Assignment: ASSIGNMENT 5

Name: Shekhar, Manish

Date: 2021-04-26

```
## Set the working directory to the root of your DSC 520 directory
## Not needed as data file is in the current working directory
## setwd("/home/jdoe/Workspaces/dsc520")

## Load the `data/r4ds/heights.csv` to
heights_df <- read.csv("./heights.csv")
head(heights_df)
```

```
##    earn   height    sex ed age  race
## 1 50000 74.42444   male 16  45 white
## 2 60000 65.53754 female 16  58 white
## 3 30000 63.62920 female 16  29 white
## 4 50000 63.10856 female 16  91 other
## 5 51000 63.40248 female 17  39 white
## 6  9000 64.39951 female 15  26 white
```

```
str(heights_df)
```

```
## 'data.frame':    1192 obs. of  6 variables:
##  $ earn  : num  50000 60000 30000 50000 51000 9000 29000 32000 2000 27000 ...
##  $ height: num  74.4 65.5 63.6 63.1 63.4 ...
##  $ sex   : chr  "male" "female" "female" "female" ...
##  $ ed    : int  16 16 16 16 17 15 12 17 15 12 ...
##  $ age   : int  45 58 29 91 39 26 49 46 21 26 ...
##  $ race  : chr  "white" "white" "white" "other" ...
```

```
summary(heights_df)
```

```
##       earn            height          sex                 ed
##  Min.   :   200   Min.   :57.50   Length:1192        Min.   : 3.0
##  1st Qu.: 10000   1st Qu.:64.01   Class :character   1st Qu.:12.0
##  Median : 20000   Median :66.45   Mode  :character   Median :13.0
##  Mean   : 23155   Mean   :66.92                      Mean   :13.5
##  3rd Qu.: 30000   3rd Qu.:69.85                      3rd Qu.:16.0
##  Max.   :200000   Max.   :77.05                      Max.   :18.0
##       age            race
##  Min.   :18.00   Length:1192
##  1st Qu.:29.00   Class :character
##  Median :38.00   Mode  :character
##  Mean   :41.38
##  3rd Qu.:51.00
##  Max.   :91.00
```

```
## Using `cor()` compute correlation coefficients for
## height vs. earn
# check if there is any NA in the data
sum(is.na(heights_df$height))
```

```
## [1] 0
```

```r
sum(is.na(heights_df$earn))
```

```
## [1] 0
```

```r
# get pearson, spearman, and kendall correlation coefficients
# no need to specify use as there are no NAs
cor(heights_df$height, heights_df$earn, method = 'pearson')
```

```
## [1] 0.2418481
```

```r
cor(heights_df$height, heights_df$earn, method = 'spearman')
```

```
## [1] 0.2682315
```

```r
cor(heights_df$height, heights_df$earn, method = 'kendall')
```

```
## [1] 0.1825669
```

```r
# below statement is throwing error
# Error in match.arg(method) : 'arg' must be of length 1
# cor(heights_df$height, heights_df$earn, method = c('pearson','spearman','kendall'))
```

```r
### age vs. earn
# check if there is any NA in the data
sum(is.na(heights_df$age))
```

```
## [1] 0
```

```r
sum(is.na(heights_df$earn))
```

```
## [1] 0
```

```r
# get pearson, spearman, and kendall correlation coefficients
# no need to specify use as there are no NAs
cor(heights_df$age, heights_df$earn, method = 'pearson')
```

```
## [1] 0.08100297
```

```r
cor(heights_df$age, heights_df$earn, method = 'spearman')
```

```
## [1] 0.1496324
```

```r
cor(heights_df$age, heights_df$earn, method = 'kendall')
```

```
## [1] 0.1101134
```

```r
### ed vs. earn
# check if there is any NA in the data
sum(is.na(heights_df$ed))
```

```
## [1] 0
```

```r
sum(is.na(heights_df$earn))
```

```
## [1] 0
```

```r
# get pearson, spearman, and kendall correlation coefficients
# no need to specify use as there are no NAs
cor(heights_df$ed, heights_df$earn, method = 'pearson')
```

```
## [1] 0.3399765
```

```r
cor(heights_df$ed, heights_df$earn, method = 'spearman')
```

```
## [1] 0.3417063
```

```r
cor(heights_df$ed, heights_df$earn, method = 'kendall')
```

```
## [1] 0.2541748
```

```r
## Spurious correlation
## The following is data on US spending on science, space, and technology in millions of today's dollar
## and Suicides by hanging strangulation and suffocation for the years 1999 to 2009
## Compute the correlation between these variables
tech_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731, 29449)
suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)
## create a data frame with two variables above
my_data <- data.frame(tech_spending, suicides)
my_data
```

```
##    tech_spending suicides
## 1          18079     5427
## 2          18594     5688
## 3          19753     6198
## 4          20734     6462
## 5          20831     6635
## 6          23029     7336
## 7          23597     7248
## 8          23584     7491
## 9          25525     8161
## 10         27731     8578
## 11         29449     9000
```

```r
## find correlation using cor() function
## as there is no NA, no need to specify use argument
cor(tech_spending, suicides)
```

```
## [1] 0.9920817
```

```r
## using data frame
cor(my_data, method = "pearson")
```

```
##                tech_spending  suicides
## tech_spending      1.0000000 0.9920817
## suicides           0.9920817 1.0000000
```