University of Tartu
Faculty of Science and Technology
Institute of Ecology and Earth Sciences
Department of Geography

Report in Spatial Data Studio

Author: Sourav Karmakar
Supervisor: Evelyn Uuemaa PhD

Tartu 2021

Table of contents

1 Introduction

This report will describe the crime point pattern analysis in Estonia as a part of spatial data quality and data aggregation task. Crime dataset is about the offences against the public order with location in 2012, that was collected from Estonian Police and Border Guard Board open data page where the source of necessary shape file of Estonian population, road network and administrative units was Estonian Statistics. Task 1 focuses on the general and statistical features of the dataset through quality checking and plotting basic histograms. Task 2 aims to figure out the spatial distribution and patterns of offences in relation with administrative units through clustering and kernel density. Furthermore, Task 3 seeks to what extents crime incidents correlates with population data by using correlation analysis. Software used for this work is QGIS 3.16.10, Microsoft Excel and Google Colab. All the maps are plotted and analysed in Estonian National Coordinate System, EPSG:3301

2 Description of work

2.1 Task 1 - Exploring general and statistical feature of the datasets

First and foremost, crime dataset was imported in QGIS and then reprojected to EPSG 3301. In order to maintain the privacy protection rules, exact location of the points was represented as a centroid of aggregated grid cell of different resolution. In the precision column of attribute table which have integer value, it is observed that, some points are collected with 500 m precision, some with 1000 m precision and some has null value whose locations were not given in the data. With the help of ‘select feature using an expression’ tool, it was observed that, the points having higher precision (500), mostly located at urban areas specially around the city centre. The dataset has 15230 crime incidents so far. As the location of 140 entries were not given in the dataset, only 15090 points were visible in the map.

Furthermore, the field type of ‘Date’ is date whereas ‘Time’, ‘WeekDay’ and ‘MonetaryDamage’ have sting value. ‘MonetaryDamage’ was intended to show the range of the damage, therefore it was an integer.

In order to visualise the crime distribution in histogram, mentioned columns were converted into ‘month’, ‘day_of_week’, ‘hour’ with field type integer. Range of ‘MonetaryDamage’ was reclassified with 5 classes and converted into integer 0 to 4

based on the range. Later the dataset was exported to excel and then analysed in 'Google-Colab' notebook.

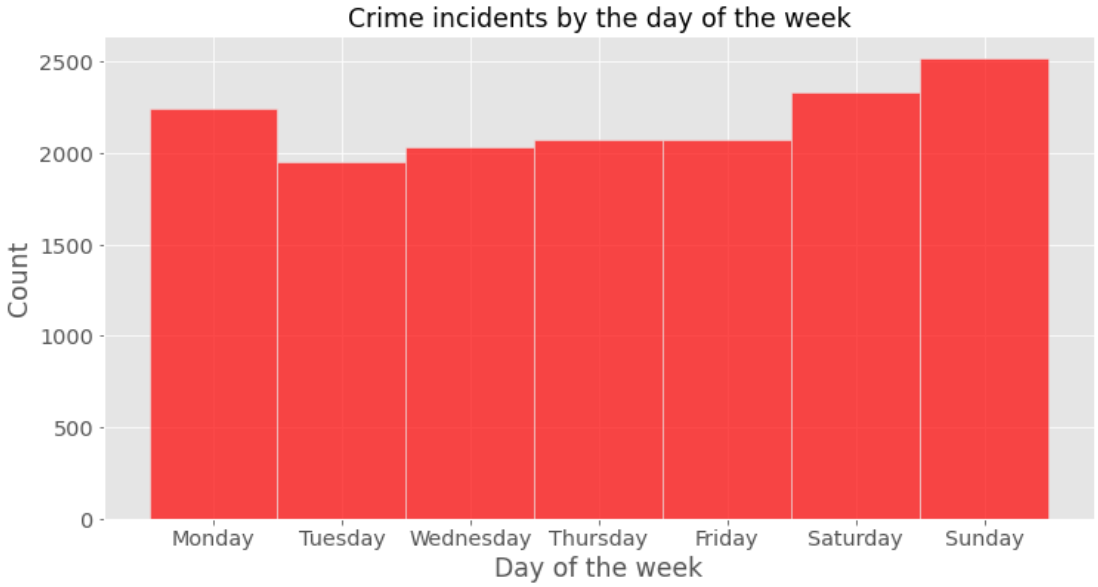

Figure 1: Crime incidents by the day of week in 2012

Figure 1 depicts the relation of crime incidents with the day of the week. Criminal offences were peaked on the weekend compared to weekdays. Around 2500 incidents has taken place on Sunday and Saturday was on the second position in 2012. On the other hand, Tuesday possessed the lowest position of below 2000 incidents. Wednesday, Thursday and Friday had almost the same number of counts, just above 2000, however, first day of the week had almost 2200 criminal offences that were taken place in Estonia in the year 2012.

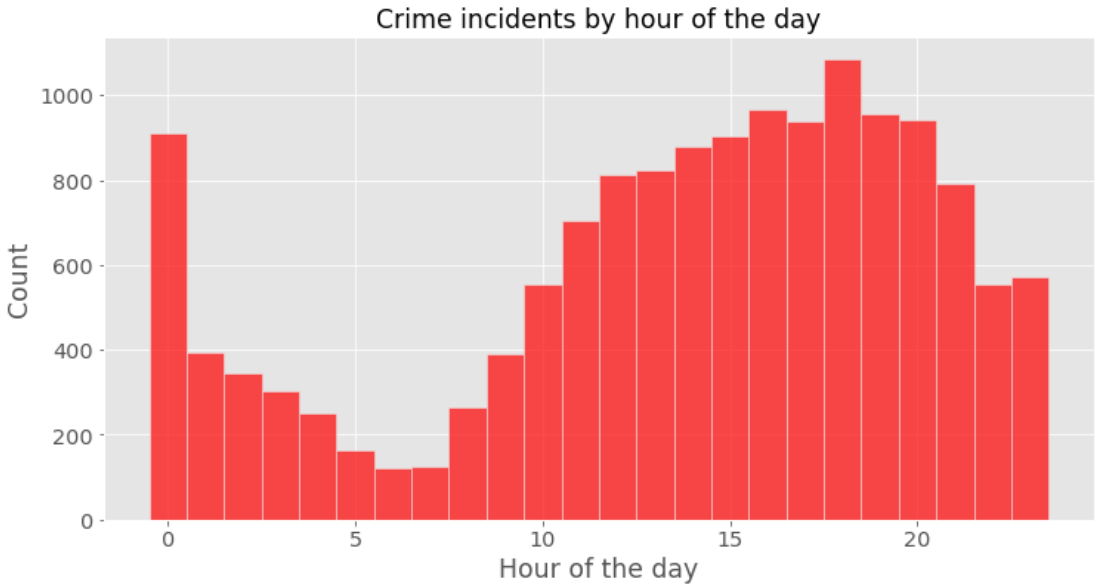

Figure 2: Crime incidents by hour in 2012

According to Figure 2, most of the offences were carried out at night rather than in day. In the morning at 6am - 8am, crime rate was the lowest and proportionally increased with the daylight. Contrarily, it was the highest at 6pm. It can be said that 4pm to 8pm was the peak hour for criminal offences in 2012. 12am to 1am was another suitable time for the miscreants and after 1am crime incidents used to fall in that year.

Figure 3: Crime incidents by month of the year 2012

Figure 3 illustrate the crime incidents that has taken place by different months of the year. Overall, most of the occurrences were in August month, where February had the

lowest recorded crimes in 2012. Interestingly, crimes started declining with the arrival of winter months after August. For the months of March, April and May, crime incident remained same to almost 1300 counts and in January, June and July it was around 1250 occurrences.

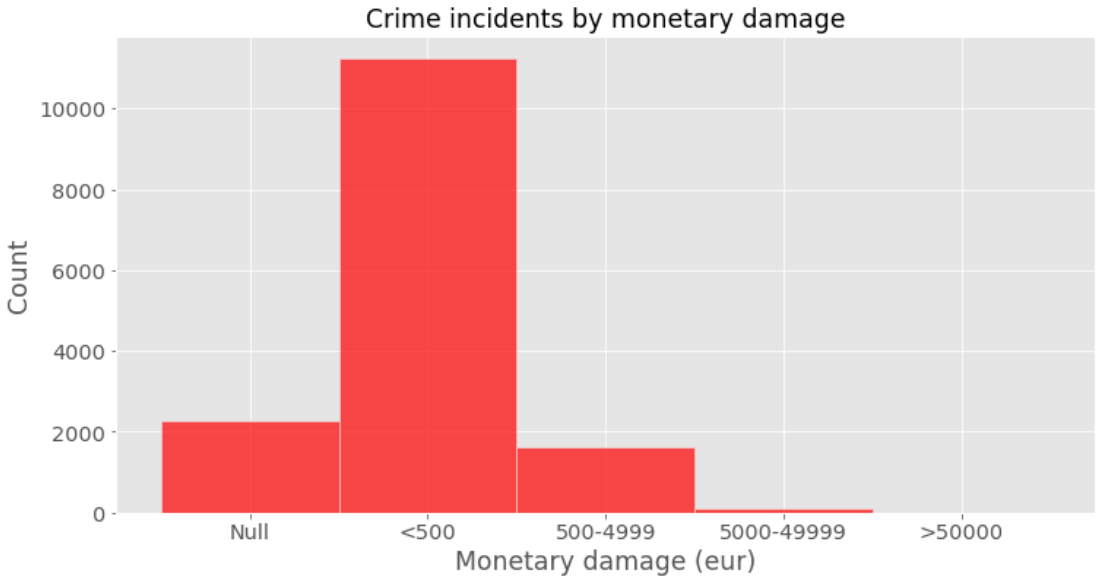

Figure 4: Crime incidents by monetary damage in 2012

Figure 4 displays the criminal offences and their adjacent economic loss in 2012. Almost 2000 incidents did not have recorded monetary damages. Most of the incidents, around 11000, costed a damage of below 500 Euro. Near about 1800 criminal offences attributed to the economic damage of 500 Euro to 4999 Euro. There were a very few crime incidents in 2012 which costs 5000 Euro to 49999 Euros. Only 4 counts had listed as monetary damage of more than 50000 Euro that were not even visible in histogram.

2.2 Task 2: Spatial patterns and distribution of the crimes

In order to get a better visualization of crime incidents, in this step, all crime points are grouped into clusters using k-means clustering. First number of clusters was set 300 which created 300 clusters for all the crime points using a cluster id and cluster size (number of points in each cluster). When the symbology was changed, all of the crime points were distributed into 300 cluster, assigned a random colour for each cluster. However, when the cluster number was reduced to 200, it grouped all crime points into 200 randomly coloured cluster. And when it reduced to 100, size of group was increased. That means, if we reduce the cluster number, it will increase the group size in each cluster.

After that, for each 300 cluster, multi points are generated in a collected cluster based on the cluster ID. As 300 clusters were set previously, the number of features was 300 in attribute table, however, in maps they did not change visually. The collected clusters grouped the points based on number of clusters. As a result, number of features changed in the attribute table.

In next step, a centroid was drawn for multi-points in each cluster. Therefore, for each cluster, we have now only one centroid points and the point holds its cluster size in the attribute table. Total number of crime points can be obtained when summing up the cluster size.

At this point, Estonian road network and population per administrative unit data were imported. In the population datasets, there were some gaps and overlaps in the shapefiles, identified using topology checker. The two gap area contains lake. So, for calculating population density, we need not to take action about this error as lake has no population. But overlaps would be problematic because it may provide faulty measure of admin unit's area. To get rid of this error, population map was digitized, and problematic islands has cut out with the help of advanced digitizing toolbar and snapping toolbar. Also, before and after digitizing, the area was calculated for all admin units so that the area differences could be obtained. Two of the polygons had issues with overlaps. The area differences (before area – after area) for 'Narva-Joesuu linn' was 635217.8 m^2 and for 'Toila vald' the area difference was 2835974.3 m^2. Then population density per sq. km was calculated for the new area and presented with colour ramp in map.

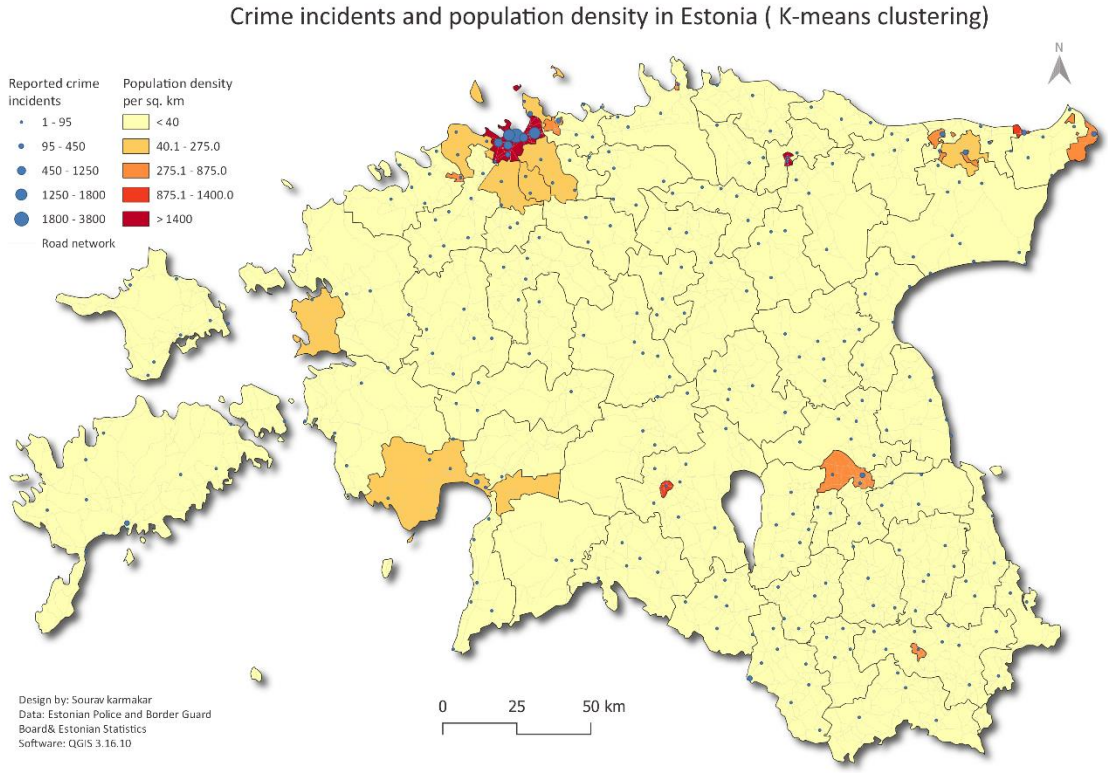

Figure 5: Map showing population density and crime hotspot in Estonia 2012

Figure 5 illustrated the population density along with the crime incidents in terms of k means clustering. The centroid of each cluster was represented here based on the cluster size. Most of the administrative units in Estonia have low population density below 40 person per square kilometres. But population density in Tallinn city was the highest more than 1400 people per km^2. Also, the size of centroid in Tallinn city was the biggest.

Next, for better visualization of crime density a heat map was created using Kernel density estimation. Radius was set to 1000 m which means within the 1000 m, all points were counted into one single cell. If the radius of heatmap was increased to 3000m, some points which were separated previously, came within the same cell.

Heatmap can also be created using symbology of the crime dataset. Figure 6 portrays the heatmap using symbology with a radius of 4 mm and maximum value 20. If the radius is increased, small areas with crime incidents were shown together with large area, they were not distinguishable separately. Further, if the kernel radius is reduced than 4, there are lots of smaller points that reduces the map visibility. Again, if the maximum number of points in the kernel increased, it took more points in group thus reduced the occurrence of colour and if the number was reduced to smaller number, the number of colour occurrence was increased. Therefore, radius 4 mm and maximum number of kernel 20 were chosen because it gave me an appropriate visual impression for a quick overview of where the crime incidents are higher and where it is lower.
Figure 6 also shows the city name that has more than 10,000 population along with Estonian border.

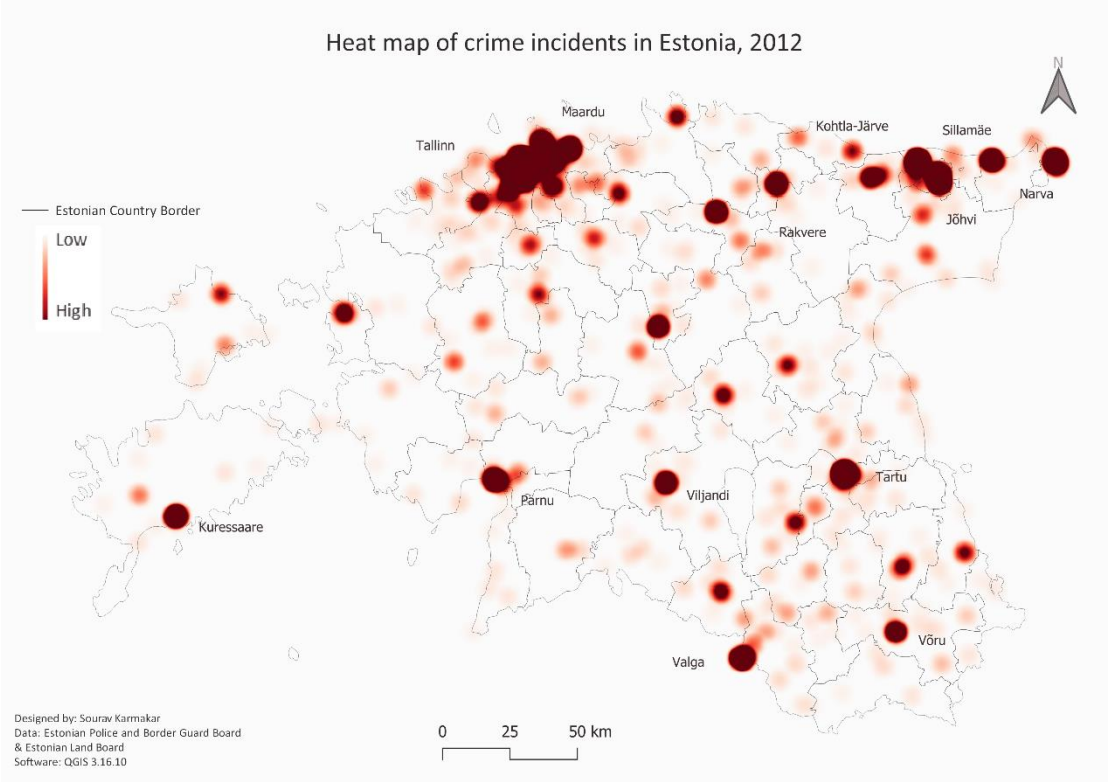

Figure 6: Heatmap of Estonian crime incidents in 2012

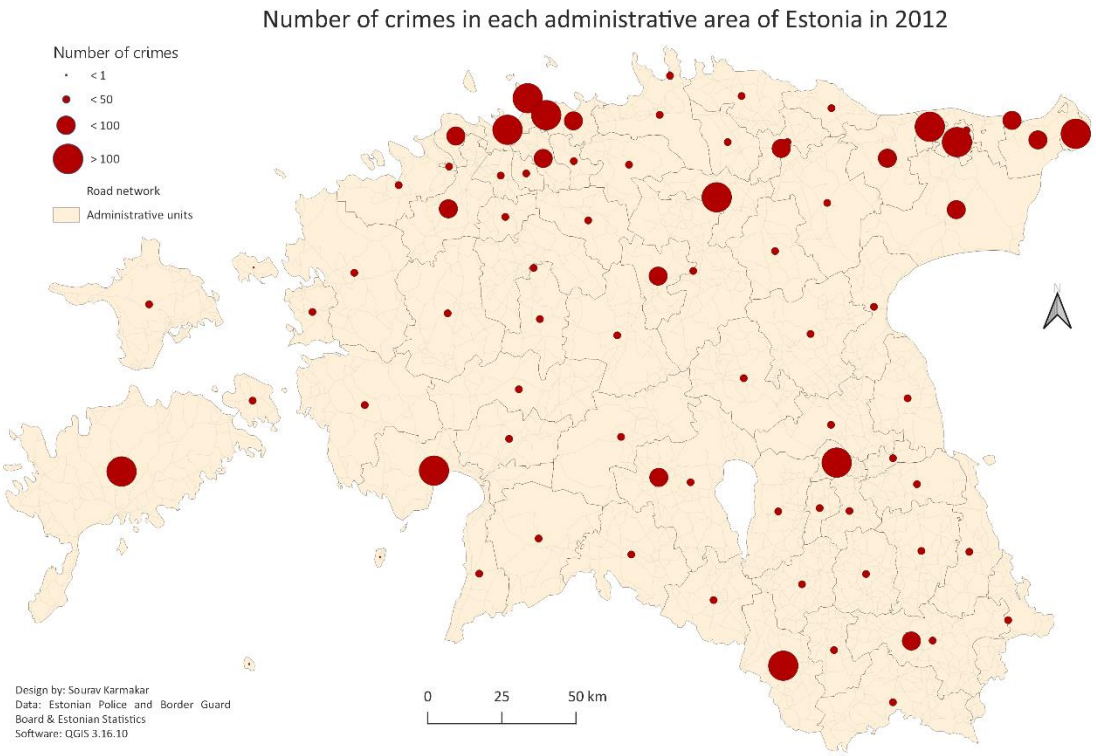

Figure 7: Crime numbers in each admin unit, Estonia 2012

Then in the next step, crime layer and population in admin unit layer were aggregated with the help of function 'aggregate' so that crime counts can be visualised in each administrative unit. After that, using the same function and layers, proportional crime incidents for each administrative area were created. Figure 7 shows the map that used proportional symbol for number of crimes per administrative areas. Most of the administrative units had less than 50 incidents in 2012. Number of crimes were higher in city areas compared to other areas. Around 11 administrative had more than 100 crime incidents in 2012.

2.3 Task 3: Analyse the crimes in relation to population

In order to test the hypothesis that most crimes have taken place in the city areas, in this task a correlation between number of crimes and the population was carried out because higher population represents the city area. Using the tool 'Join attribute by location (summary)' crime layer was joined with population layer. After joining, it was observed

from the attribute table of the joined layer that, administrative units namely ‘Vormsi vald’, ‘Ruhnu vald’ and ‘Kihnu vald’ haven’t had any crime records in 2012. Then by exporting the data into excel, the correlation was calculated.

Figure 8 shows the correlation between crimes and population size in linear scale. The R^2 value was as high as 0.95 which means 95% of the variation was explained by the model. So, there was a good correlation between the crime incident and population size. According to the linear scale, there was a proportional relation among the two variables. So it can be said that, number of crime incident increases with the increase of population size.

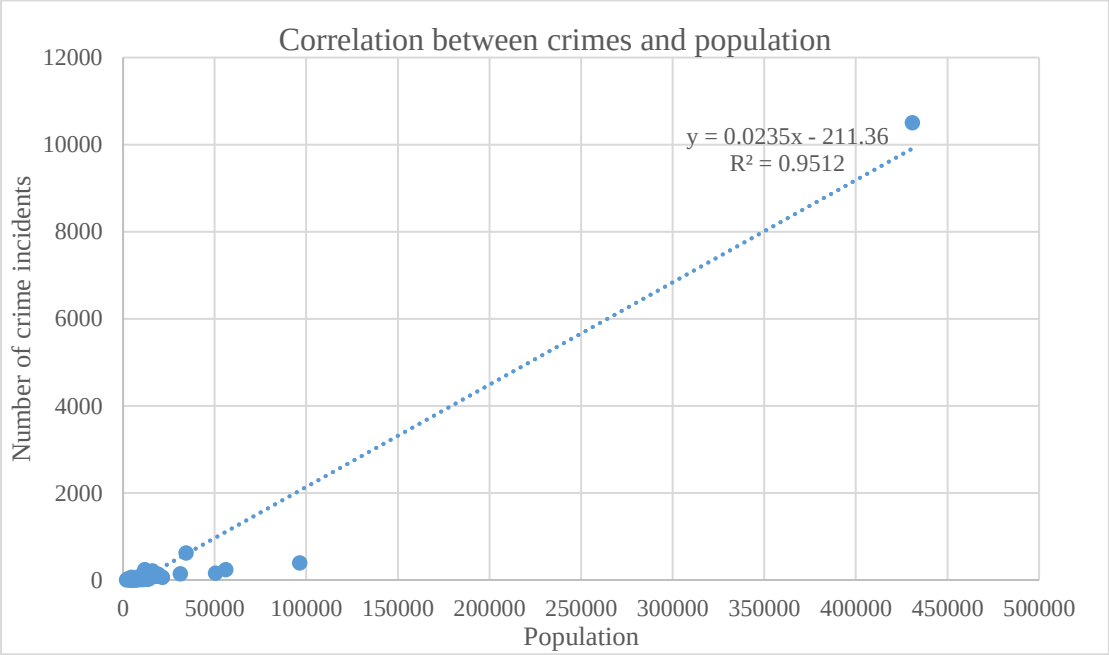

Figure 8:Relationship between population and crime incidents number (in linear scale)

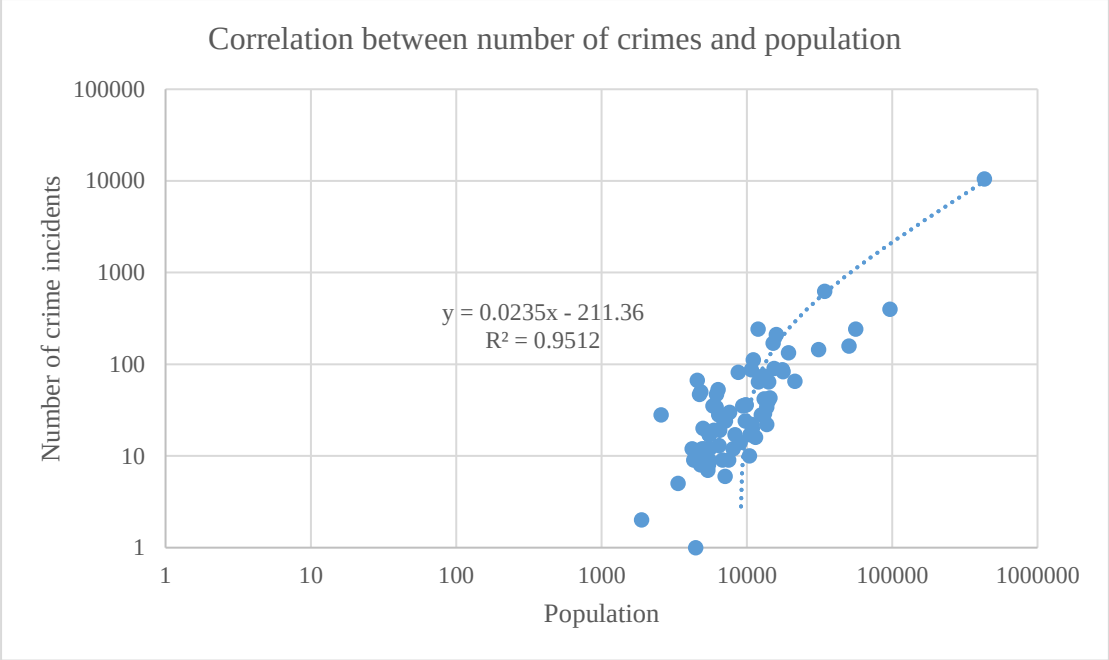

Figure 9: Relationship between population and crime incidents number (in logarithmic scale)

In linear scale, one value point stands isolated and provides a good linear relationship. However, logarithmic scale (Figure - 9) provides the better view for each point. It also explains that crime incidents increase with the increase in population size. In city area, population size and density were higher, therefore it is true that most of crime incidents has taken place in the urban areas. In previous maps it was also observed that city area has a high population density and higher crime incidents which is compatible with the correlation analysis.