Recipe for a House

MICHELLE LI

To Make and Sell a House

INGREDIENTS

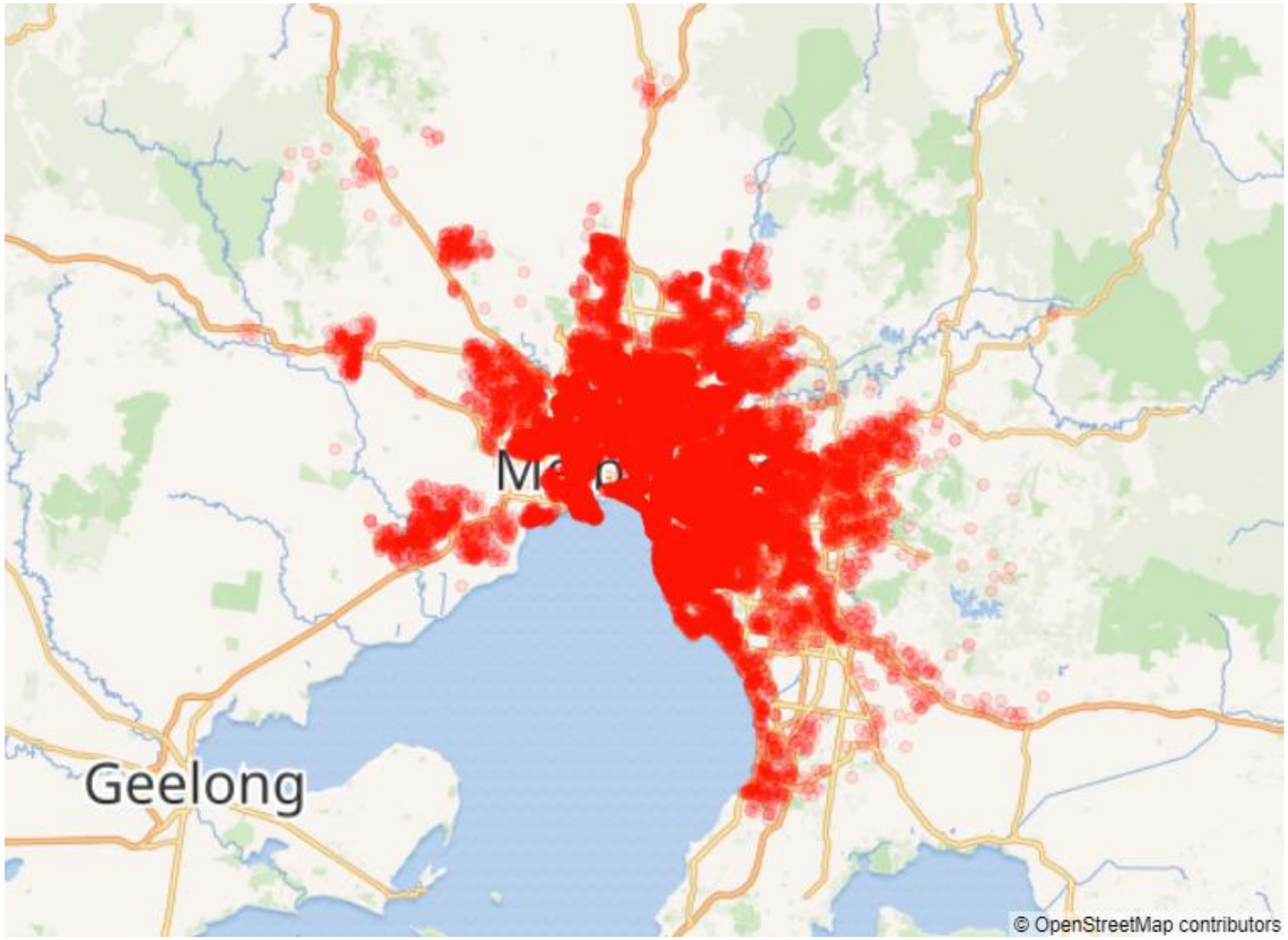
- ❑ Land
- ❑ Walls
  - ❑ Bedroom
  - ❑ Living room
  - ❑ Bathroom
  - ❑ Kitchen
- ❑ Parking space

PREPARATION

- ❑ Location
- ❑ Local building laws
- ❑ Real estate developer
- ❑ Real estate agent

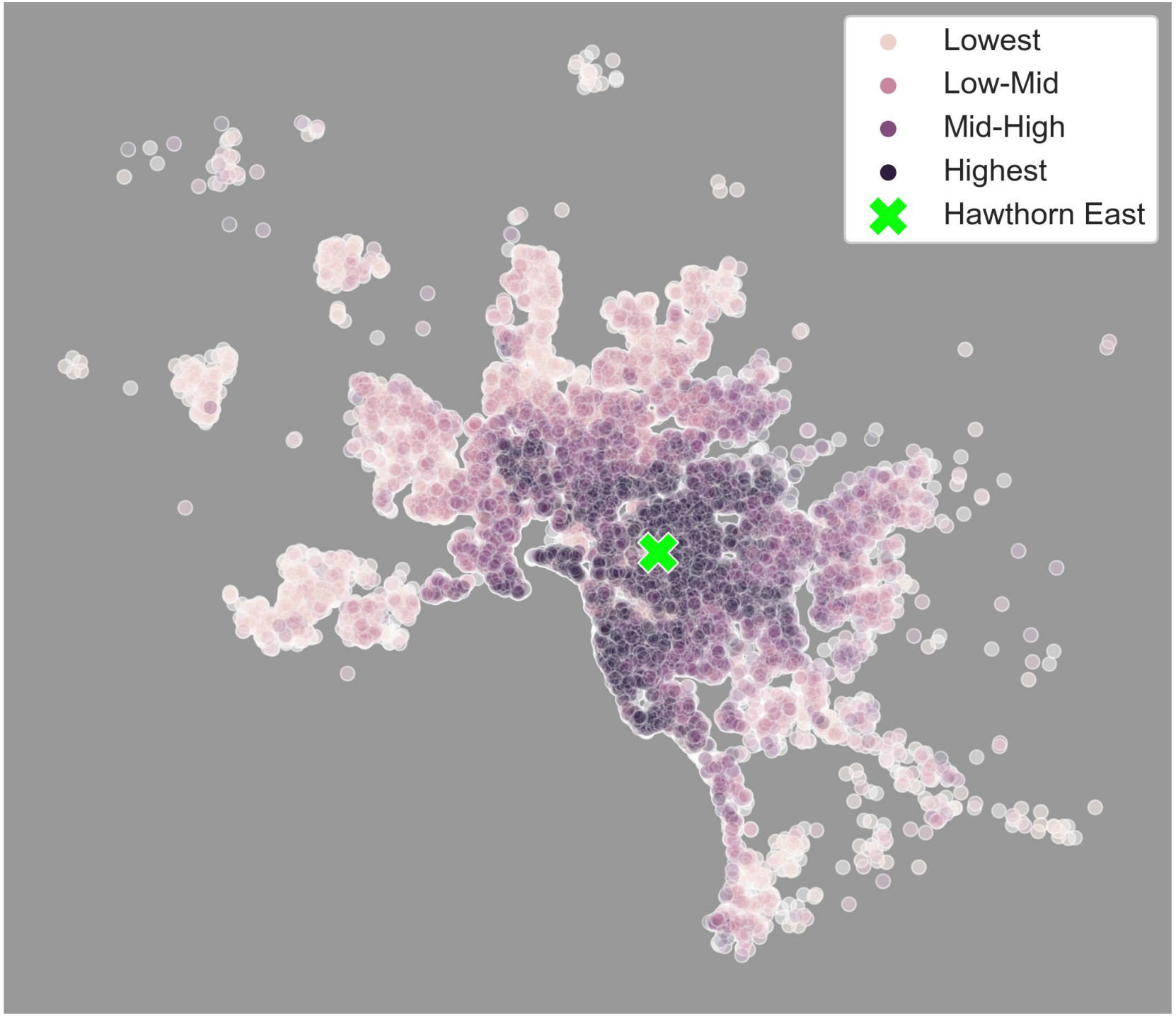
Melbourne Housing Market

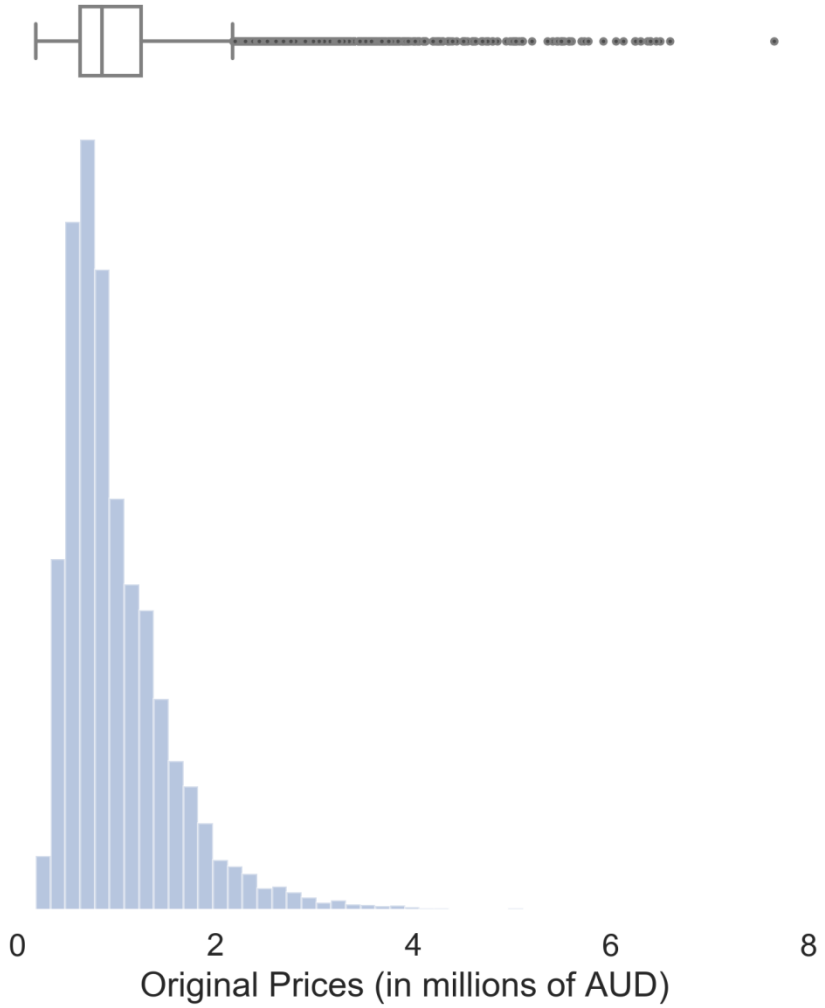
- Created by Tony Pino
- Available on Kaggle
- Auctions of houses from January 2016 to October 2018

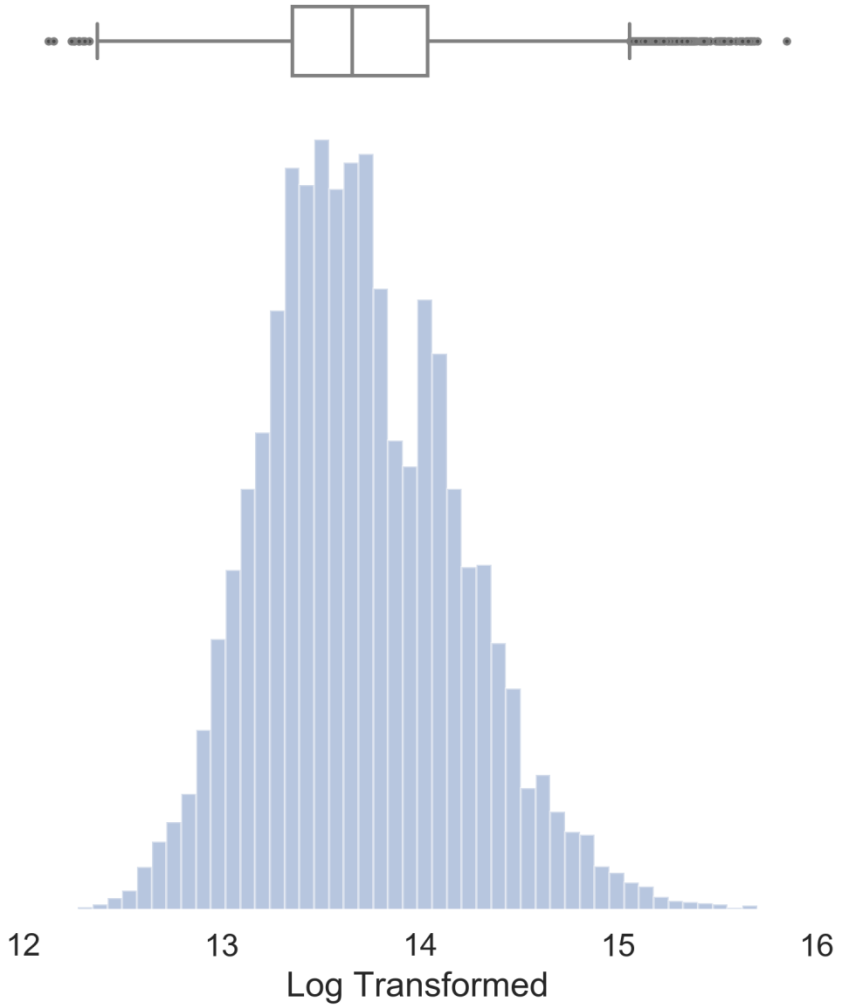


Where are the Most Expensive Houses?

- Highest prices are clustered
- Prices decrease further from Hawthorn East
- Distance from Hawthorn East captured in the “Hawthorn Distance” feature







What do House Prices Look Like?

- Strong right skew
- Log transformation works well

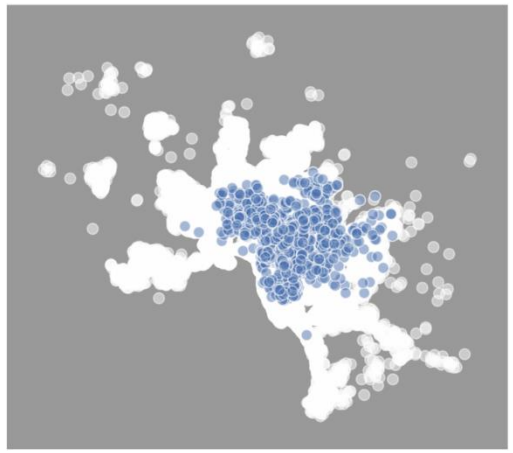
Top 10 Suburbs by Properties Sold

Top 10 Suburbs by Percent of Properties Sold

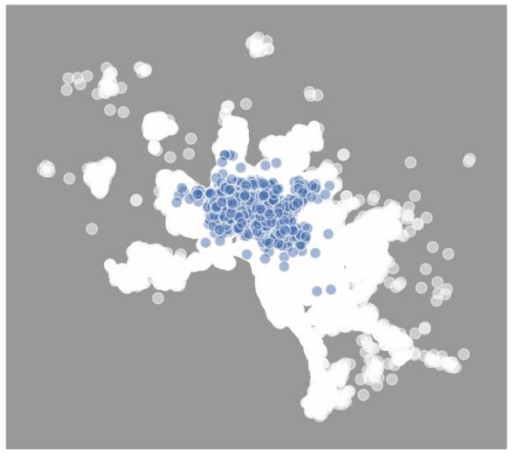
Where are the Most Sales Happening?

- Suburbs with the most properties sold are generally close to and East of the Central Business District (CBD)
- Suburbs with the highest % of their total properties sold are all about 10 km away from the CBD and in 3 clusters
- Distance from the CBD captured in the “CBD Distance” feature

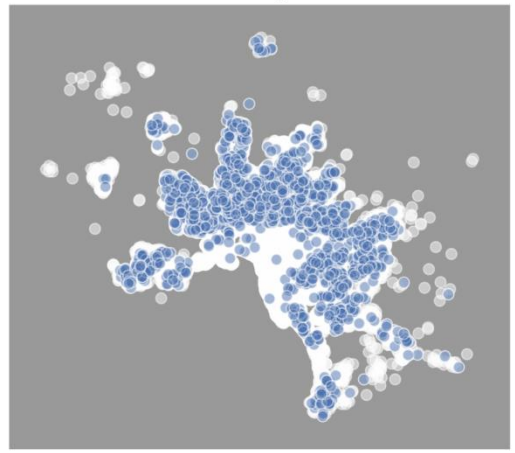
Jellis



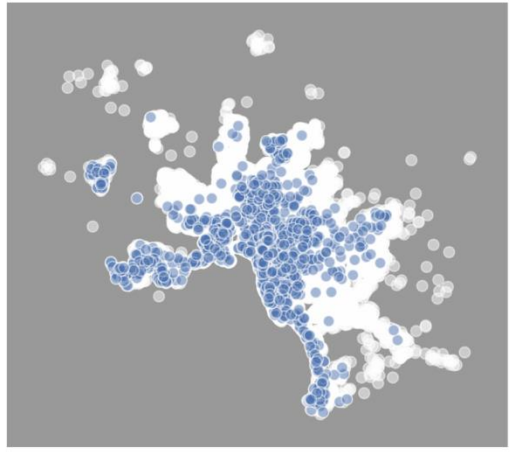
Nelson



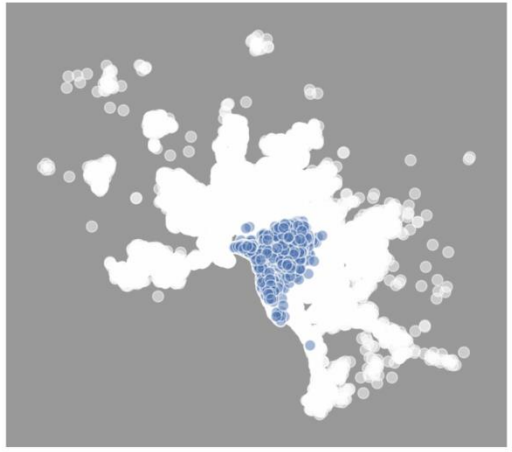
Barry



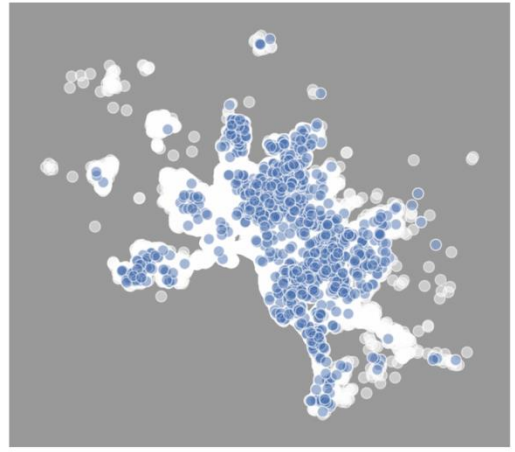
hockingstuart



Marshall



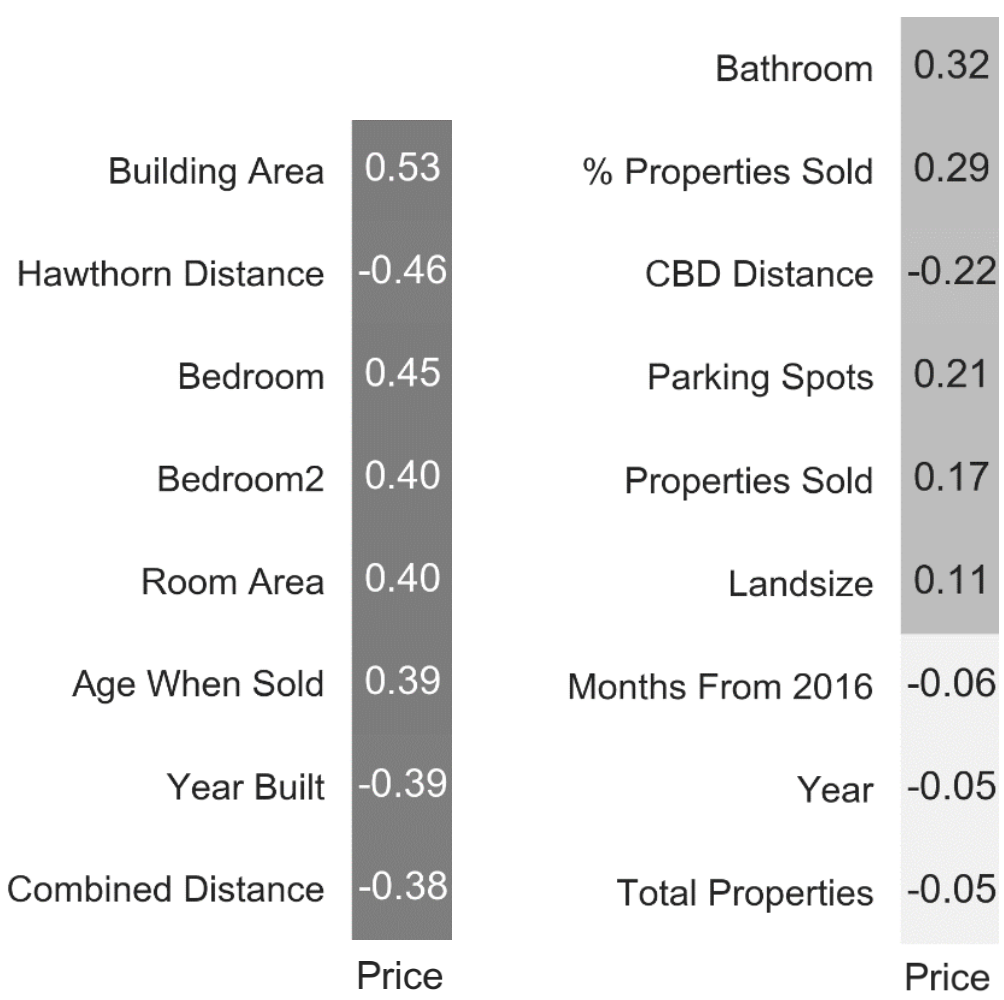
Ray

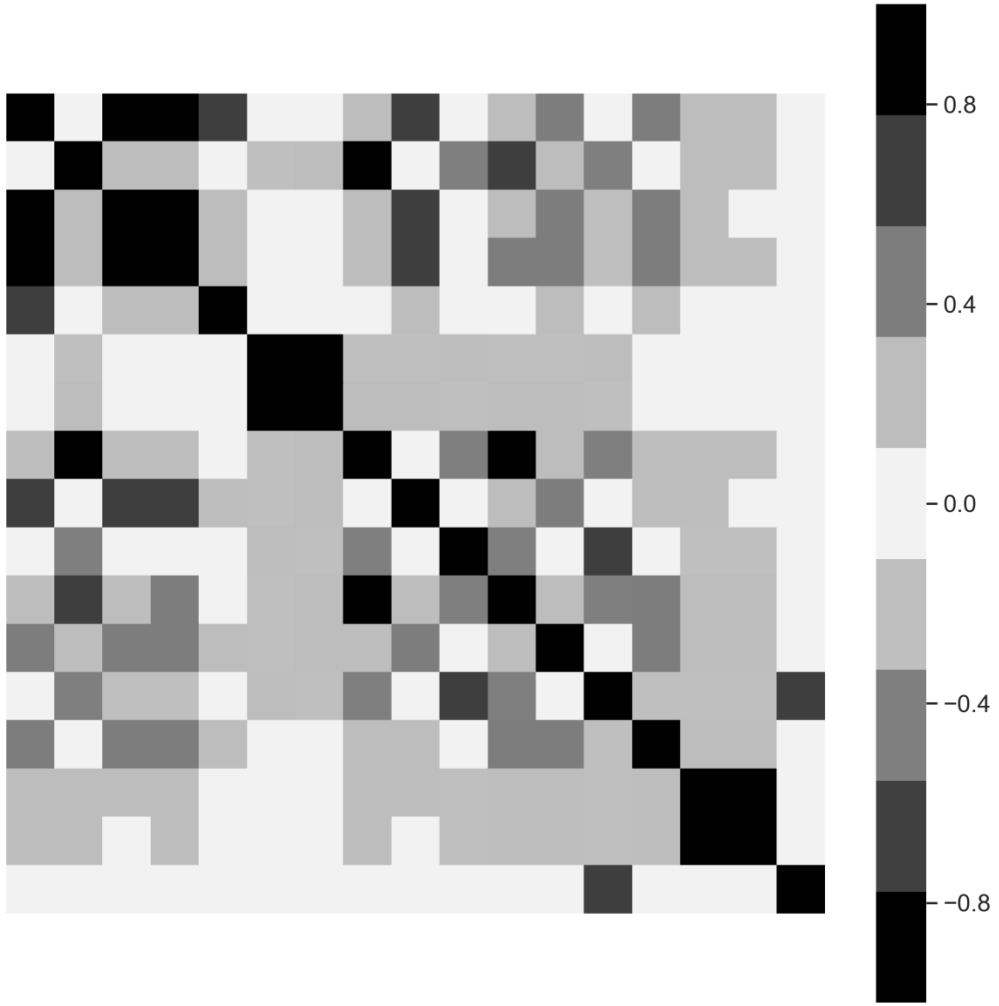


Who are the Top Sellers?

- Sellers seem to operate in defined regions

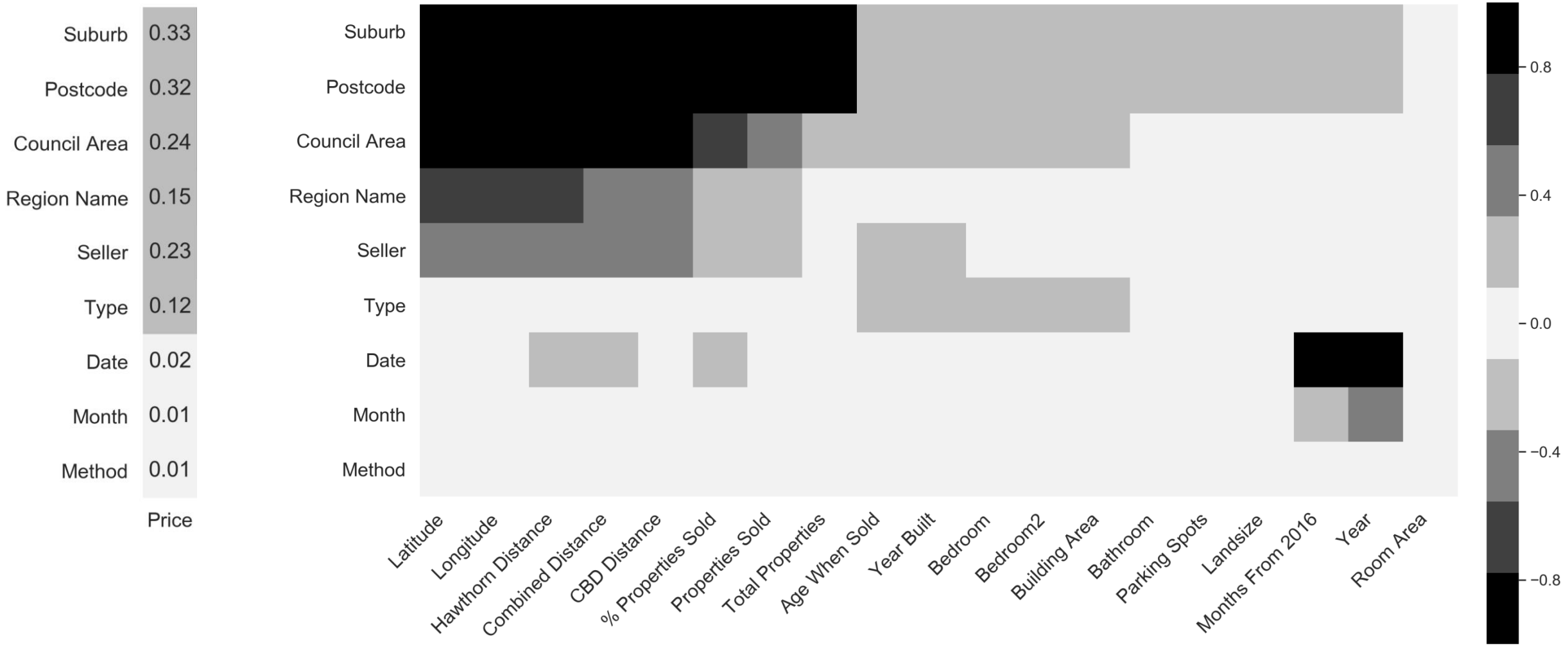
Feature Selection





Spearman's Correlations

- Assumes monotonic relationship between variables instead of a linear relationship (unlike Pearson's correlation)
- Strongest correlations are related to distances, building area, and age of the property
- Strong correlations exist among features

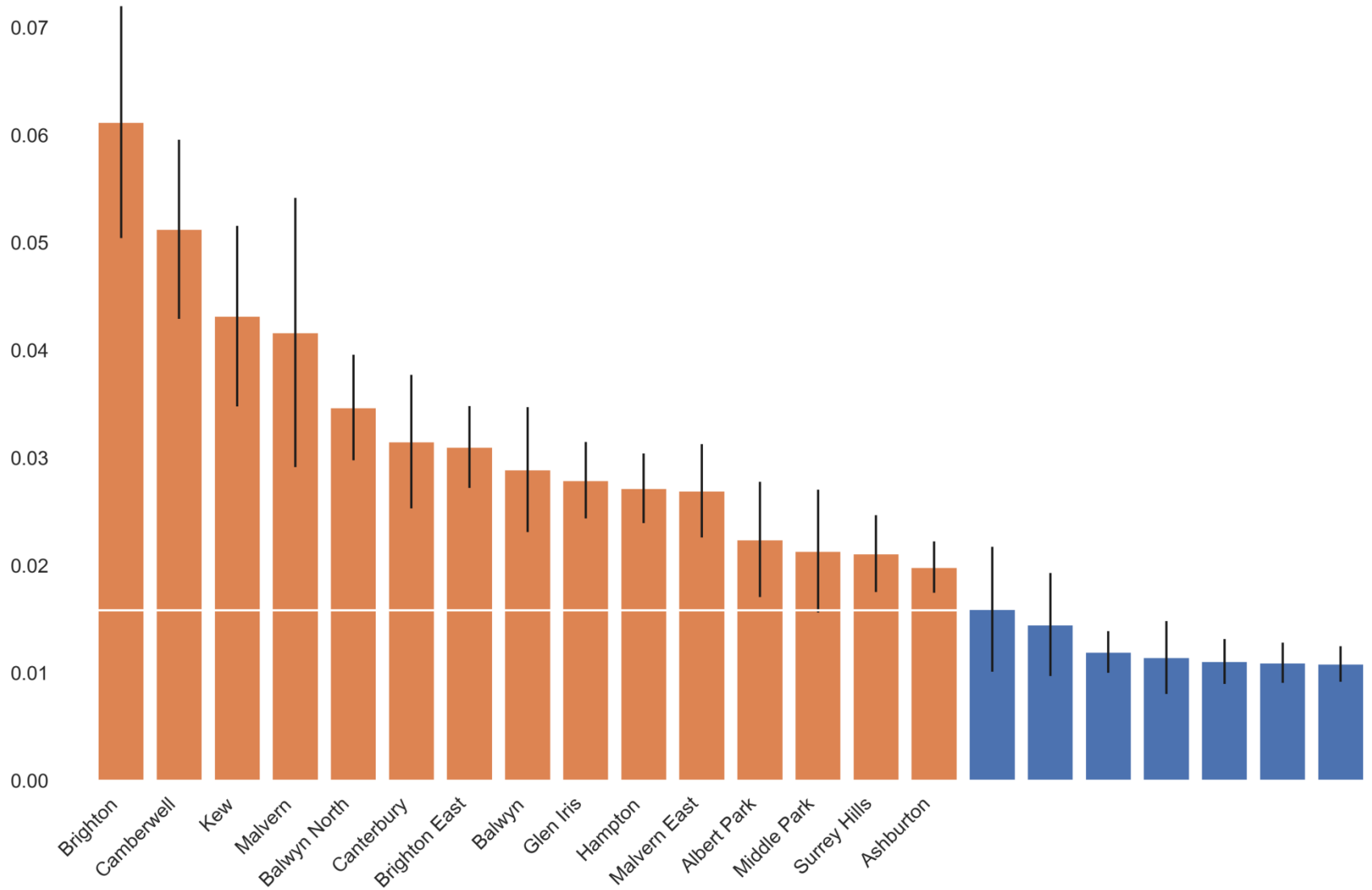


Adjusted R-squared Scores

- Used random forests to predict each numeric feature with each categorical feature
- Suburb is best overall for predicting numeric features
- Seller is surprisingly good at predicting distance-related features

Which Suburbs are Important?

- 350 unique suburbs
- Random forest ranks suburbs by their importance for predicting Price
- Select top 15 suburbs with importance above 0.015 to turn into features



Effect Coding

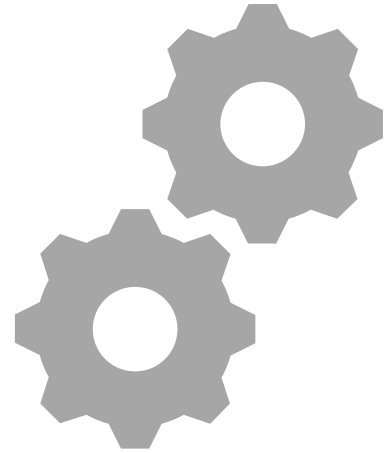
BEFORE

Index	Suburb
0	Brighton
1	Brighton
2	Brighton
3	Camberwell
4	Camberwell
5	Eaglemont
6	Mont Albert
7	Strathmore

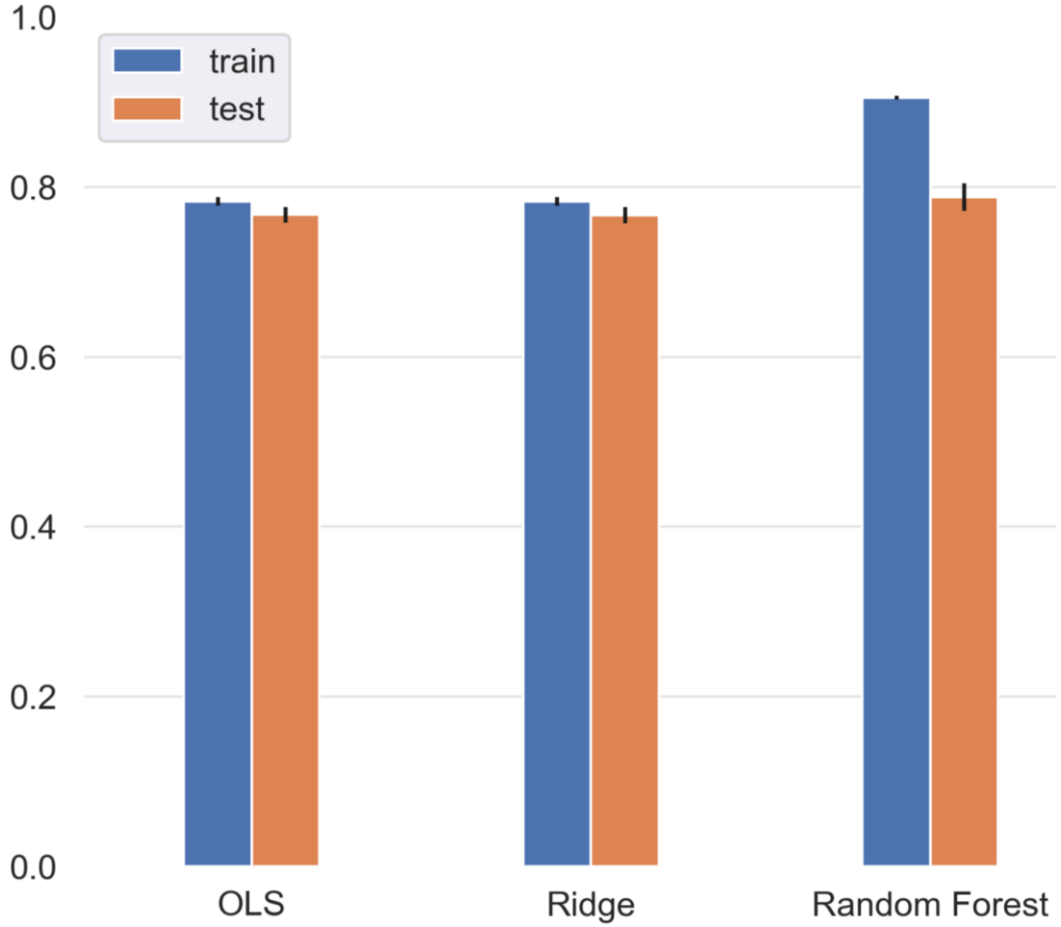
AFTER

Index	Brighton	Camberwell
0	1	0
1	1	0
2	1	0
3	0	1
4	0	1
5	-1	-1
6	-1	-1
7	-1	-1

Model Evaluation



Cross-Validation Scores



ADJUSTED R-SQUARED

Model	Train	Test
OLS	0.783 (0.773, 0.793)	0.767 (0.749, 0.785)
Ridge	0.783 (0.773, 0.793)	0.767 (0.749, 0.785)
Random Forest	0.905 (0.901, 0.910)	0.788 (0.755, 0.821)

SCIKIT-LEARN PARAMETERS

- Ridge: alpha=0.1
- Random Forest: max_depth=8, n_estimators=10

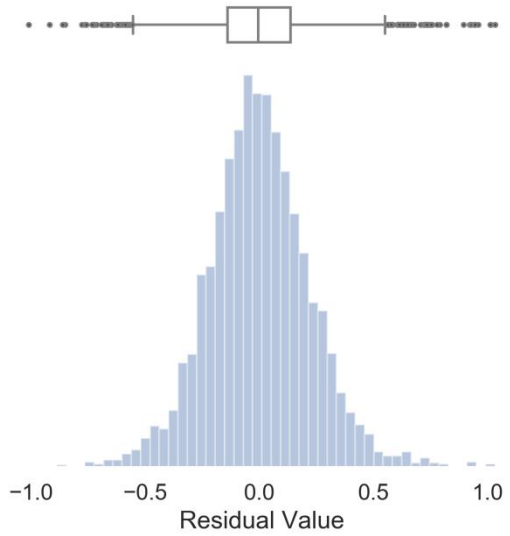
OLS or Ridge?

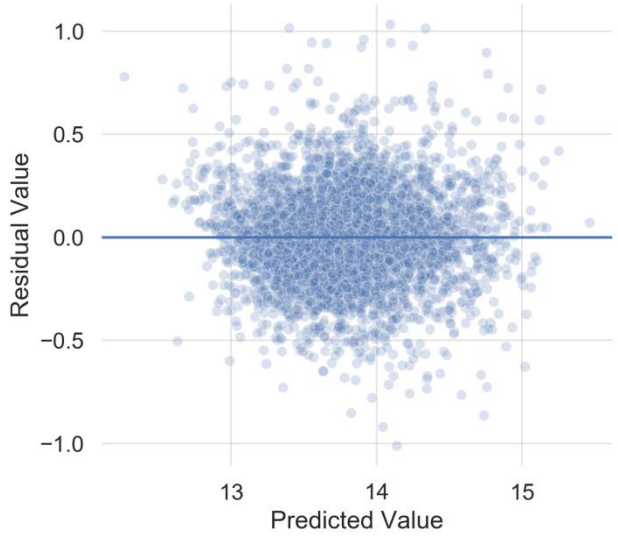
- 4 coefficients differ by over 10% between the models
  - These 4 features are strongly correlated with each other
- Choose ridge regression because it is better for analyzing data with correlated features

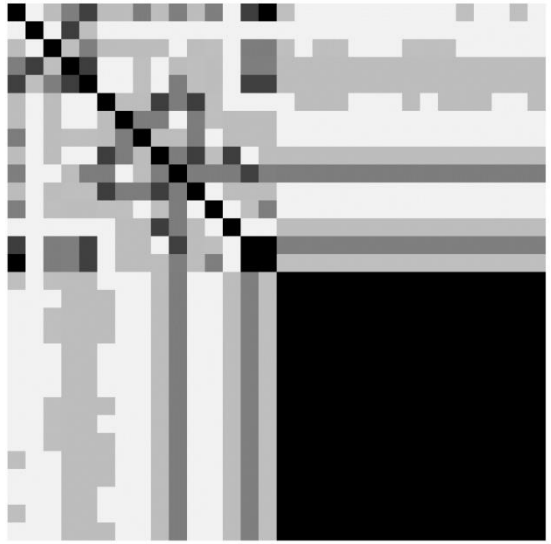
Feature	OLS Coefficient	Ridge Coefficient	% Difference
Bedroom	-0.842	-0.651	29.33
Bathroom	-0.534	-0.423	26.43
Room Area	-1.250	-1.034	20.87
Building Area	1.784	1.572	13.45
Combined Distance	-0.023	-0.022	3.38
% Properties Sold	-0.017	-0.017	2.53
Suburb Malvern East	-0.069	-0.068	2.14

Linear Regression Assumptions

- ☑ Is there a linear relationship between Price and each feature?
- ☑ Are the residuals normally distributed?
- ☑ Is the distribution of residuals consistent?
- ☒ Are the Pearson's correlations between features weak?







Which Correlated Features to Keep?

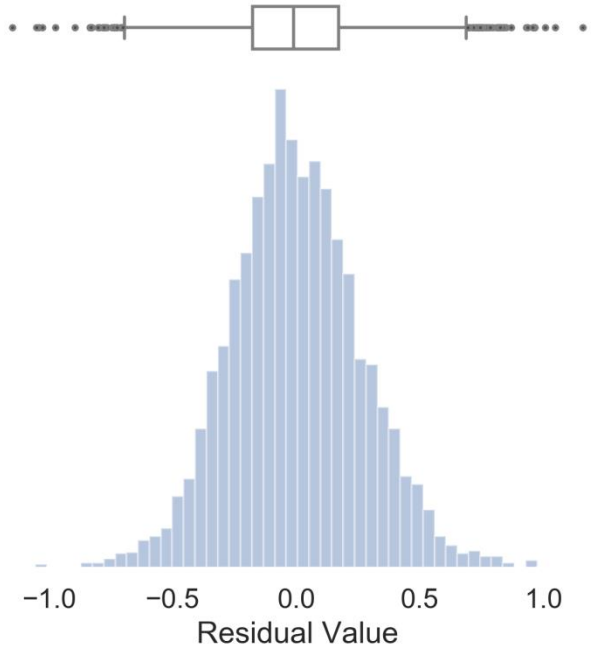
- Suburb features are all highly correlated because of effect coding
- Record the effect of removing each feature from a group on model performance
- Keep the feature whose removal most adversely affects the model

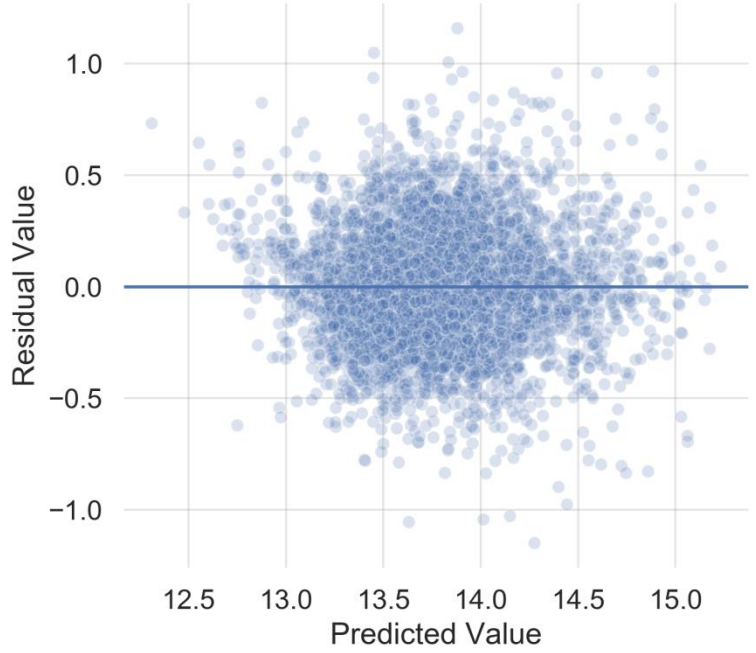
Group	Feature 1	Feature 2	Feature 3	Feature 4
1	CBD Distance	**Hawthorn Distance**	Combined Distance	
2	Properties Sold	**Total Properties**		
3	**Building Area**	Bathroom	Room Area	Bedroom
4	**Hawthorn Distance**	Months From 2016	Age When Sold	
5	Parking Spots	Landsize	**Building Area**	

- Remove **Total Properties** because the model performs about the same with or without it

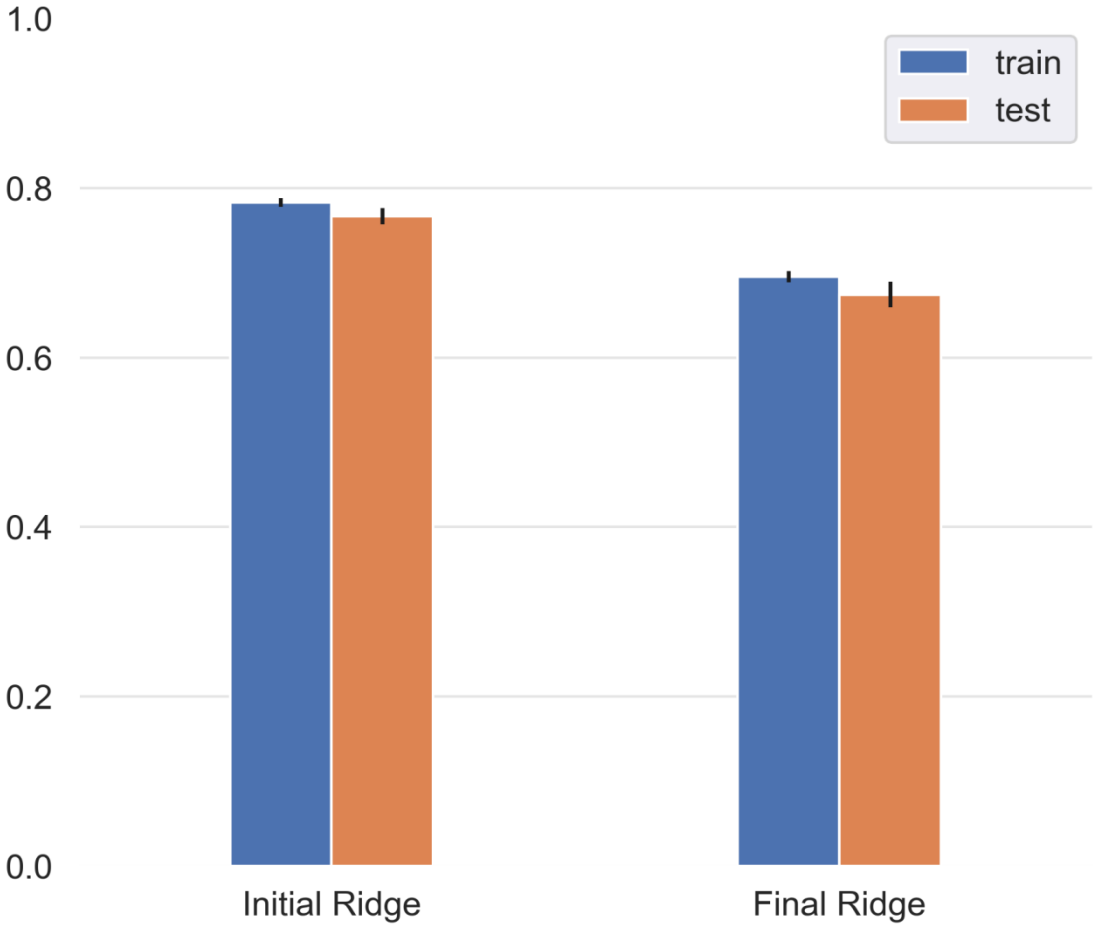
Final Ridge Model

- Keeping 2 non-suburb features out of 14
  - Building Area and Hawthorn Distance (correlation=0.04)
- Linear regression assumptions satisfied





Cross-Validation Scores



ADJUSTED R-SQUARED

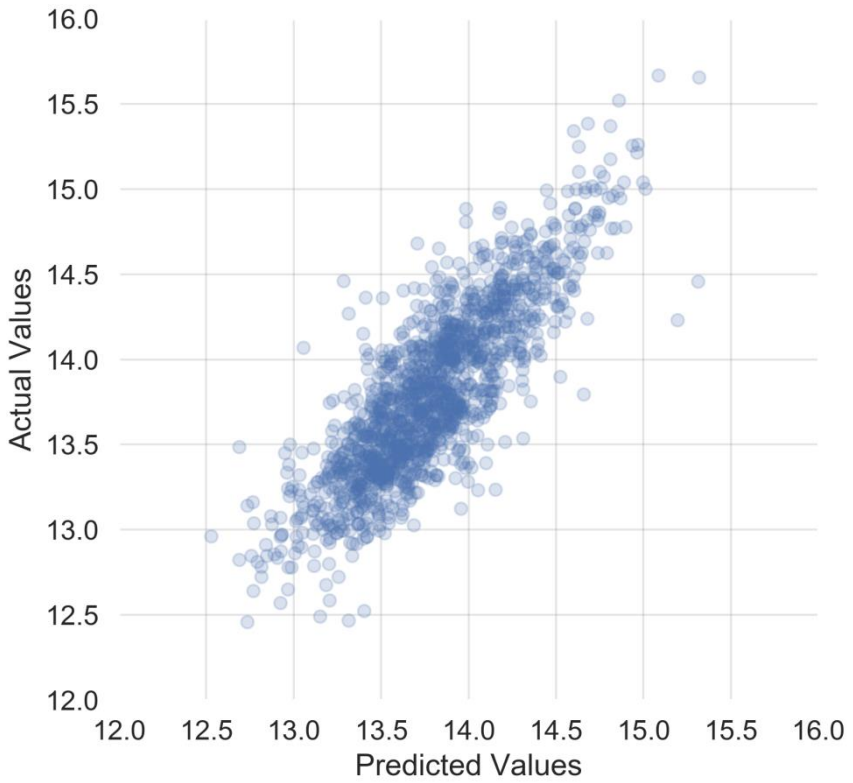
Model	Train	Test
Initial Ridge	0.783 (0.773, 0.793)	0.767 (0.749, 0.785)
Final Ridge	0.696 (0.682, 0.709)	0.674 (0.644, 0.705)

SCIKIT-LEARN PARAMETERS

- Initial Ridge: alpha=0.1
- Final Ridge: alpha=0.7

Final Results

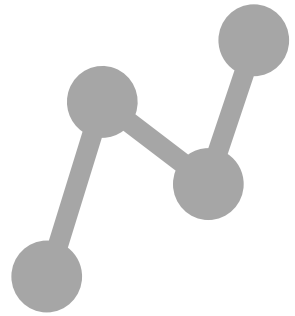
- Adjusted R-squared on holdout set: 0.695



Feature Coefficients	
Building Area	0.678
Hawthorn Distance	-0.276
Suburb Albert Park	0.427
Suburb Ashburton	-0.131
Suburb Balwyn	-0.170
Suburb Balwyn North	-0.109
Suburb Brighton	0.250
Suburb Brighton East	0.116
Suburb Camberwell	-0.264

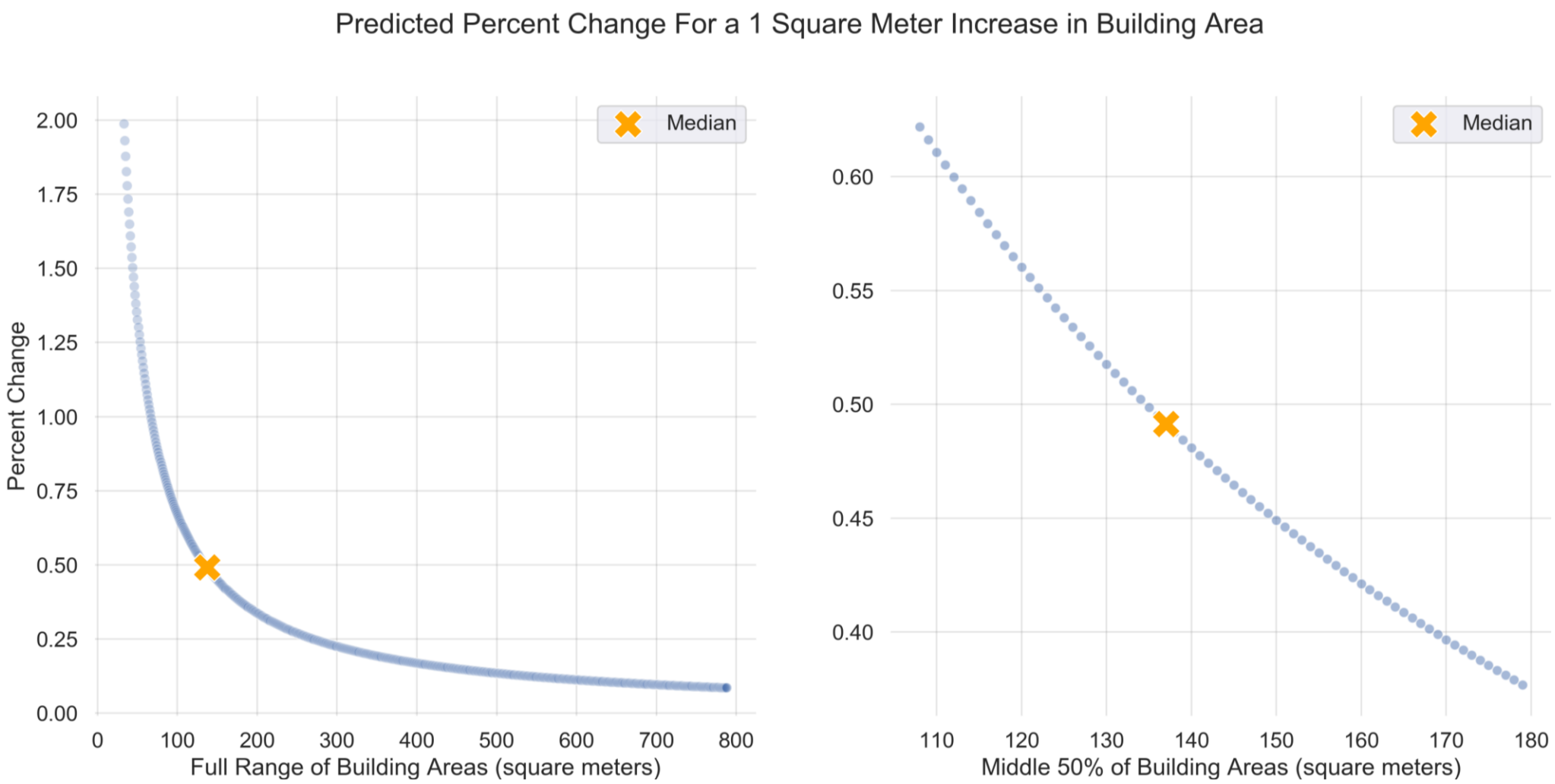
Feature Coefficients	
Suburb Canterbury	-0.152
Suburb Glen Iris	-0.190
Suburb Hampton	0.255
Suburb Kew	-0.124
Suburb Malvern	0.033
Suburb Malvern East	-0.079
Suburb Middle Park	0.470
Suburb Surrey Hills	-0.195

Model Interpretation



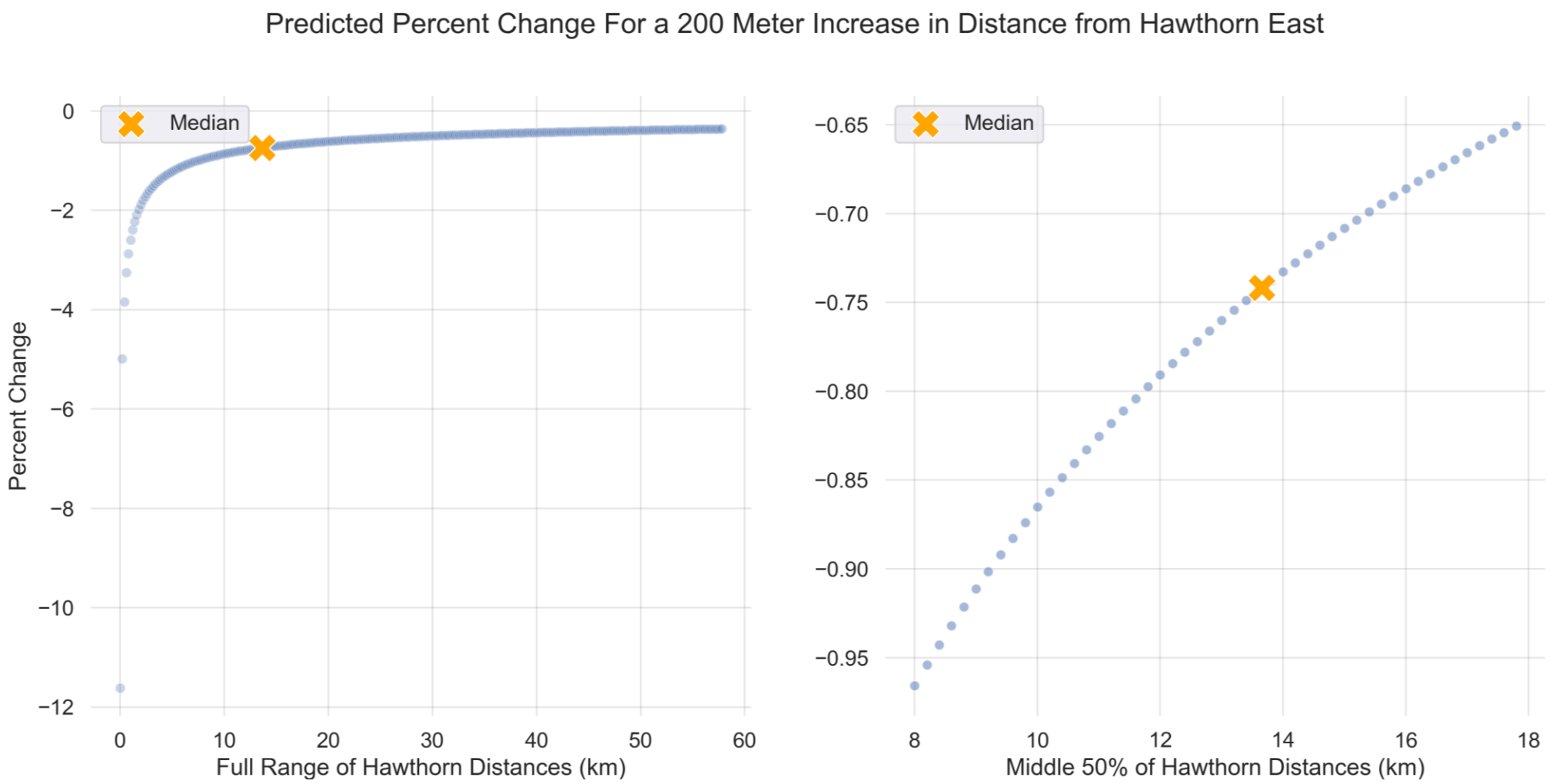
What do the Coefficients Mean?

- Interpretation isn't straightforward because Price and the features were transformed
- Log transformed: Price, Building Area
- Square-root transformed: Hawthorn Distance
- All the suburb features were effect coded



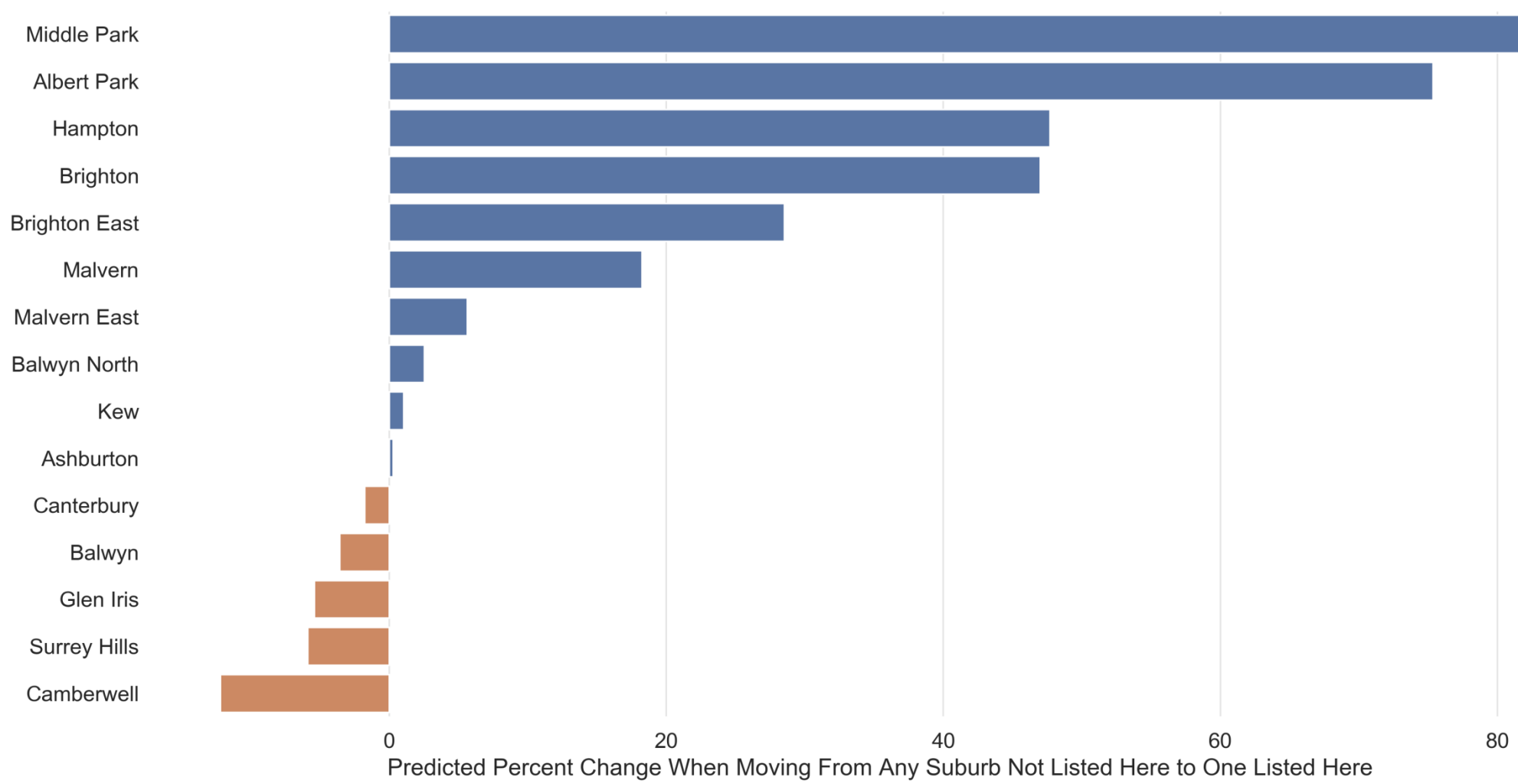
Effect of Building Area on Price

- Percent change in Price depends on the value of Building Area and its coefficient
- A property with 1 m^2 more Building Area than the median (137 m^2) is predicted to see a 0.49% increase in its Price
- Ex: Property valued at $936,926 with 137 m^2 Building Area is predicted to be worth $941,530 if it had 138 m^2 Building Area



Effect of Hawthorn Distance on Price

- Percent change in Price depends on the value of Hawthorn Distance and its coefficient
- A property that is 200 meters further than the median (13.7 km) is predicted to see a 0.74% decrease in its Price
- Ex: Property valued at $936,926 at 13.7 km away from Hawthorn East is predicted to be worth $929,976 if it were 13.9 km away



Effect of Suburb on Price

- Percent change in Price depends on all of the coefficients for the suburb features, but not the values of the suburb features
- A property that is not in any of these suburbs is predicted to see a 75.26% increase in its Price if it was in Albert Park
- Ex: Property valued at $936,926 and not in any of these suburbs is predicted to be worth $1,642,910 if it was in Albert Park

So… What Affects the Price of a House?

- An increase in building area is predicted to increase Price
- An increase in distance from Hawthorn East is predicted to decrease Price

Predicted Change at 1st, 2nd, and 3rd Quartiles			
Building Area[13]	108 m^2 → 109 m^2 +0.62% ($4,965)	137 m^2 → 138 m^2 +0.49% ($4,604)	180 m^2 → 181 m^2 +0.37% ($4,221)
Hawthorn Distance[14]	8.7 km → 8.9 km -0.93% ($10,661)	13.7 km → 13.9 km -0.74% ($6,950)	18.4 km → 18.6 km -0.64% ($5,095)

Assumptions:

[1]The property is not located in any of the suburbs that were turned into features

[3]The property is at the median Hawthorn Distance

[4]The property has the median value for Building Area

So… What Affects the Price of a House?

- Moving to one of the suburbs listed below is predicted to affect Price as follows:

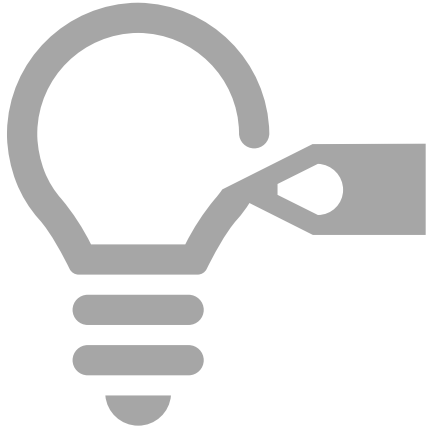
Predicted Increase		
Middle Park	+83.04%	($778,056)
Albert Park	+75.35%	($705,984)
Hampton	+47.71%	($446,998)
Brighton	+47.01%	($440,441)
Brighton East	+28.54%	($267,378)

Predicted Increase		
Malvern	+18.25%	($170,976)
Malvern East	+5.62%	($52,641)
Balwyn North	+2.54%	($23,816)
Kew	+1.03%	($9,652)
Ashburton	+0.27%	($2,526)

Predicted Decrease		
Camberwell	-12.12%	($114,244)
Surrey Hills	-5.88%	($55,121)
Glen Iris	-5.40%	($50,592)
Balwyn	-3.57%	($33,444)
Canterbury	-1.79%	($16,733)

Assumptions:
The property was not originally located in any of the suburbs listed above
The property is at the median Hawthorn Distance
The property has the median value for Building Area

Future Research



Ideas Explored and Unexplored

- ☒ Can the dataset be partitioned by CBD Distance?
  - ❑ Adjusted R-squared for one partition was slightly better
  - ❑ Score for the other partition was much worse
- ❑ Is there a pattern in the residuals that depend on suburb?
- ❑ What are the p-values of the coefficients from bootstrapping?

Appendix

Log Transformation

The dependent and independent variables are transformed, so the relation between the percent change in Price (dependent variable) given an increase of C in a feature (independent variable), assuming all the other independent variables are held constant, is:

$$100\frac{y_2 - y_1}{y_1} = 100(e^{B\ln\left(1+\frac{C}{x_1}\right)} - 1)$$

where B is the coefficient for that feature and x_1 is the starting value.

The percent change in Price depends not only on the coefficient B, but also on the starting value of the feature.

Square-Root Transformation

The dependent and independent variables are transformed, so the relation between the percent change in Price (dependent variable) given an increase of C in a feature (independent variable), assuming all the other independent variables are held constant, is:

$$100\frac{y_2 - y_1}{y_1} = 100(e^{B\sqrt{x_1 - C} - \sqrt{x_1}} - 1)$$

where B is the coefficient for that feature and x_1 is the starting value.

The percent change in Price depends not only on the coefficient B, but also on the starting value of the feature.

Effect Coding

If a property is in the suburb Albert Park, it will have a value of 1 for that suburb feature and 0 for all the other suburb features. If a property is not in any of the selected suburbs, it will have a value of -1 for all the suburb features ('Other' category).

We won't be able to change the value for one suburb feature without changing all the other suburb features. I'll interpret the coefficients using the 'Other' category as the reference point. The relation between the percent change in the Price, assuming the property was originally in the 'Other' category, and the new suburb of the property, is:

$$100\frac{y_2 - y_1}{y_1} = 100(e^{B_1 + \sum_{n=1}^{k} B_n} - 1)$$

where B_1 is the coefficient for the new suburb of the property, k is the number of suburb features, and $\sum_{n=1}^{k} B_n$ is the sum of the coefficients for all the suburb features.

The percent change in Price doesn't depend on the values of the suburb features, but it depends on all of the coefficients for the suburb features.