Case Study 1 (CS1): Visualizing Crime
Technical Report

"On my honor, I pledge that I have neither given nor received help on this assignment."

Name: Prateek Agrawal
Computing ID: pa7sb
Date: 09/13/2015
Course: Data Mining

Data:-

The data is collected from the City of Chicago data portal https://data.cityofchicago.org/Public-Safety/Crimes-2014/qnmj-8ku6

The data contains the reported cases of crime from 01/01/2014 12:00:00am to 12/31/2014 11:58:00pm. It has 22 different fields ranging from the crime ID, Case Number, ICUR code, Time, Location, Longitude, Latitude, Description of Crime etc. The data contains 273348 reported crime cases.

Situation:

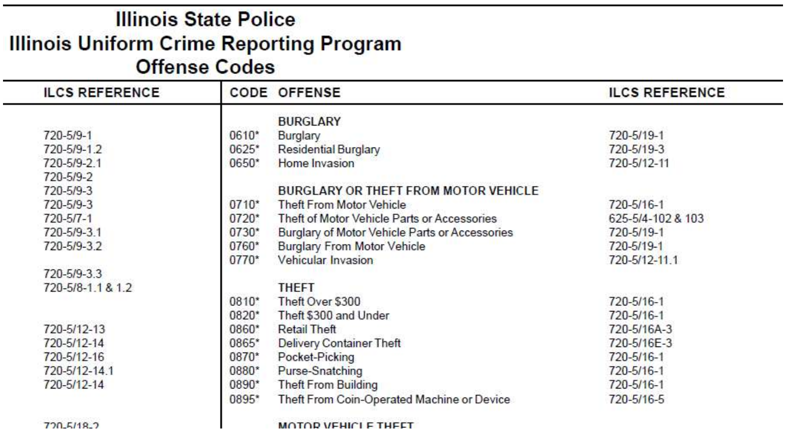
The Chicago Police Department (CPD) has limited resources (officers, patrol cars, etc.). In an effort to better deploy these resources, the CPD would like to target areas of high crime concentration. Specifically, they need answers to the following questions:

Q1.) *Is theft uniformly distributed across the city? If not, where does it concentrate?*

Before Looking at the data, the thefts will not be uniformly distributed across the city of Chicago, as there will be less cases of reported theft in the areas where O'Hare international airport i.e. the farthest north-east and Harborside international golf centre (towards south) are located, while the major theft concentration will be present in the downtown and the uptown area where major schools and landmarks are located in the city of Chicago.

Data Clean-up:-

Chicago state Police department have a created an Illinois Uniform Crime Report Program Offence codes i.e. IUCR codes (present in the data collected) which categories each type of crime reported. Therefore using the IUCR code list, I have created a sample of the data with only those IUCR codes that are only used for crimes under the "Theft", "Burglary" and "Burglary and theft from motor vehicle" type.(the required code changes has been implemented in the CrimeUtil.R file)



Illinois State Police
Illinois Uniform Crime Reporting Program
Offense Codes

ILCS REFERENCE	CODE	OFFENSE	ILCS REFERENCE
		BURGLARY	
720-5/9-1	0610*	Burglary	720-5/19-1
720-5/9-1.2	0625*	Residential Burglary	720-5/19-3
720-5/9-2.1	0650*	Home Invasion	720-5/12-11
720-5/9-2			
720-5/9-3		BURGLARY OR THEFT FROM MOTOR VEHICLE	
720-5/9-3	0710*	Theft From Motor Vehicle	720-5/16-1
720-5/7-1	0720*	Theft of Motor Vehicle Parts or Accessories	625-5/4-102 & 103
720-5/9-3.1	0730*	Burglary of Motor Vehicle Parts or Accessories	720-5/19-1
720-5/9-3.2	0760*	Burglary From Motor Vehicle	720-5/19-1
	0770*	Vehicular Invasion	720-5/12-11.1
720-5/9-3.3			
720-5/8-1.1 & 1.2		THEFT	
	0810*	Theft Over $300	720-5/16-1
	0820*	Theft $300 and Under	720-5/16-1
720-5/12-13	0860*	Retail Theft	720-5/16A-3
720-5/12-14	0865*	Delivery Container Theft	720-5/16E-3
720-5/12-16	0870*	Pocket-Picking	720-5/16-1
720-5/12-14.1	0880*	Purse-Snatching	720-5/16-1
720-5/12-14	0890*	Theft From Building	720-5/16-1
	0895*	Theft From Coin-Operated Machine or Device	720-5/16-5
720-5/18-2		MOTOR VEHICLE THEFT	

The latitude and Longitude are then converted into meters from degree using the projection "***+init = epsg: 26971***" because we have to create a map of the crime location in a 2-D diagram.

For the analysis I have used only six columns from the entire dataset namely; X-value, Y-value, Time-Stamp, Hour, Day of the week, month.

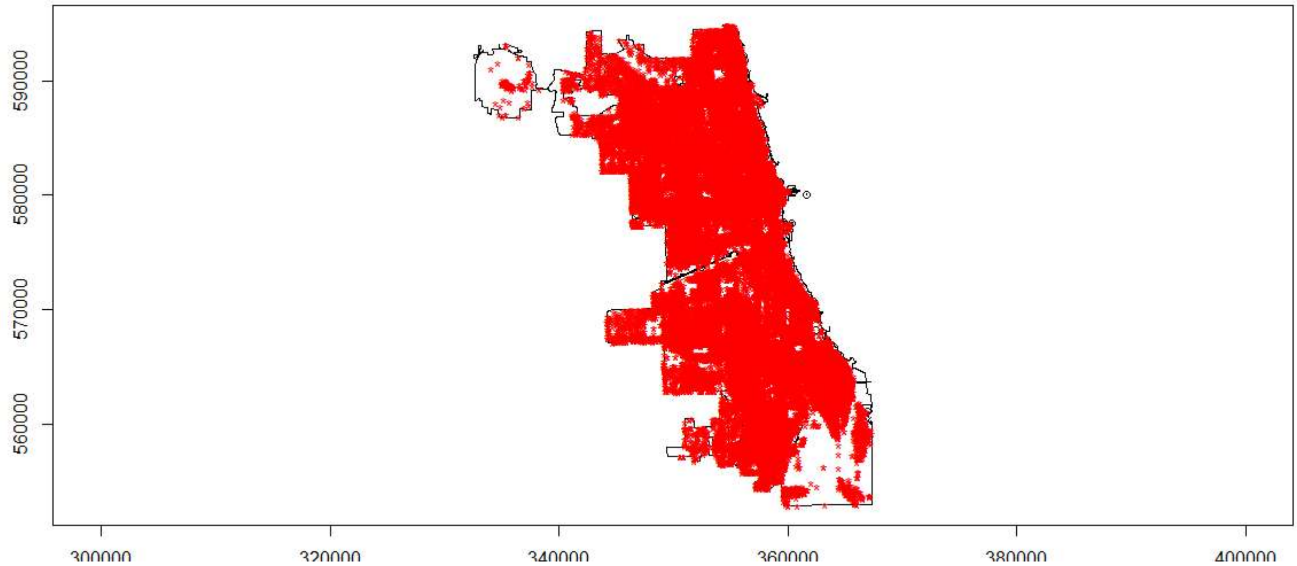
Data Analysis:-

Using the Shapefile for Chicago city, I have plotted the map of the Chicago, and highlighted the location of schools, landmarks and police stations (in green) in the city.

Then using the x and y values in the new dataset, I then plotted the location of the thefts in the city of Chicago on the same map above thus providing us an estimate of the spread of thefts in the city of Chicago.

Evidences Produced:-

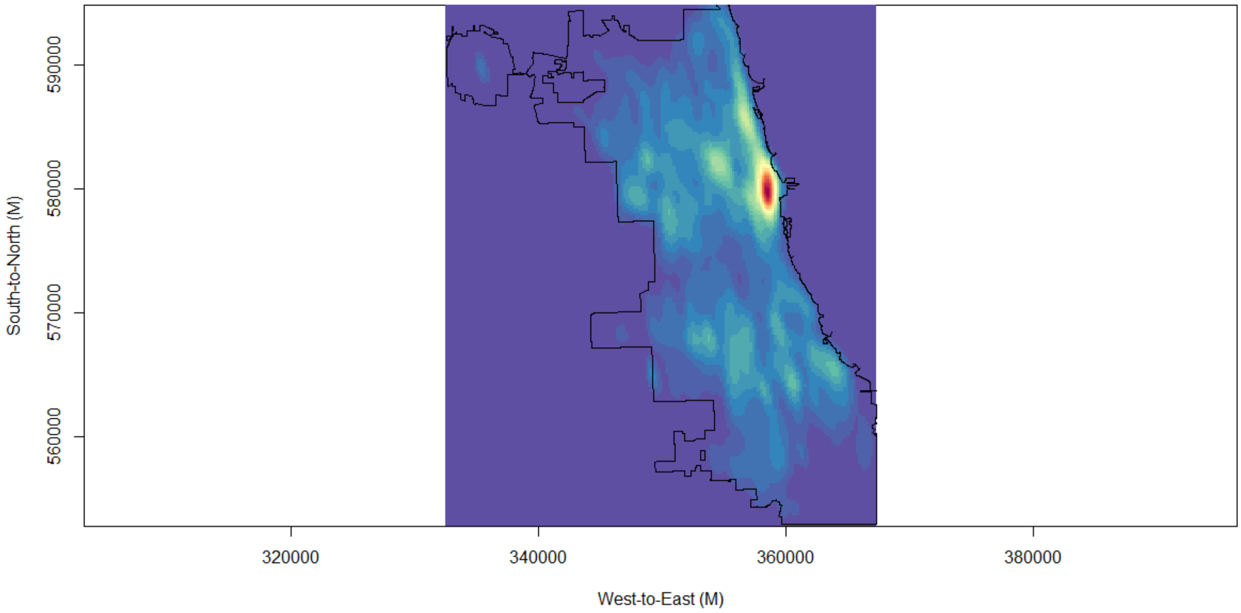
After plotting the data I saw that the spread of reported thefts was not uniform as shown in the picture below:



As guessed earlier area like the O'Hare international airport and Harborside international golf centre reporting lesser thefts as compared to the downtown area.

To find the concentration of thefts I then plotted a heat map using the 2-dimensional *Kernal Density Estimate*, taking a random 1000 sample theft reports.

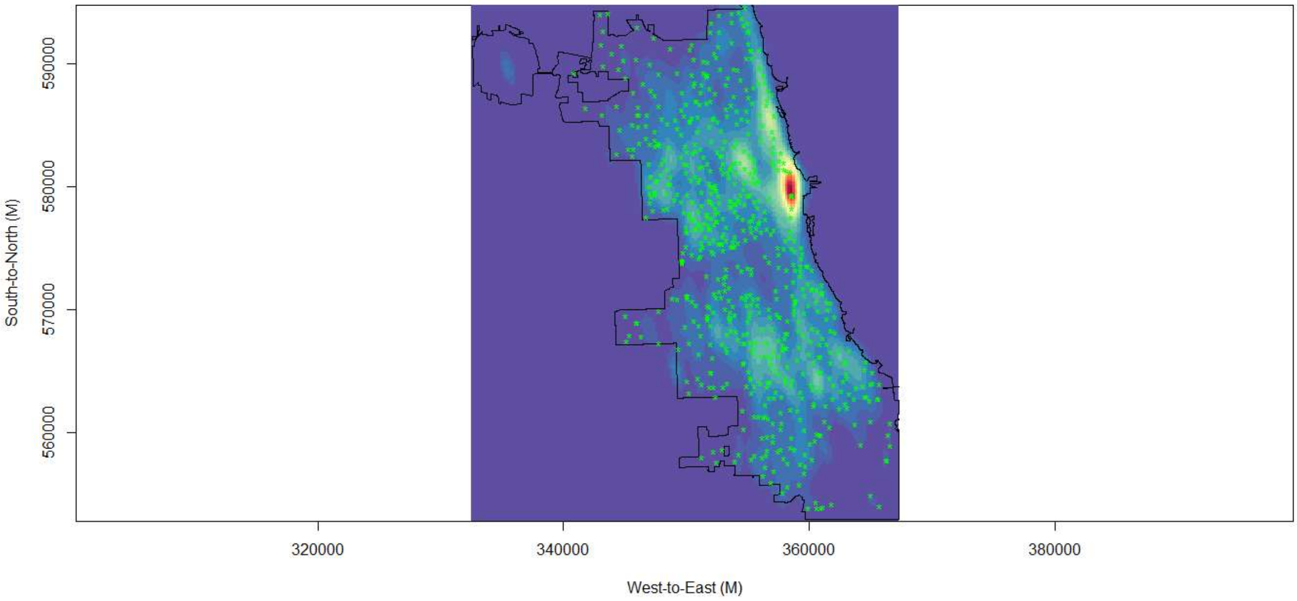
Colour scale used in the KDE is red-highest and violet-lowest.



This estimate clearly shows that the majorly of the theft concentration lies in the ***downtown area*** of the city.

Recommendation:-

After plotting the spread of police station in the city of Chicago over the heat map of the spread of thefts, it seems that the down town area of Chicago where theft concentration is present (red spot) there are not as many police station as compared to North Lawndale area of the city.



Therefore, I would recommend

1. *The CPD to establish more police stations in the high concentration areas(taking reference from the heat map)*
2. *Also, various other measures like installation of cctv cameras and spreading more awareness among the residents and tourist in the downtown area would help.*

Q2.) *Do theft concentrations look different depending on the time of day, day of week, and month of year?*

There would be more spread of the thefts during the late-night as compared to the morning, afternoon or evening hours because during those hours criminal feel that they have a better chance to getting away but overall the even in the day time higher concentration of the thefts would be found in the downtown area as the place attracts higher local and foreign tourists.

While comparing the days of the week higher thefts spread should be present on the weekends as people tend of spend their weekends outdoor, but due to higher tourist attraction the highest concentration would still be found in the downtown area of the town.

While comparing the theft concentration on months, the spread of the thefts will vary significantly because of school holidays and national holidays in some months as compared to others.

<u>*Data Cleanup*</u>:-

While comparing the time of the day, I have divided the day in 4 different time zones:

1.) The *morning time* from 5am to 11am,
2.) The *afternoon time* from 11am to 5pm,
3.) The *evening time* from 5pm to 11pm and
4.) The *late-night time* from 11pm to 5am.

Therefore, four different samples of dataset are created by dividing the ‘hour’ column in the original dataset.

Similarly, for comparing the theft concentration for different days of the week, I have created 7 different sample dataset by sub setting the original dataset based on the different values in the ‘day of the week’ column.

Where: 1 = Sunday, 2= Monday, 3= Tuesday, 4= Wednesday, 5=Thursday, 6=Friday, 7=Saturday

For comparison of theft concentration between different months of the year, 12 different datasets are created by sub-setting the dataset on basis of the values in the 'month' column of the dataset.

Data Analysis:-

Using the kernel density estimate four different heat maps are created to compare the concentration of thefts at different time of the day and are then clubbed together in the same screen for better visualization.

For different day of the week, 7 different heat maps are plotted and for comparison all the 7 plots are clubbed together.

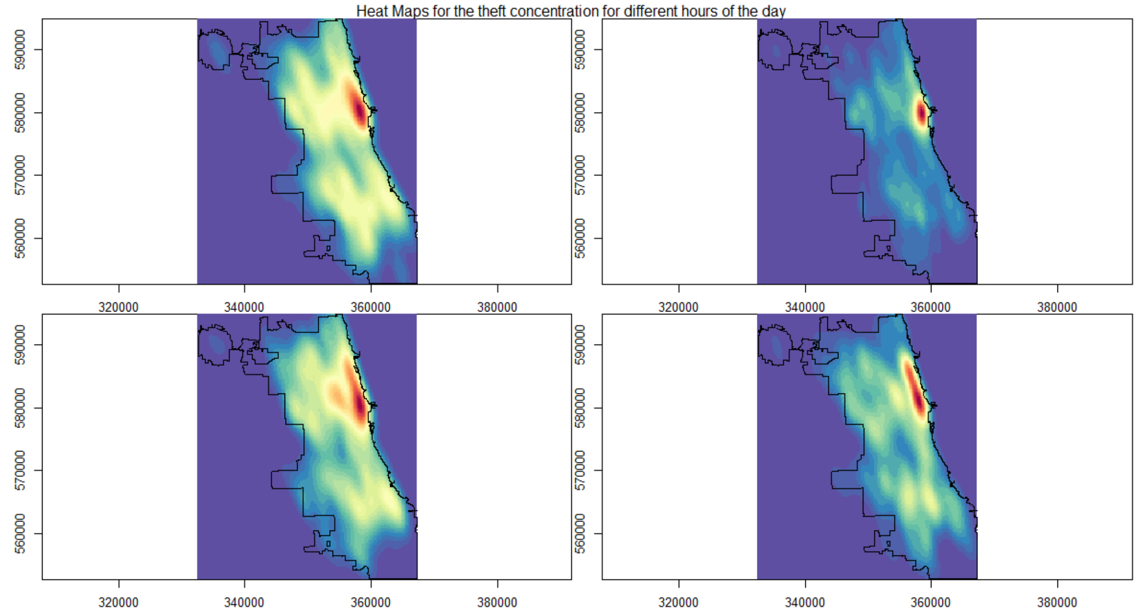
12 different heat maps for different months of year are plotted and for better visualization, heat maps are arranged in 2 different screens, in which first shows the months from January to June and the second shows the comparison for the months of July to December.

Different time of the day

Evidences Produced:-

It is seen that the spread of thefts concentration are different for different times of the day, but the highest concentration can still be found in the downtown area of the map, there are some key insights observed.

- Most spread of theft concentration is observed in the morning time and the evening time
- In the afternoon time almost all the thefts are concentrated in the downtown area of the town.
- Again the late-night time the spread of thefts concentration is more but still the higher concentration is spread across the coastal area of the town.


Heat Maps for the theft concentration for different hours of the day
590000
580000
570000
560000
320000
340000
360000
380000

The first plots is for the morning time, second for afternoon, third for evening and fourth for late-night time

Recommendations:-

I suggest the following recommendations.

1.) As highest spread of the theft concentration are spread in the morning and evening, which suggest the thefts of the form pick-pocketing, purse snatching, etc are in higher concentration as more people are outdoors during this hour, therefore, the police deploy their forces in high population concentrated zones during these hours.

2.) Proper awareness should be spread through banner and radio messages so that people are more attentive towards their belongings.

3.) Cctv cameras should be installed in areas with high population concentration.

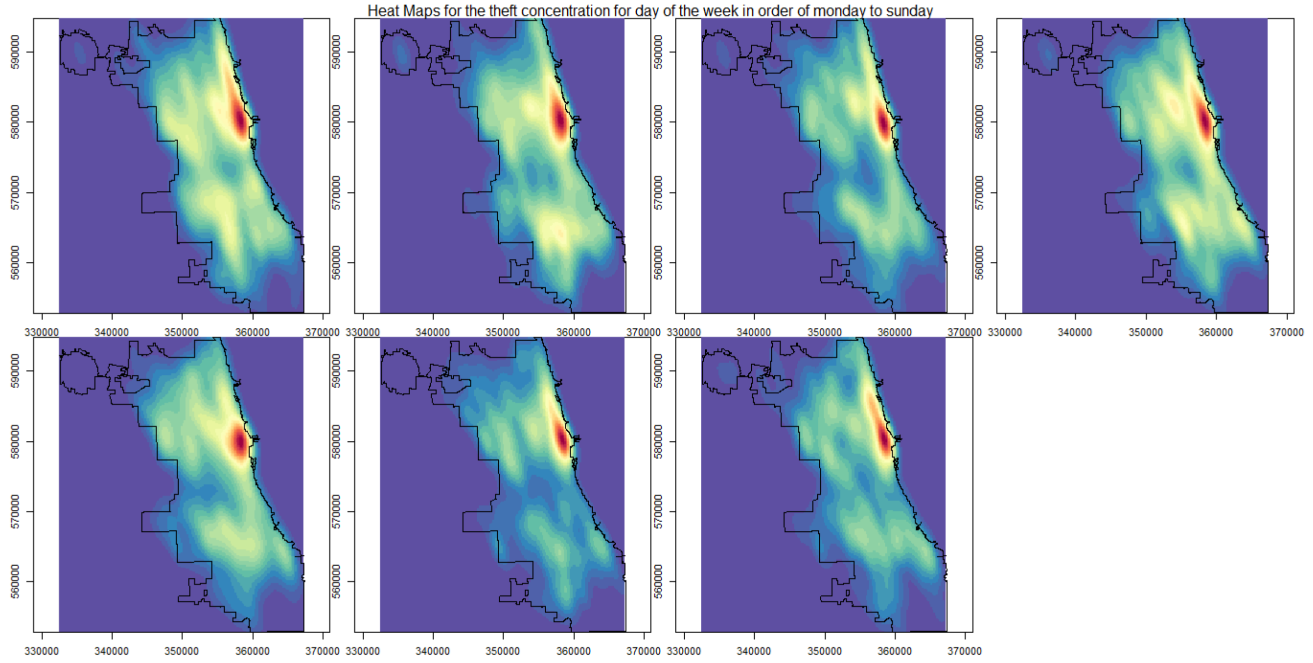
4.) As we can see the highest concentration is still in the downtown area at all hours of the day, CPD should implement all or some of the above recommendations in the downtown area at the earliest.

Different days of the week

Evidences Produced:-

It is seen that the highest concentration still remains in the downtown area of the city. But we can make some conclusions:

- On weekends the highest theft concentration is in downtown, while the spread of thefts is not high but during weekends people prefer to spend their weekends outdoors at specific landmarks and recreational location so, we can conclude that burglary is not high but other forms of thefts like motor thefts, pick-pocketing is higher.
- Higher spread can be seen on the weekdays with highest spread of theft concentration can be seen on Monday and Thursday which again implies that the concentration of thefts of the form like pick-pocketing, purse snatching, rental thefts, thefts from building etc are quite high.


Heat Maps for the theft concentration for day of the week in order of monday to sunday
590000
580000
570000
560000
330000
340000
350000
360000
370000

The plots are in order: 1.Monday, 2. Tuesday, 3. Wednesday, 4. Thursday, 5. Friday, 6. Saturday, 7. Sunday

Recommendations:-

As heat map helps us conclude that there is higher concentration of thefts like pick-pocketing, purse-snatching etc as compared to burglary, I would like to suggest following recommendations.

1. CPD needs to spread more awareness about the types of thefts prevailing in the city through radio, banner and advertisements if possible.
2. More cctv's should be installed to catch the persons of interest involved in thefts.

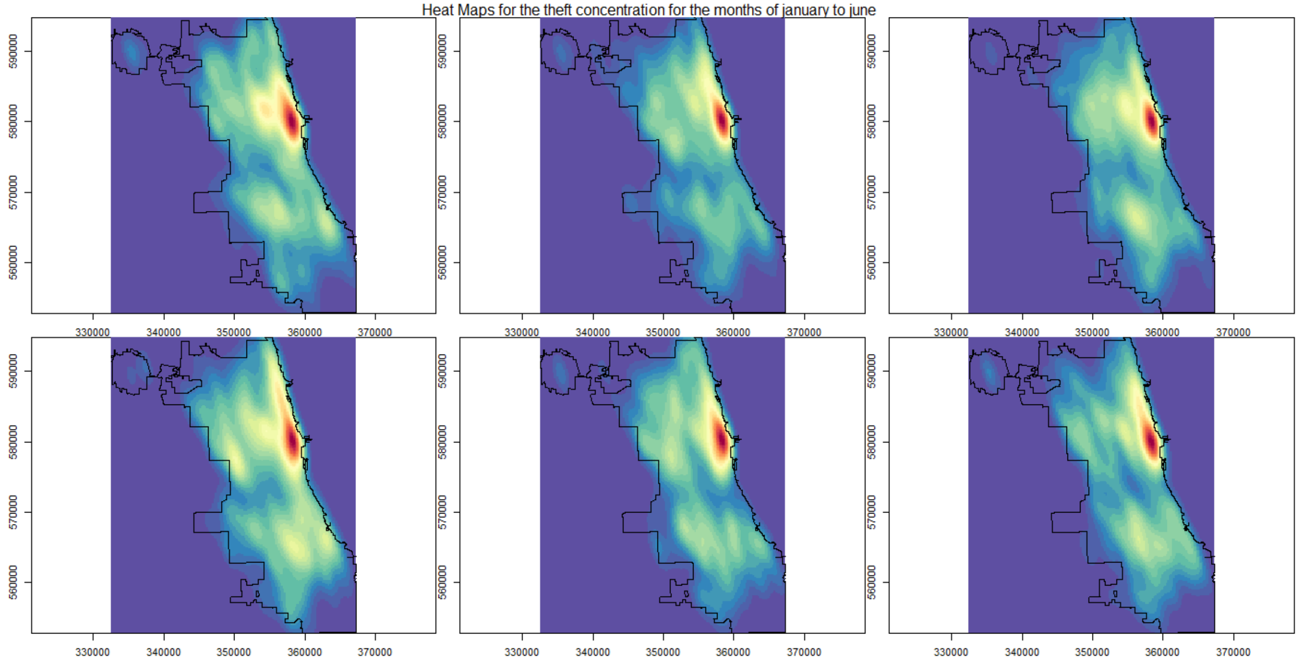
3. Police patrols should be posted in busier locations of the town in weekdays and high people concentration locations in the weekends with the help of heat maps plotted above.

Different months of the year

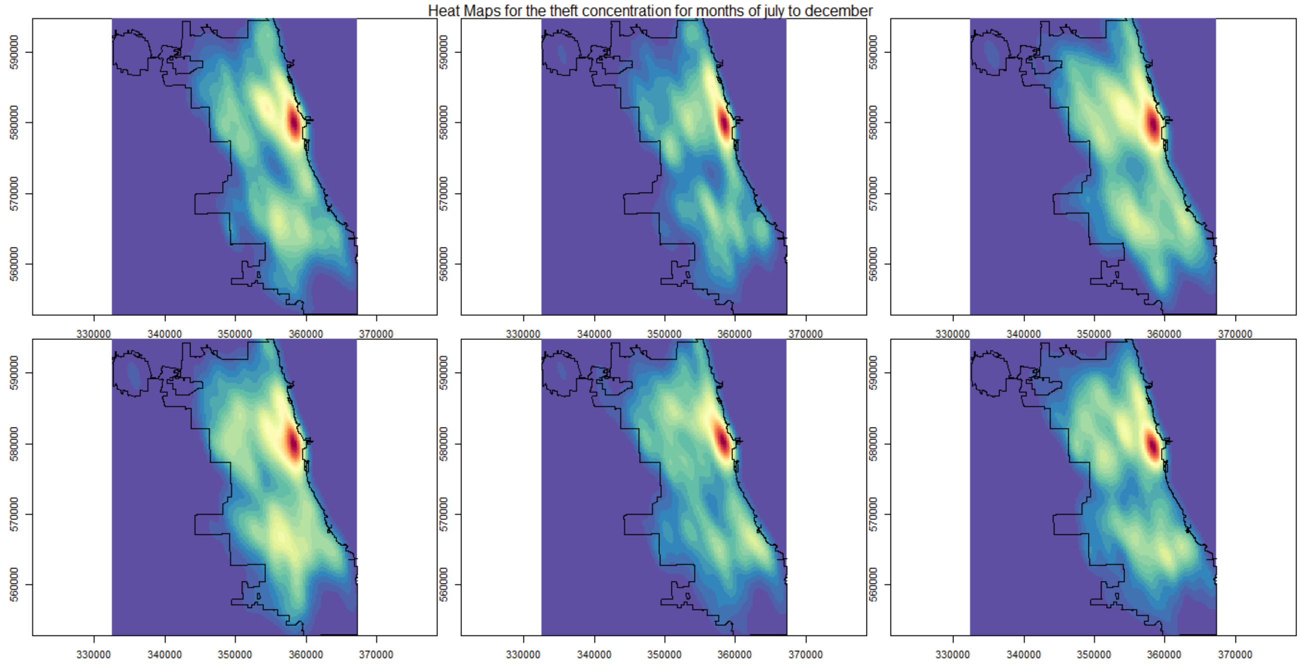
Evidences Produced:-

We can comprehend following observations from the evidences produced:

- The highest concentration hotspot still remains to be the downtown area of the town irrespective of the month of the year.
- The highest spread in the theft concentration is found in the month of April (which is the summer vacation break), September and October (These are the months in which new fall session of many schools start and lot of new and unaware students enter the city).


Heat Maps for the theft concentration for the months of january to june
590000
580000
570000
560000
330000
340000
350000
360000
370000

The plots are in order: 1. Januray, 2. Febuary, 3. March, 4. April, 5. May, 6. June.


Heat Maps for the theft concentration for months of july to december
590000
580000
570000
560000
330000
340000
350000
360000
370000

The plots are in order: 1. July, 2. August, 3. September, 4. October, 5. November, 6. December

Recommendations:-

1. As the spread of concentration of thefts are spread across the year with not much similarity but the hot spot still remains to be the downtown area of the town, we also know that the spread of police stations in the downtown area is not high as compared to the theft concentration, therefore, I recommend the CPD to establish more police stations in the downtown area.
2. In the months of April, September and October, CPD should pay extra attention to the safety of students in the different school of the city by creating awareness and establishing extra temporary police surveillance booths across the areas.

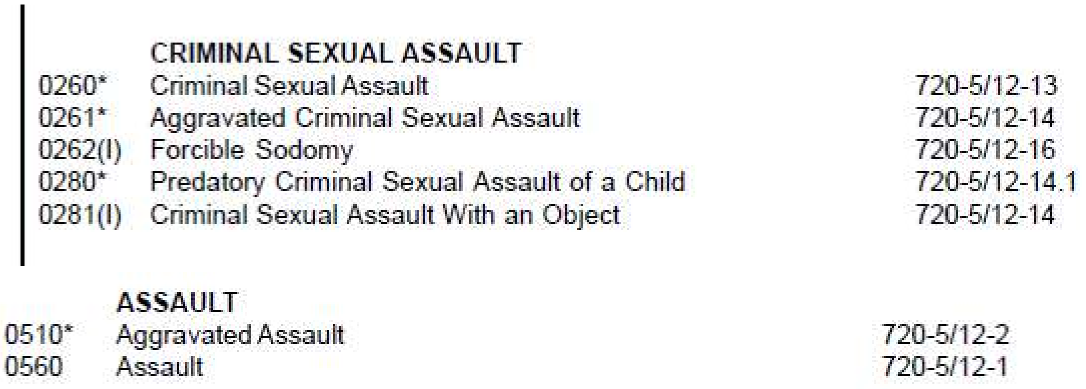
Q3.) How do assaults compare with thefts it terms of the above questions?

Prediction:

As it was the case with thefts, I can estimate that assaults will also not be uniformly distributed across the city of Chicago, as there will be less cases of reported assaults in the areas where O'Hare international airport i.e. the farthest north-east and Harborside international golf centre (towards south) are located, while the major concentration will be present in the downtown area where major schools and landmarks are located in the city of Chicago.

Data Clean-up:-

Chicago state Police department have created an Illinois Uniform Crime Report Program Offence codes i.e. IUCR codes (present in the data collected) which categories each type of crime reported. Therefore using the IUCR code list, I have created a sample of the data with only those IUCR codes that are only used for crimes under the "assault" type. (Done in the CrimeUtil.R file)



	CRIMINAL SEXUAL ASSAULT	
0260*	Criminal Sexual Assault	720-5/12-13
0261*	Aggravated Criminal Sexual Assault	720-5/12-14
0262(I)	Forcible Sodomy	720-5/12-16
0280*	Predatory Criminal Sexual Assault of a Child	720-5/12-14.1
0281(I)	Criminal Sexual Assault With an Object	720-5/12-14
	ASSAULT	
0510*	Aggravated Assault	720-5/12-2
0560	Assault	720-5/12-1

The latitude and Longitude are then converted into meters from degree using the projection "***+init = epsg: 26971***" because we have to create a map of the crime location in a 2-D diagram.

For the analysis I have used only six columns from the entire dataset namely: X-value, Y-value, Time-Stamp, Hour, Day of the week, month.

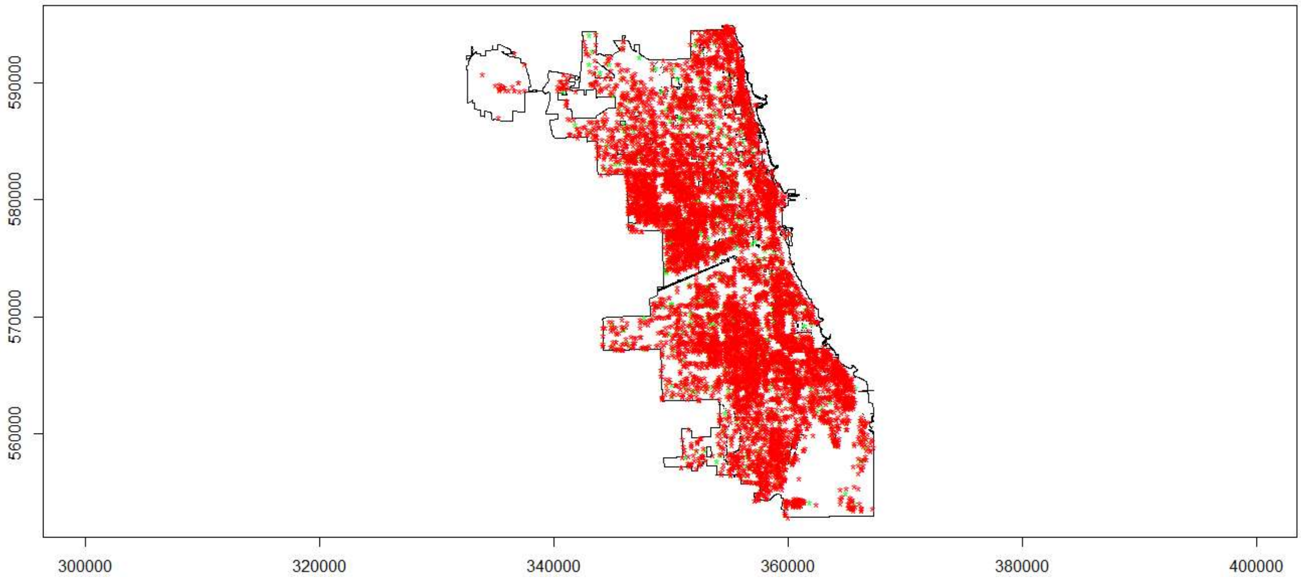
Data Analysis:-

Using the Shapefile for Chicago city, I have plotted the map of the Chicago, and highlighted the location of schools, landmarks and police stations (in green) in the city

Then using the x and y values in the new dataset, I plotted the location of the thefts in the city of Chicago on the same map above thus providing us an estimate of the spread of assaults in the city of Chicago.

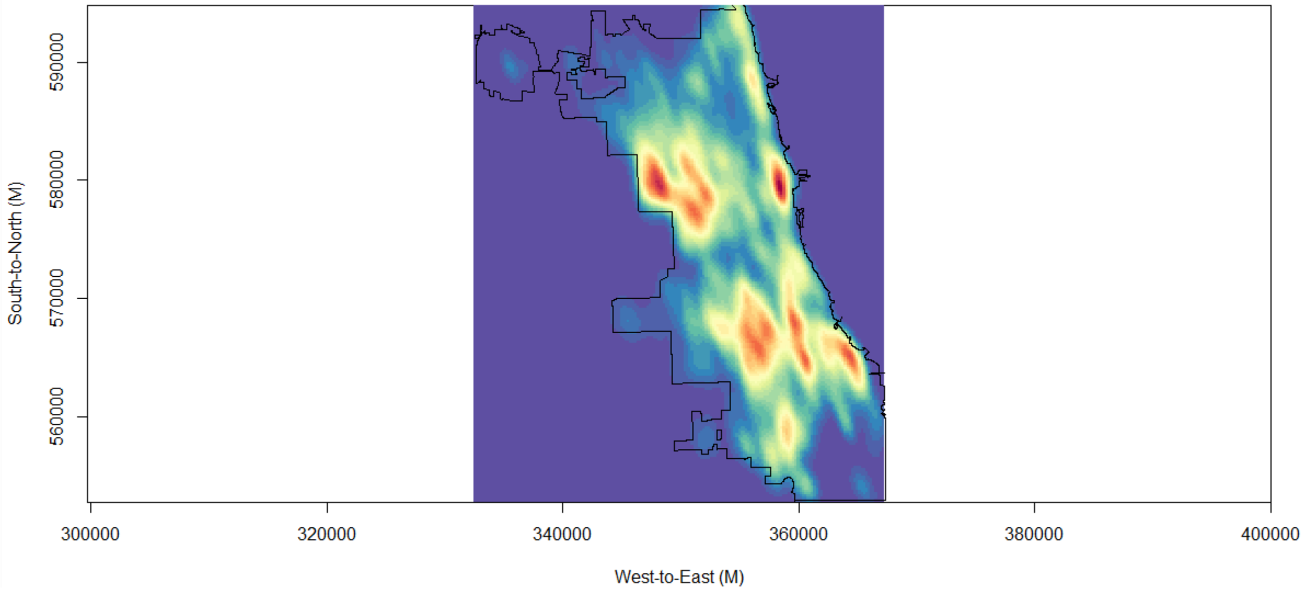
Evidences Produced:-

The plot clearly shows that assaults are not that the spread of reported assaults was not uniform as shown in the picture below:



As predicted the area like the O'Hare international airport reported lesser assaults cases as compared to the downtown area but the interesting thing is that there is a high concentration of assault reports in the Auburn Gresham (in the South Side of the city), South Austin and North Lawndale area of the city.

To find the concentration of thefts I then plotted a heat map using the 2-dimensional *Kernel Density Estimate*, taking 1000 random sample (due to processor limitations).

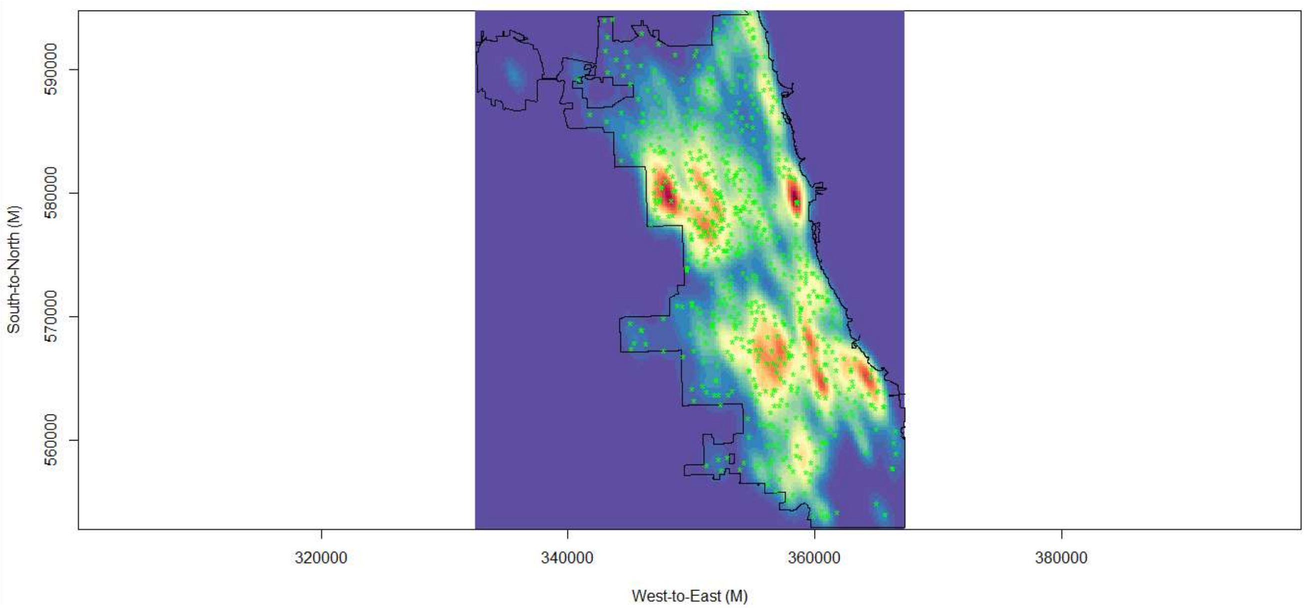


As compared to the Thefts concentrations we can observe:

- There is more number of hot spots locations in the assault crime cases as compared to the theft cases.

Recommendation:-

- Plotting the number of police stations(green stars) over the heat map, we can see that baring South Austin the other hot spot areas of the town, the number of police station are pretty less as compared to the assault concentrations. Therefore, more police stations or surveillance stations need to be set up.



- Along with establishing police stations, warning boards and tighter surveillance of the deserted locations should be looked upon.
- Even more lighting facilities should be provided in streets where lighting is not adequate.

Q2.) *Do theft concentrations look different depending on the time of day, day of week, and month of year?*

Prediction:

The hot spots locations in the assault concentration cases map will be more as compared to the theft concentration, also unlike theft concentration heat map there would be higher crime concentration spread during the late night hours as compared to the morning, afternoon or evening hours.

While comparing the day of the week, more assault concentration spread will be found in the weekends as compared to weekdays found in theft concentration spreads.

While comparing the theft concentration on months, the months of April, September and October will consist of higher spread but number of hot spots will increase, due to the same reason as mentioned in theft concentration plot.

Data Cleanup:-

While comparing the time of the day, I have divided the day in 4 different time zones:

1.) The morning time from 5am to 11am,
2.) The afternoon time from 11am to 5pm,
3.) The evening time from 5pm to 11pm and
4.) The late-night time from 11pm to 5am.

Therefore, four different samples of dataset are created on basis of the hour column in the original dataset.

Similarly, for comparing the theft concentration for different days of the week, I have created 7 different sample dataset by sub setting the original dataset on basis of the value in the day of the week column.
Where: 1 = Sunday, 2= Monday, 3= Tuesday, 4= Wednesday, 5=Thursday, 6=Friday, 7=Saturday

For comparison of theft concentration between different months of the year, 12 different datasets are created by sub-setting the dataset on basis of the values in the month's column of the dataset.

Data Analysis:-

Using the kernel density estimate four different heat maps are created to compare the concentration of thefts at different time of the day and then are clubbed together for better visualization.

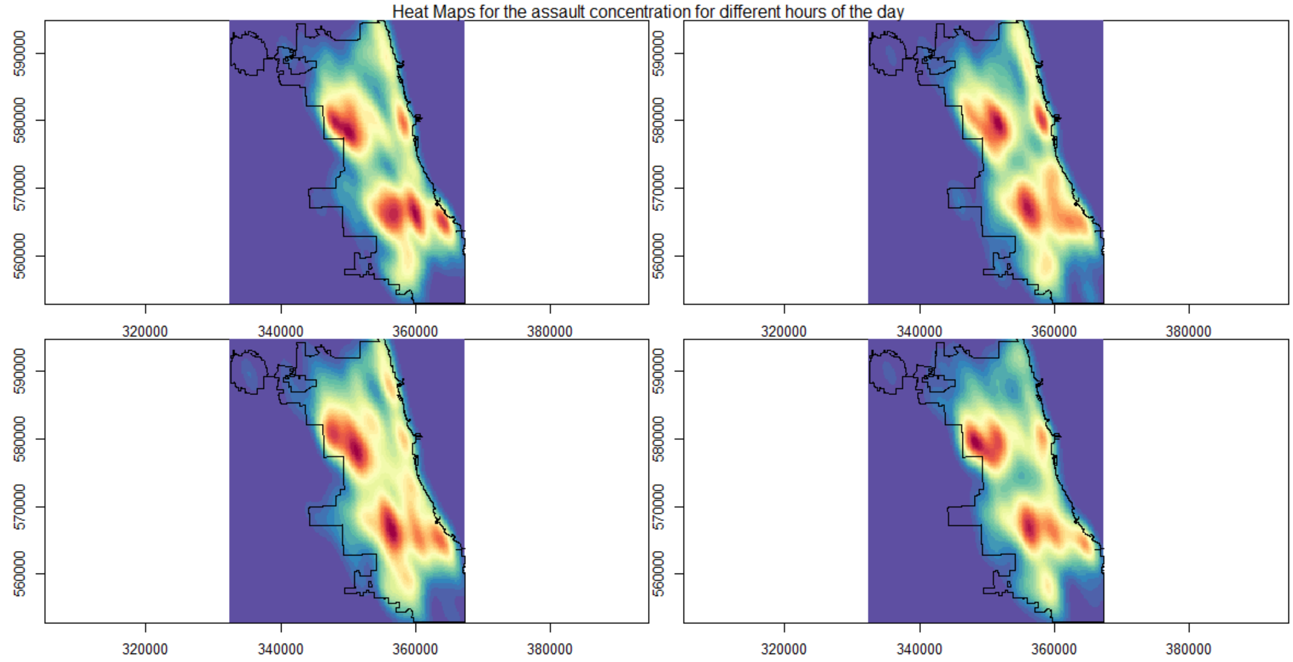
For different day of the week, heat maps are plotted and for visualization I have kept all the 7 plots in the screen.

Heat maps for the different months of year, for better visualization, heat maps are clubbed together with January to June in the first collection and July to December in the second collection.

Evidences Produced:-

Different time of the day

- It is observed that assault concentration has more than one hot-spots and that high concentration is spread in the south part and the South Austin area of the city.
- It is seen that the spread in all four times are almost evenly distributed but the assault concentration spread is a little more in the evening time as compared to other time of the day.


Heat Maps for the assault concentration for different hours of the day
590000
580000
570000
560000
320000
340000
360000
380000

The plots are in the order: 1. Morning time, 2. Afternoon time, 3. Evening time, 4. Late Night time

As compared to the theft concentrations we can observe:

- There are more high concentration (red spots) areas in assault concentration as compared to theft concentration plot.
- The highest theft concentration was in downtown, but in case of assaults south side and south Austin has higher concentration as compared to downtown.

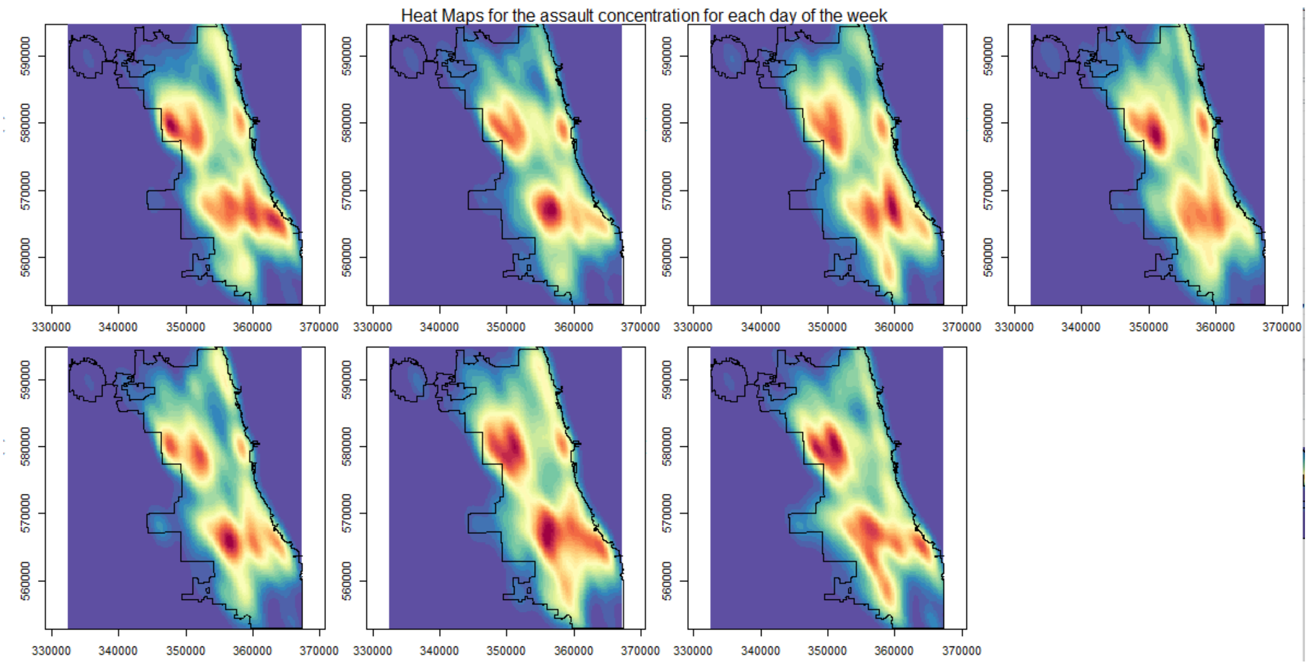
Recommendations:

The following recommendations are suggested for CPD.

- More patrolling resources should be allocated for south side of the town.
- Better surveillance systems should be deployed.
- More awareness should be spread to the people and training workshops should be set up where people should be trained on self protection.

Different days of the week

- It is observed that there is heavy concentration of assault cases in the south side and west side of the city.
- There is higher spread and more concentration in the weekends as compared to weekdays.


Heat Maps for the assault concentration for each day of the week
590000
580000
570000
560000
330000
340000
350000
360000
370000

The plots are in order: 1. Monday, 2. Tuesday, 3. Wednesday, 4. Thursday, 5. Friday, 6. Saturday, 7. Sunday

When compared to theft concentration

- As compared to thefts were concentration was in downtown, in assaults the higher concentration is in south and west side of the city.
- The spread and concentration is more in weekends in assaults spread as compared to thefts concentration where the spread was more in weekdays.

Recommendations:

1.) Higher patrolling resources should be allotted for weekends as compared to week days.

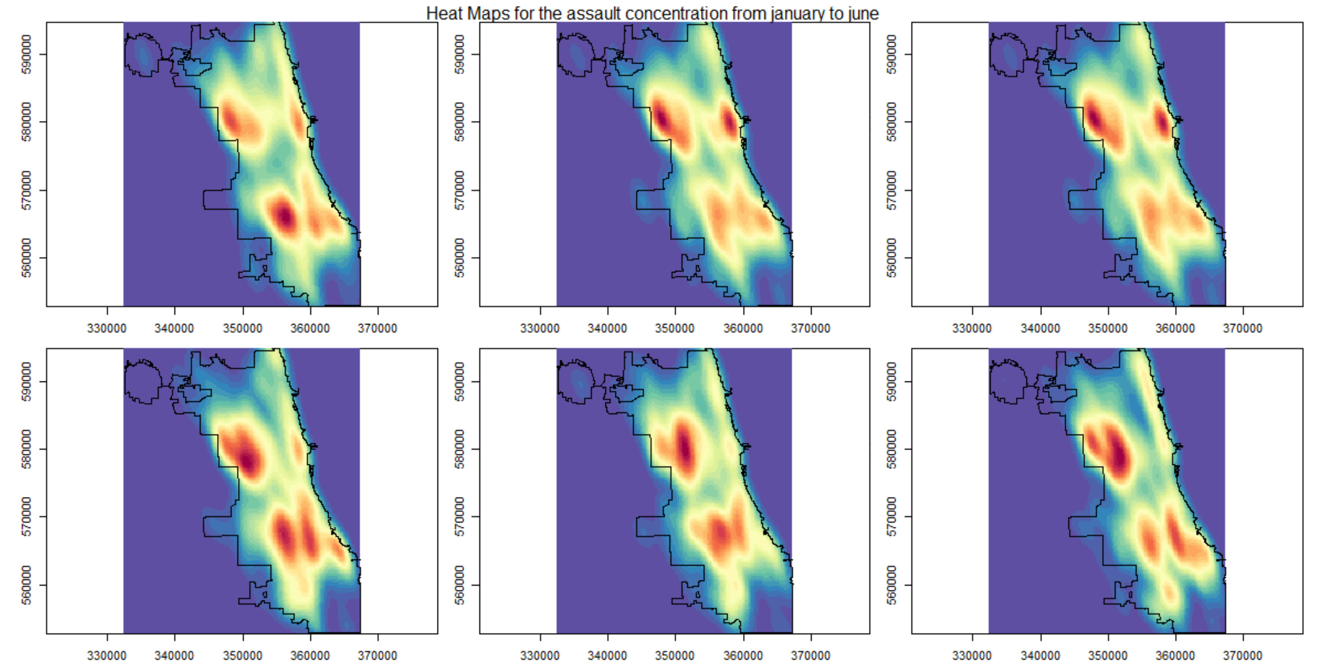
2.) As seen earlier the spread of police station is not dense in the south side of the city, therefore more number of permanent and temporary stations should be set up across the south side of the city.

3.) Better surveillance and rapid action systems should be deployed across the city.

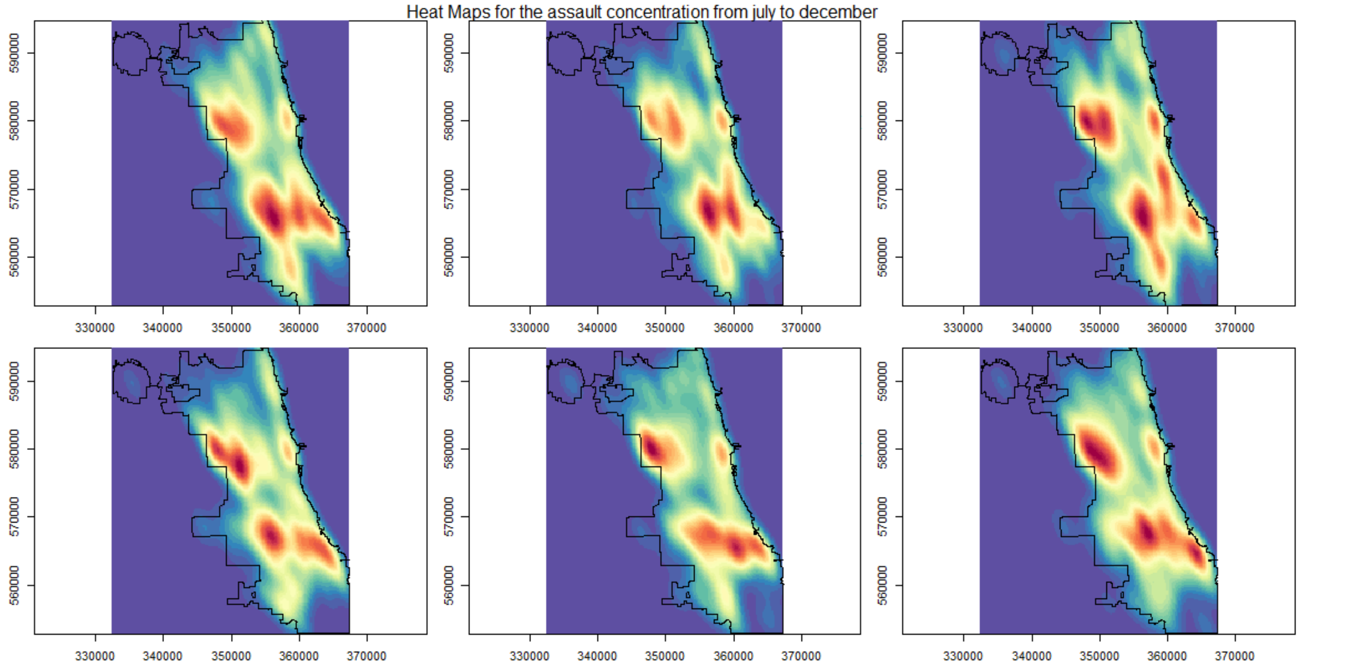
Different months of the year

Following observations can be made:

- The assault concentration is spread in the downtown area for the start of the year but as the year proceeds the concentration diminishes.
- Similarly, lesser assaults concentration can be seen in south side of the city, but the year progresses, the concentration gets darker.
- The assault cases are spread across the city with hotspots in some places.


Heat Maps for the assault concentration from january to june
590000
580000
570000
560000
330000
340000
350000
360000
370000

The plots are in order: 1. January, 2. Febuary, 3. March, 4. April, 5. May, 6. June


Heat Maps for the assault concentration from july to december
590000
580000
570000
560000
330000
340000
350000
360000
370000

The plots are in order: 1. July, 2. August, 3. September, 4. October, 5. November, 6. December

When comparing with theft concentrations we can say:-

- In heat map of crime by thefts, the major concentration was present in the downtown area of the city, but in the assault crime heat map, the spread of crime in more uniformly spread across the city with hot spots in south and west end of the city.

- The thefts concentration peaked in the months of April, September and October but assault concentration remains spread throughout the year, just the location of hot spots keeps varying a little with initial and later part of the year.

Recommendations:-

We suggest the following recommendations.

1.) As seen in the heat maps that, assault concentration decreases in the later part of the year in downtown, therefore, CPD should reallocate their resources to more assault prone areas of the city.
2.) As the year progresses the assault concentrations get higher in the south part of the city therefore better surveillance system should be deployed as early as possible.
3.) Most importantly the spread of police stations in less in the south part of the city, therefore, more stress should be laid on increasing the strength of the police in the south side of the city.

Q4.) What is the practical significance of the KDE resolution (e.g., 200m versus 1000m)?

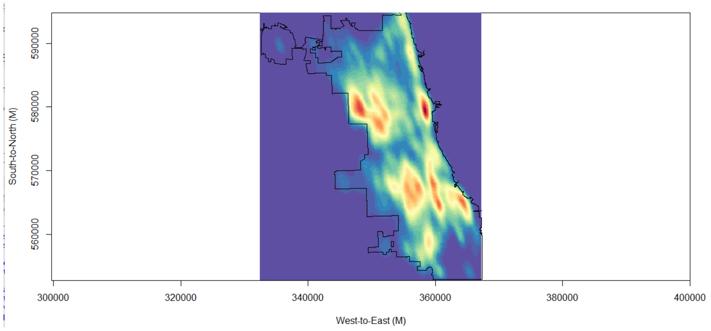
Solution:

KDE resolution divides the whole plot in grid and the value of KDE resolution is interpreted as the spacing between the points on which the estimate (heat map) is plotted. Therefore, higher the resolution, higher is the spacing between the points, so fewer number of points on which the estimate made.

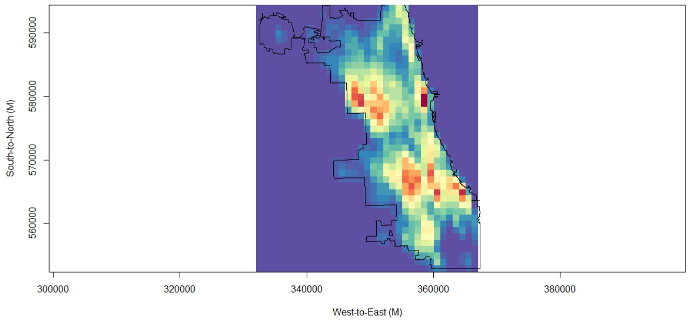
For example:

For KDE resolution= 200, the spacing between the points is 200 meter in x-axis and y-axis therefore sharper and more accurate heat map can be plotted as compared, when KDE resolution =100 which means the points are at 1000 meter difference in both x and y axis.

I have plotted the assault concentration on heat map for KDE resolution =200m and 1000m for better judge the difference.



Plot with KDE resolution=200



Plot with KDE resolution=1000

As the value of KDE resolution decreases, the plot is more accurate (higher pixel) but for plots with higher number of observations will require higher computational power, therefore, will take more time involved in calculation.

So, we have to create a balance in selecting the KDE resolution and the number of observations, as per the computational power available.

-----XXXX-----