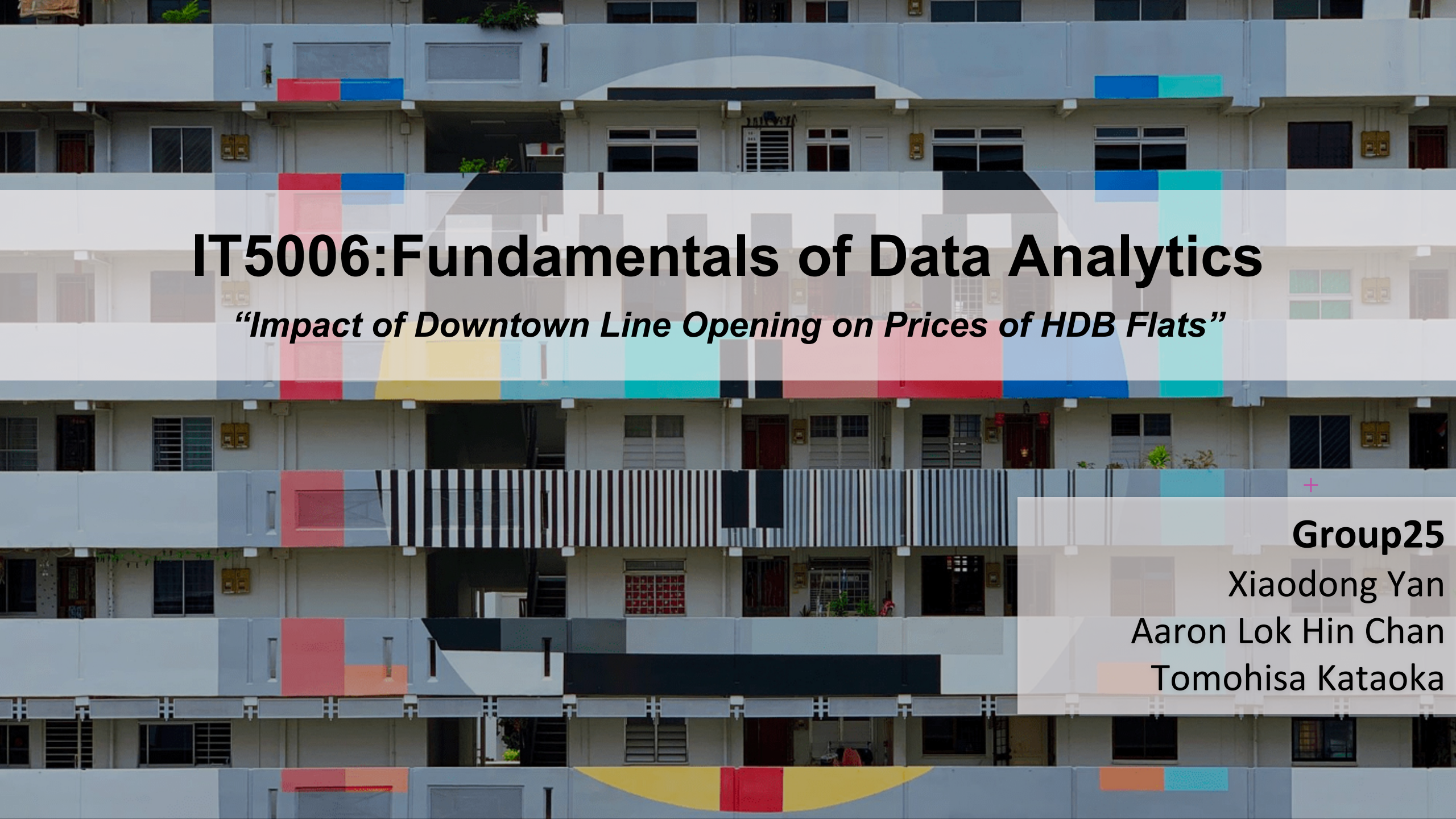


IT5006:Fundamentals of Data Analytics

"Impact of Downtown Line Opening on Prices of HDB Flats"

Group25

Xiaodong Yan

Aaron Lok Hin Chan

Tomohisa Kataoka

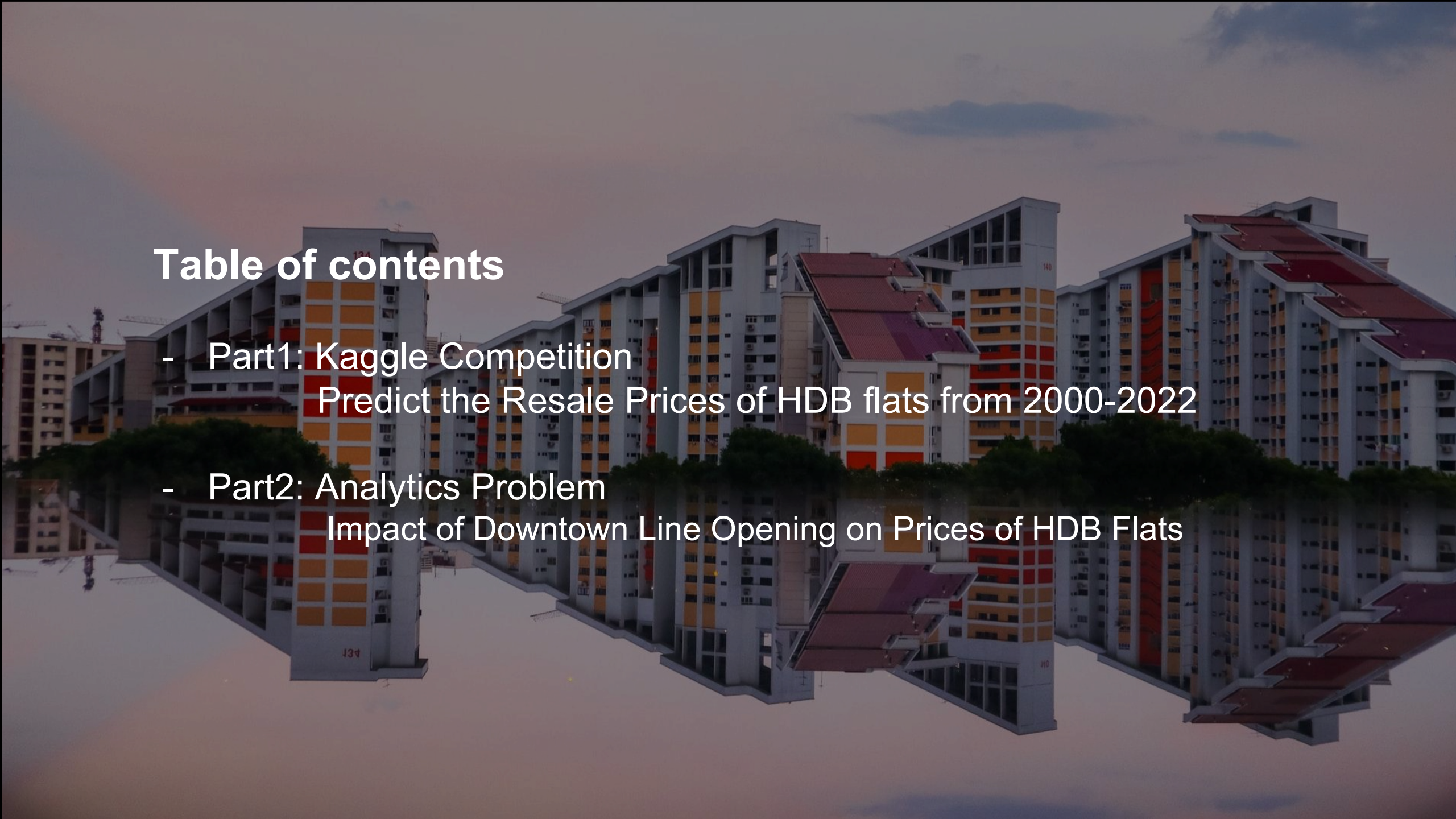


Table of contents

- Part1: Kaggle Competition
  Predict the Resale Prices of HDB flats from 2000-2022

- Part2: Analytics Problem
  Impact of Downtown Line Opening on Prices of HDB Flats

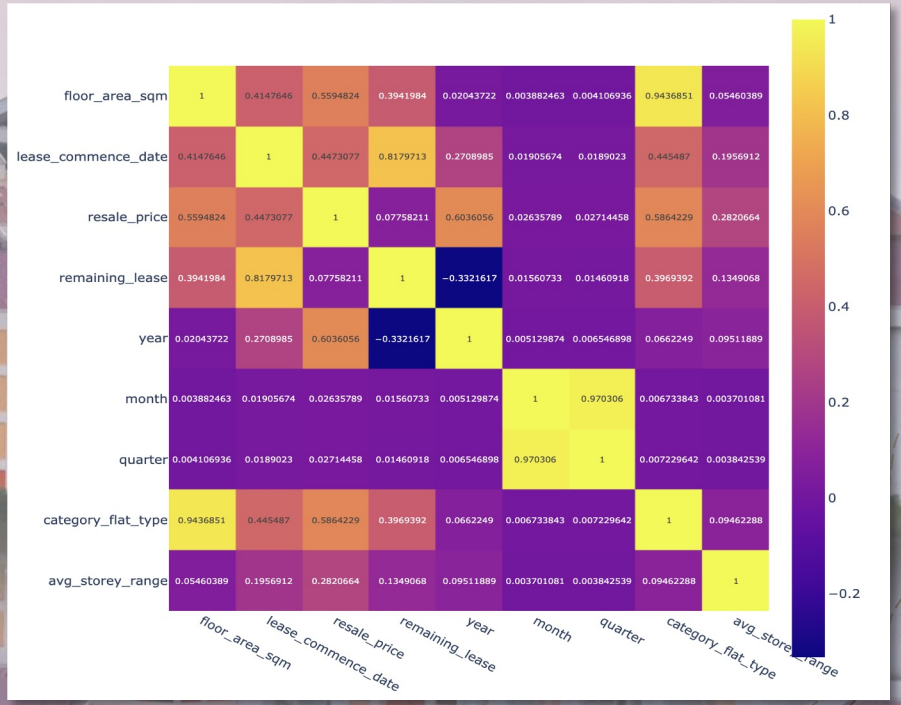
Kaggle Competition

Prediction Problem Statement:

To predict Resale Prices of HDB flats from 2000-2022

Exploratory Data Analysis

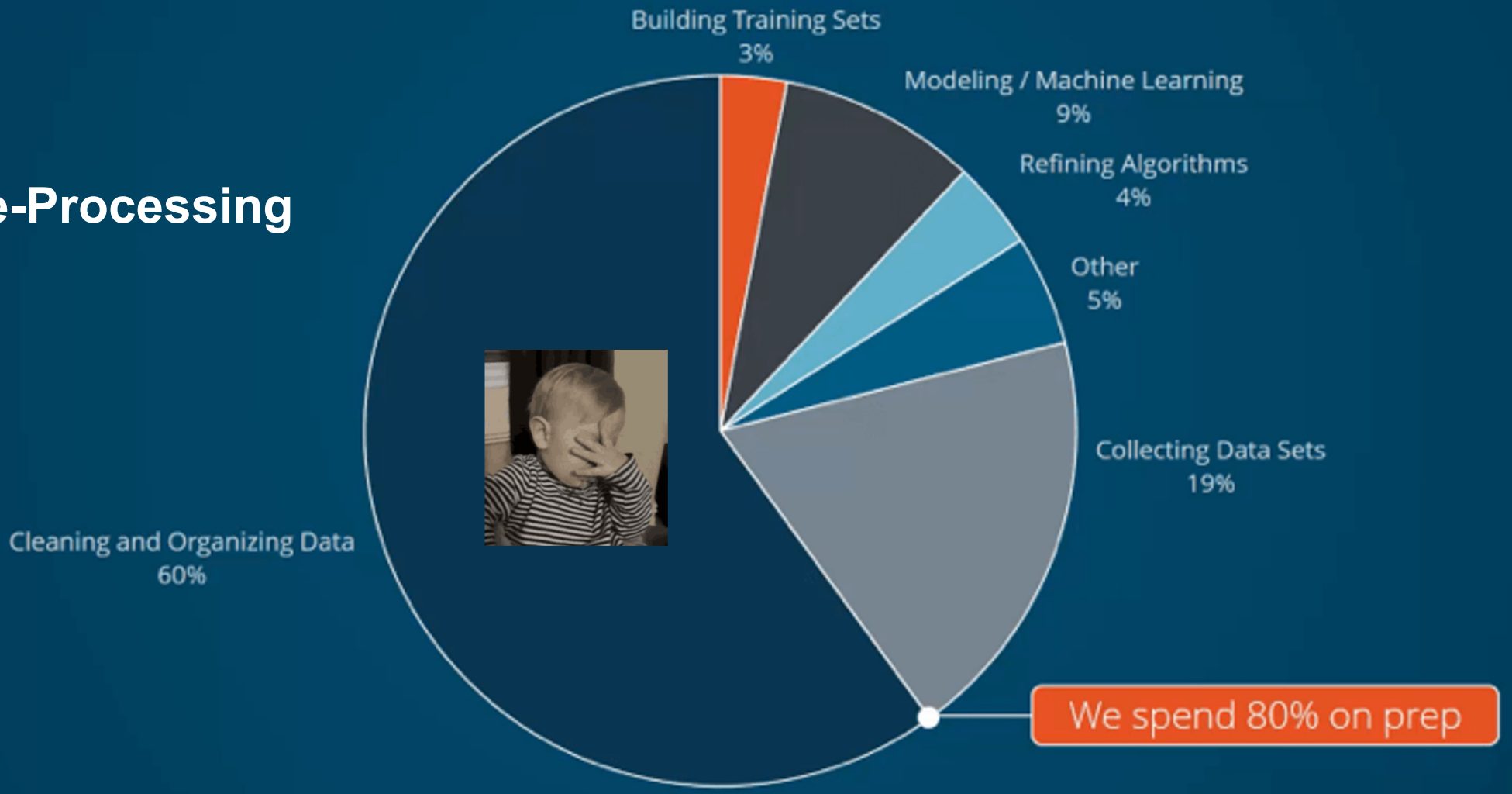
HOUSING & DEVELOPMENT BOARD



- Detect outliers, missing data
- Distributions
- Correlations, Multi-collinearity

What data scientists spend the most time doing

Data Pre-Processing



https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Pre-Processing

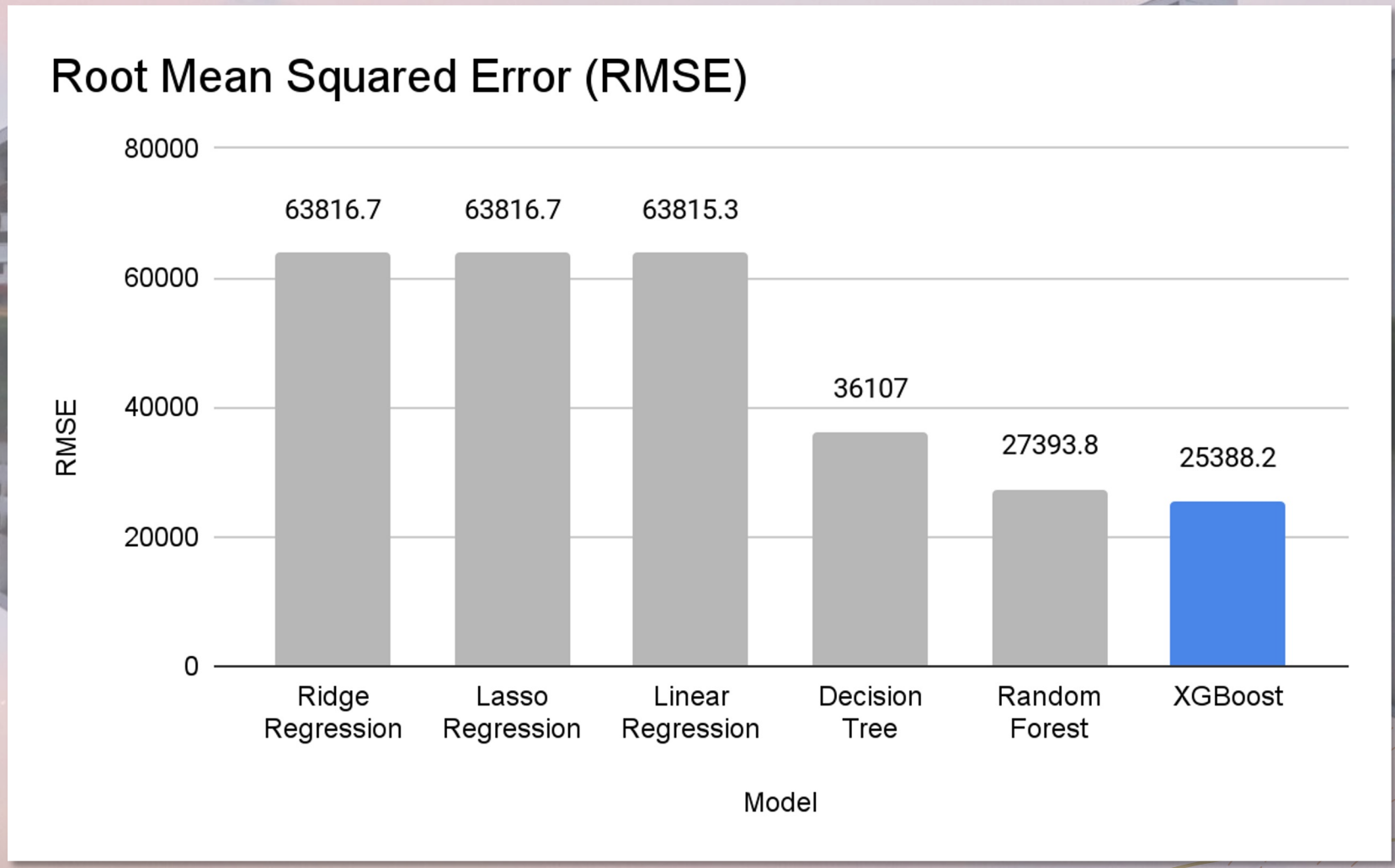
Use different processing methods depending on feature type

Feature Name	Feature Engineering	Feature Type	Processing	
			Step1	Step2
town	-	Categorical	One Hot Encoding	-
flat_model	-			-
year	Extracted from “month”		Label Encoding	MinMax Scaler
month				
quarter	Created from new “month” field			
category_flat_type	Binned to 4 categories from “flat_type”			
floor_area_sqm	-	Numerical	-	
avg_storey_range	Calculated from “storey_range“		-	
remaining_lease	Calculated from “month” and “lease_commence_date”		-	

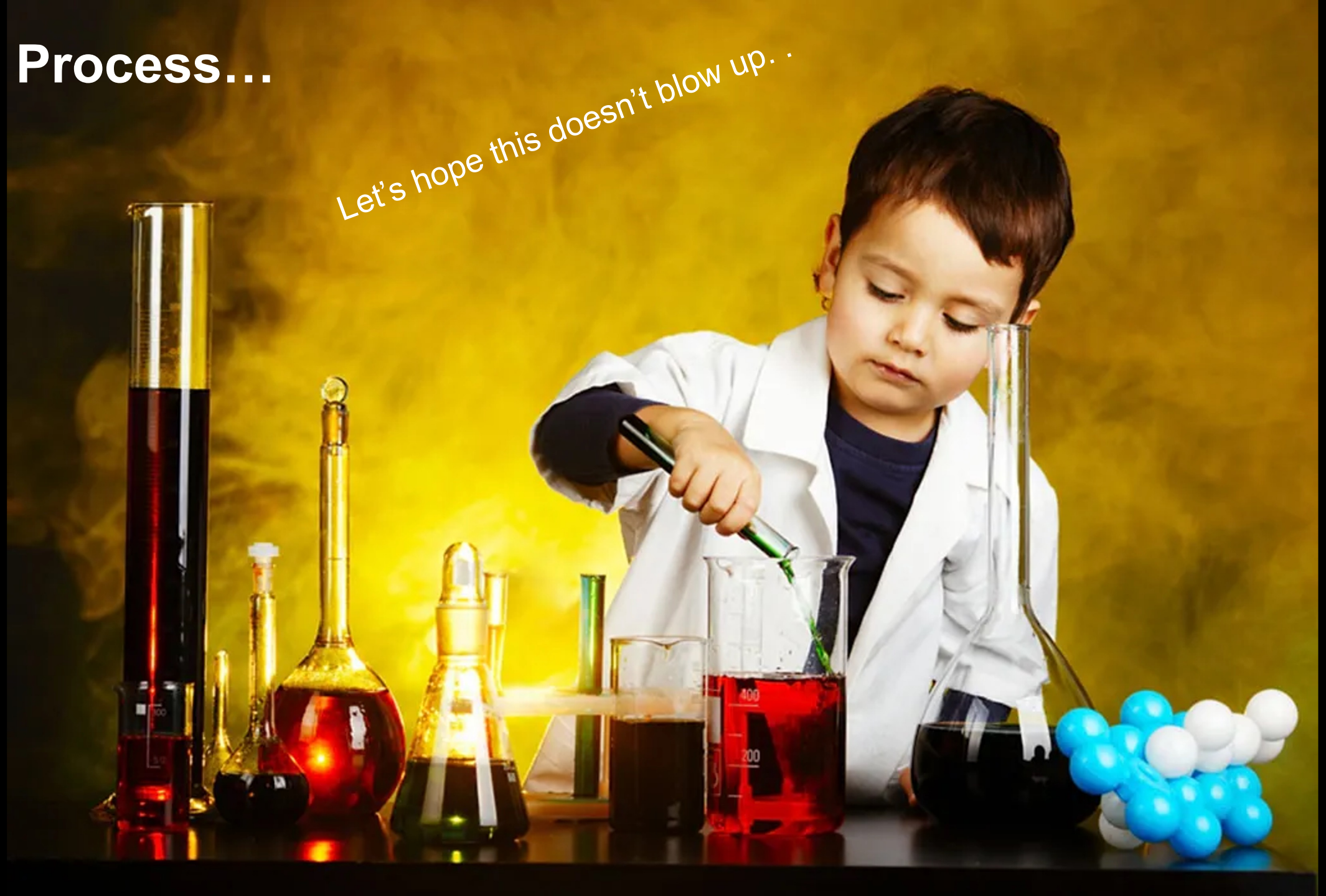
(Dropped "block” and "street_name”)

Model Selection

The best model for this problem is **XGBoost** among six models.

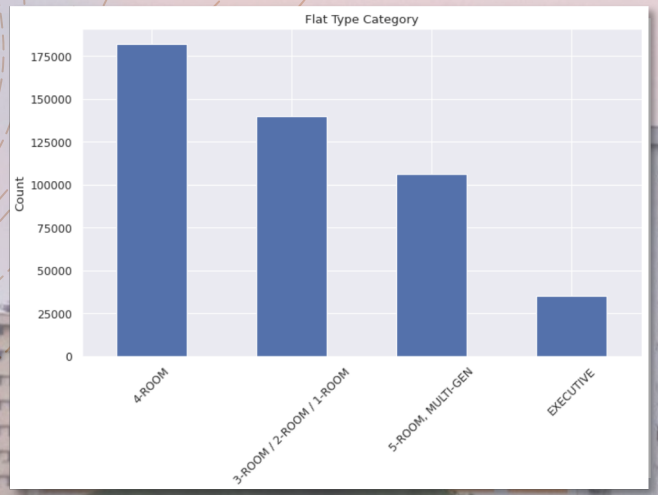


The Process…

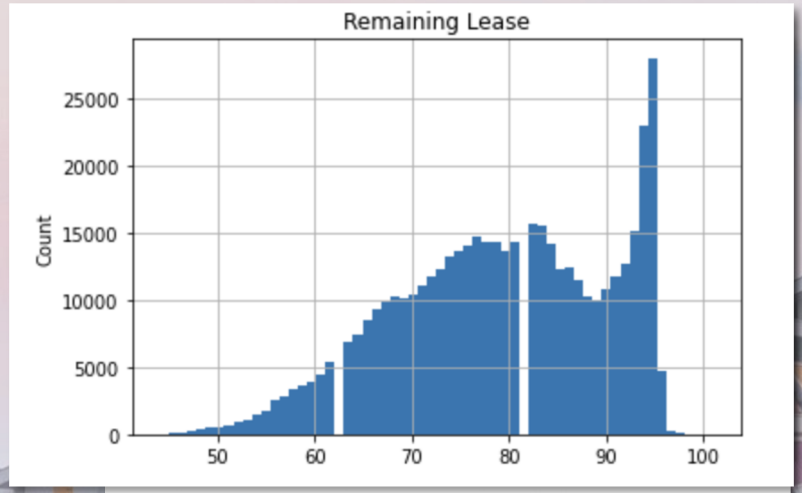


Experimentation

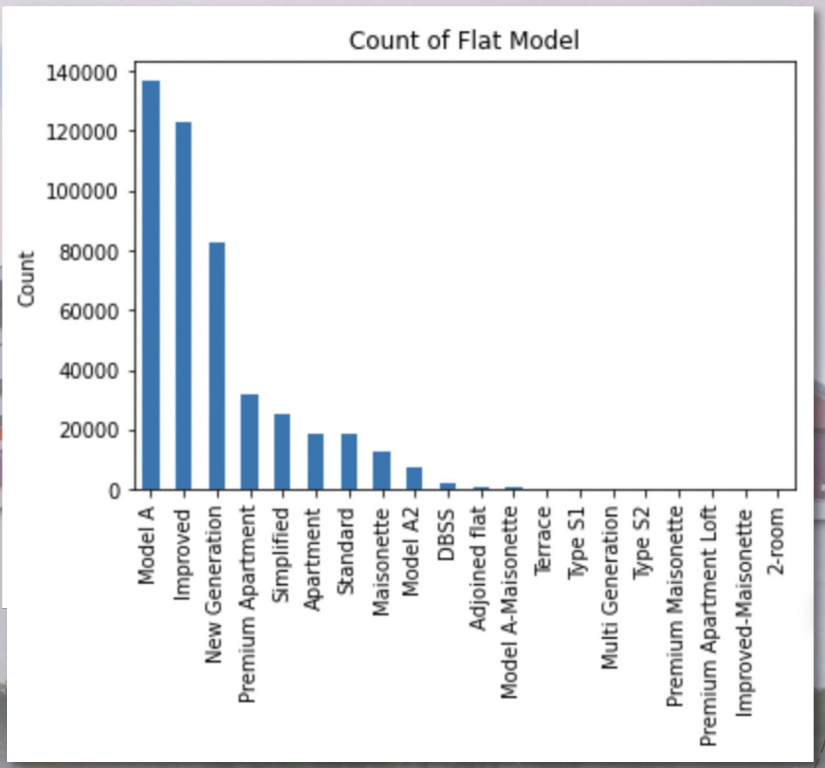
Flat Type



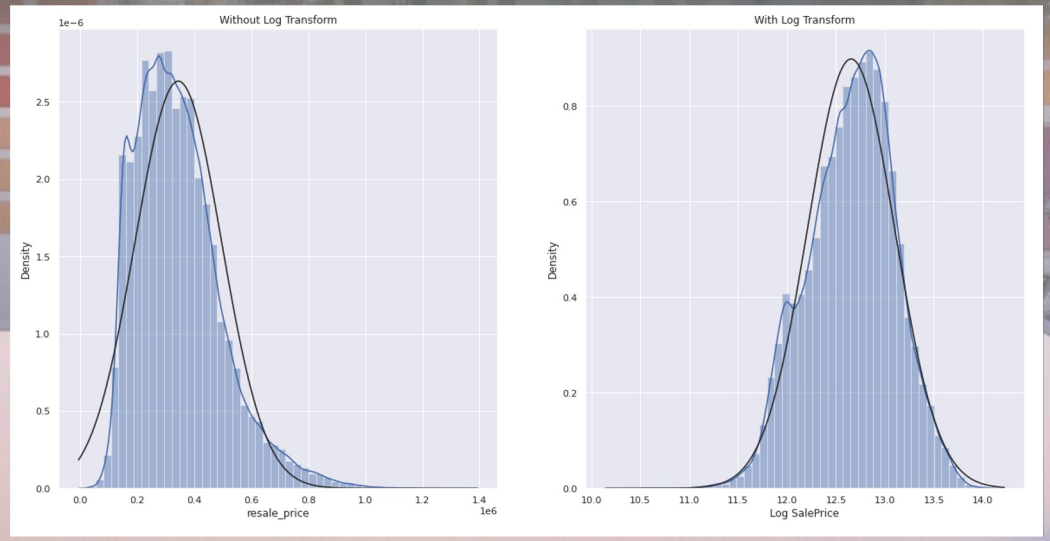
Remaining Lease



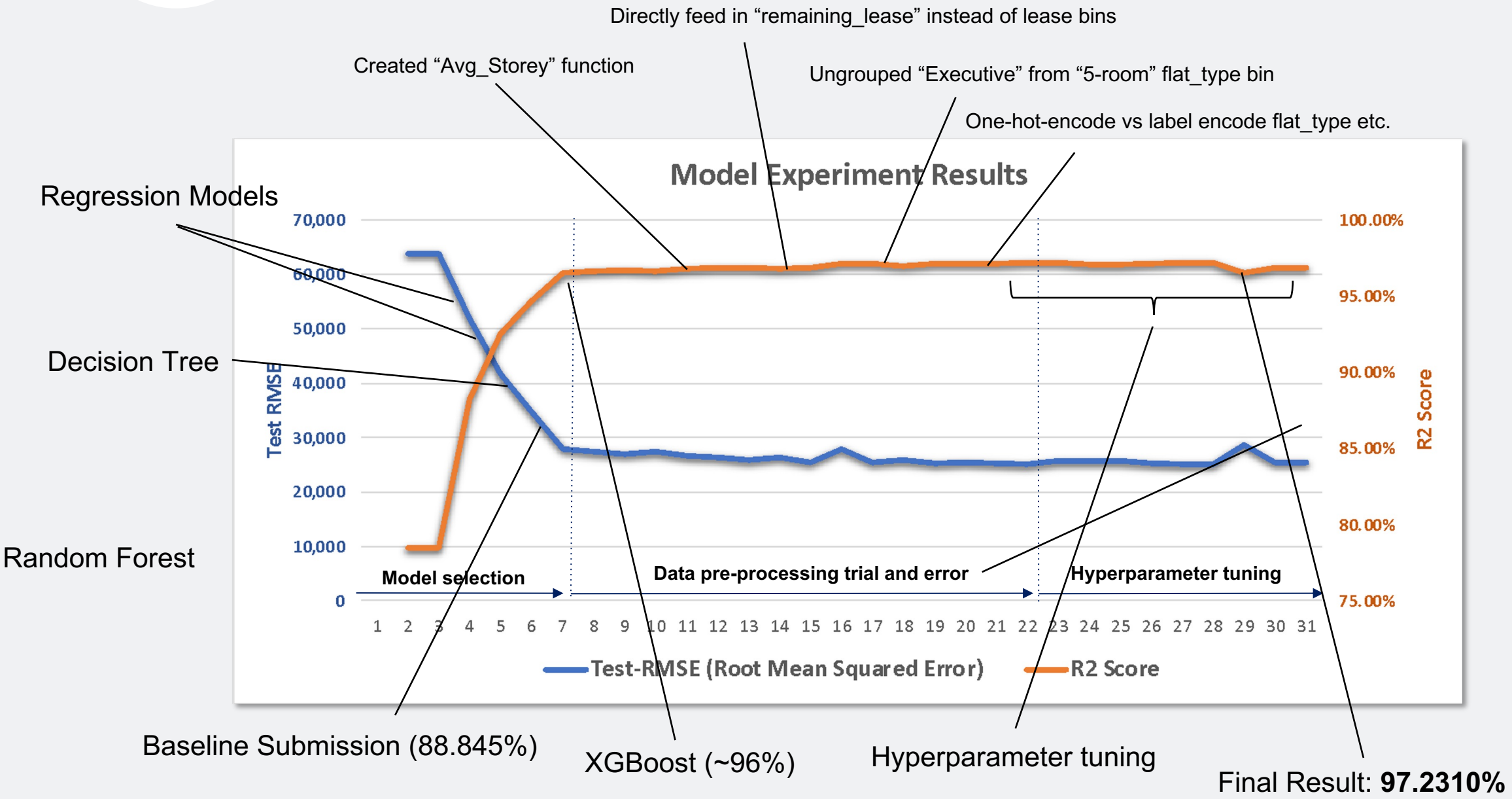
Flat Model



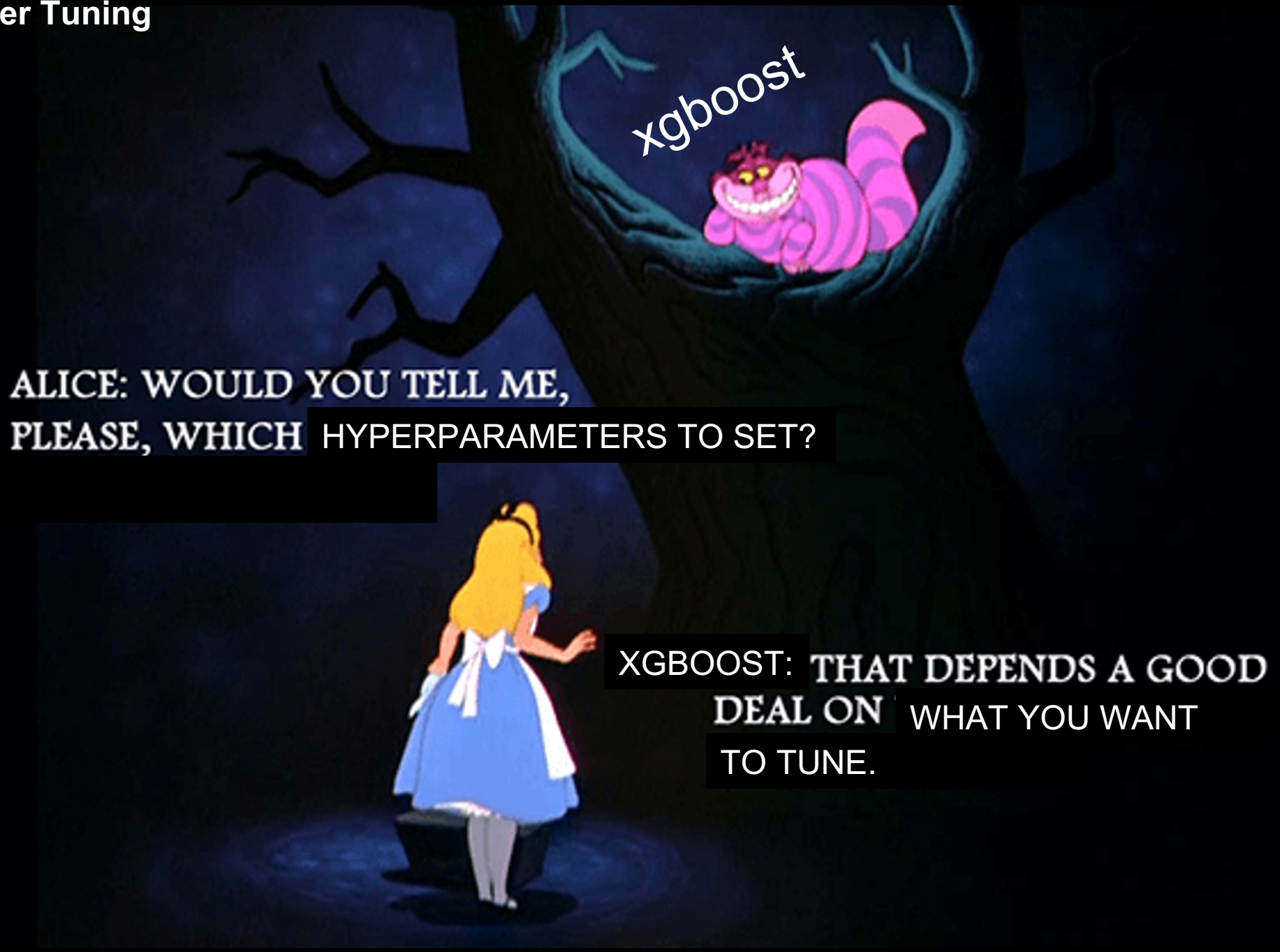
RMSE vs RMSLE



Timeline


Directly feed in “remaining_lease” instead of lease bins
Created “Avg_Storey” function
Ungrouped “Executive” from “5-room” flat_type bin
One-hot-encode vs label encode flat_type etc.
Regression Models
Decision Tree
Random Forest
Model Experiment Results
Test RMSE
R2 Score
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
100.00%
95.00%
90.00%
85.00%
80.00%
75.00%
Model selection
Data pre-processing trial and error
Hyperparameter tuning
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Test-RMSE (Root Mean Squared Error)
R2 Score
Baseline Submission (88.845%)
XGBoost (~96%)
Hyperparameter tuning
Final Result: 97.2310%

Hyperparameter Tuning



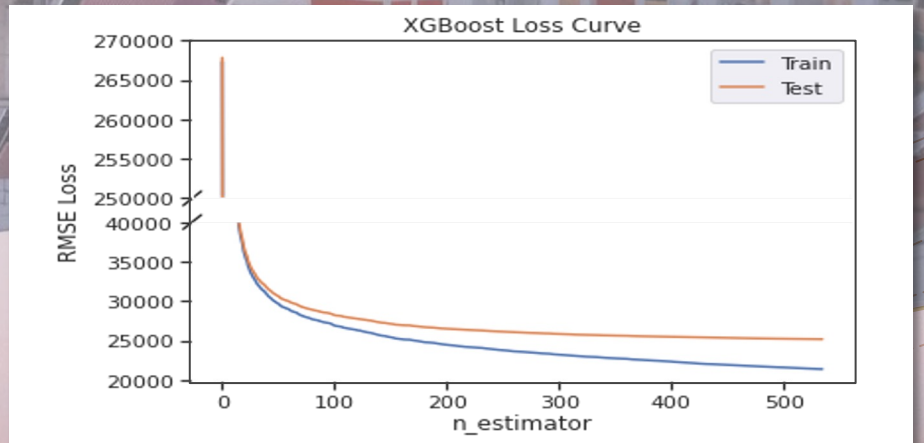
Hyperparameter Tuning

Try **GridSearchCV** and **RandomizedSearchCV** for hyperparameter tuning

Parameter Name	Default	Range of tuning	Best result
max_depth	6	5, 6, 7, 8, 9, 10	**8**
min_child_weight	1	1, 3	**1**
eta (learning rate)	0.3	0.3, 0.1	**0.3**
colsample_bytree	1.0	0.5, 1.0	**1.0**
subsample	1.0	0.5, 1.0	**1.0**
num_boost_round	10	800, 1000, 1200	**1000**

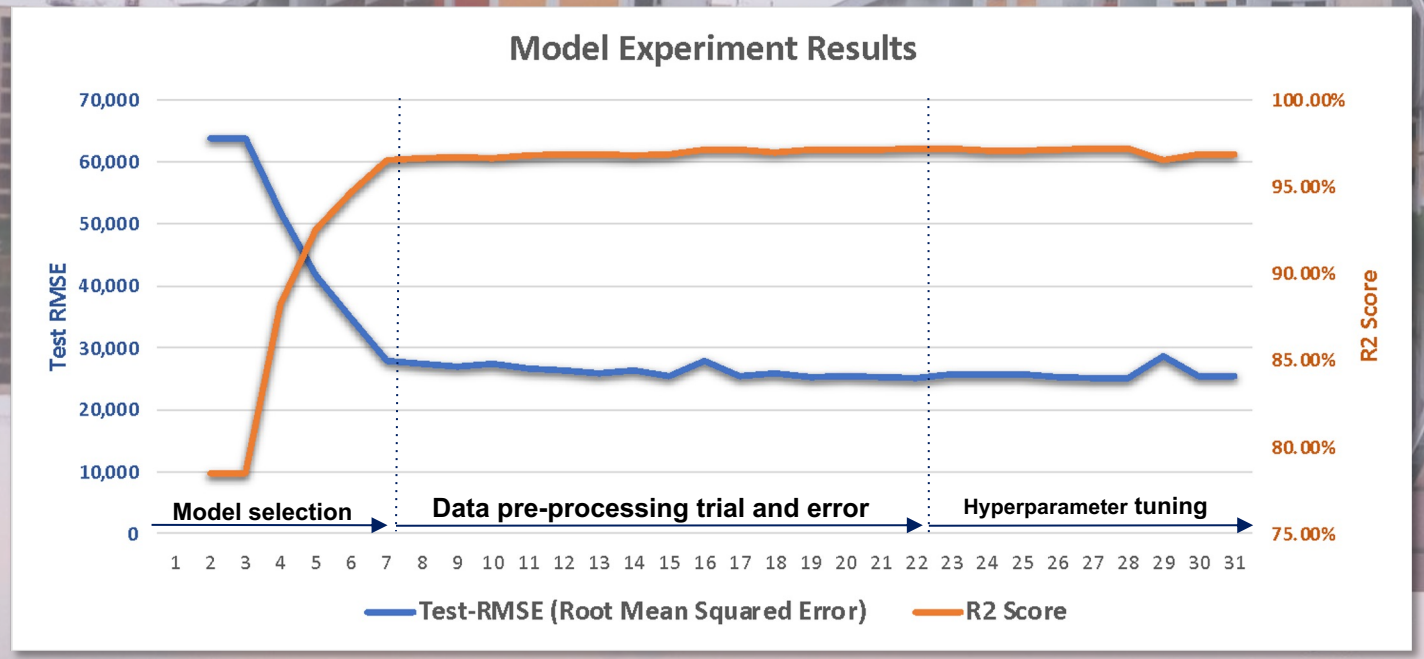
Other parameters (did not do tuning)

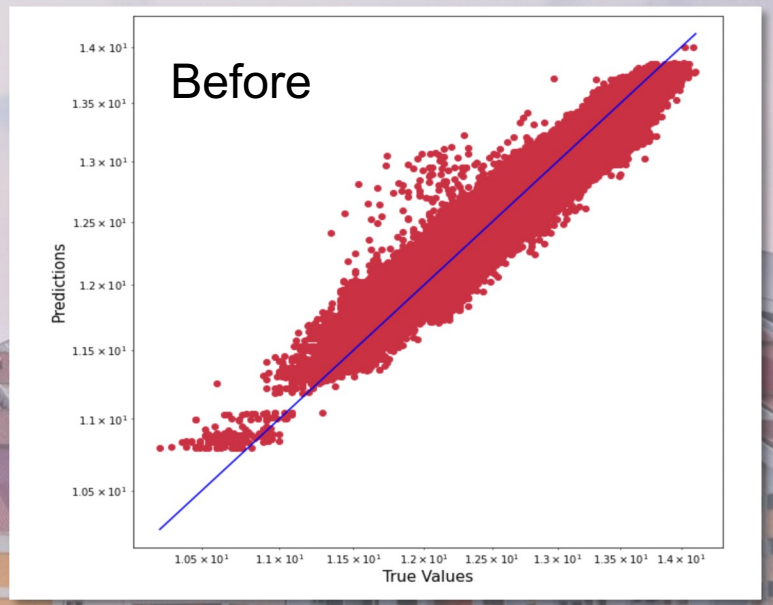
booster	gbtree
objective	reg: squarederror
eval_metric	rmse

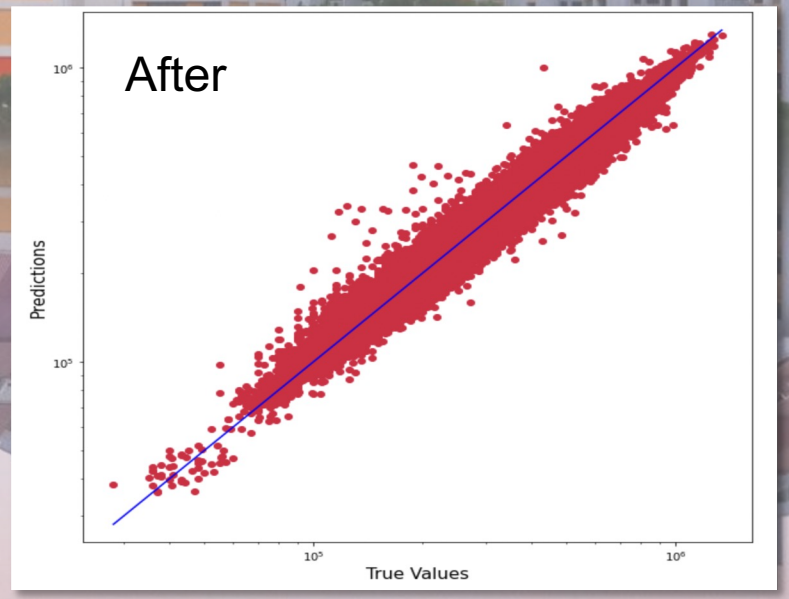


Summary

	Baseline	Final Result
Model	Random Forest	**XGBoost**
R2 Score	88.845%	**97.2310%**
Test-RMSE	63,816	**24,932.0**

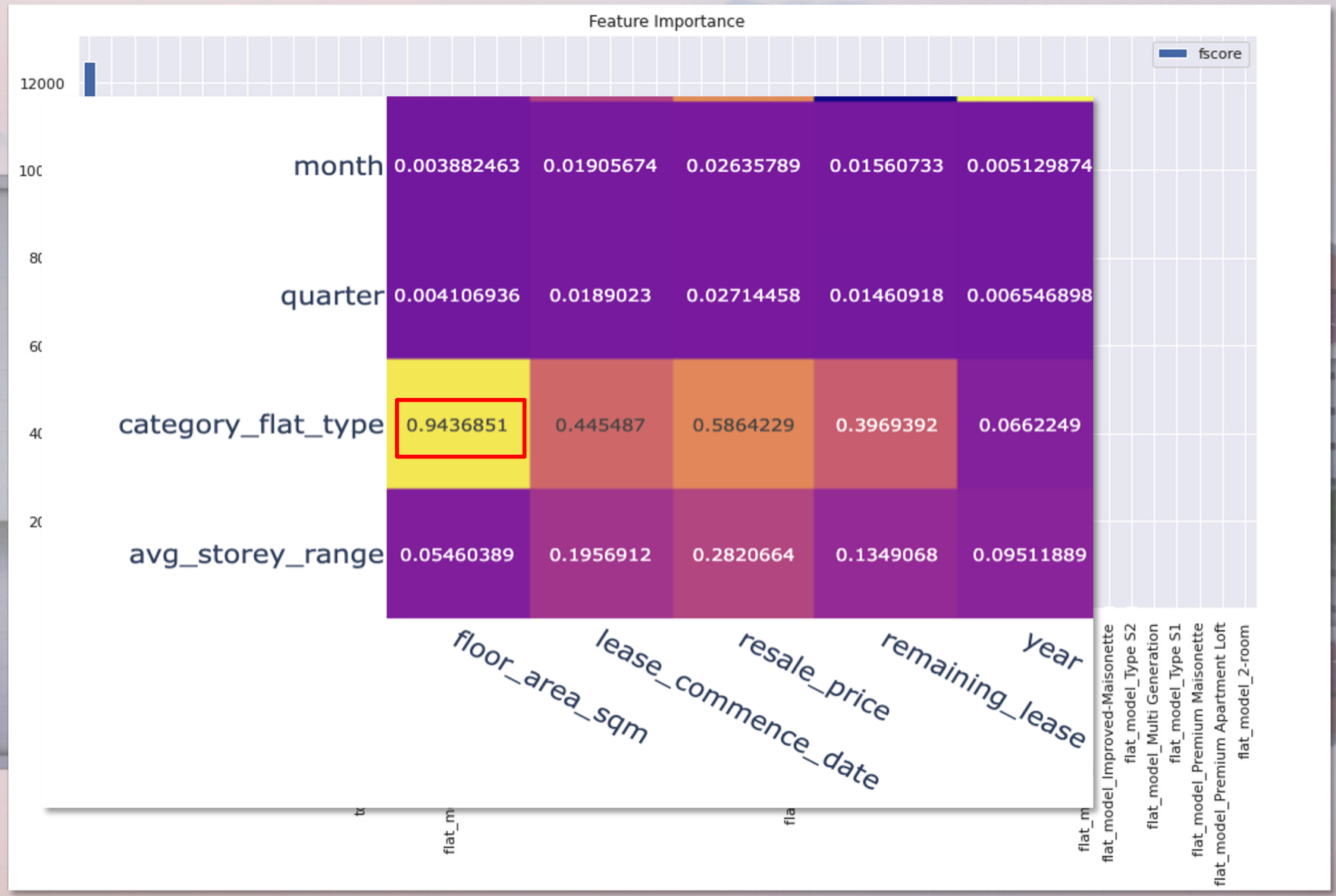






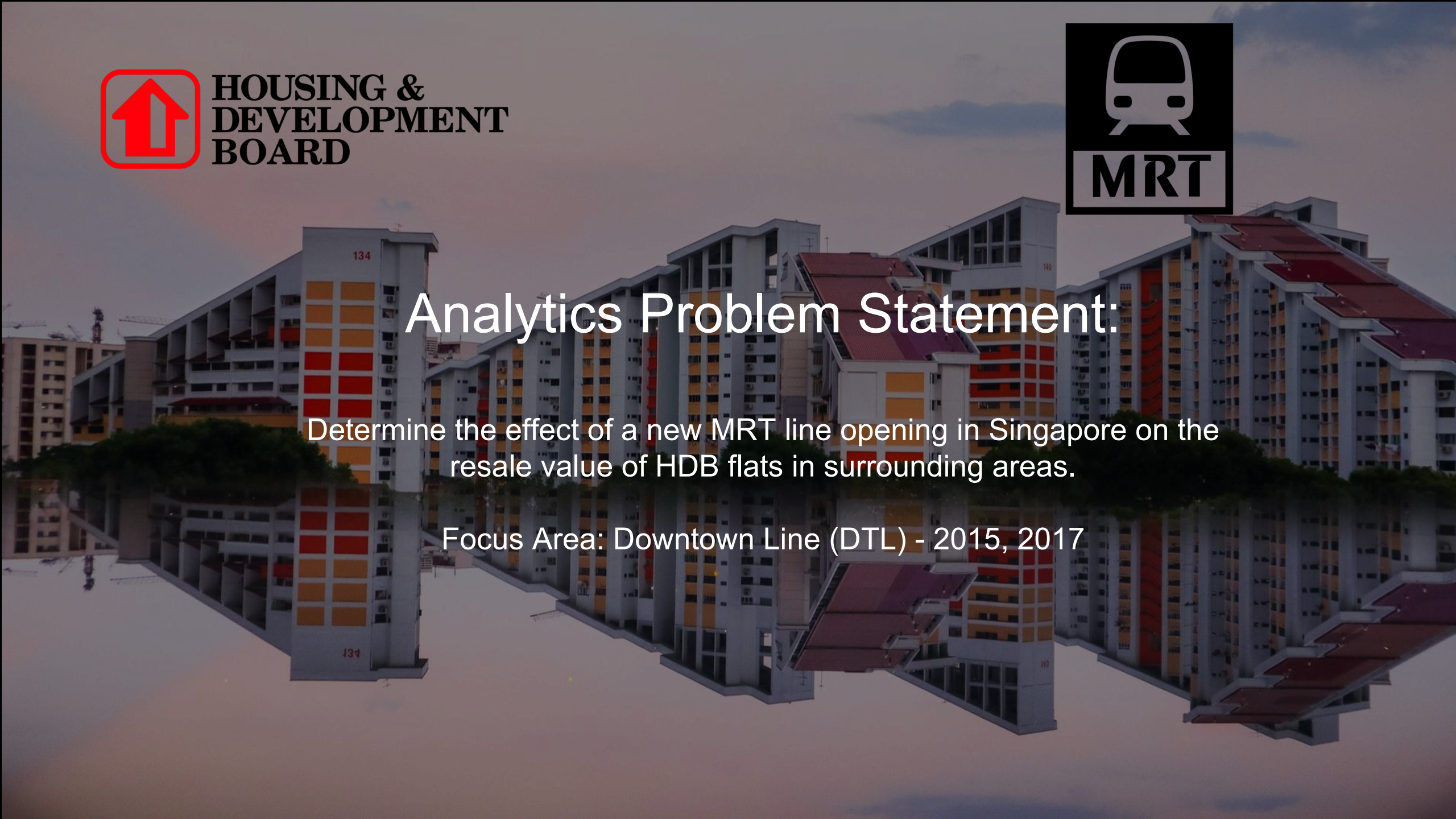
Feature Importance

	fscore
floor_area_sqm	12475
remaining_lease	11323
year	9057
month	6981
avg_storey_range	6073
town_BUKIT MERAH	645
flat_model_Improved	635
town_JURONG WEST	547
town_BEDOK	510
town_TOA PAYOH	498
flat_model_Model A	472
category_flat_type	469
town_KALLANG/WHAMPOA	447
town_HOUGANG	434
town_QUEENSTOWN	431
town_GEYLANG	382
flat_model_Premium Apartment	349
town_JURONG EAST	336
town_SERANGOON	321
town_BUKIT BATOK	309
town_CLEMENTI	305
town_WOODLANDS	295



Lessons Learned

- Data modeling is an Empirical Process
- No "one size fits all" solution
- You need to know what you're doing when doing Gridsearch (and a fast GPU)



Analytics Problem Statement:

Determine the effect of a new MRT line opening in Singapore on the resale value of HDB flats in surrounding areas.

Focus Area: Downtown Line (DTL) - 2015, 2017

Analytics Problem Statement

- Determine the effect of a new MRT line opening in Singapore on the resale value of HDB flats in surrounding areas
- Focus Area: Downtown Line (DTL) - 2015, 2017

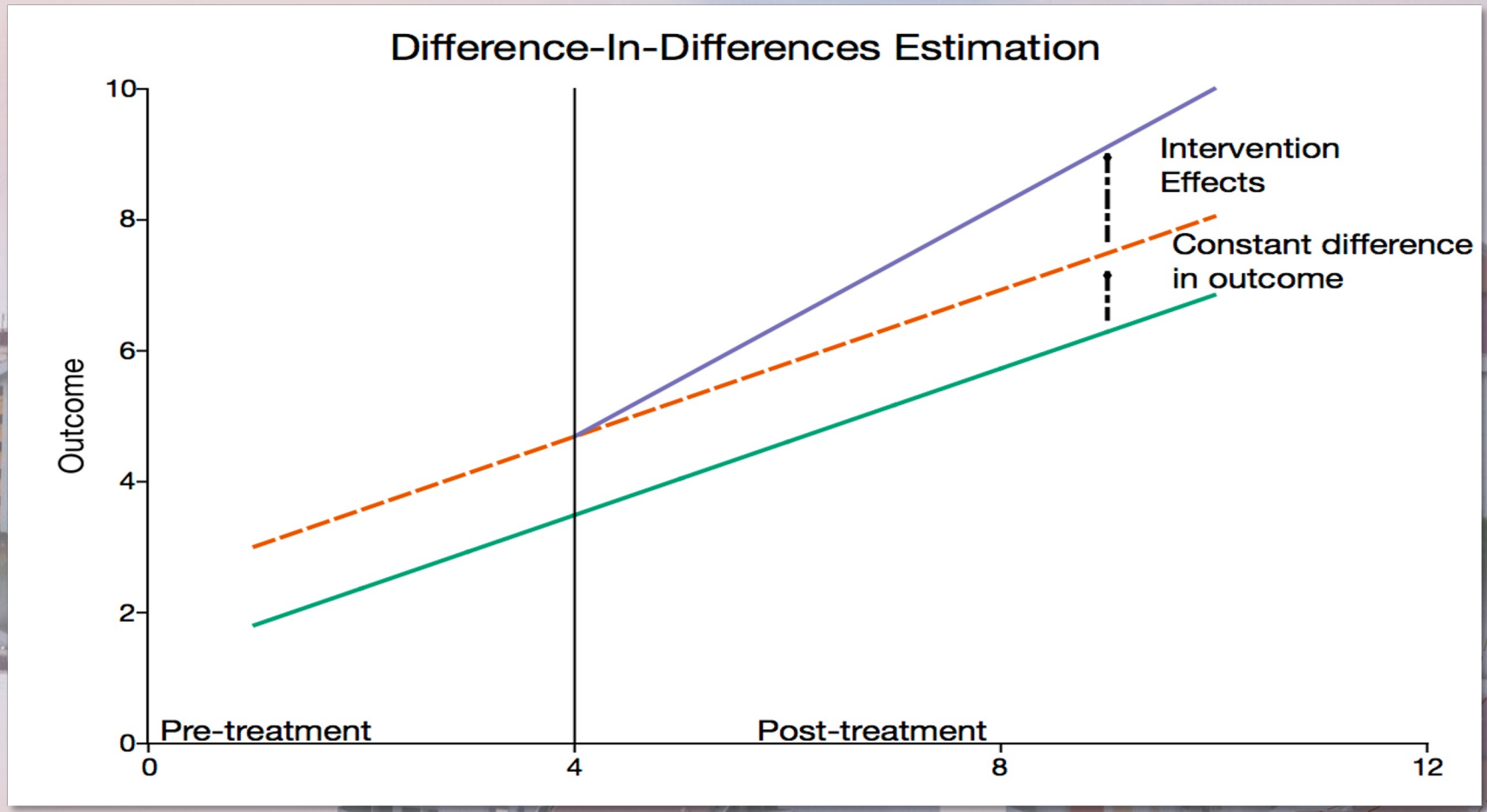
Data

Data Required	Source of Data	Details
Coordinates of each HDB Unit	OneMap API	Repeatedly call the OneMap API with each block's address to get the coordinates and postal code as a JSON object.
Coordinates of all MRT Stations	Kaggle (Link)	Readily available data, last updated as of Oct 2021.
Opening Dates of all MRT Stations	Wikipedia	Scrapped the opening dates from Wikipedia.

Methodology: Difference-in-Difference (DiD)

Akin to A/B Testing

Control for changes before and after treatment



	Before Opening of DTL	**After Opening of DTL**
Treatment (<=500m of DTL)	$Group_{Treatment,\ Before}$	$Group_{Treatment,\ After}$
Control (>500m from DTL)	$Group_{Control,\ Before}$	$Group_{Control,\ After}$

Methodology: Difference-in-Difference (DiD)

Regression Equation:

$$PSM = B_0 + B_1 TimePeriod + B_2 Treatment + B_3 TimePeriod * Treatment + B_4 OtherVariables$$

Other Variables: Remaining Lease, Flat Type, etc.

B_1: 1 if after treatment period, 0 if before

B_2: 1 if within 500m of DTL, 0 if not

B_3: Effect Variable (if B_1 and B_2 are both 1, then B_3 = 1)

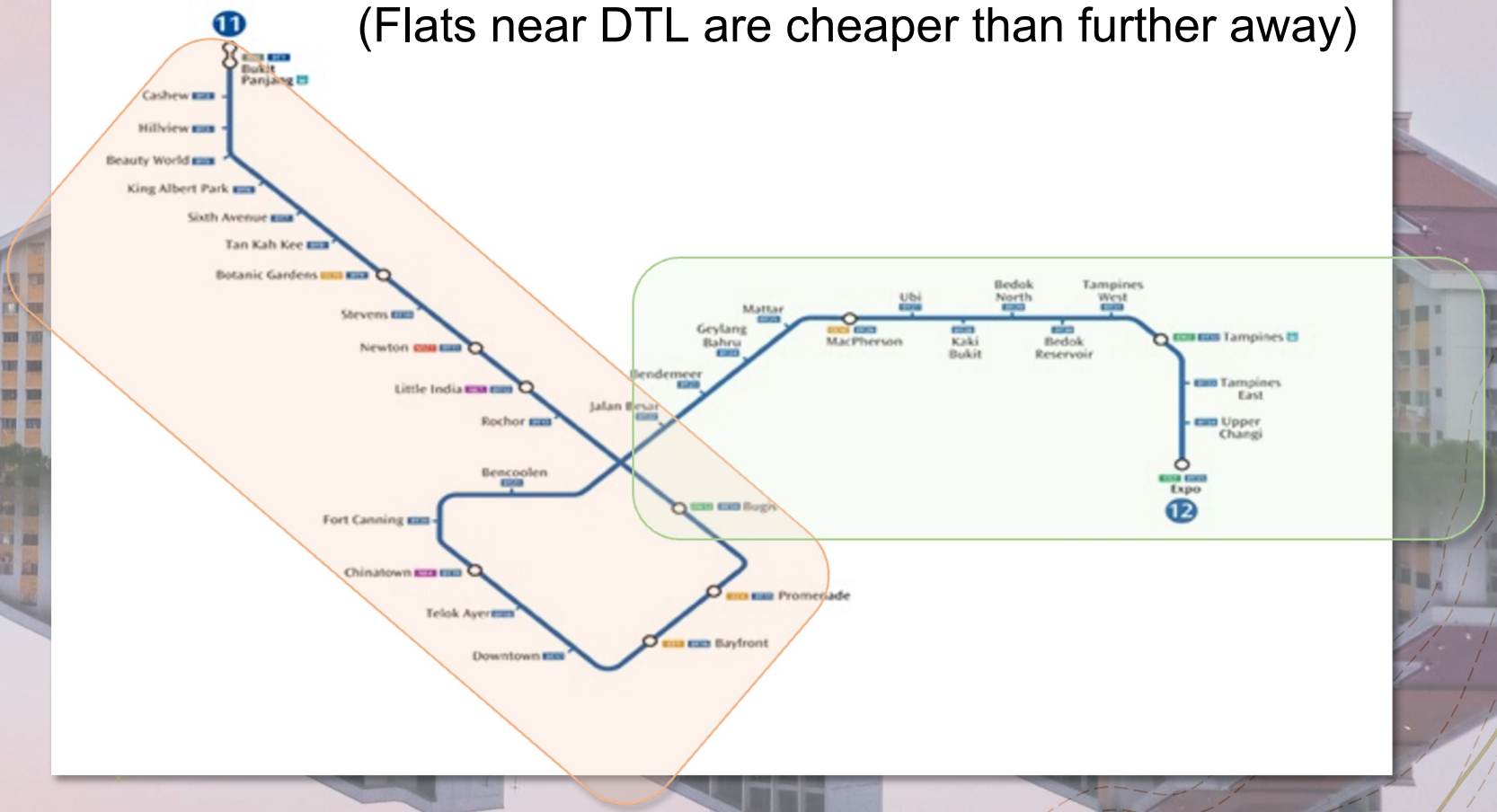
Results

	Model 1	Model 2
Adjusted R Squared	0.7613	0.7646
Intercept	2654.4998*** (118.9887)	6605.6962*** (180.2534)
Remaining Lease	54.6900*** (0.3213)	-54.7927*** (3.7851)
Remaining Lease_Squared		0.7142*** (0.0246)
Treatment	-228.2454*** (15.7654)	-214.7709*** (15.6637)
TimePeriod	-6.1537 (4.6541)	-15.4382*** (4.6331)
Treatment x TimePeriod	90.3185*** (20.6795)	**79.2672***** **(20.5405)**

Interpretation

	Model 2
Treatment	-214.7709*** (15.6637)
Time Period	-15.4382*** (4.6331)
Treatment x TimePeriod	**79.2672***** **(20.5405)**

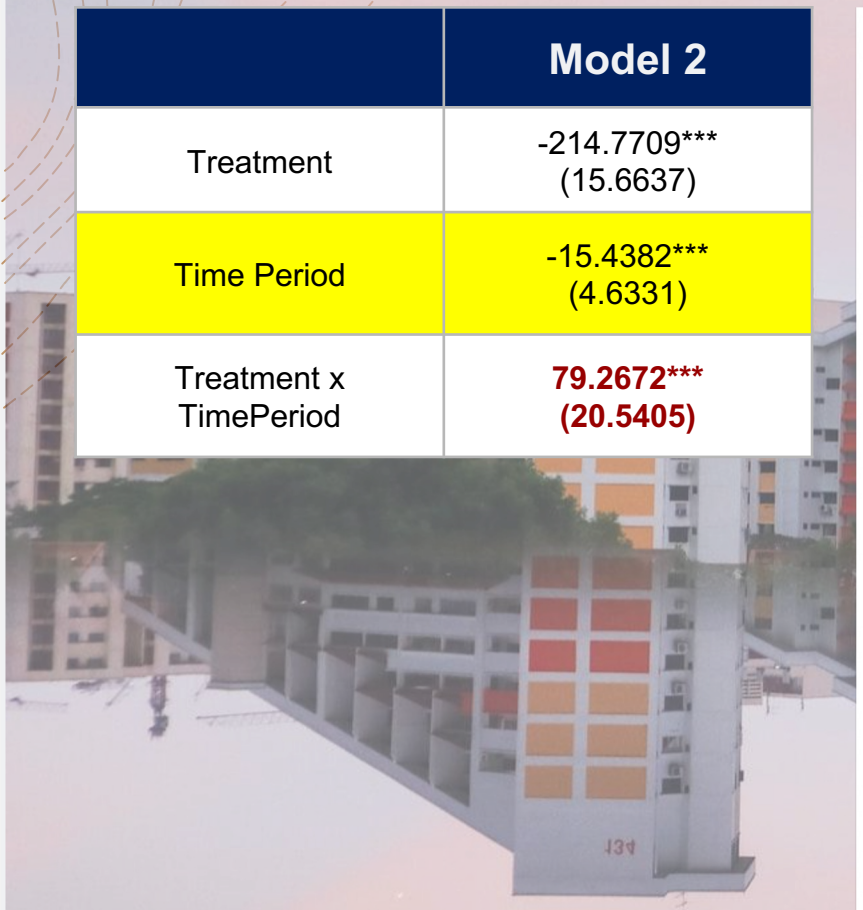
(Flats near DTL are cheaper than further away)

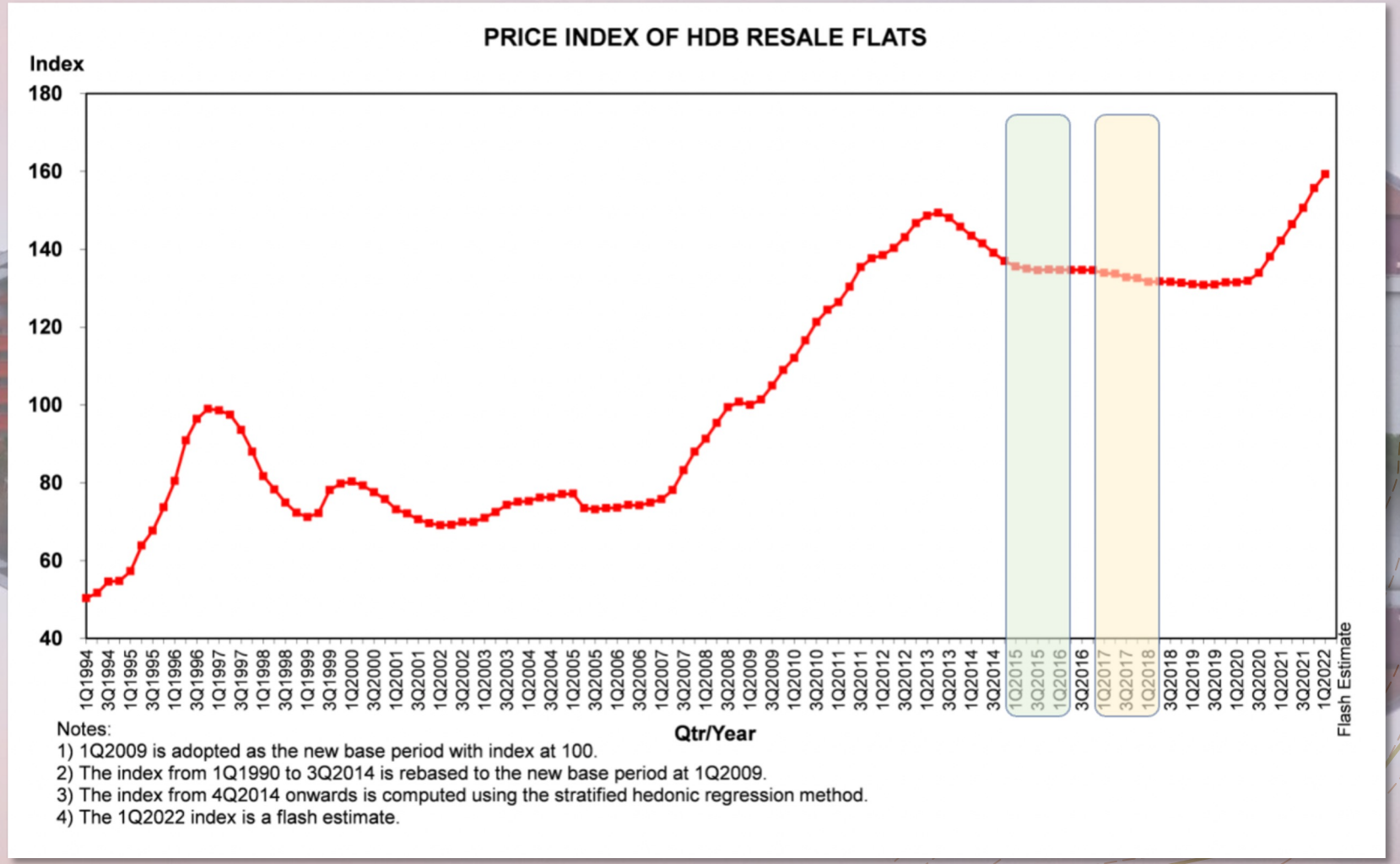


Interpretation

PSM decreased generally over time

	Model 2
Treatment	-214.7709*** (15.6637)
Time Period	-15.4382*** (4.6331)
Treatment x TimePeriod	**79.2672*** (20.5405)**

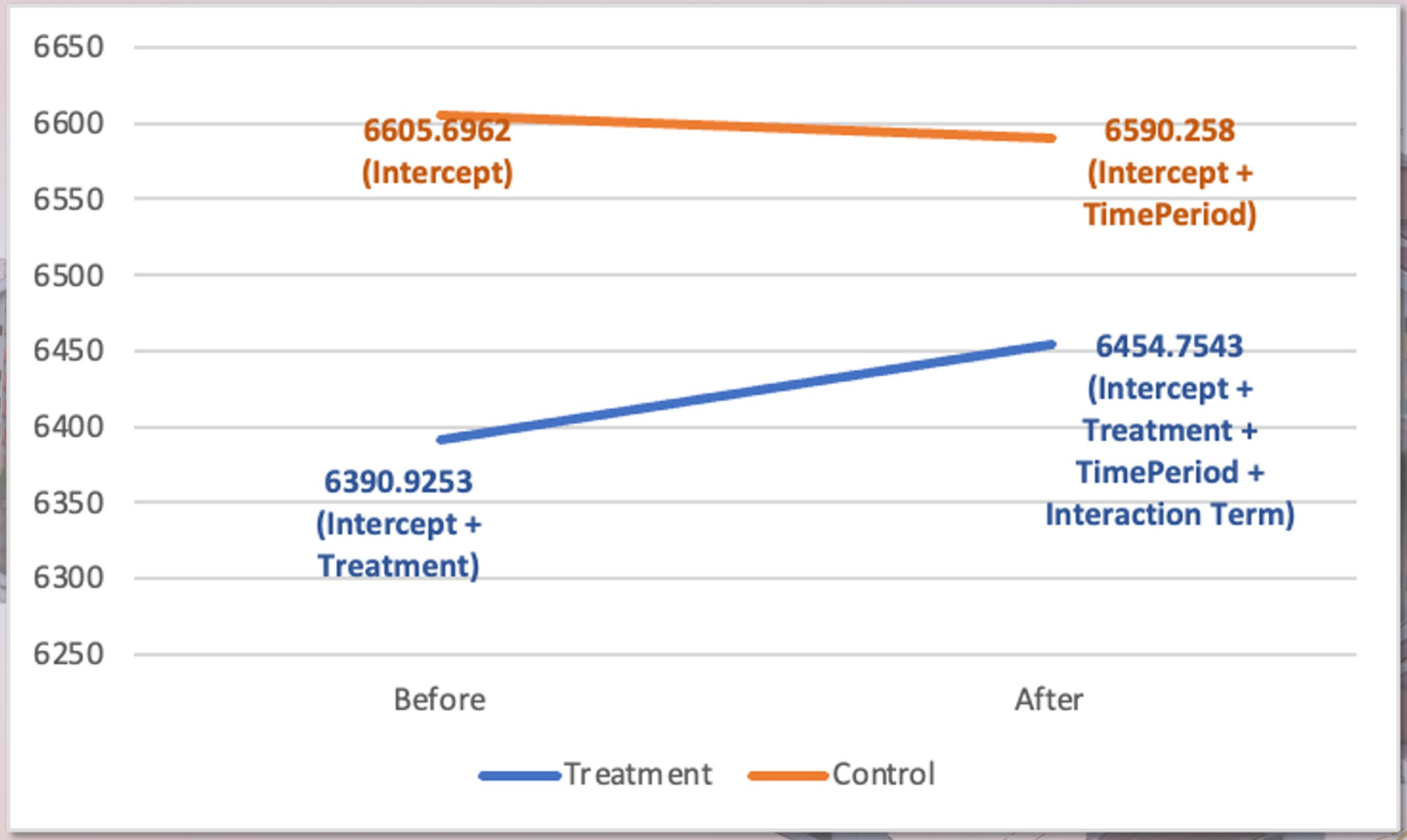




Interpretation

Net effect = $79.27

	Model 2
Treatment	-214.7709*** (15.6637)
Time Period	-15.4382*** (4.6331)
Treatment x TimePeriod	**79.2672***** **(20.5405)**



Robustness Checks

Relations are the same (signs), magnitude differs

	500m (Original)	400m	1,000m
Adjusted R Squared	0.7646	0.7640	0.7670
Intercept	6605.6962*** (180.2534)	6640.5661*** (180.470)	6515.7445*** (179.369)
Treatment	-214.7709*** (15.6637)	-148.4850*** (18.650)	-320.7570*** (12.424)
TimePeriod	-15.4382*** (4.6331)	-14.1193*** (4.601)	-15.9020*** (4.761)
Treatment x TimePeriod	79.2672*** (20.5405)	77.9901*** (24.916)	45.0120*** (14.213)

Conclusion

- Increase in PSM of flats near the DTL after the DTL opened, results are robust after adjusting assumptions
- Accounted for possible price changes island-wide through the use of DiD regression

Further Considerations

- Include more features, such as distance to other amenities, such as schools, shopping malls
- Analyze the “announcement effect”

HOUSING & DEVELOPMENT BOARD

MRT

Thank you for listening to our presentation!