Data Visualization Project

PSYC 3031 A

Jasleen Ghuman (216278962) & Anthony Griffo (216380610)

November 4th, 2020

a. **A general description of the data :**

The data in this data set describes the admissions acceptance of students based on their sex in the six largest departments at the University of California, Berkeley in 1973. There were a total of 12,763 applicants used in the study. The goal was to examine if there was discrimination of acceptance based on gender.

b. Your visualization question :

Is there a sex bias with admissions in the six largest departments for graduate studies at University of California, Berkeley?

c. A description of the goals/outcomes of your visualization (i.e., what information do you want to communicate with your data):

Our goal is to showcase if there is a sex bias in admissions at the University of California, Berkeley. It is safe to say that sex plays no role on any of the qualifications relevant to gain acceptance. We will communicate this by making a series of graphs comparing gender to admission rates across the 6 largest departments at University of California, Berkeley. As we are comparing two categorical variables (Admit, Gender) and one continuous (Frequency), the best way to graph each individual department is with a clustered bar graph.

We will be creating graphs that will showcase the total number of females and male applicants that were rejected versus accepted at UCSB's 6 largest departments. We will create clustered bar

graphs to showcase the difference between males and females based on their frequency of admitted or rejected applicants. These graphs will display if there is a difference between the frequency of admitted versus rejected for each genders in each department. We will also be focusing on the number of applicants for each gender in each department. The number of female and male applicants is important to consider because this could also be a reason to why there is more of one gender being admitted/rejected. If this is the case then we will combine all 6 departments together by creating boxplots comparing each female and males admitted or rejected frequencies.

d. Your R code with sufficient comment for anyone to replicate and fully understand your code. These comments can include information about functionality but also any technical aspects you may need to consider :

UCSB Admissions.R

Jasleen Ghuman & Anthony Griffo

2020-11-04

```
#load appropriate packages
#load psyc package for descriptive stats
library(psych)
#load tidyverse for importing, cleaning and plotting data
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────
──────── tidyverse 1.3.0 ──

## ✓ggplot2 3.3.2     ✓purrr   0.3.4
## ✓tibble  3.0.3     ✓dplyr   1.0.2
## ✓tidyr   1.1.2     ✓stringr 1.4.0
## ✓readr   1.3.1     ✓forcats 0.5.0

## ── Conflicts ────────────────────────────────────────────────────────
─── tidyverse_conflicts() ──
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
#load here package to call for data from appropriate files
library(here)
```

```
## here() starts at /Users/jasleen2000/Desktop/R Projects/data visualization
project/UCSB admissions
```

```
#use "here" function to import data file. The "here" function will tell R whe
re our data file is located in our R folder
#create a dataframe named "ucsb_admissions" that will carry the original data
from the csv document we import
#the read_csv function is from the readr package that is installed when you i
nstall tidyverse. "read_csv" will be the function that actually imports the d
ata file into R
```

```
ucsb_admissions <- read_csv(file = here ("data", "ucsbadmissions.csv"))

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   Admit = col_character(),
##   Gender = col_character(),
##   Dept = col_character(),
##   Freq = col_double()
## )

#once "ucsb_admissions" dataframe is created, check to see what R thinks each
type of variable is
#spec function will go through each varaible in data frame to tell you what e
ach varaible is imported as

spec(ucsb_admissions)

## cols(
##   X1 = col_double(),
##   Admit = col_character(),
##   Gender = col_character(),
##   Dept = col_character(),
##   Freq = col_double()
## )

#its important to see if there was any problems importing the data and we can
use the "problems" function to check this
problems(ucsb_admissions)

## [1] row      col      expected actual
## <0 rows> (or 0-length row.names)

#specify appropriate variables for each character variable
#change the levels of the variables to meaningful values
#the mutate function will allow us to transform variables
#select the "ucsb_admissions" data frame to use mutate function on each chara
cter variable
#"fct_recode" will take each level of the variable and transform it to intege
rs
#specifiy appropriate variable for "Admit" and transform each level to an int
eger value
ucsb_admissions <- ucsb_admissions %>%
  mutate(Admit = fct_recode (Admit,
                             "Admitted" = "1",
                             "Rejected" = "0"))
```

```
## Warning: Problem with `mutate()` input `Admit`.
## ℹ Unknown levels in `f`: 1, 0
## ℹ Input `Admit` is `fct_recode(Admit, Admitted = "1", Rejected = "0")`.

## Warning: Unknown levels in `f`: 1, 0
```

```
#specifiy appropriate variable for "Gender" and transform each level to an in
teger value
ucsb_admissions <- ucsb_admissions %>%
  mutate(Gender = fct_recode(Gender,
                             "Male" = "0",
                             "Female" = "1"))
```

```
## Warning: Problem with `mutate()` input `Gender`.
## ℹ Unknown levels in `f`: 0, 1
## ℹ Input `Gender` is `fct_recode(Gender, Male = "0", Female = "1")`.

## Warning: Unknown levels in `f`: 0, 1
```

```
#specifiy appropriate variable for "Dept" and transform each level to an inte
ger value
ucsb_admissions <- ucsb_admissions %>%
  mutate(Dept = fct_recode(Dept,
                           "A" = "1",
                           "B" = "2",
                           "C" = "3",
                           "D" = "4",
                           "E" = "5",
                           "F" = "6"))
```

```
## Warning: Problem with `mutate()` input `Dept`.
## ℹ Unknown levels in `f`: 1, 2, 3, 4, 5, 6
## ℹ Input `Dept` is `fct_recode(...)`.

## Warning: Unknown levels in `f`: 1, 2, 3, 4, 5, 6
```

```
#generate descriptives for categorical variables
#use the summary function to check descriptives for categorical variables
#use select function to tell R which data frame you are using and the appropr
iate variables you want decriptives on
summary(select(ucsb_admissions, Admit, Gender, Dept))
```

```
##       Admit       Gender   Dept
##  Admitted:12   Female:12   A:4
##  Rejected:12   Male  :12   B:4
##                            C:4
##                            D:4
##                            E:4
##                            F:4
```

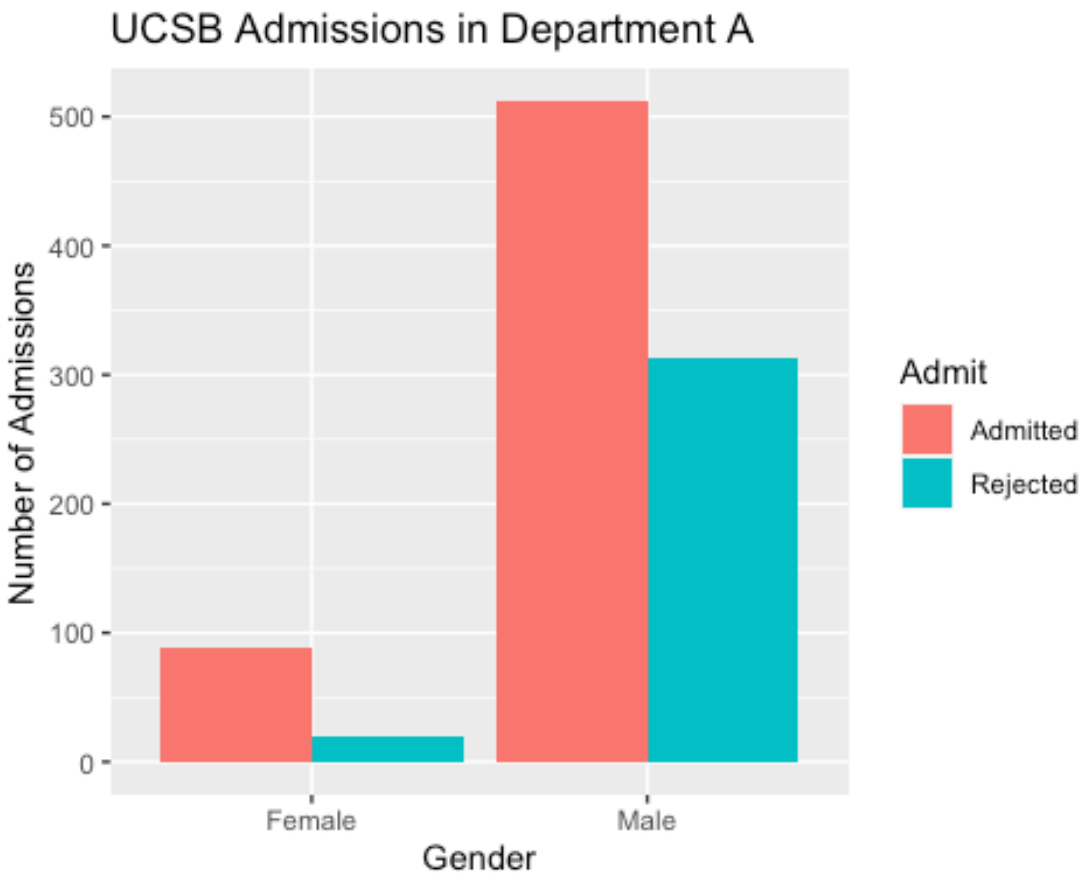
```
#generate descriptives for continous variables
#use the describe function to check descriptives for continuous variables
```

```r
#use select function to tell R which data frame you are using and the appropr
iate variables you want decriptives on
describe(select(ucsb_admissions, X1, Freq))
```

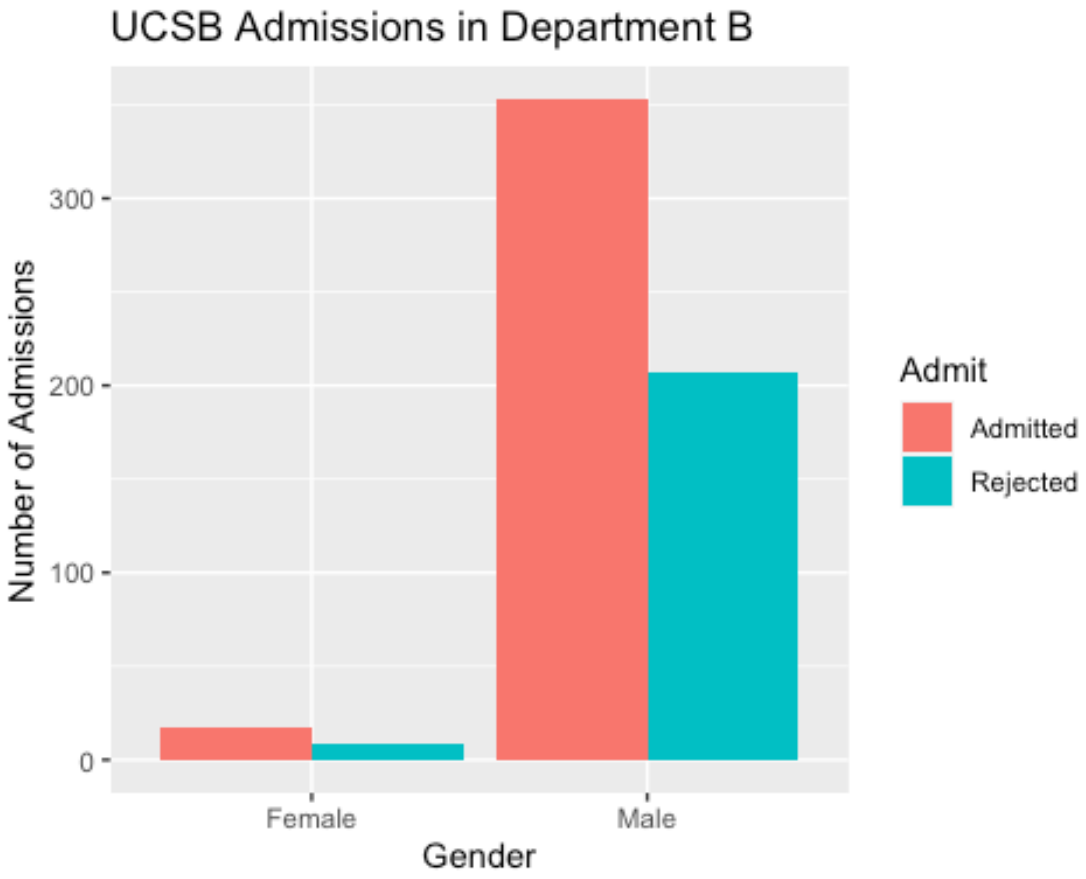
```
##      vars  n   mean     sd median trimmed    mad min max range skew kurtos
is
## X1      1 24  12.50   7.07   12.5    12.5   8.90   1  24    23 0.00    -1.
35
## Freq    2 24 188.58 140.06  170.0   179.9 182.36   8 512   504 0.41    -0.
88
##         se
## X1    1.44
## Freq 28.59
```

```r
#create graphs to interpret the "ucsb_admissions" data frame


#we want to interpret if there is a sex bias in each department at UCSB
#from original dataframe, we have to subset rows using "filter" function to c
reate seperate dataframes for each department
#the seperate data frames for each department created by the filter function
will allow us to create appropriate graphs comparing the amount of females an
d males admitted versus rejected within each department
#after creating each dataframe for the 6 departments, make graphs for each
#create a clustered bar graph for males versus females in each department tha
t were accepted or rejected
#use the ggplot function from the "tidverse" package to create the clustered
bar graphs
#select the appropriate data frame for each graph and use the "aes" function
that will allow you to pick what variables you want to use from your datafram
e on the x, y axis and the fill for the bars
#use "geom_bar" to create a clustered bar graph
#since we want to use two categorical variables (gender and admit) we have to
create a clustered bar graph
#clustered bar graphs can be created by adjusting position in geom_bar
#adjust position to "position = position_dodge()" as this will allow for a si
de by side comparison for each level of "Admit" for each Gender
#use "labs" function to rename x axis, y axis, title and legend title

# create data frame for department A by subsetting rows from "ucsb_admissions
"
ucsb_admissions_depta <- filter(ucsb_admissions, Dept == "A")

#create graph for dept A
ggplot(ucsb_admissions_depta, aes (x = Gender, y = Freq)) +
  geom_bar(
    aes(Admit = Admit, fill = Admit),
    stat = "identity" , position = position_dodge()
      )+
```

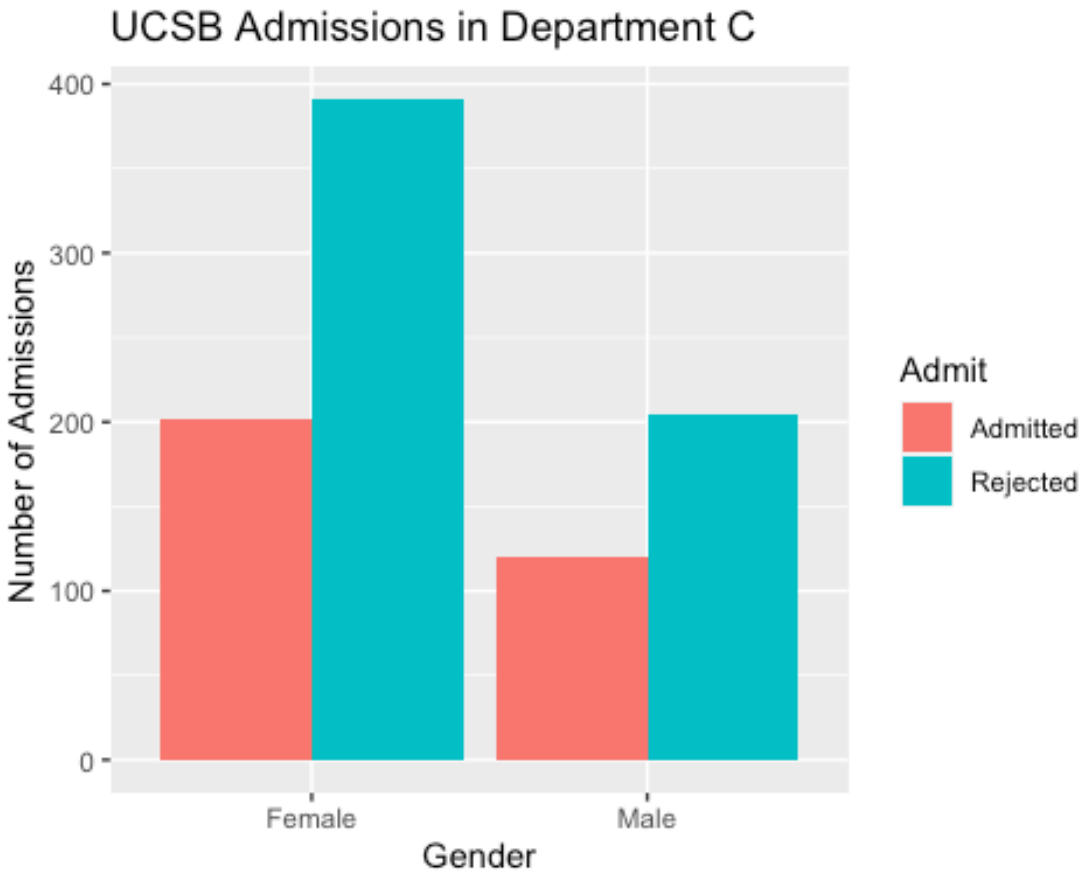
```
  labs(title = "UCSB Admissions in Department A", x = "Gender", y = "Number o
f Admissions")

## Warning: Ignoring unknown aesthetics: Admit
```



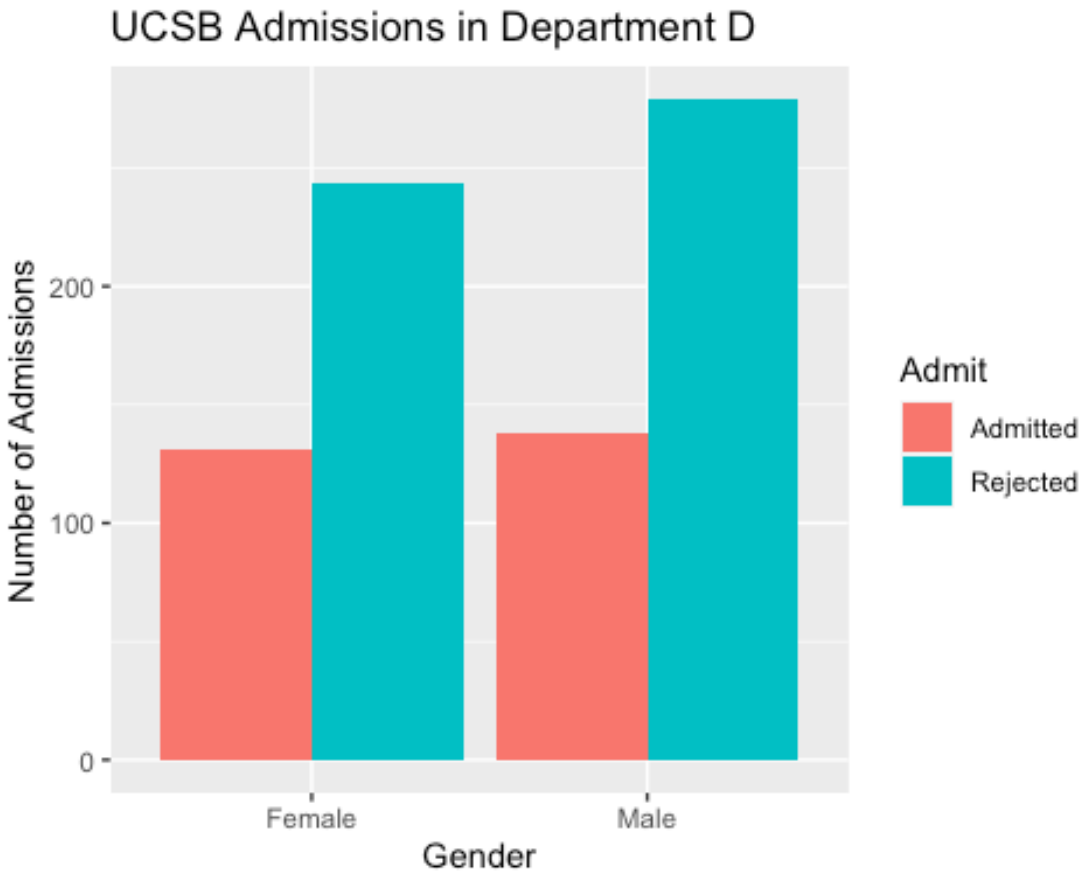

```
#create data frame for dept B
ucsb_admissions_deptb <- filter(ucsb_admissions, Dept == "B")

#create graph for dept B
ggplot(ucsb_admissions_deptb, aes(x = Gender, y = Freq)) +
  geom_bar(
    aes(Admit = Admit, fill = Admit),
    stat = "identity", position = position_dodge()
  )+
  labs(title = "UCSB Admissions in Department B", x = "Gender", y = "Number o
f Admissions")

## Warning: Ignoring unknown aesthetics: Admit
```



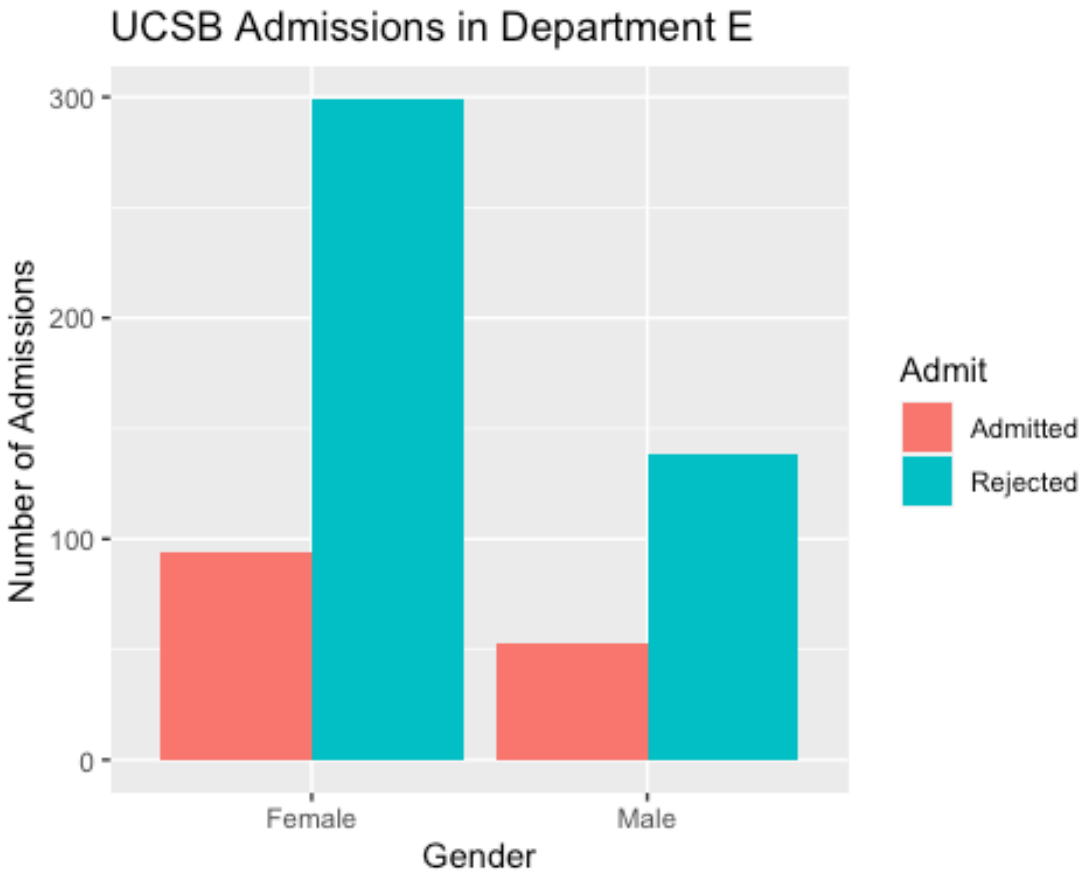

```
#create data frame for dept C
ucsb_admissions_deptc <- filter(ucsb_admissions, Dept == "C")

#create graph for dept C
ggplot(ucsb_admissions_deptc, aes(x = Gender, y = Freq)) +
  geom_bar(
    aes(Admit = Admit, fill = Admit),
    stat = "identity", position = position_dodge()
  )+
  labs(title = "UCSB Admissions in Department C", x = "Gender", y = "Number o
f Admissions")

## Warning: Ignoring unknown aesthetics: Admit
```



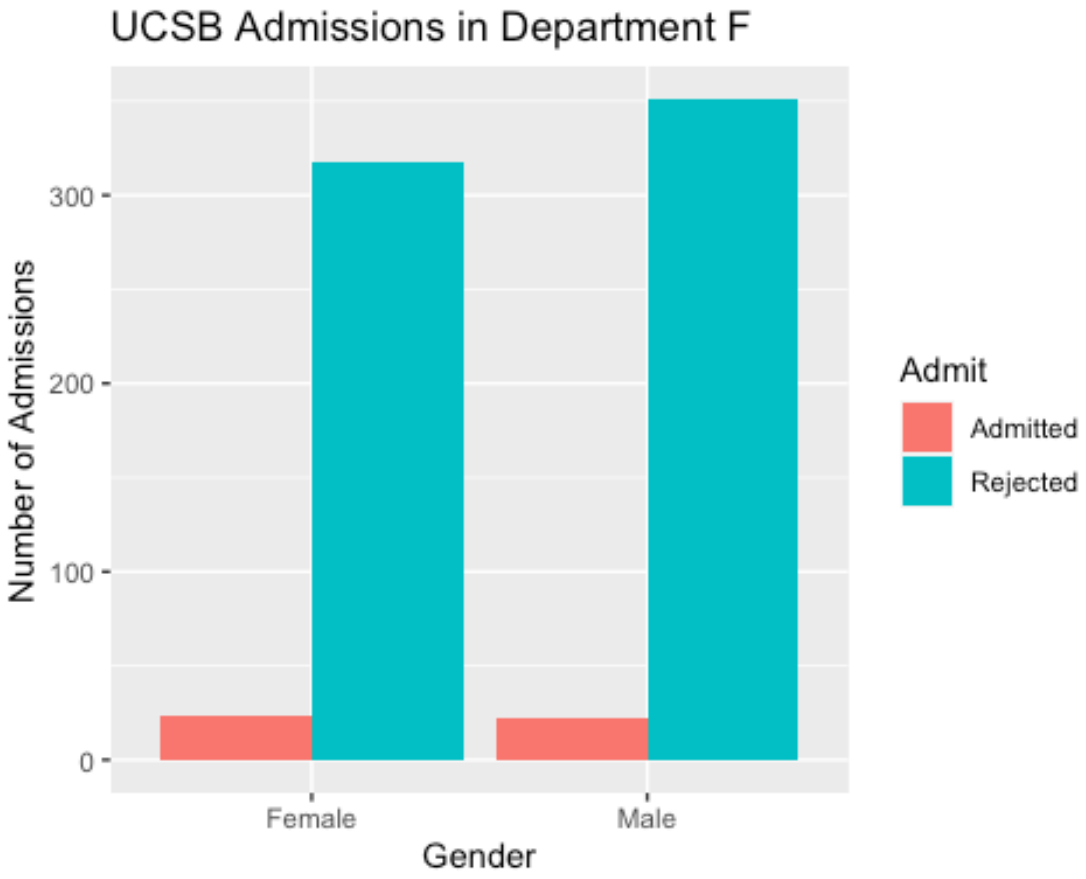

```
#create data frame for dept D
ucsb_admissions_deptd <- filter(ucsb_admissions, Dept == "D")

#create graph for dept D
ggplot(ucsb_admissions_deptd, aes(x = Gender, y = Freq)) +
  geom_bar(
    aes(Admit = Admit, fill = Admit),
    stat = "identity", position = position_dodge()
  )+
  labs(title = "UCSB Admissions in Department D", x = "Gender", y = "Number o
f Admissions")

## Warning: Ignoring unknown aesthetics: Admit
```




```
#create data frame for dept E
ucsb_admissions_depte <- filter(ucsb_admissions, Dept == "E")

#create graph for dept E
ggplot(ucsb_admissions_depte, aes(x = Gender, y = Freq)) +
  geom_bar(
    aes(Admit = Admit, fill = Admit),
    stat = "identity", position = position_dodge()
  )+
  labs(title = "UCSB Admissions in Department E", x = "Gender", y = "Number o
f Admissions")

## Warning: Ignoring unknown aesthetics: Admit
```

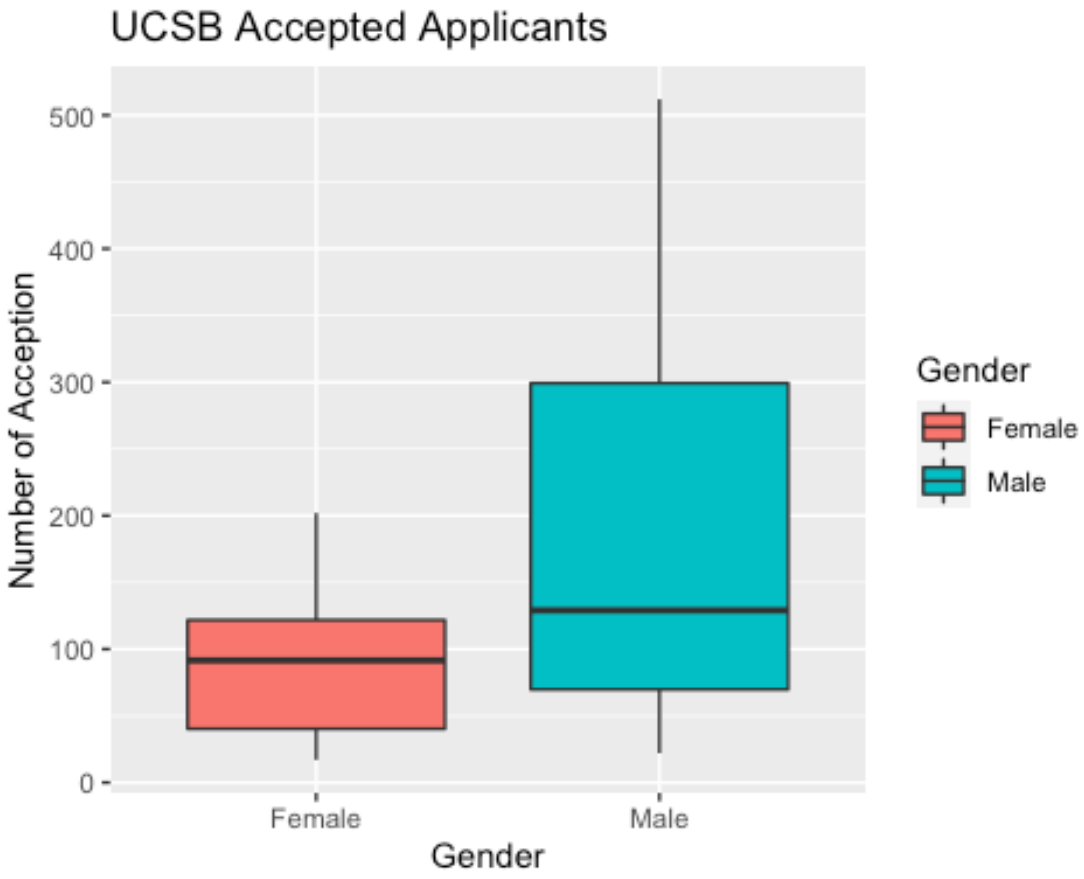



```
#create data frame for dept F
ucsb_admissions_deptf <- filter(ucsb_admissions, Dept == "F")

#create graph for dept F
ggplot(ucsb_admissions_deptf, aes(x = Gender, y = Freq)) +
  geom_bar(
    aes(Admit = Admit, fill = Admit),
    stat = "identity", position = position_dodge()
  )+
  labs(title = "UCSB Admissions in Department F", x = "Gender", y = "Number o
f Admissions")

## Warning: Ignoring unknown aesthetics: Admit
```



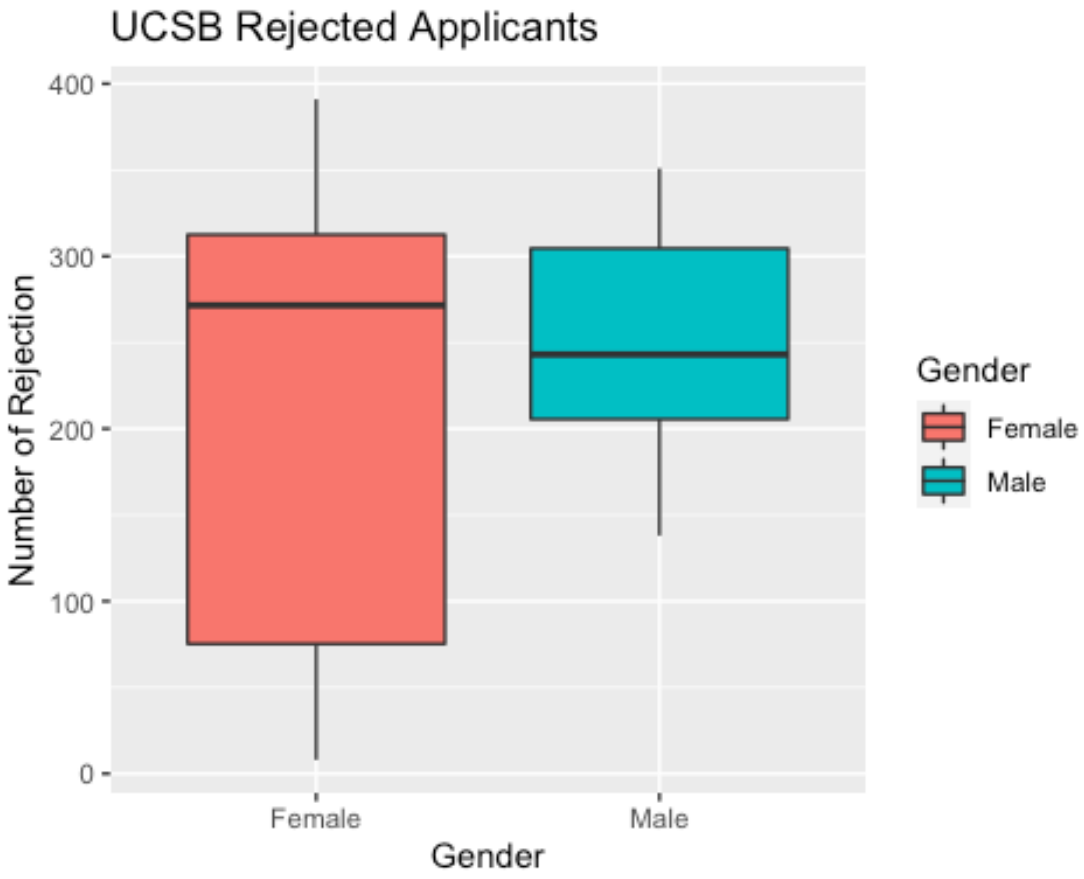

```
#looking at each of the departments, there is a significant difference in the number of applicants for females and males primarily in departments A and B

#because of this difference in the number of applicants it is difficult to state whether there is a sex bias in departments A and B because there are significantly less female applicants


#the amount of female applicants are significantly less than male applicants in departments A and B which would result in less females being accepted into the departments and males dominating the acceptance rates

#in department A and B the number of male applicants that were admitted is increasingly higher than the amount of female applicants that were admitted

#But also in departments A and B there were fewer female applicants in comparison to male applicants which is probably another reason why males were more likely to be admitted


#across departments C,D,E,F , female applicants were at a significantly higher frequency rejected than accepted
```

```r
#In departments C,D,E,F it is noted that there are more female applicants com
pared to departments A,B.

#In departments C,D,E,F, with more female applicants, there is a higher rejec
tion frequency than admitted frequency meaning that females are more likely t
o be rejected than admitted into UCSB.

#based on C,D,E,F clustered bar graphs, males are more likely to be rejected
than admitted as well

#the individual departments do not show a clear conclusion to whether there i
s a sex bias on admission rates in each department.This is due to the number
of female versus male applicants differing in each department

#we can look at USCB as a whole to determine whether females/males are more l
ikely to be accepted or rejected when applying to any of the 6 departments at
UCSB

#after comparing the frequency of admissions by gender for each department, c
reate graphs comparing the total amount of females versus males that were adm
itted across all 6 departments at UCSB


#create a data frame named "ucsb_admissions_admitted" and subset rows from "u
csb_admissions" for females and males who were only admitted
#use a ggplot function to select the appropriate data frame and variables nee
ded for each axis
#use geom_boxplot to create a graph as it will show the median of those admit
ted for females and males

#create a data frame for those admitted into UCSB

ucsb_admissions_admitted <- filter(ucsb_admissions, Admit == "Admitted")

#create a boxplot for females and males admitted into the 6 largest departmen
t at UCSB
#use ggplot function and select "ucsb_admissions_admitted" dataframe and use
aes function to select x axis to be Gender, y axis to be Freq, and fill to be
Gender
#label each axis, title and legend table appropriately by using labs function

ggplot(ucsb_admissions_admitted, aes(x = Gender, y = Freq, fill = Gender)) +
  geom_boxplot()+
  labs(title = "UCSB Accepted Applicants", x = "Gender", y = "Number of Accep
tion")
```



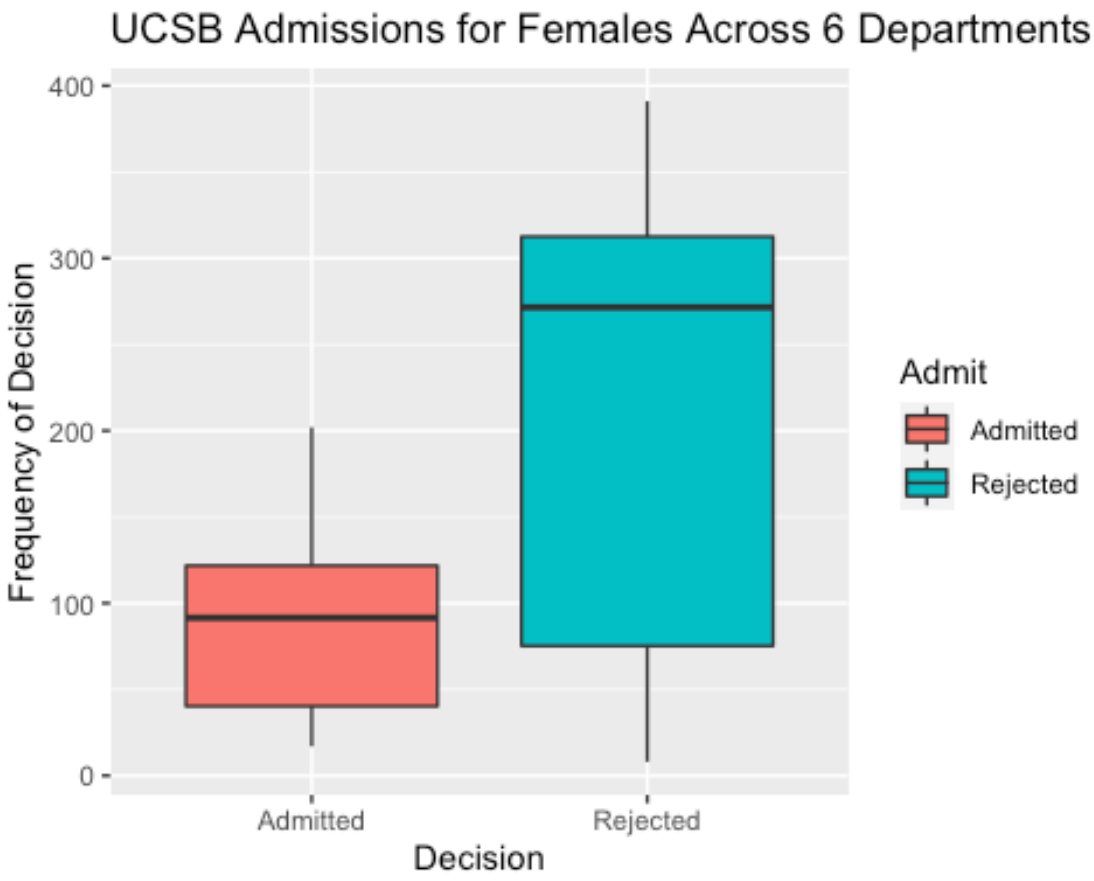

```
#run descriptives for the plots based on females that were admitted and males
that were admitted
#need to subset the "ucsb_admissions_admitted" data frame to create one for j
ust females and one for just males using the filter function
#once created each data frame for seperate genders, run the describe function
for each data frame

#create data frame for females admitted across 6 largest departments at UCSB
and run descriptives on them
ucsb_admissions_admitted_females <-filter(ucsb_admissions_admitted, Gender ==
"Female")
describe(select(ucsb_admissions_admitted_females, X1, Freq))

##      vars n  mean    sd median trimmed   mad min max range skew kurtosis
se
## X1      1 6 13.00  7.48   13.0   13.00  8.90   3  23    20  0.0    -1.80
3.06
## Freq    2 6 92.83 69.11   91.5   92.83 79.32  17 202   185  0.3    -1.54 2
8.21

#create data frame for males admitted across 6 largest departments at UCSB an
d run descriptives on them
ucsb_admissions_admitted_males <- filter(ucsb_admissions_admitted, Gender ==
```

```
"Male")
describe(select(ucsb_admissions_admitted_males, X1, Freq))

##      vars n   mean     sd median trimmed    mad min max range skew kurtosi
s
## X1      1 6  11.00   7.48     11   11.00   8.90   1  21    20 0.00    -1.8
0
## Freq    2 6 199.67 191.98    129  199.67 135.66  22 512   490 0.58    -1.5
8
##         se
## X1    3.06
## Freq 78.38



#the descriptive stats generated showcase the median which is presented on th
e line within each boxplot for the "UCSB Accepted Applicants" graph
#the amount of females admitted to UCSB's 6 largest departments is significan
tly less than males
#the median for females admitted into the 6 largest departments at UCSB is 91
.5 and the mean is 92.83
#the median for males admitted into the 6 largest departments at UCSB is 129
and the mean is 199.67
#the average for males being admitted is twice as likely than females into th
e 6 largest departments at UCSB

#after comparing the total amount of females versus males that were admitted
across all 6 departments at UCSB, compare the amounts of females versus males
rejected across the 6 departments at UCSB
#create a data frame for those rejected into UCSB

ucsb_admissions_rejected <- filter(ucsb_admissions, Admit == "Rejected")

#create a boxplot for females and males rejected into the 6 largest departmen
t at UCSB
#use ggplot function and select "ucsb_admissions_rejected" dataframe and use
aes function to select x axis to be Gender, y axis to be Freq, and fill to be
Gender
#label each axis, title and legend table appropriately by using labs function

ggplot(ucsb_admissions_rejected, aes(x = Gender, y = Freq, fill = Gender)) +
  geom_boxplot()+
  labs(title = "UCSB Rejected Applicants", x = "Gender", y = "Number of Rejec
tion")
```



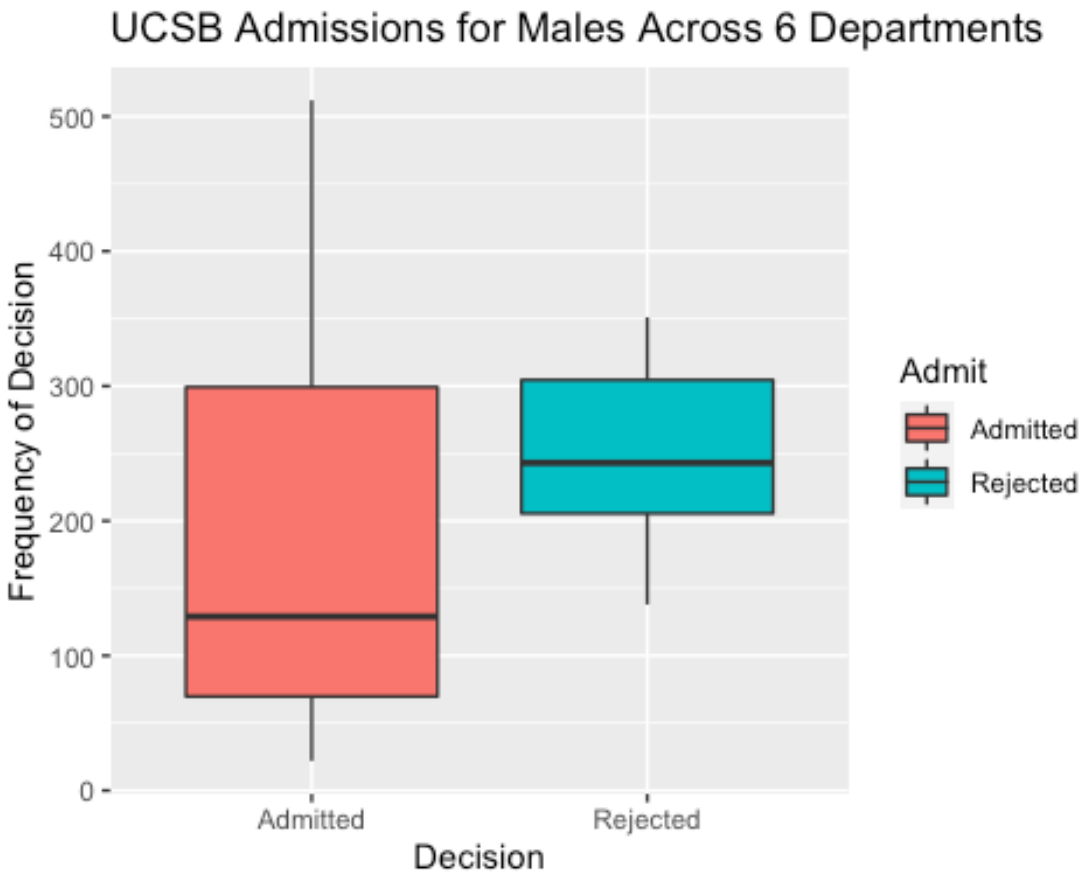

```
#run descriptives for the plots based on females that were rejected and males
that were rejected
#need to subset the "ucsb_admissions_rejected" data frame to create one for j
ust females and one for just males using the filter function
#once created each data frame for separate genders, run the describe function
for each data frame

#create data frame for females rejected across 6 largest departments at UCSB
and run descriptives on them
ucsb_admissions_rejected_females <-filter(ucsb_admissions_rejected, Gender ==
"Female")
describe(select(ucsb_admissions_rejected_females, X1, Freq))

##      vars n mean     sd median trimmed    mad min max range  skew kurtosis
## X1      1 6   14   7.48   14.0      14   8.90   4  24    20  0.00    -1.80
## Freq    2 6  213 161.57  271.5     213 122.31   8 391   383 -0.34    -1.93
##         se
## X1    3.06
## Freq 65.96

#create data frame for males rejected across 6 largest departments at UCSB an
d run descriptives on them
ucsb_admissions_rejected_males <-filter(ucsb_admissions_rejected, Gender == "
```

```
Male")
describe(select(ucsb_admissions_rejected_males, X1, Freq))

##      vars n   mean    sd median trimmed   mad min max range  skew kurtosis
## X1      1 6  12.00  7.48     12   12.00  8.90   2  22    20  0.00     -1.8
## Freq    2 6 248.83 79.27    243  248.83 80.06 138 351   213 -0.05     -1.8
##         se
## X1    3.06
## Freq 32.36

#the descriptive stats generated showcase the median which is also presented
on the line within each boxplot for the "UCSB Rejected Applicants" graph
#the median for females being rejected is significantly greater than for male
s
#median for females being rejected is 271.5
#median for males being rejected is 243
#the median for females is higher in rejected applicants at UCSB's six larges
t departments when compared to male applicants


#create seperate graphs for each gender to determine the number of applicants
based on sex


#create a data frame for each gender and subset rows for each gender from "uc
sb_admissions"
#use a ggplot function to select the appropriate data frame and variables nee
ded for each axis
#use geom_boxplot to create a graph as it will show the median of those admit
ted or rejected for each gender

#create a data frame for females who applied to the 6 largest departments at
UCSB

ucsb_admissions_female <- filter(ucsb_admissions, Gender == "Female")

#create a boxplot for females who applied to the 6 largest departments at UCS
B
#use ggplot function and select "ucsb_admissions_female" dataframe and use ae
s function to select x axis to be Admit, y axis to be Freq, and fill to be Ad
mit
#label each axis, title and legend table appropriately by using labs function

ggplot(ucsb_admissions_female, aes(x = Admit, y = Freq, fill = Admit)) +
  geom_boxplot()+
  labs(title = "UCSB Admissions for Females Across 6 Departments", x = "Decis
ion", y = "Frequency of Decision")
```




```
#use describe function to see the numerical descriptives of females that appl
ied to the 6 largest departments at UCSB
describe(select(ucsb_admissions_female, X1, Freq))

##      vars  n   mean     sd median trimmed    mad min max range skew kurtos
is
## X1      1 12  13.50   7.15   13.5    13.5   8.90   3  24    21 0.00    -1.
53
## Freq    2 12 152.92 134.07  112.5   143.6 140.11   8 391   383 0.39    -1.
48
##         se
## X1    2.07
## Freq 38.70

#create a data frame for males who applied to the 6 largest departments at UC
SB
ucsb_admissions_male <- filter(ucsb_admissions, Gender == "Male")

#create a boxplot for males who applied to the 6 largest departments at UCSB
#use ggplot function and select "ucsb_admissions_male" dataframe and use aes
function to select x axis to be Admit, y axis to be Freq, and fill to be Admi
t
#label each axis, title and legend table appropriately by using labs function
```

```
ggplot(ucsb_admissions_male, aes(x = Admit, y = Freq, fill = Admit)) +
  geom_boxplot()+
  labs(title = "UCSB Admissions for Males Across 6 Departments", x = "Decisio
n", y = "Frequency of Decision")
```




```
#use describe function to see the numerical descriptives of males that applie
d to the 6 largest departments at UCSB
describe(select(ucsb_admissions_male, X1, Freq))
```

```
##      vars  n   mean     sd median trimmed    mad min max range skew kurtos
is
## X1      1 12  11.50   7.15   11.5    11.5   8.90   1  22    21 0.00    -1.
53
## Freq    2 12 224.25 142.37  206.0   215.7 143.07  22 512   490 0.38    -0.
93
##         se
## X1    2.07
## Freq 41.10
```

```
#these two graphs show a side by side comparison for each gender on the admit
ted median and the rejected median

#for both males and females, the median for being rejected is higher than the
median for being admitted

#although for males, they have a higher median for being admitted than female
s being admitted

#for males they also have a lower median of being rejected when comparing to
the median of  females being rejected

#males in this data set are more likely to be accepted when applying at the 6
largest departments in UCSB in comparision to the female applicants

#the reason to why male applicants have a higher admitted frequency than for
females is due to the number of male applicants

#the mean and median number for male applicants for those who applied to UCSB
is significantly higher than for females

# the mean number of male applicants is 224. 25 and the mean number of female
applicants is 152.92

# the median of male applicants is also higher which is 206.0 and the median
for female applicants is also significantly lower at 112.0


#this can showcase that UCSB's six largest departments are catered towards ma
le applicants, which would mean that the departments would have a higher rate
of acceptance than rejection for males

#with less female applicants, males will naturally have a higher chance to be
accepted at UCSB in these 6 departments

#there is a reason to why there are less female applicants and that may be du
e to the time period because the data is from 1973 and at the time, females w
ere less likely to pursue a post secondary degree.

#due to the influx of male applicants, there is a natural bias that occurs as
more male applicants are likely to be accepted and female applicants are more
likely to be rejected at UCSB’s 6 largest departments

#to completely state that UCSB’s six largest departments hold a sex bias base
d on this data set would be wrong as there are not equal amounts of females a
nd male applicants

#based on this dataset, it is concluded that there are more males that are ac
cepted than females as there are more male applicants in the 6 largest depart
ments at UCSB making UCSB’s 6 largest departments more male dominated than fe
male dominated.
```

e. Discuss any limitations of the data or the visualization you applied:

The limitations of our data is that most of the variables used are categorical except for one, which is the frequency. It was hard to have a variety of graphs as our data mainly was categorical variables so we were limited in the graphs we could use. We chose to stick with graphs like boxplots and clustered bar graphs as it would show a clear depiction for the data set we had.

The way that the data is organized in the original data set made it difficult to create graphs for the specific areas we were targeting. This lead us to do an extra step by creating dataframes for each graph by subsetting data we needed from the original data set. For example, we would have to filter the original data frame to subset rows for each department to create the graphs. We made a lot of data frames to create specific graphs we needed to showcase our data.

Another limitation that we did not foresee was that the distribution between genders and departments as it was not randomized. Meaning that in a few of the six departments mostly women applied, and in the other departments, mostly men applied. Because of this, it is hard to tell if UCSB admission rates display a sex bias as the number of female to male applicants were different. There were more male applicants across all 6 departments which is why males were more likely to be accepted. If there was an equal amount of male and female applicants, we would be able to clearly define whether there was a sex bias and if certain departments held more of a sex bias than other. The only thing we could conclude from our data set was that there were more accepted males than females in the 6 largest departments at UCSB and that was because there were more male applicants than female applicants.