ANALYZING CRIME TRENDS AND CLUSTERS

SHEA YEE, KHOO

13 JULY 2024

AGENDA

- Introduction
- Objective
- About Dataset
- Methodology Overview
- Clustering Models
- Data Analysis & Insights Gathered
- Summary & Suggestion
- Model Limitation & Suggestions

AGENDA

- Introduction
- Objective
- About Dataset
- Methodology Overview
- Clustering Models
- Data Analysis & Insights Gathered
- Summary & Suggestion
- Model Limitation & Suggestions

INTRODUCTION

This project leverages a dataset from the Department of Justice press releases (2009-2018) to analyze crime trends across the United States. **By examining crime types and their geographical locations, we aim to uncover patterns and insights that can inform law enforcement strategies and public awareness.** Understanding these patterns will help in identifying crime hotspots and trends, enabling more efficient resource allocation and enhancing public safety.

OBJECTIVE

Problem Statement

- This project seeks to address the challenge of **optimizing law enforcement resource allocation** and **improving public awareness of crime trends**.

Proposed Solution

- Utilize **natural language processing (NLP) and clustering techniques** to identify crime hotspots and trends over time, thereby aiding law enforcement and public information efforts.

Stakeholders

- Law enforcement agencies, community groups & residents.

Value Proposition

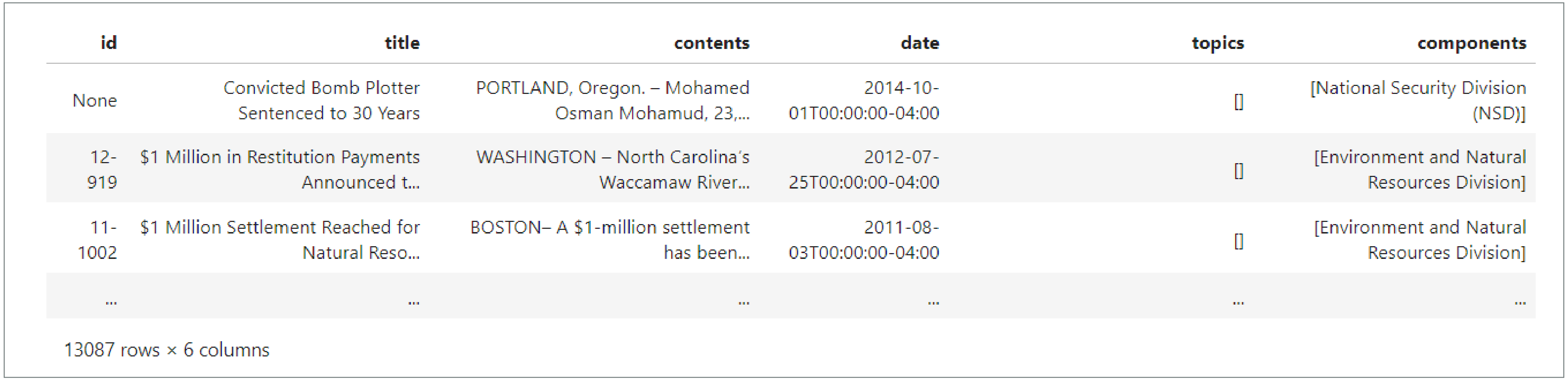
- Resource Allocation : Efficient deployment of police resources to high-crime areas.
- Public Awareness : Transparent communication about crime trends and increased community vigilance and safety.

AGENDA

- Introduction
- Objective
- **About Dataset**
- Methodology Overview
- Clustering Models
- Data Analysis & Insights Gathered
- Summary & Suggestion
- Model Limitation & Suggestions

ABOUT THE DATASET

- This is a historical dataset containing **13,087 press releases** from the Department of Justice's (DOJ) website https://www.justice.gov/news.
- Dataset is retrieved from https://www.kaggle.com/datasets/jbencina/department-of-justice-20092018-press-releases
- The contents are stored as newline delimited JSON records with the following fields:
  - **id:** Press release number
  - **contents:** Text of release
  - **topics:** Array of topic tags (if any provided)
  - **title:** Title of release
  - **date:** Posted date
  - **components:** Array of agencies & departments (if any provided)



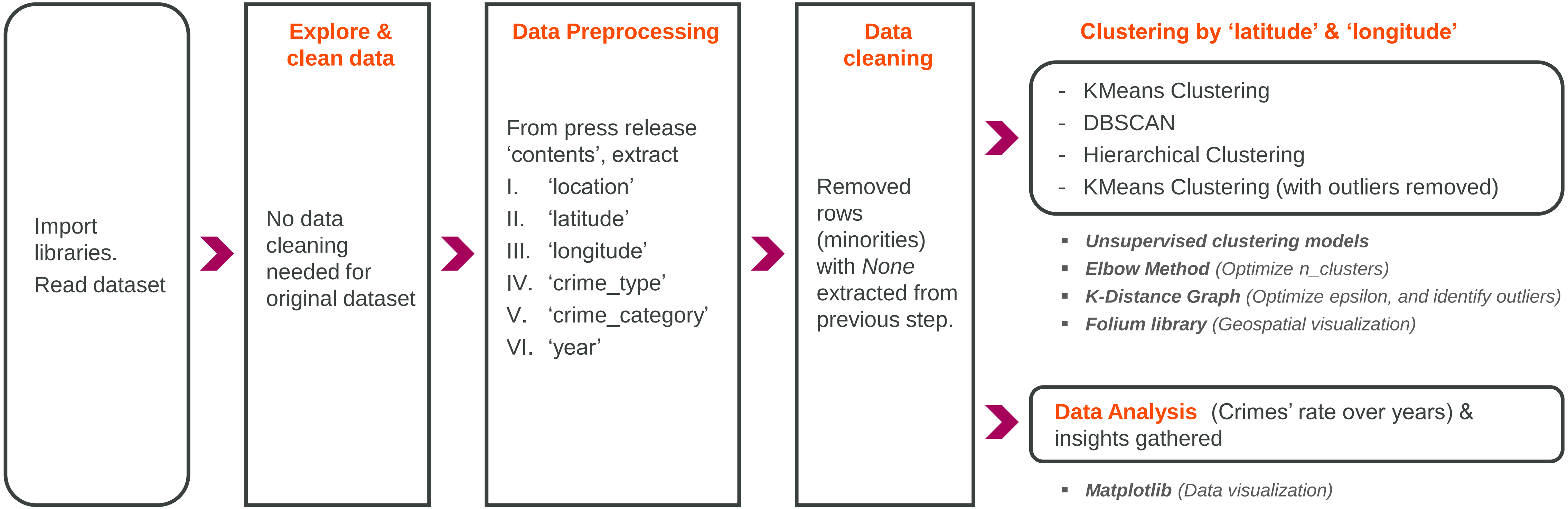
id	title	contents	date	topics	components
None	Convicted Bomb Plotter Sentenced to 30 Years	PORTLAND, Oregon. – Mohamed Osman Mohamud, 23,...	2014-10-01T00:00:00-04:00	[]	[National Security Division (NSD)]
12-919	$1 Million in Restitution Payments Announced t...	WASHINGTON – North Carolina's Waccamaw River...	2012-07-25T00:00:00-04:00	[]	[Environment and Natural Resources Division]
11-1002	$1 Million Settlement Reached for Natural Reso...	BOSTON– A $1-million settlement has been...	2011-08-03T00:00:00-04:00	[]	[Environment and Natural Resources Division]
...	...	...	...	...	...

13087 rows × 6 columns

AGENDA

- Introduction
- Objective
- About Dataset
- Methodology Overview
- Clustering Models
- Data Analysis & Insights Gathered
- Summary & Suggestion
- Model Limitation & Suggestions

METHODOLOGY OVERVIEW



Import libraries.

Read dataset

Explore & clean data

No data cleaning needed for original dataset

Data Preprocessing

From press release 'contents', extract

I. 'location'
II. 'latitude'
III. 'longitude'
IV. 'crime_type'
V. 'crime_category'
VI. 'year'

Data cleaning

Removed rows (minorities) with *None* extracted from previous step.

Clustering by 'latitude' & 'longitude'

- KMeans Clustering
- DBSCAN
- Hierarchical Clustering
- KMeans Clustering (with outliers removed)

- ***Unsupervised clustering models***
- ***Elbow Method*** *(Optimize n_clusters)*
- ***K-Distance Graph*** *(Optimize epsilon, and identify outliers)*
- ***Folium library*** *(Geospatial visualization)*

Data Analysis (Crimes' rate over years) & insights gathered

- ***Matplotlib*** *(Data visualization)*

- ***SpaCy*** *(Text extraction, lemmatization)*
- ***geopy.geocoders*** *(Generate latitude, longitude)*
- ***Others*** *(Numpy, Pandas, datetime)*

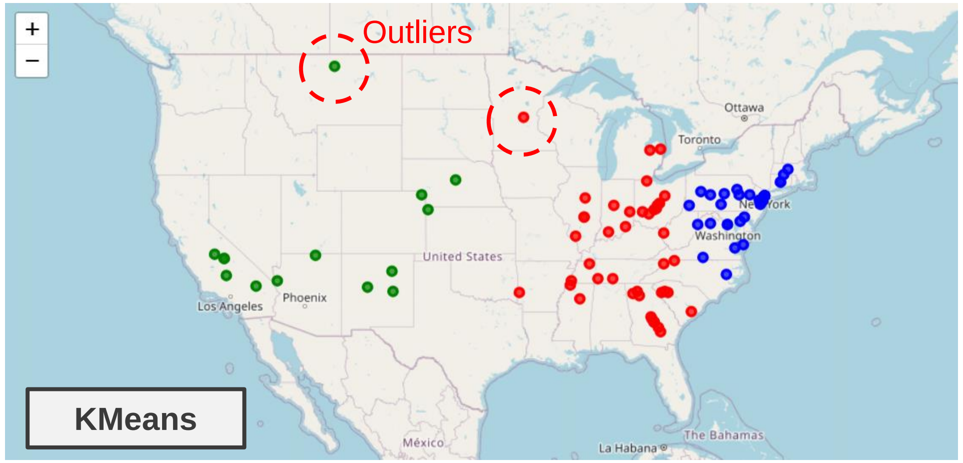
AGENDA

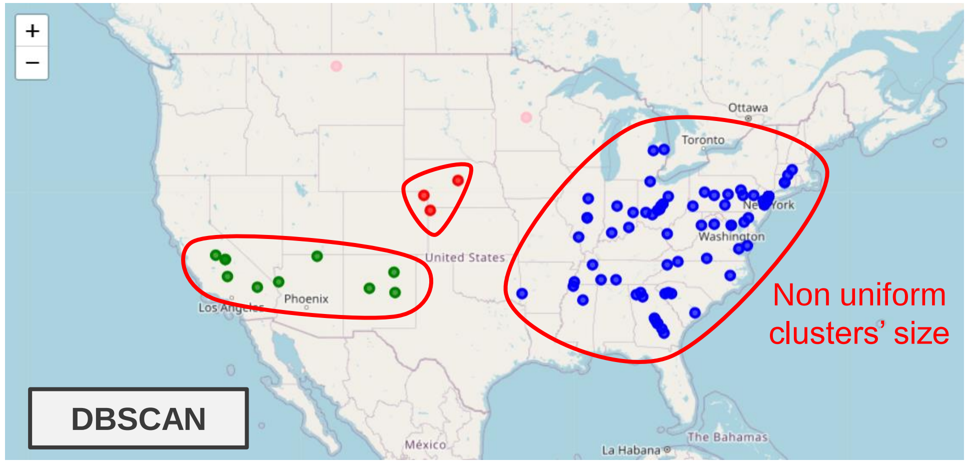
- Introduction
- Objective
- About Dataset
- Methodology Overview
- **Clustering Models**
- Data Analysis & Insights Gathered
- Summary & Suggestion
- Model Limitation & Suggestions

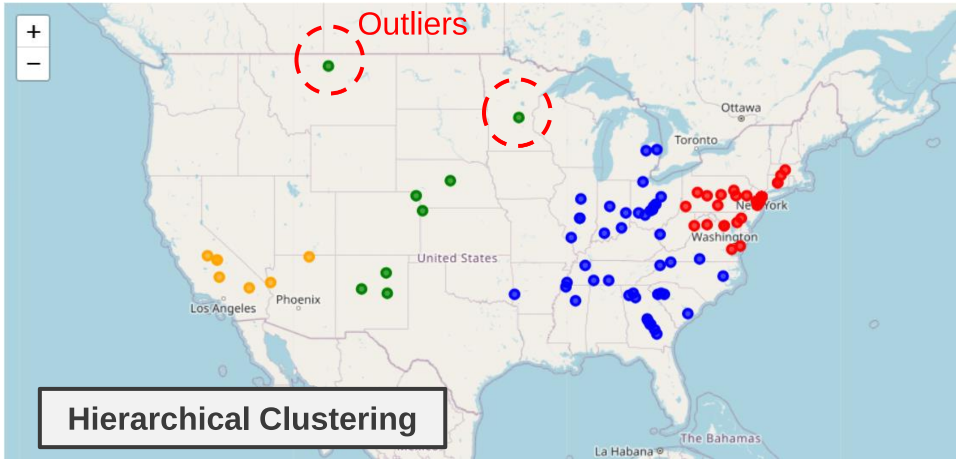
CLUSTERING MODEL

KMeans clustering by ‘latitude’ & ‘longitude’ (two features)

Example #1: Cybercrime



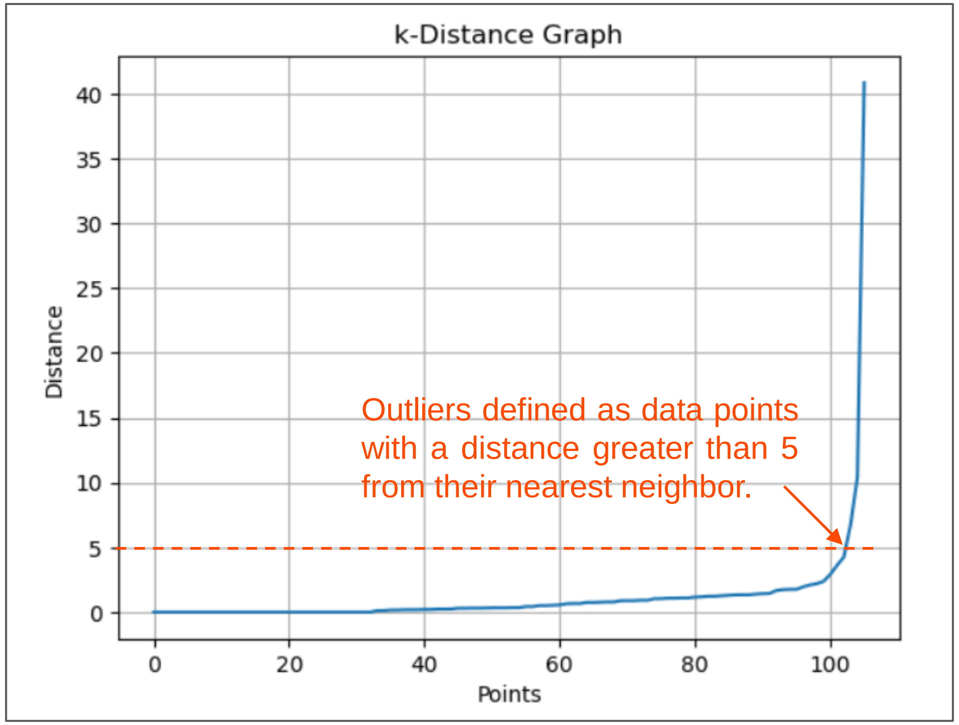




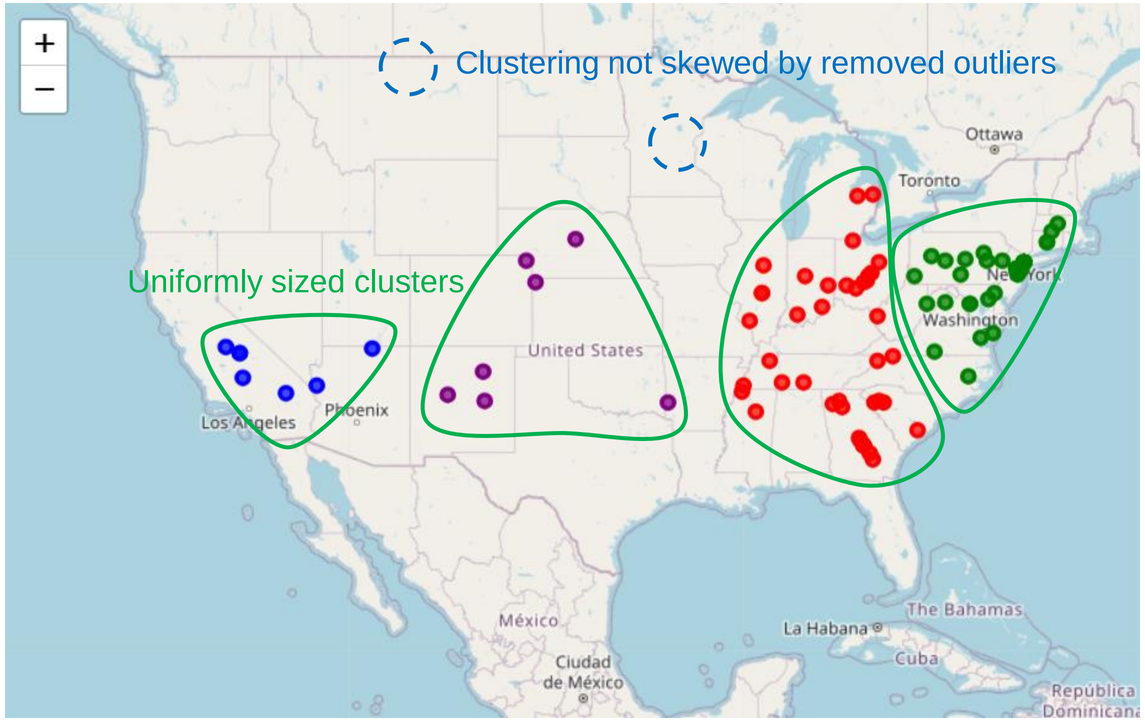
Model	Ability to isolate outliers	Clusters size
Kmeans	Poor	Uniform
DBSCAN	Good	Non uniform
Hierarchical Clustering	Poor	Uniform

None of these models are able to create clusters that are uniform in size and successfully isolate outliers.

Remove outliers identified from k-distance graph, then perform KMeans clustering



Outliers defined as data points with a distance greater than 5 from their nearest neighbor.



Clustering not skewed by removed outliers

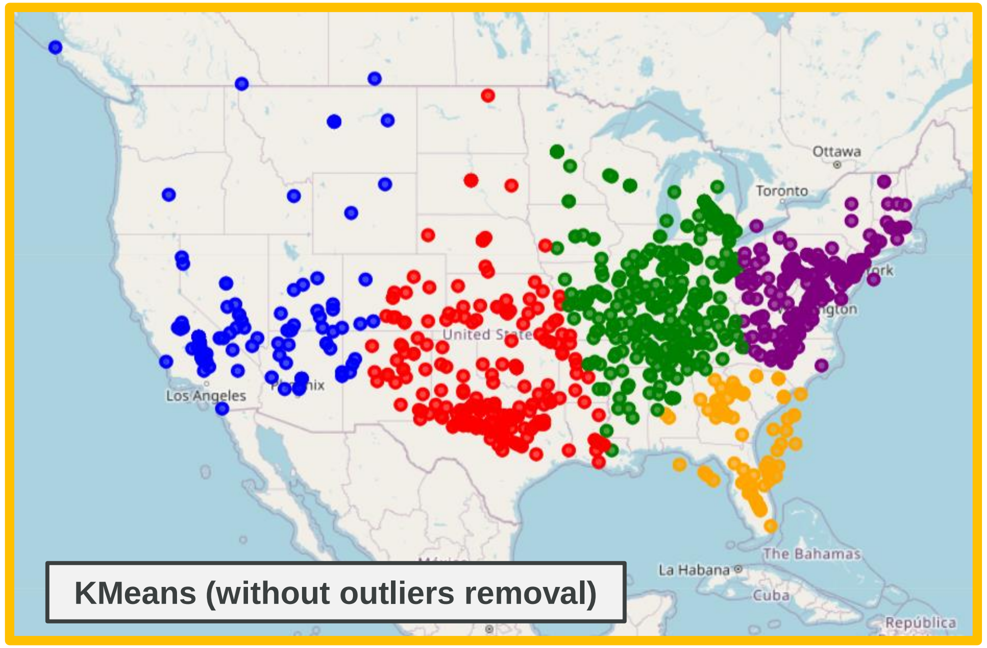
Uniformly sized clusters

- **KMeans Clustering (with prior outliers removal)** produced the best results, demonstrating higher cluster density and more uniformly sized clusters.
- Cybercrime clusters mainly in LA, Memphis, Washington, Philadelphia and NY.
- These cities are targets probably due to their high population density, significant financial and tech industries, and greater internet connectivity, which can attract cybercriminals seeking lucrative opportunities.

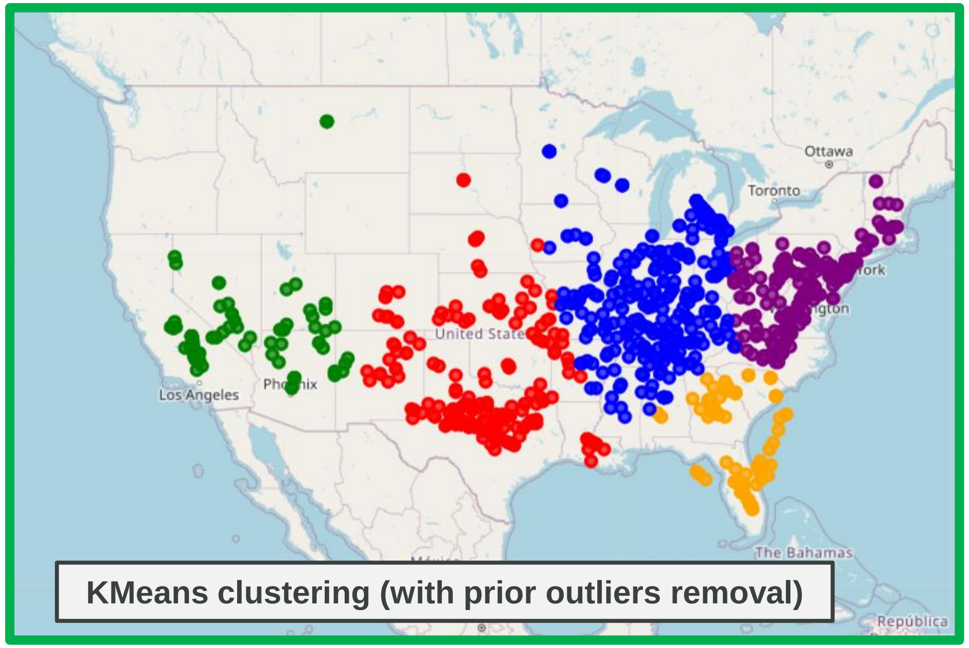
CLUSTERING MODEL

KMeans clustering by 'latitude' & 'longitude' (two features)

Example #2: Drug Offenses



KMeans (without outliers removal)



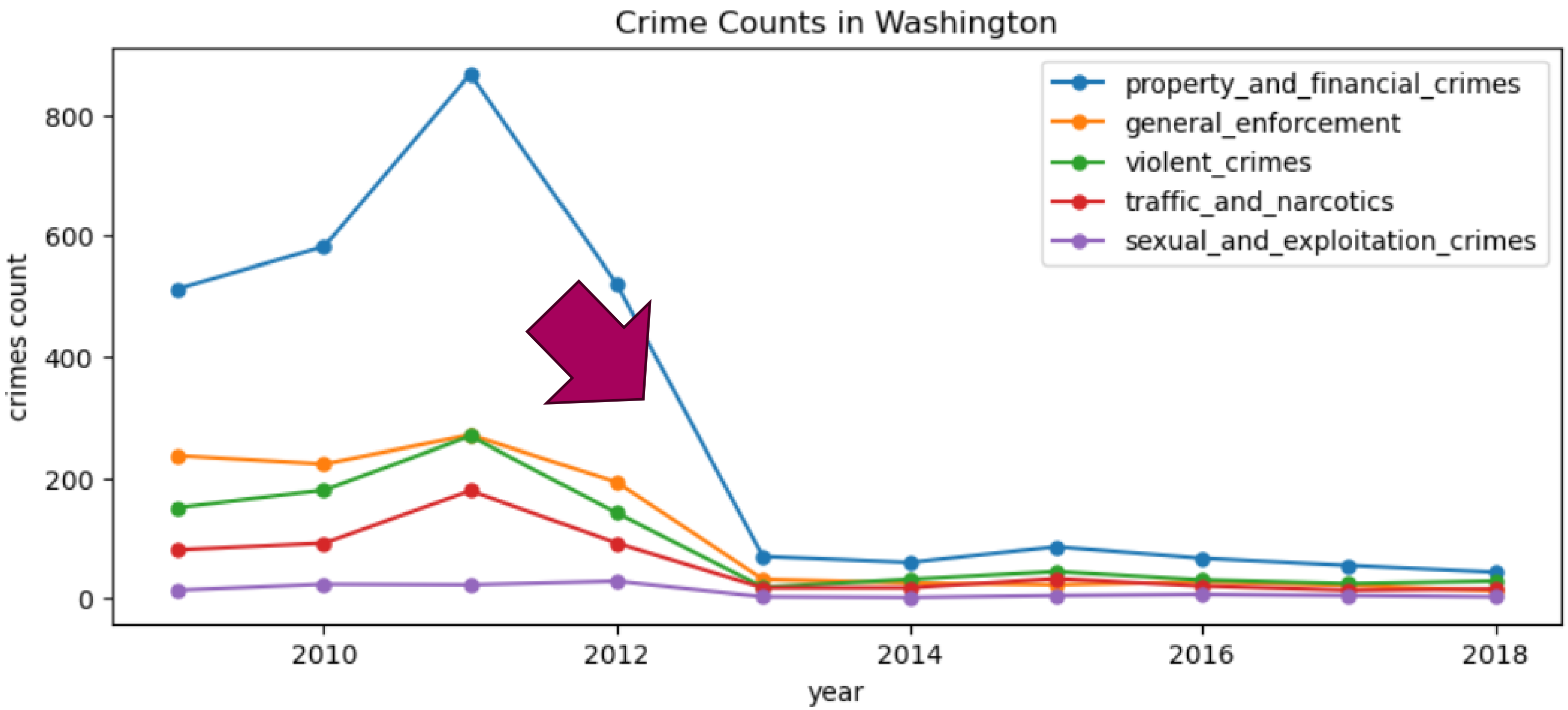
KMeans clustering (with prior outliers removal)

- The best model is validated on another crime type (drug offenses) with a different distribution of hotspots.
- **KMeans clustering (with prior outliers removal)** again performs better, demonstrating higher cluster density and more uniformly sized clusters.
- Clusters information allows law enforcement to allocate resources more effectively by targeting areas with higher crime rates, and optimizing patrols.

AGENDA

- Introduction
- Objective
- About Dataset
- Methodology Overview
- Clustering Models
- **Data Analysis & Insights Gathered**
- Summary & Suggestion
- Model Limitation & Suggestions

DATA ANALYSIS & INSIGHTS GATHERED



- Using NLP techniques to extract crime types from texts also enables us to perform EDA, such as analyzing how crime rates change over time, which is crucial for identifying trends and developing informed crime prevention strategies.
- Crime rates in Washington decreased (2011 →2013) likely due to improved law enforcement strategies like community policing and CompStat in the early 2010s, as well as the economic recovery after the Great Recession, as higher employment and better economic stability often lead to fewer economic-related crimes.

DATA ANALYSIS & INSIGHTS GATHERED

Documentation

Definition of each Crime Category:

Category	Crime Type
General enforcement	altercation, beat, violation
Violent crimes	arson, assault, attack, battery, homicide, intimidation, kidnap, kill, manslaughter, murder, robbery, terrorism, threat
Internal affairs and corruption	bribery, corruption, graft, misconduct
Property and financial crimes	burglary, cheat, collusion, conspiracy, counterfeit, damage, deceit, discrimination, dishonesty, embezzlement, extortion, forgery, fraud, game, harass, impersonation, larceny, misappropriation, mistreatment, pilfer, prejudice, scam, steal, subversion, swindle, theft
Cyber and environmental crimes	contamination, cybercrime, pollution
Traffic and narcotics	drug, dui, narcotic, traffic
Sexual and exploitation crimes	molestation, prostitution, rape, solicitation

AGENDA

- Introduction
- Objective
- About Dataset
- Methodology Overview
- Clustering Models
- Data Analysis & Insights Gathered
- **Summary & Suggestion**
- Model Limitation & Suggestions

SUMMARY & SUGGESTION

Value Proposition:

- ☑ **Optimal resources allocation** by Law enforcement agencies
- ☑ **Public awareness** on crime patterns

In summary, this project has **developed and applied advanced methodologies for analyzing crime data, focusing on NLP and clustering techniques**. By leveraging these methods, we are able to effectively identify geographical crime clusters and trends over time. These insights can significantly enhance law enforcement strategies by facilitating more targeted resource allocation and empowering proactive measures to combat crime effectively.

Moving forward, it is recommended to refine **clustering methodologies to incorporate both location and crime type** can offer nuanced insights into spatially and categorically related criminal activities. This approach allows law enforcement to **identify specific crime patterns unique to different areas and types of crime**, enabling more targeted intervention strategies.

AGENDA

- Introduction
- Objective
- About Dataset
- Methodology Overview
- Clustering Models
- Data Analysis & Insights Gathered
- Summary & Suggestion
- **Model Limitation & Suggestions**

MODEL LIMITATION & SUGGESTIONS

The current clustering method, which relies solely on "latitude" and "longitude" features, is not able to produce clear and meaningful clusters when incorporating crime type. This limitation restricts the ability to identify specific crime patterns that are unique to different areas and types of crime.

To address this, **integrating expert knowledge** can help refine the clustering process. Using predefined zones for specific crime types, for example, can enhance the accuracy of the clusters, allowing for more precise identification of crime patterns in different locations.

Additionally, **incorporating temporal features such as the time of day or day of the week** can capture temporal crime patterns that interact with location and crime type. This integration can further enhance the clustering process, leading to more insightful and actionable results.

THANK YOU