

**教学上机实验报告**

**课程名称：** 数据分析与可视化

**任课教师姓名：** 唐朝生

**学生学号：**  312220020923

**学生姓名：**  谭炜烁

**学生专业班级：** 大数据2209

**2023 ～ 2024 学年 第 二 学期**

|  |
| --- |
| **河南理工大学**  **教学上机实验报告评价分值标准** |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 序号 | 评价指标 | 分值 | 评价等级及参考分值 | | | | | 评价分 | | 优 | 良 | 中 | 合格 | 差 | | 1 | 实验报告内容完整充实 | 10 | 10 | 8 | 7 | 6 | 3 |  | | 2 | 实验内容书写规范、字迹工整认真 | 10 | 10 | 8 | 7 | 6 | 3 |  | | 3 | 实验过程叙述详细、概念正确，语言表达准确，结构严谨，调理清楚，逻辑性强，自己努力完成，没有抄袭。 | 30 | 30 | 26 | 23 | 20 | 10 |  | | 4 | 对实验过程中存在的问题分析详细透彻、深刻、全面、规范、，结合实验内容，有自己的个人见解和想法，并能结合该实验提出相关问题，给出解决方法。 | 30 | 30 | 26 | 23 | 20 | 10 |  | | 5 | 实验结果、分析和结论正确无误 | 20 | 20 | 17 | 15 | 13 | 6 |  | | 总得分 | | | | | | | |  |     签名（签章）：  日期：2023年5月25日 |
|  |

|  |
| --- |
| **河南理工大学教学上机实验报告** |
| 上机时间 2023 年 5 月 10 日 |
| **实验题目：**  **Numpy数值计算** |
| **实验目的和要求：**  1、掌握Numpy基本数据结构  2、掌握Numpy数组的创建和基础运算  3、掌握Numpy数组的统计分析方法 |
| **实验过程：**   1. **numpy数组的创建** 2. **使用np对象中的array（）方法，参数可以是列表数组。**     **2.通过np对象中其他的方法创建numpy数组：zeros(),ones(),empty(),arrange()。Np对象中linspace()方法不能传入小数但是最后一个数可以是20. 但是arange与range可以传入小数作为间隔值。**      **（二）关于ndarray对象数据类型**  **1.查看ndarray对象的数据类型：通过创建dtype对象，利用dtype对象中的name属性获取该ndarray对象的数据类型。可以直接用一个\_查看最近操作的数据数据类型。**    **2. 转换ndarray对象的数据类型：对ndarry对象用里面的astype（）方法转换ndarray对象的数据类型。被操作后返回结果原对象本身不变，被操作后没有返回结果原对象本身改变，如果数组中元素都是字符串型必须保证字符串中的内容都是整数才可以转换成整数类型。**    **（三）ndarray对象矢量运算**  **1.Ndarray对象之间的运算通过对象间的加减乘除元素级的矢量化运算操作，利用索引确定操作数位置与结果位置。**      **2.广播机制：通过对ndarray对象数组的扩展使数组对象的shape属性值相同。**    **（五）ndarray对象标量运算**  **即ndarray对象与常数的运算：产生一个与ndarray对象相同规模数组，每个元素与ndarray对象的每个元素进行运算根据索引放到新数组的相应索引位置上**    **（六）ndarray对象索引操作与切片操作**  **1.对不同维度的np对象索引的对象具有不同的结果，若是一维数组索引结果是数组中的元素，若是二维数组索引结果是一个一维数组或者是一个元素。除此之外np对象有花式索引：通过数组或者列表作为索引，根据索引数组或者索引列表中的每一个元素作为目标索引的下表进行取值。Np对象也可以通过布尔型数据进行索引，结果是布尔数组中true对应的数组中的值。**    **一维np对象的切片操作与py中列表的切片操作差不多。**      **花式索引 当有一个列表时获取的是列表中元素对应的索引的位置的值，当有两个列表时获取的是两个列表中相对应的索引位置上元素的值，返回的结果是数组类型的。**    **布尔型索引**    **（七）ndarray对象转置方法与属性**  **1.T属性对数组进行转置，通常对二维数组使用，不会改变数组中的值只会通过下标改变数组中元素的位置。**    **2.遇到高维数组np对象有transpose()方法，参数传入的是转置后的编号组成的元组。如果只需转换两个轴，可以用swapaxes()方法转化其中的某些轴。**    **（八）常见一元与二元的元素级函数运算**  **通用函数对ndarray中的数据执行元素级运算函数返回一个新的数组。**    **（九）利用ndarray中的方法处理数据**  **1.三元表达式转换为数组运算**     1. **Nd对象所具有的统计计算方法：sum(),mean(),min(),max(),argmin(),argmax(),cumsum(),cumprod().**     **对于cumsum(),cumprod()方法，产生的结果是对每个元素进行从开始到该元素相加或相乘的结果放在该元素的索引对应所在新数组的位置。**   1. **排序方法sort()，如果想要对某一个轴上的元素进行排序的话参数写想要排序的轴。**     **sort()方法本身会修改原数组中的内容，sort()方法，在二维数组中默认按照一轴进行排序，如果想要其按照指定的轴向进行排序的话可以在括号里面加上指定的轴向。**   1. **检索元素方法all()所有元素至少有一个满足条件,any()所有元素都满足条件结果返回类型都是布尔型。**      1. **集合逻辑方法，唯一化方法**   **唯一化方法中unique()函数用来找出数组中的唯一值有哪些并返回排序后的结果。**    **in1d()函数可以用来判断数组中的元素是否在另一个数组中，返回一个布尔型的数组，d前面的不是l而是数字1。**    **集合运算的函数：intersect1d(x,y)求集合的交集并返回有序的结果;union1d(x,y)求x和y的并集并返回有序的结果；setdiff1d(x,y)求集合的差返回在x中却不在y中的元素；setxor1d(x,y)求集合的对称差。** |
| **实验结果：**  **1.只有形状相同的数组才可以进行矢量化运算，如果形状不相同需要进行广播机制，进行广播机制以后不改变原数组的形状。广播机制变换方式是在想同维度上向最大值看起，之后进行拓展相同轴向直接赋值，拓展之后进行元素之间的运算。**  **2.数组与标量之间的运算还可以混合进行运算。**  **3.数组的转置并不会改变原数组。** |
| **实验分析：**  **1.通过对创建数组arane方法参数只能传元组或者列表，如果想要传进去一个字典怎么实现？显然是不能直接传进去字典类型的，既然参数里面只有列表和元组类型的参数我可以用list()函数将一个字典类型的对象取出值或者是键转化为一个列表传进去。当然也可以用字典对象的idems()方法将键值对转换为元组直接传进去。**  **2.如果astype方法中的字符不全是整数想要把数字转化成int型而字符仍保持原有类型表达的内容不变怎么办？搜尽境内全网也只能发现当数据类型带一个百分号的字符串时只能对这组数据使用lambda。**  **3.对于arr2 = np.array([1,2,3])为什么arr2.shape得到的结果是（3，）？如果shape得到的结果只有一个数字，表示数组只有1个维度，只有一个数字3表示这是一个一维数组里面有三个元素。**  **4.使用布尔型进行索引时除了可以有==运算符还可以有！=与—表示否定，也能用&与|组合多个布尔条件。**  **5.通过empty()创建出的数组的类型里面的数据都是float型，在进行对其中的值填充时填充后的值也是float型的。** |
| **实验成绩：**  日期：2023 年 5月 12 日 |
| **河南理工大学教学上机实验报告** |
| 上机时间 2023 年 5 月 10 日 |
| **实验题目：**  **Matplotlib可视化** |
| **实验目的和要求：**  1、掌握pyplot常用绘图参数调节方法  2、掌握子图布局和绘制方法  3、掌握常用图形的绘制与保存方法 |
| **实验过程：**  **（一）创建画布和常用操作**  **figure()方法创建一张空白的画布容纳图表中的组件。**  **1.figure()方法的语法：num:数字代表编号字符串代表名称。figsize:设置画布宽度高度以英寸为单位。dpi：设置图形的分辨率。facecolor：用于设置画板的背景颜色。edgecolor：用于显示边框的颜色。frameon：是否显示边框。clear：若设置为True且该图形已经存在则它会被清楚。**      **2.在一个画布上创建多个图形：figure对象中允许划分多个绘图区域每一个绘图区域是一个Axes对象有自己的坐标系统被称为子图。**  **subplot()方法语法：nrows，ncols表示子区网格的行数和列数：将整个画布分成nrows行ncols列的矩阵区域，之后从左上角子图为编号1开始按照从左到右从上到下顺序编号。index表示使用编号为多少的区域。如果这三个参数的值都小于10则它们之间可以省略逗号简写成一个实数。每调用一次只会在设定的比例区域中创建一个子图。**    **3.创建多个子图：subplots()语法格式：nrows，ncols：表示网格的行数列数默认为1。sharex，sharey表示x轴或y轴是否共享，True或all表示x轴和y轴在所有子图中所共享，False或None表示每个子图的轴是独立的，row表示沿着行方向共享轴，col表示沿着列方向共享轴。该方法返回的是一个元组第一个元素是Figure对象，第二个元素是Axes对象或者Axes对象数组。**    **4.添加和选中子图：add\_subplot()语法格式：\*args表示一个三位数或者是三个独立的实数描述子图的位置。**    **5.添加标签和图例：标签可以在绘制图形前后都可以添加但是图例只有在绘制完图形之后才可以添加。**    **（二）绘制常见图表**  **1.直方图：hist()方法：x表示输入值可以是单个数组或者不需要相同长度的数组序列。bins表示绘制条住的个数给定一个整数返回bins+1个条住默认是10。range表示bins的上下取值范围。color表示条柱的颜色默认是None。**    **2.散点图：scatter()方法，参数含义：x,y:表示x轴和y轴对应的数据。s:指定点的大小。若传人的是一维数组，则表示每个点的大小。c:指定散点的颜色，若传人的是一维数组，则表示每个点的颜色。marker:表示绘制的散点类型。alpha:表示点的透明度，接收01之间的小数。**    **3.柱状图：bar()方法，参数含义：x:表示x轴的数据一般是属性的值。height:表示条形的高度。width:表示条形的宽度，默认为0.8。color:表示条形的颜色。edgecolor:表示条形边框的颜色。**    **（三）保存本地图形：savefig()方法保存图形到本地指定的文件中。** |
| **实验结果：**  **1.直方图是一种显示数据分布的条形图，散点图用于可视化两个变量之间的关系，箱型图可以用来检测异常值情况，饼图可以用来显示百分比情况，折线图可以用来直观反映变化的趋势，条形图可以用来反应差异情况。**  **2.** **subplots() 函数是 Matplotlib 绘图的基础，用于创建由子图组成的网格布局。通过指定子图的行数和列数，可以创建不同排列的子图。**  **3. 图表中正确显示中文的方法：**    **4.matplotlib中的许多方法都是面向对象直接调用的。** |
| **实验分析：**  **1.** **虽然matplotlib在生成图例时，默认会选择一个最合适的位置，可以使用loc 自己设置图例位置。**  **2.matplotlib生成的图保存为pdf是矢量图放大不会变模糊而且占用的内存较小，而平常的png保存图片时使用位图保存，放大之后会变得比较模糊。**  **3.直方图和条形图的区别：条形图是用条形的长度表示各类别频数的多少，其宽度(表示类别)则是固定的。直方图是用面积表示各组频数的多少，矩形的高度表示每一组的频数或频率，宽度则表示各组的组距，其高度与宽度均有意义。由于分组数据具有连续性，所以直方图的各矩形通常是连续排列的，而条形图则是分开排列的。条形图主要用于展示分类数据，而直方图则主要用于展示数据型数据。** |
| **实验成绩：**  日期：2023 年 5月 12 日 |

|  |
| --- |
| **河南理工大学教学上机实验报告** |
| 上机时间 2023 年 5月 10 日 |
| **实验题目：**  **Pandas数据统计** |
| **实验目的和要求：**  1、掌握Pandas的基本数据结构  2、掌握Pandas索引的操作  3、掌握Pandas基础操作  4、掌握Pandas读写数据的方法 |
| **实验过程：**  **（一）pandas中数据结构**  **1. Series**  **Series：类似与一维数组，索引位于左边数据位于右边，Series构造方法中常用参数及其类型：data：可以是ndarray或者是list型或者时字典型的（字典型可以直接规定具体索引）；index：索引必须唯一不能有重复且于data中传入的数据长度相同；dtype：表示传入数据的类型；copy：是否复制数据默认为False。**      **Series对象的index和value属性：index属性返回的结果是一个索引列表对象，values返回的结果是一个数组，两者都可以用索引访问相应的值。**    **Series对象中的数据在进行运算之后结果仍然与这个索引保持对应的位置关系。**    **2.DataFrame**  **DataFrame对象是一个二维数组类型对象，具有行索引列索引，不同列数据类型可以不同。DataFrame的构造方法中主要的参数类型：data的数据类型可以是一个二维数组，index表示行标签，columns表示列标签。**    **每列数据的获取：两种方法，第一种可以利用列索引的方式获取，第二种可以利用对象的属性进行获取返回的结果是一个Series对象它的行索引与原来在DataFrame中一样。但是利用该对象的属性速度更快。（在实际中列名可能是空格字符所以用点的方式获取这一列时显得不太合适）**    **type()函数可以获取到对象的类型，参数传入想要获得对象类型的对象名。**    **增加某一列和删除某列：增加某列通过给索引或列名称赋值方式实现新增的列必须和原来的列的长度保持相同，删除某列时用del obj[‘列名’]。**    **（二）pandas索引操作**  **1.Pandas中的索引都是索引对象具有不可变性一旦定义之后就不会再被修改。索引对象还可以被多个对象所共用。**    **2.重置索引：对原索引和新索引进行匹配原索引数据按照新索引进行排序，如果新索引没有原索引的数据将会添加新的索引值用NaN或者是fill\_value()进行填充（fill\_value只会将所有的缺失值都填充为同一个值）。**  **reindex()参数：index，columns用于索引的新序列；method表示插值的填充方式有前向填充ffill，后向填充bfill，从最近索引值填充nearest三种方式；limit表示前向或者后向填充的最大填充量。**      **3.Series的索引操作：如果希望获取某个数据有通过索引位置获取数据和通过索引名称获取数据两种方式。如果希望获取某一个连续的数据可以通过切片操作来实现通过索引位置切片是左闭右开通过索引名称切片是左闭右闭。也可以通过列表获取多个不连续索引的值。**    **除此之外也可以通过布尔型索引筛选数据，创建布尔型对象再通过索引获取结果为True的数据。**    **4.DataFrame索引操作：每列数据都是一个Series对象可以通过列索引获取每一列数据，但是每一行数据不能直接用行索引去获取可以使用相邻数的切片来获取。**      **获取多个不连续的Series对象可以通过索引列表来实现，获取某一行某一列的数据时可以通过多个切片来操作先进行行索引取某几行的数据再进行列索引取某几列的数据。**    **除此之外pandas还提供了loc与iloc进行索引操作（对列而言），iloc基于位置进行索引（对列而言），loc基于标签进行索引。仍然满足通过索引位置切片是左闭右开通过索引名称切片是左闭右闭。**    **（三）层次化索引**  **创建具有两层索引的Series与DataFrame通过嵌套列表的方式实现但要保持两个列表的长度相同：**      **1.通过MultiIndex中的from\_tuples()将元组列表转化为MultiIndex对象，元组中第一个元素表示外层索引第二个元素表示内层索引。From\_tuples中tuples表示包含若干元组的列表，names表示两层索引的名称。**    **2.通过MultiIndex中的from\_arrays()将数组列表转化为MultiIndex对象，第一个列表表示外层索引第二个表示内层索引。From\_arrays中arrays可以接受数组也可以接受列表，names表示两层索引的名称。**    **3.通过MultiIndex中的from\_product()将多个集合的笛卡尔积转化为MultiIndex对象，iterabls中第一个集合表示外层索引第二个集合表示内层索引。From\_product中iterables接受多个集合，names表示两层索引的名称。**    **（四）层次化索引的操作**  **1.选取子集操作通过外层索引进行操作直接ser\_obj[‘外层索引名’]，如果想取某一内层索引ser\_obj[‘外层索引名’,’内层索引名’]，外层索引名必须要写如果不知道至少写一个：否则会报错。**      **2.交换分层顺序交换内层索引与外层索引的位置。**    **3.排序分层：sort\_index(),by表示按照指定的值进行排序（已经过时），ascending默认为True表示按升序排列。此方法会优先按照外层索引进行排列再按照内层索引进行排列。**      **（四）算术计算**  **先按照索引进行对齐相应索引位置上的值进行运算结果放在对应的索引位置上没有对齐位置的结果用NaN来填充。如果不用NaN填充数据可以用.add的方法里面有一个fill\_value进行运算中缺失值的填充而不是结果中缺失值直接填充。**    **（五）统计计算**  **统计计算的方法中一般都是按照列来进行运算的如果想要改变轴向可以在参数中设置axis=1。**    **可以用describe一次输出 多个统计指标**    **（六）数据排序**  **1.按索引排序：sort\_index()参数的意思：axis表示排序的方向按行索引还是列索引进行排序，level表示对指定索引级别的值进行排序，ascending默认为True表示升序排列，inplace默认为False表示不创建新的实例如果想要排序的对象本身改变需要用inplace=True，kind表示进行排序的算法可以选择merge，quicksort等。**      **2.按值排序sort\_values(),方法中的参数：by表示要排序的列名，na\_position有first和last两个取值first表示把NaN值放在开头False表示把NaN值放在结尾默认是放在末尾。**    **在dataframe中sort\_values()在by参数中传入多个列名可以将多个列中的值进行排序（但是传入的必须是一个列表）。**    **（七）读写数据操作** |
| **实验结果：**  **1.method参数中的向前和向后填充表示在axis0轴上轴所指的方向是靠后的方向。**  **2.type(\_)表示的意思是获取最近一步操作对象的类型。**  **3.loc或iloc索引操作与直接通过行索引或者是列索引进行索引操作的区别：loc或iloc的参数要写在[]之间并且逗号前面表示进行的行索引的操作逗号后面表示的是进行列索引的操作，而直接索引时第一个中括号表示的是对行进行的索引或者是切片操作第二个中括号表示的是对列索引进行的索引或者是切片操作。** |
| **实验分析：**  **1.关于Series中参数copy()的分析：拷贝分为浅拷贝于深拷贝，浅拷贝是copy中参数为False的情况，深拷贝是copy中参数为True时的情况（默认情况），浅拷贝与原始拷贝共享数据和索引，深拷贝有自己的数据和索引副本。**    **2.关于求series中每个值出现的次数的分析:可以用Series对象中的value\_count()实现对每一个值的统计。**    **3.关于排序分层时的by已过时的分析：如果想要按某一个列的值进行排序时可以用另外的sort\_values()进行传入参数by和ascending进行排序。** |
| **实验成绩：**  日期：2023 年5月 12 日 |
| **河南理工大学教学上机实验报告** |
| 上机时间 2023 年 5 月 10 日 |
| **实验题目：**  **数据预处理** |
| **实验目的和要求：**  1、掌握数据合并的基本原理与方法  2、掌握数据清洗的基本方法  3、掌握数据标准化方法  4、掌握数据转换的常用方法 |
| **实验过程：**  **（一）数据清洗**  **1.空值和缺失值处理：判断存在空值和缺失值的方法：isnull()存在空值或缺失值标记为True,notnull()存在空值或缺失值标记为False，这两个方法中参数只有一个obj表示传入的可以是DataFrame或者是Series对象。**    **统计一列中有多少个空值的方法：先判断是否为空值再用sum()方法进行统计。**    **2.处理空值的方法：填充空值或者是直接将空值或者缺失值丢弃。**  **如果丢弃的话用dropna()方法，该方法中axis取值可以是0或者1表示删除有缺失值的行还是列默认删除缺失值的行，how表示过滤的标准取值为any或者all默认为any表示存在NaN值则删除，thresh表示该行或者该列至少有thresh个非NaN值时将其保留，inplace表示是否在原数据上操作如果为True表示直接修改原数据但默认是False。**      **如果填充空值的话用fillna()方法参数有value表示用于填充的数值method与limit与之前的重置索引方法reindex()中的使用方法相同，method参数不能与value参数同时使用，fillna()中填充时填充类型可以是标量、字典、Series与DataFrame。**      **2.重复值处理：重复值的判断只要两条数据中所有的条目的值都相同就判断为重复值。标记是否有重复值的方法：duplicated()，删除重复值的方法drop\_duplicated()。**  **duplicated()方法中subset用于识别重复的行标签或者是所有的列标签序列，默认识别列标签，keep可以取first,last,False。first表示从前向后找除第一次外其余的标记为重复，last表示从后向前找除第一次外其余的标记为重复，False表示所有相同的都标记为重复。该方法的返回值是一个有布尔型数组的series对象重复标记为True不重复标记为False。**    **drop\_duplicates()方法表示删除重复值返回的结果与原来对象的类型相同，参数中ignore\_index设置为True时表示索引重新分配，其他参数与duplicated()方法中相同。**    **3.异常值处理：3σ原则检验异常值和基于箱型图检测异常值**      **检查出来异常值之后对异常值进行修改：replace()方法既可以对单个数据进行替换还可以对多个数据进行替换to\_replace表示被替换值方式，value默认为None表示用来替换to\_replace的值，limit表示前向或后向填充的最大尺寸，method表示向前填充还是向后填充，regex接受布尔值默认为false表示是否将to\_replace或value解释为正则表达式。**    **如果是替换多个值时可以用列表。**  **4.数据类型处理：两种方法：在创建对象时明确指出数据类型或者使用astype()方法或者to\_numberic()函数进行类型转换。**  **在创建pandas对象时直接通过dtype参数指定数据类型。**    **astype()方法强转类型：参数含义:dtype表示要转换为什么类型的数据，copy表示是否建立副本默认为true，errors表示错误处理方法raise或ignore抑制异常默认为raise允许引发异常。如果是有某些不是int型的数据转换为float就会发生异常errors就该起作用了。**    **to\_numeric()方法只能把非数据类型转换为数据类型：参数:arg表示要转换的数据可以是list，tuple，Series，errors表示错误的处理方式和astype中的一样。但是该方法不能直接操作DataFrame对象。**    **（二）数据合并**  **1.轴向堆叠数据：concat()方法，主要参数是axis，join和ignore\_index，join表示内连接inner还是外连接outer，默认是外连接。根据axis与join值组合有四种合并方式：横向堆叠和外连接，横向堆叠和内连接，纵向堆叠和外连接，纵向堆叠和内连接。如果是纵向堆叠当join的值是inner时表示将两个的行索引进行交运算如果join的值时outer时表示将两个的行做并运算，如果是纵向堆叠对列作相应的运算，以上运算之后如果有不存在的值将用NaN填充。**    **0轴方向上的inner与outer**    **1轴方向上的inner与outer**      **2.主键合并数据：merge()方法，方法中的参数作用：**  **首先是left和right分别表示参与连接的左侧DataFrame对象与右侧DataFrame对象。**  **其次是连接两个表中作为合并的键on，on的取值可以有多个，如果on有多个时需要用列表的形式表示出来。**  **最后是连接的方式how有四种取值分别是left,right,inner,outer当取left或right时表示左表或右表作为基准表通过on中的键进行连接on中在在左表或右表中的行会全部显示出来而另外一个表中只会显示与on所对应的基准表中的行索引所相同的行如果不存在则用NaN表示，Inner和outer则表示是以on中的键为基准的内连接还是外连接如果是inner将两个表中on相同的行提取出来，如果是outer将两个表中on中所有的行全部提取出来。**  **参数中还提供列索引没有重叠时的合并left\_index,right\_index表示以左侧或右侧行索引作为连接键。**  **如果想让整个表作为连接键的话可以设置left\_on，right\_on。**  **其他的参数有sort表示是否排序默认False，suffixes表示追加到重叠列名的末尾默认是\_x,\_y。**          **merge()方法中的默认参数：默认会使用重叠的列索引作为合并键并采用内连接的方式合并数据。**  **3.根据行索引合并数据：join()方法，该方法应该由要连接的左表进行使用，方法中other表示要进行连接的右表，on表示作为连接键的列名，how表示连接的方式分为四种使用方式左外连接、左内连接、右外连接、右内连接，当左连接时左表中的索引会成为新的表中的索引当为右连接时右表中的索引会成为合并之后表的索引，当how=inner时只有两个表中相同的行索引作为合并后表中的行索引，当how=outer时两个表中所有的行索引合并都会作为合并后表的行索引，如果合并后表某个值为空则用NaN来替换，lsuffix和rsuffix表示在左侧或右侧重叠的列名后加上后缀名。sort接收布尔值表示排序默认为false。**    **4.重叠数据的合并：combin\_first()方法，只有一个参数other，原理：根据索引对齐的方式将使用该方法的表中NaN值替换为other表中与NaN值相同索引位置上的值。**    **（三）数据重塑**  **1.重塑层次化索引：stack()与unstack()方法，前者是将数据的列旋转为行后者是将数据的行旋转为列。两个方法中的level默认为-1表示内层索引，设置为0表示外层索引。stack()中的dropna表示是否将过滤旋转后的缺失值接受布尔型数据。unstack()中fill\_value表示将旋转后的缺失值用fill\_value来填充。**    **将一个DataFrame对象可以通过stack()之后类型变为Series类型。**  **将一个Series对象可以通过unstack()之后类型变为DataFrame类型。(但在某些情况之下DataFrame仍转换为DataFrame)**  **2.轴向旋转：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。参数有index，columns，values，如果index与columns没有设置直接使用原对象中的行索引和列索引。**    **（四）数据转换:数据转换的方法有重命名轴索引，离散化和面元划分，计算指标和哑变量。**  **1.重命名轴索引：方法rename()常用的参数有index，columns表示要转换的行索引和列索引，通常用字典来表示。axis表示轴的名称可以用0，1来表示。copy表示是否复制底层数据，inplace默认False表示是否返回新的pandas对象。Level表示多级索引的级别名称内层为-1外层为0。**    **使用函数统一改变所有索引的名称：**    **使用字典改变部分索引的名称：**    **2.离散化连续数据：将离散化的数据拆分成几个区间。cut()方法实现，参数含义：x表示要分箱的数组，必须是一维的。bins接收int和序列类型的数据如果传入的是int类型的值表示在一维数组中最小值和最大值之间划分为多少个等间距区间，如果传入的是一个序列两个相邻的数为一个区间，则表示将x划分在指定的区间中，若不在序列表示的区间中，则为NaN。right:是否包含右端点决定区间的开闭默认为True，right为True时表示左开右闭right为False时表示左闭右开。labels:用于生成区间的标签。retbins:是否返回bin。precision:精度，默认保留三位小数。include\_lowest:是否包含左端点。cut()方法会返回一个Categorical 对象，表示cuts数组中的每一个值在哪一个区间中。**    **3.哑变量处理类别型数据：用来反映某个变量的不同类别将分类变量转换为哑变量矩阵或者指标矩阵矩阵的值通常用0或1来表示。将Series或DataFrame对象转换为哑变量的方法get\_dummies()参数类型：data接受要转化的对象。prefix表示列名的前缀默认是None，columns表示指定编码的列名。** |
| **实验结果：**  **1.缺失值一般用NaN表示空值一般用None表示，判断空值时可以用面向对象的方式df\_obj.isnull()也可以用 传参的方式pd.isnull(df\_obj)来进行判断。**  **2.在DataFrame中对每一列的操作都可以用索引法或者对象调用法来操作该列。比如df\_2['B'].astype(dtype = 'float')与df\_2.B.astype(dtype='float')**  **3.merge()方法中的on：如果左表或者右表有一个没有on中的键都会报错，后缀的产生条件：如果两个表中有多个相同的列采用这些列的真子集作为on的参数时其他没被选中的列会产生后缀。** |
| **实验分析：**  **1.对astype()和to\_numeric()方法的区别的分析：astype()如果是存在数值型数据之外的数据就无法转换，而to\_numeric()方法只能将非数据类型的转换为数据类型的。**  **2.轴向堆叠数据合并时与根据行索引合并数据的联系：轴向堆叠数据时将concat()方法中的参数设置为0时表示将行索引进行inner或者outer操作和根据行索引合并数据的形式一样但不同点在于依据行索引合并时有一个基准表，基准表中的数据会全部的显示在结果中。** |
| **实验成绩：**  日期：2023 年 5月 12 日 |
| **河南理工大学教学上机实验报告** |
| 上机时间 2023 年 5 月 17 日 |
| **实验题目：**  **数据分组与聚合** |
| **实验目的和要求：**  1、理解分组与聚合的原理  2、掌握groupby分组方法  3、掌握聚合操作  4、掌握自定义聚合操作 |
| **实验过程：**  **分组聚合包括三个步骤：拆分：按照指定的轴向和一定的规则将数据集拆分成若干个组。应用：将某个函数或者方法应用到每个分组中。合并：将应用后的新值整合到结果中。**  **（一）分组**  **方法：groupby()，返回的是根据调用对象返回调用对象相应的GroupBy对象。by表示按照什么样标准分组，可以接收列表或数组DataFrame中某列的名称，字典或Series对象，函数。axis表示分组的轴向默认是0。Level表示在多层次索引中指定某个级别分组。as\_index表示聚合后是否以组标签作为索引的DataFrame对象输出默认是True。sort表示是否对分组标签进行排序默认是True。**    **通过列名分组时会得到一组元组类型的数据第一个数据为该组的名称也就是by对应的一列数据第二个数据是该组的具体数据，索引也被拆分，分组后的结果不显示需要将结果遍历出来，也可以使用列表对指定列名进行分组。**    **通过列表或者Series对象进行分组时列表或者Series对象的长度必须和指定轴的长度相等列表或Series对象中的相同元素分到一组，被分组对象的索引与该组中元素索引相同的分到一组。**  **需求：原始数据很多很复杂想重新定一个标准来进行分组。前提：应该知道每行的具体内容。**  **如果Series或者列表的长度与要分组的长度不相同时只会将部分数据进行分组。**    **将字典中值相同的分为一组并以字典中的值作为分组后的组名标签。字典的键必须与列名或行名相对应。**    **分组键的函数会在相应的轴索引上都会被调用一次返回值被用作分组的名称。**  **df\_obj.groups查看分组信息，返回一个字典。**  **df\_obj.get\_group(‘A’)获取哪一组的具体信息。**    **（二）聚合**  **1.使用内置统计方法聚合：将分组后的每组数据使用内置统计方法再将每组使用方法得到的值进行聚合。**    **2.面向列的聚合方法：agg()参数func表示用于汇总数据的函数可以是一个函数也可以是一个函数列表。Axis表示函数作用的轴向默认是0轴。**        **对每一列数据应用同一个函数：自定义函数时可以不加引号但是内置函数进行分组时需要加上引号。**    **对同一列数据运用不同的函数，对每一列所用的函数在列表中写出来，如果想要给函数一个别名可以在列表中使用元组的形式将别名用字符串的形式写在第一个位置，所用函数写在第二个位置。**    **对不同列使用不同的函数：用字典的形式将列名写在键的位置上，该列所使用的函数写在值的位置上。**  **（三）数据转换**    **transform方法返回的结果有可以广播的标量值或者与数组大小相同的结果数组。transfrom方法将func()函数运用到各个分组中并在每个分组中相同列上的值相同。**  **可以使用特定的分组依据进行分组保证返回的结果与原数组形状相同。**    **（四）数据应用：拿到数据后想对数据有一个基本的了解**  **apply()方法：作用于每一行元素或每一列元素。可以在apply中使用正则表达式。** |
| **实验结果：**  **1.使用agg方法进行分组聚合运算时返回对象的形状与被分组数据集的形状是不相同的要想使分组后的数据与原来的形状保持相同可以通过transfrom的方法实现。**  **2.** **GroupBy对象其实是一个可迭代对象，无法通过直接输出的方式查看分组信息。除了通过前面介绍的for 语句循环遍历之外，还可以通过GroupBy 对象的属性或方法查看分组的信息，还可以利用apply应用lambda x:x函数进行查看分组信息但是返回一个DataFrame。**  **3.** **当使用groupby()方法对数据进行分组时，若给参数by传入的值是Series类的对象，此时会根据Series类的对象中的值决定如何分组，若值相同，则需要将原数据中与值位置对应的行拆分成一个分组；若值不同，则需要将原数据中与值位置对应的行自成一个分组。** |
| **实验分析：**  **1.** **分组与聚合是根据一定的拆分标准将原数据拆分成若干个分组，然后对每个分组应用统计运算，并把运算后的结果合并到一起。**  **2.主要有拆分，合并应用三个流程：原数据按照一定的拆分标准被拆分为若干个分组，每个分组应用某个函数或方法进行相应的操作，并产生一个标量值，分组名称与标量值整合成新的对象。** |
| **实验成绩：**  日期： 2023 年 5月 20 日 |
| **河南理工大学教学上机实验报告** |
| 上机时间 2023 年 5 月 17 日 |
| **实验题目：**  **时间序列分析** |
| **实验目的和要求：**  1、理解并掌握时间序列分析原理创建时间序列  2、掌握时间时序数据操作方法索引的切片  3、掌握时间序列数据预测方法 |
| **实验过程：**  **（一）创建时间序列**  **1.** **时间戳用Timestamp对象表示。通过to\_datetime()函数将datetime转换为TimeStamp对象。如果传入的是datetime列表，则将其强制转换为DatetimeIndex类对象里面的每一个都是TimeStamp对象。**    **时间序列类型之一是创建时间戳为索引的Series对象。将包含多个datetime对象列表传给index参数，同样能创建具有时间戳索引的Series对象。**      **DataFrame也可以有时间戳索引。**    **2.通过时间戳索引选择子集：直接使用位置索引来获取数据。使用datetime日期获取数据。直接使用日期字符串获取数据日期字符串符合可以被解析的格式。直接指定年份或者月份操作索引获取数据。通过truncate()方法截取 Series或DataFrame对象，方法中参数含义：before 表示截断此索引值之前的所有行，after表示截断此索引值之后的所有行，axis表示截断的轴，默认为行索引方向。**    **（二）固定频率的时间序列**  **1.date\_range()方法生成一个具有固定频率的DatetimeIndex对象，start：表示起始日期，默认为None。end：表示终止日期，默认为None。periods：表示产生多少个时间戳索引值。freq：用来指定计时单位。start、end、periods、freq四个参数至少要确定三个参数。**  **如果只是传入了开始日期默认freq参数为D表示按天计算的。**    **如果只传入了开始日期或结束日期则还需要用periods参数指定产生多少个时间戳。**    **如果想要时间序列中的时间戳都是每周固定的周几，则可以将freq参数设置为固定的周数。**    **如果日期中带有与时间相关的信息，可以将normalize参数的值设为True。**    **2.每个基础频率还可以跟着一个被称为日期偏移量的DateOffset对象。如果想要创建一个DateOffset对象，则需要先导入pd.tseries。**    **3.时间序列的移动：shift()方法用来前移或后移数据但数据索引保持不变。periods表示移动的幅度可以为正数也可以为负数默认值是1代表移动一次。**    **（三）时间序列数据预测方法**  **1.创建时期对象：创建Period类对象只需要在构造方法中传入一个字符串或整数的形式日期。**    **Period对象加上或者减去一个整数根据具体的时间单位进行位移操作。**    **具有相同频率的两个Period对象进行数学运算计算结果为它们的单位数量。**    **创建多个Period对象可以通过period\_range()函数实现。**    **2.时期频率转换：asfreq()方法来转换时期的频率。frep表示计时单位。how可以取值为start或end，默认为end。normalize表示是否将时间索引重置为午夜。fill\_value用于填充缺失值的值。**    **（四）重采样**  **1.** **resample()是一个对常规时间序列数据重新采样和频率转换方法rule表示重采样频率的字符串或DateOffset。fill\_method -- 表示升采样时如何插值。closed设置降采样哪一端是闭合的。**    **2.降采样：ohlc()方法**    **3.升采样：空值的填充，通过ffill(limit)或bfill(limit)方法，取空值前面或后面的值填充，limit可以限制填充的个数。** |
| **实验结果：**  **1.不同时间点上出现的不同数据形成一个序列的可能是循环的。**  **2.** **降采样时间颗粒会变大，数据量是减少的。为了避免有些时间戳对应的数据闲置，可以利用内置方法聚合数据。**  **3.** **降采样相当于另外一种形式的分组操作，它会按照日期将时间序列进行分组，之后对每个分组应用聚合方法得出一个结果。**  **4.升采样还可以通过fillna(‘ffill’)或fillna(‘bfill’)进行填充，传入ffill则表示用NaN前面的值填充，传入bfill则表示用后面的值填充。使用interpolate()方法根据插值算法补全数据。** |
| **实验分析：**  **1.时间序列分析的目标：描述：描述时间序列的主要特征，序列是递增还是递减；第二个解释变量如何影响时间序列的值？监控：检测时间序列行为何时发生变化突然出现峰值。预测：从当前值预测时间序列的未来值，并量化这些预测中的不确定性。回归：给定多个时间序列以及与这些序列对应的一个额外的值，找到其中的关系。分类：给定多个时间序列，将它们按照相似性进行分类。**  **2.时间序列中的值既可以是定期出现的也可以是不定期出现的。**  **3.滑动窗口根据指定单位长度框住时间序列，计算框内的统计指标，相当于一个长度指定的滑块在刻度尺上面滑动，每滑动一个单位即可返回框内的数据。** |
| **实验成绩：**  日期：2023 年 5月 20 日 |
| **河南理工大学教学上机实验报告** |
| **上机时间 2023 年 5 月 17 日** |
| **实验题目：**  **文本数据分析** |
| **实验目的和要求：**  1、掌握文本预处理流程  2、掌握文本情感分析  3、掌握文本分类方法 |
| **实验过程：**  **（一）文本预处理流程**  **文本预处理一般包括分词、词性标注、词形归一化、删除停用词。**  **分词：分词是指将由连续字符组成的语句，按照一定的规则划分成一个个独立词语的过程。**      **词性标注是指为分词结果中的每个词标注一个正确的词性调用pos\_tag()函数进行标注。**    **词形规范化过程主要包括两种：词干提取和词形还原，lemmatize()方法，用于获得单词的基本形式。**    **NLTK库里面有一个标准的停用词列表，在使用之前要确保已经下载了stopwords语料库，并且用import语句导入stopwords模块。**    **（二）文本情感分析**  **nltk.classify模块中NaiveBayesClassifier类实现了朴素贝叶斯分类算法，该类中有一个类方法train()用于训练模型。**      **（三）文本分类方法**  **数据准备：**    **特征抽取:**    **模型训练：** |
| **实验结果：**  **1.典型的文本处理流程：文本分词：将连续文本切分为有意义的单词或词组，这通常依赖于语言的特性和具体的分词工具。对于中文等没有明显词边界的语言，需要使用专门的分词算法或工具。词性标注：为分词后的每个词或词组标注词性（如名词、动词、形容词等）去除停用词和低频词：停用词通常对文本意义不大，可以去除。文本向量化：将文本转换为数值型数据，以便机器学习模型能够处理。特征选择：从向量化的文本中选择最相关和最有信息的特征。数据归一化/标准化：对数据进行缩放，以确保所有特征具有相似的尺度。划分数据集：将预处理后的数据划分为训练集、验证集和测试集。模型训练和评估：使用训练集训练选定的机器学习或深度学习模型。**  **2.** **选择合适的情感分析模型：**  **根据特定的任务和目标来选择模型。常用的情感分析模型包括基于规则、基于词典、基于机器学习和基于深度学习的方法。数据预处理：去除特殊字符、停用词、标点符号，以及进行词形还原、词性标注等操作。理解文本中的情感倾向：情感倾向通常可以划分为正面、负面和中性三个类别。** |
| **实验分析：**  **1.文本预处理的步骤可能因具体任务和所用模型的不同而有所调整。在某些情况下，可能需要进行更复杂的文本清洗和特征工程工作，以适应特定的数据集和任务需求**  **2.情感分析注意事项：对文档分词，找出文档中的情感词、否定词以及程度副词然后判断每个情感词之前是否有否定词及程度副词，将它之前的否定词和程度副词划分为一个组如果有否定词将情感词的情感权值乘以-1，如果有程度副词就乘以程度副词的程度值最后所有组的得分加起来，大于0的归于正向，小于0的归于负向。(得分的绝对值大小反映了积极或消极的程度)** |
| **实验成绩：**  **日期：**2023 年 5月 20 日 |
| **河南理工大学教学上机实验报告** |
| **上机时间 2023 年 5 月 17 日** |
| **实验题目：**  **数据分析实战** |
| **实验目的和要求：**  1、掌握数据读写操作  2、掌握数据过滤方法  3、掌握数据特征展示方法  4、能够进行数据结果分析 |
| **实验过程：**  **关于北京租房数据的统计分析：**  **1.数据来源和数据读取：**    **2.数据预处理：**  **重复值和空值处理：**        **数据类型转换：**    **3.数据特征和展示方法：**  **面积区间分析：**      **平均租金分析：** |
| **实验结果：**  **1.对平均租金的分析：从图中可以看出，西城区、东城区、海淀区、朝阳区的房租价格相对较高，这主要是因为东城区和西城区作为北京市的中心区，租金相比其他几个区域自然偏高一些，而海淀区租金较高的原因推测可能是海淀区名校较多，也是学区房最火热的地带，朝阳区内的中央商务区聚集了大量的世界500强公司，因此这四个区域的房租相对其他区域较高。**  **2.对面积的分析：通过图可以看出，50~70平方米的房屋在租房市场中占有率最大。总体看来，租户主要以120平方米以下的房屋为租住对象，其中50~70平方米以下的房屋为租户的首先对象。** |
| **实验分析：**  **1.** **从不同来源获取多组数据，如历史销售数据、用户调查数据等。对数据进行清洗、筛选和整理，确保数据的质量和准确性，符合研究目的。**  **2.** **运用统计分析方法对数据进行分析和处理。**  **3.** **根据数据分析结果选择合适的图表类型。**  **4.** **在生成初始图表后，进行反复的调整和优化，提高图表的可读性和效果。**  **5.** **对生成的图表进行细致的分析和解读，提取有价值的信息。** |
| **实验成绩：**  **日期：2023 年 5月 20 日** |