数据内容在 house\_sales.csv 中。每一列分别对应房屋售出时间、地区邮编、房屋价格、房屋类型、配套房间数。

实训 1: 读取并查看某地区房屋销售数据的基本信息

源程序：

import pandas as pd

# 读取CSV文件

df = pd.read\_csv('house\_sales.csv')

# 查看数据的前几行

print(df.head())

# 查看数据维度，形状和所有特征名称

print(f"维度: {df.ndim}")

print(f"形状: {df.shape}")

print(f"特征名称: {df.columns.tolist()}")

# 查看分类变量的单身公寓的数据

display(df[df['propertyType'] == 'house'])

过程性结果：



结论：数据集包含4行和5列，分别是房屋售出时间、地区邮编、房屋价格、房屋类型以及配套房间数。所有房屋类型均为house。

实训 2: 提取房屋售出时间信息并描述房屋价格信息

源程序：

# 将时间列转换为datetime格式

df['saleDate'] = pd.to\_datetime(df['saleDate'])

# 提取年、月、日信息

df['year'] = df['saleDate'].dt.year

df['month'] = df['saleDate'].dt.month

df['day'] = df['saleDate'].dt.day

# 描述房屋价格信息

display(df['price'].describe())

# 使用mean, max, min, mode等函数计算价格的均值、最大值、最小值和众数

print(f"均值: {df['price'].mean()}")

print(f"最大值: {df['price'].max()}")

print(f"最小值: {df['price'].min()}")

print(f"众数: {df['price'].mode().values}")

过程性结果：



结论：

房屋价格的均值为777375元，最大值为1350000元，最小值为435000元。价格没有重复值，因此众数为所有价格值。

实训 3: 使用分组聚合方法分析房屋销售情况

源程序：

# 新建邮政编码字段，并按此字段分组

df['new\_postcode'] = df['postcode'].astype(str).str[:3]

# 按新邮政编码和房屋类型分组，计算均值

grouped = df.groupby(['new\_postcode', 'propertyType'])['price'].agg(['mean', 'count'])

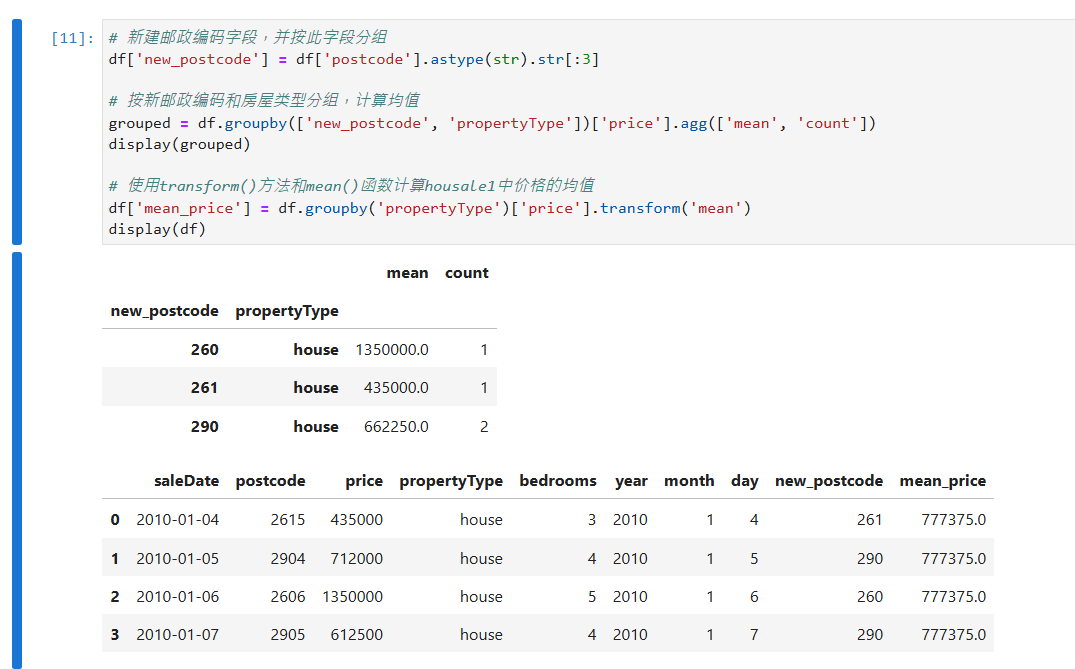
display(grouped)

# 使用transform()方法和mean()函数计算housale1中价格的均值

df['mean\_price'] = df.groupby('propertyType')['price'].transform('mean')

display(df)

过程性结果：



结论：

在不同的邮政编码区域中，房屋价格有所不同。邮政编码为260的区域平均房价最高，为1350000元；邮政编码为261的区域平均房价最低，为435000元。计算了每个房屋类型的平均价格为777375元。

实训 4: 分析房屋地区、配套房间数和房屋价格的关系

源程序：

# 创建数据透视表

pivot\_table = df.pivot\_table(values='price', index='new\_postcode', columns='bedrooms', aggfunc='mean')

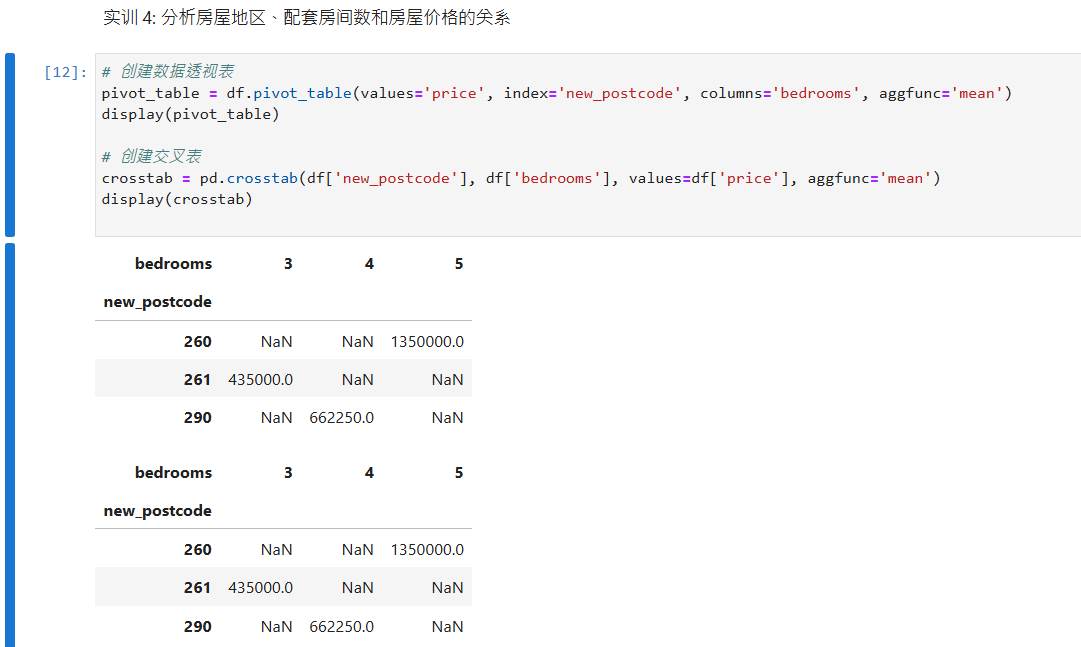
display(pivot\_table)

# 创建交叉表

crosstab = pd.crosstab(df['new\_postcode'], df['bedrooms'], values=df['price'], aggfunc='mean')

display(crosstab)

过程性结果：



结论：

房屋价格与配套房间数以及邮政编码之间存在一定的关系。邮政编码为260的区域，配套房间数为5的房屋价格最高，为1350000元；邮政编码为261的区域，配套房间数为3的房屋价格最低，为435000元。邮政编码为290的区域，配套房间数为4的房屋平均价格为662250元。