Schools

October 16, 2017

1 Read in the data

In [1]:
```
import pandas
import numpy
import re

data_files = [
    "ap_2010.csv",
    "class_size.csv",
    "demographics.csv",
    "graduation.csv",
    "hs_directory.csv",
    "sat_results.csv"
]

data = {}

for f in data_files:
    d = pandas.read_csv("schools/{0}".format(f))
    data[f.replace(".csv", "")] = d
```

2 Read in the surveys

In [2]:
```
all_survey = pandas.read_csv("schools/survey_all.txt",
                             delimiter="\t", encoding='windows-1252')
d75_survey = pandas.read_csv("schools/survey_d75.txt",
                             delimiter="\t", encoding='windows-1252')
survey = pandas.concat([all_survey, d75_survey], axis=0)

survey["DBN"] = survey["dbn"]

survey_fields = [
    "DBN",
    "rr_s",
    "rr_t",
    "rr_p",
    "N_s",
```

```
        "N_t",
        "N_p",
        "saf_p_11",
        "com_p_11",
        "eng_p_11",
        "aca_p_11",
        "saf_t_11",
        "com_t_11",
        "eng_t_10",
        "aca_t_11",
        "saf_s_11",
        "com_s_11",
        "eng_s_11",
        "aca_s_11",
        "saf_tot_11",
        "com_tot_11",
        "eng_tot_11",
        "aca_tot_11",
    ]
    survey = survey.loc[:,survey_fields]
    data["survey"] = survey
```

3 Add DBN columns

```
In [3]: data["hs_directory"]["DBN"] = data["hs_directory"]["dbn"]

    def pad_csd(num):
        string_representation = str(num)
        if len(string_representation) > 1:
            return string_representation
        else:
            return "0" + string_representation

    data["class_size"]["padded_csd"] = data["class_size"]["CSD"].apply(pad_csd)
    data["class_size"]["DBN"] = data["class_size"]["padded_csd"] + \
        data["class_size"]["SCHOOL CODE"]
```

4 Convert columns to numeric

```
In [4]: cols = ['SAT Math Avg. Score', 'SAT Critical Reading Avg. Score',
                'SAT Writing Avg. Score']
    for c in cols:
        data["sat_results"][c] = pandas.to_numeric(data["sat_results"][c],
                                                   errors="coerce")

    data['sat_results']['sat_score'] = data['sat_results'][cols[0]] + \
        data['sat_results'][cols[1]] + \
```

```
        data['sat_results'][cols[2]]

    def find_lat(loc):
        coords = re.findall("\(.+, .+\)", loc)
        lat = coords[0].split(",")[0].replace("(", "")
        return lat

    def find_lon(loc):
        coords = re.findall("\(.+, .+\)", loc)
        lon = coords[0].split(",")[1].replace(")", "").strip()
        return lon

    data["hs_directory"]["lat"] = data["hs_directory"]["Location 1"] \
            .apply(find_lat)
    data["hs_directory"]["lon"] = data["hs_directory"]["Location 1"] \
            .apply(find_lon)

    data["hs_directory"]["lat"] = pandas.to_numeric(
        data["hs_directory"]["lat"], errors="coerce")
    data["hs_directory"]["lon"] = pandas.to_numeric(
        data["hs_directory"]["lon"], errors="coerce")
```

5 Condense datasets

```
In [5]: class_size = data["class_size"]
        class_size = class_size[class_size["GRADE "] == "09-12"]
        class_size = class_size[class_size["PROGRAM TYPE"] == "GEN ED"]

        class_size = class_size.groupby("DBN").agg(numpy.mean)
        class_size.reset_index(inplace=True)
        data["class_size"] = class_size

        data["demographics"] = data["demographics"] \
                               [data["demographics"]["schoolyear"] == 20112012]

        data["graduation"] = data["graduation"][data["graduation"]["Cohort"]
                                                == "2006"]
        data["graduation"] = data["graduation"][data["graduation"]["Demographic"]
                                                == "Total Cohort"]
```

6 Convert AP scores to numeric

```
In [6]: cols = ['AP Test Takers ', 'Total Exams Taken',
                'Number of Exams with scores 3 4 or 5']

        for col in cols:
```

```
        data["ap_2010"][col] = pandas.to_numeric(data["ap_2010"][col],
                                                 errors="coerce")
```

7 Combine the datasets

```
In [7]: combined = data["sat_results"]

        combined = combined.merge(data["ap_2010"], on="DBN", how="left")
        combined = combined.merge(data["graduation"], on="DBN", how="left")

        to_merge = ["class_size", "demographics", "survey", "hs_directory"]

        for m in to_merge:
            combined = combined.merge(data[m], on="DBN", how="inner")

        combined = combined.fillna(combined.mean())
        combined = combined.fillna(0)
```

8 Add a school district column for mapping

```
In [8]: def get_first_two_chars(dbn):
            return dbn[0:2]

        combined["school_dist"] = combined["DBN"].apply(get_first_two_chars)
```

9 Find correlations

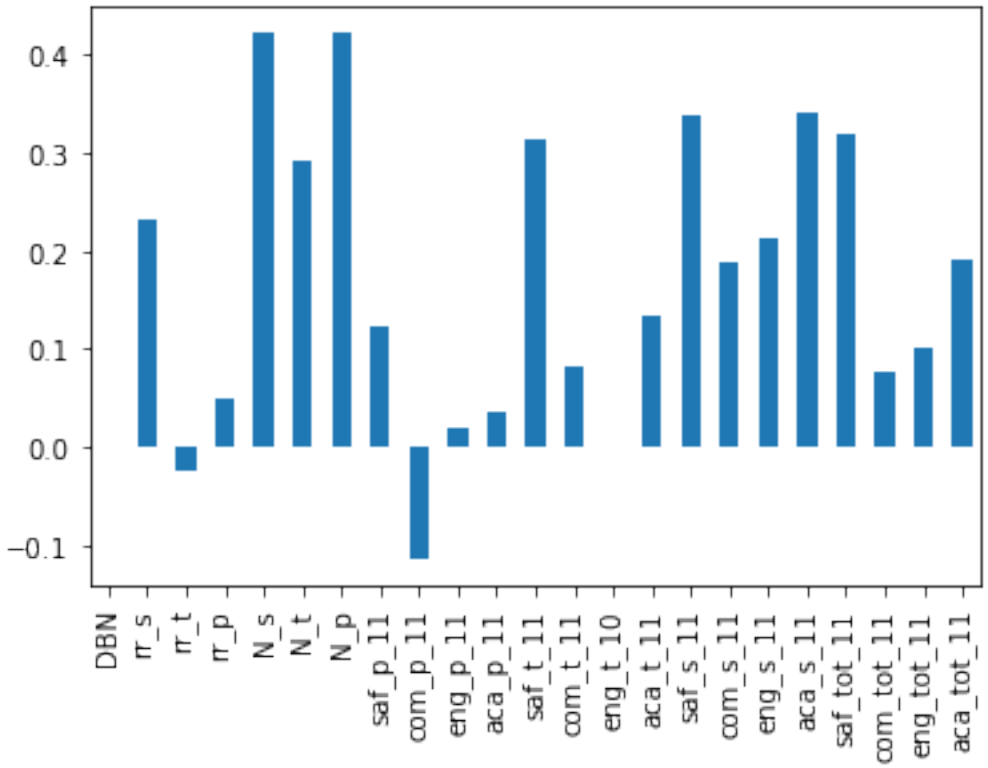
```
In [9]: correlations = combined.corr()
        correlations = correlations["sat_score"]
        print(correlations)
```

```
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
```

10 Plotting Survey Correlations

```
In [10]: %matplotlib inline
         combined.corr()["sat_score"][survey_fields].plot.bar();
```



There are high correlations between N_s, N_t, N_p and sat_score. Since these columns are correlated with total_enrollment, it makes sense that they would be high.

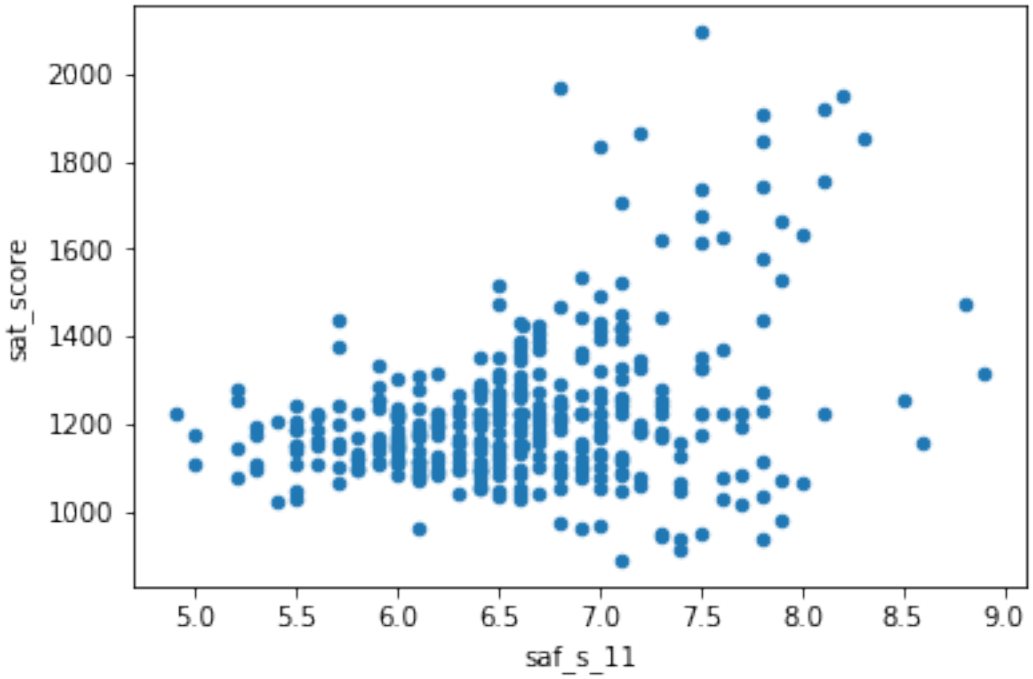
It is more interesting that rr_s, the student response rate, or the percentage of students that completed the survey, correlates with sat_score. This might make sense because students who are more likely to fill out surveys may be more likely to also be doing well academically.

How students and teachers percieved safety (saf_t_11 and saf_s_11) correlate with sat_score. This make sense, as it's hard to teach or learn in an unsafe environment.

The last interesting correlation is the aca_s_11, which indicates how the student perceives academic standards, correlates with sat_score, but this is not true for aca_t_11, how teachers perceive academic standards, or aca_p_11, how parents perceive academic standards.

11 Exploring Safety

```
In [11]: combined.plot.scatter('saf_s_11', 'sat_score');
```



There appears to be a correlation between SAT scores and safety, although it isn't that strong. It looks like there are a few schools with extremely high SAT scores and high safety scores. There are a few schools with low safety scores and low SAT scores. No school with a safety score lower than 6.5 has an average SAT score higher than around 1500.

12 Plotting Safety

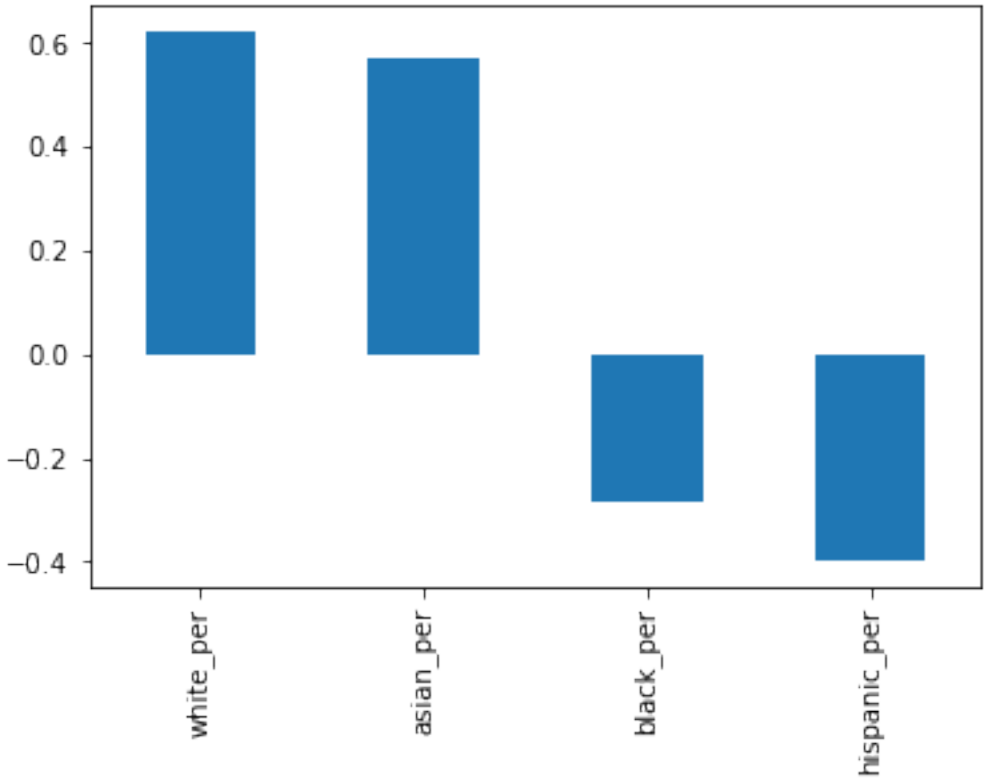
In [12]:
```
import matplotlib.pyplot as plt
from mpl_toolkits.basemap import Basemap
import numpy as np

districts = combined.groupby("school_dist").agg(np.mean)
districts.reset_index (inplace=True)

m = Basemap(
    projection='merc',
    llcrnrlat=40.496044,
    urcrnrlat=40.915256,
    llcrnrlon=-74.255735,
    urcrnrlon=-73.700272,
    resolution='i'
)

m.drawmapboundary(fill_color = "#85A6D9")
m.drawcoastlines(color = "#6D5F47", linewidth=.4)
```

```
m.drawrivers(color = "#6D5F47", linewidth=.4)
m.fillcontinents(color='white',lake_color='#85A6D9')

longitudes = districts["lon"].tolist()
latitudes = districts["lat"].tolist()

m.scatter(longitudes, latitudes, s=50, zorder=2, latlon=True,
          c=districts["saf_s_11"], cmap="summer")
plt.show()
```

It looks like Upper Manhattan and parts of Queens and the Bronx tend to have lower safety scores, whereas Brooklyn has high safety scores.

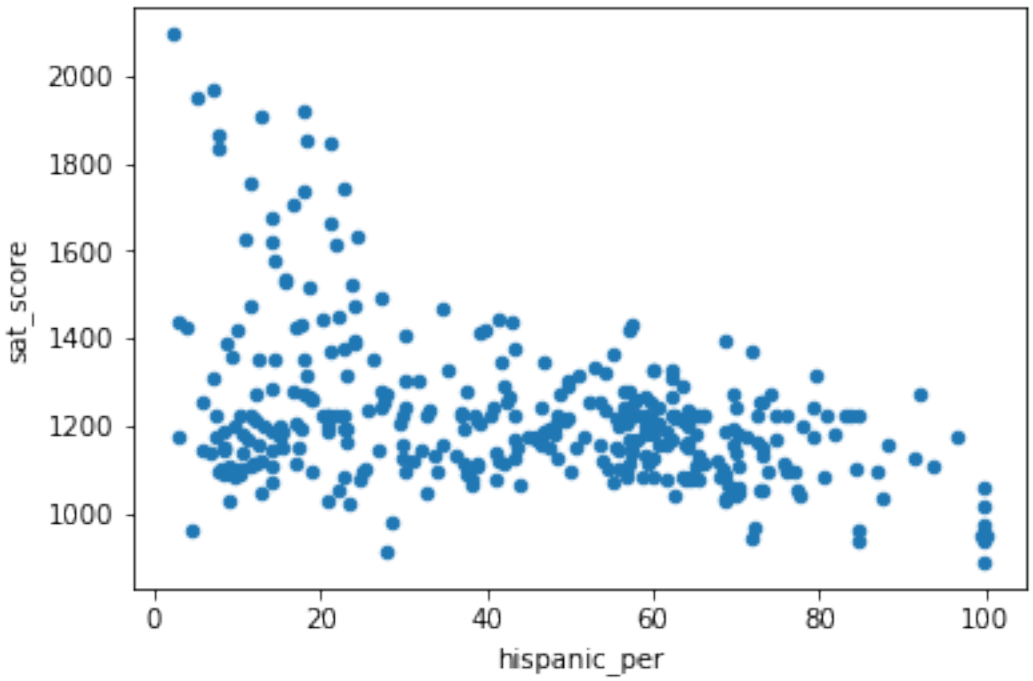
13 Racial differences in SAT Score

```
In [13]: RF = ['white_per', 'asian_per', 'black_per', 'hispanic_per']
         combined.corr()["sat_score"][RF].plot.bar();
```



It looks like a higher percentage of white or asian students at a school correlates positively with sat score, whereas a higher percentage of black or hispanic students correlates negatively with sat score. This may be due to a lack of funding for schools in certain areas, which are more likely to have a higher percentage of black or hispanic students.

```
In [14]: combined.plot.scatter('hispanic_per', 'sat_score');
```




```
In [15]: print(combined[combined["hispanic_per"] > 95]["SCHOOL NAME"])

44                          MANHATTAN BRIDGES HIGH SCHOOL
82       WASHINGTON HEIGHTS EXPEDITIONARY LEARNING SCHOOL
89     GREGORIO LUPERON HIGH SCHOOL FOR SCIENCE AND M...
125                  ACADEMY FOR LANGUAGE AND TECHNOLOGY
141                INTERNATIONAL SCHOOL FOR LIBERAL ARTS
176     PAN AMERICAN INTERNATIONAL HIGH SCHOOL AT MONROE
253                            MULTICULTURAL HIGH SCHOOL
286               PAN AMERICAN INTERNATIONAL HIGH SCHOOL
Name: SCHOOL NAME, dtype: object
```

The schools listed above appear to primarily be geared towards recent immigrants to the US. These schools have a lot of students who are learning English, which would explain the lower SAT scores.

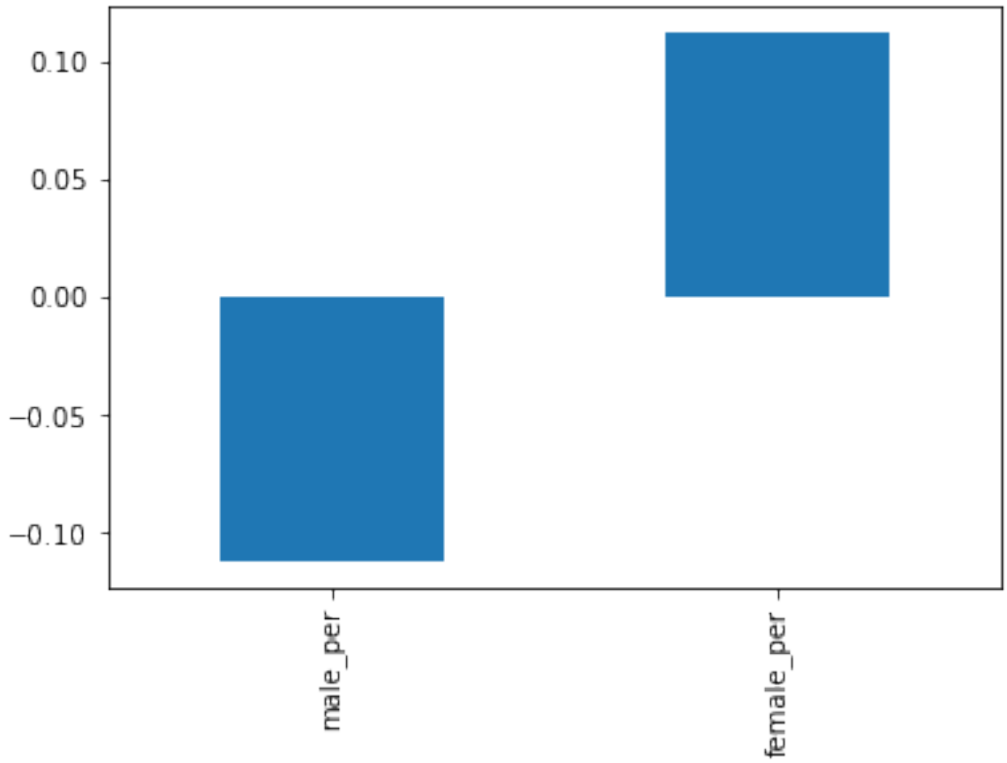
```
In [16]: print(combined[(combined["hispanic_per"] < 10) &
                        (combined["sat_score"] > 1800)]["SCHOOL NAME"])

37                                 STUYVESANT HIGH SCHOOL
151                         BRONX HIGH SCHOOL OF SCIENCE
187                       BROOKLYN TECHNICAL HIGH SCHOOL
327    QUEENS HIGH SCHOOL FOR THE SCIENCES AT YORK CO...
356                   STATEN ISLAND TECHNICAL HIGH SCHOOL
```

```
Name: SCHOOL NAME, dtype: object
```

Many of the schools above appear to be specialized science and technology schools that receive extra funding, and only admit students who pass an entrance exam. This doesn't explain the low hispanic_per, but it does explain why their students tend to do better on the SAT – they are students from all over New York City who did well on a standardized test.

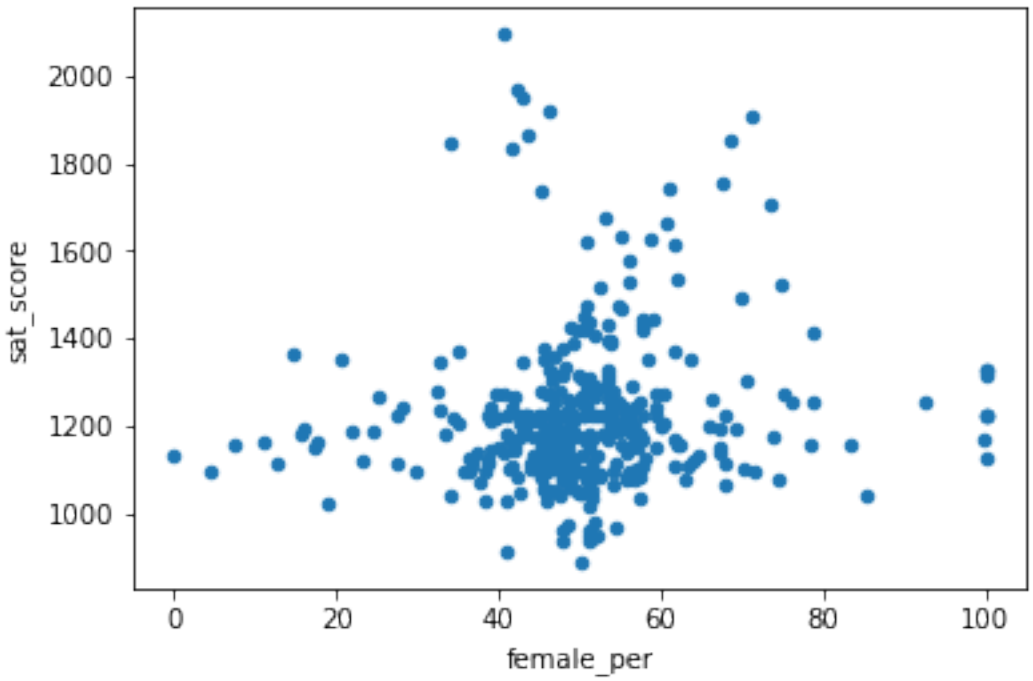
14 Gender Differences in SAT Score

```
In [17]: TG = ["male_per", "female_per"]
         combined.corr()["sat_score"][TG].plot.bar();
```



In the plot above, we can see that a high percentage of females at a school positively correlates with SAT score, whereas a high percentage of males at a school negatively correlates with SAT score. Neither correlation is extremely strong.

```
In [18]: combined.plot.scatter("female_per", "sat_score");
```



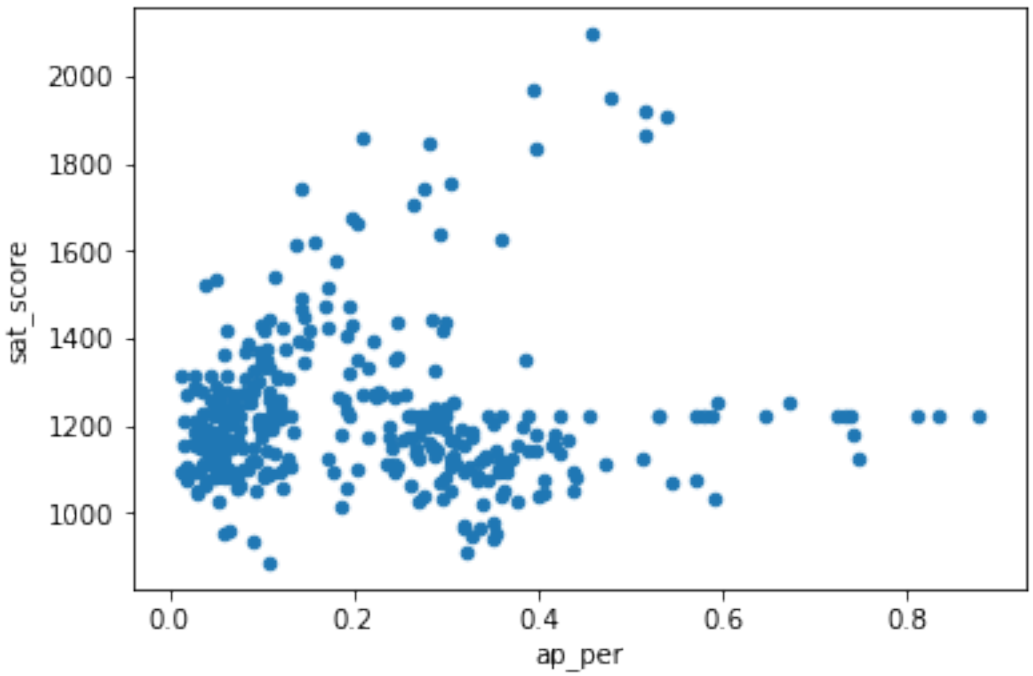
Based on the scatterplot, there doesn't seem to be any real correlation between sat_score and female_per. However, there is a cluster of schools with a high percentage of females (60 to 80), and high SAT scores.

```
In [19]: print(combined[(combined["female_per"] > 60)
                         & (combined["sat_score"] > 1700)]["SCHOOL NAME"])

5                          BARD HIGH SCHOOL EARLY COLLEGE
26                          ELEANOR ROOSEVELT HIGH SCHOOL
60                                   BEACON HIGH SCHOOL
61     FIORELLO H. LAGUARDIA HIGH SCHOOL OF MUSIC & A...
302                          TOWNSEND HARRIS HIGH SCHOOL
Name: SCHOOL NAME, dtype: object
```

These schools appears to be very selective liberal arts schools that have high academic standards.

15 AP Exams vs SAT Scores

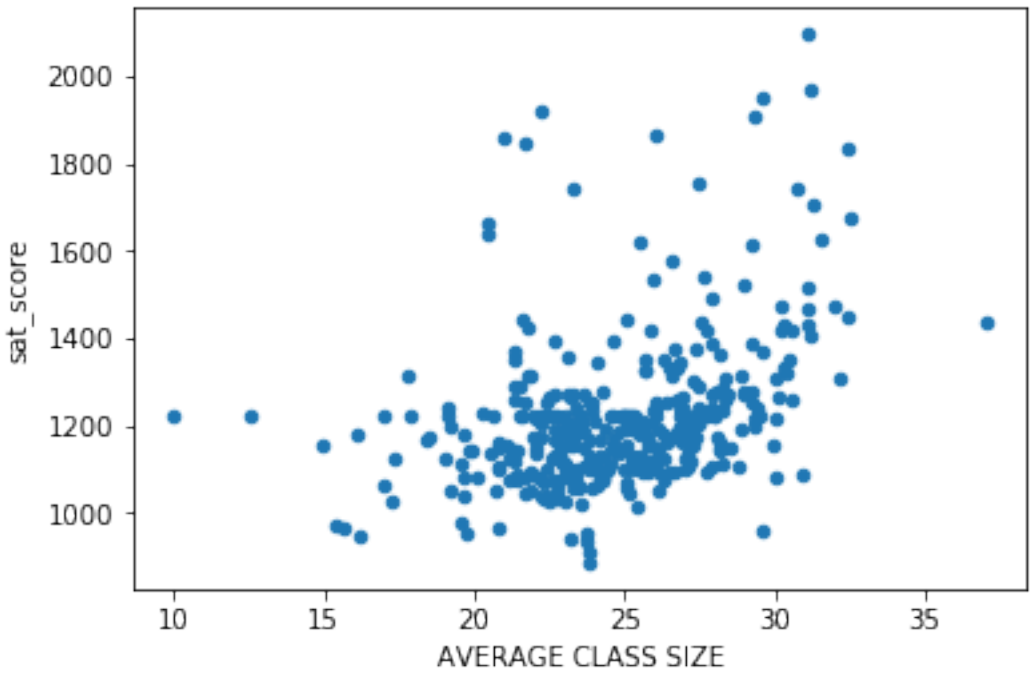
```
In [32]: combined['ap_per'] = combined["AP Test Takers "] / \
                               combined["total_enrollment"]
         combined.plot.scatter(x='ap_per', y='sat_score');
```



It looks like there is a relationship between the percentage of students in a school who take the AP exam, and their average SAT scores. It's not an extremely strong correlation, though.

16 Average Class Size vs SAT Scores

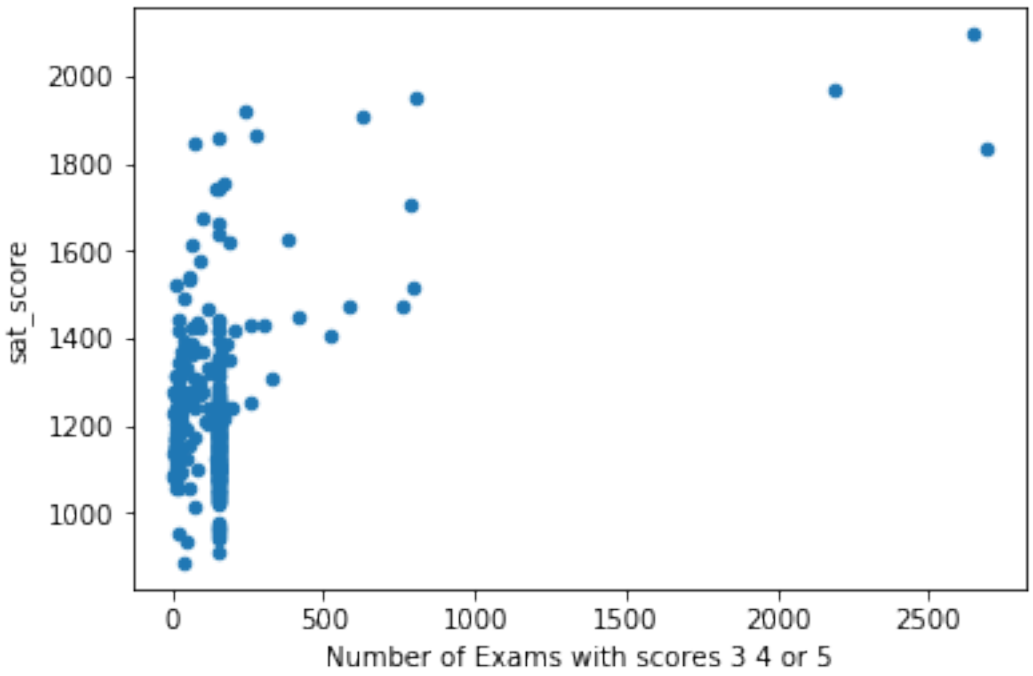
```
In [21]: combined.plot.scatter('AVERAGE CLASS SIZE', 'sat_score');
```



One can conclude from the graph that the larger the class size is, the higher the SAT score is. This is not a very strong correlation nor is this intuitive. One might guess that a smaller class size means students recieves more attention from a teacher. But one can also say that the larger class size hass a teacher who has very effective teaching methods.

17 Passed AP Exams vs SAT Scores

```
In [22]: combined.plot.scatter('Number of Exams with scores 3 4 or 5', 'sat_score')
```

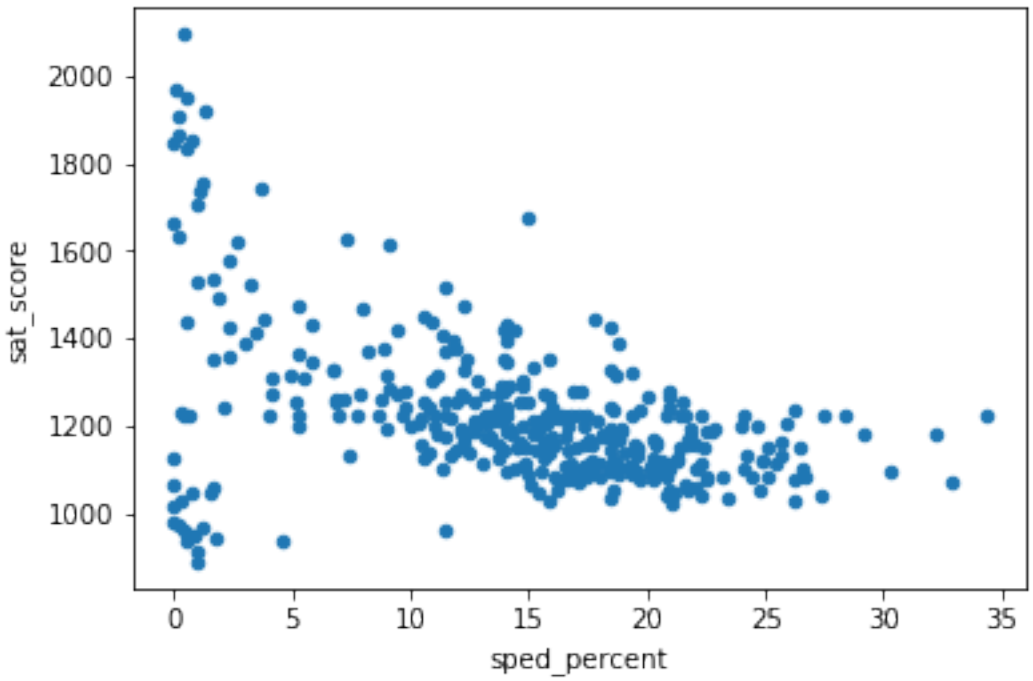



```
In [23]: print(combined[(combined['Number of Exams with scores 3 4 or 5'] > 2000)
                         & (combined['sat_score'] > 1800)]["SCHOOL NAME"])
```

```
37              STUYVESANT HIGH SCHOOL
151       BRONX HIGH SCHOOL OF SCIENCE
187     BROOKLYN TECHNICAL HIGH SCHOOL
Name: SCHOOL NAME, dtype: object
```

The above three schools have appeared earlier in this analysis and these schools have high academic standards. With that in mind, it would make sense that they have a high number of AP exams being passed (at least a 3 is required to pass) and a high SAT score. In addition,

```
In [24]: combined.plot.scatter('sped_percent', 'sat_score');
```



It would make sense that scores with a higher special education percentage would have a lower SAT score. The SAT is a exam, which measures a student's college readiness. If the student has difficulty reading, writing, or doing arithmetic (many of Sp. Ed students do,) it could explain low SAT scores.

```
In [25]: print(combined[(combined['sped_percent'] > 30) &
                         (combined['sat_score'] < 1300)]["SCHOOL NAME"])

6      47 THE AMERICAN SIGN LANGUAGE AND ENGLISH SECO...
39                    UNITY CENTER FOR URBAN TECHNOLOGIES
83             HIGH SCHOOL FOR EXCELLENCE AND INNOVATION
207                                AUTOMOTIVE HIGH SCHOOL
Name: SCHOOL NAME, dtype: object
```

The above schools either are geared toward special education students or mainstreams special education students, which puts undue stress on the general education teacher. In the case of Automotive High School, these students genuinely have no interest going to college when they can be a mechanic instead. Also, academics is often an afterthought at Automotive HS.

```
In [26]: print(combined[(combined['sped_percent'] < 12) &
                         (combined['sat_score'] < 1000)]["SCHOOL NAME"])

91                   INTERNATIONAL COMMUNITY HIGH SCHOOL
125                  ACADEMY FOR LANGUAGE AND TECHNOLOGY
126                      BRONX INTERNATIONAL HIGH SCHOOL
```

```
139               KINGSBRIDGE INTERNATIONAL HIGH SCHOOL
141               INTERNATIONAL SCHOOL FOR LIBERAL ARTS
176    PAN AMERICAN INTERNATIONAL HIGH SCHOOL AT MONROE
179                      HIGH SCHOOL OF WORLD CULTURES
188                  BROOKLYN INTERNATIONAL HIGH SCHOOL
225       INTERNATIONAL HIGH SCHOOL AT PROSPECT HEIGHTS
237                          IT TAKES A VILLAGE ACADEMY
253                           MULTICULTURAL HIGH SCHOOL
286              PAN AMERICAN INTERNATIONAL HIGH SCHOOL
Name: SCHOOL NAME, dtype: object
```

A lot of students in these schools are learning English, which explains the low SAT score despite the low percent of special education students.