Exploratory Data Analysis and Visualization on Crime Data Using PySpark

Final Project

Addis Yesserie(addisy1@umbc.edu)
MPS Data Science, UMBC

Data603 Platforms for Big Data Processing

Spring 2022

Outline

- Introduction
- Objectives
- Datasets
- Data Preprocessing and Processing
- EDA
- Conclusions
- Limitations
- References

Introduction

- Big Data Analytics (BDA) can effectively address the challenges of data
  - too vast, too unstructured, and too fast-moving to be managed by traditional methods
- BDA can aid organizations to utilize their data and facilitate new opportunities

Introduction

- BDA has become an emerging approach for:
  - Analyzing data
  - Extracting information
  - Relations in a wide range of application areas
- BDA is a systematic approach for
  - analyzing and identifying
    - Patterns
    - Relations
    - Trends within a large volume of data

Objectives

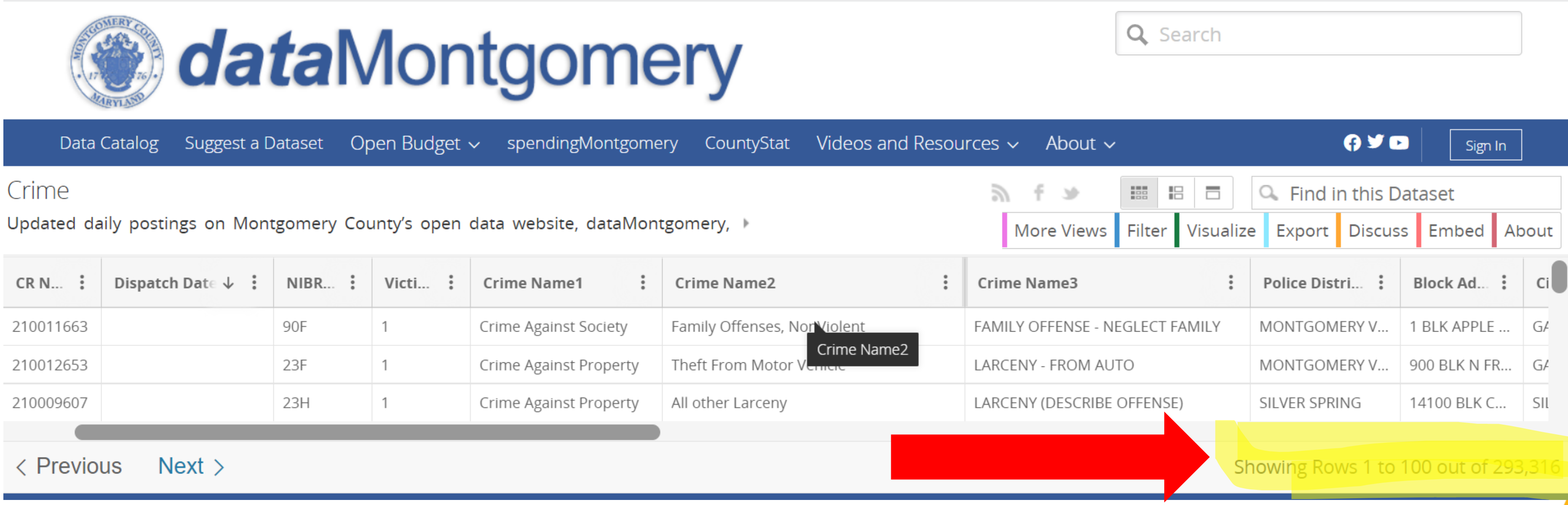
- To explore, analyze, visualize crime incidences in Montgomery County, Maryland

- The use of Big Data Analytics on this crime incident and pattern analysis will enable for hotspot detection and predictive policing

Datasets

- Publicly available datasets that consist of crime activities in Montgomery County, MD

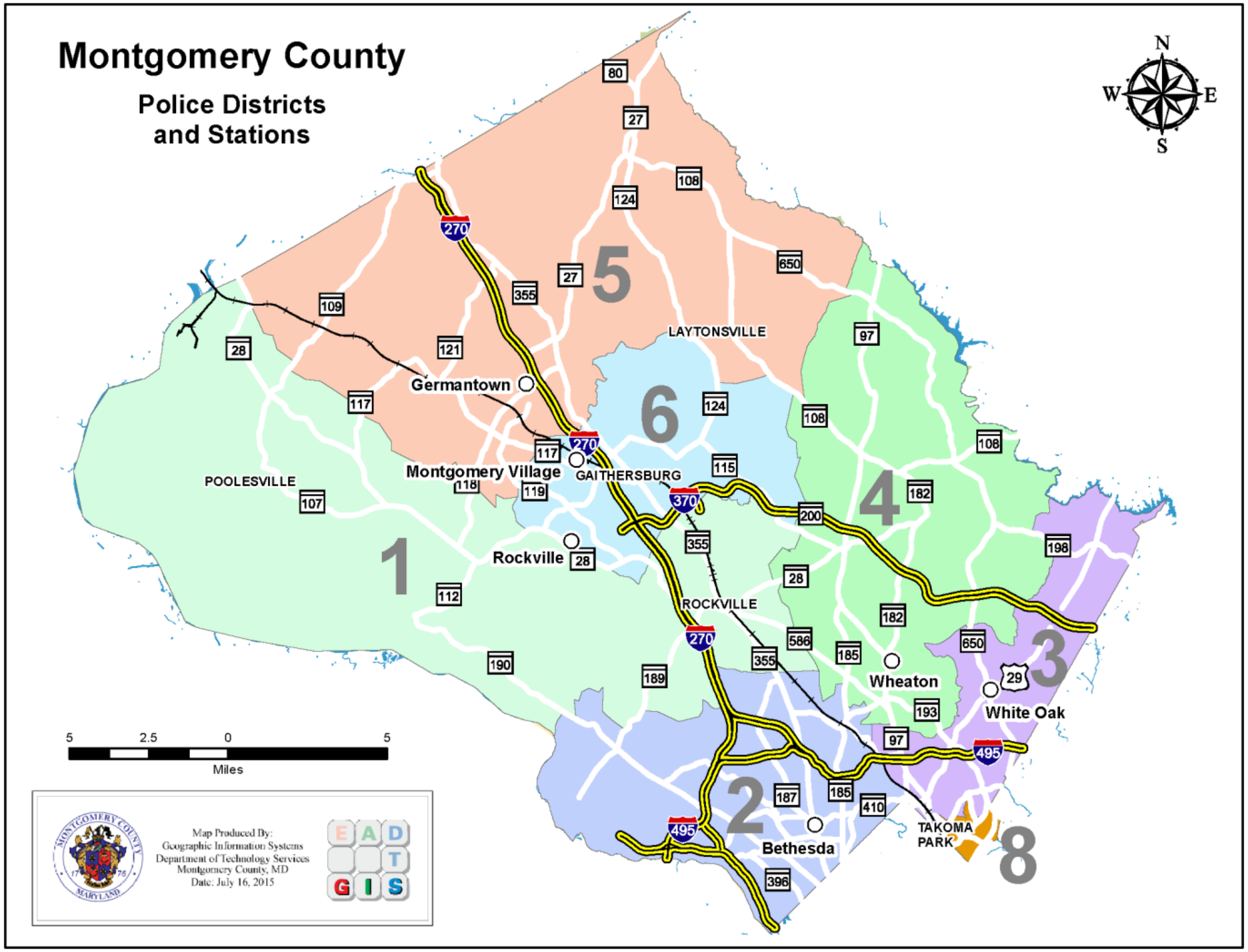
Crime Data https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3/data



Appendix – Overview of Data Included in Crime Dataset

Display Order	Column	Field Description
1	Incident ID	Police Incident Number
2	CR Number	Police Report Number
3	Dispatch Date / Time	The actual date and time a Officer was dispatched
4	Class	Four-digit code identifying the crime type of the incident
5	Class Description	Common name description of the incident class type
6	Police District Name	Name of District (Rockville, Wheaton etc.)
7	Block Address	Address in 100 block level
8	City	City
9	State	State
10	Zip Code	Zip code
11	Agency	Assigned Police Department
12	Place	Place description
13	Police Sector	Police Sector Name
14	Beat	Police patrol area subset within District
15	PRA	Police patrol are subset within Beat
16	Start Date / Time	Occurred from date/time
17	End Date / Time	Occurred to date/time
18	Latitude	Latitude
19	Longitude	Longitude
20	Police District Number	Major Police Boundary
21	Location	Location

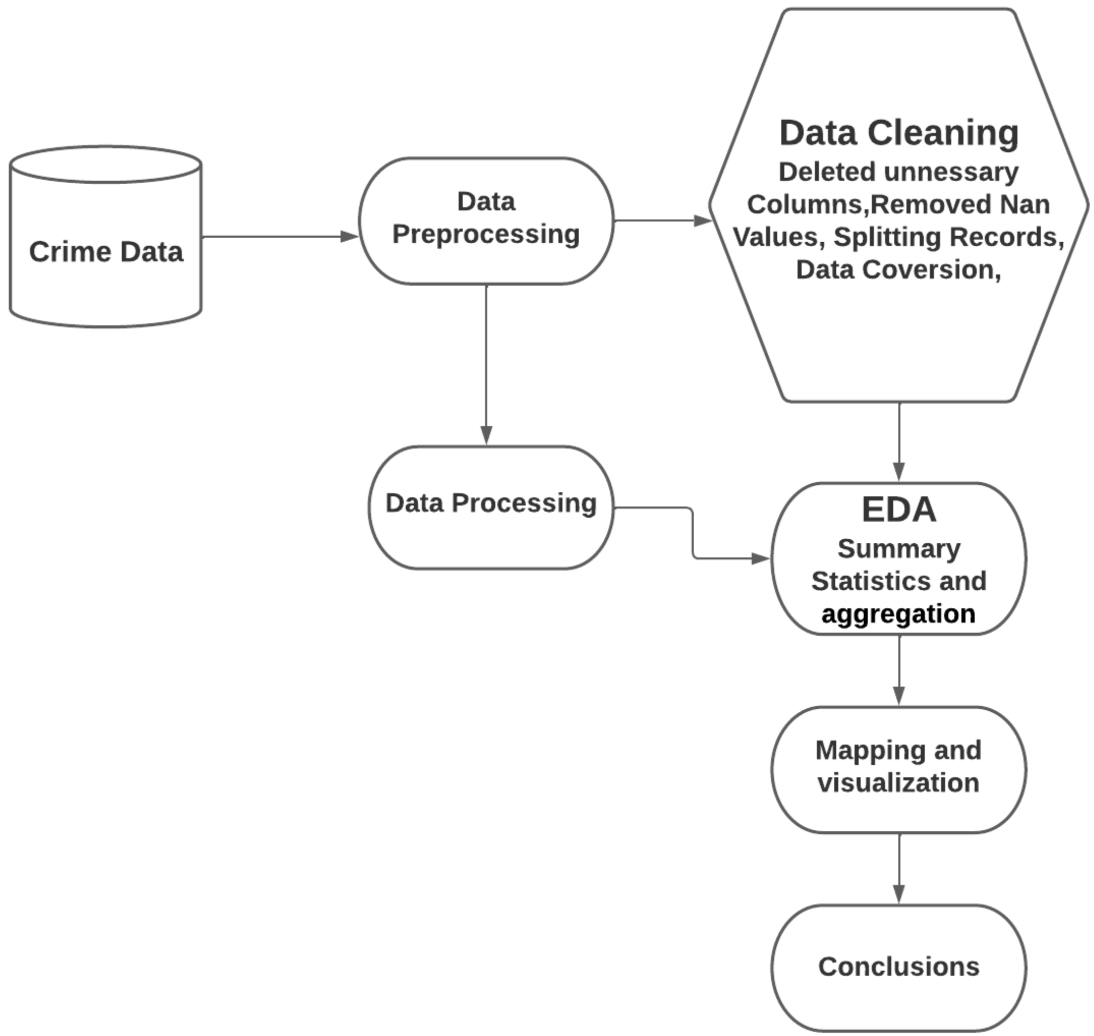
Study Area – Montgomery County Police Districts and Stations



Methods

- After acquiring the data, big data processing python package including :
  - PySpark were used to perform data reading, transforming, querying and analysis
  - Folium maps implemented to visualize crime data

Methodology



Data Preprocessing and Processing

```
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
from pyspark.sql import*
import pyspark
import datetime
from datetime import datetime
from pyspark.sql.functions import unix_timestamp, from_unixtime
from pyspark.sql.functions import year, month, dayofmonth, dayofweek, hour
from pyspark.sql.functions import when, count, col, countDistinct, desc, first, lit
from pyspark.sql.functions import split
import pandas as pd
import matplotlib.pyplot as plt
import folium
from folium import plugins
from folium.plugins import MarkerCluster
from folium.plugins import FastMarkerCluster
```

Python Libraries/ Functions

Schema

```
root
 |-- Incident ID: integer (nullable = true)
 |-- Offence Code: string (nullable = true)
 |-- CR Number: integer (nullable = true)
 |-- Dispatch Date / Time: timestamp (nullable = true)
 |-- NIBRS Code: string (nullable = true)
 |-- Victims: integer (nullable = true)
 |-- Crime Name1: string (nullable = true)
 |-- Crime Name2: string (nullable = true)
 |-- Crime Name3: string (nullable = true)
 |-- Police District Name: string (nullable = true)
 |-- Block Address: string (nullable = true)
 |-- City: string (nullable = true)
 |-- State: string (nullable = true)
 |-- Zip Code: integer (nullable = true)
 |-- Agency: string (nullable = true)
 |-- Place: string (nullable = true)
 |-- Sector: string (nullable = true)
 |-- Beat: string (nullable = true)
 |-- PRA: string (nullable = true)
 |-- Address Number: integer (nullable = true)
 |-- Street Prefix: string (nullable = true)
 |-- Street Name: string (nullable = true)
 |-- Street Suffix: string (nullable = true)
 |-- Street Type: string (nullable = true)
 |-- Start_Date_Time: timestamp (nullable = true)
 |-- End_Date_Time: timestamp (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- Police District Number: string (nullable = true)
 |-- Location: string (nullable = true)
```

EDA: Crime Analysis and Visualization

Count of Crime Categories

Crime_Category	Crime_CategoryCount
LARCENY	57723
ASSAULT	23081
POLICE INFORMATION	16378
DRUGS	16202
DAMAGE PROPERTY	13413
LARCENY (DESCRIBE...	11320
MENTAL ILLNESS	10635
FRAUD	9863
LOST PROPERTY	9817
IDENTITY THEFT	8131
BURGLARY	7724
DRIVING UNDER THE...	6664
SUDDEN DEATH	5618
AUTO THEFT	5579
PUBLIC PEACE	5563
MISSING PERSON	4340
LIQUOR	4226
TRESPASSING	3731
DAMAGE PROPERTY (...	3563
JUVENILE	3549

only showing top 20 rows

Crime_Against	Crime_AgainstCount
Crime Against Pro...	127659
Other	56799
Crime Against Soc...	48045
Crime Against Person	27212
Not a Crime	3360

Count of Crime Against

Districts with most Crime

Police_District	count
SILVER SPRING	54644
WHEATON	49860
MONTGOMERY VILLAGE	44434
BETHESDA	37598
ROCKVILLE	36614
GERMANTOWN	34061
CITY OF TAKOMA PARK	4959
TAKOMA PARK	826

Which City had the most Crime Incidents?

City_Name	count
SILVER SPRING	90343
GAITHERSBURG	37677
ROCKVILLE	37023
GERMANTOWN	25887
BETHESDA	18816
MONTGOMERY VILLAGE	8297
TAKOMA PARK	7035
POTOMAC	5694
CHEVY CHASE	5561
DERWOOD	4860
KENSINGTON	4167
OLNEY	4127
BURTONSVILLE	3236
CLARKSBURG	2910
DAMASCUS	2205
BOYDS	1843
BROOKEVILLE	817
POOLESVILLE	793
ASHTON	375
SANDY SPRING	363

Place of Incidence

Place1	Place1Count
Residence	92848
Street	43700
Parking Lot	28300
Other/Unknown	23307
Retail	22431
School/College	6300
Parking Garage	5496
Restaurant	5108
Grocery/Supermarket	4874
Commercial	3485
Convenience Store	3006
Government Building	2923
Hotel/Motel/Etc.	2351
Hospital/Emergenc...	1970
Gas Station	1892
Park	1417
School	1299
Bar/Night Club	1215
Bank	1168
Bank/S&L/Credit U...	1005

only showing top 20 rows

Top 3 place where Crime Incidences mostly happened

Top 15 Block Addresses with most Crime

Block_Address	Block_AddressCount
11100 BLK VEIRS ...	3881
20900 BLK FREDER...	2733
7100 BLK DEMOCRA...	1815
100 BLK EDISON P...	1622
700 BLK RUSSELL AVE	1483
7300 BLK CALHOUN PL	1438
8100 BLK GEORGIA...	1002
8600 BLK COLESVI...	994
11200 BLK NEW HA...	881
20000 BLK AIRCRA...	822
11200 BLK GEORGI...	803
12000 BLK CHERRY...	750
8200 BLK GEORGIA...	694
1000 BLK MILESTO...	679
22700 BLK CLARKS...	666

Which year had the most crimes?

year(Date)	yearcount
2017	50383
2018	47279
2019	45358
2020	41064
2021	39949

only showing top 5 rows

Which Month had the most crimes?

month(Date)	monthCount
10	24127
7	23274
8	23267
9	23198
3	22887
12	22712
11	22517
1	22050
2	20981
4	20569
5	19355
6	18138

UMBC

Which Day had the most crimes?

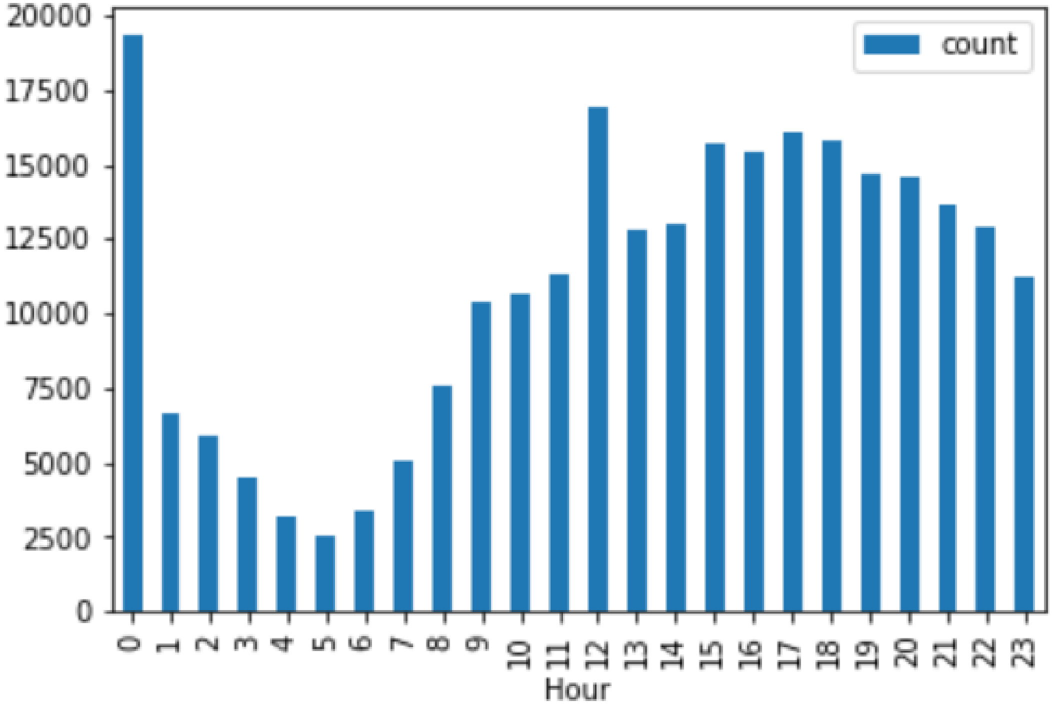
Day	count	
6	41111	Friday
4	39341	
3	38874	
5	38830	
2	37640	
7	34883	
1	32396	Sunday

What time of the day most crimes occurred?

hour(Date)	hourCount	
0	19348	Midnight
12	16897	
17	16045	
18	15776	
15	15713	
16	15381	
19	14632	
20	14585	
21	13672	
14	13037	
22	12866	
13	12775	
11	11324	
23	11248	
10	10656	
9	10375	
8	7566	Mornings
1	6613	
2	5912	
7	5027	

only showing top 20 rows

What time of the day most crimes occurred?



hour(Date)	hourCount	
0	19348	Midnight
12	16897	
17	16045	
18	15776	
15	15713	
16	15381	
19	14632	
20	14585	
21	13672	
14	13037	
22	12866	
13	12775	
11	11324	
23	11248	
10	10656	
9	10375	
8	7566	Mornings
1	6613	
2	5912	
7	5027	

only showing top 20 rows

Before Covid-19 (2017)

Place1	Place1Count
Residence	16689
Street	9862
Parking Lot	5370
Retail	4302
Other/Unknown	3745
School/College	1644
Restaurant	1160
Parking Garage	1129
Grocery/Supermarket	871
Commercial	683

only showing top 10 rows

During Covid-19 (2020)

Place1	Place1Count
Residence	15905
Street	5477
Parking Lot	4807
Other/Unknown	4721
Retail	3104
Grocery/Supermarket	837
Parking Garage	789
School/College	698
Restaurant	597
Convenience Store	497

only showing top 10 rows

Visualizing Places of Crime Incidences before and during Covid-19

Before Covid-19 (2017)

Crime_Category	Crime_CategoryCount
LARCENY	10499
DRUGS	4401
ASSAULT	4000
POLICE INFORMATION	2732
DAMAGE PROPERTY	2532
LARCENY (DESCRIBE...	2294
FRAUD	1941
LOST PROPERTY	1805
DRIVING UNDER THE...	1748
BURGLARY	1521
PUBLIC PEACE	1354
MENTAL ILLNESS	1256
MENTAL ILLNESS	1194
LIQUOR	1127
TRESPASSING	888
SUDDEN DEATH	855
AUTO THEFT	845
MISSING PERSON	795
DAMAGE PROPERTY (...	776
IDENTITY THEFT	748

only showing top 20 rows

During Covid-19 (2020)

Crime_Category	Crime_CategoryCount
LARCENY	10037
ASSAULT	3573
POLICE INFORMATION	2852
DAMAGE PROPERTY	2443
MENTAL ILLNESS	2002
FRAUD	1931
LARCENY (DESCRIBE...	1897
LOST PROPERTY	1507
IDENTITY THEFT	1442
DRUGS	1347
BURGLARY	1208
SUDDEN DEATH	1146
AUTO THEFT	1019
PUBLIC PEACE	663
MISSING PERSON	650
DRIVING UNDER THE...	558
JUVENILE	539
RECOVERED PROPERTY	524
TRESPASSING	498
DAMAGE PROPERTY (...	496

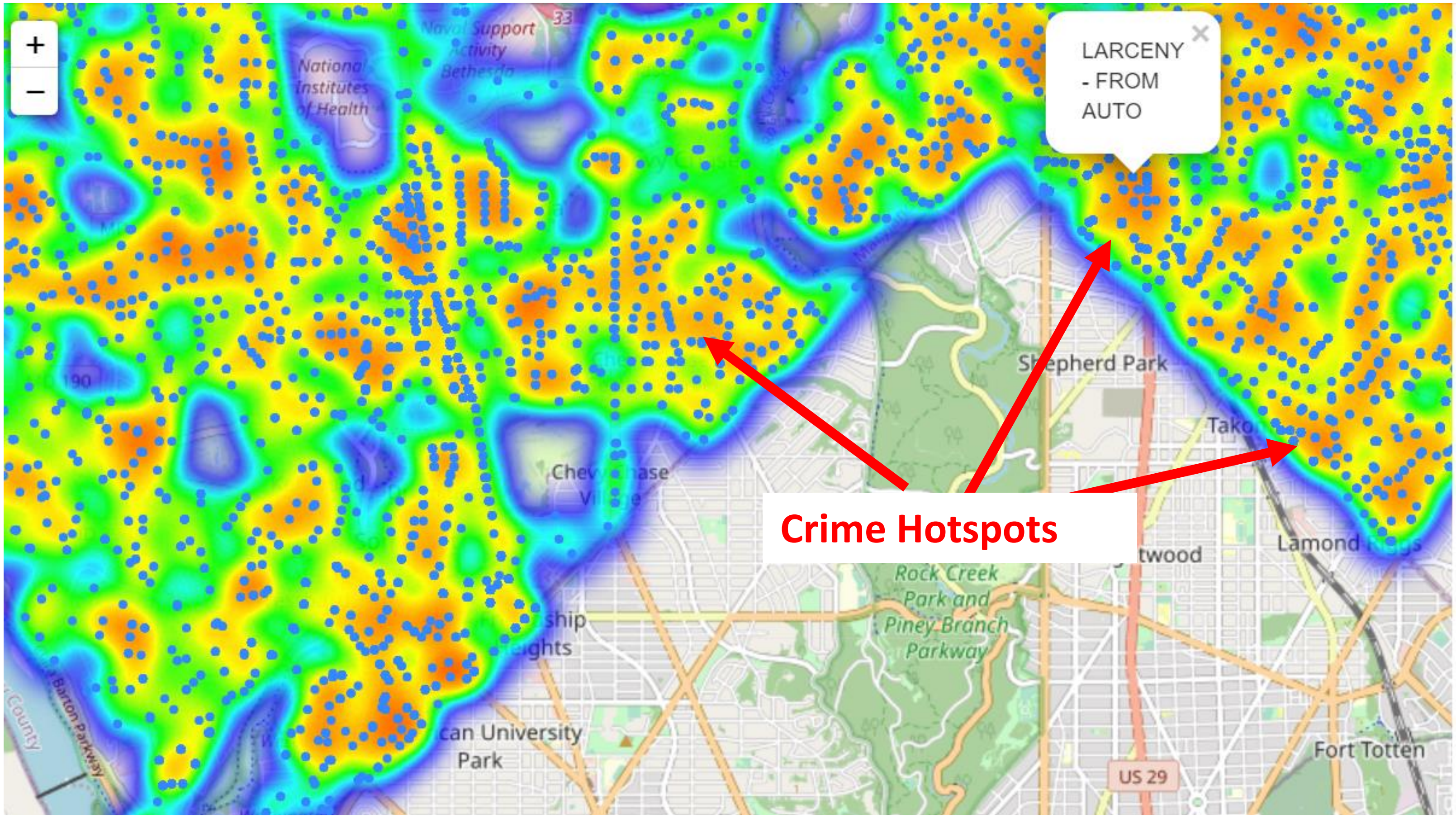
only showing top 20 rows

Visualizing Crime Types before and during Covid-19

Observations

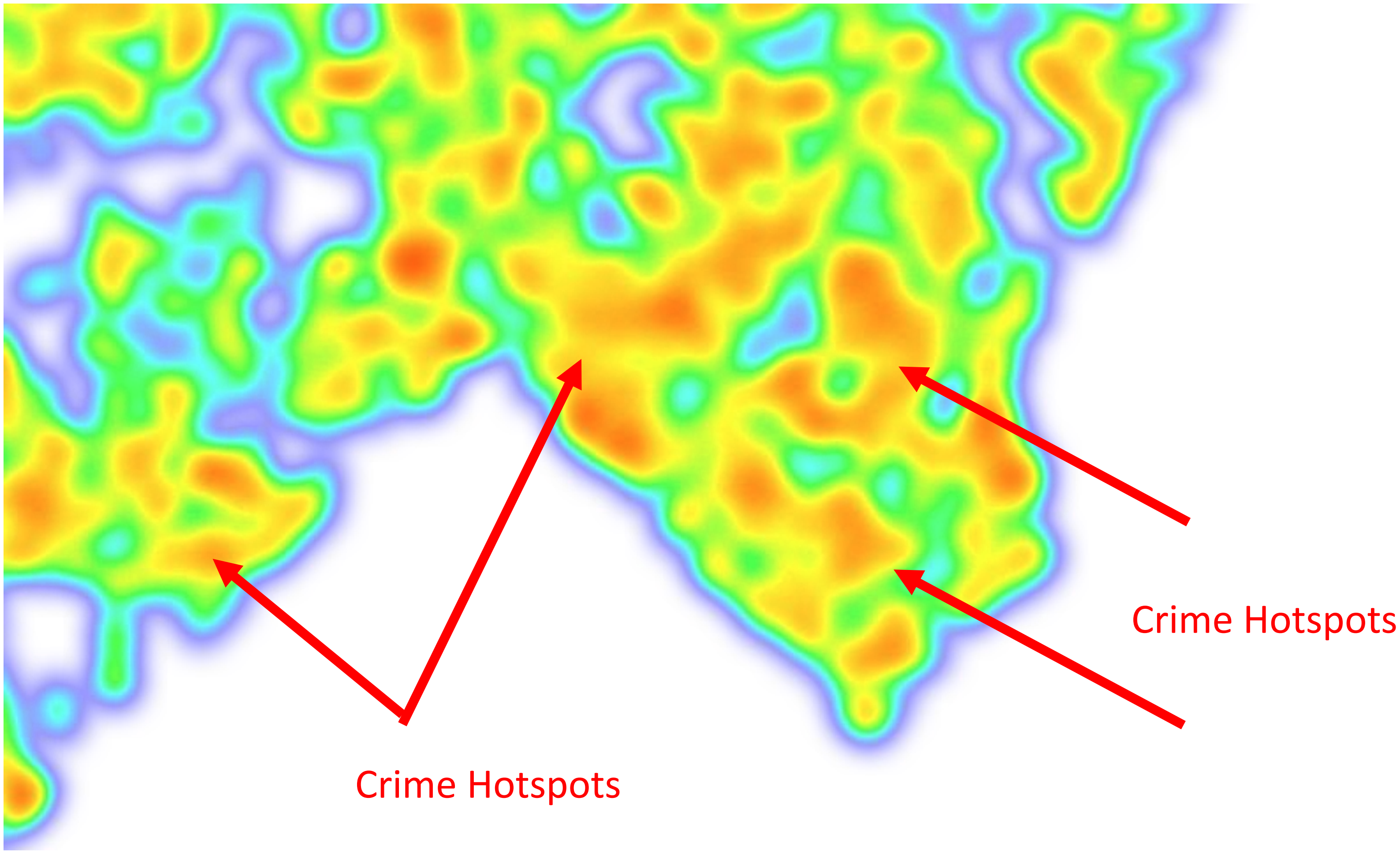
- Places and magnitude of crime has changed mainly the following:
  - Drug related crimes has declined during the Covid-19 periods
  - Crime at school/College areas has sharply declined as most students were on a virtual learning programs
  - Identity Theft has dramatically increased
  - Mental Illness has shown an increase in total
  - Sudden death has increased
  - Liquor related crimes has declined

Geospatial Visualization of Crime Patterns - Folium Maps



From the folium map, we can see, each member in one cluster are close to each other based on the same color is near to each other and clearly show Hotspots of crime Locations

Folium Maps – Crime Hotspots



UMBC

Conclusions

- Comparing crime patterns before and during Covid-19 clearly indicates an overall decline in the crime incidences
- The absence of clear crime pattern created a challenge from achieving a clearly defined clusters
- Folium Mapping was found useful tool and easy to visualize data that has been manipulated in Python on an interactive leaflet map

Limitations

- The organization and complexity of the Crime data
- Many missing Data and Null Values had impacted the analysis to certain extent
- The large dataset has hindered to run Folium maps in Docker

???

Thank you!