Case Study 1

Instructions:

Get into groups of 3 and answer the following questions. Submit your knitted report as a pdf/html file and include the original .Rmd file also. Submit one report per group with all members' names on it before the end of class. **For each question, write the name of the team member who answered/coded it.**

1a. Read in the data contained in the file "casestudy1_data.csv". This data contains your class's and others' answers to those anonymous survey questions. Print the first 6 rows of the data. What are the column names?

```
data <- read.csv("~/Desktop/repos/master-intro-ds/Data/casestudy1_data.csv")

head(data)
```

```
##   X     class tattoos sleep                                    smoke      music
## 1 1 Sophomore       1   6.5                                      No.       Rock
## 2 2    Senior       4   7.0                         Yes\\, cannabis.    Hip hop
## 3 3 Sophomore       0   7.5                           Yes, cannabis.    Hip hop
## 4 4 Sophomore       0   8.0                           Yes, cannabis.       Rock
## 5 5    Senior       0   4.0 Yes\\, vapes.,Yes\\, tobacco/cigarettes. Electronic
## 6 6 Sophomore       0   6.5                                      No.    Country
##   shoesize piercing    residentialpref number cannabis  vape tobacco heightcm
## 1      6.0        8 Rural (Small town)      4    FALSE FALSE   FALSE   157.48
## 2      9.6        0           Suburban      7     TRUE FALSE   FALSE   177.00
## 3      7.0        0           Suburban      2     TRUE FALSE   FALSE   155.00
## 4      7.5        0           Suburban      8     TRUE FALSE   FALSE   167.64
## 5      8.0        2         Urban/City      8    FALSE  TRUE    TRUE   165.00
## 6      7.5        5 Rural (Small town)      7    FALSE FALSE   FALSE   170.18
##   heightin
## 1 62.00000
## 2 70.00000
## 3 61.02362
## 4 66.00000
## 5 65.00000
## 6 67.00000
```

The column names are:

```
colnames(data)
```

```
##  [1] "X"               "class"           "tattoos"         "sleep"
##  [5] "smoke"           "music"           "shoesize"        "piercing"
##  [9] "residentialpref" "number"          "cannabis"        "vape"
## [13] "tobacco"         "heightcm"        "heightin"
```

1b. Look at the columns entitled "cannabis", "vape", and "tobacco". What data type do these columns contain? How do their values relate to the "smoke" column? Give an example reason why we may prefer to have this smoking information stored in these 3 separate columns rather than the single "smoke" column.

```
str(data)
```

```
## 'data.frame':    180 obs. of  15 variables:
##  $ X              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ class          : chr  "Sophomore" "Senior" "Sophomore" "Sophomore" ...
##  $ tattoos        : int  1 4 0 0 0 0 1 2 0 0 ...
##  $ sleep          : num  6.5 7 7.5 8 4 6.5 7 7 8 6.5 ...
##  $ smoke          : chr  "No." "Yes\\, cannabis." "Yes, cannabis." "Yes, cannabis." ...
##  $ music          : chr  "Rock" "Hip hop" "Hip hop" "Rock" ...
##  $ shoesize       : num  6 9.6 7 7.5 8 7.5 8.5 8.5 8 6.5 ...
##  $ piercing       : int  8 0 0 0 2 5 4 2 0 2 ...
##  $ residentialpref: chr  "Rural (Small town)" "Suburban" "Suburban" "Suburban" ...
##  $ number         : num  4 7 2 8 8 7 8 9 5 7.3 ...
##  $ cannabis       : logi  FALSE TRUE TRUE TRUE FALSE FALSE ...
##  $ vape           : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ tobacco        : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ heightcm       : num  157 177 155 168 165 ...
##  $ heightin       : num  62 70 61 66 65 ...
```

The data types for those 3 columns are logical. They contain `TRUE` if the `smoke` column indicates that they smoke that substance. It's easier to work with each type of substance separately and store it as a boolean if, for example, we want to calculate a percentage of students who vape.

2a. What are the unique values of the class variable? Why are there 2 Super Senior categories? Combine them into a single, consistent "Super Senior+" category.

```
unique(data$class)
```

```
## [1] "Sophomore"      "Senior"         "Freshman"       "Junior"
## [5] "Super Senior +" "Super Senior+"
```

There are 2 super senior categories because one has a space and one does not. We can combine them into a single category with the following line:

```
data[data$class == "Super Senior +",]$class = "Super Senior+"
```

2b. Convert the dataframe into a tibble. Then subset the data into 5 distinct tibbles which separate the data by classes: one with Freshman only, one with Sophomores only, etc. Which class is most represented in our original data?

```
data = as_tibble(data)

freshmen = data[data$class == "Freshman", ]
sophomores = data[data$class == "Sophomore", ]
juniors = data[data$class == "Junior", ]
seniors = data[data$class == "Senior", ]
superseniors = data[data$class == "Super Senior+", ]

nrow(freshmen)
```

```
## [1] 24
```

```
nrow(sophomores)
```

```
## [1] 94
```

```
nrow(juniors)
```

```
## [1] 32
```

```
nrow(seniors)
```

```
## [1] 25
```

```
nrow(superseniors)
```

```
## [1] 5
```

Sophomores are most represented (94 students).

3a. Which class of students gets the least sleep? Hypothesize a reason for this discrepancy.

```
mean(freshmen$sleep)
```

```
## [1] 7.054167
```

```
mean(sophomores$sleep)
```

```
## [1] 6.882483
```

```
mean(juniors$sleep)
```

```
## [1] 7.132812
```

```
mean(seniors$sleep)
```

```
## [1] 7.02
```

```
mean(superseniors$sleep)
```

```
## [1] 6.1
```

On average, Super Seniors sleep the least. This may because they are taking many classes to try to graduate.

3b. Suppose we classified students who average less than 6 hours a sleep a night as "insomniac" students and students who sleep more than 10 hours a night "sleepy" students. Create 2 tibbles: one containing all insomniac students and one containing all sleepy students. Then combine these 2 tibbles into a single tibble called "extreme".

```
insomniac = data[data$sleep < 6,]
sleepy = data[data$sleep > 10,]

extreme = rbind(insomniac, sleepy)
extreme
```

```
## # A tibble: 14 x 15
##         X class       tattoos sleep smoke music shoesize piercing residentialpref
##     <int> <chr>         <int> <dbl> <chr> <chr>    <dbl>    <int> <chr>
##  1      5 Senior            0  4    "Yes~ Elec~      8          2 Urban/City
##  2     51 Sophomore         0  5.5  "No." Pop        7.5        2 Urban
##  3     62 Freshman          0  5    "No." Rock       7          2 Urban
##  4     87 Sophomore         0  5.5  "No." Pop        8.5        0 Suburban
##  5     97 Sophomore         0  5.5  "No." Clas~      7.5        6 Suburban
##  6    122 Sophomore         0  5.82 "No." Rock       8.5        2 Suburban
##  7    124 Sophomore         0  4    "No." Pop       11          0 Suburban
##  8    129 Senior            0  5    "No." Pop        9.5        3 Urban/City
##  9    157 Super Seni~       0  5    "No." Rock      14          0 Rural (Small t~
## 10    164 Junior            2  5    "No." Pop        6         11 Urban
## 11    165 Sophomore         0  5.5  "Yes~ Indie      7          5 Rural (Small t~
## 12   166 Freshman          0  4    "No." Pop        8          2 Suburban
## 13   171 Freshman          0  5    "No." Clas~      8.5        0 Rural (Small t~
## 14    77 Junior            0 11    "No." Elec~      6.5        1 Rural (Small t~
## # i 6 more variables: number <dbl>, cannabis <lgl>, vape <lgl>, tobacco <lgl>,
## #   heightcm <dbl>, heightin <dbl>
```

4. Suppose I wanted to investigate smoking tendencies as it relates to musical preference. (In other words, I would like to know what fraction of pop/rock/hip hop/etc. music lovers smoke vapes? Smoke cannabis? Smoke tobacco? Repeated for all music categories.)

a. Which type of music lover do you think is most likely to smoke cannabis? Which genre do you think is least likely to smoke vapes? Give a potential "reason" to explain your intuition for each question.

Answers will vary. Here are some example responses:

I think electronic music lovers are most likely to smoke cannabis because that music is the most "vibey" in my opinion.

I think classical music lovers are least likely to smoke vapes because to me they seem serious and unlikely to take chances with their health.

b. Create a dataframe/tibble which for each preference of music genre, reports 1. the number of students who prefer it, 2. the fraction of those students who vape, 3. the fraction of those students who smoke cannabis, and 4. the fraction of those students who smoke tobacco.

Using the help function to understand `group_by` and `summarise`, I was able to find an example and modify it to create:

```r
data %>%
  group_by(music) %>%
  summarise(count = n(),
            vape = mean(vape),
            cannabis = mean(cannabis),
            tobacco = mean(tobacco)
            )
```

```
## # A tibble: 8 x 5
##   music      count   vape cannabis tobacco
##   <chr>      <int>  <dbl>    <dbl>   <dbl>
## 1 Classical      7 0          0      0
## 2 Country       13 0          0.154  0
## 3 Electronic    12 0.0833     0.417  0.0833
## 4 Hip hop       38 0.0789     0.368  0.0789
## 5 Indie         18 0.0556     0.389  0.111
## 6 Other         25 0.12       0.28   0.08
## 7 Pop           47 0          0      0
## 8 Rock          20 0.05       0.25   0
```

```
# A tibble: 8 × 5
  music      count   vape   cannabis tobacco
  <chr>      <int>  <dbl>  <dbl>    <dbl>
1 Classical      7    0      0          0
2 Country       13    0      0.154      0
... etc.
#This means that ~15.4% of the 13 country music lovers smoke cannabis.
```

(Hint: there are many ways to create such a table. You can do it in any way you choose. If you want to use simple tidyverse functions to do it, look at the help menu/examples for the functions `group_by` and `summarise`. You can also do this question without the help of these functions.)

c. For each of the two questions in part a. above, check whether your intuition was right or wrong using the data. If you were wrong, say what the correct answer was.

I was right about Electronic music lovers: they smoke weed at the highest rate (~42%). Many genres vape the least: Classical, Country, and Pop music lovers do not vape at all in our dataset.

5a. In the column "number", I asked each of you to provide a random number between 0-10. First, round the numbers down to the nearest whole number. (Hint: the floor() function can help with this.)

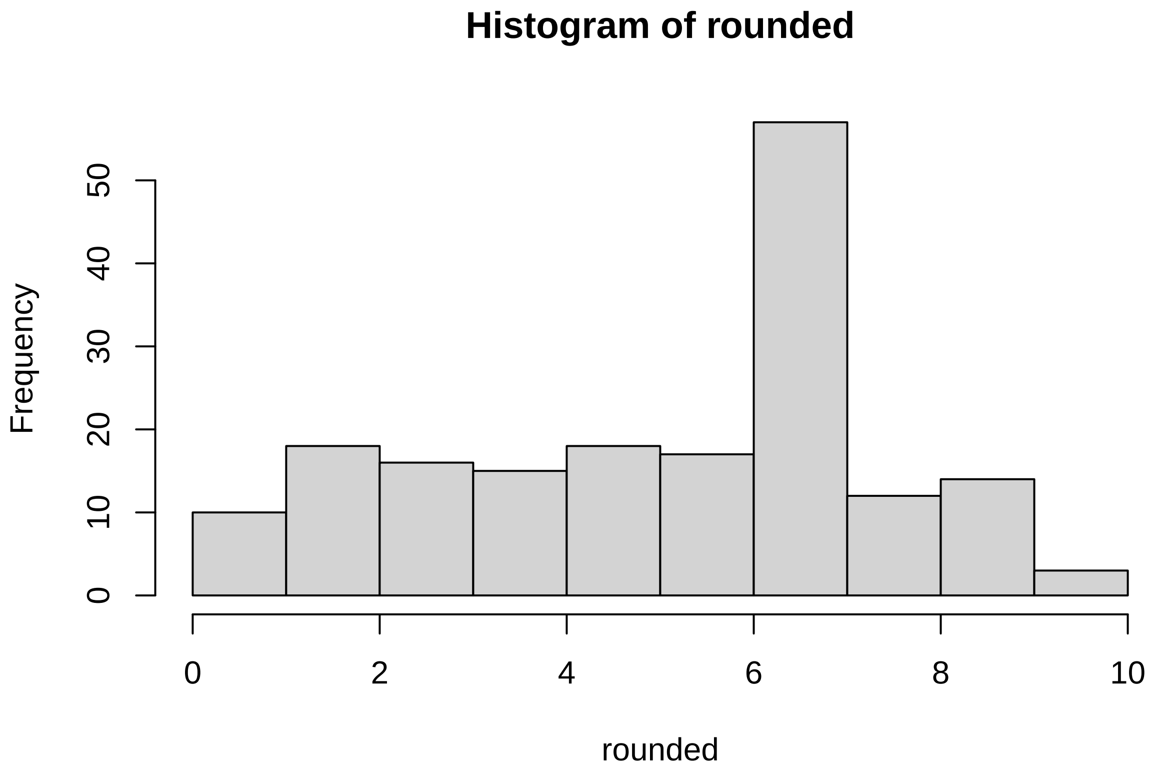
```
rounded = floor(data$number)
```

5b. If all 180 students truly chose a random number between 0-10 (say e.g. using a random number generator), approximately how many of each rounded digit would we expect to have in our data?

If we round digits down, we would expect an even number of students to be in the 0 - 9 categories. 10 may be lower since we didn't allow any decimals above 10. If we exclude 10 and think each other number will have an equal amount of students, we expect 180/10 = 18 students per digit.

5c. Look at the rounded numbers found in our actual data (either with a table of counts or with a histogram). Does the observed distribution match up with the expected distribution you described in part b.? Why or why not? Do you think that students generated truly random numbers based on your answer?

```
hist(rounded)
```



The digit 7 is way over represented compared to a truly random distribution. This indicates that students are not really picking numbers randomly as if chosen from a random number generator.

6a. Suppose a student makes the following claim: "I believe that students who prefer to live in Urban environments are more likely to have at least one tattoo than those who do not." Check this statement using the survey data. Describe your process in words alongside the steps you perform in code.

First lets look at the residentialpref categories:

```
unique(data$residentialpref)
```

```
## [1] "Rural (Small town)"         "Suburban"
## [3] "Urban/City"                 "Urban"
## [5] "Very Rural (Woods/Desert)"  "Rural (small town)"
## [7] "Very rural (woods/desert)"  "Very rural (forest/desert)"
```

There are 2 urban categories: "Urban" & "Urban/City". I could combine them like I did in part 2 or just subset on them using an OR statement and separate the data into `urban` and `not_urban` tibbles as follows:

```
urban_idx = (data$residentialpref == "Urban/City") | (data$residentialpref == "Urban")

#urban
urban = data[urban_idx,]

#others
not_urban = data[!urban_idx,]
```

Then I want to check the percentage of students which have at least 1 tattoo in each category:

```
urban_tattoo_pct = nrow(urban[(urban$tattoos>0),]) / nrow(urban)

not_urban_tattoo_pct = nrow(not_urban[(not_urban$tattoos>0),]) / nrow(not_urban)


urban_tattoo_pct / not_urban_tattoo_pct
```

```
## [1] 1.511364
```

Students who prefer urban environments are 1.5x as likely to have at least 1 tattoo according to our data.

6b. (Open-ended) Make your own hypothesis about a trend present in the data. Give a reason why you think this might be the case. Then check whether you were right or wrong according to the real observations.

Answers will vary. Here's an example answer:

Hypothesis: I think people who tend to prefer rural areas will like country music at greater rates than urban students.

First I'll use grep to find all students who have rural in their preferred areas. Because I don't want to care about caps lock, I can use `ignore.case=TRUE`:

```
rural_idx = grep("rural", data$residentialpref, ignore.case = TRUE)

rural = data[rural_idx,]
```

Then I'll compare the percentages of those who like country music from rural vs. urban:

```r
rural_country_pct = nrow(rural[rural$music == "Country",] ) / nrow(rural)


urban_country_pct = nrow(urban[urban$music == "Country",] ) / nrow(urban)

c(rural_country_pct, urban_country_pct)
```

```
## [1] 0.20930233 0.01515152
```

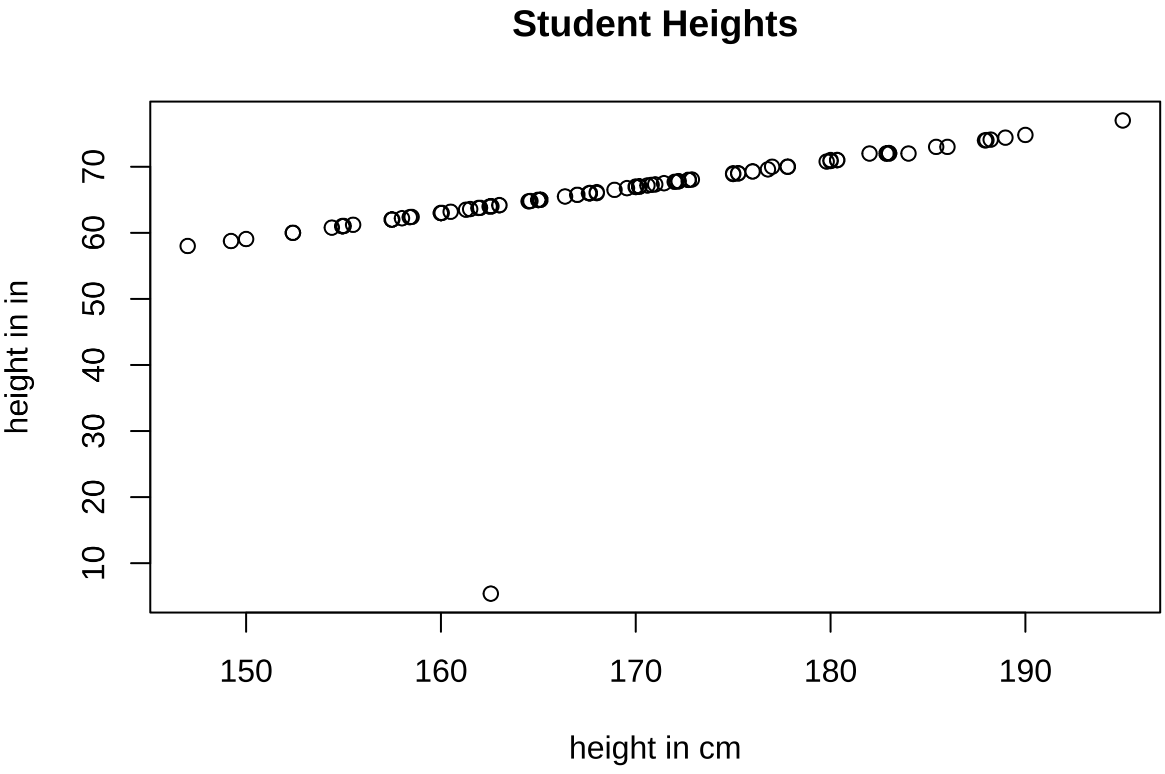
Almost 20% of students who prefer rural environments like country music best, while only 1.5% of students who prefer urban environments are country music lovers. My hunch was correct.

7a. Consider the two columns "heightcm" and "heightin". We know that 1in = 2.54cm. Because of this, what would we expect the data to look like if we plotted "heightcm" (x-axis) vs. "heightin" (y-axis)? (Answer this before continuing to part b.)

I would expect the points to fall perfectly on the line $y = 2.54x$ if there were no mistakes made in converting.

7b. Create the plot described in part a. with the `plot()` function. What do you observe? Does the plot match up with your expectations? Are there any notable exceptions/outliers? What is a potential explanation for this outlier?

```r
plot(data$heightcm,
     data$heightin,
     xlab = "height in cm",
     ylab = "height in in",
     main = "Student Heights")
```



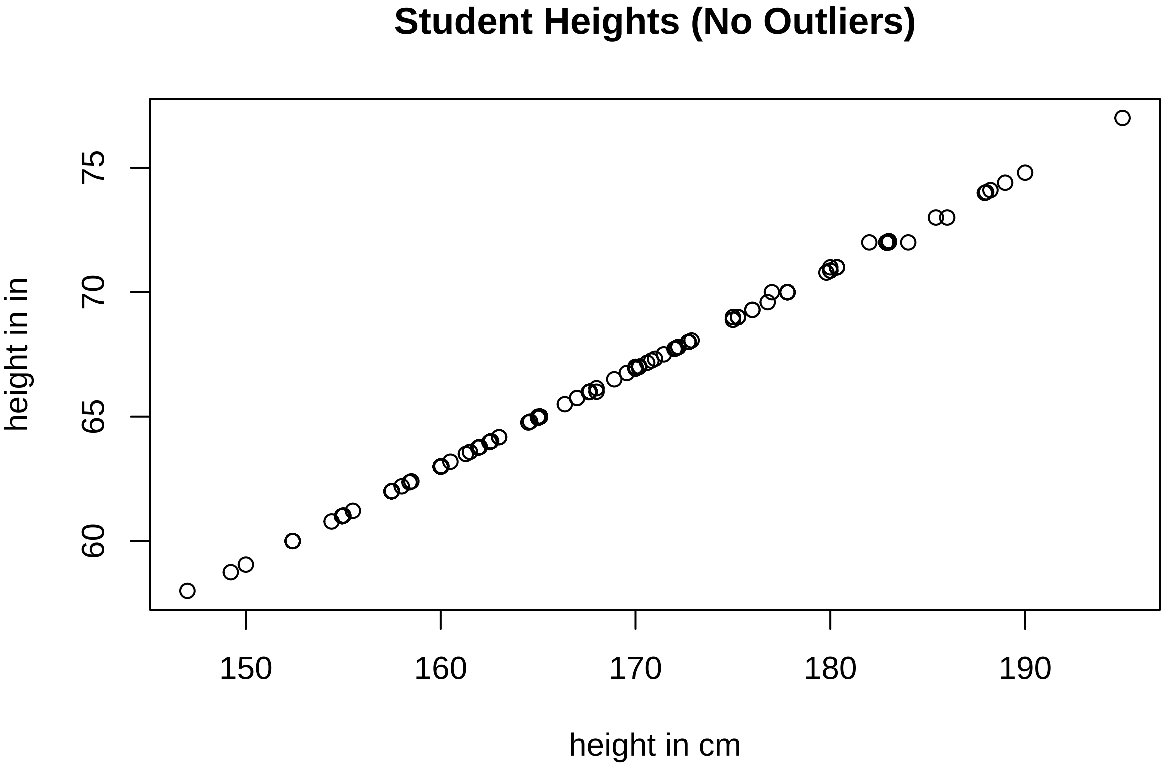
Most points fall close to a straight line except one outlier. Probably they had a typo when answering their height in inches on thes survey.

7c. If you found outliers, create a tibble with the outliers removed and remake the plot. With the outliers removed, do the points fall perfectly on a straight line now? If yes, explain why. If no, provide an explanation of a potential source of error.

To remove the outlier, I can notice that that student is the only one with a height in inches below 10in. So that will be my condition:

```
not_outliers = data[data$heightin>10, ]

plot(not_outliers$heightcm,
     not_outliers$heightin,
     xlab = "height in cm",
     ylab = "height in in",
     main = "Student Heights (No Outliers)")
```



The points are much closer to falling on the straight line. There are still some points which are not perfectly on the line. This is likely due to rounding errors made when converting.