CustomerNYCStudentData

November 23, 2018

1 Machine Learning Engineer Nanodegree

1.1 Capstone Project

- Section **??** : Problem Definition Section
- Section **??** : Analysis Conducted
- Section **??** : Methodology Adopted
- Section **??** : Results of Analysis
- Section **??** : Inferences Drawn
- Section **??** : References Used

## Definition

- Section **??**: Overview
- Section **??** : Detailed statement
- Section **??** : Evaluation Parameter

Section 1.5.3 : Back to Top

1.1.1 Project Overview: Creating Customer Segments from NYC School Data

In this project, we are conidering data from Kaggle courtesy PASSNYC, a not-for-profit organization that facilitates a collective impact that is dedicated to broadening educational opportunities for New York City's talented and underserved students. PASSNYC uses public data to identify students within New York City's under-performing school districts and, through consulting and collaboration with partners, aims to increase the diversity of students taking the Specialized High School Admissions Test (SHSAT).

The objective is to assess the needs of students by using this data to quantify the challenges students face in taking the SHSAT. By segmenting and clustering, we can identify the schools where minority and underserved students stand to gain the most from services like after school programs, test preparation, mentoring, or resources for parents.

My personal motivation in selecting this domain is two-fold. I work for a company matching high school athletes to sports programs, so I wish to understand clustering better as applicable to school programs. Also, as an online student and micro-volunteer with school programs, it is of personal interest too.

1.1.2 Problem Statement:

In this project, the objective is to describe the variation in the different types of schools present in NYC. Based on demographic diversity or lack thereof other features of the school, the idea is to create categories of that represent subsets that together represent all the schools. Doing so will better equip these programs and how best to serve the schools.

To perform clustering on the data, we shall be using two techniques :Gaussian Mixture model and K-means clustering to create the clusters and decide on the optimal number of clusters.

1.1.3 Metrics:

Here, our solution is to determine the different segments present in the dataset. We will quantify the "goodness" of the clustering by calculating each data point's silhouette coefficient. The silhouette coefficient for a data point measures how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar). Calculating the mean silhouette coefficient will provide for a simple scoring method for our clustering and we shall take the best silhouette score of the different algorithms used. Silhouette score can be computed using sklearn library. Additionally, we will also verify the optimal number of clusters using the Elbow curve method.

1.2 Analysis

- Section **??**: Exploratory Analysis
- Section **??** :Vizzes and Exploration
- Section **??** : Methods adopted
- Section **??** : Reference Model

Section 1.5.3: Back to Top

1.2.1 Data Exploration

The dataset are taken from Kaggle at: https://www.kaggle.com/laiyipeng/target-schools-action-recommended-for-passnyc/data

This is a dataset hosted by the City of New York and has demographic statistics broken down by school districts. It contains data about the percentage of students who belong to various races like White, Latino/Hispanic, African American, Pacific Islander or Alaskan, multi racial, economically disdvantaged. Their ELA and Math test scores are also provided, both averages as well as for each grade 3-8. We also have attributes such as Economic Need Index, Student Attendance Rate, Ratings on parameters such as Rigorous Instruction, Effective School Leadership, Trust, etc. By joining on the SHSAT dataset, we also get the corresponding values of how many students registered and then attempted the test, as well as how many were finally amde an offer. Please find below the detailed exploration of the datasets.

```
In [577]: #Read and explore the datasets.
```

```
In [578]: # Import libraries necessary for this project
          import numpy as np
          import pandas as pd
          import csv
```

```
# Import supplementary visualizations code visuals.py
#import visuals as vs

# Pretty display for notebooks
%matplotlib inline

df = pd.read_csv('2016 School Explorer.csv')
df.info()
df.head(3)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1272 entries, 0 to 1271
Columns: 161 entries, Adjusted Grade to Grade 8 Math 4s - Economically Disadvantaged
dtypes: float64(5), int64(123), object(33)
memory usage: 1.6+ MB
```

```
Out[578]:   Adjusted Grade New? Other Location Code in LCGMS                School Name  \
          0            NaN  NaN                          NaN  P.S. 015 ROBERTO CLEMENTE
          1            NaN  NaN                          NaN        P.S. 019 ASHER LEVY
          2            NaN  NaN                          NaN       P.S. 020 ANNA SILVER

                 SED Code Location Code  District   Latitude  Longitude  \
          0  310100010015        01M015         1  40.721834 -73.978766
          1  310100010019        01M019         1  40.729892 -73.984231
          2  310100010020        01M020         1  40.721274 -73.986315

                              Address (Full)  \
          0  333 E 4TH ST NEW YORK, NY 10009
          1   185 1ST AVE NEW YORK, NY 10003
          2  166 ESSEX ST NEW YORK, NY 10002

                                ...                       \
          0                     ...
          1                     ...
          2                     ...

            Grade 8 Math - All Students Tested  Grade 8 Math 4s - All Students  \
          0                                  0                               0
          1                                  0                               0
          2                                  0                               0

            Grade 8 Math 4s - American Indian or Alaska Native  \
          0                                                  0
          1                                                  0
          2                                                  0
```

```
            Grade 8 Math 4s - Black or African American  \
          0                                           0
          1                                           0
          2                                           0

            Grade 8 Math 4s - Hispanic or Latino  \
          0                                    0
          1                                    0
          2                                    0

            Grade 8 Math 4s - Asian or Pacific Islander  Grade 8 Math 4s - White  \
          0                                           0                        0
          1                                           0                        0
          2                                           0                        0

            Grade 8 Math 4s - Multiracial Grade 8 Math 4s - Limited English Proficient  \
          0                             0                                            0
          1                             0                                            0
          2                             0                                            0

            Grade 8 Math 4s - Economically Disadvantaged
          0                                            0
          1                                            0
          2                                            0

          [3 rows x 161 columns]
```

```
In [579]: df1 = pd.read_csv('D5 SHSAT Registrations and Testers.csv')
          df1.info()
          df1.head(3)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 140 entries, 0 to 139
Data columns (total 7 columns):
DBN                                                140 non-null object
School name                                        140 non-null object
Year of SHST                                       140 non-null int64
Grade level                                        140 non-null int64
Enrollment on 10/31                                140 non-null int64
Number of students who registered for the SHSAT    140 non-null int64
Number of students who took the SHSAT              140 non-null int64
dtypes: int64(5), object(2)
memory usage: 7.7+ KB
```

```
Out[579]:       DBN             School name  Year of SHST  Grade level  \
          0  05M046  P.S. 046 Arthur Tappan          2013            8
          1  05M046  P.S. 046 Arthur Tappan          2014            8
          2  05M046  P.S. 046 Arthur Tappan          2015            8

             Enrollment on 10/31  Number of students who registered for the SHSAT  \
          0                   91                                               31
          1                   95                                               26
          2                   73                                               21

             Number of students who took the SHSAT
          0                                     14
          1                                      7
          2                                     10
```

```
In [580]: #2017-2018_SHSAT_Admissions_Test_Offers_By_Sending_School.csv
          df3 = pd.read_csv('2017-2018_SHSAT_Admissions_Test_Offers_By_Sending_School.csv')
          df3.info()
          df3.head(3)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 594 entries, 0 to 593
Data columns (total 5 columns):
Feeder School DBN                     594 non-null object
Feeder School Name                    594 non-null object
Count of Students in HS Admissions    594 non-null int64
Count of Testers                      594 non-null int64
Count of Offers                       594 non-null int64
dtypes: int64(3), object(2)
memory usage: 23.3+ KB
```

```
Out[580]:   Feeder School DBN              Feeder School Name  \
          0            01M034  P.S. 034 FRANKLIN D. ROOSEVELT
          1            01M140          P.S. 140 NATHAN STRAUS
          2            01M184            P.S. 184M SHUANG WEN

             Count of Students in HS Admissions  Count of Testers  Count of Offers
          0                                  58                 6                5
          1                                  67                 6                5
          2                                  88                67               23
```

```
In [581]: dfx= df.merge(df1, how='left', right_on='DBN' , left_on='Location Code')
          dfx.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1364 entries, 0 to 1363
Columns: 168 entries, Adjusted Grade to Number of students who took the SHSAT
dtypes: float64(10), int64(123), object(35)
memory usage: 1.8+ MB
```

```
In [582]: dfx.shape
```

```
Out[582]: (1364, 168)
```

```
In [583]: dfy = df.merge(df1, how='left', right_on='DBN' , left_on='Location Code')
          dfy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1364 entries, 0 to 1363
Columns: 168 entries, Adjusted Grade to Number of students who took the SHSAT
dtypes: float64(10), int64(123), object(35)
memory usage: 1.8+ MB
```

```
In [584]: dfy.head()
```

```
Out[584]:   Adjusted Grade New? Other Location Code in LCGMS  \
          0            NaN  NaN                          NaN
          1            NaN  NaN                          NaN
          2            NaN  NaN                          NaN
          3            NaN  NaN                          NaN
          4            NaN  NaN                          NaN

                                School Name      SED Code Location Code  District  \
          0       P.S. 015 ROBERTO CLEMENTE  310100010015        01M015         1
          1             P.S. 019 ASHER LEVY  310100010019        01M019         1
          2            P.S. 020 ANNA SILVER  310100010020        01M020         1
          3  P.S. 034 FRANKLIN D. ROOSEVELT  310100010034        01M034         1
          4       THE STAR ACADEMY - P.S.63  310100010063        01M063         1

              Latitude  Longitude                    Address (Full)  \
          0  40.721834 -73.978766   333 E 4TH ST NEW YORK, NY 10009
          1  40.729892 -73.984231    185 1ST AVE NEW YORK, NY 10003
          2  40.721274 -73.986315   166 ESSEX ST NEW YORK, NY 10002
          3  40.726147 -73.975043  730 E 12TH ST NEW YORK, NY 10009
          4  40.724404 -73.986360   121 E 3RD ST NEW YORK, NY 10009

                             ...                    Grade 8 Math 4s - Multiracial  \
          0                  ...                                              0
          1                  ...                                              0
          2                  ...                                              0
          3                  ...                                              0
          4                  ...                                              0

             Grade 8 Math 4s - Limited English Proficient  \
          0                                             0
          1                                             0
          2                                             0
          3                                             0
```

```
4                                              0

  Grade 8 Math 4s - Economically Disadvantaged  DBN School name Year of SHST  \
0                                             0  NaN         NaN          NaN
1                                             0  NaN         NaN          NaN
2                                             0  NaN         NaN          NaN
3                                             0  NaN         NaN          NaN
4                                             0  NaN         NaN          NaN

   Grade level Enrollment on 10/31  \
0          NaN                 NaN
1          NaN                 NaN
2          NaN                 NaN
3          NaN                 NaN
4          NaN                 NaN

  Number of students who registered for the SHSAT  \
0                                             NaN
1                                             NaN
2                                             NaN
3                                             NaN
4                                             NaN

  Number of students who took the SHSAT
0                                   NaN
1                                   NaN
2                                   NaN
3                                   NaN
4                                   NaN

[5 rows x 168 columns]
```

```
In [585]: # Display a description of the dataset
          display(dfy.describe())
```

```
           SED Code     District     Latitude    Longitude           Zip  \
count  1.364000e+03  1364.000000  1364.000000  1364.000000   1364.000000
mean   3.274443e+11    15.384164    40.739964   -73.920300  10762.730938
std    1.265045e+10     9.354700     0.086044     0.078165    548.074632
min    3.075000e+11     1.000000    40.507803   -74.244025  10001.000000
25%    3.207000e+11     7.000000    40.672706   -73.955847  10304.000000
50%    3.314000e+11    14.000000    40.736511   -73.925881  10738.000000
75%    3.332000e+11    24.000000    40.816396   -73.884761  11228.000000
max    3.531009e+11    32.000000    40.903455   -73.708920  11694.000000

       Economic Need Index  Average ELA Proficiency  Average Math Proficiency  \
count          1339.000000              1309.000000               1309.000000
mean              0.677808                 2.530527                  2.667540
```

```
std               0.205944                 0.361546                  0.473053
min               0.049000                 1.810000                  1.830000
25%               0.559500                 2.260000                  2.290000
50%               0.737000                 2.450000                  2.580000
75%               0.839000                 2.740000                  2.990000
max               0.957000                 3.930000                  4.200000

       Grade 3 ELA - All Students Tested  Grade 3 ELA 4s - All Students  \
count                        1364.000000                    1364.000000
mean                           58.475806                       4.734604
std                            57.315286                       8.084927
min                             0.000000                       0.000000
25%                             0.000000                       0.000000
50%                            52.000000                       1.000000
75%                            93.000000                       6.000000
max                           356.000000                      55.000000

                       ...                    \
count                  ...
mean                   ...
std                    ...
min                    ...
25%                    ...
50%                    ...
75%                    ...
max                    ...

       Grade 8 Math 4s - Asian or Pacific Islander  Grade 8 Math 4s - White  \
count                                  1364.000000              1364.000000
mean                                      1.850440                 0.905425
std                                      12.410368                 6.648429
min                                       0.000000                 0.000000
25%                                       0.000000                 0.000000
50%                                       0.000000                 0.000000
75%                                       0.000000                 0.000000
max                                     246.000000               126.000000

       Grade 8 Math 4s - Multiracial  \
count                    1364.000000
mean                        0.002199
std                         0.081230
min                         0.000000
25%                         0.000000
50%                         0.000000
75%                         0.000000
max                         3.000000

       Grade 8 Math 4s - Limited English Proficient  \
count                                        1364.000000
mean                                            0.148827
std                                             1.276455
min                                             0.000000
25%                                             0.000000
50%                                             0.000000
75%                                             0.000000
max                                            33.000000

       Grade 8 Math 4s - Economically Disadvantaged  Year of SHST  \
count                                        1364.000000    113.000000
mean                                            3.000000   2014.601770
std                                            12.312154      1.122337
min                                             0.000000   2013.000000
25%                                             0.000000   2014.000000
50%                                             0.000000   2015.000000
75%                                             1.000000   2016.000000
max                                           196.000000   2016.000000

       Grade level  Enrollment on 10/31  \
count   113.000000           113.000000
mean      8.309735            93.486726
std       0.464444            48.200310
min       8.000000            35.000000
25%       8.000000            66.000000
50%       8.000000            86.000000
75%       9.000000           110.000000
max       9.000000           344.000000

       Number of students who registered for the SHSAT  \
count                                       113.000000
mean                                         22.716814
std                                          24.419865
min                                           0.000000
25%                                           4.000000
50%                                          17.000000
75%                                          30.000000
max                                         118.000000

       Number of students who took the SHSAT
count                             113.000000
mean                               11.442478
std                                10.398264
min                                 0.000000
25%                                 3.000000
50%                                10.000000
75%                                17.000000
max                                45.000000
```

```
[8 rows x 133 columns]
```

```
In [586]: # Display a description of the dataset
          display(dfx.describe())
```

```
           SED Code     District     Latitude    Longitude           Zip  \
count  1.364000e+03  1364.000000  1364.000000  1364.000000   1364.000000
mean   3.274443e+11    15.384164    40.739964   -73.920300  10762.730938
std    1.265045e+10     9.354700     0.086044     0.078165    548.074632
min    3.075000e+11     1.000000    40.507803   -74.244025  10001.000000
25%    3.207000e+11     7.000000    40.672706   -73.955847  10304.000000
50%    3.314000e+11    14.000000    40.736511   -73.925881  10738.000000
75%    3.332000e+11    24.000000    40.816396   -73.884761  11228.000000
max    3.531009e+11    32.000000    40.903455   -73.708920  11694.000000

       Economic Need Index  Average ELA Proficiency  Average Math Proficiency  \
count          1339.000000              1309.000000               1309.000000
mean              0.677808                 2.530527                  2.667540
std               0.205944                 0.361546                  0.473053
min               0.049000                 1.810000                  1.830000
25%               0.559500                 2.260000                  2.290000
50%               0.737000                 2.450000                  2.580000
75%               0.839000                 2.740000                  2.990000
max               0.957000                 3.930000                  4.200000

       Grade 3 ELA - All Students Tested  Grade 3 ELA 4s - All Students  \
count                        1364.000000                    1364.000000
mean                           58.475806                       4.734604
std                            57.315286                       8.084927
min                             0.000000                       0.000000
25%                             0.000000                       0.000000
50%                            52.000000                       1.000000
75%                            93.000000                       6.000000
max                           356.000000                      55.000000

                       ...                     \
count                  ...
mean                   ...
std                    ...
min                    ...
25%                    ...
50%                    ...
75%                    ...
max                    ...

       Grade 8 Math 4s - Asian or Pacific Islander  Grade 8 Math 4s - White  \
count                                   1364.000000              1364.000000
mean                                       1.850440                 0.905425
std                                       12.410368                 6.648429
min                                        0.000000                 0.000000
25%                                        0.000000                 0.000000
50%                                        0.000000                 0.000000
75%                                        0.000000                 0.000000
max                                      246.000000               126.000000

       Grade 8 Math 4s - Multiracial  \
count                    1364.000000
mean                        0.002199
std                         0.081230
min                         0.000000
25%                         0.000000
50%                         0.000000
75%                         0.000000
max                         3.000000

       Grade 8 Math 4s - Limited English Proficient  \
count                                   1364.000000
mean                                       0.148827
std                                        1.276455
min                                        0.000000
25%                                        0.000000
50%                                        0.000000
75%                                        0.000000
max                                       33.000000

       Grade 8 Math 4s - Economically Disadvantaged  Year of SHST  \
count                                   1364.000000    113.000000
mean                                       3.000000   2014.601770
std                                       12.312154      1.122337
min                                        0.000000   2013.000000
25%                                        0.000000   2014.000000
50%                                        0.000000   2015.000000
75%                                        1.000000   2016.000000
max                                      196.000000   2016.000000

       Grade level  Enrollment on 10/31  \
count   113.000000           113.000000
mean      8.309735            93.486726
std       0.464444            48.200310
min       8.000000            35.000000
25%       8.000000            66.000000
50%       8.000000            86.000000
75%       9.000000           110.000000
max       9.000000           344.000000
```

```
       Number of students who registered for the SHSAT  \
count                                       113.000000
mean                                         22.716814
std                                          24.419865
min                                           0.000000
25%                                           4.000000
50%                                          17.000000
75%                                          30.000000
max                                         118.000000

       Number of students who took the SHSAT
count                             113.000000
mean                               11.442478
std                                10.398264
min                                 0.000000
25%                                 3.000000
50%                                10.000000
75%                                17.000000
max                                45.000000

[8 rows x 133 columns]
```

```
In [587]: dfy['AllTested'] =  dfy[['Grade 3 ELA - All Students Tested',
                                   'Grade 4 ELA - All Students Tested',
                                   'Grade 5 ELA - All Students Tested',
                                  'Grade 6 ELA - All Students Tested',
                                  'Grade 7 ELA - All Students Tested',
                                  'Grade 8 ELA - All Students Tested']].mean(axis=1)

          dfy['All4'] =  dfy[['Grade 3 ELA 4s - All Students',
                              'Grade 4 ELA 4s - All Students',
                             'Grade 5 ELA 4s - All Students',
                             'Grade 6 ELA 4s - All Students',
                             'Grade 7 ELA 4s - All Students',
                             'Grade 8 ELA 4s - All Students']].mean(axis=1)

          dfy['Native4'] =  dfy[['Grade 3 ELA 4s - American Indian or Alaska Native',
                                 'Grade 4 ELA 4s - American Indian or Alaska Native',
                                'Grade 5 ELA 4s - American Indian or Alaska Native',
                                'Grade 6 ELA 4s - American Indian or Alaska Native',
                                'Grade 7 ELA 4s - American Indian or Alaska Native',
                                'Grade 8 ELA 4s - American Indian or Alaska Native']].mean(axi

          dfy['AfricanAmerican4'] =  dfy[['Grade 3 ELA 4s - Black or African American',
                                          'Grade 4 ELA 4s - Black or African American',
                                         'Grade 5 ELA 4s - Black or African American',
                                         'Grade 6 ELA 4s - Black or African American',
                                         'Grade 7 ELA 4s - Black or African American',
                                         'Grade 8 ELA 4s - Black or African American']].mean(a

          dfy['Latino4'] =  dfy[['Grade 3 ELA 4s - Hispanic or Latino',
                                 'Grade 4 ELA 4s - Hispanic or Latino',
                                'Grade 5 ELA 4s - Hispanic or Latino',
                                'Grade 6 ELA 4s - Hispanic or Latino',
                                'Grade 7 ELA 4s - Hispanic or Latino',
                                'Grade 8 ELA 4s - Hispanic or Latino']].mean(axis=1)

          dfy['Islander4'] =  dfy[['Grade 3 ELA 4s - Asian or Pacific Islander',
                                   'Grade 4 ELA 4s - Asian or Pacific Islander',
                                  'Grade 5 ELA 4s - Asian or Pacific Islander',
                                  'Grade 6 ELA 4s - Asian or Pacific Islander',
                                  'Grade 7 ELA 4s - Asian or Pacific Islander',
                                  'Grade 8 ELA 4s - Asian or Pacific Islander']].mean(axis=1)

          dfy['White4'] = dfy[['Grade 3 ELA 4s - White',
                               'Grade 4 ELA 4s - White',
                              'Grade 5 ELA 4s - White',
                              'Grade 6 ELA 4s - White',
                              'Grade 7 ELA 4s - White',
                              'Grade 8 ELA 4s - White']].mean(axis=1)

          dfy['Multiracial4'] = dfy[['Grade 3 ELA 4s - Multiracial',
                                     'Grade 4 ELA 4s - Multiracial',
                                    'Grade 5 ELA 4s - Multiracial',
                                    'Grade 6 ELA 4s - Multiracial',
                                    'Grade 7 ELA 4s - Multiracial',
                                    'Grade 8 ELA 4s - Multiracial']].mean(axis=1)

          dfy['LimitedEnglish4'] = dfy[['Grade 3 ELA 4s - Limited English Proficient',
                                        'Grade 4 ELA 4s - Limited English Proficient',
                                       'Grade 5 ELA 4s - Limited English Proficient',
                                       'Grade 6 ELA 4s - Limited English Proficient',
                                       'Grade 7 ELA 4s - Limited English Proficient',
                                       'Grade 8 ELA 4s - Limited English Proficient']].mean(ax

          dfy['Disadv4'] = dfy[['Grade 3 ELA 4s - Economically Disadvantaged',
                                'Grade 4 ELA 4s - Economically Disadvantaged',
                               'Grade 5 ELA 4s - Economically Disadvantaged',
                               'Grade 6 ELA 4s - Economically Disadvantaged',
                                'Grade 7 ELA 4s - Economically Disadvantaged',
                               'Grade 8 ELA 4s - Economically Disadvantaged']].mean(axis=1)

          dfy['AllMath4Tested'] =  dfy[['Grade 3 Math - All Students tested',
                                        'Grade 4 Math - All Students Tested',
                                       'Grade 5 Math - All Students Tested',
                                       'Grade 6 Math - All Students Tested',
                                       'Grade 7 Math - All Students Tested',
                                       'Grade 8 Math - All Students Tested']].mean(axis=1)

          dfy['AllMath4'] =  dfy[['Grade 3 Math 4s - All Students',
                                  'Grade 4 Math 4s - All Students',
                                 'Grade 5 Math 4s - All Students',
                                 'Grade 6 Math 4s - All Students',
                                 'Grade 7 Math 4s - All Students',
                                 'Grade 8 Math 4s - All Students']].mean(axis=1)

          dfy['NativeMath4'] =  dfy[['Grade 3 Math 4s - American Indian or Alaska Native',
                                     'Grade 4 Math 4s - American Indian or Alaska Native',
                                  'Grade 5 Math 4s - American Indian or Alaska Native',
                                    'Grade 6 Math 4s - American Indian or Alaska Native',
                                    'Grade 7 Math 4s - American Indian or Alaska Native',
                                    'Grade 8 Math 4s - American Indian or Alaska Native']].mea

          dfy['AfricanAmericanMath4'] =  dfy[['Grade 3 Math 4s - Black or African American',
                                              'Grade 4 Math 4s - Black or African American',
                                             'Grade 5 Math 4s - Black or African American',
                                             'Grade 6 Math 4s - Black or African American',
                                             'Grade 7 Math 4s - Black or African American',
                                             'Grade 8 Math 4s - Black or African American']].m

          dfy['LatinoMath4'] =  dfy[['Grade 3 Math 4s - Hispanic or Latino',
                                     'Grade 4 Math 4s - Hispanic or Latino',
                                    'Grade 5 Math 4s - Hispanic or Latino',
                                    'Grade 6 Math 4s - Hispanic or Latino',
                                    'Grade 7 Math 4s - Hispanic or Latino',
                                    'Grade 8 Math 4s - Hispanic or Latino']].mean(axis=1)

          dfy['IslanderMath4'] =  dfy[['Grade 3 Math 4s - Asian or Pacific Islander',
                                       'Grade 4 Math 4s - Asian or Pacific Islander',
                                      'Grade 5 Math 4s - Asian or Pacific Islander',
                                      'Grade 6 Math 4s - Asian or Pacific Islander',
                                      'Grade 7 Math 4s - Asian or Pacific Islander',
                                      'Grade 8 Math 4s - Asian or Pacific Islander']].mean(axi

          dfy['WhiteMath4'] = dfy[['Grade 3 Math 4s - White',
                                   'Grade 4 Math 4s - White',
                                  'Grade 5 Math 4s - White',
                                  'Grade 6 Math 4s - White',
                                  'Grade 7 Math 4s - White',
                                  'Grade 8 Math 4s - White']].mean(axis=1)

          dfy['MultiracialMath4'] = dfy[['Grade 3 ELA 4s - Multiracial',
                                         'Grade 4 Math 4s - Multiracial',
                                        'Grade 5 Math 4s - Multiracial',
                                        'Grade 6 Math 4s - Multiracial',
                                        'Grade 7 Math 4s - Multiracial',
                                        'Grade 8 Math 4s - Multiracial']].mean(axis=1)

          dfy['LimitedEnglishMath4'] = dfy[['Grade 3 Math 4s - Limited English Proficient',
                                            'Grade 4 Math 4s - Limited English Proficient',
                                           'Grade 5 Math 4s - Limited English Proficient',
                                           'Grade 6 Math 4s - Limited English Proficient',
                                           'Grade 7 Math 4s - Limited English Proficient',
                                           'Grade 8 Math 4s - Limited English Proficient']].mea

          dfy['DisadvMath4'] = dfy[['Grade 3 Math 4s - Economically Disadvantaged',
                                    'Grade 4 Math 4s - Economically Disadvantaged',
                                   'Grade 5 Math 4s - Economically Disadvantaged',
                                   'Grade 6 Math 4s - Economically Disadvantaged',
                                   'Grade 7 Math 4s - Economically Disadvantaged',
                                   'Grade 8 Math 4s - Economically Disadvantaged']].mean(axis=

In [588]: # drop columns
          dfy= dfy.drop(['Grade 3 ELA - All Students Tested',
                   'Grade 4 ELA - All Students Tested',
                   'Grade 5 ELA - All Students Tested',
                  'Grade 6 ELA - All Students Tested',
                  'Grade 7 ELA - All Students Tested',
                  'Grade 8 ELA - All Students Tested',
                  'Grade 3 ELA 4s - All Students',
                  'Grade 4 ELA 4s - All Students',
                 'Grade 5 ELA 4s - All Students',
                 'Grade 6 ELA 4s - All Students',
                 'Grade 7 ELA 4s - All Students',
                 'Grade 8 ELA 4s - All Students',
                'Grade 3 ELA 4s - American Indian or Alaska Native',
                 'Grade 4 ELA 4s - American Indian or Alaska Native',
                'Grade 5 ELA 4s - American Indian or Alaska Native',
                'Grade 6 ELA 4s - American Indian or Alaska Native',
                'Grade 7 ELA 4s - American Indian or Alaska Native',
                'Grade 8 ELA 4s - American Indian or Alaska Native',
                        'Grade 3 ELA 4s - Black or African American',
                                          'Grade 4 ELA 4s - Black or African American',
                                         'Grade 5 ELA 4s - Black or African American',
                                         'Grade 6 ELA 4s - Black or African American',
                                         'Grade 7 ELA 4s - Black or African American',
                                         'Grade 8 ELA 4s - Black or African American',
          'Grade 3 ELA 4s - Hispanic or Latino',
                                 'Grade 4 ELA 4s - Hispanic or Latino',
                                'Grade 5 ELA 4s - Hispanic or Latino',
                                'Grade 6 ELA 4s - Hispanic or Latino',
          'Grade 7 ELA 4s - Hispanic or Latino',
                                'Grade 8 ELA 4s - Hispanic or Latino',
          'Grade 3 ELA 4s - Asian or Pacific Islander',
                                   'Grade 4 ELA 4s - Asian or Pacific Islander',
                                  'Grade 5 ELA 4s - Asian or Pacific Islander',
                                  'Grade 6 ELA 4s - Asian or Pacific Islander',
                                  'Grade 7 ELA 4s - Asian or Pacific Islander',
                                  'Grade 8 ELA 4s - Asian or Pacific Islander',
          'Grade 3 ELA 4s - White',
                               'Grade 4 ELA 4s - White',
                              'Grade 5 ELA 4s - White',
                              'Grade 6 ELA 4s - White',
                              'Grade 7 ELA 4s - White',
                              'Grade 8 ELA 4s - White',
          'Grade 3 ELA 4s - Multiracial',
                                     'Grade 4 ELA 4s - Multiracial',
                                    'Grade 5 ELA 4s - Multiracial',
                                    'Grade 6 ELA 4s - Multiracial',
                                    'Grade 7 ELA 4s - Multiracial',
                                    'Grade 8 ELA 4s - Multiracial',
          'Grade 3 ELA 4s - Limited English Proficient',
                                        'Grade 4 ELA 4s - Limited English Proficient',
                                       'Grade 5 ELA 4s - Limited English Proficient',
                                       'Grade 6 ELA 4s - Limited English Proficient',
                                       'Grade 7 ELA 4s - Limited English Proficient',
                                       'Grade 8 ELA 4s - Limited English Proficient',
          'Grade 3 ELA 4s - Economically Disadvantaged',
                                'Grade 4 ELA 4s - Economically Disadvantaged',
                               'Grade 5 ELA 4s - Economically Disadvantaged',
                               'Grade 6 ELA 4s - Economically Disadvantaged',
                                'Grade 7 ELA 4s - Economically Disadvantaged',
                               'Grade 8 ELA 4s - Economically Disadvantaged',
          'Grade 3 Math - All Students tested',
                                        'Grade 4 Math - All Students Tested',
                                       'Grade 5 Math - All Students Tested',
                                       'Grade 6 Math - All Students Tested',
                                       'Grade 7 Math - All Students Tested',
                                       'Grade 8 Math - All Students Tested',
          'Grade 3 Math 4s - All Students',
                                  'Grade 4 Math 4s - All Students',
                                 'Grade 5 Math 4s - All Students',
                                 'Grade 6 Math 4s - All Students',
                                 'Grade 7 Math 4s - All Students',
                                 'Grade 8 Math 4s - All Students',
          'Grade 3 Math 4s - American Indian or Alaska Native',
                                     'Grade 4 Math 4s - American Indian or Alaska Native',
                                  'Grade 5 Math 4s - American Indian or Alaska Native',
                                    'Grade 6 Math 4s - American Indian or Alaska Native',
                                    'Grade 7 Math 4s - American Indian or Alaska Native',
                                    'Grade 8 Math 4s - American Indian or Alaska Native',
          'Grade 3 Math 4s - Black or African American',
                                              'Grade 4 Math 4s - Black or African American',
                                             'Grade 5 Math 4s - Black or African American',
                                             'Grade 6 Math 4s - Black or African American',
                                             'Grade 7 Math 4s - Black or African American',
                                             'Grade 8 Math 4s - Black or African American',
          'Grade 3 Math 4s - Hispanic or Latino',
                                     'Grade 4 Math 4s - Hispanic or Latino',
                                    'Grade 5 Math 4s - Hispanic or Latino',
                                    'Grade 6 Math 4s - Hispanic or Latino',
                                    'Grade 7 Math 4s - Hispanic or Latino',
                                    'Grade 8 Math 4s - Hispanic or Latino',
          'Grade 3 Math 4s - Asian or Pacific Islander',
                                       'Grade 4 Math 4s - Asian or Pacific Islander',
                                      'Grade 5 Math 4s - Asian or Pacific Islander',
                                      'Grade 6 Math 4s - Asian or Pacific Islander',
                                      'Grade 7 Math 4s - Asian or Pacific Islander',
                                      'Grade 8 Math 4s - Asian or Pacific Islander',
          'Grade 3 Math 4s - White',
                                   'Grade 4 Math 4s - White',
                                  'Grade 5 Math 4s - White',
                                  'Grade 6 Math 4s - White',
                                  'Grade 7 Math 4s - White',
                                  'Grade 8 Math 4s - White',
          'Grade 3 ELA 4s - Multiracial',
                                         'Grade 4 Math 4s - Multiracial',
                                        'Grade 5 Math 4s - Multiracial',
                                        'Grade 6 Math 4s - Multiracial',
                                        'Grade 7 Math 4s - Multiracial',
                                        'Grade 8 Math 4s - Multiracial',
          'Grade 3 Math 4s - Limited English Proficient',
                                            'Grade 4 Math 4s - Limited English Proficient',
                                           'Grade 5 Math 4s - Limited English Proficient',
                                           'Grade 6 Math 4s - Limited English Proficient',
                                           'Grade 7 Math 4s - Limited English Proficient',
                                           'Grade 8 Math 4s - Limited English Proficient',
          'Grade 3 Math 4s - Economically Disadvantaged',
                                    'Grade 4 Math 4s - Economically Disadvantaged',
                                   'Grade 5 Math 4s - Economically Disadvantaged',
                                   'Grade 6 Math 4s - Economically Disadvantaged',
                                   'Grade 7 Math 4s - Economically Disadvantaged',
                                   'Grade 8 Math 4s - Economically Disadvantaged',
                        'Adjusted Grade',
                        'New?','Other Location Code in LCGMS'], axis=1)
```

```
In [589]: dfy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1364 entries, 0 to 1363
Data columns (total 66 columns):
School Name                                        1364 non-null object
SED Code                                           1364 non-null int64
Location Code                                      1364 non-null object
District                                           1364 non-null int64
Latitude                                           1364 non-null float64
Longitude                                          1364 non-null float64
Address (Full)                                     1364 non-null object
City                                               1364 non-null object
Zip                                                1364 non-null int64
Grades                                             1364 non-null object
Grade Low                                          1364 non-null object
Grade High                                         1364 non-null object
Community School?                                  1364 non-null object
Economic Need Index                                1339 non-null float64
School Income Estimate                             906 non-null object
Percent ELL                                        1364 non-null object
Percent Asian                                      1364 non-null object
Percent Black                                      1364 non-null object
Percent Hispanic                                   1364 non-null object
Percent Black / Hispanic                           1364 non-null object
Percent White                                      1364 non-null object
Student Attendance Rate                            1339 non-null object
Percent of Students Chronically Absent             1339 non-null object
Rigorous Instruction %                             1339 non-null object
Rigorous Instruction Rating                        1288 non-null object
Collaborative Teachers %                           1339 non-null object
Collaborative Teachers Rating                      1288 non-null object
Supportive Environment %                           1339 non-null object
Supportive Environment Rating                      1284 non-null object
Effective School Leadership %                      1339 non-null object
Effective School Leadership Rating                 1291 non-null object
Strong Family-Community Ties %                     1339 non-null object
Strong Family-Community Ties Rating                1291 non-null object
Trust %                                            1339 non-null object
Trust Rating                                       1291 non-null object
Student Achievement Rating                         1278 non-null object
Average ELA Proficiency                            1309 non-null float64
Average Math Proficiency                           1309 non-null float64
Grade 3 Math 4s - Multiracial                      1364 non-null int64
DBN                                                113 non-null object
School name                                        113 non-null object
Year of SHST                                       113 non-null float64
Grade level                                        113 non-null float64
```

```
Enrollment on 10/31                                   113 non-null float64
Number of students who registered for the SHSAT     113 non-null float64
Number of students who took the SHSAT                 113 non-null float64
AllTested                                             1364 non-null float64
All4                                                  1364 non-null float64
Native4                                               1364 non-null float64
AfricanAmerican4                                      1364 non-null float64
Latino4                                               1364 non-null float64
Islander4                                             1364 non-null float64
White4                                                1364 non-null float64
Multiracial4                                          1364 non-null float64
LimitedEnglish4                                       1364 non-null float64
Disadv4                                               1364 non-null float64
AllMath4Tested                                        1364 non-null float64
AllMath4                                              1364 non-null float64
NativeMath4                                           1364 non-null float64
AfricanAmericanMath4                                  1364 non-null float64
LatinoMath4                                           1364 non-null float64
IslanderMath4                                         1364 non-null float64
WhiteMath4                                            1364 non-null float64
MultiracialMath4                                      1364 non-null float64
LimitedEnglishMath4                                   1364 non-null float64
DisadvMath4                                           1364 non-null float64
dtypes: float64(30), int64(4), object(32)
memory usage: 714.0+ KB
```

```
In [590]: dfy.describe()
```

```
Out[590]:            SED Code     District     Latitude    Longitude          Zip  \
          count  1.364000e+03  1364.000000  1364.000000  1364.000000   1364.000000
          mean   3.274443e+11    15.384164    40.739964   -73.920300  10762.730938
          std    1.265045e+10     9.354700     0.086044     0.078165    548.074632
          min    3.075000e+11     1.000000    40.507803   -74.244025  10001.000000
          25%    3.207000e+11     7.000000    40.672706   -73.955847  10304.000000
          50%    3.314000e+11    14.000000    40.736511   -73.925881  10738.000000
          75%    3.332000e+11    24.000000    40.816396   -73.884761  11228.000000
          max    3.531009e+11    32.000000    40.903455   -73.708920  11694.000000

                 Economic Need Index  Average ELA Proficiency  Average Math Proficiency  \
          count          1339.000000              1309.000000               1309.000000
          mean              0.677808                 2.530527                  2.667540
          std               0.205944                 0.361546                  0.473053
          min               0.049000                 1.810000                  1.830000
          25%               0.559500                 2.260000                  2.290000
          50%               0.737000                 2.450000                  2.580000
          75%               0.839000                 2.740000                  2.990000
          max               0.957000                 3.930000                  4.200000
```

```
       Grade 3 Math 4s - Multiracial  Year of SHST     ...       \
count                    1364.000000    113.000000     ...
mean                        0.061584   2014.601770     ...
std                         0.538215      1.122337     ...
min                         0.000000   2013.000000     ...
25%                         0.000000   2014.000000     ...
50%                         0.000000   2015.000000     ...
75%                         0.000000   2016.000000     ...
max                         8.000000   2016.000000     ...

       AllMath4Tested     AllMath4  NativeMath4  AfricanAmericanMath4  \
count     1364.000000  1364.000000  1364.000000           1364.000000
mean        55.016984    10.307674     0.021505              1.526393
std         41.252793    15.106568     0.179063              4.058560
min          0.000000     0.000000     0.000000              0.000000
25%         29.500000     1.333333     0.000000              0.000000
50%         46.166667     4.333333     0.000000              0.166667
75%         67.208333    13.833333     0.000000              1.166667
max        330.333333   151.666667     3.500000             61.166667

       LatinoMath4  IslanderMath4   WhiteMath4  MultiracialMath4  \
count  1364.000000    1364.000000  1364.000000       1364.000000
mean      2.028715       3.249756     2.288368          0.046676
std       3.480549       8.994379     6.348200          0.292881
min       0.000000       0.000000     0.000000          0.000000
25%       0.166667       0.000000     0.000000          0.000000
50%       0.833333       0.000000     0.000000          0.000000
75%       2.500000       1.500000     0.833333          0.000000
max      29.666667     107.666667    88.000000          6.666667

       LimitedEnglishMath4  DisadvMath4
count          1364.000000  1364.000000
mean              0.262830     5.353372
std               0.912010     8.654386
min               0.000000     0.000000
25%               0.000000     0.666667
50%               0.000000     2.166667
75%               0.166667     6.375000
max              14.333333    88.166667

[8 rows x 34 columns]
```

```
In [591]: dfy = dfy.merge(df3, how='left', right_on='Feeder School DBN' , left_on='Location Co
          dfy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1364 entries, 0 to 1363
```

```
Data columns (total 71 columns):
School Name                                        1364 non-null object
SED Code                                           1364 non-null int64
Location Code                                      1364 non-null object
District                                           1364 non-null int64
Latitude                                           1364 non-null float64
Longitude                                          1364 non-null float64
Address (Full)                                     1364 non-null object
City                                               1364 non-null object
Zip                                                1364 non-null int64
Grades                                             1364 non-null object
Grade Low                                          1364 non-null object
Grade High                                         1364 non-null object
Community School?                                  1364 non-null object
Economic Need Index                                1339 non-null float64
School Income Estimate                             906 non-null object
Percent ELL                                        1364 non-null object
Percent Asian                                      1364 non-null object
Percent Black                                      1364 non-null object
Percent Hispanic                                   1364 non-null object
Percent Black / Hispanic                           1364 non-null object
Percent White                                      1364 non-null object
Student Attendance Rate                            1339 non-null object
Percent of Students Chronically Absent             1339 non-null object
Rigorous Instruction %                             1339 non-null object
Rigorous Instruction Rating                        1288 non-null object
Collaborative Teachers %                           1339 non-null object
Collaborative Teachers Rating                      1288 non-null object
Supportive Environment %                           1339 non-null object
Supportive Environment Rating                      1284 non-null object
Effective School Leadership %                      1339 non-null object
Effective School Leadership Rating                 1291 non-null object
Strong Family-Community Ties %                     1339 non-null object
Strong Family-Community Ties Rating                1291 non-null object
Trust %                                            1339 non-null object
Trust Rating                                       1291 non-null object
Student Achievement Rating                         1278 non-null object
Average ELA Proficiency                            1309 non-null float64
Average Math Proficiency                           1309 non-null float64
Grade 3 Math 4s - Multiracial                      1364 non-null int64
DBN                                                113 non-null object
School name                                        113 non-null object
Year of SHST                                       113 non-null float64
Grade level                                        113 non-null float64
Enrollment on 10/31                                113 non-null float64
Number of students who registered for the SHSAT    113 non-null float64
Number of students who took the SHSAT              113 non-null float64
AllTested                                          1364 non-null float64
```

```
All4                                               1364 non-null float64
Native4                                            1364 non-null float64
AfricanAmerican4                                   1364 non-null float64
Latino4                                            1364 non-null float64
Islander4                                          1364 non-null float64
White4                                             1364 non-null float64
Multiracial4                                       1364 non-null float64
LimitedEnglish4                                    1364 non-null float64
Disadv4                                            1364 non-null float64
AllMath4Tested                                     1364 non-null float64
AllMath4                                           1364 non-null float64
NativeMath4                                        1364 non-null float64
AfricanAmericanMath4                               1364 non-null float64
LatinoMath4                                        1364 non-null float64
IslanderMath4                                      1364 non-null float64
WhiteMath4                                         1364 non-null float64
MultiracialMath4                                   1364 non-null float64
LimitedEnglishMath4                                1364 non-null float64
DisadvMath4                                        1364 non-null float64
Feeder School DBN                                  683 non-null object
Feeder School Name                                 683 non-null object
Count of Students in HS Admissions                 683 non-null float64
Count of Testers                                   683 non-null float64
Count of Offers                                    683 non-null float64
dtypes: float64(33), int64(4), object(34)
memory usage: 767.2+ KB
```

```
In [592]: list(dfy)
```

```
Out[592]: ['School Name',
           'SED Code',
           'Location Code',
           'District',
           'Latitude',
           'Longitude',
           'Address (Full)',
           'City',
           'Zip',
           'Grades',
           'Grade Low',
           'Grade High',
           'Community School?',
           'Economic Need Index',
           'School Income Estimate',
           'Percent ELL',
           'Percent Asian',
           'Percent Black',
           'Percent Hispanic',
           'Percent Black / Hispanic',
           'Percent White',
           'Student Attendance Rate',
           'Percent of Students Chronically Absent',
           'Rigorous Instruction %',
           'Rigorous Instruction Rating',
           'Collaborative Teachers %',
           'Collaborative Teachers Rating',
           'Supportive Environment %',
           'Supportive Environment Rating',
           'Effective School Leadership %',
           'Effective School Leadership Rating',
           'Strong Family-Community Ties %',
           'Strong Family-Community Ties Rating',
           'Trust %',
           'Trust Rating',
           'Student Achievement Rating',
           'Average ELA Proficiency',
           'Average Math Proficiency',
           'Grade 3 Math 4s - Multiracial',
           'DBN',
           'School name',
           'Year of SHST',
           'Grade level',
           'Enrollment on 10/31',
           'Number of students who registered for the SHSAT',
           'Number of students who took the SHSAT',
           'AllTested',
           'All4',
           'Native4',
           'AfricanAmerican4',
           'Latino4',
           'Islander4',
           'White4',
           'Multiracial4',
           'LimitedEnglish4',
           'Disadv4',
           'AllMath4Tested',
           'AllMath4',
           'NativeMath4',
           'AfricanAmericanMath4',
           'LatinoMath4',
           'IslanderMath4',
           'WhiteMath4',
           'MultiracialMath4',
           'LimitedEnglishMath4',
           'DisadvMath4',
           'Feeder School DBN',
           'Feeder School Name',
           'Count of Students in HS Admissions',
           'Count of Testers',
           'Count of Offers']
```

```
In [593]: dfy.head(3)
```

```
Out[593]:                  School Name      SED Code Location Code  District   Latitude  \
          0  P.S. 015 ROBERTO CLEMENTE  310100010015        01M015         1  40.721834
          1        P.S. 019 ASHER LEVY  310100010019        01M019         1  40.729892
          2       P.S. 020 ANNA SILVER  310100010020        01M020         1  40.721274

             Longitude                    Address (Full)      City    Zip  \
          0 -73.978766  333 E 4TH ST NEW YORK, NY 10009  NEW YORK  10009
          1 -73.984231   185 1ST AVE NEW YORK, NY 10003  NEW YORK  10003
          2 -73.986315  166 ESSEX ST NEW YORK, NY 10002  NEW YORK  10002

                            Grades       ...       IslanderMath4 WhiteMath4  \
          0  PK,0K,01,02,03,04,05       ...            0.000000   0.000000
          1  PK,0K,01,02,03,04,05       ...            0.333333   0.000000
          2  PK,0K,01,02,03,04,05       ...            4.500000   0.333333

            MultiracialMath4  LimitedEnglishMath4 DisadvMath4 Feeder School DBN  \
          0              0.0             0.000000    0.000000               NaN
          1              0.0             0.000000    2.500000               NaN
          2              0.0             0.166667    2.666667               NaN

            Feeder School Name Count of Students in HS Admissions Count of Testers  \
          0                NaN                                NaN              NaN
          1                NaN                                NaN              NaN
          2                NaN                                NaN              NaN

            Count of Offers
          0             NaN
          1             NaN
          2             NaN

          [3 rows x 71 columns]
```

```
In [594]: dfy.corr()
```

```
Out[594]:                                                  SED Code  District  Latitude  \
          SED Code                                           1.000000  0.954898 -0.661821
          District                                           0.954898  1.000000 -0.642761
          Latitude                                          -0.661821 -0.642761  1.000000
          Longitude                                          0.118464  0.178117  0.291984
          Zip                                                0.775995  0.772984 -0.604029
          Economic Need Index                               -0.311408 -0.283162  0.306643
```

```
Average ELA Proficiency                           0.128419  0.111628 -0.195998
Average Math Proficiency                          0.117325  0.108624 -0.166521
Grade 3 Math 4s - Multiracial                    -0.039204 -0.054976 -0.026301
Year of SHST                                      0.033258       NaN -0.052813
Grade level                                       0.154903       NaN -0.159922
Enrollment on 10/31                              -0.144704       NaN  0.339524
Number of students who registered for the SHSAT  0.011083       NaN  0.032252
Number of students who took the SHSAT             0.007385       NaN -0.121631
AllTested                                         0.238549  0.248214 -0.106065
All4                                              0.181169  0.176803 -0.183555
Native4                                           0.109187  0.113853 -0.056912
AfricanAmerican4                                 -0.008202  0.006479 -0.069582
Latino4                                           0.084169  0.090725  0.091737
Islander4                                         0.217653  0.220223 -0.129859
White4                                            0.109849  0.090336 -0.233920
Multiracial4                                     -0.101175 -0.103340 -0.021333
LimitedEnglish4                                   0.063333  0.051208 -0.025918
Disadv4                                           0.222882  0.238323 -0.154559
AllMath4Tested                                    0.237845  0.249322 -0.100759
AllMath4                                          0.144911  0.145041 -0.162547
NativeMath4                                       0.103002  0.107702 -0.049700
AfricanAmericanMath4                             -0.121024 -0.104757  0.030972
LatinoMath4                                       0.005167  0.016430  0.130747
IslanderMath4                                     0.222771  0.223257 -0.153146
WhiteMath4                                        0.111477  0.094061 -0.248455
MultiracialMath4                                 -0.100134 -0.097328 -0.023081
LimitedEnglishMath4                               0.101698  0.097299 -0.099548
DisadvMath4                                       0.172113  0.189524 -0.126465
Count of Students in HS Admissions                0.355074  0.331419 -0.202395
Count of Testers                                  0.281210  0.260098 -0.220273
Count of Offers                                   0.119194  0.109803 -0.145873

                                                 Longitude       Zip  \
SED Code                                          0.118464  0.775995
District                                          0.178117  0.772984
Latitude                                          0.291984 -0.604029
Longitude                                         1.000000  0.411442
Zip                                               0.411442  1.000000
Economic Need Index                               0.015074 -0.165680
Average ELA Proficiency                          -0.088961  0.059880
Average Math Proficiency                         -0.080416  0.060285
Grade 3 Math 4s - Multiracial                    -0.095886 -0.026360
Year of SHST                                      0.020974 -0.027054
Grade level                                       0.139795  0.054755
Enrollment on 10/31                               0.233728  0.367929
Number of students who registered for the SHSAT -0.399194 -0.209484
Number of students who took the SHSAT            -0.259444 -0.148144
AllTested                                         0.023724  0.160152
```

```
All4                                             -0.055796  0.129721
Native4                                           0.116855  0.111561
AfricanAmerican4                                  0.069077  0.063760
Latino4                                           0.019831  0.039055
Islander4                                         0.104443  0.211271
White4                                           -0.245419  0.002639
Multiracial4                                     -0.099501 -0.082597
LimitedEnglish4                                   0.002595  0.086197
Disadv4                                           0.046138  0.208663
AllMath4Tested                                    0.023147  0.162333
AllMath4                                         -0.063211  0.115274
NativeMath4                                       0.119862  0.103137
AfricanAmericanMath4                              0.009498 -0.064445
LatinoMath4                                       0.001571 -0.021385
IslanderMath4                                     0.084367  0.231808
WhiteMath4                                       -0.251764  0.010609
MultiracialMath4                                 -0.103052 -0.079925
LimitedEnglishMath4                              -0.011005  0.137108
DisadvMath4                                       0.025356  0.178356
Count of Students in HS Admissions                0.001274  0.241836
Count of Testers                                 -0.019732  0.231809
Count of Offers                                  -0.048907  0.107895

                                                 Economic Need Index  \
SED Code                                                   -0.311408
District                                                   -0.283162
Latitude                                                    0.306643
Longitude                                                   0.015074
Zip                                                        -0.165680
Economic Need Index                                         1.000000
Average ELA Proficiency                                    -0.800394
Average Math Proficiency                                   -0.702374
Grade 3 Math 4s - Multiracial                              -0.240659
Year of SHST                                               -0.011539
Grade level                                                -0.332009
Enrollment on 10/31                                        -0.192580
Number of students who registered for the SHSAT             0.023440
Number of students who took the SHSAT                      -0.172916
AllTested                                                  -0.180824
All4                                                       -0.510312
Native4                                                    -0.062567
AfricanAmerican4                                           -0.072707
Latino4                                                    -0.127446
Islander4                                                  -0.323675
White4                                                     -0.549216
Multiracial4                                               -0.281811
LimitedEnglish4                                             0.034956
Disadv4                                                    -0.270175
```

```
AllMath4Tested                                             -0.155151
AllMath4                                                   -0.450023
NativeMath4                                                -0.059500
AfricanAmericanMath4                                        0.004962
LatinoMath4                                                -0.029510
IslanderMath4                                              -0.286434
WhiteMath4                                                 -0.566390
MultiracialMath4                                           -0.266687
LimitedEnglishMath4                                        -0.015378
DisadvMath4                                                -0.209139
Count of Students in HS Admissions                         -0.252197
Count of Testers                                           -0.411867
Count of Offers                                            -0.421596

                                                 Average ELA Proficiency  \
SED Code                                                        0.128419
District                                                        0.111628
Latitude                                                       -0.195998
Longitude                                                      -0.088961
Zip                                                             0.059880
Economic Need Index                                            -0.800394
Average ELA Proficiency                                         1.000000
Average Math Proficiency                                        0.929950
Grade 3 Math 4s - Multiracial                                   0.196213
Year of SHST                                                    0.012700
Grade level                                                     0.243363
Enrollment on 10/31                                             0.094302
Number of students who registered for the SHSAT                 0.111542
Number of students who took the SHSAT                           0.283662
AllTested                                                       0.236735
All4                                                            0.658177
Native4                                                         0.054658
AfricanAmerican4                                                0.212182
Latino4                                                         0.326014
Islander4                                                       0.440704
White4                                                          0.566242
Multiracial4                                                    0.338812
LimitedEnglish4                                                 0.110526
Disadv4                                                         0.462595
AllMath4Tested                                                  0.206157
AllMath4                                                        0.637720
NativeMath4                                                     0.060073
AfricanAmericanMath4                                            0.161203
LatinoMath4                                                     0.240882
IslanderMath4                                                   0.407912
WhiteMath4                                                      0.568440
MultiracialMath4                                                0.327528
LimitedEnglishMath4                                             0.124413
```

```
DisadvMath4                                                     0.427474
Count of Students in HS Admissions                              0.210394
Count of Testers                                                0.465485
Count of Offers                                                 0.520695

                                                 Average Math Proficiency  \
SED Code                                                         0.117325
District                                                         0.108624
Latitude                                                        -0.166521
Longitude                                                       -0.080416
Zip                                                              0.060285
Economic Need Index                                             -0.702374
Average ELA Proficiency                                          0.929950
Average Math Proficiency                                         1.000000
Grade 3 Math 4s - Multiracial                                    0.175799
Year of SHST                                                     0.007885
Grade level                                                      0.265510
Enrollment on 10/31                                             -0.013963
Number of students who registered for the SHSAT                  0.078433
Number of students who took the SHSAT                            0.223685
AllTested                                                        0.251121
All4                                                             0.614635
Native4                                                          0.045762
AfricanAmerican4                                                 0.225106
Latino4                                                          0.328540
Islander4                                                        0.437103
White4                                                           0.483199
Multiracial4                                                     0.286675
LimitedEnglish4                                                  0.171579
Disadv4                                                          0.466593
AllMath4Tested                                                   0.232278
AllMath4                                                         0.681759
NativeMath4                                                      0.065718
AfricanAmericanMath4                                             0.263405
LatinoMath4                                                      0.326734
IslanderMath4                                                    0.442585
WhiteMath4                                                       0.501042
MultiracialMath4                                                 0.284416
LimitedEnglishMath4                                              0.224012
DisadvMath4                                                      0.517739
Count of Students in HS Admissions                               0.194387
Count of Testers                                                 0.442862
Count of Offers                                                  0.468433

                                                 Grade 3 Math 4s - Multiracial  \
SED Code                                                             -0.039204
District                                                             -0.054976
Latitude                                                             -0.026301
```

```
Longitude                                                            -0.095886
Zip                                                                  -0.026360
Economic Need Index                                                  -0.240659
Average ELA Proficiency                                               0.196213
Average Math Proficiency                                              0.175799
Grade 3 Math 4s - Multiracial                                         1.000000
Year of SHST                                                               NaN
Grade level                                                                NaN
Enrollment on 10/31                                                        NaN
Number of students who registered for the SHSAT                            NaN
Number of students who took the SHSAT                                      NaN
AllTested                                                             0.022525
All4                                                                  0.129934
Native4                                                              -0.012703
AfricanAmerican4                                                      0.001654
Latino4                                                               0.018822
Islander4                                                             0.029978
White4                                                                0.202957
Multiracial4                                                          0.300033
LimitedEnglish4                                                      -0.026729
Disadv4                                                               0.008969
AllMath4Tested                                                        0.023866
AllMath4                                                              0.116946
NativeMath4                                                          -0.013752
AfricanAmericanMath4                                                 -0.011045
LatinoMath4                                                           0.011719
IslanderMath4                                                         0.016346
WhiteMath4                                                            0.226853
MultiracialMath4                                                      0.287386
LimitedEnglishMath4                                                  -0.019547
DisadvMath4                                                          -0.006277
Count of Students in HS Admissions                                   -0.023242
Count of Testers                                                      0.022784
Count of Offers                                                       0.030693

                                                 Year of SHST  \
SED Code                                             0.033258
District                                                  NaN
Latitude                                            -0.052813
Longitude                                            0.020974
Zip                                                 -0.027054
Economic Need Index                                 -0.011539
Average ELA Proficiency                              0.012700
Average Math Proficiency                             0.007885
Grade 3 Math 4s - Multiracial                             NaN
Year of SHST                                         1.000000
Grade level                                          0.050325
Enrollment on 10/31                                 -0.061249
```

```
Number of students who registered for the SHSAT     -0.153356
Number of students who took the SHSAT               -0.045207
AllTested                                           -0.006235
All4                                                 0.009377
Native4                                                   NaN
AfricanAmerican4                                     0.083052
Latino4                                             -0.025164
Islander4                                           -0.025141
White4                                              -0.025141
Multiracial4                                        -0.025141
LimitedEnglish4                                     -0.021568
Disadv4                                              0.041013
AllMath4Tested                                      -0.009201
AllMath4                                             0.019439
NativeMath4                                               NaN
AfricanAmericanMath4                                 0.055281
LatinoMath4                                         -0.019891
IslanderMath4                                       -0.025141
WhiteMath4                                          -0.025141
MultiracialMath4                                    -0.010211
LimitedEnglishMath4                                 -0.013455
DisadvMath4                                          0.032070
Count of Students in HS Admissions                  -0.015943
Count of Testers                                     0.017560
Count of Offers                                     -0.025141

                                                      ...         \
SED Code                                              ...
District                                              ...
Latitude                                              ...
Longitude                                             ...
Zip                                                   ...
Economic Need Index                                   ...
Average ELA Proficiency                               ...
Average Math Proficiency                              ...
Grade 3 Math 4s - Multiracial                         ...
Year of SHST                                          ...
Grade level                                           ...
Enrollment on 10/31                                   ...
Number of students who registered for the SHSAT       ...
Number of students who took the SHSAT                 ...
AllTested                                             ...
All4                                                  ...
Native4                                               ...
AfricanAmerican4                                      ...
Latino4                                               ...
Islander4                                             ...
White4                                                ...
```

```
Multiracial4                                          ...
LimitedEnglish4                                       ...
Disadv4                                               ...
AllMath4Tested                                        ...
AllMath4                                              ...
NativeMath4                                           ...
AfricanAmericanMath4                                  ...
LatinoMath4                                           ...
IslanderMath4                                         ...
WhiteMath4                                            ...
MultiracialMath4                                      ...
LimitedEnglishMath4                                   ...
DisadvMath4                                           ...
Count of Students in HS Admissions                    ...
Count of Testers                                      ...
Count of Offers                                       ...

                                                 AfricanAmericanMath4  \
SED Code                                                    -0.121024
District                                                    -0.104757
Latitude                                                     0.030972
Longitude                                                    0.009498
Zip                                                         -0.064445
Economic Need Index                                          0.004962
Average ELA Proficiency                                      0.161203
Average Math Proficiency                                     0.263405
Grade 3 Math 4s - Multiracial                               -0.011045
Year of SHST                                                 0.055281
Grade level                                                  0.157422
Enrollment on 10/31                                         -0.237079
Number of students who registered for the SHSAT             -0.152962
Number of students who took the SHSAT                       -0.036843
AllTested                                                    0.133670
All4                                                         0.123543
Native4                                                      0.019590
AfricanAmerican4                                             0.839716
Latino4                                                      0.058003
Islander4                                                   -0.062393
White4                                                      -0.044113
Multiracial4                                                 0.020160
LimitedEnglish4                                             -0.037597
Disadv4                                                      0.191805
AllMath4Tested                                               0.131229
AllMath4                                                     0.269587
NativeMath4                                                  0.076112
AfricanAmericanMath4                                         1.000000
LatinoMath4                                                  0.161130
IslanderMath4                                               -0.077146
```

```
WhiteMath4                                                -0.048821
MultiracialMath4                                           0.094581
LimitedEnglishMath4                                       -0.042476
DisadvMath4                                                0.341562
Count of Students in HS Admissions                        -0.054994
Count of Testers                                           0.018657
Count of Offers                                           -0.044907

                                                 LatinoMath4  IslanderMath4  \
SED Code                                            0.005167       0.222771
District                                            0.016430       0.223257
Latitude                                            0.130747      -0.153146
Longitude                                           0.001571       0.084367
Zip                                                -0.021385       0.231808
Economic Need Index                                -0.029510      -0.286434
Average ELA Proficiency                             0.240882       0.407912
Average Math Proficiency                            0.326734       0.442585
Grade 3 Math 4s - Multiracial                       0.011719       0.016346
Year of SHST                                       -0.019891      -0.025141
Grade level                                         0.033879       0.113581
Enrollment on 10/31                                -0.055517       0.033151
Number of students who registered for the SHSAT     0.288104       0.101141
Number of students who took the SHSAT               0.080072       0.291501
AllTested                                           0.484628       0.571474
All4                                                0.356964       0.726540
Native4                                             0.004995       0.080127
AfricanAmerican4                                    0.098546      -0.054732
Latino4                                             0.833767       0.265917
Islander4                                           0.176991       0.940405
White4                                              0.136859       0.378943
Multiracial4                                        0.045885       0.146847
LimitedEnglish4                                     0.293614       0.225757
Disadv4                                             0.417239       0.761421
AllMath4Tested                                      0.493930       0.575876
AllMath4                                            0.463463       0.793577
NativeMath4                                         0.025953       0.080432
AfricanAmericanMath4                                0.161130      -0.077146
LatinoMath4                                         1.000000       0.166465
IslanderMath4                                       0.166465       1.000000
WhiteMath4                                          0.132373       0.382159
MultiracialMath4                                    0.066978       0.145016
LimitedEnglishMath4                                 0.180779       0.610268
DisadvMath4                                         0.494340       0.788233
Count of Students in HS Admissions                  0.352206       0.640299
Count of Testers                                    0.331389       0.790181
Count of Offers                                     0.157090       0.781842

                                                 WhiteMath4  MultiracialMath4  \
SED Code                                           0.111477         -0.100134
District                                           0.094061         -0.097328
Latitude                                          -0.248455         -0.023081
Longitude                                         -0.251764         -0.103052
Zip                                                0.010609         -0.079925
Economic Need Index                               -0.566390         -0.266687
Average ELA Proficiency                            0.568440          0.327528
Average Math Proficiency                           0.501042          0.284416
Grade 3 Math 4s - Multiracial                      0.226853          0.287386
Year of SHST                                      -0.025141         -0.010211
Grade level                                        0.113581          0.135376
Enrollment on 10/31                                0.033151         -0.003628
Number of students who registered for the SHSAT    0.101141          0.001986
Number of students who took the SHSAT              0.291501          0.167718
AllTested                                          0.402895          0.133575
All4                                               0.779631          0.404398
Native4                                           -0.011612         -0.010621
AfricanAmerican4                                  -0.004120          0.110634
Latino4                                            0.249102          0.121022
Islander4                                          0.443344          0.189189
White4                                             0.973047          0.488634
Multiracial4                                       0.447110          0.903375
LimitedEnglish4                                    0.009921         -0.034339
Disadv4                                            0.473337          0.135630
AllMath4Tested                                     0.386273          0.113615
AllMath4                                           0.700406          0.348437
NativeMath4                                       -0.010641         -0.012548
AfricanAmericanMath4                              -0.048821          0.094581
LatinoMath4                                        0.132373          0.066978
IslanderMath4                                      0.382159          0.145016
WhiteMath4                                         1.000000          0.466884
MultiracialMath4                                   0.466884          1.000000
LimitedEnglishMath4                                0.065524         -0.015748
DisadvMath4                                        0.388335          0.109534
Count of Students in HS Admissions                 0.443029          0.130456
Count of Testers                                   0.643815          0.302159
Count of Offers                                    0.727092          0.532177

                                                 LimitedEnglishMath4  \
SED Code                                                    0.101698
District                                                    0.097299
Latitude                                                   -0.099548
Longitude                                                  -0.011005
Zip                                                         0.137108
Economic Need Index                                        -0.015378
Average ELA Proficiency                                     0.124413
Average Math Proficiency                                    0.224012
Grade 3 Math 4s - Multiracial                              -0.019547
```

```
Year of SHST                                               -0.013455
Grade level                                                -0.016051
Enrollment on 10/31                                        -0.027152
Number of students who registered for the SHSAT             0.221893
Number of students who took the SHSAT                       0.013616
AllTested                                                   0.358098
All4                                                        0.254364
Native4                                                     0.002829
AfricanAmerican4                                           -0.059808
Latino4                                                     0.166169
Islander4                                                   0.399347
White4                                                      0.053820
Multiracial4                                               -0.019669
LimitedEnglish4                                             0.379997
Disadv4                                                     0.328336
AllMath4Tested                                              0.383147
AllMath4                                                    0.427443
NativeMath4                                                 0.007294
AfricanAmericanMath4                                       -0.042476
LatinoMath4                                                 0.180779
IslanderMath4                                               0.610268
WhiteMath4                                                  0.065524
MultiracialMath4                                           -0.015748
LimitedEnglishMath4                                         1.000000
DisadvMath4                                                 0.491741
Count of Students in HS Admissions                          0.458174
Count of Testers                                            0.441553
Count of Offers                                             0.286620

                                                 DisadvMath4  \
SED Code                                           0.172113
District                                           0.189524
Latitude                                          -0.126465
Longitude                                          0.025356
Zip                                                0.178356
Economic Need Index                               -0.209139
Average ELA Proficiency                            0.427474
Average Math Proficiency                           0.517739
Grade 3 Math 4s - Multiracial                     -0.006277
Year of SHST                                       0.032070
Grade level                                        0.111544
Enrollment on 10/31                               -0.165340
Number of students who registered for the SHSAT    0.102049
Number of students who took the SHSAT              0.071933
AllTested                                          0.697246
All4                                               0.746844
Native4                                            0.086700
AfricanAmerican4                                   0.294488
```

```
Latino4                                             0.506948
Islander4                                           0.746271
White4                                              0.377947
Multiracial4                                        0.067370
LimitedEnglish4                                     0.233133
Disadv4                                             0.906080
AllMath4Tested                                      0.707533
AllMath4                                            0.879551
NativeMath4                                         0.110006
AfricanAmericanMath4                                0.341562
LatinoMath4                                         0.494340
IslanderMath4                                       0.788233
WhiteMath4                                          0.388335
MultiracialMath4                                    0.109534
LimitedEnglishMath4                                 0.491741
DisadvMath4                                         1.000000
Count of Students in HS Admissions                  0.611494
Count of Testers                                    0.732673
Count of Offers                                     0.599504

                                                 Count of Students in HS Admissions
SED Code                                                                   0.355074
District                                                                   0.331419
Latitude                                                                  -0.202395
Longitude                                                                  0.001274
Zip                                                                        0.241836
Economic Need Index                                                       -0.252197
Average ELA Proficiency                                                    0.210394
Average Math Proficiency                                                   0.194387
Grade 3 Math 4s - Multiracial                                             -0.023242
Year of SHST                                                              -0.015943
Grade level                                                                0.161501
Enrollment on 10/31                                                        0.385689
Number of students who registered for the SHSAT                            0.058574
Number of students who took the SHSAT                                      0.144917
AllTested                                                                  0.909962
All4                                                                       0.651056
Native4                                                                    0.204923
AfricanAmerican4                                                           0.055765
Latino4                                                                    0.578781
Islander4                                                                  0.616307
White4                                                                     0.449475
Multiracial4                                                               0.103813
LimitedEnglish4                                                            0.230929
Disadv4                                                                    0.692586
AllMath4Tested                                                             0.897209
AllMath4                                                                   0.612487
NativeMath4                                                                0.198148
```

```
AfricanAmericanMath4                                                      -0.054994
LatinoMath4                                                                0.352206
IslanderMath4                                                              0.640299
WhiteMath4                                                                 0.443029
MultiracialMath4                                                           0.130456
LimitedEnglishMath4                                                        0.458174
DisadvMath4                                                                0.611494
Count of Students in HS Admissions                                         1.000000
Count of Testers                                                           0.856769
Count of Offers                                                            0.521510

                                                 Count of Testers  \
SED Code                                                 0.281210
District                                                 0.260098
Latitude                                                -0.220273
Longitude                                               -0.019732
Zip                                                      0.231809
Economic Need Index                                     -0.411867
Average ELA Proficiency                                  0.465485
Average Math Proficiency                                 0.442862
Grade 3 Math 4s - Multiracial                            0.022784
Year of SHST                                             0.017560
Grade level                                              0.159930
Enrollment on 10/31                                      0.111636
Number of students who registered for the SHSAT          0.109380
Number of students who took the SHSAT                    0.333936
AllTested                                                0.804685
All4                                                     0.852334
Native4                                                  0.107553
AfricanAmerican4                                         0.134114
Latino4                                                  0.541865
Islander4                                                0.780939
White4                                                   0.645959
Multiracial4                                             0.281024
LimitedEnglish4                                          0.194009
Disadv4                                                  0.807129
AllMath4Tested                                           0.782446
AllMath4                                                 0.813443
NativeMath4                                              0.122513
AfricanAmericanMath4                                     0.018657
LatinoMath4                                              0.331389
IslanderMath4                                            0.790181
WhiteMath4                                               0.643815
MultiracialMath4                                         0.302159
LimitedEnglishMath4                                      0.441553
DisadvMath4                                              0.732673
Count of Students in HS Admissions                       0.856769
Count of Testers                                         1.000000
Count of Offers                                          0.778770

                                                 Count of Offers
SED Code                                                0.119194
District                                                0.109803
Latitude                                               -0.145873
Longitude                                              -0.048907
Zip                                                     0.107895
Economic Need Index                                    -0.421596
Average ELA Proficiency                                 0.520695
Average Math Proficiency                                0.468433
Grade 3 Math 4s - Multiracial                           0.030693
Year of SHST                                           -0.025141
Grade level                                             0.113581
Enrollment on 10/31                                     0.033151
Number of students who registered for the SHSAT         0.101141
Number of students who took the SHSAT                   0.291501
AllTested                                               0.485583
All4                                                    0.859680
Native4                                                 0.038949
AfricanAmerican4                                        0.000243
Latino4                                                 0.285010
Islander4                                               0.829191
White4                                                  0.738594
Multiracial4                                            0.542744
LimitedEnglish4                                         0.103076
Disadv4                                                 0.678692
AllMath4Tested                                          0.456146
AllMath4                                                0.792829
NativeMath4                                             0.061064
AfricanAmericanMath4                                   -0.044907
LatinoMath4                                             0.157090
IslanderMath4                                           0.781842
WhiteMath4                                              0.727092
MultiracialMath4                                        0.532177
LimitedEnglishMath4                                     0.286620
DisadvMath4                                             0.599504
Count of Students in HS Admissions                      0.521510
Count of Testers                                        0.778770
Count of Offers                                         1.000000

[37 rows x 37 columns]
```

In this initial exploration, we can clearly observe that : Economic Need Index inversely correlated to Average ELA Proficiency and Economic Need Index inversely correlated to Average Math Proficiency

```
In [595]: df_sub=dfy
          df_sub=df_sub.drop(['District',
           'Latitude',
           'Longitude',
           'Address (Full)',
           'City',
           'Zip',
           'Grades',
           'Grade Low',
           'Grade High',
           'Rigorous Instruction %',
           'Collaborative Teachers %',
           'Supportive Environment %',
           'Effective School Leadership %',
           'Strong Family-Community Ties %',
           'Trust %',
           'Grade 3 Math 4s - Multiracial',
           'DBN',
           'School name',
           'Year of SHST',
           'Grade level',
           'Enrollment on 10/31',
           'Feeder School DBN',
           'Feeder School Name'],axis=1)
```

```
In [596]: df_sub.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1364 entries, 0 to 1363
Data columns (total 48 columns):
School Name                                         1364 non-null object
SED Code                                            1364 non-null int64
Location Code                                       1364 non-null object
Community School?                                   1364 non-null object
Economic Need Index                                 1339 non-null float64
School Income Estimate                              906 non-null object
Percent ELL                                         1364 non-null object
Percent Asian                                       1364 non-null object
Percent Black                                       1364 non-null object
Percent Hispanic                                    1364 non-null object
Percent Black / Hispanic                            1364 non-null object
Percent White                                       1364 non-null object
Student Attendance Rate                             1339 non-null object
Percent of Students Chronically Absent              1339 non-null object
Rigorous Instruction Rating                         1288 non-null object
Collaborative Teachers Rating                       1288 non-null object
Supportive Environment Rating                       1284 non-null object
Effective School Leadership Rating                  1291 non-null object
Strong Family-Community Ties Rating                 1291 non-null object
Trust Rating                                        1291 non-null object
```

```
Student Achievement Rating                          1278 non-null object
Average ELA Proficiency                             1309 non-null float64
Average Math Proficiency                            1309 non-null float64
Number of students who registered for the SHSAT    113 non-null float64
Number of students who took the SHSAT               113 non-null float64
AllTested                                           1364 non-null float64
All4                                                1364 non-null float64
Native4                                             1364 non-null float64
AfricanAmerican4                                    1364 non-null float64
Latino4                                             1364 non-null float64
Islander4                                           1364 non-null float64
White4                                              1364 non-null float64
Multiracial4                                        1364 non-null float64
LimitedEnglish4                                     1364 non-null float64
Disadv4                                             1364 non-null float64
AllMath4Tested                                      1364 non-null float64
AllMath4                                            1364 non-null float64
NativeMath4                                         1364 non-null float64
AfricanAmericanMath4                                1364 non-null float64
LatinoMath4                                         1364 non-null float64
IslanderMath4                                       1364 non-null float64
WhiteMath4                                          1364 non-null float64
MultiracialMath4                                    1364 non-null float64
LimitedEnglishMath4                                 1364 non-null float64
DisadvMath4                                         1364 non-null float64
Count of Students in HS Admissions                  683 non-null float64
Count of Testers                                    683 non-null float64
Count of Offers                                     683 non-null float64
dtypes: float64(28), int64(1), object(19)
memory usage: 522.2+ KB
```

```
In [597]: df_sub.corr()
```

```
Out[597]:                                                  SED Code  \
          SED Code                                         1.000000
          Economic Need Index                             -0.311408
          Average ELA Proficiency                          0.128419
          Average Math Proficiency                         0.117325
          Number of students who registered for the SHSAT  0.011083
          Number of students who took the SHSAT            0.007385
          AllTested                                        0.238549
          All4                                             0.181169
          Native4                                          0.109187
          AfricanAmerican4                                -0.008202
          Latino4                                          0.084169
          Islander4                                        0.217653
          White4                                           0.109849
```

```
Multiracial4                                    -0.101175
LimitedEnglish4                                  0.063333
Disadv4                                          0.222882
AllMath4Tested                                   0.237845
AllMath4                                         0.144911
NativeMath4                                      0.103002
AfricanAmericanMath4                            -0.121024
LatinoMath4                                      0.005167
IslanderMath4                                    0.222771
WhiteMath4                                       0.111477
MultiracialMath4                                -0.100134
LimitedEnglishMath4                              0.101698
DisadvMath4                                      0.172113
Count of Students in HS Admissions               0.355074
Count of Testers                                 0.281210
Count of Offers                                  0.119194

                                                 Economic Need Index  \
SED Code                                                   -0.311408
Economic Need Index                                         1.000000
Average ELA Proficiency                                    -0.800394
Average Math Proficiency                                   -0.702374
Number of students who registered for the SHSAT             0.023440
Number of students who took the SHSAT                      -0.172916
AllTested                                                  -0.180824
All4                                                       -0.510312
Native4                                                    -0.062567
AfricanAmerican4                                           -0.072707
Latino4                                                    -0.127446
Islander4                                                  -0.323675
White4                                                     -0.549216
Multiracial4                                               -0.281811
LimitedEnglish4                                             0.034956
Disadv4                                                    -0.270175
AllMath4Tested                                             -0.155151
AllMath4                                                   -0.450023
NativeMath4                                                -0.059500
AfricanAmericanMath4                                        0.004962
LatinoMath4                                                -0.029510
IslanderMath4                                              -0.286434
WhiteMath4                                                 -0.566390
MultiracialMath4                                           -0.266687
LimitedEnglishMath4                                        -0.015378
DisadvMath4                                                -0.209139
Count of Students in HS Admissions                         -0.252197
Count of Testers                                           -0.411867
Count of Offers                                            -0.421596
```

```
                                                 Average ELA Proficiency  \
SED Code                                                        0.128419
Economic Need Index                                            -0.800394
Average ELA Proficiency                                         1.000000
Average Math Proficiency                                        0.929950
Number of students who registered for the SHSAT                 0.111542
Number of students who took the SHSAT                           0.283662
AllTested                                                       0.236735
All4                                                            0.658177
Native4                                                         0.054658
AfricanAmerican4                                                0.212182
Latino4                                                         0.326014
Islander4                                                       0.440704
White4                                                          0.566242
Multiracial4                                                    0.338812
LimitedEnglish4                                                 0.110526
Disadv4                                                         0.462595
AllMath4Tested                                                  0.206157
AllMath4                                                        0.637720
NativeMath4                                                     0.060073
AfricanAmericanMath4                                            0.161203
LatinoMath4                                                     0.240882
IslanderMath4                                                   0.407912
WhiteMath4                                                      0.568440
MultiracialMath4                                                0.327528
LimitedEnglishMath4                                             0.124413
DisadvMath4                                                     0.427474
Count of Students in HS Admissions                              0.210394
Count of Testers                                                0.465485
Count of Offers                                                 0.520695

                                                 Average Math Proficiency  \
SED Code                                                         0.117325
Economic Need Index                                             -0.702374
Average ELA Proficiency                                          0.929950
Average Math Proficiency                                         1.000000
Number of students who registered for the SHSAT                  0.078433
Number of students who took the SHSAT                            0.223685
AllTested                                                        0.251121
All4                                                             0.614635
Native4                                                          0.045762
AfricanAmerican4                                                 0.225106
Latino4                                                          0.328540
Islander4                                                        0.437103
White4                                                           0.483199
Multiracial4                                                     0.286675
LimitedEnglish4                                                  0.171579
Disadv4                                                          0.466593
```

```
AllMath4Tested                                                   0.232278
AllMath4                                                         0.681759
NativeMath4                                                      0.065718
AfricanAmericanMath4                                             0.263405
LatinoMath4                                                      0.326734
IslanderMath4                                                    0.442585
WhiteMath4                                                       0.501042
MultiracialMath4                                                 0.284416
LimitedEnglishMath4                                              0.224012
DisadvMath4                                                      0.517739
Count of Students in HS Admissions                               0.194387
Count of Testers                                                 0.442862
Count of Offers                                                  0.468433

                                                 Number of students who registered fo
SED Code
Economic Need Index
Average ELA Proficiency
Average Math Proficiency
Number of students who registered for the SHSAT
Number of students who took the SHSAT
AllTested
All4
Native4
AfricanAmerican4
Latino4
Islander4
White4
Multiracial4
LimitedEnglish4
Disadv4
AllMath4Tested
AllMath4
NativeMath4
AfricanAmericanMath4
LatinoMath4
IslanderMath4
WhiteMath4
MultiracialMath4
LimitedEnglishMath4
DisadvMath4
Count of Students in HS Admissions
Count of Testers
Count of Offers

                                                 Number of students who took the SHS
SED Code                                                                     0.00738
Economic Need Index                                                         -0.1729
```

```
Average ELA Proficiency                                                       0.28366
Average Math Proficiency                                                      0.22368
Number of students who registered for the SHSAT                               0.71355
Number of students who took the SHSAT                                         1.00000
AllTested                                                                    -0.06036
All4                                                                          0.31662
Native4                                                                            Na
AfricanAmerican4                                                              0.09898
Latino4                                                                       0.25437
Islander4                                                                     0.29150
White4                                                                        0.29150
Multiracial4                                                                  0.29150
LimitedEnglish4                                                              -0.01393
Disadv4                                                                       0.30706
AllMath4Tested                                                               -0.10051
AllMath4                                                                      0.10974
NativeMath4                                                                        Na
AfricanAmericanMath4                                                         -0.03684
LatinoMath4                                                                   0.08007
IslanderMath4                                                                 0.29150
WhiteMath4                                                                    0.29150
MultiracialMath4                                                              0.16771
LimitedEnglishMath4                                                           0.01361
DisadvMath4                                                                   0.07193
Count of Students in HS Admissions                                            0.14491
Count of Testers                                                              0.33393
Count of Offers                                                               0.29150

                                                 AllTested      All4  \
SED Code                                          0.238549  0.181169
Economic Need Index                              -0.180824 -0.510312
Average ELA Proficiency                           0.236735  0.658177
Average Math Proficiency                          0.251121  0.614635
Number of students who registered for the SHSAT  -0.021819  0.200793
Number of students who took the SHSAT            -0.060369  0.316628
AllTested                                         1.000000  0.686314
All4                                              0.686314  1.000000
Native4                                           0.143783  0.089638
AfricanAmerican4                                  0.203450  0.217079
Latino4                                           0.643399  0.529362
Islander4                                         0.581499  0.813208
White4                                            0.421975  0.810412
Multiracial4                                      0.102096  0.389538
LimitedEnglish4                                   0.176506  0.128738
Disadv4                                           0.744634  0.856384
AllMath4Tested                                    0.993251  0.657475
AllMath4                                          0.681137  0.923672
NativeMath4                                       0.116180  0.085426
```

```
AfricanAmericanMath4                              0.133670  0.123543
LatinoMath4                                       0.484628  0.356964
IslanderMath4                                     0.571474  0.726540
WhiteMath4                                        0.402895  0.779631
MultiracialMath4                                  0.133575  0.404398
LimitedEnglishMath4                               0.358098  0.254364
DisadvMath4                                       0.697246  0.746844
Count of Students in HS Admissions                0.909962  0.651056
Count of Testers                                  0.804685  0.852334
Count of Offers                                   0.485583  0.859680

                                                  Native4  AfricanAmerican4  \
SED Code                                         0.109187         -0.008202
Economic Need Index                             -0.062567         -0.072707
Average ELA Proficiency                          0.054658          0.212182
Average Math Proficiency                         0.045762          0.225106
Number of students who registered for the SHSAT       NaN         -0.078216
Number of students who took the SHSAT                 NaN          0.098985
AllTested                                        0.143783          0.203450
All4                                             0.089638          0.217079
Native4                                          1.000000          0.108634
AfricanAmerican4                                 0.108634          1.000000
Latino4                                          0.035452          0.095225
Islander4                                        0.113690         -0.019605
White4                                          -0.013632          0.012959
Multiracial4                                    -0.000053          0.050383
LimitedEnglish4                                 -0.025916         -0.054510
Disadv4                                          0.146602          0.284380
AllMath4Tested                                   0.131057          0.184188
AllMath4                                         0.058387          0.240237
NativeMath4                                      0.895855          0.110782
AfricanAmericanMath4                             0.019590          0.839716
LatinoMath4                                      0.004995          0.098546
IslanderMath4                                    0.080127         -0.054732
WhiteMath4                                      -0.011612         -0.004120
MultiracialMath4                                -0.010621          0.110634
LimitedEnglishMath4                              0.002829         -0.059808
DisadvMath4                                      0.086700          0.294488
Count of Students in HS Admissions               0.204923          0.055765
Count of Testers                                 0.107553          0.134114
Count of Offers                                  0.038949          0.000243

                                                      ...         \
SED Code                                              ...
Economic Need Index                                   ...
Average ELA Proficiency                               ...
Average Math Proficiency                              ...
Number of students who registered for the SHSAT       ...
Number of students who took the SHSAT                 ...
AllTested                                             ...
All4                                                  ...
Native4                                               ...
AfricanAmerican4                                      ...
Latino4                                               ...
Islander4                                             ...
White4                                                ...
Multiracial4                                          ...
LimitedEnglish4                                       ...
Disadv4                                               ...
AllMath4Tested                                        ...
AllMath4                                              ...
NativeMath4                                           ...
AfricanAmericanMath4                                  ...
LatinoMath4                                           ...
IslanderMath4                                         ...
WhiteMath4                                            ...
MultiracialMath4                                      ...
LimitedEnglishMath4                                   ...
DisadvMath4                                           ...
Count of Students in HS Admissions                    ...
Count of Testers                                      ...
Count of Offers                                       ...

                                                 AfricanAmericanMath4  \
SED Code                                                    -0.121024
Economic Need Index                                          0.004962
Average ELA Proficiency                                      0.161203
Average Math Proficiency                                     0.263405
Number of students who registered for the SHSAT             -0.152962
Number of students who took the SHSAT                       -0.036843
AllTested                                                    0.133670
All4                                                         0.123543
Native4                                                      0.019590
AfricanAmerican4                                             0.839716
Latino4                                                      0.058003
Islander4                                                   -0.062393
White4                                                      -0.044113
Multiracial4                                                 0.020160
LimitedEnglish4                                             -0.037597
Disadv4                                                      0.191805
AllMath4Tested                                               0.131229
AllMath4                                                     0.269587
NativeMath4                                                  0.076112
AfricanAmericanMath4                                         1.000000
LatinoMath4                                                  0.161130
IslanderMath4                                               -0.077146
```

```
WhiteMath4                                                  -0.048821
MultiracialMath4                                             0.094581
LimitedEnglishMath4                                         -0.042476
DisadvMath4                                                  0.341562
Count of Students in HS Admissions                          -0.054994
Count of Testers                                             0.018657
Count of Offers                                             -0.044907

                                                 LatinoMath4  IslanderMath4  \
SED Code                                            0.005167       0.222771
Economic Need Index                                -0.029510      -0.286434
Average ELA Proficiency                             0.240882       0.407912
Average Math Proficiency                            0.326734       0.442585
Number of students who registered for the SHSAT     0.288104       0.101141
Number of students who took the SHSAT               0.080072       0.291501
AllTested                                           0.484628       0.571474
All4                                                0.356964       0.726540
Native4                                             0.004995       0.080127
AfricanAmerican4                                    0.098546      -0.054732
Latino4                                             0.833767       0.265917
Islander4                                           0.176991       0.940405
White4                                              0.136859       0.378943
Multiracial4                                        0.045885       0.146847
LimitedEnglish4                                     0.293614       0.225757
Disadv4                                             0.417239       0.761421
AllMath4Tested                                      0.493930       0.575876
AllMath4                                            0.463463       0.793577
NativeMath4                                         0.025953       0.080432
AfricanAmericanMath4                                0.161130      -0.077146
LatinoMath4                                         1.000000       0.166465
IslanderMath4                                       0.166465       1.000000
WhiteMath4                                          0.132373       0.382159
MultiracialMath4                                    0.066978       0.145016
LimitedEnglishMath4                                 0.180779       0.610268
DisadvMath4                                         0.494340       0.788233
Count of Students in HS Admissions                  0.352206       0.640299
Count of Testers                                    0.331389       0.790181
Count of Offers                                     0.157090       0.781842

                                                 WhiteMath4  MultiracialMath4  \
SED Code                                           0.111477         -0.100134
Economic Need Index                               -0.566390         -0.266687
Average ELA Proficiency                            0.568440          0.327528
Average Math Proficiency                           0.501042          0.284416
Number of students who registered for the SHSAT    0.101141          0.001986
Number of students who took the SHSAT              0.291501          0.167718
AllTested                                          0.402895          0.133575
All4                                               0.779631          0.404398
```

```
Native4                                           -0.011612         -0.010621
AfricanAmerican4                                  -0.004120          0.110634
Latino4                                            0.249102          0.121022
Islander4                                          0.443344          0.189189
White4                                             0.973047          0.488634
Multiracial4                                       0.447110          0.903375
LimitedEnglish4                                    0.009921         -0.034339
Disadv4                                            0.473337          0.135630
AllMath4Tested                                     0.386273          0.113615
AllMath4                                           0.700406          0.348437
NativeMath4                                       -0.010641         -0.012548
AfricanAmericanMath4                              -0.048821          0.094581
LatinoMath4                                        0.132373          0.066978
IslanderMath4                                      0.382159          0.145016
WhiteMath4                                         1.000000          0.466884
MultiracialMath4                                   0.466884          1.000000
LimitedEnglishMath4                                0.065524         -0.015748
DisadvMath4                                        0.388335          0.109534
Count of Students in HS Admissions                 0.443029          0.130456
Count of Testers                                   0.643815          0.302159
Count of Offers                                    0.727092          0.532177

                                                 LimitedEnglishMath4  \
SED Code                                                    0.101698
Economic Need Index                                        -0.015378
Average ELA Proficiency                                     0.124413
Average Math Proficiency                                    0.224012
Number of students who registered for the SHSAT             0.221893
Number of students who took the SHSAT                       0.013616
AllTested                                                   0.358098
All4                                                        0.254364
Native4                                                     0.002829
AfricanAmerican4                                           -0.059808
Latino4                                                     0.166169
Islander4                                                   0.399347
White4                                                      0.053820
Multiracial4                                               -0.019669
LimitedEnglish4                                             0.379997
Disadv4                                                     0.328336
AllMath4Tested                                              0.383147
AllMath4                                                    0.427443
NativeMath4                                                 0.007294
AfricanAmericanMath4                                       -0.042476
LatinoMath4                                                 0.180779
IslanderMath4                                               0.610268
WhiteMath4                                                  0.065524
MultiracialMath4                                           -0.015748
LimitedEnglishMath4                                         1.000000
```

```
DisadvMath4                                                 0.491741
Count of Students in HS Admissions                          0.458174
Count of Testers                                            0.441553
Count of Offers                                             0.286620

                                                 DisadvMath4  \
SED Code                                            0.172113
Economic Need Index                                -0.209139
Average ELA Proficiency                             0.427474
Average Math Proficiency                            0.517739
Number of students who registered for the SHSAT     0.102049
Number of students who took the SHSAT               0.071933
AllTested                                           0.697246
All4                                                0.746844
Native4                                             0.086700
AfricanAmerican4                                    0.294488
Latino4                                             0.506948
Islander4                                           0.746271
White4                                              0.377947
Multiracial4                                        0.067370
LimitedEnglish4                                     0.233133
Disadv4                                             0.906080
AllMath4Tested                                      0.707533
AllMath4                                            0.879551
NativeMath4                                         0.110006
AfricanAmericanMath4                                0.341562
LatinoMath4                                         0.494340
IslanderMath4                                       0.788233
WhiteMath4                                          0.388335
MultiracialMath4                                    0.109534
LimitedEnglishMath4                                 0.491741
DisadvMath4                                         1.000000
Count of Students in HS Admissions                  0.611494
Count of Testers                                    0.732673
Count of Offers                                     0.599504

                                                 Count of Students in HS Admissions
SED Code                                                                   0.355074
Economic Need Index                                                       -0.252197
Average ELA Proficiency                                                    0.210394
Average Math Proficiency                                                   0.194387
Number of students who registered for the SHSAT                            0.058574
Number of students who took the SHSAT                                      0.144917
AllTested                                                                  0.909962
All4                                                                       0.651056
Native4                                                                    0.204923
AfricanAmerican4                                                           0.055765
Latino4                                                                    0.578781
```

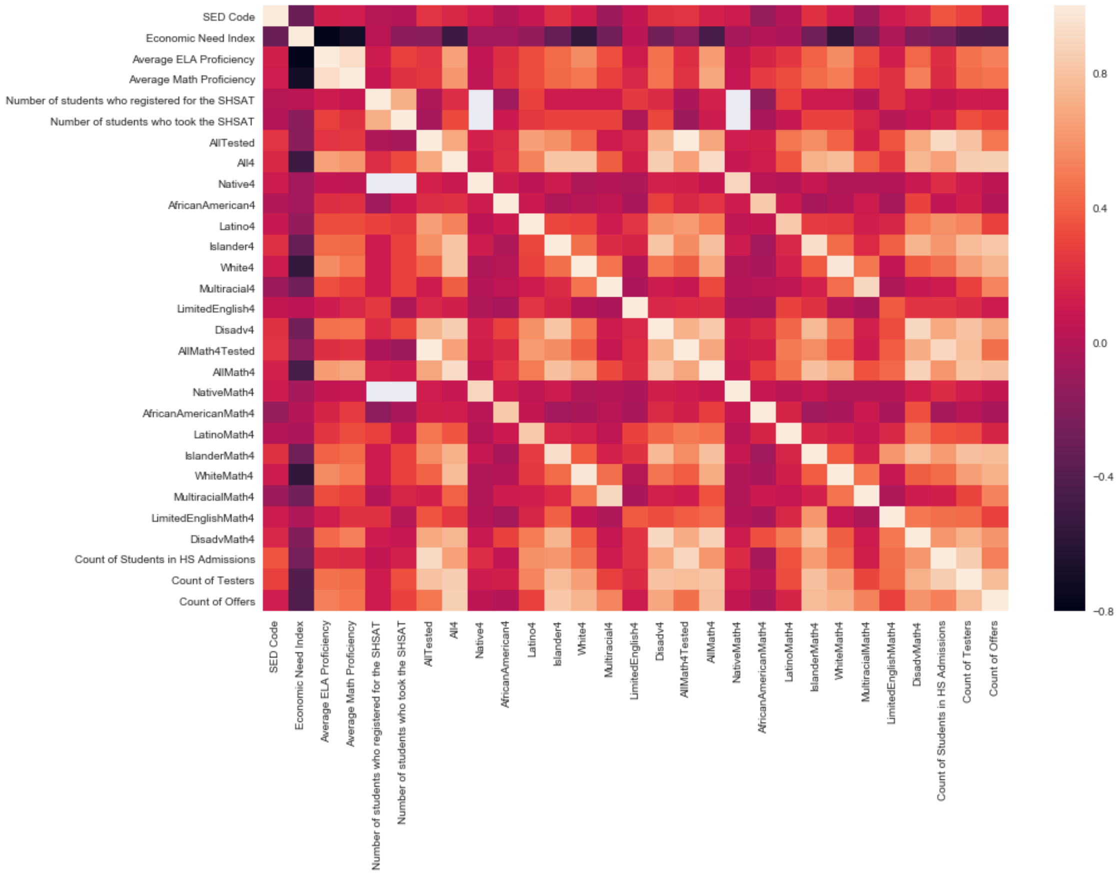
```
Islander4                                                                  0.616307
White4                                                                     0.449475
Multiracial4                                                               0.103813
LimitedEnglish4                                                            0.230929
Disadv4                                                                    0.692586
AllMath4Tested                                                             0.897209
AllMath4                                                                   0.612487
NativeMath4                                                                0.198148
AfricanAmericanMath4                                                      -0.054994
LatinoMath4                                                                0.352206
IslanderMath4                                                              0.640299
WhiteMath4                                                                 0.443029
MultiracialMath4                                                           0.130456
LimitedEnglishMath4                                                        0.458174
DisadvMath4                                                                0.611494
Count of Students in HS Admissions                                         1.000000
Count of Testers                                                           0.856769
Count of Offers                                                            0.521510

                                                 Count of Testers  \
SED Code                                                 0.281210
Economic Need Index                                     -0.411867
Average ELA Proficiency                                  0.465485
Average Math Proficiency                                 0.442862
Number of students who registered for the SHSAT          0.109380
Number of students who took the SHSAT                    0.333936
AllTested                                                0.804685
All4                                                     0.852334
Native4                                                  0.107553
AfricanAmerican4                                         0.134114
Latino4                                                  0.541865
Islander4                                                0.780939
White4                                                   0.645959
Multiracial4                                             0.281024
LimitedEnglish4                                          0.194009
Disadv4                                                  0.807129
AllMath4Tested                                           0.782446
AllMath4                                                 0.813443
NativeMath4                                              0.122513
AfricanAmericanMath4                                     0.018657
LatinoMath4                                              0.331389
IslanderMath4                                            0.790181
WhiteMath4                                               0.643815
MultiracialMath4                                         0.302159
LimitedEnglishMath4                                      0.441553
DisadvMath4                                              0.732673
Count of Students in HS Admissions                       0.856769
Count of Testers                                         1.000000
Count of Offers                                                 0.778770

                                                 Count of Offers
SED Code                                                0.119194
Economic Need Index                                    -0.421596
Average ELA Proficiency                                 0.520695
Average Math Proficiency                                0.468433
Number of students who registered for the SHSAT         0.101141
Number of students who took the SHSAT                   0.291501
AllTested                                               0.485583
All4                                                    0.859680
Native4                                                 0.038949
AfricanAmerican4                                        0.000243
Latino4                                                 0.285010
Islander4                                               0.829191
White4                                                  0.738594
Multiracial4                                            0.542744
LimitedEnglish4                                         0.103076
Disadv4                                                 0.678692
AllMath4Tested                                          0.456146
AllMath4                                                0.792829
NativeMath4                                             0.061064
AfricanAmericanMath4                                   -0.044907
LatinoMath4                                             0.157090
IslanderMath4                                           0.781842
WhiteMath4                                              0.727092
MultiracialMath4                                        0.532177
LimitedEnglishMath4                                     0.286620
DisadvMath4                                             0.599504
Count of Students in HS Admissions                      0.521510
Count of Testers                                        0.778770
Count of Offers                                         1.000000

[29 rows x 29 columns]
```

1.2.2 Exploratory Visualization

In [598]:
```
# Now we use corr() to get the feature correlations and
#then visualize them using a heatmap
#The correlation values are fed into the heatmap to gain further insight.

# Documentation referred at:
# http://seaborn.pydata.org/generated/seaborn.heatmap.html
import numpy as np; np.random.seed(0)
import seaborn as sns; sns.set()
import matplotlib.pyplot as plt

plt.subplots(figsize=(15,10))
```

```
        sns.heatmap(df_sub.corr())
```

```
Out[598]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2ad05828>
```



1.2.3 Insights:

From the correlation values and heatmap, we can see that 'Economic Need Index' is inversely correlated with high negative values to - 'Average ELA Proficiency', - 'Average Math Proficiency', - 'White4',
- 'WhiteMath4', - 'Count of Testers' and - 'Count of Offers'. Thus, the general assumption that schools with greater economic needs - perform more poorly on ELA and Math tests, - have lesser White students, - have fewer students taking the tests and - consequently getting less SHSAT offers holds true.

Now to apply clustering techniques, we will limit the number of features further. We will use only the following columns as : - 'Economic Need Index', - 'AfricanAmerican4', - 'Latino4', - 'White4', - 'AfricanAmericanMath4', - 'LatinoMath4', - 'WhiteMath4', - 'Number of students who took the SHSAT', - 'Count of Students in HS Admissions', - 'Count of Offers']

```
In [599]: sub = df_sub[['Economic Need Index',
          'AfricanAmerican4',
           'Latino4',
           'White4',
           'AfricanAmericanMath4',
           'LatinoMath4',
           'WhiteMath4',
           'Number of students who took the SHSAT',
           'Count of Offers']]
          sub.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1364 entries, 0 to 1363
Data columns (total 9 columns):
Economic Need Index                      1339 non-null float64
AfricanAmerican4                         1364 non-null float64
Latino4                                  1364 non-null float64
White4                                   1364 non-null float64
AfricanAmericanMath4                     1364 non-null float64
LatinoMath4                              1364 non-null float64
WhiteMath4                               1364 non-null float64
Number of students who took the SHSAT    113 non-null float64
Count of Offers                          683 non-null float64
dtypes: float64(9)
memory usage: 106.6 KB
```

```
In [600]: sub.corr()
```

```
Out[600]:                                        Economic Need Index  AfricanAmerican4  \
          Economic Need Index                               1.000000         -0.072707
          AfricanAmerican4                                 -0.072707          1.000000
          Latino4                                          -0.127446          0.095225
          White4                                           -0.549216          0.012959
          AfricanAmericanMath4                              0.004962          0.839716
          LatinoMath4                                      -0.029510          0.098546
          WhiteMath4                                       -0.566390         -0.004120
          Number of students who took the SHSAT            -0.172916          0.098985
          Count of Offers                                  -0.421596          0.000243

                                                  Latino4    White4  \
          Economic Need Index                   -0.127446 -0.549216
          AfricanAmerican4                       0.095225  0.012959
          Latino4                                1.000000  0.277335
          White4                                 0.277335  1.000000
          AfricanAmericanMath4                   0.058003 -0.044113
          LatinoMath4                            0.833767  0.136859
          WhiteMath4                             0.249102  0.973047
          Number of students who took the SHSAT  0.254371  0.291501
          Count of Offers                        0.285010  0.738594
```

```
                                       AfricanAmericanMath4  LatinoMath4  \
Economic Need Index                                0.004962    -0.029510   
AfricanAmerican4                                   0.839716     0.098546   
Latino4                                            0.058003     0.833767   
White4                                            -0.044113     0.136859   
AfricanAmericanMath4                               1.000000     0.161130   
LatinoMath4                                        0.161130     1.000000   
WhiteMath4                                        -0.048821     0.132373   
Number of students who took the SHSAT             -0.036843     0.080072   
Count of Offers                                   -0.044907     0.157090   

                                       WhiteMath4  \
Economic Need Index                     -0.566390   
AfricanAmerican4                        -0.004120   
Latino4                                  0.249102   
White4                                   0.973047   
AfricanAmericanMath4                    -0.048821   
LatinoMath4                              0.132373   
WhiteMath4                               1.000000   
Number of students who took the SHSAT    0.291501   
Count of Offers                          0.727092   

                                       Number of students who took the SHSAT  \
Economic Need Index                                                -0.172916   
AfricanAmerican4                                                    0.098985   
Latino4                                                             0.254371   
White4                                                              0.291501   
AfricanAmericanMath4                                               -0.036843   
LatinoMath4                                                         0.080072   
WhiteMath4                                                          0.291501   
Number of students who took the SHSAT                               1.000000   
Count of Offers                                                     0.291501   

                                       Count of Offers  
Economic Need Index                          -0.421596  
AfricanAmerican4                              0.000243  
Latino4                                       0.285010  
White4                                        0.738594  
AfricanAmericanMath4                         -0.044907  
LatinoMath4                                   0.157090  
WhiteMath4                                    0.727092  
Number of students who took the SHSAT         0.291501  
Count of Offers                               1.000000  
```

```
In [601]: sub=sub.fillna(0)
```

```
In [602]: # copy dataframe into sub1 for decision tree to get score
```

```
          sub1 = sub
          sub1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1364 entries, 0 to 1363
Data columns (total 9 columns):
Economic Need Index                         1364 non-null float64
AfricanAmerican4                            1364 non-null float64
Latino4                                     1364 non-null float64
White4                                      1364 non-null float64
AfricanAmericanMath4                        1364 non-null float64
LatinoMath4                                 1364 non-null float64
WhiteMath4                                  1364 non-null float64
Number of students who took the SHSAT       1364 non-null float64
Count of Offers                             1364 non-null float64
dtypes: float64(9)
memory usage: 106.6 KB
```

We use Decision Tree metjod to determine if indeed our selected features are relevant or not using score() methods.

```
In [603]: from sklearn.tree import DecisionTreeRegressor
          from sklearn.cross_validation import train_test_split

          # Done: Make a copy of the DataFrame, using the 'drop' function to drop the given fe
          new_data = sub1.drop('Count of Offers', axis=1)

          # Done: Split the data into training and testing sets(0.25) using the given feature 
          # Set a random state.
          X_train, X_test, y_train, y_test = train_test_split(new_data, sub1['Count of Offers']

          # Done: Create a decision tree regressor and fit it to the training set
          regressor = DecisionTreeRegressor(random_state=32)
          regressor.fit(X_train,y_train)

          # Done: Report the score of the prediction using the testing set
          score = regressor.score(X_test,y_test)

          print(round(score,4))
```

```
0.0923
```

```
In [604]: # copy dataframe into sub2 for decision tree to get score
          sub2 = sub
          sub2.info()

          from sklearn.tree import DecisionTreeRegressor
          from sklearn.cross_validation import train_test_split

          # Done: Make a copy of the DataFrame, using the 'drop' function to drop the given fe
          new_data = sub2.drop('Economic Need Index', axis=1)

          # Done: Split the data into training and testing sets(0.25) using the given feature
          # Set a random state.
          X_train, X_test, y_train, y_test = train_test_split(new_data, sub2['Economic Need Inc

          # Done: Create a decision tree regressor and fit it to the training set
          regressor = DecisionTreeRegressor(random_state=32)
          regressor.fit(X_train,y_train)

          # Done: Report the score of the prediction using the testing set
          score = regressor.score(X_test,y_test)

          print(round(score,4))
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1364 entries, 0 to 1363
Data columns (total 9 columns):
Economic Need Index                        1364 non-null float64
AfricanAmerican4                           1364 non-null float64
Latino4                                    1364 non-null float64
White4                                     1364 non-null float64
AfricanAmericanMath4                       1364 non-null float64
LatinoMath4                                1364 non-null float64
WhiteMath4                                 1364 non-null float64
Number of students who took the SHSAT      1364 non-null float64
Count of Offers                            1364 non-null float64
dtypes: float64(9)
memory usage: 146.6 KB
0.2061
```

```
In [605]: # copy dataframe into sub3 for decision tree to get score
          sub3 = sub
          sub3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1364 entries, 0 to 1363
Data columns (total 9 columns):
Economic Need Index                        1364 non-null float64
AfricanAmerican4                           1364 non-null float64
Latino4                                    1364 non-null float64
White4                                     1364 non-null float64
AfricanAmericanMath4                       1364 non-null float64
LatinoMath4                                1364 non-null float64
```

```
WhiteMath4                                  1364 non-null float64
Number of students who took the SHSAT       1364 non-null float64
Count of Offers                             1364 non-null float64
dtypes: float64(9)
memory usage: 146.6 KB
```

```
In [606]: from sklearn.tree import DecisionTreeRegressor
          from sklearn.cross_validation import train_test_split

          # Done: Make a copy of the DataFrame, using the 'drop' function to drop the given fe
          new_data = sub3.drop('Number of students who took the SHSAT', axis=1)

          # Done: Split the data into training and testing sets(0.25) using the given feature
          # Set a random state.
          X_train, X_test, y_train, y_test = train_test_split(new_data, sub3['Number of studen

          # Done: Create a decision tree regressor and fit it to the training set
          regressor = DecisionTreeRegressor(random_state=32)
          regressor.fit(X_train,y_train)

          # Done: Report the score of the prediction using the testing set
          score = regressor.score(X_test,y_test)

          print(round(score,4))
```

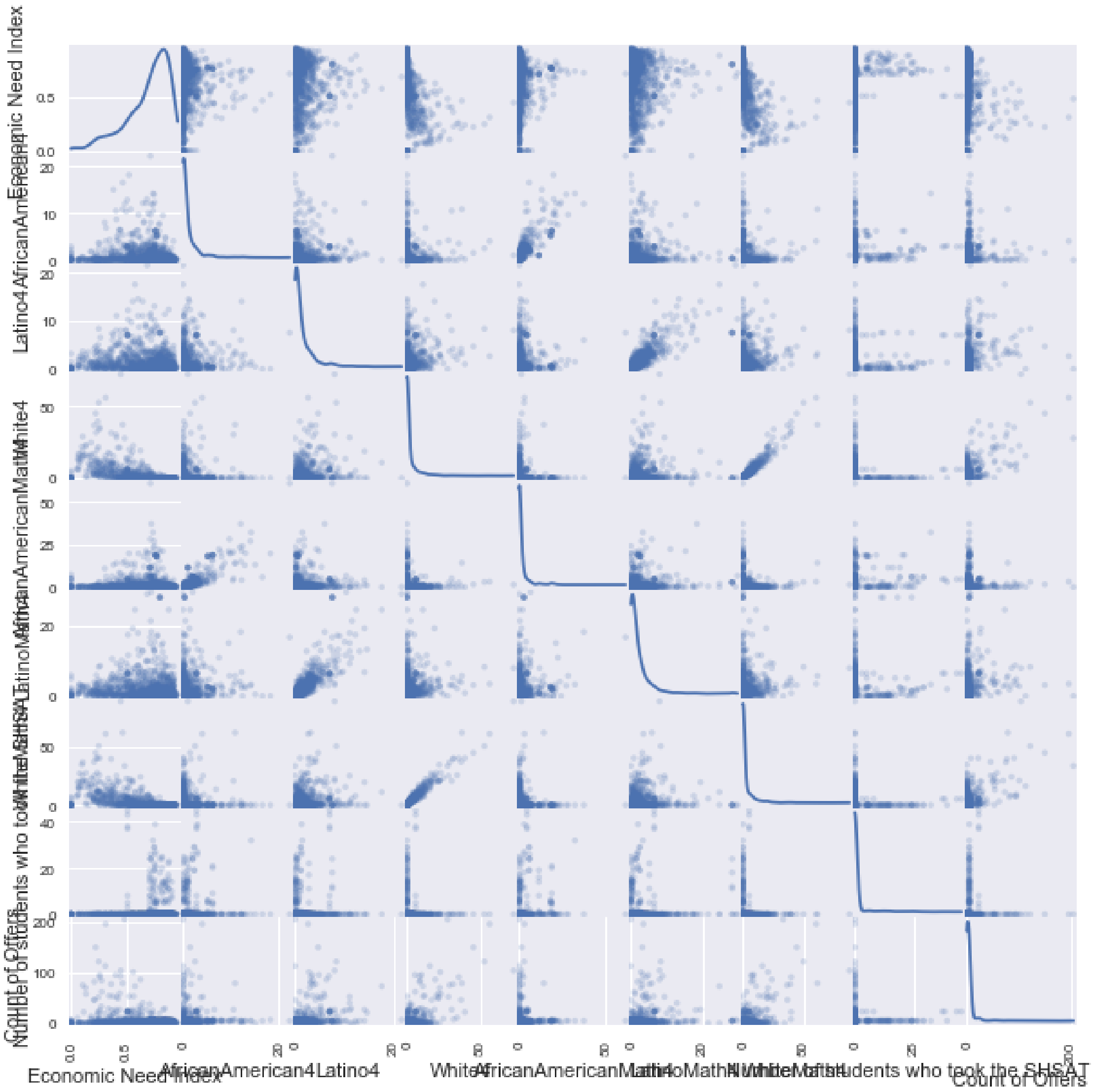
```
0.3862
```

The coefficient of determination, Rˆ2, is scored between 0 and 1, with 1 being a perfect fit. A negative Rˆ2 implies the model fails to fit the data. If you get a low score for a particular feature, that lends us to beleive that that feature point is hard to predict using the other features, thereby making it an important feature to consider when considering relevance. Thus, 'Count of Offers' and 'Economic Need Index' are important features while 'Number of students who took the SHSAT' is of medium importance and validate our choice of features in the subset we are considering.

1.2.4 Algorithms and Techniques

Here we will be using Kmeans and GMM algorithms to determine the best clustering method for the datasets. We will use scatterpplot matrix, Elbow curve and Silhouette score to evaluate if the number of clusters is optimal. The silhouette score will also be used to compare against benchmark model. We are using Kmeans clustering as it is easy to implement, provides tight clusters and with our use of Silhouette score can give us the optimal number of clusters as well. (offsets its disadvantage of difficulty in predicting number of clusters) The Gaussian mixture model (GMM) attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset. This method will also ensure that the main practical issues with k-means will be removed namely, allowing for a full covariance and not using hard cutoffs for cluster assignment within the training set. thus, we can be sure that the best clusters and numbers are obtained.

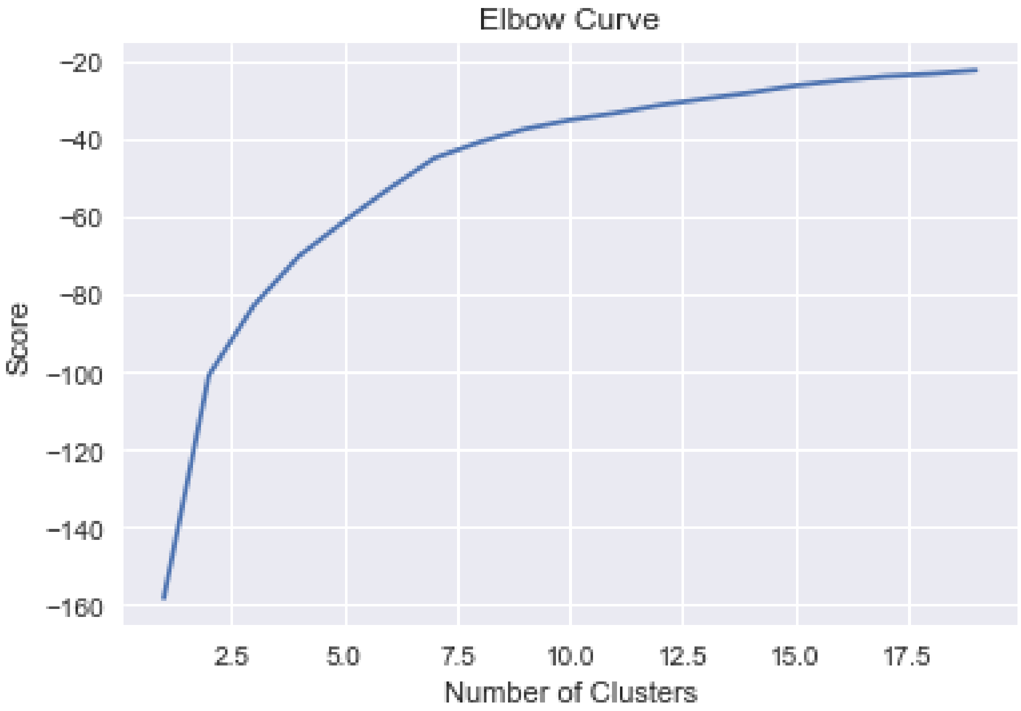
```
In [607]: from pandas.plotting import scatter_matrix

          scatter_matrix(sub, alpha=0.2, figsize=(10, 10), diagonal='kde');
```



The Elbow method is a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset.

```
In [608]: import pylab as pl
          from sklearn.preprocessing import StandardScaler
          from sklearn.preprocessing import MinMaxScaler

          #Scaling of data
          ss = StandardScaler()
          ss.fit_transform(sub)
```

```
minmax = MinMaxScaler()
minmax.fit_transform(sub)

x = sub.iloc[:, ].values
z = minmax.fit_transform(x)
Nc = range(1, 20)
kmeans = [KMeans(n_clusters=i) for i in Nc]

score = [kmeans[i].fit(z).score(z) for i in range(len(kmeans))]
pl.plot(Nc,score)
pl.xlabel('Number of Clusters')
pl.ylabel('Score')
pl.title('Elbow Curve')
pl.show()
```



Thus, we can consider 3 clusters although 7 or 8 clusters also look feasible using the Elbow Curve method.

1.3 Methodology

1.3.1 Data Preprocessing

Since the dataset is from Kaggle, only minimal preprocessing was done on the dataset. Most of the wrangling was simply joining the datasets and subsetting to include only relevant attributes. The rating percentages had converted to float values with the percent symbol stripped while test scores for all each particular race was computed as average of individual grade level scores. Feature relevance was determined by using creating a decision tree regressor and fitting it to the training set and then reporting the score of the prediction using the testing set.

1.3.2 Implementation

For execution, we will be using sklearn to determine the clusters and centers using K-means. We will then do the same with GMM algorithm and then compute the silhouette score. We shall do so for cluster numbered for 3,4,7,8. We will then compute silhouette score for clusters numbered 3 to 8 using Kmeans and additionally create cluster maps to see how each clsuter number chosen fared.

1.3.3 Refinement

The refinement was basically to evaluate which cluster number gave the best results. Hence, both GMM and Kmeans algorithms were used. Also, my proposal simply considerd silhouette score but I added using elbow curve method as well and evaluated this metric for all the clsuter numbers and for both algorithms to hone in on the right groups.

Section 1.5.3: Back to Top

1.3.4 Cluster centers for n=8 clusters using K-means

```
In [609]: from sklearn.cluster import KMeans
          import numpy as np
          X = sub
          kmeans = KMeans(n_clusters=8, random_state=0).fit(sub)
          kmeans.labels_
          #kmeans.predict([[0, 0], [4, 4]])
          cluster_centers = kmeans.cluster_centers_
          print(cluster_centers)
```

```
[[  2.86024096e-01   3.19277108e-01   1.63253012e+00   1.14959839e+01
    3.19277108e-01   2.05823293e+00   1.50582329e+01   0.00000000e+00
    2.16867470e+00]
 [  3.41000000e-01   1.80000000e+00   4.73333333e+00   5.07000000e+01
    1.86666667e+00   4.60000000e+00   5.52333333e+01   0.00000000e+00
    1.55400000e+02]
 [  7.03391595e-01   6.28621458e-01   8.40337472e-01   4.89652977e-01
    7.32569245e-01   1.19134034e+00   6.72715696e-01   1.43266476e-01
    2.26552053e+00]
 [  3.41120000e-01   1.22000000e+00   4.58000000e+00   1.92600000e+01
    1.16666667e+00   4.87333333e+00   2.15400000e+01  -2.22044605e-16
    7.18800000e+01]
 [  5.21208333e-01   1.37847222e+00   4.88888889e+00   5.78125000e+00
    1.14236111e+00   5.01041667e+00   5.55208333e+00   3.33333333e-01
    2.34166667e+01]
 [  7.61400000e-01   1.67000000e+00   1.29000000e+00   4.80000000e-01
    3.86333333e+00   2.12333333e+00   4.13333333e-01   2.05400000e+01
    6.44000000e+00]
 [  6.85725490e-01   7.95751634e+00   1.26470588e+00   7.84313725e-02
    1.78790850e+01   3.06209150e+00   8.16993464e-02   1.15686275e+00
    4.62745098e+00]
 [  7.10927273e-01   9.93939394e-01   6.83030303e+00   9.54545455e-01
    1.63939394e+00   1.27515152e+01   1.11212121e+00   7.45454545e-01
    2.65454545e+00]]
```

1.3.5 Silhouette score for n=8 clusters using GMM

```
In [610]: # Links used:
          # http://scikit-learn.org/stable/modules/clustering.html
          # http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.
          from sklearn.mixture import GMM
          from sklearn.metrics import silhouette_score
          import warnings
          warnings.filterwarnings('ignore')


          # Done: Apply your clustering algorithm of choice to the reduced data
          x3 = GMM(n_components=8)

          reduced_data=sub
          clusterer = x3.fit(reduced_data)

          # Done: Predict the cluster for each data point
          preds = clusterer.predict(reduced_data)

          # Done: Find the cluster centers
          centers = clusterer.means_

          # Done: Predict the cluster for each transformed sample data point
          sample_preds = clusterer.predict(cluster_centers)

          # Done: Calculate the mean silhouette coefficient for the number of clusters chosen
          score = silhouette_score(reduced_data, preds)

          print(round(score,4))
          print("Scores for 8 clusters is {}".format(round(score,4)))
```

```
0.0867
Scores for 8 clusters is 0.0867
```

1.3.6 Cluster centers for n=3 clusters using K-means

```
In [611]: from sklearn.cluster import KMeans
          import numpy as np
          X2 = sub
          kmeans = KMeans(n_clusters=3, random_state=0).fit(X2)
          kmeans.labels_
          cluster_centers = kmeans.cluster_centers_
          print(cluster_centers)
```

```
[[  6.73086336e-01   9.71596597e-01   1.31481481e+00   1.34034034e+00
    1.53365866e+00   1.96559059e+00   1.70170170e+00   9.70720721e-01
    3.23198198e+00]
 [  3.45518519e-01   1.15432099e+00   4.36419753e+00   1.92962963e+01
    1.10493827e+00   4.66666667e+00   2.14259259e+01  -3.33066907e-16
    6.93333333e+01]
 [  3.41000000e-01   1.80000000e+00   4.73333333e+00   5.07000000e+01
    1.86666667e+00   4.60000000e+00   5.52333333e+01   0.00000000e+00
    1.55400000e+02]]
```

1.3.7 Silhouette score for n=3 clusters using GMM

```
In [612]: # Links used:
          # http://scikit-learn.org/stable/modules/clustering.html
          # http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.
          from sklearn.mixture import GMM
          from sklearn.metrics import silhouette_score
          import warnings
          warnings.filterwarnings('ignore')


          # Done: Apply your clustering algorithm of choice to the reduced data
          x3 = GMM(n_components=3)

          reduced_data2=sub
          clusterer = x3.fit(reduced_data2)

          # Done: Predict the cluster for each data point
          preds = clusterer.predict(reduced_data2)

          # Done: Find the cluster centers
          centers = clusterer.means_

          # Done: Predict the cluster for each transformed sample data point
          sample_preds = clusterer.predict(cluster_centers)

          # Done: Calculate the mean silhouette coefficient for the number of clusters chosen
          score = silhouette_score(reduced_data2, preds)
          print(round(score,4))
          print("Scores for 3 clusters is {}".format(round(score,4)))
```

```
0.2614
Scores for 3 clusters is 0.2614
```

1.3.8 Cluster centers for n=7 clusters using K-means

```
In [613]: from sklearn.cluster import KMeans
          import numpy as np
          X2 = sub
          kmeans = KMeans(n_clusters=7, random_state=0).fit(X2)
          kmeans.labels_
          cluster_centers = kmeans.cluster_centers_
          print(cluster_centers)
```

```
[[  6.88205457e-01   4.97057250e-01   9.84216158e-01   6.23863028e-01
    7.41305511e-01   1.76404494e+00   9.21883360e-01  -4.44089210e-16
    5.50670620e-14]
 [  3.41000000e-01   1.80000000e+00   4.73333333e+00   5.07000000e+01
    1.86666667e+00   4.60000000e+00   5.52333333e+01   0.00000000e+00
    1.55400000e+02]
 [  3.41120000e-01   1.22000000e+00   4.58000000e+00   1.92600000e+01
    1.16666667e+00   4.87333333e+00   2.15400000e+01  -2.22044605e-16
    7.18800000e+01]
 [  5.19595745e-01   1.40780142e+00   4.94680851e+00   5.85460993e+00
    1.16666667e+00   5.09574468e+00   5.62765957e+00   3.40425532e-01
    2.36170213e+01]
 [  7.20851562e-01   8.88671875e-01   1.24479167e+00   4.40429688e-01
    9.38802083e-01   1.51660156e+00   5.08463542e-01   7.81250000e-01
    5.19921875e+00]
 [  2.77413333e-01   3.44444444e-01   1.63777778e+00   1.21000000e+01
    3.40000000e-01   2.02000000e+00   1.57688889e+01   4.44089210e-16
    2.13333333e+00]
 [  7.05155844e-01   5.69047619e+00   1.93290043e+00   3.63636364e-01
    1.32554113e+01   4.62121212e+00   3.22510823e-01   1.13896104e+01
    5.81818182e+00]]
```

1.3.9 Silhouette score for n=7 clusters using GMM

```
In [614]: # Links used:
          # http://scikit-learn.org/stable/modules/clustering.html
          # http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.
          from sklearn.mixture import GMM
          from sklearn.metrics import silhouette_score
          import warnings
          warnings.filterwarnings('ignore')


          # Done: Apply your clustering algorithm of choice to the reduced data
          x3 = GMM(n_components=7)

          reduced_data2=sub
          clusterer = x3.fit(reduced_data2)

          # Done: Predict the cluster for each data point
          preds = clusterer.predict(reduced_data2)

          # Done: Find the cluster centers
          centers = clusterer.means_

          # Done: Predict the cluster for each transformed sample data point
          sample_preds = clusterer.predict(cluster_centers)

          # Done: Calculate the mean silhouette coefficient for the number of clusters chosen
          score = silhouette_score(reduced_data2, preds)

          print(round(score,4))
          print("Scores for 7 clusters is {}".format(round(score,4)))
```

```
0.1008
Scores for 7 clusters is 0.1008
```

1.3.10 Cluster centers for n=4 clusters using K-means

```
In [615]: from sklearn.cluster import KMeans
          import numpy as np
          X2 = sub
          kmeans = KMeans(n_clusters=4, random_state=0).fit(X2)
          kmeans.labels_
          cluster_centers = kmeans.cluster_centers_
          print(cluster_centers)
```

```
[[  7.01517270e-01   9.85882675e-01   1.12595943e+00   5.18777412e-01
    1.60389254e+00   1.80317982e+00   7.12856360e-01   9.16118421e-01
    2.55016447e+00]
 [  3.49615385e-01   1.19230769e+00   4.49358974e+00   1.85256410e+01
    1.14743590e+00   4.81410256e+00   2.07179487e+01  -3.33066907e-16
    7.11538462e+01]
 [  3.41000000e-01   1.80000000e+00   4.73333333e+00   5.07000000e+01
    1.86666667e+00   4.60000000e+00   5.52333333e+01   0.00000000e+00
    1.55400000e+02]
 [  3.73888889e-01   8.16239316e-01   3.27492877e+00   1.02037037e+01
```

```
    7.90598291e-01   3.64387464e+00   1.23048433e+01   1.52991453e+00
    1.04786325e+01]]
```

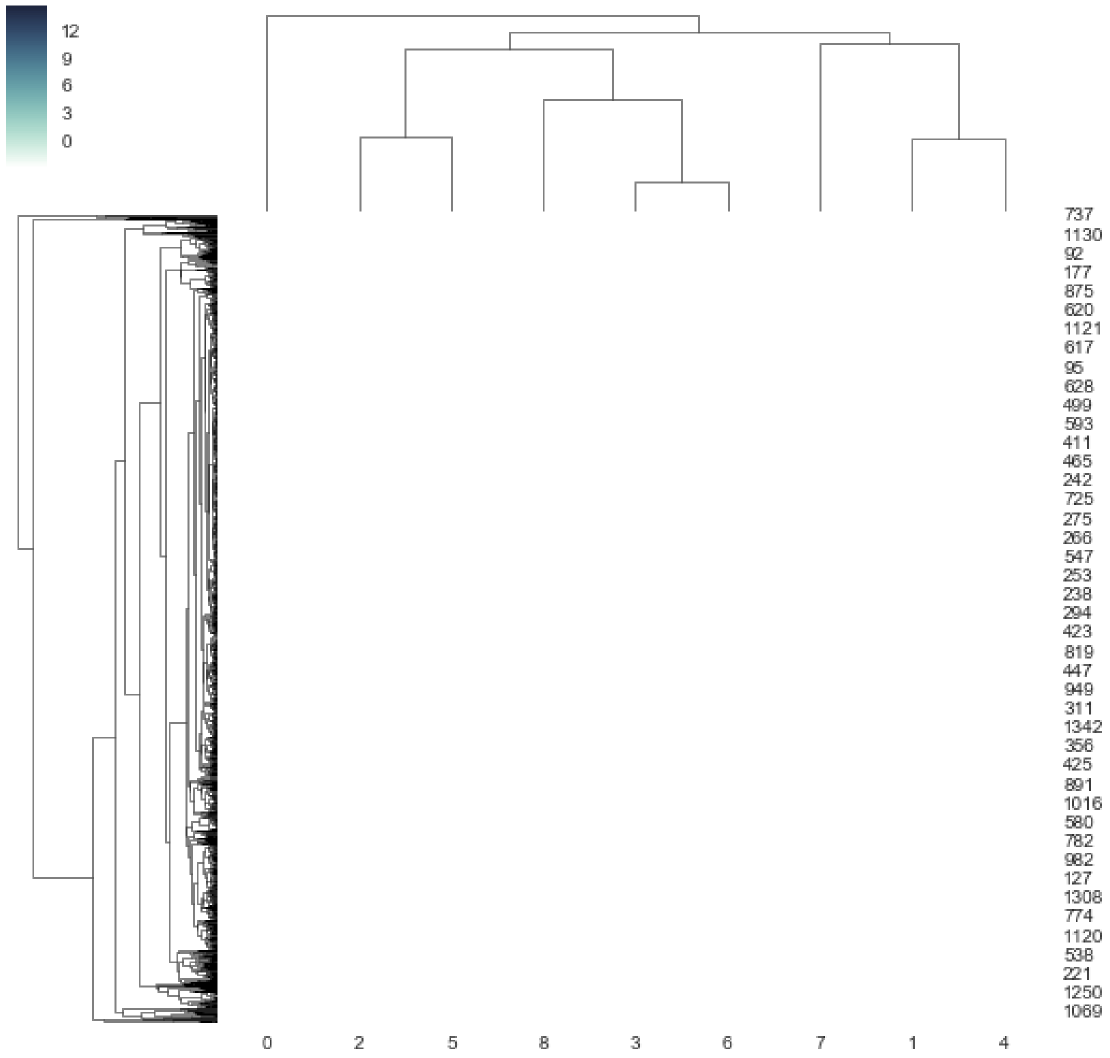
1.3.11 Silhouette score for n=4 clusters using GMM

```
In [616]: # Links used:
          # http://scikit-learn.org/stable/modules/clustering.html
          # http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.
          from sklearn.mixture import GMM
          from sklearn.metrics import silhouette_score
          import warnings
          warnings.filterwarnings('ignore')


          # Done: Apply your clustering algorithm of choice to the reduced data
          x3 = GMM(n_components=4)

          reduced_data2=sub
          clusterer = x3.fit(reduced_data2)

          # Done: Predict the cluster for each data point
          preds = clusterer.predict(reduced_data2)

          # Done: Find the cluster centers
          centers = clusterer.means_

          # Done: Predict the cluster for each transformed sample data point
          sample_preds = clusterer.predict(cluster_centers)

          # Done: Calculate the mean silhouette coefficient for the number of clusters chosen
          score = silhouette_score(reduced_data2, preds)

          print(round(score,4))
          print("Scores for 4 clusters is {}".format(round(score,4)))
```

```
0.1711
Scores for 4 clusters is 0.1711
```

```
In [617]: #Creating cluster map to visualize clusters
          from sklearn.metrics import silhouette_samples, silhouette_score
```

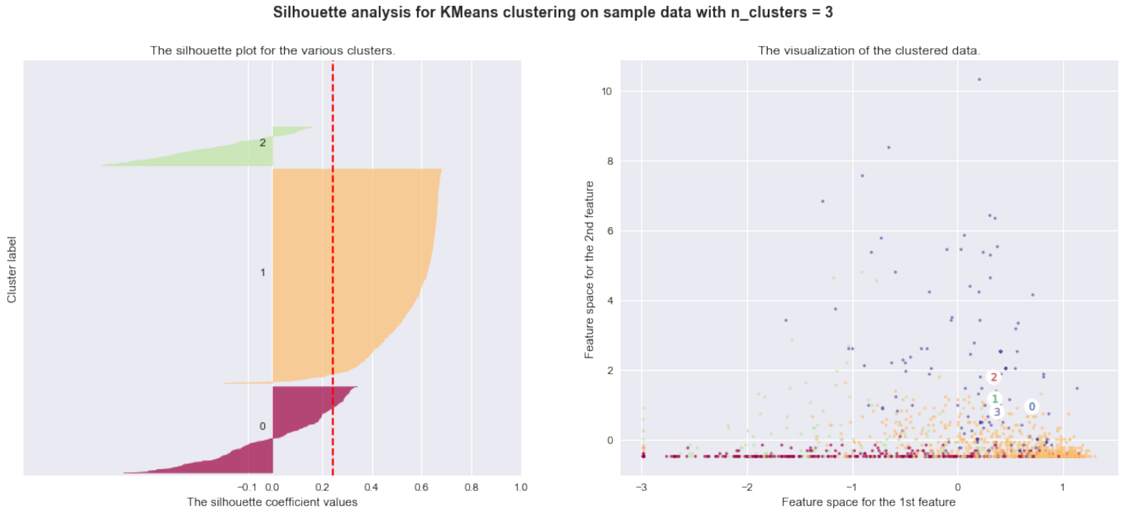
```
In [618]: X = sub[['Economic Need Index',
          'AfricanAmerican4',
           'Latino4',
           'White4',
           'AfricanAmericanMath4',
           'LatinoMath4',
           'WhiteMath4',
           'Number of students who took the SHSAT',
           'Count of Offers']]

          from sklearn.preprocessing import StandardScaler
          scaler = StandardScaler()
          X_scaled = scaler.fit_transform( X )

In [619]: import matplotlib.cm as cm
          import seaborn as sn
          cmap = sn.cubehelix_palette(as_cmap=True, rot=-.3, light=1)

In [620]: sn.clustermap(X_scaled, cmap=cmap, linewidths=.5);
```



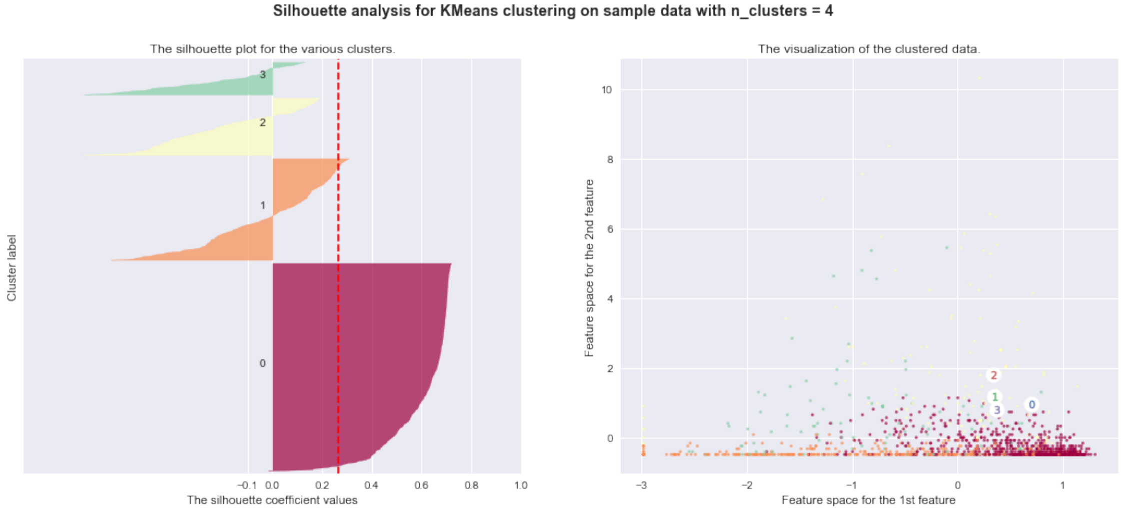
1.3.12 Cluster maps and silhouette scores using Kmeans for number of clusters ranging from 3 to 8

```
In [621]: %matplotlib inline

          cluster_range = range( 3, 9 ) #range excludes upper bound

          for n_clusters in cluster_range:
            fig, (ax1, ax2) = plt.subplots(1, 2)
            fig.set_size_inches(18, 7)

            # The 1st subplot is the silhouette plot
            # The silhouette coefficient can range from -1, 1
            ax1.set_xlim([-1, 1])

           # The (n_clusters+1)*10 is for inserting blank space between silhouette
            # plots of individual clusters, to demarcate them clearly.
            ax1.set_ylim([0, len(X_scaled) + (n_clusters + 1) * 10])

            cluster_labels = clusterer.fit_predict( X_scaled )

            # The silhouette_score gives the average value for all the samples.
            # This gives a perspective into the density and separation of the formed
            # clusters
            silhouette_avg = silhouette_score(X_scaled, cluster_labels)
            print("For n_clusters =", n_clusters,
                  "The average silhouette_score is :", silhouette_avg)

            # Compute the silhouette scores for each sample
            sample_silhouette_values = silhouette_samples(X_scaled, cluster_labels)

            y_lower = 10
            for i in range(n_clusters):
                # Aggregate the silhouette scores for samples belonging to
                # cluster i, and sort them
                ith_cluster_silhouette_values = \
                    sample_silhouette_values[cluster_labels == i]

                ith_cluster_silhouette_values.sort()

                size_cluster_i = ith_cluster_silhouette_values.shape[0]
                y_upper = y_lower + size_cluster_i

                cmap = cm.get_cmap("Spectral")
                color = cmap(float(i) / n_clusters)

                #color = cm.spectral(float(i) / n_clusters)
                ax1.fill_betweenx(np.arange(y_lower, y_upper),
                                  0, ith_cluster_silhouette_values,
                                  facecolor=color, edgecolor=color, alpha=0.7)

                # Label the silhouette plots with their cluster numbers at the middle
                ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))

                # Compute the new y_lower for next plot
                y_lower = y_upper + 10
            ax1.set_title("The silhouette plot for the various clusters.")
            ax1.set_xlabel("The silhouette coefficient values")
            ax1.set_ylabel("Cluster label")

            # The vertical line for average silhoutte score of all the values
            ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

            ax1.set_yticks([])  # Clear the yaxis labels / ticks
            ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])

            # 2nd Plot showing the actual clusters formed
            #cmap = cm.get_cmap("Spectral")
            colors = cmap(cluster_labels.astype(float) / n_clusters)
            #colors = cm.spectral(cluster_labels.astype(float) / n_clusters)
            ax2.scatter(X_scaled[:, 0], X_scaled[:, 1], marker='.', s=30, lw=0, alpha=0.7,
                        c=colors)

            # Labeling the clusters
            centers = kmeans.cluster_centers_
            #centers = clusterer.cluster_centers_
            # Draw white circles at cluster centers
            ax2.scatter(centers[:, 0], centers[:, 1],
                        marker='o', c="white", alpha=1, s=200)

            for i, c in enumerate(centers):
                ax2.scatter(c[0], c[1], marker='$%d$' % i, alpha=1, s=50)

            ax2.set_title("The visualization of the clustered data.")
            ax2.set_xlabel("Feature space for the 1st feature")
            ax2.set_ylabel("Feature space for the 2nd feature")

            plt.suptitle(("Silhouette analysis for KMeans clustering on sample data "
                          "with n_clusters = %d" % n_clusters),
                         fontsize=14, fontweight='bold')

            plt.show();
```

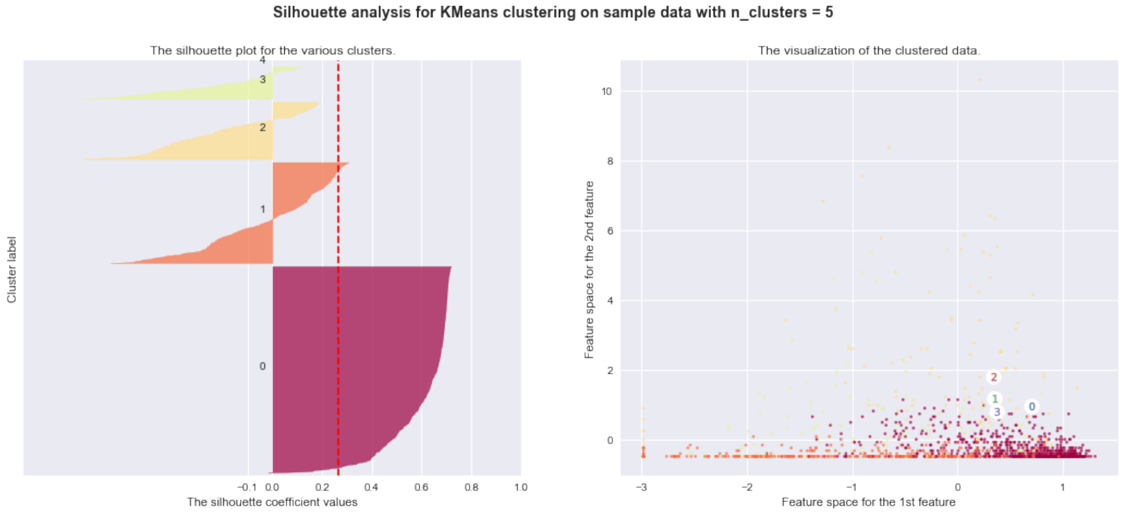
```
For n_clusters = 3 The average silhouette_score is : 0.243592391953
```



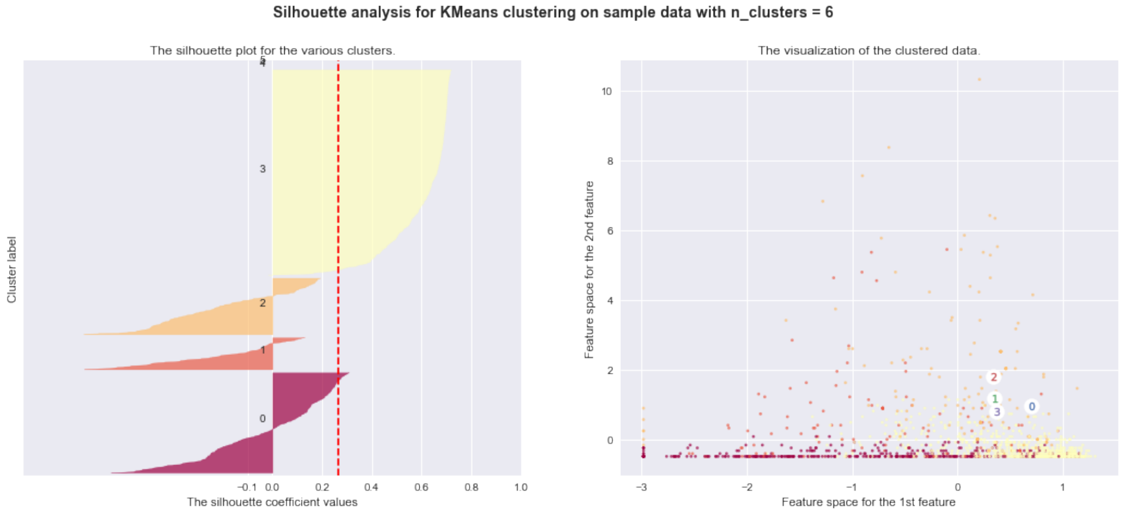
```
For n_clusters = 4 The average silhouette_score is : 0.266591843412
```



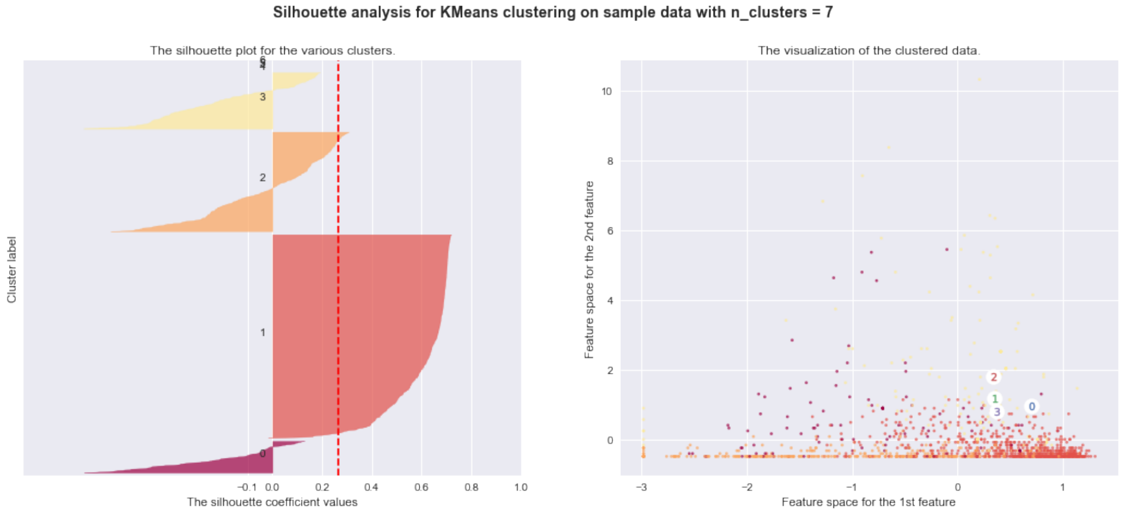
```
For n_clusters = 5 The average silhouette_score is : 0.266591843412
```



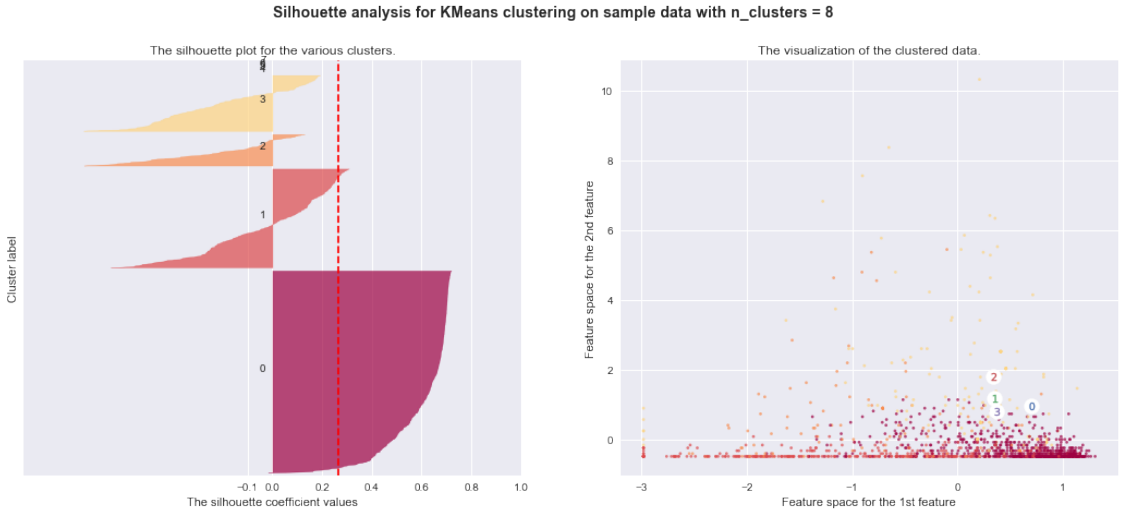
```
For n_clusters = 6 The average silhouette_score is : 0.266775421534
```



```
For n_clusters = 7 The average silhouette_score is : 0.266591843412
```



```
For n_clusters = 8 The average silhouette_score is : 0.266591843412
```



Thus, for clusters =3, we get the best results when considering clustermap, GMM and K-means techniques.

1.3.13 Benchmark:

For this project, the benchmark model is the kernel from the Kaggle dataset at: https://www.kaggle.com/laiyipeng/target-schools-action-recommended-for-passnyc It was one of the winners when the competition (using this dataset) originally ran. There are 3 clusters present in this analysis. I shall try to compare my model with this analysis.

In the benchmark model, we have: 1) New York State annual test results 2) New York City Department of Education annual quality review 3) Other

Since this original kernel is in R and approaches the problem with a different angle and does not compute silhouette score, am reproducing the dataset and joins used and computing the silhouette scores for this model.

```
In [622]: benchmark =dfy
          benchmark.shape
```

```
Out[622]: (1364, 71)
```

```
In [623]: benchmark = benchmark.drop([
                'School Name',
                'SED Code',
                'Location Code',
                'District',
                'Latitude',
                'Longitude',
                'Address (Full)',
                'City',
                'Zip',
                'Grades',
                'Grade Low',
                'Grade High',
                'Community School?',
                'Economic Need Index',
                'School Income Estimate',
                'Percent ELL',
                'Percent Asian',
                'Percent Black',
                'Percent Hispanic',
                'Percent Black / Hispanic',
                'Percent White',
                'Student Achievement Rating',
                'Grade 3 Math 4s - Multiracial',
                'DBN',
                'School name',
                'Year of SHST',
                'Grade level',
                'Enrollment on 10/31',
                'Number of students who registered for the SHSAT',
                'Number of students who took the SHSAT',
                'AllTested',
                'All4',
                'Native4',
                'AfricanAmerican4',
                'Latino4',
                'Islander4',
                'White4',
                'Multiracial4',
                'LimitedEnglish4',
                'Disadv4',
                'AllMath4Tested',
                'AllMath4',
                'NativeMath4',
                'AfricanAmericanMath4',
                'LatinoMath4',
                'IslanderMath4',
                'WhiteMath4',
                'MultiracialMath4',
                'LimitedEnglishMath4',
                'DisadvMath4',
                'Feeder School DBN',
                'Feeder School Name',
                'Count of Students in HS Admissions',
                'Count of Testers',
                'Rigorous Instruction Rating',
                'Collaborative Teachers Rating',
                'Supportive Environment Rating',
                'Effective School Leadership Rating',
                'Strong Family-Community Ties Rating',
                'Trust Rating'],
           axis=1)
```

```
In [624]: benchmark.shape
```

```
Out[624]: (1364, 11)
```

```
In [625]: list(benchmark)
```

```
Out[625]: ['Student Attendance Rate',
           'Percent of Students Chronically Absent',
           'Rigorous Instruction %',
           'Collaborative Teachers %',
           'Supportive Environment %',
           'Effective School Leadership %',
           'Strong Family-Community Ties %',
           'Trust %',
           'Average ELA Proficiency',
           'Average Math Proficiency',
           'Count of Offers']
```

```
In [626]: benchmark.head(2)
```

```
Out[626]:   Student Attendance Rate Percent of Students Chronically Absent  \
          0                     94%                                    18%
          1                     92%                                    30%
```

```
  Rigorous Instruction % Collaborative Teachers % Supportive Environment %  \
0                    89%                      94%                      86%
1                    96%                      96%                      97%

  Effective School Leadership % Strong Family-Community Ties % Trust %  \
0                           91%                            85%     94%
1                           90%                            86%     94%

   Average ELA Proficiency  Average Math Proficiency  Count of Offers
0                     2.14                      2.17              NaN
1                     2.63                      2.98              NaN
```

```
In [627]: benchmark['Student Attendance Rate']  = benchmark['Student Attendance Rate'] .str.rs
          benchmark['Percent of Students Chronically Absent']  = benchmark['Percent of Students
          benchmark['Rigorous Instruction %']  = benchmark['Rigorous Instruction %'] .str.rstri
          benchmark['Collaborative Teachers %']  = benchmark['Collaborative Teachers %'] .str.
          benchmark['Supportive Environment %']  = benchmark['Supportive Environment %'] .str.
          benchmark['Effective School Leadership %']  = benchmark['Effective School Leadership
          benchmark['Strong Family-Community Ties %']  = benchmark['Strong Family-Community Ti
          benchmark['Trust %']  = benchmark['Trust %'] .str.rstrip('%').astype('float') / 100.
```

```
In [628]: benchmark.head(2)
```

```
Out[628]:    Student Attendance Rate  Percent of Students Chronically Absent  \
          0                     0.94                                    0.18
          1                     0.92                                    0.30

             Rigorous Instruction %  Collaborative Teachers %  Supportive Environment %  \
          0                    0.89                      0.94                      0.86
          1                    0.96                      0.96                      0.97

             Effective School Leadership %  Strong Family-Community Ties %  Trust %  \
          0                           0.91                            0.85     0.94
          1                           0.90                            0.86     0.94

             Average ELA Proficiency  Average Math Proficiency  Count of Offers
          0                     2.14                      2.17              NaN
          1                     2.63                      2.98              NaN
```

```
In [629]: benchmark.fillna(0)
```

```
Out[629]:       Student Attendance Rate  Percent of Students Chronically Absent  \
          0                        0.94                                    0.18
          1                        0.92                                    0.30
          2                        0.94                                    0.20
          3                        0.92                                    0.28
          4                        0.93                                    0.23
          5                        0.92                                    0.33
```

6	0.95	0.13
7	0.91	0.36
8	0.93	0.27
9	0.92	0.27
10	0.98	0.02
11	0.91	0.37
12	0.84	0.58
13	0.90	0.35
14	0.95	0.11
15	0.92	0.26
16	0.93	0.23
17	0.93	0.23
18	0.94	0.15
19	0.92	0.27
20	0.97	0.05
21	0.97	0.03
22	0.96	0.07
23	0.95	0.18
24	0.98	0.06
25	0.95	0.13
26	0.96	0.05
27	0.96	0.10
28	0.96	0.13
29	0.96	0.07
...	...	...
1334	0.00	0.00
1335	0.94	0.19
1336	0.00	0.00
1337	0.00	0.00
1338	0.96	0.09
1339	0.93	0.26
1340	1.00	0.00
1341	0.97	0.04
1342	0.95	0.11
1343	0.95	0.12
1344	0.00	0.00
1345	0.96	0.05
1346	0.97	0.06
1347	0.93	0.23
1348	0.95	0.14
1349	0.95	0.13
1350	0.94	0.20
1351	0.00	0.00
1352	0.00	1.00
1353	0.00	1.00
1354	0.95	0.33
1355	0.96	0.10
1356	0.91	0.34

```
1357                     0.96                                     0.10
1358                     0.96                                     0.10
1359                     0.95                                     0.13
1360                     0.94                                     0.24
1361                     0.95                                     0.12
1362                     0.95                                     0.12
1363                     0.93                                     0.22

      Rigorous Instruction %  Collaborative Teachers %  \
0                       0.89                      0.94   
1                       0.96                      0.96   
2                       0.87                      0.77   
3                       0.85                      0.78   
4                       0.90                      0.88   
5                       0.93                      0.99   
6                       0.88                      0.78   
7                       0.87                      0.89   
8                       0.94                      0.91   
9                       0.92                      0.89   
10                      0.90                      0.81   
11                      1.00                      1.00   
12                      0.72                      0.77   
13                      0.84                      0.78   
14                      0.90                      0.93   
15                      0.92                      0.96   
16                      0.97                      0.97   
17                      0.99                      0.97   
18                      0.96                      0.99   
19                      0.79                      0.87   
20                      0.92                      0.95   
21                      0.81                      0.73   
22                      0.90                      0.89   
23                      0.89                      0.91   
24                      0.93                      0.86   
25                      0.97                      0.93   
26                      0.87                      0.90   
27                      0.93                      0.93   
28                      0.97                      0.90   
29                      0.97                      0.94   
...                      ...                       ...   
1334                    0.00                      0.00   
1335                    0.99                      0.97   
1336                    0.00                      0.00   
1337                    0.00                      0.00   
1338                    0.91                      0.92   
1339                    0.89                      0.90   
1340                    0.96                      0.96   
1341                    0.93                      0.92   
```

```
1342                    0.95                      0.93
1343                    0.93                      0.93
1344                    0.00                      0.00
1345                    0.77                      0.77
1346                    0.99                      0.99
1347                    0.85                      0.89
1348                    0.90                      0.95
1349                    0.80                      0.83
1350                    0.95                      0.84
1351                    0.00                      0.00
1352                    0.97                      0.92
1353                    0.99                      0.95
1354                    0.96                      0.98
1355                    0.95                      0.95
1356                    0.74                      0.83
1357                    0.78                      0.76
1358                    0.84                      0.88
1359                    0.94                      0.93
1360                    0.93                      0.90
1361                    0.97                      0.92
1362                    0.93                      0.91
1363                    0.87                      0.84

      Supportive Environment %  Effective School Leadership %  \
0                         0.86                           0.91
1                         0.97                           0.90
2                         0.82                           0.61
3                         0.82                           0.73
4                         0.87                           0.81
5                         0.95                           0.91
6                         0.95                           0.69
7                         0.88                           0.88
8                         0.85                           0.87
9                         0.90                           0.83
10                        0.91                           0.67
11                        0.99                           0.99
12                        0.77                           0.72
13                        0.81                           0.80
14                        0.94                           0.93
15                        0.92                           0.96
16                        0.97                           0.96
17                        0.95                           0.96
18                        0.96                           0.98
19                        0.82                           0.77
20                        0.92                           0.96
21                        0.85                           0.55
22                        0.88                           0.86
23                        0.95                           0.86
```

```
24                        0.94                           0.76
25                        0.97                           0.93
26                        0.95                           0.87
27                        0.98                           0.91
28                        0.94                           0.82
29                        0.99                           0.91
...                        ...                            ...
1334                      0.00                           0.00
1335                      0.97                           0.92
1336                      0.00                           0.00
1337                      0.00                           0.00
1338                      0.78                           0.86
1339                      0.93                           0.88
1340                      0.94                           0.96
1341                      0.88                           0.84
1342                      0.93                           0.90
1343                      0.79                           0.86
1344                      0.00                           0.00
1345                      0.87                           0.70
1346                      0.87                           0.94
1347                      0.83                           0.83
1348                      0.89                           0.88
1349                      0.81                           0.83
1350                      0.95                           0.80
1351                      0.00                           0.00
1352                      0.98                           0.84
1353                      0.98                           0.88
1354                      0.87                           0.91
1355                      0.96                           0.86
1356                      0.73                           0.77
1357                      0.78                           0.71
1358                      0.84                           0.85
1359                      0.94                           0.88
1360                      0.88                           0.88
1361                      0.89                           0.84
1362                      0.96                           0.89
1363                      0.84                           0.77

      Strong Family-Community Ties %  Trust %  Average ELA Proficiency  \
0                               0.85     0.94                     2.14
1                               0.86     0.94                     2.63
2                               0.80     0.79                     2.39
3                               0.89     0.88                     2.48
4                               0.89     0.93                     2.38
5                               0.88     0.97                     2.29
6                               0.87     0.78                     2.80
7                               0.79     0.94                     2.28
8                               0.83     0.93                     2.21
```

```
9                              0.89     0.95                     2.16
10                             0.83     0.85                     3.24
11                             0.92     0.99                     2.17
12                             0.76     0.87                     1.96
13                             0.74     0.87                     2.29
14                             0.97     0.96                     2.86
15                             0.86     0.96                     2.26
16                             0.93     0.94                     2.89
17                             0.95     0.99                     2.55
18                             0.96     0.98                     3.10
19                             0.76     0.88                     2.34
20                             0.88     0.95                     2.82
21                             0.81     0.74                     3.83
22                             0.83     0.94                     2.92
23                             0.87     0.94                     2.59
24                             0.83     0.89                     2.75
25                             0.96     0.95                     3.09
26                             0.92     0.90                     3.40
27                             0.93     0.95                     3.12
28                             0.92     0.89                     3.04
29                             0.96     0.96                     3.39
...                             ...      ...                      ...
1334                           0.00     0.00                     0.00
1335                           0.76     0.96                     2.99
1336                           0.00     0.00                     0.00
1337                           0.00     0.00                     0.00
1338                           0.83     0.90                     2.69
1339                           0.89     0.93                     2.57
1340                           0.91     0.98                     2.43
1341                           0.85     0.96                     2.21
1342                           0.84     0.95                     2.59
1343                           0.80     0.87                     2.55
1344                           0.00     0.00                     0.00
1345                           0.85     0.91                     2.30
1346                           0.92     0.97                     2.38
1347                           0.81     0.90                     2.23
1348                           0.90     0.94                     2.67
1349                           0.75     0.89                     2.44
1350                           0.83     0.91                     2.83
1351                           0.00     0.00                     0.00
1352                           0.93     0.97                     3.23
1353                           0.94     0.97                     3.16
1354                           0.90     0.95                     3.00
1355                           0.86     0.98                     3.03
1356                           0.86     0.91                     2.02
1357                           0.78     0.83                     2.39
1358                           0.87     0.91                     2.42
1359                           0.83     0.94                     2.48
```

```
1360                            0.88     0.93                     2.50
1361                            0.86     0.94                     2.77
1362                            0.91     0.95                     2.60
1363                            0.85     0.84                     2.74

      Average Math Proficiency  Count of Offers
0                         2.17              0.0
1                         2.98              0.0
2                         2.54              0.0
3                         2.47              5.0
4                         2.54              0.0
5                         2.48              0.0
6                         3.20              0.0
7                         2.73              0.0
8                         2.27              5.0
9                         2.31              0.0
10                        3.63             23.0
11                        2.32              5.0
12                        1.83              0.0
13                        2.00              5.0
14                        3.20              0.0
15                        2.20              5.0
16                        2.99              0.0
17                        2.68              0.0
18                        3.08              0.0
19                        2.48              5.0
20                        2.90              5.0
21                        4.03             91.0
22                        3.01             14.0
23                        3.14              0.0
24                        3.24              0.0
25                        3.41              0.0
26                        3.71              0.0
27                        3.35              0.0
28                        3.40              0.0
29                        3.65              0.0
...                        ...              ...
1334                      0.00              0.0
1335                      3.36              5.0
1336                      0.00              0.0
1337                      0.00              0.0
1338                      2.81              5.0
1339                      3.04              0.0
1340                      2.66              5.0
1341                      2.29              0.0
1342                      2.76              5.0
1343                      2.83              5.0
1344                      0.00              5.0
```

```
        1345                      2.46              5.0
        1346                      2.30              5.0
        1347                      2.48              5.0
        1348                      2.87              5.0
        1349                      2.52              5.0
        1350                      3.11              5.0
        1351                      0.00              5.0
        1352                      3.97              5.0
        1353                      3.91              6.0
        1354                      3.52              5.0
        1355                      3.18              5.0
        1356                      2.27              0.0
        1357                      2.50              5.0
        1358                      2.87              5.0
        1359                      2.60              0.0
        1360                      2.85              5.0
        1361                      3.09              5.0
        1362                      3.29              0.0
        1363                      3.19              0.0

        [1364 rows x 11 columns]
```

```
In [630]: #Check for  nulls if any
          benchmark.isnull().any()
```

```
Out[630]: Student Attendance Rate                       True
          Percent of Students Chronically Absent        True
          Rigorous Instruction %                        True
          Collaborative Teachers %                      True
          Supportive Environment %                      True
          Effective School Leadership %                 True
          Strong Family-Community Ties %                True
          Trust %                                       True
          Average ELA Proficiency                       True
          Average Math Proficiency                      True
          Count of Offers                               True
          dtype: bool
```

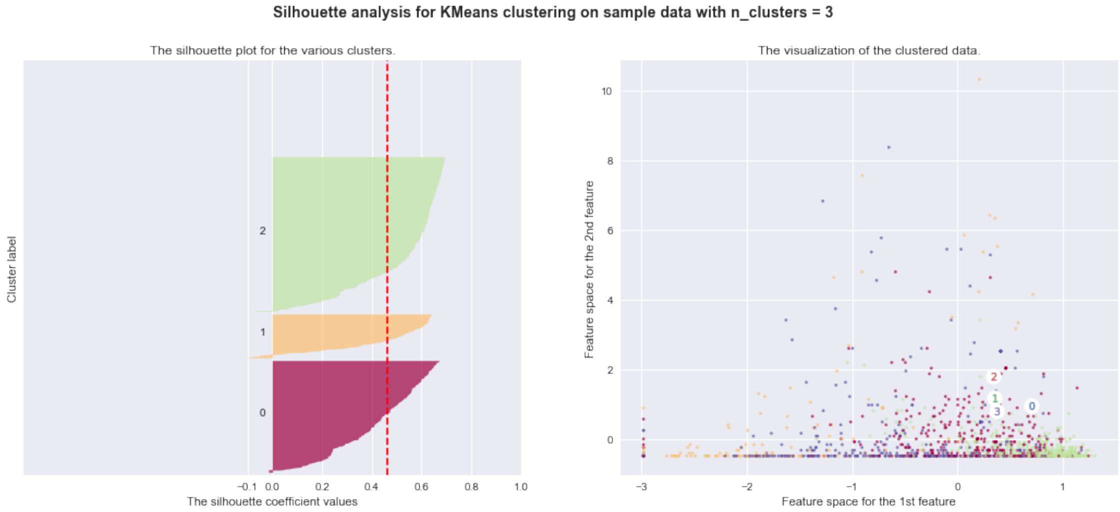
```
In [631]: #Replace nulls if any
          benchmark = benchmark.fillna(method='ffill')
```

```
In [632]: X1 = benchmark[[
           'Average ELA Proficiency',
           'Average Math Proficiency']]
```

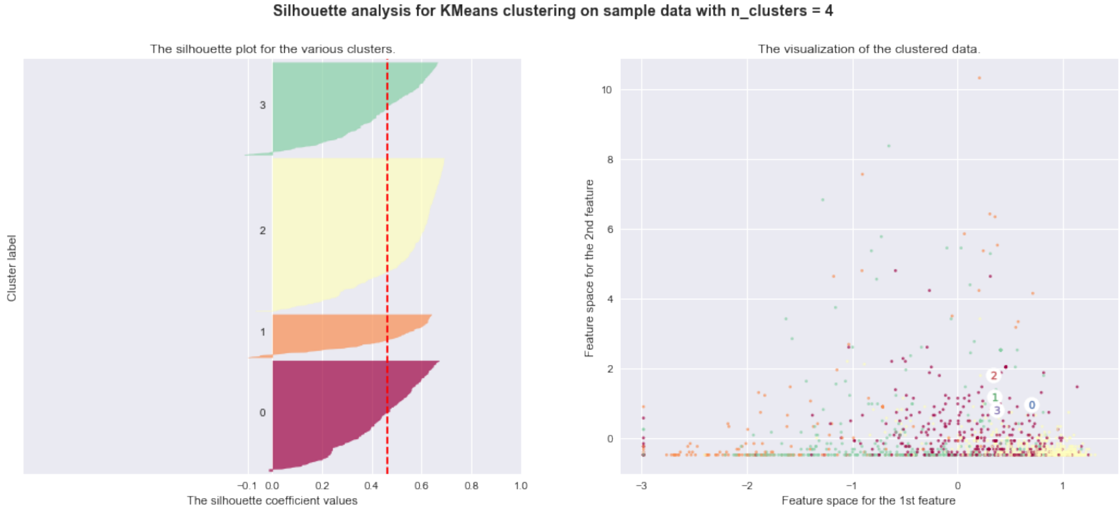
```
In [633]: from sklearn.preprocessing import StandardScaler
          scaler = StandardScaler()
          X1_scaled = scaler.fit_transform( X1 )
```

```
In [634]: %matplotlib inline

          cluster_range = range( 3, 9 ) #range excludes upper bound

          for n_clusters in cluster_range:
            fig, (ax1, ax2) = plt.subplots(1, 2)
            fig.set_size_inches(18, 7)

            # The 1st subplot is the silhouette plot
            # The silhouette coefficient can range from -1, 1
            ax1.set_xlim([-1, 1])

           # The (n_clusters+1)*10 is for inserting blank space between silhouette
            # plots of individual clusters, to demarcate them clearly.
            ax1.set_ylim([0, len(X1_scaled) + (n_clusters + 1) * 10])

            cluster_labels = clusterer.fit_predict( X1_scaled )

            # The silhouette_score gives the average value for all the samples.
            # This gives a perspective into the density and separation of the formed
            # clusters
            silhouette_avg = silhouette_score(X1_scaled, cluster_labels)
            print("For n_clusters =", n_clusters,
                  "The average silhouette_score is :", silhouette_avg)

            # Compute the silhouette scores for each sample
            sample_silhouette_values = silhouette_samples(X1_scaled, cluster_labels)

            y_lower = 10
            for i in range(n_clusters):
                # Aggregate the silhouette scores for samples belonging to
                # cluster i, and sort them
                ith_cluster_silhouette_values = \
                    sample_silhouette_values[cluster_labels == i]

                ith_cluster_silhouette_values.sort()

                size_cluster_i = ith_cluster_silhouette_values.shape[0]
                y_upper = y_lower + size_cluster_i

                cmap = cm.get_cmap("Spectral")
                color = cmap(float(i) / n_clusters)

                #color = cm.spectral(float(i) / n_clusters)
                ax1.fill_betweenx(np.arange(y_lower, y_upper),
                                  0, ith_cluster_silhouette_values,
                                  facecolor=color, edgecolor=color, alpha=0.7)
                # Label the silhouette plots with their cluster numbers at the middle
                ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))

                # Compute the new y_lower for next plot
                y_lower = y_upper + 10
            ax1.set_title("The silhouette plot for the various clusters.")
            ax1.set_xlabel("The silhouette coefficient values")
            ax1.set_ylabel("Cluster label")

            # The vertical line for average silhoutte score of all the values
            ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

            ax1.set_yticks([])  # Clear the yaxis labels / ticks
            ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])

            # 2nd Plot showing the actual clusters formed
            #cmap = cm.get_cmap("Spectral")
            colors = cmap(cluster_labels.astype(float) / n_clusters)
            #colors = cm.spectral(cluster_labels.astype(float) / n_clusters)
            ax2.scatter(X_scaled[:, 0], X_scaled[:, 1], marker='.', s=30, lw=0, alpha=0.7,
                        c=colors)

            # Labeling the clusters
            centers = kmeans.cluster_centers_
            #centers = clusterer.cluster_centers_
            # Draw white circles at cluster centers
            ax2.scatter(centers[:, 0], centers[:, 1],
                        marker='o', c="white", alpha=1, s=200)

            for i, c in enumerate(centers):
                ax2.scatter(c[0], c[1], marker='$%d$' % i, alpha=1, s=50)

            ax2.set_title("The visualization of the clustered data.")
            ax2.set_xlabel("Feature space for the 1st feature")
            ax2.set_ylabel("Feature space for the 2nd feature")

            plt.suptitle(("Silhouette analysis for KMeans clustering on sample data "
                          "with n_clusters = %d" % n_clusters),
                         fontsize=14, fontweight='bold')

            plt.show();
```

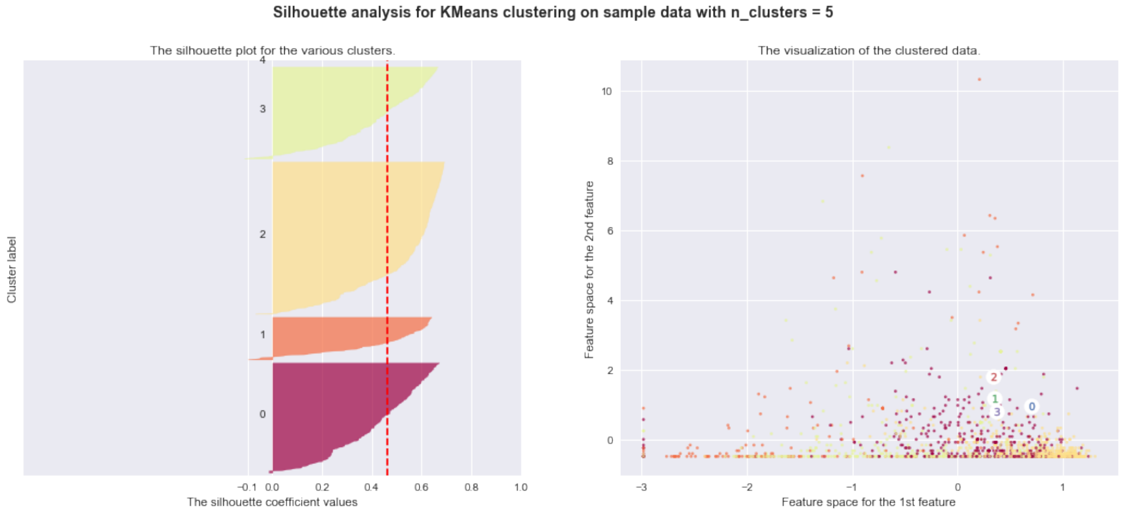
```
For n_clusters = 3 The average silhouette_score is : 0.462324434771
```



```
For n_clusters = 4 The average silhouette_score is : 0.462324434771
```



```
For n_clusters = 5 The average silhouette_score is : 0.462324434771
```

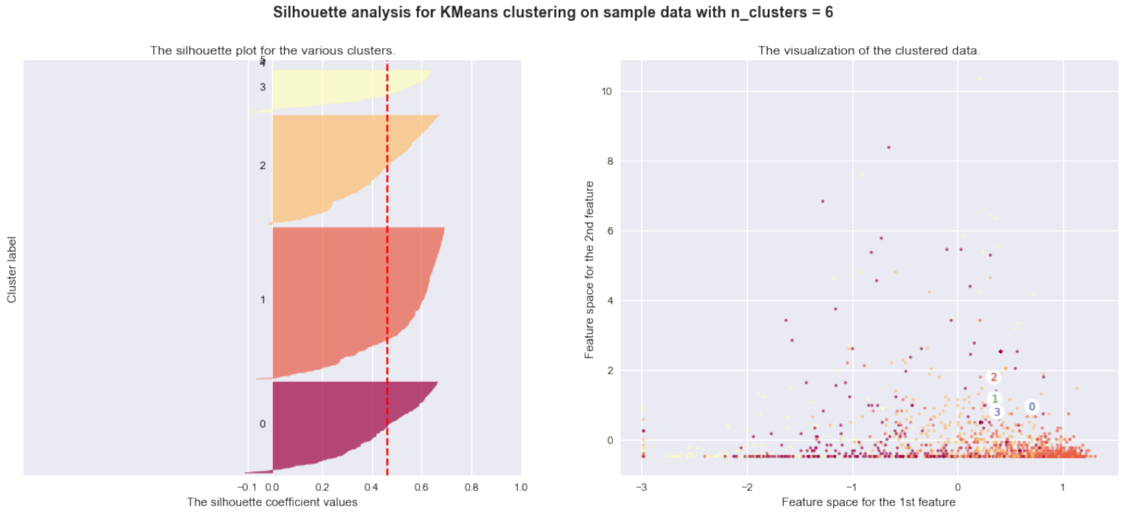


Silhouette analysis for KMeans clustering on sample data with n_clusters = 5

The silhouette plot for the various clusters.

The visualization of the clustered data.

```
For n_clusters = 6 The average silhouette_score is : 0.462324434771
```

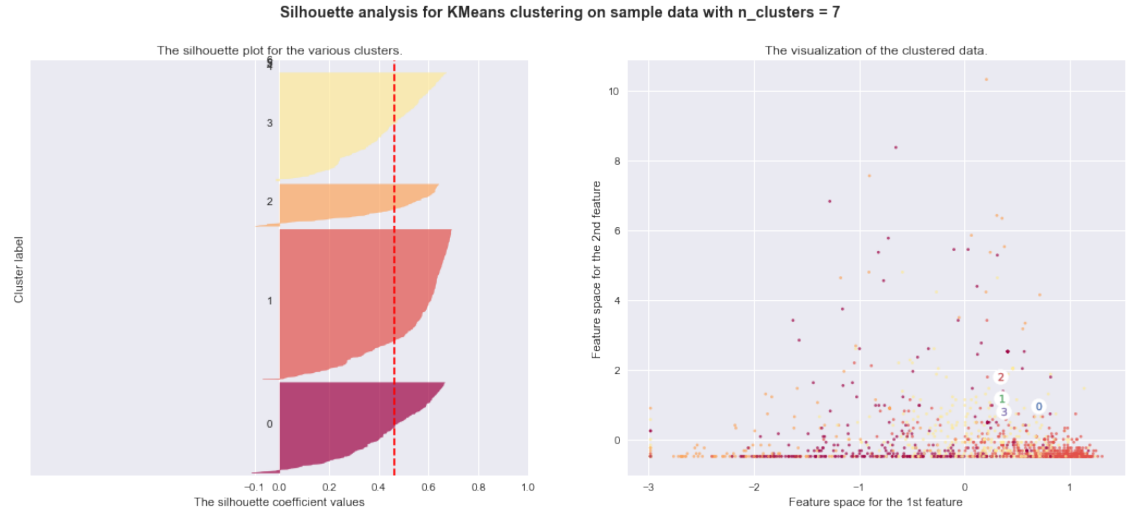


Silhouette analysis for KMeans clustering on sample data with n_clusters = 6

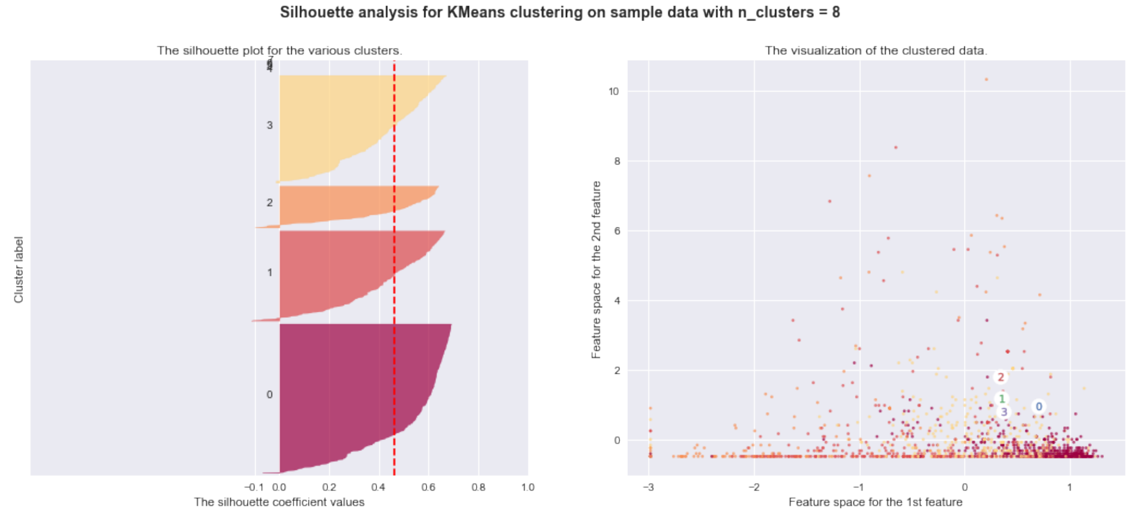
The silhouette plot for the various clusters.

The visualization of the clustered data.

```
For n_clusters = 7 The average silhouette_score is : 0.462324434771
```



```
For n_clusters = 8 The average silhouette_score is : 0.462324434771
```



1.4 Results

- Section **??**: Evaluation of Analysis
- Section **??**: Validation based final inference

Section 1.5.3: Back to Top

1.4.1 Model Evaluation and Validation

Silhouette score is the best for clusters =3 in our model using GMM technique. For K-means it shows best as clusters =4, and close to 0 for clusters =3, but the visuals clearly show that this is a result of overlapping. Given that Kmeans suffer from spherical clustering, this seems to validate the score as 3. When we compare with our benchmark model too, we get the silhouette score as same for clusters =3 to clusters =8 but the visuals clearly show the homogenous clustering for clusters =3. Thus, we can acertain that clusters =3 is the optimal solution and comparable to benchmark model.

1.4.2 Justification

For the benchmark model as well, the best scores and homogenous clusters are for clusters = 3. It also is important to note that the original benchmark solution had ascertained cluster number as 3 though not using the techniques we employ. Thus, we can see that the clustering algorithms applied using our methods provides an equal measure of accuracy and hence, we can consider out model to be a good model.

1.5 Conclusion

Section 1.5.3: Back to Top

1.5.1 Reflection

In this project, the intent was to categorize the schools based on the variables, finding the optimal number of segments along the way. We did this using Kmeans and GMM with silhouette score as the basis of evaluation and comparison against benchmark model. One aspect that was interesting and challenging was determining the best silhouette score since we had very close values for the benchmark model for clusters numbering from 3 to 8. However, it is the visuals that help to pinpoint the right size. Again, for our model, GMM matched but not K-means; where it seemed like the right size did not match since 4 seemed to have best silhouette score value. Here too, the visuals provided a better comprehension with the overlapping effect evident (score of 0 again indicates close clusters which holds true with spherical contraints of k-means).

1.5.2 Improvement

One aspect of the analysis that can be improved is to use the analysis is to evaluate the performance using different seed values. Seed values were used in kmeans clustering and feature relevance prediction.

1.5.3 References

Please find below referenced articles and links in addition to those added in relevant sections. - http://www.awesomestats.in/python-cluster-validation/ - https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py - http://scikit-learn.org/stable/modules/clustering.html - http://scikit-learn.org/stable/modules/generated/sklearn.metrics - https://en.wikipedia.org/wiki/Elbow_method_(clustering)

Section 1.5.3