assignment_07_MunjewarSheetal

Sheetal M

2023-01-29

Install and Load required packages :

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v purrr   1.0.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
```

Set the working directory to the root of your DSC 520 directory

setwd("E:\Data_Science_DSC510\DSC520-Statistics\dsc520")

Load the `data/r4ds/heights.csv` to

heights_df <- read.csv("data/r4ds/heights.csv")

Using `cor()` compute correlation coefficients for

```
## Set the working directory to the root of your DSC 520 directory
setwd("E:\\Data_Science_DSC510\\DSC520-Statistics\\dsc520")

## Load the `data/r4ds/heights.csv` to
```

```r
heights_df <- read.csv("data/r4ds/heights.csv")

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" ...
```

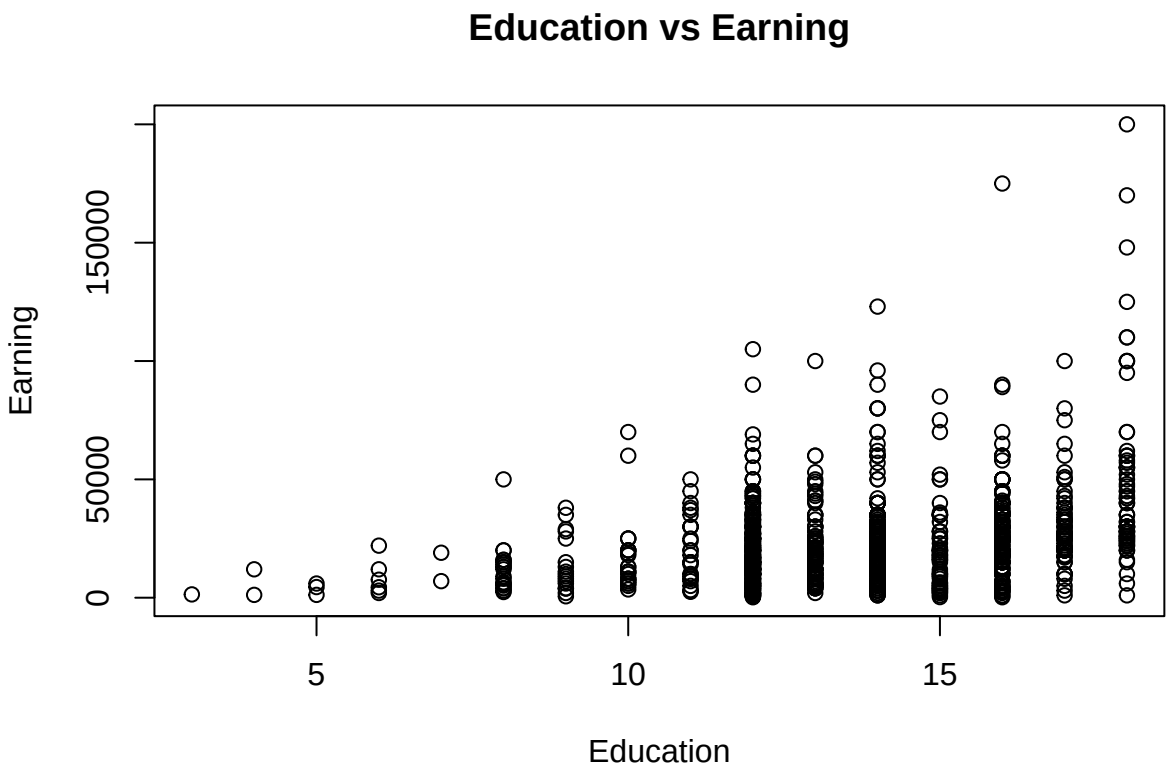
```r
### height vs. earn
cor(heights_df$earn,heights_df$height, method='pearson')
```

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

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

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

```r
### ed vs. earn
plot(heights_df$earn ~ heights_df$ed,
     data = heights_df,
     main = "Education vs Earning",
     xlab = "Education",
     ylab = "Earning")
```



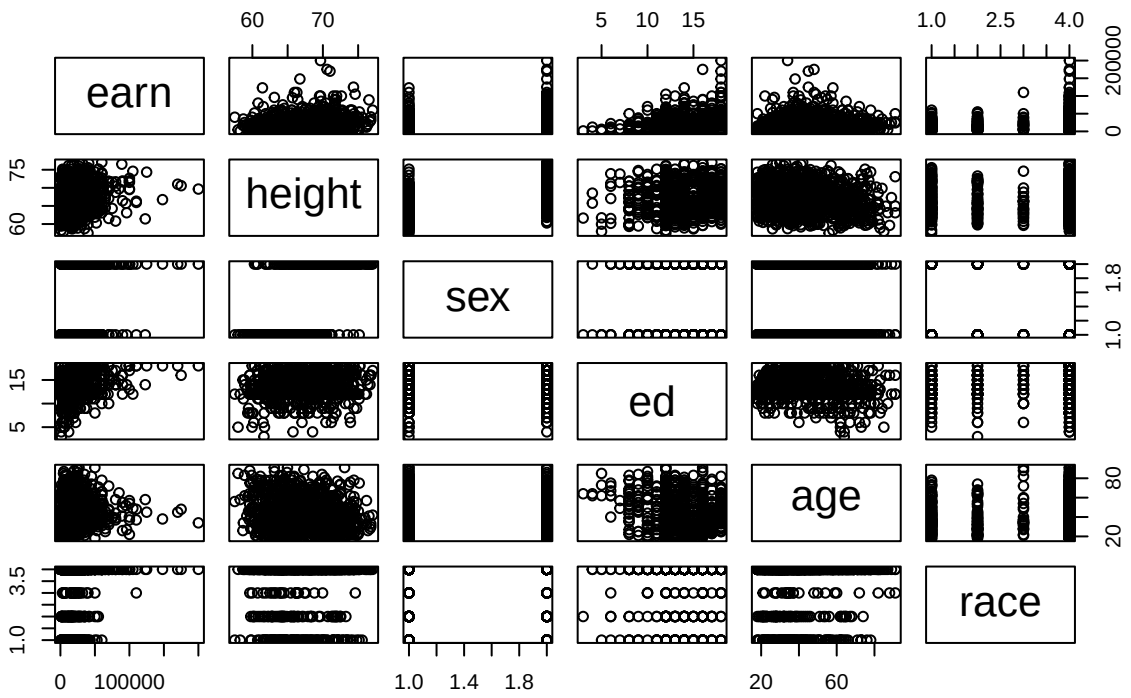

```
cor(heights_df$ed,heights_df$earn, method = 'kendall')
```

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

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

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

```
# cor(heights_df)
plot(heights_df, main = "Height vs Earning")
```



Spurious correlation

The following is data on US spending on science, space, and technology in millions of today's dollars

and Suicides by hanging strangulation and suffocation for the years 1999 to 2009

Compute the correlation between these variables

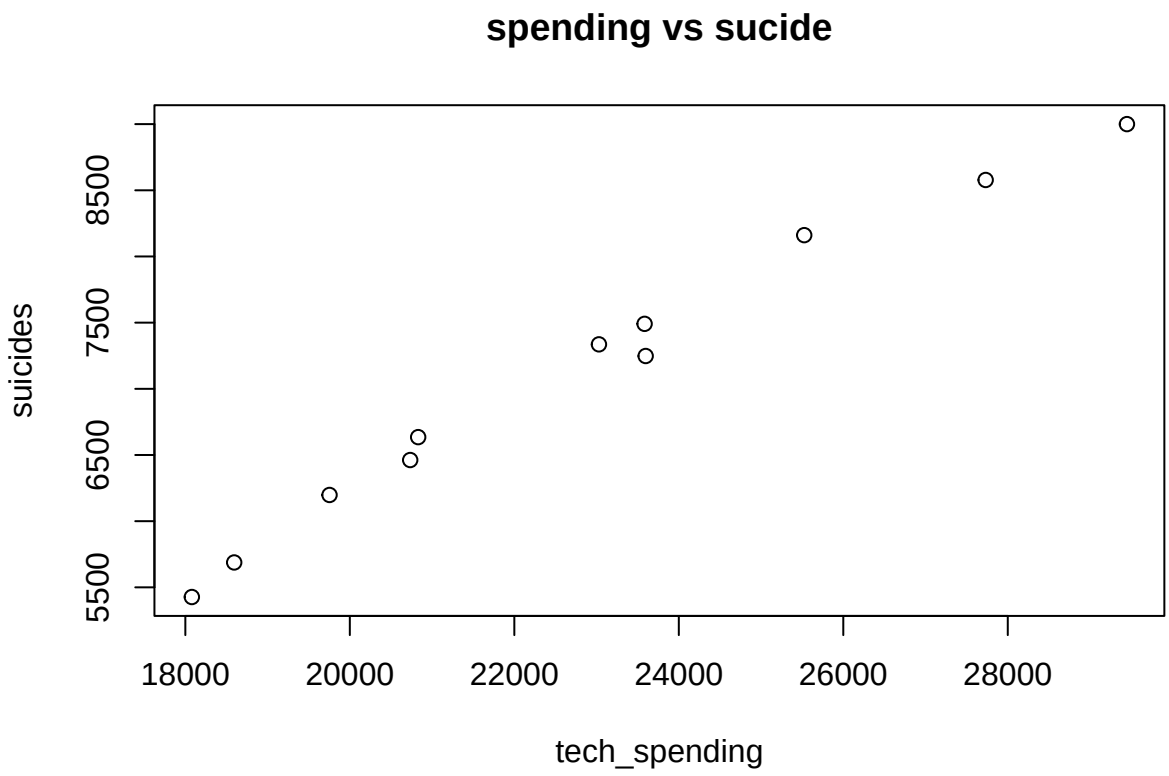
```
tech_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731, 29449)
suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)
ss_df <- data.frame(tech_spending,suicides)
str(ss_df)
```

```
## 'data.frame':    11 obs. of  2 variables:
##  $ tech_spending: num  18079 18594 19753 20734 20831 ...
##  $ suicides     : num  5427 5688 6198 6462 6635 ...
```

```
cor(ss_df)
```

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

```r
plot(ss_df, main = "spending vs sucide")
```




```r
cor(ss_df)
```

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

```r
cov(ss_df)
```

```
##                tech_spending suicides
## tech_spending      13465867  4210888
## suicides            4210888  1337883
```

```r
#calculate Pearson correlation coefficient and ignore any rows with NA
cor(ss_df$tech_spending,ss_df$suicides, use='complete.obs')
```

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