~~~
---
title: "Biometry activity 1"
author: "Madeleine Schoderbek"
date: "`r Sys.Date()`"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
Cpap <- read.csv(file = "data/CPAPAdherence_Data.csv")
```

```{r}
Cpap <- read.csv(file = "data/CPAPAdherence_Data.csv" )
```


1. There is a data entry error for finding the ages of the dataset. I find that there are
ages above 500 all the way to 600 in this data set, which is not possible.

```{r}
hist(Cpap$age)

weirdAges <- which(Cpap$age >= 100)
print(x = weirdAges)
Cpap$age[weirdAges]
removeAges <- which(Cpap$age > 110)
Cpap <- Cpap[-removeAges, ]
hist(Cpap$age)
```

2. When looking at the table for race_white, I find that there is a third column labeled
"q". When in the data dictionary it indicates that there should only be 2 columns: 1 =
yes, 0 = no

```{r}
table(Cpap$race_white)

removerace_white <- which(Cpap$race_white == "q")
Cpap$race_white[removerace_white] <- NA
table(Cpap$race_white)

```

3. When looking at the table for race_other, I find that there is a third column labeled
"p". When in the data dictionary it indicates that there should only be 2 columns: 1 =
yes, 0 = no
```{r}
table(Cpap$race_other)
removerace_other <- which(Cpap$race_other == "p")
Cpap$race_other[removerace_other] <- NA
table(Cpap$race_other)
```

4. When looking at the table for sex, I find that there are 3 columns: 0, 1, 2 with
corresponding numbers of sex under each column. When the data dictionary indicates that
there should only be 2 columns for sex: 1 = female and 0 = male.
```{r}
table(Cpap$sex)
removesex <- which(Cpap$sex == 2)
Cpap$sex[removesex] <- NA
~~~

~~~
table(Cpap$sex)
```

5. When looking at education, I find that there is a random value indicated in the table
showing a 1.
```{r}

table(Cpap$education)

removeeducation <- which(Cpap$education == "")
Cpap$education[removeeducation] <- NA
table(Cpap$education)
```

6. When looking at the histogram graph for ess (Measure of day time sleepiness), the
information for values 25-30 are inaccurate because it is indicated on the data dictionary
that the maximum possible number of daytime sleepiness is 24, when on the x-axis it
reaches over 24 to 30.

```{r}
hist(Cpap$ess)
weirdess <- which(Cpap$ess >= 24)
print(x = weirdess)
Cpap$ess[weirdess]
removeess <- which(Cpap$ess > 24)
Cpap <- Cpap[-removeess,]
hist(Cpap$ess)
```




7. When looking at the histogram graph for avg daily cpap, we find that there is a value
of over 20 hours of sleep which is an obvious outlier in the data.

```{r}
hist(Cpap$avg_daily_cpap)
weirdavg_daily_cpap <- which(Cpap$avg_daily_cpap >= 20)
print(x = weirdavg_daily_cpap)
Cpap$avg_daily_cpap[weirdavg_daily_cpap]
removeavg_daily_cpap <- which(Cpap$avg_daily_cpap > 20)
Cpap <- Cpap[-removeavg_daily_cpap,]
hist(Cpap$avg_daily_cpap)
```


8. When looking at subject ID for the dataset. I find that there is a duplicate subject ID
for lines 124 and 125. By finding the duplicate subject I can remove it from the dataset.

```{r}
table(Cpap$subject_id)
duplicated(x = Cpap$subject_id)
which(duplicated(x = Cpap$subject_id))
print(x = Cpap[120:130, ])
dup <- which(duplicated(Cpap$subject_id))
Cpap <- Cpap[-dup,]
print(x = Cpap$subject_id)

```


# RACE == COMBINING RACE VARIABLES

```{r}
~~~

~~~~
Cpap$race <- ifelse(Cpap$race_black == 1, "Black",
ifelse(Cpap$race_white == 1, "White",
ifelse(Cpap$race_other == 1, "Other", NA)))

print(x = Cpap$race)
table(Cpap$race)

```




#AVG_DAILY_CPAP == CREATING "ADHERENCE VARIABLE"

```{r}

Cpap$adherence <- ifelse(Cpap$avg_daily_cpap >= 4, "Adherent", "Non-Adherent")

print(x = Cpap$adherence)
```



# EXAMINING SHAPE OF AHI IN TERMS OF SHAPE AND SKEW

# Using the histogram function for Apnea Hypopnea Index (AHI), using base R I see that
this histogram is a positively right skewed distribution. The shape of the histogram is
regular and the frequency decreases as the apnea events per hour increases. I see that
this display is univariate because the x-axis represents only a single variable.
Furthermore, I can also tell that this is univariate being that the y axis shows frequency
and there are no groupings shown in the graph.

```{r}
hist(Cpap$ahi)

```

# EXAMINING DISTRIBUTIONS OF AGE VS ADHERENCE

# Using the box plot function for age vs adherence category I can see that this box plot
is bivariate because it is comparing 2 variables (adherent vs non-adherent under age).
Being that both box plots are about the same width, I can tell that there is not a wide
variety of age differences between the two. I can tell that there is a difference in the
IQR between the 2 box plots because the adherent boxplot is taller than the non-adherent
box plot. I can also see that for the non-adherent box plot, the top whisker is longer
than the top whisker of the adherent box plot, this could show that some non-adherent
individuals are older than the rest, and the age distribution may be skewed right.

```{r}
boxplot(Cpap$age)
boxplot(age ~ adherence,
data = Cpap,
xlab = "Adherence Category",
ylab = "Age",
main = "Age vs Adherence Category")
```


# EXAMINING DISTRIBUTIONS OF RACE VS ADHERENCE

# Using the bar plot function to compare distributions of race vs the adherence category,
I can see that this is a bivariate display because it shows 3 variables on the x axis
(black, other, white). The bar plot shows that there are differences in the height between
all 3 for the adherence category. Seeing this I can tell that the race distribution across
variables is not consistent across adherence variables.
~~~~

~~~
```{r}
tab1 <- table(x = Cpap$race)
print(tab1)
barplot(height = tab1,
horiz = FALSE,
ylim = c(0, 200),
xlab = "race",
ylab = "Adherence Category",
main = "Race vs Adherence Category")
```






#RELATIONSHIP BETWEEN ESS AND MMSE

#To determine the relationship between Ess(Epworth Sleepiness Scale) and MMSE (Mini-Mental
State Exam) to adherence and non-adherence I used the scatter plot function to compare the
variables. Doing so, I made the adherent plots blue and the non adherent plots red. From
looking at the scatter plot you can see that there is no correlation between the ESS and
MMSE and their adherent and non-adherent data. To confirm, you can see in the data set
that the numbers given for MMSE and ESS are mostly random and vastly different which makes
sense seeing the scatterplot I created. This display is Bivariate being that is includes 2
different variables compared to 2 different variables.



```{r}
plot(Cpap$ess[Cpap$adherence == "Adherent"], Cpap$mmse[Cpap$adherence == "Adherent"],
     col = "blue", pch = 16,
     xlab = "Epworth Sleepiness Scale (ESS)",
     ylab = "Mini-Mental State Exam (MMSE)",
     main = "ESS vs MMSE by CPAP Adherence")
points(Cpap$ess[Cpap$adherence == "Non-Adherent"], Cpap$mmse[Cpap$adherence == "Non-
Adherent"],
       col = "red", pch = 16)

legend("topright", legend = c("Adherent", "Non-Adherent"),
       col = c("blue", "red"), pch = 16)
```


#CREATING TABLE 1

# When finding table 1 for this data set, I can see that there are 127 adherent
invididuals and 45 non adherent individuals to CPAP. For sex, you can see that for
adherent and non adherent the percentages of male and female are similar. For race, you
can see that there is a large percentage (80%) of the adherent individuals that are white
vs 50% white for the non-adherent group. For average daily Cpap use, there is
significantly more use in the adherent group with a mean of 6.40 hours vs a mean of 1.60
hours for the non-adherent group which makes sense. The mean of ess, mme, and ahi show
little differences between adherent vs non-adherent groups. The mean for age was about the
same for adherent vs. non-adherent (about 67). The education variable between adherent vs
non-adherent showed that for both categories, about 73-80% of individuals had higher than
a high school degree.
```{r}

library(tableone)

vars <- c("age", "sex", "race", "avg_daily_cpap", "ess", "mmse", "ahi", "education",
"ethnicity")
~~~

~~~~
strata <- "adherence"

table1 <- CreateTableOne(vars = vars, strata = strata, data = Cpap, factorVars = c("sex",
"race"))

print(x = table1, showAllLevels = TRUE)
```
~~~~