**A Coursera Capstone Project**

# Battle of the Neighbourhoods

# Clustering suburbs in metropolitan Melbourne

Xiaopeng Gong

July 28, 2021

# Introduction

## **Background**

A start-up company plans to expand and grow their business in Australia in the next five years. The company is planning to establish the headquarter office in Docklands, Victoria 3008. Furthermore, twenty of the senior-level employees will be asked to relocate to Melbourne and supervise the progress of this business establishment. As a part of the plan, the company will provide accommodations to their senior-level employees to help them transit into new locations seamlessly and reduce the negative influence of relocation on their work efficiency.

## **Criteria**

According to the company's founders, their selection criteria of accommodations' locations must be in the suburbs with the following features:

1. 15 km in radius distance or equivalently 30mins travel time to the office during peak hours (Office will be in Docklands, VIC 3008)
2. An affordable rental price with a large volume of rental properties available in the market
3. A safe area with low assaulting class crime incidents
4. Easy to access venues such as restaurants, cafes, and parks.

## **Objectives**

This data science project aims to apply clustering to categorise the suburbs in Metropolitan Melbourne regarding recorded crime incidents, venues categories, and rental status in 2020. In addition, a feature analysis will be carried out on each cluster. Finally, the cluster with the desired features will be recommended to the company as suitable accommodations locations.

# Data acquisition and cleaning

In this project, all data were collected from free and publicly available datasets.

## **Data requirements**

Based on the project's objective, the datasets used in this project will cover the following sectors:

* General information by suburb including name, postcode, coordinates, and council which it belongs to
* Types of crime incidents and their frequencies recorded by suburb
* Number of rental properties available and their median prices by suburb
* Categories of venues and their numbers by suburb

## **Data source**

* General information of suburbs in Metropolitan Melbourne (https://en.wikipedia.org/wiki/List\_of\_Melbourne\_suburbs)
* Nominatim API to extract geographical coordinates of each suburb (https://nominatim.org/release-docs/develop/api/Overview/)
* Crime Statistics Agency Data Tables - Criminal Incidents (https://discover.data.vic.gov.au/dataset/crime-by-location-data-table)
* Rental Report - Quarterly: Moving Annual Rents by Suburb (https://discover.data.vic.gov.au/dataset/rental-report-quarterly-moving-annual-rents-by-suburb)
* Foursquare API to extract venues information (https://developer.foursquare.com/docs/places-api/)

## **Data cleaning**

This section aims to collect and prepare four datasets, including geographical dataset, crime record dataset, rental dataset, and venue dataset. Data downloaded or scraped from the above sources were cleaned individually and then merged into one table.

After a quick preliminary examination of each dataset, a few missing values were found. A few duplicated data entries exist in geographical attributes since some suburbs belong to two different councils simultaneously. Fortunately, the number of duplications is not significantly high, and there are only 547 instances in the table. Thus, searching for duplicated suburbs through the entire table does not increases the computational cost dramatically. Hence, the solution to this problem is to search the entire table, locate the duplicated rows, and remove them.

Furthermore, with a few searches on google, it can be found that some suburbs placed in the geographical dataset have been given the wrong postcodes. Replacing these incorrect postcodes is essential to maintain consistency throughout datasets when merging. However, it is challenging to correct them manually. One of the potential solutions is to compare the postcodes and suburb names with ones given in the crime dataset, manually check the pairs that contain differences, and replace them with the correct postcode numbers.

Regarding the crime dataset, the suburbs outside of metropolitan Melbourne were removed from the dataset. Furthermore, there are still 15 suburbs without having criminal records. However, by checking them on Google Maps individually, it is safe to say that these suburbs are too far away from the central areas of Melbourne. Therefore, the suburbs having missing crime records will be ignored in this case.

Several small suburbs were combined to represent the rental status in those local areas in the rental dataset. In order to merge with other datasets, it is vital to maintain consistency in suburbs' names with others. Therefore, the row contained multiple suburbs were separated into rows that represent several individual suburbs. The rental counts for the local area were divided evenly among each suburb, while the median price remained unchanged.

On the other hand, several suburbs have no rental data. In this case, the missing data were determined based on the surrounding suburbs' rental status. The strategy is to use the coordinates of the target suburb as the centre to draw a circle. A threshold value was defined as the circle's radius to distinguish whether the suburbs are within the selected range. Subsequently, utilise the Latitude and Longitude values to estimate the Euclidean distance from the centre to each suburb. Accordingly, Select the suburbs within the boundary and calculate the mean values of Count and Median in the surrounding suburbs. At last, replace the missing values in Count and Median using the resultant values.

In the later work, the prepared datasets are implemented in conjunction with machine learning models to explore the characteristics of suburbs in metropolitan Melbourne. Accordingly, the results are used to determine the suitable suburbs for the client which meet their requirements.

# Exploratory Data Analysis

In this section, an exploratory dataset analysis was conducted to investigate the characteristics of each dataset. Accordingly, the datasets were pre-processed based on the results of the analysis.

## **Narrow down the search scope**

The geographical dataset contains geographical identifications, including name, postcode, council, and coordinates for every suburb, in Metropolitan Melbourne. From the geographical dataset, it can be found that there are 545 suburbs distributed in 31 councils in Metropolitan Melbourne. Figure 1 demonstrates that Shire of Yarra Ranges has the most suburbs, whereas the City of Bayside has the least suburbs.

Chart

Description automatically generated with medium confidence

Figure 1 Distribution of suburbs in metropolitan Melbourne

The first task is to narrow down the search scope and find out the suburbs that have the travel time to the office that is less than 30 mins during peak hours. One solution is to determine the boundaries with the drive time equal to 30mins to the office during peak hours on the map. Accordingly, the suburbs within limits will fit under the criterion.

TravelTime's Isochrones API can determine the shape of the less-than-30mins zone. When the arrival coordinates and travel time are given, the API call can return the coordinates of the zone's shape reachable within the corresponding travel time.

Figure 2 indicates the boundaries of the reachable zone within 30mins drive time to the office during peak hours. Furthermore, an interactive choropleth map has been plotted in the notebook.

Map

Description automatically generated

Figure 2: Boundaries of the reachable zone within 30mins driving time to the office during peak hours.

According to Figure 2, it is evident that the suburbs fall into less-than-30mins zone are primarily distributed in the councils of

* City of Melbourne,
* City of Port Phillip,
* City of Hobsons Bay,
* City of Stonnington,
* City of Glen Eira,
* City of Yarra,
* City of Moreland,
* City of Darebin,
* City of Maribyrnong,
* City of Moonee Valley,
* City of Boroondara,
* City of Bayside

Therefore, after refining the search scope, the number of potentially suitable suburbs are reduced from 547 to 127.

## **Crime Incident Report 2020**

Figure 3 presents the sums of crime incidents corresponding to each crime division in 2020. From the result, it is evident that most crime incidents have been classified as Properties and deception offences in 2020. The total number of B-class crimes is more than approximately 60000 incidents. Meanwhile, the A-class crime incidents, Crimes against the person, comes as the second in total. It is also notable that the difference in number between the B class crime and the rest classes is considerable.

A picture containing chart

Description automatically generated

Figure 3: Crime incidents in each crime division for 2020

Figures 4 a-f demonstrate the histograms of crime incidents distribution of each crime division for each selected suburb. Most suburbs recorded the crime incidents in the same range within each crime division, whereas one suburb has crime incidents far more than others. Such data distributions could potentially lead to poor segmentations for some clustering algorithms. However, a scaler can be applied to transform the attributions into more bell-shaped distribution to improve the metric score in clustering.

Figure 5 shows the crime incident statistics 2020 in the sub-division classes. The incidents distribution in each crime subdivision is consistent with the results observed in Fig.3. Furthermore, it is noteworthy that sub-division class, B40 Theft, has the most frequent occurrence among all sub-divisions. In addition, the F90 Miscellaneous Offences also recorded a total number close to 15000 cases, which was placed in the second frequent crime in 2020. However, the discrepancy in numbers between the B40 and F90 classes is significant.

A picture containing shoji, crossword puzzle, window, building

Description automatically generated

Figure 4: Histogram of crime incidents distribution in each crime division

Chart

Description automatically generated

Figures 6 a-g illustrate the histograms of crime incidents distribution of each crime subdivision for the selected suburb. Again, the same tendencies were observed as the ones in Fig. 4.

A picture containing shoji, window, building, tall

Description automatically generated

Figure 6: Histograms of crime incidents in every subdivision for selected suburbs.

* 1. **Rental Report 2020**

Chart, histogram

Description automatically generated

Figure 7: Histograms of (a) average volumes and (b) average median prices of rental properties of selected suburbs.

Figures 7a and b demonstrate the histograms of the number of rental properties and the rental price, respectively. The first notable feature is that the scales of these attributes are very different. Furthermore, both histograms are tail-heavy that potentially could affect the metric score for some clustering algorithms. Therefore, the rental dataset needs to be transformed later to ensure compatibility with the selected clustering algorithm.

* 1. **Venue Categories**

From the pie chart demonstrated in Figure 8, it is notable that Cafés are the most popular venues in the selected suburbs, following by Bar and Pizza places.

From Figure 9, most suburbs have the total number of venues in the same bracket, ranging from 0 to 20. On the other hand, a small group of suburbs possess a significantly large number of venues than the rest. Furthermore, the total number of venues was capped at 100 due to the limited access through API requests. Fortunately, it does not seem to have a substantial impact on the dataset since only a few suburbs have venues that are more than 100.

From Figure 10, it is evident that the dataset has a heavy-tailed distribution. The data extends much farther to the right of the median than to the left. Again, it might be worth trying to transform the dataset into a more bell-shaped distribution to improve compatibility with some clustering techniques.

A picture containing compact disk, vector graphics

Description automatically generated

Figure 8: Pie chart of major venues (Total number > 10) of selected suburbs in metro Melbourne.

A screenshot of a computer

Description automatically generated with low confidence

Graphical user interface, application

Description automatically generated

Figure 10: Histograms of top three venues in selected suburbs.

# Clustering Suburbs in Metro Melbourne

Based on the preliminary analysis and exploration of the dataset, it is evident that most attributions are heavily tailed towards the right of the median. In this case, K-Means can be an excellent choice among the clustering algorithms. Since the K-Means is relatively simple to implement and it generalises to clusters of different shapes and sizes. More importantly, it guarantees convergence. However, K-Means is strongly dependent on the initial assumption of the number of clusters. Due to the complexity of the dataset, it is challenging to assume the number of clusters solely based on the preliminary dataset exploration. Therefore, an experiment can be used to find out the value of k.

Before starting to explore the values of k, it is essential to pre-process the dataset for training.

* 1. **Data Pre-processing**
     1. **Scaling the dataset**

Based on the preliminary analysis of the datasets, it is evident that the difference in scale between each attribute is significant.

If the data were implemented to the K-Means model without normalising, each attribute's scale would serve as a weight. As a result, the outcomes of Euclidean distance calculation are more sensitive to the attributes with a large magnitude while ignoring other attributes with relatively small values. Thus, the movement of centroids would be determined solely based on attributes with large scale values. In order to avoid such biased clustering, it is vital to ensure that the data is standardised.

In this study, the MaxAbsScaler() will transform each feature by its maximum absolute value. Unlike StandardScaler(), it does not shift or centre the data. Therefore, it does not change the sparsity in each attribute.

* + 1. **Dimensionality Reduction**

Kernel PCA was applied to transform the training dataset into a two-dimensional space to avoid suffering from the curse of dimensionality before conducting K-Means clustering. Such dimensional reduction could potentially speed up the training and improve data visualisation during clustering.

The mean squared error was implemented to measure the reconstruction pre-image error after the transformation. Additionally, the best hyperparameters for the Kernel PCA to yield the lowest reconstruction pre-image error were determined using the GridSearchCV().

Chart, scatter chart

Description automatically generated

Figure 11: Training dataset transformed into two-dimensional space using KPCA.

* 1. **K-Means Clustering**
     1. **Find the best k value for K-Means Clustering**

One of the main assumptions made before applying K-Means clustering is to assume the number of clusters. However, it is impossible to segment the data from the above graph manually. One way to optimise the k value is to find the equilibrium between k and the inertia associated with the silhouette score while iterating the k value, ranging from 2 to 30.

From Figure 12, the inertia decreases as the k increases, indicating that the distance between data points and their corresponding centroids decreases. In addition, a significant drop in inertia was observed when k = 4, indicating that k = 4 might be a potential candidate for the optimal value of k. However, relying solely on evaluating the inertia between centroids and data during K-Means clustering can be a coarse method to determine the optimal k.

Shape

Description automatically generated

Figure 12: The 'elbow' method to find the best value of k for K-Means clustering.

An alternative is to examine the silhouette coefficient, also known as silhouette score.

Chart

Description automatically generated

Figure 13: Silhouette scores of current K-Means clustering model using k values from 2 to 30.

In Figures 13, the silhouette scores experienced a significant decrease after k=2 and fluctuated in the same interval as increasing k, having a mean value of approximately 0.4. Furthermore, k=4 has the highest silhouette score, excluding the case of k=2. This finding is consistent with the inertia evaluation, suggesting that k=4 might be an optimal value for k.

A picture containing shoji, building

Description automatically generated

Figure 14: Silhouette diagrams of top ten silhouette scores with k values.

Figure 14 demonstrates the silhouette diagram. Each diagram contains one row per cluster. The row's height represents the number of instances the cluster contains, whereas the width illustrates the sorted silhouette coefficients of the instance in the cluster. In addition, the silhouette coefficient varies between -1 and 1.

+1 means the instance is well close to the centroid.

0 means the instance is at the boundary

-1 means the instance is assigned to the wrong cluster.

Meanwhile, the vertical dash line indicates the overall silhouette score. From Fig. 14, it is evident that k=2 and k=4 are the potential optimal settings for the number of clusters because the silhouette score for each cluster exceeded their corresponding overall silhouette score. Additionally, nearly 75% of instances are categorised in cluster 0 when k=2. Similarly, cluster 2 is alternatively large when k=4, suggesting it also includes significantly more instances than others. Furthermore, K=9 produces excellent segmentation since the silhouette coefficients of each cluster vastly outperform the overall silhouette score. Additionally, the distribution of instances within every cluster is very even. Therefore, although k=2 or k=4 has the higher silhouette score, k=9 is a better alternative to get the even dimension clusters.

* + 1. **K-Means Clustering when K = 9**

Figures 15 and 16 represent the segmentations of the training data projected into a two-dimensional space using K-Means Clustering with hyperparameter K = 9.

Chart, scatter chart

Description automatically generated

Figure 15: Centroids of clusters

Chart, radar chart

Description automatically generated

Figure 16: Boundaries between clusters with centroids

1. **Results and Discussion**
   1. **Feature Analysis**

In this section, the characteristics of each cluster have been recognised using the features involving geographical information, crime state, rental status, and venue activities.

**Chart

Description automatically generated**

Figure 17: The distribution of suburbs in every cluster.

Figure 17 presents the distribution of suburbs in every cluster. Clusters 3 has the least number of suburbs, whereas Cluster 2 and Cluster 6 hold 30% of suburbs, making them the clusters with most suburbs. The rest suburbs were distributed almost evenly in the rest of the clusters.

Figure 18 illustrates the average Euclidean distance to the office for every cluster. Clusters 2, 4, and 5 have significantly different average Euclidean distances compared to the other clusters. Cluster 1 has the shortest average Euclidean distance to the office, whereas Cluster 4 holds the second least. By contrast, Cluster 5 possesses the longest average Euclidean distance to the office. Meanwhile, Cluster 0, 2, and 6 have the same average Euclidean distance to the office, whereas Cluster 3, 7, and 8 have the same.

Figure 19 demonstrates the distribution of categorised crime incidents on average among clusters in 2020. Cluster 1 has the most significant number of crime incidents on average among clusters, whereas Cluster 6 contains the least. On the other hand, compared with Cluster 1, Cluster 3 only possesses half the amount of crime incidents on average. In addition, it is evident that the total number of crime incidents in Cluster 7, 4, 0, and 2 ranked in the middle section and declined linearly.

Chart

Description automatically generated

Figure 18: Avg. Euclidean distance to the office for every cluster.

Chart, funnel chart

Description automatically generated

Figure 19: The distribution of categorised crime incidents on average among clusters in 2020

Figure 20 presents the average percentage of crime occurrences for subdivisions among clusters in 2020. In terms of crime percentage distribution, crime subdivisions are relatively consistent across clusters. Variations in percentages among clusters are observable but not statistically significant.

In 2020, approximately 40%~50% of crimes incidents were classified as B40 theft, making it the most prominent crime subdivision. In addition, miscellaneous crimes are the second-most common crime in all clusters, fluctuating between 10% and 15%. Five types of crimes occur between 6-10% of the time, including assaults and related offences, property damage, burglaries, deceptions, and breaches of orders.

Chart, bar chart

Description automatically generated

Figure 20: The average percentage of crime occurrences for subdivisions among clusters in 2020

The boxplots in Figure 21 (a) depict the numbers of rental properties and rental prices within each cluster.

Among the clusters, Cluster 1 has the highest median number of rental properties at approximately 8800 units, whereas Cluster 8 has the lowest, at approximately 2500 units.

The most pronounced interquartile range belongs to Cluster 4, ranging from 2600 to 9000 units, whereas the middle 50% of data found in Cluster 3 is the most compacted, varying from 3500 to 4200 units. A boxplot works best when the sample size is at least 20. However, Cluster 3 only contains four samples, thus, making the quartiles found in Cluster 3 less meaningful. Cluster 5 also exhibits compacted interquartile ranges, which are between 2800 and 3600 units.

The data in Cluster 1, 3, 6, 7, and 8 are normally distributed. In contrast, the data obtained from Cluster 0 are negatively skewed, whereas those obtained from Clusters 2,4,5 are positively skewed.

The clusters 0, 1, 4, 5, and 7 contain outliers, respectively. Moreover, Cluster 1 and Cluster 4 have the highest outliers and exceeds approximately 23000 units, whereas Cluster 5 has the lowest outlier, at approximately 1000 units.

Figure 21 (b) illustrates the boxplots of rental prices among clusters.  Cluster 8, with a median rental price of approximately A$510, has the highest median price among clusters, while Cluster 3 has the lowest median price, at approximately A$420. Meanwhile, the median rental prices for the rest of the clusters fluctuates between A$440 and A$480.

It was evident from the interquartile ranges of clusters that there was a discrepancy among them. However, it is less apparent than those found concerning the number of rental properties. Furthermore, the middle 50% of data obtained in Cluster 0 is widely dispersed, while Clusters 2 and 4 have a more concentrated pattern.

It is important to note that the data found in Cluster 1, 2, 4, and 5 are normally distributed, whereas those obtained in the rest of the clusters are positively skewed. Furthermore, Cluster 3 and Cluster 7 are the only clusters with outliers.

A picture containing graphical user interface

Description automatically generated

Figure 21: Boxplots of (a) average volumes and (b) average median prices of suburbs in each cluster.

The following table flaunts the ten most common venues by cluster. There is no denying that residents of metro Melbourne enjoy their coffee and brunch since Café is the topmost frequented venue in nine of ten clusters. However, even though the ten most common venues have been filtered, the pattern in Table 1 is not discernible. In order to recognise the characteristics in each cluster, one of the solutions is to classify each venue into more general categories. Thus, a further labelling process is required.

* 1. **Cluster Profiles**

An initial understanding of the characteristics of each cluster has been developed after preliminary investigations regarding crime, rental, and venue activities. Now, it is essential to analyse every cluster in dimensions that focus on the corporate requirements.

Let us rewind the requirements:

* 15 km in radius distance or equivalently 30mins travel time to the office during peak hours (Office will be in Docklands, VIC 3008)
* An affordable rental price with a large volume of rental properties available in the market
* A safe area with low assaulting class crime incidents
* Easy to access venues such as restaurants, cafes, and parks.

According to the above requirements, the desired features are joined to the cluster labels to profile each cluster. Following is a list of desirable attributes:

**Requirement 1**

* Euclidean distance

**Requirement 2**

* Count
* Median

**Requirement 3**

* A20 Assault and related offences
* A50 Robbery
* A70 Stalking, harassment and threatening behaviour
* A80 Dangerous and negligent act endangering people
* B10 Arson
* B20 Property damage
* B30 Burglary/Break and enter
* B40 Theft

**Requirement 4**

* Food place
* Café
* Leisure
* Shopping
* Transport

It is relatively straightforward to select desired features among clusters for considering requirements 1 and 2. In addition to selecting the assaulting type of crime, crimes classified in subdivisions that could induce or be associated with assaulting type crimes such as arson, property damage, burglary/breaking in, and theft were also included concerning requirement 3. At last, when considering requirement 4, the patterns hidden inside of venue categories are hardly recognisable since the original categories of venues are overly sophisticated under the current scope of analysis. One of the solutions is to create a set of more general categories to ensure the extent aligns with the current scope of requirement 4. As a result, five categories regarding venues are introduced to analyse every cluster, including Food places, Café, Leisure, Shopping, Transport.

* + 1. **Feature Scaling**

It is essential to scale the features within a range so that the comparison between every cluster is meaningful. In this case, the MinMaxScale() was used to scale every selected attribute into the specific range between (0,1).

* + 1. **Profiles of Clusters**

In this section, the desired features were visualised in Figures 22-27. Accordingly, each cluster was described with references to these selected attributes to help the client find suitable suburbs to rent properties for their oncoming overseas employees.

Chart, histogram

Description automatically generated

Figure 22: Overview of desired attributes for each cluster.

A picture containing text, clipart

Description automatically generated

Figure 23: The average Euclidean Distance from each cluster to the office location.

A group of colorful pencils

Description automatically generated with low confidence

Figure 24: The average crime incidents in selected crime subdivisions for each cluster.

Chart, bar chart

Description automatically generated

Figure 25: The average volume and average median price of rental properties in each cluster.

Chart, histogram

Description automatically generated

Figure 26: The quantities of venue activities in each cluster.

**A picture containing person, outdoor, building, dark

Description automatically generated**Figure 27: Polar plots for each cluster regarding distance, crime, rental, and venues.

**Cluster 0**

Profile:

This cluster consists of suburbs possessing median stats across most sectors.

**Distance**

There is a significant distance between the suburbs and the office. Therefore, the travel time to the office is approximately between 20~30 mins during peak hours by driving. It is also worth mentioning that access to public transportations could be difficult since only a limited number of transportation facilities are available.

**Crime**

Furthermore, the crime rate reported in each category is slightly below the average.

**Rental**

Besides, affordable rental properties are relatively easy to find. There is a moderate number of properties available in the market, while the average rental price is slightly above the average compared to other suburbs.

**Venue**

Although a wide variety of activities covering all types of venues is perceived in these suburbs, the median number of venues across categories is slightly below the average, particularly for food places. The top three most popular venues are cafés, coffee shops, and supermarkets.

**Cluster 1**

Profile:

The suburbs found in Cluster 1 are highly commercialised, well entertained, yet, they have the highest crime rates on average across every crime category.

**Distance**

The suburbs are the closest to the office on average and offer easy access to public transportations. The average travel time to the office is less than 15mins by driving.

**Crime**

In addition, these suburbs have the highest average crime rates across all crime categories, suggesting a high risk of becoming a victim of crime.

**Rental**

On the other hand, these suburbs offer the biggest rental market with the third-lowest average rent prices, making affordable rental properties readily available.

**Venue**

Moreover, these suburbs have the most venues simultaneously, offering a wide variety of venue types. In contrast with the cluster with the second largest number of venues, the discrepancy of the number of venues in the same category exceeds at least 40%. The top three activities are cafés, bars, and Vietnamese restaurants.

**Cluster 2**

Profile:

In this cluster, the suburbs are 'quiet and remote'.

**Distance**

The average distance from these suburbs to the office is the second-longest among suburbs. The driving time during peak hours usually takes 25-30mins. Nevertheless, public transportation is relatively accessible, which may significantly shorten travel times.

**Crime**

The risk of being exposed to criminal activities is regarded as low. However, despite a low number of assaulting types of crime, the number of crimes such as stalking, harassment, and threatening behaviour is significant. In addition, it is notable that Burglary/Break and entry is the primary cause of crimes in these suburbs.

**Rental**

Despite a lower supply of rental properties available than average on the market, the rental price is relatively competitive compared to other suburbs.

**Venue**

There are only a few venue activities available, with the café ranking first. Meanwhile, the number of restaurants is significantly low in these suburbs. The top three popular venues are Cafés, Parks, and Pizza Place.

**Cluster 3**

Profile:

The high crime rate could be a concern for these suburbs in this cluster.

**Distance**

The travel distance to the office is slightly above average among clusters and the driving time is approximately 15-20mins on average. Furthermore, it is relatively easy to access public transportations.

**Crime**

Across all categories, the crime rate is considered high on average, ranking at the second highest among clusters. Furthermore, the likelihood of crime classified as dangerous and negligent acts endangering people is significantly higher than in other clusters. Meanwhile, the average number of criminal activities classified as stalking, harassment, threatening behaviour, arson, burglary, break, and entry is also more significant than in others.

**Rental**

This cluster possesses a relatively small rental property market. The average number of rental properties is low, while the average median price is significantly below average, with the second lowest among clusters.

**Venue**

On average, these suburbs have very few venue activities. Restaurants and recreational activities are available, but their numbers are limited. Furthermore, these suburbs have the smallest number of cafes on average compared to other clusters. In contrast, many shopping options are available in these suburbs compared with other venue categories. The most three typical venues are ice cream shops, vintage stores, bakeries.

**Cluster 4**

Profile:

The suburbs classified in this cluster are very well balanced in every sector.

**Distance**

These suburbs are the second closest to the office on average. During peak hours, driving to work takes about 10-15 minutes, but also, public transportations are very accessible in these suburbs.

**Crime**

Moreover, crime rates are consistently low across all categories. The most frequent crimes are theft, burglary, break and entry, and robbery. Additionally, it is worth mentioning that the distribution of the crime incidents across categories is reasonably like the one observed in Cluster 1 and 2.

**Rental**

Meanwhile, this cluster has the second-largest rental property market among clusters. Moreover, the rental price is at average compared with the other clusters. Therefore, it is evident that searching for a rental property at a reasonable price is notably straightforward.

**Venue**

These suburbs offer a comprehensive range of activities. Meanwhile, the average quantity of venues is the second largest amongst clusters. Additionally, apart from Cluster 1, the average volume of venues in each category is 60% more than other clusters. Especially in food places, the difference in number exceeds more than 300%. The most three typical venues are cafés, coffee shops, zoo exhibits among themselves.

**Cluster 5**

Profile:

The cluster consists of tranquil suburbs having the most distance to the office and monotonous activities.

**Distance**

The suburbs are highly distant from the office. In peak hours, driving to the office takes close to 30 minutes. Furthermore, public transportation is inaccessible in comparison with other suburbs.

**Crime**

Criminal activities are minimised to less than 5% of Cluster 1 on average across categories.

**Rental**

Meanwhile, a limited number of properties are available for rent in association with a reasonable average rental price.

**Venue**

In this cluster, the average number of venue activities is minimal. Cafés are the most common venue category. Additionally, the number of cafés is at the equivalent level as the one found in Cluster 2. The top three most common venues are cafés, pizza places, and bakeries.

**Cluster 6**

Profile:

This cluster contains suburbs that have the lowest numbers in both crime rates and rental prices on averages.

**Distance**

The average travel distance from these clusters to the office is slightly longer than the average. During peak hours, driving time is approximately 15-20 minutes. Meanwhile, the number of public transportation facilities is adequate, making public transportations relatively accessible.

**Crime**

These suburbs have the lowest average number of crime incidents amongst clusters. The only comparable crime category is dangerous and negligent acts endangering people.

**Rental**

The rental market has slightly fewer properties available than the average. However, the average rental price is the cheapest compared with other clusters.

**Venue**

While many activities can be observed in categories like cafés and shopping, dining places and leisure facilities are inadequate compared to those in other suburbs. The top three most common venues are cafés, fast food restaurants, and bakeries.

**Cluster 7**

Profile:

The character of this cluster is the high crime rate in conjunction with a high average rental price.

**Distance**

The average travel distance between the office and these suburbs is at the median level, and the driving time is between 15 and 20 minutes during peak hours. Moreover, the number of transportation facilities is adequate to provide easy access to public transportations.

**Crime**

The activity degree of criminal incidents is moderate in categories, such as stalking, harassment, threatening behaviour, burglary, and break and entry. In addition, the average numbers of crimes in the rest of the categories are not significant, yet, they are still above the median value.

**Rental**

The average number of rental properties is at the median value compared with other clusters. Meanwhile, the median rental price amongst these suburbs is comparatively higher than the median value amongst the clusters. It suggests that finding a rental property at an affordable price might be challenging.

**Venue**

The cluster contains a variety of venues. The number of restaurants is equivalent to the one obtained in Cluster 6, while the number of activities across the rest of the categories matches the numbers observed in Cluster 0. The top three most common venues are cafés, convenience stores, and Italian restaurants.

**Cluster 8**

Profile:

This cluster consists of suburbs possessing extremely high average rental prices with the best-ranking accessibility to public transportation.

**Distance**

Cluster 8 possesses an average Euclidean distance to the office equivalent to Cluster 3 and 7, slightly above the median amongst clusters. The commute takes around 15-20 minutes. In addition, public transportation is available everywhere, making them the most convenient suburb for transit.

**Crime**

These suburbs have a small number of crime incidents on average. In addition, criminal activities are evenly distributed across crime categories.

**Rental**

The highest average rental price in conjunction with the lowest rental properties makes it impossible to rent an affordable place in these suburbs.

**Venue**

A significant number of venue activities are observed in the cluster. The number of cafés, recreation activities, and shopping stores is considerably higher than in most clusters, whereas restaurants are close to the median. The top three most common venues are cafés, grocery stores, and bakeries.

* + 1. **Which clusters will be the best fit for our current objective?**

In order to answer this question, it is crucial to elucidate which sector or sectors that the client values the most. There could be multiple combinations of weights to address this question, and the client can decide the distribution of the weights based on their interests. In this case, we assumed the weights are evenly distributed across sectors.

* **Evenly distribute weights on sectors.**

Suppose the weights are evenly distributed between sectors. In that case, suburbs found in Cluster 0, 4, 6 could be the potential targets based on established profiles of each cluster above. These suburbs offer a safe, entertained, convenient, and affordable environment.

Among these suburbs, suburbs in Cluster 4 could become the best options compared with others. The most substantial advantage of suburbs in Cluster 4 is that they possess the shortest travel time in conjunction with the most convenient public transportation system and hold the most comprehensive range of venues in different categories. Furthermore, Although the suburbs in Cluster 4 have a significantly higher median rental price than the others in Cluster 6, the bigger rental market that suburbs in Cluster 4 possess offers a tremendous variety of options when hunting a rental property. Finally, regarding personal safety, the average number of crime incidents reported in the suburbs from Cluster 4 is significantly higher than those observed in Cluster 0 and 6. However, living in the suburbs from Cluster 4 is still considered safe since the average crime rates across categories are equivalent to the median value amongst clusters.

Table

Description automatically generated

1. **Conclusion**

In conclusion, based on our assumption that the weights are evenly distributed across commute time, safety, rental status, and surrounding activities, it is recommended to search the rental properties as the settlements for the coming overseas employees in the suburbs that obtained in Cluster 4. The advantage of these suburbs is a shorter distance travelling to the office with the second-best public transport system, a second-largest rental market in conjunction with a reasonable median rental price, and an enormous number of activities throughout categories. The only drawback of these suburbs is that the average crime rates across crime categories are not as low as some clusters, but living in these suburbs is still considered safe, especially against personal safety.

1. **Future Work**

A few limitations presented in this project, a plan of future works, could improve these aspects.

**Inconsistent accuracy of datasets**

The quality of datasets that we acquired from different sources is inconsistent. Especially the rental dataset has a lot of missing data across attributes. Although we managed to fill in the missing data with information obtained from surrounding suburbs, the fabricated data could inevitably cause skewness in the dataset, leading the data analysis to conclude any unrealistic insight. Furthermore, since all the datasets were collected from free and public domains, it is vital to ensure the accuracy and reliability of the data. Unfortunately, there is no validation process during data collection and clean in this practice. In future, it is possible to find more reliable data sources. Furthermore, it is also vital to implement a validation process to improve the accuracy and reliability of the data.

**K-Means clustering only**

In this practice, we only applied K-Means clustering to segment the suburbs. Although k-means clustering is an excellent choice in this simple case, it is interesting to find the best clustering method to ensure the best segmentations by exploring different clustering algorithms such as DBSCAN and hierarchy-based clustering (Agglomerative, BIRCH).