# Crime Rates Analysis in Malaysia – A Visual Analytics Approach

Ashraf Rauf

**Abstract**— This paper explores the temporal and spatial patterns of crime rates in Malaysia using district-level data. Crime rates and neighourhood characteristics data are obtained via the OpenDOSM website. The paper first analyses the temporal pattern of crime across districts. Subsequently, it attempts to use the k-means clustering method to group similar cistricts together based on their crime and neighbourhood characteristics profile. The results show that crime has generally been on a declining trend across district, with some spatial nuances. There is also limited similarity in terms of crime profile across district, even after accounting for socio-economic variables.

## 1 Problem Statement

In the World Happiness Report 2024, safety is highlighted as one of the influential factors that affects an individual's subjective well-being [1]. Recognising its importance, public security has also been incorporated as an explicit goal within the Sustainable Development Goal (SDG) framework [2]. One of the key indicators in measuring public security is the crime rate. A better understanding of how crime rates evolve through time and its spatial distribution can help inform local planning and policing strategies.

As a country that experienced rapid urbanisation but with regional disparities, Malaysia provides a good contextual basis to study the public security phenomenon. This paper thus aims to study crime rates in Malaysia and seeks to answer the following questions:

1. **RQ1**: What is the trend of crime rates in Malaysia in recent years?
2. **RQ2**: Are there similarities in crime type breakdown across districts?
3. **RQ3**: What are the district characteristics associated with variations in crime?

This analysis utilizes crime rate data sourced from OpenDOSM, a governmental platform designed to disseminate national data efficiently [3]. The dataset contains crime rates from 2016-2023 at the police district level and includes the breakdown of different criminal types. This would enable a comprehensive temporal investigation of crime rates. Additionally, district-level neighbourhood characteristics obtained from the same platform would allow the investigation of the relationship between crime rates and neighbourhood characteristics. To facilitate spatial analysis, district shapefiles are sourced from geoBoundaries, which allow for the visualization and exploration of spatial patterns in the data [4].

## 2 State of the Art

Visual analytic approaches in the crime analysis domain are mainly concerned with using visualisations to address questions like the ones posed in this paper, which are variations of:

- How are crimes distributed spatially? Are there any areas with higher concentration of crimes?
- What is the evolution of crime through time?
- Is there a relationship between crimes and neighbourhood characteristics?

Maciejewski et al. (2010) [5] utilises a dataset which contains granular information of each crime incidence such as the location and time. The authors approached spatial analysis of crime by applying density-based partitioning under the assumption that nearby crimes are related. The clusters are then visualised on a map as circles with the size proportional to the number of crimes. The authors also implemented kernel density estimation (KDE), assuming crime events are spatially correlated, and visualising the output as a heatmap. Temporal patterns are analysed using simple line charts, while spatio-temporal analysis is addressed via contour line ghosting combined with heatmap visualisation. This method is also extended to facilitate neighbourhood characteristic analysis.

While the KDE heatmap approach provides granular spatial details, its implementation is sensitive to its parameter settings [6], thus may be prone to misuse or misinterpretation. Additionally, administrative considerations may warrant a different approach. Guided by the requirements of domain experts, Garcia et al. (2021) [7] aggregated crime events at the census unit level and used a choropleth map to identify sites with higher crime rates. This approach indirectly assumes crime rates are uniformly distributed throughout the area. Temporal investigation is facilitated by various bar charts showing different cumulative periods. They also utilised a line chart to visualise the evolution of crime rates across time and the ranking by occurrences for each period. The authors also attempted to implement a spatio-temporal clustering of the census units but failed to obtain any meaningful inference. The paper also did not attempt to link crime rates and neighbourhood characteristics.

In the absence of granular datasets, Silva et al. (2017) [8] developed an approach that utilised publicly available dataset aggregated at the police district level. Like [7], the authors utilised a choropleth map to analyse spatial distribution of crime and identify crime hotspots. While the authors mentioned a clustering implementation, there is limited discussion on the justification or inference of the clusters. The authors also used parallel coordinate plots to visualise the

breakdown of crime types by districts. Temporal analysis is mainly implemented by user interactions such as filtering and brushing. A time-lapse animation of the choropleth map is utilised to help with spatio-temporal investigation, although the effectiveness is unclear. Parallel coordinate plots are also used to visualise neighbourhood characteristics by districts to facilitate multivariate analysis.

Like [8], this paper would use a choropleth map to assess spatial variation given data limitations, while cognisant of its limitations. Temporal trends can be explored using simple 2D charts. Clustering can also be implemented to identify spatially similar districts. Parallel coordinate plots are also useful to visualise the multivariate data, especially when combined with clustering methods.

## 3 Properties of the Data

The crime dataset is sourced from the Royal Malaysian Police and is made available via the OpenDOSM platform [3]. The crime data is presented as the number of crimes for each police district. The data is based on convicted cases, thus omits any false reports. The crimes are divided into two broad categories: assault and property crimes. Each of the main categories has seven and five sub-categories, respectively. However, detailed descriptions of the sub-categories are not provided. The data is reported on an annual basis for the period 2016 to 2023, resulting in a total of 1,072 district-year observations.

To ensure comparability across districts, the crime volume is divided by the population for each district. Similar to [7] and [8], this approach indirectly assumes all inhabitants are equally at-risk for crime events. The population data, supplied by the Department of Statistics Malaysia, is also available on the same platform. Given that the police and administrative districts differ slightly, adjustments were made to align the police districts with administrative districts before merging them based on district identifiers. The administrative districts are taken to be the ground truth, in line with most other statistical datasets. As the population data is only available 2020 onwards, crime data prior to 2020 is normalised using the 2020 population values.

District-level neighbourhood characteristics were similarly obtained from the same platform, available in different datasets. As some datasets had lower temporal resolutions, complete data was only available for two time periods. The most recent period is for the year 2022, while the other is a mix between 2019 and 2020 data, depending on availability. Taken together, these datasets can be seen as a comparison between pre- and post-COVID19 pandemic period. The characteristics can be categorised into four broad themes – age composition, ethnic composition, income and employment, as well as education. The datasets were merged at the administrative district level using district identifiers[1] and all variables are normalised to the district-level population.

For spatial visualisation of the data, shapefiles from the geoBoundaries website are used [4]. The shapefile contained administrative district-level polygons hence were easily matched with the other datasets.

Missing values are examined using a nullity matrix heatmap and spatial visualisation, as shown in Fig. 1. While the crime dataset by itself has no missing values, Fig. 1a shows that certain districts do not have crime data reported, which are mainly within East Malaysia. These districts may not have a dedicated police department; thus, the data may be included in nearby districts. The omission may weaken the relationship between crime rates and neighbourhood characteristics. Fig. 1b shows that some neighbourhood characteristics are absent for certain districts. As only four districts are affected, it is unlikely it will materially affect the subsequent analysis.

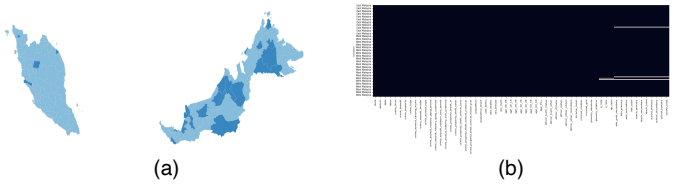


*Fig. 1: Understanding missing data. (a) Darker regions identify the administrative districts that do not have crime data reported, likely due to absence of police precinct. (b) The nullity matrix, where each observation is encoded as a line, and gaps imply missing values.*

## 4 Analysis

### 4.1 Approach

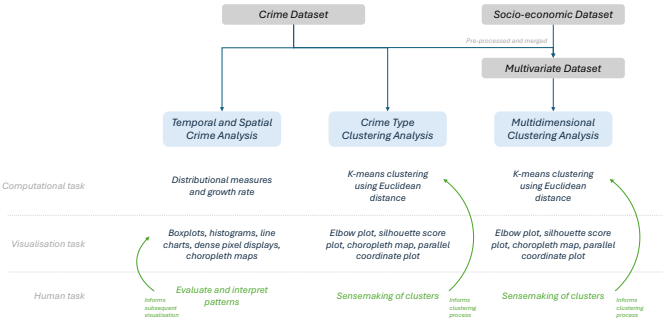


*Fig. 2: Analysis workflow overview. For each computational and visualization task, human-reasoning (highlighted in green) is required to interpret the results and refine the analysis.*

Similar to the papers mentioned earlier, this paper uses a visual analytics approach for temporal and spatial analysis of crime. The analysis workflow involves a series of

[1] Some additional data cleansing was still required due to slight discrepancies in naming convention (e.g., "Manjung (Dinding)" vs "Manjung").

computational and visualisation steps, after which human reasoning is required to interpret the results and refine the analysis (Fig. 2).

*RQ1: Temporal and Spatial Crime Analysis*

RQ1 is approached by utilising distributional plots such as boxplots and histograms across time, thus providing an overview of the temporal pattern of crime rates. Small multiples of similar plots can also be used across different crime types or regions to identify similarities or locate outliers within specific sub-groups. Spatio-temporal analysis is done by visualising district-level observations using the dense pixel display method, which is more commonly used in the financial domain. This method encodes each pixel as an observation, and hue is used to encode an appropriate value determined by the analyst, such as crime rate, crime rate growth, or district ranking of crime rate. This method is chosen instead of choropleth maps as the varying geometric sizes of the districts may affect the analyst's perception and thus influence interpretation. For all visualisations, the analyst is required to evaluate patterns and trends as well as use the insights to guide more focused exploration within the data using the relevant visual output.

*RQ2: Crime Type Clustering Analysis*

For RQ2, the multidimensional crime type data for each district is projected to a two-dimensional plane using the multidimensional scaling (MDS) method based on the Euclidean distance function. This visual representation helps the analyst to identify potential similarities in the pattern of crime types between districts. The analyst uses this information to inform the implementation of k-means clustering in the attempt of grouping similar districts together. The "elbow" method and the silhouette score are also used to guide the analyst to identify the optimal cluster. However, the final cluster count is determined by the analyst by inspecting parallel coordinate plots of the crime types and choosing clusters that are coherent and able to provide meaningful insights. If required, the analyst uses the results to refine the clustering process.

*RQ3: Multidimensional Clustering Analysis*

Analysis for RQ3 starts with analysing the correlation matrix of the crime and neighbourhood characteristic variables to identify potential relationships between the variables. The insights from this analysis will then inform the analyst on the relevant variables to be included for a similar clustering procedure as before but using both crime and socio-economic data. The potential optimal number of clusters will again be guided by the "elbow" method and silhouette score. As the aim of this paper is to provide insights on crime rates, the performance of the clusters will similarly be analysed using the parallel coordinate plots, specifically on the coherence of the clustering of crime data. The results from this exercise would be compared to the previous clustering implementation to identify any additional or supporting insights.

### 4.2 Process

*RQ1: What is the trend of crime rates in Malaysia?*

RQ1 is first approached by visualising the distribution of crime per '000 population rates across the years (Fig. 3). The distribution of crime rates has narrowed since 2016 (Fig. 3a), suggesting that areas with higher crime rates in 2016 have seen a decrease in crime. Nonetheless, since 2021, there seems to be a few districts that saw higher crime rates, shown by the outlier point marks above the boxplots. To better understand the phenomenon, the distribution of crime across years is also visualised by the different types of crime (Fig. 3b). The chart shows that the declining crime rates were mainly driven by declining property theft. Specifically, vehicle-related thefts have shown the most significant decline across the years, shown the leftward shift of the empirical cumulative distribution function.

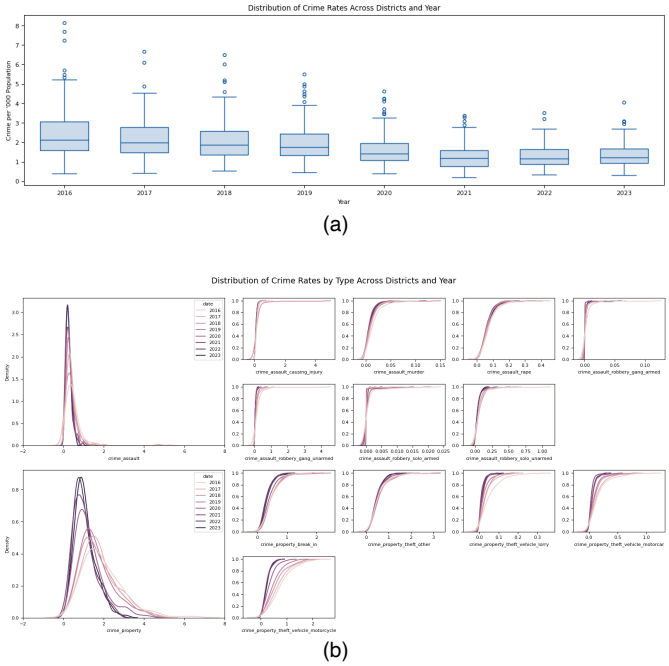


*Fig. 3: Temporal distribution of crime rates. (a) Each boxplot represents the distribution of district-level crime rate for each year. (b) Distribution of assault and property crime rates across time along with their sub-categories.*

While both charts paint a rosy picture at the aggregate level, it is also worth investigation district-level trends to assess any spatial differences in the trend of crime rates. In particular, as noted earlier, there seems to be some outlier districts that do not conform to the general trend. Fig. 4a suggests that prior to the COVID19 pandemic, districts in West Malaysia tend to experience a decline or negligible change in crime rates. However, districts in East Malaysia tend to have slightly positive crime rate growths. During the pandemic years of 2020 and 2021, all districts experienced a decline in crime rates. This could be due to the lockdown imposed as well as administrative backlog in processing reports. Post-pandemic, crime rates generally increase slightly as report backlogs are processed, with stronger growth seen

for districts within East Malaysia. Despite the differing trends in growth rates, the general ranking of districts based on crime rates remain the same. Fig. 4b shows that districts tend to remain within their crime rate quintiles throughout the years. Albeit at a different geographical scale, this observation seems to lend support to Weisburd's Law of Crime Concentration that postulates the temporal pattern of crime concentration tend to be stable [6]. Notwithstanding, there are some movements between adjacent quintiles in recent years. This could be due to the narrowing distribution shown earlier in Fig. 3a, resulting in the quintile rankings being more sensitive to smaller changes in crime rates, particularly for the observations near the quintile boundaries.

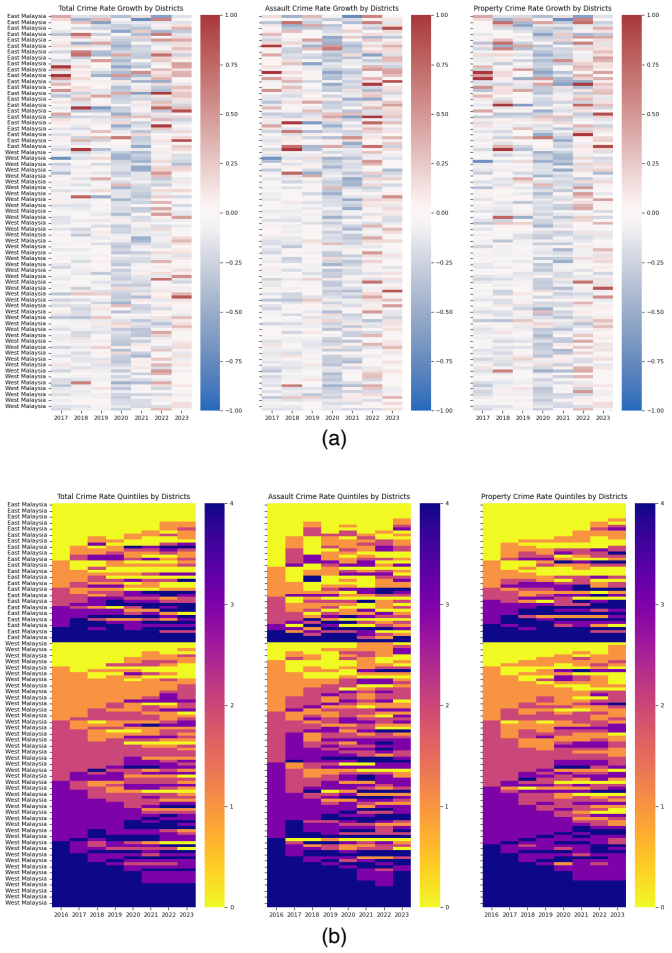


*Fig. 4: Spatio-temporal pattern of crime rates. The observations are sorts by East and West Malaysia Districts, with East Malaysia districts at the top. (a) Each cell represents a district-year observation, and the hue encodes the annual crime rate growth. Positive growth rates are shown in shades of red while negative growth rates are shown in shades of blue. (b) The hue in this visual encodes the quintile that the district belongs to for the particular year. Darker shades are equivalent to higher quintiles and thus higher crime rates.*

*RQ2: Are there similarities in crime types across districts?*

As the earlier section showed that the temporal pattern for crime concentration tend to be stable, only the most recent data is used for the MDS projection to reduce the computational complexity. The two-dimension projection results seem to imply the presence of several outliers which are mostly districts from East Malaysia (Fig. 5a). A visual spatial inspection found that these regions tend to border districts without a police precinct. One possible hypothesis is that these districts may include reports of crime events in the adjacent districts without a precinct, thus leading to a significant dissimilarity with the other districts. The projection also does not visualise any natural clusters, thus suggesting limited ability to cluster the districts into smaller coherent groups.

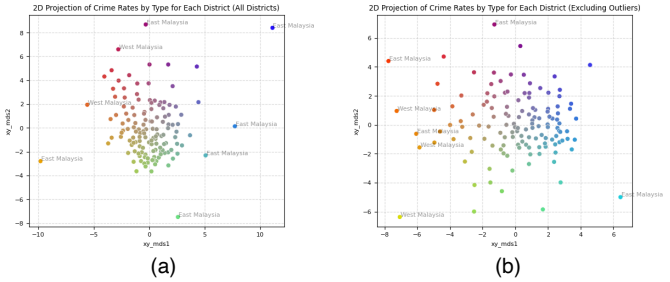


*Fig. 5: Two-dimensional MDS projection of crime rate types by district (a) Using all district data. (b) Excluding selected districts.*

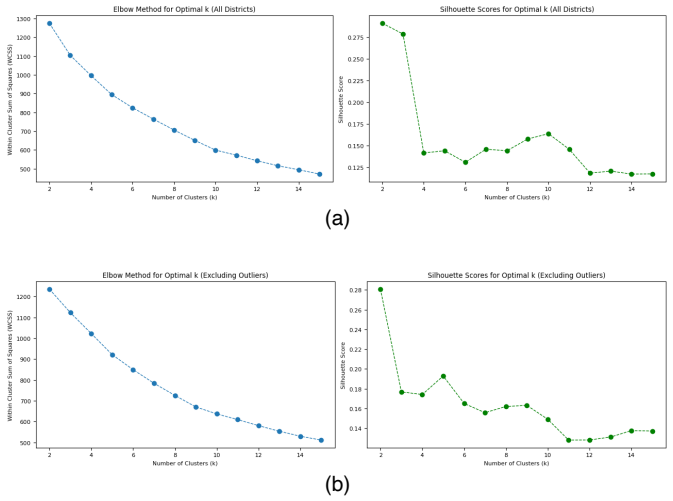


*Fig. 6: Identifying the optimal cluster using elbow method and silhouette score. (a) Using all district data. (b) Excluding selected districts.*

Nevertheless, a k-means clustering implementation was still attempted. Using the data for all districts, there is no clear optimal number of clusters based on the elbow method, while the silhouette method suggests two clusters as the best solution (Fig. 6a). However, a visual inspection of the two-cluster solution using parallel coordinate plots suggests that the districts are divided between areas of high and low crime. While this further supports the Weisburd's Law of Crime Concentration mentioned earlier, it does not provide any additional information based on crime type. Increasing the number of clusters to three led to an additional cluster identifying one of the outlier districts. Further adding the

number of clusters resulted in at least another cluster with only two observations also comprising of the outlier districts identified via the MDS projection.

Therefore, three of the outlier districts were removed and the process was repeated. However, the MDS projection (Fig. 5b) and elbow method (Fig. 6b) still do not show any obvious optimal number of clusters. The silhouette score similarly suggests two clusters as the optimal which again reflects areas with low and high crime rates. Solutions up to five clusters were inspected, as the silhouette scores declined for solutions with more than five clusters.

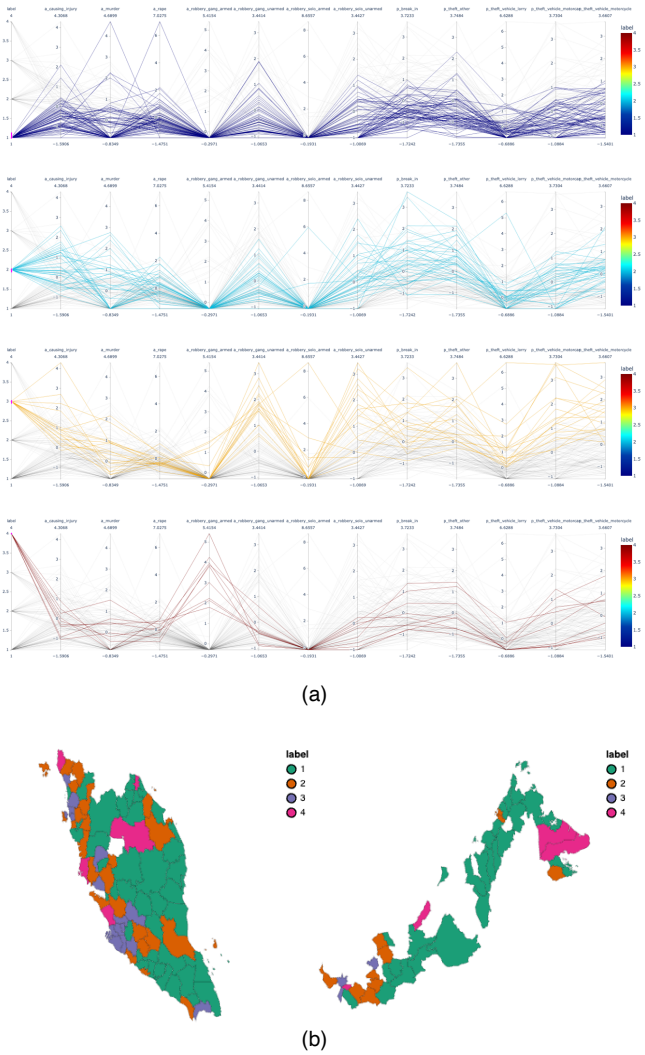


*Fig. 7: K-means clustering results using crime data. (a) parallel coordinate plots for each cluster. (b) Spatial visualization of the resulting clusters.*

An inspection of the parallel coordinate plots suggests that the four-cluster solution is the best candidate to provide further insights (Fig. 7a). The first cluster identifies areas that have generally low crime rates across all crime types, which makes up most of the districts. The second cluster appears to be districts with higher break-ins and other thefts, while the third clusters are districts where all property-related thefts are relatively higher. The fourth cluster likely identifies districts where armed gang robberies are elevated. A preliminary spatial analysis via a choropleth map (Fig. 7b) suggests that cluster three mainly comprise of districts with higher population density and greater economic activity. However, for the other clusters, geographical interpretation was unable to be established due to lack of sufficient contextual information.

The projection and clustering procedure was also repeated using spatio-temporal data with each crime type for each year treated as a separate feature. Similar to the exercise of using only spatial variation data, there were no obvious clusters identified by the projection and elbow method. Due to the interpretation challenges and significant computational resources involved, the approach did not yield any meaningful insights.

*RQ3: What are the district characteristics associated with variations in crime?*

Only some socio-economic variables exhibit a moderate relationship with crime rates (Fig 8a), which seem to weaken post-COVID19 pandemic. However, as highlighted earlier, some administrative districts in East Malaysia do not have a police precinct present. This may imply crime in these districts are reported to adjacent districts and hence distorts the relationship between crime and neighbourhood characteristics. Thus, the correlation matrix was also visualised using observations only from West Malaysia (Fig 8b). The relationships now appear slightly more prominent as shown by the darker shades in the chart. Nonetheless, no factors appear to be strongly correlated with crime, and the relationship still appears to weaken post-pandemic.

For variables that are moderately correlated with crime, the causal pathway remains to be determined. Crime rates appear to be lower where there is a higher proportion of kids and young adults. This could be attributed to the 'community effect', where more children suggest the presence of more families, fostering a stronger sense of community. Alternatively, it could reflect the 'opportunity effect,' where a higher proportion of older individuals may indicate a larger population more vulnerable to crime. Crime rates are also higher in areas with higher income and lower poverty. This suggests that crime events may be motivated more arising by 'availability of opportunity" as opposed to "the need to survive". Education-related variables are found to be negatively correlated with crime rates. This perhaps suggests education as an effective tool to lower crime.

Additionally, by inspecting the full correlation matrix of the dataset, variables that exhibit moderate correlation with crime are also observed to be moderately correlated to each other. For example, the proportion of Chinese and Indian ethnicity in the district population seem to also be moderately positively correlated with income and proportion of elderly population. Therefore, more advanced modelling methods are required to determine the potential causal effect between these variables and crime rates, which is out of scope for this paper.

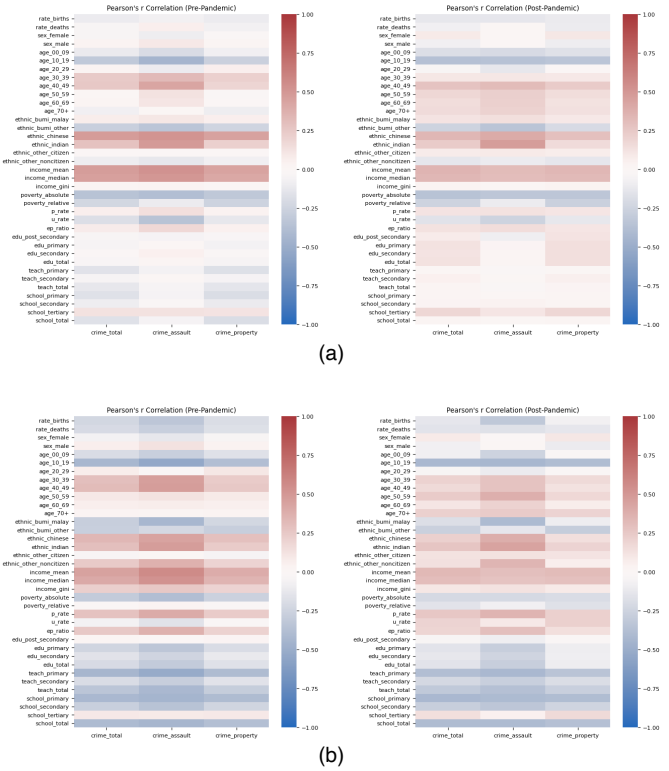


*Fig. 8: Correlation between socio-economic variables and crime rates. Positive correlations are indicated by shades of red while negative correlations are indicated by shades of blue. (a) Using all district data. (b) Using only data for districts within West Malaysia.*

The clustering procedure is then repeated using both crime and socio-economic data. This procedure is repeated twice: once using data from all districts and again using data only from districts within West Malaysia. Only moderately correlated variables were used for the clustering, identified via the correlation matrix in Fig. 8. Inspection of the parallel coordinate plots of the various cluster solutions reveals that the clustering algorithm primarily identifies similarities in terms of neighbourhood characteristics, while crime types remain relatively dissimilar within clusters. This observation further supports the hypothesis of weak or non-existent relationship between crime and neighbourhood characteristics at the district level.

This finding is further supported by examining the neighbourhood characteristics within the crime clusters identified earlier using only crime data. There is significant overlap of the distribution of neighbour characteristics across the clusters, which suggests that the crime clusters are not perfectly separable based on their neighbourhood characteristics (Fig. 9). Only the proportion of Chinese and Indian ethnicity seems to show potential to identify one of the clusters. However, as previously noted, these variables are also correlated with other factors. Therefore, it is important to explore the causal pathways more thoroughly before drawing definitive conclusions, as a premature or oversimplified interpretation could risk exacerbating social divisions and undermining social cohesion within the country.

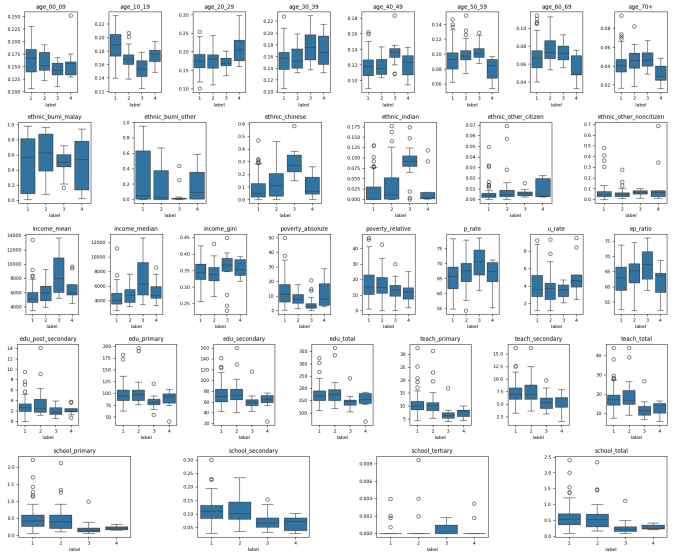

*Fig. 9: Distribution of socio-economic variables by the clusters identified using crime data only.*

### 4.3 Results

Since 2016, crime rates In Malaysia have generally declined and was most evident during the years of the COVID19 pandemic. While there are some spatial differences in the temporal pattern, areas of high crimes tend to remain the same.

There seems to be limited similarities in terms of crime types across districts. Using a k-means clustering method and after removing outliers, the districts seem to be able to be clustered into four different groups based on the profile of their crime types, although the supporting evidence for the clusters as measured by the silhouette score is relatively low.

Some socio-economic variables show a moderate correlation with crime rates. Specifically, age, income, and labour force variables are observed to be positively correlated with crime, while education-related variables are observed to be negatively correlated. There also seems to be a positive correlation between the proportion of certain ethnicities with crime rates, but the causal pathway remains uncertain as they are also correlated with other variables. No meaningful insights on crime rates were obtined based on the clustering attempt using both crime and socio-economic variables.

## 5 Critical Reflection

This paper explores the temporal and spatial patterns of crime rates in Malaysia using district-level data. The use of distributional plots together with small multiples allows for a quick overview of the aggregate and sub-group temporal trends. To utilise the full value of the dense pixel display, human guidance and reasoning is needed to sort the observations such that useful insights can be extracted.

Otherwise, a random sorting of the observations may appear as noise.

The most challenging aspect of this analysis is the clustering implementation. Although the 'elbow' method and the silhouette scores are approaches to provide an indication of the optimal cluster number, human reasoning and intuition is still required to ensure the insights are meaningful. For example, the 'elbow' charts in Fig. 6 lacks a clear "elbow", hence a literal interpretation may lead to an implementation of the highest number of cluster possible. However, interpreting the various clusters may prove to be challenging. Alternatively, the silhouette score in Fig. 6 is the highest for two clusters, thus a full-reliance on the silhouette scores would lead to an implementation of two-cluster solution. However, as mentioned earlier, the two-cluster solution is mainly reflecting areas of high and low crimes. While useful in certain contexts, it does not provide any further insights based on crime types. Additionally, there are potentially simpler methods to obtain similar insights, such as dividing the districts into quintiles.

The inclusion of socio-economic variables into the clustering further adds the interpretation complexity. An interactive parallel coordinate plot, created using the Plotly package in Python, significantly facilitates the sensemaking aspect, as users can filter individual clusters separately. Nevertheless, the inability to produce meaningful inference from the clustering process may partially reflect the limited domain knowledge to guide the human reasoning, which increases in importance as more data is introduced.

One of the main limitations of this analysis is the granularity of the data. While district-level data may be useful to provide administrative insights, one of the aspects of crime theory emphasises the importance the relationship between micro-scale locations and crime [6], [9]. For example, crime events may be concentrated within a specific street segment. This means that the relevant "population-at-risk" and the "neighbourhood characteristics" should be the ones in the immediate surrounding area. Analysing such data at a more aggregated level would likely dilute any relationship and may likely be the driving factor behind the moderate correlation seen in Fig. 8 between crime rates and the other variables.

**Table of word counts**

| | |
|---|---|
| Problem statement | 250 / 250 |
| State of the art | 500 / 500 |
| Properties of the data | 474 / 500 |
| Analysis: Approach | 480 / 500 |
| Analysis: Process | 1461 / 1500 |
| Analysis: Results | 182 / 200 |
| Critical reflection | 411 / 500 |

## REFERENCES


[1] J. Marquez, L. Taylor, L. Boyle, W. Zhou, and J.-E. De Neve, "Child and Adolescent Well-Being: Global Trends, Challenges and Opportunities," in *World Happiness Report 2024*, J.F. Helliwell, R. Layard, J.D. Sachs, J.-E. De Neve, L.B. Aknin, and S. Wang, Eds., University of Oxford: Wellbeing Research Centre, 2024, pp. 65. [Online]. Available: https://worldhappiness.report/ed/2024/

[2] United Nations. "Goal 16: Progress and Info." *Sustainable Development Goals*. Accessed: Dec. 26, 2024. Available: https://sdgs.un.org/goals/goal16#progress_and_info

[3] Department of Statistics Malaysia. "Data Catalogue." *OpenDOSM.* Accessed: Dec. 26, 2024. Available: https://open.dosm.gov.my/data-catalogue

[4] D. Runfola *et al.*, "geoBoundaries: A global database of political administrative boundaries" *PLoS ONE* 15(4): e0231866, Apr. 2020. doi: 10.1371/journal.pone.0231866.

[5] R. Maciejewski *et al.*, "A Visual Analytics Approach to Understanding Spatiotemporal Hotspots," *IEEE Trans. on Visualization and Comput. Graph.*, vol. 16, no. 2, pp. 205-220, Mar.-Apr. 2010, doi: 10.1109/TVCG.2009.100.

[6] M. Saraiva, I. Matijošaitienė, S. Mishra, and A. Amante, "Crime Prediction and Monitoring in Porto, Portugal, Using Machine Learning, Spatial and Text Analytics," *ISPRS Int. J. of Geo-Inf.,* vol. 11, no. 7, Jul. 2022, Art. no. 400, doi: 10.3390/ijgi11070400.

[7] G. Garcia *et al.*, "CrimAnalyzer: Understanding Crime Patterns in São Paulo," *IEEE Trans. on Visualization and Comput. Graph.*, vol. 27, no. 4, pp. 2313-2328, Apr. 2021, doi: 10.1109/TVCG.2019.2947515.

[8] L. Silva, S. González, C. Almeida, S. Barbosa, and H. Lopes, "CrimeVis: An Interactive Visualization System for Analyzing Crime Data in the State of Rio de Janeiro," *Proc. of the 19th Int. Conf. on Enterprise Inf. Syst.,* vol. 1, pp. 293-200, 2017, doi: 10.5220/0006258701930200.

[9] J.E. Eck, "Crime Hot Spots: What They Are, Why We Have Them, and How to Map Them" in *Mapping Crime: Understanding Hot Spots (NCJ 209393)*, A. Gonzales, and R. Schofield, S. Hart, Eds., U.S. Dept. of Justice, Office of Justice Programs, 2005, pp. 1-14. [Online]. Available: https://www.ojp.gov/pdffiles1/nij/209393.pdf