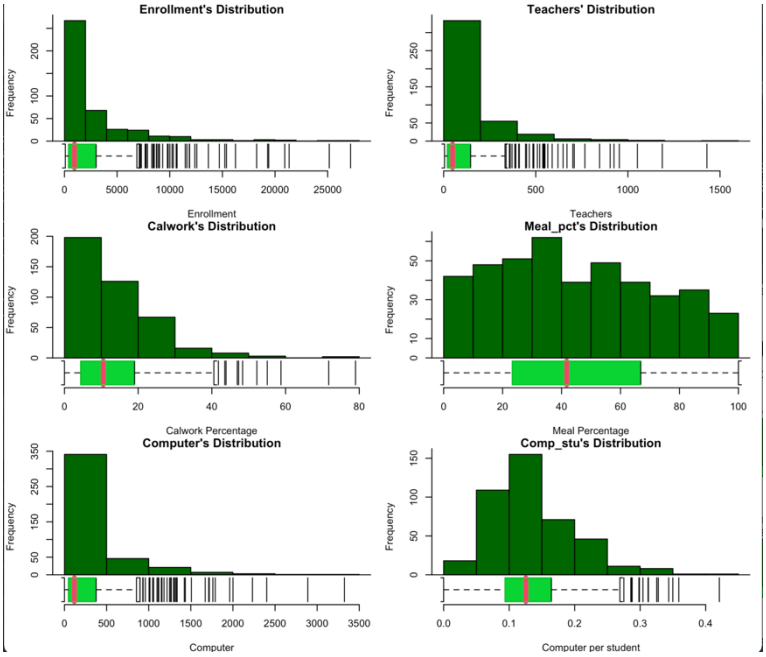
STAR - Research Question (Fernandes Michael)

The object of my study is the analysis of the STAR dataset. It contains some data of 420 school districts and is described by 14 different variables. There are no missing data.

Descriptive

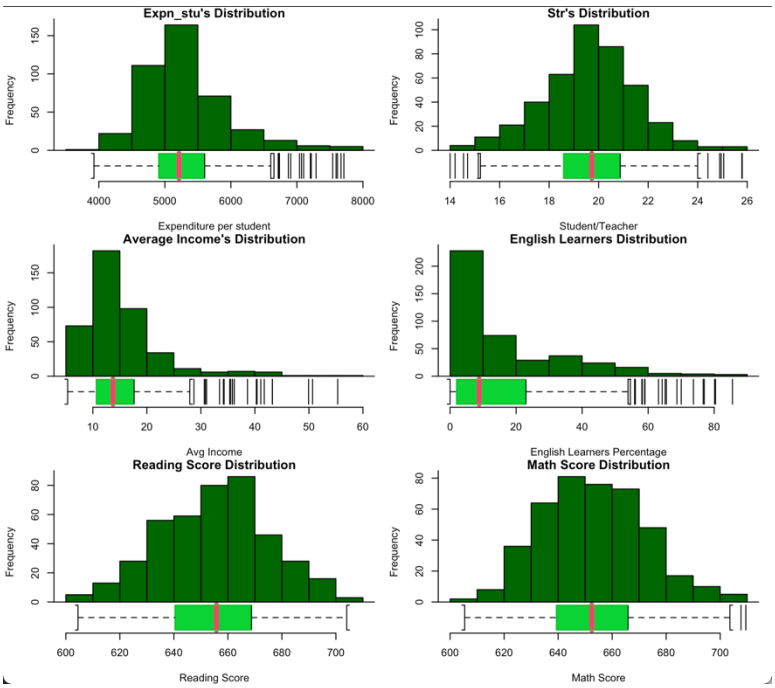
Using *summary(star)* function and some graphs I am going to briefly describe the variables of our dataset.

- **enrl_tot:** it describes the number of students in the district. The range goes from 81 to 27176. The interquartile range goes from 379 to 3008, while its median and mean are 950,5 and 2628,8. Thus, it suggests us a positive skewness (2.86), since median <mean. The variation coefficient is 1.39.
- **Teachers:** number of teachers in the district. Values range goes from 4,85 to 1429.Itss interquartile range goes from 19.66 to 146.35. Since it is highly correlated with the number of enrollment, we can also imagine median < mean (48,56 and 129,07) with a positive skewness (2,92). The variation coefficient is 1.36.

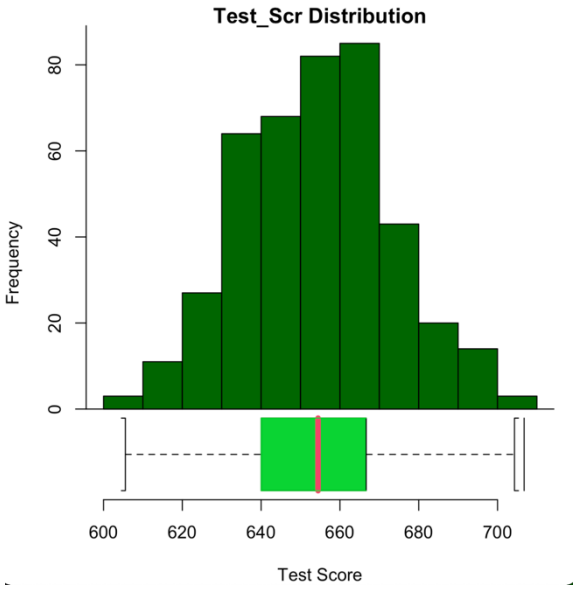


- **Str:** Student to Teacher ratio. Range (14,25.80) and interquartile range from 18.58 to 20.87. Its median and mean are almost coincident (19.72 and 19.64). In fact, skew is approximately near to 0. Its variation coefficient is also pretty small. (0.09)
- **Calw_pct :** it indicates the students' percentage qualifying for a public assistance program. Values range goes from 0 to 78.99, while interquartile range it's from 4.39 to 18.98. The median is 10.52, while its mean is 13.24, skew=1.68 and also its variation coefficient is low, compared to enrollment and teacher. (0.81)
- **Meal_pct:** it indicates the proportion of students qualifying for reduced lunch price. It is highly correlated with calw_pct but its distribution differs from a scatter point of view. Min and max are respectively 0 and 100, while 1st and 3rd quartile are 23.28 and 66.86. The median is pretty near the mean value (41.75 and 44.71), so we can expect a pretty low skew absolute value(0.18).
- **Computer:** the absolute number of computers in the district. Range from 0 to 3324, while 46 and 375 are the 1st and 3rd quartile. We also have a positive skewness, since the median is pretty smaller than the mean (117.5 and 303.4).
- **Comp_stu:** it measures the proportion of computers over the enrollment in the district. It is relevant for our analysis because we can make comparisons among the districts. Its values are included between 0 and 0.42, while the interquartile range is between 9,38% and 16,45%. Its median and mean values are 12,55% and 13,59% and the skew is 0.92. Its variation coefficient is 0.45

- **Expn_stu:** Expenditure per student. Its range goes from 3926 to 7712. 1st and 3rd quartile are 4906 and 5601, while the median and mean values are 5215 and 5312. Skew is 1.06, while VC is 0.11.
- **Avginc:** Average income in the district. It is highly correlated with calw_pct and meal_pct, since it is a relevant parameter to be eligible for the benefits. Its range goes from 5.33 to 55.33(in thousands of dollars), while the interquartile range goes from 10,64 to 17,63. It is pretty positively skewed(2.21), its median and mean value are respectively 13,73 and 15,32. Its variation coefficient is 0.44



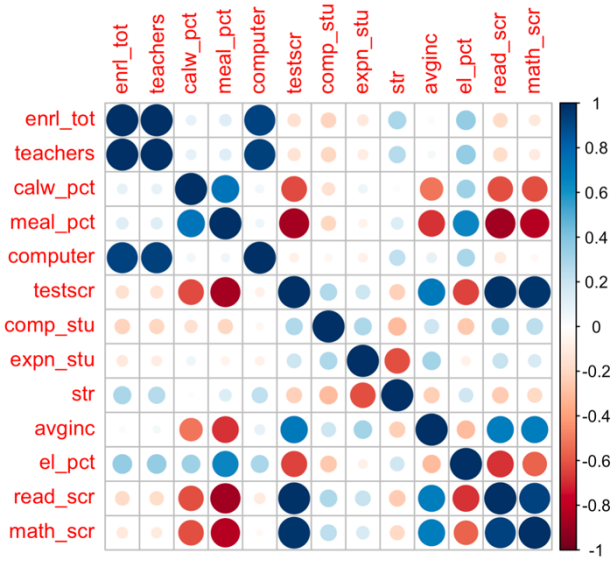
- **el_pct :** Percentage of English learners. Range from 0 to 85.54, median 8.778 and mean 15.768. Its skew value is 1.42 and its variation coefficient is 1.08.
- **Mathscr:** District's average Math score. Its range is 605.4-709.5, while the 1st and 3rd quartiles are 639,4 and 665,9. Median and simple mean are 652.5 and 653.3. Skew value is 0.25, while variation coefficient is 0.03
- **Readscr:** District's average Reading score. Its range is 604.5-704, while the 1st and 3rd quartiles are 640,4 and 668,7. Median and simple mean are 655.8 and 655. Skew value is -0.06, while the variation coefficient is 0.03.
- **Testscr:** It is the variable we want to study. It is calculated as an average between math and reading score. .Its range is 605.5-706.8, while the 1st and 3rd quartiles are 640 and 666,7. Median and simple mean are 654.5 and 654.2. Skew value is 0.09, while variation coefficient is 0.03.



Comparative analysis

What is the relationship between STAR's variables? To answer this question I simultaneously used the correlation matrix that you can see below and some scatterplots graphs, in order to analyze some trends that could be useful for the predictive phase, during regression. In particular, my goal is to identify some significant correlations with test score, but also some potential multicollinearity issues in the Multiple Linear Regression Model.

-We can see a great positive correlation among *enrl_tot, teachers* and *computer*, and I suppose that's because as long as enrollments increase, the school must provide new computers and hire new teachers.
-Another group could be identified among *calw_pct, meal_pct* and *avginc*, because they all indicate income indicators. The most correlated ones are calw_pct and meal_pct(r=0,739) with high risk of multicollinearity, then *meal_pct* and *avginc* are also negatively correlated (r= -0,684). *Calw_pct and avginc* have a r=-0,513. *Avginc* is negatively correlated to the other two variables because these public assistance programs are eligible for indigent students.



-A variable that caught my attention is also *el_pct*. It is correlated with *meal_pct*(r=0,653), *tstscr*(r= -0.644) and of course also with *read_scr*(r=-0,69) and *math_scr*(r=-0,569). What I think the data are telling us here is that a high percentage of English learners students make the score drop, especially in the read_scr, since we have a higher absolute correlation than in math_scr and we can logically suppose that they could have more difficulties than other students.

	enrl_tot	teachers	calw_pct	meal_pct	computer	testscr	comp_stu	expn_stu	str	avginc	el_pct	read_scr	math_scr
testscr	-0.154	-0.145	-0.627	-0.869	-0.074	1.000	0.271	0.191	-0.226	0.712	-0.644	0.982	0.979

-*Testscr* is the variable we want to predict in our Regression Model, but first we can analyze the parameters that influence it. Of course, it is highly explained by read_scr and math_scr since it is calculated as a simple mean between the two. Then it is pretty much correlated with our "income group": *meal_pct*(-0,869), *avginc(0,712),calw_pct(-0,627).* What we can see here is that districts with higher Income, or less Meal and CalWork percentage, averagely do better in the test. It is pretty important because we could analyze the impact of these public assistance programs over time, but it is also an incentive to understand about equal opportunity issues. We also have to consider multicollinearity among the three variables, so we'll have to choose at most two.
As I already said, el_*pct* learners surely influence the score, but also *comp_stu* and *str* could be interesting.

Predictive analysis

Our goal is to describe which phenomena can influence the students' performance in the test. We should also check the influence of *CalWorks* and *Student-Teacher Ratio* on *testscr.* I performed many simple regressions, with testscr as a regressand and the other variables as regressors, then I chose the more adequate ones for the Multiple Regression, using both a *stepwise process* and a *backward elimination.*

- testscr~meal_pct:

Testscr=681.44-0.61(meal_pct)

We can confirm our high correlation hypothesis between the two variables. We have significant coefficients t values (corresponding to low p-values), so we can reject the null hypothesis of linear correlation's lackness. Also, the determination coefficient is pretty high(R^2 adj=0.754). The F-statistic is also pretty large so our global test makes us deny the null hypothesis of linear correlation's lackness.

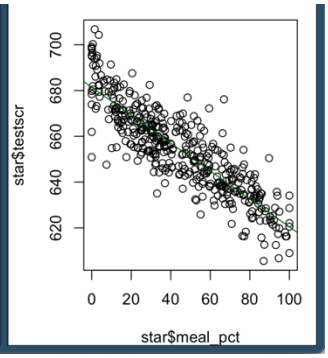
```
Call:
lm(formula = testscr ~ meal_pct, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-30.540  -6.038  -0.647   6.245  35.543

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 681.43952    0.88943  766.16 <0.0000000000000002
meal_pct     -0.61029    0.01701  -35.87 <0.0000000000000002

(Intercept) ***
meal_pct    ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.447 on 418 degrees of freedom
Multiple R-squared:  0.7548,    Adjusted R-squared:  0.7542
F-statistic:  1286 on 1 and 418 DF,  p-value: < 0.00000000000000022
```



- testscr~enrollment:

Testscr=656.13-0.00075(enrl_tot)

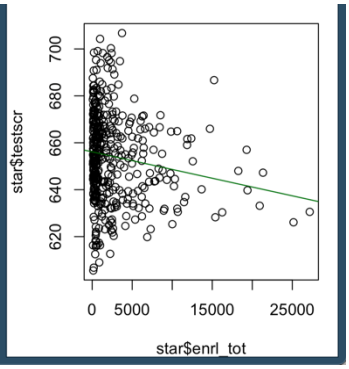
The t-value is pretty low, so we can consider the coefficients pretty reliable, but the determination coefficient R^2 is very low, so it cannot be considered a good model to explain our *testscr.*

```
Call:
lm(formula = testscr ~ enrl_tot, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-50.475 -13.308  -0.337  12.773  53.415

Coefficients:
              Estimate  Std. Error t value
(Intercept) 656.1275703   1.1083957 591.961
enrl_tot     -0.0007498   0.0002353  -3.186
                        Pr(>|t|)
(Intercept) < 0.0000000000000002 ***
enrl_tot                 0.00155 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.85 on 418 degrees of freedom
Multiple R-squared:  0.02371,   Adjusted R-squared:  0.02138
F-statistic: 10.15 on 1 and 418 DF,  p-value: 0.001549
```



- tescr~str:

Testscr=698.93-2.28(str)

str variable, like enrollment, doesn't have a high R^2, but the coefficients are definitely more significantly different from 0. As we have seen before it has not a pretty high correlation with other variables, but we could try to consider it in our Multiple Regression model to see if we have some improvements.

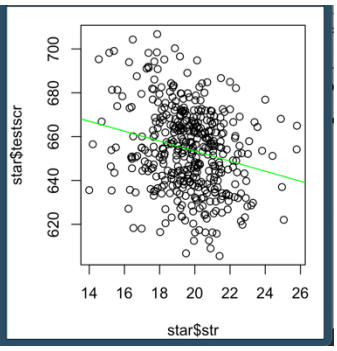
```
Call:
lm(formula = testscr ~ str, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 698.9330     9.4675  73.825 < 0.0000000000000002 ***
str          -2.2798     0.4798  -4.751          0.00000278 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,   Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 0.000002783

>
```



- testscr~computer:

Testscr=655.12-0.00318(comp)

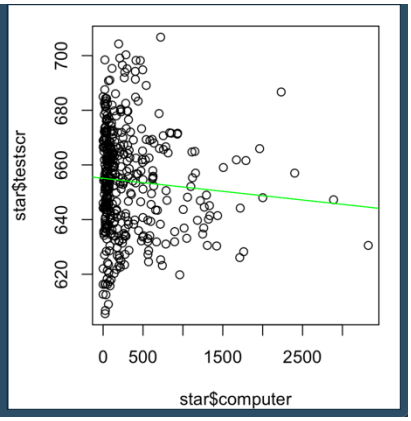
This variable also has a low R-squared, since the residual standard error is 19.02. In fact, the slope is pretty horizontal and we can expect a low Regression Sum of Squares. The coefficient is also not particularly significant and the global test doesn't give us a sufficiently low p-value to refuse our null independence hypothesis.

```
Call:
lm(formula = testscr ~ computer, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-49.493 -13.902   0.061  12.747  53.923

Coefficients:
             Estimate Std. Error t value
(Intercept) 655.122302   1.126888 581.355
computer     -0.003183   0.002106  -1.512
                       Pr(>|t|)
(Intercept) <0.0000000000000002 ***
computer                0.131
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.02 on 418 degrees of freedom
Multiple R-squared:  0.005437,  Adjusted R-squared:  0.003058
F-statistic: 2.285 on 1 and 418 DF,  p-value: 0.1314
```



- testscr ~teachers:

Testscr=656.05-0.015(teachers)

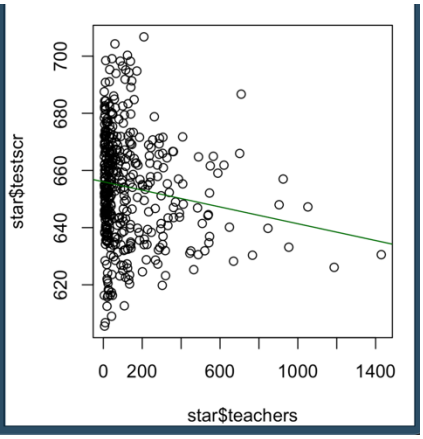
Our low correlation value confirms our assumptions. The coefficients test brings us to deny our null hypothesis in both cases, and also our global test. However, the R^2 is very small (0.02).

```
Call:
lm(formula = testscr ~ teachers, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-50.408 -13.331  -0.245  12.850  53.760

Coefficients:
             Estimate Std. Error t value
(Intercept) 656.052252   1.117751 586.940
teachers     -0.014688   0.004907  -2.993
                        Pr(>|t|)
(Intercept) < 0.0000000000000002 ***
teachers                0.00292 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.87 on 418 degrees of freedom
Multiple R-squared:  0.02098,   Adjusted R-squared:  0.01864
F-statistic: 8.959 on 1 and 418 DF,  p-value: 0.002925
```



- testscr~calw_pct:

Testscr=667.97-1.04 (calw_pct)

In the graph, I highlighted the difference between two groups in the variable. We can see that districts with at most 3% CalWork beneficiaries didn't have a testscore below the mean, while districts over 37-38% had low test scores, below the mean.

We already noticed that the correlation with CalWork percentage was pretty high, but when we see this graphically, it makes a completely different effect. In fact, the R^2 is 0.39 and the p-values are sufficiently small to confirm our y-dependence hypothesis.

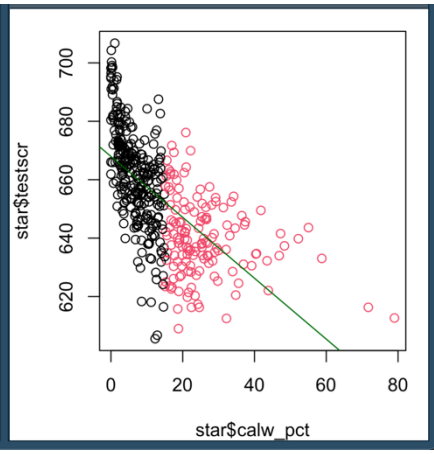
```
Call:
lm(formula = testscr ~ calw_pct, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-49.573  -9.583   0.239   9.216  39.902

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 667.96786    1.10949  602.05   <2e-16 ***
calw_pct     -1.04267    0.06339  -16.45   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.86 on 418 degrees of freedom
Multiple R-squared:  0.3929,    Adjusted R-squared:  0.3915
F-statistic: 270.6 on 1 and 418 DF,  p-value: < 2.2e-16

>
```



- testscr ~cmp_stu:

Testscr=643.36- 79.41(comp_stu)

The R-squared is higher than teachers, enrollment and computers. In fact, we can see a low positive correlation. Districts with a ratio bigger than 0.28 didn't perform badly. They have scored over the mean, while the whole point cloud seems to have an upward trend.

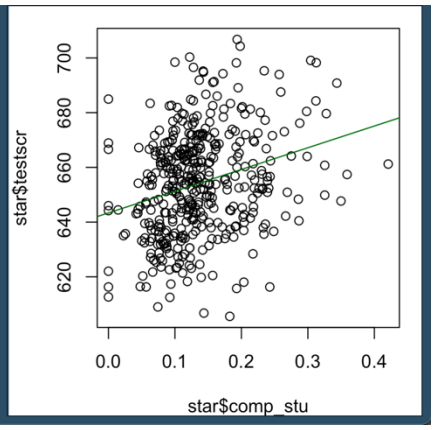
```
Call:
lm(formula = testscr ~ comp_stu, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-52.303 -13.880  -0.259  12.649  48.013

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   643.36       2.08 309.274      < 2e-16 ***
comp_stu       79.41      13.81   5.749 0.0000000173 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.36 on 418 degrees of freedom
Multiple R-squared:  0.07328,   Adjusted R-squared:  0.07106
F-statistic: 33.05 on 1 and 418 DF,  p-value: 0.00000001732

>
```



- testscr~expn_stu:

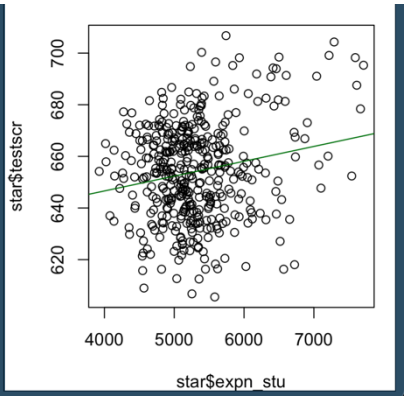
Testscr=623.62+0.0057(expn_stu)

```
Call:
lm(formula = testscr ~ expn_stu, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-50.146 -14.206   0.689  13.513  50.127

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 623.616497   7.719660  80.783   < 2e-16 ***
expn_stu      0.005749   0.001443   3.984 0.0000799 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.72 on 418 degrees of freedom
Multiple R-squared:  0.03659,   Adjusted R-squared:  0.03428
F-statistic: 15.87 on 1 and 418 DF,  p-value: 0.00007989
```



In this point cloud, we cannot recognize a clear pattern with testscore. However, we have a small group (expn > 6800) that clearly performed well in the test. It could also be casual, or influenced by the district's average income. The R^2 is not pretty high (0.034), but the coefficients are pretty significant, since the p-value is low.

- testscr~avginc:

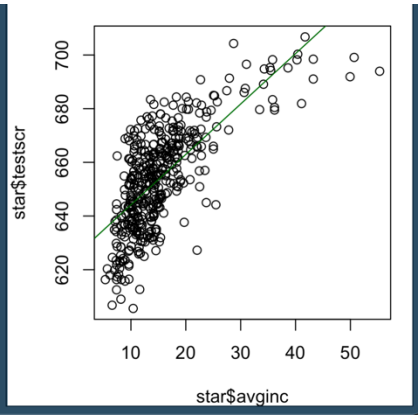
Testscr=625.38+1.878(avginc)

```
Call:
lm(formula = testscr ~ avginc, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-39.574  -8.803   0.603   9.032  32.530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 625.3836     1.5324  408.11   <2e-16 ***
avginc        1.8785     0.0905   20.76   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.39 on 418 degrees of freedom
Multiple R-squared:  0.5076,    Adjusted R-squared:  0.5064
F-statistic: 430.8 on 1 and 418 DF,  p-value: < 2.2e-16
```



It is clearly high correlated with testscr, but also in this case we can see that the point cloud explains two different trends. We have the cloud below avginc=30, where the increase of x has a great effect on the test score, so the slope will be higher in absolute terms. Over avginc=30 we are going to have a very high score, but we are not going to have a great effect on the dependent variable. However, the R^2 is 0.51 and test values of both the coefficients and the global model are significantly high to reject our null hypothesis.

- testscr~elpct:

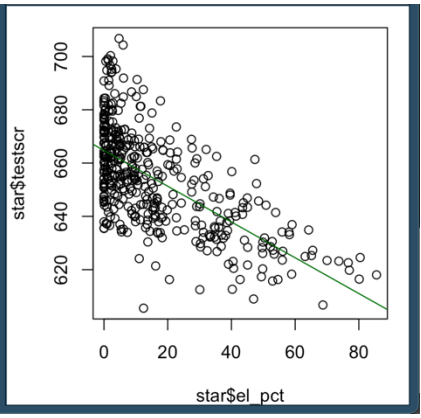
Testscr=664.74-0.671(el_pct)

```
Call:
lm(formula = testscr ~ el_pct, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-50.861 -10.183  -0.807   9.004  45.183

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 664.73944    0.94064  706.69   <2e-16 ***
el_pct       -0.67116    0.03898  -17.22   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.59 on 418 degrees of freedom
Multiple R-squared:  0.4149,    Adjusted R-squared:  0.4135
F-statistic: 296.4 on 1 and 418 DF,  p-value: < 2.2e-16
```



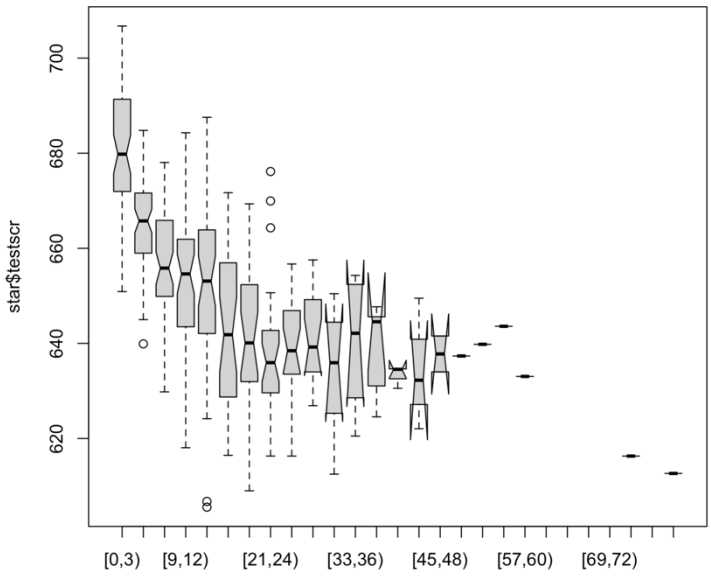
The English Learners ratio gives us some information about the testscr. In particular, we can see a downward trend as long as el_pct increases. Districts with over 40% el_pct performed below the mean and the point cloud shape could be an indicator of the dependence of the testscr from this variable. Then the error seems to decrease when the value of el_pct increases. It means that our model would report small errors for high values of the regressor(the y-range becomes smaller and smaller), while for low regressor values the range is higher (es. If el_pct=0 the range goes from 640 to 700), so the model would be less precise.

However, the coefficients test and the global test both give us low p-values, so we can conclude they are quite significant, while the R^2 remains pretty high(0,414).

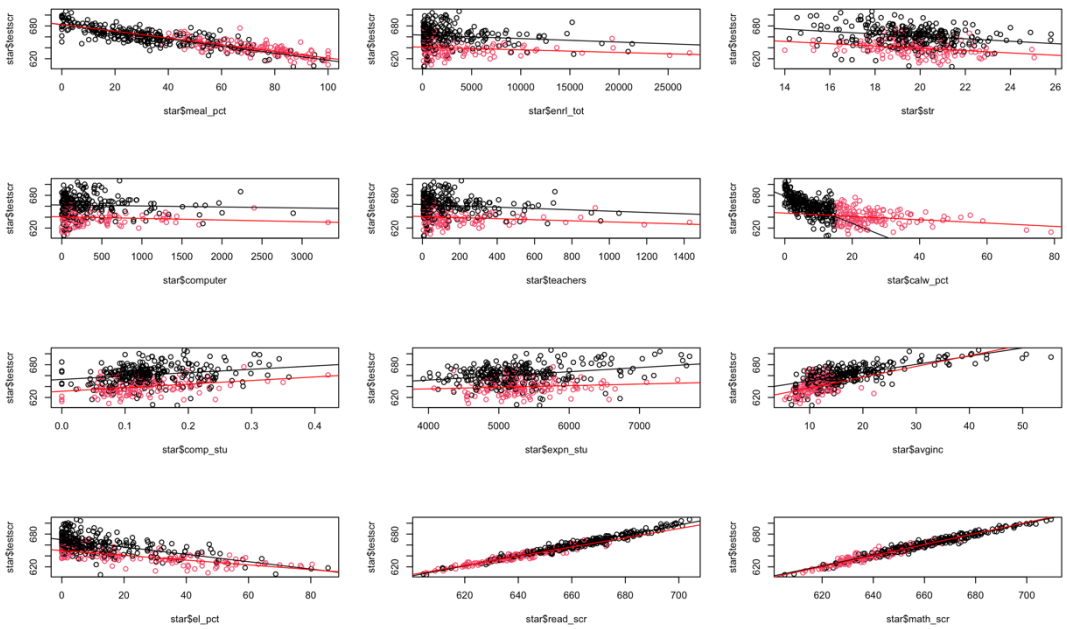
Dummy Variable

In order to evaluate the effect of CalWork on the test, I decided to introduce a new dummy variable, but I wondered about the possible values that could represent my two groups. So I transformed it from a continuous variable to a discrete one, I built a frequency table and I also print a boxplot with testscores on the y-axis to study its distribution.

```
> cbind(cw.freq,cw.cumfreq,cw.cumrelfreq)
      cw.freq cw.cumfreq cw.cumrelfreq
[0,3)      57         57        0.136
[3,6)      64        121        0.288
[6,9)      56        177        0.421
[9,12)     49        226        0.538
[12,15)    43        269        0.640
[15,18)    32        301        0.717
[18,21)    25        326        0.776
[21,24)    27        353        0.840
[24,27)    20        373        0.888
[27,30)    16        389        0.926
[30,33)     6        395        0.940
[33,36)     6        401        0.955
[36,39)     5        406        0.967
[39,42)     3        409        0.974
[42,45)     3        412        0.981
[45,48)     2        414        0.986
[48,51)     1        415        0.988
[51,54)     1        416        0.990
[54,57)     1        417        0.993
[57,60)     1        418        0.995
[60,63)     0        418        0.995
[63,66)     0        418        0.995
[66,69)     0        418        0.995
[69,72)     0        418        0.995
[72,75)     1        419        0.998
[75,78)     0        419        0.998
[78,81)     1        420        1.000
```



As we have seen in the description chapter, the median is 10,52. I wanted to have two groups with similar relative frequency, but also two groups that were significant for the performance in the test. I saw a 1st group, from 0<calw<3, that had a very high testscr, 2nd group from 3 to 15 with a medium score, and 3rd group with calw_pct over 15 that had lower scores than others. However, I decided to create two groups. *Calw_cat*=0 if *calw_pct*<15 and *Calw_cat*=1 otherwise.
We have also already seen that *calw_pct* explains pretty well our test score, but it could be more clear if we see some graphs. *(they are pretty more clear if you run the code in Rstudio)



- We have all the scatterplots with the different variables on the x-axis and the testscr on the y-axis. The red dots represent calw_cat=1, while the black ones represent calw_cat=0.
- In this case, the contribution of CalWork is pretty clear. It has an increasing effect in the intercept value, and it helps reducing the errors in our models.
- We have a significant effect on the coefficient of the two lines in *calw_pct* scatterplot. Since we have created two groups based on their performance, and we also have a difference between in the interval CalW=(0-3) and CalW=(3-15), the lines' slope is differs between the red-dotted and the black-dotted group. A decreasing in the CalW_pct will have a larger effect on testscr in the black-dotted group, than it will have on testscr in the red-dotted group.
- The dummy variable had not a great impact on *meal_pct,read_scr* and *math_scr*. Since they alreay have an high R^2, probably the information of the dummy variable is redundant and it didn't have a great effect.
- I performed the ANOVA for every simple regression and I added some comments in the code

Multiple Regression Analysis

Here I am going to illustrate *backward elimination* method, but I also tried a second time using stepwise (forward) regression, that consists in minimizing the AIC value. It helps in the evaluation of the model. Our objective is to consider both the model complexity and the efficacy in predicting the dependent variable.

The backward elimination considers an initial model with all the variables as predictors, and dropping at every step the less significant coefficient, considering a significance level of 0,05.
So, this is our starting model, and we are going to eliminate the coefficient with the higher p-value. In this case is teacher with a t-value 0,226.

```
Call:
lm(formula = testscr ~ ., data = star2[, -12])

Residuals:
     Min       1Q   Median       3Q      Max
-31.1387  -5.2680   0.1635   4.9890  26.6532

Coefficients:
              Estimate  Std. Error t value        Pr(>|t|)
(Intercept) 658.7778741   9.7479075  67.581         < 2e-16 ***
enrl_tot     -0.0005647   0.0016483  -0.343          0.7321
teachers      0.0082030   0.0363627   0.226          0.8216
calw_pct     -0.0829012   0.0583439  -1.421          0.1561
meal_pct     -0.3739002   0.0363427 -10.288         < 2e-16 ***
computer      0.0015700   0.0031115   0.505          0.6141
comp_stu     10.1769110   7.7665430   1.310          0.1908
expn_stu      0.0015969   0.0009035   1.767          0.0779 .
str          -0.1524738   0.3261523  -0.467          0.6404
avginc        0.6146885   0.0897236   6.851 0.0000000000271 ***
el_pct       -0.1994941   0.0350835  -5.686 0.0000000247429 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.417 on 409 degrees of freedom
Multiple R-squared:  0.8095,    Adjusted R-squared:  0.8048
F-statistic: 173.8 on 10 and 409 DF,  p-value: < 2.2e-16
```

After several steps we can consider our fitted model as:
-testscr=mealpct+comp_stu+expn_stu+ avginc+ el_pct
The global test and coefficient test are both quite significant. The R^adj is 0.806

```
Call:
lm(formula = testscr ~ meal_pct + comp_stu + expn_stu + avginc +
    el_pct, data = star2[, -12])

Residuals:
    Min      1Q  Median      3Q     Max
-31.574  -5.422   0.034   5.141  26.526

Coefficients:
              Estimate  Std. Error t value        Pr(>|t|)
(Intercept) 654.9726130   3.6197396 180.945         < 2e-16 ***
meal_pct     -0.4064676   0.0279520 -14.542         < 2e-16 ***
comp_stu     13.5065526   6.8001000   1.986          0.0477 *
expn_stu      0.0016874   0.0007307   2.309          0.0214 *
avginc        0.6194042   0.0877352   7.060 0.00000000000707 ***
el_pct       -0.1859544   0.0313278  -5.936 0.00000000619164 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.395 on 414 degrees of freedom
Multiple R-squared:  0.8082,    Adjusted R-squared:  0.8059
F-statistic: 348.9 on 5 and 414 DF,  p-value: < 2.2e-16
```

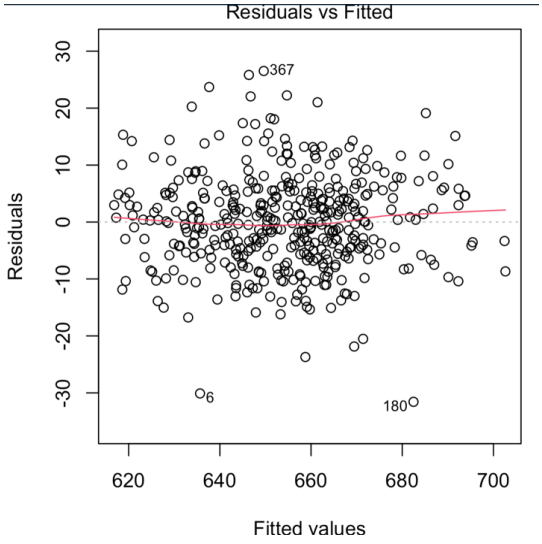
Then I wondered about the effect of adding our dummy variable *calw_cat* to our fitted model.
1) Same slope, different intercept: it did not bring a significant result, since the p-value on the coefficient test is considerably high (0.714).
2) Same intercept, different slopes. In this case, the model is much more complex, and it is better to evaluate the performance with ANOVA.

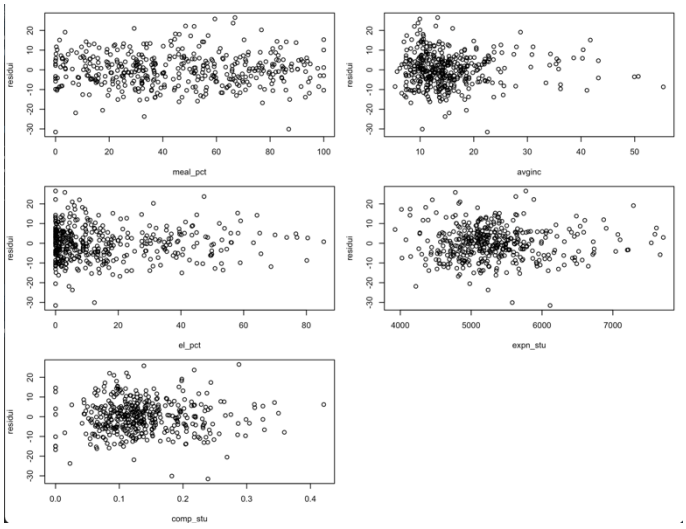
```
Analysis of Variance Table

Model 1: testscr ~ meal_pct + avginc + el_pct + expn_stu + comp_stu
Model 2: testscr ~ meal_pct:calw_cat + avginc:calw_cat + el_pct:calw_cat +
    expn_stu:calw_cat + comp_stu:calw_cat
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    414 29174
2    409 28944  5    229.72 0.6492 0.6622
```

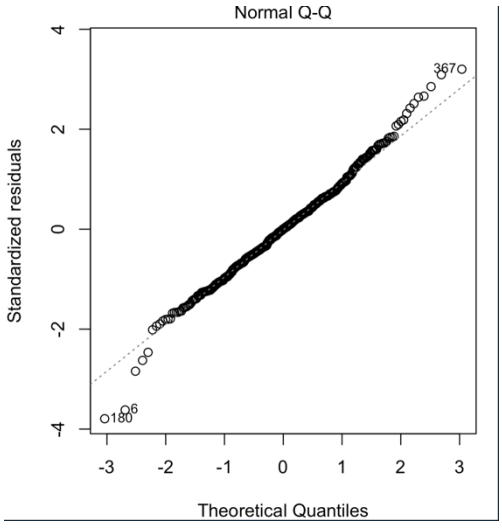
In this case, as we have already said, the "saving" in terms of decreasing of Residual Sum of Squares doesn't justify the complexity of the model. Our p-value for the model with 409 df is pretty high.

Plot(mod5)





Residuals vs.fitted = the red line is near to the dashed 0-line and the points seem quite casually located, so we can confirm linearity.

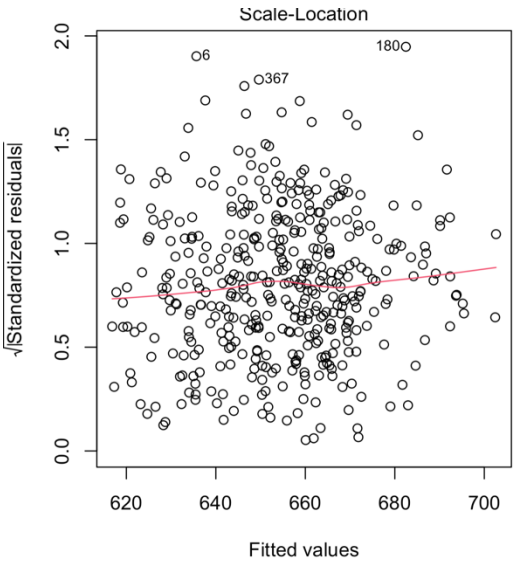


Normal qqplot
The graph shows empirical standardized residuals vs. Gaussian Quantiles, in order to confirm our errors' Gaussian distribution assumption. In this case, we can see that points approximately fall on the line, even if it shows three outliers' values(180,6,367). The normality assumption is respected from (-2,2) interval. The values in tails don't seem to be normally distributed.

Scale-Location

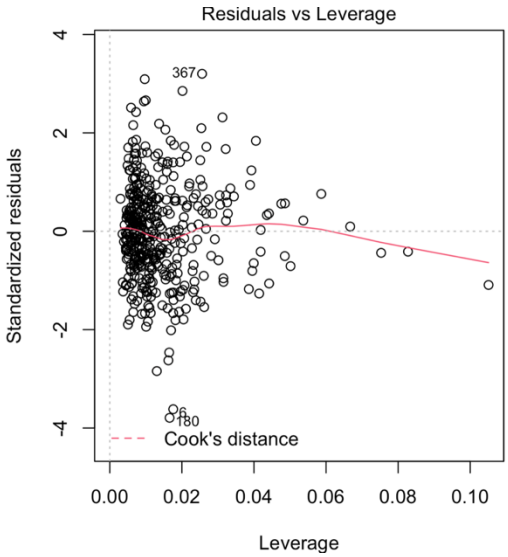
scale-location plot helps us to check homoskedasticity assumption. We have the Fitted Value on the x-axis and The square root of absolute standardized residuals on the y-axis.

Homoskedasticity seems to be respected since the red line (representing the mean of the standardized residuals for the fitted values) is quite horizontal and the vertically spread of the dots seems quite homogeneous. It doesn't seem to be respected for higher values since they seem to fall near the red line and are less scattered. This could also involve that we have a smaller frequency for values over 680. However, if we perform the studentized Breusch-Pagan test, where the null hypothesis is homoskedasticity, we fail to reject it.



Residuals vs. Leverage

This graph helps us detect some outliers that could largely influence our model. First, we can notice that the spread of points diminishes with the increase of the leverage(on the x-axis). However, we don't need to consider removing any point, since no point surpasses Cook's distance in the graph.



	obs	enrl_tot	teachers	calw_pct	meal_pct	computer	testscr	comp_stu	expn_stu	str	avginc	el_pct	read_scr	math_scr
6	6	137	6.400	12.3188	86.9565	25	605.55	0.1824818	5580.147	21.40625	10.41500	12.40876	605.7	605.4

	obs	enrl_tot	teachers	calw_pct	meal_pct	computer	testscr	comp_stu	expn_stu	str	avginc	el_pct	read_scr	math_scr
180	180	92	5.600	13.5417	0.0000	22	650.90	0.2391304	6113.860	16.42857	22.52900	0.00000	656.1	645.7

	obs	enrl_tot	teachers	calw_pct	meal_pct	computer	testscr	comp_stu	expn_stu	str	avginc	el_pct	read_scr	math_scr
367	367	139	7.350	20.8633	66.9065	40	676.15	0.2877698	5771.384	18.91157	13.27300	0.00000	680.1	672.2

These three districts were often highlighted in our regression plot, probably because they behave differently from the other ones. Let's try to understand why.

6. This district is pretty small, since it counts a few number of enrollment and teachers iit has been classified in our *"calw<15"* group, and it has a great percentage of reduced-price lunch beneficiaries. The low el_pct and calw_pct should suggest a medium-high testscr. However, despite these characteristics, our regression line fails to predict our value because the average performance on the test is 605.5, the lowest value in our dataset.

180. Also this district is pretty small and has a low calw_pct(<15). Some values are pretty significant. For example, the meal_pct is 0, average income it's over the 3rd quartile and el_pct is 0, expenditure is also over the 3rd quartile, str also is pretty low. I think that in this case, the regression line fails to predict our *testscr* because it had a 650.9 score, below mean and median values.

367. This small district has a high calw_pct and meal_pct. However, the el_pct is pretty low(if we look at the simple regression scatterplot we remember that the point clouds are more scattered near low values). It is pretty relevant because it has been classified in calw_cat=1, but it had a brilliant test score(676.15), so its predicted value is much different from the real one.