HDB resale prices

Samuel Cheah
Holmusk coding challenge

Problem Statement

HDB has provided a dataset of resale HDB transactions and would like understand the drivers of resale prices.

Date range analyzed: **2012-2020**

Dataset features

1. Month - YYYY-MM format
2. Town - AMK, Bishan, etc
3. Flat_type - 2-room, 3-room etc
4. Block
5. Street name
6. Storey range - e.g. unit is between storey 10-12
7. Floor area - in square meters (sqm)
8. Flat model - Standard, Improved, Maisonette etc
9. Lease commence date
10. Resale price

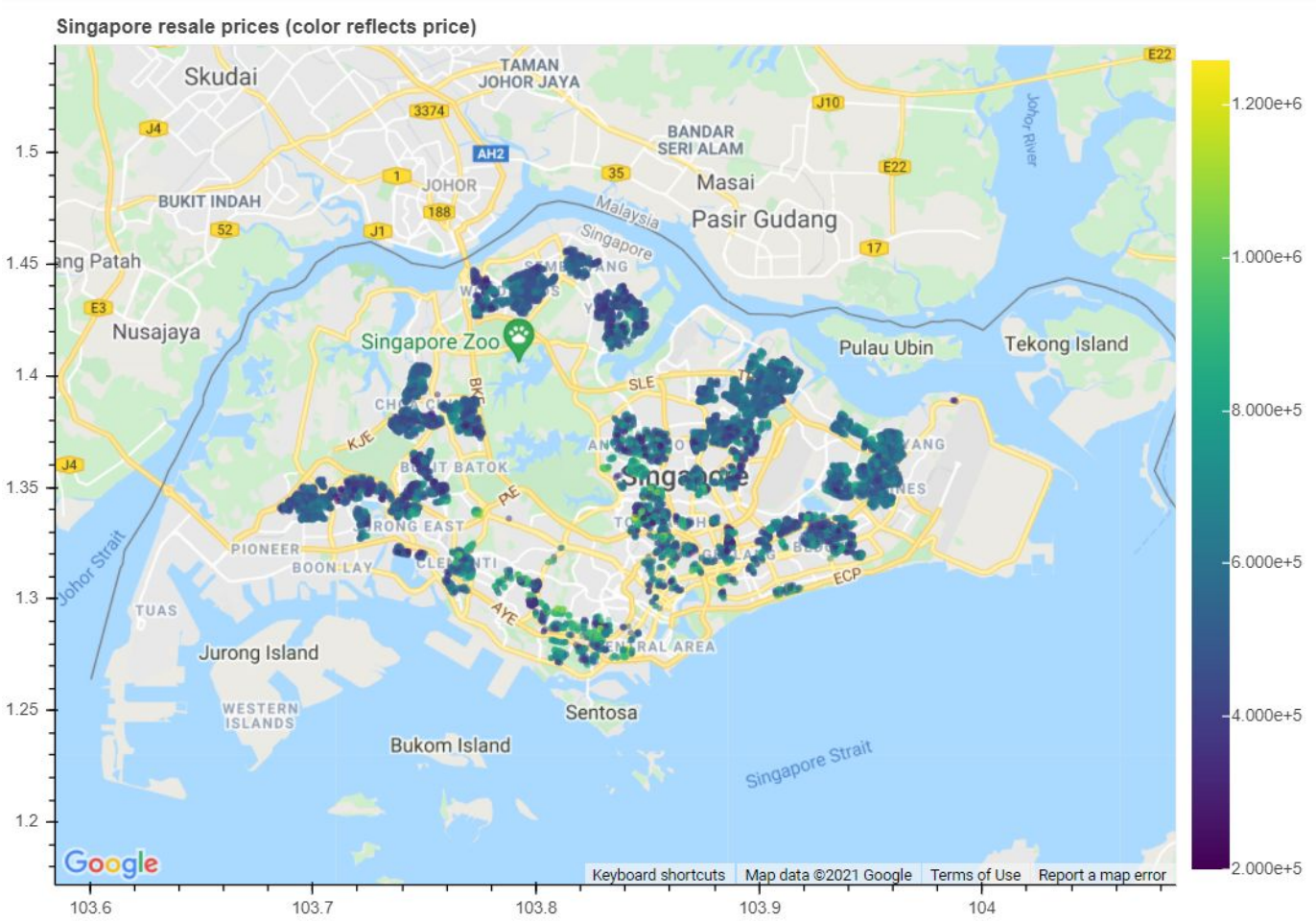
Data preprocessing

1. Resale price
    a. Adjusted for inflation using CPI for housing.
2. Month (YYYY-MM)
    a. Split into ‘year’ and ‘mth’ features.
3. Flat type
    a. Recoded on ordinal integer scale, 0-6.
4. Storey Range
    a. Converted to average of range
5. Lease commencement date
    a. Transformed into a remaining lease years feature

Data preprocessing

Combine block and street name into address

Use Onemap API to pull coordinates for each address.


Singapore resale prices (color reflects price)
Skudai
TAMAN JOHOR JAYA
BANDAR SERI ALAM
Masai
Pasir Gudang
BUKIT INDAH
JOHOR
Malaysia
Singapore
Nusajaya
Singapore Zoo
Pulau Ubin
Tekong Island
BUKIT BATOK
JURONG EAST
PIONEER
BOON LAY
TUAS
Jurong Island
WESTERN ISLANDS
Sentosa
Bukom Island
Singapore Strait
Johor Strait
Johor River
Google
Keyboard shortcuts
Map data ©2021 Google
Terms of Use
Report a map error
1.5
1.45
1.4
1.35
1.3
1.25
1.2
103.6
103.7
103.8
103.9
104
1.200e+6
1.000e+6
8.000e+5
6.000e+5
4.000e+5
2.000e+5

Feature engineering

Distance to nearest MRT/LRT

- Better MRT access -> more valuable

Distance to nearest shopping mall

- Easier to access essential services, like food, groceries, banking - > more valuable

Feature engineering

Proximity to Top 20 primary school

- Primary school priority school admission is dependent on proximity of unit.
- Might be important to parents with young children.

Feature engineering

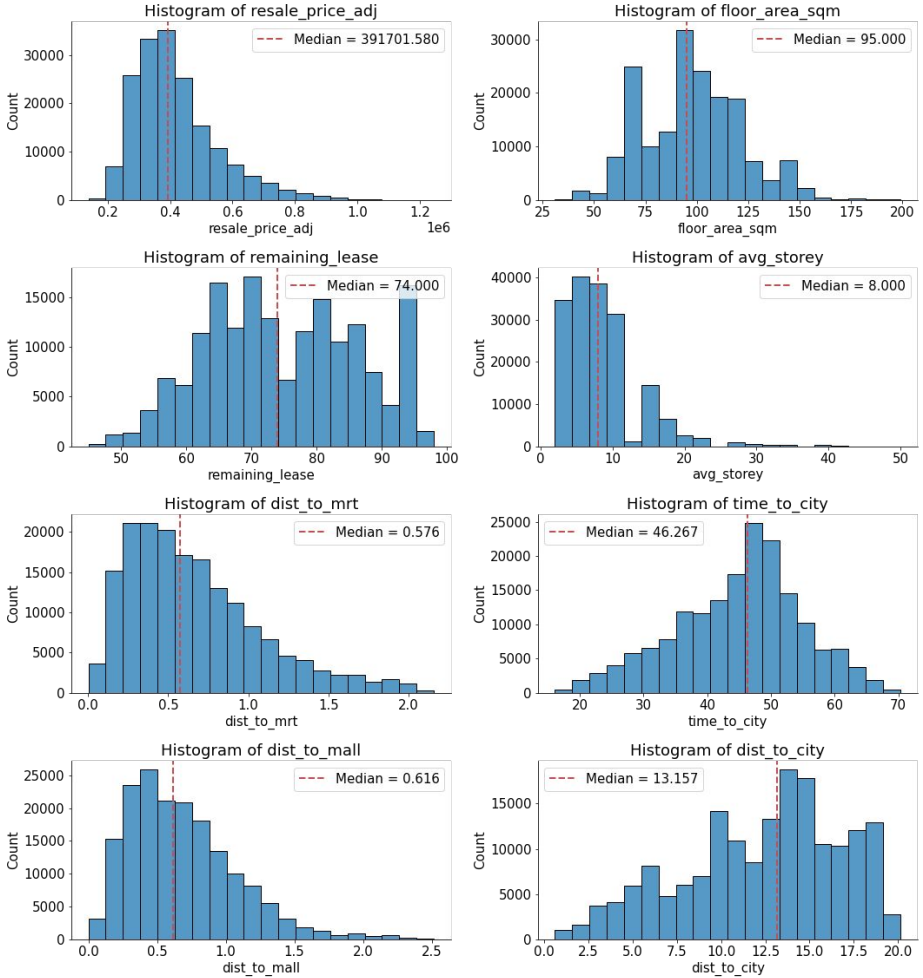
Distance/time to city centre

- Time to travel to work also important consideration
- Both variable created due to uncertainty about accuracy of Onemap API for time estimates

EDA

Examined distributions of numerical variables

- Resale price is right skewed.
- Expensive HDBs are rarer since HDBs are meant to be an affordable housing avenue.



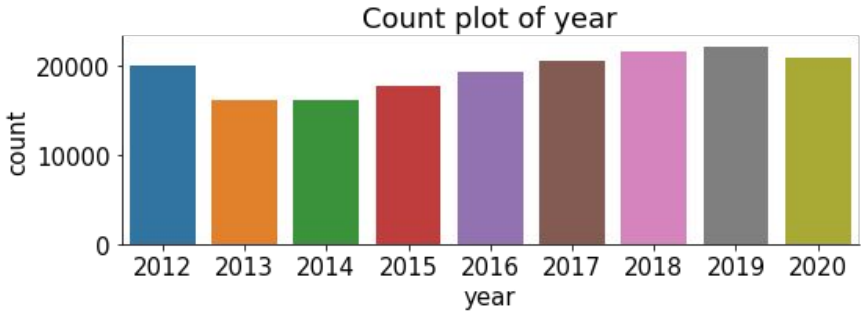
EDA

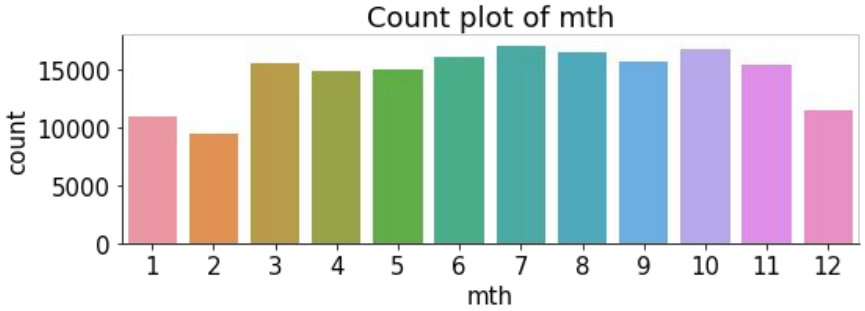
In 2013 and 2014, we see the effects of the property cooling measures on number of resale transactions.

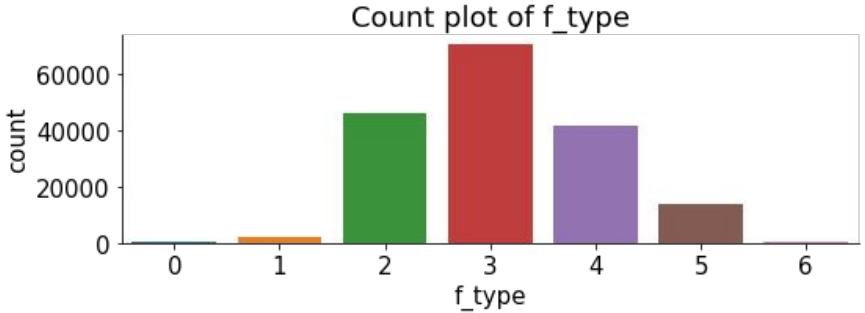
Transactions stabilize thereafter.

Transactions are lower during start and end of year

Most common flat types are 3, 4, and 5-room flats.





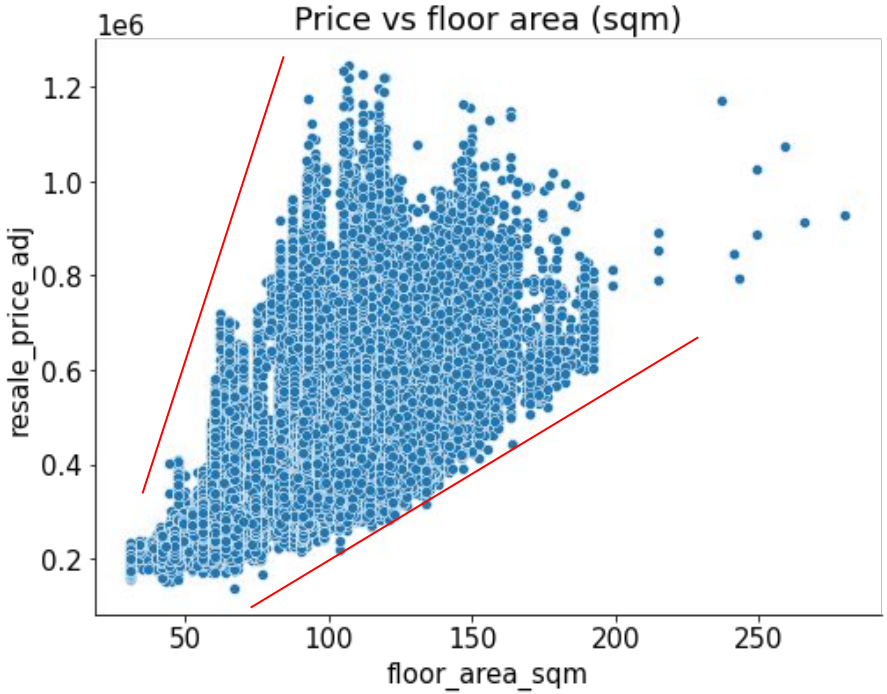


EDA

We see prices spread out as floor size increases.

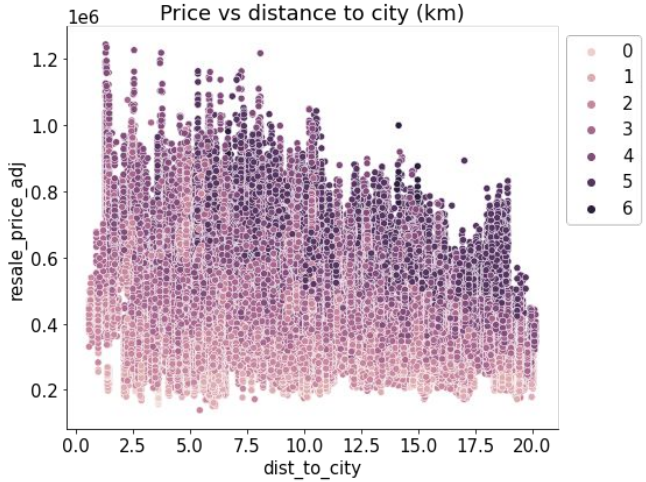
There are also some outliers with floor area >200sqm.

- Terraces and maisonettes that are likely 2 units combined into 1 to give the size.
- Dropped these units.



EDA

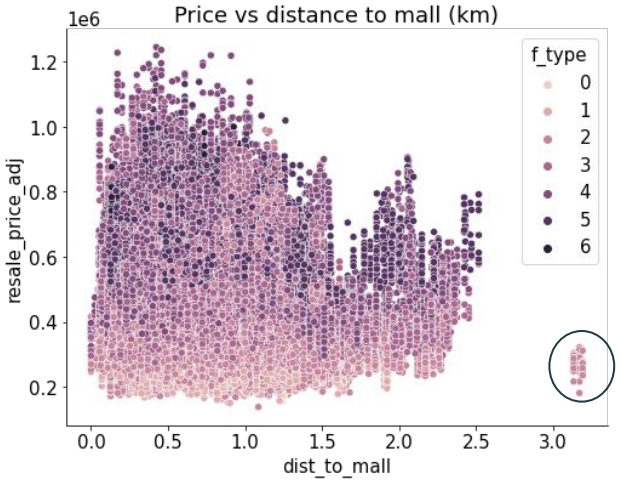
While maximum resale prices get higher as distance/time gets lower, minimum resale prices appear unaffected

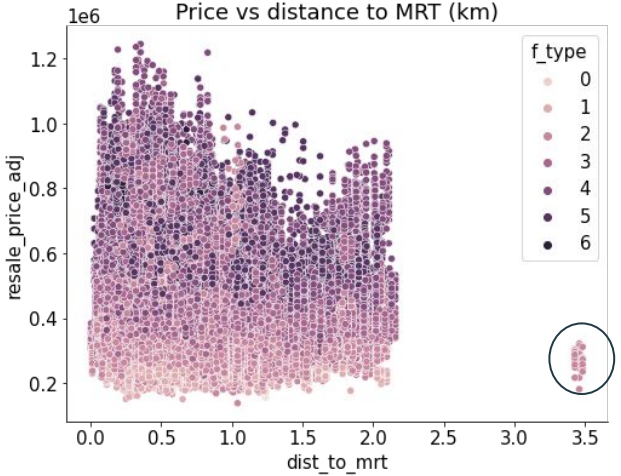


EDA

Similar distribution with distance to MRT and mall.

Outliers present very far to the right. What are they?





Changi Village!

Some old HDBs still exist here, but very far from everyone else.

Dropped these units from dataset.

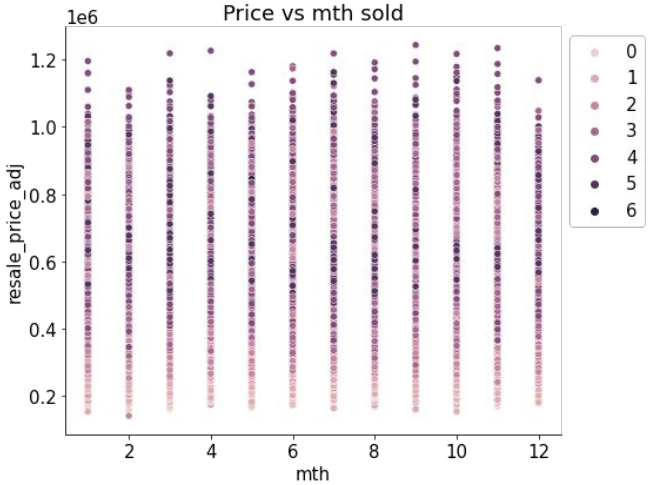
EDA

Maximum resale prices go up as years go by.

Minimum resale prices remain stable

- Housing remains affordable for majority.

No effect of month on price, dropped.



EDA

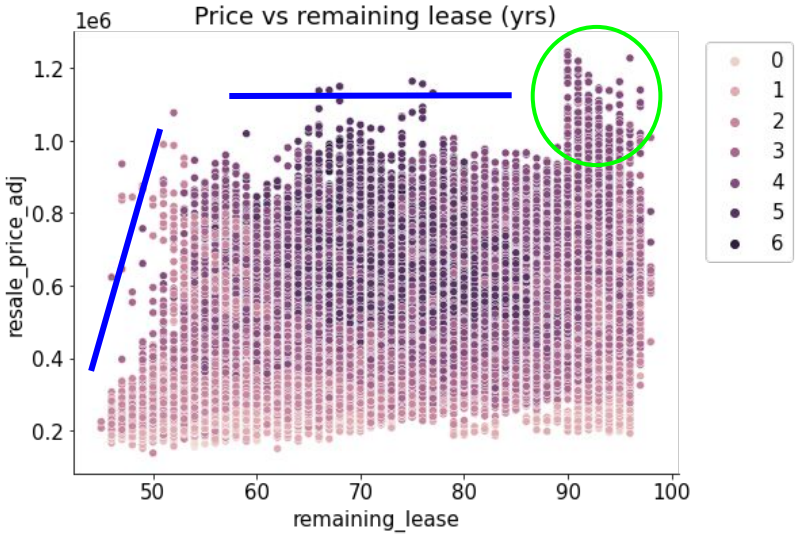
Between 60-90 years, price is quite stable.

Drops sharply below 60.

- Older houses risk lease expiry

Sudden jump in max resale price above 90 yrs

- DBSS, Pinnacle@Duxton flats being sold after Minimum Occupancy Period for profit.



EDA

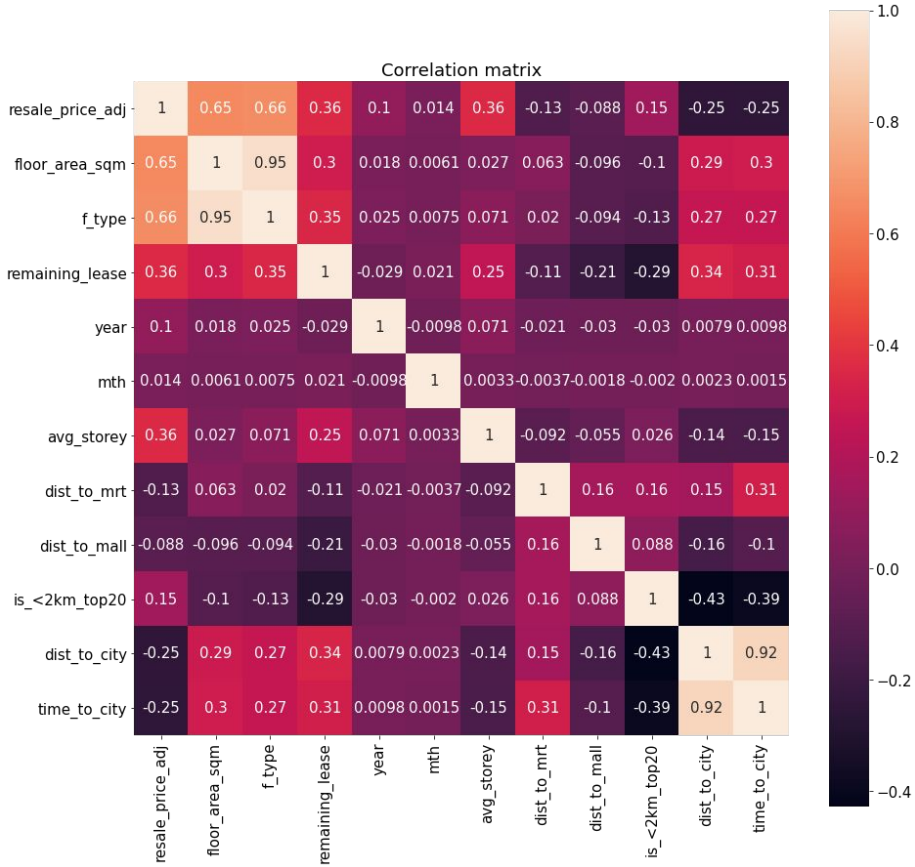
High correlation between flat type and floor area

- Dropped flat type

High correlation between dist_to_city and time_to_city

- Decided to drop dist_to_city

Only floor area has moderately high correlation with resale price.



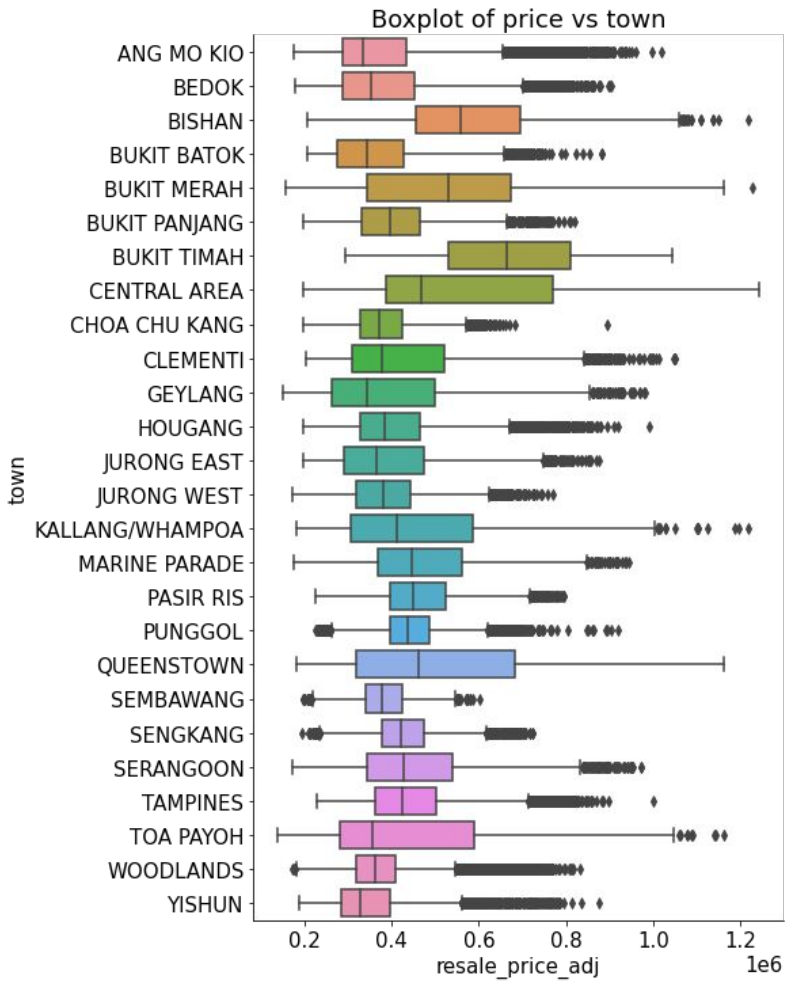
EDA

Some towns clearly more expensive than others

- Bishan, Bukit Timah, Bukit Merah

Some towns have smaller IQR, prices more consistent

- Sembawang, Woodlands, Choa Chu Kang
- These are quite far away from city centre, unlikely to command $1mil prices like other places.



Modeling

After checking for multicollinearity, these are our final predictors:

1. Floor area
2. Remaining lease
3. Year bought/sold
4. Unit storey
5. Distance to nearest MRT
6. Distance to nearest mall
7. Is near top20 primary school
8. Town (dummied)
9. Flat model (dummied)

Modeling

Multi-Linear Regression using statsmodel OLS.

Adjusted R2 = 0.861

Model is statistically significant

- Prob (F-statistic) = 0.00

Multicollinearity present due to dummied

categorical variables

Dep. Variable:	resale_price_adj	**R-squared:**	0.861
Model:	OLS	**Adj. R-squared:**	0.861
Method:	Least Squares	**F-statistic:**	2.117e+04
Date:	Sat, 19 Jun 2021	**Prob (F-statistic):**	0.00
Time:	02:17:25	**Log-Likelihood:**	-2.1361e+06
No. Observations:	174397	**AIC:**	4.272e+06
Df Residuals:	174345	**BIC:**	4.273e+06
Df Model:	51		
Covariance Type:	nonrobust		

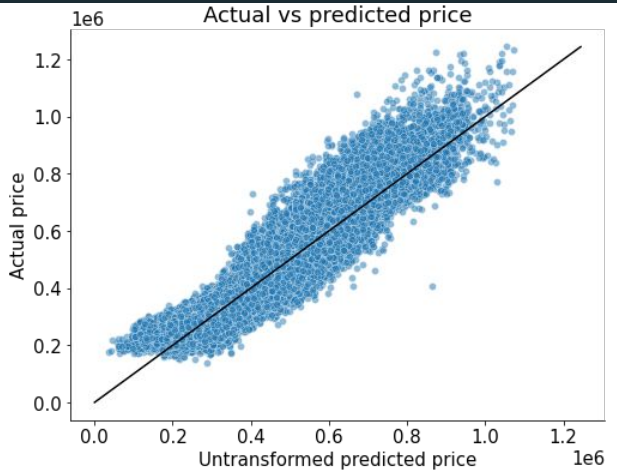
Modeling

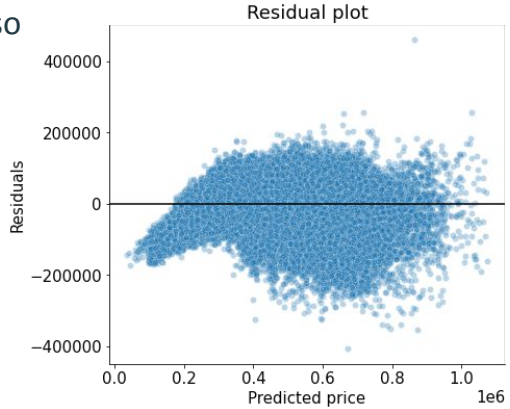
Predictions should be clustered around the diagonal line, but is not.

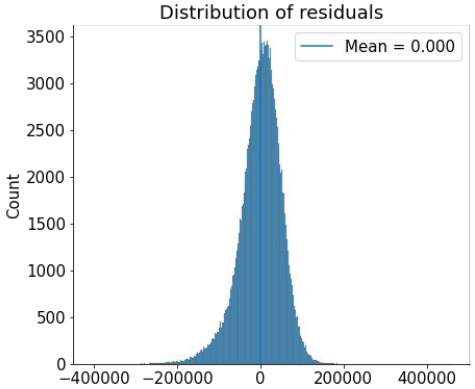
Residual plot display severe heteroscedasticity, also displays left skew.

RMSE = 50500

Solution: apply log transformation to resale price







Modeling

After logging target, this is our regression equation

$$\log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Adjusted R2 = 0.881

RMSE = 47492

Model is still statistically significant, and all individual predictors are now statistically significant at 95% confidence level

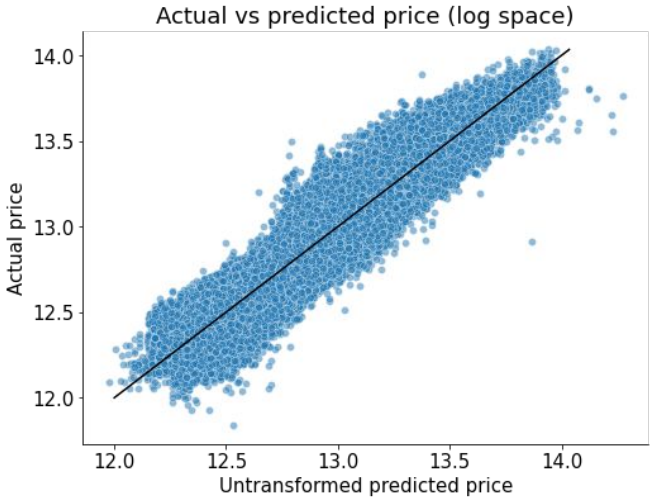
Dep. Variable:	resale_price_adj	**R-squared:**	0.882
Model:	OLS	**Adj. R-squared:**	0.881
Method:	Least Squares	**F-statistic:**	2.543e+04
Date:	Sat, 19 Jun 2021	**Prob (F-statistic):**	0.00
Time:	02:41:02	**Log-Likelihood:**	1.4781e+05
No. Observations:	174397	**AIC:**	-2.955e+05
Df Residuals:	174345	**BIC:**	-2.950e+05
Df Model:	51		
Covariance Type:	nonrobust		

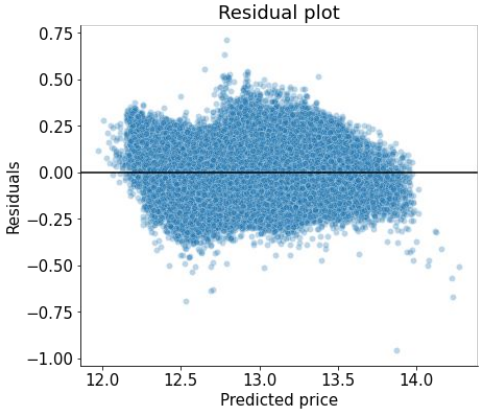
Modeling

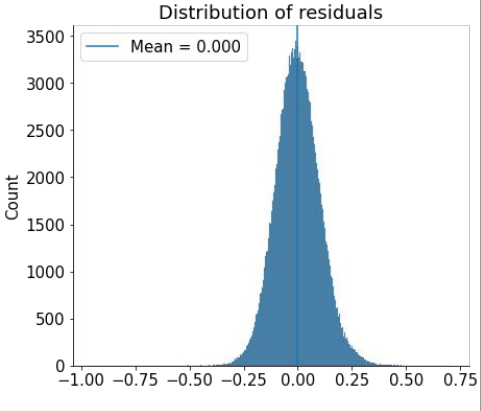
Predictions are now nicely clustered around actual price.

Residuals are largely homoscedastic, with a few outliers that we will examine later.

Residuals are normally distributed.







Interpretation of results

Our model shows a linear relationship between log(y) and our predictors.

Exponentiate the equation to return y to linear space

$$y = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}$$

Each unit increase of X1 increases y by e^β_1 ***times.***

β_X

	coef
const	12.8821
floor_area_sqm	0.0093
year	0.0085
avg_storey	0.0081
remaining_lease	0.0100
dist_to_mrt	-0.1216

Exponentiate →

e^{β_p}

const	393203.193
floor_area_sqm	1.009
year	1.009
avg_storey	1.008
remaining_lease	1.010
dist_to_mrt	0.886

Interpretation of results

When we say a unit increase of X1 increases y by e^{β_1} times, what is the base reference?

Base reference is when all predictors are 0, i.e. $y = e^{\beta_0}$, or $393,203.

This hypothetical unit, when all predictors=0, is a unit with these characteristics:

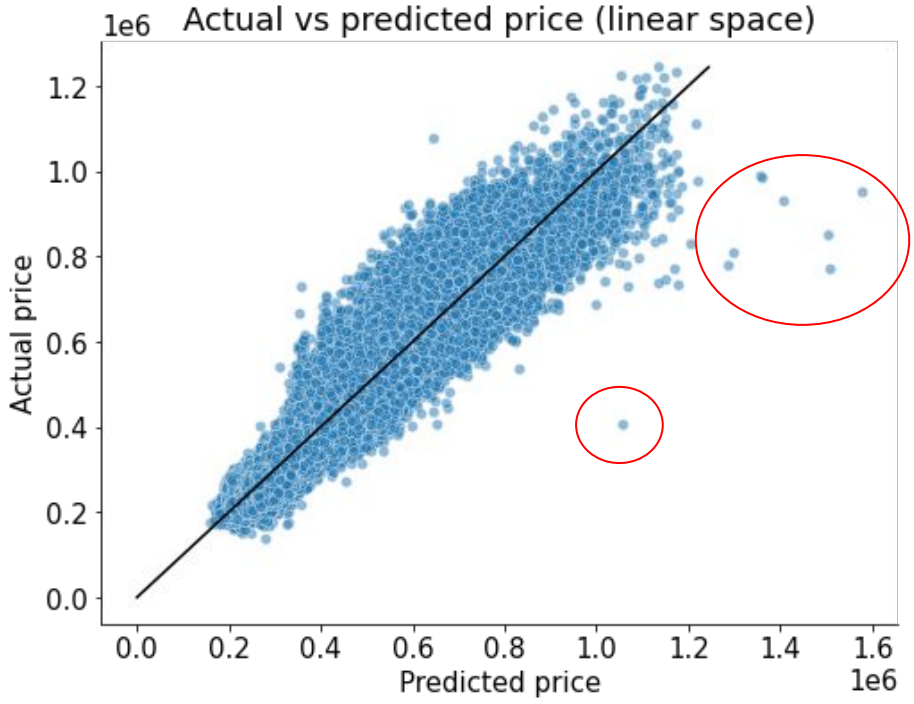
Town	F_type	Floor area	Remaining lease	Avg unit storey	Distance to MRT	Distance to mall	Near top20 pri sch	Year bought/sold
AMK	2-room	97.1 sqm	75 yrs	8th floor	0.66km	0.68km	No	2016

Interpretation of results

Examine predictions vs actual resale price back

in linear space.

Some outliers present.



Outliers

town	address	floor_area_sqm	dist_to_mrt	time_to_city	year	flat_model	f_type	avg_storey	remaining_lease	resale_price_adj	preds
bishan	240 BISHAN ST 22	156.000	1.127	38.050	2012	maisonette	5	3	79	732138.118	844732.817
bishan	247 BISHAN ST 22	146.000	1.222	37.033	2019	maisonette	5	2	72	777345.919	743959.383
bishan	443 SIN MING AVE	199.000	1.298	43.267	2013	maisonette	5	8	76	778521.411	1287096.356
bishan	445 SIN MING AVE	190.000	1.345	43.283	2016	maisonette	5	11	73	828878.499	1203376.828

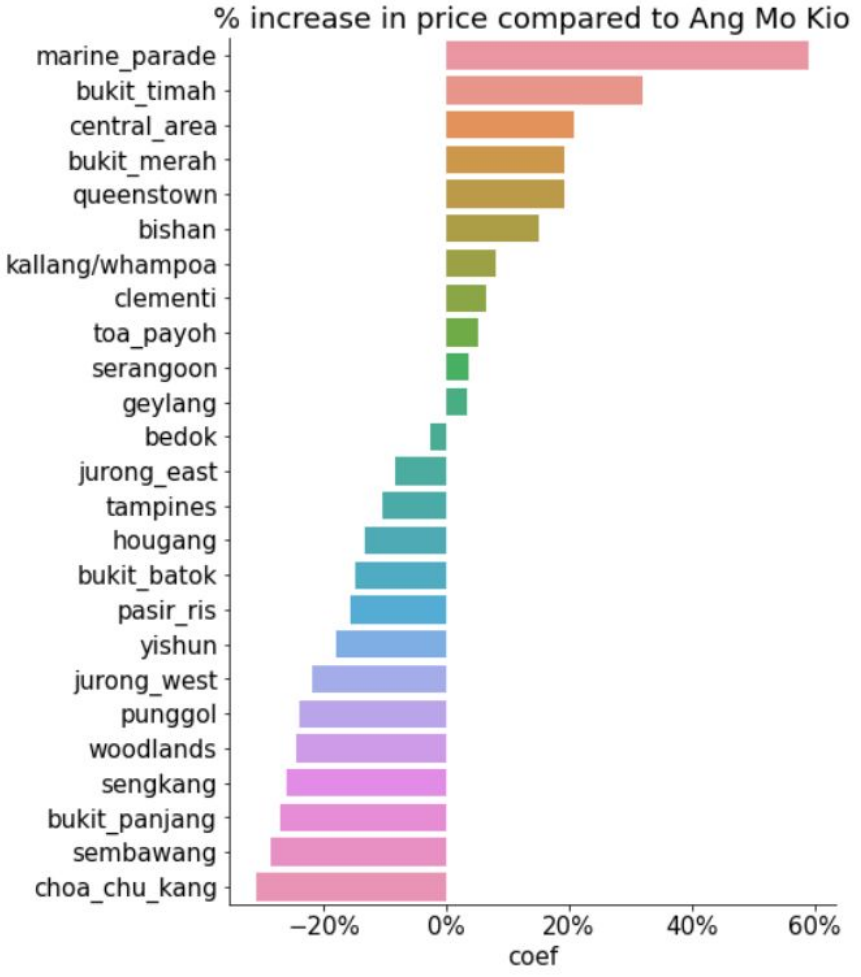
Outliers

	town	address	floor_area_sqm	dist_to_mrt	time_to_city	year	flat_model	f_type	avg_storey	remaining_lease	resale_price_adj	preds
	kallang/whampoa	53 JLN MA'MOR	111.000	1.066	35.183	2017	terrace	2	2	54	708375.836	784392.277
	kallang/whampoa	53 JLN MA'MOR	119.000	1.066	35.183	2019	terrace	2	2	52	817979.182	842718.967
	kallang/whampoa	53 JLN MA'MOR	108.000	1.066	35.183	2019	terrace	2	2	52	854804.609	760441.463
	kallang/whampoa	59 JLN MA'MOR	181.000	1.020	37.083	2017	terrace	2	2	54	770804.587	1509406.082

Interpretation of results

Different towns have different price premiums placed on them.

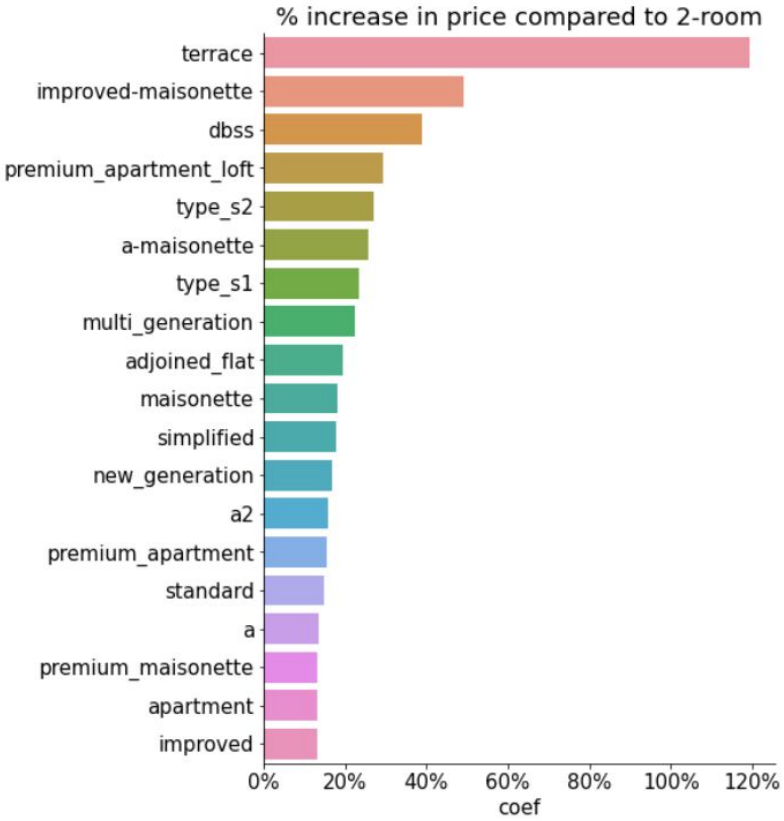
However, recall that we had to drop time_to_city due to multicollinearity. Some of the premium due to time is reflected here inside town. (Limitation)



Interpretation of results

Biggest price premium is placed on terraces, maisonettes,

DBSS, and the Pinnacle@Duxton units.



Conclusions

MLR model showed linear relationship between log of resale price and predictors.

Model is statistically significant, with adjusted R2 of 0.881, RMSE of 47492.

All predictor variables are significant at the 95% confidence level.

Conclusions

Factors with positive relationship to price:

- Floor area, remaining lease years, unit storey, year transacted, proximity to top20 primary school

Factors with negative relationship to price:

- Distance to nearest MRT, distance to nearest mall

Limitations

1. Town variable contains information about time to city.
    a. Need to isolate these effects out to better understand how town affects price
2. Flat model variable not consistent in characteristics it measures
    a. What is the difference between Model A vs Standard vs Apartment?
    b. Are flat models related to size, or secondary characteristics like terraces having backyards?
3. Model performs poorly on jumbo sized units (>180sqm)
    a. These units not much more expensive than much smaller units in same location

Recommendations

1. Engineer other variables to capture unique characteristics of town separate from time factors.
    a. Vibrancy of community, maturity of estate etc
2. Extract salient secondary characteristics of flat models that distinguish them from regular flats
    a. Maisonettes have 2 storeys, lofts have higher ceilings, terraces have backyards etc
3. Consider building separate model to account for extra large flats.
4. Interaction terms were not explored here, possible future avenue for improvements.