**ABIYAANTRIX & SAPIENCE ACADEMY**

**INTERNSHIP + TRAINING**



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 Mr. SRIKRISHNA S KASHYAP**

**DIRECTOR 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**

Acknowledgement 2

Abstract 3

Contents 4

**1.** **Introduction 5-8**

**2.** **System** **Requirement** **and** **Specification** **9**

2.1 Hardware Requirements 9

2.2 Software Requirements 9

**3.** **Testing** **and** **Results** **10**

**4.** **Implementation 11-13**

**5.** **Snapshots** **14-16**

**Future Enhancement**  **17** **Conclusion 17**

**Bibliography** **18**

**INTRODUCTION**



**Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract [knowledge](https://en.wikipedia.org/wiki/Knowledge) and insights from [data](https://en.wikipedia.org/wiki/Data) in various forms, both structured and unstructured, similar to [data mining](https://en.wikipedia.org/wiki/Data_mining).

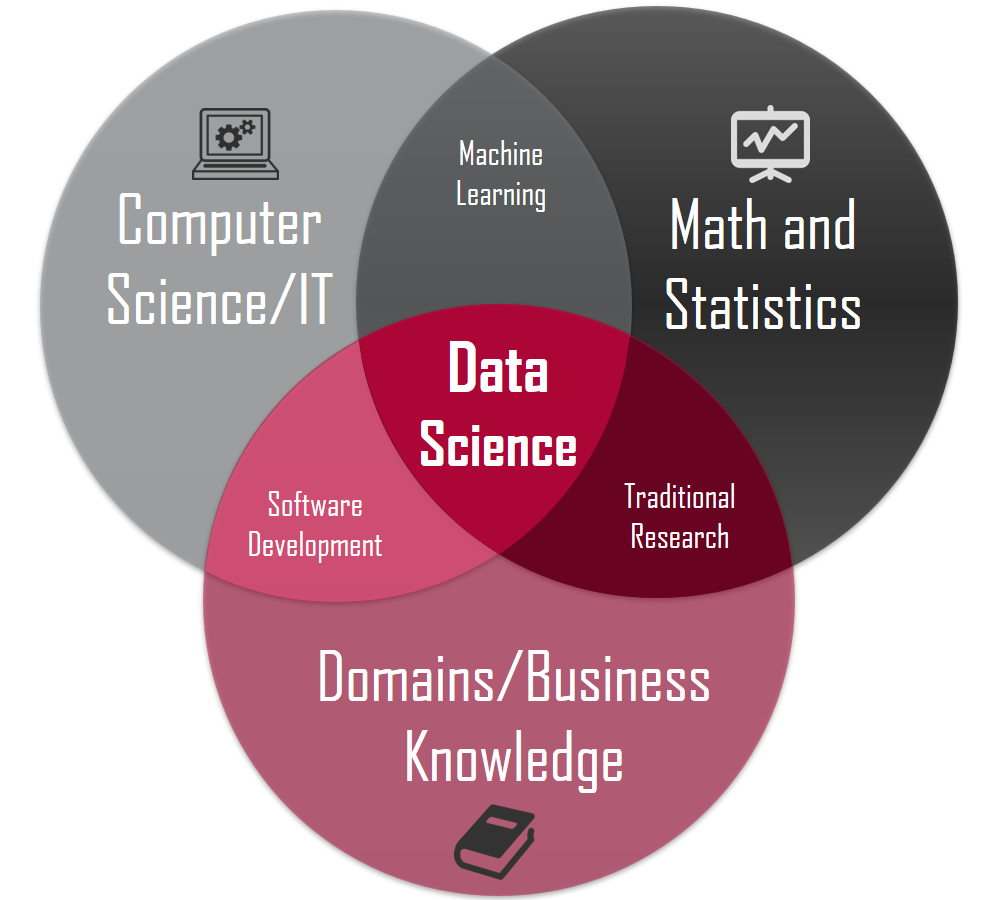
Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyse actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of [mathematics](https://en.wikipedia.org/wiki/Mathematics), [statistics](https://en.wikipedia.org/wiki/Statistics), [information science](https://en.wikipedia.org/wiki/Information_science), and [computer science](https://en.wikipedia.org/wiki/Computer_science).

[Turing award](https://en.wikipedia.org/wiki/Turing_award) winner [Jim Gray](https://en.wikipedia.org/wiki/Jim_Gray_%28computer_scientist%29) imagined data science as a "fourth paradigm" of science ([empirical](https://en.wikipedia.org/wiki/Empirical_research), [theoretical](https://en.wikipedia.org/wiki/Basic_research), computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the [data deluge](https://en.wikipedia.org/wiki/Information_explosion).

**HISTORY**

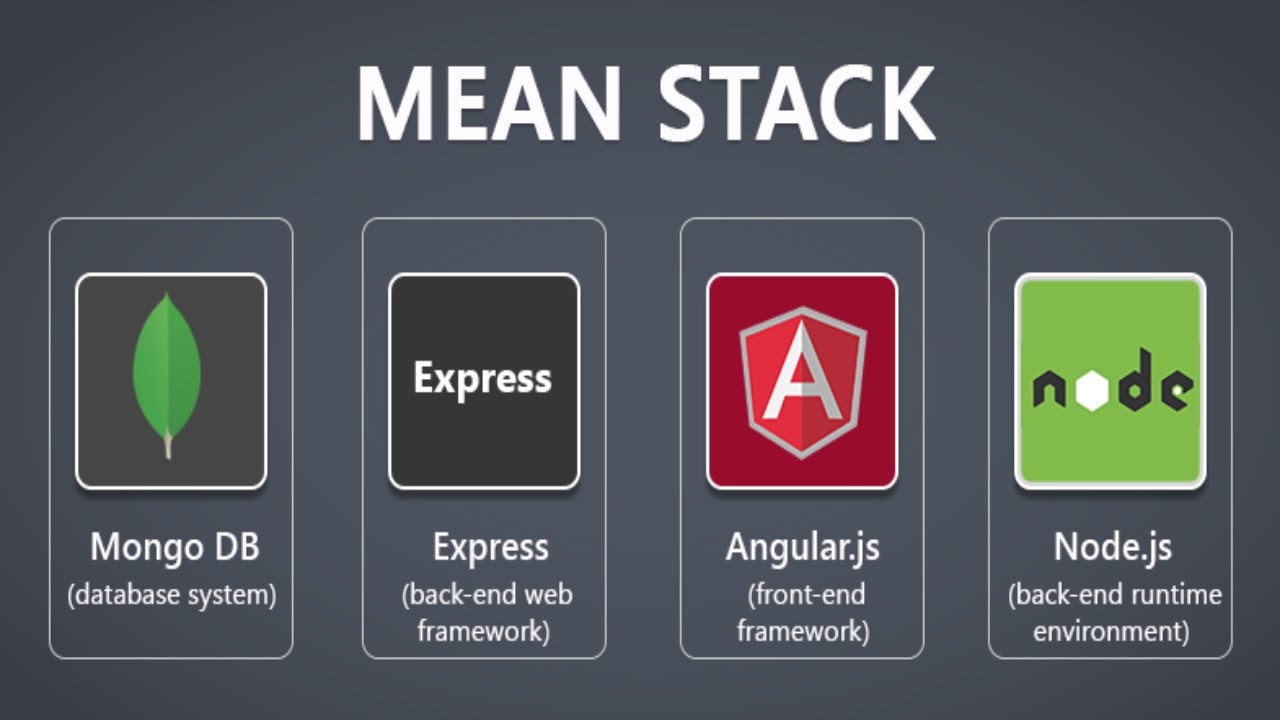
The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage it was used as a substitute for [computer science](https://en.wikipedia.org/wiki/Computer_science) by [Peter Naur](https://en.wikipedia.org/wiki/Peter_Naur) in 1960. Naur later introduced the term "[datalogy](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Statistics). Writing in [Forbes](https://en.wikipedia.org/wiki/Forbes), Gil Press argues that data science is a [buzzword](https://en.wikipedia.org/wiki/Buzzword) without a clear definition and has simply replaced “[business analytics](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/American_Statistical_Association), noted applied statistician [Nate Silver](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Software_Developer) and big data [engineers](https://en.wikipedia.org/wiki/Software_engineer). On the other hand, responses to criticism are as numerous. In a 2014 [Wall Street Journal](https://en.wikipedia.org/wiki/The_Wall_Street_Journal) article, Irving Wladawsky-Berger compares the data science enthusiasm with the dawn of [computer science](https://en.wikipedia.org/wiki/Computer_science). He argues data science, like any other [interdisciplinary](https://en.wikipedia.org/wiki/Interdisciplinarity) field, employs [methodologies](https://en.wikipedia.org/wiki/Methodology) and practices from across the [academia](https://en.wikipedia.org/wiki/Academy) and [industry](https://en.wikipedia.org/wiki/Industry), but then it will morph them into a new [discipline](https://en.wikipedia.org/wiki/Discipline_%28academia%29). He brings to attention the sharp criticisms computer science, now a well-respected academic discipline, had to once face. Likewise, [NYU](https://en.wikipedia.org/wiki/New_York_University) [Stern's](https://en.wikipedia.org/wiki/NYU_Stern_Center_for_Business_and_Human_Rights) 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](https://en.wikipedia.org/wiki/Discipline_%28academia%29), which focuses only on explaining [data sets](https://en.wikipedia.org/wiki/Data_set). Data science seeks actionable and consistent [pattern](https://en.wikipedia.org/wiki/Pattern_recognition) for [predictive uses](https://en.wikipedia.org/wiki/Predictive_modelling). This practical engineering goal takes data science beyond traditional [analytics](https://en.wikipedia.org/wiki/Analytics). Now the data in those disciplines and [applied fields](https://en.wikipedia.org/wiki/Applied_science) that lacked solid [theories](https://en.wikipedia.org/wiki/Theory), like [health science](https://en.wikipedia.org/wiki/Health_science) and [social science](https://en.wikipedia.org/wiki/Social_science), could be sought and utilized to generate powerful predictive models.

**MEAN STACK**



**MEAN** is a [free and open-source](https://en.wikipedia.org/wiki/Free_and_open-source_software) [JavaScript](https://en.wikipedia.org/wiki/JavaScript) [software stack](https://en.wikipedia.org/wiki/Software_stack) for building [dynamic web sites](https://en.wikipedia.org/wiki/Dynamic_web_page) and [web applications](https://en.wikipedia.org/wiki/Web_application).

The MEAN stack is [MongoDB](https://en.wikipedia.org/wiki/MongoDB), [Express.js](https://en.wikipedia.org/wiki/Express.js), [AngularJS](https://en.wikipedia.org/wiki/AngularJS) (or [Angular](https://en.wikipedia.org/wiki/Angular_%28application_platform%29)), and [Node.js](https://en.wikipedia.org/wiki/Node.js). Because all components of the MEAN stack support programs are written in JavaScript, MEAN applications can be written in one language for both [server-side](https://en.wikipedia.org/wiki/Server-side) and [client-side](https://en.wikipedia.org/wiki/Client-side) execution environments.

**SYSTEM** **REQUIREMENT** **AND** **SPECIFICATION**

**2.1** **Hardware** **Requirements:**

* Intel® Pentium 4 CPU and higher versions
* 256 MB RAM,
* 80GB HDD Mouse
* QWERTY Keyboard
* Standard VGA Monitor

**2.2** **Software** **Requirements:**

* Operating System : WINDOWS 10
* Language
* Tool

: PYTHON

: 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

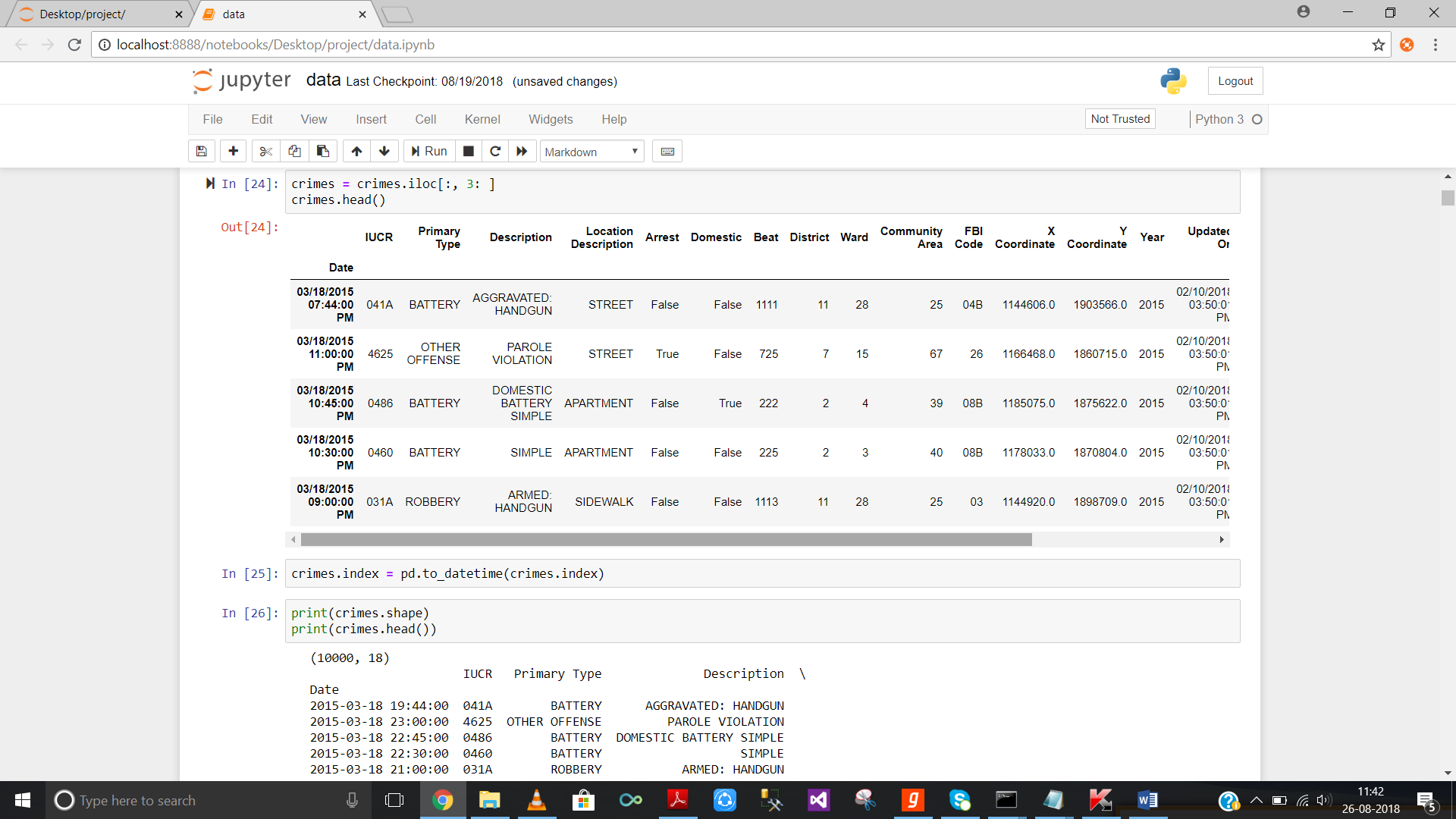
domestic\_yearly.resample('D').sum().plot()

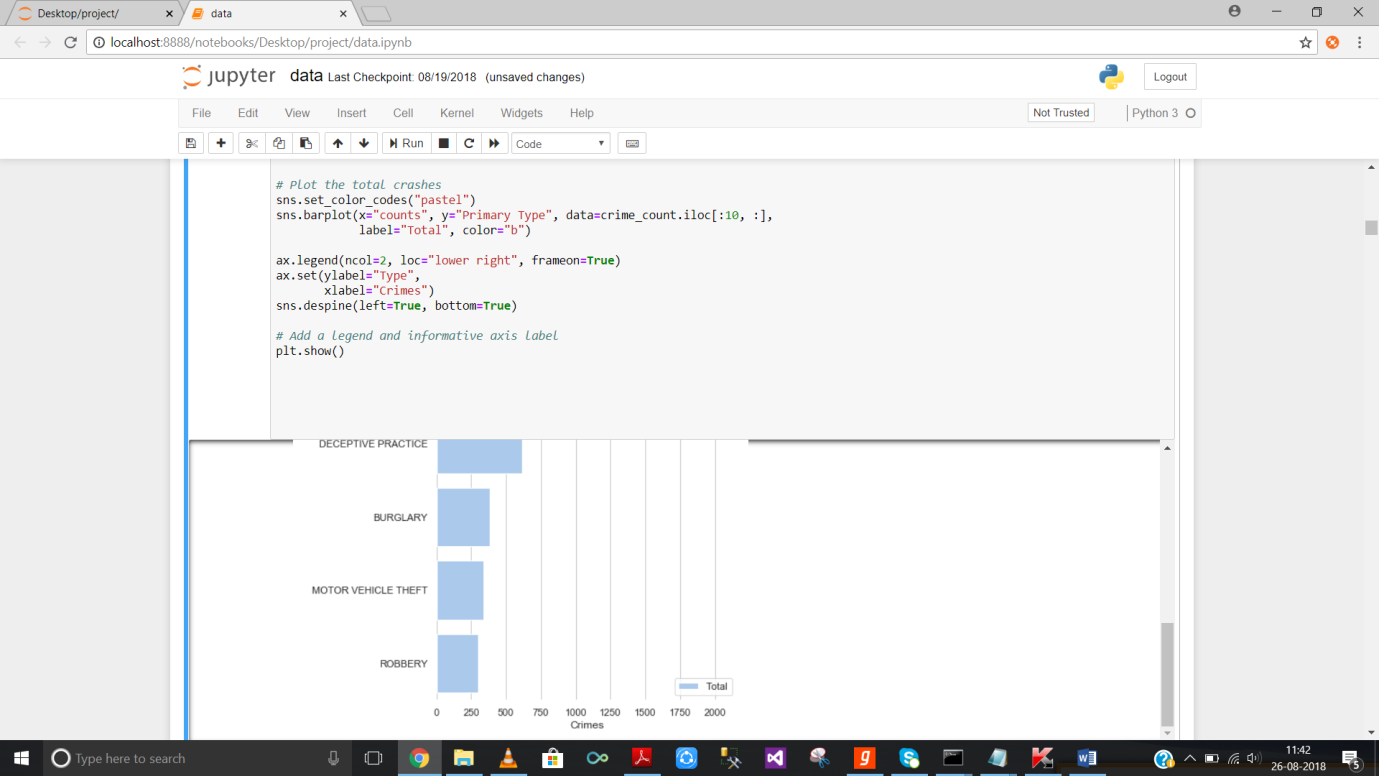
plt.title('Daily domestic violence')

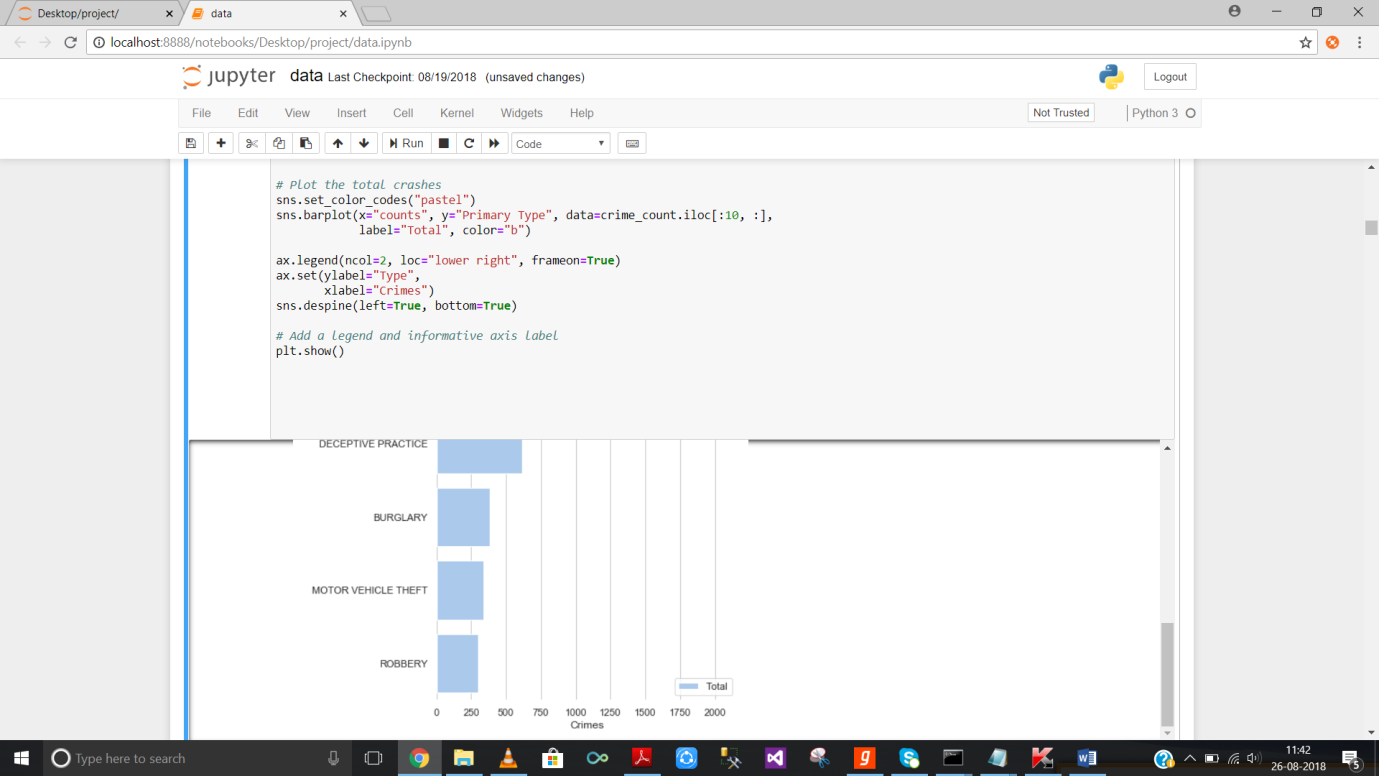
plt.show()

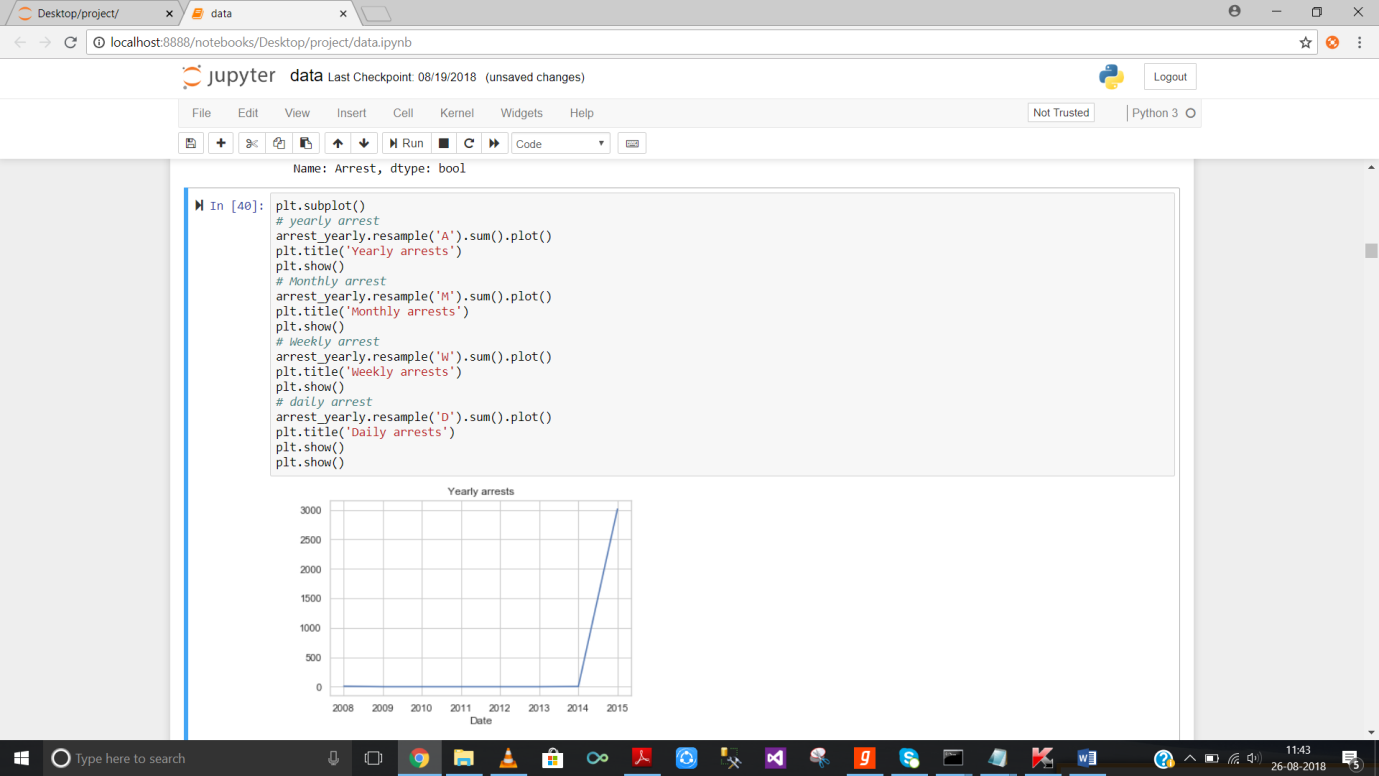
plt.show()

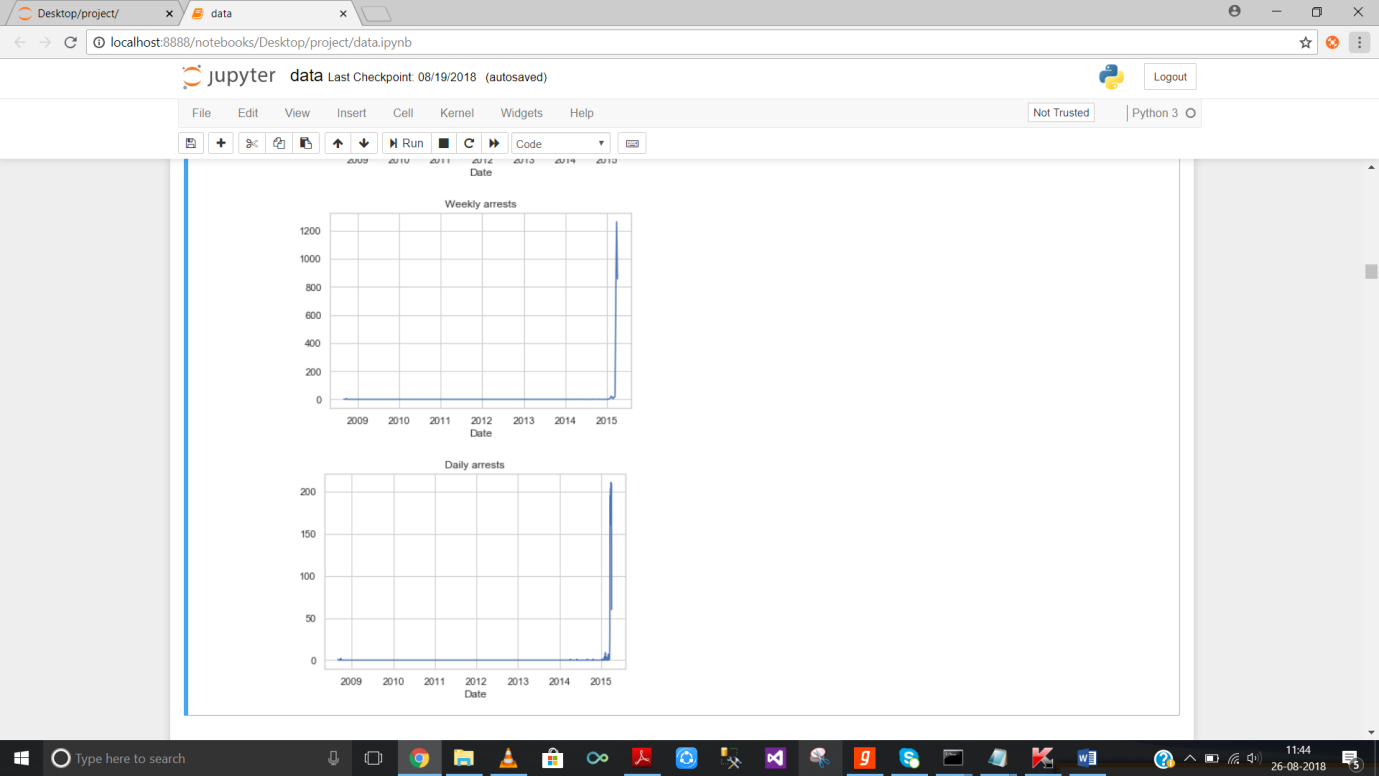
**SNAPSHOTS**

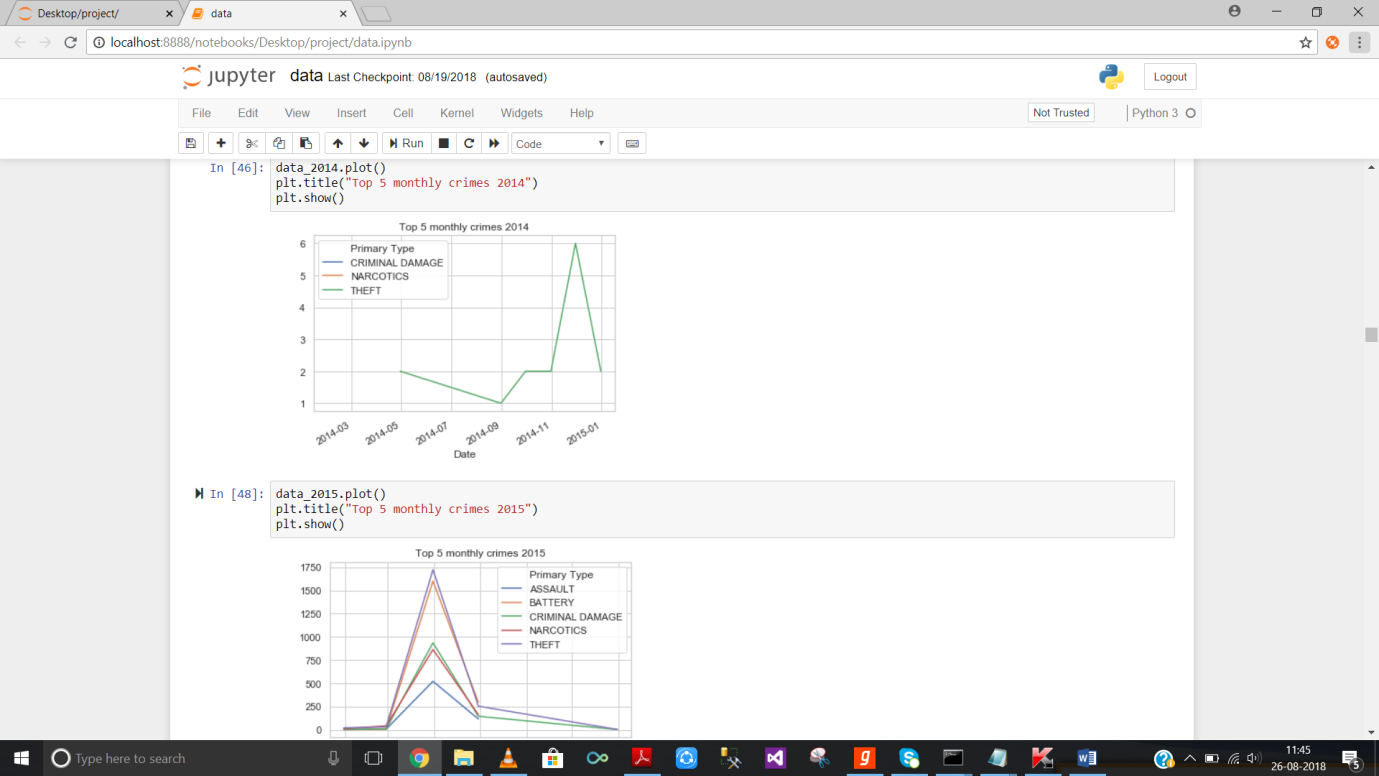


****

****

****



****

**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