Big Echo

Udit Gupta | Ahalya Sugumar | Nagashree Angadi
Department of Computer Science
Stony Brook University

SDG - Peace, Justice and Strong Institutions

Analyzing NYC Crime Data

INTRODUCTION

Project Proposal

- To analyze the Historic NYPD crime data in order to identify trends phenomenon
- Helps us in making decisions and efficiently targeting resources employing mapping and spatial analysis.

Motivation

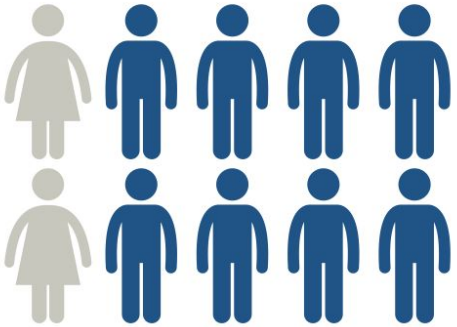
Information on patterns can help law enforcement agencies in following ways,

- deploy resources in a most effective manner
- assist agencies in identifying and apprehending suspects.

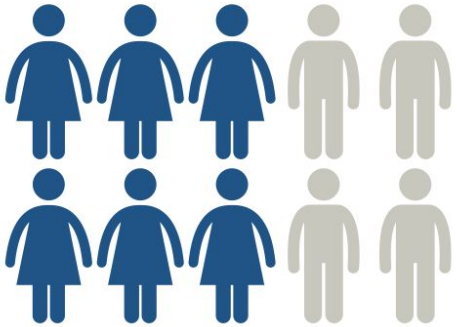
SDG - Peace, Justice and Strong Institutions

The analysis will help institutions to build effective, accountable, and inclusive institutions at all levels, thus promoting peaceful societies for sustainable development.

WHY SHOULD WE CARE



MEN
MAKE UP AROUND
80%
OF HOMICIDE VICTIMS
OVERALL

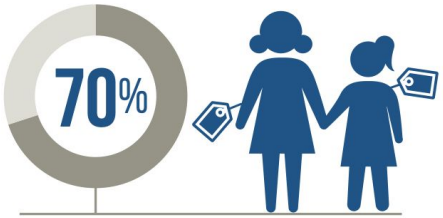


BUT WOMEN
CONSTITUTE
64%
OF HOMICIDE VICTIMS
OF INTIMATE PARTNER/
FAMILY-RELATED HOMICIDE

Violent Crime
13% of people have experienced **violent crime** in the last 12 months

Property Crime
14% of people have experienced **property crime** in the last 12 months

Mass Shootings
8% of people have been affected by a **mass shooting**



OF DETECTED VICTIMS OF
HUMAN TRAFFICKING ARE
WOMEN AND GIRLS

https://www.un.org/sustainabledevelopment/peace-justice/

BACKGROUND

01

Crime Risk Maps: A Multivariate Spatial Analysis of Crime Data

A generalized multivariate conditional autoregressive model to model the spatial dependence between sites, and the dependence between multiple crime types.

02

New York City Police Department Burrough and Precinct level Statistics

Crime Prevention data-driven performance measure taken by NYPD. Data and statistical reports of crime trends and traffic data is updated regularly

03

The Effects of Local Police Surges on Crime and Arrests in New York City

Poisson regression models to compare differences in crime and arrest counts before and after census block groups.

04

An Analysis of the New York City Police Department's Stop-and-Frisk Policy in the Context of Claims of Racial Bias

Police stop persons of racial and ethnic minority groups more often than whites relative to their proportions in the population

DATA

01

NYPD COMPLAINTS DATA

- JSON Format
- Borough, Precinct Level
- Victim and Suspect Information
- From 2006 to 2019

02

NYPD ARRESTS DATA

- JSON Format
- Borough, Precinct Level
- Suspect Information
- From 2006 to 2019

12.8 million rows
35 columns
11.2 GB volume

03

NYC DEMOGRAPHICS DATA

- CSV Format
- Borough Level
- Sub Borough Level
- From 2000 to 2018
- Poverty rate, Population, Racial diversity, Unemployment rate & Income diversity

04

SHAPE FILES

- GeoJSON Format
- Borough
- Sub Borough

TOOLS

GCP - DataProc

BIG DATA TECHNOLOGY

- HDFS
- Apache Spark

DATA SCIENCE CONCEPTS

- Multivariate Linear Regression
- Hypothesis Testing
- Stratified Sampling

Data Analysis Pipeline

1. **Data Collection** — collecting data from various courses
2. **Data Cleaning** — imputing null values, dropping unwanted columns
3. **Data Transformation** — adding custom variable, feature selection
4. **Data Aggregation** — merging all the datasets on relevant fields
5. **Distributed Storage** — storing data in Apache HDFS
6. **Exploratory Data Analysis** — performing initial investigation on data
7. **Linear Regression** — crime count vs socio-economic factors
8. **Hypothesis Testing** — find if the relation between crime count and other factors are significant
9. **Analyze Results** — analyze the results obtained from hypothesis testing

METHODS Overview

WHEN

Step 1: Temporal Analysis

- Year x Crime Type
- Month/Season x Crime Type
- Day of Month x Crime Type
- Hour of Day x Crime Type

WHERE

Step 2: Spatial Analysis

- Borough x Crime Type
- Precinct x Crime Type
- Premises x Crime Type

WHAT

Step 3: Information Related to Victims and Suspects

- Ethnicity of Suspect/Victims
- Correspondence Analysis
- Location Distribution

IMPACT

Step 4: Crime rate impacts Social and Economic factors

- Poverty Rate
- Racial Diversity
- Income Diversity
- Unemployment Rate

METHODS REGRESSION AND HYPOTHESIS TESTING

Linear Regression

The linear regression is used to **estimate** the relationship between crime count and demographic factors. The relationship can be controlled by population.

Hypothesis Testing

Analysing the **significance** of relationship between crime count and various demographic factors at various spatial levels.

Demographic factors considered,

- poverty rate
- unemployment rate
- racial diversity
- income diversity

METHODS REGRESSION AND HYPOTHESIS TESTING

Null Hypothesis 1: The demographic factors don't affect the crime count at borough level.

Null Hypothesis 2: The demographic factors don't affect the crime count controlled by population at borough level.

Null Hypothesis 3: The demographic factors don't affect the crime count at sub-borough level.

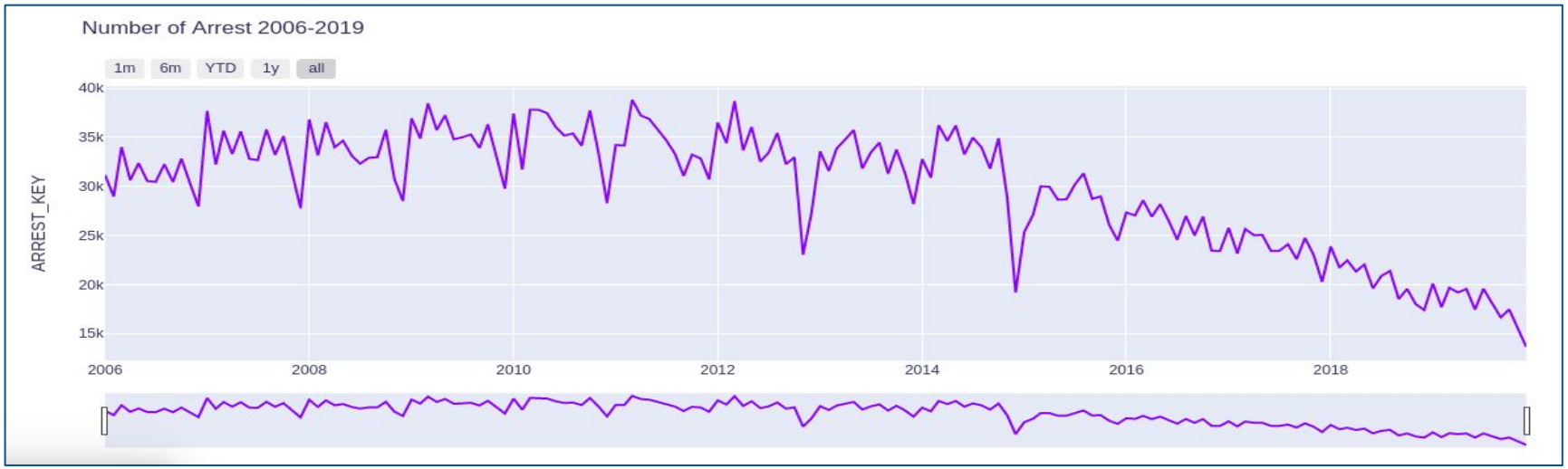
Null Hypothesis 4: The demographic factors don't affect the crime count controlled by population at sub-borough level.

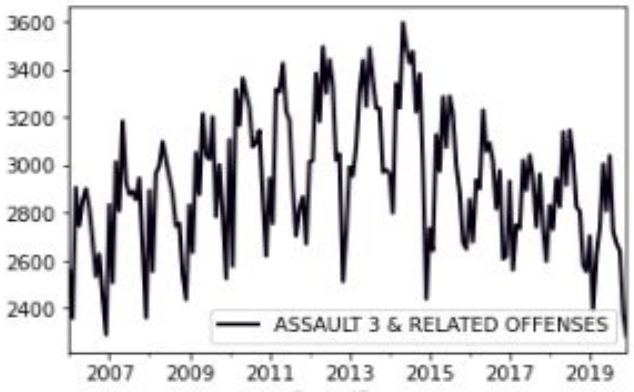
Level of Significance: 0.05

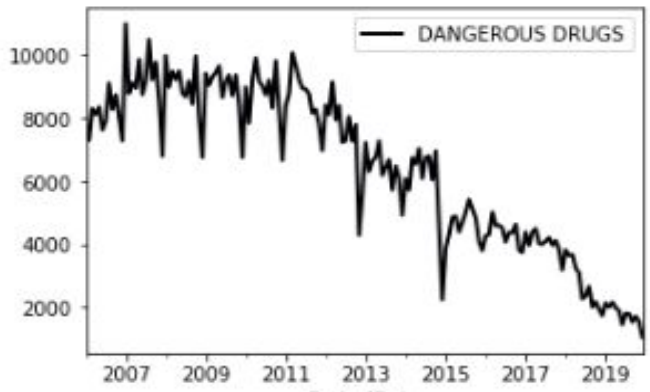
If p-value < 0.05 then
The null hypothesis is rejected
else
The null hypothesis is not rejected

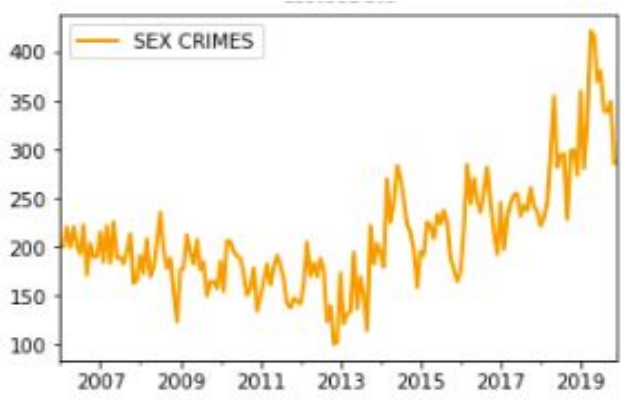
RESULTS Step 1 : Temporal Analysis

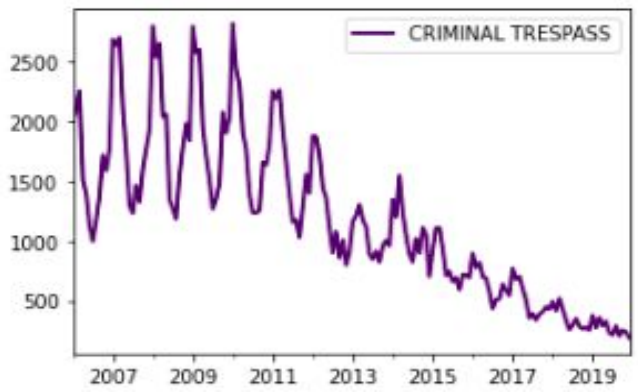
Crime Trends Over Past 10 Years

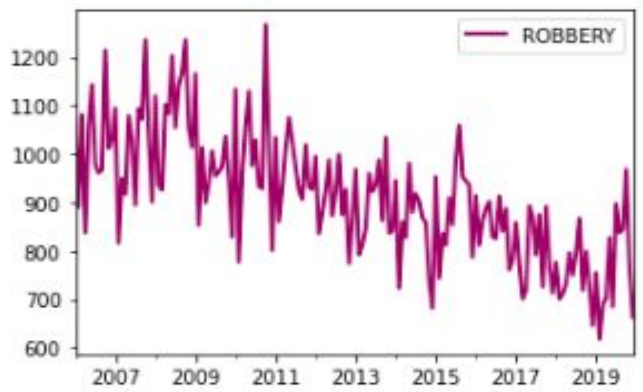






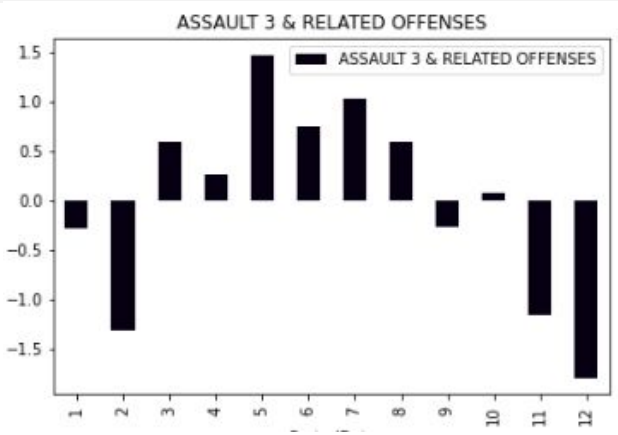


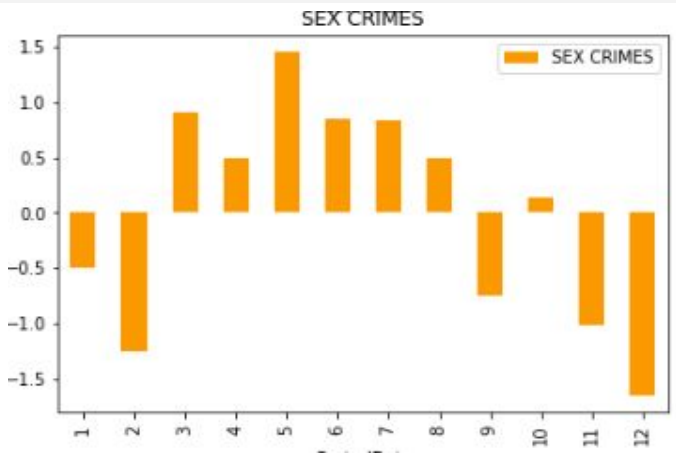


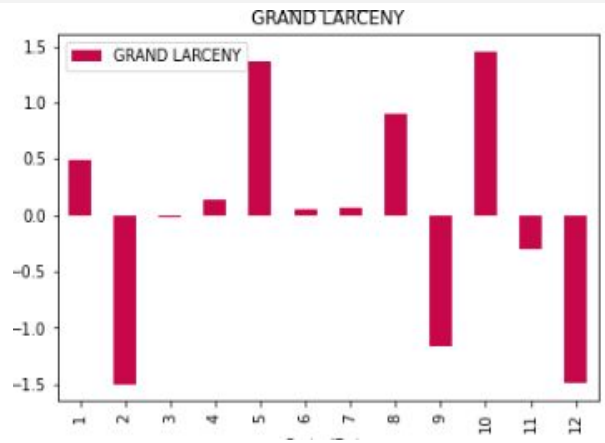


RESULTS Step 1 : Temporal Analysis

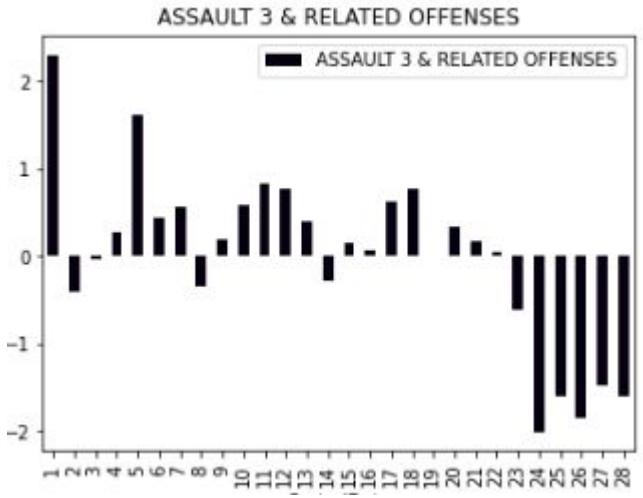
Month Level

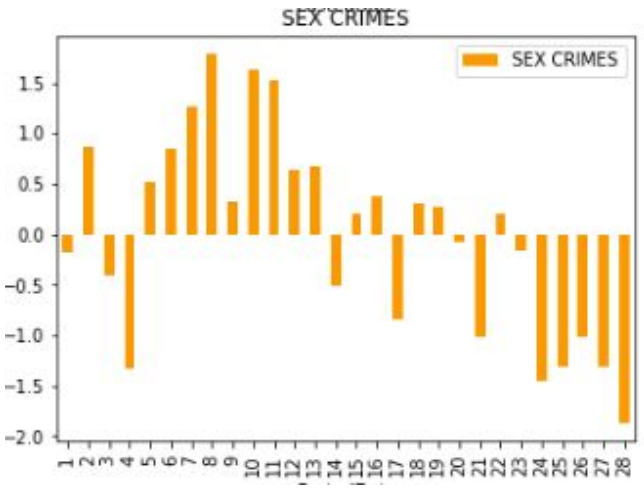


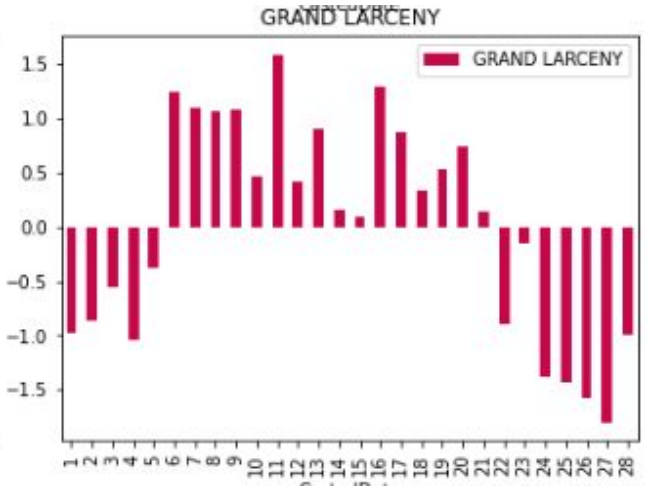




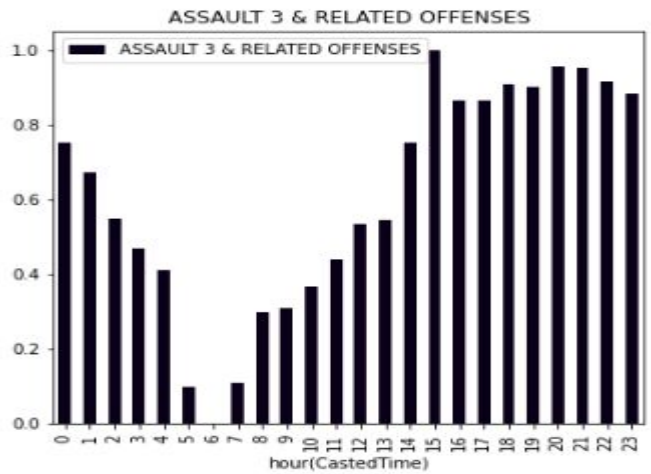
Day of Month

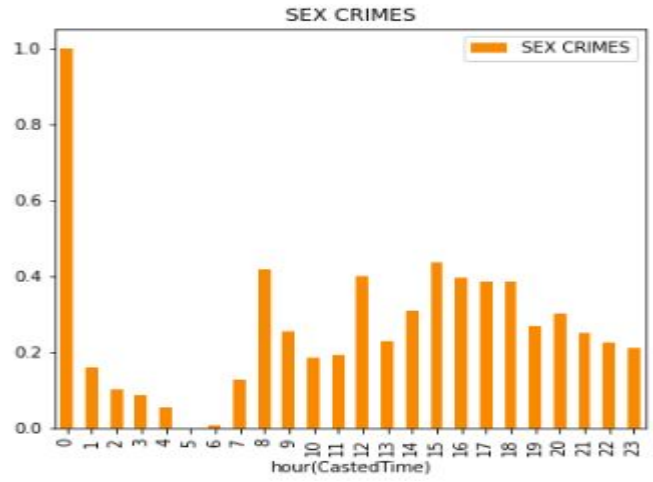


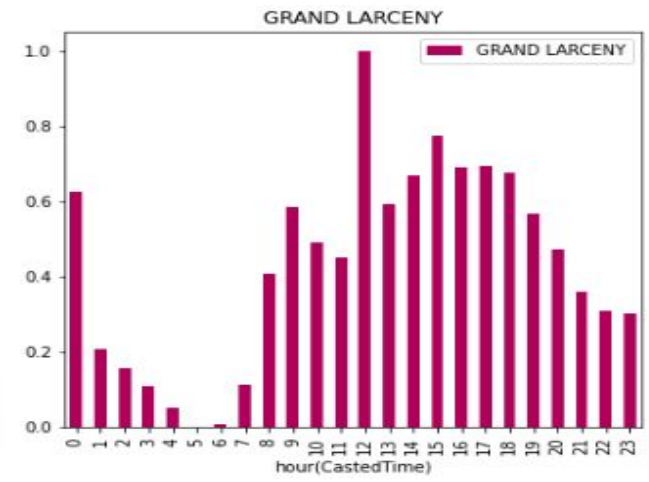




Hour of Day



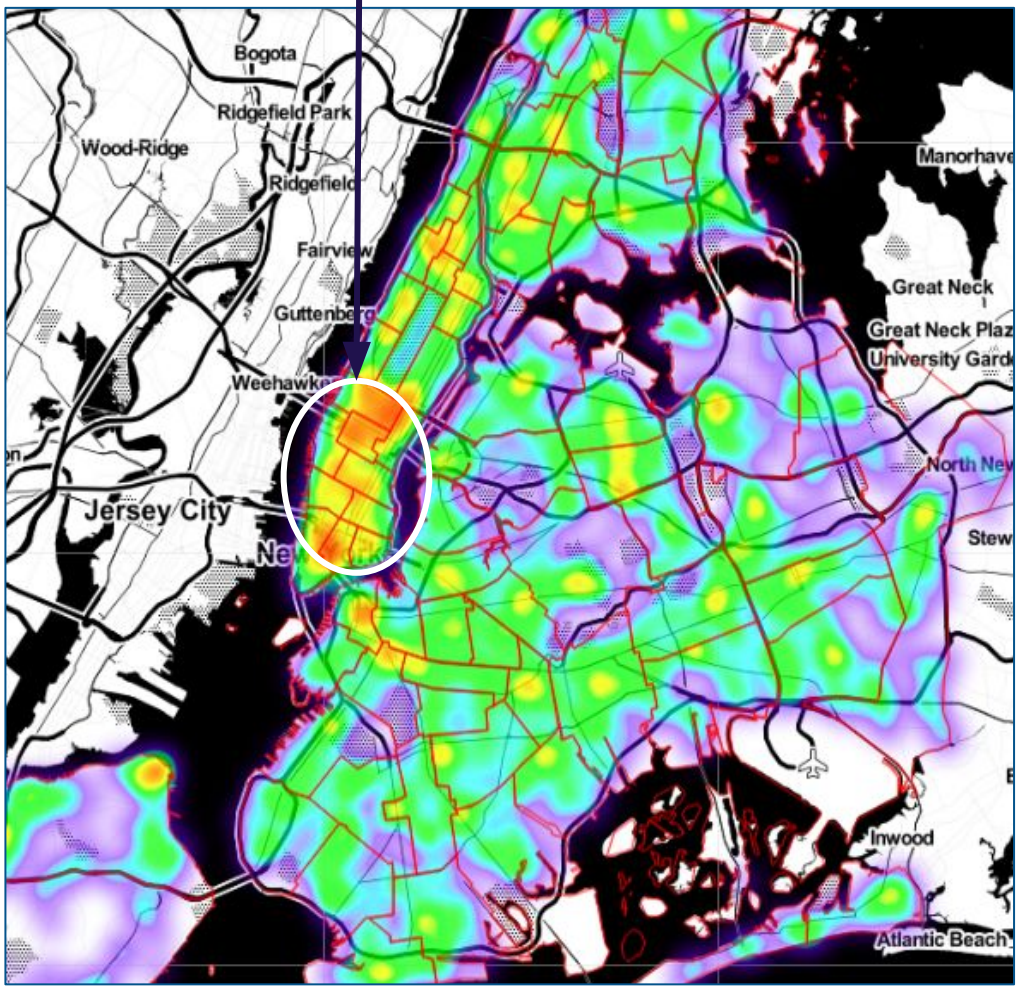




RESULTS Step 2 : Spatial Analysis

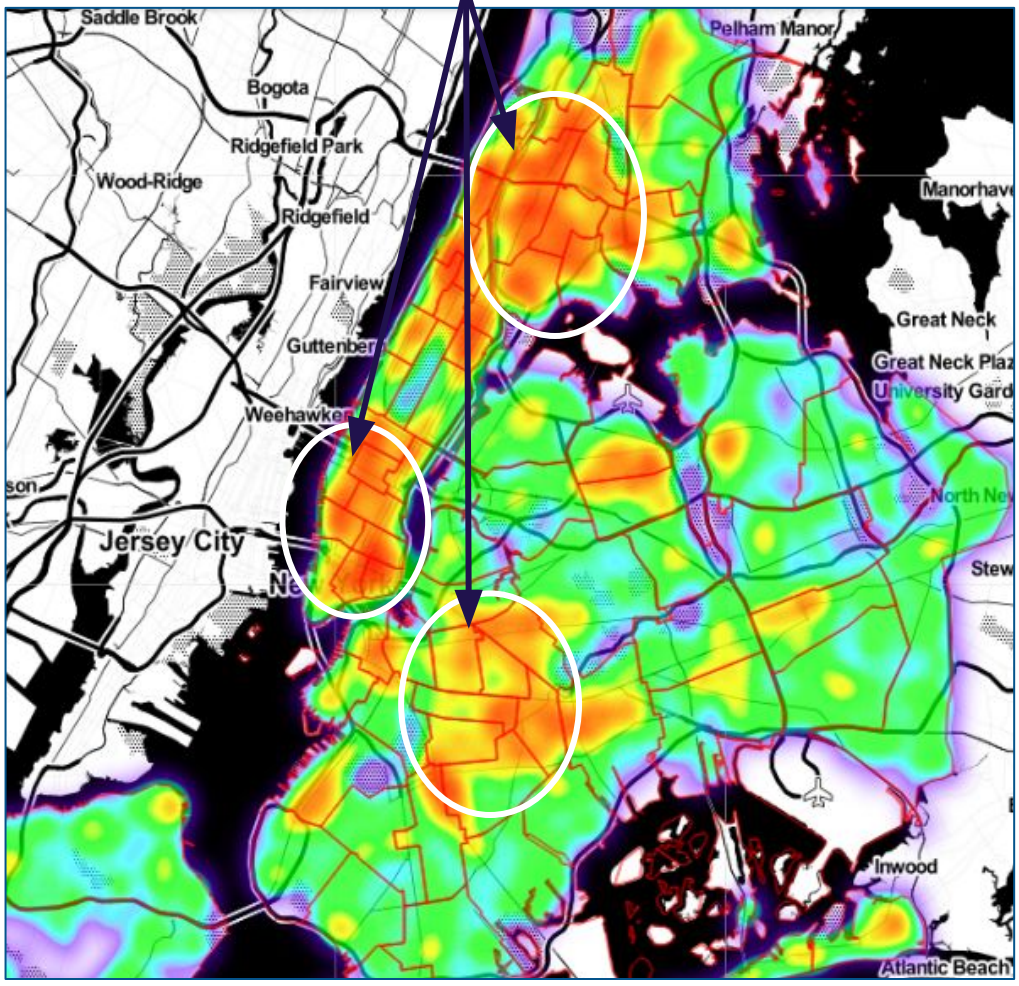
Grand Larceny in NYC - Heat Map

Precincts
Manhattan - 5, 6, 10, 13, 14



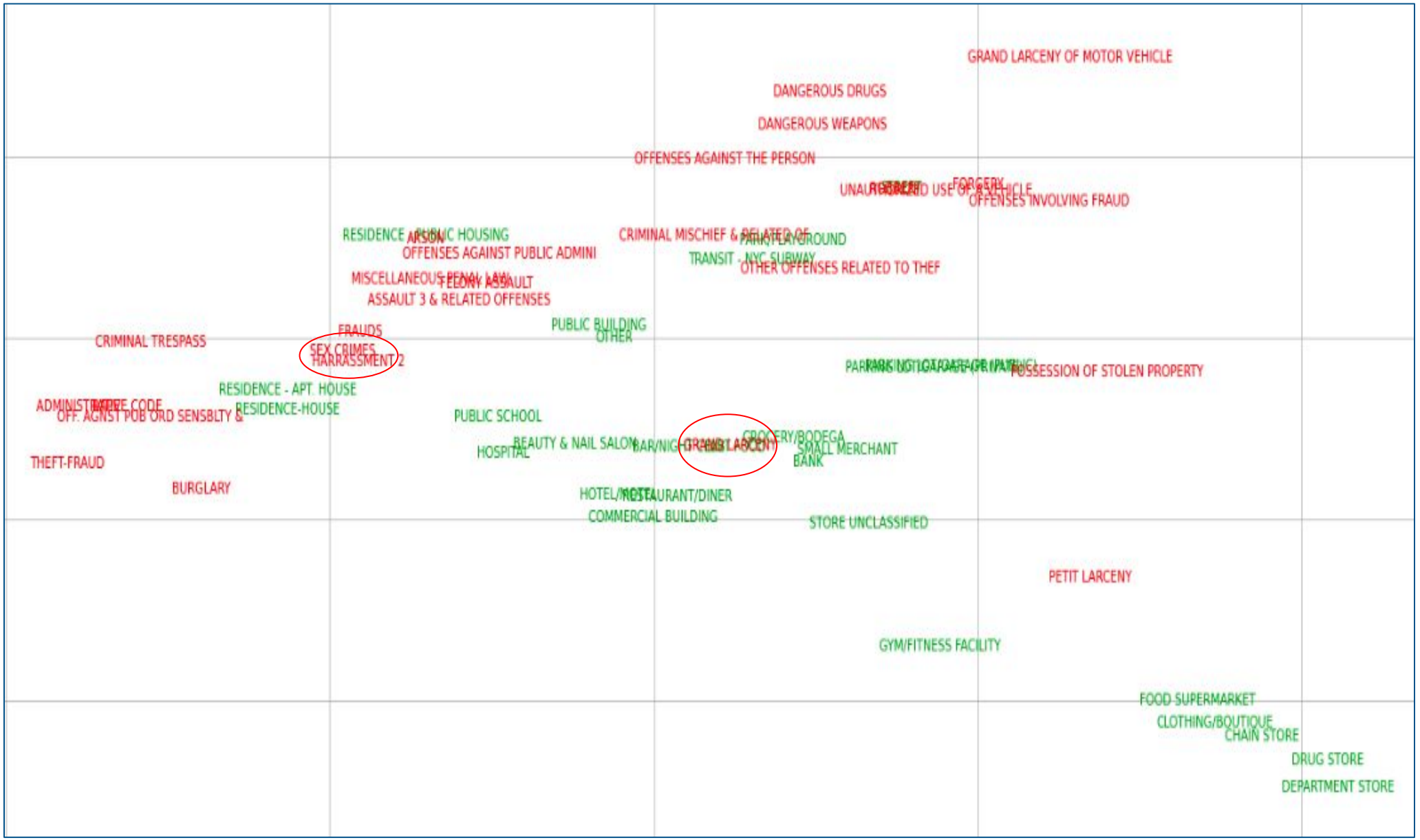
Assault in NYC - Heat Map

Precincts
Manhattan - 17, 6 ,10, 13,19
Bronx - 44, 40, 41, 42, 48, 52
Brooklyn - 83, 79, 71, 73, 77



RESULTS Step 2 : Spatial Analysis

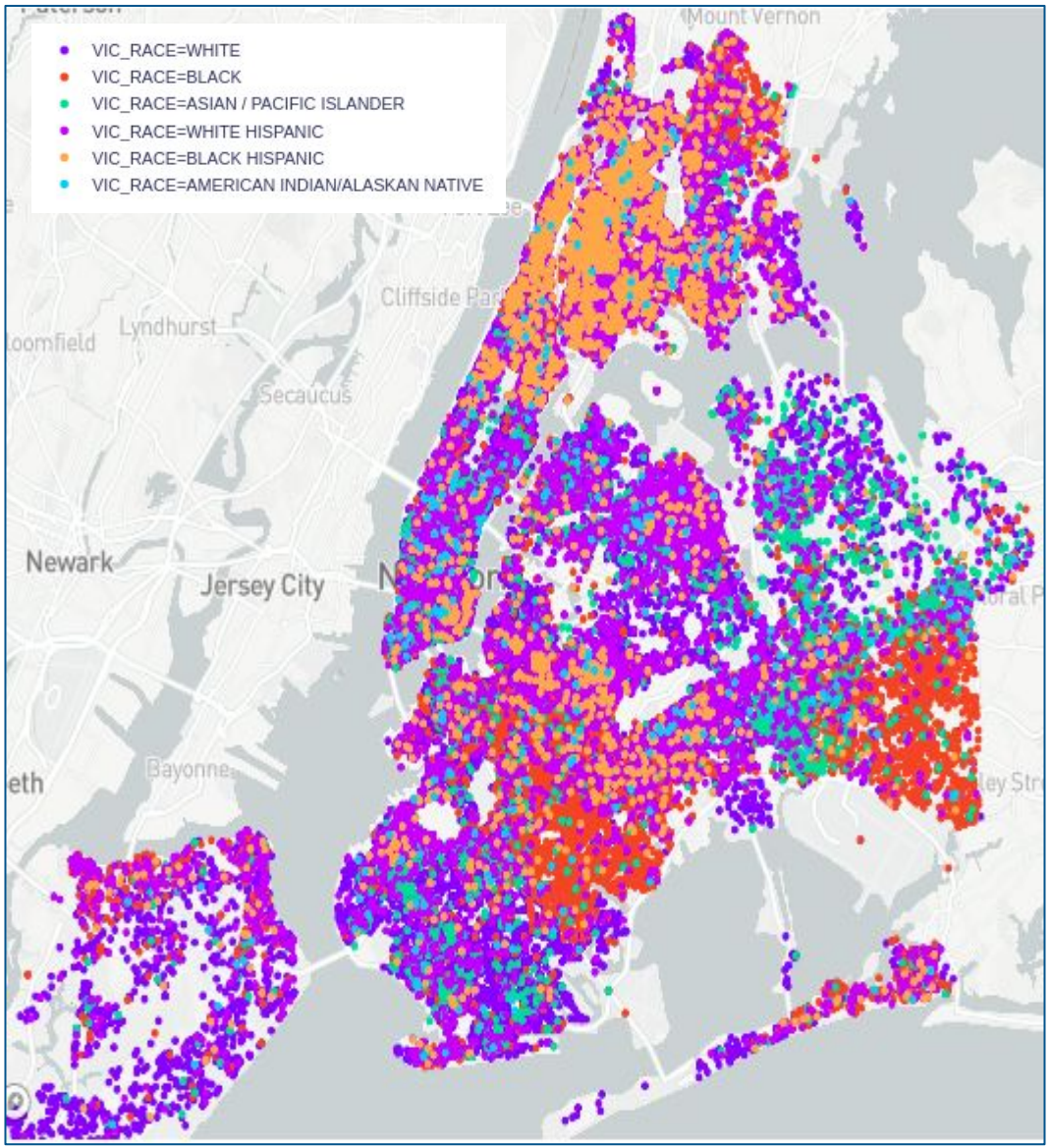
Correspondence Analysis
Premises vs Crime Type



RESULTS Step 3 : Additional Information

Correspondence Analysis of Ethnicity of Victims vs Crime Type

Location Distribution of Victims with different Ethnicities



RESULTS Step 4 : Socio-Economic Analysis

Null Hypothesis: The demographic factor don't affect the crime count at a certain spatial level (borough/ sub-borough) with or without control variable (population).

Level of Significance: 0.05

If p-value < 0.05 then
The null hypothesis is rejected
else
The null hypothesis is not rejected

	Level	Poverty Rate	Unemployment Rate	Income Diversity	Racial Diversity
No Control Variable	Borough	0.065	0.036	0.008	0.054
	Sub Borough	0.0057	0.00045	0.0125	0.0219
Controlled By Population	Borough	0.18	0.04	0.0002	0.01
	Sub Borough	6.75e-07	9.776e-07	0.002	0.007

CONCLUSION

Crime Rate is influenced by large number of development indicators

- Population density and degree of urbanization.
- Economic conditions, including income diversity racial diversity, poverty level, and job availability.
- Effective strength of law enforcement agencies.

Analysis would help in understanding patterns in Crime Statistics

- Temporal Trends in Crime Rate
- Actionable Insights for Law Authorities
- Proactive Decision Making

Thank You

WHY SHOULD WE CARE

MEN
MAKE UP AROUND
80%
OF HOMICIDE
VICTIMS
OVERALL

BUT WOMEN
CONSTITUTE
64%
OF HOMICIDE
VICTIMS
OF INTIMATE PARTNER/
FAMILY-RELATED HOMICIDE

OF DETECTED VICTIMS OF
HUMAN TRAFFICKING ARE
WOMEN AND GIRLS

MOST OF WHOM
ARE TRAFFICKED
FOR SEXUAL
EXPLOITATION