United States Public Elementary-Secondary Education spending (2016)

by Z. McLaughlin

Introduction

There's been a lot of media attention to teacher salaries. This exploration covers various elements of how money is spent on education. One area of particular interest is spending in Oklahoma which has gotten a lot of attention in the news recently. Exploration covers overall trends in spending per student across the country, but then focuses is in Oklahoma vs California vs New Jersey to see how things compare.

Example youtube video of teachers in Oklahoma changing jobs. (http://kfor.com/2018/06/06/watch-oklahoma-teachers-break-down-explaining-why-they-are-leaving-the-classroom/)

Dataset

There is a government census done of all public schools in the United States. This data for 2016 is published here:

2016 Public Elementary-Secondary Education Finance Data (https://www.census.gov/data/tables/2016/econ/school-finances/secondary-education-finance.html)

Since none of the provided tables contained all the desired information created a cleaned dataset using the following information:

- All Data Items (https://www2.census.gov/programs-surveys/school-finances/tables/2016/secondary-education-finance/elsec16.xls) - Contains the most comprehensive data.
- elsec16t.xls (https://www2.census.gov/programs-surveys/school-finances/tables/2016/secondary-education-finance/elsec16t.xls) - Contains state information for each school
- school16doc.doc (https://www2.census.gov/programs-surveys/school-finances/tables/2016/secondary-education-finance/school16doc.doc) - Contains the key for all the column names and key for states.

Tables attached to project assignment:

- elsec_16.csv
- elect16.csv
- state_info_table.csv

Cleaned data created has the following column headers:

- "STATE" - Two letter state abbreviation
- "IDCENSUS" - Census id for the school district
- "NAME" - School district name
- "CONUM" - ANSI State and County Code
- "ENROLL" - Number of students enrolled
- "TOTALREV" - Total revenue in thousands
- "TFEDREV" - Total federal revenue in thousands

- “TSTREV” - Total state revenue in thousands
- “TLOCREV” - Total local revenue in thousands
- “TOTALEXP” - Total Expenses in thousands
- “TCURISAL” - Total salaries & wages for instruction in thousands
- “TCURIBEN” - Total benefits for instruction in thousands
- “PPCSTOT” - Per student total spending
- “PPSALWG” - Per student total salaries and wages
- “PPEMPBEN” - Per student total benefit payments
- “PPITOTAL” - Per student total spending for instruction
- “PPISALWG” - Per student total spending for salaries and wages for instruction
- “PPIEMBEN” - Per student total employee benefits for instruction

After cleaning data, decided to only include school districts with at least 50 students enrolled.

Added a few more columns later in the exploration:

- “PPTISB” - Per student money spent on salaries and benefits for instruction
- “PPPOI” - Percentage of money spent on teacher compensation
- “PLOCALREV” - Local rev percentage of total revenue
- “OKCA” - Limited number of states for analysis
- “PPNORM” - Teacher compensation per student normalized for cost of living.
- “PPTEACH” - Teacher compensation per student normalized for class size

Note: Teacher compensation = Salary + benefits

Summary of cleaned data

```
##      ENROLL            TOTALREV           TFEDREV
##  Min.   :    50.0   Min.   :     234   Min.   :      0
##  1st Qu.:   452.5   1st Qu.:    6388   1st Qu.:    347
##  Median :  1163.0   Median :   16328   Median :    920
##  Mean   :  3724.4   Mean   :   51223   Mean   :   3961
##  3rd Qu.:  3059.5   3rd Qu.:   42748   3rd Qu.:   2590
##  Max.   :981667.0   Max.   :27448356   Max.   :1739101
##      TSTREV            TLOCREV           TOTALEXP
##  Min.   :       0   Min.   :       0   Min.   :     301
##  1st Qu.:    2884   1st Qu.:    2206   1st Qu.:    6245
##  Median :    7775   Median :    6105   Median :   15962
##  Mean   :   24103   Mean   :   23159   Mean   :   50922
##  3rd Qu.:   18936   3rd Qu.:   18391   3rd Qu.:   42114
##  Max.   :10568010   Max.   :15141245   Max.   :29620098
##     TCURISAL           TCURIBEN          PPCSTOT          PPSALWG
##  Min.   :       0   Min.   :      0   Min.   :     0   Min.   :     0
##  1st Qu.:    2030   1st Qu.:    692   1st Qu.:  9455   1st Qu.:  5433
##  Median :    5133   Median :   2058   Median : 11035   Median :  6347
##  Mean   :   16947   Mean   :   6928   Mean   : 12796   Mean   :  7237
##  3rd Qu.:   13947   3rd Qu.:   5795   3rd Qu.: 14345   3rd Qu.:  8031
##  Max.   :10044302   Max.   :6258743   Max.   :374873   Max.   :183063
##     PPEMPBEN         PPITOTAL          PPISALWG         PPIEMBEN
##  Min.   :    0   Min.   : -1472   Min.   :     0   Min.   :    0
##  1st Qu.: 1827   1st Qu.:  5637   1st Qu.:  3641   1st Qu.: 1200
##  Median : 2552   Median :  6566   Median :  4269   Median : 1669
##  Mean   : 2981   Mean   :  7606   Mean   :  4861   Mean   : 1995
##  3rd Qu.: 3633   3rd Qu.:  8554   3rd Qu.:  5426   3rd Qu.: 2490
##  Max.   :76688   Max.   :228102   Max.   :131903   Max.   :55750
```

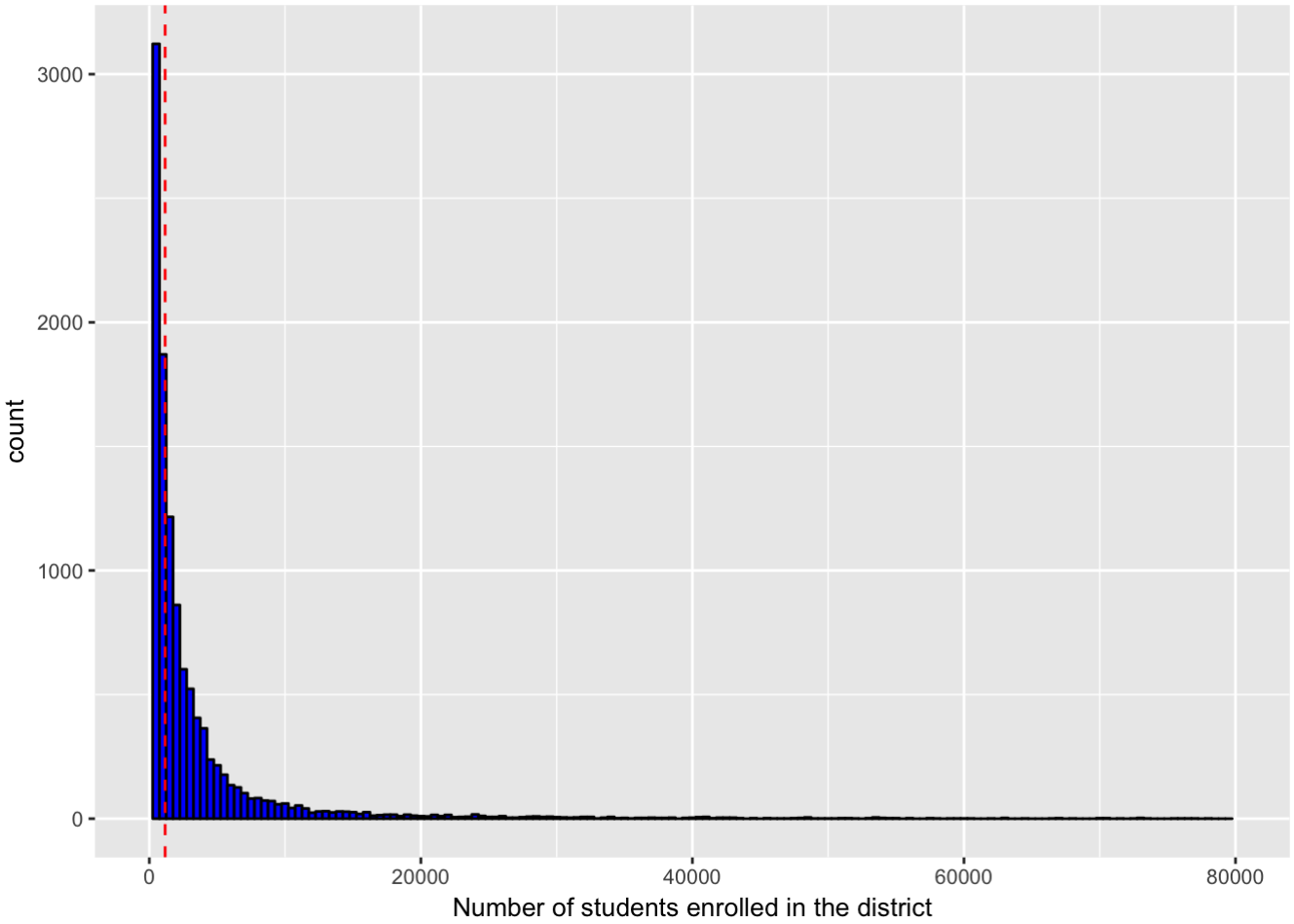
The data varies a lot:

- Enrollment - One school district has almost a million students while others have less than 50 (although anything less than 50 was filtered out.)
- Total revenue max is over $27 billion dollars
- Max spent per pupil/student is $228,102
- Just a quick look at the Revenue from federal, state and local, it’s clear that federal only plays a small part in funding education
- The per student total spending varies from -1,472 to 228,102 indicating that 1. Some data might be questionable so it’s better to look at trends rather than the outliers.

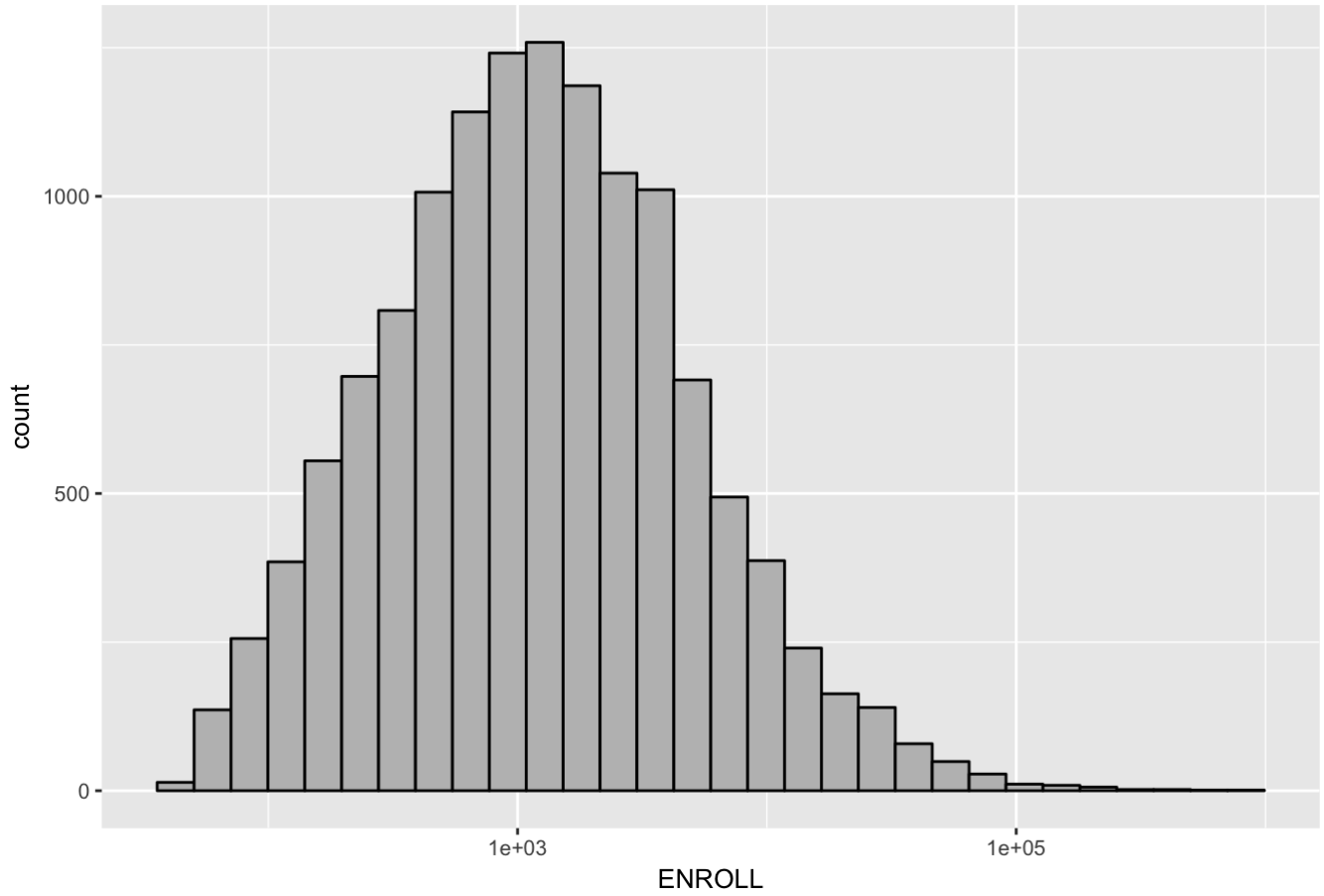
Univariate Plots Section

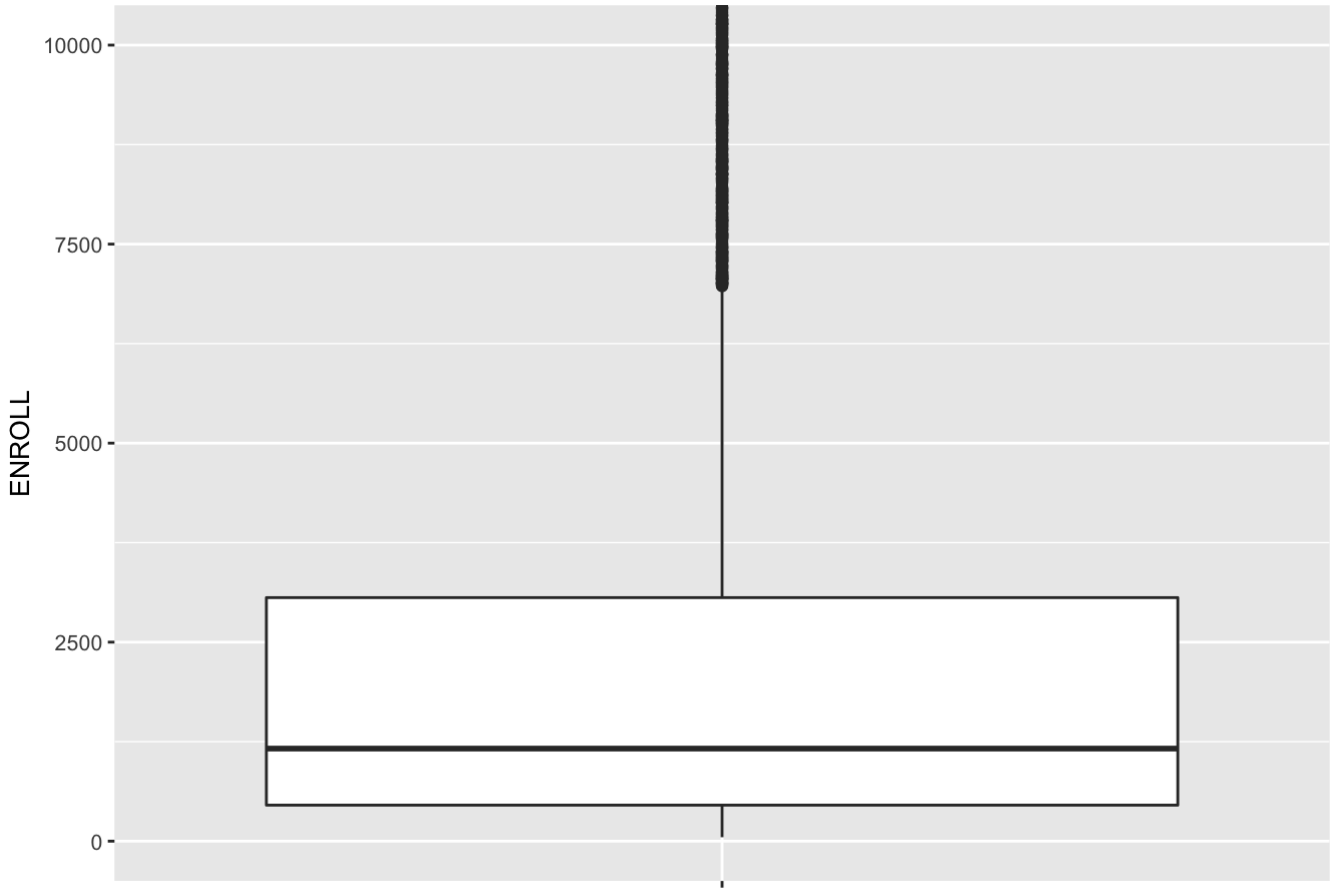
Histogram exploration

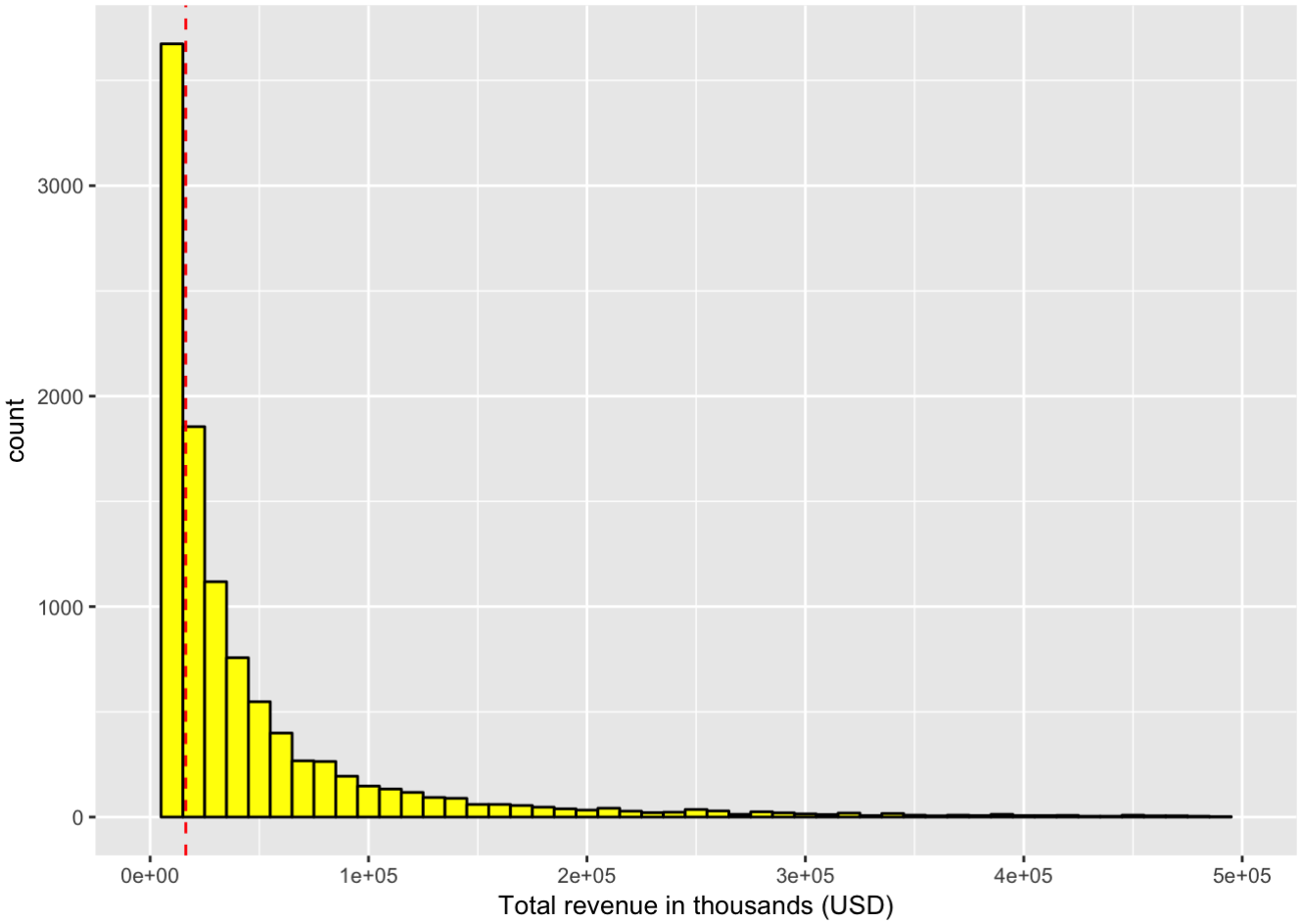
Investigating how large the school districts are (note each district may contain a different number of separate schools.) Also investigating what the per student spending looks like. Median is marked with red dashed line.

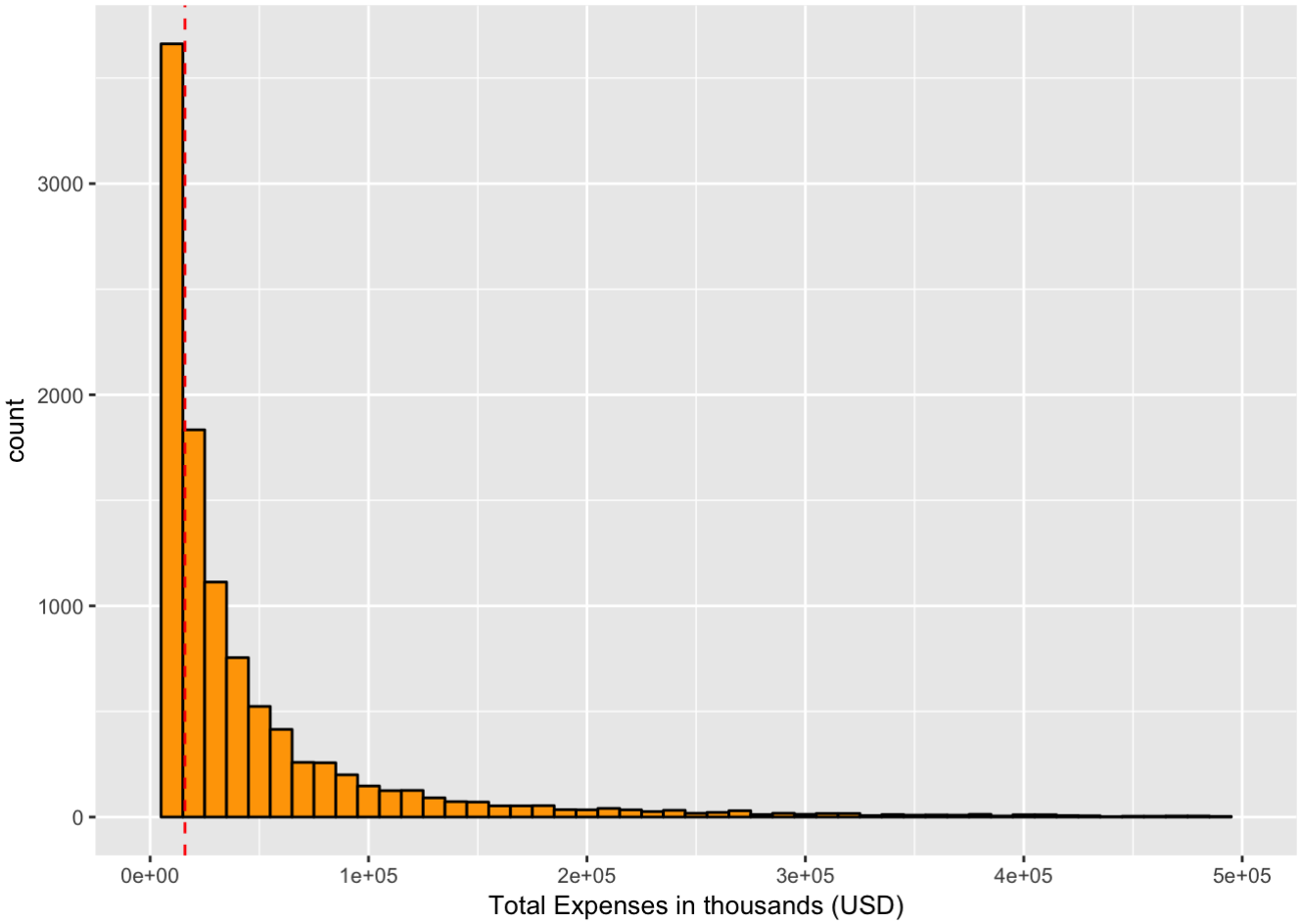


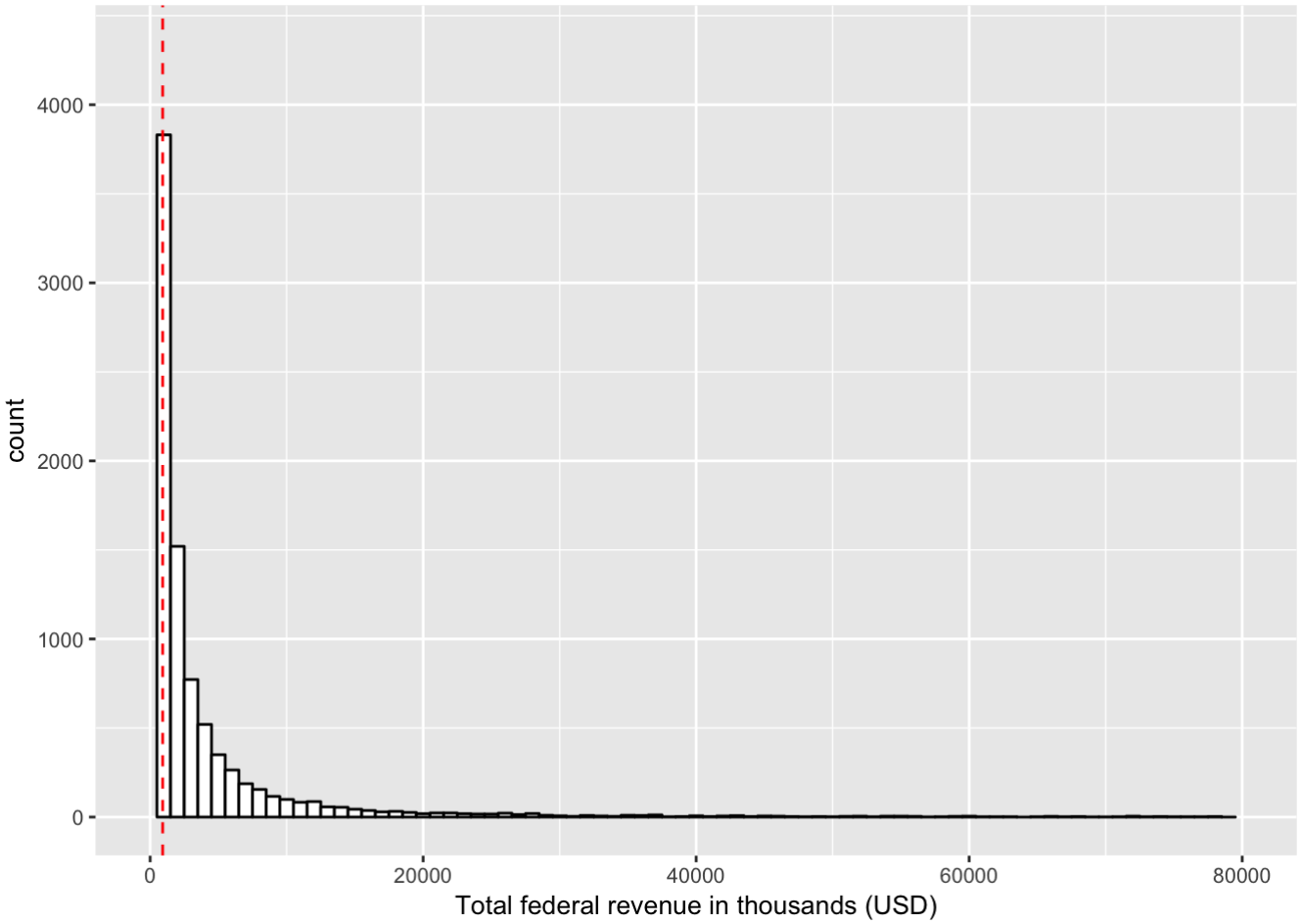
Histogram using log 10 - note 'x' axis scale

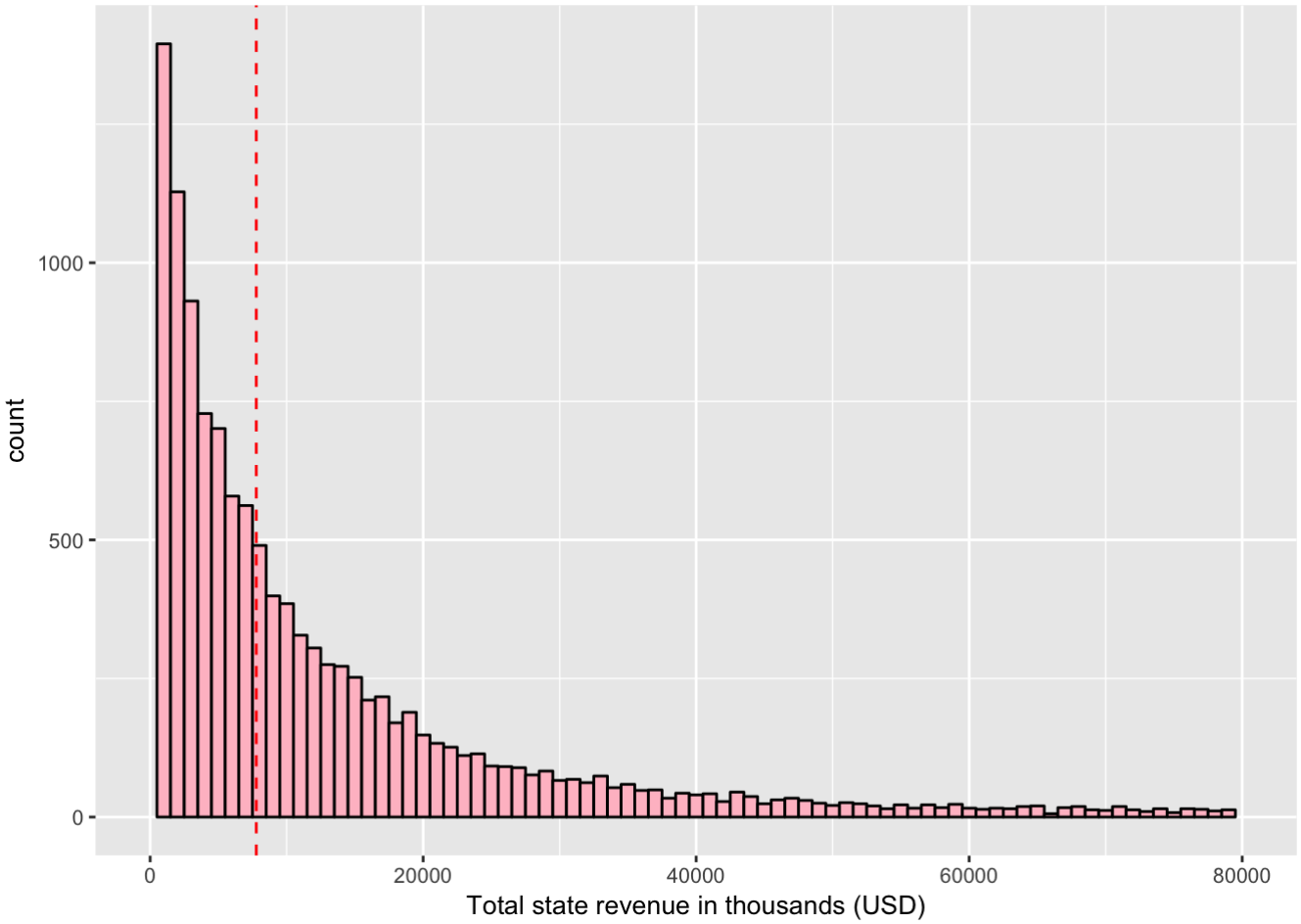


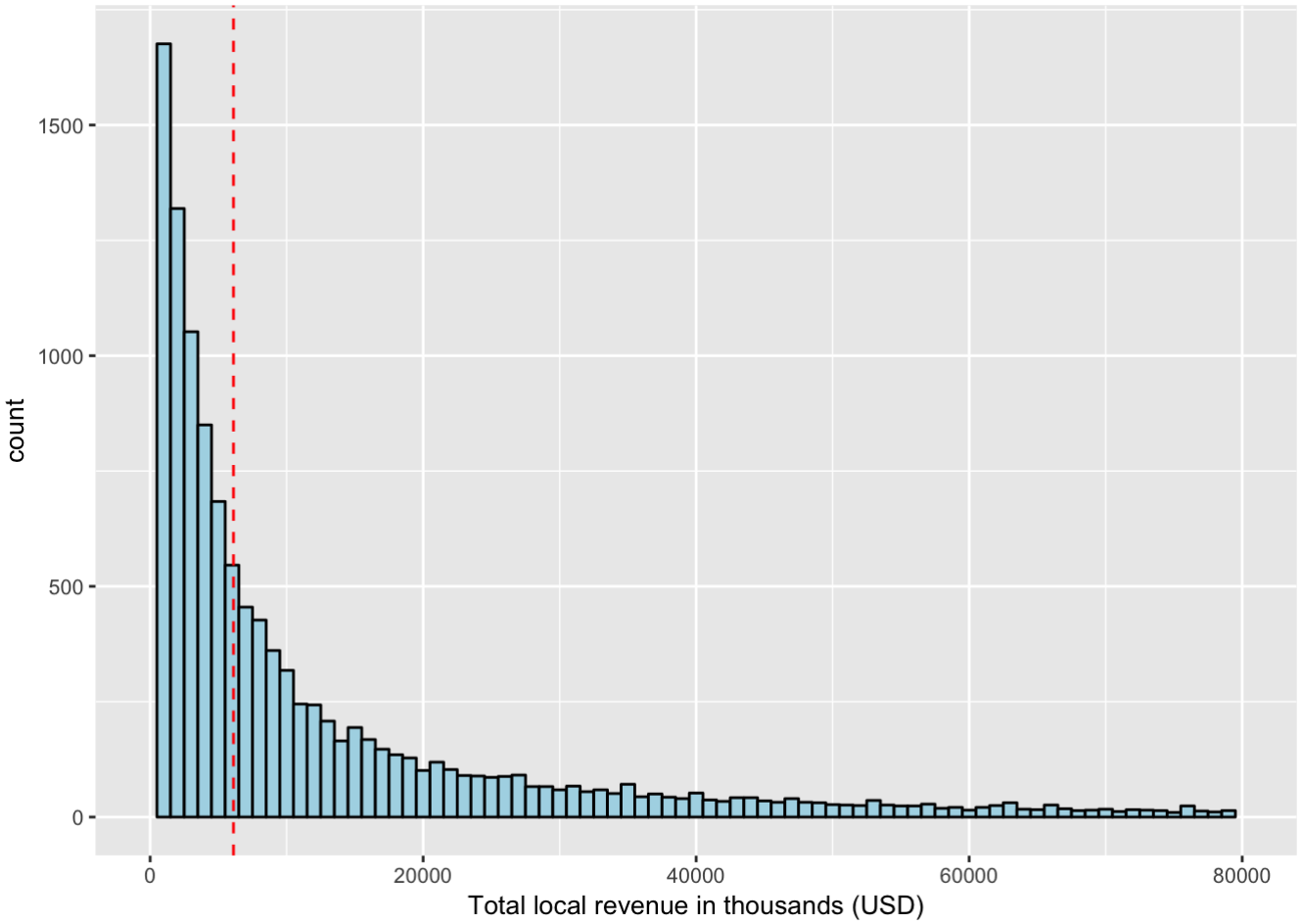


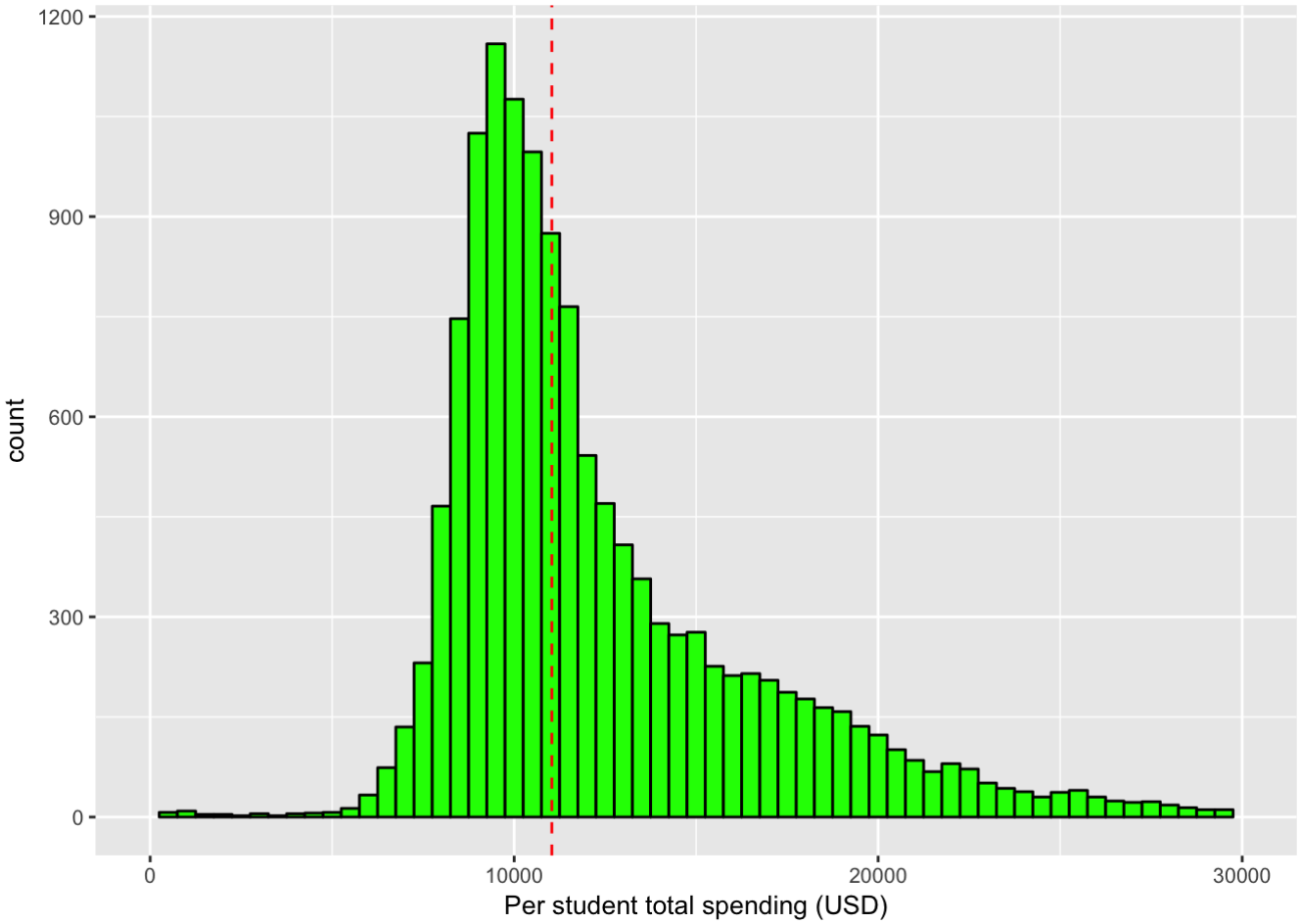

count
3000
2000
1000
0
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Total revenue in thousands (USD)

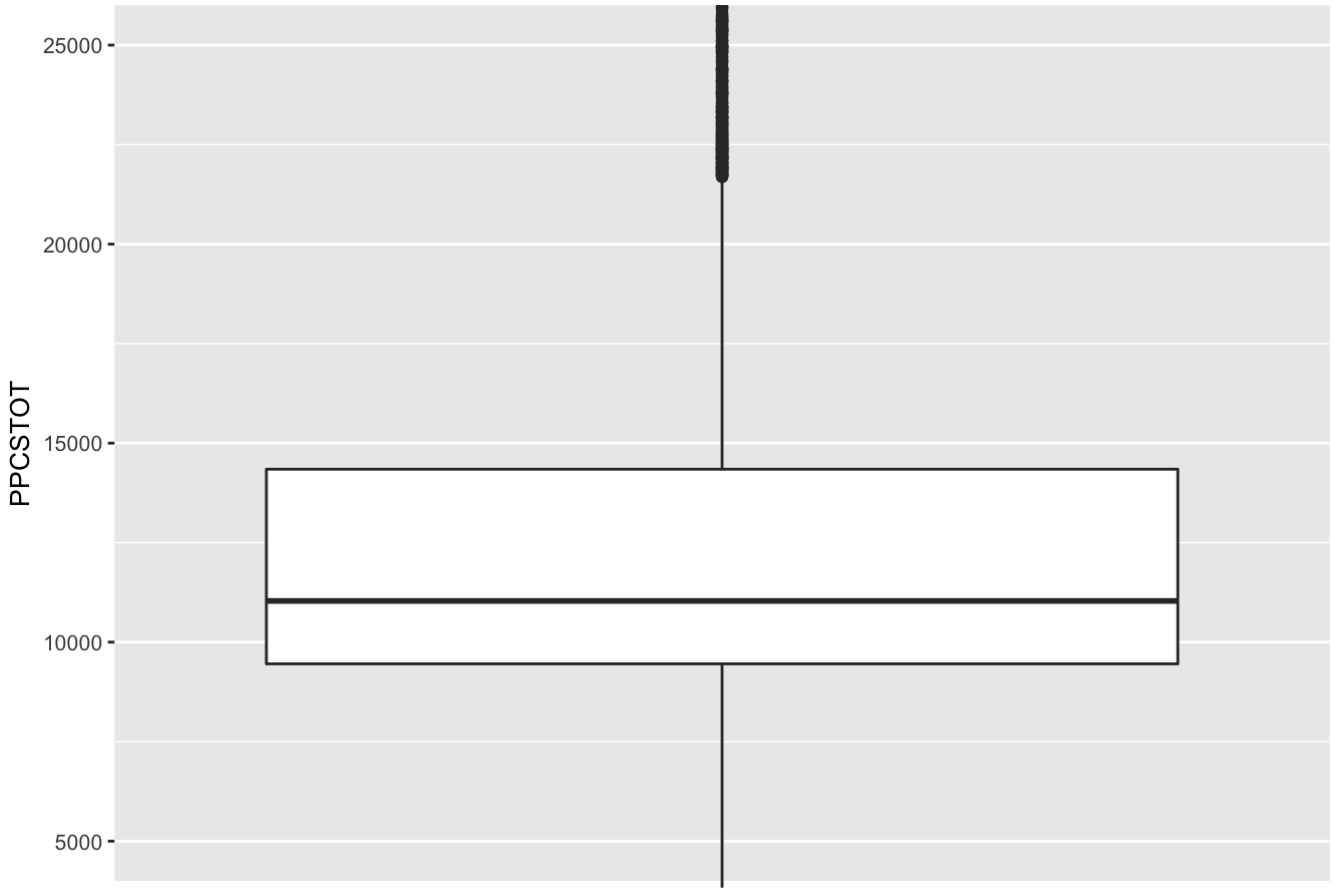

count
3000
2000
1000
0
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Total Expenses in thousands (USD)

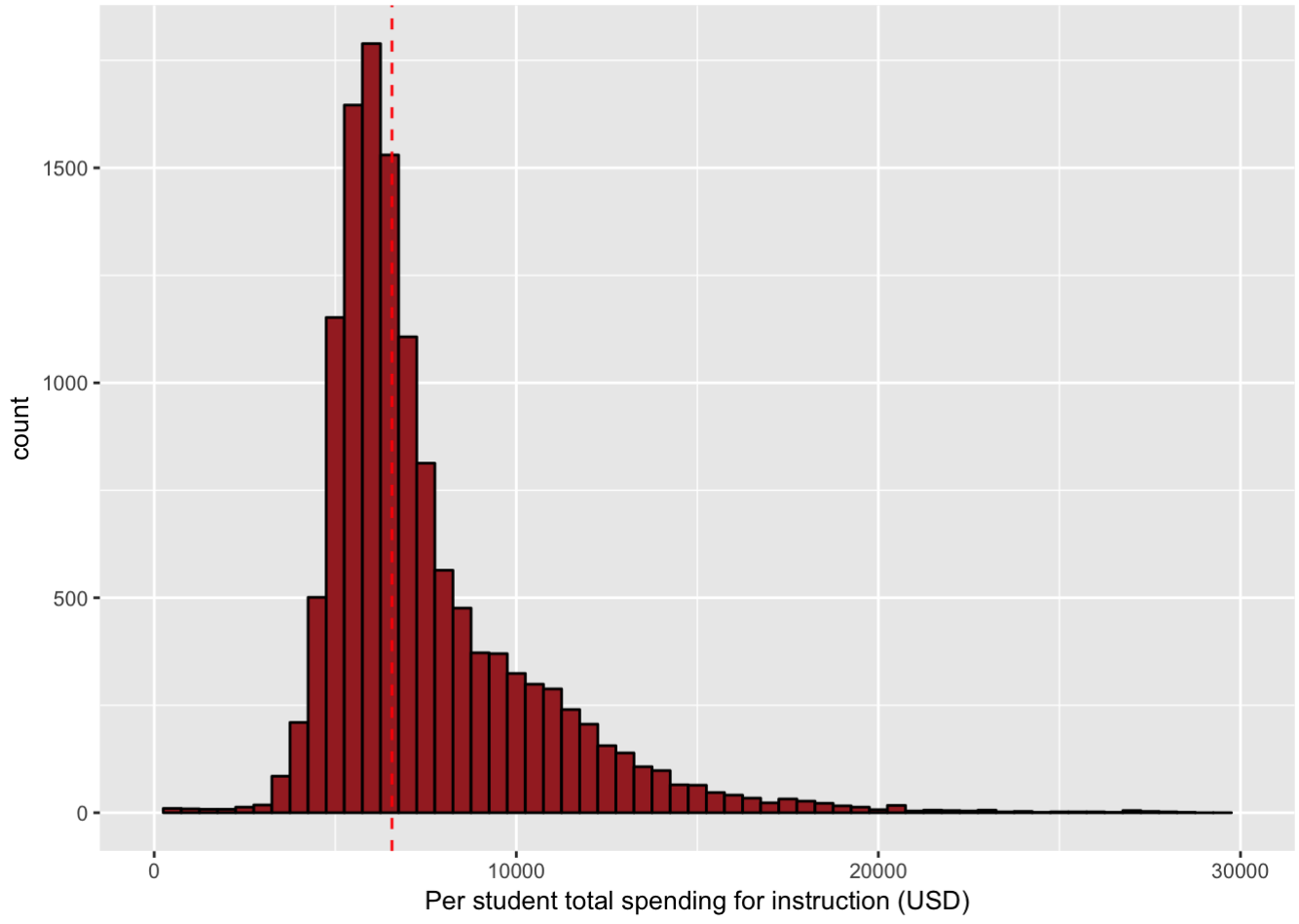

4000
3000
2000
1000
0
count
0
20000
40000
60000
80000
Total federal revenue in thousands (USD)

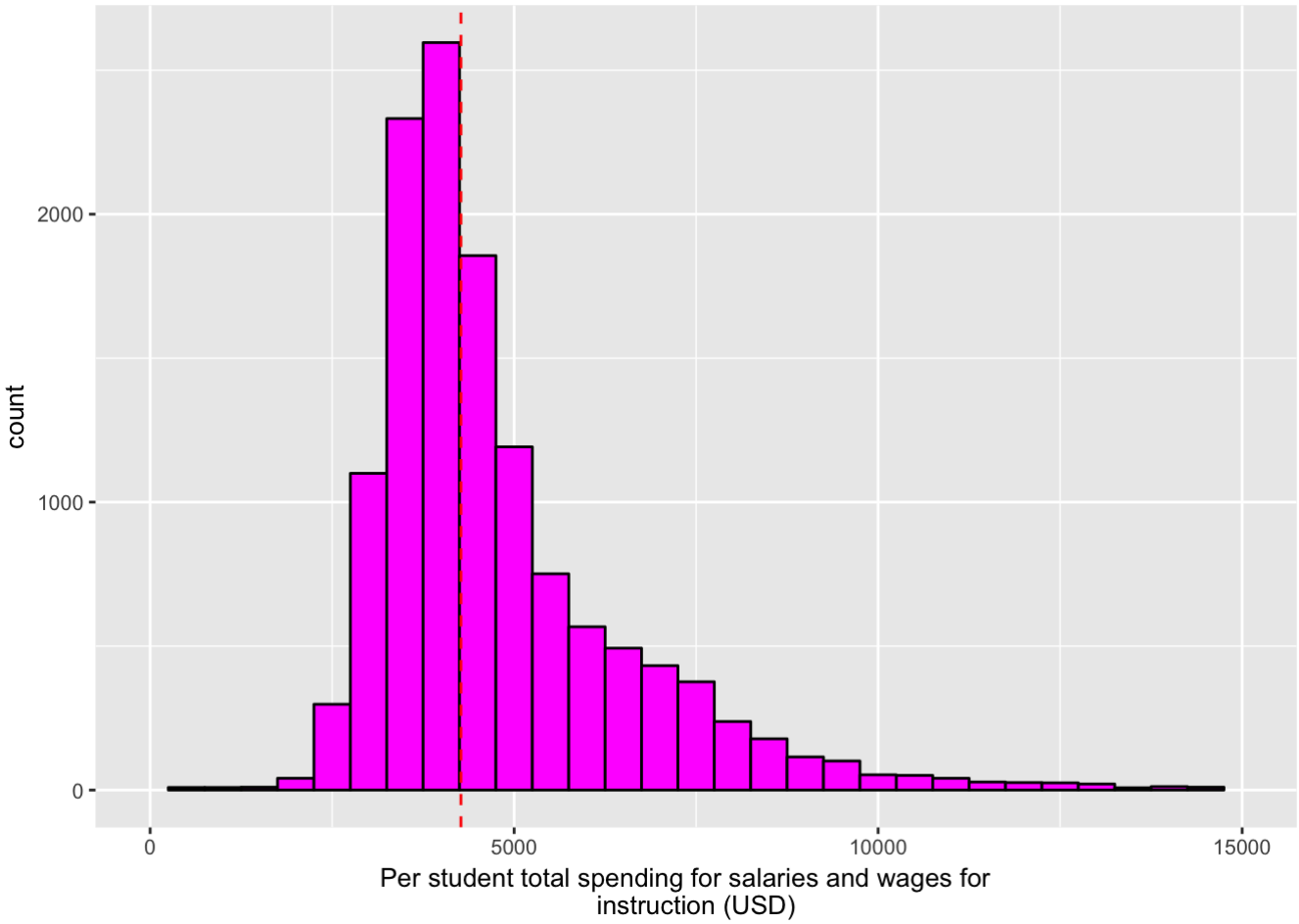

1000
500
0
count
0
20000
40000
60000
80000
Total state revenue in thousands (USD)

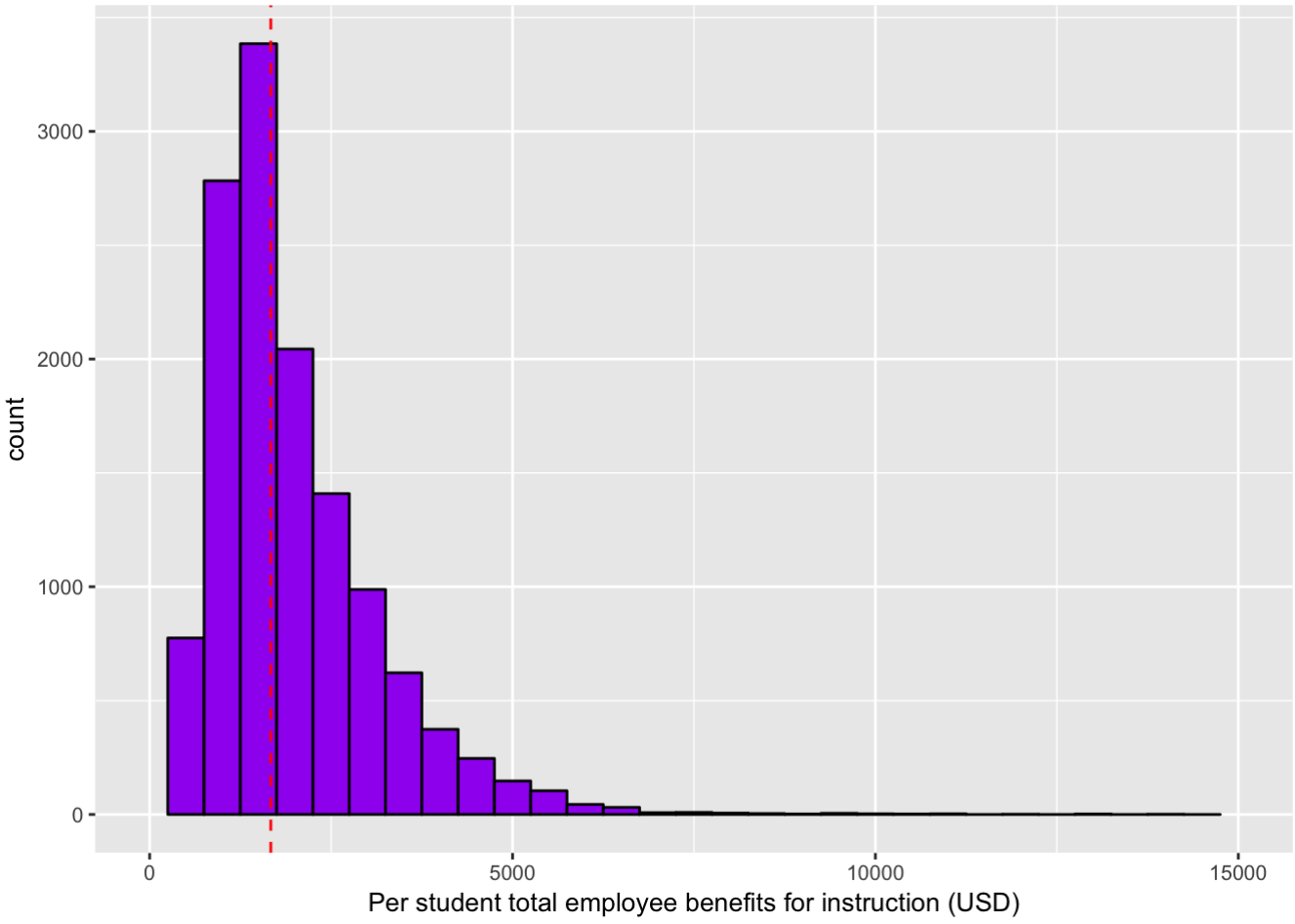

1500
1000
500
0
count
0
20000
40000
60000
80000
Total local revenue in thousands (USD)


1200
900
600
300
0
count
0
10000
20000
30000
Per student total spending (USD)


25000
20000
15000
10000
5000
PPCSTOT


1500
1000
500
0
count
0
10000
20000
30000
Per student total spending for instruction (USD)

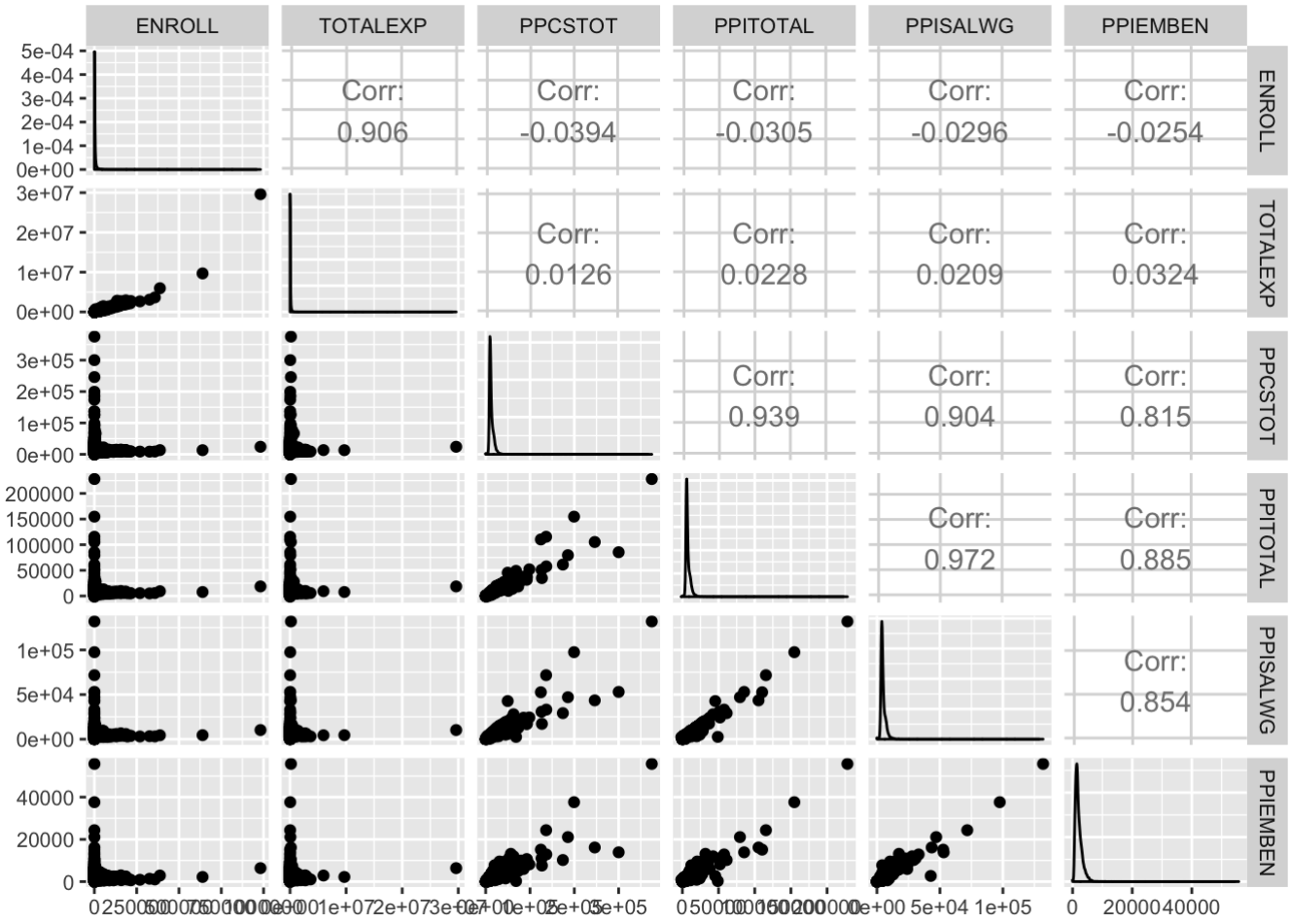

count
2000
1000
0
0
5000
10000
15000
Per student total spending for salaries and wages for instruction (USD)


count
3000
2000
1000
0
0
5000
10000
15000
Per student total employee benefits for instruction (USD)

- All the histograms are right skewed with there still being values far to the right.
- Using a histogram with log 10 was able to see the normalized distribution for enrollment. This was a good sest of data for using the log 10 because it's easy to think of numbers of people in sets of 100,1000,10000, etc.
- Using a boxplot it was possible to see that although there's a large variation in school size, there's a large number of school districts under 2500 students.
- The total amount spent per student is also skewed right, but a little more normalized. It is unexpected that the amount spent on a per student basis would differ so drastically. Possible factors: cost of living, special needs population?
- Also created a box plot for total spending per student to get an idea of the distribution.

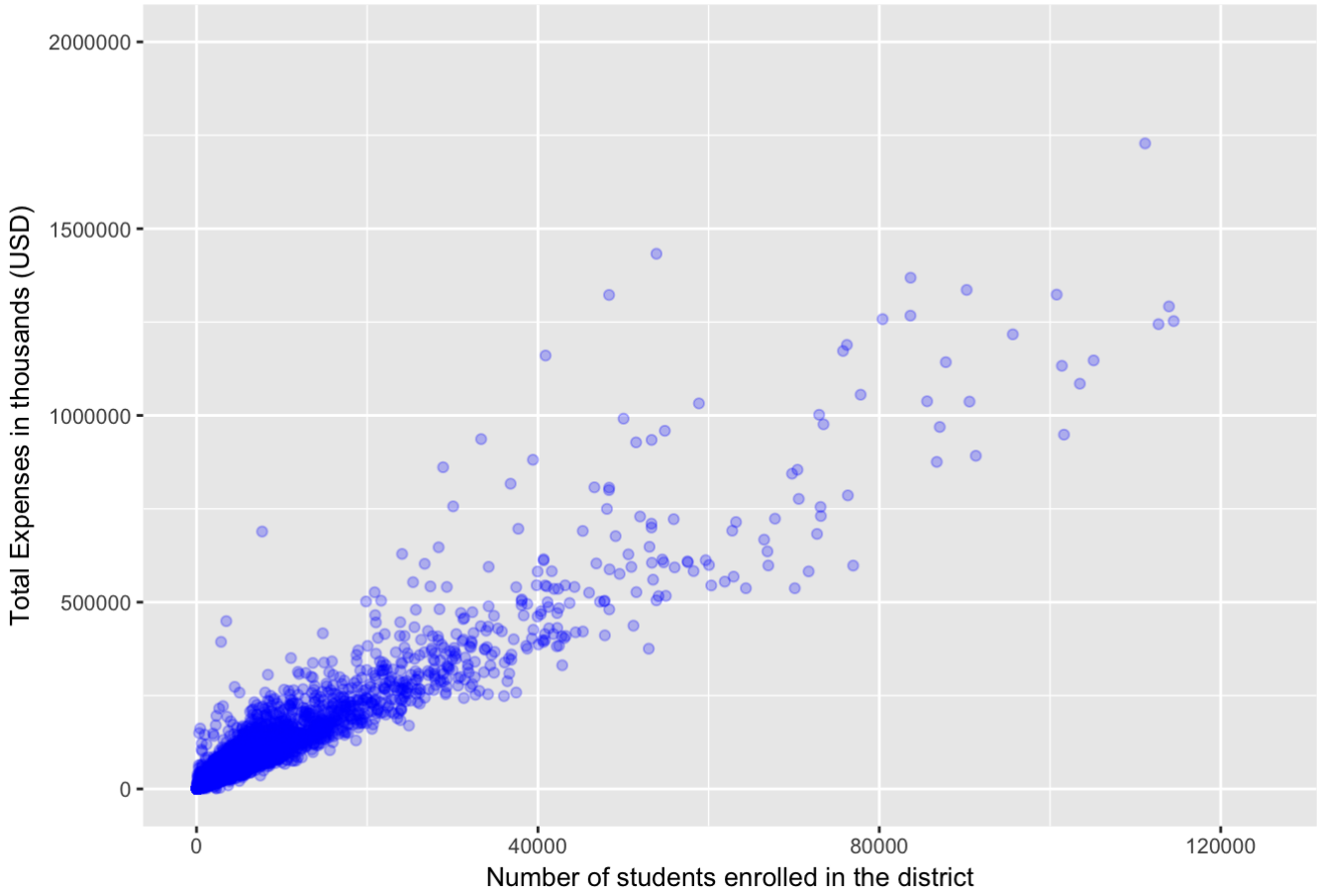
Bivariate Plots

GGPAIRS matrix - Getting some basic correlation data

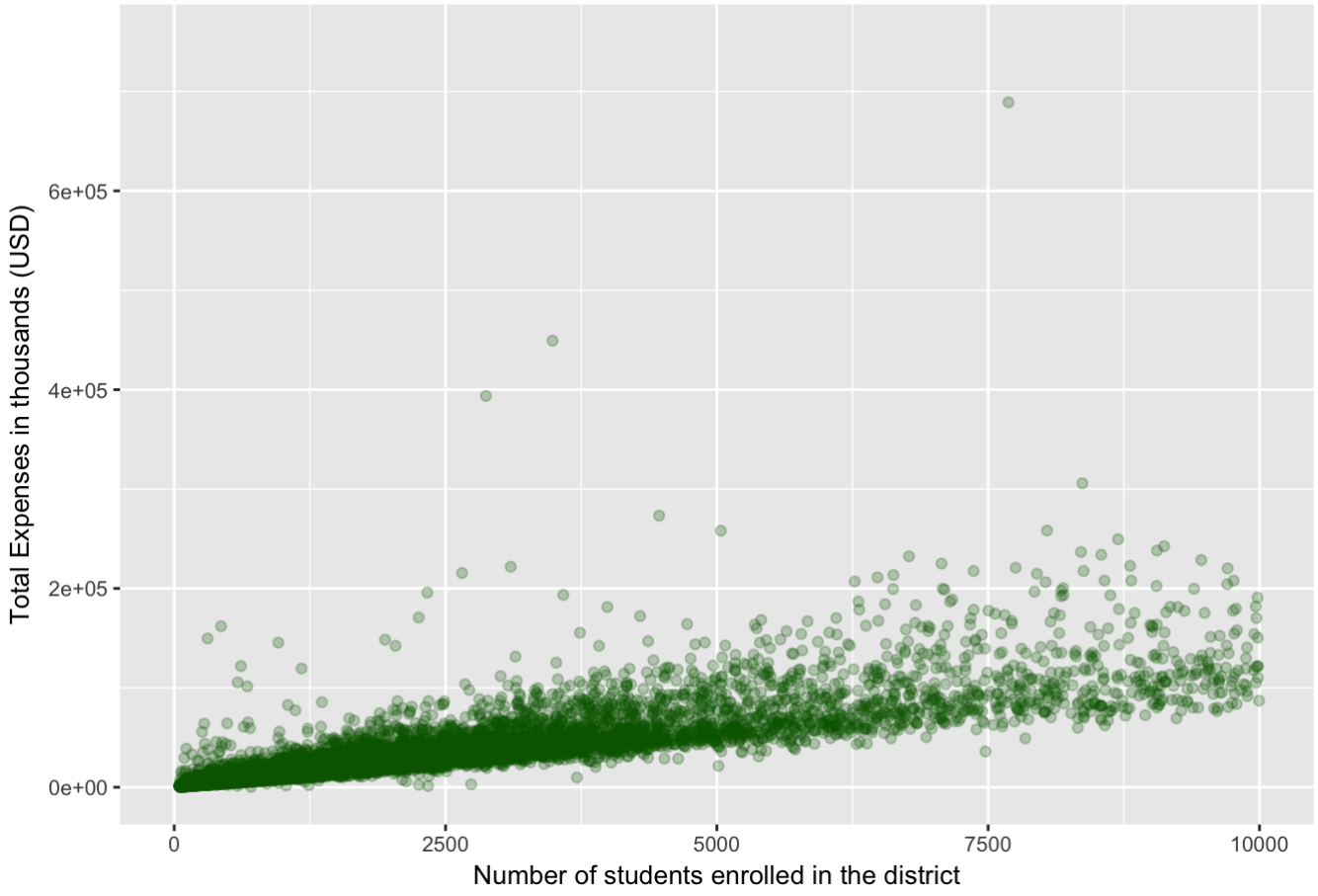


Plots of enrollment vs expenses

Includes schools up to 120,000 students



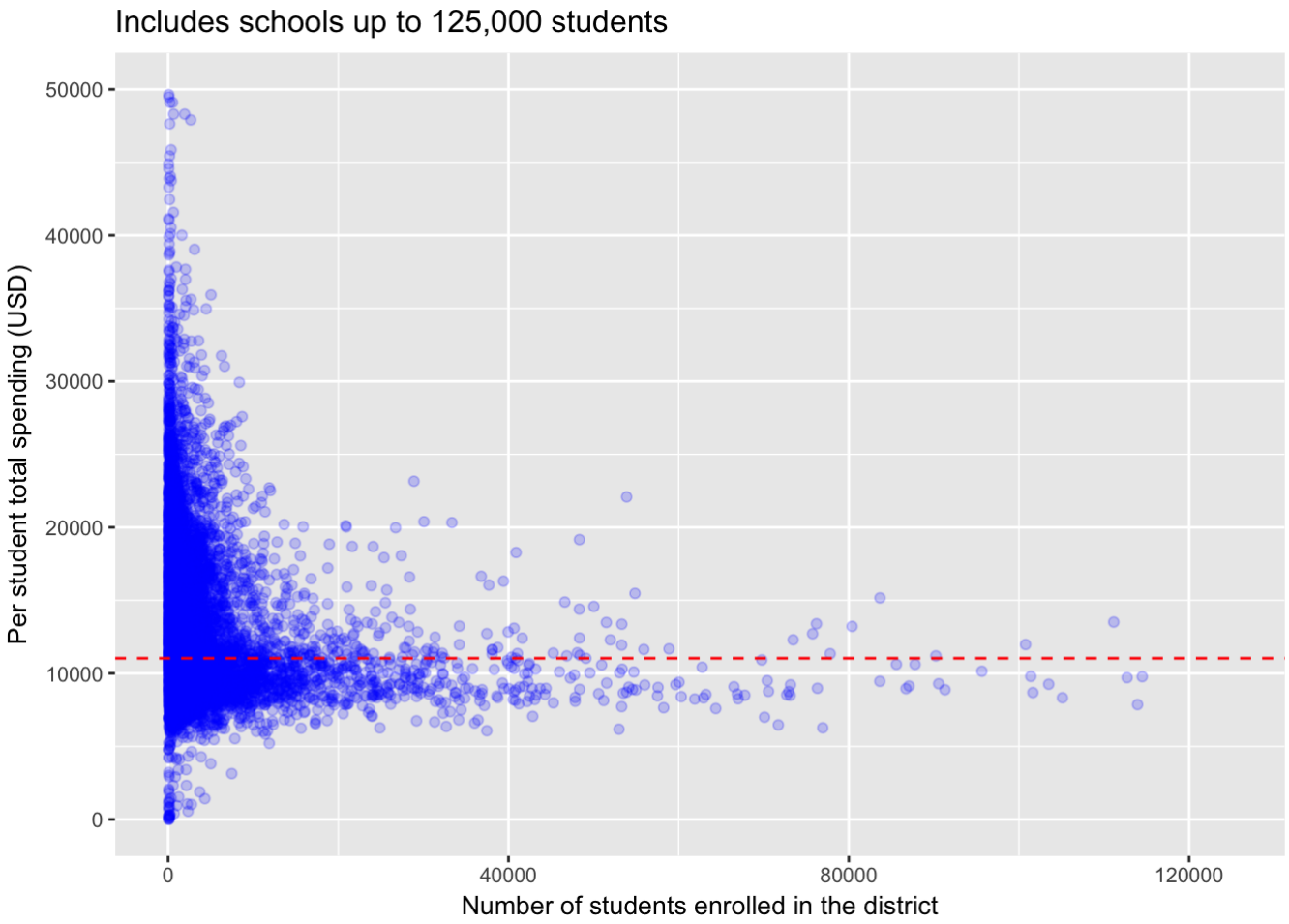
Includes schools up to 10,000 students

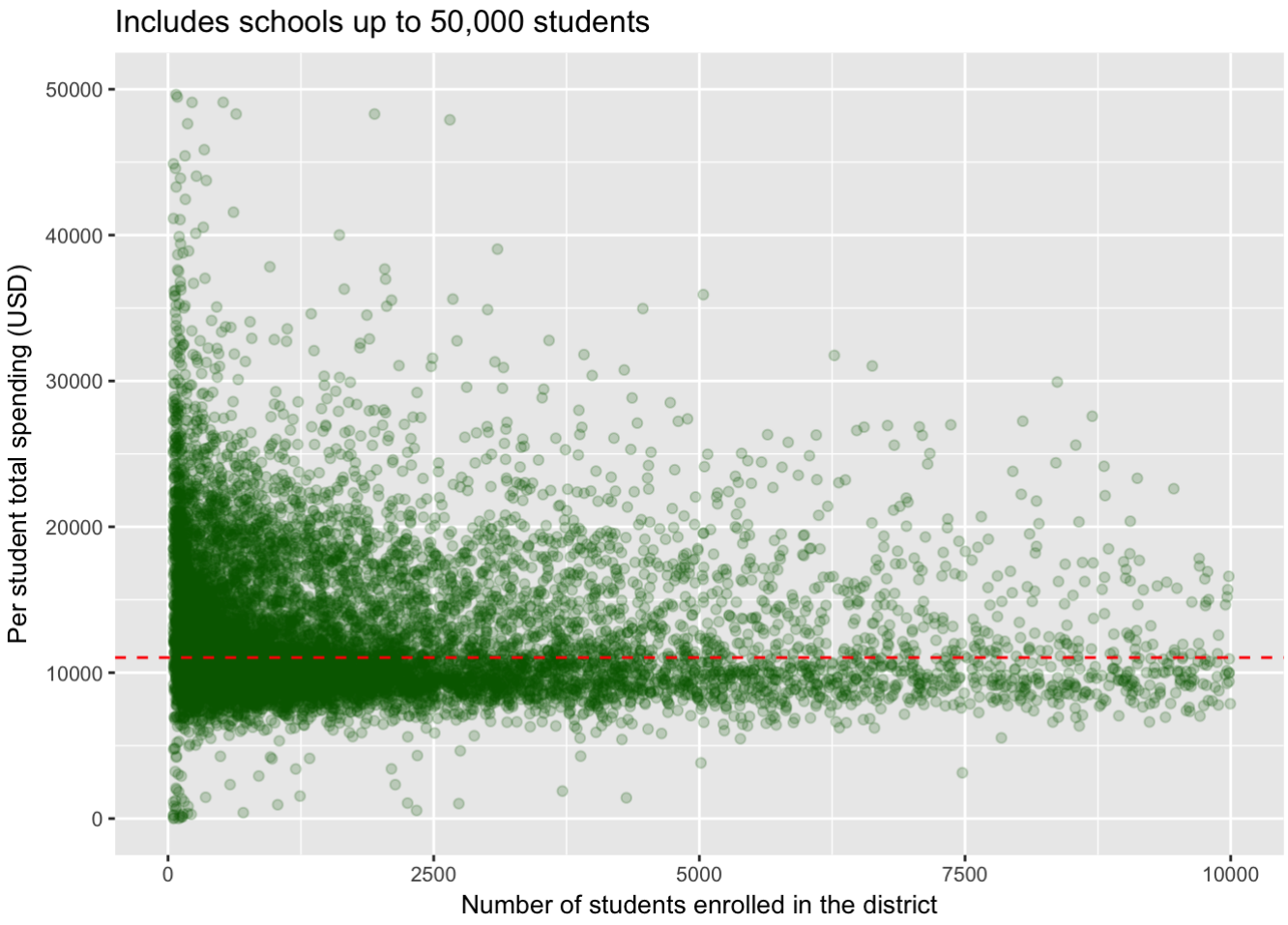


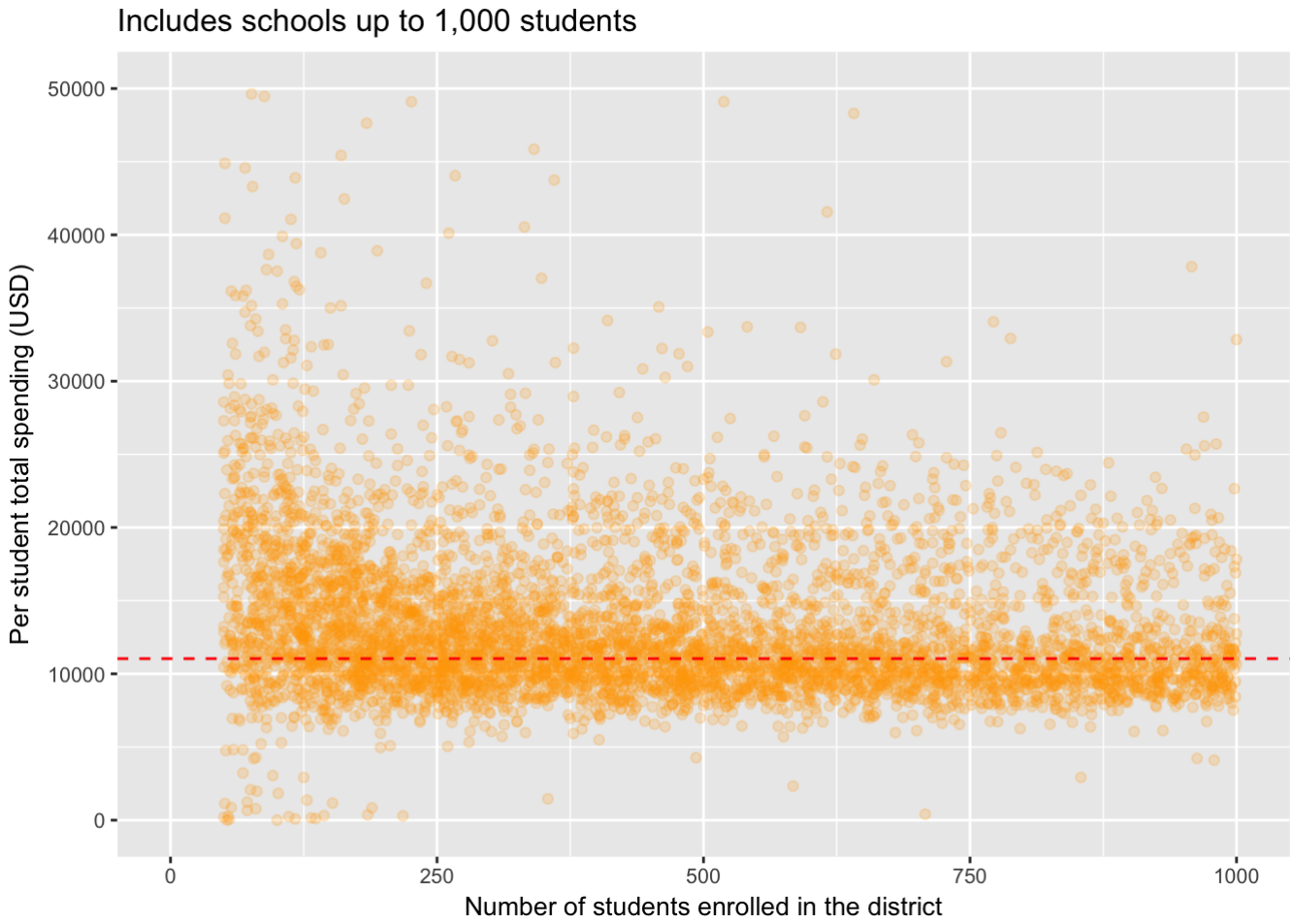
The relationship between enrollment and and total expenses looks very linear at the lower numbers, but as the enrollment goes up, the relationship scatters. Also even though the relationship is fairly linear there is quite a bit of variation and some outliers.

Note that at this point decided to look mainly at districts with less than 125,000 vs districts like NYC that has almost a million students or Hawaii where the whole state is one district.

Plots of enrollment vs total per student sending


Includes schools up to 125,000 students
Per student total spending (USD)
50000
40000
30000
20000
10000
0
0
40000
80000
120000
Number of students enrolled in the district

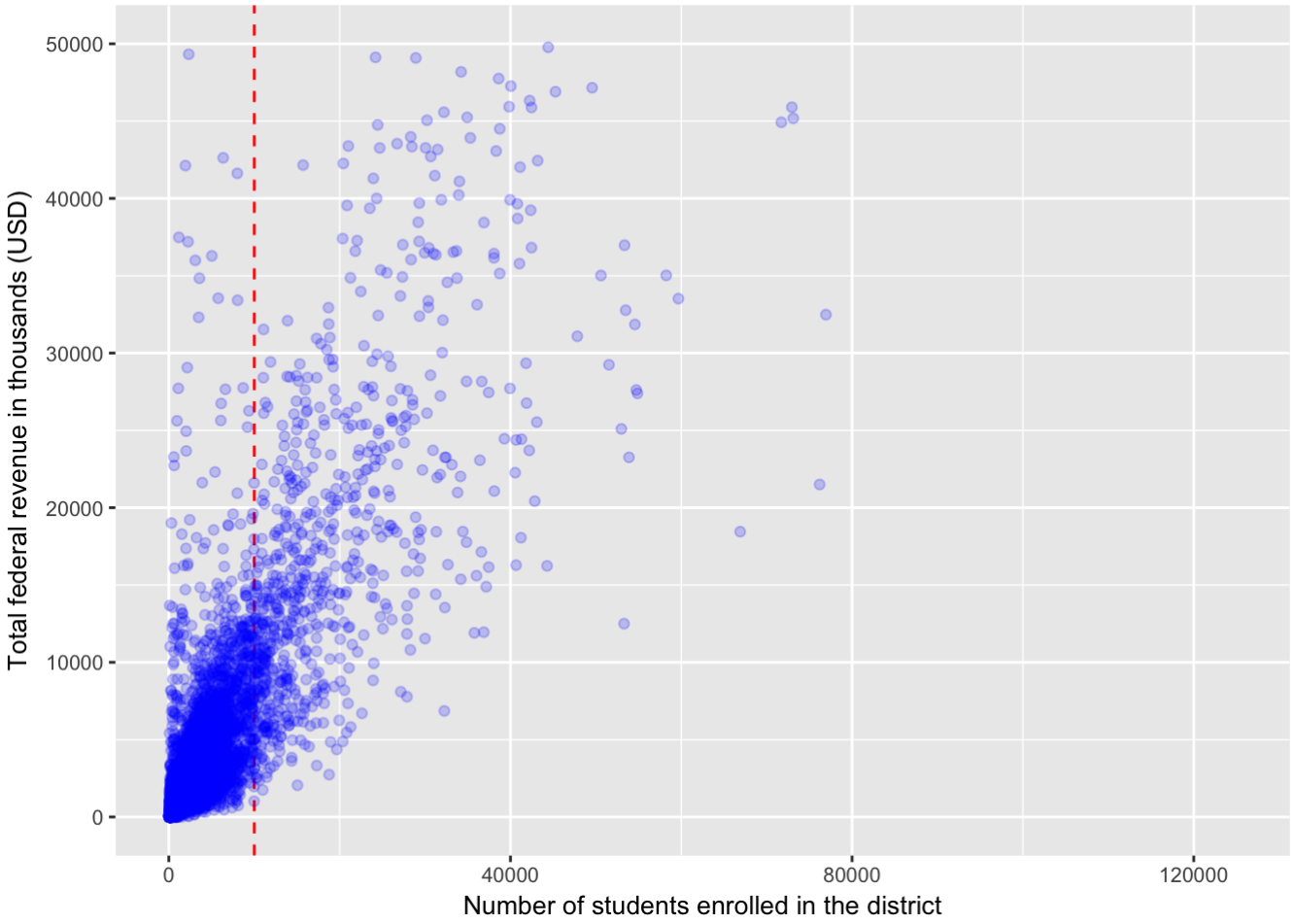

Includes schools up to 50,000 students
Per student total spending (USD)
50000
40000
30000
20000
10000
0
0
2500
5000
7500
10000
Number of students enrolled in the district

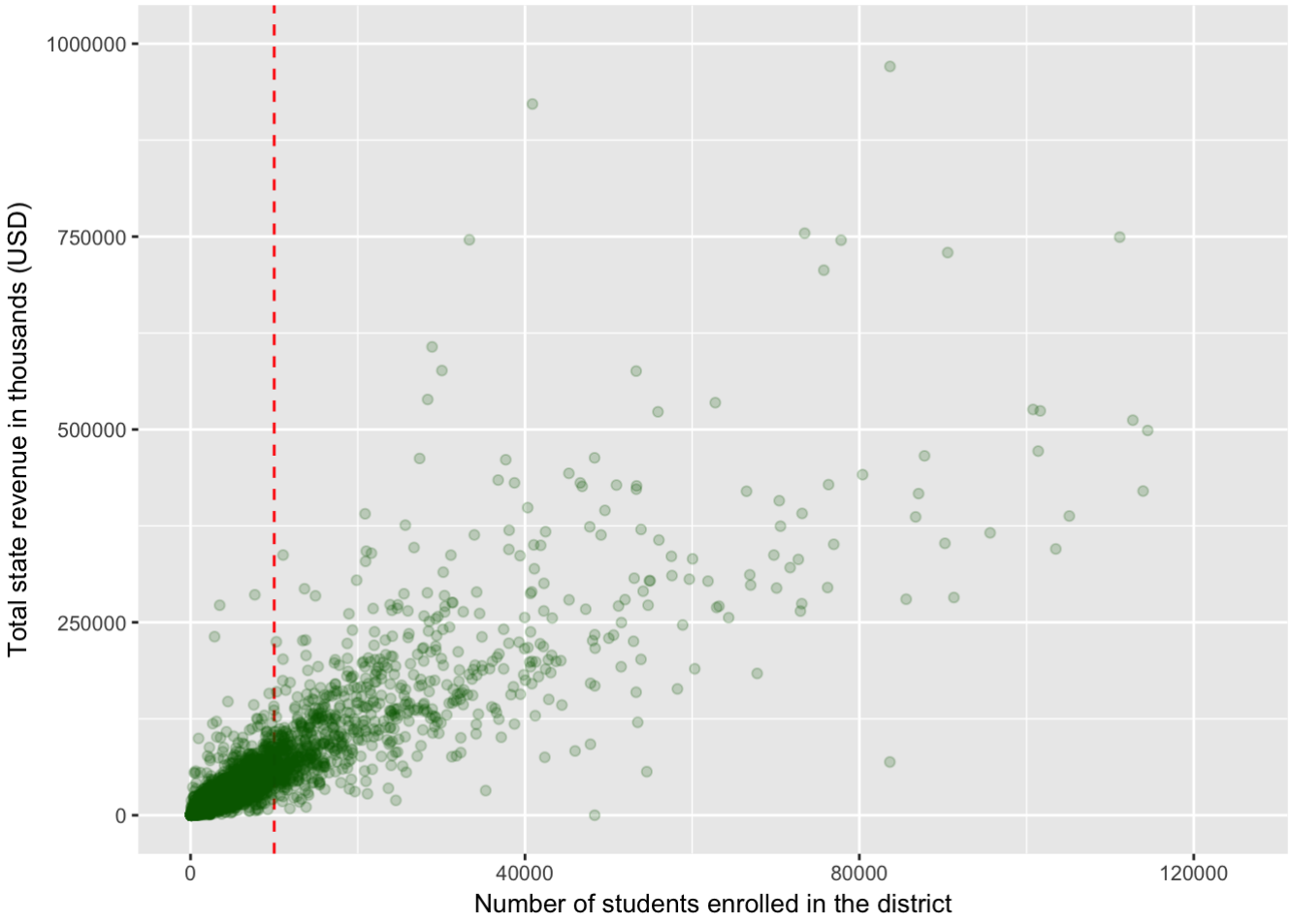


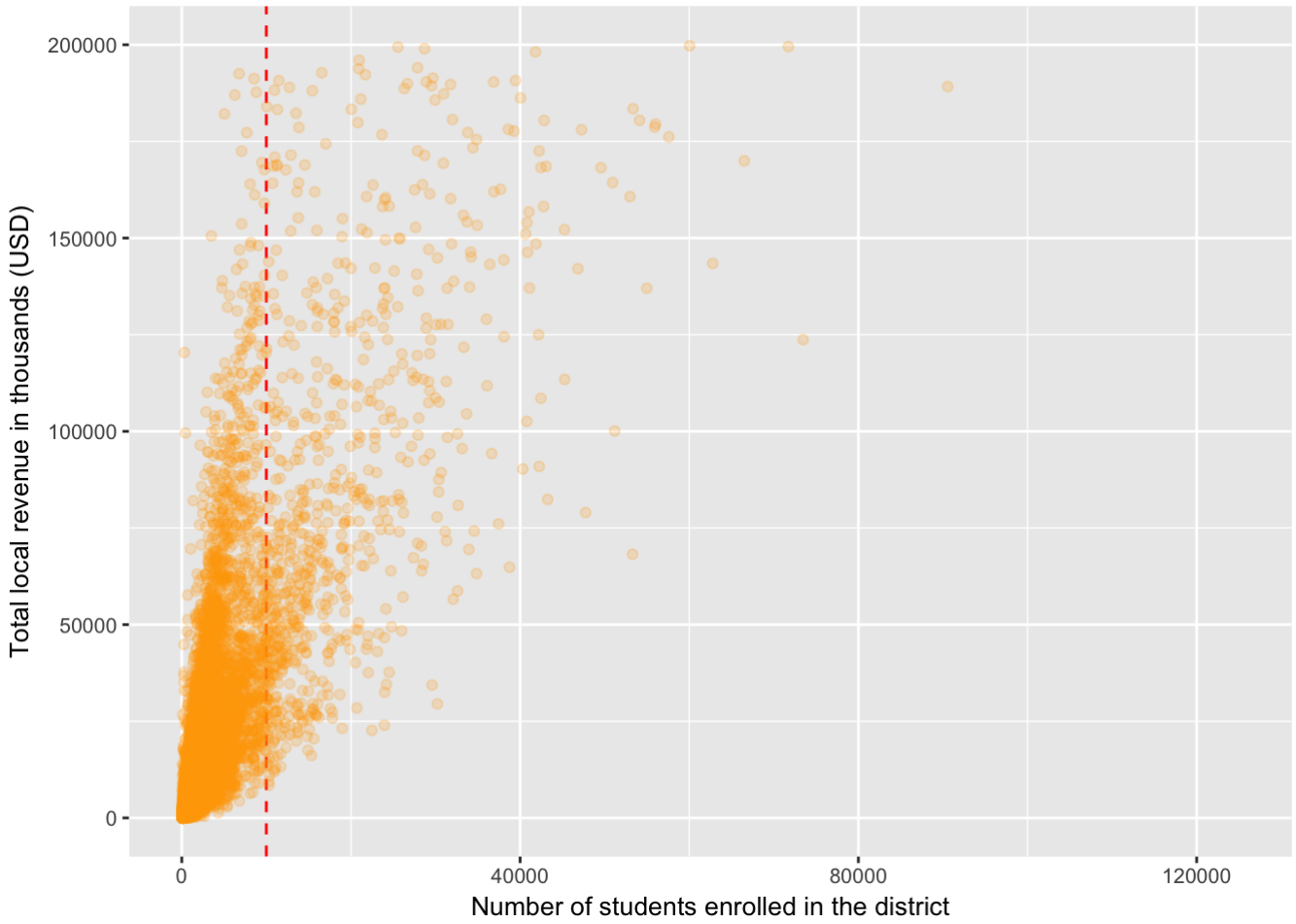
Horizontal line on plots above is the median. It appears that the spending per student is at the lower enrollments hovers around the median, but there are quite a few schools where the spending is substantially higher. Starting to bring up questions:

- What is the difference between the school that spends $7,000/student vs the one that spends $20,000/student?
- The giant school districts seem to hover closer to the median than the smaller school districts. Why?

Plots of enrollment vs revenue sources (Federal, State, Local)


Total federal revenue in thousands (USD)
50000
40000
30000
20000
10000
0
0
40000
80000
120000
Number of students enrolled in the district


Total state revenue in thousands (USD)
1000000
750000
500000
250000
0
0
40000
80000
120000
Number of students enrolled in the district



Added a red dashed line for enrollment = 10,000 students to make it easier to compare the various sources of funding.

Looks like most money comes from Where the money comes from looks interesting. While state funding looks pretty linear. Looks like most money for education comes from state and local sources.

Bivariate Plots Section 2

Based on the plots above there is a need to create additional data to get more specific plots.

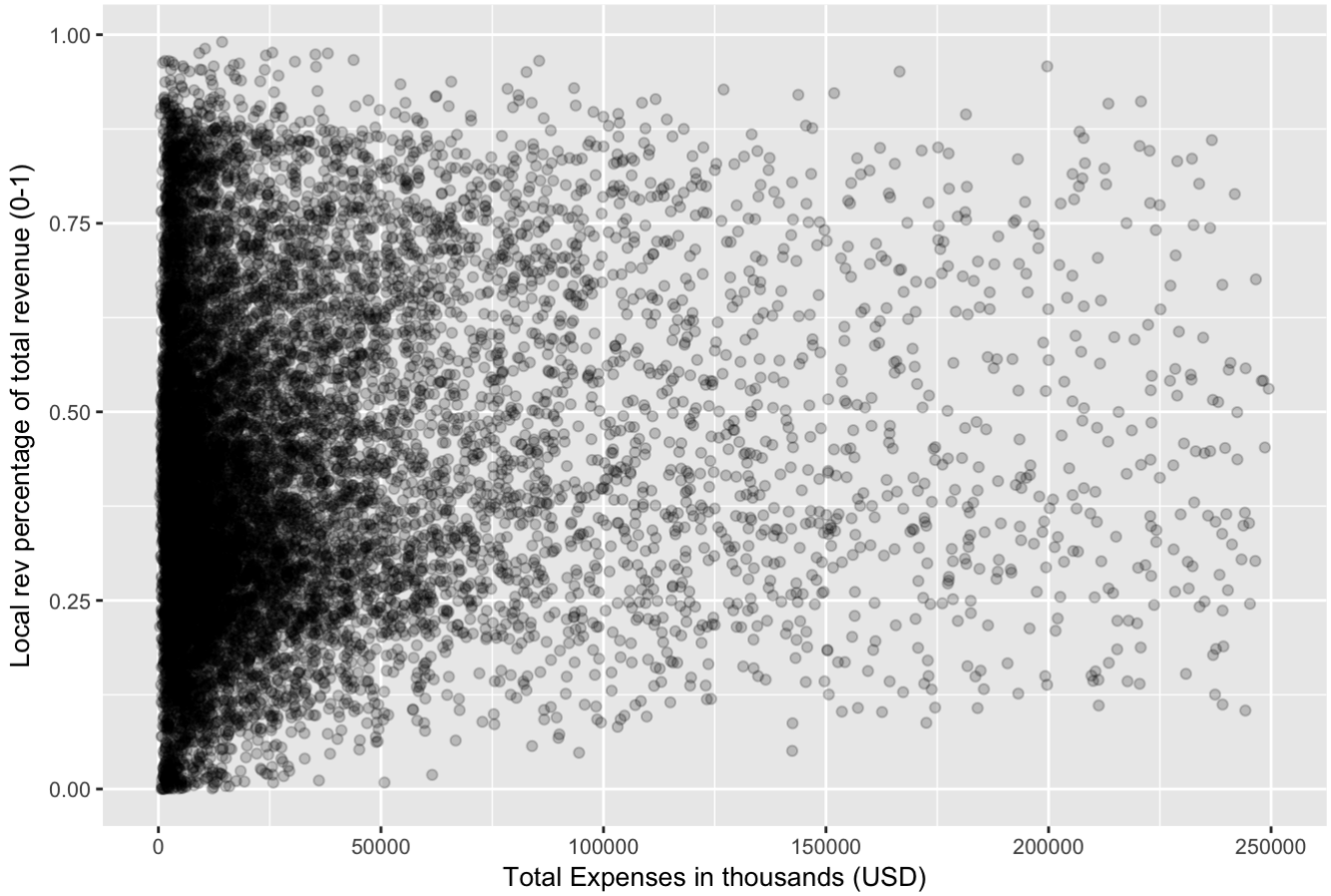
- PPTISB = PPISALWG + PPIEMBEN - Teacher compensation per student for instruction
- PPPOI = PPTISB/PPCSTOT - Percentage of money spent on teacher compensation
- PLOCALREV = TLOCREV/TOTALREV - Local rev percentage of total revenue
- OKCA = OK, CA, NY, NJ, AZ, WA, OR, MA, IA OTHER to be able to more detailed information at the state level on the same plot.

```
##      PPTISB           PPPOI          PLOCALREV
##  Min.   :     0   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  5004   1st Qu.:0.5002   1st Qu.:0.2696
##  Median :  5882   Median :0.5420   Median :0.3997
##  Mean   :  6856   Mean   :0.5371   Mean   :0.4307
##  3rd Qu.:  7757   3rd Qu.:0.5803   3rd Qu.:0.5817
##  Max.   :187653   Max.   :2.1406   Max.   :0.9904
##                   NA's   :2
```

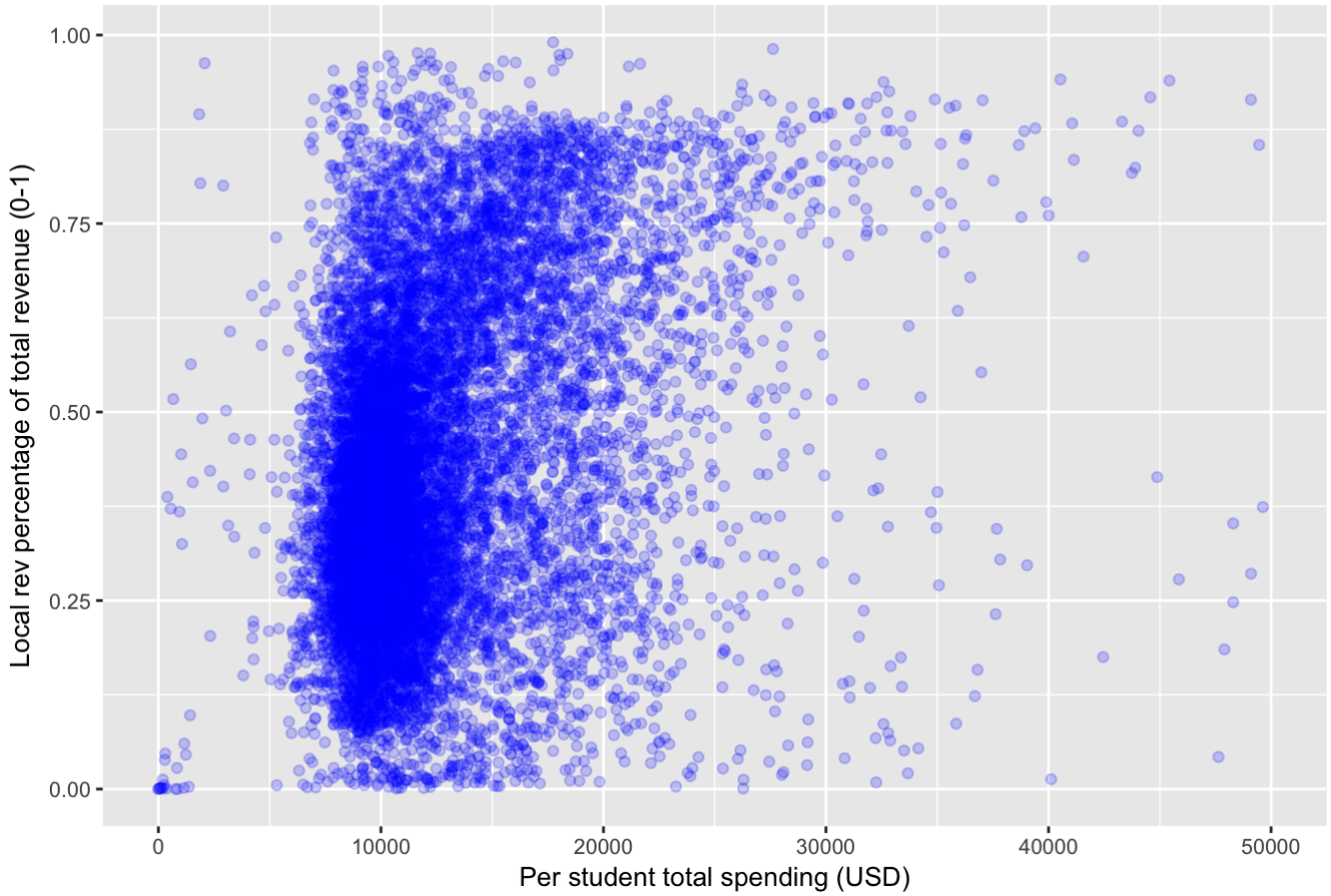
The summary showed percentages over 1 for percent of total spent per pupil for salaries, wages benefits. Did some spot checks and found that there were only a handful of schools impacted. Brings up questions as to the reliability of the data, but since this is not for professional use, will not investigate further.

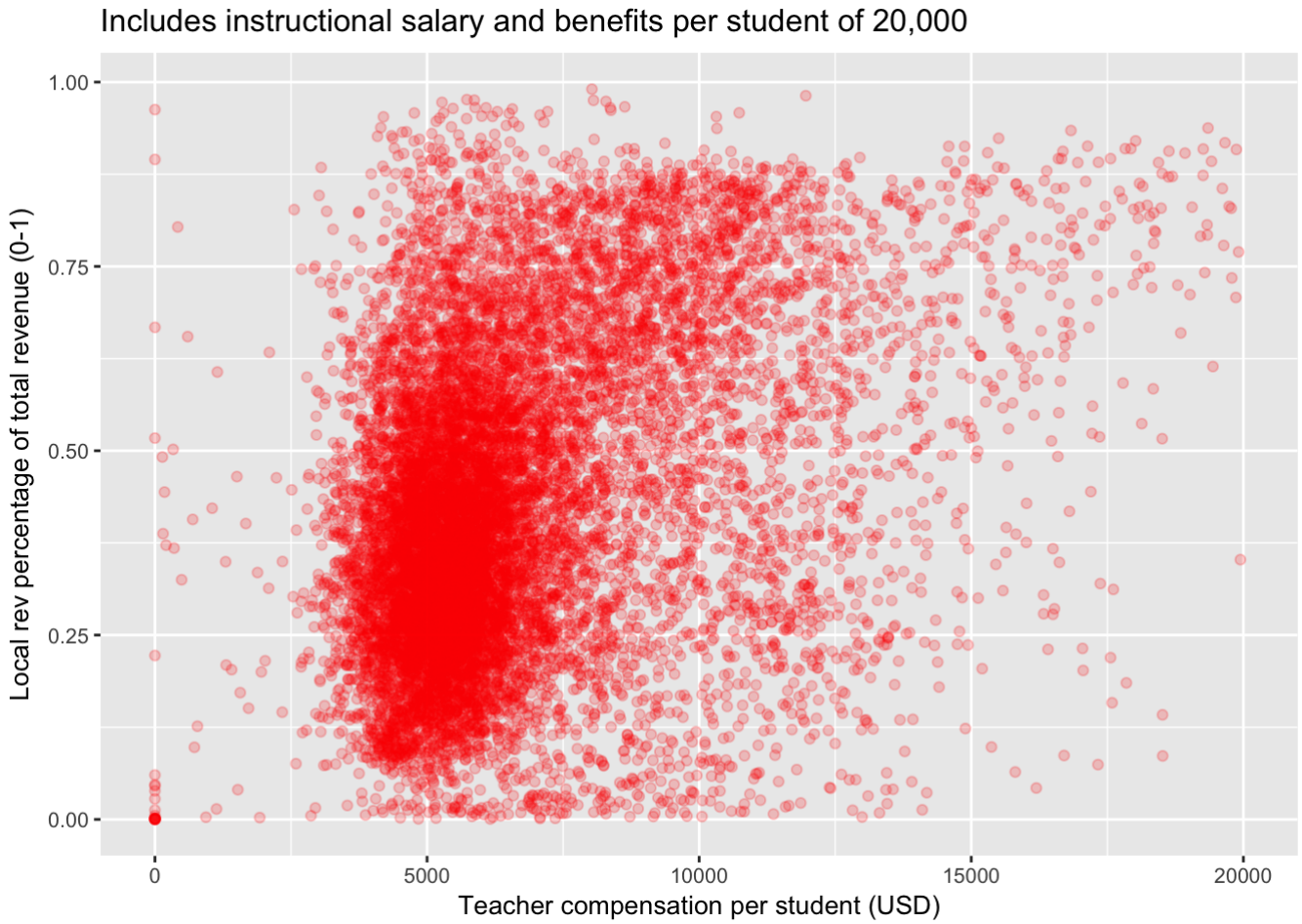
Plots Looking at percentage of local rev vs other variables

Includes revenue up to 250,000,000 dollars



Includes up to 50,000 dollars per student spending

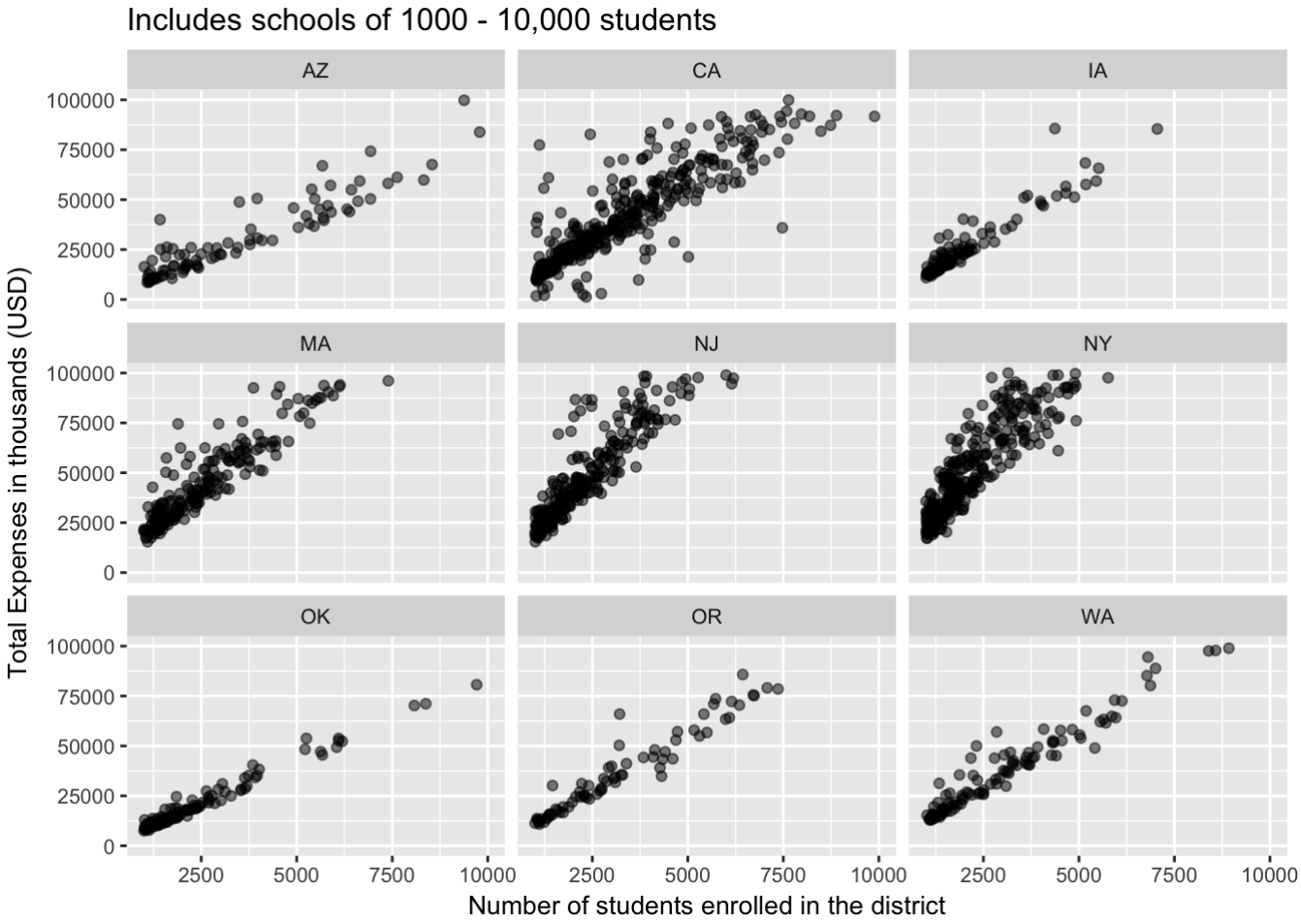


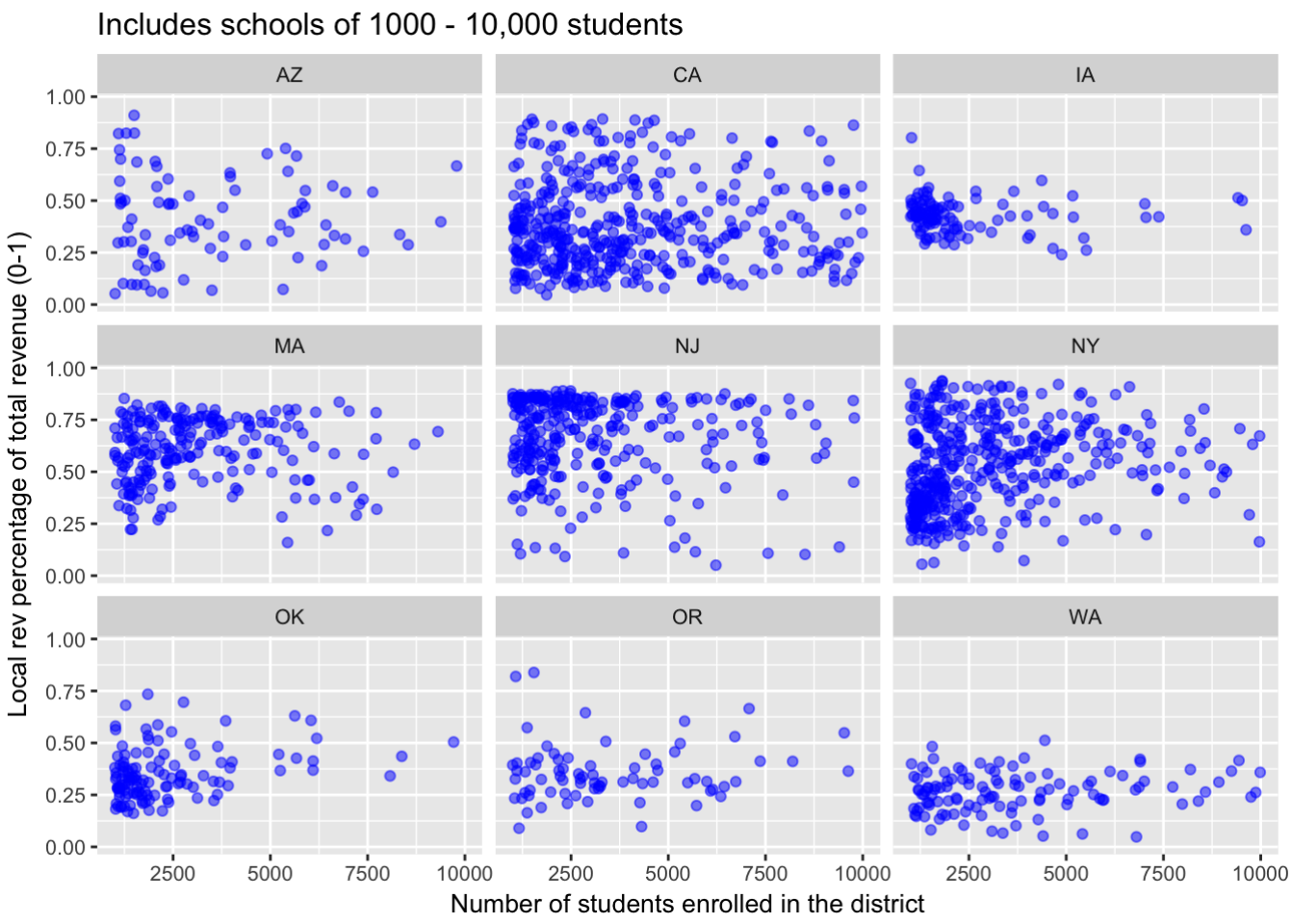


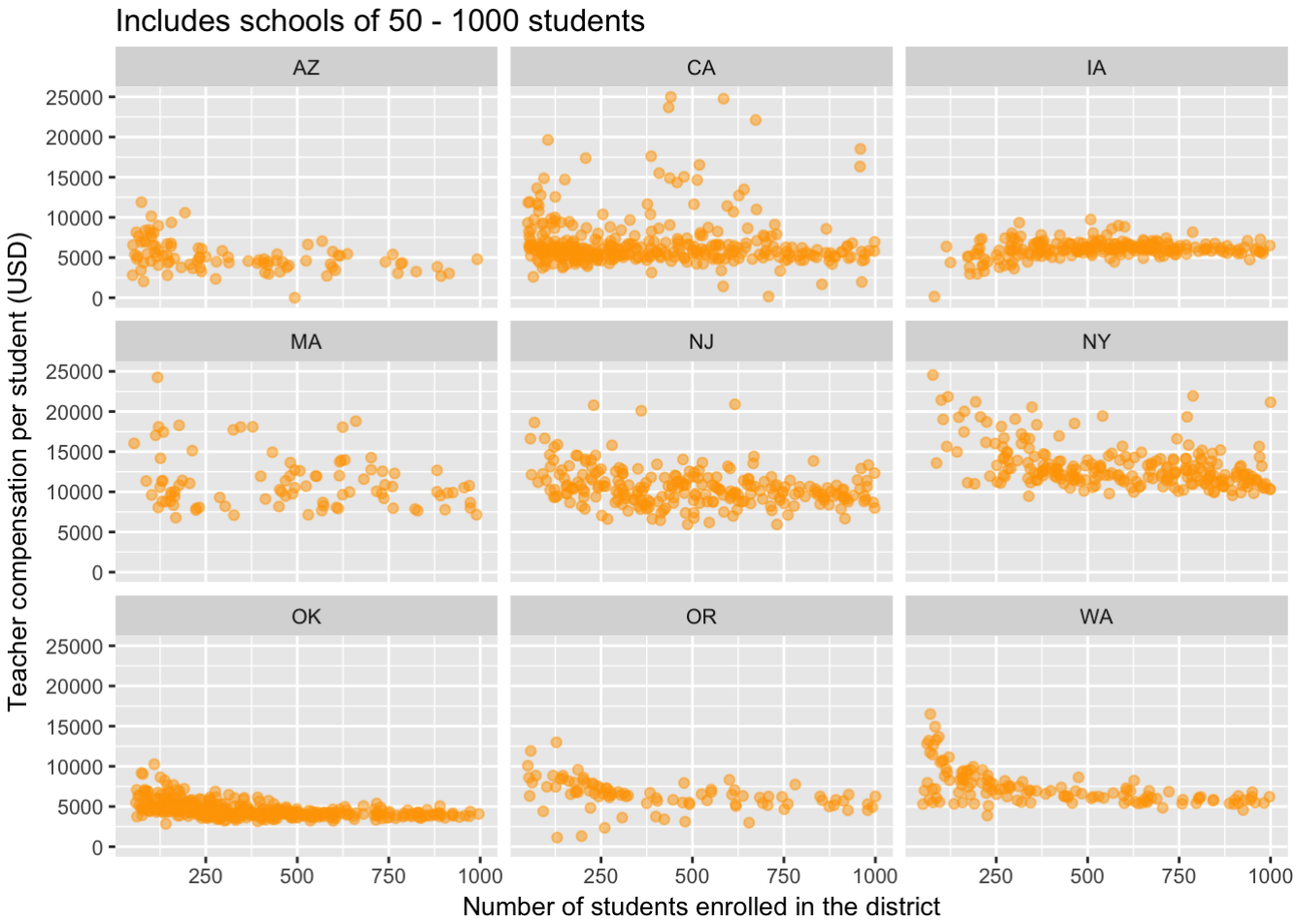
Had expected to maybe be able to see a trend, but did not see anything useful. Decided to look at a subset of data with a limited number of states.

Multivariate Plots Section

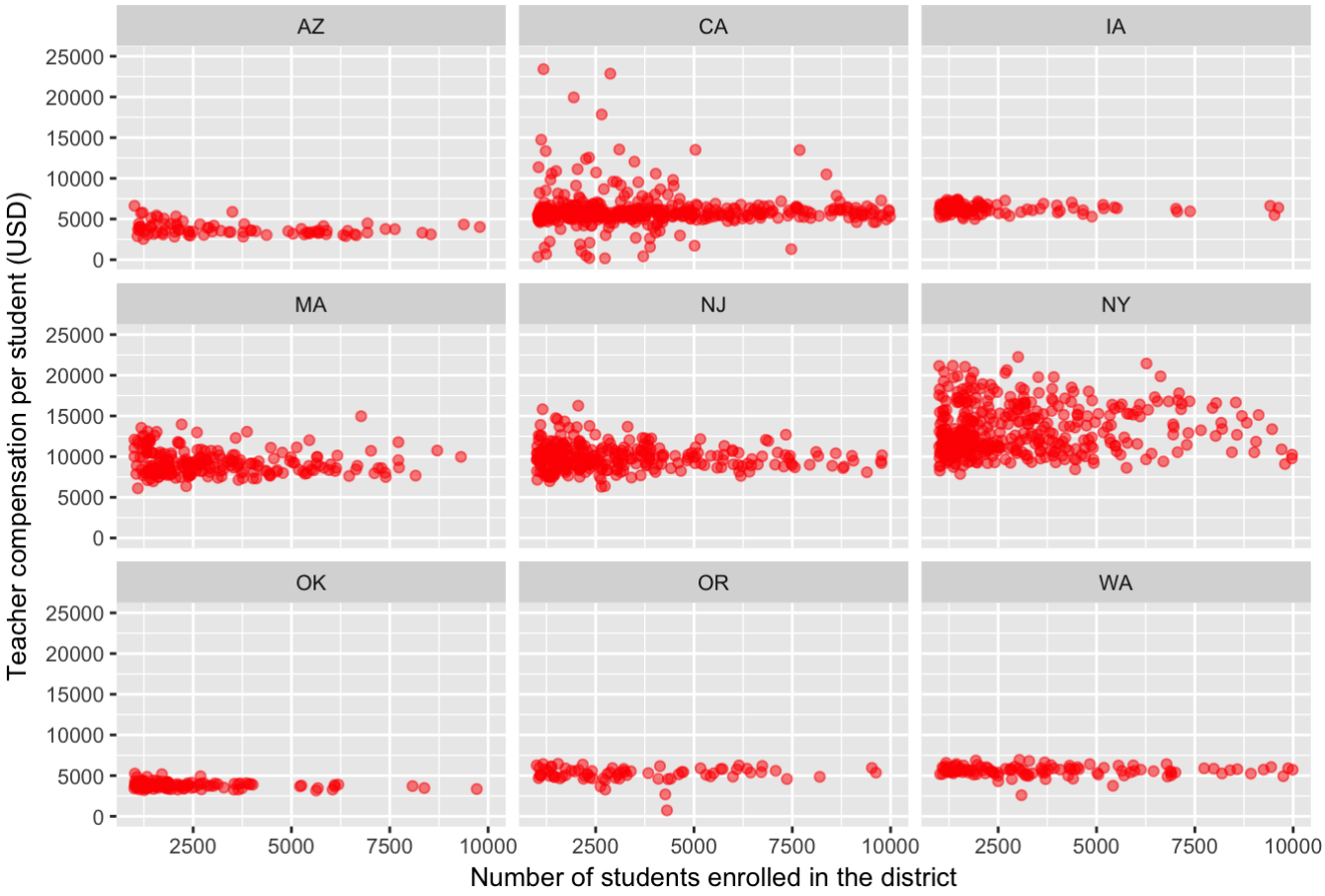
Plots Looking at data for a subset of states


Includes schools of 1000 - 10,000 students
AZ
CA
IA
MA
NJ
NY
OK
OR
WA
Total Expenses in thousands (USD)
100000
75000
50000
25000
0
2500
5000
7500
10000
Number of students enrolled in the district

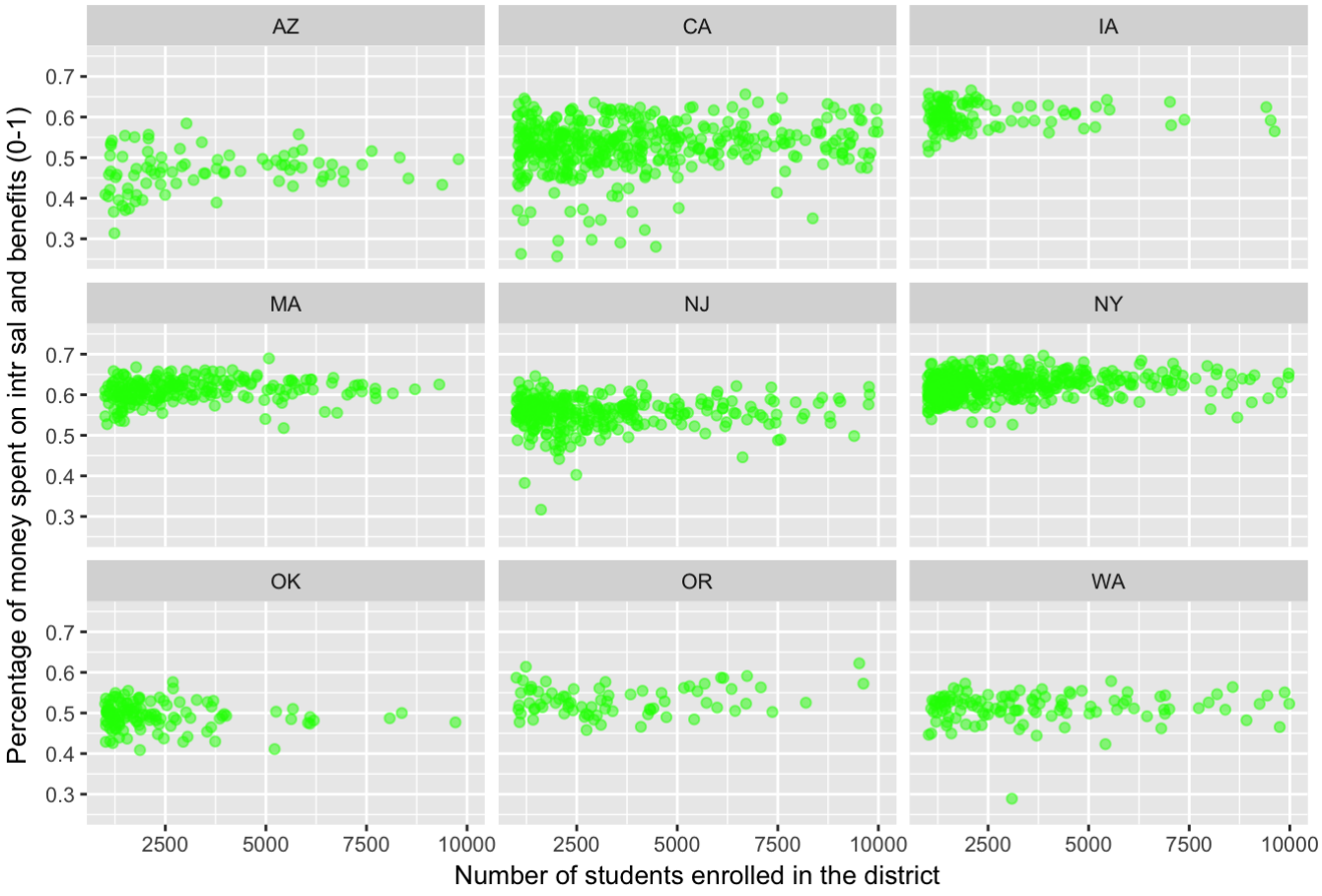

Includes schools of 1000 - 10,000 students
AZ
CA
IA
MA
NJ
NY
OK
OR
WA
Local rev percentage of total revenue (0-1)
1.00
0.75
0.50
0.25
0.00
2500
5000
7500
10000
Number of students enrolled in the district


Includes schools of 50 - 1000 students
AZ
CA
IA
MA
NJ
NY
OK
OR
WA
Teacher compensation per student (USD)
0
5000
10000
15000
20000
25000
250
500
750
1000
Number of students enrolled in the district

Includes schools of 1000 - 10,000 students



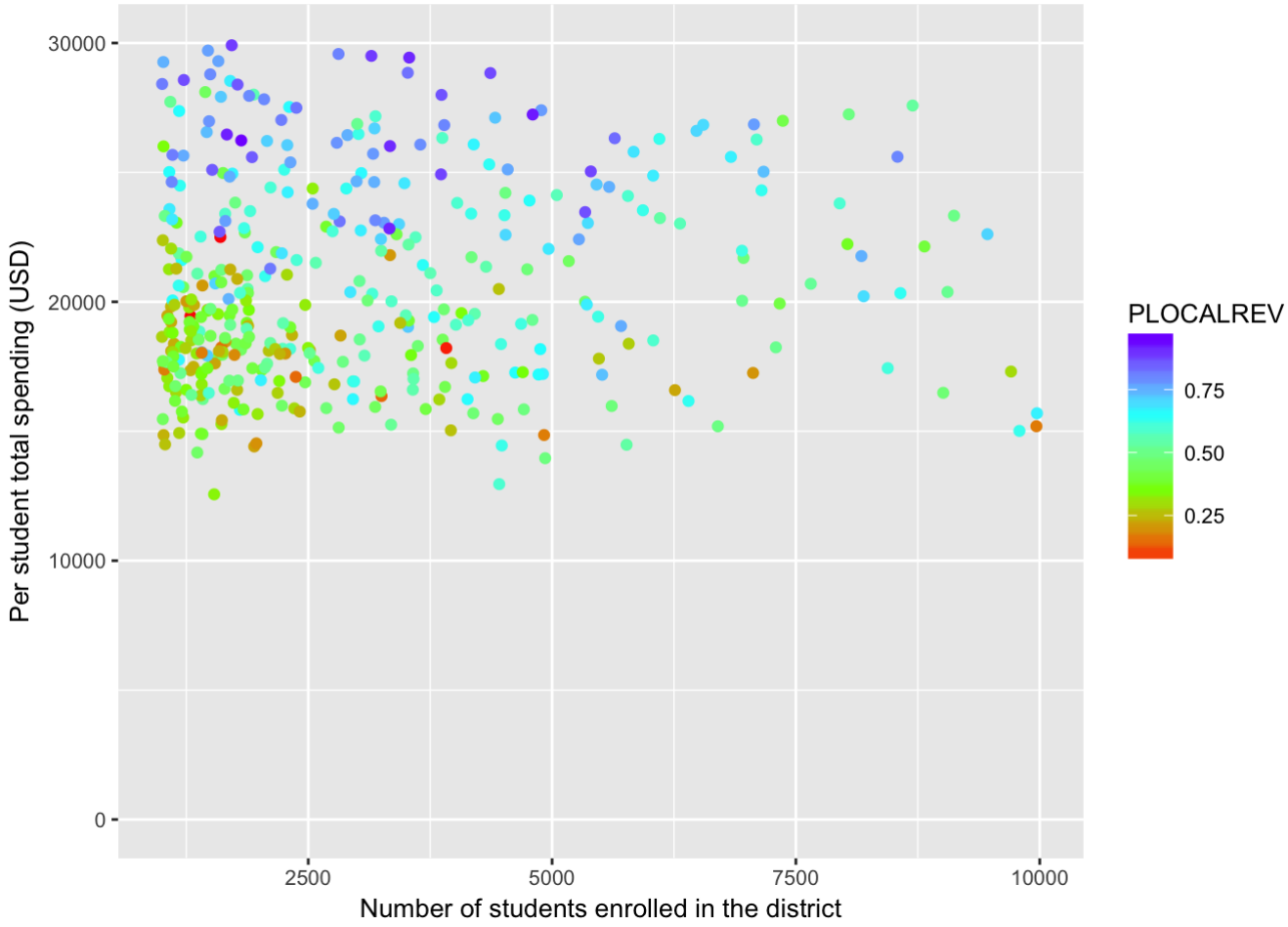
Includes schools of 1000 - 10,000 students



Starting to see some noticeable differences:

- New York spends a lot more on education than Oklahoma.
- California has a lot of variation in spending.
- It's looking pretty clear that MA, NY, and NJ are spending significantly more on teacher compensation than CA, AZ, OK.

Plot Looking at just 1 state (New York) a little deeper

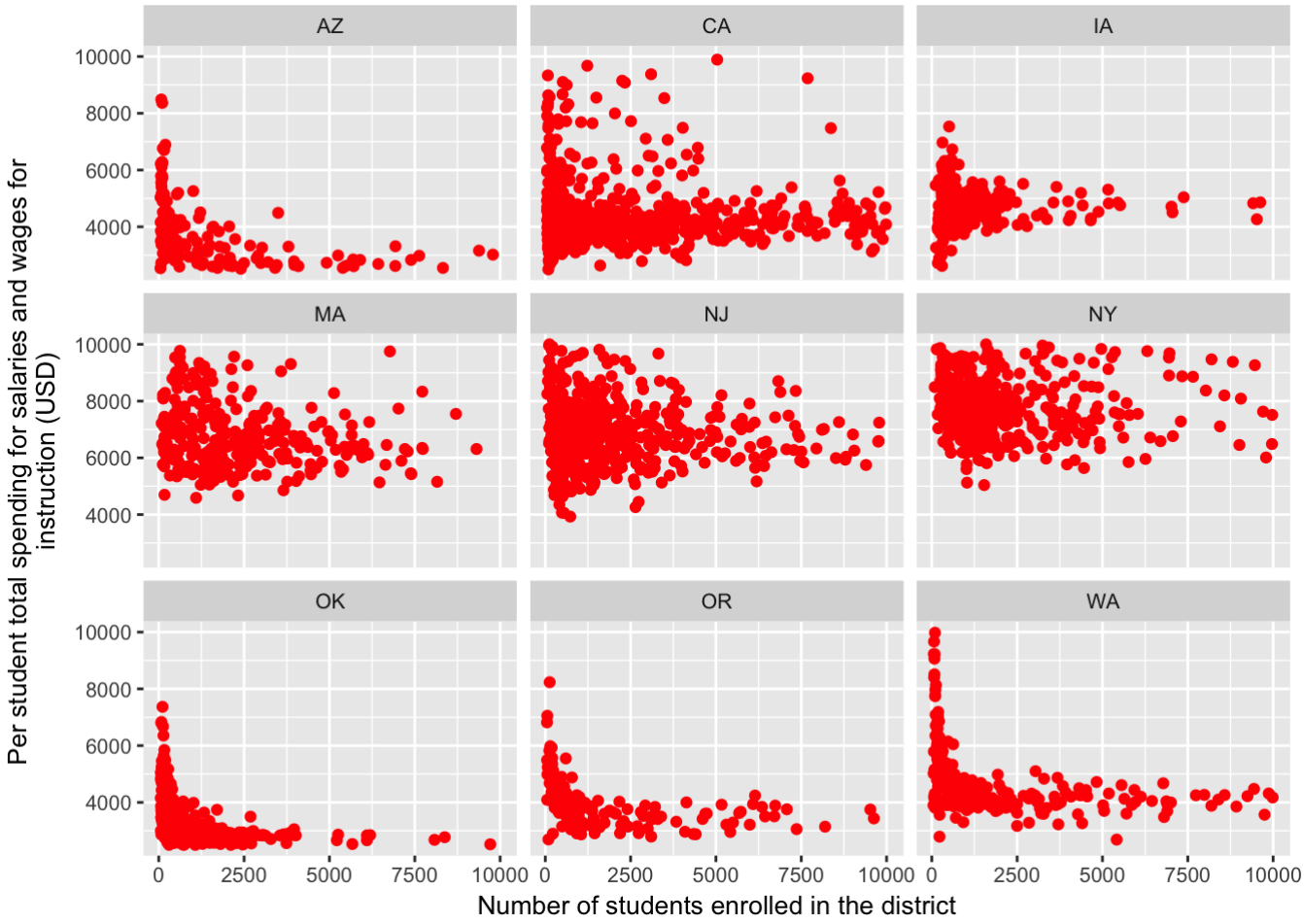


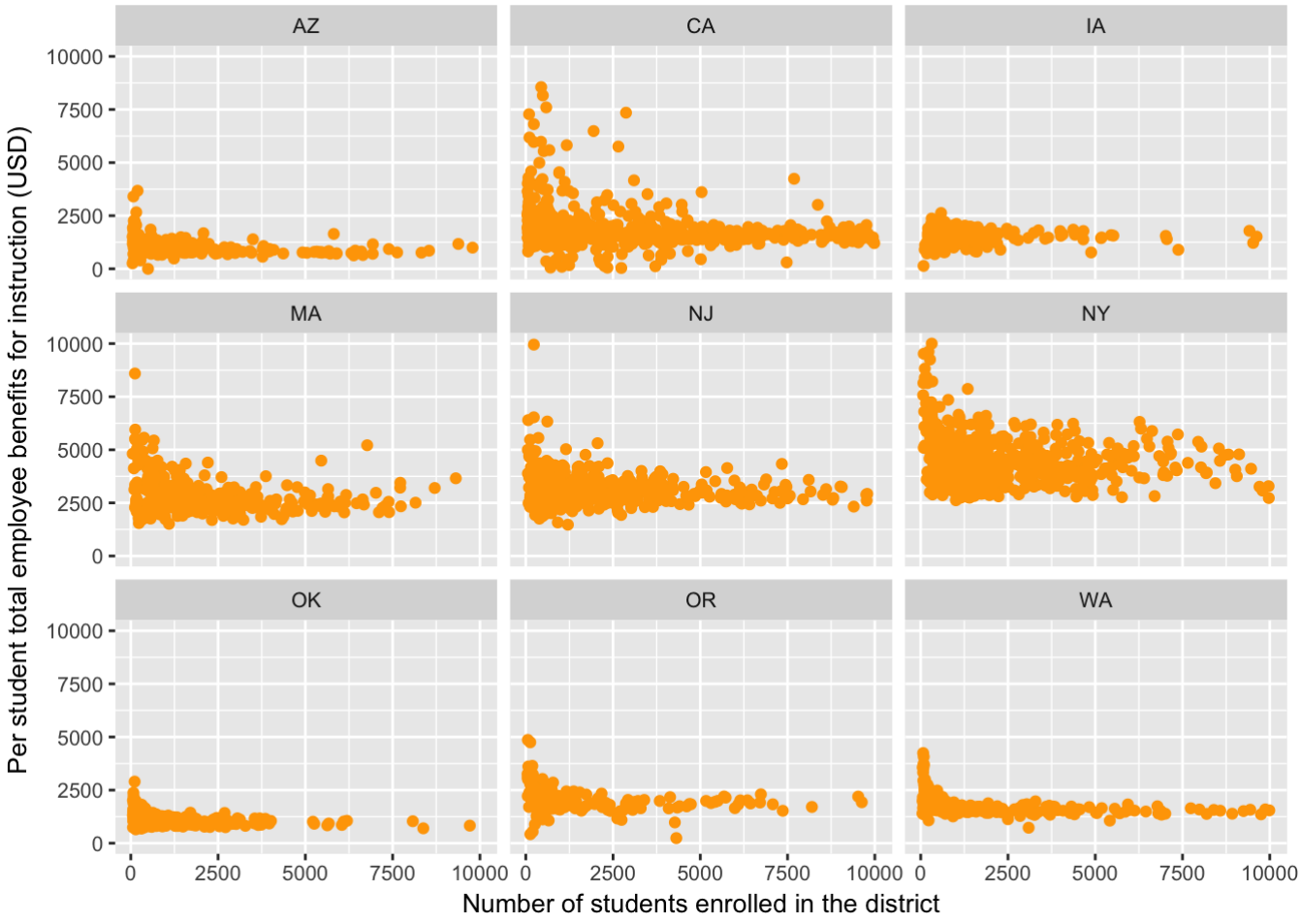
It does appear that total per student spending goes up when the percent of funding from local revenue for the school district goes up.

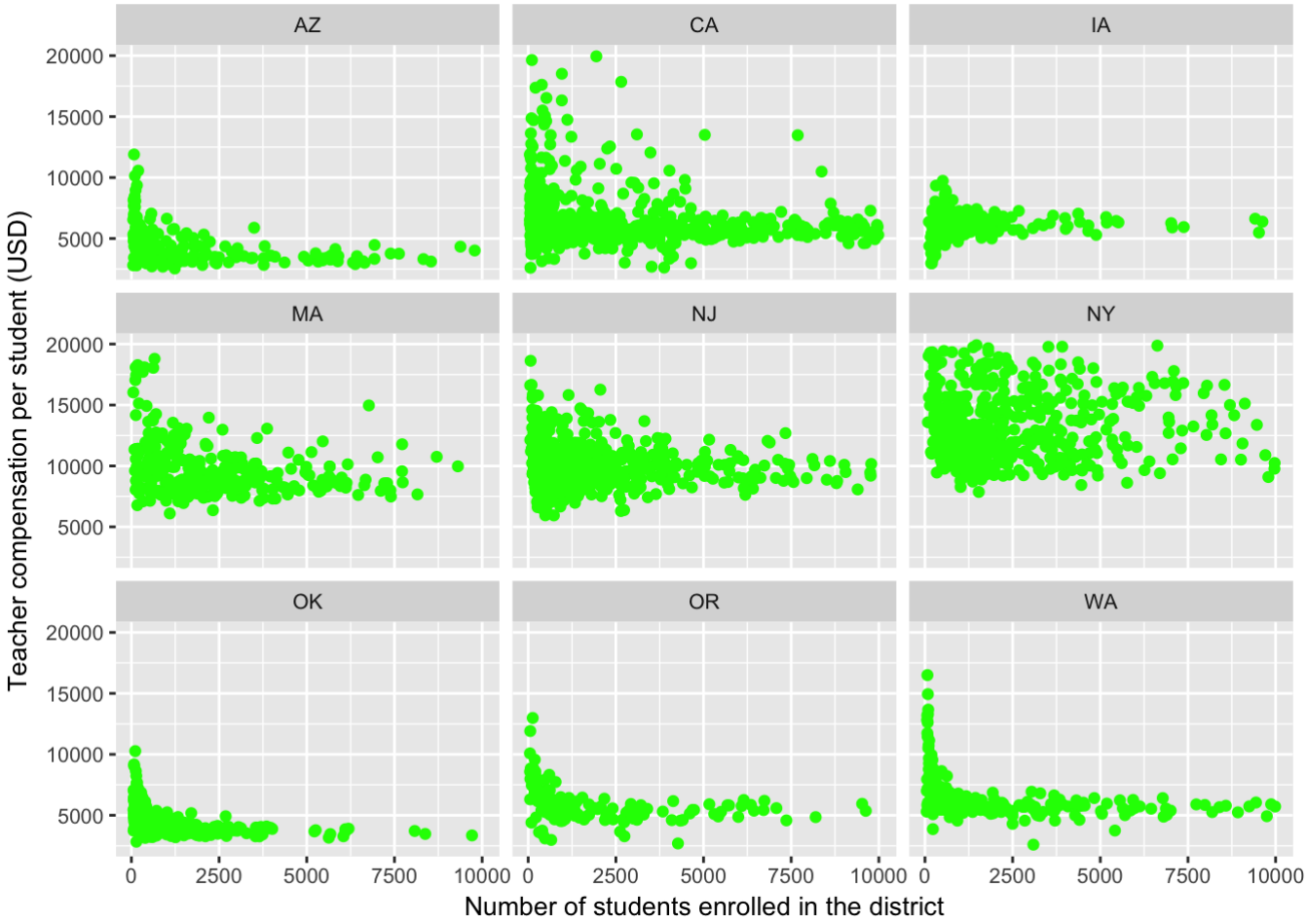
Narrowing down to instructional salary and benefits

- "PPISALWG" - Per student total spending for salaries and wages for instruction
- "PPIEMBEN" - Per student total employee benefits for instruction

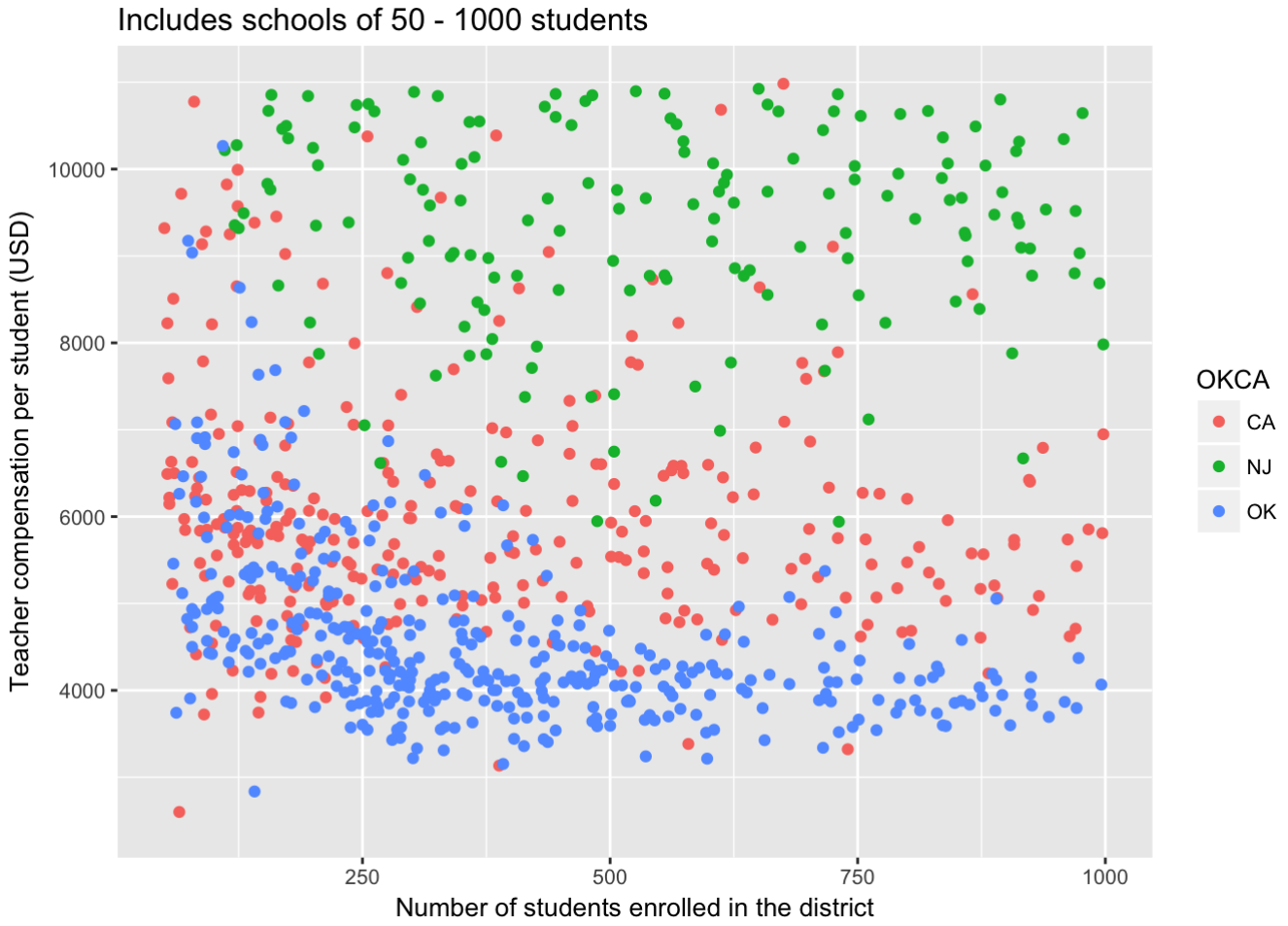
Looking at a few more factors.

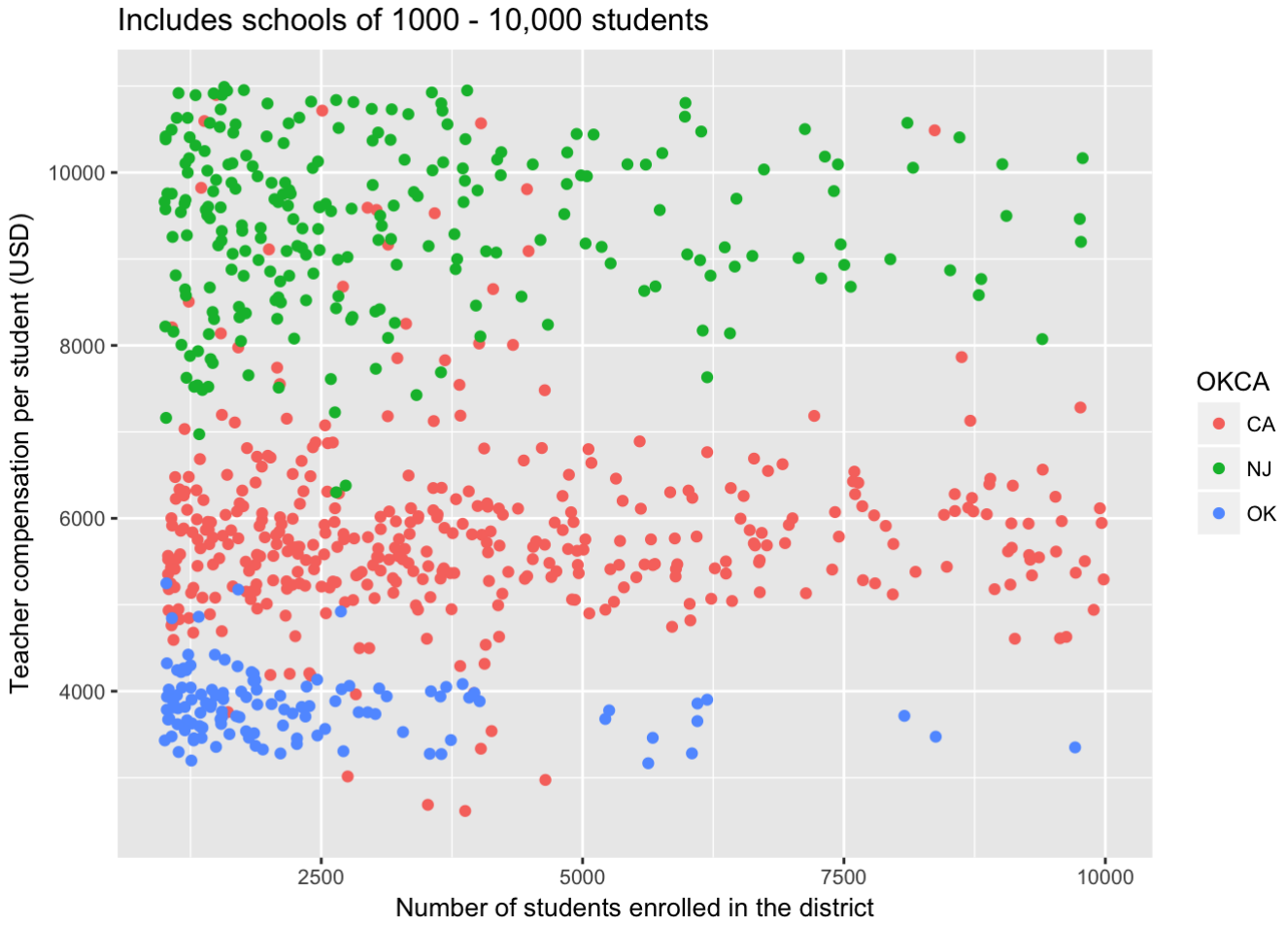

AZ
CA
IA
MA
NJ
NY
OK
OR
WA
10000
8000
6000
4000
0
2500
5000
7500
10000
Per student total spending for salaries and wages for instruction (USD)
Number of students enrolled in the district

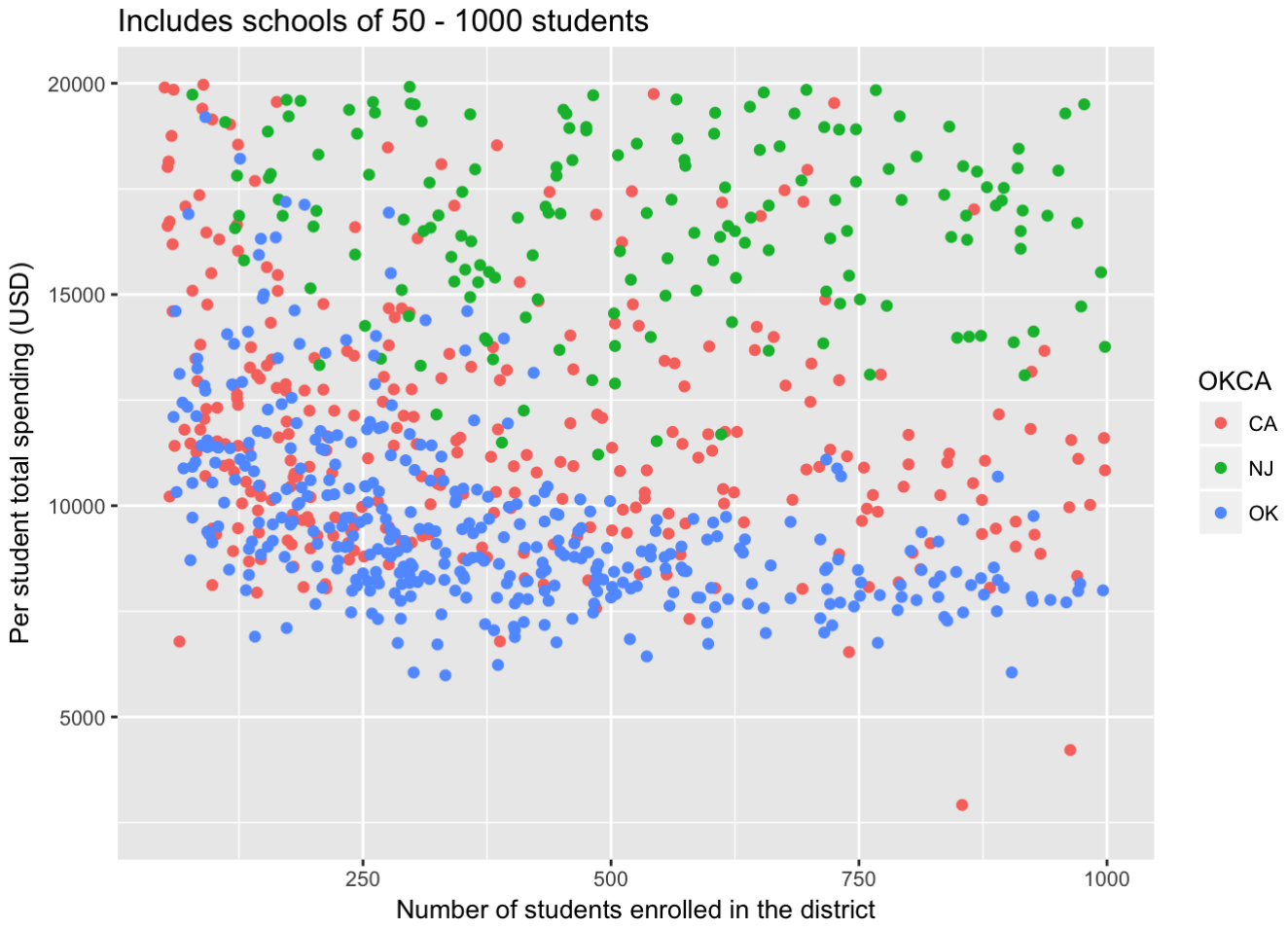

AZ
CA
IA
MA
NJ
NY
OK
OR
WA
10000
7500
5000
2500
0
Per student total employee benefits for instruction (USD)
Number of students enrolled in the district

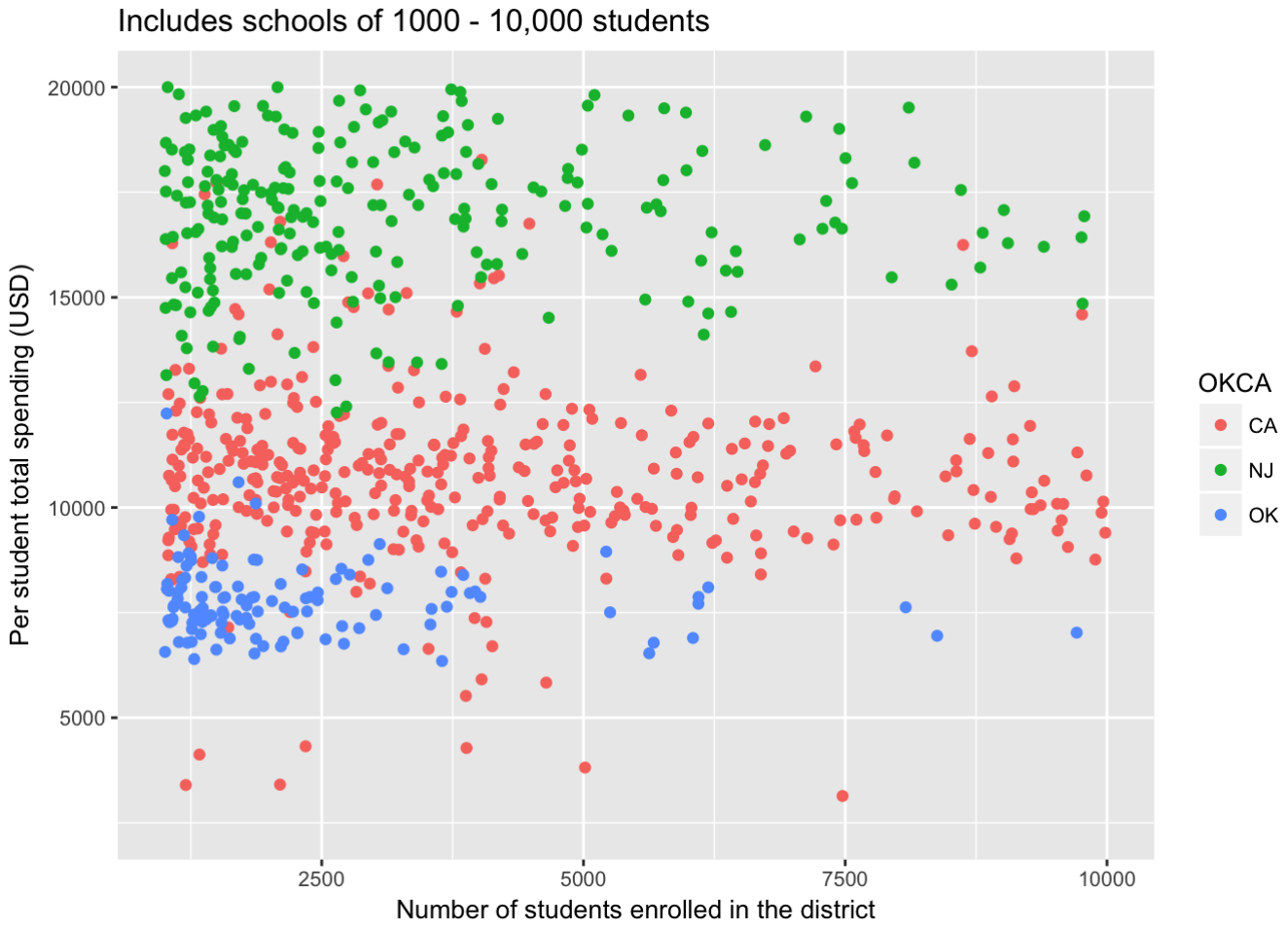


Narrowing down to just 3 states, NJ, CA, & OK


Includes schools of 50 - 1000 students
Teacher compensation per student (USD)
10000
8000
6000
4000
250
500
750
1000
Number of students enrolled in the district
OKCA
CA
NJ
OK


Includes schools of 1000 - 10,000 students
Teacher compensation per student (USD)
10000
8000
6000
4000
2500
5000
7500
10000
Number of students enrolled in the district
OKCA
CA
NJ
OK


Includes schools of 50 - 1000 students
Per student total spending (USD)
20000
15000
10000
5000
250
500
750
1000
Number of students enrolled in the district
OKCA
CA
NJ
OK


Includes schools of 1000 - 10,000 students
Per student total spending (USD)
20000
15000
10000
5000
2500
5000
7500
10000
Number of students enrolled in the district
OKCA
CA
NJ
OK

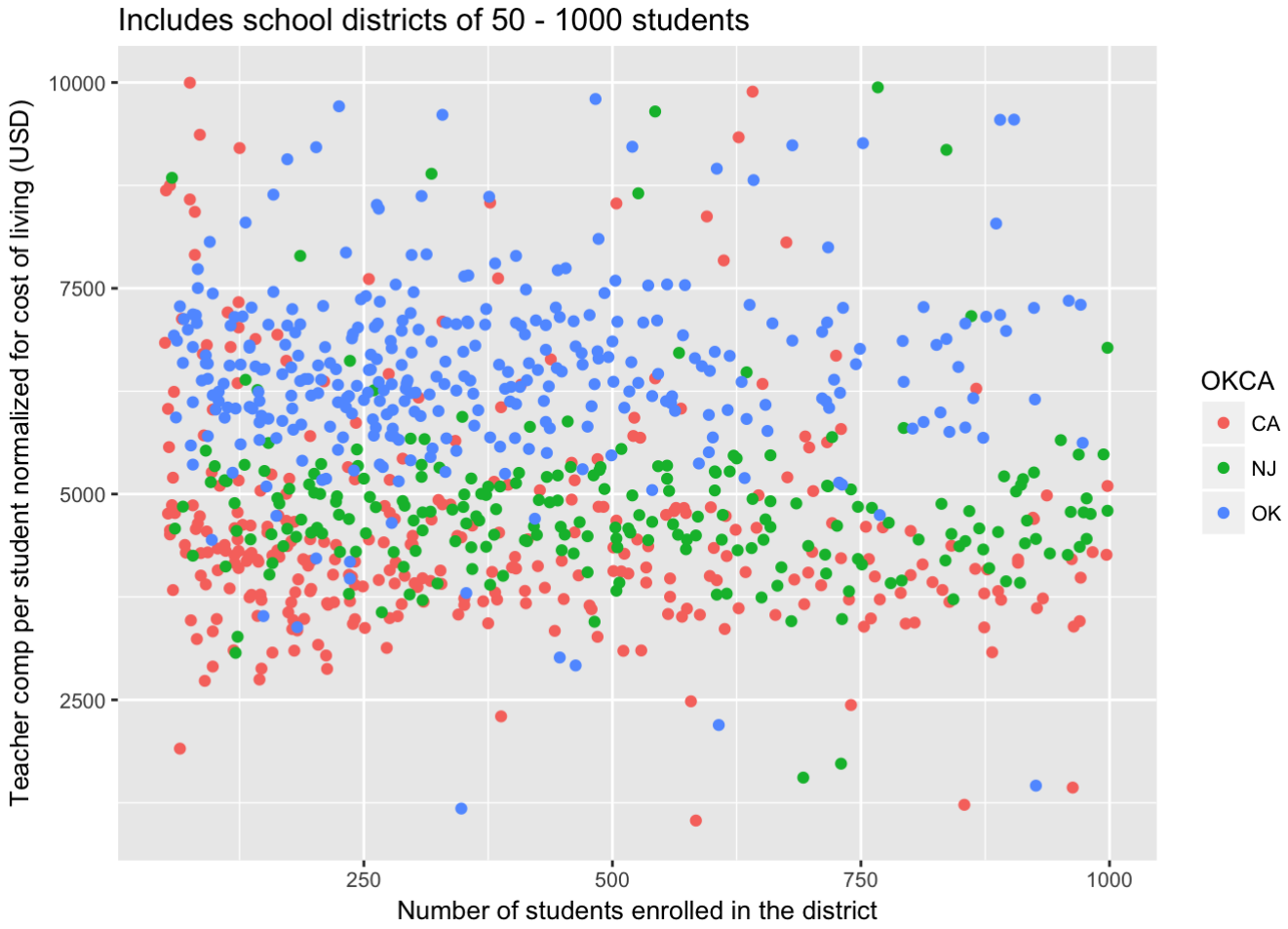
CA is way below NJ and not too far above OK although it's pretty well known that it's much more expensive to live in California. So what happens if a cost of living index is applied to the values?

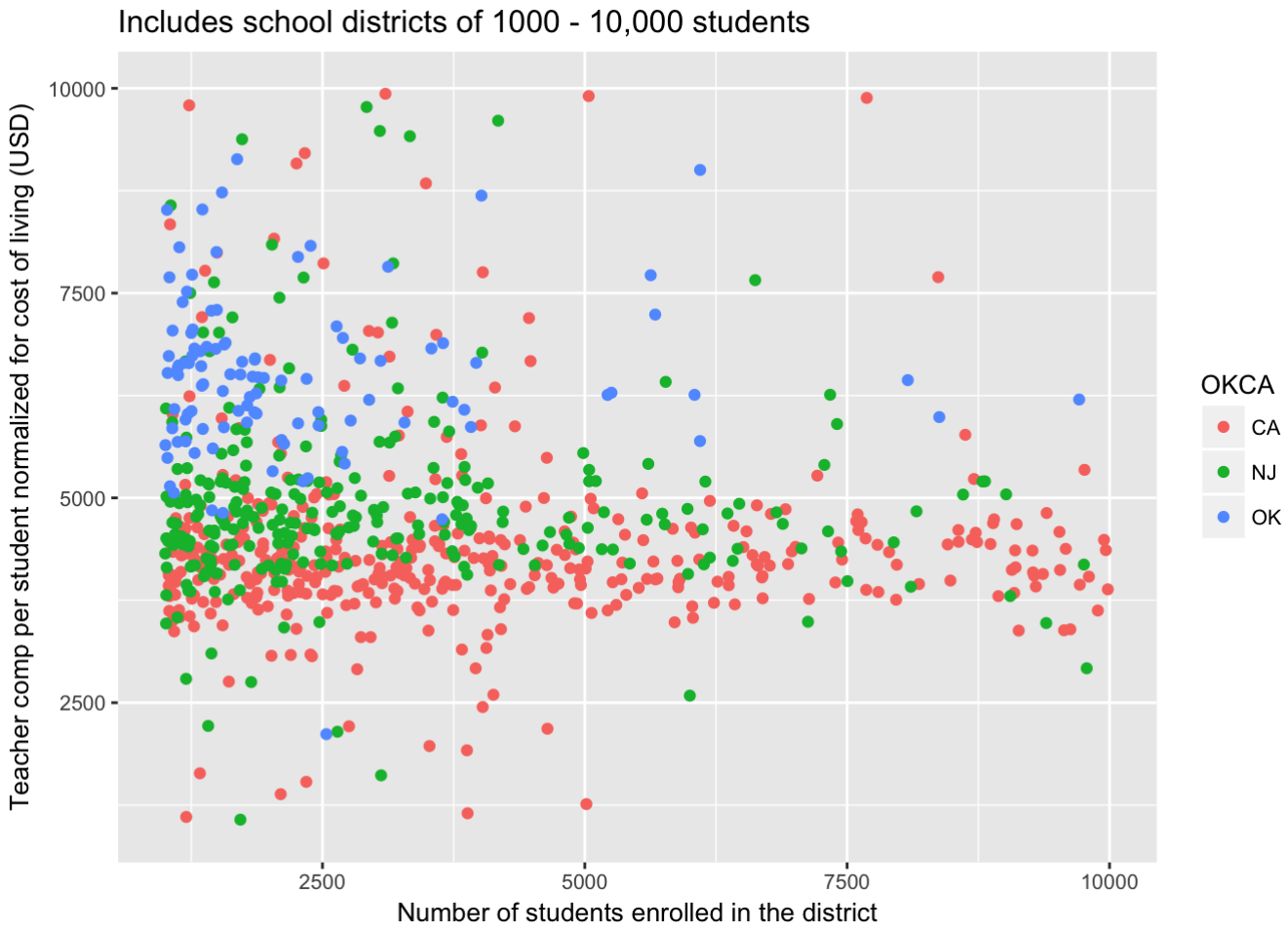
> ### Factoring in the cost of living for NJ, CA, & OK

Grabbed some data from 2017 (close enough) from: US Learning (https://www.uslearning.net/cost-of-living-by-state.html)

- CA - 136.3
- NJ - 121.2
- OK - 89.1

To normalize the data will take the total spent on salaries and benefits per pupil divided by the index/100 just to get a more dollar based amount.


Includes school districts of 50 - 1000 students
Teacher comp per student normalized for cost of living (USD)
10000
7500
5000
2500
250
500
750
1000
Number of students enrolled in the district
OKCA
CA
NJ
OK


Includes school districts of 1000 - 10,000 students
Teacher comp per student normalized for cost of living (USD)
10000
7500
5000
2500
2500
5000
7500
10000
Number of students enrolled in the district
OKCA
CA
NJ
OK

Based on the graphs above accounting for cost of living, it seems Oklahoma teachers get paid more than either New jersey or California.

Factoring in the class size for NJ, CA, & OK

In theory class size would have impact the actual salaries. If compensation is the per student money spent on instructional salaries and benefits than:

- Higher class size = higher salary
- Lower class size = lower salary

If the per student instructional compensation were the same.

(Note: This doesn't examine quality of education based or difficulty of the teaching role as class size increases.)

From:

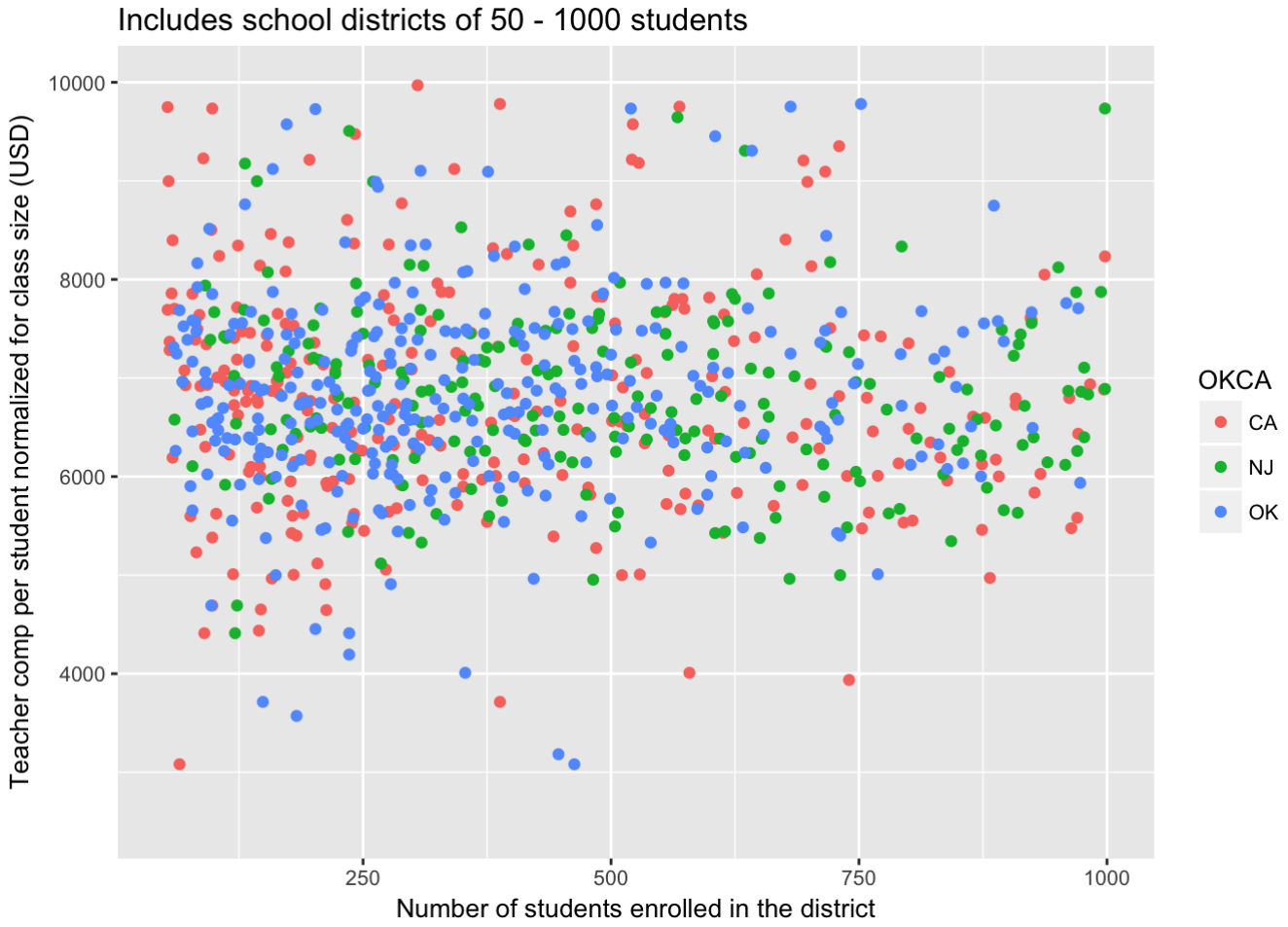
National Center for Education Statistics (https://nces.ed.gov/programs/digest/d16/tables/dt16_209.30.asp)

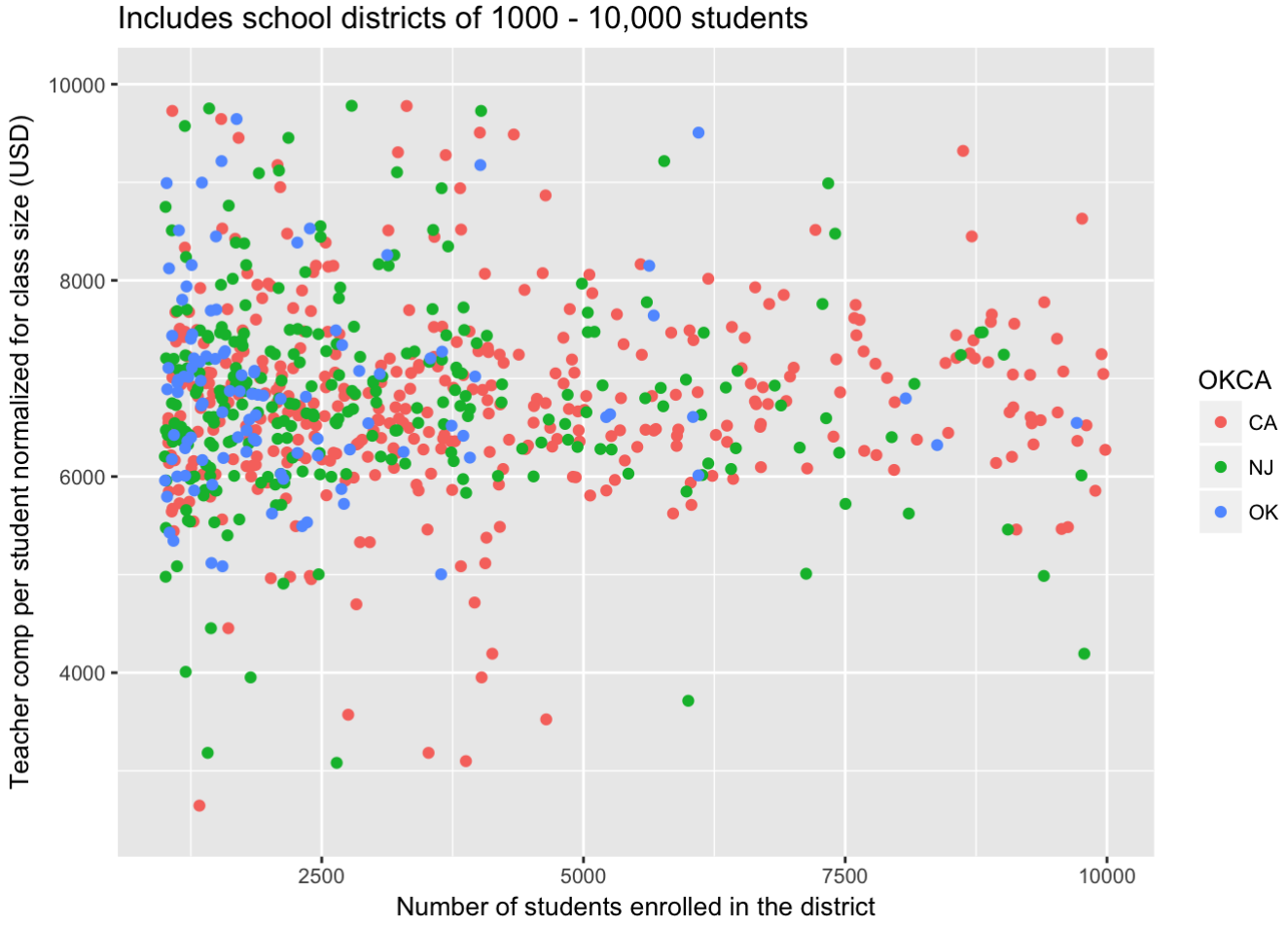
This information is from 2012 so things may be a little old. Four years is a long time. But unless the states increased or decreased disproportionately to other states the results should be similar.

Decide to use the numbers provided in the following way:

(Average class size by level of instruction Elementary + Average class size by level of instruction Secondary) divided by 2. Then normalize that using the average US class size

- CA - (25.0 + 32.0)/2
- NJ - (18.5 + 23.9)/2
- OK - (20.7 + 23.7)/2
- US - (21.3 + 26.8)/2

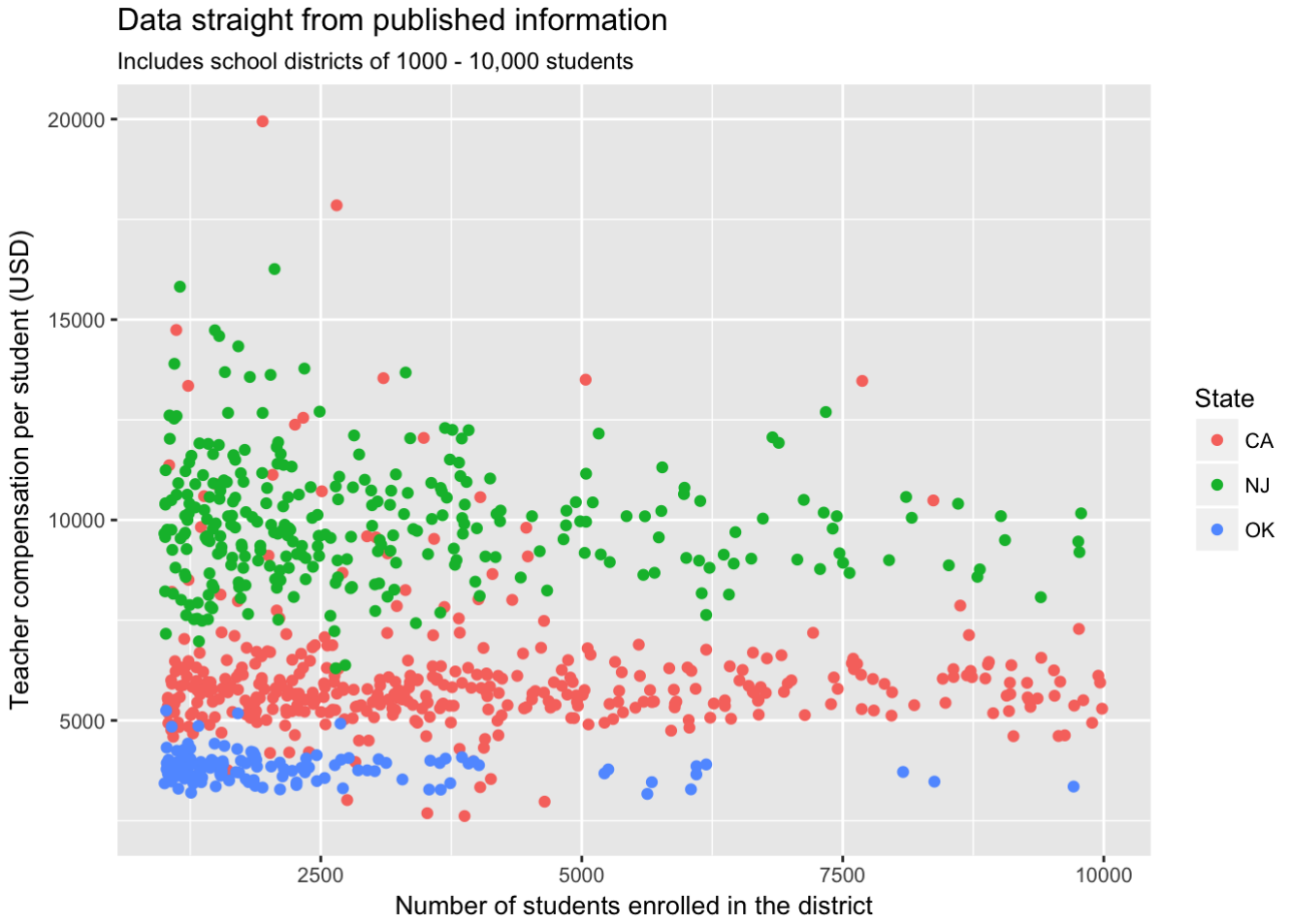

Includes school districts of 50 - 1000 students
Teacher comp per student normalized for class size (USD)
10000
8000
6000
4000
250
500
750
1000
Number of students enrolled in the district
OKCA
CA
NJ
OK


Includes school districts of 1000 - 10,000 students
Teacher comp per student normalized for class size (USD)
10000
8000
6000
4000
2500
5000
7500
10000
Number of students enrolled in the district
OKCA
CA
NJ
OK

Results: Once you add in cost of living and class size there while there is a lot of variation across states, there doesn't appear to be a big difference between California, New Jersey and Oklahoma in terms of Instructional Salaries and benefits.

Final Plots and Summary

Plot One



Description One

Using the cleaned up data from:

2016 Public Elementary-Secondary Education Finance Data (https://www.census.gov/data/tables/2016/econ/school-finances/secondary-education-finance.html)

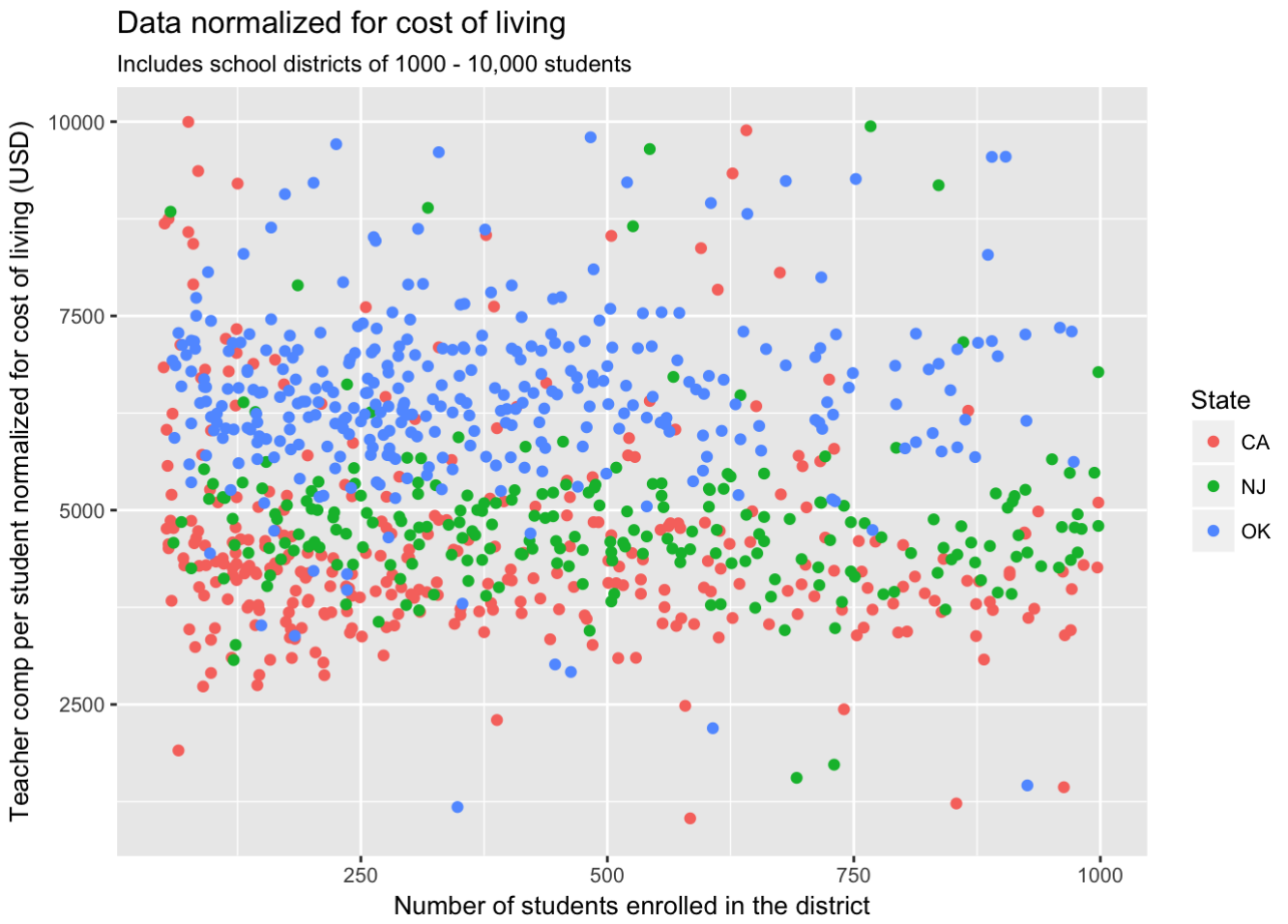
to graph the per student spend on instructional salary and benefits for California, New Jersey and Oklahoma there is a clear indication that the per student spending in Oklahoma is far less than California or New Jersey for school districts between 1000 and 10,000 students.

It also appears that the variation of what is spent per student is far greater in California and New Jersey than in Oklahoma.

Using this graph alone:

- Oklahoma teachers are justified in how upset they are regarding salary and benefits provided.

Plot Two



Description Two

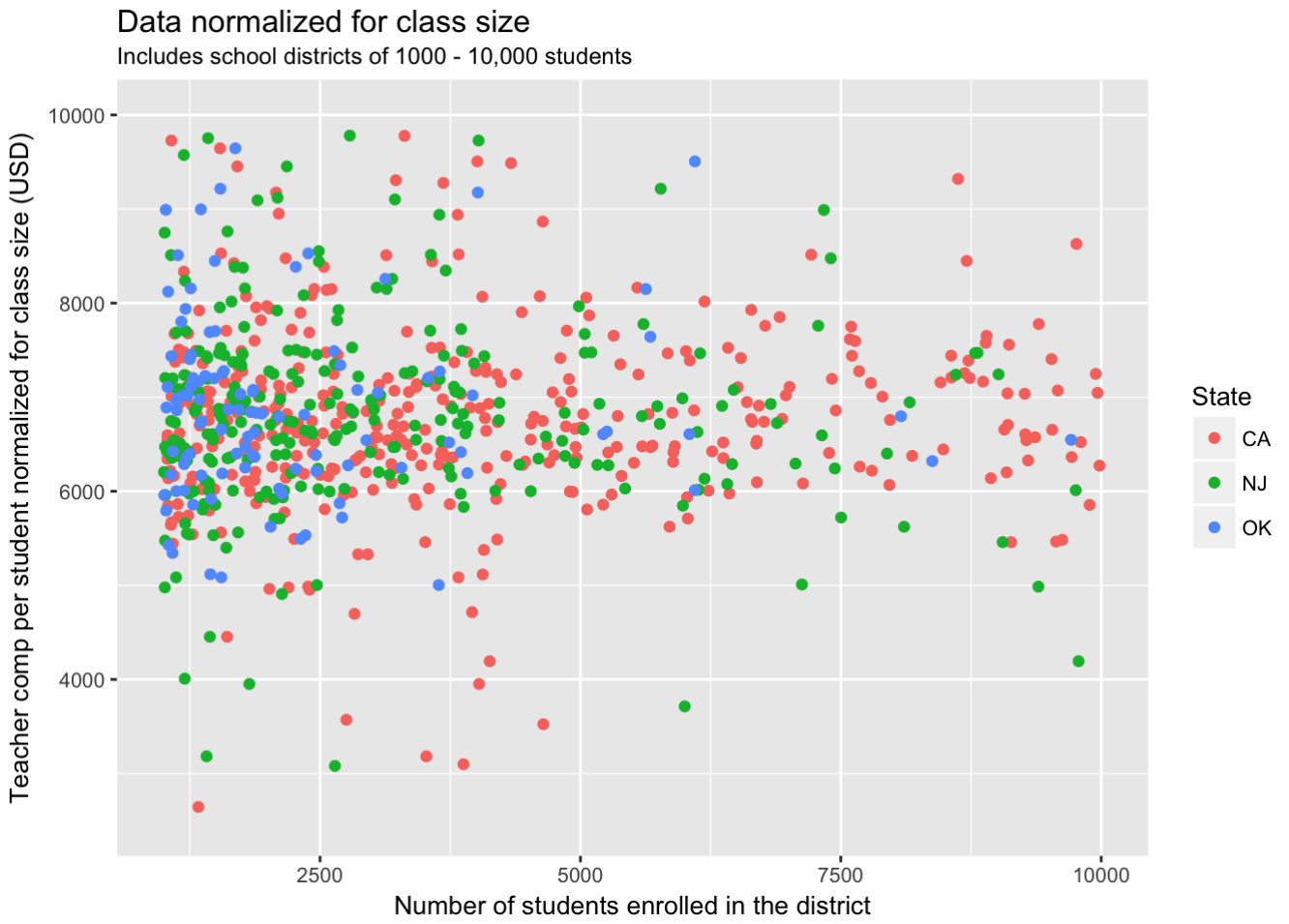
Grabbed some data from 2017 (close enough) from: US Learning (https://www.uslearning.net/cost-of-living-by-state.html)

Normalized for cost of living:

- CA - 136.3
- NJ - 121.2
- OK - 89.1

The schools switch places. California falls to the bottom. New Jersey ends up in the middle and it appears that Oklahoma teachers are the best paid teachers. Also the variation in spending per student on instructional salaries and benefits for Oklahoma appears larger once the cost of living index is applied since a dollar in Oklahoma goes farther than it does in California or New Jersey.

Plot Three



Description Three

Because the data on plot 1 and plot 2 is NOT strictly reflective of how much a particular teacher gets paid in any of the states it is necessary to use a normalizer to try to compute what the relationships might look like.

Using class size information:

From:
National Center for Education Statistics (https://nces.ed.gov/programs/digest/d16/tables/dt16_209.30.asp)

This information is from 2012 so things may be a little old. Four years is a long time

Decide to use the numbers provided in the following way:

Average class size for teachers in self-contained classes + Average class size for teachers in departmentalized instruction and divide by 2. Then normalize that using the average US class size.

- CA - (25.0 + 32.0)/2
- NJ - (18.5 + 23.9)/2
- OK - (20.7 + 23.7)/2
- US - (21.3 + 26.8)/2

and normalizing the data in plot 2 created the plot above.

Now the data from all three states starts to overlap indicating that maybe there is less of a difference on what is spent on teacher salary and benefits by state than what the original data showed alone.

Note that the final numbers on the y axis are not real numbers for any of the states, but an indication of spending relative to each other.

Reflection

What were some of the struggles?

- I thought I had a really good dataset to start out with, but the more I looked at it, the more I realized that I needed to clean it up first.
- I had an expectation that different things about educational spending would pop out right away, but actually there is a lot of variation by school district, so it was harder to see trends than I expected.
- I hadn't been using Rstudio previously in a way that could be used by knit so I had to learn a lot of the details of Rstudio while working on the project.

What went well?

- Using the full dataset and a few different graphing techniques I was able to narrow down to an area of interest.
- As I found some bad data in the dataset, I started to realize that I got lucky that the dataset was as good as it was.

What was surprising?

- That government published data could have so much detailed information, but still have errors.
- That the state of Hawaii has only one school district.
- That NYC has almost a million students and a budget of 20+ billion dollars.
- That there is so much variation on what is spent per student on across the country and across a state.
- That there are so many individual school districts

What further investigations could be done?

- Could try to find more recent class size information and see if the results are the same.
- Could could create data on a map.
- Could do a more detailed analysis of just one state and how and why spending varies within a state.