R Bridge course work week 2

Frederick Jones

2023-07-22

The data wage2 was obtained from the website http://vincentarelbundock.github.io/Rdatasets/. The data is about wages of employees and their education, experience, age etc.

The summary of the data:

```
dataset = read.csv('wage2.csv')
summary(dataset)
```

```
##        X              wage            hours             IQ       
##  Min.   :  1.0   Min.   : 115.0   Min.   :20.00   Min.   : 50.0  
##  1st Qu.:234.5   1st Qu.: 669.0   1st Qu.:40.00   1st Qu.: 92.0  
##  Median :468.0   Median : 905.0   Median :40.00   Median :102.0  
##  Mean   :468.0   Mean   : 957.9   Mean   :43.93   Mean   :101.3  
##  3rd Qu.:701.5   3rd Qu.:1160.0   3rd Qu.:48.00   3rd Qu.:112.0  
##  Max.   :935.0   Max.   :3078.0   Max.   :80.00   Max.   :145.0  
##                                                                  
##       KWW             educ           exper           tenure      
##  Min.   :12.00   Min.   : 9.00   Min.   : 1.00   Min.   : 0.000  
##  1st Qu.:31.00   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.: 3.000  
##  Median :37.00   Median :12.00   Median :11.00   Median : 7.000  
##  Mean   :35.74   Mean   :13.47   Mean   :11.56   Mean   : 7.234  
##  3rd Qu.:41.00   3rd Qu.:16.00   3rd Qu.:15.00   3rd Qu.:11.000  
##  Max.   :56.00   Max.   :18.00   Max.   :23.00   Max.   :22.000  
##                                                                  
##       age           married          black            south       
##  Min.   :28.00   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:30.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :33.00   Median :1.000   Median :0.0000   Median :0.0000  
##  Mean   :33.08   Mean   :0.893   Mean   :0.1283   Mean   :0.3412  
##  3rd Qu.:36.00   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:1.0000  
##  Max.   :38.00   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
##                                                                   
##      urban             sibs           brthord          meduc       
##  Min.   :0.0000   Min.   : 0.000   Min.   : 1.000   Min.   : 0.00  
##  1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 8.00  
##  Median :1.0000   Median : 2.000   Median : 2.000   Median :12.00  
##  Mean   :0.7176   Mean   : 2.941   Mean   : 2.277   Mean   :10.68  
##  3rd Qu.:1.0000   3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.:12.00  
##  Max.   :1.0000   Max.   :14.000   Max.   :10.000   Max.   :18.00
```

```
##                                   NA's   :83       NA's   :78
##      feduc           lwage
##  Min.   : 0.00   Min.   :4.745
##  1st Qu.: 8.00   1st Qu.:6.506
##  Median :10.00   Median :6.808
##  Mean   :10.22   Mean   :6.779
##  3rd Qu.:12.00   3rd Qu.:7.056
##  Max.   :18.00   Max.   :8.032
##  NA's   :194
```

Question 1

Mean and median of two attributes of data.

Mean and median of wages

```
mean_wage <- mean(dataset$wage)
sprintf("The mean wage is %.2f", mean_wage)
```

```
## [1] "The mean wage is 957.95"
```

```
median_wage <- median(dataset$wage)
sprintf("The median wage is %.2f", median_wage)
```

```
## [1] "The median wage is 905.00"
```

Mean and median of ages

```
mean_age <- mean(dataset$age)
sprintf("The mean age is %.2f", mean_age)
```

```
## [1] "The mean age is 33.08"
```

```
median_age <- median(dataset$age)
sprintf("The median age is %.2f", median_age)
```

```
## [1] "The median age is 33.00"
```

Question 2

Subset of data

Here rows from 1 to 100 and columns from 1 to 10 are taken in the subset dataframe df

```r
df <- dataset[1:100, 1:10]
head(df)
```

```
##   X wage hours  IQ KWW educ exper tenure age married
## 1 1  769    40  93  35   12    11      2  31       1
## 2 2  808    50 119  41   18    11     16  37       1
## 3 3  825    40 108  46   14    11      9  33       1
## 4 4  650    40  96  32   12    13      7  32       1
## 5 5  562    40  74  27   11    14      5  34       1
## 6 6 1400    40 116  43   16    14      2  35       1
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
df2 <- select(dataset,
              wage, hours, educ, exper, tenure, age, married)
head(df2)
```

```
##   wage hours educ exper tenure age married
## 1  769    40   12    11      2  31       1
## 2  808    50   18    11     16  37       1
## 3  825    40   14    11      9  33       1
## 4  650    40   12    13      7  32       1
## 5  562    40   11    14      5  34       1
## 6 1400    40   16    14      2  35       1
```

The subset of data containing all the married employees whos age is greater than 30 years.

```r
df3 <- filter(df2,
              married ==1, age >30)
head(df3)
```

```
##   wage hours educ exper tenure age married
## 1  769    40   12    11      2  31       1
## 2  808    50   18    11     16  37       1
## 3  825    40   14    11      9  33       1
## 4  650    40   12    13      7  32       1
## 5  562    40   11    14      5  34       1
## 6 1400    40   16    14      2  35       1
```

Question 3

The dataframe df3 will be used for further working with this assignement. df and df2 were just the experiemental subsets on how to take subset from given dataframe.

Renaming columns

```
df3 <- rename(df3,  Salary = wage, Hours = hours,  Education = educ, Experience =exper, Tenure =tenure,
head(df3)
```

```
##   Salary Hours Education Experience Tenure Age Married
## 1    769    40        12         11      2  31       1
## 2    808    50        18         11     16  37       1
## 3    825    40        14         11      9  33       1
## 4    650    40        12         13      7  32       1
## 5    562    40        11         14      5  34       1
## 6   1400    40        16         14      2  35       1
```

Question 4 ## Summary of the newly created dataframe df3

```
summary(df3)
```

```
##      Salary          Hours          Education       Experience       Tenure
##  Min.   : 200   Min.   :20.00   Min.   : 9.00   Min.   : 1.0   Min.   : 0.000
##  1st Qu.: 732   1st Qu.:40.00   1st Qu.:12.00   1st Qu.: 9.0   1st Qu.: 3.000
##  Median : 962   Median :40.00   Median :12.00   Median :13.0   Median : 8.000
##  Mean   :1013   Mean   :44.03   Mean   :13.49   Mean   :12.6   Mean   : 7.956
##  3rd Qu.:1202   3rd Qu.:48.00   3rd Qu.:16.00   3rd Qu.:16.0   3rd Qu.:12.000
##  Max.   :3078   Max.   :80.00   Max.   :18.00   Max.   :23.0   Max.   :22.000
##       Age           Married
##  Min.   :31.00   Min.   :1
##  1st Qu.:32.00   1st Qu.:1
##  Median :35.00   Median :1
##  Mean   :34.56   Mean   :1
##  3rd Qu.:37.00   3rd Qu.:1
##  Max.   :38.00   Max.   :1
```

Mean and median of wage and age are changed becuase df3 contains the data only for married employees whose age is greater than 30 years old.

```
mean_wage <- mean(df3$Wage)
```

```
## Warning in mean.default(df3$Wage): argument is not numeric or logical:
## returning NA
```

```
sprintf("The mean wage is %.2f", mean_wage)
```

```
## [1] "The mean wage is NA"
```

```
median_wage <- median(df3$Wage)
sprintf("The median wage is %.2f", median_wage)
```

```
## character(0)
```

```
mean_age <- mean(df3$Age)
sprintf("The mean age is %.2f", mean_age)
```

```
## [1] "The mean age is 34.56"
```

```
median_age <- median(df3$Age)
sprintf("The median age is %.2f", median_age)
```

```
## [1] "The median age is 35.00"
```

The mean and meadian of wage and age has been changed because df3 contains the data only for employees who are married and older than 30 years. So, df3 contains less number of rows as compared to original dataset. Hence, mean and median changed.

Question 5

I can replace in the column urban with the value 1 'Urban' and 0 with 'not-Urban' in a new dataset df4 which will be a copy of the original dataset.

```
df4 <- dataset
df4$urban[df4$urban== 1] <-'Urban'
df4$urban[df4$urban== 0] <-'not-Urban'
df4$married[df4$married==1] <-'Married'
df4$married[df4$married==0] <-'UnMarried'
head(df4, 10)
```

```
##     X wage hours  IQ KWW educ exper tenure age   married black south     urban
## 1   1  769    40  93  35   12    11      2  31   Married     0     0     Urban
## 2   2  808    50 119  41   18    11     16  37   Married     0     0     Urban
## 3   3  825    40 108  46   14    11      9  33   Married     0     0     Urban
## 4   4  650    40  96  32   12    13      7  32   Married     0     0     Urban
## 5   5  562    40  74  27   11    14      5  34   Married     0     0     Urban
## 6   6 1400    40 116  43   16    14      2  35   Married     1     0     Urban
## 7   7  600    40  91  24   10    13      0  30 UnMarried     0     0     Urban
## 8   8 1081    40 114  50   18     8     14  38   Married     0     0     Urban
## 9   9 1154    45 111  37   15    13      1  36   Married     0     0 not-Urban
## 10 10 1000    40  95  44   12    16     16  36   Married     0     0     Urban
##    sibs brthord meduc feduc    lwage
## 1     1       2     8     8 6.645091
## 2     1      NA    14    14 6.694562
```

```
## 3     1       2    14    14 6.715384
## 4     4       3    12    12 6.476973
## 5    10       6     6    11 6.331502
## 6     1       2     8    NA 7.244227
## 7     1       2     8     8 6.396930
## 8     2       3     8    NA 6.985642
## 9     2       3    14     5 7.050990
## 10    1       1    12    11 6.907755
```

##Question 6 all outputs contain more then 5 rows.

##Question 7 Bonus Question

```
library(readr)
##dfremote <- read_csv("https://github.com/jewelercart/R/blob/main/wage2.csv")
##head(dfremote)
```