泰迪智能科技
TIPDM INTELLIGENT TECHNOLOGY

第6章 房屋租金影响因素分析与预测

引言

- 随着信息化的发展和科学技术的进步，数据分析与挖掘技术开始得到广泛应用。人们无时无刻不面对着海量的数据，这些海量数据中隐藏着人们所需要的具有决策意义的信息。数据分析与挖掘技术的产生和发展就是帮助人们来利用这些数据并从中发现有用的，隐藏的信息。
- 在此背景下，本文主要运用数据分析与挖掘技术对房屋租金影响因素进行分析，了解不同因素对房屋价格的影响，并根据房屋信息预测房屋租金价格，希望能够给租房者提供一定的参考。

目录

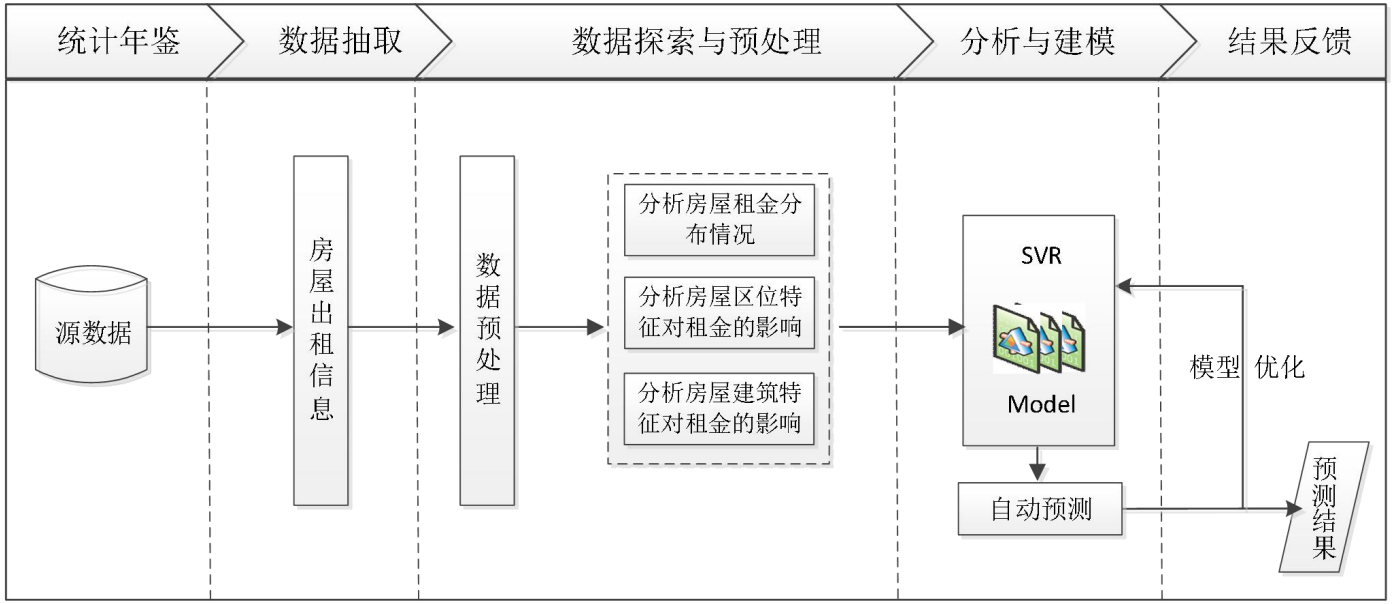
背景与挖掘目标

- 近年来，随着城市化进程的加快，越来越多的外来人口涌入城市，对租赁市场带来了较大的需求。然而，房屋租金受多种因素的影响，如供需关系、经济发展水平、政策法规等。为了更好地了解房屋租金的变化趋势，本文将采用数据分析与挖掘的方法进行预测和分析。
- 结合房屋租金影响因素分析与预测的需求分析，本次数据分析建模目标主要有以下2个。
  - 分析不同因素对房屋价格的影响。
  - 根据房屋信息预测房屋租金价格。

目录

分析方法与过程

- 本案例的总体流程如图所示，主要包括以下步骤。
  - 对原始数据进行预处理分析，提高数据质量。
  - 利用步骤1形成的数据进行数据探索分析，分析不同因素对房屋租金的影响。
  - 利用步骤1形成的数据进行构造属性，并提取关键属性。
  - 利用步骤3的数据建立支持向量回归预测模型，并对模型进行评价。



数据预处理

- 本案例的数据预处理主要包含查看数据情况、检测与处理重复值、检测与处理异常值、检测与处理缺失值、特征变换等。
- 某企业收集了某市房屋出租的相关信息，包含某主流住房租赁平台的房屋的地点、价格信息及一份地铁位置数据等，各项属性名称如表所示。

表名	属性名称	属性说明
houses	行政区	房子的行政区id
	租房类型	租房类型，包括whole、shared。其中whole表示整租；shared表示合租
	房屋面积/平方米	房子的面积大小
	租房价格/元	房子租金价格
	每平方米租金价格/（元/平方米）	每平方米租金价格
	楼层	房子楼层位置（L：低、M：中、H：高）
	最高楼层数	房屋所在建筑物的最高楼层数
	是否有电梯	建筑物是否有电梯(有，无)
	社区ID	房屋所在社区id
	地理位置	社区的地理位置
metro	LineCode	地铁系统官方线路代码
	StationCode	地铁线路的官方车站代码
	Geometry	站点的地理位置

数据预处理

1.查看数据情况

- 由代码运行结果可知：可以得到houses表和metro表各属性的基本情况。根据表可知，houses表中“楼层”“最高楼层”“是否有电梯”3个属性中存在缺失值，metro表中不存在缺失值、重复值。

#	Column	Non-Null Count	Dtype
0	行政区	11128 non-null	object
1	租房类型	11128 non-null	object
2	房屋面积/平方米	11128 non-null	int64
3	租房价格/元	11128 non-null	int64
4	每平方米租金价格/（元/平方米）	11128 non-null	int64
5	楼层	11125 non-null	object
6	最高楼层数	11127 non-null	float64
7	是否有电梯	11123 non-null	float64
8	社区ID	11128 non-null	int64
9	地理位置	11128 non-null	object

#	Column	Non-Null Count	Dtype
0	LineCode	274 non-null	object
1	StationCode	274 non-null	int64
2	Geometry	274 non-null	object

数据预处理

2.重复值检测与处理

- 对houses表进行重复值检测与处理。
- 由代码运行结果可知：houses表中存在406个重复值，并对重复值进行删除处理。

数据预处理

3.异常值检测与处理

- 对houses表中的类别型属性、数值型属性进行描述性分析，并绘制箱型图对数值型数据进行分析。
- 由代码运行结果可知：类别型属性和数值型属性的基本情况。

	行政区	租房类型	楼层	是否有电梯	社区ID
count	10722	10722	10719	10717.0	10722
unique	11	2	48	2.0	2964
top	城区D	Whole	M	1.0	498
freq	2624	10393	3806	7283.0	55

数据预处理

3.异常值检测与处理

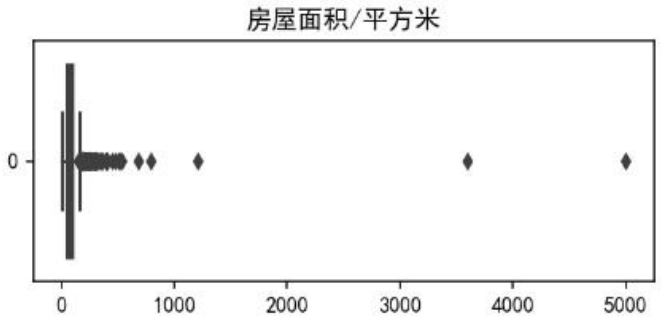
- 由表可知，“楼层”属性表示楼层位置信息（分为高中低3类），描述性统计结果显示“楼层”属性存在48个类别，可能是数据中存在具体数字楼层；
- “最高楼层数”属性中存在0层或为空白信息的情况，不符合实际情况；“房屋面积/平方米”属性中存在房子面积大小小于5平方米的情况，不符合某市房屋租赁管理规定，并且数据中存在异常离群值；“租房价格/元 ”属性和“每平方米租金价格/（元/平方米） ”属性中也存在一定的异常离群值。

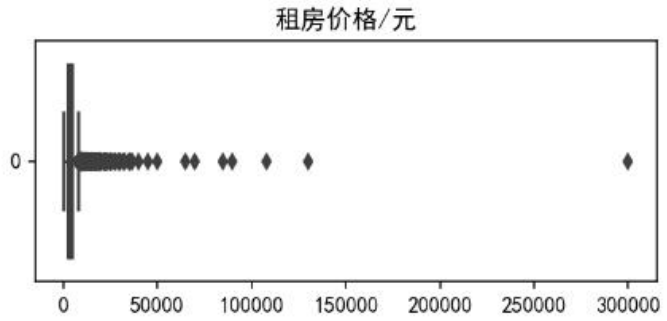
房屋面积/平方米	每平方米租金价格/（元/平方米）	最高楼层数	租房价格/元
10722.000000	10722.000000	10721.000000	10722.000000
76.432195	60.090935	19.240276	4060.263197
71.296029	78.034560	10.907930	4504.913339
1.000000	1.000000	0.000000	400.000000
50.000000	38.000000	9.000000	2500.000000
73.000000	53.000000	18.000000	3500.000000
95.000000	71.000000	29.000000	4700.000000
5000.000000	5600.000000	62.000000	300000.000000

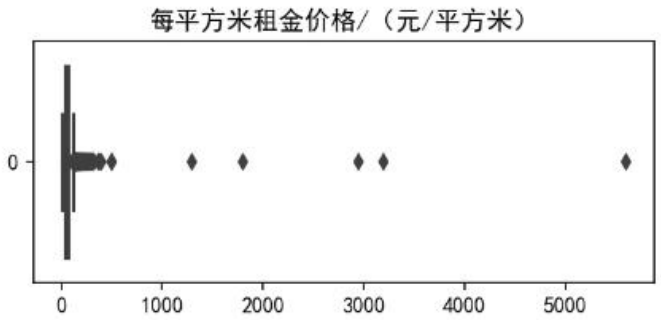
数据预处理

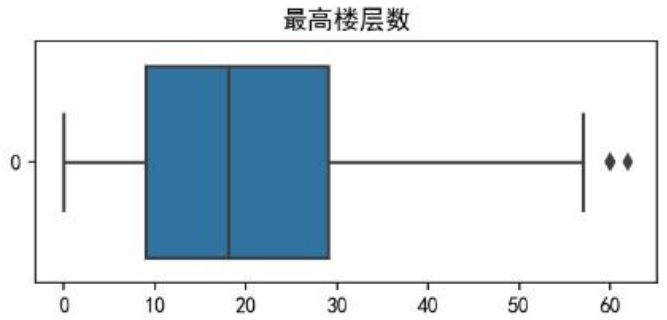
3.异常值检测与处理

- 运行得到的箱线图，如图所示。









数据预处理

3.异常值检测与处理

- 根据的运行结果得到异常值的处理原则如下。
  - 删除“最高楼层数”属性中为0或为空的数据。
  - 提取“楼层”属性中数值型数据，当楼层数小于等于最高楼层数量的1/3时，标记为“L”；当楼层数小于等于最高楼层数量的2/3时，标记为“M”；等楼层数大于最高楼层数量的2/3时，标记为“H”。
  - 删除“房屋面积/平方米”属性中面积小于5的数据。
  - 使用3sigma原则删除“房屋面积/平方米”“租房价格/元 ”“每平方米租金价格/（元/平方米）”“最高楼层”属性中的异常数据。

数据预处理

5.数据变换

为了满足后续分析和算法的需求，需要对数据进行变换，包含地理位置变换、构建距离属性、“是否有电梯”数据变换等。

（1）地理位置变换

由于hourses表、metro表中所给社区地理位置信息与地铁位置信息为字符型的“度分秒”经纬度信息，难以进行分析，所以对该列数据进行拆分，拆解为经纬度两列数据，并将度分秒数据利用正则提取数字部分进而转换为度数据，将其整理为列表格式以便后续距离计算。

数据预处理

5.数据变换

为了满足后续分析和算法的需求，需要对数据进行变换，包含地理位置变换、构建距离属性、“是否有电梯”数据变换等。

（2）构建距离属性

使用整理所得经纬度数据计算社区位置离最近地铁的直线距离，构建距离指标，并获得最近地铁的经纬度信息。

（3）“是否有电梯”数据变换

由于原始数据所给分类数据的名称是以数字编码指代，不方便后续探索性分析，因此将对“是否有电梯”进行数据变换。

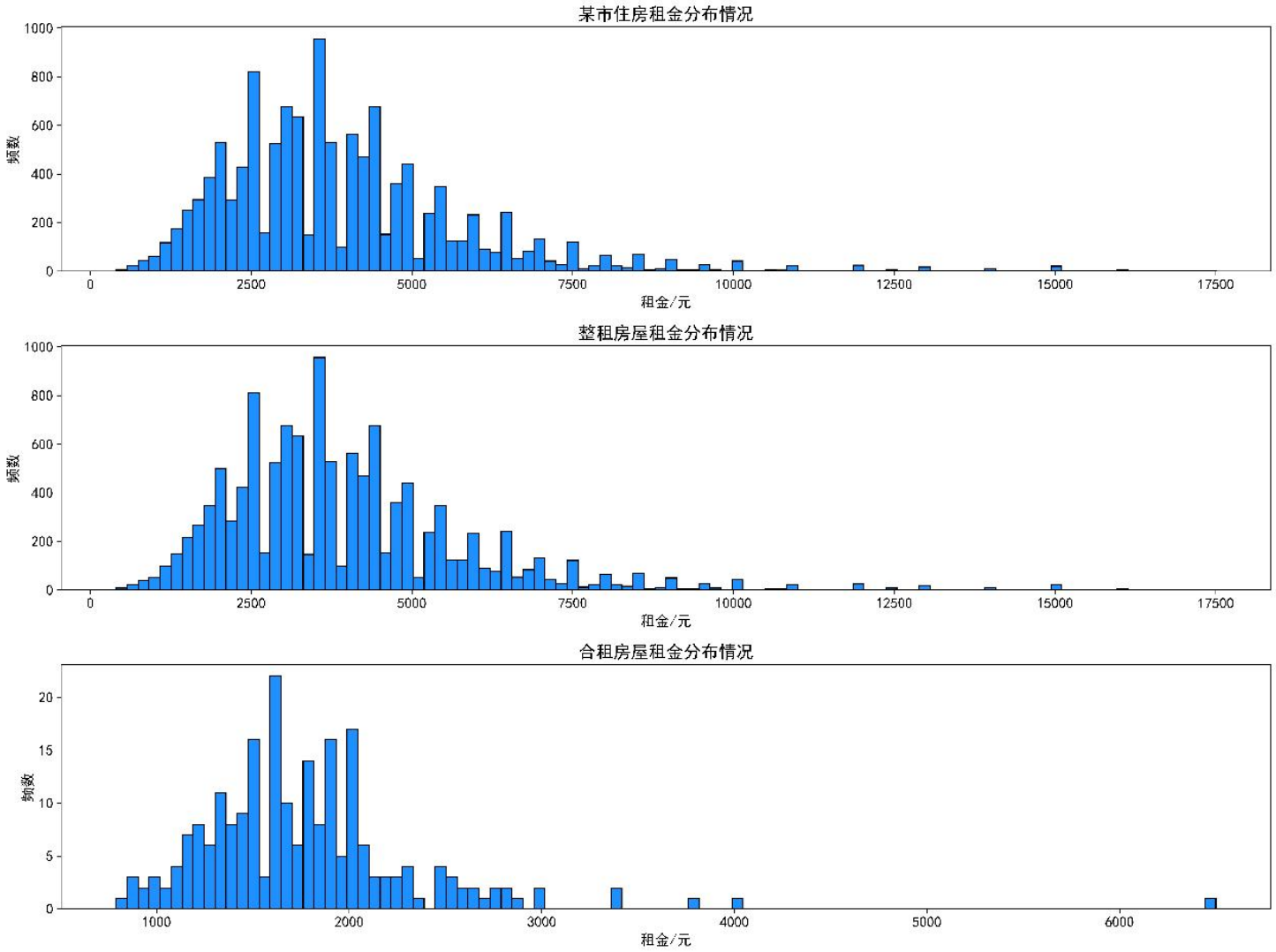
数据探索

- 通过对数据的探索性分析，可以获取数据的基本概况，发现数据间的关联关系，揭示不同变量之间的关系，提取隐藏在数据中的规律和趋势，为后续的建模和预测提供指导。
- 本小节主要探索的内容有房屋租金分布情况、房屋区位特征对租金的影响、房屋建筑特征对租金的影响。

数据探索

1.分析房屋租金分布情况

- 选取房屋价格、整租方式的房屋租金、合租方式的房屋租金进行探索分析，探索房屋租金分布情况。
- 由代码运行结果可知：得到整体房屋租金分布情况、整租方式的房屋租金分布情况、合租方式的房屋租金分布情况，如图所示。



数据探索

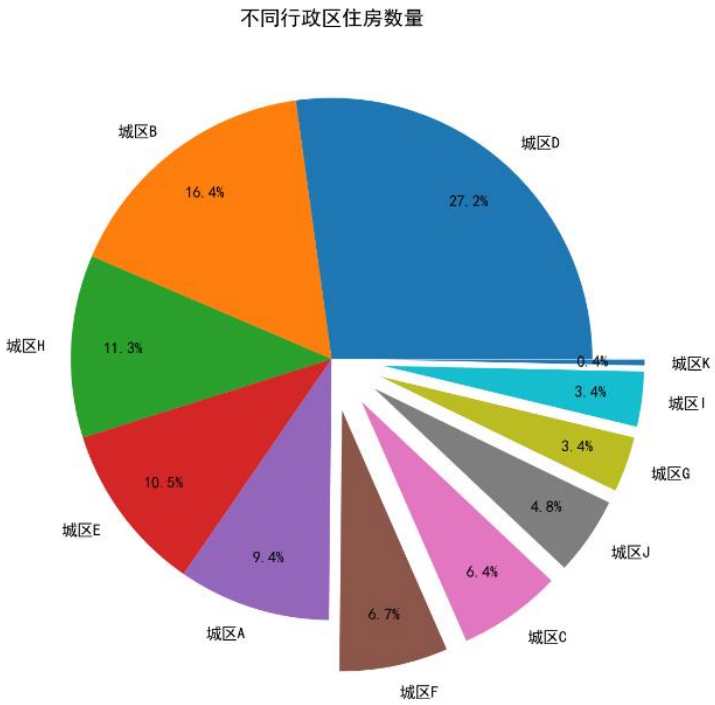
1.分析房屋租金分布情况

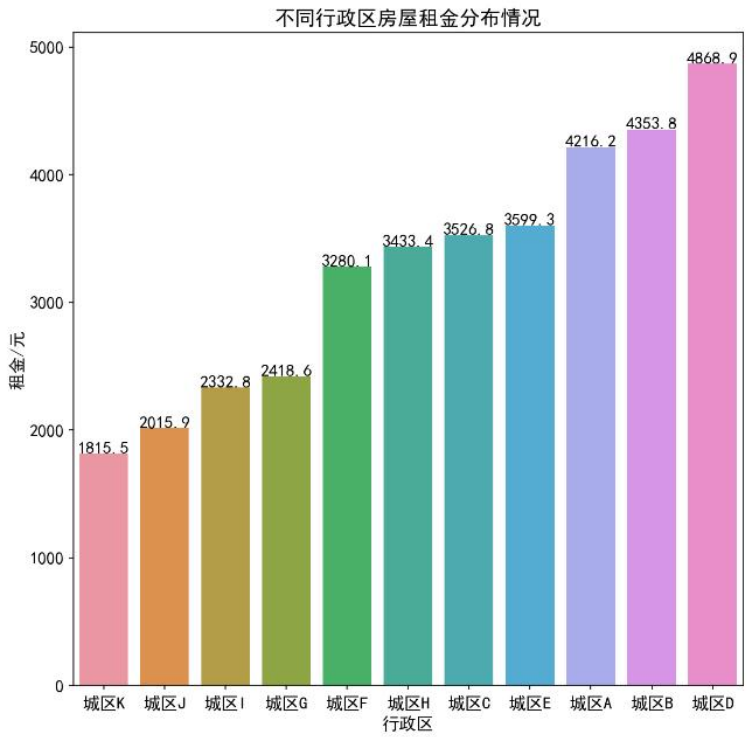
- 由图可知，住房租金呈现右偏分布，租金在1000-5000元附近分布得最为密集。两种不同租赁方式的租金分布范围具有明显的差异。
- 整租住房的租金区间主要分布在2000-5000元，而合租住房的租金区间主要分布在1000-2000元，合租房源的整体租金水平低于整租房源。

数据探索

2.分析房屋区位特征对租金的影响

- 由代码运行可知：不同房屋行政区对租金的影响，如图。
- 由图可知，城区D、城区B、城区H、城区E的房源数量较多；城区D、城区B、城区A的租金较高，平均租金均超过4000元。

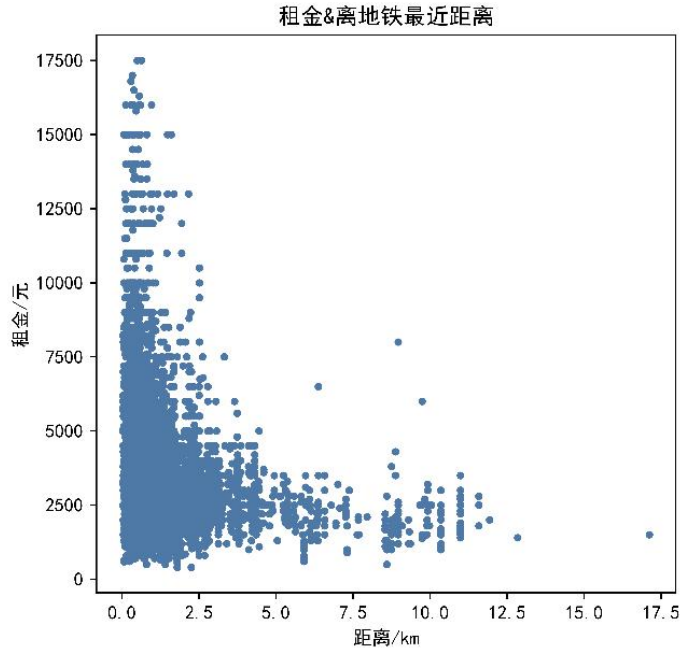


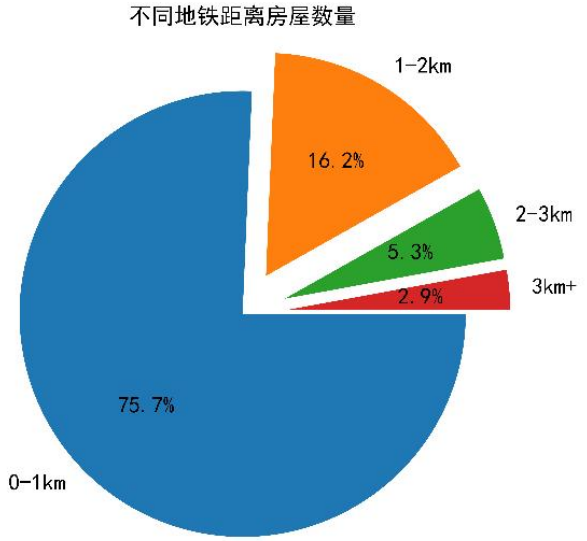


数据探索

2.分析房屋区位特征对租金的影响

- 由代码运行可知：选取距离地铁的距离、租房价格，探索地铁距离与租金的关系。
- 由图可知，地铁距离对租金有影响,社区与地铁距离小于1km的房屋数量最多，占比75.7%，同时对应的平均租金最高。当距离高于3km时，房屋需求量较少，所以相对租金较低。



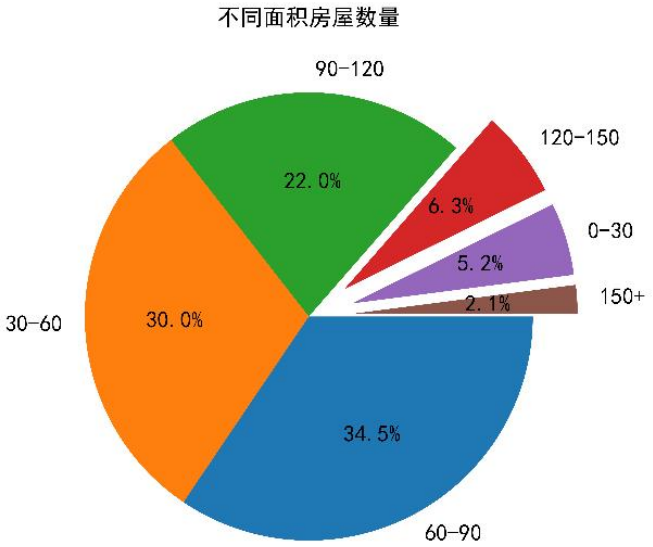


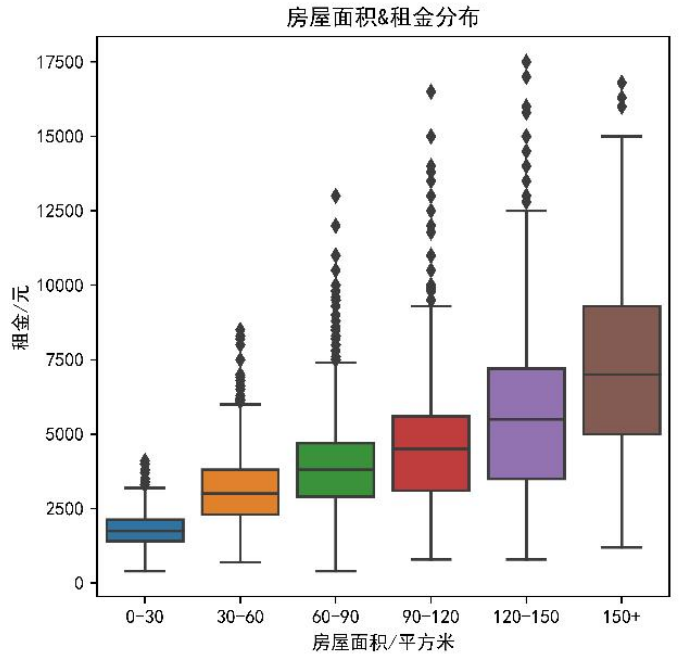
数据探索

3.分析房屋建筑特征对租金的影响

（1）分析不同房屋面积与租金的关系

选取房屋面积、租房价格，探索不同房屋面积对租金的影响，由图可知，86.5%的房屋面积集中为30-120平方米之间，符合租客整租与合租的情况。房屋面积与租金分布呈阶梯性，随着房屋面积的增大，房屋平均租金也上升。



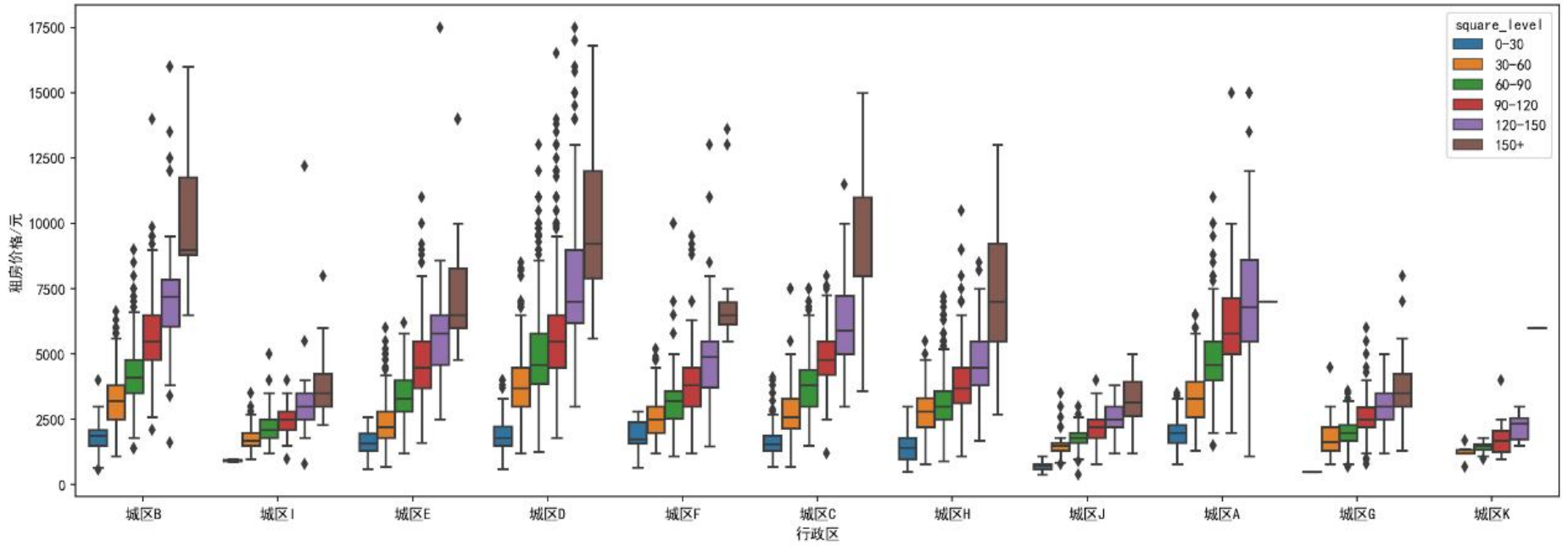


数据探索

3.分析房屋建筑特征对租金的影响

（2）分析不同行政区房屋面积与租金的关系

由代码运行可知：选取行政区、租房价格，探索不同行政区房屋面积对租金的影响。由图可知，不同行政区下房屋面积越大对应的房屋租金越高，市中心区域及附近辖区房屋租金高于远离市中心区域的房屋。



数据探索

3.分析房屋建筑特征对租金的影响

（3）分析不同租房类型与租金的关系

由代码运行可知：选取租房类型、租房价格，探索不同租房类型与租金的关系。由图可知，98.2%为整租房屋，仅1.8%为合租房屋。整租类型的租房价格比合租类型的高，且整租类型的租金范围也大于合租类型的住房。

整租类型房屋由于提供了更大的空间独立性、个人便利性以及配套设施的完善多样化，因此租金更高；而合租房屋由多个租客合租，空间共享性强，因此租金相对较低。

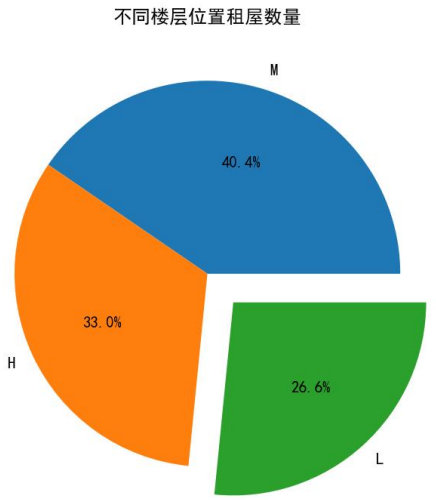
数据探索

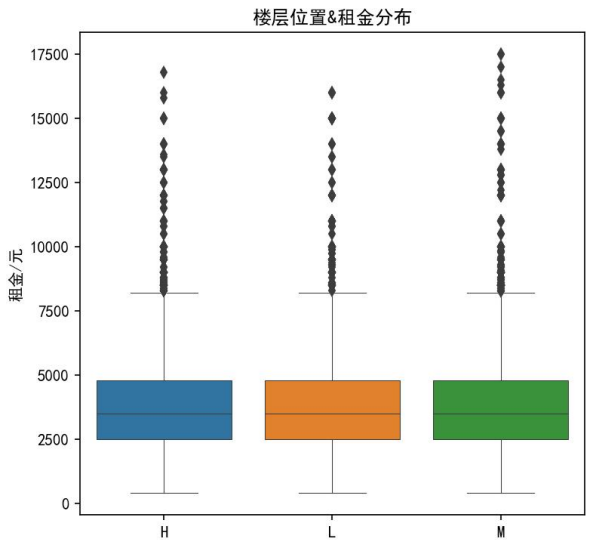
3.分析房屋建筑特征对租金的影响

（4）分析不同楼层位置与租金的关系

由代码运行可知：选取楼层、租房价格，探索不同楼层位置与租金的关系。

由图可知，不同楼层位置对房屋租金没有实质性的影响，三种楼层下房屋租金分布相似，这主要是因为不同楼层有不同的优缺点，低楼层比较方便出行但容易受噪声影响，高楼层出行不便但视野宽阔，相较于关注房屋所在楼层位置，租赁者更关心房屋的其他特征。





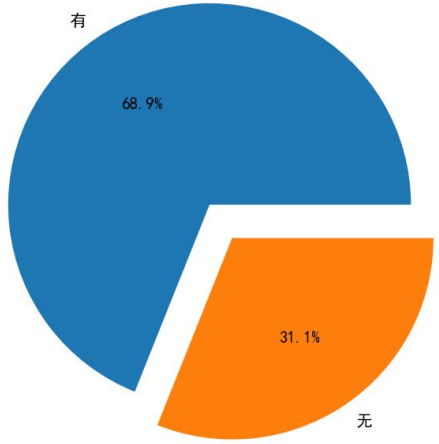
数据探索

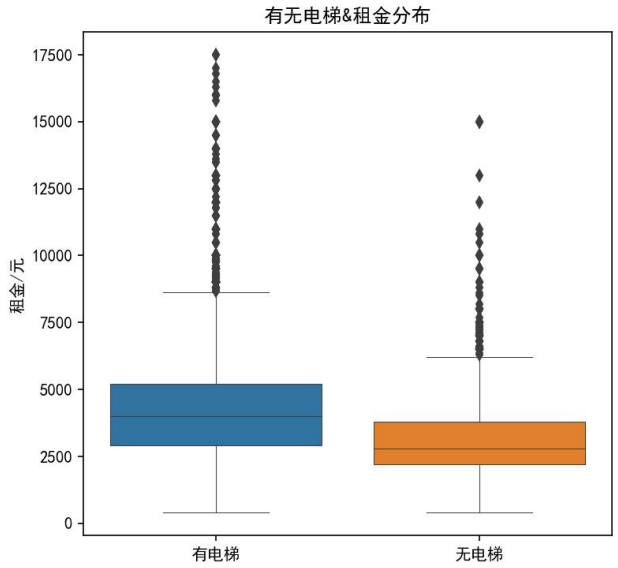
3.分析房屋建筑特征对租金的影响

（5）分析有无电梯与租金的关系

由代码运行可知：选取是否有电梯、租房价格，探索有无电梯与租金的关系。

由图可知，有电梯的租房数量相对较多，占总体租房数量的68.8%，并且有无电梯对于住房租金有较大的影响，有电梯的租屋价格明显高于无电梯租屋，无电梯租屋的楼层主要集中为低楼层，有电梯租屋相对而言出行更方便，因此租金也更高。



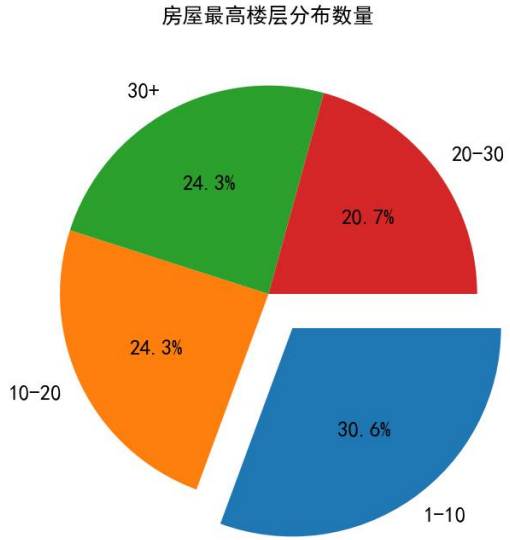


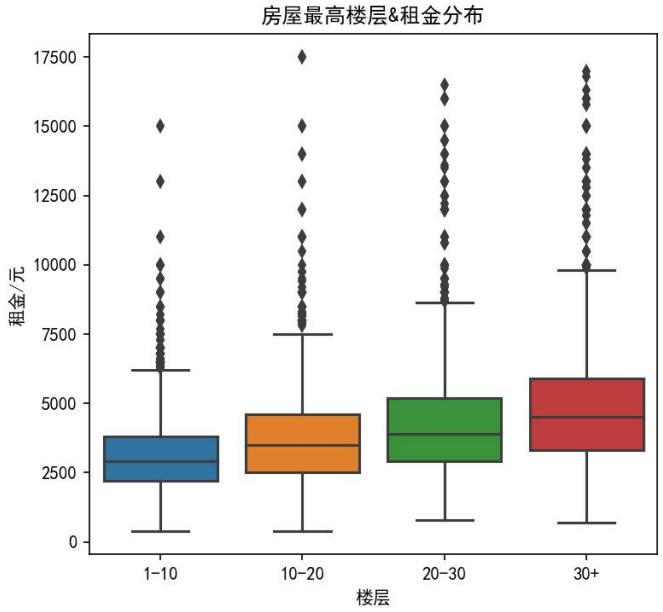
数据探索

3.分析房屋建筑特征对租金的影响

（6）分析房屋最高楼层与租金的关系

由代码运行可知：选取最高楼层、租房价格，探索房屋最高楼层与租金的关系。由图可知，建筑楼层高度对房屋租金有一定的影响。30.6%房屋所在建筑最高楼层在10层以内，同时有24.3%房屋所在建筑最高楼层在30层以上，楼层高度与租金呈阶梯形，数据上显示最高楼层数越高的房屋平均租金也越高。





模型构建

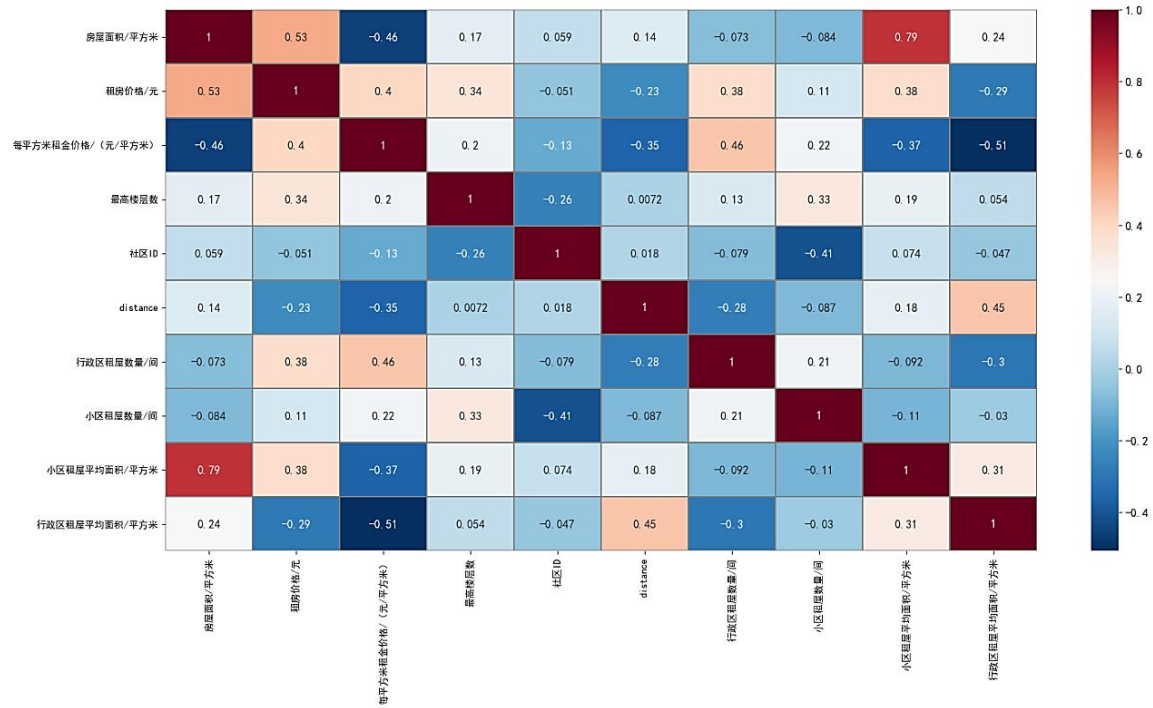
1.属性构造

- 由代码运行可知：为了提取更有用的信息，提高挖掘结果的精度，需要利用已有的属性构造新的属性。
- 本小节主要构造的属性包括不同行政区的租屋数量、不同小区的租屋数量、不同小区的租屋平均面积、不同小区的租屋平均面积、不同行政区的租屋平均面积。

模型构建

1.属性构造

- 由图可知，租房价格与房屋面积、每平方米租房价格、最高楼层、行政区租屋数量、小区租屋平均面积、行政区租屋平均数量的相关性均在0.3左右，没有很强的线性相关关系，可以说明房屋租金的价格是由多个因素影响，而不由单一特征决定，因此尽管线性相关性不高，该特征对于租金也可能存在一定影响。



模型构建

2.SVR算法

- SVR（Support Vector Regression，支持向量回归）是在做拟合时，采用了支持向量的思想，来对数据进行回归分析。给定训练数据集 $T=\{(\vec{x}_1,y_1),(\vec{x}_2,y_2),\ldots,(\vec{x}_n y_n)\}$，其中 $\vec{x}_1=(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(n)})^{\mathrm{T}}\in\mathbf{R}^n$，$y_i\in\mathbf{R}$，i=1,2,,n。对于样本通常根据模型输出 $f(\vec{x}_i)$与真实值yi之间的差别来计算损失，当且仅当 $f(\vec{x}_i)=y_i$ 时损失才为零。

模型构建

2.SVR算法

- SVR的基本思路是：允许 $f(\vec{x}_i)$ 与yi之间最多有的偏差。仅当 $f(\vec{x}_i)=y_i\ |>\varepsilon$ 时，才计算损失。当 $\left|f(\vec{x}_i)=y_i\right|\leqslant\varepsilon$ 时，认为预测准确。用数学语言描述SVR问题。

$$\min_{\vec{w},b}\frac{1}{2}\left\|\vec{w}\right\|^2+\mathrm{C}\sum_{i=1}^{n}L_\varepsilon(f(\vec{x}_i)-y_i)$$

- 其中 $\mathrm{C}\geqslant 0$ 为罚项系数，L为损失函数。

模型构建

2.SVR算法

- 更进一步，引入松弛变量i、 $\hat{\xi}_i$ ，则新的最优化问题如式所示。

$$\min_{\vec{w},b,\xi,\hat{\xi}} \frac{1}{2}\left\|\vec{w}\right\|^2 + \mathrm{C}\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i)$$

$$\begin{cases} s.t. f(\vec{x}_i) - y_i \leqslant \varepsilon + \xi_i \\ y_i - f(\vec{x}_i) \leqslant \varepsilon + \hat{\xi}_i \\ \xi_i \geqslant 0, \hat{\xi}_i \geqslant 0, i = 1,2,\ldots,n \end{cases}$$

模型构建

2.SVR算法

- 这就是SVR原始问题。类似的，引入拉格朗日乘子，$\mu_i \geqslant 0, \hat{\mu}_i \geqslant 0, \alpha_i \geqslant 0, \hat{\alpha}_i \geqslant 0$，定义拉格朗日函数如式所示。

$$L(\vec{w}, b, \vec{\alpha}, \hat{\alpha}, \vec{\xi}, \hat{\xi}, \vec{\mu}, \hat{\mu}) = \frac{1}{2}\left\|\vec{w}\right\|^2 + \mathrm{C}\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) - \sum_{i=1}^{n}\mu_i\xi_i - \sum_{i-1}^{n}\hat{\mu}_i\hat{\xi}_i +$$
$$\sum_{i=1}^{n}\alpha_i(f(\vec{x}_i) - y_i - \varepsilon - \xi_i) + \sum_{i-1}^{n}\hat{\alpha}_i(y_i - f(\vec{x}_i) - \varepsilon - \hat{\xi}_i)$$

- 根据拉格朗日对偶性，原始问题的对偶问题是极大极小问题。

$$\max_{\vec{\alpha}, \hat{\alpha}} \min_{\vec{w}, b, \vec{\xi}, \hat{\xi}} L(\vec{w}, b, \vec{\alpha}, \hat{\alpha}, \vec{\xi}, \hat{\xi}, \vec{\mu}, \hat{\mu})$$

模型构建

2.SVR算法

- 先求极小问题：根据 $L(\vec{w},b,\vec{\alpha},\hat{\alpha},\vec{\xi},\hat{\xi},\vec{\mu},\hat{\mu})$ 对 $\vec{w}$ 、b、$\vec{\xi}$ 、$\hat{\xi}$ 偏导数

$$\begin{cases}\vec{w}=\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i)\vec{x}_i \\ 0=\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i) \\ C=\alpha_i+\mu_i \\ C=\hat{\alpha}_i+\mu_i\end{cases}$$

模型构建

2.SVR算法

- 再求极大问题（取负号变极小问题）如式所示。

$$\min_{\vec{\alpha},\hat{\vec{\alpha}}} \sum_{i=1}^{n}\left[y_i(\hat{\alpha}_i-\alpha_i)-\varepsilon(\hat{\alpha}_i+\alpha_i)\right]+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{\alpha}_i-\alpha_i)(\hat{\alpha}_j-\alpha_j)\vec{\boldsymbol{x}}_i^{\mathrm{T}}\vec{\boldsymbol{x}}_j$$

$$\begin{cases} s.t.\sum_i^n(\hat{\alpha}_i-\alpha_i)=0 \\ 0\leqslant\alpha_i,\hat{\alpha}_i\leqslant C \end{cases}$$

模型构建

2.SVR算法

- *KKT*条件如式所示。

$$\begin{cases} \alpha_i(f(\vec{\boldsymbol{x}}_i)-y_i-\varepsilon-\xi_i)=0 \\ \alpha_{i(}y_i-f(\vec{\boldsymbol{x}}_i)-\varepsilon-\hat{\xi}_i)=0 \\ \alpha_i\hat{\alpha}_i=0 \\ \xi_i\hat{\xi}_i=0 \\ (\mathrm{C}-\alpha_i)\xi_i=0 \\ (\mathrm{C}-\hat{\alpha}_i)\hat{\xi}_i=0 \end{cases}$$

模型构建

2.SVR算法

- 假设最终解为 $\vec{\alpha}^* = (\alpha_1^* + \alpha_2^* + \cdots + \alpha_n^*)^{\mathrm{T}}$，在 $\hat{\vec{\alpha}} = (\hat{\alpha}_1^* + \hat{\alpha}_2^* + \cdots + \hat{\alpha}_n^*)^{\mathrm{T}}$ 中，找出 $\vec{\alpha}^*$ 的某个分量

$$\mathrm{C} > \alpha_j^* > 0$$

$$b^* = y_i + \varepsilon - \sum_{i=1}^{n} (\hat{\alpha}_i^* + \alpha_j^*) \vec{x}_i^{\mathrm{T}} \vec{x}_j$$

$$f(\vec{x}) = \sum_{i=1}^{n} (\hat{\alpha}_i^* + \alpha_i^*) \vec{x}_i^{\mathrm{T}} \vec{x} + b^*$$

模型构建

2.SVR算法

- 更进一步，如果考虑使用核技巧，给定核函数 $K(\vec{x}_i, \vec{x})$，则SVR可以表示为如式所示。

$$f(\vec{x}) = \sum_{i=1}^{n} (\hat{\alpha}_i - \alpha_i) K(\vec{x}_i, \vec{x}) + b$$

模型构建

2.SVR算法

- 由于支持向量机拥有完善的理论基础和良好的特性，人们对其进行了广泛的研究和应用，涉及分类、回归、聚类、时间序列分析、异常点检测等诸多方面。
- 具体的研究内容包括统计学习理论基础、各种模型的建立、相应优化算法的改进以及实际应用。支持向量回归也在这些研究中得到了发展和逐步完善，已有许多富有成果的研究工作。

模型构建

2.SVR算法

- 相比较于其他方法，支持向量回归优点是：支持向量回归不仅适用于线性模型，对于数据和特征之间的非线性关系也能很好抓住；
- 支持向量回归不需要担心多重共线性问题，可以避免局部极小化问题，提高泛化性能，解决高维问题；支持向量回归虽然不会在过程中直接排除异常点，但会使得由异常点引起的偏差更小。缺点是计算复杂度高，在面临数据量大的时候，计算耗时长。

模型构建

4.结果分析

- 构建支持向量机回归模型，得到部分房屋租金的预测值，如表。
- 平均绝对误差相对较小，R方值相对接近1，表明建立的支持向量回归模型拟合效果较好，但是还有一定的优化空间。

真实值	预测值	真实值	预测值
5000	5197.67459243	……	……
4300	4327.75426829	1800	1839.08881741
2200	2755.39438673	3300	3740.26655705
……	……	1800	1715.76151434

平均绝对误差	均方误差	R方值
409.7115	443927.4372	0.8865

小结

- 本章结合房屋租金影响因素分析与预测的案例，介绍了原始数据的处理、影响因素分析、构建支持向量回归预测模型、模型的评价四部分内容。
- 重点探究影响房屋租金的关键因素，在模型的构建阶段，根据筛选出的关键影响因素，建立支持向量回归模型，得到房屋租金的最终预测值。

Thank you!