MertGöksel

Mert Göksel

4/29/2021

```
df <- read.table("/media/rootsyl/Aphostrof/Downloads/cps.txt")
names <- c(
  "EDUCATION",
  "SOUTH",
  "SEX",
  "EXPERIENCE",
  "UNION",
  "WAGE",
  "AGE",
  "RACE",
  "OCCUPATION",
  "SECTOR",
  "MARR"
)
names(df) <- names
df$SOUTH <- factor(df$SOUTH)
df$SEX <- factor(df$SEX)
df$UNION <- factor(df$UNION)
df$RACE <- factor(df$RACE)
df$OCCUPATION <- factor(df$OCCUPATION)
df$SECTOR <- factor(df$SECTOR)
df$MARR <- factor(df$MARR)
```

Question 1:

```
#A:
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
df_new <- df %>% select(c("SEX", "EXPERIENCE", "WAGE", "AGE", "MARR"))

#B:

df_new %>% select(where(is.numeric)) %>% cor()
```

```
##            EXPERIENCE       WAGE       AGE
## EXPERIENCE 1.00000000 0.08705953 0.9779612
## WAGE       0.08705953 1.00000000 0.1769669
## AGE        0.97796125 0.17696688 1.0000000
```

```
#C:

df_new_3 <- df %>% select(!c("SOUTH", "UNION", "MARR")) %>%
  filter(AGE > 30 & AGE < 50 & SECTOR == 2)
head(df_new_3)
```

```
##   EDUCATION SEX EXPERIENCE  WAGE AGE RACE OCCUPATION SECTOR
## 1        12   0         20  7.61  38    1          6      2
## 2        12   0         24 10.75  42    3          6      2
## 3        10   0         27  9.00  43    3          6      2
## 4        12   0         19 12.22  37    3          6      2
## 5        11   0         29  9.50  46    3          6      2
## 6        11   0         28 10.78  45    1          6      2
```

```
#D:

df_new_3 %>% mutate(New_Column = WAGE / AGE) %>%
  filter(New_Column>0.25) %>% nrow() %>%
  cat("There are", ., "observations satisfying that condition")
```

```
## There are 2 observations satisfying that condition
```

```
#E:

df %>% filter(SEX==1) %>% aggregate(WAGE~OCCUPATION,., mean)
```

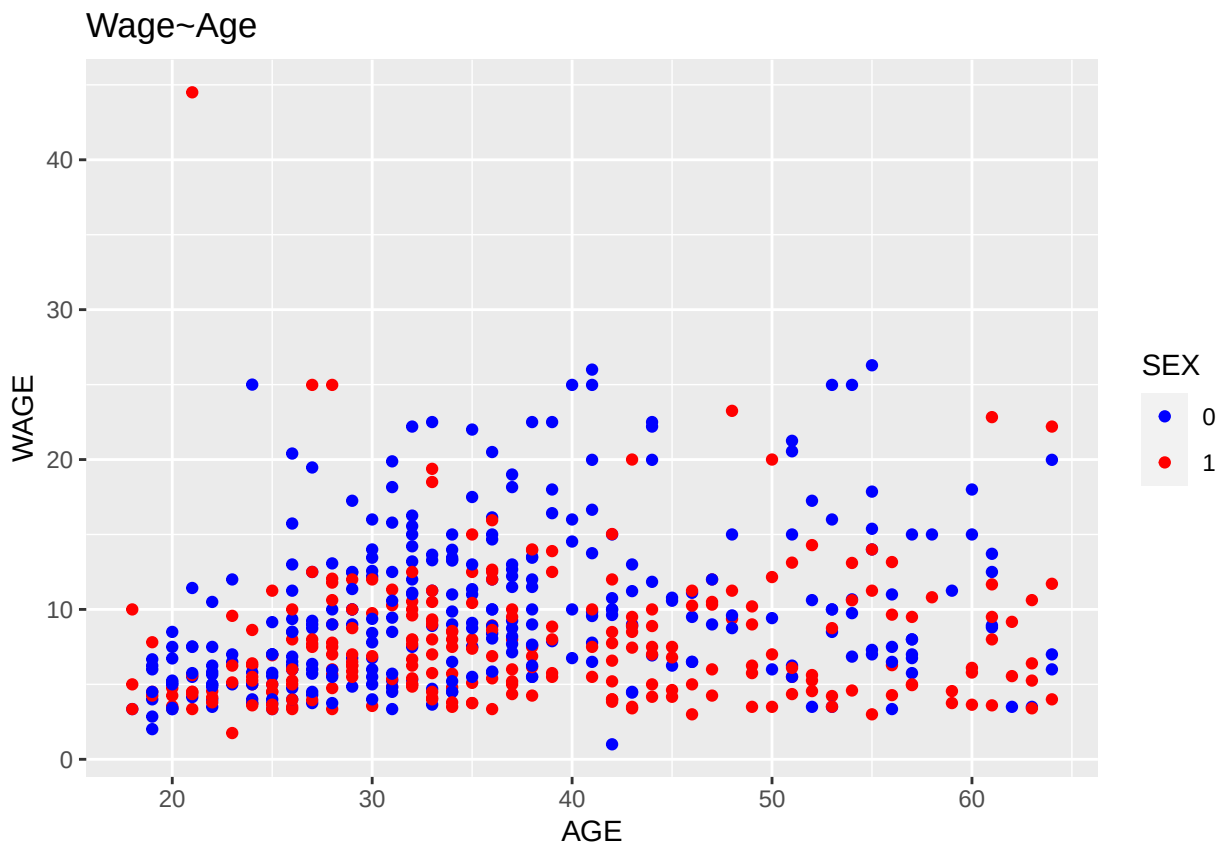
```
##   OCCUPATION      WAGE
## 1          1 11.056190
## 2          2  5.241765
## 3          3  7.404211
## 4          4  6.059388
## 5          5 11.105000
## 6          6  5.731333
```

```
#F:

df %>% aggregate(SEX~MARR,.,table) #This table works :)
```

```
##   MARR SEX.0 SEX.1
## 1    0   101    83
## 2    1   188   162
```

Question 2:

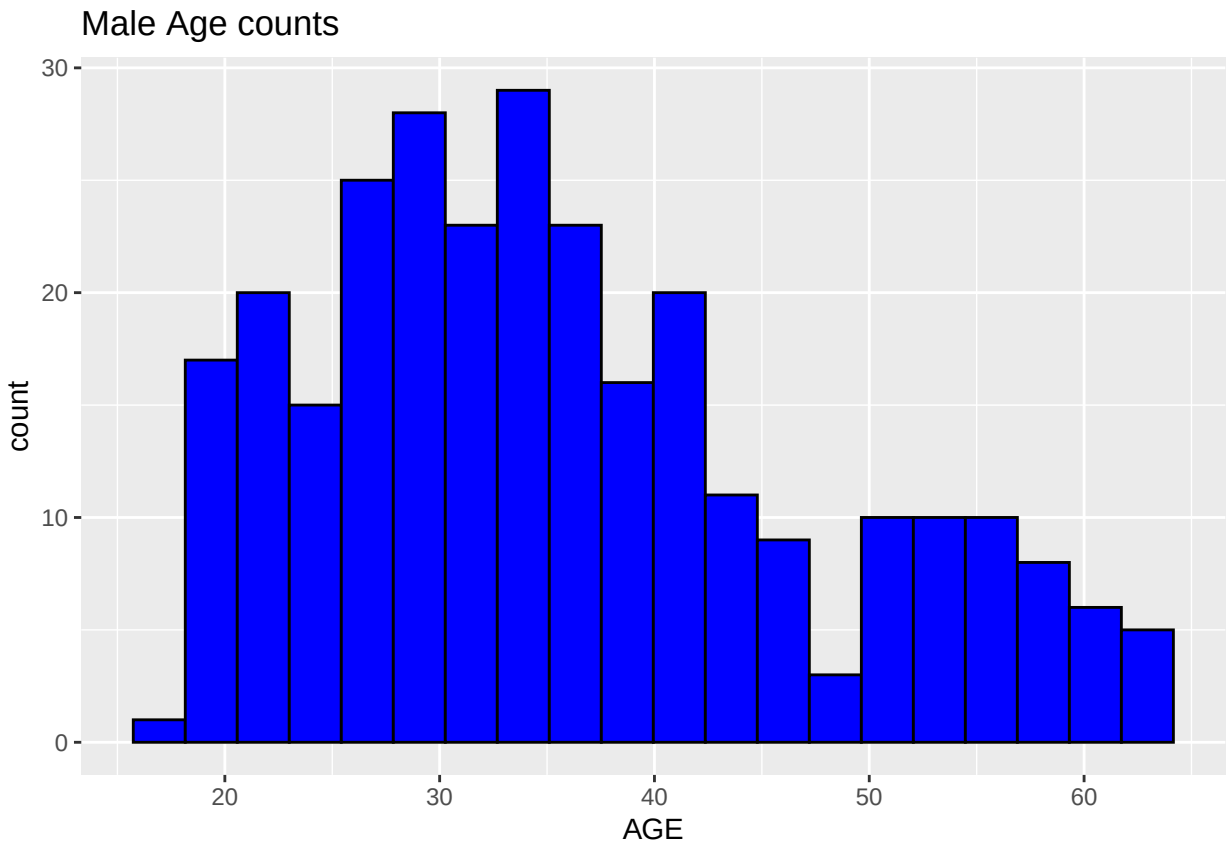
```
library(ggplot2)

#A:
ggplot(df,aes(x=AGE, y=WAGE, color=SEX)) + geom_point() +
  scale_color_manual(values=c("Blue", "Red")) + ggtitle("Wage~Age")
```



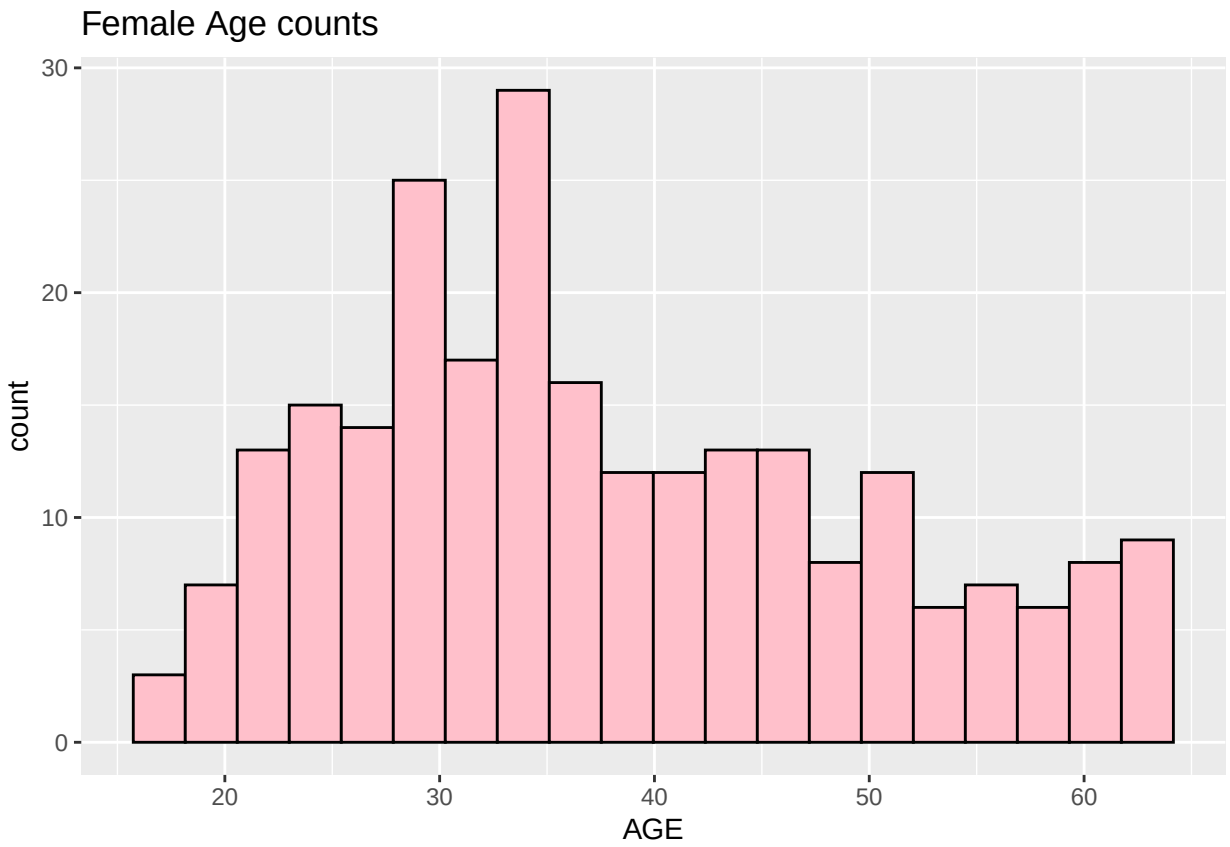
```
#I used blue and red for colors to be in sync with SEX
#Apart from the outlier in the start,
#men seems to be higher paid as their age increases.

#B:
df %>% ggplot(aes(x = WAGE, color=RACE, size=.2)) +
  geom_density(linetype = "dashed") + scale_linetype_discrete(3) +
  ggtitle("Wage density plot")
```


```
#Most abundant race seems to be number 2

#C:
library(ggpubr)
levels(df$OCCUPATION) <- c("Management", "Sales", "Clerical",
                           "Service", "Professional", "Other")
levels(df$RACE) <- c("Other", "Hispanic", "White")
df %>% ggplot(aes(x = RACE, y = WAGE, fill=RACE)) +
  geom_boxplot() + facet_wrap(vars(OCCUPATION)) + ylab("WAGE ($/hour)") +
  ggtitle("Boxplots for WAGE")
```


```
#In every plot except for "Professional" and "Service" it
#seems that White people are dominant.
#D:
a <- df %>% filter(SEX==0) %>% ggplot(aes(x=AGE)) +
  geom_histogram(fill="blue", bins = 20, color="black") +
  ggtitle("Male Age counts")
a
```



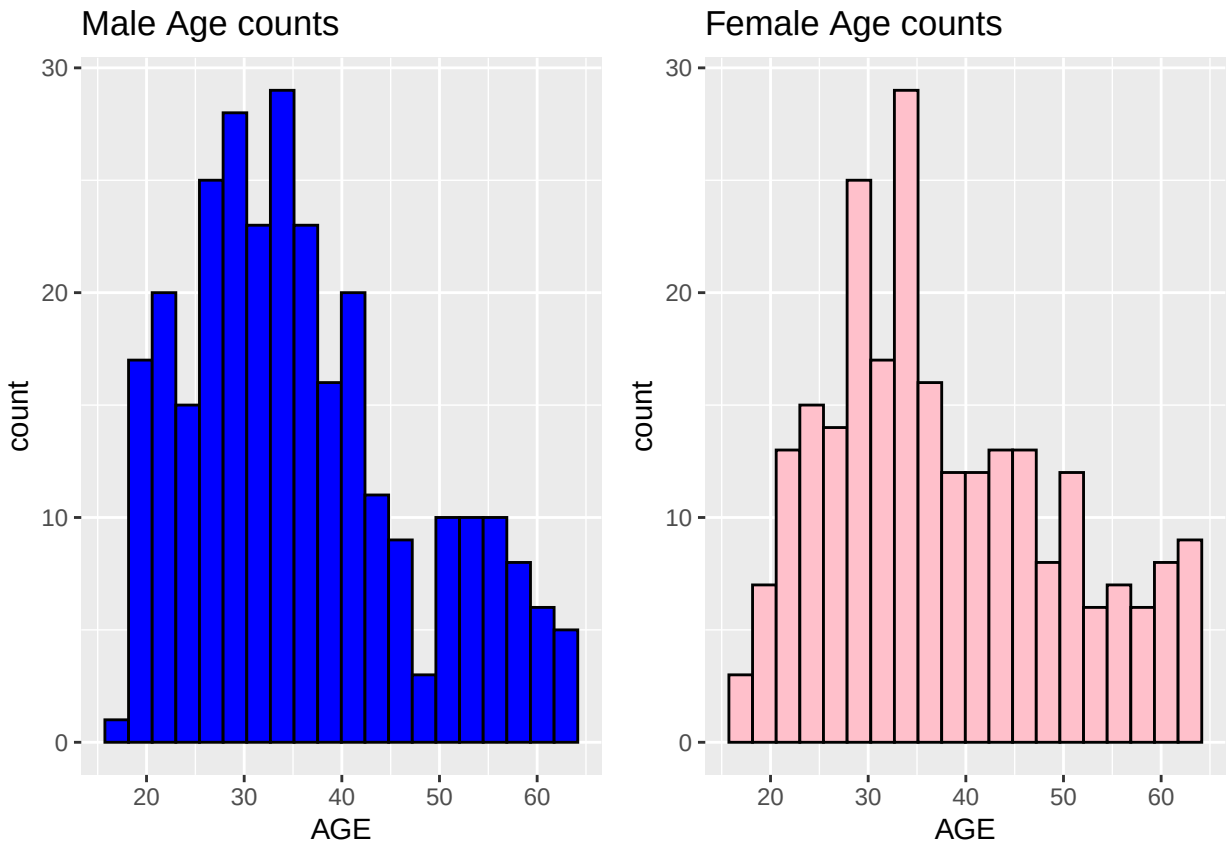

```
b <- df %>% filter(SEX==1) %>% ggplot(aes(x=AGE)) +
  geom_histogram(fill="pink", bins = 20, color="black") +
  ggtitle("Female Age counts")
b
```



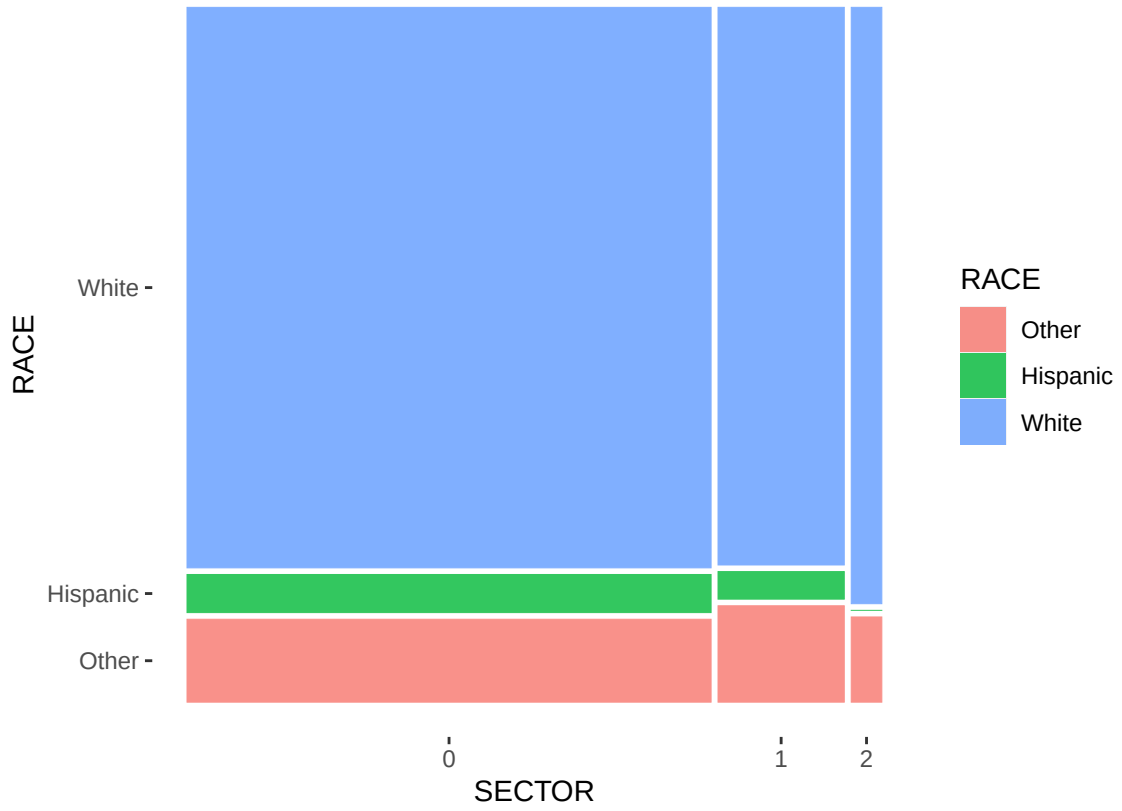

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(a, b, ncol=2)
```




```
#Female sample is more than males when it comes
#to age>40 that means females live longer?

library(ggmosaic)

#E:
df %>% ggplot() + geom_mosaic(aes(x=product(SECTOR),
                                  fill=RACE)) + theme_mosaic()
```




```
#In every sector white is overwhelming against other races.
```

Question 3:

```
set.seed(292)

#A:
obj1 <- list("X"=rnorm(1000), "Y"=rnorm(50, 10, 2), "Z"=runif(200,-5,20))
lapply(obj1, mean)
```

```
## $X
## [1] 0.01017814
##
## $Y
## [1] 9.964074
##
## $Z
## [1] 7.781811
```

```
#B:
obj2 <- matrix(obj1$X, ncol = 20, nrow=50)
apply(obj2, 2, sd)
```

```
##  [1] 0.9659344 1.0158556 1.0300657 1.0717474 1.0119949 1.1852774 0.8709294
##  [8] 0.8717674 1.0302342 0.9066866 0.9358745 1.0710708 0.9410272 0.8039254
## [15] 0.9573222 0.8201377 1.0657517 0.8759060 1.0488815 0.9832693
```

```
#C:
obj3 <- data.frame(obj1$Z, let=rep(LETTERS[1:4], each=50))
tapply(X = obj3$obj1.Z, INDEX = obj3$let, FUN=mean)
```

```
##        A        B        C        D
## 8.304322 8.090143 6.606004 8.126775
```

```
#D:
matrix(unlist(tapply(X = obj1$Y, INDEX = rep(1:5, each=10), FUN = summary)),
       ncol = 6, byrow = TRUE)[,c(1,6)]
```

```
##          [,1]     [,2]
## [1,] 6.530325 15.12361
## [2,] 7.149340 12.51637
## [3,] 5.125965 13.83990
## [4,] 6.309126 11.67108
## [5,] 7.154169 11.99404
```

```
#This may be too complicated to read, so i give another solution

ab <- obj1$Y %>% split(.,gl(5, 10)) %>% lapply(.,summary) %>% unlist(.) %>%
  matrix(ncol = 6, byrow = TRUE)
ab[,c(1,6)]
```

```
##          [,1]     [,2]
## [1,] 6.530325 15.12361
## [2,] 7.149340 12.51637
## [3,] 5.125965 13.83990
## [4,] 6.309126 11.67108
## [5,] 7.154169 11.99404
```