COMP 5070 Statistical Programming for Data Science

Take-home Exam

Introduction

https://www.rent.com.au/ is Australia's number one search site for renters and provides information on thousands of rental properties. The data that needs to be used for analysis has been scraped and subjected to data cleaning and other processing.

The report will analyse the processed data, study the distribution of price and car space number and the relationship between these two elements and bedroom and housing type and which places in Adelaide and surrounding areas are more popular will be studied, and the characteristics of houses in related areas will be explored.

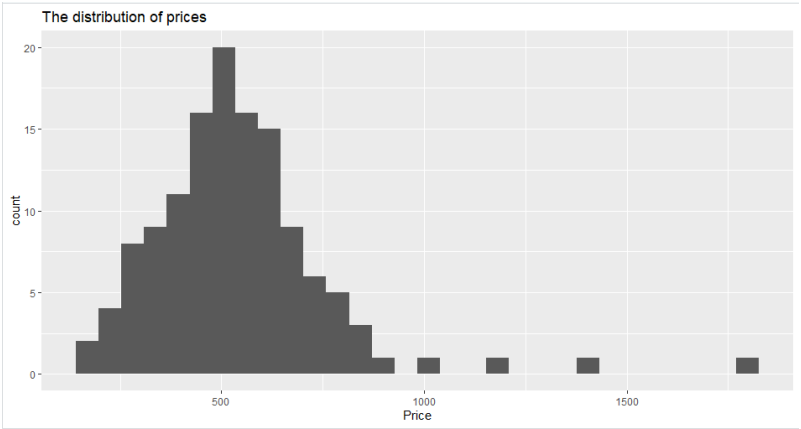
Rental Prices

For the price data in the data, a simple statistical summary result is as follows:

N	MEAN	SD	MEDIAN	MIN	MAX	SKEW	KURTOSIS	SE
129	538.75	217.76	520	170	1800	2.12	9.03	19.17

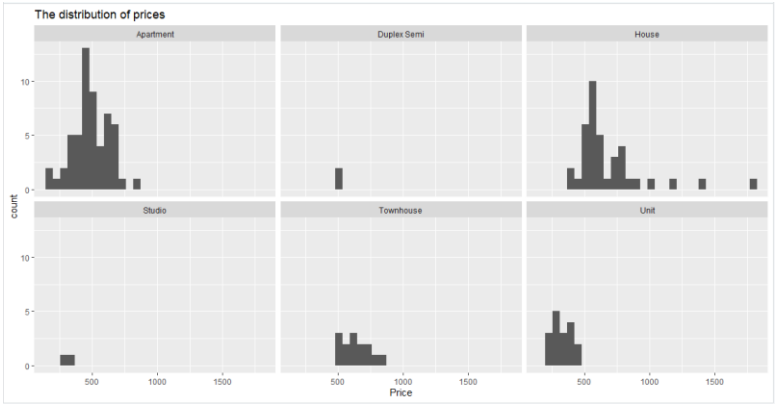
It is not difficult to see from the above table that the data contains 129 price data, the average value of these data is 538.75, the median is 520, the price distribution is between 170 and 1800, and the span is 1630. Since the value of skewness is greater than 1, it can be considered that the price data are highly skewed. Since the value of kurtosis is much greater than 1, it can be considered that the distribution is severely skewed.

The figure below shows the overall distribution of prices:

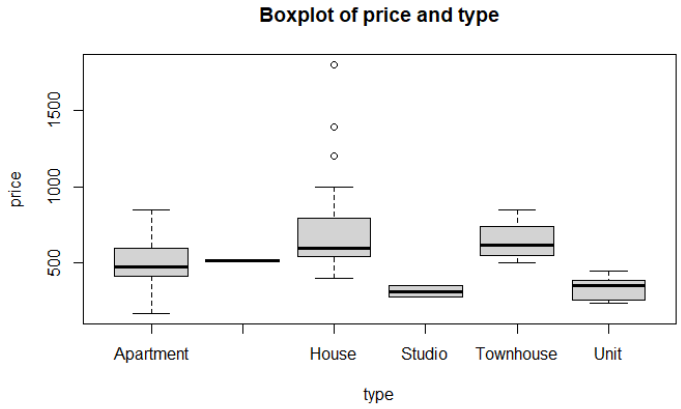


It is not difficult to find that there are some outliers in the data. If these outliers are removed, the overall distribution of price is close to the normal distribution.

The distribution of prices for different types of houses was further studied:



It is not difficult to see from the above figure that rental housing is rarely duplex semi and studio, but these types of housing are not the cause of many outliers in the overall price distribution, but houses are. The price distributions of the other four types of houses can be considered as having bimodal distributions.



If you use box plots to display the prices of different types of houses, the results are as shown above. House is indeed the cause of outliers in the overall price distribution. Without considering outliers, it can be considered that the price distribution of houses is statistically higher than that of other types of houses, and the price distribution of apartment has the highest degree of dispersion.

Since the "type" of rental properties is a category data, when studying the relationship between price and the data, the Anova test will be used first for detection. From the obtained results, since the value of p-value is less than 0.05, statistically speaking, the hypothesis that price and type are related cannot be rejected.

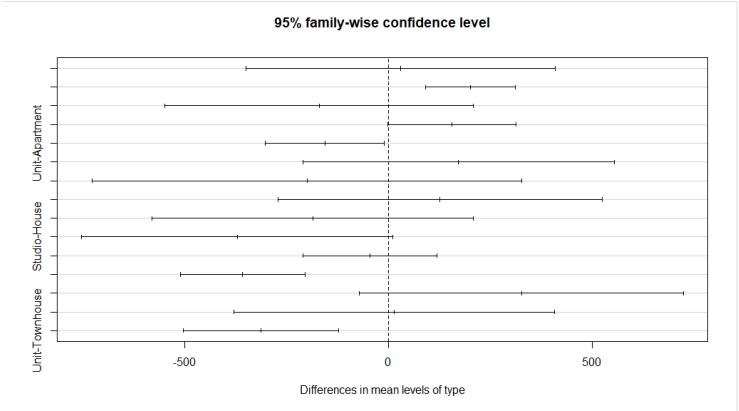
HSD test was used for further exploration, the result is shown as below:

$type

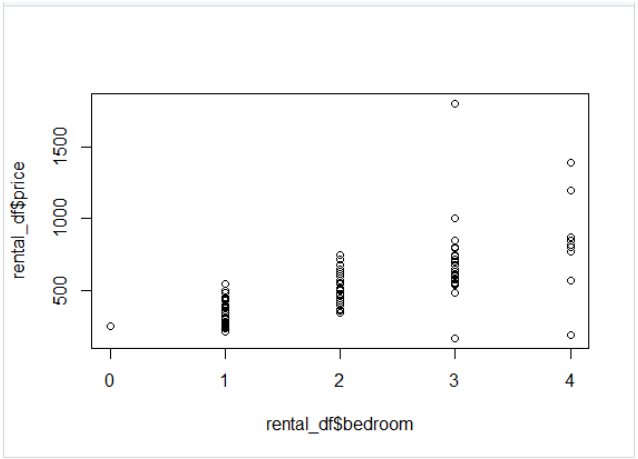
	diff	lwr	upr	p adj
Duplex Semi-Apartment	29.83929	-349.5032080	409.18178	0.9999142
House-Apartment	201.54981	90.7587758	312.34085	0.0000088
Studio-Apartment	-170.16071	-549.5032080	209.18178	0.7852664
Townhouse-Apartment	156.62500	-0.8883971	314.13840	0.0522213
Unit-Apartment	-156.33718	-302.3091536	-10.36522	0.0282694
House-Duplex Semi	171.71053	-210.7174972	554.13855	0.7845794
Studio-Duplex Semi	-200.00000	-727.1406520	327.14065	0.8811569
Townhouse-Duplex Semi	126.78571	-271.6951632	525.26659	0.9403982
Unit-Duplex Semi	-186.17647	-580.2378302	207.88489	0.7460578
Studio-House	-371.71053	-754.1385498	10.71750	0.0618853
Townhouse-House	-44.92481	-209.7305640	119.88094	0.9688961
Unit-House	-357.88700	-511.6994842	-204.07451	0.0000000
Townhouse-Studio	326.78571	-71.6951632	725.26659	0.1735170
Unit-Studio	13.82353	-380.2378302	407.88489	0.9999984
Unit-Townhouse	-312.96218	-503.2096495	-122.71472	0.0000763

The p-adj values of the four data circled in the above table are all less than 0.05, so statistically speaking, it is impossible to deny that there is a significant relationship between the corresponding two types of prices. Statistically speaking, the average price of house is 201.54981 higher than that of apartment, and there is 95% certainty that the price difference between house and apartment is distributed between 90.76 and 312.34. The other three can also be interpreted using the same logic.

The following figure is a visual display of the above content:



The following figure is a discrete diagram of the relationship between the number of the bedroom and the price of the rental property:



From the results of correlation exploration, it can be considered that the relationship between the number of bedroom and rental price is 0.6528674, which is a statistically significant strong positive correlation. Further exploration using linear regression analysis yields a model that looks like this:

$$Price = 206.95 + 156.21 * the\ number\ of\ bedroom$$

Statistically speaking, the price of rental properties will increase by 156.21 for every increase in the number of bedrooms.

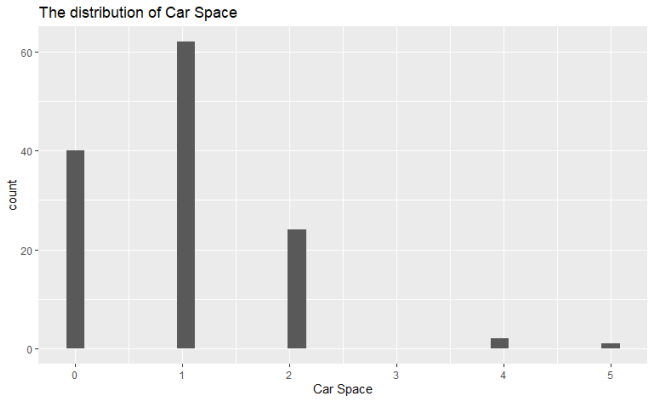
Car Space

For the number of car space, a simple statistical exploration is also performed first.

N	MEAN	MEDIAN	MIN	MAX	RANGE	SKEW	KURTOSIS
129	0.95	1	0	5	5	1.34	3.68

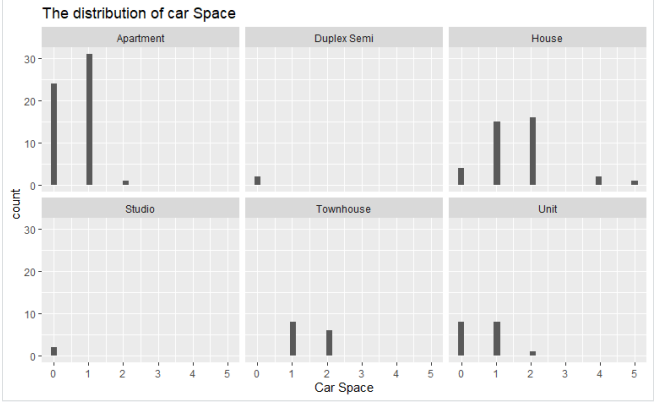
The number of observed data is also 129, the average value of the number of car space is 0.95, the median is 1, the minimum value is 0, the maximum value is 5, and the range is 5. Because skewness > 0, the value of the positive deviation is large, which is positive or right. The long tail drags on the right, and there are more extreme values at the right end of the data.

The overall breakdown of the number of car space is shown below:



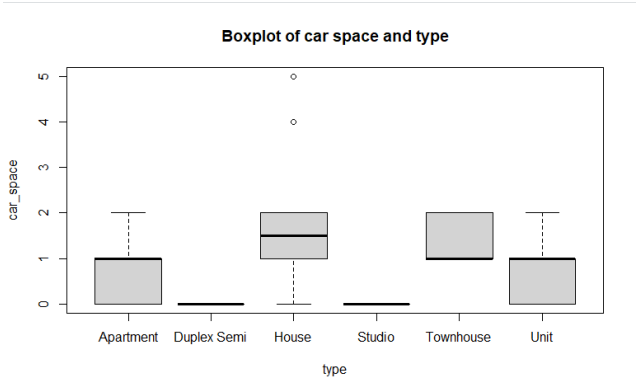
From the above figure, it can be found that there are indeed many outliers in the number of bedrooms. Overall, the distribution of the number of bedrooms presents a right-skewed distribution.

The distribution of the number of bedrooms for different types of rental properties is as follows:



It is speculated from the above figure that the outlier should be caused by the rental properties of the House type.

The following figure is a box plot, which provides the same support for this conclusion:



For the same reason, the relationship between the number of car space and type was tested by Anova test, and the result of p-adj value was 3.3e-09. Therefore, the hypothesis that the two are related cannot be rejected.

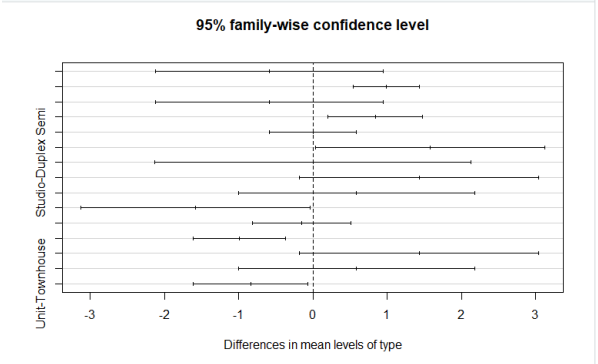
The following figure shows the results of the HSD test:

$type

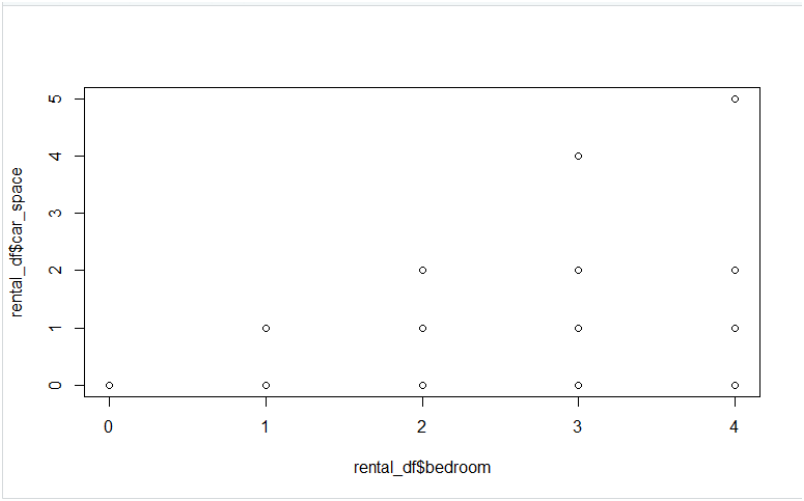
	diff	lwr	upr	p adj
Duplex Semi-Apartment	5.892857e-01	-2.11962132	0.94104989	0.8744172
House-Apartment	9.896617e-01	0.54271076	1.43661255	0.0000000
Studio-Apartment	-5.892857e-01	-2.11962132	0.94104989	0.8744172
Townhouse-Apartment	8.392857e-01	0.20384844	1.47472299	0.0027942
Unit-Apartment	-1.050420e-03	-0.58992750	0.58782666	1.0000000
House-Duplex Semi	1.578947e+00	0.03616418	3.12173056	0.0416144
Studio-Duplex Semi	2.220446e-15	-2.12657987	2.12657987	1.0000000
Townhouse-Duplex Semi	1.428571e+00	-0.17897185	3.03611471	0.1119066
Unit-Duplex Semi	5.882353e-01	-1.00147886	2.17794944	0.8918518
Studio-House	-1.578947e+00	-3.12173056	-0.03616418	0.0416144
Townhouse-House	-1.503759e-01	-0.81523188	0.51448000	0.9863812
Unit-House	-9.907121e-01	-1.61121921	-0.37020493	0.0001353
Townhouse-Studio	1.428571e+00	-0.17897185	3.03611471	0.1119066
Unit-Studio	5.882353e-01	-1.00147886	2.17794944	0.8918518
Unit-Townhouse	-8.403361e-01	-1.60782850	-0.07284377	0.0231159

The four circled can be considered to be significantly correlated, because the p-values of these four are all less than 0.05. Taking House-Apartment as an example, statistically speaking, the number of car space of House is 9.896617e-01 higher than that of Apartment on average, there is 95% certainty that the difference between the number of car space of House and Apartment is between 0.54 and 1.44.

The following figure is a visual display of the above content:



The figure below is a scatter plot between the number of bedroom and car space:



From the figure above, there is a positive correlation between the two. The more the number of bedrooms, the more the number of car space. Further research shows that the correlation coefficient of the two is 0.5673222, therefore, it can be considered that there is a strong positive correlation between the two.

Using linear regression analysis, a model can be obtained that looks like this:

$$The\ number\ of\ car\ space = -0.20342 + 0.54468 * the\ number\ of\ bedroom$$

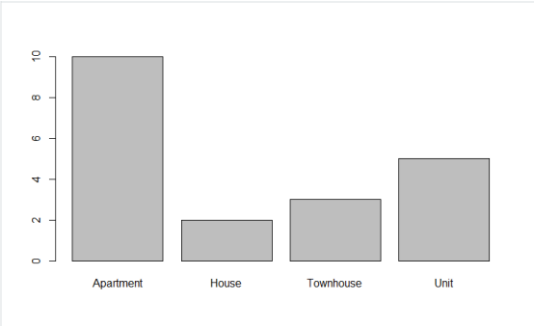
Statistically speaking, each additional bedroom will know that the number of car space will increase by an average of 0.54468.

Locations

It can simply be assumed that the street with the most homes for rent on the site is the most popular street. From this, it can be obtained that the most popular streets in the Adelaide SA 5000 part are Wright Street, Waymouth Street, Hutt Street, Hindley Street and Gray Street.

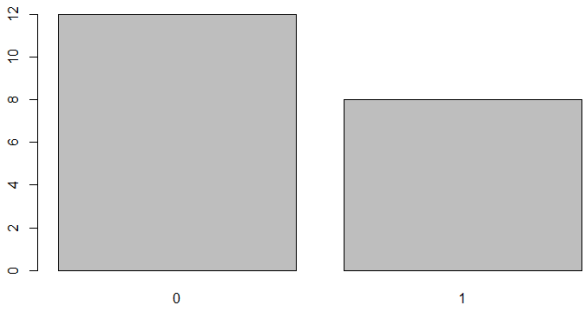
These five are equivalent in terms of ranking, so the housing information in these five streets will be analyzed to explore the characteristics of the houses in the most popular areas.

The following figure shows the distribution analysis results of the types of rental properties:



Judging from the results, the type of most houses is apartment, followed by unit and

townhouse, and the least number is house.

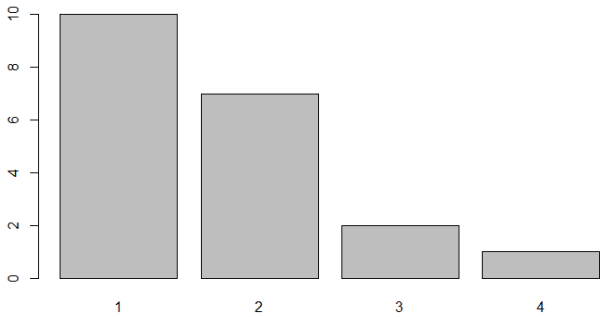


Most of these rental properties do not have a car space, and a small part has a car space.

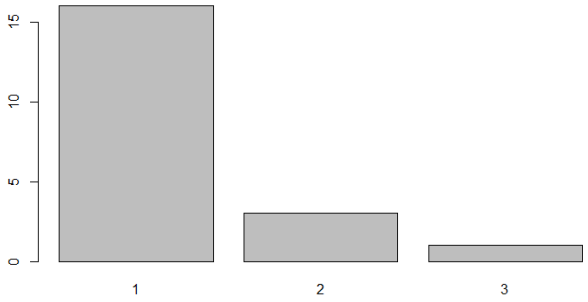
An analysis of the prices of these rental properties is as follows:

From the point of view of distribution, there is a bimodal distribution, the price is the most between 400-500, followed by 200-300, and the number of prices between 100-200 is the least.

The number of bedrooms is the highest at 1. With the increase of the number of bedrooms, the frequency of house occurrences decreases continuously:

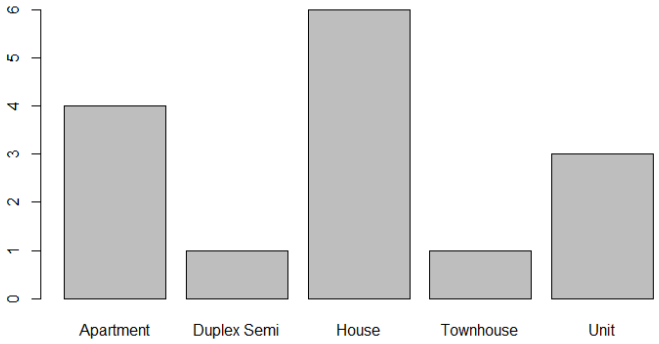


The number of bathrooms is the same:



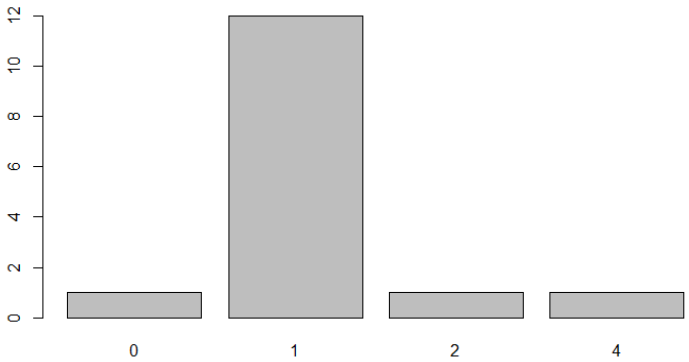
Most of the rental properties only provide 1 bathroom. It is not difficult to see from the figure that the number of rental properties with 2 bathrooms has dropped sharply, while only 1 rental property has 3 bathrooms.

Non-Adelaide area, the most popular is North Adelaide SA 5006.



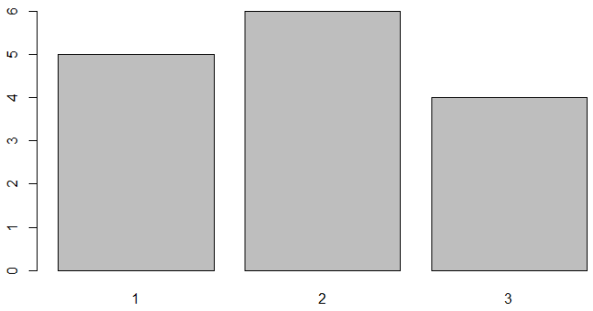
The distribution of housing types in this area is shown in the figure above. The largest number is house, followed by apartment and unit, while Duplex Semi and Townhouse are the least.

The figure below shows the distribution of the number of car space:

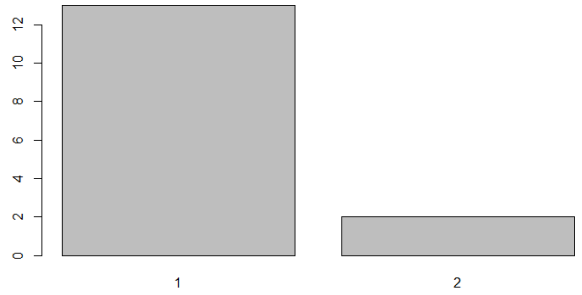


It can be found that only 1 car space is provided in the case of the largest number.

The frequency plot for the number of bedrooms is as follows:



It can be found that there is not much difference in the frequency of providing 2 bedrooms, 3 bedrooms and 1 bedroom.



As for the bathroom, most of them only provide 1 bathroom, and a few provide 2 bathrooms.

The price distribution in this area presents a bimodal distribution, with the largest number between 500 and 600, the least number between 300 and 400, and only one price between 600 and 700 and 700 and 800.

Conclusion

There are outliers in the prices of rental properties, which approximate a normal distribution without taking these outliers into account. There is a significant relationship between the price of rental properties and the type of rental properties, and the price of House is statistically higher than other types. This may be because the House has more bedrooms, bathrooms, and car space numbers on average, but whether this is really the case needs further research.

There is also a positive correlation between the number of car spaces provided by rental properties and the type of rental properties. The upper limit of the number of car spaces that House can provide is the highest, but the lower limit is basically the same as other types. This may be due to the fact that House presents more possibilities in design and planning. But from an overall and statistical point of view, if the number of bedrooms provided by the rental property is more, then the car space number will be more.

By comparing the characteristics of rental properties in the most popular street in the Adelaide region and the non-Adelaide region, it can be found that the price of the rental property in the most popular street in the non-Adelaide region is 500 ~600, and 400~500 in the Adelaide area, which may be because the corresponding rental properties provide more bedrooms, bathrooms, or car spaces. Bedrooms is most likely the reason, because the number of rental properties in the non-adelaide area offering 2 and 3 bedrooms is closer to the number of only 1, but the most rental properties in the Adelaide area still provide 1 bedroom, Provides an extremely rapid reduction in the number of two.

Some more data may be collected and aggregated for analysis, such as the distribution of people with different occupations in various regions, which can be used to analyze the living preferences of people with different wages.