31008-02: Data Mining Principles

23 May, 2018

Analysis of Melbourne Real Estate

Group 4

Cullen McNamee
Josep Nueno
WanQi (Angie) Tay

Executive Summary

Key Stakeholder:

- The analysis conducted in this project is intended on capturing key pricing forces in Australian real estate markets from 2016-2018.
- The intended audience is a real estate business person or house-flipper looking for pricing arbitrage opportunities of underpriced listings.

Goals of Analysis:

- The primary goal of the analysis is to understand which factors are most important in affecting selling price (i.e. location, number of rooms, year built), and ultimately using these to better predict potential selling price.
- In doing so, our stakeholders would be able to see a listing for a 2-bedroom home in Northern Victoria built in 1980 and determine an expected selling price. If the expected selling price is $100,000 AUD while the home is listed for $75,000 AUD this might signal a good buying opportunity.

Analytical Plan & Sampling

- We leverage the a publically available data source published to Kaggle called “Melbourne Housing Market”. The dataset includes approximately 30,000 properties of which 10,000 have complete information available.
- For our analysis we implement K-Means Clustering, Principal Components Analysis, Tree Classification Modelling, and Logistic Regression and compare the results -- along with geographical lat/long data.

Agenda

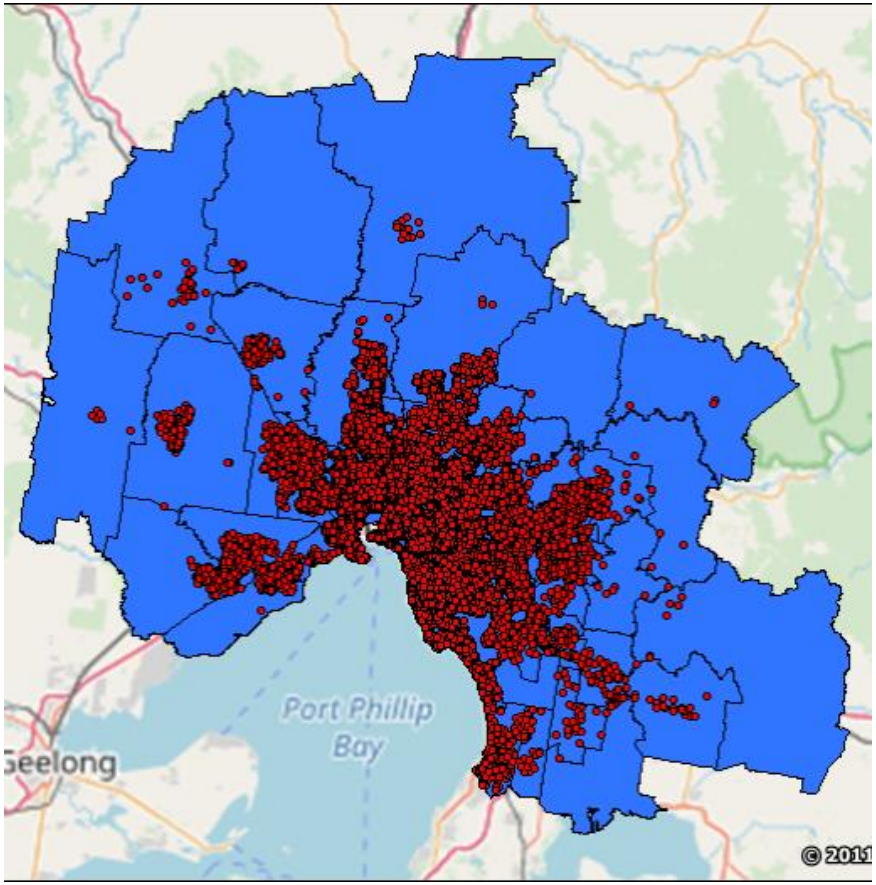
- Data overview
- Classifying Melbourne's properties for sale
- Comparison of different classification methods
- Pricing Melbourne's real estate properties
- Conclusions

Agenda

- **Data overview**
- **Classifying Melbourne's properties for sale**
- **Comparison of different classification methods**
- **Pricing Melbourne's real estate properties**
- **Conclusions**

Data

Real estate transactions in Greater Melbourne



Heatmap of population density in sample

Data Summary - Variables Analyzed

	Variable type	Description
1	Price	• **Sale price:** indicates the final sale price for all listed properties
2	Housing characteristics	• **Rooms:** total number of rooms • **Bedrooms.** • **Property area:** total property surface, built and unbuilt • **Built property area.**
3	Date	• **Date of sale:** when was the transaction finalized • **Date of construction:** year when the housing unit was built
4	Geographical information	• **Latitude/Longitude.** • **Address:** full address for the housing unit • **Community:** name of the administrative area where the unit is located • **Distance:** to nearest Central Business District

Data Cleaning

- One particularly beneficial feature of the dataset is its inclusion of Latitude and Longitude coordinates. While these variables are useful for clustering and plotting, we decided to conduct this as separate analysis and excluded the variables from the larger set when analyzing.

- Other fields excluded included: Address (unique variable), Seller, Postcode, Date, Suburb, and Council Area.

- Observations were removed where NAs were present in Price, Number of Bedrooms, Number of Bathrooms, Number of Car Parking Spots, Land-size, Building Area, and Year Built. Cost-benefit analysis was conducted and ultimately we concluded that the gain of dimensions outweighed the loss of observations from this filtering.

- Finally, for Logistic Regression we dummied Price as '1' if in the top-quartile >1,345,000 AUD and '0' otherwise.
  - *Note: This is an arbitrary threshold, but one designed to answer "how likely is the property to be among the highest price?"*

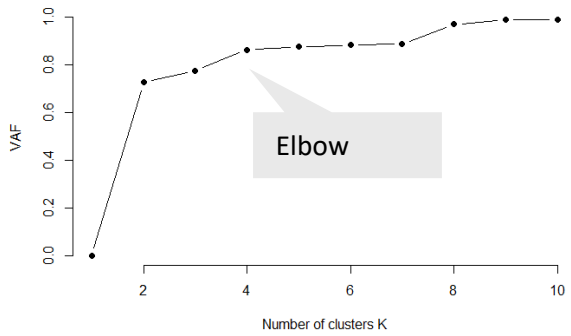
Agenda

- Data overview
- **Classifying Melbourne's properties for sale**
- Comparison of different classification methods
- Pricing Melbourne's real estate properties
- Conclusions

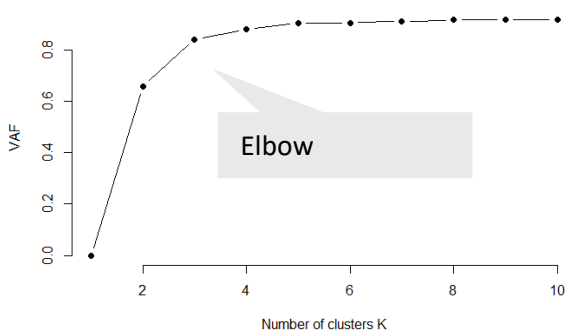
K-means clustering - Methodology

- Since K-means only accepts quantitative variables as inputs properties were segregated into the following subsets:
  - Type H: houses. Properties where the built area is different than the actual size of the plot
  - Type U: units. Apartments
  - Type T: townhouses. Houses built side by side, sometimes with a front yard. Built area is very close to the size of the plot
- We run the clustering exercise separately for each one of these 3 types

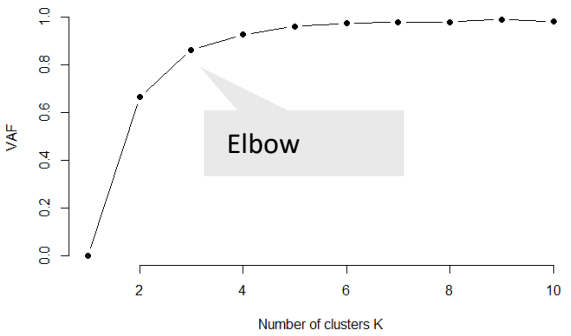
Type H – 4 clusters



Type U – 3 clusters

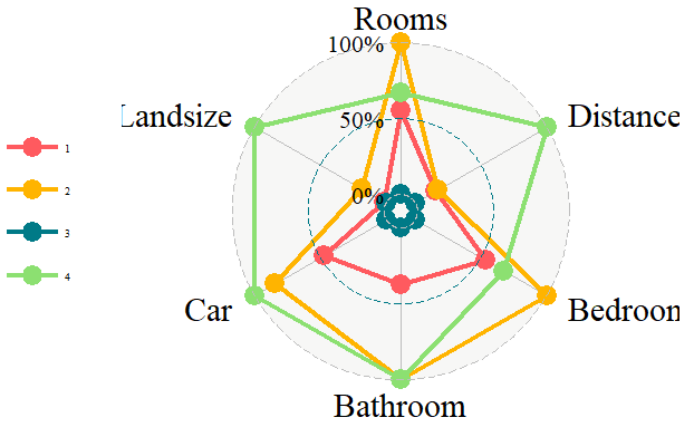


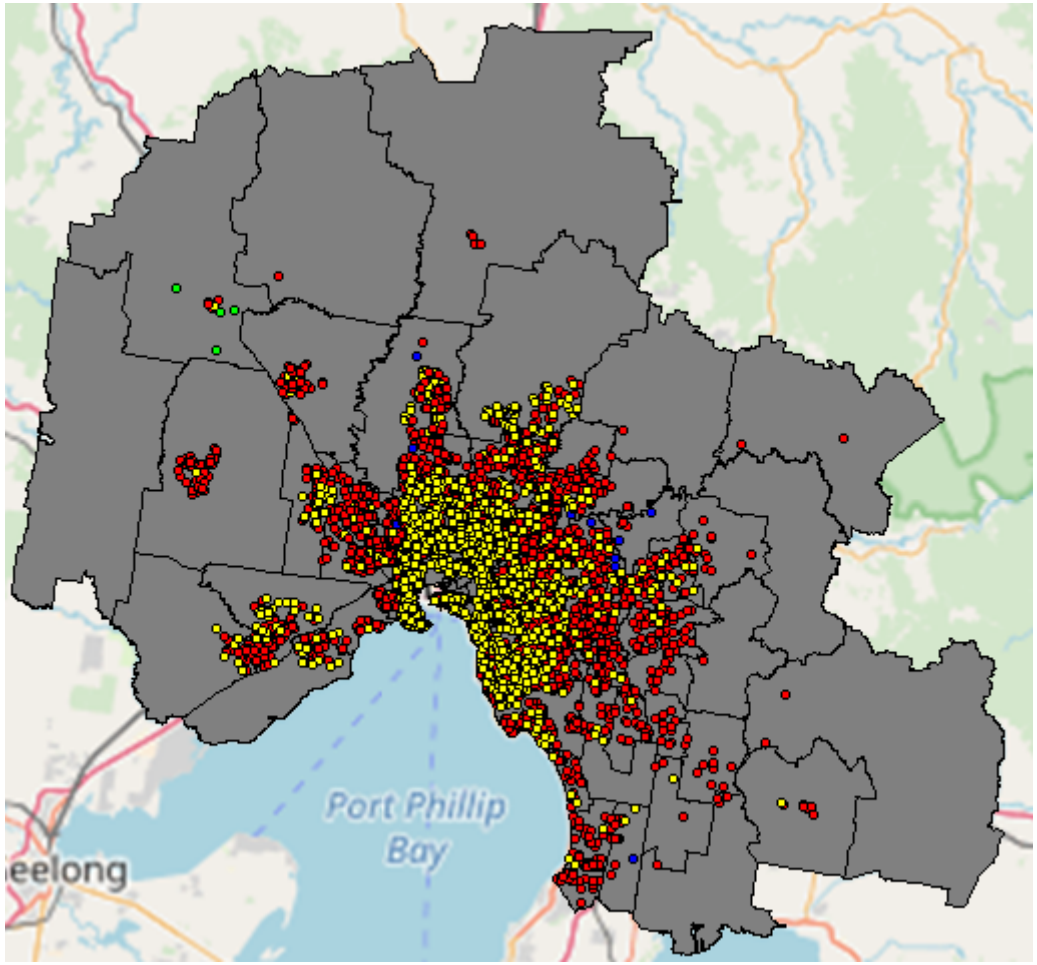
Type T – 3 clusters



K-means clustering – Results for H

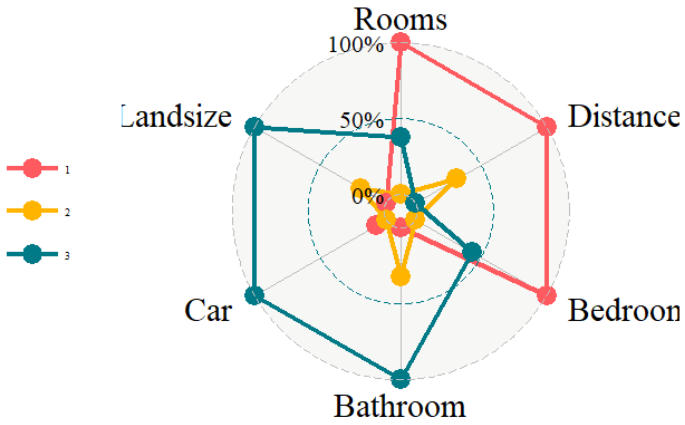
- For type H (houses) 4 clusters maximize the cluster to variance explained ratio. 3 of them can be characterized while the fourth aggregates a small number of outliers
  - Cluster 1: small houses in urban areas (smaller than the ones present in other clusters)
  - Cluster 2: large houses in urban areas
  - Cluster 3: Outliers
  - Cluster 4: large properties in the suburbs/rural areas

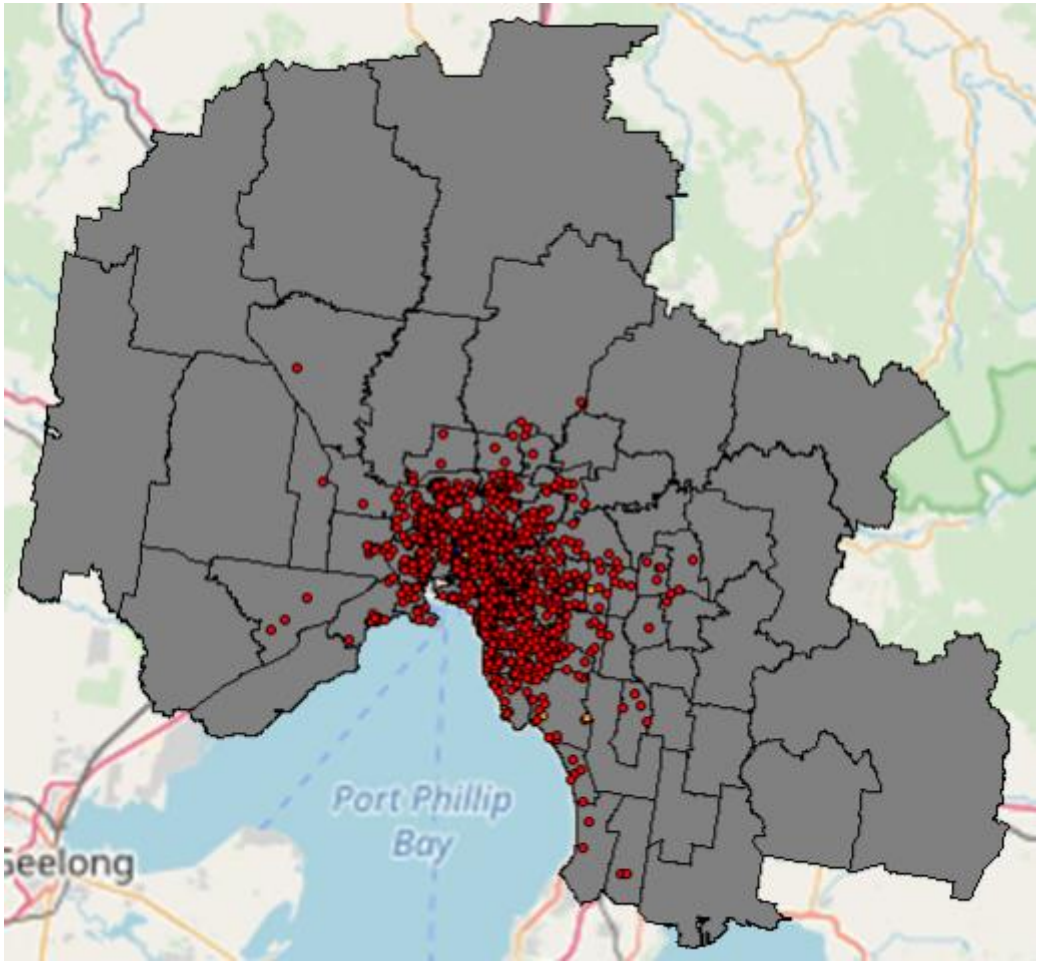




K-means clustering – Results for U

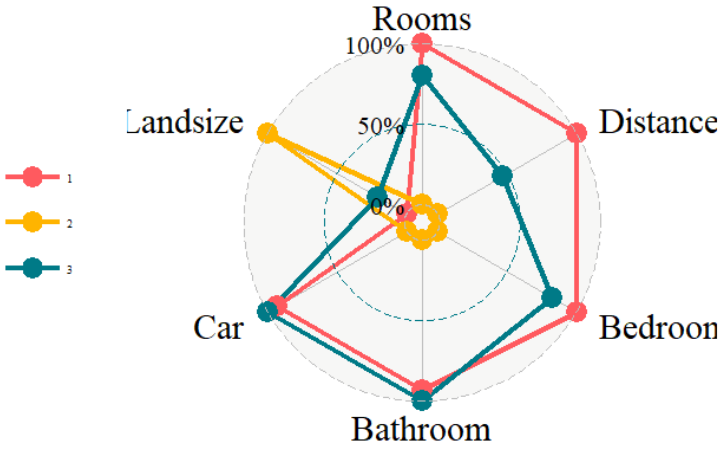
- For type U (units or apartments) we find the elbow at 3 clusters, although most members fall in cluster 1
  - Cluster 1: small apartments in all the metropolitan area
  - Cluster 2: medium apartments
  - Cluster 3: large apartments in the center

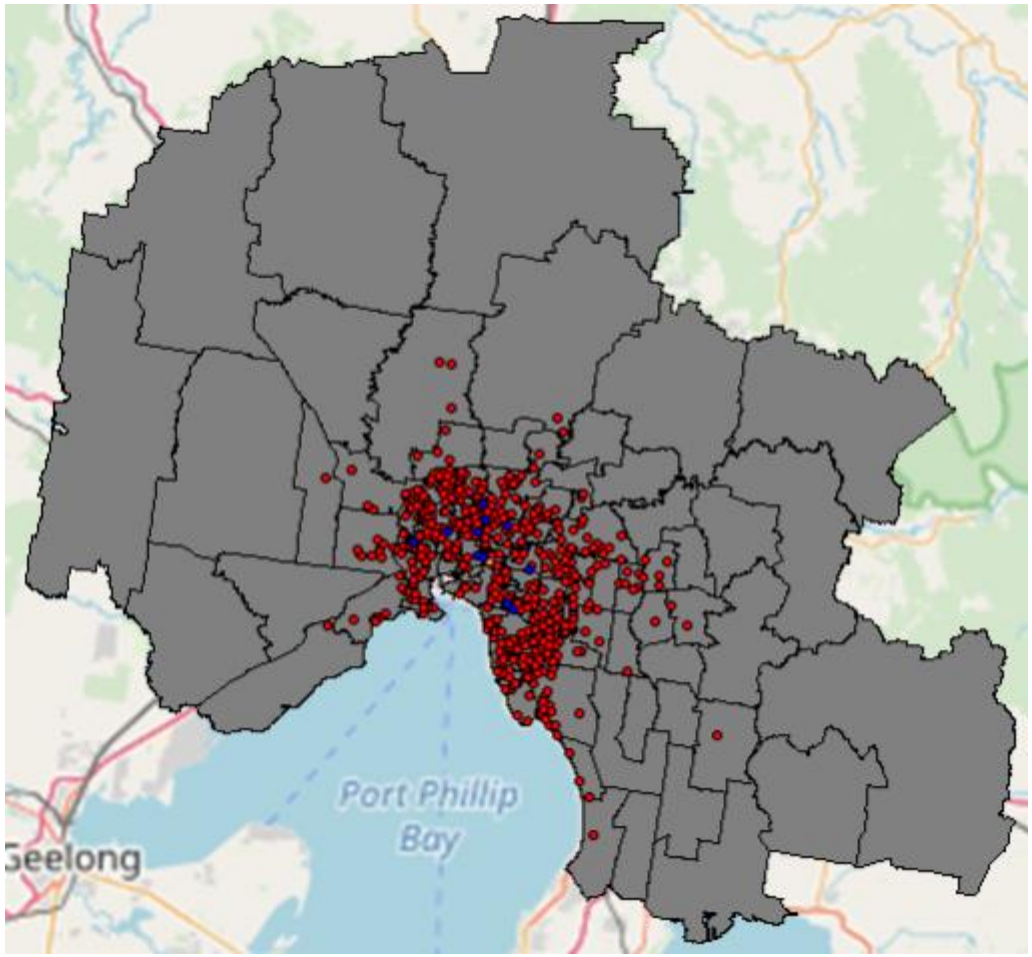




K-means clustering – Results for T

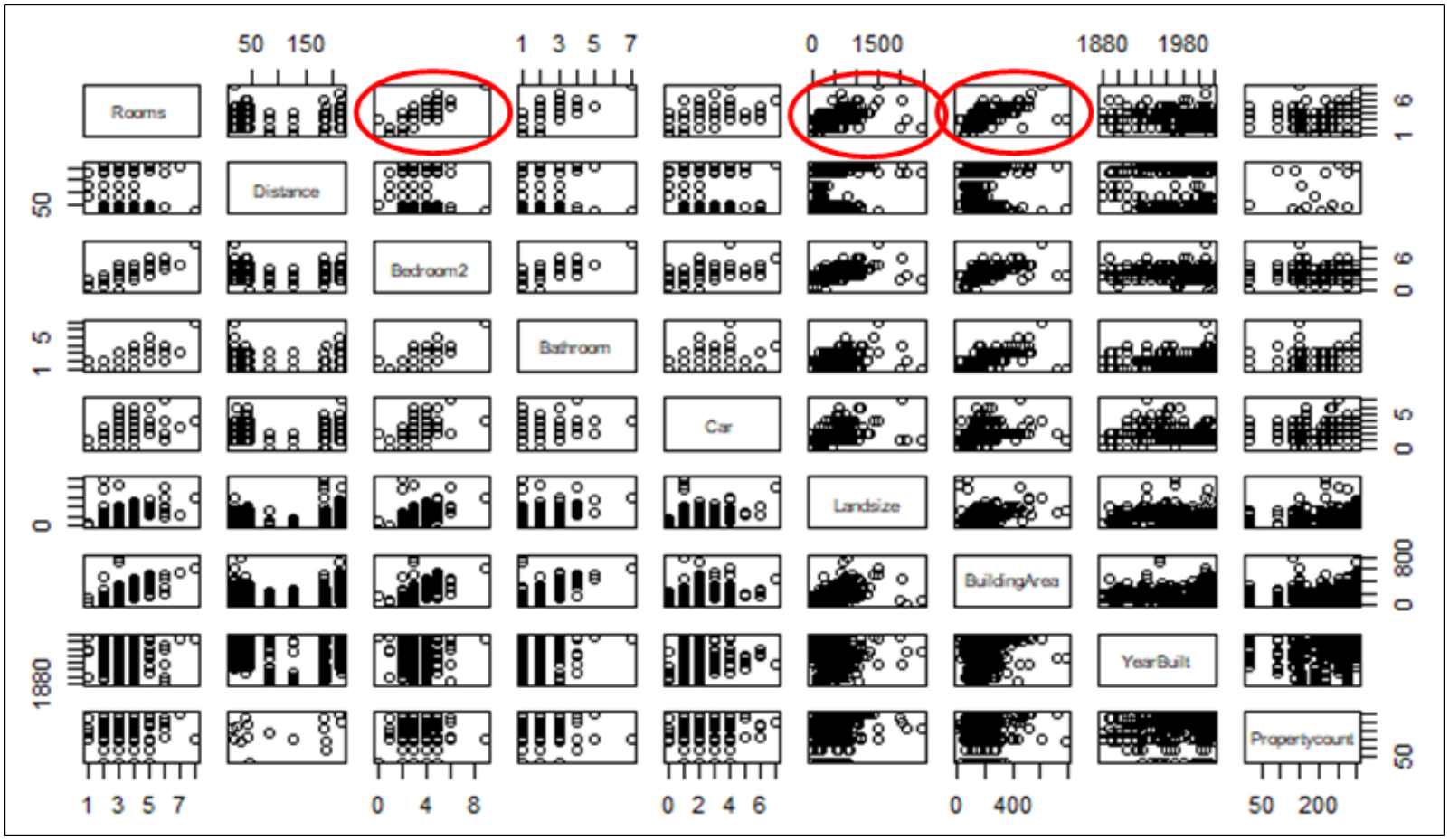
- For type T (townhouses) we find the elbow at 3 clusters, with most members assigned to cluster 1, although more balanced than for type U. Cluster 2 concentrates outliers
  - Cluster 1: townhouses in the metropolitan area
  - Cluster 2: outliers
  - Cluster 3: townhouses in City of Melbourne proper





PCA Clustering - Methodology

- Data has highly correlated variables
- Reduce dimensionality
- Convert highly correlated variables into linear uncorrelated variables

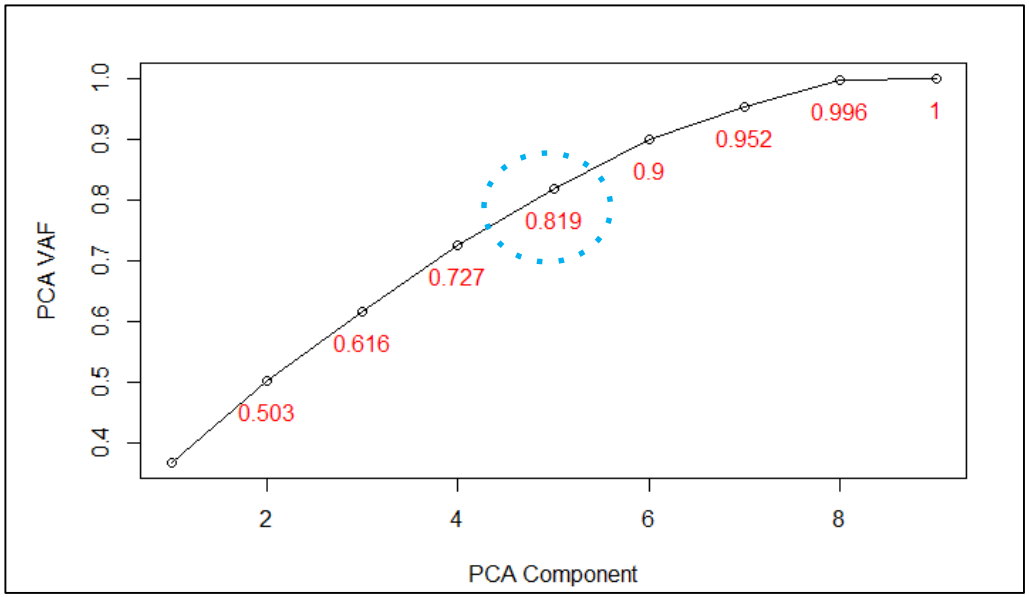


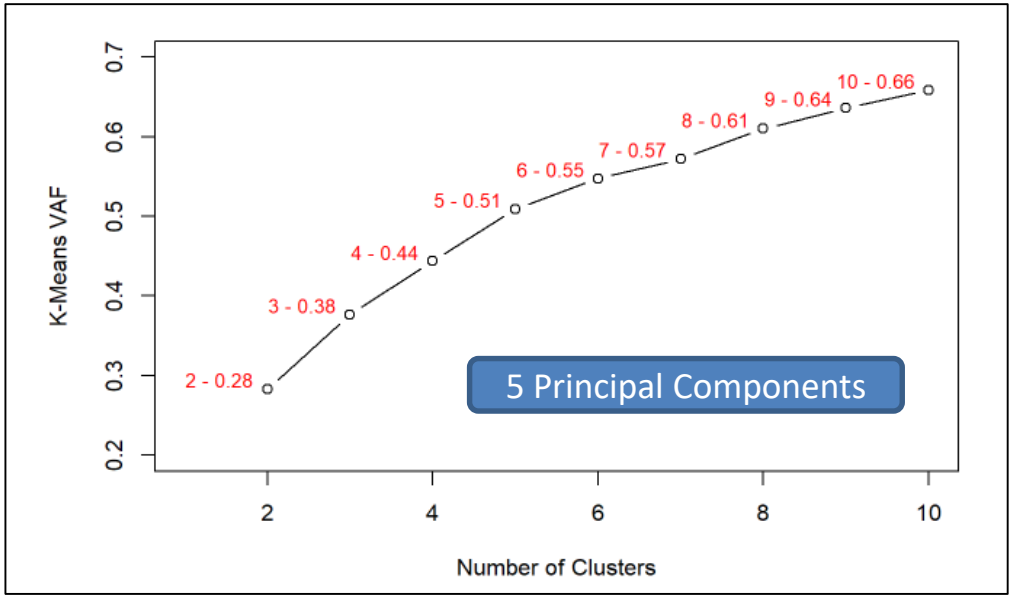
PCA Clustering - Methodology

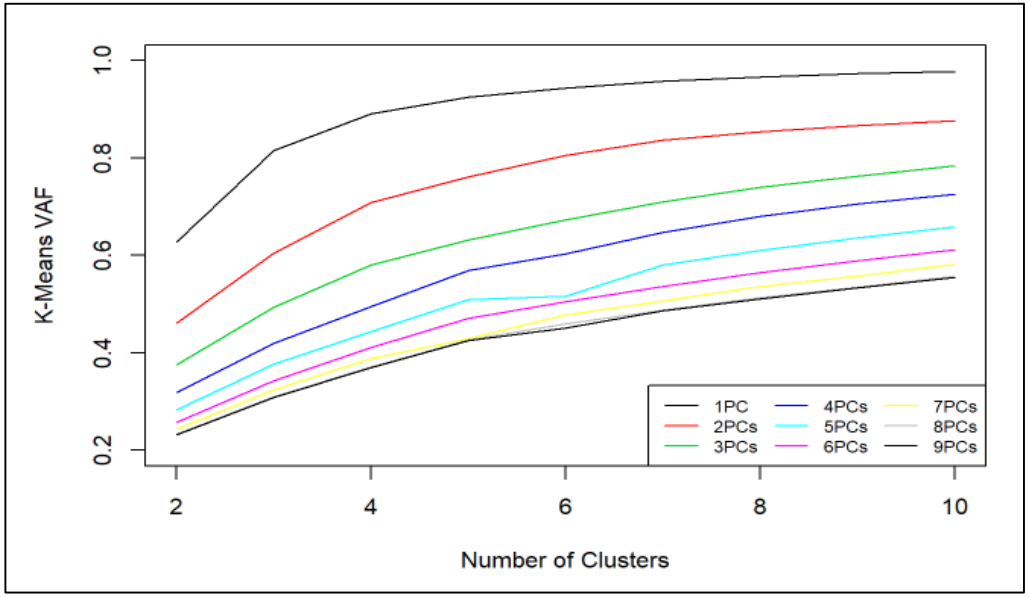
- PCA identifies correlated variables and groups them together under the same principal component.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Rooms	0.50	-0.15	-0.09	-0.04	0.10	0.07	0.41	0.14	-0.71
Distance	-0.08	-0.63	0.18	0.17	-0.70	-0.16	0.13	0.04	0.01
Bedroom2	0.50	-0.13	-0.09	-0.04	0.11	0.06	0.44	0.13	0.70
Bathroom	0.43	0.02	0.05	-0.15	-0.23	0.23	-0.23	-0.79	0.00
Car	0.31	0.18	0.13	0.21	0.06	-0.89	-0.11	-0.10	0.00
Land Size	0.10	0.16	0.20	0.91	0.04	0.32	0.00	-0.02	0.00
Building Area	0.43	-0.10	0.00	-0.06	-0.11	0.13	-0.71	0.52	0.02
Year Built	0.08	0.68	0.29	-0.19	-0.55	0.07	0.23	0.23	-0.01
Property Count	0.01	-0.19	0.90	-0.20	0.34	0.07	0.01	0.01	0.00

PCA and K-Means Clustering - Methodology







- PCA subspace is smaller than original space
- Searching in the PCA subspace is more effective and efficient.
- As dimensions reduced, the clustering results improved.
- How many principal components to retain?

PCA and K-Means Clustering - Training Results

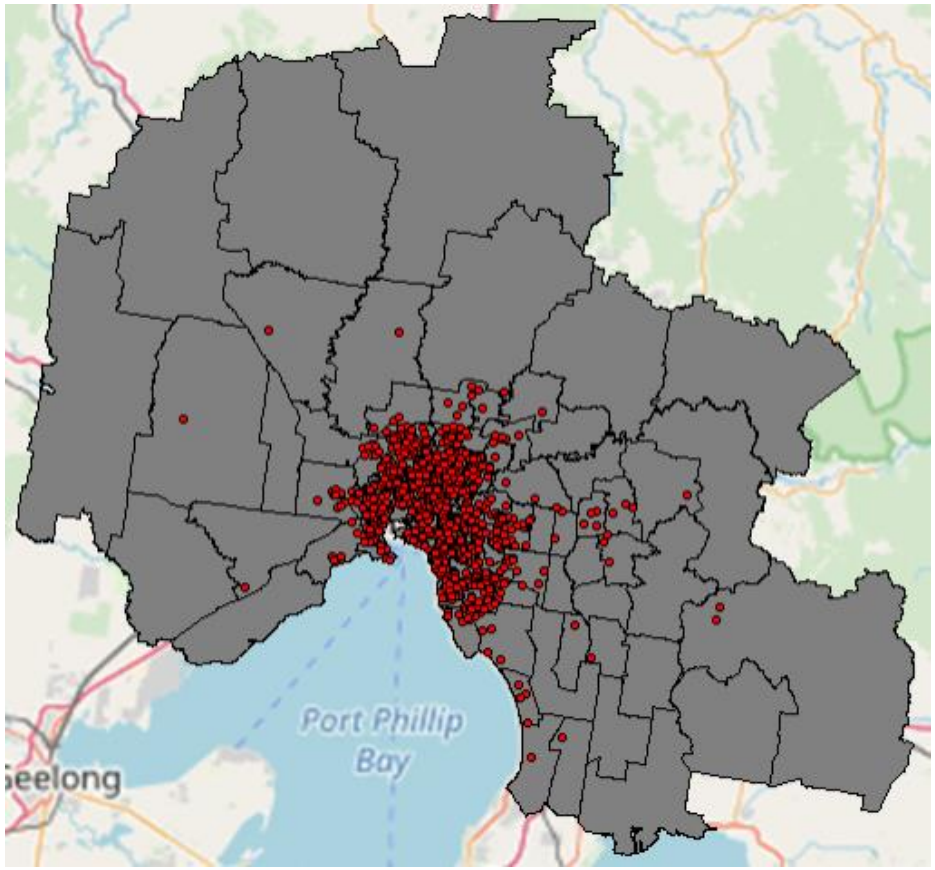
	Number of Housing Units in Cluster	Average Price ($AUD)	Average Land Area (m2)	Average Distance to CBD (km)
Middle Class Suburbs	2176	$ 880K	528	54
Big Property	24	$ 743K	12,364	121.4
High End Apartment	1707	$ 1,509K	652	94.7
Small Apartment	1647	$ 681K	286	136.9
Affluent Suburbs	1562	$ 1,346K	427	180
TOTAL	7,116	$ 1,087k	519	111

PCA and K-Means Clustering - Holdout Validation

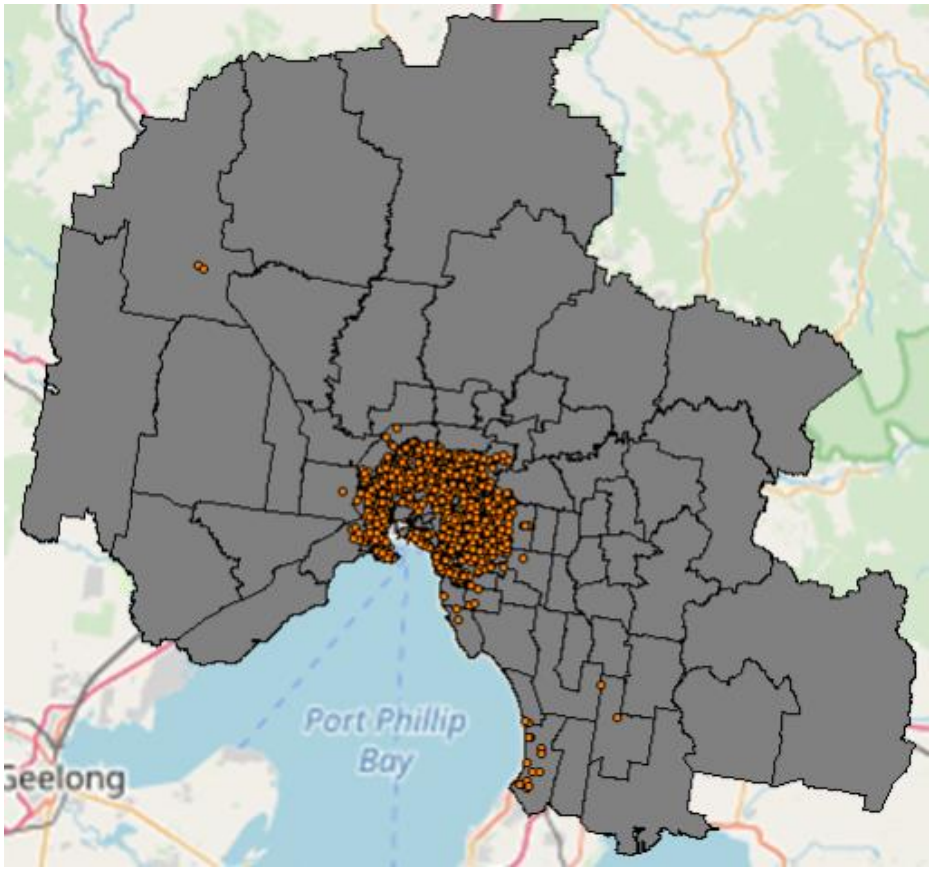
	Number of Housing Units in Cluster	Average Price ($AUD)	Average Land Area (m2)	Average Distance to CBD (km)
Middle Class Suburbs	571	$ 958K	580	52
Big Property	2	$ 725K	39900	163.5
High End Apartment	382	$ 1,595K	683	102.5
Small Apartment	445	$ 685K	283	132.3
Affluent Suburbs	379	$ 1,377K	423	182.7
TOTAL	1779	$ 1,115K	539	111

PCA and K-Means Clustering - Maps

Small Apartment

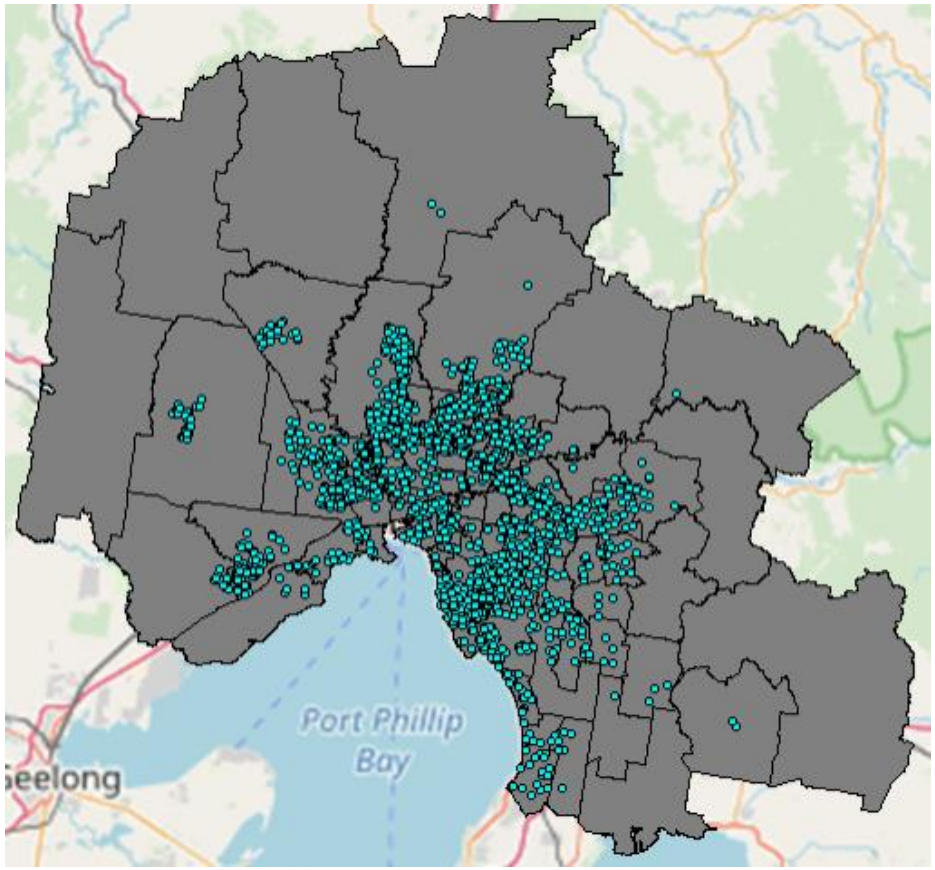


High End Apartment

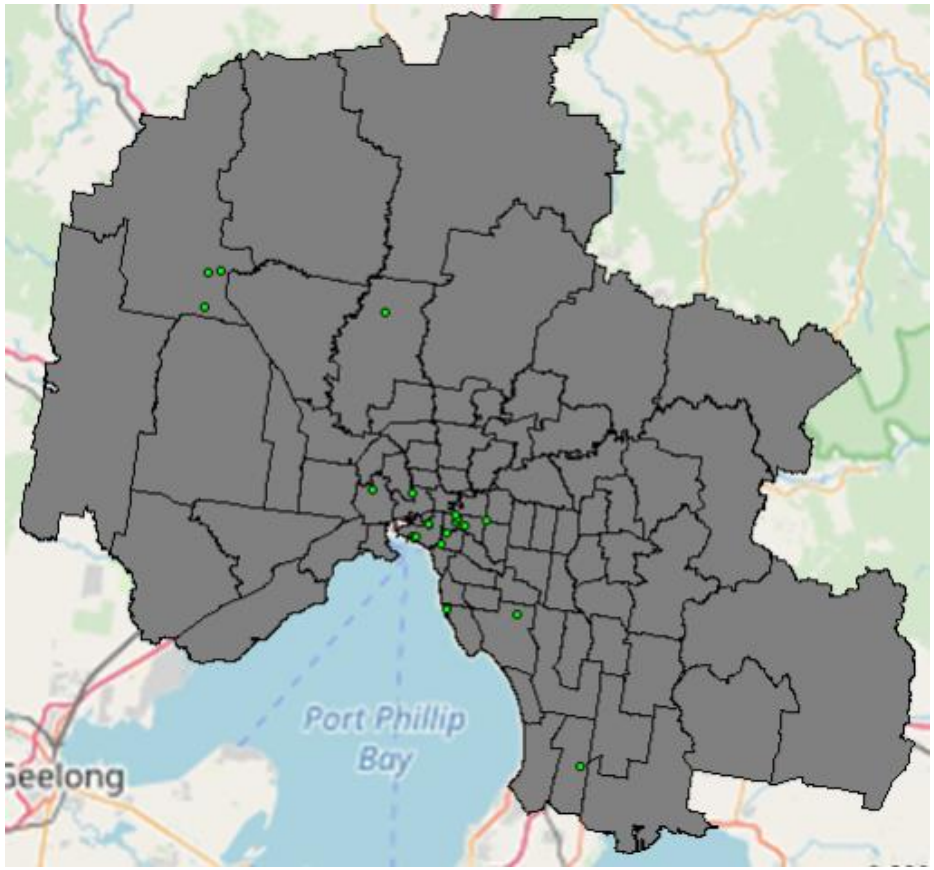


PCA and K-Means Clustering - Maps

Middle Class Suburbs

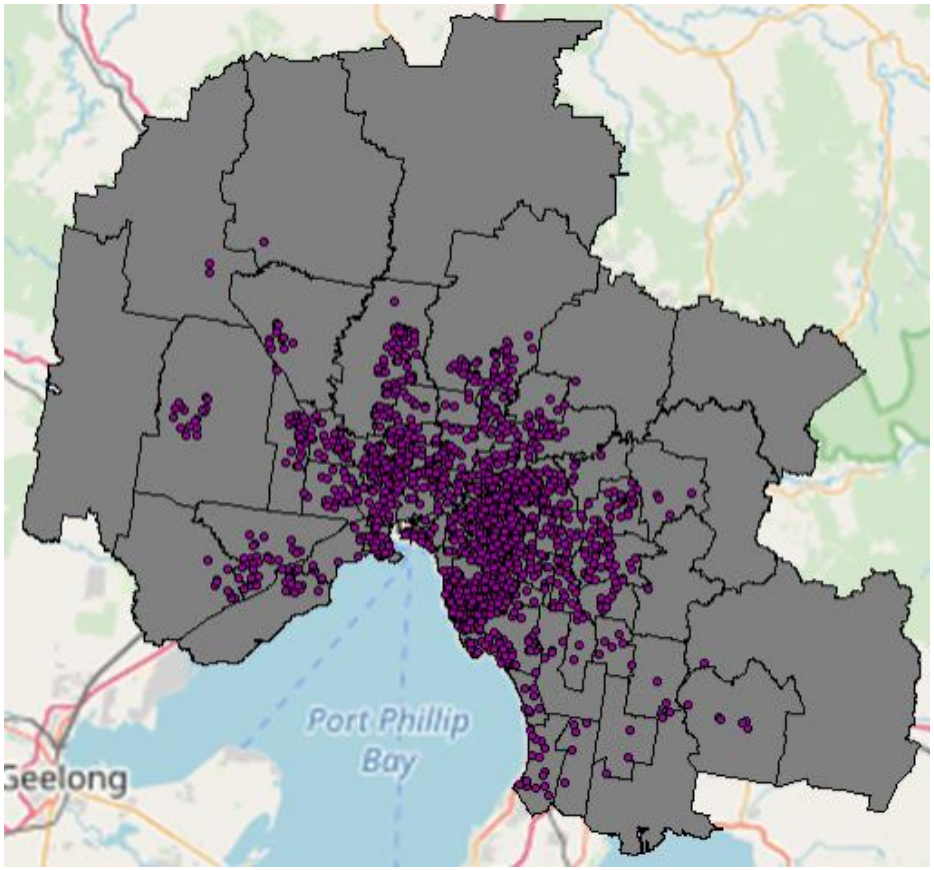


Big Property



PCA and K-Means Clustering - Maps

Affluent Suburbs



Agenda

- Data overview
- Classifying Melbourne's properties for sale
- **Comparison of different classification methods**
- Pricing Melbourne's real estate properties
- Conclusions

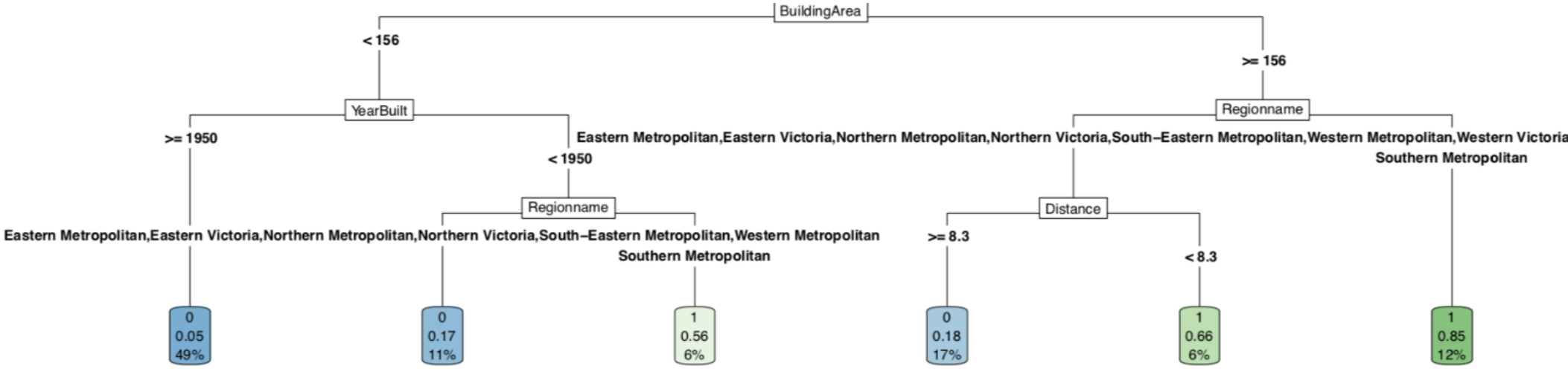
Comparison of classification methods

	Task	Description	K-means	PCA	Comments
1	Data	Limitations to the data used	✓	✓	Both methods restrict data to quantitative variables
2	Running time	Processing time required for each approach	✓	✓	Both approaches run in similar time, performance is not a bottleneck given the size of our data
3	Ease of implemen-tation	Data transformation required, steps needed to get the final results	✓	✓	K-means is easier to implement, but despite that the additional difficulty in PCA is not significant
4	Interpreta tion	Ease for characterizing the clusters based on the variables used	✗	✓	PCA complicates direct interpretation of the results.
5	Cluster differenti-ation	How distinct the resulting clusters are and can actionable data be extracted from them	✗	✓	PCA results in clusters that are more differentiated between them
6	Overall	Preferred method overall	✗	✓	PCA yields better results at a very low cost

Agenda

- Data overview
- Classifying Melbourne's properties for sale
- Comparison of different classification methods
- **Pricing Melbourne's real estate properties**
- Conclusions

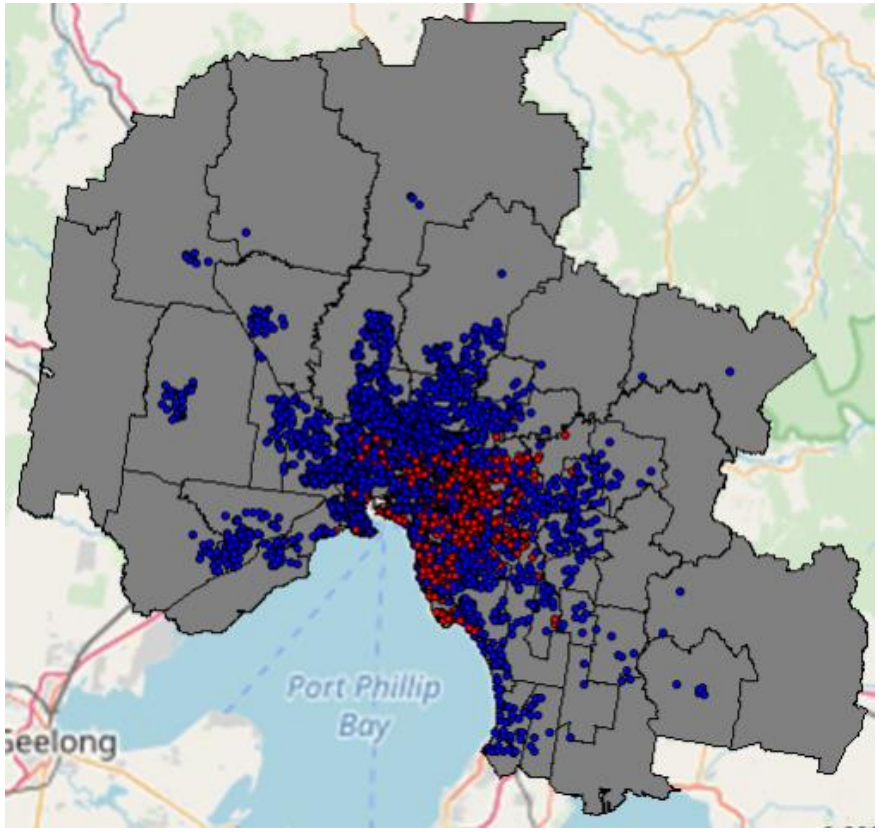
Tree classification - Results



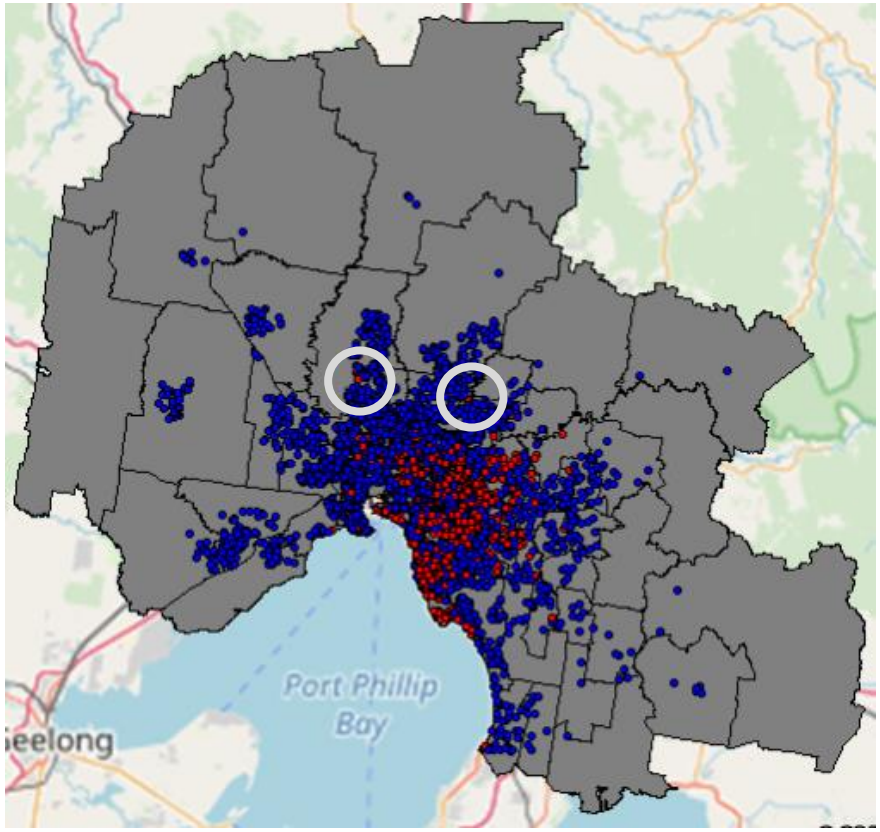
- As the tree above shows, the data can be grouped easily with simple intuition kept in tact.
- Here, 1 indicates that the property is likely to have sold for over 1,345,000 AUD (top-quartile). Properties belonging to this grouping can be described by one of the following feature sets:
  - "Small" (<156 square meters) | "Old" (built before 1950) | "Metro" (ex. Northern Metro)
  - "Large" (>156 square meters) | "Victoria" or "Metro" | "Close to CBD" (<8.3 km)
  - "Large" (>156 square meters) | "Metro" (ex. Northern Metro) or Western Victoria
- In Train 71% of top-quartile funds were correctly captured and 92% of non-top-quartile funds were correctly captured using this model (with minsplit = 500). In Holdout the captures were the same at 72% and 91%, respectively.
- Using a smaller minsplit of 30, we can increase the capture of top-quartile funds to 75% in Holdout.

Tree classification – Maps

Predicted price segment



Actual price segment



Logistic regression

- Same variables as the ones used for the tree classification model.
- After running an initial iteration for the full model a step process is run to select the best performing model.
- All variables are selected for the optimized model.

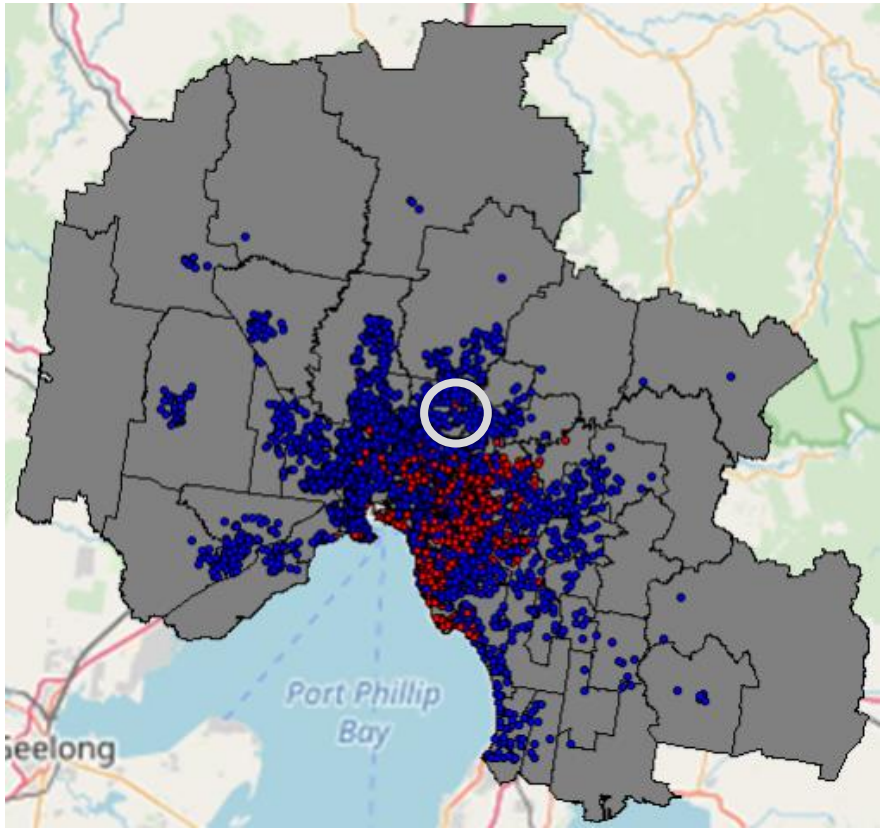
$$Price = Suburb + Rooms + Type + Method + Distance + Bathroom + Car + Landsize + YearBuilt$$

Training Conf. Matrix	Group 1	Group 0
Group 1	95%	5%
Group 0	17%	83%

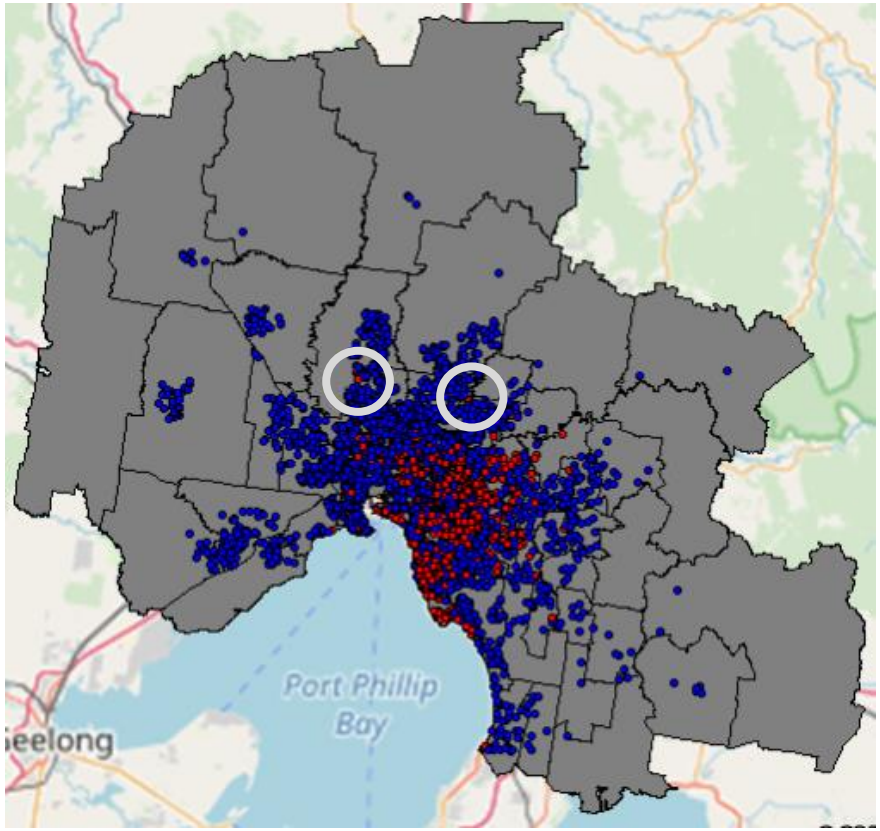
Holdout Conf. Matrix	Group 1	Group 0
Group 1	94%	6%
Group 0	28%	72%

Logistic classification – Maps

Predicted price segment



Actual price segment



Agenda

- Data overview
- Classifying Melbourne's properties for sale
- Comparison of different classification methods
- Pricing Melbourne's real estate properties
- **Conclusions**

Conclusions

- **Methodologies**:

Model	Pros	Cons
K-Means	Simple and flexible	Susceptible to outliers
PCA & K-Means	Makes K-Means clustering more effective and efficient	Trade off of choosing between loss of information and high VAF
Tree Model	Easy interpretation	Susceptible to outliers
Logistic Regression	Easily differentiate in price points including properties that are not easily be generalized	Harder to interpret contribution of each factor

Conclusions

Key Takeaways:

1. **Not surprisingly location and home size proved to be the most critical factors in determining selling price regardless of methodology.**

1. **While it is easy to use 'business sense' to categorize homes based on these two factors alone, the various approaches are helpful in drawing the line between borderline properties.**

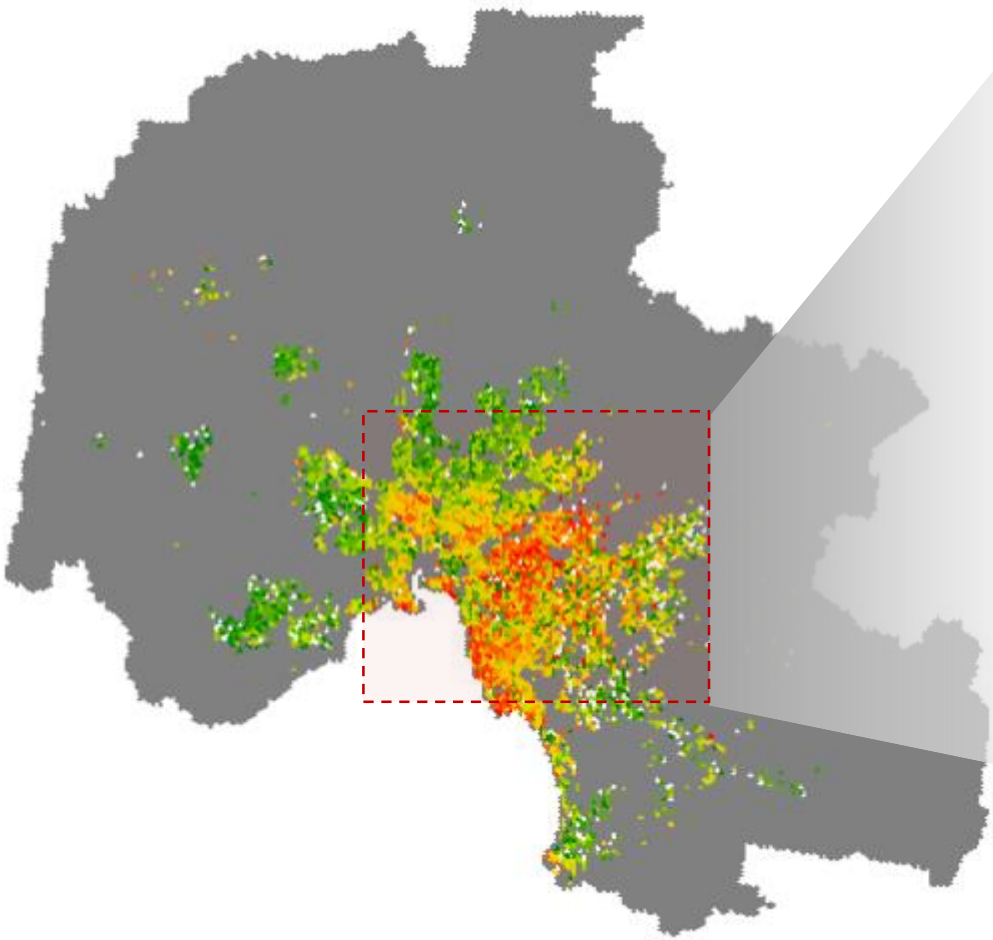
1. **Each factor observed could also be analyzed more closely on an independent basis. For example, older homes (evidenced by Year Built) were more likely to be highly priced in both Tree Classification & PCA.**

    - This may be a result of causal effects -- namely, that deteriorating homes are torn down/replaced instead of maintained, while high quality homes are better maintained.

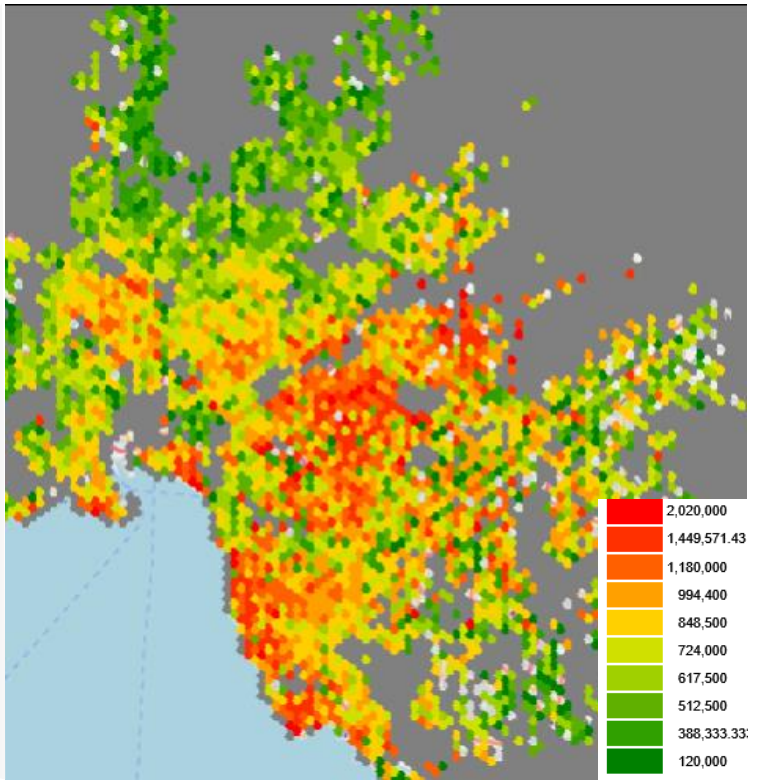
APPENDIX

Spatial distribution of price

Heatmap – Average price distribution

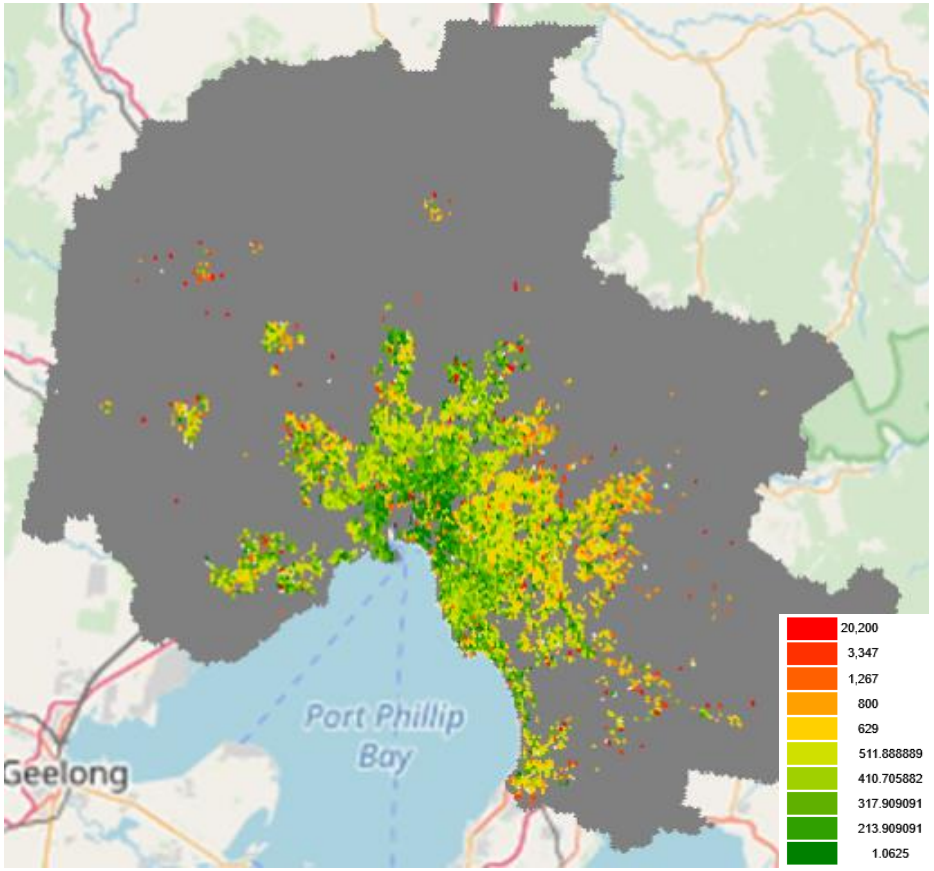


Heatmap – City of Melbourne zoom

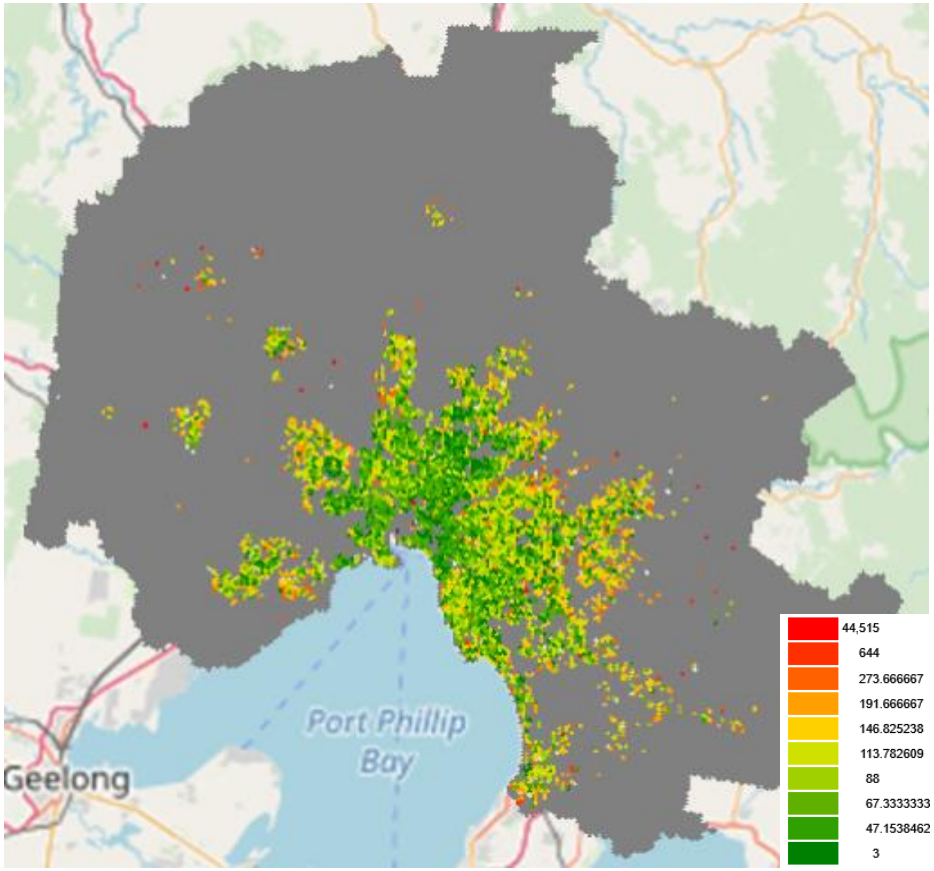


Spatial distribution of housing characteristics

Heatmap – Available land area



Heatmap – Built land area



Spatial distribution of date variables

Heatmap – Year built