Data Analytics pipeline

阶段	主要任务	细节说明	可能的错误	示例
Gathering	- 收集数据(Surveys, sensors, APIs, databases) - 确定数据来源 - 确保数据完整性和准确性 - 使用适当工具或方法采集数据	- 使用传感器记录天气数据 - 从在线数据库中提取销售记录 - 发放调查问卷收集消费者反馈	- 数据遗漏或采集不足 - 数据采集设备或工具误差 - 不相关数据的采集 - 隐私或伦理问题	- 调查问卷中遗漏了关键问题 - 从多个传感器采集温度数据时，传感器校准不当 - 使用错误的API导致数据不准确
Processing	- 数据清理（去除重复值、异常值 ） - 数据转换（格式化、归一化） - 合并或重组数据以适应分析需求	- 将日期格式统一为YYYY-MM-DD - 用均值填充缺失值 - 将分类变量转换为数值变量	- 数据丢失或错误清理 - 转换过程中引入偏差 - 使用不一致的格式 - 数据注释错误	- 删除有用的异常值（如销售高峰） - 不同数据来源的时间格式不匹配 - 在分类数据时错误分配标签
Analysing	- 探索数据模式和趋势 - 使用统计或机器学习方法得出结论 - 验证假设	- 使用t检验分析组间差异 - 绘制散点图观察变量相关性 - 使用分类模型预测客户流失	- 假设错误或模型不适用 - 忽略关键变量 - 过度拟合或欠拟合 - 计算错误	- 误用线性回归分析非线性数据 - 数据中遗漏关键变量如季节性因素 - 模型过于复杂以致无法解释分析结果
Presenting	- 可视化分析结果 - 准备报告或演示 - 确保受众理解关键见解	- 创建柱状图显示不同产品销售情况 - 在PPT中总结营销策略的效果 - 向团队展示机器学习模型的预测准确率	- 图表不清晰或误导 - 信息过于复杂难以理解 - 忽略目标受众的需求	- 使用误导性比例的饼图 - 为非技术观众准备了过于复杂的回归模型公式 - 错误解释数据模式
Preserving	- 保存数据及分析结果 - 记录元数据和分析过程 - 选择适当的存储和备份机制	- 保存数据到云存储并记录数据来源 - 使用README文件解释数据集格式和内容 - 定期检查备份是否损坏	- 数据存储格式不兼容 - 文件损坏或丢失 - 未记录关键步骤，导致分析无法复现	- 未正确保存分析代码和模型参数 - 数据备份过期或丢失 - 数据存储格式过于老旧导致难以读取

ask your client questions

1. 项目目标与预期

核心关注点：明确客户想要解决的问题和达到的目标。

- 示例问题：
    - 这个系统/报告的最终目标是什么？（比如政策评估、公众教育、资源分配等）
    - 客户希望从数据中得出哪些具体结论或见解？
    - 有没有明确的衡量指标（**KPIs**）来评估项目的成功？

---

2. 数据范围与来源

核心关注点：清楚需要处理的数据类型、数据来源和覆盖范围。

- 示例问题：
    - 您是否有现成的数据源，还是需要我们额外收集？
    - 数据需要覆盖哪些时间段？比如，是否有特定年份或时间点需要重点分析？
    - 是否有访问某些数据源的权限问题？

---

3. 目标用户与使用场景

核心关注点：了解谁会用系统/报告以及如何使用，以确保结果符合用户需求。

- 示例问题：
    - 最终用户是谁？（政府部门、普通公众、企业决策者等）
    - 用户希望通过系统/报告完成哪些具体任务？
    - 数据查询或分析结果需要以什么样的形式呈现？（如可视化图表、可下载报告等）

---

4. 技术与交付需求

核心关注点：明确项目的技术限制、交付周期以及客户的特定需求。

- 示例问题：
    - 有没有对平台或技术工具的偏好？（如必须使用 **Excel**、**Tableau** 或特定数据库）
    - 数据系统需要在线实时访问，还是离线报告生成即可？
    - 是否有明确的交付时间表或项目阶段要求？

---

5. 数据质量与维护

核心关注点：了解客户对数据完整性、更新频率和长期维护的要求。

- 示例问题：
  - 数据质量是否有最低标准？（如数据缺失率、误差范围）
  - 数据需要多长时间更新一次？系统是否需要支持自动化更新？
  - 数据结果是否需要长期保存以供将来参考？

---

6. 隐私与安全要求

核心关注点：确保数据处理符合隐私法规，避免法律和伦理问题。

- 示例问题：
  - 数据是否包含敏感或个人信息？需要如何保护？
  - 是否需要遵循特定的数据合规性要求？（如 **GDPR**、**HIPAA** 等）
  - 客户对数据存储位置和访问权限有哪些具体要求？

---

7. 成本与预算

核心关注点：了解客户的预算范围和资源限制。

- 示例问题：
  - 客户希望项目在哪些方面节约成本？
  - 是否有资源或工具支持我们使用？（如已有的服务器、许可证）

8. 数据存储需求

数据的访问频率和实时性要求是什么？

- 这个问题帮助了解数据更新的频率和实时访问的需求。例如，销售数据是否需要实时更新？如果是，存储系统需要支持实时写入和查询。

数据的规模和增长速度是多少？

- 了解数据的数量以及随着时间推移的增长趋势，这有助于判断存储系统是否能够扩展，是否需要支持大规模存储和处理。

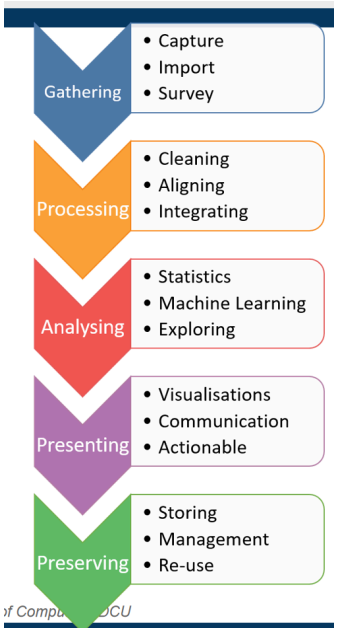
是否有数据备份和恢复的要求？

- 确保了解客户是否有严格的数据备份、恢复和保留策略，尤其是在出现系统故障时，需要如何保证数据的安全性和可恢复性。

design a system for a data collection task:

(i) List three (3) important questions you would ask your client.
这个系统/报告的最终目标是什么？
数据范围与来源?
最终用户是谁？
数据结果是否需要长期保存以供将来参考？数据需要多长时间更新一次？
数据是否包含敏感或个人信息？
(ii) Describe the data and/or specific file formats that you are likely to use in collecting and storing the data.
政府开放数据（如 pedestrian footfall 数据）
通常以 CSV、JSON 或 Excel 格式提供，便于批量处理和分析。
交通监控摄像头数据（如 vehicle count 数据）
可能为 CSV 或 SQL 数据库 格式；如果涉及视频，则可能使用 MP4 文件结合元数据文件（JSON/XML）。
消费者调查数据
CSV 或 SPSS 文件格式，因为调查数据需要统计分析工具的支持。
气象和经济数据
这些通常是时间序列数据，常见格式为 CSV 或 JSON，以便与分析工具集成。
系统设计与输出
数据结果和报告可能需要存储为 PDF、HTML 或 可交互图表文件（如 D3.js 支持的格式）。

分析关于在线教学的反馈



场景：分析**2020**年**DCU**学生关于在线教学的反馈
1. Gathering (数据收集**)**
为了分析学生对**DCU 2020**年在线教学的反馈，我们收集了问卷调查的数据。问卷包括了学生对课程质量、互动性、教师表现和技术支持的评分。数据通过**Google**表单或**SurveyMonkey**进行收集，保存为**Excel**或**CSV**文件。
2. Processing (数据处理**)**
在数据处理阶段，我们对原始调查数据进行了清理和标准化。使用**Excel**或**Python (Pandas)** 进行数据清理，删除缺失数据并统一评分标准。我们还会将一些开放式问题的回答进行分词和情感分析。
3. Analysis (数据分析**)**
在分析阶段，我们使用描述性统计分析（如均值、标准差）和推断统计（如**t**检验、卡方检验）来分析不同群体（如不同年级、不同学科）对在线教学的看法差异。使用****Python (SciPy, statsmodels)****进行统计检验，确定哪些因素对学生满意度产生了显著影响。
4. Presenting (数据展示**)**
我们使用图表和可视化工具将结果展示出来。常用的可视化方式包括条形图、折线图和热力图，这些可以帮助我们展示学生对各方面评分的分布情况。我们可能会使用**Tableau**或**Power BI**制作仪表板，展示调查结果的关键指标。
5. Preserving (数据保存**)**
分析和可视化结果会保存在云存储（如**Google Drive**、**Dropbox**）中，以便随时访问和共享。数据和代码也会使用**GitHub**进行版本控制，确保数据的可追溯性和可复现性。报告会以**PDF**格式保存，以便给相关人员（如教师、管理层）展示分析结果。

场景**:** 我的**DMV**作业
"Gathering:
从 **Common Crawl** 获取了一个关于**HTTP**响应代码和服务器类型的数据集**.**
具体工具：**Python (Requests, BeautifulSoup)**，用于从网站抓取和清理数据，**Common Crawl API** 用于获取数据集。
Processing:
使用了两个**Python**脚本来清理和提取有用信息。第一个脚本用于下载原始数据集并清洗掉无关的数据，第二个脚本提取了所有**HTTP**响应代码及其对应的服务器类型。这里使用的工具包括：**Python**，以及用于数据处理的库如 **Pandas** 和 **NumPy**，用于数据清理和转换。
Analysis:
对服务器类型及其对应的**HTTP**响应代码进行了聚合和排序，从而得到排名前几的服务器类型。使用 **Python (Pandas)** 可以通过 **groupby()** 方法来按服务器类型分组，使用 **sort_values()** 方法对数据框进行排序。
Presenting: we chose an interactive chord diagram. It uses chord thickness and color hues to represent load distributions and relationships.
Preserving：
代码在**colab**上，保存 **Colab** 文件和清洗前后的**csv**文件在**Google Drive**上。**"**

the process and rules of scraping data from a website

process:
了解目标网站
设置抓取工具
发送**HTTP**请求
使用**HTML**解析器（如`BeautifulSoup`）解析网页内容
存储数据

处理错误和反抓取措施

Rules:
尊重数据隐私
声明抓取目的
不可干扰网站正常功能
遵守网站的服务条款
遵守网站的**Robots.txt**文件

Identify how each error or artefact is most likely to have been introduced,State any assumptions.

错误或伪影	引入阶段	原因和假设
1. 年龄分布存在不正确的分组	数据处理 (Processing)	假设在对数据进行预处理时，年龄分组的规则或代码逻辑出错，例如未正确地定义年龄区间。可能是在分组时没有校验分组的连续性或重叠性，导致错误分组。
2. 0-5岁的年龄分组中出现摩托车驾驶员这种不合理数据	数据收集 (Gathering)	假设在数据采集阶段，未对参与者类型和年龄分组进行逻辑验证。例如，采集时可能将不同数据源直接合并，导致儿童被错误地归类为摩托车驾驶员。缺乏字段间的逻辑校验机制。
3. 2012年存在死亡人数为 -999 的异常值	数据处理 (Processing)	假设在处理缺失值或异常值时，错误地使用占位符如“-999”来表示未知或缺失的数据。这可能是由于开发人员或工具未能正确地处理这些值，将其保留在数据集中，而非清洗或转换为合适的值如“0”。

data quality metrics

维度	数据清洗	数据质量指标
含义	数据清洗是指识别和修复数据中的错误、缺失值、重复项或不一致的步骤，以确保数据的准确性和一致性。	数据质量指标是用来衡量数据准确性、完整性、一致性等质量属性的标准，帮助评估和改进数据的整体质量。
好处	- 提高数据分析的准确性和可靠性。 - 降低错误模型和决策的风险。	- 帮助发现数据缺陷或问题。 - 提供改进数据管理流程的依据。
区别	- 数据清洗是处理数据质量问题的具体过程。 - 关注数据本身的修复任务。	- 数据质量指标是评估数据质量的量化标准。 - 关注如何衡量和评估数据的整体状态。
常见指标	- 无特定指标（数据清洗是操作任务）。	数据完整性：检查数据是否有缺失值、空值或不一致的部分。 数据准确性：确保数据反映真实情况，并且没有错误输入或记录。 数据一致性：确保数据在不同来源或时间点之间

		的一致性。 数据时效性：确保数据是最新的，反映了实际的情况。 数据合规性：确保数据收集和处理遵循适当的法律和伦理标准，特别是在涉及个人隐私的情况下。
常见任务	1. 缺失值处理：用均值填补或删除缺失记录。 2. 去除重复值：清除重复记录。 3. 格式标准化：统一数据格式。	1. 检测错误：检查是否存在不合理值。 2. 数据对比：对比数据是否一致或重复。

数据质量的关键问题及解决方法

关键问题	解决方法
1. 数据的准确性如何确保？	- 使用数据验证规则检查输入（如年龄字段仅允许数字）。 - 定期审计和校验数据。
2. 数据的完整性如何保证？	- 设计表结构时设置必要字段为“非空”。 - 使用自动化脚本检测和补充缺失值（如用均值填补）。
3. 数据的一致性如何检测？	- 采用统一的数据格式（如日期格式设置为“YYYY-MM-DD”）。 - 数据整合时使用ETL工具（如Talend）检查字段一致性。
4. 数据的及时性如何提高？	- 实施实时数据处理系统（如Apache Spark）。 - 建立定期更新流程，确保数据不会陈旧。
5. 数据的可信度如何评估？	- 检查数据来源的可靠性（如是否为权威机构）。 - 对历史数据和新数据进行交叉验证。
6. 数据质量指标如何监控？	- 使用数据质量工具（如Informatica）设置关键指标（如缺失率、错误率）。

Identify a weakness (or important task that is not included) with the Generic Data Analytics Pipeline

一个常见的弱点或重要任务，通常不包括在 通用数据分析管道（**Generic Data Analytics Pipeline**）中的是 数据质量监控与评估（**Data Quality Monitoring and Evaluation**）.

原因：

- 数据的质量直接影响到分析结果的准确性和可靠性。无论数据收集、处理、分析还是呈现，如果基础数据质量不高，那么后续步骤的结果也将无法产生可靠的结论。
- 在通用的数据分析管道中，通常没有单独的步骤来监控和评估数据的质量，尤其是持续的质量保证和审查机制。这样可能导致数据不准确、丢失、冗余或偏差，从而影响最终的分析结果。

缺失的影响：

- 没有及时识别和修复数据问题，可能导致误导性的分析结论或决策。
- 在处理大量数据时，容易忽视质量问题，最终影响模型的预测准确度和数据的可解释性。
- 在数据分析管道的后期，错误的数据质量评估会导致错误的决策或浪费资源。

如何改进？

- 在数据处理和分析阶段之间加入 数据质量评估和修复 的步骤。
- 采用自动化工具和脚本，定期监控和报告数据质量。
- 结合 机器学习 或 统计方法，发现和纠正数据中的潜在质量问题。

属性/列的数据类型

每一列数据的类型，如分类数据（Categorical）、数值型（Numerical）、 日期时间型（Datetime）等。	- 分类数据：性别（男/女） - 数值型：年龄（整数）、工资（浮点数） - 日期时间型：2023-12-06T14:00:00

元数据

元数据类型	描述
描述性元数据	描述数据的内容、属性和特征。 例如，文档的标题、作者、发布日期、图像的分辨率。品牌型号、颜色、屏幕尺寸。
管理性元数据	描述数据的管理、存储和维护信息。 例如，数据创建日期、访问权限、文件格式、版本号。购买日期、质保信息。
结构性元数据	描述数据的结构和组织方式。例如，数据库表结构、字段名称、文件编码格式。序列号、文件系统格式、硬件组件结构、应用程序列表。

使用标准化元数据如何改善元数据数据质量？以及实施元数据标准的一项潜在困难：

标准化元数据的质量提升：

- 一致性和准确性：使用标准化元数据（如统一的属性名称和数据格式）可以确保所有衣物的数据条目具有一致性。这使得数据之间的比较和分析更加可靠，并且消除了歧义。

- 便于集成：如果所有CA682学生的衣物数据都遵循同一标准，集成不同来源的数据时会更加方便，减少因元数据格式不统一而导致的错误。
- 数据可重用性：标准化元数据使得不同系统和平台间的数据交换和共享变得更加顺畅，可以提高数据的可重用性。

潜在困难：

- 强制实施标准的困难：强制所有学生遵守一个统一的元数据标准可能会遇到实施难题，尤其是当用户缺乏技术知识或时间时。此外，不同的学生可能对标准有不同的理解，这可能导致数据收集时的偏差或错误。

Data Management Skills

技能	关键问题	适用技术/工具	解释与举例
规划数据存储	- 数据规模和复杂性如何？ - 数据查询频率高吗？ - 数据需要实时更新吗？ - 是否有隐私数据？	- 关系型数据库（如MySQL、PostgreSQL） - NoSQL数据库（如MongoDB、Cassandra） - 云存储（如AWS S3）	- 关系型数据库：适合结构化数据，如公司员工记录。 - NoSQL数据库：适合半结构化或非结构化数据，如日志文件或社交媒体内容。 - 云存储：适合存储大规模静态文件。
理解数据质量的概念	- 如何定义数据的准确性、完整性、一致性和及时性？ - 有哪些因素可能导致数据质量下降？	- 数据质量管理工具（如Talend、Informatica） - 数据验证脚本	- 测量数据质量：分析缺失值比例、不一致性（如日期格式混乱）。 - 导致数据质量差的原因：如手动输入错误、数据收集流程不规范。
数据清洗方法	- 数据中是否存在缺失值、重复值或异常值？ - 如何自动化数据清洗过程？	- Python库（如Pandas、NumPy） - 数据清洗工具（如OpenRefine）	- 清洗步骤： 1. 使用Pandas检测重复值并删除。 2. 用均值或中位数填充缺失值。 3. 检测异常值并剔除（如Z-score超过3）。

数据保护与隐私	- 数据是否包含敏感信息？ - 数据共享和存储时如何确保隐私合规性？ - 是否遵守法律法规（如GDPR）？	- 数据加密工具（如VeraCrypt） - 访问权限管理（如AWS IAM）	- 保护方法：对敏感数据（如用户密码）进行加密存储。 - 示例：公司通过访问控制限制员工查看客户数据权限。

规划数据存储的关键问题及解决方法

关键问题	解决方法
1. 数据规模和复杂性如何？	- 对于大规模和复杂数据，选择分布式数据库（如Hadoop HDFS）。 - 对于简单结构化数据，使用关系型数据库（如MySQL）。
2. 数据查询频率高吗？	- 高查询频率场景选择内存数据库（如Redis）。 - 查询不频繁但需要批量处理，选择数据仓库（如Amazon Redshift）。
3. 数据需要实时更新吗？	- 实时更新场景选择流处理数据库（如Kafka）。 - 非实时场景选择传统数据库（如PostgreSQL）。
4. 数据是否包含隐私信息？	- 使用加密工具对数据进行加密（如SSL/TLS）。 - 实施访问控制（如AWS IAM）。
5. 数据格式是什么（结构化、非结构化）？	- 结构化数据选择关系型数据库。 - 半结构化数据选择NoSQL数据库（如MongoDB）。

适用技术/工具及使用场景

存储技术/工具	适用场景	使用原因
关系型数据库（如MySQL, PostgreSQL）	适用于结构化数据，如公司员工信息、财务记录。	结构化数据：销售数据、库存数据（如产品ID、描述、单价、交易记录等）都属于结构化数据，符合关系型数据库的特点，能方便地进行表格化存储。 **ACID**特性：关系型数据库支持ACID事务特性，确保数据的一致性和完整性，特别是在处理涉及交易的销售数据时。 可扩展性：现代的关系型数据库可以根据需要扩展，适合处理跨地区的多个商店的数据。 查询和分析：关系型数据库非常适合

		进行SQL查询、报表生成和数据分析。
NoSQL数据库（如MongoDB, Cassandra）	适用于半结构化或非结构化数据，如日志文件、JSON文档、社交媒体内容。	为什么会改变？ 数据类型的不同：网站日志和社交媒体内容通常是非结构化或半结构化数据（例如，日志文件、社交媒体帖子和评论），这些与结构化的销售和库存数据有所不同。 存储需求的不同：非结构化数据的存储和处理需求更加灵活，通常需要支持大规模数据存储和分析的系统。
数据仓库（如Amazon Redshift, Snowflake）	适用于大规模批量数据分析，如企业年度销售报表、客户行为数据的离线分析。	
分布式文件系统（如Hadoop HDFS）	适用于大规模分布式存储和处理，如流媒体服务的视频文件存储。	
内存数据库（如Redis, Memcached）	适用于需要高性能、低延迟的应用，如实时缓存、电商库存管理。	
流处理数据库（如Apache Kafka）	适用于实时数据流处理，如传感器数据、交易记录的实时分析。	
云存储（如AWS S3, Google Cloud Storage）	适用于存储海量静态数据，如备份文件、归档文档、大型图片库。	
对象存储（如MinIO, Azure Blob Storage）	适用于处理非结构化数据，如音频、视频、图像文件。	

exploratory or explanatory

特性	探索性可视化 **(Exploratory Visualization)**	解释性可视化 **(Explanatory Visualization)**
定义	用于发现数据中的模式、趋势、异常点和关系，是一种数据探索的工具。	用于讲述明确的故事或解释数据分析结果，传达特定的见解或结论。
目标	发现未知，生成假设，探索数据背后的潜在模式。	解释已知，支持假设或论点，清晰地传达发现。

使用对象	分析人员、数据科学家。	客户、决策者、非技术观众。
特点	- 偏向动态和交互式 - 通常未定型，具有开放性 - 可能生成多个图表并进行迭代	- 结构化、聚焦于特定信息 - 图表简单易懂 - 风格清晰，讲述性强
使用场景	数据分析的初期阶段，探索数据分布和关系，生成假设。	向非技术人员、管理层或公众报告分析结果，提供支撑决策的证据。
工具和方法	Python 的 Matplotlib、Seaborn、Plotly，R 的 ggplot2，Tableau 等。	Excel、PowerPoint 图表，Tableau（仪表板模式），简单条形图等。
具体例子	- 检查销售数据的分布，探索销售额与季节之间的潜在关系 - 绘制散点图观察两个变量的相关性 "- 绘制热力图分析广告点击与时间段的关系 - 观察用户行为模式" "- 散点图显示学习时间与考试成绩的关系 - 检查异常数据" - 分布图探索患者年龄与某种疾病的发病率 "- 箱线图探索年度预算分布情况 - 检查可能的数据异常"	- 创建一张折线图展示公司过去5年销售增长 - 使用柱状图说明某营销策略的成效 - 使用条形图展示哪种广告形式产生了最高的转化率 - 折线图展示过去三年不同年级的平均考试成绩趋势 - 饼图显示某药物在治疗中的成功率 - 柱状图对比不同部门的年度支出情况
可视化类型	- 散点图 - 箱线图 - 热力图 - 动态交互式图表 分布图 (Distribution Plot), 例如直方图或核密度估计。 气泡图 (Bubble Chart),带有额外维度（如大小）的散点图。 平行坐标图 (Parallel Coordinates Plot) 小提琴图 (Violin Plot) 时间序列图 (Time Series Plot) 树状图 (Tree Map),用于探索数据的层次结构。 主成分分析图 (PCA Plot)	- 条形图 - 折线图 - 饼图 - 简单的静态可视化 面积图 (Area Chart) 仪表盘图表 (Dashboard Visualizations) 树状图 (Tree Diagram) 瀑布图 (Waterfall Chart) 甘特图 (Gantt Chart) 文字云 (Word Cloud)
复杂度	复杂，可能需要技术人员进一步解释。	简单易懂，面向更广泛的受众。
优先级	探索数据的完整性，检查可能存在的异常值或错误。	提炼核心信息，确保清晰、有效传达发现

Data Visualisation skills

技能	关键要点	方法/规则	举例与说明

规划可视化	1. 选择图表类型	- 根据数据类型和目标选择合适的图表： - 比较：条形图、折线图 - 分布：箱线图、直方图 - 关系：散点图	举例： 电商平台需要分析每日销售趋势，选择折线图来表示每天销售额的变化。若需要展示地区销量分布，选择柱状图。
	2. 理解沟通和人类感知	- 优先考虑用户目标：传递信息还是发现规律？ - 遵循感知原则：颜色区分有限，避免过多信息堆积	举例： 如果用户需要快速识别数据差异，用不同色调表示，但避免超过5种颜色，否则可能引起认知负担。
	3. 使用设计规则	- 颜色：使用对比鲜明的颜色，不同数据组使用一致色系 - 布局：重要信息放在视觉焦点区域（如左上角或中间）	举例： 分析收入来源的饼图，采用亮眼色调区分大类别，同时在图旁注释具体比例，保证用户快速理解。
阅读、评价和拆解可视化	1. 评估沟通效果	- 检查图表是否清晰传递了作者意图 - 是否引入不必要的复杂性	举例： 若分析每月用户增长趋势，条形图可以更清晰展示单一数值对比，而过多折线图可能引起混乱。
	2. 评估人类感知原则的应用	- 是否使用了合适的视觉编码（如大小、长度优于角度对比） - 是否避免误导信息（如扭曲比例、不等间隔坐标轴）	举例： 收入趋势图若使用截断纵轴可能造成误导，给人数据变化剧烈的假象，应选择从零起始的坐标轴。
	3. 检查设计规则是否有效	- 颜色对比是否合适，避免太亮或过于花哨 - 布局是否有逻辑，重要数据是否突出显示	举例： 产品对比图中，各品牌用统一色系渐变区分，使用户容易快速识别主次关系。

Marks是图表中的基本数据元素

图表类型 **(Chart Type)**	**Marks (**标记**)**	**Visual Attributes (**视觉属性**)**
柱状图 (Bar Chart)	条形（Bar）	- 位置：条形的长度或高度表示数值 - 颜色：用不同颜色表示不同类别
散点图 (Scatter Plot)	点（Point）	- 位置：X 和 Y 轴的坐标表示数据点 - 颜色：不同颜色表示不同类别 - 形状：不同形状表示不同类别
折线图 (Line Chart)	线（Line）	- 位置：数据点的位置由坐标表示 - 颜色：不同颜色表示不同系列 - 形状：折点的形状（例如圆点）

饼图 (Pie Chart)	弧（Arc）	- 角度：弧的大小代表数据的比例 - 颜色：每个部分用不同颜色表示
热图 (Heatmap)	热图块（Tile）	- 颜色：颜色的深浅代表数值的高低 - 位置：位置代表数据行列
面积图 (Area Chart)	区域（Area）	- 位置：X 和 Y 轴的坐标表示数据的关系 - 颜色：不同类别的区域使用不同颜色
气泡图 (Bubble Chart)	气泡（Bubble）	- 位置：气泡的位置表示数据的两个变量 - 大小：气泡的大小表示第三个变量 - 颜色：不同颜色表示不同类别
堆积柱状图 (Stacked Bar Chart)	条形（Bar）	- 位置：条形的长度表示数值的大小 - 颜色：每个部分用不同颜色表示不同类别 - 堆叠：展示各部分的堆积关系
树图 (Treemap)	矩形（Rectangle）	- 位置：矩形的布局决定数据的层次结构 - 大小：矩形的大小与数据的值成正比 - 颜色：用颜色区分不同类别
雷达图 (Radar Chart)	线（Line）或点（Point）	- 位置：数据点根据角度和数值确定位置 - 颜色：不同颜色代表不同类别
箱线图 (Box Plot)	箱子（Box）、点（Point）	- 位置：箱体和点表示数据的分布情况 - 颜色：颜色表示不同类别 - 长度：箱体的宽度代表数据的分布
弦图 (Chord Diagram)	弦（Arc）、弧（Arc）	- 角度：弧和弦的大小反映数据的比例 - 颜色：不同颜色代表不同类别
流图 (Flow Chart)	流线（Line）	- 位置：流线的起点和终点表示数据的流动路径 - 颜色：不同颜色代表不同的流向
直方图 (Histogram)	条形（Bar）	- 位置：条形的长度或高度代表数据的频数 - 颜色：条形的颜色可以区分不同的类别

visual attributes用来表示数据的特性

Visual Attribute	含义	对比	举例	使用场景	适用图表类型
位置 (Position)	数据点在图表中的具体位置，通常通过 X 和 Y 轴定义。	最精确，适合强调数量关系或趋势。	数据点 (2,5) 在散点图上表示为特定的 X 和 Y 坐标位置。	展示数值趋势、比较数值或分类。	折线图、散点图、柱状图
长度 (Length)	线段、条形或图形的长度表示数值大小。	比较直观，适合强调数值差异。	一个柱状图中，条形长度分别为 50 和 30，表示两组数据的大小差异。	数值大小的直接比较。	柱状图、条形图

方向 (Direction)	数据点的线段或趋势线的方向表示变化趋势。	清晰度高，适合展示趋势方向变化。	折线图中，趋势向上表示增长，向下表示减少。	展示变化趋势或方向。	折线图、箭头图
角度 (Angle)	扇形或角度大小表示数据比例或分布。	适合展示部分与整体的关系，但不如长度精确。	饼图中一个扇形区域的角度占比为 25%，表示该部分占总数的四分之一。	部分与整体关系展示。	饼图、雷达图
面积 (Area)	图形的面积大小表示数据的数值大小。	易引起误导（面积变化为平方关系），适合大致比较。	一个气泡图中，面积大小分别为 10 和 25，表示两组数据的大小差异。	数据相对大小的直观对比。	气泡图、树图
颜色 (Color)	通过颜色的不同或渐变表示数据分类或数值范围。	易解读，适合大范围数据或分类数据展示。	热力图中红色表示高值，蓝色表示低值。	分类、数值范围分布、强度对比。	热力图、分面图
纹理 (Texture)	用不同的图案或线型表示数据分类。	适合印刷，但不如颜色直观。	条形图中用斜线纹理区分不同类型的数据组。	数据分类展示。	条形图、区域图
形状 (Shape)	数据点的不同形状表示不同类别或分组。	适合区分类别，但精确性不如位置或长度。	散点图中圆形表示男性，三角形表示女性。	分类数据的区分。	散点图、多类别图表
明暗度 (Intensity)	数据颜色的明暗或透明度表示数据的密度或强度。	适合表现分布或密度，但不如位置或长度直观。	热力图中，颜色越深表示密度越高。	数据密度或强度的可视化。	热力图、地图图表
点 (Point)	单个数据点用来标记具体的值或位置。	精确但孤立，适合表示独立数据。	散点图中，每个点代表一个观测值（如城市人口和温度）。	标记位置或独立数据点。	散点图、地图
线 (Line)	数据点间的连接表示趋势或关系。	清晰展现趋势，但可能隐藏数据细节。	折线图中，线条表示销量随时间的变化。	表达趋势或连续性。	折线图、流图
颜色 (色相) (Hue)	用不同颜色表示分类或属性。	易于区分类别，适合少量类别数据。	条形图中用红色表示男性，蓝色表示女性。	分类数据的直观区分。	条形图、热力图、分面图
坡度 (Slope)	线条的倾斜度表示变化速率或趋势。	可表现变化速度，但对数值不敏感。	折线图中，斜率越大表示变化越快。	展示速率或比较趋势变化。	折线图、回归分析图
大小 (长度) (Size - Length)	用长度表示数值大小。	精确、直观，适合数量比较。	条形图中，长度不同的条表示不同销量。	强调数值差异。	条形图、柱状图
大小 (面积) (Size - Area)	用图形面积表示数值大小。	直观但易误导（面积变化是平方关系）。	气泡图中，较大气泡表示较高的值。	展示相对大小或视觉比较。	气泡图、树图
数量 (Quantity)	用数据点的数量表示密度或频率。	适合大范围分布数据，可能缺乏精确性。	直方图中，高柱子表示频率较高。	显示分布、频率或密度。	直方图、热力图

preattentive features和视觉特征

特征类型	定义	举例
预注意特征 (Preattentive Features)	一种视觉特征，能够在极短的时间内（通常不到250毫秒）快速被识别出来，通常用于吸引注意力并引导观众的视线。	颜色（红色背景上的绿色点），大小（一个比其他点大的圆圈），形状（不同形状的点或符号）。
视觉特征 (Visual Features)	指通过图形元素和其视觉表现方式来传达信息的任何属性，广泛包括了形状、颜色、大小、位置、对比度等，它们不仅有助于快速感知，还帮助我们深度理解数据。	形状（圆形、方形），颜色（红色、蓝色），大小（大与小），位置（图表上的位置分布）。

QualitativeQuantitativeDiscreteContinuousNominalOrdinalIntervalRatio

```
├── Qualitative（质性）
│   ├── Nominal（名义型）
│   └── Ordinal（顺序型）
└── Quantitative（量化）
    ├── Discrete（离散型）
    │   ├── Ordinal（顺序型）
    │   └── Ratio（比率型）
    └── Continuous（连续型）
        ├── Interval（区间型）
        └── Ratio（比率型）
```

	定义	例子	特点和使用场景	与其他类型的区别

定性(Qualitative)	用于描述类别或质量，而不是数值，无法进行数学运算。	性别（男/女）、职业（医生/教师）、国家（中国/美国）。	适合分类数据，用于展示类别之间的差异或分布，常用于统计分析中的频率分析。	无法进行数值计算，与定量数据不同。
定量(Quantitative)	表示数量的数值数据，可以进行数学运算。	年龄（30岁）、收入（5万元/年）、身高（175厘米）。	用于展示数值大小或趋势，可直接用于统计分析（如均值、方差）。	能进行数学计算，与定性数据不同。
离散型(Discrete)	数据是整数，表示可以数的项，没有小数或连续值。	家庭人数（3人）、订单数量（5件）、考试通过人数（25人）。	适合有限且可数的情况，如频率分布分析。	值是有限集合，与连续型数据不同。
连续型(Continuous)	数据可以是任意小数值，表示一个范围内的所有可能值。	温度（37.5℃）、长度（1.75米）、体重（65.5千克）。	适合表示范围或变化的情况，可用于精确分析。	值是无限的，精度可以随测量增加，与离散型数据不同。
名义(Nominal)	只有分类和标记功能，无大小或顺序关系。	性别（男/女）、血型（A/B/O/AB）、品牌（苹果/三星）。	适合表示无序的分类，可用于频率统计或分布分析。	无法排序，与序数型不同。
序数(Ordinal)	表示类别数据，同时具有排序关系，但没有精确数值差距。	满意度（满意/一般/不满意）、学历（小学/初中/高中/本科）、 酒店评级（1星到5星）。	可排序但不可计算，用于比较等级数据。	可排序但无数值间隔，与间隔型不同。
间隔(Interval)	数据间隔有意义，但没有绝对零点，无法进行比率计算。	温度（摄氏 0°C 和 10°C 的差距有意义，但 0°C 并不表示“无温度”）、年份（2023年和2022年的差距有意义）。	可加减但不可乘除，用于时间序列或温度分析。	没有绝对零点，与比率型不同。
比率(Ratio)	数据具有绝对零点，可以进行加减乘除运算。	收入（0元表示无收入，50元是25元的两倍）、体重（0千克表示无重量）。	适合比例分析和直接计算。	有绝对零点，与间隔型不同。

数据的格式

数据格式	数据类型	描述	常见文件格式	用途	举例
结构化数据	表格化数据，具有明确的行列结构，常用于数据库。	数据按行列组织，易于存储和查询，通常适用于关系型数据库。	CSV、Excel（.xlsx）、SQL Dump、JSON、Parquet	数据分析、统计建模、查询系统。	客户数据库、销售记录、库存管理系统

非结构化数据	没有固定格式，通常为文本、图片、视频或音频文件。	数据无固定的结构，通常是自由格式的文本或多媒体文件。	文本文件（.txt）、PDF、音频文件（.mp3）、图像文件（.jpg/.png）、视频文件（.mp4）	NLP处理、图像识别、音频分析。	社交媒体帖子、客户反馈、电影视频文件
半结构化数据	部分有结构，如带标签的数据。	包含一些结构元素（如标签），但不像结构化数据那样严格遵循表格结构。	JSON、XML、YAML	Web应用数据交换、API数据。	电子邮件内容、XML格式的产品目录
时间序列数据	时间戳记录的数据，常见于物联网、金融和传感器领域。	每个数据点都带有时间标签，用于分析随时间变化的模式或趋势。	CSV、HDF5、Time Series Database（如InfluxDB）	分析趋势、异常检测。	股票价格变化、传感器温度数据、网络流量数据
地理空间数据	包含位置坐标或地理信息的数据。	数据包含关于地理位置的信息，如经纬度、地形、地图边界等。	GeoJSON、Shapefile（.shp）、KML、GPX	地图绘制、空间分析。	GPS轨迹、城市规划数据、地理信息系统（GIS）地图
流数据	实时生成的数据流，通常来自传感器或在线服务。	数据在不断流动和更新，通常用于实时处理和分析。	Kafka Stream、JSON、Avro	实时监控、流处理。	传感器数据流、在线交易数据、实时社交媒体推文
多媒体数据	包括图像、视频和音频文件。	包含图像、音频和视频等类型的多媒体数据，通常用于视觉或听觉处理。	PNG、JPEG、MP4、AVI、WAV	图像处理、视频分析、音频处理。	电影文件、音频录音、医疗影像
日志数据	系统或应用程序生成的运行记录。	记录程序运行、事件或系统状态的文本文件。	Log文件（.log）、JSON	系统监控、故障排除。	服务器日志、应用程序错误日志、访问日志
元数据	描述数据的数据，包括创建日期、格式、大小等信息。	描述其他数据的特征信息，用于数据管理和存储优化。	JSON、XML、RDF	数据管理、数据检索。	数据库表结构定义、图像的EXIF数据、文件大小和创建日期

大数据的经典特征

特征	描述	举例
体量（Volume）	数据的规模非常庞大，通常从TB（千兆字节 ）到PB（千万兆字节 ）。	社交媒体平台（如Facebook或Twitter）每天生成数十TB的数据。
速度（Velocity）	数据生成、处理和分析的速度非常快。大数据往往是实时产生的。	股票市场数据每毫秒更新一次，或者物联网设备的传感器数据实时产生。
多样性（Variety）	数据的类型多种多样，包括结构化数据、半结构化数据和非结构化数据。	智能家居系统中的数据可能包括文本、图像、视频和传感器数据等不同类型。
真实性（Veracity）	数据的可靠性或不确定性，数据可能是杂乱的、不完整的或不一致的。	社交媒体中的数据可能包含噪声，有些帖子可能不相关或不可靠。
价值（Value）	数据的有用性，指通过数据挖掘和分析可以获得的洞察力和决策支持。	分析客户购买数据来识别趋势，从而改进市场营销策略。

most likely to be classified as big data:

A Viewing data for Netflix subscribers including the show and the date watched and social media sentiment analysis responding to the show.
B Sales data from the four DCU campus restaurants and catering facilities in 2020.
C A download of content and metadata from my personal twitter account.
D Player training data (sensors and observations) from the Irish Rugby Squad.
Answer: A

explain why your choice is most likely to produce big data.

Answer:

体量（Volume）：Netflix拥有全球数以百万计的用户，产生了大量的观看数据。同时，社交媒体情绪分析也会生成大量的数据（例如用户评论、帖子、点赞等）。这些数据的体量非常庞大，远远超过普通数据集的规模。
速度（Velocity）：Netflix的观看数据和社交媒体的情绪分析数据会实时或近实时更新。例如，Netflix用户每时每刻都在观看新节目，社交媒体上的评论也在不断更新。
多样性（Variety）：这些数据不仅包括结构化的观看记录（节目名称、日期等），还包括非结构化的社交媒体内容（用户评论、情感分析等）。数据的来源和类型非常多样。
真实性（Veracity）：这些数据可能存在一定的不确定性或噪声，尤其是社交媒体情绪分析中的评论可能带有主观性或不完全准确的信息。
价值（Value）：这些大数据可以用来分析用户行为，优化推荐算法，了解用户对节目的情感反应，进而提升用户体验和节目推荐的准确性。

分析是否符合“大数据”的经典特征。

1. Volume（数量）- 数据量

- 是否符合大数据特征？ 是的，符合。超市的忠诚卡程序通常涉及成千上万甚至更多的顾客。每个顾客的账户数据、购买记录、以及与忠诚卡相关的各种活动数据（如购物频率、购物类型等）都会被记录和存储。随着时间的推移，数据量会迅速增长，可能涵盖数百万甚至更多的交易记录和客户互动数据。由于这个数据量巨大，它非常符合“大数据”中**数量（Volume）**的定义。
- 假设：我们假设超市链在多个地点运营，且忠诚卡程序覆盖广泛的顾客群体，因此产生了大量的数据。

2. Variety（多样性）- 数据的多样性

- 是否符合大数据特征？ 是的，符合。超市的忠诚卡程序会产生多种类型的数据，包括结构化数据（如交易金额、商品种类、购买日期等）、半结构化数据（如客户的反馈或评论），以及可能的非结构化数据（如顾客参与的广告活动或促销活动数据）。这些数据的来源和形式各不相同，因此具有很高的多样性。
- 假设：我们假设数据包括从顾客的购买记录到在线互动、广告响应等各种类型的信息。

3. Velocity（速度）- 数据的速度

- 是否符合大数据特征？ 是的，符合。超市的忠诚卡程序会实时或接近实时地生成大量的数据。例如，每次交易时，顾客的购买记录都会立刻更新到数据库中。与此同时，顾客参与促销活动、回应广告的行为也会在短时间内产生大量的数据。这些数据的更新速度非常快，需要快速处理和分析，以支持即时决策和营销活动。
- 假设：我们假设该超市使用现代的技术架构来实时处理顾客的交易和互动数据，以便能够及时响应顾客需求。

个人信息和敏感信息

DCU

Personal Data

Any **information** relating to an **identified or identifiable** natural **person** (“data subject”);

an identifiable person is one who **can be identified, directly or indirectly**, in particular by reference to an identification number, location data, an online identifier or to one or more factors specific to his physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Personal data may be **processed** only if the data subject has **unambiguously** given his/her **consent** (“prior consent”).

(art. 4, n. 1), Regulation (EU) 2016/679)

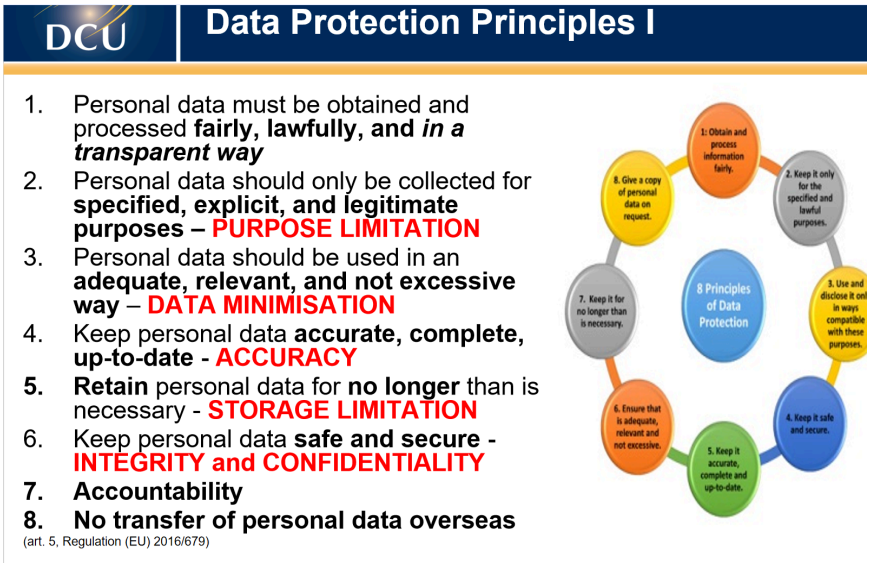
Sensitive (Personal) Data (Special categories)

Personal data revealing **racial** or **ethnic origin, political opinions, religious or philosophical beliefs**, **trade-union membership**, and the processing of **genetic data, biometric** data for the purpose of uniquely identifying a natural person, data concerning **health** or data concerning a natural person’s sex life or **sexual orientation**.

Sensitive data may be processed only if the data subject has given his/her **explicit consent** to the processing of those data (“prior written consent”).

(art. 9, Regulation (EU) 2016/679 + art. 10)

GDPR保护原则



了解并遵守数据保护法规：确保数据处理符合GDPR或相关国家/地区的法律要求。
获取明确同意或其他合法依据：明确告知数据主体如何使用其数据，并根据需要获得其同意。
必须以公平、合法和透明的方式获取和处理个人数据
个人数据的收集应仅用于特定、明确和合法的目的 – 目的限制
个人数据的使用应充分、相关且不过度 – 数据最小化
保持个人数据准确、完整、最新 – 准确性
保留个人数据的时间不得超过必要时间 – 存储限制
保持个人数据安全可靠 – 完整性和保密性
问责制
不得将个人数据转移到海外

handled differently due to European GDPR requirements. Explain why or why not.

数据类型	需要特别处理的原因	是否受**GDPR**影响	解释
客户数据（姓名、邮箱、电话、支付详情、送货地址）	客户数据属于个人可识别信息，GDPR要求对这类数据进行严格管理，确保合法性、透明性、必要性和安全性。	是	客户数据必须遵守GDPR规定，获取客户明确同意、确保数据安全、防止滥用，并且客户有权访问、修改、删除或限制处理其个人数据。
销售交易数据（与个人信息关联的交易记录）	如果销售数据与客户的个人信息有关（例如通过客户账号进行的交易），则这些数据也是个人数据，受GDPR	是	销售交易记录可能包含客户的个人信息，必须遵循GDPR的隐私保护规定，即

	保护。		使支付信息通过第三方支付平台处理，也需要确保数据的安全性和隐私权。
忠诚度计划数据（例如客户积分、购买历史）	如果忠诚度计划涉及收集客户的个人信息，GDPR适用，要求确保客户数据的合法收集、处理和存储。	是	忠诚度计划中涉及个人信息时，必须获得客户同意，并提供明确的数据用途说明。处理这些数据时，必须遵循GDPR的管理要求。
员工数据（姓名、联系方式、工资等）	员工的个人信息属于个人数据，必须符合GDPR的隐私保护要求，尤其是在工资和工作记录等敏感信息的处理上。	是	员工数据受到与客户数据类似的保护，必须确保数据的合法收集、存储和使用，遵循数据最小化原则，避免泄露员工的私人信息。
产品数据（产品ID、描述、单价等）	这类数据与任何个人身份无关，因此不涉及GDPR要求。	否	产品数据不涉及任何个人信息，因此不受GDPR管辖。

绘制图表

```
option = {
  tooltip: {
    trigger: 'axis',
    axisPointer: {
      // Use axis to trigger tooltip
      type: 'shadow' // 'shadow' as default; can also be 'line' or 'shadow'
    },
    formatter: (params) => {
      let total = params.reduce((sum, item) => sum + item.value, 0);
      let details = params
        .map(
          (item) =>
            `${item.marker}${item.seriesName}: ${item.value}万 (${(
              (item.value / total) *
              100
            ).toFixed(1)}%)`
        )
        .join('<br>');
      return `季度总支出: ${total}万<br>${details}`;
    }
  },
  title: {
    text: '2019年各季度员工等级支出分布',
```

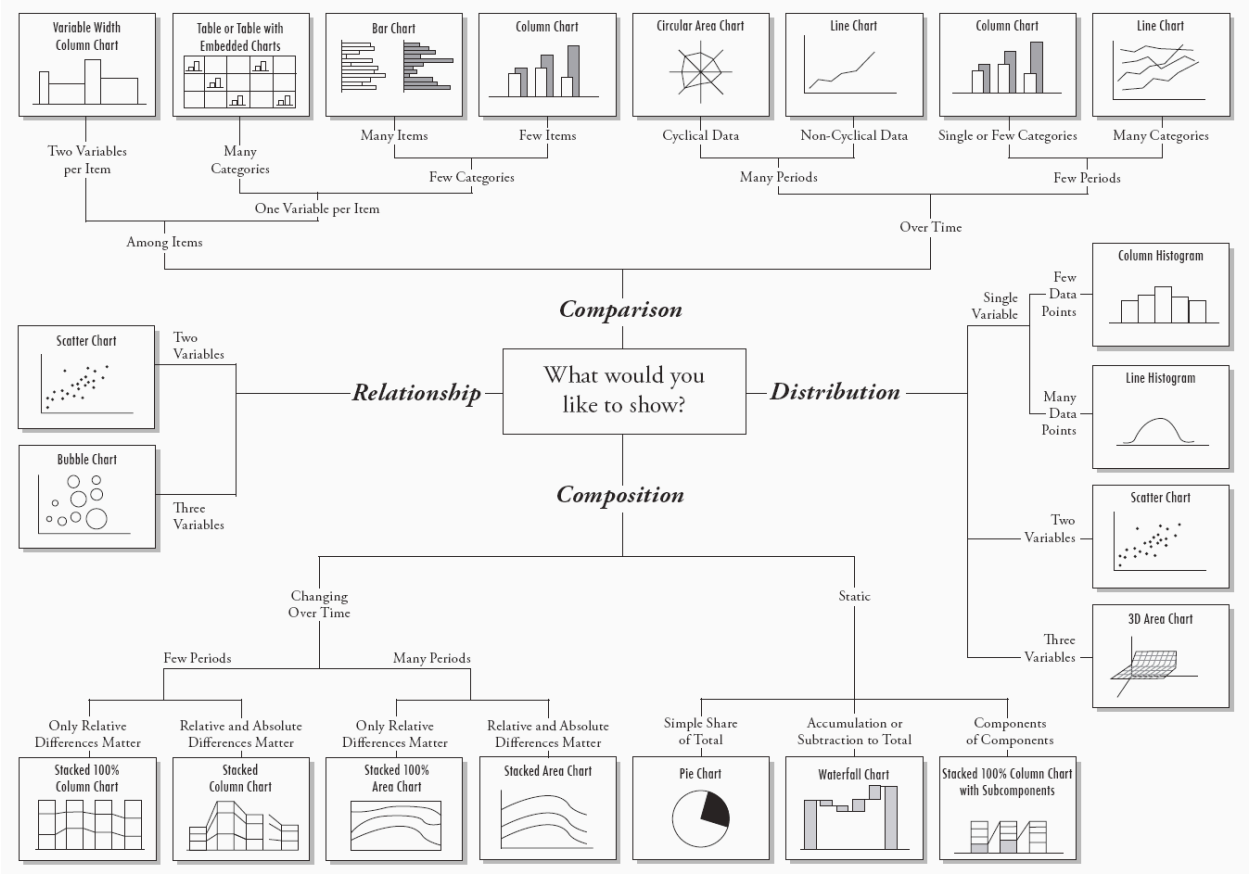
```
    subtext:'数据显示各员工等级在不同季度的支出对比(单位:万)',
    left: 'left'
  },
legend: {
      right: '5%',
      data: ['等级1', '等级2', '等级3', '等级4', '等级5']
  },
 grid: {
  left: '3%',
  right: '4%',
  bottom: '10%',//每组柱子之间的间距(分类间距)
  containLabel: true
 },
 xAxis: {
  type: 'value',
  splitLine: { show: true, lineStyle: { type: 'dashed', color: '#ccc' } }
  // splitLine: { show: false } // 隐藏网格线
 },
 yAxis: {
  type: 'category',
  data: ['Q1', 'Q2', 'Q3', 'Q4'],
  axisTick: { show: false }, // 隐藏刻度线
 },
 series: [
    {
       name: '等级1',
       type: 'bar',
       stack: '总量',
      label: {
       show: true,
       position: 'inside',
       formatter: '{c}万',
       textStyle: { fontSize: 12, fontWeight: 'bold' }
      },
      itemStyle: {
          color: '#4575b4',
          borderColor: 'black',
          borderWidth: 1,
          shadowBlur: 10, // 阴影模糊半径
          shadowColor: 'rgba(0, 0, 0, 0.5)', // 阴影颜色
          shadowOffsetX: 2, // 阴影水平偏移
          shadowOffsetY: 2  // 阴影垂直偏移
      },
      barWidth: 35, // 设置柱子宽度
```

```
    data: [10, 25, 30, 35],
    markLine: {
      data: [
        { xAxis: 70, name: '支出目标' } // 基准线
      ],
      lineStyle: {
        type: 'dashed', // 虚线
        color: '#ff0000' // 红色线条
      },
      label: {
        formatter: '目标值: {c}万', // 显示基准值
        position: 'end',
        color: '#ff0000'
      }
    },
  },
  {
    name: '等级2',
    type: 'bar',
    stack: '总量',
    label: {
      show: true,
      position: (val) => (val > 20 ? 'inside' : 'top'),
      formatter: '{c}万',
      textStyle: { fontSize: 12, fontWeight: 'bold' }
    },
     itemStyle: { color: '#74add1' },
    barWidth: 40, // 设置柱子宽度
    data: [30, 10, 25, 20]
  },
  {
    name: '等级3',
    type: 'bar',
    stack: '总量',
    label: {
      show: true,
      position: 'inside',
      formatter: '{c}万'
    },
        itemStyle: { color: '#abd9e9' },

    barWidth: 40, // 设置柱子宽度
    data: [25, 30, 20, 15]
  },
```

```
    {
      name: '等级4',
      type: 'bar',
      stack: '总量',
      label: {
        show: true,
        position: 'inside',
        formatter: '{c}万'
      },
      itemStyle: { color: '#fee090' },
      barWidth: 40, // 设置柱子宽度
      data: [15, 15, 10, 10]
    },
    {
      name: '等级5',
      type: 'bar',
      stack: '总量',
      label: {
        show: true,
        position: (val) => (val > 20 ? 'inside' : 'top'),
        formatter: '{c}万',
        textStyle: { fontSize: 12, fontWeight: 'bold' }
      },
      itemStyle: { color: '#fc8d59' },
      barWidth: 40, // 设置柱子宽度
      data: [10, 30, 10, 10]
    },
  ]
};
```

图形的选择树



suggest an appropriate graph type

A. Compare the performance of stocks in Microsoft, Apple and Samsung over

the last 5 years.

推荐图表类型：折线图（**Line Chart**）

- **CHRTS** 类别：时间序列（Time Series）
- 简要理由：
    - 折线图是比较随时间变化的数据的理想选择。在这个任务中，我们需要比较三家公司（微软、苹果和三星）股票的表现，折线图可以清晰地展示每家公司股票价格随时间的变化趋势。
    - 由于数据是时间序列数据，折线图能够直观地显示每家公司在过去五年内的股票表现变化，使得不同公司间的趋势对比更为直观和清晰。
    - 可以通过不同的线条或颜色来区分不同公司，进一步提高可视化效果。

B. Explore movie commercial performance for the IMDB top 50 by director based

on cost to make and ticket sales.

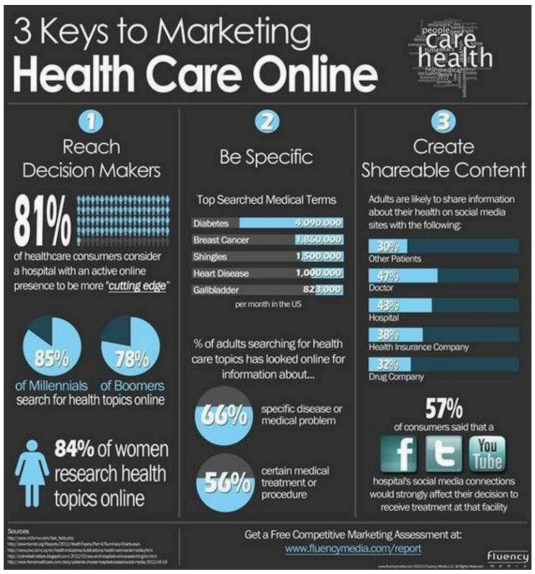
推荐图表类型：气泡图（Bubble Chart）

- **CHRTS** 类别：关系（Relationships）
- 简要理由：
    - 气泡图适用于展示三个变量之间的关系。在这个任务中，我们需要分析每部电影的制作成本、票房销售额以及导演的信息。
    - 横轴可以表示制作成本，纵轴可以表示票房销售额，而气泡的大小则可以表示每部电影的排名或其他相关的数值（例如导演的名气）。气泡图能够帮助揭示不同导演制作电影的商业表现，并通过视觉化展示票房与制作成本的关系。
    - 气泡图清晰地展示了不同电影的表现，同时由于气泡的大小和位置，可以帮助识别出高制作成本但票房表现不佳的电影，或反之。

communication purpose

类别	定义	目标	设计特点	
Information	提供事实或数据，传递信息。	让读者快速了解某个主题的核心数据或内容。	使用简洁文字、图表、重点数据，信息直观，避免复杂解释。	Design principles: Structure is key Level of detail - macro v micro Layout, Color
Persuasion	试图改变观众的观点，推动某种行为或信念。	说服观众采取行动（如购买产品、加入课程等）。	强调情感语言、号召性语句（如“立即行动”）、可能使用故事化表达。	Communication to elicit a particular response E.g., Advertising → using information to present a message Appeal: Factual (rational) vs Emotional (values, opinions, attitudes) Design principles: Research audience Illustrations, themes, colours, grouping → attract viewer’s eye
Education	系统传递知识或技能，让读者深入理解某个主题。	帮助观众学习新概念或技能，提供背景、原因和深度解释。	内容系统化，信息层次清晰，可能包括过程分析（如图示）和技能列表，支持观众自主学习。	transferring knowledge and skills Textbooks, online learning resources, brochures, movies Design principles:Divide information into chunks (hierarchy - trees, chapters, etc.) Legibility is key Progressive disclosure
Entertainment	提供愉悦或娱乐，激发情感共鸣。	让观众感到轻松、有趣或愉悦。	色彩明快、内容幽默、形式活泼（如漫画、故事化叙述等）。	pleasure, diversion, amusement art, video games, film, television, ebooks Design principles: focus on narrative how constructed (lighting, layout, multimodality) → the medium Style ...

例题01



中文回答

(i) What is the **main** communication purpose and why?
这个图片的主要沟通目的是 Persuasion（劝说/说服）。
原因：
目标明确：图片的目的是说服医疗机构关注如何通过在线营销提升影响力，并给出具体的数据和建议来支持这一观点。
激发行动：通过关键数据（如“81%的消费者认为活跃的在线形象更具吸引力”、“57%的消费者会因社交媒体影响选择医院”），试图说服医疗机构重视在线推广。
提供解决方案：图片不仅描述了现状，还给出了行动方向（如“创建可分享内容”、“关注高搜索量疾病”），进一步增强说服力。
因此，这张图片的核心目的是通过数据和建议 说服医疗机构采纳特定的在线营销策略，以改善其服务推广和患者覆盖效果。

(ii) What design choices or guidelines have been used to support this purpose?

1. 视觉层次和重点突出：
    - 大字号的数字（如“81%”、“57%”）和可视化图表（如条形图、饼图）突出了关键统计数据，使信息一目了然。
    - 像“Reach Decision Makers”（接触决策者）和“Be Specific”（明确定位）这样的标题起到导航作用，引导观众快速抓住重点。

2. 清晰分区：
    - 信息被划分为三个独立的部分（接触决策者、明确定位、创作可分享内容），使内容层次分明、易于理解。
3. 颜色搭配：
    - 一致使用蓝色和白色，与黑色背景形成鲜明对比，增强可读性，同时突出重要数据点。
    - 使用不同颜色区分数据类别（如疾病搜索量和社交媒体分享偏好）。
4. 面向特定受众：
    - 提供针对性的人群数据（如“85%的千禧一代搜索健康话题”与“78%的婴儿潮一代搜索健康话题”），确保内容与特定群体相关联。
5. 可操作性建议：
    - 每个部分都包含具体的可执行建议，如针对搜索量高的医疗话题或利用社交媒体增强患者决策影响力。
6. 字体设计：
    - 大型无衬线字体让数据更直观易懂，不同字体大小显示信息的重要性层次。
7. 图表辅助：
    - 使用图表（如饼图、条形图）直观展示复杂数据，提升观众的理解力和吸引力。

综上，这张图通过 视觉吸引力、清晰的结构和可操作性内容，成功地达到了传播目的，帮助医疗营销人员更好地制定在线策略。

英文回答

(i) Main Communication Purpose

The main communication purpose of this graphic is **to provide insights and strategies for effectively marketing healthcare services online.** It emphasizes:

- The importance of an active online presence for healthcare institutions.
- The need to focus on specific medical topics of interest to attract and engage online audiences.
- Encouraging the creation of shareable content to enhance online outreach.

This is achieved by presenting statistical evidence, trends, and actionable tips, all tailored to help healthcare marketers understand their target audience and make data-driven decisions.

(ii) Design Choices or Guidelines Used

1. **Visual Hierarchy and Emphasis**:
    - **Large bold numbers** (e.g., "81%", "57%") and graphical representations (e.g., bar charts, pie charts) emphasize key statistics and make them easy to understand at a glance.
    - Headlines like "Reach Decision Makers" and "Be Specific" act as focal points, guiding the viewer through the three main strategies.
2. **Clear Sections**:
    - The content is divided into three distinct sections (Reach Decision Makers, Be Specific, Create Shareable Content), making the information digestible and structured.
3. **Color Coding**:
    - Consistent use of **blue and white** contrasts with the black background, enhancing readability and focus on critical data points.
    - Different shades and graph types differentiate data categories (e.g., disease terms and sharing preferences).
4. **Relevance to Audience**:
    - The inclusion of demographic-specific data (e.g., Millennials vs. Boomers, women researching health topics) ensures the information resonates with specific user groups.
5. **Actionable Insights**:
    - Each section provides practical advice for healthcare marketers, such as focusing on specific medical terms or leveraging social media to influence patient decisions.
6. **Typography**:
    - Large, sans-serif fonts make the data approachable, while varying font sizes highlight the hierarchy of importance.
7. **Graphical Support**:
    - Visual aids like pie charts and bar graphs complement the textual information, making complex data easier to interpret.

In summary, the design effectively combines **visual appeal, clarity, and actionable content** to fulfill its purpose of educating and guiding healthcare marketers.

Gestalt Laws 的Proximity Similarity Enclosure Closure Continuity Connection

原则	定义	视觉特点	示例
Proximity（接近性）	彼此靠得较近的元素会被视为一个组。	空间上靠近的元素更容易被归为同一组。	一堆圆点中，如果某些圆点彼此距离更近，会被认为是一个组；例如多个图表中的分组数据。
Similarity（相似性）	外观相似的元素（如形状、颜色、大小）会被视为一个组。	在颜色、形状、大小等属性上相似的元素被自动归类。	彩色折线图中，相同颜色的线条会被认为是同一类别的数据；例如同一色调的分类标签。
Enclosure（包围性）	通过视觉边界（如框线、背景色）将元素包围起来，从而将其视为一个组。	使用边框或背景色突出某些元素，暗示它们有联系。	在信息图中，用边框圈出“重要信息”，让读者将其视为一个单独的内容块；例如用框线区分标题与正文。
Closure（闭合性）	人们倾向于将不完整的形状自动补全为一个完整图形。	未闭合的形状会被脑海中补充完整，形成感知上的闭合。	画一个半开的圆，人们会自然感知为一个完整的圆；例如公司 logo 中的半闭合图案（如 Adobe 的图标 ）。
Continuity（连续性）	人们倾向于沿着连续的路径或线条感知元素，而不会轻易中断。	元素排列成曲线或直线时，视觉上会认为它们是相关的或连续的。	一组箭头沿着一个弧线排列，人们会感知它们是连接在一起的流动；例如地图上的路径标记。
Connection（连结性）	通过视觉连线（如直线、曲线 ）连接的元素会被视为一个组。	视觉上用线条直接将元素连接起来，明确它们的关联性。	用箭头或线条连接文本框，表示步骤之间的关系；例如流程图中的步骤连接。

design rules and theories

设计规则/理论	定义	应用示例

格式塔理论（Gestalt Theory）	探讨人类如何感知整体而非独立的部分，强调元素之间的关系，帮助用户快速理解图形信息。	接近性：在信息图中，将相近的图标分为一组以展示分类数据。 相似性：使用相同颜色表示相同类别数据。 闭合性：用半圆形让用户脑补成完整的圆。
预注意特性（Preattentive Features）	人类视觉系统能快速感知颜色、形状、大小等特性，而无需集中注意力。	颜色：红色高亮异常数据点。 大小：突出较大的字体作为标题。 方向：使用箭头指示数据流向。
视觉层级（Visual Hierarchy）	调整元素显著性，引导用户注意力，确保信息按照重要性依次被感知。	使用大号字体和对比色突出标题，将次要信息用较小字体放置在下方或边缘。
信息简化原则（Simplicity & Clarity）	通过简化设计减少认知负担，避免信息过载，同时保证信息的完整性。	使用简单的柱状图代替复杂3D图，分段标注数据以减少过多信息干扰。
色彩理论（Color Theory）	利用颜色之间的关系与对比，传递信息和引发情感联想。	使用蓝色传递专业与信任感，用对比色（如红色与灰色）强调异常值或关键数据。
一致性原则 (Consistency Principle)	图表中使用的视觉元素（如颜色、形状、字体）应保持一致，避免混淆。	在多个图表中使用相同的颜色来表示同一类别的数据，如健康部门始终使用绿色，这样可以帮助用户快速识别信息。
可读性原则（Readability Principle）	可读性原则强调通过简洁、直观、清晰的设计，使得数据可视化能够快速有效地传递信息。	
泰思勒法则（Tesler’s Law）	设计的复杂性应尽量减少，但不能低于任务本质的复杂性。	在表单设计中，精简输入字段，只保留必要信息，同时确保用户操作完整。

米勒法则（Miller’s Law）	人类短时记忆一次只能处理 7±2 个信息块，因此应分组展示信息。	在信息图表中，将多个数据类别分为几组，并通过标题和颜色标识组别。
图形对齐原则（Alignment Theory）	元素的对齐能够提升设计的整洁感和可读性。	在表格和布局中使用网格系统，确保所有内容对齐，避免随意放置造成混乱。
数据可视化原则（Data Visualization Principles）	数据的图形化表示应清晰、直观，避免过多干扰，使受众快速理解数据意义。	使用折线图展示趋势，柱状图展示对比，避免使用过多颜色或复杂图案干扰数据解读。
费茨法则（Fitts’s Law）	目标越大、越近，用户点击或关注所需时间越短，因此关键内容应醒目且易于操作。	将按钮设计为大尺寸，放置在显著位置，确保用户可以快速找到并点击。
留白原则（White Space Principle）	适当留白可增强设计的舒适感和可读性，引导用户聚焦重要内容。	在文字周围留出足够空间，让页面显得整洁有序，同时引导用户关注核心内容。
对比与强调（Contrast & Emphasis）	通过颜色、大小、位置等差异突出重要内容，引导用户注意力。	使用红色标记重要数据，背景保持浅色，关键指标用大字体呈现。

批判图片1

问题 1:图表的两个可能问题及改进建议

问题	说明	改进建议	理由
1. 数据过度视觉化(过多使用小人图标)	小人图标虽然直观，但过多重复容易导致视觉混乱，不便于快速比较数据变化。	使用柱状图替代，利用简单的高度差展示增长趋势。	根据图表简洁性原则(Gestalt 简化原则)，柱状图能够清晰传达数据，避免分散观众注意力。
2. 数据变化趋势被视觉元素干扰	使用小人图标和虚线同时展示数据趋势，导致观众难以专注于数据的真实变化。	删除虚线，仅用柱状图或单一趋势线表示数据增长。	信息层级原则要求在图表中减少不必要的元素，简化视觉负担。

问题 **2**:可视化属性的使用及对注意力的影响

视觉属性

1. 形状:
    - 小人图标代表护士数量，用来直观表达人数变化。

    - 使用重复图形(多个小人叠加)传递数据大小的概念。
2. 大小:
    - 图标的数量和排列展示人数增长,但因为单位固定(一个小人代表多少人未明确),具体含义可能不直观。
3. 颜色:
    - 使用粉色表示护士数据,保持了一致性,但图表单一颜色缺乏对比,不利于强调重点。

图像如何引导注意力

1. 利用:
    - 预注意特性(**Preattentive Features**):
        - 重复(Repetition):大量的小人图标通过重复抓住观众注意力,快速传递人数变化的信息。
        - 位置(Position):图标按时间顺序排列,引导观众从左到右读取数据增长趋势。
    - **Gestalt** 理论:
        - 接近性(**Proximity**):小人图标紧密排列形成组,明确显示不同年份的人数。
        - 相似性(**Similarity**):图标形状相同,帮助观众快速识别为同一类型数据。
2. 不足:
    - 缺乏多样颜色和视觉对比,导致观众难以迅速注意到重要年份(如增长高峰年份)。
    - 预注意特性未被完全优化,比如颜色对比或显著标记未用来强调关键数据点。

改进建议

1. 优化颜色对比:
    - 对关键年份(如 2011/12 和 2013 年)使用对比色突出数据高峰。
    - 增加背景网格线,增强趋势的直观感受。
2. 简化数据呈现:
    - 替换小人图标为传统柱状图,避免信息过度视觉化,使数据的差异更加直观和易于解读。

这种改进符合图表设计原则,能够减少观众的认知负担,同时提高数据传达的效率和清晰度。

批判图片2

改进措施	合理性	原则

使用2D饼图或条形图替代3D饼图	减少扭曲，提供更清晰和准确的比较。	根据 **“Chartjunk”** 原则
调整颜色对比度	增强可读性，使相邻类别易于区分。	根据 颜色理论 原则
提供更明确的标题和详细的图例	确保观众能够更好地理解数据上下文。	根据 “可读性原则”

数据清洗工具的优劣

	Strengths	**Weaknesses**	**Resources**
Spreadsheets: Microsoft Excel, Google Sheets, Open Office.	widely available, relatively easy to use interface, support for many common data file formats, some advanced tools for data reconciliation and cleaning	memory limits (desktop or cloud), no audit trail and limited data version management options, lack of repeatability for multi-step cleaning, lack of automation	Cleaning Data with Excel and Google Sheets Top ten ways to clean your data using Excel Using a spreadsheet to clean up a dataset
OpenRefine	easier for non-programmers to use, interactive and graphical viewing of transformations, advanced options for data reconciliation (eg, geographical), auditable changes (complete provenance/undo history of all modifications), wide variety of input/output formats	standalone tool less widely available, yet another tool to learn (if you are already an experienced R/Python programmer), difficult to perform editing actions like splitting rows or inserting new rows (OpenRefine is not a spreadsheet).	Using OpenRefine (Book) Cleaning Data with OpenRefine
Python: Pandas, Numpy	powerful and flexible general purpose language, supports very wide range of input/output sources and formats, if documented and used properly then scripts are auditable and repeatable, cleaning process can be deployed to other parts of the workflow, Regular expressions are extremely powerful	not useful for non-programmers, scripting and has no graphical user interface, can be difficult to explore or view data without using extra libraries and functionality, Regular expressions require skill and practise to use well	Pythonic Data Cleaning With Pandas and NumPy – Real Python