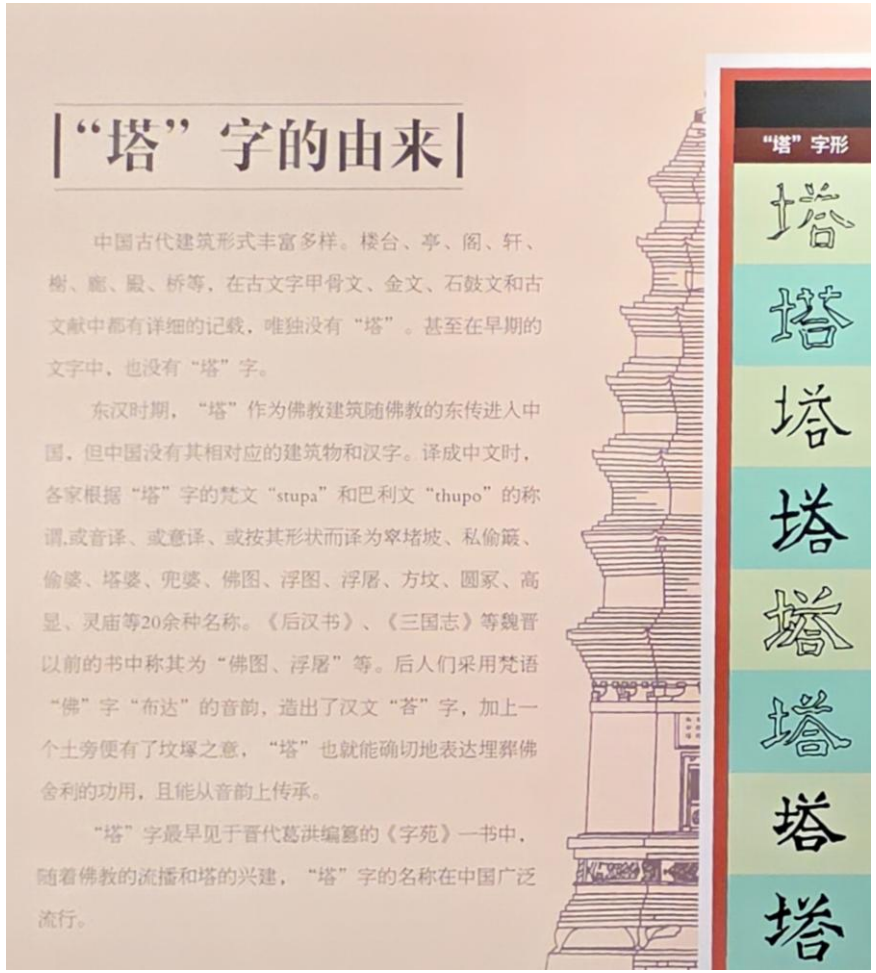
3



数据预处理：缺失值处理 (Missing Value Handling)

授课教师：[李阳源]

周末话动：许昌文峰塔



“塔”字的由来

中国古代建筑形式丰富多样。楼台、亭、阁、轩、榭、廊、殿、桥等，在古文字甲骨文、金文、石鼓文和古文献中都有详细的记载，唯独没有“塔”。甚至在早期的文字中，也没有“塔”字。

东汉时期，“塔”作为佛教建筑随佛教的东传进入中国，但中国没有其相对应的建筑物和汉字。译成中文时，各家根据“塔”字的梵文“stupa”和巴利文“thupo”的称谓，或音译、或意译、或按其形状而译为窣堵坡、私偷簸、偷婆、塔婆、兜婆、佛图、浮图、浮屠、方坟、圆冢、高显、灵庙等20余种名称。《后汉书》、《三国志》等魏晋以前的书中称其为“佛图、浮屠”等。后人们采用梵语“佛”字“布达”的音韵，造出了汉文“苻”字，加上一个土旁便有了坟冢之意，“塔”也就能确切地表达埋葬佛舍利的功用，且能从音韵上传承。

“塔”字最早见于晋代葛洪编纂的《字苑》一书中，随着佛教的流播和塔的兴建，“塔”字的名称在中国广泛流行。

“塔”字形

塔 塔 塔 塔 塔 塔 塔 塔

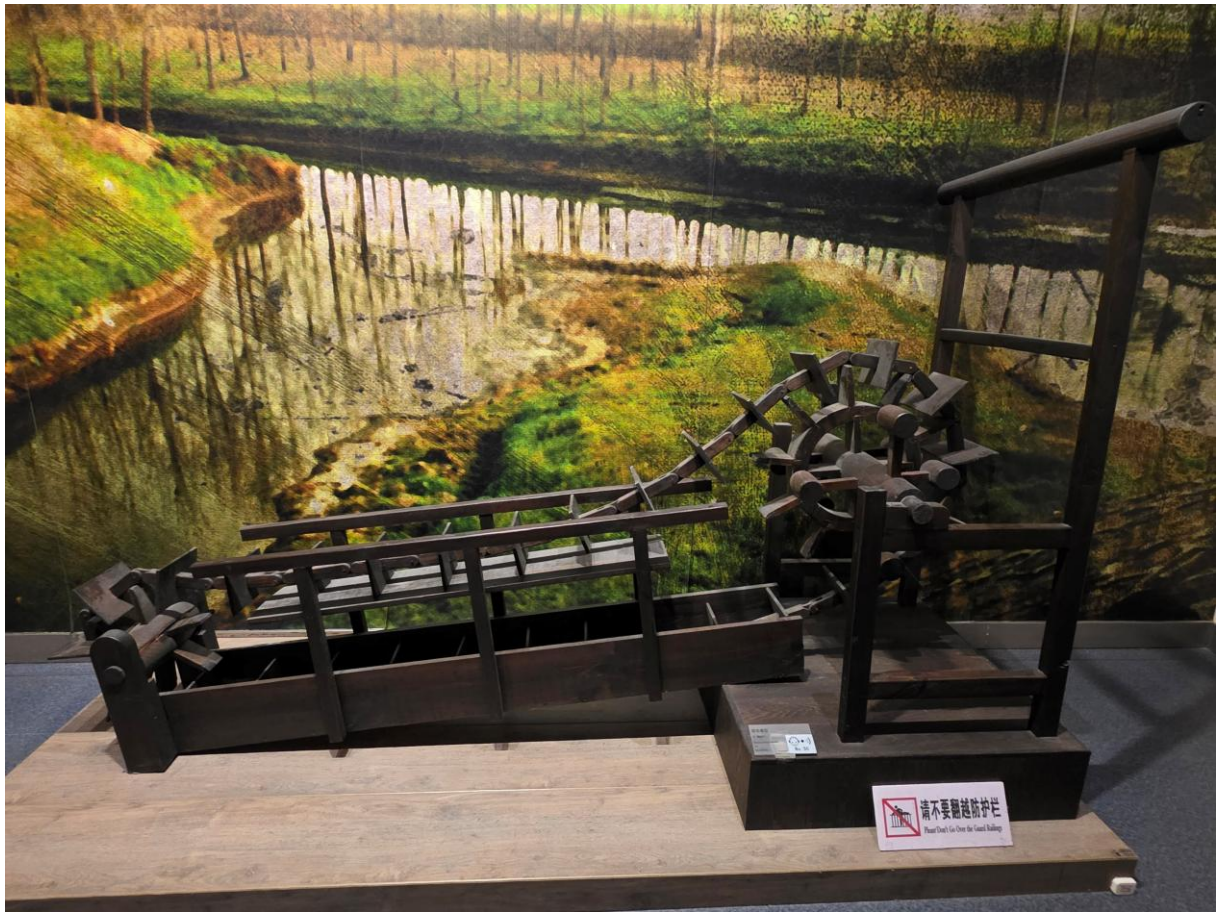


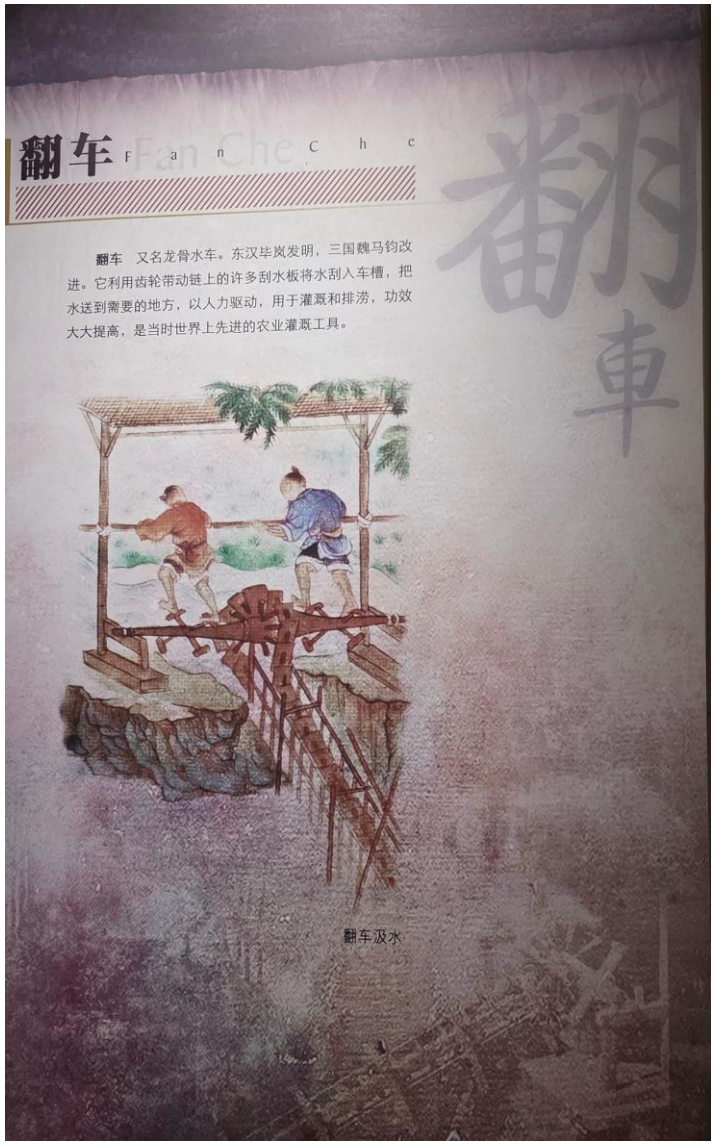
周末话动：曹丞相府访问





周末话动：许昌博物馆


请不要翻越防护栏


翻车 Fan Che
翻车 又名龙骨水车。东汉毕岚发明，三国魏马钧改进。它利用齿轮带动链上的许多刮水板将水刮入车槽，把水送到需要的地方，以人力驱动，用于灌溉和排涝，功效大大提高，是当时世界上先进的农业灌溉工具。
翻車
翻车汲水



数据预处理技术语



- Missing Value: 缺失值(Quēshī zhì)
- MCAR (Missing Completely At Random): 完全随机缺失(Wánquán suíjī quēshī)
- MAR (Missing At Random): 随机缺失(Suíjī quēshī)
- MNAR (Missing Not At Random): 非随机缺失(Fēi suíjī quēshī)
- Listwise Deletion: 列表删除法(Lièbiǎo shānchú fǎ)
- Pairwise Deletion: 成对删除法(Chéngduì shānchú fǎ)
- Imputation: 替换(Tìhuàn)
- Mean Imputation: 均值替换(Jūnzhì tìhuàn)

数据预处理技术语



- Median Imputation: 中位数替换(Zhōngwèishù tìhuàn)
- Mode Imputation: 众数替换(Zhòngshù tìhuàn)
- Forward Fill: 前值填充(Qián zhí tiánchōng)
- Backward Fill: 后值填充(Hòu zhí tiánchōng)
- Regression Imputation: 回归替换(Huíguī tìhuàn)
- Multiple Imputation: 多重替换(Duō zhòng tìhuàn)
- KNN Imputation: KNN替换法(KNN tìhuàn fǎ)
- Missing Rate: 缺失率(Quēshī lǜ)

数据预处理技术语



- Correlation: 相关性(Xiāngguān xìng)
- Conditional Randomness: 条件随机性(Tiáojiàn suíjī xìng)
- Statistical Inference: 统计推断(Tǒngjì tuīduàn)
- Data Distribution: 数据分布(Shùjù fēnbù)
- Sensitivity Analysis: 敏感性分析(Mǐngǎn xìng fēnxī)
- Data Quality: 数据质量(Shùjù zhìliàng)
- Domain Knowledge: 领域知识(Lǐngyù zhīshí)
- Bias: 偏差(Piānchà)

3



数据预处理：缺失值处理 (Missing Value Handling)

授课教师：[李阳源]

SL02: 今日学习目标



- 理解缺失值(Missing Value)的概念
- 掌握缺失值产生的原因
- 区分缺失机制的类型 (MCAR, MAR, MNAR)
  (**Classifying Types of Missing Data Mechanisms**)
- 学习缺失值处理的各种方法
- 认识数据分析中缺失值处理的重要性

SL03:什么是缺失值？



- 数据集中没有值的状态（空白、NULL、NaN等）
- 在收集、输入、传输过程中产生的信息缺失
- 直接影响数据质量的因素
- 大多数分析算法无法处理缺失值

(Most analytical algorithms cannot handle missing values)



空白 (Blank) 与 NULL 的区别详解

1. 核心概念对比

分类	空白 (Blank/空值)	NULL (空值/缺失值)
定义	"有存储空间但无内容"	"无任何值存在" (未定义状态)
表现形式	空字符串 `""`、空格 `" "`	`NULL` (SQL)、`None` (Python)、`NA` (R)
检测方法	`字段 == ""`	`字段 IS NULL` (SQL)、`pd.isna()` (Python)

SL04:缺失值产生的原因



- 数据收集时受访者不回应
- 测量设备故障(传感器, 温度计等)
- 数据输入过程中的人为错误(打字时漏掉一些数据)
- 系统错误或网络问题(有时候不是人的问题，而是技术的问题)
- 数据整合过程中的不一致

  (当我们合并来自不同来源的数据时，可能会发现某些地方信息不匹配)
- 数据转换过程中的错误(一个格式转换, 另一个格式时 丢失一些信息)

SL05:实际生活中的缺失值案例



- 问卷调查：个人信息题目无回应
- 医疗数据：未接受特定检查的患者(核磁共振检查)
- 传感器数据：设备临时故障导致的缺失
- 销售数据：特定时期销售记录缺失(某一天系统崩溃了)
- 网络日志：因网络问题导致的连接记录丢失

SL06:缺失值的影响



- 分析结果的扭曲和偏差(distortion and bias)
- 模型性能下降 (model)
- 统计推断的准确度降低
- 数据集规模缩小(行删除时)(dataset)
- 对整体数据质量信任度下降 (reliability) (The overall trust in data quality has declined.)
- 时间和资源浪费

SL07:缺失机制：MCAR



完全随机缺失 (Missing Completely At Random)

- 缺失概率与其他变量或缺失值本身**无关**
- 缺失**完全随机发生**
- 最理想的缺失形式（偏差最小化）
- 验证：Little的MCAR检验

SL08:MCAR示例



- 问卷的部分页面因印刷错误而缺失
- 随机选择的样本中特定测量被省略
- 实验中测量设备的偶然故障
- 数据输入过程中的随机错误

SL09:缺失机制：MAR



随**机缺失 (Missing At Random)**

- 缺失概率**依赖于观察到的其他变量**
- 与缺失值本身的特性无关
- 条件随机性 (conditional randomness)
- 可通过观察数据预测缺失模式

SL10:MAR示例



- 高收入群体倾向于不回答收入问题
- 特定年龄段对健康相关问题的响应率降低
- 特定地区的传感器数据更频繁缺失
- 特定产品类别的评价更频繁缺失

SL11: 缺失机制：MNAR



非随机缺失 (Missing Not At Random)

- 缺失概率**依赖于缺失值本身**
- 缺失本身包含重要信息
- 最复杂且难以处理的形式
- 缺失模式在分析中具有重要意义

SL12: MNAR示例



- 成绩低的学生拒绝公开成绩
- 有特定疾病的患者不回答相关问题
- 体重较重的人不回答体重问题
- 满意度低的客户不回应调查

SL13: 缺失机制(mechanism)比较



特性	MCAR	MAR	MNAR
缺失模式	完全随机	依赖观察变量	依赖缺失值本身
偏差风险	低	中	高
检测难度	容易	中等	困难
处理方法	可删除/替换	推荐多变量替换	需要建模
统计验证	可能	部分可能	困难

Multiple imputation

实际实列



1. 您的性别是？

☐ 男 (nán)

☐ 女 (nǚ)

☐ 其他 / 不愿透露 (qítā / bùyuàn tòulù)

2. 您的年龄是？（请填写具体数字） _____ 岁

实际实列



3. 您的月平均收入是多少？

- ☐ 3000元以下
- ☐ 3000–5000元
- ☐ 5001–8000元
- ☐ 8001–12000元
- ☐ 12000元以上
- ☐ 不愿透露

实际实列



4. 您的职业是?

- ☐ 学生
- ☐ 上班族
- ☐ 个体经营
- ☐ 退休
- ☐ 无业
- ☐ 其他: _____

实际实列



5. 您每晚平均睡眠时间是多少？

- ☐ 少于 4 小时
- ☐ 4–6 小时
- ☐ 6–8 小时
- ☐ 超过 8 小时

实际实列



6. 您是否经常感到压力？

- ☐ 是
- ☐ 否
- ☐ 不确定

实际实列



7. 过去一年中，您是否接受过心理咨询？

- □ 是
- □ 否
- □ 不愿回答

调查问卷中的三个变量



变量	完全随机缺失 (MCAR)	随机缺失 (MAR)	非随机缺失 (MNAR)
收入	因为机器问题丢失	老年人更不愿填写	收入太低或太高者拒绝填写
年龄	调查员忘记输入	女性更常省略年龄	年龄大者因敏感而不写
性别	表格印刷错误	年轻人更容易跳过	性别认同模糊者回避回答

SL14: 缺失值处理方法1：删除方法



1. 列表删除法 (Listwise Deletion)

- 删除**包含缺失值的整行**
- 优点：实现简单，获得完整数据集
- 缺点：数据损失，样本量减少

2. 成对删除法 (Pairwise Deletion)

- 仅删除**特定分析中变量有缺失的案例**
- 优点：增加数据利用率
- 缺点：每次分析样本量不同，可能产生协方差矩阵问题

SL15: 缺失值处理方法2:替换方法(简单)



1. 均值/中位数/众数替换(average/median/mode)

- 用该变量的统计量替换缺失值
- 优点：实现简单，无数据损失
- 缺点：分布扭曲，忽略变量间关系

2. 前/后值替换 (Forward/Backward Fill)

- 主要用于时间序列数据(time series data)
- 优点：保持时间连续性
- 缺点：长期缺失时不适合

SL16:缺失值处理方法3:替换方法(高级)



1. 回归替换 (Regression Imputation)

- 使用其他变量预测缺失值
- 优点：考虑变量间关系
- 缺点：过拟合可能性(overfitting risk)，低估不确定性(underestimated uncertainty)

2. 多重替换 (Multiple Imputation)

- 用多个可能值替换，生成多个数据集
- 优点：反映不确定性，减少偏差
- 缺点：计算复杂性，解释困难

Regression Imputation



Original Data

X1	X2	Y
10	5	80
12	7	95
15	6	??
11	6	85
14	7	92

X1	X2	Y	
10	5	80	
12	7	95	
15	6	84.5	회귀로 대체
11	6	85	
14	7	92	

Regression Model

$$Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + \varepsilon$$

使用无缺失的观测值训练模型

→ 示例：$Y = 20 + 3.5 * X_1 + 2 * X_2$

[4] 预测缺失值

$X_1 = 15, X_2 = 6$

→ $\hat{Y} = 20 + 3.5 * 15 + 2 * 6 = 20 + 52.5 + 12 =$ **84.5**

SL17: 缺失值处理方法4:基于算法(Algorithm)



1. KNN替换 (K-Nearest Neighbors)

- 基于相似样本值进行替换
- 优点：保留局部模式
- 缺点：计算成本，对特征选择敏感

2. 机器学习基础替换(Machine Learning)

- 利用随机森林、深度学习等(Random Forest, Deep Learning)
- 优点：可捕捉复杂模式(complex pattern)
- 缺点：模型构建复杂性，解释困难

SL18: 缺失值处理时的考虑因素



- 缺失机制(MCAR, MAR, MNAR)的判断
- 缺失率(Missing Rate)的确认
- 变量重要性评估
- 数据规模和分析目的
- 领域知识的应用
- 多种处理方法结果的比较

SL19: 缺失值处理的阶段性方法



- 了解缺失值模式并可视化
- 估计缺失机制
- 根据缺失率决定变量/案例处理策略
- 选择并应用适当的替换方法
- 评估替换结果并进行敏感性分析
- 必要时比较多种方法

SL20: 理解度确认测验



1. MCAR, MAR, MNAR的主要区别是什么？
2. 解释列表删除法(Listwise Deletion)的优缺点。
3. 均值替换方法的缺点是什么？
4. 当一个变量有20%的缺失值时，什么处理方法合适？
5. 多重替换(Multiple Imputation)优于单一替换的原因？

SL21: 总结与下次课程预告



今天学习内容

- 缺失值的概念和产生原因
- 缺失机制(MCAR, MAR, MNAR)的特性
- 各种缺失值处理方法及优缺点
- 缺失值处理的考虑因素

下节课预告

- 异常值(Outlier)处理：概念和检测方法
- 异常值处理技术和应用案例

问答时间！

数据预处理：缺失值处理

Missing Value Handling

理解度测验（完整版） · 20题四选一

授课教师：李阳源 · 2026.03.02 · 第03讲 · 含处理方法全部内容

1. 以下哪项最准确地描述了「缺失值（Missing Value）」的本质?

A. 数值为零（0）的数据

B. 数据集中没有值的状态，表现为空白、NULL、NaN等形式，直接影响数据质量

C. 超出正常范围的极端数值

D. 格式不符合规范的错误数据

2. 在Python中，检测NULL（缺失值）和检测「空白（Blank）」的方法分别是什么?

A. 两者都用 pd.isna() 检测

B. NULL用 IS NULL（SQL），空白用 pd.isna()（Python）

C. NULL用 字段 IS NULL（SQL）或 pd.isna()（Python），空白用 字段 == ""

D. 两者都用 字段 == "" 检测

3. 下列哪项描述的缺失值产生原因属于「技术层面」而非「人为层面」？

A. 受访者故意不回答敏感问题

B. 数据输入员打字时漏掉部分数据

C. 系统错误或网络问题导致数据未能传输保存

D. 数据整合时来自不同来源的信息不匹配

4. 在调查问卷案例中，「年龄」变量因「调查员忘记输入」而缺失，这属于哪种缺失机制?

A. MNAR（非随机缺失），因为老年人更敏感

B. MCAR（完全随机缺失），因为遗忘与年龄本身无关，属于纯粹随机错误

C. MAR（随机缺失），因为与调查员的习惯有关

D. 无法判断缺失机制

5. 在调查问卷案例中，「收入太低或太高的人拒绝填写收入」属于哪种缺失机制？为什么？

A. MCAR；因为是随机发生的

B. MAR；因为与年龄等其他观察变量有关

C. MNAR；因为是否缺失直接取决于收入数值本身（太低或太高），缺失依赖于缺失值本身

D. MCAR；因为机器故障导致数据丢失

6. 根据课程中调查问卷三变量案例，「女性更常省略年龄」属于哪种缺失机制?

A. MCAR；因为性别和年龄无关

B. MAR；缺失概率依赖于可观察变量「性别」，而非年龄本身

C. MNAR；因为年龄越大越不想填

D. MCAR；因为是随机遗漏

7. 关于MCAR、MAR、MNAR三种缺失机制，以下比较哪项是正确的?

A. MCAR偏差风险高，MNAR偏差风险低

B. MCAR检测最容易（可用Little检验），MAR检测中等，MNAR检测最困难

C. 三者的统计验证难度相同

D. MAR是最理想的缺失形式，偏差最小

8. 列表删除法（Listwise Deletion）的最大缺点是什么？在什么情况下这个缺点最为严重？

A. 计算复杂，难以实现

B. 删除整行导致数据损失和样本量减少；当缺失比例较大时，会严重影响分析结果的代表性

C. 每次分析使用的样本量不同，可能产生协方差矩阵问题

D. 无法处理时间序列数据

9. 成对删除法（Pairwise Deletion）与列表删除法（Listwise Deletion）的核心区别是什么？

A. 成对删除法删除整行，列表删除法只删除缺失列

B. 成对删除法仅删除特定分析中该变量有缺失的案例，数据利用率更高，但每次分析样本量可能不同

C. 两者完全相同，只是名称不同

D. 成对删除法只能用于时间序列数据

10. 使用「均值/中位数/众数替换」缺失值的主要缺点是什么？

A. 实现复杂，需要大量编程

B. 会造成数据损失，样本量减少

C. 会扭曲数据分布，并忽略变量间关系，可能导致方差被低估

D. 只能用于非数值型数据

11.「前/后值替换（Forward/Backward Fill）」最适合哪类数据？为什么？

A. 关系型数据库中的结构化数据，因为可以用SQL查询

B. 时间序列数据，因为它保持数据在时间上的连续性

C. 非结构化文本数据，因为文本具有上下文关系

D. 高维度特征数据，因为可以降低计算复杂度

12. 回归替换（Regression Imputation）的工作原理是什么？它有什么潜在风险？

A. 用均值替换缺失值；风险是分布扭曲

B. 用相邻时间点的值填充；风险是长期缺失时不适用

C. 使用其他变量建立回归模型来预测缺失值；风险是过拟合（overfitting）和低估不确定性

D. 生成多个替换值并建立多个数据集；风险是计算复杂性

13. 根据课程的回归替换示例，已知模型为 $Y = 20 + 3.5\times X_1 + 2\times X_2$。当 $X_1=15$、$X_2=6$ 时，预测替换值$\hat{Y}$是多少？

A. 76.5

B. 80.0

C. 84.5

D. 90.0

14. 多重替换（Multiple Imputation）与回归替换（Regression Imputation）相比，最核心的优势是什么？

A. 多重替换计算更简单，实现更容易

B. 多重替换生成多个可能值，能反映不确定性并减少偏差，而回归替换只给出单一预测值

C. 多重替换不需要使用其他变量

D. 多重替换只能用于时间序列数据

15. KNN替换（K-Nearest Neighbors Imputation）的基本原理是什么?

A. 用变量的均值替换缺失值

B. 用随机森林模型预测缺失值

C. 基于与缺失样本最相似的K个样本的值进行替换，保留局部数据模式

D. 通过时间序列的前后值进行插值

16. 相较于KNN替换，使用随机森林等机器学习方法处理缺失值的主要优势和劣势分别是什么?

A. 优势：计算简单；劣势：无法处理复杂模式

B. 优势：可捕捉复杂模式（complex pattern ）；劣势：模型构建复杂、解释困难

C. 优势：保留局部数据模式；劣势：对特征选择敏感

D. 优势：保持时间连续性；劣势：长期缺失时不适合

17. 某数据集中，「体重」字段缺失比例为30%，且体重较重的人更倾向于不填写。分析师应选择哪种处理策略最为合适?

A. 直接使用列表删除法删除含缺失值的行，简单有效

B. 用均值替换，保持数据完整

C. 判断为MNAR，需建立专门模型处理，不能简单删除或填均值，否则会引入严重偏差

D. 使用前值替换，保持时间连续性

18. 对于时间序列传感器数据中偶发的短暂故障导致的缺失值，哪种处理方法最为合适？为什么？

A. 列表删除法；因为删除简单

B. 均值替换；因为均值最能代表典型值

C. 前/后值替换；因为传感器数据具有时间连续性，相邻时间点的值最能反映真实状态

D. 多重替换；因为能反映不确定性

19. 以下对四类缺失值处理方法（删除法、简单替换、高级替换、算法替换）的描述，哪一项是错误的?

A. 列表删除法的优点是实现简单，能获得完整数据集

B. 均值/中位数/众数替换的优点是实现简单且无数据损失

C. 回归替换完全没有偏差风险，是最完美的方法

D. KNN替换能保留局部模式，但计算成本较高且对特征选择敏感

20. 综合本课所有内容，以下哪种说法最能体现处理缺失值的正确决策流程？

A. 发现缺失值后，首选删除法，简单快捷

B. 任何情况下都用多重替换，因为它最准确

C. 首先判断缺失机制类型（MCAR/MAR/MNAR），再根据数据特点（是否为时间序列、缺失比例、变量关系等）选择最合适的处理方法

D. 缺失比例低于10%时直接忽略，无需处理

PL01: 缺失值处理实践 (Missing Value Handling Practice)



PL02:今日实践目标



- 学习Python中识别缺失值的方法
- 利用Pandas探索和可视化缺失值
- 应用各种缺失值处理技术
- 实践缺失值删除方法 (.dropna())
- 实践缺失值替换方法 (.fillna())
- 应用于实际数据集并比较结果

PL03: 安装并导入必要的库



```
# 文件名: missing_value_practice.py

# 导入基本库
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# 可视化设置
plt.style.use('seaborn')
sns.set(font_scale=1.2)

# 忽略警告信息设置
import warnings
warnings.filterwarnings('ignore')

print("库导入完成！")
```

PL04: 创建样本数据



```
# 文件名: missing_value_practice.py

# 创建包含缺失值的样本数据
def create_sample_data():
    # 使用字典创建数据
    data = {
        '姓名': ['金哲秀', '李英姬', '朴敏洙', '崔真英', np.nan, '郑秀真'],
        '年龄': [25, 30, np.nan, 45, 33, 22],
        '性别': ['男', '女', '男', np.nan, '女', '男'],
        '收入': [250, np.nan, 350, 450, 150, np.nan],
        '支出': [200, 220, np.nan, 400, 130, 250]
    }

    # 创建数据框
    df = pd.DataFrame(data)
    return df

# 创建并输出样本数据
df_sample = create_sample_data()
print("₩n样本数据:")
print(df_sample)
```

PL05: 检查缺失值



```
# 文件名: missing_value_practice.py

# 检查缺失值方法1: isnull()
print("₩n1. 使用isnull()函数:")
print(df_sample.isnull())

# 检查缺失值方法2: isna()
print("₩n2. 使用isna()函数(与isnull相同):")
print(df_sample.isna())

# 检查非缺失值: notna()
print("₩n3. 使用notna()函数:")
print(df_sample.notna())
```

PL17: 实践: 亲自动手



```
# 导入必要的库
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# 忽略警告
import warnings
warnings.filterwarnings('ignore')

plt.rcParams['font.family'] = 'SimHei' # 黑体
plt.rcParams['axes.unicode_minus'] = False
```

PL17: 实践: 亲自动手



据创建函数

```
def create_practice_data():
        np.random.seed(42) # 设置随机种子，确保结果可重复
        n = 100
        # 创建学生数据
        data = {
        '学号': range(1001, 1001 + n),
        '姓名': ['学生' + str(i) for i in range(1, n + 1)],
        '数学': np.random.randint(0, 101, n),
        '英语': np.random.randint(0, 101, n),
        '科学': np.random.randint(0, 101, n),
        '出勤率': np.random.randint(50, 101, n) / 100
        }
```

PL17: 实践: 亲自动手



```
# 创建数据框

df = pd.DataFrame(data)

# 随机创建缺失值 (约15%)

for col in ['数学', '英语', '科学', '出勤率']:

mask = np.random.random(n) < 0.15

df.loc[mask, col] = np.nan

return df
```

PL17: 实践: 亲自动手



```
# 创建练习数据集

students_df = create_practice_data()

print("创建的学生数据:")

print(students_df.head())

print("\n缺失值统计:")

print(students_df.isnull().sum())
```

PL18: 实践任务: 处理缺失值



你的任务:

1.可视化缺失值（热图、柱状图等）
2.为每列选择合适的缺失值处理方法:
   - 数学、英语、科学: 均值、中位数或其他方法
   - 出勤率: 均值 的替换方法
3.处理缺失值后计算各科目的平均分和标准差
4.比较并可视化处理前后的结果
5.讨论处理方法对分析结果的影响

PL19: 实践任务参考答案



Titanic





2026-03-10

实践题：探索 Titanic 数据中的缺失值



今天，我们将使用著名的泰坦尼克数据集来确认缺失值并尝试各种处理方法。



准备材料:

- Python安装环境（PyCharm IDE）
- pandas, numpy, matplotlib, seaborn库
- 提供的`titanic.csv`文件

步骤01: 数据加载与基本探索



1) 加载必要的库

2) 加载'titanic.csv'文件并创建数据框

3) 确认数据的大小（行数和列数）

4) 打印前5行来了解数据结构

5) 使用df.info()函数检查数据类型和缺失值情况

6) 使用df.describe()函数查看数值型数据的统计概要

步骤02:缺失值确认和可视化



1) 确认每列的缺失值数量并说明哪些列存在缺失值

2) 计算相对于整体数据的缺失值比例(%)

3) 使用条形图(bar plot)可视化缺失值

4) 使用热图(heatmap)表示缺失值

5) 思考缺失值较多的列应该如何处理

步骤03: 删除缺失值函数(基础)



1) 删除所有包含缺失值的行并确认剩余数据的大小

2) 只删除Age列有缺失值的行并确认剩余数据的大小

3) 比较这两种方法的差异，并讨论在什么情况下各自适用

步骤04: 使用平均值替换缺失值



1) 计算Age列的平均值

2) 使用计算出的平均值替换Age列的缺失值

3) 替换后检查Age列的缺失值数量

4) 思考使用平均值替换的优缺点

步骤05: 使用分组平均(group average)值替换缺失值



1) 计算按性别(Sex)分组的平均年龄

2) 根据性别用相应组的平均值替换Age缺失值

3) 计算按客舱等级(Pclass)和性别(Sex)分组的平均年龄

4) 根据客舱等级和性别用相应组的平均值替换Age缺失值

5) 比较这两种方法的结果，讨论哪种方法可能提供更准确的替换值

步骤06: 使用众数和中位数替换缺失值



1) 找出Embarked列的众数(出现频率最高的值)

2) 使用众数替换Embarked列的缺失值

3) 计算Age列的中位数

4) 使用中位数替换Age列的缺失值

5) 讨论每种方法适用的情况

步骤07: 处理Cabin列



1) 确认Cabin列的缺失值比例

2) 提出处理Cabin列大量缺失值的可能方法

3) 创建一个新的二元变量(Has_Cabin)表示是否有Cabin信息

4) 讨论创建的变量可能具有的意义

步骤08: 综合应用缺失值处理方法



1) 综合应用学到的各种缺失值处理方法:

- Age: 按客舱等级和性别分组的平均值替换

- Embarked: 使用众数替换

- Cabin: 创建表示信息有无的新变量

2) 检查最终数据集的缺失值状态

3) 讨论缺失值处理对数据分析或机器学习模型可能产生的影响

步骤09: 缺失值处理技术比较与评估



1) 总结目前应用的各种缺失值处理方法(删除、平均值、中位数、分组平均值、众数)的优缺点

2) 解释每种方法适用的情况

3) 讨论选择缺失值处理方法时需要考虑的因素

PL20: 理解度确认测验



- Pandas中检查缺失值的函数是?

  a) isEmpty() b) isNull() c) isnull() d) isNone()

- 删除所有包含缺失值的行的Pandas代码是?

  a) df.dropna() b) df.drop_null() c) df.remove_na() d) df.delete_null()

- 用中位数替换数值变量缺失值的代码是?

  a) df['col'].fillna(mean) b) df['col'].fillna(df['col'].median()) c) df['col'].replace(np.nan, median) d) df['col'] = median

PL20: 理解度确认测验



- KNN替换法的优点是?

  a) 实现简单 b) 计算速度快 c) 考虑变量间关系 d) 始终准确

- 选择缺失值处理方法时最重要的考虑因素是?

  a) 实现难度 b) 缺失机制(MCAR, MAR, MNAR) c) 数据大小 d) 计算速度

PL21: 总结与下次课程预告



今天学习内容

- 使用Python和Pandas识别并可视化缺失值
- 各种缺失值删除方法 (dropna)
- 使用多种替换方法处理缺失值 (fillna)
- 在实际数据集上应用缺失值处理并比较结果

下节课预告

- 异常值(Outlier)处理实践: 异常值检测和处理方法
- 使用箱线图、Z-score、IQR方法识别异常值

问答时间!