Code

CS 422 HW1

Jane Downer

Part 2.1-A

Hide

```
setwd("~/Desktop")
```

```
The working directory was changed to /Users/user/Desktop inside a noteboo
k chunk. The working directory will be reset when the chunk is finished r
unning. Use the knitr root.dir option in the setup chunk to change the wo
rking directory for notebook chunks.
```

Hide

```
library(dplyr)
collegeData <- read.csv(file = 'College.csv', row.names = 1)
collegeData[c(1:5),c(1,5,8,10)]
```

	Private <fctr>	Top10perc <int>	P.Undergrad <int>	Room.Bo... <int>
Abilene Christian University	Yes	23	537	3300
Adelphi University	Yes	16	1227	6450
Adrian College	Yes	22	99	3750
Agnes Scott College	Yes	60	63	5450
Alaska Pacific University	Yes	16	869	4120

5 rows

Part 2.1-B

Hide

```
private <- nrow(collegeData %>% filter(Private == "Yes"))
public <- nrow(collegeData %>% filter(Private == "No"))
noquote(sprintf("There are %d private colleges, and %d public colleges in
the dataset.", private, public))
```

```
[1] There are 565 private colleges, and 212 public colleges in the datase
t.
```

Part 2.1-C

Hide

```
newDF <- collegeData[,c("Private", "Apps", "Accept", "Enroll",
                        "PhD", "perc.alumni", "S.F.Ratio",
                        "Grad.Rate")]
head(newDF,6)
```

	Private <fctr>	A... <int>	Acc... <int>	Enroll <int>	... <int>	perc.alumni <int>
Abilene Christian University	Yes	1660	1232	721	70	12
Adelphi University	Yes	2186	1924	512	29	16
Adrian College	Yes	1428	1097	336	53	30
Agnes Scott College	Yes	417	349	137	92	37
Alaska Pacific University	Yes	193	146	55	76	2

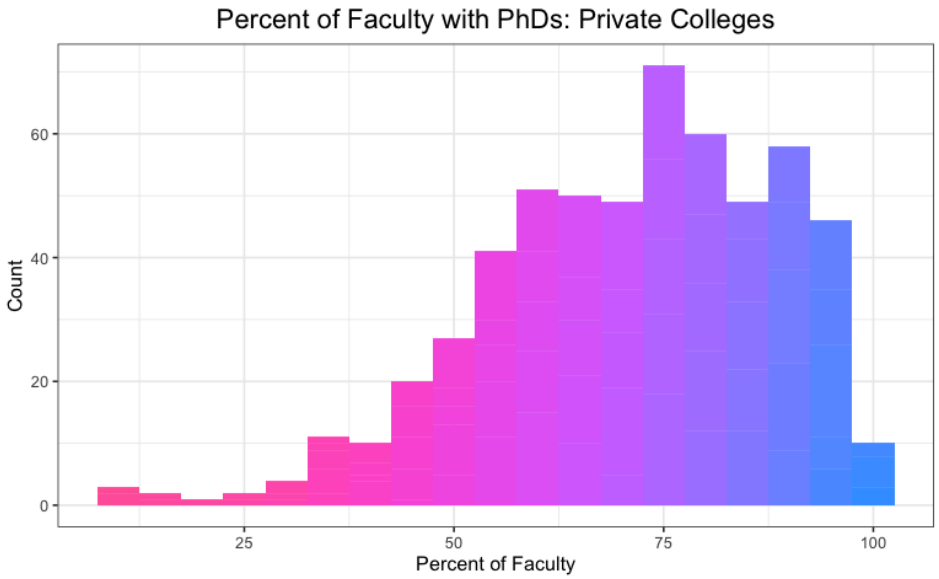
	Private <fctr>	A... <int>	Acc... <int>	Enroll <int>	... <int>	perc.alumni <int>
Albertson College	Yes	587	479	158	67	11

6 rows | 1-8 of 8 columns

Part 2.1-D-i

Hide

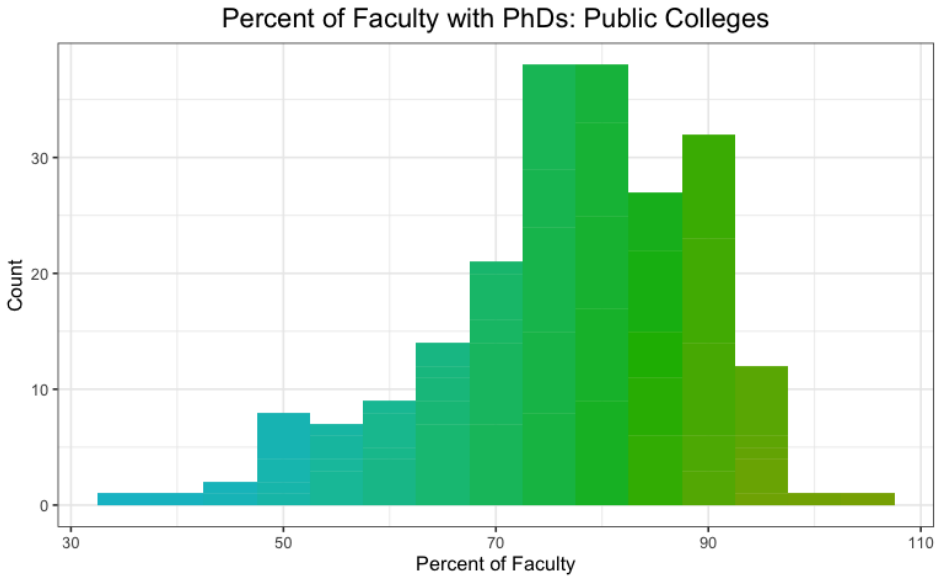
```
library(ggplot2)
p <- newDF %>% filter(Private == "Yes") %>%
  ggplot(aes(PhD, fill = cut(PhD, 3000))) +
  geom_histogram(binwidth = 5, show.legend = F) +
  ggtitle("Percent of Faculty with PhDs: Private Colleges") +
  xlab("Percent of Faculty") +
  ylab("Count") +
  scale_fill_discrete(h = c(350,250)) +
  theme_bw() +
  theme(plot.title = element_text(size = 15, hjust = 0.5))
p
```



Part 2.1-D-ii

Hide

```
p <- newDF %>% filter(Private == "No") %>%
  ggplot(aes(PhD, fill = cut(PhD, 3000))) +
  geom_histogram(binwidth = 5, show.legend = F) +
  ggtitle("Percent of Faculty with PhDs: Public Colleges") +
  xlab("Percent of Faculty") +
  ylab("Count") +
  scale_fill_discrete(h = c(200,100)) +
  theme_bw() +
  theme(plot.title = element_text(size = 15, hjust = 0.5))
p
```



Part 2.1-D-iii (extra credit)

Hide

```
print("See parts i and ii.", quote = F)
```

```
[1] See parts i and ii.
```

Part 2.1-E-i

Hide

```
# Top 5 colleges with minimum graduation rates
select(head(newDF[order(newDF$Grad.Rate),],5),Grad.Rate)
```

	Grad.Rate <int>
Texas Southern University	10
Alaska Pacific University	15
Montreat-Anderson College	15
Brewton-Parker College	18
Claflin College	21

5 rows

Part 2.1-E-ii

Hide

```
# Top 5 colleges with maximum graduation rates
select(tail(newDF[order(newDF$Grad.Rate),],5),Grad.Rate)
```

	Grad.Rate <int>
Missouri Southern State College	100
Santa Clara University	100
Siena College	100
University of Richmond	100
Cazenovia College	118

5 rows

Part 2.1-F

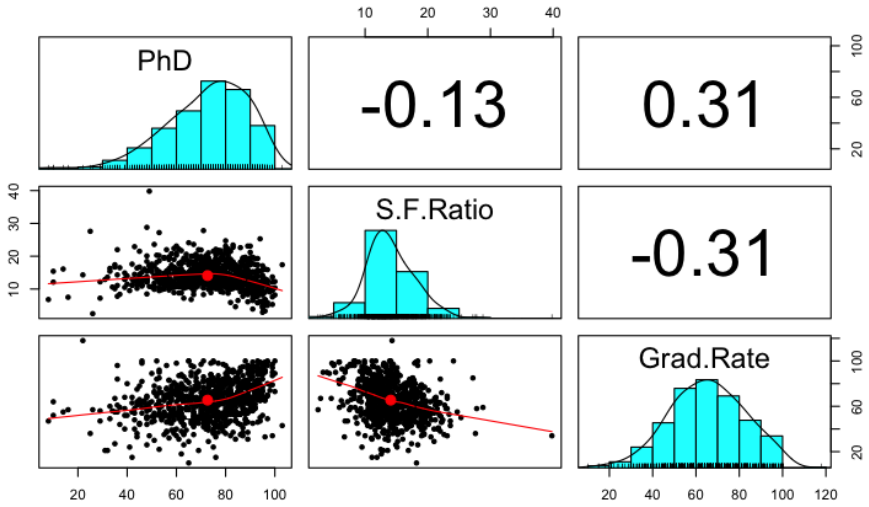
Hide

```
#install.packages("psych")
```

Part 2.1-F-i

Hide

```
library(psych)
pairs.panels(newDF[,c("PhD","S.F.Ratio","Grad.Rate")])
```



Part 2.1-F-ii

Hide

```
print("'PhD'(percentage of faculty with PhDs) and 'Grad.Rate' (Graduation
Rate) are positively correlated. This makes sense, because academically d
riven students and faculty are likely drawn to the same institutions.", q
uote = F)
```

```
[1] 'PhD'(percentage of faculty with PhDs) and 'Grad.Rate' (Graduation Ra
te) are positively correlated. This makes sense, because academically dri
ven students and faculty are likely drawn to the same institutions.
```

Part 2.1-F-iii

Hide

```
print("'S.F.Ratio' (student-to-faculty ratio) and 'Grad.Rate' (Graduation
Rate) are negatively correlated. This makes sense. This makes sense. A hi
gher S.F.Ratio implies that there are fewer resources available to studen
ts, which could explain lower graduation rates.", quote = F)
```

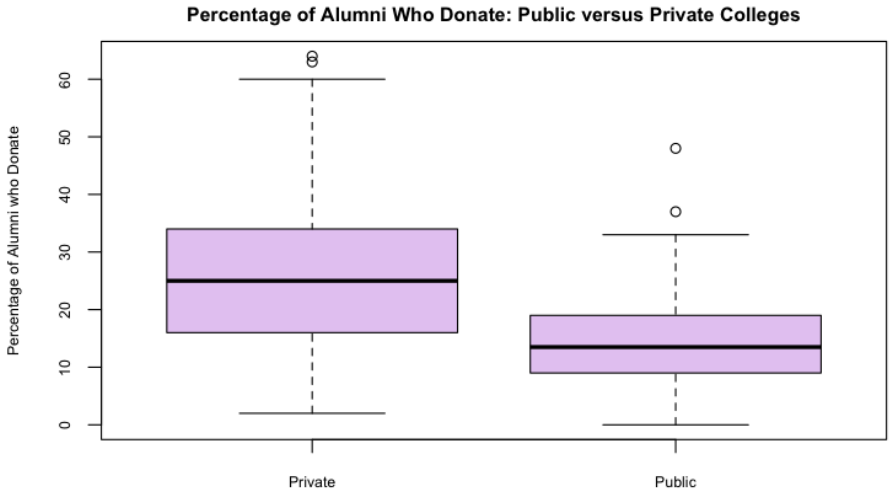
```
[1] 'S.F.Ratio' (student-to-faculty ratio) and 'Grad.Rate' (Graduation Ra
te) are negatively correlated. This makes sense. This makes sense. A high
er S.F.Ratio implies that there are fewer resources available to student
s, which could explain lower graduation rates.
```

Part 2.1-G

Hide

```
dataPrivate <- collegeData %>% filter(Private == "Yes")
dataPublic <- collegeData %>% filter(Private == "No")

boxplot(dataPrivate$perc.alumni, dataPublic$perc.alumni,
        main = "Percentage of Alumni Who Donate: Public versus Private Co
lleges",
        ylab = "Percentage of Alumni who Donate",
        names = c("Private", "Public"),
        cex.main = 0.9, cex.lab = 0.7, cex.axis = 0.7,
        col = c(rgb(0.9,0.8,0.95)), lwd = 1)
```



Hide

```
print("Based on the boxplots above, the highest donors tend to be alumni
 of private colleges.", quote = F)
```

```
[1] Based on the boxplots above, the highest donors tend to be alumni of
private colleges.
```

Part 2.1-H

Hide

```
cdfCollege <- ecdf(collegeData$Expend)
```

Part 2.1-H-iii:

Hide

```
p_color = c(rgb(0.9,0.8,0.95))

plot(cdfCollege, verticals = T, do.points = F, col = p_color, lwd=3,
     main = "Instructional Expenditure", cex.main = 0.9,
     xlab ="Expenditure per Student (in dollars)", cex.lab = 0.7,
     ylab = "Cumulative Density",
     xaxs = "i", yaxs = "i", xlim  = c(0,35000), ylim = c(0,1), xaxt =
'n', cex.axis = 0.7)

box(col = "white")
```

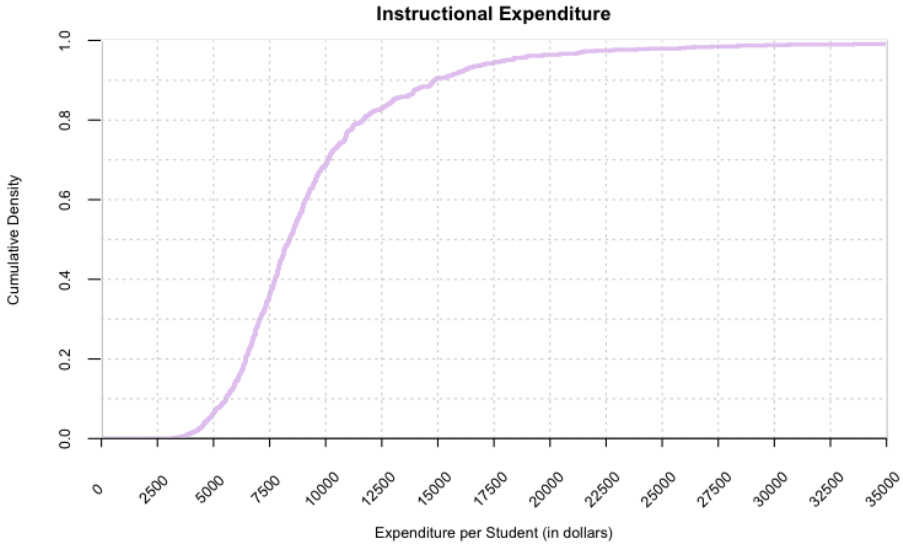
Hide

```
grid(14,10)

axis(xaxs = "i", side=1,tck=-0.02,at=c(seq(from=0,to=35000,by=2500)), lab
els = F)
```

Hide

```
text(seq(1, 35001, by=2500), par("usr")[3] - 0.2,
     labels = c(seq(from=0,to=35000,by=2500)), srt = 45, pos = 3, offset
 = 1,
     xpd = TRUE, cex =0.7)
```



Part 2.1-H-i:

Hide

```
fifty = quantile(collegeData$Expend, probs = 0.5)
noquote(sprintf("The gridlines in the above CDF plot place the median exp
enditure between $7500 and $10,000 per student -- roughly $8500. Upon per
forming the actual calculation, we find that the true value is $%0.f, whi
ch is not far off from the estimate.", fifty))
```

```
[1] The gridlines in the above CDF plot place the median expenditure betw
een $7500 and $10,000 per student -- roughly $8500. Upon performing the a
ctual calculation, we find that the true value is $8377, which is not far
off from the estimate.
```

Part 2.1-H-ii

Hide

```
eighty = quantile(collegeData$Expend, probs = 0.8)
noquote(sprintf("The gridlines in the above CDF plot suggest that 80%% of
students pay below a value that is betwen $10,000 and $12,500 -- about $1
1,500. Upon performing the actual calculation, we find that the true valu
e is $%0.f, which is not far off from the estimate.", eighty))
```

```
[1] The gridlines in the above CDF plot suggest that 80% of students pay
below a value that is betwen $10,000 and $12,500 -- about $11,500. Upon p
erforming the actual calculation, we find that the true value is $11656,
which is not far off from the estimate.
```