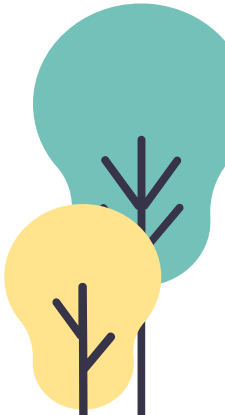
SC1015
Mini-Project

LAB GROUP SC17

Muhammad Hisham Bin Khairul Annuar (U2121992E)

Ong Li Han (U2021646C)

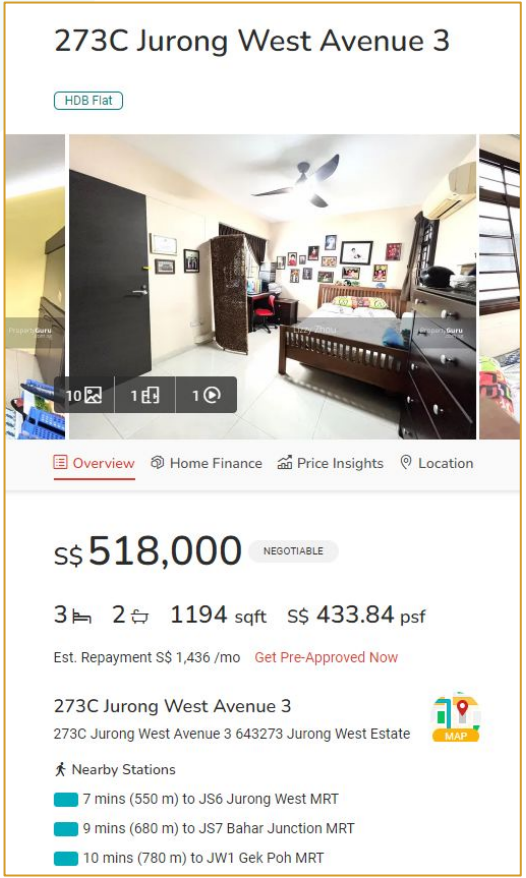
TABLE OF CONTENTS



PROBLEM

Should I buy this flat?

- Is the price of this flat listing reasonable?
- What should I look for if I want a cheaper flat?
- Does distance to nearby amenities really affect price?



Datasets

Data.gov.sg

- Resale Flat Prices
- Schools Directory

HealthHub Directory

- Hospital Addresses

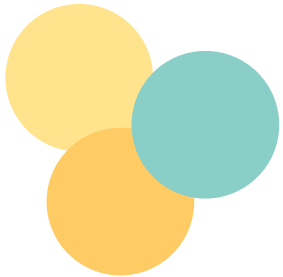
Open Source Repositories

- MRT Station Locations

OneMap Search API

- Convert local address to coordinates

CALCULATING DISTANCE TO THE NEAREST AMENITY



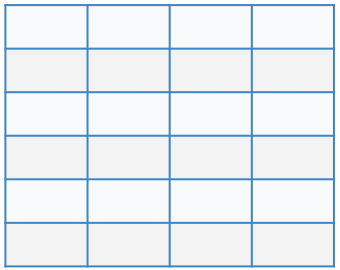
Raw Data

- 120,000+ flat listings with 9,000 unique addresses
- 550+ MRT stations, schools, hospitals
- Location given in (Lat, Long)

Haversine Distance

- Distance between two sets of coordinates
- Assumes Earth is a perfect sphere

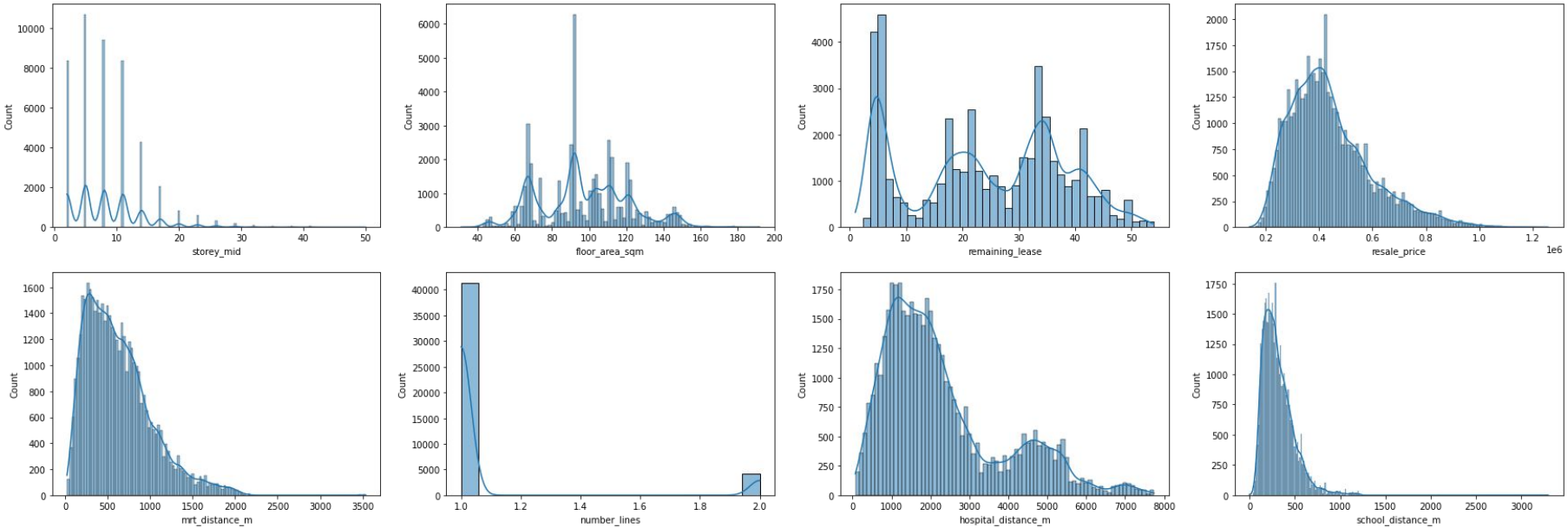
For each unique flat location, calculate the distance to each amenity, and return the one with the shortest distance



Formatted Data

- Each row contains a unique flat listing

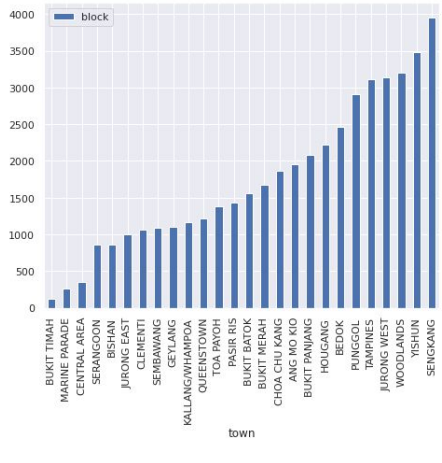
DISTRIBUTIONS (Numerical)

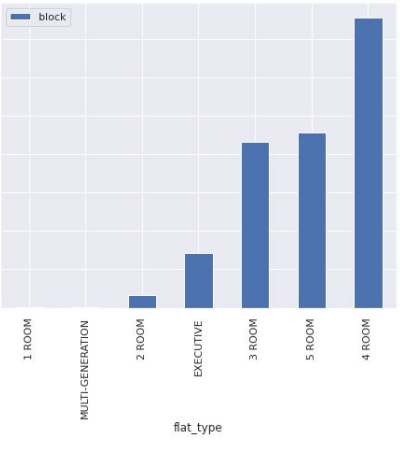


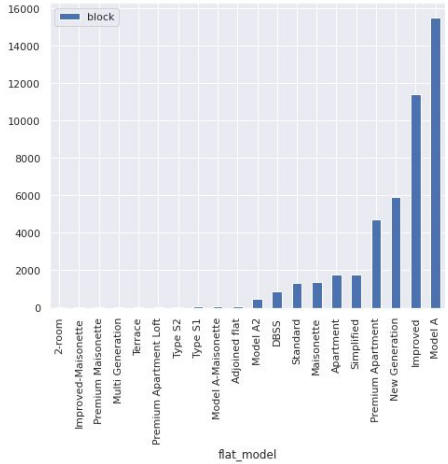
Observations:

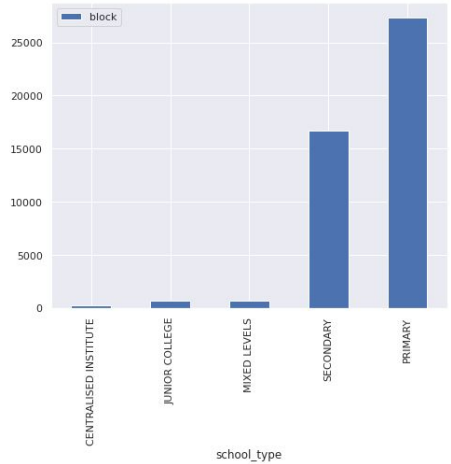
- Some variables such as resale price and distance to closest MRT, largely follow the normal distribution
- Others like floor area and remaining lease seem more random in their distribution
- Almost all flats are within 1km of their closest school, but the distances to the closest hospital are much larger

DISTRIBUTIONS (Categorical)









Observations:

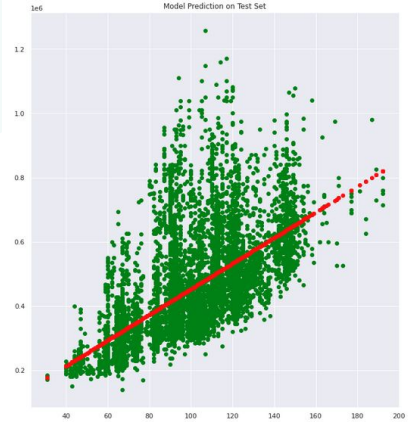
- Our classes in our categorical features are rather unevenly distributed.
- For example, there is a very low proportion of resale flats in the Bukit Timah town, since that area is known for private housing estates
- There seems to be some collinearity between flat_model and flat_type; it is likely that a flat of a certain model could only belong to a few particular types

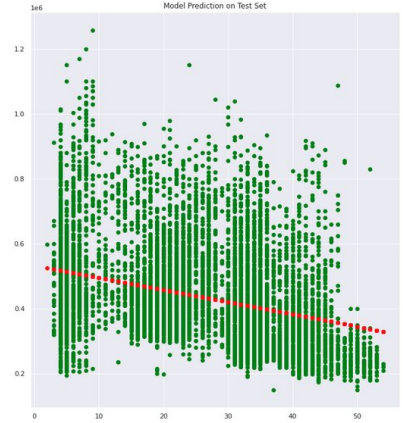
ML & DATA ANALYSIS

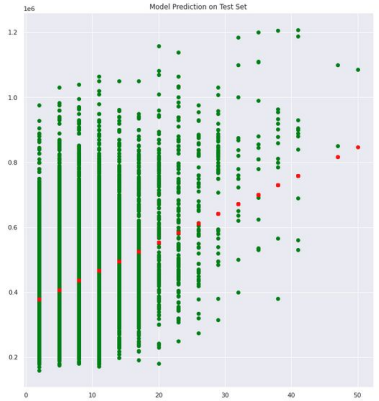
Q1) How accurately can we predict the price of a flat?

Univariate Linear Regression

- 80/20 Train/test split against floor area, remaining lease, and level of the flat
- Floor area's Train set had the highest R^2 of 0.395; it also had the highest prediction accuracy, with RMSE of 120214
- Other 2 factors had far lower R^2 (~0.1) and far higher root MSE.
- In conclusion, we can use Floor Area to predict resale price with decent accuracy







Linear Regression with All Features

- Using SKLearn pipeline and gradient boosting regression to account for all features together, the RMSE of the model further reduces to 42808
- Thus, using all the features together allows us to predict resale price the most accurately

Linear Regression

```
linr F1 score: 0.889
linr RMSE: 51809.541
linr MAE: 39312.175
```

Gradient Boosting Regression

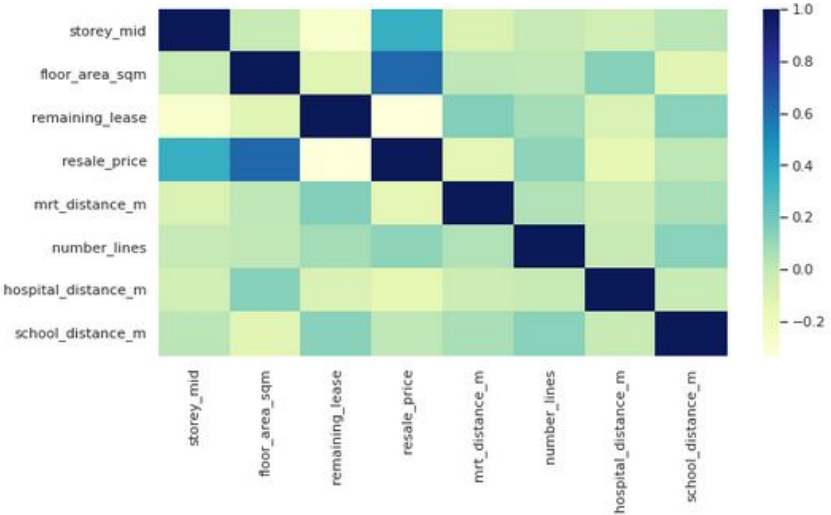
```
gbr F1 score: 0.924
gbr RMSE: 42808.130
gbr MAE: 31534.310
```

ML & DATA ANALYSIS

Q2) What features of a flat best predicts its resale price?

Correlation Matrix

- Correlation matrix for all continuous variables
- **2 variables** stand out for relationship against resale price:
  - storey_mid (0.362891), and
  - floor_area_sqm (0.629002)
- Thus, these are the best continuous predictors for resale price

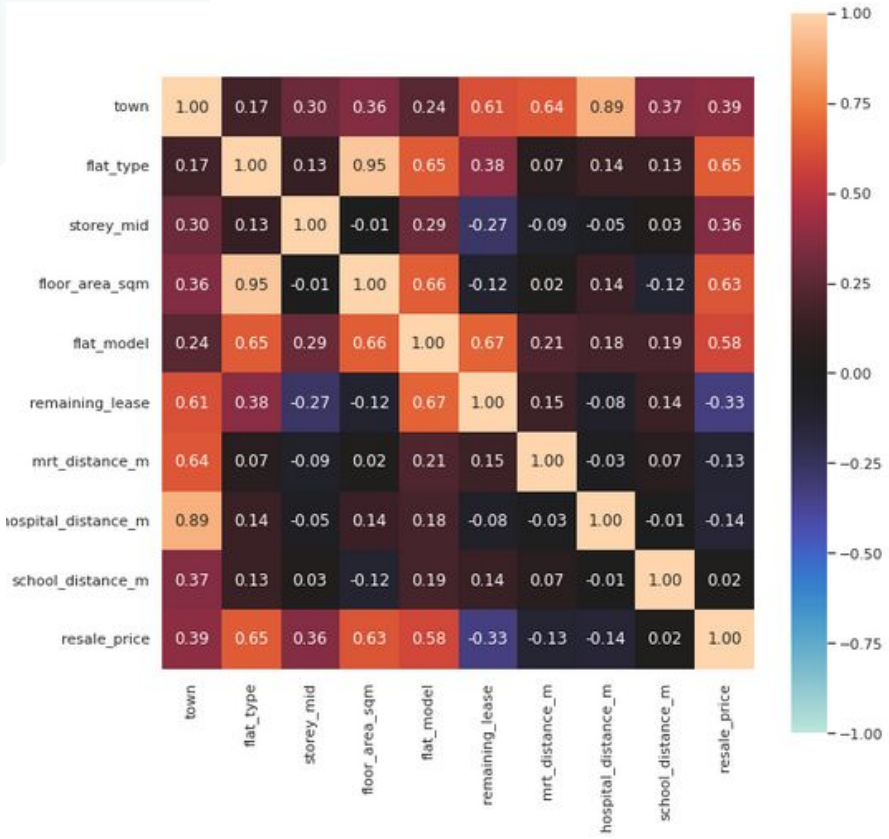


Including Categorical Variables

- Next we create a correlation matrix including categorical variables flat_type, flat_model, and town
- Note: Correlation ratio here = weighted variance of mean of each category divided by variance of all samples

- The most important categorical variable in determining resale price is flat_type (followed by flat_model), since it has the highest coefficient of 0.65

- In fact, flat_type has a slightly higher coefficient than floor_area_sqm, indicating an even stronger relationship



Using Feature Importance

- From the linear regression model, we can obtain coef_ to measure each variable's importance in determining resale price

- We can see that remaining_lease has the highest feature importance, closely followed by floor_area_sqm

- Town has the highest feature importance for a categorical variable

original_feature	importance
remaining_lease	124526.970142
floor_area_sqm	114291.071115
town	97104.583485
flat_model	85685.667667
mrt_distance_m	36252.022160
storey_mid	27223.322452
flat_type	22641.366861
hospital_distance_m	10065.767839
school_distance_m	4546.143948

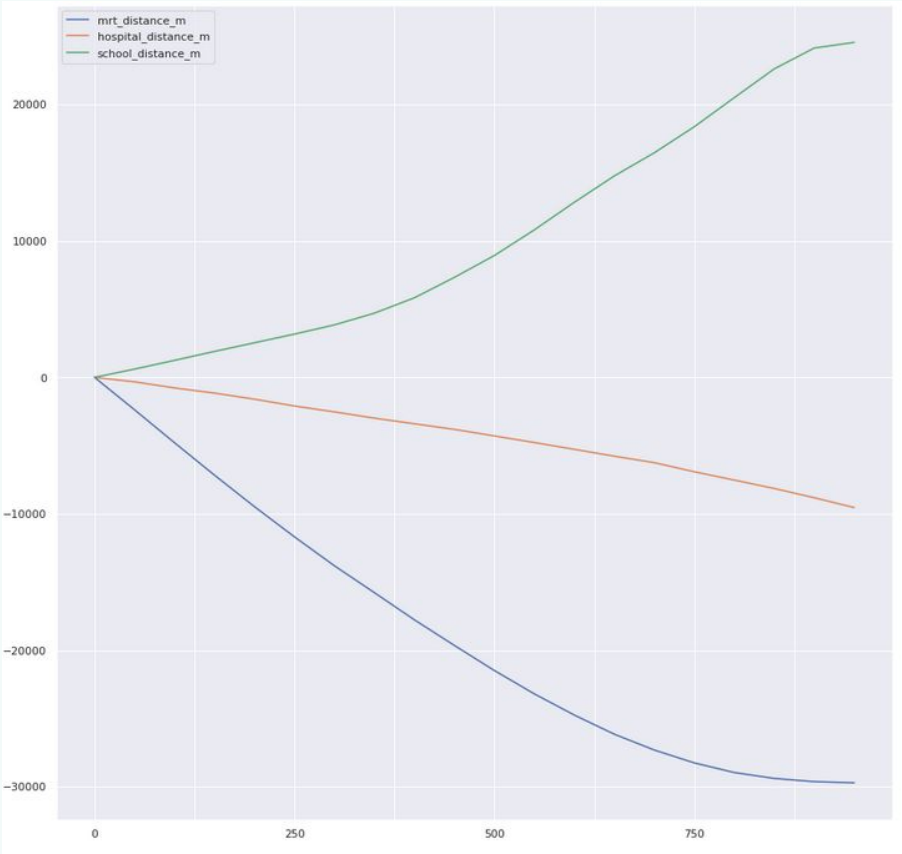
ML & DATA ANALYSIS

Q3) Which nearby amenity has the most impact on price?

Reducing Effects of Other Variables

- It seems that amenities play an insignificant effect on resale prices; however, it could be that the effect is simply dwarfed by that of other variables.

- To observe the effects of amenity distance on price, we made batch predictions on our entire dataset, using the same values for each feature, except that we vary the distances to a specific amenity type.
- Illustration:
    - Price_Base = Predict(Flat characteristics, X metres to MRT)
    - Price_+50m = Predict(Flat characteristics, X+50 metres to MRT)
    - Price_+100m = Predict(Flat characteristics, X+100 metres to MRT)



- **As distance to the closest MRT / hospital increases, price of the flat decreases**
  - Rate of decrease is stronger for MRT

- However , **the inverse is true** for schools
  - Could be because flats that are typically further away from schools are larger / more premium
  - Flats that are further from schools tend to be nearer to MRTs and other important amenities

Outcome

1) We were able to accurately predict resale price using linear regression on several features together.

2) We found that **floor area** was the most correlated variable in predicting resale price, and the 2nd most important feature; **remaining lease** seemed to have the highest feature importance despite having lower correlation to resale price.

3) By making use of our model to analyse the effects on resale price, of varying distances to each amenity, we can infer that flats which are nearer to certain amenities (ie. MRT, followed by hospital) tend to be sold at higher prices

INSIGHTS

DATA AGGREGATION

Learnt to reduce dataset size by cleaning/removing unnecessary data + organize in a useful manner

OPEN SOURCE API USAGE

Gathering further information using APIs and data we already have

CATEGORICAL ENCODING

Increase the accuracy of linear regression by accounting for categorical variables together with numeric variables

FEATURE IMPORTANCE

Utilised feature importance and permutation importance to examine how useful a feature is in determining the response variable

THANK YOU

LAB GROUP SC17

Muhammad Hisham Bin Khairul Annuar

Ong Li Han