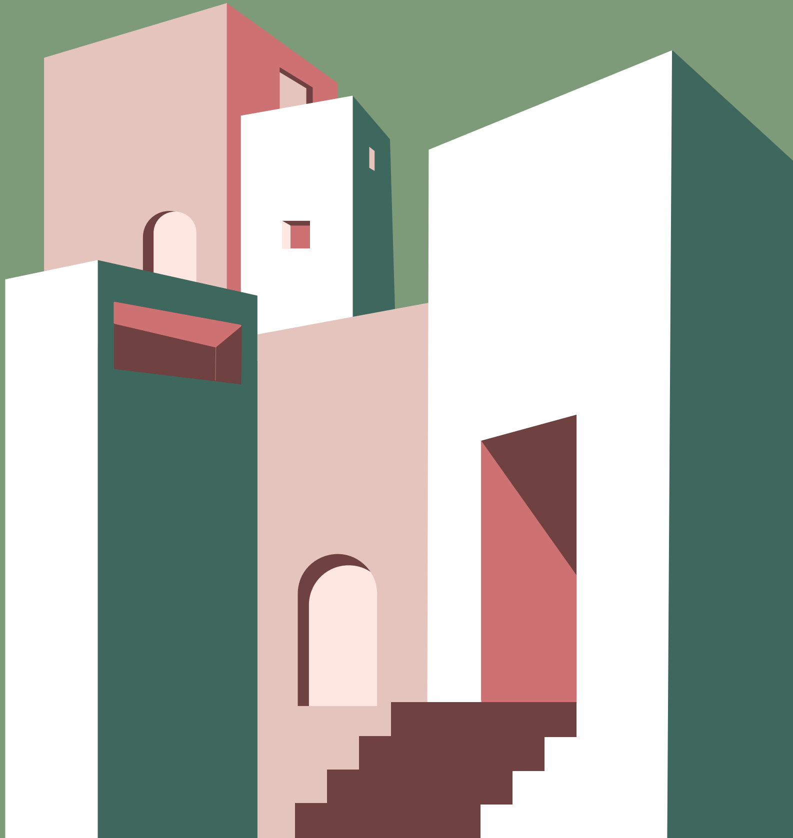
HDB RESALE PRICE PREDICTION

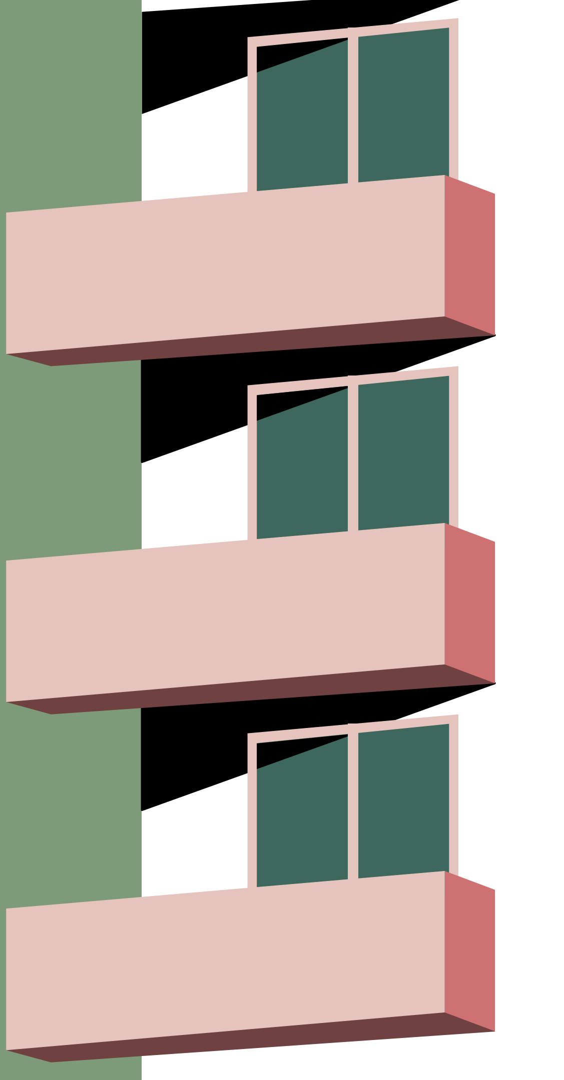
Benedict, Zach, Kai
Group 2



CONTENTS PAGE

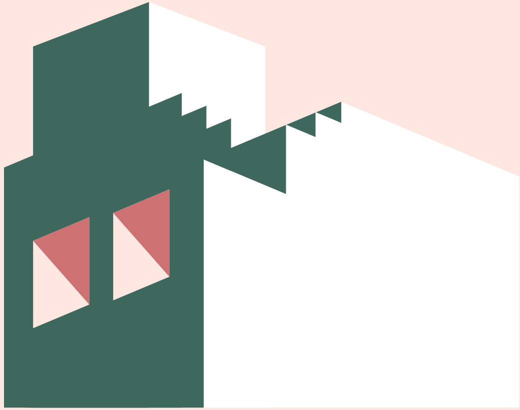
01

OVERVIEW OF PROJECT



PROBLEM STATEMENT

Using Singapore public housing data, we will be creating a regression model that predicts the price of Housing Development Board (HDB) flats in Singapore.



OVERVIEW OF THE PROJECT

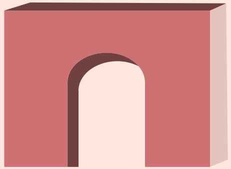
FEATURES

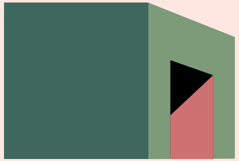
1. ***Location information*** - 8 Features:
   - e.g. Street name, town, address, postal code etc

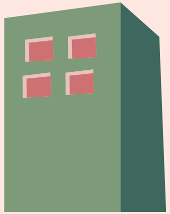
2. ***Flat information*** - 37 Features:
   - e.g. flat type, storey range, type of residential room sold (2,3,4) etc

3. ***Transport availability*** - 10 Features:
   - e.g. closest mrt station distance, closest mrt station name, closest bus distance etc

4. ***Amenities availability*** - 22 Features:
   - e.g. Primary and Secondary School closest distance, hawker centres closest distance, mall closest distance etc







TARGET

RESALE PRICE

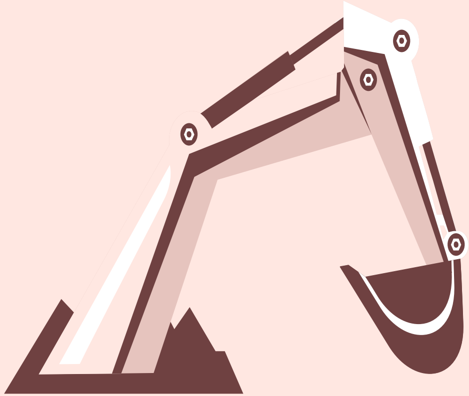
02

BASELINE MODEL

2A. BASELINE PREPROCESSING

1. **Removal of noise and composite data for technical reasons. For example:**

   - 'ID no': Noise with no value in prediction
   - 'postal code': Noise with no value in prediction
   - 'Address': Noise and composite data
   - 'Tranc_YearMonth': Composite data of features 'Year' and 'Month'
   - 'Full_flat_type': Composite data of features 'flat_type' and 'flat_model'
   - 'Bus_stop_name': Noise with no value in prediction



2A. BASELINE PREPROCESSING

2. Process cells with missing/null values to ensure the running of the model

We chose to remove the columns with null values after taking a look at the data dictionary. We found that the columns with null values were regarding:

- Flat distance from hawker(s)
- Flat distance from mall(s)

For Hawker distance we found that other columns without missing values existed, such as “hawker_nearest_distance” and “hawker_market_stalls”.

For mall distance, after taking a look at some missing values we found that the missing values were missing at random. We dropped these columns for technical reasons.(incompatible with kaggle submission)

	na values
Hawker_Within_500m	97390
Mall_Within_500m	92789
Hawker_Within_1km	60868
Hawker_Within_2km	29202
Mall_Within_1km	25426
Mall_Within_2km	1940
Mall_Nearest_Distance	829

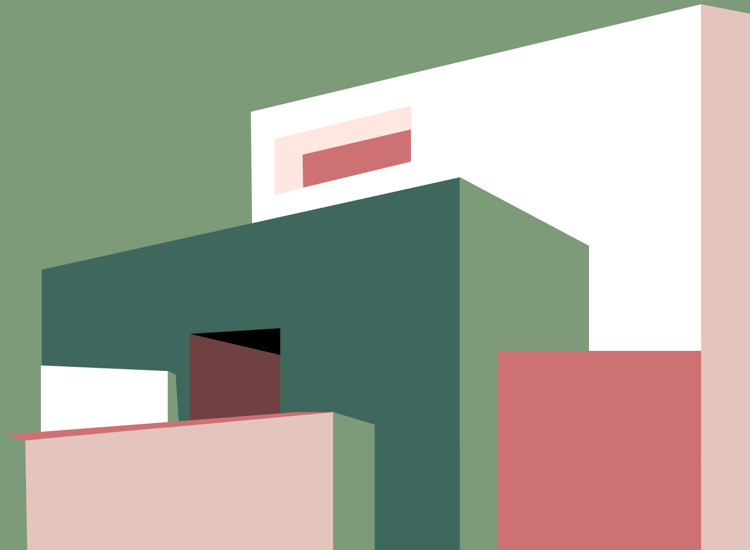
2B. BASELINE MODEL WORKFLOW

<u>Transformers used</u>:

- Standard Scalar
- One Hot Encoding

(Credit: Soon Poh)

2C. BASELINE MODEL EVALUATION

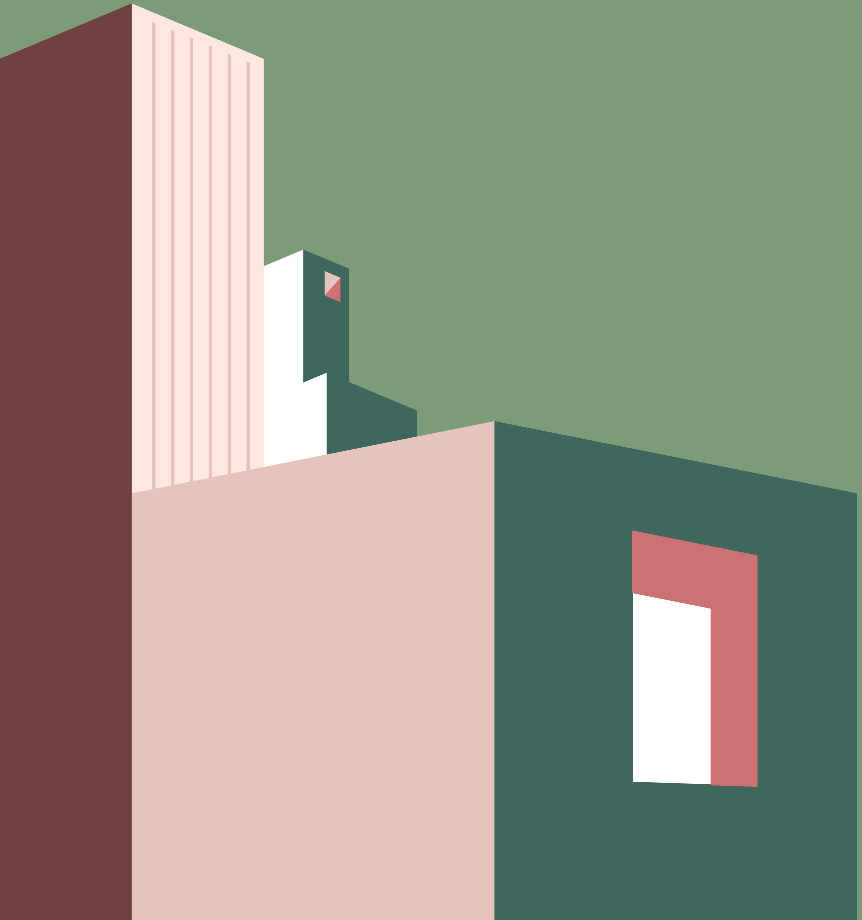


Baseline R^2 scores

TRAIN SET: 0.937
HOLDOUT SET: - 2.89 E^{14}

Evaluation

Baseline model is grossly overfitted
with high variance and low bias



03

ENHANCED MODEL

3A. ENHANCEMENTS FOR BASELINE MODEL

EDA

Understanding the data:

- Multi-colinearity via heatmap
- OLS for feature selection

FEATURE ENGINEERING

- Inputting missing data
- Dissecting of features (i.e. postal code)

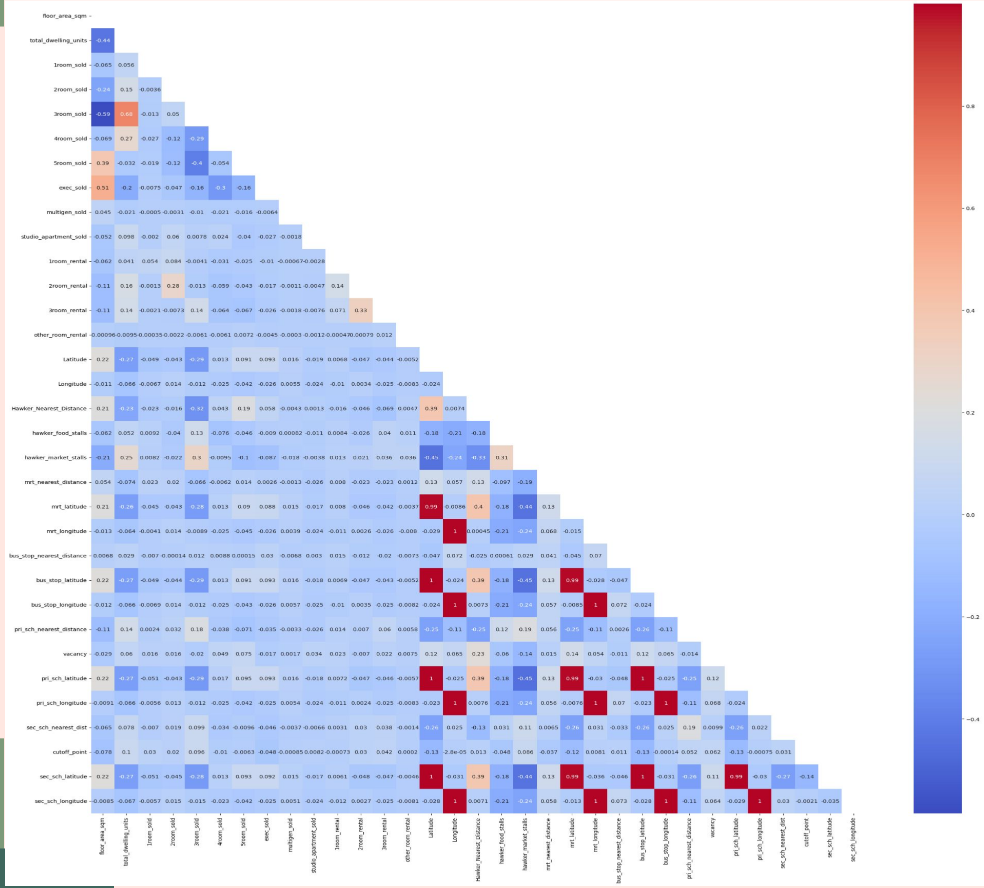
REGULARIZATION & HYPERPARAMETER TUNING

Lasso
&
Ridge

EVALUATION

Looking at the metrics to compare between different EDA/hyperparameter combination

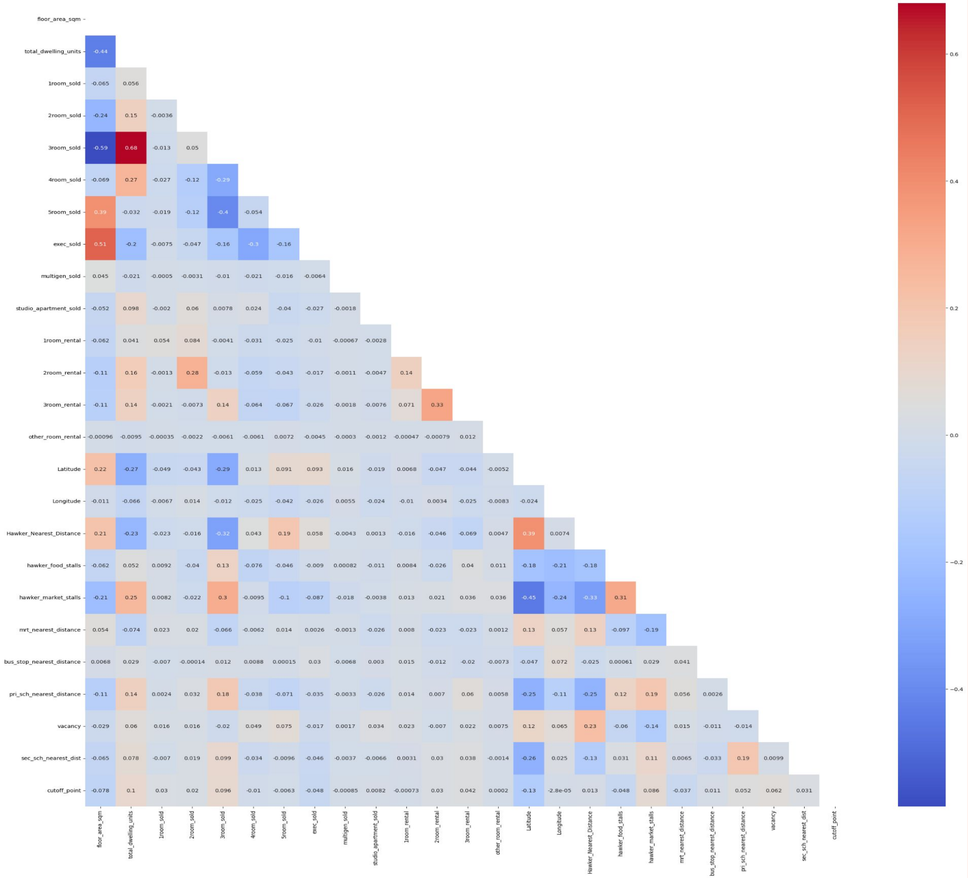
3B. EDA - MULTI-COLINEARITY



Red boxes:

- Identify features with coefficient > 0.75
- Correlation coefficient of >0.98
- All of them are features relating to longitude and latitude
- Logical to be multi-colinear as the figure should be very similar given how small Singapore is
- Decided to remove them

3B. EDA - MULTI-COLINEARITY



Post-removal

- Features removed:
  - 'sec_sch_latitude',
  - 'sec_sch_longitude',
  - 'pri_sch_latitude',
  - 'pri_sch_longitude',
  - 'Bus_stop_latitude',
  - 'bus_stop_longitude',
  - 'mrt_latitude',
  - 'mrt_longitude'
- Feature with highest correlation coefficient is 0.68 which is acceptable

3C. FEATURE SELECTION WITH OLS

Dep. Variable:	resale_price	**R-squared:**	0.655
Model:	OLS	**Adj. R-squared:**	0.655
Method:	Least Squares	**F-statistic:**	8945.
Date:	Mon, 26 Dec 2022	**Prob (F-statistic):**	0.00
Time:	17:57:11	**Log-Likelihood:**	-1.4414e+06
No. Observations:	112975	**AIC:**	2.883e+06
Df Residuals:	112950	**BIC:**	2.883e+06
Df Model:	24		
Covariance Type:	nonrobust		

Omnibus:	15685.444	**Durbin-Watson:**	2.008
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	30251.318
Skew:	0.880	**Prob(JB):**	0.00
Kurtosis:	4.825	**Cond. No.**	1.15e+16

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.391e+07	3.87e+05	-61.842	0.000	-2.47e+07	-2.32e+07
floor_area_sqm	3494.8914	17.189	203.316	0.000	3461.200	3528.582
total_dwelling_units	5350.6052	786.052	6.807	0.000	3809.956	6891.254
1room_sold	-5524.6670	786.906	-7.021	0.000	-7066.990	-3982.344
2room_sold	-4872.9410	786.228	-6.198	0.000	-6413.936	-3331.946
3room_sold	-5400.1481	785.971	-6.871	0.000	-6940.640	-3859.656
4room_sold	-4726.3802	785.956	-6.014	0.000	-6266.842	-3185.918
5room_sold	-4465.2260	786.067	-5.680	0.000	-6005.906	-2924.546
exec_sold	-4110.9900	786.114	-5.230	0.000	-5651.761	-2570.219
multigen_sold	-2439.6812	812.122	-3.004	0.003	-4031.428	-847.934
studio_apartment_sold	-3484.5915	787.190	-4.427	0.000	-5027.471	-1941.712
1room_rental	-4476.5271	793.292	-5.643	0.000	-6031.367	-2921.687
2room_rental	-5684.7768	786.613	-7.227	0.000	-7226.527	-4143.027
3room_rental	-3964.9233	874.812	-4.532	0.000	-5679.542	-2250.305
other_room_rental	5.45e+04	9427.315	5.781	0.000	3.6e+04	7.3e+04
Latitude	-1.109e+06	7250.013	-153.033	0.000	-1.12e+06	-1.1e+06
Longitude	2.45e+05	3706.201	66.098	0.000	2.38e+05	2.52e+05
Hawker_Nearest_Distance	-8.4114	0.275	-30.558	0.000	-8.951	-7.872
hawker_food_stalls	-171.8958	14.158	-12.141	0.000	-199.646	-144.145
hawker_market_stalls	-10.5714	5.617	-1.882	0.060	-21.580	0.437
mrt_nearest_distance	-40.4189	0.604	-66.928	0.000	-41.603	-39.235
bus_stop_nearest_distance	0.2419	4.549	0.053	0.958	-8.675	9.159
pri_sch_nearest_distance	-1.0460	1.155	-0.905	0.365	-3.310	1.219
vacancy	141.4560	14.687	9.632	0.000	112.670	170.242
sec_sch_nearest_dist	25.5030	0.849	30.027	0.000	23.838	27.168
cutoff_point	312.2320	12.803	24.387	0.000	287.138	337.326

FEATURE WITH P > 0.05

- ‘Hawker_market_stall’
- ‘bus_stop_nearest_distance’
- ‘pri_sch_nearest_distance’

OBSERVATION

- After removing these 3 features, the metrics and kaggle scores actually worsened
- Logically-speaking, these features are important too
- Therefore, these features are kept for analysis

3D_1. INPUTTING MISSING DATA

	town	Mall_Nearest_Distance	na_count
town			
PUNGGOL	7793	7614	179
SENGKANG	11069	10894	175
CHOA CHU KANG	6343	6216	127
QUEENSTOWN	4121	4062	59
TAMPINES	10506	10463	43
JURONG EAST	3470	3431	39
WOODLANDS	11334	11299	35
GEYLANG	3986	3951	35
BEDOK	9046	9023	23
BUKIT PANJANG	5686	5664	22
BUKIT MERAH	5854	5834	20
HOUGANG	7555	7537	18
PASIR RIS	4763	4746	17
TOA PAYOH	4817	4804	13
JURONG WEST	11451	11445	6
KALLANG/WHAMPOA	4340	4334	6
YISHUN	10042	10037	5
CLEMENTI	3633	3628	5
BISHAN	2871	2870	1
ANG MO KIO	6908	6907	1
MARINE PARADE	959	959	0

Inputting missing values for ‘Mall_nearest_distance’:

- 829 missing values for this feature - viable to try to input
- Used median distance of their respective Town to input
- Chose median as it provides the best metrics

3E. DISSECTING FEATURES

- Decided to dissect postal code to determine residential district using the first 2 digits

- ~10 entries have NIL postal code and we determine the postal code manually using their street name

6 Digit Postal System

Later in year 1995, the six digit postal system replaced the previous four digit postal system. Since then, every building has a unique postal code with the first two digits being the postal sector. For example, Tuas starts with 63xxxx.

```
array(['76', '64', '56', '73', '65', '46', '82', '32', '68', '14', '10',
       '60', '52', '54', '37', '13', '51', '12', '44', '90', '53', '75',
       '39', '27', '15', '67', '31', '35', '61', '55', '16', '79', '47',
       '57', '81', '40', '21', '38', '19', '26', '41', '43', '33', '80',
       '85', '18', '30', '50', '36', '66', '91', '59', '20', '42', 54,
       '58'], dtype=object)
```

REGULARIZATION & HYPERPARAMETER TUNING

RIDGE

Parameters :
Alpha = 1.0
CV = LOOCV

LASSO

Parameters :
Alpha = 1.0
CV = 5

** Reason for Regularization: Model is overfitted in baseline model*

Ridge R^2 and RMSE

TRAIN SET: 0.932

HOLDOUT SET: 0.934

RMSE: 36708

Lasso R^2 and RMSE

TRAIN SET: 0.932

HOLDOUT SET: 0.933

RMSE: 36823

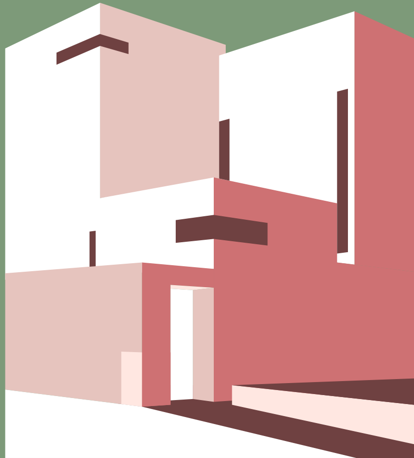
RIDGECV

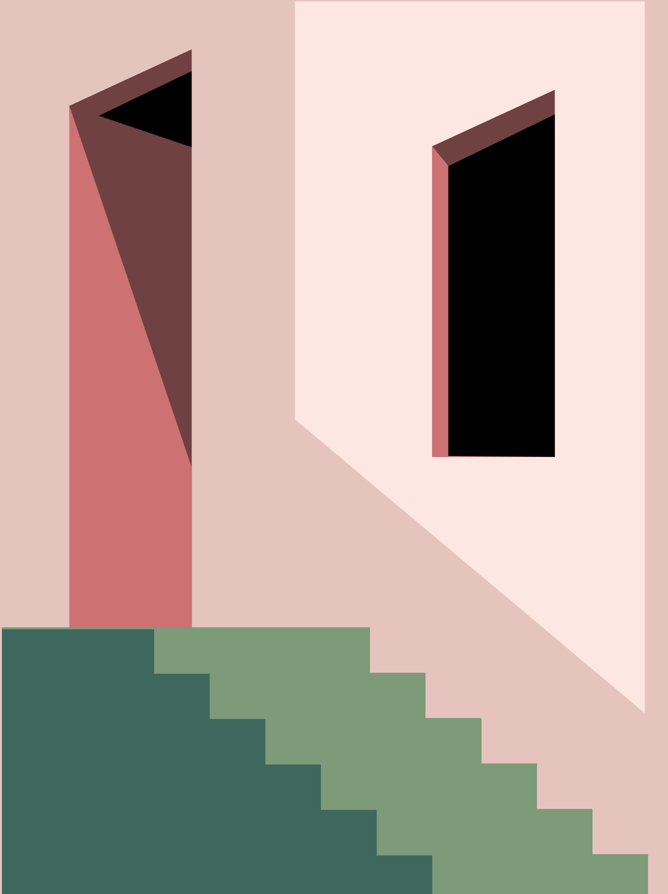
Parameters:

- Alpha range = np.logspace(0, 5,100)
- CV = 5

R2 Score: 0.927

RMSE: 38629





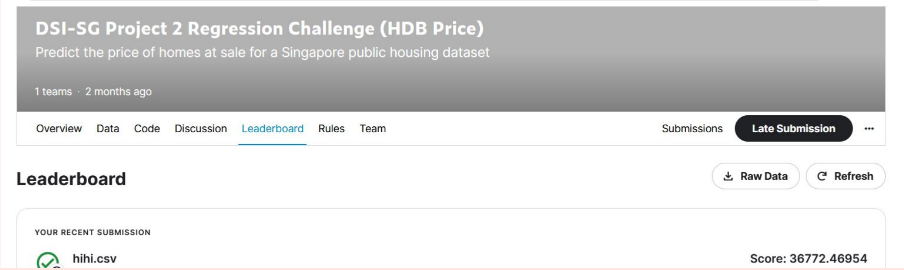
04

KAGGLE SUBMISSION

KAGGLE PREDICTION

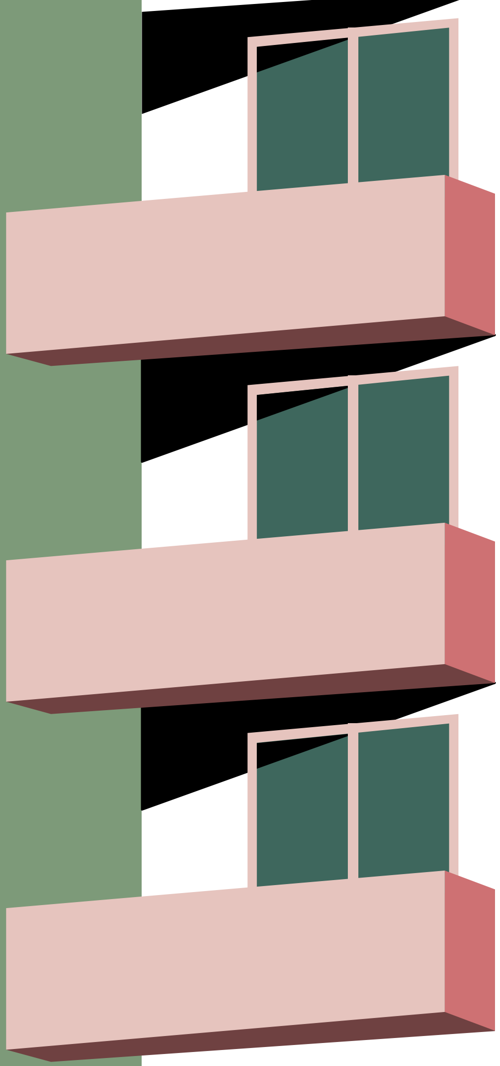
RMSE OF TEST DATA

36,772



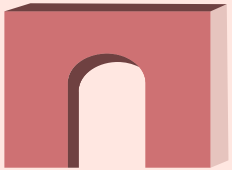
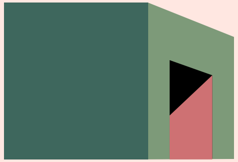
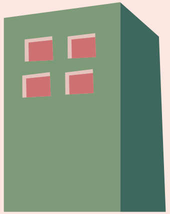
05

RECOMMENDATIONS AND CONCLUSION



5. CONCLUSION AND RECOMMENDATIONS

- This project is supplied with a vast amount of data, both in terms of number of features and number of transactions
- Therefore, the model created can provide a R2 metrics which is as high as 0.93
- This model should be further refined with more entries to increase its accuracy further
- The current model still has many features despite the efforts to reduce the model's dimensionality as much as possible
- Hopefully, the number of features could be further reduced as the model is fed with more data (i.e. increase the predictive power of certain keys features)

THANK YOU!