Educational data Exploretion

Recently, the US government released college scorecard data, a rich set of data on all US undergraduate degree-granting institutions.

library(RSQLite)

## Loading required package: DBI

library(ggplot2)  
library(tidyr)

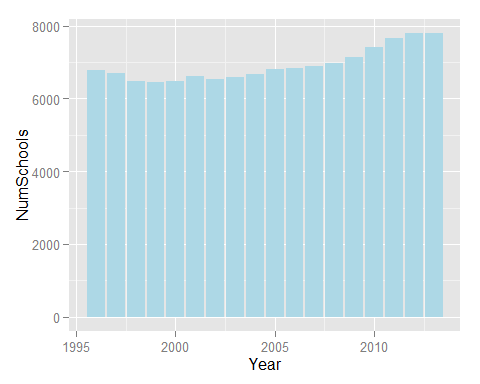
## Warning: package 'tidyr' was built under R version 3.2.3

setwd("I:/R Data")  
db = dbConnect(SQLite(), "database.sqlite")  
dbListTables(db)

## [1] "Scorecard"

Each row corresponds to data on a US school in a given year. The years covered in this data range from 1996 to 2013. Here, you can see the number of schools we have data on for each year.

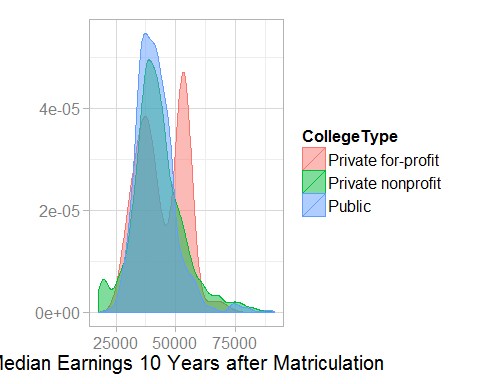
x = dbGetQuery(db, "SELECT Year, COUNT(Id) NumSchools FROM Scorecard GROUP BY Year")  
ggplot(x, aes(x = Year, y = NumSchools)) + geom\_bar(stat="identity", fill="lightblue")



This data is very wide and rich: it has 1,731 associated columns. One of the most interesting aspects of this data is the earnings information it contains on students once they graduate. This includes the median earnings years out, as well as the top/bottom quartiles and deciles. Here, we’ll query this data for earnings 10 years after matriculation.

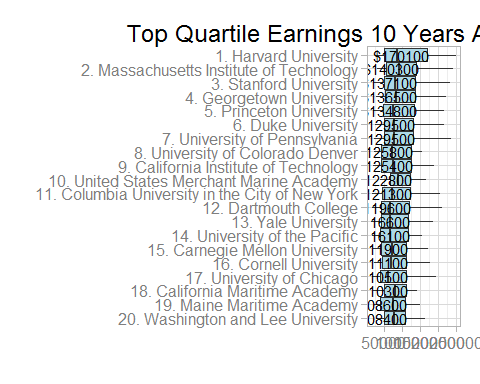
From this earnings data, we can see the distribution of earnings 10 years after matriculation at undergraduate institutions. We’ll split this into three types of schools: private non-profit, private for-profit, and public. The median earnings for public and private non-profit schools look similar over the bulk of the earnings range. At the very low end, there are more private non-profit schools than public schools. Curiously, the median earnings distribution for private for-profit schools is bimodal.

earnings <- dbGetQuery(db, "  
SELECT s11.INSTNM College,  
 s11.CONTROL CollegeType,  
 s11.md\_earn\_wne\_p10 e50,  
 s11.pct10\_earn\_wne\_p10 e10,  
 s11.pct25\_earn\_wne\_p10 e25,  
 s11.pct75\_earn\_wne\_p10 e75,  
 s11.pct90\_earn\_wne\_p10 e90  
FROM Scorecard s11  
  
INNER JOIN Scorecard s13 ON s11.UNITID=s13.UNITID  
WHERE s11.Year=2011  
 AND s13.Year=2013  
 AND s11.pct75\_earn\_wne\_p10 IS NOT NULL  
 AND s11.pct75\_earn\_wne\_p10 != 'PrivacySuppressed'  
 AND s11.PREDDEG = 'Predominantly bachelor''s-degree granting'  
 AND s13.CCBASIC NOT LIKE '%Special%'  
ORDER BY s11.pct75\_earn\_wne\_p10 DESC")  
earnings <- cbind(Rank=1:nrow(earnings), earnings)  
earnings$College <- paste(earnings$Rank, earnings$College, sep=". ")  
earnings$College <- factor(earnings$College, levels=rev(earnings$College))  
  
ggplot(earnings, aes(x=e50, color=CollegeType, fill=CollegeType)) +  
 geom\_density(alpha=0.5) +  
 theme\_light(base\_size=15) +  
 xlab("Median Earnings 10 Years after Matriculation") + ylab("")



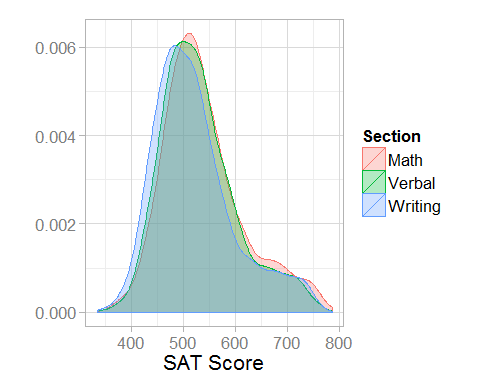
One of the first questions that comes to mind is “what colleges have students that earn the most?” The graph below ranks colleges by their top quartile of earnings ten years after the students matriculated, and plots the distribution of student earnings. Some of the usual suspects appear on top: Harvard, MIT, Stanford, Geogetown, Princeton, and Duke.

ggplot(earnings[1:20,], aes(x=College, ymin=e10, lower=e25, middle=e50, upper=e75, ymax=e90)) +  
 geom\_boxplot(stat="identity", fill="lightblue") +   
 geom\_text(aes(x=College, y=e75-2000, hjust=0.95, label=paste0("$", e75)), size=4) +   
 theme(axis.text.y = element\_text(hjust=0, color="black")) +  
 theme\_light(base\_size=15) +  
 coord\_flip() +  
 xlab("") + ylab("") +  
 ggtitle("Top Quartile Earnings 10 Years After Matriculation ($)")



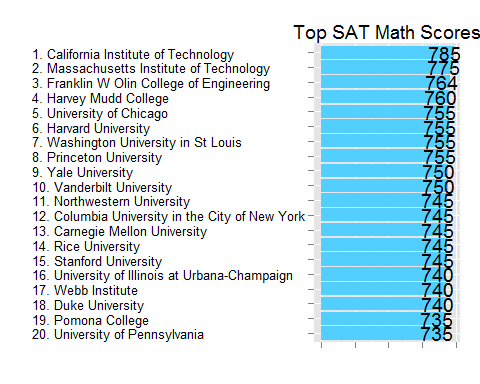
The data includes the median SAT scores of admitted students for each of the colleges. We can see the distribution of these median scores across all colleges in the data. The distribution of median section scores for admitted students at colleges is remarkably consistent between the SAT Math, Verbal, and Writing sections.

sat <- dbGetQuery(db, "  
SELECT INSTNM College,  
 SATMTMID Math,  
 SATVRMID Verbal,  
 SATWRMID Writing  
FROM Scorecard  
WHERE Year=2013  
 AND SATMTMID IS NOT NULL  
 AND SATMTMID != 'PrivacySuppressed'  
 AND SATVRMID IS NOT NULL  
 AND SATVRMID != 'PrivacySuppressed'  
 AND SATWRMID IS NOT NULL  
 AND SATWRMID != 'PrivacySuppressed'")  
  
ggplot(sat %>% gather(Section, Score, -College), aes(x=Score, color=Section, fill=Section, group=Section)) +  
 geom\_density(alpha=0.3) +  
 theme\_light(base\_size=16) +  
 xlab("SAT Score") +  
 ylab("")



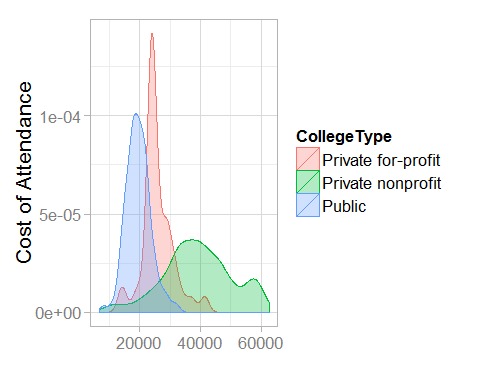
We can also see what schools have the top median SAT Math.

topmath = sat[order(sat[,2], decreasing = T)[1:20],1:2]  
topmath <- cbind(Rank=1:nrow(topmath), topmath)  
topmath$College <- paste(topmath$Rank, topmath$College, sep=". ")  
topmath$College <- factor(topmath$College, levels=rev(topmath$College))  
  
ggplot(topmath, aes(x = College, y = Math)) +  
  
 geom\_bar(stat = "identity", fill = "#53cfff") +  
 geom\_text(aes(x = College, y = Math-50,label = Math)) +   
  
 theme(axis.text.y = element\_text(hjust=0, color="black"), axis.text.x=element\_blank()) +  
 coord\_flip() +  
 xlab("") + ylab("")+  
 ggtitle("Top SAT Math Scores")



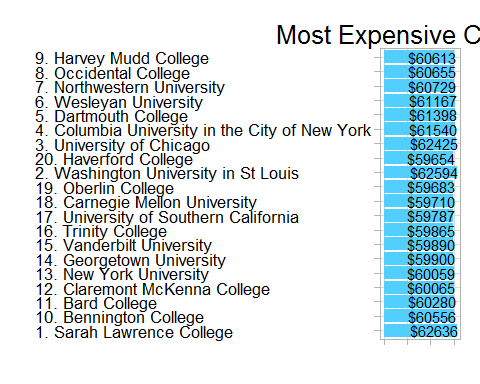
The cost here reflects the average annual total cost of attendance, including tuition and fees, books and supplies, and living expenses for all full-time, first-time, degree/certificate-seeking undergraduates who receive Title IV aid.

cost <- dbGetQuery(db, "  
SELECT INSTNM College,  
 COSTT4\_A Cost,  
 CONTROL CollegeType  
FROM Scorecard  
WHERE Year=2013  
 AND PREDDEG='Predominantly bachelor''s-degree granting'  
 AND CCBASIC NOT LIKE '%Special Focus%'  
 AND COSTT4\_A IS NOT NULL  
ORDER BY COSTT4\_A DESC")  
cost <- cbind(Rank=1:nrow(cost), cost)  
cost$College <- paste(cost$Rank, cost$College, sep=". ")  
  
ggplot(cost, aes(x=Cost, color=CollegeType, fill=CollegeType, group=CollegeType)) +  
 geom\_density(alpha=0.3) +  
 theme\_light(base\_size=16) +  
 xlab("") + ylab("Cost of Attendance")



I got sticker shock when I saw the total cost of attendance for the most expensive colleges - this has risen dramatically even since I graduated five years ago.

ggplot(cost[1:20,], aes(x=College, y=Cost)) +  
 geom\_bar(stat="identity", fill="#53cfff") +  
 geom\_text(aes(x=College, y=Cost-500, ymax=Cost, hjust=0.95, label=paste0("$", Cost)), size=4) +   
 theme\_light(base\_size=16) +  
 theme(axis.text.y = element\_text(hjust=0, color="black"), axis.text.x=element\_blank()) +  
 xlab("") + ylab("") +  
 coord\_flip() +  
 ggtitle("Most Expensive Colleges")



cost$College <- factor(cost$College, levels=rev(cost$College))