On the last screen, you should have seen output that looked like this:



SAT Critical Reading Avg. Score         0.986820

SAT Math Avg. Score                     0.972643

SAT Writing Avg. Score                  0.987771

sat\_score                               1.000000

AP Test Takers                          0.523140

Total Exams Taken                       0.514333

Number of Exams with scores 3 4 or 5    0.463245

Total Cohort                            0.325144

CSD                                     0.042948

NUMBER OF STUDENTS / SEATS FILLED       0.394626

NUMBER OF SECTIONS                      0.362673

AVERAGE CLASS SIZE                      0.381014

SIZE OF SMALLEST CLASS                  0.249949

SIZE OF LARGEST CLASS                   0.314434

SCHOOLWIDE PUPIL-TEACHER RATIO               NaN

schoolyear                                   NaN

fl\_percent                                   NaN

frl\_percent                            -0.722225

total\_enrollment                        0.367857

ell\_num                                -0.153778

ell\_percent                            -0.398750

sped\_num                                0.034933

sped\_percent                           -0.448170

asian\_num                               0.475445

asian\_per                               0.570730

black\_num                               0.027979

black\_per                              -0.284139

hispanic\_num                            0.025744

hispanic\_per                           -0.396985

white\_num                               0.449559

                                        ...

rr\_p                                    0.047925

N\_s                                     0.423463

N\_t                                     0.291463

N\_p                                     0.421530

saf\_p\_11                                0.122913

com\_p\_11                               -0.115073

eng\_p\_11                                0.020254

aca\_p\_11                                0.035155

saf\_t\_11                                0.313810

com\_t\_11                                0.082419

eng\_t\_10                                     NaN

aca\_t\_11                                0.132348

saf\_s\_11                                0.337639

com\_s\_11                                0.187370

eng\_s\_11                                0.213822

aca\_s\_11                                0.339435

saf\_tot\_11                              0.318753

com\_tot\_11                              0.077310

eng\_tot\_11                              0.100102

aca\_tot\_11                              0.190966

grade\_span\_max                               NaN

expgrade\_span\_max                            NaN

zip                                    -0.063977

total\_students                          0.407827

number\_programs                         0.117012

priority08                                   NaN

priority09                                   NaN

priority10                                   NaN

lat                                    -0.121029

lon                                    -0.132222

Name: sat\_score, dtype: float64

Unsurprisingly, SAT Critical Reading Avg. Score, SAT Math Avg. Score, SAT Writing Avg. Score, and sat\_scoreare strongly correlated with sat\_score.

We can also make some other observations:

* total\_enrollment has a strong positive correlation with sat\_score. This is surprising because we'd expect smaller schools where students receive more attention to have higher scores. However, it looks like the opposite is true -- larger schools tend to do better on the SAT.
  + Other columns that are proxies for enrollment correlate similarly. These include total\_students, N\_s, N\_p, N\_t, AP Test Takers, Total Exams Taken, and NUMBER OF SECTIONS.
* Both the percentage of females (female\_per) and number of females (female\_num) at a school correlate positively with SAT score, whereas the percentage of males (male\_per) and the number of males (male\_num) correlate negatively. This could indicate that women do better on the SAT than men.
* Teacher and student ratings of school safety (saf\_t\_11, and saf\_s\_11) correlate with sat\_score.
* Student ratings of school academic standards (aca\_s\_11) correlate with sat\_score, but this does not hold for ratings from teachers and parents (aca\_p\_11 and aca\_t\_11).
* There is significant racial inequality in SAT scores (white\_per, asian\_per, black\_per, hispanic\_per).
* The percentage of English language learners at the school (ell\_percent, frl\_percent) has a strong negative correlation with SAT scores.

Because enrollment seems to have such a strong correlation, let's make a scatterplot of total\_enrollment vs sat\_score. Each point in the scatterplot will represent a high school, so we'll be able to see if there are any interesting patterns.

We can plot columns in a dataframe using the [pandas.DataFrame.plot()](http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.plot.html" \t "_blank) accessor on a dataframe. We can also specify a certain plot type. For example, df.plot.scatter(x="A", y="b") will create a scatterplot of columns A and B.

* Create a scatterplot of total\_enrollment versus sat\_score.

**Answer**

import matplotlib.pyplot as plt

# Because enrollment seems to have such a strong correlation, let's make a scatterplot of total\_enrollment vs sat\_score.

# Each point in the scatterplot will represent a high school, so we'll be able to see if there are any interesting

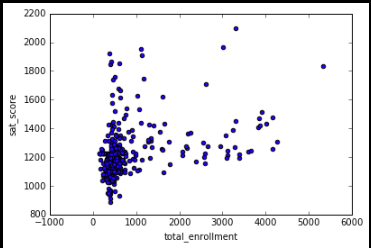
# patterns.

# We can plot columns in a dataframe using the pandas.DataFrame.plot() accessor on a dataframe. We can also specify a

# certain plot type. For example, df.plot.scatter(x="A", y="b") will create a scatterplot of columns A and B.

plot = combined.plot.scatter(x = 'total\_enrollment', y = 'sat\_score')

plt.show()



Judging from the plot we just created, it doesn't appear that there's an extremely strong correlation between sat\_score and total\_enrollment. If there was a very strong correlation, we'd expect all of the points to line up. Instead, there's a large cluster of schools, and then a few others going off in three different directions.

However, there's an interesting cluster of points at the bottom left where total\_enrollment and sat\_score are both low. This cluster may be what's making the r value so high. It's worth extracting the names of the schools in this cluster so we can research them further