ABIYAANTRIX & SAPIENCE ACADEMY
INTERNSHIP + TRAINING

Academy & Management Solutions

Internship Mini Project on

"CHICAGO CRIME DATASET"

Submitted in partial fulfillment towards Mini Project work of Internship

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE ENGINEERING

SUBMITTED BY

Vivek Pawar	**4CA15CS021**
Shantha Raman S	**4PS15CS100**
Anushree G	**4GW15CS009**
Bhavini Pahwa	**1RN16CS020**
Komal B R	**4GW15CS040**
Prakruthi S	**4GW16CS415**
Akash V	**01JST16IS051**

UNDER THE GUIDANCE OF

Mrs. ANJANA SHASHI KIRAN
DIRECTOR

Mr. SRIKRISHNA S KASHYAP
INTERNSHIP TRAINER

ACKNOWLEDGMENT

We sincerely owe our gratitude to all the persons who helped and guided us in completing this mini-project.

We are thankful to **Mrs. Anjana Shashi Kiran**, *Honorary Director*, Abiyaantrix Tech Solutions, Mysuru, for having supported in our academic endeavours.

We are thankful to **Mr. Srikrishna S Kashyap**, *Trainer*, Abiyaantrix Tech Solutions, Mysuru, for all the support he has rendered.

We are extremely pleased to thank our parents, family members and friends for their continuous support, inspiration and encouragement, for their helping hand and also last but not the least, We thank all the members who supported directly or indirectly in this internship process.

Vivek Pawar
Shantha Raman S
Anushree G
Bhavini Pahwa
Komal B R
Prakruthi S
Akash V

ABSTRACT

It is the process by which order, structure and meaning are given to the data (information).

It consists in transforming the collected data into useful and true conclusions and or lessons.

From the pre-established topics, the data are processed, looking for trends, differences and variations in the information obtained.

The processes, techniques and tools used are based on certain assumptions and as such have limitations.

The process is used to describe and summarize the data, identify the relationships and differences between variables, compare variables and make predictions.

CONTENTS

INTRODUCTION

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar todata mining.

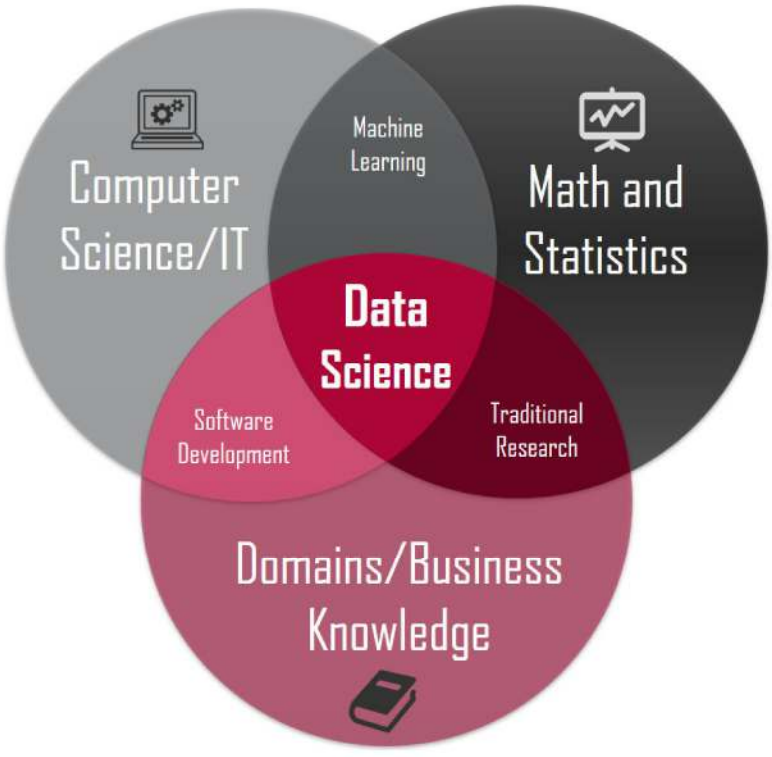
Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyse actual phenomena" with data.It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

HISTORY

The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage it was used as a substitute for computer science by Peter Naur in 1960. Naur later introduced the term "datalogy".In 1974, Naur published *Concise Survey of Computer Methods*, which freely used the term data science in its survey of the contemporary data processing methods that are used in a wide range of applications.

RELATIONSHIP TO STATISTICS



The popularity of the term "data science" has exploded in business environments and academia, as indicated by a jump in job openings. However, many critical academics and journalists see no distinction between data science and statistics. Writing in Forbes, Gil Press argues that data science is a buzzword without a clear definition and has simply replaced “business analytics” in contexts such as graduate degree programs.In the question-and-answer section of his keynote address at the Joint Statistical Meetings of American Statistical Association, noted applied statistician Nate Silver said, “I think data-scientist is a sexed up term for a statistician....Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn’t berate the term statistician.”Similarly, in business sector, multiple researchers and analysts state that data scientists alone are far from being sufficient in granting companies a real competitive advantage and consider data scientists as only one of the four greater job families companies require to leverage big data effectively, namely: data analysts, data scientists, big data developers and big data engineers.On the other hand, responses to criticism are as numerous. In a 2014 Wall Street Journal article, Irving Wladawsky-Berger compares the data science enthusiasm with the dawn of computer science. He argues data science, like any other interdisciplinary field, employs methodologies and practices from across the academia and industry, but then it will morph them into a new discipline. He brings to attention the sharp criticisms computer science, now a well-respected academic discipline, had to once face. Likewise, NYUStern's Vasant Dhar, as do many other academic proponents of data science, argues more specifically in December 2013 that data science is different from the existing practice of data analysis across all disciplines, which focuses only on explaining data sets. Data science seeks actionable and consistent pattern for predictive uses. This practical engineering goal takes data science beyond traditional analytics. Now the data in those disciplines and applied fields that lacked solid theories, like health science and social science, could be sought and utilized to generate powerful predictive models.

SYSTEMREQUIREMENTANDSPECIFICATION

2.1HardwareRequirements:

- Intel® Pentium 4 CPU and higher versions
- 256 MBRAM,
- 80GBHDD Mouse
- QWERTYKeyboard
- Standard VGAMonitor

2.2SoftwareRequirements:

- OperatingSystem:WINDOWS 10,7
- Language :PYTHON
- Tool : JUPYTER NOTEBOOK

TESTING AND RESULTS

The full creating and implementing Chicago Crime Dataset using python, in which we have used various python modules. Modules included pandas to handle dataframes, matplotlib to plot a graph ,numpy, gmaps to display heatmap.

UNIT TESTING

Here the individual components are tested to ensure that they operate correctly. Each component is tested independently, without other system components.

MODULE TESTING

Module is a collection of dependent components such as procedures and functions. Since the module encapsulates related components can be tested with our other system modules. The testing process is concerned with finding errors which results from erroneous function calls from the main function to various individual functions.

SYSTEM TESING

The Modules are integrated to make up the entire system. The testing process is concerned with finding errors with the results from unanticipated interactions between module and system components. It is also concerned with validating that the system meets its functional and non-functional requirements.

IMPLEMENTATION

- import pandas as pd

  from pandas import read_csv

  crimes = read_csv('data1.csv', index_col='Date', nrows=20000)

  print(crimes.head())

- crimes = crimes.iloc[:, 3:]
  crimes.head()

- crimes.index = pd.to_datetime(crimes.index)
  print(crimes.shape)
  print(crimes.head())

- s = crimes[['Primary Type']]
  s.head()

- crime_count=pd.DataFrame(s.groupby('Primary Type').size().sort_values(ascending=False).rename('counts').reset_index())
  print(crime_count.head())

- import seaborn as sns
  import matplotlib.pyplot as plt
  sns.set(style="whitegrid")
  # Initialize the matplotlib figure
  f, ax = plt.subplots(figsize=(6, 15))

```
      # Plot the total crashes
sns.set_color_codes("pastel")
sns.barplot(x="counts", y="Primary Type", data=crime_count.iloc[:10, :],
          label="Total", color="b")
     ax.legend(ncol=2, loc="lower right", frameon=True)
     ax.set(ylabel="Type" xlabel="Crimes")
     sns.despine(left=True, bottom=True)
# Add a legend and informative axis label
plt.show()
```

Arrests

```
        crimes_2014 = crimes.loc['2014']
        crimes_2015 = crimes.loc['2015']

        # Yearly crimes 12 to 17
        arrest_yearly = crimes[crimes['Arrest'] == True]['Arrest']
        print(arrest_yearly.head())
        plt.subplot()

        # yearly arrest
        arrest_yearly.resample('A').sum().plot()
        plt.title('Yearly arrests')
        plt.show()

        # Monthly arrest
        arrest_yearly.resample('M').sum().plot()
        plt.title('Monthly arrests')
        plt.show()
```

```
# Weekly arrest
arrest_yearly.resample('W').sum().plot()
plt.title('Weekly arrests')
plt.show()

# daily arrest
arrest_yearly.resample('D').sum().plot()
plt.title('Daily arrests')
plt.show()
plt.show()
```

Domestic violence

- domestic_yearly = crimes[crimes['Domestic'] == True]['Domestic']#print(domestic_yearly.head())
- plt.subplot()

```
# yearly domestic violence
domestic_yearly.resample('A').sum().plot()
plt.title('Yearly domestic violence')
plt.show()

# Monthly domestic violence
domestic_yearly.resample('M').sum().plot()
plt.title('Monthly domestic violence')
plt.show()

# Weekly domestic violence
domestic_yearly.resample('W').sum().plot()
plt.title('Weekly domestic violence')
plt.show()

# daily domestic violence

domestic_yearly.resample('D').sum().plot()
plt.title('Daily domestic violence')
```

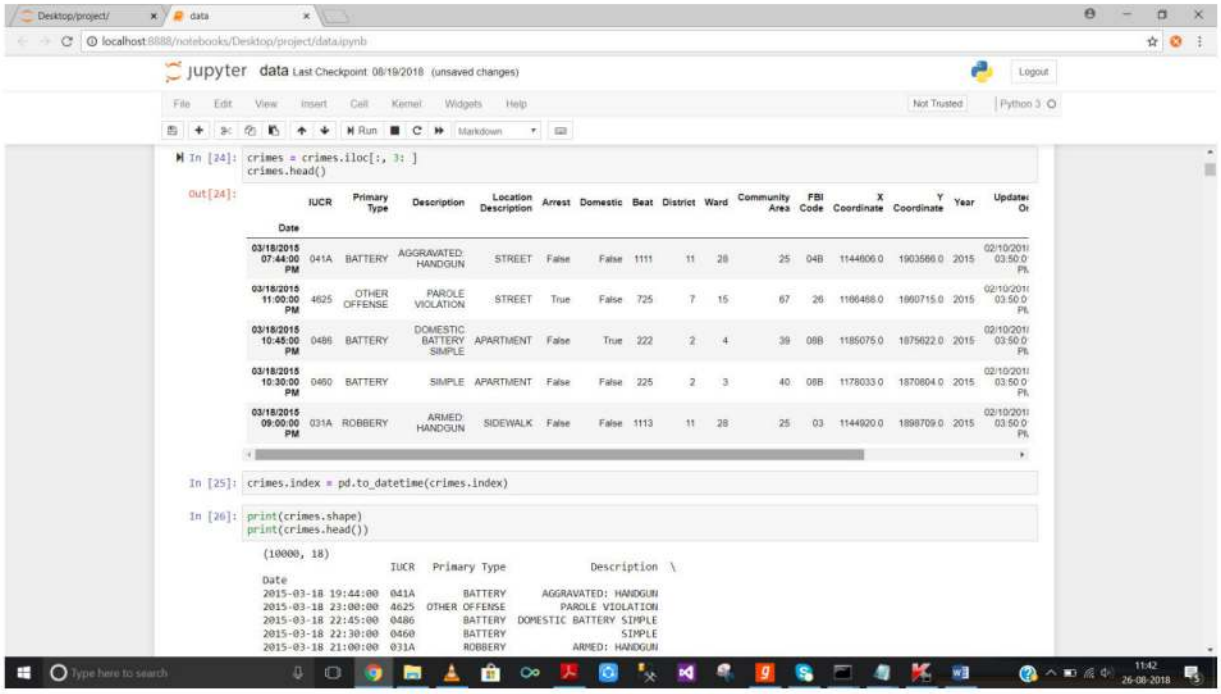
```
plt.show()
```

```
plt.show()
```

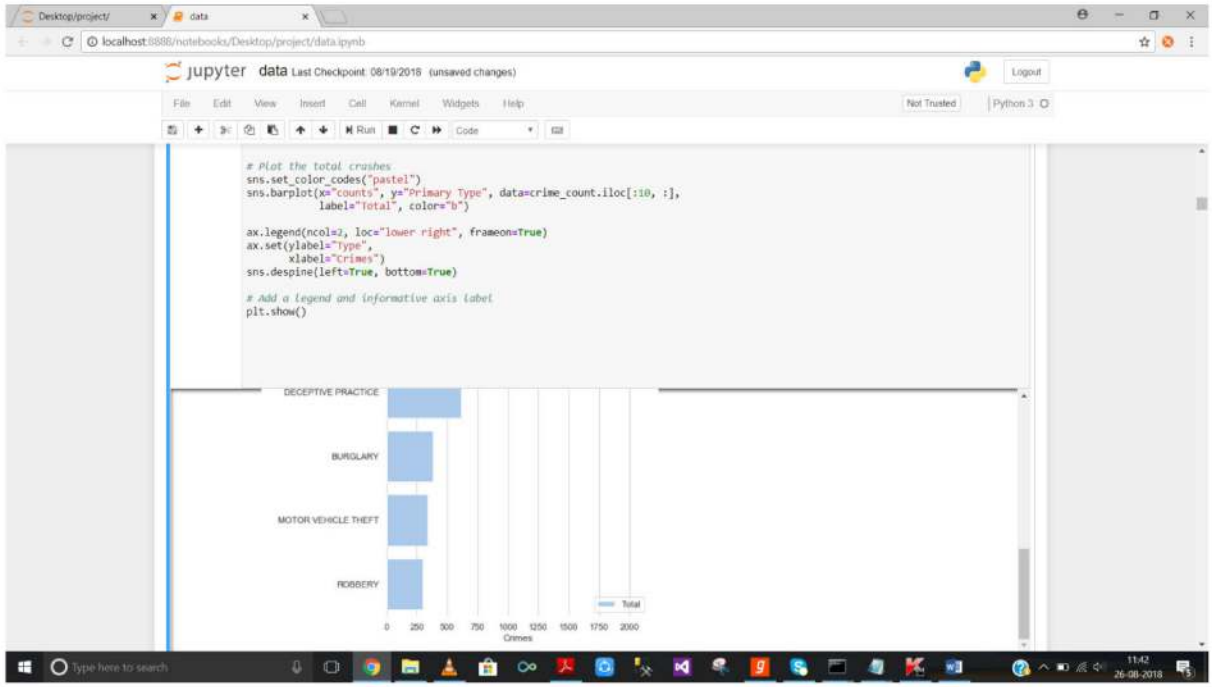
HEAT MAPS

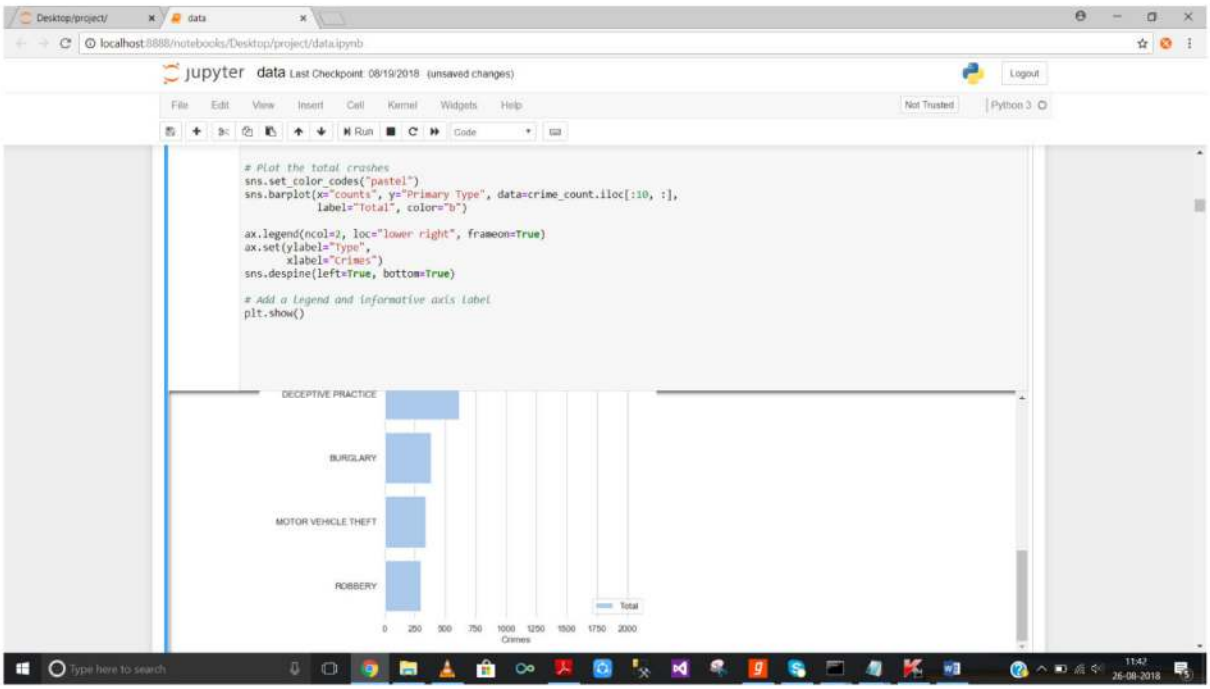
```
import gmaps
gmaps.configure(api_key='AIzaSyDy7-k3aoklchxeq2B41yH-...API KEY')
fig = gmaps.figure(map_type='SATELLITE')
heatmap_layer = gmaps.heatmap_layer(data)
fig.add_layer(heatmap_layer)
fig
```

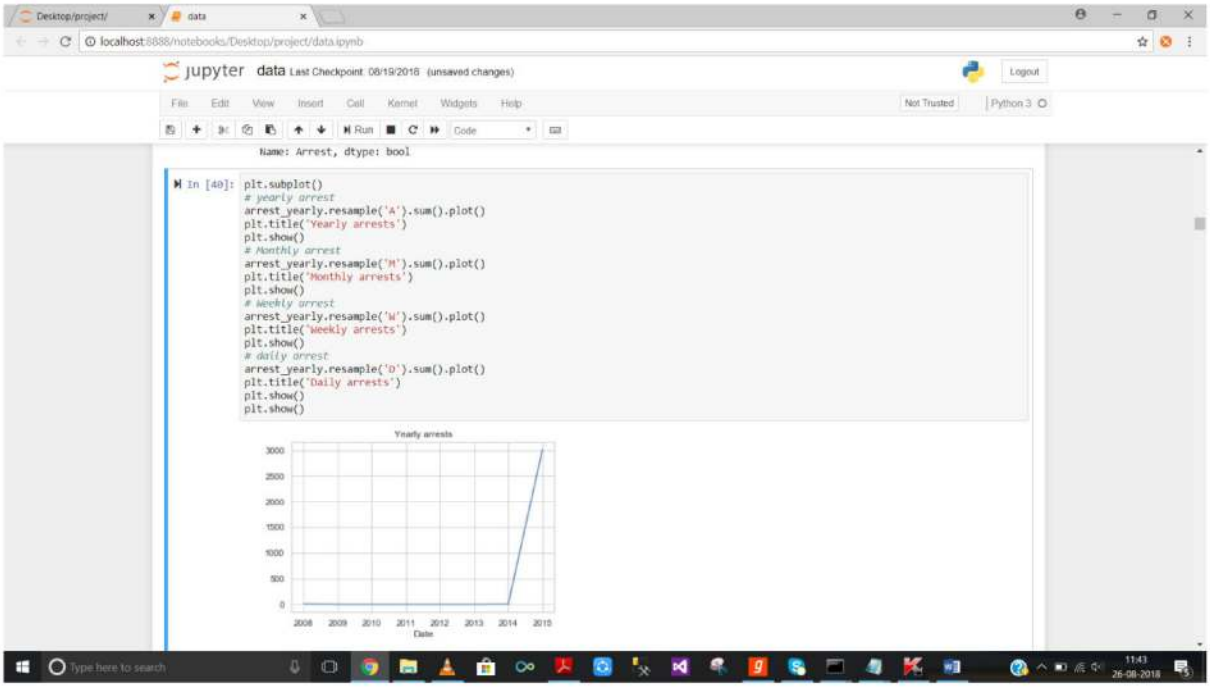
SNAPSHOTS

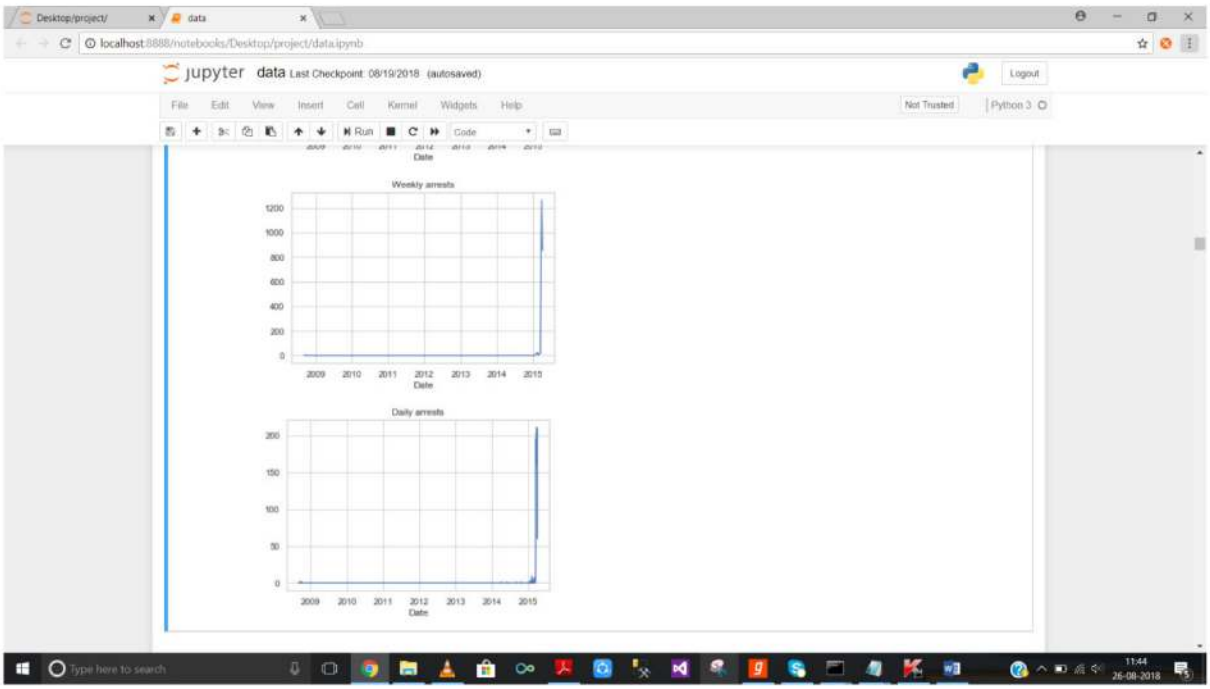


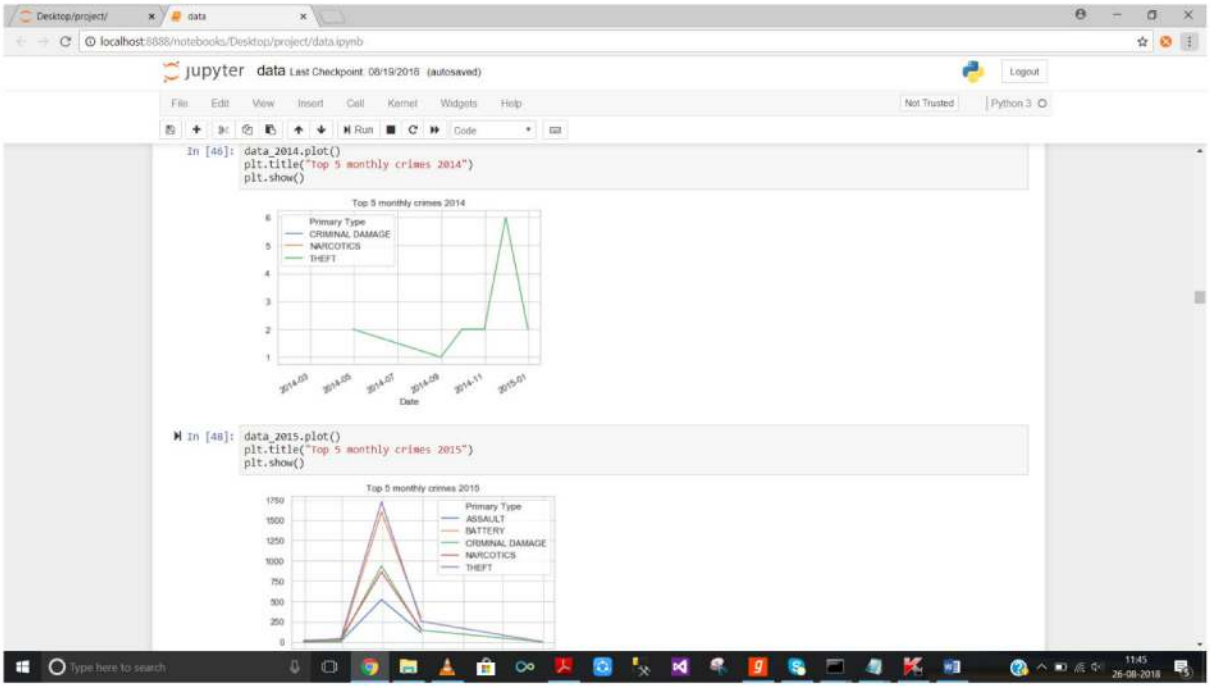
Date	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	District	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year	Updated On
03/18/2015 07:44:00 PM	041A	BATTERY	AGGRAVATED: HANDGUN	STREET	False	False	1111	11	28	25	04B	1144606.0	1903566.0	2015	02/10/2018 03:50:0 PM
03/18/2015 11:00:00 PM	4625	OTHER OFFENSE	PAROLE VIOLATION	STREET	True	False	725	7	15	67	26	1166468.0	1860715.0	2015	02/10/2018 03:50:0 PM
03/18/2015 10:45:00 PM	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	False	True	222	2	4	39	08B	1185075.0	1875622.0	2015	02/10/2018 03:50:0 PM
03/18/2015 10:30:00 PM	0460	BATTERY	SIMPLE	APARTMENT	False	False	225	2	3	40	08B	1178033.0	1870804.0	2015	02/10/2018 03:50:0 PM
03/18/2015 09:00:00 PM	031A	ROBBERY	ARMED: HANDGUN	SIDEWALK	False	False	1113	11	28	25	03	1144920.0	1898709.0	2015	02/10/2018 03:50:0 PM

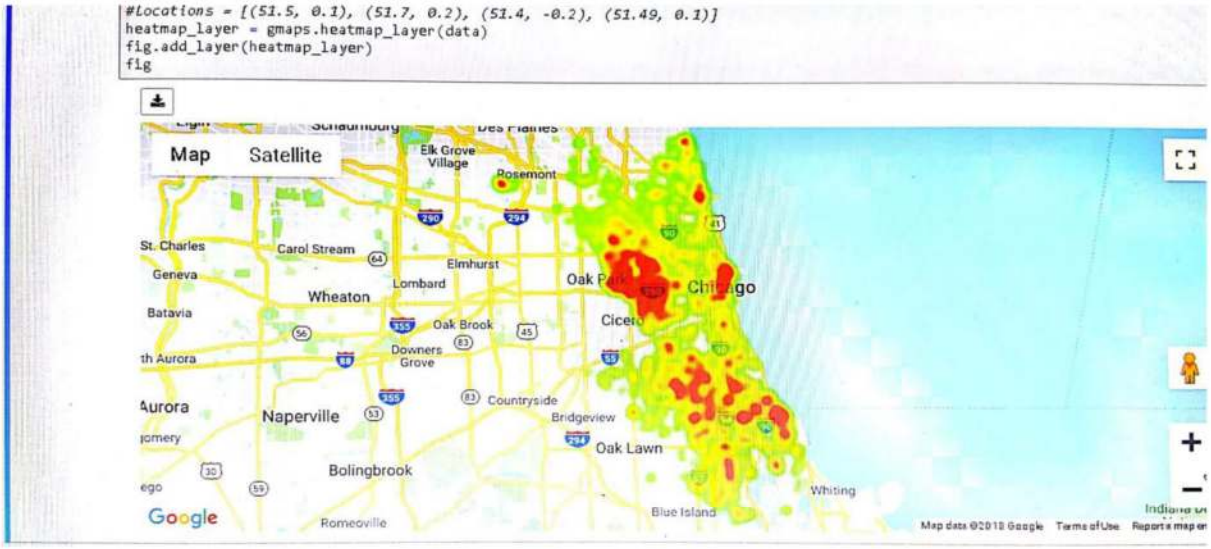



Desktop/project/
data
localhost:8888/notebooks/Desktop/project/data.ipynb
jupyter data Last Checkpoint: 08/19/2018 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
# Plot the total crashes
sns.set_color_codes("pastel")
sns.barplot(x="counts", y="Primary Type", data=crime_count.iloc[:10, :],
label="Total", color="b")
ax.legend(ncol=2, loc="lower right", frameon=True)
ax.set(ylabel="Type",
xlabel="Crimes")
sns.despine(left=True, bottom=True)
# Add a legend and informative axis label
plt.show()
DECEPTIVE PRACTICE
BURGLARY
MOTOR VEHICLE THEFT
ROBBERY
Total
0 250 500 750 1000 1250 1500 1750 2000
Crimes
Type here to search
11:42
26-08-2018


Desktop/project/
data
localhost:8888/notebooks/Desktop/project/data.ipynb
jupyter data Last Checkpoint: 08/19/2018 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Name: Arrest, dtype: bool
In [40]:
plt.subplot()
# yearly arrest
arrest_yearly.resample('A').sum().plot()
plt.title('Yearly arrests')
plt.show()
# Monthly arrest
arrest_yearly.resample('M').sum().plot()
plt.title('Monthly arrests')
plt.show()
# Weekly arrest
arrest_yearly.resample('W').sum().plot()
plt.title('Weekly arrests')
plt.show()
# daily arrest
arrest_yearly.resample('D').sum().plot()
plt.title('Daily arrests')
plt.show()
plt.show()
Yearly arrests
3000
2500
2000
1500
1000
500
0
2008 2009 2010 2011 2012 2013 2014 2015
Date
Type here to search
11:43
26-08-2018


Desktop/project/
data
localhost:8888/notebooks/Desktop/project/data.ipynb
jupyter data Last Checkpoint: 08/19/2018 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Date
Weekly arrests
1200
1000
800
600
400
200
0
2009 2010 2011 2012 2013 2014 2015
Date
Daily arrests
200
150
100
50
0
2009 2010 2011 2012 2013 2014 2015
Date
Type here to search
11:44
26-08-2018


Desktop/project/
data
localhost:8888/notebooks/Desktop/project/data.ipynb
jupyter data Last Checkpoint: 08/19/2018 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [46]:
data_2014.plot()
plt.title("Top 5 monthly crimes 2014")
plt.show()
Top 5 monthly crimes 2014
Primary Type
CRIMINAL DAMAGE
NARCOTICS
THEFT
6
5
4
3
2
1
2014-03
2014-05
2014-07
2014-09
2014-11
2015-01
Date
In [48]:
data_2015.plot()
plt.title("Top 5 monthly crimes 2015")
plt.show()
Top 5 monthly crimes 2015
Primary Type
ASSAULT
BATTERY
CRIMINAL DAMAGE
NARCOTICS
THEFT
1750
1500
1250
1000
750
500
250
0
Type here to search
11:45
26-08-2018



Google Heat maps

FUTURE ENHANCEMENTS

In order to achieve mastery over working with abundant data, this data set can serve as the ideal stepping stone in the pursuit of tackling mountainous data. We can implement this Dataset with R language which is more powerful compared to python.

CONCLUSION

An overwhelming expansion of data archives posed a challenge to various industries, as these are now struggling to make use of such enormous amount of information. Almost 90% of all data ever recorded worldwide has been created in the last decade alone.

In this project we have explored the data and it provides the insights and forecasts about crimes in Chicago. It extracts the data from Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. It contains information on reported incidents of crime in the city of Chicago from 2001 to present.

BIBLIOGRAPHY

- Newspaper article
- Wikipedia.org
- Techopedia.com
- Stackoverflow.com
- Quora.com
- Elitedatascience.com/data-cleaning
- Trifecta.com
- Mean.io
- Searchbusinessanalytics.techtarget.com
- Optimizely.com
- Geeksforgeeks.com
- Youtube.com
- Studytonight.com
- Techterms.com
- Google images
- Codingdojo.com
- Github.com
- Hackermoon.com