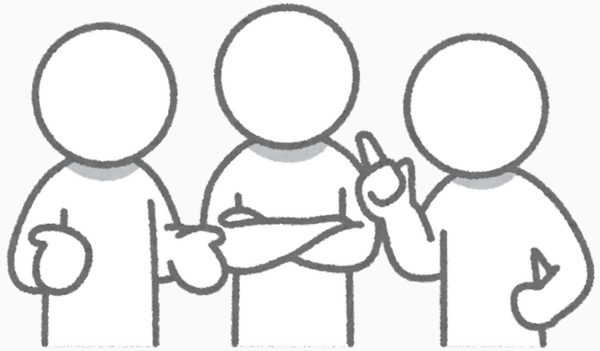
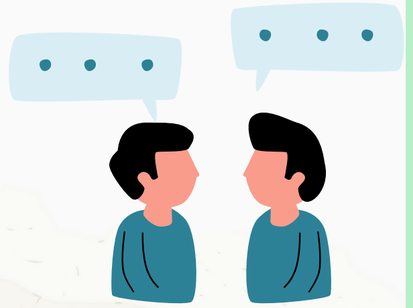
SC1015 Mini Group Project

U2320468E Shammas
U23222601C Ethan
U2320828A Max

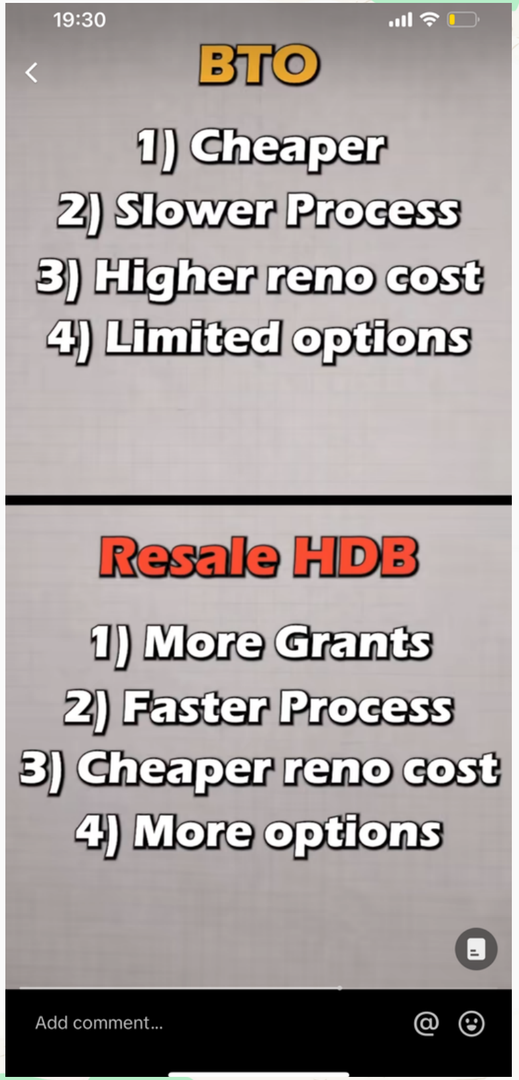
INSPIRATION



- Chit chat with our friends
- Discuss about our future after graduating
- Topic of buying a house and how it is recently a hot conversation topic due to the housing market being inflated



This is a snippet from a TikTok discussing about the differences between BTO and Resale



Problem Analysis

Recently, housing has been an issue in SIngapore

- Some people feel that the waiting time for BTO flats are too long so they are looking to purchase Resale flats
- However, Resale flats are not fixed in their prices as they are not directly sold by HDB but rather by houseowners/real estate agents
- So how does one plan their finance to purchase a Resale flat?

Solution

- Our team has come up with a solution to analyse the factors influencing Resale Flat prices and then trying to predict a Resale Flat's price based on previous trends
- We used the CSV file found from Data.gov, Singapore's Open Data Platform

INFORMATION

3

9

2

The CSV file contains the following information:

- Location (Hot-encoded into Central, North, South, East, West)
- Flat type (1 = 1 room flat ... 6 = Exec appartment)
- Storey range (1 = lower level, 2 = upper level)
- Floor Area squaremetre
- Remaining Lease (the decimal indicates the months left (.05 = 5 months or .11 = 11 months)

EXPLORATORY DATA ANALYSIS

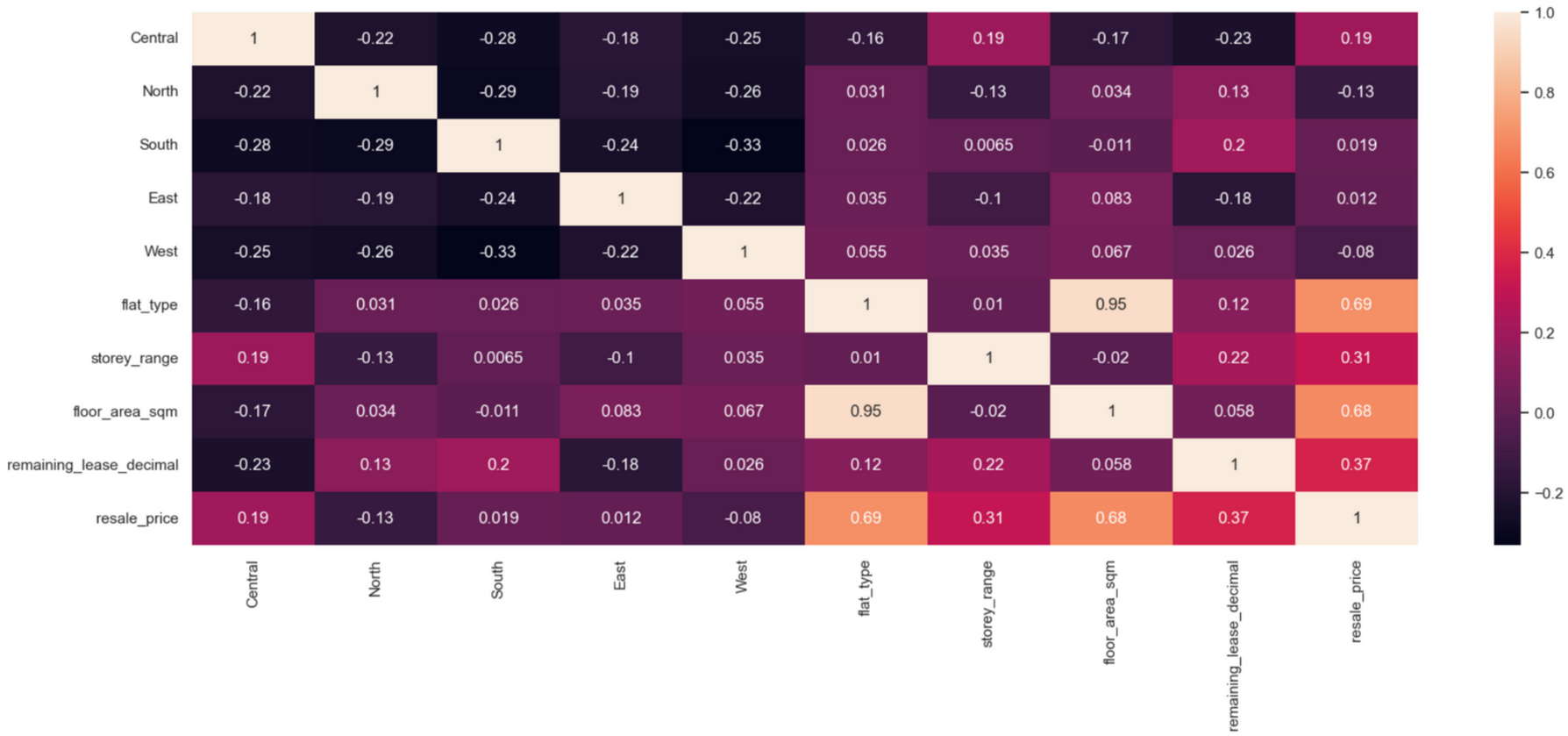
- There is a total of 25,205 rows of data across 10 columns

```
   Central  North  South  East  West  flat_type  storey_range  floor_area_sqm  \
0        0      0      1     0     0          2             1            44.0
1        0      0      1     0     0          2             1            49.0
2        0      0      1     0     0          2             1            44.0
3        0      0      1     0     0          2             1            44.0
4        1      0      0     0     0          3             1            64.0

   remaining_lease_decimal  resale_price
0                    55.05      267000.0
1                    53.06      300000.0
2                    54.01      280000.0
3                    54.01      282000.0
4                    62.01      508000.0
(25205, 10)
```

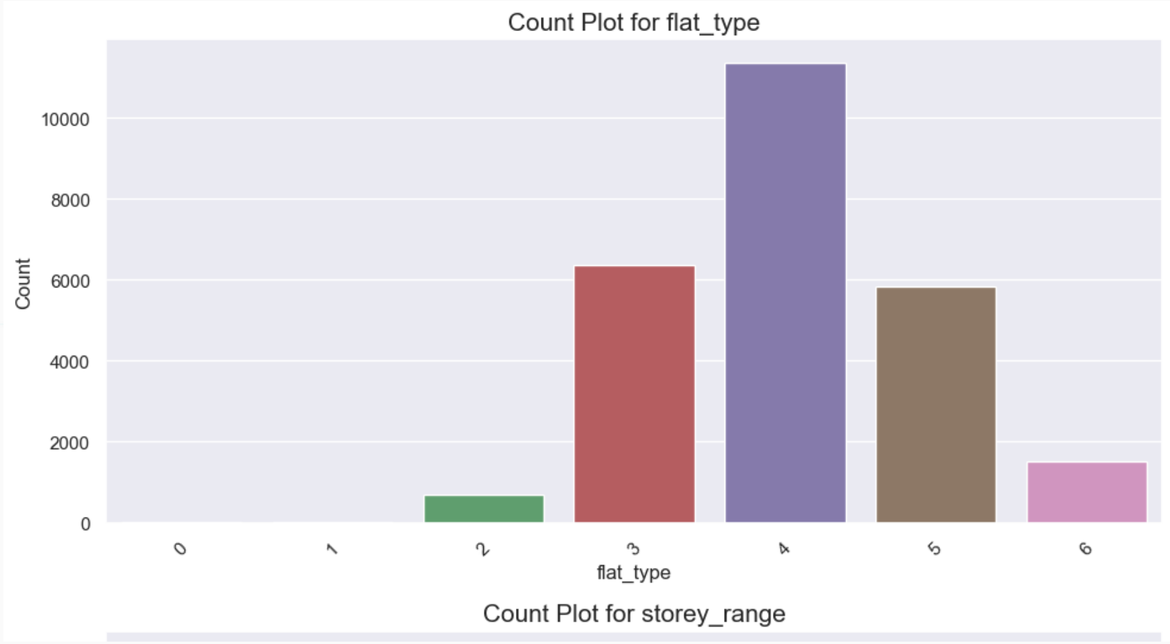
EDA

- The Correlation Heat Map

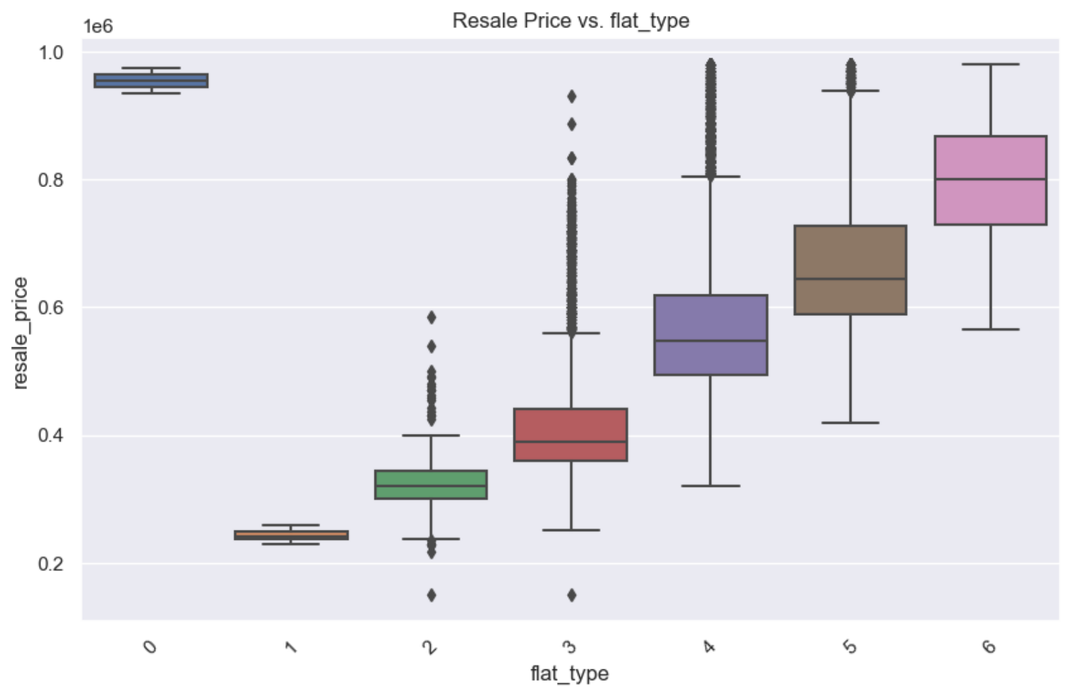


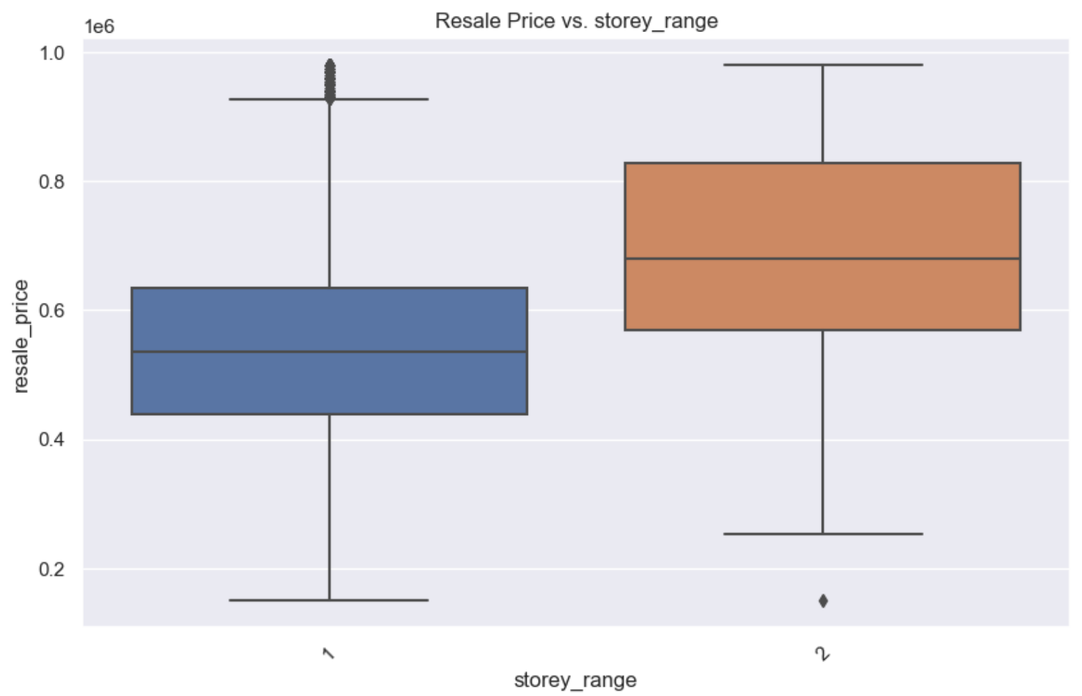
EDA

Looking at the count plot, it seems that the 4-room flat is the most sought after flat as it has the highesy frequency in our dataset

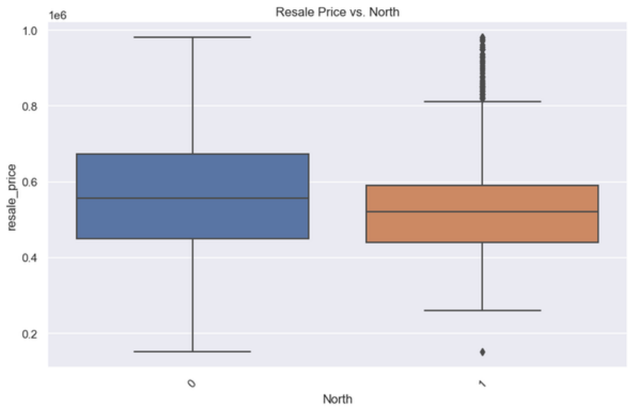


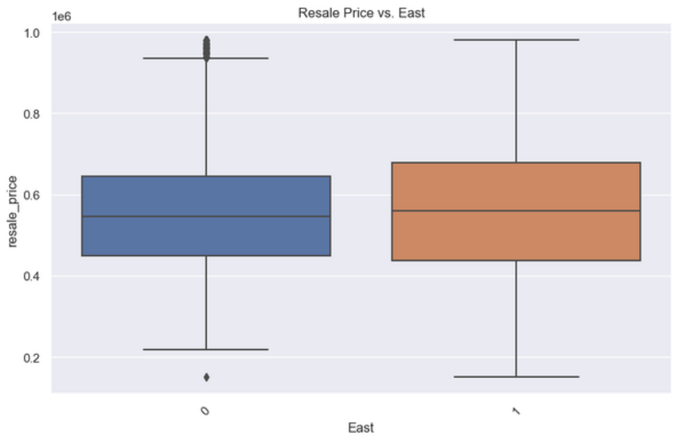
EDA

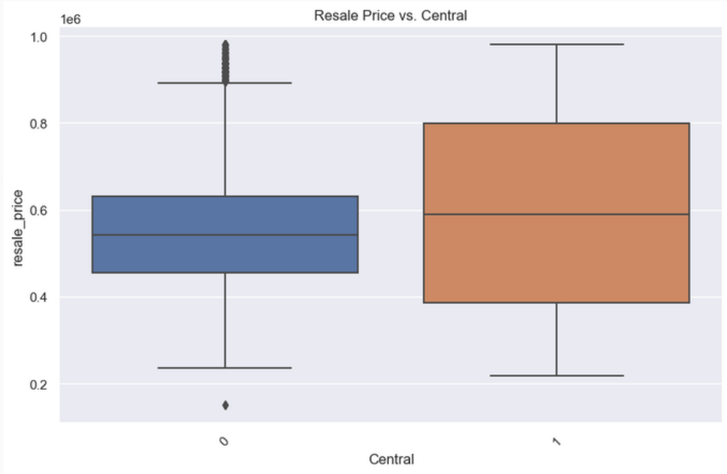

Resale Price vs. flat_type
1e6
1.0
0.8
0.6
0.4
0.2
resale_price
0
1
2
3
4
5
6
flat_type

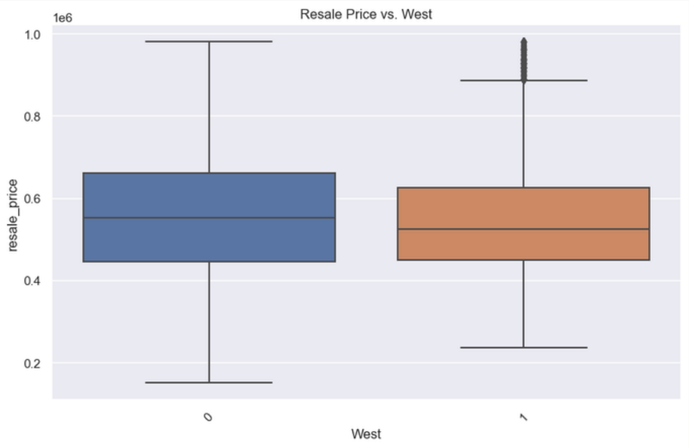

Resale Price vs. storey_range
1e6
1.0
0.8
0.6
0.4
0.2
resale_price
1
2
storey_range

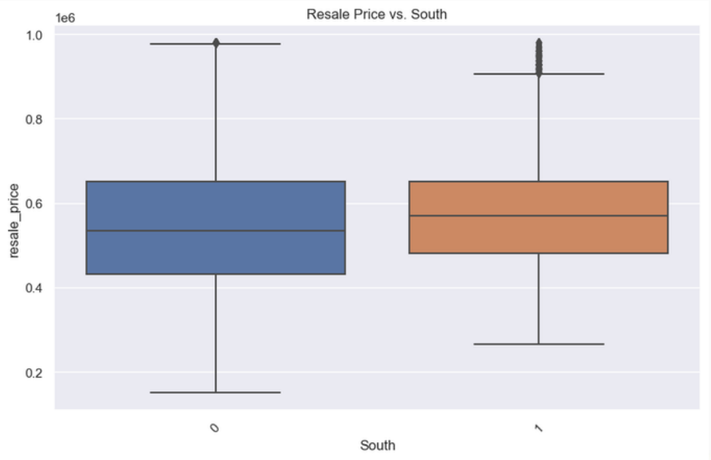
EDA


Resale Price vs. North
1e6
1.0
0.8
0.6
0.4
0.2
resale_price
0
1
North


Resale Price vs. East
1e6
1.0
0.8
0.6
0.4
0.2
resale_price
0
1
East

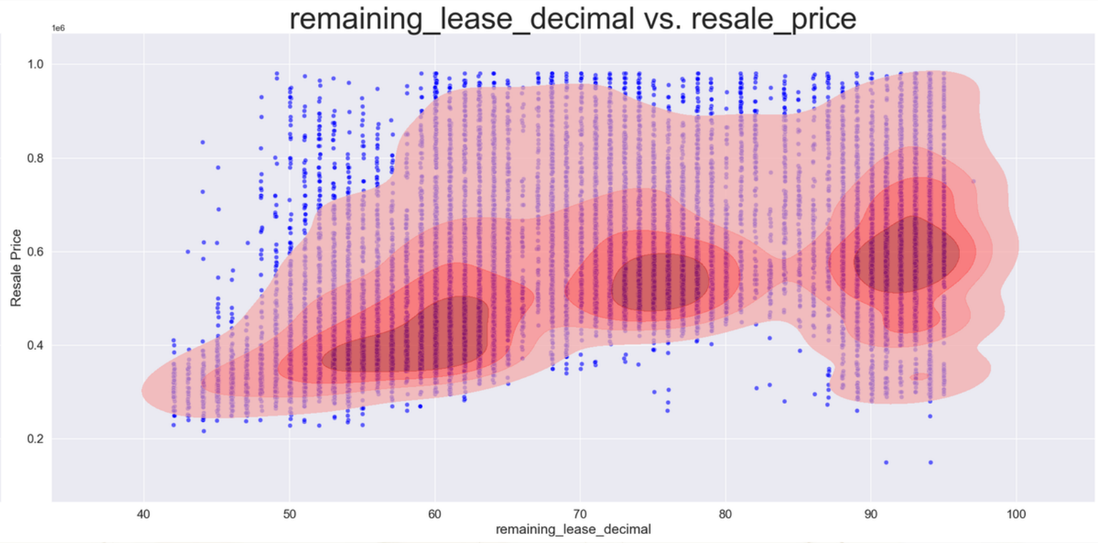

Resale Price vs. Central
1e6
1.0
0.8
0.6
0.4
0.2
resale_price
0
1
Central


Resale Price vs. West
1e6
1.0
0.8
0.6
0.4
0.2
resale_price
0
1
West


Resale Price vs. South
1e6
1.0
0.8
0.6
0.4
0.2
resale_price
0
1
South

EDA

The following graphs are the scatter and KDE distribution of the continuous data floor area square and remaining lease respectively



Machine Learning Models

Linear Regression

Decision Tree Regressor with Grid Search

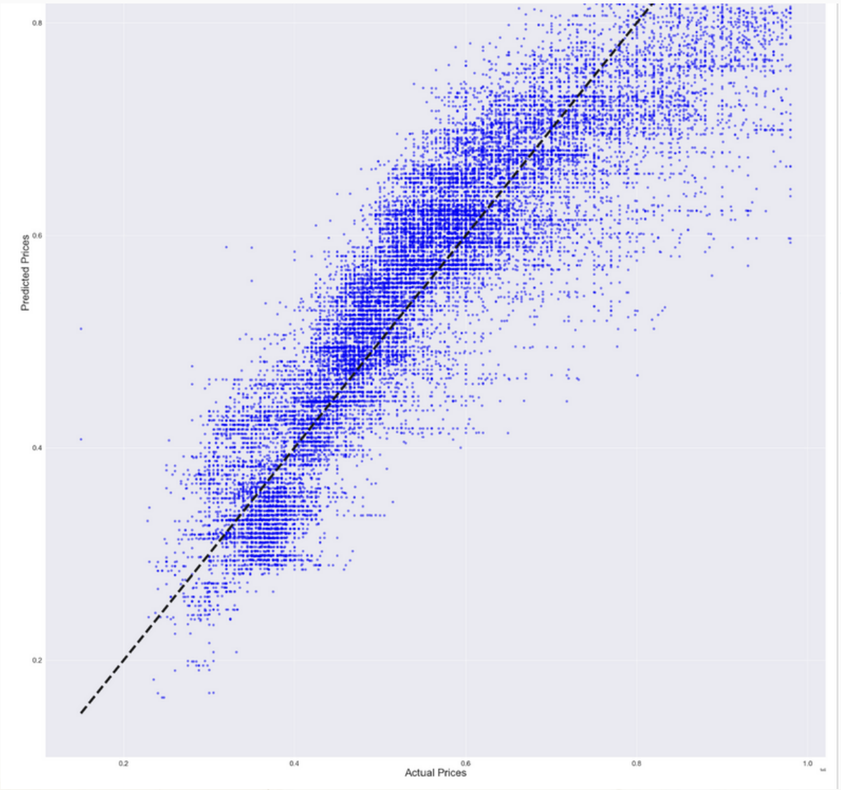
Random Forest with Grid Search

Multi-variate Linear Regression

Regression models are very effective in predicting outputs for response variables from the predictors.

- Linear relationship between independent and dependent variable

- Scatterplots are predicted values and diagonal line represent true values



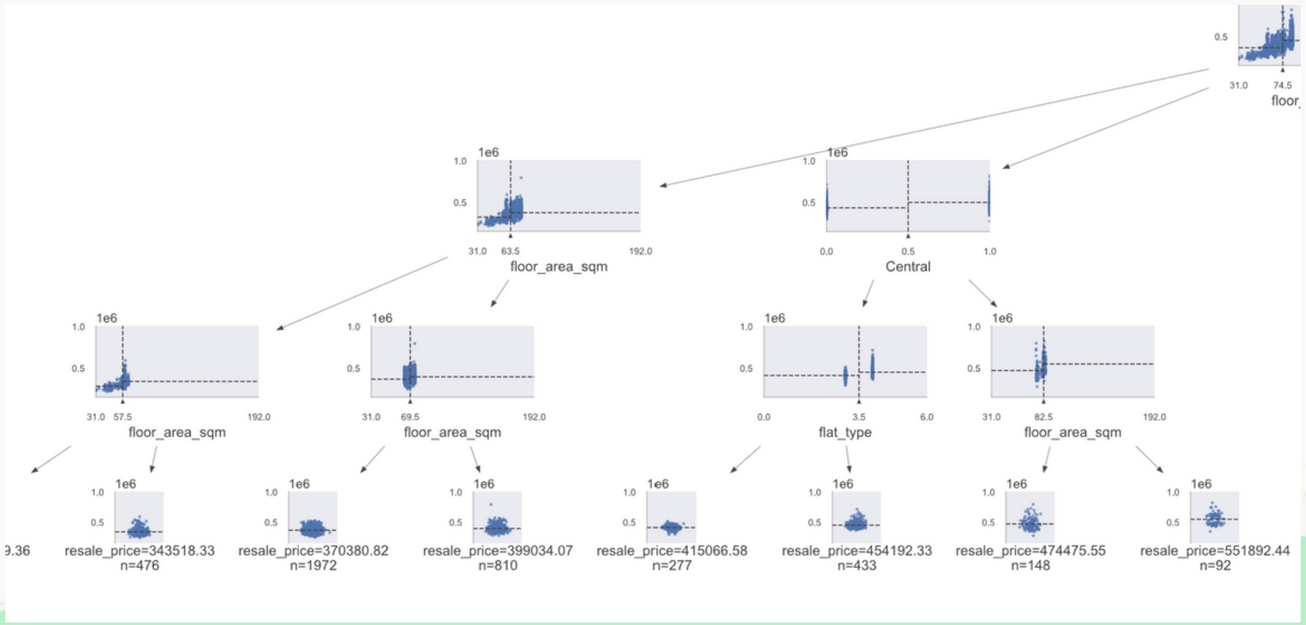
Goodness of fit

```
Metrics:
r2_train: 0.7942039518070801
Adjusted R^2: 0.7941222664493256
r2_test: 0.8039573738695427
Adjusted R^2: 0.8036457003780794
MSE: 4757397048.2488985
MAE: 52936.209277091504
```

- Goodness of fit factors: Explained Variance, Adjusted explained variance, Mean squared error and Mean absolute error
- Relatively similar goodness of fit for test and train data
- Minimal overfitting and good accuracy

Decision Tree with Grid Search

- Decision trees can handle non-linear relationships from the data.
- GridSearch allows us to fine tune our hyperparameters of our tree, improving the accuracy.



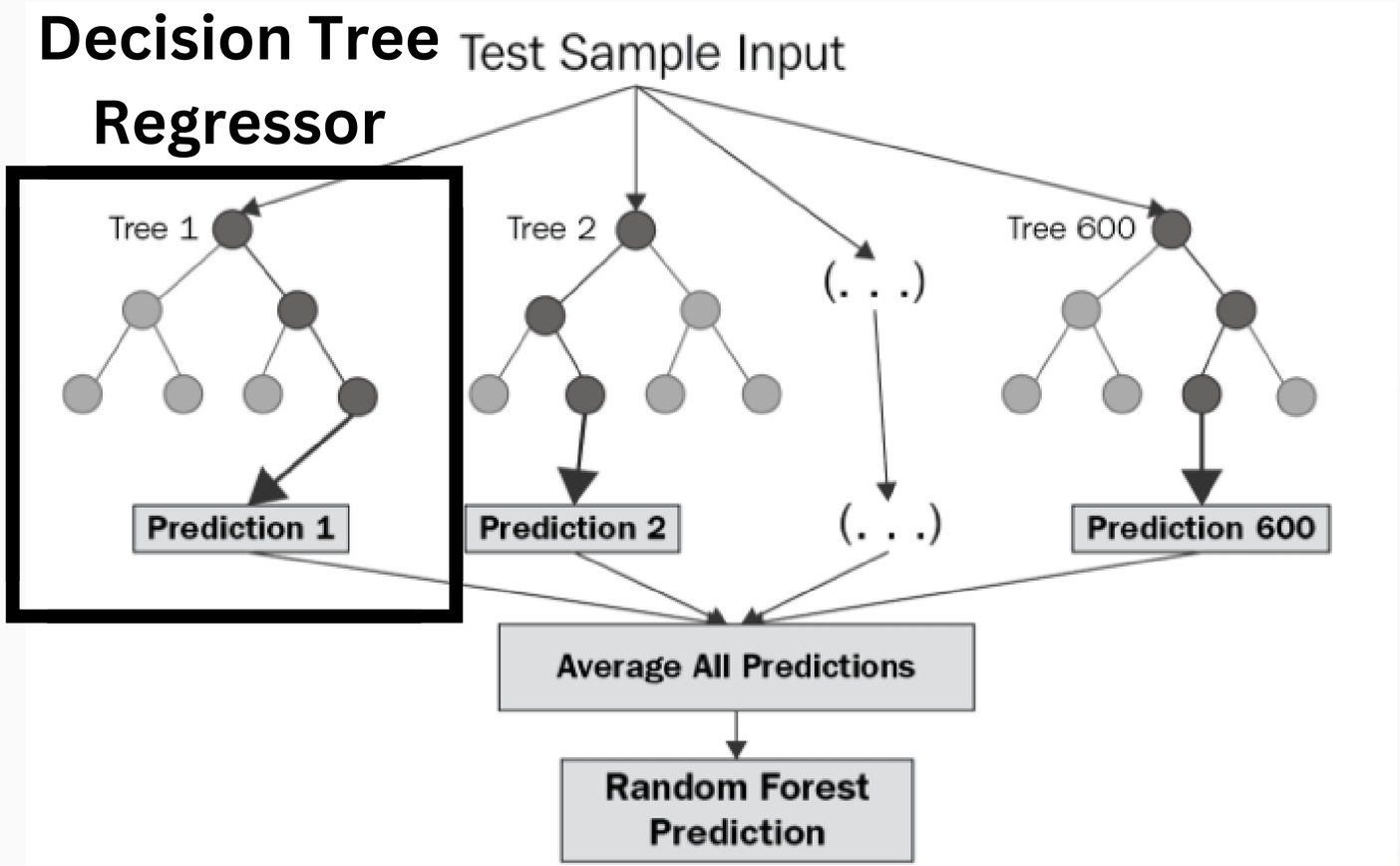
Goodness of Fit

- Similar goodness of fit between train and test data
- Decision tree have higher prediction accuracy as compared to linear regression
- Better model for predicting resale housing prices

```
Test Set Metrics:
r2_train: 0.8106034369944796
Adjusted R^2: 0.8105282609833635
r2_test: 0.811281127205257 4
Adjusted R^2: 0.8109810972008142
MSE: 4579670381.405298
MAE: 48428.98224068751
```

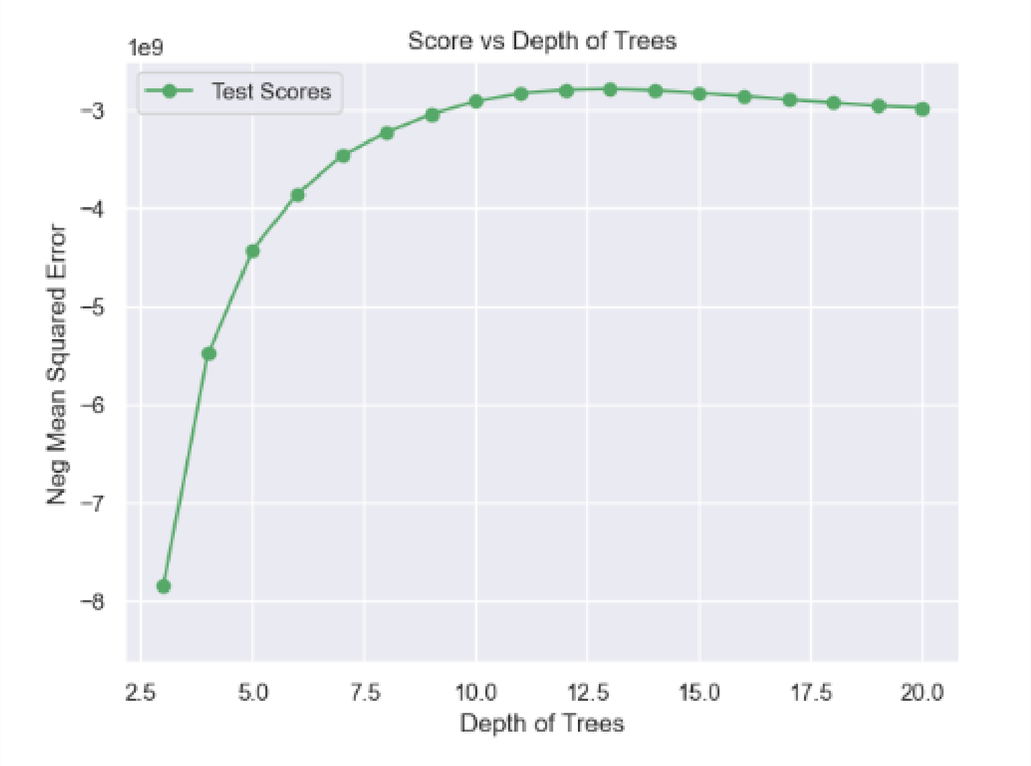
Random Forest with Grid Search

Random Forests!



- Reduces overfitting by averaging across multiple trees
- Captures complex relationships between multiple trees

Goodness of Fit




```
Test Set Metrics:
r2_train: 0.8196343172296313
Adjusted R^2: 0.819562725790179
r2_test: 0.8199464941146868
Adjusted R^2: 0.8196602405282236
MSE: 4369386568.284575
MAE: 47326.06344657531
```

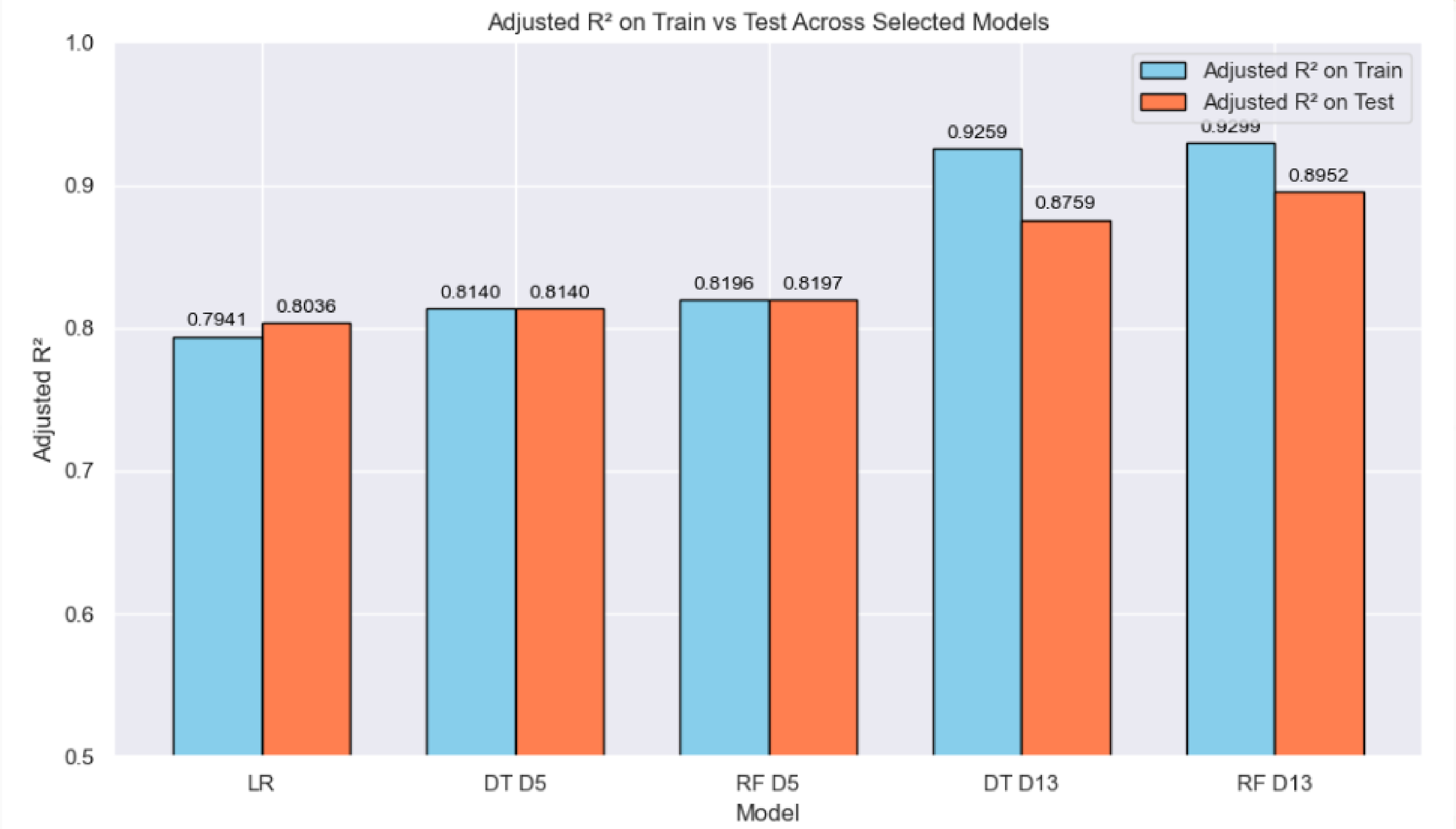
```
Test Set Metrics:
r2_train: 0.9298884438074786
Adjusted R^2: 0.9298606148593496
r2_test: 0.8953525582857773
Adjusted R^2: 0.8951861871542761
MSE: 2539495823.662074
MAE: 34856.97134631782
```

Key Findings:

Overall Goodness of Fit

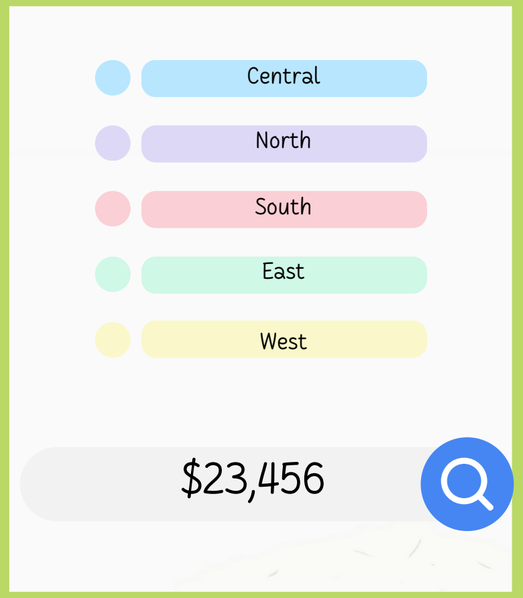
	Model	R^2 on Train	Adjusted R^2 on Train	R^2 on Test	Adjusted R^2 on Test	Mean Squared Error	Mean Absolute Error
0	LinearRegressor without Outliers	0.794204	0.794122	0.803957	0.803646	4.757397e+09	52936.209277
1	Depth05_DecisionTreeRegressor with Gridsearch	0.814026	0.813952	0.814335	0.814040	4.505568e+09	48121.339562
2	Depth05_RandomForestRegressor with Gridsearch	0.819634	0.819563	0.819946	0.819660	4.369387e+09	47326.063447
3	Depth13_DecisionTreeRegressor with Gridsearch	0.925945	0.925916	0.876062	0.875865	3.007616e+09	37319.693209
4	Depth13_RandomForestRegressor with Gridsearch	0.929888	0.929861	0.895353	0.895186	2.539496e+09	34856.971346


Mean Squared Error Across Models
1e9
Mean Squared Error
4
3
2
1
0
LR
DT D8
RF D6
DT D13
RF D13
Model


Adjusted R² on Train vs Test Across Selected Models
Adjusted R² on Train
Adjusted R² on Test
Adjusted R²
1.0
0.9
0.8
0.7
0.6
0.5
0.7941
0.8036
0.8140
0.8140
0.8196
0.8197
0.9259
0.8759
0.9299
0.8952
LR
DT D5
RF D5
DT D13
RF D13
Model

FUTURE DIRECTION

Software



Models

Bayesian Search

Neural Networks

Time Series

Thank you!