Problem Set 1

Zhengting (Johnathan) He

2021/9/2

```
# set working directory
require(tidyverse)
```

```
## Loading required package: tidyverse
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

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

```
setwd("D:/OneDrive - Johns Hopkins/Course/140.621.81 - Statistical Methods in Public Health I/Problem set/
jhsphbiostat621-assignment/Problem set 1")
```

Problem set 1: Displaying and Thinking About Public Health Data EDA

Problem 1. Air Pollution and Mortality in Baltimore

Section 1: Exploratory Data Analysis

```
# import data
ps1 <- read_csv("./data/baltps11.csv")
```

```
## Rows: 40 Columns: 2
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): group, deaths
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ps1$group <- factor(ps1$group, levels = c("1", "2"))
```

a. Create stem and leaf displays by group:

```
for (pollut in c("1", "2")) {
    print(ifelse(pollut == "Highest", "# Daily mortality count on Highest Particulate Pollution Days",
        "# Daily mortality count on Lowest Particulate Pollution Days"))
    stem(filter(ps1, group == pollut)$deaths)
}
```

```
## [1] "# Daily mortality count on Lowest Particulate Pollution Days"
## 
##   The decimal point is 1 digit(s) to the right of the |
## 
##   1 | 44
##   1 | 567778899
##   2 | 334
##   2 | 5678
##   3 | 3
##   3 | 7
## 
## [1] "# Daily mortality count on Lowest Particulate Pollution Days"
## 
##   The decimal point is 1 digit(s) to the right of the |
## 
##   1 | 134
##   1 | 5666667899
##   2 | 0134
##   2 | 577
```

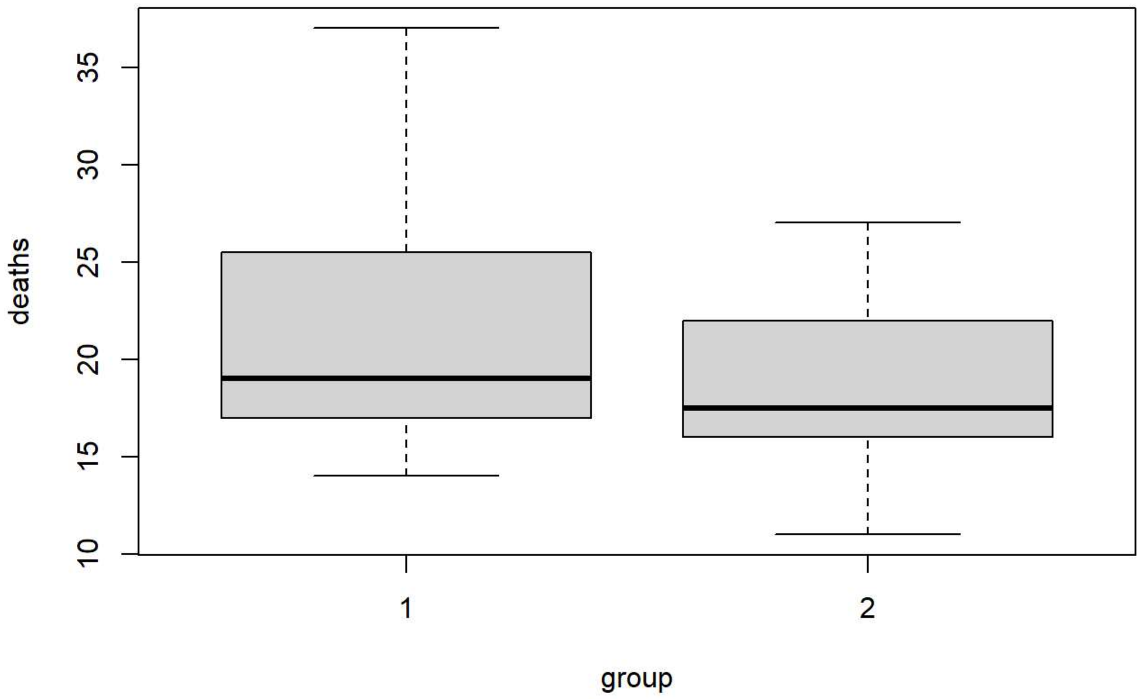
Note: R will decide the number of stems unless you control the scale. One can do this by specifying the number of lines.

```
for (pollut in c("1", "2")) {
    print(ifelse(pollut == "Highest", "# Daily mortality count on Highest Particulate Pollution Days",
        "# Daily mortality count on Lowest Particulate Pollution Days"))
    print("# Scale = 1")
    stem(filter(ps1, group == pollut)$deaths, scale = 1)
    print("# Scale = 3")
    stem(filter(ps1, group == pollut)$deaths, scale = 3)
}
```

```
## [1] "# Daily mortality count on Lowest Particulate Pollution Days"
## [1] "# Scale = 1"
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   1 | 44
##   1 | 567778899
##   2 | 334
##   2 | 5678
##   3 | 3
##   3 | 7
##
## [1] "# Scale = 3"
##
##   The decimal point is at the |
##
##   14 | 000
##   16 | 0000
##   18 | 0000
##   20 |
##   22 | 00
##   24 | 00
##   26 | 00
##   28 | 0
##   30 |
##   32 | 0
##   34 |
##   36 | 0
##
## [1] "# Daily mortality count on Lowest Particulate Pollution Days"
## [1] "# Scale = 1"
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   1 | 134
##   1 | 5666667899
##   2 | 0134
##   2 | 577
##
## [1] "# Scale = 3"
##
##   The decimal point is at the |
##
##   10 | 0
##   12 | 0
##   14 | 00
##   16 | 000000
##   18 | 000
##   20 | 00
##   22 | 0
##   24 | 00
##   26 | 00
```

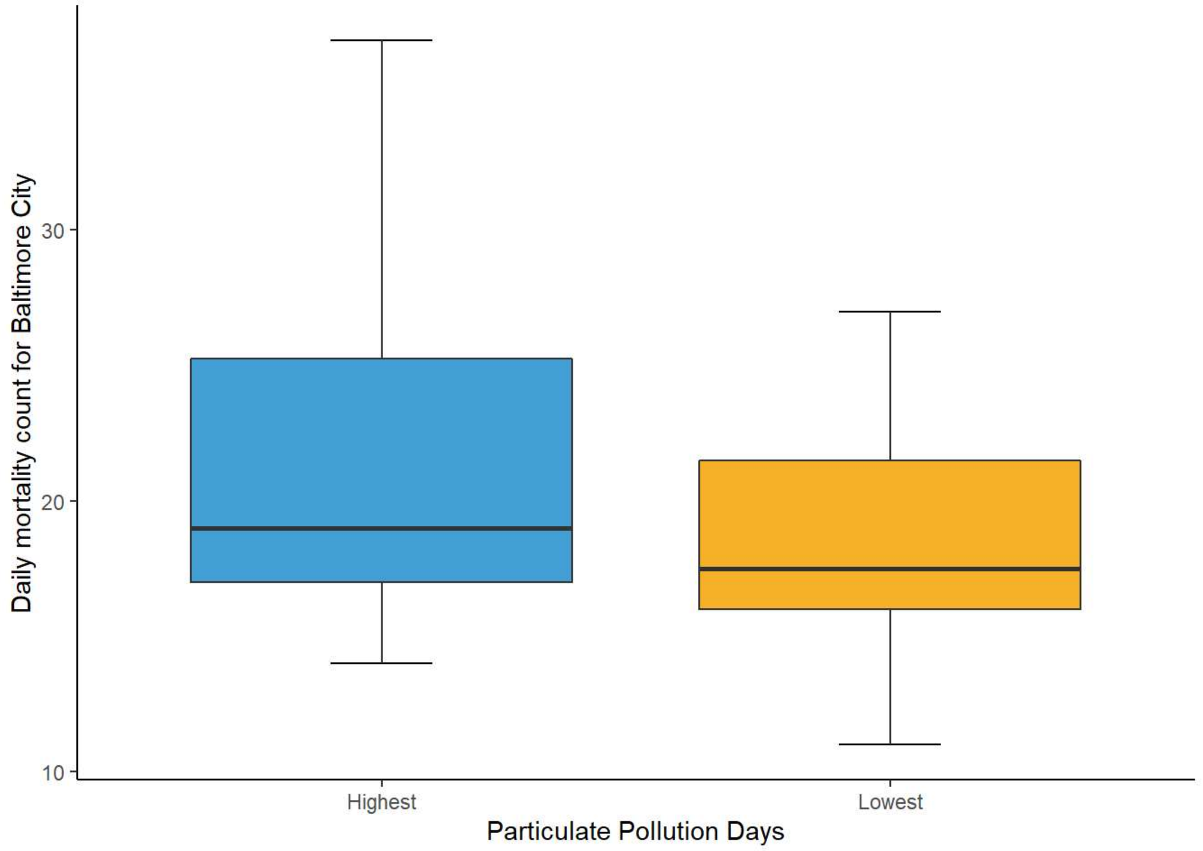
b. Create a box and whiskers plot by group:

```
boxplot(deaths ~ group, data = ps1)
```



```
ggplot(aes(x = group, y = deaths), data = ps1) + stat_boxplot(geom = "errorbar",
    width = 0.2) + geom_boxplot(aes(fill = group)) + xlab("Particulate Pollution Days") +
    ylab("Daily mortality count for Baltimore City") + scale_x_discrete(labels = c("Highest",
    "Lowest")) + scale_fill_manual(values = c("#439fd3", "#f6b128"), guide = FALSE) +
    theme_classic()
```

```
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please
## use `guide = "none"` instead.
```



c. Please check that your results for steps a. and b. give you the same results as you obtained by hand in Lab Exercise 1.

```
# Yes.
```

d. In a sentence or two, as if for a public health journal, summarize the evidence from these data relevant to whether mortality tends to be higher on high particulate pollution days, as compared to low particulate pollution days. Be quantitative, that is use numeric results, to describe your findings in the context of air pollution and mortality, the topic of greatest interest to your readers. Remember units!

```
# There is a tendency for lower death on lowest particulate pollution days, since the median daily mortali
ty count for Baltimore city is 1.5 person higher on high particulate pollution days, compared with low par
ticulate pollution days. Since the daily mortality counts on both high particulate pollution days and low
particulate pollution days are on an asymmetry distribution, it is reasonable to use median as a measureme
nt for central tendency to compare.
```

e. To assure reproducible research, save your script file. Please include the relevant parts of your code/output with your results as part of your write-up.

Problem 2. Costs of Carotid Endarterectomy in Maryland

Section 1: Exploratory Data Analysis

i. We have drawn a random sample of 100 male and 100 female patients from the HSCRC carotid endarterectomy data set which is located in the data file named ce621.csv on the course website. Refer to your *Class Dataset Code-Book* for the file format. Load the tidyverse with: `library(tidyverse)` **Open the data set and name it ce621. And, don't forget to open a script in order to save your work. Make stem and leaf plots of the male and female CE costs.**

```
# import data
ce621 <- read_csv("./data/ce621.csv")
```

```
## Rows: 200 Columns: 12
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): provnum, sex, race, volhosp, volsurg, dthstrk, smoker
## dbl (5): patid, ccscore, totchg, age, year
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ce621$provnum <- as.numeric(ce621$provnum)
ce621$sex <- factor(ce621$sex, levels = c("Male", "Female"))
ce621$race <- factor(ce621$race, levels = c("Caucas", "AfroAm", "Other"))
ce621$smoker <- factor(ce621$smoker, levels = c("No", "Yes"))
```

```
for (sex in c("Male", "Female")) {
    print(paste("# CE costs for ", sex, sep = ""))
    stem(filter(ce621, sex == sex)$totchg)
}
```

```
## [1] "# CE costs for Male"
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 29
##    2 | 5667788990002223344445666677778899999
##    4 | 01122233333344444444455555677778999000001112222223334444556666778999
##    6 | 001122233344455566667778999900122223459
##    8 | 35577889123555779
##   10 | 12524567889
##   12 | 2588908
##   14 | 57
##   16 | 177
##   18 | 12244489
##   20 |
##   22 |
##   24 | 0
##   26 |
##   28 |
##   30 | 6
##   32 | 2
##   34 |
##   36 |
##   38 |
##   40 | 1
##
## [1] "# CE costs for Female"
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 29
##    2 | 5667788990002223344445666677778899999
##    4 | 01122233333344444444455555677778999000001112222223334444556666778999
##    6 | 001122233344455566667778999900122223459
##    8 | 35577889123555779
##   10 | 12524567889
##   12 | 2588908
##   14 | 57
##   16 | 177
##   18 | 12244489
##   20 |
##   22 |
##   24 | 0
##   26 |
##   28 |
##   30 | 6
##   32 | 2
##   34 |
##   36 |
##   38 |
##   40 | 1
```

ii. Make the necessary calculations using R to complete the summary table below regarding CE costs by sex. To practice hand calculations, first verify that the mean, median, interquartile range (IQR), and standard deviation of the following sample of numbers: 1, 1, 1, 2, 2, 3, 5, 7, 11 are X = 3.67 ; median = 2; IQR = 4; and s = 3.43.

From the output generated from the above commands, complete the summary table by hand to summarize the typical value and variability of CE cost by sex.

```
# summary statistics
ce621.summary <- ce621 %>%
    group_by(sex) %>%
    summarise(Mean = mean(totchg, na.rm = TRUE), Median = median(totchg, na.rm = TRUE),
        IQR = quantile(totchg, probs = c(0.75), na.rm = TRUE) - quantile(totchg,
            probs = c(0.25), na.rm = TRUE), SD = sd(totchg, na.rm = TRUE))

# transform dataset to meet with the required format
ce621.summary <- ce621.summary %>%
    gather(Statistic, Value, c(Mean:SD)) %>%
    spread(sex, Value, c("Male", "Female")) %>%
    mutate(Statistic = factor(Statistic, levels = c("Mean", "Median", "IQR", "SD"))) %>%
    arrange(Statistic) %>%
    mutate(Male = round(as.numeric(Male), 2), Female = round(as.numeric(Female),
        2))
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
## Warning in if (!is.na(fill)) {: the condition has length > 1 and only the first
## element will be used
```

```
print(ce621.summary)
```

```
## # A tibble: 4 x 3
##   Statistic  Male Female
##   <fct>     <dbl>  <dbl>
## 1 Mean      6484.  8099.
## 2 Median    5598.  5644.
## 3 IQR       2862.  5187
## 4 SD        3278.  6679.
```

Statistic	Male	Female
Typical Value		
Mean	6484.04	8099.29
Median	5597.50	5643.50
Variability		
Interquartile Range (IQR)	2861.50	5187.00
Standard Deviation (s)	3277.94	6679.28

iii. In a brief paragraph, as if written for a health services research journal, compare the distribution of costs for men and women. Be quantitative by referring to the average cost, variability, and the shape of the distribution.

```
# The average cost of women is 1615.25 dollars higher than men, and the standard deviation of cost of fema
le is 3401.34 dollars higher than men, indicating women have a higher costs and higher variability. The sh
apes of the distribution of costs of both men and women are right skewed. Since the distribution of costs
of both men and women are asymmetry, it may be more appropriate to use median and interquartile range to c
onduct summary statistics. The median cost of female is slightly higher than male (46.00 dollars), while t
he interquartile range of women is 2325.50 dollars higher than men, indicating a similiar (slightly highe
r) central cost and large variability of women compared with men.
```

iv. Now use the entire CE data file named `CE621entire.csv` . Open the data set and name it `ce621e` . Choose only one year. Create an age category variable where age is categorized as: ≤ 50 years; 51-64; 65+.

Display side-by-side six box plots of the six sex-by-age strata (≤ 50-male, female; 51-64- male, female; 65+-male, female). Label the strata on the boxplot and order the box plots to facilitate the sex comparison within age-group.

```
# import dataset
ce621e <- read_csv("./data/ce621entire.csv")
```

```
## Rows: 9918 Columns: 8
```

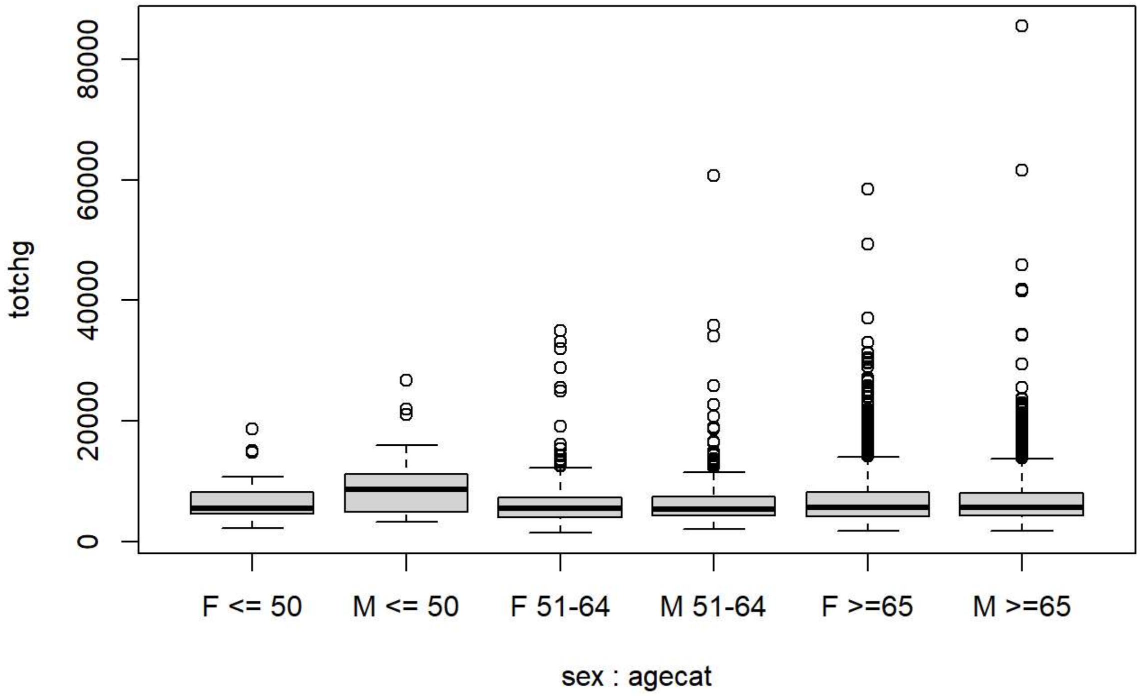
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): provnum, sex, race, smoker
## dbl (4): patid, totchg, age, year
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ce621e$provnum <- as.numeric(ce621e$provnum)
ce621e$sex <- factor(ce621e$sex, levels = c("Male", "Female"))
ce621e$race <- factor(ce621e$race, levels = c("Caucas", "AfroAm", "Other"))
ce621e$smoker <- factor(ce621e$smoker, levels = c("No", "Yes"))
```

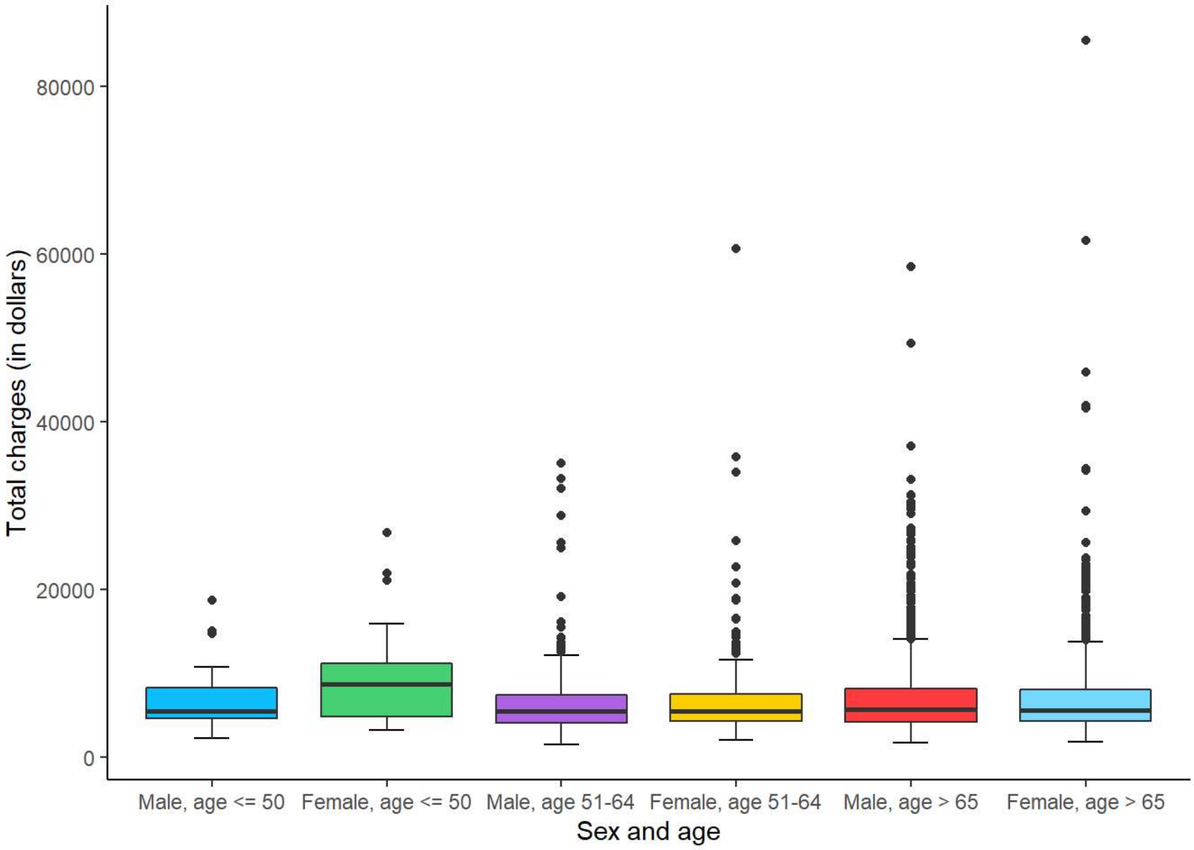
```
# > max(ce621e$age)
# [1] 92

# Categorize age and sex into composite factor variable
ce621e <- ce621e %>%
              filter(year == 1995) %>% # choose year 1995
                  mutate(agecat = cut(age, breaks = c(0, 50, 64, 92))) %>%
                      mutate(agesex = case_when(agecat == "(0,50]" & sex == "Male" ~ "1",
                                                agecat == "(0,50]" & sex == "Female" ~ "2",
                                                agecat == "(50,64]" & sex == "Male" ~ "3",
                                                agecat == "(50,64]" & sex == "Female" ~ "4",
                                                agecat == "(64,92]" & sex == "Male" ~ "5",
                                                agecat == "(64,92]" & sex == "Female" ~ "6")) %>%
                          mutate(agesex = factor(agesex, levels = c("1", "2", "3", "4", "5", "6"),
                                                         labels = c("M <= 50", "F <= 50", "M 51-64",
                                                                    "F 51-64", "M >= 65", "F >= 65")))
```

```
boxplot(totchg ~ sex + agecat, data = ce621e, names = c("F <= 50", "M <= 50", "F 51-64",
    "M 51-64", "F >=65", "M >=65"))
```



```
ggplot(aes(x = agesex, y = totchg), data = ce621e) + stat_boxplot(geom = "errorbar",
    width = 0.2) + geom_boxplot(aes(fill = agesex)) + xlab("Sex and age") + ylab("Total charges (in dollar
s)") +
    scale_x_discrete(labels = c("Male, age <= 50", "Female, age <= 50", "Male, age 51-64",
        "Female, age 51-64", "Male, age > 65", "Female, age > 65")) + scale_fill_manual(values = c("#0ebef
f",
    "#47cf73", "#ae63e4", "#fcd000", "#ff3c41", "#76daff"), guide = FALSE) + theme_classic()
```

```
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please
## use `guide = "none"` instead.
```

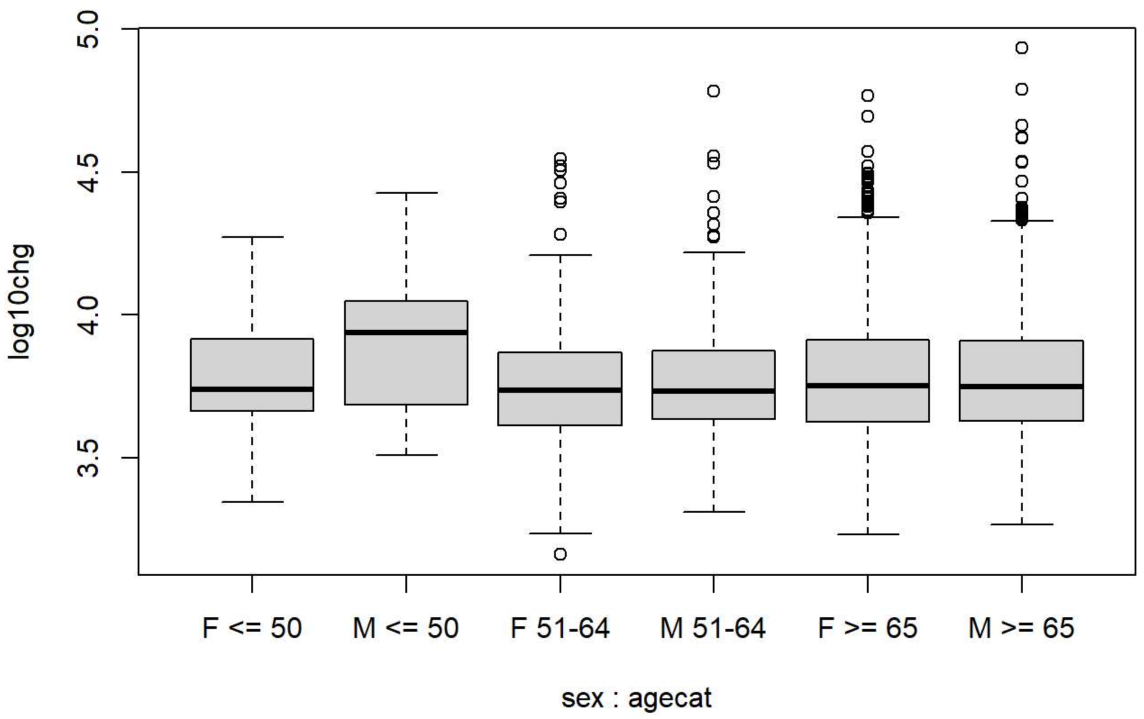


v. Write a sentence or two that describe how the CE cost distribution differs for men and women within each of the three age strata. From an equivalent but alternate perspective, describe how the distribution varies across age separately for men and women.

```
# In the year 1995, in all age and sex groups, the distributions of total charges in dollars are right ske
wed.
# As for the distribution differs for men and women within each of the three age strata, among age <= 50 y
ears, both median and interquartile range of men are lower than women, indicating a lower charges among me
n and lower variability. While among age group 51-64 and 65+, both median and interquartile range of men a
re similar (slightly higher) than women, indicating a similar (slightly higher) charges and variability am
ong men.
# As for the distribution various across age separately for men and women, among men, the median and inter
quartile range between each age group are similiar. Among women, age group 51-64 has the lowest median and
iterquartile range, follow by age group 65+ and <= 50, indicating the total charges with variability among
age groups follow a sequence of 51-64, 65+, <= 50, from small to large. In both men and women, with age in
creases, the range increases in each age group.
```

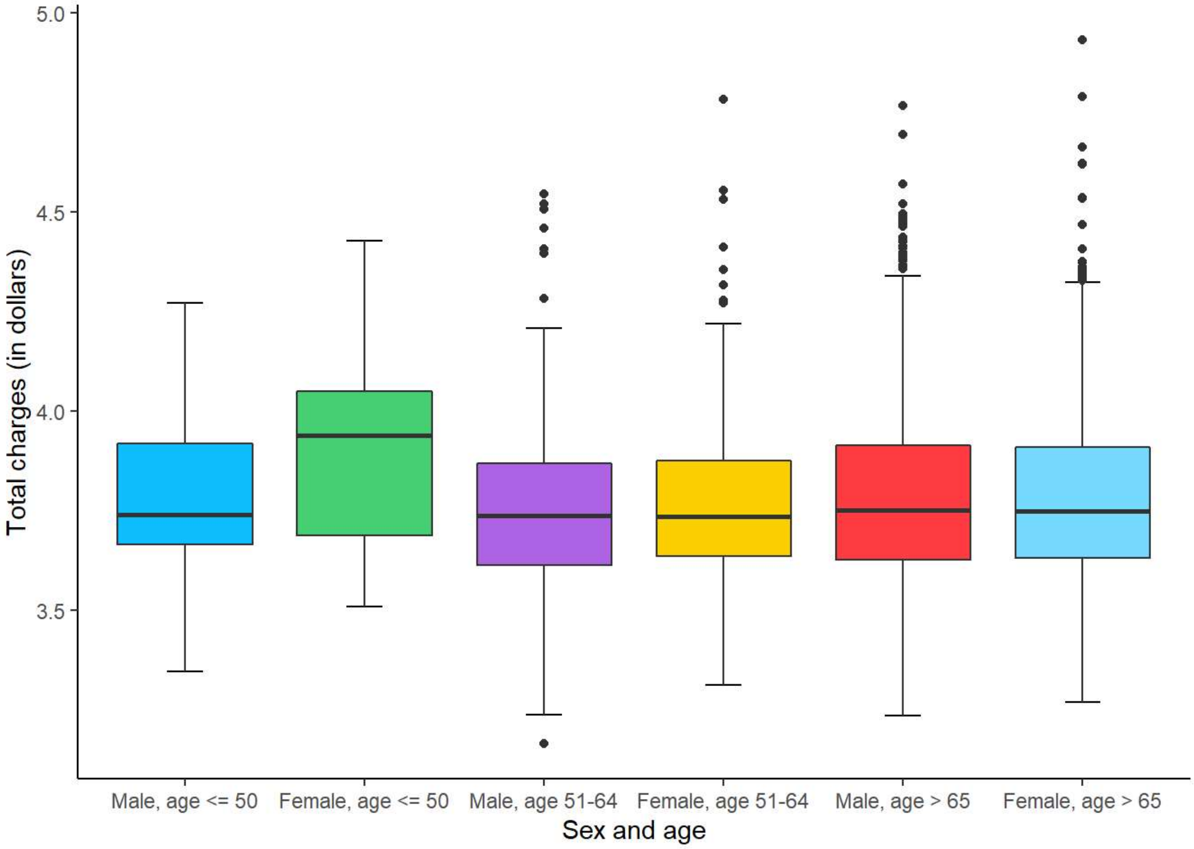
vi. Now repeat Step iv. using a logarithmic (base 10) scale for dollars. Specifically generate a new variable with log(base 10) dollars. How does this display compare with the previous one using the original scale? What additional patterns are made apparent?

```
ce621e <- ce621e %>%
    mutate(log10chg = log10(totchg))

boxplot(log10chg ~ sex + agecat, data = ce621e, names = c("F <= 50", "M <= 50", "F 51-64",
     "M 51-64", "F >= 65", "M >= 65"))
```



```
ggplot(aes(x = agesex, y = log10chg), data = ce621e) + stat_boxplot(geom = "errorbar",
    width = 0.2) + geom_boxplot(aes(fill = agesex)) + xlab("Sex and age") + ylab("Total charges (in dollar
s)") +
    scale_x_discrete(labels = c("Male, age <= 50", "Female, age <= 50", "Male, age 51-64",
        "Female, age 51-64", "Male, age > 65", "Female, age > 65")) + scale_fill_manual(values = c("#0ebef
f",
    "#47cf73", "#ae63e4", "#fcd000", "#ff3c41", "#76daff"), guide = FALSE) + theme_classic()
```

```
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please
## use `guide = "none"` instead.
```



```
# When using arithmetic scale for boxplot, the differences in central tendency and variability were not ap
prant due to outliers larger than the upper fence. When using logarithmic transformation, the differences
in median and variability in each age and sex group are made relatively more apparent compared with privou
s boxplot, since it allows plotting of numbers of different orders of magnitude on the same graph. The dis
tribution of total dollars change to relatively more symmetric in each age and sex group after logarithmic
transformation.
```

vii. To assure reproducible research, save your script file. Please include the relevant parts of your code/output with your results as part of your write-up.