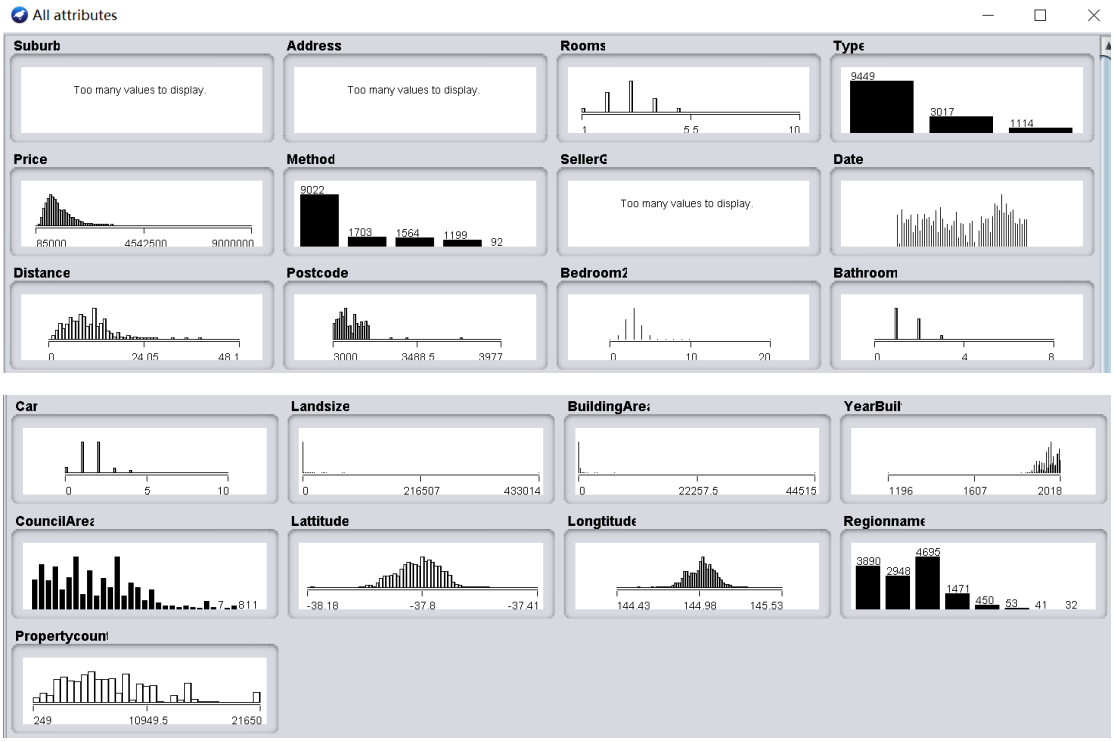
USE CASE

Measure the shape of data source

The information of the data set (melb_data.csv) is as follows:

Name	Number	Details
Attributes	21	Features: 20 Class: 1 (Price)
Instances	13850	Nominal:8 Numeric:13
Missing data	4	Car (62, 0%) BuildingArea (6480, 47%) YearBuild (5373, 40%) CouncilArea (1369, 10%)
Outliers		



Understanding what are categories and features

This file (melb_data.csv) describes the information of houses sold in Melbourne area from 2016 to 2017. The data set contains 21 columns, including 20 columns of features, and a target variable, namely Price (class).

Features: Suburb, Address, Rooms, Type, Price, Method, SellerG, Date, Distance, Postcode, Bedroom2, Bathroom, Car, Landsize, BuildingArea, YearBuilt, CouncilArea, Lattitude, Longtitude, Regionname, Propertycount.

In the above 20 feature sets, we select the subset that we focus on, and then conduct data analysis on the relevant factors that may affect the housing price. Such as:

Suburb, Regionname, CouncilArea: The area where the house is located.

Type: Three house types (h, u, t)

Date: When the house was sold

Distance: The distance between the house and a center.

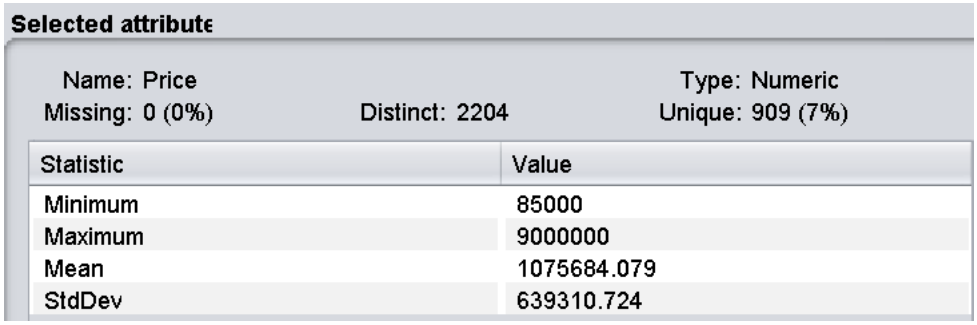
Bedroom2, Bathroom, Car, Landsize, BuildingArea: Number of rooms and construction area.

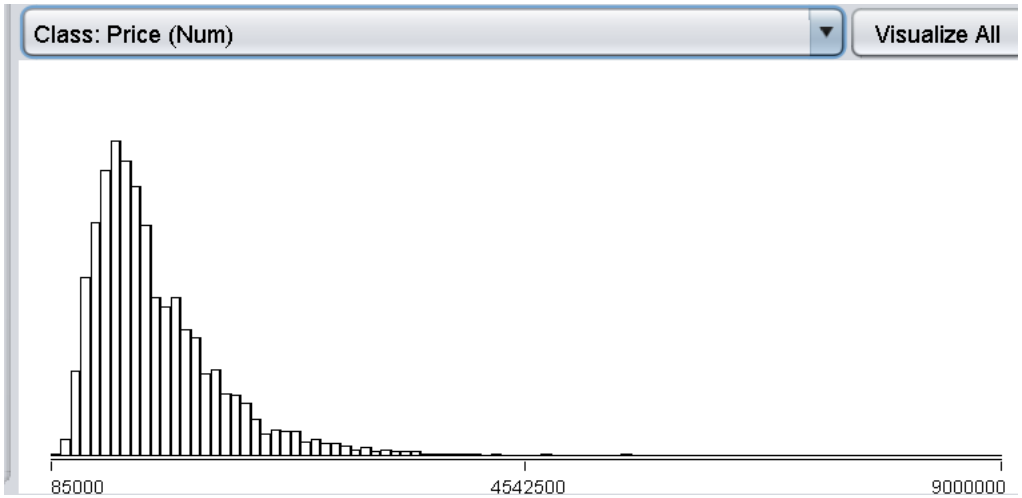
YearBuilt：Year of building.

Lattitude, Longtitude: Geographical distribution.

Price: Sale price，which is also the target variable of observation.

Target variable distribution (Price)





It can be seen that Price is not a normal distribution, but a positive skewness (right skewness).

By features, assuming what questions can be asked from users/stakeholders

1. Buyers

Buyers pay attention to the factors that affect the price negotiation and seek the best price performance housing. For example, in which areas can customers buy houses? For example, the customer demand is: the unit price is below XXX yuan, and the age of the house is not higher than x years.

2. Seller

Hot housing types, price adjustment, publicity, sales methods

3. Investors

Is the development trend of house price falling or rising? Therefore, it is necessary to analyze and predict the time development trend of house prices. Which area has more investment value?

4. Government

Data statistics, regulation of house prices and taxation.

The possible problems are listed as follows:

1. The most important factors affecting house price

2. House price ranking of each district

3. House price ranking of popular areas

4. How many houses are sold in each area? For example, the current number of houses sold in the top ten areas

5. What are the popular districts in each district? The hot area in the house source and the area

6. What are the best-selling house types?

7. According to the purchase time, statistics the transaction volume within a certain period of time, housing transaction type, can make targeted adjustments, such as promotion, building houses and investment and development

List all these questions you want to solve by analyzing the data

The purpose is to predict the sales price of the house by analyzing the data related attributes. For example, selecting multiple groups of factors that may affect the house price: YearBuilt, geographical location and type of house,etc.

The house type can be subdivided into Bedroom2, Bathroom, Car, Landsize, BuildingArea**.**

1. Data preprocessing

- Identify the type of variable: numerical variable and nominal variable need to be processed separately, or the housing building time should be changed to numerical type to facilitate the subsequent calculation of housing age.

- Missing values are filled
- Outlier handling, such as outliers

2. Feature selection

- Single feature. View the distribution of each numerical feature.
- Correlation of data features. If house price is closely related to location and house type, we could describe the correlation between several variables by using standard correlation coefficient.
- Create new features

3. Data prediction

The task could be to predict the factors related to housing prices, analyze the reasons and draw conclusions, such as the trend of housing prices, hot-selling housing types, or the influence of different regions on housing sales and prices.