HDB Resale Prices

Trends and Predictions

Problem Statement

Using HDB resale flats data from https://data.gov.sg/dataset/resale-flat-prices,

We want to find out:

1. What are the primary factors driving resale house prices?
2. Predict the direction the prices will go for the next few years.

The Data

There are 4 HDB datasets:

1. 1990 - 1999, based on approval dates. 288144 rows, 10 columns.
2. 2000 - 2012, based on approval dates. 369651 rows, 10 columns.
3. 2012 - 2014, based on registration dates. 55203 rows, 10 columns.
4. 2015 - 2018, based on registration dates. 75800 rows, 11 columns. The extra column is ‘remaining_lease’, calculating the lease years remaining at point of purchase/registration.

The Data

After calculating ‘remaining_lease’ first 3 datasets, all datasets were merged.
Note: formula used for ‘remaining_lease is 99 - (year of sale - lease_commence_date).
Sample dataset below:

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
0	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	10 TO 12	31.0	IMPROVED	1977	86	9000.0
1	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	04 TO 06	31.0	IMPROVED	1977	86	6000.0
2	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	10 TO 12	31.0	IMPROVED	1977	86	8000.0
3	1990-01	ANG MO KIO	1 ROOM	309	ANG MO KIO AVE 1	07 TO 09	31.0	IMPROVED	1977	86	6000.0
4	1990-01	ANG MO KIO	3 ROOM	216	ANG MO KIO AVE 1	04 TO 06	73.0	NEW GENERATION	1976	85	47200.0
5	1990-01	ANG MO KIO	3 ROOM	211	ANG MO KIO AVE 3	01 TO 03	67.0	NEW GENERATION	1977	86	46000.0

The Data - Assumptions

All flats are 99-year lease

- Assume all flats are 99-year lease properties.
- Used for calculating remaining_lease feature.

What is the state of the economy at time of sale?

- Assume macroeconomic factors influence resale prices as well.
- GDP per capita chosen to represent state of the economy during time of sale. Data taken from: https://data.gov.sg/dataset/per-capita-gni-and-per-capita-gdp-at-current-market-prices-annual

Resale Prices at original values.

- Assume resale prices have not yet been adjusted for inflation.
- Prices will be adjusted to 2018 values, using data from: https://data.gov.sg/dataset/consumer-price-index-annual

The Data

Ultimately, I've chosen to focus on only the last 2 datasets (2012 - 2018), due to a number of reasons:

1. 2011 was PAP's worst election performance on record, which subsequently resulted in major shifts in housing policy. This effect is not represented in the data and may result in noise.
2. Technical reasons: My laptop is not able to model over 700,000+ rows of data effectively.

The Data

Final size of data set: 130883 rows x 13 columns

Features:

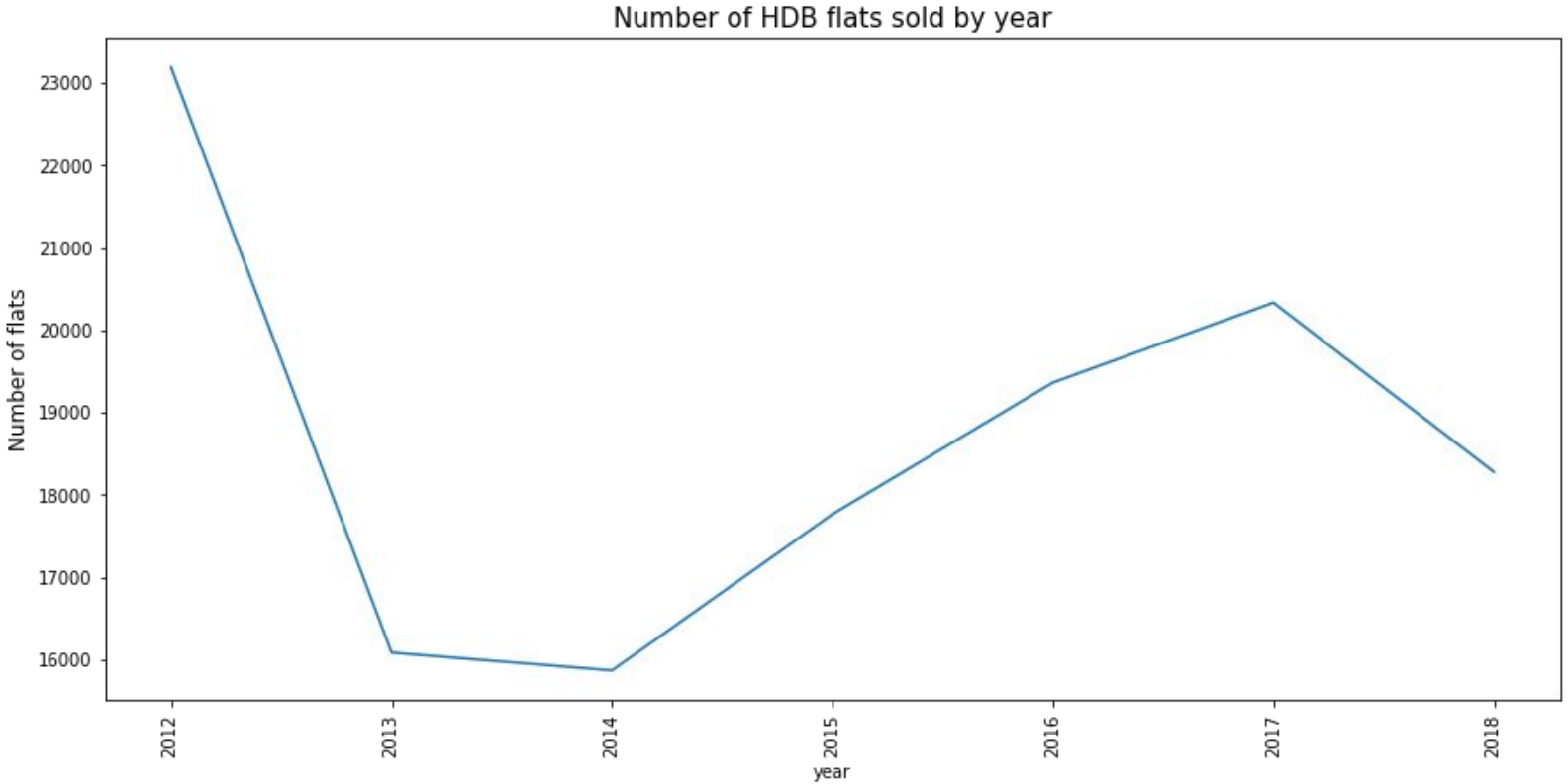
date | month | year | town | flat_type | block | street_name
storey_range | floor_area_sqm | flat_model | lease_commence_date
remaining_lease | resale_price | gdp_per_capita

EDA

Exploratory Data Analysis

1. General Resale Price Trends
2. Numerical Features
3. Categorical Features

EDA - General Resale Price Trends



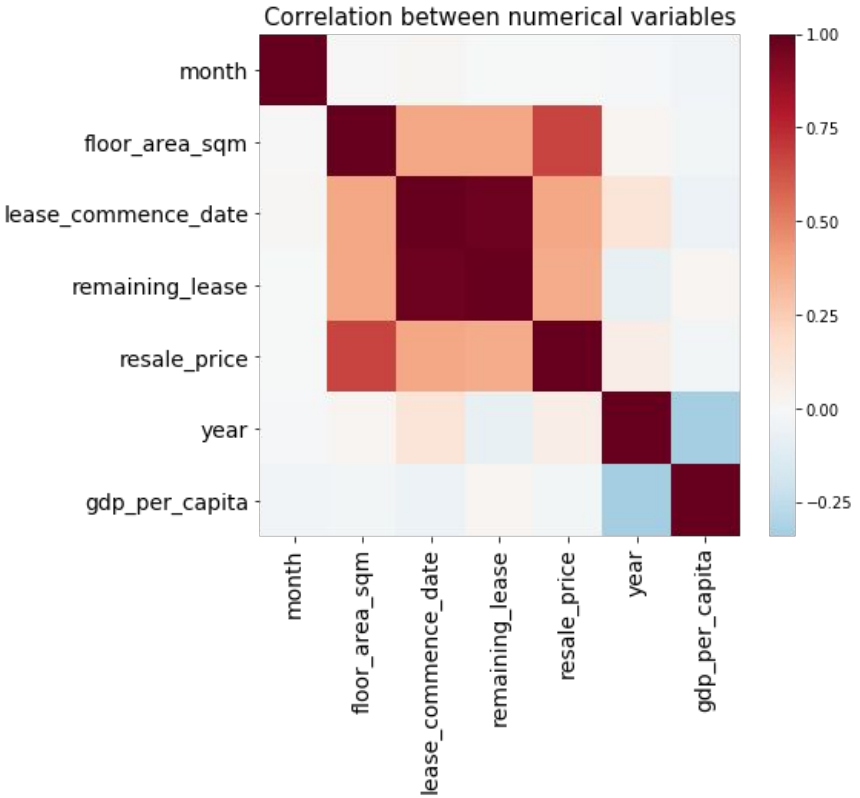
EDA - General Resale Price Trends

EDA - General Resale Price Trends

Insights:

1. Despite the huge dip in number of flats sold during 2013 and 2014, prices of flats remained relatively constant for the past 6 years.
2. Only the prices of the most expensive properties experienced an upwards trend. They also tend to be more volatile.

EDA - Numerical Features



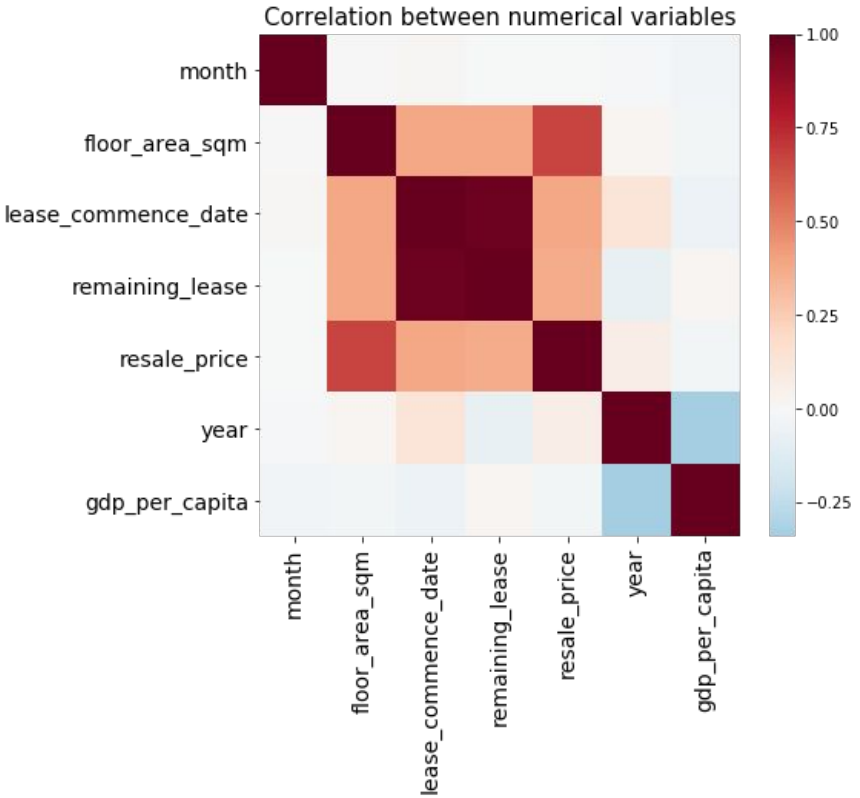
Strong positive correlation:

- remaining_lease vs lease_commence_date (expected, as the two are related).
- floor_area_sqm vs resale_price.

Weaker positive correlation:

- resale_price vs remaining_lease and lease_commence_date

EDA - Numerical Features



Almost no correlation:

- gdp_per_capita vs resale_price - My earlier assumption might be wrong!
- month vs resale_price - there may not be a seasonal trend based on month.

Negative correlation:

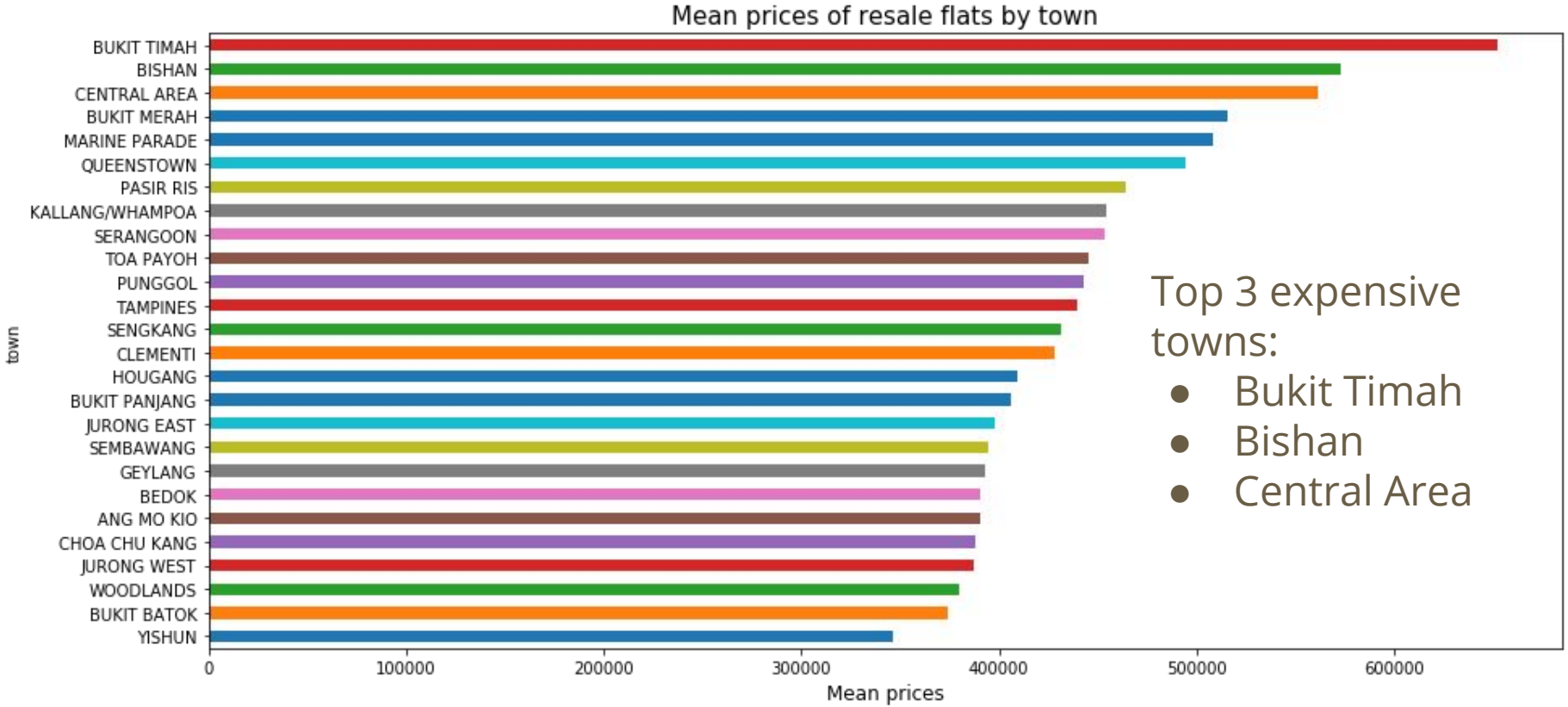
- gdp_per_capita vs year- economy was slowing down.

EDA - Numerical Features

Insights:

1. The bigger the flat, the more expensive it tends to be.
2. Time of the year and economy has very little impact on flat prices.
3. Newer houses tend to be more expensive. This corresponds to the number of years remaining on their lease.

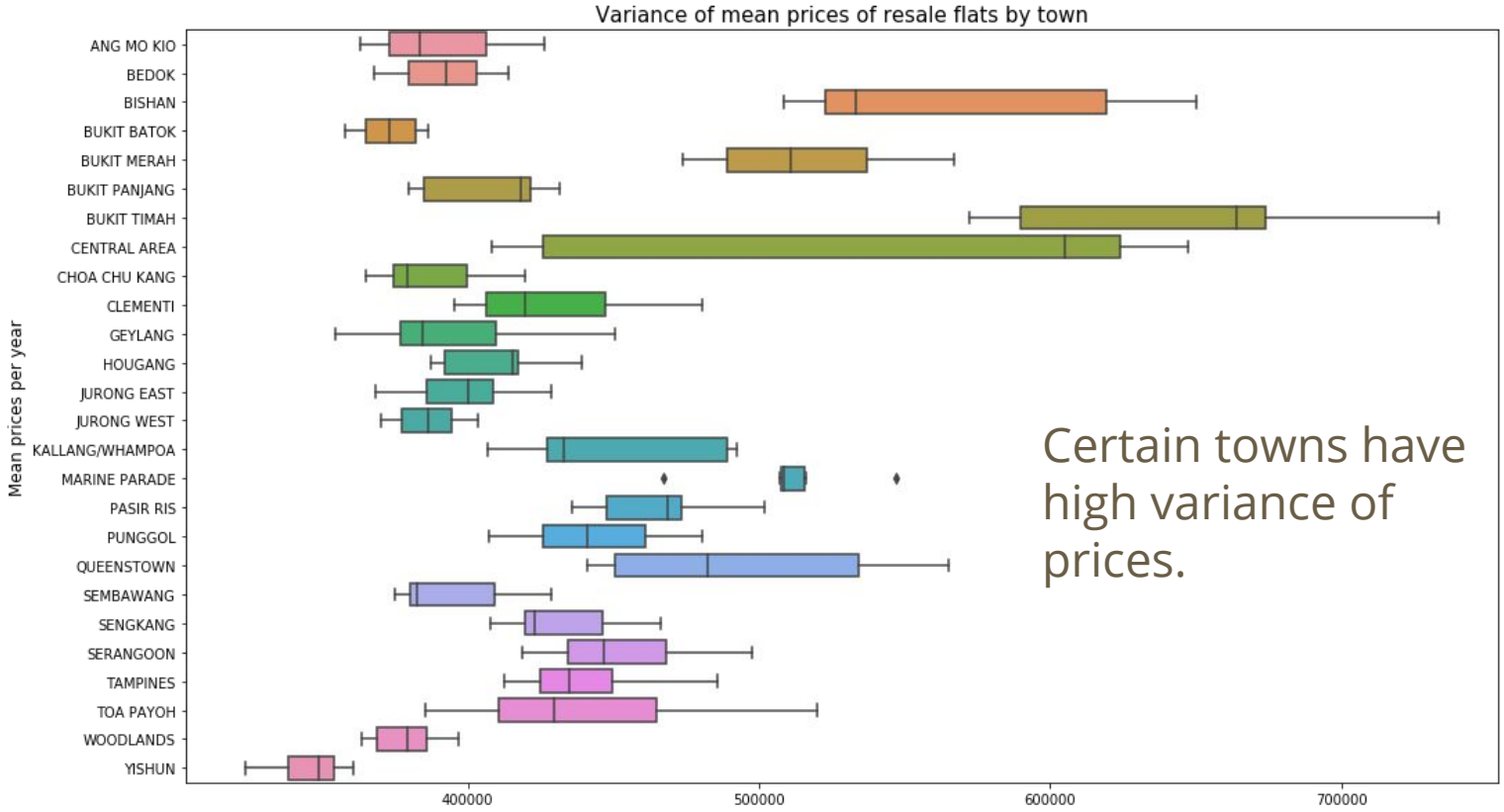
EDA - Categorical Features



Top 3 expensive towns:

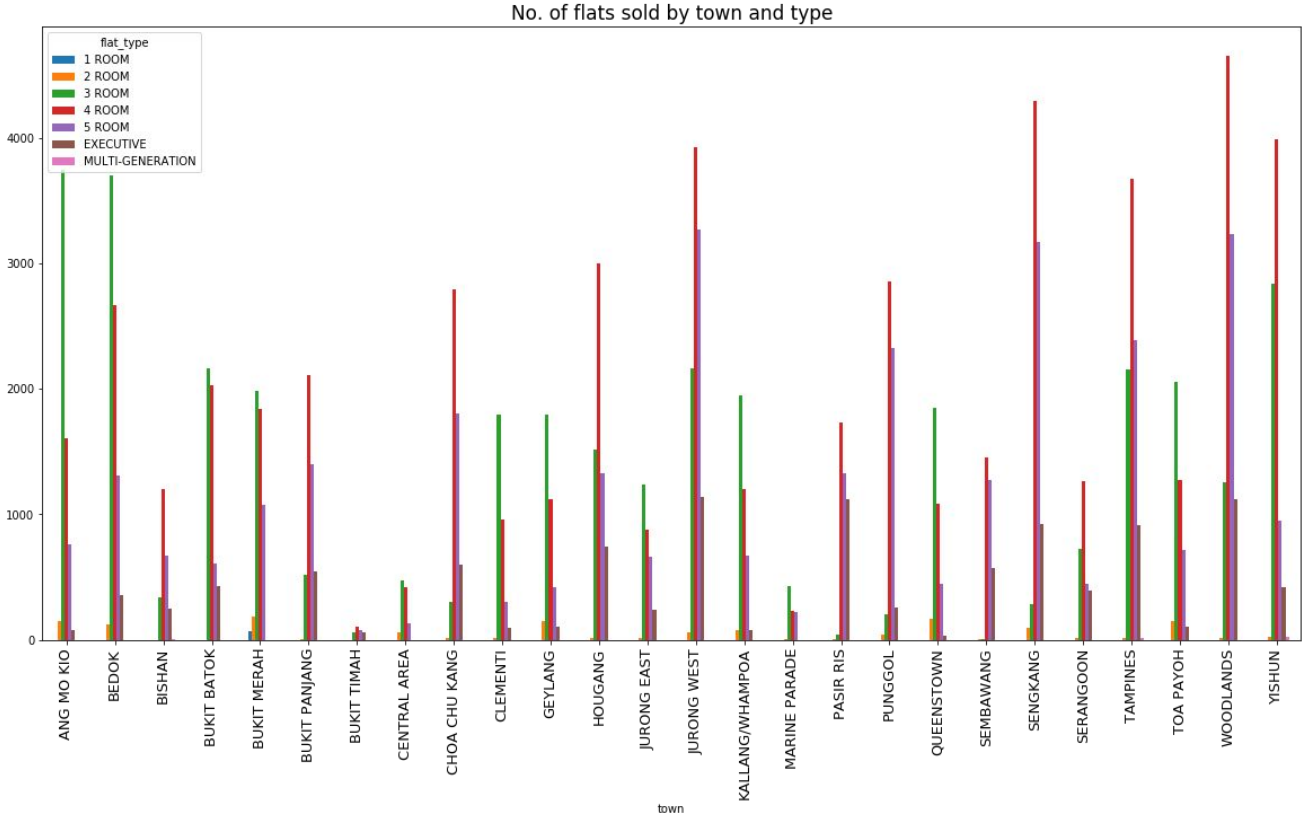
- Bukit Timah
- Bishan
- Central Area

EDA - Categorical Features



Certain towns have high variance of prices.

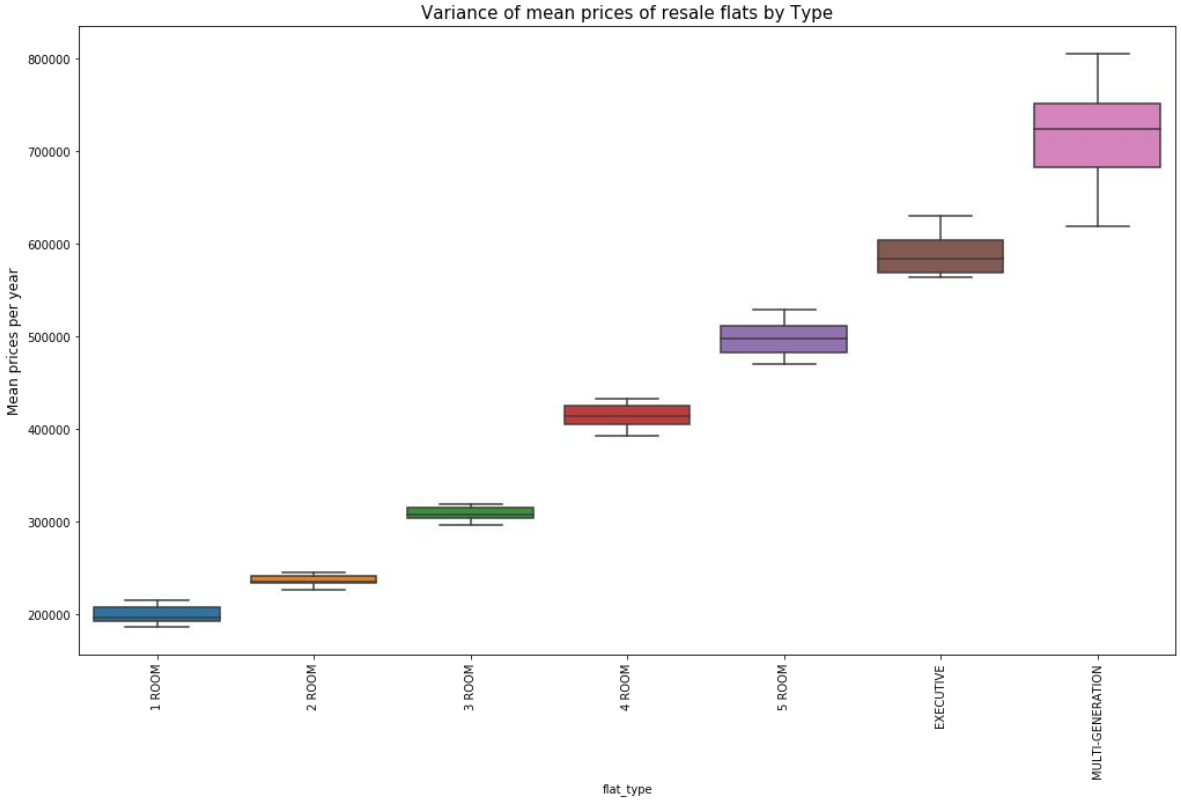
EDA - Categorical Features



Most popular flat type in cheaper towns tend to be 3 room flats.

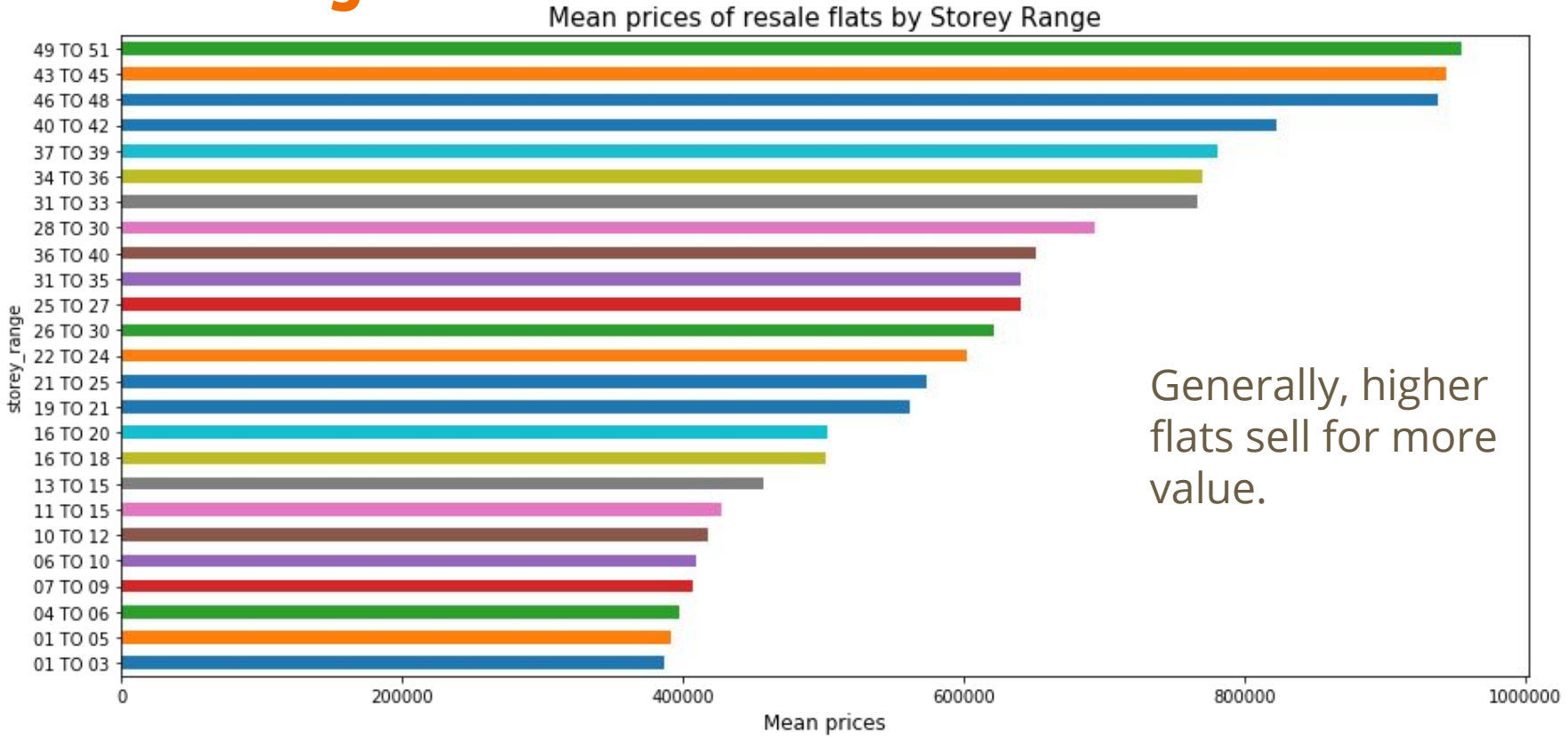
For more expensive towns, 4 room flats are more common.

EDA - Categorical Features



Range of prices between different flat types are more clearly defined, with little overlap.

EDA - Categorical Features



Generally, higher flats sell for more value.

EDA - Categorical Features

Insights:

1. Town, flat type and storey height have some effect on prices.
2. In the case of towns, certain towns have higher variance than others, meaning knowing the town alone would not guarantee a good prediction of prices.
3. Flat type, however, has clearly defined price ranges. Flat type might be a stronger indicator of resale price.

Modeling

1. Feature Engineering
2. Regression Models
3. Time Series Analysis

Feature Engineering

1. Features kept:

Month | Town | Flat Type | Storey range | Floor Area | Flat Model | Lease commence date | Remaining Lease | GDP per capita

2. Features dropped:

Date and Year - Because we're interested in predicting future sales, including year doesn't make sense.

Blk and Street Name - Without some way to quantify location (i.e, using GPS coordinates to measure distance from key landmarks), this information is useless.

Regression Models

Train-test-split with 25% test sample size used. R2 score used for scoring.

- R2 SCORE

BAD MODEL

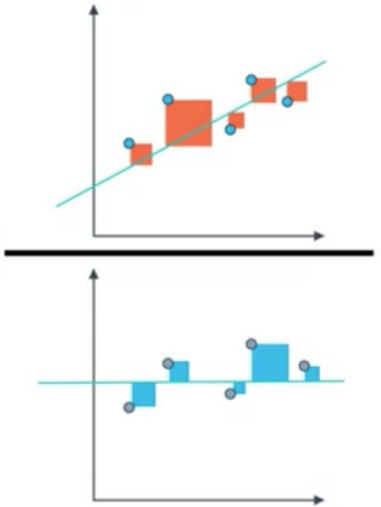
The errors should be similar.
R2 score should be close to 0.

GOOD MODEL

The mean squared error for the linear regression model should be a lot smaller than the mean squared error for the simple model.

R2 score should be close to 1.

$$R2 = 1 - \frac{}{}$$



Regression Models

By definition, a simple baseline model (all predicted values are mean values of resale price) will have an R2 score of 0. A perfect model will score a 1.

Elastic Net Regression

Linear Regression based approach

R2 score: 0.838.

Random Forest Regressor

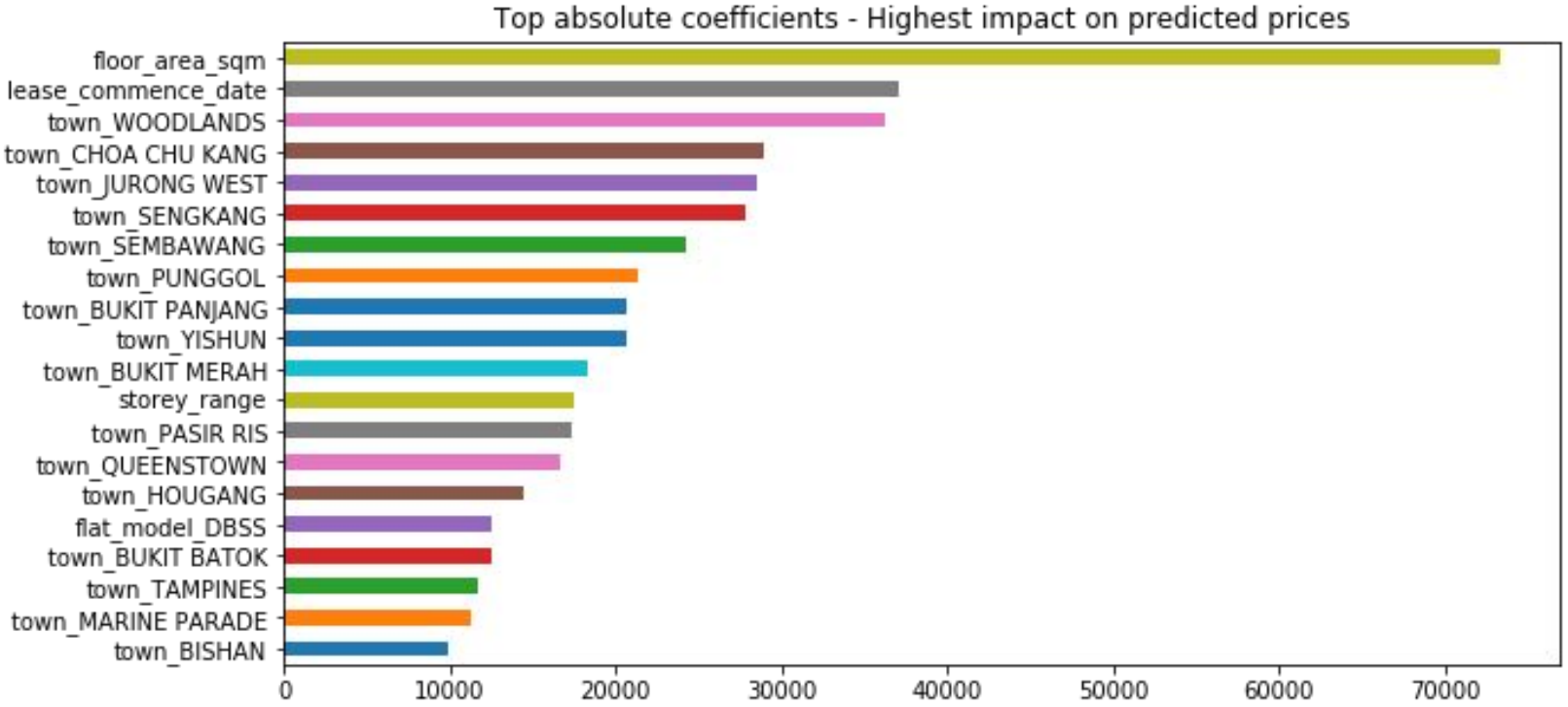
Tree-based approach

R2 score: 0.945

Random Forest Regressor performed better. This implies the chosen variables may not have a straightforward linear relationship with the resale price. A tree-based approach models it more accurately.

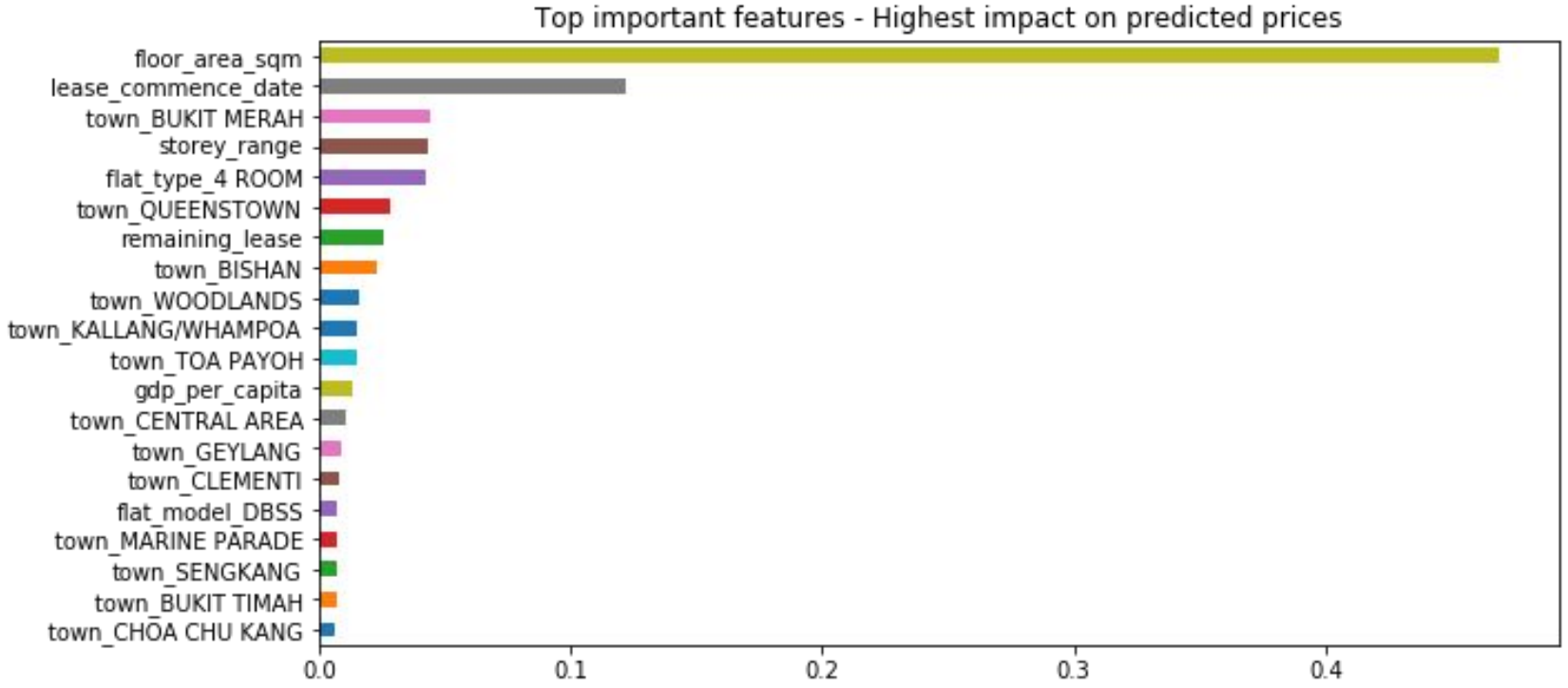
Regression Models

Elastic Net coefficients of the variables.



Regression Models

Random Forest Regressor feature importance of the variables.

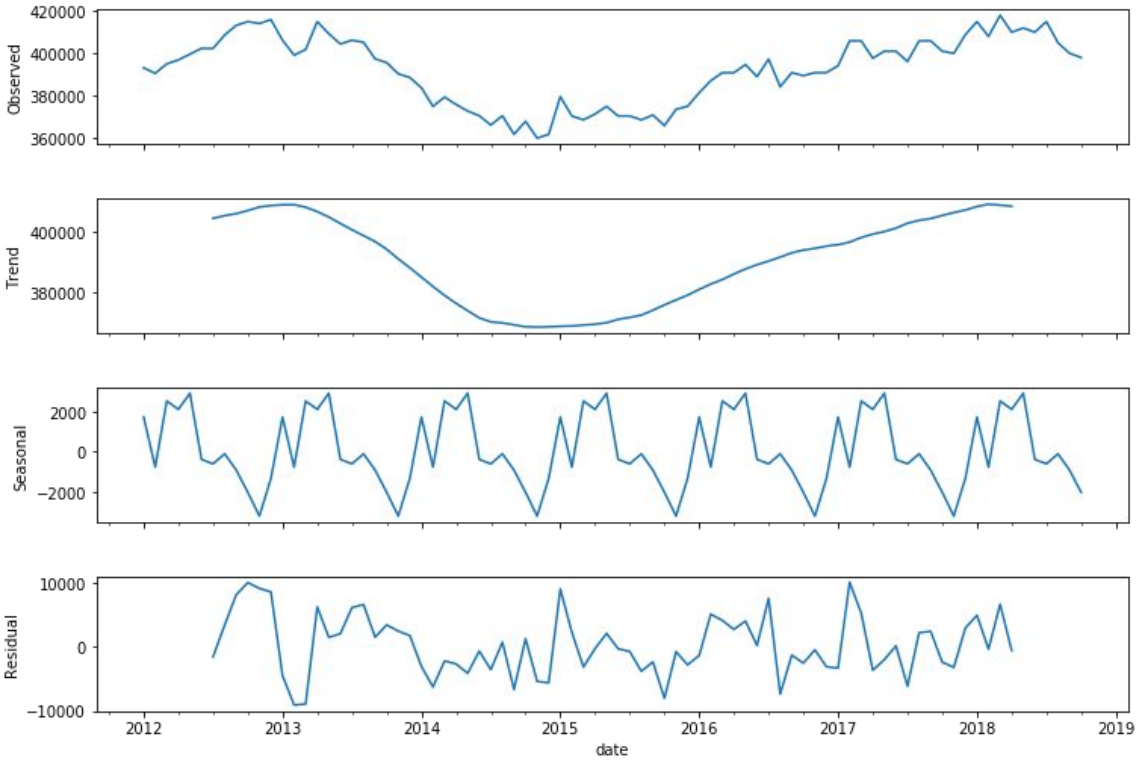


Regression Models

Insights:

1. Decision tree-based methods are better at predicting resale prices.
2. Both models agree on certain features that are important in predicting resale prices, namely floor area and lease commence date. Size of the flat is the most important feature. Storey range is important as well, indicating the higher the flat is located, the more expensive it is.
3. The models disagree on other features. Elastic Net seem to favor the different towns, while Random Forest places higher weightage on certain flat types and models. However, both models rank many towns pretty high on the feature importance list, indicating that town is also fairly important.

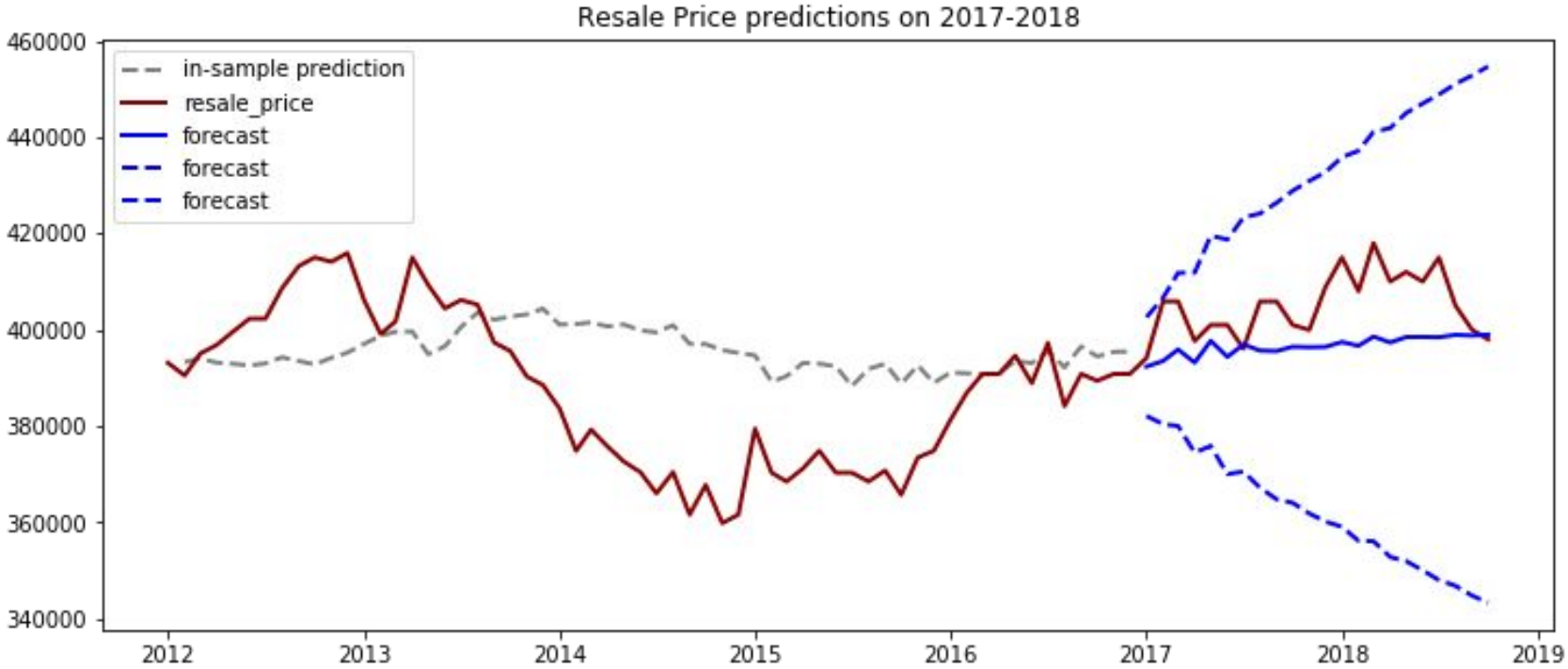
Time Series Analysis - Median Flat Prices



- No obvious long term trends, price is pretty constant between $360,000 and $420,000.
- Slight evidence of seasonality, however it only affects the price by ±$2,000.
- Unexplained residuals have a bigger effect.

Time Series Analysis - ARIMA Model

Model trained on 2012-2016 data, tested on 2017-2018 data.



Time Series Analysis - ARIMA Model

Insights:

1. ARIMA isn’t able to model resale prices accurately. It continued to predict a more or less constant price for the test set.

Conclusion

1. Summary of findings.
2. Limitations of the analysis

Summary of Findings

Insights:

1. Main driving factors behind resale prices are size and age of flats.
2. Secondary factor includes the town the flat is located in.
3. Compared to the above, flat model, type and economic factor don't matter as much.
4. Barring some fluctuation, prices are not expected to change much in the near future, with median prices continuing to be between $360,000 and $420,000.

Limitations of the Analysis

1. Block and Street Name features not used due to lack of geo-location data. If available, this can be used to engineer new features to model proximity to landmarks such as MRT stations and shopping mall, which could be important factors as well. As of now, current analysis treats the whole town in a homogeneous way.
2. ARIMA did not come up with a good prediction for time series analysis. I originally intended to use LSTM, but due to technical issues I was unable to get Tensorflow to work on my laptop. LSTM is expected to give better results.