College Scorecard Code

Yu Wang

November 25, 2016

Table of Contents

---

1. Import Libraries

The database is given in both a SQLite file and CSVs for each year per file. I will mostly be working with the SQLite because all the data is in one place.

I installed SQLiteStudio to look at the variables. The entire table crashes SQLiteStudio on my computer due to file size. But queries can still be performed to check the output. Most of the data is worked on for 2011 because that is the last dataset that has md_earn_wne_p10, the main salary component I looked at.

Please note, for every section the APPROCH was done once, to save computation time when Knitting the PDF.

First import the libraries that would be needed.

```
library(RSQLite)  # Library to work with SQLite
```

```
## Warning: package 'RSQLite' was built under R version 3.3.2
```

```
## Warning: package 'DBI' was built under R version 3.3.2
```

```
library(dplyr)    # Library for data manipulation
```

```
## Warning: package 'dplyr' was built under R version 3.3.2
```

```
library(ggplot2)  # Library for plotting
```

```
## Warning: package 'ggplot2' was built under R version 3.3.1
```

---

2. Connect to database and import

You can also embed plots, for example:

```
db <- dbConnect(dbDriver("SQLite"), "C:/Users/Adam/Desktop/CAPSTONE/output/database.sqlite")
dbGetQuery(db, "PRAGMA temp_store=2;") #Do not load everything into RAM
```

The database one have 1 table called Scorecard. What is in scorecard?

Here the head and number of columns of the table is shown. A detailed description of the column can be found within FULLDataDOCUMENTATION.PDF file.

```
scorecard_columns = dbGetQuery(db, "PRAGMA table_info('Scorecard')")
head(scorecard_columns)
```

```
##   cid   name    type notnull dflt_value pk
## 1   0     Id INTEGER       0       <NA>  1
## 2   1 UNITID INTEGER       0       <NA>  0
## 3   2  OPEID INTEGER       0       <NA>  0
## 4   3 opeid6 INTEGER       0       <NA>  0
## 5   4 INSTNM    TEXT       0       <NA>  0
## 6   5   CITY    TEXT       0       <NA>  0
```

```
nrow(scorecard_columns) # This is the number of columns within the table.
```

```
## [1] 1731
```

The number of rows in the database:

```
dbGetQuery(db, "SELECT count(*) FROM Scorecard")
```

```
##   count(*)
## 1   124699
```

So we see this is a very big data base with 1731 Columns and 124699 Rows. The dataset is separated by different years which is given in CSV format as well.

---

3. Number of Colleges though Time

How many college in US from 1996 to 2013?

```
numberOfSchools = dbGetQuery(db, "SELECT Year, COUNT(Id) NumSchools
                             FROM Scorecard GROUP by Year")
head(numberOfSchools)
```

```
##   Year NumSchools
## 1 1996       6794
## 2 1997       6699
## 3 1998       6480
## 4 1999       6466
## 5 2000       6478
## 6 2001       6619
```

Let's plot this using ggplot2

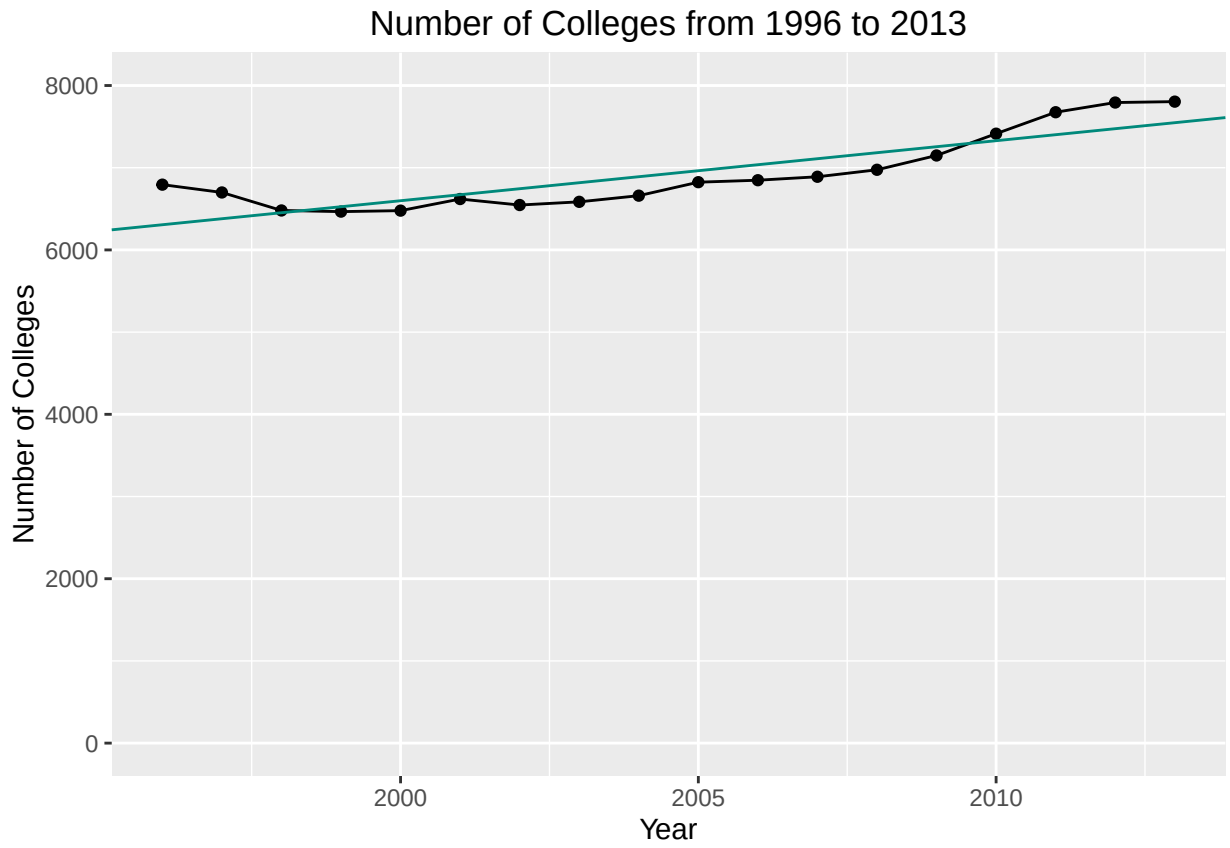
```
best_fit <- lm(numberOfSchools$NumSchools~numberOfSchools$Year)
best_fit
```

```
##
## Call:
## lm(formula = numberOfSchools$NumSchools ~ numberOfSchools$Year)
##
## Coefficients:
##          (Intercept)  numberOfSchools$Year
##           -139421.46                 73.01
```

```
summary(best_fit)
```

```
##
## Call:
## lm(formula = numberOfSchools$NumSchools ~ numberOfSchools$Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -233.21 -196.71  -83.72  212.95  486.87
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -139421.46   21775.36  -6.403 8.75e-06 ***
## numberOfSchools$Year      73.01      10.86   6.721 4.91e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 239.1 on 16 degrees of freedom
## Multiple R-squared:  0.7384, Adjusted R-squared:  0.7221
## F-statistic: 45.17 on 1 and 16 DF,  p-value: 4.913e-06
```

```
ggplot(numberOfSchools, aes(x=numberOfSchools$Year, y=numberOfSchools$NumSchools,
                            group=1)) + geom_line() +
                            geom_point() + ylim(0, 8000) + xlim(1996, 2013) +
                            geom_abline(intercept = -139421.46, slope = 73.01, colour='#00897B') +
                            labs(x="Year",y="Number of Colleges") +
                            ggtitle("Number of Colleges from 1996 to 2013")
```



From this we can predict the number of colleges in the future. There is a clear indication that the number of colleges is increasing. However, given only a few years of data points, the prediction woudn't be accuract.

---

4. SAT Scores

Distribution of SAT scores

First import database with unneeded attributes elimated, then the data was cleaned such that if any SAT score is NULL it is left out. Take a look at the head of this dataset:

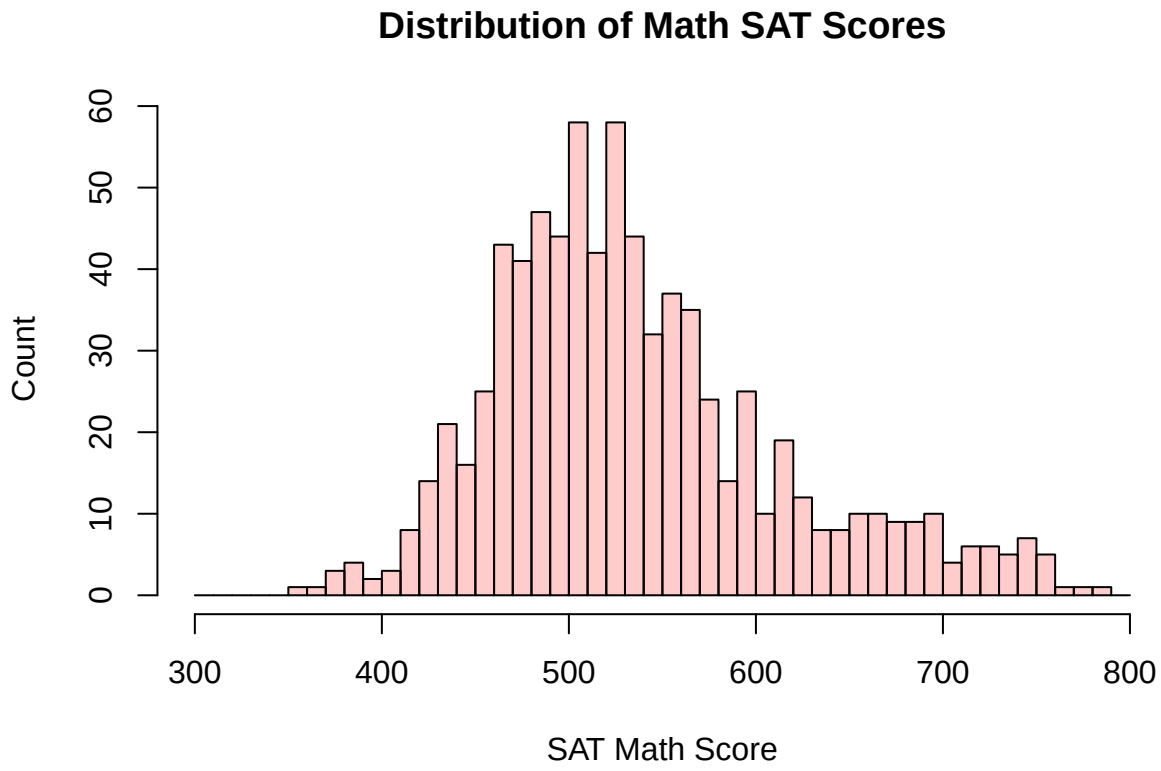
```
sat <- dbGetQuery(db, "SELECT INSTNM,
       SATMTMID,
       SATVRMID,
       SATWRMID
FROM Scorecard
WHERE Year=2013
```

```
  AND SATMTMID IS NOT NULL
  AND SATVRMID IS NOT NULL
  AND SATWRMID IS NOT NULL")

# INSTNM is College Name, SATMTMID is SAT Math Score medium,
# SATVRMID is SAT Verbal Score medium, SATWRMID is SAT Writing Score medium.

head(sat)
```

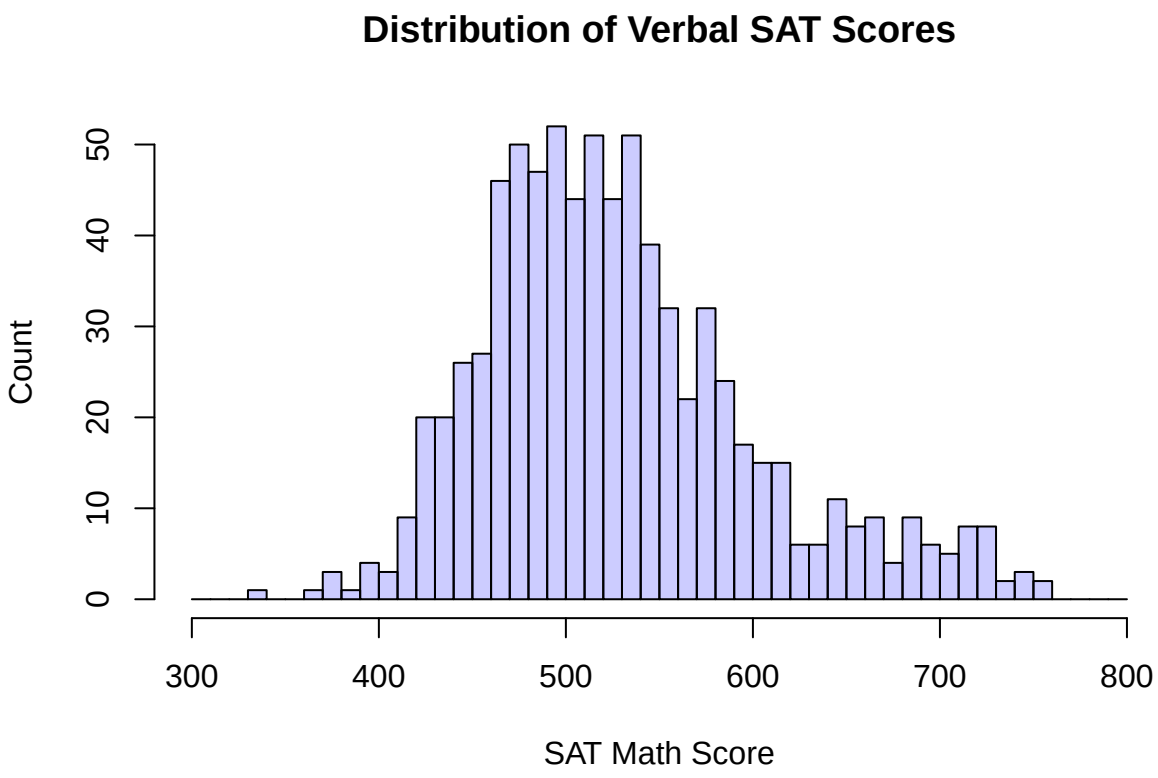
```
##                         INSTNM SATMTMID SATVRMID SATWRMID
## 1   The University of Alabama      570      555      540
## 2           Auburn University      595      570      565
## 3              Judson College      550      595      570
## 4    University of Montevallo      508      554      510
## 5          Samford University      560      565      555
## 6 University of South Alabama      500      495      470
```

Here is the distribution of Math, Verbal and Written SAT Scores averages for the Colleges

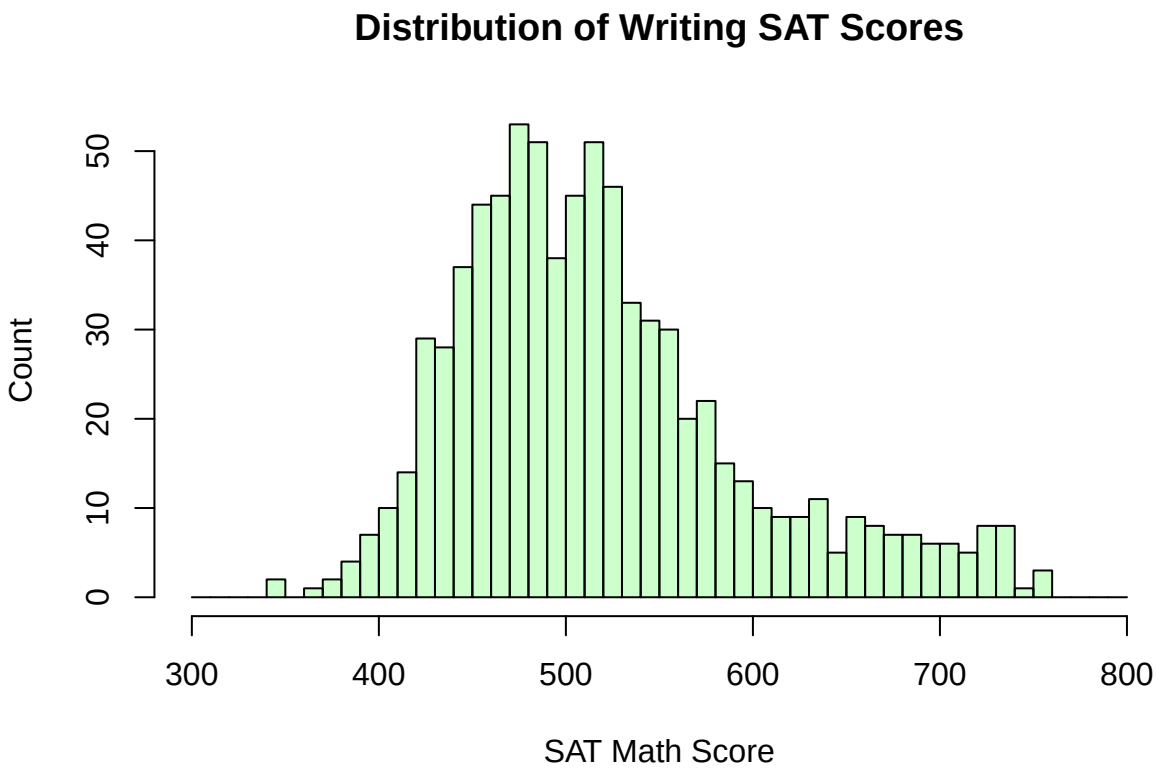
```
hist(sat$SATMTMID, main="Distribution of Math SAT Scores", xlab="SAT Math Score", ylab="Count",
     col=rgb(1,0,0,0.2), breaks = seq(300,800,by=10))
```




```
hist(sat$SATVRMID, main="Distribution of Verbal SAT Scores", xlab="SAT Math Score", ylab="Count",
     col=rgb(0,0,1,0.2), breaks = seq(300,800,by=10), add = F)
```

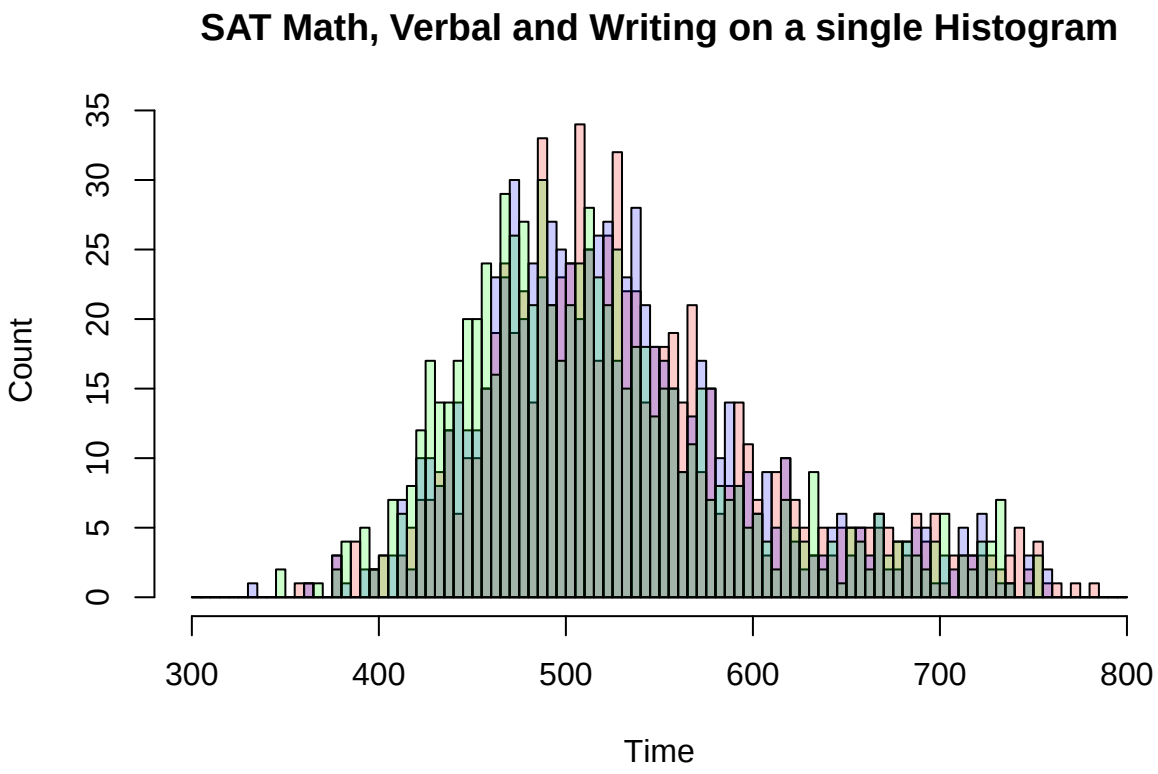



```
hist(sat$SATWRMID, main="Distribution of Writing SAT Scores", xlab="SAT Math Score", ylab="Count",
     col=rgb(0,1,0,0.2), breaks = seq(300,800,by=10), add = F)
```



Here is the distribution of Math SAT Scores averages for the Colleges

```
hist(sat$SATMTMID, main="SAT Math, Verbal and Writing on a single Histogram",
     xlab="Time", ylab="Count",
     col=rgb(1,0,0,0.2), breaks = seq(300,800,by=5))
hist(sat$SATVRMID, col=rgb(0,0,1,0.2), breaks = seq(300,800,by=5), add = T)
hist(sat$SATWRMID, col=rgb(0,1,0,0.2), breaks = seq(300,800,by=5), add = T)
```



All the distribution of SAT scores are bell curve like and has a mid point at around 500.

T-tests for SAT

We can do a t-test for SAT Math, SAT Written, SAT Verbal to see if they are similar enough to be the same data-set.

```
t.test(sat$SATMTMID, sat$SATWRMID, alternative = c("two.sided"),
       paired = F, var.equal = T, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  sat$SATMTMID and sat$SATWRMID
## t = 4.5778, df = 1564, p-value = 5.07e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  10.31592 25.78369
## sample estimates:
## mean of x mean of y
##  539.1916  521.1418
```

```
t.test(sat$SATMTMID, sat$SATVRMID, alternative = c("two.sided"),
       paired = F, var.equal = T, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  sat$SATMTMID and sat$SATVRMID
## t = 2.1898, df = 1564, p-value = 0.02869
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.8750941 15.9116237
## sample estimates:
## mean of x mean of y
##  539.1916  530.7982
```

```
t.test(sat$SATVRMID, sat$SATWRMID, alternative = c("two.sided"),
       paired = F, var.equal = T, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  sat$SATVRMID and sat$SATWRMID
## t = 2.5209, df = 1564, p-value = 0.0118
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.142851 17.170048
## sample estimates:
## mean of x mean of y
##  530.7982  521.1418
```

Since the P value is all smaller than 0.05 here we reject the Null Hypothesis that theses are from the same data-set. This questions if adding SAT scores of different subjects to get total SAT score is a good idea, it might be a good excerise in the future to do the SAT analysis by subject rather than add together.

---

5. Income of students after 10 Years and Best ROI

Salary by school

Let's first import the all the colleges and see which has the highest median earnings after 10 years.

```
salary <- dbGetQuery(db, "SELECT INSTNM College,
       CONTROL CollegeType,
       md_earn_wne_p10
FROM Scorecard
WHERE Year=2011
AND md_earn_wne_p10 IS NOT NULL
AND md_earn_wne_p10 != 'PrivacySuppressed'
ORDER BY md_earn_wne_p10 DESC")
```

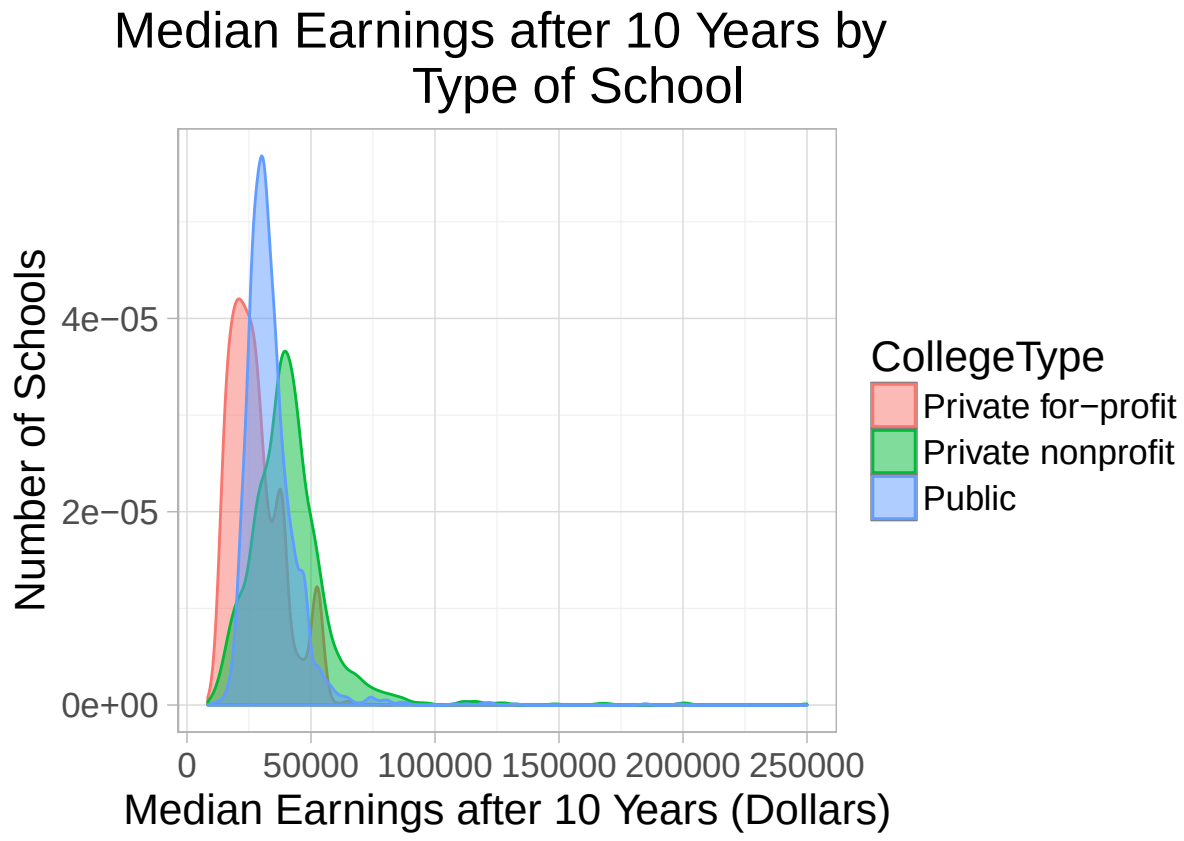
```
head(salary)
```

```
##                                          College       CollegeType
```

```
## 1                          Medical College of Wisconsin Private nonprofit
## 2                               Albany Medical College Private nonprofit
## 3             A T Still University of Health Sciences Private nonprofit
## 4         West Virginia School of Osteopathic Medicine            Public
## 5 University of Massachusetts Medical School Worcester            Public
## 6                             New York Medical College Private nonprofit
##   md_earn_wne_p10
## 1          250000
## 2          201200
## 3          199600
## 4          198300
## 5          184900
## 6          169600
```

Here it is arranged by highest earnings first. Notice the highests are all medical schools. We will fix this later on. For the plot below, we see that there are many colleges with average salary of 250k+ per year but theses are all for medical schools.

```
ggplot(salary, aes(x=salary$md_earn_wne_p10, color=CollegeType, fill=CollegeType,
                   group=CollegeType)) +
  geom_density(alpha=0.5) +
  theme_light(base_size=16) +
  xlab("Median Earnings after 10 Years (Dollars)") + ylab("Number of Schools") +
  ggtitle("Median Earnings after 10 Years by
          Type of School")
```



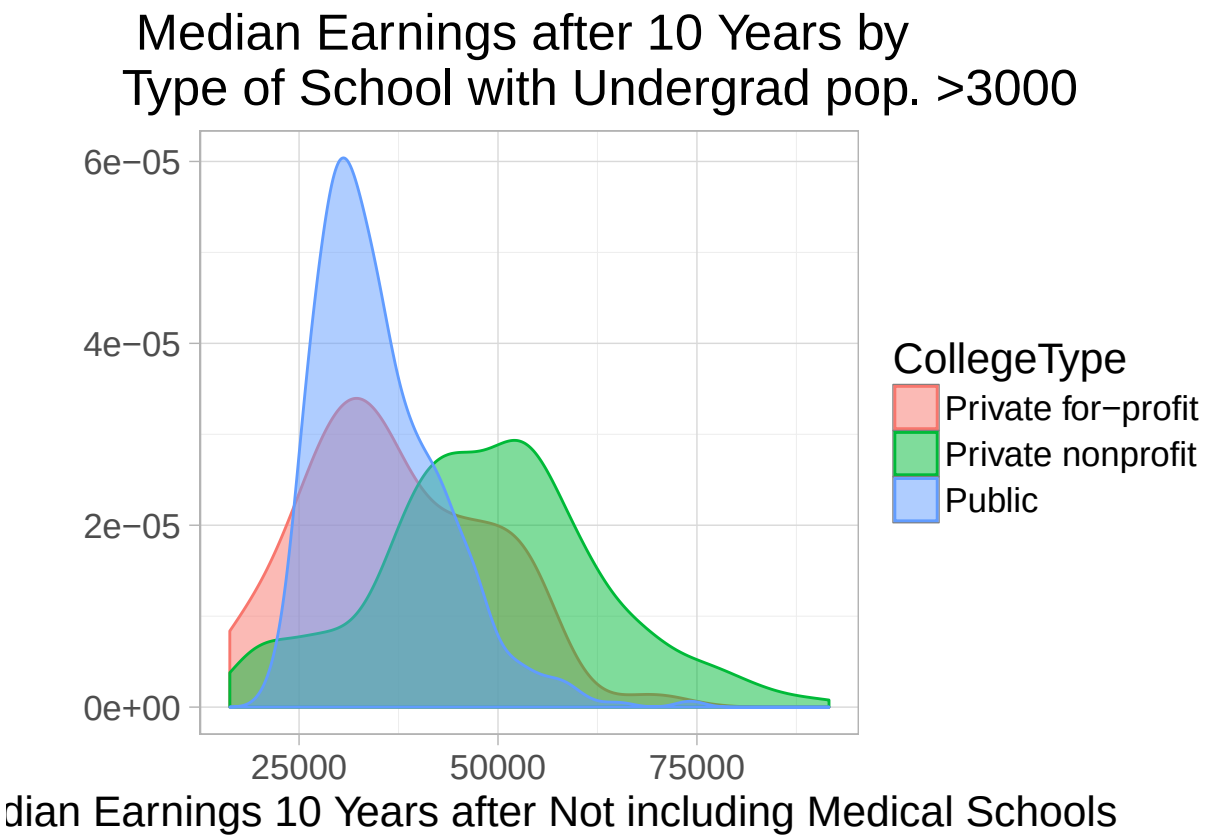
Here we gotten rid of the medical schools by making sure the college has at least 3000 students.

```
salary2 <- dbGetQuery(db, "SELECT INSTNM College,
       CONTROL CollegeType,
       md_earn_wne_p10,
       UGDS
FROM Scorecard
WHERE Year=2011
AND md_earn_wne_p10 IS NOT NULL
AND md_earn_wne_p10 != 'PrivacySuppressed'
AND UGDS IS NOT NULL
AND UGDS > 3000
ORDER BY md_earn_wne_p10 DESC")

head(salary2)
```

```
##                                 College       CollegeType md_earn_wne_p10
## 1 Massachusetts Institute of Technology Private nonprofit           91600
## 2                    Harvard University Private nonprofit           87200
## 3                  Georgetown University Private nonprofit           83300
## 4      Rensselaer Polytechnic Institute Private nonprofit           81700
## 5                    Stanford University Private nonprofit           80900
## 6            University of Pennsylvania Private nonprofit           78200
##    UGDS
## 1  4363
## 2  7245
## 3  7232
## 4  5240
## 5  6927
## 6 10720
```

The head here is more of what is expected. Top famous colleges have the highest student earnings.

```
ggplot(salary2, aes(x=salary2$md_earn_wne_p10, color=CollegeType, fill=CollegeType,
                    group=CollegeType)) +
  geom_density(alpha=0.5) +
  theme_light(base_size=16) +
  xlab("Median Earnings 10 Years after Not including Medical Schools") + ylab("") +
  ggtitle("Median Earnings after 10 Years by
          Type of School with Undergrad pop. >3000")
```



How we see a more expected earnings. From the graph, we see that private nonprofit has the best earnings. While private for-profit and public colleges have similar earnings, this peak is around 30,000$ per year.

ROI by SAT total score

First let's import the data from the data base with UGDS (The number of students) greater than 3000. And clean data with columns where any SAT score is NULL. This year we looked at is 2011 because the medium earnings data is the lastest.

```
sat_salary <- dbGetQuery(db, "SELECT INSTNM College,
       CONTROL CollegeType,
       md_earn_wne_p10,
       UGDS,
       SATMTMID,
       SATVRMID,
       SATWRMID
FROM Scorecard
WHERE Year=2011
AND md_earn_wne_p10 IS NOT NULL
AND md_earn_wne_p10 != 'PrivacySuppressed'
AND UGDS IS NOT NULL
AND UGDS > 3000
AND SATMTMID IS NOT NULL
```

```
AND SATVRMID IS NOT NULL
AND SATWRMID IS NOT NULL
ORDER BY md_earn_wne_p10 DESC")
```

See head to make sure everything is expected and create new column for total SAT score.

```
head(sat_salary)
```

```
##                                College       CollegeType md_earn_wne_p10
## 1 Massachusetts Institute of Technology Private nonprofit           91600
## 2                    Harvard University Private nonprofit           87200
## 3      Rensselaer Polytechnic Institute Private nonprofit           81700
## 4                   Stanford University Private nonprofit           80900
## 5            University of Pennsylvania Private nonprofit           78200
## 6                      Duke University Private nonprofit           76700
##    UGDS SATMTMID SATVRMID SATWRMID
## 1  4363      770      720      725
## 2  7245      750      740      740
## 3  5240      715      660      645
## 4  6927      735      720      730
## 5 10720      735      705      720
## 6  6534      735      705      720
```

```
sat_salary$total_sat <- sat_salary$SATMTMID + sat_salary$SATVRMID + sat_salary$SATWRMID
head(sat_salary)
```

```
##                                College       CollegeType md_earn_wne_p10
## 1 Massachusetts Institute of Technology Private nonprofit           91600
## 2                    Harvard University Private nonprofit           87200
## 3      Rensselaer Polytechnic Institute Private nonprofit           81700
## 4                   Stanford University Private nonprofit           80900
## 5            University of Pennsylvania Private nonprofit           78200
## 6                      Duke University Private nonprofit           76700
##    UGDS SATMTMID SATVRMID SATWRMID total_sat
## 1  4363      770      720      725      2215
## 2  7245      750      740      740      2230
## 3  5240      715      660      645      2020
## 4  6927      735      720      730      2185
## 5 10720      735      705      720      2160
## 6  6534      735      705      720      2160
```

As expected, famous Ivy league schools in the US have the highest SAT scores in all areas including total sat score.

Create column for best return, best earnings per sat score.

```
sat_salary$salary_to_sat <- sat_salary$md_earn_wne_p10 / sat_salary$total_sat
```

Rearrange dataframe by salary_to_sat with highest first.

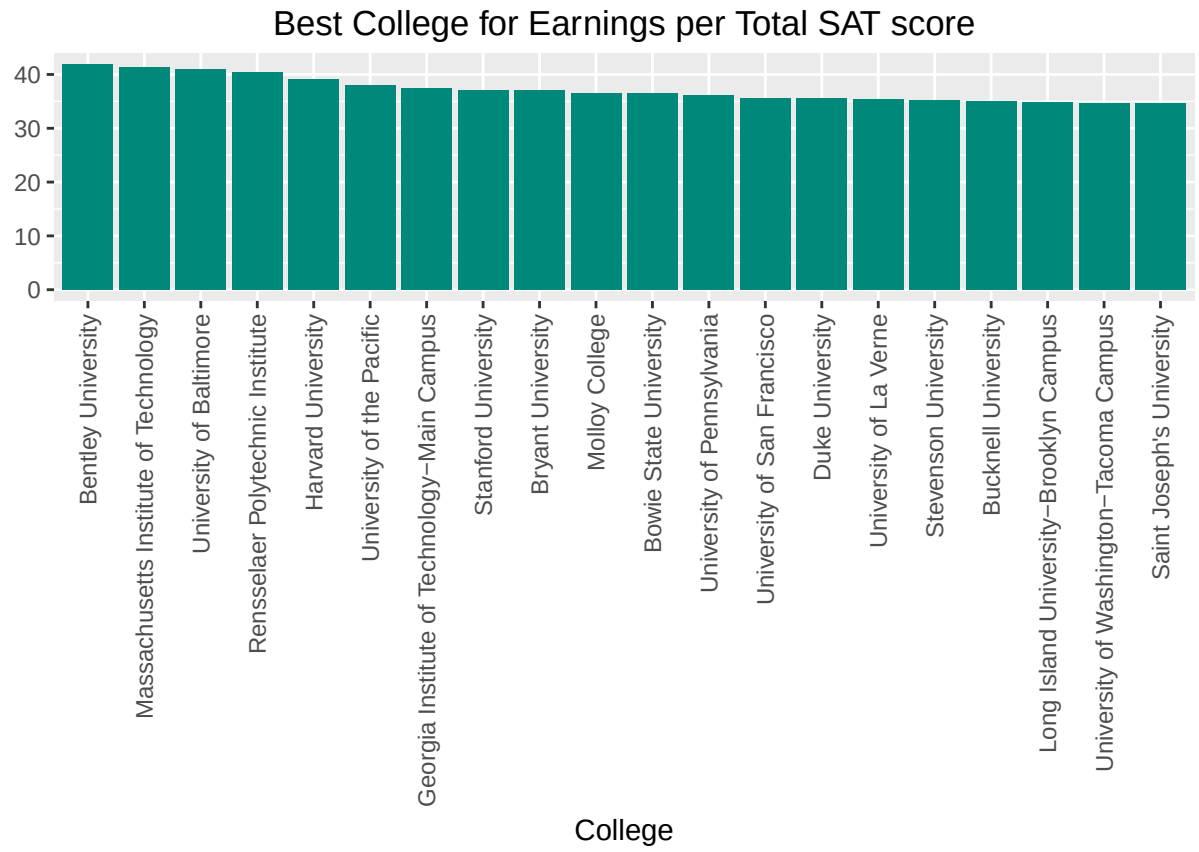
```
top_sat_salary = sat_salary[order(sat_salary$salary_to_sat, decreasing = T),]
```

See top twenty.

```
top_20_sat_salary <- top_sat_salary[1:20,]

head(top_20_sat_salary)
```

```
##                                 College       CollegeType md_earn_wne_p10
## 8                    Bentley University Private nonprofit           74900
## 1  Massachusetts Institute of Technology Private nonprofit           91600
## 44               University of Baltimore            Public           58000
## 3       Rensselaer Polytechnic Institute Private nonprofit           81700
## 2                    Harvard University Private nonprofit           87200
## 18             University of the Pacific Private nonprofit           66400
##    UGDS SATMTMID SATVRMID SATWRMID total_sat salary_to_sat
## 8  4161      600      588      600      1788      41.89038
## 1  4363      770      720      725      2215      41.35440
## 44 3230      465      485      465      1415      40.98940
## 3  5240      715      660      645      2020      40.44554
## 2  7245      750      740      740      2230      39.10314
## 18 3872      614      570      565      1749      37.96455
```

```
#Makes College into an ordered factor already so ggplot doesn't reorder it for me.

top_20_sat_salary$College <- factor(top_20_sat_salary$College, levels = top_20_sat_salary$College)

ggplot(data=top_20_sat_salary, aes(x=top_20_sat_salary$College, y=top_20_sat_salary$salary_to_sat)) +
    geom_bar(stat="identity", fill="#00897B") +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
  xlab("College") + ylab('') +
  ggtitle("Best College for Earnings per Total SAT score")
```



Top famous universities have the best earnings per total SAT score, but some not as famous universities done well as well. There aren't any community colleges that have reached the top 20 list.

Best Salary for Education Cost

```
edu_cost <- dbGetQuery(db, "SELECT INSTNM College,
       CONTROL CollegeType,
       md_earn_wne_p10,
       UGDS,
       COSTT4_A
FROM Scorecard
WHERE Year=2011
AND md_earn_wne_p10 IS NOT NULL
AND md_earn_wne_p10 != 'PrivacySuppressed'
AND UGDS IS NOT NULL
AND UGDS > 3000
ORDER BY md_earn_wne_p10 DESC")
```

20 most expensive large Colleges:

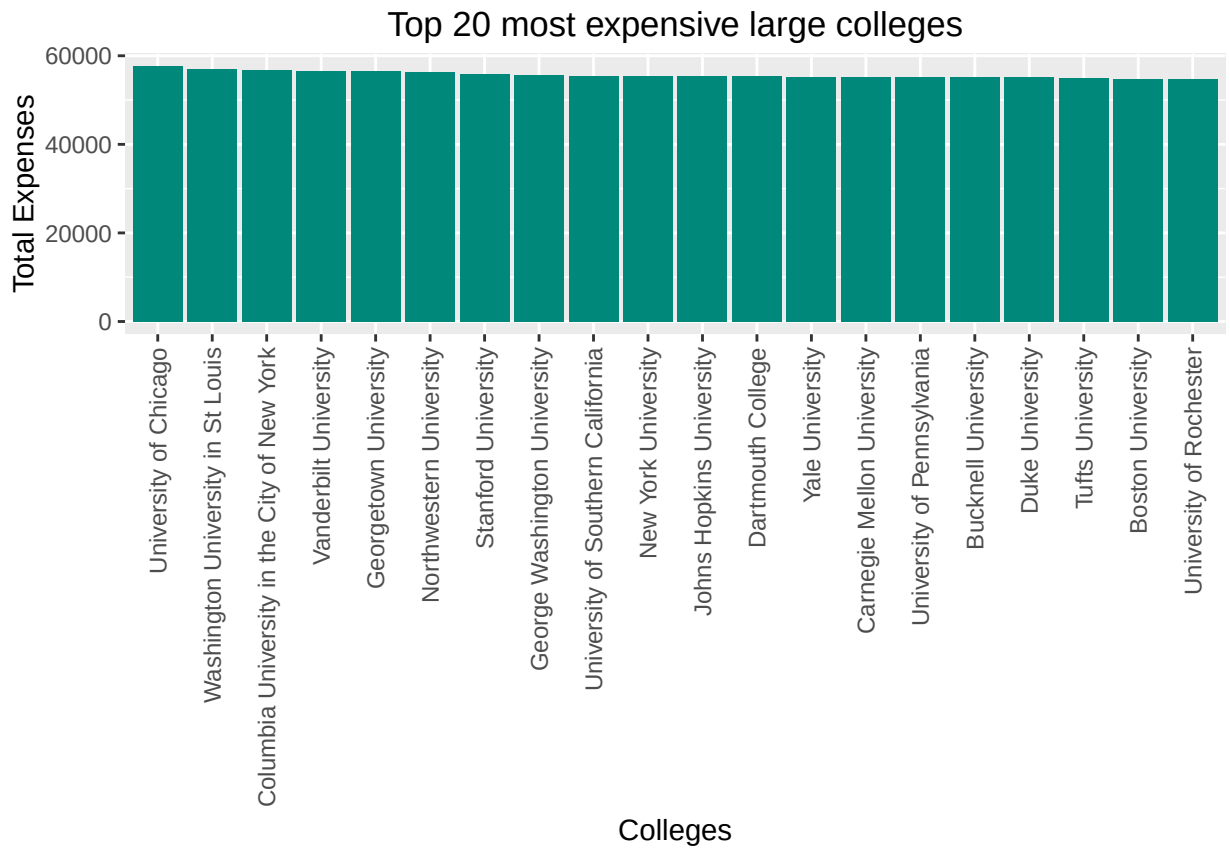
Let's take a look at the 20 most expensive large colleges before seeing which is best bang for you buck.

```r
top_edu_cost = edu_cost[order(edu_cost$COSTT4_A, decreasing = T),]

top_20_edu_cost = edu_cost[order(edu_cost$COSTT4_A, decreasing = T),][1:20,]

top_20_edu_cost
```

```
##                                       College       CollegeType
## 37                        University of Chicago Private nonprofit
## 39           Washington University in St Louis Private nonprofit
## 16 Columbia University in the City of New York Private nonprofit
## 48                          Vanderbilt University Private nonprofit
## 3                           Georgetown University Private nonprofit
## 35                        Northwestern University Private nonprofit
## 5                            Stanford University Private nonprofit
## 34                George Washington University Private nonprofit
## 30           University of Southern California Private nonprofit
## 62                          New York University Private nonprofit
## 21                       Johns Hopkins University Private nonprofit
## 26                             Dartmouth College Private nonprofit
## 31                               Yale University Private nonprofit
## 17                   Carnegie Mellon University Private nonprofit
## 6                    University of Pennsylvania Private nonprofit
## 23                           Bucknell University Private nonprofit
## 9                                Duke University Private nonprofit
## 24                             Tufts University Private nonprofit
## 51                           Boston University Private nonprofit
## 89                        University of Rochester Private nonprofit
##    md_earn_wne_p10  UGDS COSTT4_A
## 37           62800  5377    57590
## 39           62300  6658    56930
## 16           72900  8127    56681
## 48           60900  6754    56634
## 3            83300  7232    56485
## 35           64100  8991    56406
## 5            80900  6927    55918
## 34           64500 10184    55625
## 30           66100 17090    55493
## 62           58800 21820    55412
## 21           69200  5817    55390
## 26           67100  4106    55386
## 31           66000  5333    55300
## 17           72000  5848    55286
## 6            78200 10720    55250
## 23           68800  3535    55180
## 9            76700  6534    55150
## 24           67800  5136    55000
## 51           60600 16575    54836
## 89           55500  5457    54730
```

```r
top_20_edu_cost$College <- factor(top_20_edu_cost$College, levels = top_20_edu_cost$College)

ggplot(data=top_20_edu_cost, aes(x=top_20_edu_cost$College, y=top_20_edu_cost$COSTT4_A)) +
    geom_bar(stat="identity", fill="#00897B") + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0
```

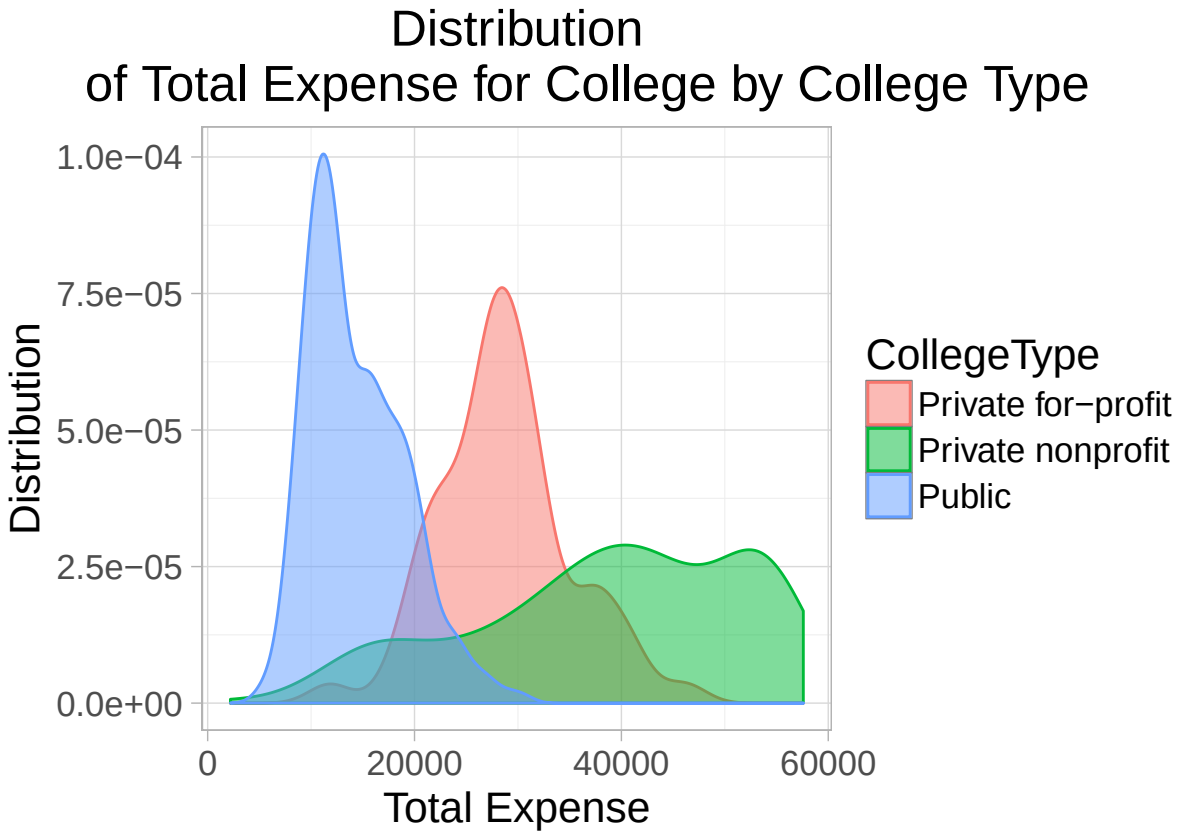
```
  xlab("Colleges") + ylab("Total Expenses") +
  ggtitle("Top 20 most expensive large colleges")
```



Here we see the top 20 most expensive colleges. Maybe of these colleges are in large cities or expensive areas.

```
ggplot(top_edu_cost, aes(x=top_edu_cost$COSTT4_A, color=CollegeType, fill=CollegeType, group=CollegeTyp
  geom_density(alpha=0.5) +
  theme_light(base_size=16) +
  xlab("Total Expense") + ylab("Distribution") +
  ggtitle("Distribution
          of Total Expense for College by College Type")
```

```
## Warning: Removed 15 rows containing non-finite values (stat_density).
```



We see that private non-profit ranges from not very expense to very expense, of almost 60k a year. While public colleges are mostly less than 20k per year. Private for-profit is more expensive than public but in most cases less than non profit.

Get the best ratio of earnings to cost

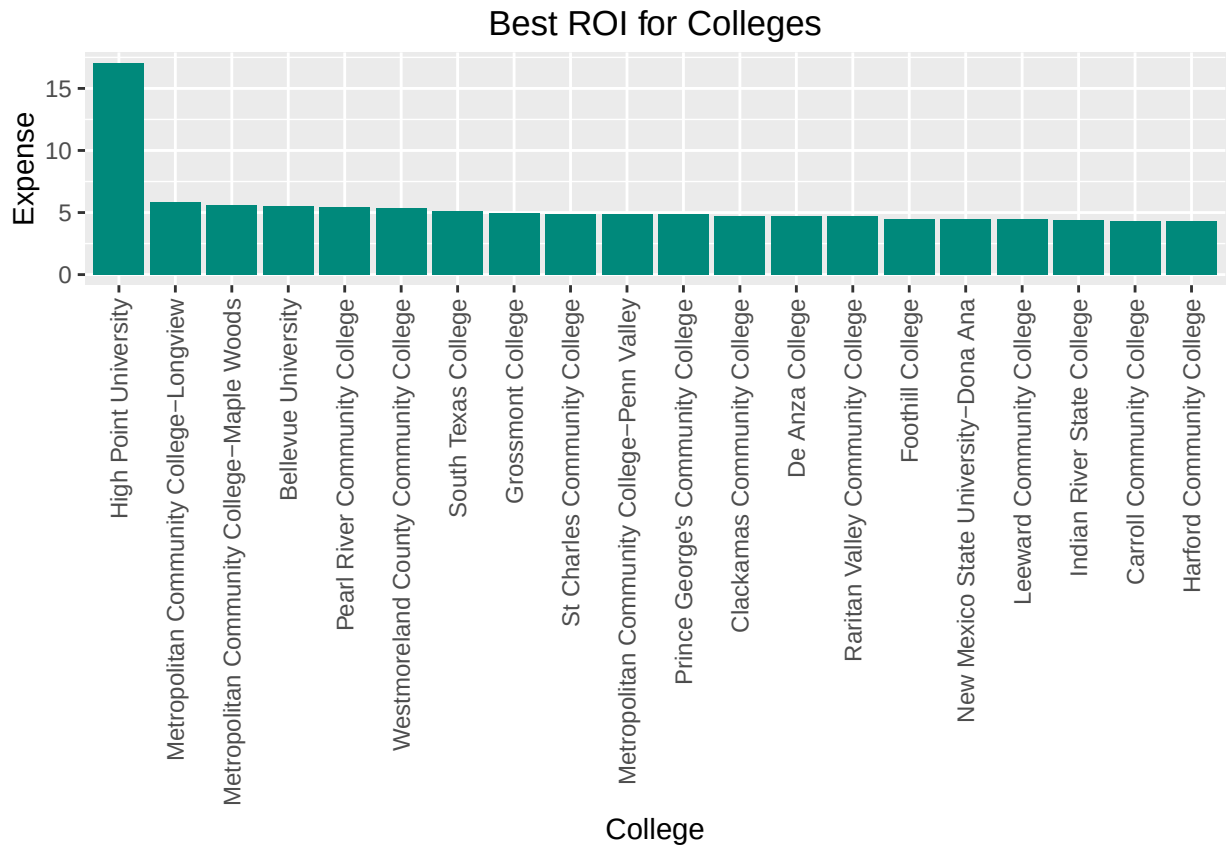
```
edu_cost$salary_to_cost <- edu_cost$md_earn_wne_p10 / edu_cost$COSTT4_A
head(edu_cost)
```

```
##                                College       CollegeType md_earn_wne_p10
## 1 Massachusetts Institute of Technology Private nonprofit           91600
## 2                   Harvard University Private nonprofit           87200
## 3                 Georgetown University Private nonprofit           83300
## 4      Rensselaer Polytechnic Institute Private nonprofit           81700
## 5                  Stanford University Private nonprofit           80900
## 6           University of Pennsylvania Private nonprofit           78200
##    UGDS COSTT4_A salary_to_cost
## 1  4363    53210       1.721481
## 2  7245    53950       1.616311
## 3  7232    56485       1.474728
## 4  5240    54035       1.511983
## 5  6927    55918       1.446761
## 6 10720    55250       1.415385
```

```
best_deal <- edu_cost[order(edu_cost$salary_to_cost, decreasing = T),]
head(best_deal)
```

```
##                                       College       CollegeType
## 556                       High Point University Private nonprofit
## 1041    Metropolitan Community College-Longview            Public
## 1042 Metropolitan Community College-Maple Woods            Public
## 138                         Bellevue University Private nonprofit
## 1135             Pearl River Community College            Public
## 1054      Westmoreland County Community College            Public
##      md_earn_wne_p10 UGDS COSTT4_A salary_to_cost
## 556            37500 3960     2200      17.045455
## 1041           29400 5066     5006       5.872952
## 1042           29400 4135     5279       5.569237
## 138            52200 6312     9495       5.497630
## 1135           28100 5390     5143       5.463737
## 1054           29300 6249     5490       5.336976
```

```
best_deal_top20 <- best_deal[1:20,]

best_deal_top20$College <- factor(best_deal_top20$College, levels = best_deal_top20$College)

ggplot(data=best_deal_top20, aes(x=best_deal_top20$College, y=best_deal_top20$salary_to_cost)) +
    geom_bar(stat="identity", fill="#00897B") +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
  xlab("College") + ylab("Expense") +
  ggtitle("Best ROI for Colleges")
```



We see that a lot of community colleges have the best of ROI but the investment isn't just cost of education,

there is also the time it cost to complete the education. However, this shows that community colleges still are a good idea due to its high ROI.

---

6. Graduation Rates

Graduation Rate and Graduation Rate to SAT ratio

```
grad_rate <- dbGetQuery(db, "SELECT INSTNM College,
       CONTROL CollegeType,
       md_earn_wne_p10,
       UGDS,
       SATMTMID,
       SATVRMID,
       SATWRMID,
       C150_4

FROM Scorecard

WHERE Year=2011
AND md_earn_wne_p10 IS NOT NULL
AND md_earn_wne_p10 != 'PrivacySuppressed'
AND UGDS IS NOT NULL
AND UGDS > 3000
AND SATMTMID IS NOT NULL
AND SATVRMID IS NOT NULL
AND SATWRMID IS NOT NULL
AND C150_4 IS NOT NULL
ORDER BY C150_4 DESC")
```

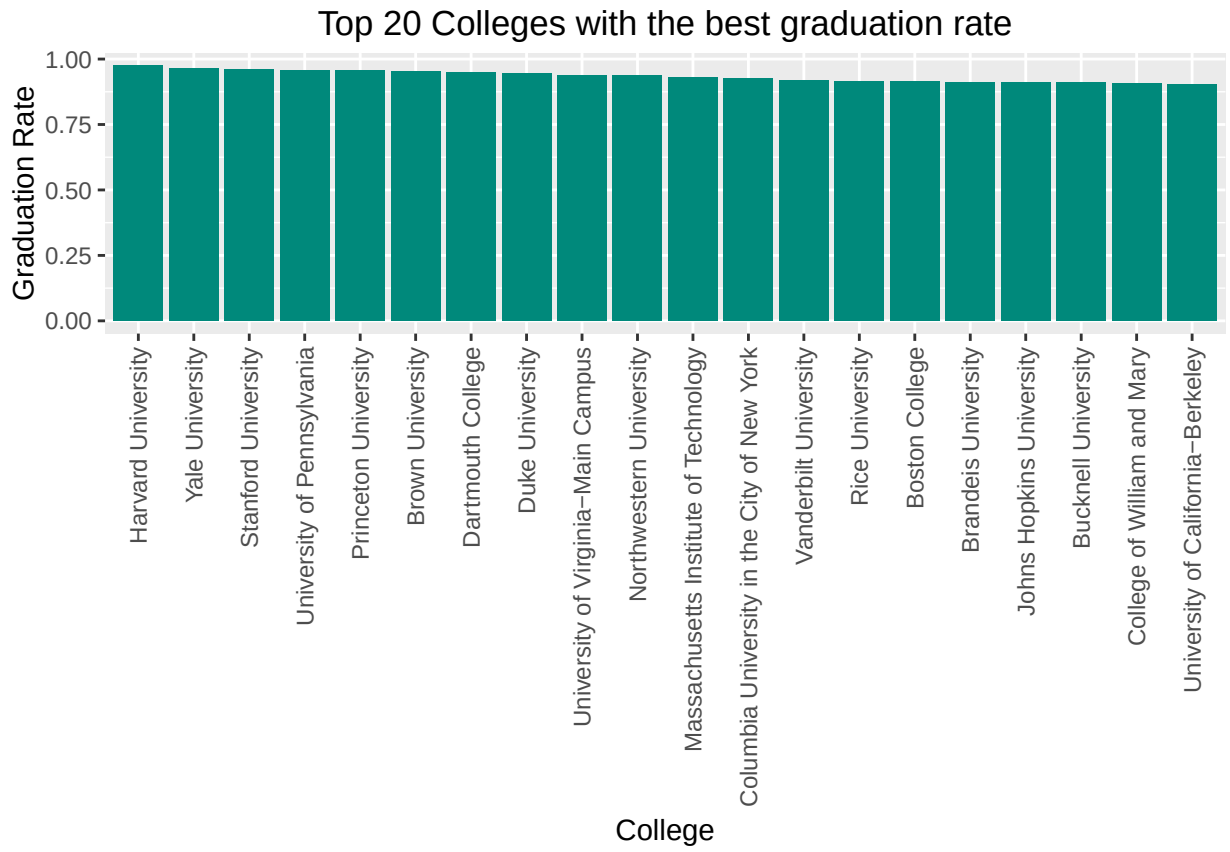
Top 20 schools with the best Graduation Rate:

```
top_grad_rate = grad_rate[order(grad_rate$C150_4, decreasing = T),]

top_20_grad_rate = top_grad_rate[1:20,]

top_20_grad_rate
```

```
##                                       College       CollegeType
## 1                           Harvard University Private nonprofit
## 2                              Yale University Private nonprofit
## 3                            Stanford University Private nonprofit
## 4                     University of Pennsylvania Private nonprofit
## 5                           Princeton University Private nonprofit
## 6                              Brown University Private nonprofit
## 7                            Dartmouth College Private nonprofit
## 8                              Duke University Private nonprofit
## 9            University of Virginia-Main Campus            Public
## 10                       Northwestern University Private nonprofit
## 11       Massachusetts Institute of Technology Private nonprofit
## 12 Columbia University in the City of New York Private nonprofit
```

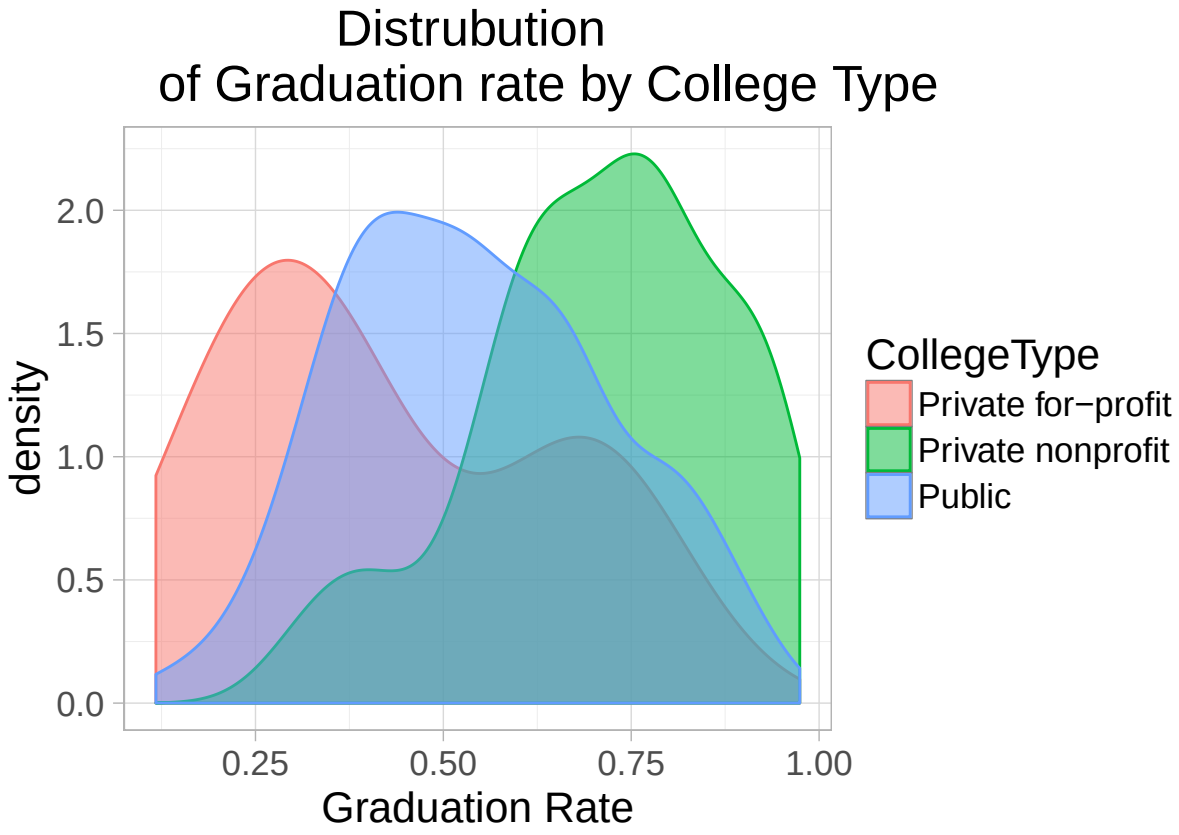
```
## 13                      Vanderbilt University Private nonprofit
## 14                            Rice University Private nonprofit
## 15                             Boston College Private nonprofit
## 16                          Brandeis University Private nonprofit
## 17                        Johns Hopkins University Private nonprofit
## 18                           Bucknell University Private nonprofit
## 19                  College of William and Mary            Public
## 20            University of California-Berkeley            Public
##    md_earn_wne_p10  UGDS SATMTMID SATVRMID SATWRMID C150_4
## 1            87200  7245      750      740      740 0.9743
## 2            66000  5333      750      750      755 0.9659
## 3            80900  6927      735      720      730 0.9614
## 4            78200 10720      735      705      720 0.9580
## 5            75100  5160      755      745      745 0.9551
## 6            59700  6118      705      685      695 0.9521
## 7            67100  4106      740      725      740 0.9497
## 8            76700  6534      735      705      720 0.9437
## 9            58600 14568      685      665      670 0.9390
## 10           64100  8991      740      715      725 0.9360
## 11           91600  4363      770      720      725 0.9286
## 12           72900  8127      745      735      735 0.9280
## 13           60900  6754      740      725      715 0.9185
## 14           59900  3708      730      700      705 0.9155
## 15           67000  9464      685      665      680 0.9153
## 16           58800  3493      685      655      675 0.9118
## 17           69200  5817      720      680      700 0.9112
## 18           68800  3535      675      635      650 0.9098
## 19           56400  6020      670      675      670 0.9074
## 20           62700 25885      695      665      675 0.9049
```

```r
top_20_grad_rate$College <- factor(top_20_grad_rate$College, levels = top_20_grad_rate$College)

ggplot(data=top_20_grad_rate, aes(x=top_20_grad_rate$College, y=top_20_grad_rate$C150_4)) +
    geom_bar(stat="identity", fill="#00897B") +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
    xlab("College") + ylab("Graduation Rate") +
  ggtitle("Top 20 Colleges with the best graduation rate")
```



The graduation rates of top university are similar and most of them approaching 100%. We will see later on that there is a significant influence of graduation rate from SAT scores.

```
ggplot(top_grad_rate, aes(x=top_grad_rate$C150_4, color=CollegeType,
                          fill=CollegeType, group=CollegeType)) +
  geom_density(alpha=0.5) +
  theme_light(base_size=16) +
  xlab("Graduation Rate") +
  ggtitle("Distrubution
          of Graduation rate by College Type")
```



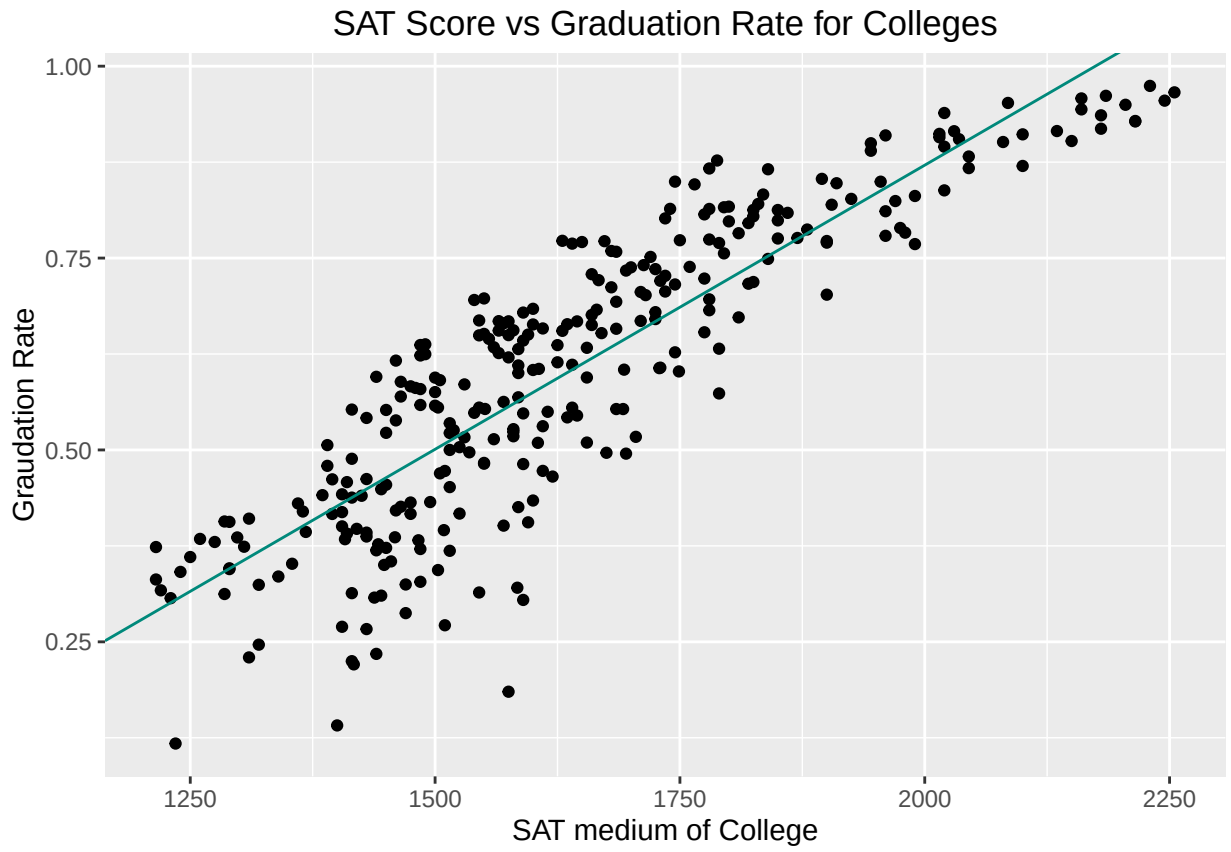
We see that private non-profit has the highest gradation rate while private for-profit has the worse graduation rate. There is a dual distribution for private for-profit colleges, one distribution has poor results and one distribution has results similar to non-profit private colleges. This dual distribution for private for-profit is seen in later results as well.

Does having higher SAT score improve graduation rate?

```
grad_rate$total_sat <- grad_rate$SATMTMID + grad_rate$SATVRMID + grad_rate$SATWRMID

fit <- lm(grad_rate$C150_4~grad_rate$total_sat)

coef(fit)
```

```
##          (Intercept) grad_rate$total_sat
##       -0.6096836814        0.0007403611
```

```
ggplot(grad_rate, aes(x=grad_rate$total_sat, y=grad_rate$C150_4, group=1)) +
  geom_point() + geom_abline(intercept = -0.6096836814, slope = 0.0007403611 , colour='#00897B') +
  xlab("SAT medium of College") + ylab("Graudation Rate") +
  ggtitle("SAT Score vs Graduation Rate for Colleges")
```




```
summary(fit)
```

```
##
## Call:
## lm(formula = grad_rate$C150_4 ~ grad_rate$total_sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37139 -0.05649  0.01177  0.06695  0.17540
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -6.097e-01  3.941e-02  -15.47   <2e-16 ***
## grad_rate$total_sat  7.404e-04  2.382e-05   31.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09304 on 292 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7671
## F-statistic: 965.8 on 1 and 292 DF,  p-value: < 2.2e-16
```

There is a clear positive relationship between SAT and graduation rate at least for colleges.

```r
predictions <- predict(fit, grad_rate)

rmse <- mean((grad_rate$C150_4 - predictions)^2)

print(rmse)
```

```
## [1] 0.008597173
```

Therefore, the Root mean square error is 0.008597 which is acceptable for this fit.

---

7. Conclusions

This was an in-depth exploratory data analysis. The idea is to understand and get familiar with this extremely large and rich dataset. Most of the understanding was conceived from plots of the data set in various ways, it is the graphs that can tell a very interesting story. This is a data set that is 1.1 GB in size with 1731 columns and 124699 rows. From 1996 to 2013, there is a trend of number of colleges is increasing. From our analysis, it is observed that SAT scores resemble a normal distribution.

Private non-profit colleges have the best earnings while public colleges have lower earnings for students. Private for-profit schools are made from two distribution, one distribution that is similar to private non-profit colleges while one distribution creates poor results. Students that are looking at Private for-profit colleges should pay extra attention.

Most expensive colleges are within large cities and have a relatively famous reputation. Private nonprofit colleges are sometimes the most expensive as well. The best ROI in terms of cost are community colleges while the best ROI in terms of SAT score are famous colleges. This shows that high school students should work harder to get into high SAT score colleges as the reward is not linear. Finally, it was observed that the graduation rate is highest in famous colleges and graduation rate of a college is positively linked with SAT scores of that college.

The success rate in private nonprofit colleges is very high but can be more expensive, this might be worth it for students that like to invest in their education. Public college for the most part creates good results and most students can be very successful. Private for-profit colleges are a hit or miss, be very careful when selecting for-profit colleges.

There are many variables within this dataset that are not investigated as there are hundreds of variables. However, the most important variables and the variables that students are most interested in are discussed here. From the conclusions here, students should have a bigger understand of the general landscape of higher education and have a good comparison when a specific school is looked at.