

# College Financial Data

Code ▼

Desiree Brinn-Rodriguez

12/5/2021

- Introduction
  - Cleaning Dataset
- Data Summary and Statistics
- Enrollment
- College Costs
  - Clustering Example
  - Revenue & Expenses
  - Faculty Salaries
- Student Loan Debt
  - Regression Example
  - Map with Loan Data
- Conclusion
- References

## Introduction

This analysis explores College Scorecard information available from the Department of Education's Open Data Platform. The College Scorecard is an aggregation of data from higher education institutions across the United States. Prospective students can go to <https://collegescorecard.ed.gov/> (<https://collegescorecard.ed.gov/>) to compare colleges in terms of several categories, such as cost, loan debt, graduation rate, and retention rate. Much of this data stems from the Integrated Post-secondary Education Data System (IPEDS), a survey instituted by the National Center for Education Statistics (NCES) that has oversight from the U.S. Department of Education. This program collects annual data from colleges and universities who participate in federal financial aid programs (ED, 2021). While the College Scorecard dataset is extensive (original dataset contained 2,392 columns), this focus has been narrowed to specific aspects of college financials: total cost of attendance, tuition revenue, average family income, average faculty salary, median student debt, and total student loan debt. The main factors for comparison include a geographical breakdown by zip code and school type: public, private non-profit, and proprietary (for-profit). Much of this information is derived from student information provided on their FAFSA (Free Application for Federal Student Aid), institutional databases, and the National Student Loan Data System (NSLDS). While the College Scorecard exists for a prospective applicant to view institutions individually or compare side-by-side, this analysis seeks to gain an understanding of financial areas of concern from the totality of colleges in the Scorecard dataset.



## FIND THE RIGHT FIT.

Search and compare colleges: their fields of study, costs, admissions, results, and more.

SEARCH SCHOOLS

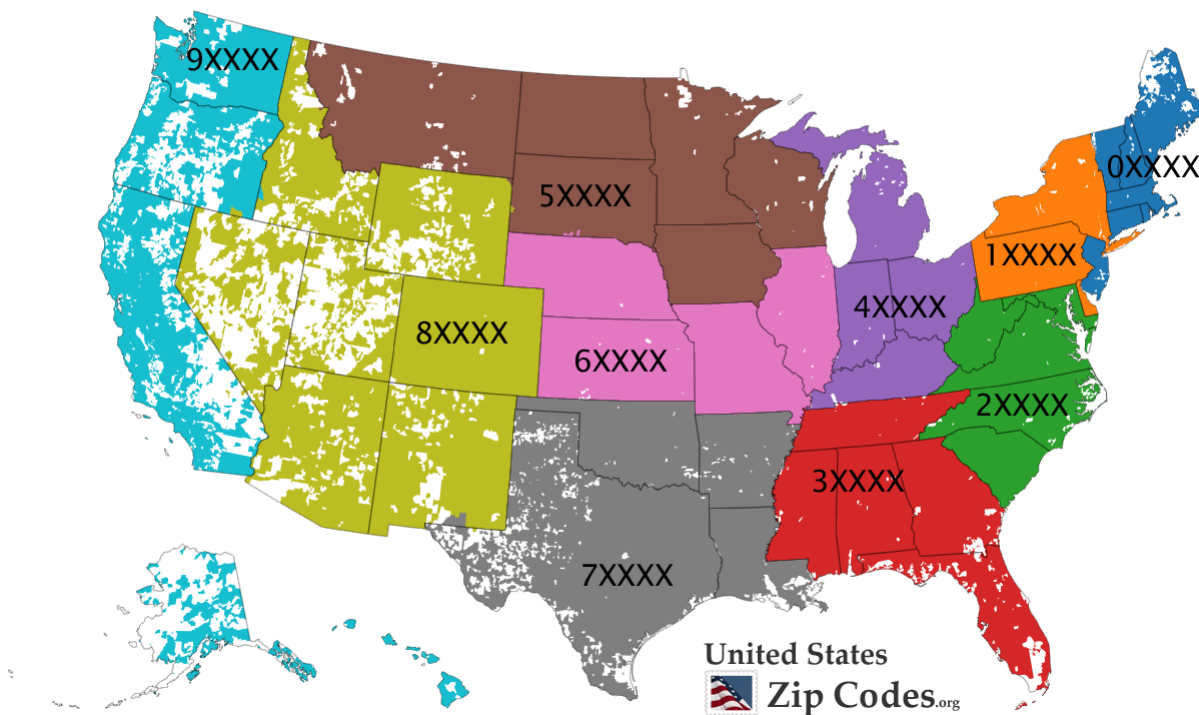
SEARCH FIELDS OF STUDY

SHOW ME OPTIONS

CUSTOM SEARCH

(U.S. Department of Education, n.d.)

## ZIP Code Zones



(UnitedStatesZipCodes.org, 2021)

Data Structure:

Hide

```
#loading file and importing ggplot2  
getwd()
```

```
## [1] "C:/Users/dbrin/Desktop/CS Big Data/Final_Project"
```

Hide

```
setwd("C:/Users/dbrin/Desktop/CS Big Data/Final_Project")
getwd()
```

```
## [1] "C:/Users/dbrin/Desktop/CS Big Data/Final_Project"
```

Hide

```
coldata<-read.csv("CollegeScorecardData.csv")
library(ggplot2)

#rename columns

colnames(coldata)<-c("IPEDSID", "OPEID", "Name", "City", "State", "ZipFactor", "Zip", "SchoolType", "Lat", "Long", "Undergrads", "TotalCost", "RevenueFT", "ExpensesFT", "SalaryAVG", "DefaultRate", "DebtMDN", "FamIncomeAVG", "Grads", "LoanCount", "LoanTotal")

#setting factors
coldata$SchoolType<-factor(coldata$SchoolType)
coldata$ZipFactor<-factor(coldata$ZipFactor)

c<-(ggplot(data=coldata)) #setting dataset for ggplot to variable

str(coldata) #review data structure
```

```
## 'data.frame': 2746 obs. of 21 variables:
## $ IPEDSID : int 426314 134130 176947 217235 231420 236133 100654 101189 101709 100724
## ...
## $ OPEID : int 147900 153500 245300 340400 370200 378300 100200 100300 100400 100500
## ...
## $ Name : chr "Embry-Riddle Aeronautical University-Worldwide" "University of Florida" "Central Methodist University-College of Liberal Arts and Sciences" "Johnson & Wales University-Providence" ...
## $ City : chr "Daytona Beach" "Gainesville" "Fayette" "Providence" ...
## $ State : chr "FL" "FL" "MO" "RI" ...
## $ ZipFactor : Factor w/ 10 levels "0","1","2","3",...: 4 4 7 1 3 10 4 4 4 4 ...
## $ Zip : chr "32114-3900" "32611-3150" "65248-1198" "02903-3703" ...
## $ SchoolType : Factor w/ 3 levels "Private-Nonprofit",...: 1 3 1 1 1 1 3 1 3 3 ...
## $ Lat : num 29.2 29.6 39.2 41.8 36.6 ...
## $ Long : num -81 -82.3 -92.7 -71.4 -79.4 ...
## $ Undergrads : int 10171 34523 1144 6033 884 903 5271 1928 2200 3750 ...
## $ TotalCost : int 23464 21151 38456 49107 46388 43017 23053 34317 24746 21866 ...
## $ RevenueFT : int 11302 7493 18439 16355 15982 17647 7870 11127 7918 9587 ...
## $ ExpensesFT : int 5919 16395 13097 8978 9449 9331 5546 5881 12066 9346 ...
## $ SalaryAVG : int 7290 11831 6388 9732 6038 6888 7709 6001 7756 7221 ...
## $ DefaultRate : num 0.047 0.02 0.079 0.094 0.089 0.041 0.176 0.098 0.077 0.18 ...
## $ DebtMDN : int 13625 14831 12317 16500 16737 14000 15500 14925 15750 18679 ...
## $ FamIncomeAVG: num 58583 58287 49458 61769 54227 ...
## $ Grads : int 3469 17002 NA 572 NA 392 899 844 317 440 ...
## $ LoanCount : int 34950 91026 12242 72353 10106 6438 31374 20118 10663 34246 ...
## $ LoanTotal : num 7.64e+08 2.75e+09 2.14e+08 1.37e+09 2.33e+08 ...
```

# Cleaning Dataset

In tackling a large dataset, I first narrowed my focus to columns related to financial information and loan debt. I removed columns related to academics. Additionally, the dataset included rows with fields marked "Privacy Suppressed" or NULL. These rows were removed from the dataset. In order to focus this presentation, columns related to specific groups for debt median (by gender, race, and first-generation college status) were removed. There were 6 duplicate OPEIDs for branch campuses; in these cases, rows were kept based on the brick-and-mortar location. Originally, I was reviewing a loan volume report for Award Year 2020-2021 Q4 from the Federal Student Aid Data Center. This included federal student loan disbursement information as well as the designation of an institution's school type (Federal Student Aid, 2021). In Excel, I used a vertical lookup function to add the column "SchoolType" to my College Scorecard data through an OPEID (Office of Post-secondary Education Identifier) match (ED, 2021). I also used an Excel LEFT function to parse out the first digit of the zip code. The resulting spreadsheet was saved as a .csv file titled "CollegeScorecardData."

## Data Summary and Statistics

In a walkthrough of the columns, the dataset includes IPEDSID and OPEID (identifiers/keys for the dataset). This is followed by general information: college name, city, state, geographical information and school type. The undergrads column indicates the total number of enrolled undergraduate students. This column is followed by the total cost of attendance, revenue per full-time student, expenses per full-time student, average faculty salary (per month), federal loan default rate, debt median, average family income, total number of graduate students, the count of federal loan recipients with outstanding balances, and the total dollar amount of these loan recipient balances. These balances do not include parent or graduate PLUS loans. According to the Scorecard Technical Documentation, this data is derived from the 2019-2020 award year from July 1, 2019 to June 30, 2020 (2021).

As seen in the summary below, this dataset provides a wide range of values. For example, the range of undergraduate students goes from 7 to over 98,360. When compared to the mean in this category of 4955, we can tell that the upper limit of this range is extreme. Other notable statistics include the average total cost of attendance (\$29,228), the debt median (\$12,500), the highest default rate (47.3%), the highest outstanding loan balances owed (\$37,427,769,280).

The total number of colleges in this dataset is 2,746. The total amount of outstanding loan balances for all institutions in the dataset is:

Hide

```
sum(coldata$LoanTotal)
```

```
## [1] 992323075073
```

This is a summary of the data:

Hide

```
summary(coldata)
```

```

##      IPEDSID      OPEID      Name      City
## Min.   :100654  Min.   : 100200  Length:2746  Length:2746
## 1st Qu.:152646  1st Qu.: 215325  Class :character  Class :character
## Median :188766  Median : 335200  Mode  :character  Mode  :character
## Mean   :203765  Mean    : 811960
## 3rd Qu.:221714  3rd Qu.: 815250
## Max.   :489937  Max.    :4281700
##
##      State      ZipFactor      Zip      SchoolType
## Length:2746    9      :381  Length:2746  Private-Nonprofit:1110
## Class :character 1      :345  Class :character  Proprietary      : 314
## Mode  :character 3      :292  Mode  :character  Public           :1322
##
##      0      :278
##      2      :277
##      4      :277
##      (Other):896
##      Lat      Long      Undergrads      TotalCost
## Min.   :13.43  Min.   :-159.40  Min.   : 7.0  Min.   : 6525
## 1st Qu.:34.24  1st Qu.: -97.16  1st Qu.: 896.2  1st Qu.:14932
## Median :39.19  Median : -86.79  Median : 2078.0  Median :24176
## Mean   :38.06  Mean    : -90.58  Mean    : 4954.9  Mean   :29228
## 3rd Qu.:41.80  3rd Qu.: -78.87  3rd Qu.: 5793.8  3rd Qu.:40467
## Max.   :64.86  Max.    : 144.80  Max.    :98630.0  Max.   :78555
##
##      RevenueFT      ExpensesFT      SalaryAVG      DefaultRate
## Min.   : 0  Min.   : 453  Min.   : 940  Min.   :0.00000
## 1st Qu.: 3939  1st Qu.: 5383  1st Qu.: 5732  1st Qu.:0.04800
## Median : 9448  Median : 7522  Median : 7078  Median :0.08700
## Mean   :10797  Mean    : 9131  Mean    : 7355  Mean   :0.09893
## 3rd Qu.:15192  3rd Qu.: 10450  3rd Qu.: 8650  3rd Qu.:0.14300
## Max.   :52786  Max.    :132974  Max.    :20988  Max.   :0.47300
##
##      DebtMDN      FamIncomeAVG      Grads      LoanCount
## Min.   : 1834  Min.   : 4181  Min.   : 1.0  Min.   : 80
## 1st Qu.: 7776  1st Qu.: 31085  1st Qu.: 186.2  1st Qu.: 3923
## Median :12500  Median : 45012  Median : 634.5  Median : 8241
## Mean   :12886  Mean    : 51328  Mean    : 1908.4  Mean   : 17534
## 3rd Qu.:17500  3rd Qu.: 66381  3rd Qu.: 2027.0  3rd Qu.: 19026
## Max.   :32500  Max.    :174263  Max.    :38561.0  Max.   :1429109
##      NA's      :1336
##      LoanTotal
## Min.   :6.653e+05
## 1st Qu.:4.965e+07
## Median :1.240e+08
## Mean   :3.614e+08
## 3rd Qu.:3.579e+08
## Max.   :3.743e+10
##

```

# Enrollment

In this analysis, geom\_point graphs with factors show the comparative differences between private, proprietary, and public institutions. Using points also illustrates the handful of high enrollment outliers (the statistical maximums) in each category. From the graphs, public institutions contain the most data points—meaning most institutions are public and service many students. In terms of enrollment, each category has a few outliers that are dominating enrollment numbers. Despite most public universities having generally greater enrollment numbers than private and proprietary colleges, there are two major outliers in the private category that have the highest numbers of students, both of which have strong online presences: Western Governors University and Southern New Hampshire University. The high number of graduate students at proprietary colleges takes them in a lead over public colleges in the total student count.

A point-to-point graph displays the geographical breakdown of student enrollment. While the “0” Zip code region/New England is an expected area of high enrollment with its concentration of colleges, the “8” Zip code region (comprised of Arizona, Colorado, Wyoming, Utah, New Mexico, and Nevada) took an unexpected lead as the region with the highest student enrollment. These colleges also tend to have large online presences, such as University of Phoenix and, again, Western Governors University (the statistical maximum).

Top 5 Undergraduate Enrollment:

Hide

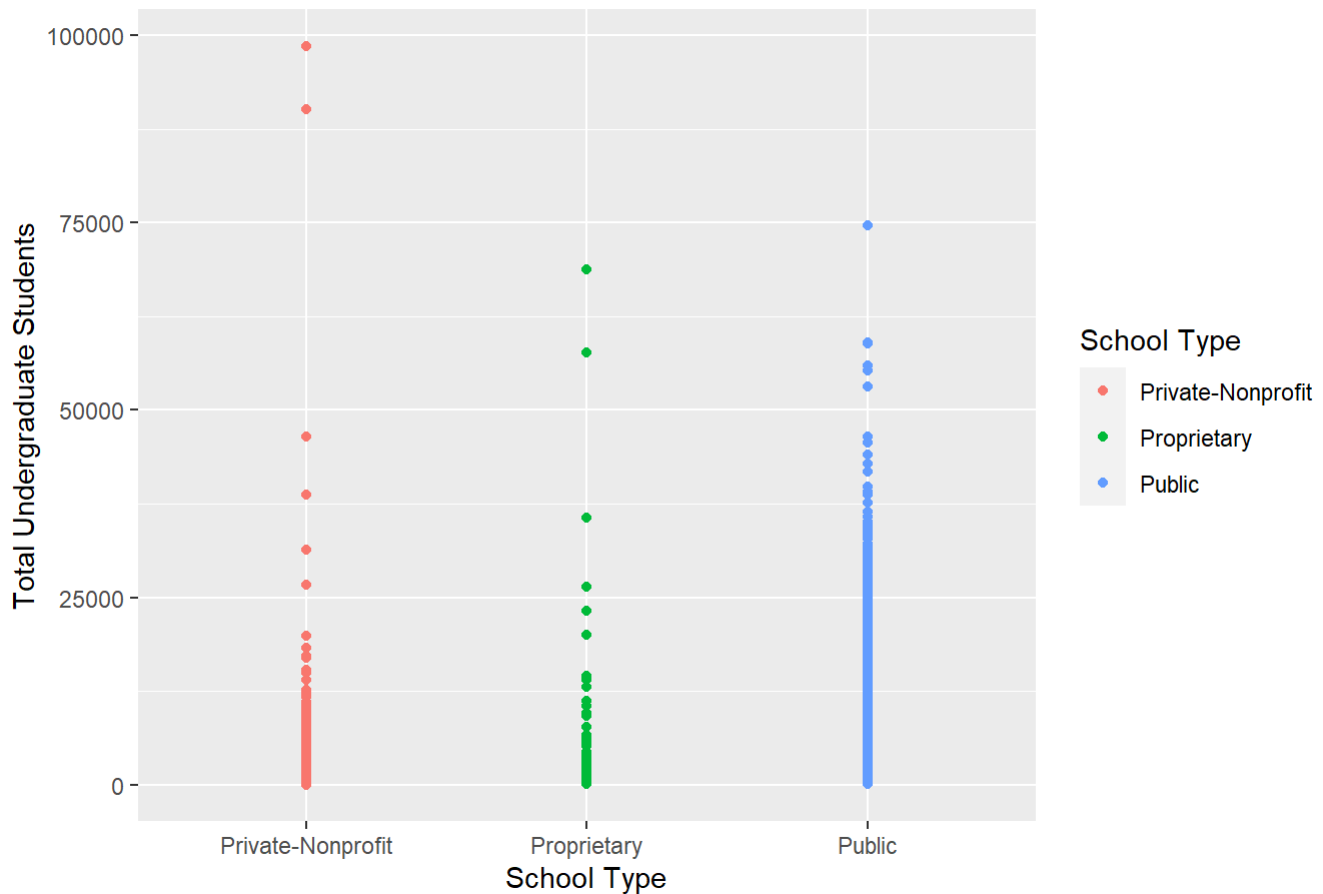
```
#sorting to display a dataframe of colleges with the most undergrads
sortug <- coldata[order( -coldata$Undergrads) , ]
head(sortug[,c(3,5, 8,11)],5)
```

##	Name	State	SchoolType
## 2570	Western Governors University	UT	Private-Nonprofit
## 945	Southern New Hampshire University	NH	Private-Nonprofit
## 1374	Pennsylvania State University-Main Campus	PA	Public
## 2341	University of Phoenix-Arizona	AZ	Proprietary
## 2167	Ivy Tech Community College	IN	Public
##	Undergrads		
## 2570	98630		
## 945	90196		
## 1374	74630		
## 2341	68833		
## 2167	58978		

Hide

```
#Total Undergrads
c + geom_point(aes(x=SchoolType, y=Undergrads, color=SchoolType)) + xlab("School Type") + ylab(
"Total Undergraduate Students") + labs(color="School Type") + ggtitle("Count of Undergraduate St
udents By School Type")
```

## Count of Undergraduate Students By School Type



Top 5 Graduate Enrollment:

Hide

```
#sorting for highest number of grads
sortgrad <- coldata[order( -coldata$Grads) , ]
head(sortgrad[,c(3,5, 8,19)],5)
```

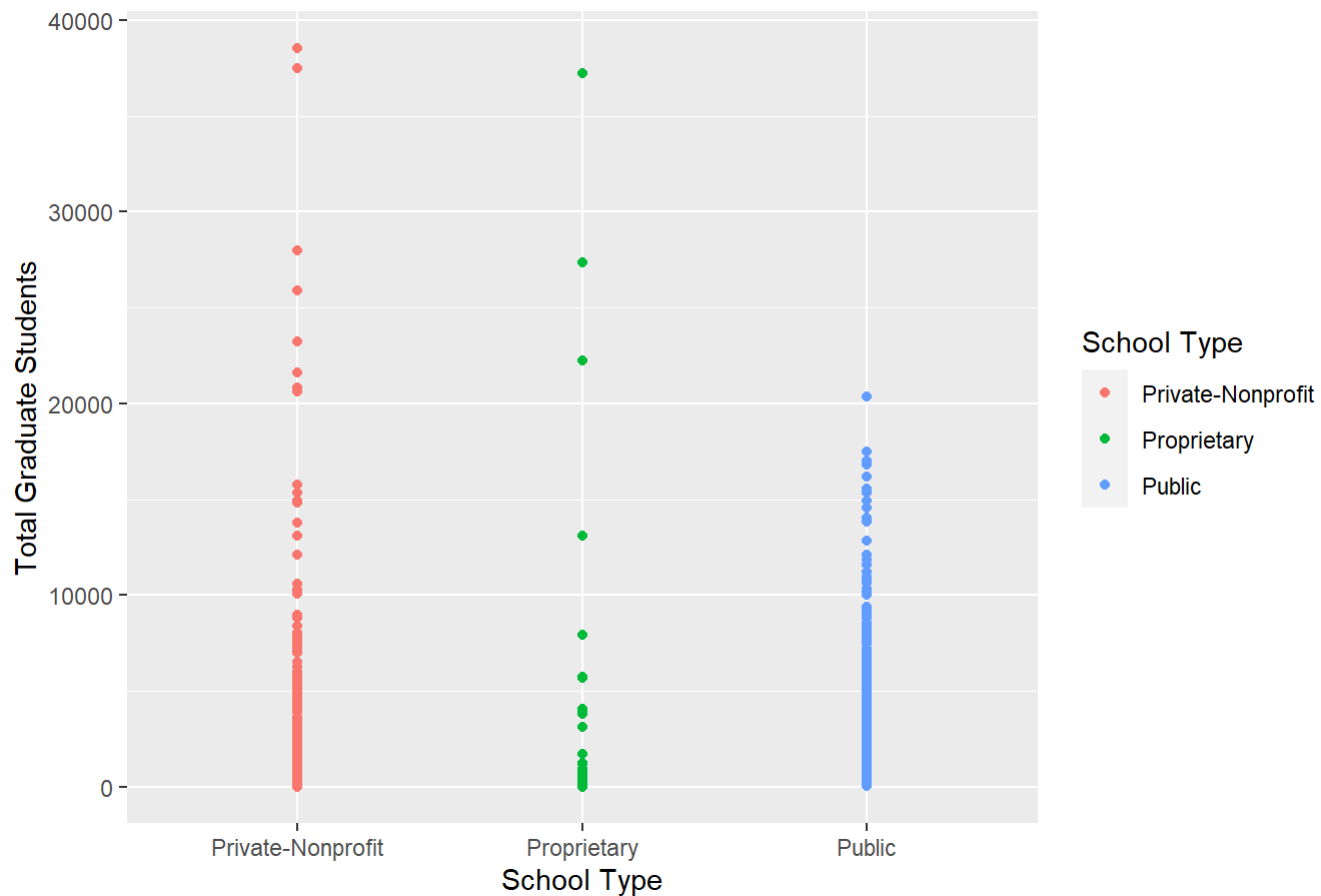
##	Name	State	SchoolType	Grads
## 2314	Liberty University	VA	Private-Nonprofit	38561
## 2570	Western Governors University	UT	Private-Nonprofit	37509
## 45	Grand Canyon University	AZ	Proprietary	37214
## 204	University of Southern California	CA	Private-Nonprofit	27970
## 2565	Capella University	MN	Proprietary	27356

Hide

```
#Total Grads
c + geom_point(aes(x=SchoolType, y=Grads, color=SchoolType)) + xlab("School Type") + ylab("Total Graduate Students") + labs(color="School Type") + ggtitle("Count of Graduate Students By School Type")
```

```
## Warning: Removed 1336 rows containing missing values (geom_point).
```

## Count of Graduate Students By School Type



Hide

```
#changing NA values in Total Grads to 0
coldata$Grads[is.na(coldata$Grads)]<-0

#Creating a column for total students, undergraduate and graduate combined
coldata$TotalStudents<-(coldata$Undergrads + coldata$Grads)
#sorting total students
sorttotal <- coldata[order( -coldata$TotalStudents) , ]
```

Top 5 Total Enrollment:

Hide

```
head(sorttotal[,c(3,5, 8,22)],5)
```

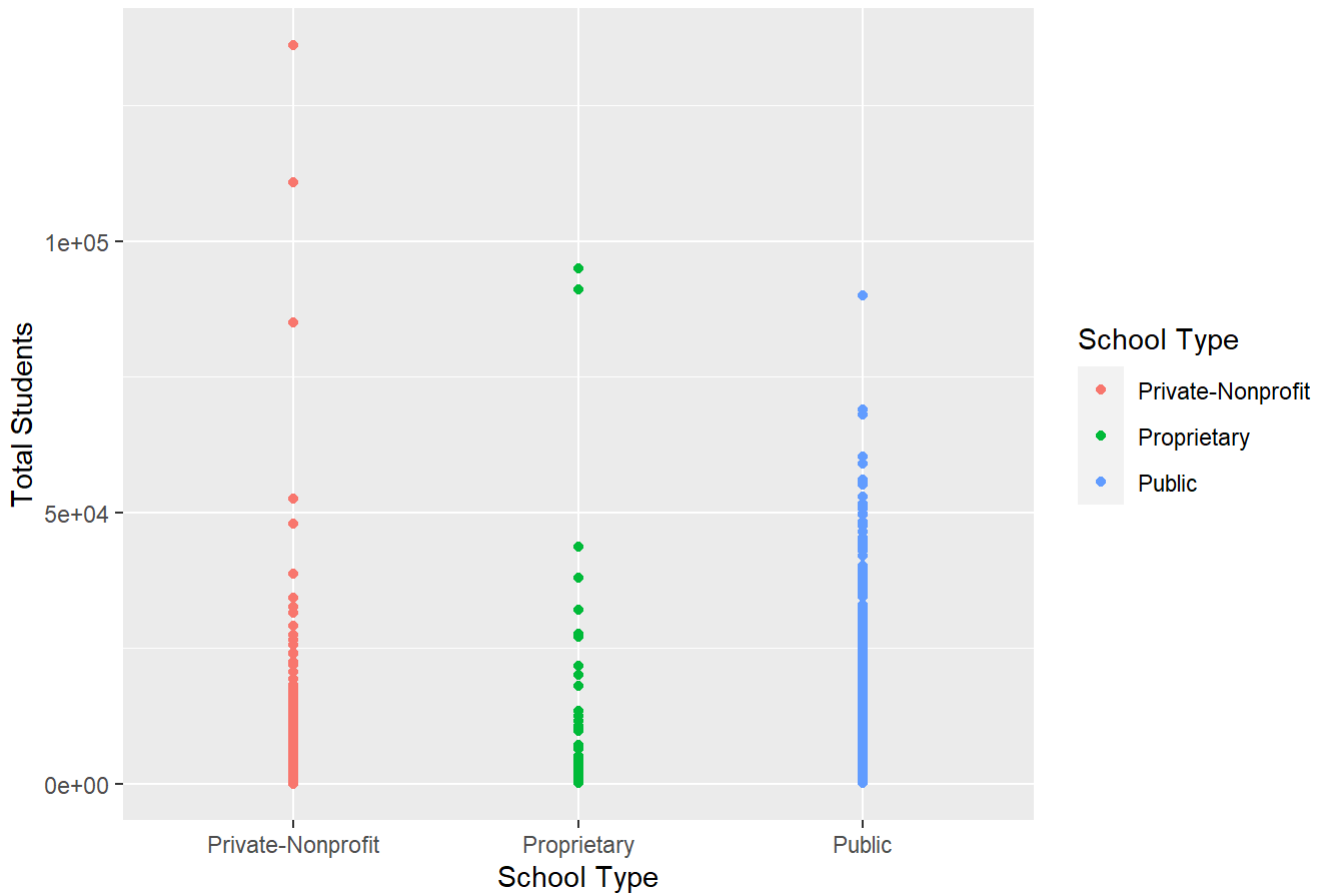


```
##           Name State      SchoolType
## 2570      Western Governors University  UT Private-Nonprofit
## 945      Southern New Hampshire University  NH Private-Nonprofit
## 45      Grand Canyon University  AZ      Proprietary
## 2341     University of Phoenix-Arizona  AZ      Proprietary
## 1374 Pennsylvania State University-Main Campus  PA      Public
##      TotalStudents
## 2570      136139
## 945      110808
## 45      94843
## 2341     91072
## 1374     89958
```

Hide

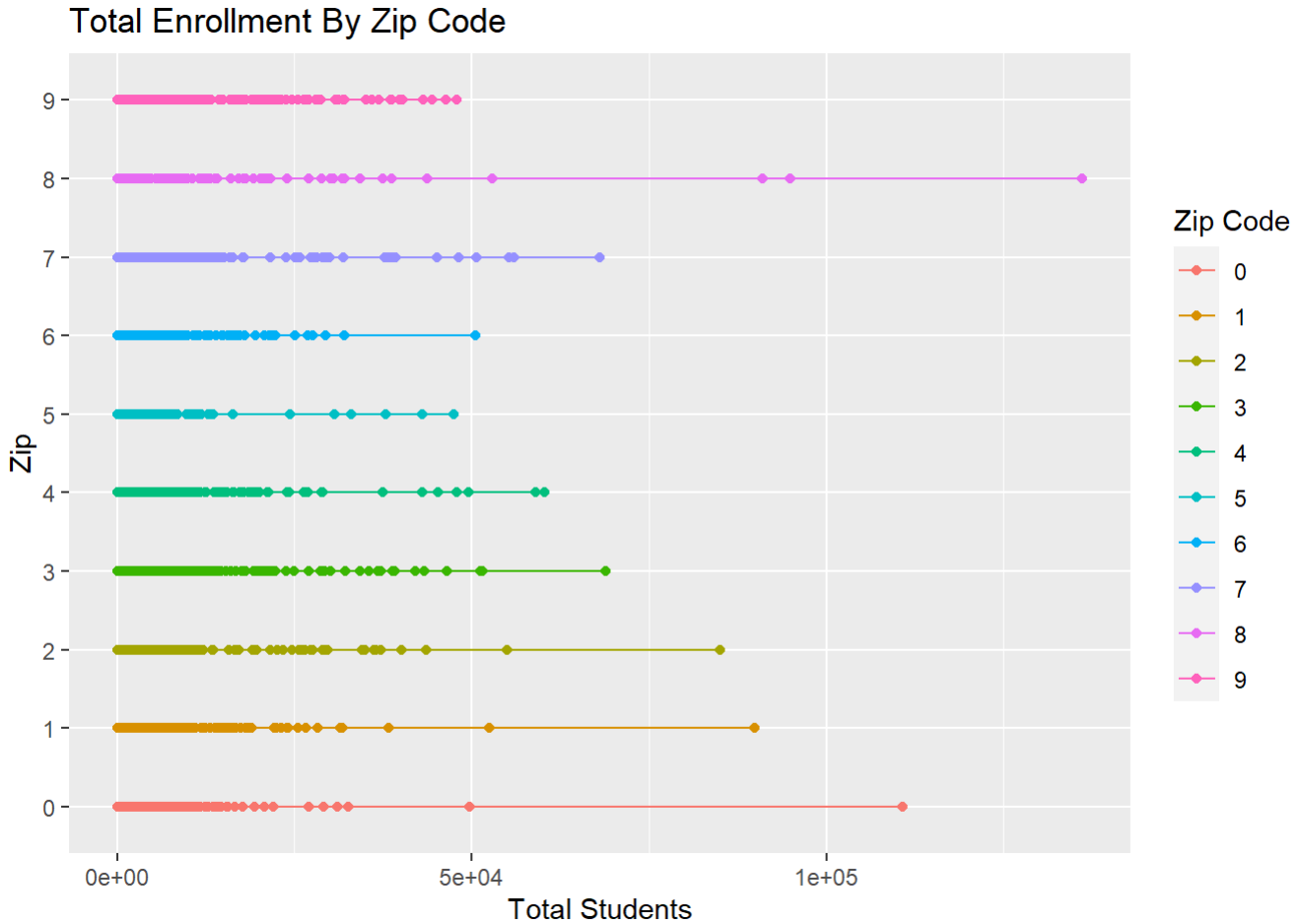
```
#Total Students
c + geom_point(aes(x=SchoolType, y=coldata$TotalStudents, color=SchoolType)) + xlab("School Type") + ylab("Total Students") + labs(color="School Type") + ggtitle("Count of Total Students by School Type")
```

Count of Total Students by School Type



Hide

```
c + geom_point(aes(x=ZipFactor, y=coldata$TotalStudents)) +
  geom_segment( aes(x=ZipFactor, xend=ZipFactor, y=0, yend=coldata$TotalStudents, col=coldata$Zip
Factor)) +
  geom_point(aes(x=ZipFactor, y=coldata$TotalStudents, color=coldata$ZipFactor)) + coord_flip()
+ xlab("Zip") + ylab("Total Students") + ggtitle("Total Enrollment By Zip Code") + labs(col =
"Zip Code")
```



# College Costs

In examining the total cost of attendance, the dataset refers to both billable and non-billable elements of a student's educational experience, such as tuition, room, board, transportation, and miscellaneous expenses (ED, 2021). In the illustrations below, we see that private colleges maintain the highest costs whereas public institutions are the most cost-effective.

The geographical comparison presents zip code region "8" (which had the highest student enrollment) as the least expensive region for students. The Northeast (zip code regions "0" and "1") have a large concentration of colleges with comparatively high costs. The jitter function on the graph illustrates the various clusters of colleges that fall outside the quartile ranges.

Top 5 Cost Examples:

Hide

```
sortcost <- coldata[order( -coldata$TotalCost) , ]  
head(sortcost[,c(3,5,8,12)],5)
```

```
##                Name State      SchoolType  
## 445      University of Chicago    IL Private-Nonprofit  
## 113      Harvey Mudd College      CA Private-Nonprofit  
## 1013 Columbia University in the City of New York NY Private-Nonprofit  
## 1390      University of Pennsylvania PA Private-Nonprofit  
## 1136      Duke University         NC Private-Nonprofit  
##      TotalCost  
## 445      78555  
## 113      76953  
## 1013     76907  
## 1390     75303  
## 1136     75105
```

Bottom 5 Cost Examples:

Hide

```
tail(sortcost[,c(3,5,8,12)],5)
```

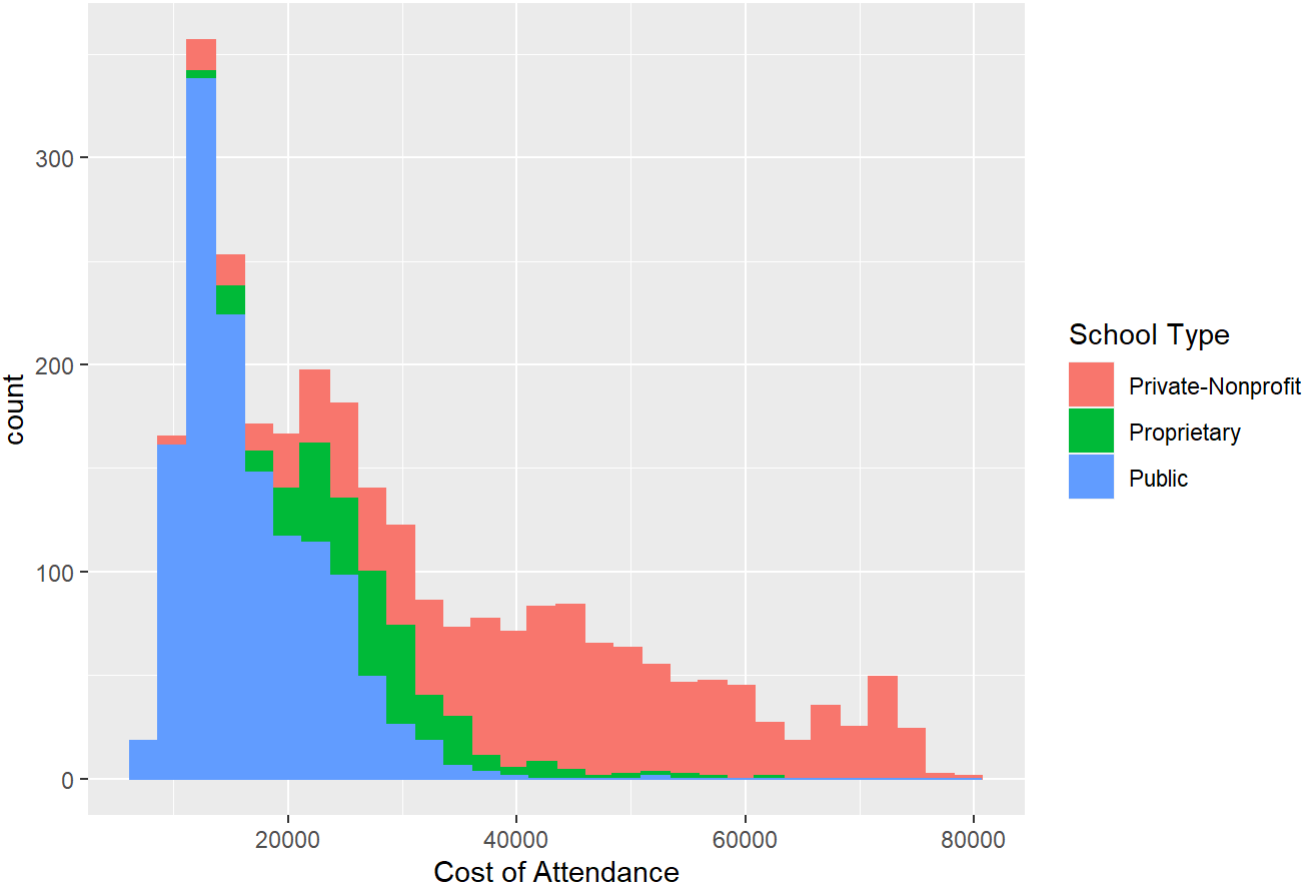
```
##                Name State SchoolType TotalCost  
## 1216 Lorain County Community College    OH    Public    7900  
## 1966      College of the Mainland      TX    Public    7871  
## 2463      St Charles Community College MO    Public    7364  
## 2281 Chattahoochee Valley Community College AL    Public    7338  
## 288      Indian River State College    FL    Public    6525
```

Hide

```
c+ geom_histogram(aes(x=TotalCost, color=SchoolType, fill=SchoolType)) +xlab("Cost of Attendance") + ggtitle("Total Cost of Attendance by School Type") + labs(color="School Type", fill="School Type")
```

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

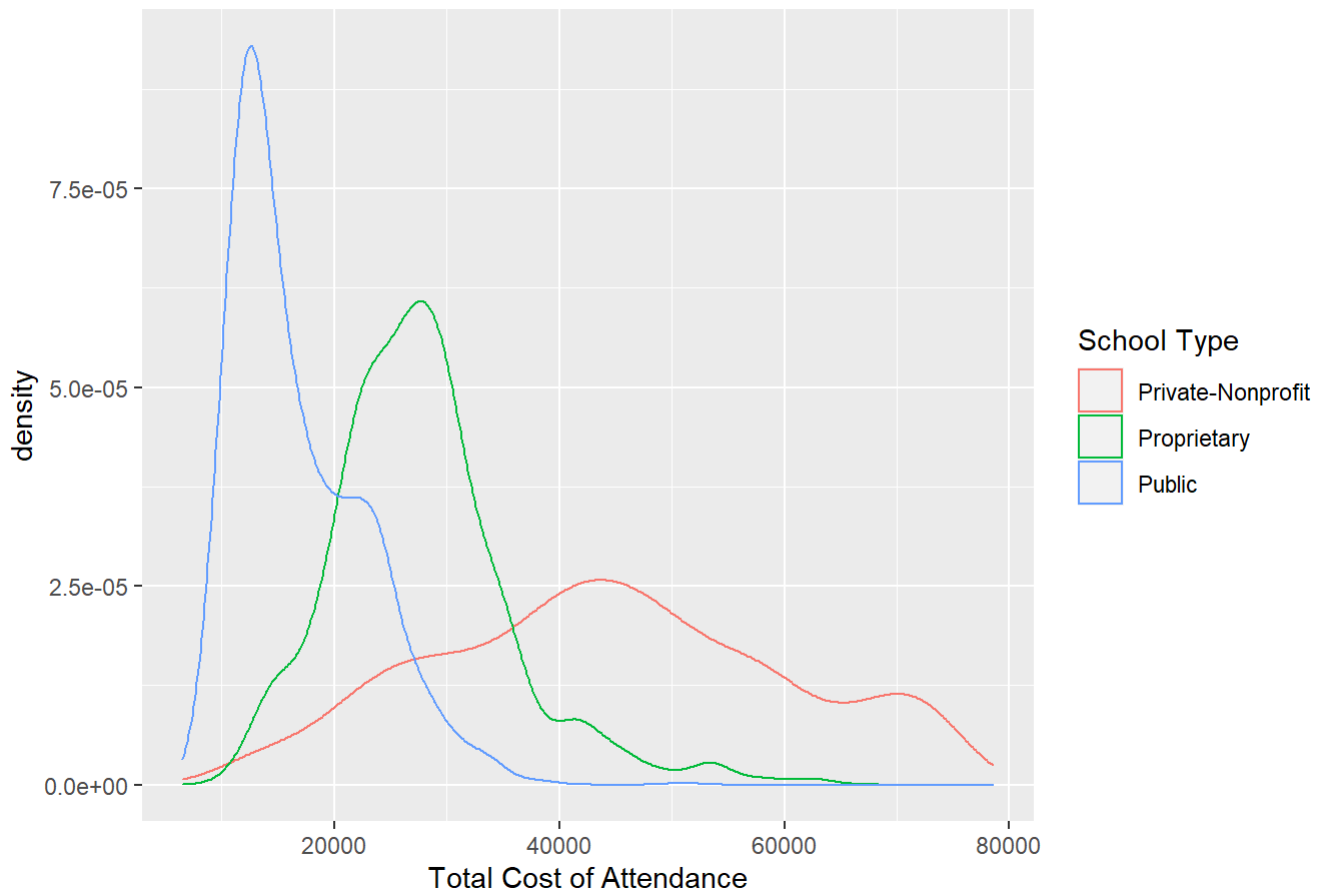
Total Cost of Attendance by School Type



Hide

```
c+ geom_density(aes(x=TotalCost, color=SchoolType)) + ggtitle("Total Cost by School Type") + xlab("Total Cost of Attendance") + labs(color="School Type")
```

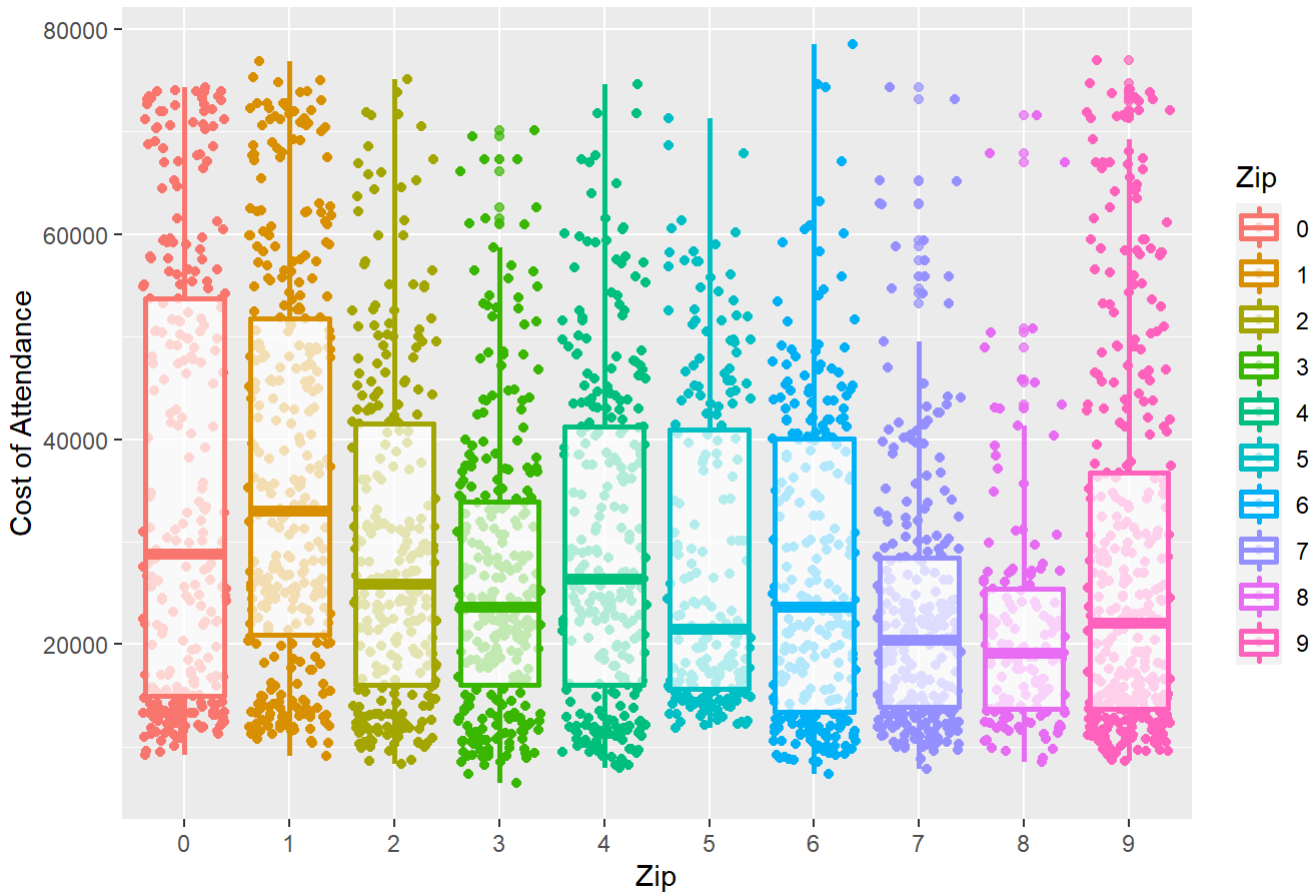
## Total Cost by School Type



Hide

```
costplot<-ggplot(data=coldata, aes(x=ZipFactor, y=TotalCost, color=ZipFactor))
costplot + geom_jitter() + geom_boxplot(size=1, alpha = 0.7) + xlab("Zip") + ylab("Cost of Attendance") + labs(color="Zip") + ggtitle("Total Cost of Attendance by Location")
```

Total Cost of Attendance by Location



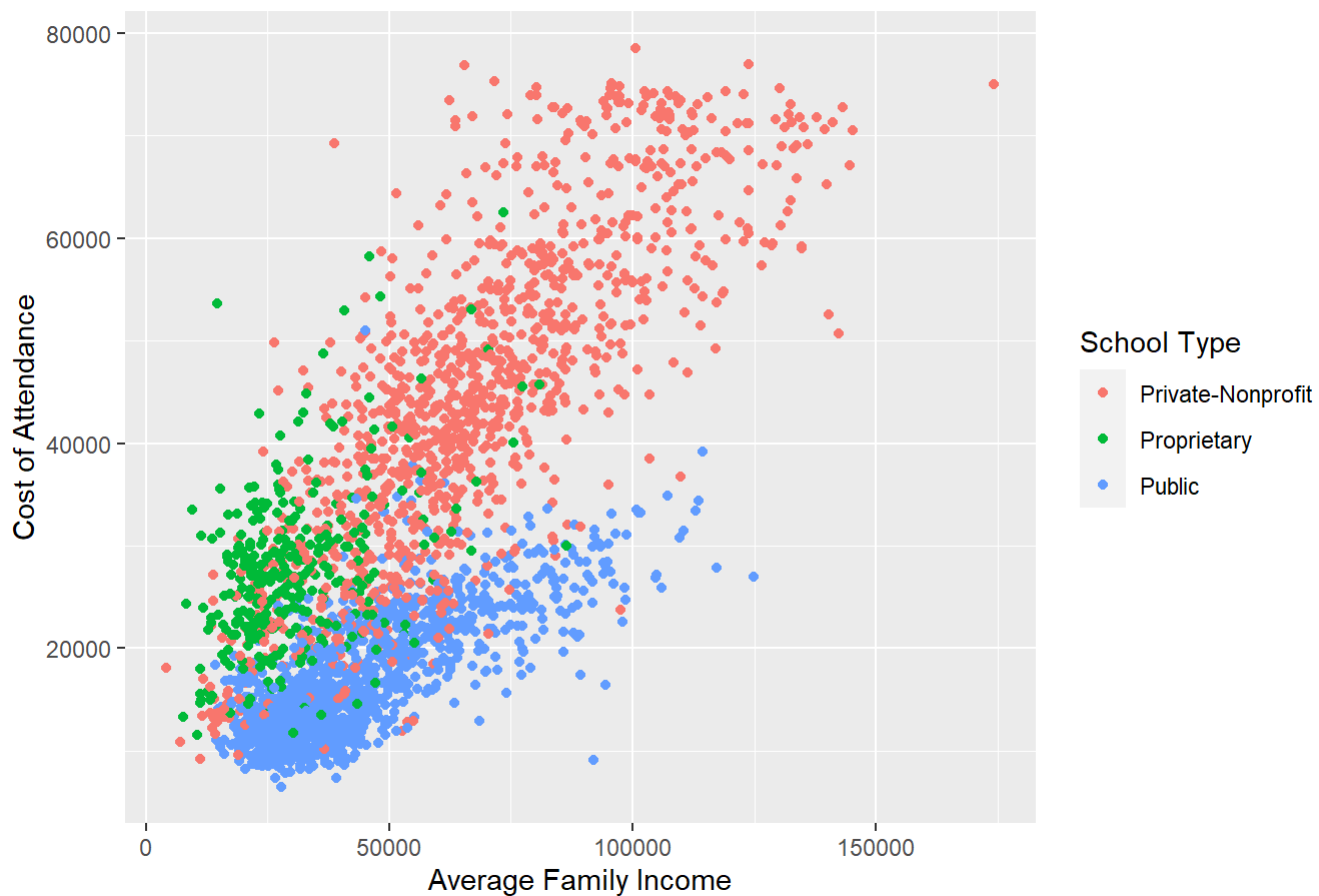
## Clustering Example

The graph below shows the relationship between average family income and college cost of attendance. Families with higher incomes are enrolling in more expensive private colleges whereas lower income households are enrolling in public universities or proprietary institutions. In clustering this graph, the end result shows a breakdown of colleges catered to three separate economic tiers: lower, middle, and upper class.

Hide

```
c+ geom_point(aes(x=FamIncomeAVG, y=TotalCost, color=SchoolType)) +xlab("Average Family Income")  
+ ylab("Cost of Attendance") +ggtitle("Cost vs. Average Family Income") + labs(color="School Type")
```

## Cost vs. Average Family Income



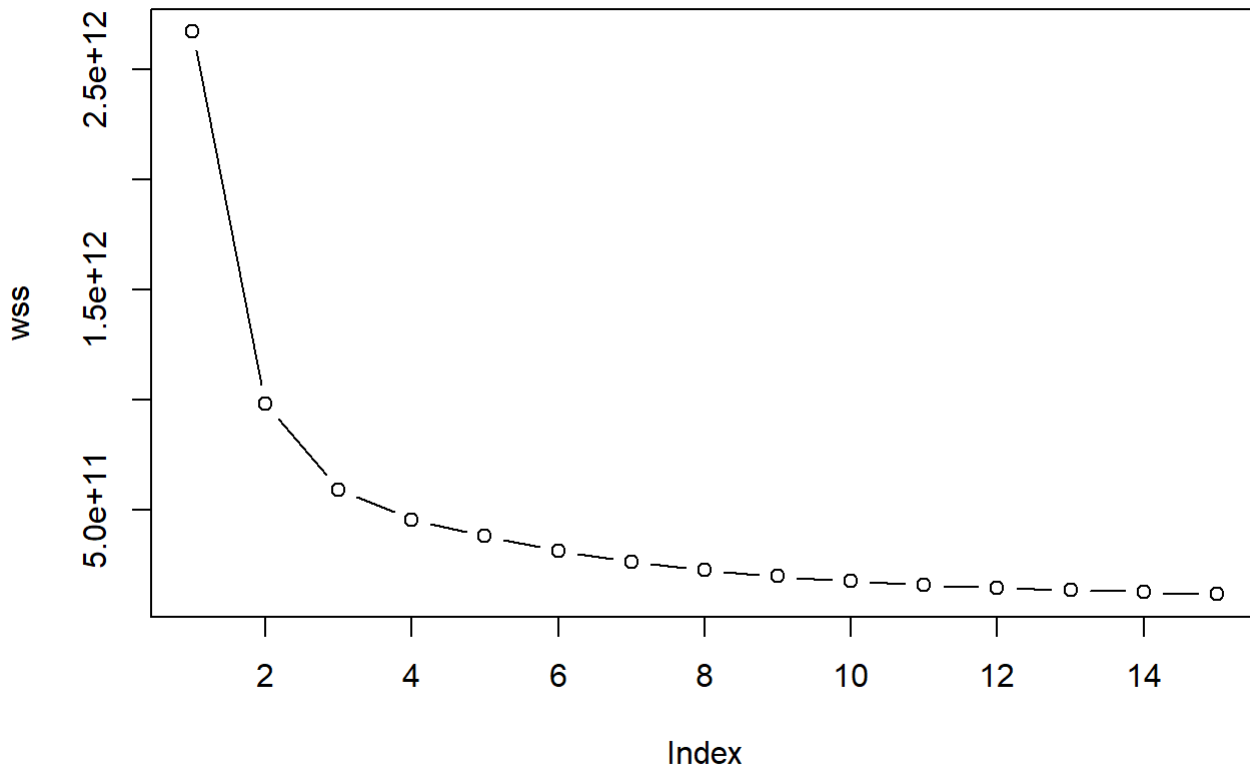
Hide

```
#clustering the data  
matcol<-cbind(coldata$FamIncomeAVG, coldata$TotalCost)  
  
wss <- rep(0,15)  
for (k in 1:15)  
  wss[k] <- sum( kmeans(matcol,centers=k, nstart=50)$withinss)
```

```
## Warning: did not converge in 10 iterations
```

Hide

```
plot(wss, type="b")
```



Hide

```
kmeans(matcol ,centers=3) #"elbow" of plot at 3
```



```
## K-means clustering with 3 clusters of sizes 880, 383, 1483
```

```
##
```

```
## Cluster means:
```

```
##      [,1]      [,2]
```

```
## 1 62643.26 35274.26
```

```
## 2 99277.62 58414.77
```

```
## 3 32230.62 18102.86
```

```
##
```

```
## Clustering vector:
```

```
## [1] 1 1 1 1 1 1 3 3 1 3 3 2 1 3 3 1 1 3 3 3 3 1 3 1 3 1 2 3 3 2 3 3 3 1 2 1 1
```

```
## [38] 1 3 1 3 3 3 3 1 3 3 3 3 1 1 1 3 3 3 1 3 3 3 1 1 1 3 1 3 2 1 3 1 3 3 1 3 1
```

```
## [75] 3 3 3 1 1 3 2 3 1 1 2 2 2 1 3 3 3 3 3 3 2 3 3 3 3 3 1 3 3 3 1 3 3 3 3 2 3
```

```
## [112] 2 2 2 2 2 3 1 3 1 3 1 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 1 1 3 3 1 3 3 3 3
```

```
## [149] 3 3 3 3 1 1 3 3 1 3 3 2 3 1 1 1 1 3 3 2 3 3 3 3 3 3 2 3 1 3 3 3 3 3 3 1
```

```
## [186] 3 1 3 2 2 3 3 1 1 1 1 1 1 1 1 2 2 2 2 1 3 3 1 2 2 1 3 3 3 2 2 1 1 1 3 1 3
```

```
## [223] 3 3 3 1 1 3 2 2 1 1 1 2 1 2 3 3 1 2 3 3 3 2 2 1 1 2 1 1 2 2 1 2 1 1 2 1 2
```

```
## [260] 1 2 3 3 2 2 1 3 3 1 3 1 1 3 3 3 3 3 3 3 1 3 3 3 2 2 1 3 3 1 3 3 3 3 3 1 3
```

```
## [297] 1 3 3 3 3 2 3 3 1 3 3 3 3 2 3 2 3 2 1 3 1 3 1 1 2 3 1 3 3 3 3 1 2 3 2 3 1
```

```
## [334] 3 3 3 1 1 1 1 3 1 1 1 3 1 1 3 1 3 1 1 1 1 3 1 1 1 1 3 1 1 3 3 3 3 3 1 1 3
```

```
## [371] 3 3 3 1 3 1 1 2 1 3 3 3 3 2 3 3 3 3 3 3 3 3 1 1 1 3 1 1 3 1 1 3 1 1 1 1 1 3
```

```
## [408] 3 2 3 1 2 3 2 1 1 3 2 3 1 1 1 3 3 2 1 1 2 1 3 1 3 1 1 1 3 2 1 1 1 1 1 1 1 3
```

```
## [445] 2 1 3 1 1 2 3 1 1 1 2 2 1 2 2 1 1 2 1 1 3 1 1 2 3 1 3 3 3 3 1 1 1 1 2 1 2
```

```
## [482] 1 3 1 2 2 1 2 2 3 2 1 1 3 2 1 2 2 3 2 2 3 3 3 1 1 2 1 1 2 2 3 3 1 1 3 2 2
```

```
## [519] 1 1 1 1 1 1 2 1 3 3 1 1 1 3 3 3 3 3 3 3 3 3 1 1 3 1 3 3 3 3 1 1 1 1 3 1 1
```

```
## [556] 3 1 3 1 1 1 1 1 1 3 1 3 1 2 3 1 1 1 2 1 1 1 1 3 1 3 1 3 1 3 3 2 1 1 3 3 3
```

```
## [593] 3 3 3 3 1 1 1 2 3 1 3 1 1 1 3 3 3 3 2 1 3 3 3 1 3 3 2 1 1 3 2 2 2 1 3 1 2
```

```
## [630] 2 2 1 1 1 3 3 1 3 3 1 3 3 3 1 2 3 3 1 2 2 2 1 2 3 1 2 2 2 1 3 1 1 3 2 2 2
```

```
## [667] 1 2 1 2 2 2 1 1 2 2 2 2 2 2 1 2 2 2 1 2 2 2 1 3 2 1 1 2 1 1 1 3 3 3 3 3 3
```

```
## [704] 3 3 3 3 3 2 1 2 1 1 1 1 1 1 1 2 2 2 2 3 3 2 2 2 1 1 2 2 2 2 1 2 2 2 2 2 2
```

```
## [741] 2 1 2 3 1 1 3 2 1 1 1 3 1 1 3 2 3 3 3 1 3 1 3 3 2 3 2 3 3 3 1 1 1 1 1 1 3
```

```
## [778] 3 3 3 1 3 3 1 1 3 1 1 3 1 3 1 1 1 2 3 3 3 3 1 3 1 3 1 1 3 2 2 1 1 2 2 1 3
```

```
## [815] 2 1 3 3 2 1 1 3 1 1 1 3 1 3 1 1 2 1 2 1 3 1 1 3 1 3 3 3 3 3 3 3 3 3 3 3 3
```

```
## [852] 3 2 1 3 3 1 3 3 3 3 3 3 1 3 3 1 3 1 3 1 3 1 1 1 1 3 3 3 1 3 1 1 3 3 3 3 1
```

```
## [889] 3 3 1 1 3 2 1 1 1 1 2 1 3 2 1 1 3 2 1 1 2 1 2 1 3 3 3 1 1 3 1 1 1 1 1 2 2
```

```
## [926] 1 1 1 1 1 1 3 3 3 1 1 1 1 1 3 2 2 2 1 1 1 3 3 1 2 2 2 1 3 3 1 1 1 2 1 1 1
```

```
## [963] 1 3 3 2 1 1 1 3 1 1 2 1 2 1 2 2 3 1 3 3 1 3 3 3 3 3 3 3 1 1 1 2 3 1 1 3 3
```

```
## [1000] 3 3 3 3 3 3 3 3 3 2 2 1 1 2 2 1 1 2 1 1 2 2 2 2 2 2 2 1 1 2 2 2 2 1 2 1 2 2
```

```
## [1037] 1 3 2 2 3 1 1 1 2 1 2 1 1 2 1 2 2 1 2 1 1 3 2 2 2 2 1 2 1 1 2 2 1 1 1 1 1
```

```
## [1074] 3 1 1 1 1 1 1 1 1 1 2 1 3 1 1 1 1 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

```
## [1111] 3 2 1 2 2 1 2 2 3 2 2 1 2 3 1 1 1 1 3 1 1 1 3 2 3 2 1 3 2 3 1 1 1 2 3 1 1
```

```
## [1148] 3 3 1 1 1 1 3 3 3 1 3 1 3 1 1 3 3 1 1 1 3 2 1 1 1 1 3 1 1 1 3 1 1 1 3 1 1
```

```
## [1185] 1 1 1 1 3 1 1 1 1 2 2 3 3 2 1 1 1 2 2 1 3 1 2 1 3 1 1 2 1 2 1 3 1 1 2 2 1
```

```
## [1222] 1 1 1 2 2 1 1 2 2 1 3 1 1 1 2 1 1 1 3 1 2 2 1 3 3 1 3 1 1 3 3 3 3 3 3 3 3
```

```
## [1259] 3 3 1 1 2 3 3 1 3 3 3 3 3 3 1 1 2 3 3 3 3 2 3 2 2 3 1 1 1 1 1 1 2 3 3 2 3
```

```
## [1296] 3 3 3 3 1 2 1 2 1 1 2 3 1 2 2 2 3 3 1 2 1 1 1 2 3 2 2 2 2 2 1 2 2 1 1 2 1
```

```
## [1333] 3 3 2 1 1 1 2 1 2 3 2 1 1 2 2 3 1 3 1 1 2 2 2 1 1 2 1 3 2 1 1 1 1 1 1 1 1
```

```
## [1370] 1 1 1 1 1 2 1 2 1 1 1 2 1 2 2 2 2 2 1 1 2 2 2 2 2 2 1 2 1 1 1 1 2 2 2 3 3
```

```
## [1407] 2 2 2 2 3 1 1 3 1 1 1 3 2 3 1 1 1 1 1 1 2 1 1 3 1 1 2 3 1 2 1 1 1 1 3 3 1
```

```
## [1444] 2 2 1 1 1 1 1 1 1 1 1 1 3 2 3 1 1 3 2 1 1 3 1 1 1 1 3 1 3 1 1 1 3 3 1 1 2
```

```
## [1481] 3 1 1 1 1 1 1 1 3 2 2 1 2 3 3 1 2 2 3 1 3 3 3 1 3 1 3 1 3 3 3 1 3 3 3 1 1
```

```
## [1518] 3 1 3 3 3 1 1 1 3 1 1 3 1 3 3 3 3 3 3 3 2 1 3 3 1 3 3 2 3 1 1 3 1 2 1 1 1
```

```
## [1555] 3 3 3 3 3 1 1 3 2 3 3 3 1 3 1 1 3 2 3 2 3 1 3 1 3 3 3 3 1 3 3 1 3 3 3 1 1
```

```

## [1592] 1 3 3 1 2 1 2 3 1 2 2 2 2 1 1 2 2 2 3 1 1 1 3 2 1 1 1 2 2 1 1 3 3 1 1 2 2
## [1629] 1 2 2 1 1 2 2 1 3 1 2 2 3 3 3 3 3 3 3 1 2 3 3 1 3 3 1 3 3 2 3 3 3 3 3 1
## [1666] 3 3 2 2 3 3 3 1 3 2 1 1 1 3 1 2 2 3 1 1 3 3 1 1 3 1 3 1 1 3 1 1 3 1 3 1 1
## [1703] 1 2 1 2 2 3 1 1 1 2 1 2 3 2 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 1 3
## [1740] 3 3 3 3 3 3 3 1 3 3 3 3 1 1 1 1 1 1 2 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [1777] 3 3 3 1 1 3 3 3 3 3 3 3 1 1 3 3 3 3 3 3 3 3 3 3 1 2 3 3 1 3 3 3 3 3 3 3
## [1814] 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3
## [1851] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [1888] 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 1 1 3 1 1 1 1 3 1 1 3 1 1 1 3 3 3 3 3 3 3
## [1925] 3 3 2 3 3 3 3 3 3 1 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 1 3 3 1 3 3 3 3 3 3 3
## [1962] 1 1 3 1 3 3 3 3 1 3 3 3 1 3 3 3 3 3 3 1 3 3 3 3 1 3 3 3 1 3 1 3 3 1 3 3
## [1999] 3 1 1 1 2 1 3 3 3 1 1 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3
## [2036] 1 3 3 3 1 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [2073] 3 3 3 3 3 1 3 3 3 3 3 3 3 3 1 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 2 1 3 3
## [2110] 3 3 3 1 3 3 3 3 3 3 3 3 3 1 3 3 1 1 3 1 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3
## [2147] 3 3 1 3 3 3 3 3 3 3 3 3 3 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [2184] 3 3 1 1 3 3 3 2 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 1 3 3 3 3 3
## [2221] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 1 3 3 3 3 1 3 3 3 3 1 3 3 3 2
## [2258] 3 3 1 1 3 3 3 3 3 3 3 3 3 3 3 1 3 1 3 3 3 3 3 3 3 2 3 3 3 1 3 3 3 3 1
## [2295] 2 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 1 3 3 3 1 3 1 2 1 1 1 3 3 3 1 3 1 1
## [2332] 1 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 1 3 3 3 3 3 3 3 1 3 3 3 3 3 3 1 3 3 2 3
## [2369] 1 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 1 3 3 1 3 3 3 3 3 3
## [2406] 1 3 1 3 3 3 3 3 3 3 1 1 3 3 3 3 3 3 1 3 3 3 3 1 3 3 3 3 3 3 1 3 1 1 3 3
## [2443] 3 3 3 3 2 3 3 1 3 3 3 3 3 3 3 3 3 3 1 3 2 3 3 3 3 3 3 1 3 3 1 3 3 3 3
## [2480] 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 1 1 3 3 3 3 3 1
## [2517] 3 1 3 3 3 3 3 3 3 3 1 3 3 3 3 3 1 3 1 3 3 1 3 3 3 3 3 3 1 3 3 3 3 3 3
## [2554] 1 3 3 3 1 3 1 3 1 3 3 3 3 3 3 3 3 3 1 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 1
## [2591] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3
## [2628] 3 3 3 3 1 3 3 3 3 3 3 2 2 3 3 3 1 3 3 3 3 1 3 3 1 3 3 3 3 3 3 3 3 1 3 2 3
## [2665] 3 3 3 3 3 3 1 3 3 3 3 3 3 1 3 3 3 3 1 3 3 3 3 1 3 3 3 1 3 3 1 3 3 3 3
## [2702] 3 3 3 3 1 3 1 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [2739] 1 3 3 3 3 1 3
##
## Within cluster sum of squares by cluster:
## [1] 220860130863 168599298773 205737186965
## (between_SS / total_SS = 77.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"    "tot.withinss"
## [6] "betweenss"   "size"         "iter"         "ifault"

```

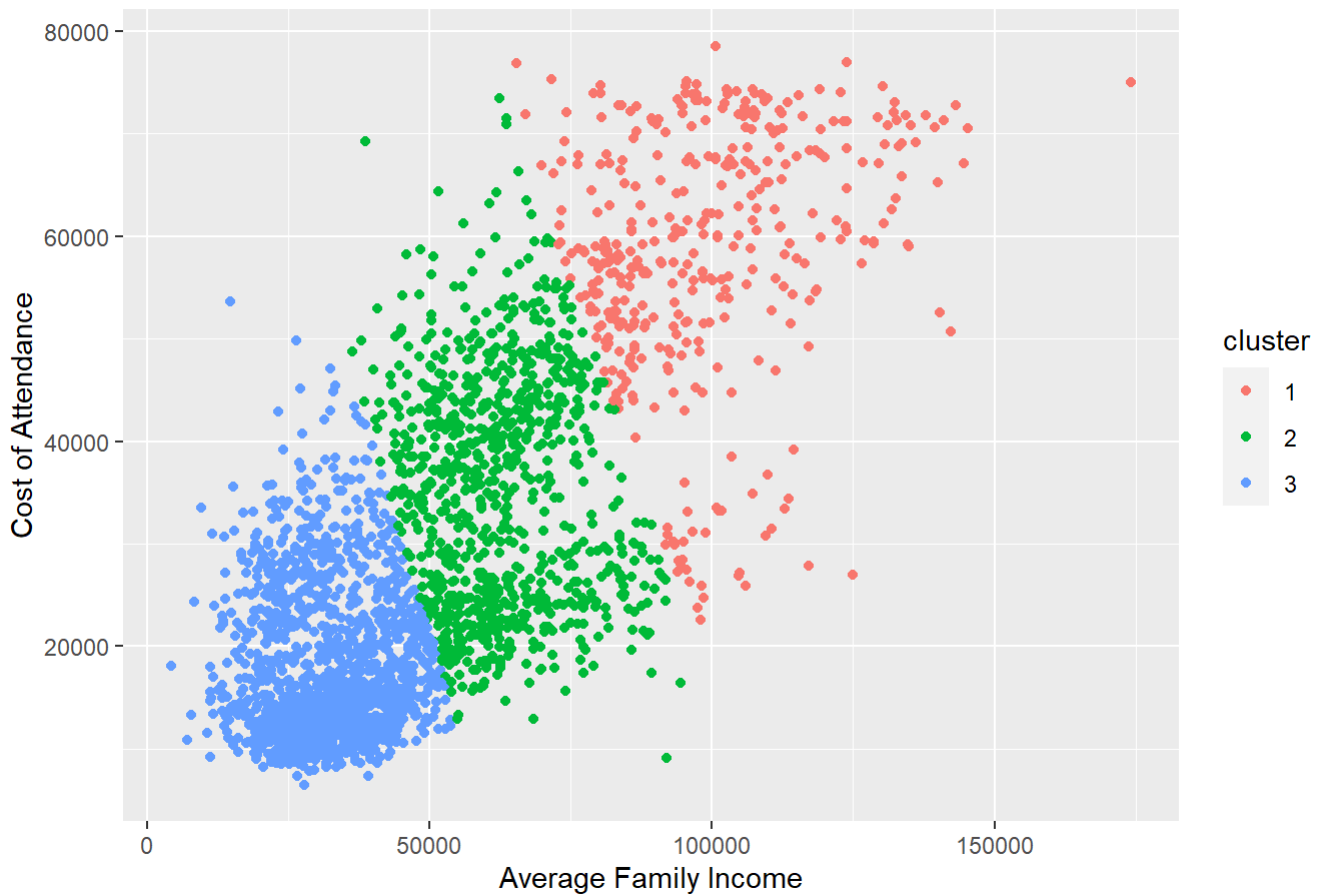
Hide

```

km = kmeans(matcol,centers=3)$cluster
coldata$cluster <- factor( km)
qplot(data=coldata, x=FamIncomeAVG, y=TotalCost, color=cluster, main="Clustered Cost vs. Average
Family Income") + xlab("Average Family Income") + ylab("Cost of Attendance")

```

## Clustered Cost vs. Average Family Income



## Revenue & Expenses

The following graph shows the breakdown of revenues and expenses per full-time student based on college type. Here, non-profit private colleges collect the most revenue compared to public universities or proprietary (for-profit) colleges. However, proprietary colleges have less expenses per student by comparison. Private colleges appear to have the highest expenses. In the case of public colleges, on average, the expenses are exceeding revenue; these colleges are supported by state taxpayer funding. By zip code, the locations with higher costs of attendance (as displayed previously) have higher revenues.

Top 5 Revenues per Full-time Student:

Hide

```
sortrev <- coldata[order( -coldata$RevenueFT) , ]  
head(sortrev[,c(3,5,8,13)],5)
```

##	Name	State	SchoolType	RevenueFT
## 114	Pitzer College	CA	Private-Nonprofit	52786
## 623	Bates College	ME	Private-Nonprofit	47882
## 1010	Colgate University	NY	Private-Nonprofit	45259
## 1601	Middlebury College	VT	Private-Nonprofit	45001
## 2464	Landmark College	VT	Private-Nonprofit	44028

Bottom 5 Revenues per Full-time Student:

Hide

```
tail(sortrev[,c(3,5,8,13)],5)
```

```
##                               Name State      SchoolType
## 388  City Colleges of Chicago-Malcolm X College  IL          Public
## 2596          Copper Mountain Community College  CA          Public
## 2272          American University of Puerto Rico  PR Private-Nonprofit
## 1121                               Webb Institute  NY Private-Nonprofit
## 1116          United States Merchant Marine Academy  NY          Public
##      RevenueFT
## 388          315
## 2596          215
## 2272           92
## 1121           55
## 1116           0
```

Hide

```
#creating a separate dataframe of average revenues & expenses
avgrev<-c(mean(coldata$RevenueFT[coldata$SchoolType=="Public"]),mean(coldata$RevenueFT[coldata$SchoolType=="Proprietary"]), mean(coldata$RevenueFT[coldata$SchoolType=="Private-Nonprofit"]))
```

```
avgrev
```

```
## [1] 5012.833 14463.627 16648.148
```

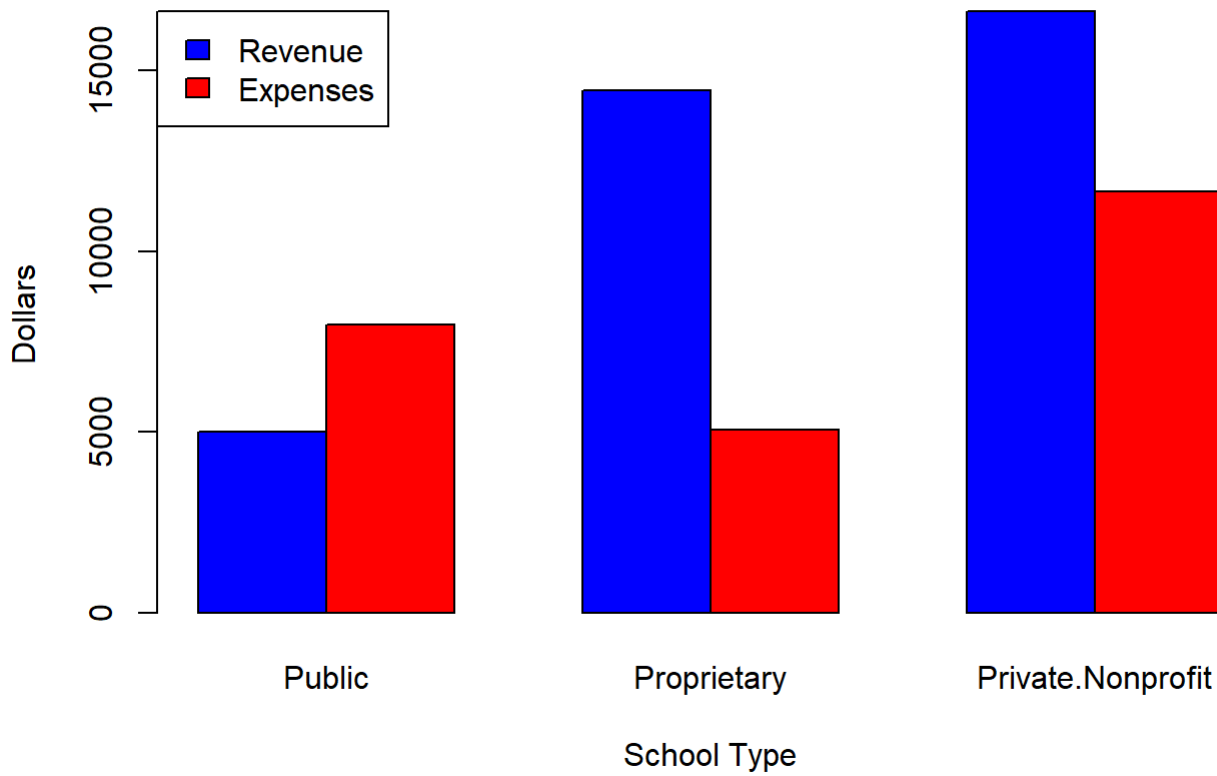
Hide

```
avgexp<-c(mean(coldata$ExpensesFT[coldata$SchoolType=="Public"]),mean(coldata$ExpensesFT[coldata$SchoolType=="Proprietary"]), mean(coldata$ExpensesFT[coldata$SchoolType=="Private-Nonprofit"]))
schoolvec<-c("Public", "Proprietary", "Private-Nonprofit")
```

```
revdfnew<-data.frame(avgrev, avgexp)
rownames(revdfnew)<-c("Public", "Proprietary", "Private-Nonprofit")
revnew<-data.frame(t(revdfnew))
rownames(revnew)<-c("Revenue", "Expenses")
```

```
barplot(as.matrix(revnew), main="Average Revenue and Expenses per Full-time Student", xlab="School Type", ylab="Dollars", col=c("blue", "red"), beside =TRUE, legend.text=rownames(revnew), args.legend= list(x="topleft", inset=c(0,0)))
```

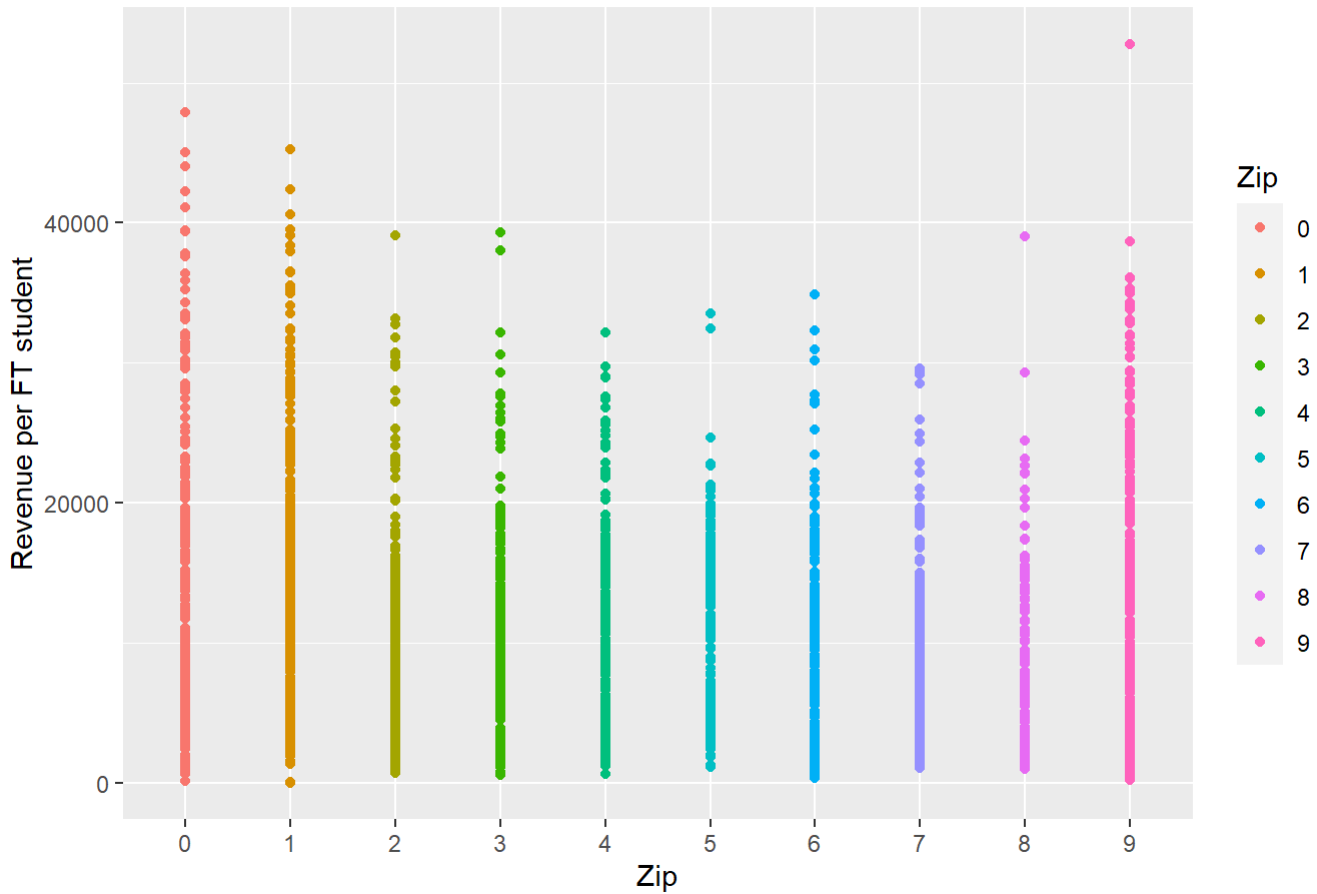
## Average Revenue and Expenses per Full-time Student



Hide

```
c + geom_point(aes(x=ZipFactor, y=RevenueFT, col=ZipFactor)) + xlab("Zip") + ylab("Revenue per FT student") + ggtitle("Revenue per Full-Time Student by Zip Code") + labs(color="Zip")
```

## Revenue per Full-Time Student by Zip Code



The following graph illustrates the increase of revenue as related to an increase in cost of attendance. In the `geom_smooth` model, public colleges follow a logistic S-shape curve whereas private colleges follow more of an exponential incline. Proprietary colleges achieve a linear pattern between revenue and cost.

Hide

```
c + geom_point(aes(x=TotalCost, y=RevenueFT, col=SchoolType)) + xlab("Total Cost") + ylab("Revenue per Full-Time Student") + ggtitle("Revenue per Full-Time Student vs. Total Cost")
```

## Revenue per Full-Time Student vs. Total Cost

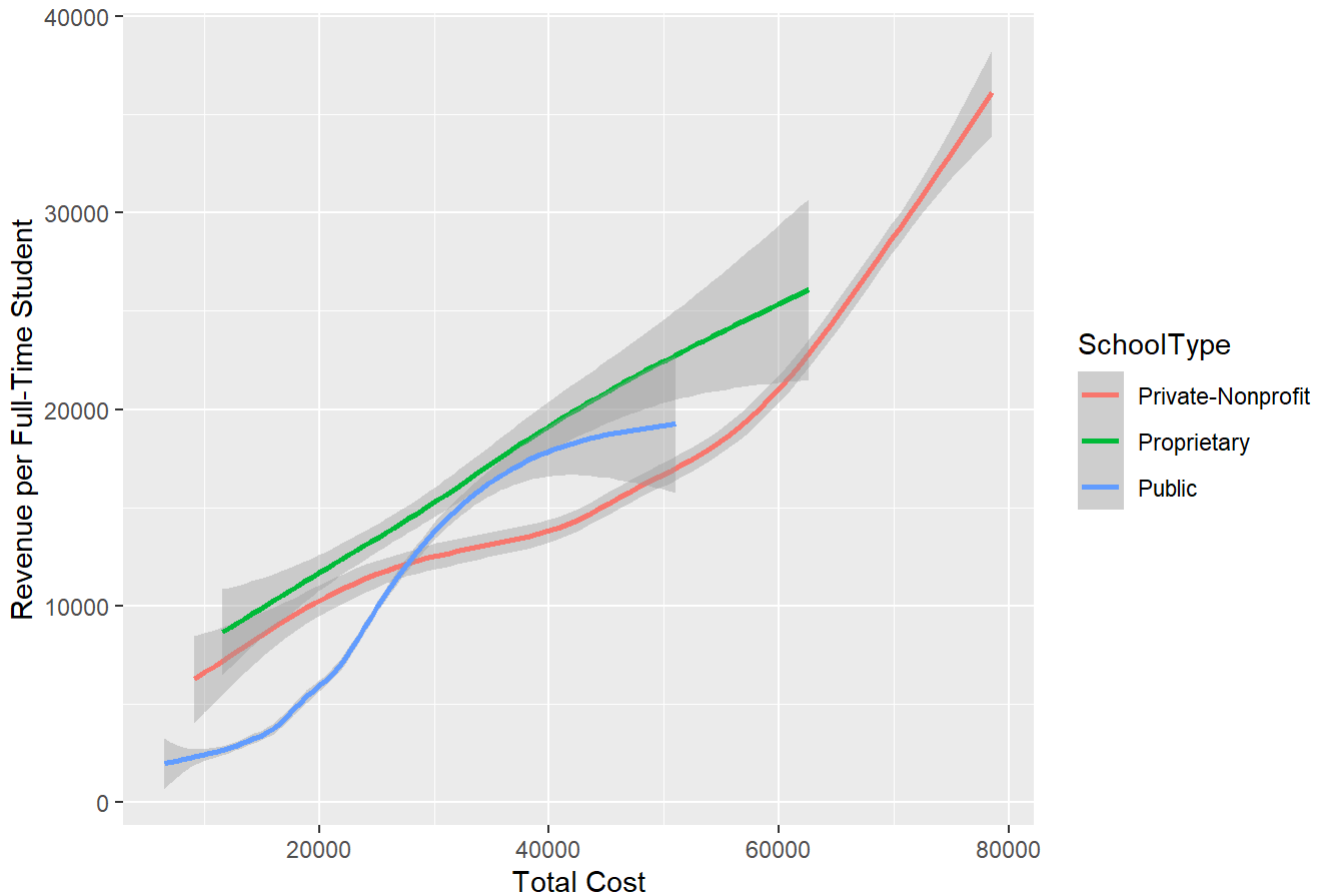


Hide

```
c + geom_smooth(aes(x=TotalCost, y=RevenueFT, col=SchoolType)) + xlab("Total Cost") + ylab("Revenue per Full-Time Student") + ggtitle("Geom_Smooth with Revenue and Total Cost")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

Geom\_Smooth with Revenue and Total Cost



## Faculty Salaries

In examining the average monthly faculty salary, the visualization below presents salary versus the revenue per full-time student. The black line indicates the average of all salaries in the dataset. Proprietary colleges seem to maintain lower monthly salaries (below the average salary) despite more revenue.

In the zip code analysis, the West, New England, and the greater NY region maintain some of the highest salaries. Region 5 (comprised of Iowa, Minnesota, Montana, North Dakota, South Dakota, Wisconsin) has a shorter range of salaries.

Top 5 Salaries:

Hide

```
sortsal <- coldata[order( -coldata$SalaryAVG) , ]
head(sortsal[,c(3,5,8,15)],5)
```

##	Name	State	SchoolType	SalaryAVG
## 693	Harvard University	MA	Private-Nonprofit	20988
## 190	Stanford University	CA	Private-Nonprofit	20865
## 972	Princeton University	NJ	Private-Nonprofit	20724
## 85	California Institute of Technology	CA	Private-Nonprofit	20595
## 254	Yale University	CT	Private-Nonprofit	19830



Bottom 5 Salaries:

Hide

```
tail(sortsal[,c(3,5,8,15)],5)
```

##	Name	State	SchoolType	SalaryAVG
## 2405	Gods Bible School and College	OH	Private-Nonprofit	1198
## 2617	Advance Science International College	FL	Proprietary	1194
## 2491	Kentucky Mountain Bible College	KY	Private-Nonprofit	1002
## 2630	Ecclesia College	AR	Private-Nonprofit	963
## 2442	Universidad Teologica del Caribe	PR	Private-Nonprofit	940

Hide

```
c+ geom_point(aes(x=RevenueFT, y=SalaryAVG, color=SchoolType)) + ggtitle("Avg Monthly Faculty Salary vs. Revenue per Full-Time Student") + xlab("Revenue") + ylab("Average Faculty Salary per Month") + labs(color="School Type") + geom_hline(yintercept = mean(coldata$SalaryAVG), color="black")
```

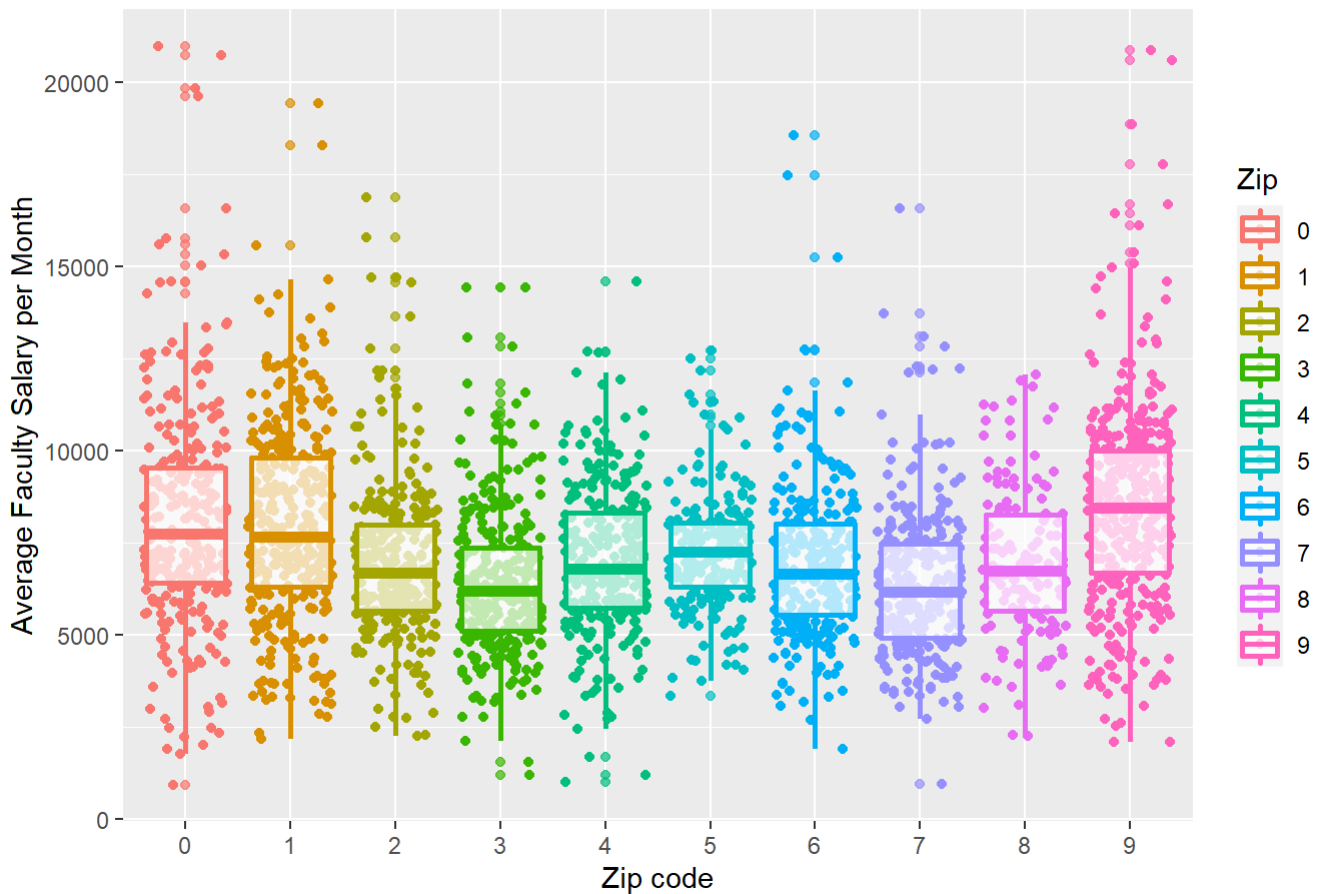
Avg Monthly Faculty Salary vs. Revenue per Full-Time Student



Hide

```
salplot<-ggplot(data=coldata, aes(x=ZipFactor, y=SalaryAVG, color=ZipFactor))  
salplot + geom_jitter() + geom_boxplot(size=1, alpha = 0.7) + ggtitle("Average Faculty Salary by Zipcode") + xlab("Zip code") + ylab("Average Faculty Salary per Month") + labs(color="Zip")
```

Average Faculty Salary by Zipcode



## Student Loan Debt

In analyzing loan debt, the first graph shows college default rate versus average family income. There appears to be a visible relationship between these two criteria; as family income increases, the default rate decreases. The next histograms show the distribution of default rate by school type and then by location. There is a higher concentration of private colleges with lower default rates. The violin graph creates a visualization of the density of median loan debts. In this example, proprietary and public institutions share a similar shape whereas private colleges have a higher median loan debt concentration.

Top 5 Default Rates:

Hide

```
#creating a column for default rate to convert to a percent
coldata$DefaultRatePercent<-coldata$DefaultRate * 100
sortdr <- coldata[order( -coldata$DefaultRatePercent) , ]
head(sortdr[,c(3,5,8,24)],5)
```

##	Name	State	SchoolType	DefaultRatePercent
## 1873	Denmark Technical College	SC	Public	47.3
## 1826	Clinton College	SC	Private-Nonprofit	37.1
## 55	Arkansas Baptist College	AR	Private-Nonprofit	37.0
## 2682	MediaTech Institute-Dallas	TX	Proprietary	35.5
## 2740	Sonoran Desert Institute	AZ	Proprietary	34.2

Top 5 Outstanding Loan Totals:

Hide

```
sortloan <- coldata[order( -coldata$LoanTotal) , ]
head(sortloan[,c(3,5,8,21)],5)
```

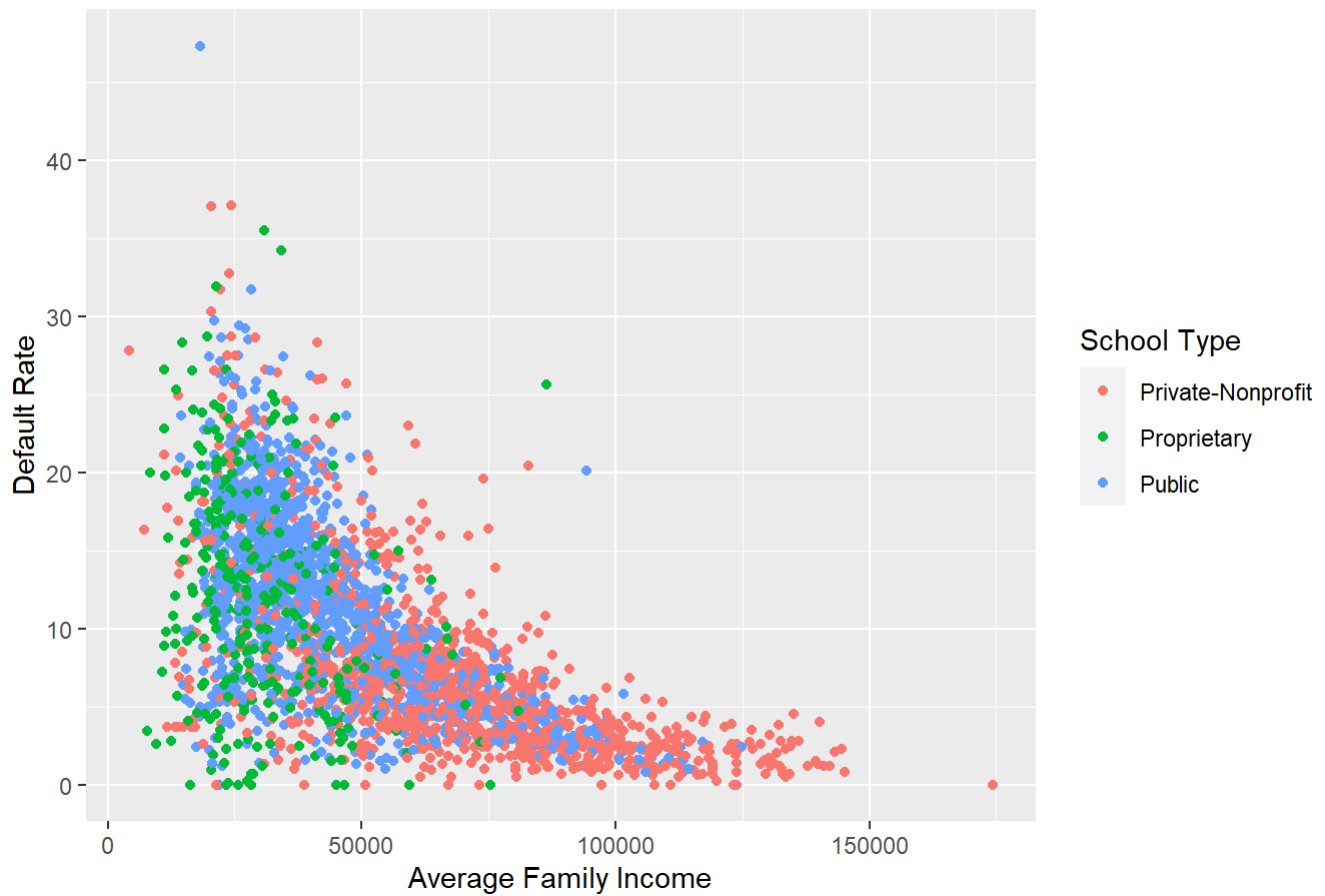
##	Name	State	SchoolType	LoanTotal
## 2341	University of Phoenix-Arizona	AZ	Proprietary	37427769280
## 2229	DeVry University-Illinois	IL	Proprietary	12690573163
## 267	Strayer University-Global Region	DC	Proprietary	9401828507
## 45	Grand Canyon University	AZ	Proprietary	8270808321
## 297	Nova Southeastern University	FL	Private-Nonprofit	8216982781

Hide

```
#calculating average default
avgdefault<-mean(coldata$DefaultRatePercent)

c+ geom_point(aes(x=FamIncomeAVG, y=coldata$DefaultRatePercent, color=SchoolType)) + ggtitle("De
fault Rate vs. Average Family Income") + xlab("Average Family Income") + ylab("Default Rate") +
labs(color="School Type")
```

## Default Rate vs. Average Family Income

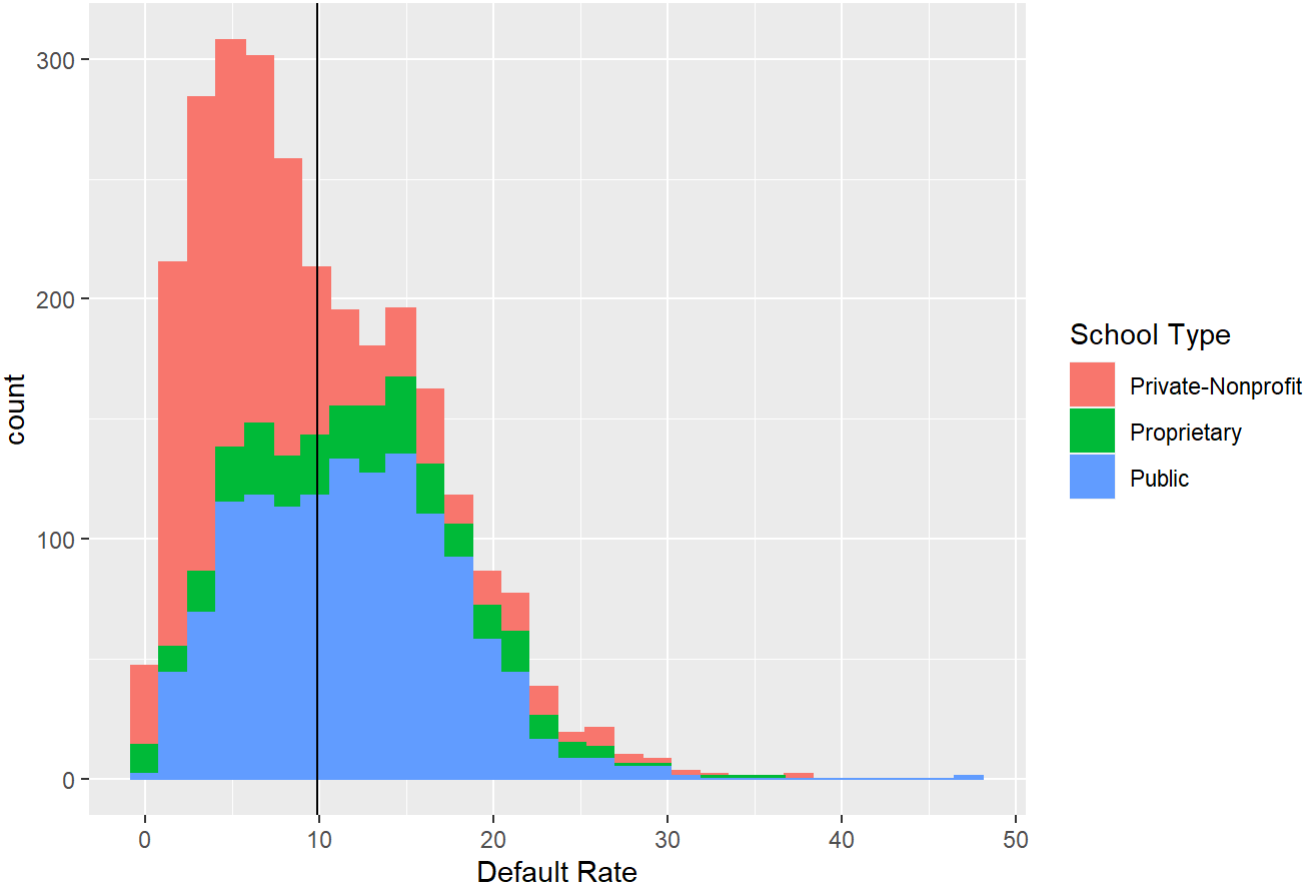


Hide

```
c+ geom_histogram(aes(x=coldata$DefaultRatePercent, color=SchoolType, fill=SchoolType)) + ggtitle("Default Rate By School Type") + xlab("Default Rate") + labs(color="School Type", fill="School Type") +geom_vline(aes(xintercept = avgdefault), color="black")
```

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

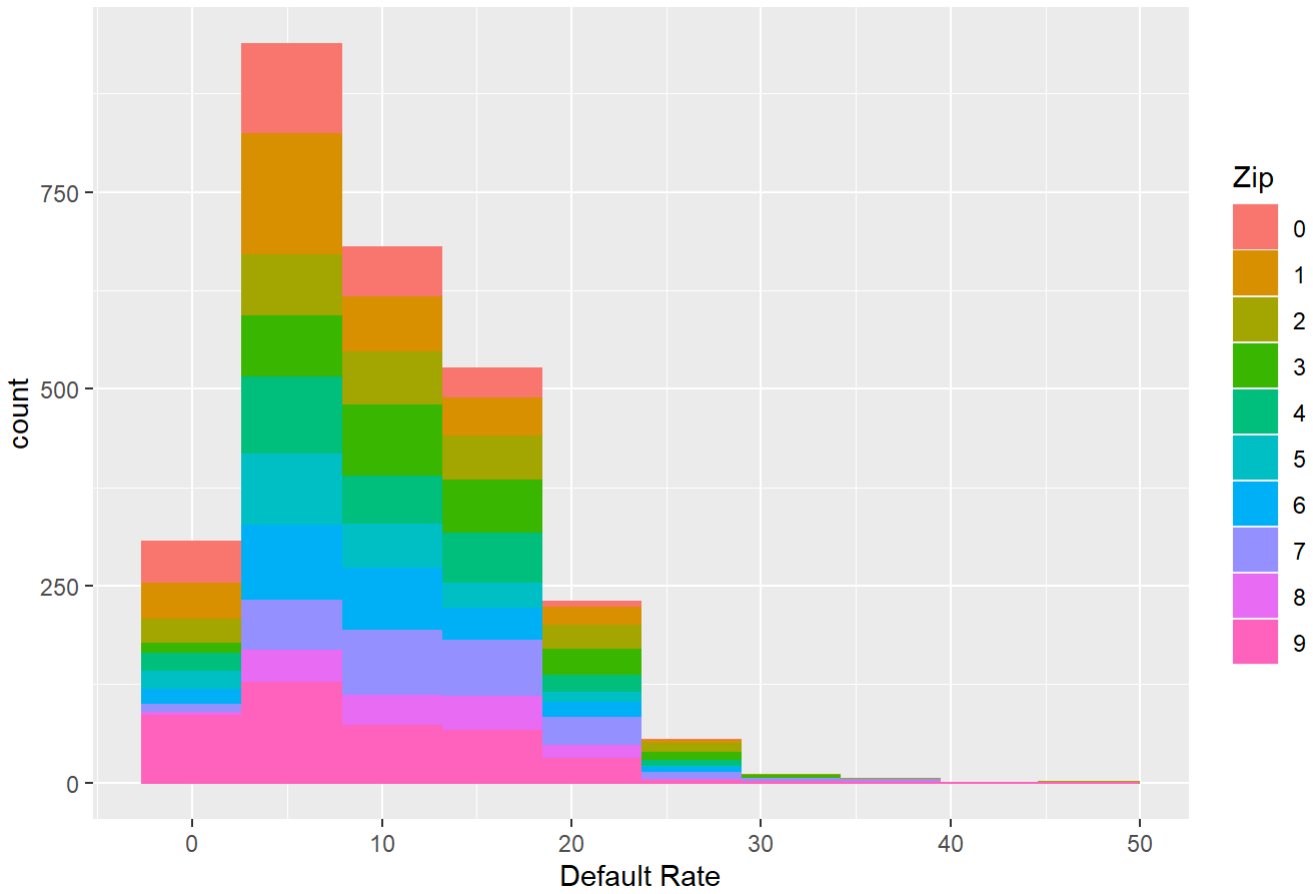
Default Rate By School Type



Hide

```
c+ geom_histogram(aes(x=coldata$DefaultRatePercent, color=ZipFactor, fill=ZipFactor), bins=10) +  
xlab("Default Rate") + labs(color="Zip", fill="Zip") + ggtitle("Default Rate per Zip Code")
```

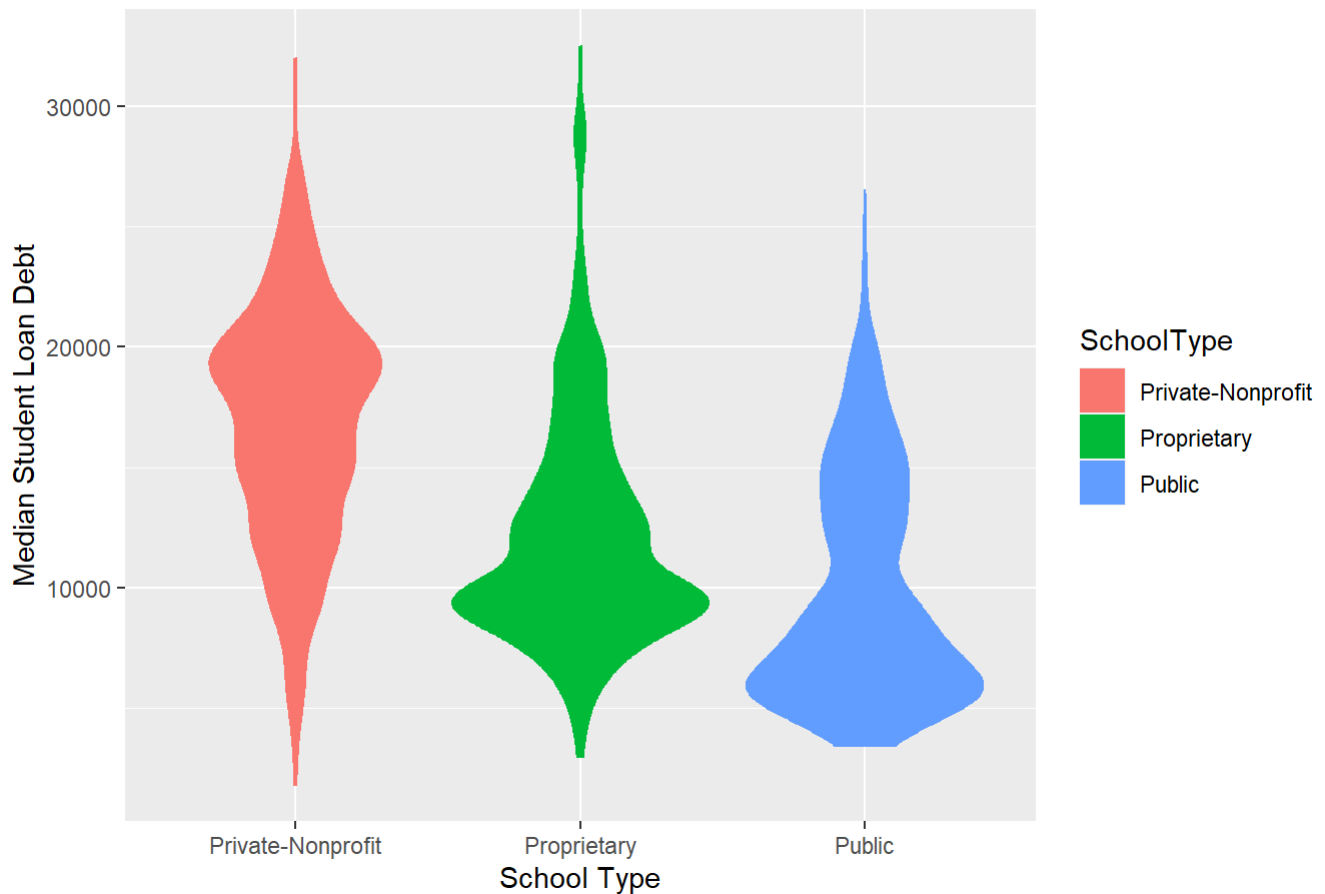
Default Rate per Zip Code



Hide

```
c+geom_violin(aes(x=SchoolType, y=DebtMDN, color=SchoolType, fill=SchoolType))+ xlab("School Type") + ylab("Median Student Loan Debt") + labs(color="SchoolType") + ggtitle("Violin Graph of Median Student Loan Debt")
```

Violin Graph of Median Student Loan Debt



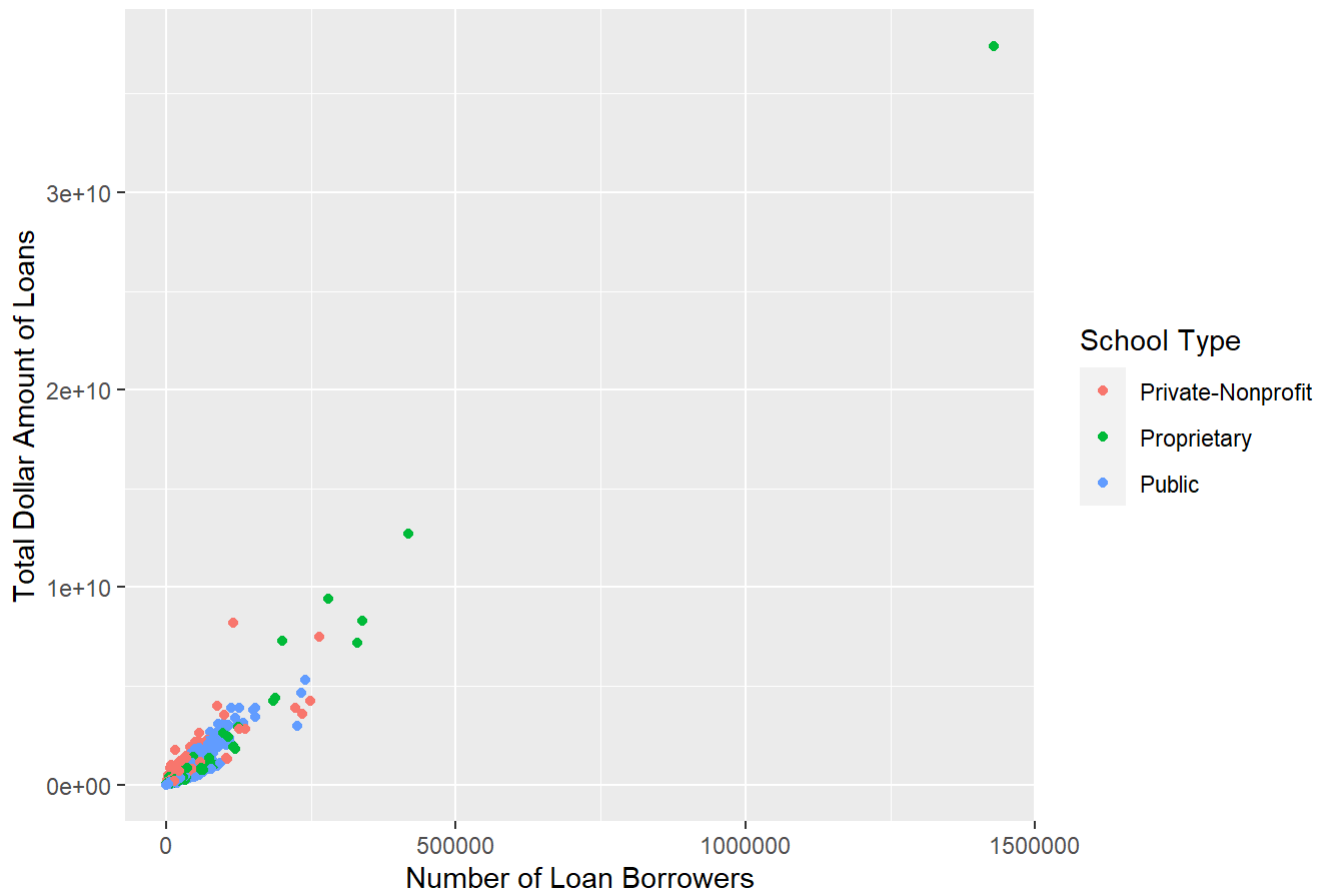
## Regression Example

In the example below, there is a strong linear relationship between the total dollar amount of federal student loans with outstanding balances compared to the number of loan borrowers with outstanding balances (ED, 2021). While the graph shows a major outlier (University of Phoenix), this outlier seemingly aligns with the regression example. The regression line can also be seen in a “zoomed in” example. In the regression formula, the multiple R-squared in this case is 93.89% (good confidence) and, because  $p$  is less than .05, the Null hypothesis is rejected; the number of borrowers with outstanding balances does impact the total dollar amount of outstanding debt.

Hide

```
lreg<-c+ geom_point(aes(x=LoanCount, y=LoanTotal, color=SchoolType)) + xlab("Number of Loan Borrowers") + ylab("Total Dollar Amount of Loans") + labs(color="School Type") + ggtitle("Outstanding Loan Total vs. Borrowers")
lreg
```

## Outstanding Loan Total vs. Borrowers



Hide

```
lr<-lm(coldata$LoanTotal~coldata$LoanCount)
lr
```

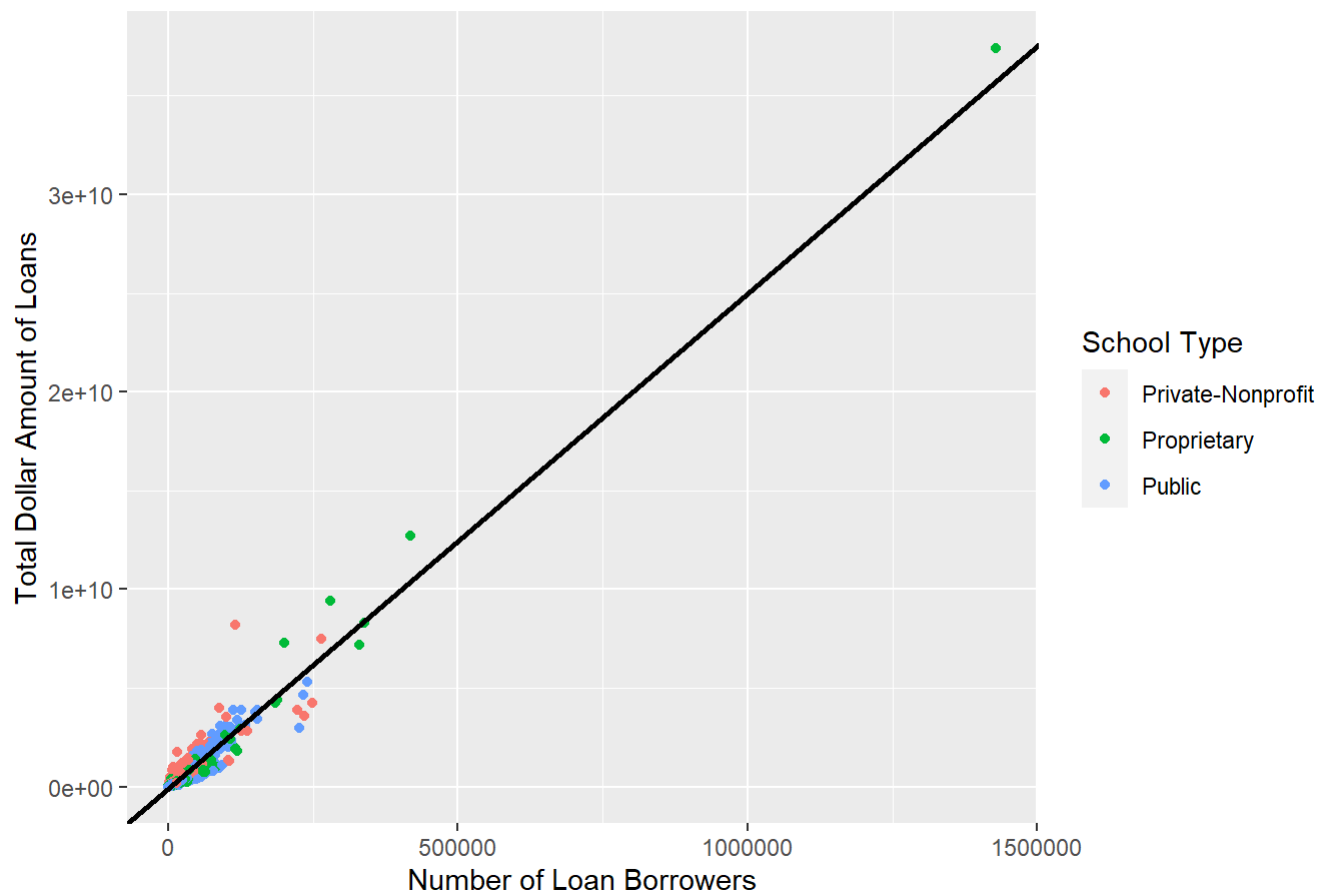
```
##
## Call:
## lm(formula = coldata$LoanTotal ~ coldata$LoanCount)
##
## Coefficients:
##      (Intercept) coldata$LoanCount
##      -78574048      25091
```

Hide

```
lreg + geom_abline(intercept=-78574048, slope=25091, color="black", size=1) + ggtitle("Loan Total vs. Borrowers with Regression Line")
```



Loan Total vs. Borrowers with Regression Line



Hide

```
lreg + xlim(0, 100000) + ylim(0, 1000000000) + geom_abline(intercept=-78574048, slope=25091, color="black", size=1) + ggtitle("Zoomed Loan Total vs. Borrowers with Regression Line")
```

```
## Warning: Removed 231 rows containing missing values (geom_point).
```

Zoomed Loan Total vs. Borrowers with Regression Line



Hide

```
summary(lr)
```

```
##
## Call:
## lm(formula = coldata$LoanTotal ~ coldata$LoanCount)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.639e+09 -4.552e+07  2.685e+07  5.936e+07  5.402e+09
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.857e+07  5.117e+06  -15.36  <2e-16 ***
## coldata$LoanCount  2.509e+04  1.222e+02  205.36  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 243500000 on 2744 degrees of freedom
## Multiple R-squared:  0.9389, Adjusted R-squared:  0.9389
## F-statistic: 4.217e+04 on 1 and 2744 DF, p-value: < 2.2e-16
```

## Map with Loan Data

In this final example, the loan and geographic information from the Scorecard dataset are combined using the leaflet library. This graphic shows a U.S. map with interactive zoom. The colors are indicative of the school type, keeping the same pattern as prior—blue (public), green (proprietary), red (private-nonprofit). Additionally, the circles indicate both the location of the school as well as the total outstanding student loans owed from borrowers at that institution. This visualization provides a deeper understanding of outstanding loan debt in the United States in terms of both location and school type.

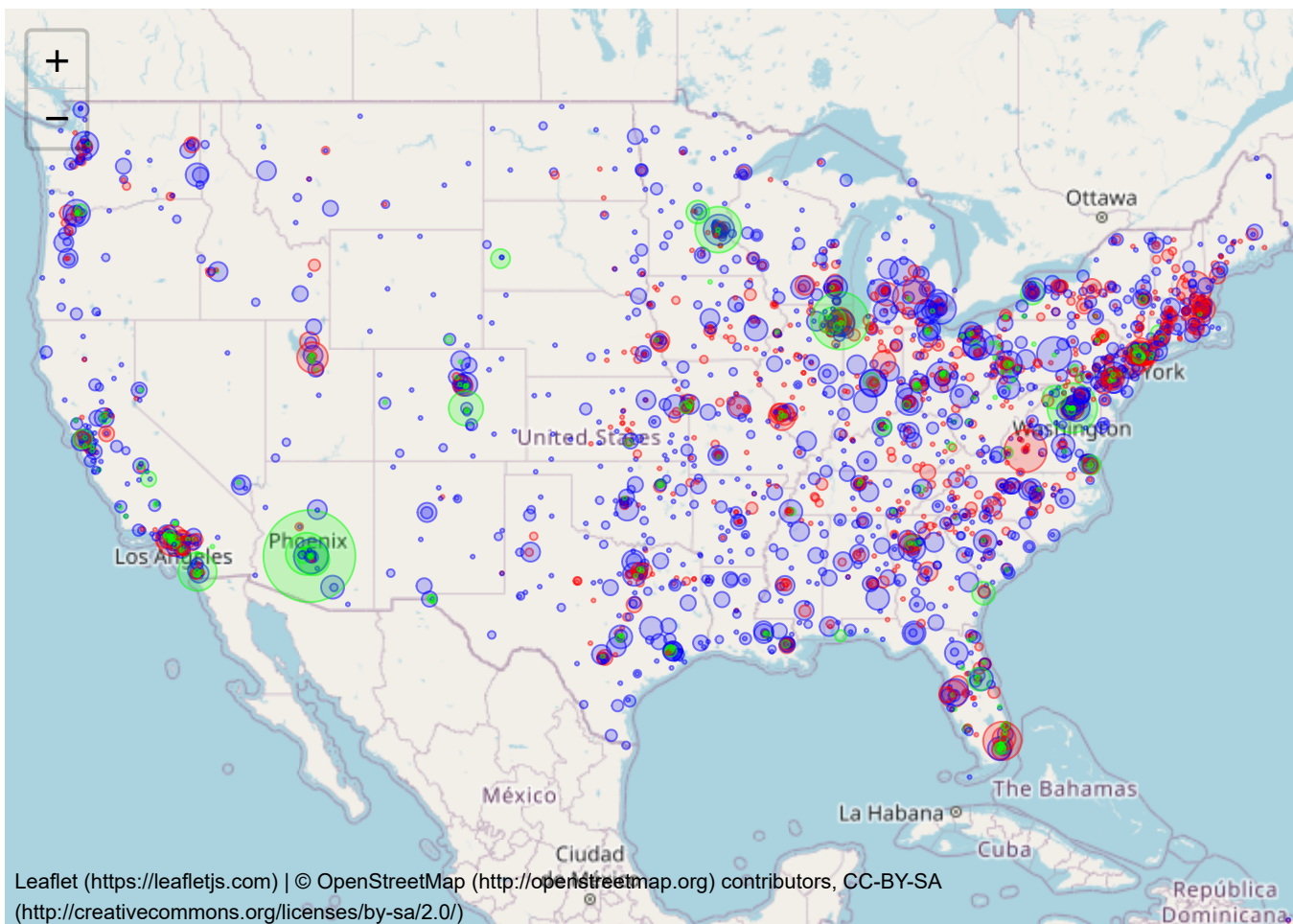
Hide

```
library("leaflet")
```

```
## Warning: package 'leaflet' was built under R version 4.1.2
```

Hide

```
pal <- colorFactor(  
  palette=c("red", "green", "blue"), # to keep color pattern  
  domain = coldata$SchoolType  
)  
  
leaflet(coldata) %>%  
  setView(lng = -96.25, lat = 39.50, zoom = 3) %>% # set viewing range  
  addTiles() %>%  
  addCircles(lng = ~Long, lat = ~Lat, weight = 1, radius = ~sqrt(LoanTotal), color = ~pal(SchoolType))
```



# Conclusion

The College Scorecard operates as a consumer guide for prospective students to obtain a data-informed snapshot of a college/university. Through data analytics on this dataset, one can see patterns in the higher education industry as a whole. These visualizations represent examples of how money shapes the higher education landscape. One can see the impact of radical outliers in comparison to the majority of institutions. The outstanding loan totals point to a seeming monopolization of the industry and a major problem in terms of education-related loan debt. Review of this data from a “big data” perspective may ultimately drive strategies for more regulatory oversight as well as solutions for growing issues related to student loan borrowing.

# References

Federal Student Aid. (2021). AY 2020-2021 Q4. *Title IV Program Volume Reports*. Retrieved November 17, 2021 from <https://studentaid.gov/data-center/student/title-iv> (<https://studentaid.gov/data-center/student/title-iv>).

Office of Planning, Evaluation and Policy Development. (2020, July 17). Open Data Platform: College Scorecard. Retrieved November 20, 2021 from <https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020/resources> (<https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020/resources>)

U.S. Department of Education. (n.d.). College Scorecard. Retrieved November 20, 2021 from <https://collegescorecard.ed.gov/> (<https://collegescorecard.ed.gov/>)

U.S. Department of Education. (2021). *Technical Documentation: College Scorecard Institution-Level Data (Version: July 2021)*. U.S. Department of Education. <https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf> (<https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf>)

UnitedStatesZipCodes.org. (2021). Zip Code Zones. Retrieved November 20, 2021 from <https://www.unitedstateszipcodes.org/> (<https://www.unitedstateszipcodes.org/>)