Big Data Course Project

Registrar Data Analysis

Tianzhixi Yin, Damian Stansbury, Dongyang Kuang

May, 3, 2014

In the beginning...

Our Goals

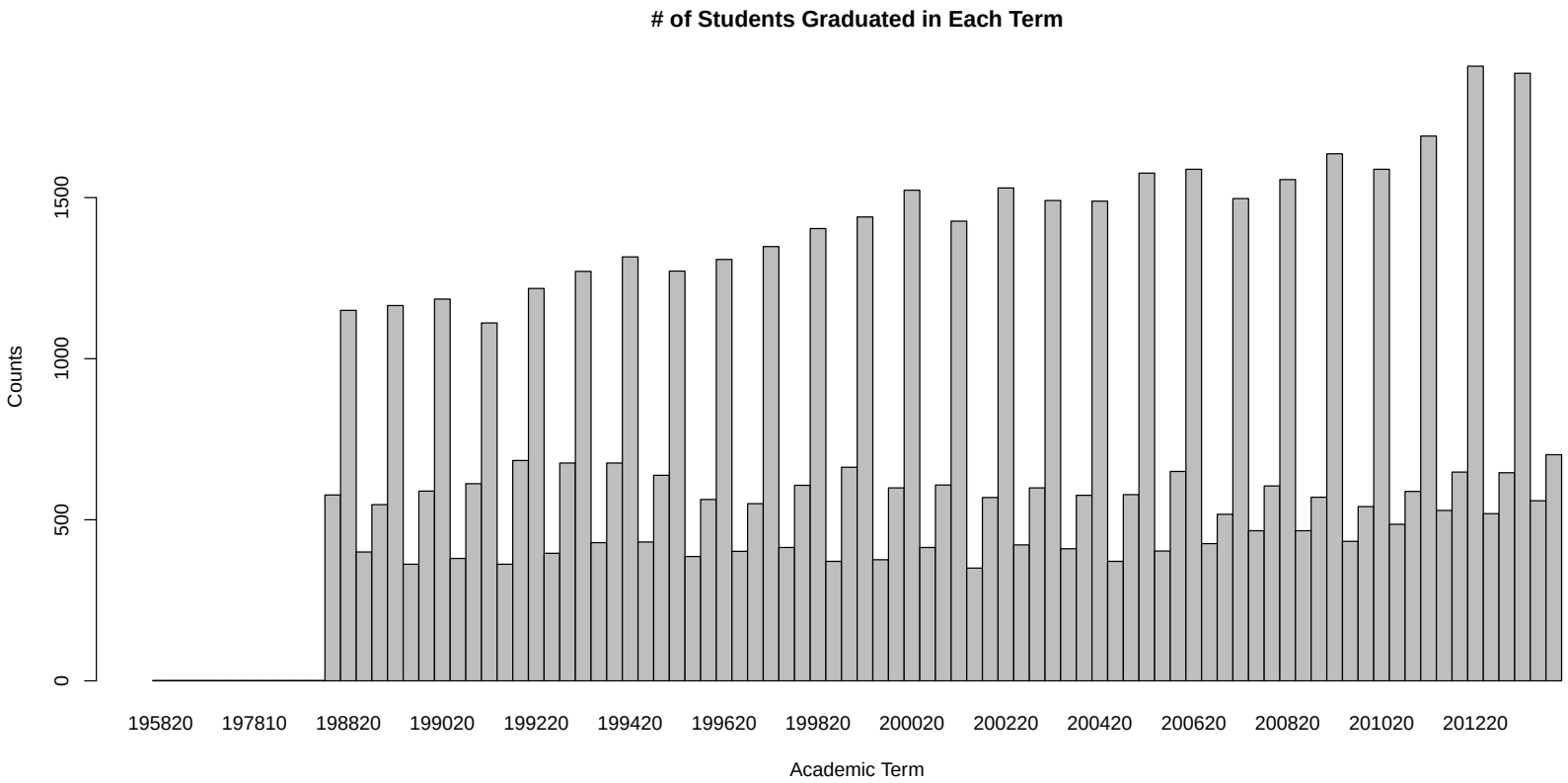
- Dig out meaningful information from the dataset.
- Make it easy for others to utilize our methods.

What is the most important issue in Data Analysis?

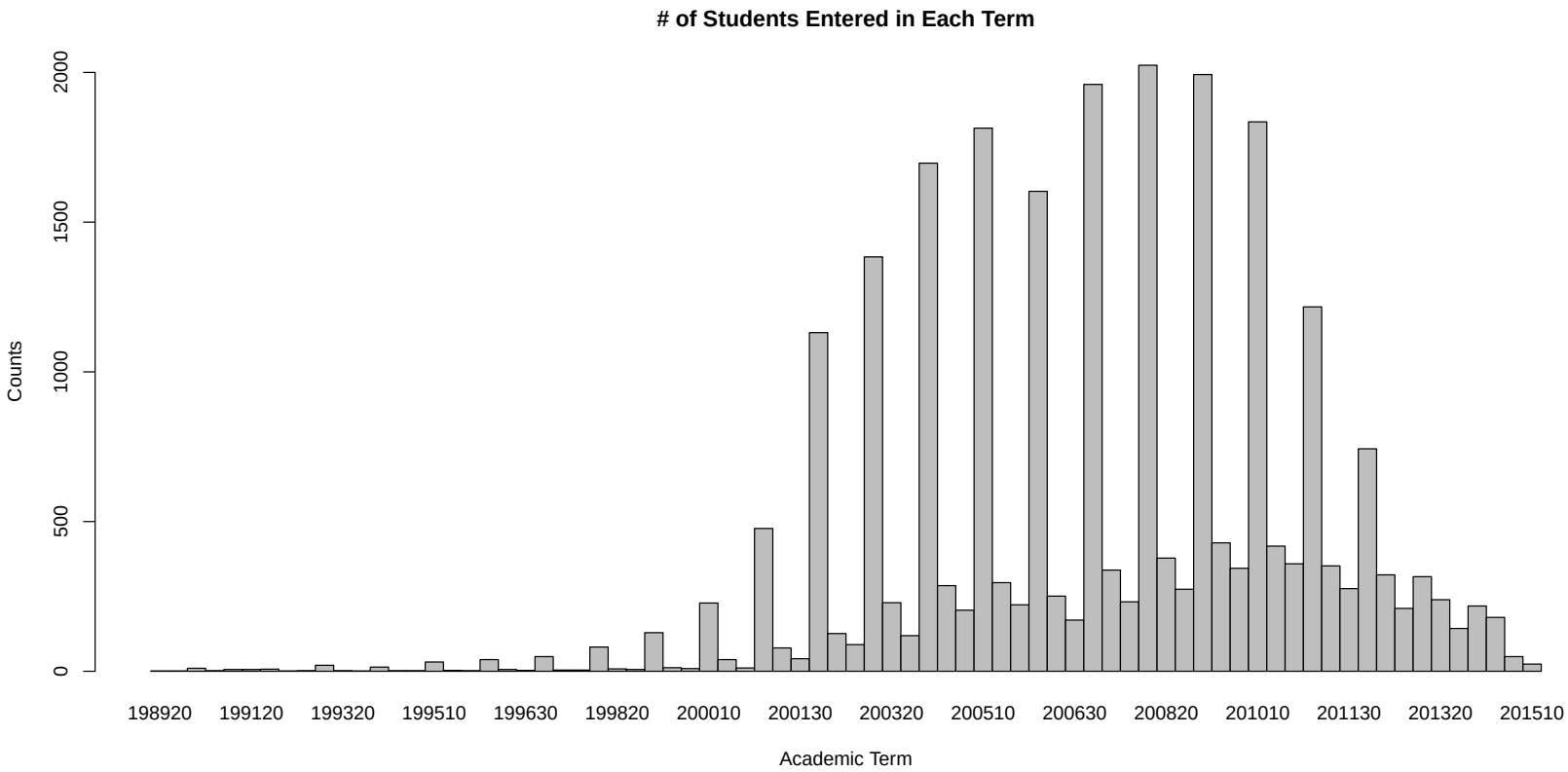
Our Major Questions

- What are the factors influencing students' performance?
- What geographic areas and high schools are sending good students to UW?
- Is Financial Aid really helping the students?

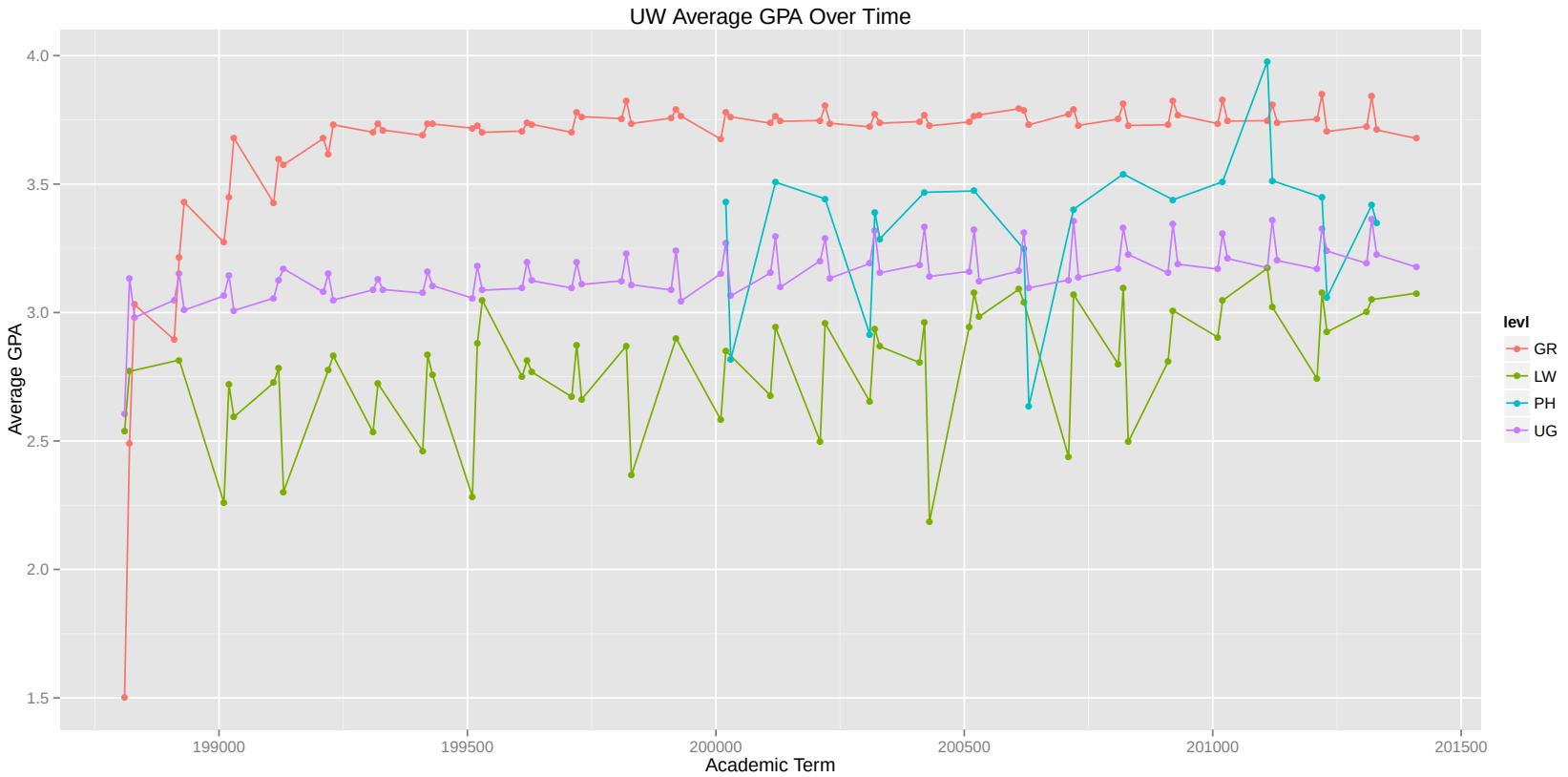
How many students have graduated from UW?



Are more and more students coming to UW?



Has GPA changed over time?



What truncated regression model thinks are most important

Predictors	Log-Likelihood
SAT Total	-2104

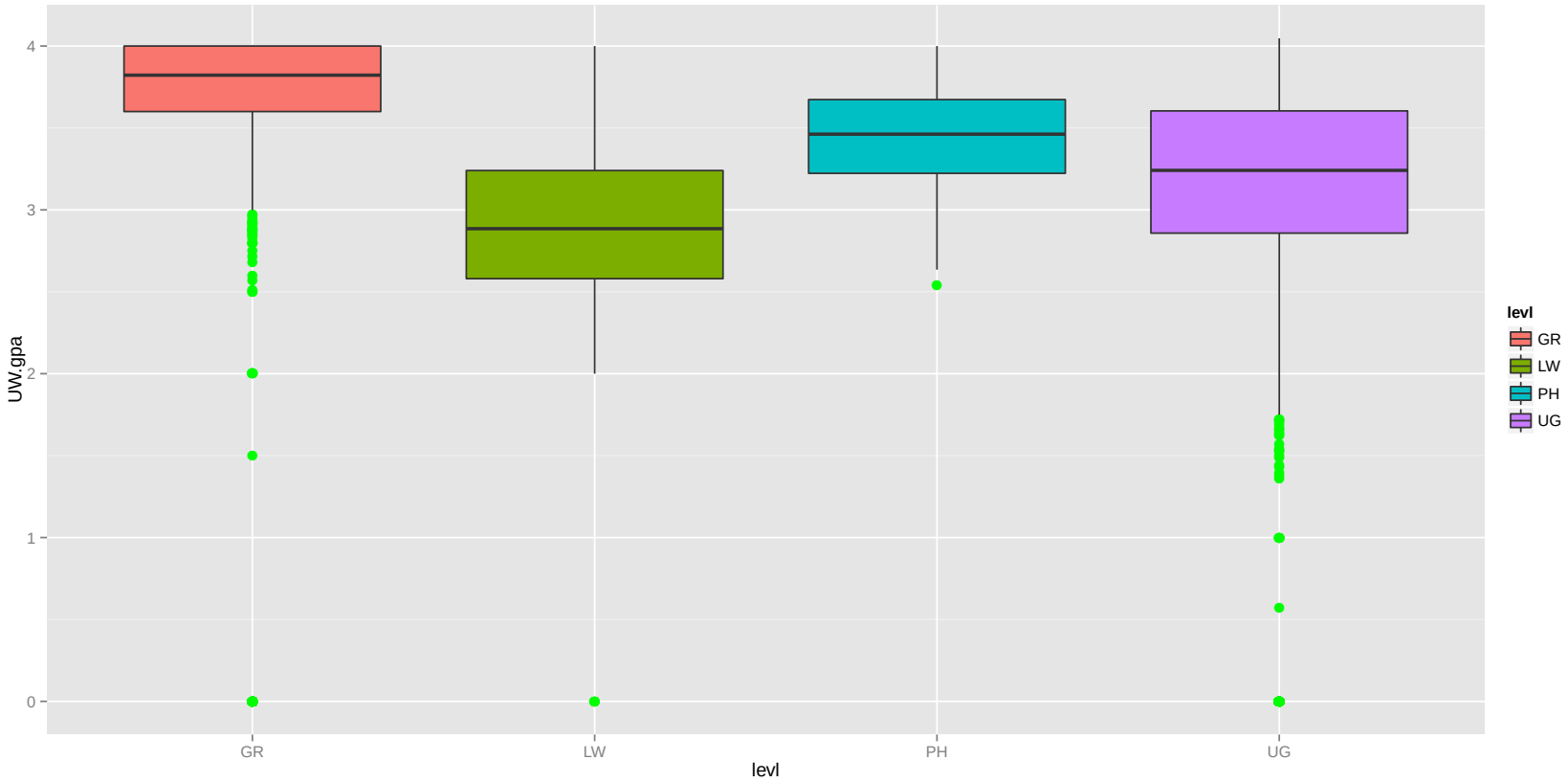
What truncated regression model thinks are most important

Predictors	Log-Likelihood
SAT Total	-2104
Highschool GPA	-5891

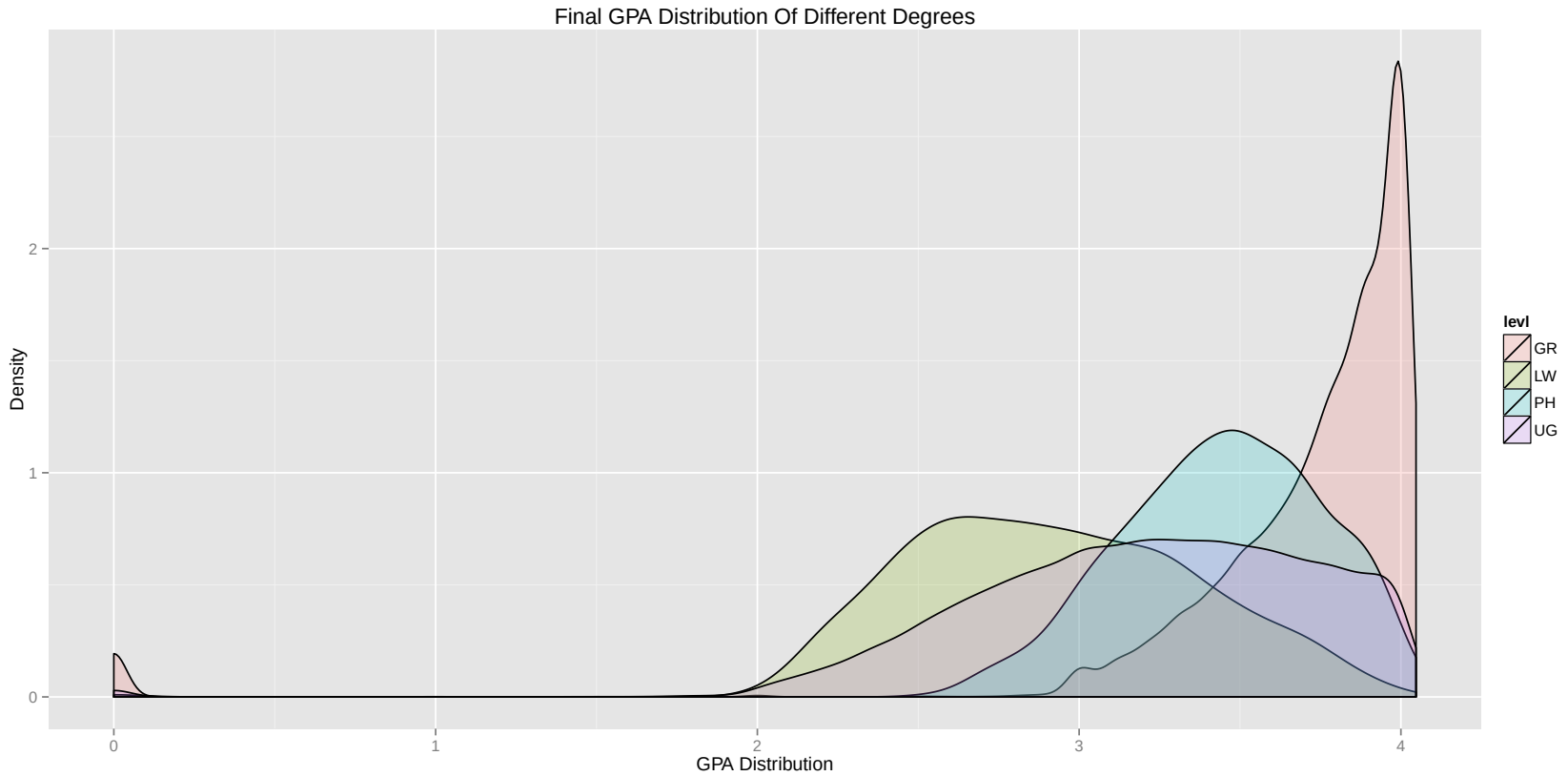
What truncated regression model thinks are most important

Predictors	Log-Likelihood
SAT Total	-2104
Highschool GPA	-5891
ACT Composite	-13021
Major	-36215
Level	-37922
College	-41629
Marital Status	-42596
Sex	-42334
Highschool State	-43207
Race	-43110
Veteran	-43354

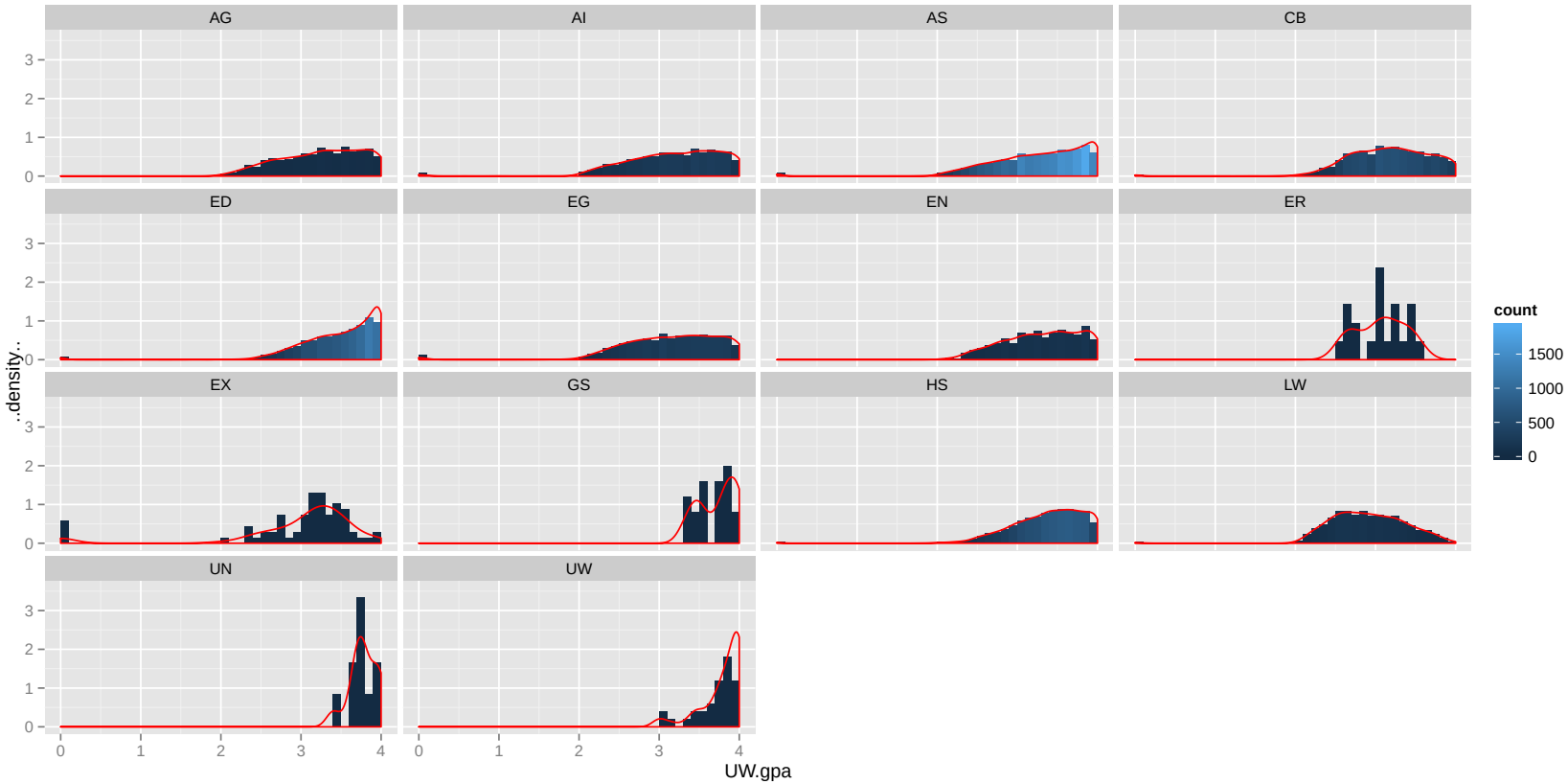
Boxplot of GPA in different degrees



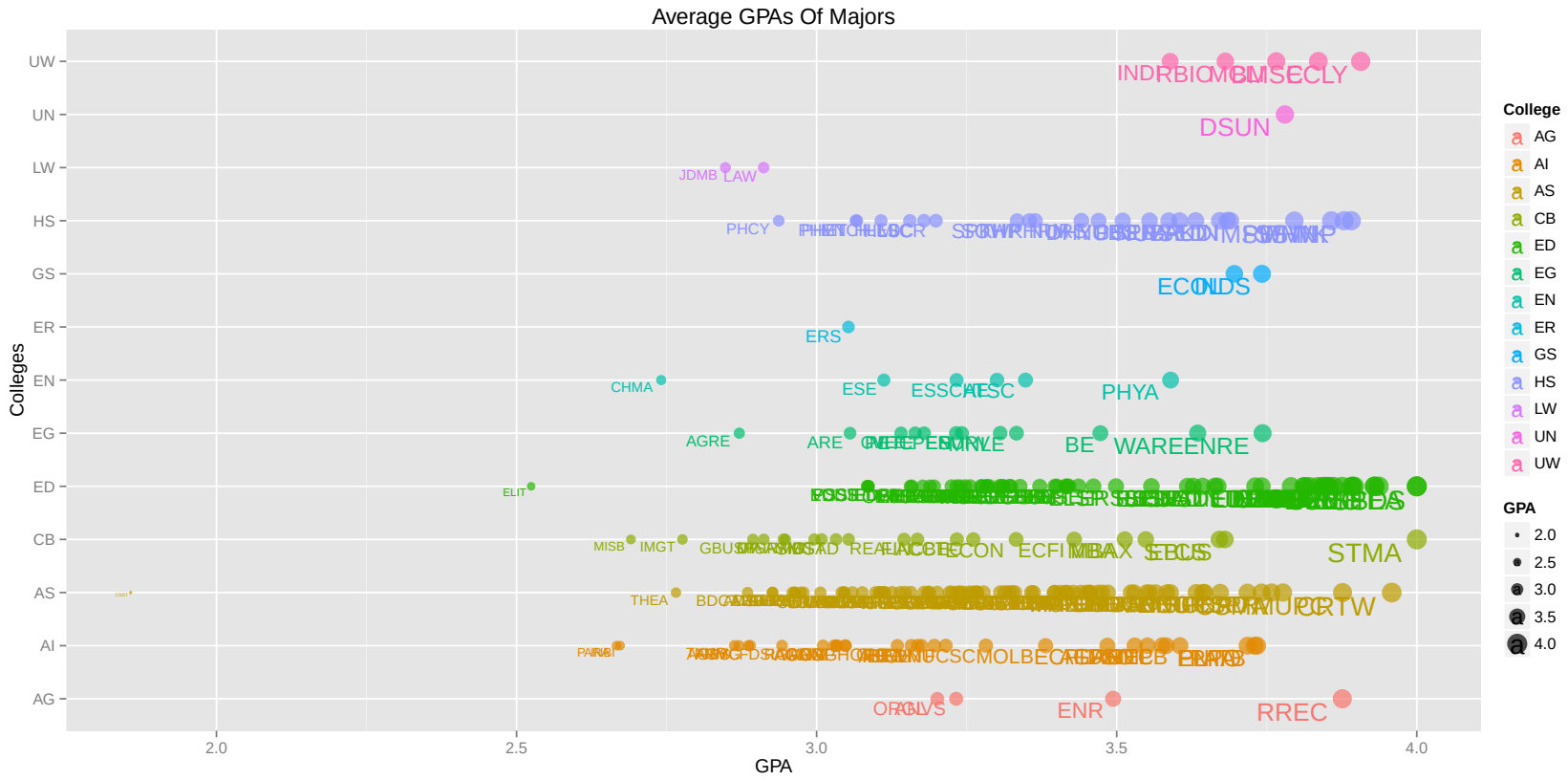
A more intuitive look



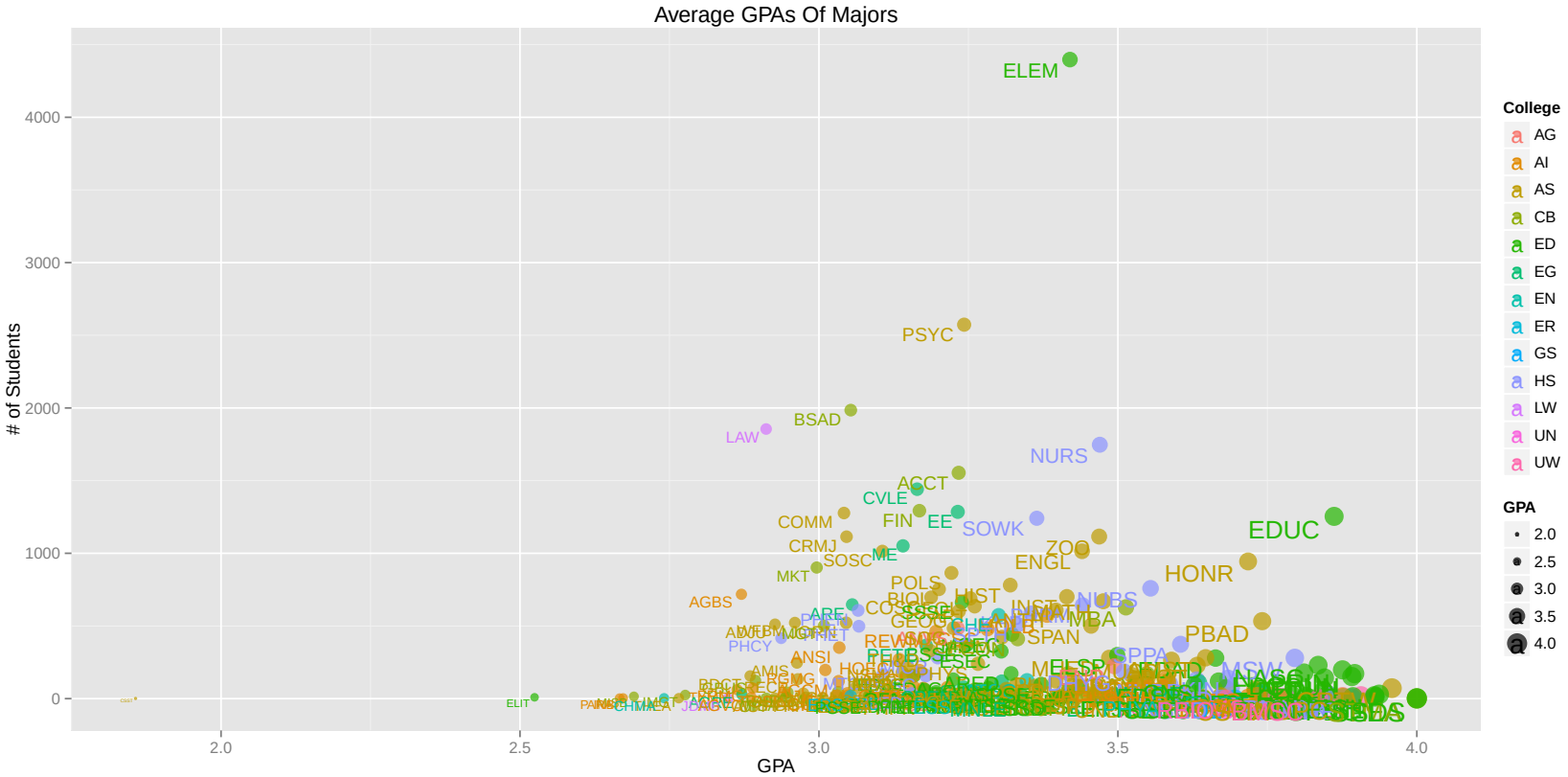
Different colleges



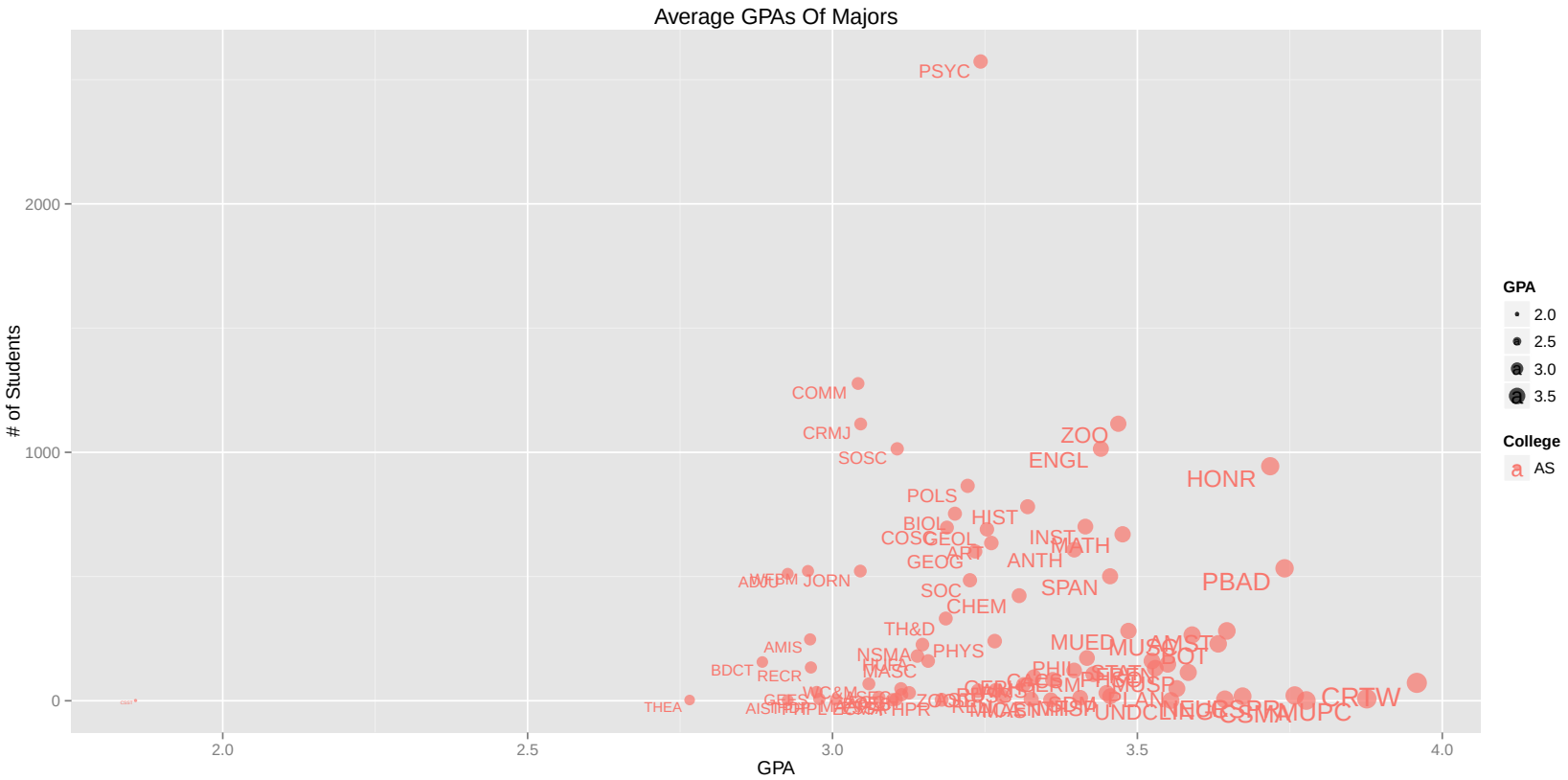
Different majors



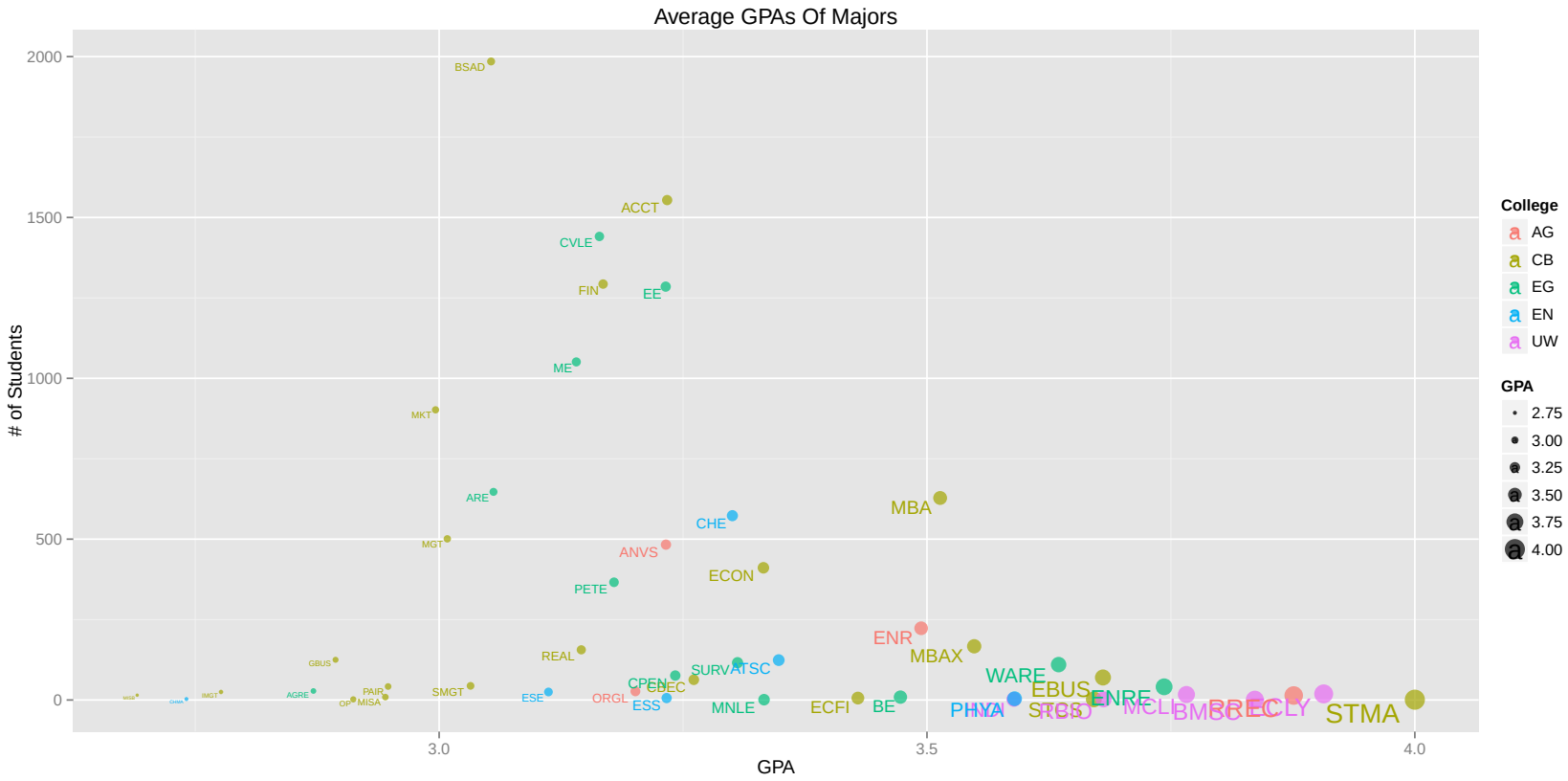
Considering the number of students



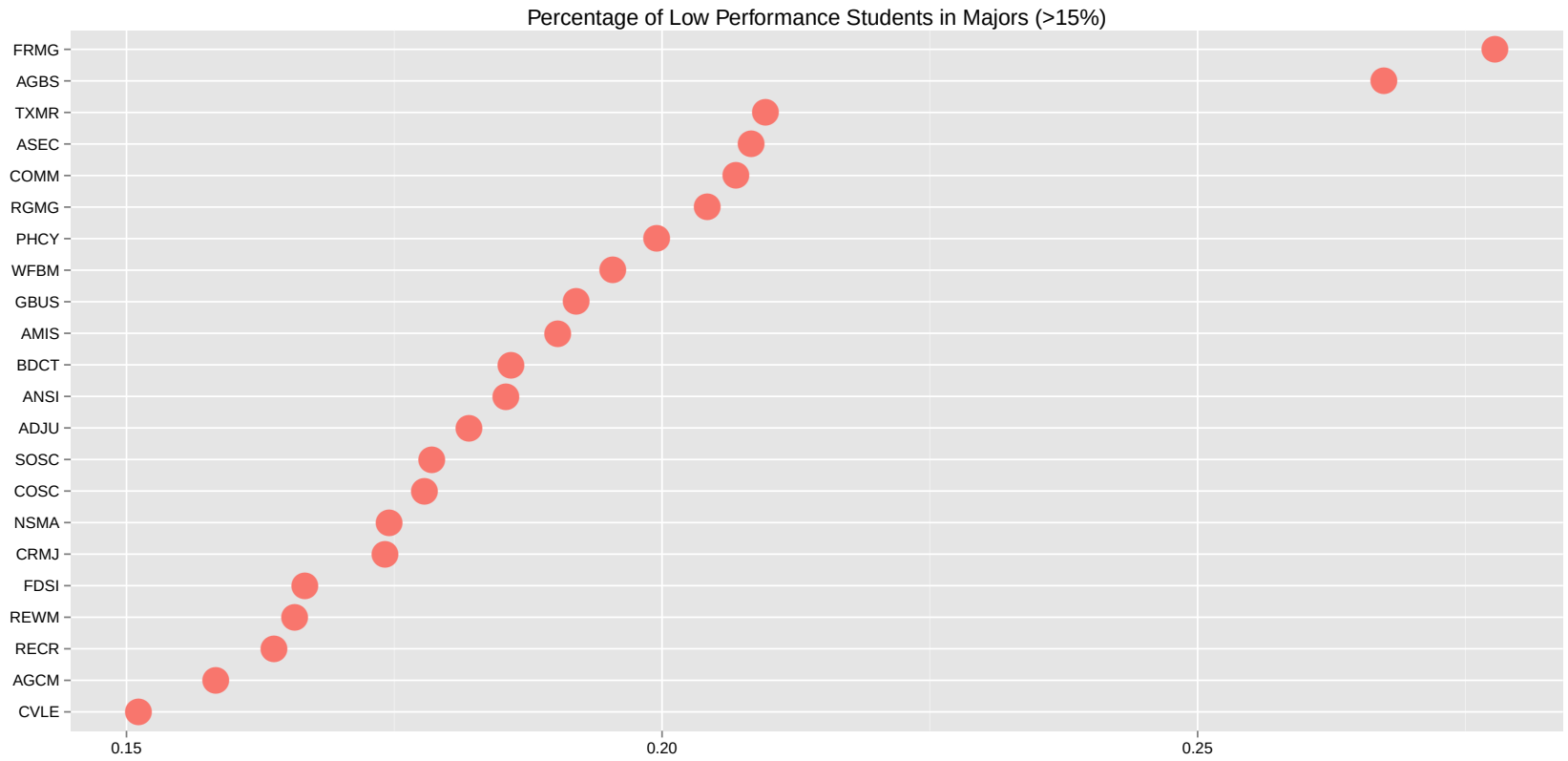
Only looking at Arts and Science majors



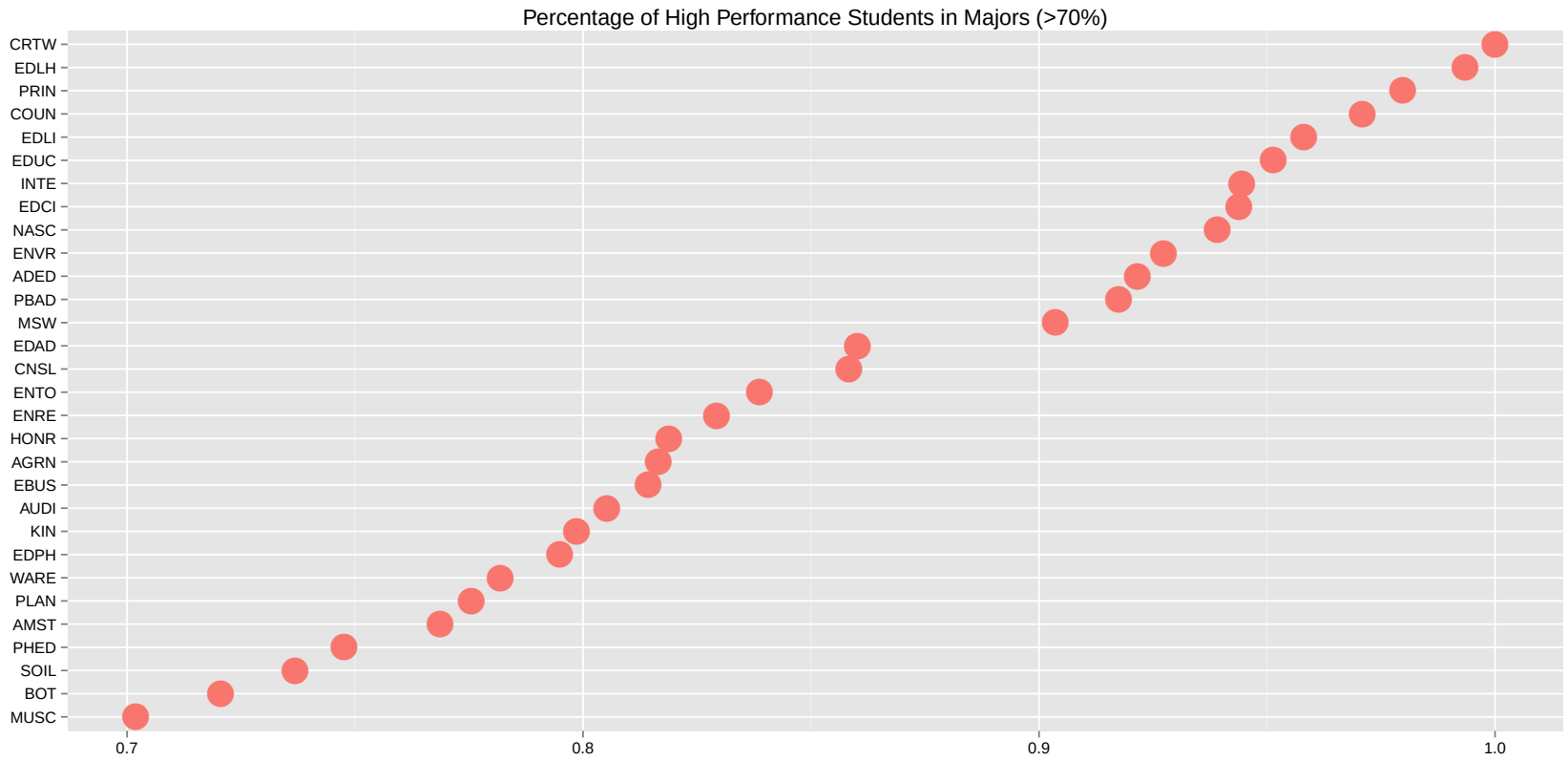
Colleges with few majors



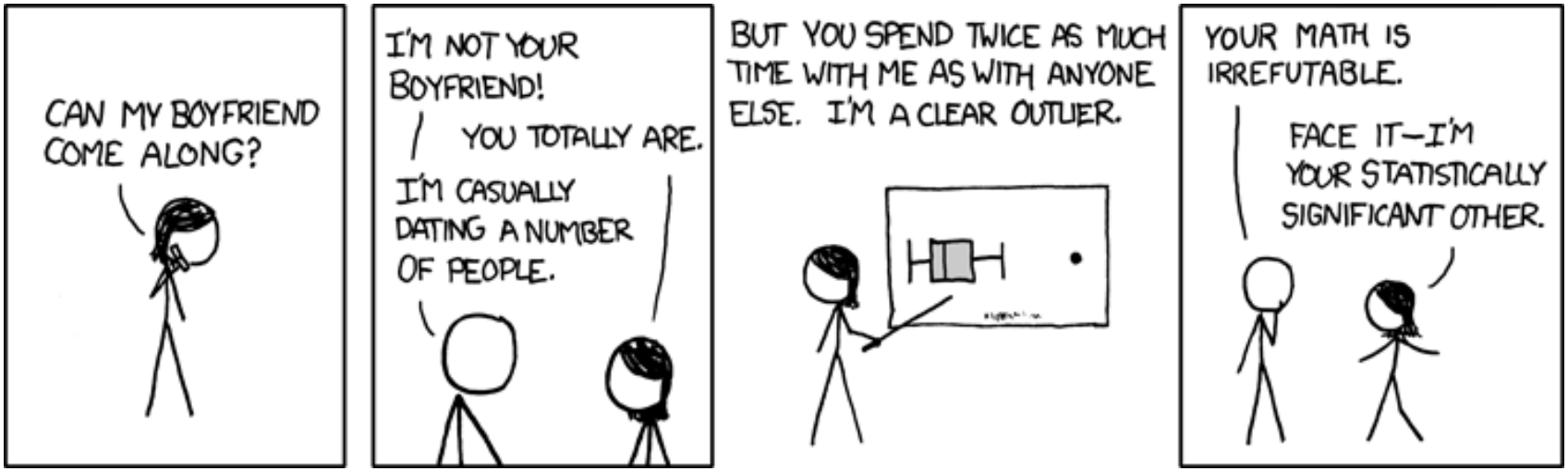
Hardest majors? (< 2.5)



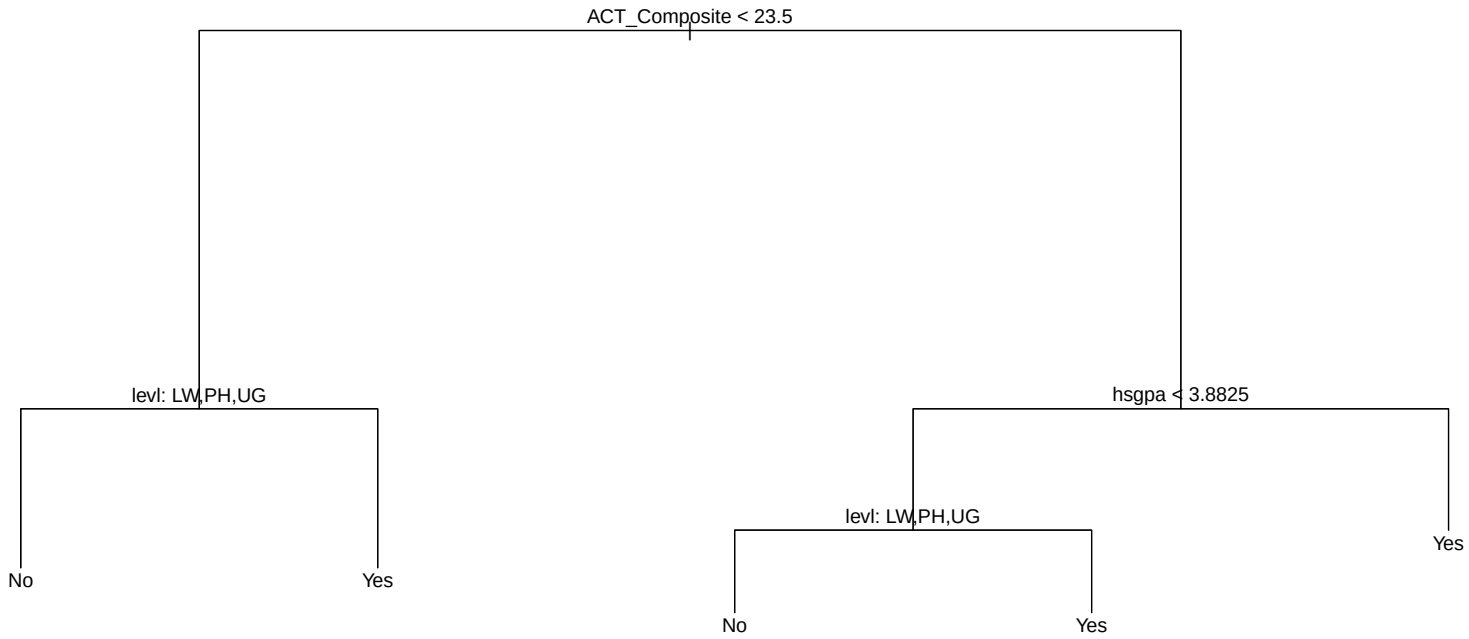
Easiest majors? (or they just work harder) (> 3.5)



About outliers



A simple way to look at it



Students with a > 3.5 GPA

Misclassification error rate: $0.2656 = 3053 \ / \ 11495$

What random forest thinks are most important

Predictors	Relative Influence
Highschool GPA	37.3484049

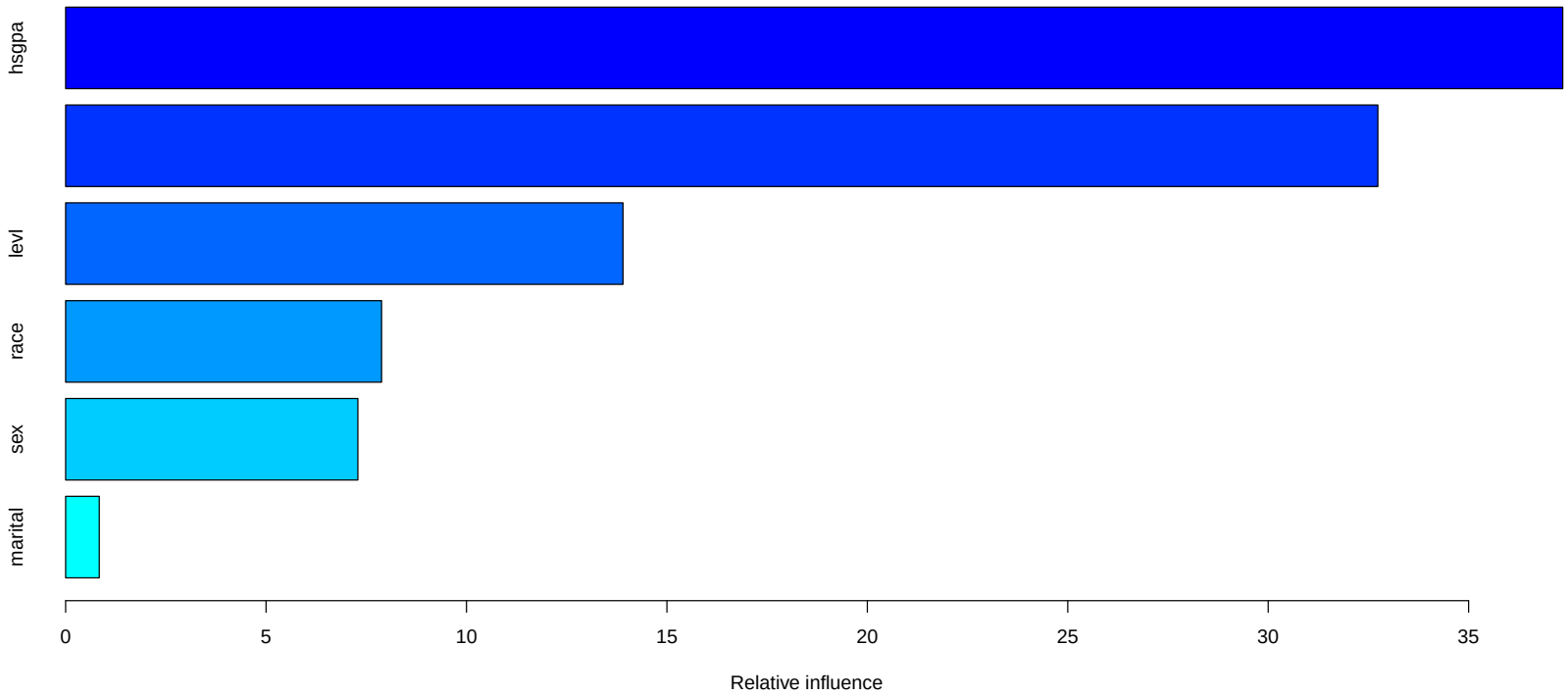
What random forest thinks are most important

Predictors	Relative Influence
Highschool GPA	37.3484049
ACT Composite	32.7361483

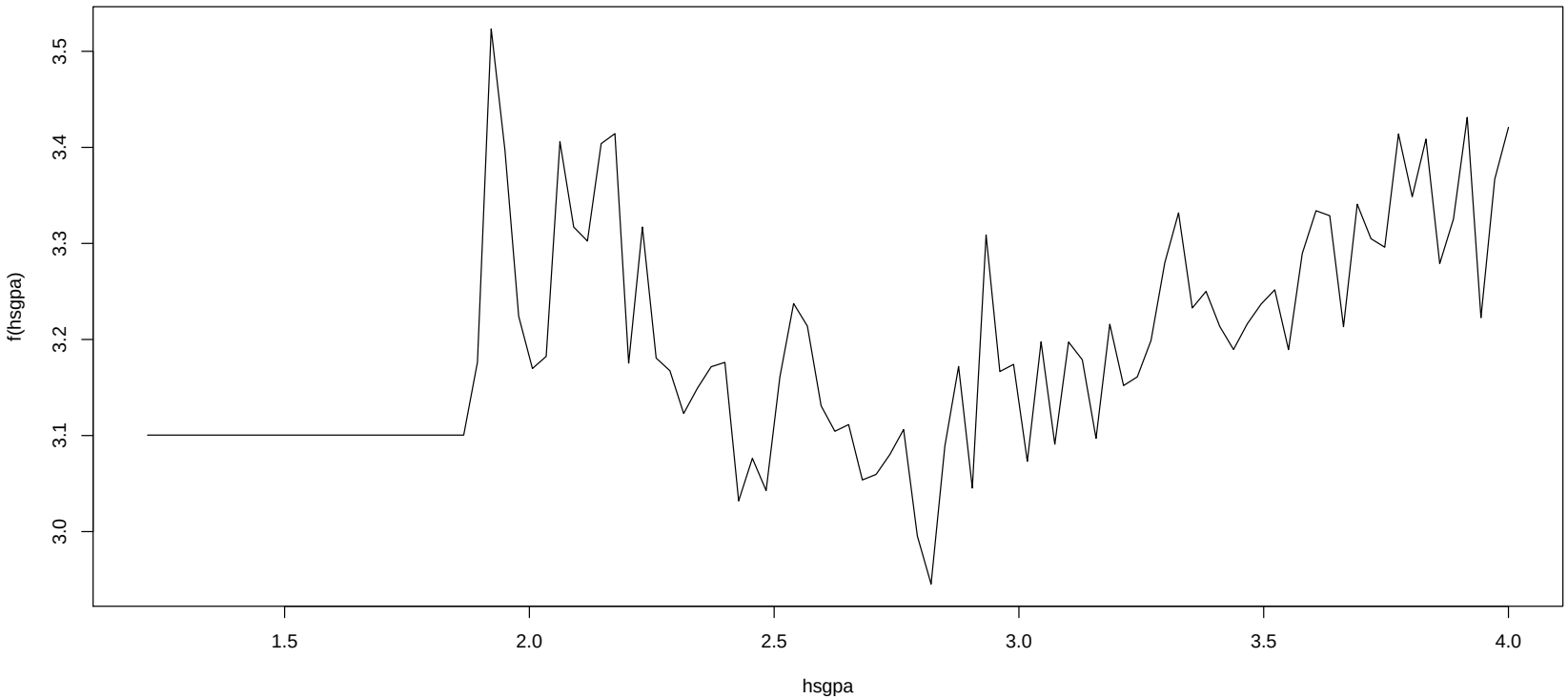
What random forest thinks are most important

Predictors	Relative Influence
Highschool GPA	37.3484049
ACT Composite	32.7361483
Level	13.9047409
Race	7.8814707
Sex	7.2903646
Marital Status	0.8388707

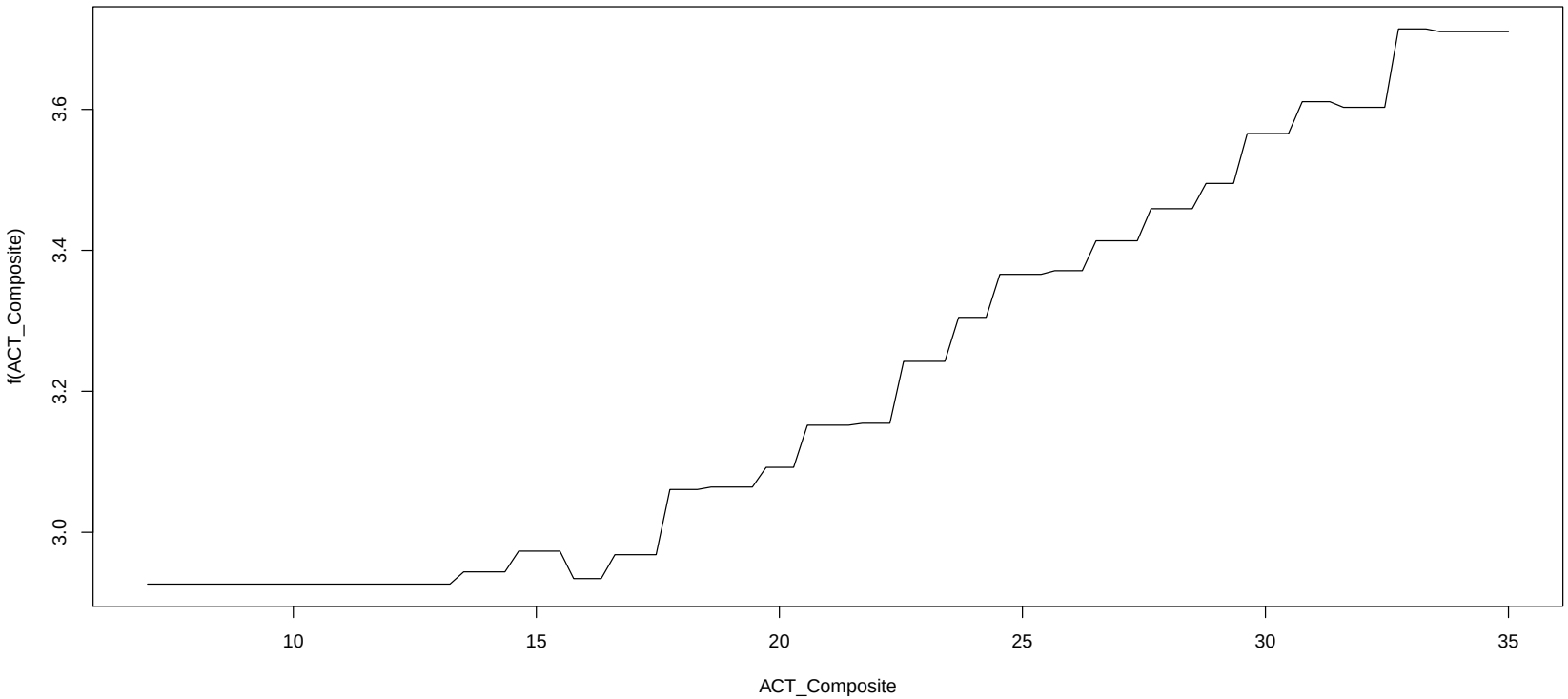
The method used here is called boosting algorithm



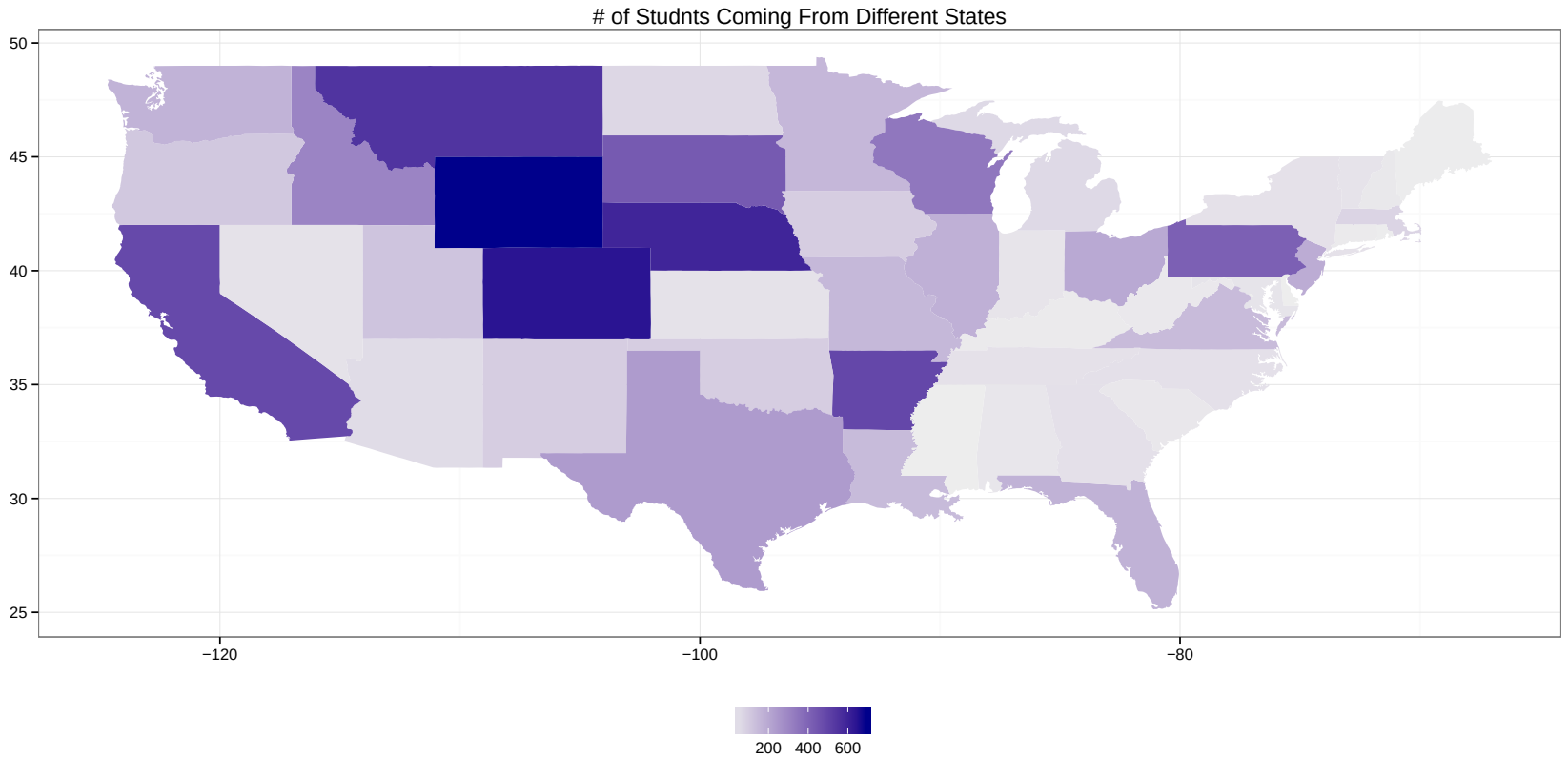
How does highschool GPA relate to UW GPA?



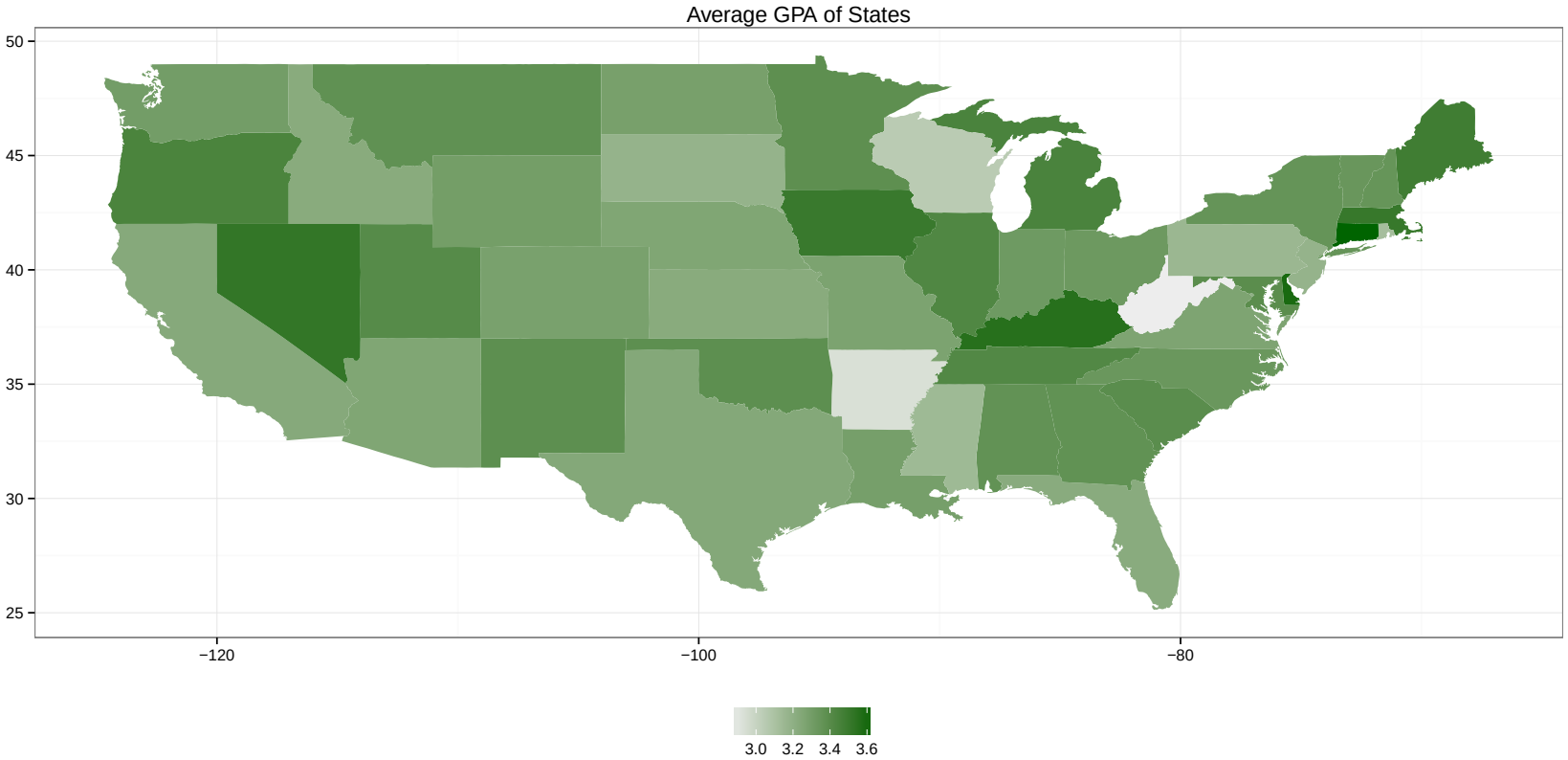
What about ACT Composite?



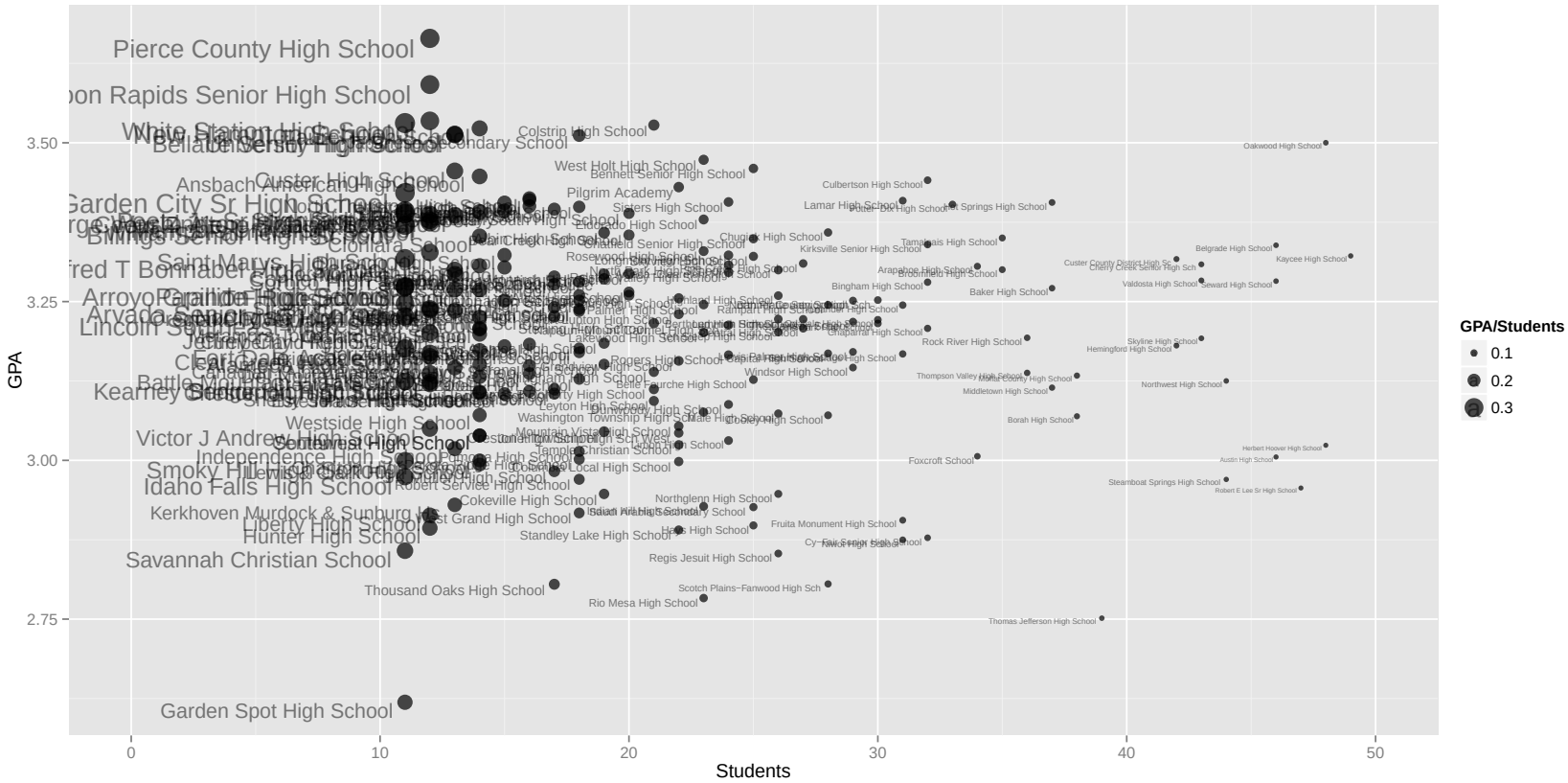
Most of our students come from...



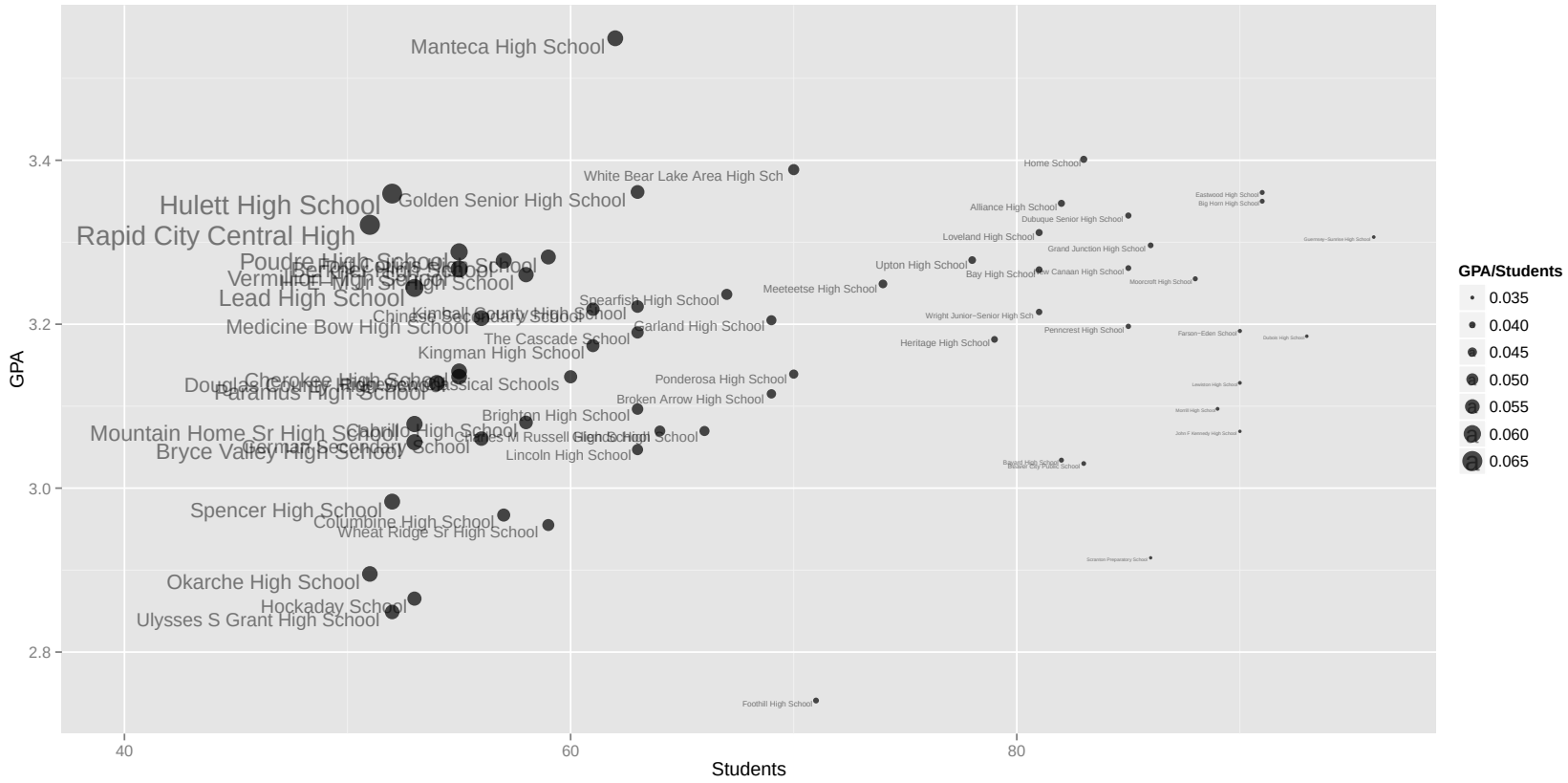
The best students come from. . .



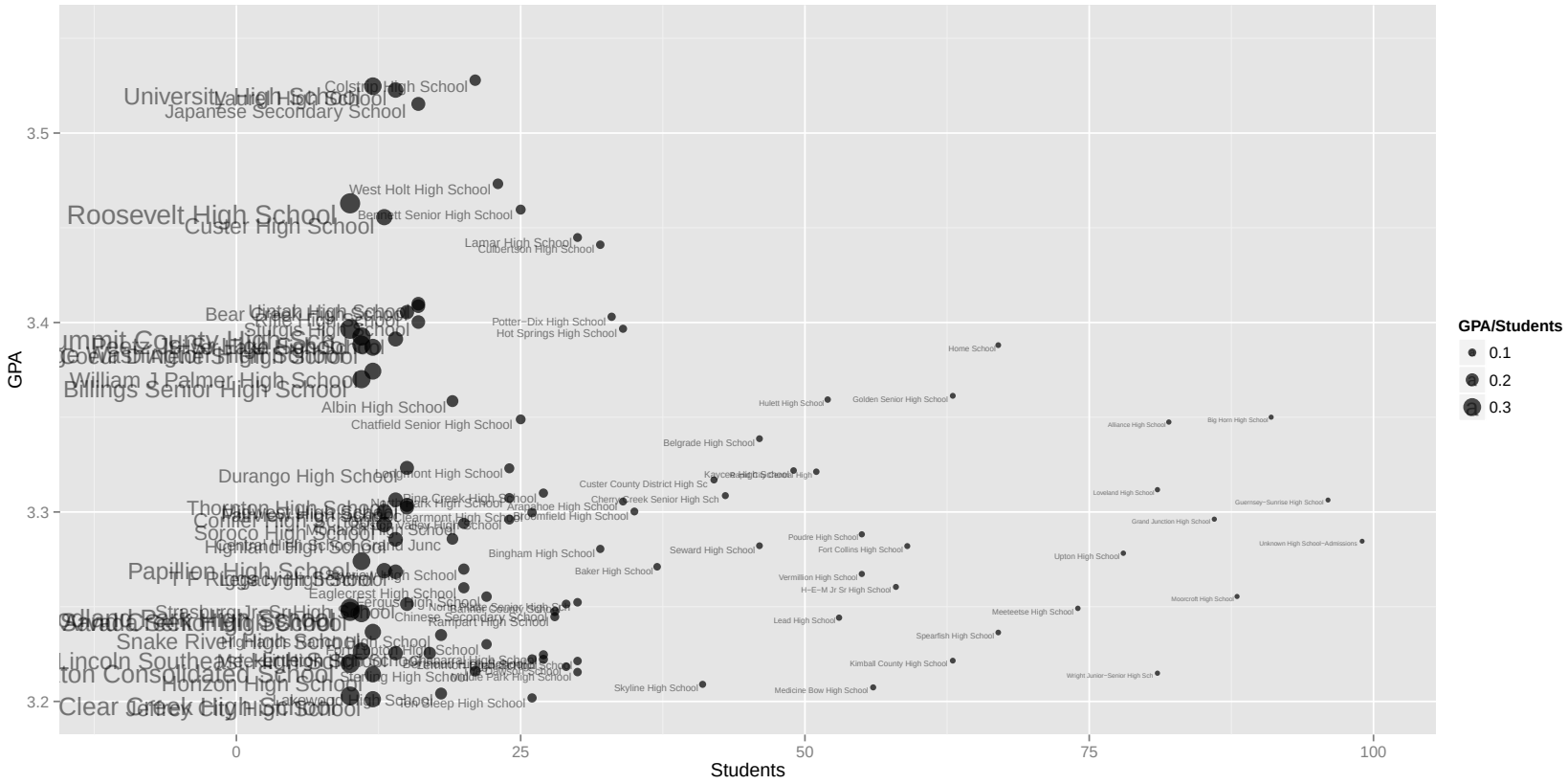
So many high schools!



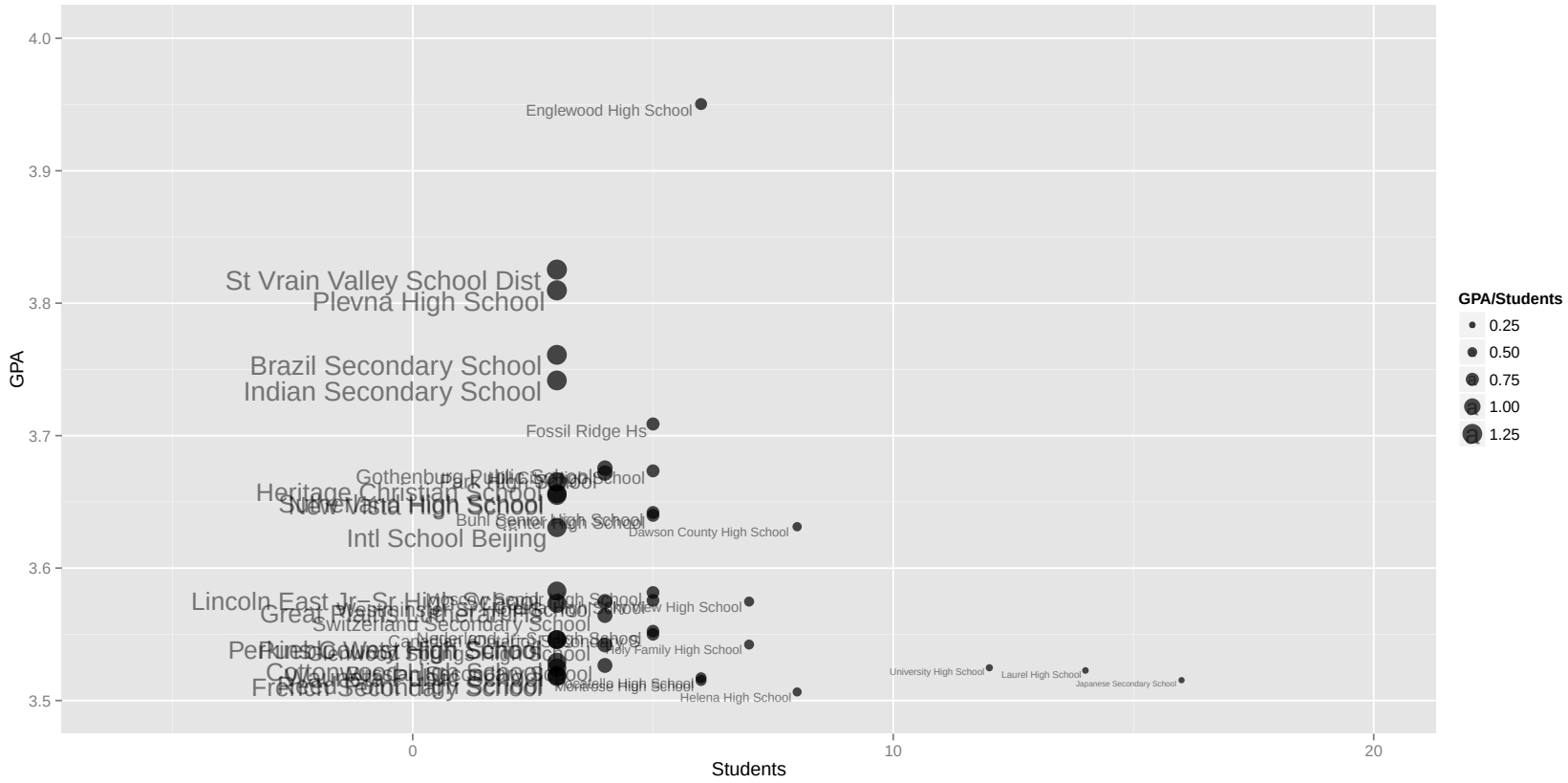
Let's take a closer look



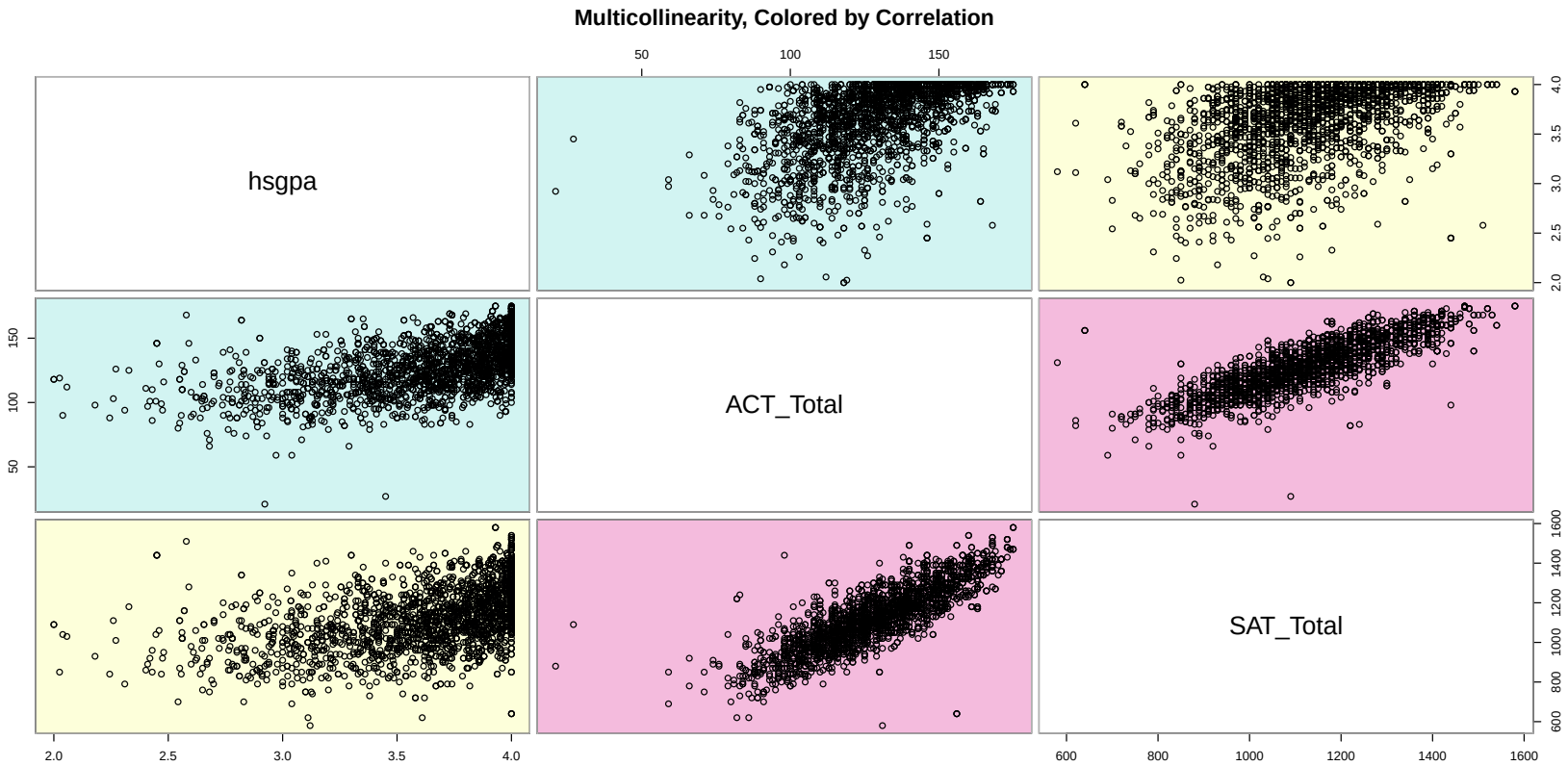
The high schools near UW



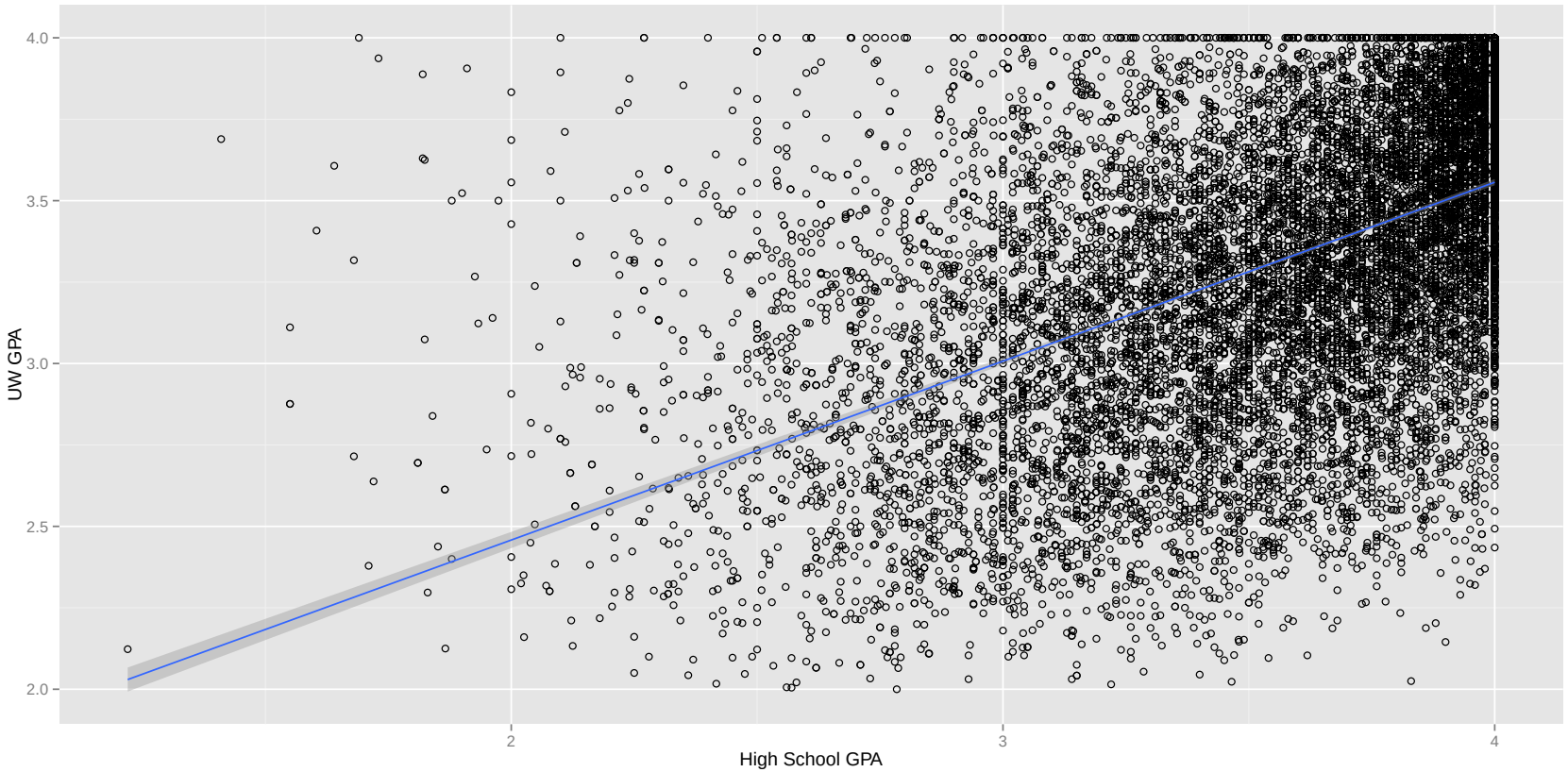
A closer look shows us. . .



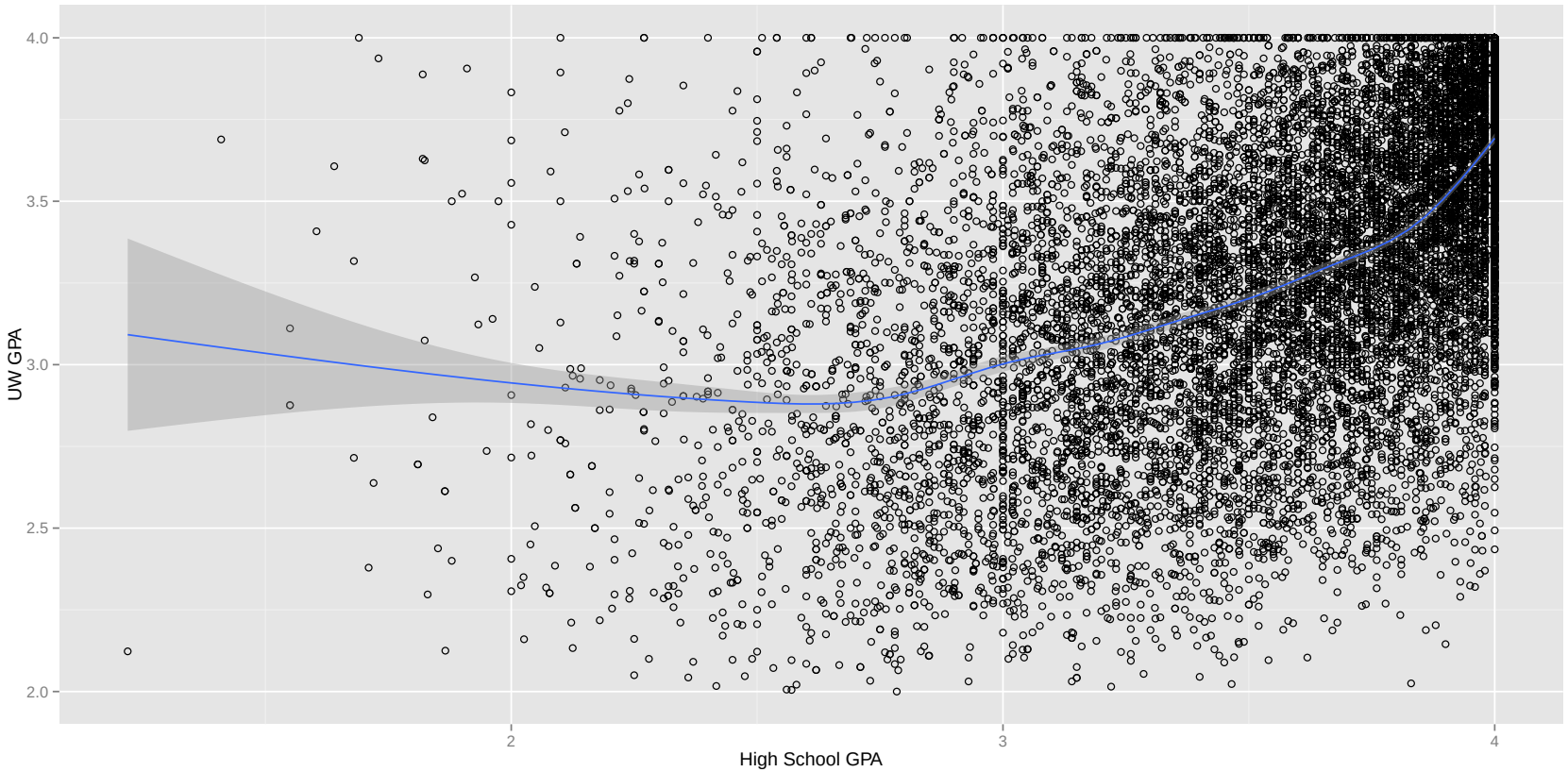
Our continuous predictors



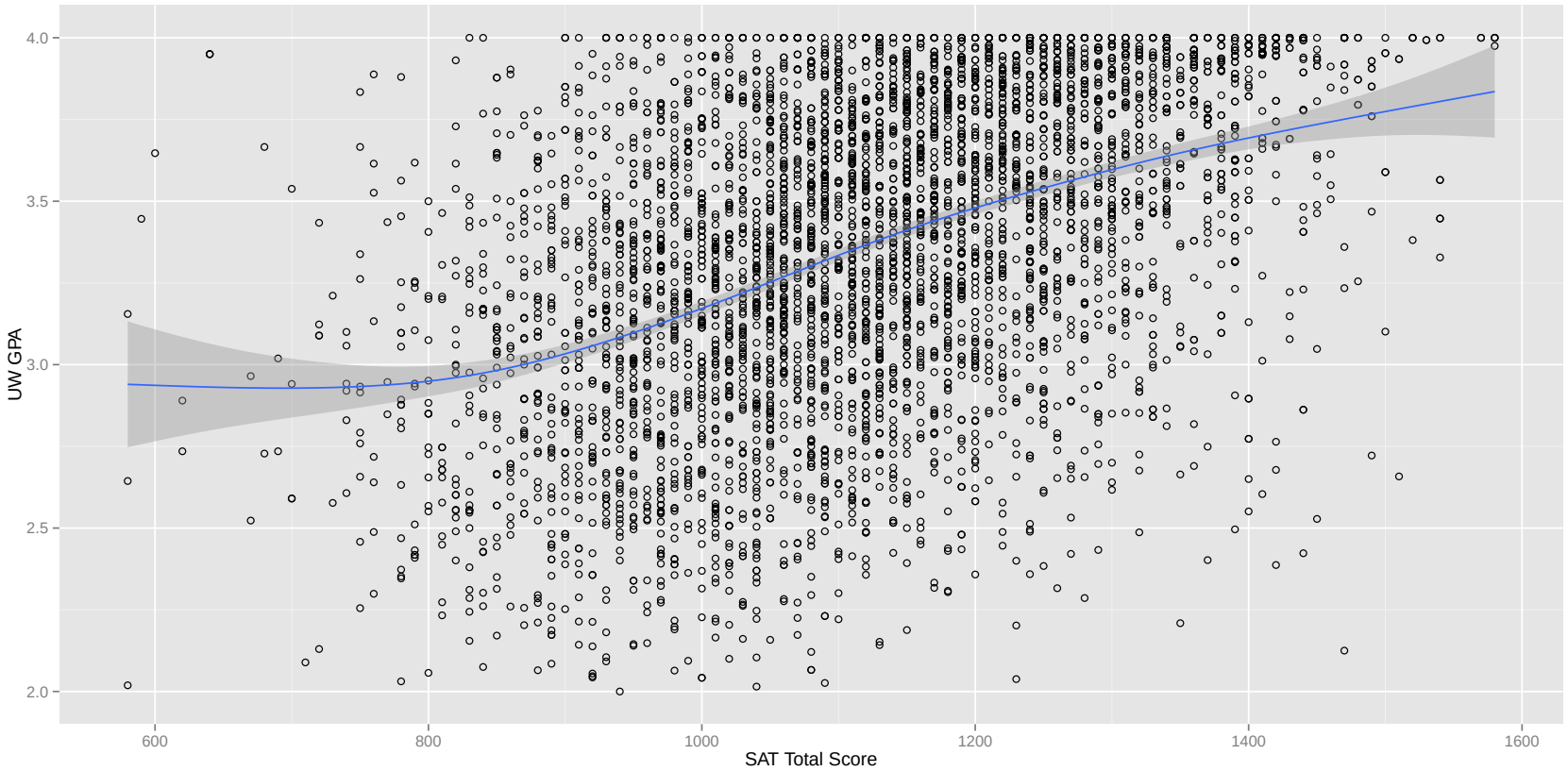
There is a positive strong correlation



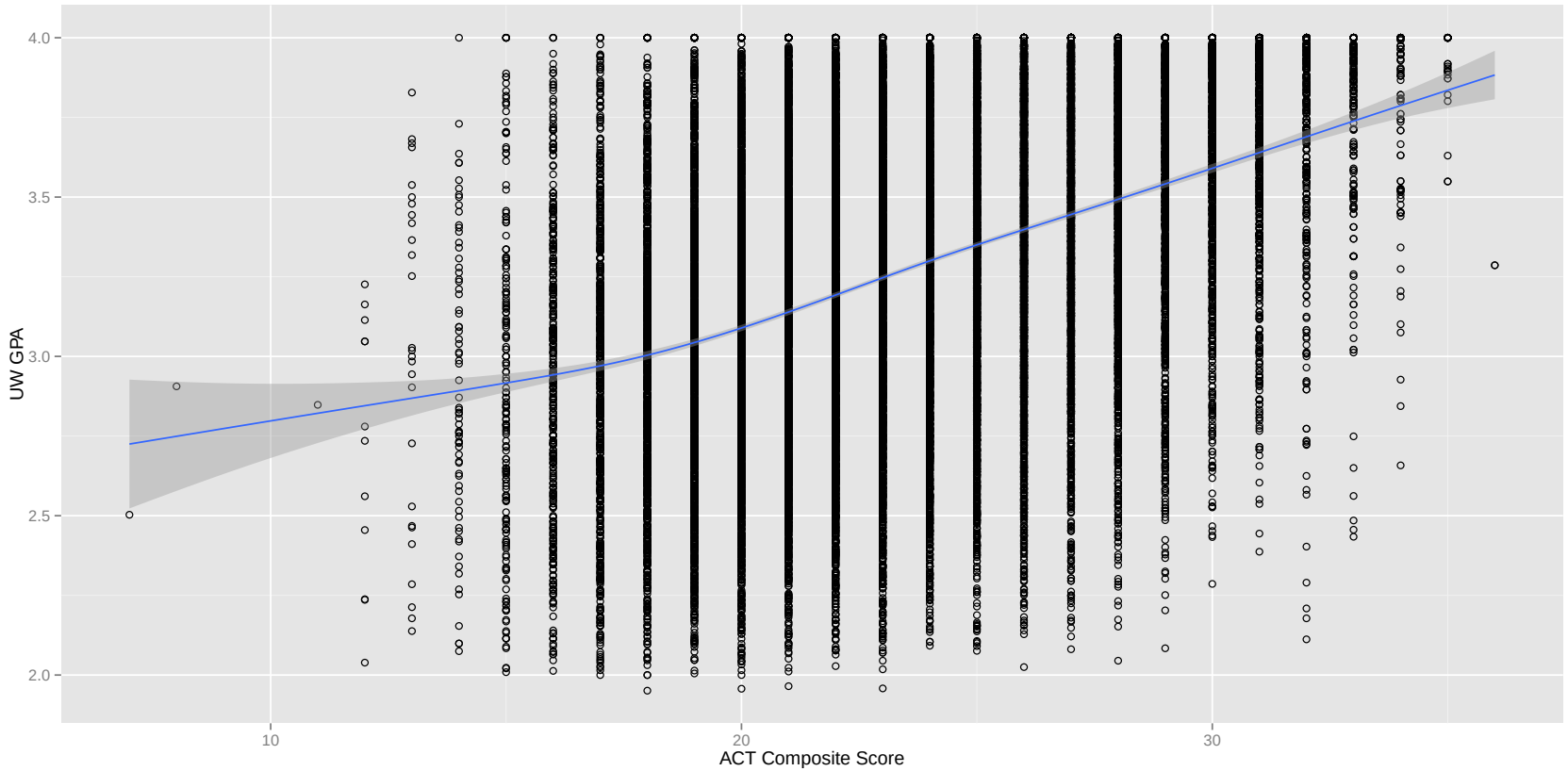
Another way to look at it



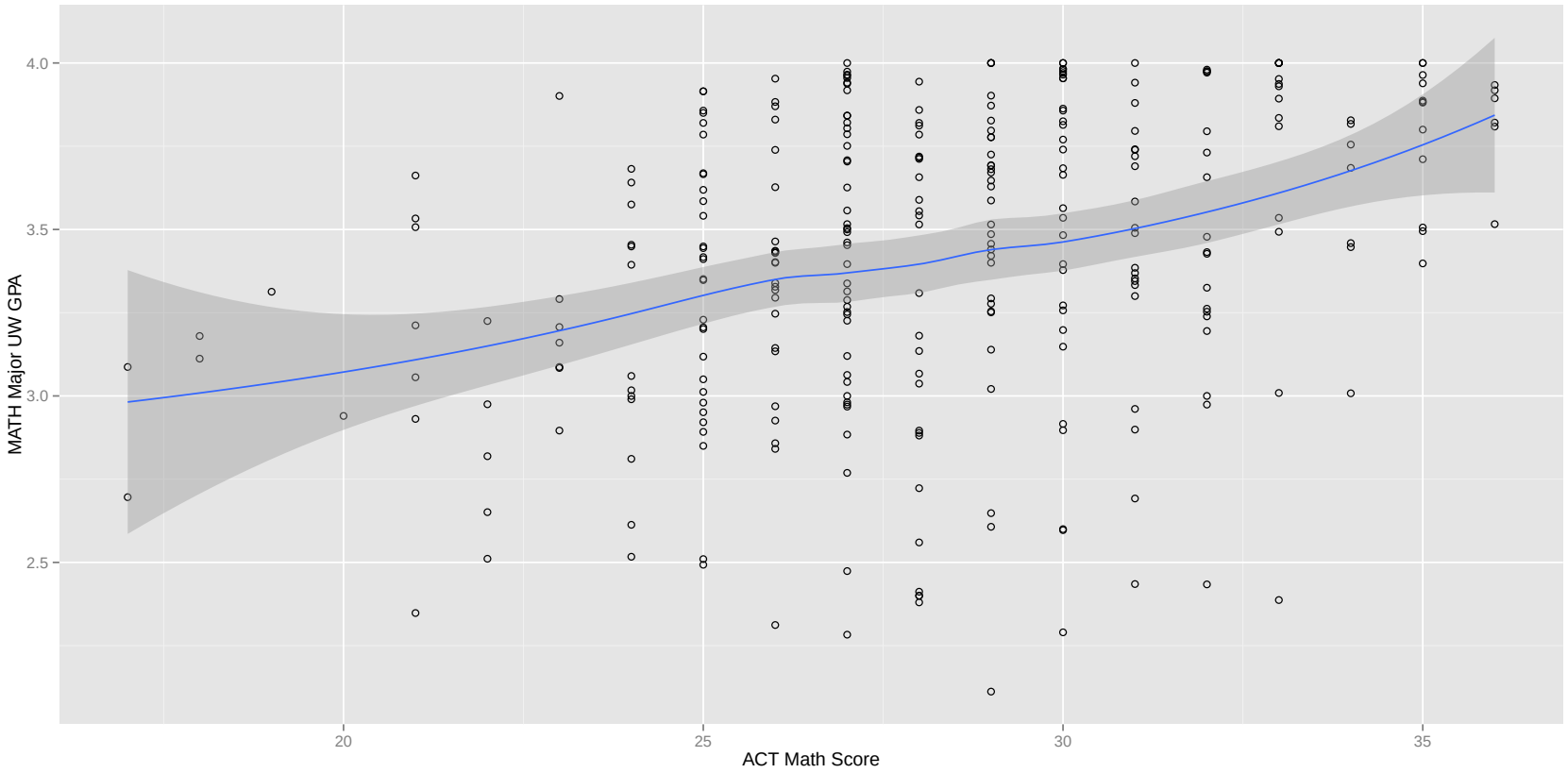
SAT


4.0
3.5
3.0
2.5
2.0
UW GPA
600
800
1000
1200
1400
1600
SAT Total Score

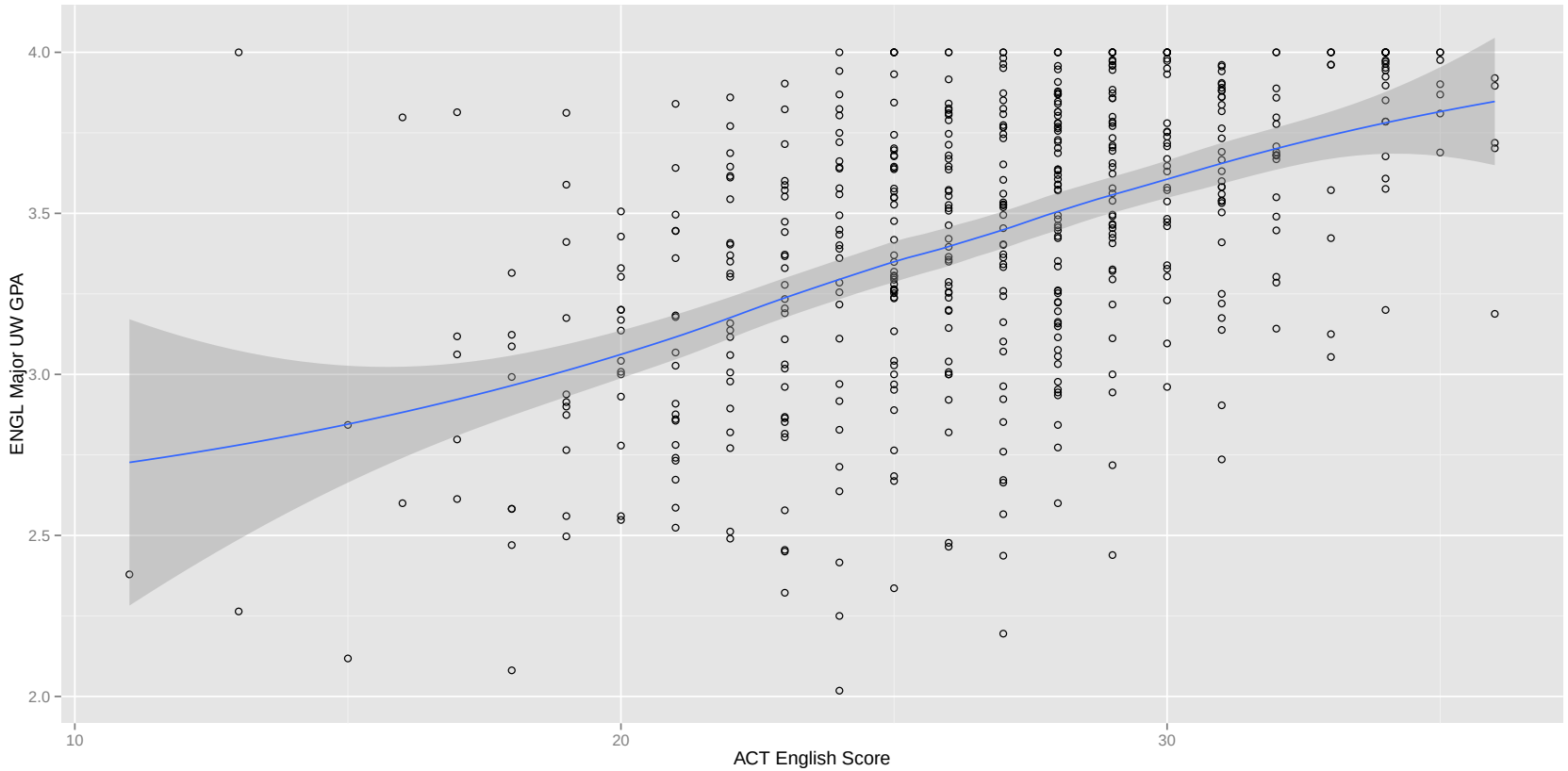
ACT



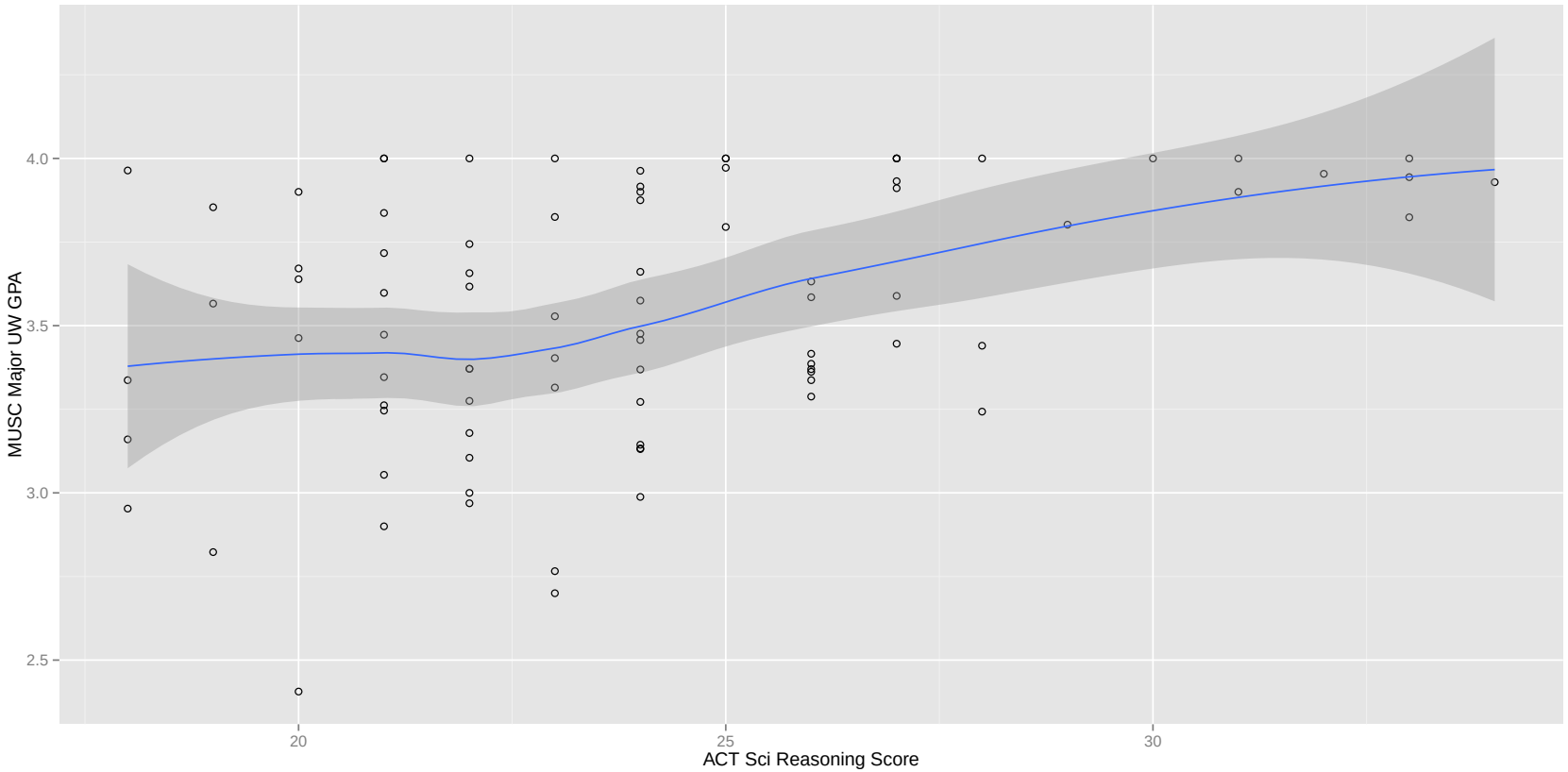
Do Math majors really need Math?



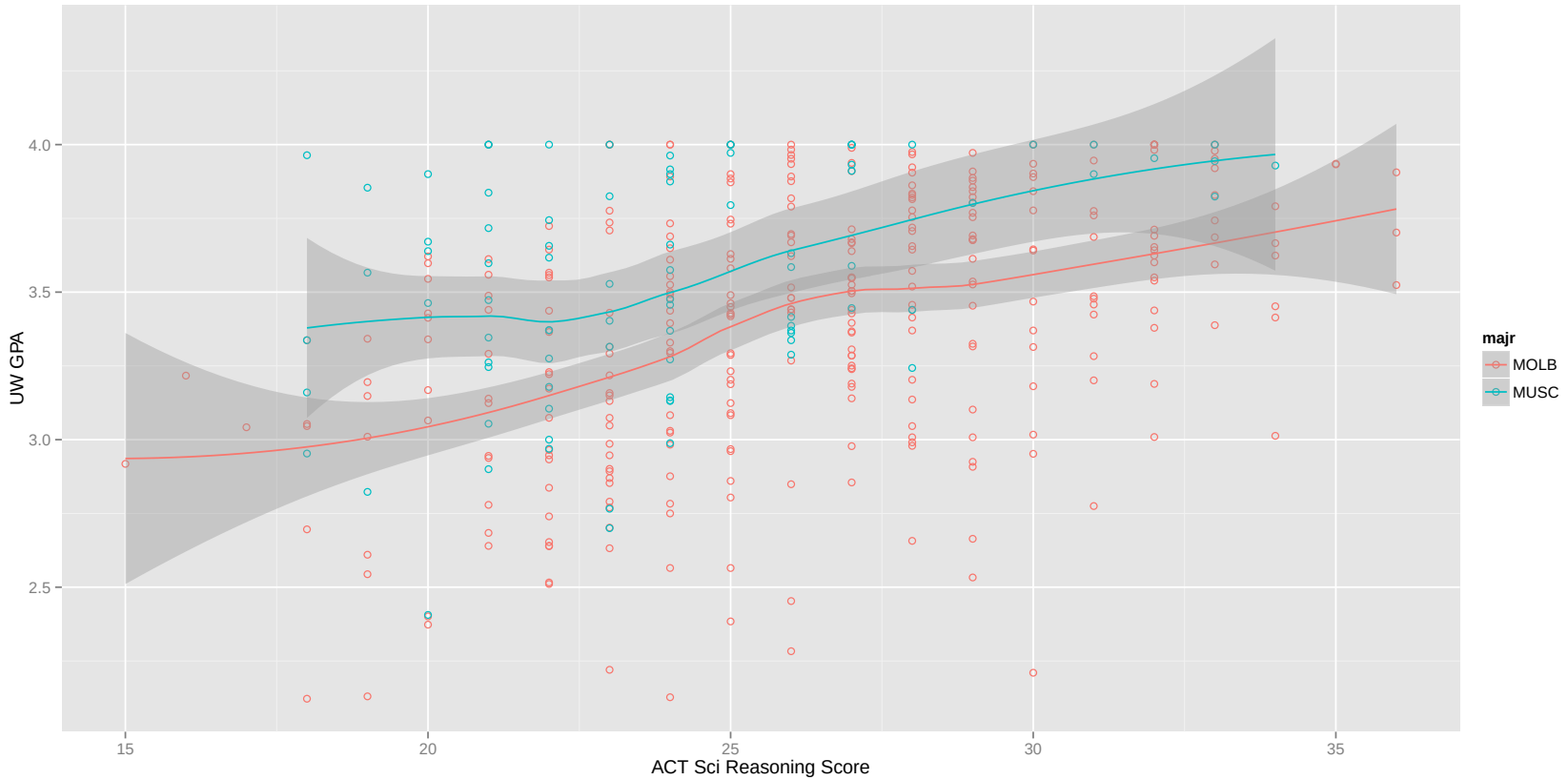
Do English majors really need English?



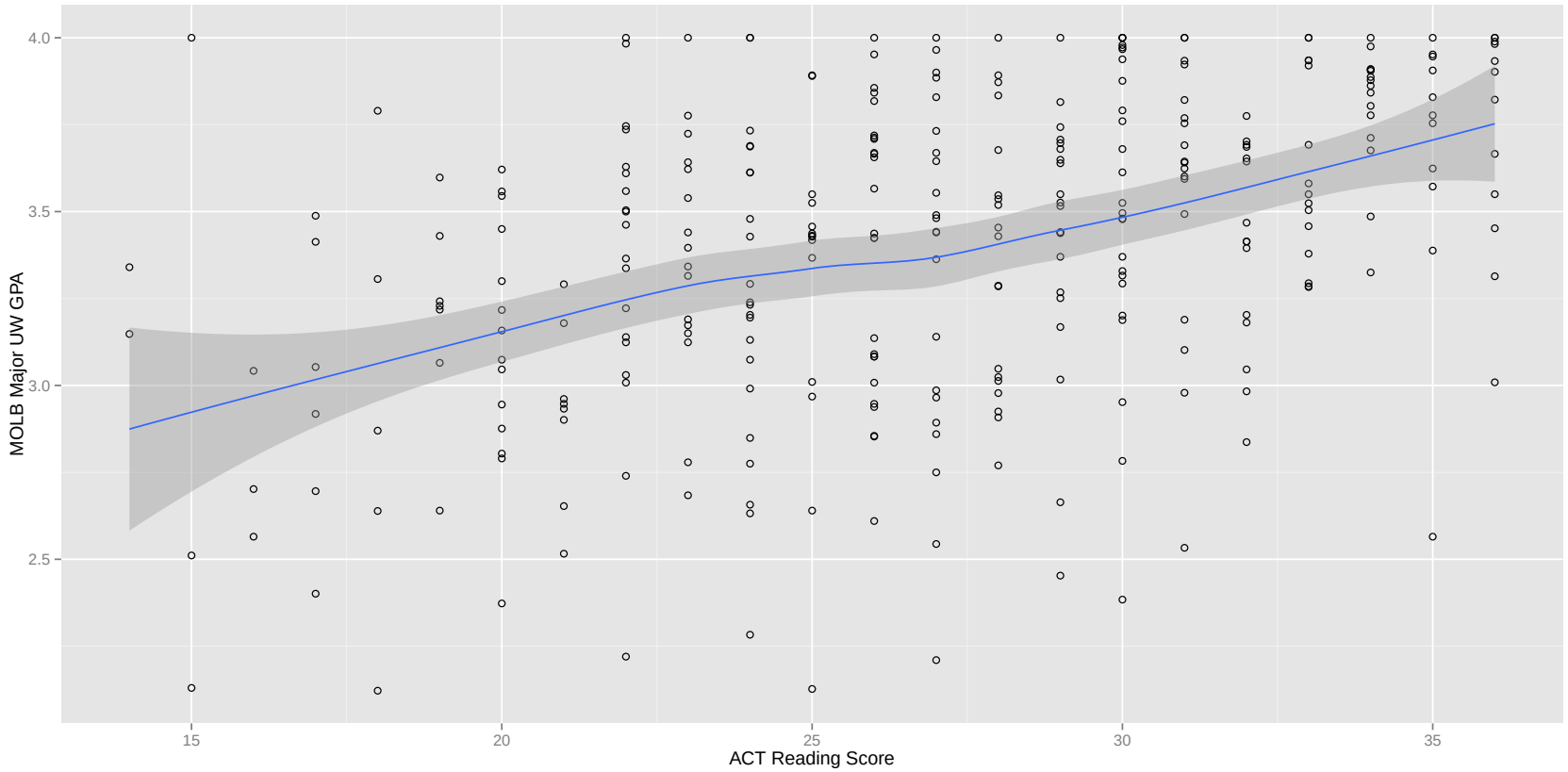
Do Music majors need science reasoning?



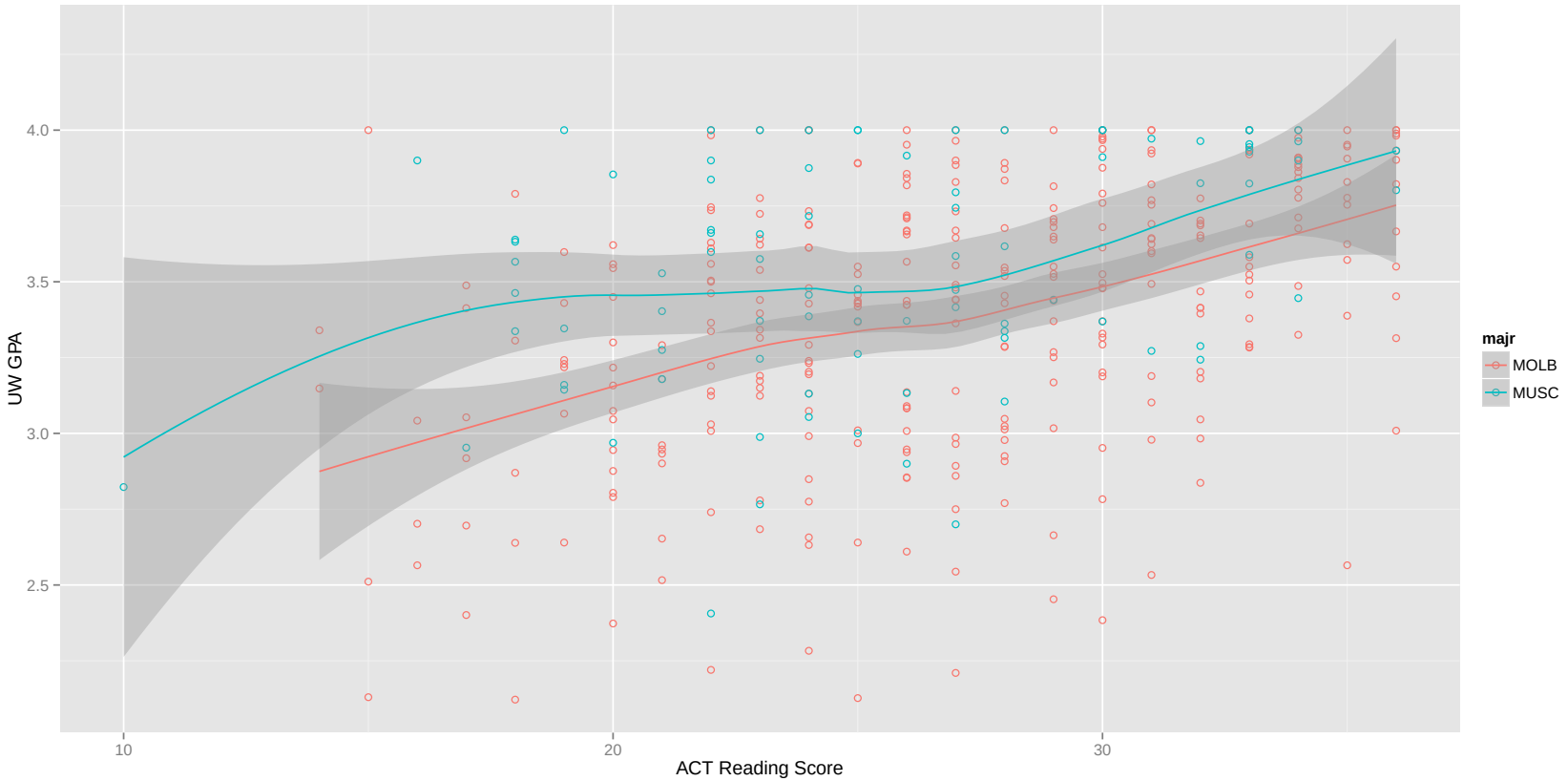
When comparing to Molecular Biology



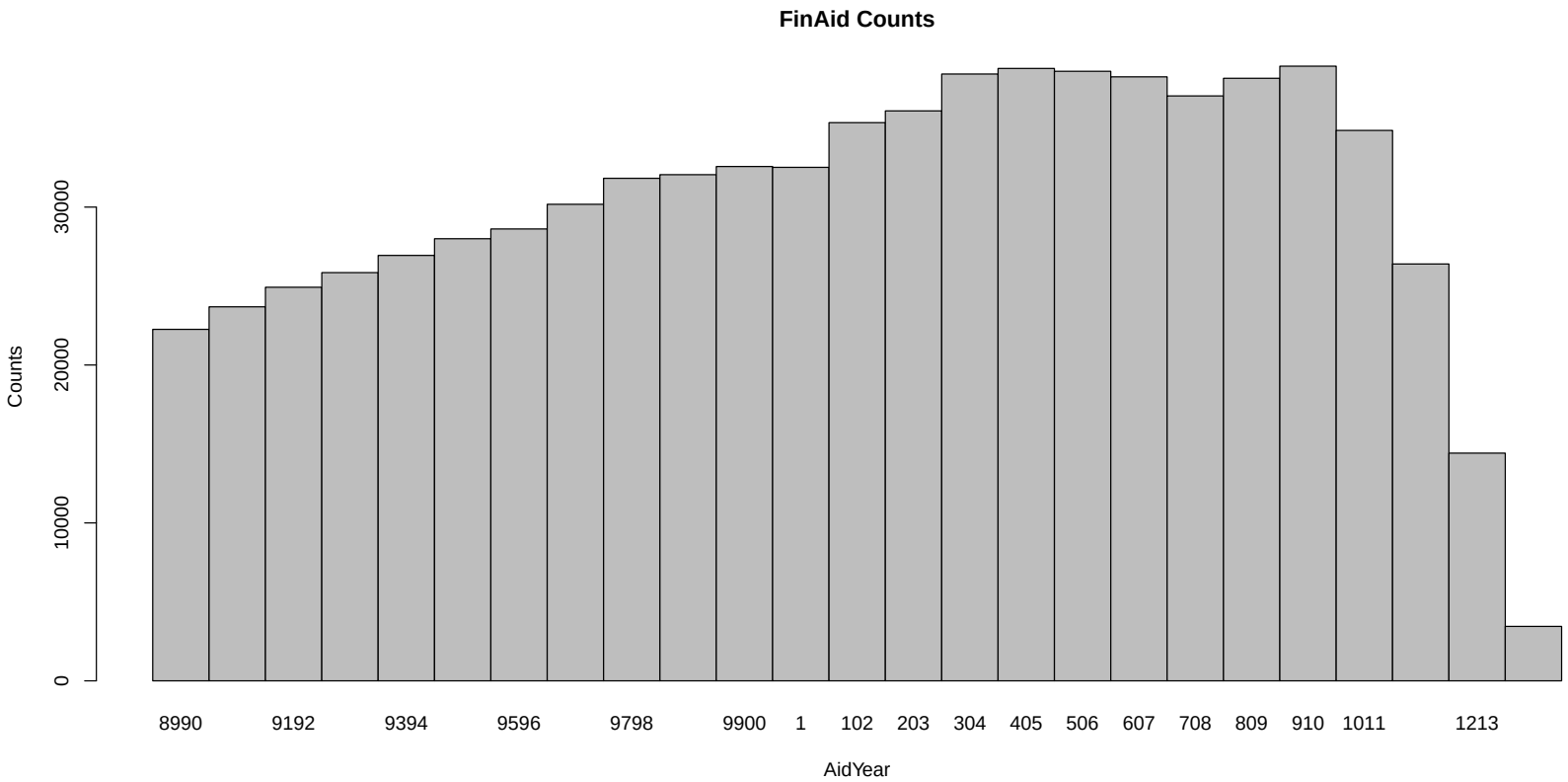
Molecular Biology again



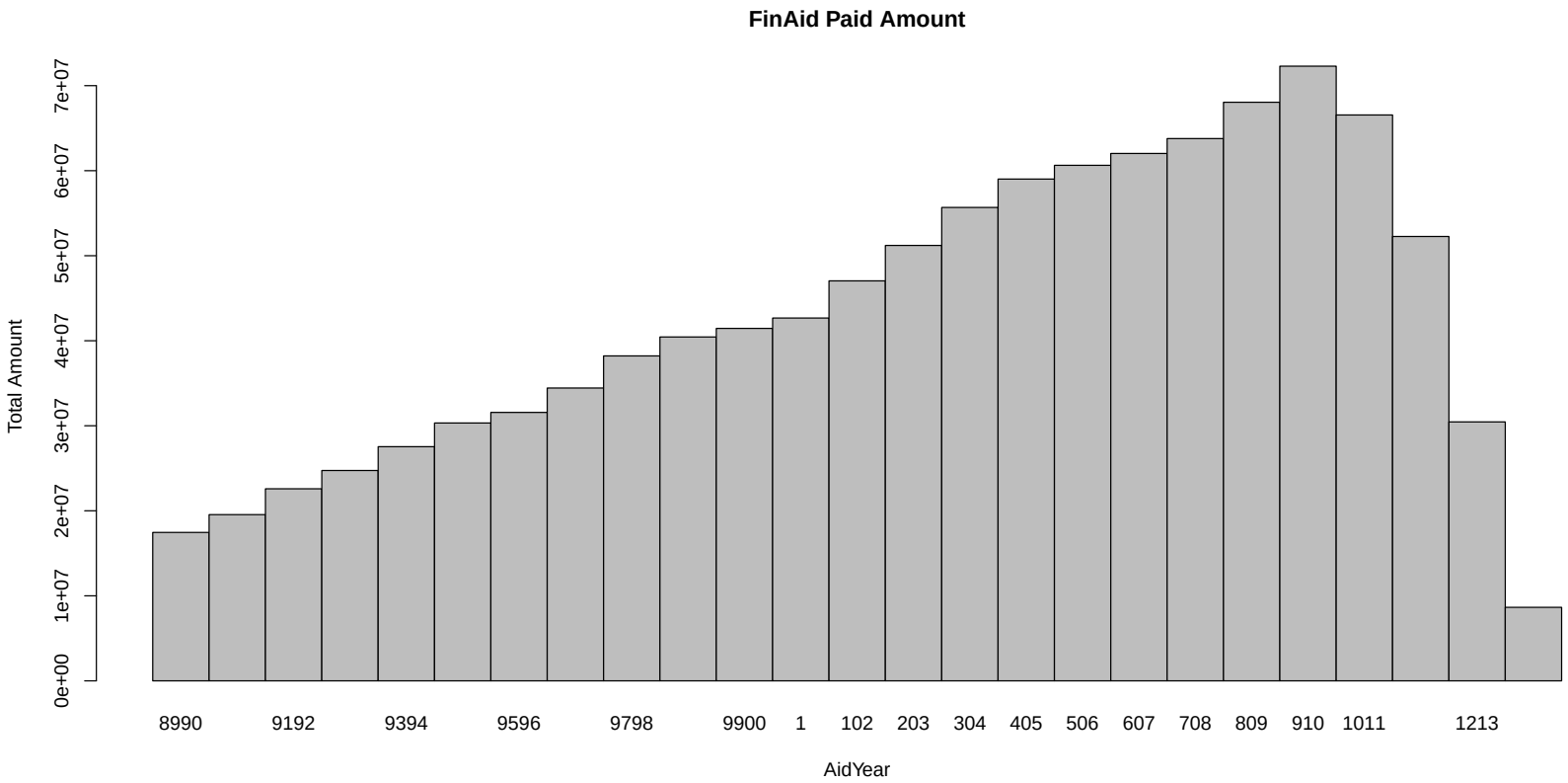
Let's add Music again



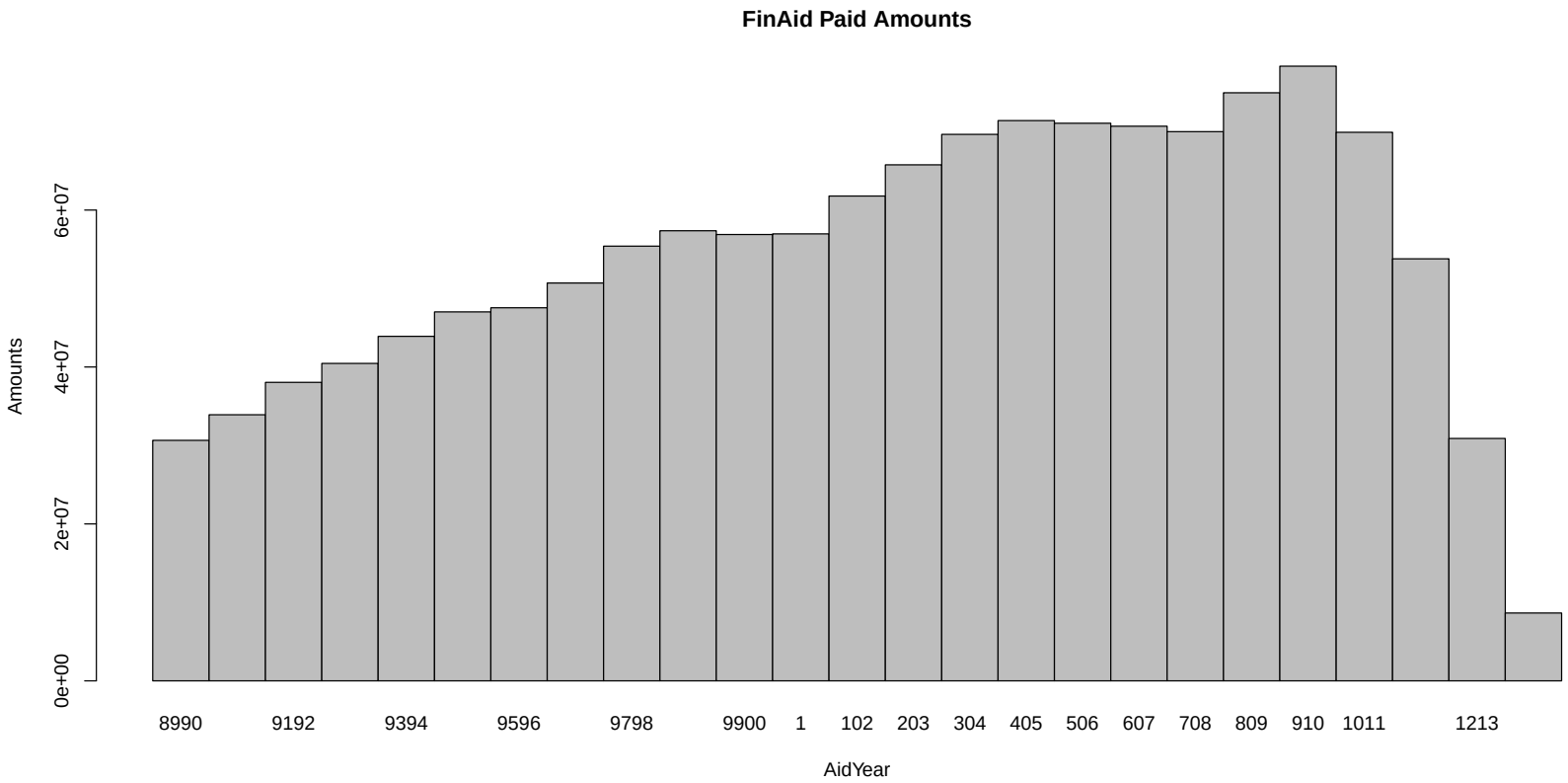
Are UW students receiving more and more FinAid?



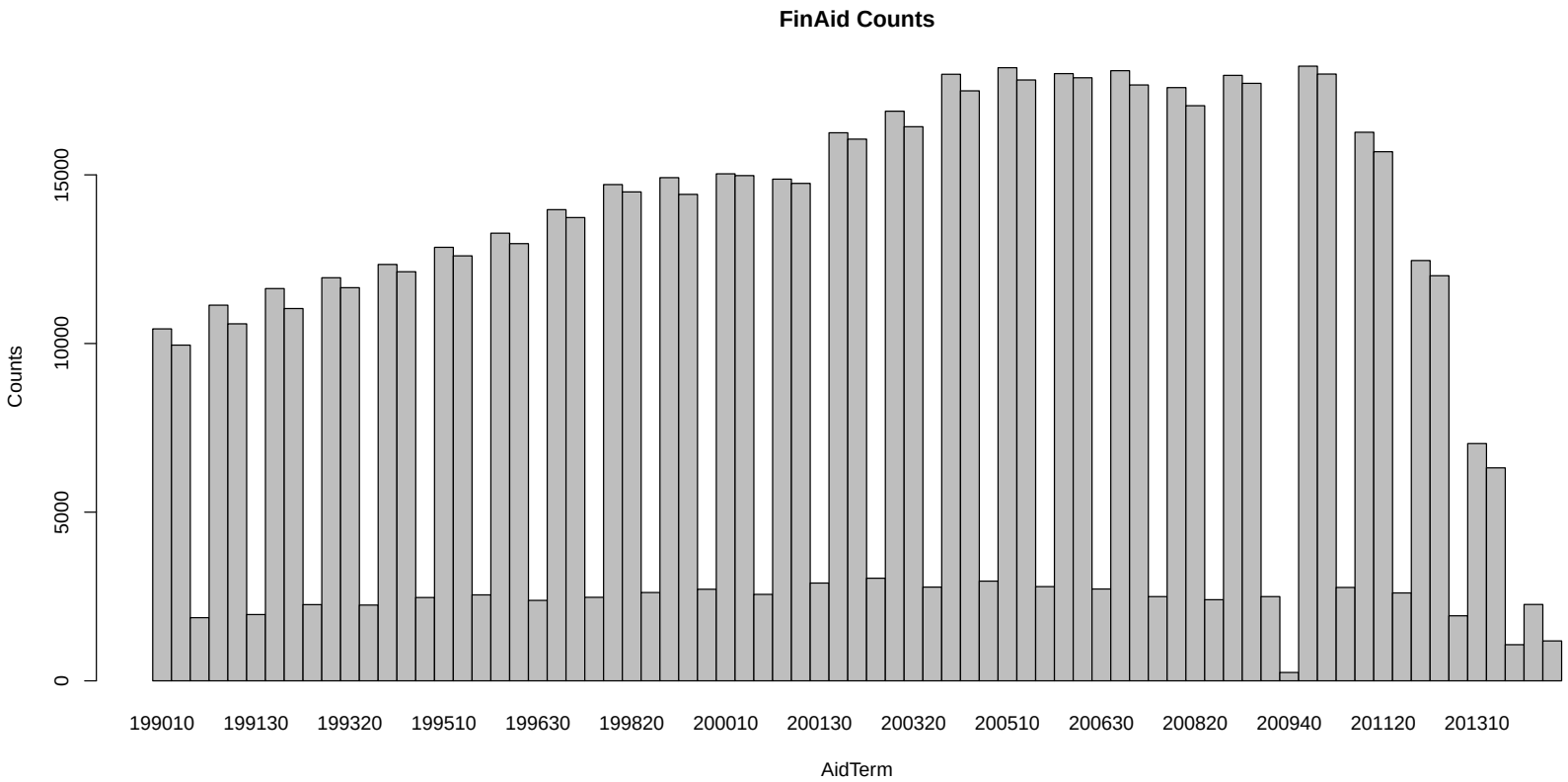
From the amount of money perspective



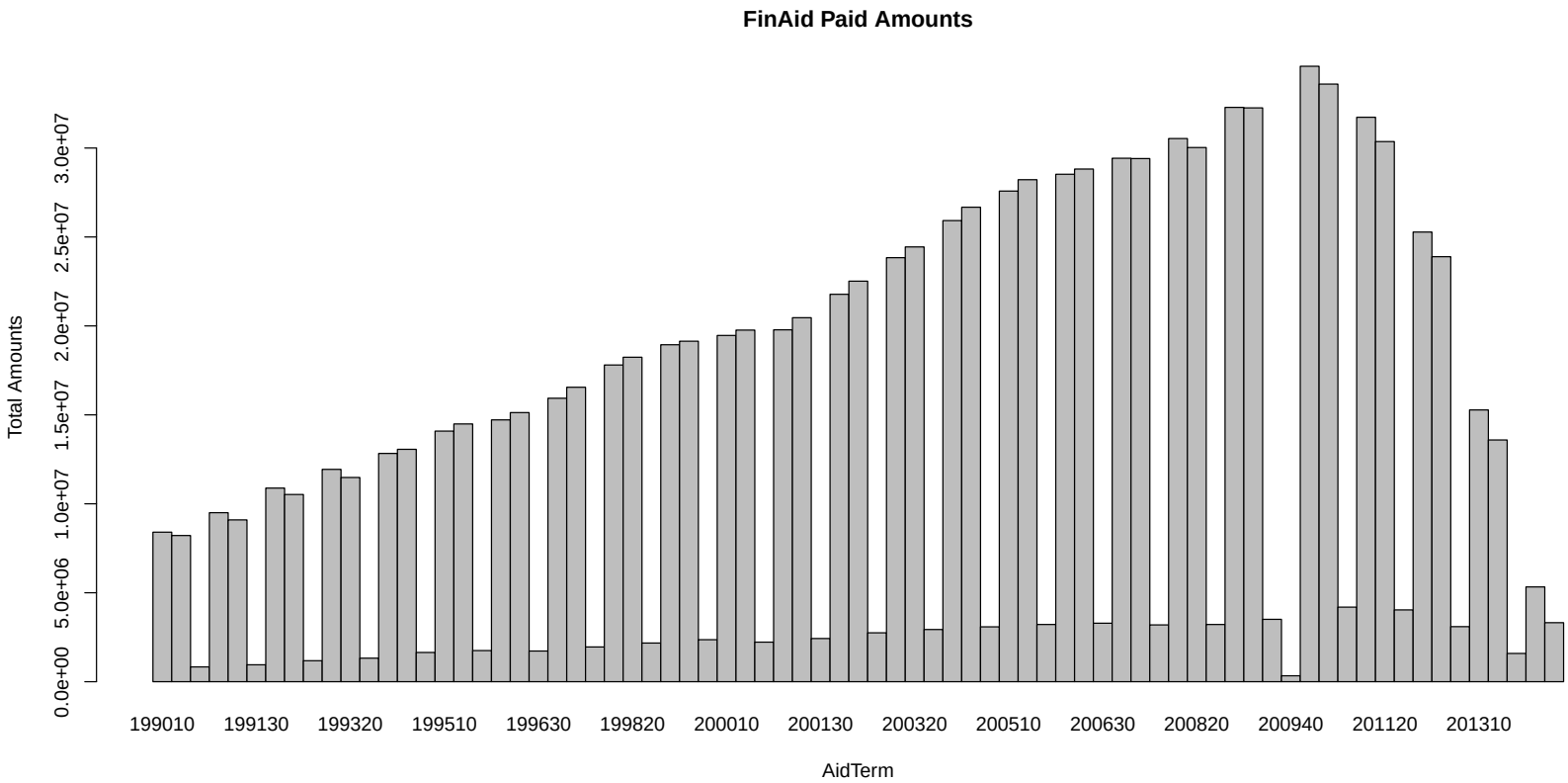
Taking into account the inflation



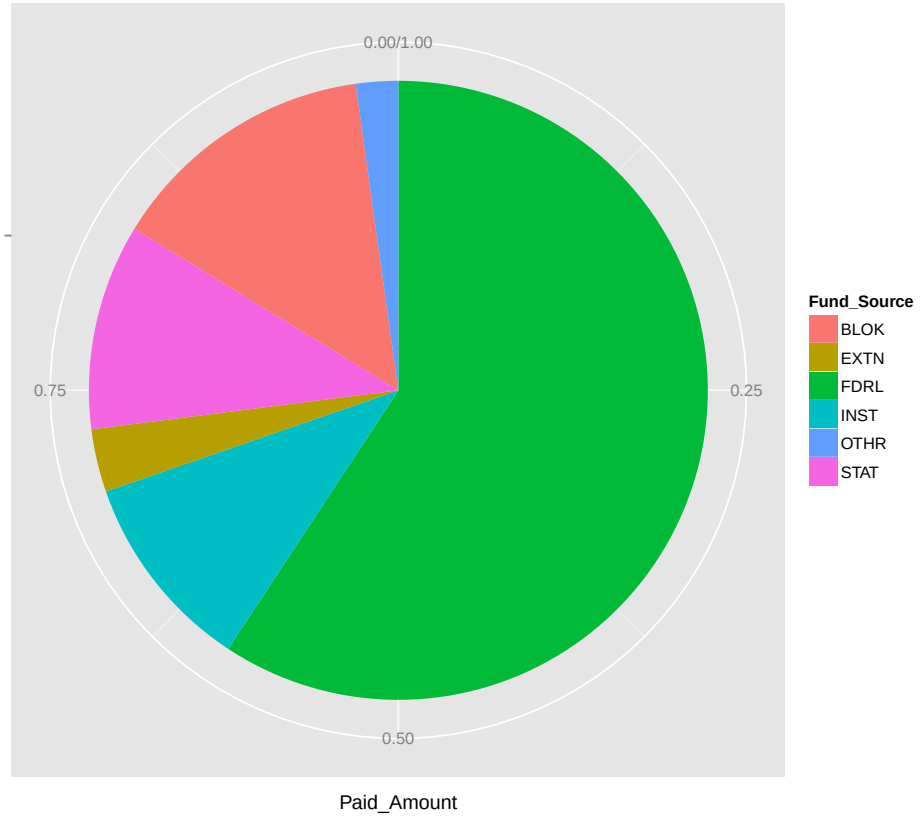
Look at it in semester terms



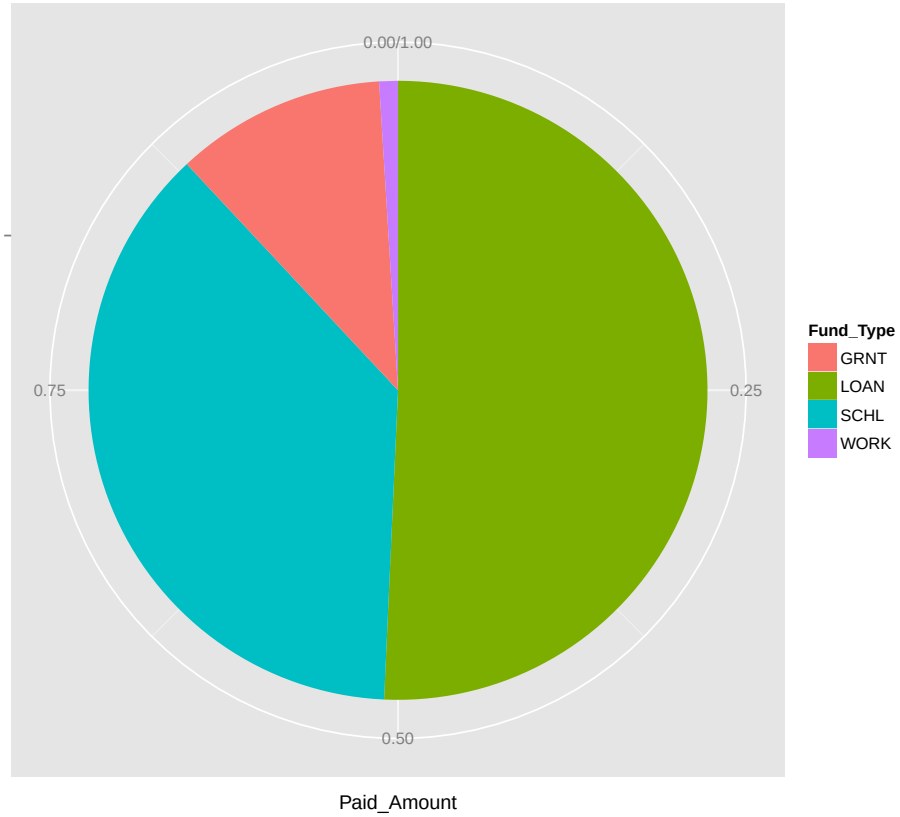
And amounts


FinAid Paid Amounts
Total Amounts
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
199010
199130
199320
199510
199630
199820
200010
200130
200320
200510
200630
200820
200940
201120
201310
AidTerm

Sources of funding



Types of funding



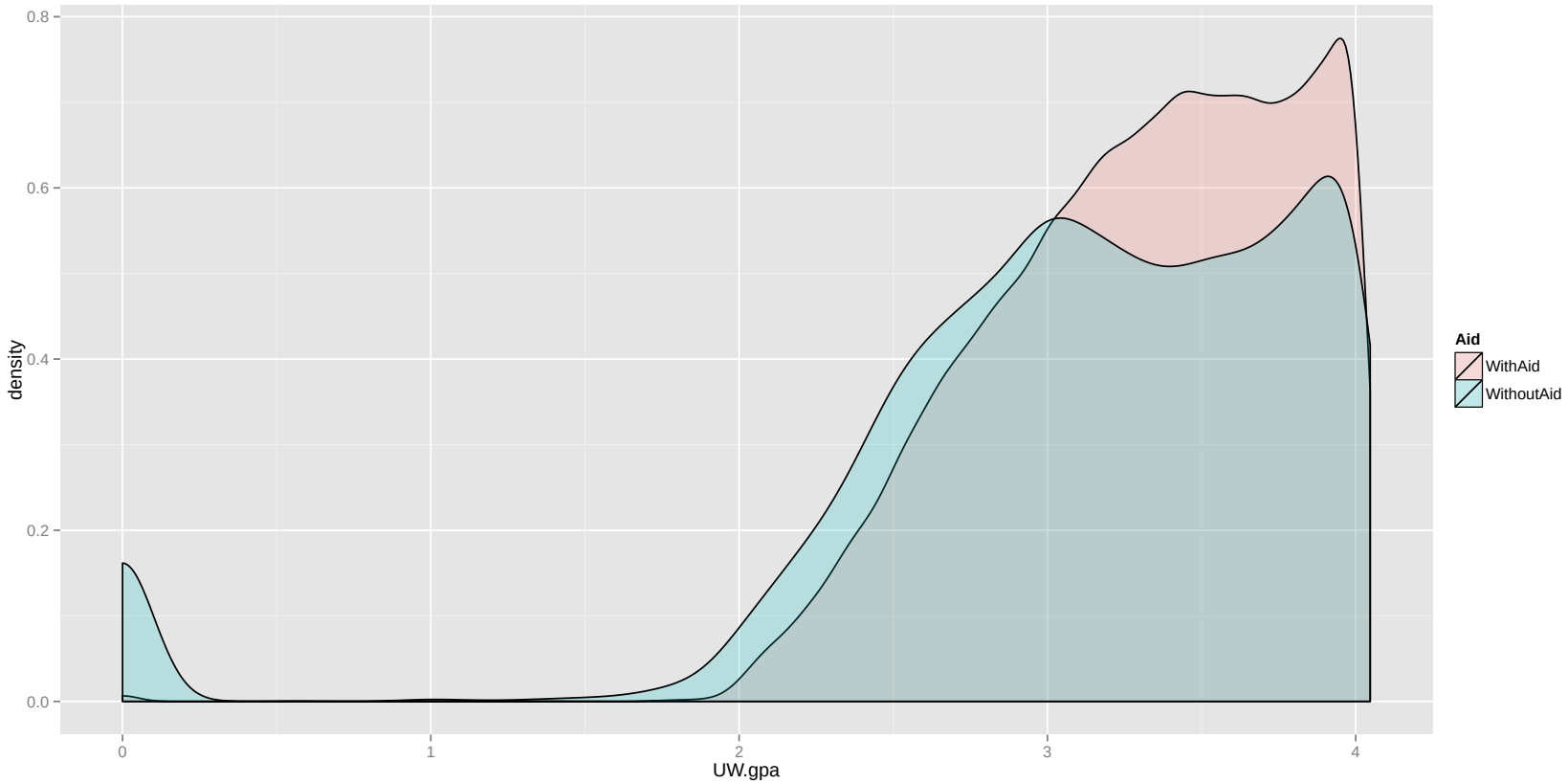
Students that have a 0 GPA

- 455 students. 47 of them have received Financial Aid.
- Or in another way, 47 of 44439 (0.106%) FinAid students withdrew.
- Among these 47, the most received $22252 (Scholarship and Loan), the least received $35 (T&F Reduction).
- Totally $155351.43.

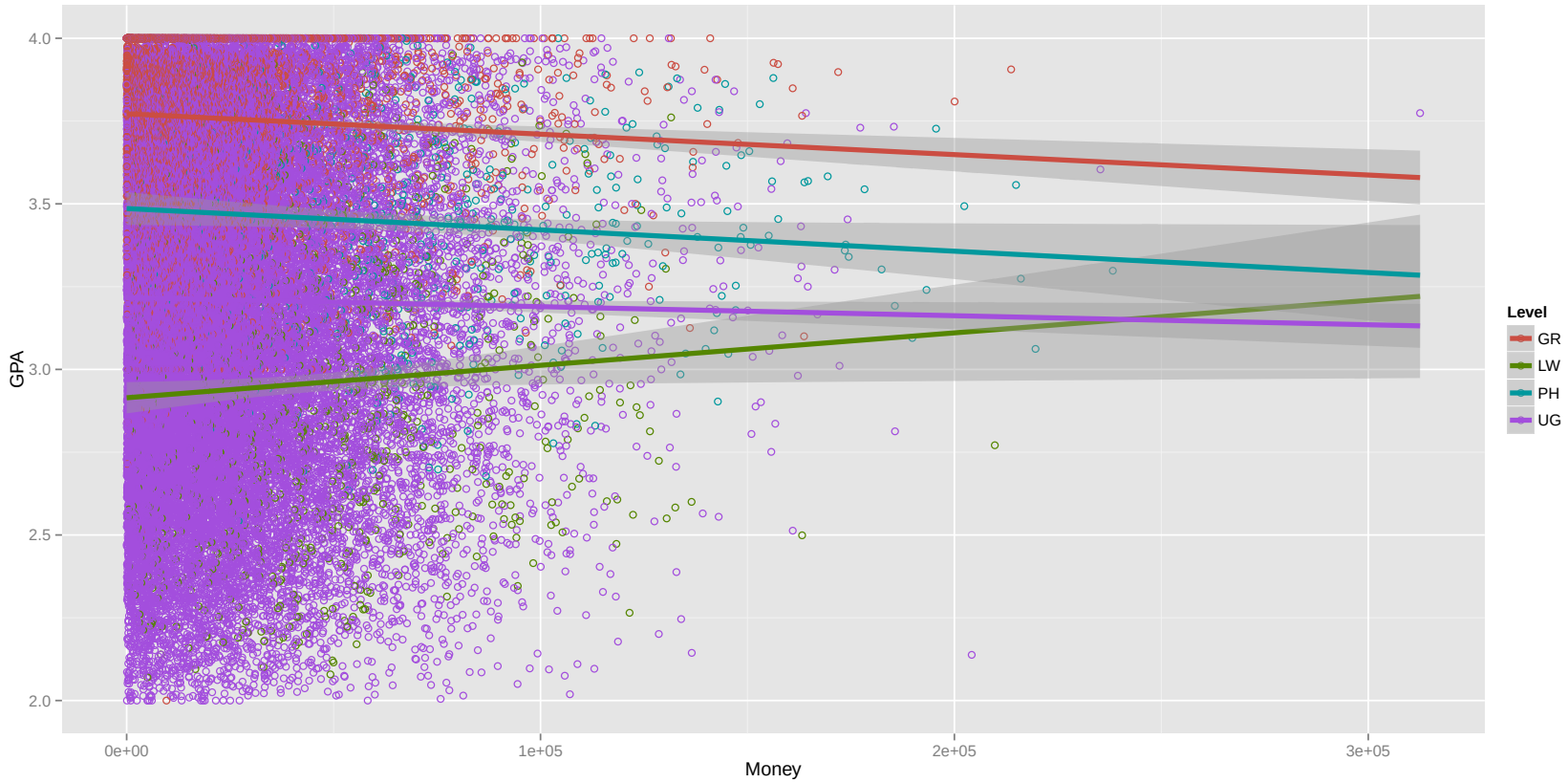
Students with a < 2.5 (Including 0) GPA

- 4697 students. 3169 of them have received Financial Aid.
- Or in another way, 3169 of 44439 (7.13%) FinAid students did not performe well.
- Among these 3169, the most received \$213708.31 (Loan and many Scholarships). Students received great amounts often have different degrees.
- Totally \$81595467.76.

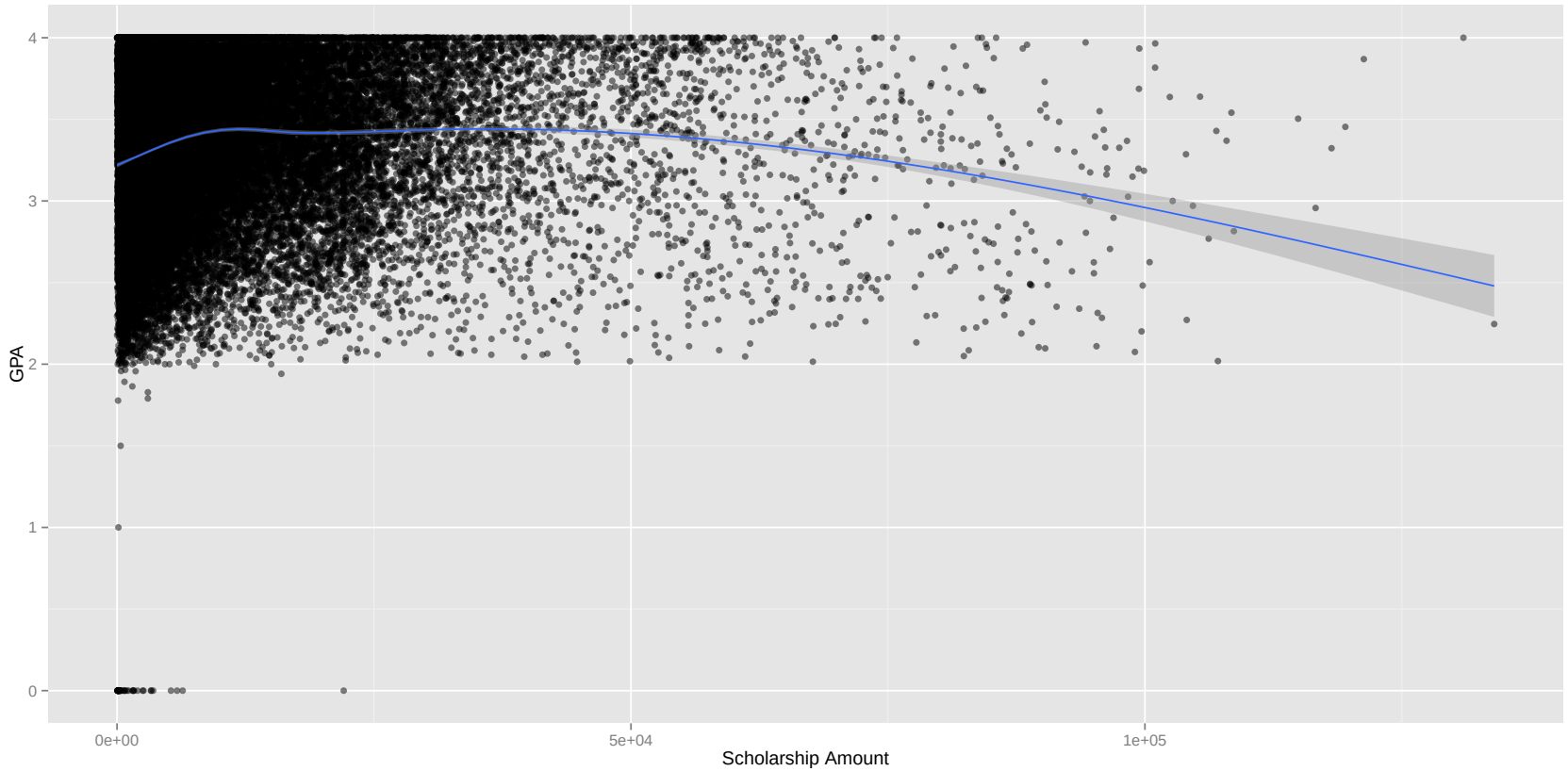
Money helps students perform better



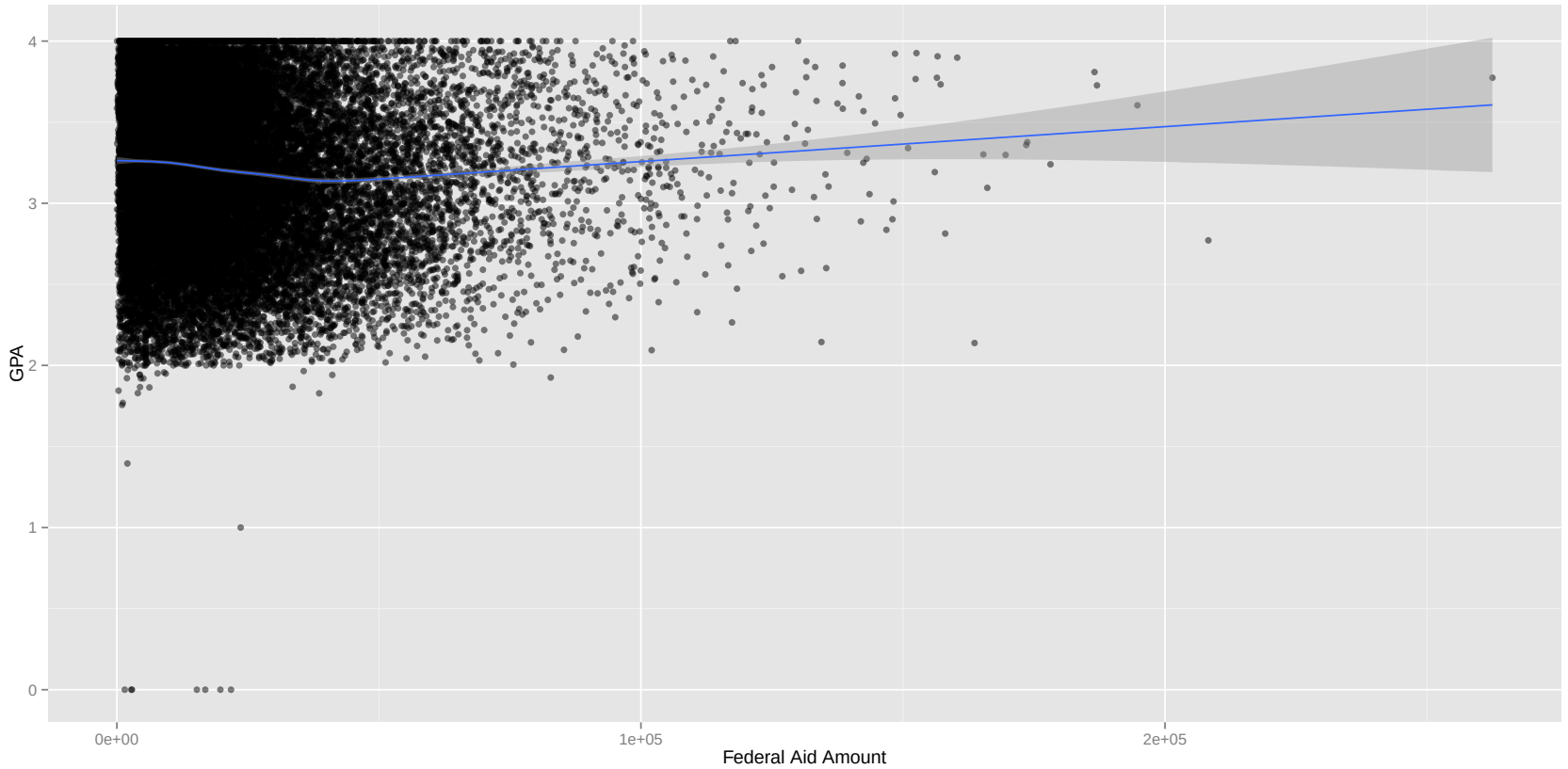
Depending on which level you are in



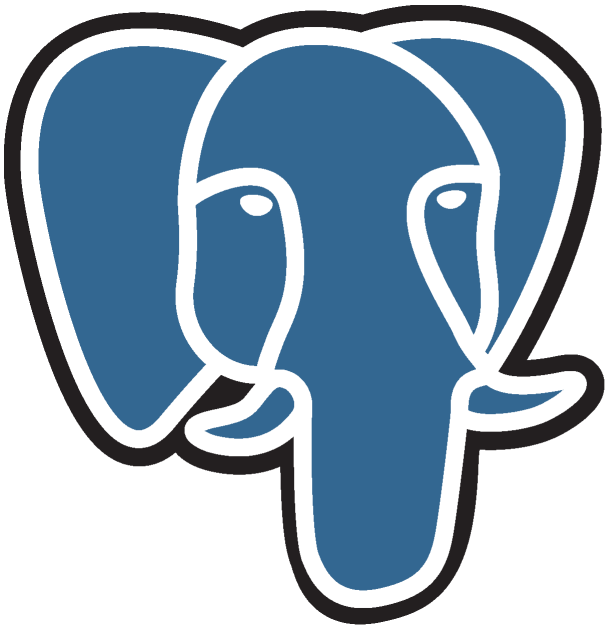
If we only look at Scholarship



If we only look at Federal Aid



PostgreSQL



PostgreSQL

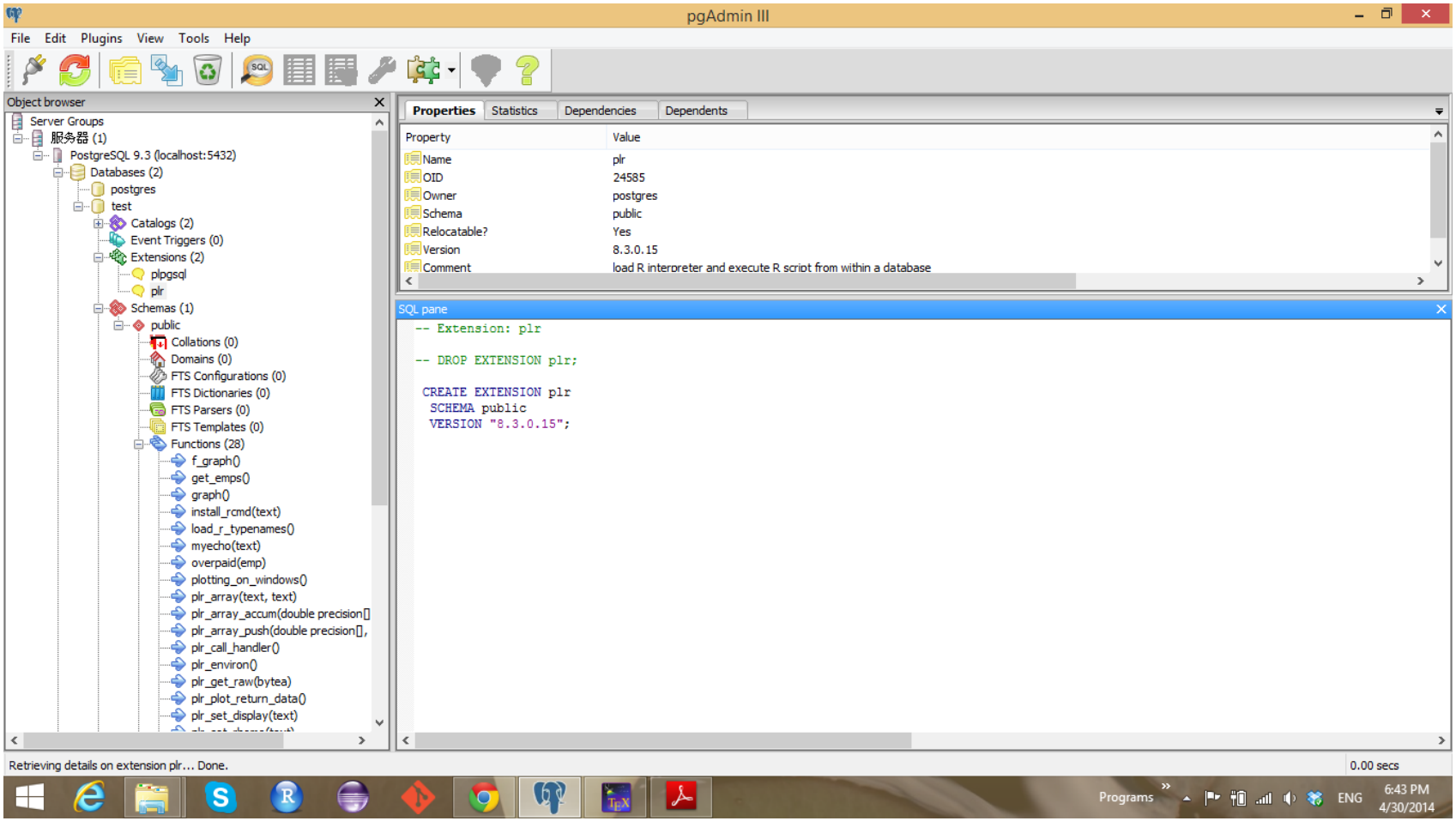
the world's most advanced open source database

Procedural Language for PostgreSQL

- PL/Java
- PL/Perl
- PL/php
- PL/Python
- And. . . our PL/R!

Created, updated and supported by one man . . . our hero Joseph E Conway http://www.joeconway.com/plr/

The graphical user interface



Why do we do this?

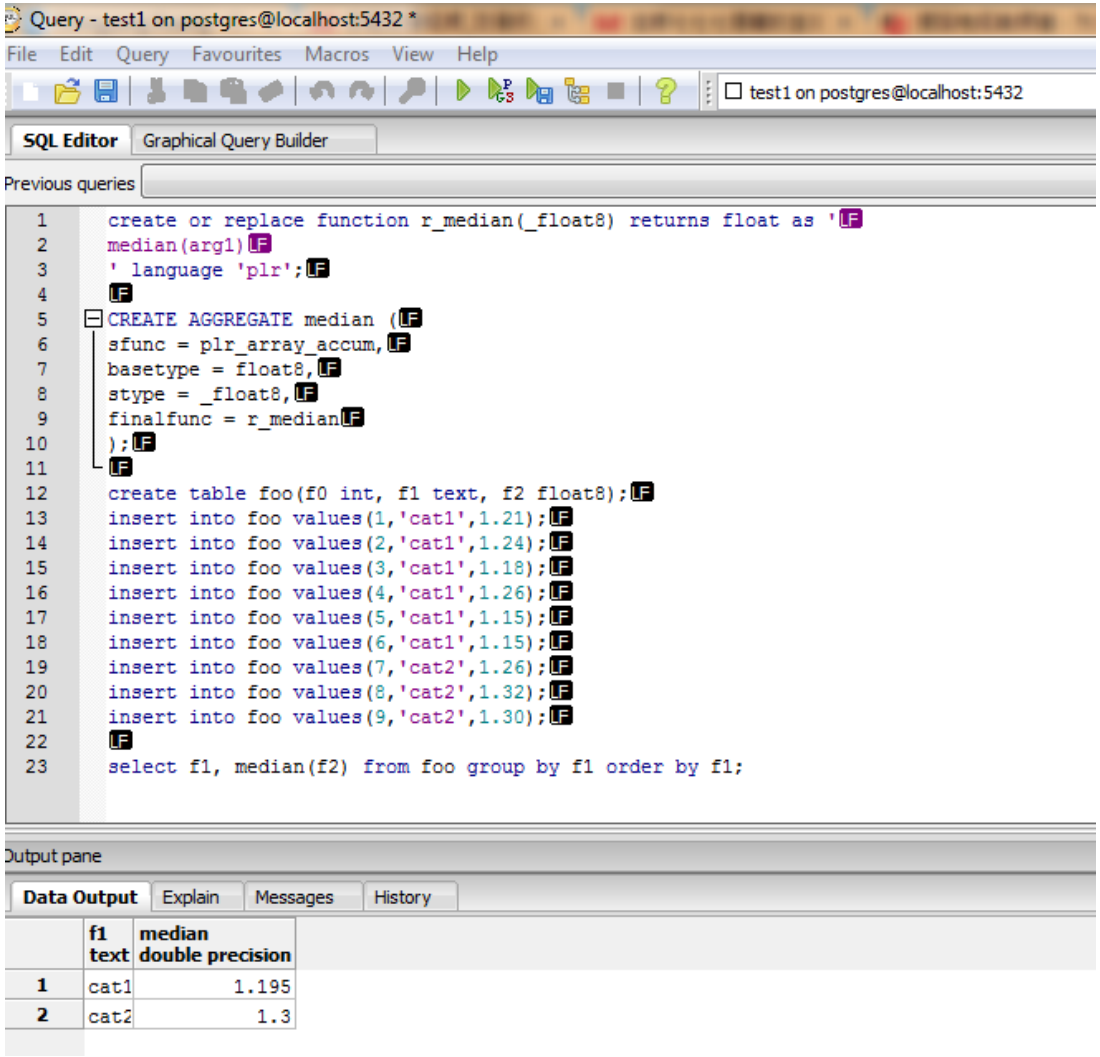
What SQL Can Do

- Count
- Sum, Average
- Max, Min
- First, Last
- . . . (not much left)

What SQL Can Not Do

- Median, Quantile
- Standard Deviation, Correlation
- Analysis of Variance
- (Generalized) Linear Regression
- All Kinds of Plotting
- Time Series Analysis
- Classification Tree, Random Forrest
- . . . basically, almost everything that R can do

Let's find the medians!

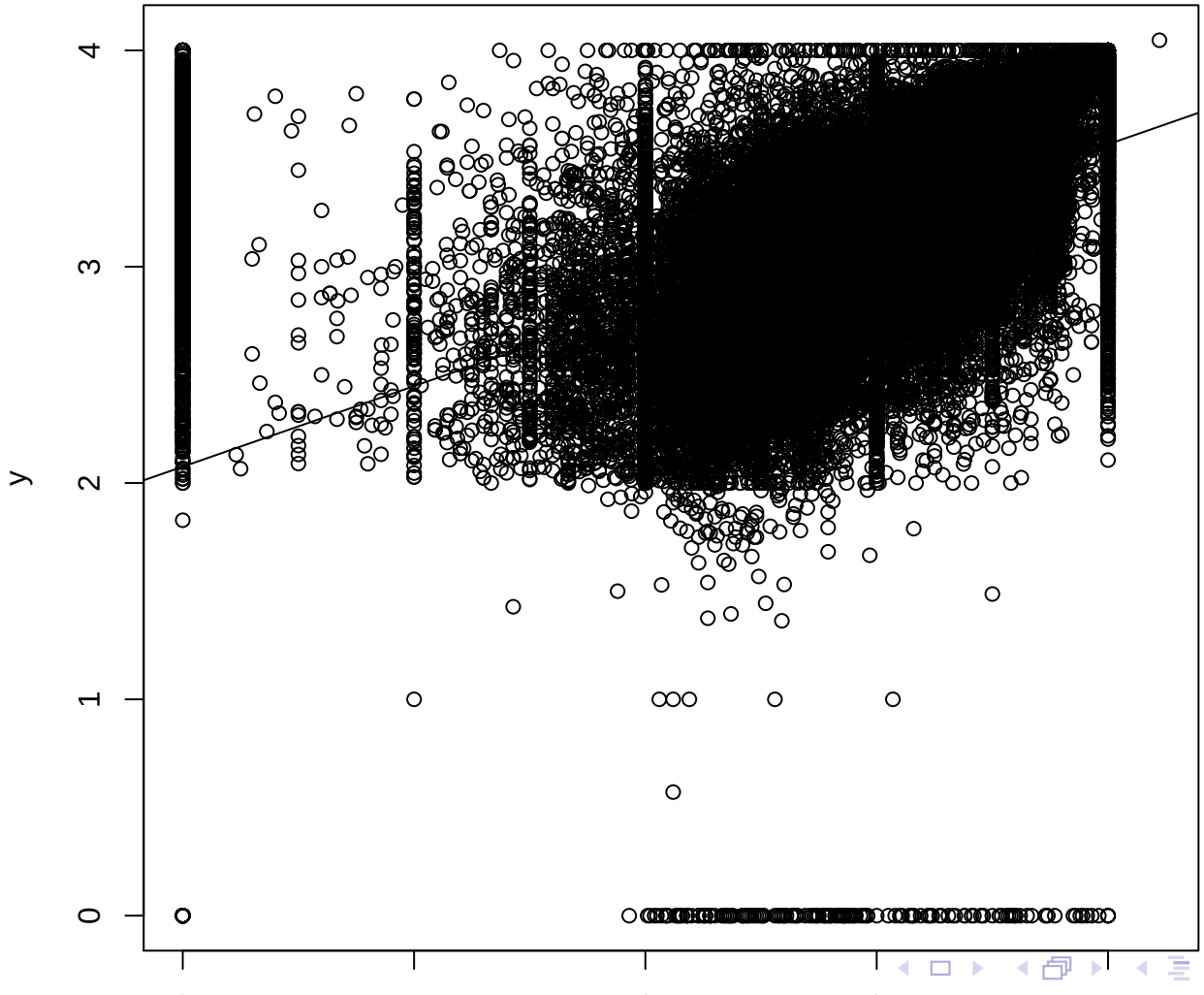


Coding is fun and easy

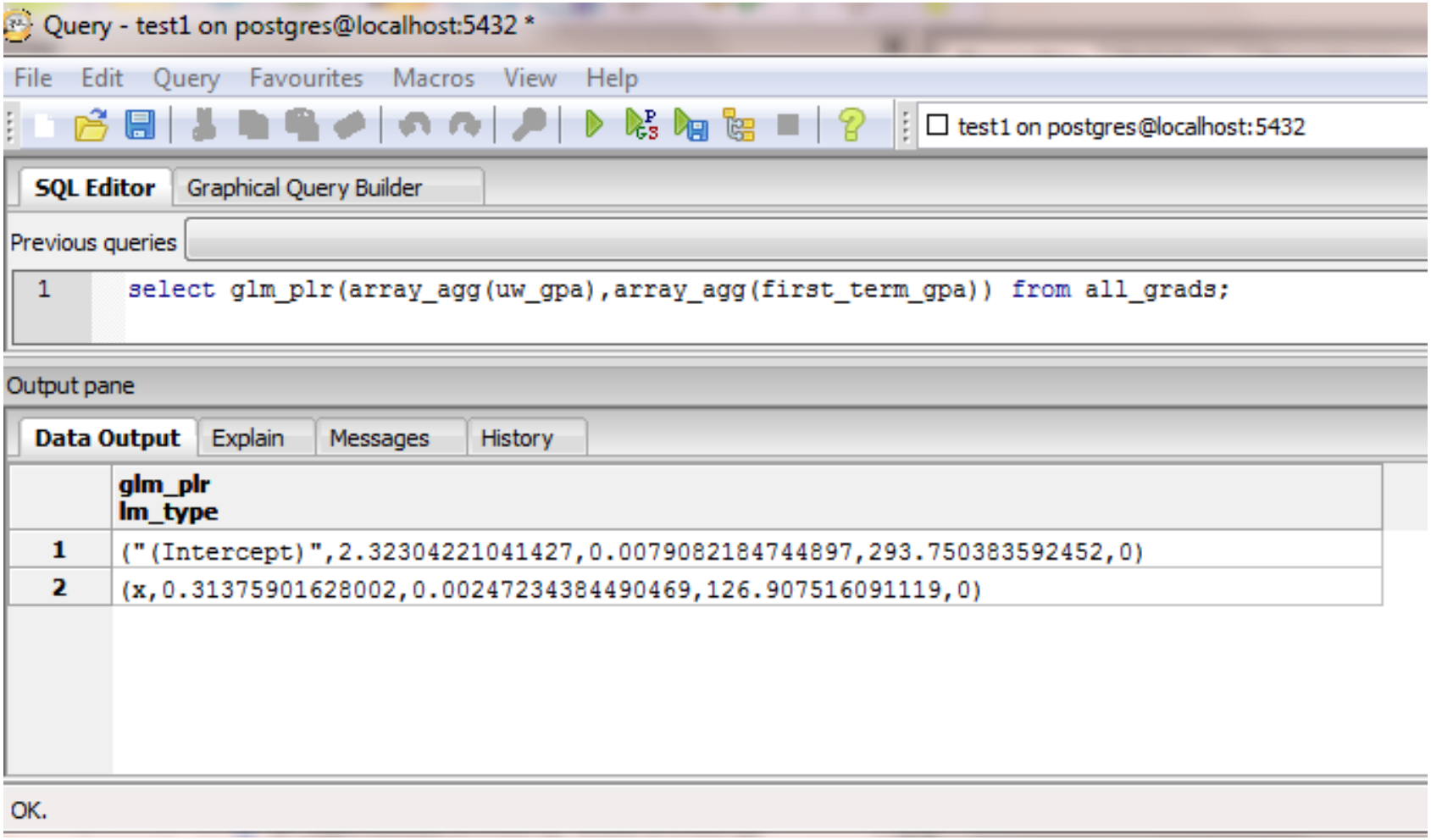
```
create or replace function boxplot(data float[])
returns text as
'
pdf("D:/boxplot.pdf");
boxplot(data);
dev.off;
print("done");
'
language 'plr' strict;

select boxplot(array_agg(uw_gpa)) from all_grads;
```

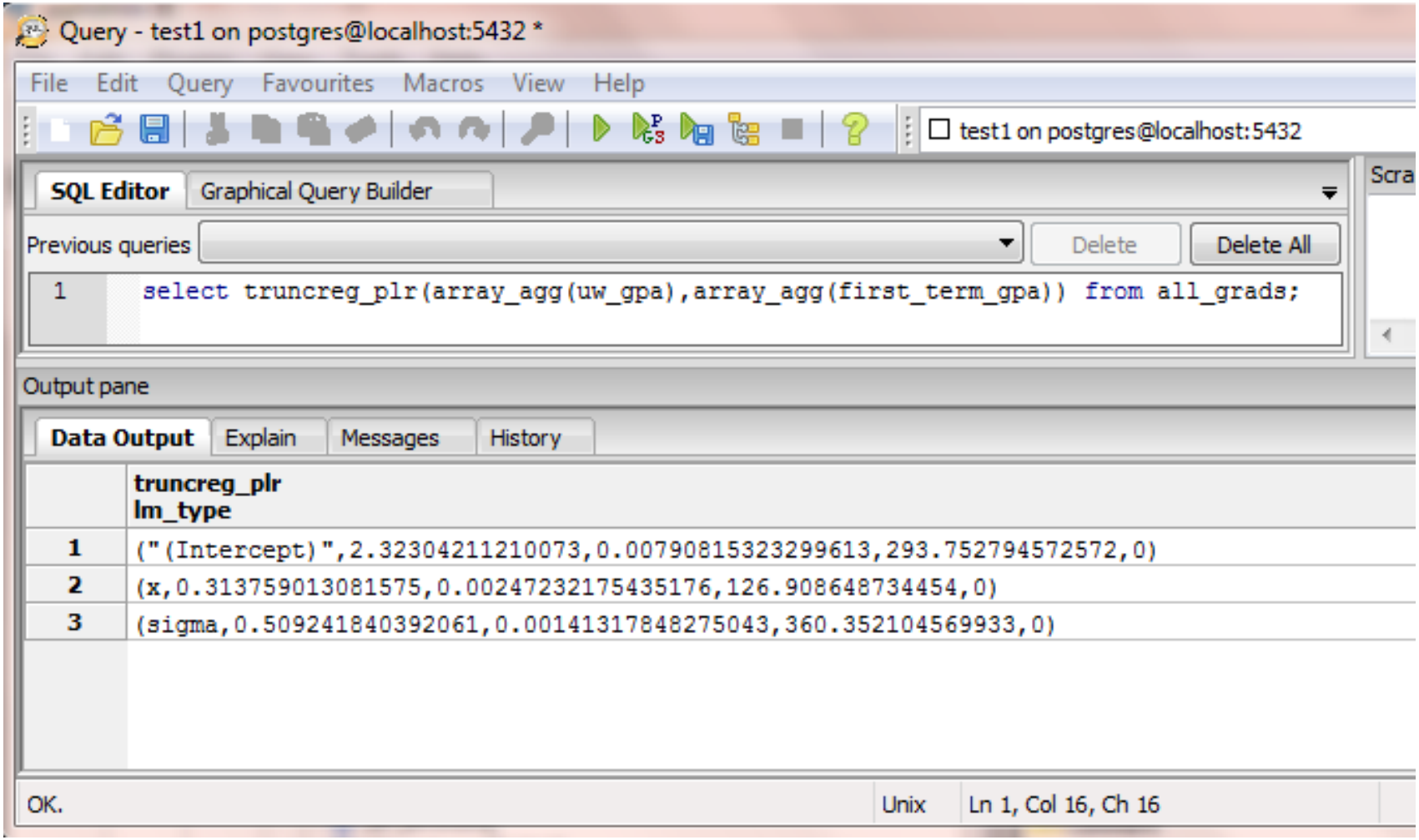
SQL can plot linear regression line now



Even generalized linear regression is possible



The truncated regression you saw before

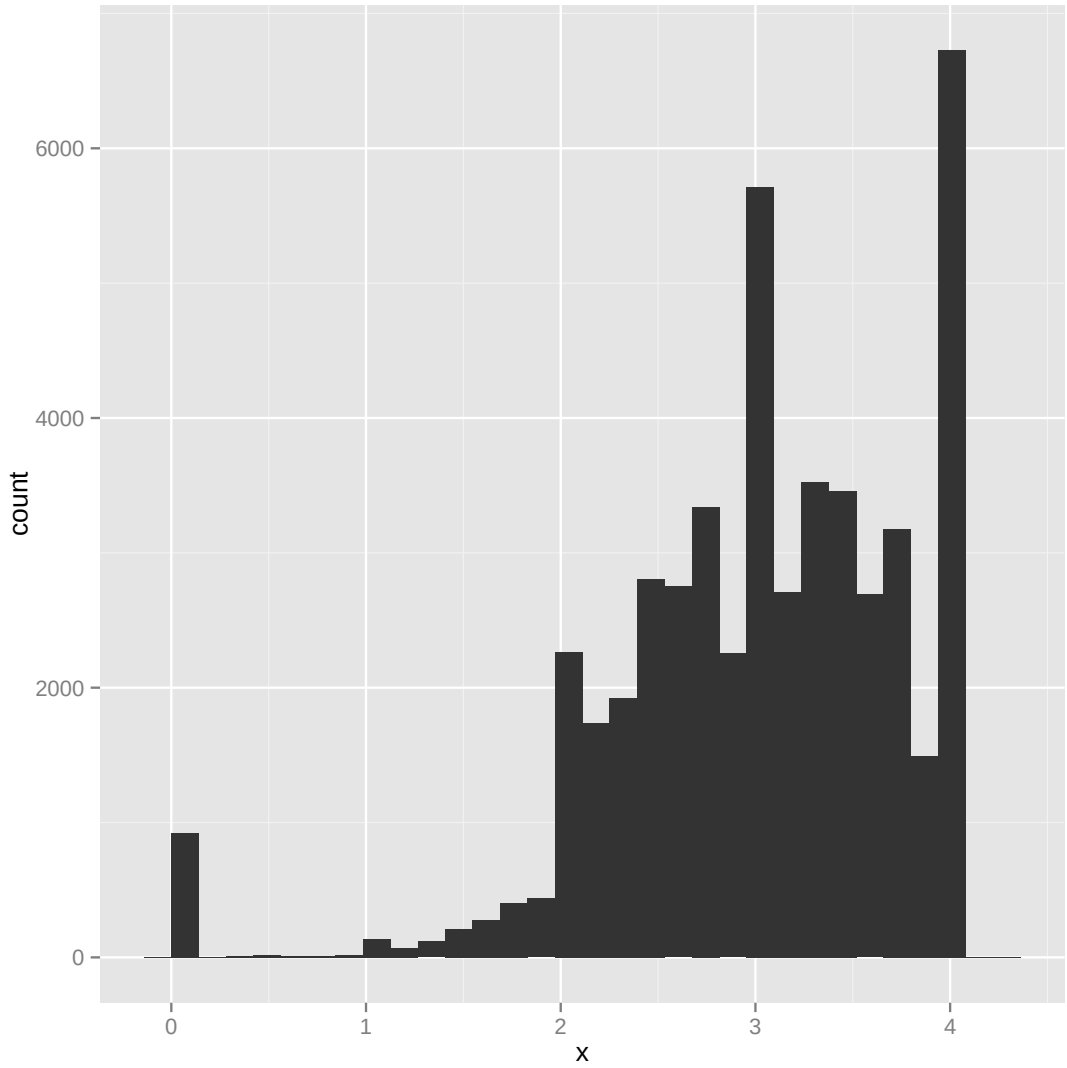


Installing R packages

```
CREATE TABLE plr_modules (
modseq int4,
modsrc text
);

INSERT INTO plr_modules
VALUES (0, 'library(Kendall)');
```

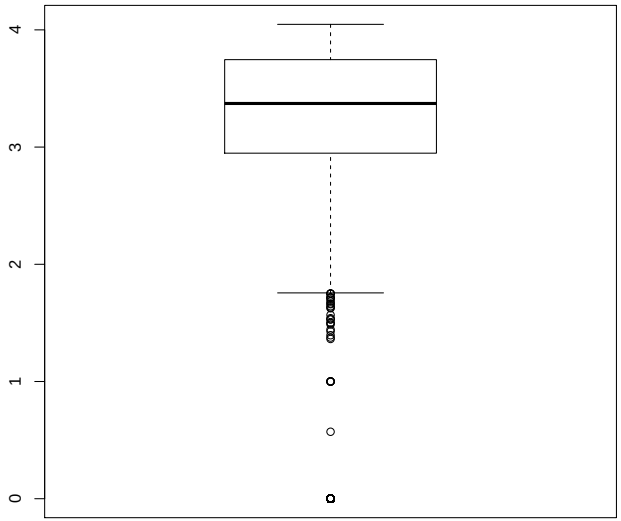
ggplot2 is working now!



Just do it!

Let's Run Some Examples!

- Correlation
- Boxplot
- Histogram by ggplot2
- Install a Package



Back to the future

Possible Future Works

- Dig deeper into the data ...
- Complete constructing and refining our R/SQL language.

Finally. . .

- Thank You All For Listening To Us!
- Thank You, Professor Douglas!