Melbourne House Price Prediction Model

Capstone Project 4

Minija Kannath

Want to buy a house in Melbourne?

Contents

1. Introduction
2. Methodology
3. Process Workflow
4. Results
5. Conclusions
6. Future Opportunities

Introduction

I am a data analyst working for a real estate company . I have a personal interest in this Melbourne dataset as both my kids are studying there

Skills

scikit learn

python™

SQL

Power BI

Excel

Representing Organization

Ocean Wave Real Estate

Real service, Real Solutions, Real Estate...

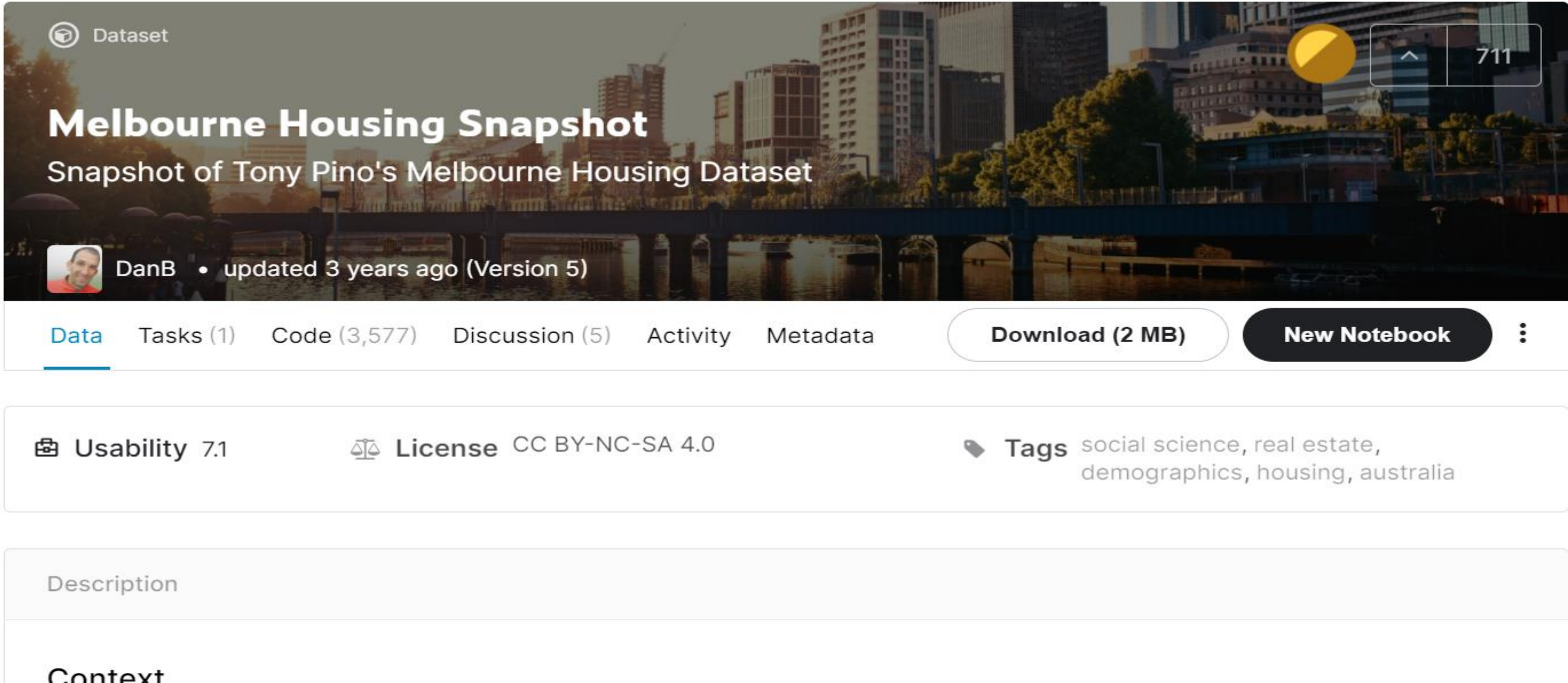
Target Audience

- Real Estate investors
- Mortgage lenders and Home insurers
- Prospective buyers

Business Objective

The objective of this project is to build a model to predict the housing prices in different Regions of Melbourne

Data Collection



Data Collection

The data set consists 13378 records and 21 features. Some of the major features with descriptions are:

- Rooms: Number of rooms
- Price: Price in dollars
- Type: h – house/cottage/villa u - unit,/duplex; t - townhouse.
- Distance: Distance from CBD
- Regionname: (West, North -West, North, North- east …etc)
- Bedroom2 : Number of Bedrooms
- Bathroom: Number of Bathrooms
- Car: Number of car spots
- Landsize: Land Size
- BuildingArea: Building Size
- CouncilArea: Governing council for the area

Methodology

Model

Baseline Model: Linear Regression

Alternative Model: Random Forest, Decision tree , Gradient Boosting regressor

Metrics

Root Mean Squared Error (RMSE)
(R-squared) metric

Tools

seaborn jupyter NumPy matplotlib scikit learn Pandas

Data Cleaning

Drop/Fill up N.A

Car	62
Landsize	0
BuildingArea	6450
YearBuilt	5375
CouncilArea	1369
Lattitude	0
Longtitude	0
Regionname	0

Bathroom	0
Car	0
Landsize	0
BuildingArea	0
YearBuilt	0
CouncilArea	0
Lattitude	0
Longtitude	0
Regionname	0

13378 Records

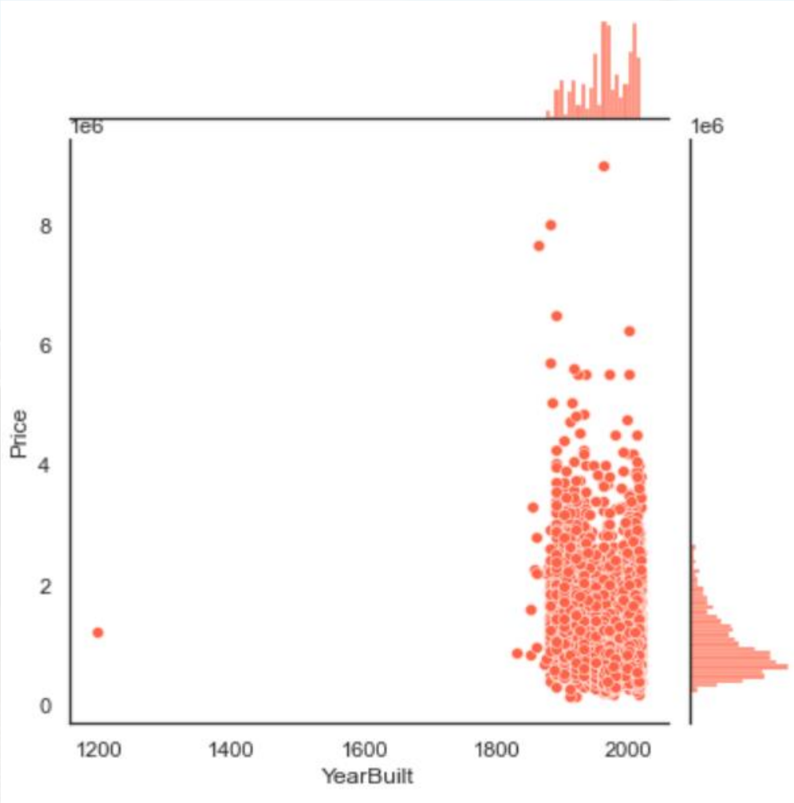
6196 Records

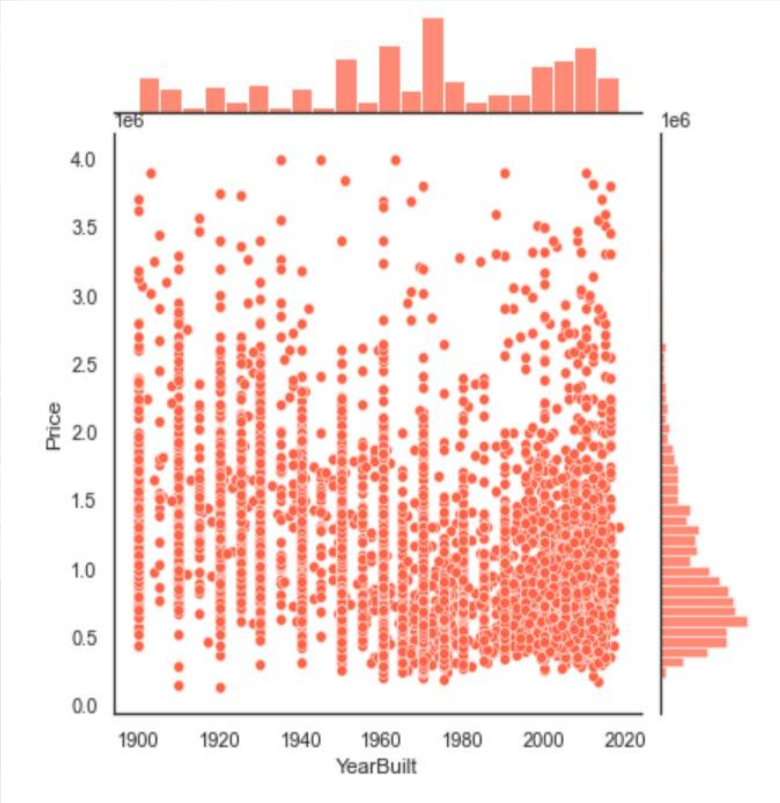
Data preparation

- Features which are insignificant to the business objective of this project are dropped.eg('Method', 'Seller G',' Property count', "Address")
- Postcode, Latitude, and Longitude: Dropped these features, because it is highly correlated to Suburb

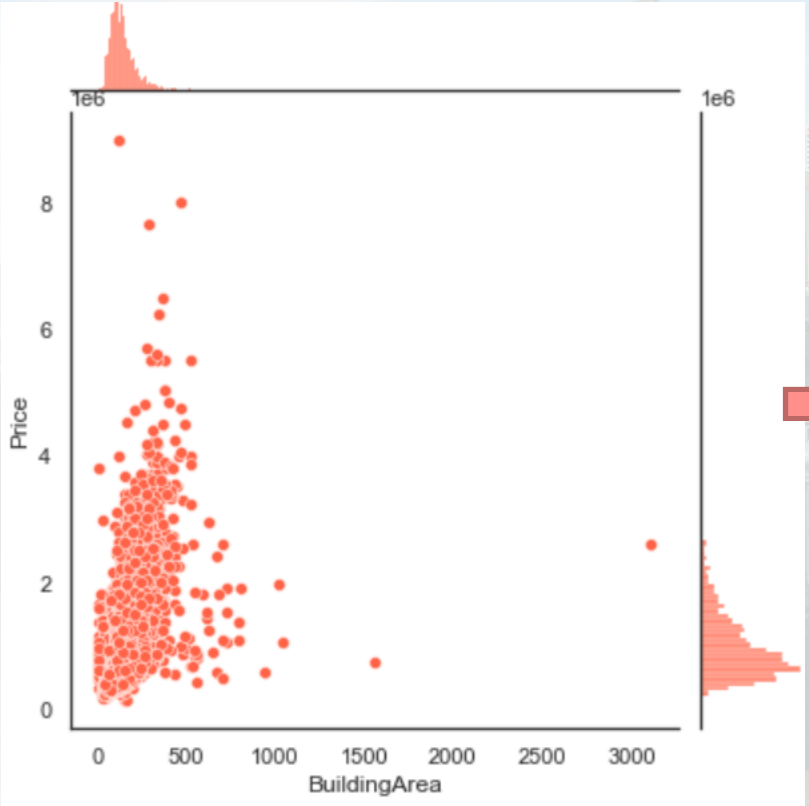
21 Features → 14 Features

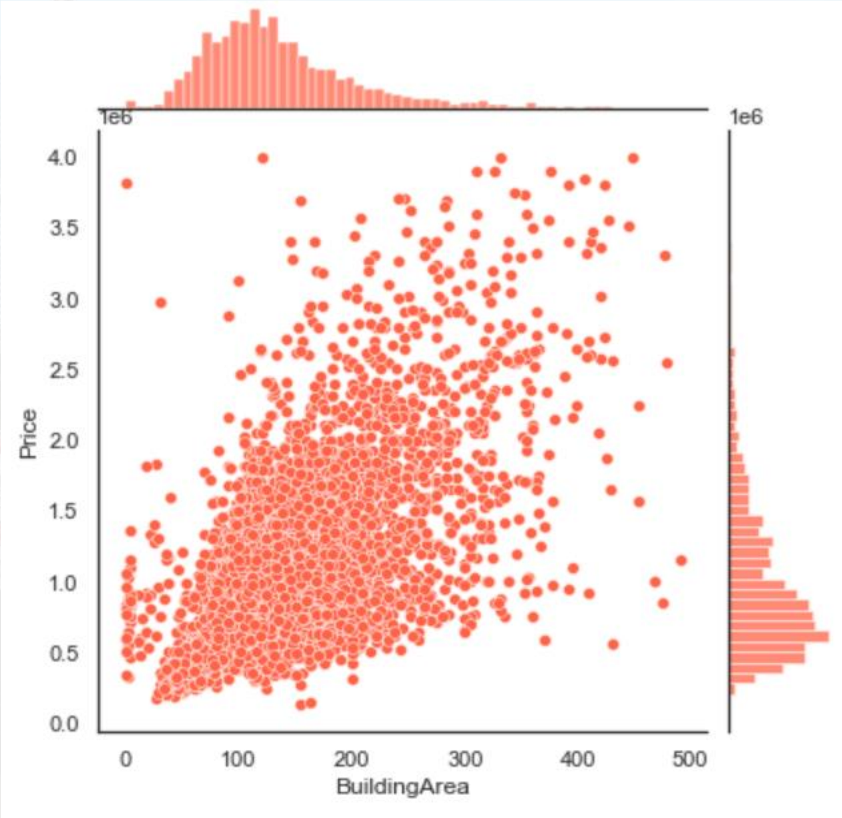
EDA/Outlier Removal



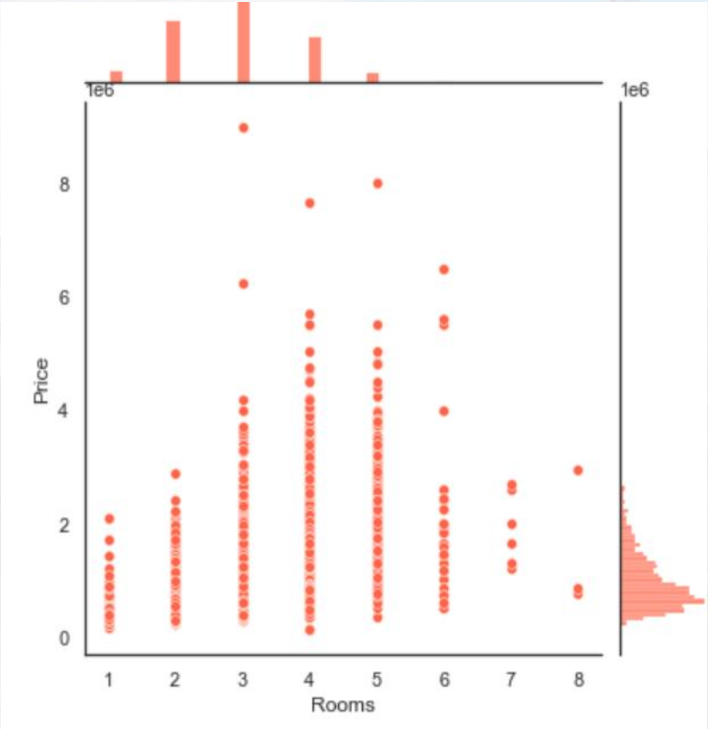


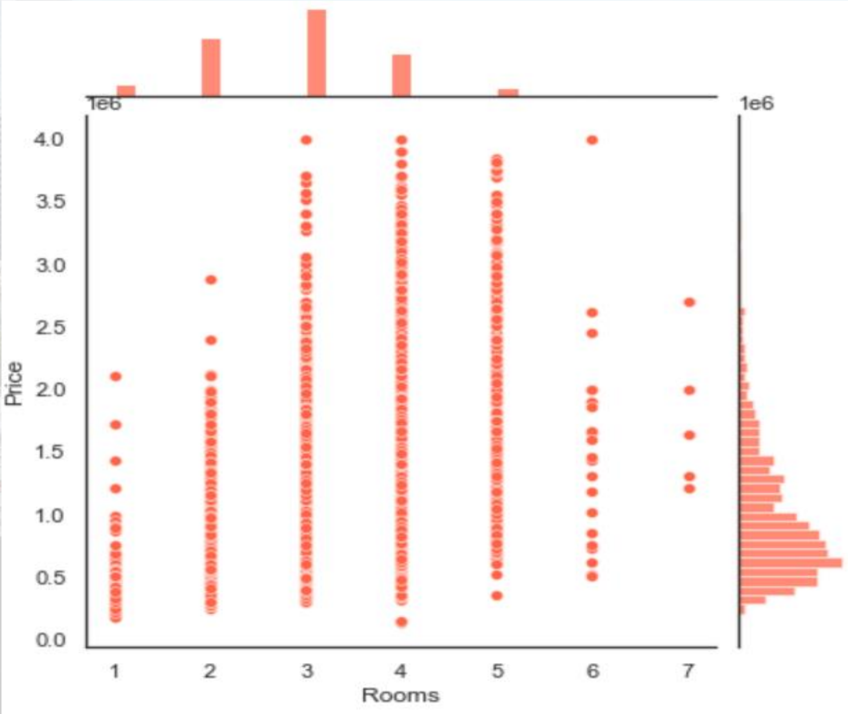
EDA/Outlier Removal





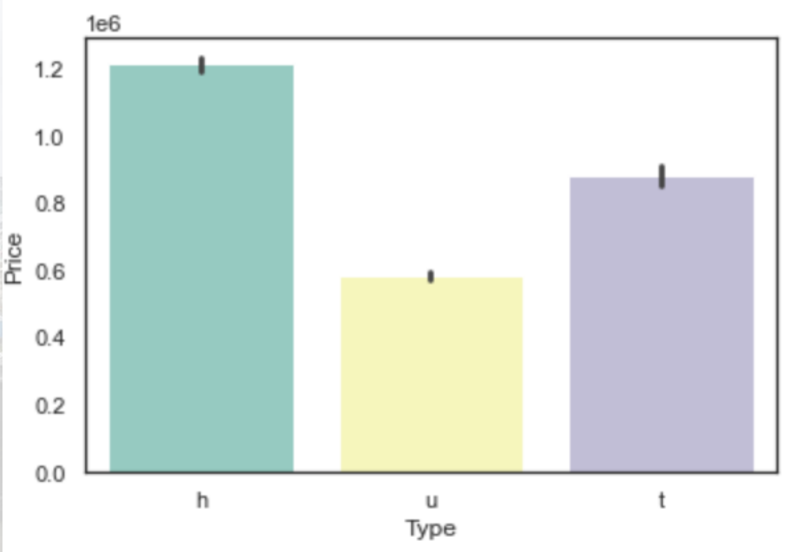
EDA/Outlier removal

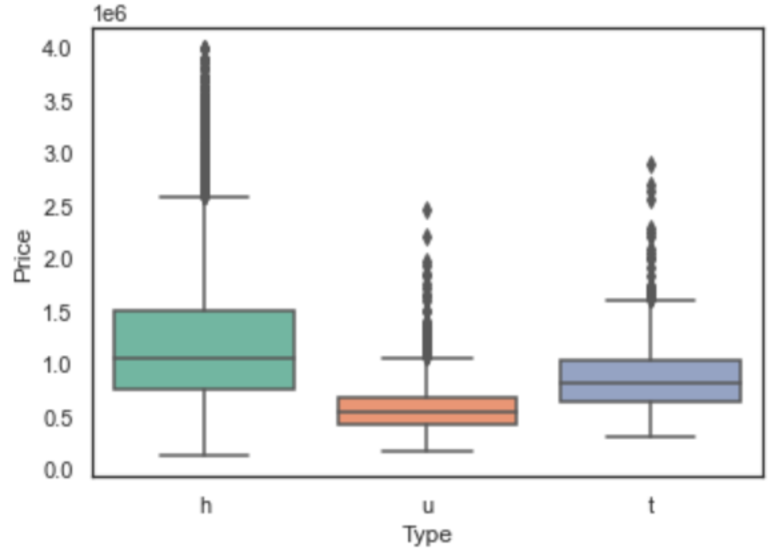




EDA /Data visualization

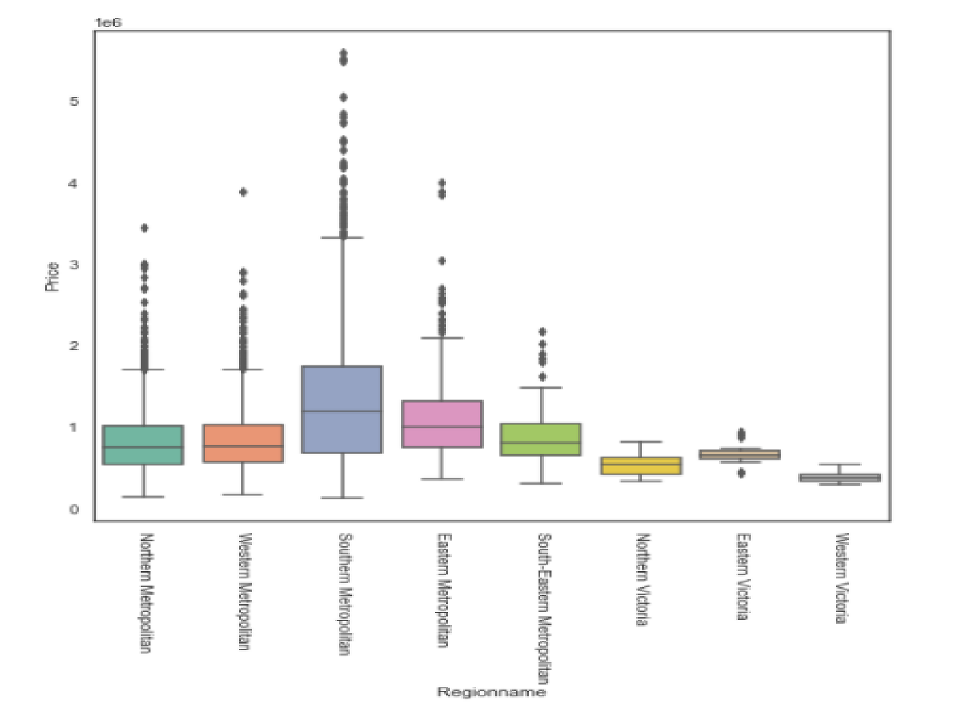
To check the relationship between the categorical feature ‘Type’ and the target value

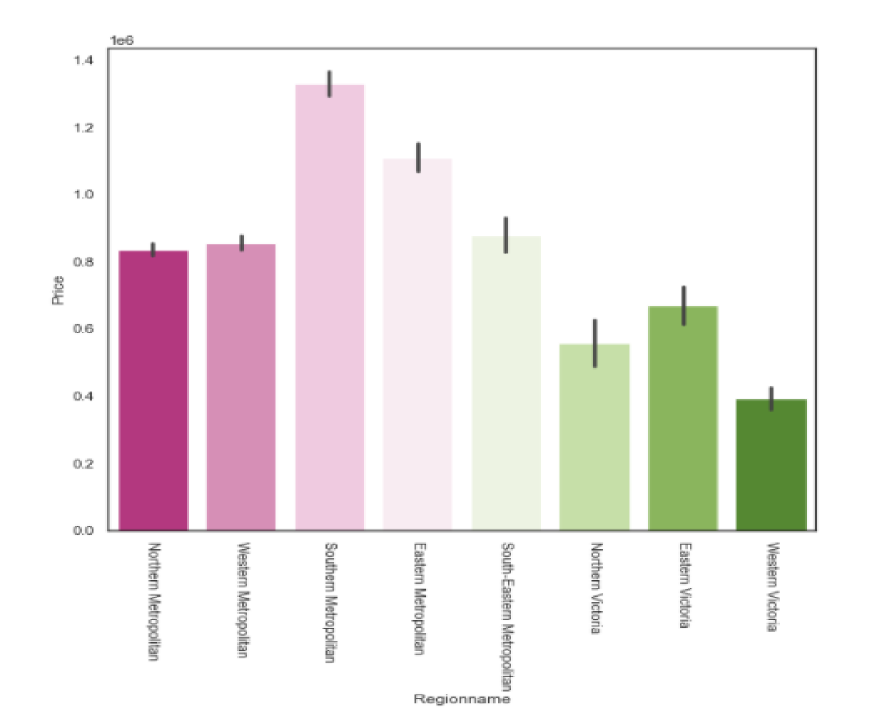




‘ h’(houses/villa/ cottage) is most expensive then comes ‘t’(town house)

EDA /Data visualization



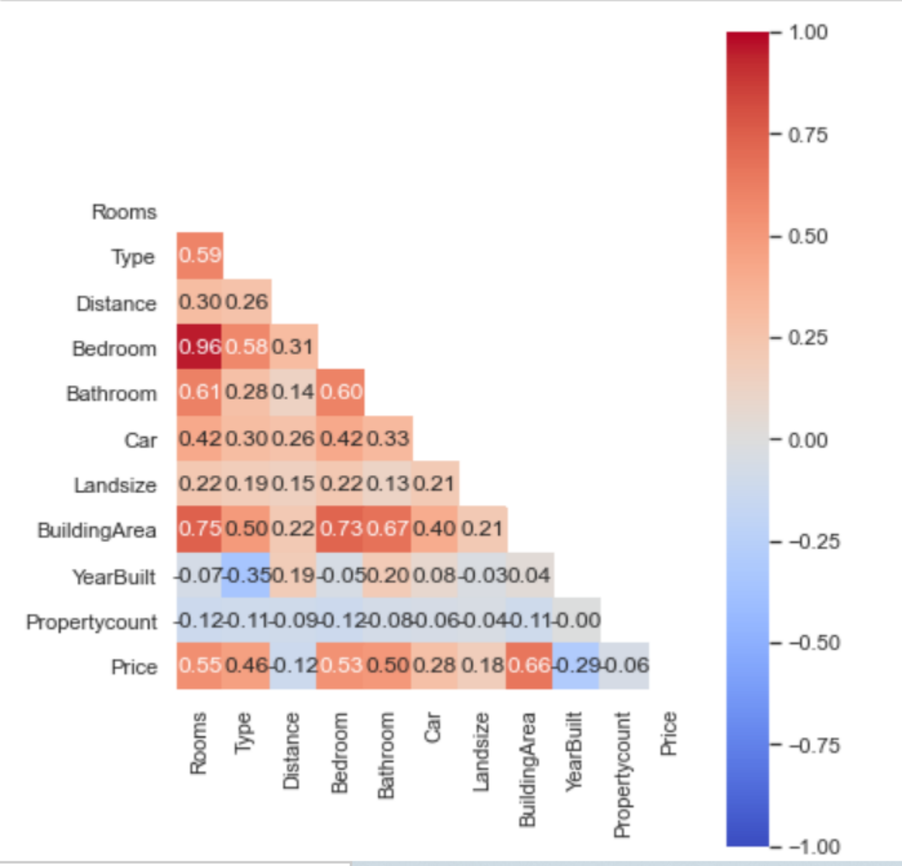


Southern Metropolitan region has the most expensive properties and is the most expensive region on average. Eastern Victoria, Northern Victoria, and Western Victoria, are the most affordable regions

Data Analysis & Feature Engineering

Only one feature from the highly correlated group will be selected to avoid bias

Features which shows weak correlation with the target value are dropped.



Encoding Categorical data/Feature Engineering

One Hot Encoding is used to transform the categorical feature ‘Regionname’ into binary form of representation. This is then used to rank each Region based on its property value.

One Hot Encoding is used to transform the categorical feature “Type” into binary form. This is then ranked based on house value

Data transformation

- Split the data into training and testing datasets
  test_size=0.2

Data Normalization

- **max min normalization** : Linear Regression, Decision Tree , Random Forest & Gradient Boosting Regressor

Machine Learning Model

MODEL

Baseline Model: Linear Regression

Alternative Model: Random Forest, Decision tree,Gradient Boosting Regressor

Machine Learning Model Training & Evaluation

Baseline Model: Linear Regression

Hyperparameter: (copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

```
Cross Validation Scores:
[0.6934026  0.66191834 0.74435393 0.7126399  0.64433696 0.58308386
 0.61434027 0.71637971 0.65266804 0.6337499  0.66169052 0.58030251]

Mean Score:
0.6582388783557784

RMSE:
0.1102766810170264
```

Machine Learning Model Training & Evaluation

Observation :
According to the R-squared score only 65.48% of the variance in the dependent variable is explained by the model.

Solution :
Alternative Model

Machine Learning Model Training & Evaluation

Alternative Model: Decision Tree model

Hyperparameter: ((criterion='friedman_mse',splitter="best"))

```
Cross Validation Scores:
[0.53733655 0.55859217 0.60627891 0.63453898 0.51005072 0.41926062
 0.57324694 0.61162811 0.61901889 0.55627301 0.55654188 0.58687296]

Mean Score:
0.5641366449273874

RMSE:
0.1246543123611106
```

Machine Learning Model Training & Evaluation

Observation :
According to the R-squared score only 56.4% of the variance in the dependent variable is explained by the model.

Solution :
Alternative Model

Machine Learning Model Training & Evaluation

Alternative Model: Random Forest Regressor

Hyperparameter: (bootstrap= True,max_depth= 20,min_samples_split= 5)

```
Cross Validation Scores:
[0.7754547  0.72324277 0.80829213 0.81362842 0.69111208 0.71922525
 0.7023033  0.79045465 0.77881318 0.75823634 0.75801469 0.78237441]

Mean Score:
0.7584293257270006

RMSE:
0.08891334851303938
```

Machine Learning Model Training & Evaluation

Observation :
According to the R-squared score only 75.8% of the variance in the dependent variable is explained by the model.

Solution :
Alternative Model

Machine Learning Model Training & Evaluation

Alternative Model: Gradient Boosting Regressor

Hyperparameter: (loss ='ls', max_depth=7)

```
Cross Validation Scores:
[0.77709918 0.73223902 0.81561186 0.83208297 0.7097644  0.69098729
 0.71625484 0.79752736 0.79519812 0.78003969 0.78768496 0.79722856]

Mean Score:
0.7693098544116702

RMSE:
0.08757514756619188
```

Machine Learning Model Training & Evaluation

Observation :

According to the R-squared score 76.8% of the variance in the dependent variable is explained by the model .Which is the best score so far

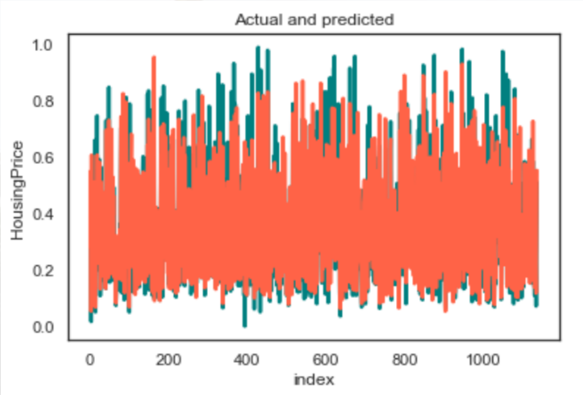
Finding the best model and Hyperparameter tuning

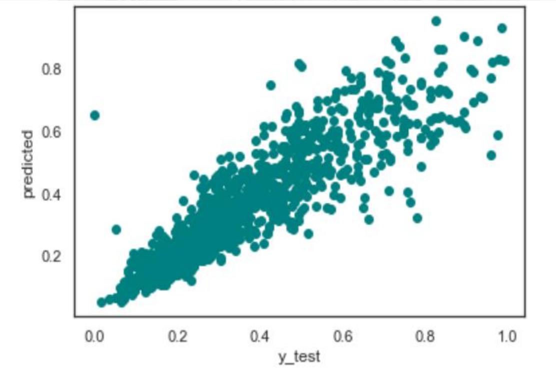
We will use GridSearchCV to find the best model and the best hyperparameter

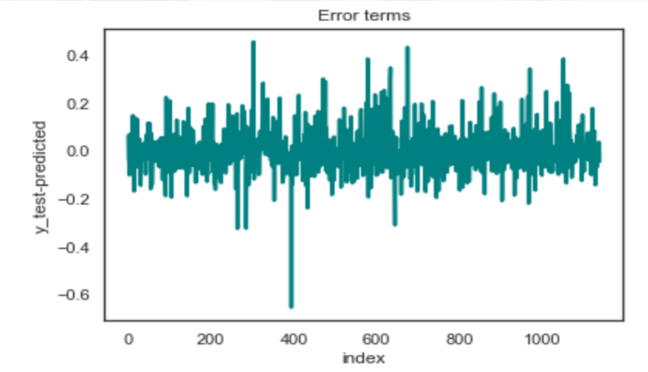
	model	best_score	best_params
0	linear_regression	0.688468	{'normalize': True}
1	random_forest_regression	0.798249	{'bootstrap': True, 'max_depth': 20, 'min_samp...
2	GradientBoostingRegressor	0.808952	{'loss': 'ls', 'max_depth': 5}
3	decision_tree	0.629313	{'criterion': 'mse', 'splitter': 'best'}

Results

Actual vs predictions and Error terms:







Interesting insights

- Median prices for houses are over $1M, townhomes are $800k - $900k and units are $500k.

- Median prices in the Metropolitan Region are higher than that of Victoria Region - with Southern Metro being the area with the highest median home price (~$1.3M).

- With an average price of $1M, historic homes (older than 50 years) are valued much higher than newer homes in the area, but have more variation in price

- Most homes in the dataset have 4 or 5 rooms.

- There is a negative correlation between Distance from Melbourne's Central Business District (CBD) and Price. The most expensive homes ($2M or more) tend to be within 20km of the CBD..

Conclusion

Gradient Boosting regressor is the best model with an accuracy of 80% to build this house prediction model and Root mean square error is .0875..

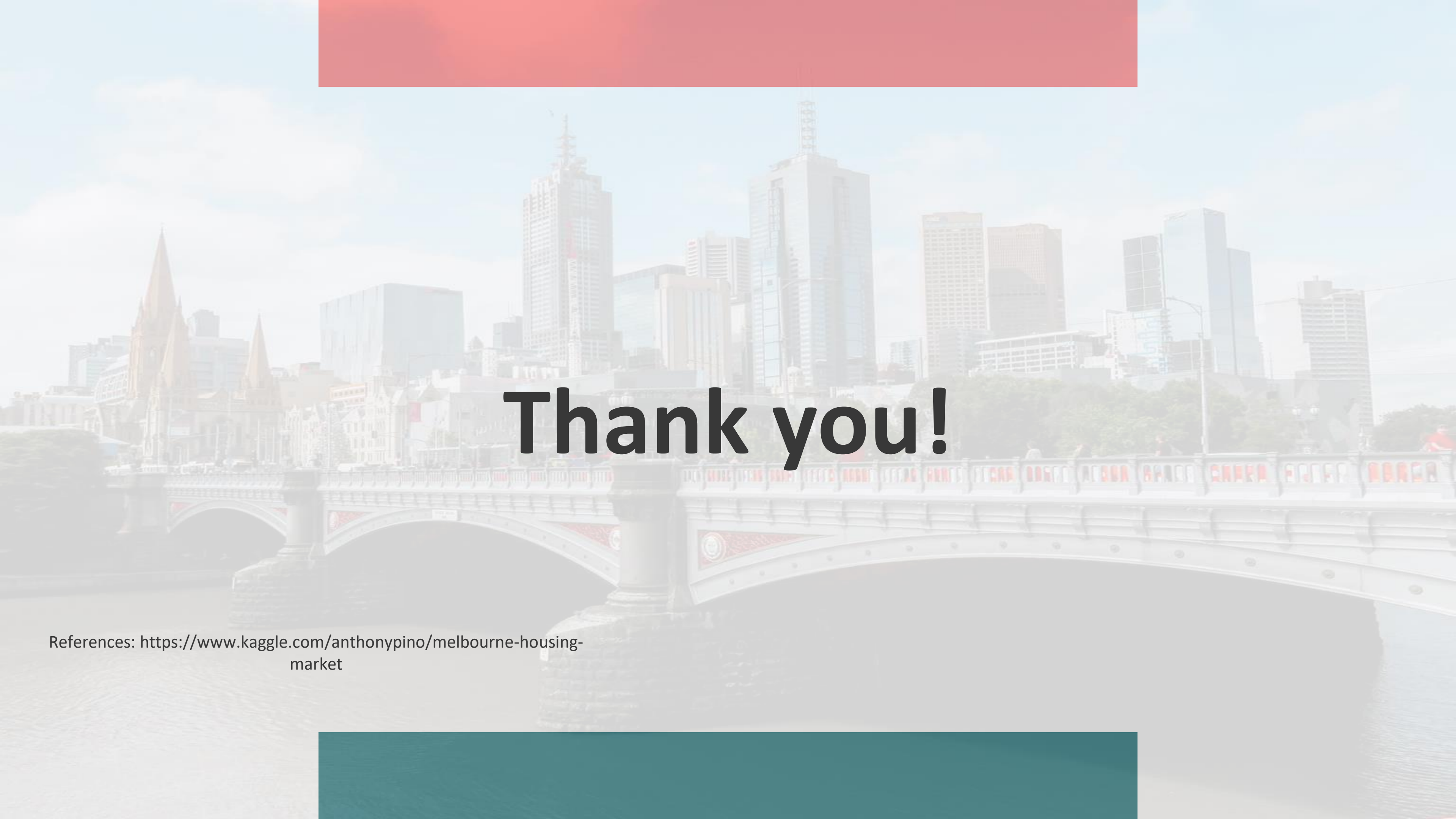
Best Hyper Parameter is ('loss': 'ls', 'max_depth’: 5)

Future Opportunities

In future I will try to get some more additional information on neighbourhood like:

- Schools in neighbourhood
- Access to shops
- Transport
- Details about traffic around the area

Also, another model can be created to give best split percentages to get maximum price by the zip code of the land.



Thank you!

References: https://www.kaggle.com/anthonypino/melbourne-housing-market