第四次作业

姓名：柳絮源

2025 年 1 月 2 日

复习题

第一题

数据全生命周期管理分为以下阶段：

- **数据采集**：收集数据，确保准确性和合法性。
- **数据传输**：确保数据流动中的安全性和完整性。
- **数据存储**：采用分布式或加密存储，确保数据可访问性和安全性。
- **数据处理**：对数据进行查询、分析和可视化。
- **数据交换/共享**：确保数据在系统间共享的合规性。
- **数据销毁**：彻底删除数据，确保不可恢复。

第二题

一、数据采集的概念

数据采集（DAQ）是从传感器或设备中自动采集信号并进行分析的过程。

二、数据采集的方法

- **问卷调查**：成本低，信息全面。
- **查阅资料**：数据经过筛选，需验证真实性。
- **实地考查**：获取第一手数据，但耗时耗力。
- **实验**：数据准确，但成本高。
- **直接观察法**：通过观察或影像收集数据。
- **网络爬虫**：自动提取网页数据，适用于大规模收集。

第三题

相同点

- **数据安全**：均需保护数据的完整性、保密性和可用性。
- **管理方便**：均支持高效的数据访问和修改。

差异点

- **处理规模**：传统数据管理处理 MB 级数据，大数据管理处理 GB、TB 甚至 PB 级数据。
- **数据类型**：传统数据管理主要处理结构化数据，大数据管理处理多种类型数据。
- **存储技术**：传统数据管理使用关系型数据库，大数据管理使用分布式存储。

第四题

大数据的计算模式包括：

- **批量计算**：处理静态数据集，如 Hadoop。
- **流式计算**：实时处理数据流，如 Storm、Spark Streaming。
- **交互计算**：通过人机交互实时查看数据。
- **图计算**：处理图结构数据，如 Pregel、GraphX。

第五题

数据分析方法

- **对比分析**：找出数据差异和规律。
- **趋势分析**：预测数据未来走向。
- **聚类分析**：将相似数据归为一类。

数据分析模型

- **SWOT 模型**：分析企业优劣势及外部机会和威胁。
- **4P 营销理论**：产品、价格、渠道、促销的组合。
- **PEST 模型**：分析政治、经济、社会和技术因素。

第六题

数据可视化的原因：

- **直观理解数据**：将复杂数据转化为图表。
- **快速识别模式**：帮助发现数据中的趋势和异常。
- **增强沟通效果**：跨越语言和文化障碍。

践习题

第七题

```
import pandas as pd
import matplotlib.pyplot as plt

# 生成数据集
data = {
    'name': ['Tom', 'Jerry', 'Mickey', 'Donald', 'Dora'],
    'age': [25, 30, 28, 32, 26],
    'score': [85, 90, 88, 92, 87],
    'gender': ['M', 'M', 'M', 'M', 'F']
}
df = pd.DataFrame(data)

# 柱状图：不同性别人数统计
gender_counts = df['gender'].value_counts()
plt.bar(gender_counts.index, gender_counts.values)
plt.xlabel('Gender')
plt.ylabel('Count')
plt.title('Number of People by Gender')

# 显示柱状图
plt.show()

# 折线图：年龄与成绩关系
plt.plot(df['age'], df['score'])
plt.xlabel('Age')
plt.ylabel('Score')
plt.title('Relationship between Age and Score')

# 显示折线图
plt.show()

# 饼图：不同年龄段人数比例
age_bins = [20, 25, 30, 35]
age_labels = ['20-24', '25-29', '30-34']
df['age_group'] = pd.cut(df['age'], bins=age_bins, labels=age_labels, right=False)
age_group_counts = df['age_group'].value_counts()
plt.pie(age_group_counts.values, labels=age_group_counts.index, autopct='%1.1f%%')
plt.title('Proportion of People by Age Group')
```

```
# 显示饼图
plt.show()
```

第八题

```
import pandas as pd
import seaborn as sns

# 生成数据集
data = {
    'name': ['Tom', 'Jerry', 'Mickey', 'Donald', 'Dora'],
    'age': [25, 30, 28, 32, 26],
    'score': [85, 90, 88, 92, 87],
    'gender': ['M', 'M', 'M', 'M', 'F']
}

df = pd.DataFrame(data)

# 箱线图：成绩分布
sns.boxplot(data=df['score'])
sns.set_title('Distribution of Scores')

# 显示箱线图
sns.show()

# 散点图：年龄与成绩关系
sns.scatterplot(data=df, x='age', y='score')
sns.set_title('Relationship between Age and Score')

# 显示散点图
sns.show()

# 小提琴图：不同性别的成绩分布
sns.violinplot(data=df, x='gender', y='score')
sns.set_title('Distribution of Scores by Gender')

# 显示小提琴图
sns.show()
```