Earning

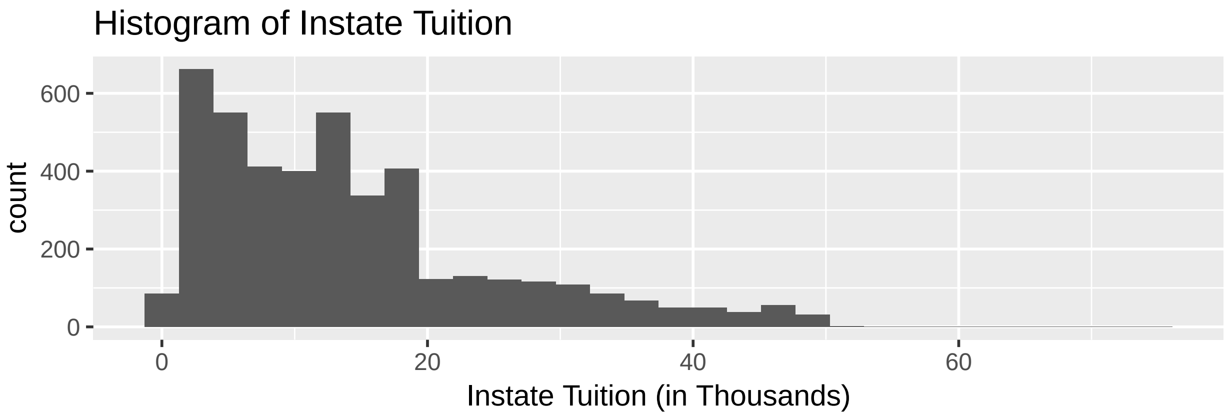
```r
setwd(("/Users/Mavis1/Documents/ColumbiaUniv/Fall2018/EDAV/"))
library(tidyverse)
```

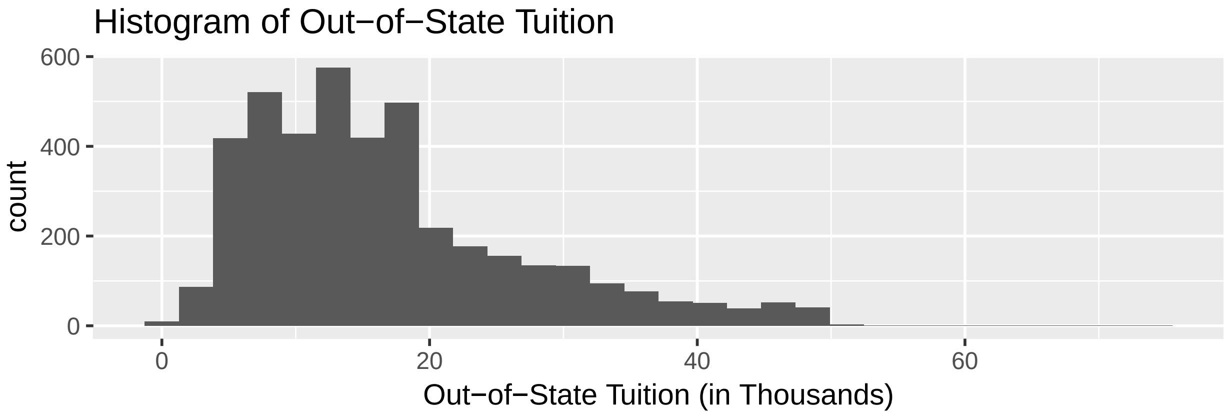
```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.2

## 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()
```

```r
library(gridExtra)
```

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

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

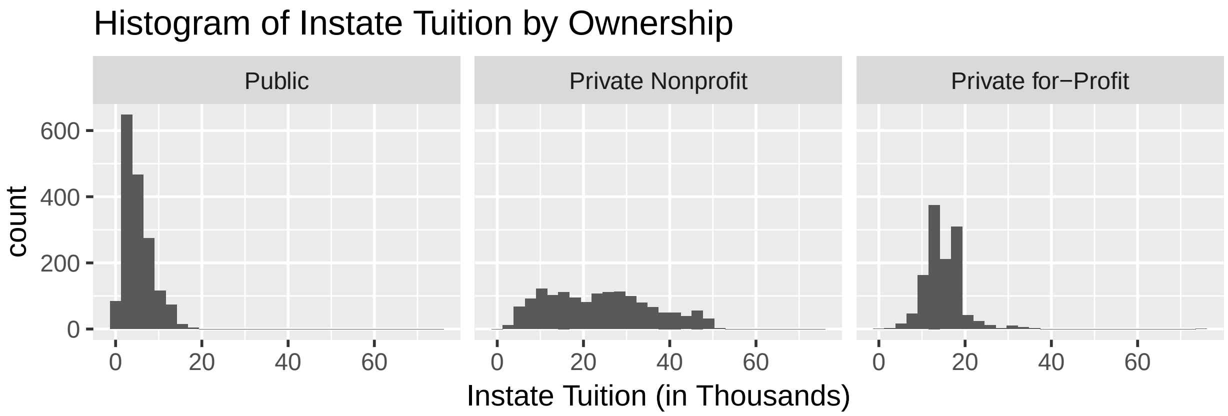
```r
college = read_rds("college.rds")
college[college=='NULL'] <- NA
#college = na.omit(college$avg_10yr_salary)
college1 =  college[!is.na(college$tuition_instate), ]
college2 =  college[!is.na(college$tuition_out), ]
#college$avg_10yr_salary = college$avg_10yr_salary/1000
college1$tuition_instate = college1$tuition_instate/1000
college2$tuition_out = college2$tuition_out/1000
recode <- c('1'='Public','2'='Private Nonprofit','3'='Private for-Profit')
college1$ownership <- recode[college1$ownership]
college1$ownership<-fct_relevel(college1$ownership,'Public','Private Nonprofit','Private for-Profit')
college2$ownership <- recode[college2$ownership]
college2$ownership<-fct_relevel(college2$ownership,'Public','Private Nonprofit','Private for-Profit')
p3<-ggplot(college1, aes(x=tuition_instate))+
  geom_histogram(bins = 30)+
  labs(x="Instate Tuition (in Thousands)")+
  ggtitle("Histogram of Instate Tuition")
p4<-ggplot(college2, aes(x=tuition_out))+
  geom_histogram(bins = 30)+
  labs(x="Out-of-State Tuition (in Thousands)")+
  ggtitle("Histogram of Out-of-State Tuition")
grid.arrange(p3,p4,ncol=1)
```

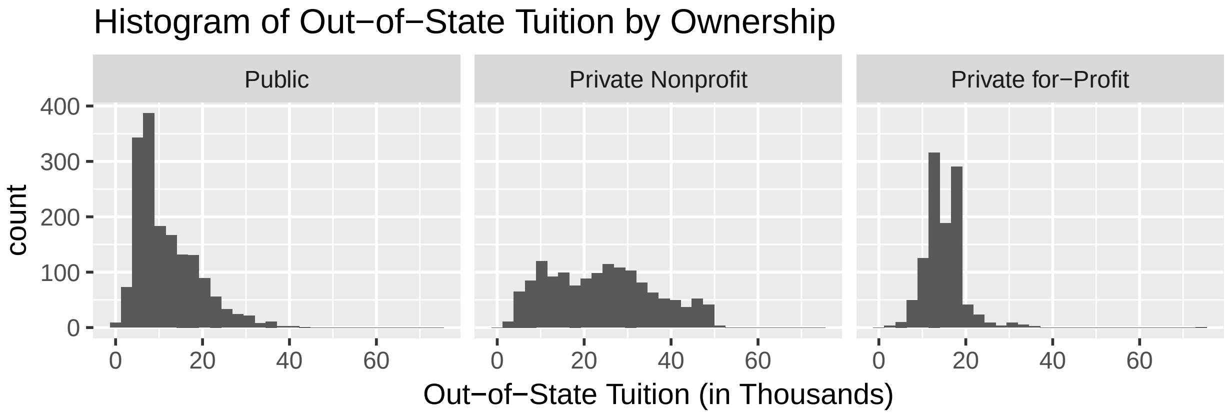




Instate and out-of-state tuition are the tuition and fees for instate, and out-of-state students, respectively. From the histogram above, we can see that the distribution of the two tuition are similar, and they both skewed right. Out-of-state students have overall higher tuition fee than instate students.

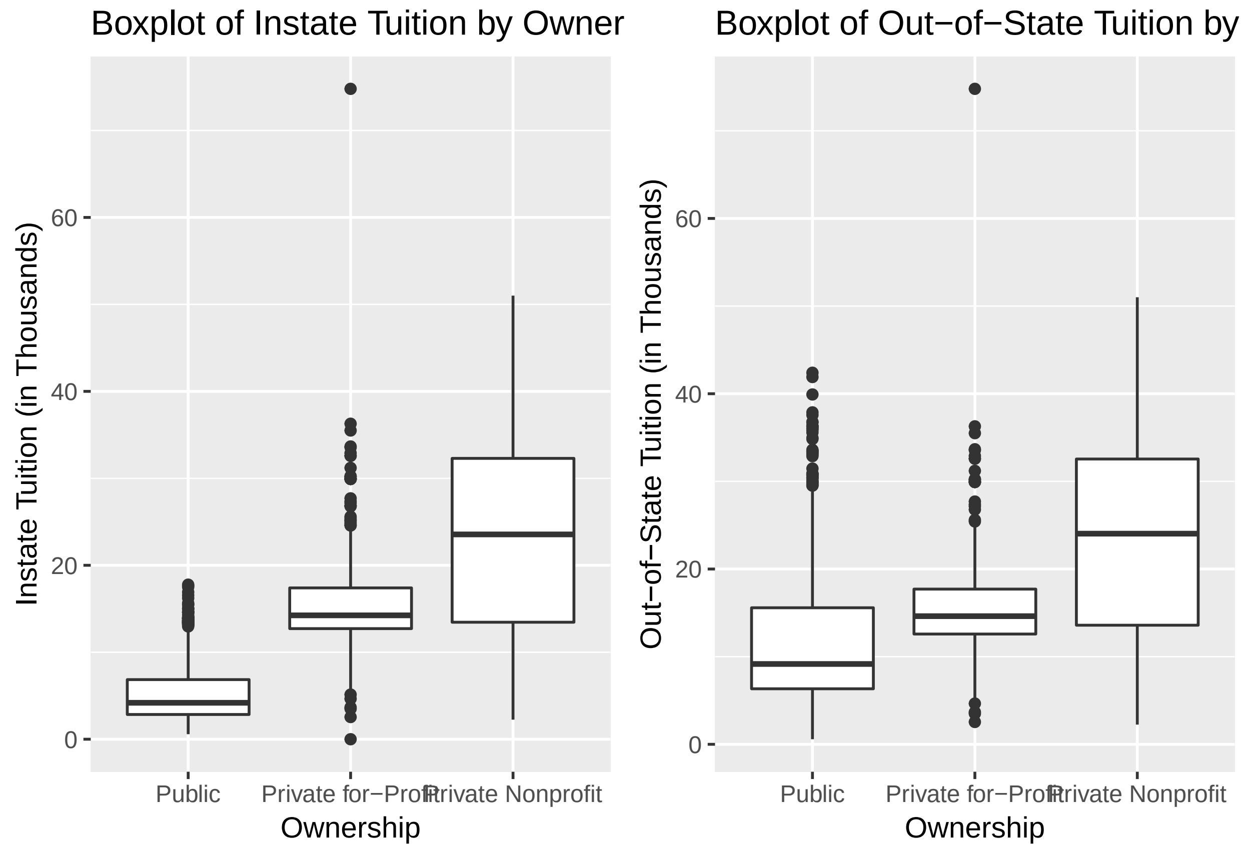
```r
p5 <- ggplot(college1, aes(x=tuition_instate))+
  geom_histogram(bins = 30)+
  facet_grid(.~ownership)+
  labs(x="Instate Tuition (in Thousands)")+
  ggtitle("Histogram of Instate Tuition by Ownership")
p6 <- ggplot(college2, aes(x=tuition_out))+
  geom_histogram(bins = 30)+
  facet_grid(.~ownership)+
  labs(x="Out-of-State Tuition (in Thousands)")+
  ggtitle("Histogram of Out-of-State Tuition by Ownership")
grid.arrange(p5,p6,ncol=1)
```





By faceting instate and out-of-state tuition into different ownership groups, we can see the different distributions for public, private nonprofit, and private for-profit schools, but the pattern for instate and out-of-state tuition is similar. For public schools, both instate and out-of-state tuition are lowest among all three ownership type, and the distribution is unimodal and skewed right. For private nonprofit schools, the tuition have larger range, and its distribution is bimodal. Therefore, the tuition differs for students from different ownership type schoools.

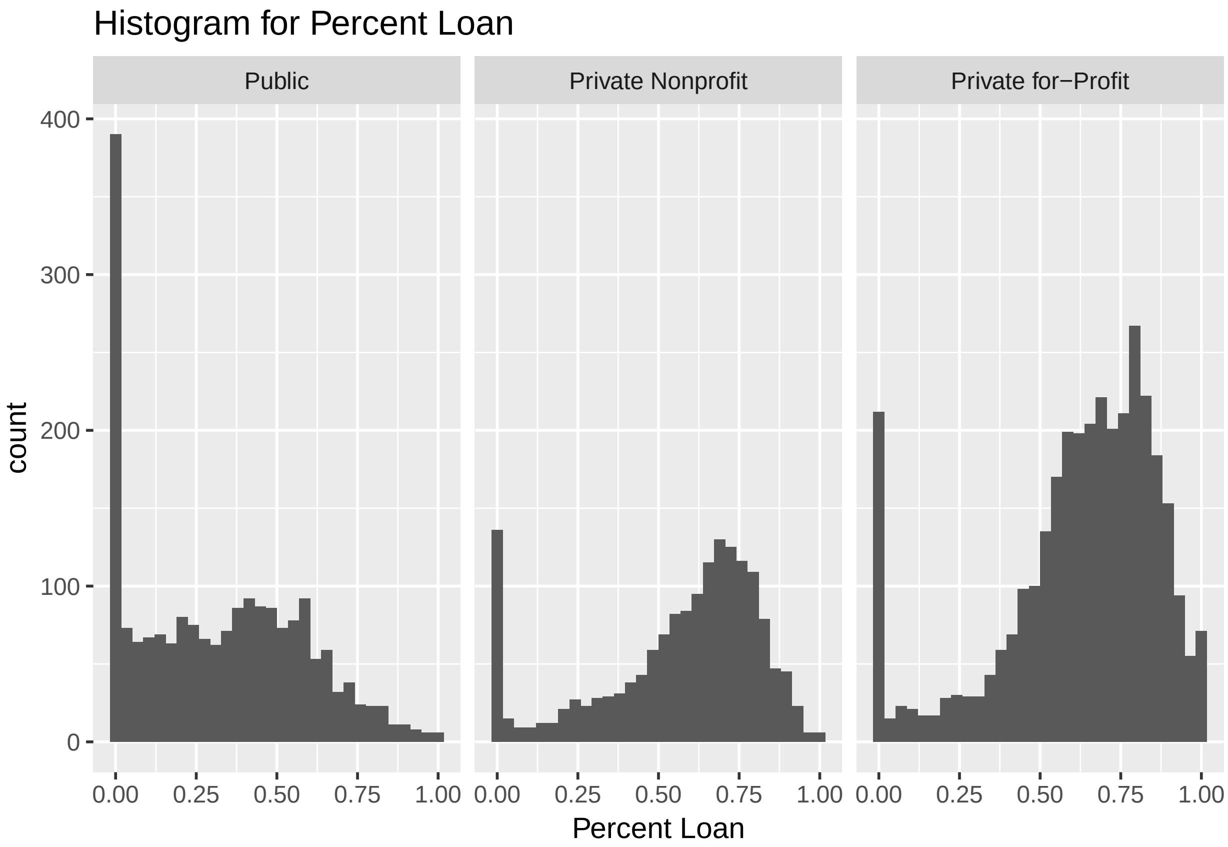
```r
p7 <- ggplot(college1, aes(x=reorder(ownership,tuition_instate),y=tuition_instate))+
  geom_boxplot()+
  labs(y="Instate Tuition (in Thousands)",x="Ownership")+
  ggtitle("Boxplot of Instate Tuition by Ownership")
p8 <- ggplot(college2, aes(x=reorder(ownership,tuition_out),y=tuition_out))+
  geom_boxplot()+
  labs(y="Out-of-State Tuition (in Thousands)",x="Ownership")+
  ggtitle("Boxplot of Out-of-State Tuition by Ownership")
grid.arrange(p7,p8,ncol=2)
```



The boxplot for tuition by ownership 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 public is lowest. Moreover, we can clearly see the outliers. The one outlier for private for-private schools with value over 70,000 dollar seems like an unusal observation.

```
college3 =  college[!is.na(college$pct_loan), ]
college3$ownership <- recode[college3$ownership]
college3$ownership<-fct_relevel(college3$ownership,'Public','Private Nonprofit','Private for-Profit')
ggplot(college3, aes(x=pct_loan))+
  geom_histogram()+
  labs(x="Percent Loan")+
  ggtitle("Histogram for Percent Loan")+
  facet_grid(.~ownership)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The percent loan variable represents percent of all undergraduate students receiving a federal student loan. By faceting into different ownership groups, the histogram shows that private for-profit schools have higher rate of students receiving a federal loan, while public schools have the lowest. For all ownership types, there are a lot of students receiving no federal loan.