**大数据预处理 复习题**

（一）单选题

1. 下列关于缺失值的形式的说法中，正确的是（ ）。

A．Python中默认的缺失值形式为NA

B．Pandas中默认的缺失值形式为NaN

C．Python中默认的缺失值形式为Null

D．Pandas中默认的缺失值形式为空字符串（""）

2．下列表述中，正确的是（ ）。

A．对字符串型变量的缺失值，需要使用众数进行填补。

B．对于缺失值，最简单的处理方法是使用均值进行填补。

C．完全变量指的包含缺失值的变量。

D．在调查过程中，因被调查者拒绝回答敏感问题而造成的数据缺失，属于人为原因。

3. 关于日期时间型数据，下列说法正确的是（ ）。

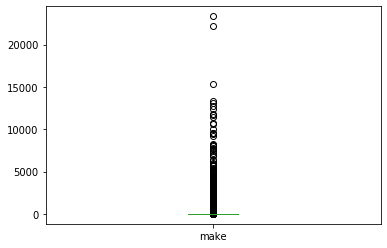
A．在计算机系统中，日期时间型数据是以文本形式存储的。

B．在计算机系统中，日期时间型数据是以数值形式存储的。

C．在计算机系统中，日期时间型数据是以因子形式存储的。

D．在计算机系统中，日期时间型数据是以缺失值形式存储的。

4. 下图是某分类变量各类别计数分布的箱线图，从图中可知（ ）。



A．少部分分类计数很低。

B．该变量类别很少，可以清晰的用箱线图展示各类别的频数。

C． 这个变量存在大量低频分类。

D．绝大多数分类计数很高。

5. 下列关于数据预处理的表述中，不正确的是（ ）。

A．具备专业经验的数据科学家可以在数据分析前忽略数据预处理。

B．数据预处理是在数据采集后，分析前这段时间里对数据进行的处理操作。

C．数据预处理的效果与数据分析顺利与否直接相关。

D．数据预处理往往占据数据分析项目总工作量的60%以上。

6. 下列数据特征缩放的公式中，正确的是（ ）。

A．数据中心化公式为：。

B．数据标准化公式为：。

C．Max-ABS缩放公式为：。

D．Robust缩放公式为：。

7. 下列关于相关系数的表述中，正确的是（ ）。

A．Pearson相关系数的值在[0,1]之间分布。

B．Spearman相关系数的值在[-1,1]之间分布。

C．变量和的Spearman相关系数的定义为：。

D．相关系数越大，则说明两个变量的相关性越强。

8.下列关于异常值的表述中，不正确的是（ ）。

A．异常值也可以称为离群值。

B．异常值的数值可能是真实可靠的。

C．异常值在数据中占得比例很小。

D．异常值完全是由于错误形成的数据。

9.下列关于哑变量的名称中，不正确的是（ ）。

A.二分类变量

B.虚拟变量

C.0-1型变量

D.数值型变量

10.下列关于异常值识别方法的表述中，不正确的是（ ）。

A.异常值的识别方式是考察变量中每一个样本值与其他样本值的相对距离。

B.通过箱线图可以识别异常值。

C.可以利用变量的样本均值和样本标准差来识别异常值。

D.异常值识别标准中使用了标准差的倍数作为距离大小的度量标准。

11. 下列表述中，正确的是（ ）。

A．含有缺失值的变量必须被剔除。

B．数据缺失的原因是该数据不存在。

C．数据缺失对于数据建模分析不存在影响。

D．随机缺失类型是指数据的缺失仅仅依赖于其它变量。

12. 下列关于数据错误的表述中，不正确的是（ ）。

A．数据集中所有错误的数据都可以被发现并予以纠正。

B．数据自身的逻辑规律可以帮助我们发现一些数据错误。

C．类别名称不统一会造成数据错误。

D．文字表述不规范会造成数据错误。

13. 在大多数箱线图绘图工具中，定义上限值位置和下限值位置分别为（ ）。

A． ，

B． ，

C． ，

D． ，

14. 下列关于数据缺失的表述中，正确的是（ ）。

A．缺失值处理的主要手段就是使用最接近的值进行填补。

B．只要数据录入人员避免漏录，数据就不会缺失。

C．在预处理阶段，遇到数据缺失情况可以忽略，不做处理。

D．某个变量的值为0，说明产生了数据缺失。

15. 下列关于数据预处理环节简化数据的表述中，不正确的是（ ）。

A．过多的数据会导致模型训练效率低下。

B．初学者进行数据分析时需要简化数据，专业数据分析师不需要简化数据。

C．简化数据时，需要保证数据集的信息不过多损失。

D．在样本量增加到一定程度后，信息含量趋近于不变。

16. 下列关于数据中心化的表述中，不正确的是（ ）。

A．中心化后，平均值变为0。

B．中心化后，标准差发生改变。

C．中心化后，极差没有发生改变。

D．中心化后，分布形状没有发生改变。

17. 下列关于数据归约的意义中，不正确的是（ ）。

A．可以降低无效、错误数据对数据建模的影响，提高建模准确性。

B．大幅缩减模型的训练时间，在需要反复训练模型的场景下能够极大地提高建模效率。

C．可以降低数据存储的空间成本。

D．有利于不同量纲数据之间的比较，也避免了自变量的不同量纲对建模的影响。

18. 下列关于异常值处理的表述中，不正确的是（ ）。

A. 异常值可能是准确的数据

B. 对于数据错误的异常值，可以删除

C. 删除异常值将改变数据集的样本量

D. 截断方法改变了样本容量

19.下列表述中，正确的是（ ）。

A．将连续型变量离散化为定性变量，其数据信息含量会增加。

B．等频法属于有监督离散化方法。

C．数据离散化是指把无限空间中有限的个体映射到有限的空间中，以此提高算法的时空效率。

D．数据离散化是指将连续型变量的部分取值映射到根据客观或主观依据事先确定好的一系列分组或分类中，从而得到另一个连续型变量的数据预处理方法。

20.关于哑变量与one-hot码，下列说法正确的是（ ）。

A．任何一个个类别的定性变量都可以转换为个哑变量。

B．任何一个个类别的定性变量都可以转换为个比特的one-hot码。

C．哑变量与one-hot码是完全相同的概念。

D．哑变量既有两个类别形式的，又有多个类别形式的。

21. 下列关于数据缺失类型的表述中，错误的是（ ）。

A．在某次调查中，在整理问卷时不小心打翻墨水瓶，导致部分问卷的部分问题污损，这种缺失记为非随机、不可忽略缺失。

B．数据缺失的三种类型是完全随机缺失、随机缺失和非随机、不可忽略缺失。

C．在某次调查中，某些消费者的“收入”变量发生了缺失，经过分析发现，没有缺失的收入数据大多属于中等收入水平，因而这种缺失属于非随机缺失。

D．随机缺失是指数据的缺失仅仅依赖于其它变量，即受本变量以外因素的影响。

22. 在大多数箱线图绘图工具中，定义上限值位置和下限值位置分别为（ ）。

A． ，

B． ，

C． ，

D． ，

23. 下列关于低频分类数据的表述中，正确的是（ ）。

A．低频分类数据都是由于数据采集是的错误导致的。

B．低频分类数据都是真实可靠的，只是其类别的样本频数比较低。

C．使用低频分类数据建立模型会使模型更简化，更精确。

D．低频分类指在分类型数据中出现一些类别频数非常低的情况。

24. 下列对定性变量的描述中，不正确的是（ ）。

A．“成绩”可以作为定性变量。

B．连续型数据很难转变为定性数据。

C．多分类定性变量可以转变为哑变量。

D．顺序型变量可以转变成得分变量。

25. 下列关于Robust缩放方法的表述中，不正确的是（ ）。

A．Robust缩放方法的理念是“首先中心化，然后除以尺度”。

B．Robust缩放方法用到的指标不易受极端值影响。

C．Robust缩放方法用中位数代替了标准化方法里的均值指标。

D．Robust缩放方法用方差代替了标准化方法里的标准差指标。

26. 下列关于变量的数据特征的表述中，正确的是（ ）。

A．原始数据的特征往往与分析方法的要求不一致。

B．变量的数据特征与测量尺度和采集方式无关。

C．两个数值型变量之间可以直接比较大小。

D．对数据进行缩放后，数据的分布形状会发生明显变化。

27. 下列关于用方差分析进行变量选择的表述中，不正确的是（ ）。

A．方差分析是用于检验两组或多组数据之间样本均值的差异是否显著的方法。

B．方差分析的检验形式是F检验。

C．P值以0.01作为筛选标准时，P值大于0.01的变量需要保留下来。

D．某些场合下通过方差分析选择变量，可以改善模型的建模效率和预测精度。

28.下列关于分箱法的表述中，不正确的是（ ）。

A.使用客观法进行数据离散化是计算机科学领域通常的做法，通常用名词“分箱”称呼这种方法。

B.分箱法主要有等宽法（等宽分箱）和等频法（等频分箱）两种形式。

C.等宽分箱法一般被称作“分类”

D.在统计分析领域，客观法和主观法中都会用到分箱法。

29.下列关于异常值处理方法的表述中，正确的是（ ）。

A．把异常值替换为某一可以接受的值，这一处理方法是标记法。

B.用一个变量将异常值标记出来的方法是截断法。

C.采用截断方法不会造成信息损失。

D.只有在研究目的是获取数据的一般规律特征时才有必要采取截断方法清楚异常值对于规律挖缺的干扰。

30.下列关于哑变量的说法中，不正确的是（ ）。

A.任何一个k个类别的定性变量都可以转换为k-1个哑变量。

B.哑变量只有两个类别，用0和1表示。

C.哑变量的0-1形态与逻辑型数据的False和True相同，0表示True，1表示False。

D.哑变量属于定性变量。

31. 下列表述中，正确的是（ ）。

A．含有缺失值的变量必须被剔除。

B．数据缺失的原因一定是该数据不存在。

C．数据缺失对于数据建模分析不存在影响。

D．随机缺失类型是指数据的缺失仅仅依赖于其它变量。

32. 下列关于数据错误的表述中，不正确的是（ ）。

A．数据集中所有错误的数据都可以被发现并予以纠正。

B．数据自身的逻辑规律可以帮助我们发现一些数据错误。

C．类别名称不统一会造成数据错误。

D．文字表述不规范会造成数据错误。

33. 下列表述中，错误的是（ ）。

A．低频分类数据都是真实的分类结果确实存在低频类别导致的。

B．可以通过绘制分类变量的各类别频数分布的箱线图来发现数据是否为低频分类数据。

C．过多的低频类别会严重影响建模的效率，在数据预处理阶段需要尽量进行处理。

D．低频分类数据通常呈现出类别众多，而且很多类别仅有几个甚至一个样本的情况。

34. 下列关于数据预处理的目的中，不正确的是（ ）。

A．完善数据，使其满足分析建模的质量和数据形式要求。

B．使数据集得以简化，提高建模分析的效率。

C．作为数据分析项目的第一步，为数据采集提供指导。

D．提高数据信息含量，进而提高数据分析模型的准确性。

35. 下列关于数据标准化的表述中，不正确的是（ ）。

A．标准化后，正负号代表了原始值是大于（+）还是小于（-）均值。

B．标准化后，可以避免数据尺度差异对后续建模造成负面影响。

C．标准化后，数据的标准差并没有变化，只有均值发生变化。

D．标准化后，通过新值可以发现原始数据中的异常值。

36. 下列关于Max-ABS缩放的表述中，不正确的是（ ）。

A．Max-ABS缩放将原值大于0的数据缩放到区间(0,1]。

B．不包含负数的变量经过Max-ABS缩放后的结果可能会出现负数。

C．原值等于0的数据缩放后还为0。

D．Max-ABS缩放的尺度就是变量绝对值的最大值。

37. 下列关于决策树模型的表述中，不正确的是（ ）。

A．越晚进入决策树的变量，其对于目标变量而言越重要。

B．该算法形象的以树状结构建立模型，再现了人类决策的过程。

C．具有建立过程直观易理解、便于可视化、应用范围广等一系列优点。

1. 存在不能保证得到全局最优决策树、容易形成复杂结构从而过拟合等缺点。

38. 下列关于数据离散化的方法中，不正确的是（ ）。

A. 数据离散化可以分为客观法和主观法两种。

B. 客观法基于研究目的确定类别。

C. 通常用名词“分箱”称呼客观法。

D. 分箱法主要有等宽法（等宽分箱）和等频法（等频分箱）两种形式。

39. 下列不属于异常值处理正确方法的是（ ）。

A. 直接删除

B. 标记

C. 截断

D. 替换为任意值

40. 确定异常值时，是比较与（ ）的距离

A. 平均值

B. 中值

C. 最大值

D. 最小值

（二）判断题

1. 若数据集中某样本包含缺失值，则该样本需要被删除。（ ）

2. 所有的数据错误都可以被发现并纠正。（ ）

3. 低频分类数据的处理方式通常是将频数过低的类别删除。（ ）

4. 数据缺失使得数据集信息含量降低，还会使一些模型无法应用。（ ）

5. 在数据预处理时，数据集中变量的数量不能发生变化。（ ）

6. Min-Max缩放可以将数据缩放至任意给定的范围内。（ ）

7. 模型预测准确度总是随着样本数量的增加而同步增加。（ ）

8. 如果输入数据是连续型数据，使用决策树模型时，就必须将连续型变量离散化为定性变量使用。（ ）

9. 数据离散化指的是将连续型变量在保留其基本数据含义的基础上转换为定性变量的操作。（ ）

10. 所有的异常值都必须删除。（ ）

11. 在数据存储过程中，由于设备故障造成存储失败而导致的数据缺失，属于客观条件原因。（ ）

12. 日期时间型数据不能进行任何计算。（ ）

13. 过多的低频类别会严重影响建模的效率。（ ）

14. 数据分析项目中，数据预处理环节不重要，可以忽略不做。（ ）

15. 经过标准化处理后，新值体现的是原值在序列中的相对位置。（ ）

16. Pearson相关系数的值均在[0,1]之间分布。（ ）

17. 变量选择的目的是删除数据集的行，样本归约的目的是删除数据集的列。（ ）

18.出现在训练集中的异常值，会干扰模型的训练。（ ）

19. 通过考察变量中每一个样本值与变量分布中心的相对距离来识别异常值。（ ）

20. “分箱”是客观法数据离散化的方法。（ ）

21. 完全随机缺失类型是指数据的缺失不受任何内部和外部因素的影响。（ ）

22. Python中，使用datetime对象的datetime.now()方法获得当前系统时间。（ ）

23. 低频分类数据的形成都是由于真实的分类结果确实存在低频类别。（ ）

24. 数据预处理是指在对数据进行分析前需要对数据进行的处理工作。（ ）

25. 数据中心化是数据标准化的第一个步骤。（ ）

26. Max-ABS缩放只可以将变量缩放至区间[0,1]。（ ）

27. 在进行变量选择时需要遵循的原则是，剔除的变量必须对数据分析影响较小。（ ）

28. 异常值的数值可能是真实的值。（ ）

29. 一旦发现异常值，就应当将其删除。（ ）

30. 数据离散化是指将定性变量转换为连续性变量。（ ）

31. 包含缺失值的数据集表现出来的不确定性与不包含缺失值的数据集相比显著增大。（ ）

32. 逻辑纠错中比较简单的方式是将不合理的值替换为缺失值，这样既保持了变量的性质不变，又避免了错误数据的危害。（ ）

33. 低频分类数据都是由于数据采集是的错误导致的。（ ）

34. 具有不同的数据特征的变量可以直接纳入到同一个分析体系中。（ ）

35. 经过Max-ABS缩放后，数据的正负可能会发生改变。（ ）

36. 数量归约是指从所有样本中选择一个有代表性的子集，因此也称为样本归约。（ ）

37. Spearman相关系数为-1时，两个变量不相关。（ ）

38. 将数据离散化后，可以克服连续型变量中隐藏的缺陷，使模型结果更加稳定，得到更加有意义的研究结论。（ ）

39. 使用客观法得到的是分类形式或顺序形式定性变量，依据的是研究者基于研究目的所确定的类别。（ ）

40. 利用箱线图可以识别出异常值。（ ）

（三）简答题

1. 请简述数据特征缩放对数据分析的意义。

2. 请简述异常值的含义。

3. 简述缺失值填补的思路。

4. 请简述使用相关系数选择变量的思想和步骤。

5. 请简述将变量缩放至区间[-1,1]时，Max-ABS缩放与Min-Max缩放的效果是否相同。

6. 简述数据离散化的操作含义。

7. 简述低频分类数据的特点、形成原因及影响。

8. 简述等宽法的适用情况。

（四）程序填空题（请在 上填写合适的代码）

1. 计算car\_data数据集中的make变量每一个类别的频数，存储在make\_count中：

make\_count = car\_data["make"].

2. 对boston数据集中的变量LSTAT进行处理，随机生成了10个缺失值：

sample = random.sample( (boston.shape[0]), 10)

boston. [sample, "LSTAT"] = np.nan

3. 对序列B进行中心化，结果保存在centralize\_b中：

centralize\_b = (B, with\_std=False)

4. 对数据集data中的所有变量进行Min-Max缩放，结果保存在data\_scale中：

mm\_scaler = ()

data\_scale = mm\_scaler. (data)

5. 通过画箱线图对数据data识别异常值：

data.plot. ()

6. 对数据data计算平均值，结果保存在result中：

result=data. ()

7. 对data数据采用等频分箱，分组数为5组，结果保存在result中：

result=pd. (data,bins=5)

8. 将data离散化为0-1型变量，结果保存在result中：

result=data. eq(0).

9. car\_data数据集中的make变量每一个类别的频数存储在make\_count中，现在将car\_data数据集中的make变量频数低于100的类别合并为一类：

car\_data["make1"] = car\_data["make"].map(

**lambda** x:"category\_under100"

**if** **else** , na\_action="ignore")

10. 将data\_1数据集age变量中的-1和999替换为缺失值：

data\_1["age"]. ([-1, 999], np.nan, inplace=True)

11. 对数据集data中的所有变量进行Max-ABS缩放，结果保存在data\_scale中：

ma\_scaler = ()

data\_scale = ma\_scaler. (data)

12. 计算变量y与x的Pearson相关系数：

Pearson\_xy = (y, x)

13. 使用客观法离散化data，设定了参数为bins = [0, 10000, 100000, 200000, np.inf]，标签为["new", "used", "old", "worn"]，结果保存在result中：

result=pd.cut(data,bins=[0,10000,100000,200000,np.inf], ,include\_lowest=True)

14.对数据data计算标准分数，data\_mean是data的平均值，data\_std是data的标准差，结果保存在result中：

result=(data- )/

15.对3倍标准差异常值分类计数，结果保存在result中：

result= data. ()

16. 查看数据集car\_data中所有变量缺失值情况，将结果保存在对象missing\_car中：

missing\_car = car\_data. .sum()

17. 使用boston数据实现线性回归模型填补。首先初始化一个线性回归模型，模型的训练集为train，将含有缺失值的变量LSTAT作为目标变量，其余变量作为自变量拟合模型：

train = boston.dropna(subset=["LSTAT"])

reg = LinearRegression()

reg.fit(X=train. ("LSTAT", ), y=train["LSTAT"])

18. 将数据集data中的所有变量缩放到区间[0,5]，结果保存在data\_scale中：

new\_scaler = ( =(0,5))

data\_scaler = new\_scaler.fit\_transform(data)

19. 使用test\_x数据集计算只包含部分变量var\_new的新模型model\_new的AUC值：

auc = (y\_true=test\_y,

y\_score=model\_new. (test\_x[var\_new])[:,1])

20. 对data数据采用等宽分箱，分组数为5组，结果保存在result中：

result=pd. (data,bins=5)

21.对数据data计算标准差，结果保存在result中：

result=data. ()

22.使用截断方法，将大于data\_mean+5\*data\_std赋值为data\_mean+5\*data\_std ：

data[data. (data\_mean+5\*data\_std)] = data\_mean+5\*data\_std

1. 将数据集data中的变量v1的缺失值使用其均值进行填补，将填补完的变量保存在v1\_fill中：

v1\_fill = data["v1"]. (data["v1"]. )

1. 截取addr的前三个字，并查看唯一值：

addr. (0, 3).unique()

1. 对数据集data中的所有变量进行Robust缩放，结果保存在data\_scale中：

rob\_scaler = ()

data\_scale = rob\_scaler. (data)

1. 计算变量y与x的Spearman相关系数：

Spearman\_xy = (y, x)

27. 找出data中3000到10000的值，结果保存在result中：

result=data. &data.

28. 计算离散化变量的各类频数，结果保存在result中：

data.

29. 对group\_a组和group\_b组样本进行方差分析的F检验：

anova = (group\_a, group\_b)

（五）程序理解题（根据所给的代码，写出代码所能完成的功能，每题10分，共20分）

1. 请写出下方代码的功能

scaler = StandardScaler()

new\_data = scaler.fit\_transform(mydata)

print(pd.DataFrame({"Scale":scaler.scale\_,"Mean":scaler.mean\_}, index= mydata.columns))

2. 请写出下方代码的功能

train, test = train\_test\_split(data, test\_size=0.3)

3.请写出下方代码的功能

car\_data = pd.read\_csv(r"d:cardata.csv",header=0)

v1\_fill = car\_data["v1"].fillna(car\_data["v1"].median())

4. 请写出下方代码的功能

data\_1 = copy.deepcopy(data)

5. 请写出下方代码的功能

dt\_example = pd.Series("")

dt\_example[0] = "2020/01/20"

dt\_example[1] = "2020/02/20"

dt\_example = pd.to\_datetime(dt\_example, format = "%Y/%m/%d")

6. 请写出下方代码的功能

pearson = pd.Series(name="pearson correlation")

for i in train\_x:

pearson[i] = pearsonr(train\_y, train\_x[i])[0]

var\_cor = pearson.abs() > 0.5

7. 请写出下方代码的功能

model\_all = GradientBoostingClassifier(random\_state=0)

model\_all.fit(X=train\_x, y=train\_y)

feature\_imp = pd.Series(model\_all.feature\_importances\_,

index=train\_x.columns)

var\_tree = feature\_imp.sort\_values(ascending=False).head(8).index

print(feature\_imp.sort\_values(ascending=False))

print("\n结果为：\n%s" % var\_tree.values)

8. 请写出下方代码的功能

bin\_1=pd.cut(price,bins=5)

d1={“price”:car\_data[“price”],”bin”:bin\_1}

p1=pd.DataFrame(data=d1)

print(“等宽分箱结果:\n%s”%p1[0:20])

print(“等宽分箱频数分布:\n%s”%bin\_1.value\_counts())

**参考答案**

（一）单选题（每题2分）

1．B 2．A 3．B 4．C 5．A

6．C 7.B 8. D 9. D 10.A

11．D 12．A 13．D 14．A 15．B

16．B 17. D 18.D 19. C 20.A

21．A 22．D 23．D 24．B 25．D

26．A 27.C 28. C 29. D 30.C

31．D 32．A 33．A 34．C 35．C

36．B 37.A 38. B 39. D 40.A

（二）判断题（每题2分）

1． 2． 3． 4． 5．

6． 7. 8. 9. 10.

11． 12． 13． 14． 15．

16． 17. 18. 19. 20.

21． 22． 23． 24． 25．

26． 27. 28. 29. 30.

31． 32． 33． 34． 35．

36． 37. 38. 39. 40.

（三）简答题（每题10分）

1. 意义共有三点：一是，多数数据特征缩放的方法可以消除数据的量纲，而保留其数据分布特征，这样更有利于不同量纲数据之间的比较，也避免了自变量的不同量纲对建模的影响；二是，数据特征缩放可以提高梯度下降求解（迭代运算）的收敛速度，提高建模效率；三是，数据特征缩放可以提高一些模型的预测精度。

2. 异常值也可以称为离群值，指在一个数据序列中与大多数值相比特别大或特别小的值。

3. 第一个思路是利用包含缺失值的变量自身的信息进行填补，主要形式为使用该变量的非缺失部分构造简单统计量，并用该统计量填补缺失部分。第二个思路是同时利用包含缺失值的变量自身的信息和其他变量的信息，建立机器学习模型，基于该模型对缺失值变量的预测结果进行填补。

4. 无论是Pearson相关系数还是Spearman相关系数，其值均在[-1,1]之间分布。当值为0时，两个变量不相关；当其值为1或-1时，表示两个变量完全正相关或完全负相关。相关系数的绝对值越大，则说明两个变量的相关性越强。运用相关系数进行变量选择的步骤是，分别计算每个变量与目标变量的相关系数，保留与目标变量相关系数大的变量。

5.缩放效果不同，因为二者采取缩放方式不同。Min-Max缩放将原数据的所有值整体缩放至[-1,1]。Max-ABS缩放是将原值大于0的数据缩放到(0,1]，将原值小于0的数据缩放到[-1,0)，原值等于0的数据缩放后还为0。

6. 数据离散化是指将连续型变量的每个取值映射到根据客观或主观标准事先确定好的一系列分组或分类中，从而得到定性变量的数据预处理方法。

7. 低频分类数据通常呈现出类别众多，并且很多类别仅有几个甚至一个样本的情况。低频分类数据形成原因大致可以分为两种：第一，真实的分类结果中确实存在低频分类。第二，采集数据时格式不规范或错误，从而形成一些频数极小（如仅有一个样本）的类别。过多的低频分类会严重影响建模的效率。

8. 等宽法适用于对数据分布较为均匀的连续型变量进行离散化，根据连续型变量的取值范围，建立若干个宽度相等且首尾相连的区间，将连续型变量的每个值映射到相应的区间，并以区间名称作为新的离散型变量的值。

（四）程序填空题（每空2分，共20分）

1. value\_counts()
2. range loc
3. scale
4. MinMaxScaler fit\_transform
5. box
6. mean
7. qcut
8. astype(int)
9. make\_count[x]<100 x
10. replace
11. MaxAbsScaler fit\_transform
12. Pearsonr
13. labels=["new", "used", "old", "worn"]

14. data\_mean data\_std

15. value\_counts

1. isna()
2. drop axis=1
3. MinMaxScaler feature\_range
4. roc\_auc\_score predict\_proba
5. cut
6. std
7. gt
8. fillna mean()
9. str.slice
10. RobustScaler fit\_transform
11. spearmanr
12. ge(3000) le(10000)
13. value\_counts()
14. f\_oneway

（五）程序理解题（每题10分）

1. 使用StandardScaler模块对数据集mydata中的所有变量进行标准化，得到新数据集new\_data。调用该模块的scale\_参数和mean\_参数，通过print输出对mydata中变量进行标准化时所依据的标准差和均值。

2. 对data数据划分训练集和测试集，测试集比例为0.3.

3. 从数据文件cardata.csv中读取数据，保存在数据框car\_data中，使用car\_data中变量v1的非缺失部分的中位数填补变量v1的缺失部分，并把填补好的结果保存在数据框v1\_fill中。

4. 将data数据深复制给data\_1。

5. 构造一个数组存储时间，第一个时间是2020年1月20日，第二个时间是2020年2月20日，将日期时间从字符串状态转换为标准时间日期格式数据，时间格式为年-月-日。

6. 建立pearson序列，使用for循环，调用scipy.stats库中的pearsonr()函数，分别计算train\_x中的每个变量与train\_y的相关系数，并存入pearson序列中。判断pearson序列中符合“绝对值大于0.5”这一条件，得到逻辑值序列var\_cor。

7. 使用scikit-learn库中的GradientBoostingClassifier()函数，基于训练集train\_x和train\_y数据，建立GBDT模型model\_all，设定随机种子random\_state为0。提取model\_all的feature\_importances属性并保存为feature\_imp，它度量的是变量在GBDT模型所包含的所有决策树上的平均重要性。然后根据feature\_imp的数值，提取了平均重要性得分较大的前8个变量，其变量名保存在var\_tree中。第一句print输出所有变量名称及其重要性，按降序排序。第二句print输出排名前8的变量名称。

8. 将变量price进行等宽分箱，设定组数为5，将原变量和等宽分箱结果合并进一个数据框，输出前20位等宽分箱结果，输出等宽分箱频数分布。