## 一、数据分析测试题

1、下列关于数据分析概念的描述错误的是（ ）。

A、使用适当的统计分析方法对收集来的大量数据进行分析

B、.数据分析可以从数据中提炼出有用的信息并形成结论

C、.数据分析可以对数据进行更深层的研究

D、.数据分析不能够在杂乱无章的数据中提取有用的数据

参考答案: D

3、关于数据分析的说法，下列描述错误的是（）。

A、数据分析可以通过计算机工具和数学知识处理数据

B、可以做出具有针对性的决策

C、数据分析没有实际的使用意义

D、在大数据环境下，数据分析能够挖掘出更有价值的信息

参考答案:C

6、数据分析的一般流程为（ ）。

A、明确目的和思路、数据收集、数据处理、数据分析、数据展现

B、.明确目的和思路、数据收集、数据分析、数据展现

C、.数据收集、数据处理、数据分析、数据展现

D、.明确目的和思路、数据处理、数据分析、数据展现

参考答案:A

7、下列选项中，属于数据处理的目的是（ ）。

A、使数据更加适用于分析

B、.对无意义的数据进行清洗

C、.整合数据

D、.提高数据质量

参考答案: A,B,C,D

8、下列选项中属于NumPy功能的是（）。

A、快速高效的多维数据对象ndarray

B、.高性能科学计算和数据分析的基础包

C、.多维数组（矩阵）具有矢量运算能力，快速、节省空间

D、.线性代数、随机数生成以及傅里叶变换功能

参考答案: A,B,C,D

答案解析:

Numpy是大部分Python科学计算的基础，它具有以下功能：

(1) 快速高效的多维数据对象ndarray。

(2) 高性能科学计算和数据分析的基础包。

(3) 多维数组（矩阵）具有矢量运算能力，快速、节省空间。

(4) 矩阵运算。无需循环，可完成类似Matlab中的矢量运算。

(5) 线性代数、随机数生成以及傅里叶变换功能。

9、关于Pandas特点的描述中，下列描述正确的是（）。

A、智能数据对齐和缺失数据的集成处理

B、基于标签的切片、精确索引和大数据集的子集

C、按数据分组进行聚合和转换

D、具有时间序列功能

参考答案: A,B,C,D

14、下列工具中，能够用于数据查看和分析的是（ ）。

A、NumPy

B、Pandas

C、Matplotlib

D、OpenCV

参考答案: A,B,C

18、Python是一门胶水语言，可以轻松地操作其它语言编写的库。

对

错

参考答案:

对

19、Jupyter Notebook的优点是可以重现整个分析过程，并将说明文字、代码、图表、公式和结论都整合在一个文档中。

对

错

参考答案:

对

20、只要当前的系统中安装了Anconda，则默认就已经拥有了Jupyter Notebook，不需要再另行下载和安装。

对

错

参考答案:

对

21、Numpy是Python开源的数值计算扩展工具。

对

错

参考答案:

对

23、Pandas是一个基于NumPy的数据分析包，它是为了解决数据分析任务而创建的。

对

错

参考答案:

对

25、Matplotlib是一个用在 Python 中绘制数组的3D 图形库。

对

错

参考答案:

错

答案解析:

Matplotlib是一个用在 Python 中绘制数组的2D 图形库。

26、使用Anconda进行开发，可以有效的解决包配置与包冲突的问题。

对

错

参考答案:

对

27、conda是一个在Windows、Mac OS、和Linux上运行的开源软件包管理系统和环境管理系统。

对

错

参考答案:

对

简述：

1、请简述为什么选择使用Python做数据分析。（至少3个）

参考答案:

1语法简单精炼，适合初学者入门

2拥有一个巨大且活跃的科学计算社区

3拥有强大的通用编程能力

4人工智能时代的通用语言

5方便对接其他语言

2、简述Anconda的特点。

参考答案:

1包含了众多流行的科学、数学、工程和数据分析的Python库

2完全开源和免费

3额外的加速和优化是收费的，但对于学术用途，可以申请免费的License

4全平台支持Linux、Windows、Mac OS X、支持Python2.6、2.7、3.4…

答案解析:

暂无解析

3、请简述什么是Anconda。

参考答案:

Anconda是一个可以便捷获取和管理包，同时对环境可以统一管理的发行版本，它包含了conda、Python在内的超过180个科学包及其依赖项。

4、请举例说明数据分析的应用领域。（至少3个）

参考答案:

数据分析可以应用在下列领域：

1电子商务

2医疗领域

3商业零售

4网络安全

5交通

6物流

## 二、科学计算库Numpy

1、下列选项中，不属于ndarray对象属性的是（ ）。

A、shape

B、dtype

C、ndim

D、map

参考答案:D

2、下面程序运行后的结果是（ ）。

import numpy as np

np.arange(1, 10, 3)

A、array([1, 4, 7, 10])

B、array([1, 4, 7])

C、array([2, 5, 8])

D、array([3, 6, 9])

参考答案: B

3、下列关于ndarray对象描述正确的是（ ）。

A、ndarray对象中可以存储不同类型的元素

B、.ndarray对象中存储元素的类型必须是相同的

C、. ndarray对象不支持小数操作

D、.ndarray对象不具备高维运算能力

参考答案: B

答案解析:

根据ndarry的特点可知，该对象中元素的类型必须是相同的

4、关于ndarray对象属性，下列描述错误的是（ ）。

A、ndim属性表示数组轴的个数

B、shape属性表示每个维度上数组的大小

C、size属性表示数组元素的总个数，等于shape属性元组元素的和

D、dtype属性表示数组中元素类型的对象

参考答案: C

答案解析:

size属性表示数组元素的总个数，等于shape属性元组元素的乘积

6、下列关于ndarray对象索引的描述错误的是（ ）。

A、ndarray对象中的元素可以通过索引和切片来访问和修改

B、精确索引可以将整数数组或列表进行索引，然后将数组或列表中的每个元素作为下标进行取值

C、布尔索引是将一个布尔数组做为数组索引，返回的数据是布尔数组中True对应位置的值

D、ndarray对象的多维数组索引和切片的使用方式与列表的使用方式完全相同

参考答案: D

答案解析:

例如想要获取二维数组中的某个数，需要使用“arr[x，y]”形式来获取

7、下列的数组统计计算中，用于计算数组中最大值的方法是（ ）。

A、max

B、maximum

C、min

D、maximal

参考答案: A

8、 下面的程序运行后，输出的结果为（ ）。

import numpy as np

arr1 = np.array([[0], [1], [2]])

arr2 = np.array([1, 2])

result = arr1 + arr2

print(result.shape)

A、(3, 2)

B、(2, 3)

C、(3, 0)

D、(2, 0)

参考答案: A

10、下列程序运行后，最终执行的结果为（ ）。

arr = np.arange(12).reshape(3, 4)

arr.shape

A、3

B、4

C、(3, 4)

D、(1, 2)

参考答案: C

12、已知，有如下一个二维数组：

arr2d = np.array([[1, 2, 3],[4, 5, 6],[7, 8, 9]])

如果希望获取元素5，则可以使用（ ）实现。

A、arr2d[1, 1]

B、.arr2d[1]

C、.arr2d[2]

D、.arr2d[1, 0]

参考答案: A

13、现有程序如下：

import numpy as np

arr = np.array([[6, 2, 7], [3, 6, 2], [4, 3, 2]]

arr.sort()

arr

对数组arr执行sort()方法后，程序运行结果是（ ）。

A、[[2 6 7] [2 3 6]]

B、[[2 6 7] [6 3 2]]

C、[[7 6 2] [6 3 2]]

D、[[7 6 2] [2 3 6]]

参考答案: A

14、在创建ndarray对象时，可以使用（ ）参数来指定元素类型。

A、dtype

B、dtypes

C、type

D、types

参考答案:A

15、要想创建一个3 \* 4的数组，下列选项正确的是（ ）。

A、np.arange(12).reshape(4, 3)

B、np.arange(12).reshape(3, 4)

C、np.arange(7).reshape(4, 3)

D、np.arange(7).reshape(3, 4)

参考答案:A

16、在NumPy通用函数中，用于计算元素级最大值的函数是（ ）。

A、max

B、maximum

C、min

D、maximal

参考答案:B

20、下列关于ndarray索引说法正确的是（ ）。

A、可以使用整数进行索引

B、可以使用整数数组进行索引

C、可以使用元组进行索引

D、可以使用布尔数组进行索引

参考答案: A,B,D

答案解析:

ndarray支持使用整数、整数数组、布尔数组进行索引

21、下列关于随机数模块描述正确的是（ ）。

A、NumPy的random模块具有更多的随机数功能

B、random模块中包括生成服从多种概率分布随机数的函数

C、numpy.random模块中有用于生成大量样本值的函数

D、rand()函数隶属于numpy.random模块

参考答案: A,B,C,D

24、下列不属于NumPy的数据类型的是（ ）。

A、char

B、varchar

C、int64

D、float32

参考答案: A,B

26、关于数组统计运算，下列说法中正确的是（ ）。

A、sum()方法表示对数组中全部或某个轴向的元素求和

B、mean()方法表示计算平均值

C、argmin()方法表示计算数组中的最小值

D、argmax()方法表示计算数组中的最大值

参考答案: A,B

27、ndarray对象的数据类型可以通过type()方法进行转换。

对

错

参考答案:错

答案解析:

ndarray对象的数据类型可以通astype()方法进行转换

28、数组使用切片和索引的方式与列表完全一样。

对

错

参考答案:错

30、ndarray对象中存储元素的类型必须是相同的。

对

错

参考答案: 对

43、

32、如果希望创建一个数组，则只能用array()函数实现。

对

错

参考答案:错

34、ndarray对象不支持切片操作。

对

错

参考答案:错

36、假设当前有一个3行3列的ndarray数组，如果想要获取第3行第2列的元素，可以使用ndarray[3,2]。

对

错

参考答案:错

答案解析:

索引从0开始计数

37、创建ndarray数组时可以不指定数据的类型。

对

错

参考答案:对

简述

2、请举例说明什么是数组的转置。

参考答案:

数组的转置指的是将数组中的每个元素按照一定的规则进行位置变换。NumPy提供了transpose()方法和T属性两种实现形式。其中，简单的转置可以使用T属性，它其实就是进行轴对换而已。

示例:暂无

## 三 Pandas

1、关于Series结构，下列描述正确的是（ ）。

A、Series是一个类似于二维数组的对象

B、Series由一组数据和与之相关的索引两部分构成

C、Series只能保存整数和字符串类型的数据

D、Series的索引默认是从1开始

参考答案: B

2、下面程序运行后，最终输出的结果为（ ）。

import pandas as pd

ser\_obj = pd.Series(range(1, 6), index=[5, 3, 1, 3, 2])

print(ser\_obj)

A、a 3.0d 2.0c 1.0b NaN

B、a 3.0b NaNc 1.0d 2.0

C、程序出现异常

D、c 1d 2a 3

参考答案: B

3、下列关于Pandas库的说法中正确的是（ ）。

A、Pandas中只有两种数据结构

B、Pandas不支持读取文本数据

C、Pandas是在NumPy基础上建立的新程序库

D、.Pandas中Series和DataFrame可以解决数据分析中一切的问题

参考答案:C

答案解析:

Pandas中还有另一种数据结构Panel

8、请阅读下面一段程序:

import pandas as pd

df\_obj = pd.DataFrame([[4, -1, -3, 0],

[2, 6, -1, -7],

[8, 6, -5, 1]])

print(df\_obj.sort\_values(by=1))

执行上述程序后，最终输出的结果为（ ）。

A、 0 1 2 30 4 -1 -3 01 2 6 -1 -72 8 6 -5 1

B、 0 1 2 30 4 -1 -3 01 2 6 -1 -72 8 6 -5 1

C、 0 1 2 32 8 6 -5 11 2 6 -1 -70 4 -1 -3 0

D、 3 2 1 00 0 -3 -1 41 -7 -1 6 22 1 -5 6 8

参考答案: B

9、Pandas在执行算术运算时，没有对齐的位置会使用（ ）进行补齐。

A、Null

B、0

C、NaN

D、null\_values

参考答案: B

10、下列选项中，不能创建一个Series对象的是（ ）。

A、ser\_obj = pd.Series([1, 2, 3, 4, 5])

B、ser\_obj = pd.Series({2001: 17.8, 2002: 20.1, 2003: 16.5})

C、ser\_obj = pd.Series((1,2,3,4))

D、ser\_obj = pd.Series(1,2)

参考答案: D

11、关于Pandas数据读写的说法中，下列描述错误的是（）。

A、read\_csv()能够读取所有文本数据

B、read\_sql()可以读取数据库中的数据

C、to\_csv()能够将结构化数据写入到csv文件中

D、to\_excel()能够将结构化数据写入到excel文件中

参考答案: A

12、下列关于DataFrame说法正确的是（ ）。

A、DataFrame结构是由索引和数据组成

B、DataFrame的行索引位于最右侧

C、创建一个DataFrame对象时需要指定索引

D、DataFrame每列的数据类型必须是相同的

参考答案: C

13、Pandas的数据结构有（ ）。

A、Series

B、DataFrame

C、Panel

D、Vector

参考答案: A,B,C

14、下列关于DataFrame说法正确的是（ ）。

A、DataFrame是一个类似二维数组的对象

B、DataFrame是由数据和索引组成

C、DataFrame有行索引与列索引

D、默认情况下DataFrame的行索引在最右侧

参考答案: A,B,C

答案解析:

默认情况下DataFrame的行索引在最左侧

15、关于读写excel文件的说法中，下列描述正确的是（ ）。

A、to\_excel()方法表示将结构化数据写入到Excel文件中

B、to\_excel()方法表示将结构化数据读取到DataFrame中

C、read\_excel()函数表示将结构化数据读取到DataFrame中

D、read\_excel()函数表示将结构化数据写入到DataFrame中

参考答案: A,C

16、下列Pandas方法中，用于求最大值和最小值的是（ ）。

A、sum

B、mean

C、max

D、min

参考答案: C,D

17、请阅读下面一段程序

import pandas as pd

ser\_obj = pd.Series(range(1, 6), index=[5, 3, 0,4, 2])

print(ser\_obj.sort\_index())

执行上述程序后，最终输出的结果是（ ）。

A、5 13 20 34 42 5

B、0 32 53 24 45 1

C、5 14 43 22 50 3

D、2 54 40 33 25 1

参考答案: B

18、下列关于Series说法正确的是（ ）。

A、Series是一个类似一维数组的对象

B、Series可以保存任何类型的数据

C、Series由数据和索引构成

D、Series结构的数据位于索引的左侧

参考答案: A,B,C,D

19、列关于Pandas的索引说法正确的是（ ）。

A、Pandas中的索引都是Index类对象

B、索引对象是可修改的

C、索引对象是不可修改的

D、Index对象是可以共享的

参考答案: A,C,D

答案解析:

索引对象是不可以修改的

20、在DataFrame中每列的数据都可以看做是一个Series对象。

对

错

参考答案: 对

21、DataFrame的结构是由索引和数据组成的。

对

错

参考答案:对

22、Series可以保存任何数据类型。

对

错

参考答案:对

23、在操作DataFrame对象时，可以通过指定索引名的方式获取数据。

对

错

参考答案:对

24、Series是一个类似于一维数组的对象。

对

错

参考答案:对

25、Series和DataFrame都支持切片操作。

对

错

参考答案:对

26、read\_csv()和read\_table()函数没有任何区别。

对

错

参考答案:错

答案解析:

read\_csv()与read\_table()函数的区别在于使用的分隔符不同，前者使用“，”作为分隔符，而后者使用“\t”作为分隔符。

27、Pandas只有Series和DataFrame两种数据结构。

对

错

参考答案:错

答案解析:

Pandas有Series、DataFrame、Panel三种数据结构

简述

1、请列举几个常用的pandas统计计算方法，并说明它们的作用。

参考答案:

常用的统计方法有sum、mean、median、max、min等，其中sum表示计算运算结果的和；mean表示计算运算结果的平均值；median表示计算运算结果的中位数；max/min表示计算运算结果的最大值/最小值。

2、请简述Pandas中有哪些常见的读写操作，并说明其作用。

参考答案:

常用的读写操作有read\_csv、to\_csv、read\_sql、to\_sql、read\_html、read\_excel、to\_excel，其中read\_csv和to\_csv表示对csv文件的读取与写入；read\_sql和to\_sql表示对数据库的读取与写入；read\_excel和to\_excel表示对excel文件的读取与写入；read\_html表示对网页中table标签数据的读取。

3、请简述什么是Series。

参考答案:

Series是一个类似于一维数组的对象，它能够保存任何类型的数据，主要由一组数据和与之相关的索引两部分构成。

4、请简述什么是DataFrame。

参考答案:

DataFrame是一个类似于二维数组或表格（如excel）的对象，它每列的数据可以是不同的数据类型。与Series的结构相似，DataFrame的结构也是由索引和数据组成的，不同的是，DataFrame的索引不仅有行索引，还有列索引。

## 四 数据预处理

1、下列选项中，用于删除缺失值的方法是（ ）。

A、isnull()

B、delete()

C、dropna()

D、fillna()

参考答案: C

答案解析:

isnull方法用于检测，当返回结果中有True时，表示有重复的数据；fillna方法用于填充缺失数据。

2、下列关于缺失值检测的说法中，正确的是（）。

A、null()和notnull()可以对缺失值进行处理

B、dropna()方法既可以删除观测记录，也可以删除特征

C、fillna()方法中用来替换缺失值的值只能是DataFrame对象

D、Pandas库中的interpolate模块包含了多种插值方法

参考答案: B

3、下列选项中，关于fillna()方法描述正确的是（ ）。

A、fillna()方法只能填充替换值为NaN的数据

B、只支持前向填充方式

C、默认可支持填充的最大数量为1

D、fillna()方法可以填充替换值为NaN和None的数据

参考答案: D

6、下列关于数据预处理的说法中，描述正确的是（ ）。

A、数据预处理的目的是为了提高数据质量

B、数据预处理是为了增加数据量

C、数据预处理是为了让图表更美观

D、 数据预处理的主要目的是删除数据

参考答案: A

7、关于空值和缺失值，下列说法中正确的是（ ）。

A、NaN和None是完全一样的

B、使用isnull()可以检测数据中是否存在空值或缺失值

C、notnull()与isnull()方法都可以判断数据中是否存在空值或缺失值

D、dropna()方法可以删除空值和缺失值

参考答案: B,C,D

8、检测出异常值之后，以下哪些处理方式比较合理？（ ）

A、直接将含有异常值的记录删除

B、用具体的值来进行替换，可用前后两个观测值的平均值修正该异常值

C、不处理，直接在具有异常值的数据集上进行统计分析

D、视为缺失值，利用缺失值的处理方法修正该异常值

参考答案:A,B,C,D

9、下列方法中，可以修改数据类型的是（ ）。

A、astype()

B、shape()

C、dim()

D、type()

参考答案:A

10、fillna()方法处理缺失数据时可以使用Series对象填充，但不可以使用DataFrame对象填充。

对

错

参考答案: 错

12、缺失数据都是人为有意造成的。

对

错

参考答案:错

简述：

1、请简述数据清洗的目的。

参考答案:

数据清洗的目的在于提高数据质量，将脏数据清洗干净，使原数据具有完整性、唯一性、权威性、合法性、一致性等特点。

## 五 数据可视化

1、下列pyplot模块的函数中，可以绘制条形图的是（）。

A、bar()

B、hist()

C、pie()

D、scatter()

参考答案: A

2、下列pyplot模块的函数中，可以绘制直方图的是（）。

A、bar()

B、hist()

C、pie()

D、scatter()

参考答案: B

3、下列图表中，适用于比较跨类别数据的是（）。

A、直方图

B、折线图

C、饼图

D、散点图

参考答案: D

4、下列关于常见图表说法正确的是（）。

A、散点图不能在子图中绘制

B、散点的x轴刻度必须为数值

C、折线图可以用作查看特征间的趋势关系

D、饼图可以用来查看特征间的相关关系

参考答案: C

5、下列函数中，可以为图表设置标题的是（）。

A、xlim()

B、xlabel()

C、title()

D、legend()

参考答案:C

6、下列参数中，调整后显示中文的是（）。

A、linestyle

B、linewidth

C、font.sans-serif

D、axes

参考答案: C

8、下列选项中，可以一次性创建多个子图的是（）。

A、figure()

B、subplot()

C、add\_subplot()

D、subplots()

参考答案:D

9、下列pyplot模块的函数中，可以绘制饼图的是（）。

A、bar()

B、hist()

C、pie()

D、scatter()

参考答案:C

10、下列函数中，可以设置当前图形x轴范围的是（）。

A、xlabel()

B、xlim()

C、title()

D、legend()

参考答案: B

11、下列pyplot模块的函数中，可以绘制散点图的是（）。

A、bar()

B、hist()

C、pie()

D、scatter()

参考答案:D

12、在创建Figure对象时，可以指定哪个参数来设置画布的尺寸？（）

A、num

B、dpi

C、figsize

D、facecolor

参考答案: C

13、下列函数中，用于保存当前生成的图表的是（）。

A、figure()

B、hist()

C、savefig()

D、show()

参考答案:C

14、下列图表中，可以清晰地反映出各数据系列的百分比情况的是（）。

A、直方图

B、折线图

C、饼图

D、散点图

参考答案: C

15、下列代码中，绘制散点图的是（）。

A、plt.scatter(x,y)

B、plt.plot(x,y)

C、plt.legend('upper left')

D、plt.xlabel('散点图')

参考答案: A

16、下列图表中，可以清晰地反映出数据变化趋势的是（）。

A、直方图

B、折线图

C、饼图

D、散点图

参考答案: B

17、关于Matplotlib的说法中，下列描述正确的是（）。

A、Matplotlib是一个Python 3D绘图库

B、它使用简单，仅用几行代码就能生成各种图表

C、仅仅适用于IPython平台

D、只能输出PNG或PDF格式

参考答案: B

答案解析:

Matplotlib是一个Python 2D绘图库，作图风格接近MATLAB，它已经成为Python中公认的数据可视化工具，适用于各种平台上。

18、如果希望一次性创建一组子图，则可以通过subplot()函数进行实现。

对

错

参考答案: 错

19、散点图包含的数据点越少，比较的效果就会越好。

对

错

参考答案:错

答案解析:

散点图是指数据点在直角坐标系平面上的分布图，通常用于比较跨类别的数据。散点图包含的数据点越多，比较的效果就会越好。

20、Matplotlib的多个子图中，位于左上角的子图编号为1。

对

错

参考答案:对

21、Figure对象允许划分为多个绘图区域。

对

错

参考答案:对

答案解析:

Figure对象允许划分为多个绘图区域，每个绘图区域都是一个Axes对象，它拥有属于自己的坐标系统，被称为子图。

22、subplot(323)和subplot(3,2,3)是等价的。

对

错

参考答案: 对

23、直方图可以利用方块的高度来反映数据的差异。

对

错

参考答案:对

24、Figure对象可以理解为一张空白的画布，用于容纳图表的各种组件。

对

错

参考答案:对

25、pyplot模块中默认拥有一个Figure对象。

对

错

参考答案:对

26、每调用一次add\_subplot()方法会规划画布划分子图，且只会添加一个子图。

对

错

参考答案:对

27、

直方图使用一系列高度不等的纵向条纹或线段表示数据分布的情况。

对

错

参考答案:对

简述

1、请简述数据可视化的意义。

参考答案:

数据可视化旨在借助图形化手段，清晰有效地将数据中的各种属性和变量呈现出来，使用户可以从不同的维度观察数据，从而对数据进行更深入地观察和分析。

2、请简述什么是数据可视化。

参考答案:

数据可视化是指将数据以图表的形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程。

3、请简述直方图与条形图的区别。

参考答案:

（1）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，因此其高度与宽度均有意义。

（2）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。

（3）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。