SC1015 Mini Project

Resale HDB Flat Prices Dataset

JOEY LIM HUI MING (U2120303G), LIM SHI TONG (U2120704H), TOH JING SHENG (U2120444J)

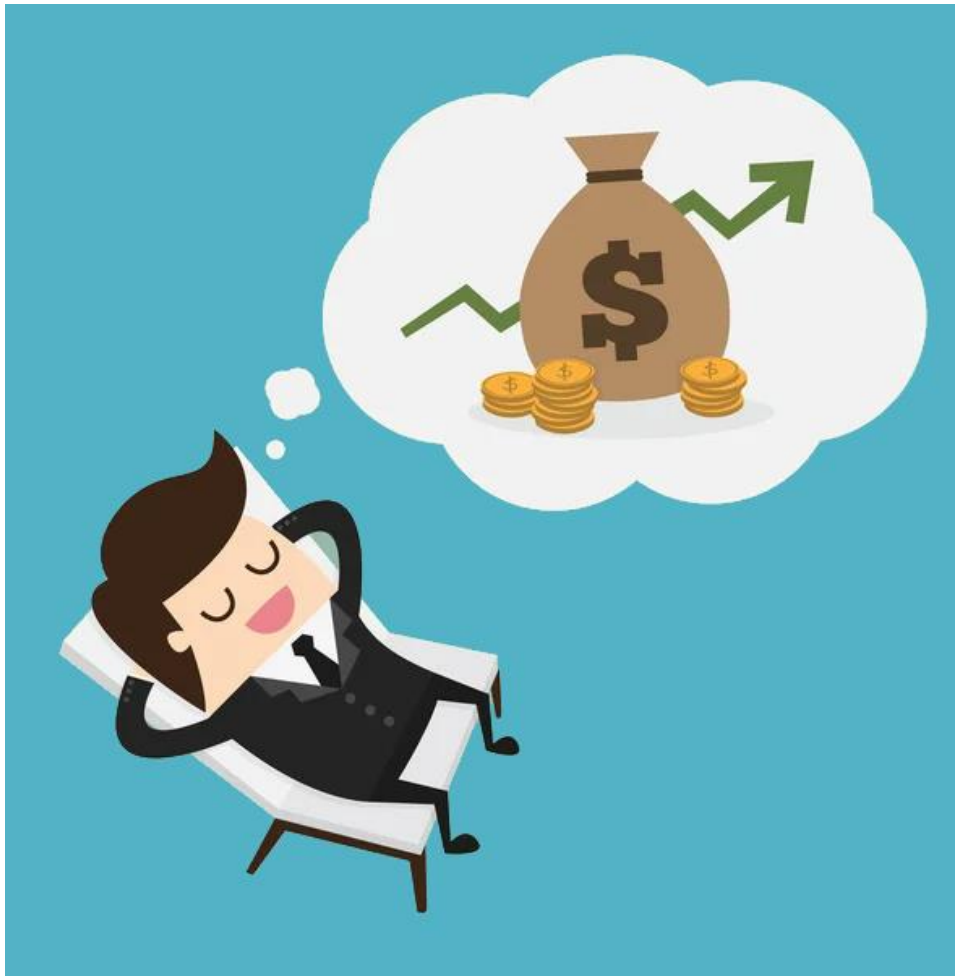
Dataset

Prices of Resale flat from Jan 1990 to Mar 2022

Problem Definition

- Derive the drivers for each generations (1990s, 2000s, 2010s, 2020s)
- To observe if recession affect resale prices for specific flat types and models

Motivation



Future financial planning

Data Cleaning

- Consulted Professional
- Individual Town to Region
- Combined Storey Range
- Sorted Years into Decades
- Calculation of Lease
- Cleaning of Flat Type and Flat Models

Dataset

Before

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price	remaining_lease
0	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	10 TO 12	31.0	IMPROVED	1977	9000.0	NaN
1	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	04 TO 06	31.0	IMPROVED	1977	6000.0	NaN
2	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	10 TO 12	31.0	IMPROVED	1977	8000.0	NaN
3	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	07 TO 09	31.0	IMPROVED	1977	6000.0	NaN
4	1990-01	ANG MO KIO	3 ROOM	216	ANG MO KIO AVE 1	04 TO 06	73.0	NEW GENERATION	1976	47200.0	NaN

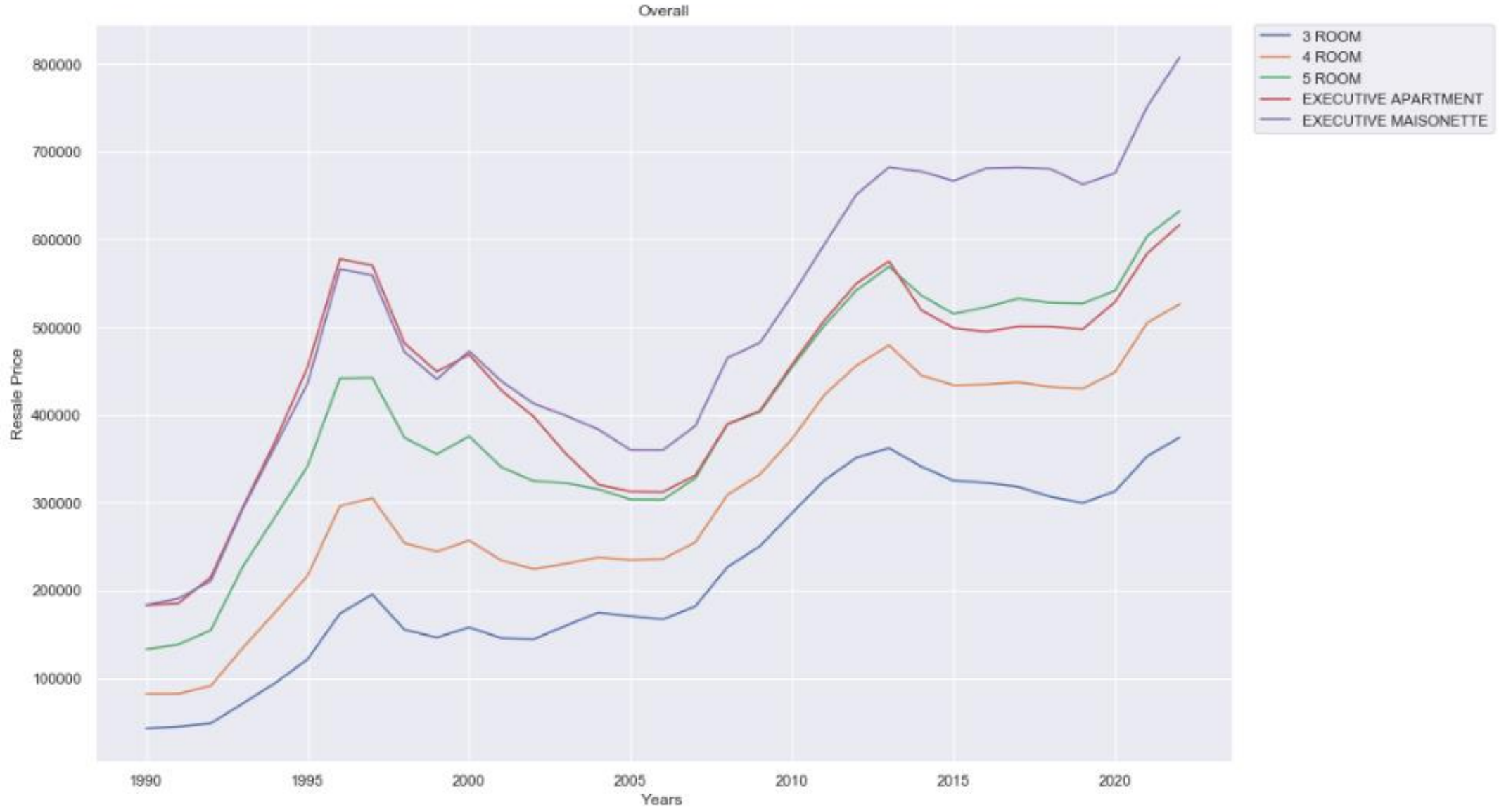
After

	flat_type	storey_range	floor_area_sqm	flat_model	resale_price	year	lease_left	region
0	1 ROOM	01 TO 15	31.0	IMPROVED	9000.0	1990	86	NORTH-EAST REGION
1	1 ROOM	01 TO 15	31.0	IMPROVED	6000.0	1990	86	NORTH-EAST REGION
2	1 ROOM	01 TO 15	31.0	IMPROVED	8000.0	1990	86	NORTH-EAST REGION
3	1 ROOM	01 TO 15	31.0	IMPROVED	6000.0	1990	86	NORTH-EAST REGION
4	3 ROOM	01 TO 15	73.0	NEW GENERATION	47200.0	1990	85	NORTH-EAST REGION

Exploratory Data Analysis

- Resale Price of Flat Type over the Years
- Distribution Over Flat Type vs Region
- Map Visualization of Average Resale Price per Region
- Distribution Over Resale Price vs Region
- Distribution Over Resale Price vs Flat Model
- Correlation of Floor Area Sqm vs Lease Left vs Resale Price

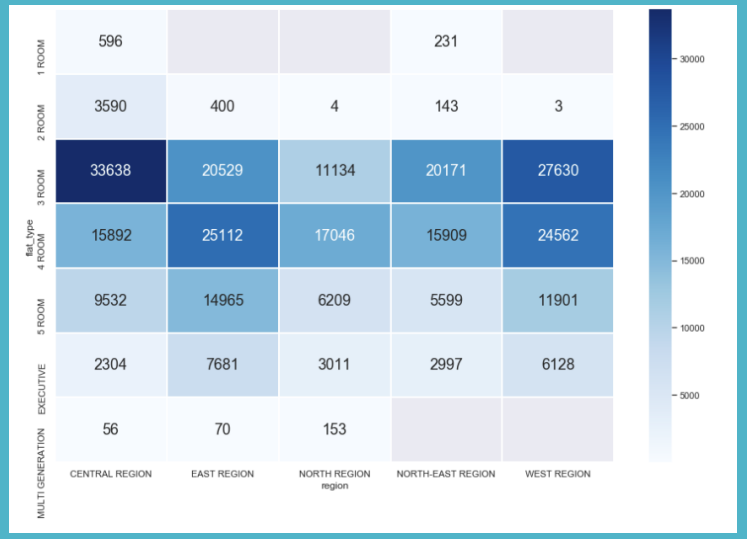
Resale Price of Flat Type over the Years



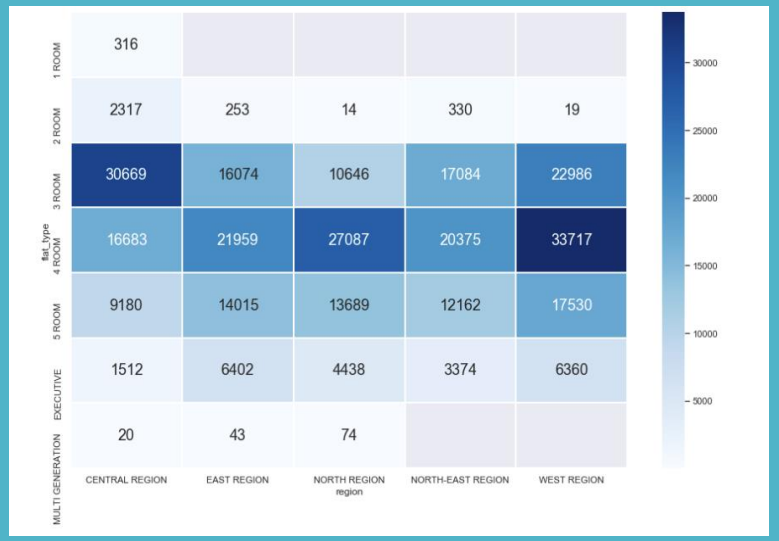
Across the Decades

Distribution Over Flat Type vs Region

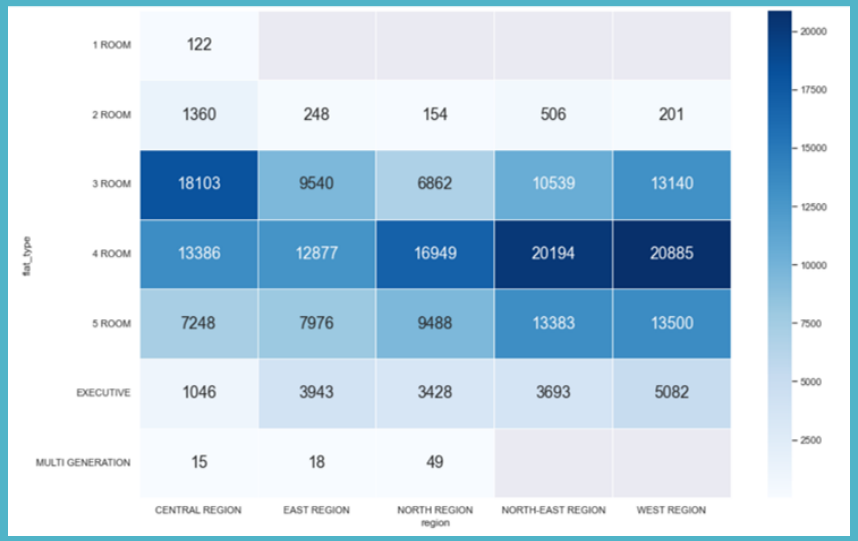
1990 to 1999



2000 to 2009

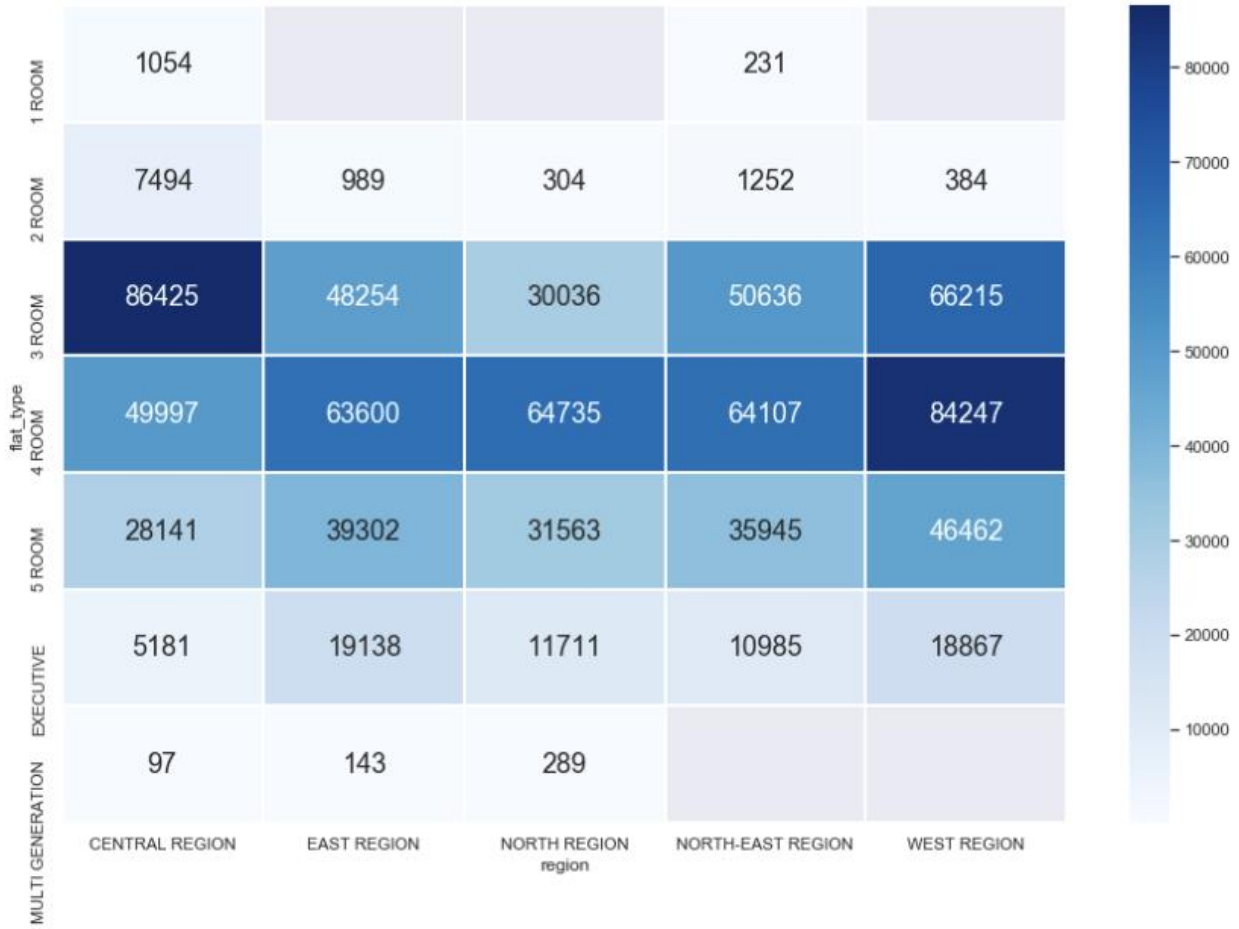


2010 to 2019



Overall

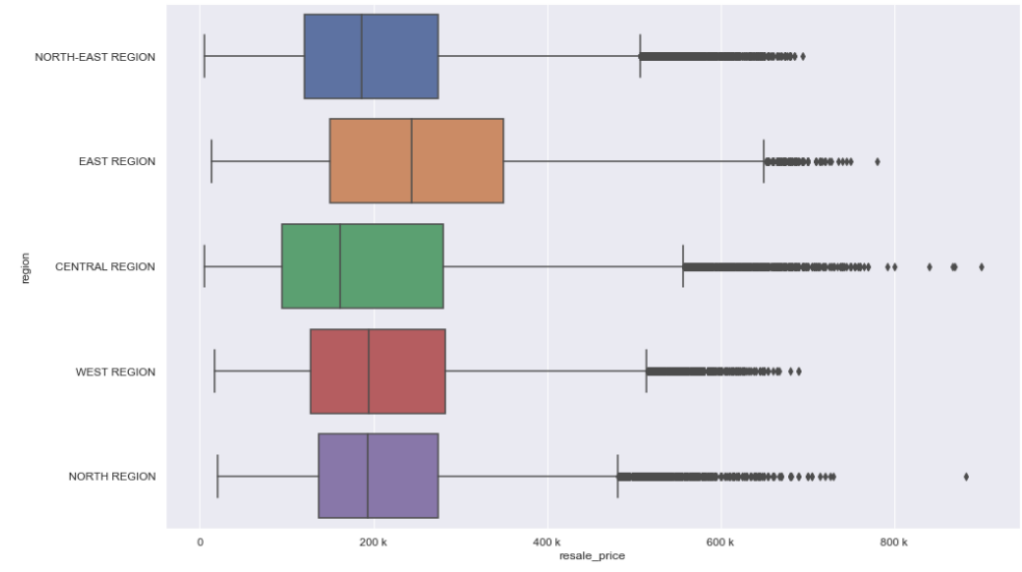
Distribution Over Flat Type vs Region



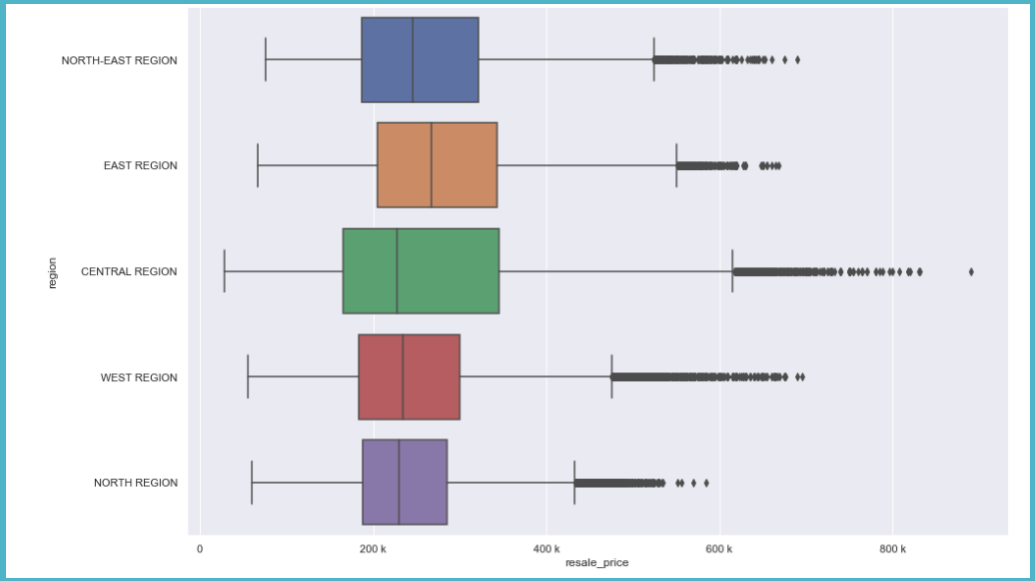
Across the Decades

Distribution Over Resale Price vs Region

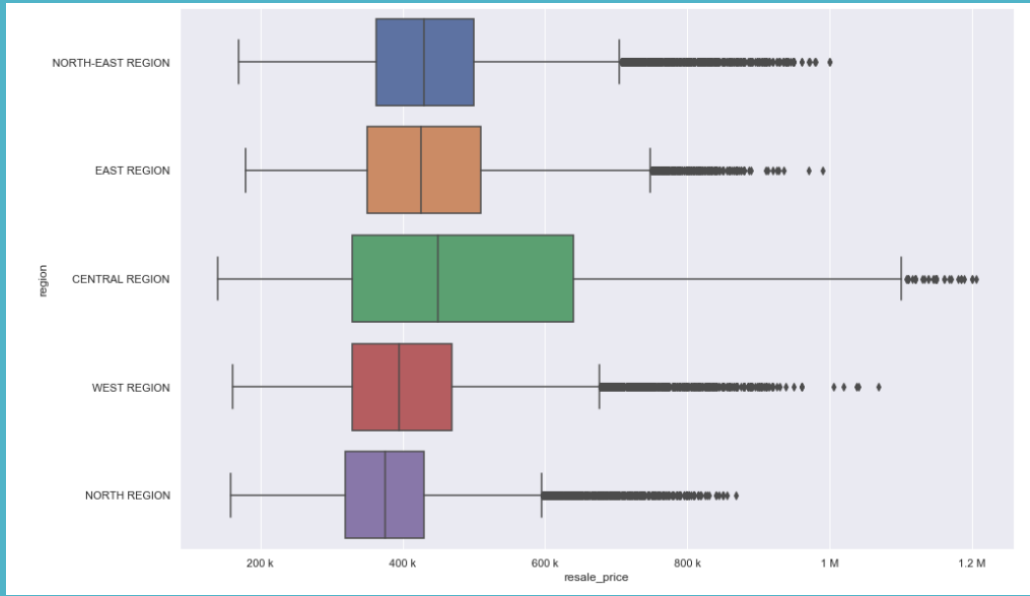
1990 to 1999



2000 to 2009

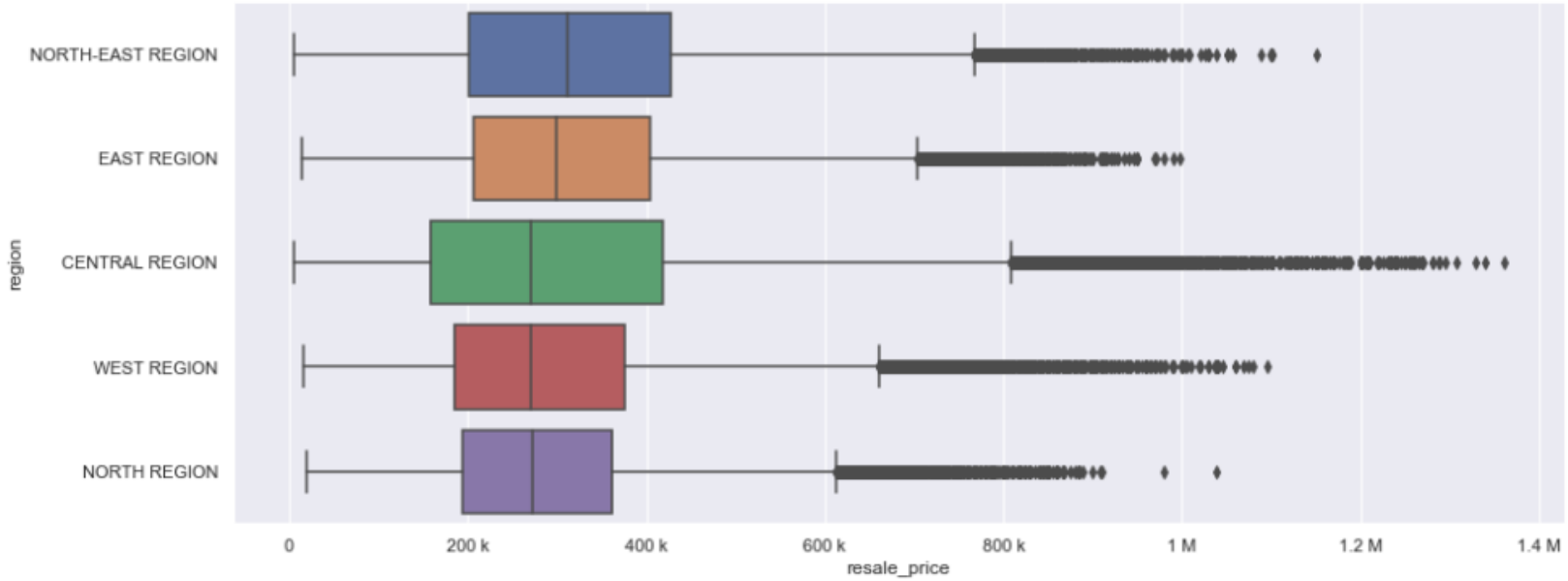


2010 to 2019



Overall

Distribution Over Resale Price vs Region



Across the Decades

Map Visualization of
Average Resale Price per Region

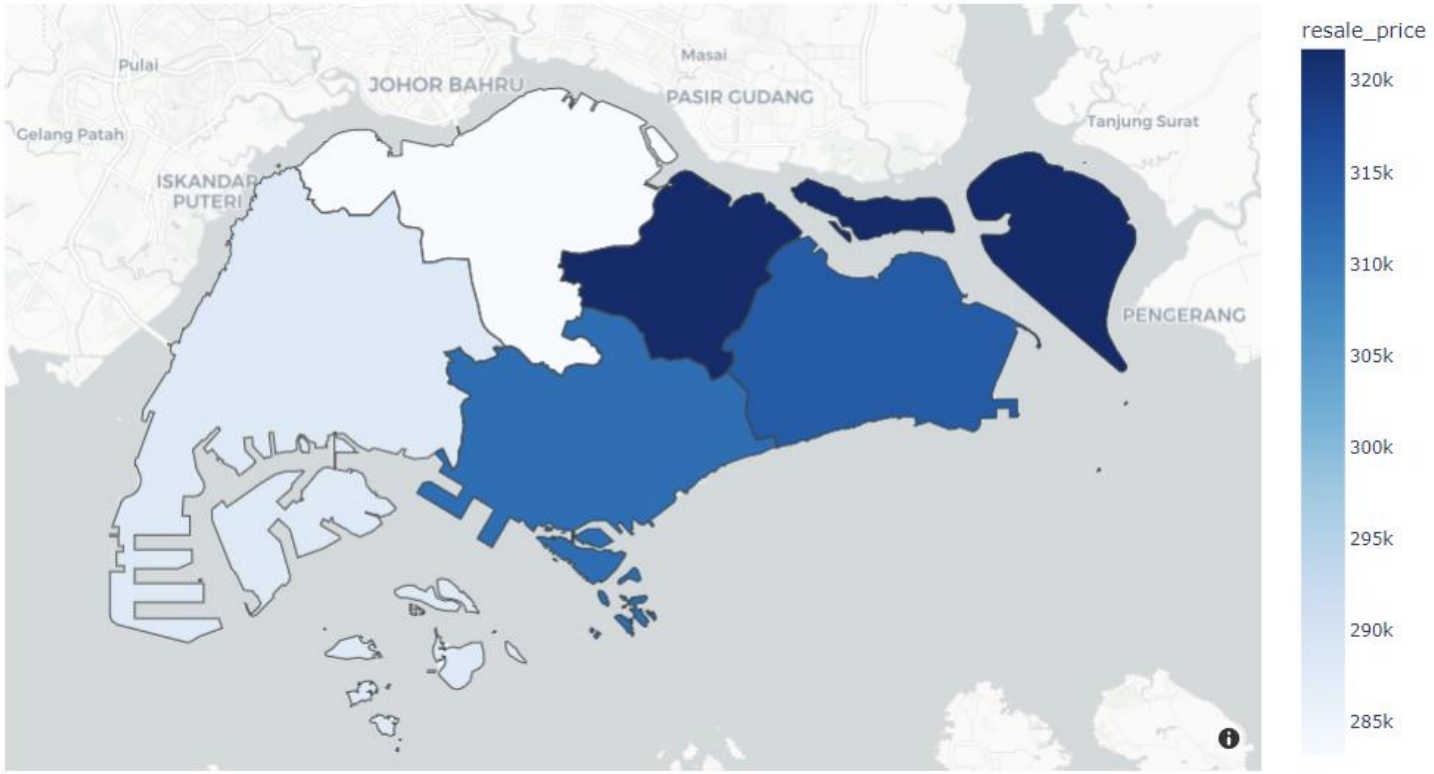
1990 to 1999

2000 to 2009

2010 to 2019

Overall

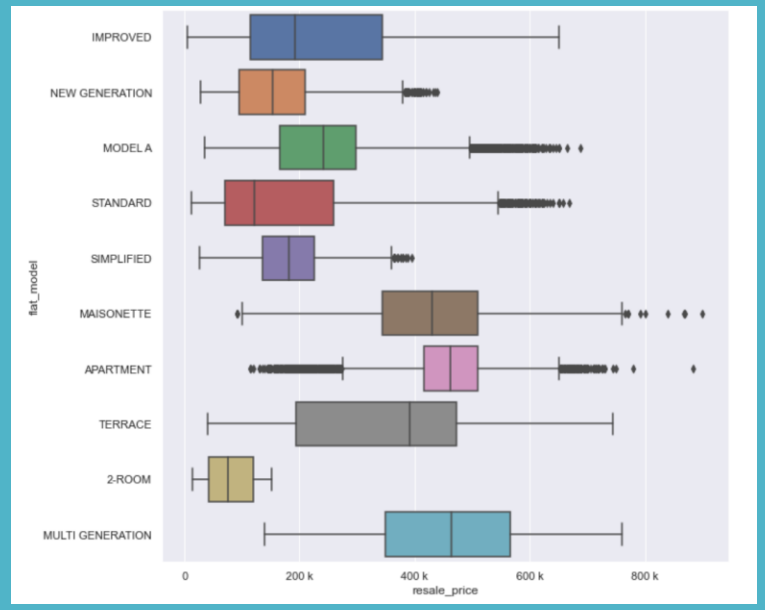
Map Visualization of
Average Resale Price per Region



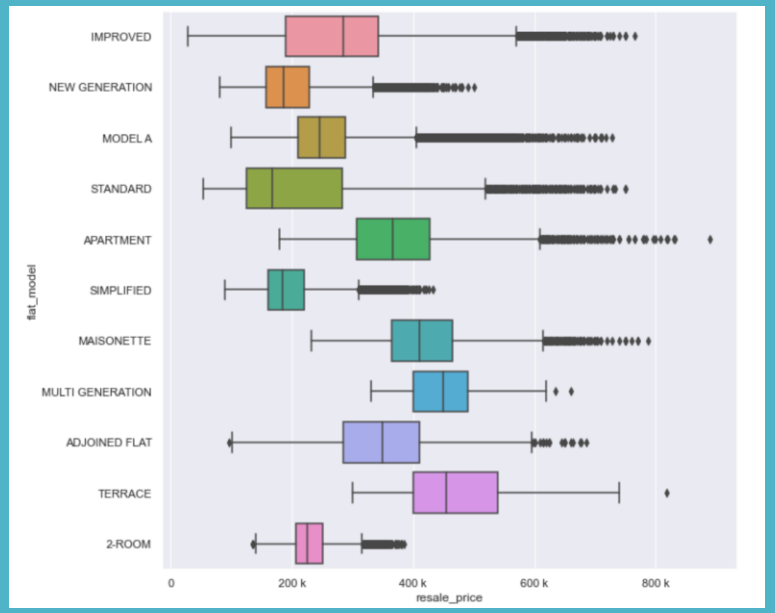
Across the Decades

Distribution Over Resale Price vs Flat Model

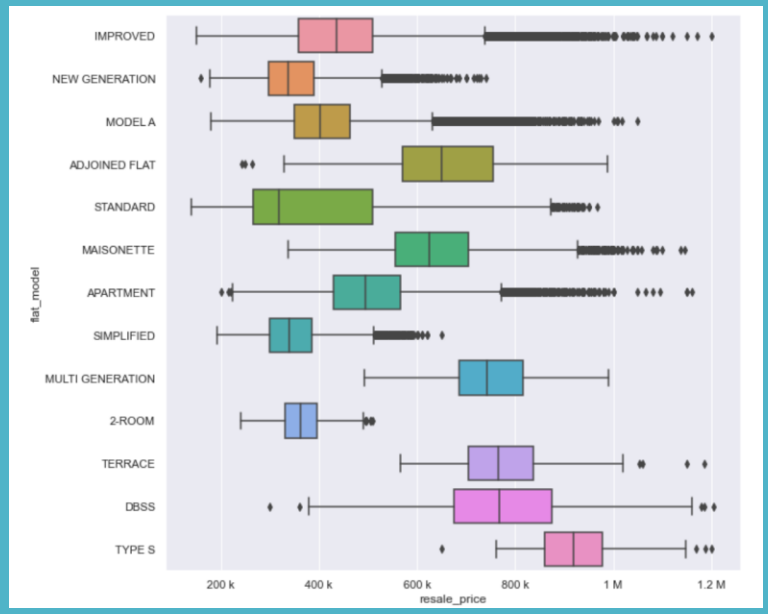
1990 to 1999



2000 to 2009

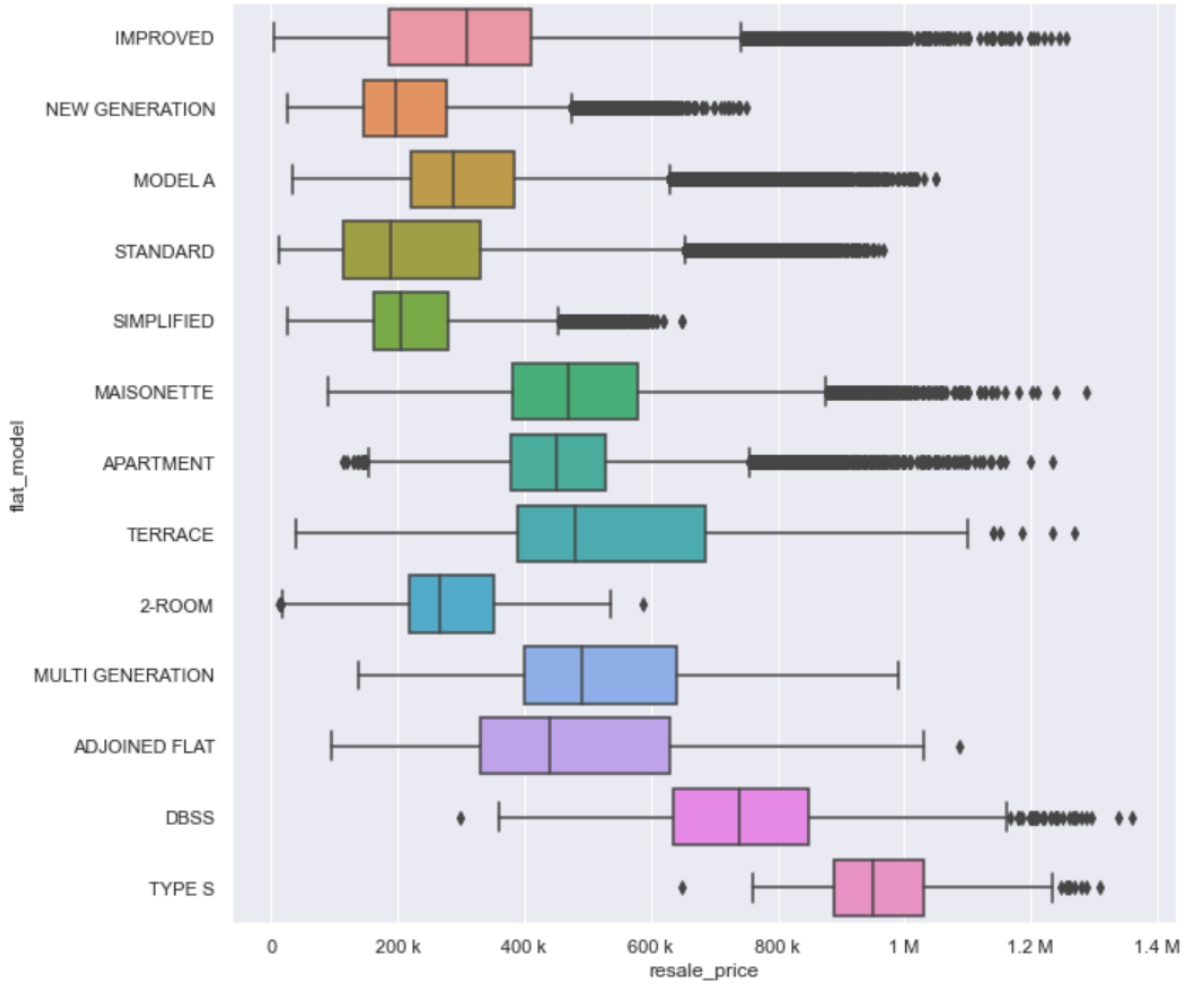


2010 to 2019



Overall

Distribution Over Resale Price vs Flat Model

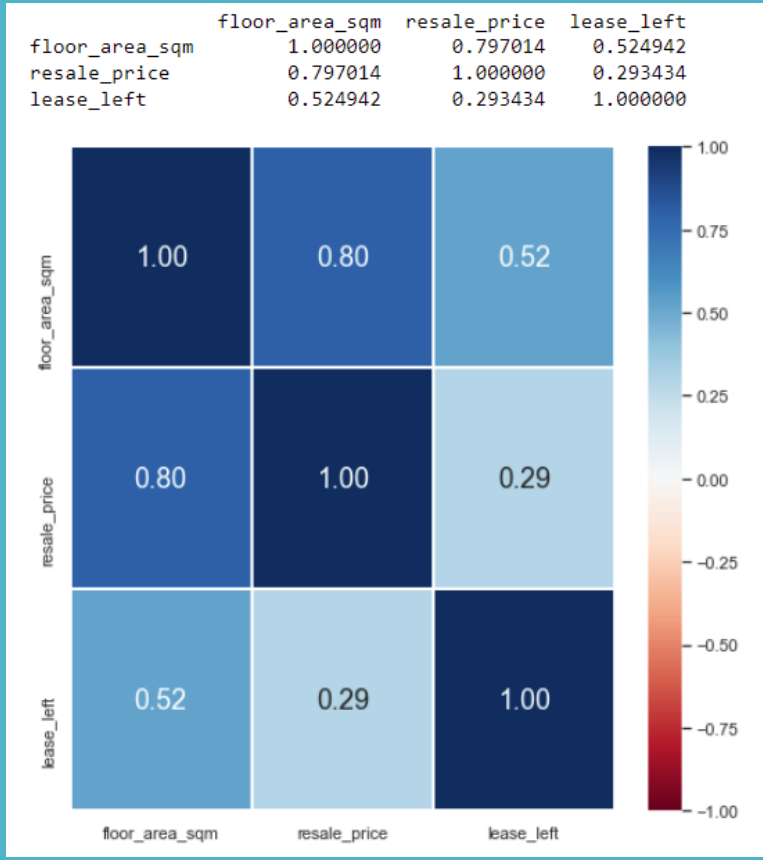


Across the Decades

Correlation of Floor Area Sqm vs Lease Left vs Resale Price

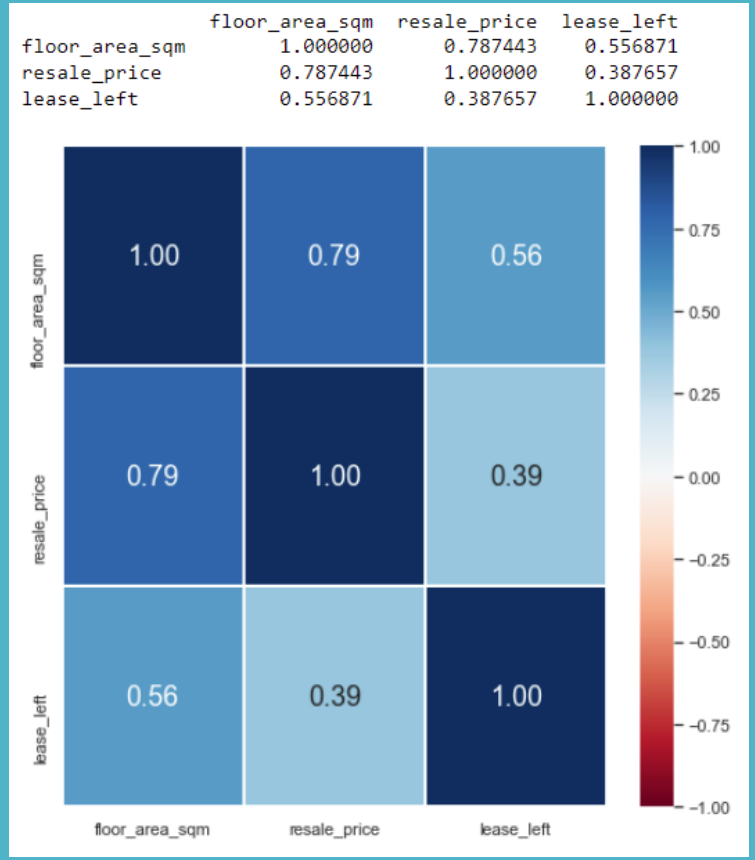
1990 to 1999

	floor_area_sqm	resale_price	lease_left
floor_area_sqm	1.000000	0.797014	0.524942
resale_price	0.797014	1.000000	0.293434
lease_left	0.524942	0.293434	1.000000



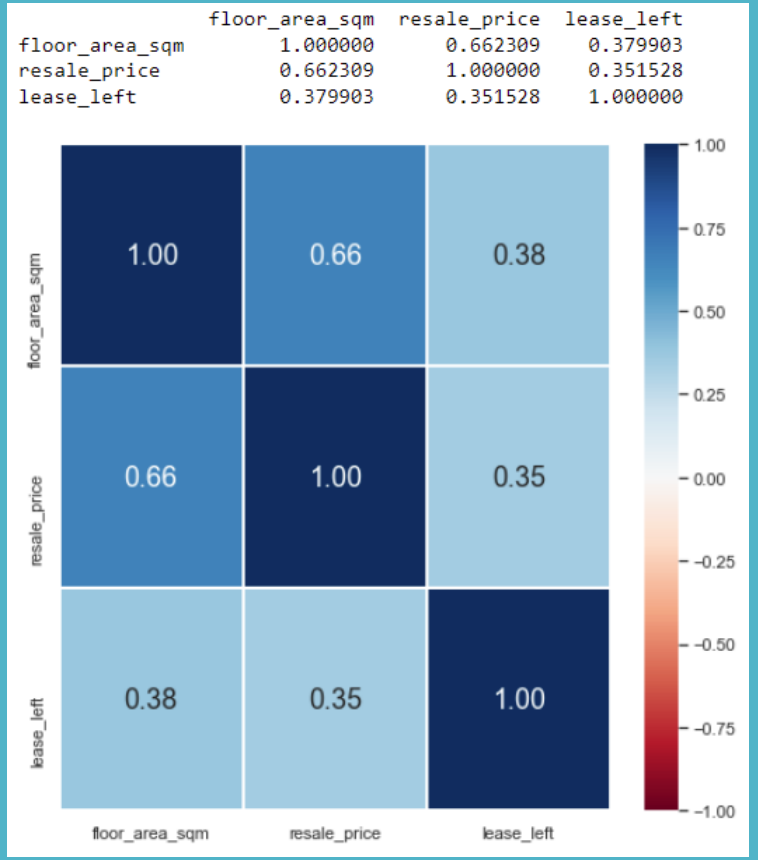
2000 to 2009

	floor_area_sqm	resale_price	lease_left
floor_area_sqm	1.000000	0.787443	0.556871
resale_price	0.787443	1.000000	0.387657
lease_left	0.556871	0.387657	1.000000



2010 to 2019

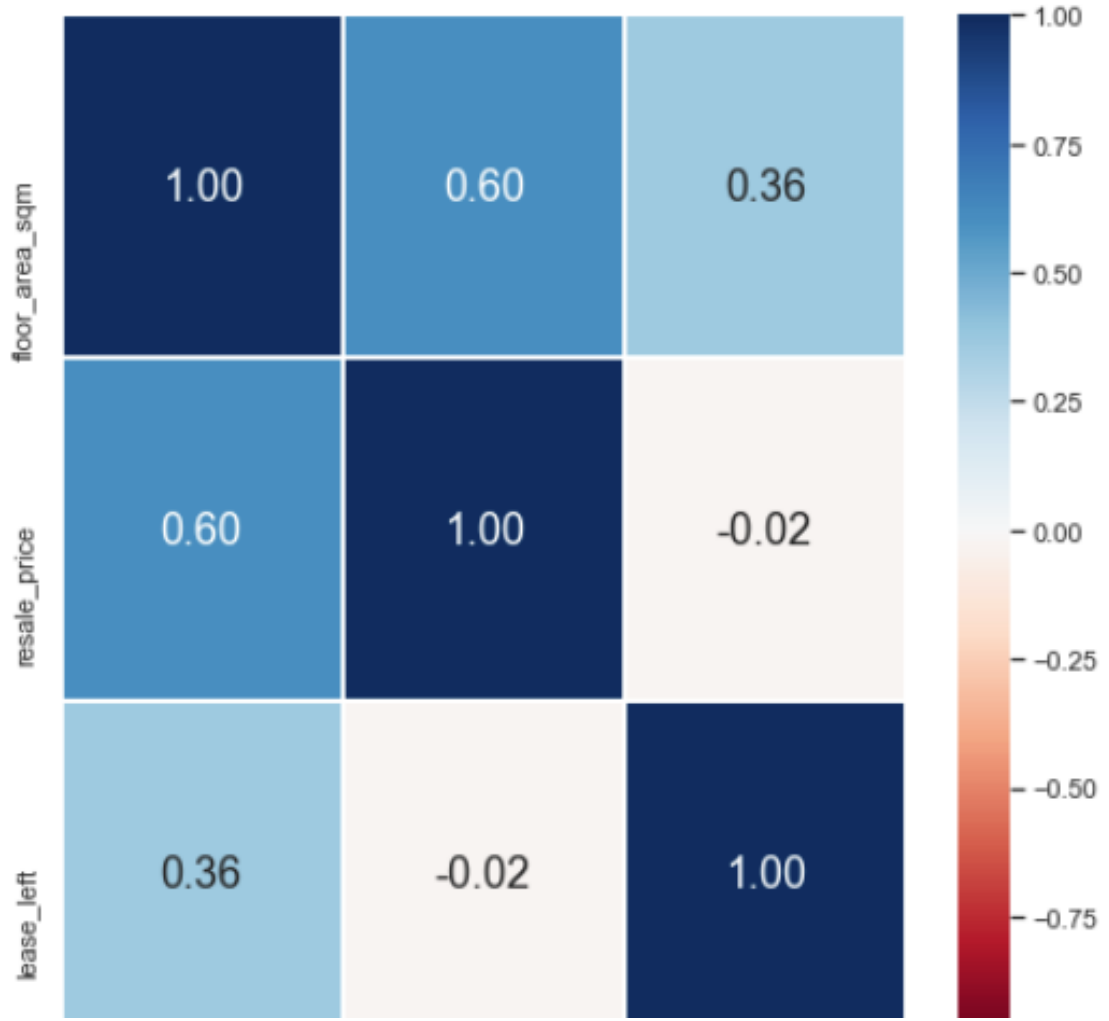
	floor_area_sqm	resale_price	lease_left
floor_area_sqm	1.000000	0.662309	0.379903
resale_price	0.662309	1.000000	0.351528
lease_left	0.379903	0.351528	1.000000



Overall

Correlation of
Floor Area Sqm vs Lease Left vs Resale Price

floor_area_sqm	1.000000	0.603984	0.362372
resale_price	0.603984	1.000000	-0.022325
lease_left	0.362372	-0.022325	1.000000



Machine Learning

- Linear Regression
- Random Forest

Linear Regression:

Model Evaluations per Decade

1990 to 1999

	Coefficient
floor_area_sqm	4156.738117
lease_left	-3686.150345

2000 to 2009

	Coefficient
floor_area_sqm	3138.643817
lease_left	-789.975217

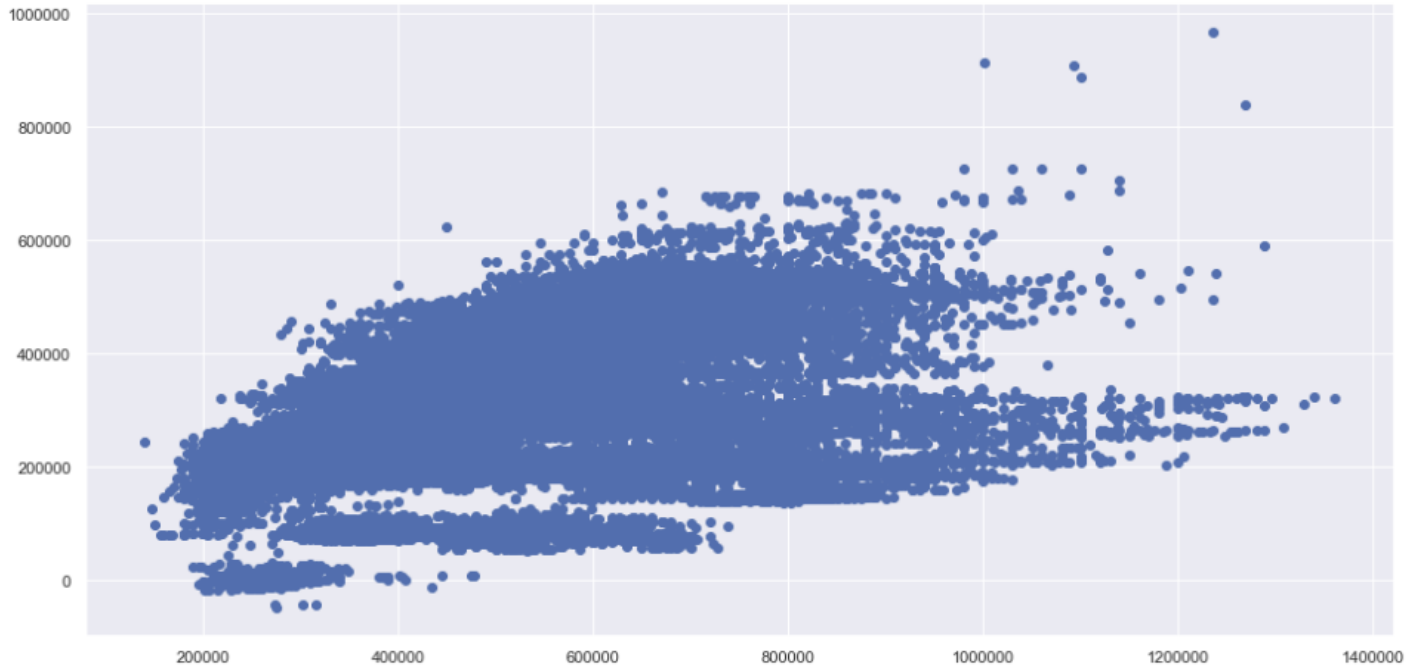
2010 to 2019

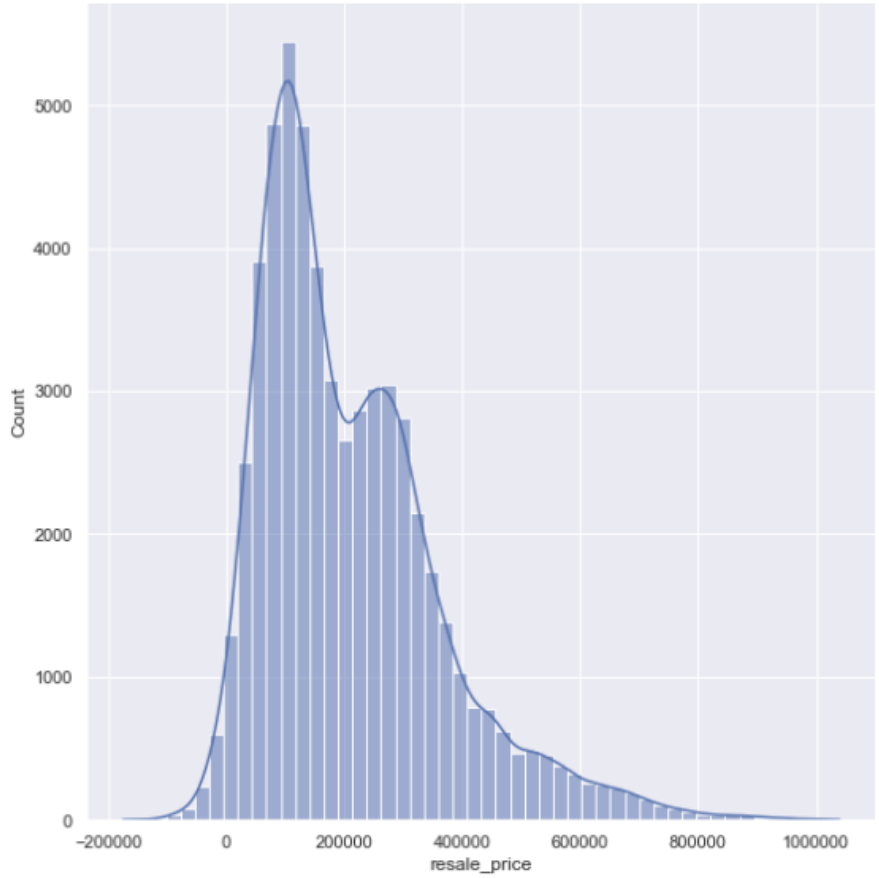
	Coefficient
floor_area_sqm	3330.597262
lease_left	1412.420532

Linear Regression:

Model Predictions per Decade

1990 to 1999



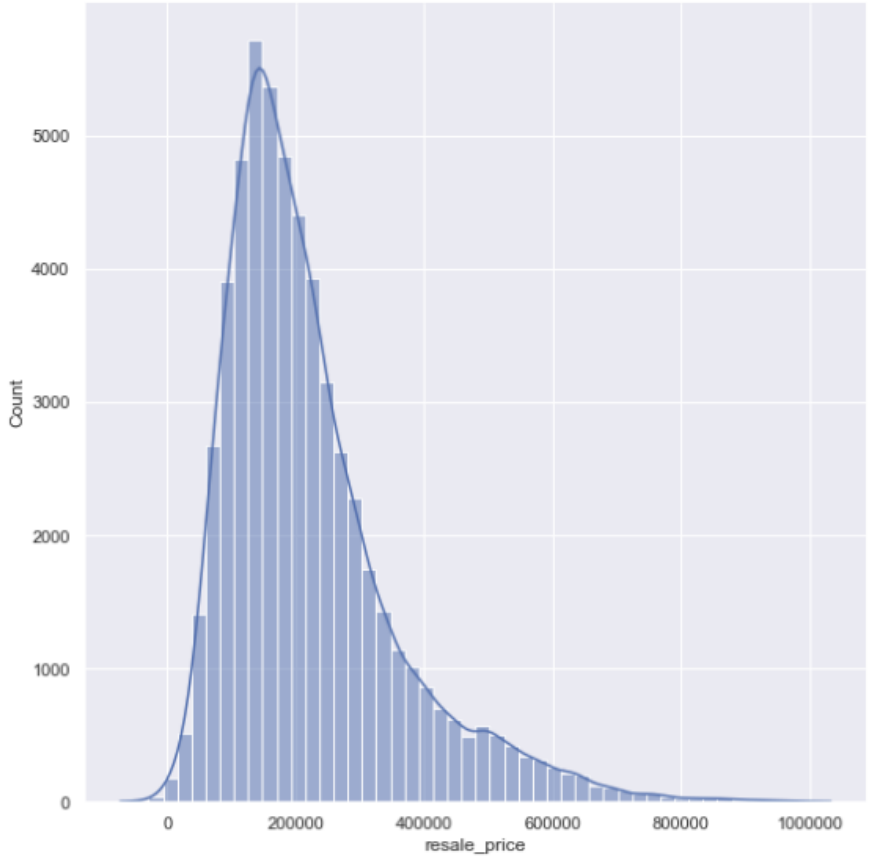


	predictions	y_test_price
predictions	1.000000	0.404902
y_test_price	0.404902	1.000000

Linear Regression:

Model Predictions per Decade

2000 to 2009

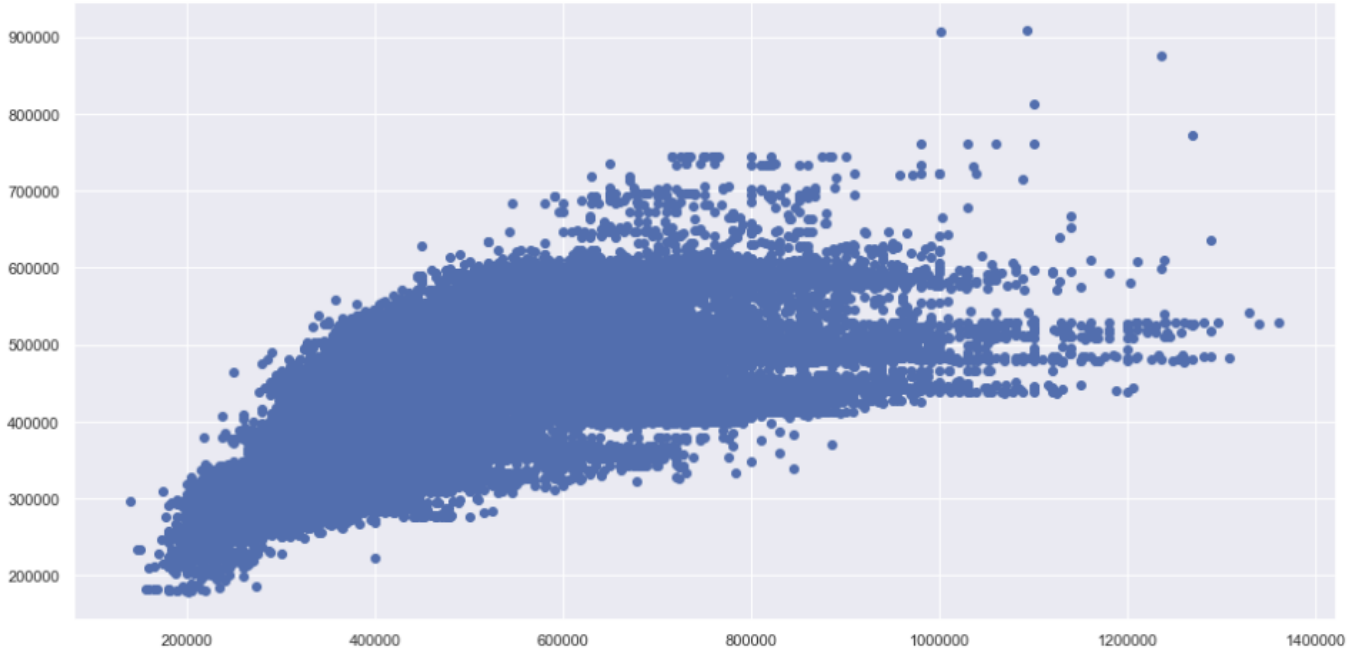


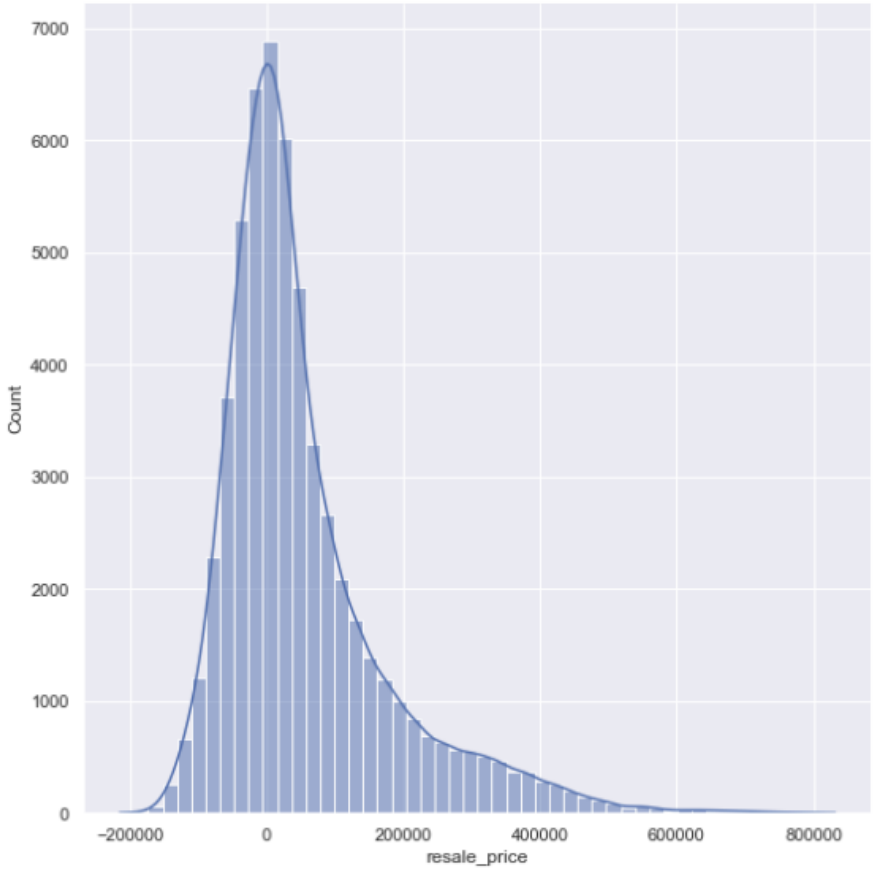
	predictions	y_test_price
predictions	1.00000	0.57337
y_test_price	0.57337	1.00000

Linear Regression:

Model Predictions per Decade

2010 to 2019





	predictions	y_test_price
predictions	1.000000	0.675781
y_test_price	0.675781	1.000000

Linear Regression

Model Comparison

	R2	MSE	RMSE	Period
0	0.656783	6.790244e+10	260580.958106	90
1	0.623813	6.673895e+10	258338.824738	00
2	0.450320	1.798608e+10	134112.200217	10

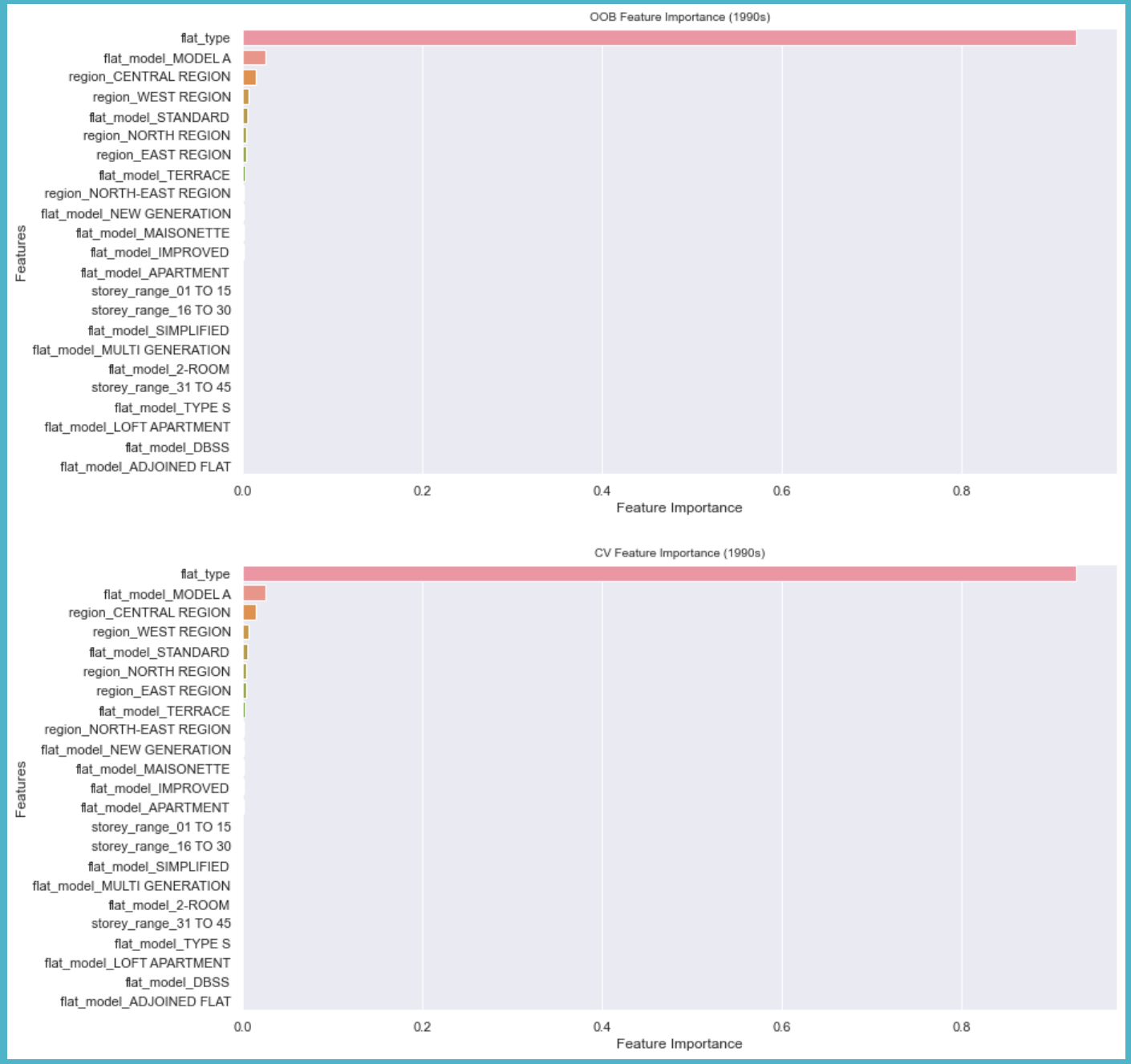
Random Forest Regression:

Training Score

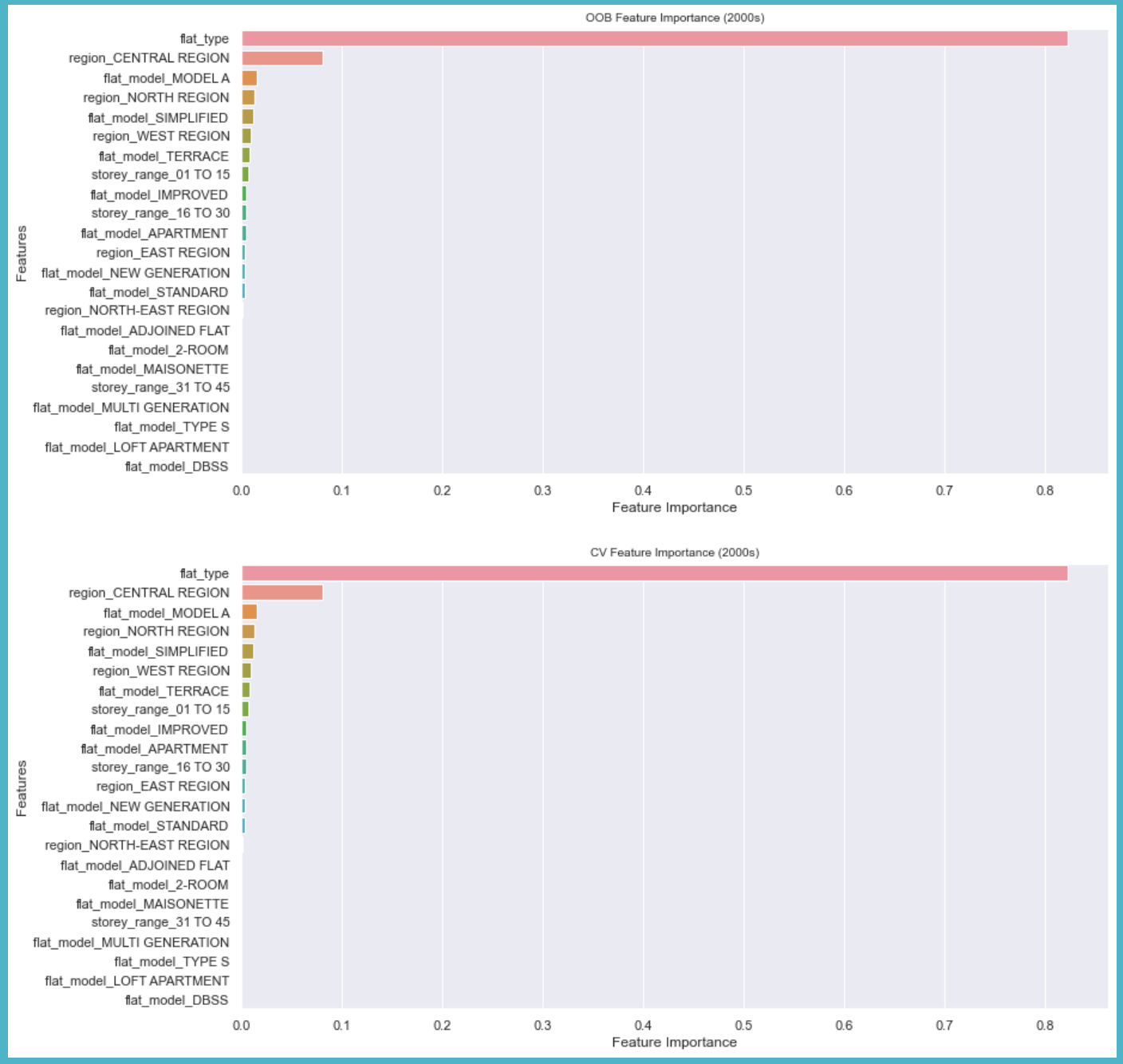
```
Out-of-bag R² score estimate (1990s): 0.672
Cross Validation R² score (1990s): -0.455

Out-of-bag R² score estimate (2000s): 0.769
Cross Validation R² score (2000s): 0.692

Out-of-bag R² score estimate (2010s): 0.766
Cross Validation R² score (2010s): 0.715
```



Random Forest Regression: Feature Importance (1990s)



Random Forest Regression: Feature Importance (2000s)

Random Forest Regression: Feature Importance (2010s)

Random Forest Regression:

Testing on 2020s data

Out-of-bag

1990s

Test data R^2 score: -1.6
Test data Spearman correlation: 0.761
Test data Pearson correlation: 0.711
Test data Mean Absolute Error: 235741

2000s

Test data R^2 score: -1.05
Test data Spearman correlation: 0.833
Test data Pearson correlation: 0.816
Test data Mean Absolute Error: 208237

2010s

Test data R^2 score: 0.701
Test data Spearman correlation: 0.87
Test data Pearson correlation: 0.876
Test data Mean Absolute Error: 64259

K-fold cross validation w/grid search

1990s

Test data R^2 score: -1.61
Test data Spearman correlation: 0.759
Test data Pearson correlation: 0.709
Test data Mean Absolute Error: 236162

2000s

Test data R^2 score: -1.05
Test data Spearman correlation: 0.836
Test data Pearson correlation: 0.817
Test data Mean Absolute Error: 208073

2010s

Test data R^2 score: 0.7
Test data Spearman correlation: 0.869
Test data Pearson correlation: 0.876
Test data Mean Absolute Error: 64407

Conclusion

- Drivers for each generation
- Recession

- Linear Regression
- Random Forest

References

- Resale Flat Prices

https://data.gov.sg/dataset/resale-flat-prices

- Singapore Regions

https://data.gov.sg/dataset/master-plan-2019-subzone-boundary-no-sea

- Map Visualization

https://plotly.com/python/mapbox-county-choropleth/

- Correlation and Linear Regression

http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/

- Residual Plot Histogram

https://www.originlab.com/doc/Origin-Help/Residual-Plot-Analysis

- Linear Regression

https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/