Patel's 01 Intro To Data Handout Part 2

Harsh Patel

2024-09-06

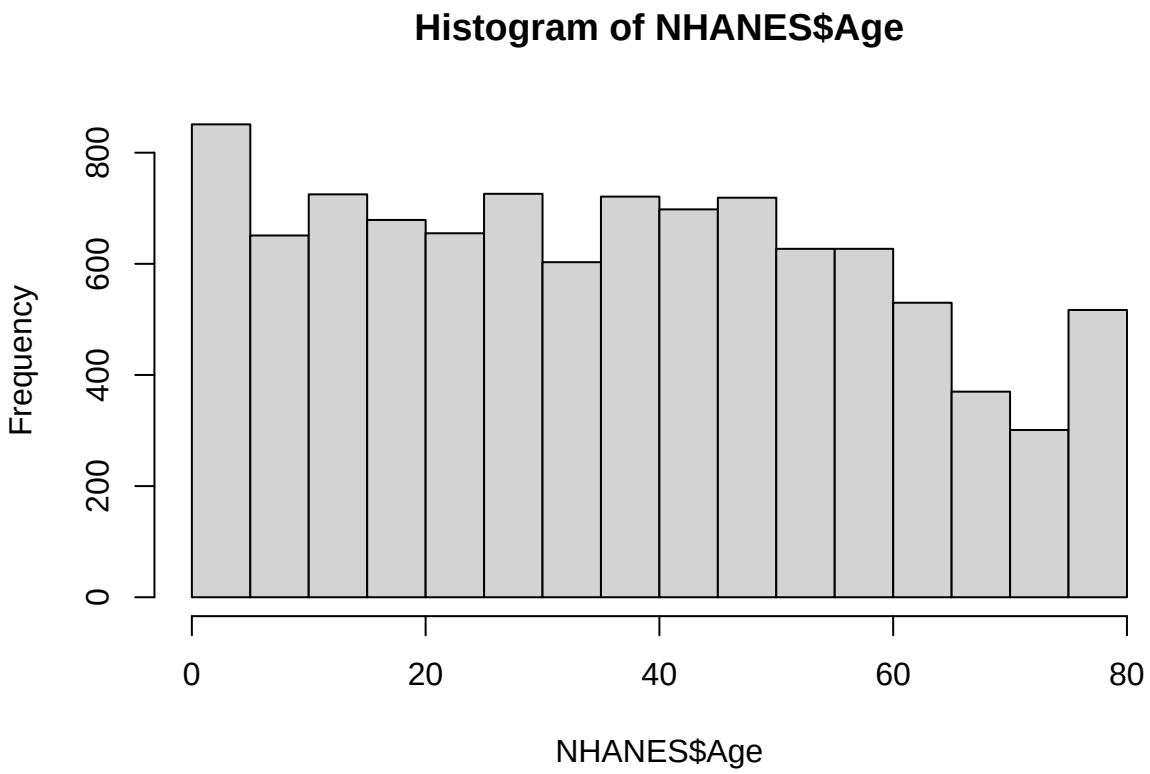
```
#Section 3: NHANES

#Question 1

#a)

#The age distribution is relatively symmetric, with 50% of respondents under 36
#And the middle 50% between 17 and 54.
#Note that ages 80 and above were recorded as 80.

#load the NHANES package and dataset
library(NHANES)
data(NHANES)

#numerical summaries
summary(NHANES$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   17.00   36.00   36.74   54.00   80.00
```

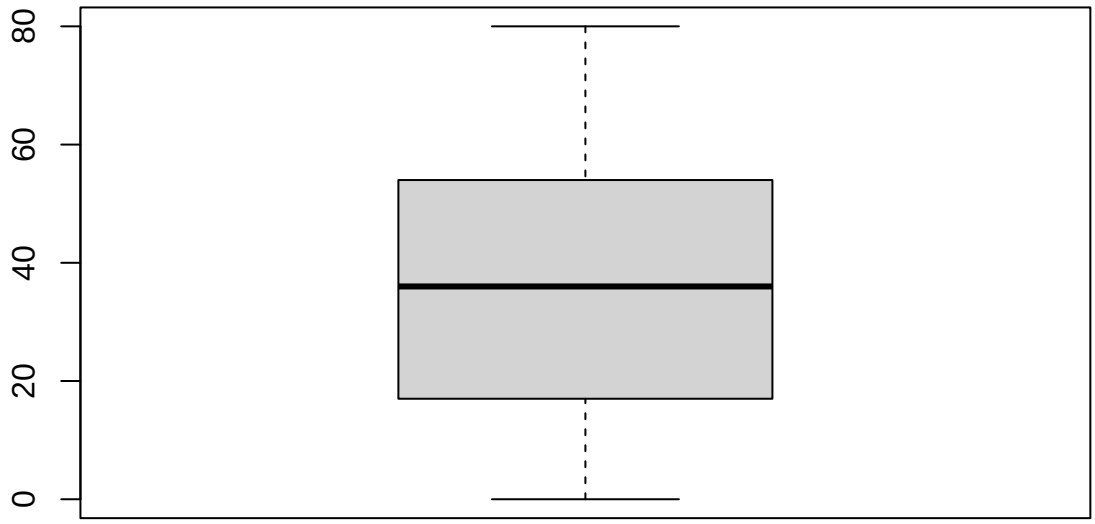
```
sd(NHANES$Age)
```

```
## [1] 22.39757
```

```
#graphical summaries
hist(NHANES$Age)
```




```
boxplot(NHANES$Age)
```



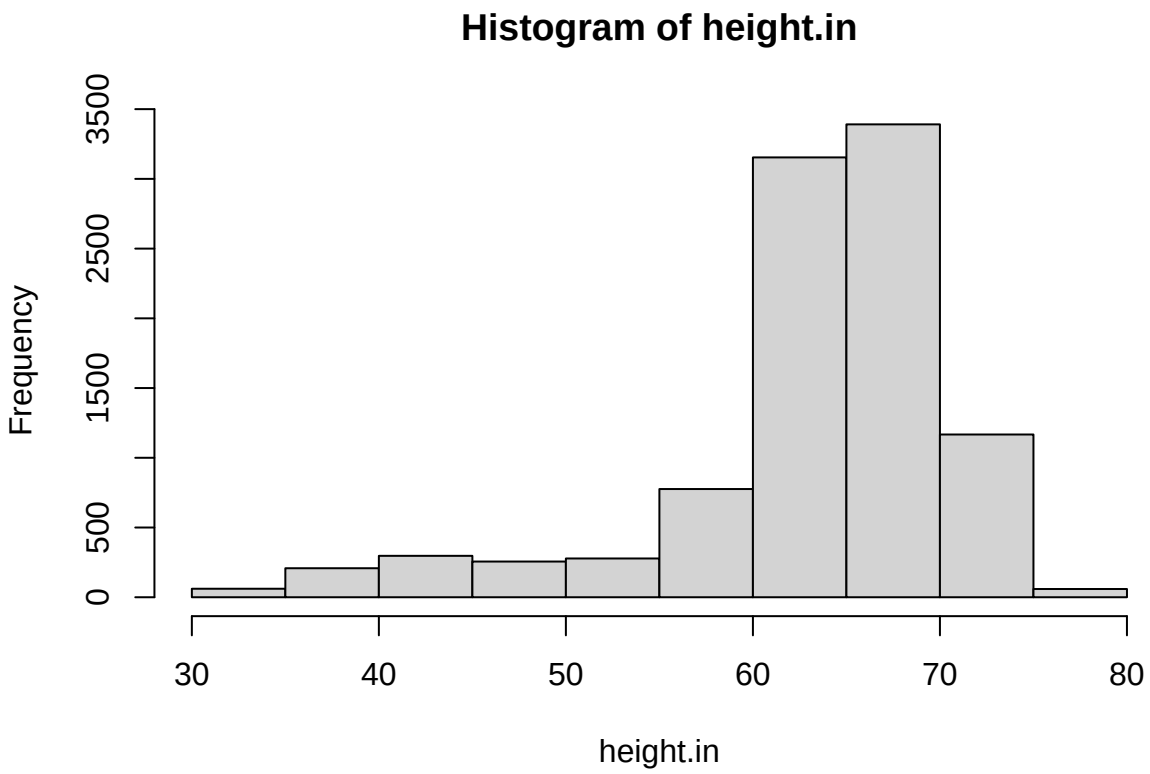

```
#b)

#The height distribution is highly left-skewed, with more individuals having taller heights.
#The median is around 65 inches (5.5 feet), and the boxplot shows this skew with dots on the lower end.

#convert to inches
height.in = 0.39*NHANES$Height

#numerical summaries
summary(height.in)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   32.60   61.15   64.74   63.13   68.06   78.16     353
```

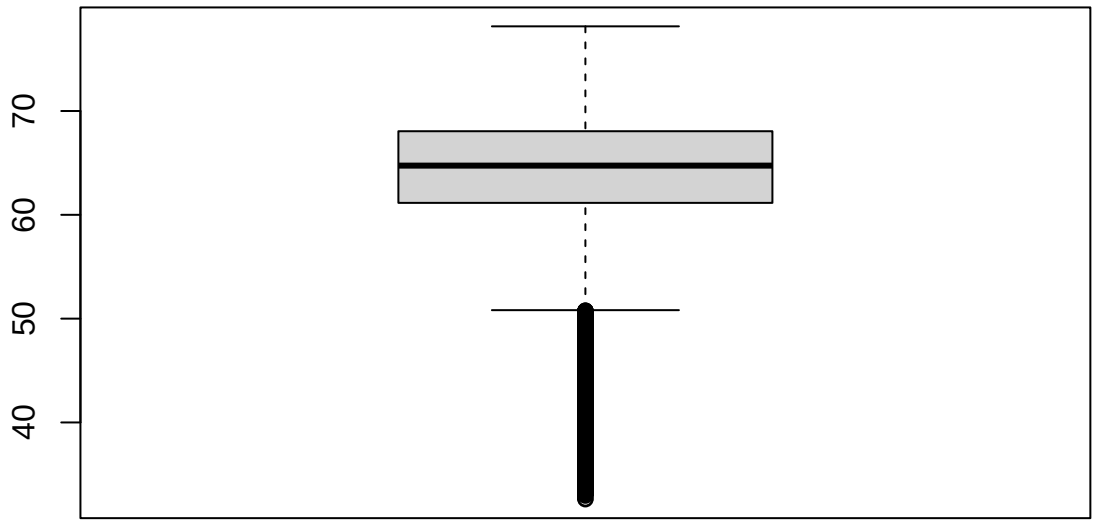
```
sd(height.in, na.rm = TRUE) #na.rm = TRUE instructs R to ignore missing values (NA's)
```

```
## [1] 7.872761
```

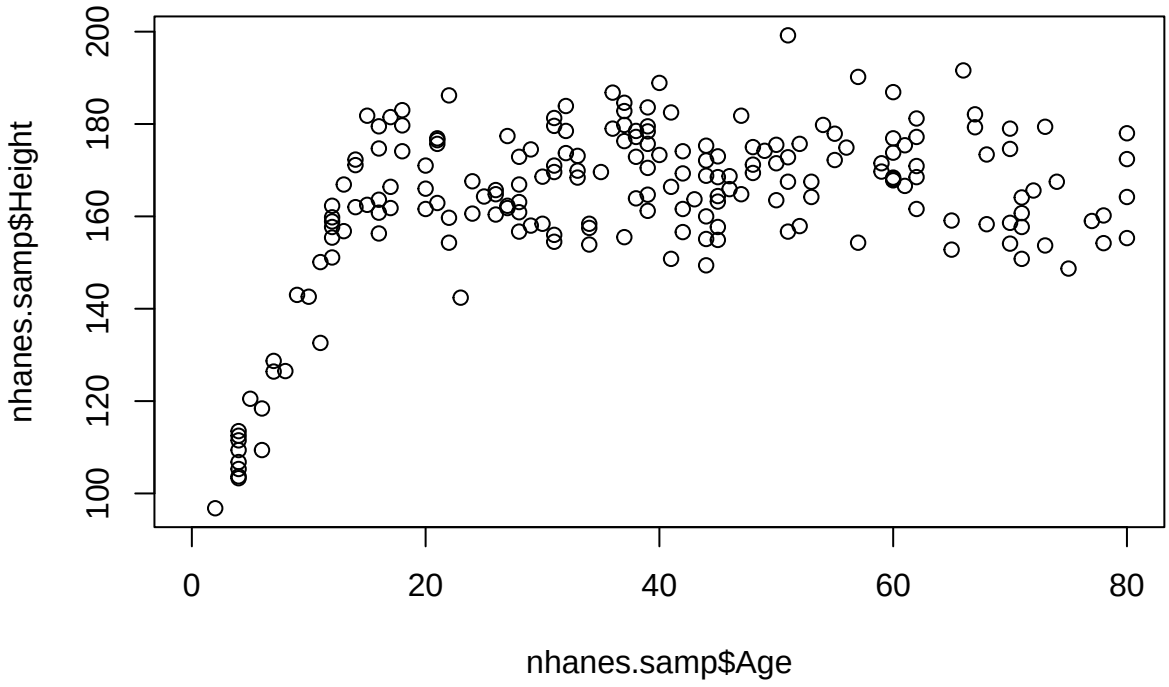
```
#graphical summaries
hist(height.in)
```



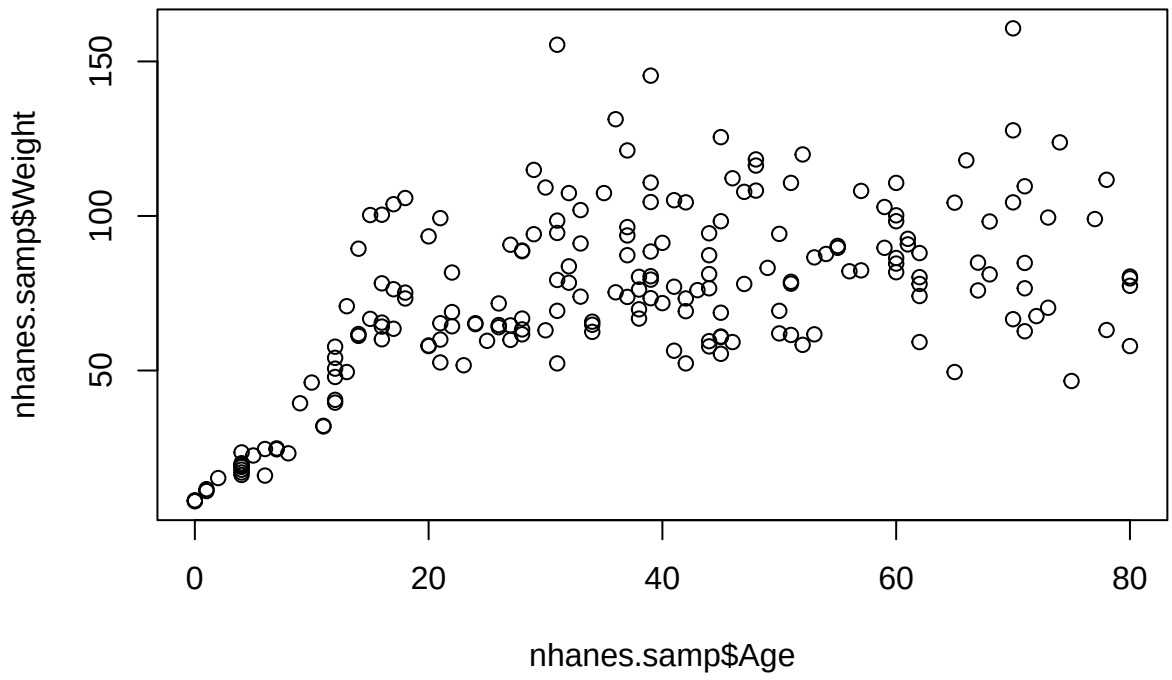

```
boxplot(height.in)
```




```r
#c)

#The scatterplot shows adults generally reach their height by age 20.
#It's not possible to determine a specific age for weight.
#As it fluctuates throughout adulthood.

#draw a random sample
set.seed(5011)
row.num = sample(1:nrow(NHANES), 200, replace = FALSE)
nhanes.samp = NHANES[row.num, ]

#investigate age and height
plot(nhanes.samp$Age, nhanes.samp$Height)
```




```
#investigate age and weight
plot(nhanes.samp$Age,nhanes.samp$Weight)
```




```
#Question 2

#a)What proportion of Americans at least 25 years of age are college graduates?

#subset the number of Americans at least 25 years of age
adults = NHANES[NHANES$Age >= 25, ]

#age and education
table(adults$Education)#summary(adults$Education) also works
```

```
##
##      8th Grade 9 - 11th Grade    High School   Some College   College Grad
##            435             814           1345           1951           2016
```

```
total.adults = length(adults$Education)

#calculations
2016/total.adults
```

```
## [1] 0.3068026
```

```
#0.307 of Americans at least 25 years of age are college graduates.

#b)What Proportion of Americans at least 25 years of age with a high school degree.
```

```r
#are high school graduates?

#calculations
(1345)/(1345 + 1951 + 2016)
```

```
## [1] 0.2532003
```

```r
#About 25.3% are Americans aged 25 and older with a high school degree who are high school graduates.

#Question 3

#a)

#numerical summary
summary(NHANES$Poverty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.240   2.700   2.802   4.710   5.000     726
```

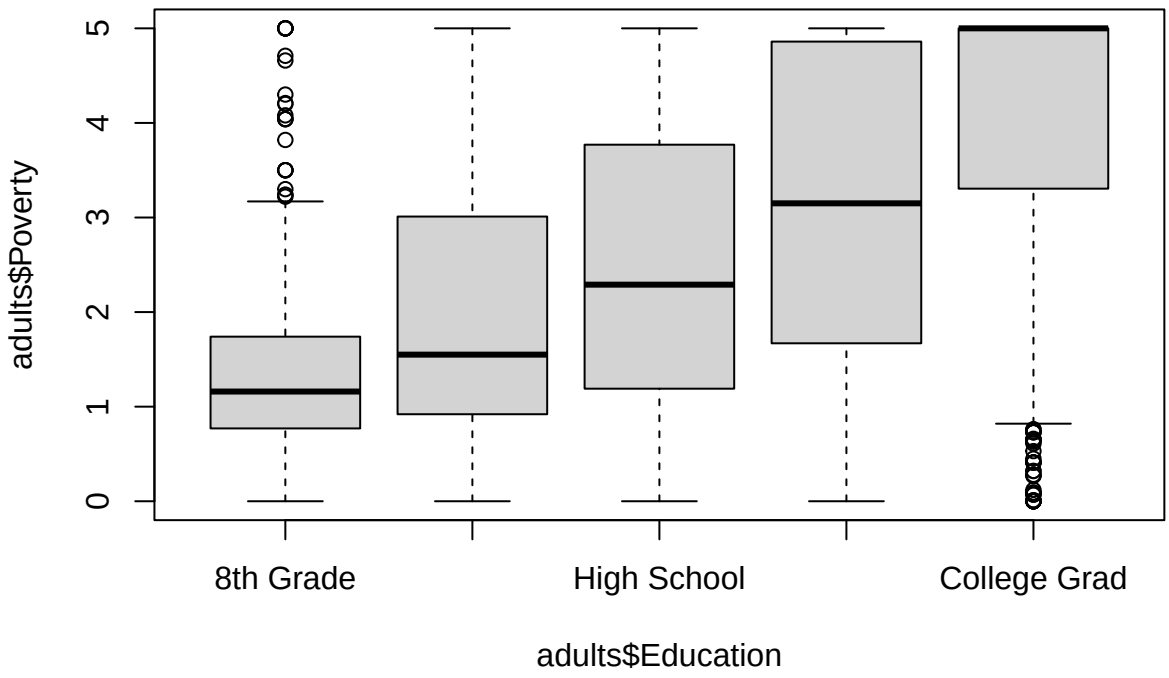
```r
#alternatively, directly use median() and IQR()
#na.rm = TRUE instructs R to disregard the missing values (NA's)
median(NHANES$Poverty, na.rm = TRUE)
```

```
## [1] 2.7
```

```r
IQR(NHANES$Poverty, na.rm = TRUE)
```

```
## [1] 3.47
```

```r
#The median is 2.7, meaning 50% of individuals have a poverty ratio above this value.
#This indicates an income level 2.7 times the poverty level.

#b)

#Median poverty rises with education level, from around 1.1.
#For those with 8th grade education to 5 for college graduates.
#The data also reveal that some 8th grade graduates are relatively wealthy.
#while some college graduates fall below the poverty level.

#graphical summary
boxplot(adults$Poverty ~ adults$Education)
```




```
#Question 4
#a) Construct a two-way table, with PhysActive as the row variable and Diabetes as the column
#variable. Among participants who are not physically active, what proportion have diabetes?
#What proportion of physically active participants have diabetes?

#Among inactive participants, 30% have diabetes, compared to 6% among those who are active.

#create table
addmargins(table(PhysActive=NHANES$PhysActive, Diabetes=NHANES$Diabetes))
```

```
##           Diabetes
## PhysActive   No  Yes  Sum
##        No  3203  472 3675
##        Yes 4361  285 4646
##        Sum 7564  757 8321
```

```
#calculations
diabetes.not.active = 472/3675
diabetes.active = 285/4646
diabetes.not.active
```

```
## [1] 0.1284354
```

```
diabetes.active
```

```
## [1] 0.06134309
```

```
#Calculate the relative risk of diabetes for inactive versus active participants.
#Does this indicate that physical activity reduces the risk of diabetes?

#From these calculations, is it possible to conclude that being physically active.
#Reduces one's chance of becoming diabetic?

#calculations
rr.diabetes = diabetes.not.active/diabetes.active
rr.diabetes
```

```
## [1] 2.093722
```

```
#With a relative risk of 2.09, inactive individuals are twice as likely to have diabetes.
#As those who are active. Though causation cannot be confirmed.
```