CRIME RATE ESTIMATION

PROJECT REPORT

TABLE OF CONTENTS

1. PROBLEM STATEMENT

Crimes have been severely increased in past few years, the Problem Statement includes analysis of crimes with different perspectives including utmost attributes possible and predicting via the study of nature of crimes committed. The problem statement is described to initially predict the crime-type based on location and time. I have worked on data about historical crimes in California.

I had close to 13,000 records of crimes with data on the date and time of the crime, its location, and its type. Common types of crime include theft, criminal damage, criminal trespass, and assault. This project took on the task of predicting the type of crime that was committed given a police report in two ways one according to time that is when crime took place and another is location that is where crime took place. From a small number of overly detailed features, in time it will give the detail that at which time slot which crime is maximum and in location it will tell at which place which type of crime is maximum. They then trained various diagram based models (Graphs and Pie charts) to classify crimes by type using the generated features.

Finally, they tested the performance of their models on testing data. They conclude that predicting the type of crimes committed by time and location alone is quite difficult, but that the feature engineering greatly increases predictive power. Predictions will be made to provide local authorities with an upper hand on crime and help them plan a better strategy to tackle the same.

2. PROPOSED IMPLEMENTATION FRAMEWORK

At a glance, here's what we did:

• Data Set Collection

• Implementation of our code

• Displaying results according to different objectives

• Copying results and pasting in excel sheet

• Creating corresponding graphs related to it

Data was collected from https://data.gov.org (2017 Dataset Crime). Data was pre-processed and cleaned to remove missing values and garbage values at various positions. Then, it was given as input to our code in mapper code. The output from mapper was sent to reducer code and the corresponding results were printed. The results were saved in a text file as the code was run in Hadoop. The part of code was then selected and pasted in excel file. Corresponding graphs were created to display the results more beautifully.

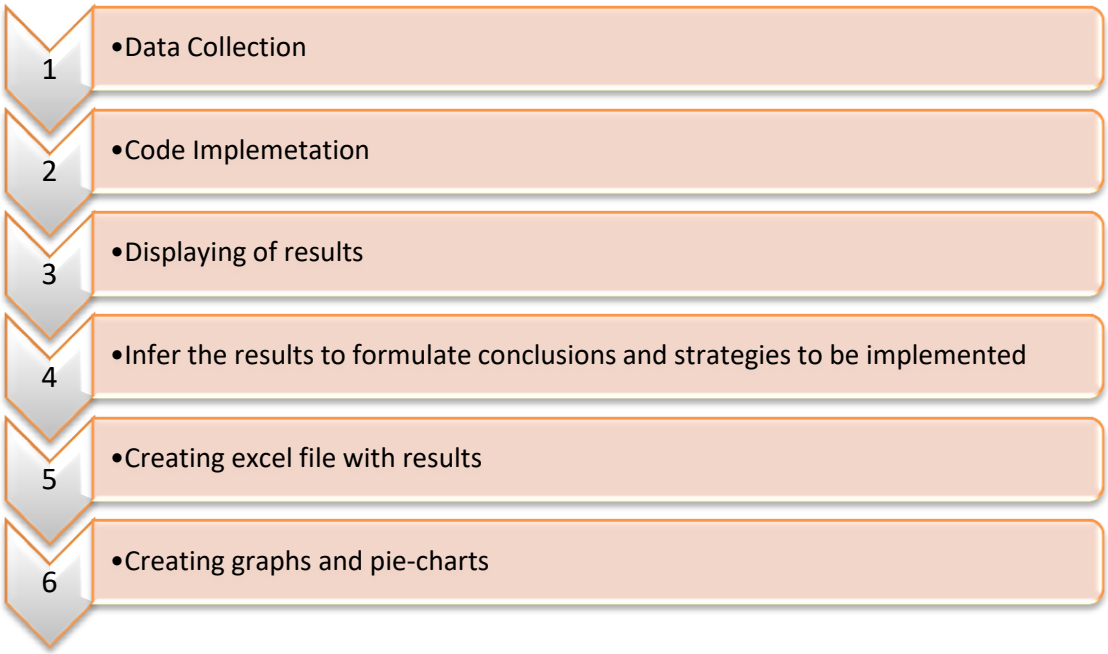
Output of the code:

- ***Based on Time of Crime*** – Analysis and predictions will be based on time of crimes i.e. which time has the maximum crime rates and needs to be inspected more efficiently.

- ***Based on Location of Crime*** – The most Crime prone locations will provide local authorities to target specific area clusters to counter crime.

Tools used for this project :

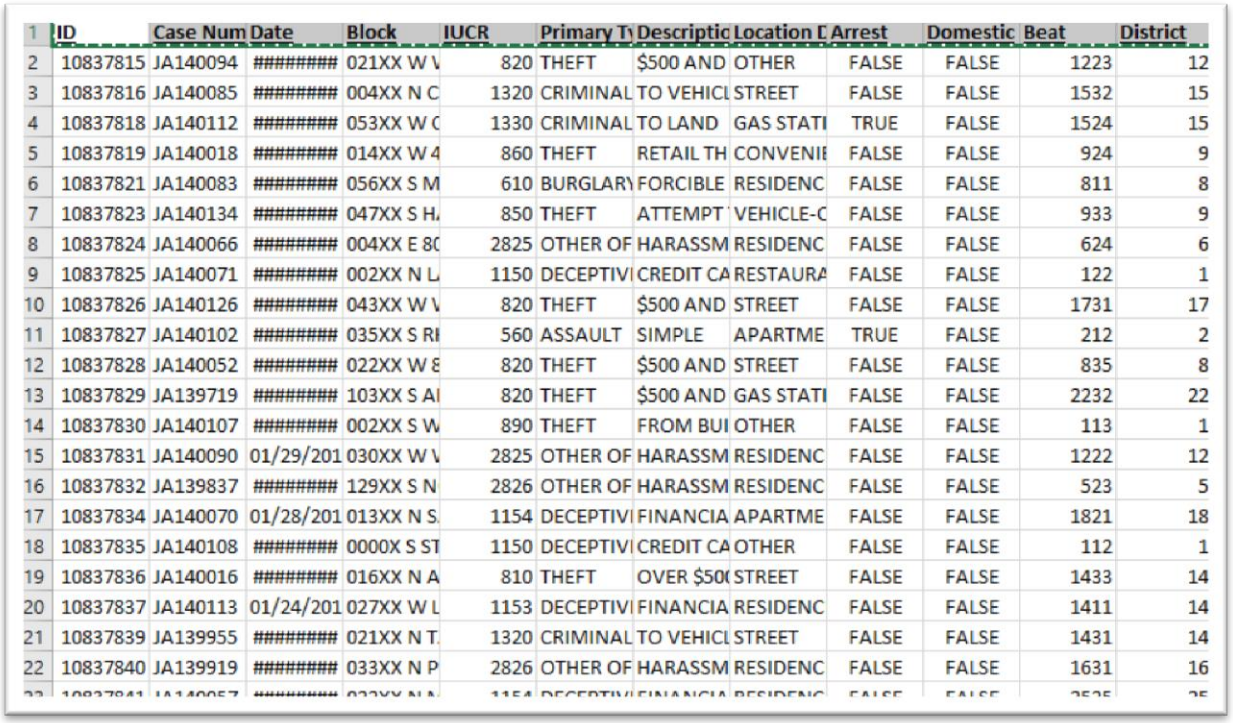
- VMware Virtual Workstation
- Windows and Cloudera OS
- Rapid Miner
- Hadoop HDFS
- Python
- Microsoft Excel



3. DATA SET DESCRIPTION

Dataset contains the following attributes:

1. *ID*
2. *Case Number*
3. *Date*
4. *Block*
5. *IUCR*
6. *Primary Type Description*
7. *Location Description*
8. *Arrest*
9. *Domestic*
10. *Beat*
11. *District Ward*
12. *Community Area*
13. *FBI Code*
14. *X Coordinate*
15. *Y Coordinate*
16. *Year*
17. *Updated On*
18. *Latitude*
19. *Longitude*
20. *Location*



1	ID	Case Num	Date	Block	IUCR	Primary Ty	Descriptic	Location D	Arrest	Domestic	Beat	District
2	10837815	JA140094	########	021XX W V	820	THEFT	$500 AND	OTHER	FALSE	FALSE	1223	12
3	10837816	JA140085	########	004XX N C	1320	CRIMINAL	TO VEHICL	STREET	FALSE	FALSE	1532	15
4	10837818	JA140112	########	053XX W C	1330	CRIMINAL	TO LAND	GAS STATI	TRUE	FALSE	1524	15
5	10837819	JA140018	########	014XX W 4	860	THEFT	RETAIL TH	CONVENIE	FALSE	FALSE	924	9
6	10837821	JA140083	########	056XX S M	610	BURGLARY	FORCIBLE	RESIDENC	FALSE	FALSE	811	8
7	10837823	JA140134	########	047XX S H	850	THEFT	ATTEMPT	VEHICLE-C	FALSE	FALSE	933	9
8	10837824	JA140066	########	004XX E 8(	2825	OTHER OF	HARASSM	RESIDENC	FALSE	FALSE	624	6
9	10837825	JA140071	########	002XX N L	1150	DECEPTIV	CREDIT CA	RESTAURA	FALSE	FALSE	122	1
10	10837826	JA140126	########	043XX W V	820	THEFT	$500 AND	STREET	FALSE	FALSE	1731	17
11	10837827	JA140102	########	035XX S RI	560	ASSAULT	SIMPLE	APARTME	TRUE	FALSE	212	2
12	10837828	JA140052	########	022XX W 8	820	THEFT	$500 AND	STREET	FALSE	FALSE	835	8
13	10837829	JA139719	########	103XX S AI	820	THEFT	$500 AND	GAS STATI	FALSE	FALSE	2232	22
14	10837830	JA140107	########	002XX S W	890	THEFT	FROM BUI	OTHER	FALSE	FALSE	113	1
15	10837831	JA140090	01/29/201	030XX W V	2825	OTHER OF	HARASSM	RESIDENC	FALSE	FALSE	1222	12
16	10837832	JA139837	########	129XX S N	2826	OTHER OF	HARASSM	RESIDENC	FALSE	FALSE	523	5
17	10837834	JA140070	01/28/201	013XX N S	1154	DECEPTIV	FINANCIA	APARTME	FALSE	FALSE	1821	18
18	10837835	JA140108	########	0000X S ST	1150	DECEPTIV	CREDIT CA	OTHER	FALSE	FALSE	112	1
19	10837836	JA140016	########	016XX N A	810	THEFT	OVER $50(	STREET	FALSE	FALSE	1433	14
20	10837837	JA140113	01/24/201	027XX W L	1153	DECEPTIV	FINANCIA	RESIDENC	FALSE	FALSE	1411	14
21	10837839	JA139955	########	021XX N T	1320	CRIMINAL	TO VEHICL	STREET	FALSE	FALSE	1431	14
22	10837840	JA139919	########	033XX N P	2826	OTHER OF	HARASSM	RESIDENC	FALSE	FALSE	1631	16

Fig 1. Dataset Image1

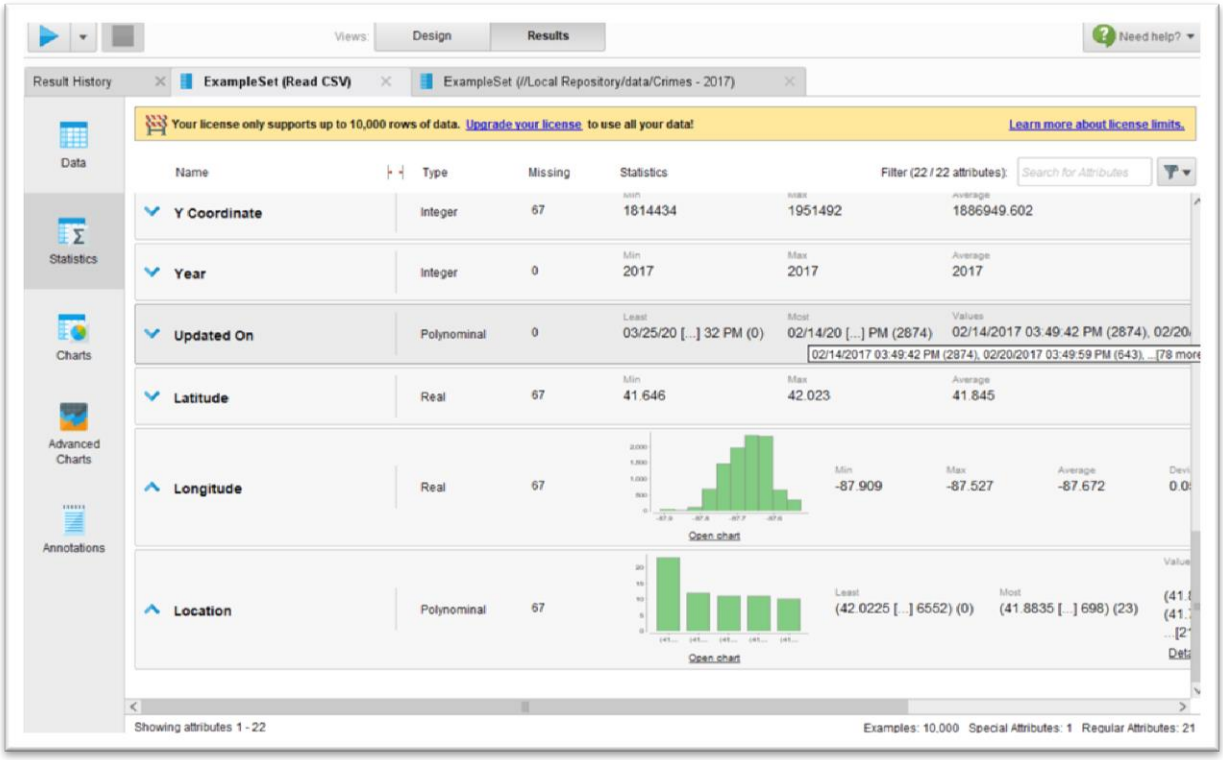


Fig 2. Dataset Description before cleaning

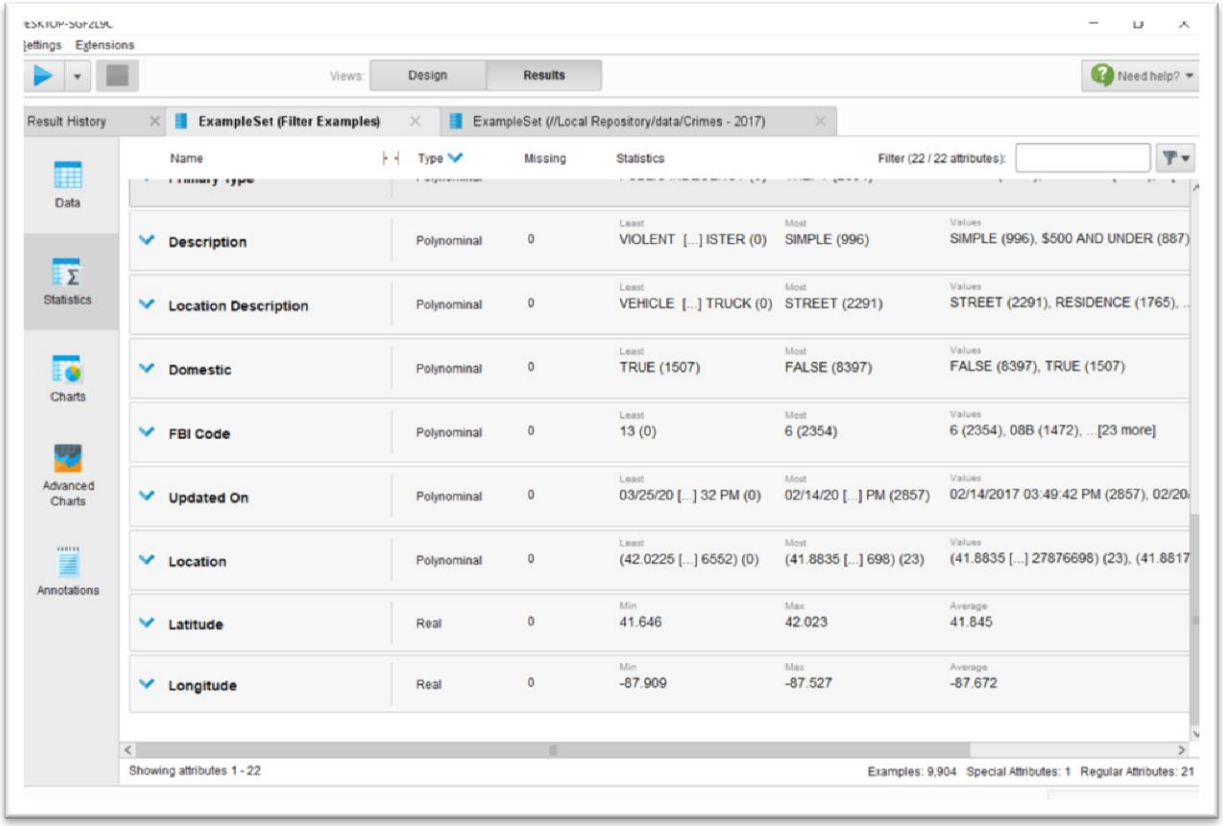


Fig 3. Dataset Description after cleaning

Data Description:

1. ID, Case No.: Attributes represents unique ID and number to a particular crime. Can be used as and identifier or primary key.
2. Primary Type: Attribute represents the type of crime that was committed.
3. FBI code: Attribute represents code assigned by FBI to the case file.
4. Date, Year: This represents the date and Year of Crime.
5. Block, Location, ward, District, Community: Represents Address at which crime was committed.
6. Updated: Represents the date on which record was last updated.
7. Arrest: Represents the arrests carried out in that particular crime or not.
8. Description: Give the description of crime committed.
9. Latitude, Longitude: Provides Specific coordinates for of Crime committed which can be clustered using algorithms for targeting areas.
10. Location: This is the (X, Y) format of location for the crime committed.

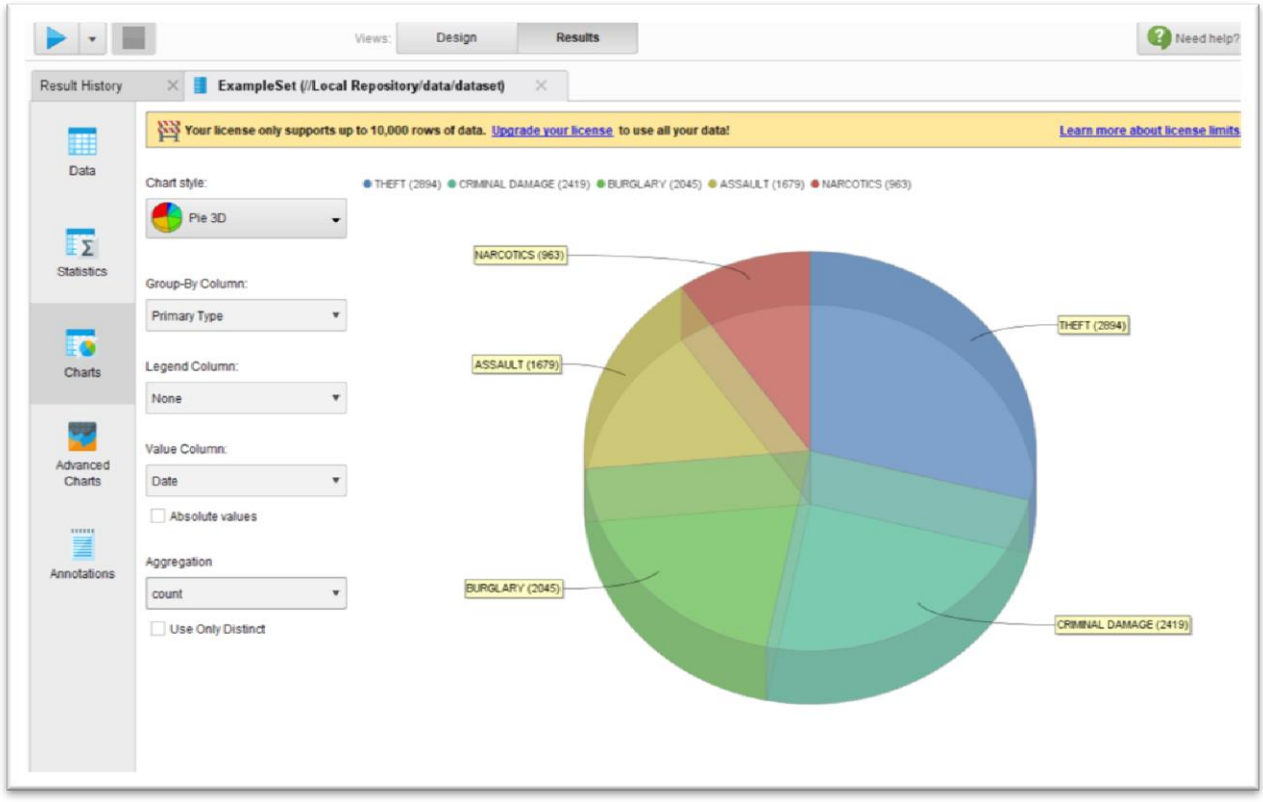


Fig 3. Total amounts of Crime

4. IMPLEMENTATION

CODE & OUTPUT:

location_mapper.py :

```
#!/usr/bin/python
Import sys
for input_line in sys.stdin:
        line = input_line.strip().split(“,”)
        #print ‘Line : ‘+str(line)
        x = line[12]     # x-coordinate
        y = line[13]     # y-coordinate
        #print ‘Time : ‘+str(time[0])
        #print ‘line ‘
        #print time[0]
        #count=count+1
        #print ‘count’+str(count)
        print “{0}\t{1}\t{2}”.format(x,y,str(input_line.strip()))
```

location_reducer.py :

```
#!/usr/bin/python
Import sys
tab = []
quad = []
count=0
for i in range(5):
        quad.append([])
        tab.append([])
for input_line in sys.stdin:
        #print “Output : “+str(input_line)
```

```
        line = input_line.strip().split("\t")                # X-coordinate \t Y-coordinate \t Tuple
        x = float(line[0])        # x-coordinate
        y = float(line[1])        # y-coordinate
        count=count+1
        #print 'count'+str(count)

        if x>=1095000 and x<1117000:# and y>=1836000 and y<1856600:
                quad[0].append(line[2])
        elif x>=1117000 and x<1139000:# and y>=1856600 and y<1877200:
                quad[1].append(line[2])
        elif x>=1139000 and x<1161000:# and y>=1877200 and y<1897800:
                quad[2].append(line[2])
        elif x>=1161000 and x<1183000:# and y>=1897800 and y<1918400:
                quad[3].append(line[2])
        elif x>=1183000 and x<1205000:# and y>=1918400 and y<1939000:
                quad[4].append(line[2])
for i in range(0,5):
        #theft_c=0,murder_c=0,robbery_c=0,rape_c=0,drugs_c=0
        theft_c = 0
        criminal_c=0
        burglary_c=0
        assault_c=0
        narcotics_c=0
        for j in range(len(quad[i])):
                crime_array = quad[i][j].strip().split(',')
                crime_type = str(crime_array[5])
                if crime_type == 'THEFT':
                        theft_c = theft_c + 1
                elif crime_type=='CRIMINAL DAMAGE':
```

```
                criminal_c = criminal_c + 1
            elif crime_type==’BURGLARY’:
                burglary_c = burglary_c + 1
            elif crime_type==’ASSAULT’:
                assault_c = assault_c + 1
            elif crime_type==’NARCOTICS’:
                narcotics_c = narcotics_c + 1


        total_crimes = theft_c + criminal_c + burglary_c + assault_c + narcotics_c
        print ‘\n\nFor Area ‘ + str(i+1) + ‘ – Occurrence of Crime types are:  ‘
        print ‘------------------------------------------------------------‘
        print ‘\t\tTHEFT\t\t\t: ‘+str(theft_c)
        print ‘\t\tCRIMINAL DAMAGE\t\t: ‘+str(criminal_c)
        print ‘\t\tBURGLARY\t\t: ‘+str(burglary_c)
        print ‘\t\tASSAULT\t\t\t: ‘+str(assault_c)
        print ‘\t\tNARCOTICS\t\t: ‘+str(narcotics_c)
        most = max(int(theft_c),int(criminal_c), int(burglary_c), int(assault_c),
int(narcotics_c))
        if most==theft_c:
            most_crime=’THEFT’
        elif most==criminal_c:
            most_crime=’CRIMINAL DAMAGE’
        elif most==burglary_c:
            most_crime=’BURGLARY’
        elif most==assault_c:
            most_crime=’ASSAULT’
        elif most==narcotics_c:
            most_crime=’NARCOTICS’
```

```
        print '\n\t\t=> Total Crimes 11ccurred and reported in above Area is : ' +
str(total_crimes) + ' reports.'

        Print '\n\t\t=> Also the most 11ccurred crime in above Area is : ' + most_crime + ' – '
+str(most) + ' times.'

        Tab[i].append(theft_c)

        tab[i].append(criminal_c)

        tab[i].append(burglary_c)

        tab[i].append(assault_c)

        tab[i].append(narcotics_c)

print
'\n\n*************************************************************************
*****************************'

print '\n\t\t\tTHEFT\t\tCRIMINAL\tBURGLARY\tASSAULT\t\tNARCOTICS'

for i in range(0,5):

        print "\nArea
"+str(i+1)+"\t\t\t"+str(tab[i][0])+"\t\t"+str(tab[i][1])+"\t\t"+str(tab[i][2])+"\t\t"+str(tab[i][3])
+"\t\t"+str(tab[i][4])

print
'\n\n*************************************************************************
*****************************'

print '\n\t* Total number of Tuples analysed are : ' + str(count)

print
'\n*************************************************************************
*****************************\n\n'

        #print "{0}\t{1}".format(output, 1)

        #print "{0}\t{1}".format(output, 1)
```

Output :

```
For Area 1 - Occurrence of Crime types are:
------------------------------------------------------------
                THEFT                   : 22
                CRIMINAL DAMAGE         : 7
                BURGLARY                : 2
                ASSAULT                 : 7
                NARCOTICS               : 0

                => Total Crimes occured and reported in above Area is : 38 reports.

                => Also the most occured crime in above Area is : THEFT - 22 times.


For Area 2 - Occurrence of Crime types are:
------------------------------------------------------------
                THEFT                   : 164
                CRIMINAL DAMAGE         : 176
                BURGLARY                : 182
                ASSAULT                 : 112
                NARCOTICS               : 37

                => Total Crimes occured and reported in above Area is : 671 reports.

                => Also the most occured crime in above Area is : BURGLARY - 182 times.
```

```
For Area 3 - Occurrence of Crime types are:
------------------------------------------------------------
                THEFT                   : 1054
                CRIMINAL DAMAGE         : 1112
                BURGLARY                : 999
                ASSAULT                 : 708
                NARCOTICS               : 748

                => Total Crimes occured and reported in above Area is : 4621 reports.

                => Also the most occured crime in above Area is : CRIMINAL DAMAGE - 1112 times.


For Area 4 - Occurrence of Crime types are:
------------------------------------------------------------
                THEFT                   : 2265
                CRIMINAL DAMAGE         : 1413
```

```
                BURGLARY                : 233
                ASSAULT                 : 266
                NARCOTICS               : 65

                => Total Crimes occured and reported in above Area is : 1251 reports.

                => Also the most occured crime in above Area is : CRIMINAL DAMAGE - 390 times.


****************************************************************************************************
```

Consolidated Output :

```
*****************************************************************************************************

                    THEFT           CRIMINAL        BURGLARY        ASSAULT         NARCOTICS
Area 1              22              7               2               7               0
Area 2              164             176             182             112             37
Area 3              1054            1112            999             708             748
Area 4              2265            1413            1156            1038            402
Area 5              297             390             233             266             65

*****************************************************************************************************
        * Total number of Tuples analysed are : 12855
*****************************************************************************************************
```

time_mapper.py :

```
#!/usr/bin/python

#  Write a MapReduce program which will display the number of hits for each different file on the Web site.

Import sys

for input_line in sys.stdin:

        line = input_line.strip().split(“,”)

        time = line[2].split(“ “)[1].split(“:”)   #time   this is the real time

        hour = float(str(time[0]))

        print “{0}\t{1}”.format(hour,str(input_line.strip()))
```

time_reducer.py :

```
#!/usr/bin/python

import sys

quad = []

tab = []

index=-1

count=0

theft=’THEFT’

for i in range(8):

        quad.append([])

        tab.append([])

for input_line in sys.stdin:

        line = input_line.strip().split(“\t”)                # Time \t Tuple

        time = str(line[0])       #time

        count=count+1

        #print ‘count’+str(count)

        hour = 0 + float(time)

        if hour>=0 and hour<3:
```

```
                quad[0].append(line[1])
        elif hour>=3 and hour<6:
                quad[1].append(line[1])
        elif hour>=6 and hour<9:
                quad[2].append(line[1])
        elif hour>=9 and hour<12:
                quad[3].append(line[1])
        elif hour>=12 and hour<15:
                quad[4].append(line[1])
        elif hour>=15 and hour<18:
                quad[5].append(line[1])
        elif hour>=18 and hour<21:
                quad[6].append(line[1])
        elif hour>=21 and hour<24:
                quad[7].append(line[1])
for i in range(0,8):
        theft_c = 0
        criminal_c=0
        burglary_c=0
        assault_c=0
        narcotics_c=0
        if i==0:                time_slot='00:00 – 02:59'
        if i==1:                time_slot='03:00 – 05:59'
        if i==2:                time_slot='06:00 – 08:59'
        if i==3:                time_slot='09:00 – 11:59'
        if i==4:                time_slot='12:00 – 14:59'
        if i==5:                time_slot='15:00 – 17:59'
        if i==6:                time_slot='18:00 – 20:59'
        if i==7:                time_slot='21:00 – 23:59'
```

```
        for j in range(len(quad[i])):
                crime_array = quad[i][j].strip().split(‘,’)
                crime_type = str(crime_array[5])
                if crime_type == ‘THEFT’:                       theft_c = theft_c + 1
                elif crime_type==’CRIMINAL DAMAGE’: criminal_c = criminal_c + 1
                elif crime_type==’BURGLARY’:                    burglary_c = burglary_c +
1
                elif crime_type==’ASSAULT’:                     assault_c = assault_c + 1
                elif crime_type==’NARCOTICS’:                   narcotics_c = narcotics_c
+ 1
        total_crimes = theft_c + criminal_c + burglary_c + assault_c + narcotics_c
        print ‘\n\nFor Time slot { ‘ + time_slot + ‘ } Occurrence of Crime types are:  ‘
        print ‘------------------------------------------------------------‘
        print ‘\t\tTHEFT\t\t\t: ‘+str(theft_c)
        print ‘\t\tCRIMINAL DAMAGE\t\t: ‘+str(criminal_c)
        print ‘\t\tBURGLARY\t\t: ‘+str(burglary_c)
        print ‘\t\tASSAULT\t\t\t: ‘+str(assault_c)
        print ‘\t\tNARCOTICS\t\t: ‘+str(narcotics_c)
        most = max(int(theft_c),int(criminal_c), int(burglary_c), int(assault_c),
int(narcotics_c))
        if most==theft_c:               most_crime=’THEFT’
        elif most==criminal_c:                  most_crime=’CRIMINAL DAMAGE’
        elif most==burglary_c:                  most_crime=’BURGLARY’
        elif most==assault_c:           most_crime=’ASSAULT’
        elif most==narcotics_c:                 most_crime=’NARCOTICS’
        print ‘\n\t\t=> Total Crimes 16ccurred and reported in above time slot is : ‘ +
str(total_crimes) + ‘ reports.’
        Print ‘\n\t\t=> Also the most 16ccurred crime in above time slot is : ‘ + most_crime +
‘ – ‘ +str(most) + ‘ times.’
        Tab[i].append(theft_c)
        tab[i].append(criminal_c)
```

```
        tab[i].append(burglary_c)

        tab[i].append(assault_c)

        tab[i].append(narcotics_c)

print '\n\n*************************************************************************************************************'

print '\n\t\t\tTHEFT\t\tCRIMINAL\tBURGLARY\tASSAULT\t\tNARCOTICS'

for i in range(0,8):

        #for j in range(0,5):

        print "\nSlot "+str(i+1)+"\t\t\t"+str(tab[i][0])+"\t\t"+str(tab[i][1])+"\t\t"+str(tab[i][2])+"\t\t"+str(tab[i][3])+"\t\t"+str(tab[i][4])

print '\n\t* Total number of Tuples analysed are : ' + str(count)

print '\n*************************************************************************************************************\n\n'
```

Output :

```
For Time slot { 00:00 - 02:59 } Occurrence of Crime types are:
-------------------------------------------------------------
                THEFT                   : 396
                CRIMINAL DAMAGE         : 415
                BURGLARY                : 320
                ASSAULT                 : 255
                NARCOTICS               : 119

                => Total Crimes occured and reported in above time slot is : 1505 reports.

                => Also the most occured crime in above time slot is : CRIMINAL DAMAGE - 415 times.


For Time slot { 03:00 - 05:59 } Occurrence of Crime types are:
-------------------------------------------------------------
                THEFT                   : 476
                CRIMINAL DAMAGE         : 426
                BURGLARY                : 408
                ASSAULT                 : 309
                NARCOTICS               : 85

                => Total Crimes occured and reported in above time slot is : 1704 reports.

                => Also the most occured crime in above time slot is : THEFT - 476 times.


For Time slot { 06:00 - 08:59 } Occurrence of Crime types are:
-------------------------------------------------------------
                THEFT                   : 516
                CRIMINAL DAMAGE         : 567
                BURGLARY                : 459
                ASSAULT                 : 351
                NARCOTICS               : 190

                => Total Crimes occured and reported in above time slot is : 2083 reports.

                => Also the most occured crime in above time slot is : CRIMINAL DAMAGE - 567 times.


For Time slot { 09:00 - 11:59 } Occurrence of Crime types are:
-------------------------------------------------------------
                THEFT                   : 668
                CRIMINAL DAMAGE         : 603
```

```
BURGLARY                : 187
ASSAULT                 : 125
NARCOTICS               : 114

=> Total Crimes occured and reported in above time slot is : 925 reports.

=> Also the most occured crime in above time slot is : THEFT - 266 times.
```

****

	THEFT	CRIMINAL	BURGLARY	ASSAULT	NARCOTICS
Slot 1	396	415	320	255	119
Slot 2	476	426	408	309	85
Slot 3	516	567	459	351	190
Slot 4	668	603	519	435	287
Slot 5	644	323	297	268	186
Slot 6	443	224	193	205	91
Slot 7	393	307	189	183	180
Slot 8	266	233	187	125	114

```
        * Total number of Tuples analysed are : 12855
```

****

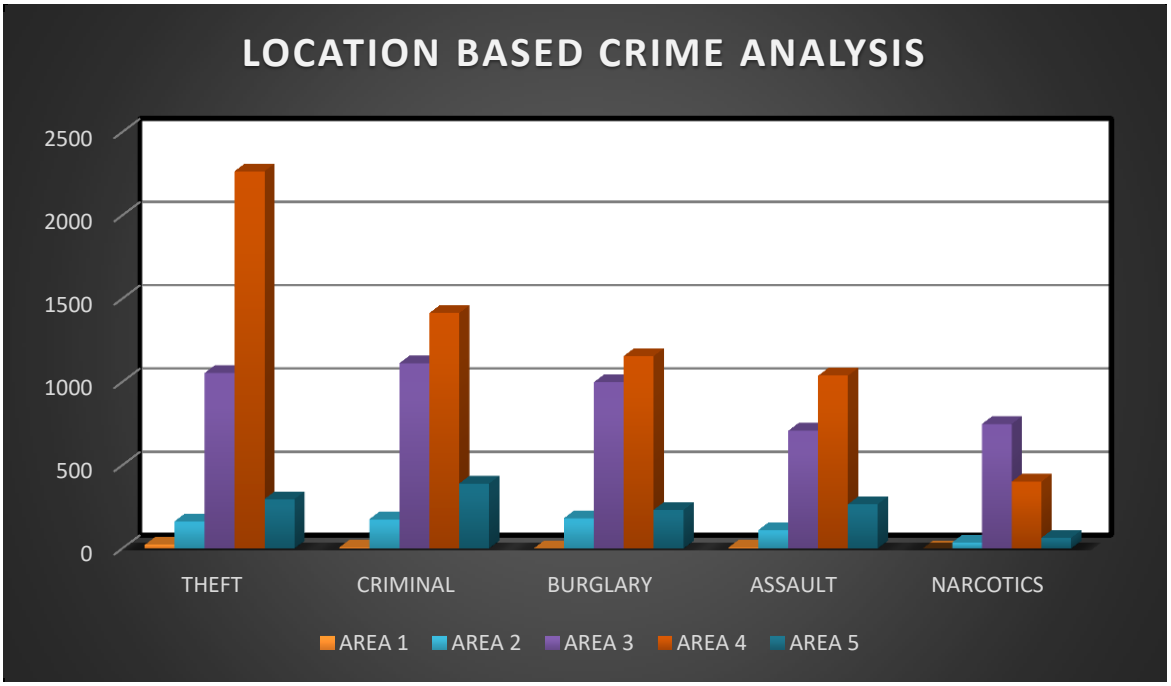
5. RESULTS

Location Based Analysis:

Table

	THEFT	CRIMINAL	BURGLARY	ASSAULT	NARCOTICS
AREA 1	22	7	2	7	0
AREA 2	164	176	182	112	37
AREA 3	1054	1112	999	708	748
AREA 4	2265	1413	1156	1038	402
AREA 5	297	390	233	266	65

Bar – Graph

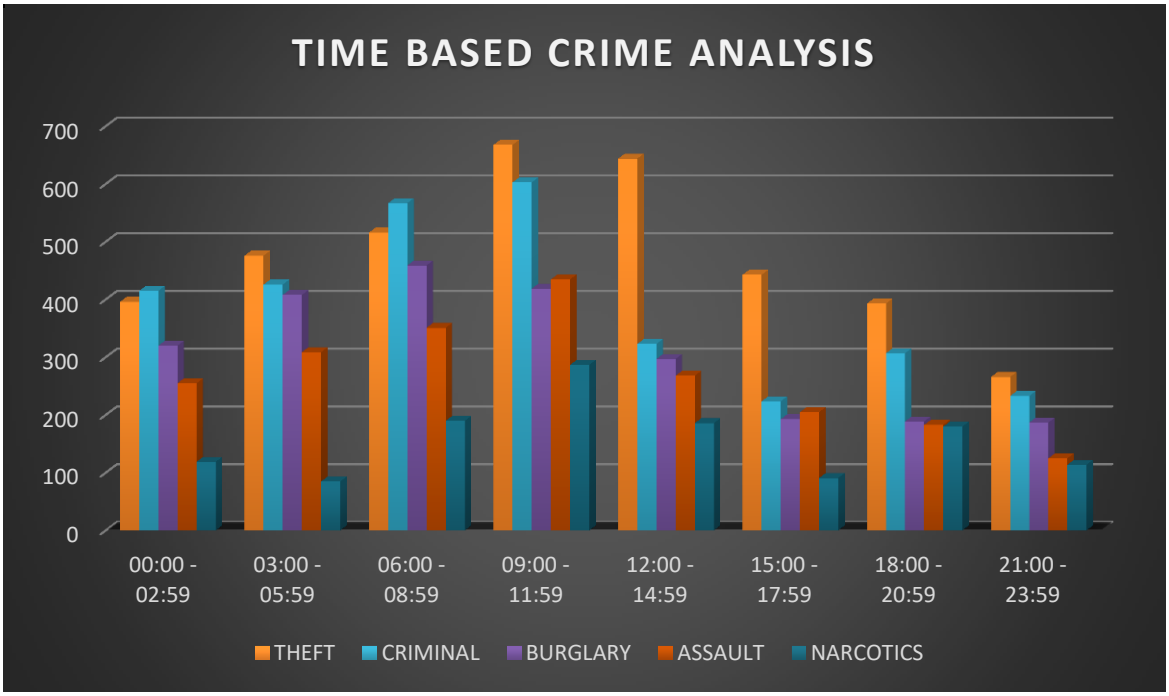


Time Based Analysis:

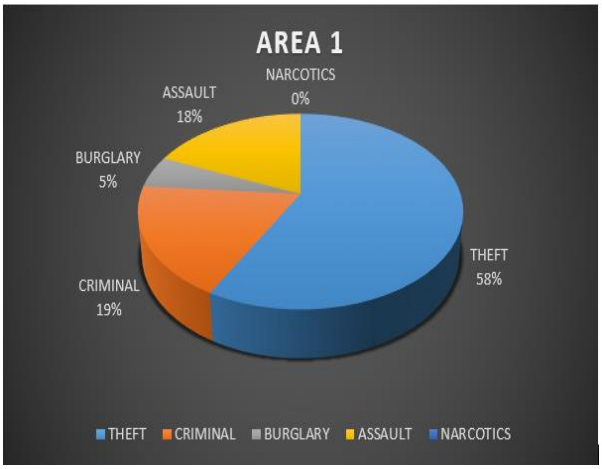
Table

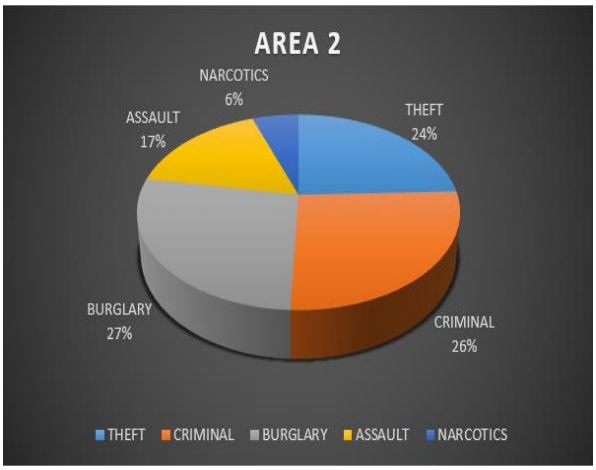
	THEFT	CRIMINAL	BURGLARY	ASSAULT	NARCOTICS
00:00 – 02:59	396	415	320	255	119
03:00 – 05:59	476	426	408	309	85
06:00 – 08:59	516	567	459	351	190
09:00 – 11:59	668	603	419	435	287
12:00 – 14:59	644	323	297	268	186
15:00 – 17:59	443	224	193	205	91
18:00 – 20:59	393	307	189	183	180
21:00 – 23:59	266	233	187	125	114

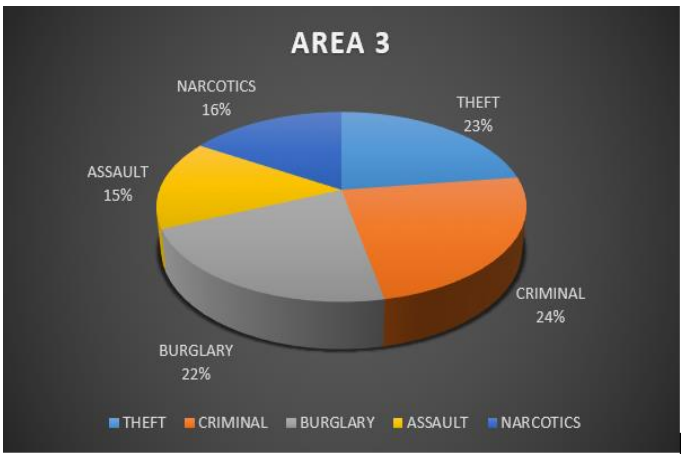
Bar – Graph

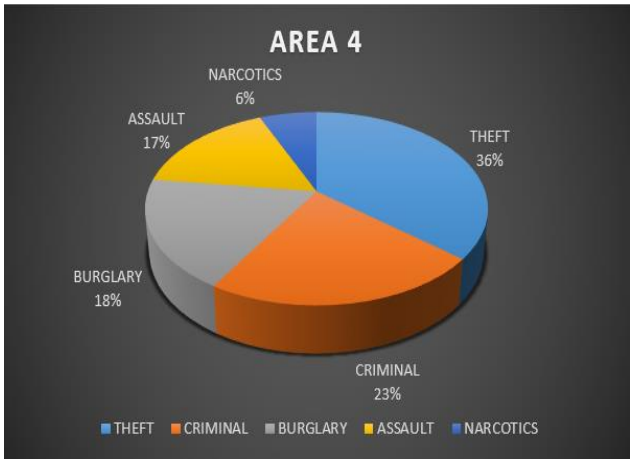


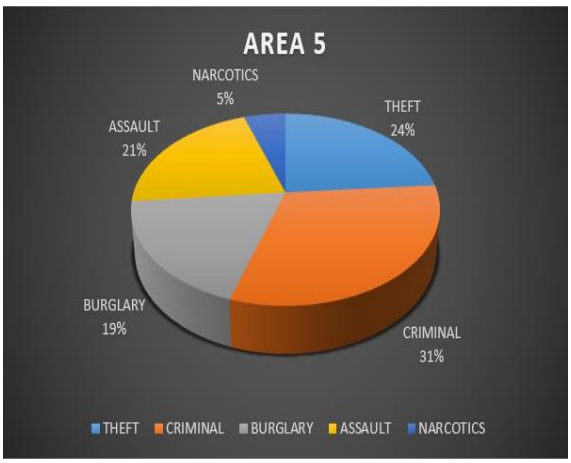
Location Based





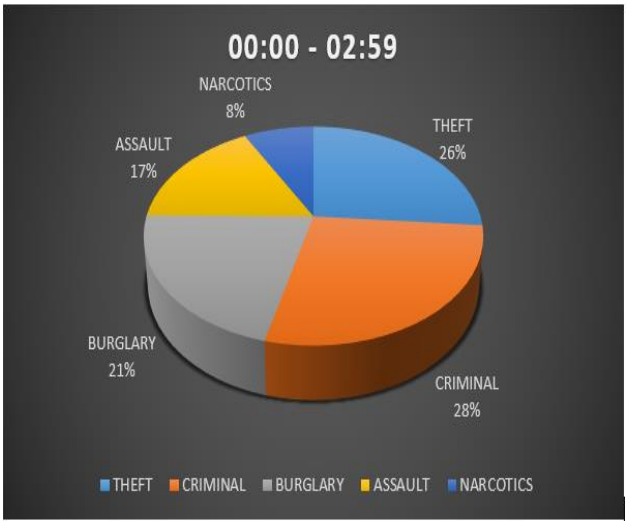




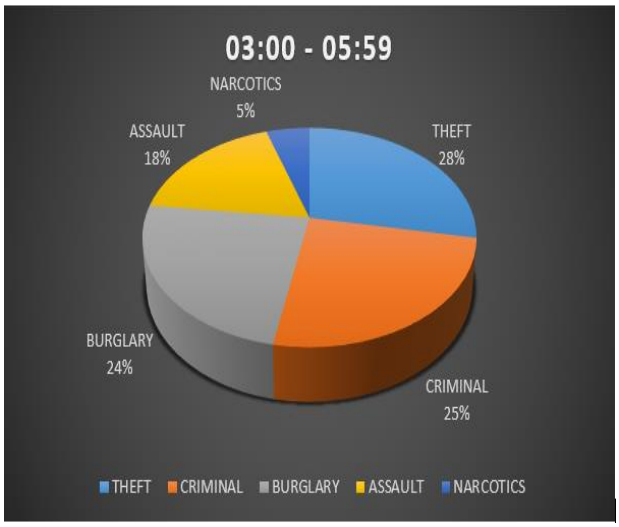


Time Based

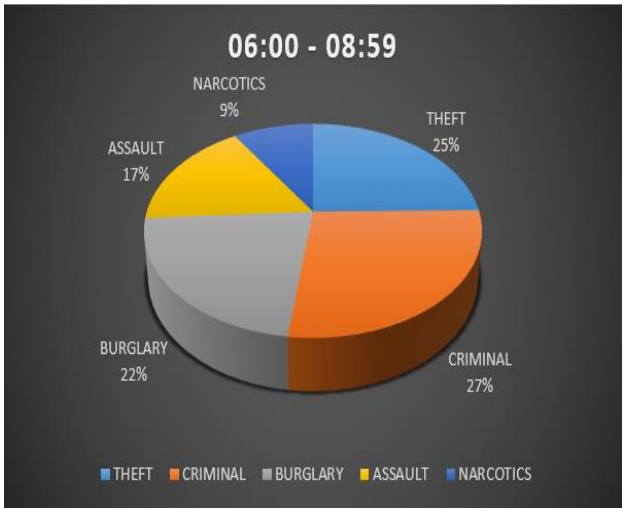
12:00 am – 02:59 am



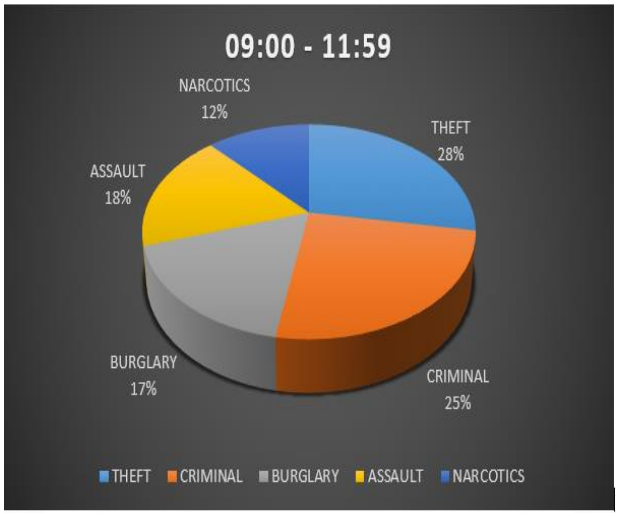
03:00 am – 05:59 am



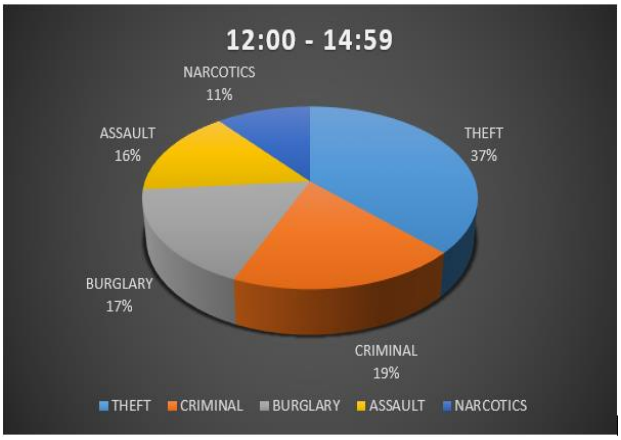
06:00 am – 08:59 am



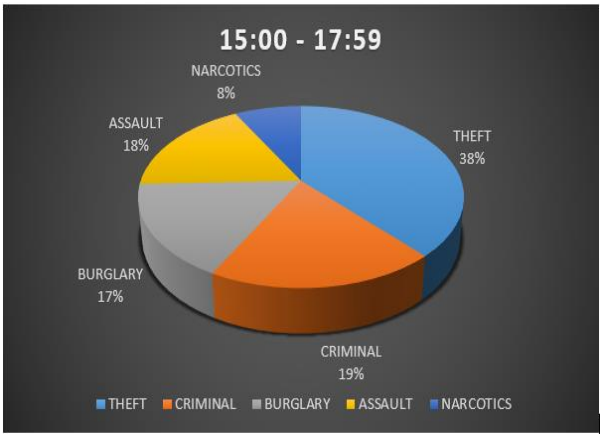
09:00 am – 11:59 am



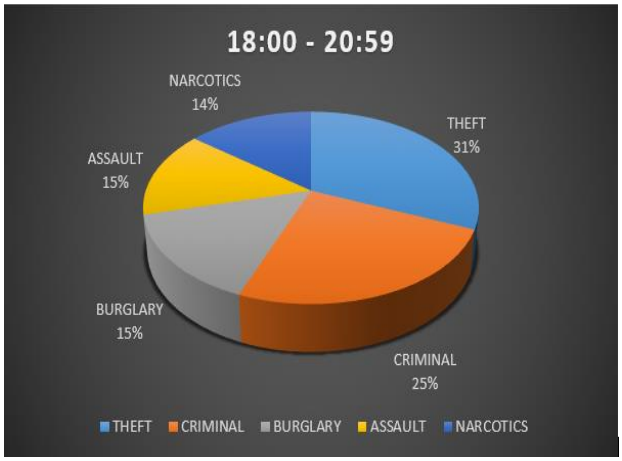
12:00 pm – 02:59 pm



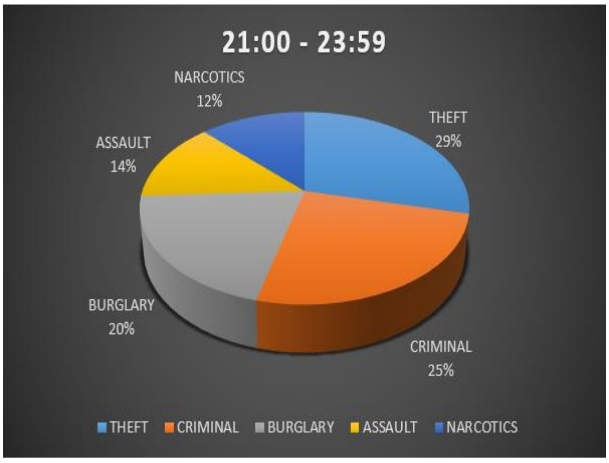
03:00 pm – 05:59 pm



06:00 pm – 08:59 pm



09:00 am – 11:59 am



6. CONCLUSION AND FUTURE WORK

Hence, I have made a project which helps to analyse the crimes based on time and location. Suggestions are provided on what policing practice should be implemented based on what crime is taking place according to the time and location.

In future, I plan to extend our project to a larger scale to implement analysis based on more different aspects. The different aspects include the prediction of the maximum crime taking place with respect to time and location together, etc.

I have done my best in coding and implementing the mapper and reducer to predict the crimes taking place with respect to time and location in process of completion of this project.

6. REFERENCES

- Han J., Kamber M., “Data Mining: Concepts and Techniques”, Morgan Kaufmann (Elsevier), 2006.
- Ricci and F. Del Missier, “Supporting Travel Decision making Through Personalized Recommendation,” Design Personalized User Experience for e-commerce, pp. 221-251, 2004.
- Giles C.L., Bollacker K.D., and Lawrence S., “CiteSeer: An automatic citation indexing system,” in Proceedings of the third ACM conference on Digital libraries, 1998, pp. 89–98.
- https://data.gov.org (2017 Dataset Crime)