传智播客

《Python数据分析与应用：

从数据获取到可视化》

教学设计

**课程名称： Python数据分析**

**授课年级： 2018年级**

**授课学期： 2018学年第一学期**

**教师姓名： 某某老师**

2018年09月09日

|  |  |  |  |
| --- | --- | --- | --- |
| 课题名称 | 第4章 数据预处理 | 计划学时 | 6 课时 |
| 内容分析 | 前期采集到的数据，或多或少都存在一些瑕疵和不足，比如数据缺失、极端值、数据格式不统一等问题。因此，在数据分析之前需要对数据进行预处理，包括数据的清洗、合并、重塑与转换。Pandas中专门提供了用于数据预处理的很多函数与方法，用于替换异常数据、合并数据、重塑数据等。 |
| 教学目标及基本要求 | 1. 掌握数据清洗
2. 掌握数据合并
3. 掌握数据重塑
4. 掌握重命名轴索引
5. 熟悉离散化连续数据
6. 了解哑变量处理类别型数据
7. 完成案例—预处理部分地区信息
 |
| 教学重点 | 1. 掌握数据清洗
2. 掌握数据合并
3. 掌握数据重塑
4. 掌握重命名轴索引
5. 完成案例—预处理部分地区信息
 |
| 教学难点 | 1. 掌握数据清洗
2. 掌握数据合并
3. 掌握数据重塑
4. 掌握重命名轴索引
 |
| 教学方式 | 教学采用教师课堂讲授为主，使用教学PPT讲解 |
| 教学过程 | **第一课时****（空值和缺失的处理、重复值的处理、异常值的处理）****一、创设情境，导入为什么要使用数据清洗**1. 教师通过生活中的举例，引出数据清洗的知识。

例如问卷调查信息，在获取这些数据后并不能立即用于分析，而是需要将这些数据进行一些清洗操作。1. 数据清洗常用的操作

常的数据清洗操作包括空值和缺失值处理、重复值的处理、异常值的处理等。1. 明确学习目标
* 要求学生掌握空值和缺失值的处理。
* 要求学生掌握重复值的处理
* 要求学生掌握异常值的处理

**二、进行重点知识的讲解**1. 教师根据课件，讲述空值和缺失值的处理。

Pandas提供了一些用于检查或处理空值和缺失值的函数，其中，使用isnull()和notnull()函数可以判断数据集中是否存在空值和缺失值，对于缺失数据可以使用dropna()和fillna()方法对缺失值进行删除和填充。1. 教师根据课件，讲述重复值的处理。

Pandas提供了两个函数专门用来处理数据中的重复值，分别为duplicate()和drop\_duplicates()。其中，前者用于标记是否有重复值，后者用于删除重复值，它们的判断标准是一样的，即只要两条数据中所有条目的值完全相等，就判断为重复值。1. 教师根据课件，讲述异常值的处理。

异常值指样本中的个别值，其数值明显偏离它所属样本的其余观测值，这些数值是不合理的或错误的。要想确认一组数据中是否有异常值，常用的检测方法有3σ原则和箱形图，对于检测出的异常值可以使用replace方法进行替换。**三、归纳总结，布置作业/随堂练习**1. 回顾上课前的学习目标，并对本节课的内容进行总结。

 教师总结本节课需要掌握的知识点，包括空值和缺失值的处理、重复值的处理、异常值的处理。1. 使用博学谷系统下发课后作业。

**第二课时****（更改数据类型、轴向堆叠合并、主键合并数据）****一、回顾上节课的内容，继续讲解本课时的知识**1. 教师对学生们的疑问进行统一答疑。
2. 回顾总结上节课内容，继续介绍本课时的内容。

上节课介绍了数据清洗的内容，接下来，本节课继续介绍数据清洗内容以及数据合并操作。1. 明确学习目标
* 要求学生掌握更改数据类型的方法
* 要求学生掌握轴向堆叠合并
* 要求学生掌握主键合并数据

**二、进行重点知识的讲解**1. 为什么要更改数据类型？

在处理数据时，可能会遇到数据类型不一致的问题，例如通过爬虫采集到的数据都是整型的数据，在使用数据的时希望保留两位小数点，这时需要将这些数据的类型转换为浮点型。针对这种问题，既可以在创建Pandas对象明确指定数据类型，也可以使用astype和to\_numberic进行转换。1. 教师根据课件，讲解数据合并——轴向堆叠数据的操作。

在对数据进行整合时，经常会用到数据合并。接下来，介绍常用的几种数据合并操作。concat()函数可以沿着一条轴将多个对象进行堆叠，其使用方式类似数据库中的数据表合并。1. 教师根据课件，讲解数据合并——主键合并数据的操作。

主键合并类似于关系数据库库的连接方式，它是根据一个或多个键将不同的DataFrmae对象连接起来，大多数是将两个DataFrame对象中的列作为合并键，Pandas主键合并方式使用merge()函数实现。**三、归纳总结，布置作业**1. 回顾学习目标，对本节课的内容进行总结。

教师带领学生总结本节课需要掌握的内容，包括更改数据类型、轴向堆叠数据、主键合并数据。1. 布置随堂练习，检查学生掌握情况。

 根据博学谷和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。 1. 使用博学谷系统下发课后作业。

**第三课时****（根据行索引合并数据、合并重叠数据、重塑层次化索引）****一、回顾上节课内容，继续讲解数据合并内容以及数据重塑中的重塑层次化索引**1. 教师对学生们的疑问进行统一答疑。
2. 回顾总结上节课内容，继续介绍数据合并的知识以及数据重塑的知识。

在上节课中，介绍了更改数据类型的方法以及数据合并的两种方法。1. 明确学习目标
* 要求学生掌握根据行索引合并数据、
* 要求学生掌握合并重叠数据
* 要求学生掌握重塑层次化索引

**二、进行重点知识的讲解**1. 教师通过示例代码，讲述join()方法的使用。

join()方法能够通过索引或指定列来连接DataFrame。1. 教师通过示例代码，讲述combine\_first()方法的使用。

在处理数据的过程中，当一个DataFrame对象中出现了缺失数据，而这些缺失数据我们希望使用其它DataFrame对象中的数据填充，这时可以通过combine\_first()方法为缺失数据填充。1. 教师通过示例代码，讲述stack()和unstack()方法的使用。

Pandas中重塑层次化索引操作主要是stack()和unstack()，前者表示将数据的列“旋转”为行，后者表示将数据的行“旋转”为列。**三、归纳总结，布置作业**1. 回顾学习目标，对本节课的内容进行总结

教师带领学生总结join()、combine\_first()、stack()、unstack()的使用方法。1. 布置随堂练习，检查学生掌握情况。

 根据博学谷和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。1. 使用博学谷系统下发课后作业。

**第四课时****（轴向旋转、重命名轴索引、离散化连续数据、哑变量处理类别数据）****一、回顾上节课内容，继续讲解本节课的内容**1. 教师讲解上节课有难度的作业，并对学生的疑问进行统一答疑。
2. 回顾上节课的内容，继续介绍数据重塑的操作以及数据转换的部分操作。

在上节课中，介绍了数据合并和数据重塑的操作，接下来，继续介绍数据重塑的操作以及数据转换的操作。1. 明确学习目标
* 要求学生掌握轴向旋转操作
* 要求学生掌握重名轴索引的操作
* 要求学生熟悉离散化连续数据
* 要求学生了解哑变量处理类别数据

**二、进行重点知识的讲解**1. 教师通过教材中的示例，引出轴向旋转的操作。

在Pandas中pivot()方法会根据给定的行索引或列索引重新组织一个DataFrame对象。1. 教师通过设定需求，引出重命名轴索引的使用。

Pandas中提供一个rename()方法来重名列索引或行索引。1. 教师通过教材中的示例，引出离散化连续数据。

有时候我们会碰到这样的需求，例如，将有关年龄的数据进行离散化（分桶）或拆分为“面元”，直白来说，就是将年龄分成几个区间。 Pandas 的 cut()函数能够实现离散化操作。1. 教师根据课件，讲解哑变量处理类别型数据。

哑变量又称虚拟变量、名义变量，从名称上看就知道，它是人为虚设的变量，用来反映某个变量的不同类别。使用哑变量处理类别转换，事实上就是将分类变量转换为哑变量矩阵或指标矩阵，矩阵的值通常用“0”或“1”表示。**三、归纳总结，布置作业**（1）回顾学习目标，对本节课的内容进行总结。教师带领学生总结数据重塑和数据转换操作。 （2）使用博学谷系统下发随堂练习，检测学生对课堂内容的掌握情况。**第五课时****（上机练习）**上机练习主要针对本章中需要重点掌握的知识点，以及在程序中容易出错的内容进行练习，通过上机练习可以考察同学对知识点的掌握情况，对代码的熟练程度。**上机一：（练习教材中的示例代码）****形式：单独完成****第六课时****（上机练习）**上机练习主要针对本章中需要重点掌握的知识点，以及在程序中容易出错的内容进行练习，通过上机练习可以考察同学对知识点的掌握情况，对代码的熟练程度。**上机一：（练习教材中的示例代码）****形式：单独完成****上机二：（完成案例—预处理部分地区信息）****形式：单独完成** |
| 思考题和习题 | 见教材第4章配套的习题 |
| 教学后记 |  |