一、基于Veeva Pulse数据的商业策略分析

职位：商业智能数据分析师 公司背景（虚拟）：您即将加入“卓越制药”（Pharma Company A），一家专注于肿瘤科（Oncology）创新药物的公司。公司购买了Veeva Pulse行业基准数据，希望借助外部数据视角，优化销售和市场策略，以应对日益激烈的市场竞争。

您的角色：作为新加入的数据分析师，您的第一个任务就是深入分析Pulse数据，并为下一季度的战略规划会议，向销售和市场部负责人提供一份数据驱动的策略建议报告。

提供的数据集：

您将收到以下8个模拟的CSV数据集（与我们之前分析中使用的数据结构相同）：

1. `hcptobrick.csv`: 医生（HCP）及其所在群体（Brick）的映射关系。
2. `brickmetrics.csv`: 每个Brick的行业平均互动指标（如平均拜访数、平均接触公司数）。
3. `hcpbrickmetrics.csv`: 将Brick的平均指标附加到每个医生上。
4. `hcpcompanymetrics_A.csv`: “卓越制药”与医生的互动数据。

请你基于数据，参考以下维度进行分析（不限于）：
1.帮助 Company A发现潜力医生客户，可以推荐top10 医生
2.帮我Company A优化当前拜访策略，可以减少部分医生的拜访，增加部分已经覆盖的医生拜访。

结果呈现 **(沟通与表达能力)**

请准备一个简短的摘要（不超过5页PPT或一页备忘录的篇幅），向非技术背景的销售总监汇报您的核心发现和三大关键建议。

二、多源医生主数据匹配 **(Doctor Mapping)**

背景：在深入分析Veeva Pulse数据之前，公司面临一个基础但至关重要的数据治理挑战：我们需要将内部来自不同业务渠道（如CRM系统、市场活动）的医生数据，与Veeva提供的医生主数据 (`veeva_master_doctors.csv`) 进行精准匹配，以建立一个统一、无重复的黄金客户视图 (Golden Customer View)。

由于不同来源的数据在录入时存在差异（例如，医院名称不统一、医生姓名包含额外字符、科室层级不同等），我们需要开发一套可靠的匹配算法来解决这个实体解析 (Entity Resolution) 问题。

提供的数据集：您将收到以下3份关于北京地区医生的数据文件：

1. **`veeva_master_doctors.csv` (100条记录)**: Veeva的医生主数据，作为本次匹配的“黄金标准”。
    - `doctor_id, name, hospital, department`
2. **`customer_A_doctors.csv` (45条记录)**: 客户A的内部CRM数据，存在医院名称缩写、科室名称不规范等问题。
    - `internal_id, doctor_name, work_unit, dept`
3. **`customer_B_doctors.csv` (55条记录)**: 客户B的市场活动数据，存在医生姓名包含后缀、医院名称包含额外地区信息等问题。
    - `id, physician_name, hospital_name, specialty`

1. 产出匹配算法设计：

请设计并详细阐述一个能够将 `customer_A_doctors.csv` 和 `customer_B_doctors.csv` 两份数据与 `veeva_master_doctors.csv` 进行匹配的算法或流程。

您的阐述应至少覆盖以下几点：

- 数据预处理: 您会采取哪些步骤来清洗和标准化每一份数据？（例如：去除多余字符、大小写转换、处理别名如“301医院”等）。
- 关键匹配字段: 您会选择哪些字段（或字段组合）作为匹配的核心依据？为什么？
- 匹配逻辑: 您会采用什么样的匹配规则？（例如：是直接进行精确匹配，还是会引入模糊匹配技术？如果使用模糊匹配，您会考虑哪种算法，如 Levenshtein distance, Jaro-Winkler 等？）
- 匹配评分与阈值: 您是否会设计一个匹配得分（Confidence Score）来评估匹配的可信度？如果是，您会如何设定一个合理的阈值来判定两条记录是否为同一人？

2. 给出匹配分析报告：

请基于您设计的算法，对客户A和客户B的数据分别进行匹配，并给出一个量化的匹配比例分析报告。

报告应包含以下内容：

- 客户**A**数据: 总记录数、成功匹配上的记录数、未匹配上的记录数，以及最终的匹配成功率**(%)**。
- 客户**B**数据: 总记录数、成功匹配上的记录数、未匹配上的记录数，以及最终的匹配成功率**(%)**。
- 未匹配原因分析: 随机抽取几个未成功匹配的案例，分析可能的原因（例如：是Veeva主数据中不存在的新医生，还是因为数据质量问题导致算法未能识别？）。
- 优化建议: 基于您的分析，如果想在下一阶段提升匹配率，您会提出哪2-3条具体的优化建议？