THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

CIVL 4210 - Advanced Construction with AI and Robotics

Guidebook: Housing Price Regression

Prof. Yu

HUANG Xuhong

Introduction

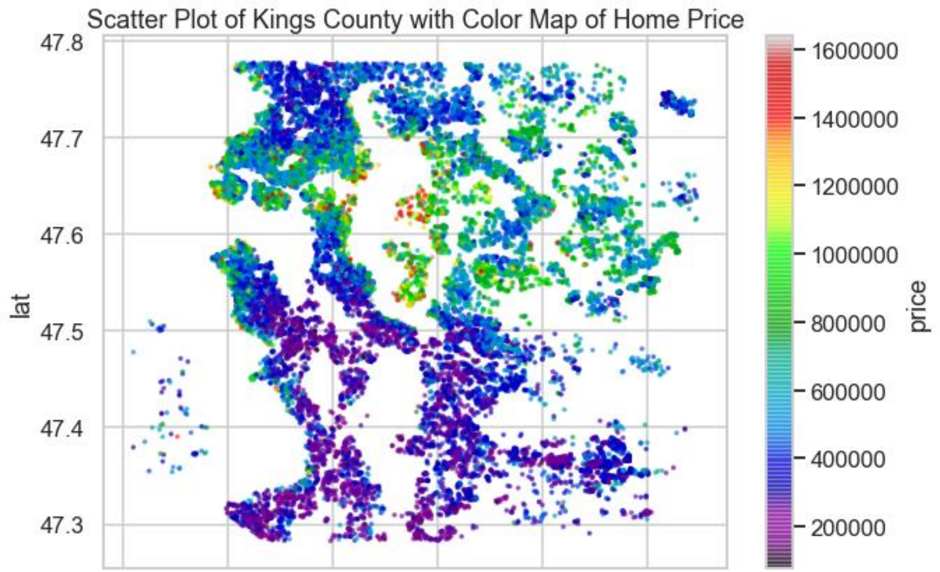
THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

The objective is to create the **linear regression model** and the **multiple regression** for a given dataset(House Sales in King County, USA).

The overall idea of regression is to examine two things:

(1) Does a set of predictor variables do a good job of predicting an outcome (dependent) variable?
(2) Which variables are significant predictors of the outcome variable, and how do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?



These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

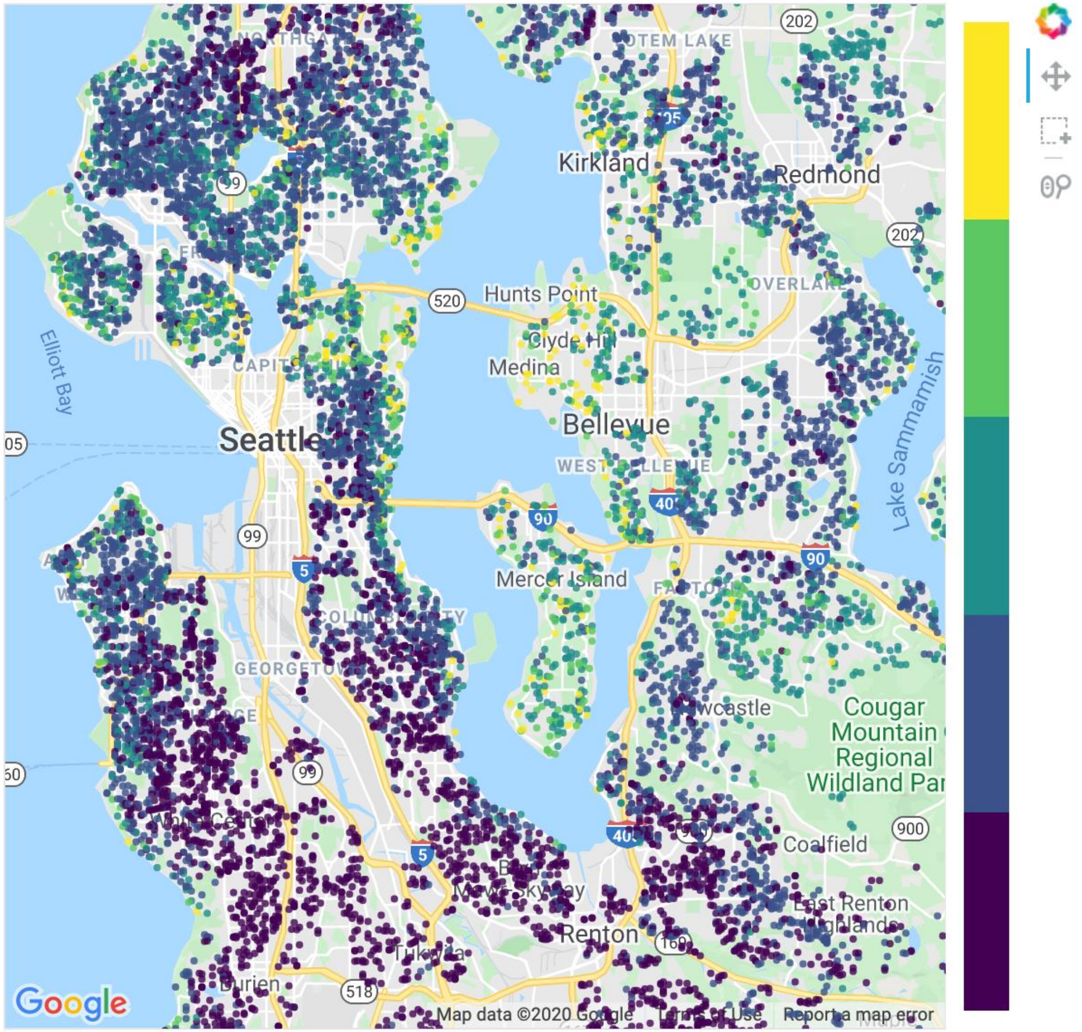
Data

In this dataset, we need to predict the **sales price of houses in King County, Seattle**. It includes homes sold between May 2014 and May 2015.

Before doing anything, we should first know about the dataset, what it contains, what its features are, and what the structure of the data is.

The dataset contains 20 house features plus the price, along with 21613 observations.

Home Sale Prices in Kings County from May 2014 - May 2015



Data

The description for the 20 features is given below:

(1) **id** :- It is the unique numeric number assigned to each house being sold.
(2) **date** :- It is the date on which the house was sold out.
(3) **price**:- It is the price of house which we have to predict so this is our target variable and aprat from it are our features.
(4) **bedrooms** :- It determines number of bedrooms in a house.
(5) **bathrooms** :- It determines number of bathrooms in a bedroom of a house.
(6) **sqft_living** :- It is the measurement variable which determines the measurement of house in square foot.
(7) **sqft_lot** : It is also the measurement variable which determines square foot of the lot.
(8) **floors** : It determines total floors means levels of house.
(9) **waterfront** : This feature determines whether a house has a view to waterfront 0 means no 1 means yes.
(10)**view** : This feature determines whether a house has been viewed or not 0 means no 1 means yes.

Data

The description for the 20 features is given below:

(11) **condition** : It determines the overall condition of a house on a scale of 1 to 5.
(12) **grade** : It determines the overall grade given to the housing unit, based on King County grading system on a scale of 1 to 11
(13) **sqft_above** : It determines square footage of house apart from basement.
(14) **sqft_basement** : It determines square footage of the basement of the house.
(15) **yr_built** : It detrmines the date of building of the house.
(16) **yr_renovated** : It detrmines year of renovation of house.
(17) **zipcode** : It determines the zipcode of the location of the house.
(18) **lat** : It determines the latitude of the location of the house.
(19) **long** : It determines the longitude of the location of the house.
(20) **sqft_living15** : Living room area in 2015(implies-- some renovations)
(21) **sqft_lot15** : lotSize area in 2015(implies-- some renovations)

Data

By observing the data, we can know that the price is dependent on various features like **bedrooms**(which is the most dependent feature), **bathrooms**, **sqft_living**(second most important feature), **sqft_lot**, **floors**, etc. The price is also dependent on the location of the house where it is present. The other features, like the waterfront view, are less dependent on the price. Of all the records, there are no missing values, which helps us create a better model.

First, we import the required libraries like pandas, numpy, seaborn, and matplotlib. Now, import the CSV file. Now, we should get to know how the data is and what data type uses the info function. We observe that the date is in 'object' format. To see the no of rows and columns, we use the shape function. Describe the data frame to know the mean, minimum, maximum, standard deviation, and percentiles.

Then, we plot graphs for visualization, and then we do simple regression using 'bedrooms,' multiple regression, and polynomial regression.

Step 1: Check the data in Colab

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

Step 1: Import and initialization

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

csvData = 'https://raw.githubusercontent.com/Yokhong/CIVL4210/main/H3_housing_price_regression/kc_house_data.csv’

df = pd.read_csv(csvData)

Step 2: Check the data

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             21613 non-null  int64
 1   date           21613 non-null  object
 2   price          21613 non-null  float64
 3   bedrooms       21613 non-null  int64
 4   bathrooms      21613 non-null  float64
 5   sqft_living    21613 non-null  int64
 6   sqft_lot       21613 non-null  int64
 7   floors         21613 non-null  float64
 8   waterfront     21613 non-null  int64
 9   view           21613 non-null  int64
 10  condition      21613 non-null  int64
 11  grade          21613 non-null  int64
 12  sqft_above     21613 non-null  int64
 13  sqft_basement  21613 non-null  int64
 14  yr_built       21613 non-null  int64
 15  yr_renovated   21613 non-null  int64
 16  zipcode        21613 non-null  int64
 17  lat            21613 non-null  float64
 18  long           21613 non-null  float64
 19  sqft_living15  21613 non-null  int64
 20  sqft_lot15     21613 non-null  int64
dtypes: float64(5), int64(15), object(1)
memory usage: 3.5+ MB
```

df.head()

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...

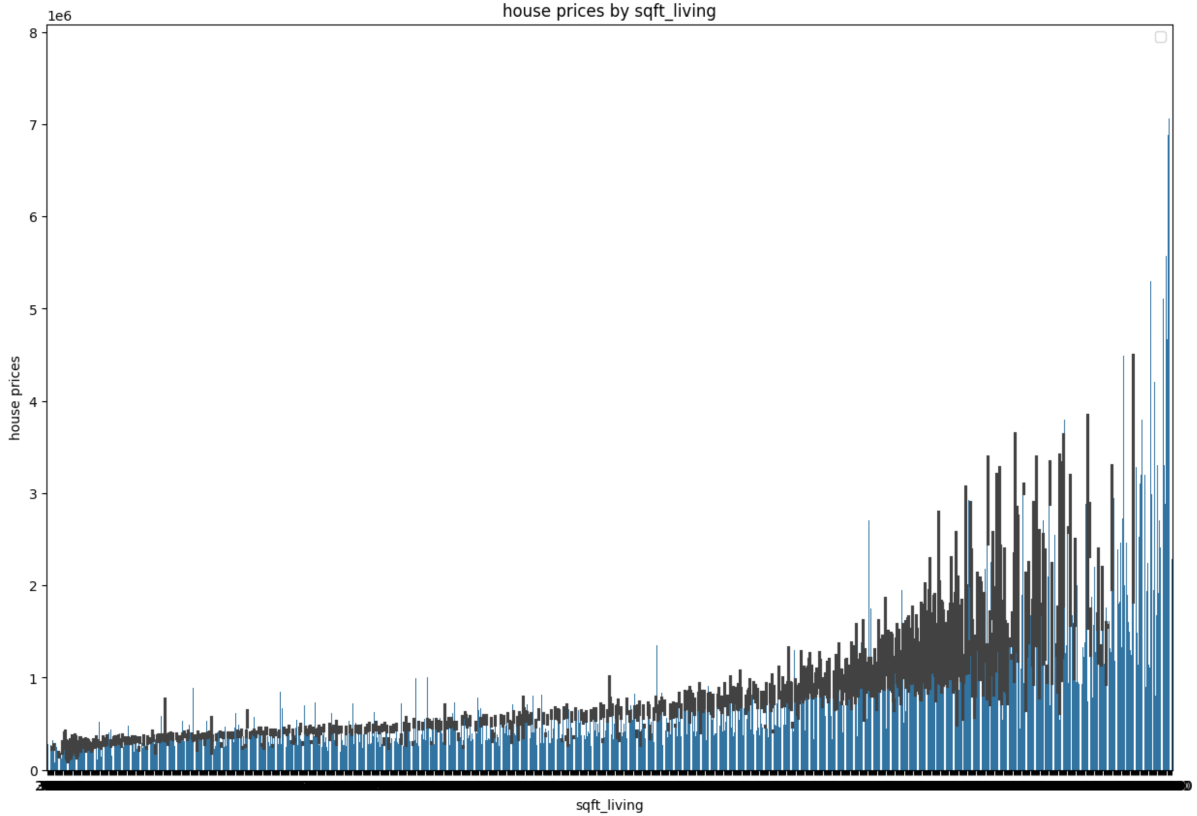
5 rows × 21 columns

⋮

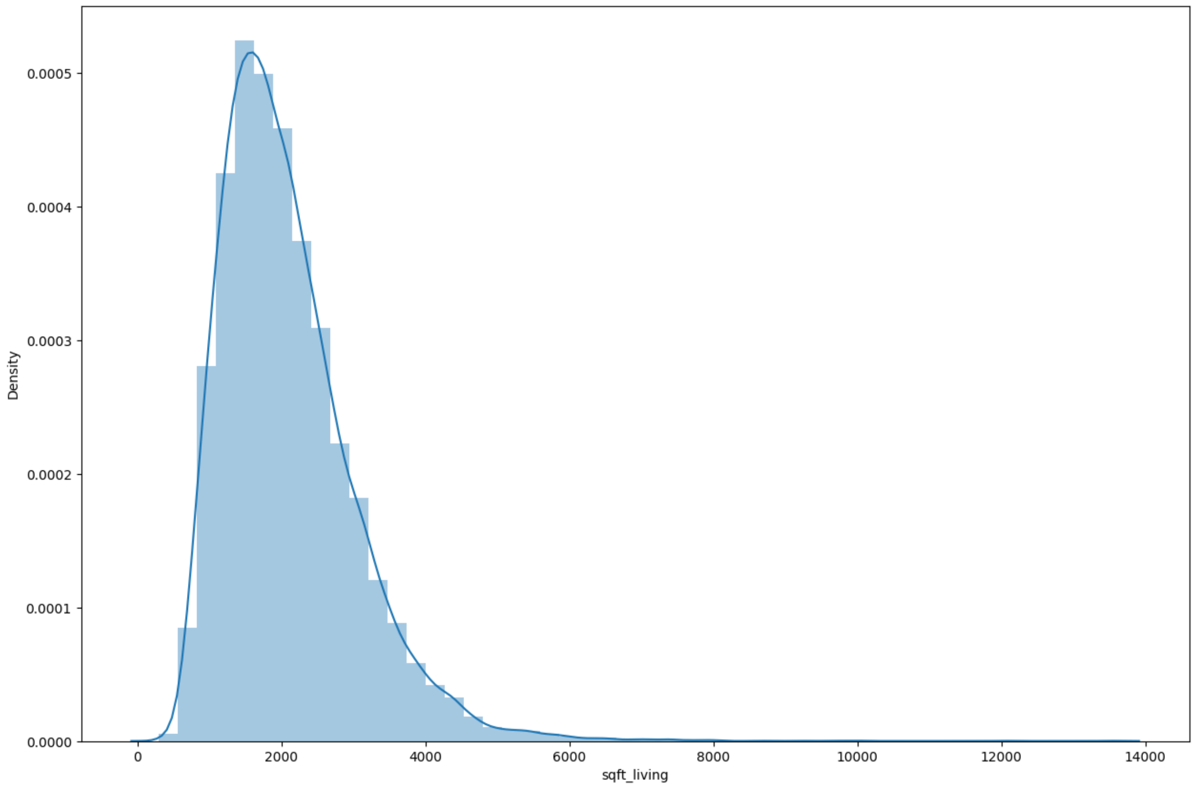
Step 2: Statistics and graphing

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING



House prices by sqft_living



Density by sqft_living

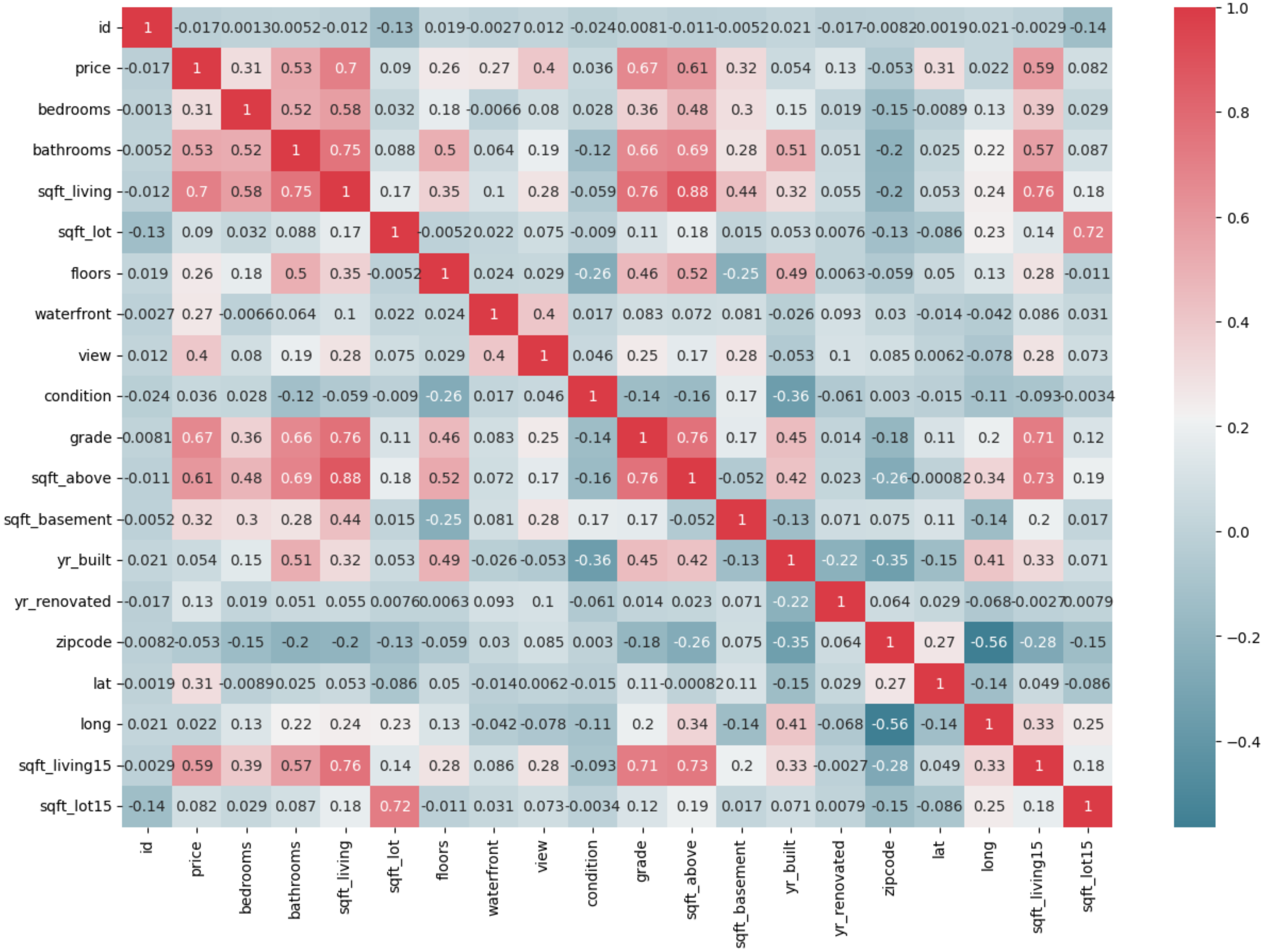
Step 2: Statistics and graphing

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

What is this?

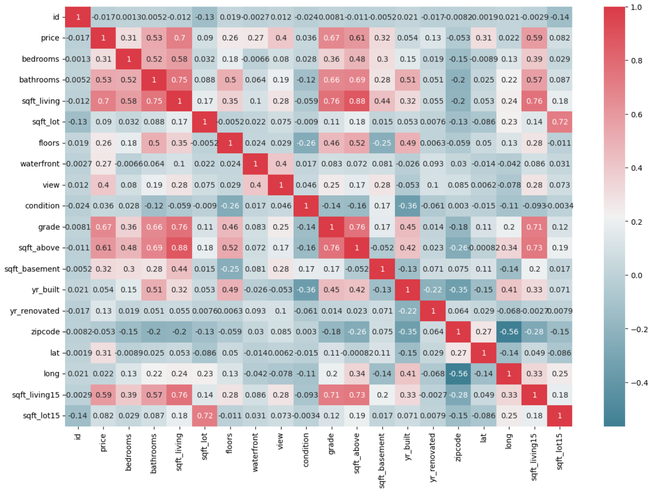
A house price correlation heatmap shows the **relationships between different factors** that can affect house prices, such as location, size, number of bedrooms, or amenities.



Step 2: Statistics and graphing

What is this?

1. **Variables Analyzed**:
   Common variables include square footage, number of bedrooms, bathrooms, age of the home, location, and features like pools or garages.
2. **Correlation Coefficient**:
   The heatmap uses a correlation coefficient (usually between -1 and 1) to show relationships:
   1 means a perfect positive correlation (as one increases, the other does too).
   -1 means a perfect negative correlation (as one increases, the other decreases).
   0 means no correlation.
3. **Color Coding**:
   Different colors represent different levels of correlation. For example:
   1. **Darker colors** may indicate stronger correlations.
   2. **Lighter colors** may indicate weaker correlations.



Step 2: Statistics and graphing

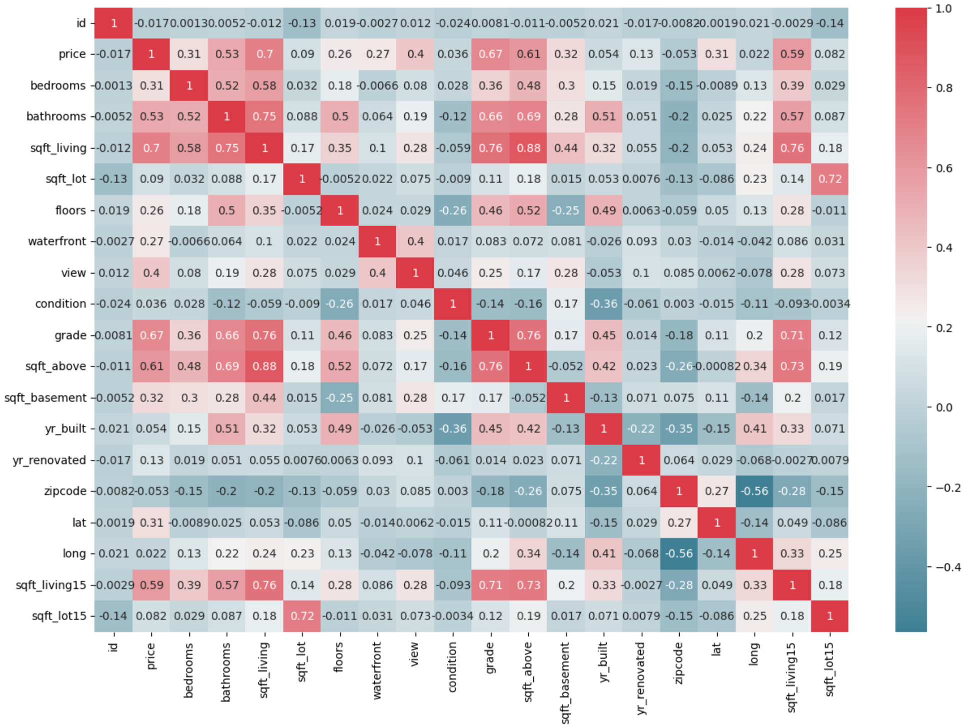
THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

What can it do?

It helps real estate professionals and buyers understand which features most influence house prices. Buyers can use this information to make informed decisions about what features to prioritize. Sellers can understand what aspects of their home might increase its value.

For example, if the heatmap shows a strong correlation between square footage and price, it suggests larger homes tend to sell for more.



Step 3: Simple Linear Regression

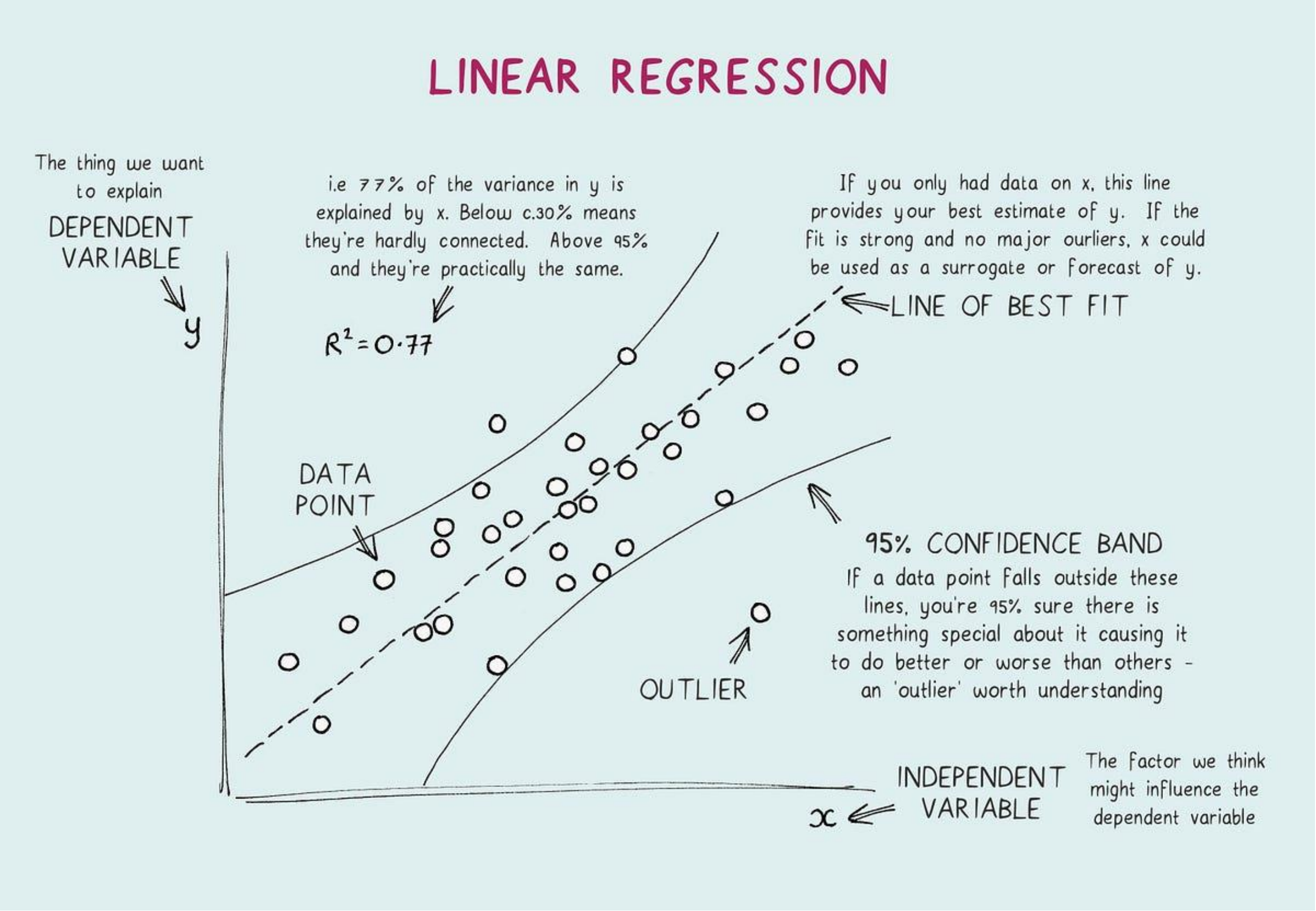
THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

What is this?

Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages.

(1) Analyzing the correlation and directionality of the data
(2) Estimating the model, i.e., fitting the line
(3) Evaluating the validity and usefulness of the model



Step 3: Simple Linear Regression

How can we build this?

1.Collect Data: **(done)**

Gather data that includes the independent variable and the dependent variable. For example, as shown on the right, you might have living space(X) and house price (Y).

2.Visualize the Data: **(done)**

Examining data features, counting the density distribution of each feature, plotting statistics between two features to determine if there is a linear relationship, and creating heat maps to determine the relationship between different features.

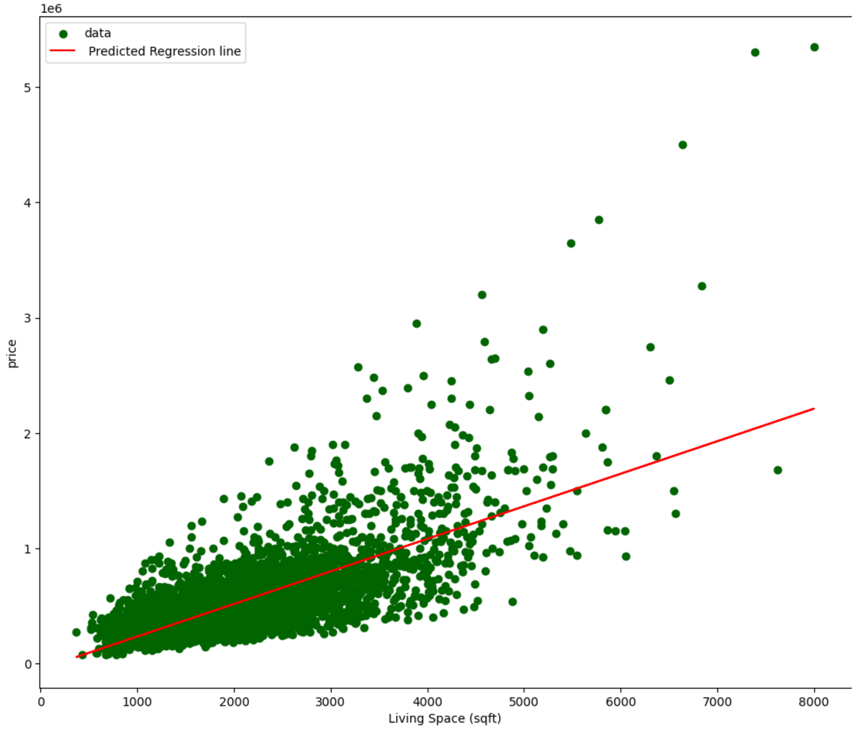
3.Prepare the Data: **(done)**

Clean the data by handling missing values and outliers if necessary.

4.Split the Data:

Divide your data into training and testing sets (e.g., 80% for training and 20% for testing).

train_data,test_data=train_test_split(df,train_size=0.8,random_state=3)



Step 3: Simple Linear Regression

How can we build this?

5. Fit the Model:

Use statistical software or programming language (like Python or R) to fit the linear regression model.

The formula for simple linear regression is: **[Y = b_0 + b_1X]** Where:

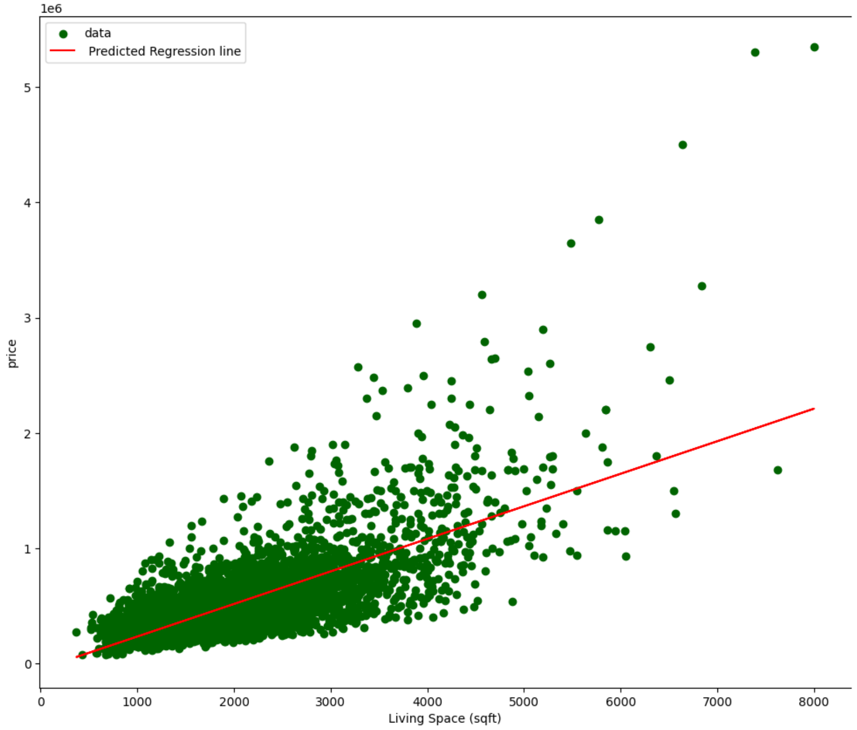
1. (Y) is the predicted value (house price).
2. (b_0) is the y-intercept.
3. (b_1) is the slope of the line (how much Y changes for a unit change in X).
4. (X) is the independent variable (square footage).

6. Evaluate the Model:

Check the model's performance using R-squared, Mean Absolute Error (MAE), or Mean Squared Error (MSE). In our case, we use **R-squared** to evaluate the model.

The R-squared value represents the proportion of the variance in the target variable explained by the model's independent variables.

The R-squared value can range from **0** (can not explain any variance) to **1** (perfectly fitting), with higher values indicating better performance.



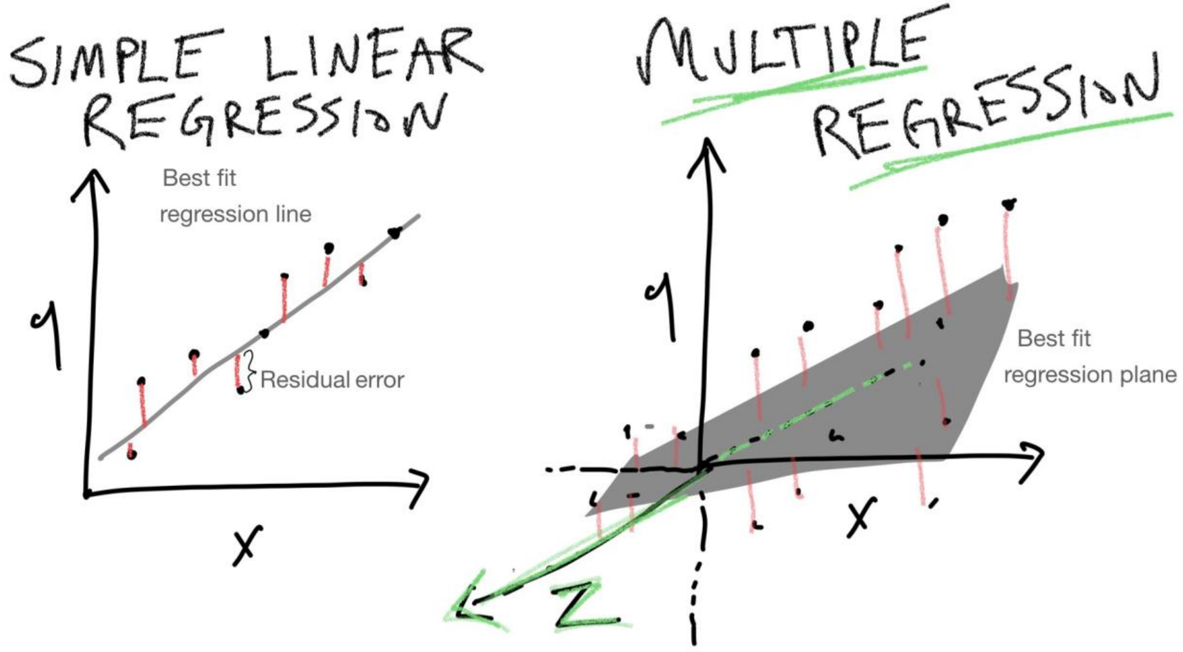
7. Make Predictions:

Use the model to make predictions on new data.

Step 4: Multiple Regression

What is this?

Multiple regression is a statistical tool that helps us understand the relationship between one dependent variable and two or more independent variables. It extends simple linear regression, which involves only one independent variable.

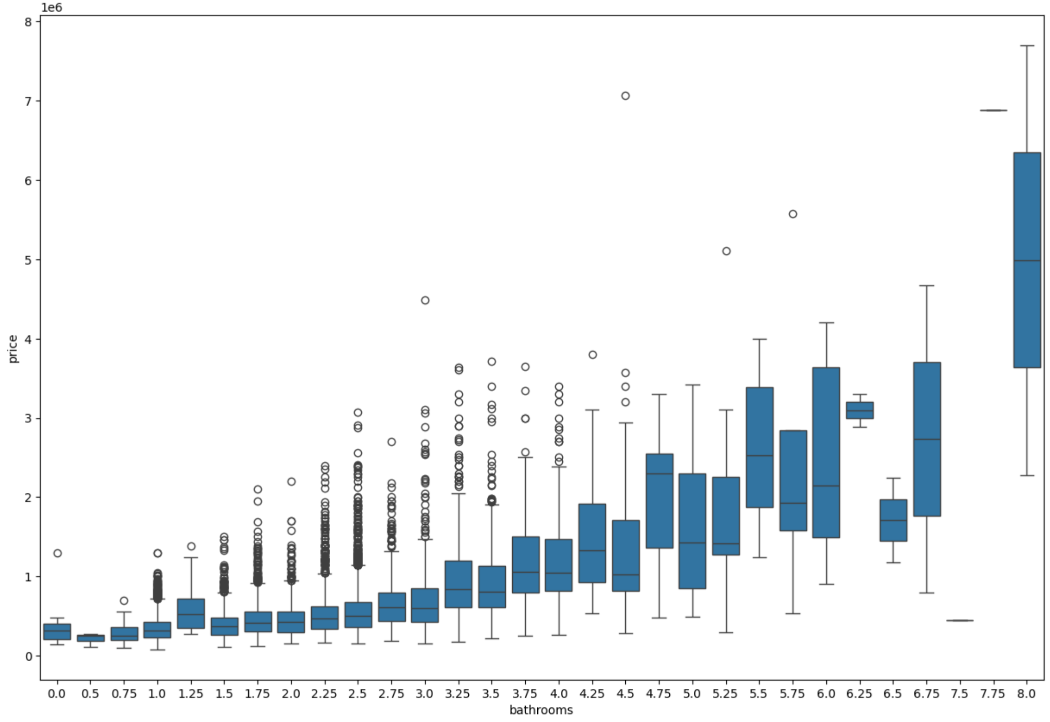


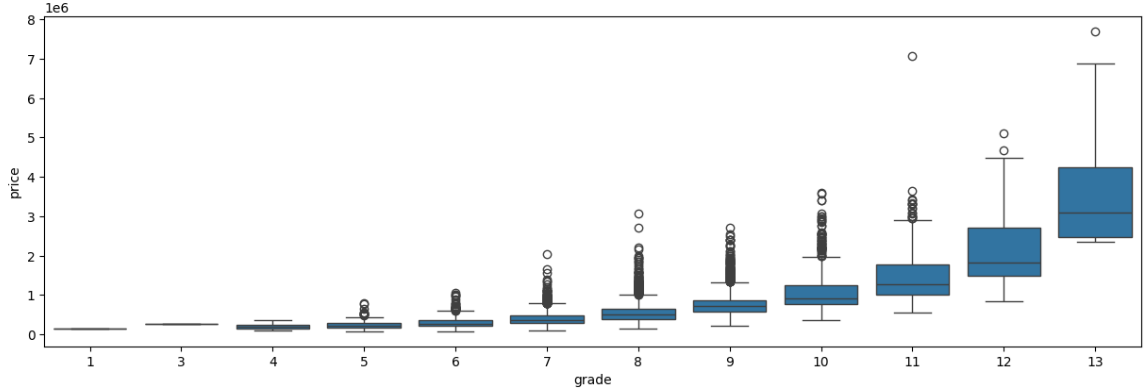
Dependent Variable: This is the outcome or the variable you want to predict (e.g., house price).
Independent Variables: These are the predictors or factors that influence the dependent variable (e.g., square footage, number of bedrooms, location).
The multiple regression equation can be expressed as:[Y = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n + \epsilon] (\epsilon) is the error term

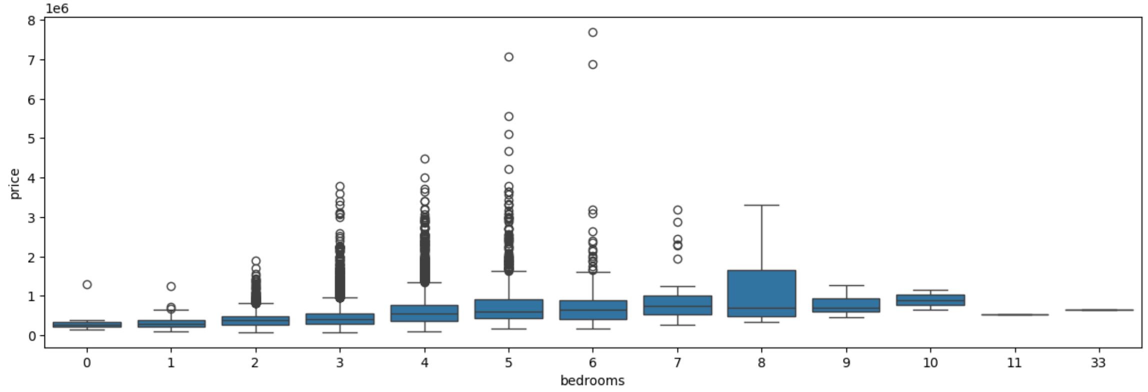
Step 4: Multiple Regression

Before we build the model

Draw a **box plot** to visualize the relationship between different factors (e.g., house class, number of bedrooms, number of bathrooms) and price.







Step 4: Multiple Regression

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

Before we build the model

Box: The central part of the box plot represents the interquartile range (IQR), which contains the middle 50% of the data. The bottom of the box represents the first quartile (Q1, 25th percentile), and the top represents the third quartile (Q3, 75th percentile).

Median Line: A line inside the box that indicates the dataset's median (Q2, 50th percentile). It divides the box into two parts, showing where the middle of the data lies.

Whiskers: Lines that extend from the top and bottom of the box to the highest and lowest values within 1.5 times the IQR from the quartiles. They help visualize the range of the data outside the IQR.

Outliers: Data points outside the whiskers (beyond 1.5 times the IQR). These are typically plotted as individual points or dots. Outliers indicate values significantly higher or lower than the rest of the data.

Axes: The x-axis (horizontal) typically represents the categorical variable (e.g., groups or categories), while the y-axis (vertical) represents the numerical variable (e.g., values or measurements).

Interpretation:

Box Height: The height of the box shows the variability of the data. A taller box indicates greater variability, while a shorter box indicates less variability.

Median Position: The position of the median line within the box indicates the skewness of the data. If the median is closer to the bottom, the data is positively skewed; if it is closer to the top, it is negatively skewed.

Outliers: The presence of outliers can indicate variability in the data or potential anomalies that may require further investigation.

Step 4: Multiple Regression

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

How can we build this?

1.Features Selection:

```
features1 = ['bedrooms', 'grade', 'sqft_living', 'sqft_above']
```

2.Creating the linear regression model:

```
reg = linear_model.LinearRegression()
```

3.Fitting the model:

```
reg.fit(train_data[features1], train_data['price’])
```

A linear regression model is fitted using features1 and the target variable price from the training data. Here train_data[features1] is the features data part.

4.Prediction:

```
pred = reg.predict(test_data[features1])
```

5.Evaluation:

```
mean_squared_error = metrics.mean_squared_error(y_test, pred) print('mean squared error(MSE)',
round(np.sqrt(mean_squared_error), 2))
print('R squared training', round(reg.score(train_data[features1], train_data['price']), 3)) print('R
squared testing', round(reg.score(test_data[features1], test_data['price']), 3))
```

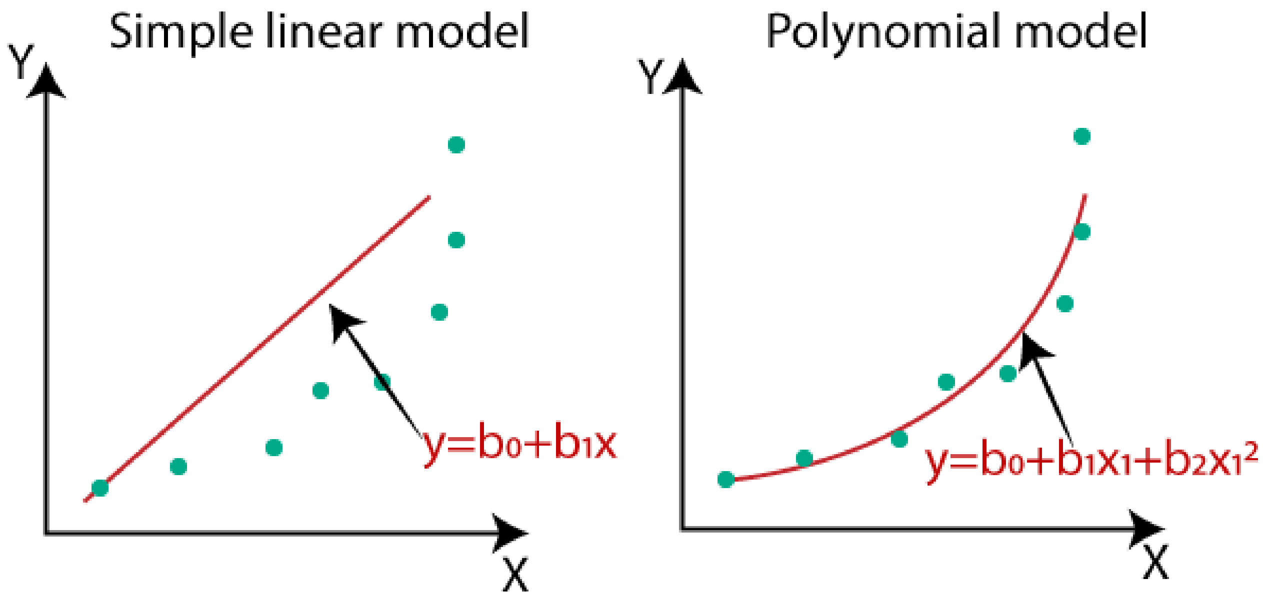
Calculate the Root Mean Square Error (RMSE) between the predicted and actual test set prices. The R-squared value reflects the explanatory power of the model on the test data and ranges from 0 to 1. The closer the value is to 1, the better the model fits the data.

Step 5: Polynomial Regression

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY
SCHOOL OF ENGINEERING

What is this?

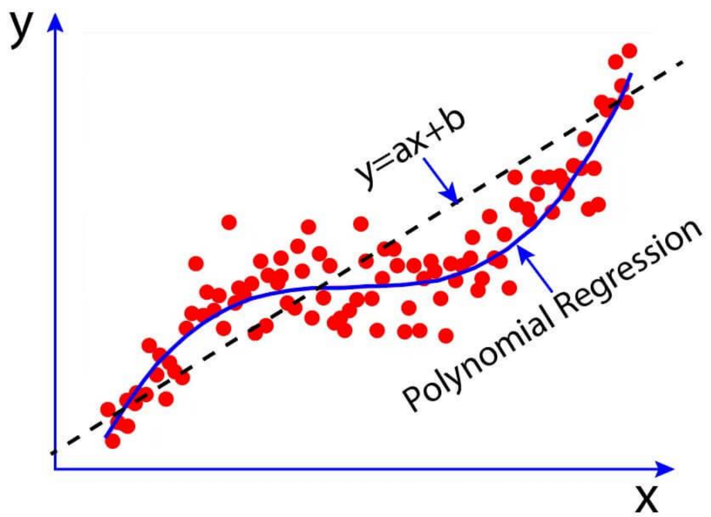
Polynomial regression is a type of regression analysis used to model the relationship between a dependent variable and one or more independent variables by fitting a polynomial equation to the data.



Why do we need it?

Let's consider a case of Simple Linear Regression.

We made our model and found out that it performs very poorly. We observe between the actual value and the best-fit line, which we predicted, and it seems that the actual value has some kind of curve in the graph, and our line is nowhere near cutting the mean of the points.
This is where polynomial Regression comes into play. It predicts the best-fit line that follows the pattern(curve) of the data, as shown in the pic below:



Step 5: Polynomial Regression

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

Polynomial Form: Instead of fitting a straight line (as in linear regression), polynomial regression fits a curve.
The equation takes the form: [y = a + b_1x + b_2x^2 + b_3x^3 + ⋯ ⋯ + b_nx^n]
where (y) is the dependent variable, (x) is the independent variable, and (a) and (b) are coefficients.

Higher-Degree Polynomials: By increasing the degree of the polynomial (e.g., quadratic, cubic), the model can capture more complex relationships in the data.

Use Cases: It's useful when the relationship between the variables is not linear, such as in cases of curvature or when the data shows increasing or decreasing trends.

Overfitting Risk: Higher-degree polynomials can lead to overfitting, where the model fits the noise in the data instead of the underlying trend.

Model Evaluation: Just like with linear regression, Polynomial Regression models can be evaluated using metrics such as **Mean Squared Error** and **R-squared** values to assess their performance and predictive power.

Step 5: Polynomial Regression

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

How can we build this?

1.Creating Polynomial Features:

```
polyfeat = PolynomialFeatures(degree=2)
```

the input features will be transformed to include all polynomial combinations of the features up to the second degree

2.Transforming Training and Testing Data:

```
xtrain_poly = polyfeat.fit_transform(train_data[features1])
xtest_poly = polyfeat.fit_transform(test_data[features1])
```

Here, the training set and test set features are transformed into polynomial features. Note that, ideally, you should fit on the training set and transform the test set separately to avoid data leakage.

3.Creating and fitting:

```
poly = linear_model.LinearRegression()
poly.fit(xtrain_poly, train_data['price'])
```

4.Prediction:

```
polypred = poly.predict(xtest_poly)
```

5.Evaluation:

```
mean_squared_error = metrics.mean_squared_error(test_data['price'], polypred)
print('Mean Squared Error (MSE) ', round(np.sqrt(mean_squared_error), 2))
print('R-squared (training) ', round(poly.score(xtrain_poly, train_data['price']), 3))
print('R-squared (testing) ', round(poly.score(xtest_poly, test_data['price']), 3))
```

Assignment requirement

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

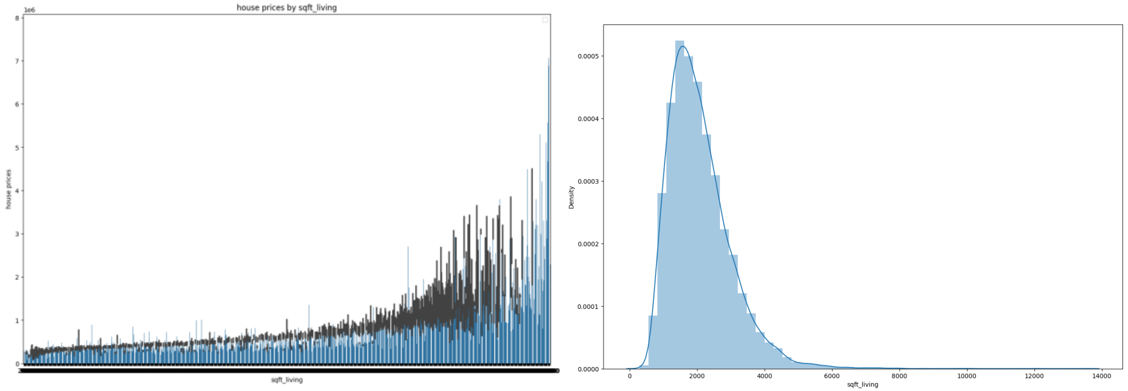
Q1 (**1 point**): count the number of occurrences of each unique value in "**condition.**"
Q2 (**1 point**): please draw a **bar plot** of **'house prices by sqft_above**' and a **density plot** of '**sqft_above**.'
Q3 (**1 point**): please draw a Simple Linear Regression plot of '**house prices by sqft_above.**'
Q4 (**1 point**): what is the **R-squared value** on **testing data** in **complex model 3**?
Q5 (**1 point**): which model is the **best** among the above **complex models**? Why?

Out[8]:

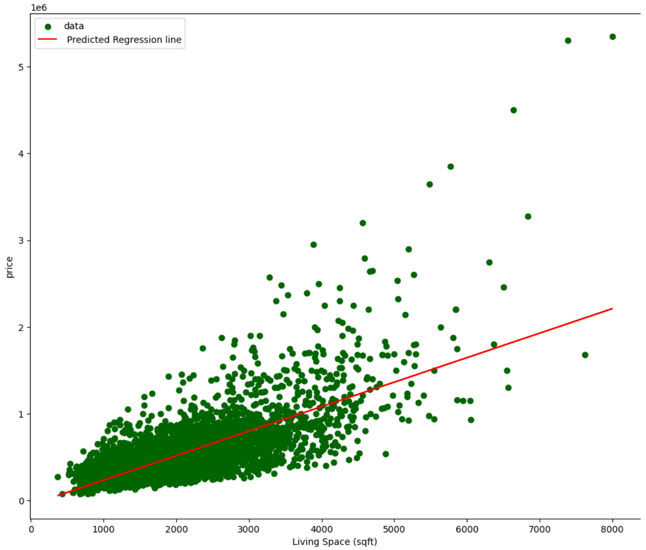
bedrooms	count
3	9824
4	6882
2	2760
5	1601
6	272
1	199
7	38
0	13
8	13
9	6
10	3
11	1
33	1

dtype: int64

Q1 sample



Q2 sample



Q3 sample

Assignment requirement

THE HONG KONG UNIVERSITY OFSCIENCE AND TECHNOLOGY

SCHOOL OF ENGINEERING

What you need to submit to Canvas is a PDF file named "Assignment 2 + your name".

名稱	修改日期	類型	大小
3.ipynb	28/9/2024 13:06	Jupyter 來源檔案	789 KB
3_full.ipynb	28/9/2024 12:59	Jupyter 來源檔案	901 KB
3_old.ipynb	18/9/2023 6:08	Jupyter 來源檔案	883 KB
Assignment_2.ipynb	28/9/2024 14:22	Jupyter 來源檔案	652 KB
Assignment_2_answer.docx	28/9/2024 14:38	Microsoft Word ...	212 KB
Assignment_2_requirement.docx	28/9/2024 14:18	Microsoft Word ...	180 KB
Assignment_2 HUANG Xuhong.pdf	7/10/2024 17:32	Microsoft Edge P...	181 KB

Assignment_2 HUANG Xuhong.pdf

Assignment 2:

Q1 (**1 point**): count the number of occurrences of each unique value in "**condition**."

Your code:

(Copy your core code here)

Your result:

(Screenshot your results here)

Q2 (**1 point**): please draw a **bar plot** of '**house prices by sqft_above**' and a **density plot** of '**sqft_above**.'

Your code:

(Copy your core code here)

Your result:

(Screenshot your results here)

Q3 (**1 point**): please draw a Simple Linear Regression plot of '**house prices by sqft_above**.'

Your code:

(Copy your core code here)

Your result:

(Screenshot your results here)

Q4 (**1 point**): what is the **R-squared value** on **testing data** in **complex model 3**?

Your answer:

(Write down your answer here)

Q5 (**1 point**): which model is the **best** among the above **complex models**? Why?

Your answer:

(Write down your answer here)

Assignment 2 template