Earning

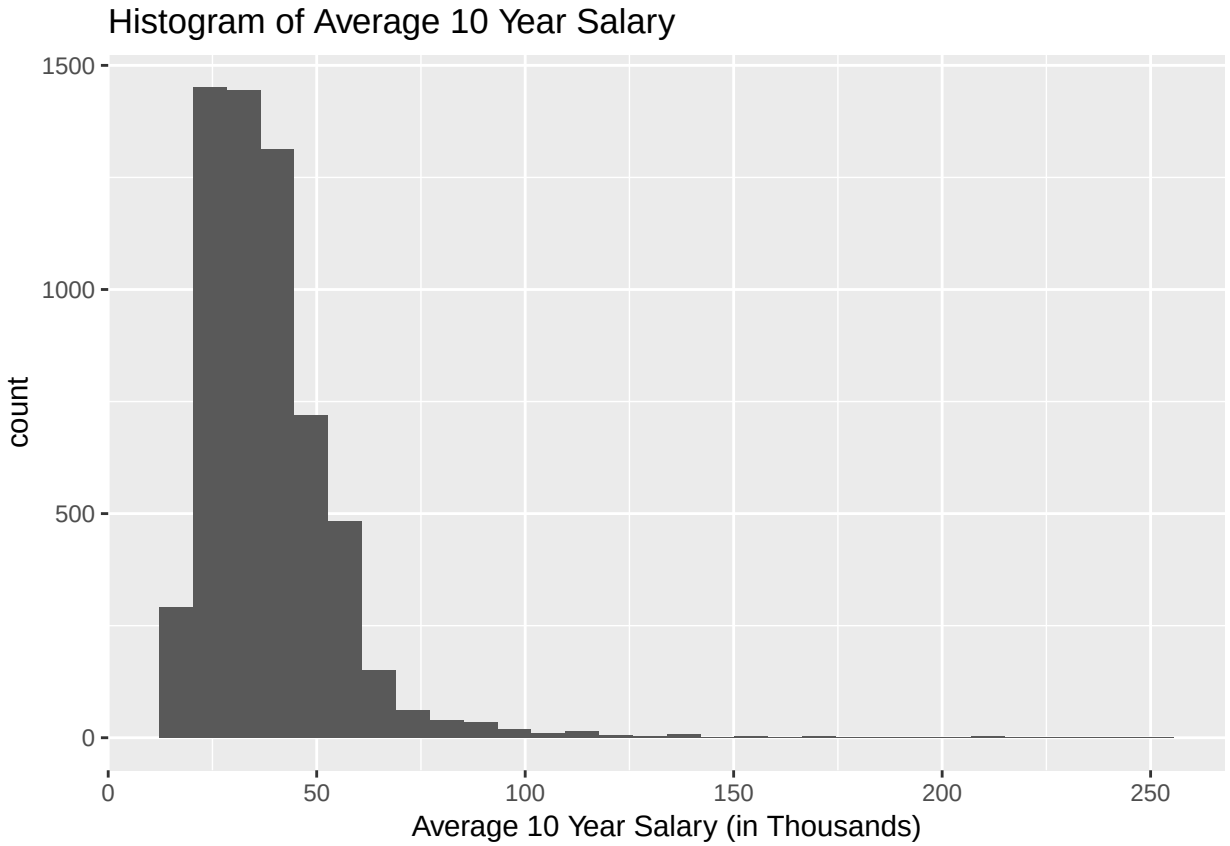
```
setwd(("/Users/Mavis1/Documents/ColumbiaUniv/Fall2018/EDAV/"))
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------- tidyverse 1.2
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

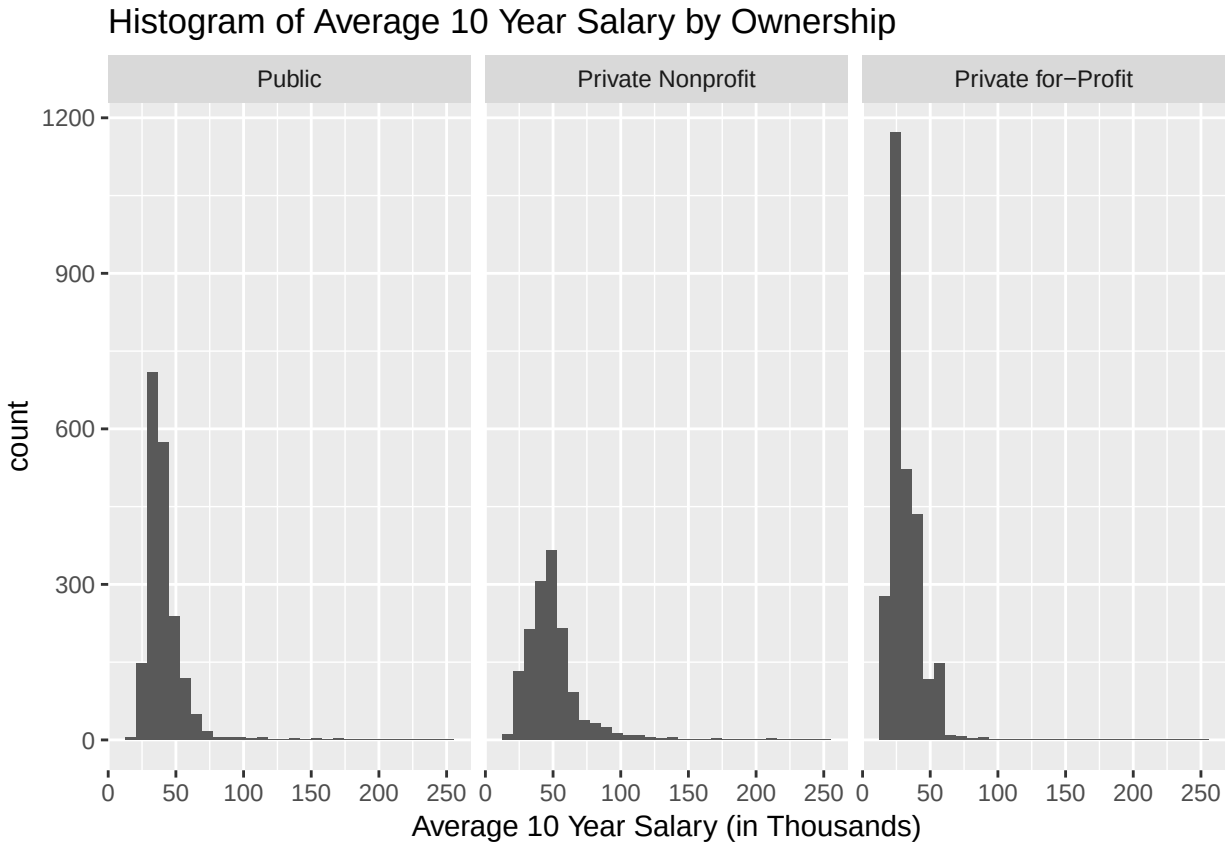
```
## -- Conflicts -------------------------------------------------------------------- tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
college = read_rds("college.rds")
college[college=='NULL'] <- NA
#college = na.omit(college$avg_10yr_salary)
college =  college[!is.na(college$avg_10yr_salary), ]
college$avg_10yr_salary = college$avg_10yr_salary/1000
college$tuition_instate = college$tuition_instate/1000
college$tuition_out = college$tuition_out/1000
recode <- c('1'='Public','2'='Private Nonprofit','3'='Private for-Profit')
college$ownership <- recode[college$ownership]
college$ownership<-fct_relevel(college$ownership,'Public','Private Nonprofit','Private for-Profit')
ggplot(college, aes(x=avg_10yr_salary))+
  geom_histogram(bins = 30)+
  labs(x="Average 10 Year Salary (in Thousands)")+
  ggtitle("Histogram of Average 10 Year Salary")
```



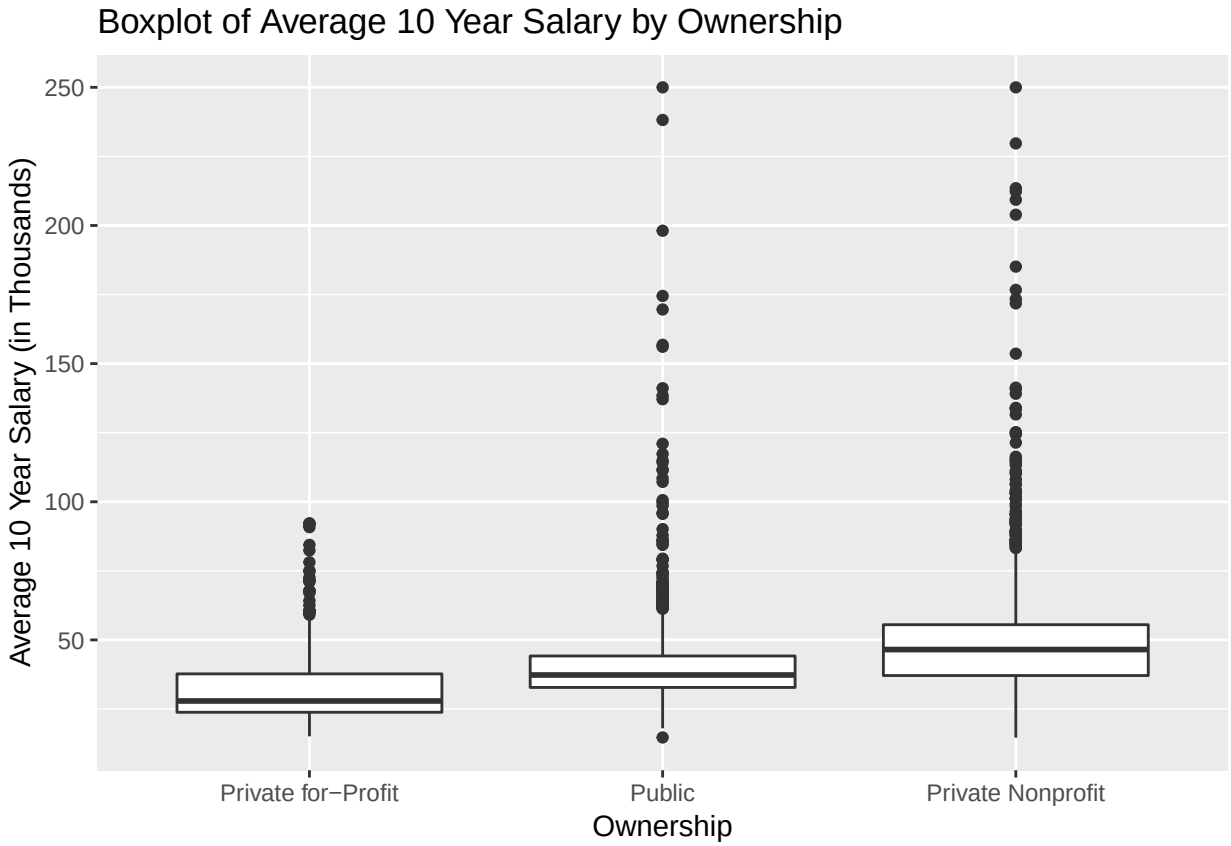
The average 10 year salary variable represents the mean earnings of students working and not enrolled at school 10 years after entry. The overall distribution of the average 10 year salary is skewed to the right. Most of our observations have an average 10 year salary of around 20,000 to 40,000 dollars. There are also a couple observations with higher average 10 year salary than 100,000 dollars.

```
a=count(college[ which(college$ownership=='Public'),])
b=count(college[ which(college$ownership=='Private Nonprofit'),])
c=count(college[ which(college$ownership=='Private for-Profit'),])
ggplot(college, aes(x=avg_10yr_salary))+
  geom_histogram(bins = 30)+
  facet_grid(.~ownership)+
  labs(x="Average 10 Year Salary (in Thousands)")+
  ggtitle("Histogram of Average 10 Year Salary by Ownership")
```



By faceting average 10 year salary into different ownership group, we can see the different distribution for public, private nonprofit, and private for-profit schools. The total numbers of observations for the three ownership type are 1890, 1481, 2691. Among the three types, although private nonprofit schools have the least observations, they have the distribution most skewed right, with highest mean and mode. This indicates students from private nonprofit schools have highest average 10 year salary among the threee school types. While Public schools have the mode for average 10 year salary at around 30,000 dollars, private for-profit schools have the mode at around 20,000 dollars, and overall, students from public schools rank the second in average 10 year salary. Therefore, the average 10 year salary differs for students from different ownership type schoools.

```
ggplot(college, aes(x=reorder(ownership,avg_10yr_salary),y=avg_10yr_salary))+
  geom_boxplot()+
  labs(y="Average 10 Year Salary (in Thousands)",x="Ownership")+
  ggtitle("Boxplot of Average 10 Year Salary by Ownership")
```



The boxplot for average 10 year salary gives us a better understanding of the median, as well as outliers of the data. We can see that the median for private nonprofit schools is highest, while that of private for-profit ones is lowest. Moreover, we can clearly see the outliers. Comparing to other ownership types, private for-profit schools have the least number of outliers, and the value of outlier is lowest (no outliers with value over 100,000). The boxplot also indicates students from private nonprofit schools have higher average 10 year salary than those from public and private for-profit schools.

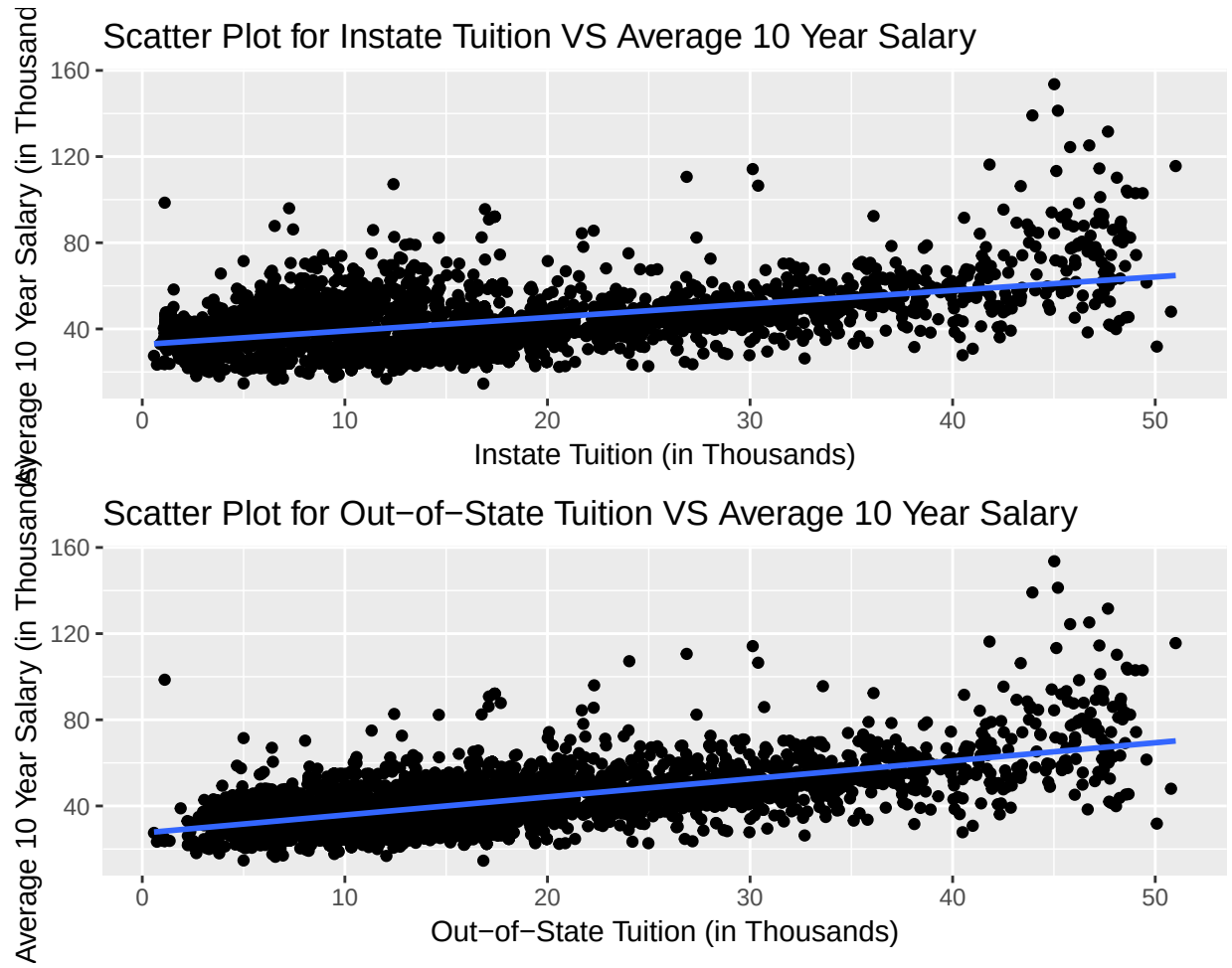
```
college =  college[!is.na(college$tuition_out), ]
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
p1 <- ggplot(college, aes(x=tuition_instate,y=avg_10yr_salary))+
  geom_point()+
  labs(x="Instate Tuition (in Thousands)",y="Average 10 Year Salary (in Thousands)")+
  ggtitle("Scatter Plot for Instate Tuition VS Average 10 Year Salary")+
  geom_smooth(method='lm',formula=y~x)
p2 <- ggplot(college, aes(x=tuition_out,y=avg_10yr_salary))+
  geom_point()+
  labs(x="Out-of-State Tuition (in Thousands)",y="Average 10 Year Salary (in Thousands)")+
  ggtitle("Scatter Plot for Out-of-State Tuition VS Average 10 Year Salary")+
  geom_smooth(method='lm',formula=y~x)
r2_1 = format(summary(lm(college$avg_10yr_salary~college$tuition_instate))$r.squared*100, digits = 3)
r2_2 = format(summary(lm(college$avg_10yr_salary~college$tuition_out))$r.squared*100, digits = 3)
```

```
grid.arrange(p1,p2,ncol=1)
```



We can see from the scatter plot that there is some positive correlation between both instate, out-of-state tuition and average 10 year salary. The more expensive the tuition is, the higer the average 10 year salary is. 28, 40.9 percent of the variation in average 10 year salary is explained by the variation in predictor instate tuition, and out-of-state tuition, respectively.