hw_wk1.r

yunting

2020-08-30

```
# title: "hw_wk1"
# author: "Yunting Chiu"
# date: "8/28/2020"

# 1. Ans: I discussed with Sihyuan Han.
# 2. Ans: In Case Study 1.2, it has allocated two segments: male and female in the initial stage.
          #Compared to Case study 1.1, it is randomly sampled and not classified at the beginning.

# 3. Using R find numerical and graphical summaries of this data.
#Use these to describe the distribution of the starting salaries for both males and females.

# Libraies and read Data

library(readr)
library(tidyverse)
```

```
## -- Attaching packages ------ tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v dplyr   1.0.1
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.1     v forcats 0.5.0
## v purrr   0.3.4
```

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

```
library(ggplot2)
library(mosaic)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggformula
```

```
## Loading required package: ggstance
```

```
##
## Attaching package: 'ggstance'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

```
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
```

```
## Loading required package: mosaicData
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.
##
## Have you tried the ggformula package for your plots?
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

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

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
SDE <- read_csv(file = "../data/case0102.csv")
```

```
## Parsed with column specification:
## cols(
##   Salary = col_double(),
##   Sex = col_character()
## )
```

```
tail(SDE)
```

```
## # A tibble: 6 x 2
##   Salary Sex
##    <dbl> <chr>
## 1   6600 Male
## 2   6600 Male
## 3   6840 Male
## 4   6900 Male
## 5   6900 Male
## 6   8100 Male
```

```
# 3a.   Give and interpret the mean salary and standard deviation of salaries for females.
#Do this also for males.
#mean salary of female and male
SDE %>%
  filter(Sex == "Female") -> SDE_Female
mean(SDE_Female$Salary)
```

```
## [1] 5138.852
```

```
SDE %>%
  filter(Sex == "Male") -> SDE_Male
mean(SDE_Male$Salary)
```

```
## [1] 5956.875
```

```
# SD
sd(SDE_Female$Salary)  #SD for Female
```

```
## [1] 539.8707
```

```
sd(SDE_Male$Salary) # SD for Male
```

```
## [1] 690.7333
```

```
# 3b.   Give and interpret the median salary and the IQR of salaries for females.
# Do this also for males.
# find interquartile range of female in R
summary(SDE_Female$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3900    4800    5220    5139    5400    6300
```

```
# find interquartile range of male in R
summary(SDE_Female$Salary)
```

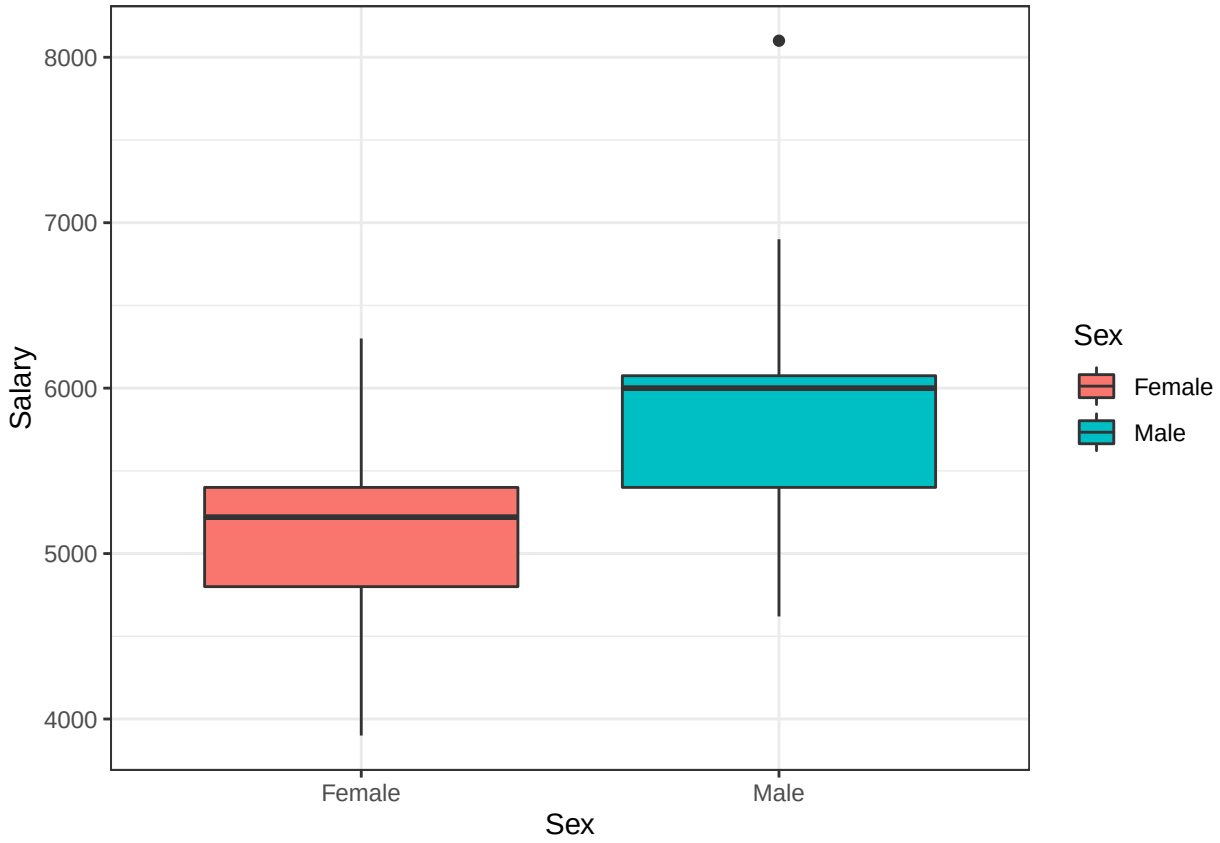
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3900    4800    5220    5139    5400    6300
```

```
# 3c.   Give a histogram of salaries for each group.
hist(SDE_Female$Salary)
```


```
hist(SDE_Male$Salary)
```


```
# 3d.   Give side-by-side boxplots of salaries.
SDE %>%
  ggplot(mapping = aes(x = Sex, y = Salary, fill = Sex))+
```

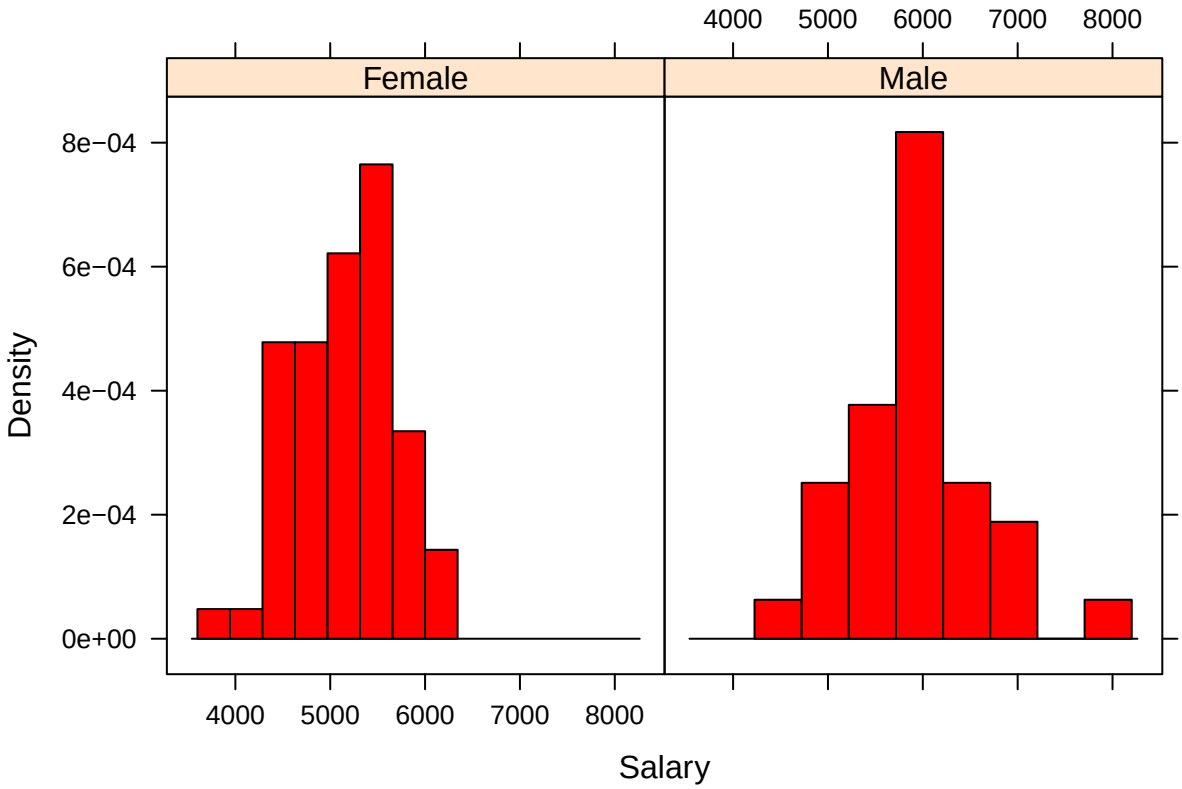
```
  geom_boxplot()+
  theme_bw()
```




```
# 3e. Use a. to d. to describe the distribution of salaries for each group.
favstats(Salary ~ Sex, data = SDE)
```

```
##      Sex  min   Q1 median   Q3  max     mean       sd  n missing
## 1 Female 3900 4800   5220 5400 6300 5138.852 539.8707 61       0
## 2   Male 4620 5400   6000 6075 8100 5956.875 690.7333 32       0
```

```
histogram(~Salary | Sex, data = SDE, col = "red")
```




```
# 3f. Ans: Female group's salaries are larger since the sample size is bigger,
      #but male's average salaries are larger,
      #because 5956.875 (male's mean salaries) > 5138.852 (female's mean salaries)

# 4. Ans: No errors on my R code

# 5. Ans: There are still some other factors we should consider,
    # such as male may have had more years of previous experience.
```