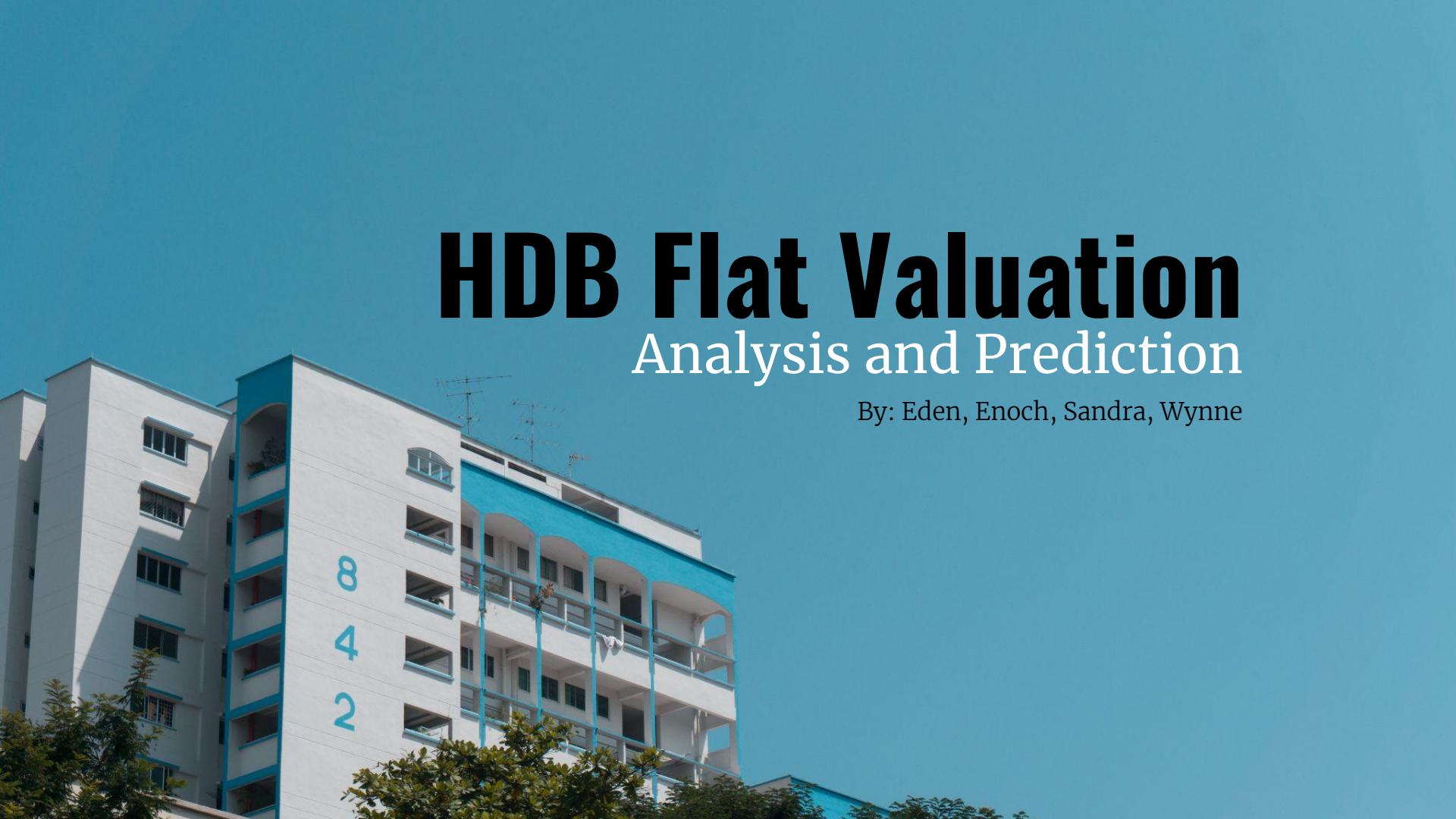
HDB Flat Valuation

Analysis and Prediction

By: Eden, Enoch, Sandra, Wynne



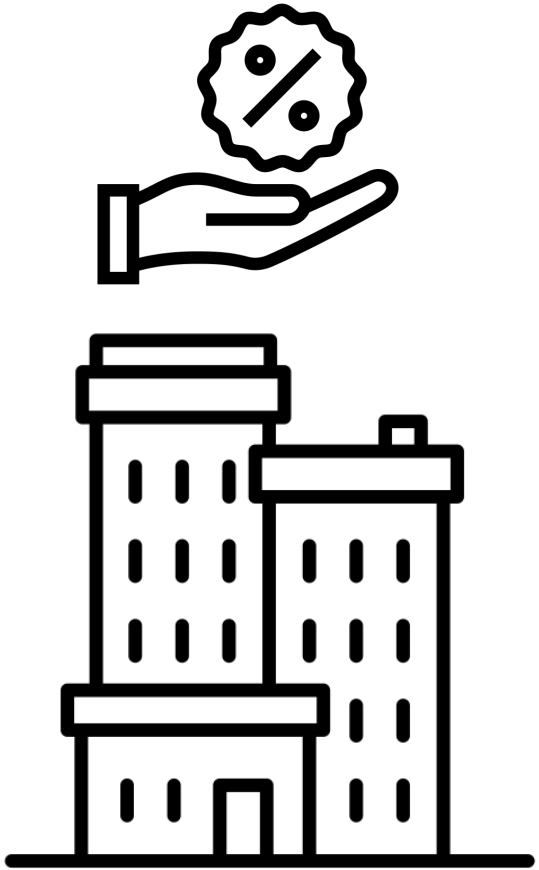
Contents

1 INTRODUCTION

Problem Statement

Our team of data scientists has been engaged by a group of real estate agents who are planning to establish their own real estate agency. They have specifically identified HDB flats as their primary focus and want us to develop a price prediction model tailored to this housing type.

The aim of our data science project is to create a robust and accurate **price prediction model** for HDB flats.



Context

$58.7 BILLION

Size of SG real estate market in 2023

$81.8 BILLION

Projected Size of SG real estate market in 2028

Source: https://www.mordorintelligence.com/industry-reports/singapore-real-estate-market

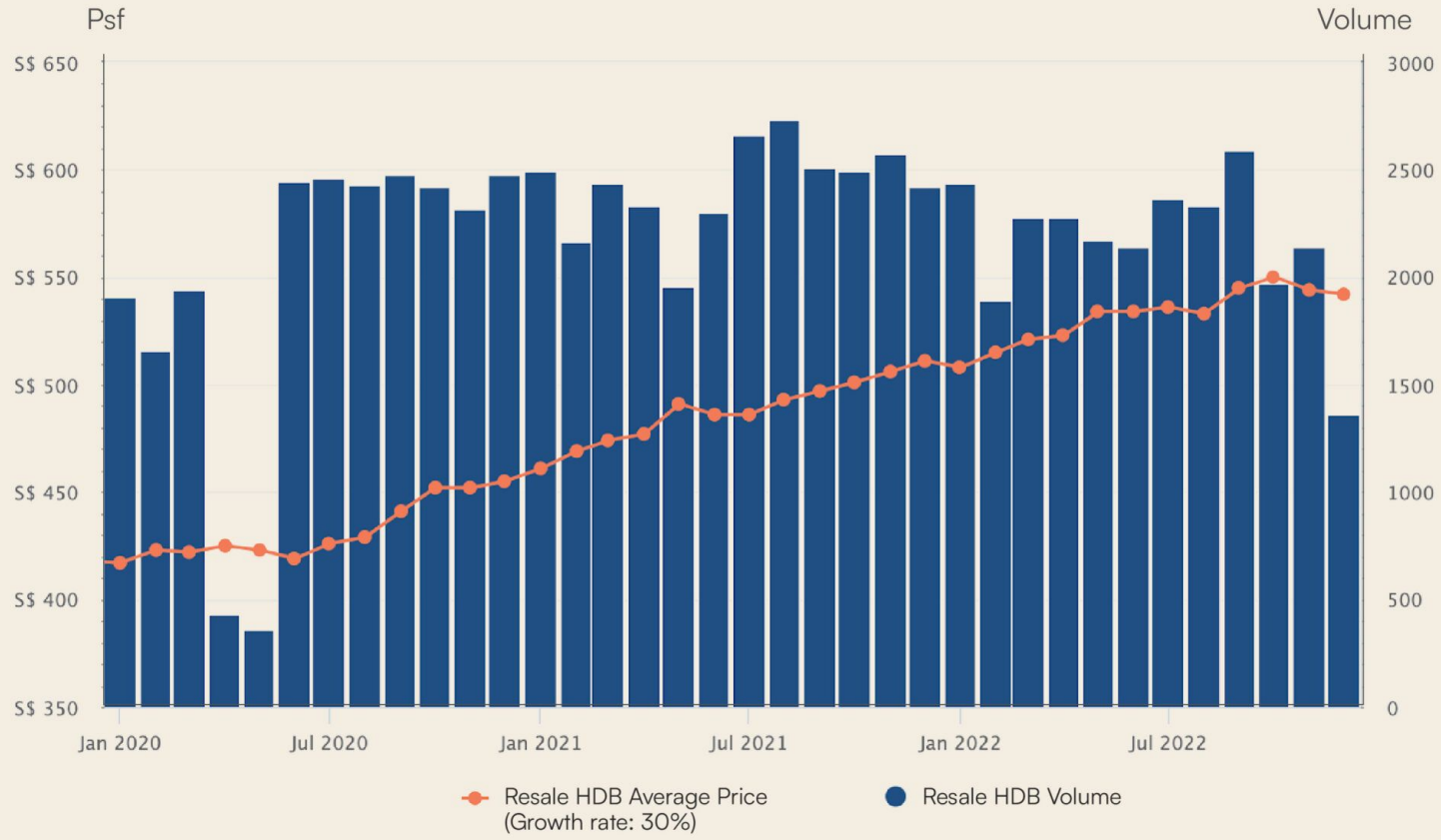
Context

Asking price for HDB resale flats increased for

16 straight quarters

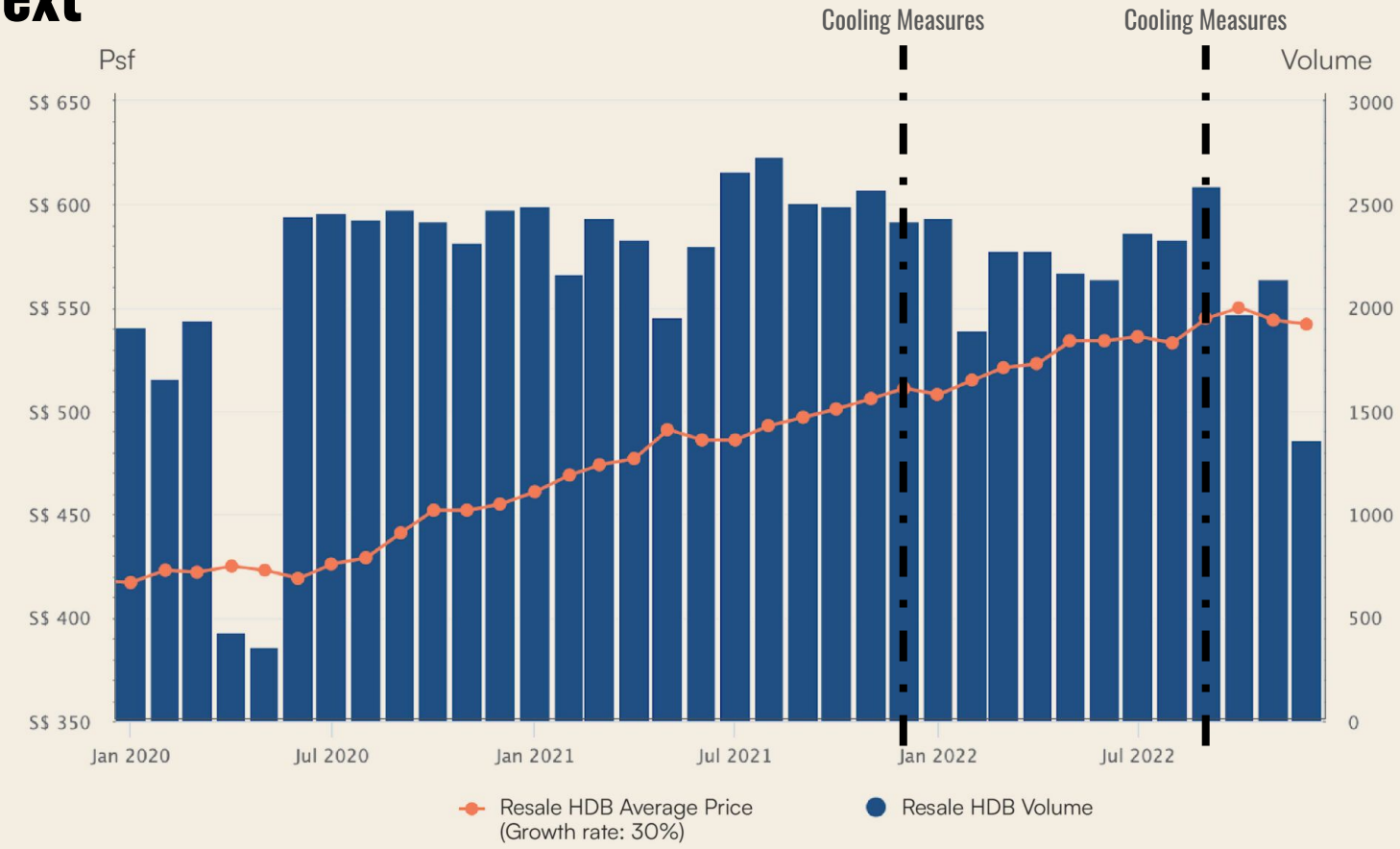
Source: https://cdn-cms.pgimgs.com/static/2022/08/PGSG-singapore-property-market-report-q3-2022-powered-by-propertyguru-datasense-full-report.pdf

Context



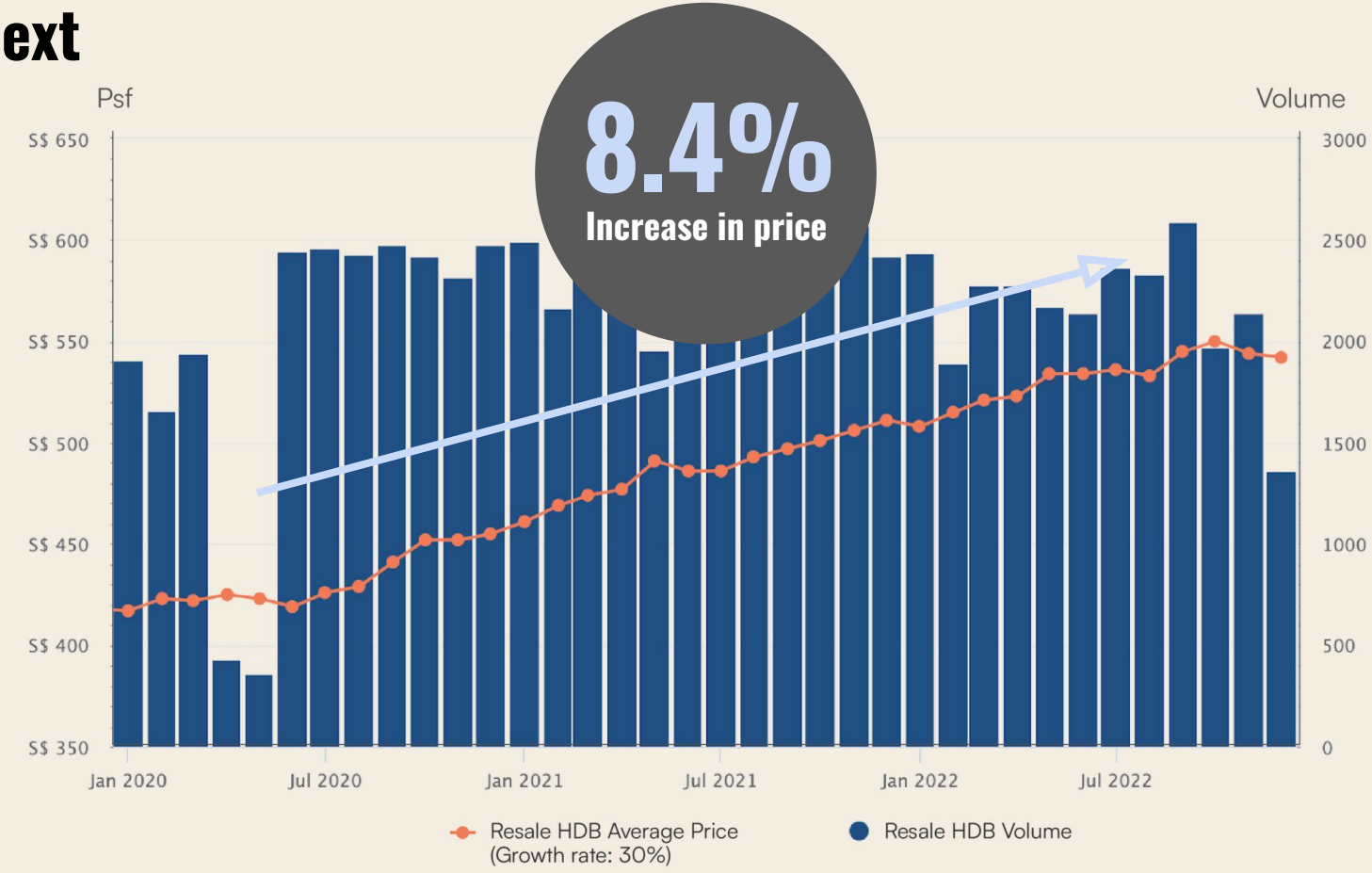
Source: https://www.propertylimbrothers.com/2022-singapore-real-estate-market-wrapped-key-statistics-and-trends/

Context



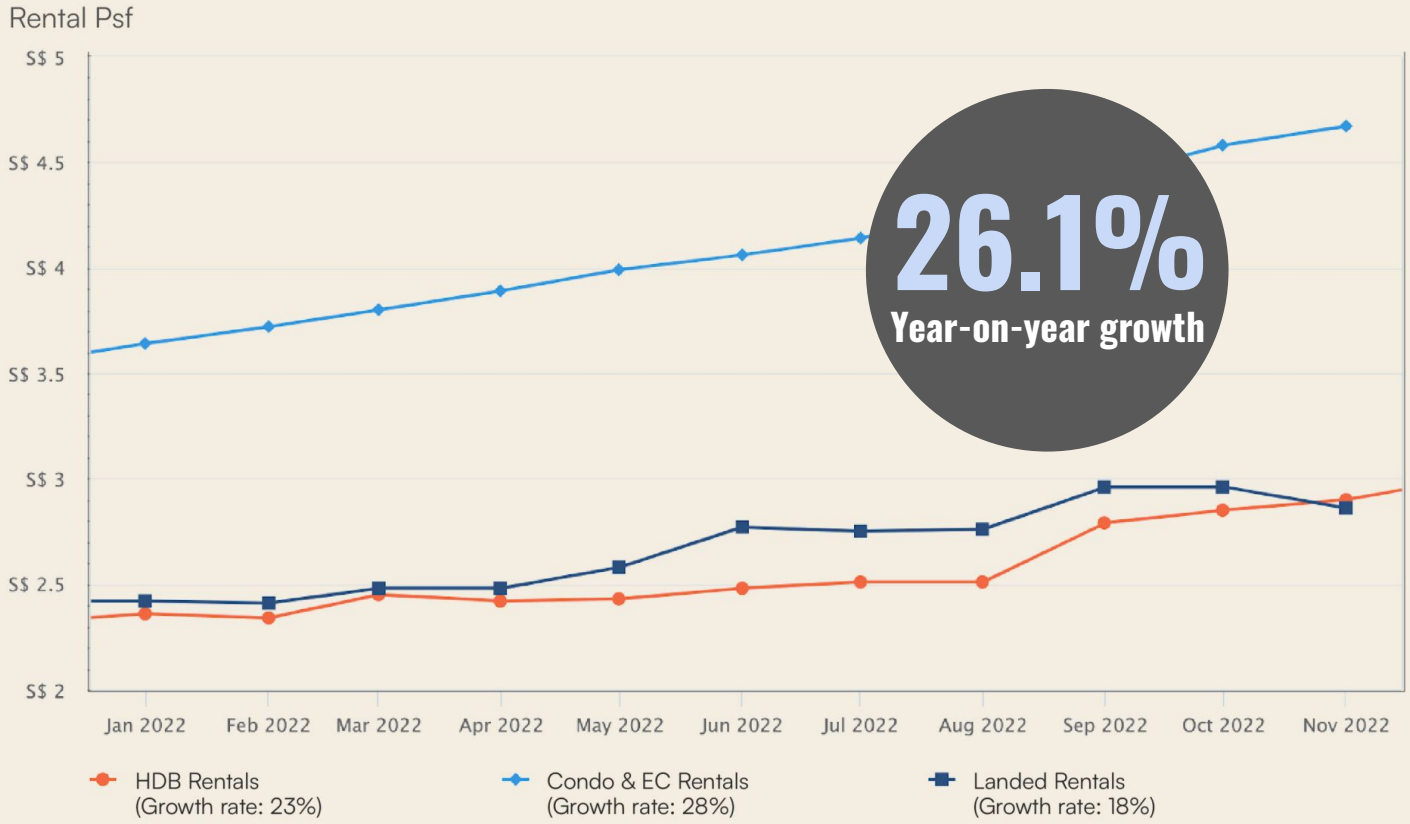
Source: https://www.propertylimbrothers.com/2022-singapore-real-estate-market-wrapped-key-statistics-and-trends/

Context



Source: https://www.propertylimbrothers.com/2022-singapore-real-estate-market-wrapped-key-statistics-and-trends/

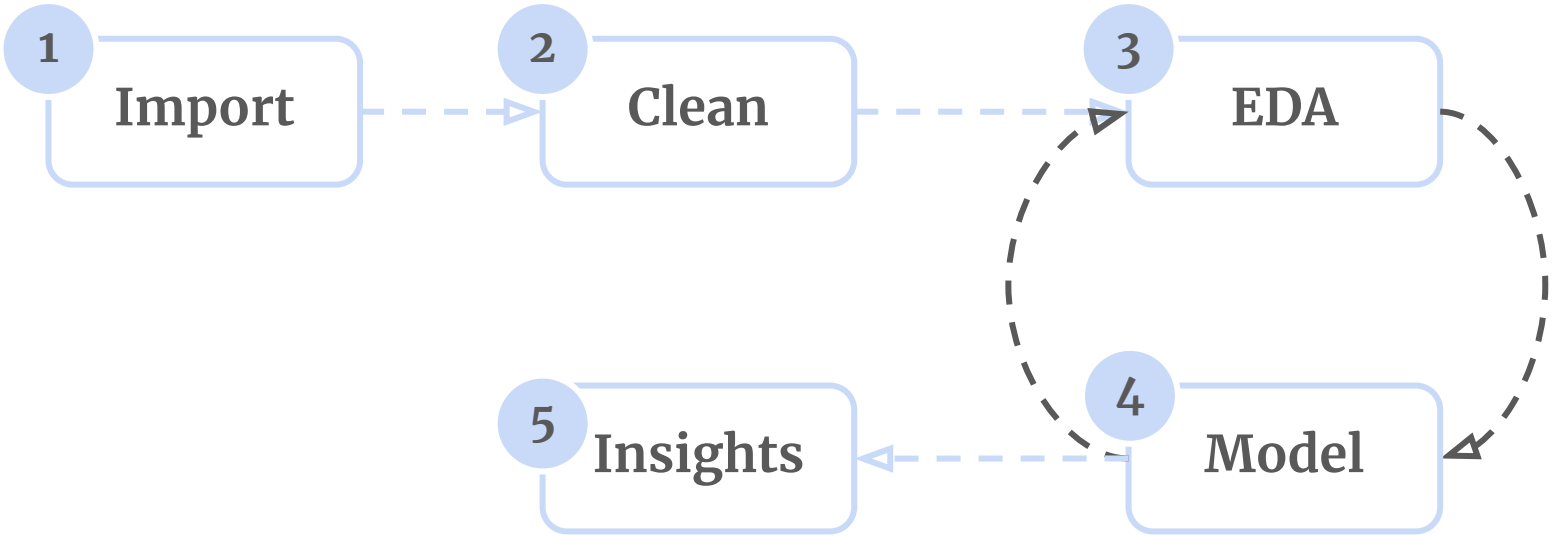
Context



Source: https://www.propertylimbrothers.com/2022-singapore-real-estate-market-wrapped-key-statistics-and-trends/


2
DATA
EXPLORATION

Workflow



Workflow

1

Import

First Look at Data

Number of Data Points

150,634

Number of Features

77

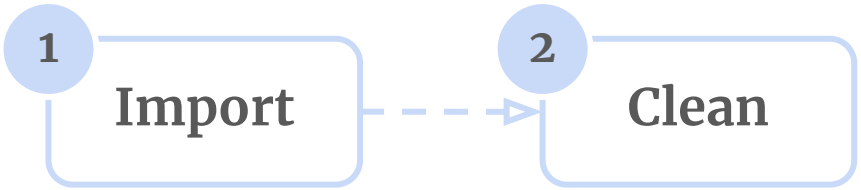
Time Period

March 2012 to April 2021

Null Values

Observed in 7 features

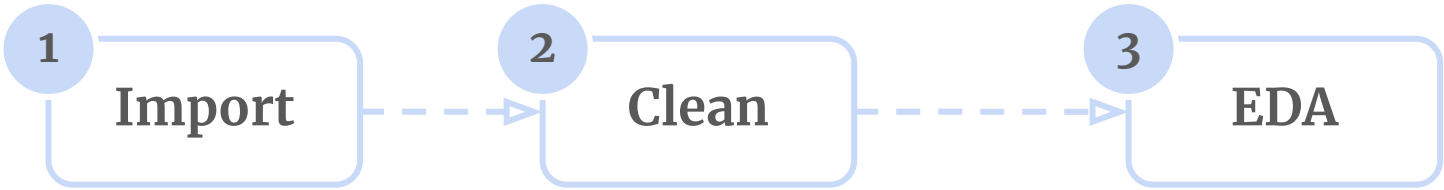
Workflow



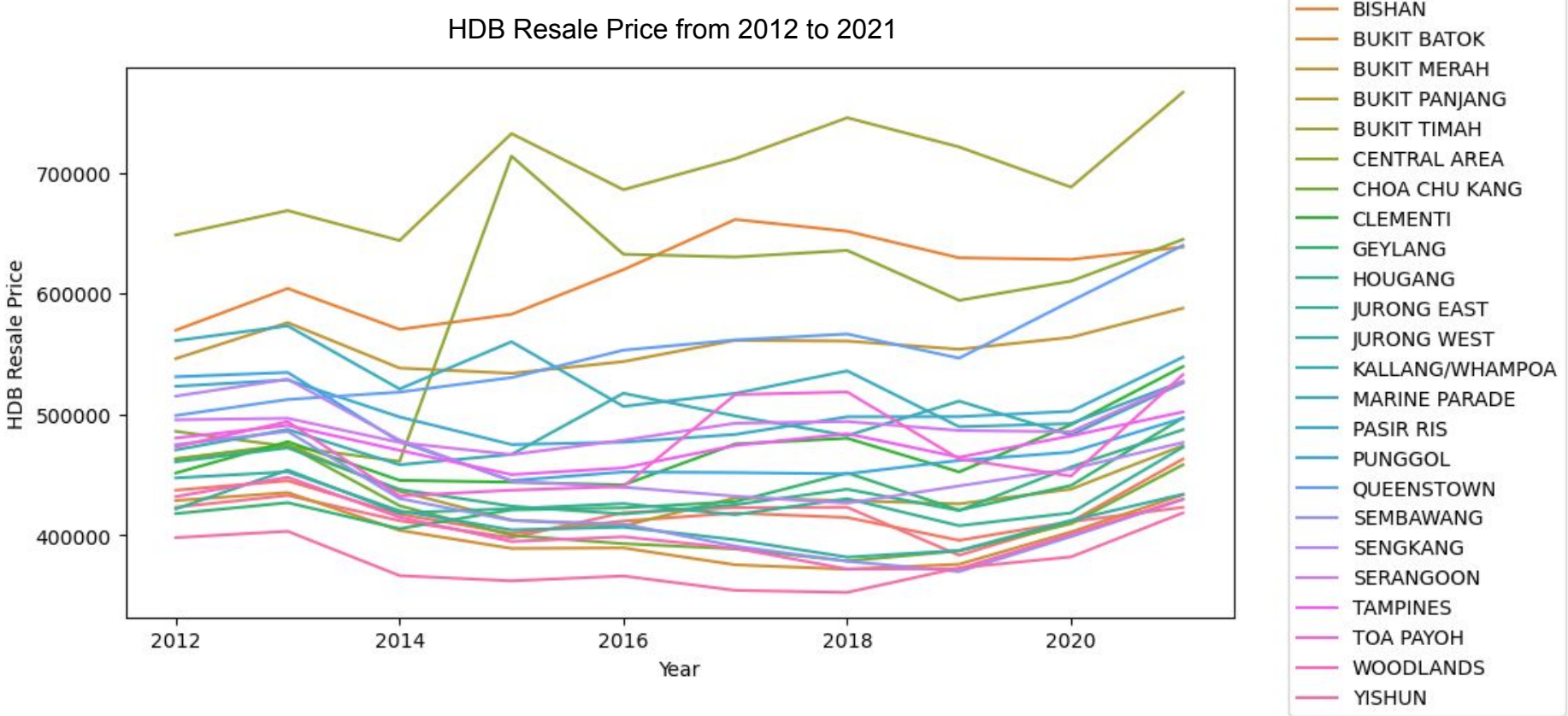
Missing Values

	No. of Null Values	Pct. of Null Values	
mall_nearest_distance	750	0.50%	Populate NaN with Pythagoras
mall_within_2km	1,940	1.29%	Populate NaN with 0
mall_within_1km	25,426	16.88%	
hawker_within_2km	29,202	19.39%	
hawker_within_1km	60,868	40.41%	
mall_within_500m	92,789	61.60%	
hawker_within_500m	97,390	64.65%	

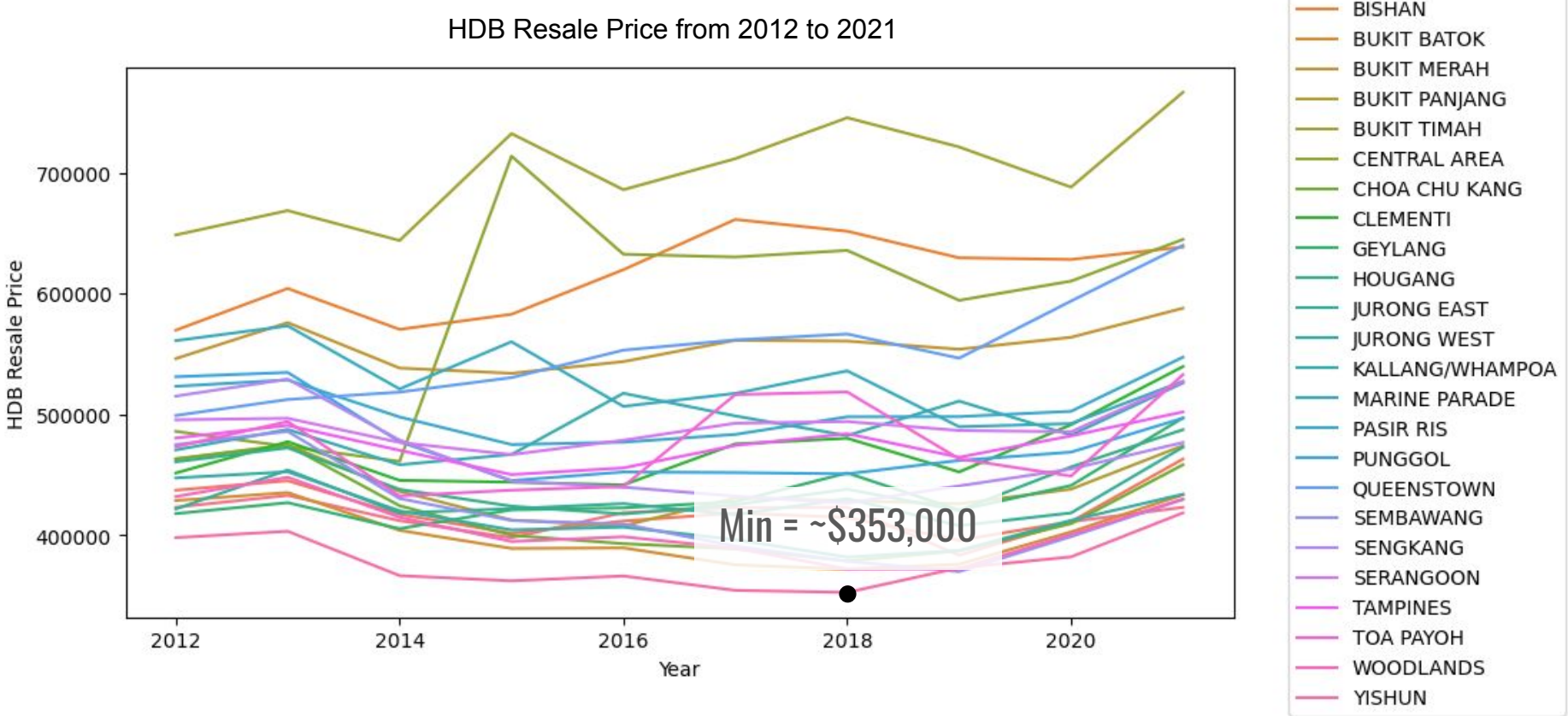
Workflow



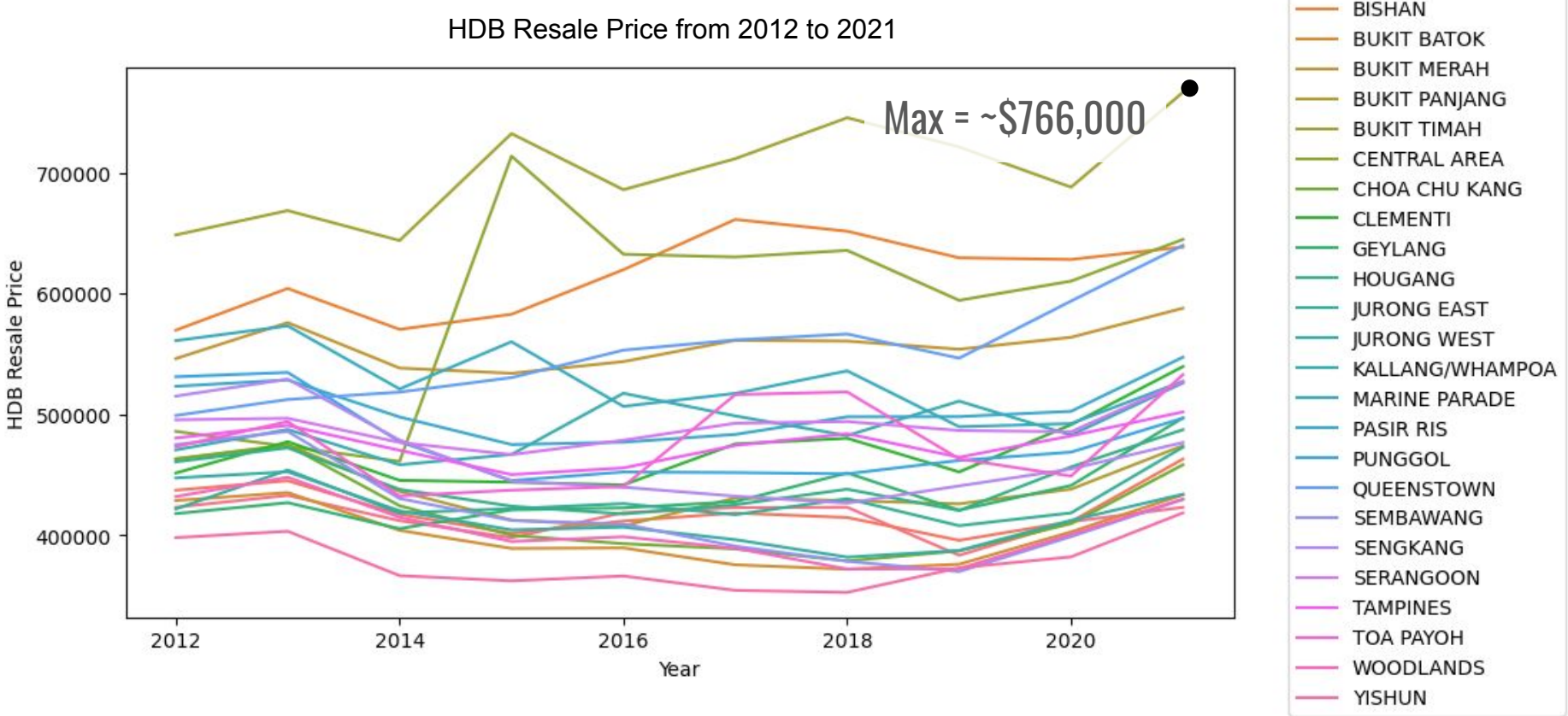
Average Resale Price Over Time



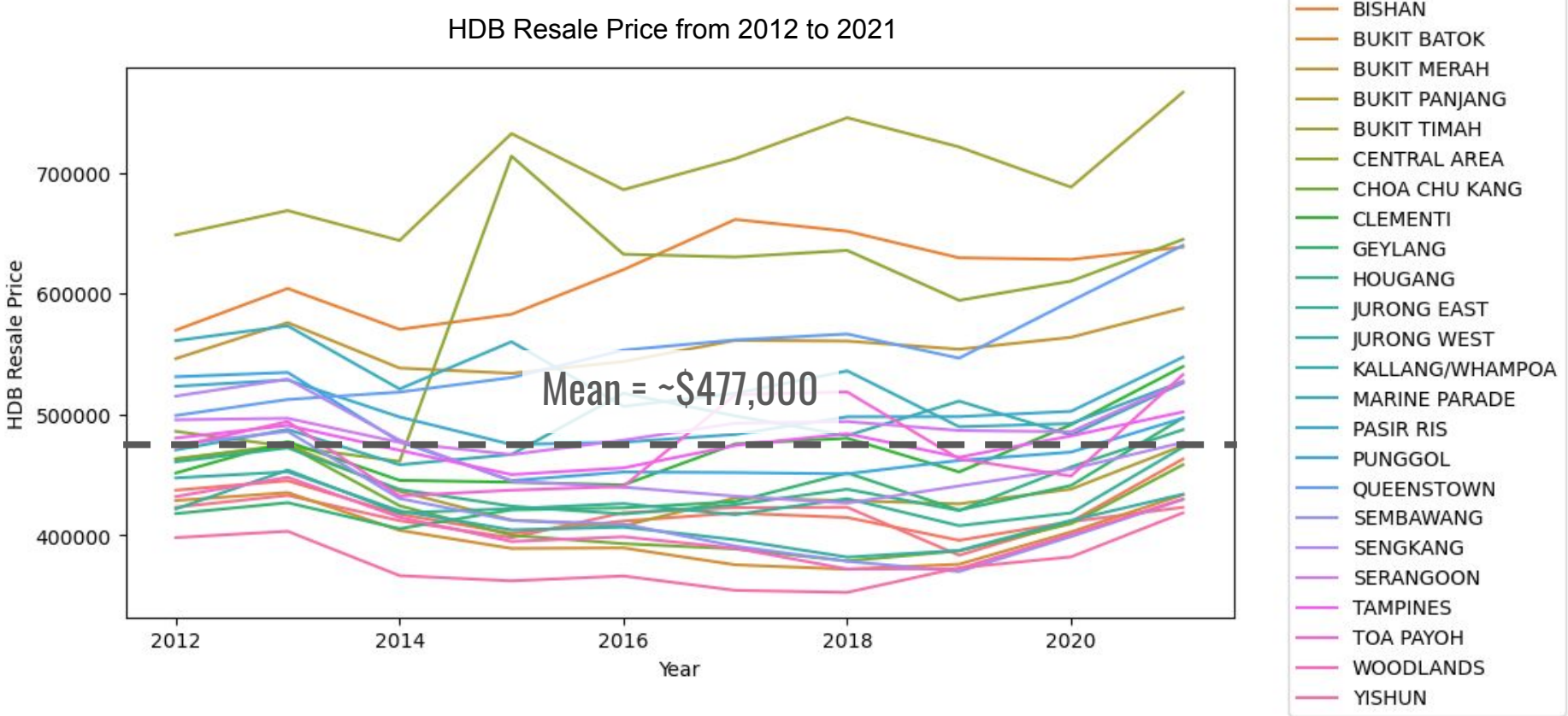
Average Resale Price Over Time



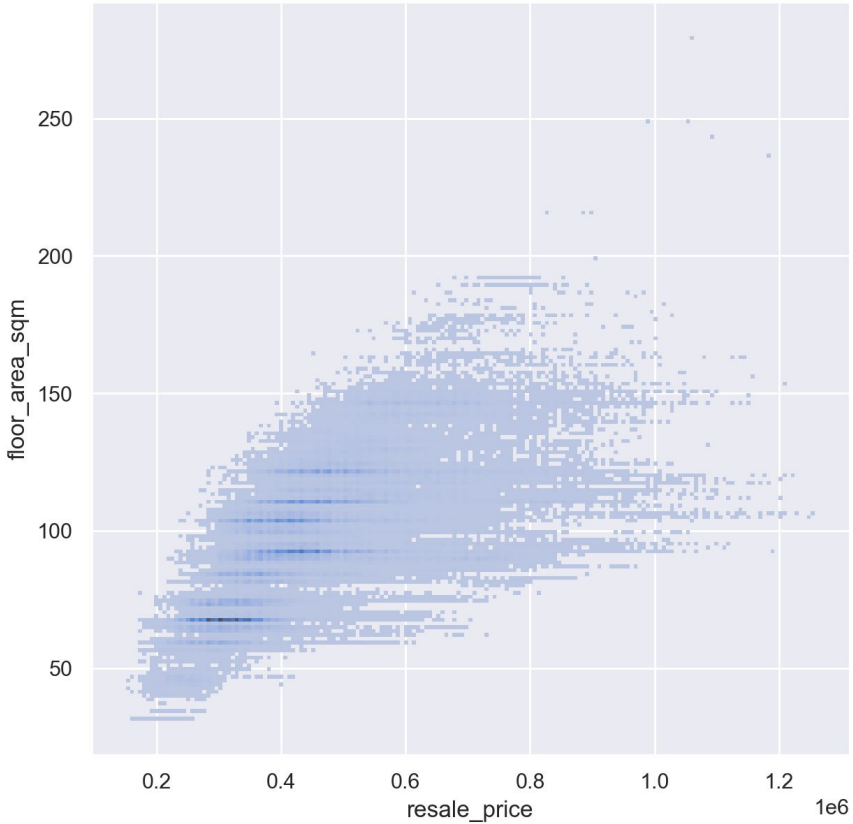
Average Resale Price Over Time



Average Resale Price Over Time



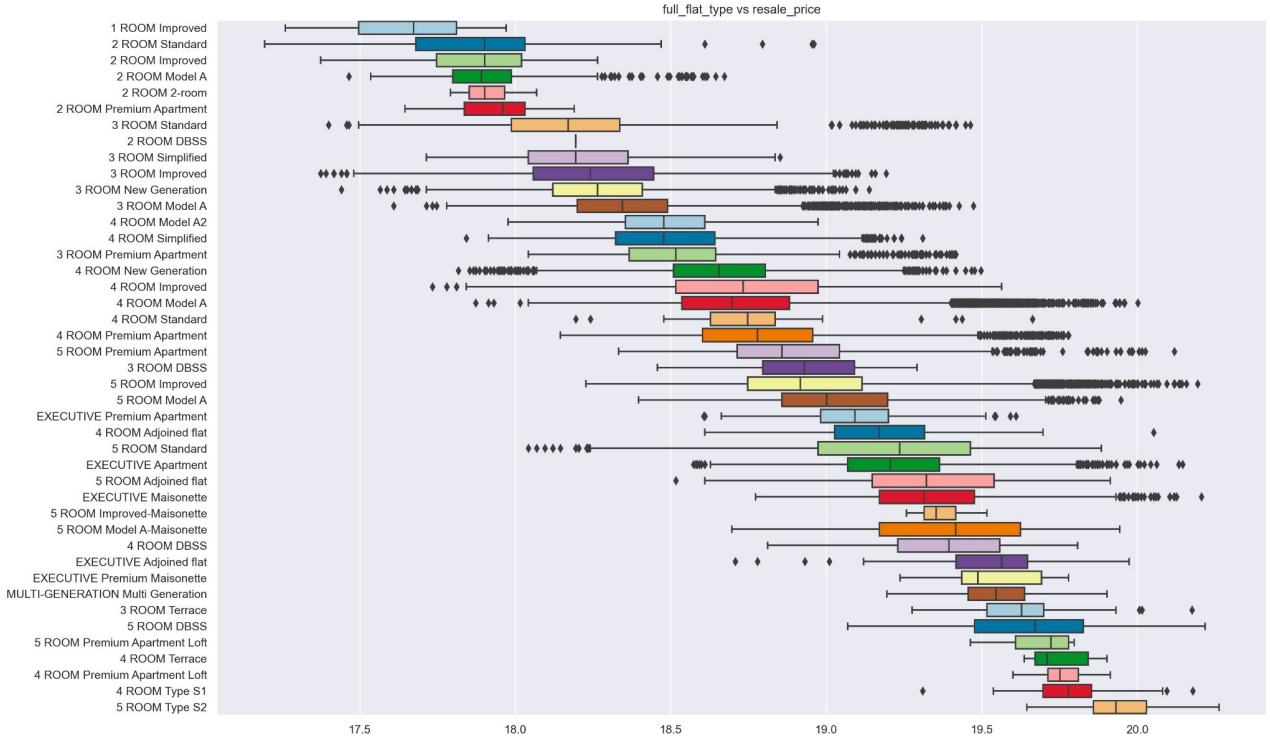
Correlation Between Resale Price and Floor Area (sqm)



Correlation

0.654

Full Flat Type vs Resale Price



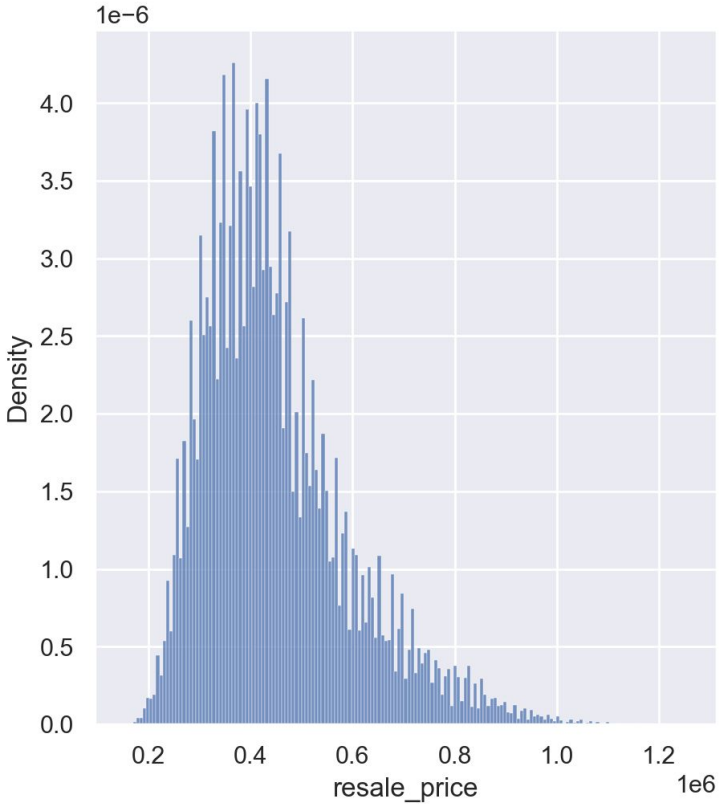
Correlation

0.744

Variability

High

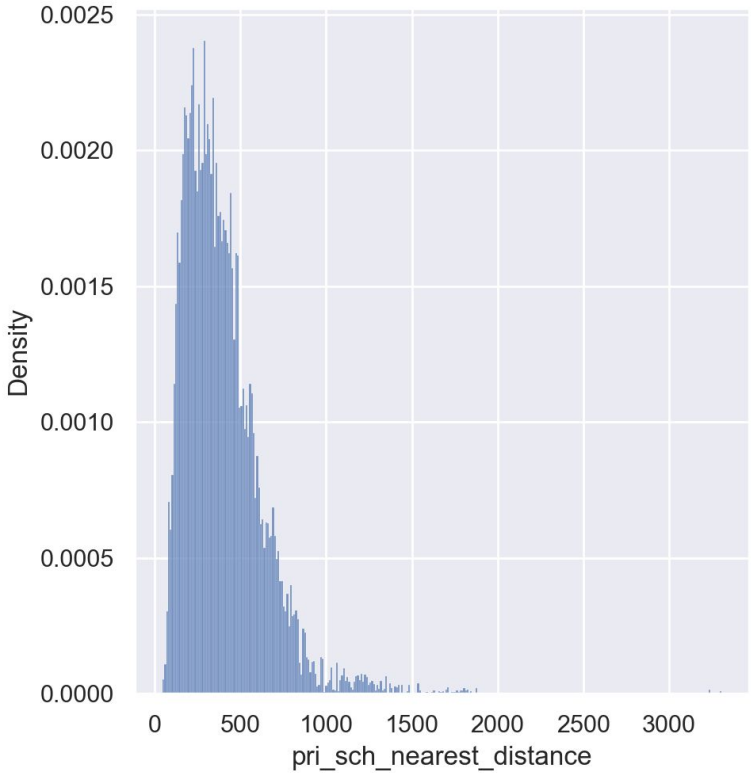
Distribution of Resale Price



Skew

1.084

Distribution of Primary School Nearest Distance

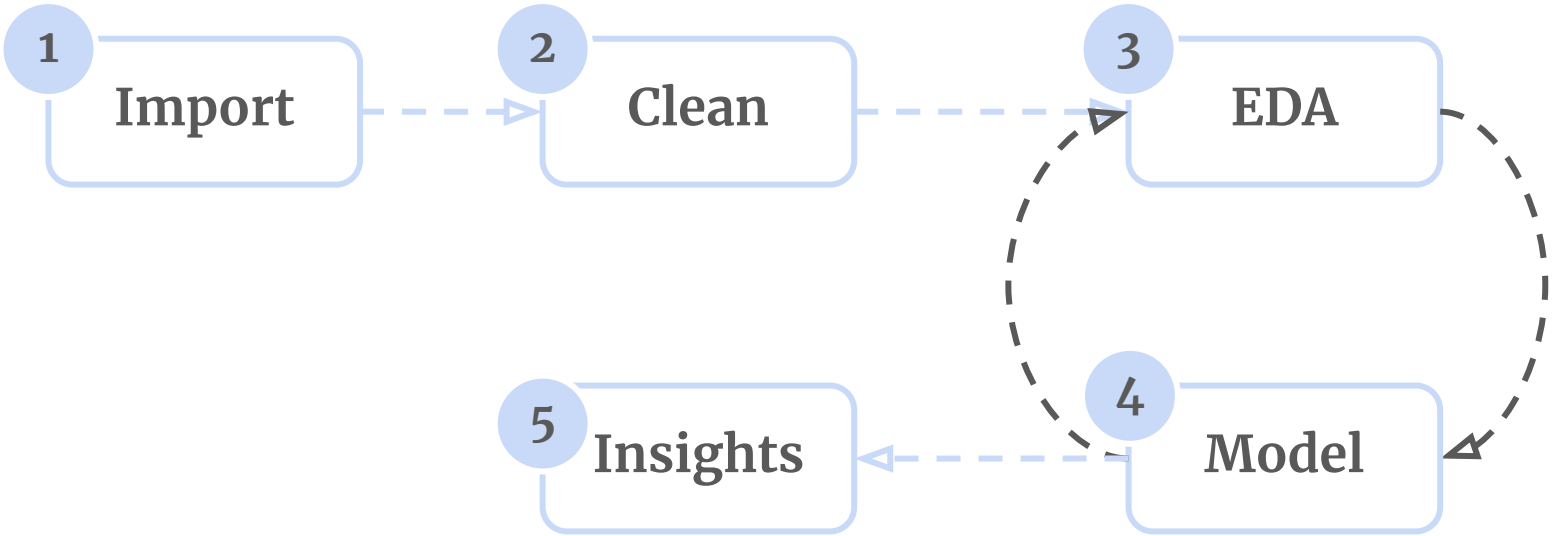


Skew

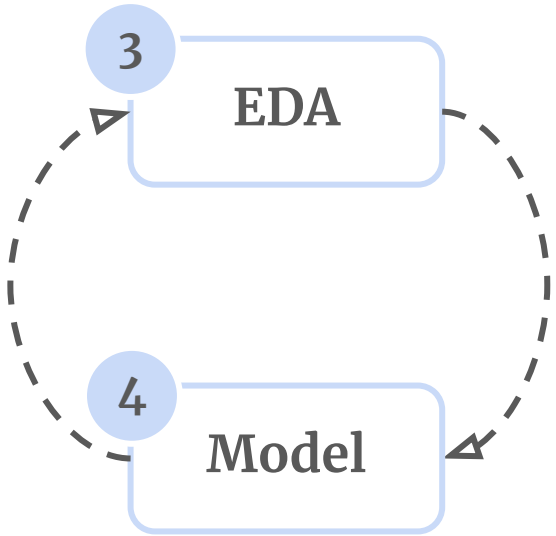
1.999


8
MODELLING
PART I

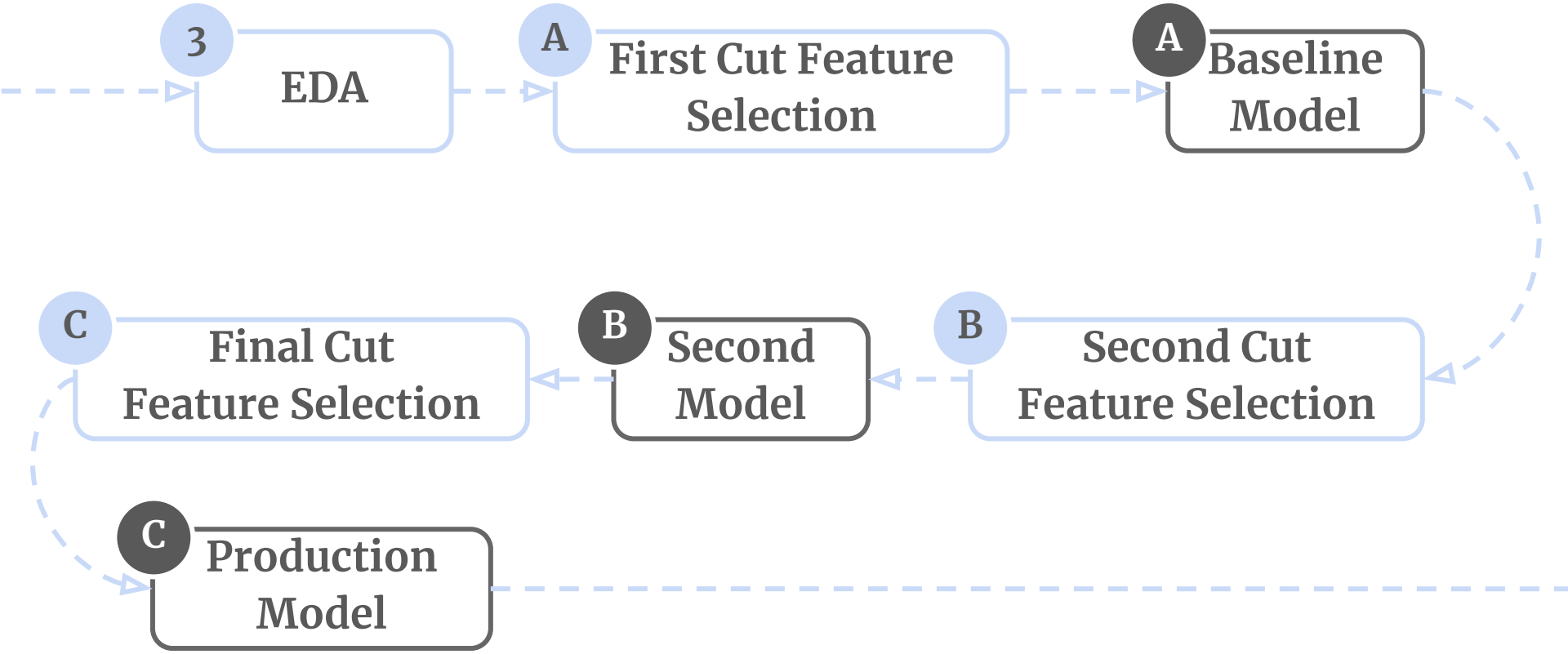
Workflow



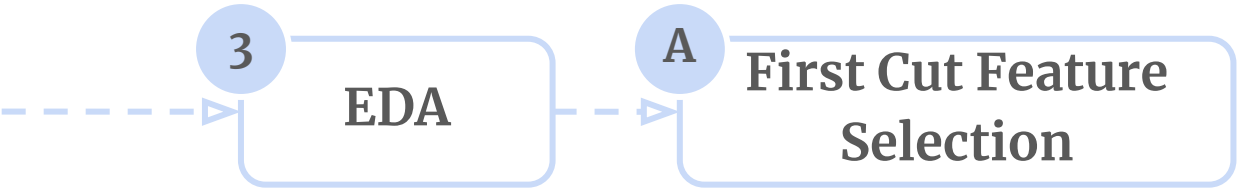
Workflow



Modelling Workflow



Modelling Workflow



First Cut Feature Selection Strategies

1

Domain Knowledge

- Unlikely Predictors
- Likely Predictors

2

High Collinearity

- HDB Age
- Year Completed
- Lease Commence Date

3

Overlapping Features

- Floor Area, in sqm
- Floor Area, in sqft

Original Number of Features

77

Number of features

>30,000

Number of features after one-hot encoding

One-hot encoding in machine learning is the conversion of categorical information into a format that may be fed into machine learning algorithms to improve prediction accuracy.
Source: https://deepchecks.com/glossary/one-hot-encoding/#:~:text=One%2Dhot%20encoding%20in%20machine,categorical%20data%20in%20machine%20learning.

Number of Features After First Cut

44

Number of features

91

Number of features after one-hot encoding

One-hot encoding in machine learning is the conversion of categorical information into a format that may be fed into machine learning algorithms to improve prediction accuracy.
Source: https://deepchecks.com/glossary/one-hot-encoding/#:~:text=One%2Dhot%20encoding%20in%20machine,categorical%20data%20in%20machine%20learning.

Modelling Workflow

4 MODELLING PART II

Modelling: Preprocessing

1

Dummify

- Transform categorical variables into numeric variables

2

Train-Test Split

- Split our data into 2 sets, one for training and one for testing

3

Scale

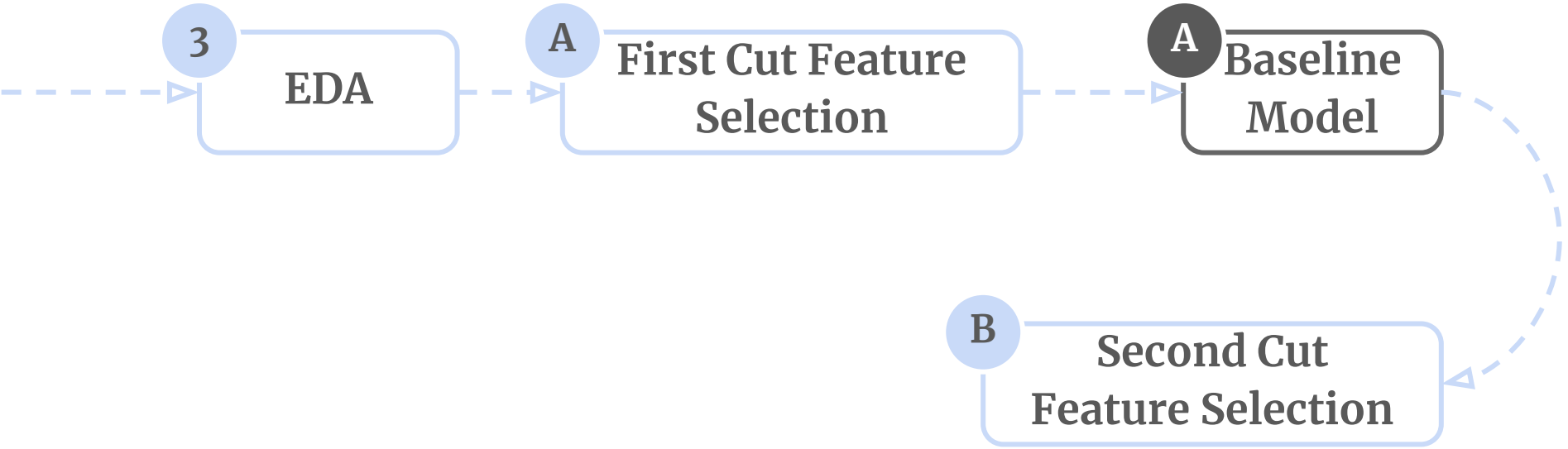
- To convert variables to be on a similar scale

Baseline Model: RMSE

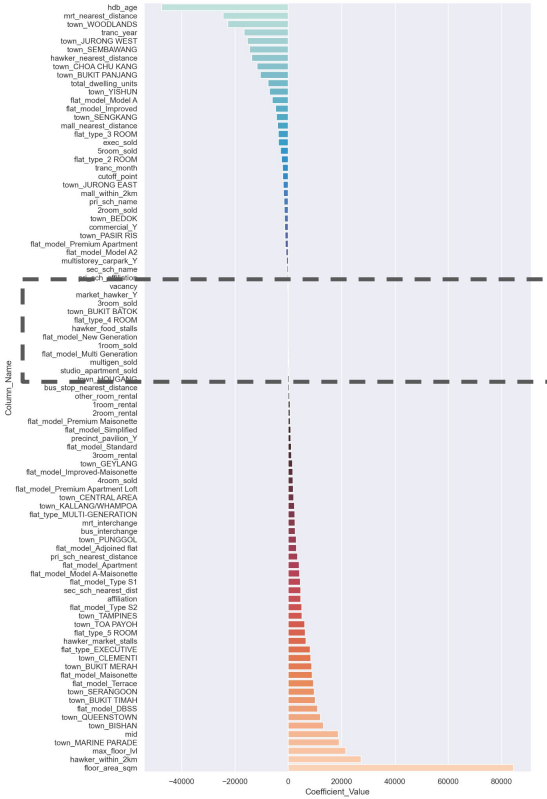
	Train	Test
Linear	47,727	47,249
Ridge	47,727	47,249
Lasso	47,745	47,273

RMSE (Root mean-squared-error) is a measure of approximate error for each prediction. The lower, the better.

Modelling Workflow

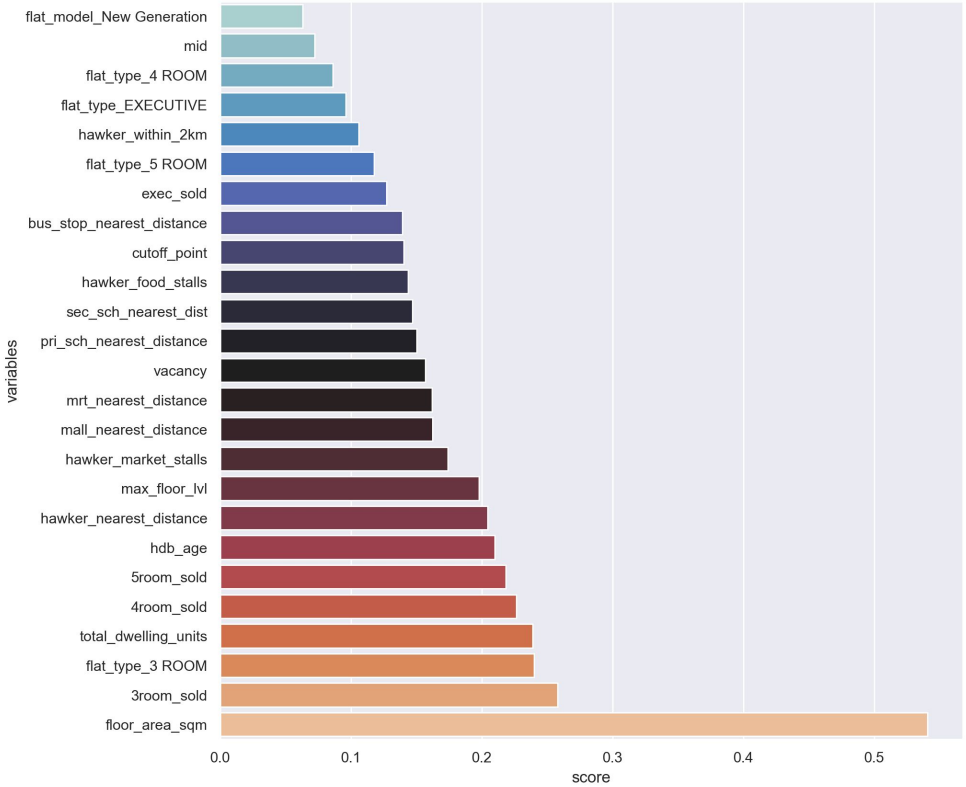


Second Cut Feature Selection Strategies



- Lasso
- Insignificant Variables

Second Cut Feature Selection Strategies



- **Automatic Feature Selection**
  Select K Best

Second Cut Feature Selection Strategies

Number of Features After Second Cut

16

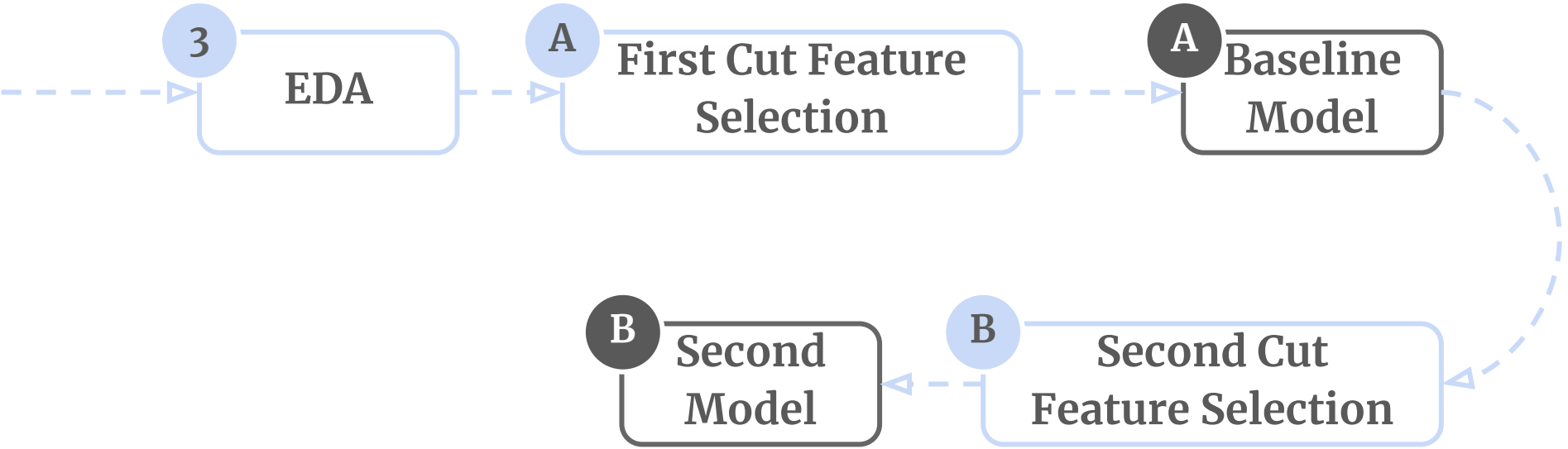
Number of features

83

Number of features after one-hot encoding

One-hot encoding in machine learning is the conversion of categorical information into a format that may be fed into machine learning algorithms to improve prediction accuracy.
Source: https://deepchecks.com/glossary/one-hot-encoding/#:~:text=One%2Dhot%20encoding%20in%20machine,categorical%20data%20in%20machine%20learning.

Modelling Workflow

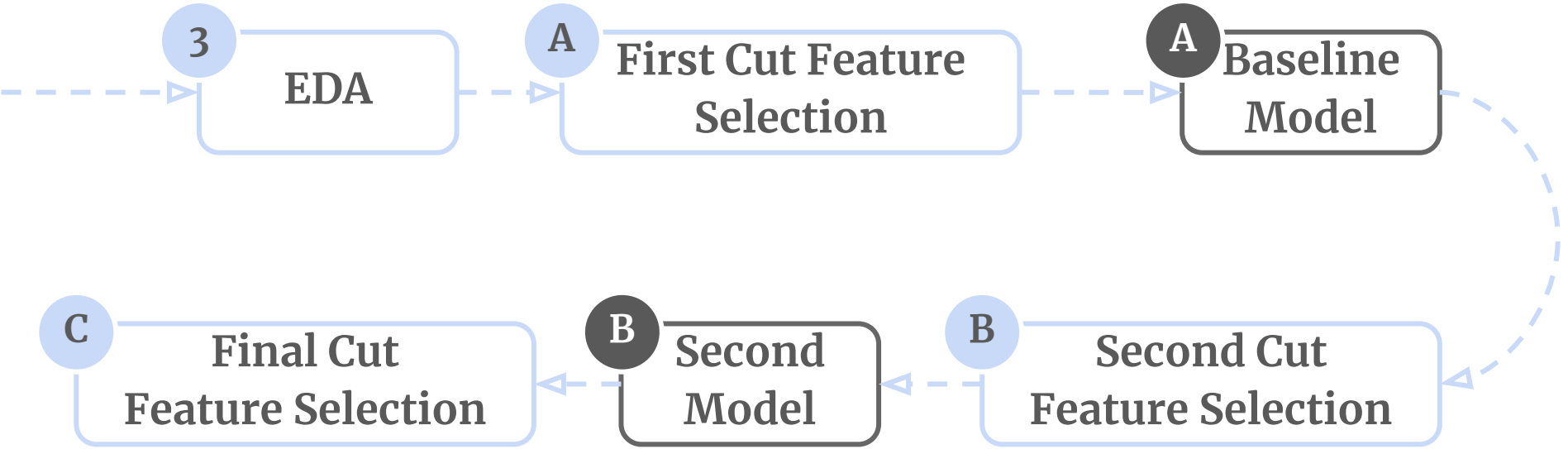


Second Model: RMSE

	Train	Test
Linear	47,831	47,598
Ridge	47,830	47,599
Lasso	47,848	47,621

RMSE (Root mean-squared-error) is a measure of approximate error for each prediction. The lower, the better.

Modelling Workflow



Final Cut Feature Selection Strategies

	Column_Name	Coefficient_Value
3	hdb_age	-48298.746776
10	mrt_nearest_distance	-24024.064419
Group 1 38	town_WOODLANDS	-21132.970301
Group 2 33	town_SEMBAWANG	-14333.776410
27	town_JURONG WEST	-13168.031203
Group 3 22	town_CHOA CHU KANG	-10892.484487
19	town_BUKIT PANJANG	-9445.409603
Group 4 39	town_YISHUN	-6192.472629

Final Cut Feature Selection Strategies

Categorical Features

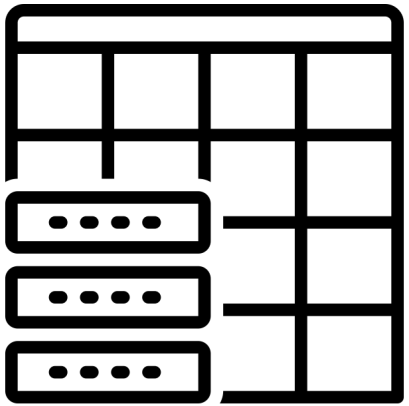
Town, Full Flat Type

Number of Unique Features in **Town**

25

Number of Unique Features in **Full Flat Type**

42



Final Cut Feature Selection Strategies

Categorical Features

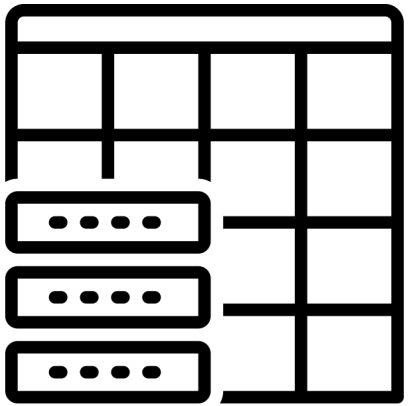
Town, Full Flat Type

Town Unique Features Reduced to

9

Full Flat Type Unique Features Reduced to

11



Number of Features After Final Cut

16

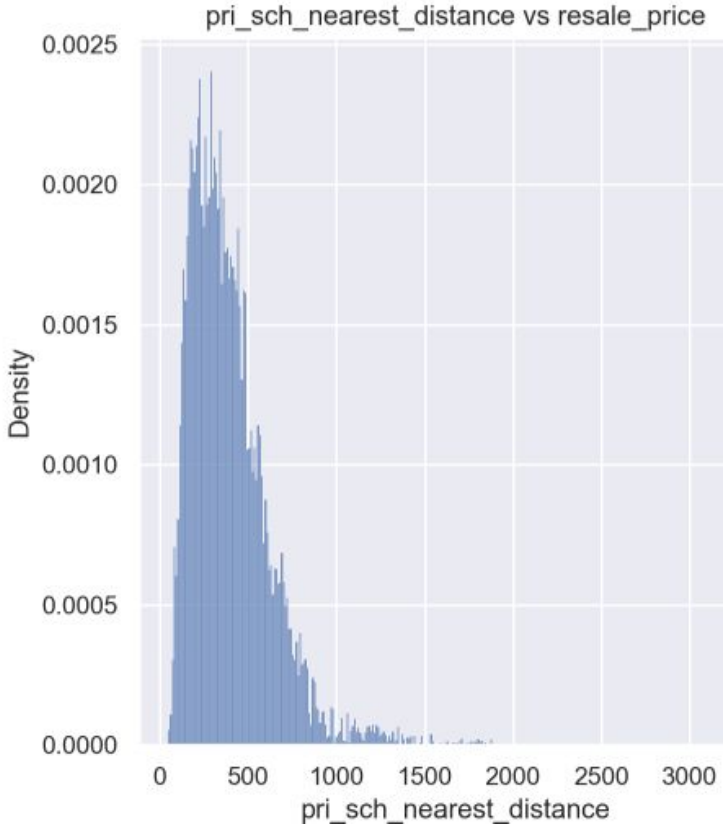
Number of features

34

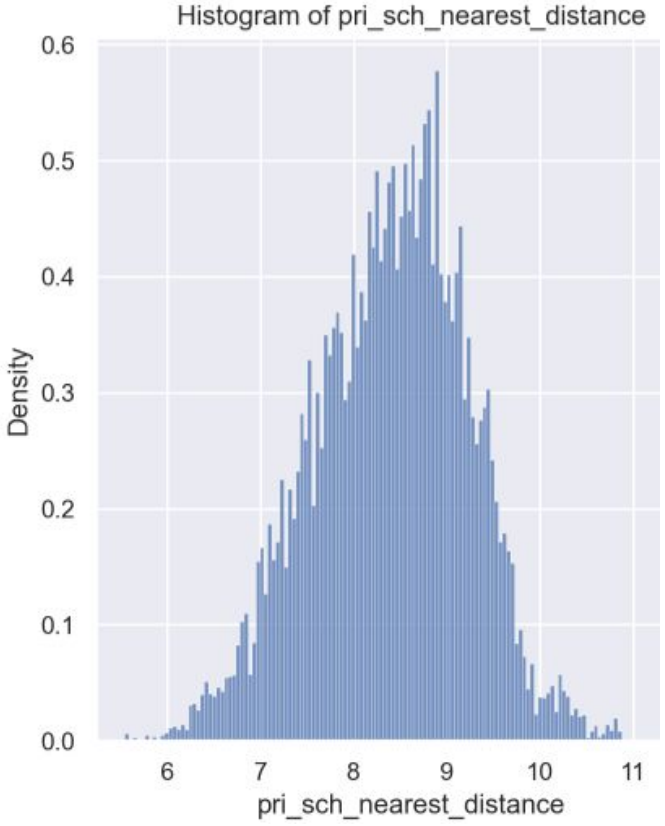
Number of features after one-hot encoding

One-hot encoding in machine learning is the conversion of categorical information into a format that may be fed into machine learning algorithms to improve prediction accuracy.
Source: https://deepchecks.com/glossary/one-hot-encoding/#:~:text=One%2Dhot%20encoding%20in%20machine,categorical%20data%20in%20machine%20learning.

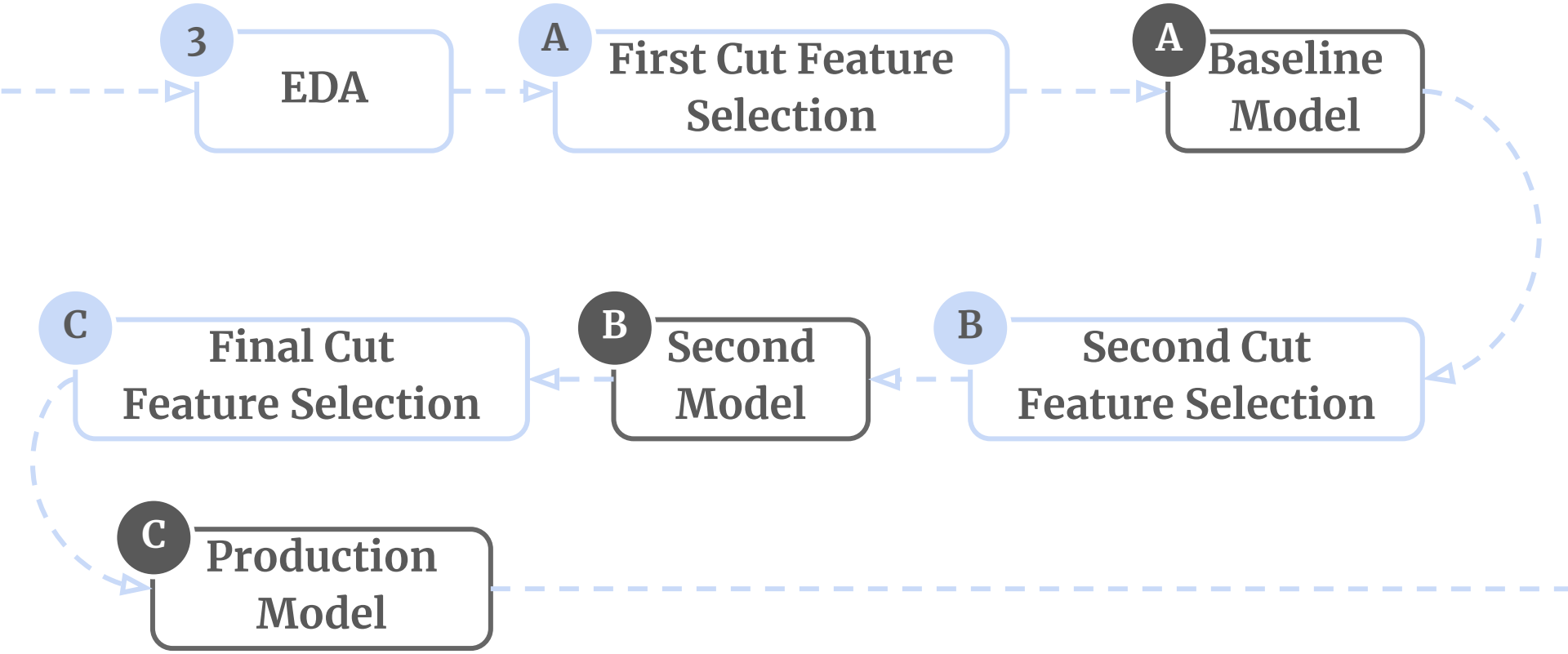
Feature Transformation: Logging non-normal features



Log Transformation



Modelling Workflow



StandardScaler vs PowerTransformer

	Standard Scaler Train	Standard Scaler Test	Power Transformer Train	Power Transformer Test
Linear	47,006	46,406	46,298	46,088
Ridge	47,001	46,405	46,298	46,088
Lasso	47,011	46,429	46,301	46,105

Power Transformer performed **very slightly better** than Standard Scaler

RMSE (Root mean-squared-error) is a measure of approximate error for each prediction. The lower, the better.

Production Model: RMSE

	Train	Test
Linear	46,298	46,088
Ridge	46,298	46,088
Lasso	46,301	46,105

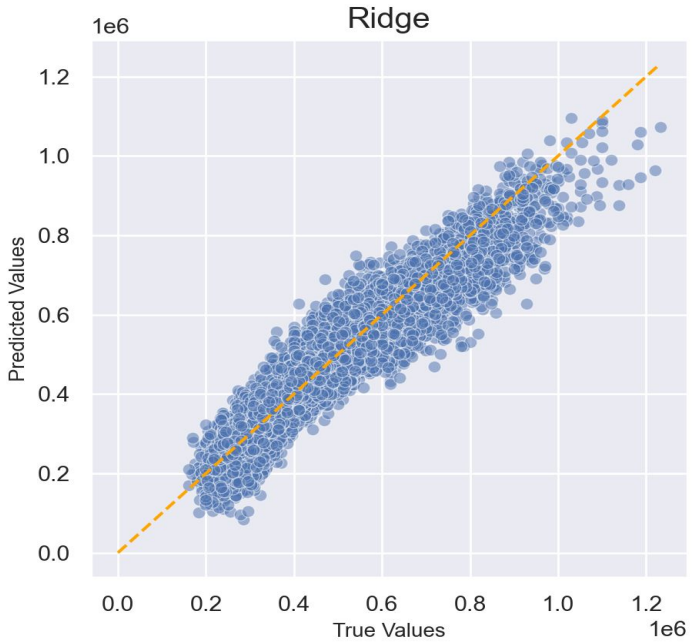
RMSE (Root mean-squared-error) is a measure of approximate error for each prediction. The lower, the better.

Modelling Summary

	No. of Features	No. of Features (After Dummifying)	R^2 Train	R^2 Test	RMSE Train	RMSE Test
Baseline	44	91	0.889	0.890	47,727	47,249
Second Model	16	83	0.889	0.889	47,830	47,599
Production Model	16	34	0.896	0.897	46,298	46,088

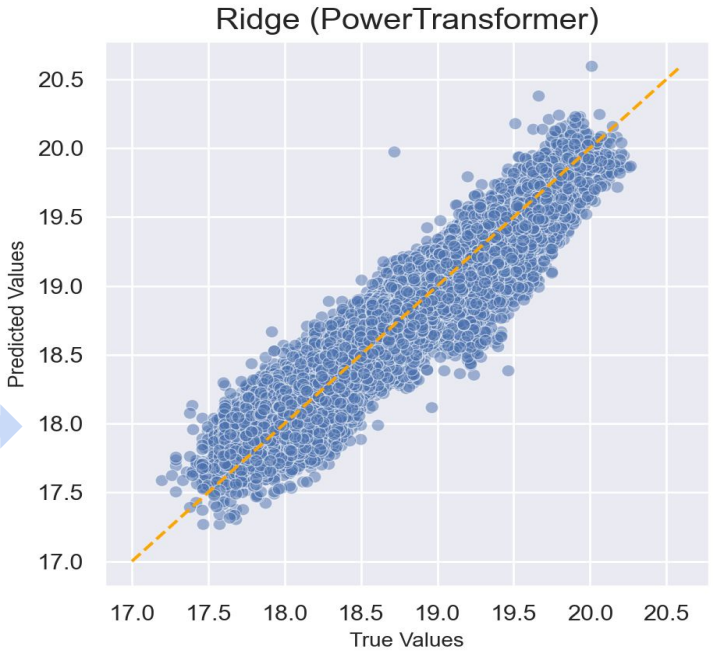
RMSE (Root mean-squared-error) is a measure of approximate error for each prediction. The lower, the better.
R^2 ranges from 0-1 (R^2 = 1 means our model explains 100% of the variance in resale price).

Modelling Summary



Baseline model

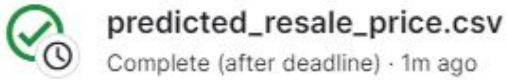
Improvement with **FEWER** Features



Production model

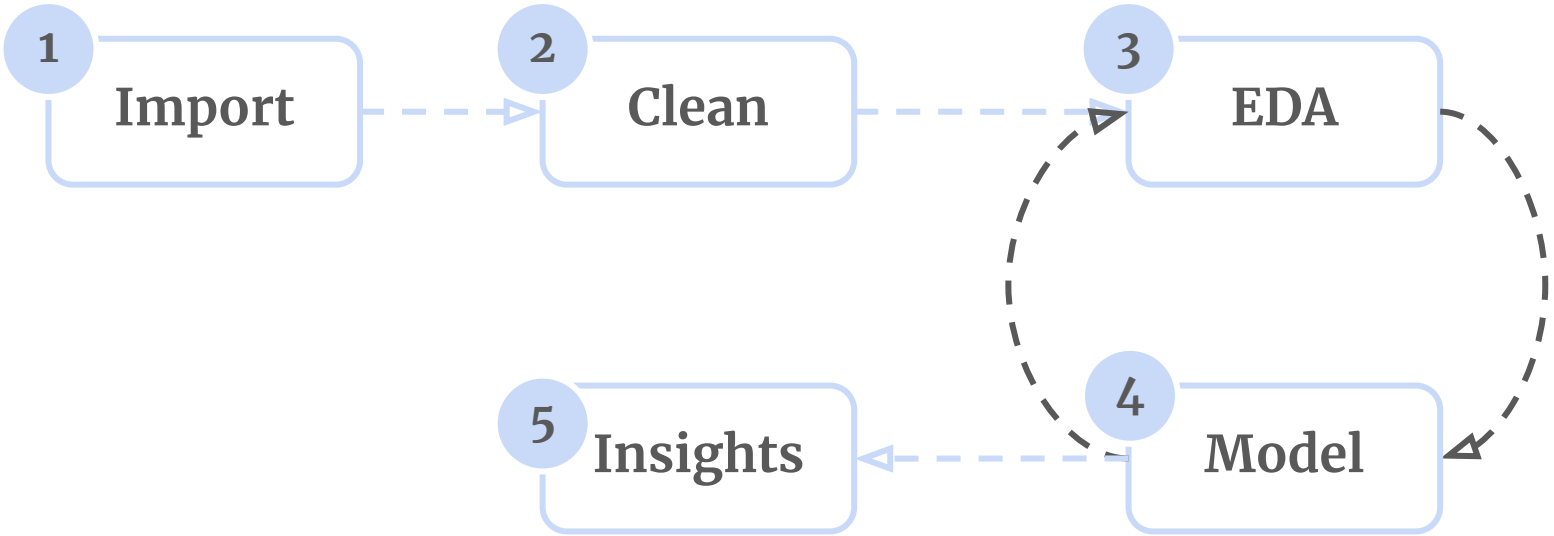
PowerTransformer provides non-linear transformations in which data is mapped to a normal distribution to stabilize variance and minimize skewness. Source: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

Kaggle Submission



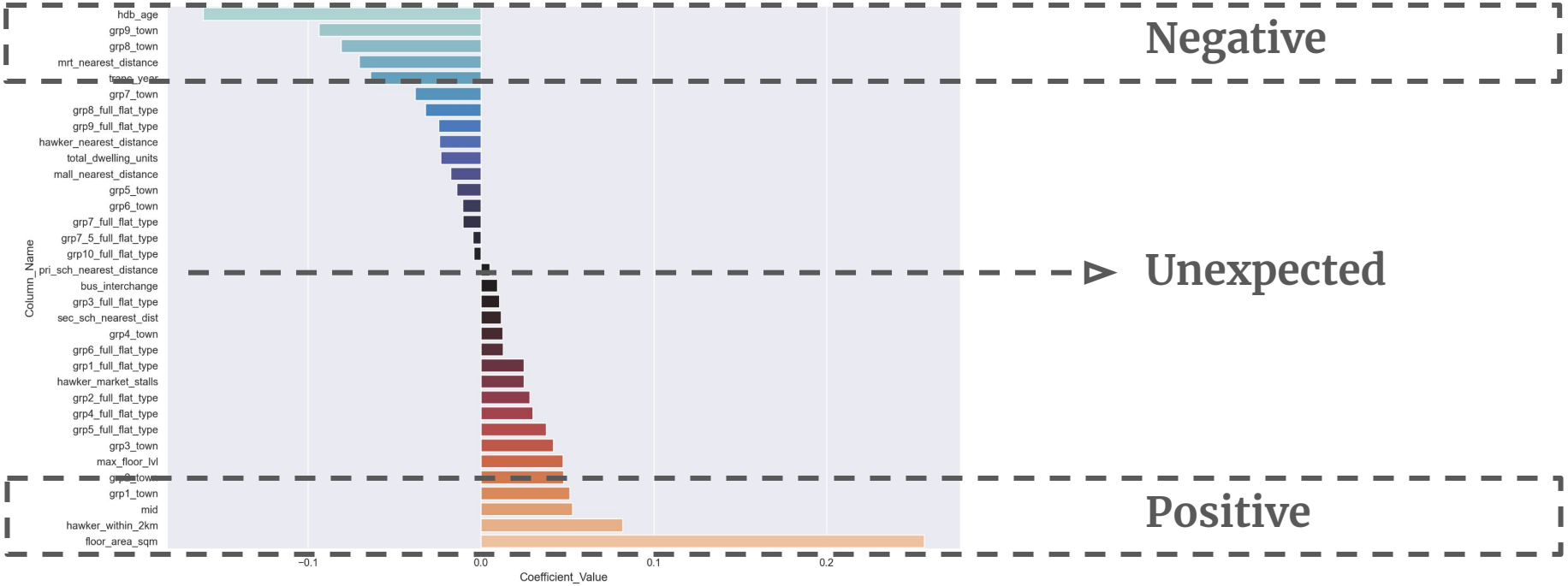
- RMSE: $46,000 to $47,000
- Similar to our train / test RMSE
- **Model is generalisable to unseen datasets!**

Workflow



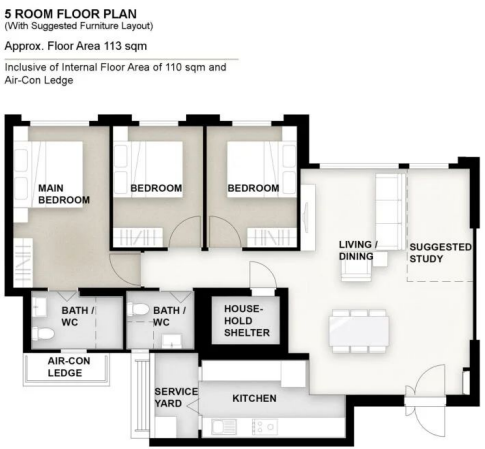
4 CONCLUSION

Key Predictors



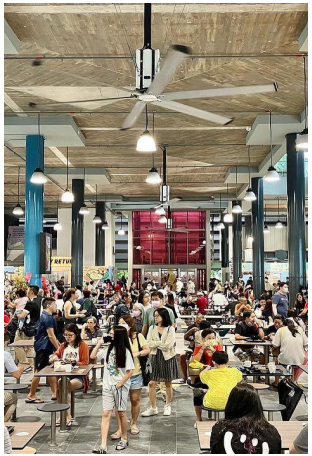
Prioritise

Bigger Floor Area



Higher Floor

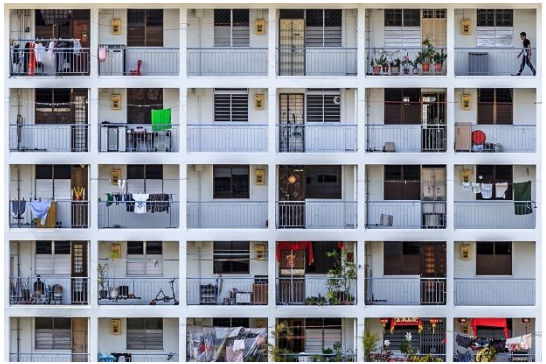
Hawker Centres



Source: https://www.hdb.gov.sg/cs/infoweb/homepage

Deprioritise

Older Flats

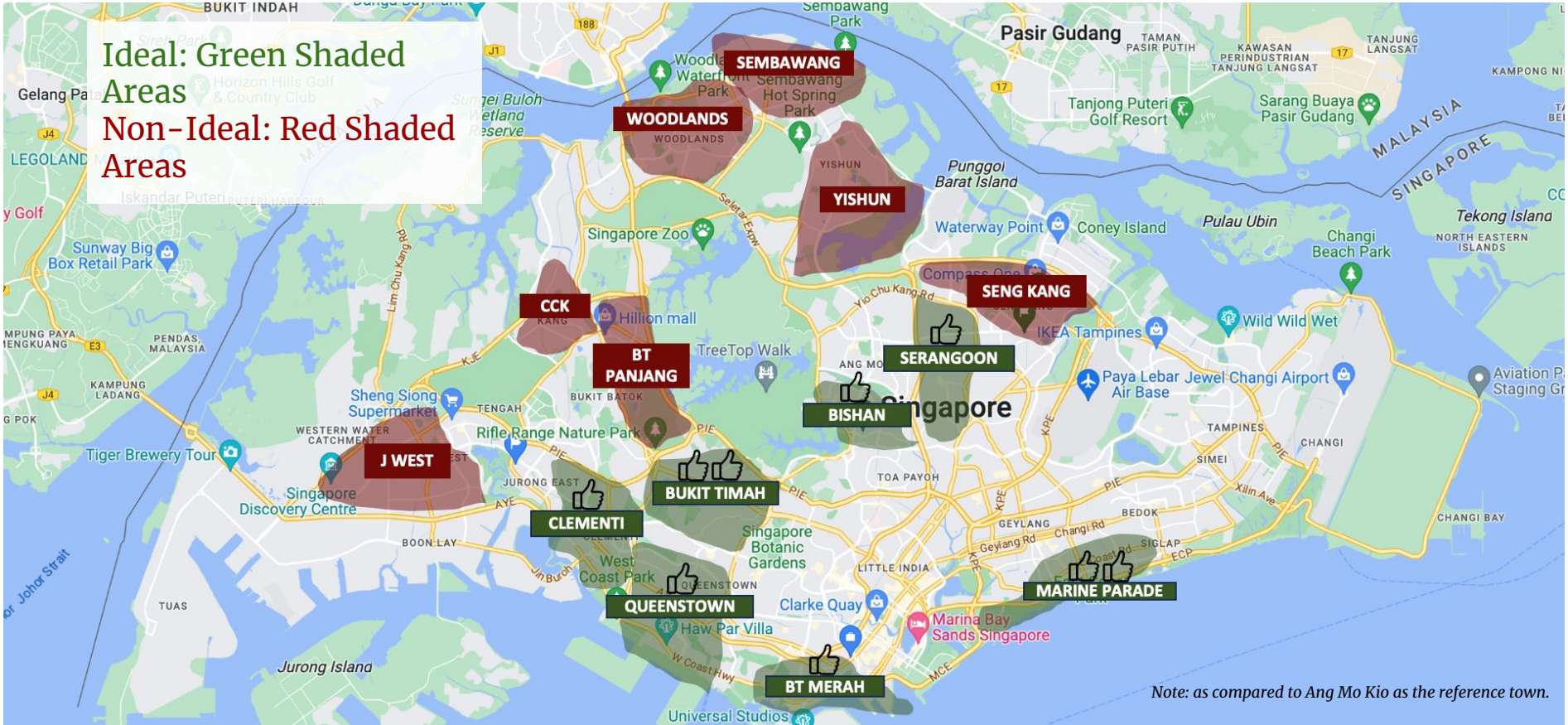


Far From MRT

Location

- Sembawang
- Woodlands
- Choa Chu Kang
- Jurong West
- Bukit Panjang

Ideal/Non-Ideal Locations



Note: as compared to Ang Mo Kio as the reference town.

Business Recommendations

1

Strategic Marketing

- Town
- Neighbourhood
- Block / Unit

2

Play Up Positive Features

- Sellers
- Buyers

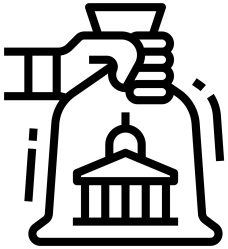
3

Build Reputation

- Regular Stream of Interesting and Relevant Content

Further Steps

Demographics



Government Subsidies

Interest Rates

Economy


5
PRODUCT


DEMO


Q&A