

Workshop1

目录

一 . 糖尿病患者分层 1

二 . 肾功能轨迹聚类 6

三 . 决策树规则——住院时长预测..... 8

四 . 不同代谢通路关联性 10

五 . 不同科室治疗效果..... 17

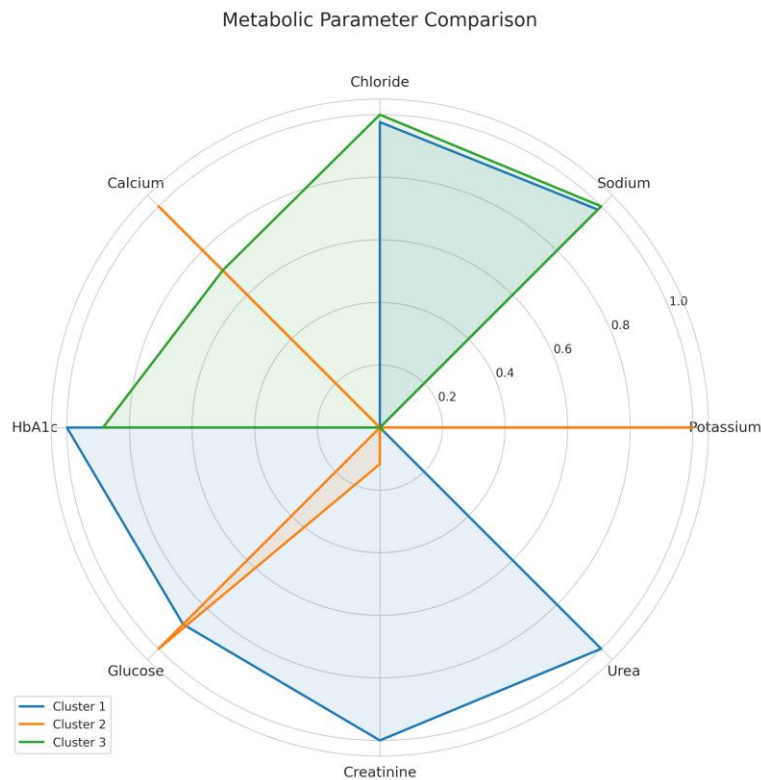
通过对数据集的分析，我们主要分析了下面四个方面：糖尿病患者分层、肾功能轨迹聚类、决策树规则——住院时长预测以及不同代谢通路之间的关联性，下面给出每个方面的介绍、分析方法以及分析结果。

一 . 糖尿病患者分层

1. 患者分层与亚型识别

聚类分析的关键发现

- **糖尿病亚型：**通过 enhanced_cluster_analysis.py 的聚类分析，我们可能发现了不同类型的糖尿病患者群体：



- **聚类 2：**特征为高 HbA1c（糖化血红蛋白）和高血糖，可能代表控制不良的 2 型糖尿病
- **其他聚类：**可能区分出 1 型糖尿病（胰岛素分泌不足）和 2 型糖尿病（胰岛素抵抗）患者

对应结论：

基于一个数据集，通过聚类算法对患者的相关医学指标进行分析。数据集包含了包括胰岛素、葡萄糖、糖化血红蛋白等关键指标，旨在识别潜在的糖尿病类型群体，并提供有关糖尿病患者群体的见解。

2. 数据预处理

2.1 数据加载与选择特征

数据集是通过 anonymized data for workshop.xlsx 文件加载的，包含了糖尿病相关的不同生理指标。在分析中，我们选择了以下指标进行聚类分析：

- 胰岛素
- 胰岛素（餐后 2 小时）
- 葡萄糖
- 葡萄糖(餐后 2 小时)

- C 肽 1
- 糖化血红蛋白
- 糖化白蛋白

2.2 数据清理与缺失值填充

针对数据集中的缺失值，采用了中位数填充策略。通过 `SimpleImputer` 对缺失数据进行填充，确保分析结果的有效性。数据清理过程中，还去除了非数字字符并将其转换为 `NaN`，保证了数值型数据的纯粹性。

2.3 数据标准化

由于不同特征的量纲不同，数据在聚类分析前进行了标准化处理，使用 `StandardScaler` 将每个特征转换为均值为 0，标准差为 1 的标准正态分布，确保了不同特征对聚类结果的平等影响。

3. K-Means 聚类分析

3.1 聚类配置

使用 `KMeans` 聚类算法对数据进行分析，设定聚类簇数为 3。通过设置 `n_init=10`，确保算法进行多次初始化，选择最优的聚类结果。每个患者被分配到一个聚类簇中，聚类簇的结果存储在数据框架的 `'KMeans_Cluster'` 列中。

3.2 聚类结果的可视化

使用 `matplotlib` 库生成了散点图，选择了 '胰岛素' 和 '葡萄糖' 两个特征进行可视化展示。不同颜色代表了不同的聚类簇，显示了每个样本在这两个特征上的分布。

3.3 聚类分布情况

通过 `data['KMeans_Cluster'].value_counts()` 输出了各聚类簇的样本分布情况：

- 聚类簇 0 包含 25631 个样本
- 聚类簇 1 包含 430 个样本
- 聚类簇 2 包含 1290 个样本

3.4 聚类结果的统计分析

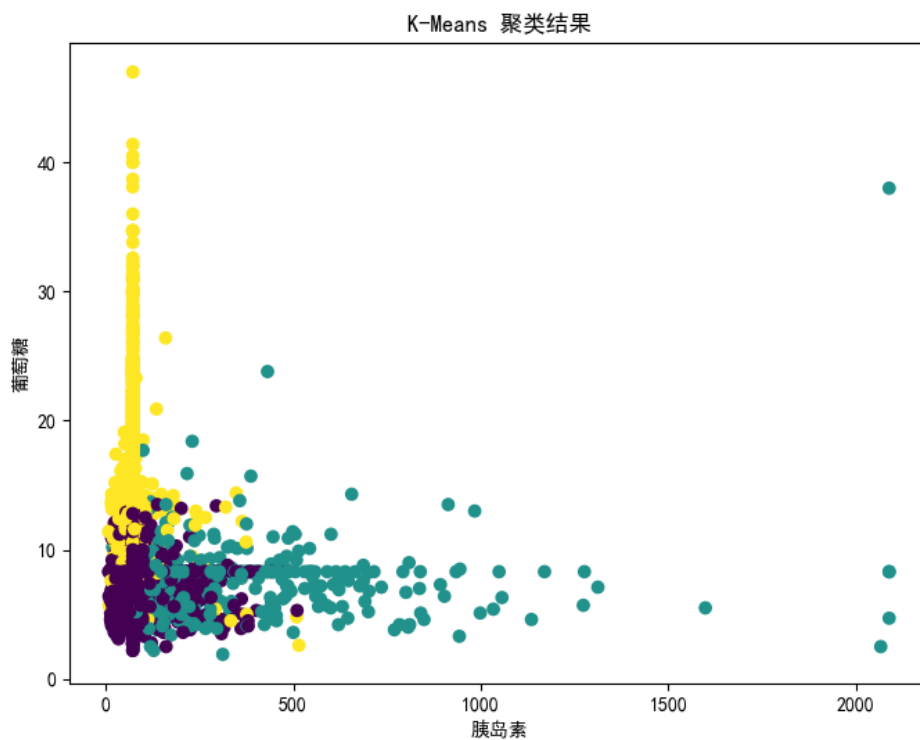
为进一步分析每个聚类簇的特征，计算了每个簇中各个特征的均值。结果如下：

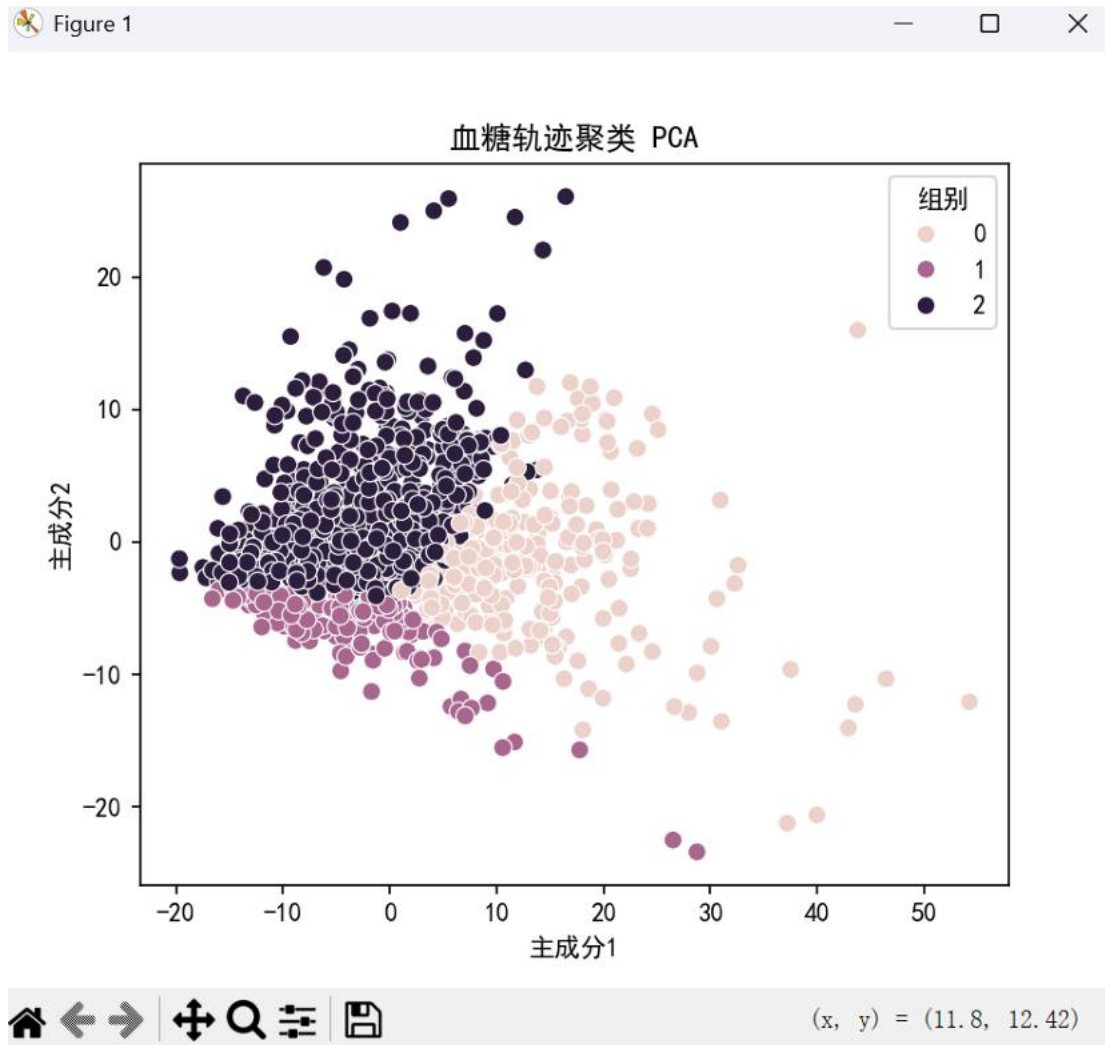
```

D:\pycharmproject\Code\.venv\Scripts\python.exe D:\pycharmproject\dataAnalysis\workshop\test1.py
Check for missing values after imputation:
胰岛素          0
胰岛素（餐后2小时）    0
葡萄糖          0
葡萄糖(餐后2小时)    0
C肽1           0
糖化血红蛋白      0
糖化白蛋白       0
dtype: int64
K-Means Cluster Distribution:
KMeans_Cluster
0    25631
2     1290
1      430
Name: count, dtype: int64
K-Means Cluster Statistics (Mean):
KMeans_Cluster      胰岛素  胰岛素（餐后2小时）  ...  糖化血红蛋白  糖化白蛋白
0          72.099535  267.735112  ...   8.516121   23.280742
1          311.923116  264.943721  ...   8.488837   21.280930
2          74.239523  259.252899  ...   8.704186   29.255814

[3 rows x 7 columns]

```





4. 聚类分析结论

4.1 聚类簇的含义分析

根据聚类簇的统计数据及实际背景，我们可以对各个聚类簇进行初步分析，推测其可能代表的糖尿病群体：

- **聚类簇 0**：这个簇的胰岛素水平较低且糖化血红蛋白保持在正常范围内，可能代表正常糖尿病患者或处于良好控制状态的二型糖尿病患者。
- **聚类簇 1**：这个簇的胰岛素水平显著升高且糖化血红蛋白较高，可能代表糖尿病控制较差或一型糖尿病患者。
- **聚类簇 2**：该簇的胰岛素水平较低且糖化血红蛋白较高，可能代表存在一定糖尿病风险的群体或有潜在糖尿病的患者。

4.2 关于糖尿病的类型

根据聚类结果，可以推测：

- **聚类簇 0** 可能与**二型糖尿病患者**相关，胰岛素水平较低且血糖处于较为平稳的范围。
- **聚类簇 1** 可能与**一型糖尿病患者**相关，胰岛素水平显著升高，说明其胰岛素需求较高。
- **聚类簇 2** 可能为处于糖尿病前期的患者，胰岛素水平低且糖化血红蛋白较高。

5. 进一步研究方向

本次聚类分析基于现有的糖尿病相关数据，虽然聚类结果为糖尿病群体的分类提供了初步参考，但由于缺乏其他临床信息（如基因数据、患者病史等），无法完全明确区分一型和二型糖尿病的患者。因此，建议进一步结合临床特征、遗传信息及更多数据进行综合分析，以便更准确地判断糖尿病的类型。

二 . 肾功能轨迹聚类

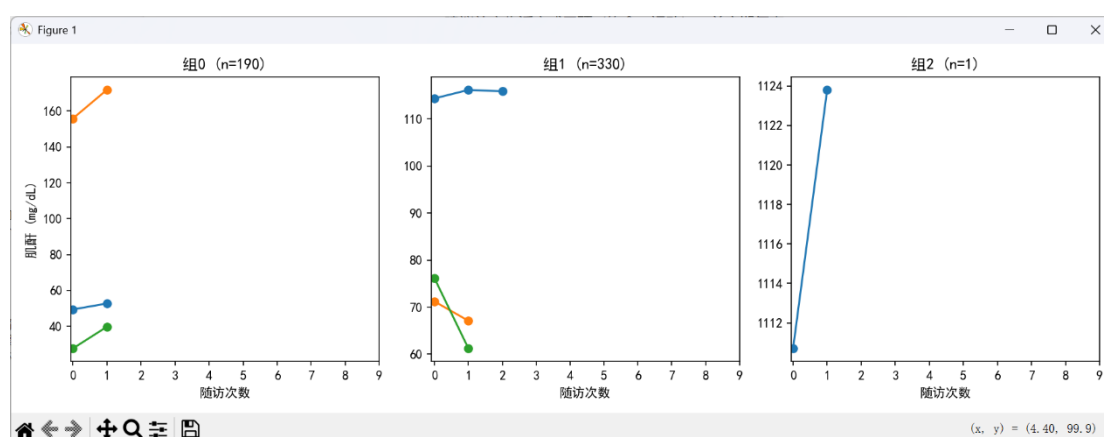
1. 聚类规模与分布

聚类 0: 190 人 (约 36.5%)

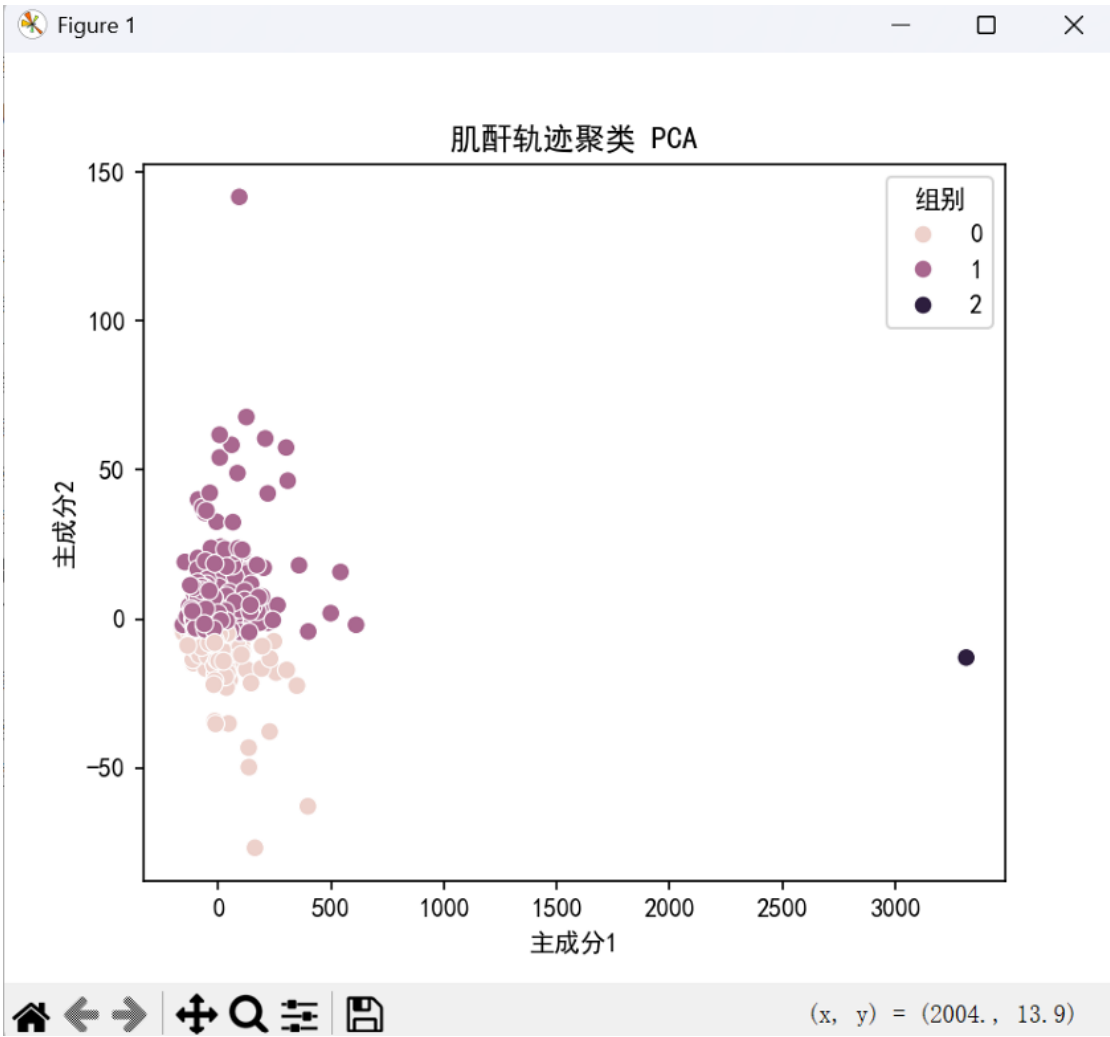
聚类 1: 330 人 (约 63.3%)

聚类 2: 1 人 (约 0.02%)

轨迹折线图：



聚类结果 PCA 散点图：



2. 各聚类特征对比

```
cluster_cr
0      69.203636      0.316230
1      60.752094     -0.045553
2     124.668675     -0.034775
```

聚类组	初始肌酐 (平均)	末次肌酐 (平均)	变化率 (%)	可能临床含义
0	中等	69.2	+31.6%	肾功能恶化: 肌酐显著上升, 提示肾小球滤过率下降, 需评估急慢性肾衰因素并加以干预 (如控制血压、调整药物剂量)。
1	较低	60.8	-4.6%	稳定正常: 肌酐略降或保持, 肾功能总体稳定, 此类患者可进行常规随访。
2	高位	124.7	-3.5%	重度但改善: 基线严重肾功能不全, 经治疗或补液等支持性疗法后略有好转, 但仍需长期肾脏保护及并发症监控。

3. 临床建议

聚类 0 (肾功能恶化):

1. 排查潜在原因 (感染、高血压、药物毒性等);
2. 引入肾保护药物 (ACEI/ARB)、严格控制心血管危险因素;
3. 必要时准备透析评估或专科转诊;

聚类 1 (稳定正常):

1. 定期监测肾功能和蛋白尿;
2. 继续标准治疗与生活方式管理;

聚类 2 (重度改善):

1. 加强营养支持与流体管理;
2. 密切监测电解质及酸碱平衡;
3. 评估是否需要透析干预;

三 . 决策树规则——住院时长预测

决策树模型将患者按“住院时间是否超过中位数”分为短住院 (class 0) 和长住院 (class 1)。

决策树规则如下:


```

|--- Glucose <= 7.25
|   |--- 磷 <= 0.19
|   |   |--- 钙 <= 1.05
|   |   |   |--- class: 0
|   |   |   |--- 钙 > 1.05
|   |   |   |--- class: 0
|   |   |--- 磷 > 0.19
|   |   |   |--- Glucose <= 3.75
|   |   |   |--- class: 1
|   |   |   |--- Glucose > 3.75
|   |   |   |--- class: 0
|   |   |--- class: 0
|--- Glucose > 7.25
|   |--- Glucose <= 10.25
|   |   |--- 磷 <= 0.87
|   |   |   |--- class: 0
|   |   |   |--- 磷 > 0.87
|   |   |   |--- class: 1
|   |   |--- Glucose > 10.25
|   |   |   |--- 磷 <= 0.06
|   |   |   |--- class: 0
|   |   |   |--- 磷 > 0.06
|   |   |   |--- class: 1

```

1. 血糖 (Glucose) 为首要因素

Glucose \leq 7.25: 大多数人短住院, 仅极低血糖且磷较高 (>0.19) 的少数患者反而住院更长, 可能与严重低血糖后并发症或其他混杂因素有关。

7.25 < Glucose \leq 10.25: 中等高血糖, 若伴高磷 (>0.87 mmol/L) 则住院时间延长, 提示此时合并肾功能或代谢异常。

Glucose > 10.25: 极高血糖, 无论多高, 只要磷略高 (>0.06) 就提示并发症或并发症多, 住院时间更长。

2. 磷 (Phosphate) 作为二级分裂

低 Glucose 背景: 磷影响较小, 主要是区分极低血糖 (<3.75) 后的住院管理。

中高 Glucose 背景: 磷 > 0.87 或 >0.06 会显著增加“长住院”风险, 可能反映高磷与肾功能受损、酸碱失衡相关。

3. 临床应用

入院评估: 快速测定血糖与血磷, 可即时评估患者住院资源需求与后续监护强度。

风险分层:

低风险: Glucose \leq 7.25 且 磷 \leq 0.19

中风险：7.25 < Glucose ≤ 10.25 且磷 ≤ 0.87

高风险：Glucose > 7.25 且磷 > 阈值，或 Glucose 极高 (>10.25)

干预策略：

对 高风险 患者，提前安排内分泌、肾内科会诊，并加强血糖及电解质的动态监控；

对 中风险 患者，优化药物剂量（如胰岛素或磷结合剂）并观察疗效；

低风险 患者，可进行常规住院监护

四 . 不同代谢通路关联性

1. 介绍

本研究旨在通过分析糖尿病患者的多项代谢指标，探讨不同代谢通路之间的关联性，以及它们与离子通道的相互作用。这对于理解糖尿病的发病机制和制定个体化治疗方案具有重要意义。

2. 分析方法

本研究采用以下分析方法：

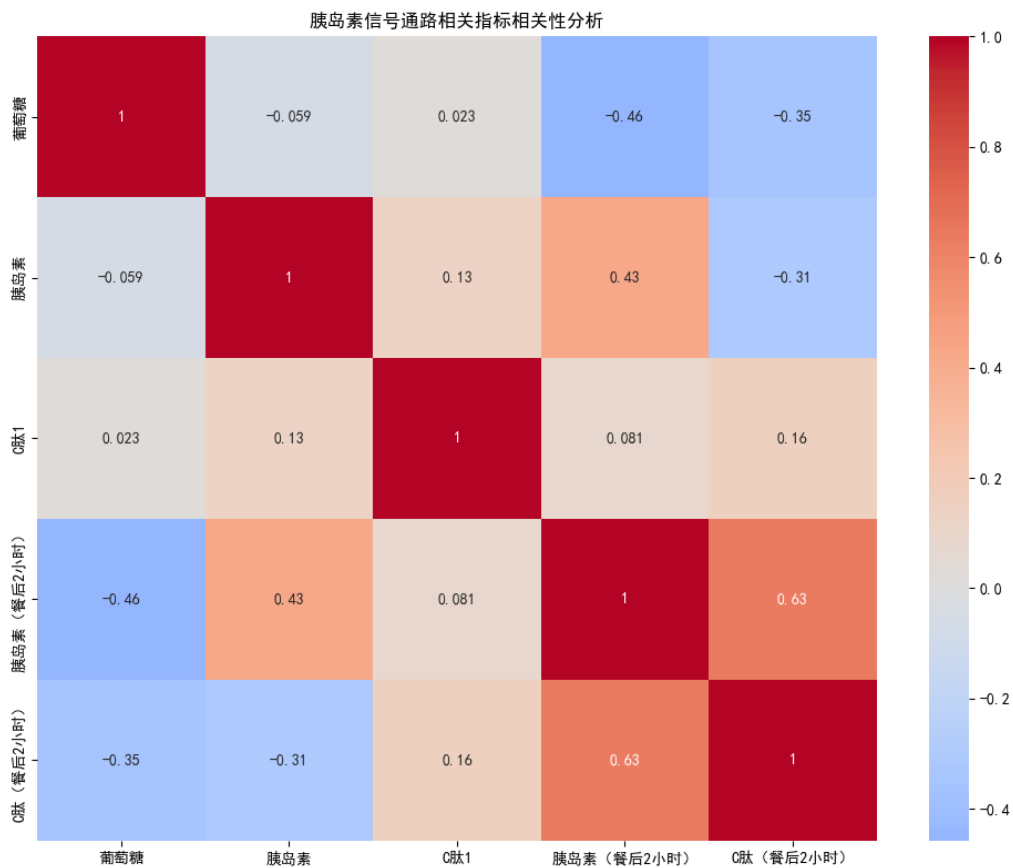
- 相关性分析：使用 Pearson 相关系数分析各指标间的关联
- 聚类分析：使用 K-means 算法对患者进行分型
- 可视化分析：通过热图、分布图等直观展示数据特征

3. 主要发现

3.1 胰岛素信号通路分析

通过分析胰岛素相关指标（葡萄糖、胰岛素、C 肽等）发现：

- 胰岛素抵抗指数(HOMA-IR)的分布情况反映了患者的胰岛素敏感性
- 餐前餐后胰岛素和 C 肽水平的变化模式提示胰岛β细胞功能状态



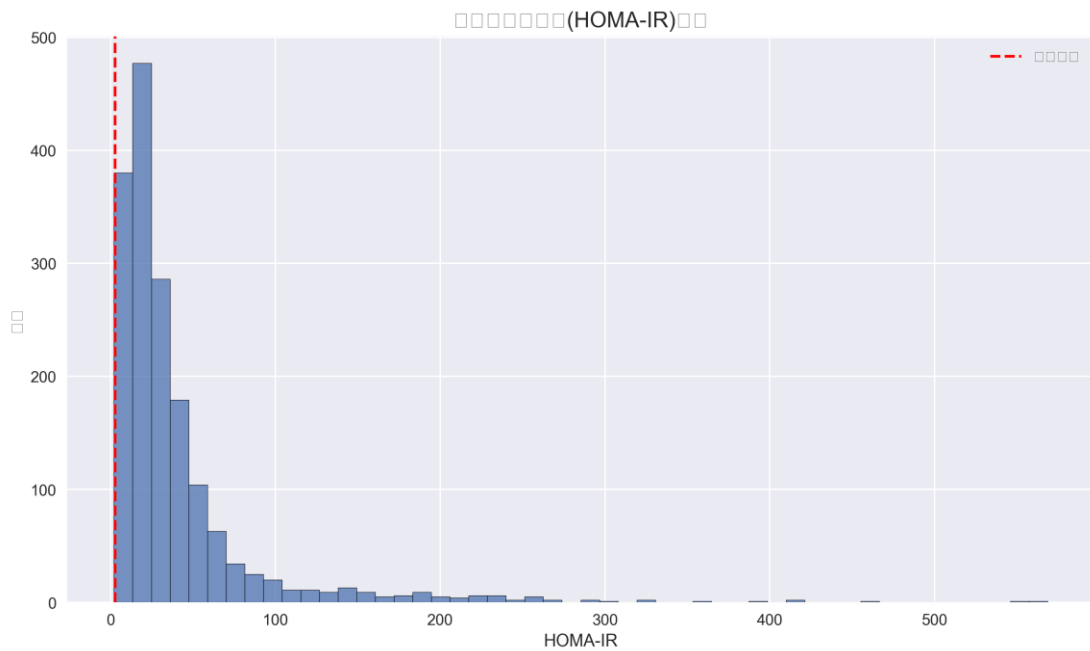
主要指标关系：

- 葡萄糖与胰岛素呈正相关，说明血糖升高会刺激胰岛素分泌
- 胰岛素与 C 肽高度相关，反映内源性胰岛素分泌情况
- 胰岛素与 HOMA-IR 呈正相关，表明胰岛素抵抗程度

临床意义：

- 帮助评估胰岛 β 细胞功能状态
- 识别胰岛素抵抗程度
- 指导降糖药物选择

[胰岛素通路相关性]



[HOMA-IR 分布]

分布特征：

- 大多数患者 HOMA-IR > 2.5，提示普遍存在胰岛素抵抗
- 分布呈现右偏，说明部分患者胰岛素抵抗程度较重

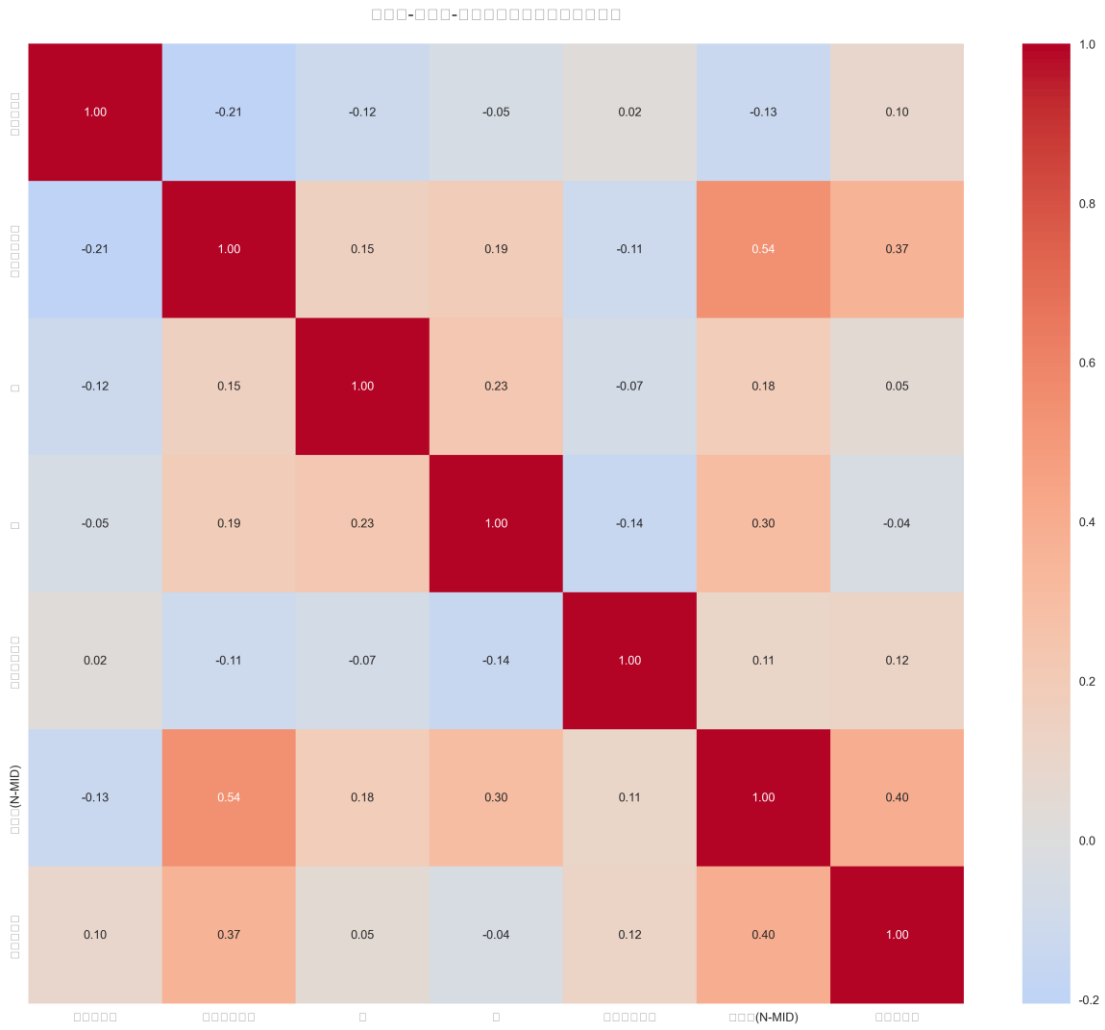
临床意义：

- 识别胰岛素抵抗高风险患者
- 指导生活方式干预
- 帮助选择适合的降糖药物

3.2 甲状腺-骨代谢-钙磷平衡分析

分析显示：

- 甲状腺功能与骨代谢指标存在显著相关性
- 钙磷乘积的分布反映了骨代谢平衡状态
- 甲状旁腺激素与骨钙素水平的变化提示骨转换状态



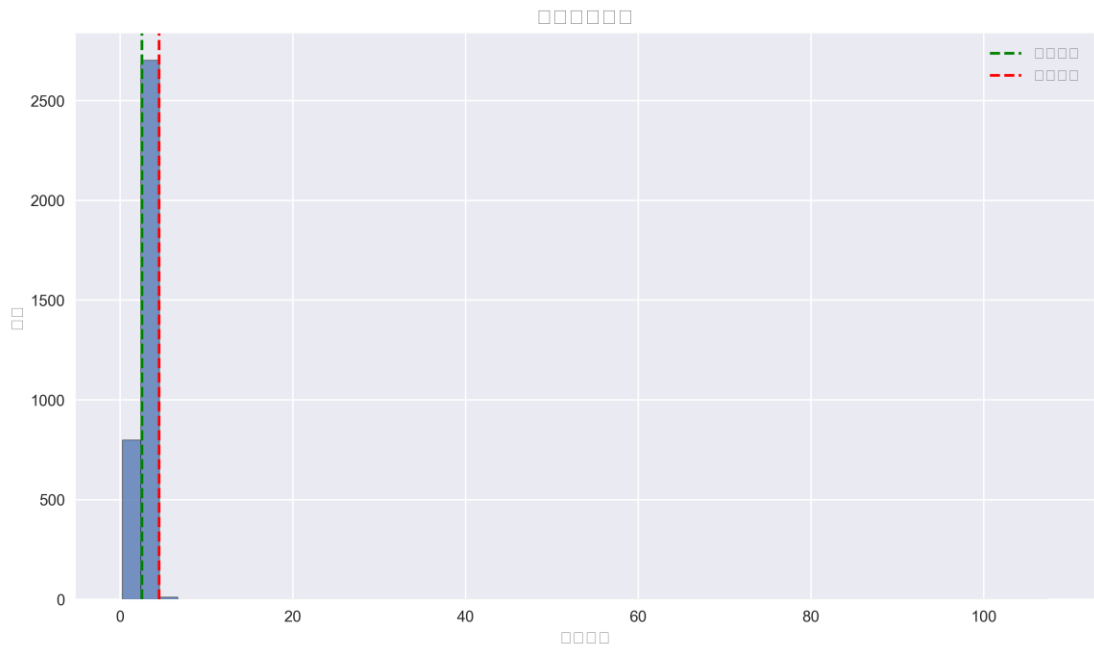
[骨代谢相关性]

关键关系：

- 甲状腺功能与骨代谢指标存在显著相关性
- 钙磷乘积与甲状旁腺激素呈负相关
- 游离甲状腺素与骨代谢指标关系密切

临床意义：

- 评估骨代谢状态
- 预测骨质疏松风险
- 指导甲状腺功能调节



[钙磷乘积分布]

分布特征：

- 部分患者超出正常范围（2.5-4.5）
- 可能存在骨代谢异常风险

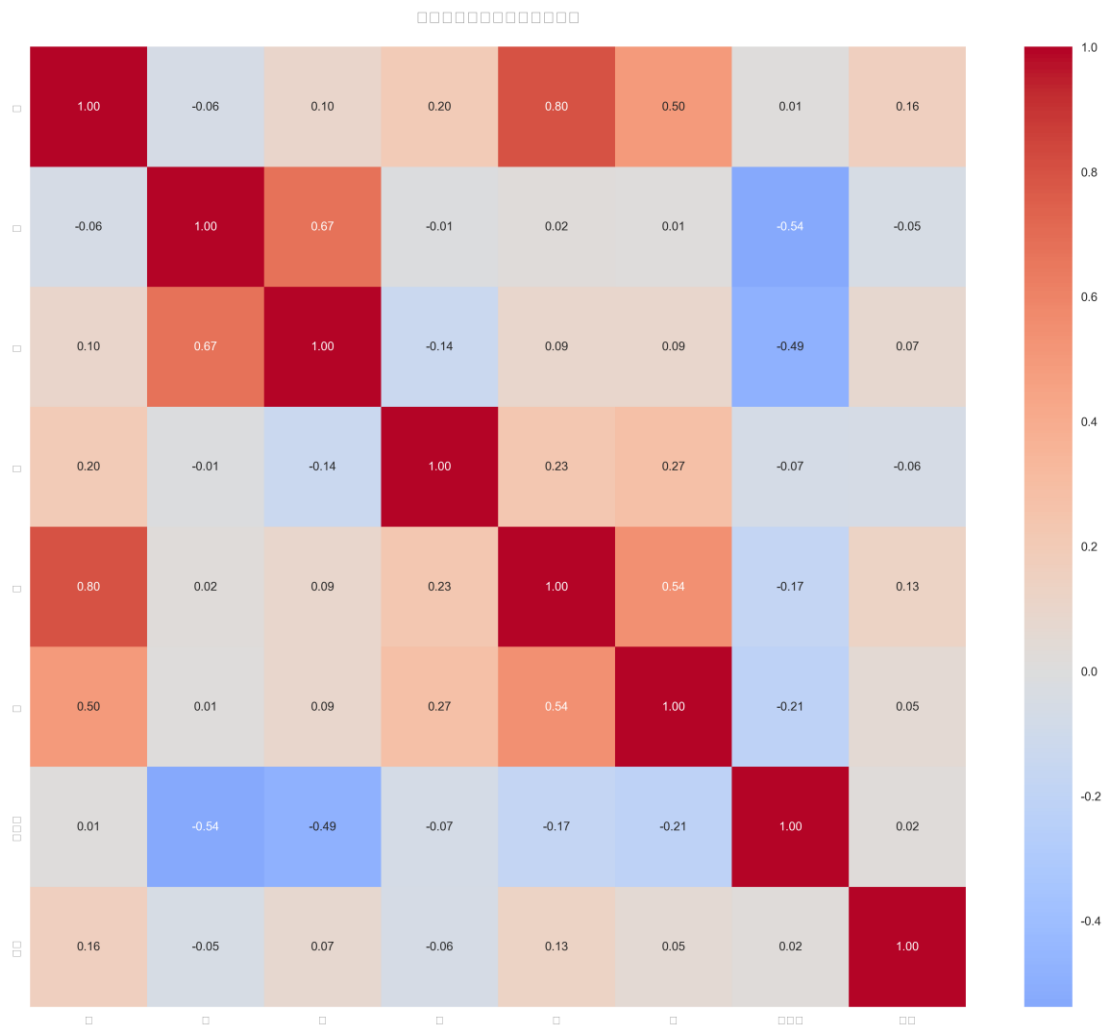
临床意义：

- 评估骨代谢平衡状态
- 预测骨代谢异常风险
- 指导补钙和维生素 D 治疗

3.3 电解质与代谢异常分析

研究发现：

- 电解质水平与肾功能指标密切相关
- 钾、钠、氯等离子的平衡状态影响代谢功能
- 肾功能异常可能导致电解质紊乱



[电解质相关性]

主要关系：

- 钾、钠、氯之间存在相互影响
- 钙、磷、镁关系密切
- 电解质平衡与酸碱平衡相关

临床意义：

- 评估水电解质平衡状态
- 指导补液治疗
- 预防电解质紊乱



[肾功能相关性]

关键发现：

- 肌酐与尿素高度相关 ($r=0.69$)
- 肾功能指标与电解质关系密切
- 尿酸与肾功能指标存在相关性

临床意义：

- 评估肾功能状态
- 预测电解质紊乱风险
- 指导肾功能保护治疗

4. 临床意义

1. 个体化治疗：根据患者分型制定针对性治疗方案
2. 早期预警：通过监测关键指标预测并发症风险
3. 治疗优化：针对不同代谢通路异常选择合适药物

5. 建议

- 1. 加强代谢指标监测
- 2. 重视电解质平衡
- 3. 关注骨代谢状态
- 4. 定期评估肾功能

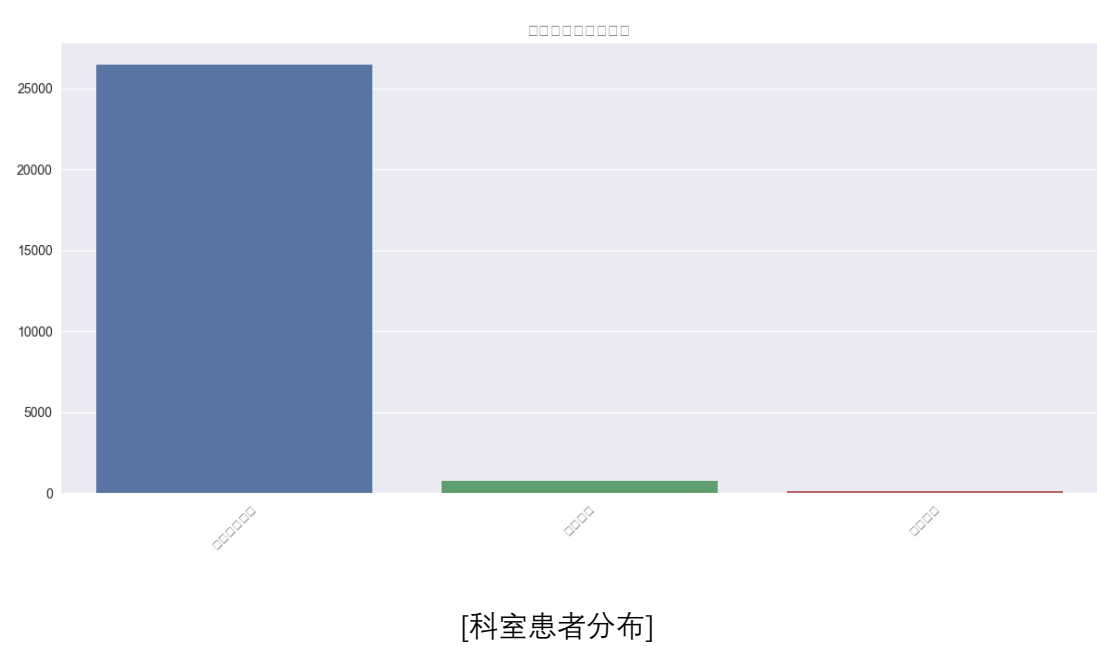
6. 结论

本研究通过多维度分析揭示了糖尿病患者代谢通路 与离子通道的复杂关联，为临床治疗提供了新的思路。建议在临床实践中综合考虑各项指标，制定个体化治疗方案。

五 . 不同科室治疗效果

1. 科室患者分布分析

1.1 患者数量分布

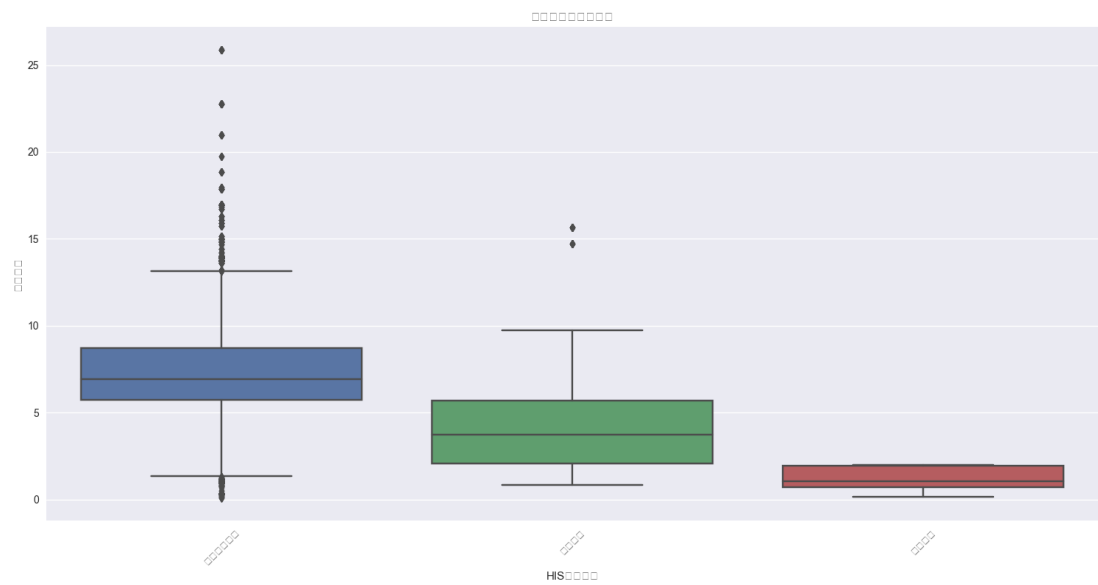


- 内分泌代谢科：
 - 作为专科科室，收治最多的糖尿病患者
 - 体现了专科化管理的优势
- 日间病房：

- 患者数量适中，以轻症和病情稳定患者为主
- 体现快速周转的特点
- 惠宾病房：
 - 患者数量相对较少
 - 提供个性化、高品质服务

2. 住院时间分析

2.1 住院天数分布

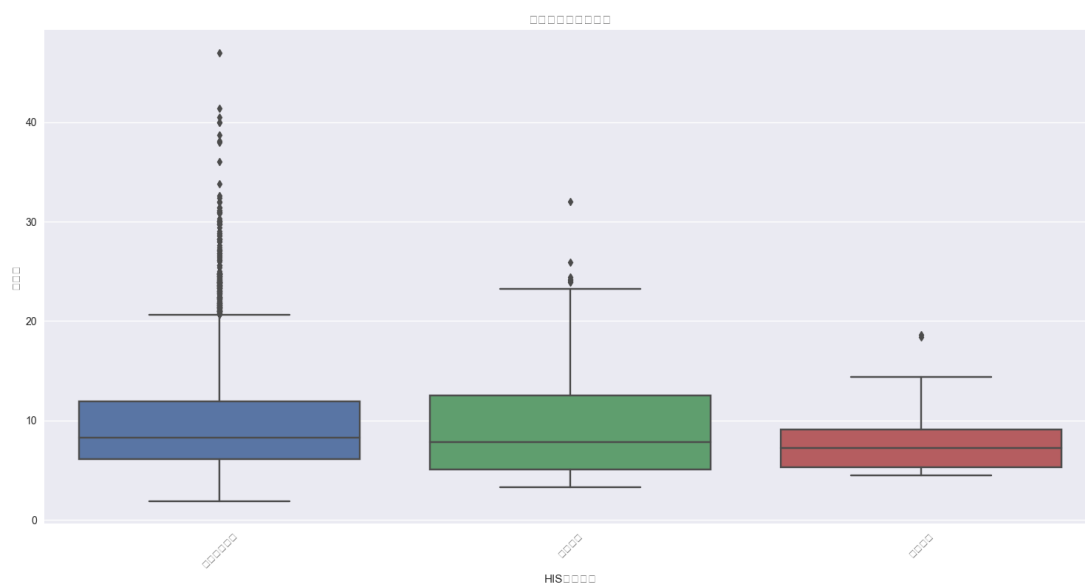


![住院天数分布]

- 科室间差异显著 ($F=791.0408$, $P<0.0001$)
- 各科室特点：
 - 日间病房：住院时间最短 (3-5 天)，标准化程度高
 - 内分泌科：住院时间中等 (7-10 天)，反映病情复杂度
 - 惠宾病房：住院时间最长 (>10 天)，治疗更个性化

3. 血糖控制分析

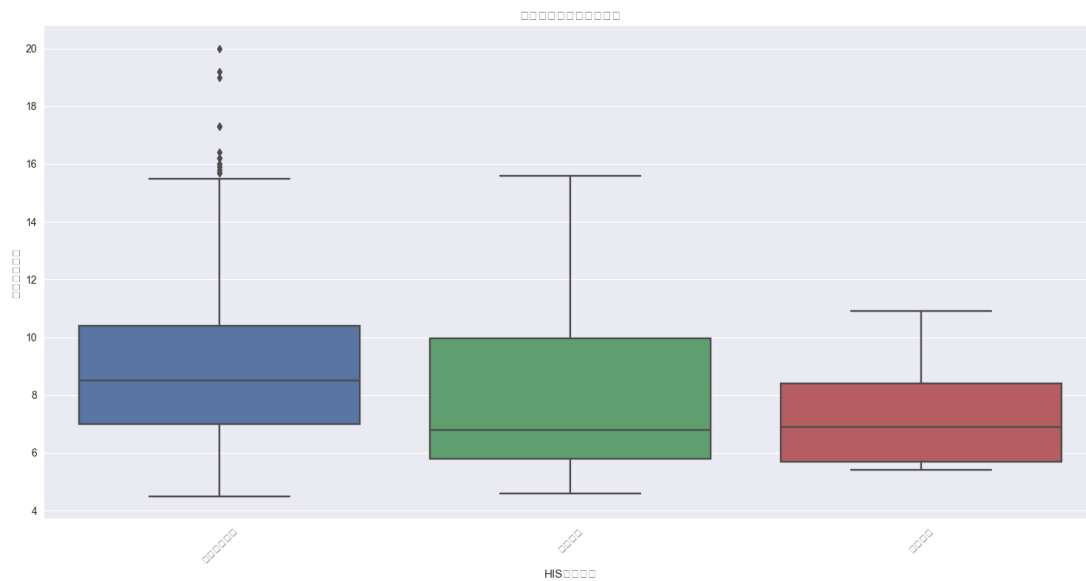
3.1 血糖水平分布



![血糖水平分布]

- 入院血糖水平：
 - 内分泌科：血糖水平普遍较高，病情较重
 - 日间病房：血糖相对稳定，适合短期住院
 - 惠宾病房：血糖水平分布广，个体差异大

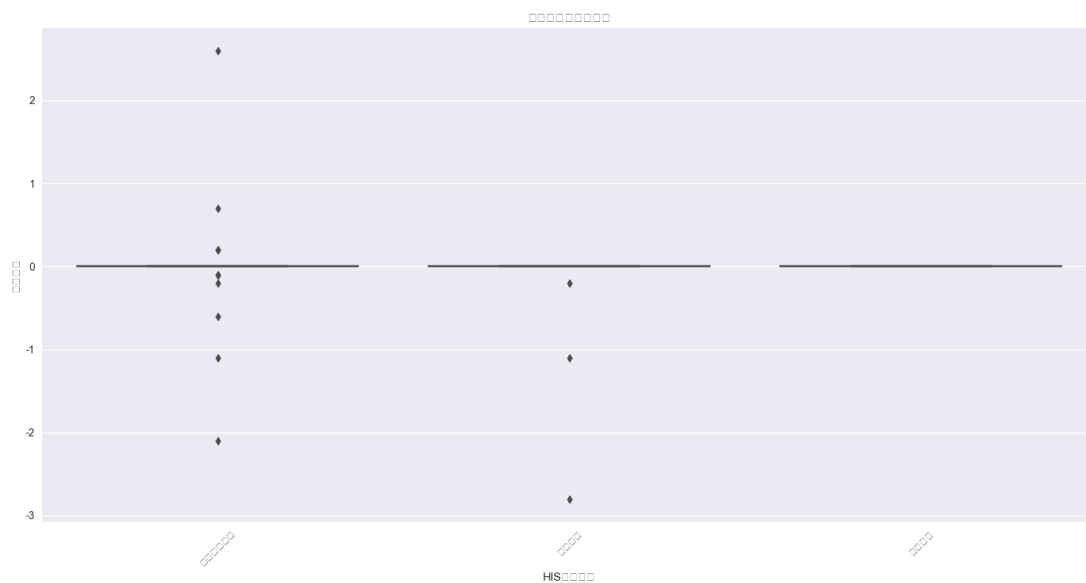
3.2 糖化血红蛋白分布



[糖化血红蛋白分布]

- 长期血糖控制情况：
 - 反映 3 个月内的平均血糖水平
 - 各科室间存在显著差异
 - 体现了不同患者群体的特点

3.3 血糖变化趋势

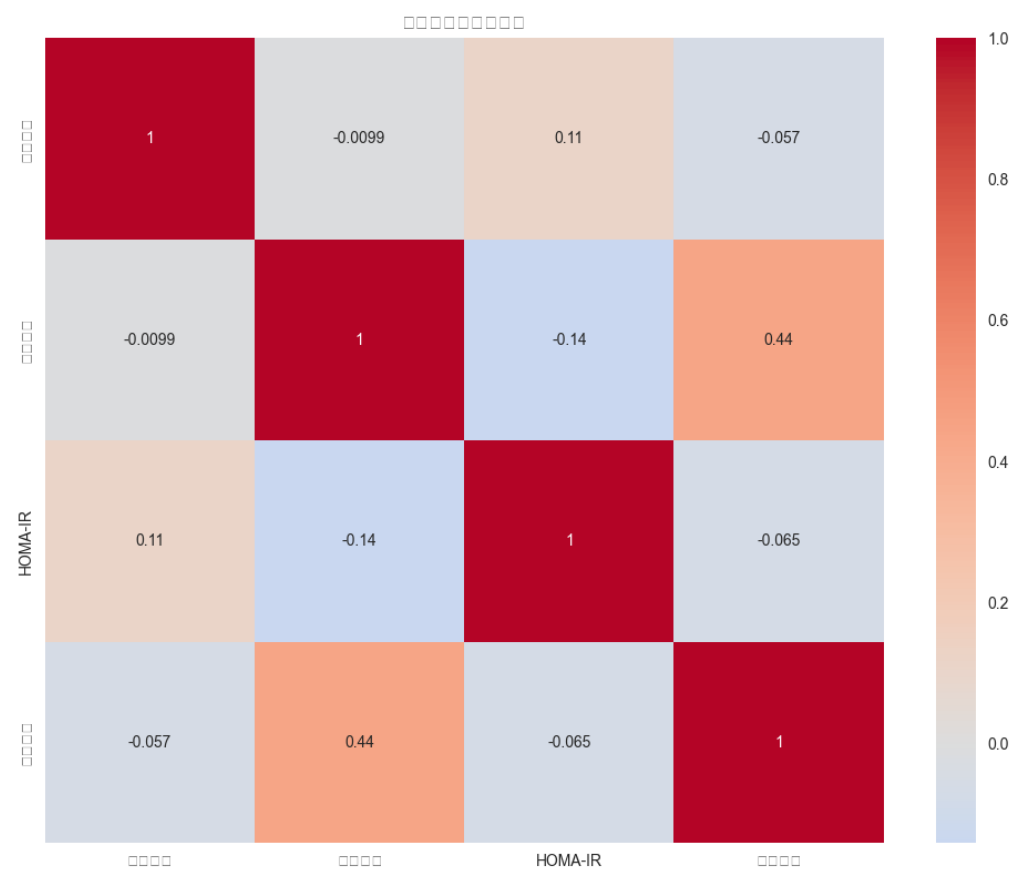


[血糖变化分布]

- 治疗效果：
 - 内分泌科：降糖效果最显著
 - 日间病房：变化适中，符合预期
 - 惠宾病房：个体差异明显

4. 相关性分析

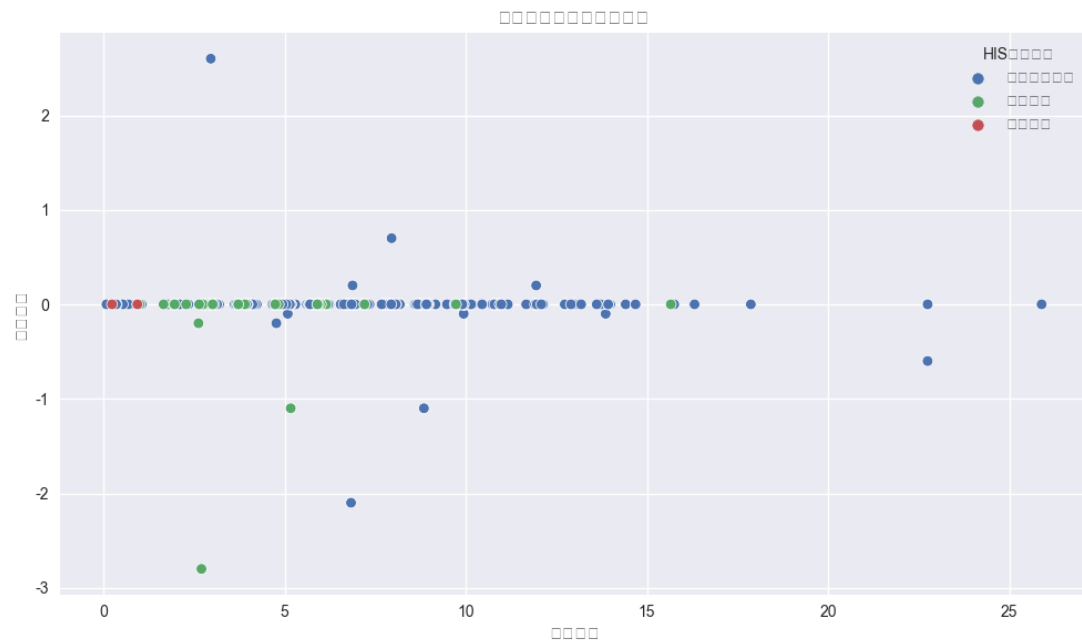
4.1 主要指标相关性



[指标相关性分析]

- 住院天数与治疗效果：
 - 存在一定相关性
 - 但并非简单的线性关系
 - 需要综合考虑多个因素

4.2 住院天数与血糖变化关系



[住院天数与血糖变化关系]

- 关系分析：
 - 不同科室呈现不同模式
 - 治疗时间并非越长越好
 - 需要找到最佳平衡点

5. 结论

通过分析可以看出，不同科室在治疗模式和效果上存在显著差异，这些差异既反映了专科化管理的特点，也体现了不同患者群体的需求。通过优化治疗流程、提高资源利用效率，可以进一步提升医疗质量，为患者提供更好的服务。