Advances in Education 教育进展, 2022, 12(8), 2861-2867
Published Online August 2022 in Hans. http://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2022.128433

Hans汉斯

Python在数据分析教学中的应用

——以Pandas进行某超市数据预处理为例

沈心雨[1]，赵胜利[1]，吕林黛[1]，钟妤玥[2]

[1]重庆理工大学，重庆
[2]重庆渝高中学，重庆

收稿日期：2022年7月11日；录用日期：2022年8月11日；发布日期：2022年8月17日

摘　要

数据分析是有组织有目的地收集数据、分析数据，使之成为信息的过程。从数据收集到数据可视化的过程中，数据预处理是极为重要的一环，需要花费大量的时间和精力。pandas是Python的核心数据分析支持库，提供了快速、灵活、明确的数据结构，旨在简单、直观地处理关系型、标记型数据，数据预处理中，pandas将承担最为重要的角色。本文以某超市数据为例，使用pandas进行数据分析处理，加强Python在数据分析教学中的应用，提高学生的数据分析能力。

关键词

Pandas模块，数据分析，数据预处理

The Application of Python in the Teaching of Data Analysis

—Taking Pandas for a Supermarket Data Preprocessing as an Example

Xinyu Shen[1], Shengli Zhao[1], Lindai Lyu[1], Yuyue Zhong[2]

[1]Chongqing University of Technology, Chongqing
[2]Chongqing Yugao Middle School, Chongqing

Received: Jul. 11th, 2022; accepted: Aug. 11th, 2022; published: Aug. 17th, 2022

Abstract

Data analysis is the process of collecting data purposefully, analyzing it, and turning it into information. From data collection to data visualization, data preprocessing is an extremely important

文章引用: 沈心雨, 赵胜利, 吕林黛, 钟妤玥. Python 在数据分析教学中的应用[J]. 教育进展, 2022, 12(8): 2861-2867.
DOI: 10.12677/ae.2022.128433

part that requires a lot of time and effort. Pandas is Python's core data analysis support library, providing a fast, flexible and clear data structure designed to simply and intuitively process relational and labeled data, and pandas will play the most important role in data preprocessing. This paper takes a supermarket data as an example, uses pandas for data analysis and processing, strengthens the application of Python in data analysis teaching, and improves students' data analysis ability.

Keywords

Pandas Module, Data Analysis, Data Preprocessing

1. 引言

数据分析能力是高等院校在培养数据人才时一项很重要的参考指标，大数据时代已经到来，大数据的典型特点是数据海量和数据处理流程复杂，这对学生的数据分析能力提出了更高的要求，目前，Python凭借简单易学、功能强大、高可扩展性等特点，已经成为数据分析教学上的首选工具。为了满足时代发展的需要和提高学生解决实际问题的能力，各大高校陆续开设了Python语言的基础教学课程，然而，传统的程序设计教学中主要教授理论知识，忽略了对学生应用程序语言解决实际问题能力的培养，不适合新背景下Python语言基础的教学[1]。为了让学生更加深刻的掌握Python语言基础的教学，激发学生用Python进行数据分析的学习兴趣，教师在进行教学时使用一系列实际例子是很有必要的。在教学领域，刘青云等进行了Python交叉融合案例教学实践，激发了学生使用Python进行数据分析的兴趣[2]；董付国作为多本Python教材的编写者，为Python教学提供了丰富的教学资源和应用实例[3]。总体来说，Python教学一般是理论与案例结合，因此，本文将以某超市数据为例，将数据预处理的相关知识融入到教学实例中，对超市数据进行相应的数据预处理，展示数据预处理的基本流程，让学生逐渐掌握如何用Python对数据进行初步处理。

2. 数据预处理概述

目前，在各领域都存在数据分析需求，对于数据分析而言，数据是显而易见的核心，但是并不是所有的数据都是有用的，大多情况下数据是不完整的、不一致的、极易受到噪声影响的，这会给后面的数据分析带来很大的麻烦，所以有必要进行数据预处理[4]。数据预处理是指对所收集的数据进行分类或分组前所做的审核、筛选、排序等必要的处理，接下来将从以下五个方面进行介绍。

2.1. 重复值的处理

当记录失误时，可能会导致存在重复数据，数据集中的重复值包括以下两种情况：1) 数据值完全相同的多条数据纪录；2) 数据主体相同但匹配到的唯一属性值不同。

2.2. 缺失值的处理

由于人为失误或机器故障的缘故，可能会导致某些数据丢失，在数据分析时应注意检查有没有缺失的数据如果有则将其删除或替换为特定的值，以减少对最终数据分析结果的影响[5]。

2.3. 异常值的处理

异常值是指严重超出正常范围的数值，这样的数据一般是数据采集错误或类似的原因造成的，在数据分析时，需要把这些数据删除或者替换为特定的值(例如认为设定的正常范围边界值)，减少对最终数据分析结果的影响[6]。异常值处理的关键是根据实际情况准确定义正常范围边界值，超出正常范围的数值认为是异常值。

2.4. 查看数据特征和统计信息

在分析数据时，有时需要查看数据的数量、平均值、标准差、最大值、最小值、四分位数等特征，pandas 对于这些操作都提供了良好的支持。

2.5. 数据的标准化处理

不同特征之间往往具有不同的量纲，由此所造成的数值之间的差异可能很大，为了消除特征之间的量纲和取值范围差异可能会造成的影响，需要对数据进行标准化处理。

3. 数据预处理实例——以超市数据为例

本文以某超市销售数据为例来进行数据预处理，该数据包含工号、姓名、日期、时段、交易额、柜台这六个特征，日期范围从 2019 年 3 月 1 日至 2019 年 3 月 31 日。使用 pandas 对超市数据进行数据清洗修复，对修复后的数据进行统计分析，利用处理后的数据做进一步的分析。

3.1. 重复值的处理

当存在重复数据的时候，一般采取的处理方法是直接丢弃重复数据。Python 中的 duplicated()可以用来检测哪些行是重复的，其语法格式为：duplicated(subset=None,keep='frist')。其中第一个参数 subset 对应值是列名，将列对应值相同的行进行去重，默认值 None，即考虑所有列；第二个参数 keep='frist'表示除了第一次出现外，其余相同的值被标记为重复。本文首先查找重复行，一共有一条数据重复，然后查找一个人同时负责多个柜台的排班，得到表 1，可以得到，该超市排班分配存在不合理的情况。代码为：

```
print(df[df.duplicated()])
dff3=df[['工号','姓名','日期','时段']]
dff3=dff3[dff3.duplicated()]
for row in dff3.values:
    print(df[(df.工号==row[0])&(df.日期==row[2])&(df.时段==row[3])])
```

Table 1. One person is responsible for the scheduling of multiple counters at the same time
表 1. 一个人同时负责多个柜台的排班

索引	工号	姓名	日期	时段	交易额	柜台
49	1002	李四	2019-03-07	14:00~21:00	1199.0	化妆品
55	1002	李四	2019-03-07	14:00~21:00	831.0	蔬菜水果
103	1006	钱八	2019-03-13	14:00~21:00	1609.0	蔬菜水果
104	1006	钱八	2019-03-13	14:00~21:00	1609.0	蔬菜水果
171	1006	钱八	2019-03-22	9:00~14:00	1555.0	蔬菜水果

Continued

175	1006	钱八	2019-03-22	9:00~14:00	1503.0	食品
201	1004	赵六	2019-03-26	9:00~14:00	1599.0	化妆品
210	1004	赵六	2019-03-26	9:00~14:00	1257.0	化妆品

3.2. 缺失值的处理

Python 中使用 dropna()丢弃带有缺失值的数据行，或者使用 fillna()方法对缺失值进行批量替换。本文原始数据有 249 行，有 3 行数据缺失，用每人交易额的均值来替换缺失值，得到表 2。代码为：

```
print(len(df))
    print(len(df.dropna())
dff=deepcopy(df)
    for i in dff[dff.交易额.isnull].index;
dff.loc[i,'交易额']=round(dff.loc[dff.姓名==dff.loc[i,'姓名'],'交易额'].mean()
    print(dff.iloc([110,124,168,;])
```

Table 2. Replace missing values with the average transaction value per person
表 2. 使用每人交易额均值替换缺失值

索引	工号	姓名	日期	时段	交易额	柜台
110	1005	周七	2019-03-14	14:00~21:00	1195.0	化妆品
124	1006	钱八	2019-03-16	14:00~21:00	1323.0	食品
168	1005	周七	2019-03-21	14:00~21:00	1195.0	食品

上述代码中，主要使用 dropna(axis=0, how='any', thresh=None, subset=None, inplace=False)函数，其中第一个参数 axis=0 表示删除带有缺失值的行，axis=1 表示删除带有缺失值的列；第二个参数 how='any'表示删除只要含有缺失值的行(列)，how='all'表删除全是缺失值的行(列)；第三个参数 thresh 表示非空元素最低数量，若该行(列)中，非空元素数量小于这个值，就删除该行(列)；第四个参数 subset 表示行(列)的索引；第五个参数 inplace 表示是否在原 DataFrame 上进行修改。

3.3. 异常值的处理

异常值的处理有以下几种方法：

1) 删除：先将异常值替换为 Na，然后使用 dropna()删除。

2) 视为缺失值：先将异常值替换为 Na，然后用缺失值的方法处理(填充、插值等)。

3) 平均值修正：如果数据量较小，也可以用前后两个观测值的平均值修正该异常值。

本文首先筛选出交易额小于 200 的数据，如表 3，一共有三条数据。然后筛选出交易额大于 3000 的数据，如表 4，一共有两条数据，最后使小于 200 的交易额等于 200，大于 3000 的交易额等于 3000。代码为：

```
df.loc[df.交易额<200,"交易额"]=200
df.loc[df.交易额>3000,"交易额"]=3000
```

Table 3. Data on transactions with transactions of less than 200
表 3. 交易额小于 200 的数据

索引	工号	姓名	日期	时段	交易额	柜台
76	1005	周七	2019-03-10	9:00~14:00	53.0	日用品
97	1002	李四	2019-03-13	14:00~21:00	98.0	日用品
194	1001	张三	2019-03-25	14:00~21:00	114.0	化妆品

Table 4. Data on transactions greater than 3000
表 4. 交易额大于 3000 的数据

索引	工号	姓名	日期	时段	交易额	柜台
105	1001	张三	2019-03-14	9:00~14:00	12100.0	日用品
223	1003	王五	2019-03-28	9:00~14:00	9031.0	食品

Pandas 有两个主要的数据结构：Series 和 DataFrame，其中 Series 是带标签的一维数组，DataFrame 是一个二维结构，DataFrame 中提供了 loc、iloc、at、iat 等访问器来访问指定的数据。其中，iloc 和 iat 使用整数来指定行、列的下标，而 loc 和 at 使用标签来指定要访问的行和列，使用这些操作能够快速的查找到需要的行和列。

3.4. 查看数据特征和统计信息

对交易额进行汇总类统计，得到表 5，可以得出，该超市在这一时间段交易额的均值为 1330.3130，找出交易额最小的 3 条记录，如表 6 所示，可以得到，交易额的最小值为 53.0，代码为：

```
df['交易额'].describe()
df.nsmallest(3,'交易额')
```

Table 5. Summary statistics on transaction value
表 5. 交易额的汇总统计

数据特征	交易额
Count	246.0000
Mean	1330.3130
Std	904.3007
Min	53.0000
25%	1031.2500
50%	1259.0000
75%	1523.0000

Table 6. The 3 records with the minimum transaction value
表 6. 交易额最小的 3 条记录

索引	工号	姓名	日期	时段	交易额	柜台
76	1005	周七	2019-03-10	9:00~14:00	53.0	日用品
97	1002	李四	2019-03-13	14:00~21:00	98.0	日用品
194	1001	张三	2019-03-25	14:00~21:00	114.0	化妆品

3.5. 数据的标准化处理

数据标准化操作是将数据按比例缩放，使之落入一个小的特定区间，用 pandas 模块实现数据标准化主要有以下方法：

1) 离差标准化：(x − min)/(max − min)，这是对原始数据的一种线性变换，结果是将原始数据的数值映射到[0, 1]区间上。

2) 标准差标准化：$(x-\mu)/\delta$。

3) 小数定标标准化：x/10**k,k=np.ceil(log10(max(|x|))，这是将数据映射到[−1, 1]区间上，移动的小数位数取决于数据绝对值的最大值。代码为：

```
def minmaxscale(data):
    data=(data-data.min())/(data.max()-data.min())
    return data
data1=minmaxscale(df['交易额'])
data1.head()
def standardscaler(data):
    data=(data-data.mean())/data.std()
return data
data2=standardscaler(df['交易额'])
data2.head()
```

标准化后前 5 行的结果如表 7，由表可知，这三种标准化处理的结果有所区别，其中离差标准化是将数据进行线性变换，但是过大或者过小的异常值都会对结果产生影响，离差标准化是消除量纲影响最简单的方法；标准差标准化后的数据符合标准正态分布，适用于本身服从正态分布的数据；小数定标标准化后的数据分布规律不变，适用于数据分布比较分散，尤其是数据分布多个数量级的情况。

Table 7. Standardized data
表 7. 标准化后的数据

索引	离差标准化	标准差标准化	小数定标标准化
0	0.1337	0.3690	0.0166
1	0.0748	−0.4161	0.0095
2	0.1123	0.0848	0.0141
3	0.1052	−0.0114	0.0132
4	0.0781	−0.3719	0.0100

4. 结语

本文利用 pandas 对某超市的销售数据进行了异常值处理、重复值处理等操作，pandas 是一款基于numpy、专门为了解决数据分析任务的工具，其提供了大量标准数据模型，具有高效操作大量数据集所需要的功能[7]。目前，在各个领域都存在数据分析需求，pandas 提供了大量函数和对象方法来支持这些操作，可以说，pandas 是使 Python 能够成为高效且强大的数据分析行业首选语言的重要因素之一。数据预处理是数据分析的第一步，在数据分析中占有重要地位。通过在 Python 语言基础教学过程中贯穿数据预处理案例，结合各个知识点的综合应用分析讲解演示，不仅能够加深学生对理论知识的理解，还能提高学生运用所学知识解决复杂数据问题的能力[8]。同时，以案例为依托的研讨式教学也对教师的专业能力提出了更高的要求，研讨式教学极大地提高了学生的自主性，教师与学生一起努力，共同完成整个教学过程。由于数据分析广阔的应用前景，可能会有来自其他专业的学生进行 Python 基础课程学习，让不同学习背景的学生掌握数据分析的基础入门知识，是我们的挑战，我们将为之不断的努力。

参考文献

[1] 高望. 基于数据预处理的 Python 课程教学案例设计研究[J]. 信息与电脑(理论版), 2022, 34(4): 254-256.

[2] 刘青云, 焦铬, 陈坚祯. MIMPS 教学法在网络编程实践课程中的应用研究[J]. 福建电脑, 2018, 34(7): 79-80+128. https://doi.org/10.16707/j.cnki.fjpc.2018.07.042

[3] 杨彩云, 詹国华. 引导性问题案例在 Python 数据分析基础课程的教学[J]. 计算机教育, 2021(1): 154-157+162. https://doi.org/10.16512/j.cnki.jsjjy.2021.01.037

[4] 许辉. 数据挖掘中的数据预处理[J]. 电脑知识与技术, 2022, 18(4): 27-28+31. https://doi.org/10.14004/j.cnki.ckt.2022.0262

[5] 高鸿斌, 申肖阳. Python 数据分析技术综述[J]. 邯郸职业技术学院学报, 2018, 31(4): 49-51.

[6] 张治斌, 刘威. 浅析数据挖掘中的数据预处理技术[J]. 数字技术与应用, 2017(10): 216-217. https://doi.org/10.19695/j.cnki.cn12-1369.2017.10.114

[7] 徐文昭. 运用 Python 及 Pandas 库分组统计“最值”记录方法探讨[J]. 内蒙古科技与经济, 2021(21): 73-74.

[8] 刘新鹏, 高斌. 利用 Python 和 Pandas 进行学生成绩处理[J]. 信息与电脑(理论版), 2020, 32(7): 41-43.