Task 2: Policing Equity

Course name – Machine Learning - Unsupervised Learning And Feature Engineering (DLBDSMLUSL01)

A course of Study – Bachelor of Science in Applied Artificial Intelligence

Author Name – Fotimakhon Gulamova

Matriculation Number – 92116230

Tutor's Name – Christian Müller-Kett

Table of Content

Introduction.

Problem Statement

The purpose of this task is to evaluate a large dataset of policing activities collected over a few years in order to discover patterns and find homogeneous categories of incidents. Also, it aims to reduce the dataset's complexity, produce meaningful visualizations that capture its key features, and provide descriptive statistics for each cluster.

Objectives and Goals

- **Data Exploration:** Understand the structure, variables, and patterns within the policing dataset.
- **Cluster Identification:** Group similar policing incidents into homogeneous clusters.
- **Visualization Creation:** Develop visualizations that illustrate the main characteristics and clusters within the dataset.
- **Descriptive Statistics:** Provide statistical insights into each cluster.

In this project, I will help handle the issues of policing equity in the local community as a freelancing data scientist. However, due to the dataset's size and complexity, it is challenging to gain meaningful insights and identify homogeneous categories of policing incidents. My goal for this use case is to reduce the complexity of the dataset and create visualizations that capture the main characteristics of the policing activities.

Firstly, I explored the data and checked it for missing values, outliers, and duplicates. Then, I encoded the data to improve the quality of the data and prepare it for analysis. Also, I will give a well-organized representation of the data to make it easier to understand. Finally, to organize unsorted data, I used the K-Means clustering method.

Step 1: Data Exploration and Preprocessing

Data Exploration

For this project, I used the data from the Kaggle dataset webpage at the Data Science for Good: Center for Policing Equity. The name of the dataset is '**11-00091_Field-Interviews_2011-2015.csv**'. I used the pandas library in Python to read a CSV

file into a DataFrame and then create a new DataFrame by sampling 10% of the data from the original DataFrame. Now, the dataset has 15,223 rows and 34 columns.

Data Cleaning

I converted the **'NO DATA ENTERED'**, '**UNKNOWN**', **'0'**, **etc** values in columns into **NaN** values. This approach allowed me to identify columns with excessive missing data. I identified columns with excessive missing data (more than 4,000 missing values) and a high number of unique values (more than 800) and dropped those columns. Then, I used the '**SimpleImputer**' method from the **scikit-learn** library with the **'*most_frequent*'** strategy, the missing values were replaced with the most frequently occurring value in each corresponding column. The dataset has duplicated rows to remove them, I used the '**drop_duplicates()**' function. Finally, I deleted all the high-correlated columns. All these actions were taken to optimize the dataset and focus on more informative features. Now the dataset contained 14,761 rows and 14 rows.

Data Preprocessing

The **'INCIDENT_DATE'** column combines year, month, day, and time in the same cell. It would be helpful to extract dates to improve the performance of machine learning methods. I converted this column to the datetime type using the **to_datetime()** method from the **pandas** class. I created two new columns: '**INCIDENT_YEAR**' and **'INCIDENT_MONTH**'. Next, I found and handled outliers in the **'INCIDENT_YEAR'** column. To replace the outlier years, I calculated a replacement value. This value is the rounded mean (average) of the **'INCIDENT_YEAR'** for the years within the range 2011 to 2015 using the **.replace()** function. After extracting the year and month information into separate columns, I no longer needed the original '**INCIDENT_DATE**' column so I dropped this column.

The dataset contains a column named **'INCIDENT_REASON.1'**, containing various reasons for incidents. I increased the clarity of the causes of these incidents by grouping them into broader categories and simplifying the rarely occurring causes.

Next, to deal with outliers in the **'OFFICER_AGE'** column I converted it to integers and replaced them with the rounded mean of ages falling within the range of 15 to 75. Then, I categorized them into specific age groups for better analysis and interpretation.

The **‘LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION’** column had 9,161 unique values. To reduce the unique values, I split the address at 'at' and extracted the non-numeric part at the end of the first part using re module. The street names in the dataset are consistent, cleaned, and stored in a new column **‘LOCATION_STREET_ADDRESS’.**

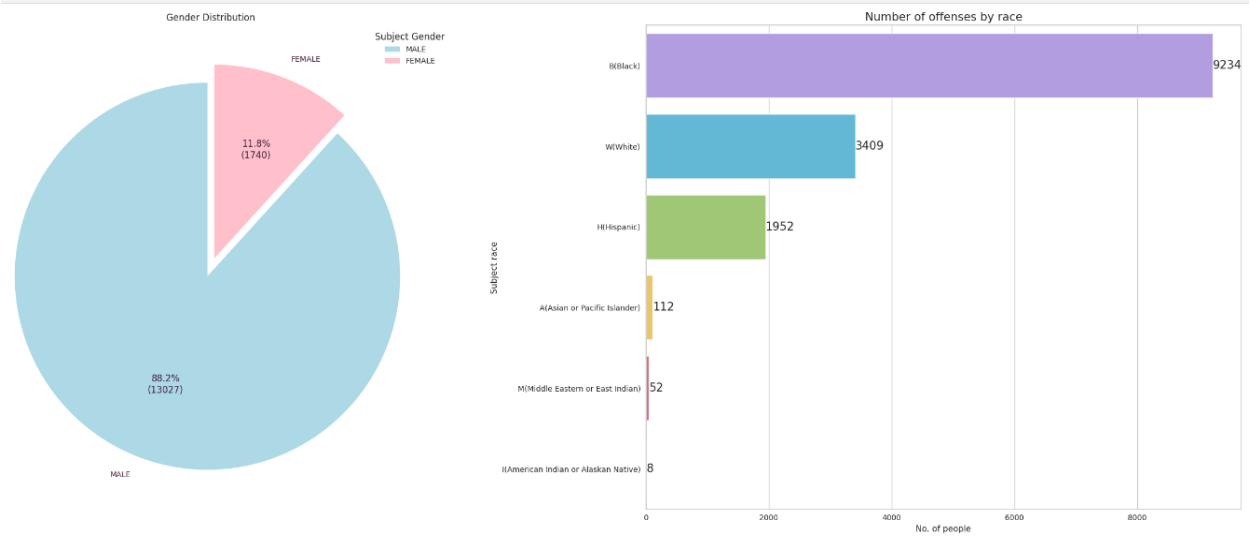
Data Encoding

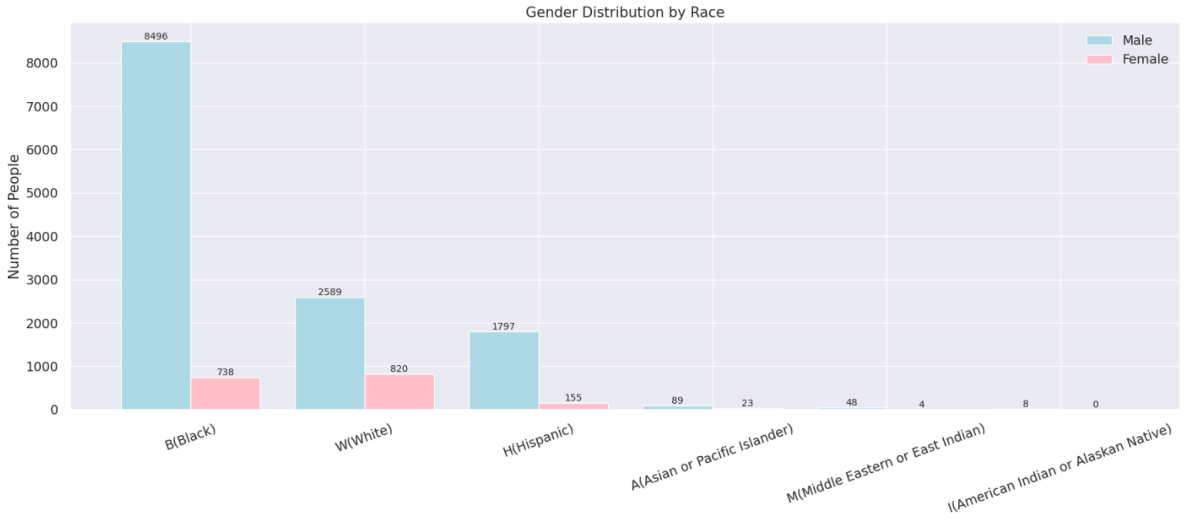
I used the **One-Hot Encoding technique** to convert categorical columns into numerical ones so that they can be fitted by machine learning models.

Step 2: Exploratory Data Analysis (EDA)

Number of offenses distributed by gender

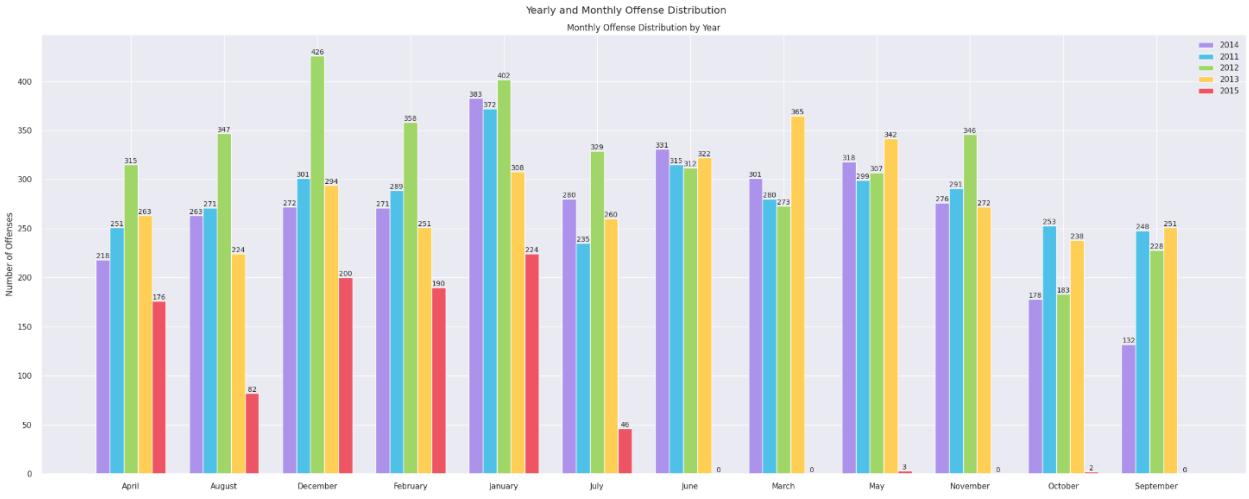
- Based on the dataset findings, the demographic breakdown of offenders indicates that approximately **88% are male**, with the remaining **12% being female**. The dataset illustrates that the majority of offenders fall within racial categories, notably Black (8496 male and 738 female), White (2589 male and 820 female), and Hispanic (1797 male and 155 female) groups.

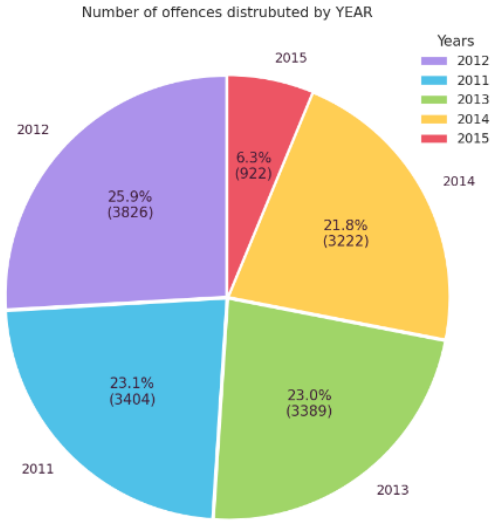


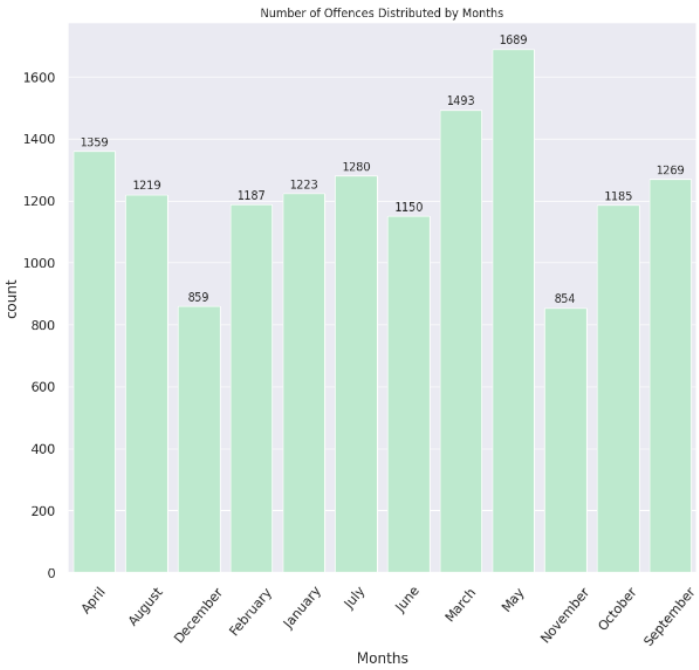


Number of offenses distributed by year and month

- The dataset provides insights into incidents that occurred between 2011 and 2015. The distribution of incidents demonstrates that from 2011 to 2014, there was a relatively consistent occurrence rate, ranging between 22% and 25%. However, only 6.3% of incidents were recorded in the year 2015. Furthermore, the data highlights a seasonal trend, with a significant concentration of incidents happening during the spring months. Specifically, March recorded 1502 incidents, April had 1388, and May saw 1732 incidents. The remaining incidents were distributed across the summer season, accounting for 3699, fall with 3436, and winter with 3386 occurrences.

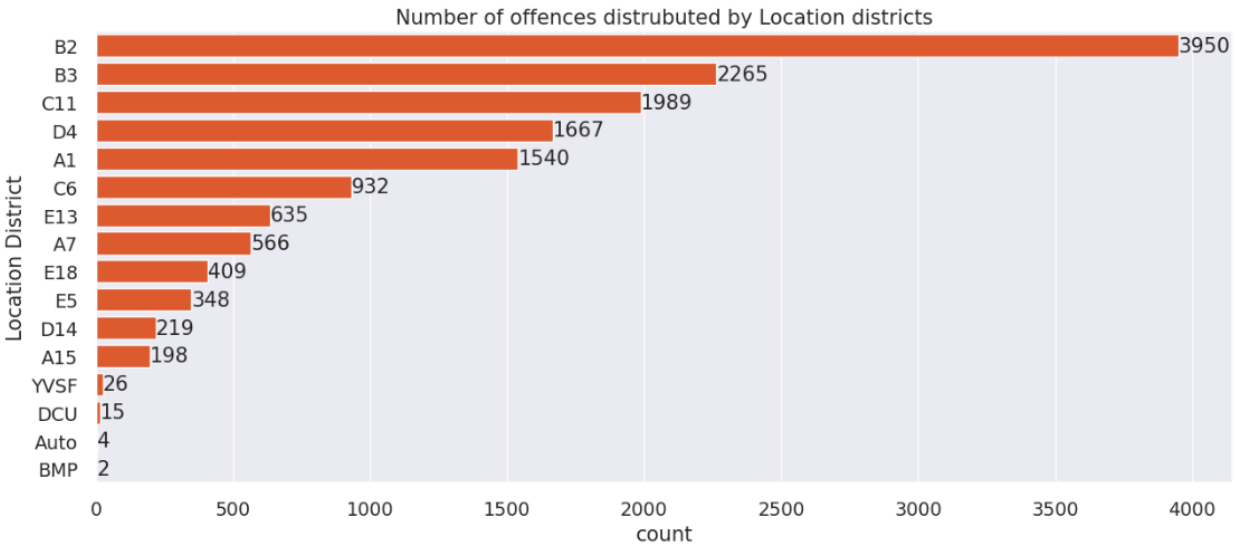


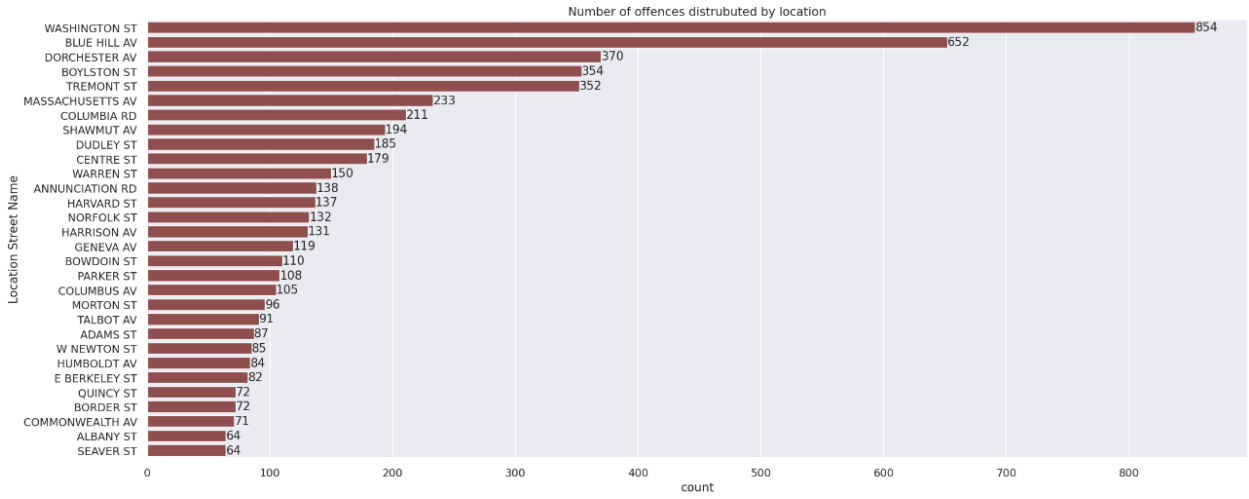




Number of offenses distributed by location

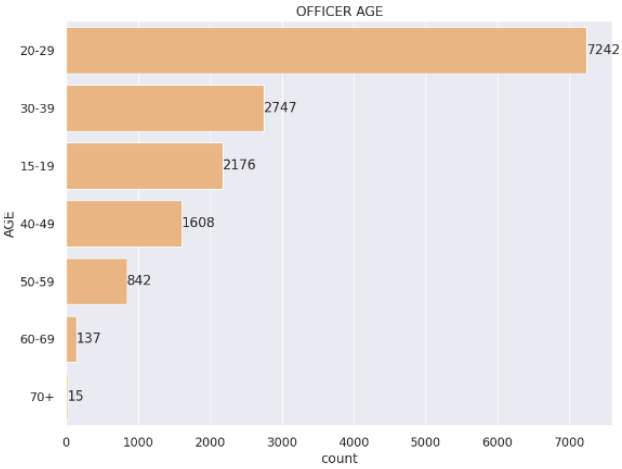
- Regarding the locations where the majority of offenses occurred, the dataset points out that Washington St (854 incidents), Blue Hill Ave (622 incidents), Dorchester Ave (370 incidents), Boylston St (354 incidents), and Tremont St (352 incidents) were among the areas with higher occurrences. Furthermore, the dataset highlights that Districts B2 (3950 incidents), B3 (2265 incidents), C11 (1989 incidents), D4 (1667 incidents), and A1 (1540 incidents) registered a notable number of offenses.

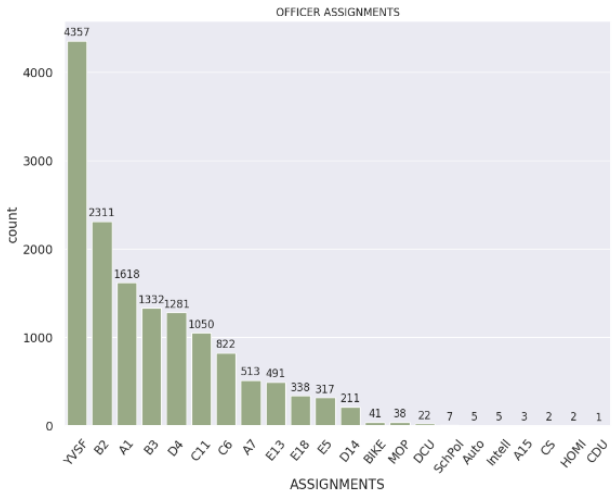




Officer Related Information

- The two images presented provide a visual representation of officers' ages and their respective assignments. The data showcases that the predominant age brackets among officers are in their twenties (20-29 years), totaling 7242 individuals, followed by those in their thirties (30-39 years), amounting to 2747. In terms of assignments, a significant number of officers, approximately 4350, are engaged in YVSF assignments. Additionally, 2311 officers are assigned to B2, while 1618 officers are allocated to A1 duties.

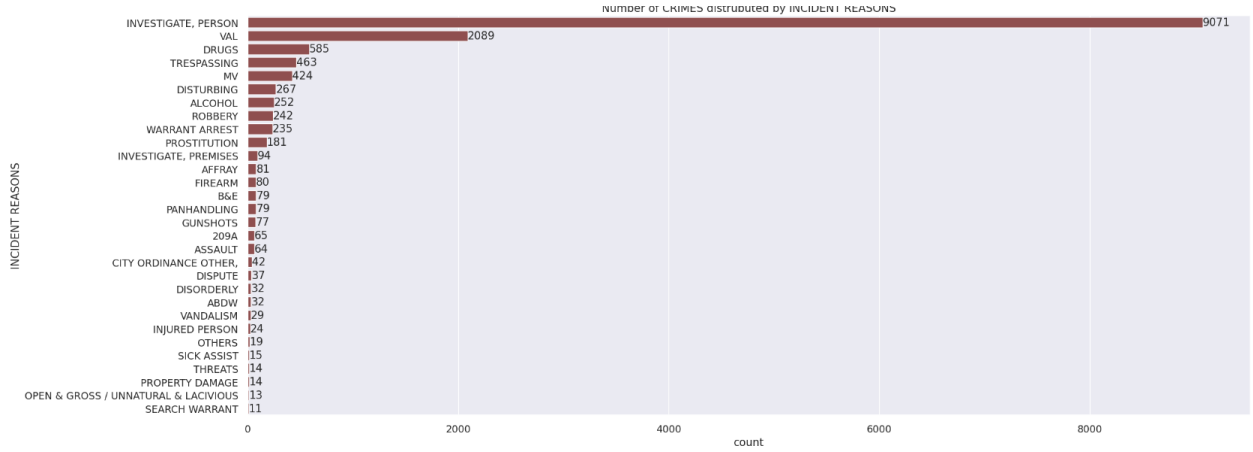




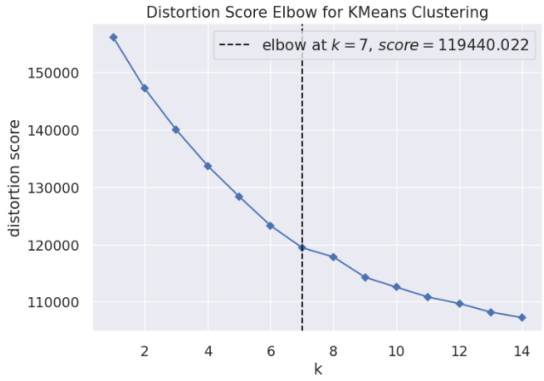
Number of offenses distributed by incident reasons

- Regarding the motives behind these incidents, the dataset indicates that the most frequent reasons were related to investigating persons (9071 incidents), followed by VAL

(2089 incidents), drug-related offenses (585 incidents), trespassing (463 incidents), and motor vehicle-related incidents (424 incidents).

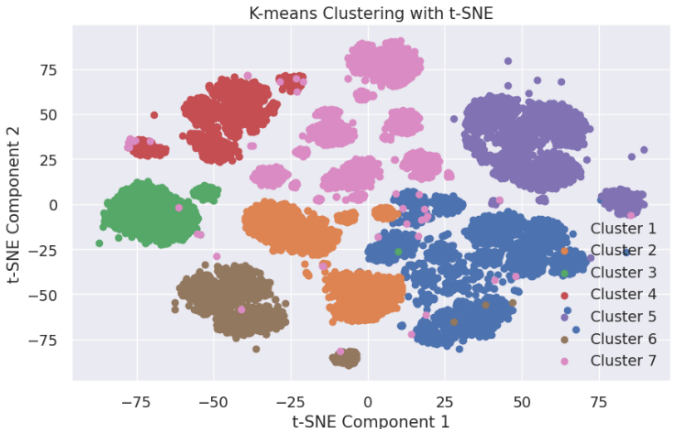


Step 3:Cluster Analysis



I used the **KMeans** clustering technique for grouping the dataset and to choose the optimal k value I used the **KElbowVisualizer** class from *yellowbrick.cluster* module. Based on the distortion score it shows that seven clusters are the optimal k value for KMeans.

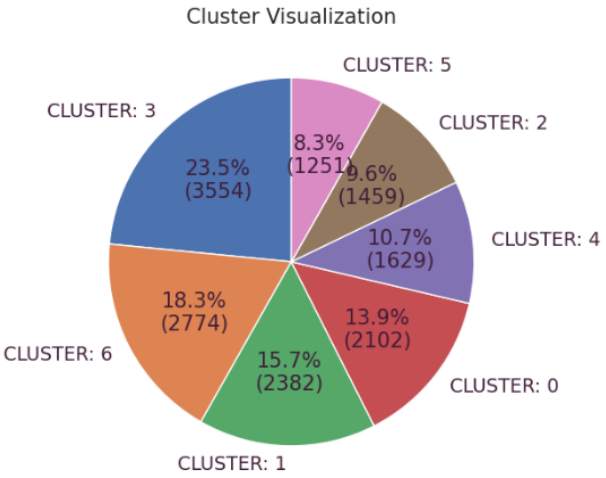
I used **T-distributed neighbor embedding (TSNE)** to provide a visual representation of the results of the clusters. And to assess the quality of clustering results I calculated the evaluation metrics.



- Silhouette Score (0.1121) means there might be some overlapping in cluster assignments.
- Davies-Bouldin Index (2.8716) suggests a moderate level of separation between clusters.

- Calinski-Harabasz Index (763.2342) is quite high, the data points within clusters are spread out from their centroids.

Results of Clusters



Cluster 1: 13.9% of the population (2102 individuals), mainly Black males from Blue Hill AV in District B2. Officers, aged 20-29, on their fourth assignment. Crimes mostly in May 2014.

Cluster 2: 15.7% of the population (2382 individuals), primarily Black males from Dorchester AV in District C11. Officers, aged 20-29, handling their 5th assignment. Crimes mainly in March 2011.

Cluster 3: 9.6% of the population (1459 individuals), mainly Black males from Massachusetts AV in District D4. Officers, aged 20-29, on their 8th assignment. Crimes mostly in April 2011.

Cluster 4: 23.3% of the population (1354 individuals), mainly White males from Boylston ST in District A1. Officers, aged 20-29, on their first assignment. Crimes mainly in May 2011.

Cluster 5: 10.7% of the population (1629 individuals), mainly Black males from Blue Hill AV in District B2. Officers, aged 20-29, on their third assignment. Crimes mainly in May 2013.

Cluster 6: 8.3% of the population (1251 individuals), mainly Black males from Blue Hill AV in District B3. Officers, aged 20-29, on their 4th assignment. Crimes mainly occurred in January 2013.

Cluster 7: 18.3% of the population (2774 individuals), mainly White males from Centre ST in District C6. Officers, aged 20-29, on their 6th assignment. Crimes mainly occurred in May 2012.

Links to the code

- **GitHub:**https://github.com/fatimagulomova/iu-projects/blob/main/DLBDSMLUSL/task-2-policy-equity.ipynb
- **Kaggle:** https://www.kaggle.com/code/fotimakhongulomova/task-2-policy-equity

Conclusion

In conclusion, an attempt to analyze and understand a complex dataset on police activity has become an important step toward identifying patterns and insights that are crucial to solving problems related to ensuring fairness in police work in the local community.

Overall, this analytical task aimed to reduce the complexity of the dataset, extract meaningful patterns through clustering, and present a comprehensive overview of policing activities.