Research on how New York City schools' demographic influence the students chronical absent rate

Project ONE

Introduction

PASSNYC is a not-for-profit organization dedicated to promoting educational opportunities for New York City's talented and underserved students. In 2016, PASSNYC collected 1273 schools' data to identify students within New York City's under-performing school districts, and aims to increase the diversity of students taking the Specialized High School Admissions Test (SHSAT). The `2016 School Explorer` dataset contains 1273 New York schools and each school's specific demographics like absent rate, races distribution, location and Collaborative Teachers Rating etc.

Using PASSNYC dataset, we can measure schools' performance in education. This project aims to analyze how New York City schools' location and Economic Need Index influence the students chronically absent rate. In general, we will select 7 areas with the greatest number of schools of New York City and then look at the Economic Need Index within 7 areas to analyze their influence on absent rate. Based on the analysis result, the PASSNYC can identify those targeted areas and implement related policy to decrease the absent rate of those targeted areas.It will promote better reallocation of educational resources.

In the following analysis, we will use the summary table, boxplot, bar graph, and correlation table to better illustrate the factors that affect the absence of the student.

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as matplot
import seaborn as sns
import plotly.graph_objs as go
import geopandas as gpd

from plotly.offline import iplot
from shapely.geometry import Point

%matplotlib inline

import qeds
qeds.themes.mpl_style();

import warnings
warnings.filterwarnings("ignore")
```

Read data

First, we read the data in python and return the first 5 rows of data frame. We have the information with 1272 rows and 161 columns.

In [2]:
```
nyc=r'C:\Users\WLJY8\Desktop\Courses\YEAR 4\ECO225\Project 1\2016 School Explorer.csv'
df = pd.DataFrame(pd.read_csv(nyc))
```

In [3]:
```
df.info()
```

```
<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
```

In [4]:
```
df.iloc[:, 3:26] .head(2)
```

Out[4]:

	School Name	SED Code	Location Code	District	Latitude	Longitude	Address (Full)	City	Zip	Grades	..
0	P.S. 015 ROBERTO CLEMENTE	310100010015	01M015	1	40.721834	-73.978766	333 E 4TH ST NEW YORK, NY 10009	NEW YORK	10009	PK,0K,01,02,03,04,05	..
1	P.S. 019 ASHER LEVY	310100010019	01M019	1	40.729892	-73.984231	185 1ST AVE NEW YORK, NY 10003	NEW YORK	10003	PK,0K,01,02,03,04,05	..

2 rows × 23 columns

Data Clean

`Percent of Students Chronically Absent` is the dependent value that resprsent each school's absence rate. In this column, we have 25 `NaN` missing data which is small group compare to total of 1272 information. In order to keep the dataset integrity, drop all `NaN`. Moreover,we use `Absent rate` to instead `Percent of Students Chronically Absent` for simplifying.

In [5]:
```
df['Percent of Students Chronically Absent'].isnull().sum()
```

Out[5]:
```
25
```

In [6]:
```
df = df.dropna(subset=['Percent of Students Chronically Absent'])
df.rename(columns={'Percent of Students Chronically Absent':'Absent Rate'},
                  inplace=True)
```

After reading the dataset, we noticed that the value of `'Absent Rate'`, and independent variable - `'Percent Asian','Percent Black','Percent Hispanic','Percent White','Collaborative Teachers %'` are recorded in percentage and stored as `objects`. We preprocess those data and create a function to convert the percentage to a fraction(e.g.8% into 0.08). Thus, we keep those point as numerical values `float64`.

In [7]:
```
df['Absent Rate'].dtype
```

Out[7]:
```
dtype('O')
```

In [8]:
```
df['Percent Asian'].dtype
```

Out[8]:
```
dtype('O')
```

In [9]:
```
def p2f(x):
    return float(x.strip('%'))/100

df['Absent Rate']=df['Absent Rate'].astype(str).apply(p2f)
df['Percent Asian']=df['Percent Asian'].astype(str).apply(p2f)
df['Percent Black']=df['Percent Black'].astype(str).apply(p2f)
df['Percent Hispanic']=df['Percent Hispanic'].astype(str).apply(p2f)
df['Percent White']=df['Percent White'].astype(str).apply(p2f)
df['Collaborative Teachers %']=df['Collaborative Teachers %'].astype(str).apply(p2f)
```

In this case our dependent variable is `absent rate` and independent variables are `City` and `Economic Need Index`. Since our original dataframe contains 161 columns, we need to reduce it into a smaller dataframe.

In [10]:
```
df2=df[['School Name','Absent Rate','City','Longitude','Latitude','Percent Asian','Percent Black',
        'Percent Hispanic','Percent White','Economic Need Index','Collaborative Teachers %','Zip']]
```

There are more than 40 areas of New York City; it is unnecessary to research all areas' influence on `absent rate`. We use the bar graph to show the number of schools in each city. We found that `Brooklyn`, `Bronx` and `New York` have more than 200 schools. However, some area like: `ROOSEVELT ISLAND`, `BROAD CHANNEL`, `SOUTH RICHMOND HILL` and `DOUGLASTON` only have 1-3 schools.

In [11]:
```
def plot_city_hist(df, title_str):
    layout = go.Layout(
        title=title_str,
        xaxis=dict(
            title='City',
            titlefont=dict(
                family='Arial, sans-serif',
                size=12,
                color='black'
            ),
            showticklabels=True,
            tickangle=315,
            tickfont=dict(
                size=10,
                color='grey'
            )
        )
    )
    data = [go.Histogram(x=df['City'])]
    fig = go.Figure(data=data, layout=layout)
    return fig

fig = plot_city_hist(df2, 'City Wise School Distribution')

fig.update_layout(
    yaxis_title="Count",font=dict(
        family='Arial, sans-serif',
        size=10,
        color='black'
    ))
iplot(fig)
```

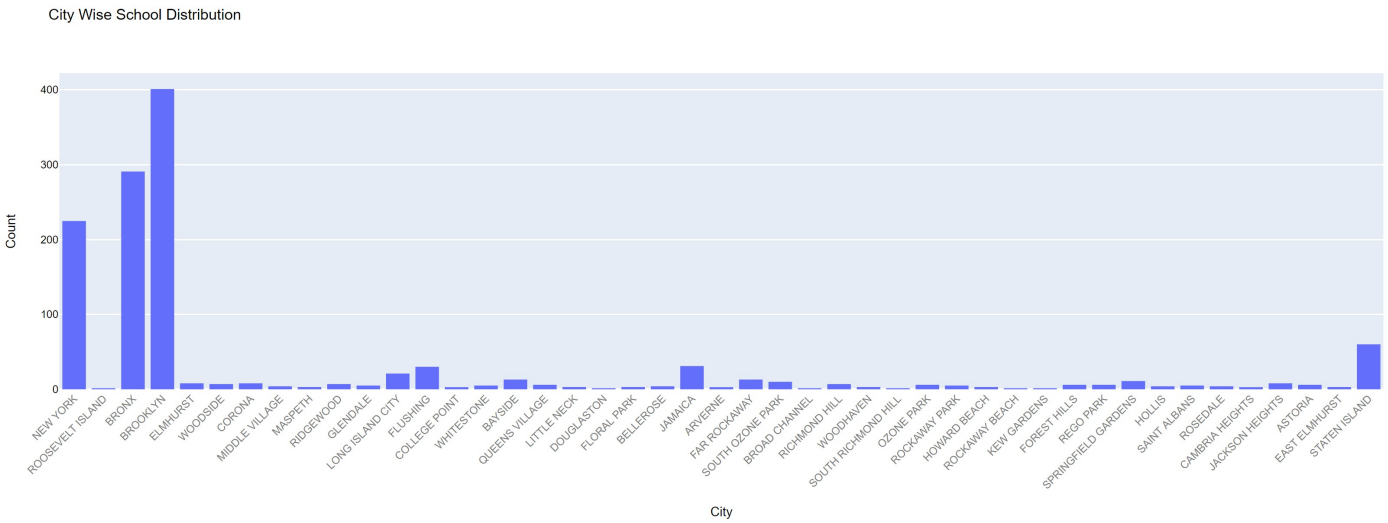


Figure 1-1: Bar graph: City Wise School Distribution

We will drop the area with less than 20 schools since their sample dont have any representativeness. And then, we only keep the information of schools within those 7 areas (`BROOKLYN` , `BRONX` , `NEW YORK` , `STATEN ISLAND` , `JAMAICA` , `FLUSHING` , `LONG ISLAND CITY`). Those areas occupy about 85% of schools in NYC.

In [12]:
```
cities = ['BROOKLYN','BRONX','NEW YORK','STATEN ISLAND','JAMAICA','FLUSHING','LONG ISLAND CITY']
df2=df2[df2['City'].isin(cities)]
df2['City'].value_counts()
```

Out[12]:
```
BROOKLYN            401
BRONX               291
NEW YORK            225
STATEN ISLAND        60
JAMAICA              31
FLUSHING             30
LONG ISLAND CITY     21
Name: City, dtype: int64
```

We check the variables type in new dataframe and make sure that they all satisfy our requirment.

In [13]:
```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1059 entries, 0 to 1271
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   School Name               1059 non-null   object
 1   Absent Rate               1059 non-null   float64
 2   City                      1059 non-null   object
 3   Longitude                 1059 non-null   float64
 4   Latitude                  1059 non-null   float64
 5   Percent Asian             1059 non-null   float64
 6   Percent Black             1059 non-null   float64
 7   Percent Hispanic          1059 non-null   float64
 8   Percent White             1059 non-null   float64
 9   Economic Need Index       1059 non-null   float64
 10  Collaborative Teachers %  1059 non-null   float64
 11  Zip                       1059 non-null   int64
dtypes: float64(9), int64(1), object(2)
memory usage: 107.6+ KB
```

The following is the sample from my new dataframe.

In [14]:
```
df2.head()
```

Out[14]:

	School Name	Absent Rate	City	Longitude	Latitude	Percent Asian	Percent Black	Percent Hispanic	Percent White	Economic Need Index	Collaborative Teachers %
0	P.S. 015 ROBERTO CLEMENTE	0.18	NEW YORK	-73.978766	40.721834	0.05	0.32	0.60	0.01	0.919	0.94
1	P.S. 019 ASHER LEVY	0.30	NEW YORK	-73.984231	40.729892	0.10	0.20	0.63	0.06	0.641	0.96
2	P.S. 020 ANNA SILVER	0.20	NEW YORK	-73.986315	40.721274	0.35	0.08	0.49	0.04	0.744	0.77
3	P.S. 034 FRANKLIN D. ROOSEVELT	0.28	NEW YORK	-73.975043	40.726147	0.05	0.29	0.63	0.04	0.860	0.78
4	THE STAR ACADEMY - P.S.63	0.23	NEW YORK	-73.986360	40.724404	0.04	0.20	0.65	0.10	0.730	0.88

Variables Analysis

First, we compute the descriptive statistics to summarize the central tendency, dispersion and shape of `Absent Rate` distribution. For all schools within 7 area, the mean(the average of the data)and median of the absent rate is 0.23. The variable scale from 0 to 1.

In [15]:
```
df2['Absent Rate'].describe()
```

Out[15]:
```
count    1059.000000
mean        0.228876
std         0.144385
min         0.000000
25%         0.120000
50%         0.220000
75%         0.320000
max         1.000000
Name: Absent Rate, dtype: float64
```

And we also draw a boxplot graph which provides a graphical summary of the distribution of a sample. Most of values are concentrated in small rate.

In [16]:
```
plt.boxplot(df2['Absent Rate'],vert = False)
plt.title('Boxplot graph')
plt.xlabel('Range')
plt.ylabel('Absent Rate')
#plt.text(-0.1,0,"Figure 1-2:Boxplot graph")
```

Out[16]:
```
Text(0, 0.5, 'Absent Rate')
```

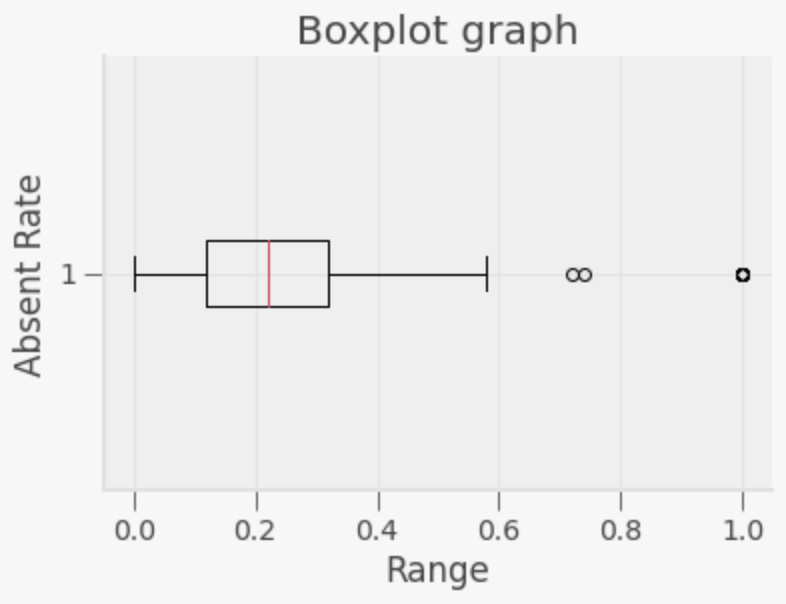


Figure 1-2: Boxplot graph for absent rate

Then, we use the same method to find the `Economic Need Index` distribution. For all schools within 7 area, the mean(the average of the data) of the Economic Need Index is 0.699212 with median 0.76.

In [17]:
```
df2['Economic Need Index'].describe()
```

Out[17]:
```
count    1059.000000
mean        0.699212
std         0.203915
min         0.049000
25%         0.600000
50%         0.760000
75%         0.852000
max         0.957000
Name: Economic Need Index, dtype: float64
```

Then, we draw a boxplot of Economic Need Index to check the graphical summary of the distribution of a sample. The distribution of Economic Need Index is not uniform, and there exists several extremely small values.

In [18]:
```
plt.boxplot(df2['Economic Need Index'],vert = False)
plt.title('Boxplot graph')
plt.xlabel('Range')
plt.ylabel('Economic Need Index')
#plt.text(-0.1,0,"Figure 1-2:Boxplot graph")
```

Out[18]:
```
Text(0, 0.5, 'Economic Need Index')
```

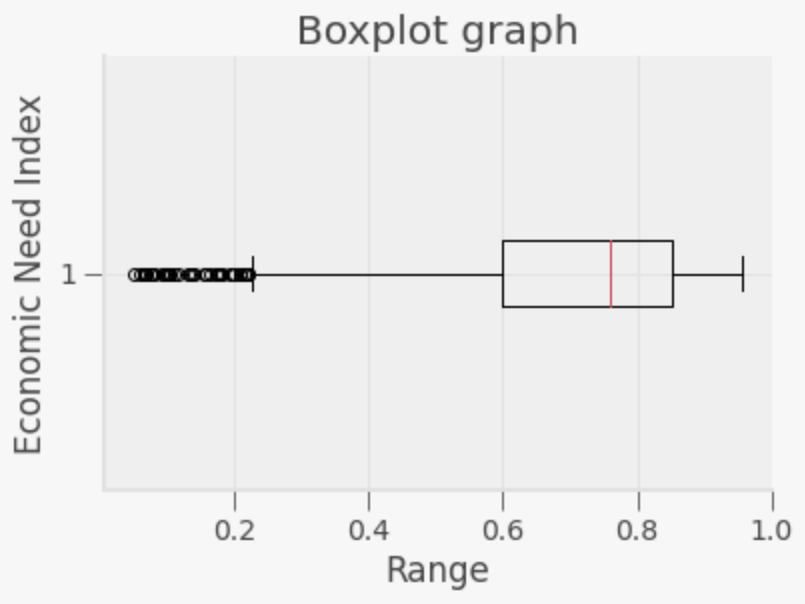


Figure 1-2: Boxplot graph for Economic Need Index

Final, we can look at the correlation between each area's Economic Need Index and each area's absent rates. For example, we noticed that in Staten Island, the Economic Need Index is highly positively correlated with absent rate compare to other areas (0.810831). And the correlation value in Bronx and Flushing is much smaller around 0.45.

In [19]:
```
df3 = df2[['Economic Need Index','Absent Rate','City']]
df3.groupby('City').corr()
```

Out[19]:

City		Economic Need Index	Absent Rate
BRONX	**Economic Need Index**	1.000000	0.431347
	Absent Rate	0.431347	1.000000
BROOKLYN	**Economic Need Index**	1.000000	0.568199
	Absent Rate	0.568199	1.000000
FLUSHING	**Economic Need Index**	1.000000	0.454123
	Absent Rate	0.454123	1.000000
JAMAICA	**Economic Need Index**	1.000000	0.665658
	Absent Rate	0.665658	1.000000
LONG ISLAND CITY	**Economic Need Index**	1.000000	0.701863
	Absent Rate	0.701863	1.000000
NEW YORK	**Economic Need Index**	1.000000	0.453115
	Absent Rate	0.453115	1.000000
STATEN ISLAND	**Economic Need Index**	1.000000	0.810831
	Absent Rate	0.810831	1.000000

Conclusion

In project 1, the data cleansing process improves our data quality, increases overall productivity and helps analysis accuracy. We prepare our data by removing some NA; checking the type of data and transfer them into

numeric or character; and creating a new subset which contains variables we will need in the analysis.

We generate the descriptive statistics and boxplot for our dependent variable `Absent Rate` and one independent variable `Economic Need Index` to see their distribution. Later, we add the correlation table of `Economic Need Index` and `Absent Rate` which grouped by `City` . It helps us to find the correlation coefficients between variables.We find that the correlation value is varies and it is highest in STATEN ISLAND and lowest in BRONX.

In next step, we repeat the steps and measure more variables effect on absent rate.

Project 2

THE MESSAGE AND UPDATED INTRODUCTION INFORMATION

The primary purpose of this research is to focus on how schools' demographics influence the absent rate.

In Project 1, we clean the data and select the variables we may use in following analysis. We shown the summary statistic table of one potential independent variable `Economic Need Index` and our dependent variable `Absent Rate` .

In Project 2, we will figure out how different areas influence the absent rate.Moreover, we will focus on the student `race percent` and other educational indicator like `Collaborative Teachers %` .I would like to use some bar graph, scatter plot,correlation table and map to research on how those independent variables related to the `absent rate` .

We expect the `race black` and `race hispanic` has the most significant absent rate since it is more likely for them to lack pre-school eduation and face greater discriminationin in comparison to other races in general due to historical reason. And we also predict that the area with more schools may have larger absent rate since the sample size for those areas is larger (i.e.more students)

Analysis

In [20]:
```
plt.figure(figsize = [8,6])

temp = sns.distplot(df2['Absent Rate'], kde=False)
temp = plt.title('Distribution of schools based on chronically absent students')
temp = plt.xlabel("Absent Rate")
temp = plt.ylabel("Count")
```

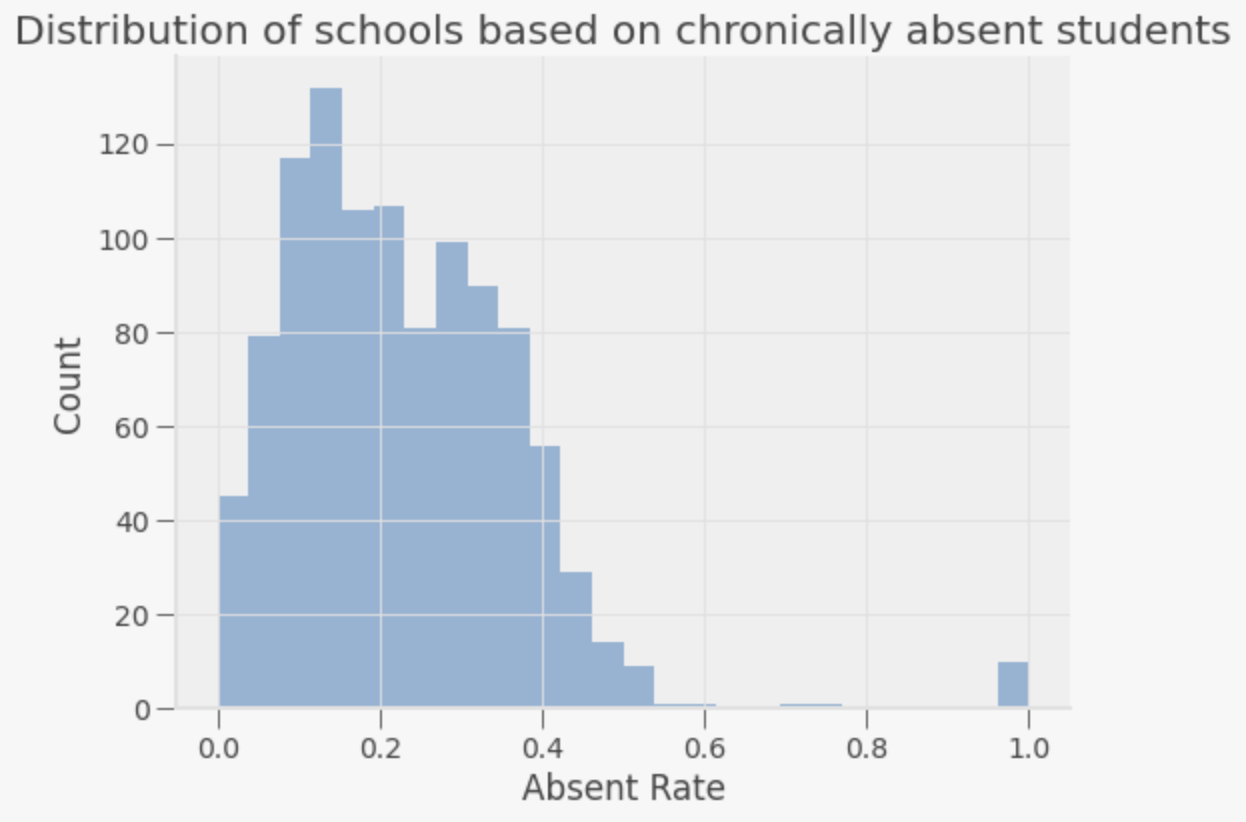


Figure 2-1: Histogram graph - the Distribution of schools based on chronically absent students

We draw the histogram to see the shape of the `absent rate` distribution. The data is right-skewed distribution and it is more concentrate on 0 to 0.4.

In [21]:
```
df_citygroup = df2.groupby('City')['Absent Rate']
df_city_sta = df_citygroup.agg([np.min, np.max, np.mean,np.median, np.std])
df_city_sta
```

Out[21]:

City	amin	amax	mean	median	std
BRONX	0.00	1.00	0.278144	0.280	0.121737
BROOKLYN	0.00	1.00	0.221546	0.200	0.139872
FLUSHING	0.02	0.29	0.099000	0.090	0.062442
JAMAICA	0.01	0.40	0.214194	0.220	0.098920
LONG ISLAND CITY	0.00	0.45	0.152381	0.130	0.102757
NEW YORK	0.00	1.00	0.217111	0.180	0.181362
STATEN ISLAND	0.05	0.41	0.182333	0.155	0.084620

We use the aggregate and group-by function to see 7 areas schools' performance on the absent rate. From the summary statistics table, we noticed that Flushing has the smallest aggregate max absent rate and smallest mean. The Bronx has the largest mean and median. Long Island City and Staten Island's performance are quite similar, and they have the a small rate range.

In [22]:
```
plt.figure(figsize=(20,8))
sns.barplot(x='City',y='Absent Rate',data=df2,ci=None,color=(0.2, 0.4, 0.6, 0.6))
plt.title('Average Absent Rate of 7 Main Areas')
```

Out[22]:
```
Text(0.5, 1.0, 'Average Absent Rate of 7 Main Areas')
```

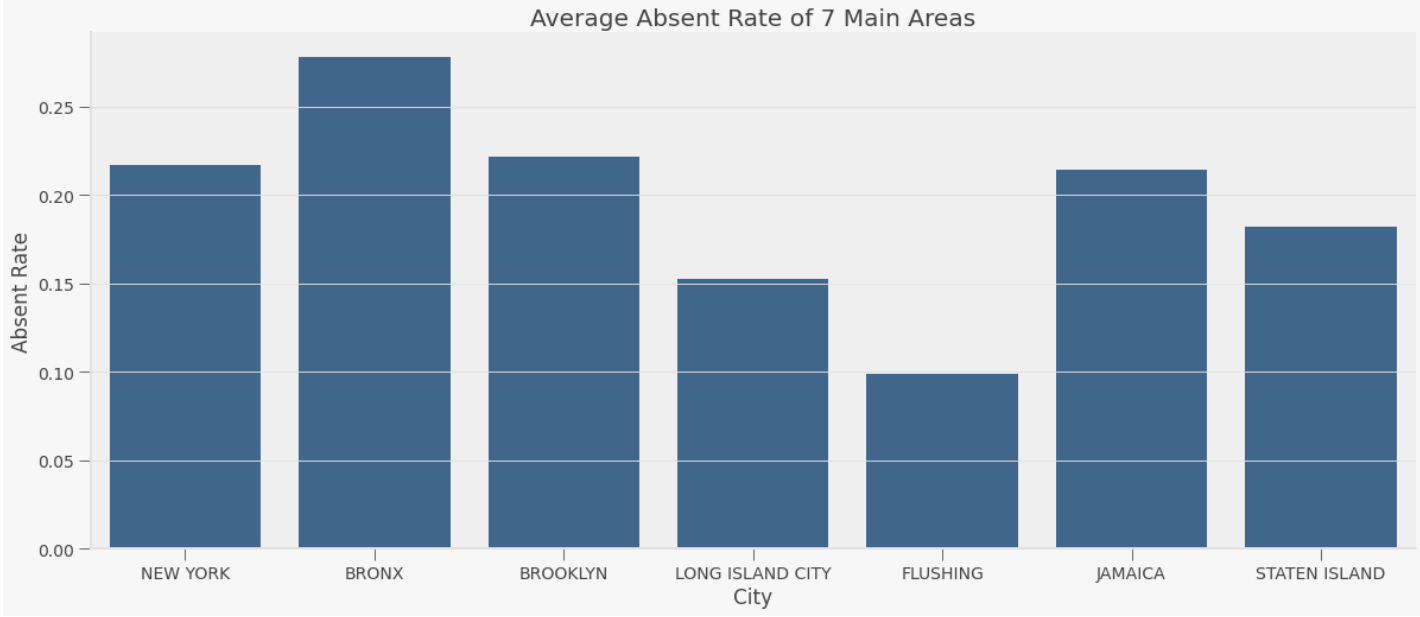


Figure 2-2: Bar graph - Average Absent Rate of 7 Main Areas

The bar graph visualizes the absent student rate is highest in Bronx (with mean 0.28), led by Brooklyn (with mean 0.22) and New York (with mean 0.21). Jamaica (with mean 0.20), Staten Island (with mean 0.18). Flushing (with mean 0.01) and Long Island City (with mean 0.13) have a relatively low absent rate.

In [23]:
```
nyc_boros = gpd.read_file(gpd.datasets.get_path("nybb"))
#boro_locations = gpd.tools.geocode(boros.BoroName)

df["Coordinates"] = list(zip(df.Longitude, df.Latitude))
df["Coordinates"] = df["Coordinates"].apply(Point)
gdf_ca = gpd.GeoDataFrame(df, geometry="Coordinates")
gdf_ca
fig, gax = plt.subplots(figsize=(20,8))
nyc_boros.to_crs("EPSG:4326").plot(ax=gax, color="white", edgecolor="k")
gdf_ca.plot(ax=gax,edgecolor="face",column='Absent Rate',legend=True,
            cmap="RdPu",s=16,vmin=0,vmax=1)
gax.annotate("Absent rate",xy=(0.8,0.06),xycoords='figure fraction')

gax.set_xlabel('longitude')
gax.set_ylabel('latitude')
plt.title("Absent rate (scale from 0-1) of NYC schools")


df_new = pd.DataFrame({
'Boroughs' : ['Bronx', 'Queens', 'New York' ,'Staten Island', 'Brooklyn',
              'Jamaica','Long Island City','FLUSHING' ],
'Latitude': [40.837048, 40.742054, 40.785091, 40.579021, 40.650002,
             40.694854,40.75855,40.721159],
'Longitude': [-73.865433, -73.769417, -73.968285, -74.151535,-73.949997,
              -73.806837,-73.939237,-73.823164] })

df_new["Coordinates"] = list(zip(df_new.Longitude, df_new.Latitude))
df_new["Coordinates"] = df_new["Coordinates"].apply(Point)
gdf = gpd.GeoDataFrame(df_new, geometry="Coordinates")

gdf.plot(ax=gax, color='blue', alpha = 0.6)

for x, y, label in zip(gdf['Coordinates'].x, gdf['Coordinates'].y, gdf['Boroughs']):
    gax.annotate(label, xy=(x,y), xytext=(4,4), textcoords='offset points')

plt.show()
```

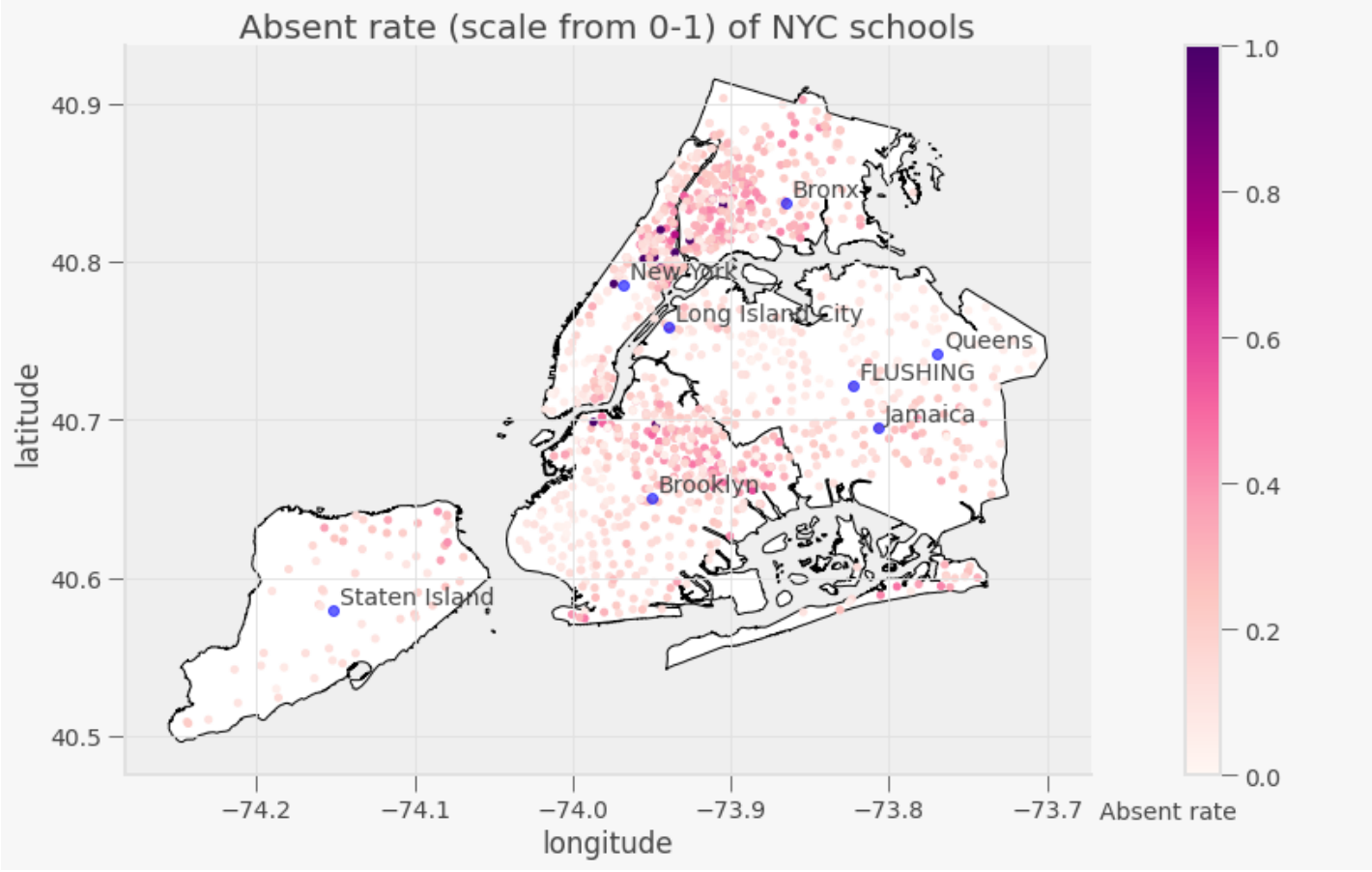


Figure 2-3: Map - Absent rate (scale from 0-1) of NYC schools

In the map, the pink points reflect the absent rate of schools in NYC. The darker points appear, the greater the absent rate. I label the 7 main areas' location by blue points. As we discussed in the last part, `New York` . `Bronx` and `Brooklyn` have a serious problem with students' absent rates. Moreover, those 3 areas have the most significant number of schools. It matches our assumption that the area with more schools will have the larger absent rate.

Next, we focus on how races influence the absent race. We create a correlation matrix to find the correlation between variables.

In [24]:
```
plt.figure(figsize=(11,8))
corrMatrix = df2.corr()
sns.heatmap(corrMatrix, annot=True)
plt.title('Correlation Matrix')
plt.show()
```

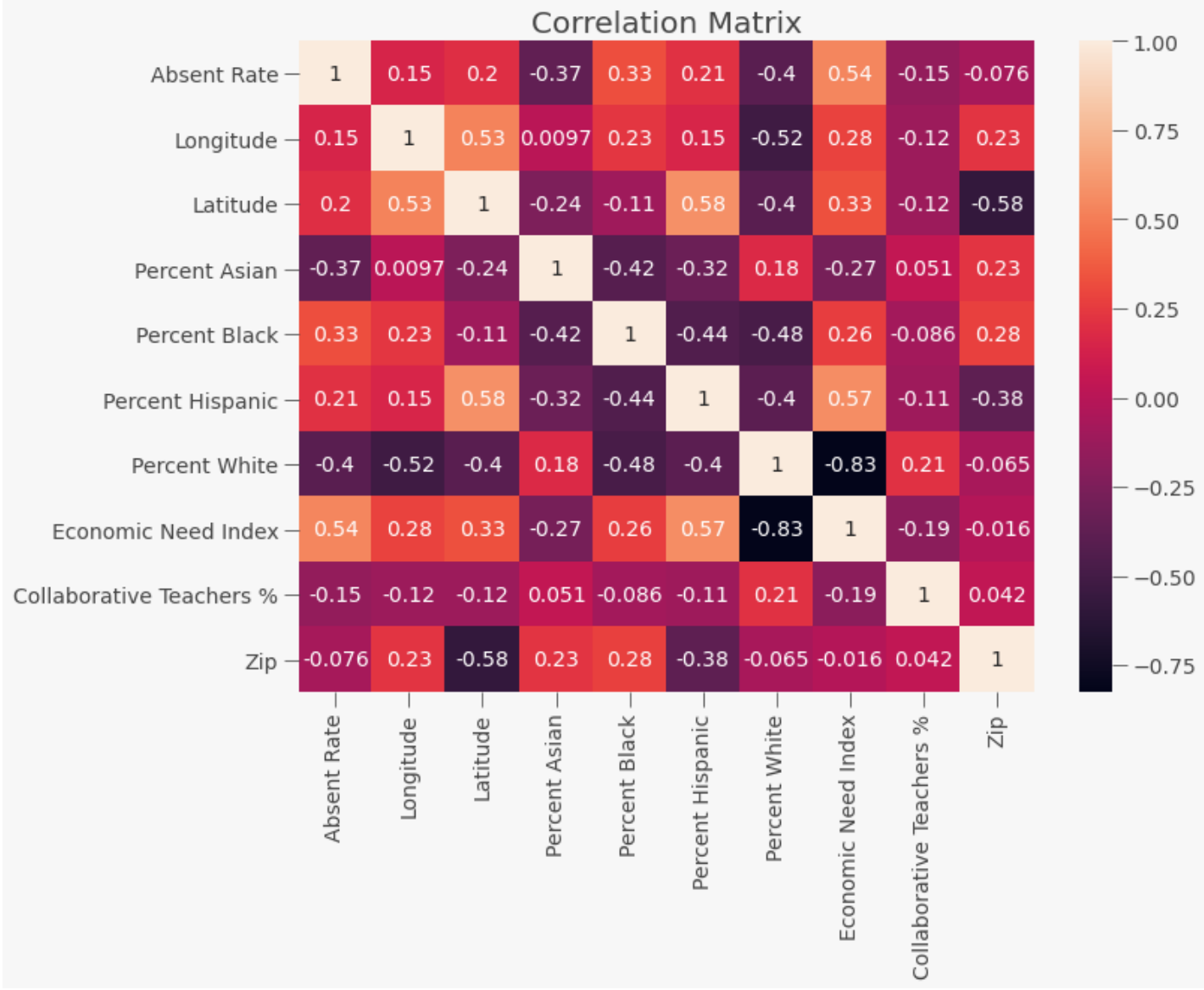


Figure 2-4: Heat map - Correlation Matrix

From the correlation table, we noticed that `Absent rate` is negatively correlated with `Percent Asian` and `Percent White`. However, `Absent rate` is positively correlated with `Percent Black` and `Percent Hispanic`. In other words, schools with large percent of Black and Hispanic have greater absent rates.

Moreover, we find that the `Collaborative Teachers %` are negatively correlated with absent rate(-0.15) and `Economic Need Index` are strongly positively correlated with absent rate(0.54).

Next steps, I want to find the relationship between each school's `Economic Need Index` and `Absent rate`. Thus, I draw a regression plot;

In [25]:
```
f, axes = plt.subplots(figsize=(8, 5))

f1=sns.regplot('Economic Need Index','Absent Rate',df2,
               scatter_kws={"color": "blue"}, line_kws={"color": "red"})
axes.set_title('Economic Need Index vs Absent Rate')
```

Out[25]:
```
Text(0.5, 1.0, 'Economic Need Index vs Absent Rate')
```

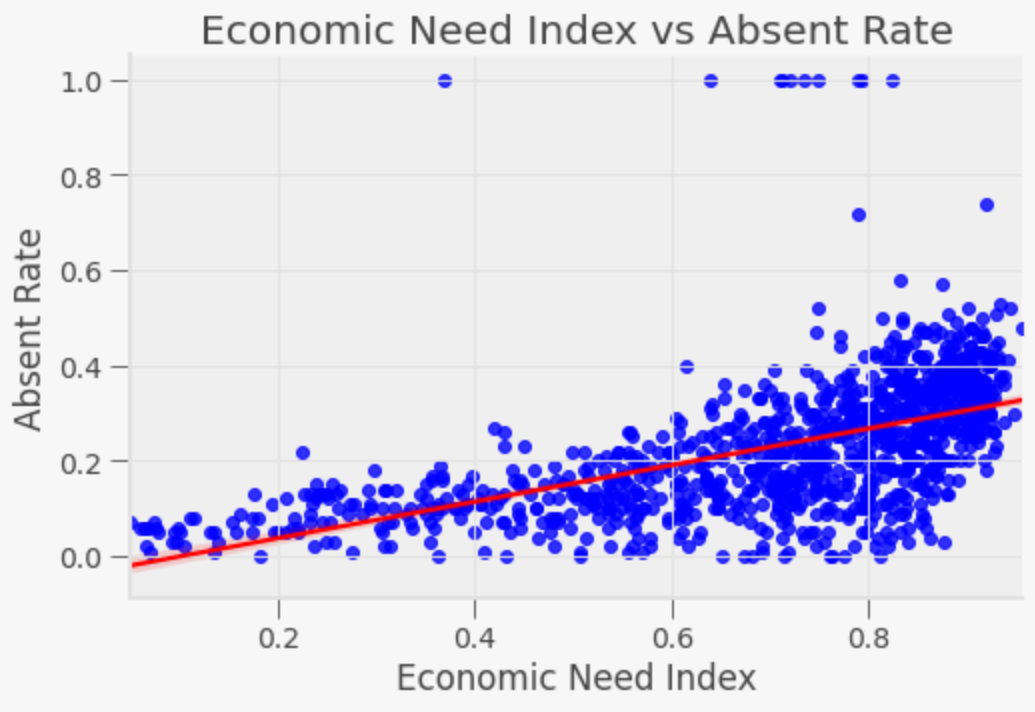


Figure 2-6: Regression plot - Economic Need Index vs Absent Rate

The positive slope indicates that the schools with high economic need index are likely to have large absent rate.This result matches our correlation table's result.

In [26]:
```
f, axes = plt.subplots(2, 2, figsize=(19, 9), sharex=True)

f1=sns.regplot('Economic Need Index','Percent Asian',df2,ax=axes[0,0],
               scatter_kws={"color": "black"}, line_kws={"color": "red"})
f2=sns.regplot('Economic Need Index','Percent White',df2,ax=axes[0,1],
               scatter_kws={"color": "green"}, line_kws={"color": "red"})
f2=sns.regplot('Economic Need Index','Percent Black',df2,ax=axes[1,0],
               scatter_kws={"color": "purple"}, line_kws={"color": "red"})
f2=sns.regplot('Economic Need Index','Percent Hispanic',df2,ax=axes[1,1],
               scatter_kws={"color": "brown"}, line_kws={"color": "red"})


axes[0,0].set_title('Economic Need Index(scale from 0-1)and Percent Asian')
axes[0,1].set_title('Economic Need Index(scale from 0-1)and Percent White')
axes[1,0].set_title('Economic Need Index(scale from 0-1)and Percent Black')
axes[1,1].set_title('Economic Need Index(scale from 0-1)and Percent Hispanic')
```

Out[26]:
```
Text(0.5, 1.0, 'Economic Need Index(scale from 0-1)and Percent Hispanic')
```

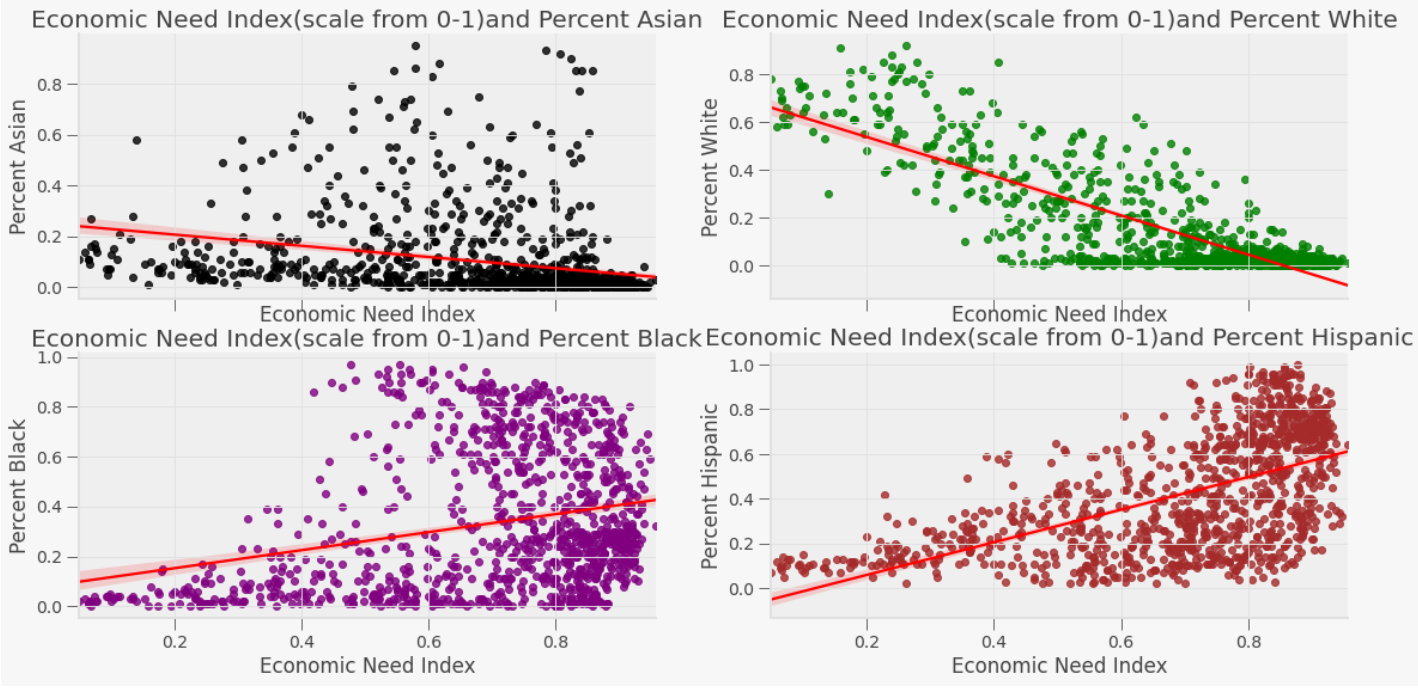


Figure 2-7: Regression plot - Economic Need Index vs Race

The above graph shows the Student Race vs Economic Need Index. Schools with large percent of White or Asian Students tend to have lower Economic Need Index. Oppositely, schools with higher percentage of Blacks or Hispanic Students tend to have higher Economic Need Index.

Then, we use the summary statistics to find the distribution of `Collaborative Teachers %` points.

In [27]:
```
df2['Collaborative Teachers %'].describe()
```

Out[27]:
```
count    1059.00000
mean        0.88406
std         0.07509
min         0.00000
25%         0.84000
50%         0.90000
75%         0.94000
max         1.00000
Name: Collaborative Teachers %, dtype: float64
```

The points are more concentrated in large values. We have a large mean and median. This observation a good indicator because it shows that most schools have a great percentage of collaborative teachers, which help student engagement.

In [28]:
```
f, axes = plt.subplots(figsize=(8, 5))

f1=sns.regplot('Collaborative Teachers %','Absent Rate',df2,
               scatter_kws={"color": "blue"}, line_kws={"color": "red"})
axes.set_title('Collaborative Teachers % vs Absent Rate')
```

Out[28]:
```
Text(0.5, 1.0, 'Collaborative Teachers % vs Absent Rate')
```

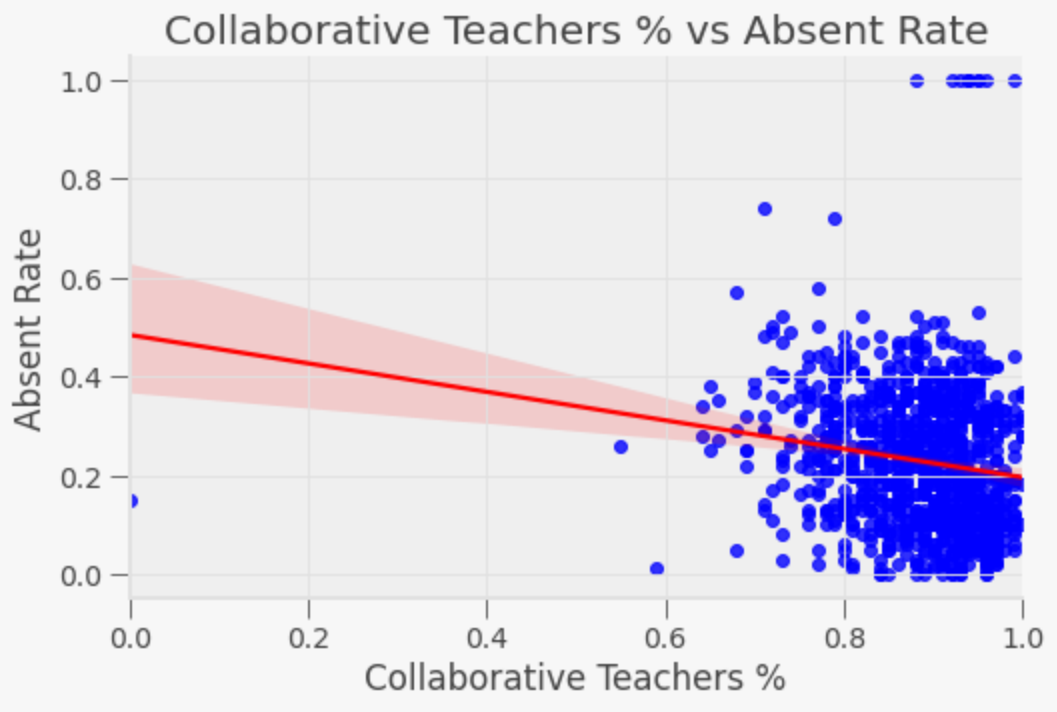


Figure 2-8: Regression plot - Collaborative Teachers % vs Absent Rate

Moreover, we fit the regression plot and see how the Collaborative Teachers % influence the Absent Rate. There exists a negatively linear realtionship between Collaborative Teachers % and the Absent Rate.This result is similar to our correlation table result.

Conclusion

In project 2 analysis, we can conclude that the areas with a large number of schools have a greater absent rate. Moreover, the race White and Asian are negatively correlated with absent rate, which is good. However, the race Black and Hispanic are positively correlated with absent rate, which brings up a concern. Both results match our intuition about the possible answer to our question.Moreover, the more ELL students appears will lead to higher choroncially absent rate.

Moreover, we find that the `Collaborative Teachers %` are negatively correlated with absent rate. And we also do some in-depth research about how economic need index related to absent rate/different race.

Based on the result I listed in the analysis part, in order to decrease the absent rate, the bureau of education should supervise the schools' teaching quality and provide more economic support for those schools with high economic need index, especially for those schools with large percent of race Black and Hispanic students.

Project 3

Introduction

Our paper's main question is to analyze how the school demographic influences the students chronically absent rate. From project one and project two, we use `2016 School Explorer` to see how the school's location, student race, collaborative teachers % and economic need index affect the absent rate based on the summary statistics, visualized graphs and map. In project 3, I will add more information to the dataset using web scraping. After merging these two datasets, we will build a new relationship between population and absent rate. The hypothesis is that if an area with a large population, that area's absence rate will be much greater.

Web Scraping

1) The data `New York Zip Codes by Population` is from the US Census. It included 1753 New York zip codes and collected accurate population information based on the zip code in New York City. The population is another key indicator of school demographics which can measure an area's prosperity and infrastructure level. Thus,it has impact on students' behavior. Based on this data, I can research how population influences the absence rate in New York City. We can measure the Average Population of 7 Main Areas. Furthermore, we may use it to investigate the potential correlation between the economic need index and population.
2) We can scrape the data from https://www.newyork-demographics.com/zip_codes_by_population.
3) Since my original data and the new data contain columns with zip code information, I will merge two datasets by `zip` and add the population information into the original data. In the analysis part, I will build a histogram and regression graph to find the relationship between population and absent rate.

In this case, we don't need to run our program over time to generate data. Scape the data from the website is not a complicated coding. We need to be careful about selecting the tables' rows and strip the unnecessary blank space and comments.
However, the scraped table is not prepared for merging. Cleansing the dataframe is the most challenge part. For example, there have more than 2 zip codes connected by `and` in one row; then, we need to separate them, rematch the population data and then create a new data frame to store the information. It better for us learn more about how to efficiently use the python code techniques to achieve the result we want.

Scrape data

First, we import the packages that we need to use for web scarping.

In [29]:
```
import re
import requests
import pandas as pd
from bs4 import BeautifulSoup
```

The second step is parsing HTML and accessing different elements. The response content can be passed to a `BeautifulSoup()` method to obtain a soup object which is structured. We can also uncomment the soup_object to explore the schema and understand the structure of the web page, which helps us extract relevant data from the web page.
In our case, the data is enclosed in the `<table>` HTML tag with the class name `ranklist`.
Meanwhile, every row of data is enclosed under a `<tr>` HTML tag. All these row values can be extracted into a list of values by finding the `<tr>` values from our newly created soup object `data_table`.
In this case, the last row of the table records the reference info, which will not be used in the merging process, so we will not scrape it.

In [30]:
```
web_url = 'https://www.newyork-demographics.com/zip_codes_by_population'
```

In [31]:
```
response = requests.get(web_url)
```

In [32]:
```
soup_object = BeautifulSoup(response.content)
# Uncomment the below line and look into the contents of soup_object
# soup_object
data_table = soup_object.find_all('table', 'ranklist')[0]
# Uncomment the below line and look into the contents of data_table
#data_table
all_values = data_table.find_all('tr')
#We don't need the last row for the website table
all_values= all_values[:-1]
```

The first element of the list contains the column names 'Rank, Zip Code and Population'. The following elements of the list have soup objects which contain the population data. This data can be extracted in a loop since all the

list elements' structure is the same. When we are using the `for loop` , We need to use `strip()` to delete the empty space in front of/behind the value and the unnecessary info at the same time.

In [33]:
```
# Create an empty dataframe
Zip_Codes_by_Population = pd.DataFrame(columns = ['Rank', 'Zip Code', 'Population'])
ix = 0 # Initialise index to zero

for row in all_values[1:]:
    values = row.find_all('td') # Extract all elements with tag <td>
    # Pick only the text part from the <td> tag
    # we use text.strip() to delete the empty space in front of/behind the value
    # we use rstrip('\n          \n            TIE') to delete
    # the unnecessary info in Rank
    Rank = values[0].text.strip().rstrip('\n          \n            TIE')
    Zip_Code = values[1].text.strip()
    Population = values[2].text.strip()
    Zip_Codes_by_Population.loc[ix] = [Rank, Zip_Code, Population]
    # Store it in the dataframe as a row
    ix += 1

# Print the first 10 rows of the dataframe
Zip_Codes_by_Population
```

Out[33]:

	Rank	Zip Code	Population
0	1	11368	112,088
1	2	11385	107,796
2	3	11211	103,123
3	4	11208	101,313
4	5	10467	101,255
...	...	...	...
1579	1,580	13826	15
1580	1,581	13353	13
1581	1,582	14854	12
1582	1,583	13352	11
1583	1,584	12007	9

1584 rows × 3 columns

Recoding process

The scraped data is not perfect. We noticed that there have more than two zip codes in one row. Our next step is to build a new data frame that each zip code matches one population number based on the scraping table.

In [34]:
```
# Select the rows which zipcode contains 'and'
df_has_and = Zip_Codes_by_Population[Zip_Codes_by_Population['Zip Code'].str.contains('and')]
# Select the rows which zipcode do not contain 'and'
df_no_and = Zip_Codes_by_Population[~Zip_Codes_by_Population['Zip Code'].str.contains('and')]

# create a new df that has single zipcode in one row (change from the rows with 'and')
idx = 0
new_df = pd.DataFrame(columns=Zip_Codes_by_Population.columns)
for index, row in df_has_and.iterrows():
    split_zip_codes = filter(lambda code: code!= '',
                             [code.strip() for code in
```

```
                                re.split(r',|and',row['Zip Code'])])
    for code in split_zip_codes:
        new_df.loc[idx]=[row['Rank'],code,row['Population']]
        idx+=1
new_df

# recombine the the 2 dataframes
zipdata=df_no_and.append(new_df)
```

Moreover, we need to check the type of data points and transfer them into numeric.

In [35]:
```
#rename the column
zipdata.rename(columns={'Zip Code': 'Zip'}, inplace=True)
#change the type of values in each columns
zipdata['Rank'] = zipdata['Rank'].apply(lambda x: x.replace(',',''))
zipdata['Rank']=zipdata['Rank'].astype(int)
zipdata['Zip']=zipdata['Zip'].astype(int)
zipdata['Population'] = zipdata['Population'].apply(lambda x: x.replace(',',''))
zipdata['Population']=zipdata['Population'].astype(int)

zipdata.sort_values(by='Rank', ascending=True)
zipdata
```

Out[35]:

	Rank	Zip	Population
0	1	11368	112088
1	2	11385	107796
2	3	11211	103123
3	4	11208	101313
4	5	10467	101255
...	...	...	...
310	1569	11931	39
311	1574	13692	30
312	1574	14168	30
313	1576	12490	27
314	1576	13860	27

1752 rows × 3 columns

The last step is to store the data frame as a csv file. Pandas has a `to_csv` method which can be used to save the data into the file.

In [36]:
```
zipdata.to_csv('zipdata.csv', index=False) # convert dataframe into csv file
```

Merge Data

Since both data has `Zip` column, we can easily merge the new scraped data with our original data by using function `merge()`.

In [37]:
```
df_combine=pd.merge(df2, zipdata, on="Zip",how="left")
df_combine.head()
```

Out[37]:

	School Name	Absent Rate	City	Longitude	Latitude	Percent Asian	Percent Black	Percent Hispanic	Percent White	Economic Need Index	Collaborative Teachers %
0	P.S. 015 ROBERTO CLEMENTE	0.18	NEW YORK	-73.978766	40.721834	0.05	0.32	0.60	0.01	0.919	0.94
1	P.S. 019 ASHER LEVY	0.30	NEW YORK	-73.984231	40.729892	0.10	0.20	0.63	0.06	0.641	0.96
2	P.S. 020 ANNA SILVER	0.20	NEW YORK	-73.986315	40.721274	0.35	0.08	0.49	0.04	0.744	0.77
3	P.S. 034 FRANKLIN D. ROOSEVELT	0.28	NEW YORK	-73.975043	40.726147	0.05	0.29	0.63	0.04	0.860	0.78
4	THE STAR ACADEMY - P.S.63	0.23	NEW YORK	-73.986360	40.724404	0.04	0.20	0.65	0.10	0.730	0.88

In [38]:
```
df_combine['Population'].dtypes
```

Out[38]:
```
dtype('float64')
```

Analysis

In project 2, we noticed that the mean of absent student rate is highest in the Bronx, followed by Brooklyn, New York, Jamaica and Staten Island. And the absence rate is the smallest in Flushing and Long Island City.

In [39]:
```
plt.figure(figsize=(20,8))
sns.barplot(x='City',y='Population',data=df_combine,ci=None,color=(0.2, 0.4, 0.6, 0.6))
plt.title('Average Population of 7 Main Areas')
```

Out[39]:
```
Text(0.5, 1.0, 'Average Population of 7 Main Areas')
```

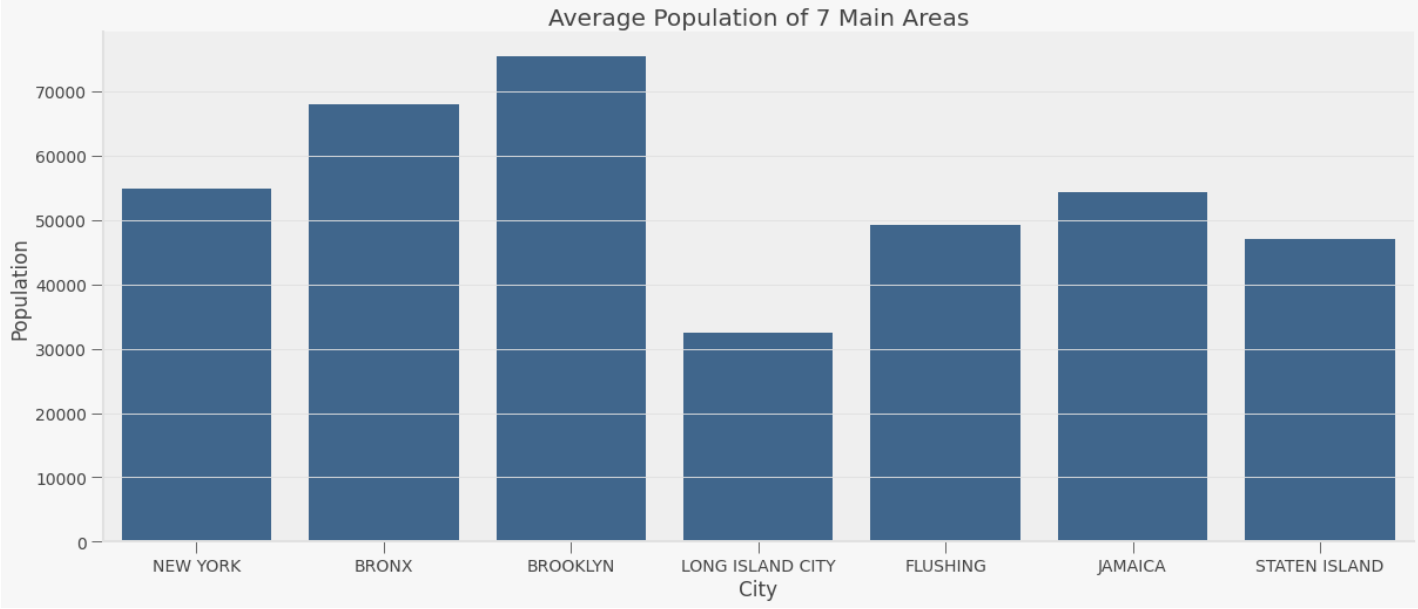


Figure 3-1: Bargraph - Average Population of 7 Main Areas

When we look that the mean of the population in each city, we found that Brooklyn, Bronx, New York and Jamaica are still in the lead place. LongIsland City has the smallest value. Thus, there may have some potential connections between absent rate and population.

In [40]:
```
df_test = pd.DataFrame(df_combine,columns=['Absent Rate','Population','Economic Need Index'])
corrMatrix = df_test.corr()
print (corrMatrix)
```

	Absent Rate	Population	Economic Need Index
Absent Rate	1.000000	0.075855	0.541120
Population	0.075855	1.000000	0.231684
Economic Need Index	0.541120	0.231684	1.000000

Then we run a correlation table to find the relevance between absent rate, population and economic need index. It is clear to see that there exists a positive correlation between absence rate and population. However, this relationship is relatively weak since the value(0.071402) is less than 0.1.
The correaltion between Population and Economic Index Rate is also positive with value of 0.226890.The population has stronger impact on Economic Need Index.

In [41]:
```
f, axes = plt.subplots(figsize=(8, 5))

f1=sns.regplot('Population','Absent Rate',df_combine,
               scatter_kws={'s':30,"color": "black"}, line_kws={"color": "red"})
axes.set_title('Population vs Absent Rate')
```

Out[41]:
```
Text(0.5, 1.0, 'Population vs Absent Rate')
```

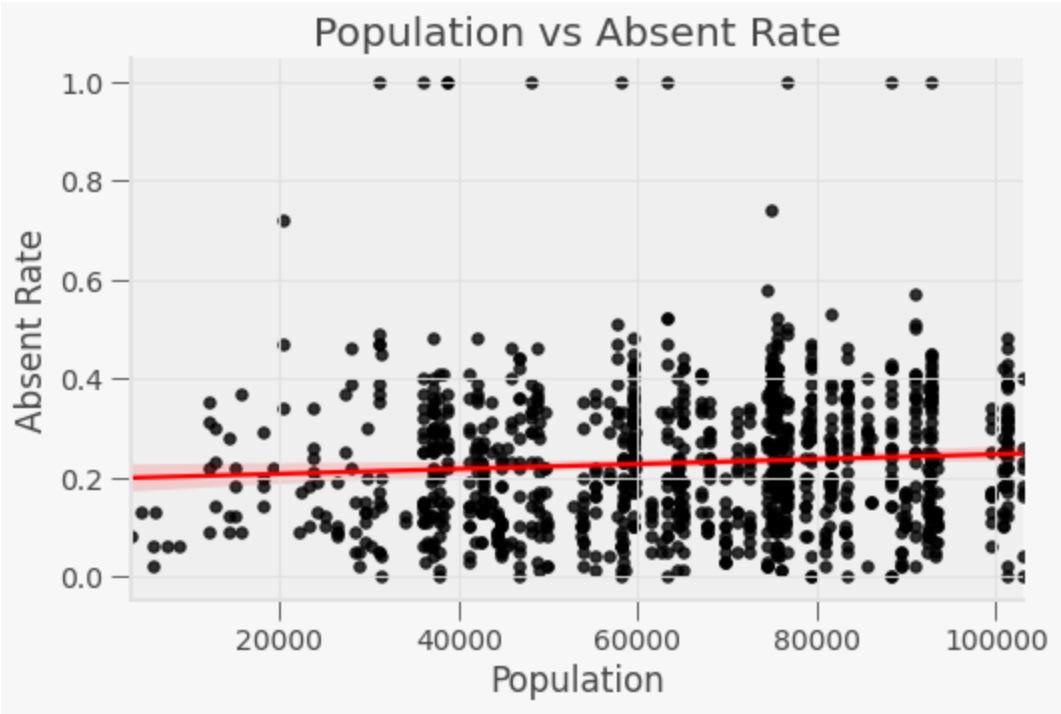


Figure 3-2: Regression plot-Population vs Absent Rate

The following graph shows the Population vs Absent Rate. We have a gradual slope, which proves that the population's influence on the absent rate is week. Thus, our hypothesis is not exactly right.

In [42]:
```
f, axes = plt.subplots(figsize=(8, 5))

f2=sns.regplot('Population','Economic Need Index',df_combine,
               scatter_kws={'s':30,"color": "black"}, line_kws={"color": "red"})
axes.set_title('Economic Need Index vs Population')
```

Out[42]:
```
Text(0.5, 1.0, 'Economic Need Index vs Population')
```

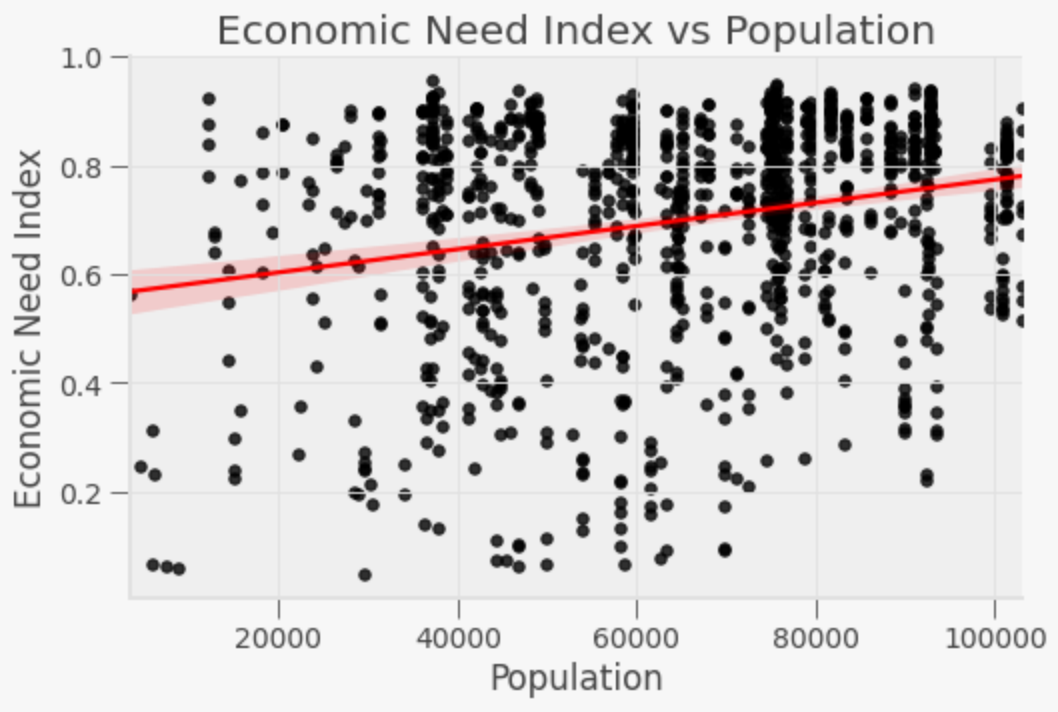


Figure 3-3: Regression plot-Economic Need Index vs Population

The above graph shows the Population vs Economic Need Index. We have a much steeper slope, which proves that the population's influence on the Economic Need Index is apparent. Since from Project 2, we know that the Economic Need Index and Absent Rate are highly positively related. Thus, I conclude that the population's direct influence on the absent rate is insignificant. However, its may exists some potential indirect effects (through impact the Economic Need Index first, then influence Absent Rate).

Conclusion and Summary

For our findings so far, in project one and project two, we noticed that the areas with a large number of schools have a greater absence rate. And the absent rate is positively correlated with race Black and Hispanic and negatively correlated with race White and Asian.

In project 3, we use web scraping to get more useful information (i.e. population) to our dataset. From the above analysis, We find that the direct effects of population on absence rate are inapparent. Thus, I do not support the hypothesis that if an area with a large population, that area's absence rate will be much greater. Nevertheless, there may exist the endogeneity problem with Population and Economic Need Index. The population may indirectly affect the absent rate, which acts based on the population's effects on the Economic Need Index. In other words, the increase in population leads to the rise in Economic Need Index, and then the increase in Economic Need Index causes an increase in the absent rate.

Final Project

Introduction(Updated from all previous projects)

Nowadays, more and more students in schools are increasingly impacted by chronic absent rate. Chronic absenteeism is defined as student miss at least 15 days of school in a year—are at serious risk of falling behind in school. Student chronic absent rate is an important an important measure of educational success. (Brendan Bartanen, 2020) For instance, we can use it to predict the graduation rates, school readiness and enrollment number. According to the 2015-16 Civil Rights Data Collection, over 7 million students, corresponding to 16 percent of the student population—or about 1 in 6 students, missed 15 or more school days in 2015-16. (U.S.

Department of Ed¬ucation, 2019). In other words, there were about more than 100 million school days lost, and this excessive amount of absenteeism phenomenon was a nationwide crisis.

While in New York City, the rate of chronic absenteeism, around 26 percent, is much higher compare to U.S. average rate. (Tina Posterli,2018) It is worth reviewing the causes for chronic absenteeism. Based on this, we can help all students have a better chance of reaching their full potential, and the government and the schools can know how to reduce the students' long-term chronic absent rate. In existing literature review, researchers mainly focus on how chronic absence influence the Student Achievement. Students who are repeatedly absent from school, caused the missing of important learning and developmental opportunities, are more likely to have negative consequences on their future outcomes such as difficulties in academic achievement, learning, sociability, and mental health. (London, Sanchez, & Castrechini, 2016) However, we cannot find any research that builds a statistical model to measure the level of influence of each independent variable on the absent rate.

Our research will mainly research how New York City schools' demographics influence students' chronically absent rate. We can use those variables to answer three key questions: (1) What percentage of students are chronically absent in 2016? (2) How each school environment and characteristics variables influence the chronically absent rate? (3) How much of those independent variables impact the chronic absent rate?

In the following analysis, we will use the summary table, boxplot, bar graph, and correlation table to illustrate better the factors that affect the student's absence and visualize our initial findings. Then, we will build the statistical model based on the variables we have in previous part and see at what level those independent variables impact the chronic absent rate. In this part, firstly, we will build a model to find how the percent of chronically absent rate varies with one specific variable (the Economic Need Index) while holding other factors stays the same. Considering Economic Need Index as one main effect, we will add other independent variables like Races, Population, and Collaborative Teachers % into the model.

In general, the ideal model is adding two characteristics variables (e.g., Race, Population), and two environment variables (e.g., Collaborative Teachers %, Economic Need Index). And we may adjust our model based on the model selection results. We will use the R-squared statistic to compare regression models. A larger R-squared value model means that a larger percentage of the variation in the dependent variable that is explained by independent variables. Thus, we expect that the R-square value increases as more variables add to the model, which better predicts the chronically absent rate. If the R-squared value is the same for the two models, we will consider finding the model with the lowest value of the Akaike Information Criterion (AIC) and Schwarz' Bayesian Information Criterion (BIC). AIC and BIC are both penalized-likelihood criteria.

Analysis

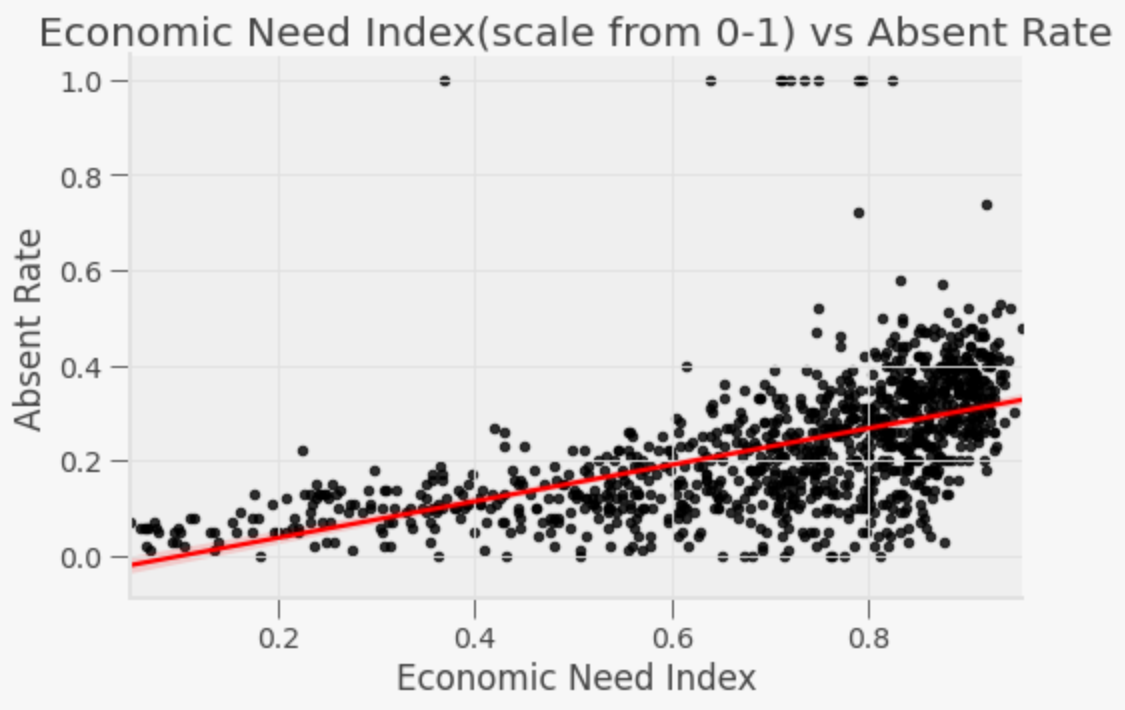
In [43]:
```
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.iolib.summary2 import summary_col
from linearmodels.iv import IV2SLS
```

First, we figure the economic relationship between Economic Need Index and Absent Rate. To check whether our Y(Absent rate) and X(Economic Need Index)'s economic relationship is linear, we start by creating a scatterplot and study the overall pattern of the plotted points. In this case, we will use the plot from project 2.

In [44]:
```
f, axes = plt.subplots(figsize=(8, 5))

f2=sns.regplot('Economic Need Index','Absent Rate',df_combine,
               scatter_kws={'s':20,"color": "black"}, line_kws={"color": "red"},)
axes.set_title('Economic Need Index(scale from 0-1) vs Absent Rate')
```

Out[44]:
```
Text(0.5, 1.0, 'Economic Need Index(scale from 0-1) vs Absent Rate')
```



The plot shows a reasonably positive linear relationship between `Economic Need Index` and `Absent Rate` since there exists an upward slope and a straight-line pattern in the plotted data points. Also, in part 2, we noticed that the correlation between those two variables are 0.54.

Specifically, if higher Economic Need Index students in school, the chronically absent rate will be much higher. Given the plot, choosing a linear model to describe this relationship seems like a reasonable assumption.

In [45]:
```
df_combine.rename(columns={"Absent Rate": "Chronic_Absent_Rate"}, inplace = True)
df_combine.rename(columns={"Percent ELL": "Percent_ELL"}, inplace = True)
df_combine.rename(columns={"Percent Asian": "Percent_Asian"}, inplace = True)
df_combine.rename(columns={"Percent Black": "Percent_Black"}, inplace = True)
df_combine.rename(columns={"Percent Hispanic": "Percent_Hispanic"}, inplace = True)
df_combine.rename(columns={"Percent White": "Percent_White"}, inplace = True)
df_combine.rename(columns={"Collaborative Teachers %": "Percent_Collaborative_Teachers"}, inplace =
df_combine.rename(columns={"Economic Need Index": "Economic_Need_Index"}, inplace = True)
```

Ordinary least squares (OLS) regression helps to estimate the relationship between one or more independent variables and a dependent variable. In this case, the independent variable is `Economic_Need_Index` and dependent variable is `Chronic_Absent_Rate` .

In [46]:
```
results1 = ols('Chronic_Absent_Rate ~ Economic_Need_Index', data=df_combine).fit()
print(results1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Chronic_Absent_Rate   R-squared:                       0.293
Model:                             OLS   Adj. R-squared:                  0.292
Method:                  Least Squares   F-statistic:                     437.6
Date:                 Sat, 19 Dec 2020   Prob (F-statistic):           1.36e-81
Time:                         05:45:14   Log-Likelihood:                730.75
No. Observations:                 1059   AIC:                            -1457.
Df Residuals:                     1057   BIC:                            -1448.
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.0390      0.013     -2.926      0.004      -0.065      -0.013
Economic_Need_Index     0.3831      0.018     20.920      0.000       0.347       0.419
==============================================================================
Omnibus:                      612.953   Durbin-Watson:                   1.431
```

Prob(Omnibus):	0.000	Jarque-Bera (JB):	9647.819
Skew:	2.331	Prob(JB):	0.00
Kurtosis:	17.032	Cond. No.	7.37

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 4-1: OLS Regression Result of Model1

We can write our simple linear regression model as

$$ChronicAbsentRate_i = \beta_0 + \beta_1 EconomicNeedIndex_i + u_i$$

where:

- β_0 is the intercept of the linear trend line on the y-axis
- β_1 is the slope of the linear trend line, representing the marginal effect of Economic_Need_Index against Chronic_Absent_Rate
- u_i is a random error term (deviations of observations from the linear trend due to factors not included in the model)

From our results, we see that

- The intercept $\hat{\beta}_0 = -0.0390$.
- The slope $\hat{\beta}_1 = 0.3831$.
- The positive $\hat{\beta}_1$ parameter estimate implies that. Economic_Need_Index has a positive effect on outcomes, as we saw in the figure.
- The p-value of 0.000 for $\hat{\beta}_1$ implies that the effect of Economic_Need_Index is statistically significant (using p < 0.05 as a rejection rule).
- The R-squared value of 0.293 indicates that around 29% of variation in Chronic_Absent_Rate is explained by Economic_Need_Index.

This model looks good. And it definitely matches our theory that there exists positive linear relationship between `Economic_Need_Index` and `Chronic_Absent_Rate` and this effect is noteworthy. In the next step, we will consider the Economic Need Index as one main effect and introduce more explanatory variables to predict a response variable's outcome. More variation is supposed to be explained.

We introduced lots of variables from project 1 to project 3, including Races, Population, and Collaborative Teachers %. And see how they impact the absent rate individually based on the correlation graph and bar graph. Noticed that the economic relationship between those independent variables and chronic absenteeism rate are also linear(project 2). We can define those variables into two groups: characteristics and environment. Characteristics variables include Races, Population; the Environment variables include Economic Need Index, and Collaborative Teachers %.

Then, we use those variables to answer our final questions: How much of those independent variables impact the chronic absent rate?

Before we fit a multiple linear regression model. In project 3. we find a positive linear relationship between population and Economic Need Index. We use the OLS model to check whether there is strong multicollinearity between those two variables.

In [47]:
```
reg0 = ols('Chronic_Absent_Rate ~ Economic_Need_Index+Population', data=df_combine).fit()
print(reg0.summary())
```

OLS Regression Results

```
Dep. Variable:     Chronic_Absent_Rate   R-squared:                       0.295
Model:                             OLS   Adj. R-squared:                  0.294
Method:                  Least Squares   F-statistic:                     221.0
Date:                 Sat, 19 Dec 2020   Prob (F-statistic):           7.08e-81
Time:                         05:45:14   Log-Likelihood:                731.51
No. Observations:                1058   AIC:                            -1457.
Df Residuals:                    1055   BIC:                            -1442.
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        -0.0227      0.016     -1.451      0.147      -0.053       0.008
Economic_Need_Index     0.3917      0.019     20.816      0.000       0.355       0.429
Population      -3.406e-07   1.73e-07     -1.968      0.049    -6.8e-07   -9.54e-10
==============================================================================
Omnibus:                      609.038   Durbin-Watson:                   1.435
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9484.693
Skew:                           2.317   Prob(JB):                         0.00
Kurtosis:                      16.917   Cond. No.                     4.15e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.15e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Table 4-2: OLS Regression Result of Model with Endogeneity

The model results indicate that there are strong multicollinearity or other numerical problems. To eliminate the multicollinearity problem, we need to recode the variable `Population` . Since the Population variables has a range from 3335 to 103123 with median 67948 , we will consider turn it into a categorical variable based on the value of it(value > 67948 : pop_high = 1, value <=67948 : pop_high = 0).We gradually include more variables in the model. The `pop_high=1` represents that the area has a significant number of population.

In [48]:
```
df_combine['Population'].describe()
```

Out[48]:
```
count      1058.000000
mean      65222.784499
std       22168.013882
min        3335.000000
25%       46843.000000
50%       67948.000000
75%       81716.000000
max      103123.000000
Name: Population, dtype: float64
```

In [49]:
```
df_combine.loc[df_combine['Population'] <=67948, 'pop_high'] = 0
df_combine.loc[df_combine['Population'] > 67948, 'pop_high'] = 1
```

Then,we gradually include more variables in the model. And see how the regression coefficient changes and whether the model is improved after we add more variables. We expect R^2 value increase since more independent variables can explain the variation of dependent variables. The coefficient between `Economic_Need_Index` and `Chronic_Absent_Rate` should remain positive. In other words, we expand the simple regression model into the multiple regression model.

In this way, our model becomes more powerful, and the estimation ability increases. New models can help researchers to earn more information about how different features affect the absence rate. The Bureau of Education can use this data to make policies that reduce the rate of truancy. The decreasing absent race provides more supports for students' future success.

The following are the four models that we may use to predict the absent rate. We will compare them and select the most appropriate model by R^2, AIC and BIC.

$$\widehat{Model2}: \widehat{ChronicAbsentRate_i} = \beta_0 + \beta_1 EconomicNeedIndex_i + \beta_2 PercentELL_i$$

$$\widehat{Model3}: \widehat{ChronicAbsentRate_i} = \beta_0 + \beta_1 EconomicNeedIndex_i + \beta_2 PercentELL_i + \beta_3 PercentAsian_i + \beta_4 PercentBlack_i + \beta_5 PercentHispanic_i + \beta_6 PercentWhite_i$$

$$\widehat{Model4}: \widehat{ChronicAbsentRate_i} = \beta_0 + \beta_1 EconomicNeedIndex_i + \beta_2 PercentELL_i + \beta_3 PercentAsian_i + \beta_4 PercentBlack_i + \beta_5 PercentHispanic_i + \beta_6 PercentWhite_i + \beta_7 pophigh_i$$

$$\widehat{Model5}: \widehat{ChronicAbsentRate_i} = \beta_0 + \beta_1 EconomicNeedIndex_i + \beta_2 PercentELL_i + \beta_3 PercentAsian_i + \beta_4 PercentBlack_i + \beta_5 PercentHispanic_i + \beta_6 PercentWhite_i + \beta_7 pophigh_i \beta_8 PercentCollaborativeTeachers_i$$

In [50]:
```
reg1 = ols('Chronic_Absent_Rate ~ Economic_Need_Index', data=df_combine).fit()
reg2 = ols('Chronic_Absent_Rate ~ Economic_Need_Index+Percent_Asian+Percent_Black+Percent_Hispanic+P
          data=df_combine).fit()
reg3 = ols('Chronic_Absent_Rate ~ Economic_Need_Index+Percent_Asian+Percent_Black+Percent_Hispanic+P
          data=df_combine).fit()
reg4 = ols('Chronic_Absent_Rate ~ Economic_Need_Index+Percent_Asian+Percent_Black+Percent_Hispanic+P
          ,data=df_combine).fit()
```

In [51]:
```
from statsmodels.iolib.summary2 import summary_col
info_dict={'R-squared' : lambda x: f"{x.rsquared:.2f}",
          'No. observations' : lambda x: f"{int(x.nobs):d}"}


results_table = summary_col(results=[reg1,reg2,reg3,reg4],
                            float_format='%0.3f',
                            stars = True,
                            model_names=['Model 2',
                                         'Model 3',
                                         'Model 4',
                                         'Model 5'],
                            info_dict=info_dict,
                            regressor_order=['Intercept',
                                             'Economic_Need_Index'
                                             'Percent_Asian',
                                             'Percent_Black',
                                             'Percent_Hispanic',
                                             'Percent_White',
                                             'Percent_Collaborative_Teachers'])

results_table.add_title('Table 4-2 - OLS Regressions Summaries for Model 1-5')

print(results_table)
```

Table 4-2 - OLS Regressions Summaries for Model 1-5

	Model 2	Model 3	Model 4	Model 5
Intercept	-0.039***	0.497***	0.447**	0.561***
	(0.013)	(0.173)	(0.174)	(0.179)
Percent_Black		-0.605***	-0.552***	-0.560***
		(0.185)	(0.186)	(0.185)
Percent_Hispanic		-0.724***	-0.677***	-0.686***
		(0.184)	(0.184)	(0.184)
Percent_White		-0.486***	-0.430**	-0.429**
		(0.183)	(0.184)	(0.183)
Percent_Collaborative_Teachers				-0.121**
				(0.047)
pop_high			-0.019**	-0.019***
			(0.007)	(0.007)
Percent_Asian		-0.835***	-0.781***	-0.788***

		(0.183)	(0.184)	(0.184)
Economic_Need_Index	0.383***	0.547***	0.562***	0.562***
	(0.018)	(0.037)	(0.037)	(0.037)
R-squared	0.293	0.388	0.391	0.395
R-squared Adj.	0.292	0.385	0.388	0.391
R-squared	0.29	0.39	0.39	0.40
No. observations	1059	1059	1058	1058

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

Table 4-3: OLS Regressions Summaries for Model 2-5

Table 4-3 summarize the coefficient of linear regression model 2-5. There is an increase of R^2 as more variables are added. Based on the method we mentioned in the introduction part, we will select model 6 since its R^2 value is the largest. Then, we print the result from model 5.

In [52]:
```
reg4 = ols('Chronic_Absent_Rate ~ Economic_Need_Index+Percent_Asian+Percent_Black+Percent_Hispanic+P
print(reg4.summary())
```

OLS Regression Results

Dep. Variable:	Chronic_Absent_Rate	R-squared:	0.395
Model:	OLS	Adj. R-squared:	0.391
Method:	Least Squares	F-statistic:	97.94
Date:	Sat, 19 Dec 2020	Prob (F-statistic):	4.91e-110
Time:	05:45:14	Log-Likelihood:	812.28
No. Observations:	1058	AIC:	-1609.
Df Residuals:	1050	BIC:	-1569.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5610	0.179	3.135	0.002	0.210	0.912
Economic_Need_Index	0.5618	0.037	15.209	0.000	0.489	0.634
Percent_Asian	-0.7881	0.184	-4.288	0.000	-1.149	-0.427
Percent_Black	-0.5601	0.185	-3.023	0.003	-0.924	-0.197
Percent_Hispanic	-0.6861	0.184	-3.731	0.000	-1.047	-0.325
Percent_White	-0.4294	0.183	-2.343	0.019	-0.789	-0.070
pop_high	-0.0187	0.007	-2.587	0.010	-0.033	-0.005
Percent_Collaborative_Teachers	-0.1215	0.047	-2.570	0.010	-0.214	-0.029

Omnibus:	686.013	Durbin-Watson:	1.504
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14317.560
Skew:	2.627	Prob(JB):	0.00
Kurtosis:	20.239	Cond. No.	199.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 4-4: OLS Regressions Summaries for Model 5

This model looks good and it avoid endogeneity problem. Thus, our final model can be concluded as

$$\begin{aligned} Model5: \widehat{ChronicAbsentRate_i} = {} & 0.5610 + 0.5618 EconomicNeedIndex_i - 0.7881 PercentAsian_i - 0.560 \\ & - 0.6861 PercentHispanic_i - 0.4294 PercentWhite_i - 0.0187 pophigh_i \\ & - 0.1215 PercentCollaborativeTeachers_i \end{aligned}$$

From our results, we see that $R^2 = 0.395$ indicates that 39.5% of the sample variance is captured in the model.In general,all independent variables are statistically significant since p-value < 0.05.

The percentage of each race, higher population, and Percent Collaborative Teachers will negatively impact Chronic_Absent_Rate, which is good news. For the characteristic variables, we cannot control students' race and

the population around the school, so the smaller the effects of the characteristic variables on the absence rate will help us put more effort into the environment factors that we can make change.

For our two environment variables, we noticed that a higher Percentage of Collaborative Teachers would decrease the absent rate, representing that teachers' behavior will influence students' engagement. The larger the Economic Need Index leads to an significant increase in the chronic absent rate. Thus, financial support is vital for those schools with high economic index. If a school has sufficient financial support, it will provide a better infrastructure that improves the learning environment and decreases the absence rate.

The following are specific coefficient interpretations.

Togethered with Intercept, we can interpret that if the population around school is less than 67948(`pop_high=0`), holding other variables constant,the `Chronic_Absent_Rate` is around 0.5610 on average.

Interpretation of Economic_Need_Index coefficient: Holding other variables constant,1 percentage increase in `Economic Need Index` will lead to $0.5618 \times 0.01 \approx 0.006 = 0.6$ percentage increase in absent rate on average.

Interpretation of Percent_Asian coefficient: Holding other variables constant,1 percentage increase in `Percent Asian` will lead to $0.7881 \times 0.01 \approx 0.008 = 0.8$ percentage decrease in absent rate on average.

Interpretation of Percent_Black coefficient: Holding other variables constant,1 percentage increase in `Percent Black` will lead to $0.5601 \times 0.01 \approx 0.006 = 0.6$ percentage decrease in absent rate on average.

Interpretation of Percent_Hispanic coefficient: Holding other variables constant,1 percentage increase in `Percent Hispanic` will lead to $0.6861 \times 0.01 \approx 0.007 = 0.7$ percentage decrease in absent rate on average.

Interpretation of Percent_White coefficient: Holding other variables constant,1 percentage increase in `Percent White` will lead to $0.4294 \times 0.01 \approx 0.004 = 0.4$ percentage decrease in absent rate on average.

Interpretation of pop_high coefficient: Holding other variables constant,if the population around school is greater than 67948, the `Chronic_Absent_Rate` will decrease by $0.0187 \times 0.01 \approx 0.0002 = 0.02$ percentage increase in absent rate on average.

Interpretation of Percent_Collaborative_Teachers coefficient: Holding other variables constant,1 percentage increase in `Collaborative Teachers` will lead to $0.1215 \times 0.01 \approx 0.001 = 0.1$ percentage decrease in absent rate on average.

Conclusions and next steps(Updated from all previous projects)

This literature mainly researches how New York City schools' demographics influence students' chronically absent rate. Unlike other literature, we mainly focus on find the variables that affect the chronic absent rate and fit the model to measure the level of those independent variables' effects on absent rate. We use those variables to answer the following three questions: (1) What percentage of students are chronically absent in 2016? (2) How each school environment and characteristics variables influence the chronically absent rate? (3) How much of those independent variables impact the chronic absent rate?

In our analysis, we mainly focus on the 7 areas (Brooklyn, Bronx, New York, Staten Island, Jamaica, Flushing, Long Island City) with the greatest number of schools of New York City. The average percentage of chronically absent rate is around 23%. And this excessive value leads to society's concern. We find that the areas (i.e., New York,

Bronx, and Brooklyn) with more schools lead to a greater absent rate. Based on the correlation table, we find that Percent Asian, Percent White, and Collaborative Teachers % is negatively correlated with absent rate. However, the Percent Black, Percent Hispanic, and Economic Need Index are positively correlated with absent rate.

Then, we add the population into the dataset. The regression plot indicates that the effects of population on absence rate are inapparent and there may exists endogeneity problem between Economic Need Index and Population. Thus, we need to cautious about this when we fit the model. Later, we fit a regression model to find how much of those independent variables impact the chronic absent rate. In this part, we test the economic relationship between our independent variables and dependent variables. The scatter plots indicate a strong linear relationship between our Xs and Y. We recode the population variables into categorical variables to avoid endogeneity problem. The OLS regression results show an increase in R-squared value as we add more variables to the model. In other words, a larger percentage of the variation in the absent rate is explained by our independent variables. After comparing 4 potential multiple regression models, we selected the one with largest R-squared value. The model is:

$$\begin{aligned} Model5: \widehat{ChronicAbsentRate}_i = {} & 0.5610 + 0.5618 EconomicNeedIndex_i - 0.7881 PercentAsian_i - 0.560 \\ & - 0.6861 PercentHispanic_i - 0.4294 PercentWhite_i - 0.0187 pophigh_i \\ & - 0.1215 PercentCollaborativeTeachers_i \end{aligned}$$

All coefficients in this model are statistically significant since the P-value < 0.05. The characteristics variables (percent race and pop_high) and Percent Collaborative Teachers have a negative impact on Chronic Absent rates. The Economic Need Index is the only variable that has a positive effect on Chronic Absent Rate. It is hard to change the characteristic variables, so we want to minimize the adverse effects of environment variables (Economic Need Index and Percent Collaborative Teachers) on the absent rate. Based on the result I listed in the analysis part, to decrease the absent rate, the bureau of education should supervise more on the teachers' collaboration work and provide more economic support for those schools with a high economic need index.

There have several limitations to this model. The regression model results vary from the correlation table results because there are some unobservable effects on races and the absent rate. For instance, in our data, if we add the percent of each race in one school, the result is more than 100%. Some mixed-blood students may identify by two or more races that influence data accuracy. In future work, we may build another model to test those unobservable effects and use 2SLS method to deal with endogeneity problems.

We also have a large dataset that can support us apply the machine learning methods such as text analysis, regression trees, or random forests in future studies. We can use text analysis to research students' attitudes towards the school environment from online postings. We can tackle both the econometrics of the OLS and regression trees and the economic intuition behind both models.

If applicable, we can fit our model on other cities' dataset. The Economic Need Index and Collaborative Teacher Percent may or may not be the key issues for those cities' chronic absent rate. Adding other environment variables into the model will significantly improve our model's accuracy and explanation ability.

Resources:

1. U.S. Department of Education. (2019). Chronic absenteeism in the nation's schools: An unprecedented look at a hidden educational crisis. Accessed January 2019, at https://www2.ed.gov/datastory/chronicabsenteeism.html

2. Bartanen, Brendan. (2020). Principal Quality and Student Attendance. Educational Researcher. 49. 0013189X1989870. 10.3102/0013189X19898702. Accessed at https://www.researchgate.net/publication/338563399_Principal_Quality_and_Student_Attendance

3. Posterli, T. (2018, September 14). Did You Know That 28 Percent of New York Students Are Chronically Absent? What's The Solution? Retrieved December 18, 2020, from https://newyorkschooltalk.org/2018/09/know-28-percent-new-york-students-chronically-absent-whats-solution/

4. London, R.A., Sanchez, M., & Castrechini, S. (2016). The dynamics of chronic absence and student achievement. Education Policy Analysis Archives, 24(112). http://dx.doi.org/10.14507/epaa.24.2741