# Residential building price prediction

DSA5101

| | | | |
|---|---|---|---|
| Cai Yusen | A0251268A | Liu Boyu | A0177847J |
| Deng Yiyang | A0251343N | Liu Yu | A0177906R |

NUS
National University
of Singapore

# Content

1 Overview of the Business Problem

2 Data Analysis

3 Sale Price Prediction

4 Recommendations

# Overview of the Business Problem

# Problem Statement

**Housing market** is tightly connected with the **state of economy** in general. For example, the precursor to the recent global economic crisis was the bust in the real estate market. **Predicting building sales price** is a crucial step for **construction companies and investors** in supporting budget assignment, finding property financing stratagems and deciding appropriate arrangements. However, the stakeholders involved might be unaware of the **statistical techniques** available to predict the building sales price, which could facilitate their investment decision.

This presentation making use of **data science analysis techniques** to find out the **driving factors of building sales price.** Through understanding **impacts and correlation** of different factors, **building models** to predict the sales price, the necessary of the construction investment can be determined.

# 01 Overview of the Business Problem

## Stakeholders

Our stakeholders are **construction companies and investors.**

The purpose is to build a model that can be used by stakeholders to gauge the sale market before they start a new construction and decide whether to build.

# Summary of Project

- **Methods**

The provided data set includes several factors related to the single-family residential apartments in Tehran, Iran. Through analyzing the data to understand the drivers and their impacts on sales price, a linear regression model is built to predict the housing price.

- **Deliverables**

We managed to build a regression model with high interpretability and prediction accuracy which can help construction companies to make wise decisions.

# Business Benefits

Stakeholders could understand the feasibility of the construction project and reduce the risk of investments.

- ✓ In a down market, the builder may decide to postpone the beginning of construction.
- ✓ In an up market, the builder may prefer to sell the units at the time of completion instead of the presale to maximize the profit.

A construction company with multiple projects in different regions can use the tool for resource scheduling and allocation.

- ✓ For example, they can allocate large cranes to localities with a better price prospect.

# Summary of Project

Team capabilities:

- Business analyst: Deng Yiyang, Cai Yusen, Liu Yu
- Data scientist: Cai Yusen, Deng Yiyang, Liu Boyu
- Code Reviewer: Cai Yusen, Deng Yiyang

# Data Analysis

# Basic Information of the Dataset

The dataset has 372 samples and 29 variables.

No null and duplicate values.

The variables can be divided into two parts below:

- Physical and financial factors about the single-family residential apartments in Tehran, Iran.
- Economic indicators and indices.

# Variables Type

**Output:** Actual Sales Prices (V.9)

**Unordered categorical variables:**

Project locality (V.1), Type of residential building (V.10)

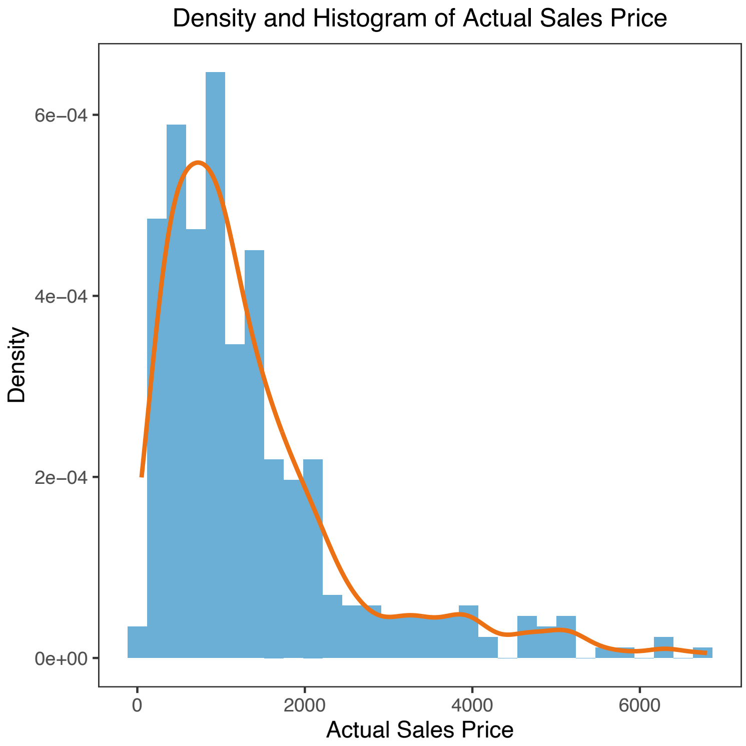
**Ordinal categorical variable:**

The interest rate for loan in a time resolution (V.20)

The interest rate is a numerical variable based on its meaning. However, since it only has 4 values, we consider it as an ordinal categorical variable.

**Numerical variables:** Others

# Actual Sales Price



| | |
|---|---|
| mean | 1387.4 |
| std | 1206.1 |
| min | 50 |
| 25% | 577.5 |
| 50% | 1000 |
| 75% | 1700 |
| max | 6800 |

# Actual Sales Price

The distribution of actual sales price is **unsymmetric**. From the histogram and the table, we can find that prices are concentrated in the range from 50 to 2000.

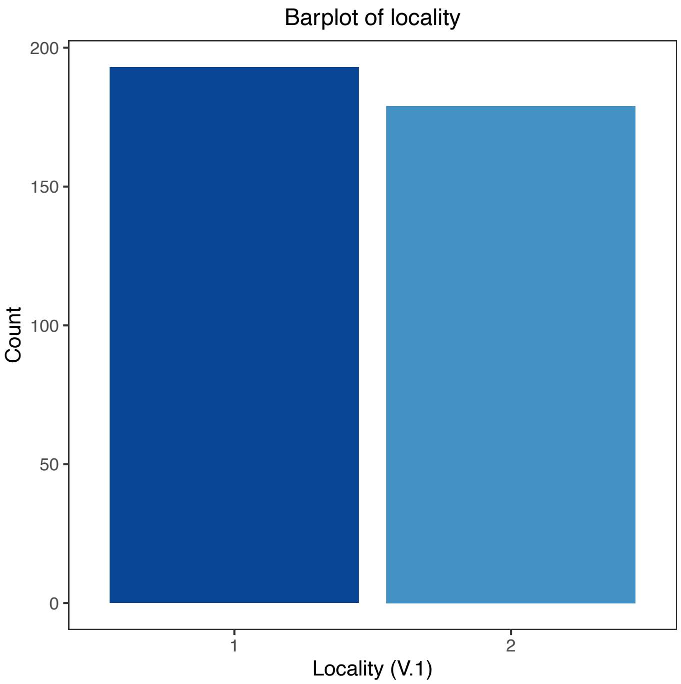
## 02 Data Analysis

# Who affects Actual Sales Price?

02

## Data Analysis

# Locality



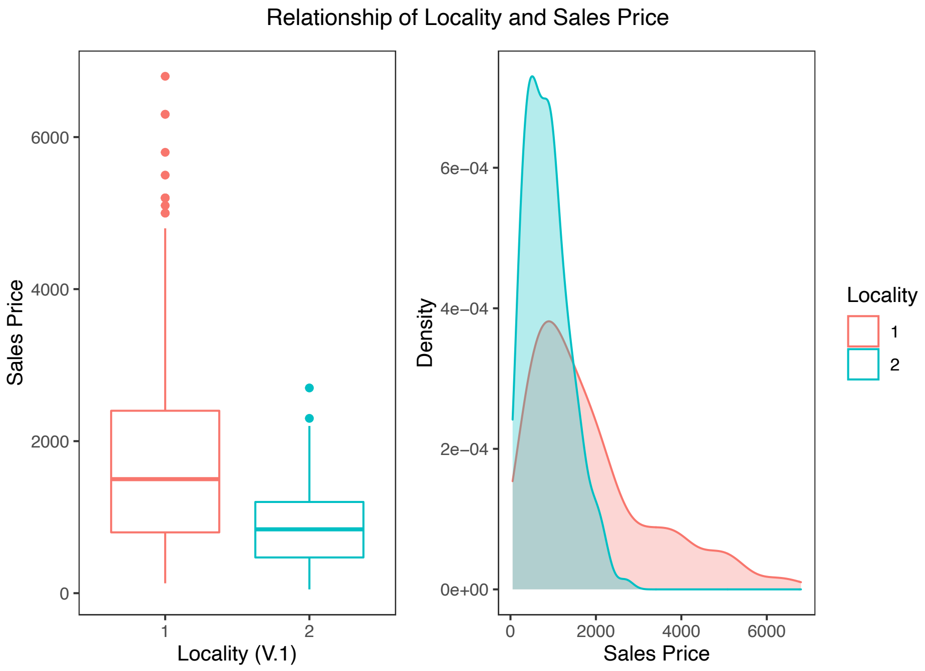
| Type | 1 | 2 |
|---|---|---|
| V.1 | 193 | 179 |

It shows that the number of apartments in different localities is similar.

02

## Data Analysis

# Locality

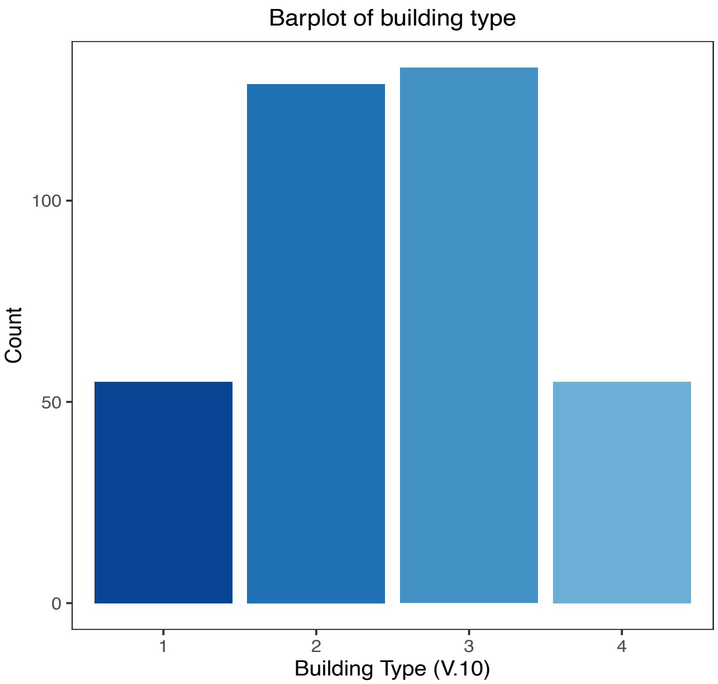


- Purpose: Determine whether the locality actually has an effect on the sales price.
- Statistical Test: T-test
- H0: Sales prices in two locality are not significantly different.
- Result: t = 8.749, df = 245.1, p-value = 3.532e-16
- Conclusion: Locality has influence on the sales price. Apartments in locality 1 are more expensive.

02

# Data Analysis

## Building Type

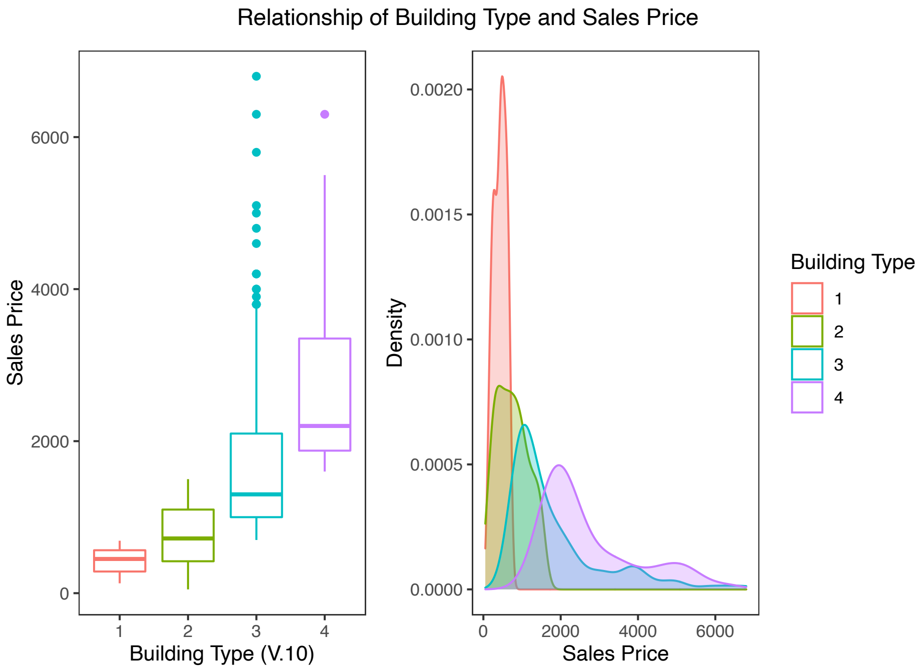


| Type | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| V.10 | 55 | 129 | 133 | 55 |

Most of the apartments belong to type 2 and 3.

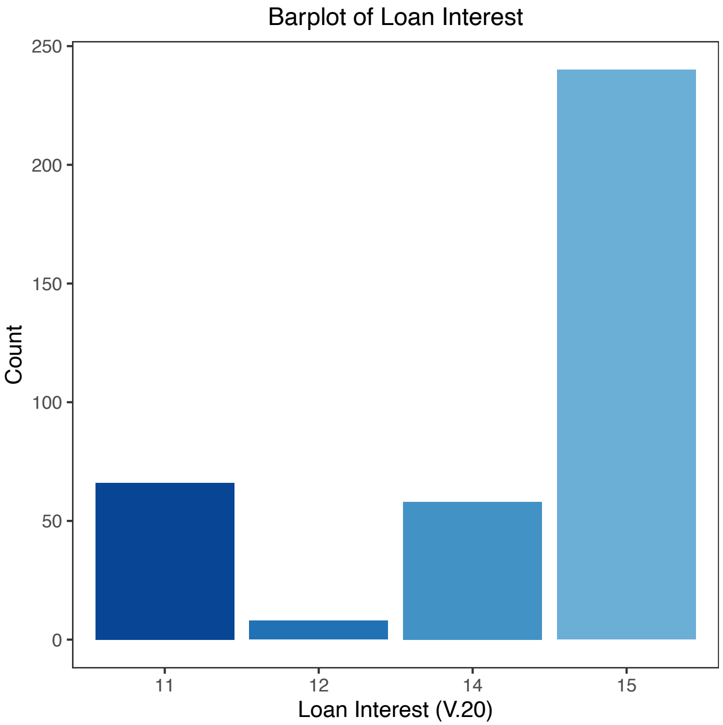
02 **Data Analysis**

# Building Type



- Purpose: Determine whether the building type actually has an effect on the sales price.
- Statistical Test: ANOVA test
- H0: Sales Price stays the same under different Building Type.
- Result: df = 3, F = 90.6, p-value < 2e-16
- Conclusion: Building type has influence on the sales price, which increases from type 1 to type 4.

# Loan Interest Rate



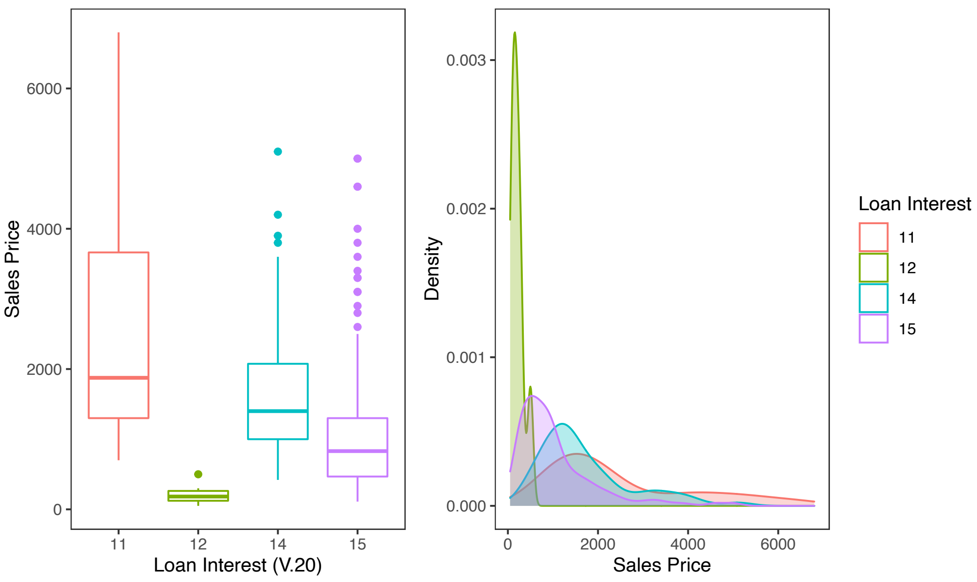
| Type | 11 | 12 | 14 | 15 |
|---|---|---|---|---|
| V.20 | 66 | 8 | 58 | 240 |

Usually, the interest rate for loan is 15%, while 12% in rare case.

Changes in interest rates can sensitively reflect the supply and demand of funds in the financial market.

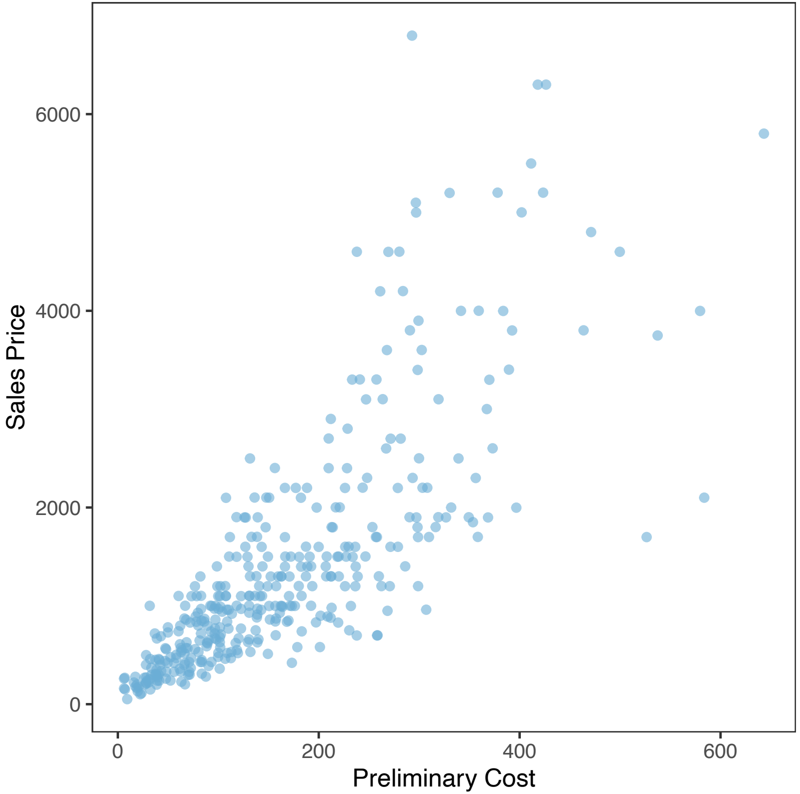
# Loan Interest Rate

Relationship of Loan Interest and Sales Price



- Purpose: Determine whether the interest rate for loan has an effect on the sales price.
- Statistical Test: ANOVA test
- H0: Sales Price stays the same under different Loan Interest
- Result: df = 3, F = 36.13, p-value < 2e-16
- Conclusion: The higher the interest rate, the lower the price. There are few samples with interest rate equal to 12 so that the corresponding price distribution is special.

# Preliminary Estimated Construction Cost

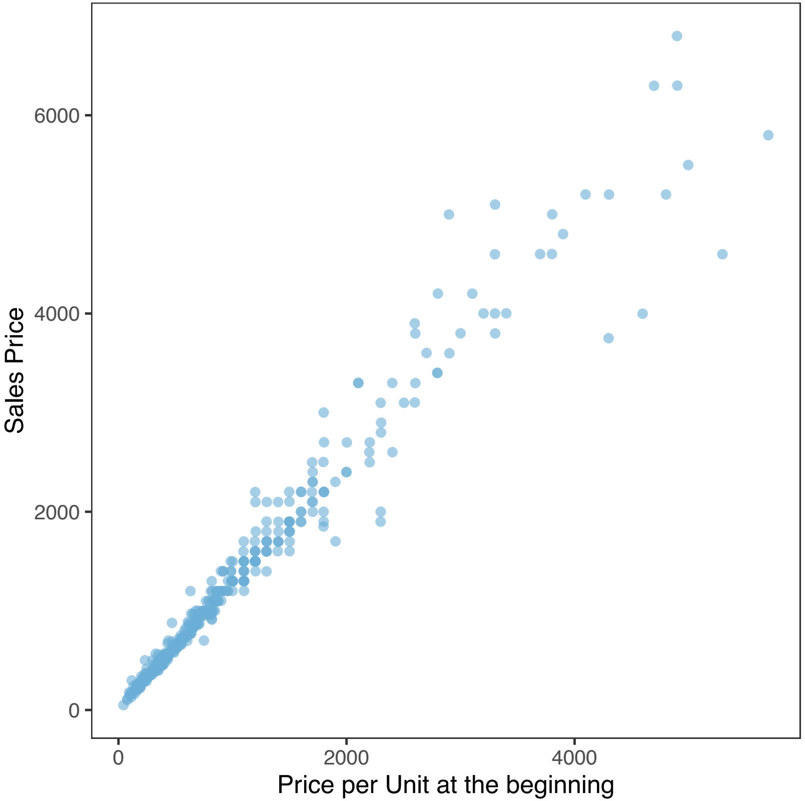


- Statistical Test: T test
- H0: Preliminary Estimated Construction Cost has no influence on Sales Price
- Result: t = 24.373, p-value < 2e-16
- Conclusion: Preliminary Estimated Construction Cost has influence on Sales Price. As it increases, the sales price will be higher as well. Precisely, their linear association is more obvious when the construction costs are relatively low.

02

## Data Analysis

# Price of the unit at the beginning



- Statistical Test: T test
- H0: Price of the unit at the beginning has no influence on Sales Price
- Result: t = 87.023, p-value < 2e-16
- Conclusion: Price of the unit at beginning has influence on Sales Price. Also, the adjusted R square hits 0.9533, which means only this variable explain over 95.33% of the variance of the Sales Price.

## 02 Data Analysis

Is there redundancy in variables?

# Locality VS Building Type

| V.1<br>V.10 | 1 | 2 |
|---|---|---|
| 1 | 16 | 39 |
| 2 | 57 | 72 |
| 3 | 75 | 58 |
| 4 | 45 | 10 |

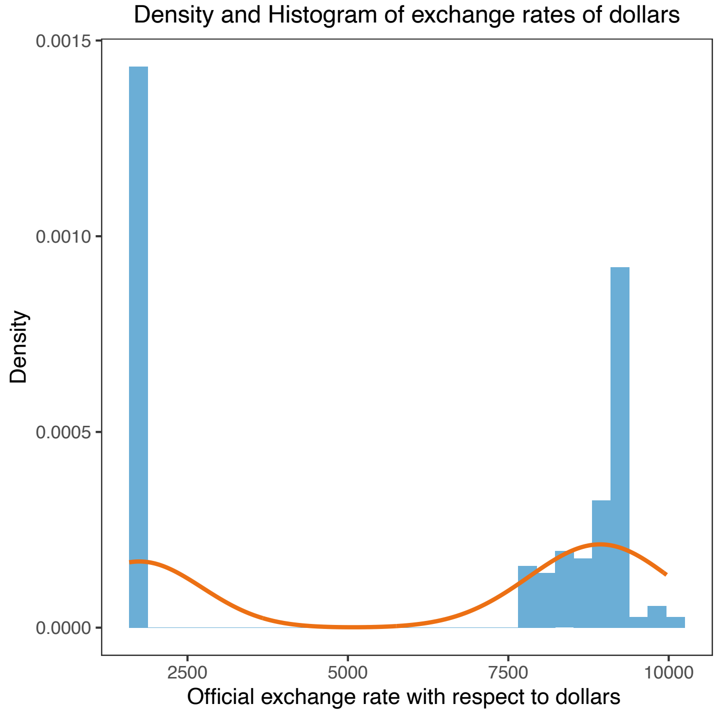
- Purpose: Since these two variables both have influence on price, we test their relationship.
- Statistical Test: Chi-square test
- H0: Locality and Building Type are independent.
- Result: df = 3, X-square = 35.33, p-value =1.037e-07
- Conclusion: The locality and building type are dependent.

# Numerical Variables

There are 25 numerical variables. The most obvious feature is the **high correlation** between variables which means we have to deal with **collinearity** when building a model.

02 Data Analysis

# Official exchange rate with respect to dollars



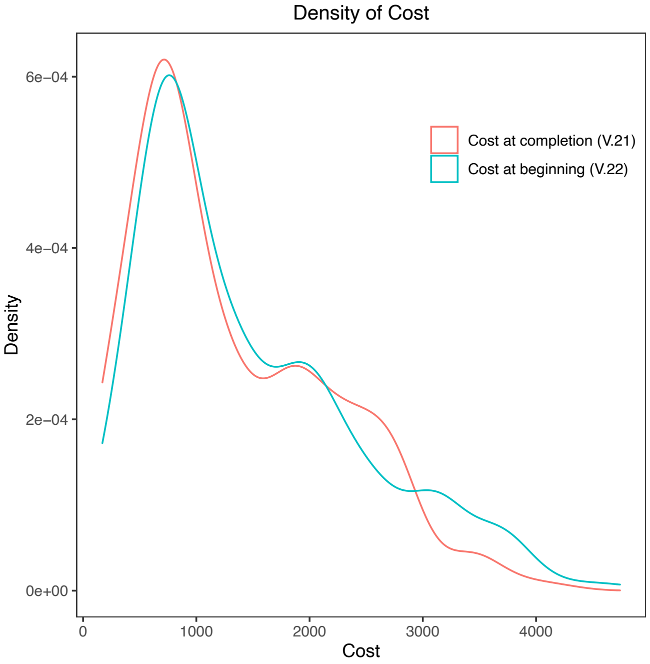
| | |
|---|---|
| mean | 5934.19 |
| std | 3543.38 |
| min | 1591.75 |
| 25% | 1755 |
| 50% | 8209.9 |
| 75% | 9137.91 |
| max | 9967.33 |

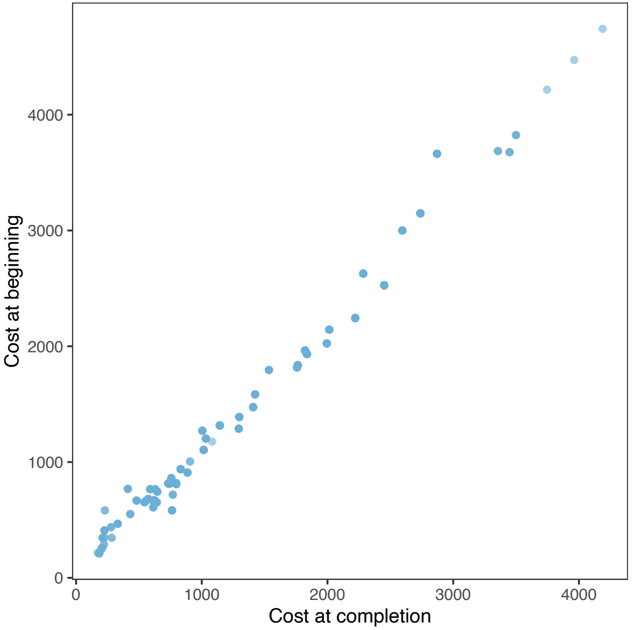
The distribution of official exchange rate is relatively sparse.

02

## Data Analysis

# Cost at different Phases



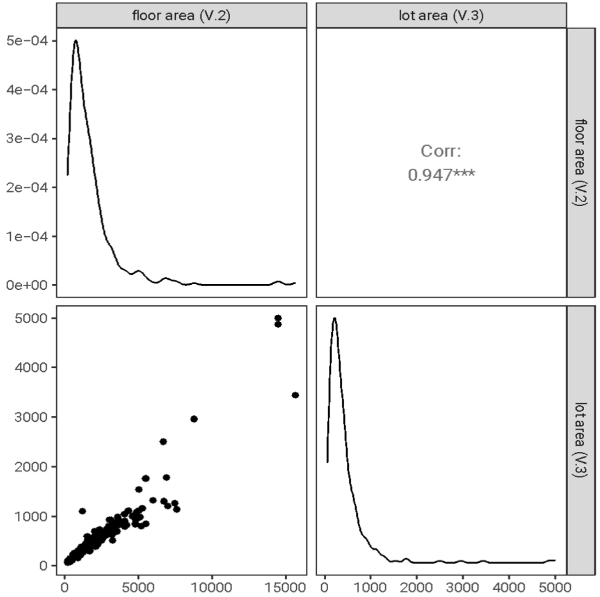


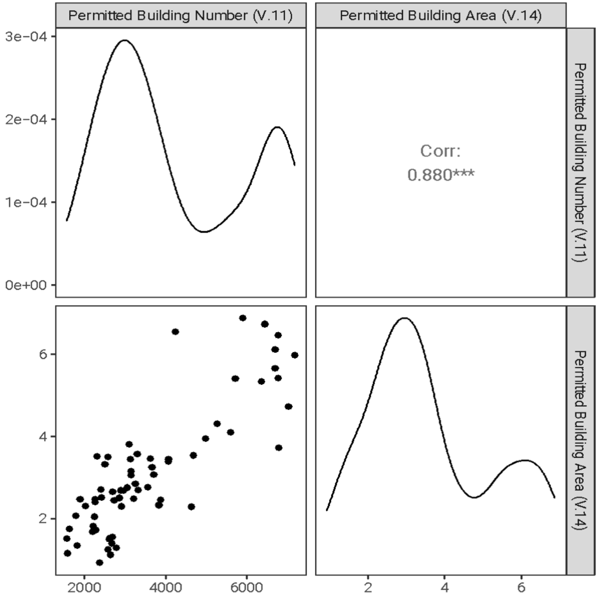
| | Completion Cost | Beginning Cost |
|---|---|---|
| mean | 1327.5 | 1466.3 |
| std | 868.5 | 957.2 |
| min | 170.3 | 211.1 |
| 25% | 641.5 | 744.5 |
| 50% | 1023.7 | 1203.3 |
| 75% | 1994.6 | 2025 |
| max | 4188.6 | 4741.6 |

# Cost at different Phases (V.21 and V.22)

They are the average of construction cost of buildings by private sector at the time of completion and beginning respectively. It shows that their distributions are similar and totally the average of cost is more expansive at the beginning than the end.

# Correlations between two variables



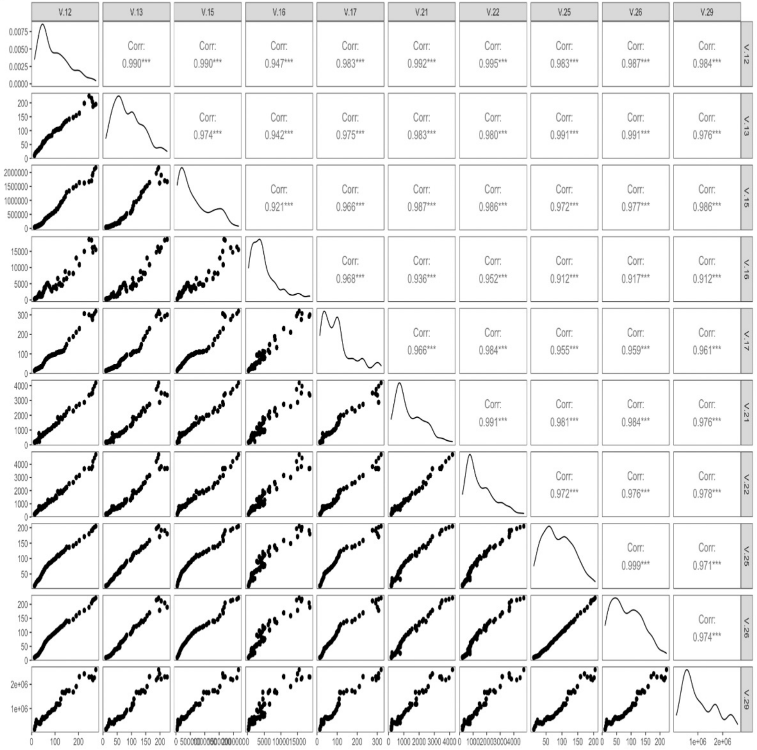


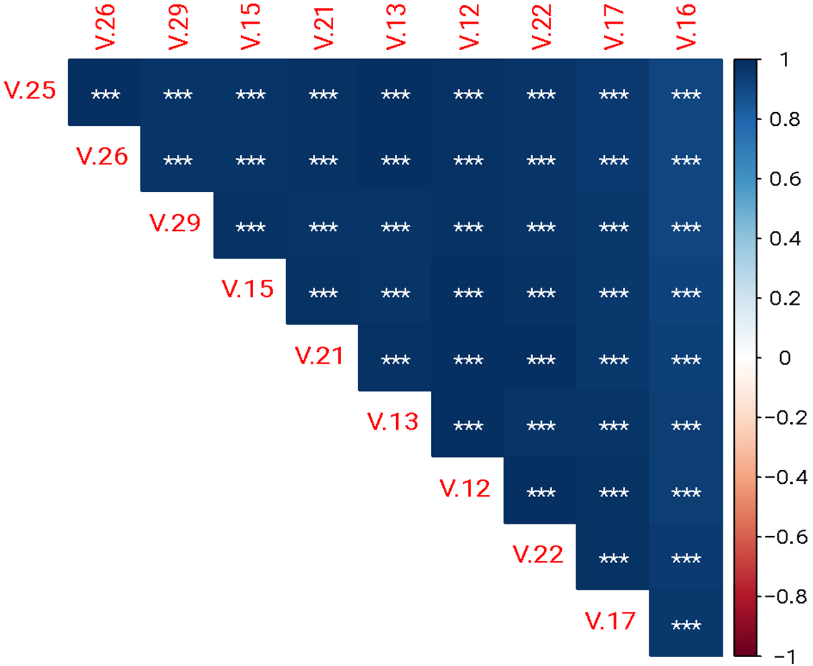
There are two positive correlated groups: floor area and lot area, permitted building number and permitted building area. The correlation coefficient both achieve above 0.88.

02

## Data Analysis

# Correlations among several variables





# Correlations among several variables

One of the characteristics of this dataset is the high correlation between variables. From the pairwise plot and heatmap plot, we can conclude that any two of V.12, 13, 15, 16, 17, 21, 22, 25, 26, 29 are positively correlated. This will cause collinearity when building a model. Thus, we have to remove some variables or perform an analysis designed for highly correlated variables, such as PCA.

# Sale Price Prediction

# Linear regression model

- Build a linear regression model with high interpretability and prediction accuracy.
- Considering the fact that some variables are **highly correlated** as concluded in the data analysis section, we remove variables whose correlation coefficient is larger than 0.85.

# Final model

| variable | coefficient | variable | coefficient |
|---|---|---|---|
| Locality 2 | -101.8 | **The amount of loans** | 0.007 (***) |
| Total estimated construction cost | 0.154(**) | The interest rate equal to 12 | -388.6(**) |
| **Estimated construction cost** | -1.72 (***) | The interest rate equal to 14 | -149.3(*) |
| Estimated construction cost in a base year | 0.165(*) | **Private sector investment in new buildings** | -0.056 (***) |
| **Duration of construction** | 42.56 (***) | **The interest rate equal to 15** | -287.3 (***) |
| **Price at the beginning of project** | 1.176 (***) | **Stock market index** | 0.024 (***) |
| Building Type 4 | 132.5(*) | Population of the city | -0.004 |
| Total floor areas of building permits issued | 27.36(**) | | |

Significant codes : p value of t test 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’

# Model explanations

Variables in the table above are all effective drivers of price. The practical meaning of coefficient is the magnitude of the effect of one variable on price. In particular, the interest rate has a negative and relatively great influence on price while the price at the beginning of the project has a significantly positive effect on the actual price.

# How good our model fits

$R^2$ is a common factor to evaluate the performance of model and is referred to as the "**goodness of fit**". It provides a measure of how well observed outcomes are replicated by the model.  In addition, adjusted $R^2$ is a tradeoff between $R^2$ and model complexity. The higher the adjusted $R^2$, the better the model. Commonly, a model achieving adjusted $R^2$ above 0.8 is considered a very good model. Here, we achieve adjusted $R^2$ 0,9769, indicating our model is a good fit.

| | $R^2$ | Adjusted $R^2$ |
|---|---|---|
| train | 0.9786 | 0.9769 |
| test | 0.9653 | 0.9590 |

# How accurate our model predicts

Root-Mean-Square error (RMSE) is a frequently used measure of the differences between values predicted by a model and the values observed. It measures how much our predicted value deviate the true value.

Normalized Root-Mean-Square error (NRMSE) is a normalized version of RMSE, which eliminates the influence of the scale. The lower NRMSE is, the more accurate our model predicts. A value of 0 indicates a perfect fit.

Here, we achieve 0.0351 NRMSE, which indicates highly accuracy of our model.

| | RMSE | NRMSE |
|---|---|---|
| train | 171.549 | 0.0277 |
| test | 237.230 | 0.0351 |

# Recommendations

# Recommendations

The actual sales prices can be influenced by several factors and may vary from year to year due to inflation, deflation, or other economic conditions. There are some regularities:

- In general there is an **inverse relationship** between the interest rates and the price.
- There are several important drivers that companies and investors should notice including the **interest rate, construction cost, duration, price at the beginning of the project, stock market index** etc.

# THANK YOU