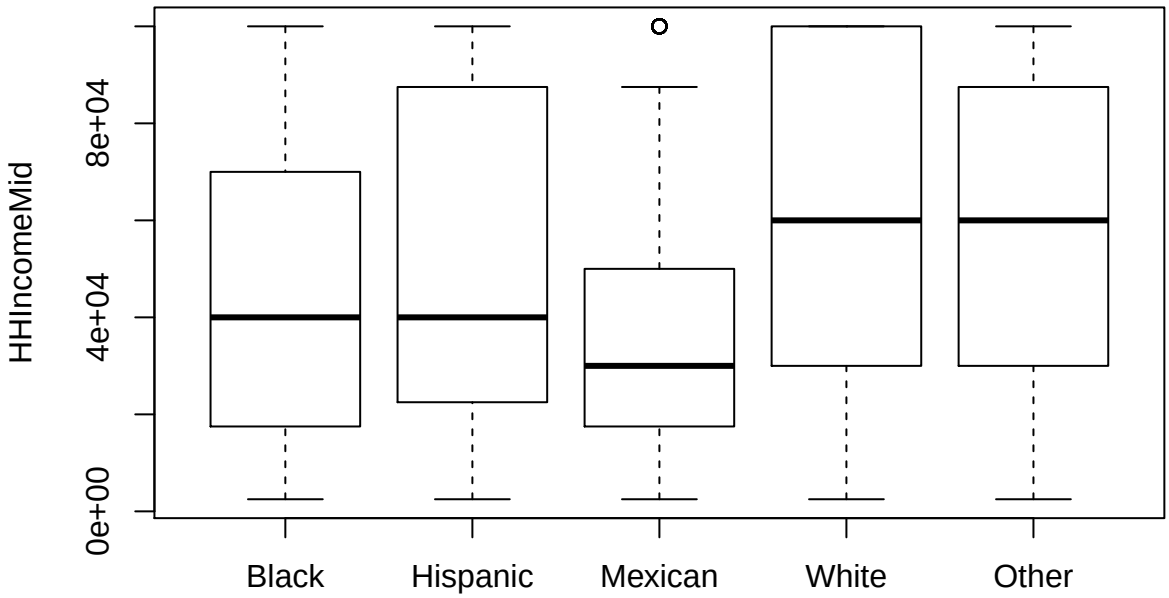
(i) Amongst each grouping (*Demographic Variables* and *Physical Measurements*) identify the one variable with the highest proportion of missing cases. (Do not consider variables that were available for only one of the survey rounds or not recorded for this age range when making your assessment.)

- Demographic Variables go from Gender to HomeOwn. Using 'summary()', HHIncomeMid have the greatest number of NAs exlcuding variables only available for one survey round.
- Physical Measurements go from Weight to Testosterone. Again using 'summary()', BPSys1 and BPDia1 have joint greatest number of NAs (consecutive BP readings).

(ii) For each of these two variables, in this sample, is there any relationship between `Gender` or `Race1` and the probability that data on the outcome is missing?

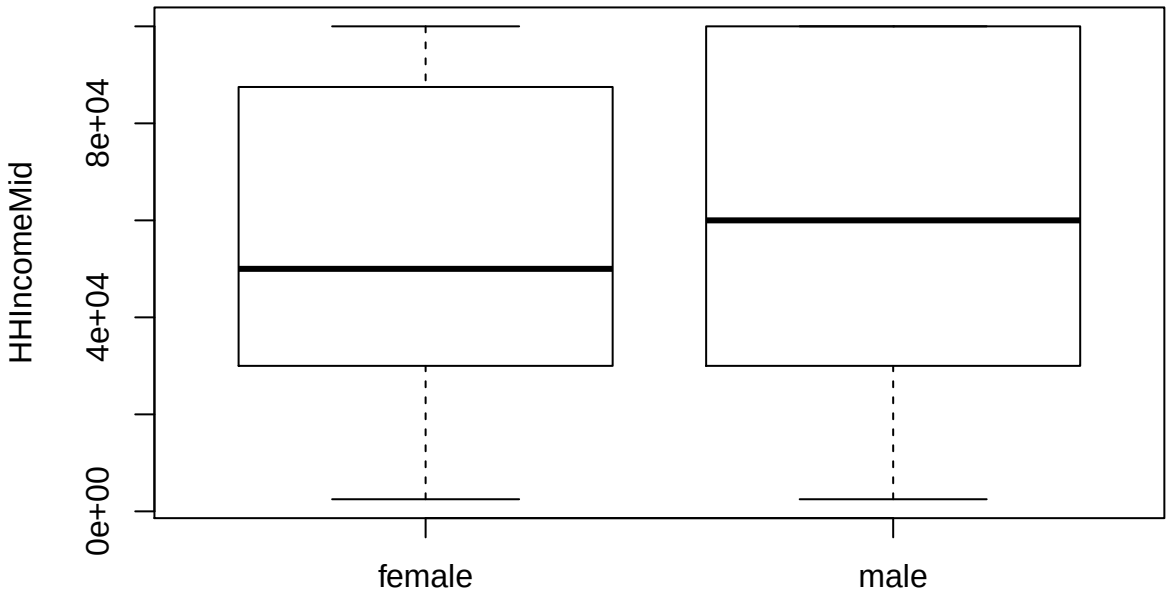
-

```
plot(nhanes20pl$Race1,nhanes20pl$HHIncomeMid,xlab="",ylab="HHIncomeMid")
```




White and other higher annual gross income than remaining 3 categories, Mexican particularly low.

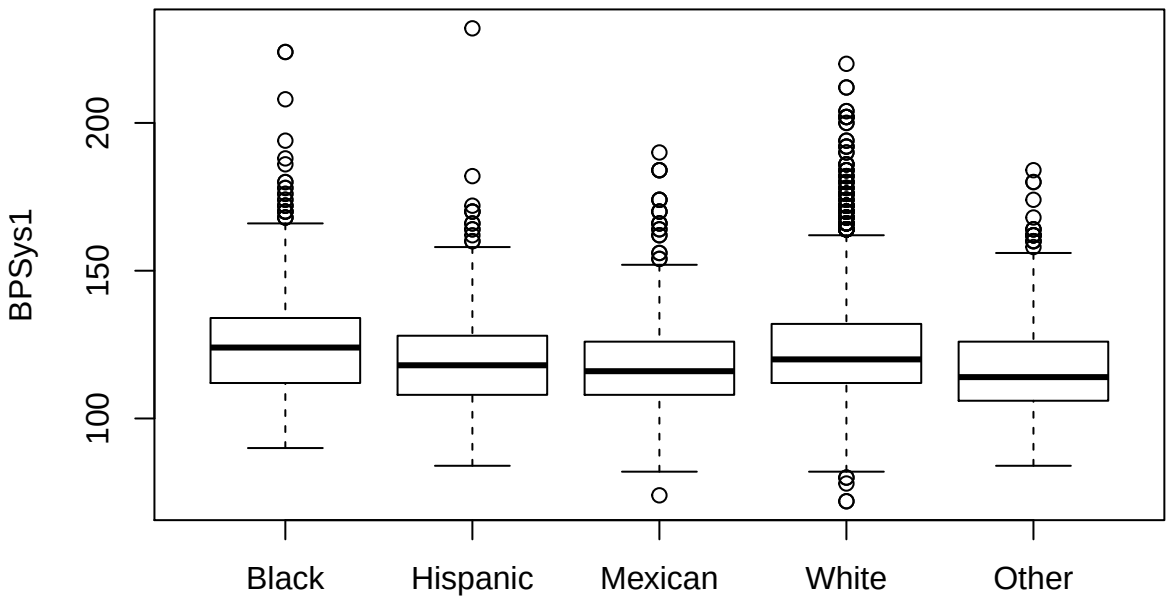
```
plot(nhanes20pl$Gender,nhanes20pl$HHIncomeMid,xlab="",ylab="HHIncomeMid")
```




Males slightly higher than females but not as pronounced as race.

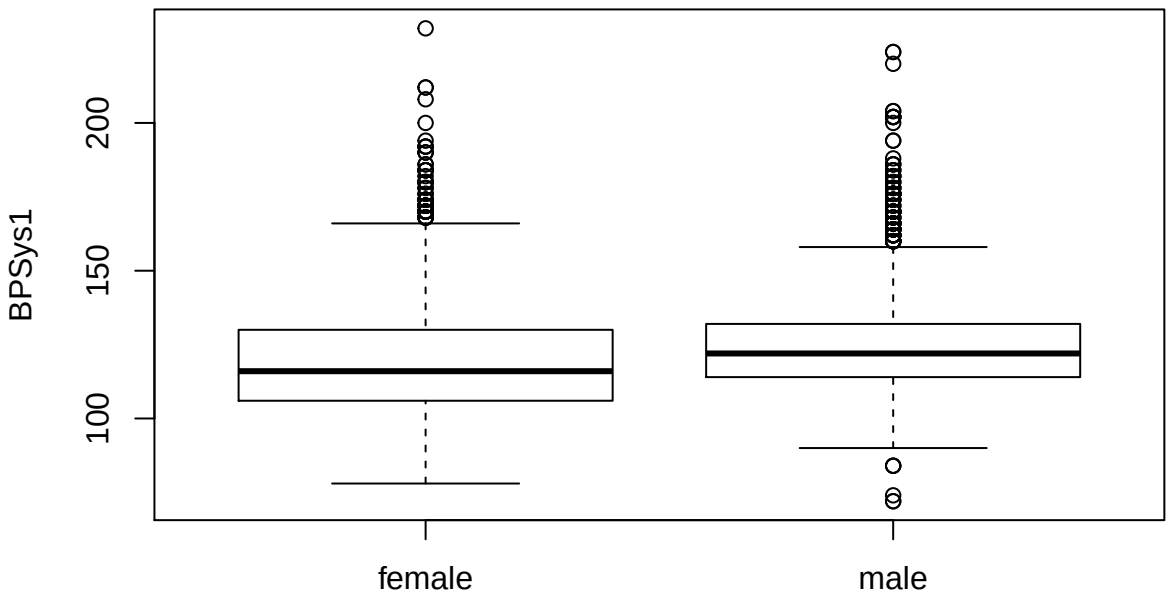
- Use BPSys1 as this is the first reading of the two.

```
plot(nhanes20pl$Race1,nhanes20pl$BPSys1,xlab="",ylab="BPSys1")
```




Fairly similar in terms of median and all have heavy tails. Black is noticeably higher than the other groups.

```
plot(nhanes20pl$Gender,nhanes20pl$BPSys1,xlab="",ylab="BPSys1")
```




Very little difference between the two.