Does the Borough or Ethnicity affect a student's overall SAT score in NYC

Project 1

Introduction

With the use of high school data compiled from the New York City Department of Education, and the SAT score averages, and testing rates provided by the College Board, we analyze and discuss whether the Borough or high school a student goes to affect their overall SAT score in NYC. Our emphasis in this paper is to analyze whether we can see differences in high schools of each borough and whether ethnicity plays any role in variations among SAT scores.

SAT scores are standardized tests taken by the college board (an independent board) that are used by students to apply for college after high school. The score a child gets greatly impacts his/her chance in getting into college and their future. Though the new SAT tests have changed their format to 1600 points (Math: 800, Reading: 400, Writing: 400), however, we will be using data from the older format which gives us almost the same results (Math: 800, Reading: 800, Writing: 800). The main emphasis of this project is to find connections in whether some boroughs or high schools in New York City do better than others and discuss the factors that lead to these differences where we analyze ethnicity, student enrollment and percentage of students being tested. Finding differences would be extremely interesting as one can infer how getting into certain public high schools or going for schooling in certain areas of the city would impact the college a child gets into.

With our question on the effect of SAT using boroughs and ethnicity, we will consider all 5 boroughs of NYC as well as divide ethnicity in to 4 sub-groups; White, Asian, Black, and Hispanic. We will also try to add and merge further data on private schools to see whether such variations are only seen in public schools and not private.

Since this is a very sensitive topic as it is related to a child's future and college, there have been many research conducted to see whether there is any relation of ethnicity/race with college applications, SAT score and degree. The 2005 paper on predicting college grades and SAT scores using student ethnicity (Rebecca Zwick, Jeffrey C. Sklar 2005) is one literature that discusses the variations among SAT due to some variables. Similarly other research has gone in to another of our variable; borough, where we see papers on Exploring School Effects on SAT Scores (Howard T. Everson & Roger E. Millsap 2010) that rigorously dives into other factors other than individuals that may cause SAT scores to vary. These research, and literature are key foundations to the variations of SAT scores in the US, and we use them as an inspiration and combine them to see how NYC varies when it comes to SAT using boroughs and ethnicities as key variables. With our paper we will further their research as we will combine both variations in areas, and ethinicities along with other variables such as student enrollment and testing percent in trying to solve our economic question that belonging to a specific borough or ethnic group impacts the SAT score a student receives.

We will conduct statistical analysis, visualize our finding, run regressions of our different explanatory variables to see if we can see any correlation between our dependent and independent variables.

Our data is very comprehensive, and to answer our question: (Y Variable) Average SAT Scores in each borough, we will use different variables. One of our variable (X_1) will be the differnt boroughs i.e. Manhattan, Queens, while (X_2) will be the different ethinic groups. With the use of these variables we will try to develop an analysis on whether the borough you live in around NYC would give certain students a natural advantage.

In [1]:

```
import pandas as pd
import qeds
import geopandas as gpd
import seaborn as sns
import json
import folium
%matplotlib inline
# activate plot theme
import qeds
qeds.themes.mpl_style();
from shapely import wkt
from shapely.geometry import Point
from IPython.display import display
import ssl
```

```
from sentinelsat import SentinelAPI
import geojsonio
import requests  # to call data from a url
from sentinelsat import geojson_to_wkt, read_geojson
import matplotlib.pyplot as plt
ssl._create_default_https_context = ssl._create_unverified_context
```

First we are Reading the data and creating a dataframe 'df' and then running our dataframe.

In [2]:
```
df = pd.read_csv(r"C:\Users\Anwar Malik\Desktop\UofT\Year 2\ECO225\ECO225Project\Data\scores.csv")
```

Data Cleaning

To start our analysis we need to do certain steps to clean our data and convert it into something that we will find easier to interpret. Since we care about SAT scores and the distribution of different ethnic groups in each high school, we will need to clean this data and remove the '%' sign and convert all percentages and SAT scores in to float.

In the cell below we clean the data remove the '%' sign from the data. Using string methods we removed the percentage sign.

In [3]:
```
df2 = df.replace('%','',regex=True)
```

Our emphasis is to determine whether certain boroughs may have higher SAT scores, as a result, we calulate the percentage of each borough in our data.

In [4]:
```
df2['Borough'].value_counts()
```

Out[4]:
```
Brooklyn         121
Bronx            118
Manhattan        106
Queens            80
Staten Island     10
Name: Borough, dtype: int64
```

In [5]:
```
df2['Borough'].value_counts(normalize= True)
```

Out[5]:
```
Brooklyn         0.278161
Bronx            0.271264
Manhattan        0.243678
Queens           0.183908
Staten Island    0.022989
Name: Borough, dtype: float64
```

We will now remove any rows that contain missing data. This will allow us to interpret our data much easily when we run our analysis.

In [6]:
```
df2 = df2.dropna()
```

Converting all our new variables that are without the % sign to floats since they were previosuly as strings. We also make a new variable 'Non-White' which includes the Black, Hispanic and Asian population.

In [7]:
```
df2['Percent White'] = df2['Percent White'].astype(float)
df2['Percent Black'] = df2['Percent Black'].astype(float)
df2['Percent Hispanic'] = df2['Percent Hispanic'].astype(float)
df2['Percent Asian'] = df2['Percent Asian'].astype(float)
df2['Non-White'] = df2['Percent Black'] +  df2['Percent Hispanic'] + df2['Percent Asian']
df2['Percent Tested'] = df2['Percent Tested'].astype(float)
```

Summary Statistics

Part of our analysis is to determine whether certain boroughs have lower SAT scores than others. An important inferance we hope to make is whether regions with higher immigrated or non-white populations might be at a disadvantage and getting lower SAT scores. The reasons could be poor infrastructure, lack of resources and equal opportunities. Hence, we will analyse and create a summary statistics.

As a result, we will use the data we have of percentage of each populations and the average SAT scores to see whether

there is correlation between a borrough with a high white population and a low white population.

In [8]:
```
df2['Average_SAT_Score'] = df2['Average Score (SAT Math)'] + df2['Average Score (SAT Reading)'] + df2['Ave
```

As we run 5 different summary statistics of each of our 5 boroughs, we can make initial comments on how the SAT scores vary among the boroughs. Regions that tend to have higher white populations such as Staten Island, Queens and Manhattan have done considerably better an acheiving higher SAT scores which are over 1350 on average. Other regions such as Bronx and Brooklyn tend to have SAT scores close to 1200. During a very competitive college application period where each point matters, a difference of 150 points shows how extremely distributed SAT scores can be within a single city.

In [9]:
```
df3 = df2
df3 = df3.drop(columns=df3.columns[:2])
df3 = df3.drop(columns=df3.columns[1:21])
```

In [10]:
```
df3.groupby(['Borough']).describe()
```

Out[10]:

	Average_SAT_Score							
	count	mean	std	min	25%	50%	75%	max
Borough								
Bronx	98.0	1202.724490	150.393901	924.0	1131.0	1190.0	1245.0	2041.0
Brooklyn	109.0	1230.256881	154.868427	926.0	1141.0	1186.0	1298.0	1896.0
Manhattan	89.0	1340.134831	230.294140	1005.0	1173.0	1284.0	1415.0	2144.0
Queens	68.0	1343.426471	195.953796	978.0	1217.0	1288.0	1427.5	1981.0
Staten Island	10.0	1439.000000	222.303596	1258.0	1346.5	1382.0	1441.0	2041.0

The summary above shows the statistical summary of the Total Average SAT Score in each of the 5 boroughs in NYC. This summary helps us to see how each borough has different mean, standard deviations, and quartiles which really helps us to have a better understanding of our data.

Since our overall aim or Y variable is to determine the variability of boroughs, hence we will also compute the summary statistics of the Total Average Score of the SAT

In [11]:
```
df2['Average_SAT_Score'].describe()
```

Out[11]:
```
count     374.000000
mean     1275.347594
std       194.866056
min       924.000000
25%      1157.000000
50%      1226.000000
75%      1327.000000
max      2144.000000
Name: Average_SAT_Score, dtype: float64
```

Our summary statistics of the Total Average SAT Score shows us the average SAT score for all highschools for whom data is collected is 1276 points approximatly. This result is interesting as we see how some boroughs like staten island and Manhattan where white population is in majority expereince relatively higher scores on average than other boroughs.

To visualize our data we will, in the initial stage of our project, show the SAT spread of average SAT scores in each borough to give a sense of the variability among different highschools and boroughs in NYC. For better visual representation, we will be renaming some of our columns to easily fit our data and key in the plots.

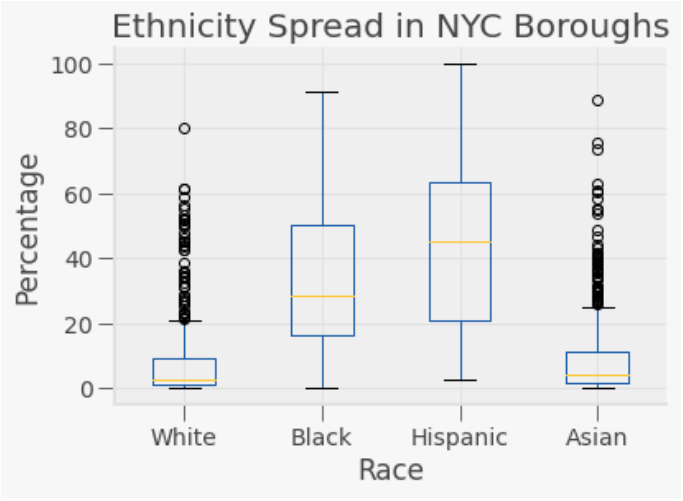
In [12]:
```
df2.rename(columns={'Average Score (SAT Math)': 'Math', 'Average Score (SAT Reading)': 'Reading', 'Average
```

In [13]:
```
df2.rename(columns={'Percent White': 'White', 'Percent Black': 'Black', 'Percent Hispanic': 'Hispanic', 'P
```

In [14]:
```
boxplot = df2.boxplot(column=["White", "Black", "Hispanic", "Asian"])
boxplot.set_ylabel('Percentage')
```

```
boxplot.set_xlabel('Race')
boxplot.set_title('Ethnicity Spread in NYC Boroughs')
```

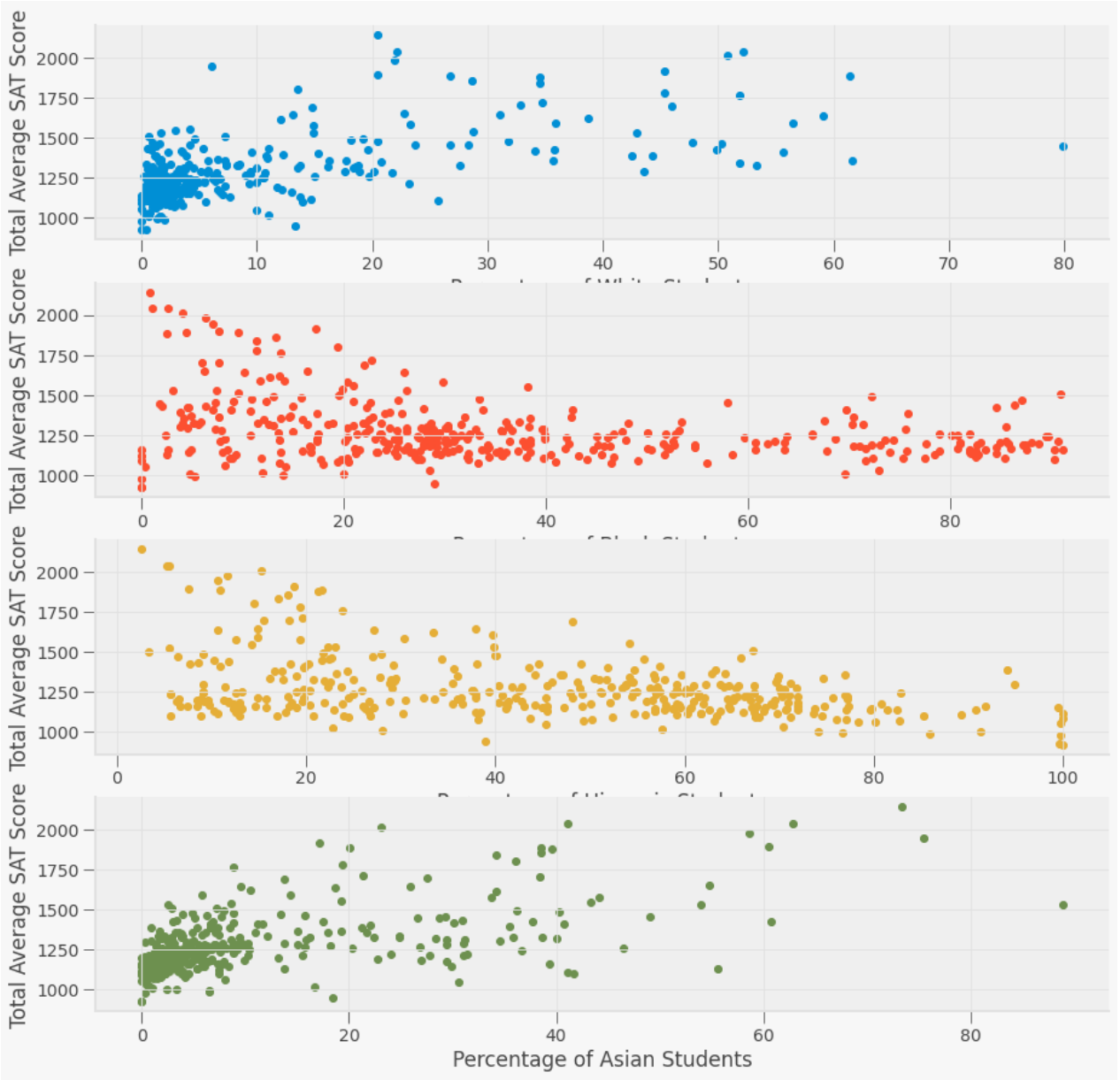
Out[14]:
```
Text(0.5, 1.0, 'Ethnicity Spread in NYC Boroughs')
```



These boxplots shows us a simple visual representation of some our variables. From a casual overview, we see how on average students gets relatively higher points in the math section, and most average SAT scores in highschools are close to 1250 with a few outliers that go over 1700. Our second boxplot shows the ethnicity spread in all these boroughs. We can see how white and asian populations are very low on average while there is a clear dominanca of black and hispoanic population. This is a key insight since in our analysis later we will interpret whether belonging from a borough with a black/hispanic majority actually might be at a disadvantage when comes to obtaining SAT scores.

We will also plot a scatter plot to see the spread of different ethinic groups.

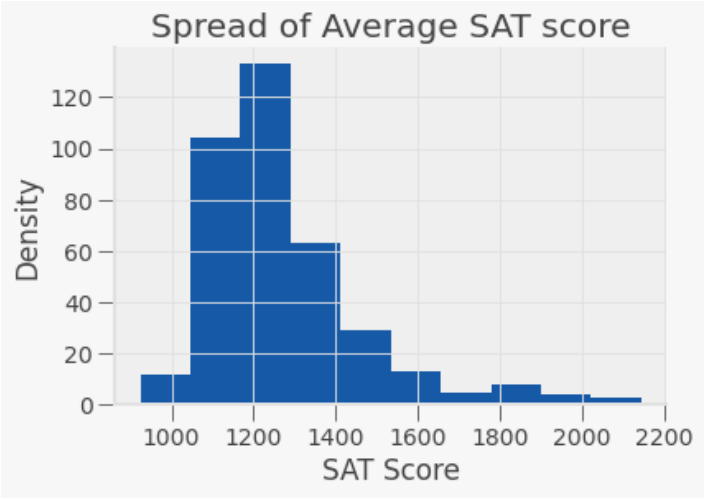
In [15]:
```
fig, ax = plt.subplots(4, figsize=(15, 15))
ax[0].scatter(x = df2['White'], y = df2['Average_SAT_Score'], color = '#008fd5')
ax[0].set_xlabel("Percentage of White Students")
ax[0].set_ylabel("Total Average SAT Score")

ax[1].scatter(x = df2['Black'], y = df2['Average_SAT_Score'], color = '#fc4f30')
ax[1].set_xlabel("Percentage of Black Students")
ax[1].set_ylabel("Total Average SAT Score")

ax[2].scatter(x = df2['Hispanic'], y = df2['Average_SAT_Score'], color = '#e5ae37')
ax[2].set_xlabel("Percentage of Hispanic Students")
ax[2].set_ylabel("Total Average SAT Score")

ax[3].scatter(x = df2['Asian'], y = df2['Average_SAT_Score'], color = '#6d904f')
ax[3].set_xlabel("Percentage of Asian Students")
ax[3].set_ylabel("Total Average SAT Score")

plt.show()
```



The multiple scatter plots above show each ethinic group with respect to average SAT scores. We see ethinic groups on the x-axis and Average SAT scores on the y-axis. Through observation we can see that white and Asian are heavily consentrated at 10% while black and hispanic are much widely spread and vary. These variations are interesting and is something we will ponder over.

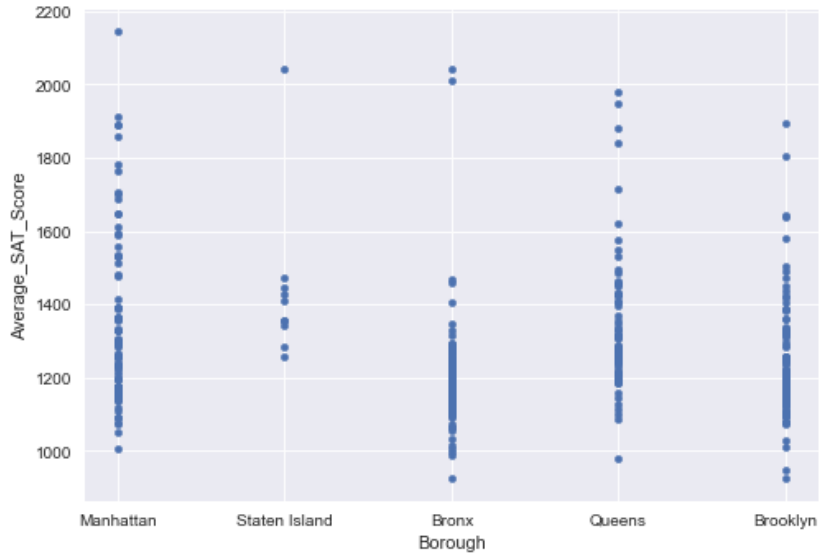
In [16]:
```
df2['Average_SAT_Score'].hist()
plt.title('Spread of Average SAT score')
plt.ylabel('Density')
plt.xlabel('SAT Score')
```

Out[16]:
```
Text(0.5, 0, 'SAT Score')
```



For further visual representation between the different boroughs and SAT scores, we are going to develop a histogram that shows how our sample looks like. We see a positive skew towards the right tail of the histogram which indicates that most of the SAT scores lie close to the mean at 1250 while only a few highschools in these boroughs have extremely high scores. The long skew could also refer to outliers. Our emphasis will be to see whether most boroughs with SAT scores close to or lower 1200 are from which ethinic group.

In [17]:
```
plt.style.use('seaborn')

df2.plot(x='Borough', y='Average_SAT_Score', kind='scatter')
plt.show()
```



After collecting data of different public schools we can plot a scatter plot that indicates the spread of SAT scores in each highschool of each borough. Our result shows that the sample size of Manhattan, Bronx, Queens and Brooklyn are large compared to Staten Island, which means that their results will not be prone to as much sampling error as Staten Island. With Outliers in all 5 boroughs, the outlier in Staten Island significantly increases the average SAT score in that Borough. However, regardless of that Borough we still see that the mean of the remaining public schools is still higher than others.

Future Steps

With our initial anaylsis we see the spread of SAT scores, the distribution and statistics of SAT scores in each borough and the divide in ethinic groups. In future steps we will further find correlations between these variables and a develop more comprehensive and elaborative models that will help in answering our question.

Project 2

The Message

Our main aim in this project is to assess whether some boroughs in NYC have a higher SAT score than others. As a result, we want to develop visual graphs and analyse data that show us how certain boroughs benefit from higher SAT scores which plays an essential part in college application process. With the current analysis, our message is to show how certain areas such as Staten Island and Queens have 20% higher SAT scores than others boroughs. As a result, our main message is that some boroughs in nyc have a much higher average SAT score compared to others.

In [18]:
```
df5 = pd.DataFrame({
'Borough':['Queens', 'Manhattan','Brooklyn', 'Bronx', 'Staten Island'],
'Latitude': [40.711111, 40.753210, 40.652112, 40.853829, 40.578410],
'Longitude': [-73.820001, -73.997860, -73.98111, -73.906455,-74.20010]
})
df5["Coordinates"] = list(zip(df5.Longitude, df5.Latitude))
df5["Coordinates"] = df5["Coordinates"].apply(Point)
gdf = gpd.GeoDataFrame(df5, geometry="Coordinates")
gdf.head()
```

Out[18]:

	Borough	Latitude	Longitude	Coordinates
0	Queens	40.711111	-73.820001	POINT (-73.82000 40.71111)
1	Manhattan	40.753210	-73.997860	POINT (-73.99786 40.75321)
2	Brooklyn	40.652112	-73.981110	POINT (-73.98111 40.65211)
3	Bronx	40.853829	-73.906455	POINT (-73.90645 40.85383)
4	Staten Island	40.578410	-74.200100	POINT (-74.20010 40.57841)

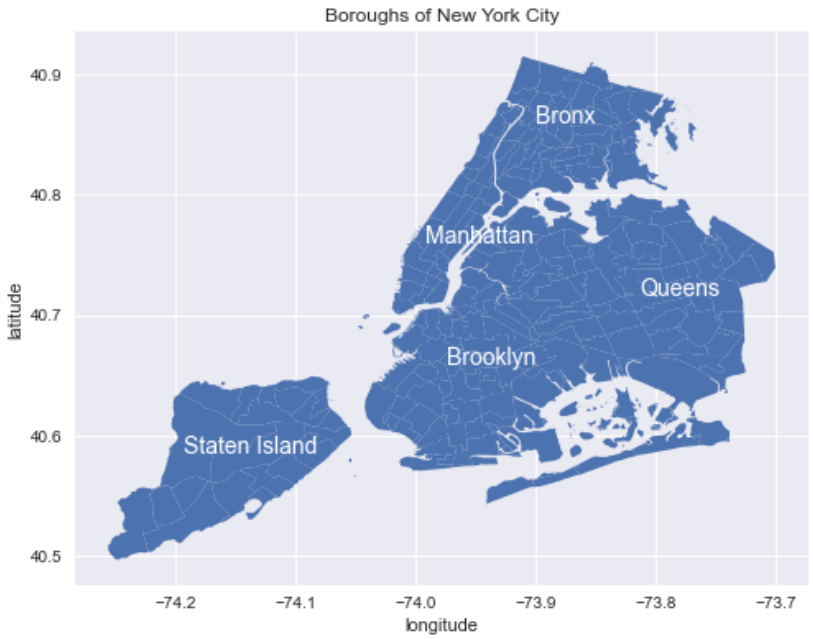
In [19]:
```
nbhoods = pd.read_csv(r'C:\Users\Anwar Malik\Downloads\nynta (3).csv')
nbhoods.head(5)
```

Out[19]:

	the_geom	BoroName	BoroCode	CountyFIPS	NTACode	NTAName	Shape_Leng	Shape_Area
0	MULTIPOLYGON (((-73.80379022888246 40.77561011...	Queens	4	81	QN51	Murray Hill	33,266.9048559	52,488,277.4492
1	MULTIPOLYGON (((-73.86109724335655 40.76366447...	Queens	4	81	QN27	East Elmhurst	19,816.7118942	19,726,845.691
2	MULTIPOLYGON (((-73.77757506882061 40.73019327...	Queens	4	81	QN41	Fresh Meadows-Utopia	22,106.4312724	27,774,853.5522
3	MULTIPOLYGON (((-73.97301487176121 40.76427887...	Manhattan	1	61	MN17	Midtown-Midtown South	27,032.7003748	30,191,534.2409
4	MULTIPOLYGON (((-73.88063708265133 40.81852042...	Bronx	2	5	BX09	Soundview-Castle Hill-Clason Point-Harding Park	67,340.9776258	51,983,797.3364

In [20]:
```
#Then, since this is a csv file, convert the geometry column text into well known text, this will allow yo
nbhoods['geom'] = nbhoods['the_geom'].apply(wkt.loads)

#Now convert the pandas dataframe into a Geopandas GeoDataFrame
nbhoods = gpd.GeoDataFrame(nbhoods, geometry='geom')
```

In [21]:
```
fig,ax = plt.subplots(1,1, figsize=(8,8))

plt.title("Boroughs of New York City")
plt.xlabel('longitude')
plt.ylabel('latitude')
nbhoods.plot(ax=ax)


for x, y, label in zip(gdf['Coordinates'].x, gdf['Coordinates'].y, gdf['Borough']):
    ax.annotate(label, xy=(x,y), xytext=(5,5), textcoords='offset points', color = 'white')
```

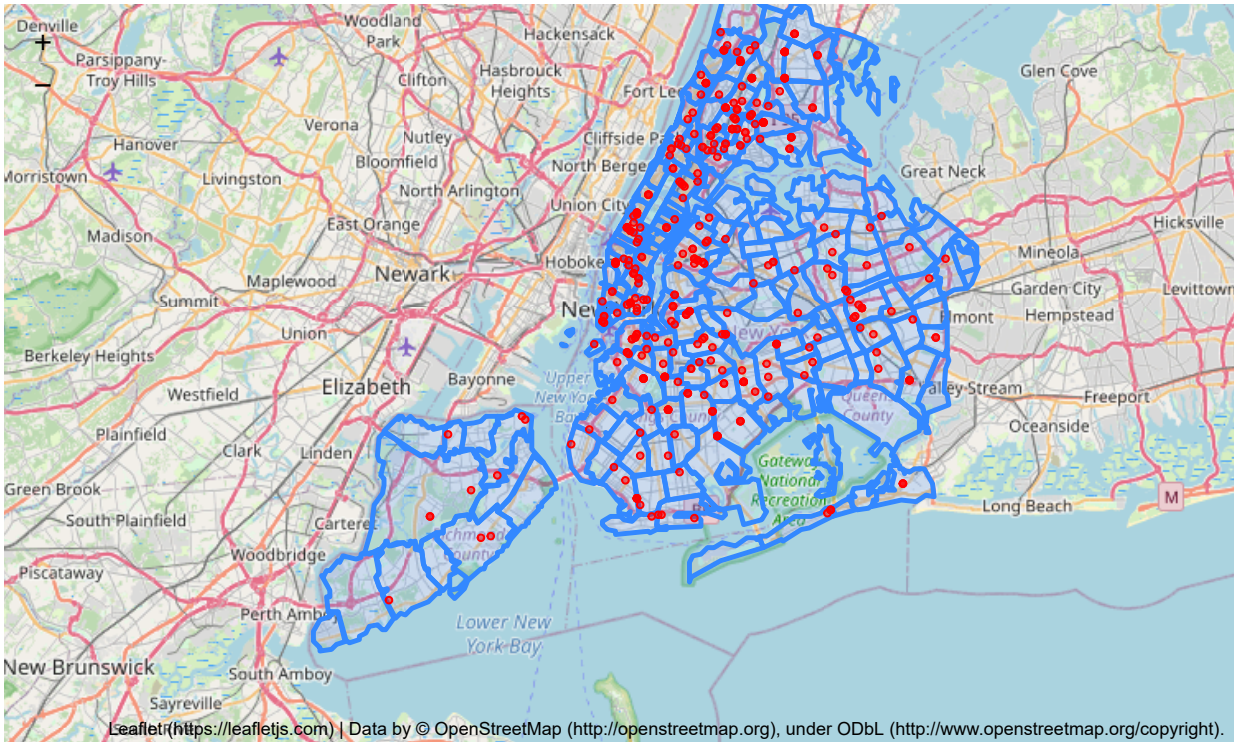


The graph above gives us an understanding of one of our X-Variable which is the borough in NYC. We see how New York is divided into 5 distintive boroughs. Each borough has a sample of highschools which will be used to see whether SAT scores in them vary. We will also see the ethnicity difference in these boroughs to see whether areas with certain ethnicity have higher/lower SAT scores.

In [22]:
```
nycArea = json.load(open(r'C:\Users\Anwar Malik\Downloads\zip_code_040114.geojson', 'r'))
```

In [23]:
```
nycMap = folium.Map(location= [40.730610, -73.935242], width = '100%', height = '100%',
                    min_lat=40, max_lat=75, tiles='openstreetmap', zoom_start = 10)


#add the shape of NYC to the map
folium.GeoJson(nycArea).add_to(nycMap)

#for each row in the dataset, plot the corresponding latitude and longitude on the map
for i,row in df2.iterrows():
    folium.CircleMarker((row.Latitude,row.Longitude), radius=1.5, weight= 1.2, color='red', \
                        fill_color='red', fill_opacity=.5).add_to(nycMap)

nycMap
```

Out[23]:



This hover satellite graph shows us more specific x-values from our data main x-value which is the borough. The graph above shows each individual highschool (red dots) in our data according to our sample. This is essential since we will use the Average SAT Scores from these highschools to further determine which borough has a higher average SAT score.

In [24]:
```
df2.sort_values(by=['Average_SAT_Score'], ascending = False, inplace = True)
df2[['Average_SAT_Score', 'School Name','Zip Code', 'Borough']].head(7)
```

Out[24]:

	Average_SAT_Score	School Name	Zip Code	Borough
105	2144.0	Stuyvesant High School	10282	Manhattan
203	2041.0	Bronx High School of Science	10468	Bronx
110	2041.0	Staten Island Technical High School	10306	Staten Island
208	2013.0	High School of American Studies at Lehman College	10468	Bronx
385	1981.0	Townsend Harris High School	11367	Queens
424	1947.0	Queens High School for the Sciences at York Co...	11433	Queens
7	1914.0	Bard High School Early College	10002	Manhattan

Project 3

Data that can enhance my Research

One important detail that we can think about is comparing the SAT scores of public and private schools in these boroughs. Since currently our data is only based on public schools. Adding and comparing this with data from private schools will help us understand whether only public schools experience differences in scores in each borough or do private schools have similar results. We can also compare public and private schools to see whether one has on average higher SAT scores compared to the other.

The following website has SAT scores information of both public and private schools in New York City: https://www.privateschoolreview.com/sat-score-stats/new-york/high

This data hence enhances our paper as it incorporates private schools too. Further we can merge this data with our current data by merging on the zip code and filtering only those zip codes that contain both a public and private school in their area. This will help us to compare SAT scores in public and private schools more affectively and reducing any other uncontrollable factor. This resource will help me not only answer one part of my research questions that some boroughs score higher than others, but will open up a new door towards comparing public and private schools in these areas.

Another data that will be extremely useful is the population data of each borough in new york city. A link to this data is as follows: https://worldpopulationreview.com/zips/new-york

This data contains the population of each of our neccassary zip codes which will help us to make an inference whether some borough that have more number of people may have a higher number of highschools, may expereince more diversity in ethnicity and hence more variability in SAT scores hence enhancing our research.

I would need to run this program yearly since we will compare annual SAT scores. Further scraping this will require a lot of skill and further learning since the data in the form of an interactive table and its code does not have any table element which makes it difficult to scrape it at this time. Future learning will be required to understand data that are not under the class of table. My other data set on populations however is something that I will be able to scrape and use it in my data. This data is collected through census hence it too will be collected annualy.

In [25]:
```
import requests
import pandas as pd
from bs4 import BeautifulSoup
web_url = 'https://worldpopulationreview.com/zips/new-york'
response = requests.get(web_url)
```

Here we are calling on our URL to start the process of scraping.

In [26]:
```
soup_object = BeautifulSoup(response.content)
data_table = soup_object.find_all('table', 'jsx-2744221037 table table-striped tp-table-body')[0]
all_values = data_table.find_all('tr')
```

We call on the neccassry table through inspecting the HTML code and put the relevant information in our variable all_values

In [27]:
```
nyc_borough = pd.DataFrame(columns = ['Zip_Codes', 'Population']) # Create an empty dataframe
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
    Zip_Codes = values[0].text
    Population = values[3].text

    nyc_borough.loc[ix] = [Zip_Codes, Population] # Store it in the dataframe as a row
    ix += 1

# Print the first 5 rows of the dataframe
nyc_borough.head()
```

Out[27]:

	Zip_Codes	Population
0	11368	112,088
1	11385	107,796
2	11211	103,123
3	11208	101,313
4	10467	101,255

Here we scrape the data using a for loop. We only take the relevant columns of the table which our Zip Codes and Populations.

In [28]:
```
df2.rename(columns={'Zip Code': 'Zip_Code'}, inplace=True)
```

In [29]:
```
nyc_pop = pd.concat([df2, nyc_borough], axis=1, join="inner")
nyc_pop = nyc_pop.dropna()
```

In [30]:
```
nyc_pop['Zip_Code'] = nyc_pop['Zip_Code'].astype(int)
```

Since this dataset may contain some zip codes which we do not have SAT information of, we are dropping them and only using the Zip Codes that are relevant to our data. We also merge our new data set with our original data set. Furthermore, we are converting are zip codes to type int.

In [31]:
```
nyc_pop['Borough'].value_counts()
```

Out[31]:
```
Brooklyn         109
Bronx             98
Manhattan         89
Queens            68
Staten Island     10
Name: Borough, dtype: int64
```

In [32]:
```
nyc_df = nyc_pop[['Zip_Codes', 'Borough','Population']]
```

We make a new dataframe which only consists of certain columns from our main dataframe.

In [33]:
```
nyc_df = nyc_df.replace(',','', regex=True)
```

Here we remove the comma from the Population to help us do anaylsis on the data.

In [34]:
```
nyc_df["Population"] = nyc_df["Population"].astype(str).astype(int)
```

In [35]:
```
nyc_pop_avg = nyc_df.groupby(['Borough']).mean()
```

Changing Population from type Object to type Int. We then then group each population in each zip code by borough and

take the mean.

In [36]:
```
data = {"Borough":["Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island"],

        "Avg Population in each Zip Code":[34091,20540,66723,17725,42935]

        };
```

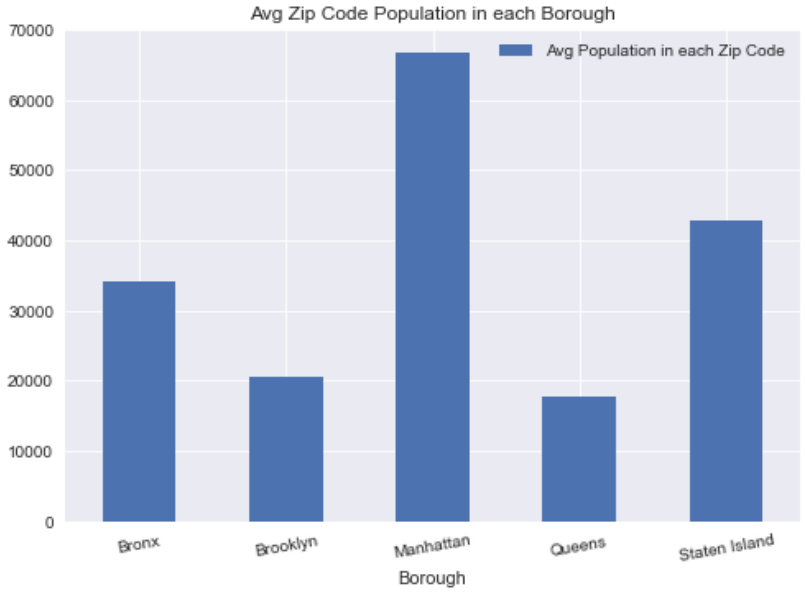
In [37]:
```
nyc_avg_pop = pd.DataFrame(data=data)
nyc_avg_pop
```

Out[37]:

	Borough	Avg Population in each Zip Code
0	Bronx	34091
1	Brooklyn	20540
2	Manhattan	66723
3	Queens	17725
4	Staten Island	42935

We create a new dataframe with rounded values of population and 2 keys: Borough and Avg Population

In [38]:
```
import matplotlib.pyplot as plot


nyc_avg_pop.plot.bar(x="Borough", rot=10, title="Avg Zip Code Population in each Borough");
plot.show(block=True);
```



Finally we plot a bar chart of our new merged data, where we find the average population of each of the zip code where we have a public highschool. Our results show that the Manhattan has the highest population density in each Zip Code followed by Staten Island. Queens and Brooklyn have the lowest Population density in each Zip Code.

Our analysis on average Zip Code population in each borough and Average SAT scores in each borough give us an interesting correlation, as we see public highschools with in higher population densitiy areas to have higher average SAT scores. This is an interesting finding as it shows correlation between two of our very important variables which we can further develop upon in future research.

OLS Regression

When we analyse are dependent and independent variables which are Boroughs and Average SAT Scores Respectively. Since there is only a limited number of highschools in each borough and the variation of SAT scores of each borough is 100 greater or lower, as a result we are likely to see a more linear relation between the boroughs and Average SAT score.

Since most of our analysis is base on SAT scores of each borough, it is essential for us to run our regression on boroughs to see whether SAT scores in each deviate or not. Since we can see differences in SAT scores, and our research questions is

about whether there are significant changes in scores for each borough, hence it is vital that we take boroughs as our x-variable and use that to determine changes in average SAT score. In other other regressions we have added ethnicity which helps us indicate if being from different ethinic group can increase or decrease the overall SAT score in the highschool. We also analyse Reading, Writing and Math scores since they are key components of SAT and can give us a better understanding to why some highschools may have a higher SAT.

In [39]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import pandas as pd
import statsmodels.api as sm
from statsmodels.iolib.summary2 import summary_col
from linearmodels.iv import IV2SLS
```

In [40]:
```
df2['const'] = 1
```

In [41]:
```
df3 = pd.get_dummies(df2, columns=['Borough'], drop_first=True)
```

In [42]:
```
df2['Student Enrollment']
```

Out[42]:
```
105    3296.0
203    3015.0
110    1247.0
208     376.0
385    1132.0
        ...
204     461.0
389     378.0
342     414.0
289     229.0
217     428.0
Name: Student Enrollment, Length: 374, dtype: float64
```

In [43]:
```
reg1 = sm.OLS(endog=df3['Average_SAT_Score'], exog=df3[['const', 'Borough_Brooklyn', 'Borough_Manhattan',
                                                        'Borough_Staten Island']], \
    missing='drop')
type(reg1)
```

Out[43]:
```
statsmodels.regression.linear_model.OLS
```

In [44]:
```
results = reg1.fit()
type(results)
```

Out[44]:
```
statsmodels.regression.linear_model.RegressionResultsWrapper
```

$$AvgSATScore_i = \beta_0 + \beta_1 Brooklyn_i + \beta_2 Manhattan_i + \beta_3 Queens_i + \beta_4 StatenIsland_i + u_i$$

The OLS equation above tells helps us analyse whether Average SAT score differs in each borough. Here each borough is a dummy variable where 1 is if they are for example Brooklyn and 0 other wise for β_1. Same is applied to all other boroughs except for Bronx which is our omitted category and is β_0.

In [45]:
```
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Average_SAT_Score   R-squared:                       0.120
Model:                            OLS   Adj. R-squared:                  0.110
Method:                 Least Squares   F-statistic:                     12.54
Date:                Sat, 16 Apr 2022   Prob (F-statistic):           1.41e-09
Time:                        15:34:29   Log-Likelihood:                -2478.2
No. Observations:                 374   AIC:                             4966.
Df Residuals:                     369   BIC:                             4986.
Df Model:                           4
Covariance Type:            nonrobust
===============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
```

```
--------------------------------------------------------------------------------
const                  1202.7245     18.569     64.771      0.000    1166.210    1239.239
Borough_Brooklyn         27.5324     25.589      1.076      0.283     -22.787      77.852
Borough_Manhattan       137.4103     26.916      5.105      0.000      84.482     190.339
Borough_Queens          140.7020     29.013      4.850      0.000      83.651     197.753
Borough_Staten Island   236.2755     61.024      3.872      0.000     116.278     356.273
==============================================================================
Omnibus:                      144.020   Durbin-Watson:                   0.252
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              480.522
Skew:                           1.757   Prob(JB):                    4.53e-105
Kurtosis:                       7.300   Cond. No.                         7.19
==============================================================================

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

$$AvgSATScore_i = 1202.72 + 27.5Brooklyn_i + 137.4Manhattan_i + 140.7Queens_i + 236.28StatenIsland_i + u_i$$

Our OLS equation adds the coeffiecients to each of our Beta. Each coefficient (which is a dummy) gives us the how high on average SAT scores in each borough is compared to Bronx. Here, we can run an example to see our results.

- If we are considering the Borough Queens, we will use 1 for the variable Queens and 0 for the others.
- This will give us an AvgSATScore of 1202.7 + 27.5(0) + 137.4(0) + 140.7(1) + 236.28(0) = 1343
- We can interpret this as, We are 99% confident that after controlling for the different boroughs, Queens on average has a SAT score that is 140 points higher than Bronx. With the Use of the dummy we get an Avg SAT Score of 1343 in Queens.

$$AvgSATScore_i = \beta_0 - \beta_1 Black_i - \beta_2 White_i - \beta_3 Hispanic_i - \beta_4 Asian_i + u_i$$

In [46]:

```
reg2 = sm.OLS(endog=df3['Average_SAT_Score'], exog=df3[['const', 'Black', 'White', 'Hispanic', 'Asian']],
    missing='drop')
type(reg2)

results1 = reg2.fit()
type(results1)

print(results1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Average_SAT_Score   R-squared:                       0.603
Model:                            OLS   Adj. R-squared:                  0.599
Method:                 Least Squares   F-statistic:                     140.2
Date:                Sat, 16 Apr 2022   Prob (F-statistic):           9.85e-73
Time:                        15:34:29   Log-Likelihood:                -2329.2
No. Observations:                 374   AIC:                             4668.
Df Residuals:                     369   BIC:                             4688.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3327.8053    387.312      8.592      0.000    2566.191    4089.420
Black        -21.7509      3.960     -5.493      0.000     -29.537     -13.964
White        -15.7008      4.057     -3.870      0.000     -23.679      -7.723
Hispanic     -22.2907      3.904     -5.710      0.000     -29.968     -14.614
Asian        -16.2967      4.013     -4.061      0.000     -24.189      -8.405
==============================================================================
Omnibus:                       14.817   Durbin-Watson:                   1.233
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               28.894
Skew:                           0.184   Prob(JB):                     5.32e-07
Kurtosis:                       4.311   Cond. No.                     3.61e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.61e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

$$AvgSATScore_i = 3327.81 - 21.75Black_i - 15.7White_i - 22.29Hispanic_i - 16.29Asian_i + u_i$$

Our OLS equation adds the coeffiecients to each of our Beta. Each coefficient is the affect of adding 1% of each ethinic group on our Average SAT score. Through observation we can see that a 1% increase in Hispanic and Black gives us the

highest reduction in SAT score where the AvgSATScore decreases by 22.3 and 21.8 respectively. Here our intercept 3327.8 has no interpretation on its own. Here, we can run an example to see our results.

- Consider a highschool with 35% White, 20% Black, 30% Asian, 15% Hispanic, then:
- This will give us an $AvgSATScore_i = 3327.81 - 21.75 * 20 - 15.7 * 35 - 22.29 * 15 - 16.29 * 20 = 1682.8$
- We can interpret this as, a highschool with 35% White, 20% Black, 30% Asian, 15% Hispanic students, the average SAT score will approximately be 1682.8 points.

$AvgSATScore_i = \beta_0 - \beta_1 Black_i - \beta_2 White_i - \beta_3 Hispanic_i - \beta_4 Asian_i + \beta_5 Brooklyn_i + \beta_6 Manhattan_i + \beta_7 Qu$

In [47]:
```
reg10 = sm.OLS(endog=df3['Average_SAT_Score'], exog=df3[['const', 'Black', 'White', 'Hispanic', 'Asian','B
                                                         'Borough_Staten Island']], \
    missing='drop')
type(reg10)

results20 = reg10.fit()
type(results20)

print(results20.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Average_SAT_Score   R-squared:                       0.643
Model:                            OLS   Adj. R-squared:                  0.635
Method:                 Least Squares   F-statistic:                     82.05
Date:                Sat, 16 Apr 2022   Prob (F-statistic):           7.70e-77
Time:                        15:34:29   Log-Likelihood:                -2309.6
No. Observations:                 374   AIC:                             4637.
Df Residuals:                     365   BIC:                             4673.
Df Model:                           8
Covariance Type:            nonrobust
============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                   2960.0896    380.298      7.784      0.000    2212.240    3707.939
Black                    -17.2306      3.903     -4.414      0.000     -24.907      -9.555
White                    -10.9388      4.031     -2.714      0.007     -18.866      -3.012
Hispanic                 -18.7787      3.822     -4.913      0.000     -26.295     -11.263
Asian                    -11.8648      3.917     -3.029      0.003     -19.567      -4.162
Borough_Brooklyn         -74.7913     19.612     -3.814      0.000    -113.358     -36.224
Borough_Manhattan         22.5981     17.973      1.257      0.209     -12.746      57.942
Borough_Queens           -73.4481     21.609     -3.399      0.001    -115.941     -30.955
Borough_Staten Island   -130.4489     46.354     -2.814      0.005    -221.604     -39.294
==============================================================================
Omnibus:                       14.231   Durbin-Watson:                   1.328
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               30.256
Skew:                           0.111   Prob(JB):                     2.69e-07
Kurtosis:                       4.376   Cond. No.                     3.72e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.72e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Our OLS equation adds the coeffiecients to each of our Beta. Each coefficient is the affect of adding 1% of each ethinic group on our Average SAT score and/or to which borough our data belongs to. Through observation we can see that a 1% increase in Hispanic and Black gives us the highest reduction in SAT score where the AvgSATScore decreases by 18.8 and 17.7 points respectively. Here our intercept 2960 has no interpretation on its own. Here, we can run an example to see our results.

- Consider a highschool with 35% White, 20% Black, 30% Asian, 15% Hispanic, in Brooklyn then:
- This will give us an $AvgSATScore_i = 2960 - 17.3 * 20 - 10.94 * 35 - 18.8 * 15 - 11.86 * 30 - 74.7 * 1 - 22.6 * 0 - -73.4 * 0 - 13$
- We can interpret this as after controlling for boroughs and ethnicity, a highschool in brooklyn with 35% White, 20% Black, 30% Asian, 15% Hispanic students, the average SAT score will approximately be 1517 points.

In [48]:
```
reg4 = sm.OLS(endog=df3['Average_SAT_Score'], exog=df3[['const', 'Student Enrollment', 'Percent Tested']],
```

```
    missing='drop')
type(reg4)

results3 = reg4.fit()
type(results3)

print(results3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Average_SAT_Score   R-squared:                       0.471
Model:                            OLS   Adj. R-squared:                  0.468
Method:                 Least Squares   F-statistic:                     165.2
Date:                Sat, 16 Apr 2022   Prob (F-statistic):           4.95e-52
Time:                        15:34:29   Log-Likelihood:                -2382.9
No. Observations:                 374   AIC:                             4772.
Df Residuals:                     371   BIC:                             4784.
Df Model:                           2
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                      835.4946     26.614     31.394      0.000     783.162     887.827
Student Enrollment           0.0818      0.010      8.540      0.000       0.063       0.101
Percent Tested               5.8388      0.393     14.848      0.000       5.066       6.612
==============================================================================
Omnibus:                      108.071   Durbin-Watson:                   0.866
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              314.587
Skew:                           1.336   Prob(JB):                     4.88e-69
Kurtosis:                       6.613   Cond. No.                     3.92e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.92e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

$$AvgSATScore_i = 835.5 + 0.08StudentEnrollment_i + 5.84PercentTested_i + u_i$$

Our OLS equation adds the coeffiecients to each of our Beta. The regression shows that after controlling for Student Enrollment, a 1% increase in percent tested increases the average SAT score by 5.84 points on average. Here our intercept 835.5 has no interpretation on its own since as no school will have 0 student enrollment and no school will have 0% student tested. Here, we can run an example to see our results.

- Consider a highschool with 1200 students and 80% tested then:
- This will give us an $AvgSATScore_i = 835.5 + 0.08 * 1200 + 5.84 * 80 = 1398.7$
- We can interpret this as, a highschool with 1200 students and 80% students tested, the average SAT score will approximately be 1398.7 points.

Justification

Running these 4 regressions was essential for our research and analysis, as we needed to see how our Dependent Variable would vary with different explanotary variables. Here our first regression focuses mainly on the affect of Boroughs on the Average Score, showing variations among boroughs. Our second Regression takes into account ethnicity and how each ethenic group affects the Average SAT score in each highschool. This gives us a relation between how ethnicity is linked with SAT scores. Our third regression combines our first and second to give us a better understanding of how controlling for both ethnicity and borough could change our results. Our last regression uses student enrollment in highschool and percentage of students opting for the SAT exam which gives us a relation between how enrollment and SAT attempts are related.

Perferred Specification

One regression that is extremely interesting is our third regression that takes in to account ethnicity and borough when evaluating average SAT scores. The regression is intriguing as it shows a relation between how different ethinic group in each borough affects our average scores. We see that hispanic and black population have a negative relation to SAT scores as an increase in those ethnicity generally leads to a lower SAT score in that particular highschool. We also notice that Manhattan is the only Borough that after controlling for ethnicity would have a positive affect on Avg SAT scores.

Evaluating Regression

As we evaluate are regressions, we can see certain key aspects that can help us understand if our regressions are

statistically and economically significant. When we look at our regression for ethnicity (regression 2) and combined ethnicity and borough (regression 3) we see a very high R^2 which indicates that a high number of variations in Avg SAT score can be explained through variations in Ethnicity. The F-Statistic is also very high with a very small P-Value indicating our model is statistically significant overall. Similar results can be seen in our regressions for Percent tested and studend enrollment where we again see a high f-test and a 0.50 R^2 which shows that these are relatively good indicators of average SAT score. Regression 1 on the other hand show low R^2 which means that it is not doing a good job in explaining variation in SAT score. Hence, here we can make an inferenance and reiterate our initial claim of how ethnicity plays a vital role in student SAT scores.

Analysis

Our general analysis from running these regression particularly shows that on average highschools in Bronx tend to have the lowest SAT Score, while an increase percentage of White and Asian population shows an increase in SAT score in that highschool. Another interesting finding is that though Staten Island has the highest avearage SAT score, however after controlling for ethnicity and borough, we see that it actually does not infer positive affect towards SAT score. We further observe that high white and Asian populated schools in manhattan will tend to have the highest SAT scores. This is a key finding as it indicates the importance of student race and its proportion in a high school to explain the SAT scores.

In [49]:
```
X = df2.drop(["Street Address", "School ID", "Building Code", "School Name", "City", "Borough", "State", "
# convert everything to be a float for later on
for col in list(X):
    X[col] = X[col].astype(float)
X.head()
```

Out[49]:

	Latitude	Longitude	White	Black	Asian
105	40.71775	-74.01405	20.4	0.8	73.4
203	40.87706	-73.88978	22.1	2.6	62.8
110	40.56791	-74.11536	52.2	1.0	41.1
208	40.87126	-73.89752	50.8	4.0	23.1
385	40.73441	-73.82142	21.9	6.3	58.6

Regression Tree

Objective Function

$$\min_{AvgSATScore \subset S} \sum (\hat{f}\,(x) - y)^2 + \alpha|\text{Ethnicity}|$$

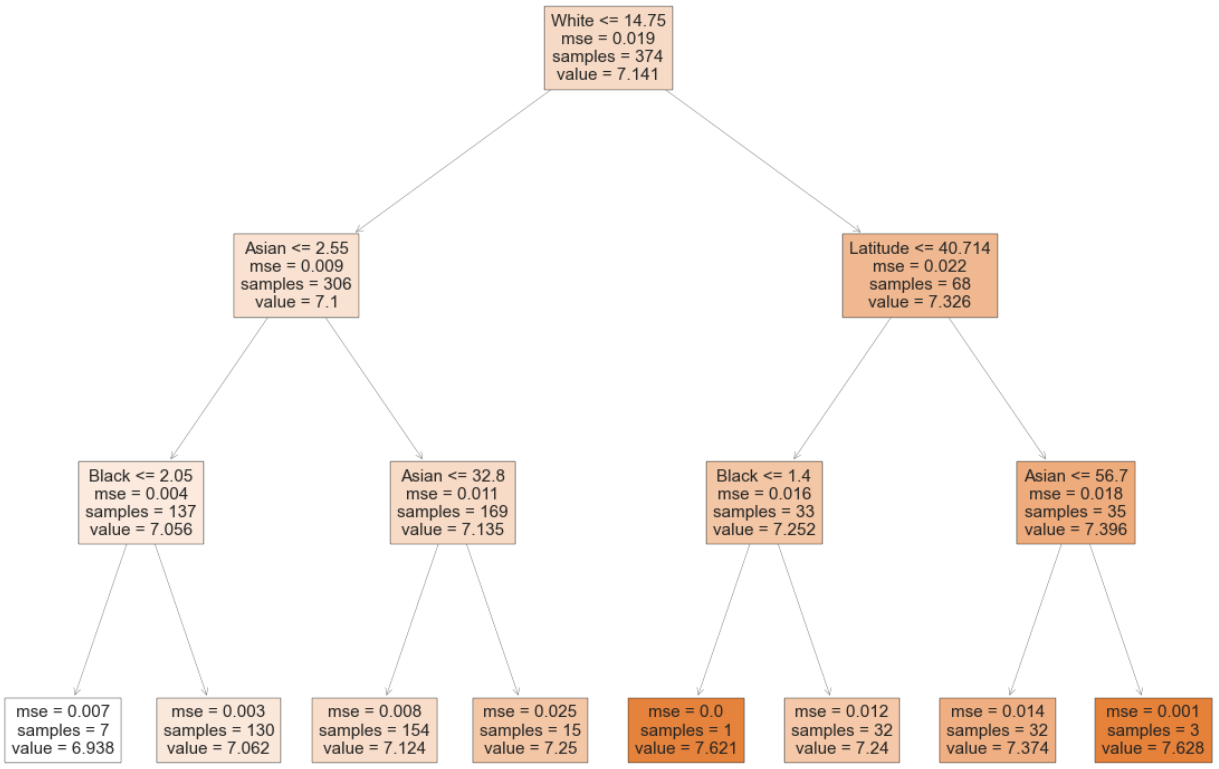
- Where $\hat{f}\,(x)$ is predicted SAT Score
- y is actual SAT Score
- The location (latitude and longitude) and α serve as our regularization parameters.

Our regularization parameters are mainly ethncity and location. We dropped some key numeric parameters such as Writing, Reading and Math, mainly because they give us a very uninteresting result where an increase in one componenet will result in an increase in overall SAT score. By discarding those we will be able to find a model that interprets ethinic groups and how they may affect SAT scores. Hence, Black, Asian, Hispanic and White, latitude and longitude are our regularization parameters

In [50]:
```
y = np.log(df2["Average_SAT_Score"])
df["log_sat"] = y
```

In [51]:
```
from sklearn import tree
sqft_tree = tree.DecisionTreeRegressor(max_depth=3).fit(X,y)
```

In [52]:
```
sqrf_fig = plt.figure(figsize=(20,15))
sqrf_fig = tree.plot_tree(sqft_tree, feature_names=X.columns, filled=True)
```



Our regression tree gives us very similar finidngs to our regression 2 where we interpret how controlling for ethnicity changes Avg SAT scores, Our regression tree shows us how a school with high white >= 14.75% generally gives us a higher SAT score. Low percent of Black and high percent of Asian also indicates higher SAT score which is similar to our findings as before since increase in black in our regression 2 results the highest decrease in SAT score while an increase in white and asian causes the least. This hence reitereates the idea that SAT scores are widely affected by ethnicities and some play a much more major role than others.

The regression tree adds to our multiple regressions as it reportd a combination of location and ethnicity while also giving the sample size of case along with the mean squared error (mse). These extra details for each of our 8 different cases help us to further see which combination of location and ethnicity plays a much more important role in determing Avg SAT Score.

Our MSE or mean squared errors are extremely small. This means that the average difference between the actual SAT scores and the predicted score is very small especially for our results that show higher white/asian and low black porportioned highschools, the MSE is small indicating that it is doing a good job in estimating SAT scores.

Conclusion

With our statistical and visual analysis, we have tried to analyze whether SAT scores vary over boroughs in NYC and if students of certain ethnicity tend to do better than others. With our initial analysis, we see that borough such as Staten Island gets 20% higher SAT scores on average compared to the Bronx. Such a finding is very interesting and important since it emphasizes how going to school in certain boroughs may help children be better prepared for SAT which will affect their future goals.

As our main question was to see whether borough or ethnicity play a key role in variations in SAT scores, we saw some interesting results where our regression and regression tree both complimented our claim that schools with higher white and Asian population tend to do much better than schools with the higher Black student population. This idea can also be interpreted through boroughs where we saw boroughs with a higher proportion of white and Asian populations such as Staten Island and Manhattan did considerably better than Brooklyn and Bronx which had higher black/Hispanic populations. However, some of our main findings tend to go against our initial claims, as after we controlled for region and ethnicity, we saw how Staten Island tends to reduce our SAT scores. This finding could be due to sampling error or lurking unobserved variables since our sample size for Staten Island was very small. However, we did see some of our

initial claims to be complemented by our regressions as high black and Hispanic proportions in Brooklyn and Bronx regions did show lowest SAT scores while majority white and Asian dominated regions such as Manhattan showed higher average SAT scores. These results are specifically interesting as prior papers such as Rebecca Zwick and Jeffrey C. Sklar's 2005 paper on ethnicity compared the role of first language and ethnicity on high school success and grades but did not do a comparative analysis of whether regions and ethnicity combine play a key role in variations in SAT scores.

As we still build on our project to further find correlations, it is still very interesting to see such high differences in tests within a single city. With such variability, one can only imagine how SAT scores may differ in different cities, states, and even countries. Furthermore, it is hard to determine a single factor for these results, whether it is schooling, ethnicity, or natural factors.

As we move forward, adding and incorporating the same data for private schools in NYC would help us add a different dimension and make us analyze if these differences in ethnicities and boroughs are only present in public schooling or also private. These data are not as easily available currently but will be something interesting to add and build upon in the near future.