

# 初阶-技术创意 2-实验一-实验报告

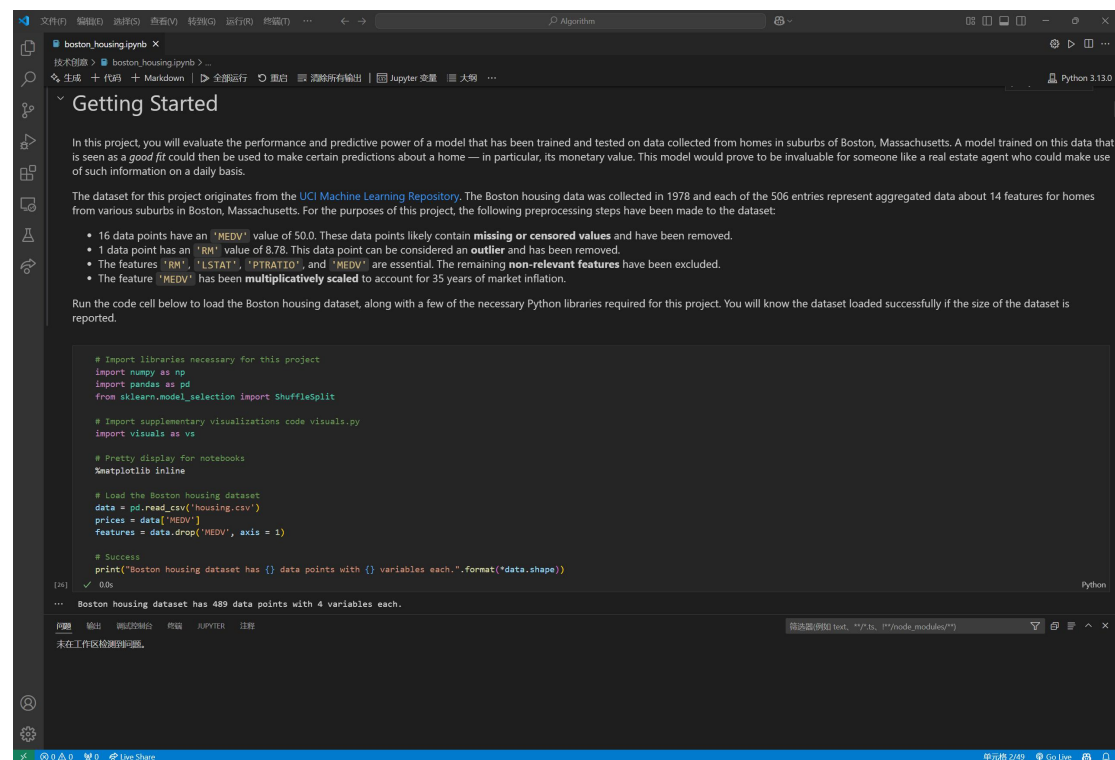
## (1) 实验目的

熟练掌握安装 *python* 和 *machine learning* 需要的环境，理解如何去对数据集进行分析。

## (2) 实验仪器/设备

vscode

## (3) 实验过程



The screenshot shows a Jupyter Notebook titled "boston\_housing.ipynb" in the VS Code editor. The notebook is in the "Getting Started" section, which provides an overview of the project and the dataset. The text describes the dataset's origin from the UCI Machine Learning Repository and lists preprocessing steps: removing 16 data points with a 'MEDV' value of 50.0, removing 1 outlier with an 'RM' value of 8.78, and excluding non-relevant features 'RM', 'LSTAT', and 'PTRATIO'. The 'MEDV' feature is noted as being multiplicatively scaled. A code cell below the text loads the dataset using pandas and sklearn, and prints the dataset's shape. The output of the code cell is displayed in the console, showing that the dataset has 489 data points with 4 variables each.

```
# Import libraries necessary for this project
import numpy as np
import pandas as pd
from sklearn.model_selection import ShuffleSplit

# Import supplementary visualizations code visuals.py
import visuals as vs

# Pretty display for notebooks
%matplotlib inline

# Load the Boston housing dataset
data = pd.read_csv('housing.csv')
prices = data['MEDV']
features = data.drop('MEDV', axis = 1)

# Success
print("Boston housing dataset has {} data points with {} variables each.".format(data.shape))
```

[24] ✓ 0.0s Python

... Boston housing dataset has 489 data points with 4 variables each.

文件的 编辑器 查看(V) 转到(G) 运行(R) 终端(T) ...

Algorithm

Python 3.13.0

boston\_housing.ipynb X

技术部直 > boston\_housing.ipynb > ...

生成 代码 + Markdown | 全部运行 重启 清除所有输出 | Jupyter 变量 大纲 ...

🔍

🔗

📁

📄

🔍

🔗

📁

📄

## Data Exploration

In this first section of this project, you will make a cursory investigation about the Boston housing data and provide your observations. Familiarizing yourself with the data through an explorative process is a fundamental practice to help you better understand and justify your results.

Since the main goal of this project is to construct a working model which has the capability of predicting the value of houses, we will need to separate the dataset into **features** and the **target variable**. The **features** "RM", "LSTAT", and "PTRATIO", give us quantitative information about each data point. The **target variable**, "MEDV", will be the variable we seek to predict. These are stored in **features** and **prices**, respectively.

### Implementation: Calculate Statistics

For your very first coding implementation, you will calculate descriptive statistics about the Boston housing prices. Since **numpy** has already been imported for you, use this library to perform the necessary calculations. These statistics will be extremely important later on to analyze various prediction results from the constructed model.

In the code cell below, you will need to implement the following:

- Calculate the minimum, maximum, mean, median, and standard deviation of "MEDV", which is stored in **prices**.
  - Store each calculation in their respective variable.

```
# TODO: Minimum price of the data
minimum_price = None
minimum_price = np.min(prices)
maximum_price = np.max(prices)
mean_price = np.mean(prices)
median_price = np.median(prices)
std_price = np.std(prices)

# # TODO: Maximum price of the data
# maximum_price = None

# # TODO: Mean price of the data
# mean_price = None

# # TODO: Median price of the data
# median_price = None

# # TODO: Standard deviation of prices of the data
# std_price = None

# Show the calculated statistics
print("Statistics for Boston housing dataset:\n")
print("Minimum price: {}".format(minimum_price))
print("Maximum price: {}".format(maximum_price))
print("Mean price: {}".format(mean_price))
print("Median price {}".format(median_price))
print("Standard deviation of prices: {}".format(std_price))
```

[27] ✓ 0.0s

Python

```
...
Statistics for Boston housing dataset:

Minimum price: $150000.0
Maximum price: $10248000.0
Mean price: $454342.9447852761
Median price $438900.0
Standard deviation of prices: $165171.13154429474
```

### Question 1 - Feature Observation

As a reminder, we are using three features from the Boston housing dataset: "RM", "LSTAT", and "PTRATIO". For each data point (neighborhood):

- "RM" is the average number of rooms among homes in the neighborhood.
- "LSTAT" is the percentage of homeowners in the neighborhood considered "lower class" (working poor).
- "PTRATIO" is the ratio of students to teachers in primary and secondary schools in the neighborhood.

\*\* Using your intuition, for each of the three features above, do you think that an increase in the value of that feature would lead to an **increase** in the value of "MEDV" or a **decrease** in the value of "MEDV"? Justify your answer for each.\*\*

**Hint:** This problem can be phrased using examples like below.

- Would you expect a home that has an "RM" value(number of rooms) of 6 be worth more or less than a home that has an "RM" value of 7?
- Would you expect a neighborhood that has an "LSTAT" value(percent of lower class workers) of 15 have home prices be worth more or less than a neighborhood that has an "LSTAT" value of 20?
- Would you expect a neighborhood that has an "PTRATIO" value(ratio of students to teachers) of 10 have home prices be worth more or less than a neighborhood that has an "PTRATIO" value of 15?

**Answer:** 一般来说, 房间更多的房屋面积更大, 价值也更高。因此, 邻里中房屋的平均房间数 ("RM") 增加, 应该会导致中位房价 ("MEDV") 增加。更大的房屋通常提供更多的空间、更多的设施和更理想的居住环境, 因此价格更高。邻里中低收入业主百分比 ("LSTAT") 较高可能意味着生活设施较差, 以及可能不太理想的居住环境。因此, "LSTAT" 增加应该会导致 "MEDV" 下降。较高的学生与教师比例可能意味着每个学生得到的个性化关注较少, 教育质量可能较低。因此, "PTRATIO" 增加应该会导致 "MEDV" 下降。

File Edit View Help Jupyter Notebook

Run Output Console Jupyter Notebook

未在工作区检测到此文档。

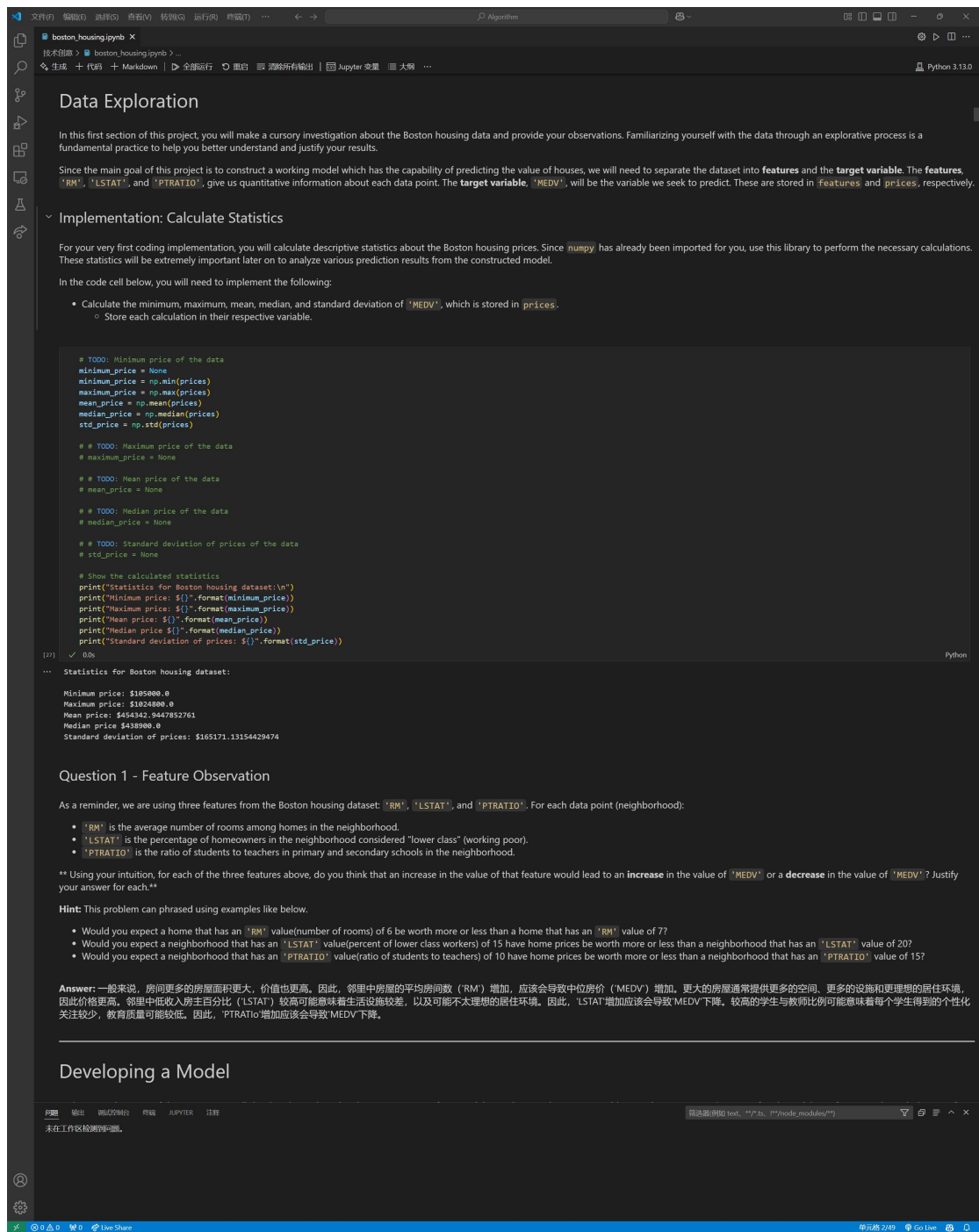
快速返回和文本, "Ctrl + J" 或 "mode\_modules()"

单元格 2/49 Go Live

## (4) 实验结果及结果分析

同上





## (5)实验总结

掌握安装 *python* 和 *machine learning* 需要的环境；提升了英语水平；了解了 *ipynb* 在 *vscode* 的使用；了解如何去对数据集进行分析。