School of Engineering

Master of Information Technology

COMP90055 Computing Project

25 point Research Project Report

Exploring Determinants of Housing Prices in Victoria State Using Open Data

October 2018

Wuang Shen

Student number: 716090

Supervisor: Professor Richard Sinnott

Acknowledgements

I would first like to thank my supervisor Prof. Richard Sinnott for giving me the opportunity to develop a project with my previous knowledge on the real estate industry and providing me with the support, guidance and materials.

As this project is built based on previous works done by Zhanyi (Zoe) Qi and David John Wang, so I would also like to thank them for providing me with the skeleton to develop the system on a different research topic.

Abstract

The real estate market in Australia grows every year since the financial crisis in 2008. There are a lot of factors needs to be considered when people invest in real estate. The hedonic house price model (HPM) estimates the value of properties by adding separate properties' attributes. Locational attributes and neighboured characteristics are major concerns when applying HPM on the study of the association between external factors and the house value. By using techniques and tools like the AURIN platform, Jupyter notebook, this project provides a system to understand the status of the current house market in Victoria state, and the associations between house price increase rate and locational attributes. Multiple statistical methods like Maximum information coefficient, ordinary least squares, Pearson's correlation coefficient are used in the analysis to test the correlation between house price increase rate and targeted factors. The results of this project show that some locational variables and neighbourhood characteristics are correlated with the house price increase rate in Victoria state.

Certification

I certify that

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.

- where necessary I have received clearance for this research from the University's Ethics Committee and have submitted all required data to the School

- the thesis is 5841 words in length (excluding text in images, table, bibliographies and appendices).

Wuang Shen

Contents

Introduction

House’s location plays an essential role in the real estate market. Location can mean heaps of things, and investors value various things in a location. With a great location, residents can easily get access to public transportations, supermarkets and other necessary infrastructures. Thus, this paper examines how the increase in house price and investors’ buying behaviours varies by different locations.

The focus of this paper is to analyse how different locational attributes and neighbourhood characteristics associates with the house price increase rate. This project will first explore the status and the trending of current house market in Victoria state. The price movement in Victoria state will be captured and analysed. Secondly, top 10 areas in Victoria state with high house price increase rate are extracted from the datasets. Thirdly, various locational features will be selected and tested against the house price increase rate. Attributes are evaluated by statistical testing. Neighbourhood characteristics and locational effects are measured to determine which factors explain the difference in price increase rate among different areas.

Hedonic House price model

HPM is a price model used to estimate the price of commodities by adding up values of their constitute properties. The demand for a commodity can be determined by HPM model. This method is extensively used in the real estate related topic. (Herath & Maier, 2010)

In Hedonic house price model, the price of houses can be estimated by their structural attributes, locational variables, neighbourhood characteristics and time. (Hearth, 2015)

$$P = f(S, L, N, t)$$

(Where P is house prices; S is structural attributes; L is locational variables; N is neighbourhood characteristics; t is an indicator of time.)

Structural attributes are houses’ physical features like the size of a house, car parking space, built-in cooling systems, etc. They are internal factors determining a house’s price. Locational variables are features consequent from the coordinates of the house, like the commuting time from your home to your workplace, or the distance from your home to CBD. Neighbourhood characteristics are similar to locational variables, but it mainly explains the part of house estimated by the community characteristics like transportation, residency culture background, or even whether people feel safe living in the area. The indicator of time explains the difference in house prices in a different time dimension.

Since the aim of this project is to discover how housing price varies by areas. Locational variables and neighbourhood characteristics are the main focuses of this paper.

Datasets

AURIN

Datasets used in this project are collected from the AURIN platform. AURIN is a collaborative national project provides researchers, designer and city planners with a distributed network of urban and environment related information and datasets. Users can get access to AURIN datasets through AURIN Portal and AURIN API. Resources shared in AURIN platform play an essential role to understand the status of current urban development and direction of future urban growth in Australia. (AURIN, 2018)

AURIN Portal is a workbench tool that integrates different urban related datasets into a one powerful online analytical platform. AURIN Portal is free and open to all researchers and staff from the education and government sectors. Users can get access to various licensed, spatially-enabled datasets through AURIN Portal. It also enables users to visualise and perform analysis datasets on the platform. (AURIN, 2018)

AURIN API is an alternative to get access to AURIN datasets. It is an application programming interface(API) that enables researchers to reach to a variety of open source datasets without console login. Although AURIN API is less restricted than AURIN Portal, available datasets through AURIN API are limited. (AURIN, 2018)

House price dataset

House price dataset used in this project is assessed through the AURIN Portal, which is published by Australia Property Monitors (APM). APM is a property intelligence platform that provides property information to banks, government, media and real estate industries. APM house price dataset covers property's sale, rent and sold time-series data across Australia for the period from 01/01/1986 to 31/10/2017, with 12-month aggregations. The house price datasets are aggregated at SA2 (Statistical Area 2), SA3, SA4 and State level. In this project, dataset aggregated at the SA2 level is used, as this aggregation level is close to the definition of suburbs. (Aggregation will be explained in a separate section in this paper.)

Tools and Technology used in this project

Jupyter Notebook

Jupyter Notebook is a web application that allows users to put codes, program executions, texts and visualisations into a single document. Users can share notebook across various platforms like Github, Dropbox and email. Jupyter Notebook is widely used in data science project, as it leverages big data tools. It supports more than 40 programming languages, like Python, R and Scala (Jupyter, 2018). The programming language used in this project is Python 2.7.

ASGS Aggregation

To assist users' analysis, visualisation, integration on statistics, Australian Statistical Geography Standard (ASGS) designed a framework of statistical areas that provides users with an integrated set of standard areas. The ASGS consists of the ABS structures and the Non-ABS Structures. The ABS structures are designed to provide statistics output. Areas aggregated with ABS structures is to aid the ABS in collecting statistical information. Aggregation levels like Statistical Area (SA), Greater Capital City Statistical Areas (GCCSA), and State/Territory are defined by ABS structure. This structure provides a stable solution to ensure the accuracy, relevance and confidentiality of the collected data. The Non-ABS structure represents administrative areas which can change regularly. Regular updates to the areas improve the relevance of data collected on these areas. Aggregation levels like Local Government Areas (LGAs), Postal Areas (POA) and State Suburbs (SSC) are defined in this structure. (Australian Bureau of Statistics, 2018)

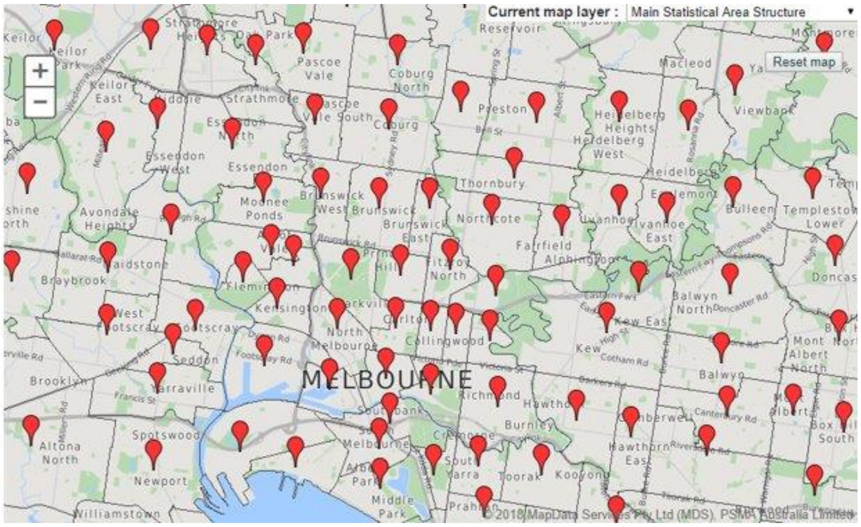
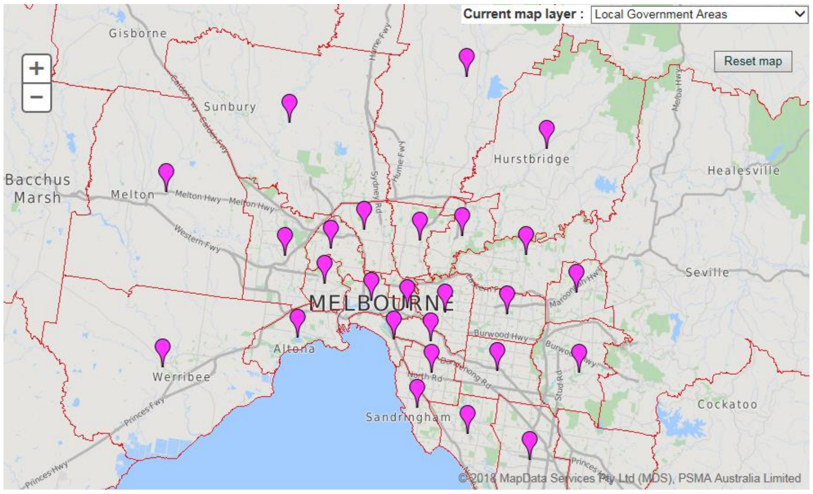


Figure 1: SA2 Aggregation Level



Current map layer : Local Government Areas

Figure 2: LGA Aggregation Level

ASGS Correspondences

The area aggregation level of house price dataset is at SA2 level (Figure 1). Selected locational attributes and neighbourhood characteristics datasets involved in this project are at SA2 and LGA level (Figure 2). Since the definition of the area in SA2 level are different from the one in LGA level, in order to study how locational attributes affect the house price in an area, it is necessary to keep all datasets at same aggregation level. To make the process simple, locational features are converted to SA2 level.

To aid users in converting statistical data to different geographic regions defined in ASGS, ABS provides a broad range of correspondences file, which can mathematically convert data from one aggregation to another aggregation based on a weighting calculation on the location of the population (Australian Bureau of Statistics, 2018).

Statistical Testing

Ordinary least squares linear regression

Ordinary least square (OLS) is a traditional statistical method used to models the relationship between a set of independent variables and a dependent variable. A dependent variable Y is defined as

$$Y = \beta_0 + \ \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

(Where βs are the regression coefficients, Xs are a collection of independent variables, ε represents residual errors that can not be explained by the model). The regression coefficients represent the significance of the associated independent variable's effects on the dependent variable (Pohlman & Leitner, 2003).

The aim of this project is to study the effects of different locational attributes rather than propose a linear model of the house price. Thus, the equation above will be modified to

$$Y = \beta_0 + \ \beta_1 X_1 + +\epsilon$$

, which only captures a single independent variable explaining the value of the dependent variable. The coefficient of determinationR^2 is used to measure the strength of the linear regression model, which indicates the degree of data fitting in the linear model (Hayashi, 2000). The coefficient of determination is defined as

$$R^2 = \frac{\sum(\widehat{y_i} - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

(where $\widehat{y_i}$ is the estimated value of y, $\bar{y}\ is\ the\ mean\ of\ y$).

R^2 ranges from 0 to 1. A higher value of a coefficient of determination suggests the existence of a strong linear relationship between the dependent variable and independent variable.

Pearson correlation coefficient

As the regression coefficients from OLS model are not constrained within the interval of [-1,1], the significance of the independent variable will be hard to interpret. Pearson correlation coefficient (PCC) is an alternative to measures the linear correlation between variables X and Y, and it ranges from -1 to +1, where -1 stand for a strongest negative correlation, 0 stands for no linear correlation, and +1 stands for strongest positive correlation. Pearson correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - x_i)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Maximum information coefficient (MIC)

The dependent variable and independent variable will not always be linearly associated. In some case, variables will be non-linearly correlated. Maximal information coefficient is a statistical method that captures both linear and non-linear relationship between variables X and Y. For variables that have a functional association, the MIC is close to the value of Coefficient of determination. (Reshef et. Al., 2011) MIC ranges from 0 to 1, where the value of 0 suggests that variables are statistically independent, and the value of 1 suggests that variables are highly dependent.

Design of the system

The house price analysing system used in this project has six stages (Figure 3).

Stage 1. The python script in Jupyter Notebook loads APM housing prices datasets into the system. The system displays each area's rank and price information. Users can enter an area ID to view its time-series graph.

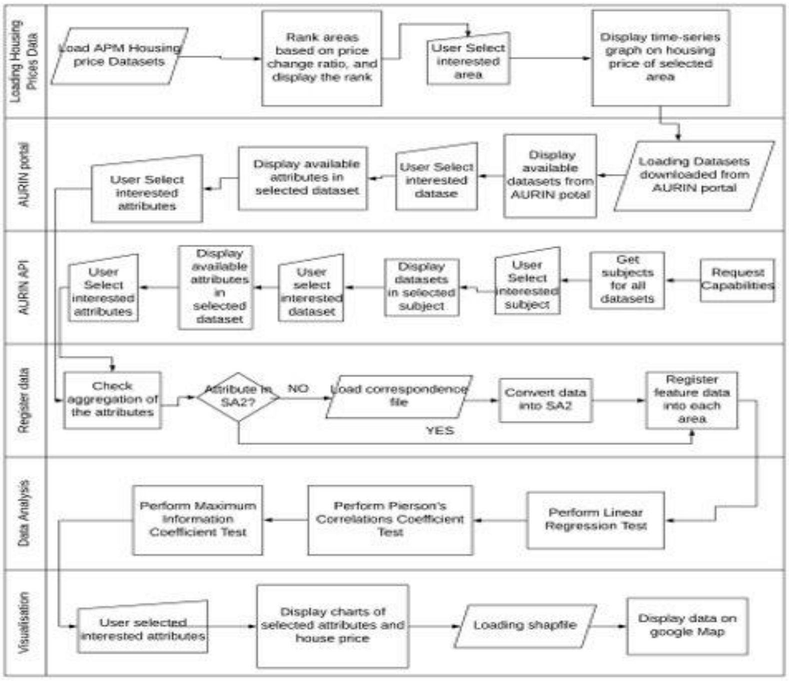
Stage 2. Attributes datasets from downloaded from AURIN Portal will be loaded into the system. The system outputs available datasets name and attributes name. Users need to select interested dataset index and attribute index manually.

Stage 3. The system request capabilities to AURIN platform through AURIN API. The system retrieves available dataset topics, datasets name and attributes. Users need to select topics, datasets and attributes manually.

Stage 4. After loading both house price datasets and locational attributes datasets, the system will check each attribute dataset's aggregation level. If datasets are not aggregated at the SA2 level, the system will perform aggregation conversion on those datasets. Then the system combines all datasets.

Stage 5. The system asks the user to choose a feature name to perform data analysis.

Stage 6. Statistical result and scatter plot from stage 5 will be returned. The system loads shapefile. The user needs to choose a feature name to display on google map. Map visualisation on selected feature will be displayed.



Loading Housing Prices Data
Load APM Housing price Datasets
Rank areas based on price change ratio, and display the rank
User Select interested area
Display time-series graph on housing price of selected area
AURIN portal
Loading Datasets downloaded from AURIN portal
Display available datasets from AURIN potal
User Select interested datase
Display available attributes in selected dataset
User Select interested attributes
AURIN API
Request Capabilities
Get subjects for all datasets
User Select interested subject
Display datasets in selected subject
User select interested dataset
Display available attributes in selected dataset
User Select interested attributes
Register data
Check aggregation of the attributes
Attribute in SA2?
NO
Load correspondence file
Convert data into SA2
Register feature data into each area
YES
Data Analysis
Perform Linear Regression Test
Perform Pierson's Correlations Coefficient Test
Perform Maximum Information Coefficient Test
Visualisation
User selected interested attributes
Display charts of selected attributes and house price
Loading shapfile
Display data on google Map

Figure 3: Design of the house price analysis system

The current state of the housing market in Victoria, Australia

A few areas’ time series data from 2010 to 2017 were retrieved. Most Victoria areas' house price trending is similar. There are increases in house price for most areas in Victoria after 2014. This trend is still growing. According to the rank output from Jupyter, the area with the highest house price increase rate from 2010 to 2017 is East Melbourne, which is increased by 138.4%. The median house price increase rate is 70.5%. Besides, most top 10 areas (Table 1) are near Melbourne and located at the south-east of Melbourne, and their house prices are all doubled.

Area Name	Increase Rate	Previous Price	Current Price
East Melbourne	138.39%	1,336,000	3,185,000
Carlton	134.67%	620,000	1,455,000
Box Hill	127.58%	620,000	1,411,000
St Kilda	125.65%	672,500	1,517,500
Ashburton (Vic.)	122.61%	785,000	1,747,500
Alphington- Fairfield	121.39%	673,000	1,490,000
Hawthorn East	118.43%	1,057,500	2,310,000
South Yarra- East	115.06%	756,500	1,627,000
Hughesdale	111.69%	633,000	1,340,000
Ararat Region	110.19%	130,000	273,250

Table 1: Top 10 Areas with high increase rate

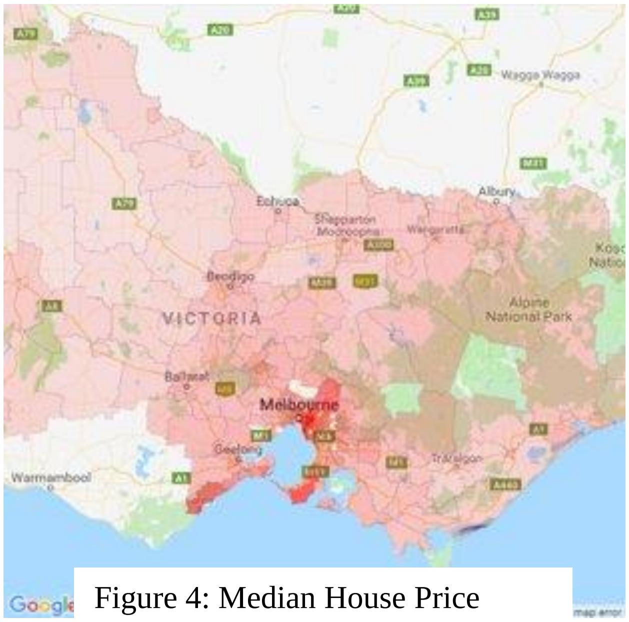


Figure 4: Median House Price Distribution in 2010

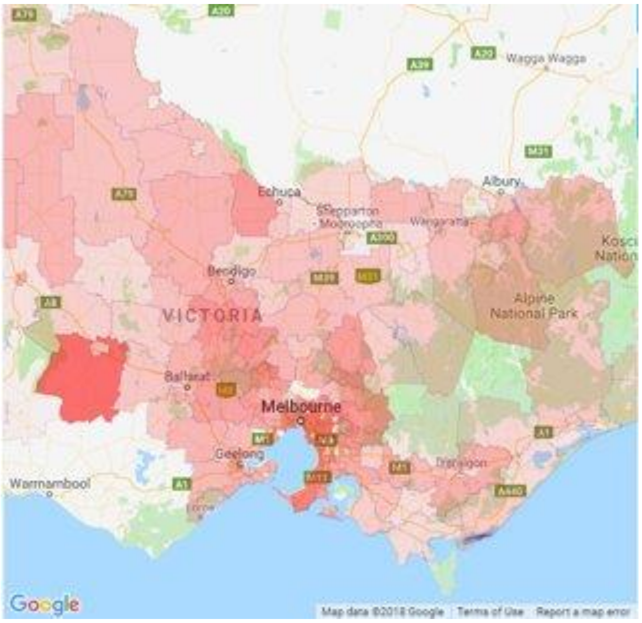


Figure 5: Distribution of House Price increase rate from 2010 to 2017

Figure 4 is a map shows median house price distribution in Victoria in 2010, while figure 5 represents house price increase rate in Victoria from 2010 to 2017. Darker red colour means those areas are with higher house prices and higher increase rates in figure 4 and figure 5 respectively. By comparing both figures, it is easy to see there exist some similar patterns in both figures. Areas located around Melbourne CBD or south-east part of Melbourne tend to be areas with the darker colour in both figures. Besides, the distribution of median house price is similar to the distribution of house price increase rate in Victoria state. This indicates that areas with higher house prices are most likely to have high increase rates in property price. This fact can be explained by locational and neighbourhood characteristics of an area. As the house price is a combination of values of structural features and locational features in Hedonic House Price model, areas with higher house price are normally equipped with more mature fundamental facilities and services like public transportation and schools. As HPM defines people's buying behaviour – their willingness to purchase commodities, thus, when the house market boom starts, the demand for houses located in areas with mature infrastructure will be higher than others, which in turn makes their increase rate higher than other areas.

Scenarios involved in this project

- Neighbourhood characteristics – Family weekly income, Crime rate, public transportation, green space, born overseas population, commercial land.
- Locational variables – Distance to CDB.

As mentioned above, areas with high house prices are highly likely to be with high house increase rates. Since weekly family income represents a family's affordability on the property, it is expected to be positively related to the house price increase rate. People with high income have more financial abilities to invest their money in the real estate. Meanwhile, areas with better infrastructures are more attractive to property investors, who possess great wealth.

In this paper, the public transportation level of an area is measured by the amount of tram, train and bus stops available in the area. As obtained datasets do not have attributes that describe the transportation level, the transportation level will be defined as

$$\text{Public transportation score } = C * \left(\frac{number\ of\ Tram\ stops}{Total\ number\ of\ Tram\ stops\ in\ Vic} + \frac{number\ of\ Bus\ stops}{Total\ number\ of\ Bus\ stops\ in\ Vic} + \frac{number\ of\ Train\ station}{Total\ number\ of\ Train\ station\ in\ Vic}\right)$$

(Where C is a constant number, C = 1000 is used in this project).

The public transportation level determines the convenience for residents' daily travels. The importance of the public transportation may vary by person to person. Different from people who commute with public transportations, residents who travel with their cars, are likely to value this factor less. In most cases, the public transportation level is expected to be positively associated with the house price increase rate.

The crime rate is a negative factor associating with the house price. There are several types of research found the existing of an inverse association between the local crime rate and the house value. According to Gibbons (2004), every increase of one-standard-deviation in local crime results in a 10% decrease in house values. To reduce the exposure to crime risk, residents may pay a higher price to live in an area with the low crime rate and sell property with high crime risk (Linden & Rockoff, 2008). Thus, the relationship between crime rate and increase rate in housing price is expected to be negative.

Green space like parks and reserves make the neighboured environment more pleasant to live. Trees and plants improve an area's air quality and reduce noise in an area. Thus, adequate green spaces normally add value to local properties. In this paper, the amount of green space in an area is expected to have a positive association with the price increase in the local house.

Born overseas population represent the immigration level in an area. Australia accepts immigrants from all over the world every year. Immigrants increase the demand for the property while contributing to multicultural Australia. To settle down in Australia, immigrants have incentives to purchase properties. Besides, people tend to live in neighbourhoods with a larger portion of residents from same cultural background or race. (Havekes et. Al., 2016). Hence, high immigration level in an area makes local properties more valuable to immigrants. The house price increase rate is expected to be positively related to the percentage of born overseas population.

Same as public transportation, commercial land usage is another indicator that shows services and facilities in the neighbourhood that are available to residents. Facilities like supermarkets, shopping malls provide residents with convenience to get access to essential everyday products. This convenience will reduce residents' cost and time if commercial facilities are close to residents' properties. Therefore, commercial land usage adds value to local properties price, which is expected to have a positive association with house price increase rate in the local area.

Locational amenities like distance to CBD is the consequence of an area's spatial attributes. Normally, people prefer areas close to CBD, as these areas are equipped with better infrastructures and close to most people's working place. Residents can save a lot of time from daily commuting if they live and work around CBD. Therefore, distance to CBD is expected to be negatively related to the increase in house price.

Results

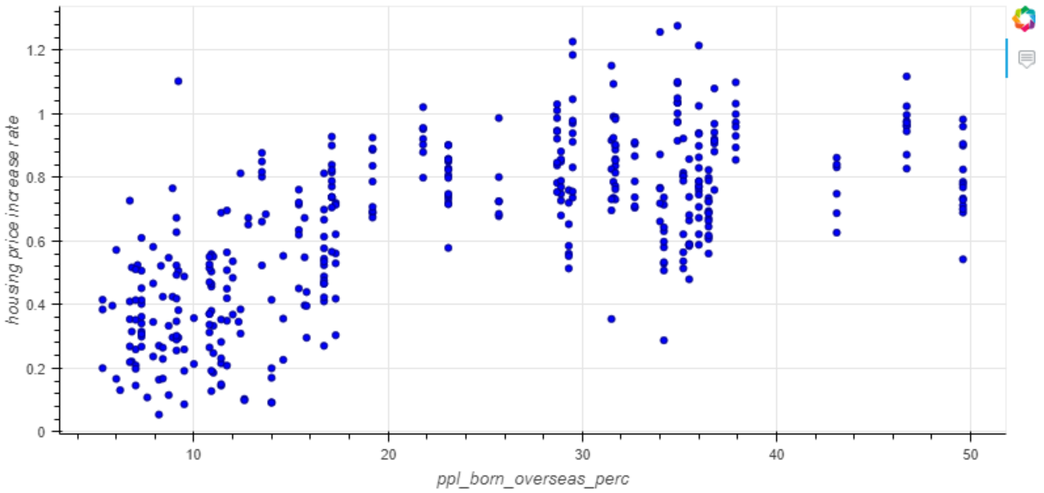
Followings are statistical results (Table 2) generated through the Jupyter notebook. To improve results from statistical testing, outliers like extreme values are removed. (Any attribute value x higher than $\bar{x} + 2 * \sigma$, or lower than $\bar{x} - 2 * \sigma$ will be determined as outlier data.)

	Family weekly income	**Born overseas population**	**Public transport**	**Commercial land**	**Crime rate**	**Distance to CBD**	**Green Space**
R^2	0.774	0.881	0.509	0.565	0.7389	0.201	0.634
Pierson Correlation Coefficient	0.274	0.673	0.472	0.564	-0.201	-0.704	0.0389
Maximal information coefficient	0.263	0.577	0.514	0.590	0.393	0.574	0.190

Table 2: Statistical Results for locational and neighborhood attributes

According to attributes' Coefficient of determination R^2, most features except distance to CBD, have a value of R^2 higher than 0.5, which indicates relatively high linear correlation with house price increase rate. Pierson Correlation Coefficient shows the extent of the relationship between attributes and the increase rates in house price. Attributes like percentage of born overseas population, commercial land and distance to CDB have a relatively significant relationship with house price increase rate, while associations with green space, crime rate and family income and public transportation are relatively less obvious. Same as what we expect, the crime rate and distance to CBD are negatively associated with house price increase rate. Maximal information coefficient provides consistent results with R^2 and PCC. MIC does not only explain the significance of locational attributes in a linear model but also reflect the importance of non-linearly associated features.

Figure 6: Scatter plot of the percentage of born overseas population



Percentage of the born overseas population has relatively high R^2, PCC and MIC among selected neighbourhood attributes, which indicates the immigration level in an area has a significant association with house price growth. Two clusters are existing in scatter plot (Figure 6). The first cluster's percentage of born overseas population roughly ranges from 6% to 18%, while the second cluster ranges from 29% to 38%. Besides, the mean value of the increase rate in the first cluster is around 50%, while the mean value of the increase rate in the second cluster is around 80%. It proves the significance of immigration level on house price increase rate. Moreover, in the first cluster, the house price increase rate rises with the growth of the immigration population. However, the second cluster shows a different pattern. There is no strong linear association in the second cluster; most areas float with house price increase rate around 80%. Figure 7 shows the percentage of the born overseas population in each area. From facts shown in Figure 6 and Figure 7, we can infer that the areas in the first cluster are outer areas, while the second cluster represents inner Melbourne. Thus, the relationship between the percentage of born overseas population and house price increase rate can be concluded into two points. First, areas with higher immigration level normally tend to be in higher increase rate. But it is not clear if it is the immigration level cause the area with higher house price increase rate, or immigrants prefer to live in the areas with great potentials to have house price growth. Second, positive association between house price increase rate and percentage of the born overseas population are more obvious in the outer area rather than the inner area.

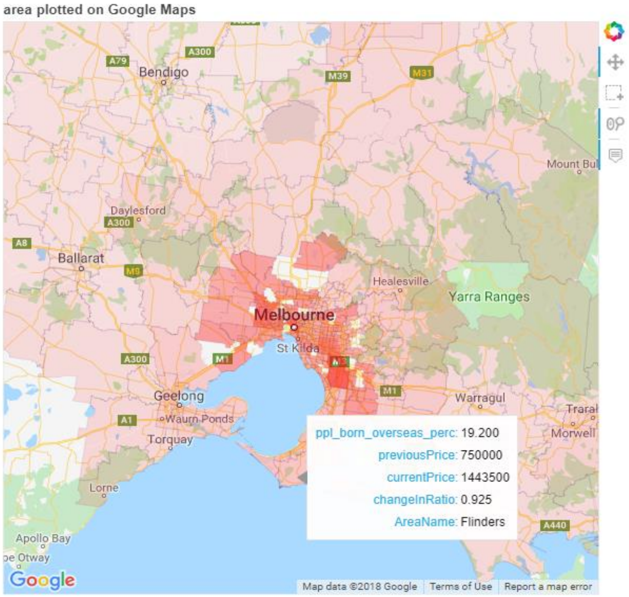


Figure 7: Distribution of the percentage of born overseas population

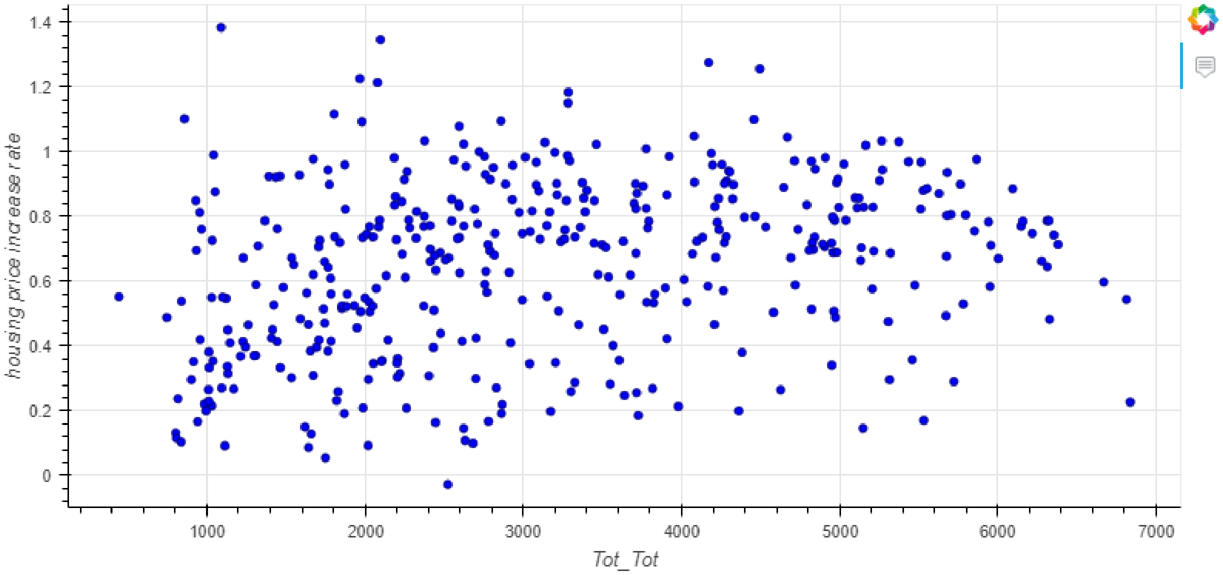


Figure 8: Scatter plot of the family weekly income

Resident’s weekly family income does have a strong linear association with house price increase rate (Figure 8). However, intuition here is not that high weekly family income results in the increase in their house price. The fact is that people with higher family income have more financial flexibilities to buy properties with better structural, locational, environmental features, which have great potentials to be at a higher price. The positive linear relationship is clearer in the interval between $800 to $3,600 per week than others. For residents whose weekly family income located in this interval, every $1,000 increase in the weekly family area associated with rough 28% increase in their property value. For most people who have the weekly family income higher than $3,600, house price increase rate in their property are clustered at the house price increase rate of 80% The increase in weekly family income does not have an obvious association with higher increase in house price once resident's weekly family income higher than $3,600. This fact explains low values of PCC and MIC for the association between family weekly income and house price increase rate

.

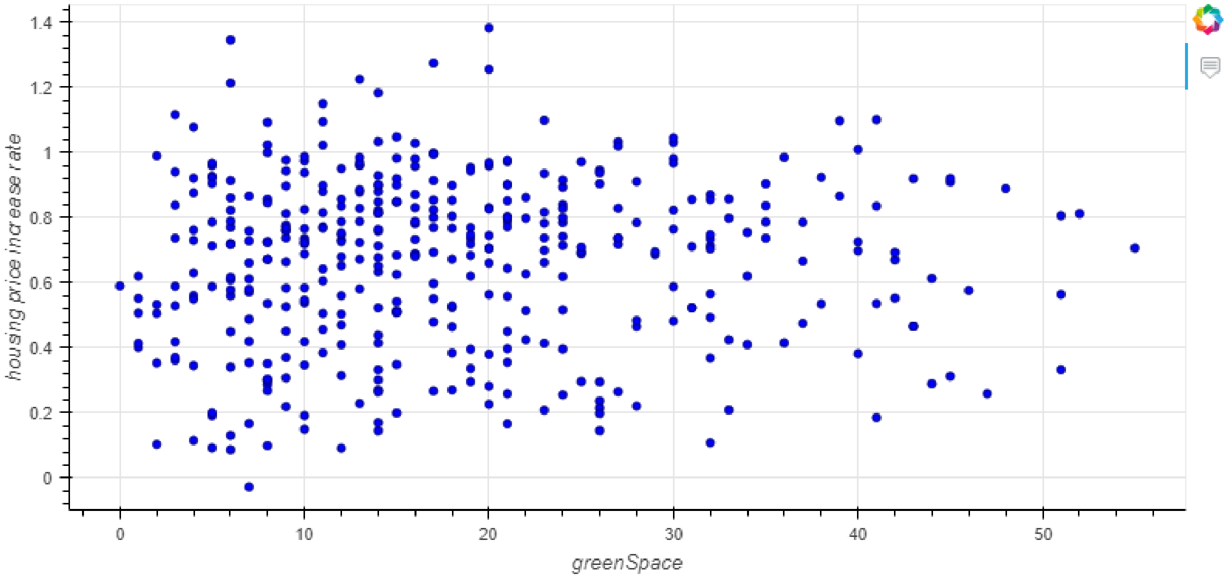


Figure 9: Scatter plot of the green spaces

Apart from weekly family income, green space (Figure 9) is another feature valued with a low MIC. Although green space is with high R^2, the low value of PCC reveals the fact that house price increase rate does not vary by the amount of green space in an area. The high value of R^2 and low value of PCC shows that data are fit into a constant function, where the constant function, in this case, is linear. The amount of green space does not correlate with house price increase rate, in spite of the association between green space and house price increase rate may look linear. In fact, residents in Victoria can easily get access to green space, even though the number of green spaces may vary by different areas. Thus, green space does not add value to house price when comparing areas located in Victoria. Hence, house price increase rate does not reflect the effects from the number of green spaces. However, if the studied subject is in another state or country that is not implemented with great urban planning on green spaces, the results of this factor may be different.

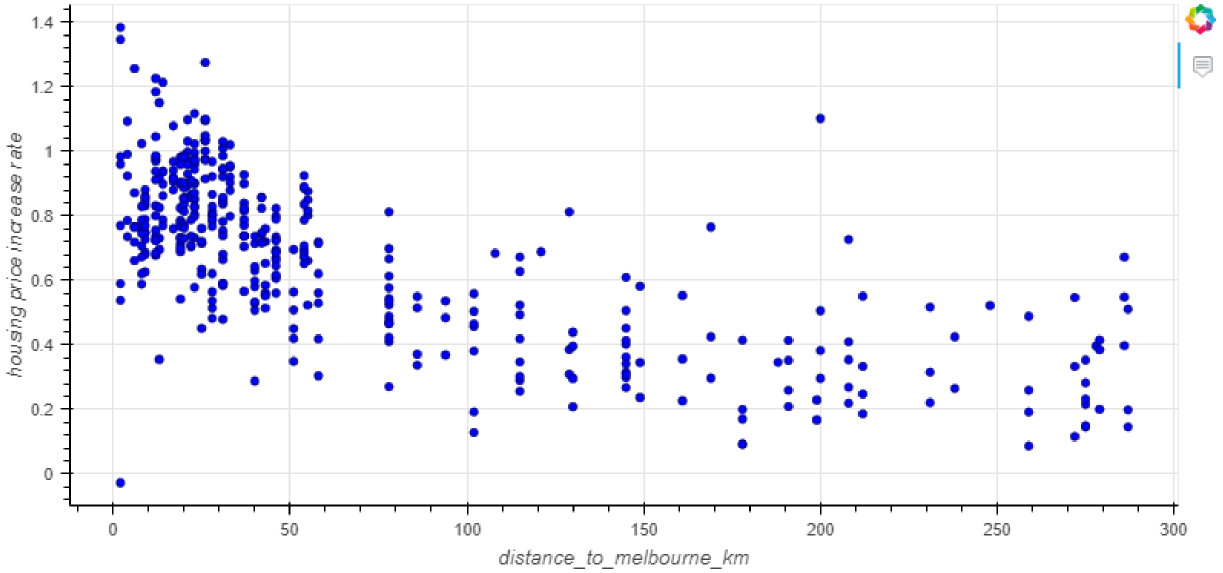


Figure 9: Scatter plot of the distance to Melbourne CBD

Distance to CBD is the only feature that is associated with low R^2, which suggests it’s not linearly related to the increase rate in house price. However, high values of PCC and MIC indicates the significance effects from Distance to CBD on house price increase rate. As shown in figure 9, areas whose distance to CBD are within 60kms are clustered together, and the rest of areas range from 60kms to 300kms. There exists a clear negative linear relationship between house price increase rate and distance to CBD for clustered areas. However, the house price increase rates in the rest of the areas are not sensitive to the distance to CBD. It makes sense that people choose to live close to CBD is because they work around CBD and have needs to go to CBD frequently. Thus, short distance to CBD is a valuable item of an area when those people plan to purchase properties. On the contrary, people live far from CBD normally do not visit CBD every day. As they have fewer incentives to go to CBD, long distance between their home and CBD does not affect their life quality much. Whether they are 100kms far from CBD or 200kms far from CBD are almost the same for people who live in regional areas. Those people do not value this feature much. Thus, for areas with distance to CBD more than 60 kms, there is no any correlation between house price increase rate and distance to CBD, which in turn makes the association less linearly correlated.

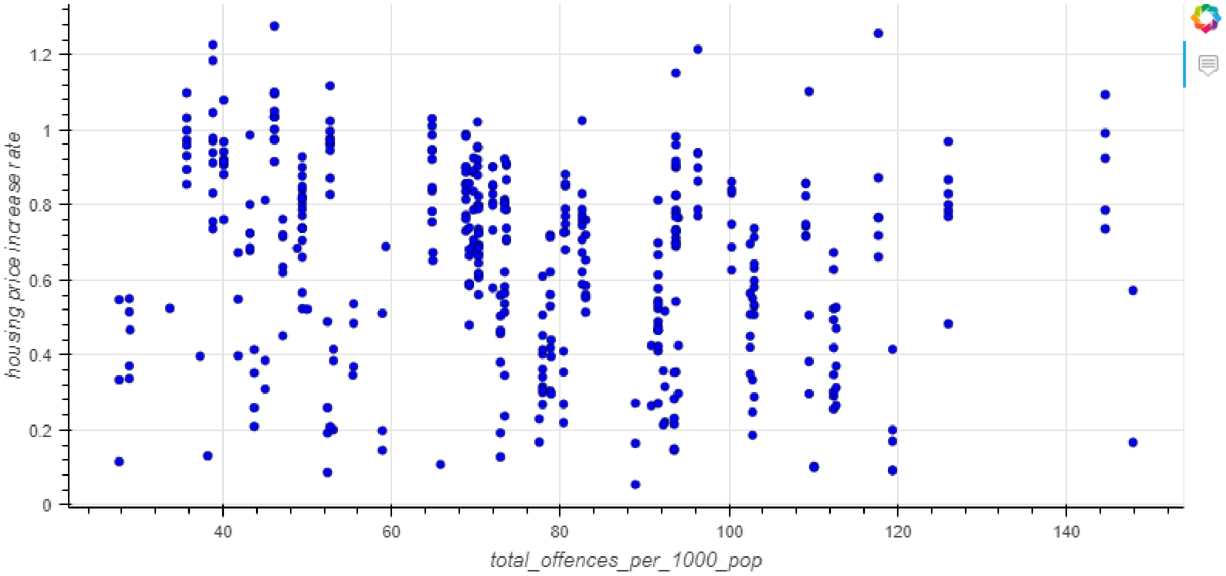
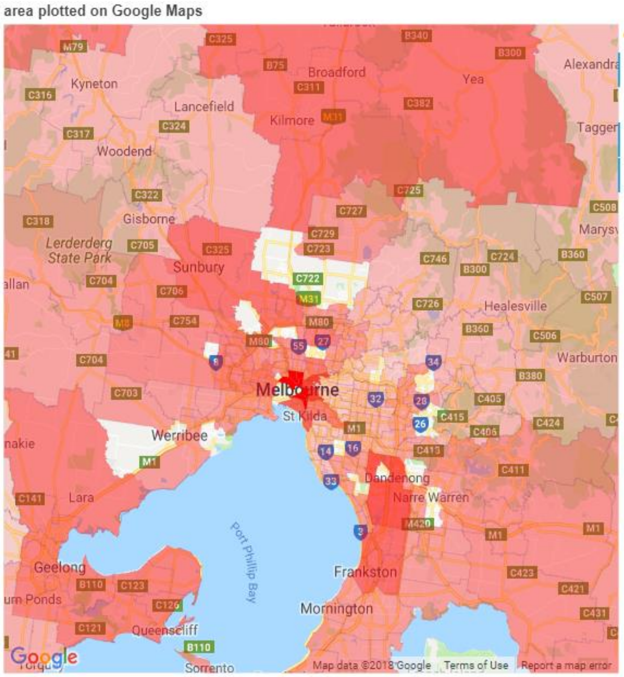


Figure 10: Scatter plot of the crime rate

The crime rate is another factor has a negative impact on house price growth. The total offences in most areas are from 40 to 120 per 1000 population. As shown in Figure 10, the house price increase rate drops with the increase in crime rate. However, this association is not so obvious. The crime rate in an area with high house price increase rate can still be high. As shown in the figure 11, areas like North Melbourne and East Melbourne are with high crime rate and high house price increase rate. According to Braithwaite (1975), areas close to inner city has a higher crime rate than smaller communities, as more crime opportunities provided in larger cities. Therefore, the negative relation between crime rate and house price increase rate in inner-city areas may not be so clear. Hence, the value of MIC and PCC for crime rate are relatively low.



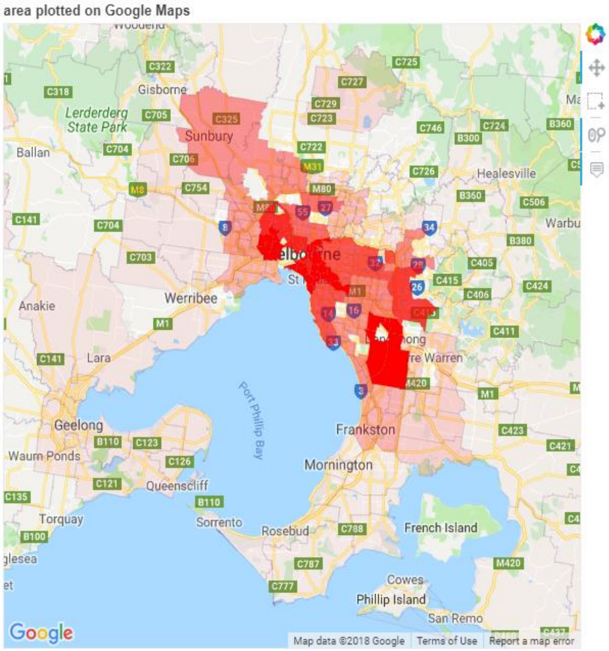


Figure 11: Distribution of Crime rate

Figure 12: Distribution of Commercial land usage

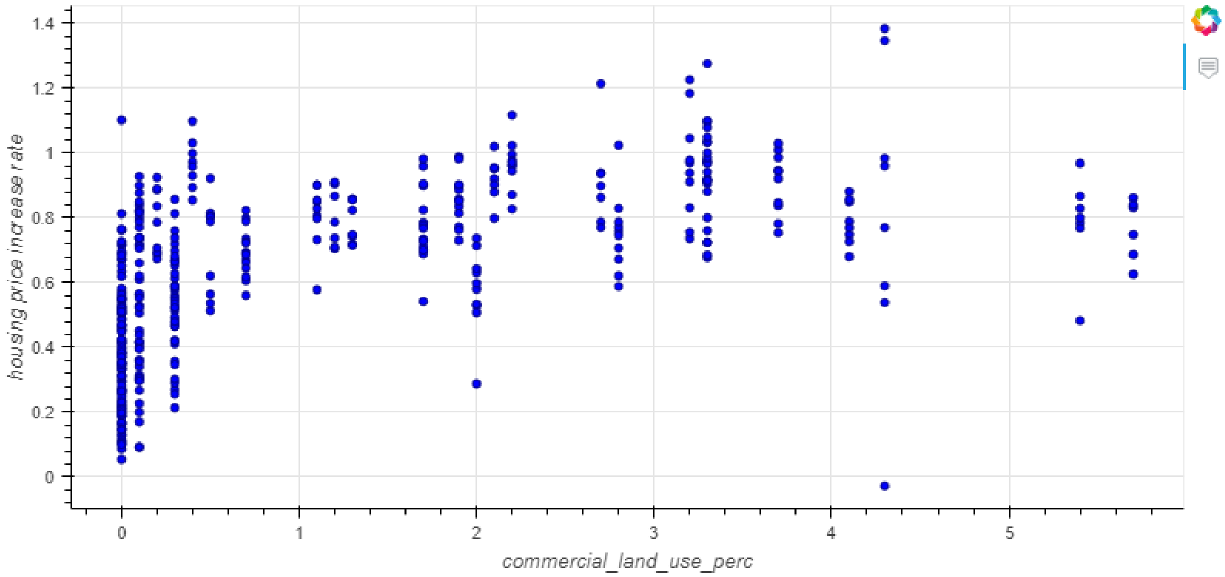
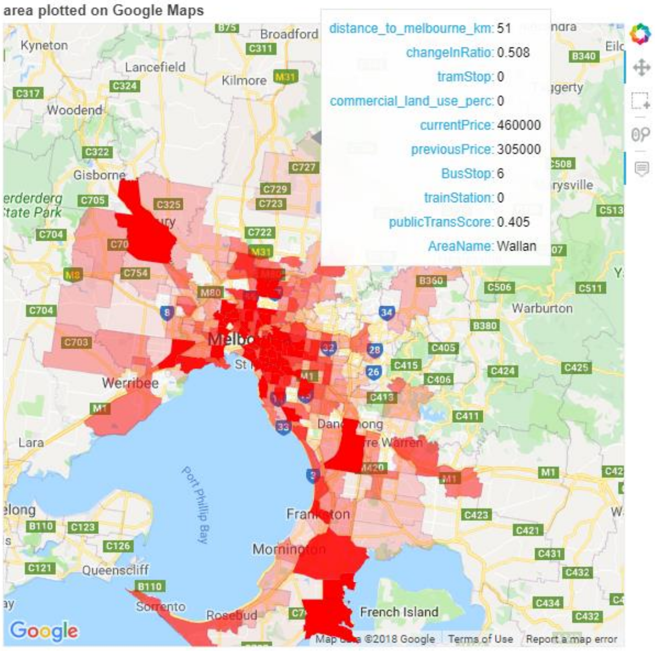


Figure 13: Scatter plot of the Commercial land usage

Commercial land usage is the factor with the highest MIC. Similar to the percentage of the born overseas population, there are different patterns for the inner area and outer area. As shown in figure 12, inner areas are with a higher percentage of commercial land usage, while outer areas are with lower usage. In figure 13 , the house price increase rate for areas with commercial land usage lower than 1%, are ranged from 20% to 100%, while the increase rate for areas with commercial land usage higher than 1% are fluctuating in the range from 60% to 120%. Although commercial land usage and house price increase rate are not linearly related to each other, the fluctuation of house price increase rate depends on whether an area's commercial land usage higher than 1% or not. Besides, as shown in figure 12, an area's commercial land usage is also correlated with the area's distance to CBD. The difference in house price increase rate can also be explained by the distance to CBD. Thus, it is not clear how much of the growth in house price is actually associated with commercial land usage rather than the distance to CBD.



distance_to_melbourne_km: 51
changeInRatio: 0.508
tramStop: 0
commercial_land_use_perc: 0
currentPrice: 460000
previousPrice: 305000
BusStop: 6
trainStation: 0
publicTransScore: 0.405
AreaName: Wallan

Figure 14: Distribution of Public transportation

Public transportation has a result of the high value of R^2, PCC and MIC. As shown in Figure14, we found that only areas distant to Melbourne within 50kms have public transportation (Tram, Train and Bus). Therefore, areas being divided into inner areas and outer areas again. Consistent with Figure 15, data in Figure 15 being divided into two parts. The house price increase rate of areas located in outer Melbourne ranges from 10% to 100%, with the mean value of 50%. However, house price increase rates for areas located in inner Melbourne fluctuate around 80%. Same as what happened in commercial land usage, the positive relationship between house price increase rate and public transportation scores only exist when we compare an inner area with an outer area. This association is weak when people are comparing an inner area with an inner area or an outer area with an outer area. Although statistical results show a strong correlation between house price increase rate and public transportation score, it is doubtful whether house price increase rate is highly related to public transportation level, or it is just related to the distance to CBD, as the association between house price increase rate and public transportation level does not exist in inner area.

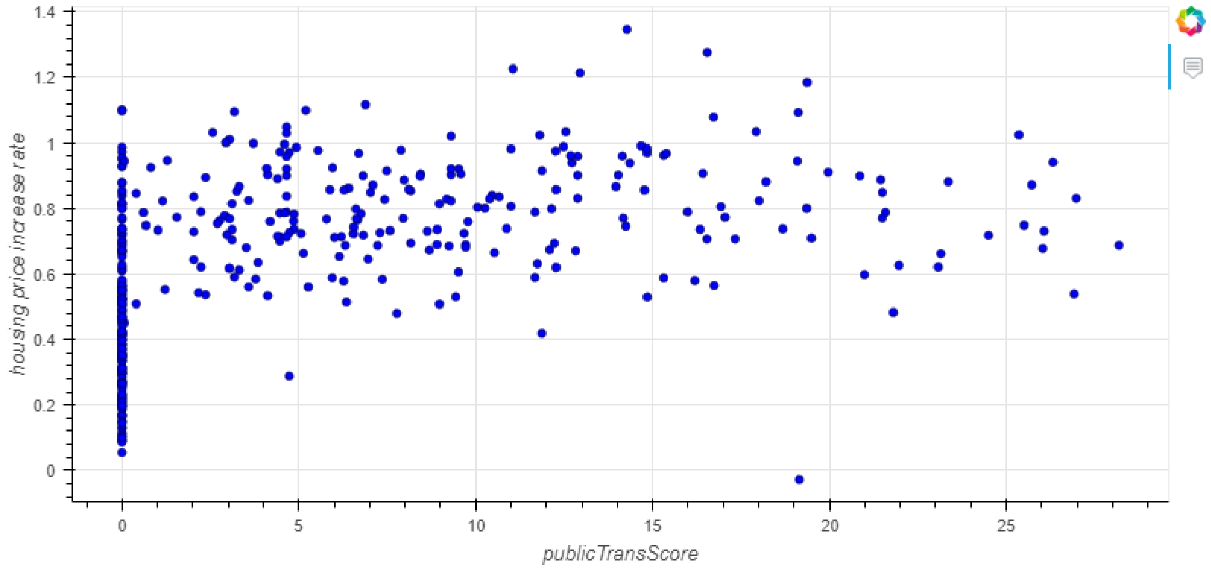


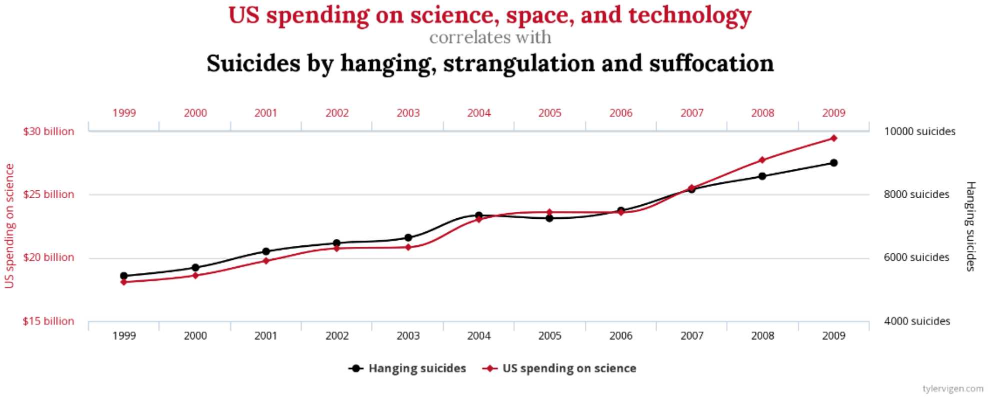
Figure 15: Scatter plot of the public transportation score

Limitation and improvements

As this study is based on Hedonic house price model, there are a few limitations resulted from this price model. First, the scope of HPM is limited. The price model only captures locational, and environmental factors that related to house price but does not fully explain all components in forming house price. External factors like interest rates, inflation rate, taxation that highly represent local economic environment are not well reflected from this model. Second, to use HPM, it requires prior knowledge on effects of locational amenities on house price, but unknown factors are normally not well handled. For instance, in this project, the potential positive and negative effect of locational and neighbourhood features like public transportation are previously known, better public transportation normally adds value to local properties. To use this model, digital datasets extracted from the AURIN platform is not enough. It also requires researchers with extra information on the potential effects of locational amenities on house price. Third, this model is mainly used to estimate people's buying behaviour and their willingness to pay for commodities. If people are not conscious of potential consequence from a locational or neighbourhood factor to their living quality, then the house price model will not reflect the value of this amenity.

Besides, the regression analysis can only explain the association between two variables. It can not be used to conclude cause-and-effect relationships between two variables. Because correlations like spurious correlations that have a mathematical relationship are not usually related to each other. For example, shown in figure 16, Vigen (2015) found that US spending on science is highly positively related to Hanging suicides. Although those two variables are mathematically highly related to each other, it is not right to conclude a cause-and-effect relationship between Hanging suicides and US spending on science. Thus, regression analysis can only show how much variables are related to each other rather than causality.

Figure 16: Spurious Correlations between spending on Science and suicides rate



In this paper, only 7 locational features are involved in the study of area's house price increase rate. There are other relevant datasets available in AURIN platform. Results derived from this paper are limited by the number of features being considered. Thus, this paper can only provide an answer to the question like "Which feature is relatively significant among those seven factors?". Questions like "Which locational feature has the most significant impact on house price increase rate" can not be answered in this paper. To make the results of this paper more practical, it's necessary to retrieve all types of locational feature datasets available in AURIN and process a feature selection process in future research.

Most locational attributes datasets in the AURIN platform is not updated every year. It is necessary to assume that targeted features keep constant over 2010 to 2017. However, this assumption brings in errors, which reduces the accuracy of results in this paper. In future work, datasets whose features change frequently each year, like personal income, should be available each year. To meet this requirement, datasets from other resources should be explored.

Moreover, the targeted aggregation level in this paper is the SA2 level, which means all datasets used in this paper have to be in SA2 aggregation level. In fact, some datasets used in this study are not aggregated at the SA2 level. Although those datasets are converted to SA2 level in aggregation conversion step, the accuracy of conversion highly depends on correspondence files provided by ABS.

Future Research Direction

In this paper, people’s willingness to purchase properties in an area has been estimated by Hedonic house price model. This estimation can only indirectly reflect people’s valuation on properties location. Besides, data involved in this paper are not available every year. Some datasets may be out-dated. To directly assess people’s willingness on buying properties in an area, big data analysis on Twitter data is a possible research direction. Twitter users’ feedbacks and estimations in each area contain rich information for providing insight on people’s buying behaviour in the real estate industry. Natural language processing technology like sentiment analysis can be used to extract relevant information from tweets data.

Conclusion

With the aid of the Jupyter notebook and AURIN platform, the state of the current house market in Victoria has been revealed in this paper. Houses located in inner Melbourne and south-east Melbourne have the highest growth in house price over eight years from 2010 to 2017. Hedonic House price model is used to estimate the value of environmental factors on property price. Locational and neighbourhood attributes in areas aggregated at the SA2 level are examined by statistical methods in this paper. Using statistical results, scatter plots and data visualization on google maps, we are able to find that, in Victoria state, an area’s percentage of born overseas population, commercial land, public transportation and distance to CBD are highly related to the increase in an area’s house price, while attributes like weekly family income, crime rate and green space have less association with the growth in house price.

Reference:

Aurin.org.au. (2018). About AURIN | AURIN. Australian Urban Research Infrastructure Network. [online] Available at: https://aurin.org.au/about/ [Accessed 21 Oct. 2018

Australian Bureau of Statistics (2018). Australian Statistical Geography Standard (ASGS). [online] Available at: http://www.abs.gov.au/websitedbs/d3310114.nsf/home/australian+statistical+geography+standard+%28asgs%29 [Accessed 28 Oct. 2018].

BRAITHWAITE, J. (1975). PopuLation Growth and Crime. AUST. & N.Z. JOURNAL OF CRIMINOLOGY, 8(1), pp.57-61.

Gibbons, S. (2004). The Costs of Urban Property Crime*. The Economic Journal, 114(499), pp.F441-F463.

Havekes, E., Bader, M. and Krysan, M. (2015). Realizing Racial and Ethnic Neighborhood Preferences? Exploring the Mismatches Between What People Want, Where They Search, and Where They Live. Population Research and Policy Review, 35(1), pp.101-126.

Hayashi, Fumio (2000). Econometrics. Princeton University Press. ISBN 0-691-01018-8.

Herath, S. (2015). Modelling urban house prices using open data. Available at: https://www.be.unsw.edu.au/sites/default/files/upload/research/clusters/Modelling%20Urban%20House%20prices%20Using%20Open%20Data.pdf. [Accessed 21 October 2015].

Herath, S. K. & Maier, G. (2010). The hedonic price method in real estate and housing market research. A review of the literature.. Institute for Regional Development and Environment (pp. 1-21). Vienna, Austria: University of Economics and Business.

Jupyter.org. (2018). Project Jupyter. [online] Available at: http://jupyter.org/ [Accessed 21 Oct. 2018].

Linden, L. and Rockoff, J. (2008). Estimates of the Impact of Crime Risk on Property Values from Megan's Laws. American Economic Review, 98(3), pp.1103-1127.

Pohlman, Leitner, Dennis, W.A., & The (2017). A Comparison of Ordinary Least Squares and Logistic Regression.

Qi, Z. (2017). Analyzing Type-1 Diabetes and AURIN Data. Melbourne: University of Melbourne.

Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M. and Sabeti, P. (2011). Detecting Novel Associations in Large Data Sets. Science, 334(6062), pp.1518-1524.

Vigen, T. (2015). Spurious correlations. New York: Hachette.

Wang, D. J. (2017) GEOT1D. Melbourne: University of Melbourne.

Appendix:

Source code and demonstration

Github Address: https://github.com/Animalone/Data-Analytics-Project-on-Housing-Prices-

Video Demonstration address:

https://youtu.be/BC7oga3ll1I

Datasets

Following are locational and neighbourhood amenities datasets and attributes used in this study.

Family Weekly Income:

Dataset name: SA2-based_B26_Total_Family_Income__Weekly__by_Family_Composition_as_at_2011-08-11

Attribute Name: Total family weekly income (2011)

Crime Rate:

Dataset name: Local_Government_Area_LGA_profiles_data_2015_for_VIC

Attribute Name: total offences per 1,000 population (2015)

Public Transportation (Output Count Points in Polygons from AURIN):

Dataset name: Public Transport Victoria (PTV) – Metro Tram Stops

Attribute Name: Tram Stop

Dataset name: Public Transport Victoria (PTV) – Metro Bus Stops

Attribute Name: Bus Stop

Dataset name: Public Transport Victoria (PTV) – Train Station Platform

Attribute Name: Train Station

Green Space (Output Count Points in Polygons from AURIN):

Dataset name: PSMA Greenspace (Point) (August 2017)

Attribute Name: Green Space

Born Overseas Population:

Dataset name: Local_Government_Area_LGA_profiles_data_2015_for_VIC

Attribute Name: Percentage of Population Born Overseas (2011)

Commercial Land:

Dataset name: Local_Government_Area_LGA_profiles_data_2015_for_VIC

Attribute Name: Percentage of Commercial Land Use (2016)

Distance to CDB:

Dataset name: Local_Government_Area_LGA_profiles_data_2015_for_VIC

Attribute Name: Distance to Melbourne (km) (2015)

List of Table

Table 1: Top 10 Areas with high increase rate

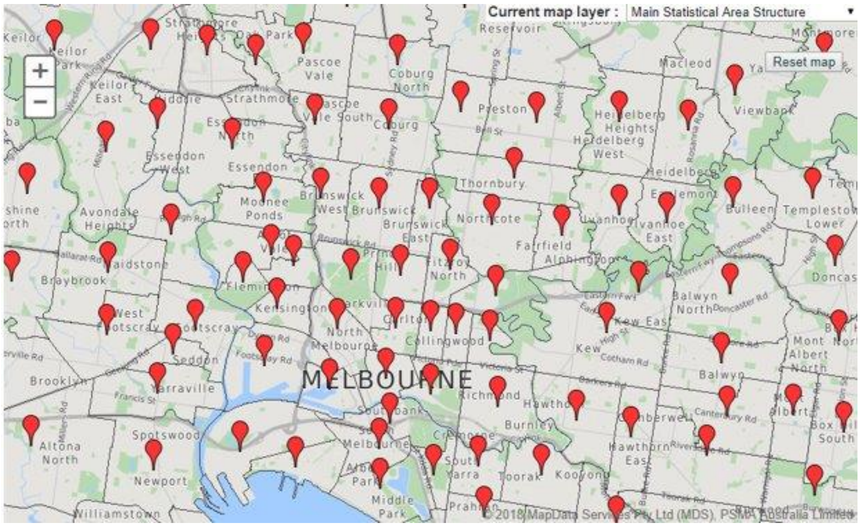
Area Name	Increase Rate	Previous Price	Current Price
East Melbourne	138.39%	1,336,000	3,185,000
Carlton	134.67%	620,000	1,455,000
Box Hill	127.58%	620,000	1,411,000
St Kilda	125.65%	672,500	1,517,500
Ashburton (Vic.)	122.61%	785,000	1,747,500
Alphington- Fairfield	121.39%	673,000	1,490,000
Hawthorn East	118.43%	1,057,500	2,310,000
South Yarra- East	115.06%	756,500	1,627,000
Hughesdale	111.69%	633,000	1,340,000
Ararat Region	110.19%	130,000	273,250

Table 2: Statistical Results for locational and neighborhood attributes

	Family weekly income	Born overseas population	Public transport	Commercial land	Crime rate	Distance to CBD	Green Space
R^2	0.774	0.881	0.509	0.565	0.7389	0.201	0.634
Pierson Correlation Coefficient	0.274	0.673	0.472	0.564	-0.201	-0.704	0.0389
Maximal information coefficient	0.263	0.577	0.514	0.590	0.393	0.574	0.190

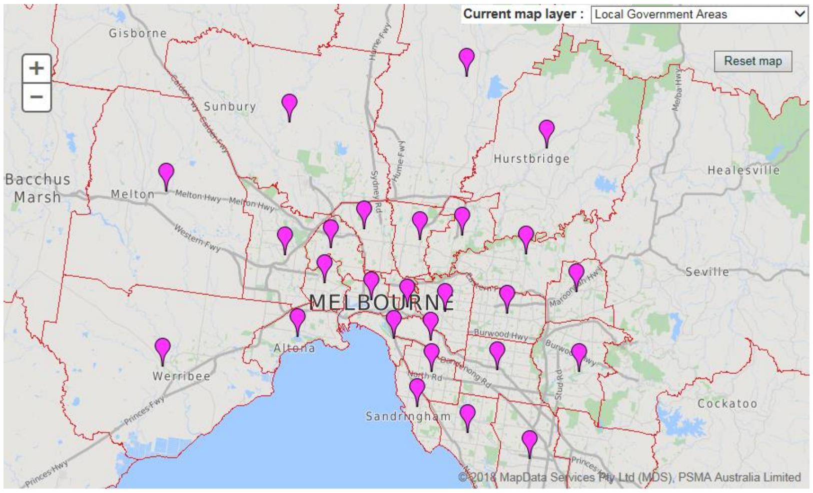
List of Figures

Figure 1: SA2 Aggregation Level



Current map layer : Main Statistical Area Structure

Figure 2: LGA Aggregation Level



Current map layer : Local Government Areas

Figure 3: Design of the house price analysis system

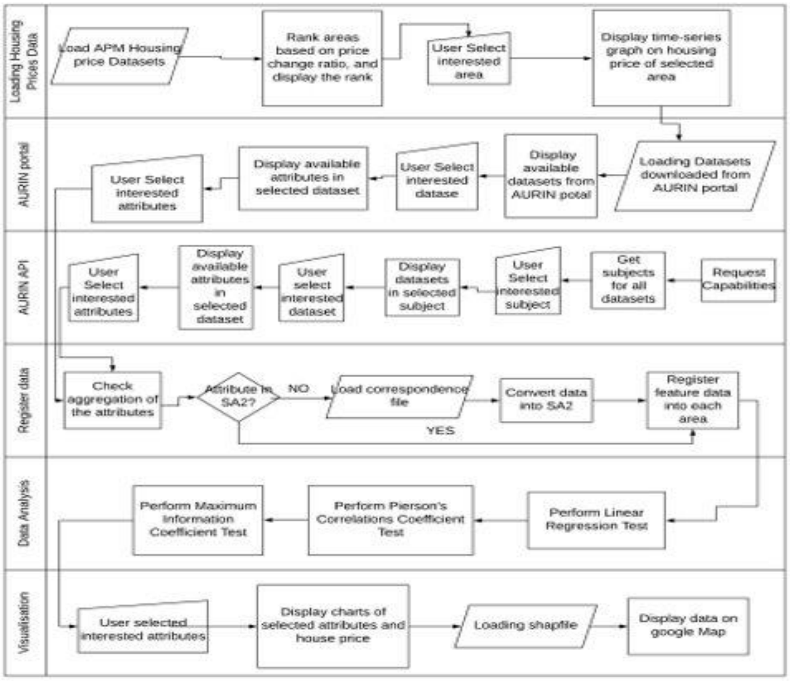

Loading Housing Prices Data
Load APM Housing price Datasets
Rank areas based on price change ratio, and display the rank
User Select interested area
Display time-series graph on housing price of selected area
AURIN portal
Loading Datasets downloaded from AURIN portal
Display available datasets from AURIN potal
User Select interested datase
Display available attributes in selected dataset
User Select interested attributes
AURIN API
Request Capabilities
Get subjects for all datasets
User Select interested subject
Display datasets in selected subject
User select interested dataset
Display available attributes in selected dataset
User Select interested attributes
Register data
Check aggregation of the attributes
Attribute in SA2?
NO
Load correspondence file
Convert data into SA2
Register feature data into each area
YES
Data Analysis
Perform Linear Regression Test
Perform Pierson's Correlations Coefficient Test
Perform Maximum Information Coefficient Test
Visualisation
User selected interested attributes
Display charts of selected attributes and house price
Loading shapfile
Display data on google Map

Figure 4: Median House Price Distribution in 2010

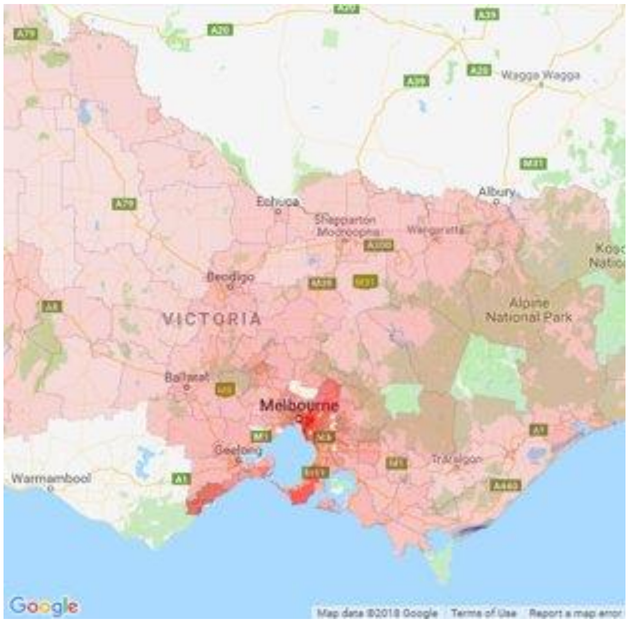


Figure 5: Distribution of House Price increase rate from 2010 to 2017

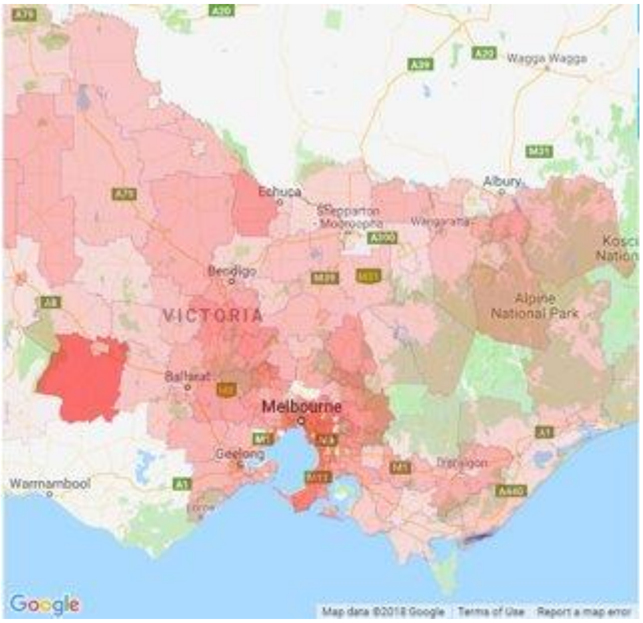


Figure 6: Scatter plot of the percentage of born overseas population

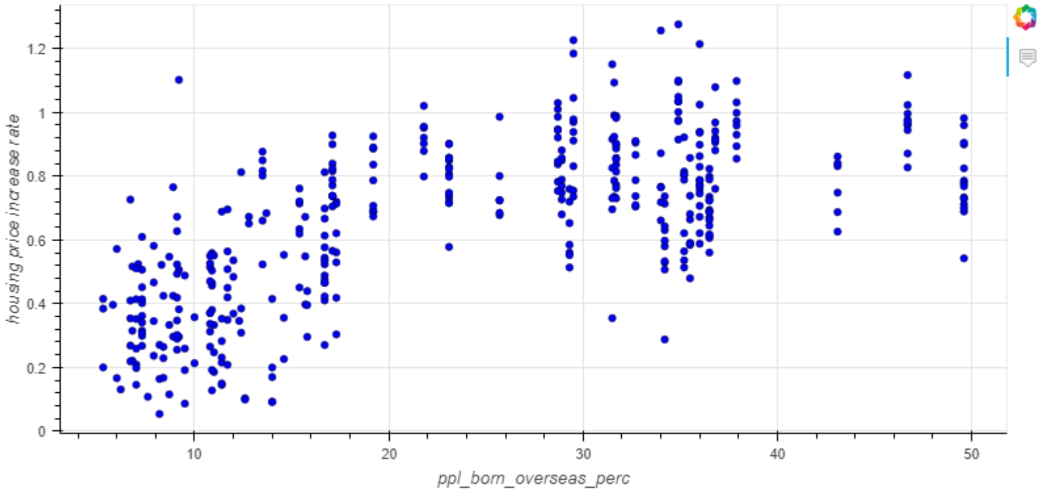
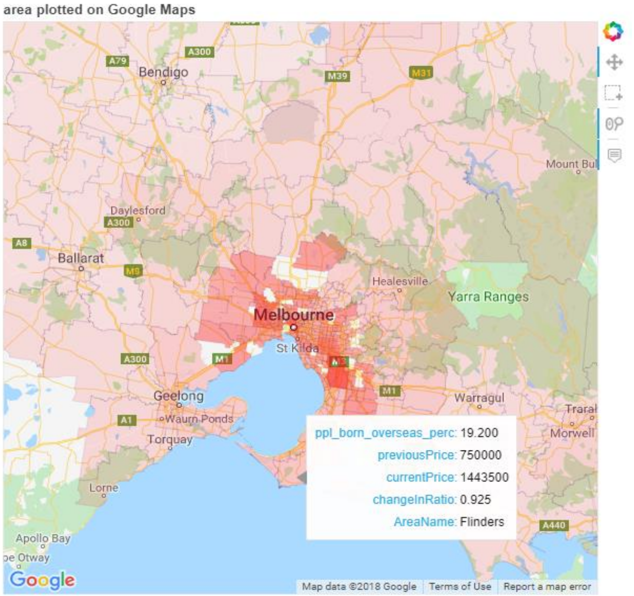


Figure 7: Distribution of the percentage of born overseas population



area plotted on Google Maps

ppl_born_overseas_perc: 19.200
previousPrice: 750000
currentPrice: 1443500
changeInRatio: 0.925
AreaName: Flinders

Figure 8: Scatter plot of the family weekly income

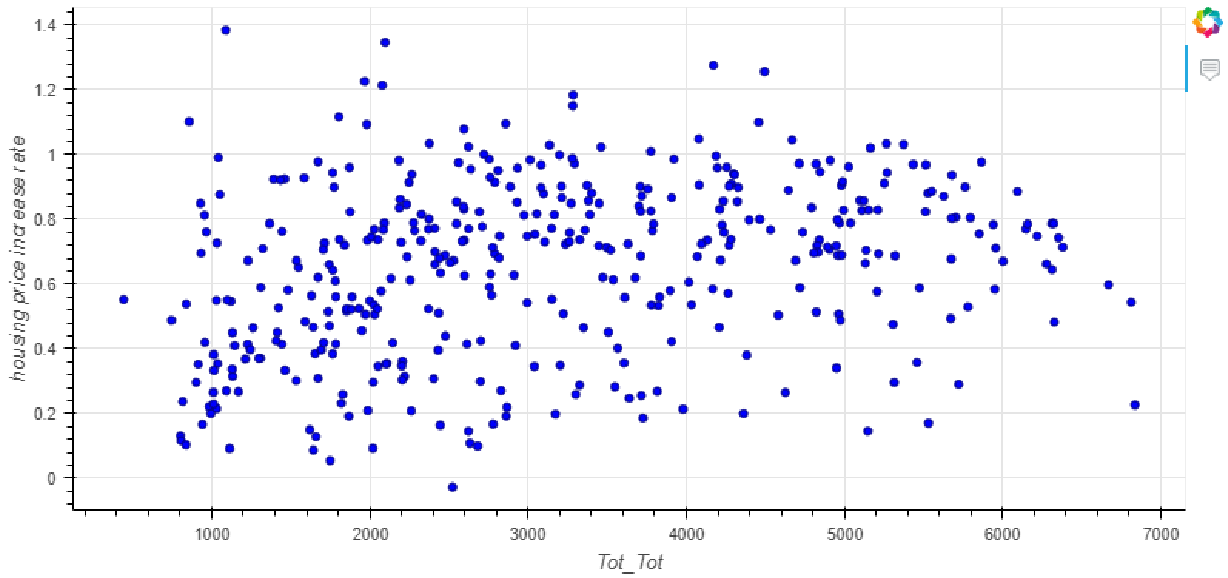


Figure 9: Scatter plot of the green spaces

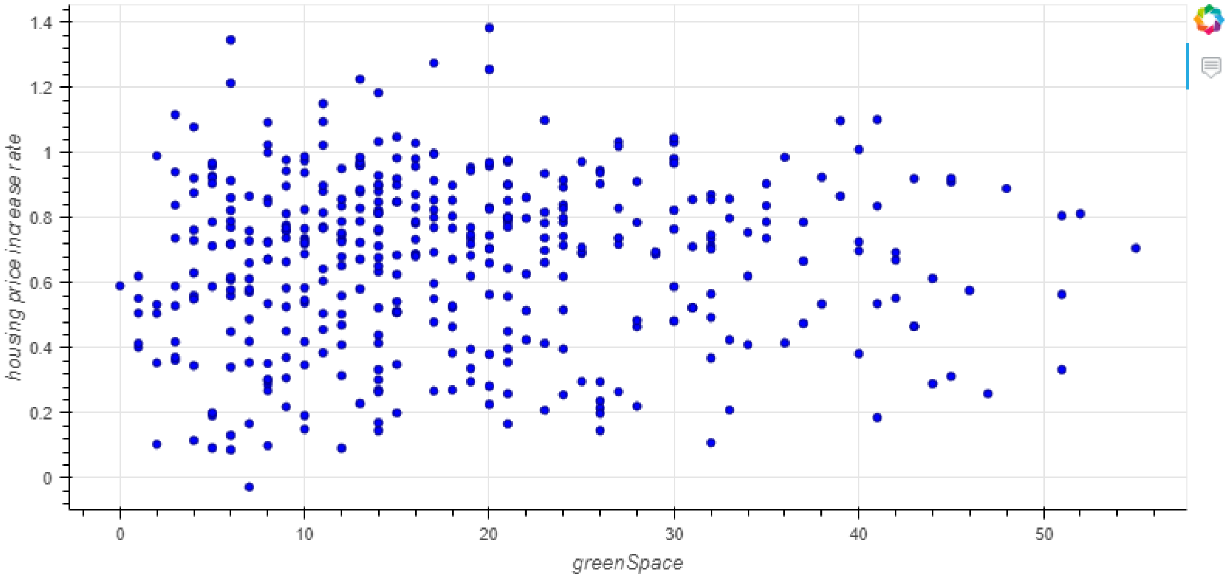


Figure 10: Scatter plot of the distance to Melbourne CBD

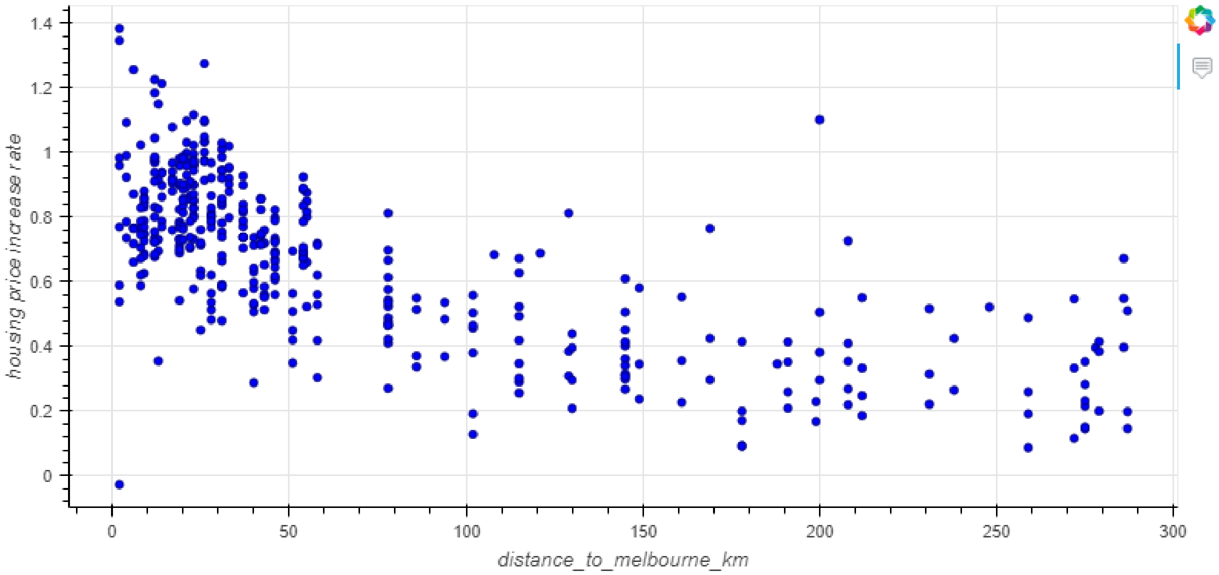


Figure 11: Scatter plot of the crime rate

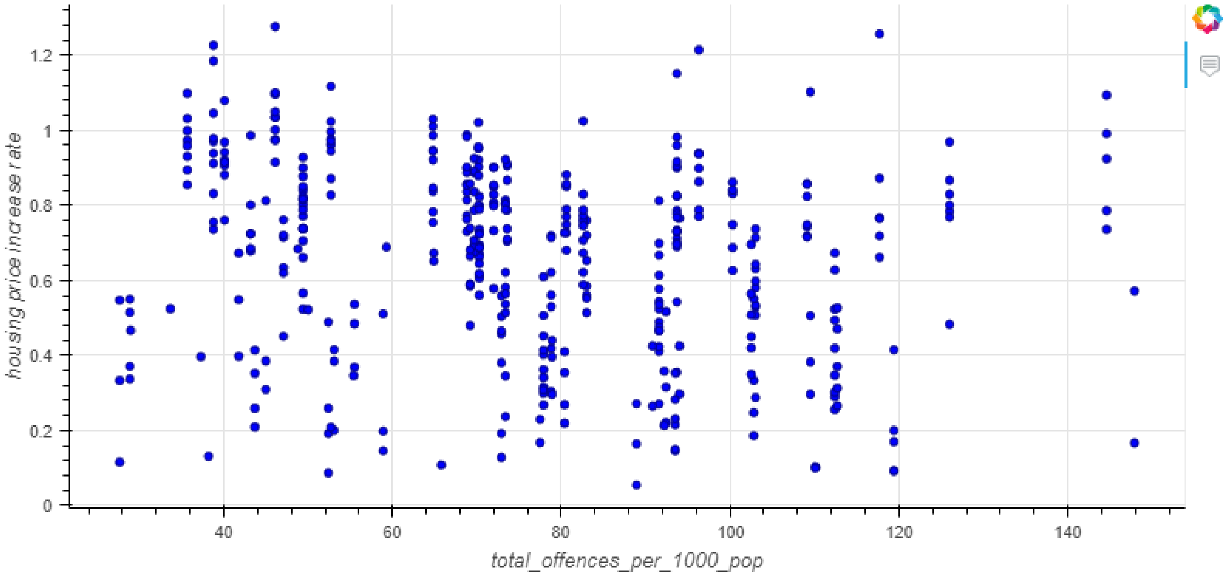


Figure 12: Distribution of Crime rate

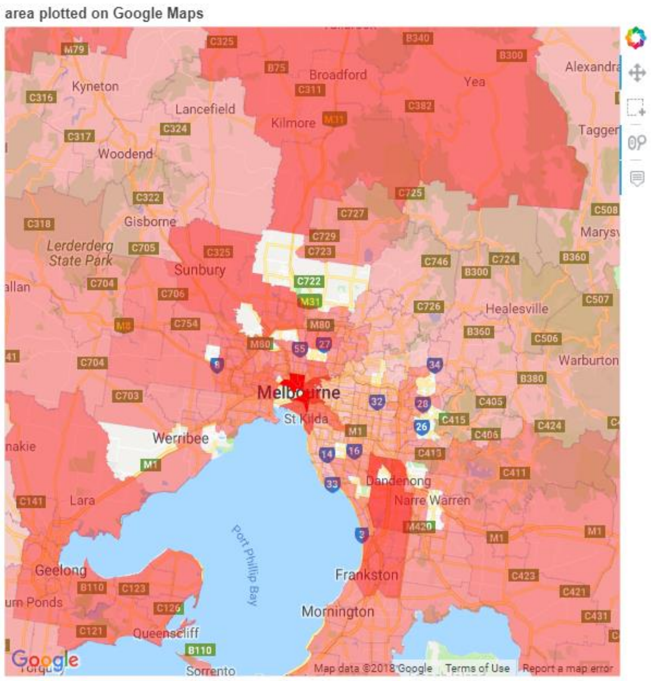


Figure 13: Distribution of Commercial land usage

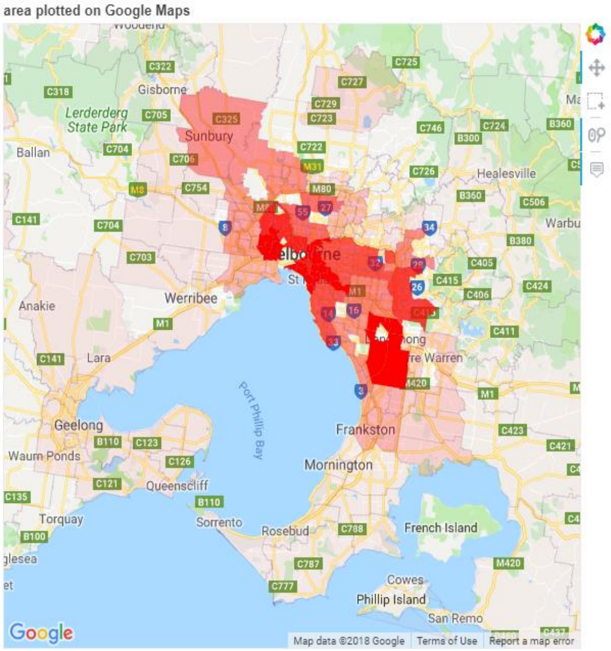
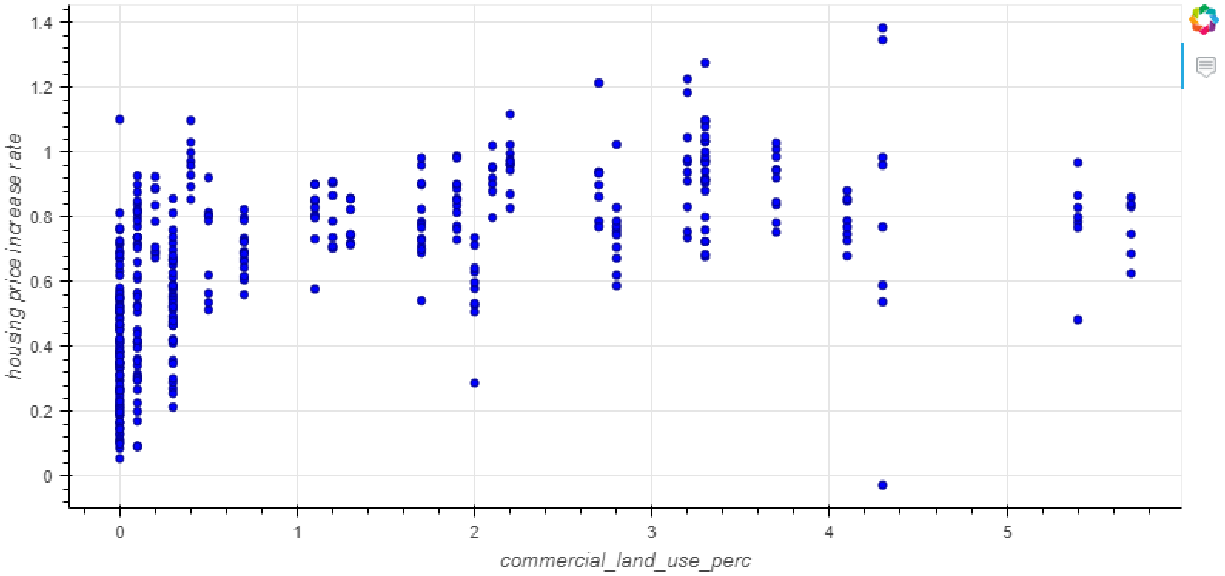


Figure 14: Scatter plot of the Commercial land usage


housing price increase rate
commercial_land_use_perc