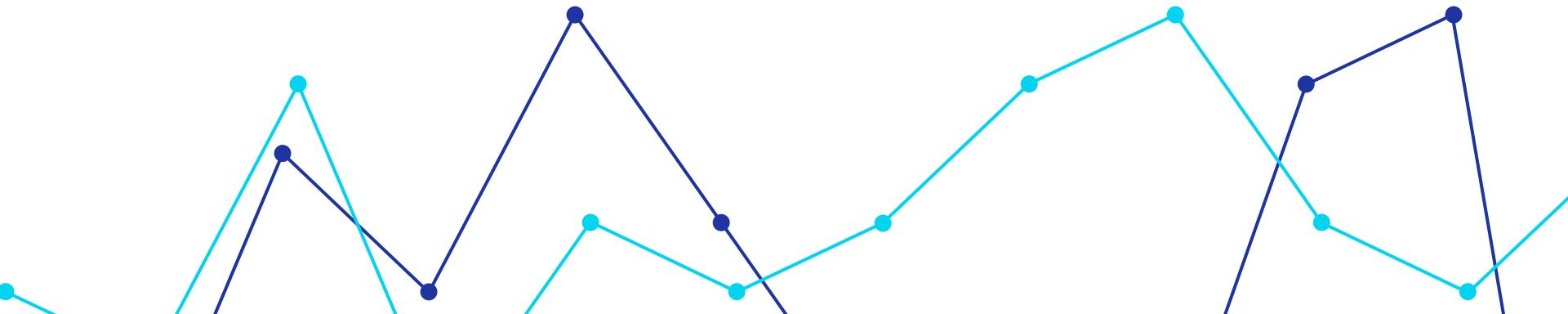


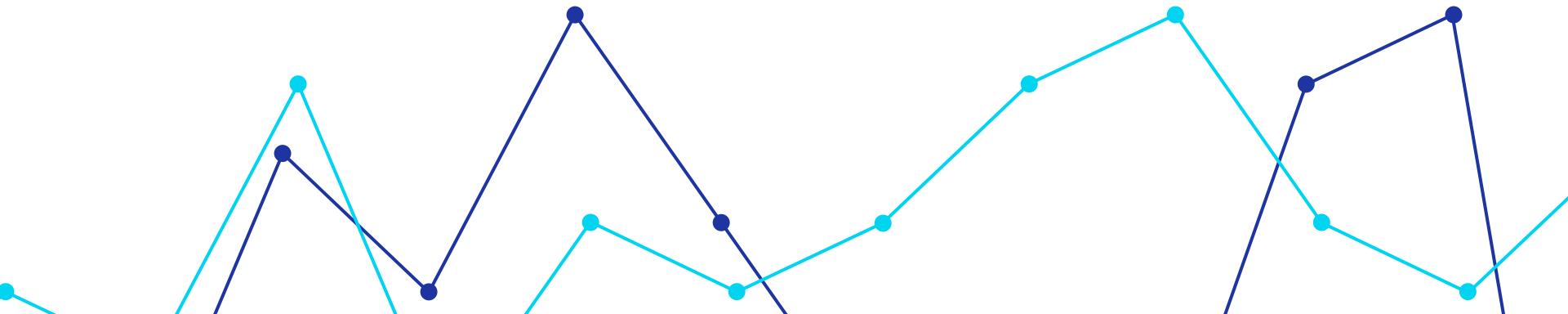
SC1015 Mini Project: Melbourne Housing

REP2 Group 7
Lex Tan Pengqin
Tan Yu Xiu
Lim Jun, Shawn



1

Practical Motivation + Problem Formulation



Practical Motivation - Relatability to Singapore

- Resilient real estate sector
- Stable economy

2024

9,103

41.3%

2025

5,438

completed private property units

Supply Constraints

Strong Demand



Practical Motivation - Relatability to Singapore

Analysis

Understand how different features and locations **impact property prices**

Methodology

Explore methods to **identify emerging high-value areas** and markets

Models

Develop models to **forecast property trends**, helping stakeholders **make decisions**



Comparable Challenges

Limited Land Supply

Rising Prices

Demographic Shifts



Trend Analysis

Problem Definition

1

Which features have the greatest impact on price?

2

How has the price trend varied across different regions over time, and can we identify emerging high-value areas?

3

How does property age and size influence prediction accuracy in models?

PREPARATION

Problem FORMULATION

Sample Collection

Raw data is accessed from the Kaggle datasets and extracted via the `read_csv` Pandas function

A total of 13580 entries, with 21 columns

Contains data on the price, features, location and other attributes

The screenshot shows the Kaggle interface. On the left, there's a sidebar with a navigation menu including 'kaggle', 'Create', 'Home', 'Competitions', 'Datasets' (which is selected), 'Models', 'Code', 'Discussions', 'Learn', 'More', 'Your Work', and a 'VIEWED' section with items like 'Melbourne Housing S...', 'Error in filepath. Synta...', 'How can I get my first...', 'Hello People, Check t...', 'Suspect building area...'. The main content area has a search bar at the top. Below it, a dataset card for 'Melbourne Housing Snapshot' by Tony Pino is displayed. The card includes a thumbnail image of a city skyline, a 'Data Card' tab (selected), and other tabs for 'Code (6620)', 'Discussion (10)', and 'Suggestions (0)'. The 'About Dataset' section contains 'Context' (describing Melbourne real estate as 'BOOMING'), 'Content' (noting it's a snapshot by Tony Pino), and 'Notes on Specific Variables' (mentioning 'Rooms: Number of rooms' and 'Price: Price in dollars'). To the right of the card, a detailed description of the DataFrame is shown:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13580 entries, 0 to 13579
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype  
--- 
 0   Suburb          13580 non-null   object  
 1   Address          13580 non-null   object  
 2   Rooms            13580 non-null   int64  
 3   Type             13580 non-null   object  
 4   Price            13580 non-null   float64 
 5   Method           13580 non-null   object  
 6   SellerG          13580 non-null   object  
 7   Date             13580 non-null   object  
 8   Distance         13580 non-null   float64 
 9   Postcode          13580 non-null   float64 
 10  Bedroom2         13580 non-null   float64 
 11  Bathroom          13580 non-null   float64 
 12  Car               13518 non-null   float64 
 13  Landsize          13580 non-null   float64 
 14  BuildingArea      7130 non-null   float64 
 15  YearBuilt         8205 non-null   float64 
 16  CouncilArea        12211 non-null  object  
 17  Latitude           13580 non-null   float64 
 18  Longitude          13580 non-null   float64 
 19  Regionname         13580 non-null   object  
 20  Propertycount      13580 non-null   float64 
dtypes: float64(12), int64(1), object(8)
memory usage: 2.2+ MB
None
```

PREPARATION

Problem FORMULATION

Sample Collection

Raw data is accessed from the Kaggle datasets and extracted via the `read_csv` Pandas function

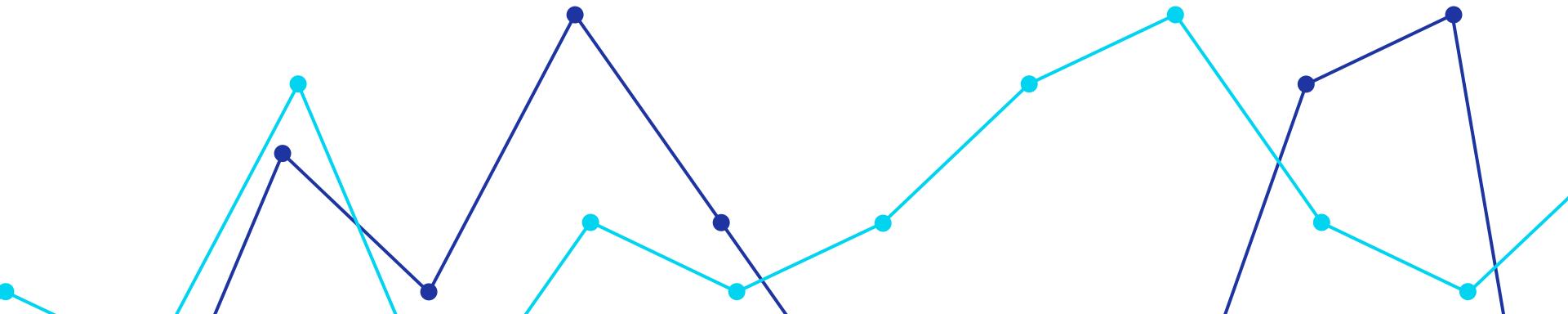
A total of 13580 entries, with 21 columns

Contains data on the price, features, location and other attributes

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R		
1	Suburb	Address		Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt	CouncilArea	
2	Abbotsford	85 Turner St		2	h	1480000	S	Biggin	3/12/2016	2.5	3067	2	1	1	202		Yarra	-37.7996	14
3	Abbotsford	25 Bloomberg St		2	h	1035000	S	Biggin	4/2/2016	2.5	3067	2	1	0	156	79	1900 Yarra	-37.8079	14
4	Abbotsford	5 Charles St		3	h	1465000	SP	Biggin	4/3/2017	2.5	3067	3	2	0	134	150	1900 Yarra	-37.8093	14
5	Abbotsford	40 Federation La		3	h	850000	PI	Biggin	4/3/2017	2.5	3067	3	2	1	94		Yarra	-37.7969	14
6	Abbotsford	55a Park St		4	h	1600000	VB	Nelson	4/6/2016	2.5	3067	3	1	2	120	142	2014 Yarra	-37.8072	14
7	Abbotsford	129 Charles St		2	h	941000	S	Jellis	7/5/2016	2.5	3067	2	1	0	181		Yarra	-37.8041	14
8	Abbotsford	124 Yarra St		3	h	1876000	S	Nelson	7/5/2016	2.5	3067	4	2	0	245	210	1910 Yarra	-37.8024	14
9	Abbotsford	98 Charles St		2	h	1636000	S	Nelson	8/10/2016	2.5	3067	2	1	2	256	107	1890 Yarra	-37.806	14
10	Abbotsford	6/241 Nicholson St		1	u	300000	S	Biggin	8/10/2016	2.5	3067	1	1	1	0		Yarra	-37.8008	14
11	Abbotsford	10 Valiant St		2	h	1097000	S	Biggin	8/10/2016	2.5	3067	3	1	2	220	75	1900 Yarra	-37.801	14
12	Abbotsford	411/8 Grosvenor St		2	u	700000	VB	Jellis	12/11/2016	2.5	3067	2	2	1	0		Yarra	-37.811	14
13	Abbotsford	40 Nicholson St		3	h	1350000	VB	Nelson	12/11/2016	2.5	3067	3	2	2	214	190	2005 Yarra	-37.8085	14
14	Abbotsford	123/56 Nicholson St		2	u	750000	S	Biggin	12/11/2016	2.5	3067	2	2	1	0	94	2009 Yarra	-37.8078	14
15	Abbotsford	45 William St		2	h	1172500	S	Biggin	13/8/2016	2.5	3067	2	1	1	195		Yarra	-37.8084	14
16	Abbotsford	7/20 Abbotsford St		1	u	441000	SP	Greg	14/5/2016	2.5	3067	1	1	1	0		Yarra	-37.8016	14
17	Abbotsford	16 William St		2	h	1310000	S	Jellis	15/10/2016	2.5	3067	2	1	2	238	97	1890 Yarra	-37.809	14
18	Abbotsford	42 Henry St		3	h	1200000	S	Jellis	16/7/2016	2.5	3067	3	2	1	113	110	1880 Yarra	-37.8056	1
19	Abbotsford	78 Yarra St		3	h	1176500	S	LITTLE	16/7/2016	2.5	3067	2	1	1	138	105	1890 Yarra	-37.8021	14
20	Abbotsford	196 Nicholson St		3	h	955000	S	Collins	17/9/2016	2.5	3067	3	1	0	183		Yarra	-37.8022	14
21	Abbotsford	42 Valiant St		2	h	899000	S	Biggin	17/9/2016	2.5	3067	2	1	1	150	73	1985 Yarra	-37.8011	14
22	Abbotsford	3/72 Charles St		4	h	1330000	PI	Kay	18/3/2017	2.5	3067	4	2	2	780	135	1900 Yarra	-37.8073	14
23	Abbotsford	13/11 Nicholson St		3	t	900000	S	Beller	18/3/2017	2.5	3067	3	2	2	0		2010 Yarra	-37.8093	14
24	Abbotsford	138/56 Nicholson St		3	u	1090000	S	Jellis	18/3/2017	2.5	3067	3	2	2	4290	27	Yarra	-37.8078	14
25	Abbotsford	6/219 Nicholson St		2	u	500000	S	Collins	18/6/2016	2.5	3067	2	1	1	0	60	1970 Yarra	-37.8015	14
26	Abbotsford	52a William St		2	h	1100000	PI	Biggin	18/6/2016	2.5	3067	2	2	1	124	135	2013 Yarra	-37.8079	14
27	Abbotsford	49 Park St		2	h	1315000	S	Marshall	19/11/2016	2.5	3067	2	1	0	147	85	1900 Yarra	-37.808	1

2

Data Processing + Exploratory Data Analysis

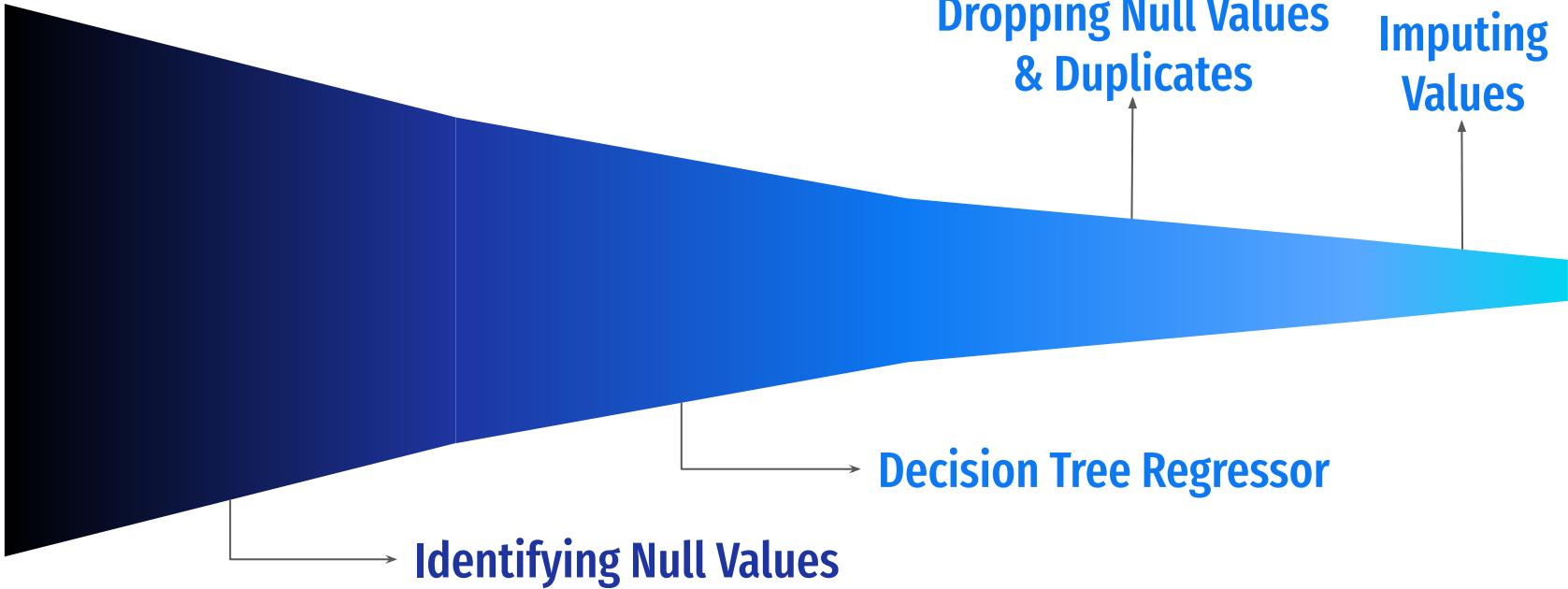


Exploratory ANALYSIS

Statistical DESCRIPTION

1. Preliminary Exploration

We adopted a 4 tiered approach for data exploration using statistical tools, before conducting targeted data preparation and cleaning



a. Identifying Null Values

```
print(housingdata.isnull().sum())
```



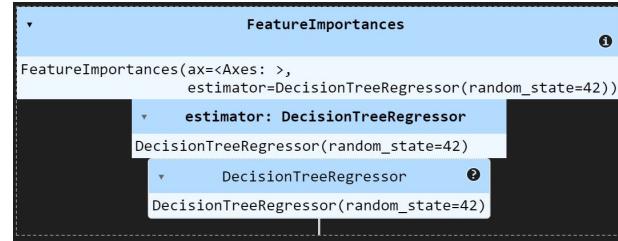
Suburb	0
Address	0
Rooms	0
Type	0
Price	0
Method	0
SellerG	0
Date	0
Distance	0
Postcode	0
Bedroom2	0
Bathroom	0
Car	62
Landsize	0
BuildingArea	6450
YearBuilt	5375
CouncilArea	1369
Lattitude	0
Longtitude	0
Regionname	0
Propertycount	0
dtype: int64	

b. Decision Tree Regressor

```
encoder = OrdinalEncoder()
housingdata[categorical] = encoder.fit_transform(housingdata[categorical].astype(str))

X1 = housingdata.drop(columns=["Price"]) # Features
y1= housingdata["Price"] # Target

model1 = DecisionTreeRegressor(random_state=42)
viz1 = FeatureImportances(model1)
viz1.fit(X1, y1)
```

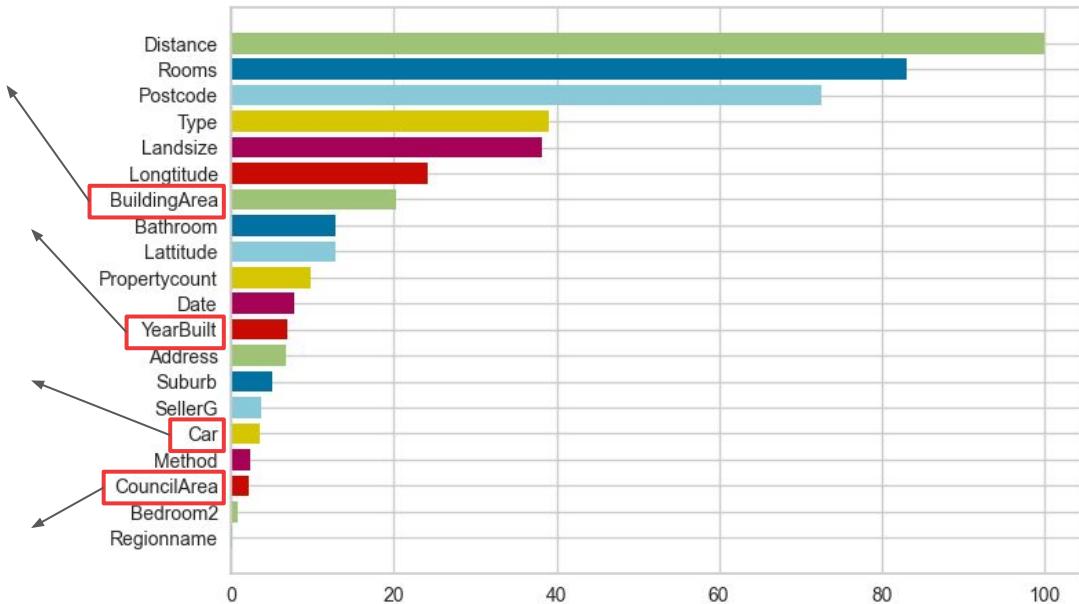


47.5% of rows
#7 in feature importance

39.6% of rows
#12 in feature importance

0.46% of rows
#16 in feature importance

10.1% of rows
#18 in feature importance



c. Handling NULL Values

Dropping Values

Variable: Car

Drop rows to reduce complexity without sacrificing much information, as its importance in predicting Price is low.

```
# drop rows accordingly
housingdata = housingdata.dropna(subset=['Car'])
```

Imputing Values

Variables: BuildingArea, YearBuilt, CouncilArea

Impute the missing values using the median/mode, grouped by relevant contextual property characteristics (Suburb, Rooms, Type) The number of NULL rows take up a notable percentage of the entire dataset, and the importance of these variables are moderate, hence we opt to impute.

```
# Impute BuildingArea based on median within Suburb + Distance + Rooms
housingdata['BuildingArea'] = housingdata.groupby(['Suburb', 'Rooms', 'Type'])['BuildingArea'].transform(
    lambda x: x.fillna(x.median()))
)

# Impute YearBuilt based on Suburb + Type
housingdata['YearBuilt'] = housingdata.groupby(['Suburb', 'Type'])['YearBuilt'].transform(
    lambda x: x.fillna(x.median()))
)

# Impute CouncilArea (categorical) based on Suburb
housingdata['CouncilArea'] = housingdata.groupby('Suburb')['CouncilArea'].transform(
    lambda x: x.fillna(x.mode()[0] if not x.mode().empty else 'Unknown'))
)
```

2. Analysis of Numerical Variables

We will analyse the below numerical variables using a Box Plot, Histogram, and Violin Plot.



	Bathroom	Car	Landsize	BuildingArea	YearBuilt
count	13580.00000	13518.00000	13580.00000	7130.00000	8205.00000
mean	1.534242	1.610075	558.416127	151.967650	1964.684217
std	0.691712	0.962634	3990.669241	541.014538	37.273762
min	0.000000	0.000000	0.000000	0.000000	1196.000000
25%	1.000000	1.000000	177.000000	93.000000	1940.000000
50%	1.000000	2.000000	440.000000	126.000000	1970.000000
75%	2.000000	2.000000	651.000000	174.000000	1999.000000
max	8.000000	10.000000	433014.000000	44515.000000	2018.000000

2. Analysis of Numerical Variables

We will analyse the below numerical variables using a Box Plot, Histogram, and Violin Plot.

	Latitude	Longitude	Propertycount
count	13580.000000	13580.000000	13580.000000
mean	-37.809203	144.995216	7454.417378
std	0.079260	0.103916	4378.581772
min	-38.182550	144.431810	249.000000
25%	-37.856822	144.929600	4380.000000
50%	-37.802355	145.000100	6555.000000
75%	-37.756400	145.058305	10331.000000
max	-37.408530	145.526350	21650.000000

2. Analysis of Numerical Variables

We will analyse the below numerical variables using a Box Plot, Histogram, and Violin Plot.

	Rooms	Price	Distance	Postcode	Bedroom2
count	13580.00000	1.358000e+04	13580.00000	13580.00000	13580.00000
mean	2.937997	1.075684e+06	10.137776	3105.301915	2.914728
std	0.955748	6.393107e+05	5.868725	90.676964	0.965921
min	1.000000	8.500000e+04	0.000000	3000.00000	0.000000
25%	2.000000	6.500000e+05	6.100000	3044.00000	2.000000
50%	3.000000	9.030000e+05	9.200000	3084.00000	3.000000
75%	3.000000	1.330000e+06	13.000000	3148.00000	3.000000
max	10.000000	9.000000e+06	48.100000	3977.00000	20.000000

2. Analysis of Numerical Variables

We will analyse the below numerical variables using a Box Plot, Histogram, and Violin Plot.

Numeric features:

```
['Rooms', 'Price', 'Distance', 'Postcode', 'Bedroom2', 'Bathroom', 'Car', 'Landsize',  
 'BuildingArea', 'YearBuilt', 'Latitude', 'Longitude', 'Propertycount']
```

Box Plot

Shows the distribution of price within each variable

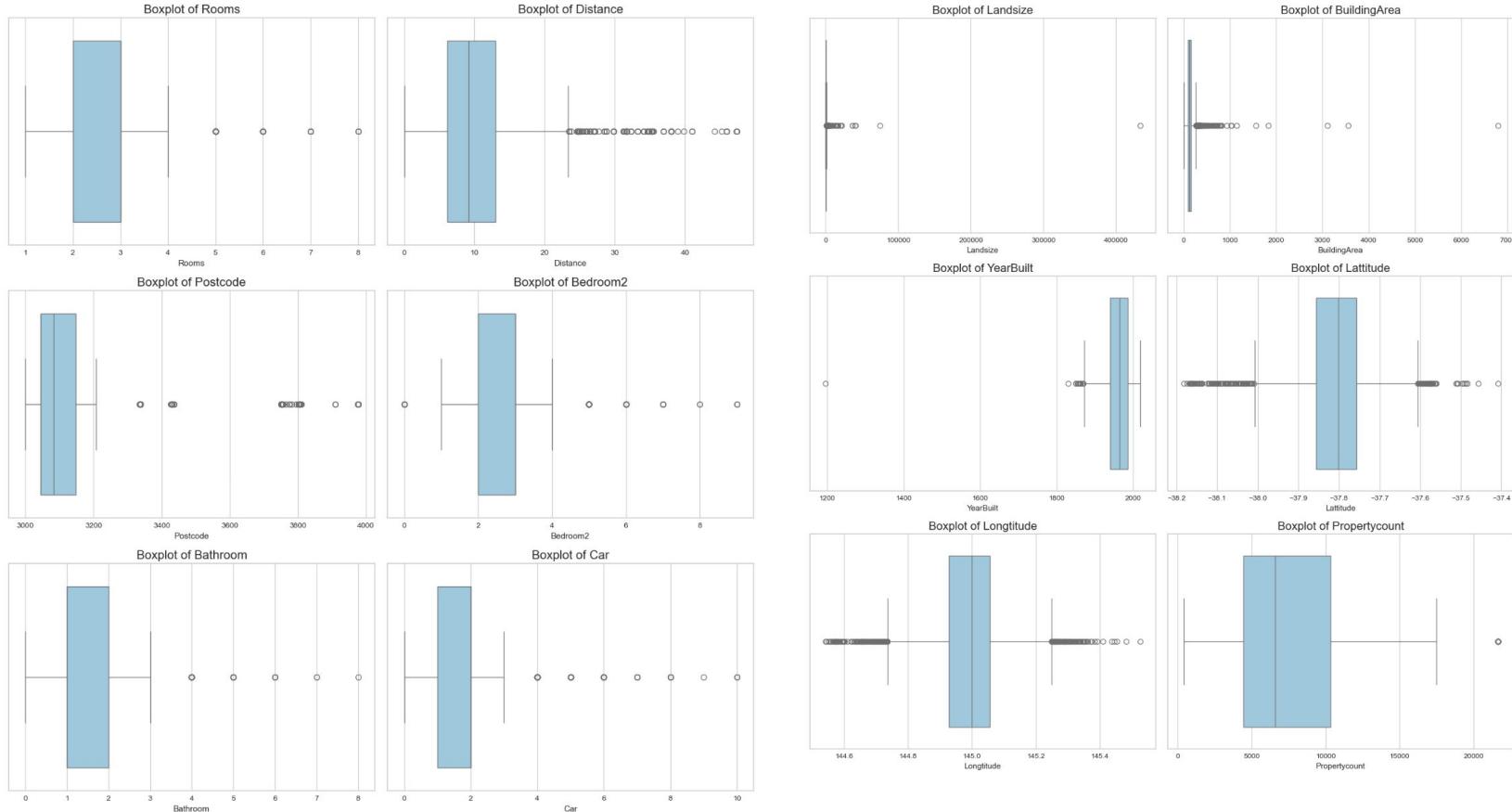
Histogram

Understand the underlying distribution (e.g., normal, skewed) and spread of the data.

Violin Plot

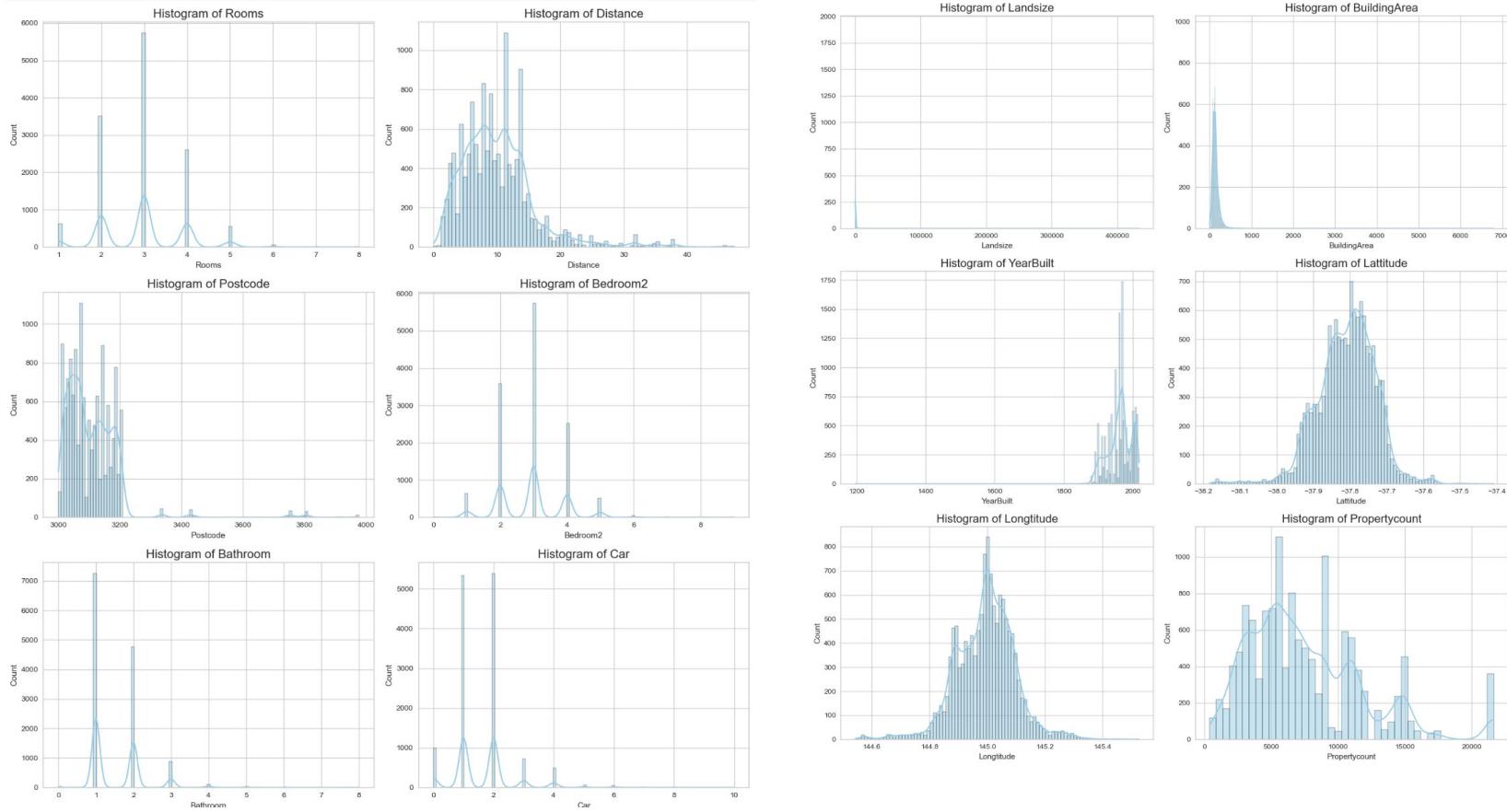
Better visualize the summary statistics (like median and IQR) and the full distribution shape of the variable

Univariate Visualisation: Boxplot



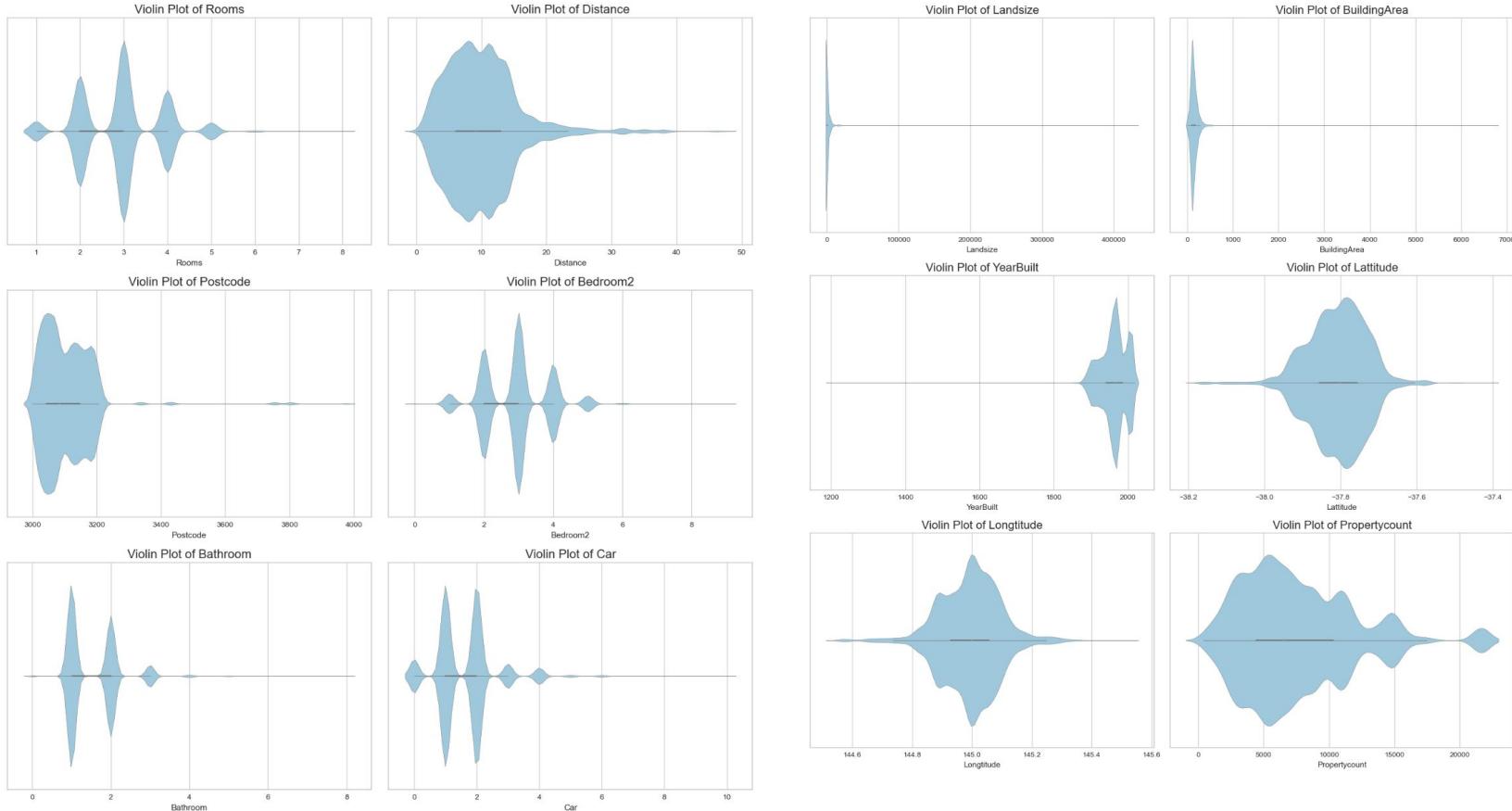


Univariate Visualisation: Histogram



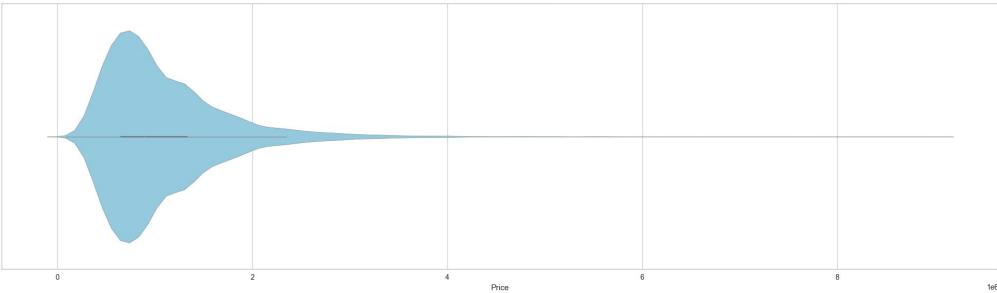
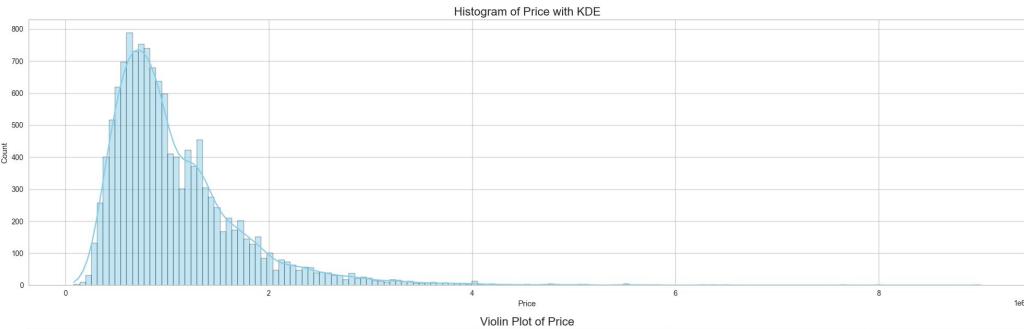
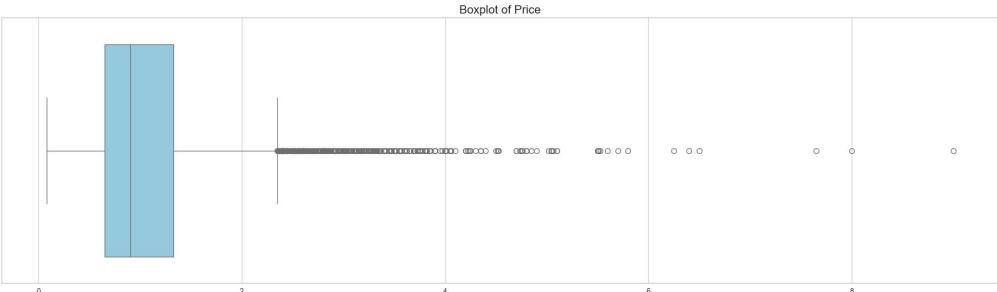


Univariate Visualisation: Violinplot



Univariate Visualisation: Response Variable

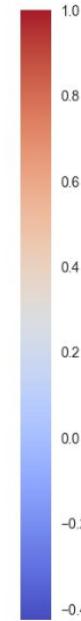
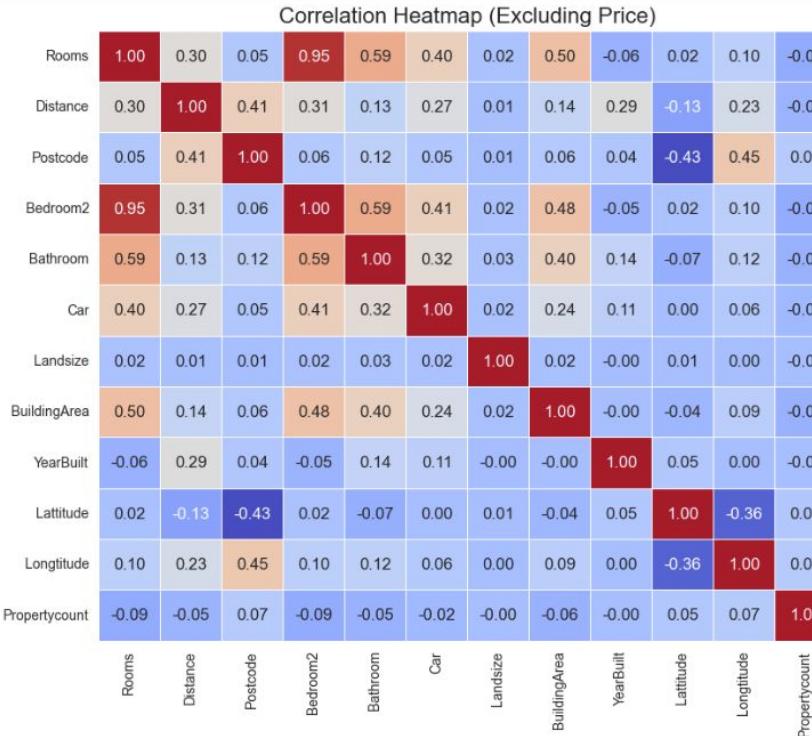
Response Variable: Price





Bivariate Visualisation

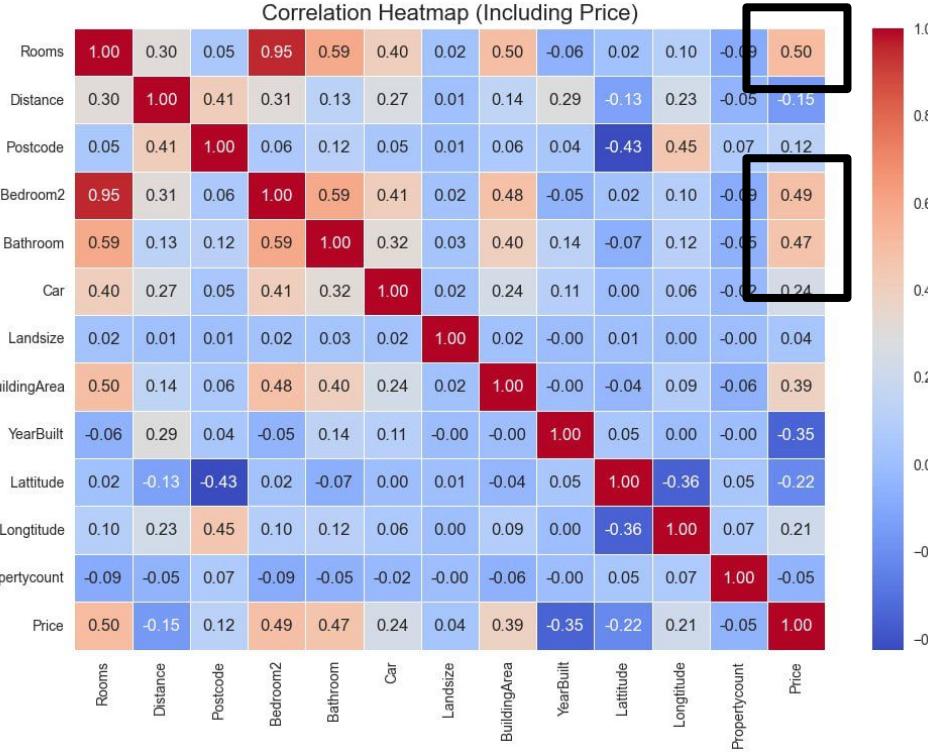
Correlation Heatmap





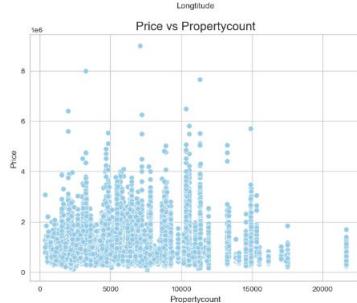
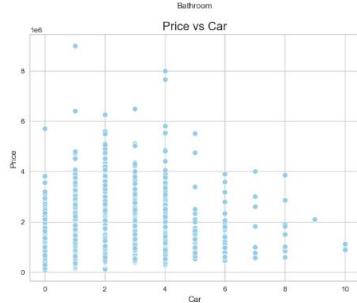
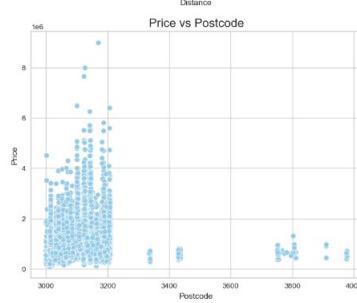
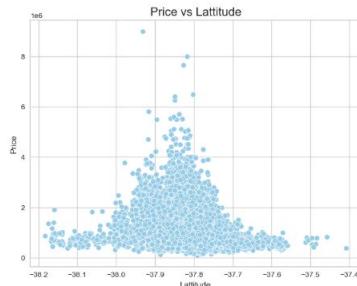
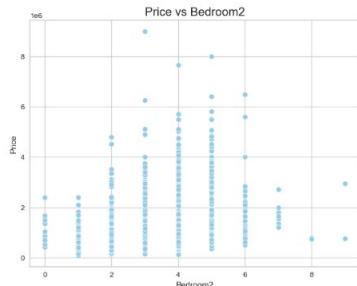
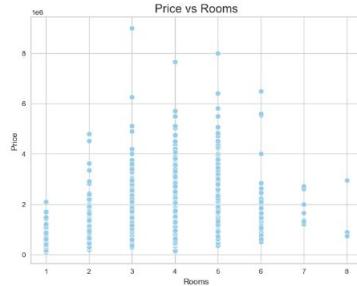
Bivariate Visualisation

Correlation Heatmap





Bivariate Visualisation: Scatter Plot



3. Analysis of Categorical Variables

We will analyse the below categorical variables using a Count Plot, Average Price for each category, as well as a Box Plot.

Categorical features:

```
['Suburb', 'Address', 'Type', 'Method', 'SellerG', 'Date', 'CouncilArea', 'Regionname']
```

Count Plot

Visualizes the frequency and distribution of each category.

Avg Price

Reveal how different categories influence the response variable (Price)

Box Plot

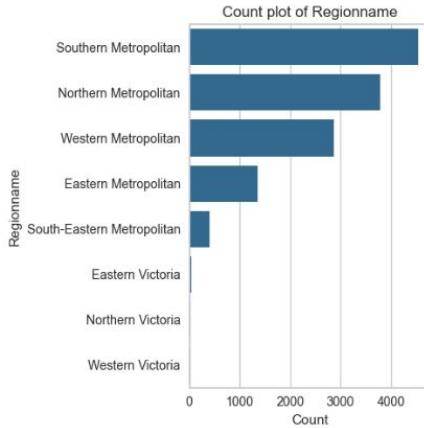
Shows the distribution of price within each category

Univariate Visualization

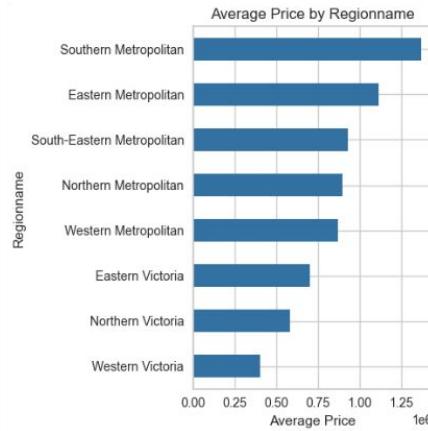
Predictor Variable: Regionname

(using 1 variable as an example)

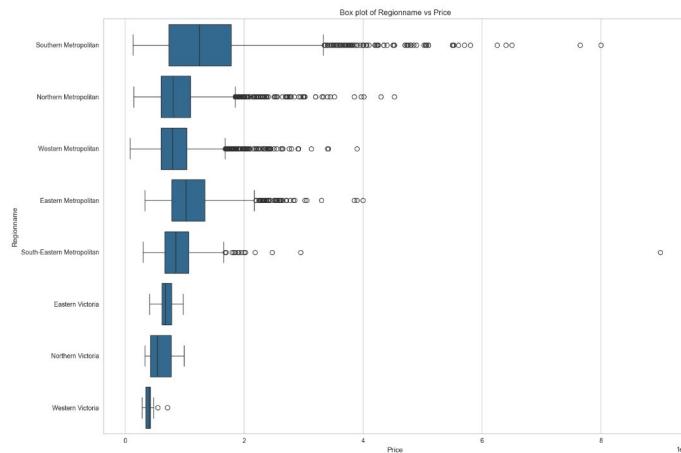
Count Plot



Avg Price

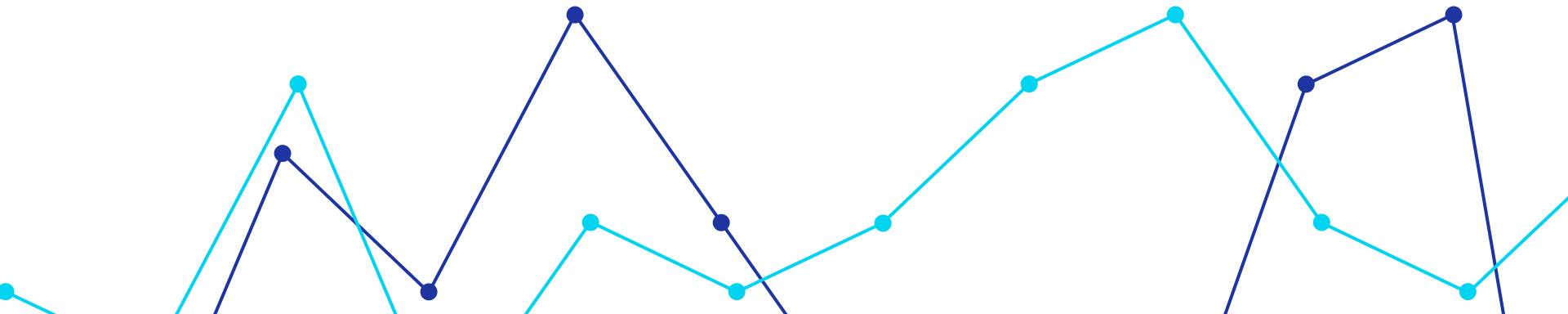


Box Plot



3

Preparation of Variables for Analysis



1. Dimensionality Reduction

How can we reduce the number of unique categories?

		count	unique	top	freq
Suburb	13066	300		Reservoir	359
Address	13066	12874		28 Blair St	3
Type	13066	3		h	9142
Method	13066	5		S	8701
SellerG	13066	263		Nelson	1532
Date	13066	58		27/05/2017	445
CouncilArea	13066	33		nan	1252
Regionname	13066	8	Southern Metropolitan	4547	

1. Dimensionality Reduction

How can we reduce the number of unique categories?



```
def group_rare_categories(df, column, threshold=0.01):
    # Calculate category counts
    category_counts = df[column].value_counts(normalize=True)

    # Identify categories with frequency less than the threshold
    rare_categories = category_counts[category_counts < threshold].index

    # Replace rare categories with "Other"
    df[column] = df[column].apply(lambda x: 'Other' if x in rare_categories else x)

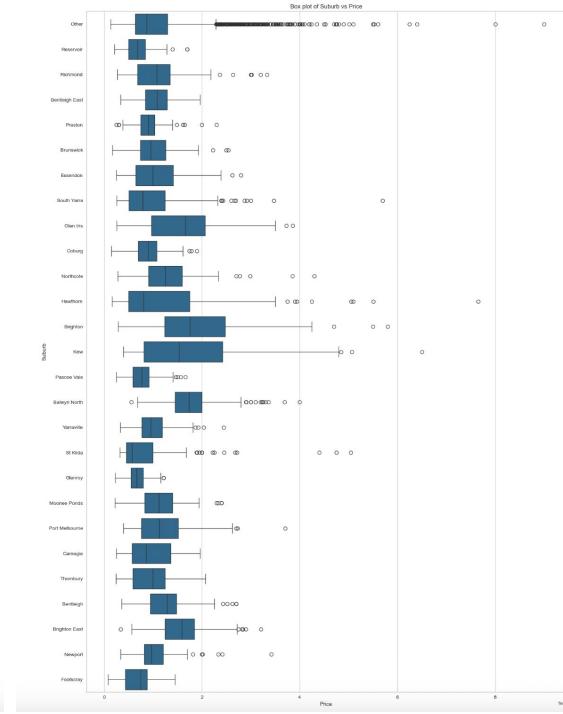
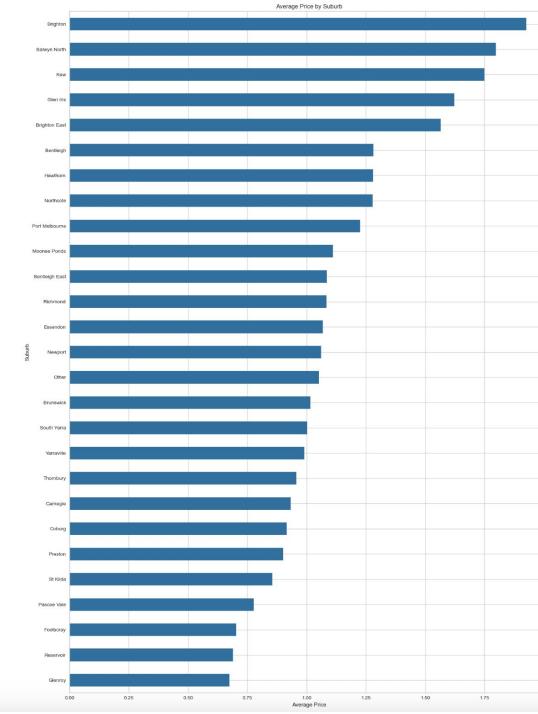
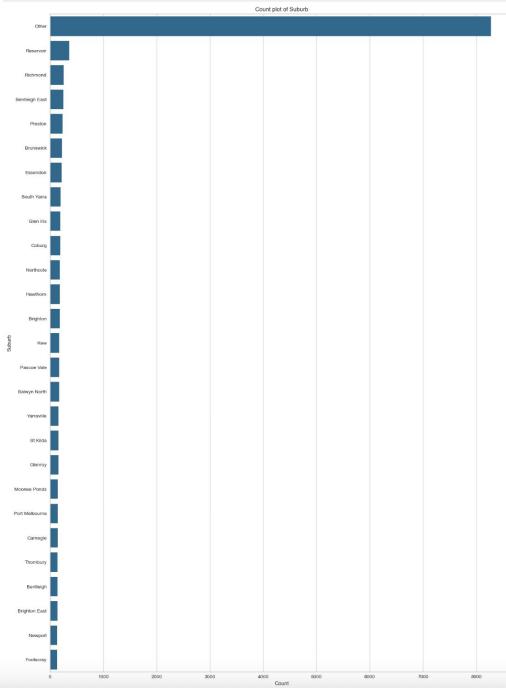
    return df
```

ENGINEERING

Dimensionality REDUCTION

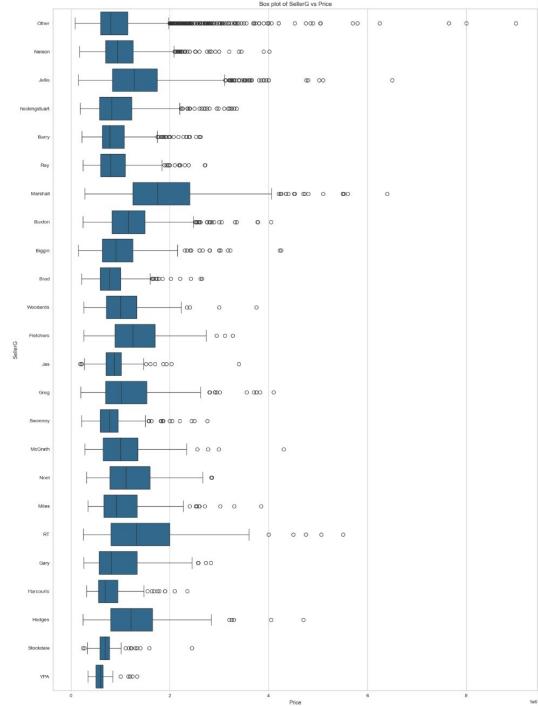
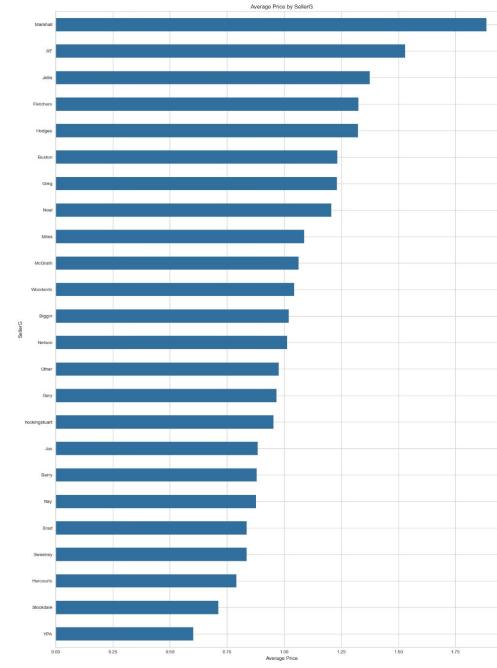
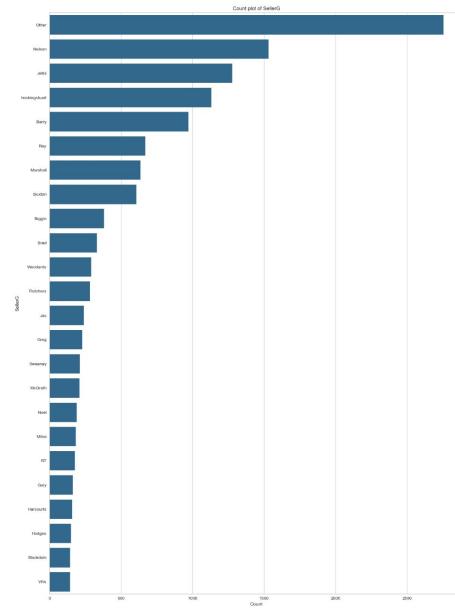
1. Dimensionality Reduction: Suburb

How can we reduce the number of unique categories?



1. Dimensionality Reduction: SellerG

How can we reduce the number of unique categories?



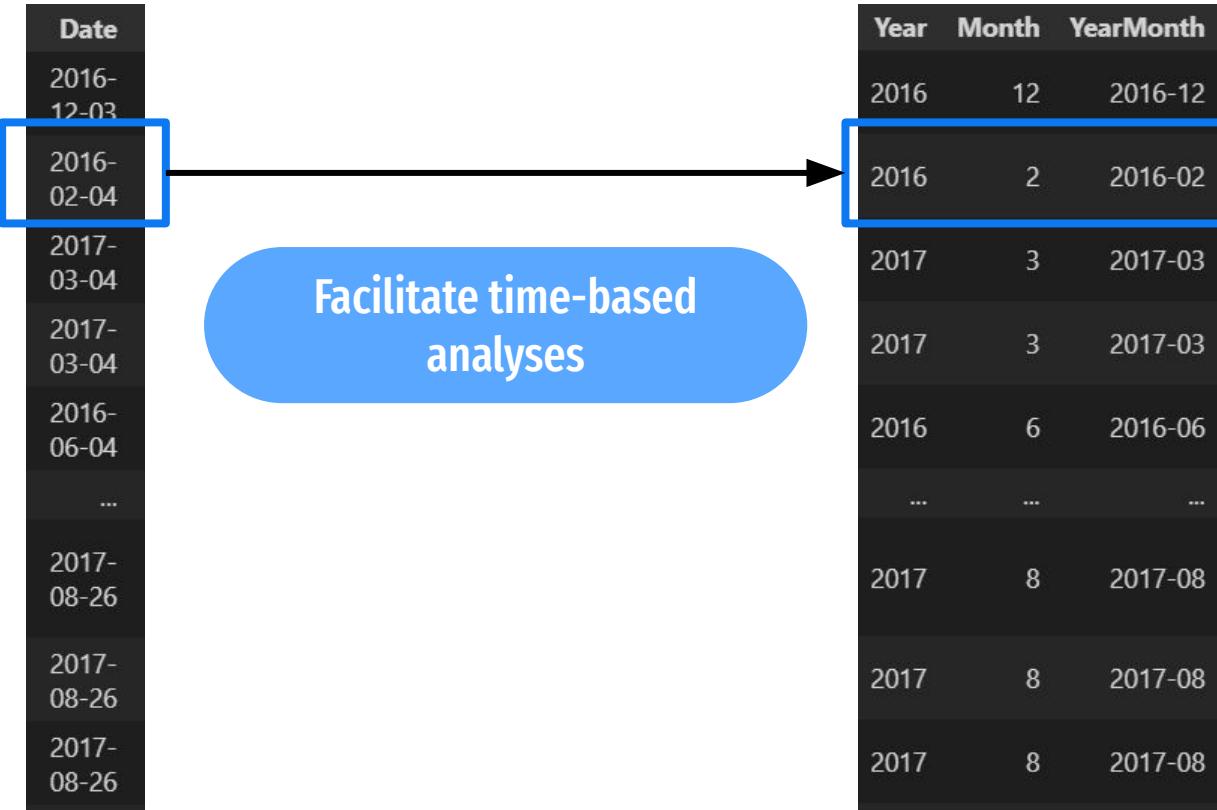
1. Dimensionality Reduction: Feature Elimination

How can we reduce the number of unique categories?



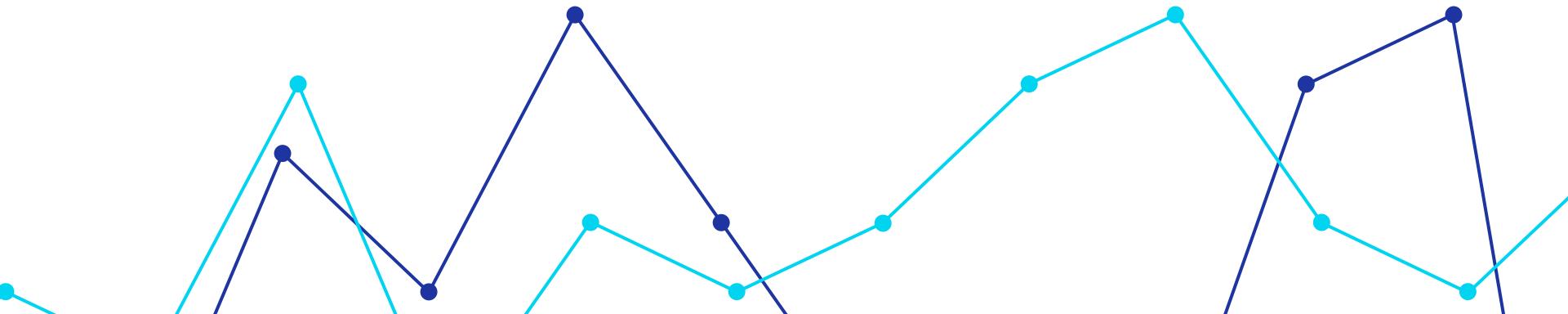
```
housingdata = housingdata.drop(columns=['Address'])
```

2. Feature Engineering of 'Date' Variable



4

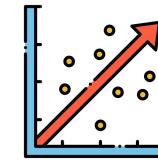
Core Analysis



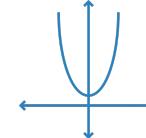
1a. Analysis of Price Predictors

Which features have the highest impact on price based on feature importance?

Linear Regression



Quadratic Regression



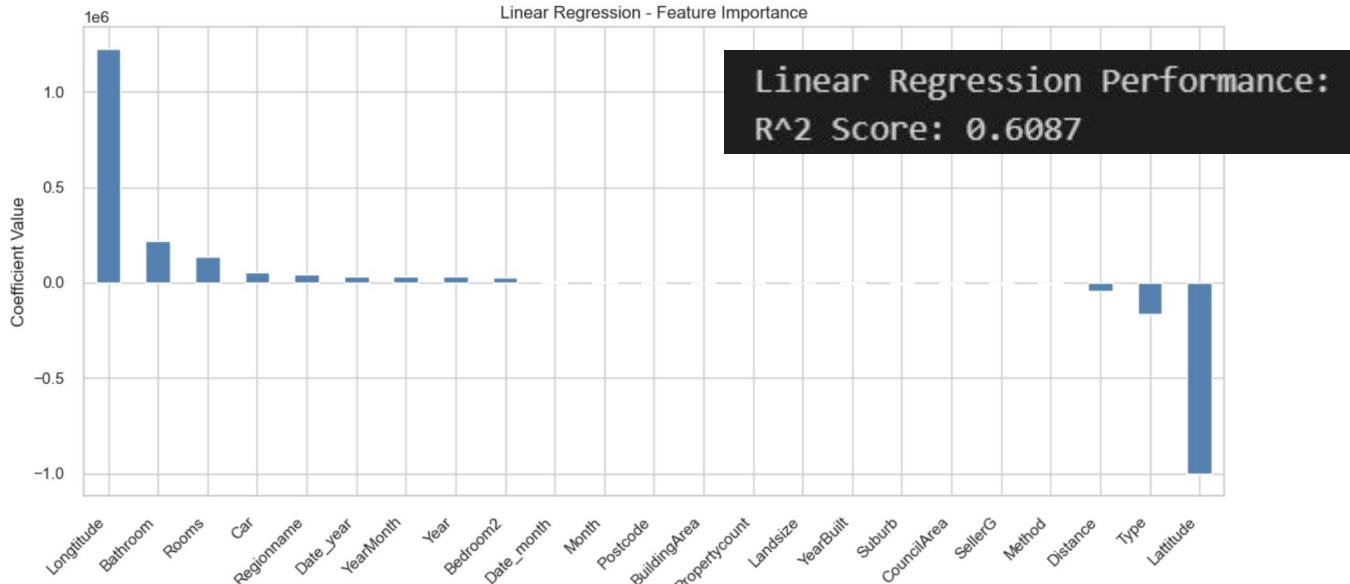
Random Forest Regression



1a. Analysis of Price Predictors

Which features have the highest impact on price based on feature importance?

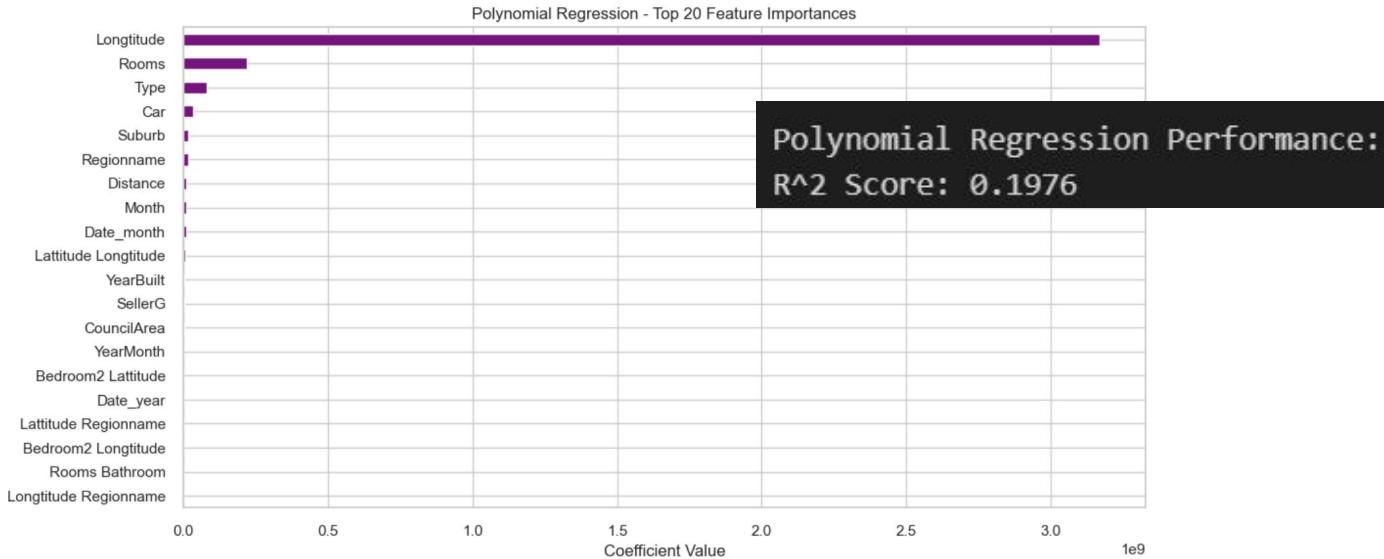
Linear Regression



1a. Analysis of Price Predictors

Which features have the highest impact on price based on feature importance?

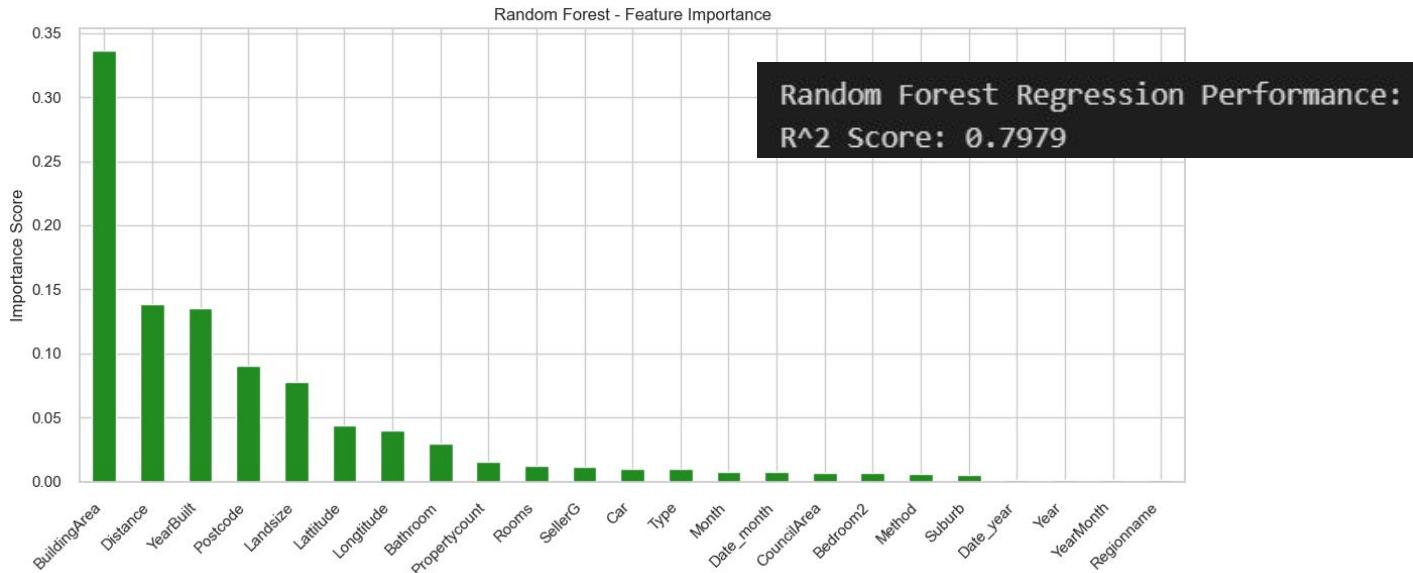
Quadratic Regression



1a. Analysis of Price Predictors

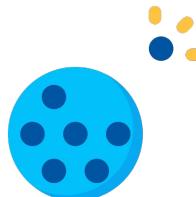
Which features have the highest impact on price based on feature importance?

Random Forest Regression



1b. Improvements to the Model

Which features have the highest impact on price based on feature importance?



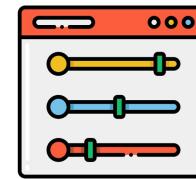
Outliers

- Remove outliers



Feature Engineering

- Encoding variables



Hyperparameter Tuning

```
param_grid = {  
    'n_estimators': [100, 200],  
    'max_depth': [None, 10, 20],  
    'min_samples_split': [2, 5],  
}
```

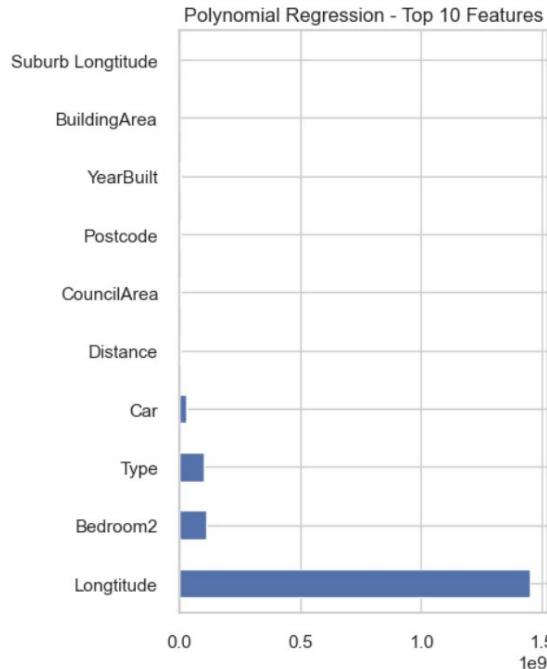
1b. Improvements to the model

Which features have the highest impact on price based on feature importance?

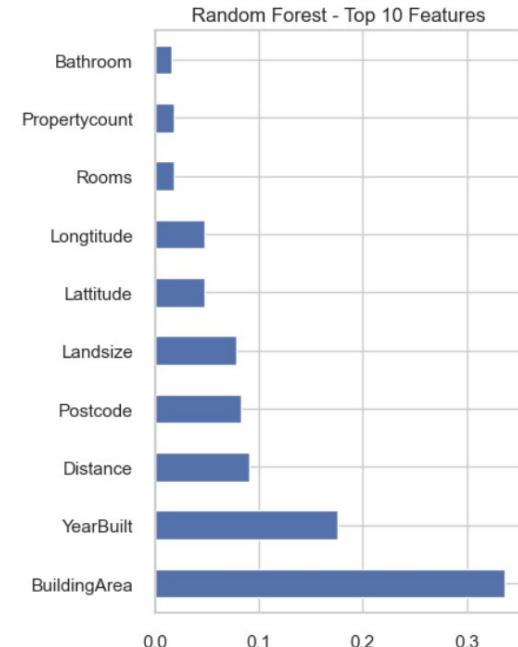
Fitting 5 folds for each of 12 candidates, totalling 60 fits



Improved Linear Regression Performance:
 R^2 Score: 0.6575



Improved Polynomial Regression Performance:
 R^2 Score: 0.7450



Tuned Random Forest Performance:
 R^2 Score: 0.8000



K-Means

ANALYSIS

CLUSTERING

2. K-Means Clustering of Regions

How has the price trend varied across different regions over time, and can we identify emerging high-value areas?

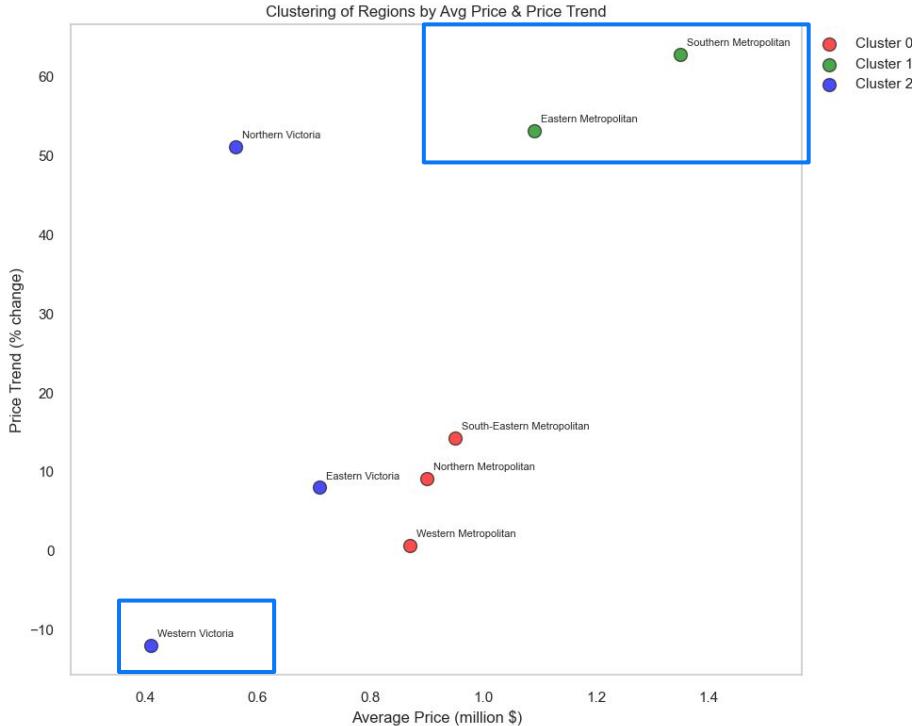
Average Price

% Change in Price

Regionname	FirstDate	LastDate	AvgPrice_M	PctChange	Cluster
Southern Metropolitan	2016-01-28	2017-09-23	1.35	62.85	1
Eastern Metropolitan	2016-02-04	2017-09-23	1.09	53.17	1
Northern Victoria	2017-05-27	2017-09-23	0.56	51.14	2
South-Eastern Metropolitan	2016-02-04	2017-09-23	0.95	14.28	0
Northern Metropolitan	2016-02-04	2017-09-23	0.90	9.12	0
Eastern Victoria	2017-05-27	2017-09-23	0.71	8.02	2
Western Metropolitan	2016-02-04	2017-09-23	0.87	0.69	0
Western Victoria	2017-05-27	2017-09-23	0.41	-11.91	2

2. K-Means Clustering of Regions

How has the price trend varied across different regions over time, and can we identify emerging high-value areas?



2. Improvements to the Model

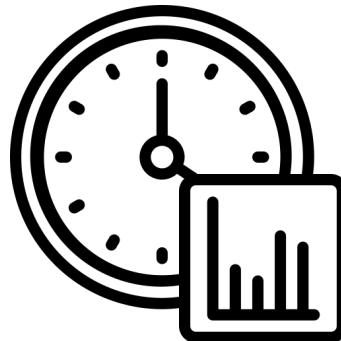
Current results only reflect the first date and last date, as well as absolute price change, but instead it should show the changes of the clusters over time

Richer Dataset

Long-term Trends



Time-series Clustering



3. Influence of Variables on Prediction Accuracy

How does property age and size influence prediction accuracy in models?

Random Forest Regression



Prediction Errors

```
Price  
count 1.358000e+04  
mean 1.075684e+06
```

```
Initial MAE: $176,788.78  
Initial MSE: 1.07e+11  
RMSE: $326584.15
```

16.5% of average Price

30.5% of average Price

3. Influence of Variables on Prediction Accuracy

How does property age and size influence prediction accuracy in models?

Property Age

Property Age Error Analysis:

PropertyAge	
(-2.001, 9.0]	71067.410175
(9.0, 19.0]	79499.810786
(19.0, 37.0]	66585.052251
(37.0, 46.0]	65962.906458
(46.0, 51.0]	64726.198076
(51.0, 57.0]	73008.901682
(57.0, 67.0]	94876.985238
(67.0, 86.0]	131123.198594
(86.0, 106.0]	122819.075716
(106.0, 821.0]	143251.234171

Property Size

Landsize Impact:

Smaller properties ($<177m^2$) MAE: \$66,342.32
Larger properties ($>650m^2$) MAE: \$121,294.82
Error increase: 82.8%

BuildingArea Impact:

Smaller properties ($<94m^2$) MAE: \$58,343.37
Larger properties ($>164m^2$) MAE: \$142,290.94
Error increase: 143.9%



3. Improvements to the Model

How does property age and size influence prediction accuracy in models?

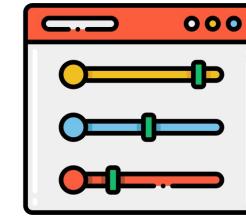


Feature Engineering

```
housingdata['Rooms_Bathroom']
```

```
housingdata['Landsize_BuildingArea']
```

- Better capture nuance



Hyperparameter Tuning

```
param_grid = {  
    'n_estimators': [200, 300],  
    'max_depth': [15, 25, None],  
    'min_samples_split': [2, 5]  
}
```

- Reduce overfitting

3. Improvements to the Model

How does property age and size influence prediction accuracy in models?

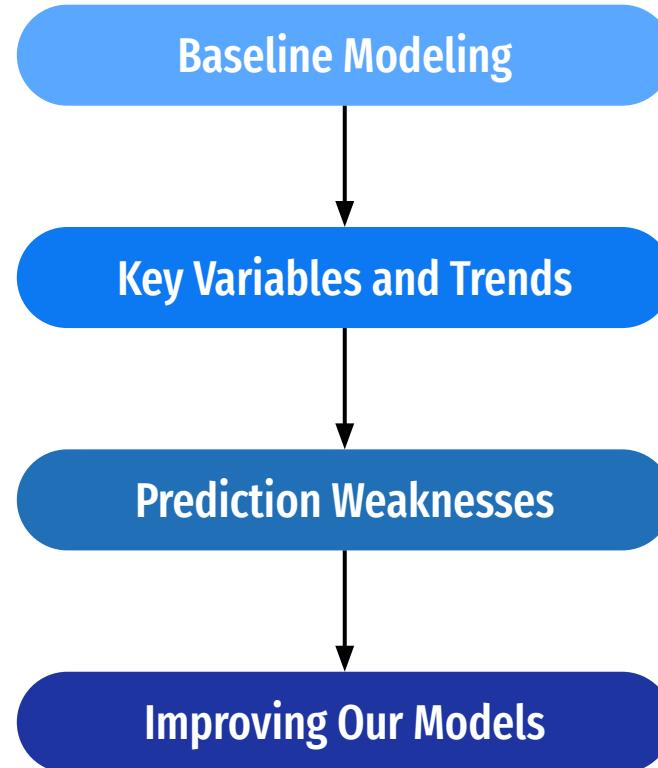
Improved MAE: \$80,438.03
Improved MSE: 2.34e+10
MAE Improvement: 54.5%



Benchmark Conditions

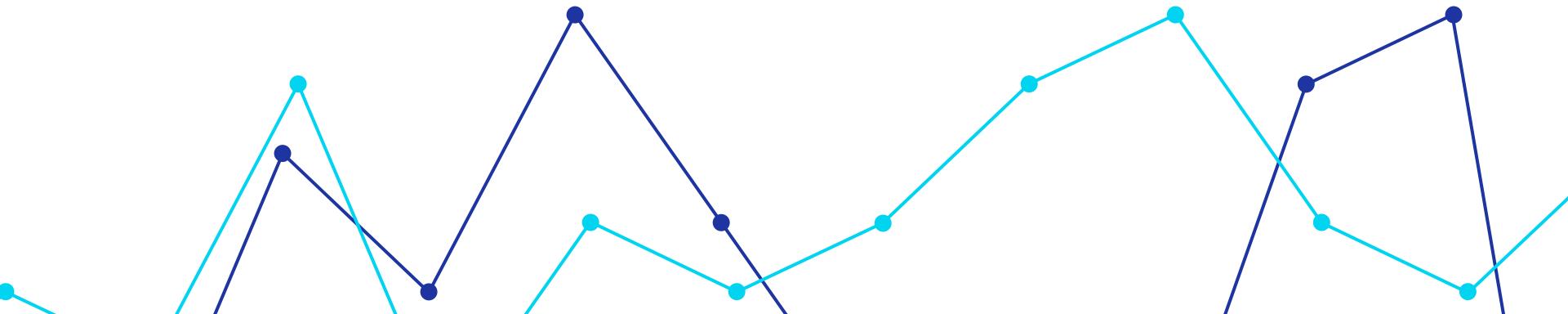
Further Refinement

Wrapping Up



4

Conclusions

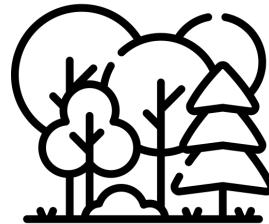


Outcome

1

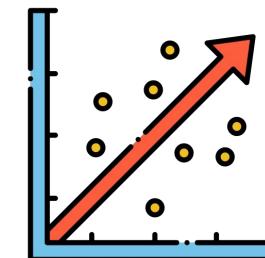
Which features have the greatest impact on price?

Influential
Predictors
Of Price



Random Forest Regression

1. BuildingArea
2. Location-based Features



Linear Regression

Outcome

2

How has the price trend varied across different regions over time, and can we identify emerging high-value areas?

K-Means Clustering

High-Growth Regions

1. Southern Metropolitan
2. Eastern Metropolitan

Declining Regions

1. Western Victoria

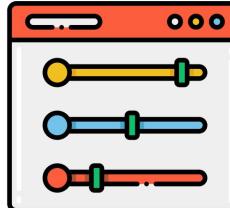
Outcome

3

How does property age and size influence prediction accuracy in models?

Model Weaknesses

1. Older Properties
2. Larger Properties



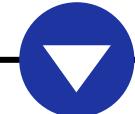
Feature Engineering

Hyperparameter Tuning

**Model
Robustness
and
Reliance**

Conclusions

- 1 Stakeholders should take note of size and age-related price deviations.
- 2 Stakeholders should monitor and anticipate growth in certain regions.



Proactive Interventions in Singapore

Thank You!

REP2 Group 7
Lex Tan Pengqin
Tan Yu Xiu
Lim Jun, Shawn

