Background Information

In this project, I explored patterns of crime in Los Angeles, California, specifically examining the relationships among crime location, time of day, and victim age. With a close friend attending UCLA, my goal was to create a cluster analysis to identify areas he could avoid at certain times to enhance his safety. I used a dataset recording crime logs from 2020-2024, focusing on variables such as latitude, longitude, area code, victim age, and time of occurrence. Through analyzing these factors, I aimed to reveal any patterns that could provide insights into specific times and places to avoid.

Initially, I intended to work with either a neural network or a clustering problem, as most of my previous experience was with regression analysis in R. I saw this project as an opportunity to challenge myself with clustering techniques. After discovering the extensive LA crime dataset, which included numerous continuous numerical variables, I decided that K-means clustering would be the best approach. With Los Angeles being one of the more dangerous cities in the United States, this project held particular significance for my friend's safety during his freshman year.

Problem Statement

The purpose of this project is to cluster the crimes in Los Angeles based on time of day, geographic area, latitude, longitude, and victim age to uncover any underlying crime patterns. This is a clustering problem, using K-means clustering to identify and analyze these patterns geographically and demographically.

Methods

In starting the project, I identified latitude and longitude as key variables for clustering analysis. Using the info() function, I examined other numerical values that could potentially be used in the analysis and chose *TIME OCC*, *AREA*, and *Vict Age* as additional variables since they were all continuous numerical values. I anticipated that *Vict Age* would provide insights related to college-aged individuals, and *TIME OCC* would help identify times to avoid certain areas.

To understand the distributions of the selected variables, I first created histograms for *Vict Age*, *TIME OCC*, and *AREA*. I found that *Vict Age* included NA values stored as "Age = 0," so I removed these entries. *TIME OCC* and *AREA* had no significant outliers, so I retained all values. I also used a scatter plot with longitude and latitude to visualize crime locations and found that coordinates of (0,0) indicated missing values, which I removed. After these adjustments, I had a clean dataset ready for clustering.

To find the optimal *k* value for K-means clustering, I applied the elbow method. After scaling the data with StandardScaler(), I calculated the Within-Cluster Sum of Squares (WCSS) for values of *k* from 1 to 10. Initially, I chose *k = 4* based on the elbow plot, but after visualizing the clusters, I decided to explore larger *k* values. I used a for loop to add columns of k values 4-9 to my data set. Afterwards, I graphed the results with *k* values from 4 to 9 in a grid.

Results and Discussion

The K-means clustering algorithm proved to be the most effective method for segmenting geographic crime data. By grouping data based on longitude, latitude, and demographic characteristics, I was able to create distinct clusters associated with unique crime patterns across Los Angeles. Overlaying these clusters on a map of Los Angeles helped me identify specific areas within the city that may present varying degrees of risk based on age demographics and time of occurrence.

Initially, my scatter plots showed a more dispersed distribution than expected, making it hard to read. I later realized this was likely due to including *TIME OCC* as a variable, which led to overlapping data points across different times of day on the same streets/areas. To improve clarity, I adjusted my approach by calculating average *Victim Age* and *Time of Occurrence* within each cluster, which allowed me to assign meaningful labels based on victim demographics and timing patterns.

The silhouette coefficient was essential in confirming $k = 9$ as the optimal cluster count. Initially, I relied on the elbow method to select the k value, but found it unreliable due to the gradual slope in the WCSS plot. Testing k values from 4 to 9, I observed a silhouette coefficient of 0.65 at $k = 9$, indicating well-defined clusters. The silhouette coefficient provided a valuable quantitative measure of clustering quality, showing how compact and well-separated the clusters were. This was huge, as previously I had thought that $k = 4$ was the optimal cluster amount. Now confident in $k = 9$, I proceeded to label and analyze each cluster.

With $k = 9$, I calculated average values for *Victim Age* and *Time of Occurrence* within each cluster, assigning labels that captured demographic and temporal characteristics. The clusters are as follows: Cluster 0: *Evening Crimes - Younger Victims (~30 years)*, occurring around 4:47 PM, Cluster 1: *Midday Crimes - Young Adults (~33 years)*, occurring around 12:21 PM, Cluster 2: *Early Afternoon Crimes - Older Victims (~57 years)*, occurring around 1:21 PM, Cluster 3: *Midday Residential Crimes - Senior Victims (~61 years)*, occurring around 12:59 PM, Cluster 4: *Evening Social Crimes - Younger Victims (~30 years)*, occurring around 4:40 PM, Cluster 5: *Late-night Crimes - Young Adults (~33 years)*, occurring around 3:52 AM, Cluster 6: *Afternoon Crimes - Middle-aged Victims (~40 years)*, occurring around 1:30 PM, Cluster 7: *Late Afternoon Crimes - Younger Victims (~33 years)*, occurring around 4:35 PM and Cluster 8: *Early Afternoon Crimes - Older Victims (~58 years)*, occurring around 2:08 PM. From these labels, I identified the clusters most dangerous for a 19-year-old. Based on the results, Clusters 0, 4, and 7, which presented higher risks during the early evening and late afternoon in social areas; additionally, Cluster 5, which showed heightened crime risk in the early morning, was also flagged for caution. These insights allowed me to provide specific recommendations to my friend on which areas to avoid and when for enhanced safety.

Sources

https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/#:~:text=K%2Dmeans%20clustering%20is%20a%20popular%20unsupervised%20machine%20learning%20algorithm,or%20structures%20within%20the%20data.

https://getsafeandsound.com/blog/most-dangerous-city-in-america/

https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

https://www.geeksforgeeks.org/clustering-metrics/#silhouette-score

https://www.w3schools.com/python/python_for_loops.asp

https://www.digitalocean.com/community/tutorials/standardscaler-function-in-python