COMP20008 Assignment 2

Research Question

What is the relationship between the weekly rent and crime rate for suburbs in Victoria?

This research question would help understand the liveability for communities as both rent prices and crime rate are very important statistics for the residents of an area. Understanding the relationship between these two variables would help inform potential renters and home buyers what suburbs may be more liveable. Suburbs having low crime rates may also reflect greater inclusiveness; having residents that are looking out for each other is an effective method for deterring crime.

Datasets

We have used two datasets in total. Both data sets split the state of Victoria by LGA (Local Government Area). There are 79 LGAs in total.

The first dataset is Rental Report - Quarterly: Quarterly Median Rents by LGA, produced by the Department of Health and Human Services. It contains median weekly rent for properties in each LGA, along with the count of properties that were taken to calculate the median. It contains quarterly data for 6 different property types (1 bedroom flats, 2 bedroom flats, 3 bedroom flats, 2 bedroom houses, 3 bedroom houses, 4 bedroom houses) from 1999 to 2020.

The second dataset is Crime Statistics Agency Data Tables - Criminal Incidents, produced by the Crime Statistics Agency. It contains the number of criminal incidents per year in each LGA, along with the rate per 100,000 population. The data ranges from 2011 to 2020.

Preprocessing and wrangling

The first step was to convert the file type from .xlsx to .csv so that our program could take in the data and store them as dataframes. This was done in excel by saving each sheet within the excel file as csv files.

1 bedroom flats

		Jun 1999		Sep 1999		Dec 1999		Mar 2000		Jun 2000	
		Count	Median	Count	Median	Count	Median	Count	Median	Count	Medi
Barwon South We	Colac-Otway	6	$68	6	$75	7	$80	6	$80	12	
	Corangamite	-	-	5	$60	-	-	-	-	-	
	Glenelg	7	$65	-	-	13	$70	8	$80	9	
	Greater Geelong	202	$80	205	$80	200	$80	255	$85	185	
	Moyne	-	-	-	-	-	-	-	-	-	
	Queenscliffe	-	-	-	-	-	-	-	-	-	
	Southern Grampians	9	$65	10	$68	7	$70	10	$68	12	
	Surf Coast	9	$100	9	$92	10	$100	8	$118	10	
	Warrnambool	28	$75	22	$75	28	$75	35	$80	18	
	Group Total	***264***	***$80***	***261***	***$80***	***269***	***$80***	***327***	***$85***	***249***	
Grampians	Ararat	-	-	-	-	-	-	-	-	5	
	Ballarat	84	$80	83	$80	77	$80	146	$80	77	
	Golden Plains	-	-	-	-	-	-	-	-	-	
	Hepburn	-	-	-	-	-	-	-	-	-	

1br Flat | 2br Flat | 3br Flat | 2br House | 3br House | 4br House | All Properties

Figure 1: Raw data for Median Rent Prices

The rental price data was formatted with LGAs as rows, and median rent and counts and sub columns with the main columns containing the quarter in which the data was recorded. This was reformatted to a dataframe with rows for data from each year in each LGA. It was important to aggregate and average the data per year from the quarterly data to match the crime dataset. Rows that contained total counts and rent price for the larger regions were also omitted. Finally, all the data from before 2011 was dropped as there was no crime data to relate it to. These preprocessing steps were applied to all house types included in the rental price dataset.

Year	Year ending	Police Region	Local Government Area	Incidents Recorded	Rate per 100,000 population
2020	December	1 North West Metro	Banyule	6,717	5,074.3
2020	December	1 North West Metro	Brimbank	15,399	7,325.9
2020	December	1 North West Metro	Darebin	12,046	7,265.3
2020	December	1 North West Metro	Hobsons Bay	4,497	4,561.1
2020	December	1 North West Metro	Hume	16,216	6,717.0
2020	December	1 North West Metro	Maribyrnong	7,342	7,676.2
2020	December	1 North West Metro	Melbourne	26,850	14,320.0
2020	December	1 North West Metro	Melton	10,113	5,880.0
2020	December	1 North West Metro	Moonee Valley	6,463	4,894.6
2020	December	1 North West Metro	Moreland	10,829	5,738.6
2020	December	1 North West Metro	Nillumbik	1,632	2,504.1
2020	December	1 North West Metro	Whittlesea	12,155	5,129.4
2020	December	1 North West Metro	Wyndham	13,246	4,672.8
2020	December	1 North West Metro	Yarra	11,403	11,055.2
2020	December	1 North West Metro	Total	154,908	6,697.2
2020	December	2 Eastern	Alpine	413	3,212.0
2020	December	2 Eastern	Bass Coast	2,053	5,538.3
2020	December	2 Eastern	Baw Baw	3,690	6,749.4
2020	December	2 Eastern	Benalla	996	7,077.6

Contents | Table 01 | Table 02 | Table 03 | Table 04 | Table 05

Figure 2: Raw data for Crime rate statistics

The crime data was already in a much more workable format. The cumulative data over larger regions were omitted and the years were sorted in ascending order. Once this was done, both datasets contained rows for each year and LGA over 10 years, with 790 rows in total.

As many of the rows in the rental dataset had empty entries with a count of zero, these points were cleaned from the dataframes. It was decided that using regression to fill in these entries would lead to higher inaccuracies in the data as we observed that the surrounding data around the missing entries also had very low counts and thus were likely unreliable.

Both rent price and crime rate statistics were also discretised into categories of “low”, “medium”, and “high” in a separate column.

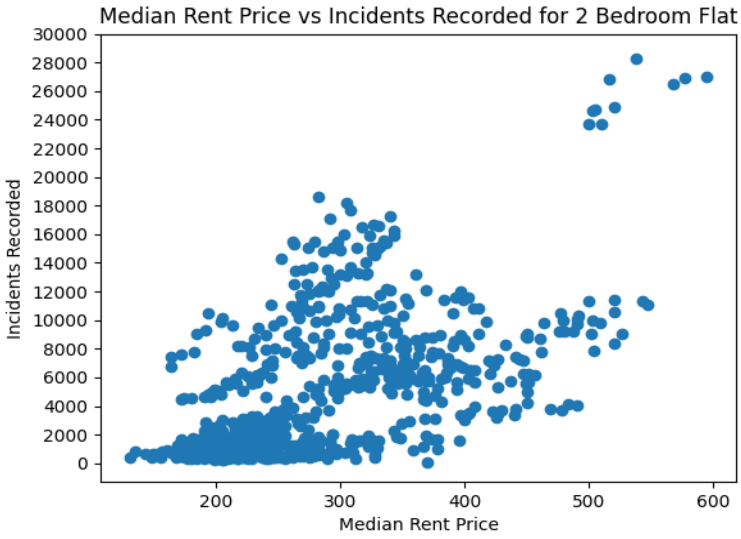
Analysis and Discussion of Data

Scatter plots

Scatter plots of “Median Rent Price” vs (Crime) “Incidents Recorded” as well as “Median Rent Price” vs (Crime) “Rate per 100,000 population” were produced for each housing type using all data from 2011 - 2020, as well as another plot for data in 2020 exclusively. This produced 4 plots for each housing type, making 28 plots in total. In general, a positive correlation between the median rent price and incidents reported can be visually observed. Only Figures for 2 bedroom houses and 2 bedroom flats have been included as representations of the other plots as they look very similar to the two.

In Figure 3 and 5, a positive correlation can be observed in an outward cone shape starting from the lower end of Median Rent Price, suggesting areas with lower rent had overall less incidents. However, when considering the population of each area (population per 100,000 plots), such as in Figure 4 and 6, there is a much weaker correlation between the two variables. From this, it is likely that the positive correlation seen in Figure 3 and 5 is a consequence of more populated areas simply having more crime. This would be logical as densely populated areas also have higher rent prices in comparison to sparsely populated areas.

This inconsistency suggests a lack of correlation between crime rate and rent price. Furthermore, in Figure 4 the plot seems to form a “U” shape with the crime rate peaking at both the lower and higher quartiles of Rent Price, perhaps informing home buyers/renters that areas near the extreme ends of Rent Price are less safe than buying a home closer to the average.



Pearsons correlation: 0.59584
Normalized Mutual Information: 0.2157194
Figure 3: Price v Incidents - 2 bedroom flat

Pearsons correlation: 0.31584
Normalized Mutual Information: 0.011258
Figure 4: Price v Rate per 100,000 - 2 bedroom flat

Pearsons correlation: 0.564593
Normalized Mutual Information: 0.212061
Figure 5: Price v Incidents - 2 bedroom house

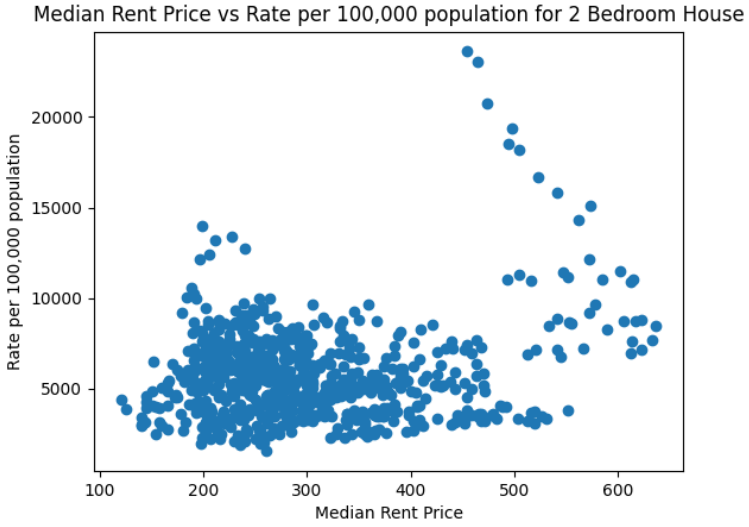


Figure 6: Price v Rate per 100,000 - 2 bedroom house
Pearsons correlation: 0.301339
Normalized Mutual Information: 0.00470358

Correlation and Mutual Information

Additionally, the Pearson's correlation coefficients and normalized mutual information for each respective plot were calculated and stored in a text file. The Pearsons correlation coefficients were moderately high (>0.30) for the population per 100,000 graphs. However, the trends in linearity are neither increasing or decreasing, and thus we cannot conclude anything about the trend between rent prices and crime rate. The normalized mutual information calculated were mostly low e.g. NMI = 0.011258 for 2 bedroom flat per 100,000 (Figure 3), further supporting the lack of correlation between crime rate and rent price. Based on these coefficients and the visualisations, we can conclude that it is very difficult to tell what the crime rate of an area would be based on the rent price.

Ranking of suburbs

The calculation of livability was done by taking the inverse of the product of crime rate and rent price. The statistics in Figure 7 could help home buyers/renters with different budgets when deciding on a location, in terms of the safety of the area. While these results are not extremely helpful on their own when deciding where to rent, they are significant and applicable to real world situations.

```
Top 10 (safe) Areas Ranked on Crime and Rent Price
Low Rent Price
('Towong', 3.6744307388177885e-06)
('Buloke', 3.1511029490542277e-06)
('Hindmarsh', 3.0503556398908057e-06)
('Indigo', 2.789856683795256e-06)
('Alpine', 2.5522441203969614e-06)
('Corangamite', 2.009488029809986e-06)
('Gannawarra', 1.9027608552872096e-06)
('Moira', 1.8144491538760267e-06)
('Yarriambiack', 1.806200810784861e-06)
('South Gippsland', 1.7694601934630554e-06)

Mid Rent Price
('Nillumbik', 1.5911001522935658e-06)
('Macedon Ranges', 1.5316056505495743e-06)
('Mount Alexander', 1.3041399585936583e-06)
('Hepburn', 1.2886047421323977e-06)
('Surf Coast', 1.210787136841835e-06)
('Yarra Ranges', 1.1912443807742192e-06)
('Moorabool', 1.0446580618457592e-06)
('Whitehorse', 1.0323107093823742e-06)
('Cardinia', 1.0261479940470042e-06)
('Glen Eira', 1.021484790586087e-06)

High Rent Price
('Manningham', 1.0734451240057539e-06)
('Boroondara', 9.476977372416822e-07)
('Monash', 8.567129148148774e-07)
('Bayside', 8.13306962970358e-07)
('Moonee Valley', 6.228299108532059e-07)
('Melton', 5.679321162577872e-07)
('Moreland', 5.395784708473804e-07)
('Stonnington', 4.022291176824319e-07)
('Port Phillip', 3.229847536845091e-07)
('Yarra', 2.464920711339441e-07)
```

Figure 7: Top 10 unsafe areas sorted by price range

```
Top 10 (Unsafe) Areas Ranked on Crime and Rent Price
Low Rent Price
('Latrobe', 6.352648554646513e-07)
('Mildura', 6.926677153889073e-07)
('Horsham', 7.223957453946699e-07)
('Greater Shepparton', 7.22676796465211e-07)
('Ballarat', 7.414318100692331e-07)
('Warrnambool', 8.43248955436913e-07)
('Ararat', 8.55795377495148e-07)
('East Gippsland', 8.70558275918602e-07)
('Wodonga', 9.344803649326403e-07)
('Wellington', 9.38776985380511e-07)

Mid Rent Price
('Maribyrnong', 4.7390373560946077e-07)
('Darebin', 4.860973781941574e-07)
('Greater Dandenong', 5.172554641677974e-07)
('Frankston', 5.78668858333759e-07)
('Hume', 6.102088413811213e-07)
('Brimbank', 6.251122138512715e-07)
('Whittlesea', 6.597147282941584e-07)
('Banyule', 6.686727754033669e-07)
('Hobsons Bay', 6.718932952061723e-07)
('Kingston', 7.150446015480841e-07)

High Rent Price
('Melbourne', 1.444778939423985e-07)
('Yarra', 2.464920711339441e-07)
('Port Phillip', 3.229847536845091e-07)
('Stonnington', 4.022291176824319e-07)
('Moreland', 5.395784708473804e-07)
('Melton', 5.679321162577872e-07)
('Moonee Valley', 6.228299108532059e-07)
('Bayside', 8.13306962970358e-07)
('Monash', 8.567129148148774e-07)
('Boroondara', 9.476977372416822e-07)
```

Figure 8: Top 10 unsafe areas sorted by price range

Clustering

Both k means and agglomerative clustering were performed on the complete dataset. First the number of clusters was determined using the elbow rule.

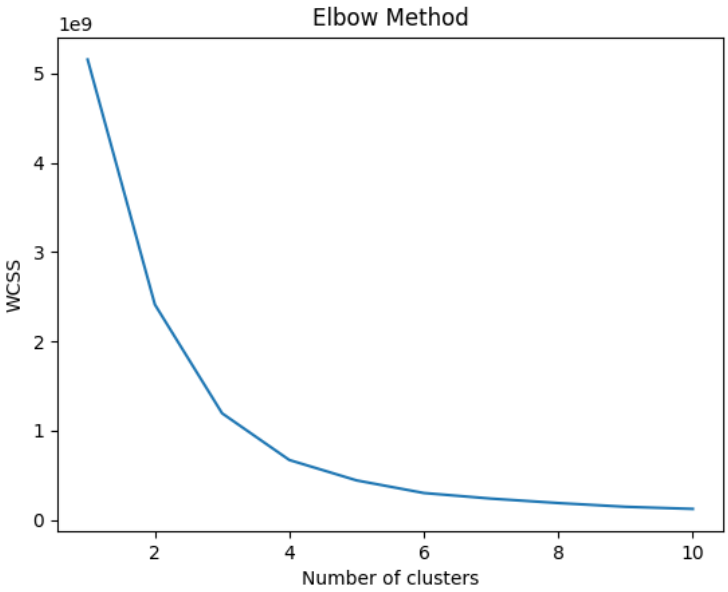
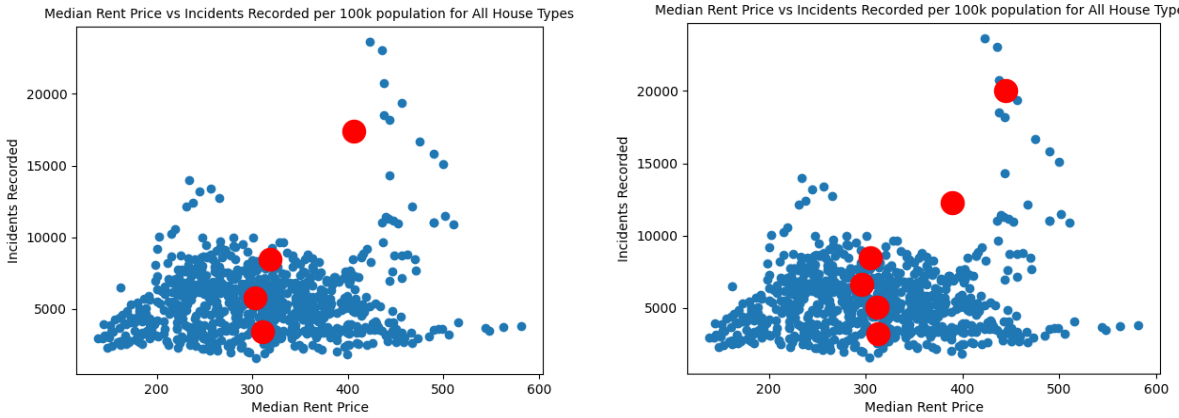


Figure 9: WCSS elbow method plot

From Figure 6, we found the optimal number of clusters to split the data would be around 4, and any more than that gives diminishing returns. We hypothesised that each cluster could represent larger areas of Victoria or group it into metropolitan and non-metropolitan areas. Clustering was also run for 6 clusters as there are 6 types of properties, to see if the clusters would represent the house types.

Figures 10 and 11: Agglomerative clustering for 4 and 6 clusters



Figures 12 and 13: k means cluster centroids for 4 and 6 clusters

As can be seen from the plots above, the data was not split into groups to begin with, the clustering simply split the data vertically. This was not very useful as we cannot deduce a lot from these clusters or if there are certain groups of suburbs that have certain characteristics. The red cluster in Figure 12 likely represents the inner city suburbs of Melbourne where rent prices and crime are both high, but other than this, the clustering did not produce valuable results as a result of the spread out and uncorrelated nature of the data.

Limitations and Future Improvements

The limitations we encountered are regarding our data. First of all, the incidents recorded are given as the number of incidents recorded, meaning that they include varying crimes ranging from simple theft to homicide. This limits us from producing more specific results targeting different types of incidents for each area or only selecting those that potentially have a correlation with rent prices.

However, with only a few datasets we still would not be able to form any conclusion about the correlation, if there are any, as there is a possibility that other factors are affecting the correlation. And as we hold no control over the variables, it is not possible for us to manipulate any of the data in an experimental manner, thus, limiting us from indicating any apparent factor.

Lastly, with all of the data originating only within Victoria, the results could only be applied mostly to Victorians as we are limited by the sample size and the variety of samples.

With all the limitations, there are improvements for future projects. A more indepth datasets could be used for a defined targeting, or more narrowing down of the research question could be beneficial. In addition, incorporating other datasets such as population density or income will possibly produce more insight on the correlation. And for the results to be more generalized, a larger sample size must be used, including datasets from outside of Victoria.