Problem Set Week 1

Francesca Lionetti

15/01/2022

Directions for the student

- Put all R code in code chunks and verbal answers outside code chunks.
- If you get a piece of R code to work, set the code chunk option eval=FALSE to ensure the document can still be knitted.
- Use tidyverse functions whenever possible.
- Comment your code to communicate your intentions.
- Ensure that the R Markdown document knits without problems into a PDF or Word document.
- Submit the R Markdown document on Canvas (under Assignments) before the deadline.

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## Warning: il pacchetto 'tidyverse' è stato creato con R versione 4.1.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2

## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.1.3

## Warning: il pacchetto 'tibble' è stato creato con R versione 4.1.3

## Warning: il pacchetto 'tidyr' è stato creato con R versione 4.1.3

## Warning: il pacchetto 'readr' è stato creato con R versione 4.1.3

## Warning: il pacchetto 'purrr' è stato creato con R versione 4.1.3

## Warning: il pacchetto 'dplyr' è stato creato con R versione 4.1.3

## Warning: il pacchetto 'stringr' è stato creato con R versione 4.1.3

## Warning: il pacchetto 'forcats' è stato creato con R versione 4.1.3

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

Data

applications.csv. Admission or rejection (variable *admit*: 1 = Admitted, 0 = Rejected) of applicants to different departments (variable *dept*) of a graduate school with applicant's sex (variable *gender*) and application date (variable *date*).

Questions

1. Import the data in R with `readr::` and determine the overall admission rate (proportion of admitted applicants) using R code.

```
#Add your R code for answering this question here.
Data<-read_csv("~/applications (3).csv")
```

```
## Rows: 4526 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): gender, dept
## dbl  (1): admit
## date (1): date
##
## 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.
```

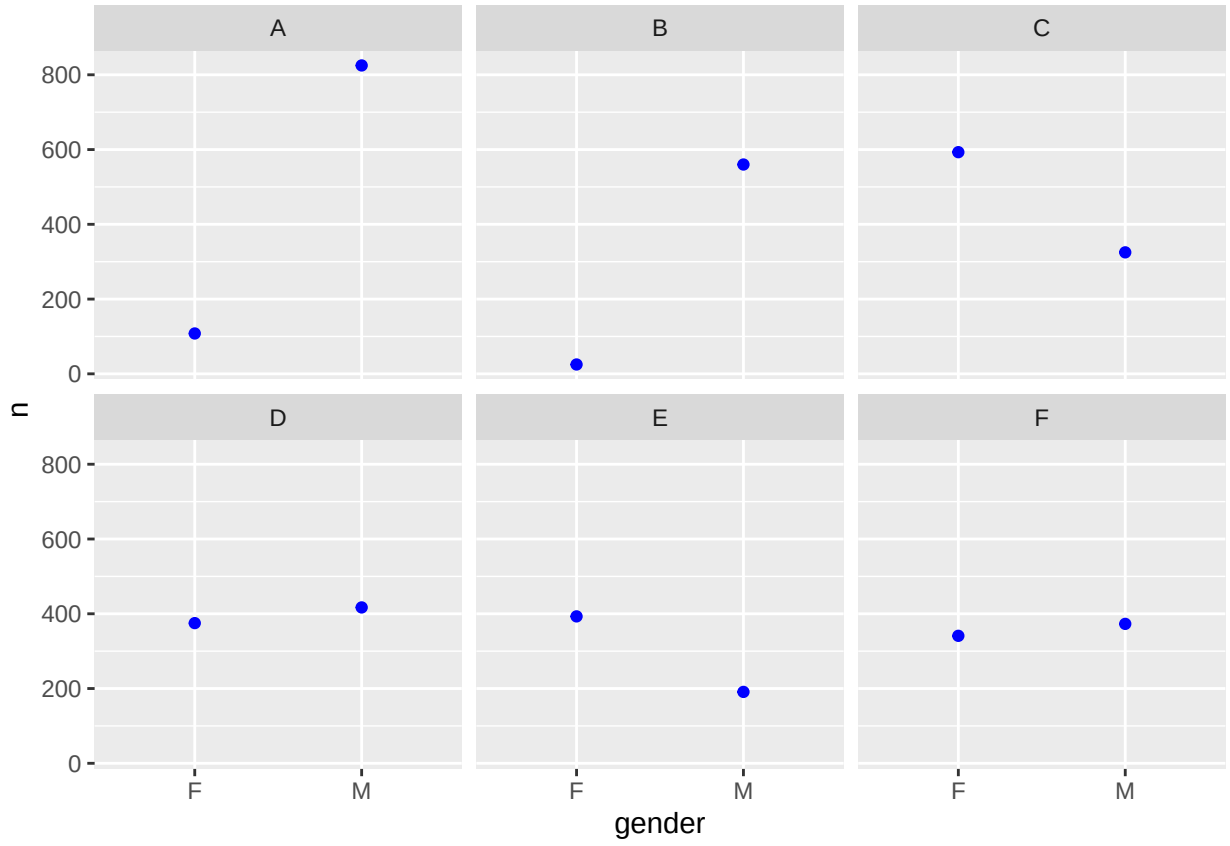
```
Data %>%
  group_by(admit)%>%
  count(admit)%>%
  ungroup()%>%
  mutate(perc = n /sum(n))
```

```
## # A tibble: 2 x 3
##   admit     n  perc
##   <dbl> <int> <dbl>
## 1     0  2771 0.612
## 2     1  1755 0.388
```

Grading	Max points	Awarded
Ex. 1	1	

2. Use `ggplot2::` to create a figure showing the relation between department and gender; set all colors to blue. Which departments are relatively popular among male applicants in comparison to female applicants?

```
Data %>%
  group_by(dept, gender) %>%
  count(gender)%>%
  ggplot( aes(x=gender, y=n, group=dept))+
       geom_point(color="blue")+
         facet_wrap(~dept)
```



Grading	Max points	Awarded
Ex. 2	1	

3. With `dplyr::` calculate the admission rate for each department for female applicants and for male applicants.

```
#Add your R code for answering this question here.
Data %>%
  group_by(dept, gender) %>%
  count(admit) %>%
  group_by(dept) %>%
  mutate(perc = n /sum(n))
```

```
## # A tibble: 24 x 5
## # Groups:   dept [6]
##    dept  gender admit     n   perc
##    <chr> <chr>  <dbl> <int>  <dbl>
##  1 A     F          0    19 0.0204
##  2 A     F          1    89 0.0954
##  3 A     M          0   313 0.335
##  4 A     M          1   512 0.549
##  5 B     F          0     8 0.0137
##  6 B     F          1    17 0.0291
##  7 B     M          0   207 0.354
```

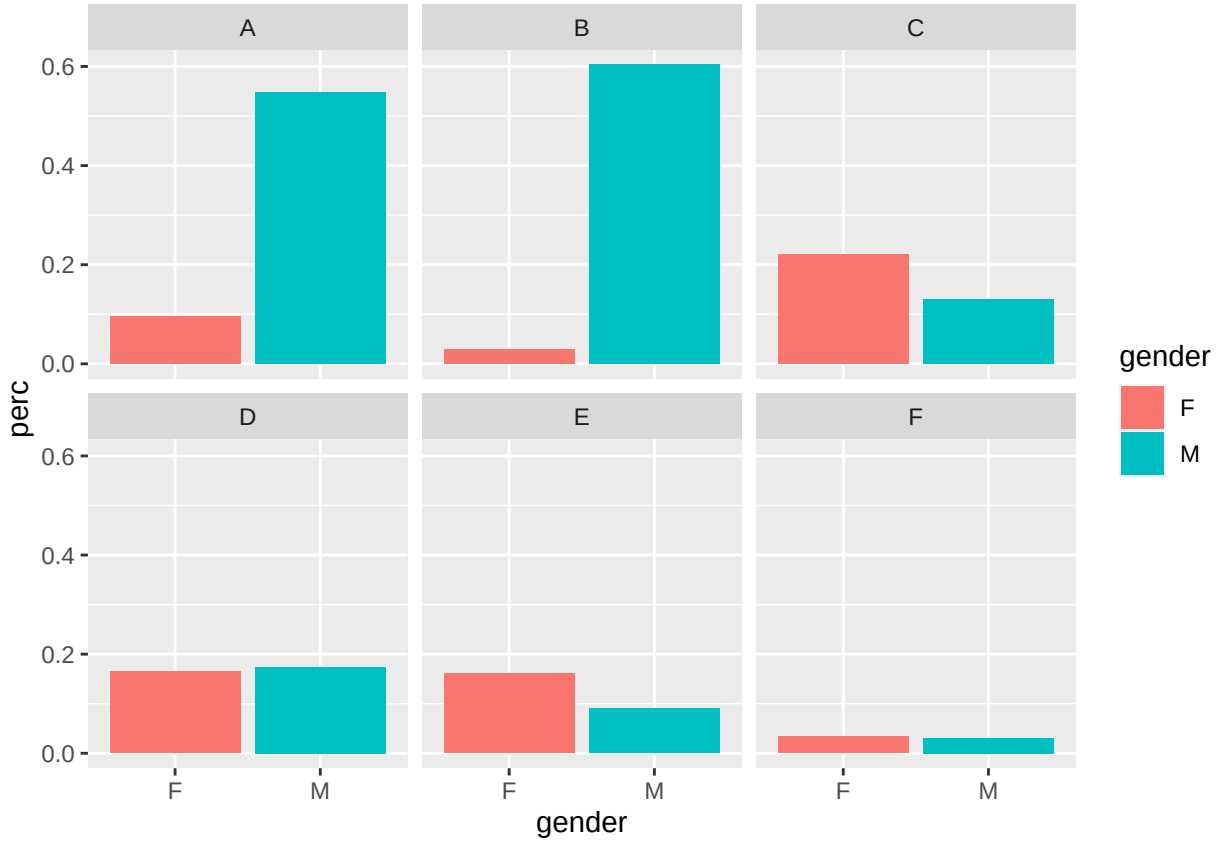
```
##  8 B     M          1   353 0.603
##  9 C     F          0   391 0.426
## 10 C     F          1   202 0.220
## # ... with 14 more rows
```

Grading	Max points	Awarded
Ex. 3	1	

4. Display your answer to Question 3 as a bar chart which makes it easy to compare female and male applicants within each department. Is the admission rate always higher for females or for males?

```
#Add your R code for answering this question here.
data_3 = Data %>%
  group_by(dept, gender) %>%
  count(admit) %>%
  group_by(dept) %>%
  mutate(perc = n /sum(n))

ggplot(data_3 %>% filter(admit==1), aes(x=gender, y=perc, group=dept, fill=gender))+
  geom_col()+
  facet_wrap(~dept)
```



```
"Not always"
```

```
## [1] "Not always"
```

```
#admission rate is higher for males in department A,B, D. While it is higher for females in C, E, F
```

Grading	Max points	Awarded
Ex. 4	2	

5. For each department, calculate the cumulative mean of admissions (variable *admit*) over time. Show the first 8 results for Department B. What do the cumulative mean scores mean here?

```
#Add your R code for answering this question here.
Data %>%
  group_by(dept, date)%>%
  count(admit)%>%
  group_by(dept)%>%
  mutate(n = cumsum(n))%>%
  group_by(dept, date)%>%
  mutate(perc = n /sum(n))%>%
  filter(admit==1) %>%
  filter(dept=="B")%>%
  head(8)
```

```
## # A tibble: 8 x 5
## # Groups:   dept, date [8]
##   dept  date       admit     n  perc
##   <chr> <date>     <dbl> <int> <dbl>
## 1 B     1975-03-03     1     3 0.75
## 2 B     1975-03-04     1    11 0.688
## 3 B     1975-03-05     1    21 0.583
## 4 B     1975-03-06     1    27 0.551
## 5 B     1975-03-07     1    37 0.529
## 6 B     1975-03-08     1    43 0.518
## 7 B     1975-03-09     1    53 0.53
## 8 B     1975-03-10     1    62 0.525
```

```
#they show the average admission rate per day
```

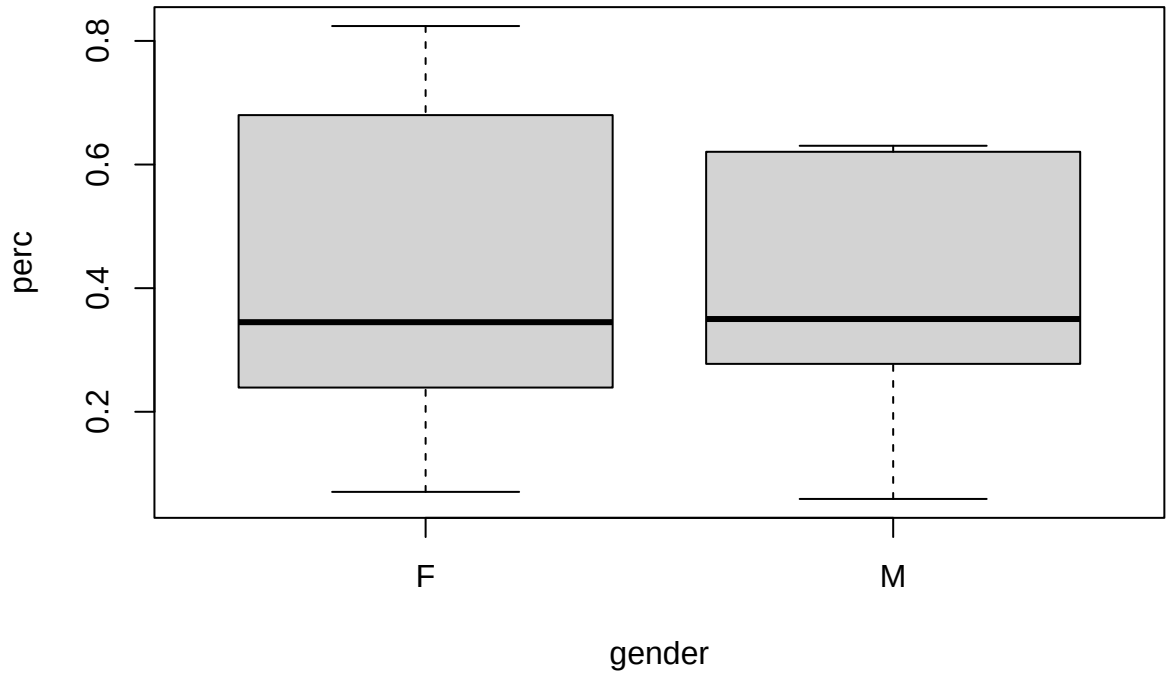
Grading	Max points	Awarded
Ex. 5	2	

6. Create a plot to explore the covariation (association) between a continuous variable and a categorical variable in *your project's data set* that are of interest to you. Use comments in the R code to justify the choices that you made to create this plot.

Note: You may use grouped summaries (aggregation) to create continuous variables or group continuous variables to obtain a categorical variable.

```r
data_box <- Data %>%
  group_by(admit, gender, dept)%>%
  count(admit)%>%
  ungroup()%>%
  group_by(gender, dept)%>%
  mutate(perc = n /sum(n))%>%
  filter(admit==1)


boxplot(perc~gender, data_box)
```



Grading	Max points	Awarded
Ex. 6	2	
Flawless knitting	1	
Total	10	