**第3章：数据预处理与可视化**

**3.0 本章学习重点**

1.基本了解数据预处理的动机和主要方法。

2.掌握数据清洗的主要方法。

3.掌握数据规范化的主要方法。

4.掌握数据离散化的主要方法。

5.基本了解什么是数据可视化及其作用。

6.掌握常见的数据可视化统计图表。

**3.1数据预处理**

数据在收集时由于各种原因可能存在缺失、错误、不一致等问题，另外实验、模拟和数据分析过程中不可避免存在误差，这种问题通常被称为不确定问题。这种不确定性主要包括数据不确定性和属性值不确定性，将影响数据的分析结果：

**数据不确定性**涉及到数据本身的完整性、准确性和可信度。数据可能存在缺失、错误或不完整的情况，这些都会导致数据的不确定性。例如，在数据收集过程中可能会发生错误或遗漏，或者在数据传输和存储过程中可能会出现损坏或丢失。这种不确定性通常与数据的来源、收集方法和处理过程有关。

**属性值不确定性**指的是数据中各个属性值的可信程度或准确性。即使数据本身完整，但属性值的真实性和准确性仍然可能受到影响。例如，某个属性的值可能是基于估计、推断或不完全的信息得出的，或者是由于测量误差或主观判断而产生的。在数据分析和决策过程中，需要考虑这种属性值的不确定性，以确保对数据的正确解释和使用。

另外一方面，用于描述对象的数据有可能不能很好地反映潜在的模式，因此需要进行有效属性的提取以及构造。描述对象的属性的数量可能有很多，针对某一类分析问题，有些属性是无用的或者冗余的，正确识别出无用的、冗余的属性不仅可以提升分析效果，发现更有意义的、更便于理解的知识，同时也可以节省分析时间，提高算法的运行效率。

数据预处理是指在对数据进行进一步分析或建模之前，对原始数据进行清洗、转换和整理的过程。这个过程旨在使数据更适合用于特定的分析任务或模型构建，并且有助于提高分析的准确性和可靠性。通过数据预处理，可以提高数据的质量和可用性，减少分析和建模过程中的错误和偏差，从而更有效地挖掘数据的潜力，得出准确、可靠的结论或模型。数据预处理是商业智能应用中的一个重要环节，通常将占用大量的时间。

**3.2数据预处理方法**

数据预处理的方法通常包括数据清洗、数据转化、数据降维和特征筛选等。

**3.2.1数据清洗**

数据清洗的主要任务是检查异常值、处理缺失的数据和检测数据的不一致和错误等，异常值可以通过检测和修正来处理，缺失值可以通过填充、删除或插值等方法进行处理，而错误值或不一致等则需要根据具体情况进行修正或排除。这些工作是进行下一步数据处理和分析的前置任务。

**（1）异常值检测**

异常值检测是数据清洗过程中的一个重要步骤，旨在识别和处理数据中的异常值或离群点。这些异常值可能是由于录入错误、测量误差、数据处理过程中的问题或者是真实的、非典型的观测结果。异常值检测对于确保数据分析的准确性和可靠性至关重要。

异常值检测方法包括基于统计方法的异常值检测、基于距离的异常值检测以及基于密度的异常值检测方法等。

1. **基于统计方法的异常值检测**

基于统计方法的异常值检测指的是基于数据的统计属性，通过比较数据点与其它数据点的统计特征，来确定是否存在异常值。常见的基于统计方法的异常值检测技术包括：

**标准差方法（Z-score方法）**：基于数据的标准差来度量数据点与均值的偏离程度，超出阈值范围的数据点被认为是异常值。对于数据集的平均值和标准差，一个数据点x的Z-score计算公式为：

在异常值检测中，Z-score的绝对值大于某个阈值（通常是3）的数据点被认为是异常值。这意味着这些数据点与平均值相比的偏离程度超过了3个标准差。

Z-score方法是一种简单直观、可解释性强且具有很强通用性的常用异常值检测方法之一。在许多情况下，Z-score方法能够提供有效的异常值识别指标，帮助分析人员快速识别和理解数据中的异常情况。然而，我们也必须注意到一些问题。首先，Z-score假设数据服从正态分布，虽然它并不要求数据严格符合这种分布，但如果数据分布差异过大，可能会导致不准确的结果。此外，Z-score方法对数据缩放及数据尺度非常敏感，对于异常值的阈值没有一个统一的标准，而且在小样本情况下可能不够稳健。因此，在实际使用时，我们需要谨慎对待这些问题，并综合考虑其他异常值检测方法来提高结果的准确性和可靠性。

**IQR（四分位距）方法**：是一种基于箱线图的异常值检测方法，其中箱线图用于可视化数据的分布情况，而IQR则是其中的一个关键指标。四分位距（IQR）是指数据集的第三四分位数（Q3）和第一四分位数（Q1）之间的距离，表示了数据集的分散程度。具体计算公式如下：

在使用IQR方法进行异常值检测时，一般会将数据集中的异常值定义为在和 之外的数据点，通常被视为异常值。此方法也是箱线图设计的原理，详见本章3.5.1箱线图示例。

假设我们有一个关于客户年收入的数据集（见表3.1），包含‘客户编号’和‘年收入（万元）’两个字段。我们用IQR方法基于‘年收入’字段进行异常值检测。

表3. 1客户年收入表

|  |  |
| --- | --- |
| 客户编号 | 年收入(万元) |
| 1 | 86 |
| 2 | 65 |
| 3 | 90 |
| 4 | -50 |
| 5 | 82 |
| 6 | 91 |
| 7 | 300 |
| 8 | 40 |

首先将年收入数据进行排序：

然后计算第一四分位数（Q1）和第三四分位数（Q3）：

那么由(3-2)得IQR为38。根据IQR方法，我们认为成绩低于Q1 - 1.5 \* IQR = 52.5 - 1.5 \* 38 =-4.5或高于Q3 + 1.5 \* IQR = 90.5 + 1.5 \* 38 = 147.5的分数为异常值。这个数据集中编号为7的客户的年收入为300，远高于147.5，编号为4的客户的年收入为-50，远低于-4.5，因此我们可以将这两个数据点标记为异常值。

IQR方法相对简单，不受数据分布的假设影响，并且能够较好地处理偏态分布的数据，适用于非正态分布和包含离群点的数据集。然而，它可能无法有效地识别某些特定形状的异常值，也无法处理连续异常值，并且依赖于数据分位数的准确计算等。这也是我们在实际应用时需加以注意和调整的地方。

**Grubbs' （格拉布斯）检验：**是基于假设检验的方法，用于识别连续数据中的单个异常值。它假设样本服从正态分布，对于数据集的平均值和标准差，对每一个数据点x计算其G值：

然后计算Grubbs'统计量GG：

是所有样本中最大的Grubbs统计量（通常是极小值或者极大值），最后在所设定显著性水平（通常是0.05或者0.01）和样本大小下，查找Grubbs’检验的临界值，将GG与其进行比较，如果GG大于临界值，则认为最大G值对应的数据点是异常值。

Grubbs’检验具有广泛的适用性和操作简便。但是需要注意，它通常适用于样本大小较小且数据近似正态分布的情况下。此外，Grubbs’检验仅适用于识别单个异常值。

1. **基于距离的异常值检测**

基于距离的异常值检测方法也是常用的异常值检测技术之一，它通过计算数据点之间的距离来识别异常值。常用的检测方法有基于KNN的异常值检测算法：

**基于KNN（K近邻）的异常值检测算法**：使用数据点与其最近的K个邻居之间的距离来判断其是否异常。该算法的基本思想是，正常数据点周围的K个最近邻居应该彼此非常接近，而异常值可能与其最近的邻居之间的距离较远。具体步骤为：

（1）**选择K值**，首先选择一个合适的K值，即确定每个数据点周围要考虑的最近邻居的数量。较大的K值将考虑更广泛的邻域，而较小的K值将更关注于局部结构。

（2）**计算距离**：对于每个数据点，计算它与其他数据点之间的距离。通常使用曼哈顿距离、欧氏距离或闵可夫斯基距离等度量，分别对应公式（3-5）中的情况。

其中n为每个数据点的特征维度数量，代表数据点x和y之间的距离。

（3）**确定邻居**：对于每个数据点，找到其与之最近的K个邻居。

（4）**计算异常值分数**：对于每个数据点，计算其异常值分数。一种常见的方法是将其与第K个最近邻居的距离作为异常值分数。距离较大的数据点可能更有可能是异常值。

（5）**识别异常值**：将异常值分数与阈值进行比较。超过阈值的数据点可能被识别为异常值。

需要注意的是，KNN算法的性能取决于K值的选择和距离度量的方式。较大的K值可能会导致较低的灵敏度，而较小的K值可能会导致对局部异常的过度敏感。因此，在选择K值和距离度量时需要进行实验和调优。此外，KNN方法对于高维数据可能会面临维度灾难的问题，因此在高维空间中需要特别小心。

1. **基于密度的异常值检测方法**

基于密度的异常值检测算法是一类通过评估数据点周围的密度来识别异常值的方法。这些算法通常假设异常值在数据集中的密度较低，而正常值则具有更高的密度。基于密度的异常值检测算法主要关注数据点周围的局部密度，通过与其邻居的密度比较来确定异常值。常见的基于密度的异常值检测算法包括局部离群因子（LOF）算法、基于密度的聚类空间扫描算法（DBSCAN）、基于可达性的聚类算法（OPTICS）以及Mean Shift算法等。

**局部离群因子（LOF）算法**：是一种基于密度的异常检测方法，由Markus M. Breunig于2000提出，通过计算和比较每个数据点周围的局部可达密度（LRD）与其邻居的局部密度可达之比来确定异常值。它认为局部可达密度（LRD）远低于其邻居的样本点为异常值。该算法的基本思想和步骤如下：

1. 计算每个点的第距离邻域内各点的第可达距离：对于每个数据点，首先找到其第k距离邻域 。其中，数据点第k距离定义为，而的第k距离邻域指的是点x的第距离内的所有点的集合，包含第距离上的点。易知，有。然后，计算其第k距离邻域内各点的第k可达距离：

其中为邻域内点到数据点x的第k可达距离，为数据点的第距离。

1. 计算每个点x的第局部可达密度：

数据点x的第k局部可达密度，即数据点x的第k距离邻域内的所有点到点x的平均第k可达距离的倒数。它表征了点x的密度情况，点x与周围点密集度越高，各点的可达距离越可能是较小的各自的第k距离，LRD值越大；点x与周围点的密集度越低，各点的可达距离越可能是较大的两点间的实际距离，LRD值越小。

1. 计算每个点的第局部离群因子且识别异常值:

数据点x的第k局部离群因子，意为将点x的邻域内所有点的平均局部可达密度与点x的局部可达密度作比较，这个比值越大于1，表明x点的密度越小于其周围点的密度，x点越可能是离群点；这个比值越小于1，表明x点的密度越大于其周围点的密度，x点越可能是正常点。

LOF算法的优点包括对数据分布的不敏感性，可以处理不规则形状的簇以及多密度区域的数据。然而，它也有一些限制，如需要选择合适的k值，算法复杂度较高等。

**其他基于密度的异常值检测算法（具体算法细节可见聚类算法章节）**：

* **DBSCAN**（基于密度的聚类算法）：定义领域半径R和最少点数目MinPoints。邻域半径R内样本点的数量大于等于Minpoints的点叫做核心点，不属于核心点但在某个核心点的邻域内的点叫做边界点，既不是核心点也不是边界点的是噪声点。噪声点可能是异常值。
* **OPTICS**（基于可达距离的聚类算法）：考虑DBSCAN算法对参数R和MinPoints过于敏感，OPTICS在此基础上做出改进。如LOF算法一样，其也通过计算数据点的可达性距离来确定其相对密度OPTICS通过分析数据点之间的可达性关系来识别数据集中的簇和异常值。
* **Mean Shift：**Mean Shift算法使用核密度估计来识别数据集中的密集区域，并将密度较低的点视为异常值。

总的来说，这些基于密度的方法在实践中异常检测任务被广泛应用。它们对数据的分布情况和数据点的相对关系非常敏感，因此在一些情况下能够提供准确的异常值检测结果。然而，这些方法也存在一些限制，例如对参数的敏感性和计算复杂度较高。因此，在使用时需要根据具体情况进行适当的调整和选择。

1. **填补数据缺失**

缺失值处理是在数据预处理中的一个重要步骤，用于处理数据中存在的缺失数值或者缺失观测的情况，以便进行后续的分析和建模。

缺失数据的类型主要有三种：

* 完全随机缺失（Missing Completely at Random，MCAR）：意味着数据的缺失是完全随机的，不受任何变量（包括已观测到的变量或未观测到的变量）的影响。换句话说，缺失值的出现是纯粹随机的，与数据集中的任何特定模式或属性无关。在这种情况下，缺失数据的缺失与样本的无偏性没有关系，可以直接忽略或者用简单的填充方法进行处理。
* 随机缺失（Missing at Random，MAR）：数据的缺失是随机的，与缺失的数据本身无关，而是与其他已观测到的数据相关。换句话说，缺失值的出现不是完全随机的，是由于其他已知的变量导致的，而不是缺失值本身的特性引起的。在这种情况下，缺失数据的模式可能与观测到的数据模式类似，因此可以使用已观测到的数据来进行填充或处理缺失值。
* 非随机缺失（Missing Not at Random，MNAR）：非随机丢失意味着数据的缺失与未观测到的变量自身的取值有关。换句话说，缺失值的出现与缺失值本身的特性或与其他未观测到的变量相关（例如，高收入人群通常不希望在调查中透露他们的收入；又如假设女性通常不想透露年龄，则年龄变量的缺失值受性别变量的影响）。在这种情况下，缺失数据的缺失与数据集中的其他因素有关，可能会导致数据的偏差。处理非随机丢失通常需要更复杂的方法，例如利用模型来估计缺失值，或者进行敏感性分析来评估缺失值对结果的影响。

在前两种情况下可以根据其出现情况删除缺失值的数据，同时，随机缺失可以通过已知变量对缺失值进行估计。在第三种情况下，删除包含缺失值的数据可能会导致模型出现偏差，同时，对数据进行填充也需要格外谨慎。正确判断缺失值的类型，能给我们的工作带来很大的便利，但目前还没有一套规范的缺失值类型判定标准。大多是依据经验或业务进行判断。

对缺失值进行处理时，最简单的方法是直接删除包含缺失值的行（样本）或列（属性），但当缺失数据所占比例较大，特别是当缺失的数据非随机分布时，直接删除可能导致数据发生偏离，引发错误的结论。另外一个方法则是缺失值填充，将数据集中的缺失值用适当的值进行替换，而在这其中，最简单的方法是用特殊值填充，如某个常数或者‘unknown’（或0），但这也可能导致严重的数据偏离，一般不推荐。

因此，根据数据集的特性和不同情况，常见的缺失值填充方法包括均值/中位数/众数填充、前向/后向填充、插值法填充等。

1. **均值/中位数/众数填充**

对于数值型属性，其空值可以用该属性在其他所有样本的均值、中位数或者其他描述数据集中趋势的统计量来填充缺失值。

对于非数值型属性（如类别变量或者有序变量），其空值可以使用该属性在其他所有样本的出现频率最高的取值（众数）来填充缺失值。

**示例：**表3.2中，在收集客户信息时，如果属性设计不合适，也可能造成数据的缺失。例如，若将“婚姻”属性改为“已婚否”,则对于离异的客户，此属性的值可能为空。当前表格存在两个缺失值，第4个客户的年收入和第8个客户的年龄。根据均值/众数分别填充数值型和非数值型属性，则第4个客户的年收入由豪华车类型为“否”的客户的年收入的均值：(86+65+90+40+20+96+80+80)/8=58来填补。对于年龄段属性，根据众数原理，根据豪华车类型为“否”的客户来判断，第8个客户的年龄则可以填补为“<30”。

表3. 2含有缺失值的数据集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 客户编号 | 年龄(岁) | 性别 | 年收入(万元) | 婚姻 | 豪华车 |
| 1 | <30 | 女 | 86 | 已婚 | 否 |
| 2 | <30 | 男 | 65 | 单身 | 否 |
| 3 | <30 | 男 | 90 | 离异 | 否 |
| 4 | <30 | 女 |  | 已婚 | 否 |
| 5 | 30～50 | 女 | 82 | 已婚 | 是 |
| 6 | 30～50 | 男 | 91 | 已婚 | 是 |
| 7 | 30～50 | 女 | 200 | 离异 | 是 |
| 8 |  | 女 | 40 | 单身 | 否 |
| 9 | 30～50 | 男 | 20 | 离异 | 否 |
| 10 | >50 | 女 | 96 | 离异 | 否 |
| 11 | >50 | 女 | 80 | 单身 | 否 |
| 12 | >50 | 男 | 50 | 单身 | 是 |
| 13 | >50 | 女 | 80 | 离异 | 否 |
| 14 | >50 | 男 | 92 | 离异 | 是 |

1. **前向填充/后向填充**

前向填充（Forward Fill）和后向填充（Backward Fill）是两种常见的缺失值填充方法，通常是针对**时间序列**数据或有序数据的最基本的方法。

* **前向填充（Forward Fill）**：使用缺失值前面最近的已知值来填充缺失值。也就是说，缺失值被其前面最近的非缺失值所填充。在时间序列数据中，这意味着缺失值后面的值会重复使用来填充缺失值，直到遇到下一个非缺失值为止。
* **后向填充（Backward Fill）**：与前向填充相反，后向填充使用缺失值后面最近的已知值来填充缺失值。也就是说，缺失值被其后面最近的非缺失值所填充。在时间序列数据中，这意味着缺失值前面的值会重复使用来填充缺失值，直到遇到下一个非缺失值为止。这是一种lookahead行为，只有当不需要预测未来数据的时候才能使用。

这两种填充方法简单直观。但是在使用的时候要注意在不同情况下可能产生不同的效果。当时间序列数据具有明显的趋势或季节性变化，则可能更倾向于使用前向填充或后向填充来保持这些趋势的连续性。但是，如果数据的变化比较平稳或随机，则这两种方法的效果可能相似。

1. **插值方法**

插值方法是一种常用的缺失值填充方法，它通过已知数据点之间的关系来估计缺失值。常见的插值方法包括线性插值、拉格朗日插值、热卡（Hot Deck）填充、K最近距离邻法、回归以及期望值最大化方法等。

**线性插值**：是一种广泛使用的简单插值方法。它根据假设已知数据点之间的线性关系来估计缺失值，是指使用连接两个已知量的直线来确定这两个已知量之间的一个未知量的关系。对于某数据点，其中为缺失值，已知数据点（x1, y1）和（x2, y2），且位于区间内，则采用线性插值的方法估计y：

**拉格朗日插值**：是一种常用的多项式插值方法。它使用拉格朗日多项式来近似估计缺失值。该方法假设已知数据点之间存在一个多项式函数，并通过多项式函数来估计缺失值：

假设是已知数据点。对于某数据点，其中自变量已知，而为缺失值，采用拉格朗日插值的方法计算y：

其中是拉格朗日基函数：

拉格朗日插值法的优点是简单易懂。然而，由于拉格朗日插值多项式的高阶特性，当数据点较多或数据点分布不均匀时，可能会导致插值结果出现振荡或不稳定的情况。因此，在应用拉格朗日插值法时，需要谨慎选择数据点。

**最近邻插补**：是一种常见的缺失值填充方法，根据“物以类聚，人以群分”这样的规律，借助与缺失样本最邻近（相似度最高）的K个已观测的样本的属性值来，通过平均或加权平均，来进行填充。当K=1时，最近邻插补法又被称为**热卡**填充法。

最近邻插补的优点是简单易行，并且能够保留数据的分布特征。它适用于各种数据类型，包括数值型和分类型数据。此外，由于它基于样本之间的相似性进行填充，因此通常不需要额外的模型假设。然而，最近邻插补也有一些局限性。首先，它对于维度较高的数据集可能效果不佳，因为在高维空间中找到最近邻样本会变得更加困难。其次，如果数据集中存在大量的噪声或异常值，最近邻插补可能会导致填充值的偏差。因此，在使用最近邻插补时，需要仔细选择距离或相似度的度量方法，并对数据进行预处理以减少噪声和异常值的影响。

**其他插值方法**：除了以上三个广泛使用的方法，常用的插值方法还包括基于机器学习的方法（如回归模型、时间序列模型等）来预测缺失值、使用矩阵分解（如奇异值分解将数据矩阵分解成低秩近似来估计缺失值）、多重插补（基于多重插补的方法，通过多次模拟估计缺失值，并对多次填充结果进行汇总来得到最终填充值）等。

**（3） 数据不一致检测**

数据不一致检测是数据清洗过程中的关键步骤，它涉及到识别和纠正数据集中的矛盾或冲突信息。不一致性可能源于数据录入错误、数据错坏、数据来源差异、数据更新不同步或数据格式不统一等问题。数据不一致性如果不被及时处理，可能会导致数据分析结果不准确，影响决策质量。

数据不一致检测通常包括以下几种情况：

* 重复数据：数据集中存在完全相同或高度相似的记录；
* 格式不一致：同一类型的数据以不同的格式或单位存储，例如日期格式的多样性；
* 数据来源冲突：来自不同数据源的信息相互矛盾，例如两个数据源提供的同一实体信息不一致。
* 逻辑错误：数据违反了业务规则或逻辑，例如一个学生的出生年份早于其父母的出生年份。

数据不一致检测和清洗通常需要结合具体的业务规则和数据质量标准来进行，可以具体采用数据对比、数据校验、重复性检查、异常值检测等方法。在实际操作中，可能需要使用数据清洗工具或编写自定义脚本来自动化这一过程。

**3.2.2数据转换**

数据转换是数据预处理中非常重要的一个过程，它是指通过一定的策略改变数据的属性值、结构或者格式，以使数据更适合后续的分析、建模或者可视化。常见的数据转换包括数据规范化、数据离散化等。

**（1）数据规范化**

数据规范化又称标准化(Standardization), 通过将原始**数值型属性**转换为一定范围内的特定规范或标准的过程。其主要目的是消除数值型属性之间量纲和数量级差异，便于数据更易于比较和分析。例如，在利用欧式距离计算对象的相异度时，如果描述对象的属性包括年龄和收入(以元为单位),则“收入”在计算距离时起到关键的作用，掩盖了年龄的作用。如果将两者均规范化到同样的区间，如 则可以避免此问题。

标准化的前提是尽量不改变原始数据的规律。常见的数据规范化方法主要包括最小-最大规范化、Z-score规范化以及小数定标规范化方法等。

1. **最小-最大规范化（Min-Max Normalization）**

最小-最大规范化是最常用的一种规范化方法，又被称为range scaling，它将原始数据线性地转化到指定的映射范围内，通常是[0,1]或[-1,1]。假设，对于给定一个属性A，映射目标区间为，原来的取值范围为点x映射到新区间转换公式如式(3-12)所示：

其中为规范化的数据。例如，对于描述客户的属性“年收入(万元)”,如果原来的取值范围为[3,200],新的取值范围为[0,1],则若某客户的年收入为60万元，规范化后的年收入为(60-3)/(200-3)\*(1-0)+0=0.29。

1. **Z-score规范化（Z-score Normalization）**

Z-score规范化又称零均值规范化(Zero-Mean Normalization)，将原始数据转化为其标准分数（也称为Z分数）。给定一个属性A,设其原始数据取值的均值为,标准差为, A的某个取值x规范化后的值的计算方法如式 (3-13) 所示：

式中，均值为和标准差为通过已有样本的属性值进行计算。规范化后的属性A其取值满足均值为0，标准差为1的正态分布。

例如，描述客户群体的年收入（万元）属性的均值为82,标准差为39,则某客户其年收入为60万元，通过Z-score规范化后，年收入属性对应的值为 -0.56。

1. **小数定标规范化 (Decimal Scaling)**

小数定标规范化是通过移动要转换的属性A的值的小数点位置来进行规范化。小数点的移动位置依赖于A的最大绝对值。对于A的某个取值x规范化后的值的计算方式如式 (3-14) 所示：

其中是一个常数，通常选择使得规范化的数据位于[-1, 1]之间。

除了以上三种，还包括对数变换（缩小数据的绝对范围）、box-cox变换（适合目标变量不满足正态分布的情况）、幂次变换（可以改变数据的分布形态，使其更满足正态分布等）、分位数转换（根据分位数等，将数据映射到介于0和1之间的均匀分布）等。

选择适当的数据规范化方法取决于数据的分布特性、数据分析和模型的要求以及具体问题的特点。在进行规范化之前，通常需要对数据进行探索性分析，了解数据的分布情况和特点，以便选择合适的规范化方法。

**（2）数据离散化**

数据离散化是将连续型数据转换为离散型数据的过程，其目的是将连续的数值型变量转换为一组有限的取值范围内的离散值，以便于分析和建模。一方面，可以更好地适应不同算法的需求，提高模型的稳定性。比如某些算法无法处理连续类型的属性，例如，关联规则的挖掘算法只能处理离散类型的取值。为此，离散化是某些知识发现过程的必要步骤。另一方面，将无限的连续型数据转化维有限数量的离散值，从而简化了数据的复杂性，减少计算复杂度。第三，离散化处理能够在一定程度上减少异常值的影响，提高对于异常数据的鲁棒性。

离散化的常用方法主要分成两类：无监督的离散化和有监督的离散化，这主要是根据是否需要使用数据的标签信息来区分的，当不需要标签信息时则是无监督的离散化方法，而需要标签信息时则是有监督的方法。无监督离散化最常用的方法是分箱法(Binning)，有监督离散化方法常用的有基于熵的方法和基于卡方统计量的ChiMerge方法。

**a) 无监督离散化方法**

分箱离散化（Binning Discretization），也称为分段离散化或分桶离散化，是一种常见的离散化方法，它将连续型数据划分为多个区间（箱子或桶），每个区间对应一个离散值或类别。分箱离散化是一种无监督离散化方法，利用这种方法为一个属性离散化时，不需要考虑其他属性的取值。分箱离散化方法分为三类：等距离(Equal-Distance)分箱、等频率(Equal-Frequency)分箱以及基于聚类分箱方法。

**等距离分箱**，又称为等宽度分箱(Equal Width Binning),是将连续型数据的取值范围划分为若干个等宽的区间，每个区间的宽度相同。其具体算法和步骤如下：

1. 对于给定属性A，首先确定分箱的数量k。分箱的数量k可以根据数据的分布情况、业务需求和分析目的来确定。通常可以通过直方图、箱线图等方式来观察数据的分布情况，从而确定合适的分箱策略。
2. 确定分箱边界。若属性A的其最小和最大取值分别为min和max，若分箱个数为k，则每个区间的间距为。
3. 对数据进行分箱。按照确定的分箱边界，将每个数据点按照属性A的取值划分到对应的分箱中。对分箱后的数据进行编码，将每个分箱映射为一个离散值或者类别。

**等频率分箱**，又称为等深度分箱(Equal-Depth Binning)。它将连续型数据划分为若干个区间，使得每个区间内包含的数据点数量大致相等。这种方法的主要目的是确保每个分箱内的数据分布均匀，以减少离散化后数据的不平衡性。其具体算法和步骤跟等距离分箱类似，只是分箱边界的确定是根据分位数来判断的。

**示例**：假设14个客户的属性“年收入”的取值按顺序为: 20, 40, 50, 58, 65, 80, 80, 82, 86, 90, 96,105, 120, 200。若确定分箱的个数为4，利用等距离分箱，则区间间距为，则4个箱的区间分别为 ，，，。根据此区间将原来的14个客户，映射到四个分箱区间为，，，。由此可见，等距离分箱可能导致属于某些分箱区间的数据点非常多，而某些区间的取值又非常少。等频率分箱方法则能够一定程度解决此问题。利用等频率分箱，每箱至少3个值，则4个箱分别为，，, 。

映射到每个分箱区间之后，属于一个区间的每个属性值都可以用区间代替，或者用一个类别代替。例如，在等距离分箱中，属性值20、40、50可以用代替，或者将4个区间分别对应类别1、2、3和4，则这3个取值均用类别1代替。

**基于聚类的分箱**，是指对连续型属性执行聚类算法（k-means），将样本划分为若干个簇，然后，我们可以将每个簇视为一个分箱，将原始的连续型数据转换为离散的箱子编号或者簇标签。

需要注意的是，以上三种分箱离散化方法对于分布不均匀的数据并不适用。

**b) 有监督离散化方法**

分箱离散化是一种无监督的数据预处理方法，但它可能会导致属性取值的离散化不利于有意义模式的发现。举例来说，当我们对属性“年收入”进行分箱离散化时，可能将取值20、40、50映射到同一个区间中。然而，实际上，“年收入小于等于40万元的客户其类别为‘豪华车=否’”可能是一个重要的模式，但由于离散化的处理，这种模式变得更加难以发现。

因此，在面对分类问题时，有监督的离散化往往更为合适。有监督的离散化可以根据分类属性的取值进行连续属性的离散化，以尽量保持映射到同一区间的对象的类别一致。

常见的有监督离散化方法包括基于熵的离散化和卡方分箱（ChiMerge）算法。

**基于熵的离散化方法**：基于熵的离散化方法是一种利用信息熵（Entropy）来进行特征分箱的有监督离散化方法。它的基本思想是在连续特征的取值范围内，寻找最佳的分割点，使得分割后的子集中的类别信息最大化地纯净或最小化地混合。具体步骤如下：

* + - 1. **计算初始熵：**首先，计算连续特征的初始熵，即将整个数据集视为一个子集时的熵值。熵可以用于度量分类属性取值的纯度，因而可以用于衡量一个区间的优劣，映射到一个区间的对象的类别纯度越高，此离散化的结果越好。给定一个数据集D及分类属性的取值，即类别集合C={c1,c2,…,ck},数据集D的信息熵entropy(D)的计算方法如式(3-15)所示：

式中， , 表示类别ci在数据集D中出现的次数，|D|代表D中的对象个数。信息熵的取值越小，类别分布越纯；反之越不纯。

* + - 1. **尝试分割点且计算信息增益**（或信息增益比）：在连续特征A的取值范围内，尝试不同的分割点，将数据集分为两个子集和。对每个分割点，同样按照式（3-15）计算分割后的两个子集的熵，并计算信息增益或信息增益比。信息增益或信息增益比的计算依赖于具体的离散化算法，如信息增益对应于ID3算法，信息增益比对应于C4.5算法。

以信息增益为例，假设分割点为，分割后的信息熵的计算如式（3-16）所示：



则按照此分割点的信息增益，记为，其计算公式如式（3-17）所示：

* + - 1. **选择最佳分割点：**选择使得信息增益或信息增益比最大的分割点作为最佳分割点。
      2. **递归地进行分割**：根据最佳分割点将数据集分为两个子集和，对每个子集重复步骤2至步骤3，直到满足停止条件为止。停止条件可以是子集大小小于预先设定的阈值或者信息增益小于某个阈值。

通过这种方式，基于熵的离散化方法能够有效地将连续特征分割为多个离散的区间，从而在保持信息丰富性的同时减少了特征的复杂度。这种方法常用于决策树等基于信息论的分类算法中，为特征工程提供了一种重要的选择。

示例：为了解释此方法，假设要离散化的属性为“年收入”,分类属性为“豪华车”,数据如表3.1所示。

表3. 3离散化数据集示例

|  |  |
| --- | --- |
| 年收入(万元) | 豪华车 |
| 20 | 否 |
| 40 | 否 |
| 50 | 是 |
| 58 | 是 |
| 65 | 否 |

表3.3中数据的信息熵，Entropy(D) = =0.97。

在比较不同的分割条件信息增益时，不必对A的每个取值均进行计算，已经证明只需要在类别发生变化的取值处进行计算即可。例如，对于表3.3中的数据，只需要比较“年收入≤40万元”和“年收入≤58万元”两种情况。其信息增益计算方法如下：

显然，分割点为40比58的信息增益大，其为最佳分割点，因此，最好将年收入的取值分为两个区间：小于等于40万元和大于40万元。接着，可以对这两个区间继续分裂。

**b) 卡方分箱离散化方法**

卡方分箱（ChiMerge）方法也是一种常用的有监督离散化方法。如果基于熵的方法可以被看做是自顶向下的分裂方法，ChiMerge则属于自底向上的合并方法。基于熵的方法是从一个大区间开始，不断分裂成小区间，而ChiMerge则是从每个值都是一个小区间开始，不断合并相邻区间成为大区间，它是基于统计量卡方检验实现的。卡方检验可以验证属性之间的独立性，用在离散化时，可以衡量相邻两个区间与分类属性的独立性，如果是独立的，则说明两个区间对于类别的取值没有显著差异，因此可以合并。具体步骤如下：

1. **初始化：**将连续型属性的取值按照从小到大排列，并且初始化每个取值区间为单独的箱子；
2. **计算卡方值**：对于相邻的取值区间，计算其对应的卡方值。卡方值用于衡量这两个区间的相似性，卡方值越小表示这两个区间之间的差异越小，合并后对目标变量的影响也越小。假设有两个区间1和2，以及分类属性的类别取值为{}，先计算得到列联表3.4，其中代表属于区间i和类别的样本数，是区间i的总样本数，是属于这两个区间的类别的总样本数，则为这两个区间的总样本数。

表3.4 列联表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 类别 | 类别 | … | 类别 | 求和 |
| 区间1 |  |  | … |  |  |
| 区间2 |  |  | … |  |  |
| 求和 |  |  | … |  |  |

根据列联表，这两个区间的卡方值的计算方法如式(3-18)所示：

其中

若式(3-19)，则将其设为很小的数(如0.1),以防式(3-18)中分母为零。

1. **合并经卡方检验独立的区间**：找到相邻取值区间中卡方检验独立的两个区间，将其合并为一个箱子。这样做可以将相似的取值区间合并在一起，减少离散化后的特征的取值个数。自由度=),其中r为行的个数，即区间数，k类别数。设定显著性水平,根据自由度查卡方分布表得到阈值β,若计算所得卡方值小于此值，则合并这两个区间。
2. **重复步骤2和3**：不断重复计算卡方值和合并箱子的步骤，直到满足停止条件为止。通常的停止条件包括合并后的区间数目达到预设阈值或者相邻区间不满足合并要求。

**示例：**以表3.3中数据为例，在将待离散化属性“年收入”的取值排序之后，生成只含有单个取值的区间，以相邻两个值的中点为分界，初始区间为[0,30],(30,45],(45,54],(54,61.5],(61.5,]。对两个相邻区间构建列联表，表3.5给出的是区间[0,30]和(30,45]的列联表：

表3.5基于表3.3数据且考虑区间[0,30]和(30,45]的列联表

|  |  |  |  |
| --- | --- | --- | --- |
|  | 豪华车=是 | 豪华车=否 | 合计 |
| [0,30] | 0 | 1 | 1 |
| (30,45] | 0 | 1 | 1 |
| 合计 | 0 | 2 | 2 |

对于表3.4,可以计算得，查卡方分布表，α=0.1时，β=2.706。因为，0.2<2.706,因此可将这两个区间合并[0,45]

**3.2.3数据特征提取**

数据特征提取主要包括数据降维和特征选择两方面。数据降维是对高维数据进行降维操作，在尽可能保持原始数据的信息的基础上，减少特征的数量和复杂度。而特征选择是指选择对分析或建模任务最具有代表性和预测能力的特征，剔除无关或冗余的特征。数据降维和特征选择的目的都是为了减少特征的数量，以便更好地适应模型和分析需求。它们可以提高模型的效率、泛化能力，同时还有助于降低过拟合的风险，简化模型的理解和解释。

**（1）数据降维**

常用的数据降维方法包括主成分分析（Principal Component Analysis，PCA）、线性判别分析（Linear Discriminant Analysis，LDA）以及t分布邻域嵌入（t-Distributed Stochastic Neighbor Embedding，t-SNE）等。

**a) 主成分分析**

主成分分析(Principle Component Analysis, PCA)最早由Karl Pearson于1901年提出，后经Harold Hotelling发展，是一种经典的统计方法。它通过对原有变量（属性和特征）进行线性变换提取数据中的主要特征，并以此实现降维、去除去除冗余和噪音的目的。

PCA 的核心思想是找到数据中最显著的方向（主成分），通过线性变换，将高维数据映射到低维空间，在保留数据的重要信息的同时减少数据的维度。

给定一个包含n个数据对象点的数据集，每个数据点由m个数据属性、、…、 描述。也就是说，每个原始数据对象都可看做是m维空间中的一个点。运用PCA（基于特征值分解协方差矩阵）对其进行降维和特征提取的主要步骤如下所示：

1. **数据中心化**：首先，对原始数据进行中心化处理，使得每个变换后的属性的均值为0，即将每个属性的取值减去该属性的均值：

其中元素代表第个对象第个属性中心化后的取值，代表其原值，是属性的均值，中心化后的数据用表示，。

2. **计算协方差矩阵**：其中元素是属性和之间的协方差，即：

则说明属性和正相关，代表负相关，而则说明互相独立。可利用中心化的数据直接进行计算协方差矩阵，即： 。

3. **特征值分解**：用特征值分解方法求协方差矩阵的特征根和特征向量，保留前个最大的特征根及对应的特征向量，其中最大特征根对应的特征向量称为第一主成分，第二大特征根对应的是第二主成分，以此类推。构造主成分矩阵,其中第个列向量则对应第个主成分。

假设降序排列的特征根为,第个主成分的贡献率的计算方法如式(3-22)所示：

前个主成分的贡献率的计算方法如式(3-23)所示：

通常，取累计贡献率达到85%～95%的前q个特征值对应的特征向量构造主成分矩阵

4. **数据转换**：将数据转换到维特征向量构造的新空间，得到降维后的数据集，,，其中**P**是主成分矩阵，是步骤(1)中得到的矩阵。

以上主成分提取以及线性变换的基本原理可这样来解释：协方差矩阵中对角线上的元素是各个属性的方差，非对角线上的元素是任意两个属性之间的协方差。为了去除冗余和降低噪音，希望将协方差矩阵中的非对角线元素变为0,即将协方差矩阵对角化，对角化后的矩阵对角线上的元素为协方差矩阵的特征根，也是转换后新特征的方差，方差大的特征具有更多的信息。如果一个特征(属性)的取值之间变化很小，即方差很小，那么它含有的信息量就很小。以分类问题为例，这样的特征对于区分不同的类别是没有价值的，因此可以舍弃。因此，最终应选择那些大的方差，即大的特征根对应的特征向量作为主成分，按照这些正交的主成分(即互相独立的)对原来的特征进行变换，每个新的特征是原有特征的线性组合。这样，既抓住了主要变量，又降低了维度、去除了噪音和冗余，对于数据的进一步分析奠定了良好的基础。

对于分类问题，通过将训练数据集进行PCA降维提取主成分之后，利用降维后的数据集进行分类器的构建。对于测试集，也需要用主成分矩阵**P**进行转换，然后利用分类器进行类别的预测。

**示例**：以kaggle竞赛官网中房价预测比赛为例1，取数据集中4个属性(MSSubClass、LotFrontage、WoodDeckSF、OpenPorchSF)分别用于指代房屋占地面积、整体质量评分、整体状况评分、地下室未完工面积。表3.5是其中的一个子集。完整训练数据集包含1460个样本。表3.6是中心化后的数据集。

表3. 5房价预测数据集子集

|  |  |  |  |
| --- | --- | --- | --- |
| MSSubClass | LotFrontage | WoodDeckSF | OpenPorchSF |
| 60 | 65.0 | 0 | 61 |
| 20 | 80.0 | 298 | 0 |
| 60 | 68.0 | 0 | 42 |
| 70 | 60.0 | 0 | 35 |
| 60 | 84.0 | 192 | 84 |
| 50 | 85.0 | 40 | 30 |
| 20 | 75.0 | 255 | 57 |
| 50 | 51.0 | 90 | 0 |
| 190 | 50.0 | 0 | 4 |
| 20 | 70.0 | 0 | 0 |

表3. 6中心化数据集

|  |  |  |  |
| --- | --- | --- | --- |
| MSSubClass | LotFrontage | WoodDeckSF | OpenPorchSF |
| 0.0 | -3.8 | -87.5 | 29.7 |
| -40.0 | 11.2 | 210.5 | -31.3 |
| 0.0 | -0.8 | -87.5 | 10.7 |
| 10.0 | -8.8 | -87.5 | 3.7 |
| 0.0 | 15.2 | 104.5 | 52.7 |
| -10.0 | 16.2 | -47.5 | -1.3 |
| -40.0 | 6.2 | 167.5 | 25.7 |
| -10.0 | -17.8 | 2.5 | -31.3 |
| 130.0 | -18.8 | -87.5 | -27.3 |
| -40.0 | 1.2 | -87.5 | -31.3 |

该数据集的协方差矩阵如下：

特征根、贡献率及累积贡献率如表3.7所示。

表3. 7特征根、贡献率及累积贡献率

|  |  |  |
| --- | --- | --- |
| 特征根 | 贡献率 | 累积贡献率 |
| 12958 | 83.39% | 83.39% |
| 1713 | 11.02% | 94.42% |
| 804 | 5.17% | 99.59% |
| 64 | 0.41% | 100.00% |

从表3.7中可以看作前2个特征根已经达到94.42%的贡献率，因此可以选择这两个特征根对应的特征向量，

新的特征是原有属性的线性组合，设新的特征为, 则：

**b) 线性判别分析**

线性判别分析（Linear Discriminant Analysis, LDA）是一种*监督学习*的降维技术，由R. A. Fisher在1936年提出。LDA考虑类别信息（即标签），将特征空间投影到低维空间中，同时最大化类别之间的距离，也就是保持类内方差最小，类间方差最大。因此，LDA不仅能够降低数据的维度，还能够提高分类器的准确性。

运用LDA进行数据降维的主要步骤如下所示：

给定一个包含n个数据对象点的数据集，每个数据点由m个数据属性、、…、 描述，即为m维向量，样本类别样本分类。定义为第j类样本的个数，为第类样本的集合，为第j类样本的均值向量，为第j类样本的协方差矩阵。

求解矩阵的特征值和对应的特征向量w，即：

其中，类内散度矩阵和类间散度矩阵为：

为所有样本的均值向量。

根据特征值的大小，选择最重要的几个特征向量作为主成分。这些特征向量代表了数据在新特征空间中的最佳投影方向，它们能够最大化类别间的分离。

按特征值从大到小排列，选择前d维主成分构造投影矩阵，则每个样本点在新特征空间中的投影坐标为：

**c）t分布邻域嵌入**

t分布邻域嵌入t-SNE（t-distributed Stochastic Neighbor Embedding）是一种非线性的降维技术，特别适用于高维数据的可视化。它由Laurens van der Maaten和Geoffrey Hinton在2008年提出。t-SNE的主要目标是在低维空间中（通常是二维或三维）保持高维空间中的局部结构，使得原始数据中相似的点在低维映射中也保持相似，而不相似的点则远离。在这个过程中，t-SNE采用了一种称为“t-分布”的概率分布，来表示高维空间和低维空间中样本点的相似度。

t-SNE的工作原理可以分为以下几个步骤：

**1. 相似度计算**：在高维空间中，t-SNE首先计算每个数据点之间的条件概率（表示在领域的概率），这个概率反映了一个点选择另一个点作为其邻居的可能性。这种概率通过高斯分布（正态分布）来建模，并且对每个点使用以该点为中心的方差，来保证每个点在高维空间中的局部邻域密度相似：

表示原高维数据中数据点，表示以为中心的方差。

但在高维空间中这样定义会带来异常值的问题。假设点 是一个噪声点，那么的平方会很大，那么对于所有的j，的值都会很小，导致在低维映射下的该点对整个损失函数的影响很小。因此，对于异常值，我们需要得到一个更大的惩罚，于是对高维空间中的联合概率修正为:

**2. t分布**：初始化低维空间中的样本点。在低维空间中，t-SNE使用自由度为1的t分布（也称为柯西分布）来计算点之间的相似度。这种分布的长尾特性有助于在低维空间中分开距离较远的点，避免不同类别的点在降维后重叠。

表示降维后低维数据中的数据点。

**3. Kullback-Leibler (KL) 散度最小化**：t-SNE的目标是使得低维空间中的条件概率分布尽可能接近高维空间中的条件概率分布。为此，它通过最小化两个分布之间的KL散度来调整低维空间中点的位置。

**4. 梯度下降**：使用梯度下降算法来最小化KL散度。在迭代过程中，它会逐步调整低维空间中点的位置，直到找到一组位置使得低维空间中的概率分布与高维空间中的分布足够接近。

t-SNE特别适用于数据可视化，尤其是在高维数据集中寻找和展示局部结构和聚类模式时。它可以揭示数据中的聚类结构，帮助我们理解数据的内在特性。

**（2）特征选择**

特征选择和降维都是为了减少特征数量。但与降维不同的是，特征选择只是从原有特征中进行选择或排除，不涉及原有特征的转变。常见的特征选择方法主要包括过滤式特征选择（Filter Method）、包裹式特征选择（Wrapper Method）以及嵌入式特征选择（Embedded Method）三种。

1. **过滤式特征选择**

过滤式特征选择Filter Feature Selection）是一种特征选择方法，通过某些统计量或度量来评估每个特征与目标变量之间的关系，然后根据这些评估结果来选择特征。这种方法的优点是计算速度快，可以处理高维数据集，并且可以减少后续学习算法的维度灾难问题。

常见的用来评估特征的统计方法有：

1. **相关系数（Correlation Coefficient）**：相关系数衡量两个变量之间的线性关系强度和方向。最常见的是皮尔逊相关系数（Pearson correlation coefficient），通常表示为r。其值的范围在-1到1之间，其中1表示完全正相关，-1表示完全负相关，0表示没有线性相关性。在特征选择中，可以通过计算每个特征与目标变量之间的相关系数来评估它们的线性关系。具有高绝对值的相关系数的特征可能与目标变量更相关。

2. **互信息**（Mutual Information）：互信息衡量两个变量之间的相互依赖性，它可以用于评估特征与目标变量之间的非线性关系。互信息的值越高，表示两个变量之间的依赖性越强。

互信息不局限于线性关系，也捕捉到特征与目标变量之间的复杂非线性关系，因此它在特征选择中特别有用。计算互信息通常需要估计概率分布，这可能在数据量较少时变得不准确。

3.**卡方检验**（Chi-Square Test）：卡方检验是一种统计方法，用于检验两个分类变量之间的独立性。在特征选择中，卡方检验可以用来评估特征的类别与目标变量的类别之间是否存在显著的关联。如果卡方检验的p值小于某个显著性水平（如0.05），则可以拒绝原假设，认为特征与目标变量之间存在关联。

4. **ANOVA（方差分析）**：ANOVA是一种统计方法，用于检验三个或更多组之间的均值是否存在显著差异。在特征选择中，ANOVA可以用来评估特征的不同水平是否与目标变量的均值有显著关联。如果ANOVA的p值小于显著性水平，这表明至少有一个组的均值与其他组显著不同，从而表明该特征可能与目标变量相关。

1. **包裹式特征选择**

包裹式特征选择（Wrapper Feature Selection）是一种基于机器学习模型进行特征选择的方法，尤其适用于特征数量相对较少的情况。给定一个机器学习模型，通过遍历所有可能的特征子集，找到最佳的子集组合，使得模型在该子集上的性能达到最优。具体步骤如下：

**1. 确定评估标准**

在开始特征选择之前，需要确定一个或多个评估标准来衡量模型的性能。这些标准可以是分类准确率、精确度、召回率、F1分数、ROC曲线下面积（AUC）、均方误差（MSE）等。评估标准具体取决于模型任务是分类、回归还是其他类型。

**2. 选择初始特征子集**

从原始特征集中选择一个初始子集作为起点。这个子集可以是全部特征，或者是通过其他方法（如过滤式特征选择）筛选出的一部分特征。

**3. 生成特征子集**

使用某种搜索策略来探索可能的特征子集。常见的搜索策略包括：

穷举搜索：尝试所有可能的特征子集组合。

递归特征消除（RFE）：递归地构建模型并移除权重最小（或贡献最小）的特征，直到达到所需数量的特征。

前向选择：从一个空集开始，逐步添加对模型性能提升最大的特征。

后向消除：从包含所有特征的集合开始，逐步移除对模型性能影响最小的特征。

启发式搜索：使用遗传算法、模拟退火等启发式方法来搜索最优特征子集。

**4. 训练和评估模型**

对于每个生成的特征子集，使用相应的机器学习算法来训练模型，并使用交叉验证等方法来评估模型的性能。这一步是包裹式特征选择的核心，因为它直接涉及到模型性能的评估。

**5. 选择最佳特征子集**

比较所有已评估的特征子集的性能，并选择性能最好的子集。通常，这可以通过比较交叉验证的平均分数或其他评估指标来完成。

**6. 验证最终模型**

使用选定的最佳特征子集训练最终模型，并在一个独立的测试集上进行验证。这一步可以帮助评估模型的泛化能力，并确保选择的特征子集不会过拟合。

1. **嵌入式特征选择**

嵌入式特征选择（Embedded Feature Selection）是一种在模型训练的过程中同时进行特征选择的方法。在训练过程中，模型会根据特征对预测结果的贡献自动调整特征的权重。一些特征可能会获得较高的权重，而其他特征的权重可能接近于零。通过分析这些权重，我们可以确定哪些特征对模型的预测性能更为重要。

支持嵌入式特征选择的机器学习模型通常具有能够评估特征重要性的机制，如决策树、随机森林、梯度提升机（Gradient Boosting Machine，GBM）、正则化线性模型（如Lasso）等。

**1. 决策树（Decision Trees）**

决策树是一种基本的分类与回归方法。在嵌入式特征选择中，决策树通过选择最佳分割特征来构建树的节点。特征的重要性可以通过以下方式评估：

**信息增益（对于分类树）**：在每个节点，计算使用特征进行分割前后的熵（或基尼不纯度）的减少量。信息增益越大的特征，其重要性越高。

**平均绝对误差减少**（对于回归树）：计算特征分割引起的平均绝对误差的减少量。减少量越大，特征越重要。

**2. 随机森林（Random Forest）**

随机森林是一种集成学习方法，通过构建多个决策树并进行投票（分类问题）或平均（回归问题）来提高预测性能。在随机森林中，特征的重要性可以通过以下方式评估：

**平均不纯度减少**：对于每个特征，计算它在所有树上的平均不纯度减少量。不纯度减少量越大，特征越重要。

**平均精确度增加**：在构建树的过程中，通过随机选择特征的子集来评估特征的重要性。特征在随机选择时对模型性能的提升越大，其重要性越高。

**3. 梯度提升机（GBM）**

梯度提升机（Gradient Boosting Machine）是也是一种集成学习算法，通过逐步添加弱学习器（通常是决策树）来最小化损失函数。在GBM中，特征的重要性通常通过以下方式评估：

梯度变化：在每次迭代中，特征对损失函数梯度的影响可以作为其重要性的指标。影响越大，特征越重要。

分裂次数：特征在树中作为分裂节点出现的次数也可以作为重要性的一个指标。分裂次数越多，特征越重要。

**4. 正则化线性模型（如Lasso）**

Lasso（Least Absolute Shrinkage and Selection Operator）是一种线性回归模型，它通过对系数施加L1正则化来进行特征选择。在Lasso模型中，特征的选择通过以下方式实现：

系数缩减：L1正则化会推动模型将不重要的特征的系数缩减至零。因此，非零系数的特征被认为是重要的。

路径追踪：在Lasso路径追踪过程中，可以观察到特征系数从非零减小到零的过程。在这个过程中，最后进入模型的特征被认为是最重要的。

在实际应用中，嵌入式特征选择可以与其他特征选择方法结合使用，以提高特征选择的效果。例如，可以先使用过滤式方法进行初步筛选，然后再应用嵌入式方法进行细化。此外，对于复杂的数据集，可能需要尝试多种模型和特征选择策略，以找到最适合问题的解决方案。通过综合这些方法，可以更有效地提高模型的性能和解释性。

**3.3数据可视化**

**3.3.1数据可视化简介**

数据可视化是一种利用图形、图表、地图等视觉形式将数据和信息呈现出来的技术，以增强人类的认知能力。人眼是一个具有高带宽的巨量视觉信号输入并行处理器，有很强的模式识别能力，对可视符号的感知速度要远远高于对数字或者文本。因此，数据可视化利用人类视觉感知的高速和高效特性，将不可见或者难以有效显示的数据转换为可感知的图形、符号、颜色、纹理等，增强数据的识别效率，使得复杂的数据集易于理解。

数据可视化不仅可以展示数据的趋势、模式、异常和相关性，还能帮助用户更快速地做出决策和发现洞见。从宏观的角度看，可视化的作用主要体现在以下三个方面：

* **记录信息**：可视化最直接的作用是作为一种信息记录手段。通过图表、图形和图像等形式，复杂的数据和概念得以以视觉化的形式呈现。相较于纯文本，这种形式的信息记录更易于被大脑识别和记忆，因为它充分利用了人类的视觉感知能力。例如，时间序列图可以帮助追踪数据随时间的变化，地图则展示了地理位置相关的信息。可视化能够将庞大的数据浓缩成简洁、直观的图形，方便存储和回顾。
* **分析推理**：可视化在数据分析和推理过程中扮演着关键角色。将信息以可视符号的方式呈现给分析者，将提高分析者分析数据的效率。它有助于发现数据中的模式、趋势和异常，从而引导提出假设、进行探索性分析和验证结果。通过可视化，我们可以更直观地理解数据之间的关系。比如，使用散点图帮助探索两个变量之间的相关性，或者使用热图来识别数据中的热点区域。此外，可视化还能够帮助在数据挖掘和机器学习过程中选择和调整模型，通过可视化模型的性能指标（如精确度、召回率）来优化模型参数。
* **信息传播与协同**：人的视觉感知输入占据了人从外界获取信息的很大比例（70%以上）。因此，在信息传达中，利用可视化技术来呈现数据和信息是非常重要的。可视化能够有效传播信息，并促进团队或组织之间的协同工作。优秀的可视化能够跨越语言和专业知识的障碍，使得非专业人士也能理解复杂的数据和分析结果。对于决策者来说尤其重要，因为他们需要基于数据分析的结果做出战略性决策。此外，在团队协作中，可视化还能够帮助成员共享发现、讨论结果和协作解决问题。例如，仪表板（Dashboard）就是一种集成了多种可视化元素的工具，它可以实时展示关键指标和分析结果，使得团队成员能够及时了解项目状态并作出响应。

**3.3.2商业智能过程中的数据可视化**

数据可视化与商业智能的关系是相辅相成的。在整个商业智能流程中，可视化不仅是一个独立的步骤，而且与商业智能的其他环节紧密相连，为商业智能的每个阶段提供支持和洞察。

**1. 数据收集和整理：**数据准备阶段，目标是收集、清洗和整理数据，确保数据质量满足分析要求。数据管理涉及到数据的存储、备份和安全性等方面。在这个阶段，可视化可以用来识别数据中的异常值、缺失值和数据分布情况。例如，通过绘制箱线图可以发现异常值，通过直方图或密度图可以了解数据的分布情况。

**2. 数据探索性分析：**探索性分析是数据分析的前期阶段，通过可视化和统计方法来探索数据的特征和模式。例如，散点图矩阵可以揭示变量之间的关系，热图可以展示变量间的相关性，而箱线图或小提琴图可以展示不同类别的分布特征。另外，可视化在这个阶段有助于形成假设和指导后续的数据挖掘算法选择。例如，通过可视化数据的时间序列趋势，可以选择适合时间数据的挖掘算法。

3. **数据分析和建模**：在收集和整理数据后，商业智能系统会对数据进行分析和建模，以发现数据中的模式、趋势和关联。这可能涉及使用统计分析、机器学习和预测建模等技术。分析的结果可以用来帮助组织理解当前业务状况、预测未来趋势，并做出决策。数据可视化在这个阶段可以用来展示算法的结果，帮助我们理解模型的性能和发现的模式。例如，混淆矩阵图可以展示分类算法的准确性，而ROC曲线可以评估模型的区分能力。

**4. 分析报告和沟通：**数据可视化的结果通常会被整合到报告和沟通中，以便与利益相关者分享分析结果和洞见。这些报告可能是定期的业务报告，也可能是针对特定问题或项目的临时报告。通过报告和沟通，组织可以将分析结果转化为实际行动，并促进组织内部的知识共享和决策制定。

在整个商业智能系统中，可视化是一个重要的工具，它帮助我们更好地理解数据，指导数据分析算法的选择和调整，并有效地传达分析结果。数据分析算法提供了从数据中提取知识和模式的能力，而可视化则使得这些知识和模式更容易被理解和解释。两者的结合使得数据分析更加高效和有洞察力。数据可视化在商业智能过程中的重要性体现在以下几个方面：

* **增强理解和洞察**：数据可视化使复杂的数据信息更易于理解和解释，帮助用户发现数据中的模式、趋势和关联，从而提供洞察力支持决策制定。
* **提升决策效率**：通过直观地呈现数据和分析结果，数据可视化可以帮助决策者更快速地理解数据，并做出准确的决策，从而提升决策的效率。
* **促进沟通和共享**：数据可视化的结果可以被整合到报告和沟通中，促进组织内部的知识共享和沟通，从而推动组织的发展和创新。
* **提高透明度和信任**：通过公开透明地呈现数据和分析结果，数据可视化可以增强组织内部和外部利益相关者对决策和行动的信任和认可，从而提高组织的声誉和竞争力。

**3.4数据可视化基本图表**

统计图表作为数据可视化的主要形式之一，也是各个可视化呈现中的基本元素。常用的统计图表包括柱状图、折线图、饼图、面积图、散点图、箱线图、箱线图等，他们主要通过可视化原始数据的属性值来直观呈现数据特征。不同的图表用来主要揭示数据的四个不同方面的特征或应用：比较与比例、关系、趋势和模式、以及分布等。

**趋势特征**：描述数据随时间或其他连续变量的变化趋势，反映了数据在长期内的变化趋势，可以是上升、下降或波动。例如，销售额随季节变化的趋势、股票价格随时间的趋势等都是趋势特征。

**关系特征**：描述不同变量之间的关系或相关性，表现为变量之间的相互依赖或影响关系，可以是正相关、负相关或无关。例如，气温与冰淇淋销量之间的关系、广告投入与销售额之间的关系等都是关系特征。

**分布特征**：描述数据在特定范围内的分布情况，展示了数据的集中程度、离散程度和分布形状，可以是正态分布、偏态分布或其他类型的分布。例如，人口年龄分布、产品销售区域分布等都是分布特征。

**比较/比例特征：**涉及对不同部分在整体中的占比进行比较。比较将不同组或类别的数据进行对比，以找出它们之间的差异或相似之处。而比例描述数据的相对大小或占比关系，常用来表示一个数值相对于另一个数值的大小关系，或者表示一个数值在整体中的占比情况。

本节将介绍一些有代表性的基本统计图表（如下图），并介绍其适用场景。随着技术发展，越来越多的可视化工具还能支持组合图、动态图和交互图等。

图 基本统计图表及类型

**3.4.1数据可视化的常用图表**

**（1）折线图（Line Plot）**

使用直线段来连接一系列的点，用于显示数据随时间或其他连续变量的**变化趋势**。折线图通常用于展示时间序列数据或其他连续变量之间的关系。折线图通常由横轴（X轴）和纵轴（Y轴）组成，横轴表示时间或其他连续变量，纵轴表示相应的数值。折线图的绘制方式是通过在坐标系中连接数据点，并用直线将它们连接起来，从而形成折线。每个数据点的位置由相应的时间或连续变量的值决定。折线图可以同时显示多个数据序列，每个序列通常用不同的颜色或线型来区分。

（2）**柱状图（Bar Chart）和堆叠的柱状图（Stacked Bar Chart）**

采用长方形的形状和颜色编码数据的属性，用于比较不同类别或组之间的数据。柱状图通常由横轴（X轴）和纵轴（Y轴）组成，横轴表示不同的类别或组，纵轴表示相应的数值。

柱状图的绘制方式是在横轴上为每个类别或组绘制一个垂直的柱子，并用柱子的高度表示相应类别或组的数值大小。每个柱子的高度通常对应着数据的数量、频率或比例。如图3.2所示，房屋建造时间大多集中在2000年至2010年，在后续预测房价时，可以依据这个特征对数据集进行划分。此图为原数据集的部分数据。

**示例1**：柱状图分析房屋建造年份情况

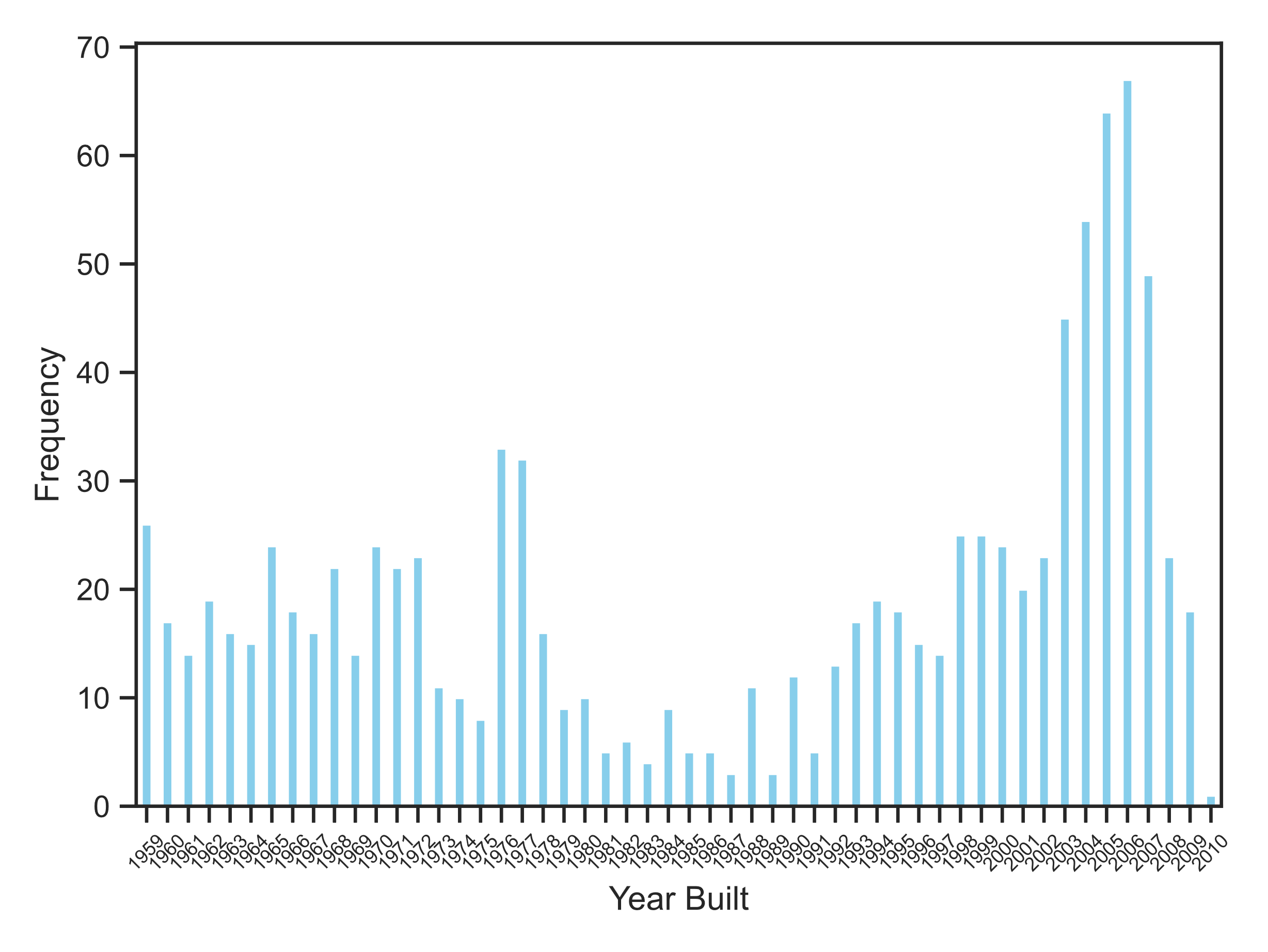


图3. 1 不同年份建造房屋（YearBuild）的分布柱状图

**堆叠的柱状图**是柱状图的一种变体，将每根直柱内部也可用像素图的方式编码，每个柱子代表一个整体，柱子中不同颜色代表不同类别。因此，它除了用于比较不同类别或组的数据，可同时显示每个类别或组内部各部分的比例关系。绘制堆叠柱状图时，每个柱子的高度仍然代表着整个类别或组的总数值，但是柱子被分割成多个部分，每个部分的高度表示该子类别或部分数据在整体中的比例。不同子类别或部分的高度叠加在一起，形成整个柱子的高度。如图3.2所示，高于平均售价的房子视为high level，低于平均售价的房子视为low level。堆积柱状图反映了不同建造年份的不同类型房子的分布情况。

**示例2**：堆积柱状图分析不同类型房屋建造年份情况

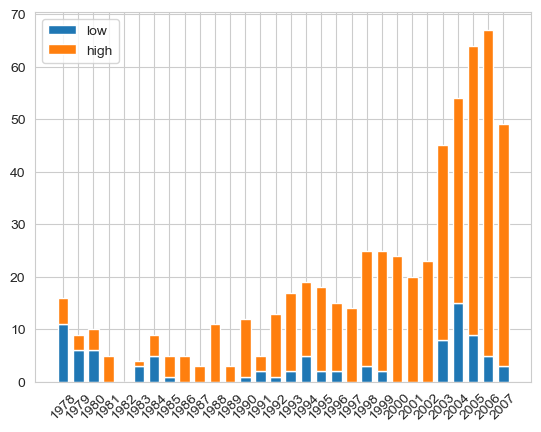


图3. 2 显示不同类型房屋在不同年份建造的分布的堆积柱状图

**（3）直方图（Histogram）：**用于表示数据的分布情况，直方图将数据分成若干个等距的区间（也称为“箱”或“柱”），并统计每个区间中数据的频数或频率，然后将这些频数或频率用柱子的高度表示。直方图通常由横轴（X轴）和纵轴（Y轴）组成，横轴表示数据的取值范围，纵轴表示数据在每个区间中的频数或频率。

绘制直方图时，需要确定数据的区间宽度，并将数据按照这些区间划分。然后统计每个区间中数据的频数或频率，并将其用柱子的高度表示。通常，相邻柱子之间没有间隙，以突出数据的连续性。

在绘制直方图时，需要选择合适的区间宽度和柱子数目，以确保图表清晰易懂，同时需要注意选择合适的标签和标题，以便清晰地传达数据的含义。

**（4）散点图（Scatter Plot）、气泡图（Bubble Chart）和散点图矩阵（Scatter Plot Matrix）**

散点图（Scatter Plot）用于展示两个变量之间的关系。在散点图中，每个数据点由两个数值变量的数对（x, y）表示，其中一个变量对应横轴（X轴），另一个变量对应纵轴（Y轴）。通过在坐标系中绘制每个数据点，可以清晰地观察到两个变量之间的关系，如线性相关性、非线性关系、聚集趋势等。如图3.3所示，房屋售价总体与建造年份成正相关。2000年前后建造的房屋售价方差较大，相对地，1920年到1960建造的房屋售价相差不大。

**示例3**：散点图分析房屋建造年份和售价关系

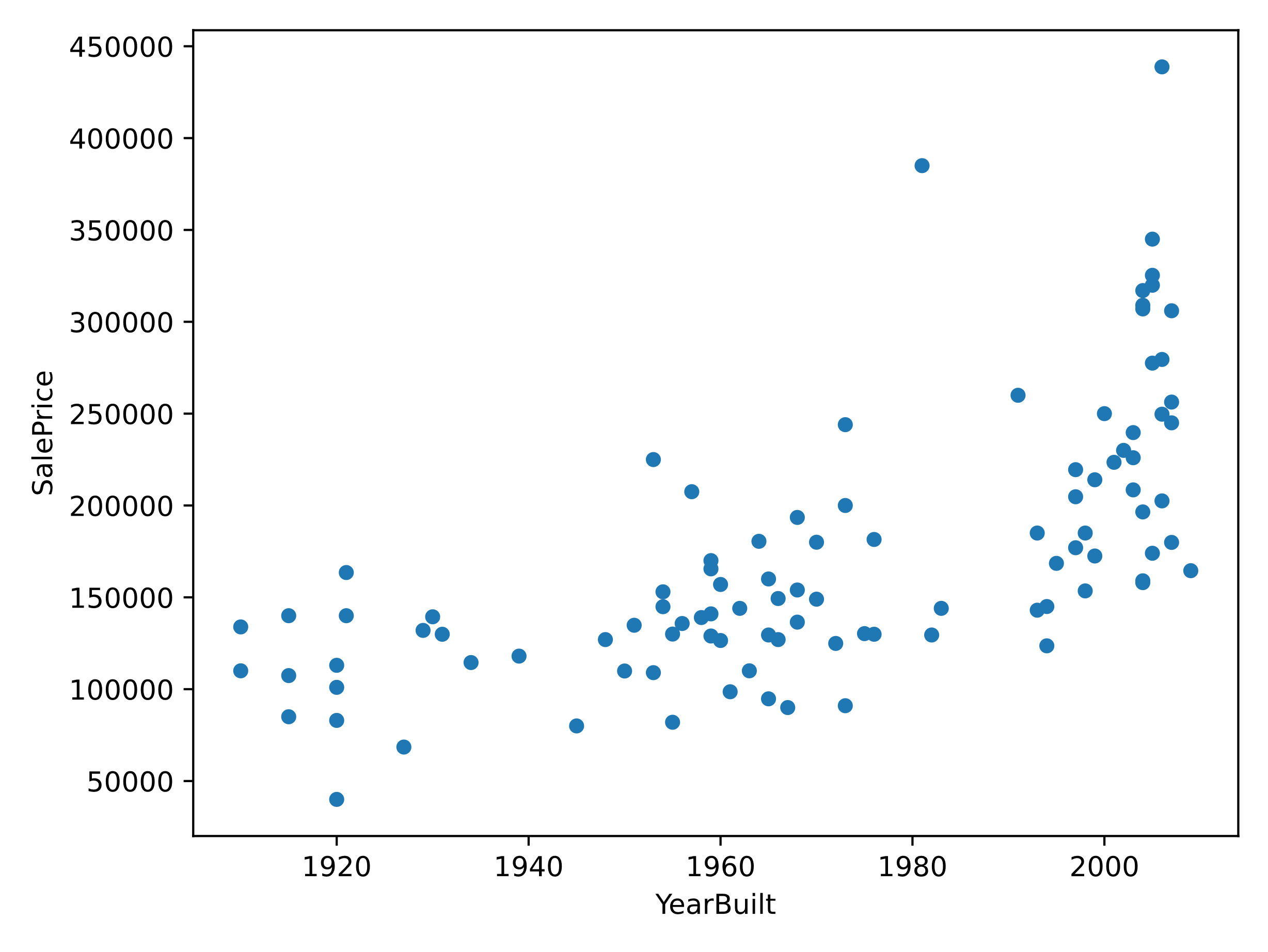


图3. 3房屋建造年份与售价关系的散点图

**气泡图**为散点图的一种变体，可以用气泡的大小表示多一个维度的信息。在气泡图中，每个数据点由三个数值变量（x, y, z）表示，其中一个变量对应横轴（X轴），另一个变量对应纵轴（Y轴），而第三个变量则通过气泡的大小来表示。

绘制气泡图时，数据点以圆形气泡的形式展示在坐标系中，气泡的位置由前两个变量决定，而气泡的大小则由第三个变量的数值大小决定。通常情况下，数据值较大的气泡会显示为较大的圆形，而数据值较小的气泡则显示为较小的圆形。如图3.4所示，气泡的大小反应了房屋售价的高低，可以看出，房屋越新，售价越高。

**示例4**：房屋建造年份和出售年份与房价的关系

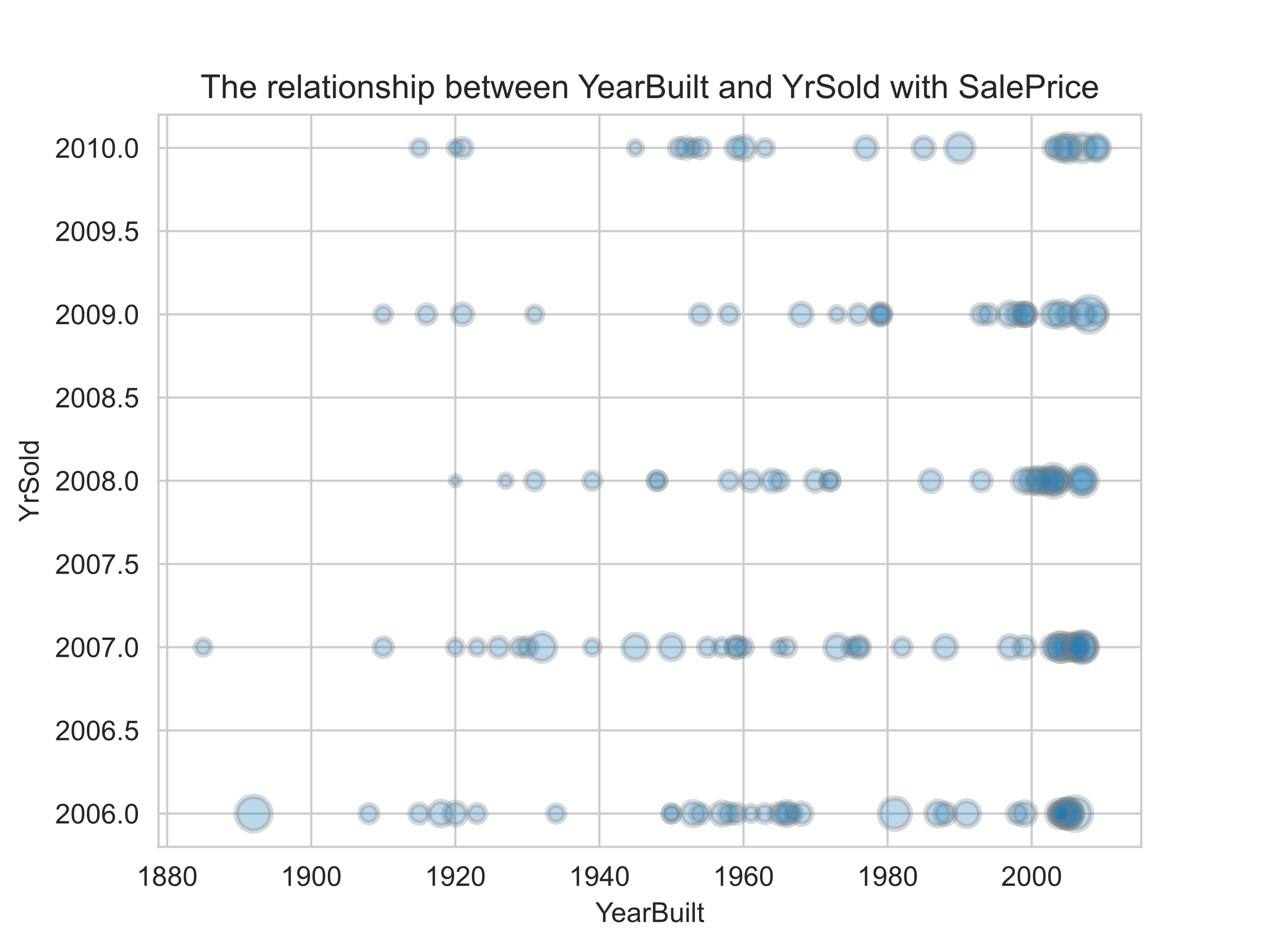


图3. 4房屋的建造年份与出售年份及买卖价格（气泡大小）的关系

**散点图矩阵（Scatter Plot Matrix）**是一种特殊的散点图，用于同时显示多个变量之间的关系。在散点图矩阵中，每个变量都与其他变量两两组合，形成一个矩阵，矩阵中的每个小格子显示对应两个变量之间的散点图。矩阵对角线呈现对应属性的数据分布情况。

**示例5：**四个属性的散点矩阵图

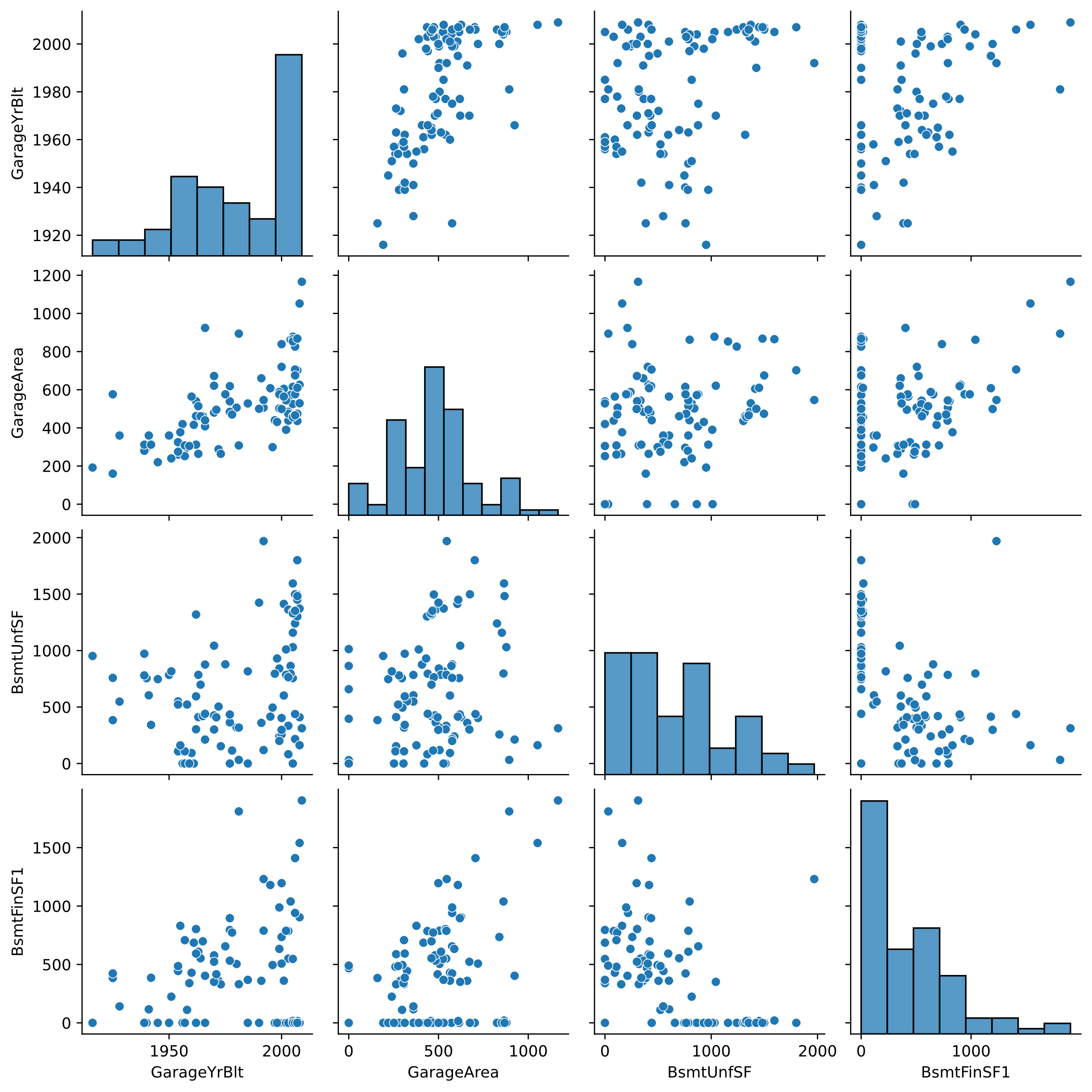


图3. 5散点矩阵图

**（5）箱线图（Boxplot）和小提琴图**（Violin Plot）

箱线图（**Boxplot）**，也称为盒须图，是一种用于显示数据分布情况的统计图表，适合展示数据的分散程度和异常值情况。它将数据按照五个统计量进行展示：最小值、第一四分位数（Q1）、中位数（Q2）、第三四分位数（Q3）和最大值。

在箱线图中，通常绘制一个矩形箱子，箱子的上边界为第三四分位数（Q3），下边界为第一四分位数（Q1），箱子中间的线表示中位数（Q2）。箱子内部的长度代表了数据集的中间50%范围，即四分位距（IQR）。箱子外部的线段称为“须”，延伸到最小值和最大值，除去异常值。异常值通常被单独标记出来，位于须之外。如图3.6所示，第一个箱子显示的是房屋建造年份（YearBuilt）的分布情况，中位数Q2大概在1970-1980年中间。箱线图中的横线线从上到下依次表示：最大值、四分之三分位数、中位数、四分之一分位数、最小值。

**示例5**：箱线图展示房屋建造年份

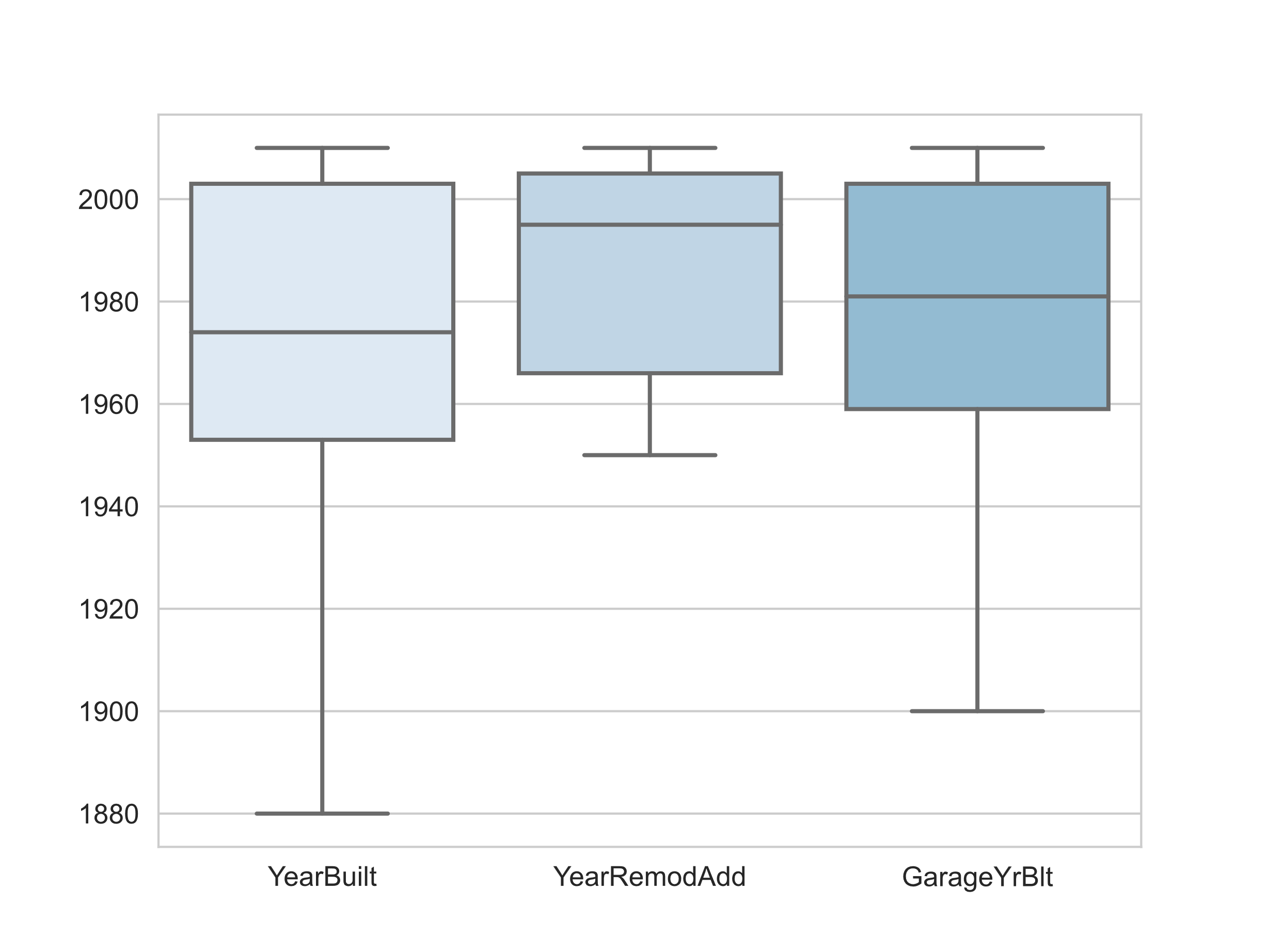


图3. 6箱线图

**小提琴图（Violin Plot**）是属于箱线图的变体，其呈现结合了箱线图和核密度估计图的特点，用于显示数据分布情况。它与箱线图类似，但提供了更丰富的信息，特别适用于展示数据的密度分布和比较不同类别或组之间的数据分布。

在小提琴图中，每个数据分布被绘制成一个“小提琴”形状的曲线，曲线的宽度表示该位置上数据的密度，而曲线的高度则反映了数据的频率分布。小提琴图的中心部分通常是箱线图，显示了数据的四分位数、中位数和异常值，而曲线的延伸部分则显示了数据的密度分布情况。图3.7用小提琴图展示了房屋建造年份的分布情况。小提琴身展示数据点的分布密度，琴弦展示数据的四分之三分位数、中位数、四分之一分位数、95%置信区间，以及最大最小值。

**示例6**：小提琴图展示房屋建造年份

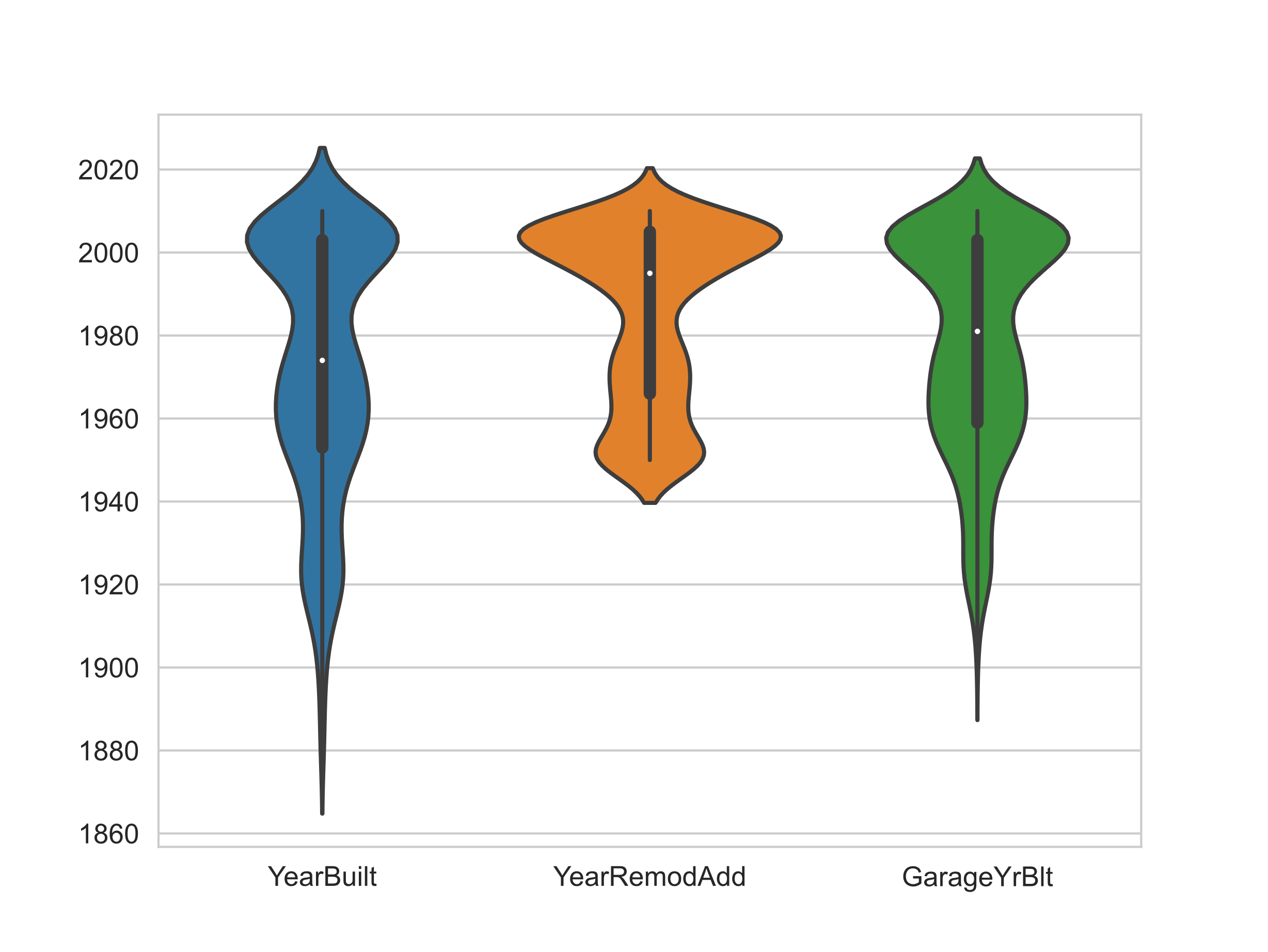


图3. 7小提琴图

（6）**热力图（Heatmap）**

热力图（Heatmap）是一种用颜色编码来展示数据矩阵的图表，其中数据值用颜色的不同深浅程度来表示。热力图通常用于可视化二维数据，其中行和列分别代表数据集中的变量或类别，而每个单元格的颜色则表示对应变量之间的关系或相似度。

在热力图中，数据值越大的单元格通常显示为较深的颜色（如红色或深蓝色），而数据值较小的单元格则显示为较浅的颜色（如浅黄色或浅蓝色）。通过颜色的变化，观察者可以直观地识别出数据集中不同变量之间的关系强度和模式。图3.8用热力图展示了Kaggle房屋数据数据的相关性矩阵，如图所示，GarageYrBlt和GarageArea有较强的正相关性，他们对房价的影响可能相似，在特征分析过程中可以做降维处理。而BsmtUnfSF和BsmtFinSF1有较强的负相关性。

**示例7**：热力图分析房屋预测数据相关性

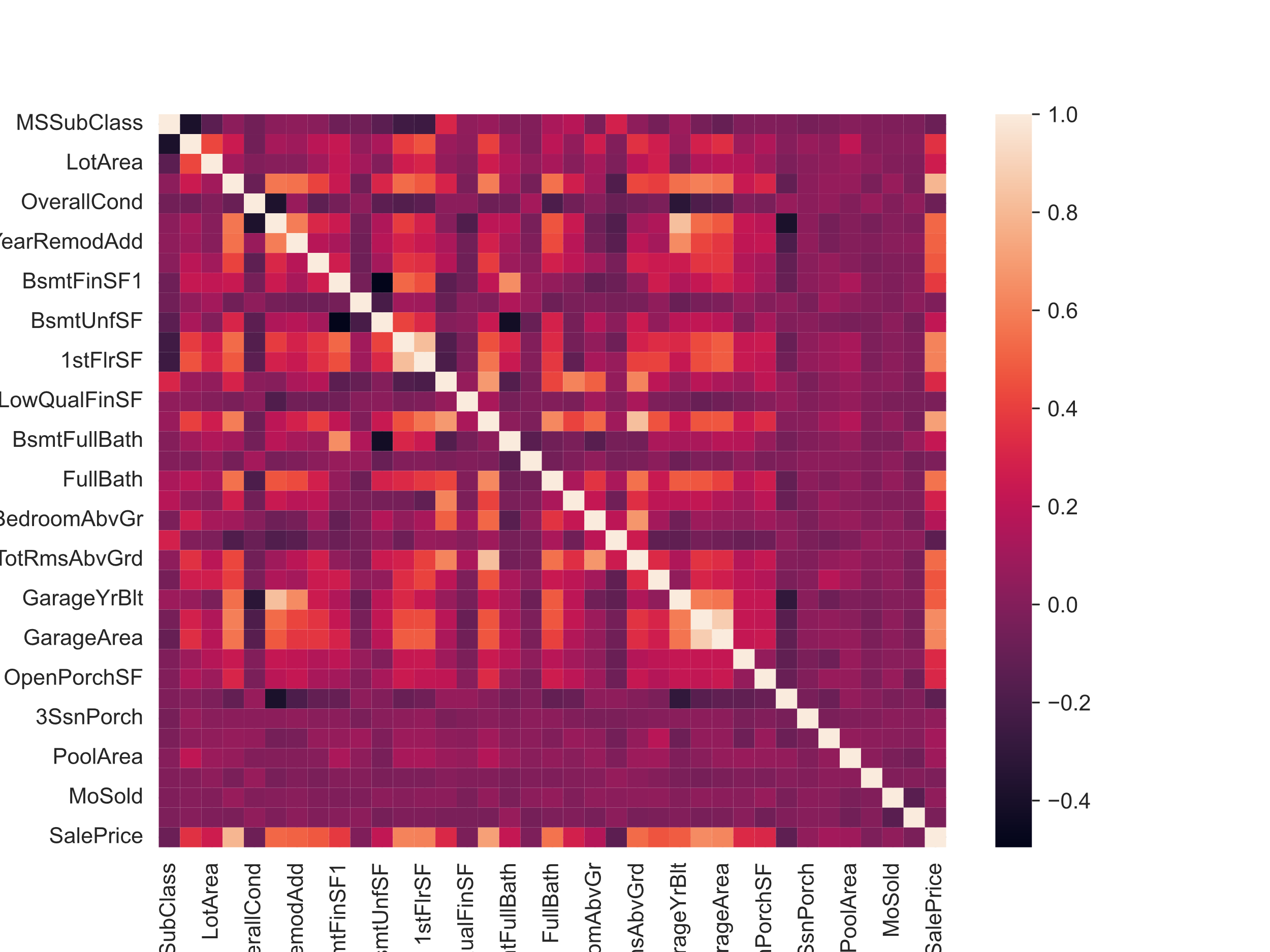


图3. 8房屋预测数据热力图

**（7）饼图（Pie Chart）**

饼图采用了饼干的隐喻，用于展示各部分占整体的比例关系。饼图将数据分成多个部分，每个部分的大小与其所代表的数据值的比例成正比，通常以扇形的形式呈现。

在饼图中，整个圆形代表了总体数据，而每个扇形代表了数据中的一个部分或类别。每个扇形的角度大小与其所代表的数据值所占比例成正比，因此大的数据值对应的扇形角度较大，而小的数据值对应的扇形角度较小。图3.9用饼图分析地上房间数量比例情况。

**示例8：**饼图分析地上房间数量情况

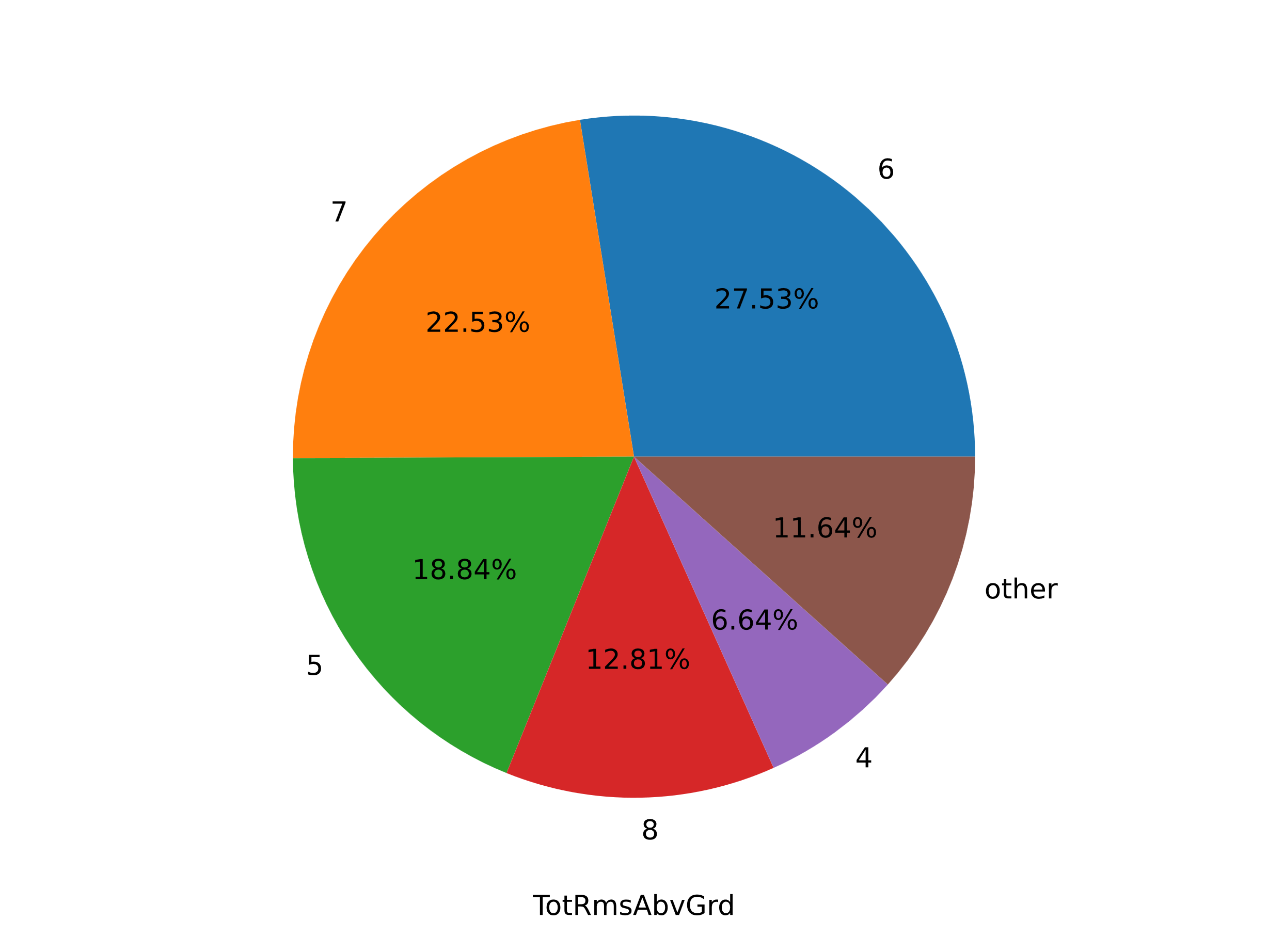


图3. 9地上房间数量饼图

**（8）面积图（Area Chart）和堆叠的面积图（Stacked Area Chart）**

面积图（Area Chart）是一种用来展示数据随时间或其他连续变量变化趋势的图表。它与折线图类似，但是将折线下方的区域填充，用颜色或阴影表示数据的大小，从而形成一个填充的面积。面积图通常用于展示数据随时间的变化趋势，以及不同类别或组之间的比较。

堆叠的面积图（Stacked Area Chart）则是在面积图的基础上进行了扩展，将多个数据序列的面积堆叠在一起，形成一个总体的填充面积。每个数据序列的面积都叠加在前一个序列的上方，因此整个图形呈现出分段堆叠的效果。堆叠的面积图常用于展示不同类别或组之间的总体趋势以及各个部分的相对贡献。

需要注意的是，在使用堆叠的面积图时，需要确保各个部分之间的叠加顺序正确，并且要避免过多的堆叠部分导致图表难以理解。此外，为了图表的清晰和易读，应该为每个数据序列选择明显区别的颜色或图案，并添加适当的标签和标题。

**示例9：**不同年份不同类型房屋建造数量的面积堆叠图

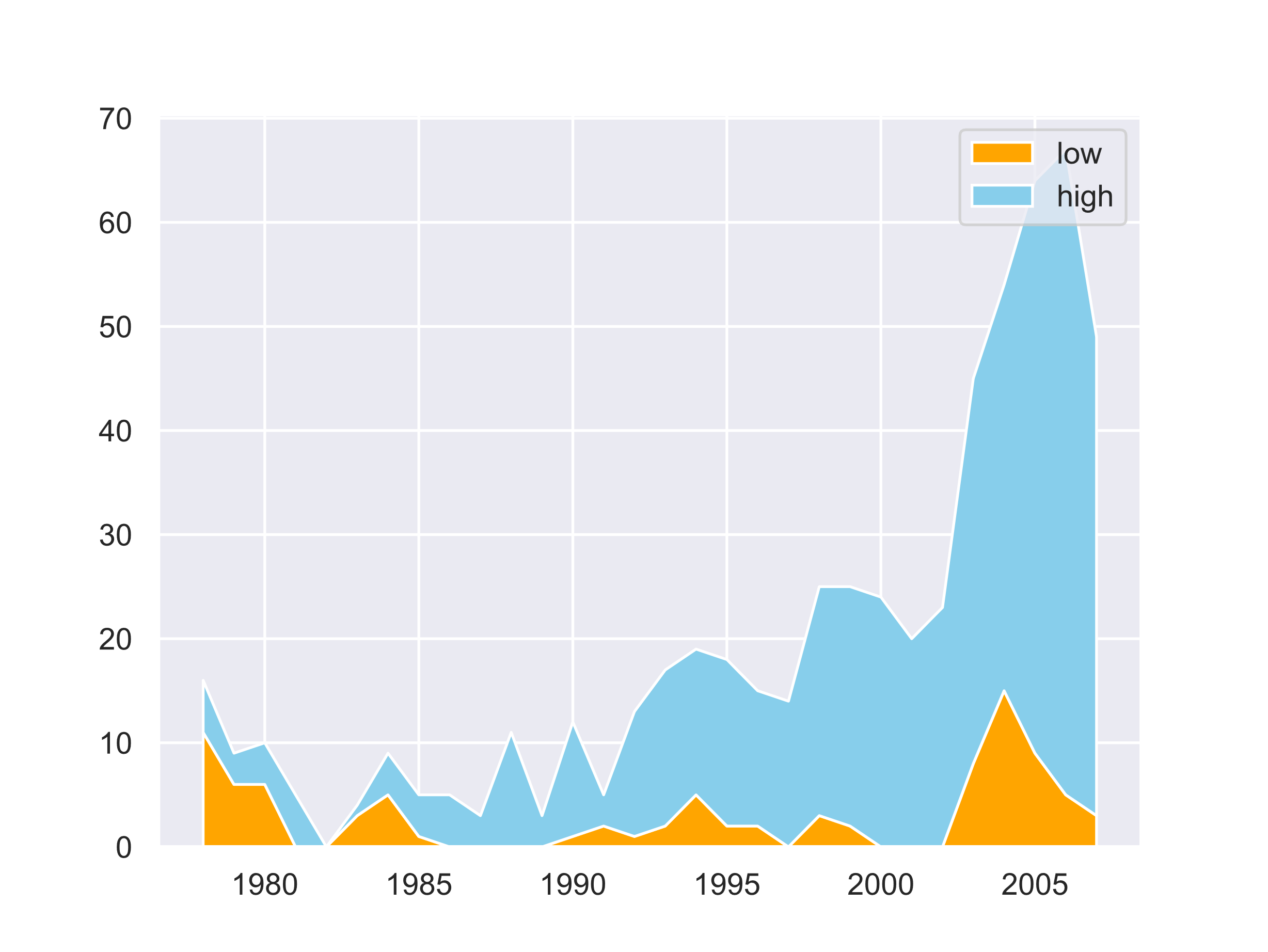


图3. 10显示不同类型房屋在不同年份建造的分布的面积堆叠图

**（9）雷达图（Radar Chart）**

雷达图（Radar Chart），也称为极坐标图或蜘蛛图，是一种多变量数据可视化图表，用于展示多个变量之间的关系和比较。雷达图通过在一个圆形或多边形的坐标系中绘制多个轴线，并沿着这些轴线绘制数据点，来展示多个变量的值。

在雷达图中，每个变量通常沿着坐标系的一条轴线表示，而数据点的位置则代表该变量的值。多个变量的数据点连接起来，形成一个封闭的多边形或区域，这个多边形的形状和大小反映了各个变量之间的关系和相对重要性。

尽管雷达图在多变量比较和特征分析方面具有一定优势，但也存在一些局限性。例如，雷达图不擅长展示大量的变量或数据，过多的变量会导致图表复杂度增加，难以理解。此外，雷达图的数据点连接线可能会交叉，使得图表难以阅读。

**示例10：**不同类型房屋各维度评分雷达图



图3. 11房屋各维度评分雷达图

在解读雷达图时，需要注意各个变量的位置和值，以及图表中的标签和标题，以确保清晰地传达多变量之间的关系和比较情况。

**（10）平行坐标图（Parallel Coordinates Plot）**

平行坐标图（Parallel Coordinates Plot）是一种用于可视化多维数据的图表，特别适用于展示多个变量之间的关系和趋势。在平行坐标图中，每个变量用坐标系中的一条平行线表示，而数据点则沿着这些平行线进行连接，形成一条折线。

在平行坐标图中，每个变量对应于坐标系中的一条垂直线，而每个数据点则表示一个样本或观测值。通过连接各个数据点形成的折线，可以清晰地展示出多个变量之间的关系、趋势和模式。

尽管平行坐标图在展示多维数据方面具有一定优势，但也存在一些局限性。例如，当变量较多时，平行坐标图可能会变得混乱和难以理解。此外，平行坐标图中的数据点连接线可能会交叉，使得图表难以阅读。

**示例11：**部分数据建造年份、改建年份、车库建造年份和销售年份平行坐标图

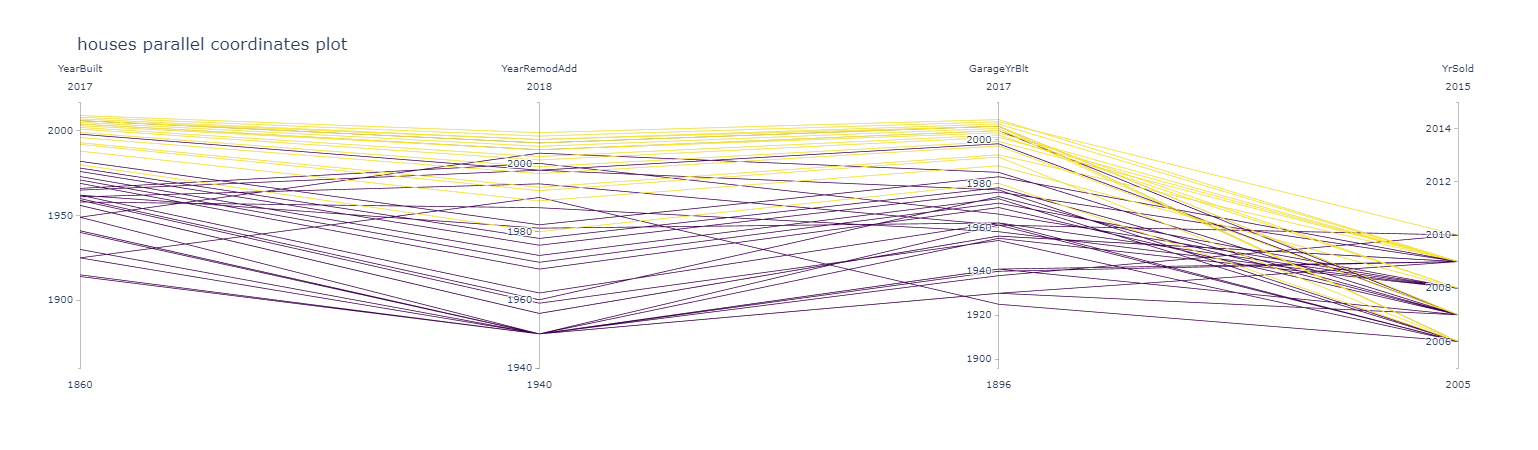


图3. 12平行坐标图

在解读平行坐标图时，需要注意各个变量的位置和值，以及数据点之间的连接情况，以确保清晰地传达多维数据之间的关系和趋势。

**3.4.2交互的可视化（D3.js）**

D3.js（Data-Driven Documents）是一个基于Web的JavaScript库，用于使用HTML、SVG和CSS创建动态、交互式的数据可视化。D3允许将数据绑定到DOM（文档对象模型），然后应用数据驱动的转换到文档。它提供了强大的可视化组件和丰富的API，让使用者能够创建复杂的数据可视化效果。

除了D3.js工具，还有很多其他工具可以支持交互可视化。这些工具涵盖了各种语言和平台，包括Python和R的开源包，如Plotly、Bokeh和ggplot2，它们提供了丰富的交互功能。此外，还有许多其他JavaScript工具，如Highcharts、Chart.js和Leaflet.js，以及商用工具如Power BI和Tableau，它们都拥有强大的可视化功能，适用于不同的需求和场景。

案例：图3.6是汉斯-罗斯林（Hans Rosling）著名的调节器可视化。它按地区动态地显示了180个国家在过去 209 年中的人均收入（x）、预期寿命（y）和人口（圆圈面积）。

|  |  |
| --- | --- |
|  |  |
| 1821年国家财富、预期寿命分布图 | 1900年国家财富、预期寿命分布图 |
|  |  |
| 1977年国家财富、预期寿命分布图 | 2009年国家财富、预期寿命分布图 |

图3. 13 D3动态散点图实例

**3.5 本章小节**

本章主要介绍了数据预处理和数据可视化的重要性以及其主要方法，涵盖了数据清洗、数据规范化和数据离散化等方面的内容。具体来说，本章的重点包括：

1. 数据预处理动机与主要方法： 我们首先了数据预处理的动机，即为了提高数据质量、减少噪声和错误，以便更有效地进行数据分析和建模。主要方法包括数据清洗、数据规范化和数据离散化。
2. 数据清洗方法：数据清洗是数据预处理的第一步，其目的是识别和处理数据中的错误、缺失值和异常值。主要的数据清洗方法包括删除错误数据、填补缺失值和识别处理异常值。
3. 数据规范化方法：数据规范化是将数据转换为统一的尺度或范围，以便更好地进行比较和分析。主要的数据规范化方法包括最小-最大规范化、Z-score规范化和小数定标规范化。
4. 数据离散化方法：数据离散化是将连续型数据转换为离散型数据的过程，常用于数据挖掘和机器学习中。主要的数据离散化方法包括等宽离散化、等频离散化和基于聚类的离散化。
5. 数据可视化及其作用：我们学习了数据可视化的概念及其作用，包括记录信息、分析推理和信息传播与协同。数据可视化通过图表、图形和图像等形式展示数据，使得复杂数据易于理解，并能揭示数据中的趋势、模式和相关性，从而帮助用户更快地做出决策和发现洞见。
6. 常见的数据可视化工具和方法：我们介绍了一些常见的数据可视化工具和方法，包括折线图、柱状图、散点图、雷达图等，以及如何根据不同的数据特征选择合适的可视化方法。

通过本章的学习，读者可以对数据预处理的基本概念和常用方法有所了解，并掌握一些常见的数据可视化工具和方法，为后续的数据分析和建模打下良好的基础。习题：

1. 简述数据预处理的必要性和主要任务。

2. 请将表3.8中数据进行规范化，规范化到区间[0,1]。

表3. 8习题2对应的数据

|  |  |  |  |
| --- | --- | --- | --- |
| 身高(m) | 体重(kg) | 身高(m) | 体重(kg) |
| 1.62 | 55 | 1.68 | 62 |
| 1.65 | 57 | 1.75 | 60 |
| 1.60 | 45 | 1.80 | 90 |
| 1.72 | 65 | 1.76 | 70 |
| 1.73 | 70 | 1.82 | 75 |

3.将表3.9中的数据分别用分箱法、基于熵的方法以及ChiMerge方法离散化属性“年 收入(万元)”为3个区间。其中属性“车型”为分类属性。

表3. 9习题3对应的数据集

|  |  |
| --- | --- |
| 年收人(万元) | 车型 |
| 10 | 普通 |
| 25 | 普通 |
| 30 | 高级 |
| 60 | 高级 |
| 40 | 中档 |
| 20 | 普通 |
| 50 | 中档 |
| 30 | 普通 |
| 100 | 高级 |
| 120 | 高级 |

4.下载kaggle房屋预测数据，绘制YearRemodAdd 、YrSold与SalePrice的散点图，并做简单分析。

1. 李航. (2019). 统计学习方法（第2版）. 清华大学出版社.
2. 周志华. (2016). 机器学习. 清华大学出版社.
3. 王元卓, 刘晓东, 张宇. (2018). 大数据可视化. 电子工业出版社.
4. 刘红岩.（2020）. 商务智能方法与应用. 清华大学出版社.
5. 王佳冬, 王文信. （2020）. 商业智能工具应用与数据可视化. 电子工业出版社.
6. 韩家炜, 米歇尔·J·普劳特, 马朝利, & 王立旺. (2016). 数据挖掘：概念与技术（第3版）. 机械工业出版社.
7. Rosling, H. (n.d.). D3 dynamic scatter plot example. Retrieved from [https://www.example.com](https://www.example.com/)
8. D3.js. (n.d.). Data-D riven Documents. Retrieved from <https://d3js.org/>
9. Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.
10. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (2nd ed.). Springer New York.