第5章 pandas数据预处理

教案

**课程名称：**Python数据分析与应用

**课程类别：**必修

**适用专业：**大数据技术类相关专业

**总学时：**64学时（其中理论36学时，实验28学时）

**总学分：**4.0学分

**本章学时**：7学时

# 材料清单

* 1. 《Python数据分析与应用》教材。
  2. 配套PPT。
  3. 数据。
  4. 代码。
  5. 引导性提问。
  6. 探究性问题。
  7. 拓展性问题。

# 教学目标与基本要求

### 教学目标

以菜品数据为例子，介绍数据分析的数据预处理过程，即数据清洗、数据合并、数据标准化和数据转换。数据清洗主要介绍对重复数据、缺失值和异常值的处理。数据合并介绍将多个数据源中的数据合并存放到一个数据存储的过程。数据标准化介绍如何将不同量纲的数据转化为可以相互比较的标准化数据。数据转换介绍如何从不同的应用角度对已有特征进行转换。

### 基本要求

1. 掌握数据清洗的基本方法。
2. 掌握数据合并的原理与方法。
3. 掌握数据标准化的概念与方法。
4. 掌握常用的数据变换方法。

# 问题

### 引导性提问

引导性提问需要教师根据教材内容和学生实际水平，提出问题，启发引导学生去解决问题，提问，从而达到理解、掌握知识，发展各种能力和提高思想觉悟的目的。

1. 数据质量不高的情况下如何提高数据质量？
2. 数据预处理包含哪些内容？
3. 数据预处理各个步骤是否有先后？

### 探究性问题

探究性问题需要教师深入钻研教材的基础上精心设计，提问的角度或者在引导性提问的基础上，从重点、难点问题切入，进行插入式提问。或者是对引导式提问中尚未涉及但在课文中又是重要的问题加以设问。

1. 重复值是否一定要做清洗？
2. 时间数据中存在哪些信息？
3. 数据变换的目的是什么？

### 拓展性问题

拓展性问题需要教师深刻理解教材的意义，学生的学习动态后，根据学生学习层次，提出切实可行的关乎实际的可操作问题。亦可以提供拓展资料供学生研习探讨，完成拓展性问题。

1. 除了哑变量处理外，还有那些方法可以处理离散型特征？
2. 能否将这些方法写成自定义函数？

# 主要知识点、重点与难点

### 主要知识点

1. 数据清洗的基本方法。
2. 数据合并的原理与方法。
3. 数据标准化的概念与方法。
4. 常用的数据变换方法。

### 重点

1. 数据合并。
2. 数据清洗。
3. 数据标准化。
4. 数据转换。

### 难点

1. 数据清洗。
2. 数据转换。

# 教学过程设计

### 理论教学过程

1. 堆叠合并数据。
2. 主键合并数据。
3. 重叠合并数据。
4. 检测与处理重复值。
5. 检测与处理缺失值。
6. 检测与处理异常值。
7. 离差标准化数据。
8. 标准差标准化数据。
9. 小数定标标准化数据。
10. 哑变量处理类别型数据。
11. 离散化连续型数据。

### 实验教学过程

1. 堆叠、主键、重叠合并数据。
2. 检测与处理重复值，缺失值，异常值。
3. 离差标准化、标准差标准化。
4. 小数定标标准化数据。
5. 哑变量处理类别型数据。
6. 离散化连续型数据。

# 教材与参考资料

### 教材

黄红梅，张良均．Python数据分析与应用[M]．北京：人民邮电出版社．2018．

### 参考资料

[1] 张良均．Python数据分析与挖掘实战[M]．北京：机械工业出版社．2015．

[2] 张良均．Python与数据挖掘[M]．北京：机械工业出版社．2016．