**Crime Pattern Analysis Using Unsupervised Clustering**  
*Allan Shibu Thomas*

*3715 Principles of Data Science*

*Hongchang Gao*

**Introduction**

Understanding crime and its pattern is a critical goal for safety and resource allocation in law enforcement agencies. With large scale crime data now being publicly available, there is an opportunity to extract actionable insights using data-driven approaches. However crime data is complex and often lacks clear labels, that makes supervised learning difficult for application without extensive preprocessing or domain specific labeling.

This project tries to investigate the use of unsupervised machine learning and data science techniques, specifically clustering algorithms, to identify meaningful groupings within crime incidents reported in Los Angeles between 2020 and the present. The aim of the project is to discover hidden patterns in when, where and what types of crimes occur, based on features such as crime types, time of day. Victim characteristics like ages, and geographical locations.

The dataset we used, obtained from Candace Gostinski through Kaggle, consists of hundreds of the thousands of crime reports. Through data pre-processing, dimensionality reduction, clustering, and spatial analysis, we seek to segments this data into meaningful groups that reveal patterns across crime types, times and areas. The hope is that these insights could help with patrol planning, resources allocation and publicity policy decisions and hopefully, crime prevention or crime reduction.

This report outlines the methods used to pre-process the data, apply and compare the appropriate clustering models and interpret the resulting clusters. By evaluating multiple clustering algorithms and selecting the model with the highest and best silhouette score, I aim to provide visual results among quantitative results that support patterns discovery in crime.

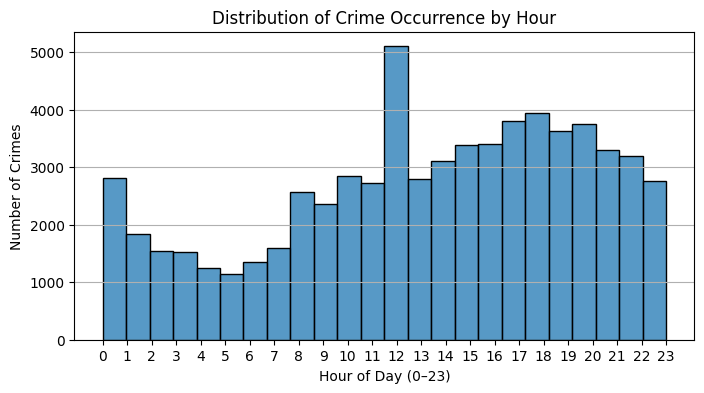
**Approach**

The project was executed with four phases in mind. These phases are data preprocessing, dimensionality reduction, clustering model selection, and cluster interpretation. Each phase, along with its purpose, forms a key cornerstone of this project.

**Phase 1, Data Pre-Processing:** The raw dataset contained records of crime incidents, each with attributes such as the crime code, description, date and time, victim demographics, and location. The diverse preprocessing steps were done through the following:

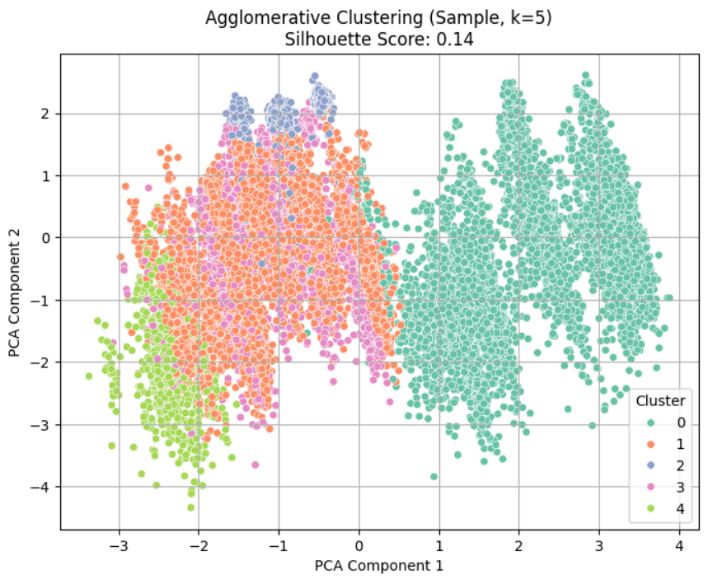
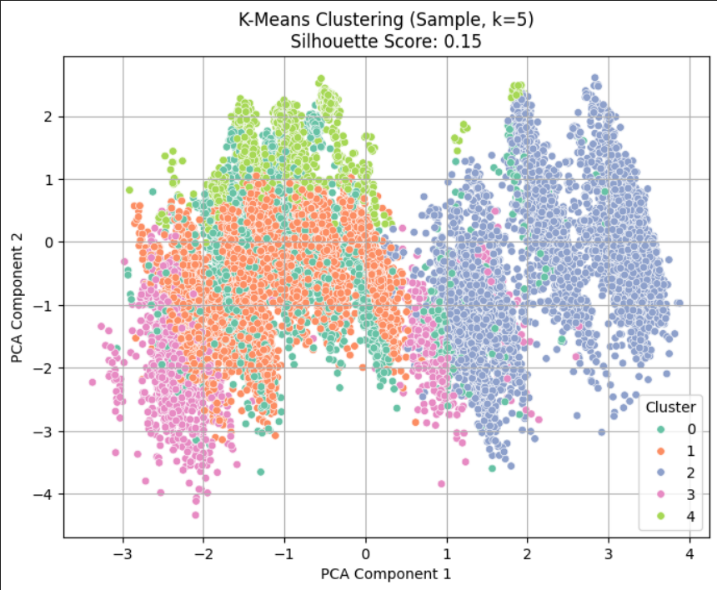
* Removing duplicate and unwanted columns
* Parsing the time of Crime data to derive the hour of day
* A derived feature, 'time of day,' was created by categorizing timestamps into morning, afternoon, evening, and night.
* Handling any missing values in fields such as Victim age, Crime Description and Weapons used.

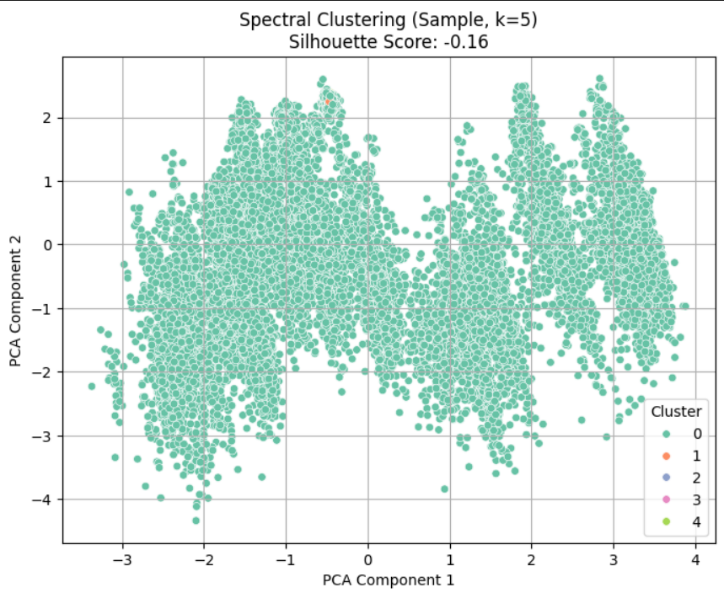
**Phase 2, EDA (Exploratory Data Analysis) and PCA (Principal Component Analysis):** EDA was conducted to understand distributions and relationships among features. All Categorical features were label-encoded while numerical features were scaled using Standard Scaler. Then PCA was applied to reduce the dimensionality while preserving most of the important data. This allowed for cluster separation and visualizations.



*Figure 1 : Distribution of Crime Occurrence by hour. This figure is part of the exploratory data analysis. This figure shows that crimes are not uniformly distributed throughout the day. There is a peak occurring around noon, followed by sustained activity in the late afternoon and evening hours.*

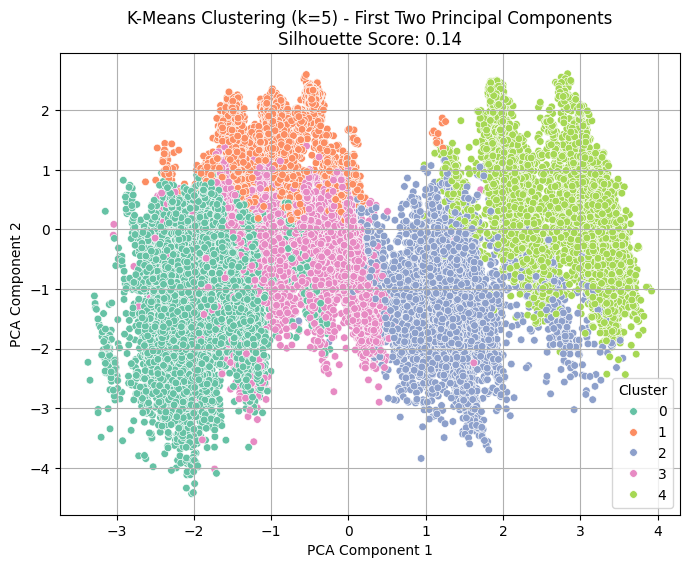
**Phase 3, Clustering Model Selection and Application:** The three clustering models we compared used were K-means, Agglomerative Clustering and Spectral Clustering. There were initial attempts made to run clustering models on the full dataset, but the models encountered a memory issue specifically for Agglomerative and Spectral clustering. As a solution for determining a fair assessment of each model, a representative sample of 25,000 records was used during model selection. The number of clusters was fixed at k = 5, and each model was evaluated using silhouette score. Based on figure 2, K-means seemed the most reliable, and so it was selected for the project.





*Figure 2: Comparison of Clustering Results Using PCA Projections (Sample, k=5).* This figure compares the clustering outcomes of three models: K-Means, Agglomerative Clustering, and Spectral Clustering, each applied to a 25,000-record sample and visualized using the first two principal components. K-Means clustering (left) shows clear and well-separated clusters, corresponding to the highest silhouette score (0.15). Agglomerative Clustering (middle) achieves moderate separation with some overlapping clusters, reflected by a slightly lower silhouette score (0.14). Spectral Clustering (right) fails to form distinct clusters, resulting in significant overlap and a negative silhouette score (-0.16).

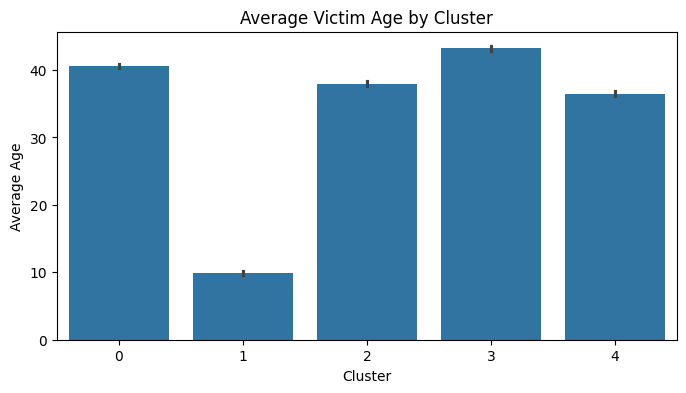
**Phase 4, Results and Evaluation:** After applying K-Means clustering to the full dataset, We evaluated the resulting clusters by analyzing crime type, time of day, victim age and crime distribution by cluster and area. Visualizations and grouped statistics were generated to support the interpretation and assessment of the clustering results.



*Figure 3: PCA Scatter Plot of K-Means Clustering on Full Dataset (k=5).* This figure shows K-Means clustering applied to the full dataset, projected onto the first two PCA components. Cluster boundaries remain distinct even after scaling to the full dataset, and the silhouette score remained strong.

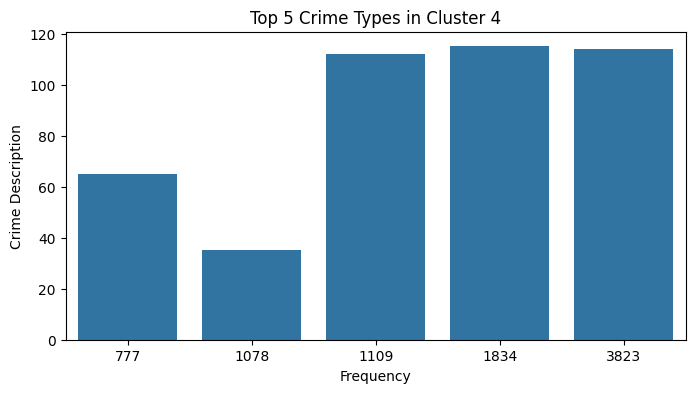
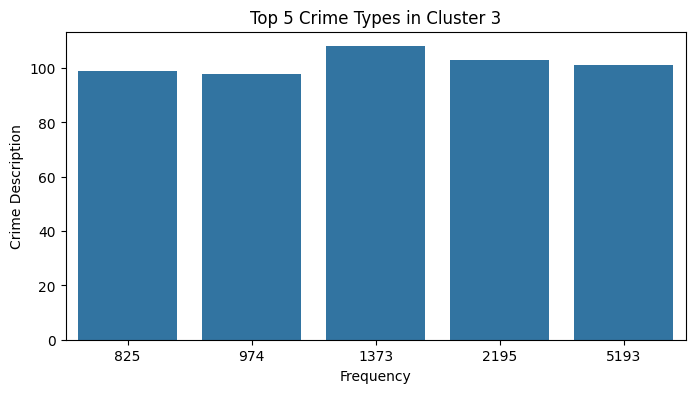
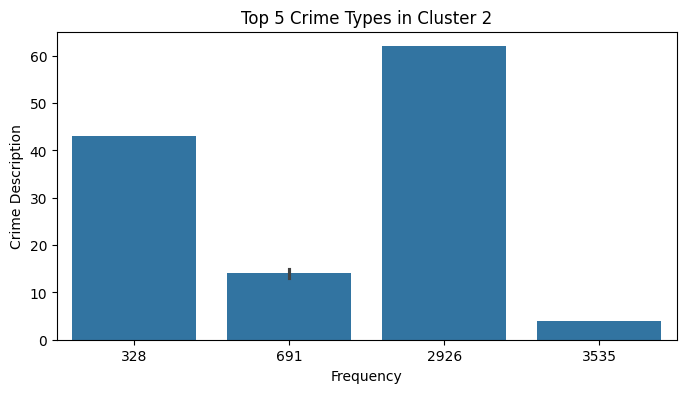
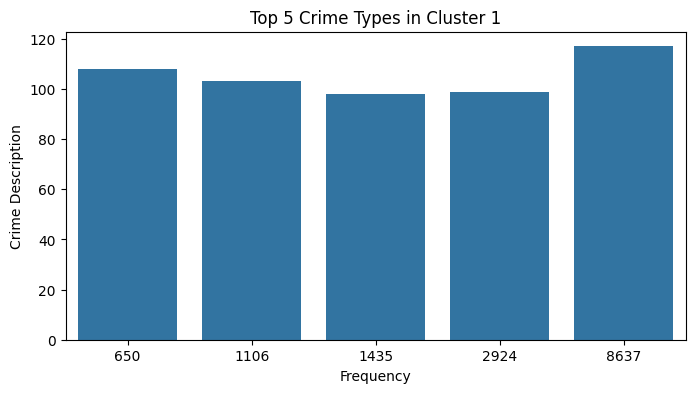
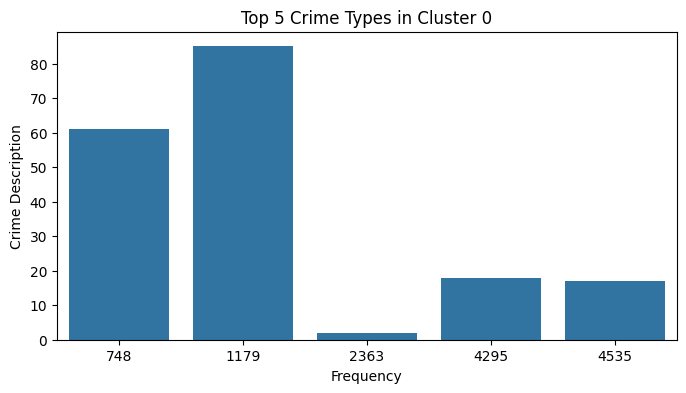
**Results**

The application of K-Means clustering to the full dataset revealed distinct patterns in crime data, supported by strong visual separation in the PCA projection and consistent quantitative evaluation through silhouette scores.



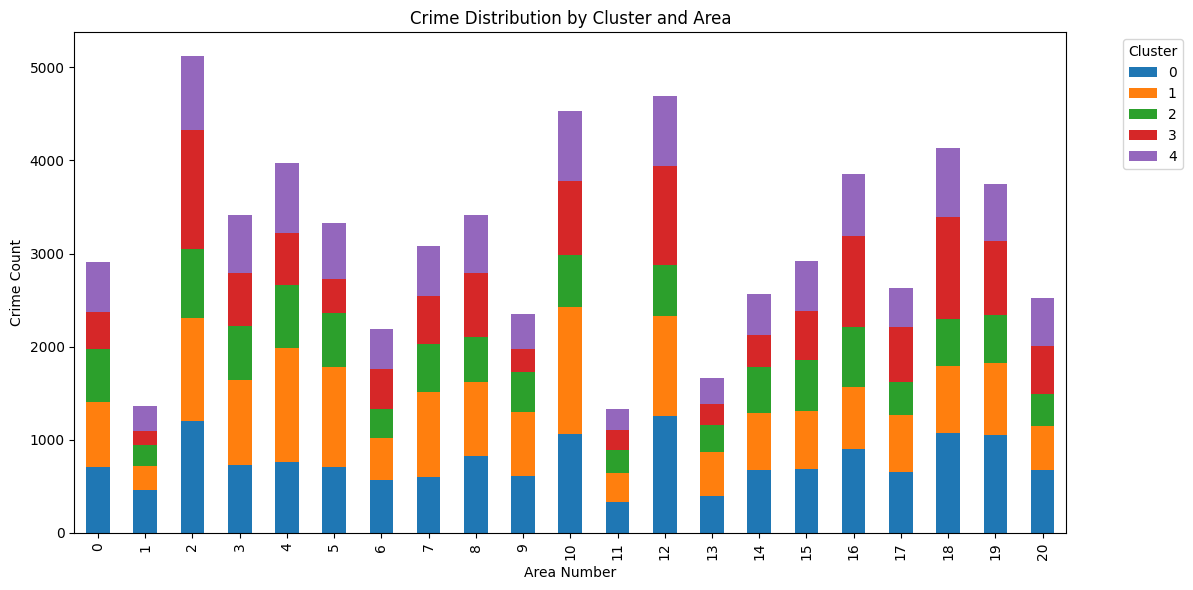
*Figure 4: The Average Victim age for each cluster.* This figure shows the average victim age across different clusters. This image was acquired from clusters formed by the K-Means clustering Model.

Clusters displayed significant variation in victim ages, with one cluster heavily skewed toward younger victims and others more associated with older individuals. This difference shows that different types of crimes tend to affect different age groups. In Cluster 0, the average victim age was around 40 years, similar to Clusters 2 and 4, suggesting that these clusters represent crimes affecting the general adult population. Cluster 3 appears to represent crimes against elderly individuals. In contrast, Cluster 1 had an average victim age of 10, indicating that it likely includes crimes against children, such as domestic abuse, vehicular assaults, accidents, or even more severe offenses.



*Figure 5: Top 5 Crime Types in Each Cluster (Clusters 0-4).* Bar plots summarize the five most frequent crime types for each cluster separately.

This analysis revealed that clusters differed sharply in their dominant crime types. In Cluster 0, crime code 1179 (vehicle theft) occurred most frequently, indicating a prevalence of property crimes. Cluster 1 was dominated by crime code 8637, representing criminal threats through verbal intimidation. Cluster 2 showed a high frequency of crime code 2926 (fraud), while Cluster 3 was associated with crime code 1373 (miscellaneous). Finally, Cluster 4 was dominated by crime code 1109 (grand theft auto). Understanding these dominant crime types helped clarify the behavior patterns and risk profiles within each cluster.



*Figure 6: Crime Distribution by Cluster and Area*. This stacked bar chart shows the distribution of crimes across Los Angeles areas, broken down by cluster.

The bar chart represents the number of crimes occurring across different parts of Los Angeles, with each number on the x-axis corresponding to a specific police division. For example, 0 represents the Central Division, while 1 corresponds to the Rampart Division. From Figure 6, it can be observed that Area 2, which represents the Southwest Division, has the highest crime count, while Area 11, representing the Northeast Division, has the lowest crime count.

**Conclusion**

This project successfully utilized unsupervised machine learning techniques, specifically clustering methods, to uncover hidden patterns within the Los Angeles crime dataset. Through systematic data preprocessing, dimensionality reduction, and clustering model comparison, meaningful groupings of crime incidents were identified based on time of day, victim demographics, crime type, and geographical distribution.

K-Means clustering with **k = 5** was selected after comparing multiple models under a supervised test on a small sample of 25,000 records. This model had the best cluster representation both visually and through the silhouette score. The K-Means clustering of the full dataset revealed important insights, such as the variation in victim age profiles across clusters, the dominance of specific crime types within each cluster, and the differences in crime distribution across police divisions in Los Angeles.

When comparing the cluster information of average age with the types of crimes occurring within those clusters, meaningful patterns emerged. For example, Cluster 2 had an average age above 40, indicating that senior citizens were primarily affected, making the most common crime in that cluster fraud more plausible. Similarly, Cluster 1, which included young children, showed a prevalence of domestic abuse-related crimes.

These findings could be instrumental in guiding law enforcement agencies in planning patrol strategies, allocating resources more effectively, and developing targeted crime prevention programs. This further highlights the power of unsupervised learning to extract actionable knowledge from complex real-world datasets.

Overall, this work demonstrates that with careful preprocessing and model evaluation, clustering techniques can provide valuable interpretations and practical recommendations even in domains where ground-truth labels are unavailable.

**Acknowledgements**

I would like to acknowledge the publicly available LAPD crime data hosted on Kaggle and maintained by Candace Gostinski, which provided the foundation for this analysis. Also I would like to acknowledge sources like google, YouTube, AI-assisted writing and coding tools and other internet sources with help for coding.