CSCI446/946 Assignment 1 - Task 1: Problem Analysis

CSCI446/946 作业1 - 任务1：问题分析

1. Problem Analysis | 问题分析

1.1 Analytics Objectives | 分析目标

The NewChic dataset analytics aims to find:
NewChic数据集分析旨在找到：

- **Top 10 products** from all selected categories **前10个产品**：来自所有选定类别
- **Best category** among the selected categories **最佳类别**：在所选类别中

1.2 Dataset Selection Rationale | 数据集选择依据

Why choose ALL data from NewChic dataset: 为什么选择NewChic数据集的所有数据：

- **Comprehensive Coverage | 全面覆盖**: Using all 9 categories (accessories, bags, beauty, house, jewelry, kids, men, shoes, women) provides complete market view 使用所有9个类别（配饰、包包、美妆、家居、珠宝、儿童、男装、鞋子、女装）提供完整的市场视角
- **Statistical Significance | 统计显著性**: Larger dataset (74,999 total products) ensures robust statistical analysis 更大的数据集（总共74,999个产品）确保稳健的统计分析
- **Cross-Category Insights | 跨类别洞察**: Enables meaningful comparison across different product categories 能够进行不同产品类别间的有意义比较
- **Business Value | 商业价值**: Complete dataset reflects real e-commerce platform diversity 完整的数据集反映了真实电商平台的多样性

Categories Selected: | 选定的类别： All 9 available categories are included to maximize data utilization and meet assignment requirements. 包含所有9个可用类别，以最大化数据利用率并满足作业要求。

1.3 Definition of "Top 10" and "Best" | "前10"和"最佳"的定义

1.3.1 "Top 10 Products" Definition | "前10个产品"定义

Products ranked by **Composite Score** calculated as: 产品按**复合得分**排名，计算公式为：

```
Composite Score = 0.50 × Likes Score + 0.30 × Discount Score + 0.20 × Price Score
复合得分 = 0.50 × 点赞得分 + 0.30 × 折扣得分 + 0.20 × 价格得分
```

Rationale: | 依据：

- **Customer Engagement (50%) | 客户参与度 (50%)**: `likes_count` reflects genuine customer interest and product appeal `点赞数`反映真实的客户兴趣和产品吸引力
- **Value Proposition (30%) | 价值主张 (30%)**: `discount` percentage indicates good deals for customers `折扣`百分比表明对客户的优惠程度
- **Affordability (20%) | 可负担性 (20%)**: Lower `current_price` makes products accessible to more customers 较低的`当前价格`使更多客户能够购买产品

Normalization Method: | 标准化方法： All components normalized to 0-1 scale using min-max normalization to ensure fair weighting. 所有组件使用最小-最大标准化归一化到0-1范围，确保公平加权。

1.3.2 "Best Category" Definition | "最佳类别"定义

Category ranked by **Final Category Score** calculated as: 类别按**最终类别得分**排名，计算公式为：

```
Final Category Score = 0.60 × Average Composite Score + 0.25 × Average Likes + 0.15 × Product Count
最终类别得分 = 0.60 × 平均复合得分 + 0.25 × 平均点赞数 + 0.15 × 产品数量
```

Rationale: | 依据：

- **Product Quality (60%) | 产品质量 (60%)**: Average composite score of products in category 类别中产品的平均复合得分
- **Customer Engagement (25%) | 客户参与度 (25%)**: Average customer engagement across category 类别中平均客户参与度
- **Product Variety (15%) | 产品多样性 (15%)**: Number of products available in category 类别中可用产品数量

1.4 Column Selection Strategy | 列选择策略

1.4.1 Columns for Clustering and Classification | 用于聚类和分类的列

Selected Numeric Columns (excluding 'id'): | 选定的数值列（排除'id'）：

- `current_price` | `当前价格`: Product pricing information | 产品定价信息
- `raw_price` | `原始价格`: Original pricing before discounts | 折扣前的原始定价
- `discount` | `折扣`: Discount percentage offered | 提供的折扣百分比
- `likes_count` | `点赞数`: Customer engagement metric | 客户参与度指标
- `is_new` | `是否新品`: Product newness indicator | 产品新品指示器

Rationale for Selection: | 选择依据：

- **Price-related features | 价格相关特征** (`current_price`, `raw_price`, `discount`): Core business metrics affecting purchasing decisions 影响购买决策的核心业务指标
- **Engagement feature | 参与度特征** (`likes_count`): Customer preference indicator 客户偏好指标
- **Product status | 产品状态** (`is_new`): Market positioning factor 市场定位因素
- **Exclusion of 'id' | 排除'id'**: As per assignment requirements, ID columns don't contribute to analytical insights 根据作业要求，ID列不对分析洞察有贡献

1.4.2 Columns for Result Discussion | 用于结果讨论的列

Categorical Columns Retained: | 保留的分类列：

- `category` | `类别`: Primary grouping variable for analysis | 分析的主要分组变量
- `subcategory` | `子类别`: Detailed product classification | 详细的产品分类
- `name` | `名称`: Product identification for result interpretation | 用于结果解释的产品标识
- `brand` | `品牌`: Brand analysis and recommendations | 品牌分析和推荐
- `currency` | `货币`: Financial context for price analysis | 价格分析的财务背景

Usage in Discussion: | 在讨论中的用途：

- **Business Context | 商业背景**: Enable meaningful interpretation of clustering/classification results 支持对聚类/分类结果的有意义解释
- **Recommendation Generation | 推荐生成**: Support actionable insights for NewChic business strategy 支持NewChic商业策略的可行性洞察
- **Result Validation | 结果验证**: Allow manual verification of algorithmic findings 允许对算法发现进行人工验证

2. Experimental Design | 实验设计

2.1 Data Processing Pipeline | 数据处理流程

1. **Data Integration | 数据整合**: Combine all 9 CSV files into unified dataset 将所有9个CSV文件合并为统一数据集
2. **Data Cleaning | 数据清洗**: Handle missing values with domain-appropriate strategies 使用领域适当的策略处理缺失值
3. **Outlier Management | 异常值管理**: Remove statistical outliers using IQR method 使用IQR方法移除统计异常值
4. **Feature Engineering | 特征工程**: Create composite scoring system 创建复合评分系统
5. **Normalization | 标准化**: Standardize features for machine learning algorithms 为机器学习算法标准化特征

2.2 Analysis Framework | 分析框架

1. **Descriptive Analytics | 描述性分析**: Data distribution and summary statistics 数据分布和汇总统计
2. **Clustering Analysis | 聚类分析** (Task 2): Identify natural product groupings 识别自然的产品分组
3. **Classification Analysis | 分类分析** (Task 3): Predict product categories 预测产品类别
4. **Ranking Analysis | 排名分析**: Determine top products and best category 确定顶级产品和最佳类别

2.3 Expected Outcomes | 预期结果

- Clear identification of top-performing products 清晰识别表现最佳的产品
- Data-driven category performance ranking 数据驱动的类别性能排名
- Actionable business recommendations for NewChic 对NewChic的可行性商业建议
- Validated analytical framework for e-commerce product analysis 验证的电商产品分析框架

3. Data Preprocessing Justification | 数据预处理说明

3.1 Missing Value Strategy | 缺失值策略

- **Price columns | 价格列**: Median imputation (robust to outliers) 中位数填充（对异常值稳健）
- **Discount/Likes | 折扣/点赞**: Zero imputation (reasonable business defaults) 零值填充（合理的业务默认值）
- **Categorical | 分类列**: "Unknown" category (preserves data points) "未知"类别（保留数据点）

3.2 Outlier Removal Rationale | 异常值移除依据

- **Method | 方法**: Interquartile Range (IQR) with 1.5×IQR bounds 四分位距(IQR)与1.5×IQR边界
- **Justification | 说明**: Removes extreme values that could skew clustering/classification results 移除可能影响聚类/分类结果的极端值
- **Business Impact | 商业影响**: Focuses analysis on mainstream product ranges 将分析聚焦于主流产品范围

3.3 Normalization Necessity | 标准化必要性

- **StandardScaler | 标准缩放器**: Ensures equal weighting across features with different scales 确保不同尺度特征的相等权重
- **Clustering Requirement | 聚类要求**: Distance-based algorithms need normalized features 基于距离的算法需要标准化特征
- **Classification Enhancement | 分类增强**: Improves convergence and performance 改善收敛性和性能

4. Success Metrics | 成功指标

4.1 Data Quality Metrics | 数据质量指标

- Zero missing values after preprocessing 预处理后零缺失值

- Reasonable outlier removal rate (<20% of data) 合理的异常值移除率（<20%的数据）
- Successful normalization (mean ≈ 0, std ≈ 1) 成功的标准化（均值≈0，标准差≈1）

4.2 Business Value Metrics | 商业价值指标

- Clear top 10 product identification 清晰的前10个产品识别
- Definitive best category determination 明确的最佳类别确定
- Actionable insights for NewChic strategy 对NewChic策略的可行性洞察

This comprehensive problem analysis establishes the foundation for subsequent clustering and classification tasks, ensuring alignment with assignment objectives and business requirements.

这一全面的问题分析为后续的聚类和分类任务奠定了基础，确保与作业目标和业务需求保持一致。