**一、单项选择题**

1. 在统计分析过程中，缺失数据是我们重点关注的对象之一。Pandas提供了强大的函数来处理缺失数据。假设数据表为“df”,下列哪一种语法能够统计df表中的缺失数据的数量（ ）。

A．df.notnull().mean()

B．df.isnull().mean()

C．df.isnull().sum()

D．df.notnull().sum()

2. Pandas中能对各变量进行基本统计量汇总的方法是( )。

A．describe()

B．head()

C．tail()

D．groupby()

3. 选择合适的统计学步骤取决于数据类型，在NumPy的数据类型中哪一种是错误的表示形式（ ）。

A．np.float32

B．np.float8

C．np.int8

D．np.int16

4. 数据分析的第一步应该总是目视检查原始数据，matplotlib可视化工具为我们提供了众多图形类型。下列哪一个matplotlib函数能够绘制出散点图（ ）。

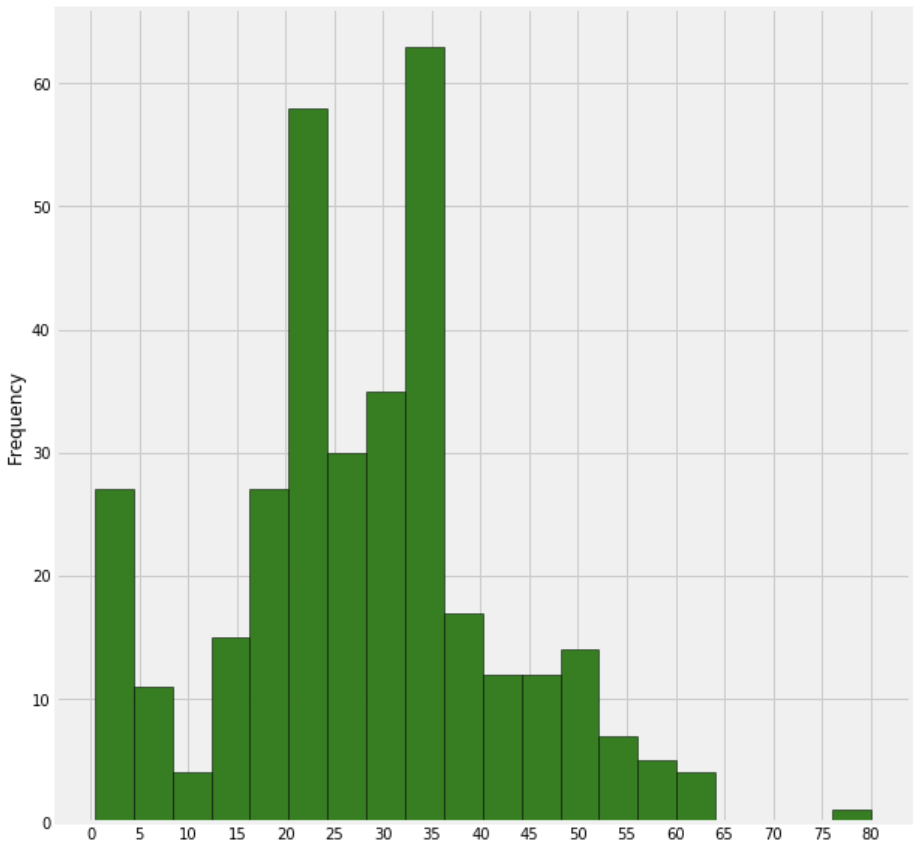
A. plt.bar()

B. plt.hist()

C. plt.box()

D. plt.scatter()

5. 下图是用matplotlib工具描绘的年龄的数量分布图，这种图形我们称作( )。



A. 密度图

B. 饼图

C. 直方图

D. 组合图

6. 下列哪一项不是箱线图的统计指标（ ）。

A. 标准差

B. 四分位数

C. 中位数

D. 最大值

7. 数据的可变性我们用方差来表示，下列哪项命令能够得到随机变量的方差 ( )。

A. np.mean()

B. np.var()

C. np.std()

D. np.median()

8. 在计算均值时，我们可以使用命令np.mean。现有一数组要计算其均值，我们可写作np.mean([0,10,20,30,np.nan]) ，它的结果为（ ）。

A、nan

B、15

C、20

D、报错

9. 极差反映了最高值和最低值之间的差异，在NumPy的命令中哪一项是极差的命令语法（ ）。

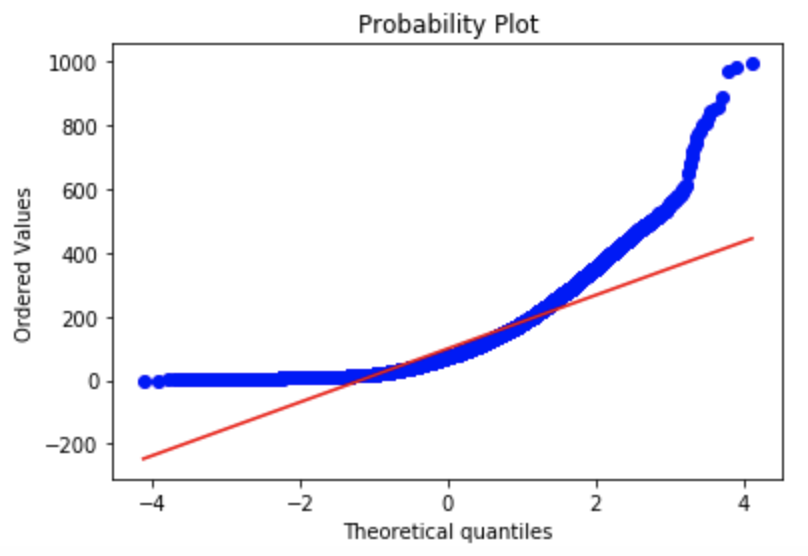
A、np.mean()

B、np.std()

C、np.diff()

D、np.ptp()

10. 在进行正态性检验时， probplot函数能够计算出的当前样本最可能的线性分布，我们可以直观地看到下图的线性拟合程度并不好。假设当前标签分布为右偏分布，这时为了提升拟合效果，我们可以采用下列哪种方法（  ）。



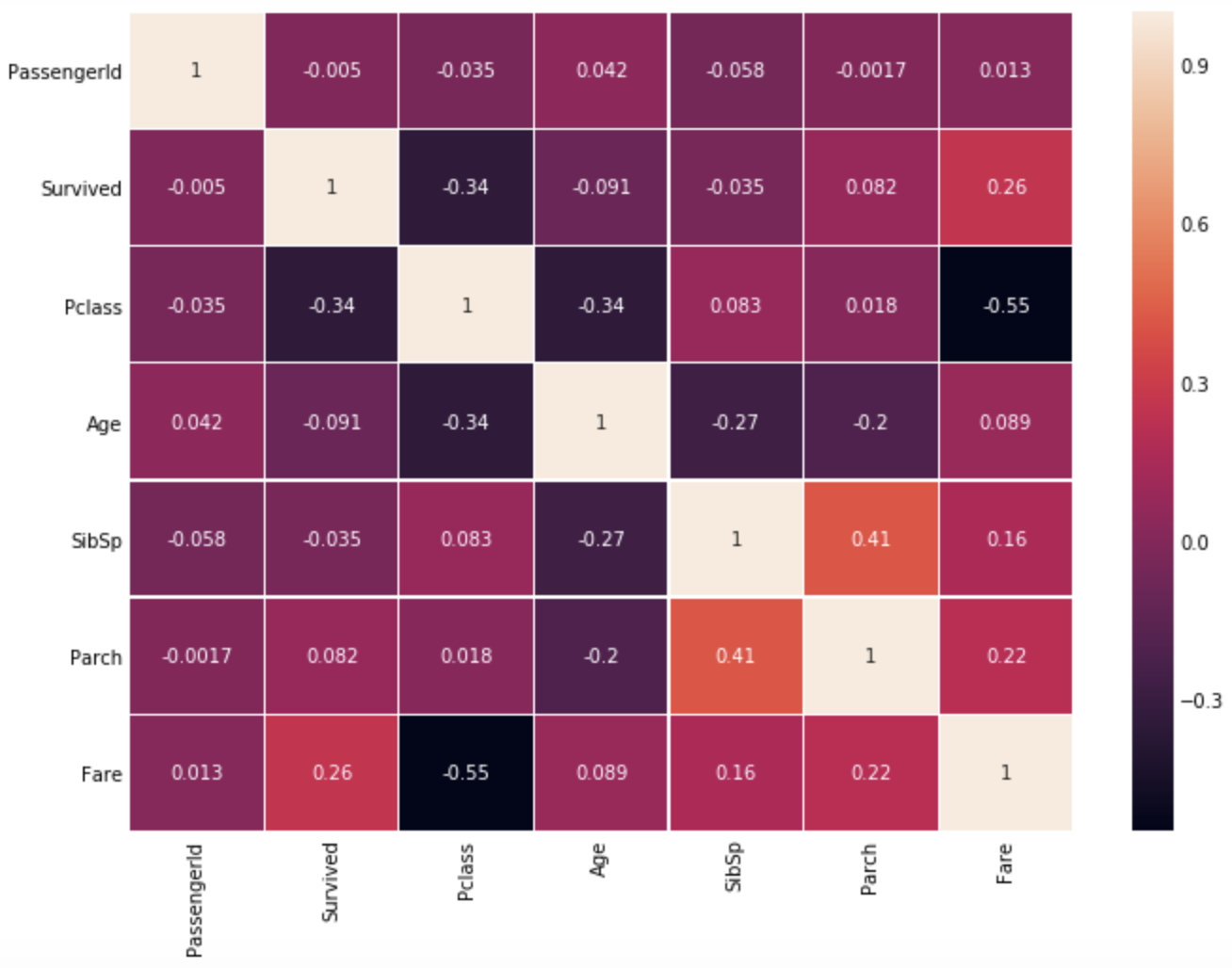
A. np.log()

B. np.exp()

C. np.mean()

D. np.quantile()

11. 在泰坦尼克号的数据挖掘过程中，我们对各变量进行了相关性分析。根据下图所呈现的相关性数据，我们可以认为（ ）



A. 变量之间存在高度相关性

B. 变量之间基本上相互独立

C. 负数值表示变量之间为正相关  
D. 该图能够反映变量间的多重共线性

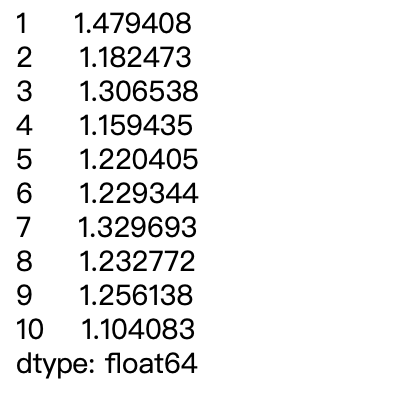
12. 在统计分析statsmodel模块中，我们可以使用variance\_inflation\_factor计算每个预测变量的方差膨胀系数，此系数是为了检验回归模型是否存在（ ）。

A. 参数的显著性

B. 线性相关性

C. 多重共线性  
D. 拟合优度

13. 下面的Series反映的是数据集中11个自变量的方差膨胀系数VIF，从数据上看，我们可见（   ）。



A. 自变量间不存在多重共线性

B. 自变量间存在严重的多重共线性

C. 自变量间存在高度的线性相关性

D. 自变量间不存在线性相关性

14. 在Pandas命令中，下列哪项命令能够计算出数据表df中变量间的皮尔逊相关系数（   ）

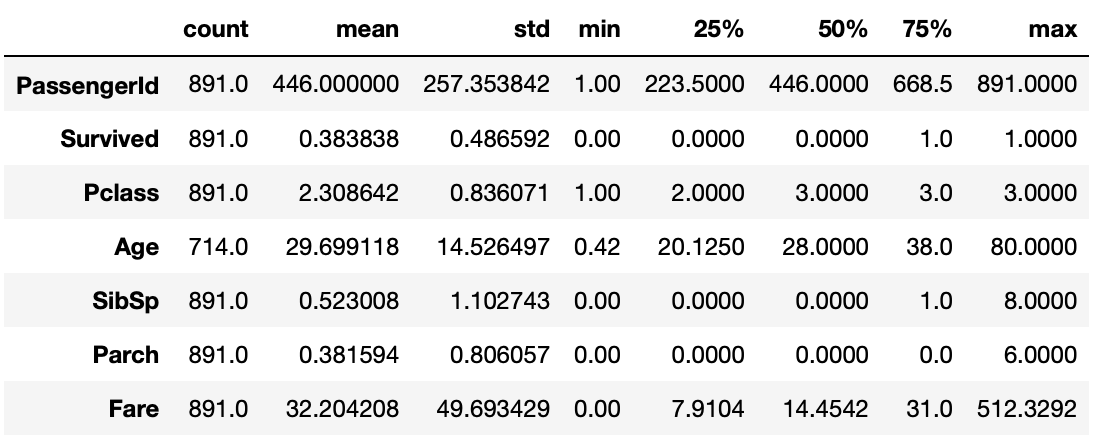
A. df.diff()

B. df.corr()

C. df.same()

D. df.value\_counts()

15. 下图为泰坦尼克号数据集中各变量的统计量。下列选项中说法错误的是（ ）。



A. 变量Age 在数据集中的最小值为0.42

B. 变量Age存在缺失数据

C. 变量Fare的上四分位数为7.9104

D. 变量间存在量纲的不统一，后续可以对其做标准化处理

16. 在数据预处理中，我们可以使用sklearn算法库中的PCA方法进行主成分分分析。下列关于主成分分析说法错误的是（ ）。

A. 我们一般保留的前k个主成分累计能够解释数据80%以上的变异

B. 主成分分析是一类常用的针对连续变量的降维方法

C. 主成分分析降维后会存在一定的信息损失

D. 主成分分析一定能够提升模型的表现

17. 现在我们想调查人们每天花多少时间看电视，按照“男性”和“女性”分组，我们可以使用下列哪项命令能够实现分组的操作（ ）。

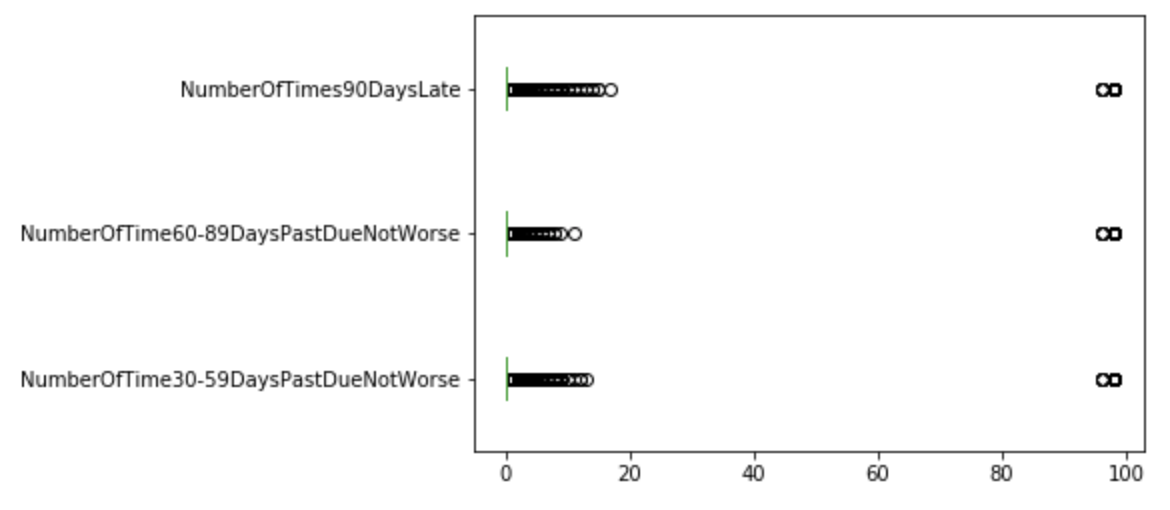
A. df.groupby()

B. df.cut()

C. df.qcut()

D. pd.value\_counts()

18. 下图所展现的是数据集中三个变量取值的箱线图的展示情况，可以看到三个变量都存在异常值。下列选项中对于异常值的说法错误的是（ ）。



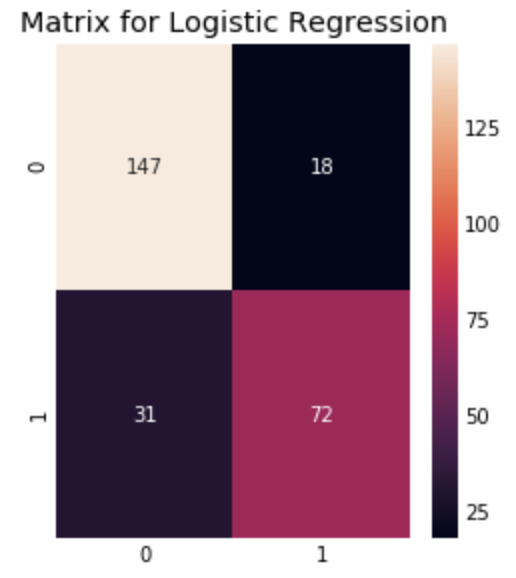
A．当异常值较少时，可以考虑删掉存在异常值的样本。

B．当需要保留所有样本时，可以使用盖帽法对异常值进行处理。

C．异常值一定是样本出现了记录错误。

D．在特殊情况下，异常值可以当做是个例单独处理。

19. 下图是逻辑回归模型预测后所输出的混淆矩阵结果，下列说法正确的是（ ）。



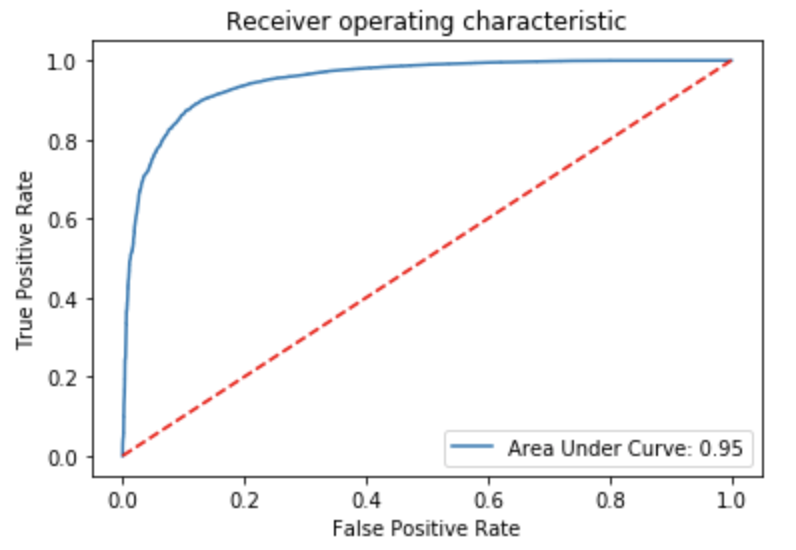
A. 左上角数字147数字表示将0这一类别的样本预测正确了147个。

B. 右上角数字18表示将0这一类别的样本预测正确了18个。

C. 左下角数字31表示将1这一类别预测正确了31个。

D. 右下角数字72表示将1这一类别预测错误了72个。

20. 下图为受试者操作特征（ROC）曲线。关于ROC曲线下列说法错误的是（ ）。



A. 和灵敏度、特异度紧密相关的是ROC曲线。

B. ROC曲线展示的横轴为假阳性率，即特异度。

C. ROC曲线展示的纵轴为真阳性率，即精确度。   
D. ROC曲线最理想的状态是逼近左上角。

**二、多项选择题**

1. Python中能够使用许多统计分析的算法库，下列选项中不属于建模工具的算法库有（ ）。

A．NumPy

B．statsmodels

C．sklearn

D．Seaborn

2.在逻辑回归模型中，我们时常对连续变量进行离散化处理，这么做可能会出现以下哪些结果（ ）。

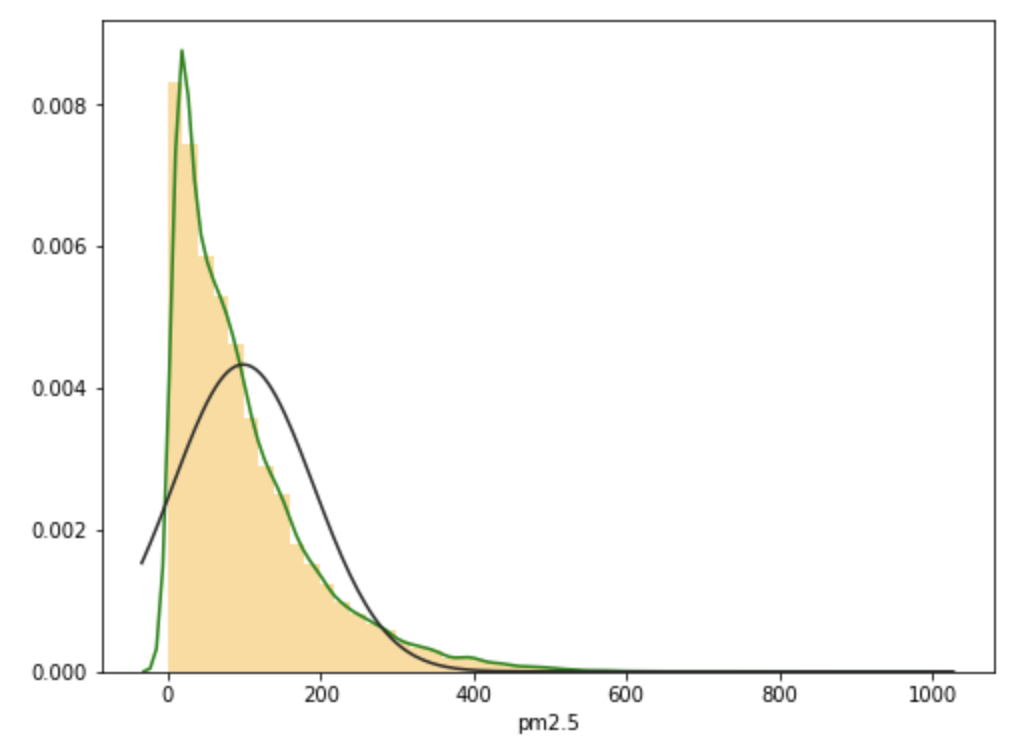
A．消除了异常值对模型的影响

B．造成了更大的误差

C．一定程度上提高了连续变量估计参数的可解释性

D．一定会提升模型的预测表现

3.下图是针对pm2.5浓度值的正态分布可能性检测，其中绿色曲线表示原始数据分布，黑色曲线表示原始数据在正态分布下的可能性。下列选项中说法错误的是（ ）。



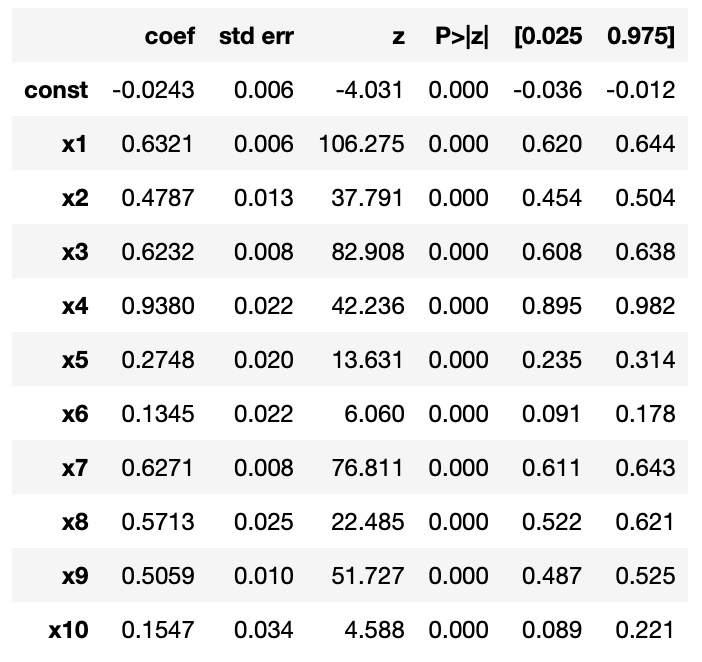
A. 该数据分布的偏度为右偏

B. 该数据分布的偏度为左偏

C. 想要使得该数据分布变换为近似正态分布，可以使用log对数转换

D. 想要使得该数据分布变换为近似正态分布，可以使用exp指数转换

4. 下表为利用statsmodels统计工具进行逻辑回归建模后得到的结果。下列说法正确的有（ ）。



A. 表中所有变量的p值均为0，则说明各变量通过了显著性检验。

B. 表中所有变量的p值均为0，则说明各变量的估计参数不显著。

C. 表中第一列表示的是各变量的参数估计值。

D. 表中第二列表示的是各变量参数的标准误。

5. 在统计学的数据类型中，数据可分为“分类型”或“数值型”。下列关于数据类型说法正确的有（ ）。

A. 离散的数值型数据只能取整数，例如子女的个数0，1，2，3 等。

B. 布尔类型的数据只有两个可能的取值，在Python中用True和False表示。

C. 当分类变量不止两个类别时，我们叫作名义类数据。比如，已婚/单身/离异。

D. 和名义变量相对的是等级变量，它是有序的，并且有逻辑顺序。比如，不满意/一般/满意。