Exercise #2

Plot and Statistical Reasoning.

Brian Schweigler; 16-102-071

16/03/2022

Preliminaries

Loading the education dataset:

```
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

education_df = read.csv("education.txt", header = TRUE, sep = "\t")
```

And handle the outliers:

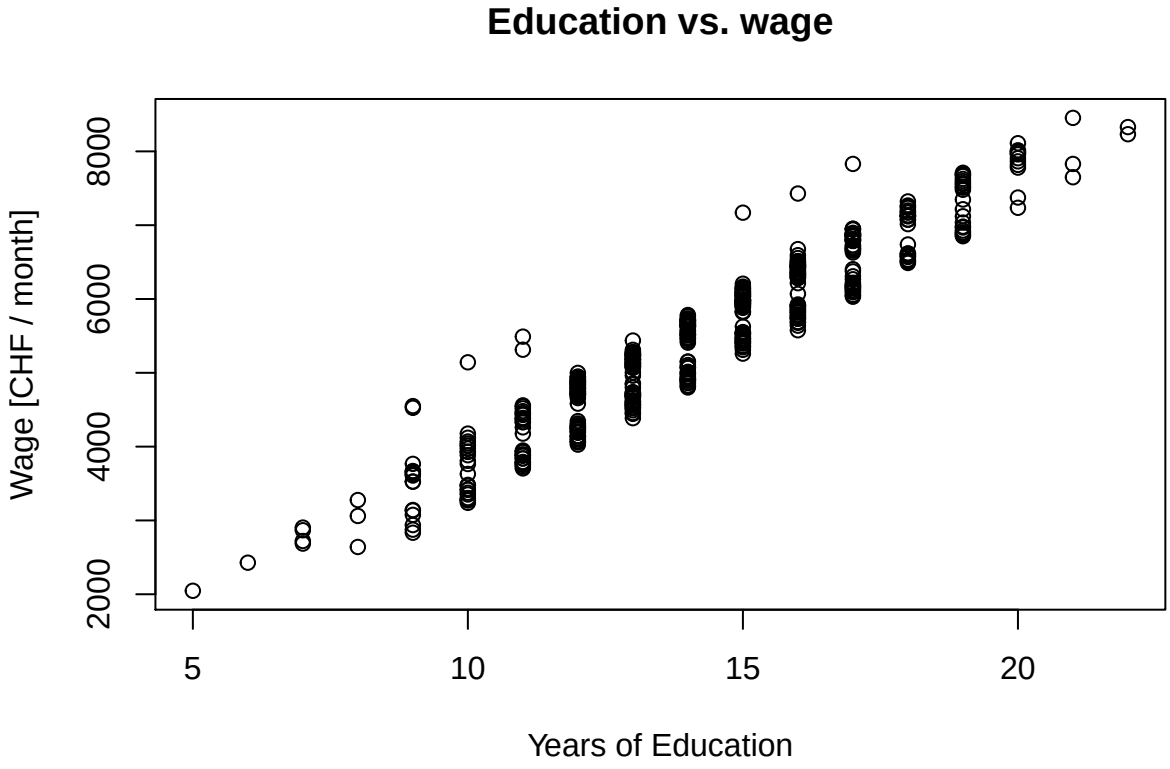
- A negative Education value (-12) in row 234;
- A non binary value for the Gender (20) in row 107;
- A very small Wage (41.8) in row 435.

```
cleaned_education_df <- education_df[-c(234, 107, 435), -1]
summary(cleaned_education_df)
```

```
##    Education         Gender           Wage
##  Min.   : 5.00   Min.   :1.000   Min.   :2047
##  1st Qu.:12.00   1st Qu.:1.000   1st Qu.:4696
##  Median :14.00   Median :1.000   Median :5520
##  Mean   :14.26   Mean   :1.398   Mean   :5479
##  3rd Qu.:16.00   3rd Qu.:2.000   3rd Qu.:6332
##  Max.   :22.00   Max.   :2.000   Max.   :8454
```

1. Using plot(); show possible relationship between independent variable Education and dependent variable Wage. Also, change the main title and and axes labels

```
plot(
    cleaned_education_df$Education,
    cleaned_education_df$Wage,
    xlab = "Years of Education",
    ylab = "Wage [CHF / month]",
    main = "Education vs. wage",
)
```

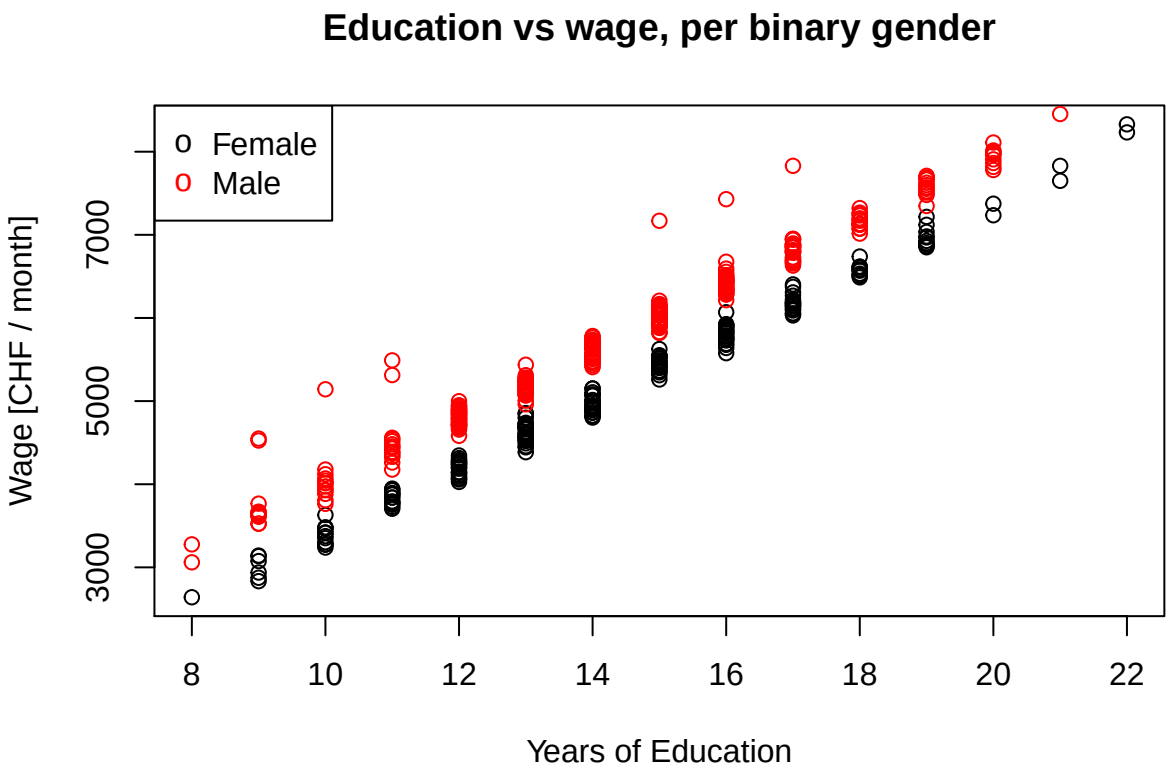


2. Split the above observation for men and women, showing both on the same plot

Splitting the dataset for the two gender.

```
female = cleaned_education_df[which(cleaned_education_df$Gender == 2),]
male = cleaned_education_df[which(cleaned_education_df$Gender == 1),]
```

```
plot(
  female$Education,
  female$Wage,
  col = "black",
  xlab = "Years of Education",
  ylab = "Wage [CHF / month]",
  main = "Education vs wage, per binary gender",
)
points(
  male$Education,
  male$Wage,
  col = "red",
)
legend(
  "topleft",
  legend = c("Female","Male"),
  col = c("Black", "red"),
  pch = c("o", "o"),
  ncol=1,
)
```



Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

---

3. Considering only the variables Result and Treatment what do you suggest as the most promising treatment for this person diagnosed with a tumor?

Loading the hospital dataset:

```
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

hospital_df = read.csv("Simpson.txt", header = TRUE, sep = " ")
```

```
treatment_options = unique(hospital_df$Treatment)

tumor_sizes = unique(hospital_df$Size)
levels(hospital_df$Size) = list(small = 1, large = 2)

treatment_outcomes = unique(hospital_df$Result)
levels(hospital_df$Result) = list(cured = 1, failed = 2)

cured = hospital_df[which(hospital_df$Result == 1), ]
cured_drugs = hospital_df[which(hospital_df$Result == 1 & hospital_df$Treatment == "drugs"), ]
cured_surgery = hospital_df[which(hospital_df$Result == 1 & hospital_df$Treatment == "surgery"), ]

failed = hospital_df[which(hospital_df$Result == 2), ]
failed_drugs = hospital_df[which(hospital_df$Result == 2 & hospital_df$Treatment == "drugs"), ]
failed_surgery = hospital_df[which(hospital_df$Result == 2 & hospital_df$Treatment == "surgery"), ]

success_ratio_general = dim(cured)[1] / (dim(cured)[1] + dim(failed)[1])
success_ratio_general
```

```
## [1] 0.7095
```

```
success_ratio_drugs = dim(cured_drugs)[1] / (dim(cured_drugs)[1] + dim(failed_drugs)[1])
success_ratio_drugs
```

```
## [1] 0.761
```

```
success_ratio_surgery = dim(cured_surgery)[1] / (dim(cured_surgery)[1] + dim(failed_surgery)[1])
success_ratio_surgery
```

```
## [1] 0.658
```

With no further information, the treatment via drugs has a higher likelihood of working than treatment via surgery. But what should be noted according to the dataset, is that surgery is denoted as the "2nd treatment". This likely means that if the drugs fail, surgery is the next step; so if the drugs have no effect, surgery would be the next step.

4. The Simpson's hospital discovers that the size of the tumor is small for this patient. Do you change your suggestion?

For this, we need to discern the impact of the tumor size on the treatment success:

```
small_cured = hospital_df[which(hospital_df$Result == 1 & hospital_df$Size == 1), ]
small_cured_drugs = hospital_df[which(hospital_df$Result == 1 & hospital_df$Treatment == "drugs" & hospi
small_cured_surgery = hospital_df[which(hospital_df$Result == 1 & hospital_df$Treatment == "surgery" & h

large_cured = hospital_df[which(hospital_df$Result == 1 & hospital_df$Size == 2), ]
large_cured_drugs = hospital_df[which(hospital_df$Result == 1 & hospital_df$Treatment == "drugs" & hospi
large_cured_surgery = hospital_df[which(hospital_df$Result == 1 & hospital_df$Treatment == "surgery" & h

small_failed = hospital_df[which(hospital_df$Result == 2 & hospital_df$Size == 1), ]
small_failed_drugs = hospital_df[which(hospital_df$Result == 2 & hospital_df$Treatment == "drugs" & hosp
small_failed_surgery = hospital_df[which(hospital_df$Result == 2 & hospital_df$Treatment == "surgery" &

large_failed = hospital_df[which(hospital_df$Result == 2 & hospital_df$Size == 2), ]
large_failed_drugs = hospital_df[which(hospital_df$Result == 2 & hospital_df$Treatment == "drugs" & hosp
large_failed_surgery = hospital_df[which(hospital_df$Result == 2 & hospital_df$Treatment == "surgery" &

small_success_ratio_general = dim(small_cured)[1] / (dim(small_cured)[1] + dim(small_failed)[1])
small_success_ratio_general
```

```
## [1] 0.8288191
```

```
small_success_ratio_drugs = dim(small_cured_drugs)[1] / (dim(small_cured_drugs)[1] + dim(small_failed_di
small_success_ratio_drugs
```

```
## [1] 0.8202934
```

```
small_success_ratio_surgery = dim(small_cured_surgery)[1] / (dim(small_cured_surgery)[1] + dim(small_fai
small_success_ratio_surgery
```

```
## [1] 0.8952381
```

```
large_success_ratio_general = dim(large_cured)[1] / (dim(large_cured)[1] + dim(large_failed)[1])
large_success_ratio_general
```

```
## [1] 0.6072423
```

```
large_success_ratio_drugs = dim(large_cured_drugs)[1] / (dim(large_cured_drugs)[1] + dim(large_failed_di
large_success_ratio_drugs
```

```
## [1] 0.4945055
```

```
large_success_ratio_surgery = dim(large_cured_surgery)[1] / (dim(large_cured_surgery)[1] + dim(large_fai
large_success_ratio_surgery
```

```
## [1] 0.6301676
```

Small tumors are more likely to be successfully treated than large tumors, regardless of treatment method. While small tumors are more likely to be successfully treated by drugs rather than surgery, surgery is still defined as the "second" treatment. One could vouch for going straight to surgery, as it is has a slightly higher success rate, but the difference is only half as large as in the case for large tumors (~7% instead of ~14%)

5. And if the Simpson's hospital discover that the size of the tumor is large, do you change your conclusion?

The recommendation is stil the same, as drugs are used before surgery. We can say that surgery is much more likely to work if the drugs do not work. Furthermore, if the lethality of the tumor would increase rapidly over time, we could go straight for the surgery, as it is 14% more likely to work than drugs for large tumors.

6. Are you consistent in the two cases above? If no, why?

Mostly consistent, simply because surgery is defined as "2nd treatment". Thus, without further information, the likelihood of surgery working includes having tried drugs beforehand.

A case can be made that surgery is significantly (14%) more likely to work than drugs for large tumors, but it is unclear to what extent the outcome will be different if drugs have not been tried before without more information on the dataset.