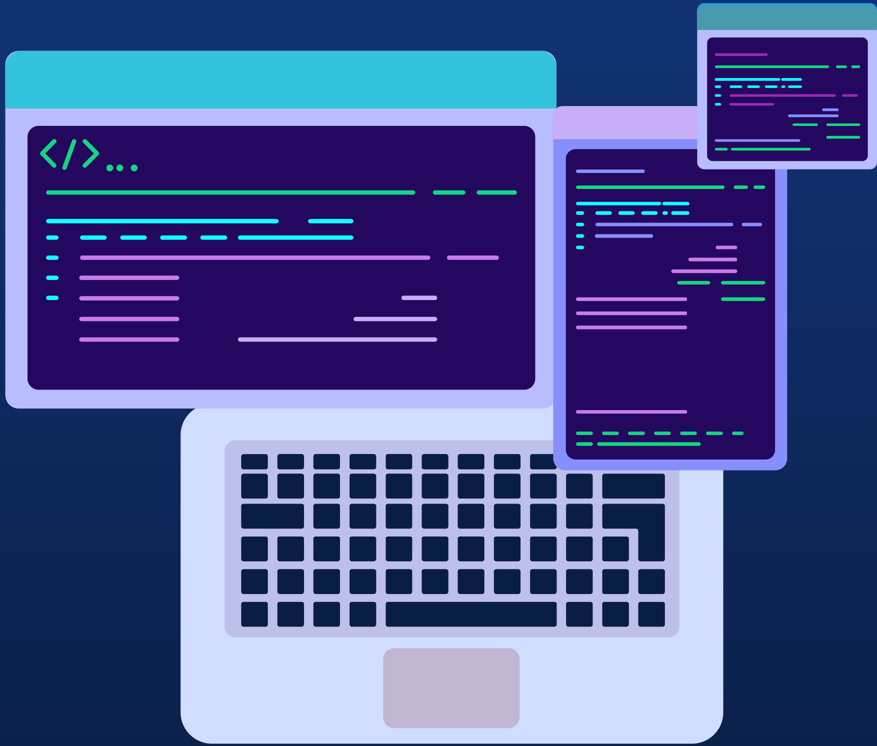
Final Project: Predicting Home Resale Prices

Jorge Bello, Sinéad McKeon, Nicolette Orlando and Chandler Stegen
Professor Platt
BSAD 399-101
December 12, 2022



AGENDA

1
INTRO & BUSINESS OBJECTIVE

A brief overview of what we do, and what we aim to do with this model

2
DATA DICTIONARY

Attributes used in our model with their descriptions

3
APPROACH AND DESIRED OUTCOME

The original steps we took in our model making process, and what we hope to gain from it

4
GRAPHS AND VISUALS

Graphs with Model data, along with further visuals to help illustrate our findings

5
MODEL STATISTICS

All of the data predicted by the model

6
SUMMARY

If our model was successful, along with concluding words on the project as a whole

1

INTRODUCTION AND BUSINESS OBJECTIVE

Introduction → Problem Statement

We're realtors, looking to apply our extensive data science knowledge into our career.

To do this, we've decided to try something new - a way to buy and resell houses.

However, to do this, we needed to create a model that could help us predict the resale prices of these houses before we decide which ones to buy.

We hope to gain profit from these investments and predictions.

Business Objective

Predict Resale Prices

Create a model that helps us predict the resale prices of houses.

Purchase and Resell

Buy homes based on predictions, and resell them

Make a Profit

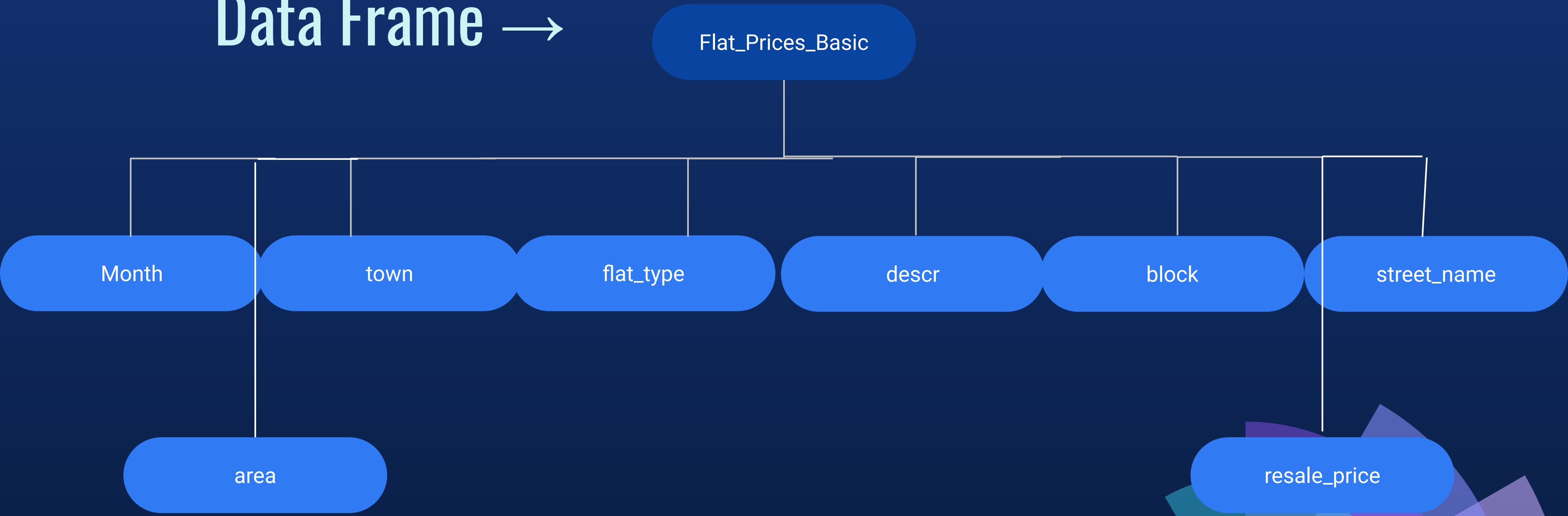
Make good business decisions using the model

2

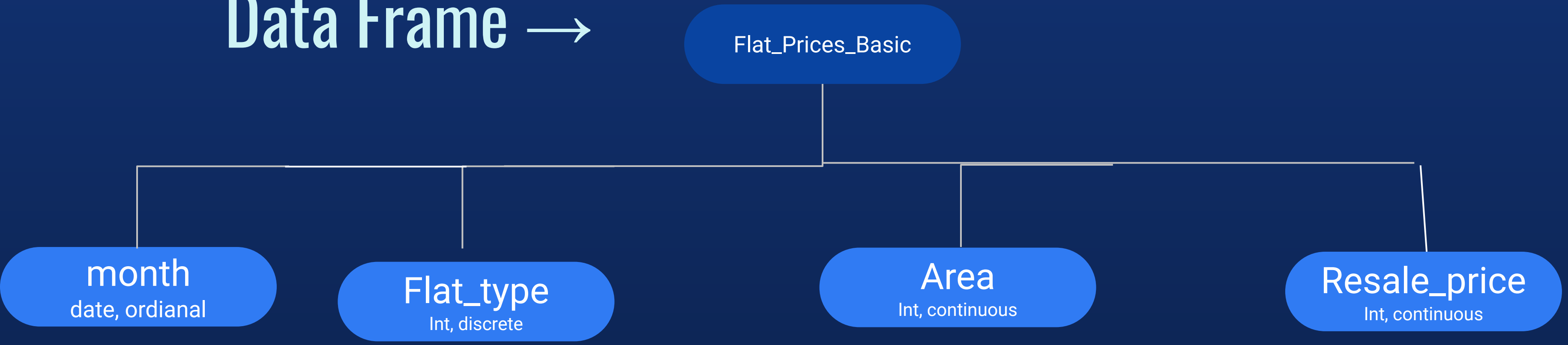
DATA DICTIONARY

Data Frame	General Info	Field	Description	Field Type/values	Data Type	Sample Data	Other
Flat_Prices_Basic		month	month and year of listing on market	date	Ordinal	2017-01	
		town	name of town	string	Nominal	ANG MO KIO	
		flat_type (Number of Rooms)	number of rooms in flat	int	Discrete	2	
		descr	description of the listing	string	Other		
		block	number associated with the block	string	Ordinal		
		street_name	name of the street	string	Nominal		
		Area (square ft)	area in square feet	int	Continuous		
		resale_price (in Thousands USD)	resale price in thousands USD	int	Continuous		
Flat_Model		flat_model	model of the flat	string	Nominal	Sunshine	
Lease_Time		remaining_lease	months remaining on lease	int	Continuous	12	
Location_and_Storey_Range		latitiude	latitudinal coordinate of listing	Int	Continuous		
	d into a specficied to belong	storey_range	range of storeys	int	Nominal	1_to_3	

Data Frame →


Flat_Prices_Basic
Month
town
flat_type
descr
block
street_name
area
resale_price

Data Frame →



Flat_Prices_Basic
month
date, ordianal
Flat_type
Int, discrete
Area
Int, continuous
Resale_price
Int, continuous

Data Frame → Flat_Model

flat_model

Data Frame → Location_and_Storey_Range

latitude

storey_range

Data Frame → Lease_Time

remaining_lease

Data Frame → Flat_Model

Flat_model
String, nominal

Data Frame → Location_and_Storey_Range

Storey_range
Int, nominal

Data Frame → Lease_Time

Remaining_lease
Int, continuous

3

APPROACH AND DESIRED OUTCOME

Desired Outcome

- We're hoping to create a predictive model allows us to:
  - See prices of houses
  - Compare those prices to the actual prices of homes we find on the market

- IF we deem house to be undervalued…
  - We can purchase/invest in it!
  - SO, later we can sell it at predicted price for profit

Raw Data

Excel Data Files

- 4 excel data files
  - Each contained 92,270 rows
    - Flat Prices Basic
      - 19 columns
      - Missing values
      - Empty/No variable
    - Flat Model
    - Location and Storey Range
    - Lease Time

Data Sets

- Flat Prices
  - Block number, street name, town name,
  - Number of rooms, area sq. ft., and current resale price
- Flat Model
  - Type of home
- Location and Storey Range
  - Latitude and how many storey's
- Lease Time
  - Time remaining on each lease

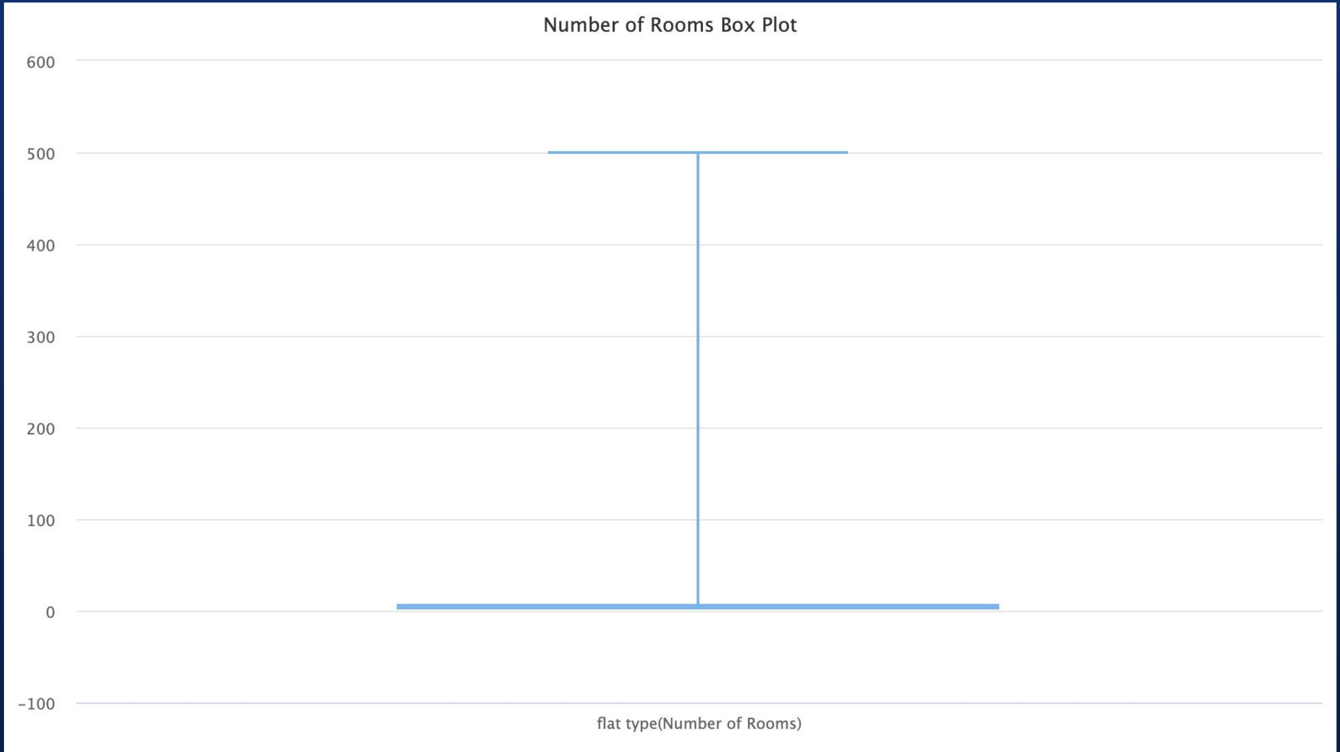
Initial Approach

Initial Approach

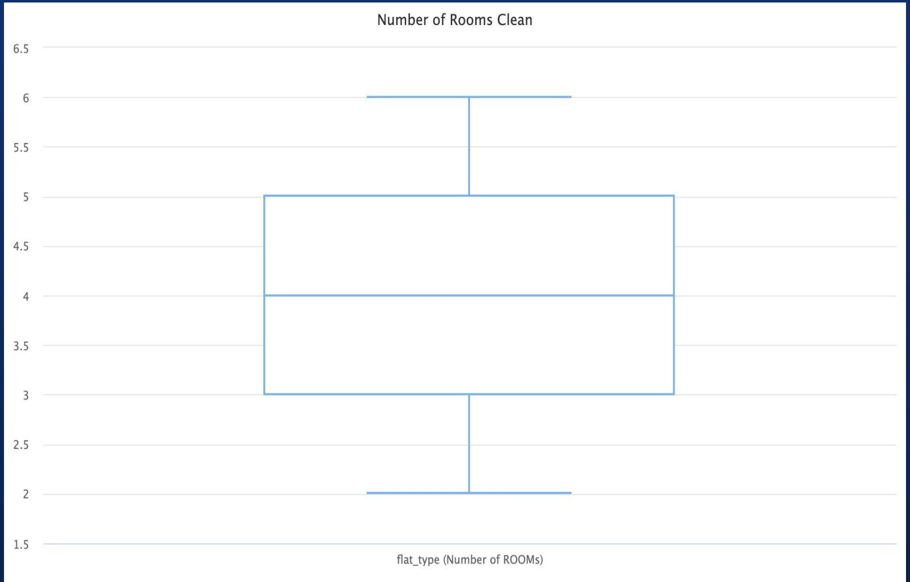
- Clean
  - Replacing values
  - Outliers
  - Imputing Missing Values
- Feature Selection
- Feature Engineering
  - Normalization
  - Encoding
- Considered doing square feet divided by number of rooms, chose against it

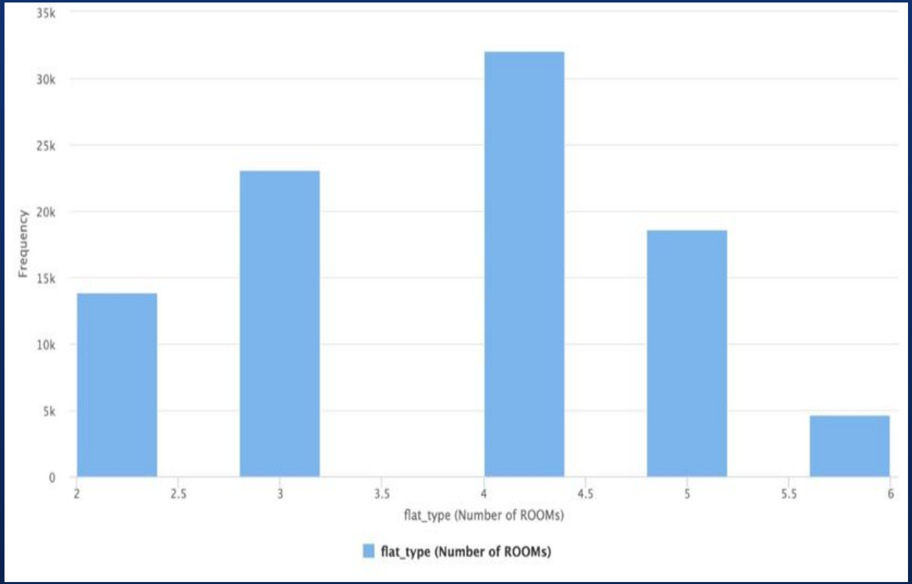
Data Cleaning: Outliers

Number of Rooms Raw

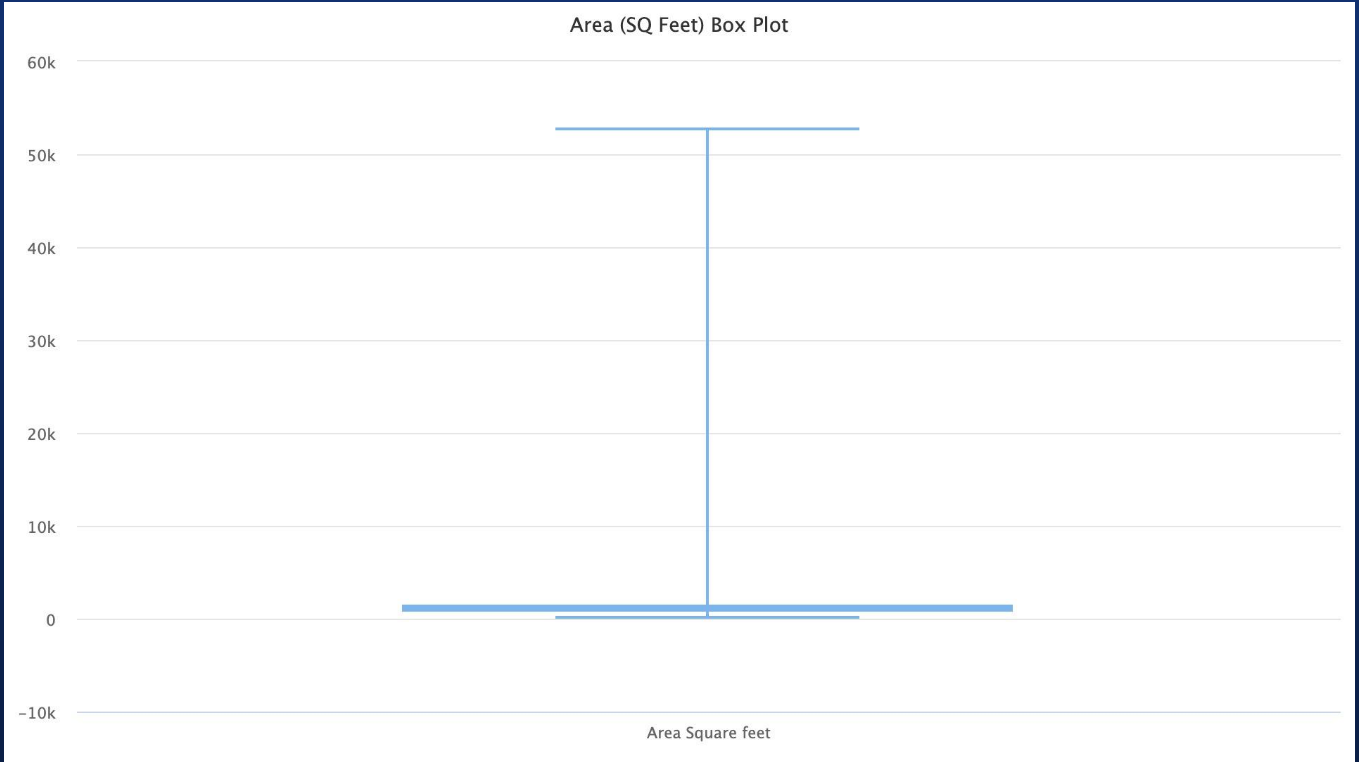


Number of Rooms Clean

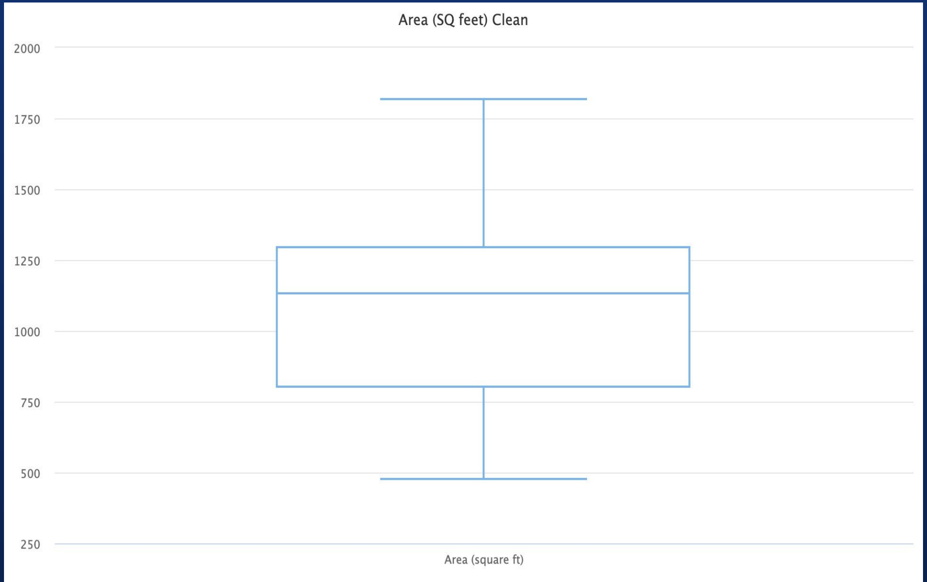


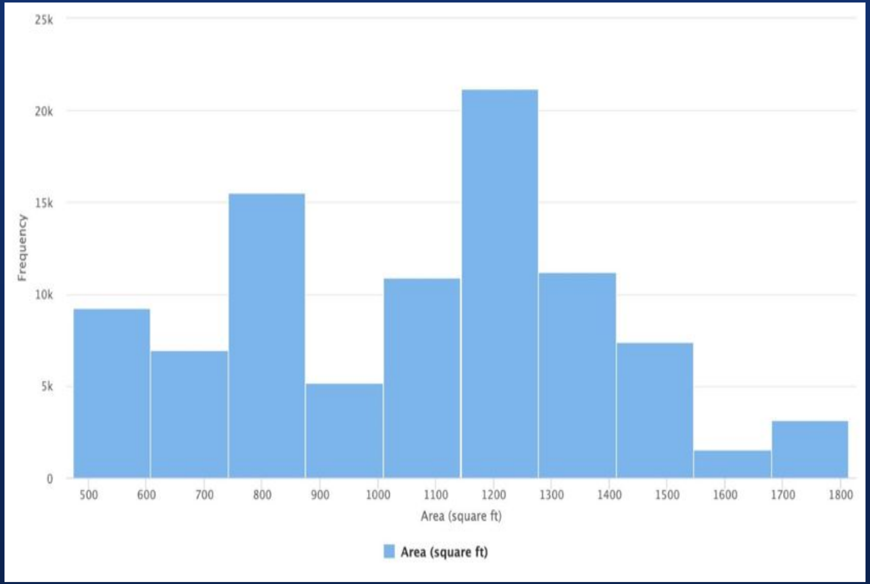


Area (Square feet) Raw

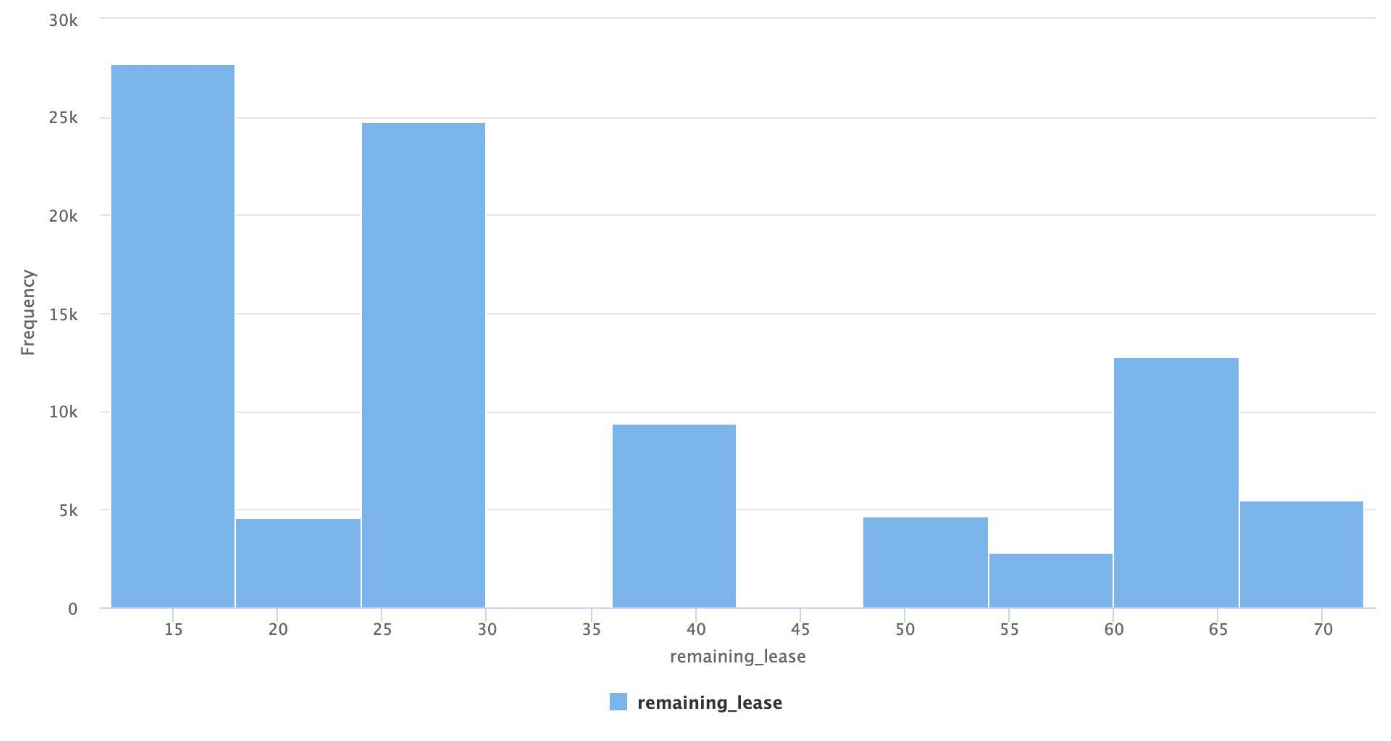


Area (Square feet)Clean

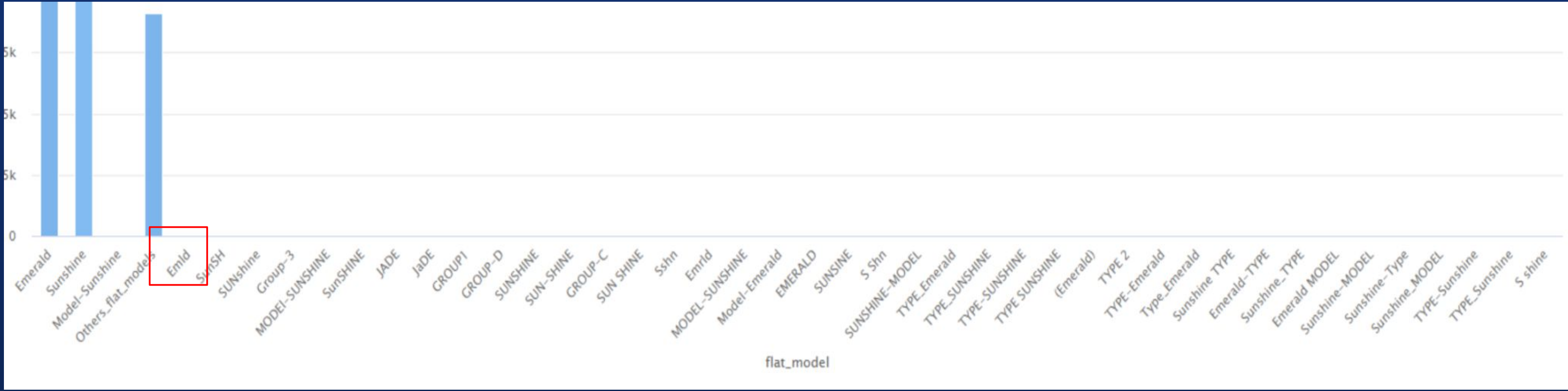


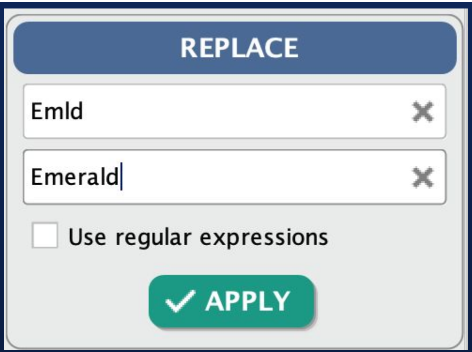


Remaining Lease

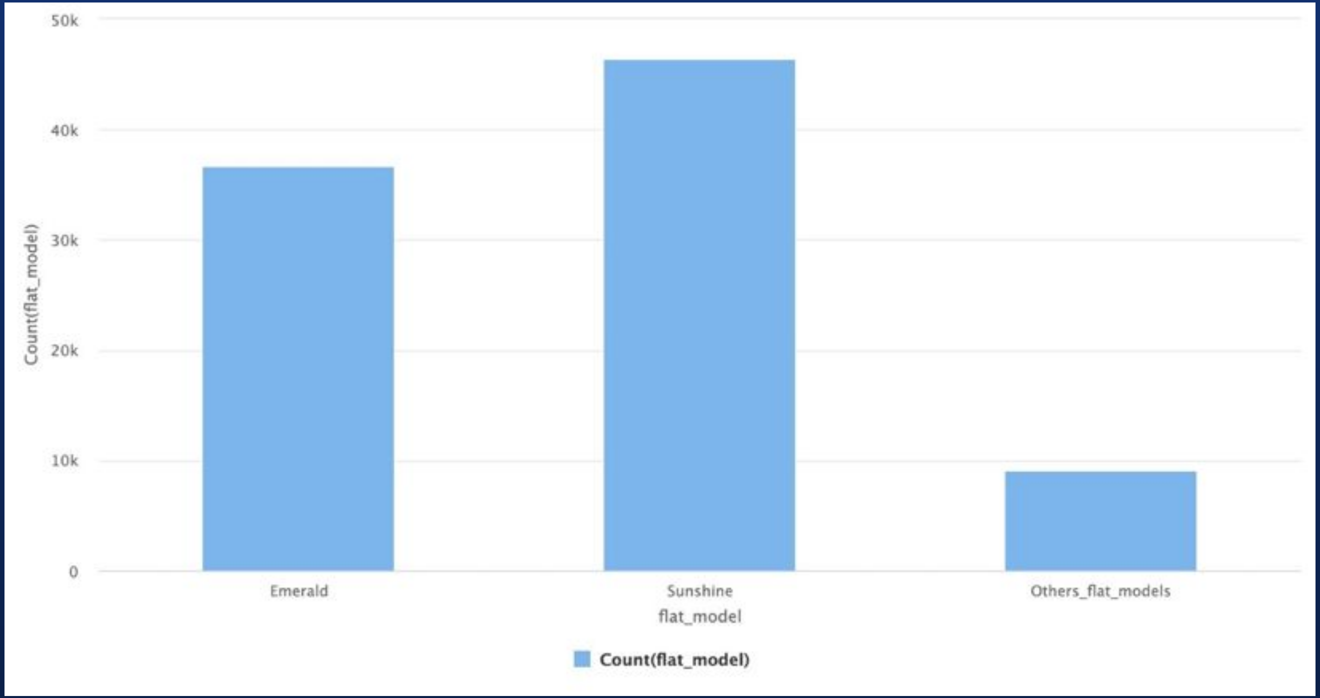


Flat Model Raw

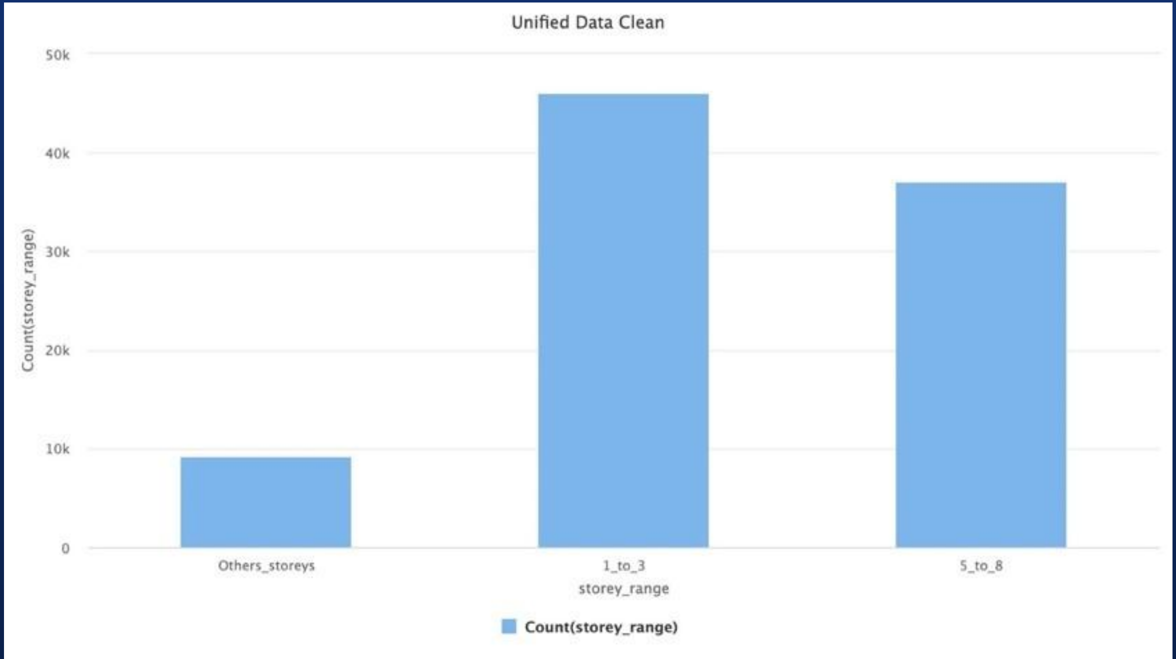




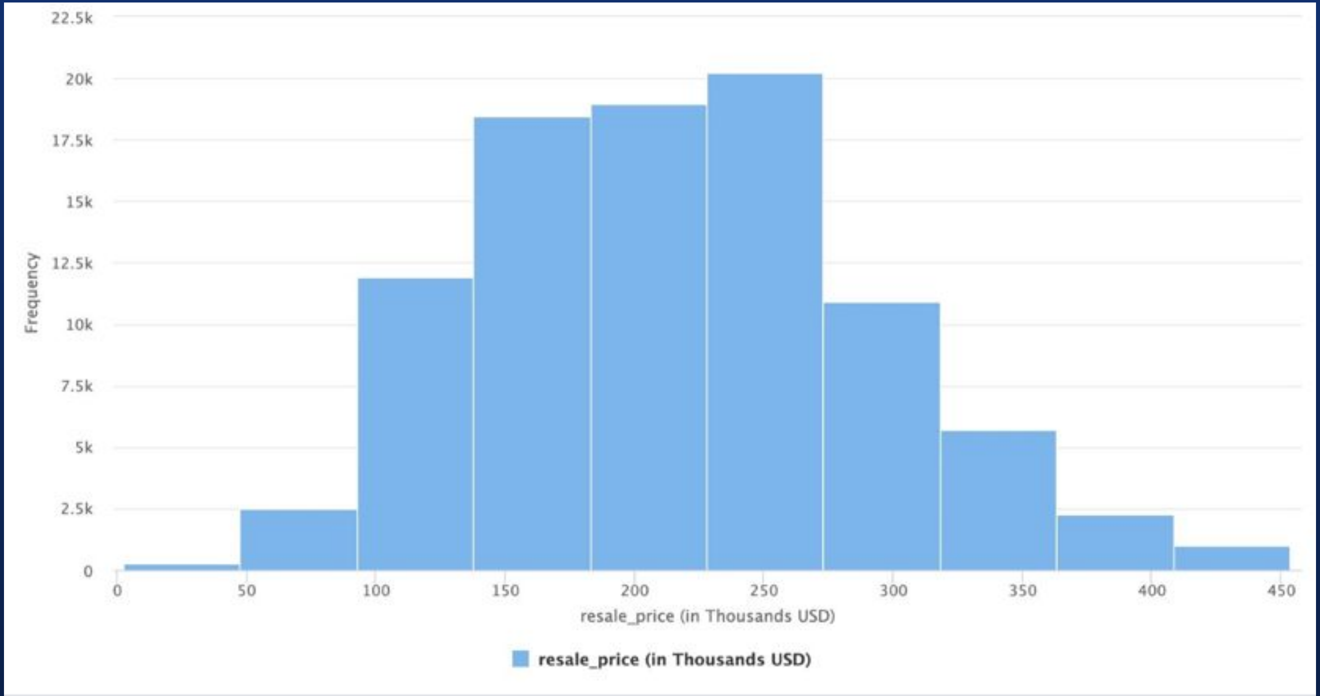
Flat Model Clean



Storey Range Clean



Resale Price



Feature Selection

Label: Resale Price

Flat Prices Basic

Area(Square Feet)
Flat Type (Number of Rooms)
Month
Descr
Street Name
Town
Block

Flat Model

Flat Model
- Emerald
- Sunshine
- Others

Location and Storey Range

Storey Range
- 1_to_3
- 5_to_8
- Others

Latitude

Lease Time

Remaining Lease

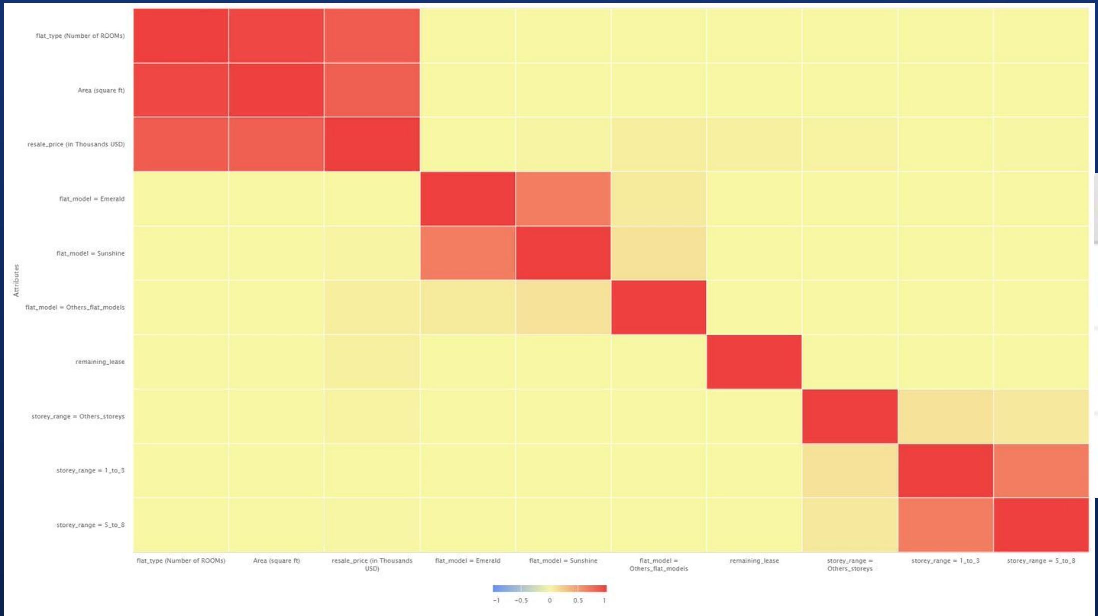
Feature Engineering: Normalization

Area (Square Feet)

	Min	Max
Area (square ft)	473.6257	1814.548

Normalized the range of Area (square ft) between 0 and 1

EDA: Correlation Matrix



Attribut...	flat_typ...	Area (s...	resale_...
flat_type ...	1	0.958	0.847
Area (sq...	0.958	1	0.823
resale_p...	0.847	0.823	1

Potential instance of multicollinearity

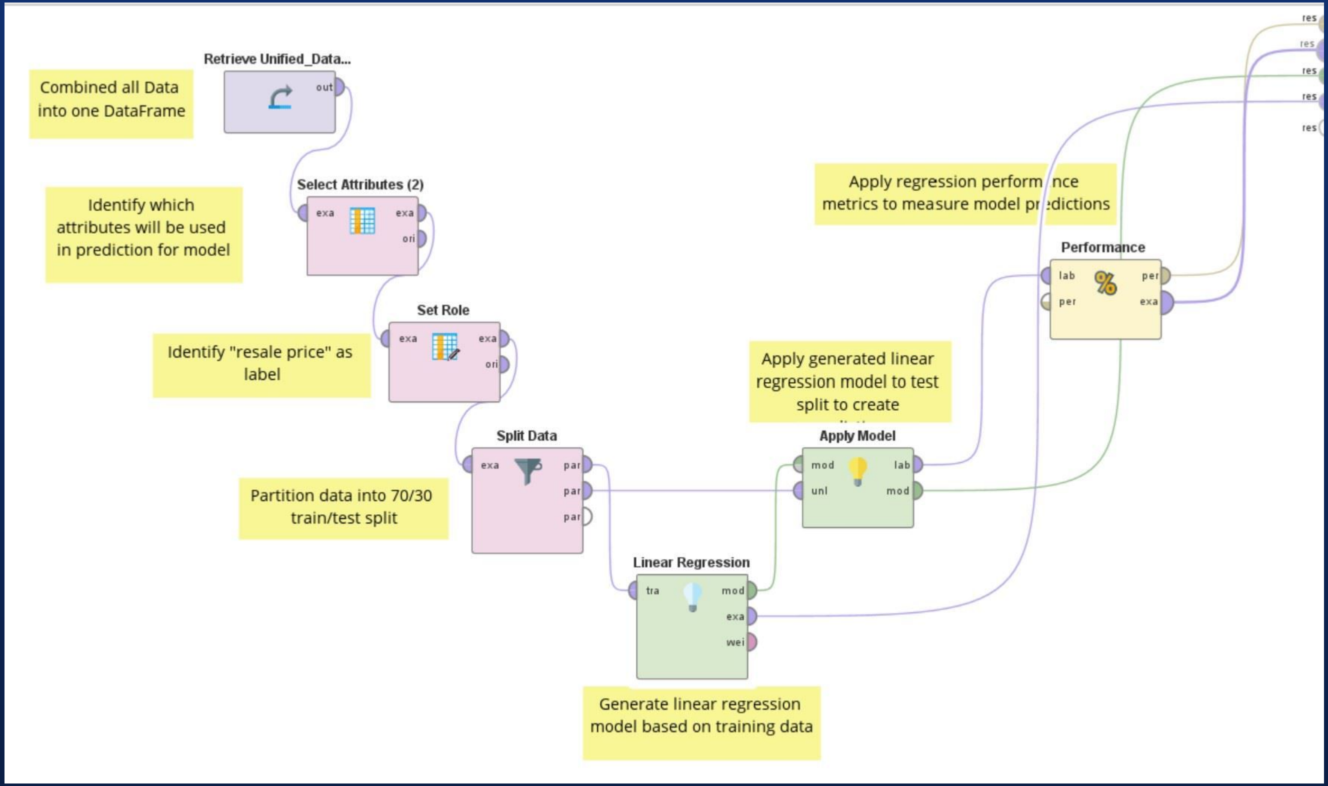
Feature Engineering: Encoding

flat_model = Emerald	flat_model = Sunshine	flat_model = Others_flat_models	storey_range	storey_range	storey_range = 5_to_8
1.0	0.0	0.0	1.0	0.0	0.0
1.0	0.0	0.0	1.0	0.0	0.0
1.0	0.0	0.0	1.0	0.0	0.0
1.0	0.0	0.0	1.0	0.0	0.0
0.0	1.0	0.0	0.0	1.0	0.0
0.0	1.0	0.0	0.0	1.0	0.0
0.0	1.0	0.0	0.0	1.0	0.0
0.0	1.0	0.0	0.0	1.0	0.0
0.0	1.0	0.0	0.0	1.0	0.0
0.0	1.0	0.0	0.0	1.0	0.0
0.0	0.0	1.0	0.0	0.0	1.0
0.0	0.0	1.0	0.0	0.0	1.0
0.0	0.0	1.0	0.0	0.0	1.0
0.0	0.0	1.0	0.0	0.0	1.0
0.0	0.0	1.0	0.0	0.0	1.0
0.0	0.0	1.0	0.0	0.0	1.0
0.0	0.0	1.0	0.0	0.0	1.0

4

GRAPHS AND VISUALS

Model Process



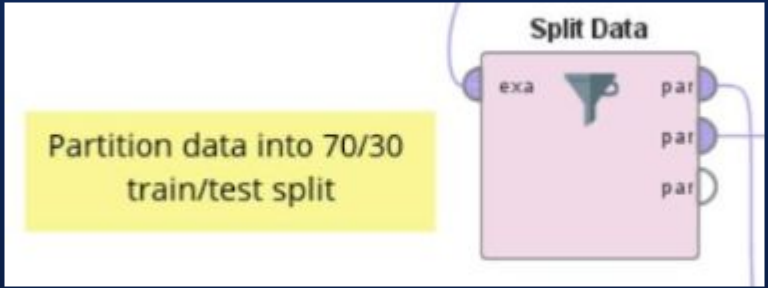
Train/Test Split

Training Set - 70%

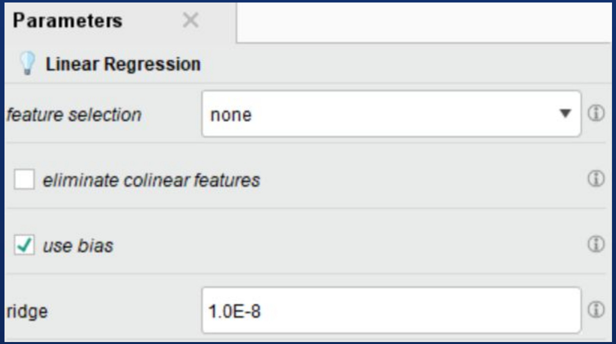
ExampleSet (64,589 examples, 1 special attribute, 7 regular attributes)

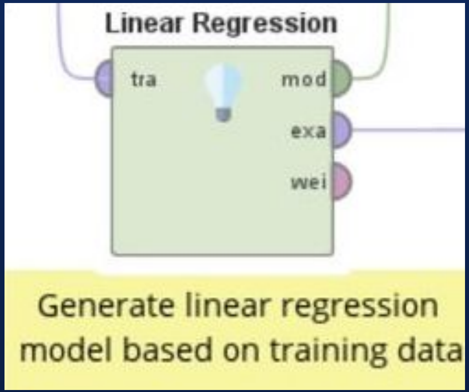
Test Set - 30%

ExampleSet (27,681 examples, 2 special attributes, 7 regular attributes)



Model Hyperparameters





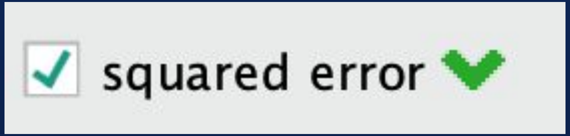
- ❏ Feature Selection was set to none
- ❏ Deselected eliminate collinear features
- ❏ Use bias selected
- ❏ Used default ridge (1.0^-8)

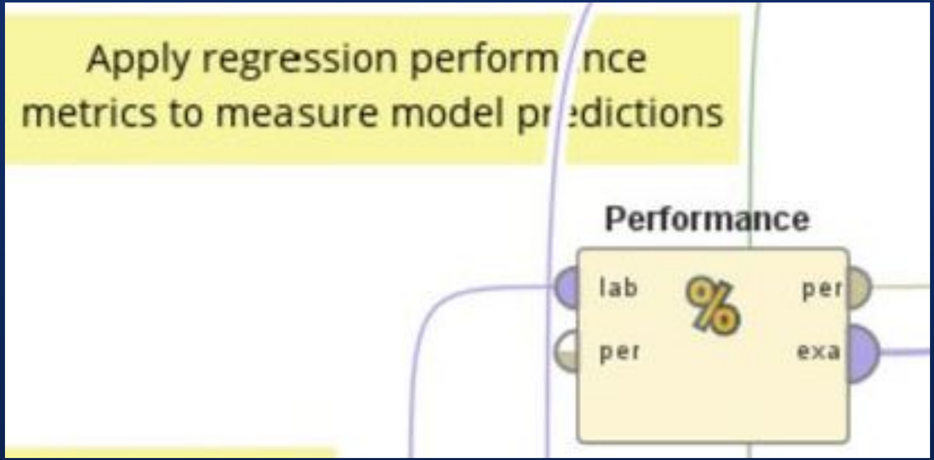
Model Performance Metrics

Root Mean Squared Error (RMSE)

Squared Correlation (R^2)

(Did not have access to adjusted R^2)





5

MODEL: Linear Regression

Model 1

Datasets Used

- Flat_Prices_Basic

Predictors Used

- Area(Sqare Feet)
- Flat_type (Number of Rooms)

Model Performance

```
root_mean_squared_error: 29.410 +/- 0.000
squared_correlation: 0.848
```

Model 2

Datasets Used

- Flat_Prices_Basic
- Flat_Model

Predictors Used

- Area(Sqare Feet)
- Number of Rooms
- Flat Model = Sunshine
- Flat Model = Emerald

Model Performance

```
root_mean_squared_error: 24.067 +/- 0.000
squared_correlation: 0.898
```

Model 3

Datasets Used

- Flat Prices Basic
- Flat Model
- Lease Time

Predictors Used

- Area(Square Feet)
- Number of Rooms
- Flat Model = Sunshine
- Flat Model = Emerald
- Remaining Lease

Model Performance

```
root_mean_squared_error: 18.991 +/- 0.000
squared_correlation: 0.937
```

Model 4

Datasets Used

- ❑ Flat Prices Basic
- ❑ Flat Model
- ❑ Lease Time
- ❑ Location and Storey Range

Predictors Used

- ❑ Area(Sqare Feet)
- ❑ Number of Rooms
- ❑ Flat Model = Sunshine
- ❑ Flat Model = Emerald
- ❑ Remaining Lease
- ❑ Storey Range = 1_to_3
- ❑ Storey Range = 5_to_8

Model Performance

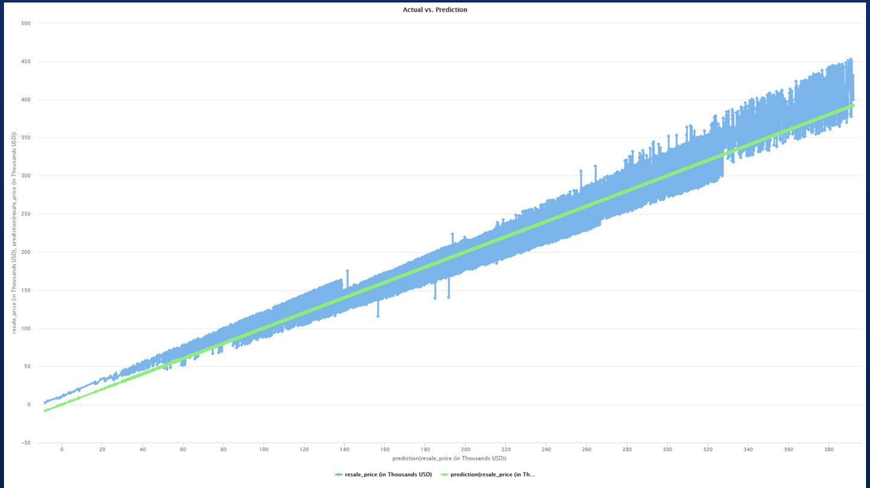
```
root_mean_squared_error: 15.162 +/- 0.000
squared_correlation: 0.960
```

Model Selection

FINAL MODEL: Model 4

- Utilized features from all given datasets
- Highest Model Performance
- Low required computational power
- Quick production speed
- More involved data transformation process

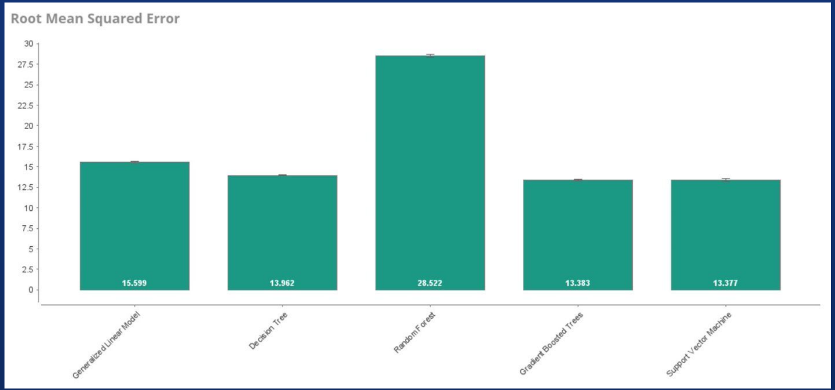
```
root_mean_squared_error: 15.162 +/- 0.000
squared_correlation: 0.960
```



For Fun !!!!!!!!!!!

Compared selected attributes through Rapidminer Auto Model (No Train/Test Split)

- SVM highest RMSE, significantly longer scoring time



Model	Root Mean Squared Error	Total Time
Generalized Linear Model	15.599	7 s
Decision Tree	13.962	8 s
Random Forest	28.522	55 s
Gradient Boosted Trees	13.383	51 s
Support Vector Machine	13.377	8 min 54 s

6

SUMMARY

Conclusion

Clean the Data & Creating Models

- ❏ Sort and Clean data we were given
- ❏ Use cleaned data to create standard linear regression models

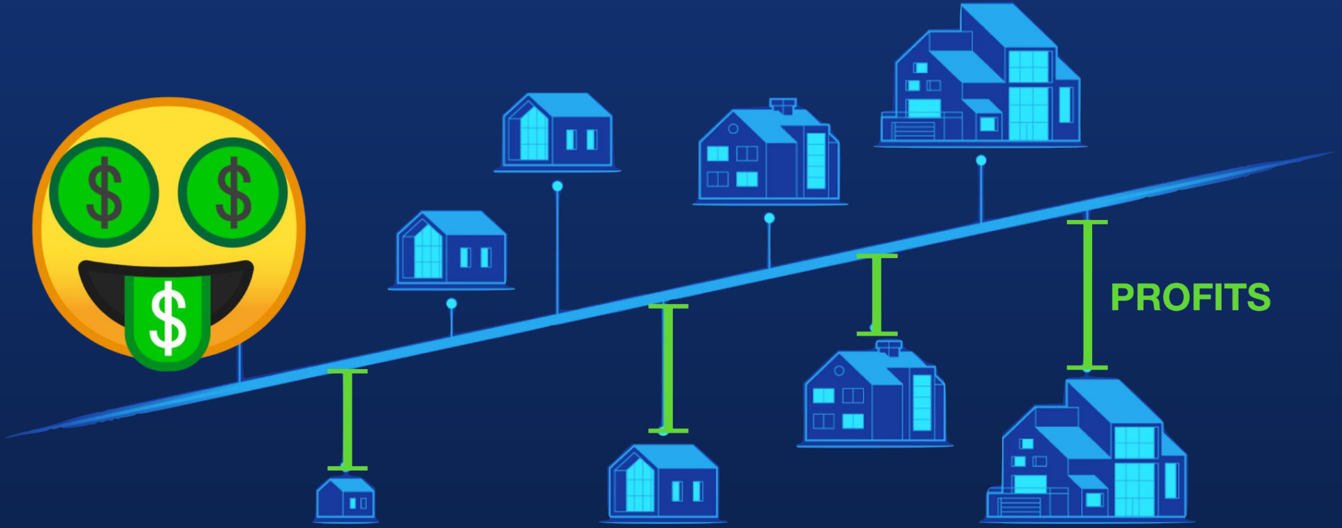
Testing the Models

- ❏ Tested 4 models
- ❏ Added more predictor variables with each
- ❏ 1 predicted: resale price

Choosing a Model

- ❏ Our last model had 7 predictor variables
- ❏ Most accurate

What Now?



Questions?