W205 Project

Cassie

August 6, 2017

Introduction

This investigation aims to determine whether there is a link between educational funding and outcomes.

The data for our research comes from the National Center for Educational Statistics and is stored in a PostgreSQL database.

This document is a record of our exploratory analysis of the data. We are examining measures of funding, expenditure and test scores and their relationships to identify trends and insights which will help answer our research question.

Exploratory Analysis

For our analysis we have the following variables available in the database:

```
(dbGetQuery(con, "SELECT tablename FROM pg_tables where schemaname='public' order by tablename"))
```

```
##      tablename
## 1       fiscal
## 2     fiscal_1
## 3     fiscal_2
## 4     fiscal_3
## 5     fiscal_4
## 6     fiscal_5
## 7        naep8
## 8      naep8_1
## 9      naep8_2
## 10     naep8_3
## 11     naep8_4
## 12     naep8_5
## 13   nonfiscal
## 14 nonfiscal_1
## 15 nonfiscal_2
## 16 nonfiscal_3
## 17 nonfiscal_4
## 18 nonfiscal_5
```

```
dbListFields(con, c("public", "fiscal"))
```

```
## [1] "survey_year"          "state"                "state_revenue"
## [4] "local_revenue"        "federal_revenue"      "total_revenue"
## [7] "teacher_salaries"     "teacher_benefits"     "current_expenditures"
```

```
dbListFields(con, c("public", "nonfiscal"))
```

```
## [1] "survey_year"     "state"           "total_teachers"  "grade8_students"
## [5] "total_students"
```

```
dbListFields(con, c("public", "naep8"))
```

```
## [1] "test_year"     "state"         "math_score"    "reading_score"
```

We have fiscal and test score data for the following years:

```
dbGetQuery(con, "SELECT DISTINCT survey_year from fiscal order by survey_year desc")
```

```
##    survey_year
## 1         2014
## 2         2013
## 3         2012
## 4         2011
## 5         2010
## 6         2009
## 7         2008
## 8         2007
## 9         2006
## 10        2005
## 11        2004
## 12        2003
## 13        2002
```

```
dbGetQuery(con, "SELECT DISTINCT test_year from naep8 order by test_year desc")
```

```
##   test_year
## 1      2015
## 2      2013
## 3      2011
## 4      2009
## 5      2007
## 6      2005
## 7      2003
```

We will start by examining the measures of revenue and test scores per state over time.

Test Scores by Year and State

Math and reading test scores both fall within the 0-500 range as expected so there are no anomolous values. Reading scores range from 238.2 to 277.0 while math scores range from 243.1 to 300.6 so the math scores are generally higher with a wider range of values.

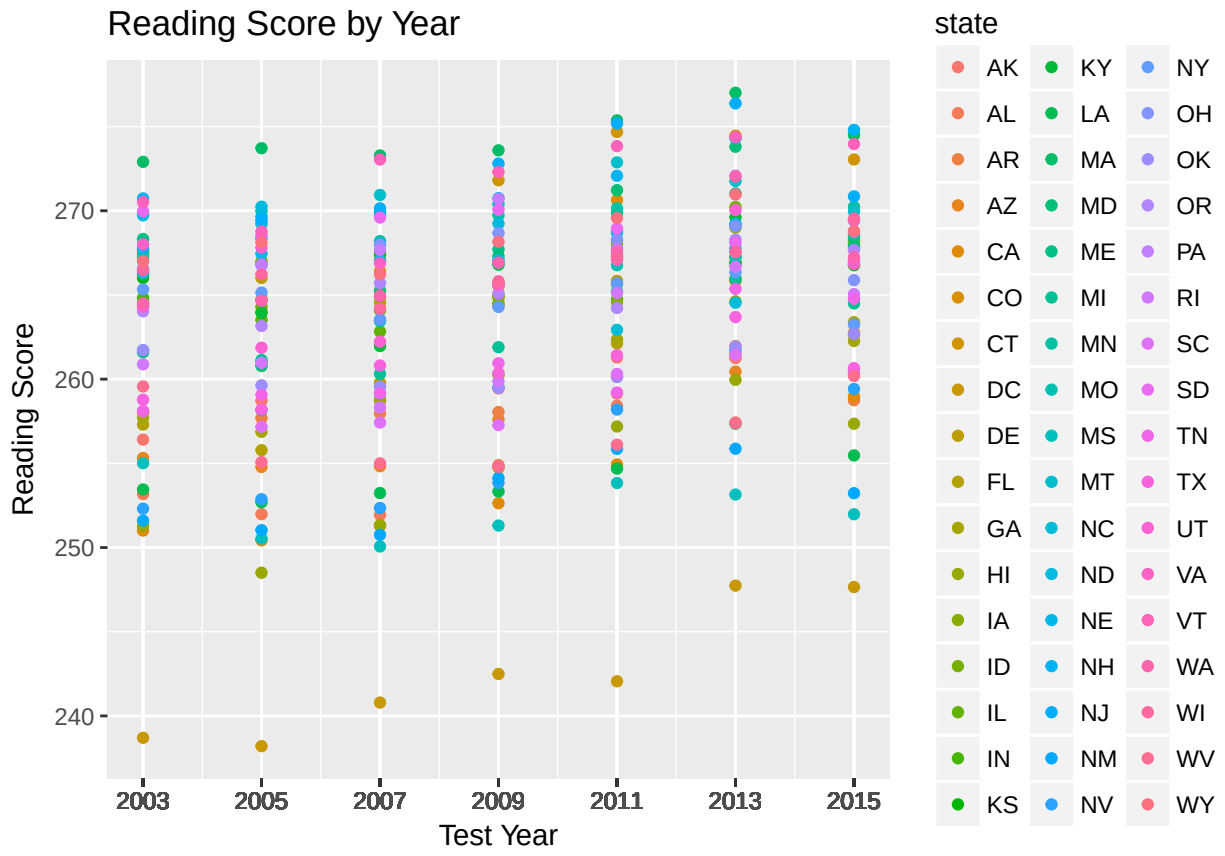
```
df_naep_reading <- dbGetQuery(con, "SELECT test_year,state,reading_score from naep8;")
df_naep_math <- dbGetQuery(con, "SELECT test_year,state,math_score from naep8;")
summary(df_naep_reading$reading_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   238.2   259.4   265.0   263.6   268.1   277.0
```

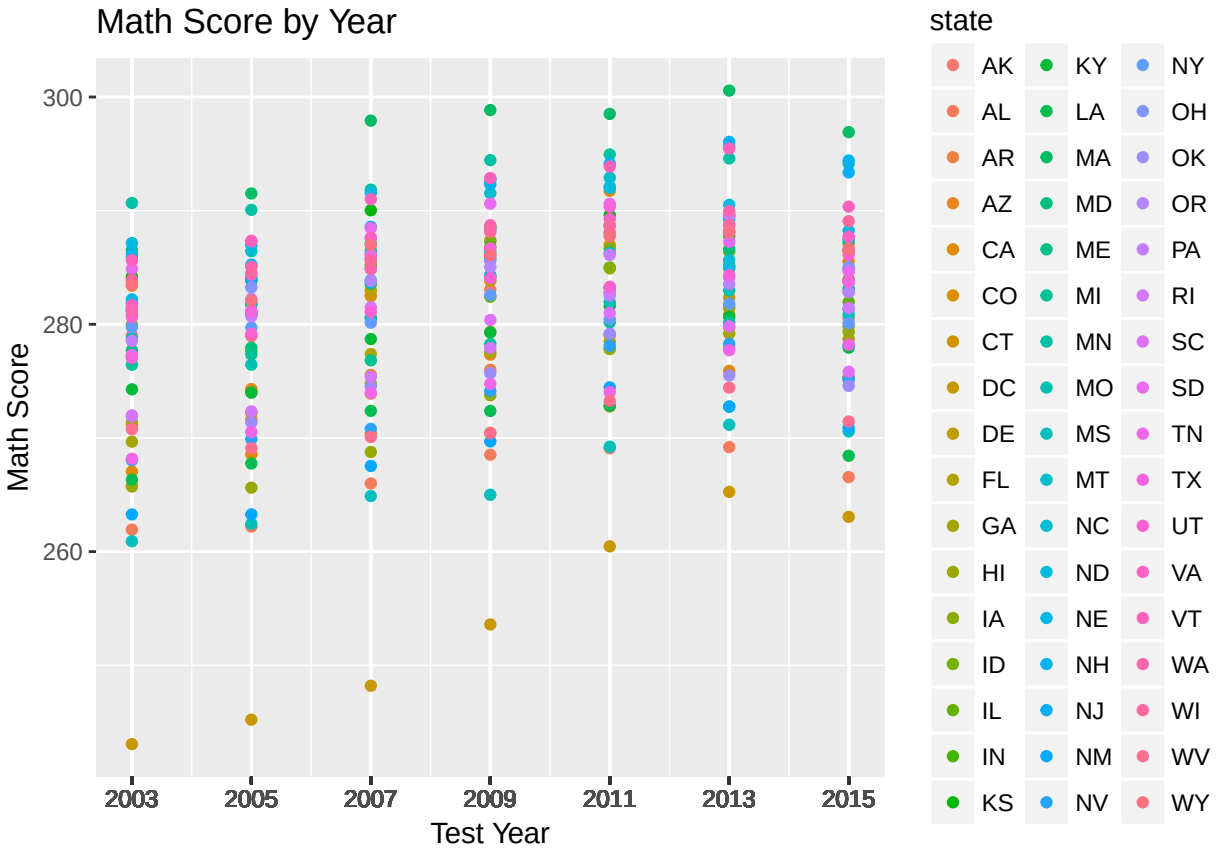
```
summary(df_naep_math$math_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   243.1   276.4   282.2   281.0   286.5   300.6
```

```
ggplot(df_naep_reading, aes(x = test_year, y = reading_score, colour = state)) + geom_point() +
  labs(x = "Test Year", y = "Reading Score", title = "Reading Score by Year") +
  scale_x_continuous(breaks = df_naep_reading$test_year)
```




```
ggplot(df_naep_math, aes(x = test_year, y = math_score, colour = state)) + geom_point() +
  labs(x = "Test Year", y = "Math Score", title = "Math Score by Year") +
  scale_x_continuous(breaks = df_naep_math$test_year)
```



The outlying values for both the math and reading scores are from DC:

```
(dbGetQuery(con, "SELECT distinct state from naep8 where math_score < 260;"))
```

```
##   state
## 1    DC
```

```
(dbGetQuery(con, "SELECT distinct state from naep8 where reading_score < 240;"))
```

```
##   state
## 1    DC
```

Revenue by Year and State

The revenue summaries show that there are some -2 values for state_revenue for DC which means "not applicable" according to the data dictionary.

```
df_funding <- dbGetQuery(con, "SELECT survey_year, state, state_revenue, local_revenue,
                         federal_revenue, total_revenue from fiscal order by survey_year, state;")
summary(df_funding$total_revenue)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 7.940e+08 2.671e+09 6.268e+09 1.070e+10 1.236e+10 7.122e+10
```

```
summary(df_funding$state_revenue)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -2.000e+00  1.393e+09  3.011e+09  4.964e+09  6.323e+09  4.366e+10
```

```
summary(df_funding$local_revenue)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 3.518e+07 1.103e+09 2.416e+09 4.677e+09 5.741e+09 3.226e+10
```

```
summary(df_funding$federal_revenue)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 6.891e+07 2.922e+08 6.321e+08 1.025e+09 1.154e+09 9.249e+09
```

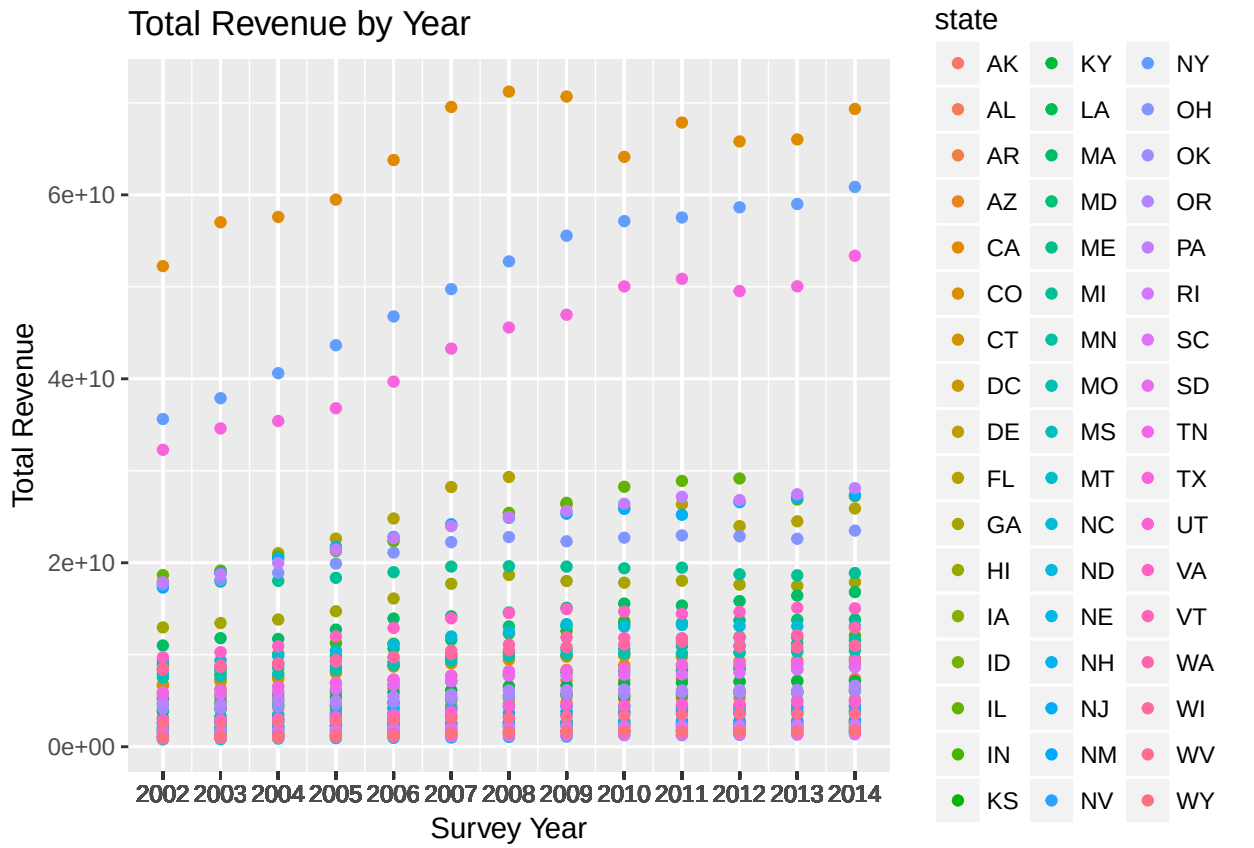
```
(dbGetQuery(con, "SELECT distinct state from fiscal where state_revenue = -2;"))
```

```
##   state
## 1    DC
```

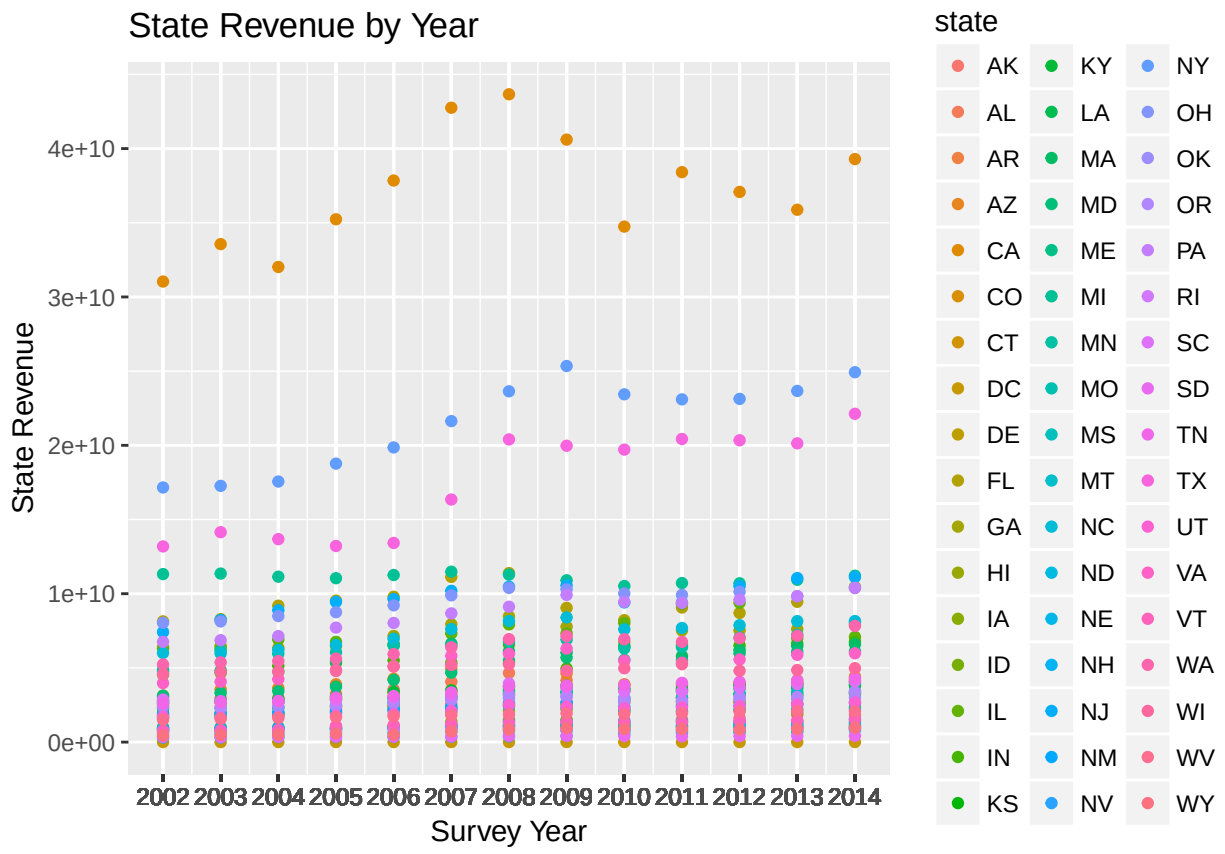
We can see that the states of CA, NY and TX have higher revenue from all sources than the other states.

It also looks like some states have seen a steady increase in local revenue over the past 10 years which suggests that this revenue source might be worth a closer look to see whether it has an impact on educational outcomes.

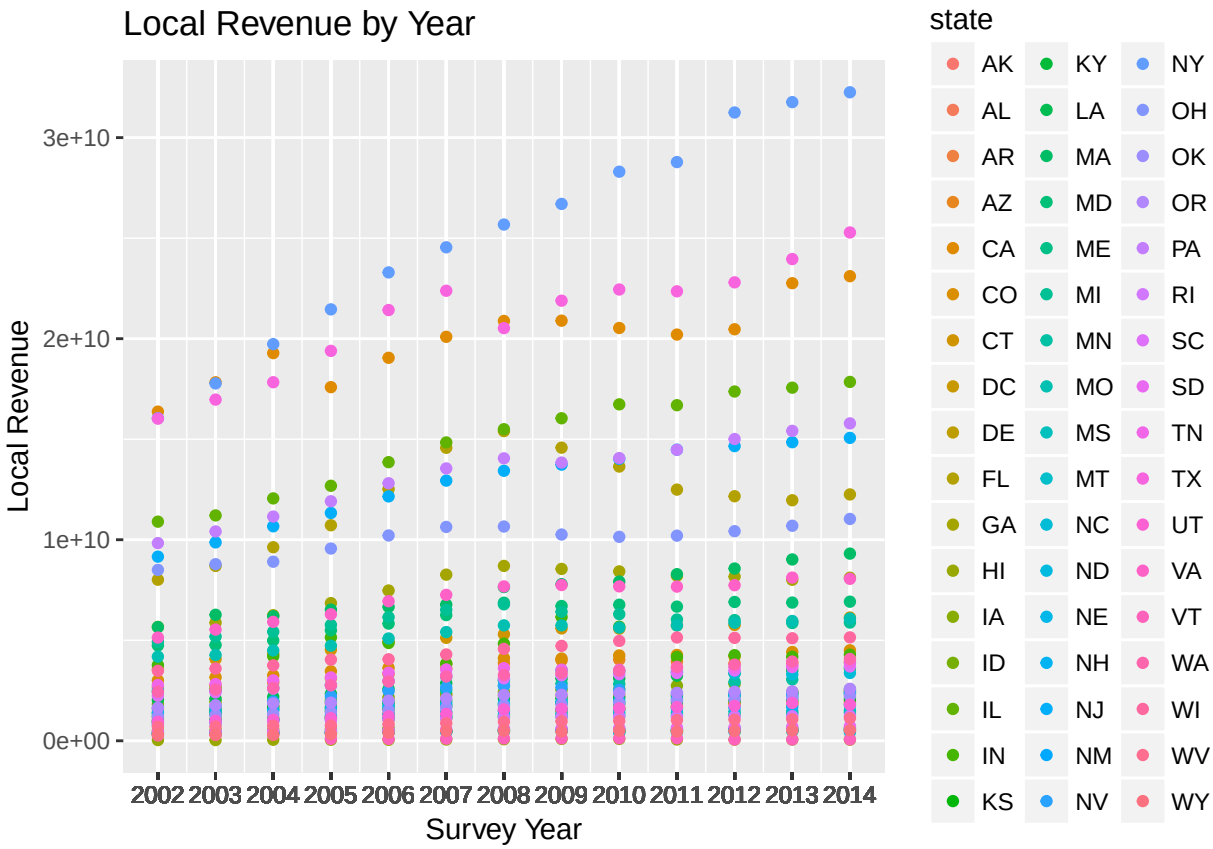
```
ggplot(df_funding, aes(x = survey_year, y = total_revenue, colour = state)) + geom_point() +
  labs(x = "Survey Year", y = "Total Revenue", title = "Total Revenue by Year") +
  scale_x_continuous(breaks = df_funding$survey_year)
```



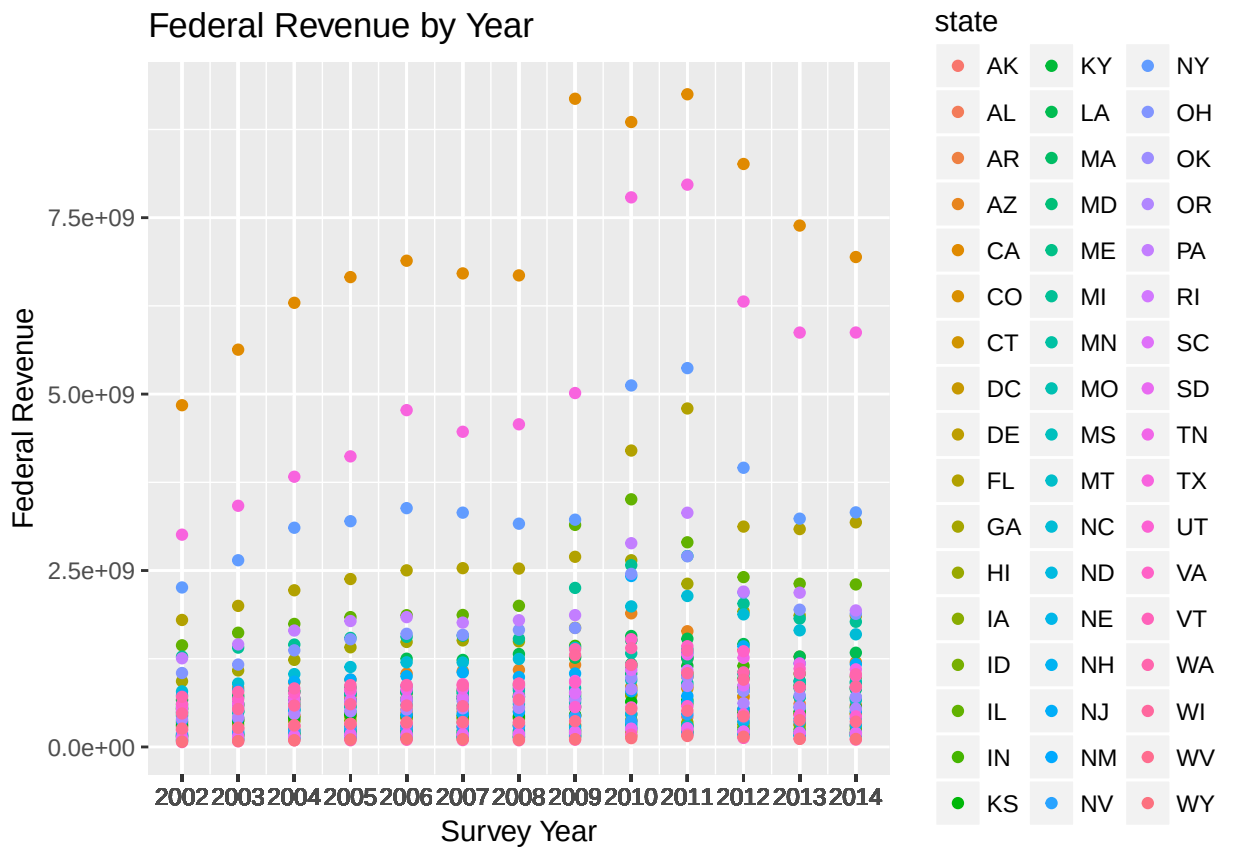

```
ggplot(df_funding, aes(x = survey_year, y = state_revenue, colour = state)) + geom_point() +
  labs(x = "Survey Year", y = "State Revenue", title = "State Revenue by Year") +
  scale_x_continuous(breaks = df_funding$survey_year)
```




```
ggplot(df_funding, aes(x = survey_year, y = local_revenue, colour = state)) + geom_point() +
  labs(x = "Survey Year", y = "Local Revenue", title = "Local Revenue by Year") +
  scale_x_continuous(breaks = df_funding$survey_year)
```




```
ggplot(df_funding, aes(x = survey_year, y = federal_revenue, colour = state)) + geom_point() +
  labs(x = "Survey Year", y = "Federal Revenue", title = "Federal Revenue by Year") +
  scale_x_continuous(breaks = df_funding$survey_year)
```




```
(dbGetQuery(con, "SELECT distinct state from fiscal where total_revenue > 50000000000;"))
```

```
##   state
## 1    CA
## 2    NY
## 3    TX
```

```
(dbGetQuery(con, "SELECT distinct state from fiscal where state_revenue > 20000000000;"))
```

```
##   state
## 1    CA
## 2    NY
## 3    TX
```

```
(dbGetQuery(con, "SELECT distinct state from fiscal where local_revenue > 20000000000;"))
```

```
##   state
## 1    CA
## 2    NY
## 3    TX
```

```
(dbGetQuery(con, "SELECT distinct state from fiscal where federal_revenue > 5000000000;"))
```

```
##   state
## 1    CA
## 2    NY
## 3    TX
```

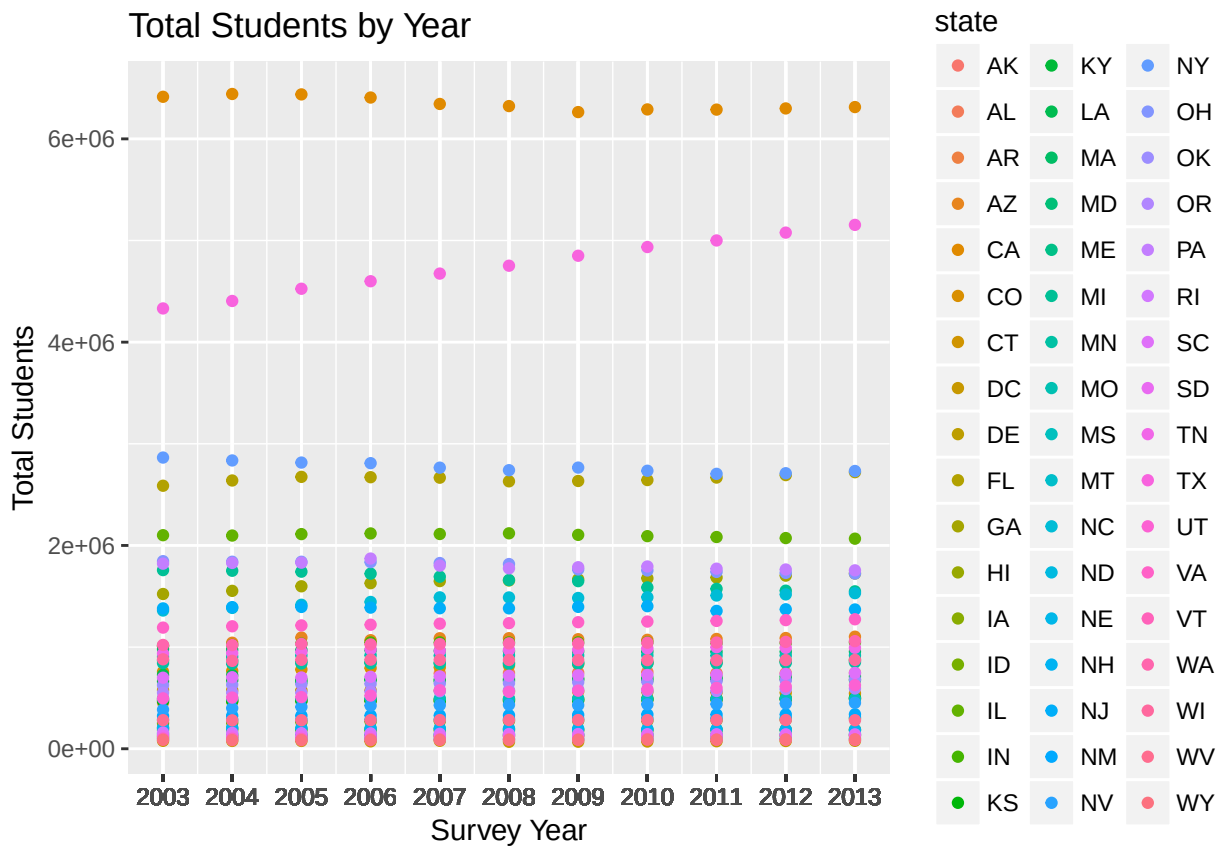
Looking at the number of students in each state, we see that the states with the highest revenue are also the

states with the largest volume of students. We will look at calculating revenue per student.

```
df_students <- dbGetQuery(con, "SELECT survey_year, state, total_students from nonfiscal
                          order by survey_year, state;")
summary(df_students$total_students)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68680  276300  673500  968700 1072000 6442000
```

```
ggplot(df_students, aes(x = survey_year, y = total_students, colour = state)) + geom_point() +
  labs(x = "Survey Year", y = "Total Students", title = "Total Students by Year") +
  scale_x_continuous(breaks = df_students$survey_year)
```



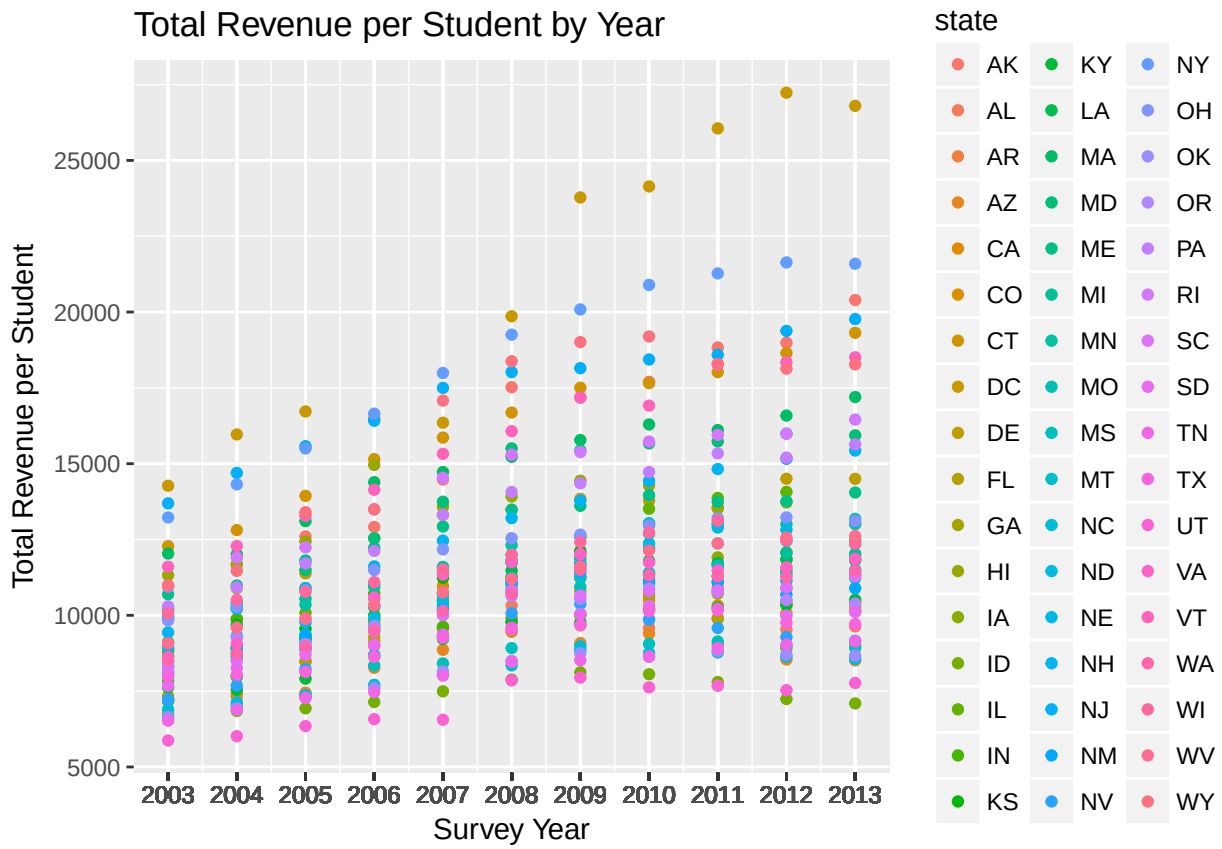

```
(dbGetQuery(con, "SELECT distinct state from nonfiscal where total_students > 2000000;"))
```

```
##   state
## 1    CA
## 2    FL
## 3    IL
## 4    NY
## 5    TX
```

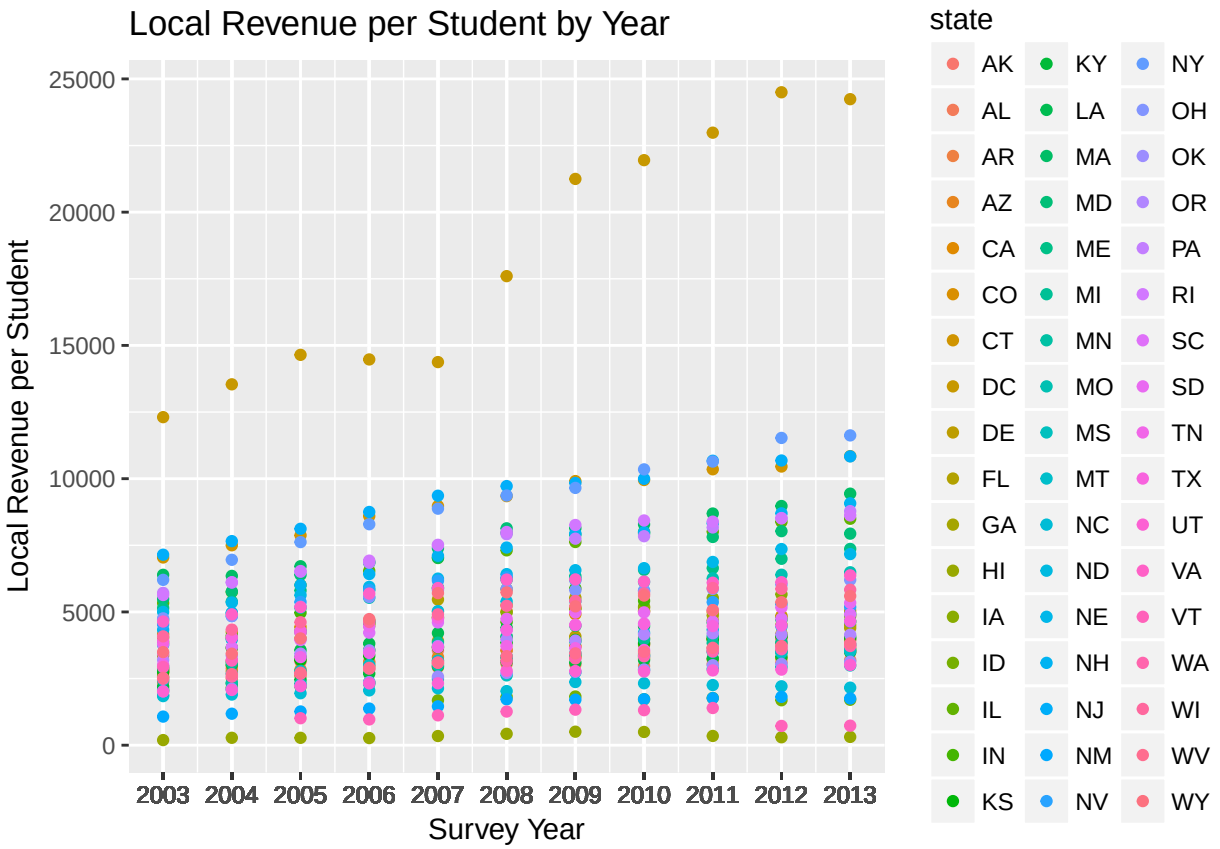
Revenue per Student

```
df_revenue_student <- dbGetQuery(con, "SELECT f.survey_year, f.state,
                                 (f.total_revenue/nf.total_students) as total_revenue_per_student,
                                 (f.local_revenue/nf.total_students) as local_revenue_per_student
```

```
                                from fiscal f, nonfiscal nf
                                where f.state = nf.state and f.survey_year = nf.survey_year
                                order by survey_year, state;")
ggplot(df_revenue_student, aes(x = survey_year, y = total_revenue_per_student, colour = state)) +
  labs(x = "Survey Year", y = "Total Revenue per Student", title = "Total Revenue per Student by Year") +
  geom_point() + scale_x_continuous(breaks = df_revenue_student$survey_year)
```




```
ggplot(df_revenue_student, aes(x = survey_year, y = local_revenue_per_student, colour = state)) +
  labs(x = "Survey Year", y = "Local Revenue per Student", title = "Local Revenue per Student by Year") +
  geom_point() + scale_x_continuous(breaks = df_revenue_student$survey_year)
```

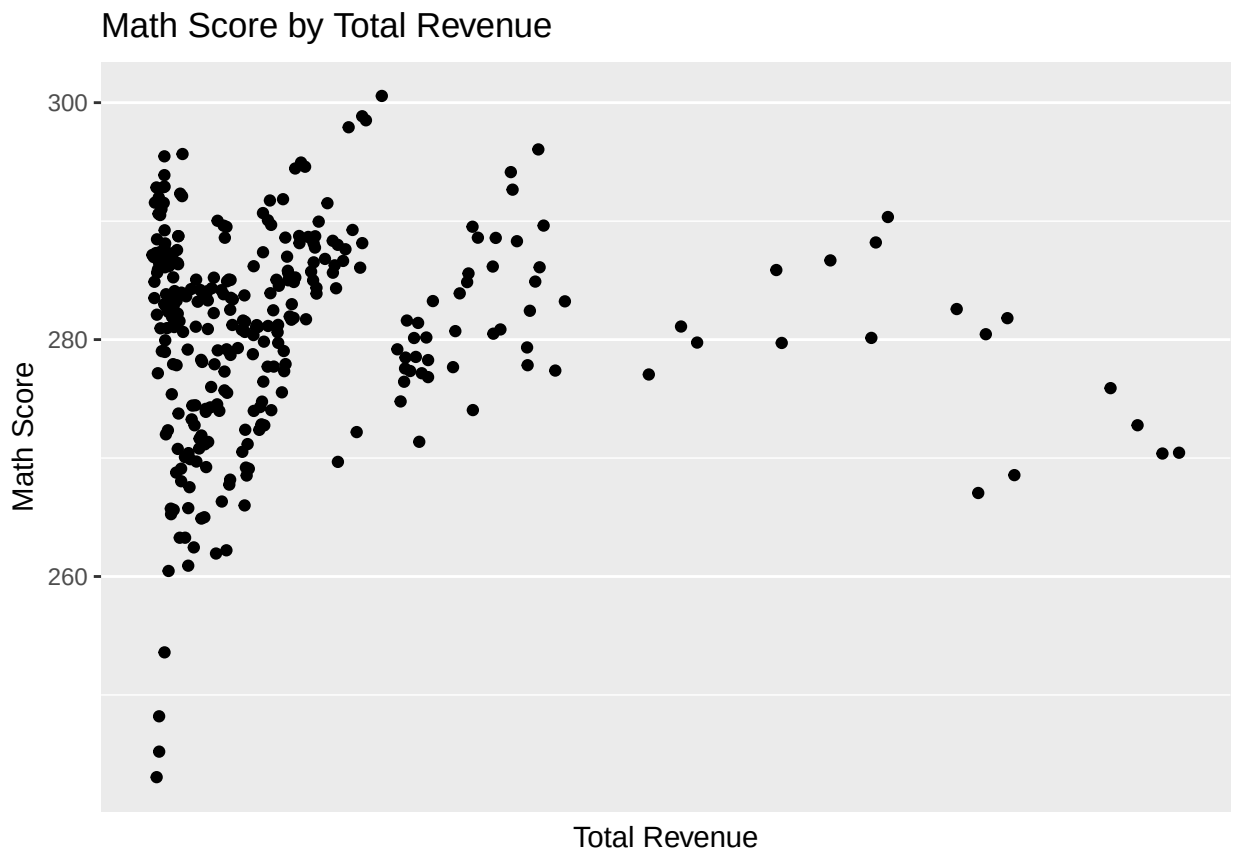


Score by Revenue

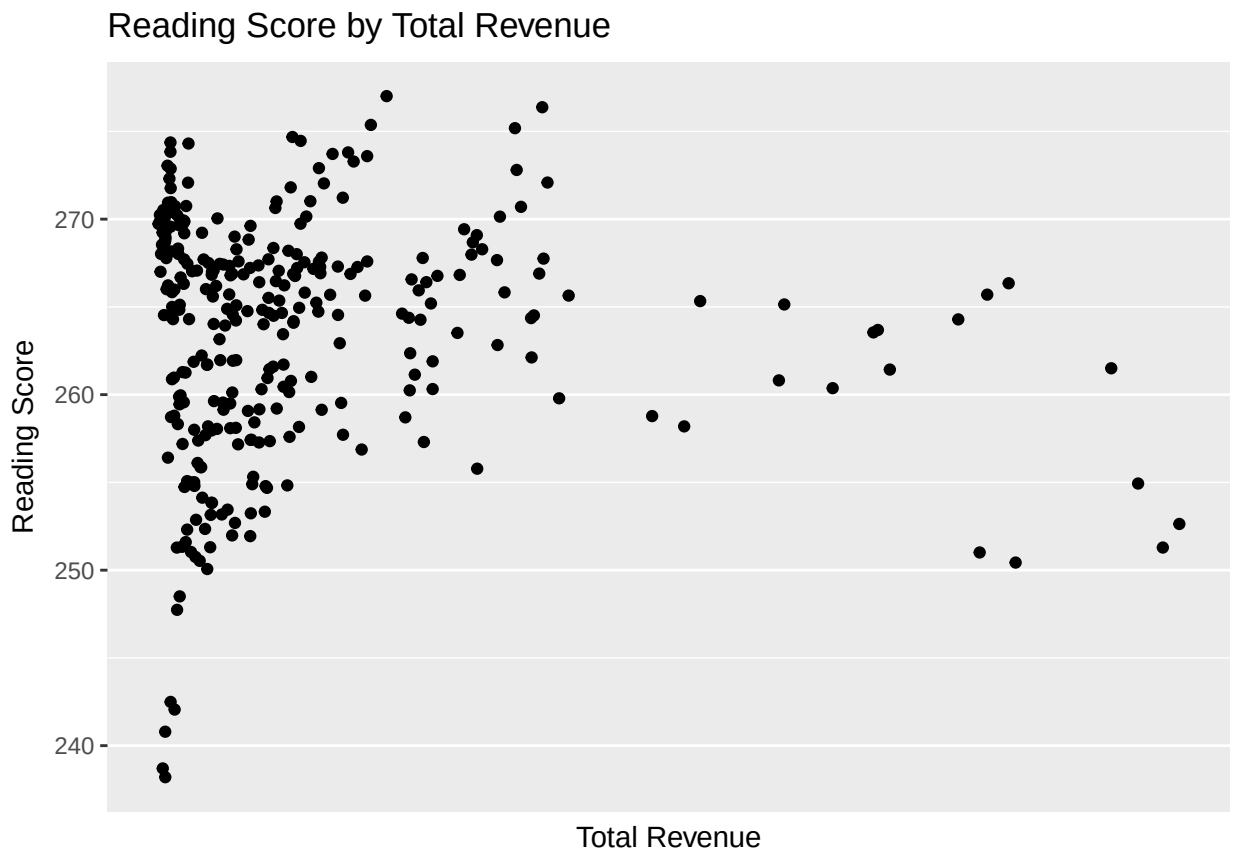
Now we will look at the values of revenue and academic scores together to determine whether there is a relationship.

Although the lowest scores in reading and math occur when funding is lowest, there does not appear to be a direct relationship between revenue and reading or math scores based on this data.

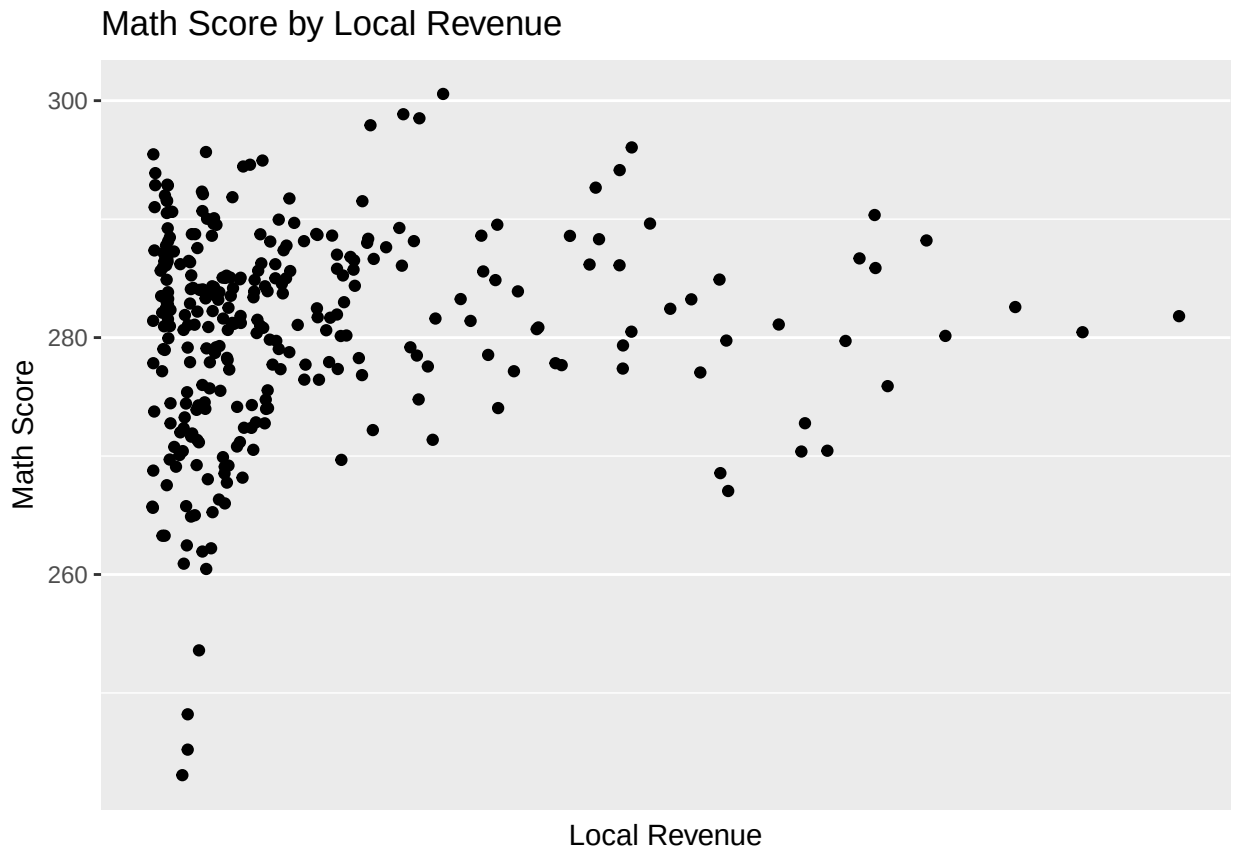
```
df_score_funding <- dbGetQuery(con, "SELECT f.total_revenue, local_revenue, s.math_score,
                               s.reading_score from fiscal f, naep8 s
                               where s.state=f.state and s.test_year=f.survey_year;")
ggplot(df_score_funding, aes(x = total_revenue, y = math_score)) + geom_point() +
  labs(x = "Total Revenue", y = "Math Score", title = "Math Score by Total Revenue") +
  scale_x_continuous(breaks = df_score_funding$survey_year)
```



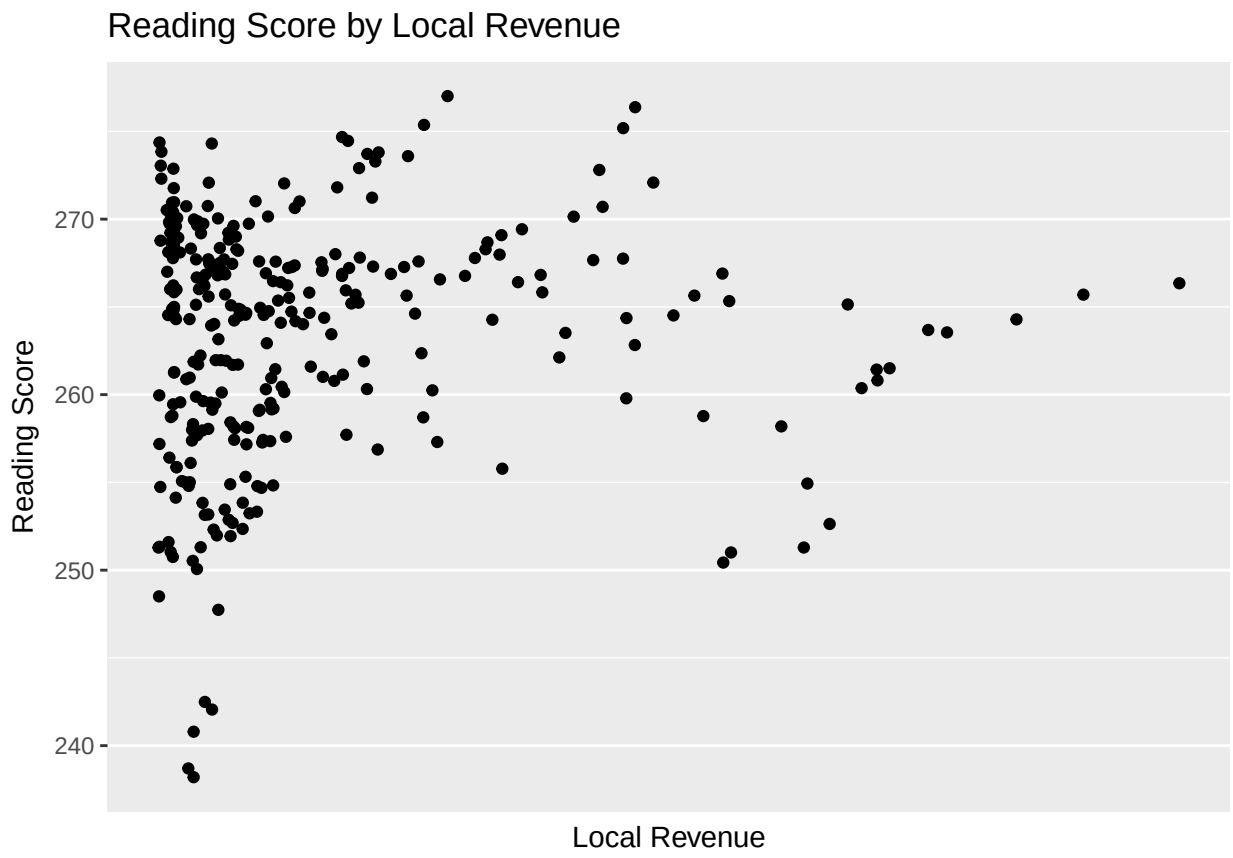

```
ggplot(df_score_funding, aes(x = total_revenue, y = reading_score)) + geom_point() +
  labs(x = "Total Revenue", y = "Reading Score", title = "Reading Score by Total Revenue") +
  scale_x_continuous(breaks = df_score_funding$survey_year)
```




```
ggplot(df_score_funding, aes(x = local_revenue, y = math_score)) + geom_point() +
  labs(x = "Local Revenue", y = "Math Score", title = "Math Score by Local Revenue") +
  scale_x_continuous(breaks = df_score_funding$survey_year)
```




```
ggplot(df_score_funding, aes(x = local_revenue, y = reading_score)) + geom_point() +
  labs(x = "Local Revenue", y = "Reading Score", title = "Reading Score by Local Revenue") +
  scale_x_continuous(breaks = df_score_funding$survey_year)
```

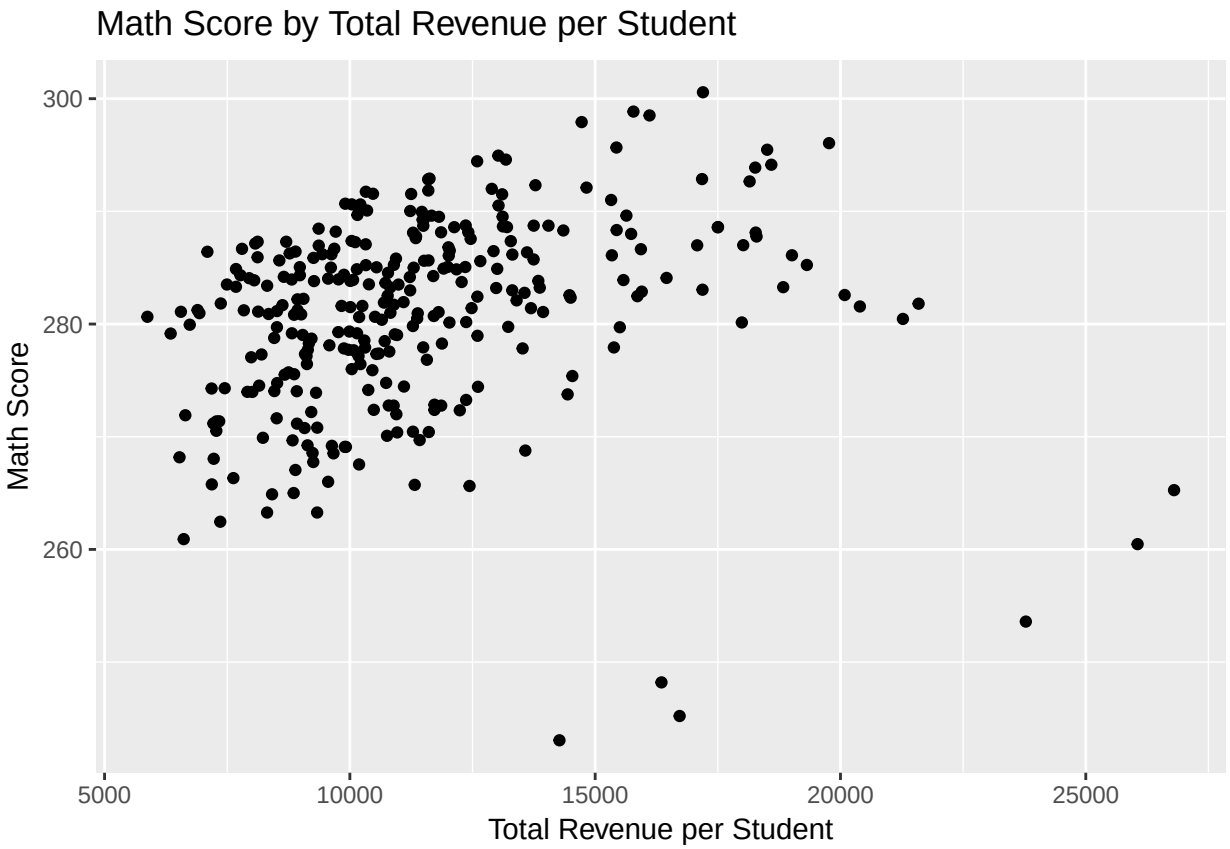


Score by Revenue per Student

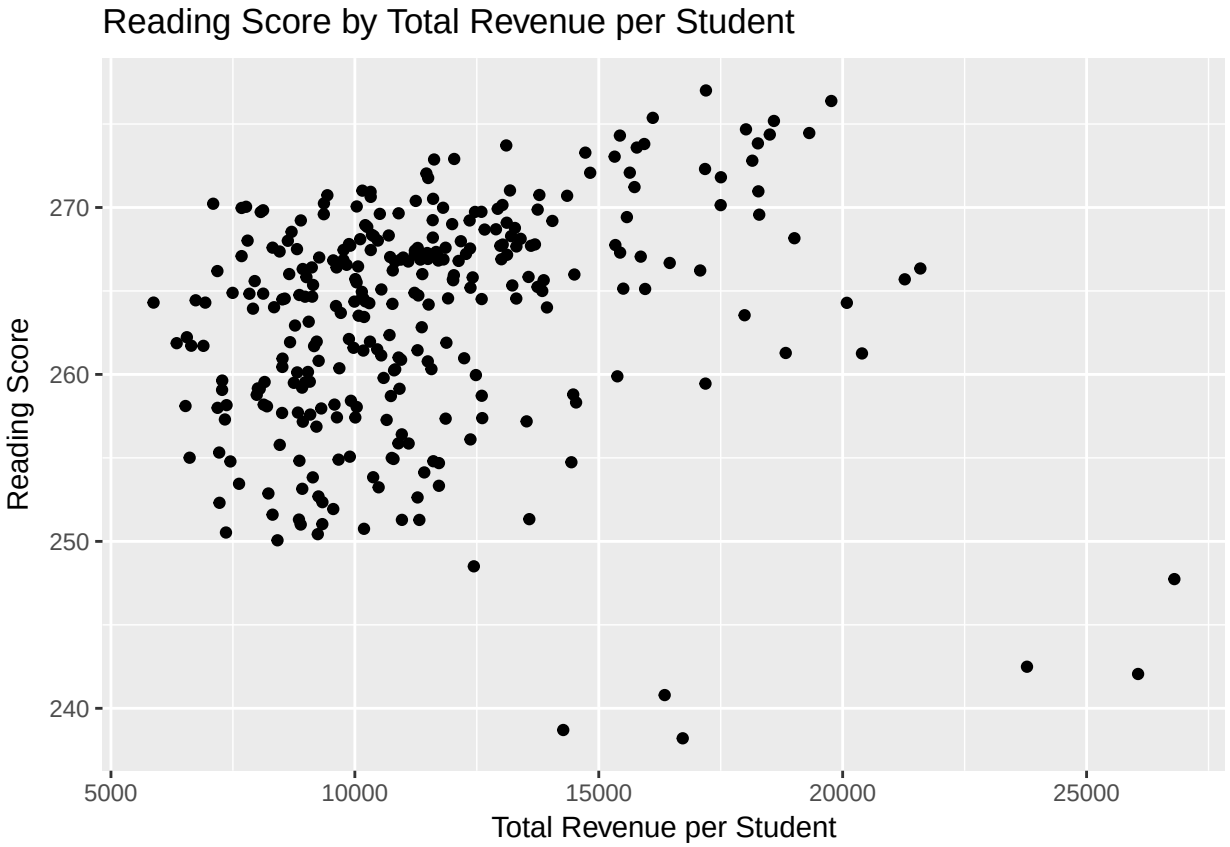
Let's check to see whether the answer is any different when we calculate funding per student.

When we look at total revenue per student we do appear to see a

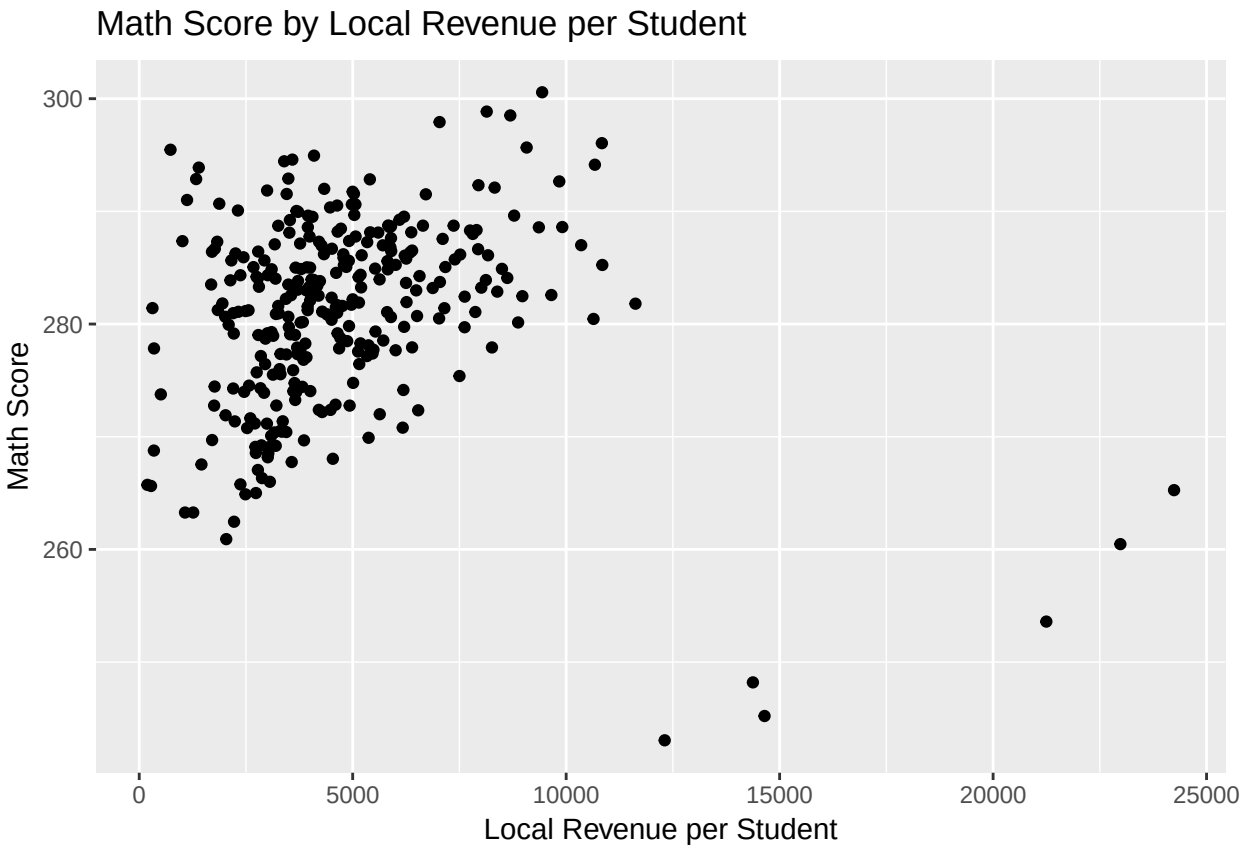
```
df_score_funding_student <- dbGetQuery(con, "SELECT (f.total_revenue/nf.total_students) as total_revenue
                                        (f.local_revenue/nf.total_students) as local_revenue_per_student
                                        s.math_score, s.reading_score
                                        from fiscal f, nonfiscal nf, naep8 s
                                        where s.state=f.state and f.state = nf.state and
                                        s.test_year=f.survey_year and f.survey_year = nf.survey_year;")
ggplot(df_score_funding_student, aes(x = total_revenue_per_student, y = math_score)) + geom_point() +
  labs(x = "Total Revenue per Student", y = "Math Score", title = "Math Score by Total Revenue per Stude
```



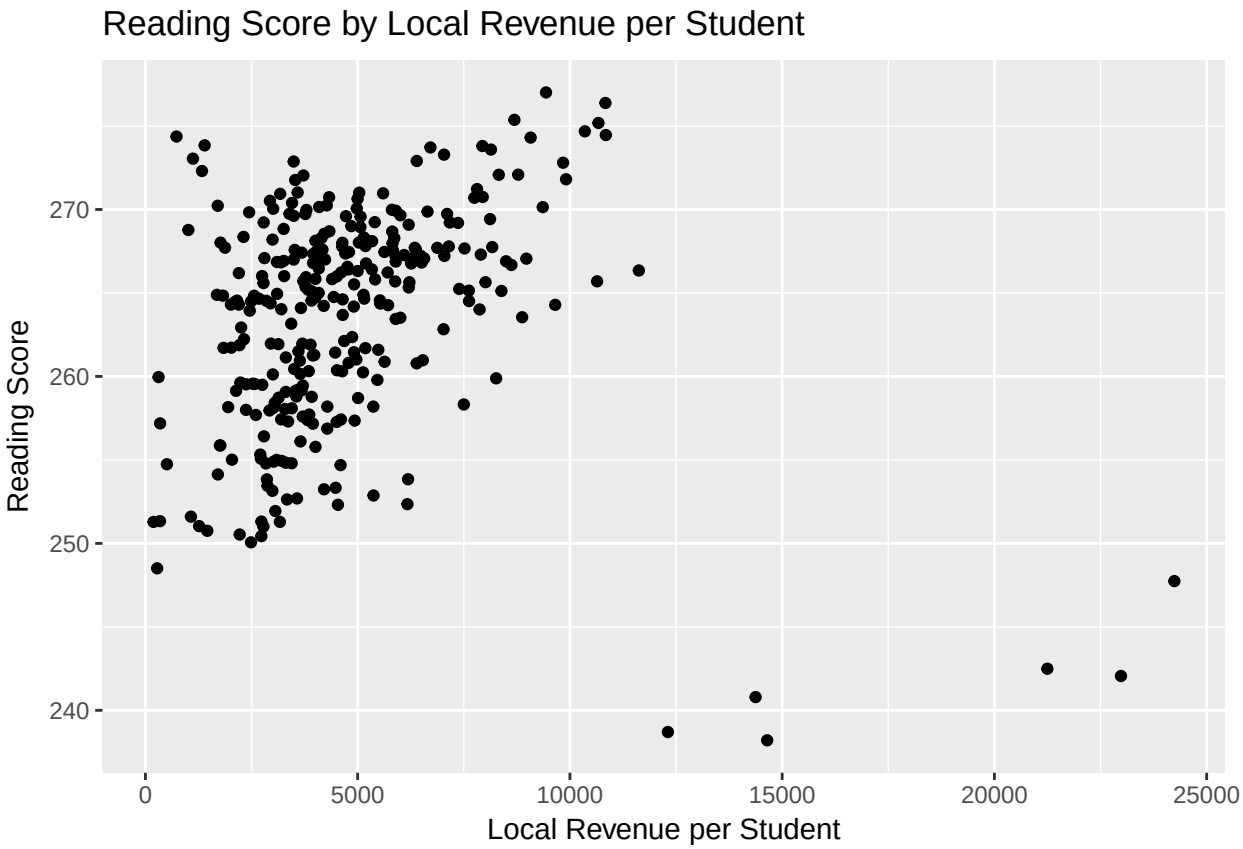

```
ggplot(df_score_funding_student, aes(x = total_revenue_per_student, y = reading_score)) + geom_point() +
  labs(x = "Total Revenue per Student", y = "Reading Score", title = "Reading Score by Total Revenue per
```




```
ggplot(df_score_funding_student, aes(x = local_revenue_per_student, y = math_score)) + geom_point() +
  labs(x = "Local Revenue per Student", y = "Math Score", title = "Math Score by Local Revenue per Stud
```




```
ggplot(df_score_funding_student, aes(x = local_revenue_per_student, y = reading_score)) + geom_point() +
  labs(x = "Local Revenue per Student", y = "Reading Score", title = "Reading Score by Local Revenue per
```



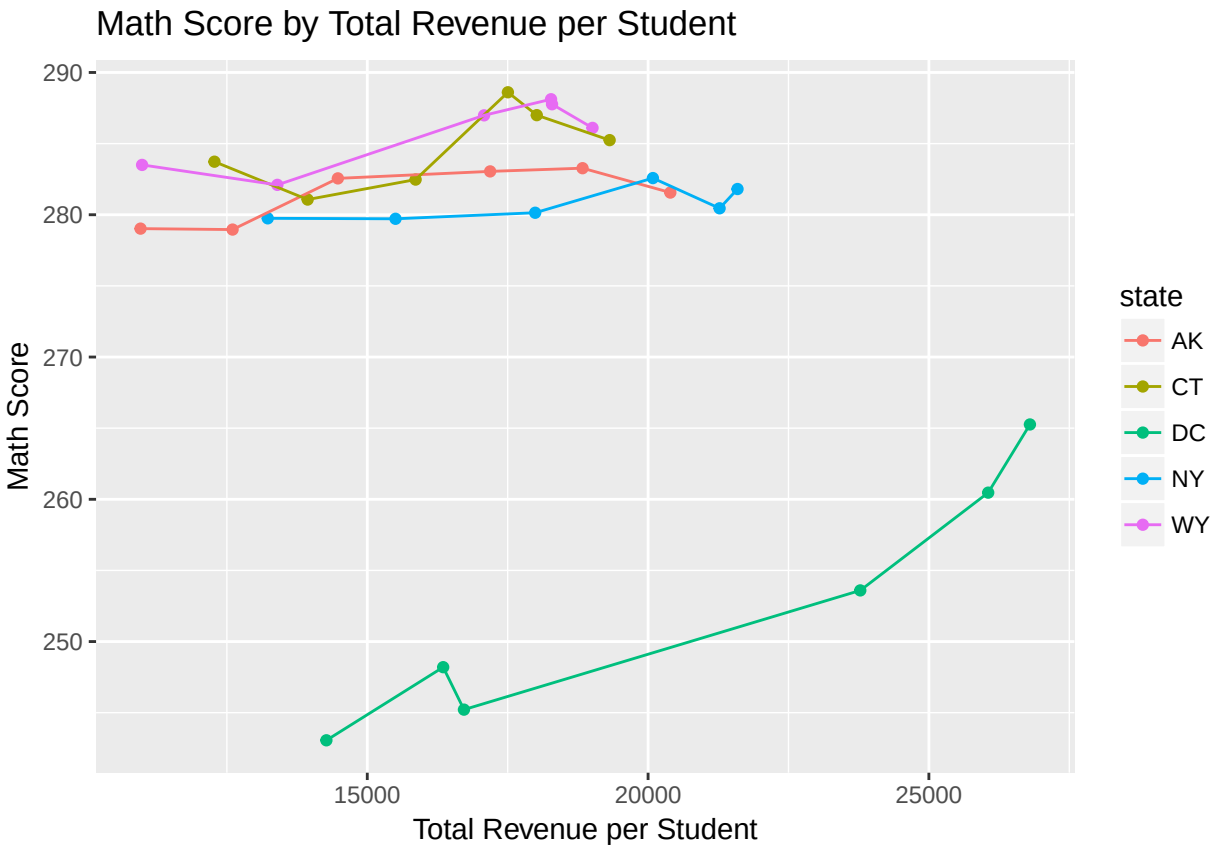
States with the Largest Change in Revenue per Student

Let's have a look at what happens when revenue changes. We will focus on the states which had the largest increase in revenue and look for any changes in educational outcome.

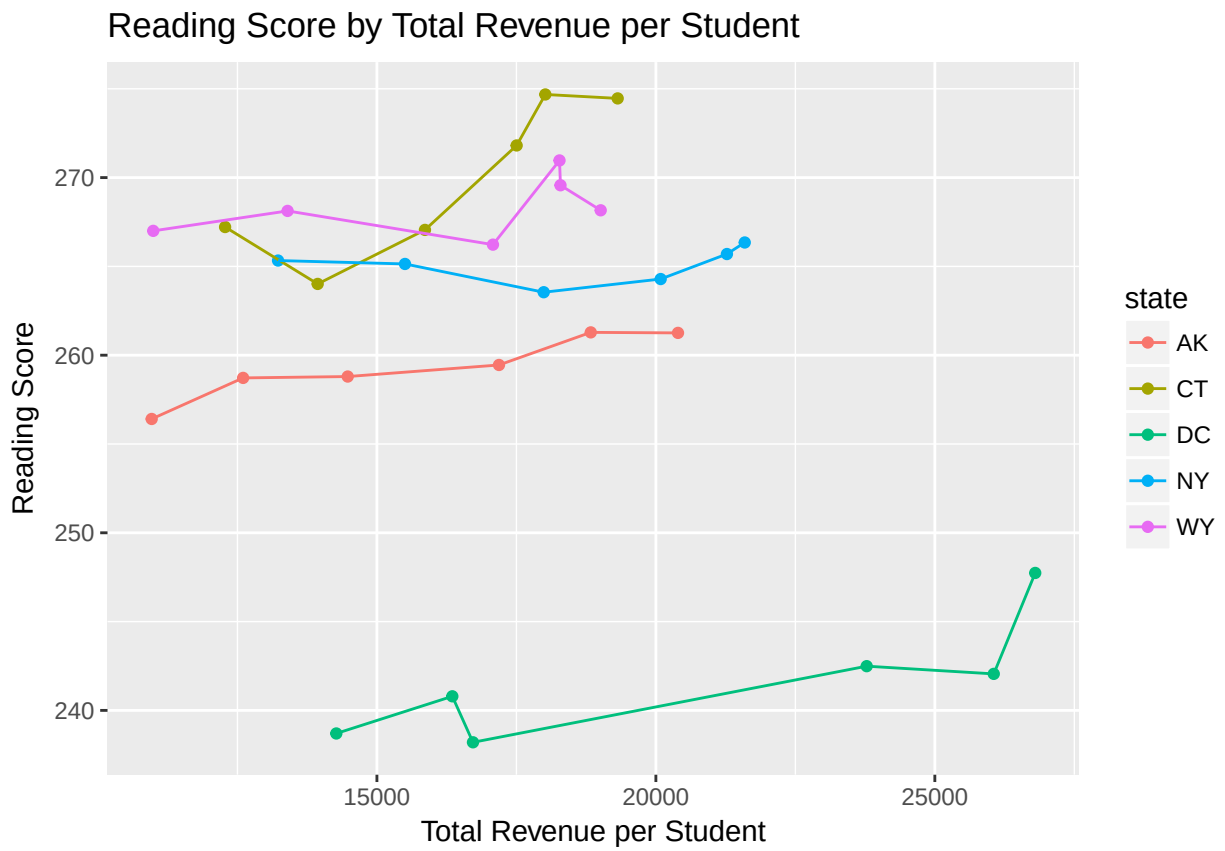
```
df_total_revenue_range <- dbGetQuery(con, "SELECT
        (max(f.total_revenue/nf.total_students)-min(f.total_revenue/nf.total_students)) as total_revenue
        f.state from fiscal f, nonfiscal nf where nf.state=f.state and nf.survey_year=f.survey_year
        group by f.state order by total_revenue_per_student_range desc limit 5;")
(df_total_revenue_range$state)
```

```
## [1] "DC" "AK" "NY" "WY" "CT"
```

```
qry <- stri_paste("SELECT (f.total_revenue/nf.total_students) as total_revenue_per_student,
                  s.math_score, s.reading_score, s.state from fiscal f, nonfiscal nf, naep8 s
                  where s.state=f.state and f.state=nf.state and s.test_year=f.survey_year
                  and f.survey_year = nf.survey_year and s.state in
                  ('", df_total_revenue_range$state[1], "','", df_total_revenue_range$state[2],"','",
                  df_total_revenue_range$state[3],"','", df_total_revenue_range$state[4],"','",
                  df_total_revenue_range$state[5],"')",collapse="")
df_total_revenue_range_score <- dbGetQuery(con, qry)
ggplot(df_total_revenue_range_score, aes(x = total_revenue_per_student, y = math_score, colour = state))
  geom_line() + labs(x = "Total Revenue per Student", y = "Math Score", title = "Math Score by Total Re
```



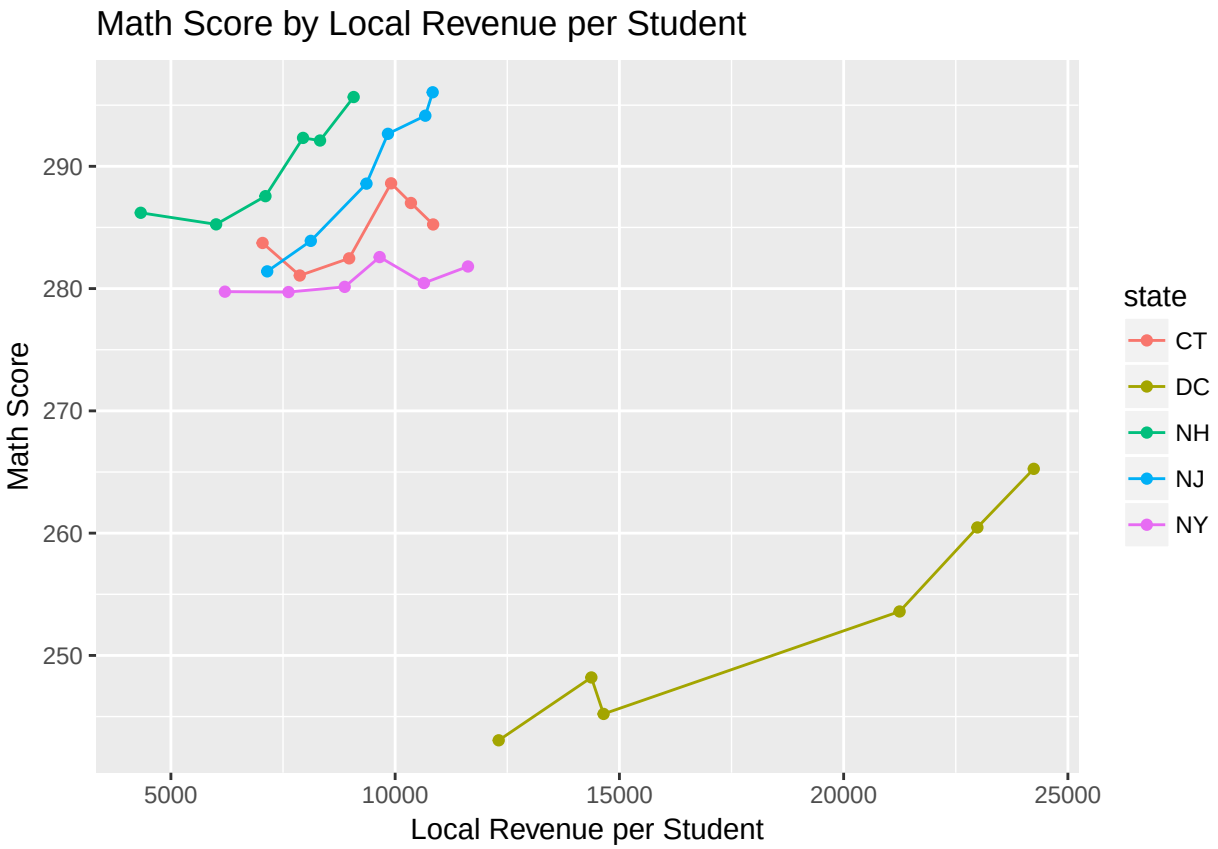

```
ggplot(df_total_revenue_range_score, aes(x = total_revenue_per_student, y = reading_score, colour = sta
  geom_line() + labs(x = "Total Revenue per Student", y = "Reading Score", title = "Reading Score by To
```



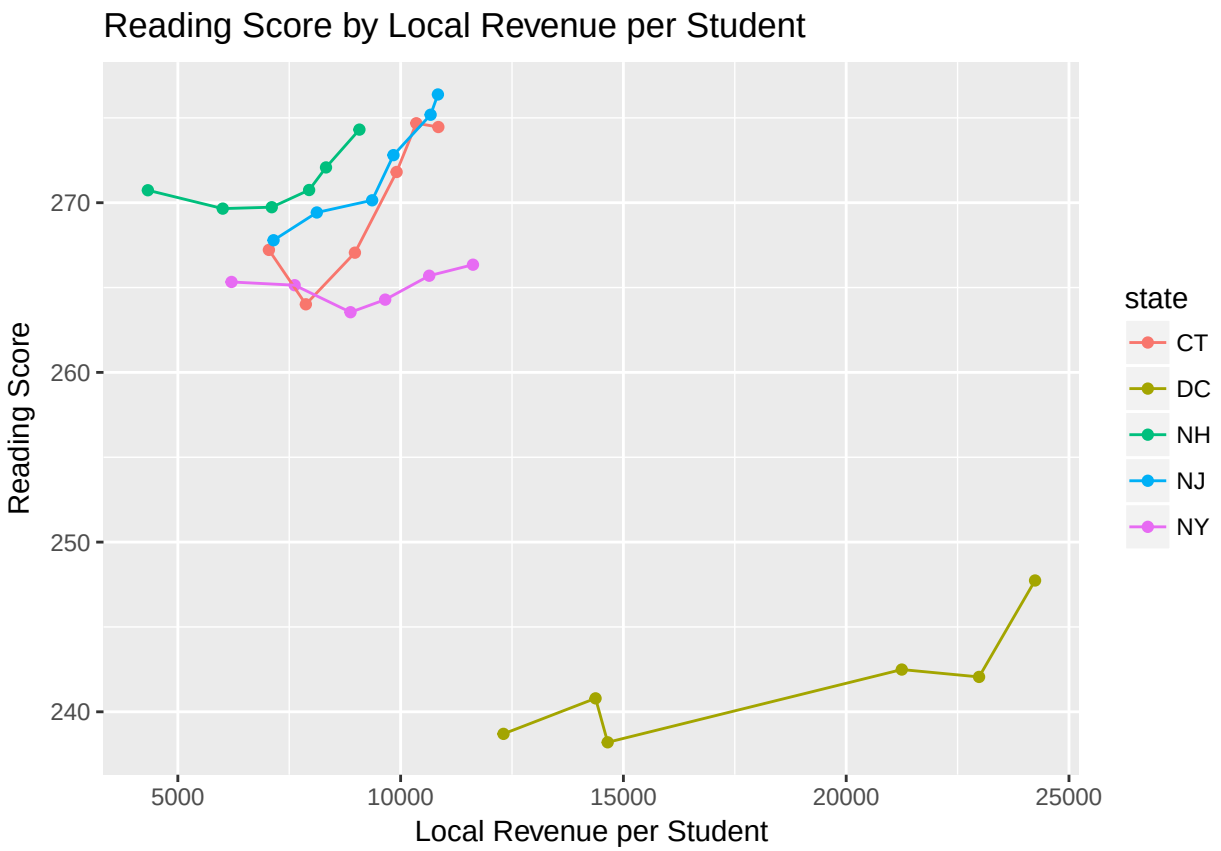

```
df_local_revenue_range <- dbGetQuery(con, "SELECT
          (max(f.local_revenue/nf.total_students)-min(f.local_revenue/nf.total_students)) as local_reve
          f.state from fiscal f, nonfiscal nf where nf.state=f.state and nf.survey_year=f.survey_year
          group by f.state order by local_revenue_per_student_range desc limit 5;")
(df_local_revenue_range$state)
```

```
## [1] "DC" "NY" "NH" "CT" "NJ"
```

```
qry <- stri_paste("SELECT (f.local_revenue/nf.total_students) as local_revenue_per_student, s.math_scor
                  from fiscal f, nonfiscal nf, naep8 s
                  where s.state=f.state and f.state=nf.state and s.test_year=f.survey_year and
                  f.survey_year = nf.survey_year and s.state in ('", df_local_revenue_range$state[1], "
                  df_local_revenue_range$state[2],"','", df_local_revenue_range$state[3],"','",
                  df_local_revenue_range$state[4],"','", df_local_revenue_range$state[5],"')",collapse=
df_local_revenue_range_score <- dbGetQuery(con, qry)
ggplot(df_local_revenue_range_score, aes(x = local_revenue_per_student, y = math_score, colour = state))
  geom_line() + labs(x = "Local Revenue per Student", y = "Math Score", title = "Math Score by Local Re
```




```
ggplot(df_local_revenue_range_score, aes(x = local_revenue_per_student, y = reading_score, colour = sta
  geom_line() + labs(x = "Local Revenue per Student", y = "Reading Score", title = "Reading Score by Lo
```



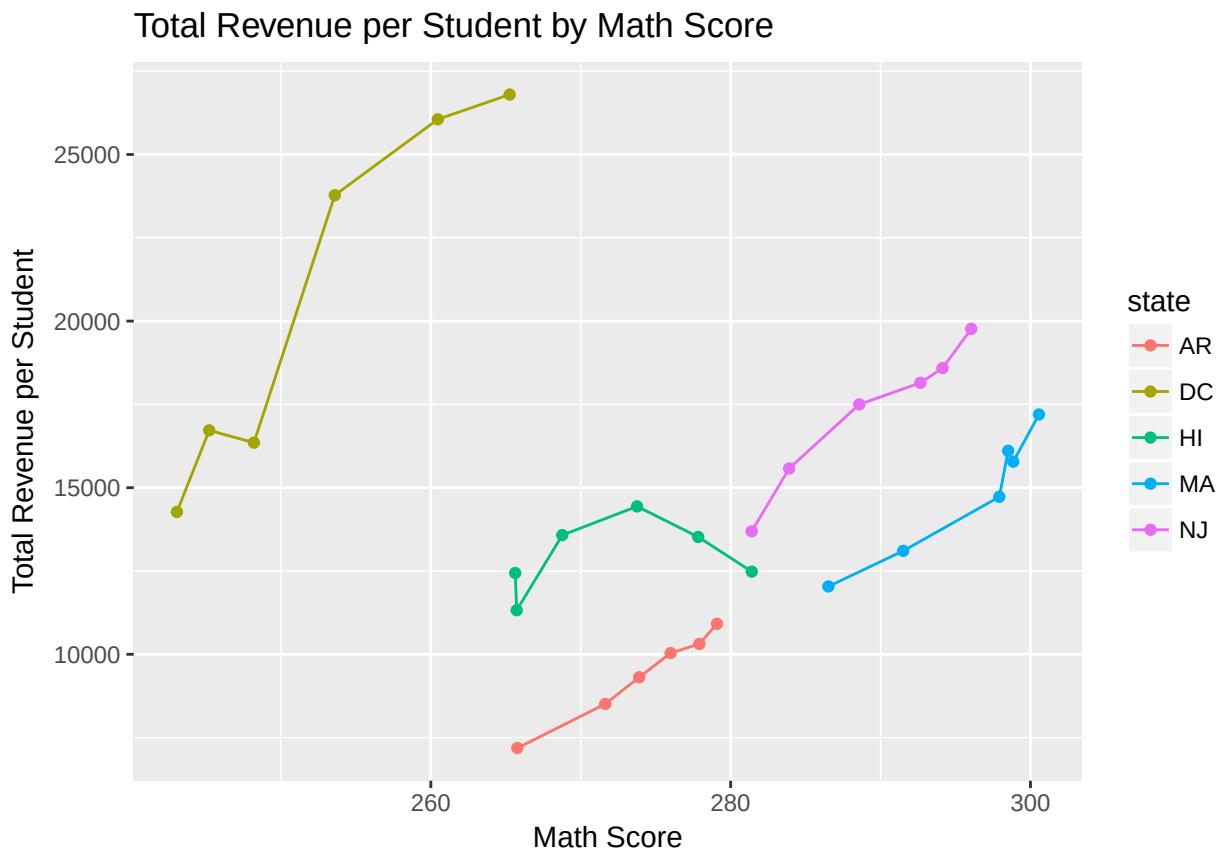
States with the Largest Change in Math Score

Conversely, lets have a look at the states with the largest change in math scores. Can we see a corresponding increase in revenue?

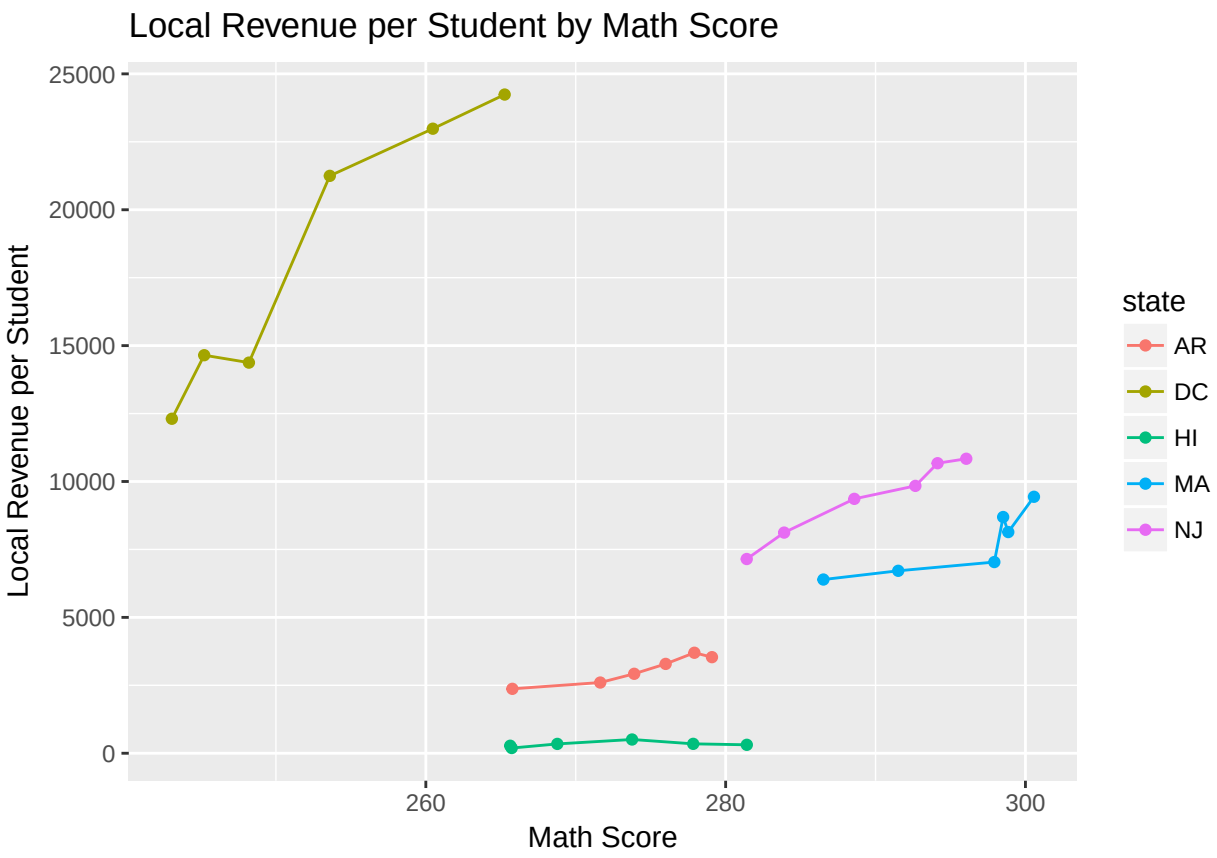
```
df_math_score_range <- dbGetQuery(con, "SELECT (max(math_score)-min(math_score)) as math_score_range, s
                                  from naep8 group by state order by math_score_range desc limit 5;")
(df_math_score_range$state)
```

```
## [1] "DC" "HI" "NJ" "MA" "AR"
```

```
qry <- stri_paste("SELECT (f.total_revenue/nf.total_students) as total_revenue_per_student, s.math_scor
                  from fiscal f, nonfiscal nf, naep8 s
                  where s.state=f.state and f.state = nf.state and s.test_year=f.survey_year and
                  f.survey_year = nf.survey_year and s.state in ('", df_math_score_range$state[1], "','
                  df_math_score_range$state[2],"','", df_math_score_range$state[3],"','",
                  df_math_score_range$state[4],"','", df_math_score_range$state[5],"')",collapse="")
df_math_score_range_revenue <- dbGetQuery(con, qry)
ggplot(df_math_score_range_revenue, aes(x = math_score, y = total_revenue_per_student, colour = state))
  geom_line() + labs(x = "Math Score", y = "Total Revenue per Student", title = "Total Revenue per Stud
```



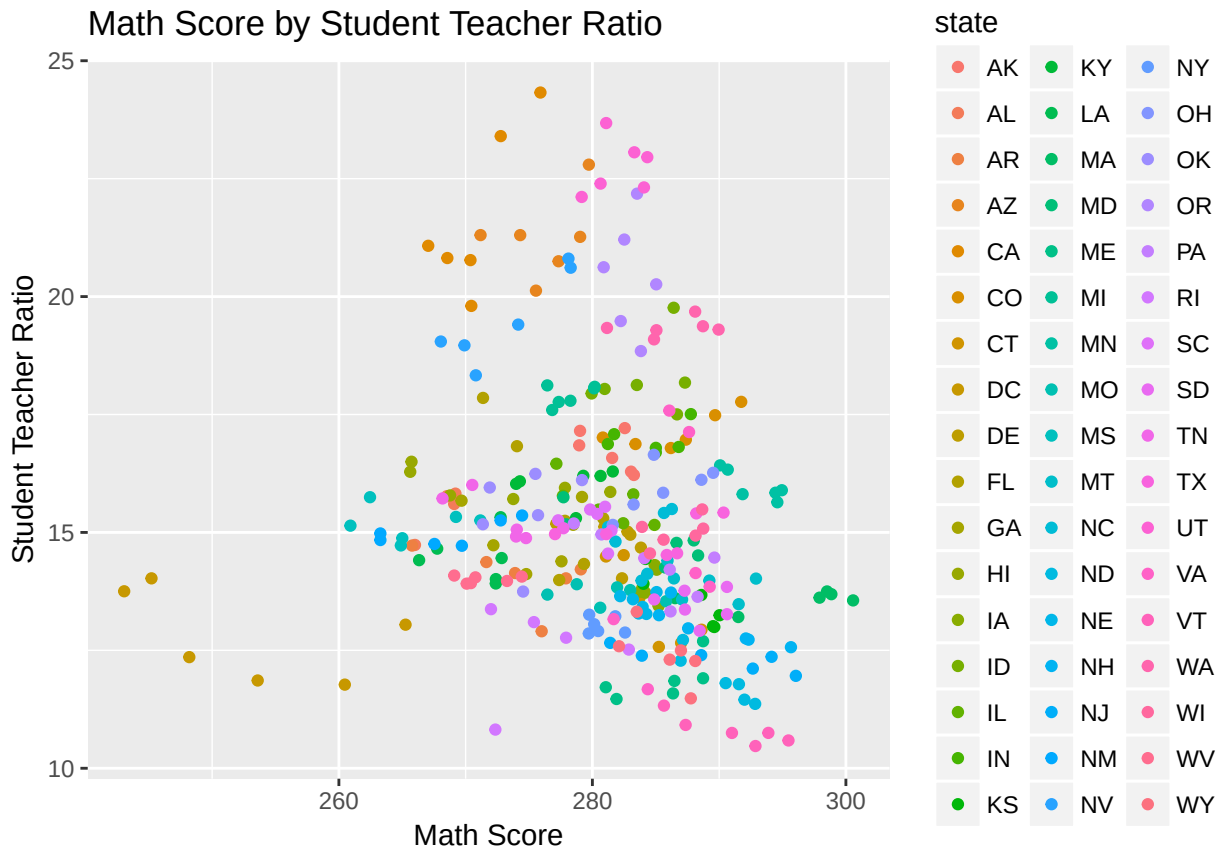

```
qry <- stri_paste("SELECT (f.local_revenue/nf.total_students) as local_revenue_per_student, s.math_score
                  from fiscal f, nonfiscal nf, naep8 s
                  where s.state=f.state and f.state = nf.state and s.test_year=f.survey_year and
                  f.survey_year = nf.survey_year and s.state in ('", df_math_score_range$state[1], "','"
                  df_math_score_range$state[2],"','", df_math_score_range$state[3],"','",
                  df_math_score_range$state[4],"','", df_math_score_range$state[5],"')",collapse="")
df_math_score_range_revenue <- dbGetQuery(con, qry)
ggplot(df_math_score_range_revenue, aes(x = math_score, y = local_revenue_per_student, colour = state))
  geom_line() + labs(x = "Math Score", y = "Local Revenue per Student", title = "Local Revenue per Stud
```



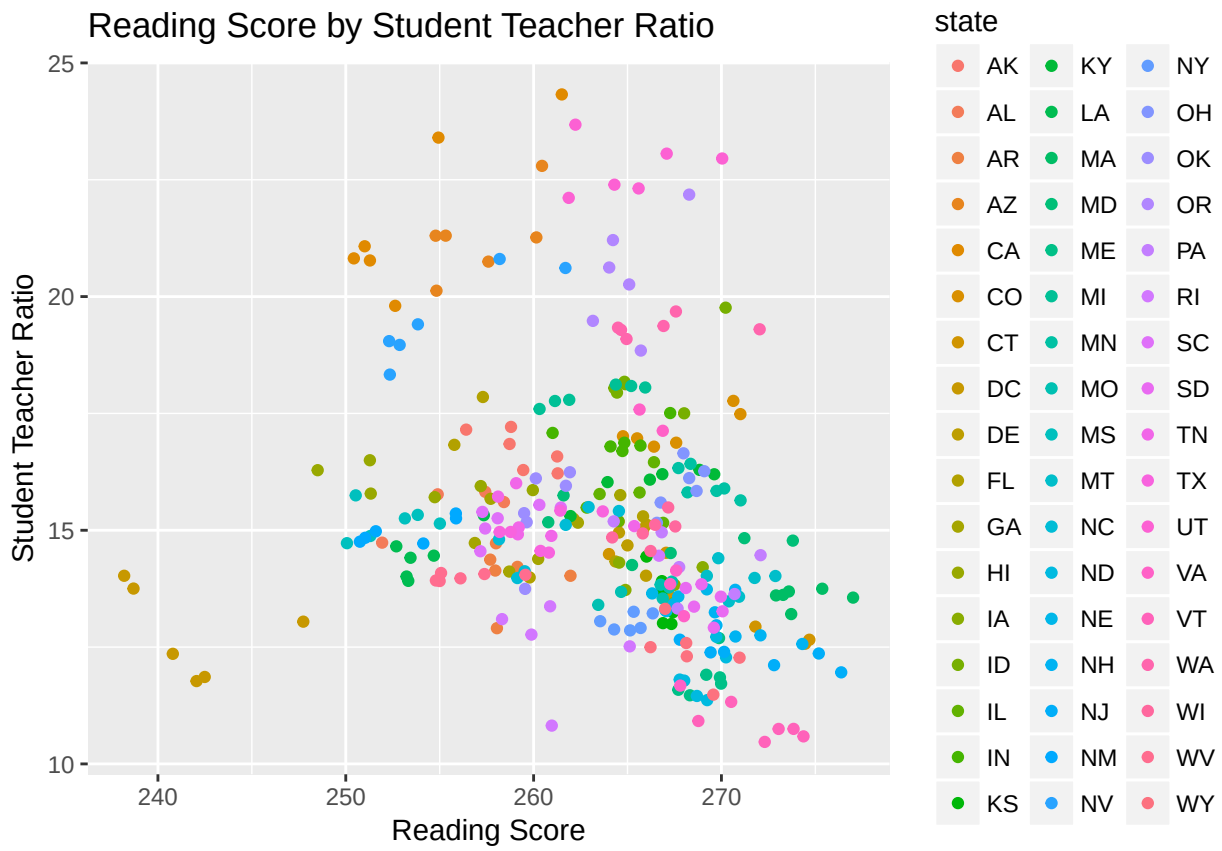
Student Teacher Ratios and Scores

Let's look at where the funding is being spent, and see whether the number of teachers makes a difference by calculating the student teacher ratio and comparing to test scores. If the lower ratio of students to teachers improves educational outcome then we would expect to see a negative slope on the graph. There may be a slight indication of this relationship in these graphs.

```
student_teacher <- dbGetQuery(con, "SELECT (nf.total_students/nf.total_teachers) as student_teacher_rati
                              s.math_score, s.reading_score, s.state from nonfiscal nf, naep8 s
                              where nf.state=s.state and nf.survey_year = s.test_year;")
ggplot(student_teacher, aes(x = math_score, y = student_teacher_ratio, colour = state)) + geom_point() +
  labs(x = "Math Score", y = "Student Teacher Ratio", title = "Math Score by Student Teacher Ratio")
```



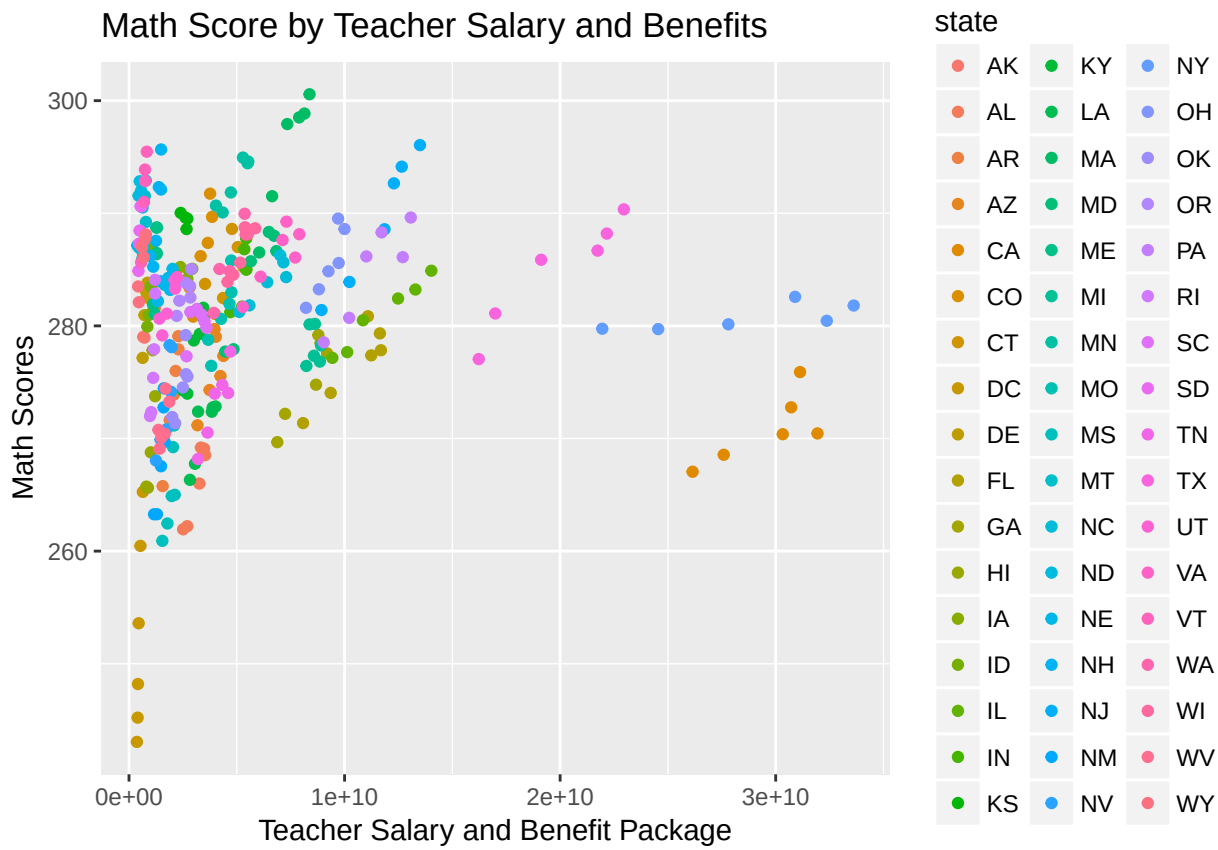

```
ggplot(student_teacher, aes(x = reading_score, y = student_teacher_ratio, colour = state)) + geom_point
  labs(x = "Reading Score", y = "Student Teacher Ratio", title = "Reading Score by Student Teacher Ratio
```



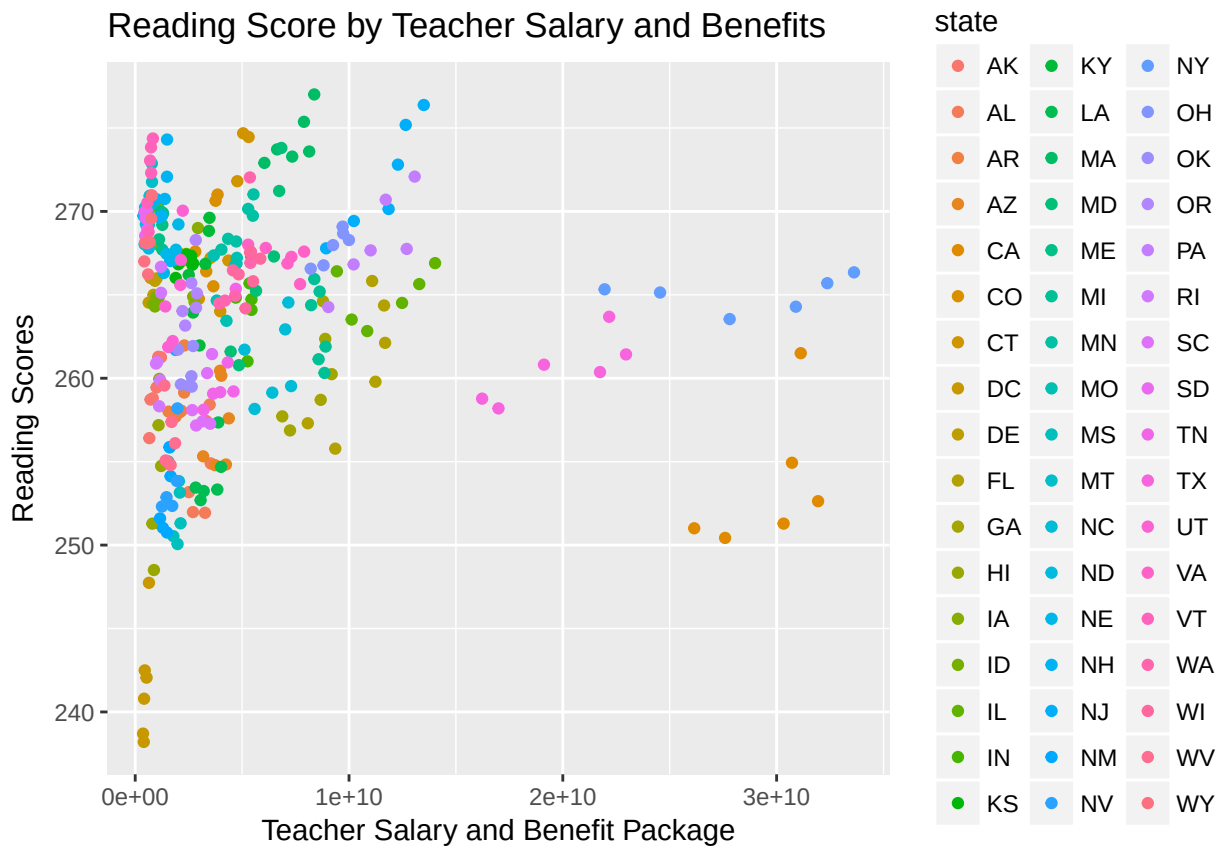
Teacher Salaries/Benefits and Scores

Let's further examine where the funding is being spent and look at teacher salaries and benefits and their relationship to educational outcomes.

```
teacher_salary <- dbGetQuery(con, "SELECT (f.teacher_salaries + f.teacher_benefits) as teacher_pay,
                             s.math_score, s.reading_score, s.state from fiscal f, naep8 s
                             where f.state=s.state and f.survey_year = s.test_year;")
ggplot(teacher_salary, aes(x = teacher_pay, y = math_score, colour = state)) + geom_point() +
  labs(x = "Teacher Salary and Benefit Package", y = "Math Scores", title = "Math Score by Teacher Sala
```

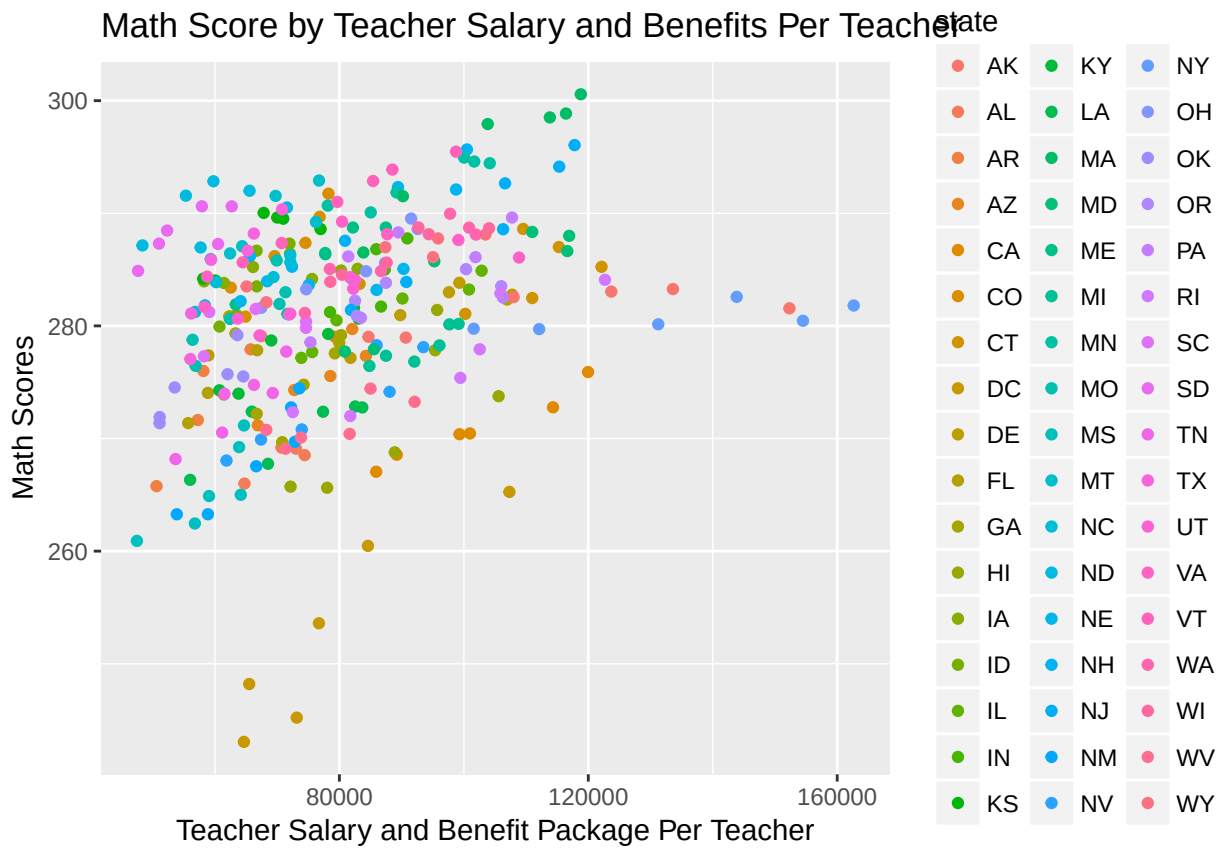



```
ggplot(teacher_salary, aes(x = teacher_pay, y = reading_score, colour = state)) + geom_point() +
  labs(x = "Teacher Salary and Benefit Package", y = "Reading Scores", title = "Reading Score by Teache
```

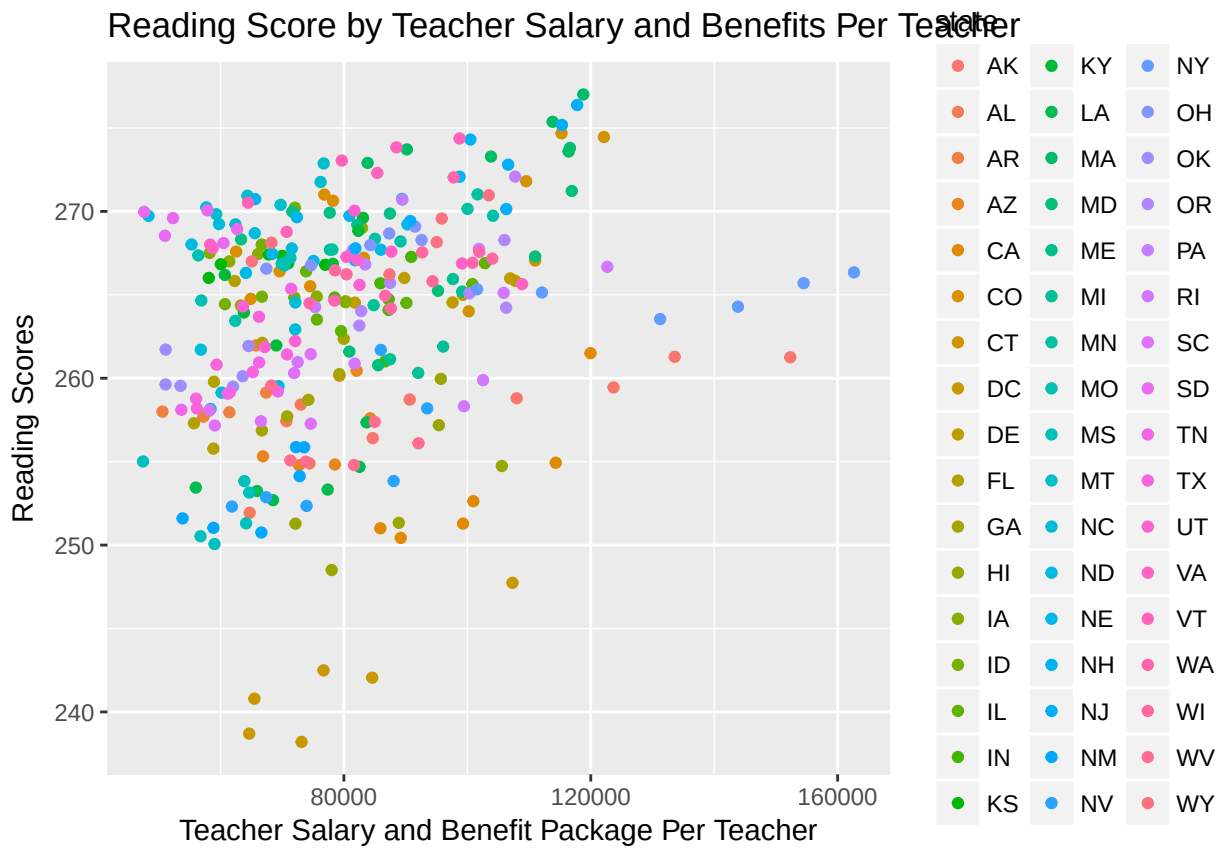


Let's check whether the result is any different when we calculate salary and benefit package per teacher.

```
teacher_salary_per_teacher <- dbGetQuery(con, "SELECT (f.teacher_salaries + f.teacher_benefits)/nf.total
                              s.math_score, s.reading_score, s.state from fiscal f, naep8 s, nonfiscal n
                              where f.state=s.state and f.state = nf.state and
                              f.survey_year = s.test_year and f.survey_year = nf.survey_year;")
ggplot(teacher_salary_per_teacher, aes(x = teacher_pay_per_teacher, y = math_score, colour = state)) + g
  labs(x = "Teacher Salary and Benefit Package Per Teacher", y = "Math Scores", title = "Math Score by T
```




```
ggplot(teacher_salary_per_teacher, aes(x = teacher_pay_per_teacher, y = reading_score, colour = state))
  labs(x = "Teacher Salary and Benefit Package Per Teacher", y = "Reading Scores", title = "Reading Sco
```



Discussion

Our exploratory analysis identified the following:

1. higher values of total revenue per student appear to correspond to better math and reading scores
2. local revenue appears to be more closely tied to better math and reading scores than total revenue
3. improvements in math and reading scores appear to correspond to increases in revenue, again more closely linked to local revenue
4. smaller teacher/student ratios correspond to better math and reading scores
5. higher teacher salary and benefit packages per teacher correspond to better math and reading scores for students

We also found that the data for DC included outlier values and not applicable values so we might want to focus on the 50 states in our final presentation for more consistent data.