Capstone Project – The Battle of Neighborhoods

Introduction

The Region Stockholm consist of 26 counties with different demographics as well as number of crimes. The population is growing and individuals, couples and families are moving to the region as a whole and within the region between the counties. How can citizens know which of the 26 counties is better than the other in terms of crime, based on the county's demographic?

My Capstone Project aims to explore the 26 counties and a number of their demographic variables as possible predictors to a number of reported crime variables. E.g. is the number of reported thefts and robberies in one county correlated by its unemployment rate or average age? Also, by calling Foursquare to get the number of police stations in the region, can one draw any conclusions between number of reported crimes and the number of available police stations near a county?

I believe this problem has many potential stakeholders. Some of the obvious ones are e.g.; the existing and coming citizens of The Region Stockholm, politicians, police department, urban developers and more.

Data

The data to be used are all statistics from 2019. The data are received from open sources such as; SCB (Statistics Sweden) and BRA (Swedish National Council for Crime Prevention).

1. Demographic data

This data is received from SCB and aims to capture 2019's demographic data for each of the 26 counties.

- **Name of county**
- **Latitude and longitude coordinates**
- **Population:** Number of individuals living in the county (by 31st of December)
- **Average age:** The average age of the individuals living in the county (by 31st of December)
- **High education rate:** The population rate that has at least a 3-years university degree (25-64 years old individuals)
- **Unemployment rate:** The employment rate of the labor force (16-64 years old individuals)
- **Early retirees' rate:** The population rate that has activity and sickness compensation (16-64 years old individuals)
- **Supported rate:** The population rate that lives by subsidy and support from government (25-64 years old individuals)
- **Business climate ranking:** A county's total business climate ranking from 1-290 based on 18 variables, made by the business industry.
- **Tax revenue per capita:** A county's total tax income divided by population
- **State subsidies per capita:** A county's total state subsidy divided by population
- **Costs per capita:** A county's total costs divided by population

2. Crime data

This data is received from BRA and aims to capture (some of) 2019's crime data for each of the 26 counties.

- **Total crimes:** Total number of reported crimes in 2019
- **Total crimes per 100.000 capita:** Total number of reported crimes in 2019 per 100.000 in capita
- **Violation of life and death:** Total number of reported crimes in category "Violation of life and death" 2019
- **Violation of life and death per 100.000 capita**: Total number of reported crimes in category "Violation of life and death" 2019 per 100.000 capita
- **Violation of freedom and peace:** Total number of reported crimes in category **"**Violation of freedom and peace" 2019
- **Violation of freedom and peace per 100.000 capita:** Total number of reported crimes in category **"**Violation of freedom and peace" 2019 per 100.000 capita
- **Theft, robbery and more:** Total number of reported crimes in category **"**Theft, robbery and more" 2019
- **Theft, robbery and more per 100.000 capita:** Total number of reported crimes in category **"**Theft, robbery and more" 2019 per 100.000 capita

Methodology

The Capstone Project was executed through a number of main components based on the Data Science Methodology presented earlier in this course.

1. **From problem to approach:** Within the frames of the project description, I found it interesting to explore the counties of Stockholm, the city where I moved to a few years ago and still live in. I thought it interesting to explore different demographical data together with reported crime data from 2019. Is there any correlations and interesting findings when working with the data?

2. **Working with the data:** Data to answer this question was quite easy to find due to many open data sources provided by Swedish's Authorities. The demographic data was received by SCB (Statistics Sweden) and crime data from BRA (Swedish National Council for Crime Prevention). The GeoJson data for the Folium visualizations was received from another open data source.

   Some work was needed to manipulate, merge and clean the data from the data sets. For example, I choose a few crime variables out of many different to decrease the number of dependent variables. Also, I merged the demographic and crime data into one csv file to ease the importing stage.

3. **Deriving the answer:** Then the exploration and analysis work started. I began understanding the dataset a bit more such as checking the data types and the variables correlation. Also, I picked a few pairs of dependent and independent variables with high correlation and plotted its linear relationship. Furthermore, I explored the dataset's basic statistics including mean, standard deviation and more.

For each pair of dependent and independent variable previously chosen, more in depth statistics where performed by calculating the Pearson Correlation Coefficient and P-value in order to confirm its significance.

A map visualization using Folium was performed to more easily see the counties' localization and difference in total crimes. The last step of the project was to call the Foursquare API service and receive all police stations in the region and add its localization on the folium map. This helps one to understand how the number of police stations and its localization can affect the counties total crime rate.

Results

Analyzing individual feature patterns, correlation, map visualization and Foursquare
This section will begin the journey of understanding the data by using correlations, visualizations and other individual feature patterns.

The performed correlation table shows the correlation between all variables. There are a number of positive, negative and non-correlated relationships between the independent variables (demographic data) and the dependent variables (crime data). By looking through the correlation result in the table, I will choose a number of paired variables to explore further in the next section.

1. Individual demographic and crime feature patterns and correlation

This sub-section will do some individual feature exploration and visualization on a number of paired features chosen based on the previous performed correlation table.

1.1 "UnemploymentRate" as potential predictor variable of "TotalCrimesPer100000Capita"

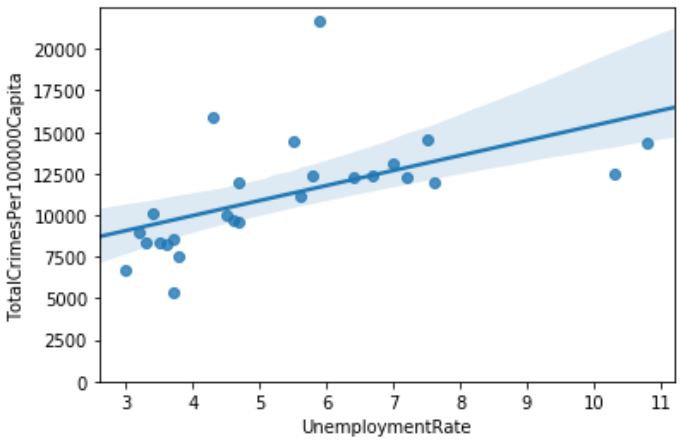

Figure 1.1: Scatterplot over "UnemploymentRate" and "TotalCrimesPer100000Capita".

	UnemploymentRate	TotalCrimesPer100000Capita
UnemploymentRate	1.000000	0.558053
TotalCrimesPer100000Capita	0.558053	1.000000

Table 1.1: Correlation table of "UnemploymentRate" and "TotalCrimesPer100000Capita".

```
The Pearson Correlation Coefficient is 0.5580529511839847  with a P-val
ue of P = 0.003051726824910994
```

Printscreen 1.1: Pearson Correlation coefficient and P-value table of “UnemploymentRate” and “TotalCrimesPer100000Capita”.

Discussion 1.1: The scatterplot shows a positive linear relationship between "UnemploymentRate" and "TotalCrimesPer100000Capita". The correlation confirms that. Meaning that when a county's unemployment rate increases the reported total crimes (per 100.000 capita) tends to increase as well.

Since the p-value is < 0.05 it is moderate evidence that the correlation between UnemploymentRate and TotalCrimesPer100000Capita is statistically significant, although the linear relationship isn't extremely strong (~0.558)

1.2 CostPerCapita" as potential predictor variable of "TheftRobberyEtcPer100000Capita"

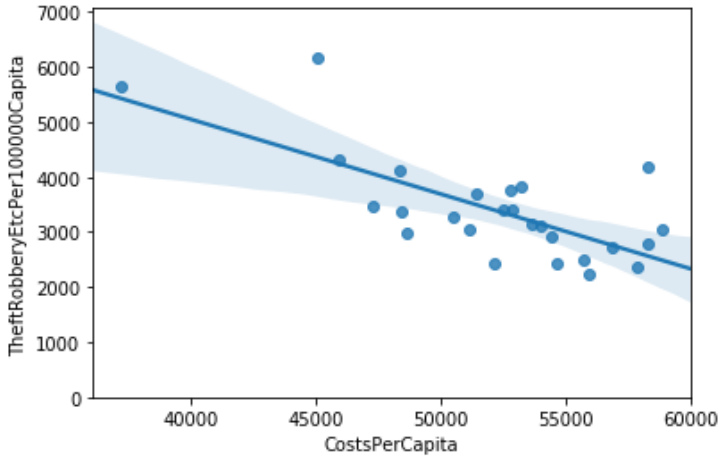

Figure 1.2 Scatterplot over “CostsPerCapita” and “TheftRobberyEtcPer100000Capita”.

	CostsPerCapita	TheftRobberyEtcPer100000Capita
CostsPerCapita	1.000000	-0.712666
TheftRobberyEtcPer100000Capita	-0.712666	1.000000

Table 1.2: Correlation table of “CostsPerCapita” and “TheftRobberyEtcPer100000Capita”.

```
The Pearson Correlation Coefficient is -0.7126664964464178  with a P-va
lue of P = 4.4070626415430645e-05
```

Printscreen 1.2: Pearson Correlation coefficient and P-value table of “CostsPerCapita” and “TheftRobberyEtcPer100000Capita”.

Discussion 1.2: The scatterplot shows a negative linear relationship between "CostsPerCapita" and "TheftRobberyEtcPer100000Capita". The correlation confirms that. Meaning that when a county's cost per capita increases the reported thefts, robberies and similar crimes (per 100.000 capita) tends to decrease as well.

Since the p-value is < 0.001 there is strong evidence that the correlation between CostsPerCapita and TheftRobberyEtcPer100000Capita is statistically significant, the linear relationship is quite strong (~-0.713)

1.3 "SupportedRate" as potential predictor variable of "ViolationLifeDeathPer100000Capita"

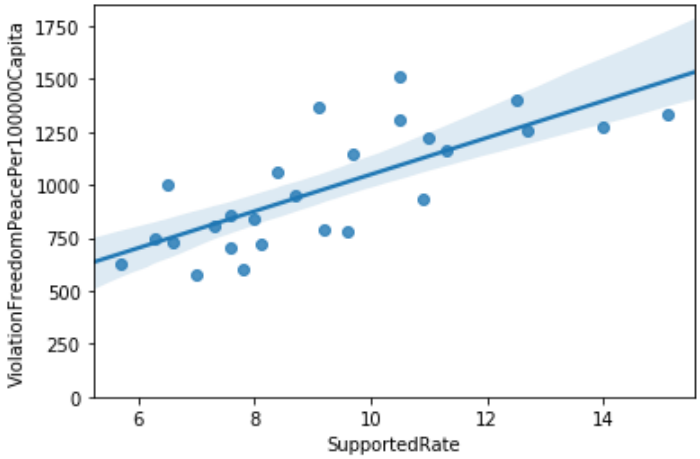

Figure 1.3: Scatterplot over “SupportedRate” and “ViolationLifeDeathPer100000Capita”.

	SupportedRate	ViolationLifeDeathPer100000Capita
SupportedRate	1.000000	0.817849
ViolationLifeDeathPer100000Capita	0.817849	1.000000

Table 1.3: Correlation table of “SupportedRate” and “ViolationLifeDeathPer100000Capita”.

```
The Pearson Correlation Coefficient is 0.8178488264591168  with a P-val
ue of P = 3.36186644284745e-07
```

Printscreen 1.3: Pearson Correlation coefficient and P-value table of “SupportedRate” and “ViolationLifeDeathPer100000Capita”.

Discussion 1.3: The scatterplot shows a strong positive linear relationship between "SupportedRate" and "ViolationLifeDeathPer100000Capita". The correlation confirms that. Meaning that when a county's supported rate increases the reported violation of life and death crimes (per 100.000 capita) tends to increase as well.

Since the p-value is < 0.001 there is strong evidence that the correlation between SupportedRate and ViolationLifeDeathPer100000Capita is statistically significant, the linear relationship is strong (~0.818)

1.4 "UnemploymentRate" as potential predictor variable of "ViolationFreedomPeacePer100000Capita"

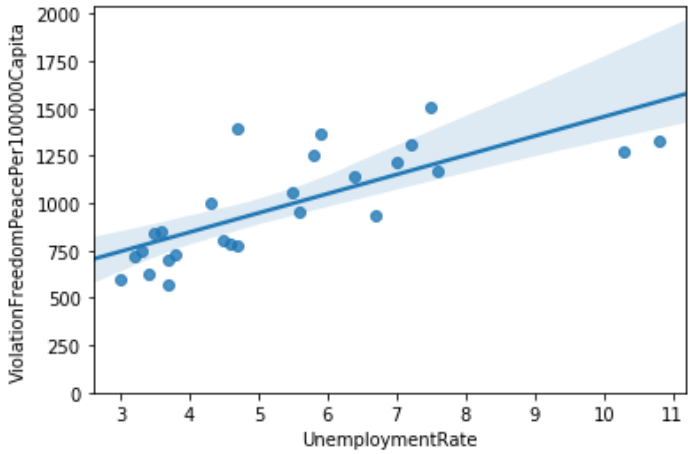

Figure 1.4: Scatterplot over "UnemploymentRate" and "ViolationFreedomPeacePer100000Capita".

	UnemploymentRate	ViolationFreedomPeacePer100000Capita
UnemploymentRate	1.000000	0.758752
ViolationFreedomPeacePer100000Capita	0.758752	1.000000

Table 1.4: Correlation table of "UnemploymentRate" and "ViolationFreedomPeacePer100000Capita".

```
The Pearson Correlation Coefficient is 0.7587519649376838  with a P-val
ue of P = 7.041049805811567e-06
```

Printscreen 1.4: Pearson Correlation coefficient and P-value table of "UnemploymentRate" and "ViolationFreedomPeacePer100000Capita".

Discussion 1.4: The scatterplot shows a strong positive linear relationship between "UnemploymentRate" and "ViolationFreedomPeacePer100000Capita". The correlation confirms that. Meaning that when a county's unemployment rate increases the reported violation of freedom and peace crimes (per 100.000 capita) tends to increase as well.

Since the p-value is < 0.001 there is strong evidence that the correlation between UnemploymentRate and ViolationFreedomPeacePer100000Capita is statistically significant, the linear relationship is strong (~0.759)

1.5 Four crime features correlation to each other

	TotalCrimesPer100000Capita	ViolationLifeDeathPer100000Capita	ViolationFreedomPeacePer100000Capita	TheftRobberyEtcPer100000Capita
CrimesPer100000Capita	1.000000	0.756402	0.773029	0.788633
eDeathPer100000Capita	0.756402	1.000000	0.932766	0.280327
lPeacePer100000Capita	0.773029	0.932766	1.000000	0.405169
eryEtcPer100000Capita	0.788633	0.280327	0.405169	1.000000

Table 1.5: Correlation table of four crime features

Discussion 1.5: The correlations between the DataFrame's five crime features; 1. "TotalCrimesPer100000Capita", 2. "ViolationLifeDeathPer100000Capita", 3. "ViolationFreedomPeacePer100000Capita" and 4. "TheftRobberyEtcPer100000Capita" are shown above.

The result shows, as can be expected, that there are high positiv correlation between 1. and 2., 3. and 4. respectively. Meaning that if 1. "TotalCrimesPer100000Capita" increases, each of the four other features tends to increase as well. That makes sense.

Interesting to note is that crime feature 4. has a low positive correlation with 2. and 3. Meaning that an increase of 4. "TheftRobberyEtcPer100000Capita" leads to a slightly increase of 2. and 3. but not as much as one could expect.

2. Folium map visualization and Foursquare police station locations

This section will use Folium to visualize the 26 counties and grade the color of its region based on the four different crime data. Also, by calling Foursquare and limit the search for a specific category id, all police stations in the area of the 26 counties was received. After some manipulation and cleaning of the response, a total of 13 police stations was added to the four Folium maps to show the crime data in relation to the police station locations.

2.1 Folium map of "TotalCrimesPer100000Capita" with Foursquare police station locations

This section creates a basic map over the 26 counties and color graded by its "TotalCrimesPer100000Capita" based on GeoJson data.

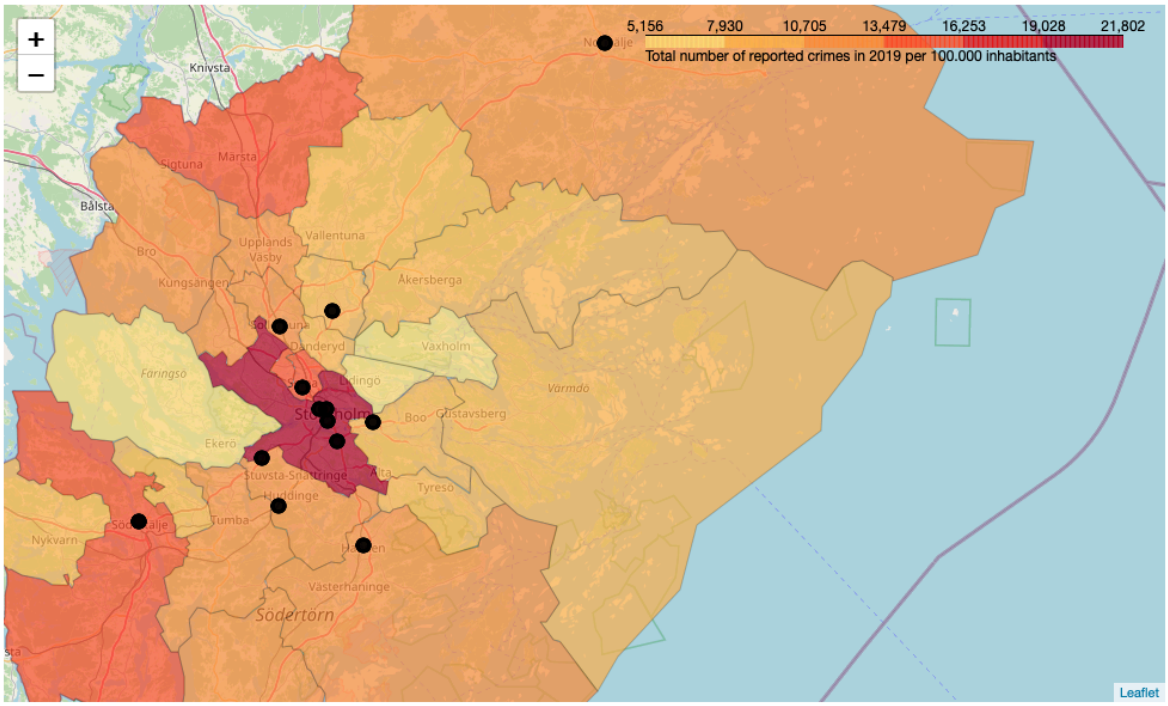

Figure 2.1: Folium map to visualize the 26 counties, color graded by "TotalCrimesPer100000Capita".

2.2 Folium map of "TheftRobberyEtcPer100000Capita" with Foursquare police station locations

This section creates a basic map over the 26 counties and color graded by its "TheftsRobberyEtcPer100000Capita" based on GeoJson data.

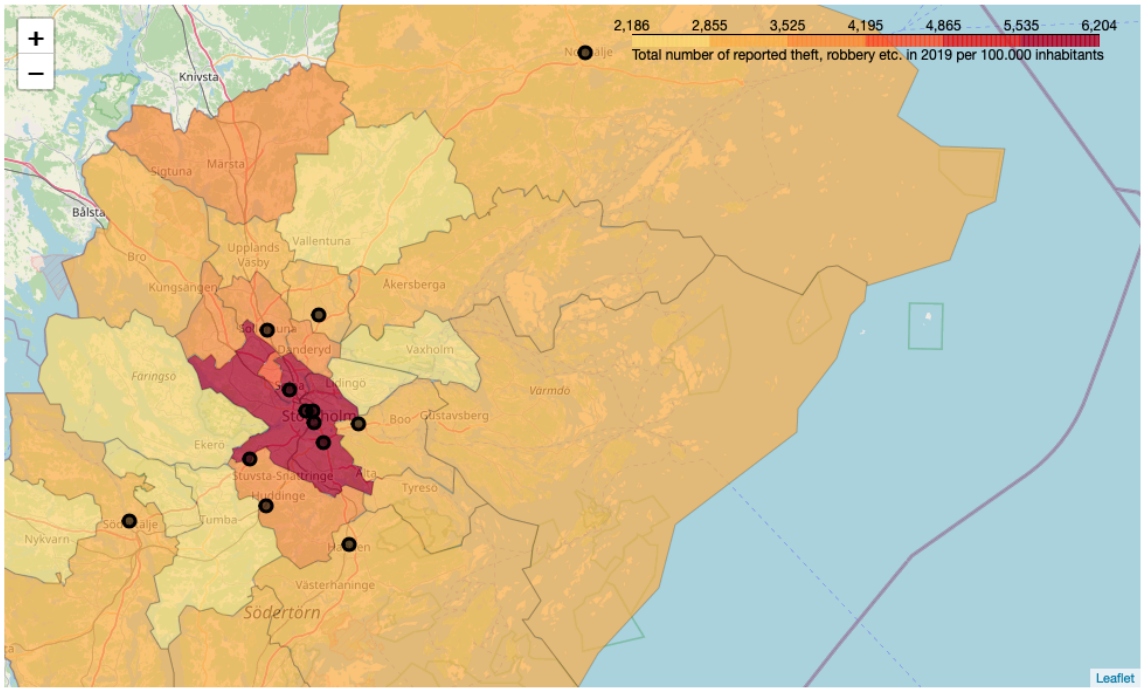

Figure 2.2: Folium map to visualize the 26 counties, color graded by "TheftRobberyEtcPer100000Capita".

2.3 Folium map of "ViolationLifeDeathPer100000Capita" with Foursquare police station locations

This section creates a basic map over the 26 counties and color graded by its "ViolationLifeDeathPer100000Capita" based on GeoJson data.

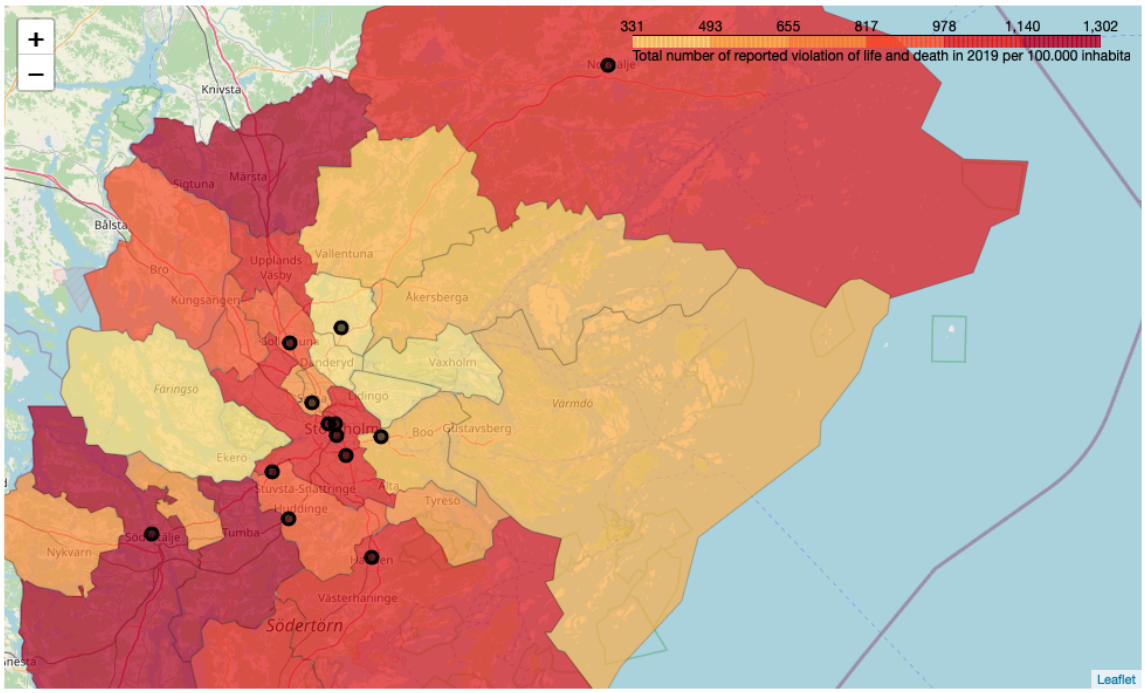

Figure 2.3: Folium map to visualize the 26 counties, color graded by "ViolationLifeDeathPer100000Capita".

2.4 Folium map of "ViolationFreedomPeacePer100000Capita" with Foursquare police station locations

This section creates a basic map over the 26 counties and color graded by its "ViolationFreedomPeacePer100000Capita" based on GeoJson data.

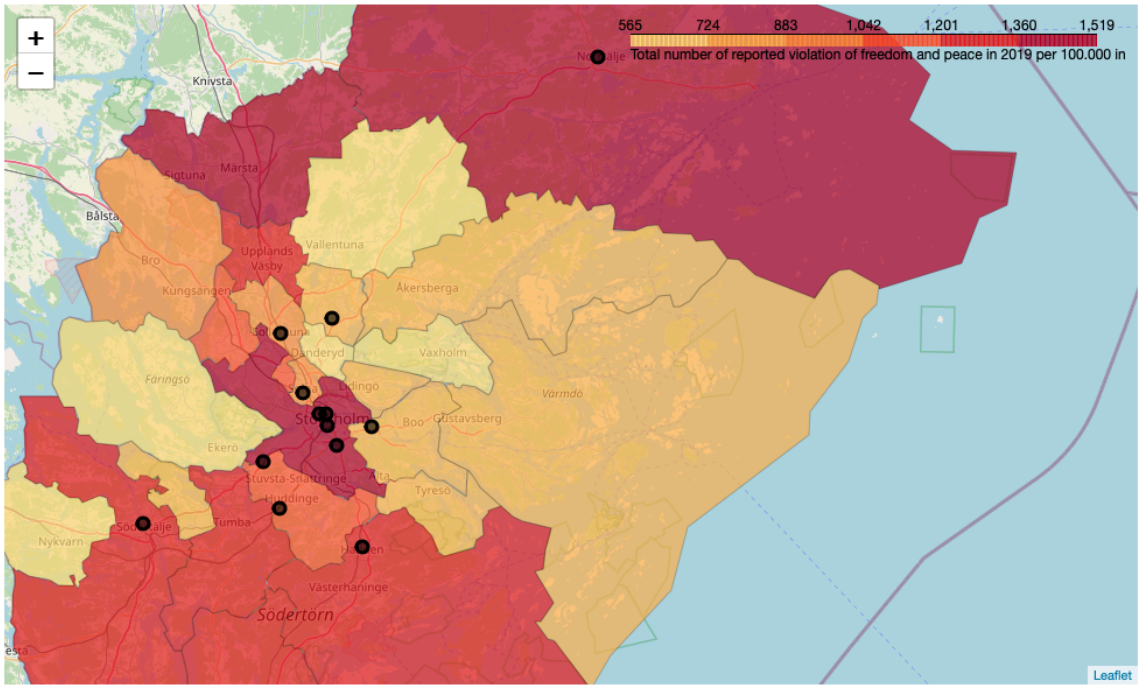

Figure 2.4: Folium map to visualize the 26 counties, color graded by "ViolationFreedomPeacePer100000Capita".

Discussion 2: The Folium maps with the additional Foursquare police station locations shows that there is a high concentration of police stations near the center of Stockholm City. That makes sense since the population is by far the highest there (~975.000). Hence, there are large areas, especially in the eastern and north western counties, where there are no police stations and large distances to the nearest one.

When taking the color of each county in consideration for the four different crime data there are some interesting findings.

Figure 2.1 shows that the highest number of reported crimes in 2019 per 100.000 inhabitants are in county Stockholm (central), followed by Södertälje (south west) and Sigtuna (north west). The later county does not have a police station nearby.

Figure 2.2 shows that highest number of reported thefts, robbery etc. in 2019 per 100.000 inhabitants is in county Stockholm (central). The other 25 counties have relatively low rates within this category of crime.

Figure 2.3 shows that the highest number of reported violations of life and death in 2019 per 100.000 inhabitants is in county Södertälje (south west) and Sigtuna (north west), closely followed by several other counties.

Figure 2.4 shows that the highest number of reported violations of freedom and peace in 2019 per 100.000 inhabitants is in county Stockholm (central), Sigtuna (north west) and Norrtälje (north), closely followed by several other counties.

Discussion

The introduction section explained the project's background and purpose of finding a way to understand which of the 26 counties in Region Stockholm that is better to move to in terms of crime rates explained by demographic data.

Four different crime data was explored during this project; "TotalCrimesPer100000Capita", "TheftsRobberyEtcPer100000Capita", "ViolationLifeDeathPer100000Capita" and "ViolationFreedomPeacePer100000Capita".

"TotalCrimesPer100000Capita" has a positive linear relationship with "UnemploymentRate" with a Pearson correlation value of ~0.558 and a P-value of <0.05 that gives a moderate evidence. The county of Stockholm has a high degree of total crimes but also a high number of police stations.

"TheftsRobberyEtcPer100000Capita" has a negative linear relationship with "CostsPerCapita" with a Pearson correlation value of ~-0.714 and a P-value of <0.001 that gives a strong evidence. The county of Stockholm has a high degree of thefts, robbery etc. but also a high number of police stations. The other counties have relatively low rates of crimes within this category.

"ViolationLifeDeathPer100000Capita" has a positive linear relationship with "SupportedRate" with a Pearson correlation value of ~0.818 and a P-value of <0.001 that gives a strong evidence. The counties of Södertälje (south west) and Sigtuna (north west), followed by a several other counties, have a high rate within this category of crime. The latter county does not have a police station nearby.

"ViolationFreedomPeacePer100000Capita" has a positive linear relationship with "UnemploymentRate" with a Pearson correlation value of ~0.759 and a P-value of <0.001 that gives a strong evidence. The county of Stockholm (central), Sigtuna (north west) and Norrtälje (north) have high rates within this category of crime, closely followed by other counties.

Conclusion

The conclusion of this project is that the crime rate in several categories varies between the 26 counties. Hence, the counties of Stockholm, Sigtuna and Södertälje tends to have the highest crime rates of the four chosen and exploring crime categories. Furthermore, a county's unemployment rate, costs per capita and supported rate are demographic features that has a high correlation with the four crime categories. These conclusions can be of interest for individuals, couples and families moving to any of the 26 counties. There are many possibilities for improvement in this exploring project and it should be seen as a first step of understanding demographic and crime data for the 26 counties in Region Stockholm.