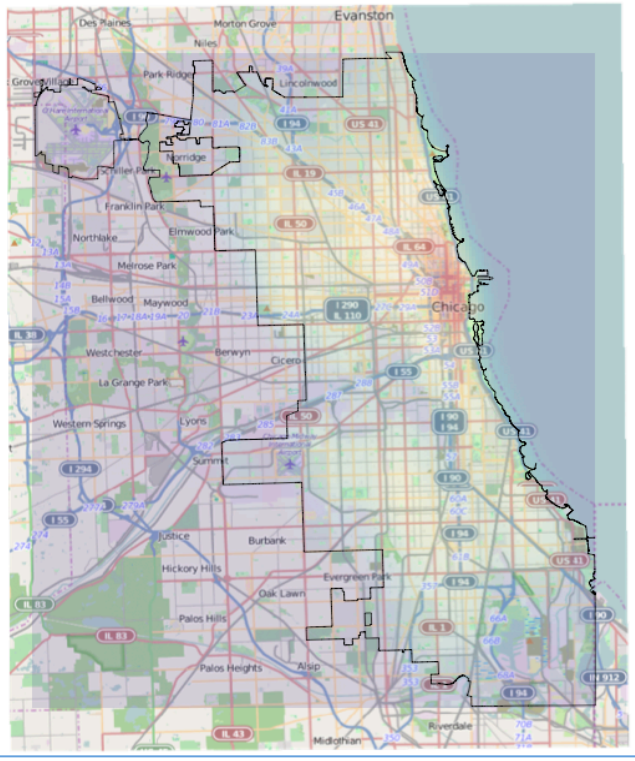
Case Study 1 (CS1): Visualizing Crime

"On my honor, I pledge that I have neither given nor received help on this assignment."
Name: Sharath Chand PV
ID: vp4pa

Introduction:

The study is about analyzing the occurrence of crime incidents (Theft and Assault) in the city of Chicago. We tried to understand areas of high crime concentration for different types of crimes. We also tried to understand whether the areas of high crime concentration varies based on factors such as time of day, day of week or month / season

Following is the open street map view of Chicago overlaid with a kde estimate plot of theft cases (transparent background) to get a better idea on locations. It will be used as reference while discussing about variations in crime concentration based on type and time of incidents



Analysis of Theft & Assault cases:

Our hypothesis is that distribution of crime incidents will not be uniform and depends on type of the crime and time of its occurrence.

- In case of regular thefts, the heavily populated areas such as malls, downtown areas, movie theaters, bus/metro stations should have high concentration.
- In case of Assaults, isolated and sparsely populated areas could be the target areas.

Datasets:

2014_THEFT.csv – All the theft incidents occurred in 2014 (Date, time, location coordinates, location description, incident description, address, arrest made etc. are the important features)

2014_ASSAULT.csv – All the assault incidents occurred in 2014 (Date, time, location coordinates, location description, incident description, address, arrest made etc. are the important features)

Analysis Methodology:

The approach is to understand the distribution of crime, we applied kde estimate for estimating the distribution of the crime in a specified boundary.

For each dataset, following data **preprocessing activities** are done:

- Step1: Coordinate Ref system of city boundary is in meters. So the location coordinates of our data to be converted to the same CRS (epsg:26971)
- Step2: parse the date variable (ex: 12/04/2014 11:30:00 PM) to extract Hour of day, day of week and month
- Step3: Addition of Extracted features to the dataset for further analysis

Sub-setting and KDE Estimate:

Each dataset (Theft and Assault) is filtered on the following conditions to understand the crime profile in different scenarios (times, days and months)

- Based on hour of day – four subsets are created to understand crime distribution on a given day
  - 1.00 to 5.59 Early morning hours
  - 6.00 to 11.59 Before Noon hours
  - 12.00 to 17.59 After Noon hours
  - 18.00 to 00.59 Night hours
- Day of week – 7 different subsets created one for each day of week (Sunday to Saturday)
- Monthly distribution – 12 subsets created one for each month
- For every subset created, we plot the kde estimate for 1000 sample size with a resolution of 200 to better understand the distribution

Functions Created:

- concentration_analysis(dataset, ID, samplesize): The function takes the dataset(Theft or Assault), Identifier (T – for theft, A – Assault) and sample

size (no. of records used for KDE estimation, 1000 in our case) and generates kde plots for various subsets described above

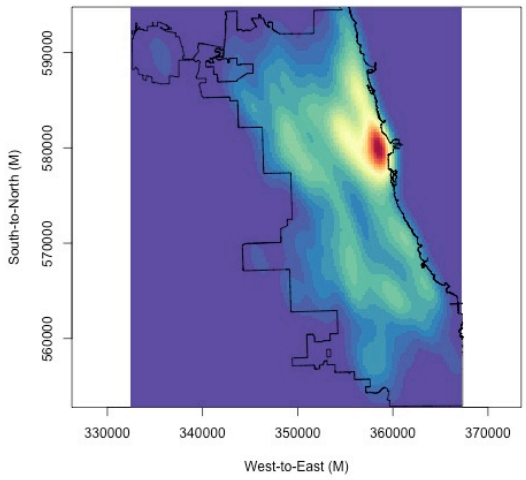
- mykde <- function(data,n): The function takes the subset data and sample size as input and generated kde plot. This function is called for multiple times from the concentration_analysis() function.

Distribution of Theft cases:

Following table gives the frequency of the most common type of theft cases. Small theft seems to be more frequent type among all.

$500 AND UNDER	OVER $500	RETAIL THEFT
26669	13804	7352
FROM BUILDING	POCKET-PICKING	PURSE-SNATCHING
5726	1620	660

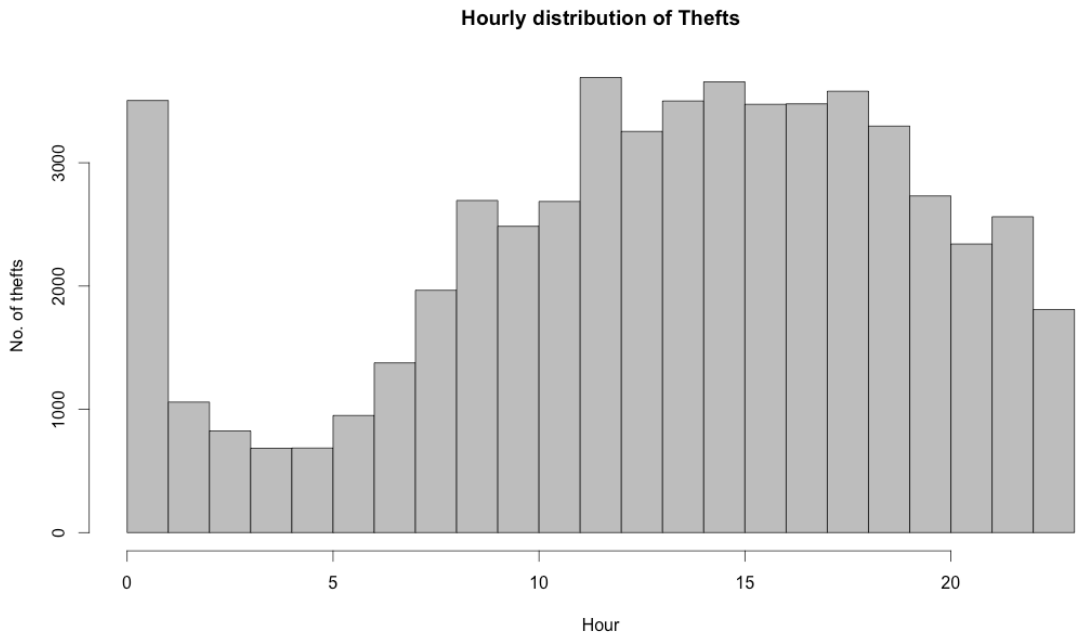
To understand how the theft cases distributed across Chicago, we have plotted the kernel density estimate of theft cases on its boundary map.



Color scale used in KDE plots (Red – Highest, Violet – Lowest)

In the above plot the crime distribution seems to be concentrated more at a particular location (yellow cloud with red core). As hypothesized, that happened to be the downtown area of Chicago.

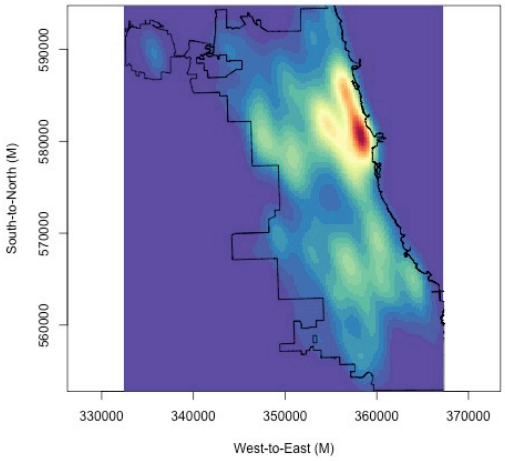
Does it depend on the hour of day?
Following histogram gives us an idea on number of theft cases in a given hour of day. The frequency looks more from 11.00 to 19.00 hrs and (00.00 to 1.00?)



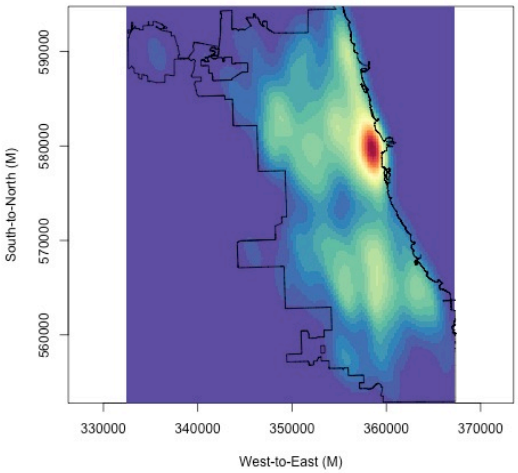
Let us look at how the distribution in different times of a day

For easy interpretation, we have divided the day into 4 parts and plotted KDE for each part of the day.

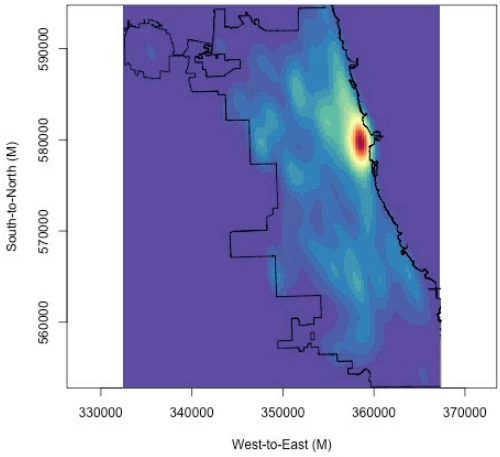
1.00 to 5.59	Early morning hours
6.00 to 11.59	Before Noon hours
12.00 to 17.59	After Noon hours
18.00 to 00.59	Night hours



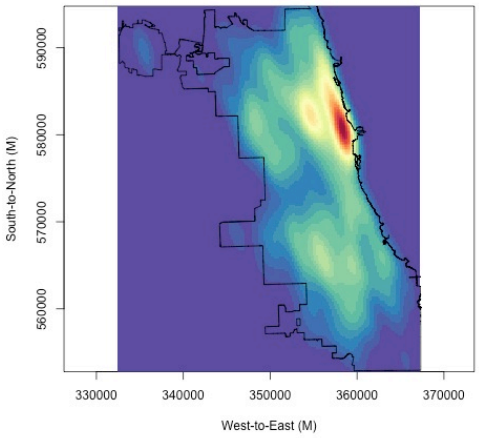
Early Morning



Before Noon



After Noon



Night Hours

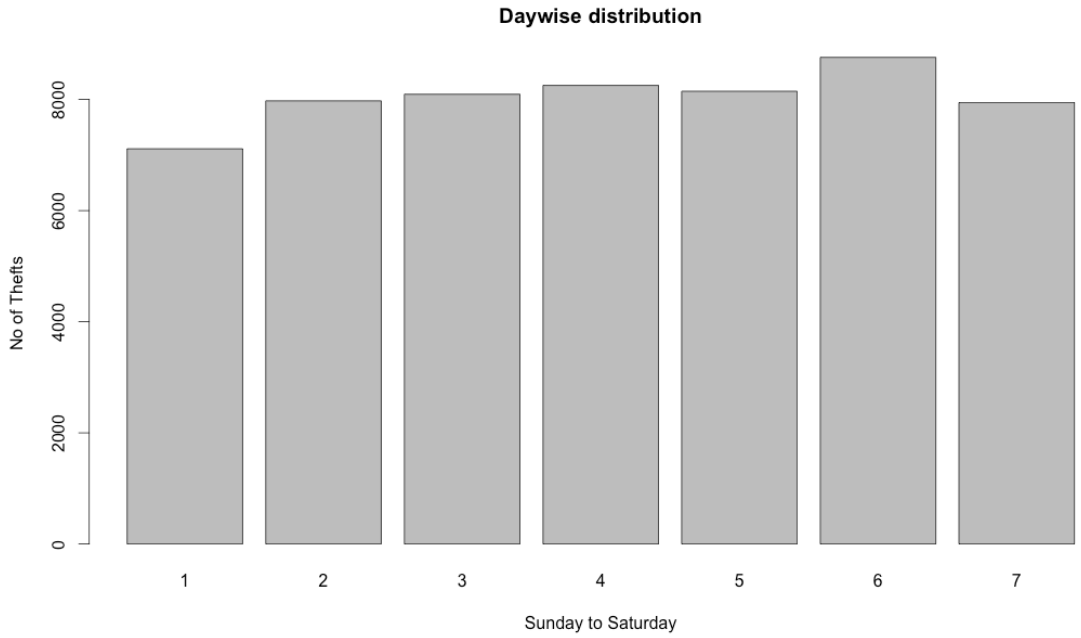
Interpretation & Recommendations:

- We can see that after noon time (1PM to 6 PM) the theft is highly concentrated at downtown area. That is the time when the downtown is densely populated.
- During night and early hours, it is spread out a bit.
- Instead of placing the Theft specialist teams everywhere, the department can use these finding and position those teams in the hotspot areas of the plot
- Warning sign boards, regular announcements in the densely populated areas about thefts, pick pocketing would keep people cautious about their belongings

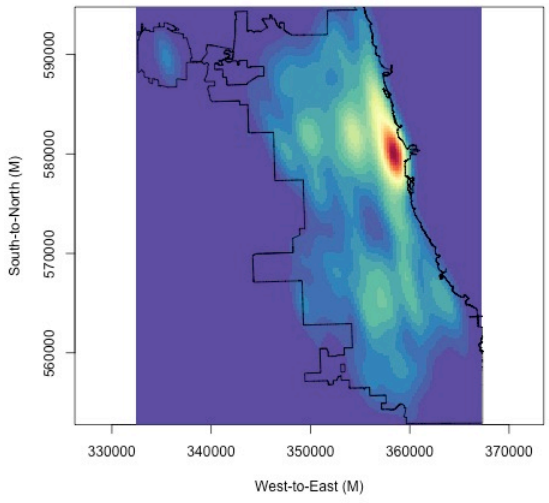
- Police compliant desks, installation of cc cameras in downtown areas (The hotspot in our plot) would further help decreasing the Theft

Does it depend on day of week? (Ex: the distribution on Sunday/Saturday is different from the distribution on other weekdays)

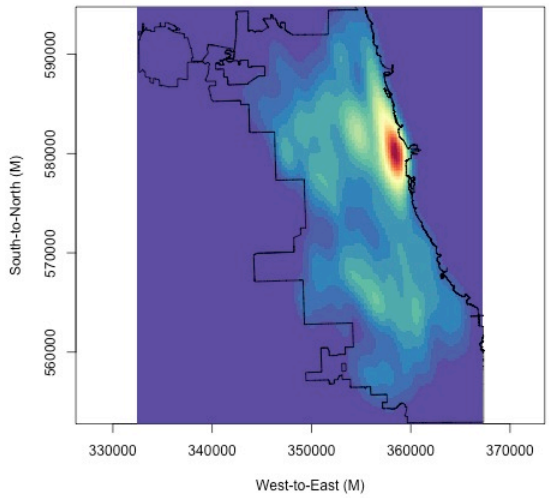
No. of thefts as per the day of week. Incidents are higher on weekdays



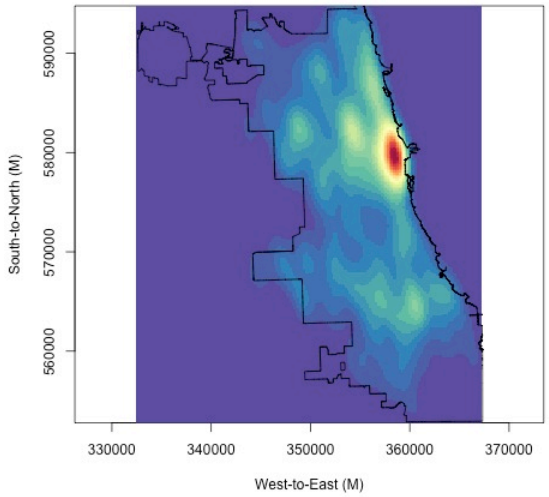
Distribution of thefts on different days of week



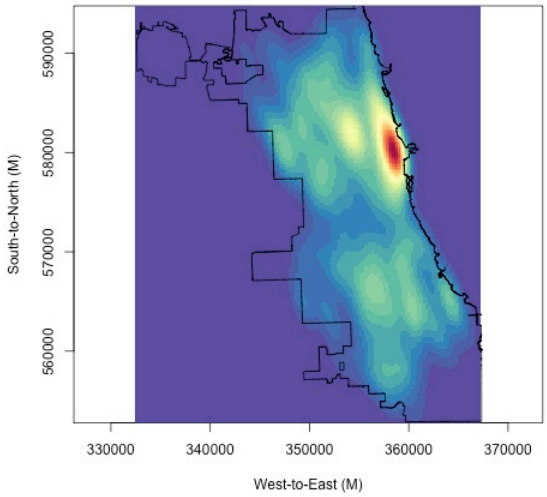
Sunday



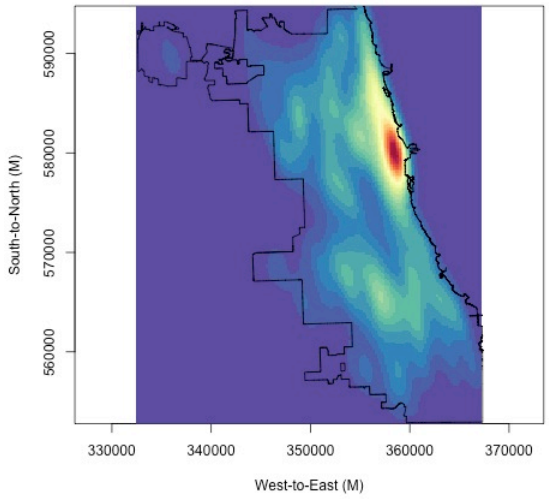
Monday



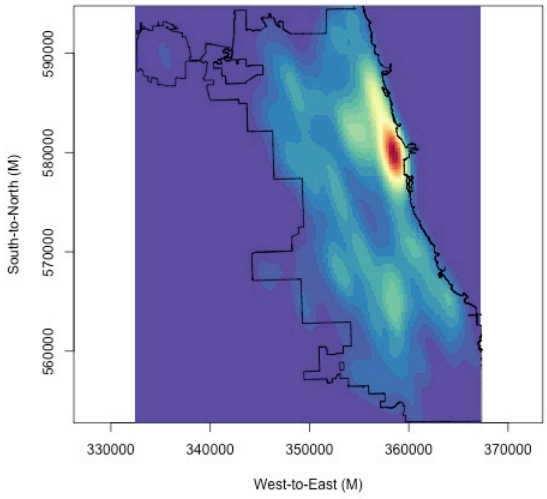
Tuesday



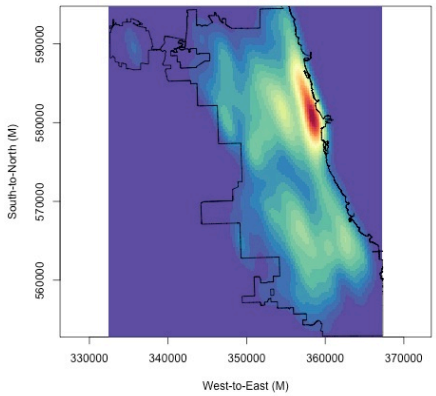
Wednesday



Thursday



Friday



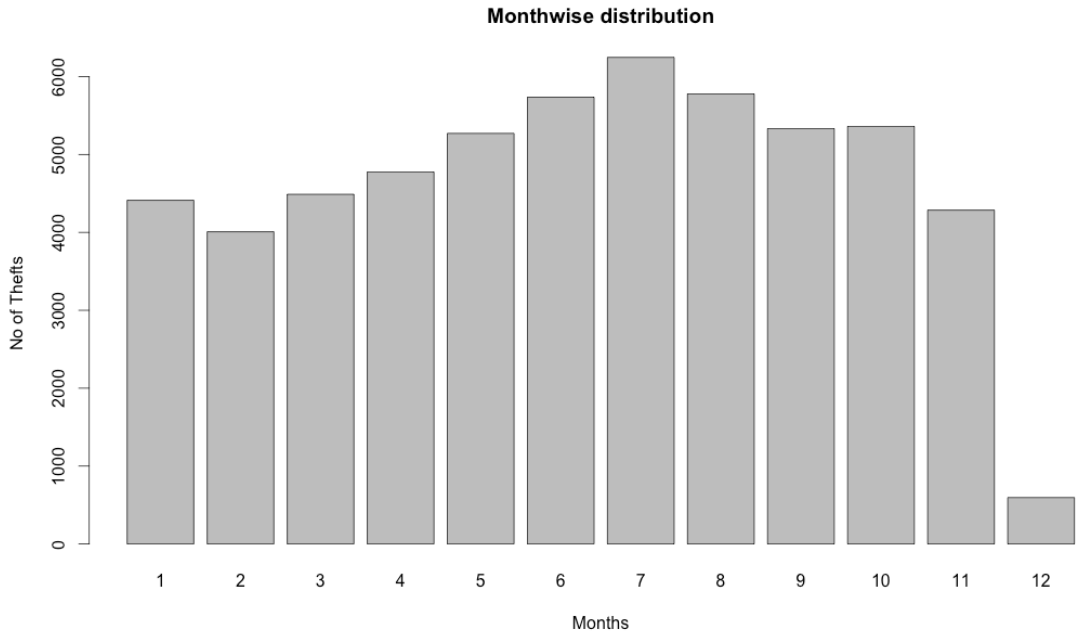
Saturday

Interpretation:

- Except Saturdays, all other days of week has similar distribution. The spread is slightly wider on Saturday and it is more concentrated in one location on other days of week
- The department can continue positioning their theft squads at the hotspots in all days of week

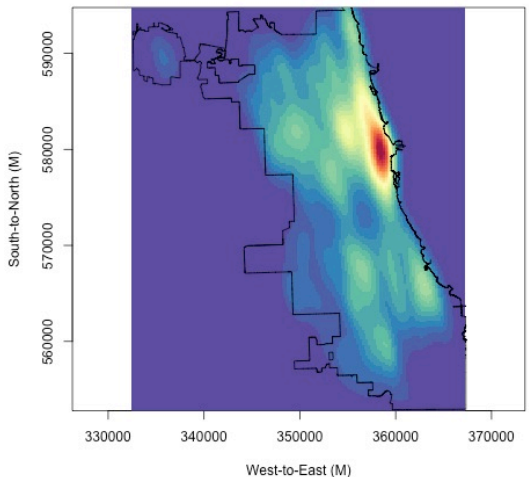
Does the Distribution of theft cases depend on Month?

Frequency is more during the summer months. The theft cases in December are only 13% of the average monthly count. It is because data for only first 4 days of December is available

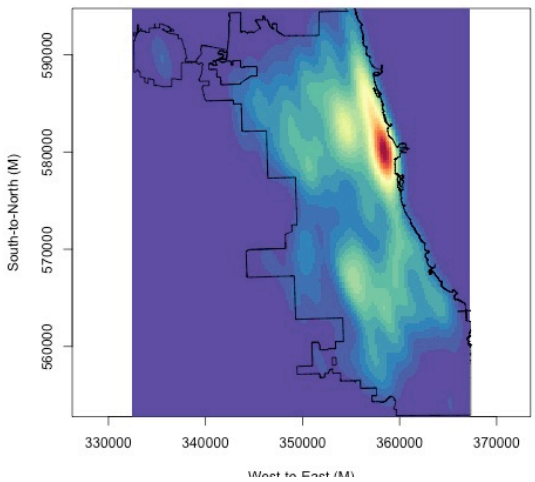


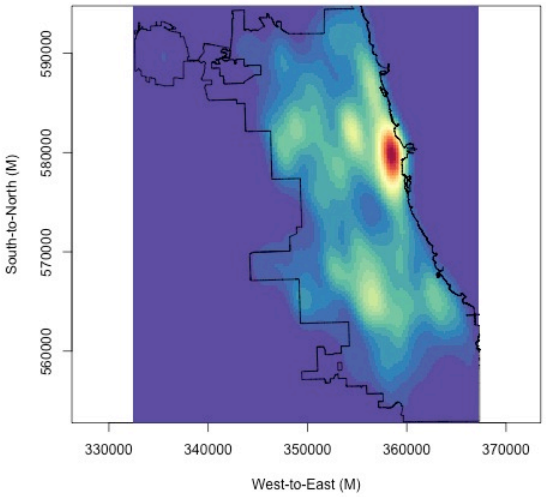
Distribution of thefts in different months

January

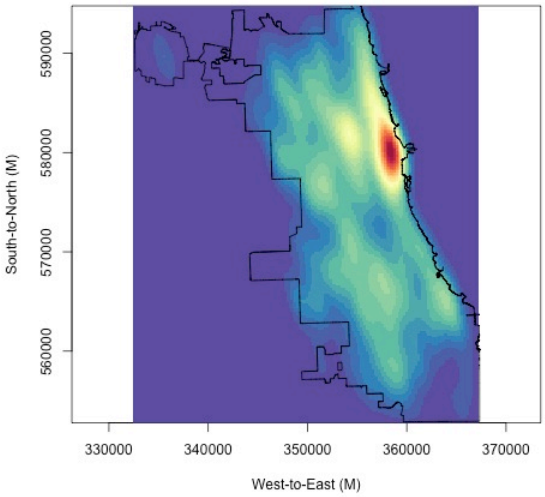


February

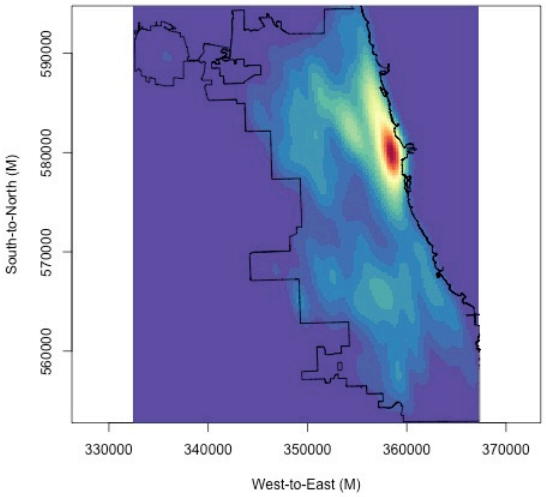




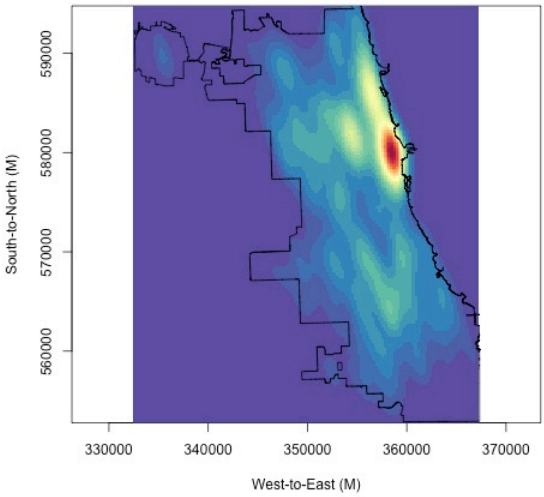
March



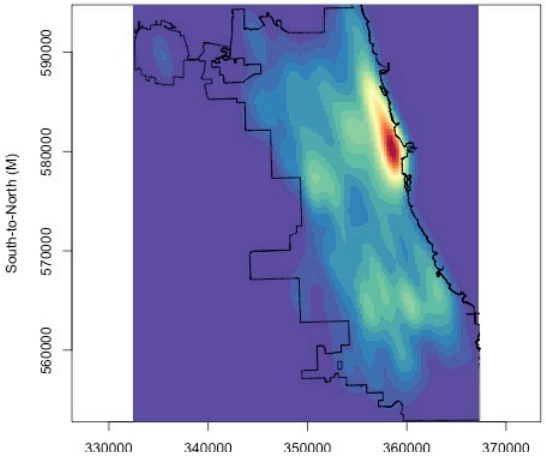
April

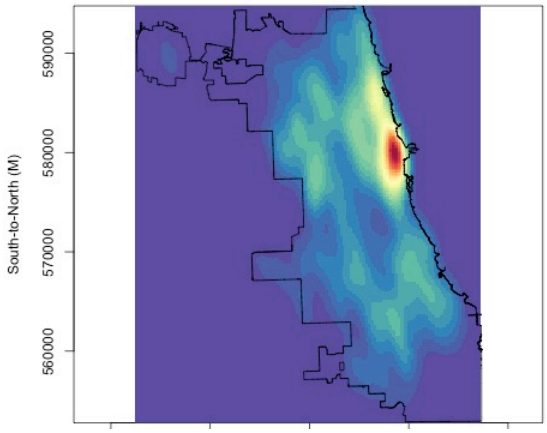


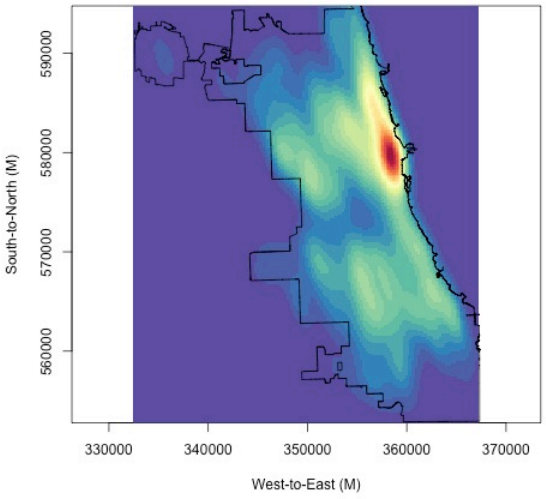
May



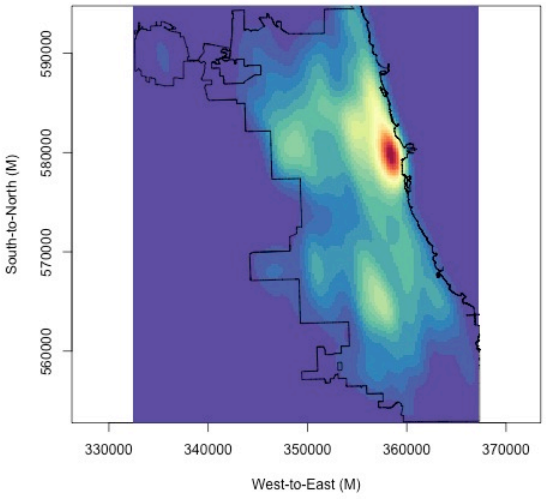
June



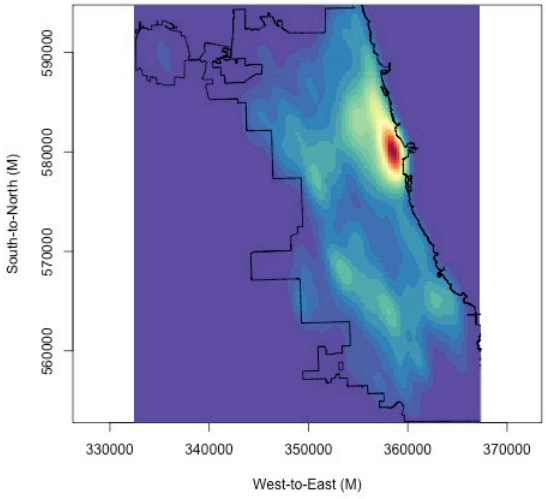




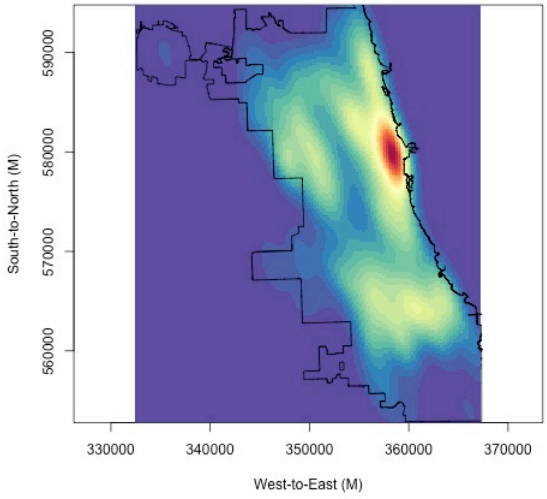
September



October



November



December

Interpretation & Recommendations:

- In Summer months thefts are concentrated mainly at and around downtown area
- In non-summer months especially in April and December months, it is spread out to other areas though more concentration is still at downtown/ city center area. December month cannot be considered because the data for only 4 days is available

- We can observe a good amount of differences among months with respect to distribution of thefts, even though the concentration seems to be at one place
- Since the hotspot is at one location, the department can focus on that area for tackling the theft cases

Distribution of Assault cases:

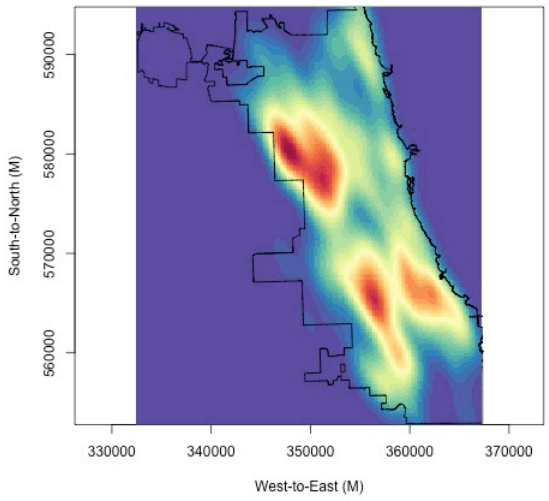
Similar analysis for Assault has been carried out as in Theft cases. We will try to compare the distribution spectrum of Assault with the corresponding spectrum of Theft.

Following table gives the frequency of the most common type of Assault cases. 2/3rd of the assaults cases in the data seems to be simple assault type.

SIMPLE	AGGR: HANDGUN	AGGR: KNIFE/CUTTING INSTR
10528	1668	1143

AGGR: DANG WEAPON	EMP HANDS NO/MIN INJURY	AGG PO HANDS NO/MIN INJURY
858	667	449

Let us plot KDE plot to understand the Assault spread



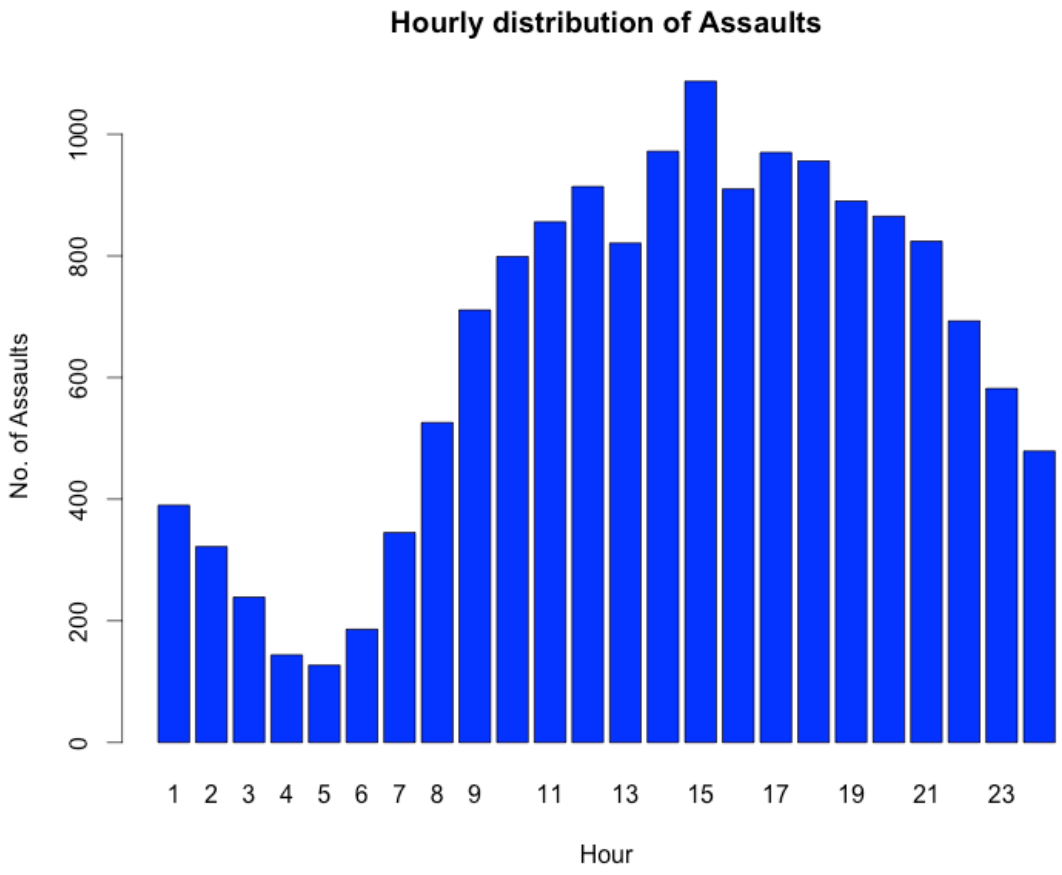
Interpretation & Recommendations:

- The spread of Assault is completely different from that of the Theft. It seems to be concentrated more in sparsely populated areas where carrying of assaults is easy and risk-free.
- The concentration is more towards west and south of Chicago (yellow clouds with red core).
- There are three or more hotspots when compared to theft in which there is only one major hotspot

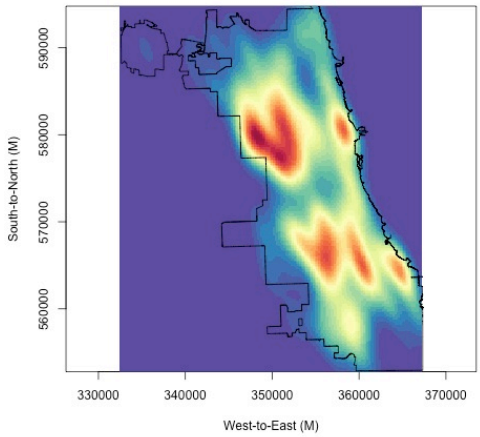
- The department can position assault specific resources in the high concentrated areas.
- The further analysis in terms of lighting, graffiti, isolated / masked streets, unused pathways can be carried out and warning boards, streets/walls cleanup activities, sufficient lighting arrangements, cc cameras can be planned to minimize the incidents

Distribution of Assaults – Hour of Day

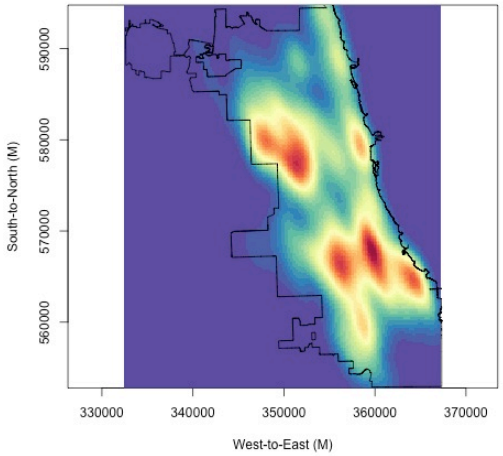
Hourly frequency shows that Assaults happen more during afternoon and evening times with the highest rate at 3PM



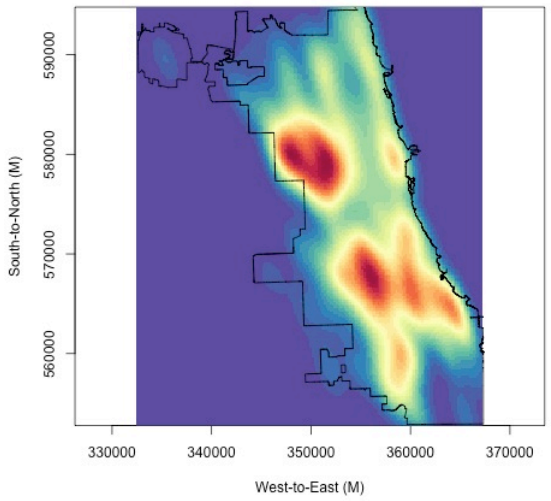
Let us look at distribution in different times of day



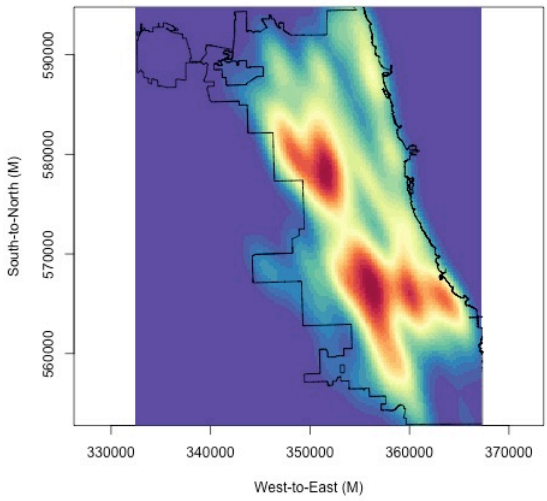
Early Hours (1AM to 6 AM)



Before Noon (6 to 12.00 PM)



Afternoon (12 PM – 6 PM)



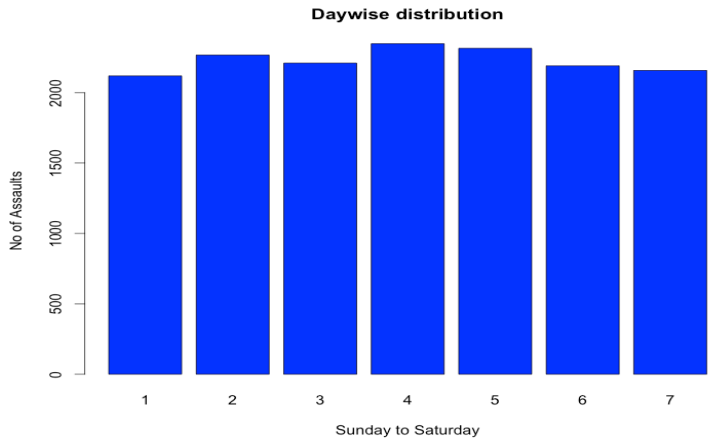
Night hours (6PM – 12 AM)

Interpretation & Recommendations:

- Assault spread is more in south and west parts of the city. Evening and night hours the downtown area where the theft is highly concentrated has very less Assault incidents
- The patrolling activities can be increased during afternoon and night hours (12PM to 12 AM) in the southern parts since the concentration is higher when compared to early and before noon hours

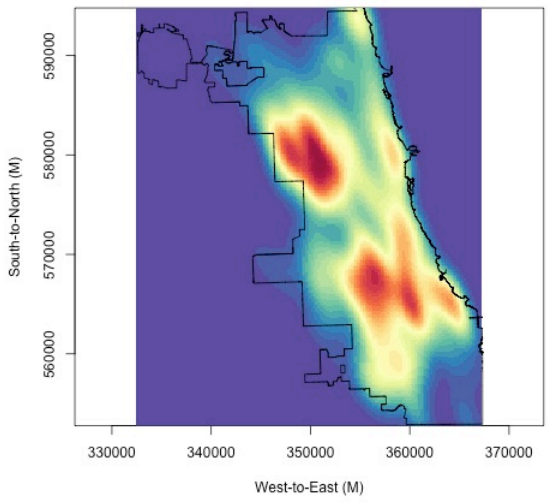
Distribution of Assaults – Day of Week

The frequency looks more or less same across days of week, slightly less on weekends

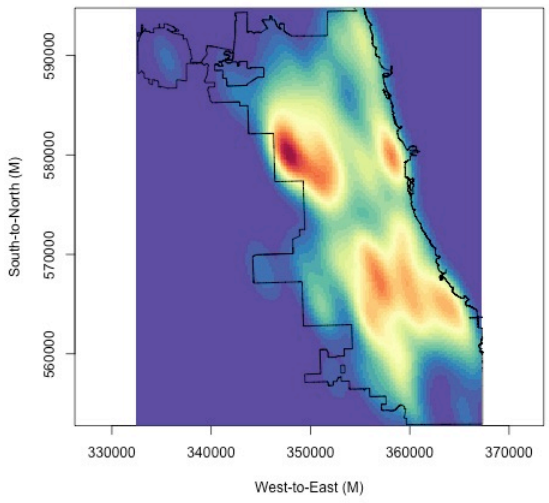


Let us look at how it is spread on various days of week:

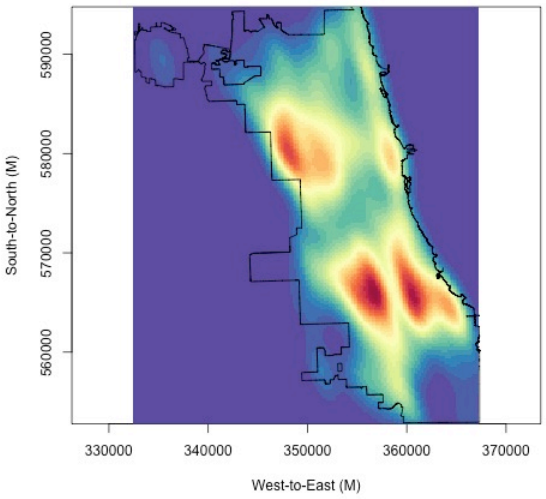
Sunday



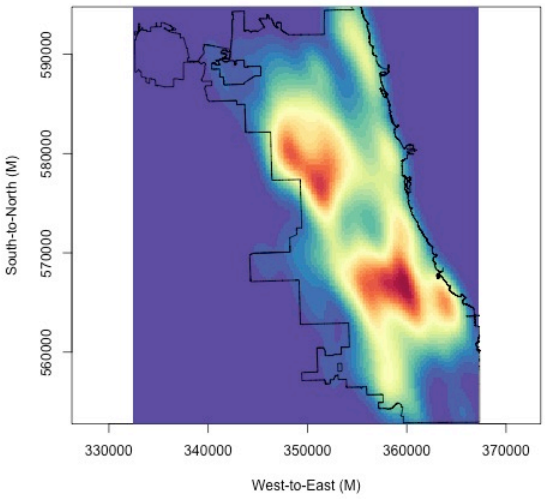
Monday

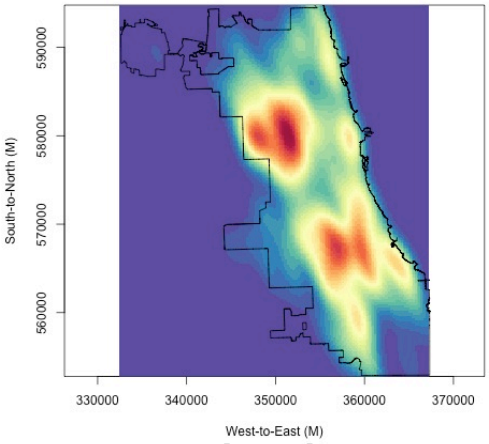


Tuesday

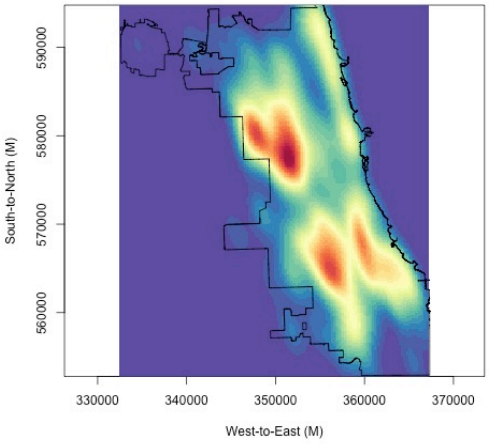


Wednesday



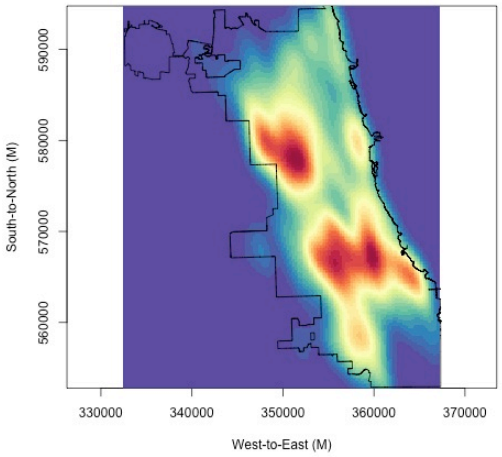


Thursday



Friday

Saturday

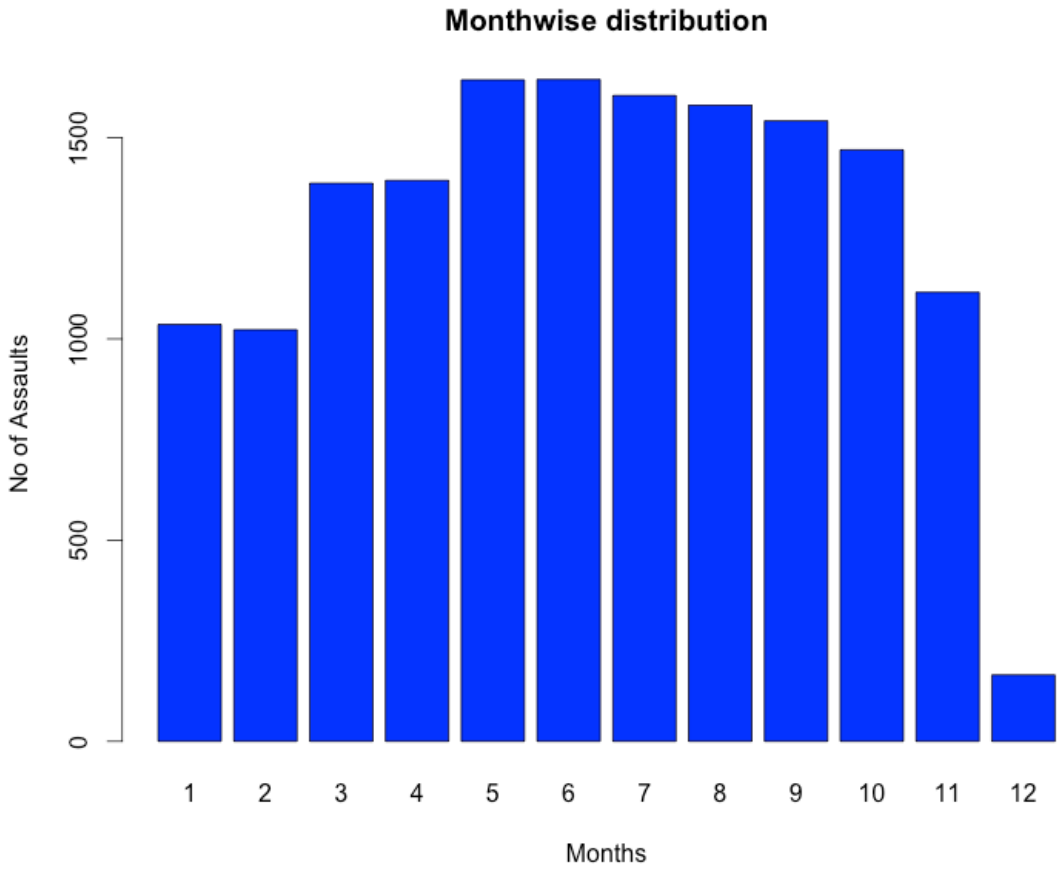


Interpretation & Recommendations:

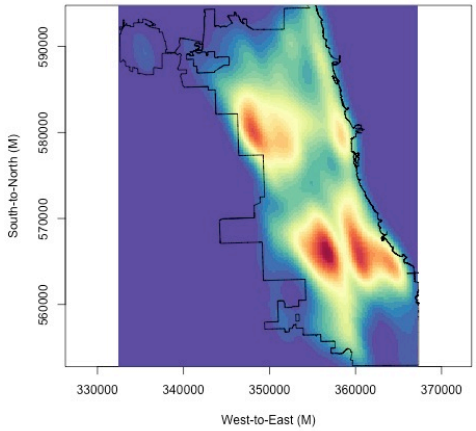
- Unlike Theft, Assault cases show different patterns on different days of week. During weekends assaults concentration is more in south and west. But during weekdays it heavily concentrated in west than south.
- It can be due to industrial corridor that south has, which would be operative during weekdays but isolated during weekends
- The department can focus their patrolling activities in western and southern parts where the concentration is higher.
- Warning sign boards on the streets of these hotspots would make costumers more cautious

Distribution of Assaults in Different Months:

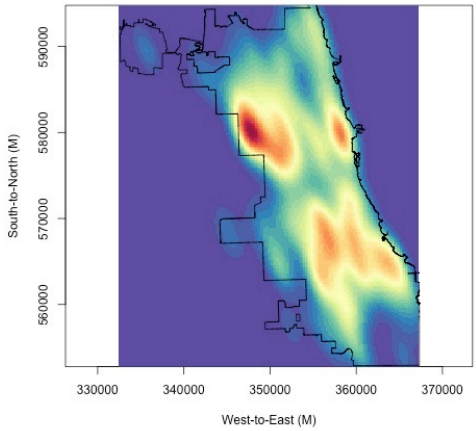
Frequency of Assaults is more in summer when compared to other months. December data is not complete, that is why we can see a sharp drop in number of Assaults.



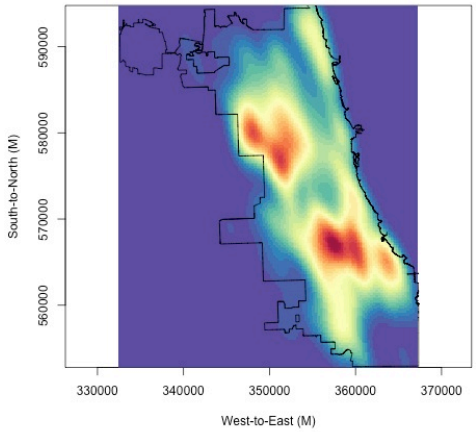
Spread of Assaults in Chicago for different months



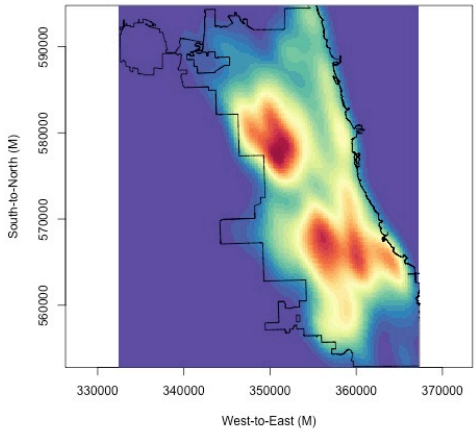
Jan



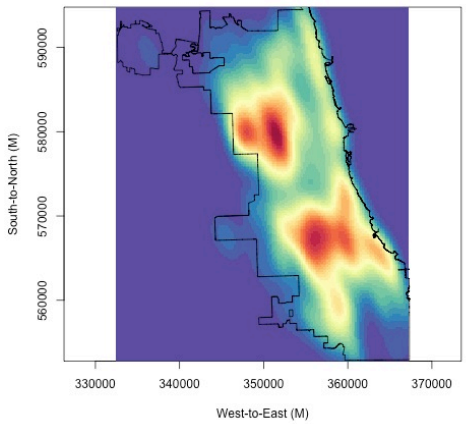
Feb



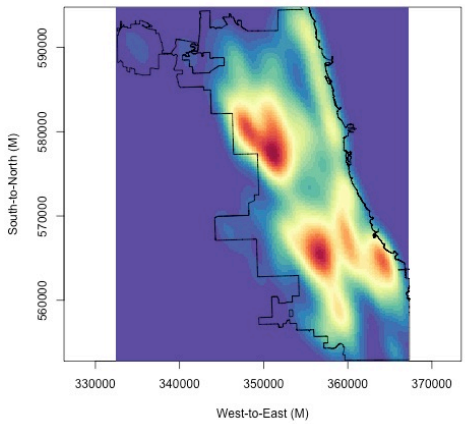
Mar



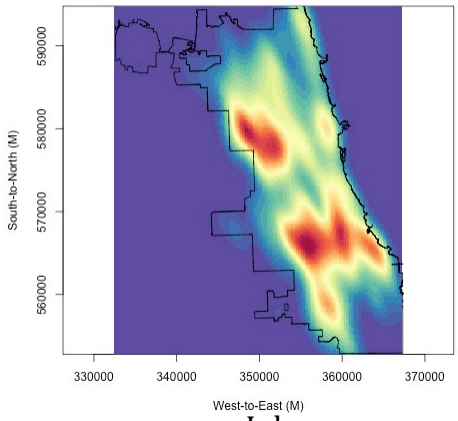
Apr



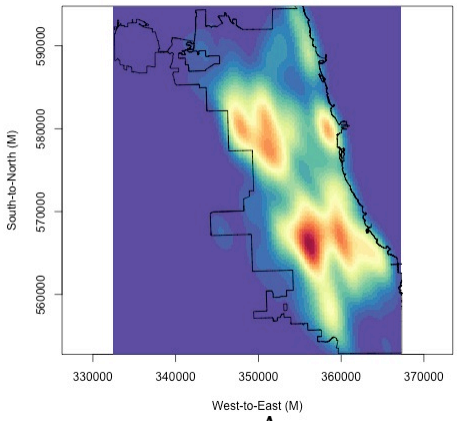
May



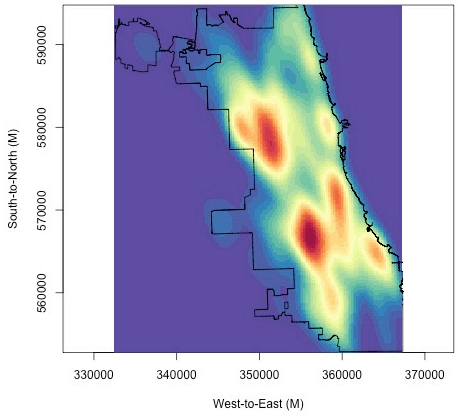
June



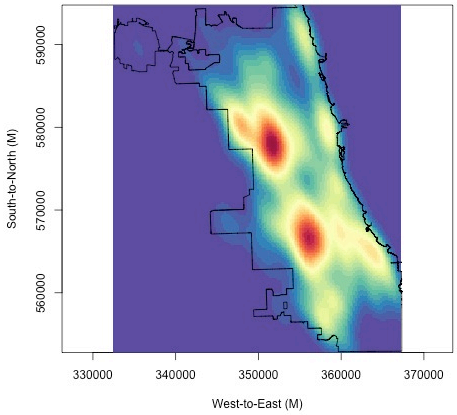
Jul



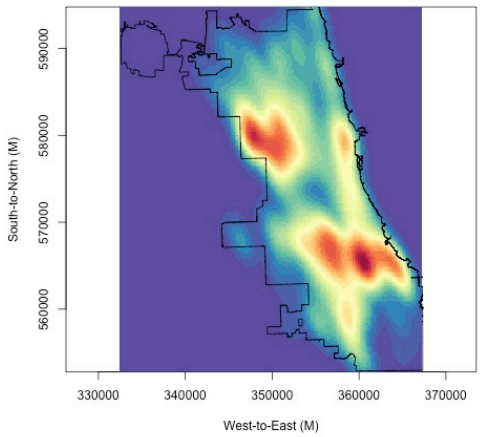
Aug



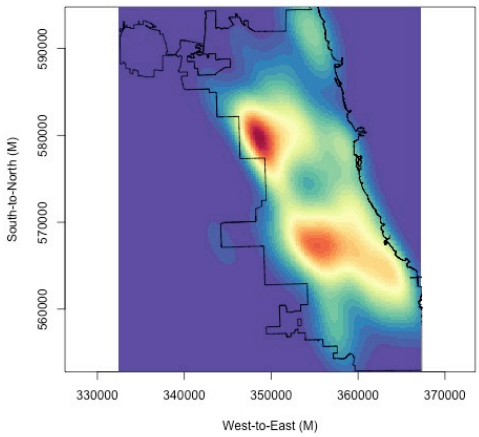
Sep



Oct



Nov



Dec

Interpretation & Recommendations:

- Monthly distribution of Assault is quite different to that of the Theft, more spread out while the theft is highly concentrated in one part of the city.
- In most of the months (Except Feb, Aug and Nov), Assault concentration is quite low in the parts where the theft is highest.
- Though west and south parts of the city are hotspots for Assault, the concentration levels vary in between those two locations in different months. While Feb and Dec have lower concentration levels in south than west, Jan and Aug have the more concentration in west than south. Other months have more or less equal levels of concentration in south and west but the spread in southern parts is more. That means more locations in southern parts are prone to assault incidents

What is the practical significance of KDE resolution?

KDE resolution can be practically interpreted as the spacing between the points at which the estimate is evaluated. The higher the resolution the larger the spacing between points and the fewer the number of points (x, y coordinates) at which kde estimates are calculated.

Ex: Resolution value 200 means one point for every 200 meters on x-axis (x-coordinate) and every 200 meters on y-axis is considered for kde estimate evaluation

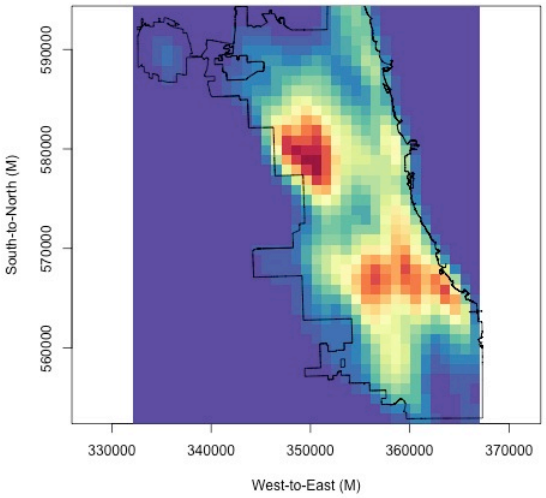
Following table enumerates the no. of points evaluated for different resolution values

Resolution	X-Axis start	X-Axis End	X-Range	Y- Axis start	Y -Axis End	Y-Range	No. of Points X-Ax	No. of Points Y-Ax	Total points estimated
10	332577	367345	34768	552875	594870	41995	3477	4199	14600648
200	332577	367345	34768	552875	594870	41995	174	210	36502
1000	332577	367345	34768	552875	594870	41995	35	42	1460

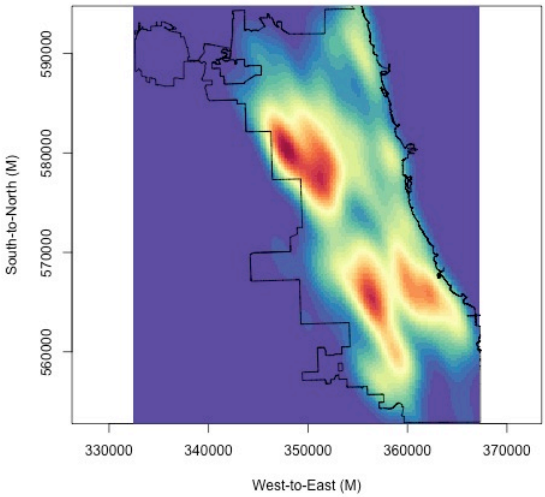
We can see that increase in resolution value drastically decreases the number of points to be estimated.

It can be interpreted as lower resolution values cause larger pixel value hence the higher clarity of the plot. But the time required for the estimating would increase.

Following are the plots with 200 and 1000 resolution values



Resolution 1000



Resolution 200