The Relationship between SAT Scores and Ethnicity - A NYC Study

SAT scores are predominantly used in US to a judge a student's math, reading and writing skills. Universities use these scores to judge whether the students that applied to their university will be a good fit. The higher the SAT score, the better chance a student has to enter a top notch university.

The following study uses data about average SAT scores for NYC public school from NYC open data. This data was gathered for the 2014-2015 school year. The data set includes information about school name, zipcode level characteristics, average SAT scores for math, reading and writing, as well as percentage of different ethnicities in the school and the percentage of students taking SAT in that school. The important variables for this study are average SAT scores in Math and Writing for each school. As well as the percentage of different ethnicities; namely, White, Black, Hispanic and Asian in the school.

The aim of this study is to analyze the relation between a different ethnicities which are the independent variables and the average Math and Writing SAT scores which are the dependent variables of NYC public schools. This will be done through tables of key summary statistics and graphs of the important vairbales.

Data

For the purpose of this report, all NaN values have been removed as they had missing information on the important variables.The total number of observations decreased from 435 to 374. The value of 0 is not treated as NaN because 0 signifies that a certain school had 0% a certain ethnicity. The data has been thoroughly cleaned to remove any '%' signs for the purpose of analysis.

In [1]:

```python
#import the libraries needed
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
from IPython.display import display
```

In [2]:

```
#understand the dataset and what columns are provided
nyc_hs = pd.read_csv(r"C:\Users\anusha\Desktop\eco225\scores.csv")
nyc_hs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435 entries, 0 to 434
Data columns (total 22 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   School ID                    435 non-null    object 
 1   School Name                  435 non-null    object 
 2   Borough                      435 non-null    object 
 3   Building Code                435 non-null    object 
 4   Street Address               435 non-null    object 
 5   City                         435 non-null    object 
 6   State                        435 non-null    object 
 7   Zip Code                     435 non-null    int64  
 8   Latitude                     435 non-null    float64
 9   Longitude                    435 non-null    float64
 10  Phone Number                 435 non-null    object 
 11  Start Time                   431 non-null    object 
 12  End Time                     431 non-null    object 
 13  Student Enrollment           428 non-null    float64
 14  Percent White                428 non-null    object 
 15  Percent Black                428 non-null    object 
 16  Percent Hispanic             428 non-null    object 
 17  Percent Asian                428 non-null    object 
 18  Average Score (SAT Math)     375 non-null    float64
 19  Average Score (SAT Reading)  375 non-null    float64
 20  Average Score (SAT Writing)  375 non-null    float64
 21  Percent Tested               386 non-null    object 
dtypes: float64(6), int64(1), object(15)
memory usage: 49.3+ KB
```

In [3]:

```
nyc_hs = pd.DataFrame(nyc_hs)
```

In [4]:

```
# drop Na values and create a dataframe
nyc_hs_new = nyc_hs.dropna()
nyc_hs_new = pd.DataFrame(nyc_hs_new)
```

In [5]:

```
%%time
white = nyc_hs_new["Percent_White_str"] = nyc_hs_new["Percent White"].str.replace("%", "")

black = nyc_hs_new["Percent_Black_str"] = nyc_hs_new["Percent Black"].str.replace("%", "")

hispanic = nyc_hs_new["Percent_Hispanic_str"] = nyc_hs_new["Percent Hispanic"].str.replace("%", "")

asian = nyc_hs_new["Percent_Asian_str"] = nyc_hs_new["Percent Asian"].str.replace("%", "")
```

```
Wall time: 9 ms
```

In [6]:

```
# convert the values to numeric after removing the % sign above
nyc_hs_new["WhitePercent_numeric"] = pd.to_numeric(nyc_hs_new["Percent_White_str"])
white_tonumeric = nyc_hs_new["WhitePercent_numeric"] = pd.to_numeric(nyc_hs_new["Percent_White_str"])

nyc_hs_new["BlackPercent_numeric"] = pd.to_numeric(nyc_hs_new["Percent_Black_str"])
white_tonumeric = nyc_hs_new["BlackPercent_numeric"] = pd.to_numeric(nyc_hs_new["Percent_Black_str"])

nyc_hs_new["HispanicPercent_numeric"] = pd.to_numeric(nyc_hs_new["Percent_Hispanic_str"])
white_tonumeric = nyc_hs_new["HispanicPercent_numeric"] = pd.to_numeric(nyc_hs_new["Percent_Hispanic_str"])

nyc_hs_new["AsianPercent_numeric"] = pd.to_numeric(nyc_hs_new["Percent_Asian_str"])
white_tonumeric = nyc_hs_new["AsianPercent_numeric"] = pd.to_numeric(nyc_hs_new["Percent_Asian_str"])
```

Summary of Key Variables

In [7]:

```
# create a new data frame with all SAT score averages
Avg_SAT_score = nyc_hs_new[["Average Score (SAT Math)", "Average Score (SAT Reading)", "Average Score (SAT Wr
iting)"]]
Avg_SAT_score = pd.DataFrame(Avg_SAT_score)
Avg_SAT_score
Avg_SAT_score.describe()
```

Out[7]:

	Average Score (SAT Math)	Average Score (SAT Reading)	Average Score (SAT Writing)
count	374.000000	374.000000	374.000000
mean	432.719251	424.342246	418.286096
std	71.916833	61.884529	64.548388
min	317.000000	302.000000	284.000000
25%	386.000000	386.000000	382.000000
50%	414.000000	412.500000	402.500000
75%	457.250000	444.500000	436.000000
max	754.000000	697.000000	693.000000

These are the summary statistics for Average SAT scores. Some interesting points to note are that the maximum score for Math is higher than Reading and Writing. However, the mean scores are almost in the same range of 415 to 435. The minimum for Writing is the lowest. Generally, students do better in Math and worse in Writing. For the purpose of this report we will focus on Average SAT Math and Writing scores to see if there is any correlation between different ethnicities and the average scores.

In [8]:

```
# create a new data frame with all ethnicities
ethnicities = nyc_hs_new[["WhitePercent_numeric", "BlackPercent_numeric", "HispanicPercent_numeric","AsianPer
cent_numeric" ]]
ethnicities = pd.DataFrame(ethnicities)
ethnicities
ethnicities.describe()
```

Out[8]:

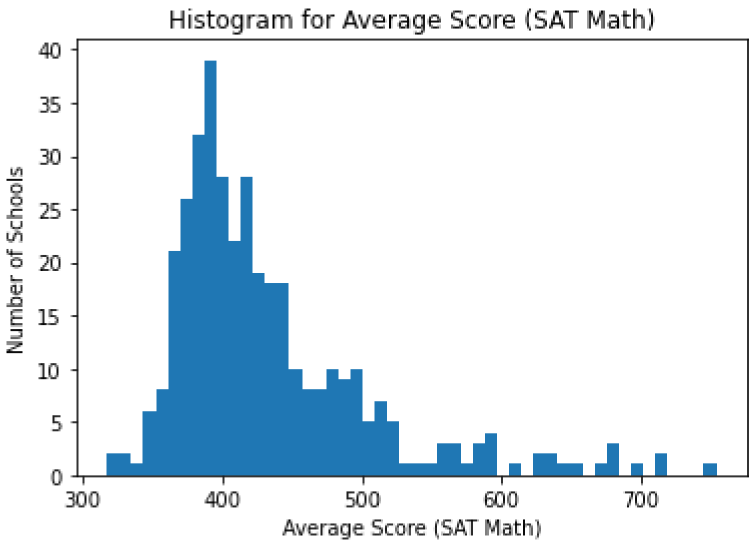
	WhitePercent_numeric	BlackPercent_numeric	HispanicPercent_numeric	AsianPercent_numeric
count	374.000000	374.000000	374.000000	374.000000
mean	8.524599	35.387166	43.929679	10.412567
std	13.359205	25.367159	24.495584	14.400556
min	0.000000	0.000000	2.600000	0.000000
25%	1.300000	16.400000	20.825000	1.600000
50%	2.600000	28.750000	45.300000	4.200000
75%	9.375000	50.100000	63.375000	11.150000
max	79.900000	91.200000	100.000000	88.900000

These are the summary statistics for various ethnicities in the NYC public schools. Some interesting points to note are that some schools have 0% of White, Black and Asian ethnicities while Hispanics are almost 100%. The mean varies a lot as well, some schools have around 8.5% White ethnicity while the mean for Hispanic ethnicity is 44%.

Visual Summary of the Key Variables

```
In [9]: plt.hist(nyc_hs_new["Average Score (SAT Math)"], bins = 50)
        plt.title("Histogram for Average Score (SAT Math)")
        plt.xlabel("Average Score (SAT Math)")
        plt.ylabel("Number of Schools")

Out[9]: Text(0, 0.5, 'Number of Schools')
```



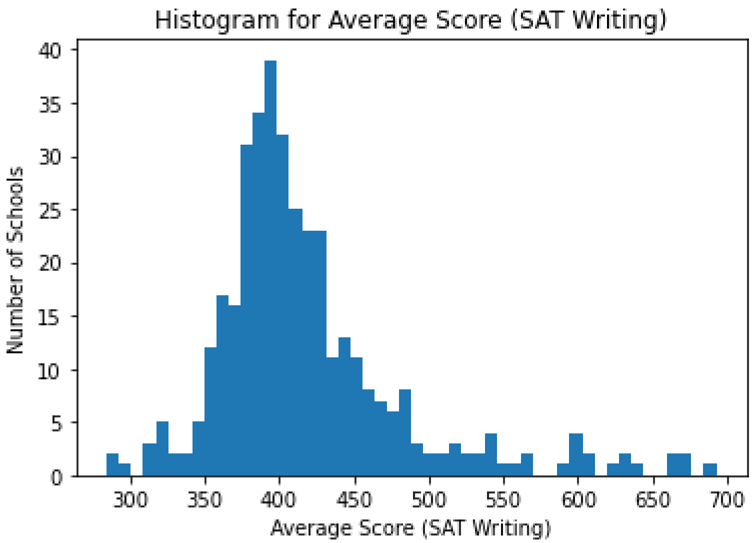
In the histogram above we can see the distribution of average SAT Math scores in NYC public schools. This tells us that many schools had an average score between 350 - 400. Schools did not score worse than 300 and some schools did really well with an average score above 700.

In [10]:

```
plt.hist(nyc_hs_new["Average Score (SAT Writing)"], bins = 50)
plt.title("Histogram for Average Score (SAT Writing)")
plt.xlabel("Average Score (SAT Writing)")
plt.ylabel("Number of Schools")
```

Out[10]:

```
Text(0, 0.5, 'Number of Schools')
```



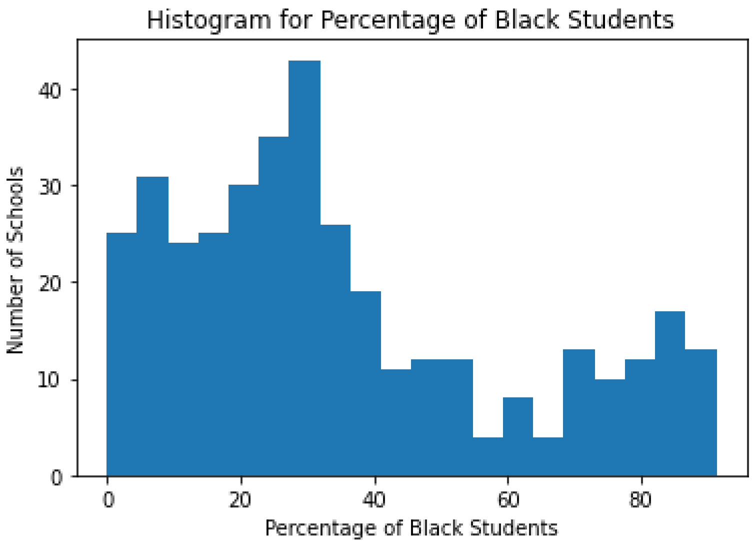
In the histogram above we can see the distribution of average SAT Writing scores in NYC public schools. This tells us that many schools had an average score between 375 - 425. Some schools did score worse than an average of 300 and very few schools did really well with an average score ranging from 650 - 700. Comparing with the Math average histogram above we can see that the mean distribution for Writing scores is actually slightly better than Math scores. This was not clear in the summary table under "Summary of Key Variables".

In [11]:

```
plt.hist(nyc_hs_new["BlackPercent_numeric"], bins = 20)
plt.title("Histogram for Percentage of Black Students")
plt.xlabel("Percentage of Black Students")
plt.ylabel("Number of Schools")
```

Out[11]:

```
Text(0, 0.5, 'Number of Schools')
```

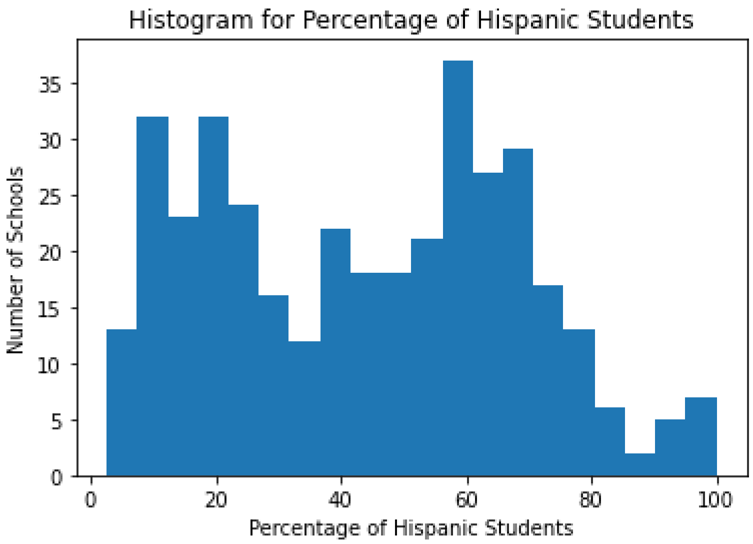


In [12]:

```
plt.hist(nyc_hs_new["HispanicPercent_numeric"], bins = 20)
plt.title("Histogram for Percentage of Hispanic Students")
plt.xlabel("Percentage of Hispanic Students")
plt.ylabel("Number of Schools")
```

Out[12]:

```
Text(0, 0.5, 'Number of Schools')
```



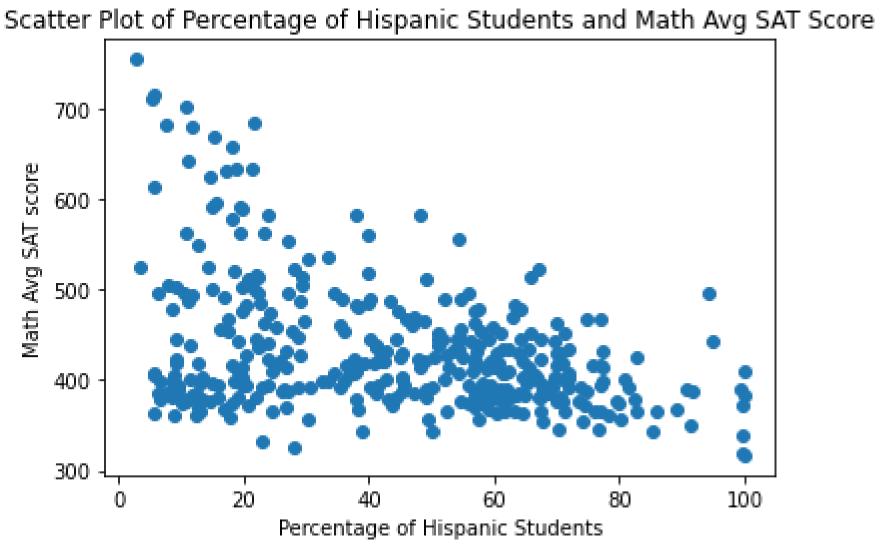
From the two histograms above it is clear that NYC Public schools have a great proportion of Black and Hispanic ethnicities in their schools.

In [13]:

```
# creating variables that define the specific ethnicity column - without specifying I'd only get one point on scatter plot
w = nyc_hs_new["WhitePercent_numeric"]
b = nyc_hs_new["BlackPercent_numeric"]
h = nyc_hs_new["HispanicPercent_numeric"]
a = nyc_hs_new["AsianPercent_numeric"]
y1 = nyc_hs_new["Average Score (SAT Math)"]
```

In [14]:
```
plt.scatter(h, y1)
plt.title("Scatter Plot of Percentage of Hispanic Students and Math Avg SAT Score")
plt.xlabel("Percentage of Hispanic Students")
plt.ylabel("Math Avg SAT score")
```

Out[14]:
```
Text(0, 0.5, 'Math Avg SAT score')
```

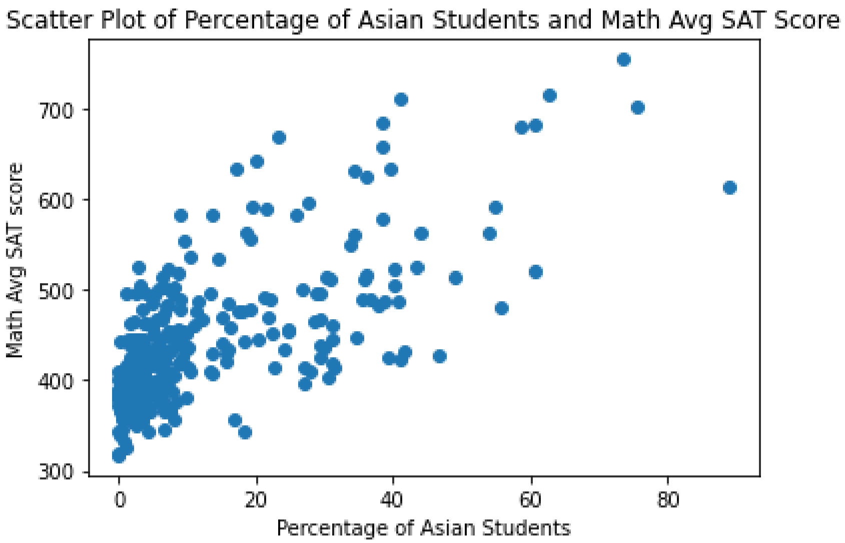


In [15]:

```
plt.scatter(a, y1)
plt.title("Scatter Plot of Percentage of Asian Students and Math Avg SAT Score")
plt.xlabel("Percentage of Asian Students")
plt.ylabel("Math Avg SAT score")
```

Out[15]:

```
Text(0, 0.5, 'Math Avg SAT score')
```



Comparing the scatter plots for White, Black, Hispanic and Asians for average SAT Math score, a clear upward trend is observed for a school with higher percentage of Asian students. The higher percentage of Asians, the higher the average SAT score for a public school. A similar result is observed for White ethinicity. However, a clear downward trend is observed for a school with higher percentage of Hispanic students. The higher percentage of Hispanic students, the lower the average SAT score for a public school. Similar results are observed for Black students.

In conclusion, it is noticeable that schools with a higher Asian and White ethnicity tend to do better at Math compared to schools with Hispanic and Black ethnicity. The result is similar for Writing scores. It will be naive to say that there is a definite correlation, for that a model needs to be created and studied. For the purpose of policy making it will be interesting to study whether these trends are directly related to the ethnicity a school belongs to or is there some discrimination in the facilities being provided to schools with a majority of Black and Hispanic ethnicity.