Assignment 1: Part D
(Comprehensive Final Report)

Lalitphan Sae-teoh a1932456

The University of Adelaide

COMP_SCI_7209 Big Data Analysis and Project

Trimester 2, 2025

Course Coordinator: Dr Hussain Ahmad

Tutor: Haider Ali Lokhand

Tutor: Manish

Table of Contents

Abstract **2**
1. Introduction **2**
2. Literature Review **3**
3. Research Methodology **5**
4. Experimental Evaluation **6**
4.1 Experimental Setup 6
4.1.1 Data Preprocessing 6
4.1.2 Data Exploration and Analysis 7
4.1.3 Feature Engineering 12
4.1.4 Model Implementation 12
4.2 Experimental Results 19
5. Discussion **20**
5.1 K-means Clustering Profile 20
5.2 Agglomerative Clustering Profile 21
5.3 Data Profiling Discussion 22
6. Limitations **22**
7. Conclusion **23**
8. Replication Package **23**
9. References **24**
10.Appendix **26**
10.1 SA Crime Level Type 26
10.2 Univariate Analysis - Crime Statistics SA Year Trends 27
10.3 Bivariate Analysis - Crime Statistics SA & Metro Median House Sales 28
10.4 Spatial Heatmaps 29
10.5 Evaluation Metrics 31
10.5.1 Silhouette Score 31
10.5.2 Calinski-Harabasz Index 31
10.5.3 Davies-Bouldin Index 31
10.6 K-Means Clusters Distribution & Trend Over Time 32
10.7 Agglomerative Clusters Distribution & Trend Over Time 33

Abstract

Urban inequality poses a significant challenge in rapidly growing cities, where disparities in crime rates and housing affordability shape the socio-economic landscape. Understanding these spatial dynamics is essential for evidence-based policymaking and urban planning. However, traditional analyses often fail to capture the complex, non-linear patterns underlying urban inequality. Many existing studies rely on a single analytical method, which limits their capacity to capture the complex interactions between crime rates, housing prices, and affordability across suburbs. To address this gap, this study employs multiple clustering techniques, including K-Means, DBSCAN, and Agglomerative Clustering, to explore socio-spatial inequalities across Adelaide suburbs. The models are evaluated using Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index, with results visualised through Principal Component Analysis. The study contributes by offering a comparative evaluation of clustering techniques within the context of urban inequality, highlighting how variation of algorithms reveals complementary insights into socio-spatial structures. The findings reveal distinct clusters, indicating that Agglomerative Clustering is particularly effective in identifying distressed areas characterised by high crime and low property values, while K-Means excels in market segmentation. These insights provide researchers and practitioners with valuable tools to address inequality hotspots and develop targeted urban planning strategies.

1. Introduction

Urban inequality is a growing concern in many Australian cities, including Adelaide, where both crime and housing affordability critically influence neighbourhood desirability, real estate investment, and residents' overall quality of life. Numerous studies have explored the relationship between local crime and property values. For instance, some researchers found substantial correlations between crime rates and housing prices across metropolitan areas (Margaretic, P & Sosa, JB 2025) [1]. In the UK, Motevali et al. (2025) observed that property prices were positively associated with non-violent crimes such as bicycle theft and car break-ins, while violent crime had a negative impact [2]. In the context of Los Angeles, de La Paz et al. (2022) showed that properties located further from crime incidents tended to have higher values, highlighting the spatial dimension of this relationship [3].

While many studies have investigated crime-price dynamics in major cities, fewer focus on medium-sized cities, such as Adelaide, where unique socio-economic patterns and geographic constraints may influence the relationship differently. Understanding the relationship between crime rates and property values in Adelaide's suburbs is essential for identifying areas that may be experiencing increasing socio-economic challenges. This research not only pinpointed the existence of these relationships but also their spatial distribution, highlighting possible "hotspots" of urban inequality. A better understanding of these patterns can help policymakers, urban planners, and community stakeholders address challenges related to affordability and safety more effectively, especially in regions experiencing rapid demographic or market shifts.

The core question leads to a more refined inquiry: **Are there spatial clusters where high crime rates and low housing affordability coincide, indicating zones of urban inequality?** To answer this, the project integrates spatial, temporal, and socio-economic datasets to examine the interaction between crime and housing dynamics in shaping urban inequality in Adelaide. Unlike broader studies that rely on aggregated crime indices, this research examines different types of crime while accounting for spatial and temporal dynamics. The findings are intended to inform a range of stakeholders, including policymakers, real estate professionals, and community advocates, by highlighting areas of socio-economic vulnerability and opportunity. Also, this study contributes to equitable urban development and evidence-based planning, providing a foundation for targeted safety interventions and housing strategies.

This research offers both methodological insights and practical value to the study of urban inequality. First, it designs and implements a comprehensive analytical framework that integrates multi-source datasets, including crime records, housing sales, property prices, and geographic boundaries, into a suitable spatially enabled processing pipeline. Second, it applies big data techniques for cleaning, normalising, and log-transforming skewed variables, ensuring statistical robustness and enabling cross-variable comparability. To identify socio-spatial patterns, advanced clustering techniques, such as K-Means, DBSCAN, and Agglomerative Clustering, are deployed, along with Principal Component Analysis (PCA) for cluster visualisation. Third, the evaluation utilised clustering validity measurements, including the Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index, thereby enabling evidence-based model selection. The framework offers a flexible and scalable approach that can be applied to other cities, supporting data-driven urban policy and promoting long-term socio-economic resilience.

2. Literature Review

The notion of urban inequality, especially in relation to housing and crime, is significant. Research conducted in major cities like Barcelona has demonstrated that areas perceived as less safe tend to have lower housing prices (Buonanno et al. 2013) [13]. Similarly, studies in England emphasise a notable relationship between street-level criminal offenses and property values (Braakmann 2017) [14]. Additionally, Baird (2020) contributes to this discourse by examining how public and private investments impact residential and commercial property prices, which can ultimately lead to a decrease in crime rates [15].

Methodological studies are increasingly utilising spatial clustering and data science techniques to gain insights into urban inequality. Cueva et al. (2024) proposed employing Time-Series Decompositions, Spatial Autocorrelations, and Regression to measure urban criminal activity, influenced by various socio-economic indicators [16]. These methods facilitate a deeper exploration of spatial areas and help analyse complex spatial patterns. Additionally, advanced machine learning models, such as decision trees and random forests, are gaining traction. In the UK, Motevali et al. (2025) have applied these machine learning techniques to investigate the relationship between house prices and crime rates [2].

Table 1: Literature Review Comparison

Author	Title	Datasets	Techniques	Findings
Buonanno, P, Montolio, D & Raya-Vílchez, JM (2013)	Housing Prices and Crime Perception	Barcelona - Crime Statistics and Property Sales (2004 - 2006)	Ordinary Least Squares (OLS) and Quantile Regressions	The findings found that in less safe areas, house rates are lower.
Braakmann, N (2017)	The Link between Crime Risk and Property Prices in England and Wales: Evidence from Street-Level Data	UK - Crime Statistics and Property Sales (2011 - 2013)	Regression	The results of the relationship between street-level criminal offences and property prices are significant.
Baird, MD, Schwartz, H, Hunter, GP, Gary-Webb, TL, Ghosh-Dastidar, B, Dubowitz, T	Does Large-Scale Neighborhood Reinvestment Work? Effects of Public-Private Real Estate Investment on	Pittsburgh, Pennsylvania - Property Sales, Property Rental, Crime Statistics, and Public	Regression	The significant effects of public–private investments on residential and commercial prices can lead to reduced crime rates.

& Troxel, WM (2020)	Local Sales Prices, Rental Prices, and Crime Rates	Investment (1990 - 2015)		
Cueva, D & Cabrera-Barona, P (2024)	Spatial, Temporal, and Explanatory Analyses of Urban Crime	Quito, Ecuador – Crime Statistics and Socio-Economic Indicators (2014 - 2020)	Time-Series Decompositions, Spatial Autocorrelations, and Regression	The findings revealed that urban criminal activity exhibits distinct spatial and temporal characteristics, and incidents of crime can be understood in relation to urban socioeconomic factors.
Motevali, S, Aljawawdeh, H, Abuezhayeh, S & Qaddoumi, E (2025)	Explore the Relationship between House Prices and Crime Rate in the UK Using Machine Learning Techniques	UK - Crime Stats and Property Sales (2020 – 2022)	Decision Tree and Random Forest	The results of the relationship between different kinds of crime and the characteristics of the housing market.

The reviewed literature highlights significant progress in both theoretical and methodological approaches for studying urban inequality. Theoretical contributions emphasise the relationship between crime and house rates, often lacking spatial analysis. Meanwhile, methodological developments primarily focus on applying regression techniques and socio-economic indicators to explore urban housing and crime contexts. However, much of this research tends to either generalise inequality patterns across metropolitan areas or examine single dimensions, such as infrastructure access or housing prices, without simultaneously analysing crime and affordability. In contrast, this study uniquely integrates clustering methods with a dual focus on crime rates and housing affordability within Adelaide's suburban landscape. Furthermore, it utilises clustering techniques to identify socio-spatial inequalities and urban segmentation. This approach not only enhances existing theoretical frameworks but also addresses methodological gaps by combining clustering with PCA visualisations and profiling based on recent data from 2024, offering a more localised and multidimensional perspective on urban inequality research.

3. Research Methodology

Clustering techniques are utilised to conduct spatial clustering for this research. Clustering is an essential unsupervised learning technique that helps identify patterns and groupings within multivariate datasets. This research aims to categorise suburbs in South Australia based on aggregated crime statistics and housing market indicators. For this purpose, three clustering models have been selected for comparison, including K-Means, DBSCAN, and Agglomerative clustering. They were chosen due to their diverse strengths and suitability for different cluster shapes, densities, and levels of interpretability. The research involves six phases: Data Preprocessing, Data Exploration and Analysis, Feature Engineering, Model Implementation, Model Evaluation, and Data Profiling.

Figure 1: 6 Phases of Research Process

The first phase, Data Preprocessing, contains data ingestion, cleaning, and integration. To conduct the analysis, three data sources must be integrated: crime statistics, housing markets, and geographic boundaries, each possessing distinct data sources and types. All datasets were sourced from data.sa.gov.au [4], which compiles data from various organisations within South Australia. After ingestion, each dataset underwent initial cleaning, aggregation, and preparation for subsequent processing stages.

Next, the Data Exploration and Analysis phases are executed to uncover insights from each dataset and examine the relationships between crime statistics, housing rates, and suburb locations. This stage is crucial for identifying key features necessary for the feature engineering and modelling phase, where both univariate and bivariate analyses are conducted through statistical methods and visualisations.

Following this, the Feature Engineering phase prepares the data for model implementation, ensuring statistical robustness and facilitating cross-variable comparability. The Model Implementation phase involves the independent execution of various clustering techniques, from which the best-performing model for each technique is selected for comparative performance assessment, particularly in terms of how effectively it clusters crime and housing rates across Adelaide suburbs. The goal is to identify the most meaningful clusters for each technique.

Subsequently, the Model Evaluation phase assesses the identified clusters using metrics such as the Silhouette Score, Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI). Finally, the Data Profiling phase focuses on profiling the best-performing clustering model from the implementation phase to address the research question.

4. Experimental Evaluation

This section contains the Experimental Setup and Experimental Results. The Experimental Setup details the process from the initial stages through to model implementation, with the goal of achieving optimal clustering performance for comparison in the Experimental Results. The Experimental Results examine the clustering outcomes and identify the most effective clustering method for Data Profiling.

4.1 Experimental Setup

The experimental setup encompasses four phases: data preprocessing, data exploration and analysis, feature engineering, and model implementation.

4.1.1 Data Preprocessing

There are three primary data sets to be ingested and integrated, as shown in Table 2. Crime statistics refer to the statistics on crimes committed against individuals and property within the Adelaide suburbs. All crimes against people and property reported to the police during the specified financial year are included in the crime statistics, which cover the period from 2010 to 2025 (Q1-Q3). It originated from the South Australian Police (SAPOL) department. Metro Median House Sales will represent the primary property price data, which is the quarterly median house prices for metropolitan Adelaide by suburb, sourced from the Department for Housing and Urban Development. It is available for almost 10 years of historical data from Q1 2015 to Q1 2025. Lastly, Geographic Boundaries data is used for visualisation to explore the relationship between crime rate and housing price. The primary source of this data set is Suburb data from Department for Housing and Urban Development. As they have different histories, the analysis scope of data for this project will be 9 years from Q1 2015 to Q4 2024. The data processing tasks are defined in Table 3.

Table 2: Data Sources

Data Type	Dataset	Source	Data Format
Crime Data	Crime Statistics SA	data.sa.gov.au	CSV
Properties Price	Metro Median House Sales	data.sa.gov.au	XLSX (excel file)
Geographic Boundaries	Suburb	data.sa.gov.au	CSV, GeoJson, Shapefile

Table 3: Data Processing Tasks

Data Processing Tasks	Description
Data Ingestion	Download and aggregate data sets.
Data Cleaning	Handle missing values, manipulate data, fix format, and normalise suburb locations.
Data Integration	Join all data sets at the suburb level and align them into the same time frames (monthly or quarterly).

After ingestion, each data source was initially cleaned to remove missing records, remove specific columns, and rename columns. Table 4 shows the initial cleaning results of each data source.

Table 4: Cleaned Data Sets Dictionary

Data Source Name	Column Name	Description	Data Type	Example
SA Crime Statistics	report_date	Report Date	Date	1/07/2013
	suburb	Suburb Name	String	ABERFOYLE PARK
	postcode	Postcode	String	5159
	offence_lv1	Offence Level 1 Description	String	OFFENCES AGAINST PROPERTY [Appendix 10.1]
	offence_lv2	Offence Level 2 Description	String	SERIOUS CRIMINAL TRESPASS [Appendix 10.1]
	offence_lv3	Offence Level 3 Description	String	SCT - Non Residence [Appendix 10.1]
	offence_count	Offence count	Float	3.0
SA House Price	city	City Name	String	ADELAIDE HILLS
	suburb	Suburb Name	String	ALDGATE
	sales	Sales	Float	17.0
	median	Median	Float	685000.0
	median_change	Median Change	Float	0.132231
	year	Year	Integer	2015
	quarter	Quarter (Not Australian Financial Year)	String	Q1

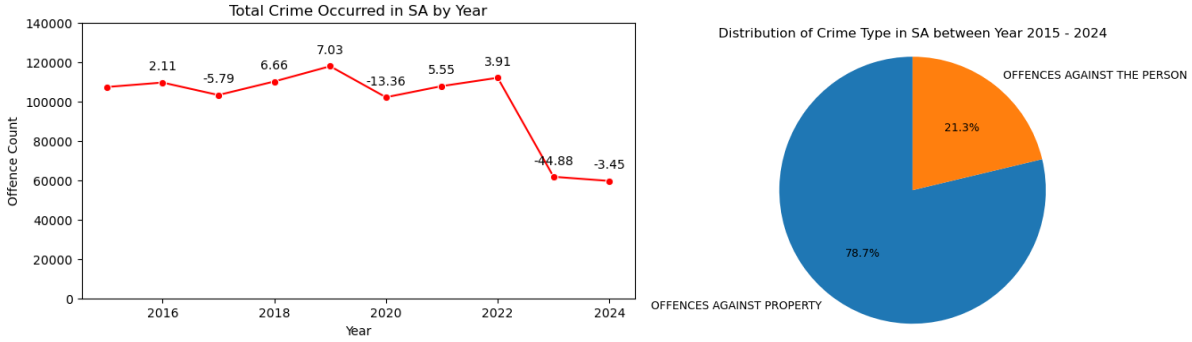
During data integration, the key variables, including total offence count, median housing price, and property sales, are integrated with suburb information and the same time frame. The suburb's name is standardised and ensures that there is no whitespace. The missing values are removed. Data is queried between Q1 2015 and Q4 2014 for a yearly full-time frame.

4.1.2 Data Exploration and Analysis

The exploration and analysis aim to identify insights from each dataset and the relationship between Crime Statistics, House Prices, and Suburb location. To find the essential features for feature engineering and the model implementation stage.

Start with the analysis of crime statistics displayed in Figure 2, which found that the overall rate has decreased significantly over time since 2022, particularly in 2023, with a decline of 44.88%. Almost 80% of crimes that occur in South Australia are property offences, and about 20% are offences against the person. For further analysis of the SA crime rate trends by offence level, in detail, can be found in [Appendix 10.2].

Figure 2: Total Crime Occurred in SA by Year & Distribution of Crime Type in SA



For offences against property, as shown in Figure 3, theft and property damage are the top two crimes found in South Australia, accounting for 25% and 21.3%, respectively. Shoplifting accounts for approximately 13%, and motor vehicle theft accounts for 11%. At the same time, Figure 4 shows that serious assaults without injury are almost half the proportion that happen for offences against the person and followed by common assault with about 24%.

Figure 3: Number of Offences against property in SA by offence type level 3

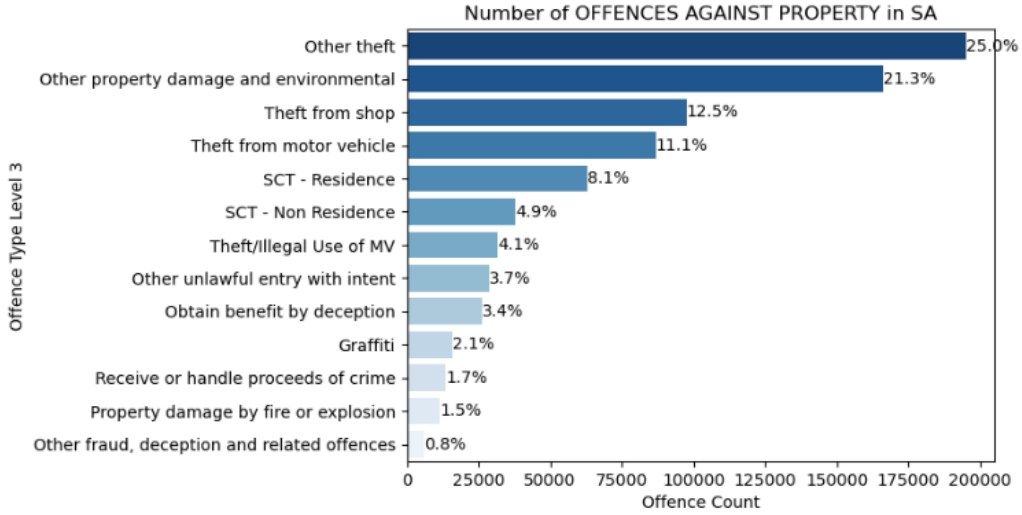
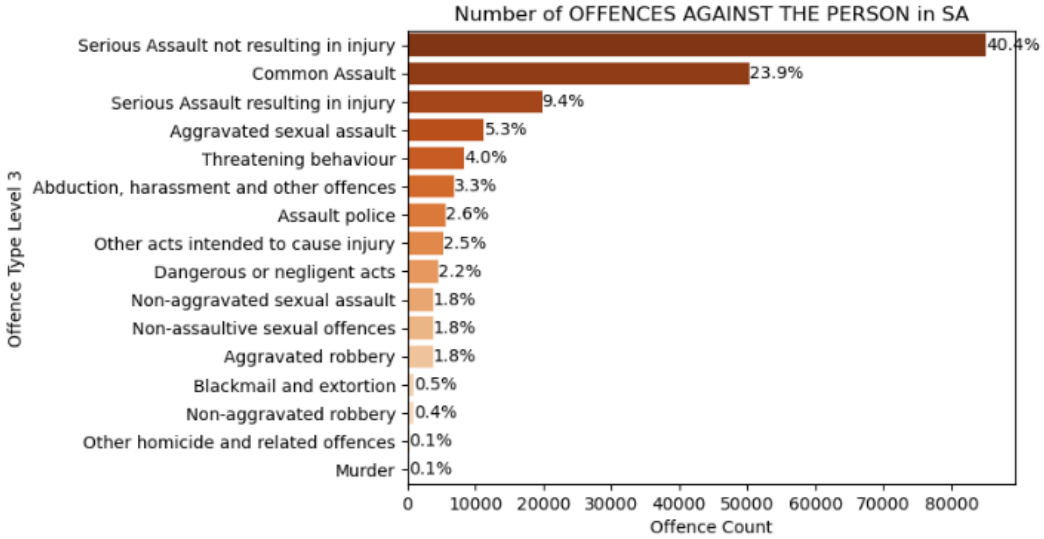
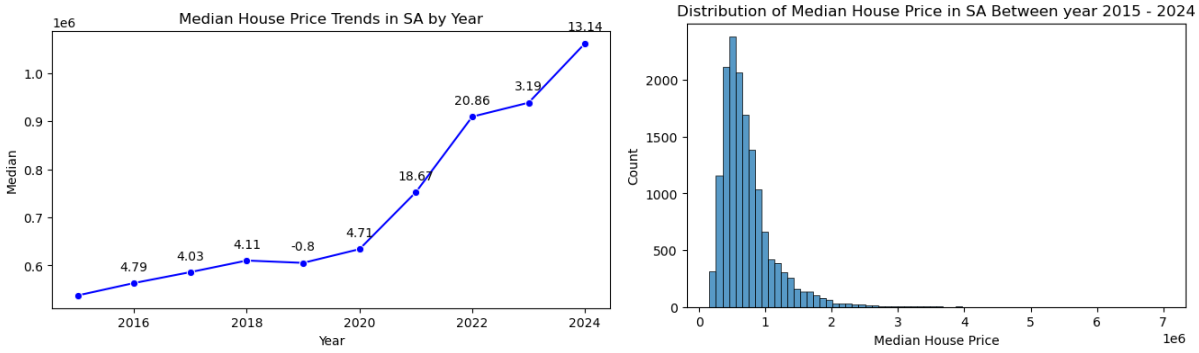


Figure 4: Number of Offences against property and the person in SA by offence type level 3



Next, the analysis of house rates in SA, as shown in Figure 5, revealed a significant increase since 2020, with increases of approximately 19% in 2021, 21% in 2022, 3% in 2023, and 13% in 2024. Additionally, the distribution of median house prices in SA is right-skewed, with an average house price of approximately $700,000 AUD. The maximum house price is around $7 million AUD, while the minimum house price is around $155,000 AUD.

Figure 5: Median House Price Trend in SA by Year & Distribution between the year 2015 – 2024



The boxplot of median house price (Figure 6) shows the range of median house prices compared to cities in SA. Median house prices in some cities exhibit slight variations, for example, in Mount Barker, Playford, Gawler, and Tea Tree Gully. In contrast to some cities that have significant differences in house prices within the area, such as Norwood Payneham & St Peters, Walkerville, Burnside, and Unley, etc. While the housing market (Figure 7) in Onkaparinga experienced the most considerable number of house sales between 2014 and 2015, with an increase of approximately 16%, this indicates the impact of population growth on house sales. The following are the most popular areas: Salisbury and Port Adelaide Enfield, with 11.2% and 11.0%, respectively. Due to a slight increase in house prices in the top 3 house sales areas, as indicated by the heatmap in Figure 8, this may make them more popular for house sales.

Figure 6: Boxplot of Median House Price in SA between the year 2015 – 2024 by City

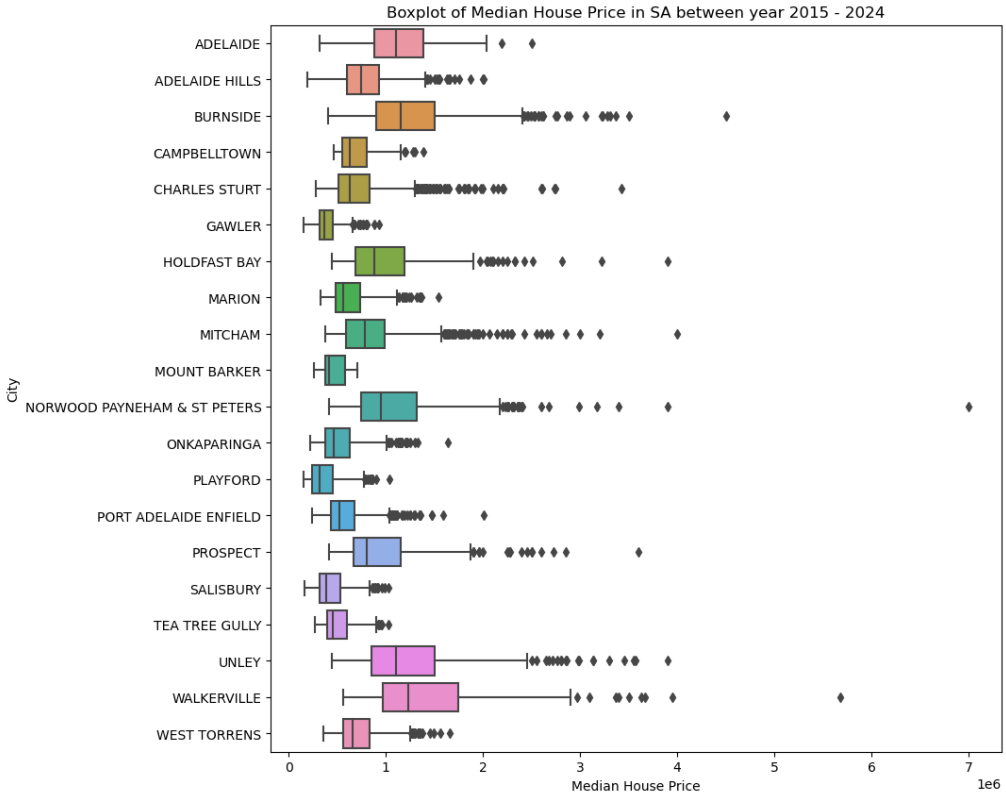
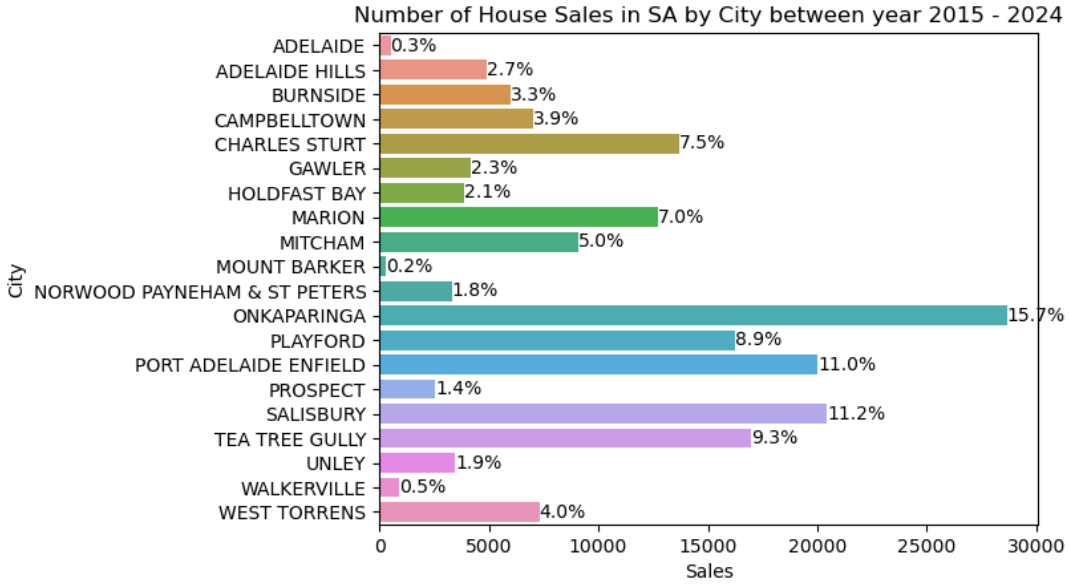
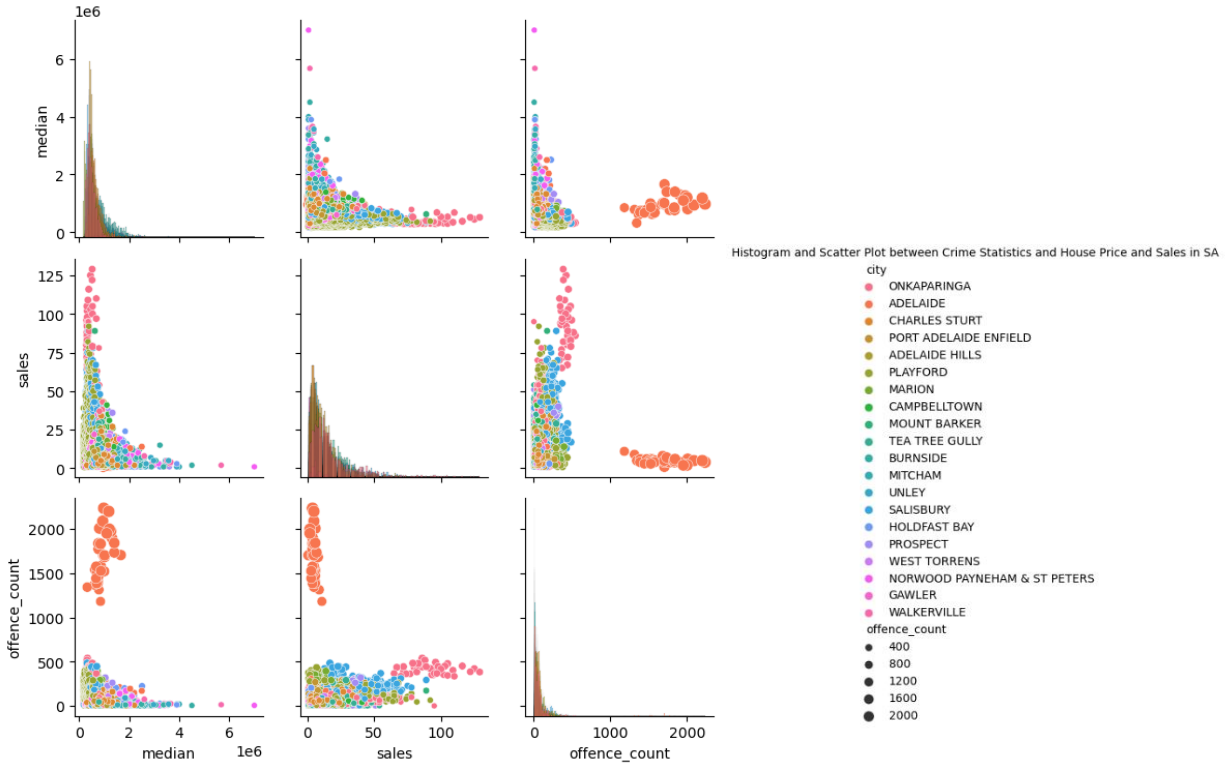


Figure 7: Number of House Sales in SA between the year 2015 - 2024 by City



When analysing crime statistics, house prices, and sales, the pairplot in Figure 8 revealed a clear negative correlation between offence count and both property prices and sales. Suburbs with higher crime rates, such as the Adelaide CBD and parts of Playford, tend to exhibit lower housing prices and reduced market activity. In contrast, affluent eastern suburbs such as Burnside, Unley, and Mitcham consistently exhibit low crime levels alongside high property values. Further correlation analysis, see [Appendix 10.3].

Figure 8: Histogram and Scatter Plot between Crime Statistics and House Price, and Sales in SA



Spatial heatmaps further reinforced these findings [Appendix 10.4]. The crime statistics map in Figure 9 showed a concentration of criminal activity in central and northern cities, whereas the housing price heatmap highlighted elevated property values in the eastern and coastal regions. Despite extending this visual analysis across multiple years, the spatial distribution of both crime (Figure 9) and property values (Figure 10) remained remarkably stable over time, suggesting persistent socio-economic patterns and limited mobility across housing or safety indicators.

Figure 9: SA Crime Statistics Heatmap over time between 2018 – 2020

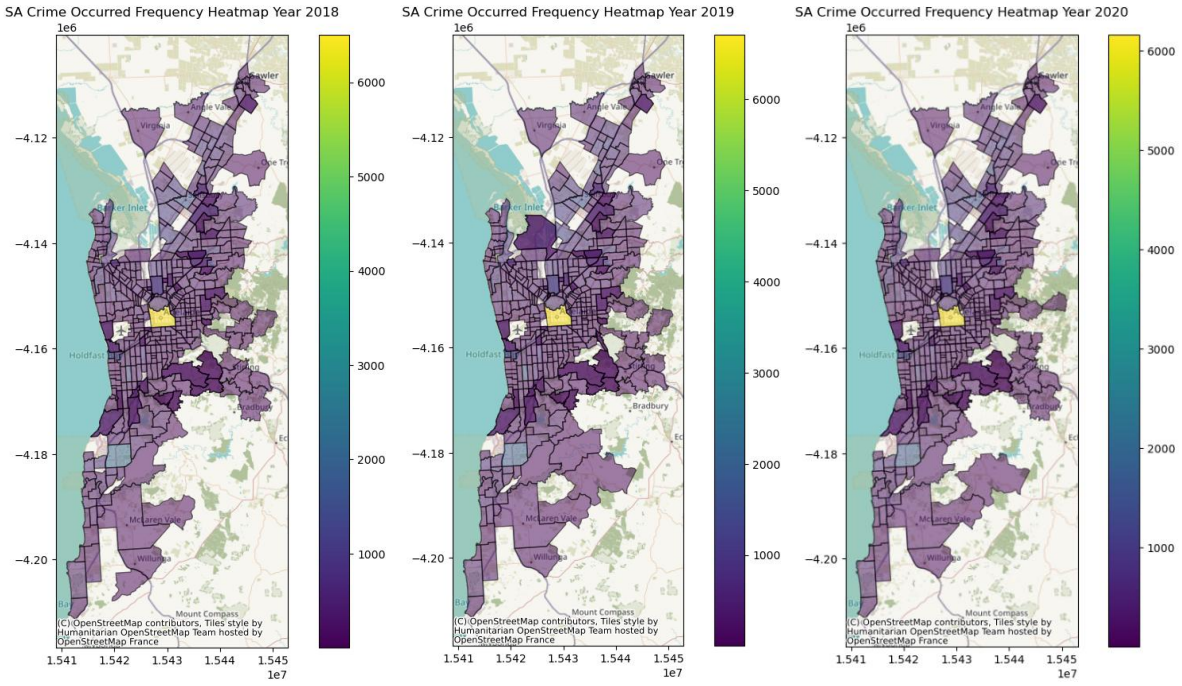
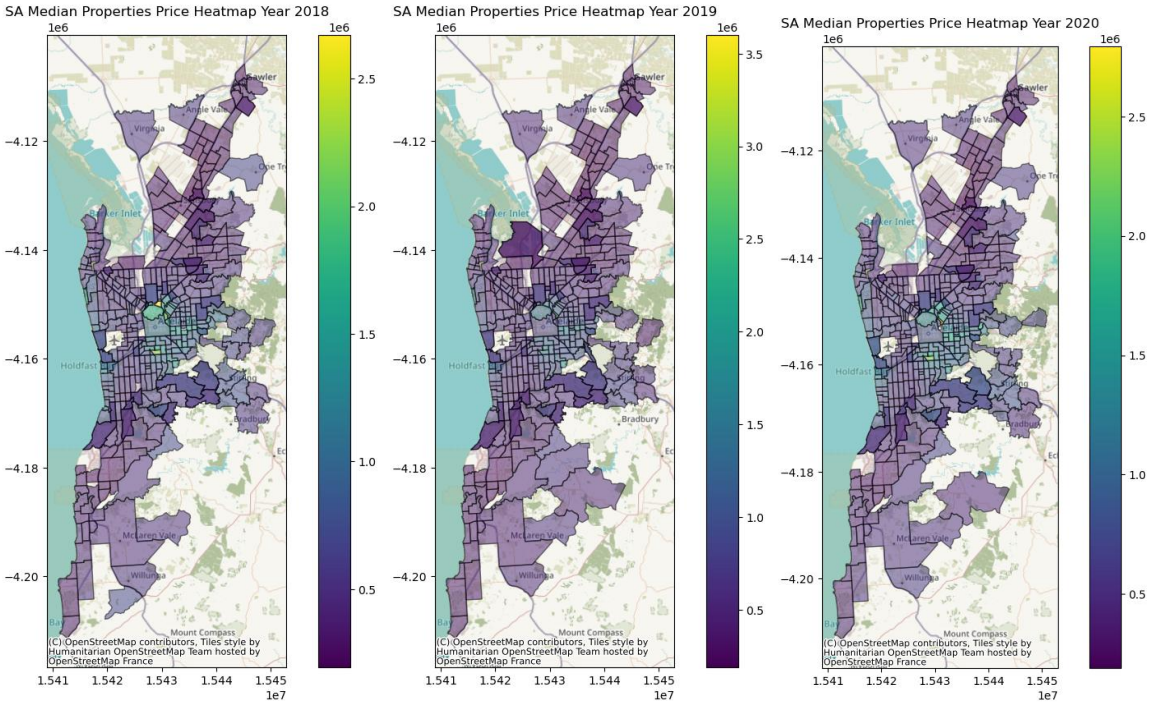


Figure 10: SA Median House Price Heatmap over time between 2018 - 2020



4.1.3 Feature Engineering

As the spatial analysis results in the Data Exploration and Analysis stage, trends remained consistent over multiple years. Therefore, all key features, including total offence count, median housing price, and property sales, are aggregated yearly. Offence counts are aggregated to level 1, which includes offences against property and offences against the person, because they have different trends.

For cluster features preparation, the aggregate data are pivoted to key variables and year. All features are applied log-transformed to correct for skewed distributions and normalised in the Min-Max scalar range between 0 and 1 to prevent being sensitive to differences in scale distance and outliers in clustering, especially those that rely on calculating means or distances. Outliers can significantly distort the cluster centres and boundaries, which might lead to inaccurate clustering results.

4.1.4 Model Implementation

Clustering techniques are used to perform spatial clustering for this research. Clustering is an essential unsupervised learning technique used to identify patterns and groupings within multivariate datasets. We aim to categorise suburbs in South Australia based on aggregated crime statistics and housing market indicators, three clustering models were selected for comparison: K-Means, DBSCAN, and Agglomerative clustering. These models were chosen for their complementary strengths and suitability to different cluster shapes, densities, and interpretability.

K-Means is one of the most widely used centroid-based clustering algorithms due to its simplicity, scalability, and speed. The algorithm partitions the dataset into k clusters by minimising the sum of squared distances between each point and its assigned cluster centroid (MacQueen, 1967) [6]. The strengths of K-Means are fast and suitable for well-separated subclusters. However, K-Means assumes clusters are spherical and evenly sized, which may limit its performance when applied to complex spatial distributions.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that identifies clusters as high-density regions separated by areas of low density (Ester, M., Kriegel, H.-P., Sander, J., & Xu, X., 1996) [7]. Unlike K-Means, DBSCAN does not require specifying the number of clusters in advance and can detect clusters of arbitrary shapes. It also identifies outliers or noise points, which are valuable when working with urban data that may include irregular suburbs or inconsistent statistics. Therefore, it might be useful for identifying core clusters of similar suburban profiles and distinguishing outlier suburbs with unique characteristics. Regardless, DBSCAN parameters are sensitive, including the neighborhoods radius (eps) and the minimum number of points required to form a dense region (min_samples).

Agglomerative Clustering is a bottom-up hierarchical clustering algorithm that builds clusters iteratively by merging the closest pairs based on a linkage criterion, such as Ward's method, average linkage, or complete linkage (Müllner, 2011) [8]. The result is a dendrogram, a tree-like structure, that reveals the hierarchical relationships between clusters and allows for flexible selection of the number of clusters based on dissimilarity thresholds. For this assignment, agglomerative clustering might provide valuable insights by revealing how suburbs group progressively. However, Agglomerative computation is expensive.

Evaluating the performance of clustering models is essential to ensure that the generated groupings are meaningful, well-separated, and compact. Unlike supervised learning, clustering does not rely on ground truth labels. Instead, internal validation metrics are used to assess the quality of the clustering structure. The Silhouette Score, Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI) are widely used evaluation metrics for clustering models, which are employed for performance comparison between the model selections in this research. [Appendix 10.5]

Each clustering technique is implemented independently, and the best-performing model from each technique was selected to compare its performance with that of another method and assess how well it clusters the Crime-House rate in Adelaide suburbs. The objective was to identify at least three meaningful clusters per technique. The potential clusters are evaluated, where optimal clustering should yield a high Silhouette Score, a high CHI, and a low DBI. Furthermore, to better understand the spatial and statistical separation of the clusters, Principal Component Analysis (PCA) was used to reduce dimensionality and visualise the cluster distribution in a 2D space. This visualisation helps to observe how well the clusters are separated and whether there is any significant overlap. Finally, each selected clustering result was further profiled using the most recent available data, specifically for 2024.

4.1.4.1 K-means Implementation

The implementation of K-Means clustering began with determining the optimal number of clusters (k) value using the elbow method. This approach involves plotting the Within-Cluster Sum of Squares (WCSS) against a range of K values and identifying the point where the rate of decrease sharply slows down, which results in an elbow shape on the graph. In this case, the elbow plot indicated the potential clusters are 3, 4, 5, and 6.

Figure 11: Visualisation of the elbow method to find optimal K for K-Means

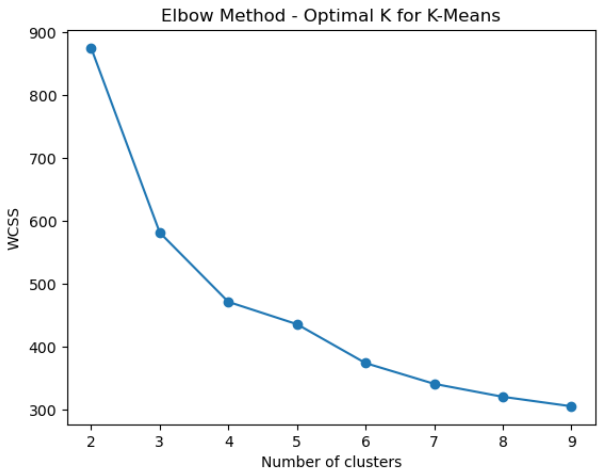


Table 5: K-Means Clusters Evaluation Results

Number of Clusters	Silhouette Score	CHI	DBI
3	**0.657**	**5662.970**	**0.666**
4	**0.601**	**4794.394**	**0.867**
5	0.588	3922.968	1.003
6	0.447	3714.324	1.119

According to the evaluation results in Table 5, clusters with optimal k values of 3 and 4 are selected for comparison to assess their effectiveness in clustering.

Figure 12: 2D PCA Visualisation of K-Means Clustering

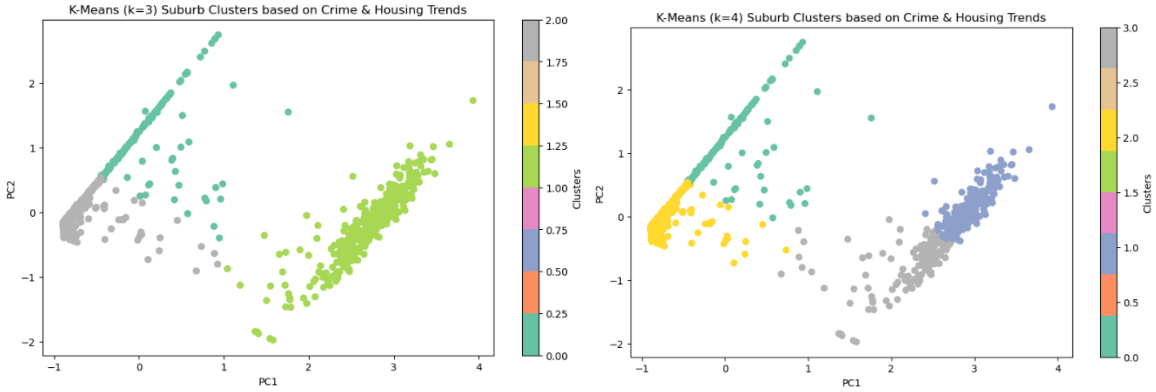


Table 6: K-Means (k=3) Clusters Profile

Variables	Statistics	Cluster 0	Cluster 1	Cluster 2
Suburbs	Count	164 (9%)	362 (20%)	512 (28%)
Crime Against Property	Average	56.75	94.86	2.46
	Median	20.5	59.5	1.0
	Min	0	1	0
	Max	1016	3048	34
Crime Against Person	Average	29.68	25.84	0.91
	Median	7.0	13.0	0
	Min	0	0	0
	Max	1154	1101	8
House Sales	Average	0.82	51.14	0.19
House Price	Average	$ 55.6 k	$ 1.1 million	$ 14.9 k

Table 7: K-Means (k=4) Clusters Profile

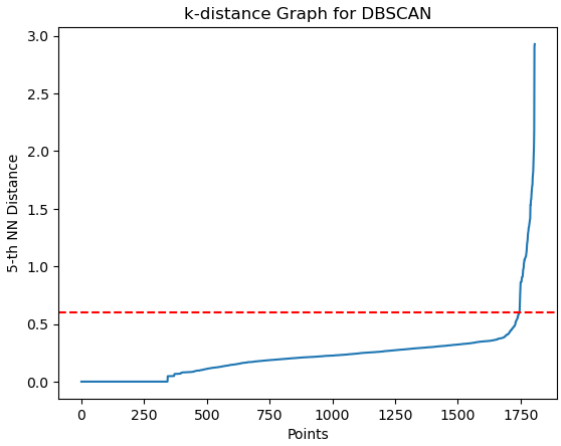
Variables	Statistics	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Suburbs	Count	161 (9%)	232 (13%)	510 (28%)	135 (7%)
Crime Against Property	Average	57.53	134.33	2.43	23.99
	Median	21.0	86.5	1.0	20.0
	Min	0	20	0	1
	Max	1016	3048	34	103
Crime Against Person	Average	30.17	37.67	0.92	4.58
	Median	7.0	22.5	0	4
	Min	0	5	0	0
	Max	1154	1101	8	24
House Sales	Average	0.81	69.15	0.08	18.76
House Price	Average	$ 43.2 k	$ 915 k	$ 13.4 k	$ 1.35 million

The above PCA visualisation and cluster profile results indicate that K-Means with 4 clusters has more interesting insights, although the evaluation results of 3 clusters are better. Clusters 1 and 3 in K-Means with 4 clusters differ from those in K-Means with 3 clusters. They capture the suburbs that have high median house prices but different crime rates, where cluster 1 has a high crime rate, while cluster 3 has a low crime rate. Therefore, K-Means with 4 clusters is the selected model for model comparison in the next step.

4.1.4.2 DBSCAN Implementation

The implementation of DBSCAN clustering began with determining the optimal eps value, which is the parameter to indicate maximum distance between points, using the k-distance graph to find the elbow point where the slope sharply increases. As DBSCAN is a distance-based model, it is susceptible to scaling. If we didn't tune the eps parameter well, the model might result in one big cluster for too large an eps, on the other hand, too much noise for an eps that is too small. The k-distance graph below indicates that the optimal value of eps for DBSCAN is between 0.4 and 0.7.

Figure 13. Visualisation k-distance for DBSCAN



Not only eps that need to be tuned, but also min_samples, which are the minimum points to form a dense region parameter. After we get the ranges of potential optimal eps, we experiment with tuning these parameters with several min_samples to assess the best performance.

Table 8: DBSCAN Clusters Performance

Eps	Min Samples	Number of Clusters	Silhouette Score	CHI	DBI
0.4	3	5	0.544	1450.405	1.198
0.4	**5**	**3**	**0.665**	**2366.689**	**1.182**
0.4	8	2	0.679	3446.105	1.278
0.5	3	6	0.553	1272.367	1.285
0.5	**5**	**4**	**0.637**	**1879.643**	**1.200**
0.5	8	2	0.683	3602.531	1.284
0.6	3	5	0.580	1541.687	1.347
0.6	5	5	0.583	1536.697	1.347
0.6	8	2	0.680	3601.135	1.327
0.7	3	4	0.594	1859.827	1.432
0.7	5	4	0.594	1859.827	1.432
0.7	8	2	0.676	3587.539	1.337

DBSCAN with 0.4 eps and 5 min_samples and DBSCAN with 0.5 eps and 5 min_samples were selected to visualise the clustering and profiling. Although they do not have the best performance in Table 8 results. However, they provide at least three clusters and have better performance compared to other parameters with the same cluster results.

Figure 14: 2D PCA Visualisation of DBSCAN Clustering

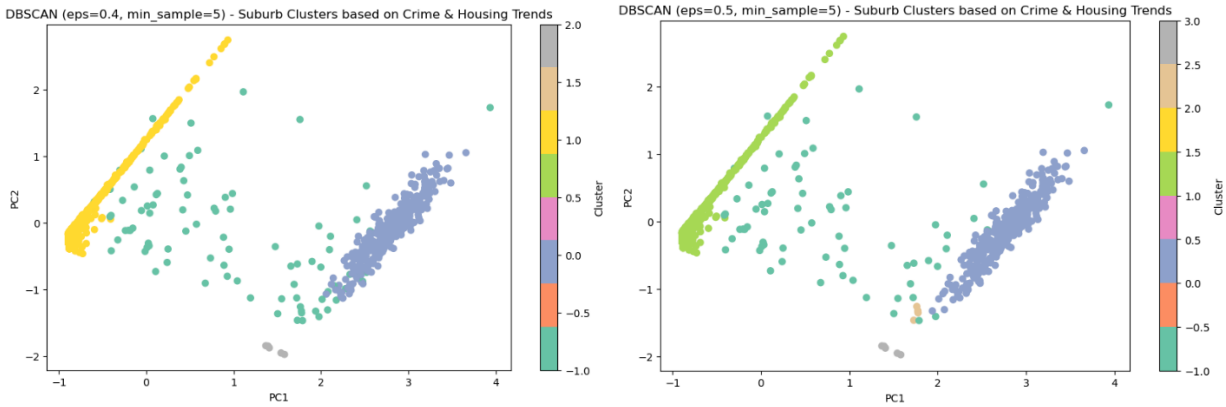


Table 9: DBSCAN (eps=0.4, min_samples=5) Clusters Profile

Variables	Statistics	Cluster -1 (Noise)	Cluster 0	Cluster 1
Suburbs	Count	88 (5%)	326 (18%)	624 (35%)
Crime Against Property	Average	61.94	92.79	14.75
	Median	7	64	2
	Min	0	4	0
	Max	3048	704	1016
Crime Against Person	Average	31.84	24.80	6.09
	Median	2	14	1
	Min	0	0	0
	Max	1154	181	436
House Sales	Average	11.47	54.41	0
House Price	Average	$ 681.6 k	$ 1.1 million	$ 0

Table 10: DBSCAN (eps=0.5, min_samples=5) Clusters Profile

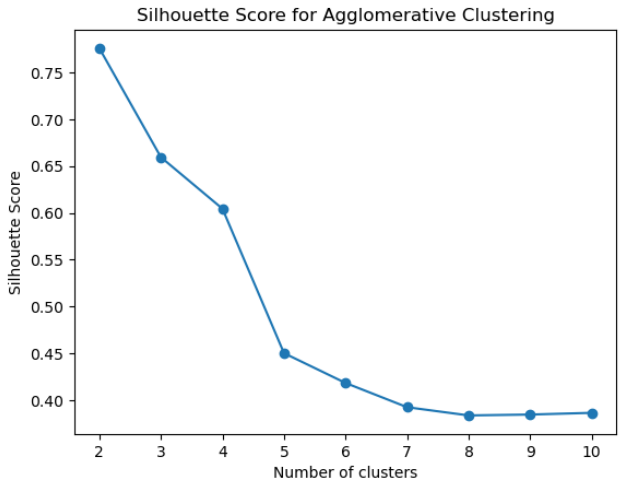
Variables	Statistics	Cluster -1 (Noise)	Cluster 0	Cluster 1	Cluster 2
Suburbs	Count	69 (4%)	338 (19%)	627 (35%)	4 (0.2%)
Crime Against Property	Average	72.49	90.56	14.81	1.75
	Median	7	62.5	2	2
	Min	0	1	0	1
	Max	3048	704	1016	2
Crime Against Person	Average	39.64	24.10	6.08	2.00
	Median	2	13.5	1.0	2.0
	Min	0	0	0	1
	Max	1154	181	436	3
House Sales	Average	8.32	53.71	0	4.25
House Price	Average	$ 515.0 k	$ 919.7 k	$ 0	$ 1.1 million

The final model for DBSCAN is the one with 0.4 eps and 5 min_samples, as it accurately represents the proportion of each cluster and clusters are well-represented in the PCA visualisation. In contrast with the profile of cluster 2 in DBSCAN 0.5 eps and 5 min_samples are too small.

4.1.4.3 Agglomerative Implementation

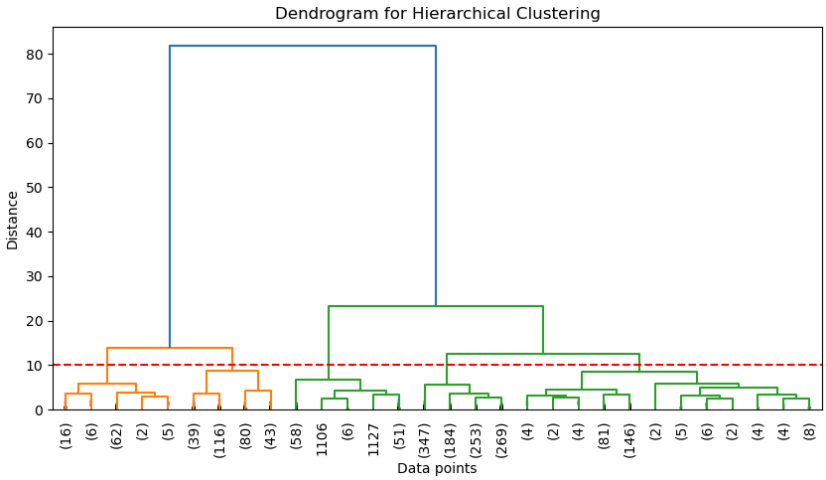
To determine the optimal number of clusters (k) for Agglomerative Clustering, several effective methods have been developed. Unlike K-Means, Agglomerative Clustering does not use the "WCSS" metric. However, we can utilise the silhouette score, dendrogram, and Calinski-Harabasz (CHI) or Davies-Bouldin (DBI) scores to determine the optimal number of clusters.

Figure 15: Silhouette Score for Agglomerative Clustering



Another way to visualise hierarchical clustering is through dendrograms, which provide visual clues to the optimal k. The number of vertical lines the red line cuts is the optimal cluster k.

Figure 16: Dendrogram of Agglomerative (Hierarchical) Clustering



However, identifying the optimal k with Silhouette Score, Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI) is easier. Figure 15 shows that the optimal k value for Agglomerative clustering is between 3 and 7.

Table 11: Agglomerative Clusters Performance

Number of Clusters	Silhouette Score	CHI	DBI
3	**0.659**	**5336.091**	**0.548**
4	**0.604**	**4346.696**	**0.811**
5	0.450	3920.592	1.159
6	0.418	3466.192	1.215
7	0.392	3202.230	1.290

According to the evaluation results in Table 11, clusters with optimal k values of 3 and 4 are selected for comparison to assess their effectiveness in clustering.

Figure 17: 2D PCA Visualisation of Agglomerative Clustering

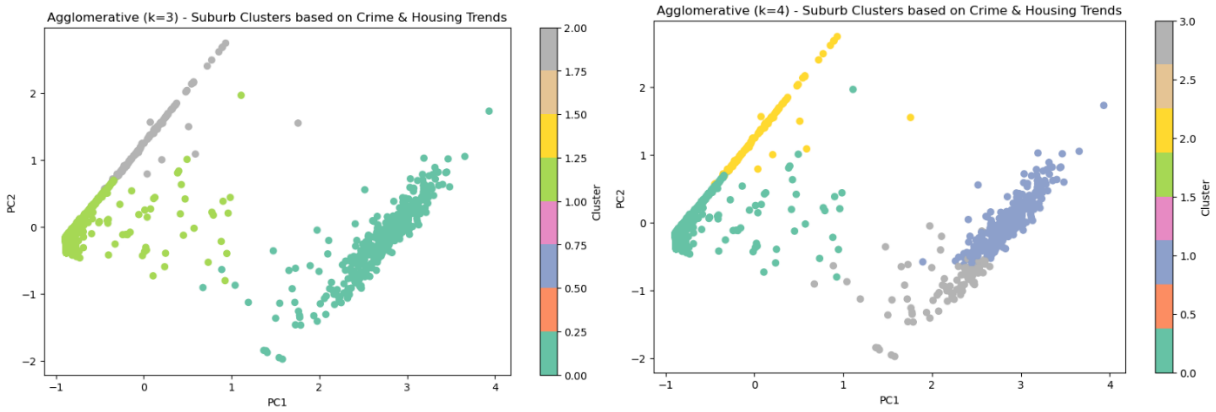


Table 12: Agglomerative (k=3) Clusters Profile

Variables	Statistics	Cluster 0	Cluster 1	Cluster 2
Suburbs	Count	364 (20%)	557 (31%)	117 (6%)
Crime Against Property	Average	94.35	3.66	72.82
	Median	59	2	27
	Min	1	0	0
	Max	3048	300	1016
Crime Against Person	Average	25.70	1.27	39.53
	Median	13	1	10
	Min	0	0	0
	Max	1101	80	1154
House Sales	Average	50.86	0.41	0.05
House Price	Average	$ 1.1 million	$ 27.8 k	$ 11.0 k

Table 13: Agglomerative (k=4) Clusters Profile

Variables	Statistics	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Suburbs	Count	557 (31%)	278 (15%)	117 (6%)	86 (5%)
Crime Against Property	Average	3.66	117.49	72.82	19.56
	Median	2	75.5	27	15
	Min	0	9	0	1
	Max	300	3048	1016	113
Crime Against Person	Average	1.27	32.66	39.53	3.19
	Median	1	19	10	3
	Min	0	1	0	0
	Max	80	1101	1154	19
House Sales	Average	0.41	60.09	0.05	21.03
House Price	Average	$ 27.8 k	$ 971.6 k	$ 10.9 k	$ 1.4 million

The PCA results of Agglomerative clustering with k = 3 and k = 4 are almost identical to those of K-Means with k = 3 and k = 4. However, the profile is quite different. Agglomerative with k=4 has more interesting insight than k=3, where cluster 3 has the pattern of high median house price with low crime rate. Therefore, the Agglomerative with k = 4 will be compared with the final models of K-Means and DBSCAN.

4.2 Experimental Results

This section is the Model Evaluation phase, where we compare the best clusters from each clustering technique and select the optimal clusters to proceed to the final stage, which is the Data Profiling phase, to address the research question.

The clustering performance was assessed using three internal validation metrics: Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. These metrics collectively offer insight into the cohesion and separation of the resulting clusters. The comparison results for K-Means, DBSCAN, and Agglomerative Clustering are presented below.

Table 14: Evaluation results of each technique's best models.

	K-Means	DBSCAN	Agglomerative
Parameters	k=4	Eps=0.4, min_samples=5	k=4
Clusters	4	3	4
Silhouette Score	0.601	0.665	0.604
Calinski-Harabasz Score	4794.394	2366.689	4346.696
Davies-Bouldin Score	0.867	1.182	0.811

K-Means clustering produced 4 clusters with a Silhouette Score of 0.601, which is considerably lower than that of the other models. The Calinski-Harabasz Index of 4794.394, which achieves the highest score, reflects excellent cluster compactness and clear boundaries. The Davies-Bouldin Score of 0.867 is relatively low, reflecting well-separated and dense clusters. Overall, K-Means demonstrates strong clustering performance, especially in terms of intra-cluster similarity and inter-cluster separability.

DBSCAN yielded only 3 clusters, where one cluster is an outlier, with the highest Silhouette Score of 0.665, suggesting firm cluster cohesion and separation. Its Calinski-Harabasz Score of 2366.689 is also significantly lower, indicating less distinct clustering. Furthermore, the Davies-Bouldin Score of 1.182 is the highest among the three methods, pointing to less compact and poorly separated clusters. This result suggests that DBSCAN may not be well-suited to this dataset in its current configuration, possibly because the dataset lacks a clear density-based structure.

Agglomerative Clustering produced 4 clusters, similar to K-Means. It achieved a Silhouette Score of 0.604, which is nearly the same as that of K-Means. The Calinski-Harabasz Score of 4346.696 indicates that the clusters are well-defined and compact. Additionally, the Davies-Bouldin Score of 0.811 is the lowest among all models, further confirming the superior cluster quality.

Among the three models, the results indicate that K-Means and Agglomerative Clustering are the best-performing models for this dataset. However, the best model isn't selected based on the best evaluation performance, but rather on how well it clusters the Crime-House trend in Adelaide suburbs and the fascinating insights and meaningful patterns we get from the cluster profile.

5. Discussion

This section describes the Data Profiling phase, in which we found that K-Means and Agglomerative Clustering are the best-performing models for this dataset, as determined by the Model Evaluation stage.

5.1 K-means Clustering Profile

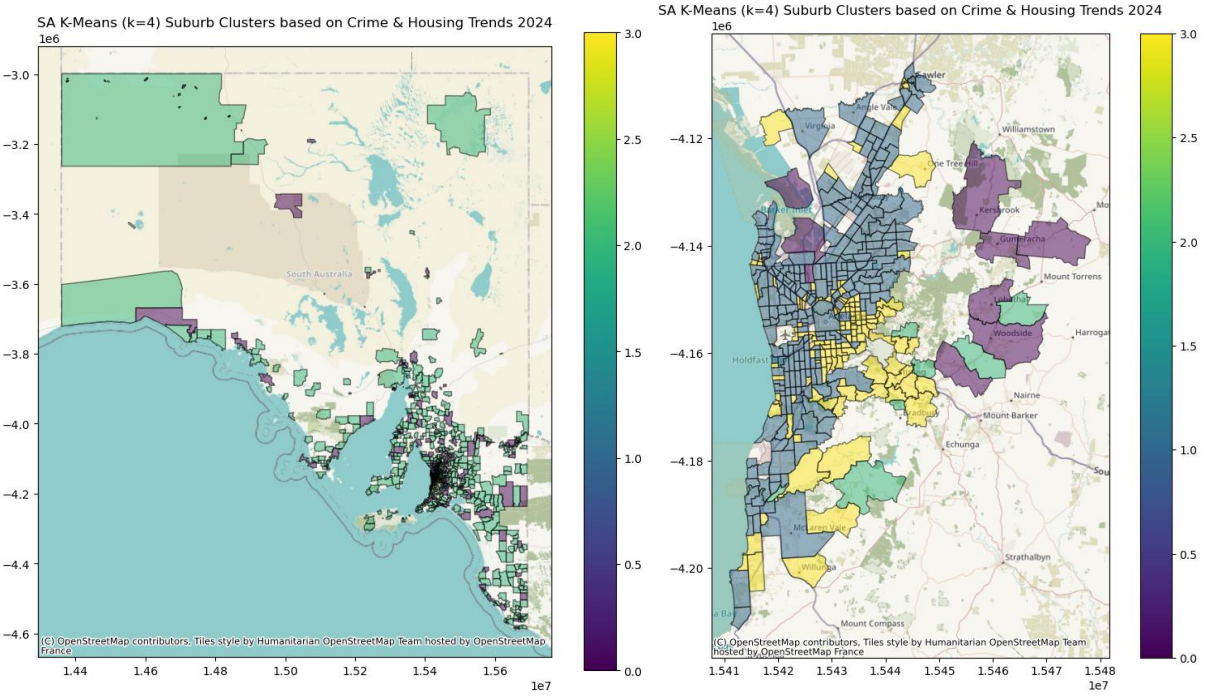
According to the profile result from recent 2024 data in Table 15, Cluster 0 (9% of suburbs) represents moderate crime, low market areas and low house sales, indicating areas with limited market activity and low property values. Cluster 1 (13% of suburbs) exhibits the highest crime levels and high housing market activity, which may represent densely populated areas with high market turnover, despite the presence of crime. Cluster 2 (28% of suburbs), the largest group, has very low crime rates and almost no house sales, with very low house prices. It may be low-activity suburbs. Cluster 3 (7% of suburbs) exhibits low crime and high-value suburbs, where the average house price is the highest at $1.35 million, indicating affluent neighbourhoods. Further visualisation of K-Means clustering distribution and trends over time can be found in [Appendix 10.6].

Table 15: Summary of K-Means (k=4) Clusters Profile

Cluster	Description
0	Moderate crime, Low house prices, Low house sales
1	High crime, High house prices, High house sales
2	Low crime, Low house prices, Extremely Low house sales
3	Low crime, High house prices, Moderate house sales

Figure 18 below presents a GIS-based visualisation of K-Means (k=4) clustering for suburbs in South Australia, including a zoomed-in view of the Adelaide metropolitan area. The map reveals that Clusters 0 (purple regions) and 2 (green regions), characterised by low housing market activity, are predominantly located in outer suburban areas. Cluster 1 (dark blue regions), which exhibits high crime rates alongside active housing market activity, corresponds to inner-city suburbs and the western coastal areas. In contrast, the affluent neighbourhoods of Cluster 3 (yellow regions) are concentrated in the eastern and southeastern suburbs near the inner city.

Figure 18: K-Means (k=4) Suburbs Clustering 2024



5.2 Agglomerative Clustering Profile

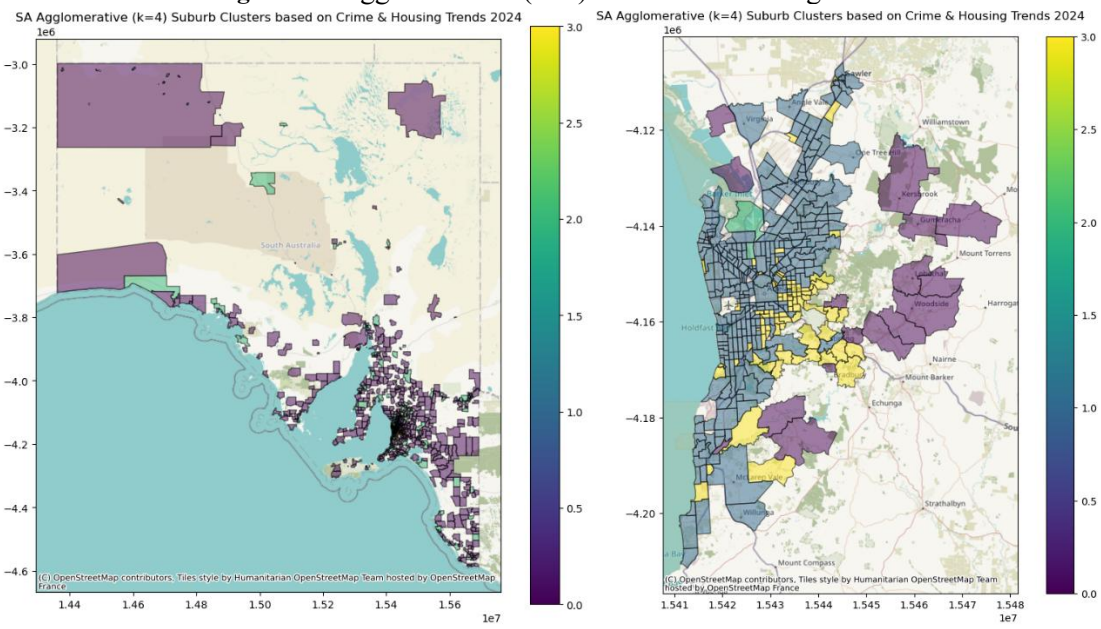
According to the profile results from the recent 2024 data in Table 16, Cluster 1 (31% of suburbs) is the largest and safest group, characterised by low crime rates, low sales, and a low average house price. It is likely to include outer suburbs with minimal market activity. Cluster 2 (15% of suburbs) has the highest property crime rate and the strongest sales, with a high average house price. These could be inner-city, high-demand areas. Cluster 3 (6% of suburbs) displays the highest person crime rate and high property crime, resulting in extremely low sales and very low average house prices. These may be distressed areas where crime affects investment. Cluster 4 (5% of suburbs) reveals low crime and high value suburbs, characterised by moderate property crime and low person crime. Moreover, it has the highest average house price, with relatively moderate sales, suggesting it is an elite residential zone. Further visualisation of Agglomerative clustering distribution and trends over time can be found in [Appendix 10.7].

Table 16: Summary of Agglomerative (k=4) Clusters Profile

Cluster	Description
0	Low crime, Low house prices, Low house sales
1	High crime, High house prices, High house sales
2	High crime, Low house prices, Extremely Low house sales
3	Low crime, High house prices, Moderate house sales

Figure 19 illustrates the spatial distribution of suburbs in South Australia using Agglomerative clustering (k=4), with a focused view of the Adelaide metropolitan area. The visualisation shows that Cluster 0 (purple areas), defined by low housing market activity, is primarily situated in the outer suburban zones. Cluster 2 appears less prominently and is scattered in limited pockets near Adelaide's urban core. Cluster 1 (dark blue areas), marked by elevated crime rates and an active housing market, is concentrated within inner-city districts and extends toward the western coastal suburbs. In contrast, Cluster 3 (yellow areas) represents more affluent suburbs and is mainly located in the eastern and southeastern parts close to the city center.

Figure 19: Agglomerative (k=4) Suburbs Clustering 2024



5.3 Data Profiling Discussion

The clustering analysis reveals distinct spatial patterns across Adelaide's suburbs, addressing the research question: **Are there spatial clusters where high crime rates and low housing affordability coincide, indicating zones of urban inequality?** The best performance models from K-Means and Agglomerative methods have similar performance, profiles and trends. Although they have a slightly different profile in some clusters, both clustering methods can identify clusters that reflect socio-spatial inequality, especially in Clusters 1 and 3. Cluster 1 indicates high crime areas, high housing market activity, and prices, whereas Cluster 3 exhibits elite residential zones with a low crime rate and the highest average housing prices, as well as moderate sales volumes. High-value suburbs tend to have lower crime and moderate sales, while high-crime areas can still attract real estate activity, particularly in inner-city zones. However, Cluster 2 of Agglomerative Clustering stands out. It more clearly separates distressed, high-crime, low-value zones. At the same time, K-Means provides a more precise market-based segmentation, especially in differentiating high-demand areas (Cluster 1) from those that are safe but stationary (Cluster 3). For this research, Agglomerative Clustering better highlights urban inequality, while K-Means excels at identifying market segmentation.

The findings carry important implications for both researchers and practitioners. For researchers, this study illustrates the importance of utilising various clustering methods when analysing urban socio-economic data, demonstrating how different algorithms can uncover complementary aspects of inequality. Agglomerative Clustering reveals distressed and at-risk suburbs, whereas K-Means offers more precise insights into market-driven trends. This suggests that subsequent studies on urban inequality should avoid relying on a single method and instead adopt a comparative, multi-method approach to accurately capture the complexities of socio-spatial patterns.

For practitioners, including urban planners, policymakers, and real estate professionals, the findings highlight that urban inequality in Adelaide is not a uniform issue, but rather a segmented one. Certain suburbs face both high crime and low affordability, while others feature exclusive housing markets with low crime rates. The ability to differentiate between struggling zones and stable, high-value neighborhoods provides a practical evidence base for targeted policy measures, resource allocation, and urban development strategies. Agglomerative Clustering provides policymakers with clearer signals for identifying and addressing inequality hotspots, whereas K-Means offers investors and developers with insights into market segmentation. These approaches can support more equitable and effective decision-making in housing and urban planning.

6. Limitations

While clustering analysis provides valuable insights into urban socio-economic issues, several limitations should be noted. The first limitation is that clustering techniques are highly sensitive to input features and their scaling. The effectiveness of clustering is significantly influenced by feature selection and data preprocessing. Variables such as crime rates, housing prices, and sales volumes exist on different scales, and even after normalisation, the choice of features can introduce biases into the clustering results.

The second limitation is that some clustering methods require pre-specification of the number of clusters, such as the K-Means algorithm. This necessity can lead to oversimplification of complex socio-spatial patterns. Although validation metrics like the Silhouette Score, CHI, and DBI can help assess cluster performance, they may not fully capture the complicated boundaries of real-world urban inequality, which is a complex phenomenon.

The last limitation is that it is essential to consider the limitations associated with spatial context. While suburb-level data provides a geographical basis, clustering approaches treat suburbs as independent data points. This treatment overlooks the implications of spatial autocorrelation and neighbourhood effects, which are crucial in urban studies, where adjacent suburbs often share similar socio-economic conditions and crime dynamics.

7. Conclusion

The clustering analysis of Adelaide's suburbs investigates spatial patterns related to crime rates and housing affordability, aiming to identify zones of urban inequality. Both K-Means and Agglomerative Clustering effectively highlight socio-spatial inequality; they can clearly identify high-crime, low-value zones. While the K-Means cluster profile provides better market segmentation insights, the Agglomerative cluster profile offers more effective highlights of urban inequality.

The findings highlight that researchers should use multiple clustering methods to understand urban socio-economic data, avoiding dependence on a single technique to capture complexities. For practitioners, identifying the segmented nature of urban inequality enables the development of targeted policy measures, effective resource allocation, and informed urban planning. Agglomerative Clustering is particularly suited to identifying inequality hotspots, while K-Means informs market-driven analyses.

Nonetheless, the study is subject to several limitations, including sensitivity to feature selection and scaling, the need for predefined cluster numbers, and the neglect of spatial autocorrelation, all of which may deny the robustness and interpretability of the findings.

To enhance the robustness of this research, future work could benefit from integrating demographic and socio-economic indicators, such as population density, age groups, and income levels. These indicators are often correlated with crime patterns and the housing market. Moreover, integration of transportation access, education facilities, and employment hubs could uncover further dimensions of spatial inequality (Randolph, B & Tice, A 2014) [12]. Overall, this study lays a foundational framework for identifying spatial injustices and guiding equitable urban development strategies.

8. Replication Package

All ingestion data sets and codes can be found in the GitHub repository (Lalitphan S, 2025)[5].

9. References

[1] Margaretic, P & Sosa, JB 2025, 'How Local is the Crime Effect on House Prices?', The Journal of Real Estate Finance and Economics, vol. 70, no. 4, pp. 754–793.

[2] Motevali, S, Aljawawdeh, H, Abuezhayeh, S & Qaddoumi, E 2025, 'Explore the Relationship between House Prices and Crime Rate in the UK Using Machine Learning Techniques', International Arab Journal of Information Technology, vol. 22, no. 3, pp. 411–428.

[3] de La Paz, PT, Berry, J, McIlhatton, D, Chapman, D & Bergonzoli, K 2022, 'The impact of crimes on house prices in LA County', Journal of European Real Estate Research, vol. 15, no. 1, pp. 88–111.

[4] South Australia Government Data Dictionary, Data.SA – Government of South Australia, viewed 15 June 2025, https://data.sa.gov.au/data/dataset/.

[5] Lalitphan Sae-teoh Github 2025, online repository, Github, https://github.com/lalitphan-rainie/COMPSCI_7209_Big_Data_Project_Analytics .

[6] MacQueen, J 1967, 'Some Methods for Classification and Analysis of Multivariate Observations', Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 5, no. 1, pp. 281–297.

[7] Ester M, Kriegel H-P, Sander J, and Xu X 1996, 'A density-based algorithm for discovering clusters in large spatial databases with noise', In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), AAAI Press, pp. 226–231.

[8] Müllner, D 2011, 'Modern hierarchical, agglomerative clustering algorithms'.

[9] Rousseeuw, PJ 1987, 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65.

[10] Kozak, M 2012, 'A Dendrite Method for Cluster Analysis' by Caliński and Harabasz: A Classical Work that is Far Too Often Incorrectly Cited', Communications in Statistics. Theory and Methods, vol. 41, no. 12, pp. 2279–2280.

[11] Davies, DL & Bouldin, DW 1979, 'A Cluster Separation Measure', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227.

[12] Randolph, B & Tice, A 2014, 'Suburbanizing Disadvantage in Australian Cities: Sociospatial Change in an Era of Neoliberalism', Journal of Urban Affairs, vol. 36, no. s1, pp. 384–399.

[13] Buonanno, P, Montolio, D & Raya-Vílchez, JM 2013, 'Housing prices and crime perception', Empirical Economics, vol. 45, no. 1, pp. 305–321.

[14] Braakmann, N 2017, 'The link between crime risk and property prices in England and Wales: Evidence from street-level data', Urban Studies (Edinburgh, Scotland), vol. 54, no. 8, pp. 1990–2007.

[15] Baird, MD, Schwartz, H, Hunter, GP, Gary-Webb, TL, Ghosh-Dastidar, B, Dubowitz, T & Troxel, WM 2020, 'Does Large-Scale Neighborhood Reinvestment Work? Effects of Public-Private Real Estate Investment on Local Sales Prices, Rental Prices, and Crime Rates', Housing Policy Debate, vol. 30, no. 2, pp. 164–190.

[16] Cueva, D & Cabrera-Barona, P 2024, 'Spatial, Temporal, and Explanatory Analyses of Urban Crime', Social Indicators Research, vol. 174, no. 2, pp. 611–629.

10. Appendix

10.1 SA Crime Level Type

Table 17: Offence Type Description

Offence Level 1 Description	Offence Level 2 Description	Offence Level 3 Description
OFFENCES AGAINST PROPERTY	FRAUD DECEPTION AND RELATED OFFENCES	Obtain benefit by deception
OFFENCES AGAINST PROPERTY	FRAUD DECEPTION AND RELATED OFFENCES	Other fraud, deception and related offences
OFFENCES AGAINST PROPERTY	PROPERTY DAMAGE AND ENVIRONMENTAL	Graffiti
OFFENCES AGAINST PROPERTY	PROPERTY DAMAGE AND ENVIRONMENTAL	Other property damage and environmental
OFFENCES AGAINST PROPERTY	PROPERTY DAMAGE AND ENVIRONMENTAL	Property damage by fire or explosion
OFFENCES AGAINST PROPERTY	SERIOUS CRIMINAL TRESPASS	Other unlawful entry with intent
OFFENCES AGAINST PROPERTY	SERIOUS CRIMINAL TRESPASS	SCT - Non Residence
OFFENCES AGAINST PROPERTY	SERIOUS CRIMINAL TRESPASS	SCT - Residence
OFFENCES AGAINST PROPERTY	THEFT AND RELATED OFFENCES	Other theft
OFFENCES AGAINST PROPERTY	THEFT AND RELATED OFFENCES	Receive or handle proceeds of crime
OFFENCES AGAINST PROPERTY	THEFT AND RELATED OFFENCES	Theft from motor vehicle
OFFENCES AGAINST PROPERTY	THEFT AND RELATED OFFENCES	Theft from shop
OFFENCES AGAINST PROPERTY	THEFT AND RELATED OFFENCES	Theft/Illegal Use of MV
OFFENCES AGAINST THE PERSON	ACTS INTENDED TO CAUSE INJURY	Assault police
OFFENCES AGAINST THE PERSON	ACTS INTENDED TO CAUSE INJURY	Common Assault
OFFENCES AGAINST THE PERSON	ACTS INTENDED TO CAUSE INJURY	Other acts intended to cause injury
OFFENCES AGAINST THE PERSON	ACTS INTENDED TO CAUSE INJURY	Serious Assault not resulting in injury
OFFENCES AGAINST THE PERSON	ACTS INTENDED TO CAUSE INJURY	Serious Assault resulting in injury
OFFENCES AGAINST THE PERSON	HOMICIDE AND RELATED OFFENCES	Murder
OFFENCES AGAINST THE PERSON	HOMICIDE AND RELATED OFFENCES	Other homicide and related offences
OFFENCES AGAINST THE PERSON	OTHER OFFENCES AGAINST THE PERSON	Abduction, harassment and other offences
OFFENCES AGAINST THE PERSON	OTHER OFFENCES AGAINST THE PERSON	Dangerous or negligent acts
OFFENCES AGAINST THE PERSON	OTHER OFFENCES AGAINST THE PERSON	Threatening behaviour
OFFENCES AGAINST THE PERSON	ROBBERY AND RELATED OFFENCES	Aggravated robbery
OFFENCES AGAINST THE PERSON	ROBBERY AND RELATED OFFENCES	Blackmail and extortion
OFFENCES AGAINST THE PERSON	ROBBERY AND RELATED OFFENCES	Non-aggravated robbery
OFFENCES AGAINST THE PERSON	SEXUAL ASSAULT AND RELATED OFFENCES	Aggravated sexual assault
OFFENCES AGAINST THE PERSON	SEXUAL ASSAULT AND RELATED OFFENCES	Non-aggravated sexual assault
OFFENCES AGAINST THE PERSON	SEXUAL ASSAULT AND RELATED OFFENCES	Non-assaultive sexual offences

10.2 Univariate Analysis - Crime Statistics SA Year Trends

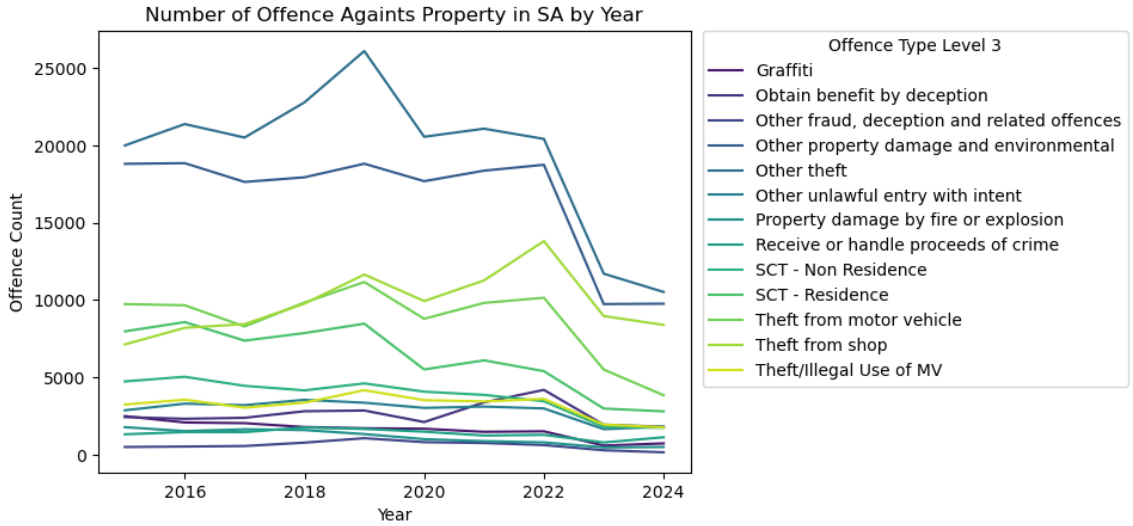


Figure 20: Number of Offences against property in SA by offence type level 3 by year

When exploring year trends in more detail, we found that over 25,000 theft cases occurred in 2019, but this number has since significantly dropped to around 12,000 cases in 2013. Theft/illegal Use of MV has increased in 2022. However, most of the offences against property decreased in 2024.

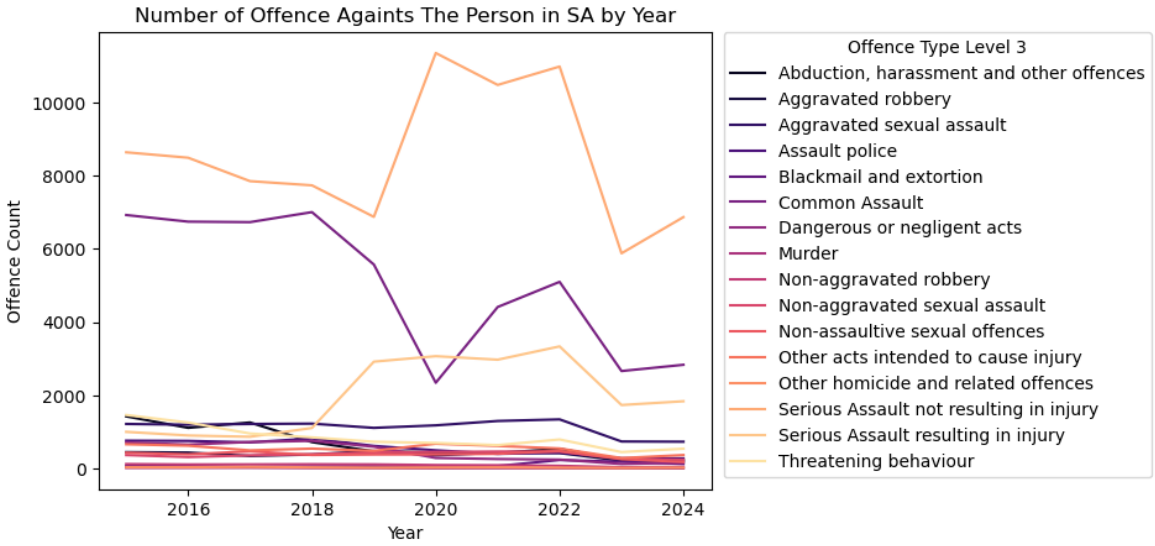


Figure 21: Number of Offences against the person in SA by offence type level 3 by year

For offences against the person by year trends, the number rose by more than 10,000 cases of serious assault without injury in 2020. In contrast, the common assault has remarkably dropped from around 7,000 cases to approximately 2,000 cases. However, the number of severe assault cases without injury declined significantly in 2023 and gradually increased in 2024. The serious assault resulting in injury increased from 1,000 to 3,000 in 2019 and continued until slowly dipping in 2023.

10.3 Bivariate Analysis - Crime Statistics SA & Metro Median House Sales

To explore the relationship between crime statistics and median house sales in SA, the bivariate analysis is performed as follows:

Figure 22: Correlation matrix between Crime Statistics and House Price and Sales in SA

Firstly, the correlation matrix between crime statistics and median house price and sales has low correlations. The median house price and sales show a negative correlation, which makes sense because low house prices should be easier to sell, while high house prices are harder to sell. However, the number of offences has a very low correlation to house prices and sales.

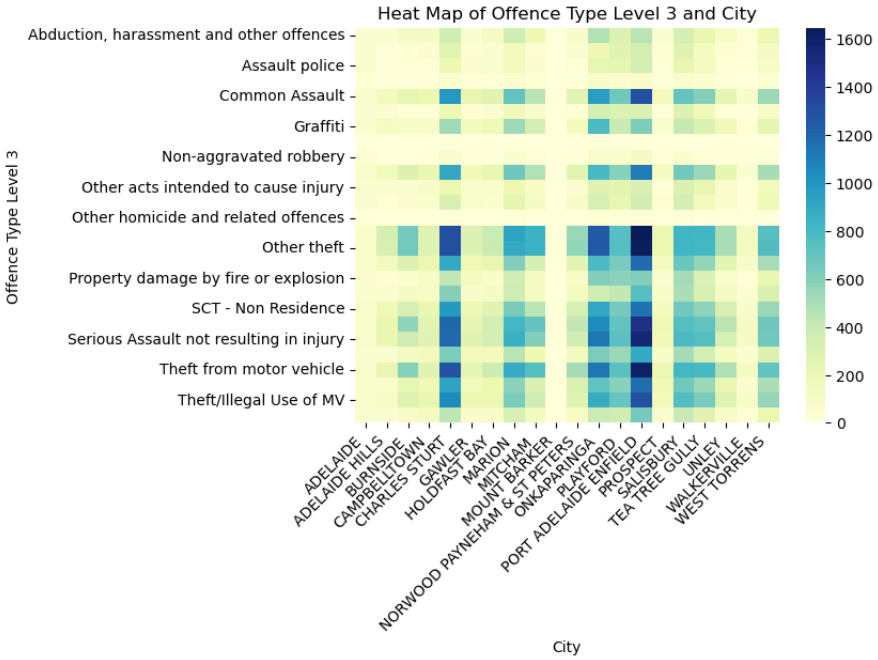


Figure 23: Heat map of the Offence Type Level 3 in SA and City

Next, the heatmap of the offence type level 3 shows the frequency of crimes that occurred in the city. Port Adelaide Enfield, Charles Sturt and Onkaparinga show some interesting information. They have high crime rates in offences against property, for instance theft and property damage. According to the boxplot in Figure 12, it exhibits slight variations in prices for areas except Charles Sturt, which might indicate that high crime rates affect the variance of house prices over time.

10.4 Spatial Heatmaps

The following visualisations are examples of heatmaps on the SA suburbs map over time.

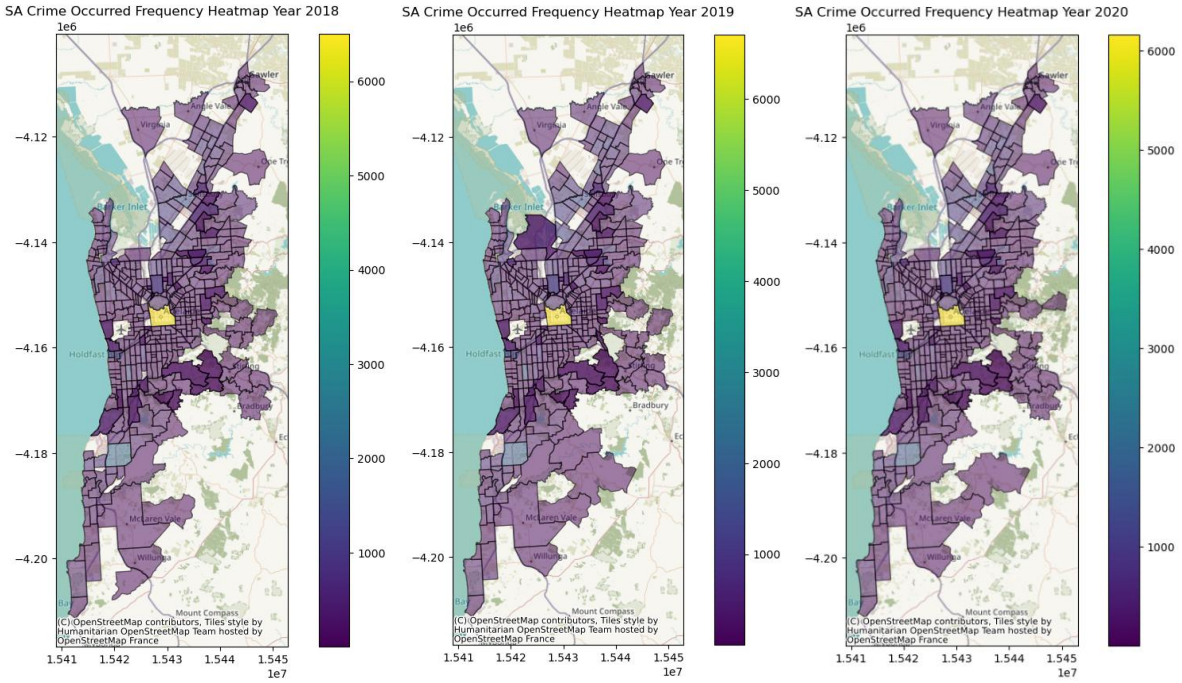


Figure 24: SA Crime Statistics Heatmap over time between 2018 – 2020

Bright yellow cluster in central Adelaid is shows the CBD and surrounding suburbs, with crime counts exceeding 6000 incidents. While dark purple outer suburbs, for instance, southern and eastern suburbs, regional fringe areas like McLaren Vale or Virginia, shows very low crime (<1000). High crime zones appear to concentrate in the north-west belt and some parts of the inner west (e.g., Port Adelaide Enfield, Playford, parts of Charles Sturt).

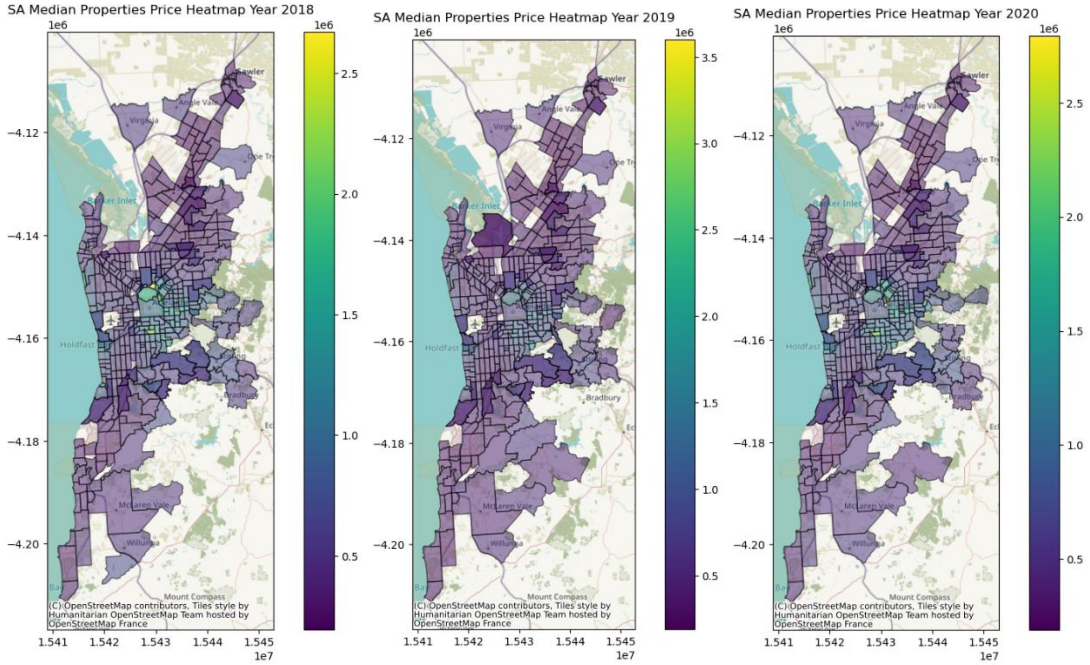


Figure 25: SA Median House Price Heatmap over time between 2018 - 2020

Brightest areas, which indicated the high house price, are not the same as high-crime zones, including eastern suburbs, such as Burnside, Unley, Walkerville and Mitcham. Central Adelaide has lower prices compared to the eastern suburbs, despite high population density. Northern suburbs like Elizabeth and Playford exhibit low housing prices.

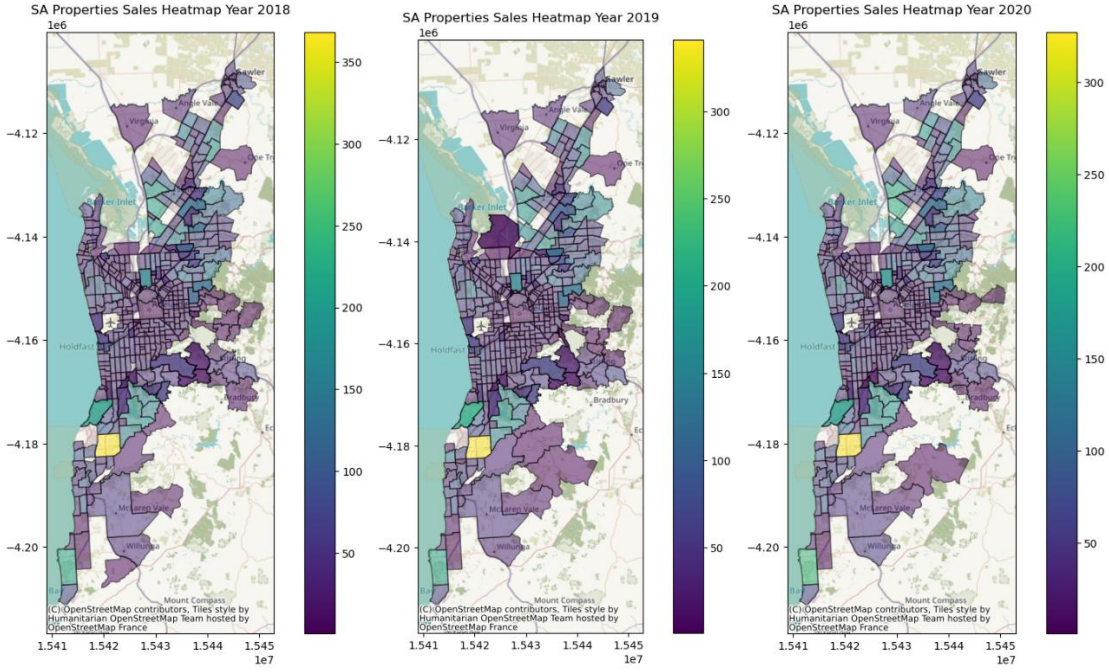


Figure 26: SA House Sales Heatmap over time between 2018 – 2020

The heatmaps from 2018 to 2020 reveal consistent patterns in property sales, with central urban areas showing stable and relatively high sales volumes throughout the period. Notably, western and southwestern suburbs experienced a gradual increase in sales, suggesting growing demand or ongoing development in those regions. In contrast, some outer southern areas saw reduced activity in 2020, possibly reflecting the economic impact of COVID-19. Overall, the data indicates a strong preference for properties in inner and middle-ring suburbs, while outer areas show more variability in sales trends over time.

10.5 Evaluation Metrics

10.5.1 Silhouette Score

The Silhouette Score measures how similar each point is to its cluster, referred to as cohesion, compared to other clusters, which is known as separation (Rousseeuw, PJ 1987) [9]. The score ranges from -1 to +1, where +1 indicates that the sample is far from neighboring clusters and well-matched to its cluster. While -1 suggests that the sample may have been assigned to the wrong cluster, and 0 indicates that clusters overlap. The average silhouette score across all data points provides an overall assessment of clustering quality. A higher silhouette score suggests better-defined and more distinct clusters.

10.5.2 Calinski-Harabasz Index

The Calinski-Harabasz Index (CHI) evaluates cluster validity based on the ratio of between-cluster dispersion to within-cluster dispersion (Kozak, M 2012) [10]. It is defined as follows:

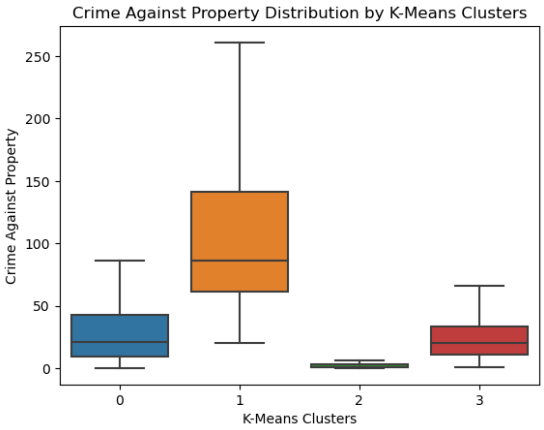
$$CHI = \frac{Between\ Cluster\ Variance}{Within\ Cluster\ Variance} \times \frac{(N-k)}{(k-1)}$$

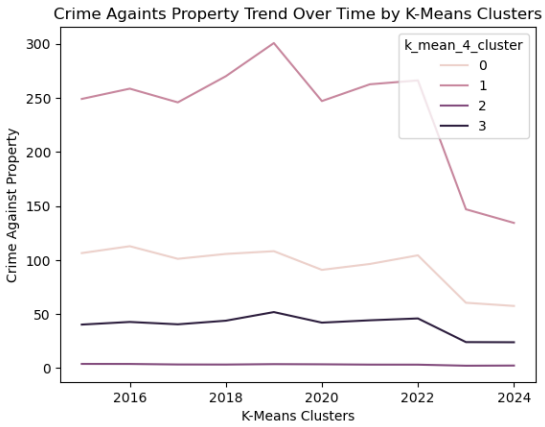
Where, N is the number of data points and k is the number of clusters. A higher Calinski-Harabasz score indicates better-defined clusters, as it reflects compact clusters that are well separated from one another.

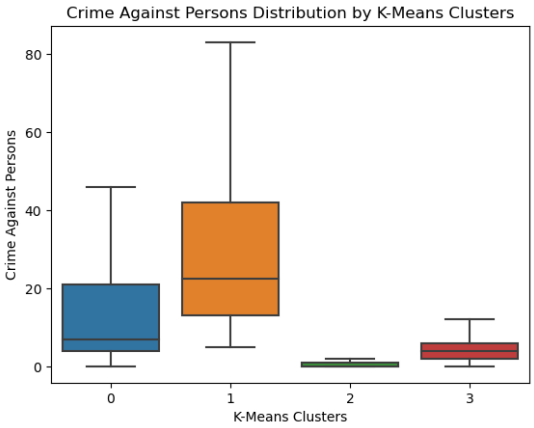
10.5.3 Davies-Bouldin Index

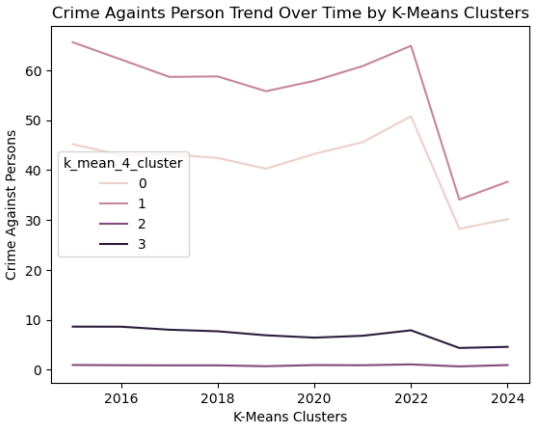
The Davies-Bouldin Index (DBI) assesses intra-cluster similarity and inter-cluster differences. It is defined as the average ratio of the sum of within-cluster scatter to between-cluster separation for each cluster (Davies, DL & Bouldin, DW 1979) [11]. Lower DBI indicates that clusters are far from each other, while higher DBI suggests overlapping or less distinct clusters. Therefore, lower DBI values are better. This index is particularly useful when comparing clustering models with different numbers of clusters.

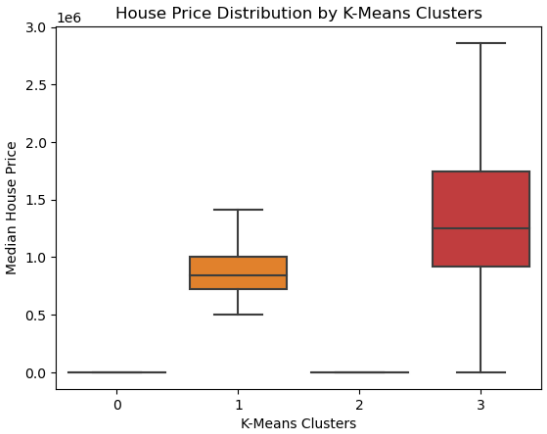
10.6 K-Means Clusters Distribution & Trend Over Time

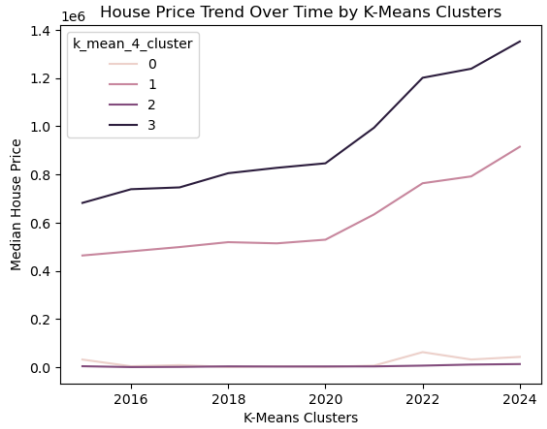


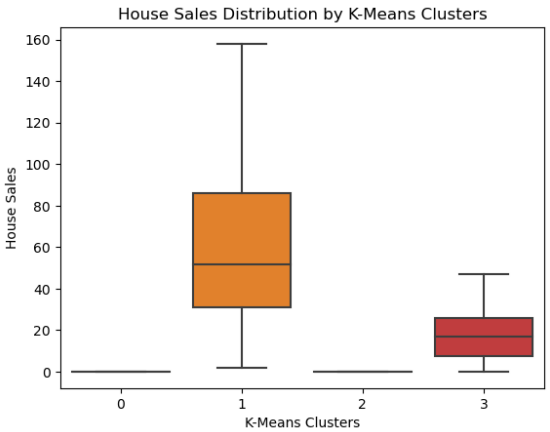


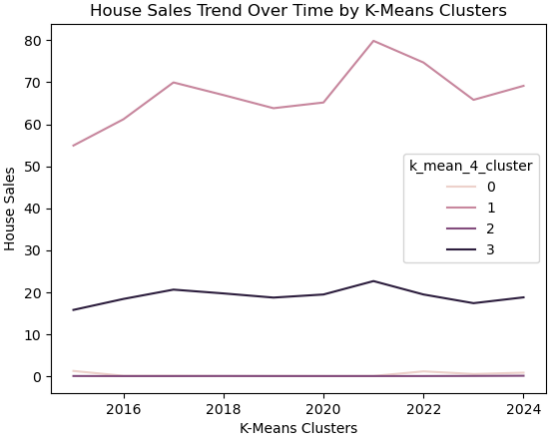












10.7 Agglomerative Clusters Distribution & Trend Over Time