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：

实训 1: 读取并查看某地区房屋销售数据的基本信息

[9]:
```
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'])
```

```
        saleDate  postcode    price propertyType  bedrooms
0  2010/1/4 0:00      2615   435000        house         3
1  2010/1/5 0:00      2904   712000        house         4
2  2010/1/6 0:00      2606  1350000        house         5
3  2010/1/7 0:00      2905   612500        house         4
维度: 2
形状: (4, 5)
特征名称: ['saleDate', 'postcode', 'price', 'propertyType', 'bedrooms']
```

	saleDate	postcode	price	propertyType	bedrooms
0	2010/1/4 0:00	2615	435000	house	3
1	2010/1/5 0:00	2904	712000	house	4
2	2010/1/6 0:00	2606	1350000	house	5
3	2010/1/7 0:00	2905	612500	house	4

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:

实训 2: 提取房屋售出时间信息并描述房屋价格信息

[13]:
```
# 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}")
```

```
count    4.000000e+00
mean     7.773750e+05
std      3.985715e+05
min      4.350000e+05
25%      5.681250e+05
50%      6.622500e+05
75%      8.715000e+05
max      1.350000e+06
Name: price, dtype: float64
均值: 777375.0
最大值: 1350000
最小值: 435000
众数: [ 435000  612500  712000 1350000]
```

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:

[11]:
```
# 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)
```

		mean	count
new_postcode	**propertyType**		
260	**house**	1350000.0	1
261	**house**	435000.0	1
290	**house**	662250.0	2

	saleDate	postcode	price	propertyType	bedrooms	year	month	day	new_postcode	mean_price
0	2010-01-04	2615	435000	house	3	2010	1	4	261	777375.0
1	2010-01-05	2904	712000	house	4	2010	1	5	290	777375.0
2	2010-01-06	2606	1350000	house	5	2010	1	6	260	777375.0
3	2010-01-07	2905	612500	house	4	2010	1	7	290	777375.0

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')

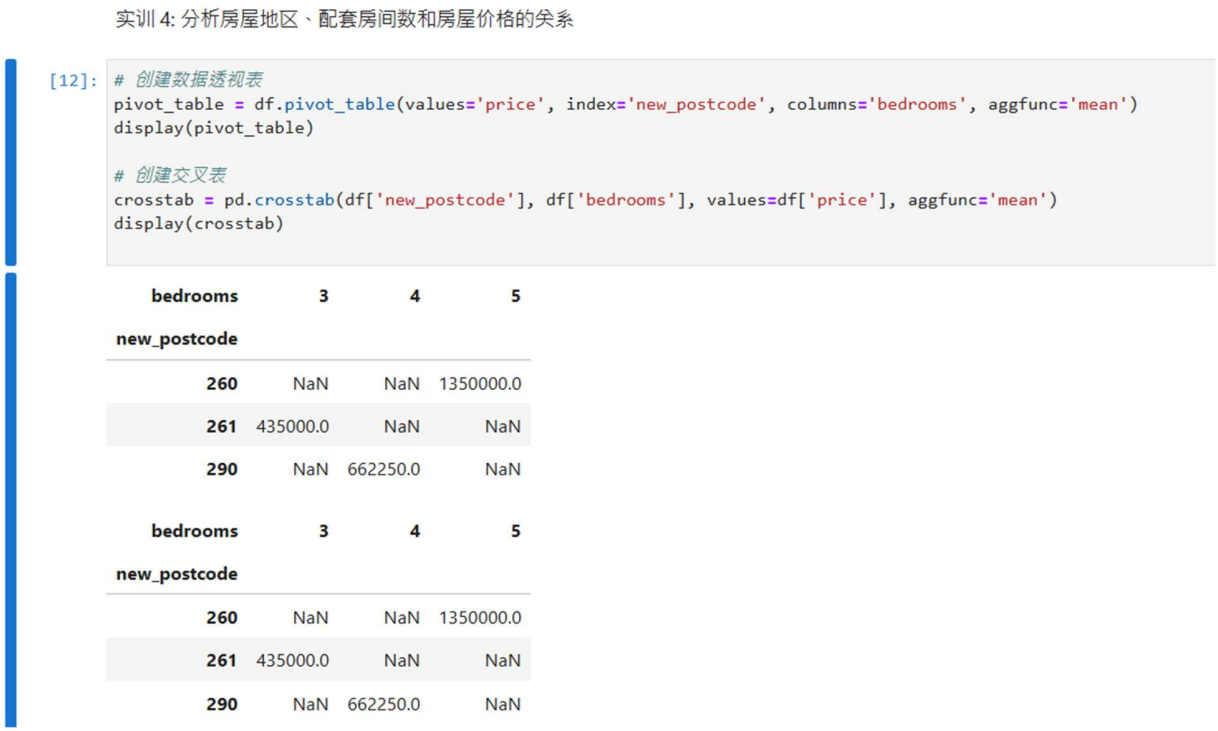
display(pivot_table)

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。