UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

ASDA

Project 2

Spatial Point Pattern and Spatial Lattice Data Analysis

Graham Davies
DVSGRA012

Department of Statistical Sciences
University of Cape Town
South Africa
November 4, 2021

Plagiarism Declaration Form

A copy of this form, completed and signed, to be attached to all coursework submissions to the Statistical Sciences Department.

COURSE CODE:	**STA4010W**
COURSE NAME:	**ASDA**
STUDENT NAME:	**Graham Davies**
STUDENT NUMBER:	**DVSGRA012**

PLAGIARISM DECLARATION

- I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
- I have used a generally accepted citation and referencing style. Each contribution to, and quotation in, this tutorial/report/project from the work(s) of other people has been attributed, and has been cited and referenced.
- This tutorial/report/project is my own work.
- I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.
- I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.
- Agreement to this statement does not exonerate me from the University's plagiarism rules.

Signature:  Date: November 4, 2021

Contents

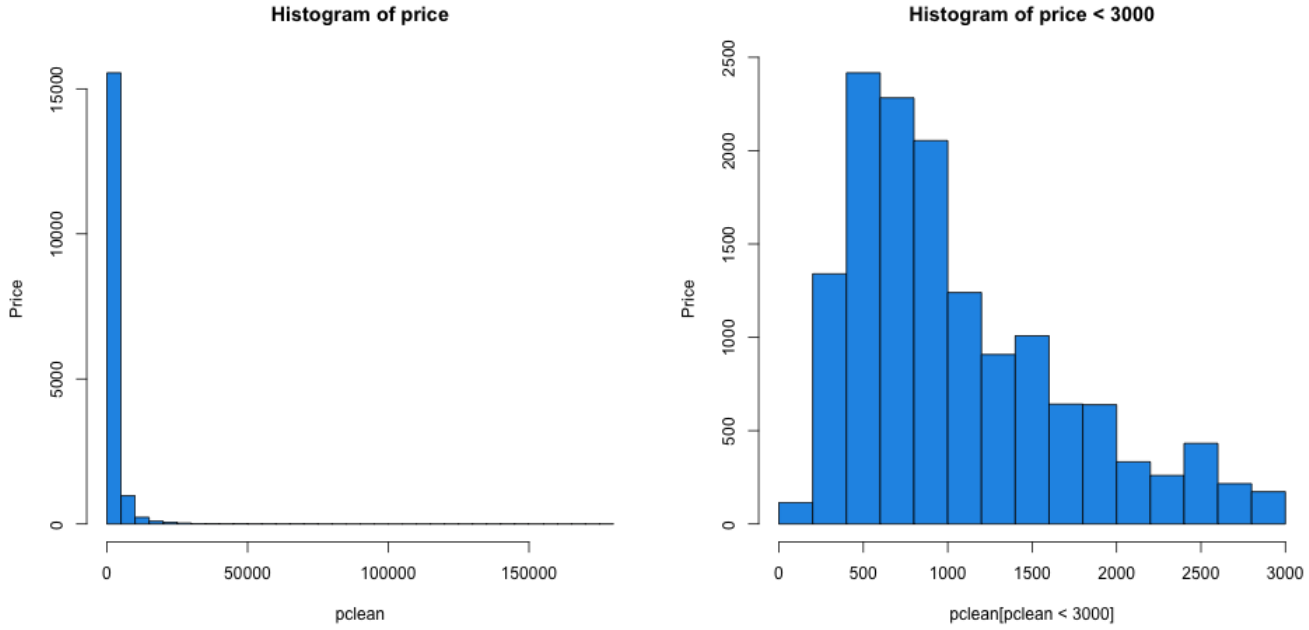
Introduction

This project serves to spatially explore the AirBnB data set through the use of spatial point pattern analysis (SPPA) and spatial lattice data analysis (SPDA). I will begin with exploratory data analysis to better understand the data. Following this, I will move onto spatial point pattern analysis and finally the spatial lattice data analysis.

Part 1: Exploratory Data Analysis

detect any discrepancies/anomalies in your dataset using any or all of the following: tables, summaries, frequencies, plots. There are missing values, outliers in your dataset.

Due to the nature of the housing market there is a broad range of values and variables for houses. I start with looking at the price variable

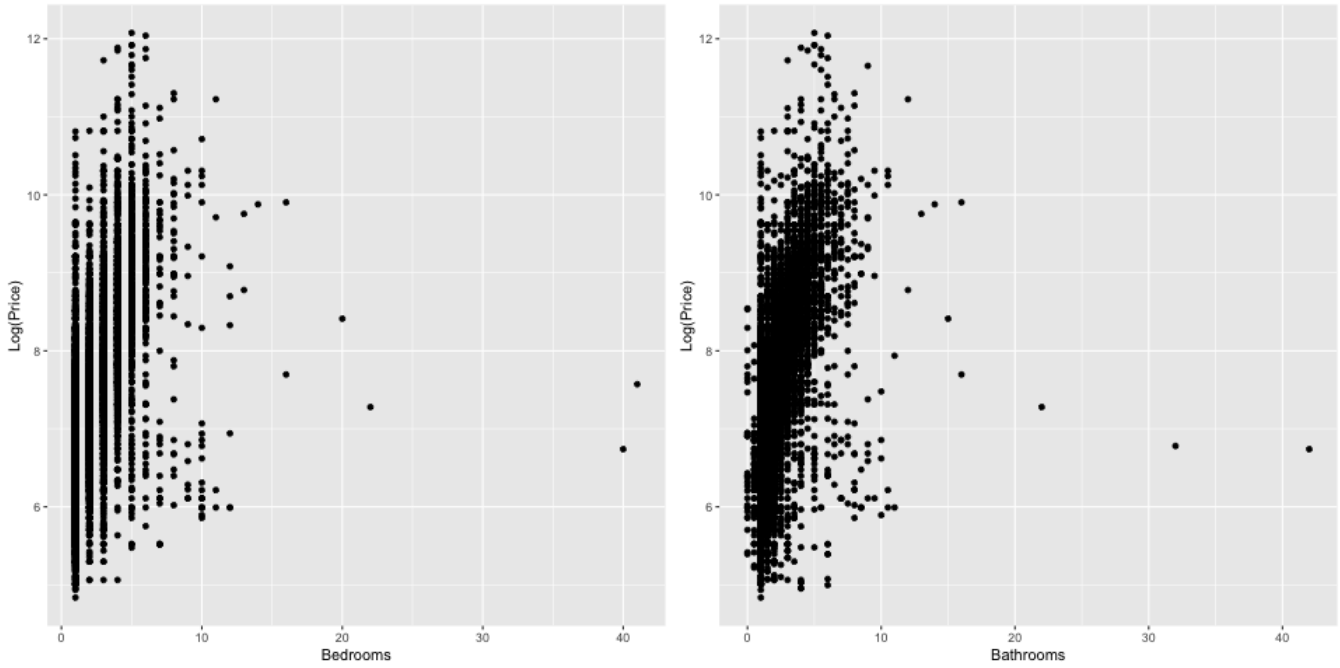


(a) Histogram of Price variable

(b) Histogram of Price variable

Figure 1: Univariate Analysis of Price

Inititally I plot a histogram of the price variable. We can see an extreme tail towards to the right which suggests there are some very expensive listings. As I will be investigating the ward data later on, and I believe price may be a differentiating factor in the wards, I keep all these variables and don't remove outliers. However to understand the majority of the price in the listings, I plotted the price that is less than $3000. Here we can see most listings are priced between $500 and $1000. Few listings are priced less than this, and there is a gradual decline in the number of listings priced more than $1000. I next look at the relationship between price and other variables



(a) Log(Price) vs. number of bedrooms

(b) Log(Price) vs. number of bathrooms

Figure 2: Bivariate analysis of Price

From these log(price) plots we note that log(price) and bedrooms/bathrooms each have outliers where there is an extreme number of bathrooms and bedrooms. Again I believe these may impact the wards so I do not remove them. I next look at a variable which calculates the price per person the listing accommodates. The formula for this is: $\frac{\text{Accommodates}}{Price}$.

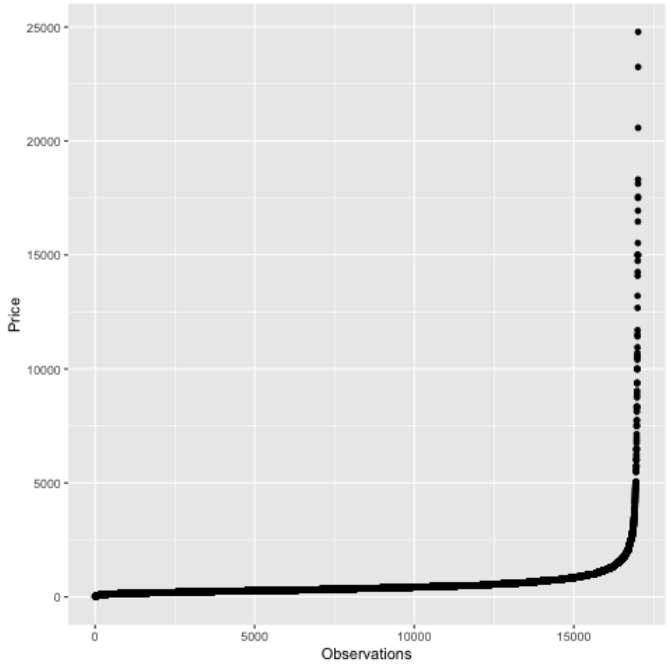
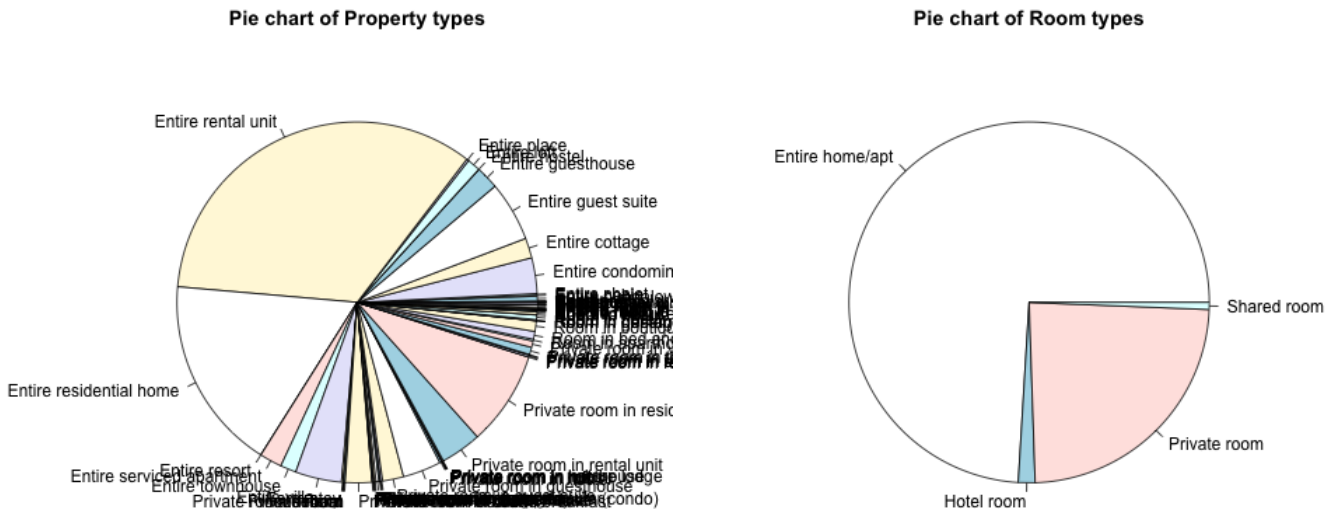


Figure 3: Price per person plot (sorted)

We note a similar result as seen earlier where the majority of listings are below $2000. There is then a very steap

increase in price per person where the maximum price per person lies around $25000.

Next I look at the categorical variables "room_type" and "property_type". I plotted pie charts to see how many categories and the proportion of each category.

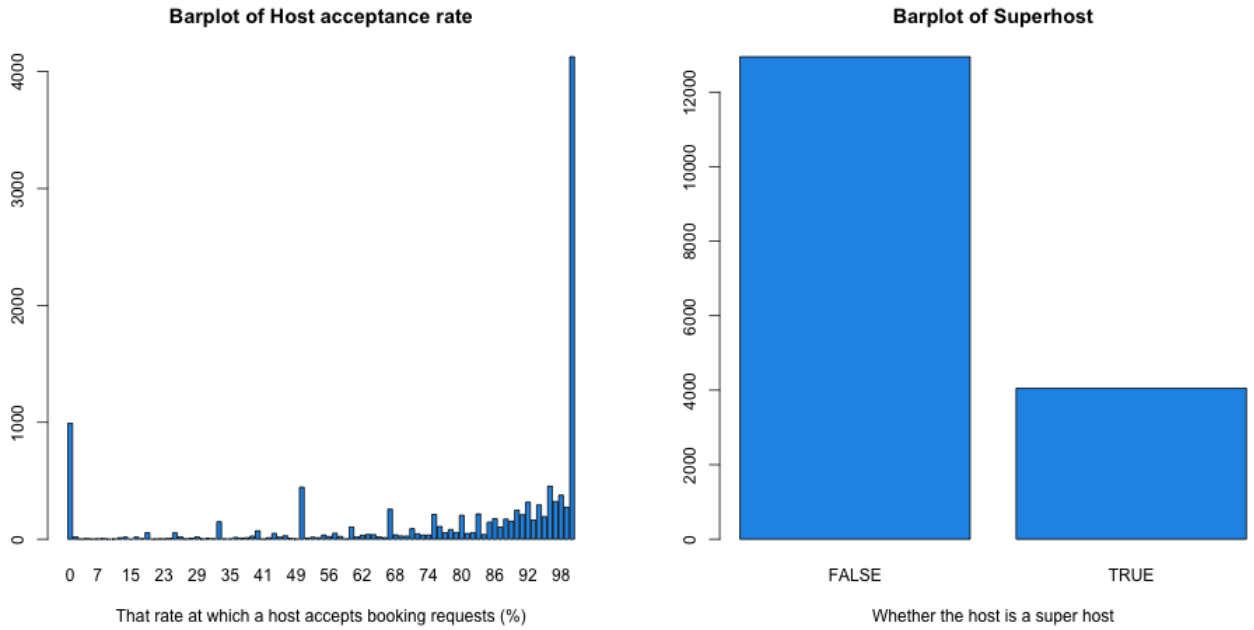


(a) Property Type

(b) Room Type

Figure 4: Pie charts comparing Room and Property Type

From these figures, we can see there are plenty of categories in the property_type variable. The "Entire rental unit" and "Entire residential home" variables are the property types that make up the majority of the observations. The third most frequent property type is a "Private room in residence". Looking at the "Room Type" pie chart we see very similar results with the "Entire home" variable making up the majority and "Private room" taking up less than a quarter of the proportion. Next I look at hosting variables:



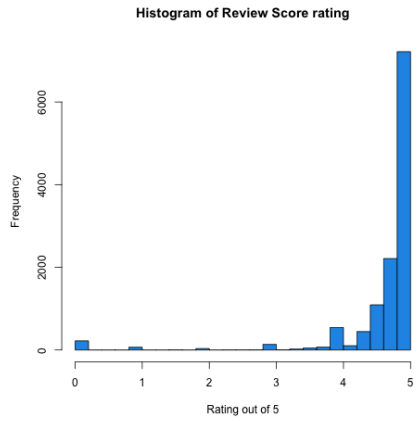
(a) Acceptance Rate

(b) SuperHost comparison

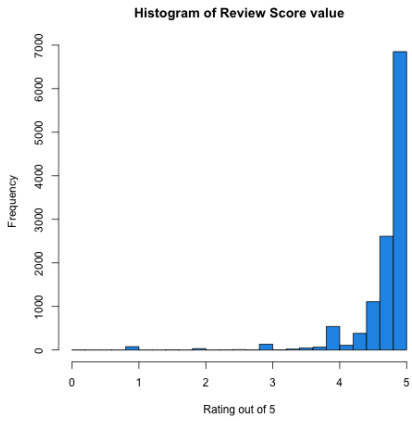
Figure 5: Bar plots comparing Acceptance Rate and Response Rate

From figure 5a we see the barplot of host acceptance rate. Many hosts accept 100% of the clients, some accept 0% but most are above 80%. If we look at figure 5c, we note that there are much fewer superhosts compared to non-superhosts. There are 16 missing values, that are included in the "False" graph but they are ignored as they have an insignificant influence.

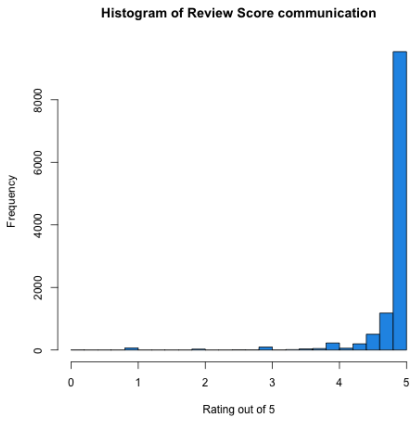
There are many review variables which relate to how the client reviews different aspects of the listings. I plotted some of these variables as well the log of reviews per month.



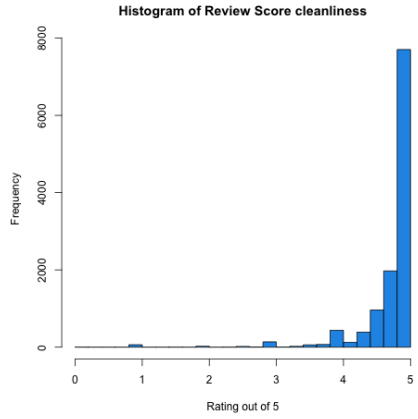
(a) Rating



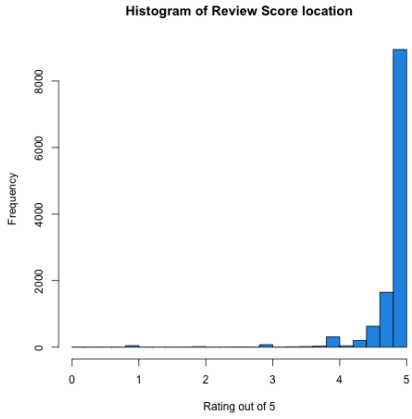
(b) Value



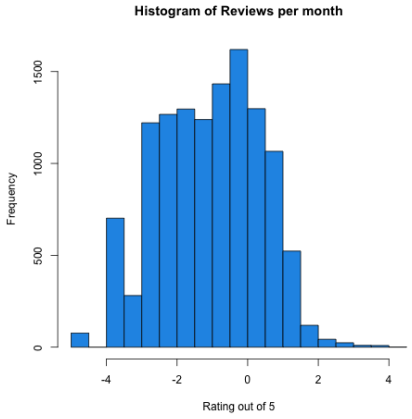
(c) Overall



(d) Cleanliness



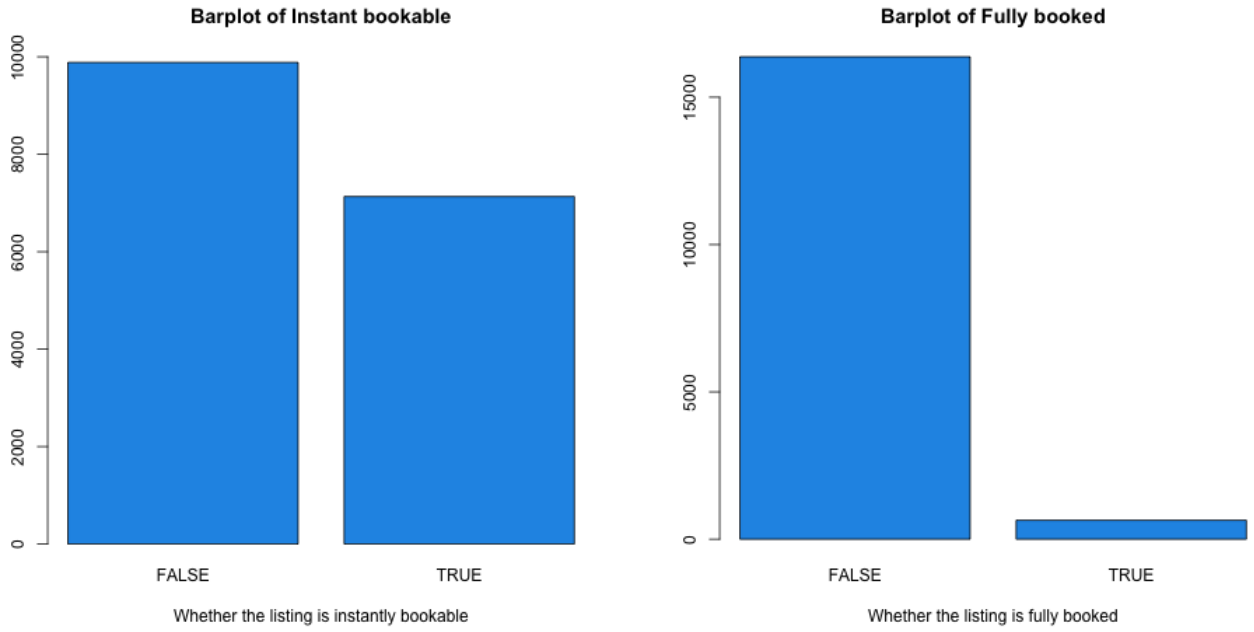
(e) Location



(f) Log(reviews per month)

Figure 6: Plots of review variables

From these plots we can see very similar results which is understandable. Intuitively, a high rating in one aspect is likely to have a high rating in other aspects. We note that most listings contain a reviews greater than a "4" for all the different review aspects. I looked at the log(reviews per month) to gauge how often listings are reviews. These values ranged from 0 to 73, NA values are removed. We can note a tail on the right which suggests that few listings have plenty of reviews.



(a) Instantly Book

(b) Fully Booked

Figure 7: Availability plots

From Figure 7a we note that more places are not instantly bookable. In Figure 6b we note that very few places are fully booked.

I next plotted the wards to gain an understanding of the longitude and latitude variables. Knowing the range of wealth across Cape Town, I thought these variables may be of value later on.

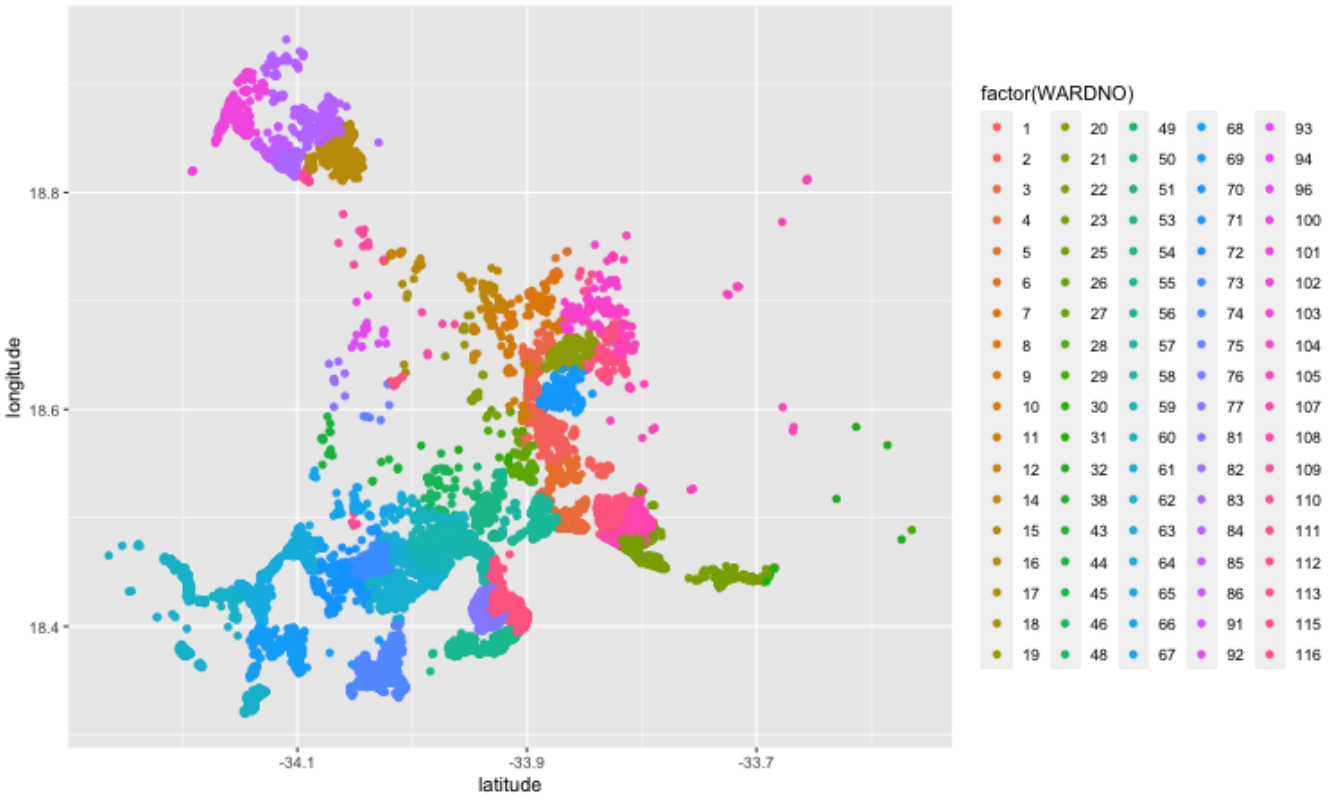


Figure 8: Map of Wards, relative to longitude and latitude

From the above plot we can the understand Cape Town map much better. The wards are colour coded, and even though it is difficult to see the exact colour, we can infer that the nearer the Ward label(e.g. ward 84 and ward 85), the nearer the ward location. Generally the blue wards are closer together to each other compared to the rest of the colours. Using domain knowledge of the areas around Cape Town, we can infer that there are listing in very wealthy areas, as well as in the poorer areas.

Part 2: Spatial Point Pattern Analysis

In Spatial Point Pattern Analysis we try to understand the shape and distribution of the points. Through the use of graphs and the Clark-Evans test I will scientifically classify distribution of points in two wards that were given to me.

Allocated Ward Numbers		
User ID	Ward Number 1	Ward Number 2
dvsgra012	62	84

EDA and data preparation

I will start with understanding ward 62 and ward 84 through some maps.

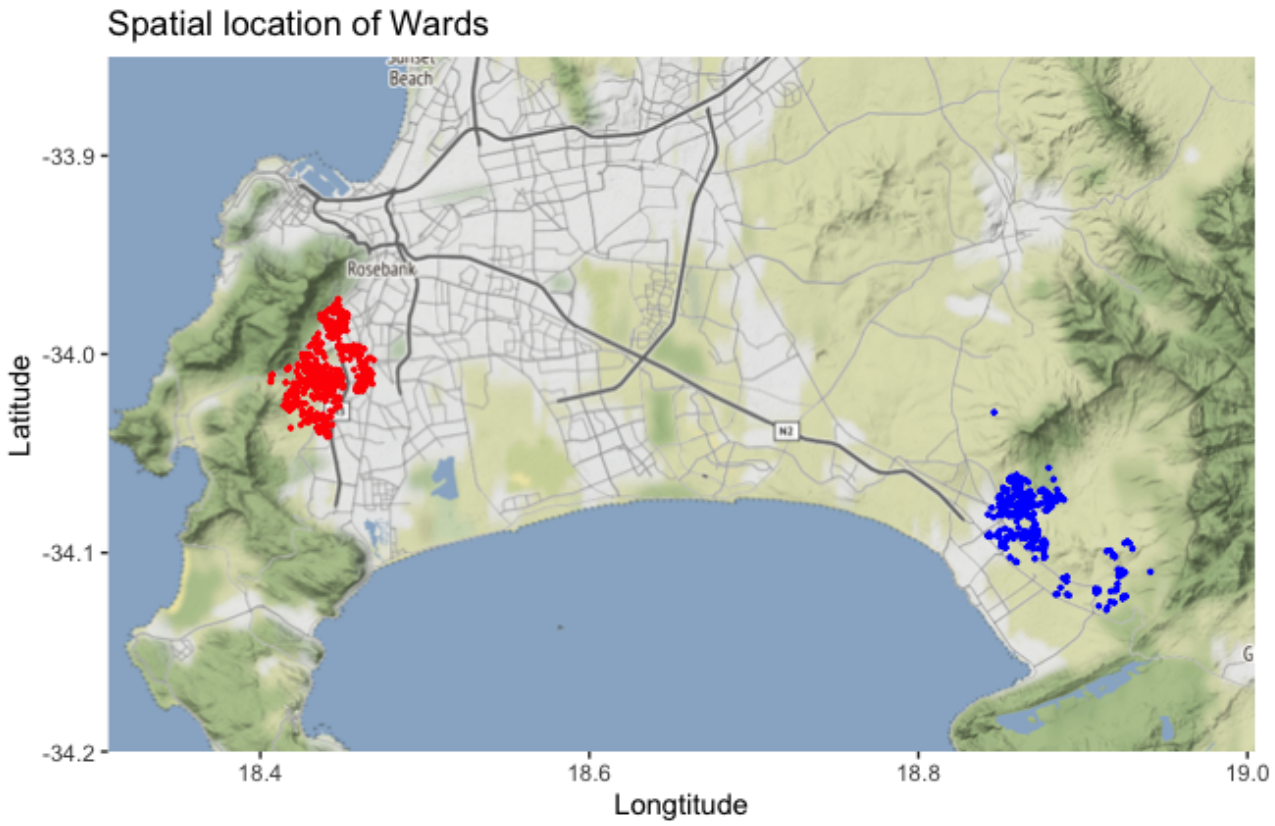


Figure 9: Ward listings mapped on Google

The above map plots the listings for both wards on a google map. The ward 62 listings are plotted in red and the ward 84 listings are plotted in blue. This is plotted to understand where the wards are located relative to each other. We note that ward 62 seems to be in a much more urban environment, although it includes parts of Table Mountain National Park. Ward 84 is near Somerset West which is known to have larger properties and many farms/plots in the areas.

I now look at each ward individually.

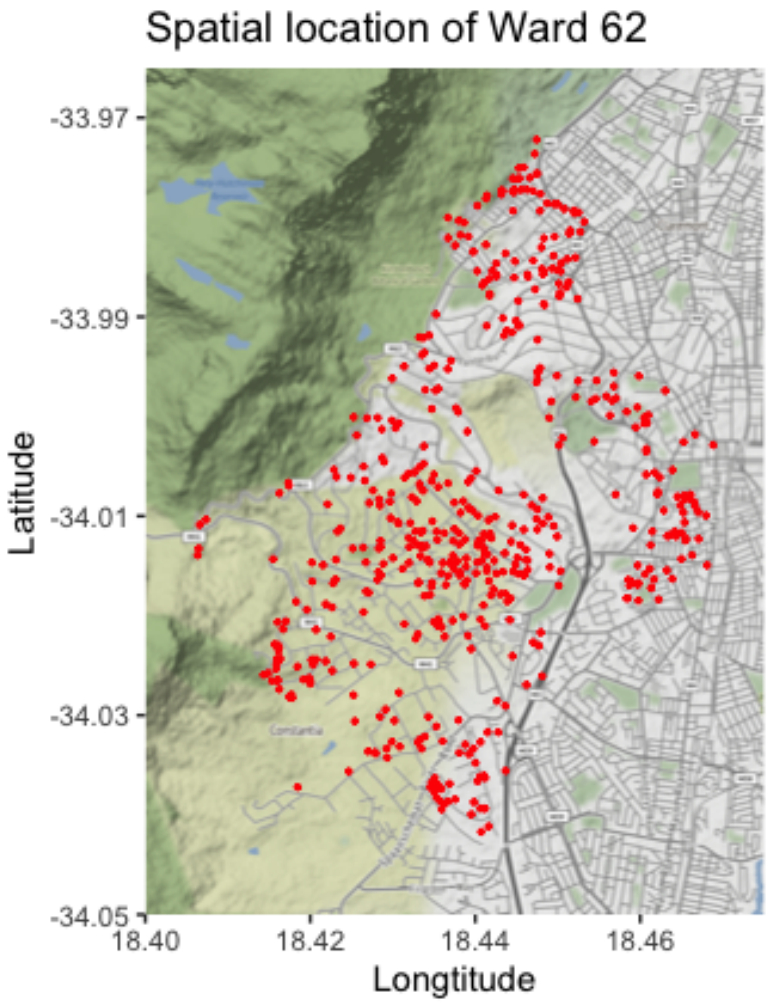


Figure 10: Ward 62 Listings

Ward 62 is located in a mixture of areas with 514 listings. We note the a high density of streets on the East side of the map and fewer streets on the North West side of the map. Ward 62 also appears to have a park in between in where there are no listings.

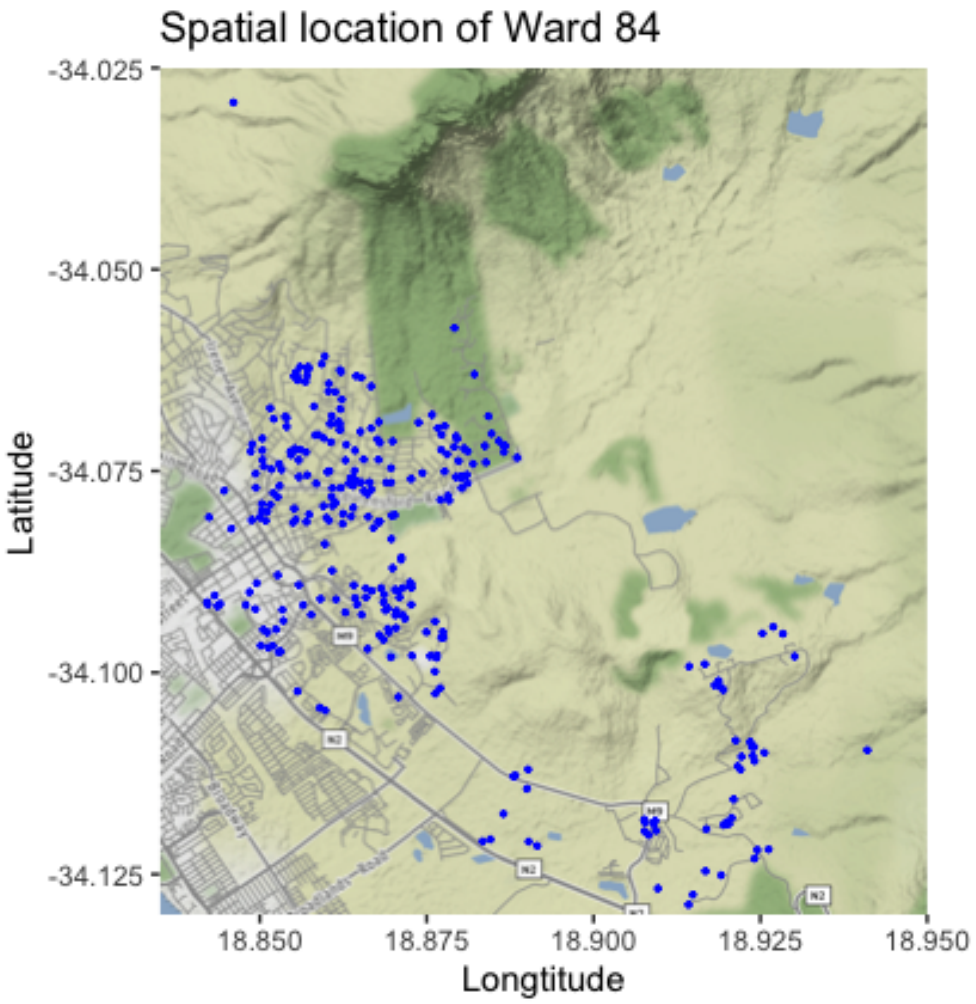


Figure 11: Ward 84 Listings

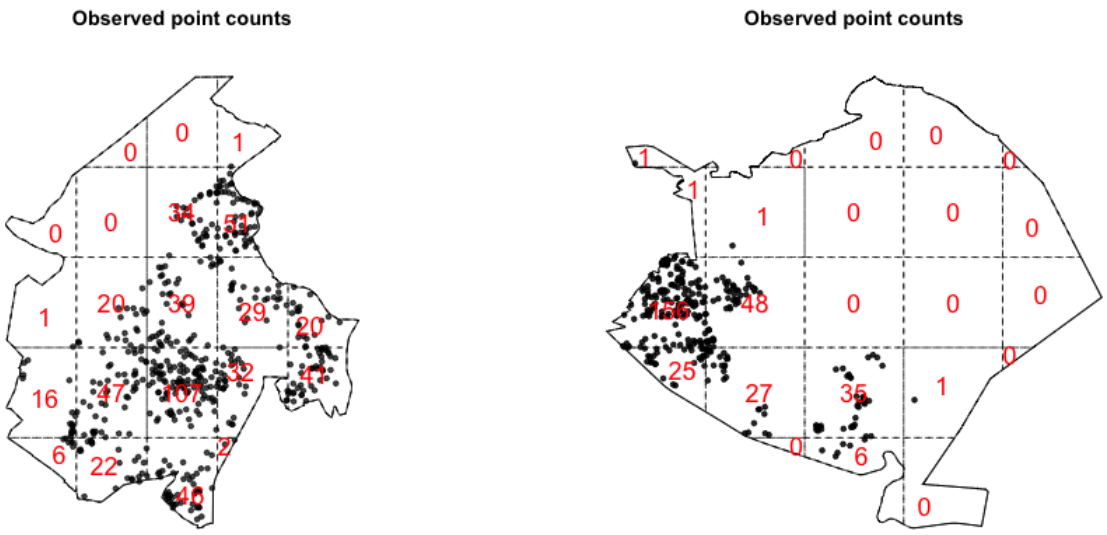
Ward 84 is also located with both suburban areas and rural areas and contains 301 listings. We note the streets on the West side contain many listings yet the open space on the East side of the map contains fewer listings. We also note a single Northern listing.

From an initial observation, I would classify the point pattern data structure as clustered.

Quadrat Test

I start my point pattern analysis by performing quadrat tests and investigating intensity.

The grid size for the quadrat tests are very important an appropriate size needs to be chosen to balance the trade-off of bias and variances. Larger quadrats reduces the relative error but removes spatial variation in intensity which in each quadrat. I experimented with a 4x4 grid, 5x5 grid and 6x6 grid and believe a 5x5 grid works best for both plots. The 4x4 plot has the right balance between few zero quadrats, and not containing most points in the same quadrat.



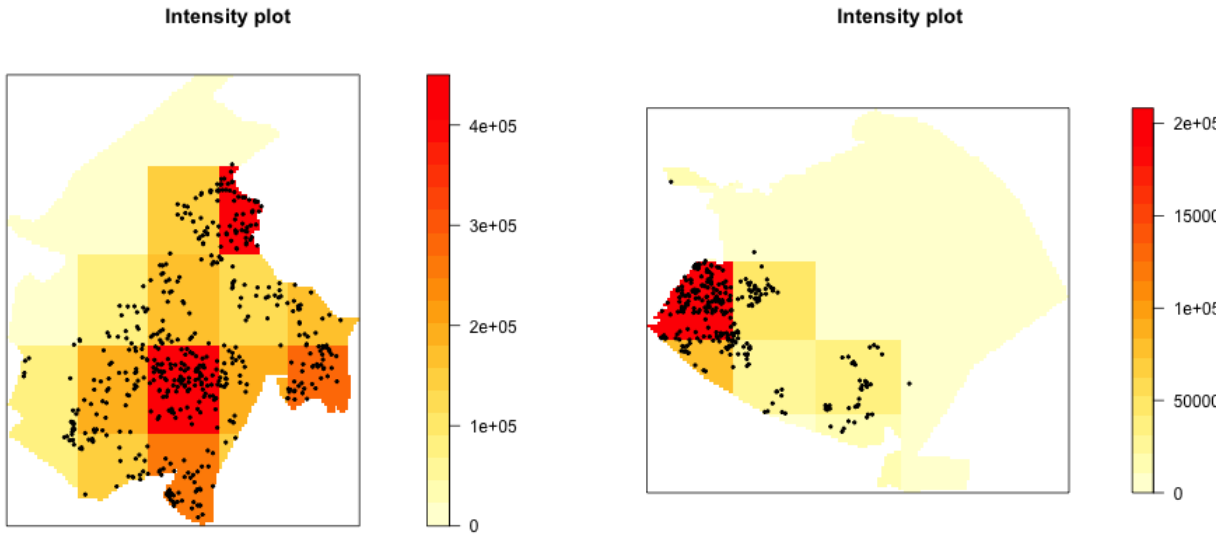
(a) Ward 62

(b) Ward 84

Figure 12: 5x5 grid counts

We can now mathematically see where the listings are found in each ward. Both wards have areas with very high concentration and parts with very low concentration of listings.

Intensity is the expected number of points per unit area and it is the first moment of a point pattern distribution. Using the same grid as before, I plot the intensity per quadrat or quadrat intensity.



(a) Ward 62

(b) Ward 84

Figure 13: 5x5 grid intensity

In Figure 13a we note two quadrats with very high intensity (dark coloured quadrats) and low intensity (light coloured quadrats) to the North West. If these points were structure or completely spatially random, we would not see much change in the colour of the quadrats. In figure 84 we see one quadrat with very high intensity, some less intense areas around that and then some very baron quadrats as we move East. From this quadrat intensity analysis, it would appear that the listings points are patterned in a cluster.

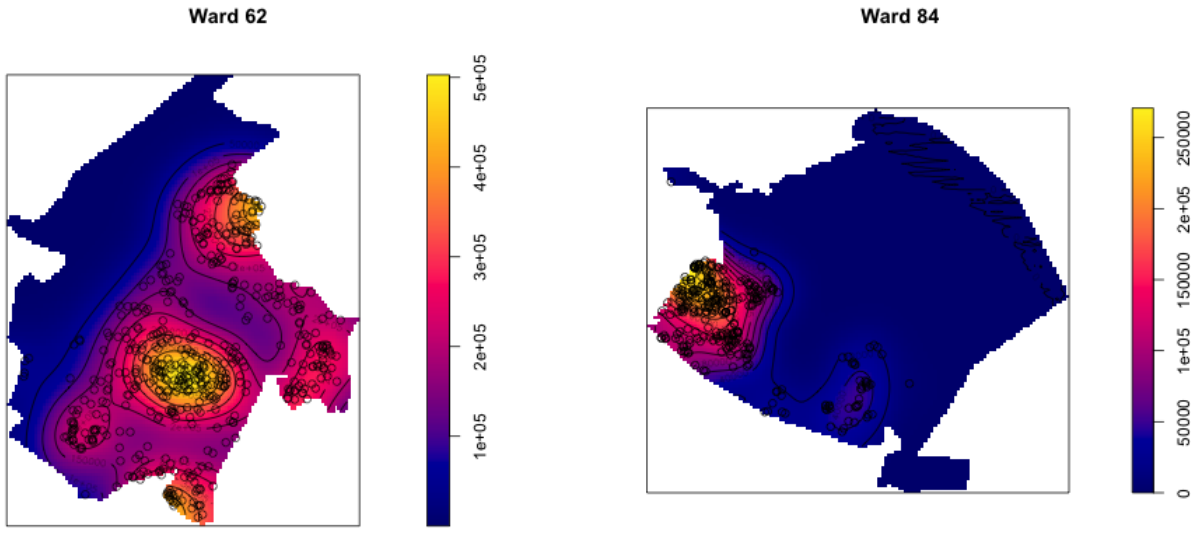
Kernel Density Smoothing

I next move onto kernel density smoothing plots. Kernel density smoothing uses the following estimator to create an estimate:

$$\hat{\lambda}(x) = \frac{1}{h^2} \sum_{i=1}^{n} \kappa \left(\frac{\|x - x_i\|}{h} \right) / q(\|x\|)$$

Where κ is a bivariate and symmetrical kernel function and $x_1, x_2, ..., x_n$ are the data points.

The bandwidth, h, measures the level of smoothing. I experimented with a few values and found that $h = 0.005$ was best for Ward 62 an $h = 0.008$ was best for Ward 84.



(a) Ward 62

(b) Ward 84

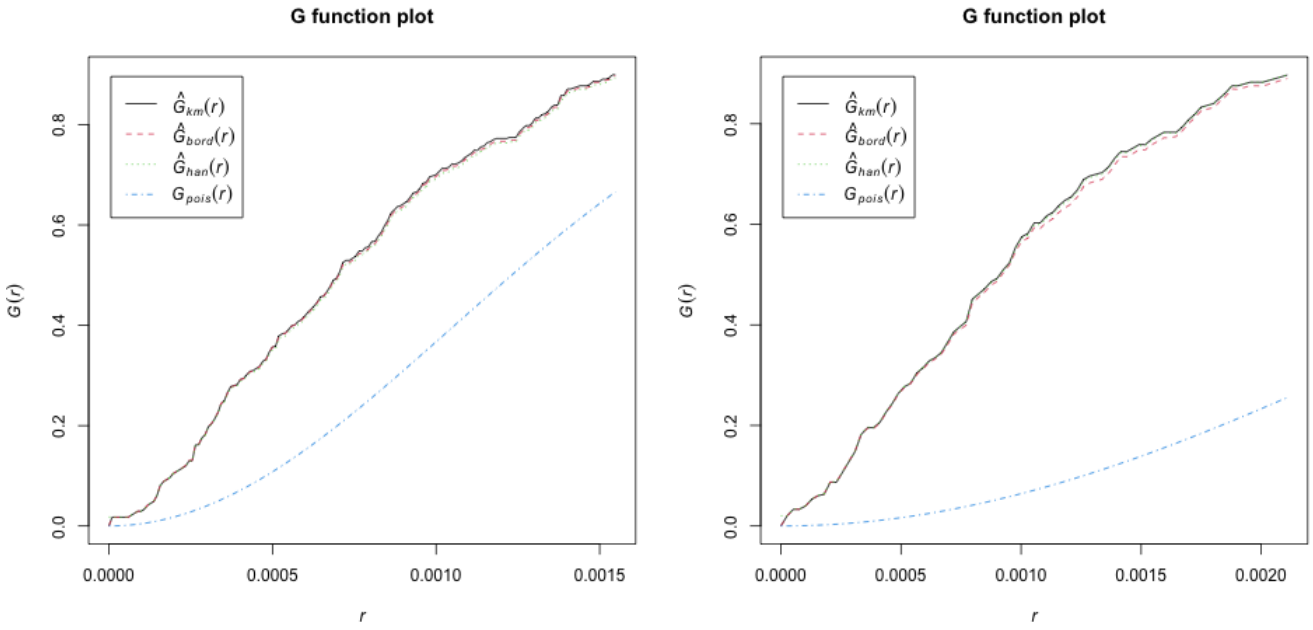
Figure 14: Kernel Density Smoothing Plots

In Ward 62 we note very few observations in the North West section of the map. This makes sense as the ward contained parts of Table Mountain National Park. We note two peaks in density with a trough in between them where the park was noted. Due to the range in the kernel density, Ward 62 appears to be clustered.

In Ward 84 we note a single peak which was expected with baron land on the East side of the Ward. Again I would say the listings are clustered based on kernel density.

G, F and Ripley's K Function

Next I look at the G-Function and F Function plots which are cumulative frequency distribution of nearest neighbourhood distances. First the G function compares observed and expected distributions of the event-to-event (e2e) distance. These are plotted below:

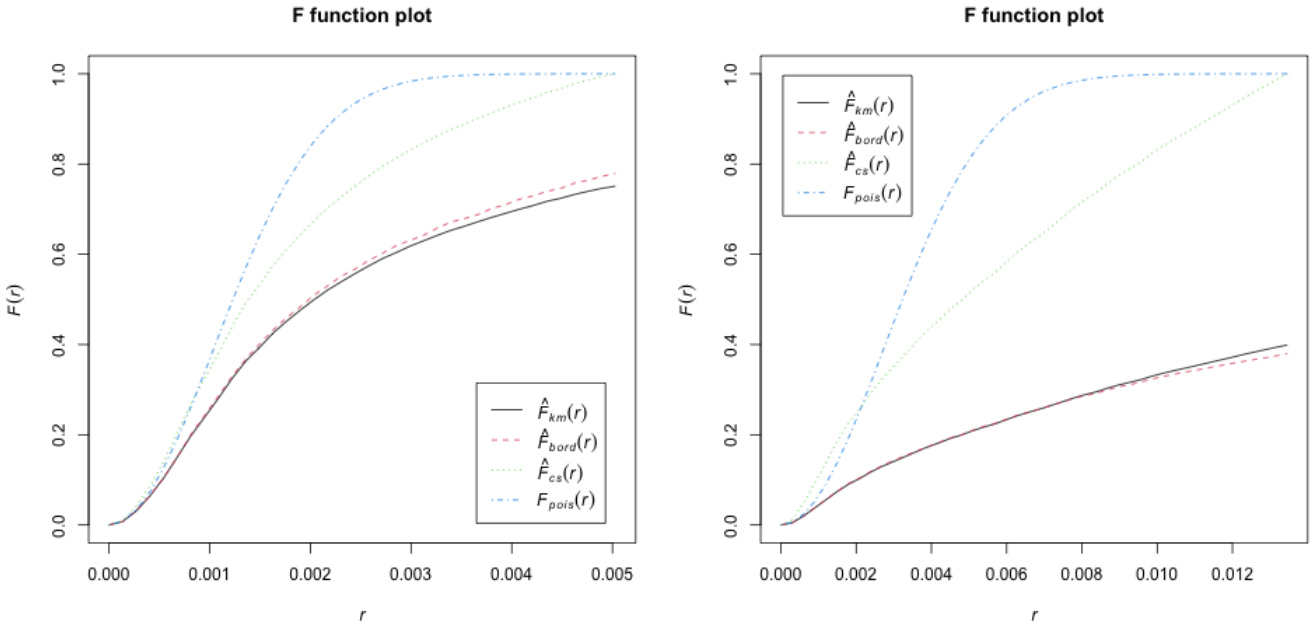


(a) Ward 62

(b) Ward 84

Figure 15: G Function

The expected distribution for the event to event distances is defined by the blue line "$G_{pois}(r)$. From this we note the observed lines are well above the expected line which indicated that the points are clustered. This result is inline with what is expected from earlier analysis. Ward 84 seems to show stronger evidence of clustered where the observed and expected plots are very far from each other.



(a) Ward 62

(b) Ward 84

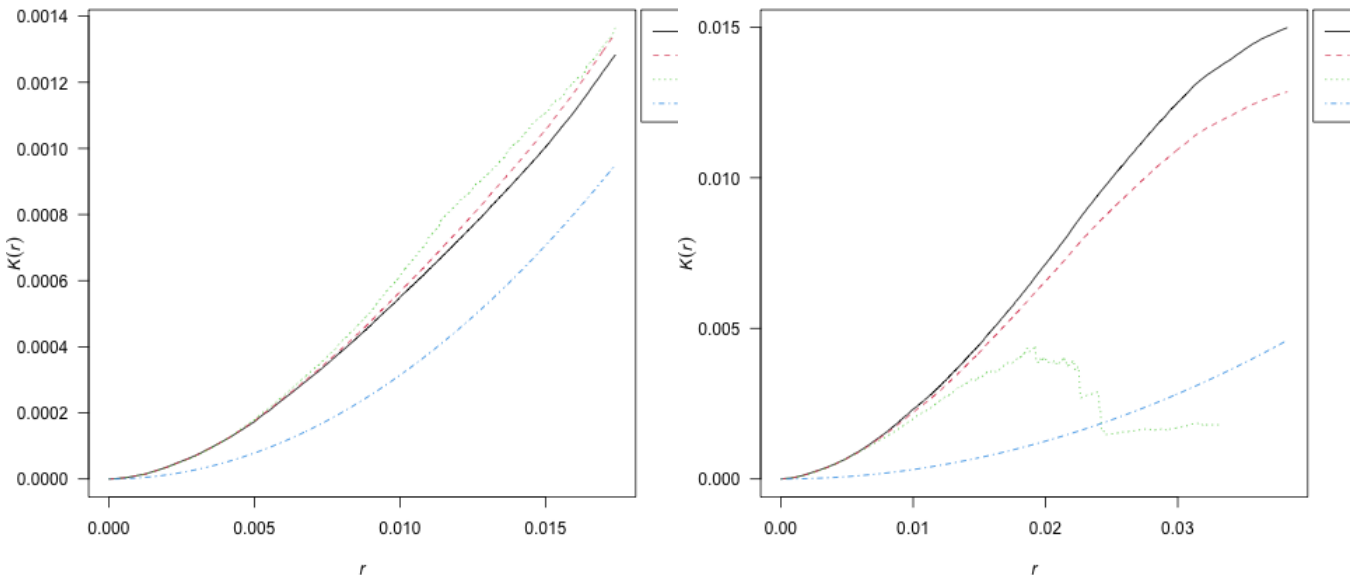
Figure 16: F Function

The F function used point to event distance instead of event to event distance like in the G function. For the points to be of a clustered structure the points would need to be below the expected line. Once again, we note that in both wards, the listings are in the shape of a clustered structure.

Next I observe second order properties through the use of Ripley's K function. This function measures the expected number of events found up to a given distance of any particular event. It is defined as

$$\mathrm{K}(r) = \lambda^{-1} E[N_0(r)]$$

where $E[.]$ denotes the expectation and $N_0(r)$ represents the number of further events up to a distance r around an arbitrary event.



(a) Ward 62

(b) Ward 84

Figure 17: Ripley's K Function

The results from Ripley's K function again demonstrate that both wards contain clustered listings as both observed lines are greater than the expected lines. Ward 84 again shows stronger evidence of being clustered.

Clark-Evans Test

My final analysis on the point pattern for the wards will be in the form of a Clark-Evans test for complete spatial randomness. The formulation for this test is as follows:

Start by taking the average of the e2e distances for m randomly sampled points in a point pattern, $\bar{d} = \frac{\sum d_{ij}}{n}$.

Then calculate the expected value $E[D]$ for a complete spatial randomness with the same intensity λ, to obtain an index of spatial regularity: $E[D] = \frac{1}{(2\sqrt{\lambda})}$

Finally, divide the average observed e2e distance by the expected value under complete spatial randomness.

$$R = \frac{\bar{d}}{E[D]}$$

Then we test by approximating the distribution of R under complete spatial randomness by a normal distribution with mean $= 1$, and variance $s^2 = (\frac{1}{\pi} - \frac{1}{4})(n\lambda)$. The results are in the following table:

Clark-Evans test		
Ward 62	R	= 0.6478
	p val.	< 2.2e-16
Ward 84	R	= 0.3664
	p val.	< 2.2e-16

We interpret the results as follows:

$R > 1$, suggests regularity

$R = 1$, is consistent with a completely random pattern

$R < 1$, suggests clustering

We obtain an R figure of 0.6478 and 0.3663 for ward 62 and 84 respectively. Thus, by the Clark-Evans test, the listings are in a clustered pattern type. It is noted that a weakness of this test is that it assumes that the point process is stationary. An inhomogeneous point pattern will typically give $R < 1$ and can produce false significance. The result of this test falls inline with the results from other measures to understand the point pattern. Therefore I am happy with this result.

Part 3: Spatial Lattice Data Analysis

This section looks at spatial lattice data analysis, where our primary focus will be on the price variable.

EDA and data preparation

To obtain the mean variables per ward, I use the listings dataset and group the variables by ward number. I then take the mean of the respective variables. I also decide which of these variables I wish to include in my analysis. I remove the average acceptance rate and average review scores as there are too many NA values. Finally, I perform principal component analysis to help decide which variables I would like to supplement the price variable.

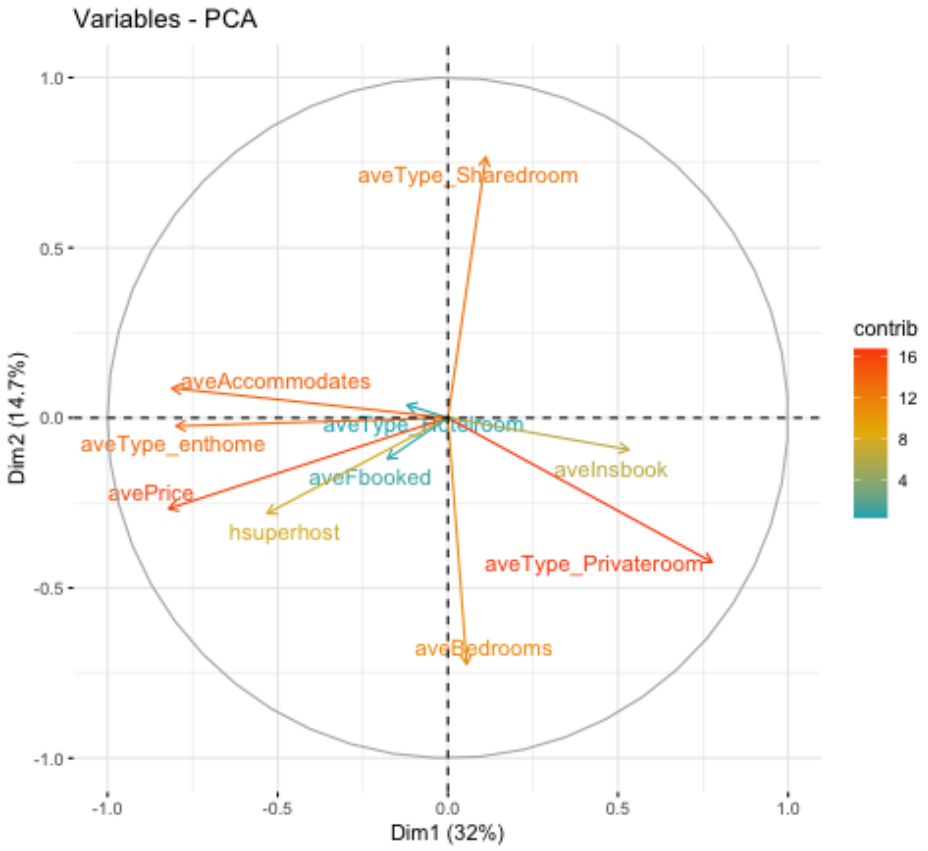


Figure 18: Principal Component Analysis

From the PCA, I remove variables the average fully booked, average roomtype of hotel room and average instantly bookable variable and these do not seem important.

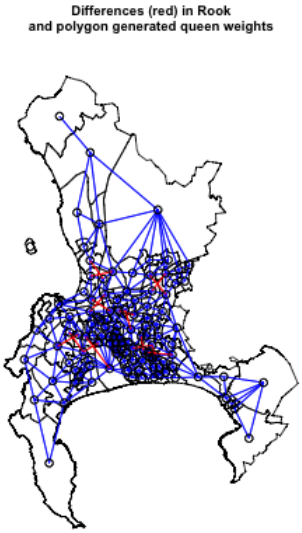
This concludes my data preparation and I now move onto obtaining the neighbourhood relationship.

Contiguity Based Neighbours

The first type of relationship I look at are contiguity based neighbours, namely the rook and queen neighbour criterion. Like the moves in chess, rook criterion can move to a neighbour that shares a side where as a queen can move to a neighbour that shares a side and a vertex. I plot the rook and queen criterion and then compare the difference between the two of them.

(a) Rook criterion

(b) Queen criterion



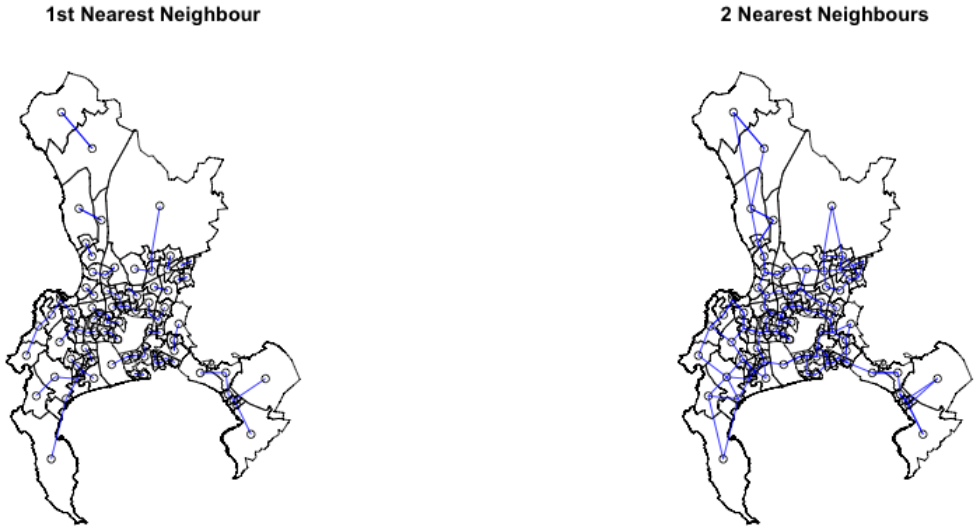
(c) Differences

Figure 19: Queen and rook neighbours

At a glace, it is very difficult to note the difference between the rook style neighbours and the queen style neighbours. Since the rook criterion is a subset of the queen criterion, I plot the additional neighbours below in figure (c). One can note very few additional naighbours when changing from rook to queen criterion. This is understandable because as the wards are shaped very differently and it is not common to share a vertex. We can therefore expect very similar Moran I results.

Distance based neighbours

In this section I look at three kinds of distance based neighbours, the first nearest neighbour, two nearest neighbours and the cut-off method. All of these methods are measured between the centroid of each ward. The first nearest neighbours measures the nearest one neighbour, and the 2 nearest neighbours method has neighbours with the two nearest neighbours. These are plotted below:



(a) 1st nearest neighbour

(b) 2 nearest neighbours

Figure 20: Queen and rook neighbours

We note in the first nearest neighbour that some are cut off from the rest of the wards as they each share a nearest neighbour with each other. This is not the case for two nearest neighbours. The two nearest neighbours plot as twice the number of neighbours.

Next I move onto the critical cut-off method which creates neighbours for each ward that is less than a cut-off distance away.

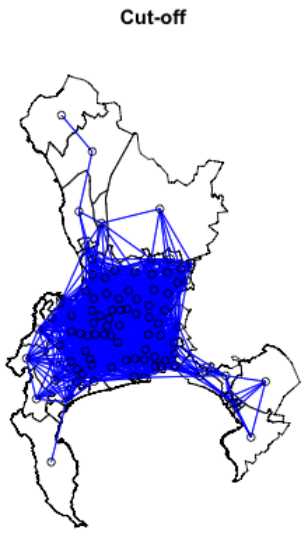


Figure 21: Critical Cut-off

From the plot we can see a dense blue shape where all the smaller wards are neighbours with each other. The Northern most ward only has one neighbour, and if in fact, the wards on the outskirts have fewer neighbour than those that are central.

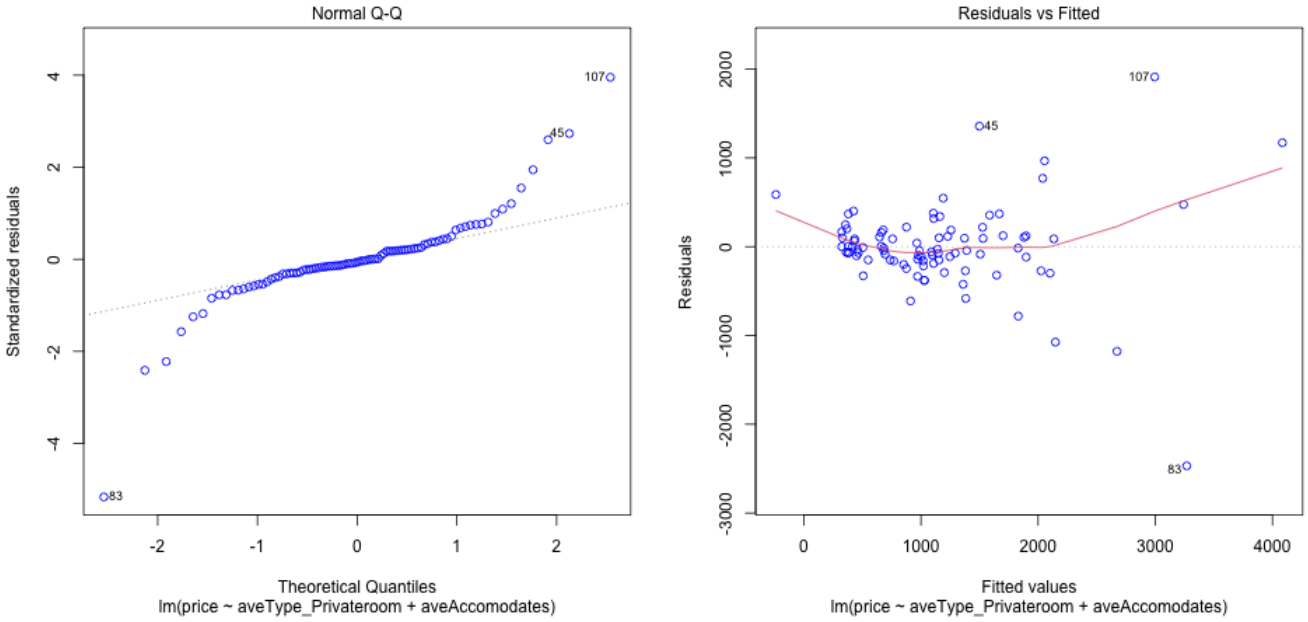
To compare these distance measures I perform Moran's I test for each neighbourhood method. The results are shown in the table:

	Rook	**Queen**	**1 NN**	**2 NN**	**Cut-off**
MI	0.50445	0.49495	0.36487	0.46028	0.22029
E[MI]	-0.0114	-0.0114	-0.0114	-0.0114	-0.0114
Var[MI]	0.00504	0.00487	0.01530	0.00805	0.00106
z-value	7.2692	7.2545	3.0403	5.2567	7.095
p-value	1.808e-13	2.016e-13	0.00118	7.333e-08	6.468e-13

Above are the results from after performing Moran's I test using five different neighbour methods. The best method would be the one that achieved the lowest p value, while simultaneously keeping variance low. The rook, queen and cut-off methods achieved the lowest p values. I will choose the rook style neighbours method as it has the greatest Z-value and lowest p value while still having low variance. Now that I have a neighbourhood, I move onto building spatial lag and error models to predict the average price.

Model building

Using the rook style neighbours, I now create model my models. I start by performing Ordinary least squares (OLS) regression.

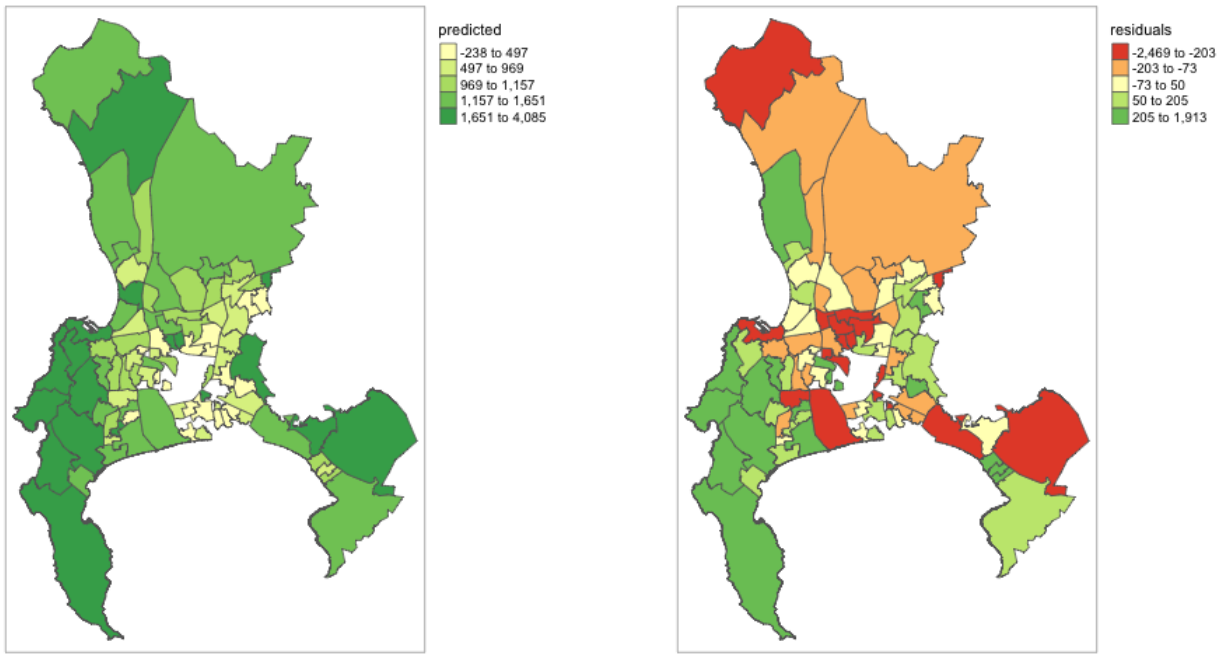


(a) QQ Plot

(b) Residuals

Figure 22

By the QQ plot we note that the dataset takes the form of a normal distribution with a tail on either end. The residuals vs. fitted suggests that a linear model will be appropriate.



(a) Predicted

(b) Residuals

Figure 23

By OLS we note that the Cape peninsula wards tend to be prices higher than those in the city. Looking at the

residuals we note that the residuals are greater in more densely populated areas. This could be due to there being many listings so lots of variability.

The models are as follows:

Model 1 :First Order Spatial Lag
Model 2 :Spatial Lag
Model 3 :Spatial Error Model
Model 4 :Spatial Lag and Spatial Error Model

The spatial lag models are defined as:

$$\boldsymbol{y} = \rho \boldsymbol{W}^s \boldsymbol{y} + \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$$
$$\varepsilon \sim N\left(0, \sigma^2 I_n\right) \text{ and } |\rho| < 1$$

The spatial error model is defined as:

$$y = \boldsymbol{X}\beta + u$$
$$u = \lambda \boldsymbol{W}_2^s u + \varepsilon$$
$$\varepsilon \sim N\left(0, \sigma^2 I_n\right)$$

The results for the above defined models are as follows;

Model	Coefficient	Estimate	Std. Error	z-value	$Pr(>\|z\|)$
1	(intercept)	396.26	119.33	3.3208 7	0.000897
2	(intercept)	-671.09186	192.18853	-3.4918	0.0004797
	privateRoom	211.54369	206.70097	1.0234	0.3061052
	aveAccommodates	3.94180	0.38556	10.2235	$< 2.2e-16$
3	(intercept)	-385.47640	207.97869	-1.8534	0.06382
	privateRoom	287.96407	218.94643	1.3152	0.18843
	aveAccommodates	4.15998	0.38843	10.7096	$< 2.2e-16$
4	(intercept)	-691.57491	180.28173	-3.8361	0.000125
	privateRoom	188.94057	197.08746	0.9587	0.337728
	aveAccommodates	3.75498	0.41373	9.0758	$< 2.2e-16$

Table 2

In the first order spatial lag model we find the price variable to be significant with a high z-value of 3.32 and a p value < 0.001. Comparing the variables in the dependent models we note that the privateRoom variable does not show significance to any of the models. In the contrary, the aveAccommodates variable is significant in all the models with a p value of $< 2.2e^{-6}$. This low p value is inline with the high z-values obtained of 10.22, 10.71 and 9.08 in model 2, 3, and 4 respectively. The intercept (price) shows significant in model 1, 2 and 4 at the 5% significance level but does not show significance in spatial error model. Significance means price is very dependent on the price defined by the neighbours, where the neighbours are defined by rook criterion. Therefore in model 1, price is very dependent on the neighbours price, in model 2 price and aveAccommodates is very dependent on the neighbours, in model 3 only the aveAccommodates is very dependent on the neighbours (at 5% significance level) and finally in model 4 price and aveAccomodates are very dependent on the neighbours.

Next I look at the ρ and/or λ values, the LR test value, Wald statistic and log-likelihood of all the models.

Model	ρ/λ	LR test value:	p-value:	Wald statistic:	p-value	log-likelihood
1	0.64693 (ρ)	37.837	7.6909e-10	56.034	7.1276e-14	-719.5236
2	0.34415 (ρ)	17.139	3.4749e-05	16.709	4.3564e-05	-677.7786
3	0.46431 (λ)	9.9689	0.0015921	17.124	3.5024e-05	-681.3634
4	0.42649 (ρ) -0.22433 (λ)	18.028	0.00012167			-677.3337

Table 3

$\boldsymbol{\rho/\lambda}$: The ρ and λ are the spatial autoregressive coefficients therefore it is important to assess these be the means of an likelihood ratio test (LR). The LR test tests for the absence of spatial dependence in spatial lag or error models therefore a significant figure means there is spatial dependence occurs in the models. Given this, all four models are found to significant as $p < 0.001$ by the LR test meaning spatial dependence exists. This corresponds to a high LR test values.

Wald Statistic: The Wald statistic also tests for the absence of spatial dependence in spatial lag or error models. All the model's were found to be significant as they obtain high Wald statistics and low p-values, which means spatial dependence exists in the models.

Log-likelihood: The log-likelihood is where the model is maximized therefore a greater log-likelihood figure is indicative of a better model. We note that there model that performed the worst was model 1 with a log-likelihood of -719.52. This is understandable as it is simple a first order spatial lag model and does thus does not contain any dependent variables. In model 2 we see an improvement in log-likelihood with a figure of -677.78 which means adding dependent variables improve the model. Finally in model 4, where we take into account spatial error, we obtain our best model with a log-likelihood value of -677.33.

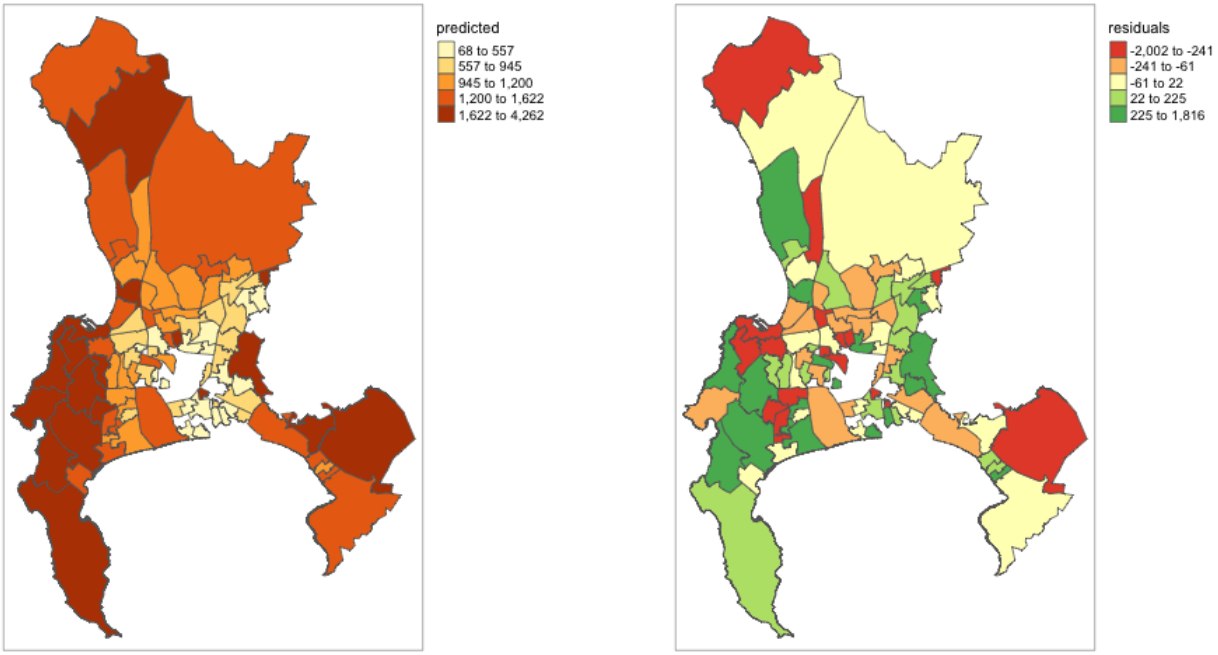
Next I perform Lagrange Multiplier (LM) tests for spatial auto correlation:

	Statistics	**df**	**p-value**
LMerr	10.676733	1	0.0010849
LMlag	10.184791	1	0.0014160
RLMerr	2.999514	1	0.0832895
RLMlag	2.507573	1	0.1133003
SARMA	13.184306	2	0.0013711

Table 4

We start by looking at the simple LM error and LM lag test and find them to both be significant at a 1% significance level. We then perform robust LM error and LM lag tests and find that both are not significant at a 5% level. There we reject both the spatial lag and the spatial lag and the spatial error models and keep the OLS regression instead.

After comparing different models I plot the predicted price for each wards using the spatial lag model:



(a) Predicted

(b) Residuals

Figure 24

From the predicted results we note that the Cape peninsula is still predicted to have much higher prices than the central wards. There are also wards with higher average prices towards Somerset West. The residuals seem to vary and is difficult to understand a trend. This concludes the spatial lattice data analysis.

Conclusion

In part 2 we looked at the point pattern of the listings in ward 62 and ward 84. From all the tests and graphs we concluded that the listings in both wards were found to be clustered. In part 3 we looked at spatial lattice data analysis where we tested for spatial dependence between the wards. Through building build spatial lag and error models it was found that spatial dependence does exist were able to

- END -

Refereces

All slides given by Dr Sebnem Er. R code given by Dr Sebnem er: `https://sebnemer.github.io/english/courses/asda2020/index.html`