《Python数据处理编程》课内实验指导书

内容 简 介

Python数据处理编程是计算机和大数据专业学生的一门专业必修课程，着重讲述基于Python的数据处理的基本概念和技术方法。对于学生从事大数据分析、数据挖掘、机器学习、深度学习等系统的研发、使用和维护有重要意义。本课程概念多、内容涉及面广、系统性强。通过本课程的学习，学生应能从数学模型、编程实现的角度去了解、分析和研究基于Python的数据处理应用系统，建立起对基于Python的数据处理应用系统的全面认识，树立全面地、发展地看问题的观点，从而加深对各种基于Python的数据处理应用系统的了解。

本课程的学习应注重理论与实践相结合，因此实验教学是教学环节中必不可少的重要内容。通过实验教学的学习，使学生熟练掌握有关基于Python的数据处理编程的基本概念、基本原理和基本思想，掌握对基于Python的数据处理系统进行设计、分析和计算的方法。

实验部分包括实验目的、实验内容和实验所需环境等，介绍了实验所需的一些基础知识和技巧。在实验中给出的实验题，跟课堂教学的内容都有密切的关系，所以需要将课堂上讲授的例子程序融会贯通，掌握实验所需的一些基本方法和工具，并在吃透例子程序的基础上，积极独立思考设计和编写满足实验要求的程序。

中南大学计算机学院鲁鸣鸣制定

上机实验要求及规范

Python数据处理编程课程具有比较强的实践性。上机实验是一个重要的教学环节。一般情况下学生能够重视实验环节，对于编写程序上机练习具有一定的积极性。但是容易忽略实验的总结，忽略实验报告的撰写。对于一名大学生必须严格训练分析总结能力、书面表达能力。需要逐步培养书写科学实验报告以及科技论文的能力。拿到一个题目，一般不要急于编程。正确的方法是：首先理解问题，明确给定的条件和要求解决的问题，然后按照自顶向下，逐步求精，分而治之的策略，按照面向对象的程序设计思路，逐一地解决子问题。

一、实验报告的基本要求：

一般性、较小规模的上机实验题，必须遵循下列要求。养成良好的习惯。

姓名 班级 学号 日期 题目

i. 问题描述

ii. 设计简要描述

iii. 程序清单（带有必要的注释）

iv. 结果分析（原始图示，测试数据与运行记录，分析正确性；）

v. 调试报告：

实验者必须重视最后这两个环节，否则等同于没有完成实验任务。这里可以体现个人特色、或创造性思维。具体内容包括：测试数据与运行记录；调试中遇到的主要问题，自己是如何解决的；经验和体会等。

二、实验报告的提高要求：

阶段性、较大规模的上机实验题，应该遵循下列要求。养成科学的习惯。

（1）问题描述

（2）需求和规格说明

（3）描述问题，简述题目要解决的问题是什么。规定软件做什么。原题条件不足时补全。

（4）概要设计：功能模块的划分

（5）详细设计：每部分模块的设计，含数据结构的设计，算法的描述（流程图或PDL）

a.设计思想：主要算法基本思想。

b.设计表示：每个函数的头和规格说明；列出每个函数所调用和被调用的函数，也可以通过调用关系图表达。

（6）实现注释：各项功能的实现程度、在完成基本要求的基础上还有什么功能。

（7）用户手册：即使用说明书。

（8）调试报告：调试过程中遇到的主要问题是如何解决的；设计的回顾、讨论和分析；时间复杂度、空间复杂度分析；改进设想；经验和体会等。

## 实验三 员工离职调查报告的数据清洗与分析

一、实验目的

通过本项目将Pandas库中的一些函数知识应用在实践案例中，需要熟悉和掌握的函数包括：可用来清理字符串类型的矢量化的字符串方法；数据格式转换函数apply、map()、applymap()等；删除缺失值或不需要值的函数fillna()、dropna()、drop()；重塑数据函数melt()；合并数据函数concat() 和merge()。

二、实验开发环境和工具

可以在Windows或Linux操作系统上搭建开发环境，使用Anaconda科学计算包安装Python 3，里面会包含大多数科研常用的库。.ipynb文件推荐使用Anaconda内置的web版Jupyter Notebook。

三、实验内容

在本实验中，我们将对澳大利亚昆士兰教育培训就业部（DETE）和职业技术教育学院（TAFE）的员工进行离职调查。你可以在这里找到TAFE离职调查和这里的DETE调查。

TAFE的离职报告数据：

<https://data.gov.au/dataset/ds-qld-89970a3b-182b-41ea-aea2-6f9f17b5907e/details?q=exit%20survey>

DETE的离职报告数据：

<https://data.gov.au/dataset/ds-qld-fe96ff30-d157-4a81-851d-215f2a0fe26d/details?q=exit%20survey>

我们将这些数据集进行了一些细微的修改，保证它们更易于使用。

修改内容包括：将编码格式更改为UTF-8（原来的编码格式为cp1252）

具体任务如下：

1. 了解项目和数据集：

本任务将主要按照如下内容展开分析：

* 那些只在研究院工作了很短一段时间的员工是否是由于存在某些不满而辞职?工作的更久一点的员工是否因存在不满而辞职？
* 年轻员工辞职是因为存在某种不满吗?老员工辞职是因为存在某种不满吗?

我们应当能够对这两个调查数据的结果进行综合考评，然后再回答上面的问题。我们的目标：完成大部分数据清理工作，并保证你可以开始分析第一个问题。

数据集中没有提供数据字典。因此对于这个项目，将使用我们的个人常识来定义列。下面是我们将在dete\_survey.csv中使用的一组列的解释:

* ID：用于识别调查参与者的id
* SeparationType：雇员雇佣关系终止的原因
* Cease Date：雇员结束工作的年或月
* DETE Start Date：雇员开始受雇于DETE的哪一年

下面是我们将在tafe\_survey.csv中使用的一组列的基本解释：

* Record ID: 用于标识调查参与者的id
* Reason for ceasing employment: 雇员雇佣关系终止的原因
* LengthofServiceOverall：学院总工作年限（单位：年）：员工的工作年限（单位：年）

具体编程任务如下：

1. 首先在markdown单元格中编写一段介绍项目和数据集的注释。
2. 导入pandas和NumPy库。
3. 通过pandas读取dete\_survey.csv文件，并赋值给变量dete\_survey。
4. 通过pandas读取tafe\_survey.csv文件，并赋值给变量tafe\_survey。
5. 使用DataFrame.info() 和DataFrame.head()方法查看这两个DataFrame的基本信息（打印前6行）。使用其他数据探索方法，比如：Series.value\_counts()和DataFrame.isnull()方法来查看数据并确定一些后续步骤。
   1. 另起markdown单元格，简要描述你的观察结果。
6. 识别缺失的数据和丢弃不必要的列：

通常，确定下一步如何对数据进行清洗和重塑都是比较关键且最困难的部分。但也不用太过担担心，后续我们将会提供一些步骤，并留有一些您自己的空间进行拓展。

首先可以进行如下观察：

* dete\_survey中包含了值为"Not Stated"，这些值表示缺失值，但它们并不是NaN。
* dete\_survey和tafe\_survey中都包含许多我们不需要分析的列。
* 这两个dataframe都包含许多类似的列，但列名不同。
* 有多少个列/答案表示员工因不满而辞职。

首先，我们将处理两个问题。使用pd.read\_csv()确定应该表示为NaN的值。我们将首先对缺失值进行处理。然后，我们再删除那些在分析中不需要的列。

具体编程任务如下：

1. 通过pandas读取dete\_survey.csv文件，但是本次读取时，将值为"Not Stated"的地方都修改为"NaN"

* 可通过设置pd.read\_csv()方法中的参数na\_values为Not Stated实现
* 将数据读取结果赋值给变量dete\_survey

1. 通过pandas读取tafe\_survey.csv文件，但是本次读取时，将值为"Not Stated"的地方都修改为"NaN"

* 通过设置pd.read\_csv()方法中的参数na\_values为Not Stated实现
* 将数据读取结果赋值给变量tafe\_survey

1. 然后，删除dataframe中一些不需要的列，保证dataframe更易于使用

* 使用dataframe .drop()方法从dete\_survey中删除以下列:

dete\_survey.columns[28:49]，请将删除维度参数axis设置为1

* + 将删除后的结果赋值给变量dete\_survey\_updated
* 使用dataframe .drop()方法从tete\_survey中删除以下列:

tete\_survey.columns[17:66]，请将删除维度参数axis设置为1

* + 将删除后的结果赋值给变量tafe\_survey\_updated

1. 新建markdown单元格，阐述你对数据集所做的改变，以及原因
2. 清洗列名：

接下来，让我们将注意力转向列名。这两个dataframe包含许多类似的列，但是列名不同。下面是一些想用于最终分析的列：

|  |  |  |
| --- | --- | --- |
| dete\_survey | tafe\_survey | Definition |
| ID | Record ID | 用于识别调查参与者的id |
| SeparationType | Reason for ceasing employment | 参与者终止雇佣关系的原因 |
| Cease Date | CESSATION YEAR | 参与者就业结束的年份或月份 |
| DETE Start Date |  | 参与者开始受雇于DETE的年份 |
|  | LengthofServiceOverall. Overall Length of Service at Institute (in years) | 参与者的工作年限（年） |
| Age | CurrentAge. Current Age | 参与者的年龄 |
| Gender | Gender. What is your Gender? | 参与者的性别 |

因为最终想要合并这两个DataFrame，所以必须对列名进行标准化。回想一下，可以使用DataFrame.columns以及向量化的字符串方法，以一次性更新所有列。下面是之前的一个实例：

|  |
| --- |
| happiness2017.columns =  happiness2017.columns.str.replace('.', ' ').str.replace('\s+', ' ').str.strip().str.upper() |

具体编程任务如下：

1. 对dete\_survey\_updated中的剩余列重命名

* 按照以下要求更新列名：
  + 所有大写字母均改为小写字母
  + 删除字符串结尾的所有空格。
  + 用下划线('\_')替换单词之间的空格。
* 例如：Cease Date应当更新为cease\_date
* 可以用DataFrame.columns属性，以一个数组的形式输出现有的列名。

1. 使用datafame.rename()方法更新tafe\_survey\_updated中的以下列。

|  |  |
| --- | --- |
| Record ID | id |
| CESSATION YEAR | cease\_date |
| Reason for ceasing employment | separationtype |
| Gender. What is your Gender? | gender |
| CurrentAge. Current Age | age |
| Employment Type. Employment Type | employment\_status |
| Classification. Classification | position |
| LengthofServiceOverall. Overall Length of Service at Institute (in years) | institute\_service |
| LengthofServiceCurrent. Length of Service at current workplace (in years) | role\_service |

1. 使用datafame.head()方法查看已更新的dete\_survey\_updated和tafe\_survey\_updated的更新情况，确保更新列名准确。
2. 新建markdown单元格，阐述你对数据集所做的改变，以及原因。
3. 滤除数据：

在上一关中，对需要使用的列进行了重命名。后面则是删除不需要的数据。回想一下，我们的最终目标是回答以下问题：

* 那些只在研究院工作了很短一段时间的员工是否是由于存在某些不满而辞职?工作的更久一点的员工是否因存在不满而辞职？

如果我们查看这两个dataframe的separationtype列中的惟一值情况，可以看到这两个DataFrame都各自包含一组不同的离职类型。对于本项目，我们将只分析辞职的员工，所以他们的离职类型的值中会包含字符串'Resignation'。

如果想挑战难度，请尝试使用所有的离职类型来完成项目——这将发现在数据清洗过程中会有更多的问题需要解决。

注意，dete\_survey\_updated中包含多种包含字符串'Resignation'的离职原因类型，这些值中都包含字符串'Resignment'：

* Resignation-Other reasons
* Resignation-Other employer
* Resignation-Move overseas/interstate

我们必须考虑到每个变量，这样我们才不会无意中丢失数据！

注：无论进行什么操作，都请记住先对DataFrame进行副本保存，然后进行修改。

具体编程任务如下：

1. 使用Series.value\_counts()方法分别查看dete\_survey\_updated和tafe\_survey\_updated的separationtype列中每一种值的数量情况。
2. 在这两个dataframes中，筛选出辞职原因类型值中包含Resignation字符串的调查对象对应的数据。

* dete\_survey\_updated中有三种离职原因都包含Resignation字符串，这三种都需要考虑进来，作为离职原因为Resignation的数据。
  + 将dete\_survey\_updated中separationtype进行切片，拿到第一个单词并赋值回separationtype，从而使得三种都包含Resignation字符串统一为一种。
* 筛选dete\_survey\_updated 中separationtype列的值为Resignation的项，所查找出的结果通过DataFrame.copy()方法赋值给dete\_resignations
* 筛选tafe\_survey\_updated中 separationtype列的值为Resignation的项，所查找出的结果通过DataFrame.copy()方法赋值给tafe\_resignations

1. 新建markdown单元格，阐述你对数据集所做的改变，以及原因。
2. 验证数据：

在开始清洗和操作其余的数据之前，我们还需要验证数据是否存在一些问题。当你处理真正的数据时，请小心，不要猜测你正在分析的数据肯定是未被污染的！

或许无法发现所有错误，但是通过确保数据在我们所知的范围内看起来是合理的，以免在进行一个数据分析项目时，这个项目最终因为糟糕的数据而变得毫无用处。

在本关，我们将重点检查“cease\_date”和“dete\_start\_date”列中的年份是否有意义。也鼓励同学们自行检查其他数据。

* 因为“cease\_date”是员工工作的最后一年，而“dete\_start\_date”是员工工作的第一年，因此在当前日期之后日期是没有意义的。
* 考虑到这个领域的大多数人都是在20多岁开始工作的，所以dete\_start\_date的值不太可能早于1940年。

如果数据中存在比当前年份更大或者比1940年的年份更小的数据，就应当人停止对数据的分析，因为这可能意味着数据有一些不合乎常理的地方。如果有少量的值过高或过低，那就直接删除这些数据。

具体编程任务如下：

1. 检查dataframe中的年份是否存在逻辑问题

* 首先，清洗dete\_resignations中的cease\_date列的数据
  + 使用Series.value\_counts()方法查看cease\_date列中的每一种值的数量情况
  + 使用向量化的字符串方法提取年份。
  + 使用Series.astype()方法将原类型转换为浮点型。
* 使用Series.value\_counts()查看dete\_resignations 中cease\_date 列和dete\_start\_date列的值；以及tafe\_resignations中cease\_date列的值。
  + 因为Series.value\_counts()返回一个series对象，所以可以使用Series.sort\_index()方法查看数量最多或是数量最少的值。通过修改参数ascending为True或是False即可设置升序或者为降序输出。
* 可以使用箱形图绘制任何数字列的值，以查看任何看起来存在问题的值。

1. 编写一个新的markdown单元格，阐述你在结果中的发现。
2. 创建新列:

上一关，我们可以看出：

1. 年份数值不存在问题。
2. 两个DataFrame中的年份跨度不完全相同。因此需要放弃一些不需要分析的年份。

现在已经证实了dete\_resignations中年份是没有问题的，将使用它们创建一个新的列。回想一下，我们的最终目标是回答以下问题：

* 那些只在研究院工作了很短一段时间的员工是否是由于存在某些不满而辞职?工作的更久一点的员工是否因存在不满而辞职？

在人力资源部门有一个专门的称呼，员工在工作地点停留的时间被称为他们的工作年限。可以发现，tafe\_resignations中已经包含一个service列。将其重命名为institute\_ service。为了同时分析这两份调查报告数据，必须在dete\_resignations中创建一个相应的institute service列。

我们是否有数据可以用来计算员工在工作地点停留的时间?花一分钟再次回顾dete\_resignations的列，是存在两个关键的列可以计算的。

具体编程任务如下：

1. 在dete\_resignations中创建一个institute\_service列

* 用cease\_date列的值减去dete\_start\_date列的值，并将结果赋值给列institute\_service列。

1. 另起一个markdown单元格，阐述你对数据的改变，以及原因 。
2. 识别不满的雇员:

上一关中，我们新建了一个institute\_service列，用于保存被调查者的工作入职时长（单位：年份）。接下来我们将会找出那些因为不满而辞职的员工。

我们将对dataframe中员工进行分类，

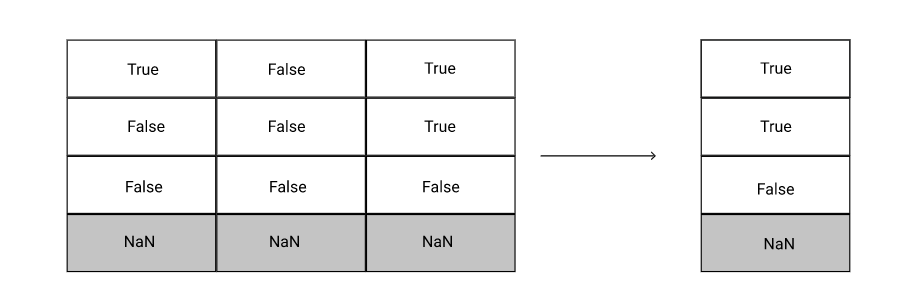
下面是我们将用于从dataframe中将员工分类为“不满意”的列。如果你不认可这种分类策略，可以自行修改！但是需要解释原因。

1. tafe\_survey\_updated：
   * + Contributing Factors. Dissatisfaction
     + Contributing Factors. Job Dissatisfaction
2. Dete\_survey\_updated：
   * + job\_dissatisfaction
     + dissatisfaction\_with\_the\_department
     + physical\_work\_environment
     + lack\_of\_recognition
     + lack\_of\_job\_security
     + work\_location
     + employment\_conditions
     + work\_life\_balance
     + workload

如果员工指出以上任何导致他们辞职的因素，则在新列中标记为dissatisfied。

创建新列需要执行以下操作：

1. 将tafe\_resignations中'Contributing Factors. Dissatisfaction'和'Contributing Factors. Job Dissatisfaction'这两列的值进行转换，转换为True、False或者NaN。
2. 如果以上列的名单中，任意一个列的值为True，则个新增的dissatisfied列填上True。将通过DataFrame.any()方法完成此工作，具体如下：
   * + 如果所选中的列的元素中有任意一个为True，返回True。
     + 如果所选中的列的元素中一个True都不存在，返回False。
     + 如果为NaN，则直接返回NaN。



DataFrmae.any()方法的用法为：

|  |
| --- |
| df.any(axis=1, skipna=False) |

详细链接可参考：

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.any.html>

经过以上操作之后，dissatisfied列必须只包含以下几个值：

* + - True：表示员工辞职是因为对工作不满
    - False：表示员工辞职不是因为对工作不满，而是其他原因
    - NaN：表示该值缺失

具体编程任务如下：：

1. 使用Series.value\_counts()方法查看tafe\_resignations中以下两个列的内容。
   1. Contributing Factors. Dissatisfaction
   2. Contributing Factors. Job Dissatisfaction
2. 将这两列的值更新为仅含True、False、NaN三种情况
   1. 声明一个叫做update\_vals的方法用于进行一下变换：
      1. 如果值为NaN，则返回np.nan（可以通过方法pd.isnull(val)判断一个值是否为NaN）
      2. 如果值为`-`, 返回False
      3. 如果是其他的值，则返回True
   2. 使用DataFrame.applymap()方法，将所声明的函数应用在上述两列中
      1. 传入update\_vals到方法applymap()中不需要括号
3. 使用df.any()方法在上述tafe\_resignations和dete\_resignations中分别创建一个名为dissatisfied的列
4. 使用df.copy()方法对结果进行备份，并将结果分别赋值为dete\_resignations\_up和tafe\_resignations\_up
5. 另起一个markdown单元格，阐述你对数据的改变，以及原因 。
6. 合并数据：

对已经完成的任务进行如下总结：

1. 给列进行重命名
2. 删除我们分析中不需要的列
3. 检查了数据的正确性
4. 创建了一个新列institute\_service
5. 对Contributing Factors列的值进行了数据清洗
6. 创建一个新列，表明员工是否因某些方面的不满而辞职

现在，可以将数据集进行合并了。最终，需要根据institute service列将数据合并，因此在组合数据时，请考虑如何将数据转换为易于聚合的形式。

具体编程要求如下：

1. 首先，在这两个dataframe中添加一列，更加易于区分这两列的数据。
   1. 添加一个institute列到dete\_resignations\_up中。该列的值均设为DETE
   2. 添加一个institute列到tafe\_resignations\_up中。该列的值均设为DETE
2. 合并数据集，将结果赋值给combined
3. dataframe中还有一些列不需要完成分析。使用datafame.dropna()方法删除任何非空值小于500的列
   1. 可以通过设置thresh参数，删除非空值小于一定数量的列。
   2. 将结果赋值给combined\_updated
4. 另起一个markdown单元格，阐述你对数据的改变，以及原因。：
5. 清洗Service列：

完成了之前的任务之后，现在已经很接近可以开始分析了。不过，首先，我们必须清理institute\_service列的值。因为它当前包含两个不同形式的值，所以清理过程相对复杂。



为了分析数据，将把这些数字转换为类别。我们将在这篇文章的基础上进行分析，这篇文章的论点是：根据职业阶段而不是年龄来了解员工的需求更有效。

我们将使用下面稍微修改过的定义

* + - New：在一家公司少于3年的员工
    - Experienced：在一家公司工作3-6年的员工
    - Established：在一家公司工作7-10年的员工
    - Veteran：在一家公司工作大于等于11年的员工

让我们使用上面的定义对institute\_service列中的值进行分类。

具体编程要求如下：

1. 首先，我们将从institute\_service列中的每个值中提取工作年限
   1. 使用Series.astype()方法将类型变为str类型
   2. 使用向量化的字符串方法从每个pattern中提取工作年限。可以在这里找到向量化字符串方法的完整列表。

<https://pandas.pydata.org/pandas-docs/stable/user_guide/text.html>

* 1. 仔细检查是否存在漏掉数字的情况。
  2. 使用Series.astype()方法将类型转化为float类型。

1. 接下来，将把每个值都分别映射到上面的某个职业阶段中。
   1. 创建一个函数，将每一个年份值都分别映射到上面的一个职业阶段。
      1. 您必须单独处理缺失值。可以通过pd.isnull(val)方法检测变量val的值是否为NaN
   2. 使用Series.apply()方法将声明的函数应用于institute\_service列上。并将结果赋值给一个新列，名为service\_cat
2. 另起一个markdown单元格，阐述你对数据的改变，以及原因。
3. 进行初步分析：

在之前的任务中，创建了一个service\_cat列，这是根据以下工作年限对员工进行分类的：

* + - New：在一家公司少于3年的员工
    - Experienced：在一家公司工作3-6年的员工
    - Established：在一家公司工作7-10年的员工
    - Veteran：在一家公司工作大于等于11年的员工

不仅在dissatisfied列汇总仍然存在缺失值，需要进行填充；还存在其他的缺失值需要处理。这意味着还需要进行一定的数据处理，保证后续工作正常开始。

Dissatisfied列中全为True或者False这种的布尔值。df.pivot\_table()会将布尔值当做整型值，将True当做1，False当做0处理。这意味着我们可以把dissatisfied那一列进行累加，从而计算出每一组的人数，每一组的人数百分比等等。

具体编程要求如下：

1. 使用Series.value\_counts()方法查看dissatisfied列中True和False各自的数量，设置参数dropna为False，保证也能看到缺失值的数量。
2. 使用DataFrame.fillna()方法替换dissatisfied列汇总的缺失值，替换策略为：是用出现次数最多的值进行填充（True或False）。
3. 使用DataFrame.pivot\_table()方法计算dissatisfied员工在每一个service\_cat组中的百分比。
   1. 因为True被看作是1，计算平均值也会计算不满意员工的百分比。默认的聚合函数是平均值，因此可以不使用aggfunc参数。
4. 使用DataFrame.plot()方法对结果进行制图。将kind参数绘制为bar，从而绘制一个柱状图
   1. 请在你的jupyter中顶部添加%matplotlib inline，保证你的图能够正常显示
5. 另起一个markdown单元格，阐述你对数据的改变，以及原因。