Module 6, Project: Final Submission—Data Analysis Report

Module 6

Author: Krishna Sathyamurthy (ks2025)

Module 6, Project

Contents

1 Introduction

In this project, we will be analyzing the NYPD historic complaint records to identify the following trends, in no particular order,

- What are the most common offenses performed by the offenders?
- Heat map of these offenses across the city.
- Heat map of the educated and employed populous across the city.
- Correlation between education, employment and crime.
- Offense frequency through the day across various precincts.
- Offense, Education, Employment count through the years.

For this project, we will only be dealing with the records from 2011-2019. This document has been split into 3 parts, data collection/cleansing, data processing, and data analysis.

2 Data Collection/Cleansing

2.1 Data Collection

The data for this project was collected from multiple open source websites. Thanks to NYC having an open database for accessing and downloading all the files, the data could be obtained easily. The NYPD historic complaint records were obtained from NYC Open website. The US-Census website was used for collecting the information about NYC population with an employment, and basic education. Finally, to identify the zip-code of the city, US-Census was again used to download the shape file.

2.2 Data Cleansing

While NYPD historic had a lot of data, some of the entries were missing, or some new columns had to be added. Additionally, the NYPD historic dataset spanned between 2006 to 2023, while the US-Census data was only available between 2011-2019. Hence, it was decided to truncate the NYPD historic dataset between the years 2011-2019. Similarly, we noticed that the education and employment dataset does not contain latitude, longitude details, but contained the zip-codes. Thus, we decided to add the zip-code to the NYPD historic dataset.

Adding zip-code to the NYPD historic dataset was one of the first challenge we came across. We tried many methods, such as polling a server for getting the zip-code, searching through the entire shape file polygon to triangulate the zip-code pertaining to our row. Unfortunately, all of these were too time consuming. To address the issue at the earliest, it was decided to take a novel approach, and truncate the shape file data for just the counties that belong to NYC. Thus, after identifying that NYC consists of 5 counties, namely, New York County, Bronx County, Kings County, Queens County, Richmond County, this was used to trim the zip-codes and Polygon geometry data for the shape file. Later, this data was used in the already truncated NYPD historic data on python. Python was preferred due to the ability to easily process our data across multiple processors.

Likewise, python was also used for concatenating all the education and employment files into a single csv file. Thus we were able to collect, cleanse and coalesce all our data.

2.3 Data Processing

For processing this data, R studio was used due to the ease at which we can make graphs and infer some exciting information from our data. The data processing can be split into multiple parts, to make some graphs for the analysis we were planning to perform. The analysis for each topic is as follows, in no particular order.

2.3.1 What are the most common offenses performed by the offenders?

Processing this data was straight forward. All we are required to do is get the count of each unique offense made. Similarly, this can also be done through the years 2011-2019. This way, we should be able to make a more informed decision on whether the same crimes remained on the top 10, or if there were any changes.

2.3.2 Heat map of these offenses across the city.

To perform this, the zip-code in our NYPD dataset was used for merging with the shapefile dataset. Since the shapefile contained the geometry details of the NYC landscape, it was easy to plot the map on the graph. The challenge we faced was plotting the offense committed as legend across the various counties all on the same map, while distinguishing the various counties. This was accomplished by adding the *new_scale* function, which helped us in adding a new fill layer on top of the existing layer. Thus, we added the new fill layer with some transparency to see the main count legend.

2.3.3 Heat map of the educated and employed populous across the city.

This again follows the same steps followed in our previous section, thus making it easier to process the data. The only additional task which was performed in this section, was to take an aggregate of the values across the years and use the mean of it for plotting the values on the graph.

2.3.4 Correlation between education, employment and crime.

To identify the correlations between the various columns, we had to use the aggregate function again. The only difference now being, instead of doing it on the zip-codes like we did earlier, now it is done on year and county. Similarly, a table function is used for the NYPD dataset to identify the count of the crimes committed through the various years across the different counties. Finally, all of this data was merged using the columns county and year to plot the correlation graphs.

2.3.5 Offense frequency through the day across various precincts.

To perform this, we first had to convert the time from character to a posix format. While this is not necessary for the graph, this will be required when updating the y axis sequence with 1 hour breaks in-between. For this section, we find the frequency of the crime committed across various precincts at that particular time, and later plot it on the graph. Finally, we will differentiate the precincts based on the counties for a better understanding on which county, precinct gets more number of offenses.

2.3.6 Offense, Education, Employment count through the years.

This is just a culmination of what we have done so far, i.e., we take an aggregate of the data which we require, and then plot it across various years. To make an inference from this graph, the graphs has been split across various counties, and all the different datasets are plotted on the same graph.

3 Data Analysis

3.1 What are the most common offenses performed by the offenders?

The first plot1 gives an idea on which of the offenses were most performed from 2011-2019. But, by zooming into our dataset and identifying the top 10 crimes in each of the year, we were able to find something new. While Burglary was the most common offense through the years 2011-2014, we see a new paradigm, and find that more and more people are doing drugs in the later part of the years. While some of the crimes seem to decrease as the years go by, unfortunately, we notice some of the more serious crimes like Assault, Harassment is seeing a steady increase.

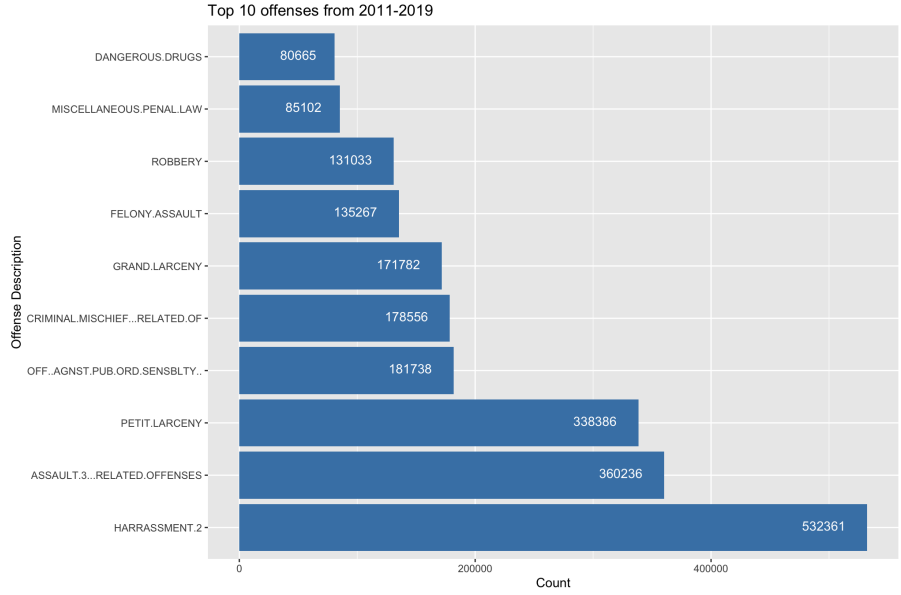

Figure 1: Top 10 Offenses

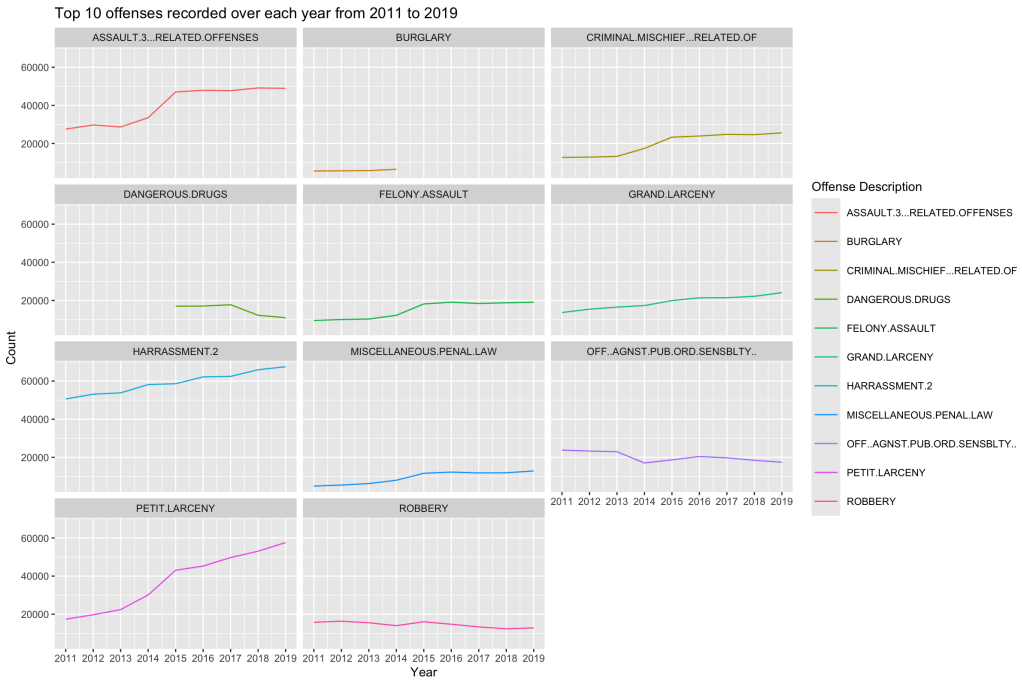

Figure 2: Top 10 offenses through the year 2011-2019

3.2 Offense frequency through the day across various precincts.

While the individual points almost transform into a line in most of the precincts, we are able to find that some of the neighborhoods are actually safer, due to absence of the clear line through the day. Another inference that we make is, more number of precincts can be opened up across the city since there appears to be a lot of crime, occurring almost all through the day. While this is not an frequency average of the crimes through the day, by increasing the number of precincts, the load on these police officers can be reduced, and we can also ensure better safety for the citizens.

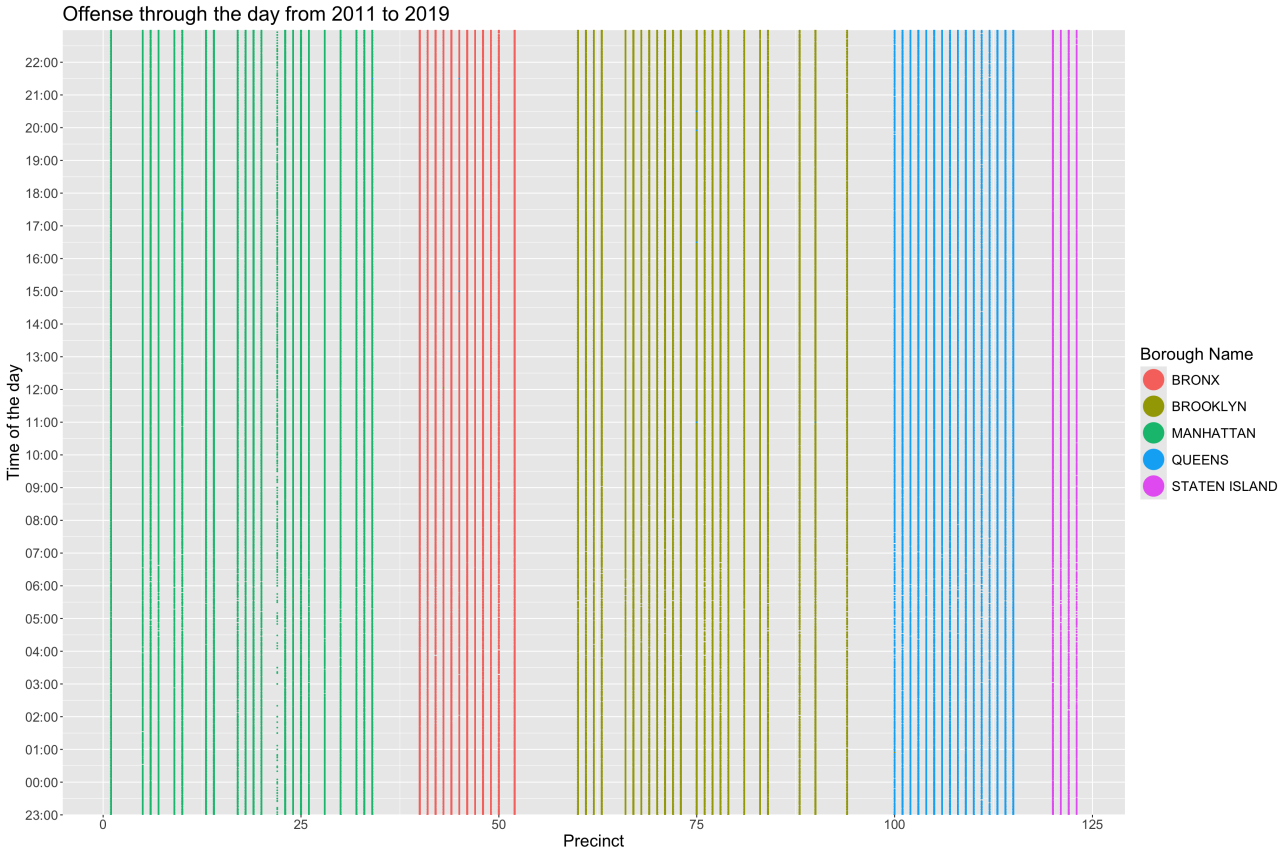

Figure 3: Offenses recorded through day across various precincts

3.3 Offense frequency through the years.

We can notice a steady increase in the number of crimes performed year over year, and another analysis we find is that, the number of crimes usually peak around the summer seasons, i.e., when people mostly prefer roaming around on the streets, and we notice a significant drop in the crimes committed during the winter season. While it is quite possible for the number of crimes to be reported during the winter season to be small, it is no surprise that the number of crimes reduce during this season. Likewise, we also notice there has been a steady increase in the number of crimes committed, and which again requires more police force. While FBI reported lower murder rates in 2014, we notice that the number of crimes has still increased dramatically after the fall of 2014. This also happens to be an outlier, since this is when the number of crimes performed during winter is almost same or more than summer per day.
By plotting a box-plot, we can better infer the increase in the number of crimes performed year over year, and the distribution also remains constant, thus explaining a common trend across each year.

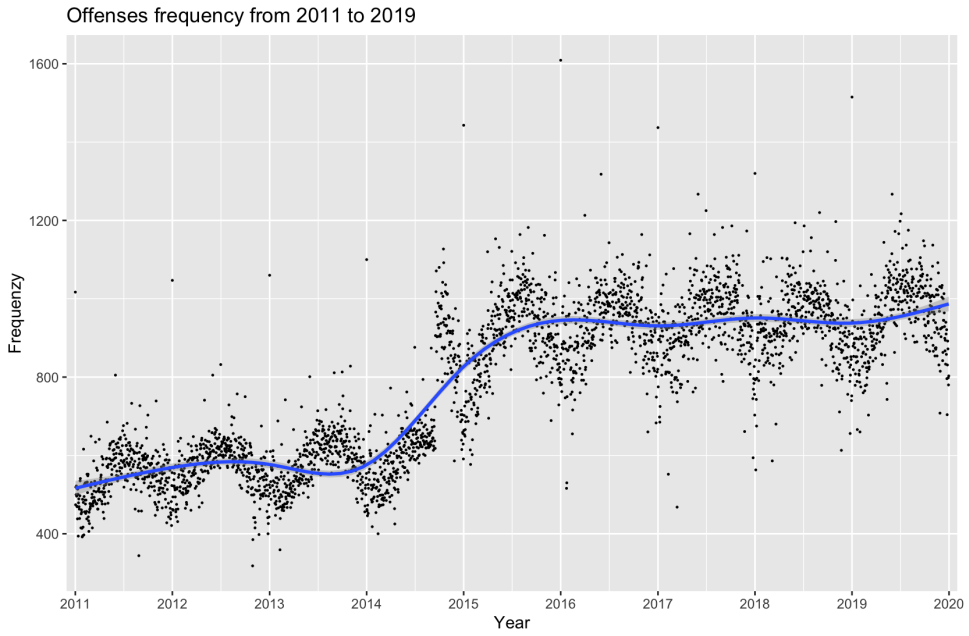

Figure 4: Scatter plot of the frequency of crimes across each year

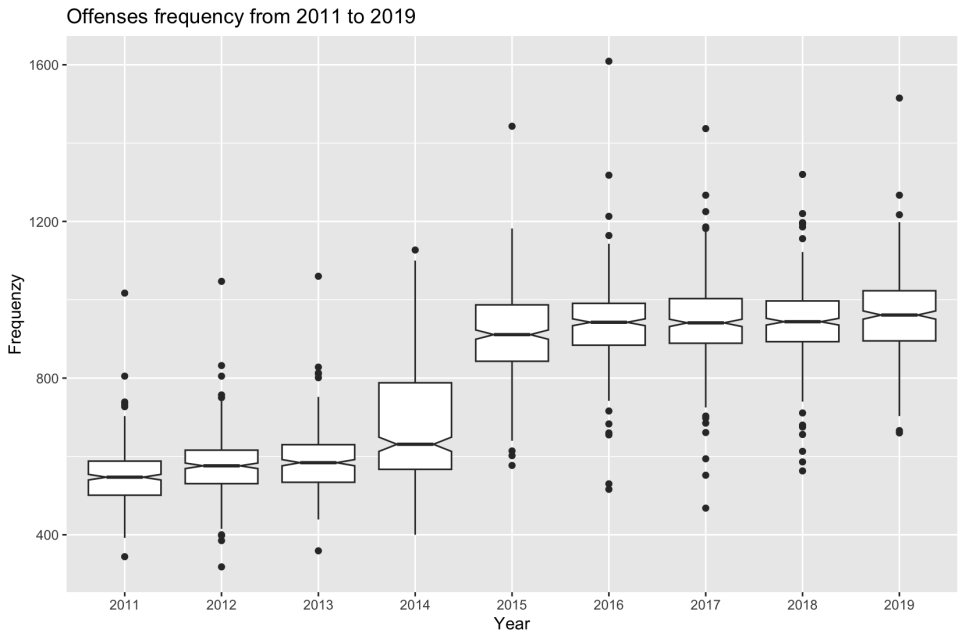

Figure 5: Box plot of the frequency of crimes across each year

3.4 Heat map of these offenses, population, educated populous and employed populous across the city.

We infer that the number of crimes is related to places that are more populated/more dense. This is no surprise, since as the number of crimes increases, the number of petty crimes might also, such as skipping a traffic signal. We also happen to infer that, the regions which are more educated, are the places where lesser crimes usually occur. Of course, as more people are educated, the more likely that they are taught about morals, and the more likely that they are to find a stable income, thereby making those regions less likely for crime to occur. But is there a cause and correlation between education, employement and crime numbers?

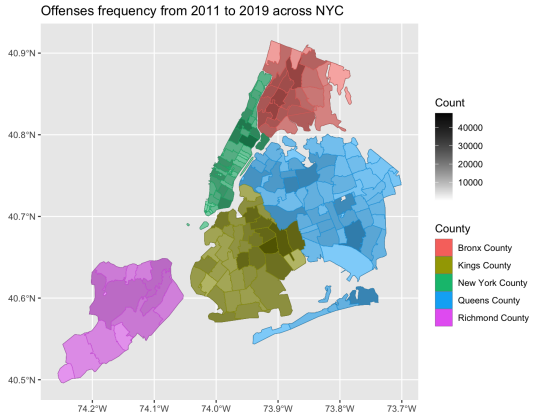

Figure 6: Offense frequency

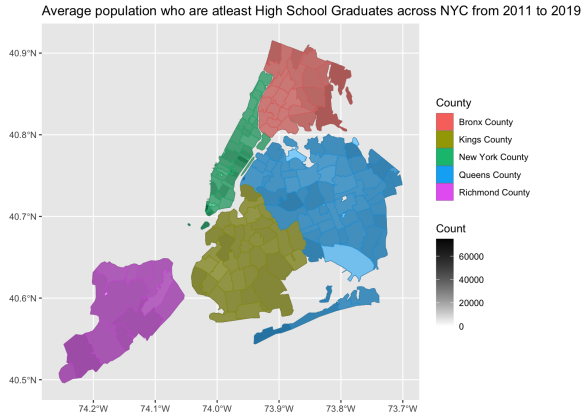

Figure 7: Population with at least a High school degree

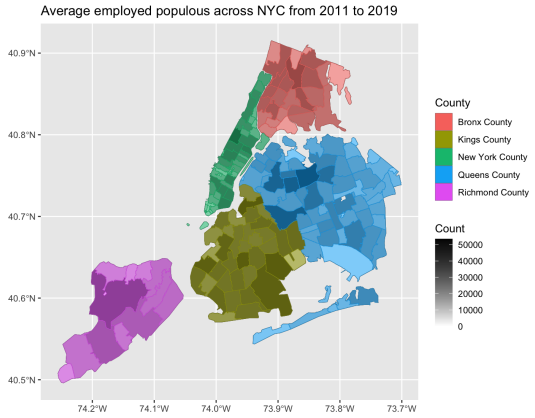

Figure 8: Employed Population

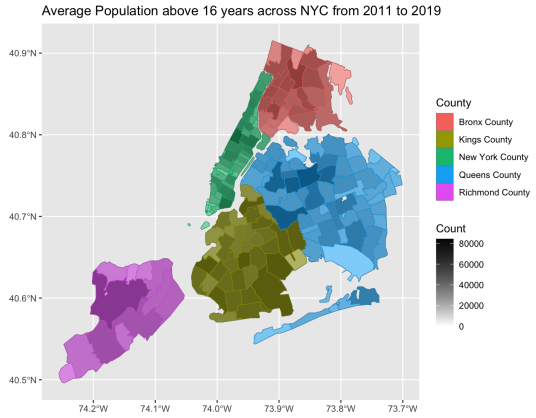

Figure 9: Population

3.5 Offense vs other factors across various counties through the years

We made a inference earlier about education having negative correlation to crimes committed, and it is very clear in this graph, since, as there is steep increase in the crimes committed as the number of educated people reduce in 2015. The increase in crimes is also recorded by the 2015 recap article from the WSP. But the correlation between employment and crime is not to be found on this graph, since there was no major shift in 2015 on the employment graph, while the other two faced drastic changes.
Another interesting inference that we make from this graph, is that, the number of crimes committed in some of the crimes are reducing over time, while some are increasing faster than before. While Bronx is infamously known for being unsafe, instead, we infer that Kings and New York County (otherwise Manhattan) are more unsafe in comparison. Kings having nearly 100,000 cases on an average throughout the year is rather alarming.

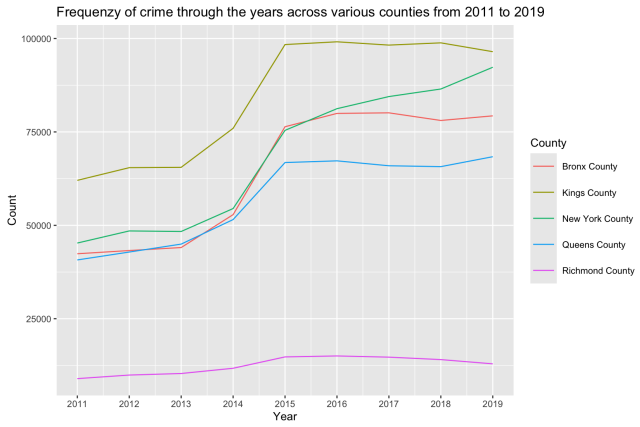

Figure 10: Crime count over the years

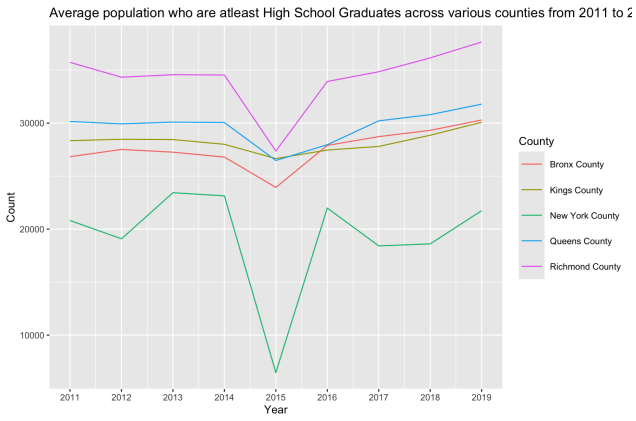

Figure 11: Educated populous over the years

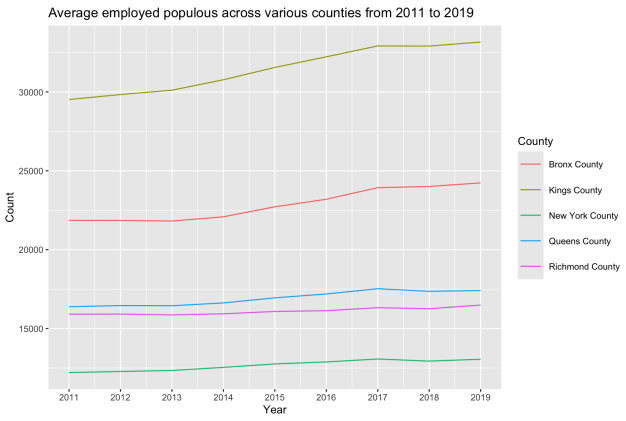

Figure 12: Employed populous over the years

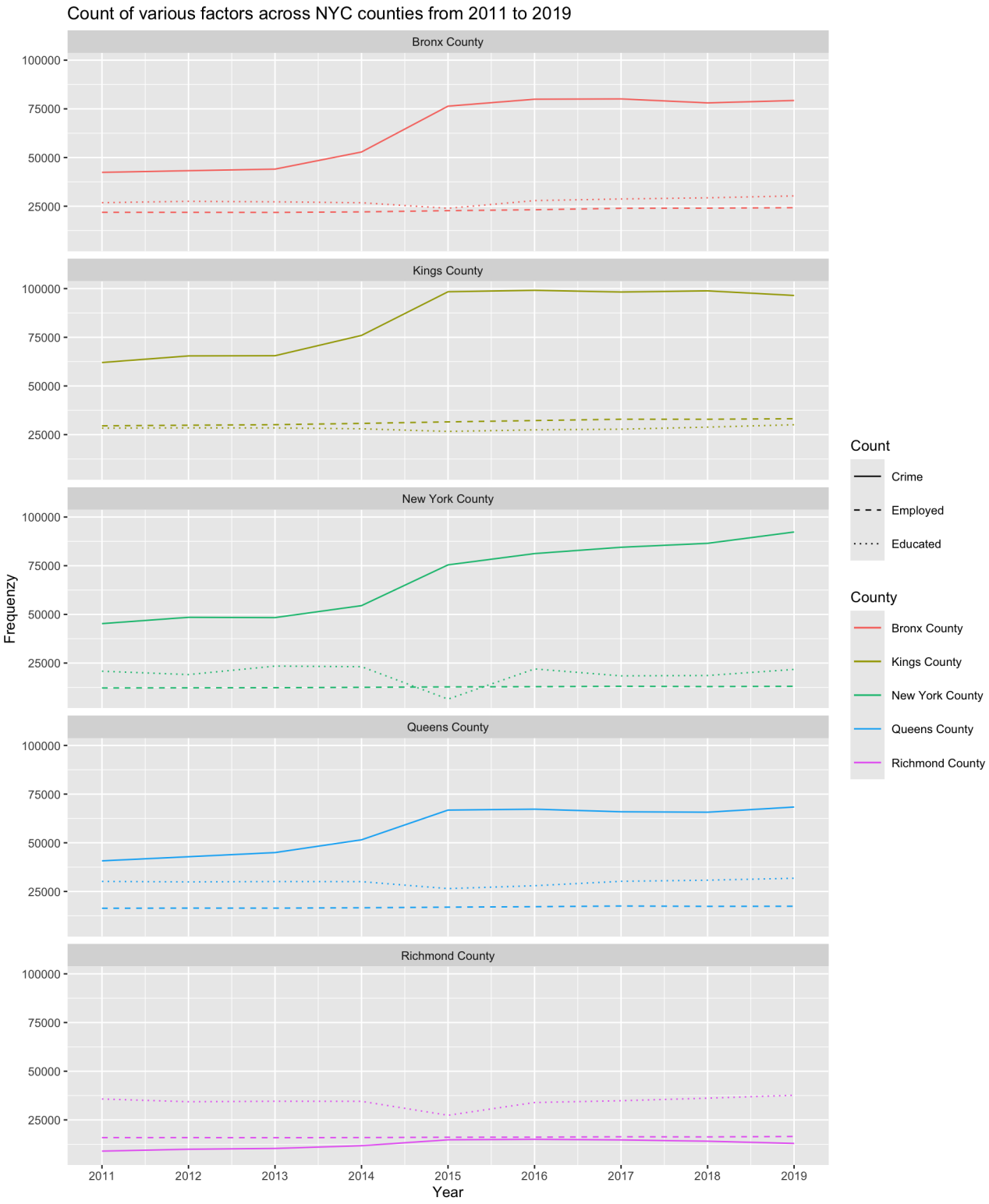
Count of various factors across NYC counties from 2011 to 2019
Bronx County
Kings County
New York County
Queens County
Richmond County
100000
75000
50000
25000
Frequenzy
2011
2012
2013
2014
2015
2016
2017
2018
2019
Year
Count
Crime
Employed
Educated
County
Bronx County
Kings County
New York County
Queens County
Richmond County

Figure 13

3.6 Correlation between education, employment and crime.

By plotting the correlation index, we are able to confirm the assumptions we have made so far based on the inference made. There indeed is a negative correlation between crimes committed, and educated people, and there is also a direct correlation between a more populous place risking more cases on an average.
By plotting the graphs for the various counties, we find that while some of the county have a very high correlation between the different columns, Richmond does not follow the same pattern with employment, whereas Bronx doesn't follow the same pattern with education. While our initial assumption of more education leading to lesser crimes might hold true, the positive correlation might give us an idea on why Bronx is typically labelled to be unsafe.

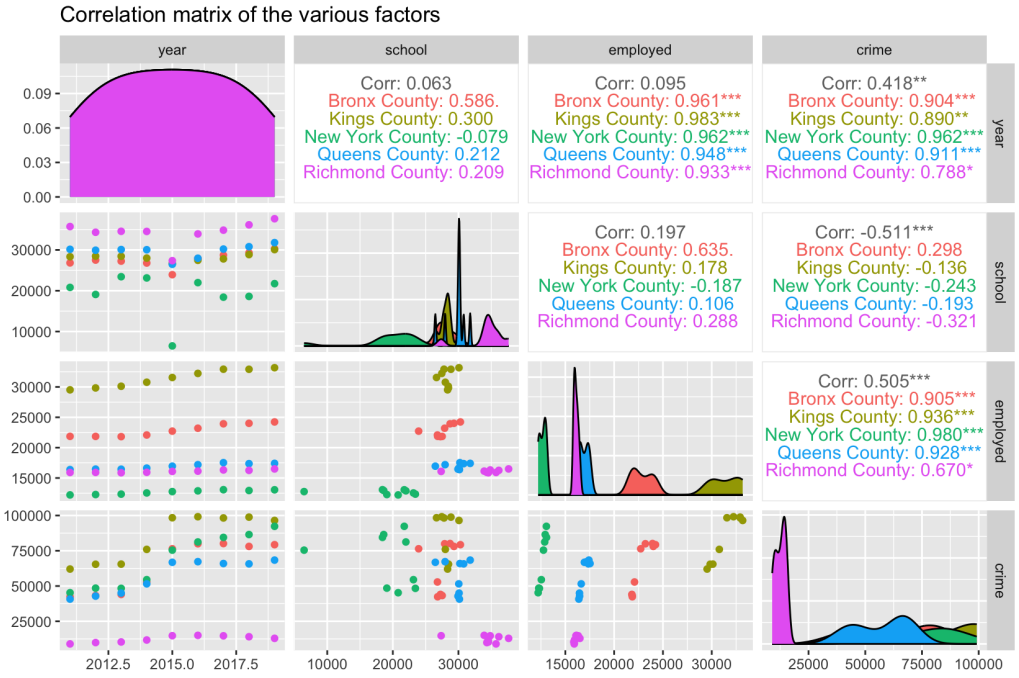

Figure 14

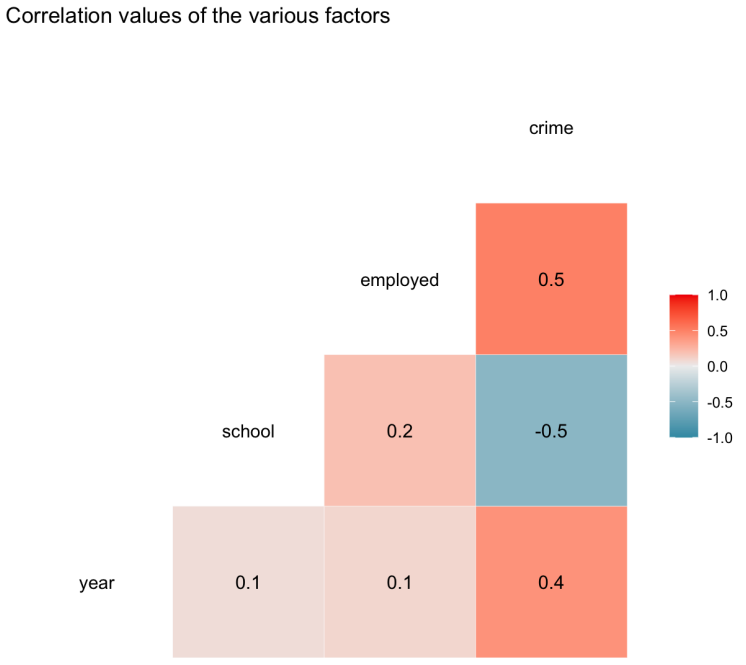

Figure 15

4 Conclusion

From the analysis performed thus far, we can conclude that the following.

- Drug department has to become more stringent, and more people should be educated about the dangers of using drugs.
- There appears to be lesser crime occurring on an average in Richmond county, thus some of the police force can be transferred to New York County, which is seeing a steady increase in the number of crimes committed each year.
- We need to ensure that the majority of populous is educated, since there appears to be some amount negative correlation between that and the number of crimes committed. Education doesn't necessarily mean just a degree, but morals should also be taught to ensure that this helps in reducing the crimes committed.