# SC1015 MINI-PROJECT

PRESENTED BY : TEAM 9
K HARSHINI DEVI U2223242E
LOH JOO ANN U2221649A

# Table of contents

# 01

# Problem Definition

Help individuals who are looking to buy a home to make a informed decision.

# Purpose

Housing prices are increasing rapidly

Not aware of the pricing range

Easier to estimate the price

# 02

# Data Processing

# Data Cleaning

Ensures the data is accurate and reliable for analysis

**Data Cleaning**

In [8]:
```
data = pd.read_csv('price-range-of-hdb-flats-offered.csv')
```

In [10]:
```
missing_values = data.isnull().sum()
print(missing_values)
```

```
financial_year                  0
town                            0
room_type                       0
min_selling_price               0
max_selling_price               0
min_selling_price_less_ahg_shg  0
max_selling_price_less_ahg_shg  0
dtype: int64
```

In [11]:
```
clean_data = data.dropna()
```

In [14]:
```
clean_data = data.drop_duplicates()
```

In [22]:
```
#Checking for the duplicate values
print(data.duplicated().sum())
```

```
0
```

In [8]:
```
#Checking all the missing values
print(df.isnull().sum())
```

```
financial_year                  0
town                            0
min_selling_price               0
max_selling_price               0
min_selling_price_less_ahg_shg  0
max_selling_price_less_ahg_shg  0
room_type_2-room                0
room_type_2-room                0
room_type_3-room                0
room_type_4-room                0
room_type_5-room                0
dtype: int64
```

In [10]:
```
#Checking for the inconsistent values
print(df['town'].value_counts())
```

```
Punggol          47
Sengkang         39
Woodlands        35
Sembawang        31
Yishun           30
Choa Chu Kang    21
Jurong West      17
Bukit Panjang    17
Bukit Batok      16
Tengah           16
Hougang           4
Jurong East       4
Name: town, dtype: int64
```

In [16]:
```
clean_data.to_csv('cleaned_data.csv', index=False)
```

# Data Normalization

**Data Normalization**

In [19]:
```
data = pd.read_csv('price-range-of-hdb-flats-offered.csv')
columns_to_normalize = ['min_selling_price', 'max_selling_price', 'min_selling_price_less_ahg_shg', 'max_selling_price_less_ahg_s
scaler = StandardScaler()
data[columns_to_normalize] = scaler.fit_transform(data[columns_to_normalize])
data.to_csv('normalized_data.csv', index=False)
```

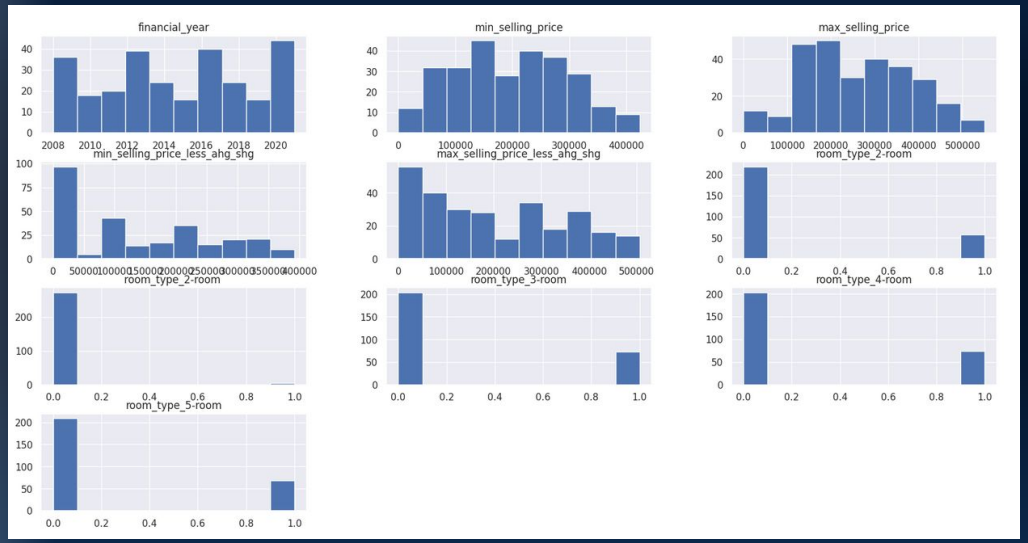
To ensure that different features have equal weight in the analysis and to reduce the effect of outliers.

# Data Transformation

Categorical variables such as "room_type" and "town"have been converted to numerical features

| max_selling_price | min_selling_price_less_ahg_shg | max_selling_price_less_ahg_shg | room_type_2-room | room_type_2-room | room_type_3-room | room_type_4-room | room_type_5-room |
|---|---|---|---|---|---|---|---|
| 107000 | 0 | 0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 211000 | 0 | 0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 327000 | 0 | 0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 428000 | 0 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 160000 | 0 | 0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |

# Data Visualization




Representing data in a graphical or pictorial format to help better understand and analyse the data

# 03

# Methodology

# Random Forest Algorithm

- Popular and powerful machine learning model that is used for regression and classification tasks.
- It constructs multiple decision trees and combines their predictions to make a final prediction.
- The main advantages of using Random Forest are that it is able to handle large datasets with high dimensionality, it is robust to noise and outliers, and it is resistant to overfitting.

1. Encode the categorical variable(s), if any
2. Split the data into training and testing sets
3. Create a Random Forest model
4. Train the Random Forest model
5. Evaluate the performance of the Random Forest model

04

# Experiment

- The performance of the Random Forest model is evaluated using mean squared error (MSE) and R-squared (R^2) metrics.
- MSE is a measure of the average squared difference between the predicted and actual values, and it is a commonly used metric for regression tasks.
- R^2 is a measure of how well the model fits the data compared to a baseline model that always predicts the mean value of the target variable.
- A higher R^2 value indicates a better fit of the model to the data.
- By comparing the performance metrics of the Random Forest model to a baseline model,  we can determine if the model is performing well or not.
- We can also compare the performance of different models to choose the best one for a given task.

# CONCLUSION

# What we learn.....

- Learned about the random forest model
- How to analyse a data and come up with a solution to solve the problem

# In Conclusion.....

- Potential home buyers are able to make a more informed decision on their purchase
- It can be seen that there have been an increase in prices of the flats over the years

# Thanks!

Please keep this slide for attribution