W200 Project 2 Report
New York City Education and Crime Patterns

Alexandra Drossos, Hassan Saad, Karina Tsang
GitHub Repo: Project2_Drossos_Saad_Tsang

1: Introduction

In 2020, COVID forced new dynamics within populations that had never been seen before. Record breaking unemployment rates, evictions, and poverty spread throughout the country. New York City, with the highest population density in the country, felt the effects of this more dramatically. The country watched as crime rates skyrocketed in 2020. As data analysts, a natural question to ask is, why did this happen and what can we do to remediate the situation? City politicians, sociologists, and federal government representatives would all be keen on the answers to these questions, so we plan to help provide them with solutions. To do this, we will be analyzing data for the state of New York around high school graduation rates and income to see the impact they have on crime rates, specifically NYPD shooting incident rates. This overall casual relationship will be picked apart by addressing six sub research questions to develop an overall conclusion.

2: Data Sources

To help answer this question, we will have to pull in multiple data sources to join together. To get accurate graduation rate data, we will pull from the New York State Education Department's website (see links below). They are mandated by the state to provide a plethora of datasets around education, one of which being data around high school enrollment, graduation, and dropout rates for each year. As we'll dive into later on, this mandate was only established in 2015, so that's as far back as the data goes. While we do wish that we had more data points to look at, the resulting data points still do provide interesting, actionable insights.

Secondly, for our outcome data, we will pull in NYPD Shooting incident data. This is pulled from data.gov so to the extent that the NYPD is accurately reporting their shootings, we can trust the quality and the validity of the data we're working with. This dataset goes back all the way to 2006, but because of our limit on the graduation rate data, we will only be pulling values between 2015 and 2020.

The last data source we implement as an independent variable is New York Median Income by County, which we acquired from the U.S. Bureau of Economic Analysis. We dig into a few caveats surrounding this data when we answer question 5 below. It's important to make sure that we uncover independent groupings and clusters that may be buried in our previous data sets. By introducing this third variable, we can attach another explanation to some of the correlations we model through our first four questions. For example, if certain counties have significantly higher dropout rates, it would also be relevant to study if students within these communities are at an inherent disadvantage due to financial burdens.

Links:

NYC Public School Graduation Rate Data between 2015-2020

NYPD Shooting Incident Data from 2015-2020

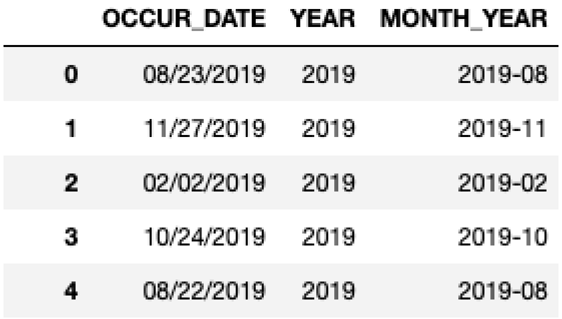
NY County Income Data from 2017-2019, listed on the U.S. Bureau of Economic Analysis Website

3: Questions

Question 1

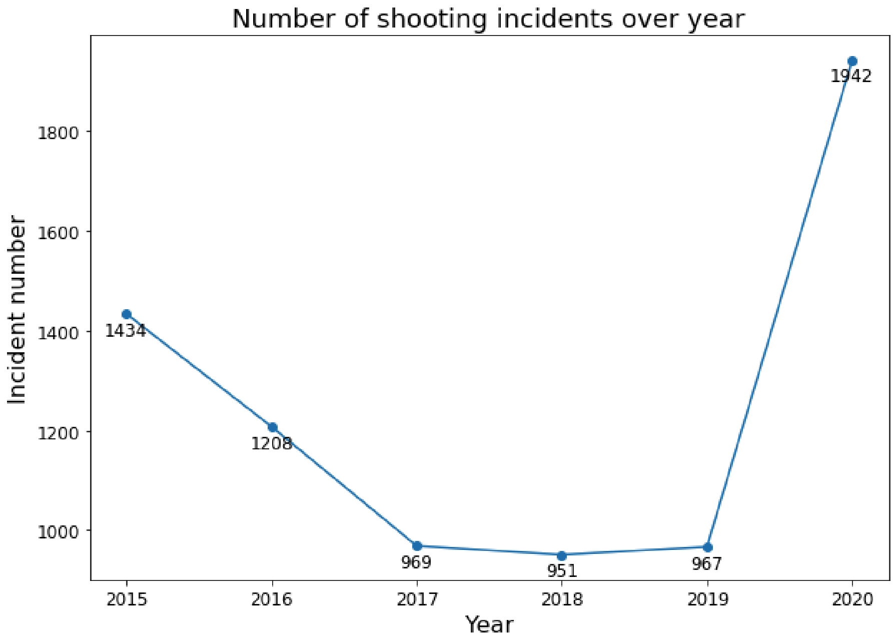
To get a better understanding of shooting incidents in New York City, we first investigate *the overall trend of shootings year over year in New York City*. In order to do so, we will be using the historical NYPD Shooting Incident Data. This data is extracted quarterly from 2006 to 2020. However, for the purpose of comparing with the graduation rate data, we will be looking at the data collected from 2015 to 2020.

To analyze trends, we first have to look at the date data which is OCCUR_DATE. It is structured in a MM/DD/YYYY format. For a more high level understanding of patterns, day to day data might be too up close. Therefore, we first create a YEAR column derived from the OCCUR_DATE column using the datetime method in Pandas. To provide slightly more granularity, we also created a MONTH_YEAR column which will present the month and year data and extracted with a similar method. As mentioned above, we are only looking at incidents from 2015 to 2020. Therefore, we slice the dataframe into the relevant data with the newly created YEAR column. This gives us the base data to answer our first research question. Below shows the OCCUR_DATE data and the derived the YEAR and MONTH_YEAR data:



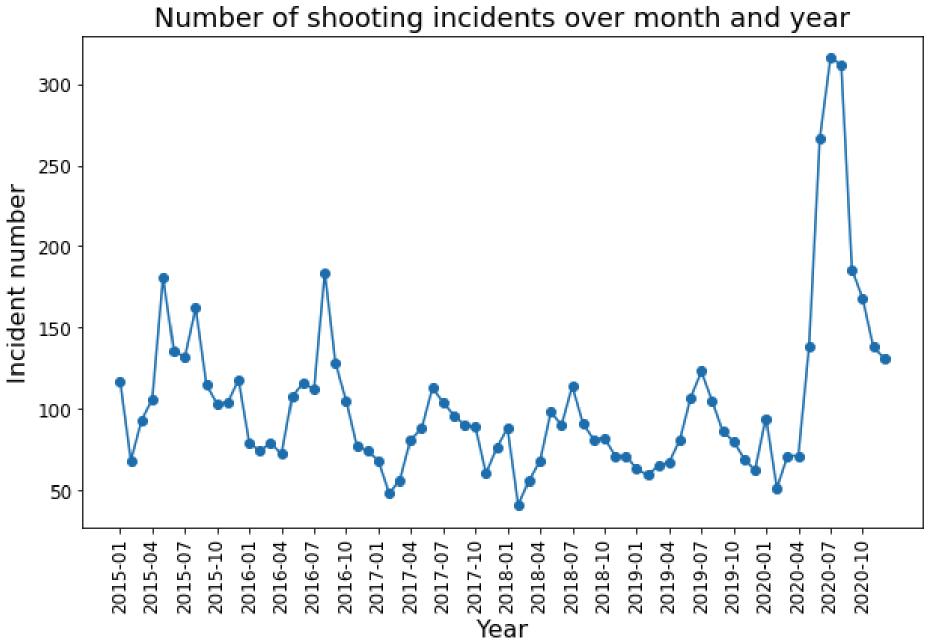
	OCCUR_DATE	YEAR	MONTH_YEAR
0	08/23/2019	2019	2019-08
1	11/27/2019	2019	2019-11
2	02/02/2019	2019	2019-02
3	10/24/2019	2019	2019-10
4	08/22/2019	2019	2019-08

To look at trends by year, we use the groupby function to aggregate sum of incident count by the year column. This sum of incident data is then plotted against the year data into a time series line chart to show the year by year trend:



In this plot, we can see that the number of shooting incidents steadily decreased from 2015 to 2017 and then maintained at a similar level ~960 cases from 2017 to 2019. However, in 2020, the number of incidents drastically increased to 1,942.

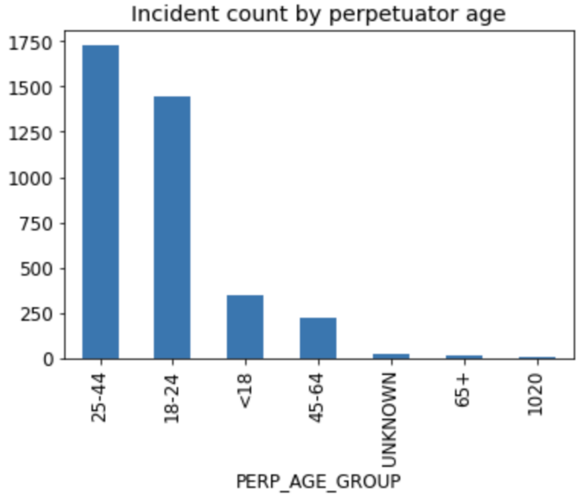
To get a better view of the trend, a month by month plot can be created by plotting the number of incidents against the MONTH_YEAR column. The number of accidents is again is again aggregated through the groupby function and gives us the below graph:

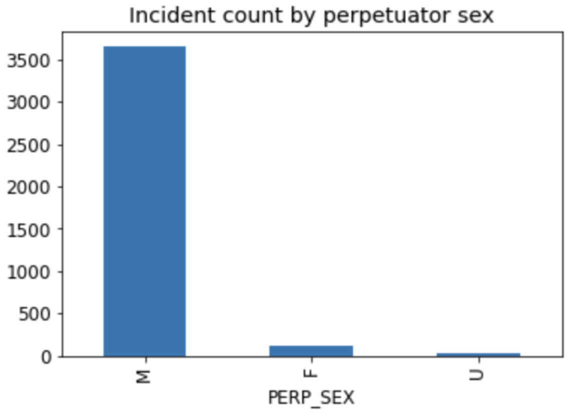


To make the graph more viewer friendly, the y-axis is labelled by quarter and the line is labelled by dots that indicate monthly incident number. We can observe a rough seasonal pattern where in the summer shooting incidents usually surge. This is especially prevalent in the year of 2020 where we see the monthly number of incidents goes above 300 in the summer.

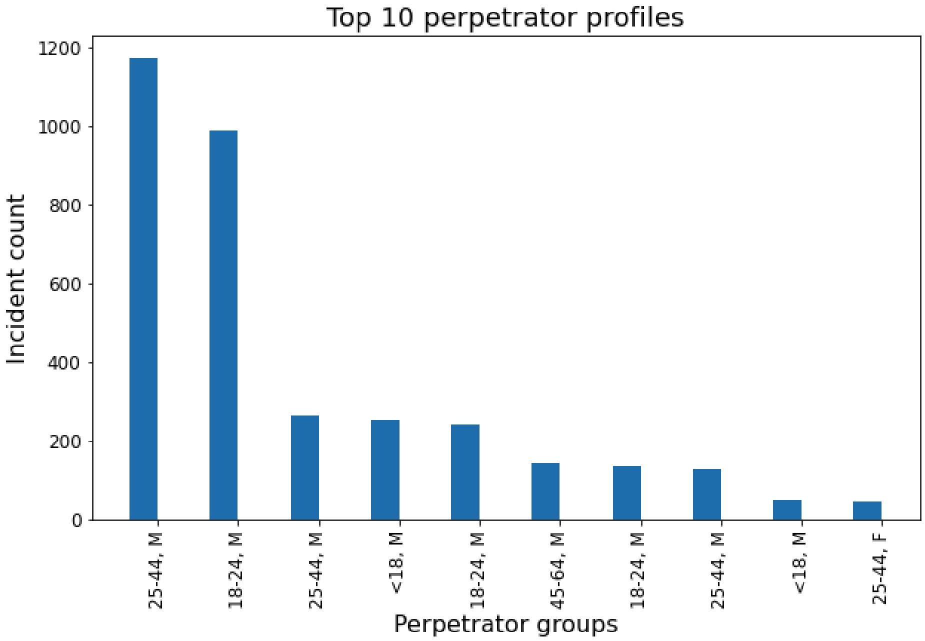
Question 2

To understand whether shooting incidents are related to graduation rate, another preliminary investigation we could do is looking at the *profile of perpetrators of shootings in New York City*, especially whether they are in the age groups where they should be in school or just graduating from school. In the data, there is information about the perpetrators of the shooting incidents. We can simply use the group by function to investigate what the most prominent groups are among these three perpetrator attributes:





By looking at the bar charts, the most prominent attributes of perpetrators are males of age 25-44. However, what these three individual plots do not tell us is the intersection of those attributes. To get a holistic picture of a perpetrator's profile, we have to combine the two attributes. A PERP column is created by concatenating those two attributes. With the new column, we can again plot a bar chart of the 10 most prominent profiles of shooting incident perpetrators:

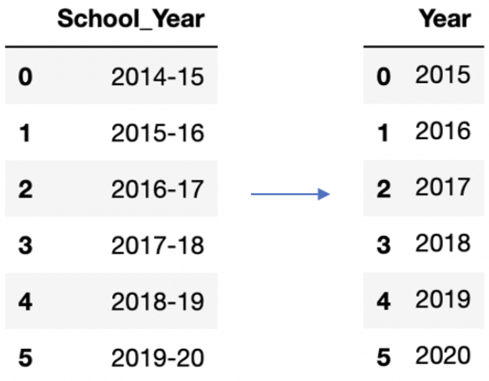


This bar chart shows that the most prominent perpetuator groups are males of age 25-44 and 18-24. They are responsible for a significantly higher number of shootings than the other groups that follow. The sum of the groups that rank 3rd, 4th and 5th is still lower than the 2nd ranking group.

Question 3

With the goal of eventually answering the question of how graduation rate and NYPD shooting incident rate are related, we first wanted to assess *what the overall trend of graduation rate year over year has been in New York City by gender.* Because we had dissected from the data what the typical profile of perpetrators in shootings was, we thought that gender was an interesting classifier to look at graduation rates against. To do this, we had to bring in data from the NYC Graduation Rate Database. As discussed in section 2, this data comes from the New York State Education Department, which is mandated by the government to produce accurate figures around high school enrollment, graduation, dropout rates. This database is the only source needed for this particular research sub question.

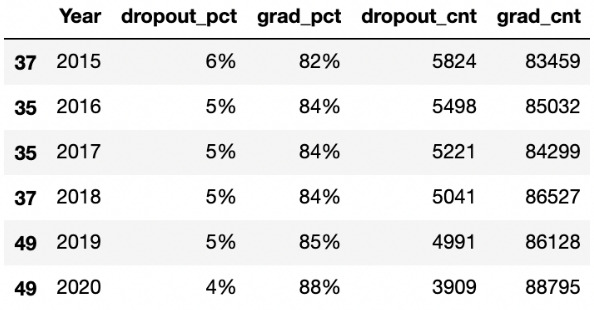
The NYSED Graduation Rate Database currently produces a separate CSV for each year's data between 2015 and 2020. Having only 5 data points to work with is not ideal, but we should still be able to see if a trend in graduation rate exists year over year. To begin the data cleaning process, we had to read in each CSV file and concatenate them. However, after looking at the data documentation, we noticed that some years' data had lowercase column names and other years had uppercase column names. To remedy this, we converted each CSV's columns to lowercase and produced unique values for the report_school_year column as a sanity check that the concatenation worked. The current format of the year being 20XX-XX wasn't going to align with the police shooting data, which is just represented as 20XX. Given this, we had to alter the School_Year column data to just use the year of graduation as the reference point. For example, the school year marked 2014-15 will be changed to 2015. We then altered those row values for the report_school_year column and checked that the output was appropriate for our x axis. We also changed the column name to ‘Year’ so that we’d be able to successfully join it to the shooting data. See the output from the transformation below:



	School_Year
0	2014-15
1	2015-16
2	2016-17
3	2017-18
4	2018-19
5	2019-20

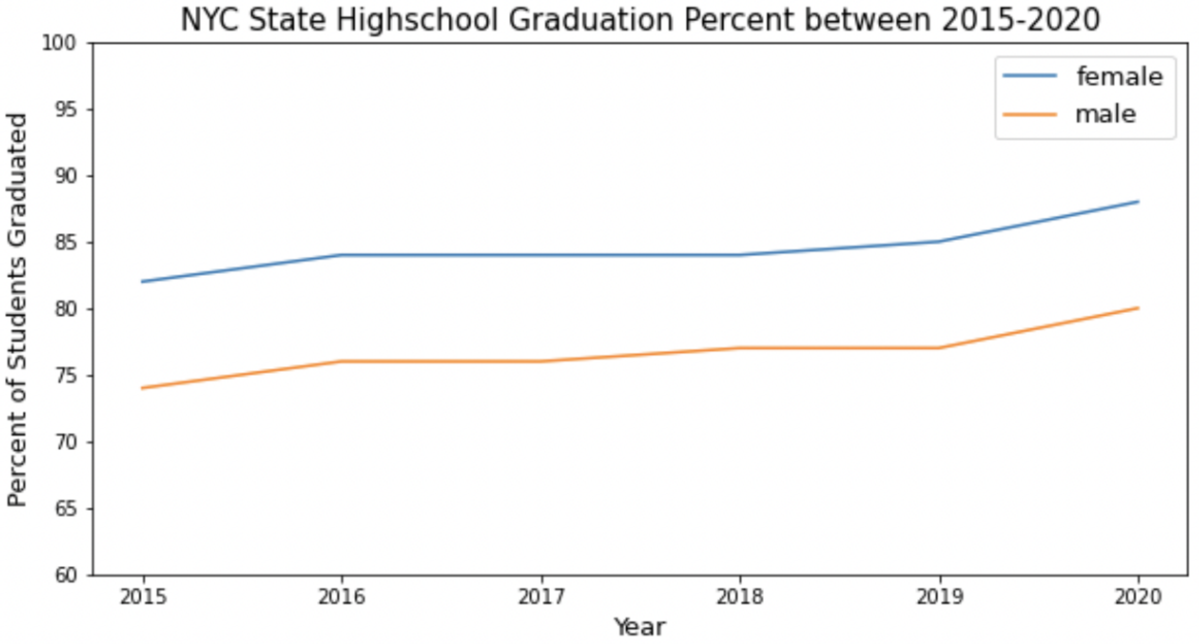
	Year
0	2015
1	2016
2	2017
3	2018
4	2019
5	2020

After the datasets were concatenated, we began applying the necessary filters to produce the desired plot. For this analysis, we wanted to look at the graduation rate aggregated at the state level and segment it by gender. To accomplish this, our first step was to only select the rows for each year that are aggregated at the state level, and this could be done by filtering on rows with column aggregation_code = 0. We determined from reading the code book for this dataset that was the correct index for state level aggregation. We also found from the documentation that to return rows representing graduation rate for students completing their degree in 4 years, we had to filter on the column membership_code = 9. Now the last step was to separate the existing dataframe into figures for male and female, giving us 2 dataframes with 5 data points in each. From the code book we knew that the segmentation description was contained in the subgroup_name and subgroup_code column, so we printed the unique values of those columns and determined subgroup_code value of 2 represents the graduation rate among female students, and a subgroup_code value of 3 represents the graduation rate among male students. Now that all the filtering is done, we can conduct some sanity checks on the outcome variables that we’ll be using for the plot. See the output of the dataframe representing Female students below as an example:



	Year	dropout_pct	grad_pct	dropout_cnt	grad_cnt
37	2015	6%	82%	5824	83459
35	2016	5%	84%	5498	85032
35	2017	5%	84%	5221	84299
37	2018	5%	84%	5041	86527
49	2019	5%	85%	4991	86128
49	2020	4%	88%	3909	88795

For readability, we chose to use the grad_pct variable rather than the count for our plot. However, we first had to strip the percentage sign from the variable name and then convert the column type to an integer. After doing these variable transformations for both the male and female dataframes, we were ready to plot our figure.



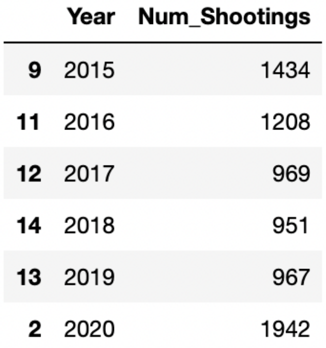
There are two main insights we can gather from this plot. Firstly, it's clear that the count of female students graduating is consistently higher than the count of male students graduating year over year. Secondly, both graduation counts for male and female have risen between 2015 and 2020. However, this increase is not very drastic. In the case of female students, between 2015 and 2020 the graduation count only rose from 83,459 to 88,795 which constitutes a 6% increase. Similarly, the male student graduation count only rose from 79,281 to 85,476 which is also a 6% increase. So, this is not a very notable increase.

Question 4

Now that we've seen the trend of NYC Graduation Rates year over year, we can get to the real causal question that we're after: What relationship do graduation rate and dropout rate have on shooting incident rates? For this question, we already have the graduation rate and dropout rate data cleaned. Now we just need to bring in the NYPD shooting incident data and clean it up for our purposes. To answer this question we'll have two separate plots in our figure, one with dropout count vs. number of police shootings by year and the other with graduation count. We will then plot the regression line on each figure to see the year over year trend.

Firstly, we have to manipulate the shooting incident data for our purposes. After reading in the head of the dataframe, we saw that each row accounts for a shooting incident along with associated data about the shooting such as time, victim race, sex, etc. which was all leveraged for earlier research questions. Because we're really only interested in the count of shootings each year, we'll need to refactor this data into a count of rows by year for our plot. To do that, I'll need to pull the year out of the date column and then use value_counts(). We also need to filter this data to only show values starting in 2015, because this data goes back a lot farther.

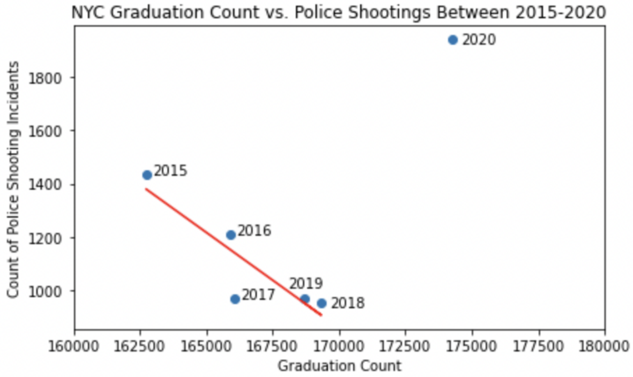
Before making these transformations, there are only two columns that I'll need to run some sanity checks on - the incident key and date columns. For the incident column we just need to ensure that there's no null values. To check for this, we ran isnull().values.any(), which returned False meaning that there were no null values. For the date column, there are a few sanity checks that need to be done. We had to check that there were no invalid values for the month, day, or year. We accomplished this by printing the sorted unique values for each, and they all only had valid values.

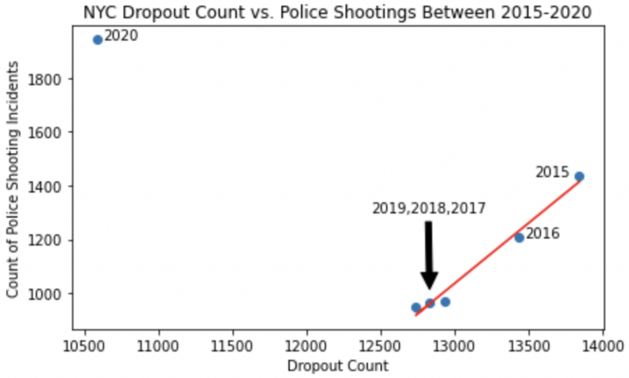


	Year	Num_Shootings
9	2015	1434
11	2016	1208
12	2017	969
14	2018	951
13	2019	967
2	2020	1942

Now that the sanity checks have been done, we can print the refactored data that reflects the number of shooting incidents that occurred each year between 2015 and 2020. At first glance we can see that the number of shootings decreased fairly substantially between 2015 and 2017, but remained more stagnant between 2017 and 2019 before drastically increasing by more than 100% in 2020. We will make more sense of these numbers after generating the plots.

Now we can join this data to the graduation rates data for our output plot. Because the y axis is going to be in terms of the *number* of shootings each year, it makes more sense for us to use the *number* of graduating students and students dropping out each year rather than the percentage that we used for the Question 3 plot. We're plotting both to make sure that we see inverse trends and confirm our data accuracy. After joining the datasets together, these were the resulting plots:



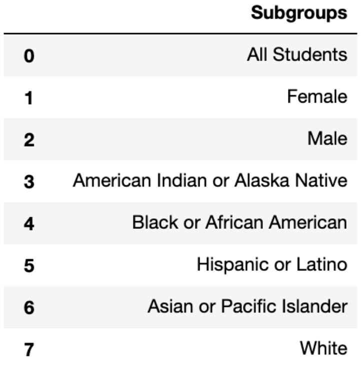


There are, again, a couple insights we can pull from this plot. Firstly, the regression lines were done using the polyfit function, which is described in further details within questions 5 and 6. We purposefully left the 2020 data points out of the regression line because it's an explainable outlier. But the general finding from the plot is that as graduation count increases, police shootings decrease. This is a very actionable insight. While we don't have nearly enough data points to really generate a statistically significant finding, we could use this as a basis to perform further analysis.

Question 5

For the last couple questions, we continued our study of shooting and graduation trends, but we implemented another important variable- median household income. The first of these was: *How does the high school dropout rate compare to median household income, aggregated across each of the 62 counties in the state of New York?*

There was a small amount of overlap in the data required for questions 3 and 5. However, the operationalization of this data quickly changed after the first couple steps; question 3 focuses on graduation rates over the last 5 years while we now shift the focus to trends across state county lines.



	Subgroups
0	All Students
1	Female
2	Male
3	American Indian or Alaska Native
4	Black or African American
5	Hispanic or Latino
6	Asian or Pacific Islander
7	White

We started by concatenating the 6 data frames for graduation rates for the years 2015-2020. At first, it was difficult to understand what portions of the data were important for this question. It became clear after some exploratory data analysis that we needed to filter by ‘all_students’ in the ‘subgroup’ series (see table to the right), and as noted above, by a specific membership code as well. This then gave us a 372 column data frame, one row for each of the 62 counties and each year 2015-2020. We then grouped the dataframe by county and for each one we calculated the average dropout percentage over the last 5 years.

After this was completed, we were ready to merge this data frame with a new one that gave information about each county’s median household income across the state of New York. This data was quite easy to obtain from the Bureau of Economic Analysis, with a small caveat- the data only displayed the median income for the year 2019.

At first we thought it would be important to account for this, and we considered dividing by a small percentage per year (to account for inflation) then taking the average income between 2015 and 2020. However, there are more complex metrics internal to median household income that would not have been captured by this calculation anyway. In the end, it also became more important to answer the question at hand, which was how the graduation rate differed across counties as a function of their different household incomes. Our charts would be unnoticeably different had we factored inflation into the data.

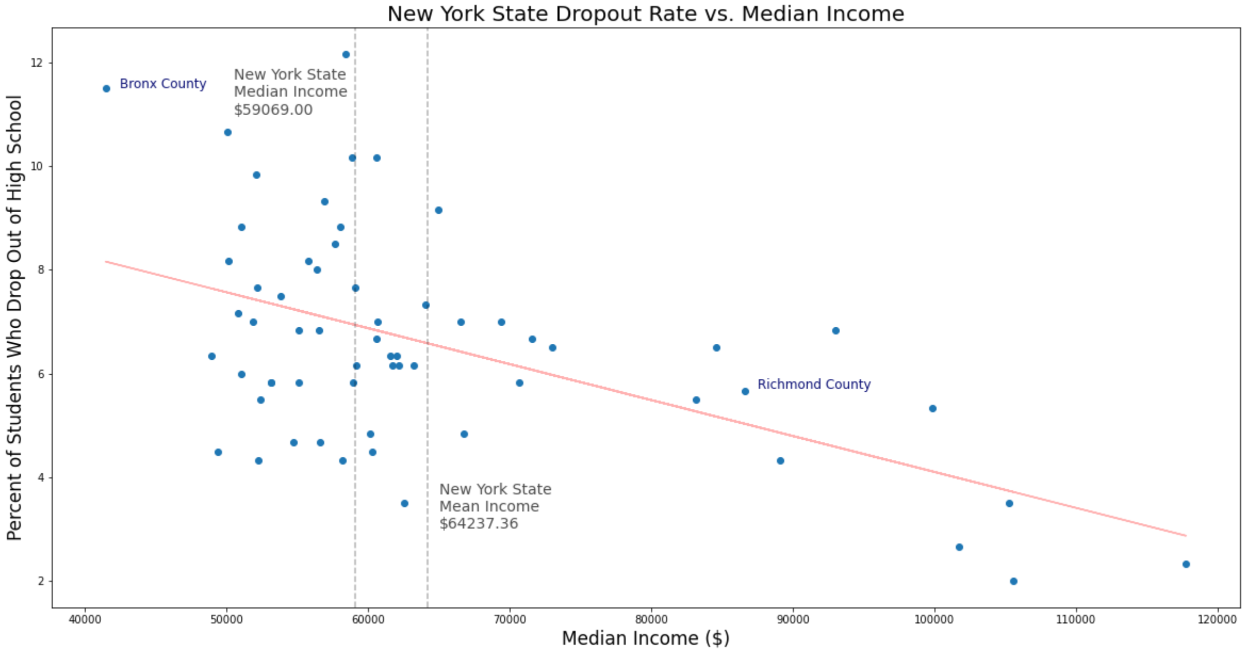
Data cleaning for the income dataframe was a little more involved. The counties listed were all lowercase and in a different format. For example, where the above graduation dataframes state ‘ALBANY’ as the county, the income data uses ‘Albany County, NY.’ This wasn’t a huge problem- we just had to use a .str method to remove the last 11 or so characters from each string. Also there was a percentage symbol in one of the variables of interest, which we deleted via a string method. We then type-casted the data into float format.

The merging process was quite straightforward. We checked to make sure that there are no duplicate county names in New York (just in case) so that we could merge the graduation data with median household income data by county name. We had to use a .str method to transform the names in the income data to uppercase letters, which was easy enough. Our independent variable was represented in two ways:

1) Median household income in dollars
2) Median household income as a percentage of the New York state average income

We elected to represent the graph with the former of these since it seems a little more intuitive.

We created a scatter plot of 62 points (again, one for each county), which produced the figure below.



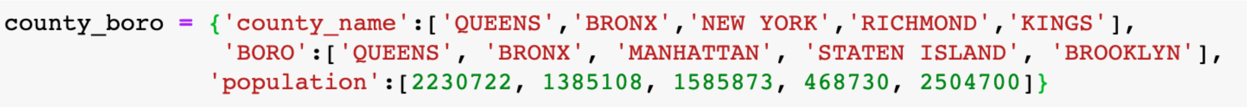
It is apparent that there is a decreasing correlation between median household income and dropout percentage. To make the chart more effective, it was important for us to add a trendline using the np.polyfit method. This took in the two data series from the scatter plot as arguments, and returned values for slope and intercept of the least squares linear model (per the numpy.org documentation). In addition, we added vertical lines at the median and mean household incomes for the state of New York. We did this using the plt.axvline() method, and we used the pd.series.mean() and pd.series.median() methods to set the value correctly for each one. These add a lot of context and allow us to see how few counties are above the average income level across the state of New York.

Question 6

Question 6 also incorporates household income, this time comparing it to shooting rates in the five major boroughs of New York City proper. Unfortunately, we were unable to acquire data about shooting rates for all 62 New York counties. Most focused on just New York City as noted above. We therefore refined our question to: *How does the shootings rate per 1000 citizens compare to median household income, aggregated across each of the 5 boroughs in the city of New York?*

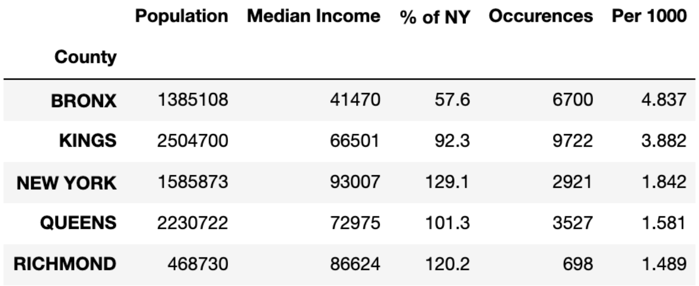
We built this data cleaning process off of that done for question 5. We already created a dataframe that contained the median household income represented in two ways.

The next piece consisted of the shooting data, but since we were measuring it across locations, we had to standardize it by population. We wanted to take a different approach rather than just importing another .csv file as a new dataframe. Since there were only 5 data points to begin with, we created a dictionary that consisted of the boroughs, their respective counties, and their populations, which we then converted into its own standalone dataframe.

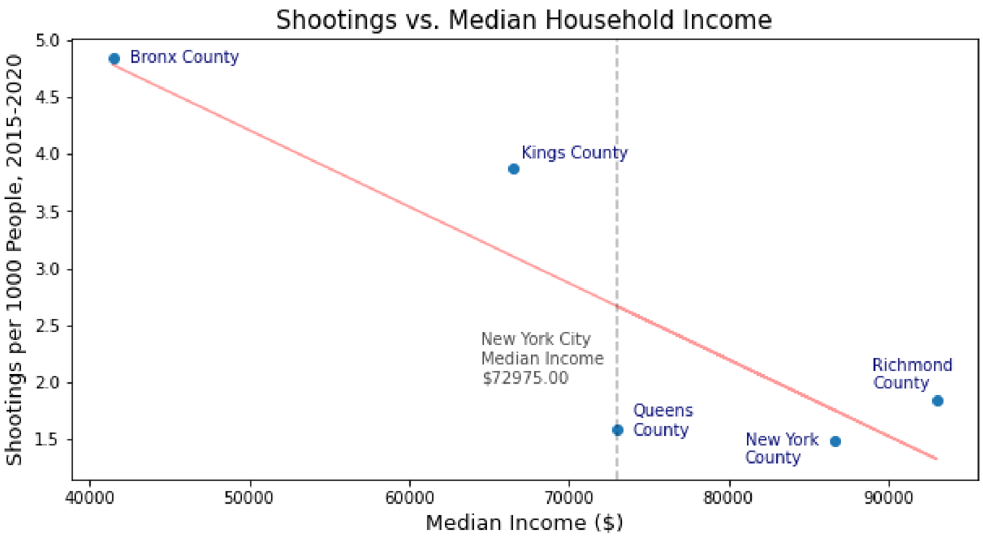


```
county_boro = {'county_name':['QUEENS','BRONX','NEW YORK','RICHMOND','KINGS'],
                'BORO':['QUEENS', 'BRONX', 'MANHATTAN', 'STATEN ISLAND', 'BROOKLYN'],
               'population':[2230722, 1385108, 1585873, 468730, 2504700]}
```

By incorporating a key called county_name in this dictionary, we were able to directly merge to the existing income and shootings dataframes. This resulted in the following table, with the occurrences variable effectively being a representation of the value counts for each county. This was good practice in applying different functions to each variable of a grouped dataset.



	Population	Median Income	% of NY	Occurences	Per 1000
County					
BRONX	1385108	41470	57.6	6700	4.837
KINGS	2504700	66501	92.3	9722	3.882
NEW YORK	1585873	93007	129.1	2921	1.842
QUEENS	2230722	72975	101.3	3527	1.581
RICHMOND	468730	86624	120.2	698	1.489



We created a scatter plot of the above table, once again deciding to use median income in dollars as the independent variable. Although we have significantly fewer data samples, it would be highly coincidental for the above pattern to be due to a coincidence in sampling variance (i.e. we think there is a true underlying correlation that is represented by the data).

4: Conclusion

The analysis above serves us well in uncovering general trends and groupings within our raw data. We start by modeling shooting patterns in New York over the last 6 years. As anticipated, we see a noticeable spike in incident rates in 2020, likely associated with increased political turmoil and strife surrounding the COVID-19 pandemic.

We then explore characteristics of shooting perpetrators to see if we could attach large proportions of these incidents to those within a specific age group (18-24 year olds who should have just graduated from high school). Our bar plots in question 2 do a good job of enforcing this trend and suggesting that there is a correlation between this age range and likelihood of being involved in a shooting incident.

We did not note any surprising patterns in graduation rates with respect to time; however, when we combined this graduation data with shootings trends, we saw that the years with increased numbers of high school dropouts also contained more gun violence incidents. We acknowledge that the sample size for this model is somewhat small (n = 5) , but it would be highly coincidental to develop the trends we see in the plots if there were no true implicit relationship.

Lastly, we found substantial evidence indicating that both shootings incidents and dropout rates have a negative correlation with a county's median income (i.e., wealthier communities have fewer shootings and fewer dropouts). It was important for us to attach this third metric to some of the relationships we had uncovered in questions 1 through 4. This added independent variable serves as an explanation for the previously-found patterns, and it is suggestive of which jurisdictions should be prioritized in terms of implementing corrective action.