

College Scorecard R Script

2025-06-25

To start we read in the data and put the tidyverse and other packages in the library. Remember you must install each package once and only once on your machine.

```
# install.packages("tidyverse")
# install.packages("reshape")
# install.packages("RColorBrewer")
# install.packages("scales")

library(tidyverse)
library(reshape)
library(RColorBrewer)
library(scales)

# install.packages("ggrepel")
library(ggrepel)
```

Read in the college scorecard data and take a look at it.

```
csc <- read.csv("Most-Recent-Cohorts-Institution.csv", header = TRUE,
               stringsAsFactors = FALSE)
View(csc)
dim(csc)
```

Now let's make the first scatterplot using the base R plot function.

```
# Lets compare verbal and math SAT scores Using 25th percentile by Institution
df_SAT <- data.frame(csc$INSTNM, csc$SATMT25, csc$SATVR25)
View(df_SAT)
colnames(df_SAT) <- c("INSTNM", "SATMT25", "SATVR25")

df_SAT <- select(csc, INSTNM, SATMT25, SATVR25)
View(df_SAT)

plot(df_SAT$SATMT25, df_SAT$SATVR25, xlim= c(200,800), ylim = c(200,800),
     xlab = "Math SAT 25th Percentile",
     ylab = "Verbal SAT 25th Percentile")
```

To find the minimum we need to be careful of missing values.

```
# Find the minimum but watch out for NA's
min(df_SAT$SATMT25)
summary(df_SAT$SATMT25)
which(df_SAT$SATMT25 == 225)
df_SAT[1686, ]
```

Make another scatterplot with median lines and $y = x$.

```
# let's add some lines and use na.rm = TRUE

plot(df_SAT$SATMT25, df_SAT$SATVR25, xlim= c(200,800), ylim = c(200,800),
     xlab = "Math SAT 25th Percentile",
     ylab = "Verbal SAT 25th Percentile", pch = 19, cex = 0.6)
abline(0,1)
abline(v=median(df_SAT$SATMT25, na.rm = TRUE), lty = "dashed", col = "red")
abline(h=median(df_SAT$SATVR25, na.rm = TRUE), lty = "dashed", col = "red")
```

Make the same plot using ggplot this time, and add in smoother and linear trendline.

```
# just with a smooth geom showing pattern
ggplot(df_SAT, aes(SATMT25, SATVR25)) +
  geom_point() + xlim(200,800)+ ylim(200,800) +
  labs(x = "Math SAT 25th Percentile",
       y = "Verbal SAT 25th Percentile") +
  geom_smooth()

# now with a linear model to get linear trendline
ggplot(df_SAT, aes(SATMT25, SATVR25)) +
  geom_point() + xlim(200,800)+ ylim(200,800) +
  labs(x = "Math SAT 25th Percentile",
       y = "Verbal SAT 25th Percentile") +
  geom_smooth(method = "lm")

# To see the coefficients (slope and y-int) of linear trendline use summary function again
summary(lm((df_SAT$SATVR25) ~ (df_SAT$SATMT25)))
```

Let's conduct some exploratory data analysis using the power of R!

```
# Exploratory data analysis
# What about R? Ha!
# Note na.rm does not work.
# cor(df_SAT$SATMT25, df_SAT$SATVR25, na.rm = TRUE)
r = cor(df_SAT$SATMT25, df_SAT$SATVR25, use = "complete.obs")
r
r^2

#now let's compare cost of attendance and admissions rate
summary(csc$COSTT4_A)
summary(csc$ADM_RATE_ALL)
summary(csc$ADM_RATE)

which(csc$COSTT4_A == max(csc$COSTT4_A, na.rm = T))
csc$INSTNM[2094]

which(csc$INSTNM == "Harvard University")
csc$COSTT4_A[1421]

#how to search for part of a string
grep("Harvard", csc$INSTNM)
```

```
length(grep("University", csc$INSTNM))
length(grep("College", csc$INSTNM))

# let's subset only the universities using this grep vector!
Univ_df <- csc[grep("University", csc$INSTNM), ]
View(Univ_df)

# let's subset for bachelor's degree awarding institutions
bach <- filter(csc, PREDDEG == 3)
View(bach)
dim(bach)
summary(bach$COSTT4_A)
summary(bach$ADM_RATE_ALL)
summary(bach$ADM_RATE)

#what is Jewish Theo Sem adm rate?
csc$ADM_RATE_ALL[2094]
```

Next scatterplot.

```
#scatterplot using full csc df since there are so many missing values
plot(csc$ADM_RATE_ALL, csc$COSTT4_A, xlab = "Admissions Rate",
     ylab = "Total Cost")

#Now let's highlight Maine schools
#First change size of points
cexVals <- rep(0.5, nrow(csc))
cexVals[csc$STABBR == "ME"] = 1

#Next change shape of dots
pchVals <- rep(3, nrow(csc))
pchVals[csc$STABBR == "ME"] = 19

#Now change colors of dots
colVals <- rep(grey(0.8), nrow(csc))
colVals[csc$STABBR == "ME"] = grey(0.2)

plot(csc$ADM_RATE_ALL, csc$COSTT4_A, xlab = "Admissions Rate",
     ylab = "Total Cost",
     cex = cexVals, pch = pchVals, col = colVals)
```

To get the Maine school data labels we create a subset.

```
#Now subset Maine schools for labels using subset instead of filter
Maine <- subset(csc, STABBR == "ME")
Maine <- Maine[, c("INSTNM", "ADM_RATE_ALL", "COSTT4_A", "MD_EARN_WNE_P10")]
View(Maine)
dim(Maine)

# What schools did NOT get plotted in the scatterplot ??

#Next get only complete cases or observations
complete <- complete.cases(Maine$ADM_RATE_ALL, Maine$COSTT4_A)
```

```
complete[1:6]

Maine2 <- Maine[complete, ]
View(Maine2)

#Now add labels
plot(csc$ADM_RATE_ALL, csc$COSTT4_A, xlab = "Admissions Rate",
     ylab = "Total Cost",
     cex = cexVals, pch = pchVals, col = colVals)
text(Maine2$ADM_RATE_ALL, Maine2$COSTT4_A, labels = Maine2$INSTNM,
     pos = 1, cex = 0.5, col = "red")
```

Then we can use ggplot.

```
#let's try ggplot
ggplot(csc, aes(ADM_RATE_ALL, COSTT4_A)) +
  geom_point(shape = pchVals, size = cexVals, color = colVals) +
  labs(x = "Admissions Rate", y = "Total Cost") +
  geom_text(data = Maine, aes(y = COSTT4_A - 2000, label = INSTNM),
           size = 2, vjust = 0, color = "red")

# Going to use ggrepel package for data labels and scales package for axis labels

ggplot(csc, aes(ADM_RATE_ALL, COSTT4_A)) +
  geom_point(shape = pchVals, size = cexVals, color = colVals) +
  labs(x = "Admissions Rate", y = "Total Cost") +
  geom_text_repel(data = Maine, aes(label = INSTNM),
                 size = 2.5, color = "red")

# try getting dollars and percentages on axis scales

ggplot(csc, aes(ADM_RATE_ALL, COSTT4_A)) +
  geom_point(shape = pchVals, size = cexVals, color = colVals) +
  labs(x = "Admissions Rate", y = "Total Cost") +
  geom_text_repel(data = Maine, aes(label = INSTNM),
                 size = 2.5, color = "red") +
  scale_y_continuous(labels = label_dollar()) +
  scale_x_continuous(labels = label_percent(accuracy = 0.1))
```

The End!