## 1. 报告摘要

研究目标：分析某省市万级别小区在 2021 年春节前后的流量与用户分布规律。

核心发现：简述识别出的主要小区画像（如办公区、住宅区）及其在节假日的行为特征。

实践意义：基于分析结论，为运营商提供基站扩容或节能的决策建议。

## 数据集概述与预处理

### 2.1 数据集介绍

数据规模：包含万级别小区的时序指标与属性数据 。

关键指标：业务流量（FLOW\_SUM）、用户数（USER\_COUNT）及小区 ID 。

静态属性：小区经纬度（LATITUDE/LONGITUDE）、覆盖类型（TYPE）和覆盖场景（SCENE）  
。

### 2.2 预处理流程

数据清洗：剔除流量或用户数为负的异常记录，处理异常值，处理缺失值，处理重复数据。

多表关联：利用 CELL\_ID 将时序数据与 attributes  
属性表进行左连接，补全场景标签 。

特征构造：根据 DATETIME\_KEY 构造"小时"、"星期"、"是否节假日"等关键特征。

## 3. 描述性统计分析

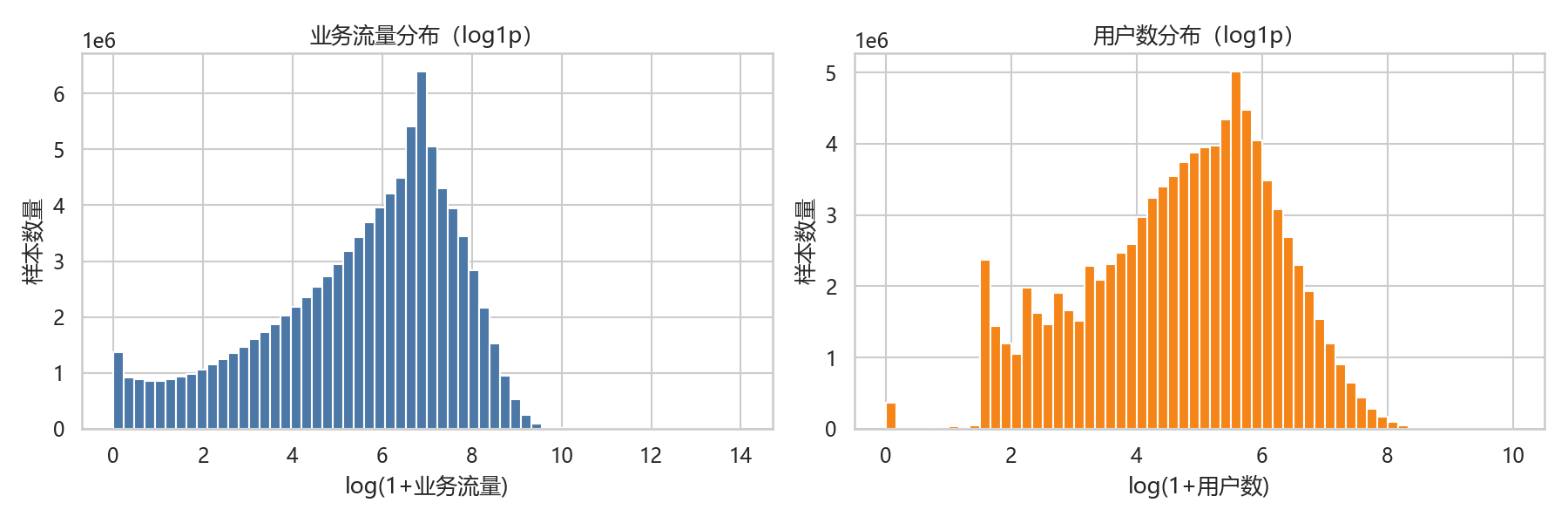
说明：本节必须使用 all\_final\_data\_with\_attributes.csv，该数据已包含属性字段与派生指标（日期、小时、星期、是否节假日、flow\_per\_user、PAR、ActivityScore 等）。在剔除负值、保留缺失为 NaN 的前提下完成统计与可视化。

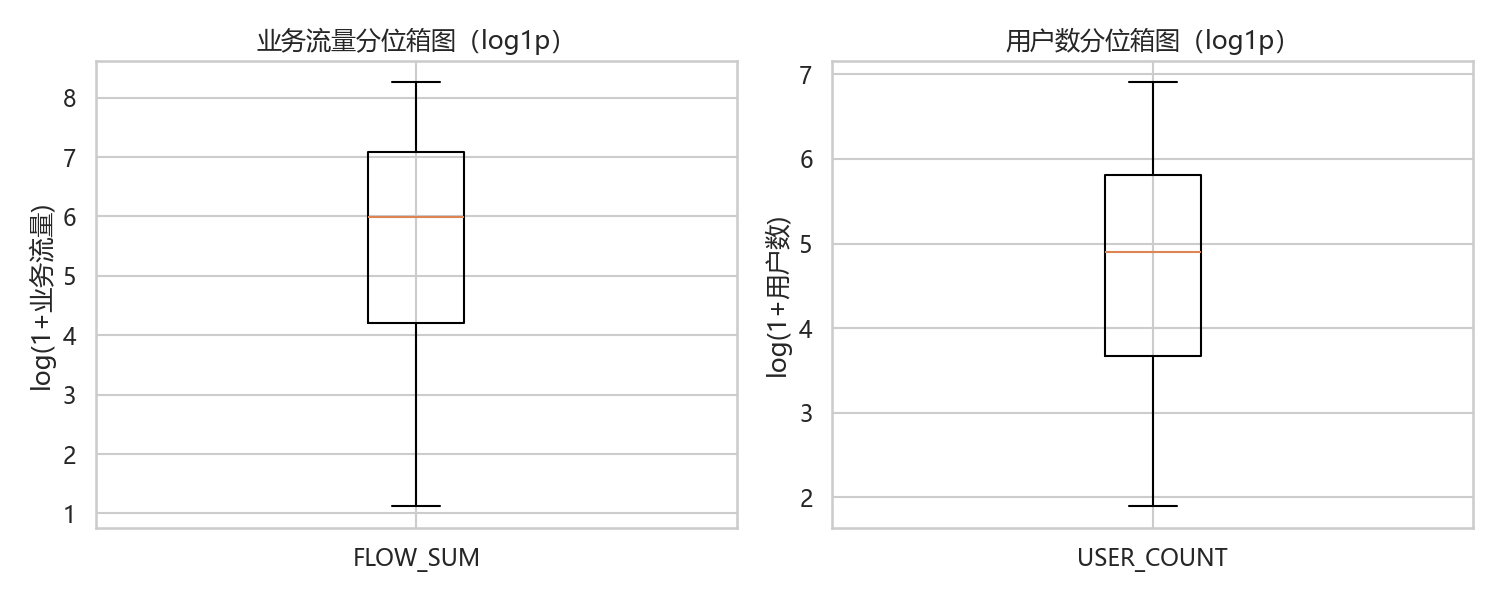
### 3.1 全局概览

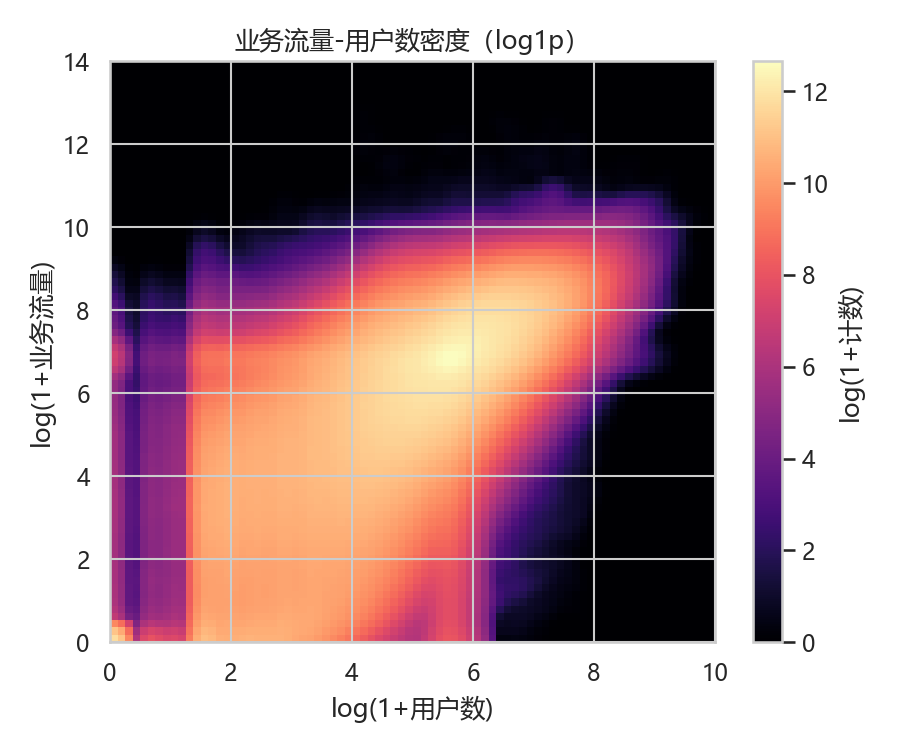
关键数据： total\_records=93,850,560; total\_flow=8.977e10 MB; date\_range=2021-02-09~2021-04-09 (60 days); flow\_mean=956.57, flow\_std=1493.66 (CV=1.56); user\_mean=269.12, user\_std=393.58 (CV=1.46); cell\_corr(flow\_sum,user\_sum)=0.669; daily\_max=1,797,291,580 (2021-04-09), daily\_min=1,073,917,202 (2021-02-15), diff=67.4%; cells=65174; type share: T1=38702 (59.4%), T0=25108 (38.5%), T2=1364 (2.1%); top scenes: S2=18544 (28.5%), S5=16185 (24.8%), S6=8746 (13.4%); top combos: S2-T1=14512 (22.3%), S5-T0=11400 (17.5%), S6-T1=8560 (13.1%).

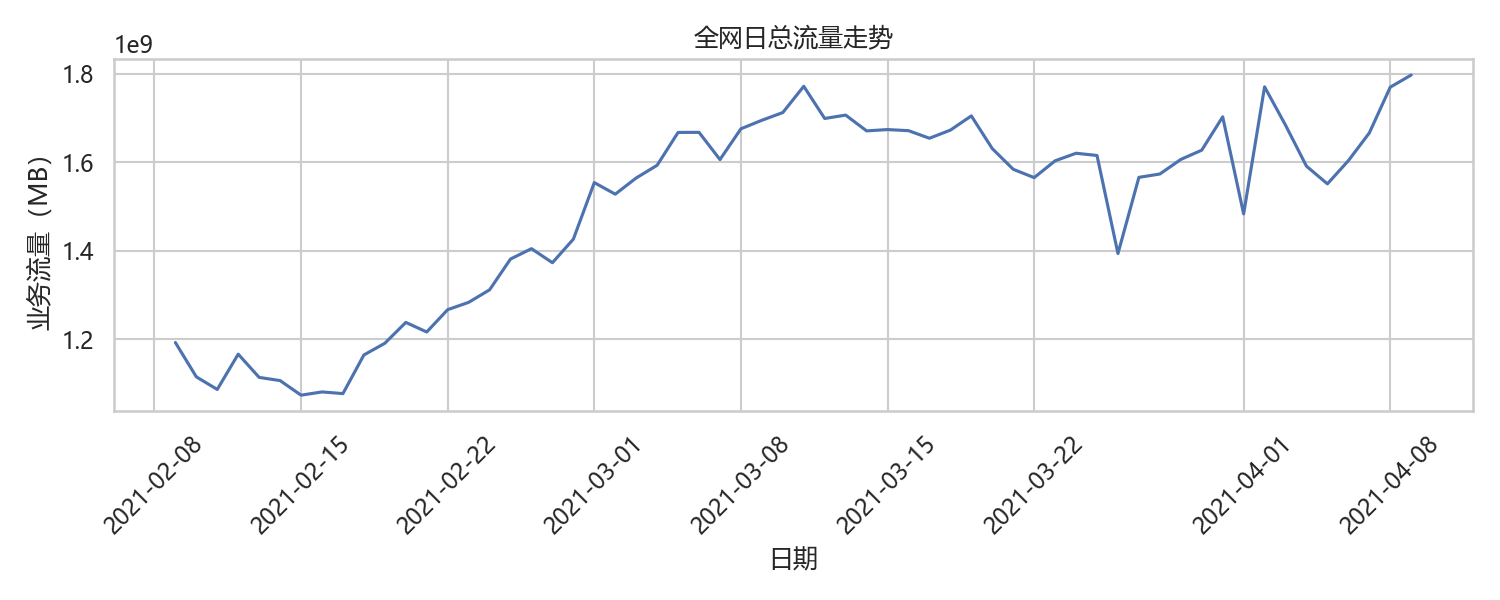
全量有效记录数为 93,850,560 条（FLOW\_SUM/USER\_COUNT 均为非负），核心统计量如下：

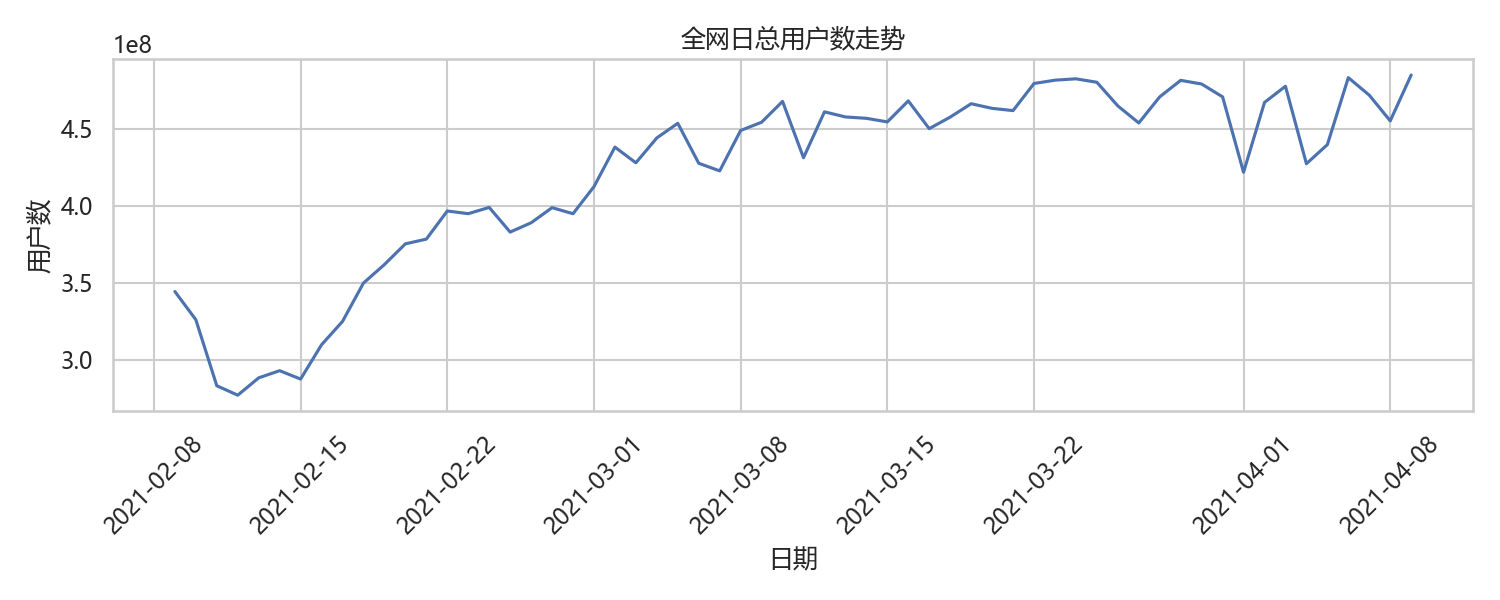
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 指标 | 均值 | 标准差 | 最小值 | 最大值 | 有效记录数 |
| 业务流量（FLOW\_SUM, MB） | 956.57 | 1493.66 | 0.00 | 329743.91 | 93,850,560 |
| 用户数（USER\_COUNT） | 269.12 | 393.58 | 0.00 | 15925.00 | 93,850,560 |
| 人均流量（flow\_per\_user, MB/人） | 6.30 | 18.07 | 0.00 | 7050.71 | 93,488,836 |
| PAR | 4.78 | 3.66 | 1.27 | 104.91 | 93,850,560 |
| ActivityScore | 0.0099 | 0.0138 | 0.0000 | 0.5419 | 93,850,560 |

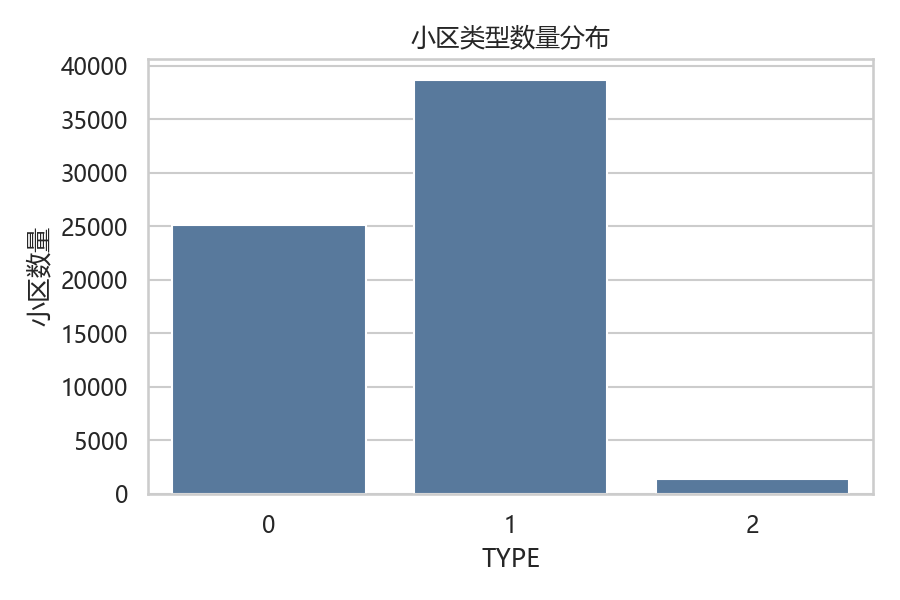
  
含义：该图展示业务流量与用户数分布。  
意义：业务流量与用户数分布, 判断集中程度、偏态与长尾特征。  
用法：业务流量与用户数分布, 用于设置阈值、识别异常或分层。

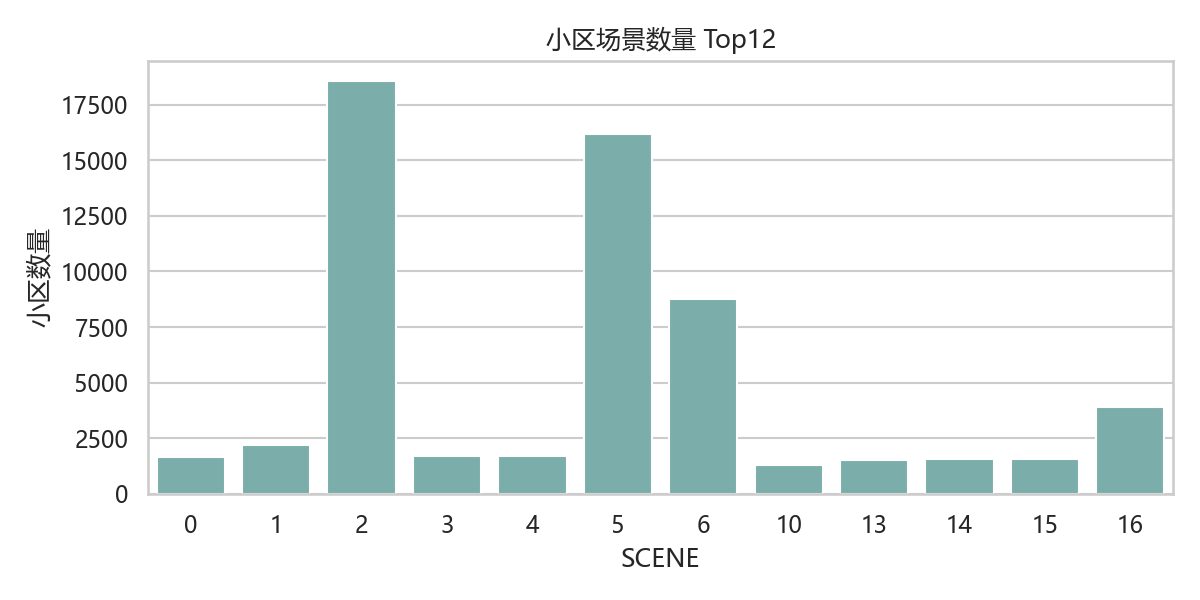
  
含义：该图展示业务流量与用户数箱线图。  
意义：业务流量与用户数箱线图, 刻画分位区间与离散程度，便于稳健比较。  
用法：业务流量与用户数箱线图, 比较中位数与波动，识别高波动或异常组。

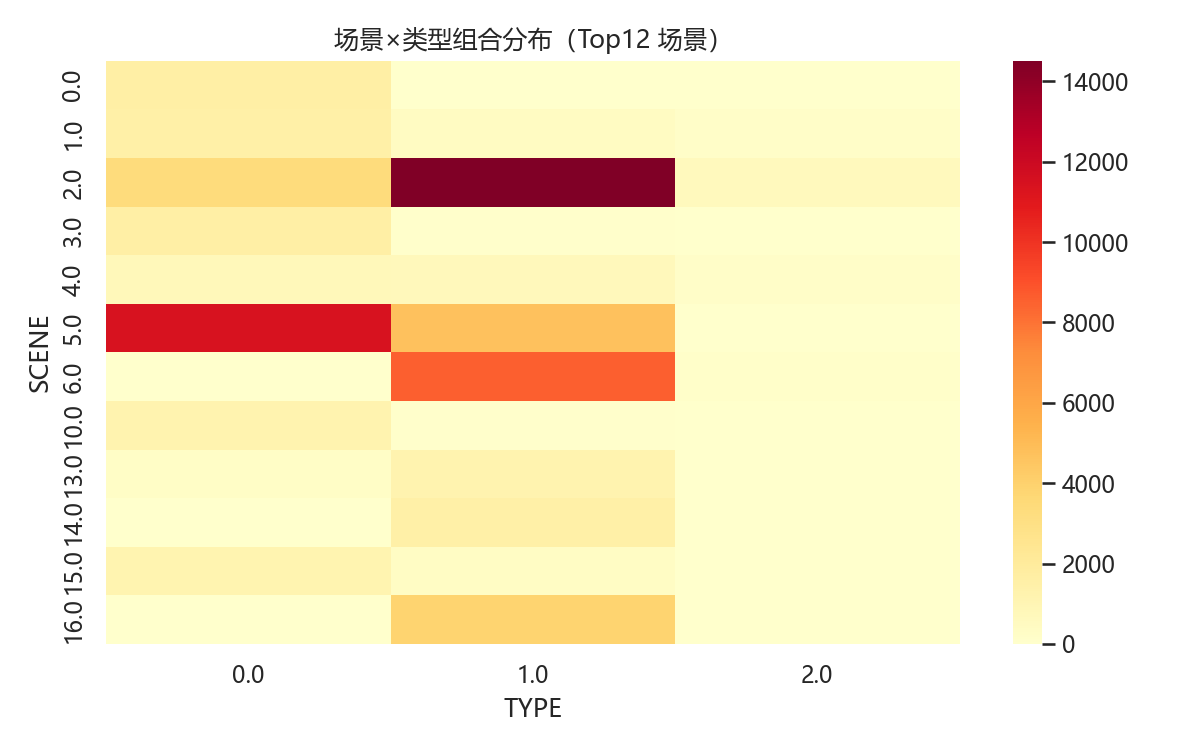
  
含义：该图展示业务流量-用户数密度。  
意义：业务流量-用户数密度, 提供整体结构与差异的直观视角。  
用法：业务流量-用户数密度, 辅助定位重点区间与异常样本。

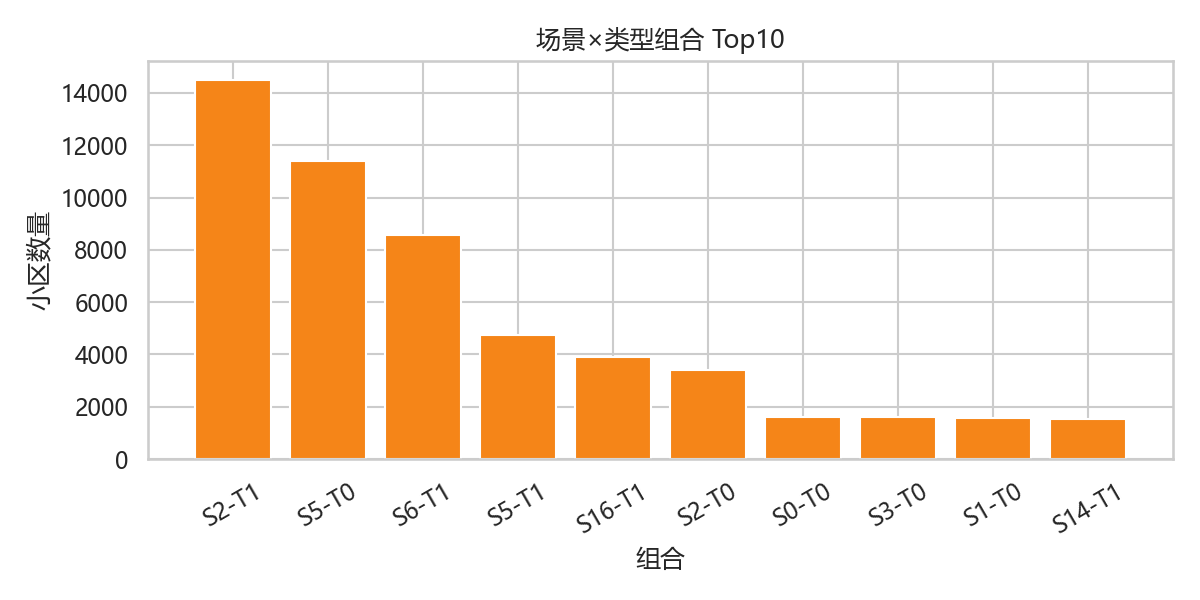
  
含义：该图展示全网日总流量走势。  
意义：全网日总流量走势, 刻画随时间的变化与周期性规律。  
用法：全网日总流量走势, 识别峰谷与突变点，支持容量与资源配置。

  
含义：该图展示全网日总用户数走势。  
意义：全网日总用户数走势, 刻画随时间的变化与周期性规律。  
用法：全网日总用户数走势, 识别峰谷与突变点，支持容量与资源配置。

  
含义：该图展示小区类型数量分布。  
意义：小区类型数量分布, 量化不同 TYPE 的规模差异。  
用法：小区类型数量分布, 用于资源分层与对比分析基线。

  
含义：该图展示小区场景数量 Top12。  
意义：小区场景数量 Top12, 揭示场景集中度与主体来源。  
用法：小区场景数量 Top12, 作为场景分层与对比优先级依据。

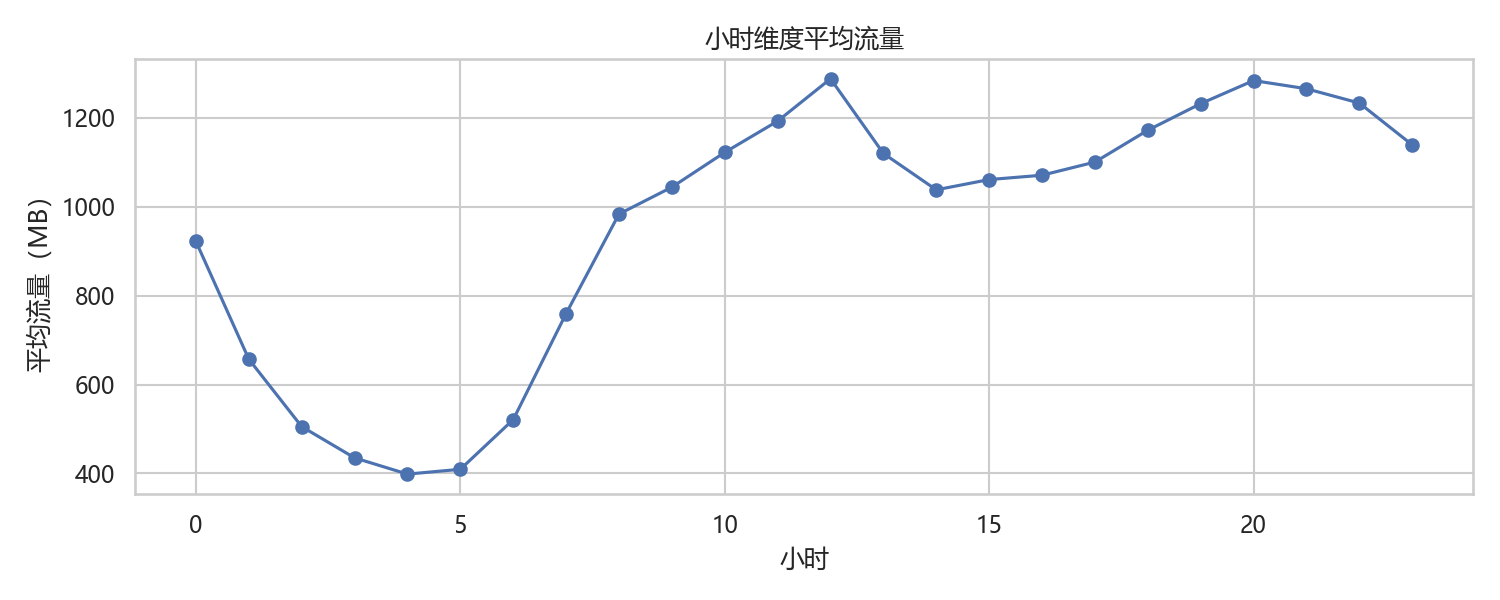
  
含义：该图展示场景×类型组合分布（Top12 场景）。  
意义：场景×类型组合分布, 显示场景与类型的耦合结构。  
用法：场景×类型组合分布, 识别主导组合与稀缺组合。

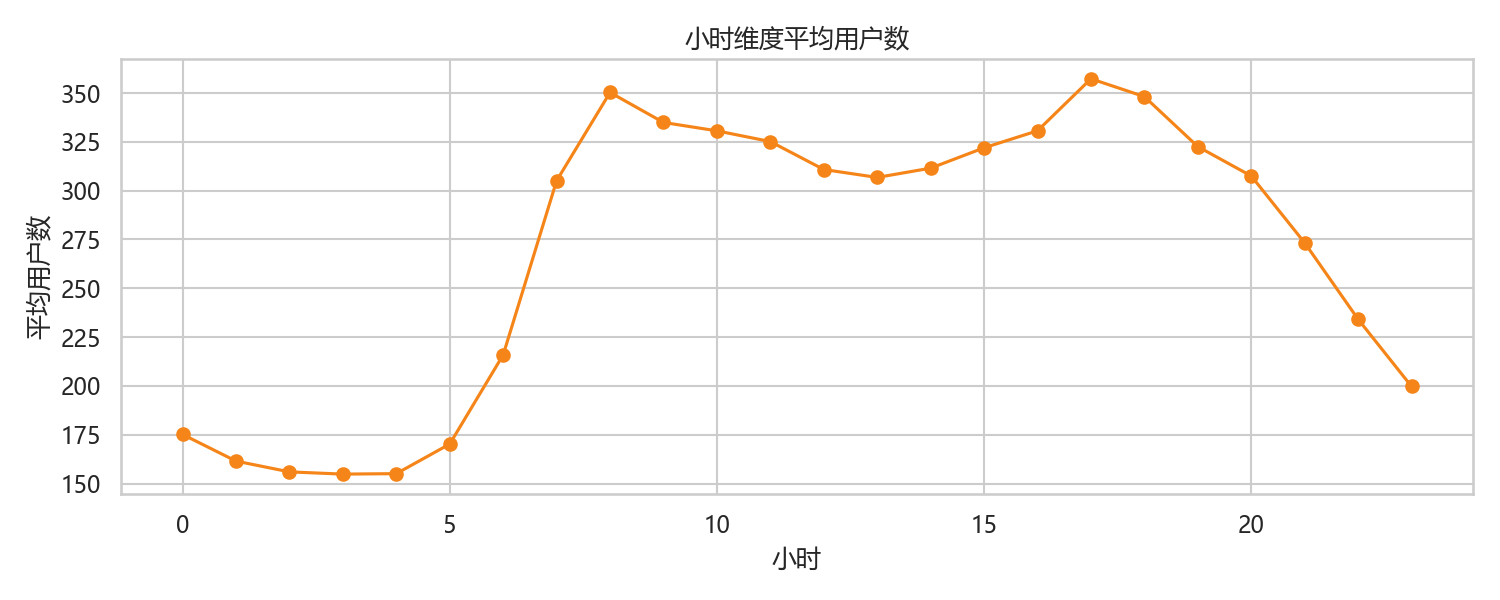
  
含义：该图展示场景×类型组合 Top10。  
意义：场景×类型组合 Top10, 量化头部组合的规模优势。  
用法：场景×类型组合 Top10, 作为资源倾斜与精细化管理目标。

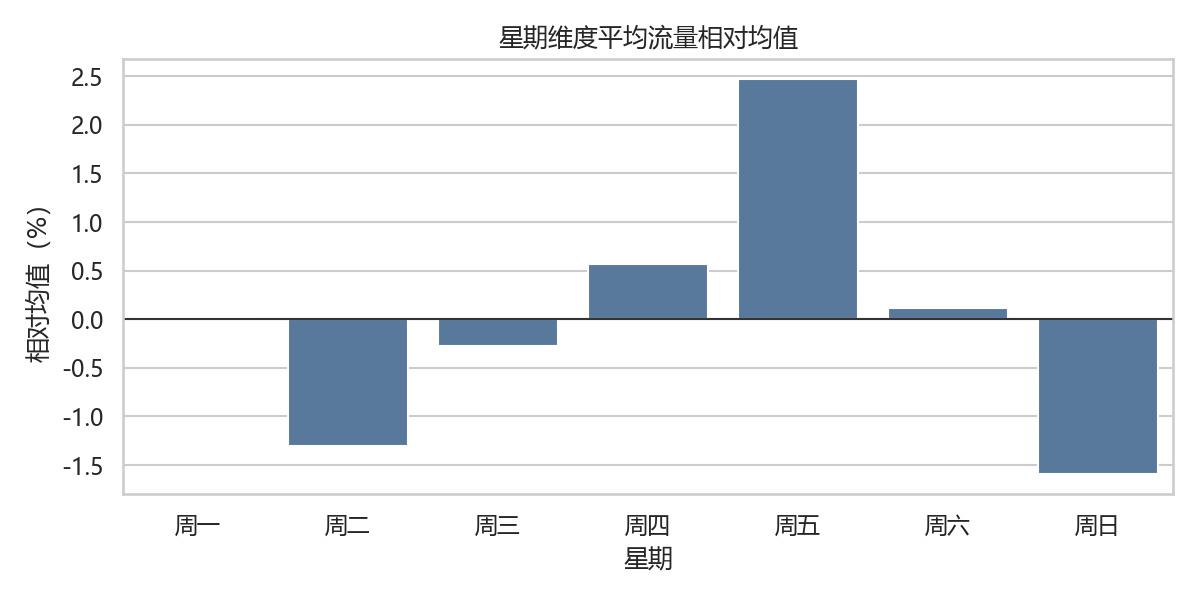
### 3.2 时间维度与节假日差异

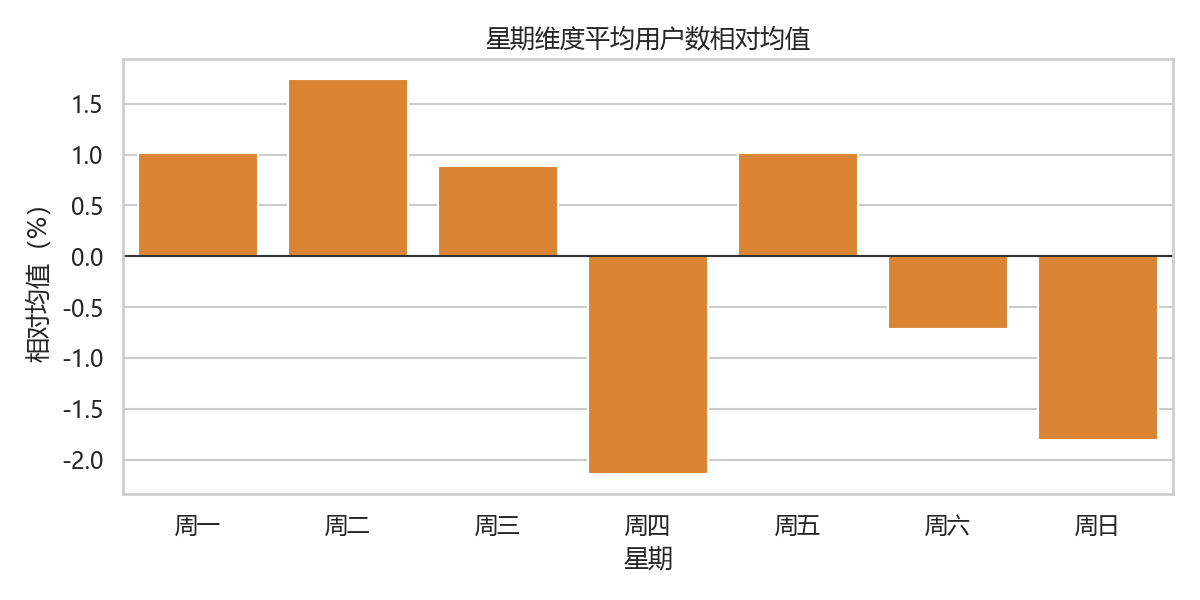
关键数据： hourly flow peak=1288.04 @12, trough=398.60 @4 (peak/trough=3.23); hourly user peak=357.30 @17, trough=154.82 @3 (peak/trough=2.31); weekday delta vs mean: flow Fri +2.47%, Sun -1.59%; user Thu -2.14%, Sun -1.81%; holiday daily mean: flow 1.253e9 vs 1.545e9 (-18.9%), user 3.41e8 vs 4.37e8 (-22.0%).

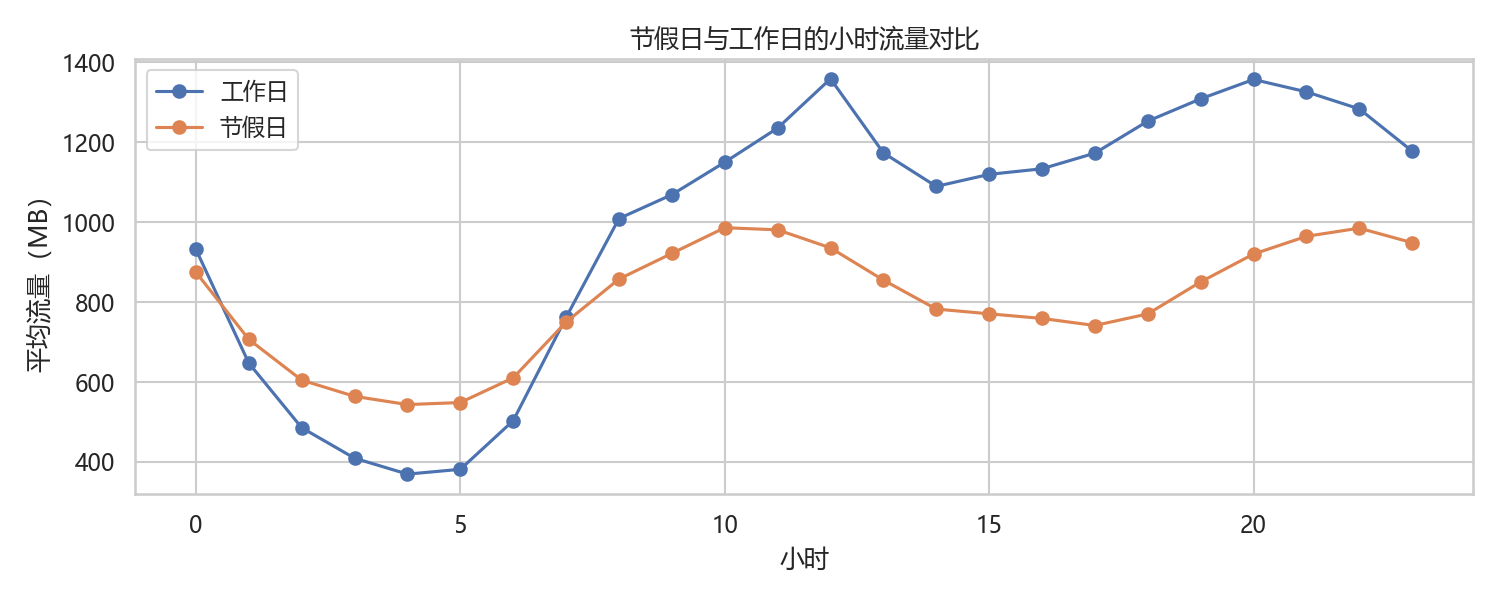
从小时、星期与节假日维度观察网络潮汐规律（节假日按 2021 年法定假日表重新标注），并用热力图揭示“小时-星期”交互特征。

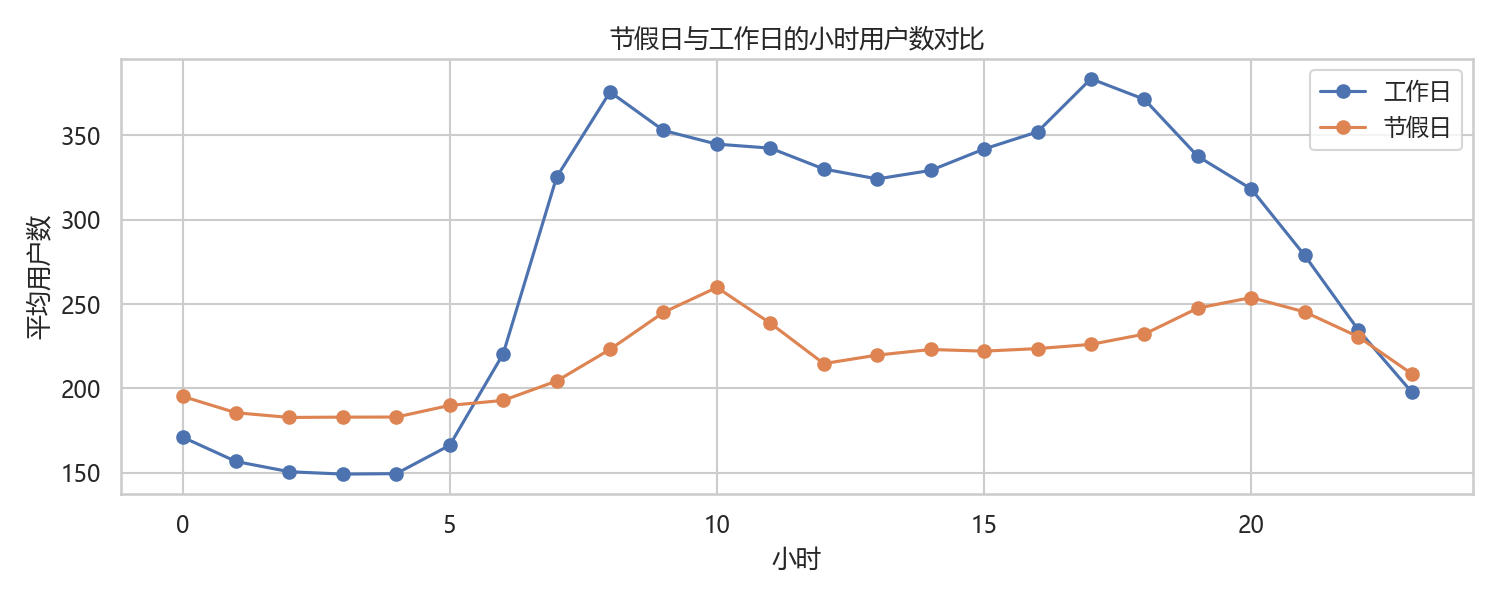
  
含义：该图展示小时维度平均流量。  
意义：小时维度平均流量, 提供整体结构与差异的直观视角。  
用法：小时维度平均流量, 辅助定位重点区间与异常样本。

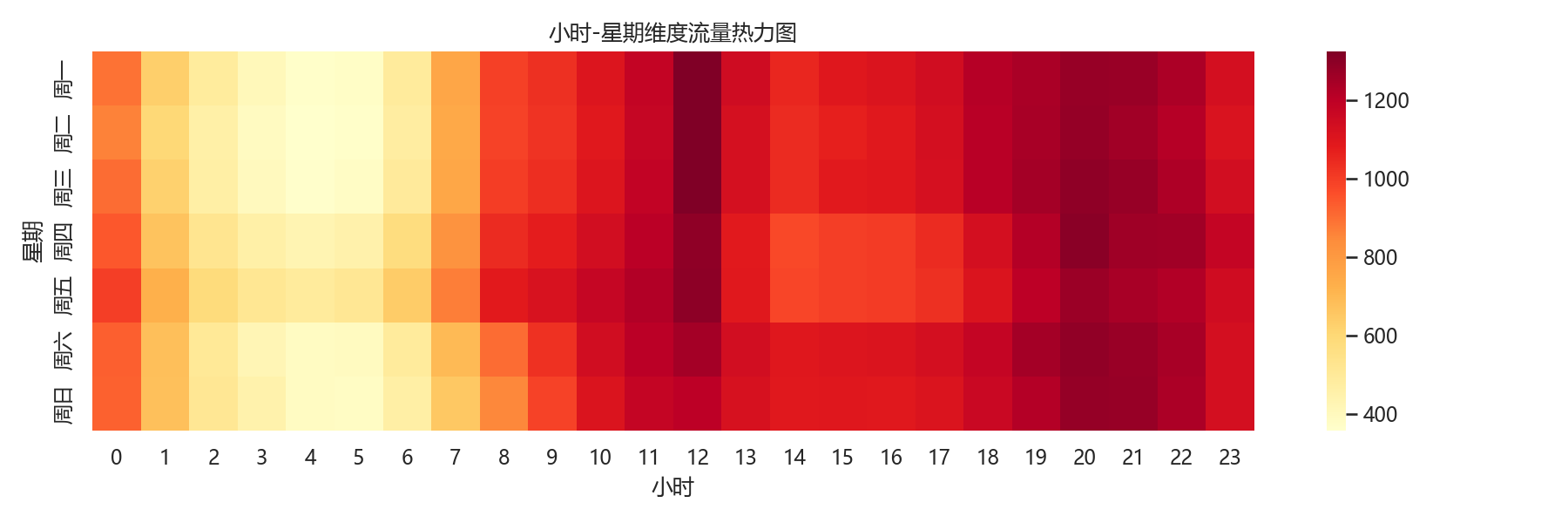
  
含义：该图展示小时维度平均用户数。  
意义：小时维度平均用户数, 提供整体结构与差异的直观视角。  
用法：小时维度平均用户数, 辅助定位重点区间与异常样本。

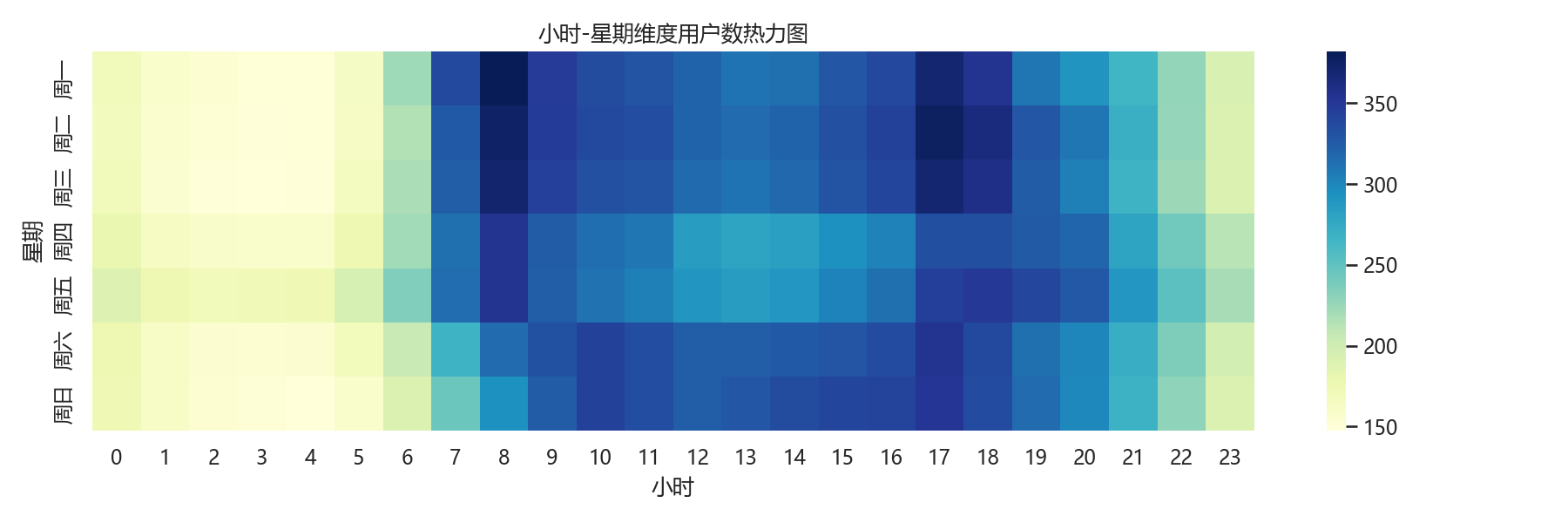
  
含义：该图展示星期维度平均流量。  
意义：星期维度平均流量, 提供整体结构与差异的直观视角。  
用法：星期维度平均流量, 辅助定位重点区间与异常样本。

  
含义：该图展示星期维度平均用户数。  
意义：星期维度平均用户数, 提供整体结构与差异的直观视角。  
用法：星期维度平均用户数, 辅助定位重点区间与异常样本。

  
含义：该图展示节假日与工作日的小时流量对比。  
意义：节假日与工作日的小时流量对比, 量化节假日与工作日差异，体现节假日效应。  
用法：节假日与工作日的小时流量对比, 用于节假日保障与弹性资源配置。

  
含义：该图展示节假日与工作日的小时用户数对比。  
意义：节假日与工作日的小时用户数对比, 量化节假日与工作日差异，体现节假日效应。  
用法：节假日与工作日的小时用户数对比, 用于节假日保障与弹性资源配置。

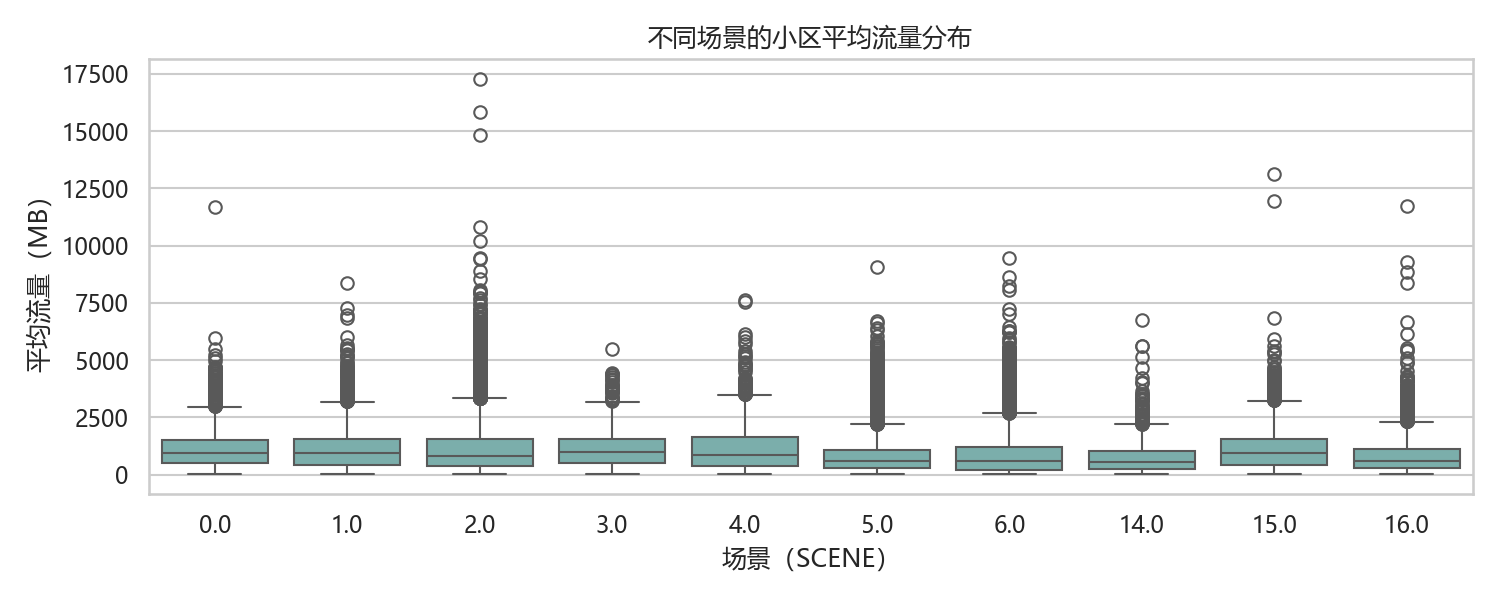
  
含义：该图展示小时-星期维度流量热力图。  
意义：小时-星期维度流量热力图, 揭示不同维度组合下的强度分布与热点区域。  
用法：小时-星期维度流量热力图, 定位高值/低值区间，用于时段、场景或类型优化。

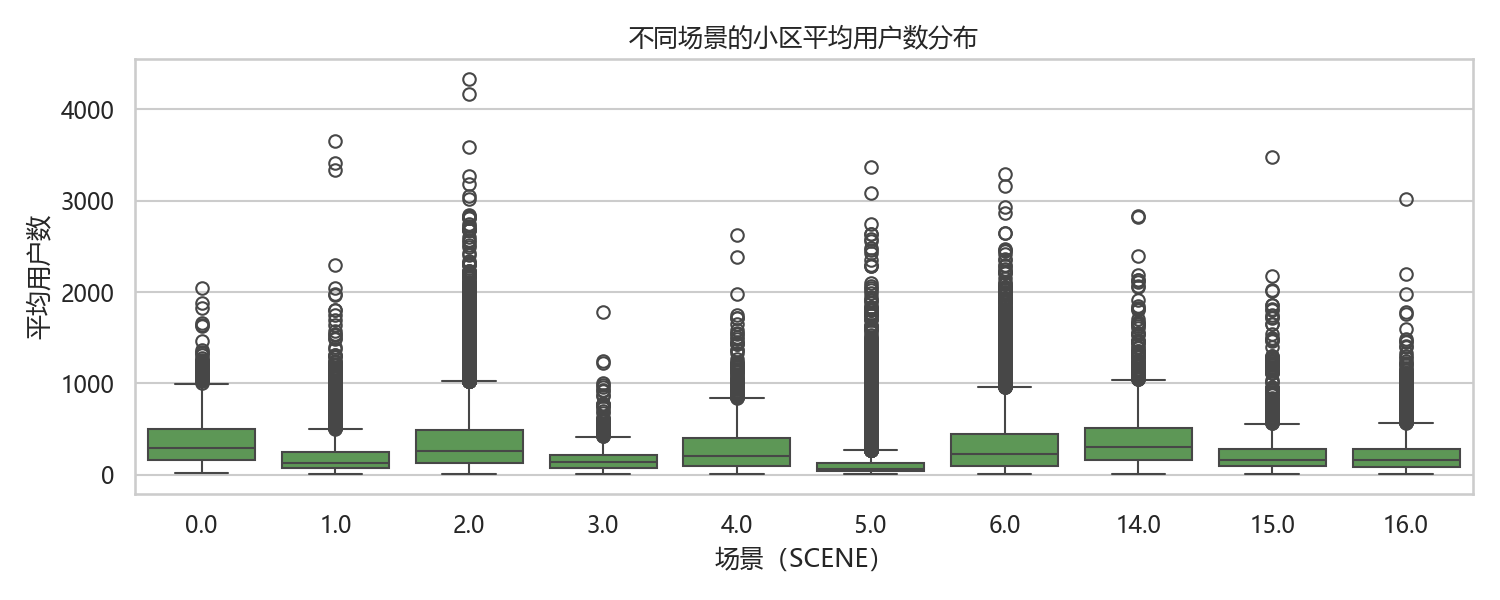
  
含义：该图展示小时-星期维度用户数热力图。  
意义：小时-星期维度用户数热力图, 揭示不同维度组合下的强度分布与热点区域。  
用法：小时-星期维度用户数热力图, 定位高值/低值区间，用于时段、场景或类型优化。

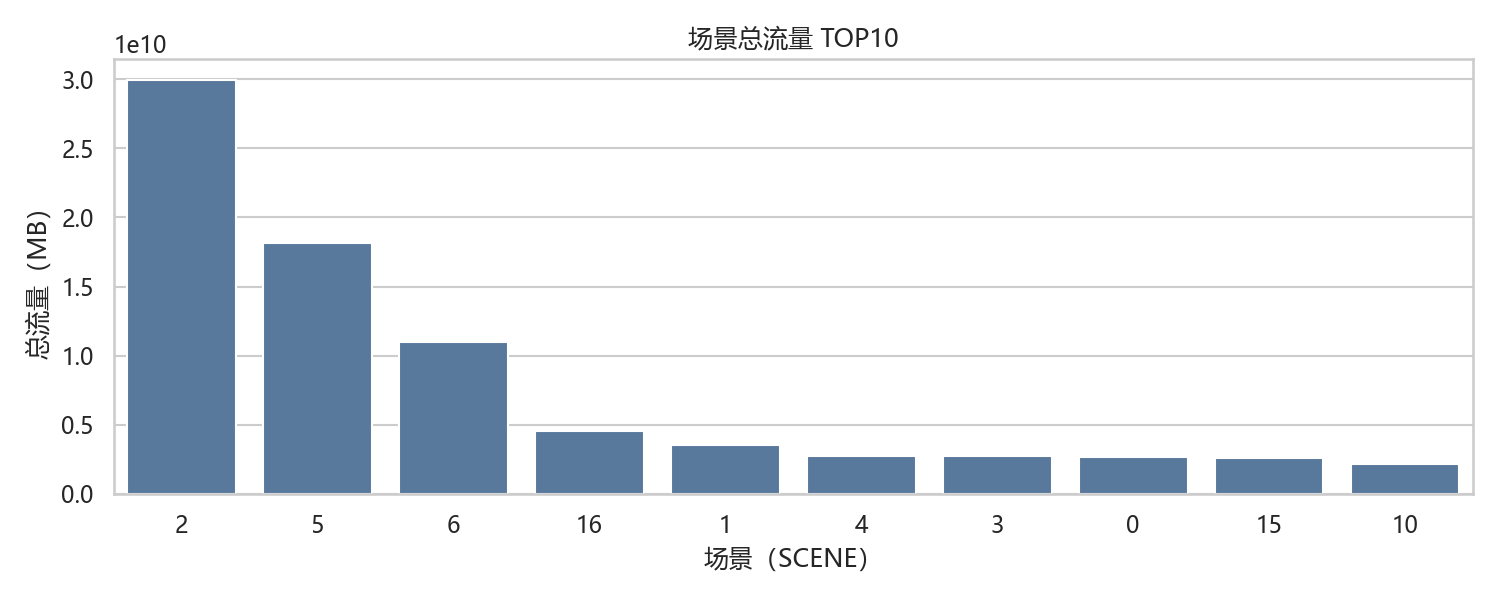
### 3.3 场景与类型差异

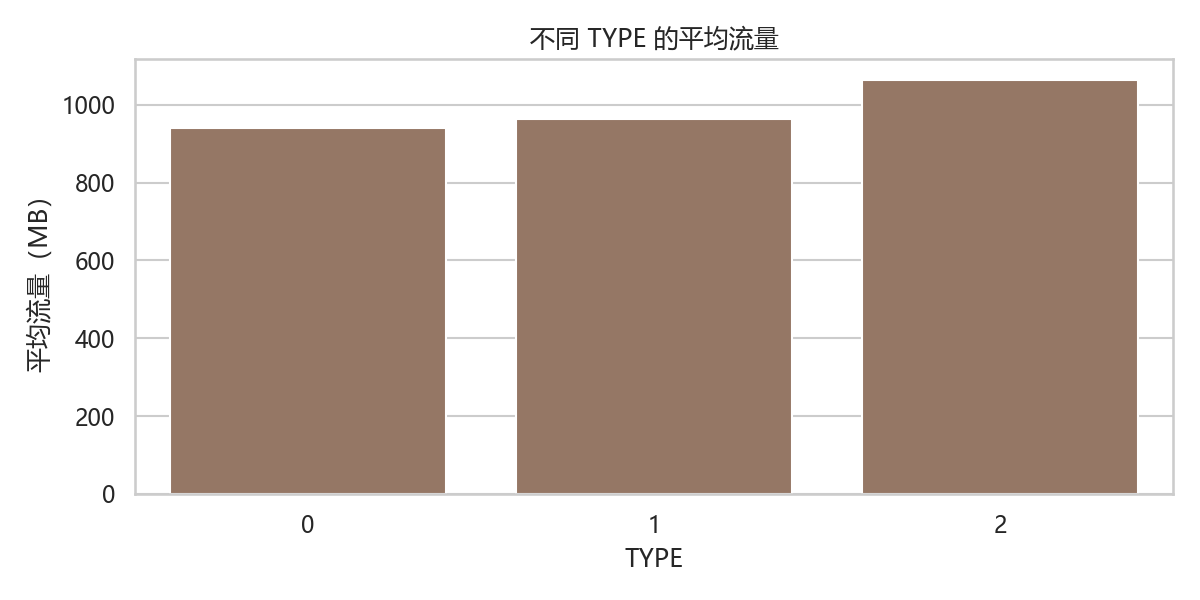
关键数据： scene share of total flow: SCENE2=33.38%, SCENE5=20.20%, SCENE6=12.24% (top3=65.8%); type means: TYPE2 flow\_mean=1064.74, TYPE0 flow\_per\_user=8.67 (highest).

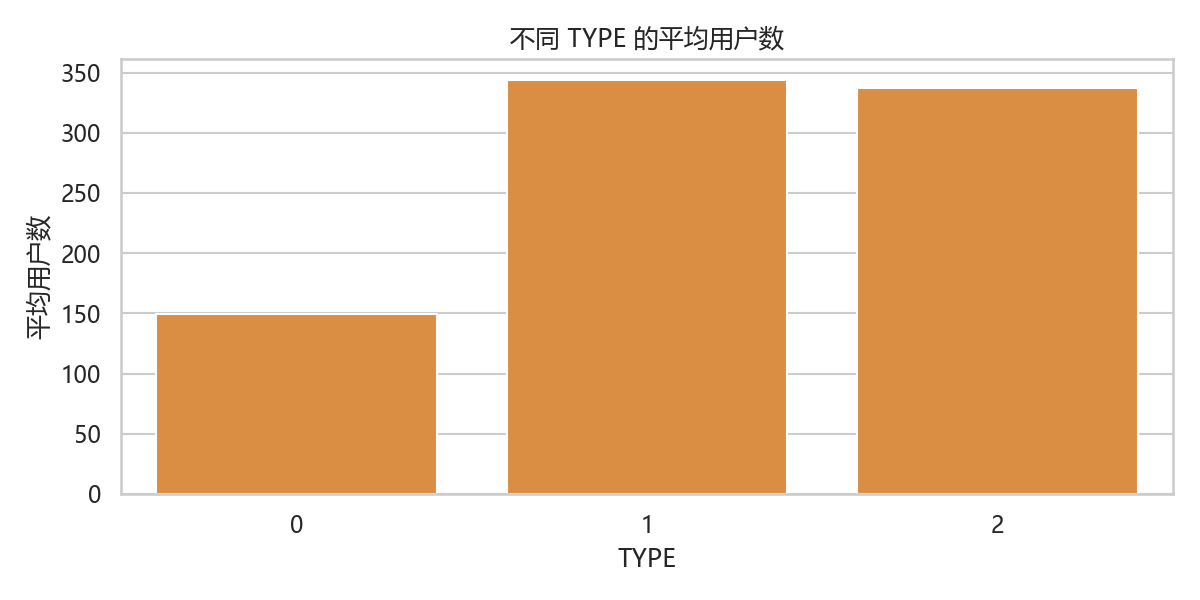
按场景（SCENE）与类型（TYPE）分析小区平均水平和总量结构，识别核心价值场景。

  
含义：该图展示不同场景的小区平均流量分布。  
意义：不同场景的小区平均流量分布, 判断集中程度、偏态与长尾特征。  
用法：不同场景的小区平均流量分布, 用于设置阈值、识别异常或分层。

  
含义：该图展示不同场景的小区平均用户数分布。  
意义：不同场景的小区平均用户数分布, 判断集中程度、偏态与长尾特征。  
用法：不同场景的小区平均用户数分布, 用于设置阈值、识别异常或分层。

  
含义：该图展示场景总流量 TOP10。  
意义：场景总流量 TOP10, 揭示头部集中度与贡献结构。  
用法：场景总流量 TOP10, 确定重点对象或场景优先级。

  
含义：该图展示不同 TYPE 的平均流量。  
意义：不同 TYPE 的平均流量, 提供整体结构与差异的直观视角。  
用法：不同 TYPE 的平均流量, 辅助定位重点区间与异常样本。

  
含义：该图展示不同 TYPE 的平均用户数。  
意义：不同 TYPE 的平均用户数, 提供整体结构与差异的直观视角。  
用法：不同 TYPE 的平均用户数, 辅助定位重点区间与异常样本。

### 3.4 TOP 分析

关键数据： top10 flow sum=189,914,493 MB (0.21% of total); top1% share=5.57%; top10 flow\_per\_user all in SCENE=5, TYPE=0.

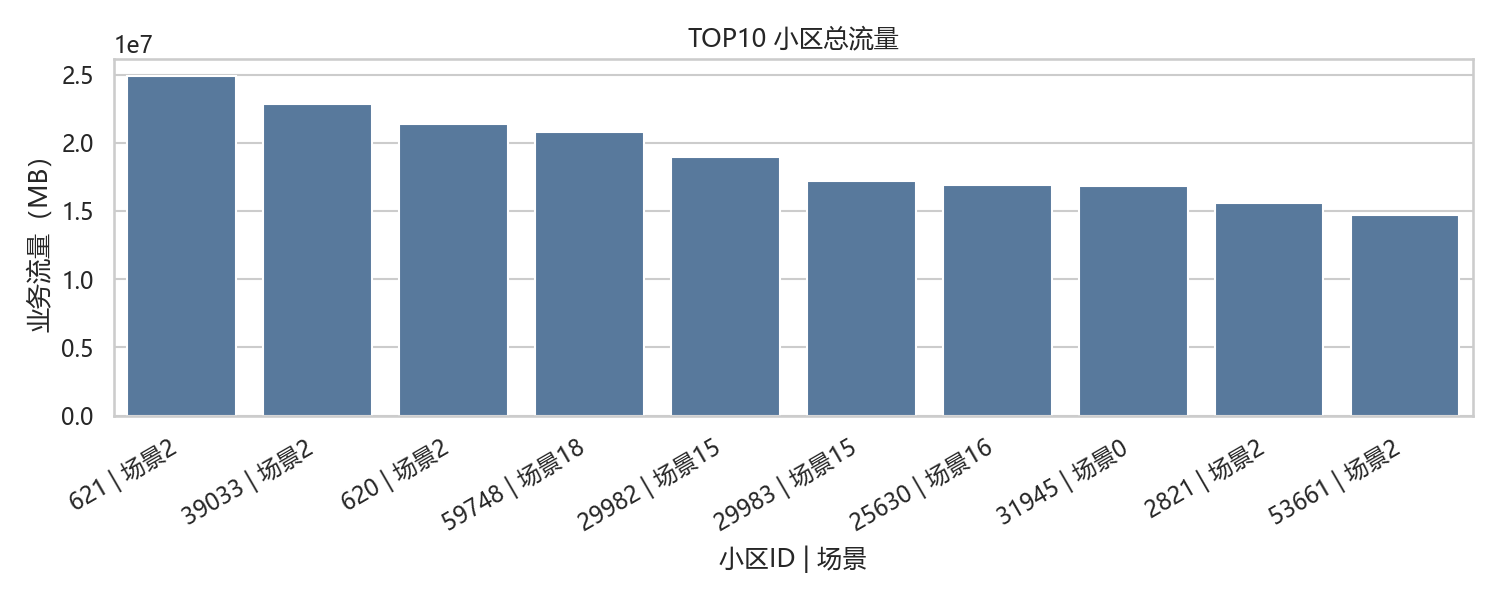
从小区总流量与人均流量两个维度进行排名，识别核心贡献与高价值小区。

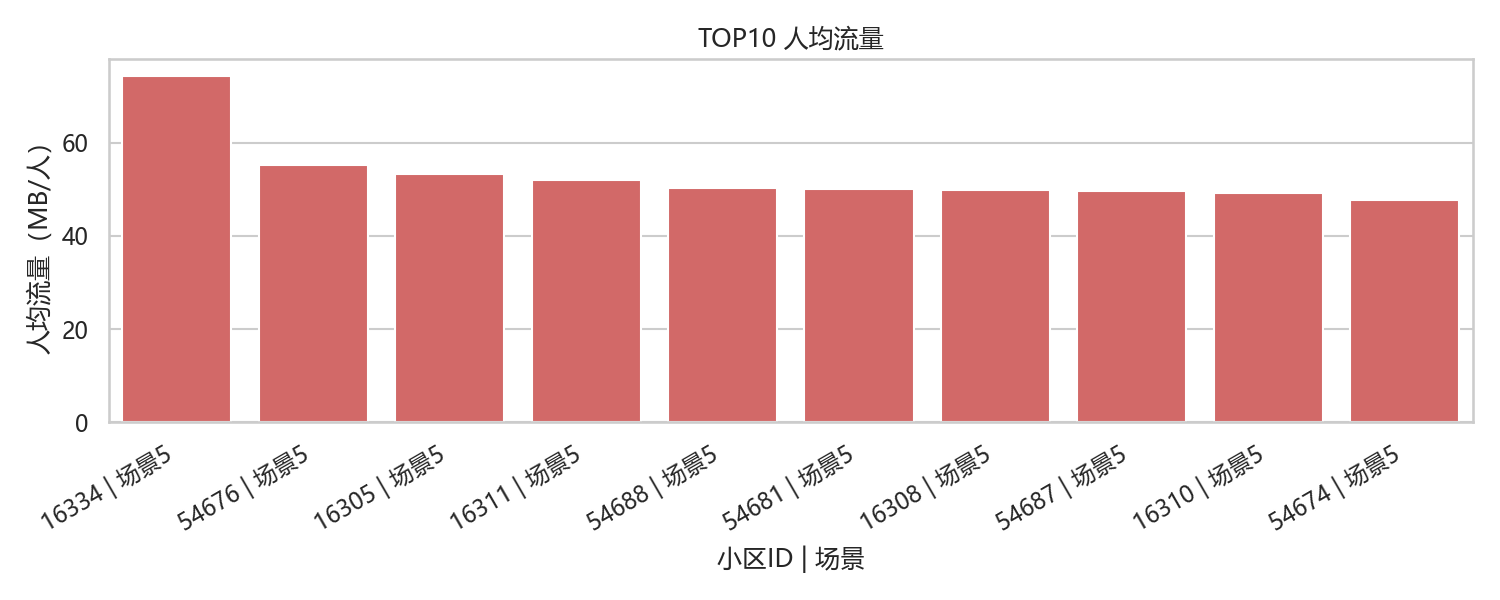
TOP10 小区总流量（单位：百万 MB）：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 排名 | 小区ID | 场景 | 类型 | 总流量（百万MB） |
| 1 | 621 | 2 | 1 | 24.89 |
| 2 | 39033 | 2 | 1 | 22.82 |
| 3 | 620 | 2 | 1 | 21.36 |
| 4 | 59748 | 18 | 1 | 20.81 |
| 5 | 29982 | 15 | 0 | 18.93 |
| 6 | 29983 | 15 | 0 | 17.18 |
| 7 | 25630 | 16 | 1 | 16.87 |
| 8 | 31945 | 0 | 0 | 16.83 |
| 9 | 2821 | 2 | 1 | 15.55 |
| 10 | 53661 | 2 | 1 | 14.68 |

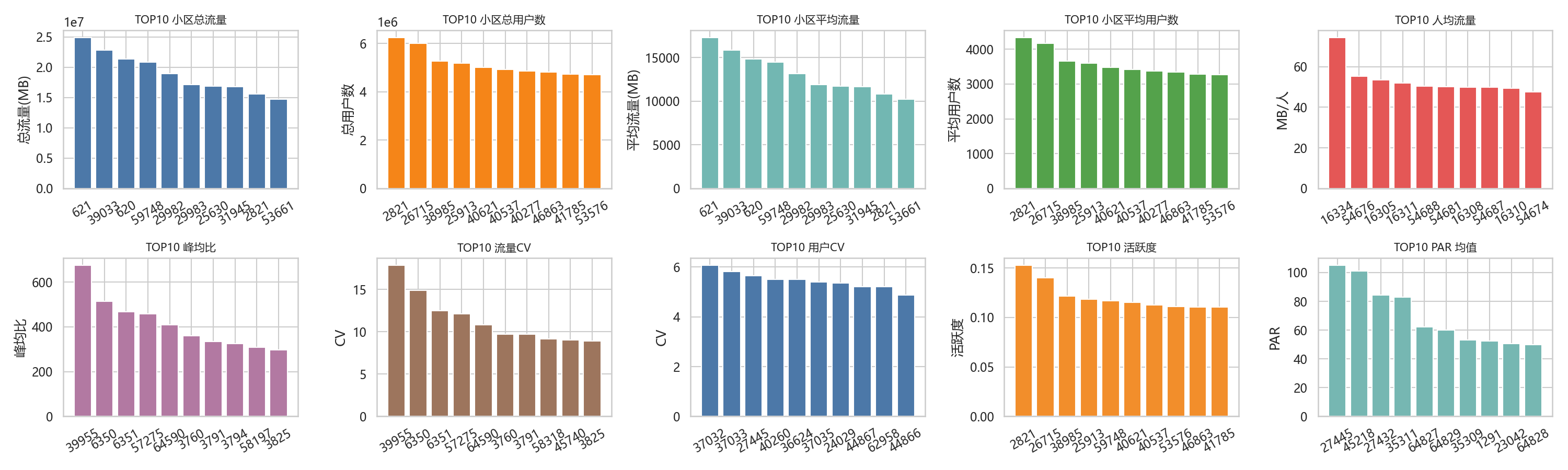
TOP10 人均流量（单位：MB/人）：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 排名 | 小区ID | 场景 | 类型 | 人均流量（MB/人） |
| 1 | 16334 | 5 | 0 | 74.29 |
| 2 | 54676 | 5 | 0 | 55.25 |
| 3 | 16305 | 5 | 0 | 53.40 |
| 4 | 16311 | 5 | 0 | 51.92 |
| 5 | 54688 | 5 | 0 | 50.29 |
| 6 | 54681 | 5 | 0 | 50.19 |
| 7 | 16308 | 5 | 0 | 49.84 |
| 8 | 54687 | 5 | 0 | 49.74 |
| 9 | 16310 | 5 | 0 | 49.26 |
| 10 | 54674 | 5 | 0 | 47.65 |

  
含义：该图展示TOP10 小区总流量。  
意义：TOP10 小区总流量, 揭示头部集中度与贡献结构。  
用法：TOP10 小区总流量, 确定重点对象或场景优先级。

  
含义：该图展示TOP10 人均流量。  
意义：TOP10 人均流量, 揭示头部集中度与贡献结构。  
用法：TOP10 人均流量, 确定重点对象或场景优先级。

补充绘制 10 项关键指标的 TOP10 网格图，从多指标视角识别头部小区结构与差异。

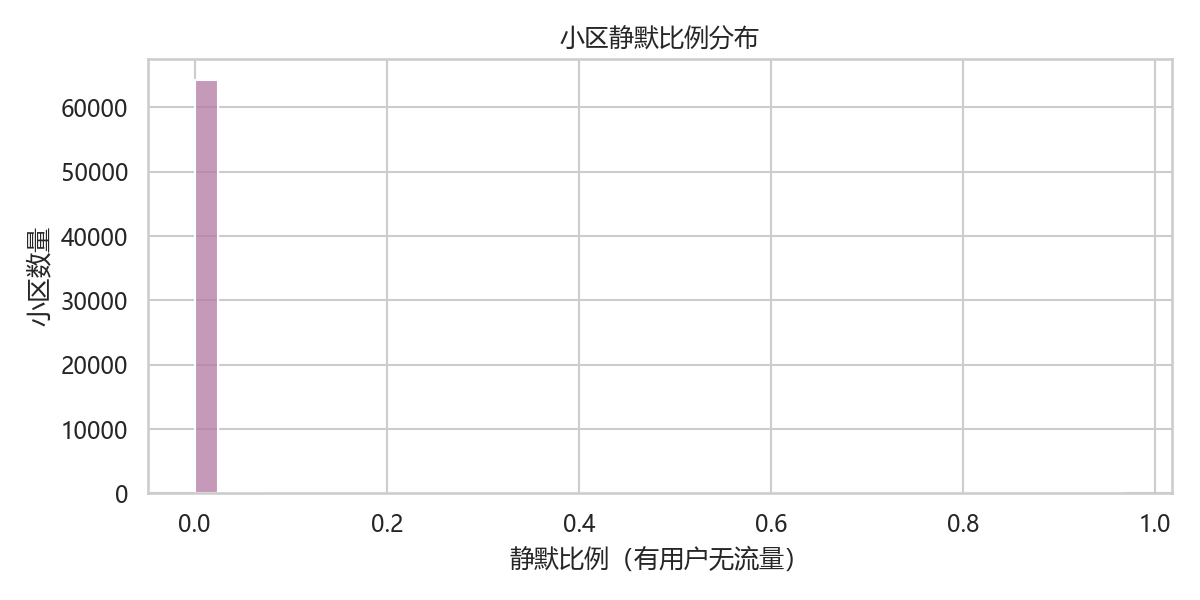
  
含义：该图展示TOP 指标网格对比。  
意义：TOP 指标网格对比, 便于横向比较不同指标的头部小区差异。  
用法：TOP 指标网格对比, 用于综合评估重点小区与资源优先级。

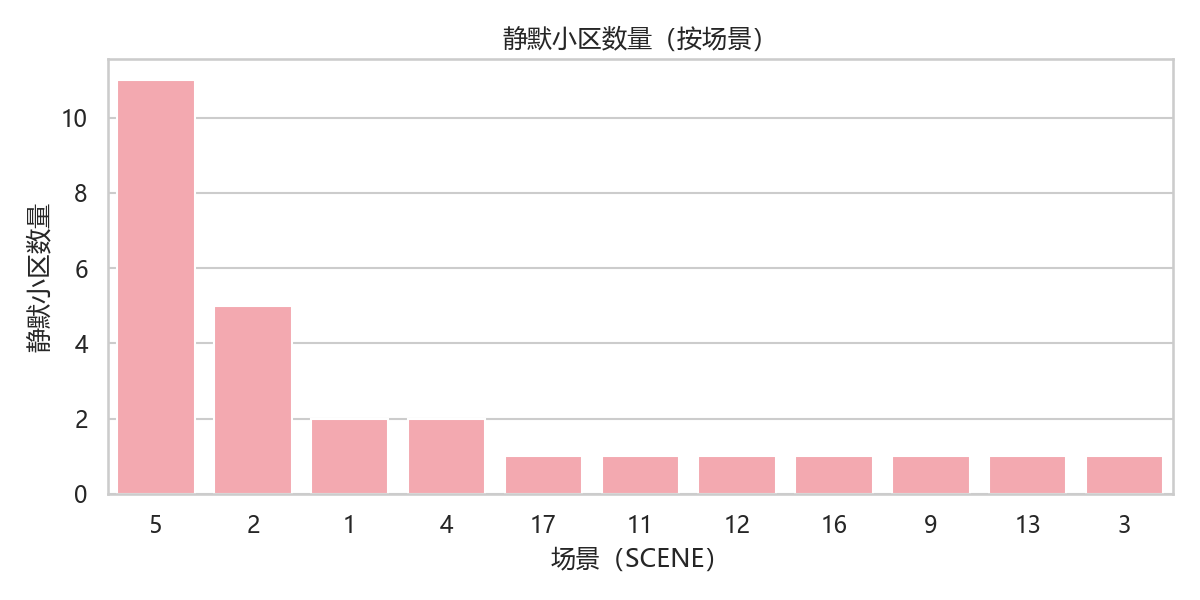
### 3.5 异常分析

关键数据： silent\_ratio>0 cells=7.53%; silent\_ratio>=0.3 cells=27 (0.041%); highload threshold=6,093,868.32 MB; highload cells=652 (1.00%).

定义“静默小区”为“有用户无流量”的时间占比 ≥ 30% 的小区；定义“高负荷小区”为全量小区总流量位于前 1% 的小区（阈值 6,093,868.32 MB）。识别结果如下：

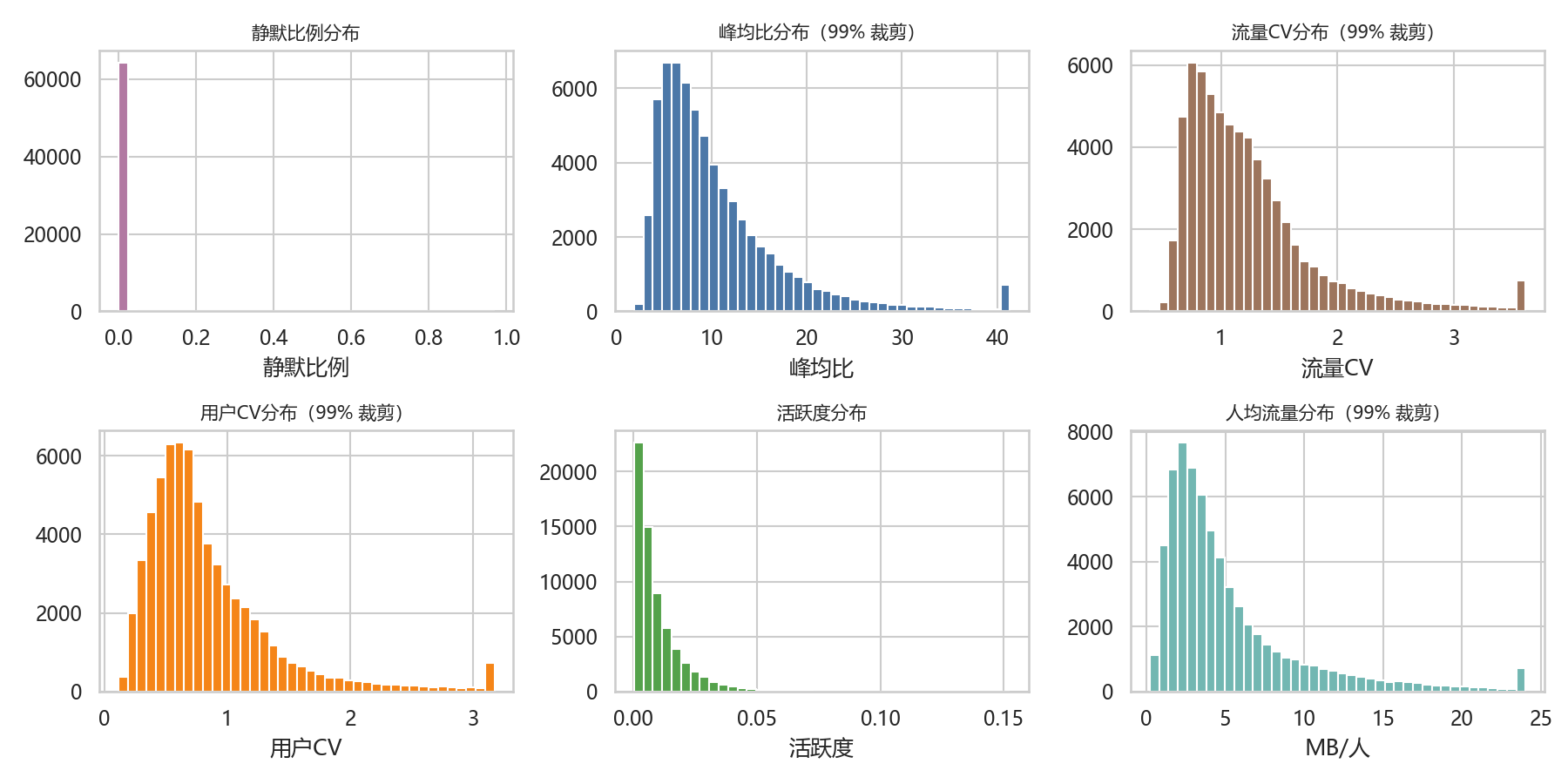
* 静默小区：27 个（占比极低，属于边缘异常）。
* 高负荷小区：652 个（核心承载区，需重点关注扩容与保障）。

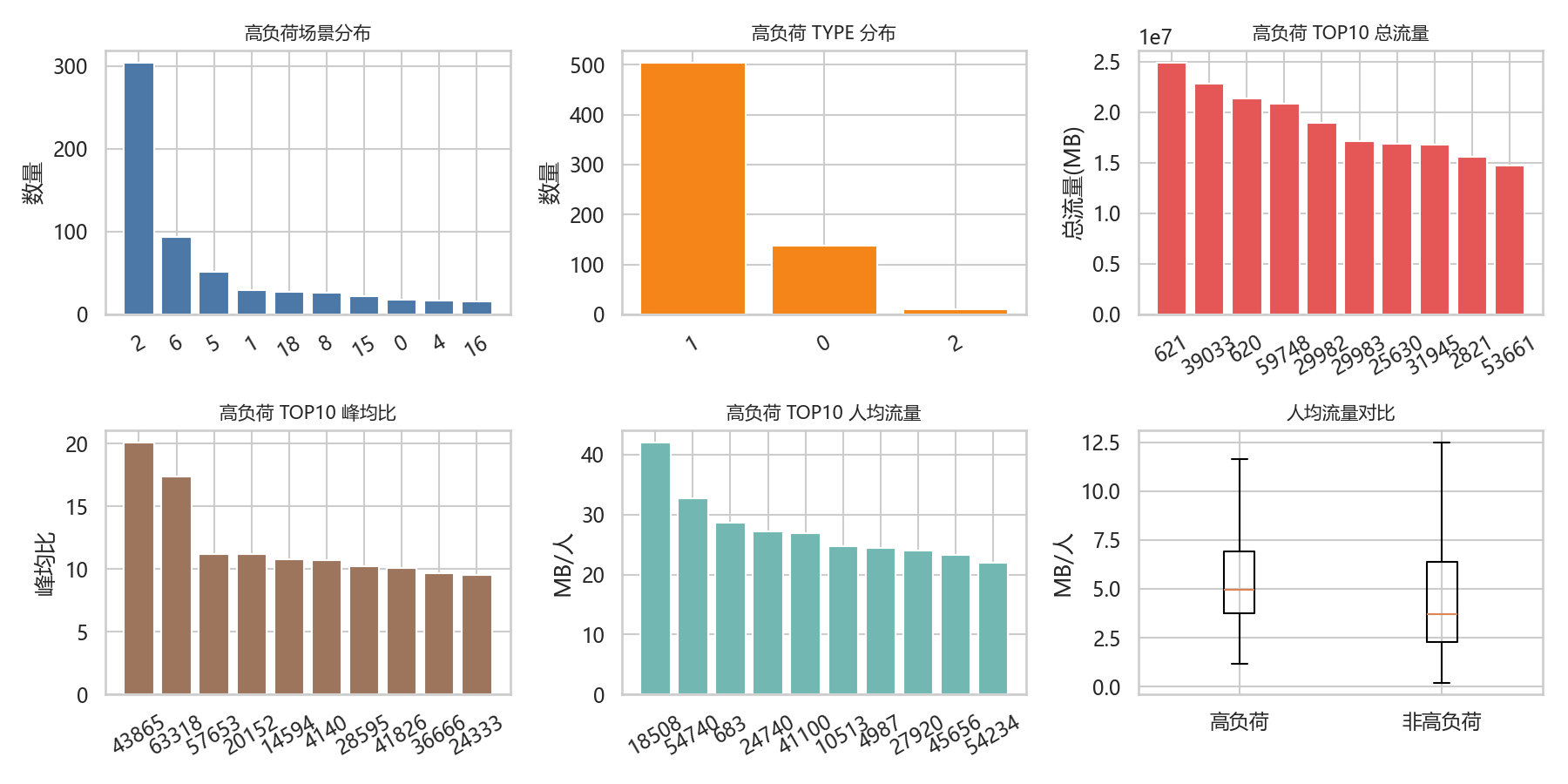
  
含义：该图展示静默比例分布。  
意义：静默比例分布, 判断集中程度、偏态与长尾特征。  
用法：静默比例分布, 用于设置阈值、识别异常或分层。

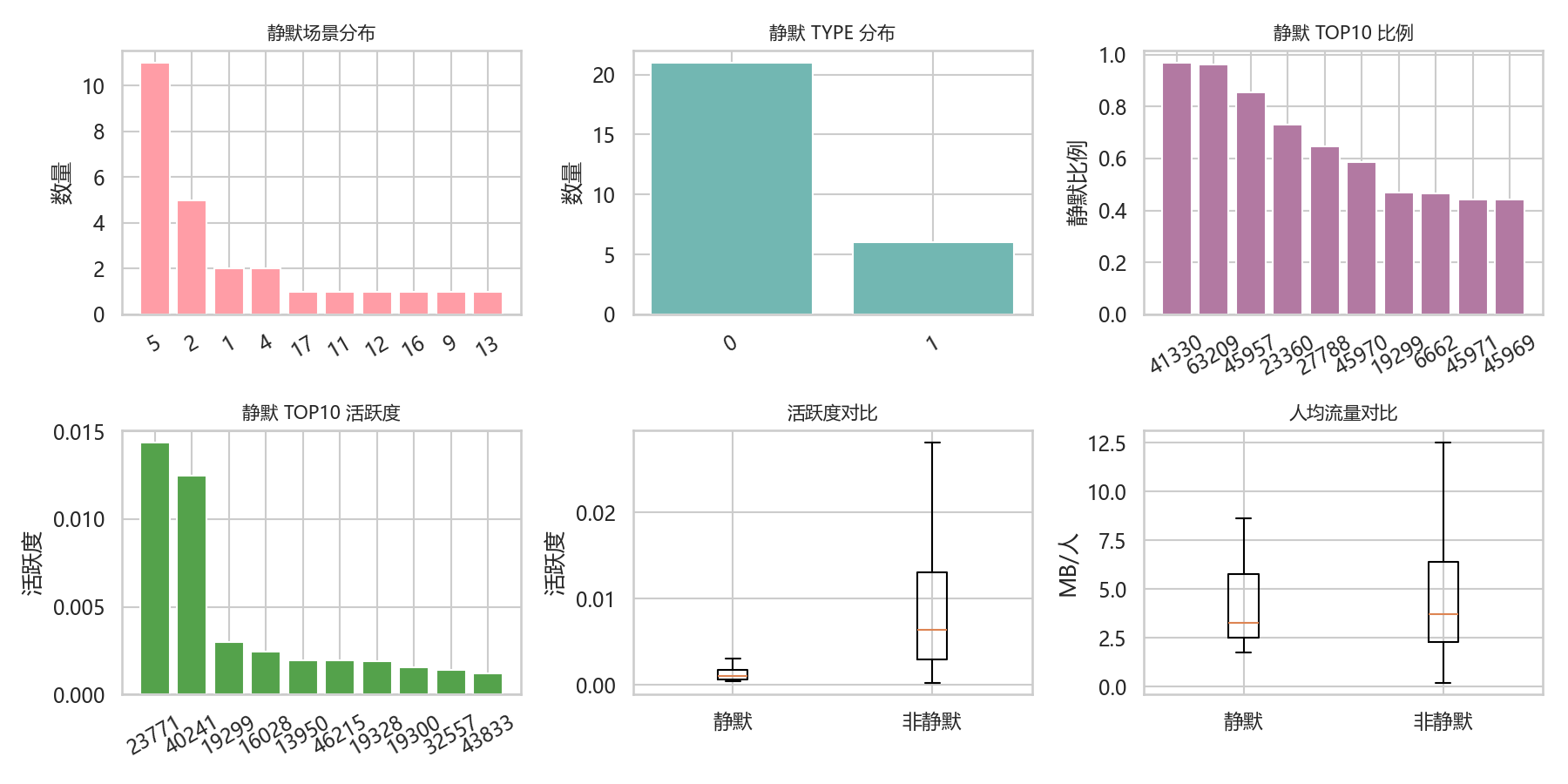
  
含义：该图展示静默小区数量（按场景）。  
意义：静默小区数量（按场景）, 提供整体结构与差异的直观视角。  
用法：静默小区数量（按场景）, 辅助定位重点区间与异常样本。

  
含义：该图展示高负荷小区数量（按场景）。  
意义：高负荷小区数量（按场景）, 提供整体结构与差异的直观视角。  
用法：高负荷小区数量（按场景）, 辅助定位重点区间与异常样本。

补充绘制异常分析分组网格图：分布类、 高负荷画像类、静默画像类分别成组展示，避免不同类型混杂在同一图中。

  
含义：该图展示异常分布网格图。  
意义：异常分布网格图, 对比静默比例、峰均比、CV 等分布差异。  
用法：异常分布网格图, 用于阈值选择与异常分层。

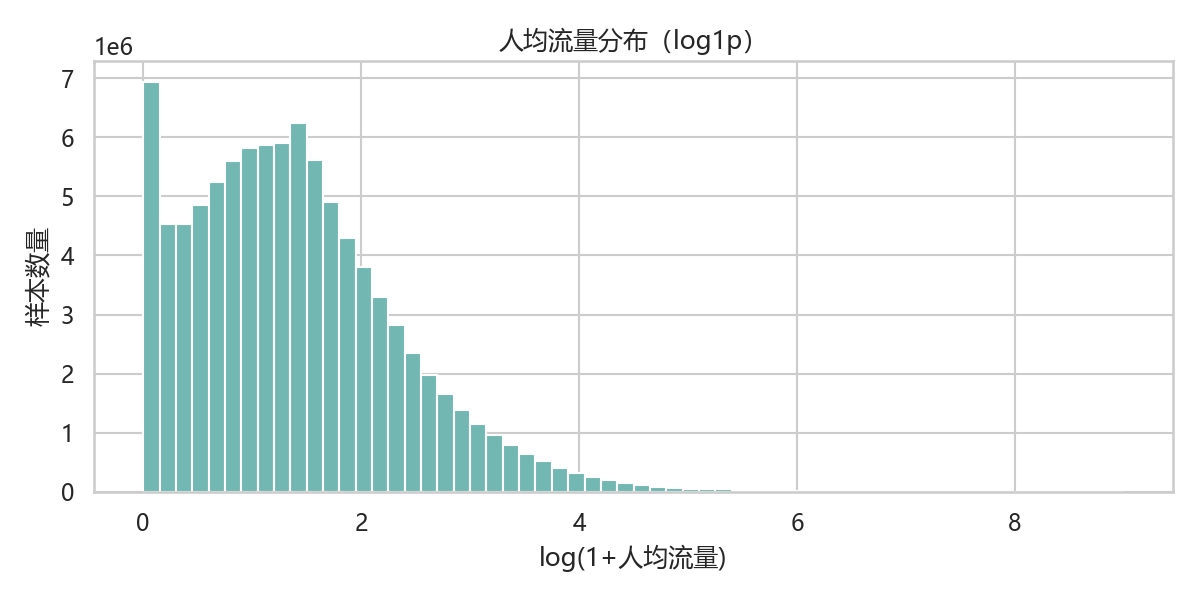
  
含义：该图展示高负荷画像网格图。  
意义：高负荷画像网格图, 综合刻画高负荷小区的场景、类型与指标特征。  
用法：高负荷画像网格图, 用于扩容优先级与重点保障清单。

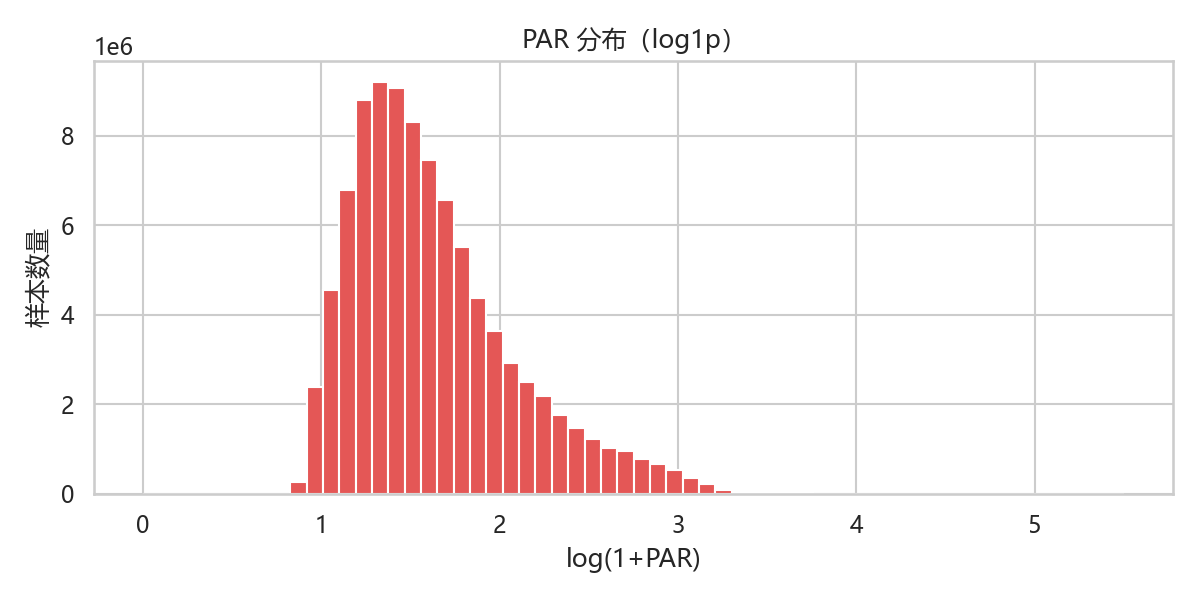
  
含义：该图展示静默画像网格图。  
意义：静默画像网格图, 综合刻画静默小区的场景、类型与指标特征。  
用法：静默画像网格图, 用于故障排查与整改排序。

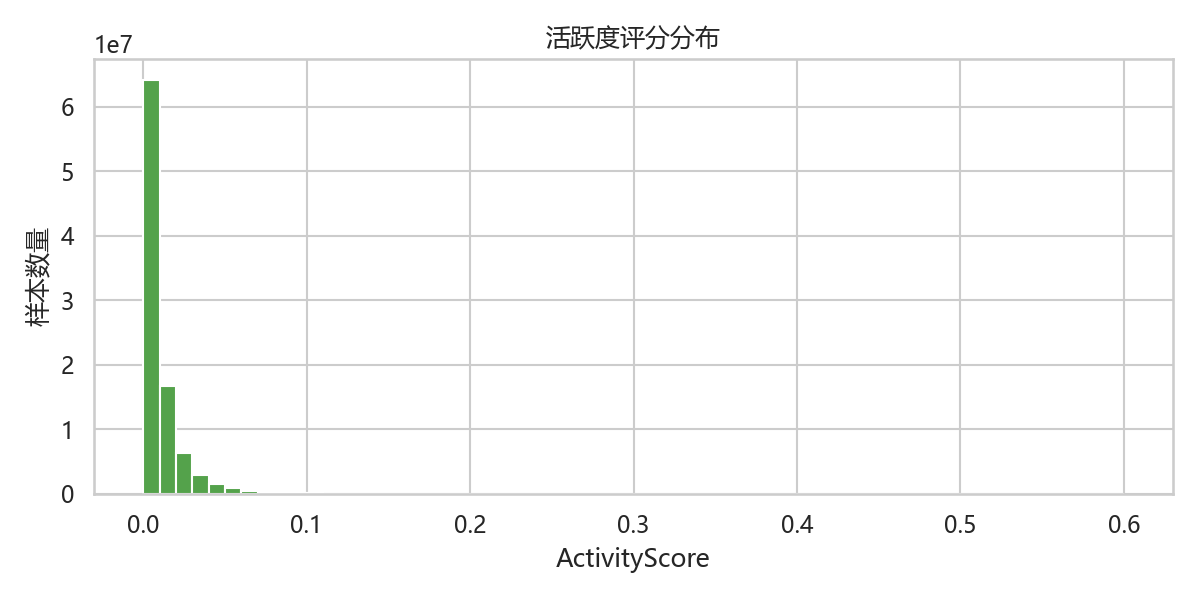
### 3.6 指标扩展与空间特征

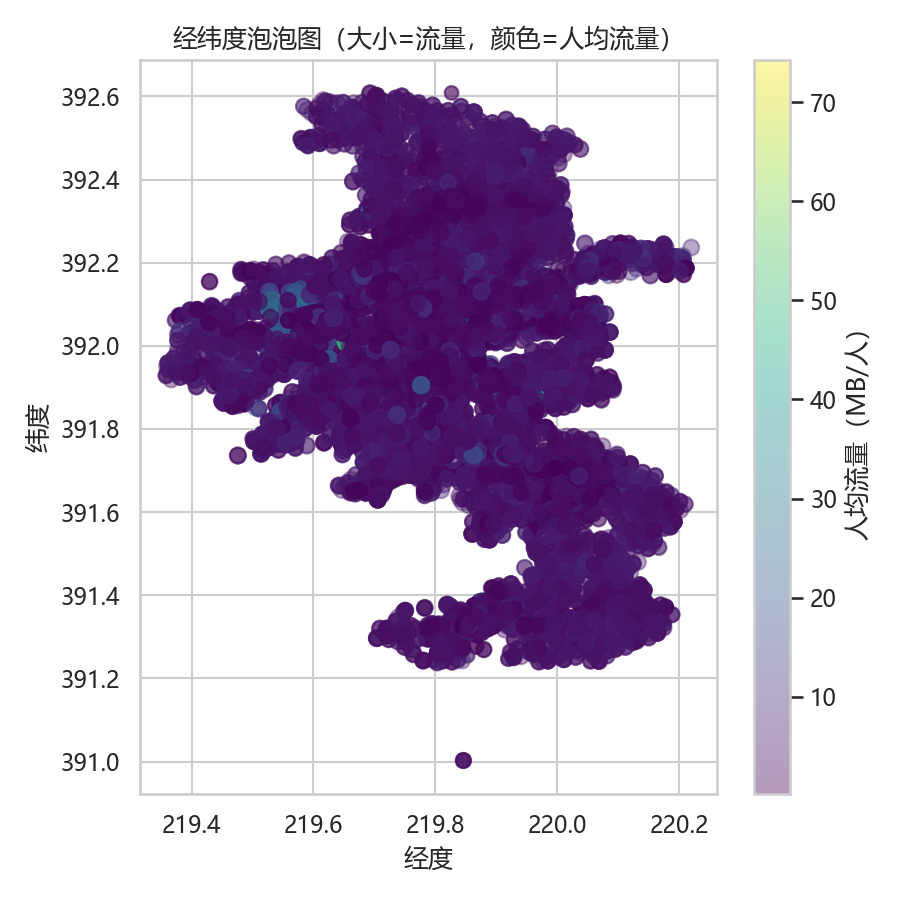
关键数据： flow\_mean median=696.44, p90=2101.40; user\_mean median=163.35, p90=638.71; flow\_per\_user median=3.72, p90=11.25; peak\_ratio median=8.63, p90=18.84; activity\_mean median=0.00634, p90=0.02283.

基于 flow\_per\_user、PAR、ActivityScore 与经纬度特征补充多维分析。

  
含义：该图展示人均流量分布。  
意义：人均流量分布, 判断集中程度、偏态与长尾特征。  
用法：人均流量分布, 用于设置阈值、识别异常或分层。

  
含义：该图展示PAR 分布。  
意义：PAR 分布, 判断集中程度、偏态与长尾特征。  
用法：PAR 分布, 用于设置阈值、识别异常或分层。

  
含义：该图展示活跃度评分分布。  
意义：活跃度评分分布, 判断集中程度、偏态与长尾特征。  
用法：活跃度评分分布, 用于设置阈值、识别异常或分层。

  
含义：该图展示经纬度泡泡图（大小=流量，颜色=人均流量）。  
意义：经纬度泡泡图（大小=流量，颜色=人均流量）, 揭示空间分布与热点区域。  
用法：经纬度泡泡图（大小=流量，颜色=人均流量）, 定位重点区域进行扩容或优化。

### 3.7 流程与派生指标说明

本节分析基于 all\_final\_data\_with\_attributes.csv，节假日字段按 2021 年法定假日表重新计算，在剔除负值、保留缺失的前提下完成统计与可视化，输出 20+ 张中文图表与小区聚合结果。派生指标口径如下：

* 小区平均业务流量：flow\_mean = flow\_sum / flow\_count
* 小区平均用户数：user\_mean = user\_sum / user\_count
* 人均流量：flow\_per\_user = flow\_sum / user\_sum
* 流量变异系数：flow\_cv = flow\_std / flow\_mean
* 静默比例：silent\_ratio = silent\_count / record\_count
* 峰均比：peak\_ratio = flow\_max / flow\_mean
* 活跃度均值：activity\_mean = activity\_sum / activity\_count
* PAR 均值：par\_mean = par\_sum / par\_count

对应的产出文件：section3\_descriptive.py、report\_assets/section3/section3\_stats.json、report\_assets/section3/section3\_cell\_agg.csv 与 report\_assets/section3/fig01\_flow\_user\_hist.png 至 report\_assets/section3/fig35\_scene\_type\_top10.png。

## 4. 多维度对比与趋势分析

4.1 日内潮汐规律分析

小时分布：绘制全天 24 小时的流量走势图，分析不同 TYPE  
小区的早晚高峰特征。

4.2 场景、类型差异对比

场景画像：对比不同 SCENE（如场景 2 与场景  
6）的流量密度，分析哪些场景是运营商的核心价值区 。

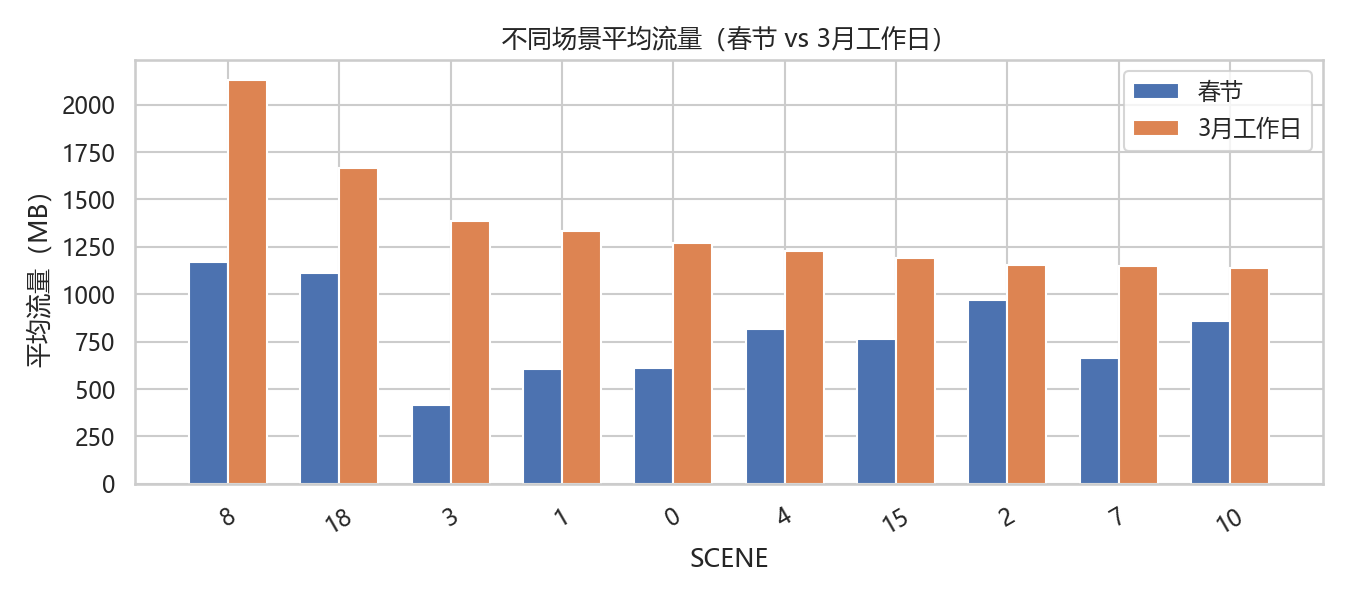
Type画像

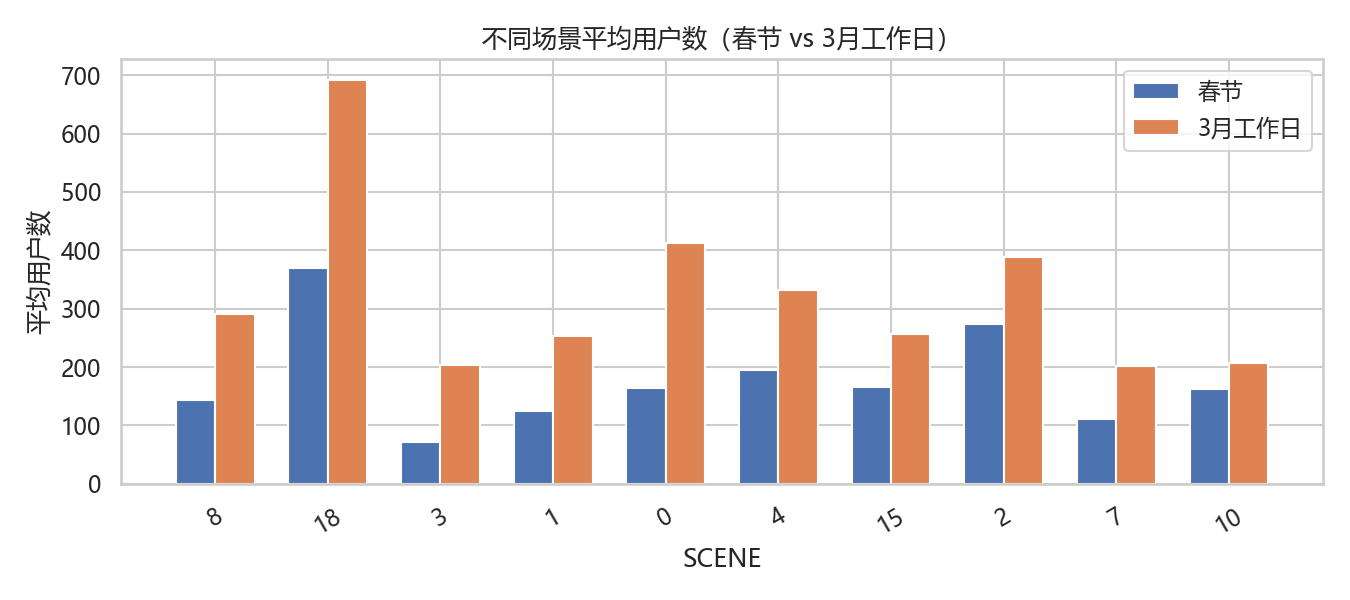
4.3 节假日专项分析（核心点）

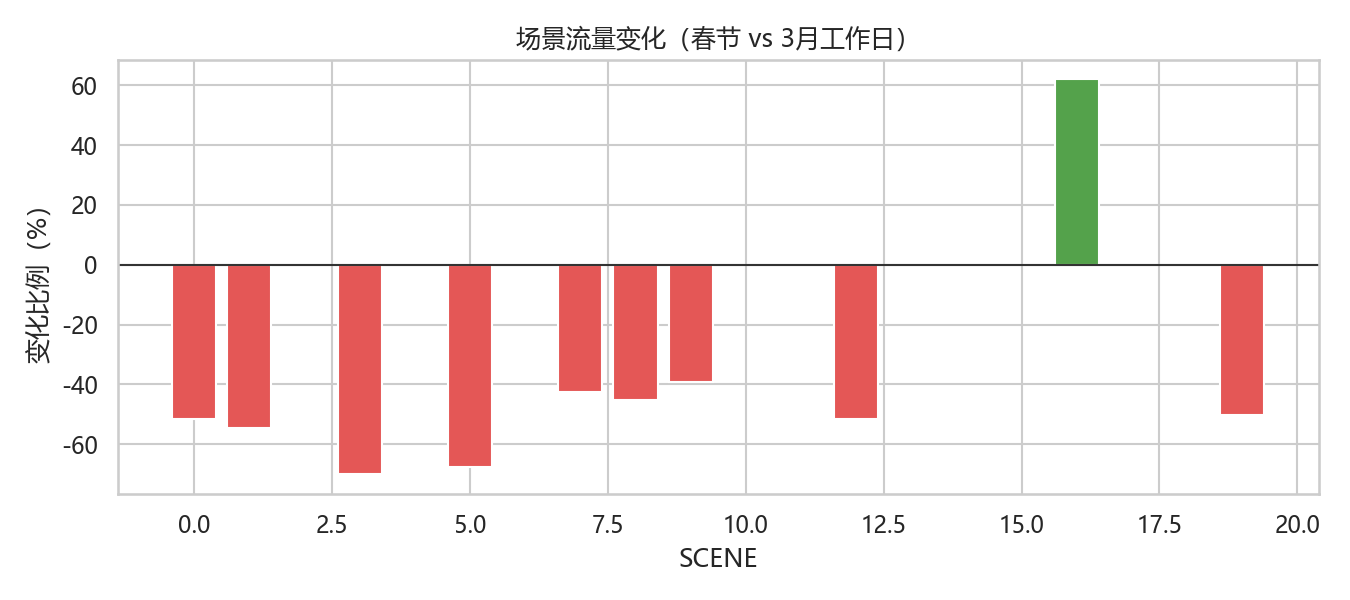
关键数据： spring daily mean flow=1.101e9 vs March weekday=1.634e9 (-32.6%); spring daily mean user=2.949e8 vs 4.590e8 (-35.7%); top drop scenes: S3 -70.0%, S5 -67.5%; top rise: S16 +62.0%; target\_cell=621 (radar dates: 2/11, 2/12, 3/07).

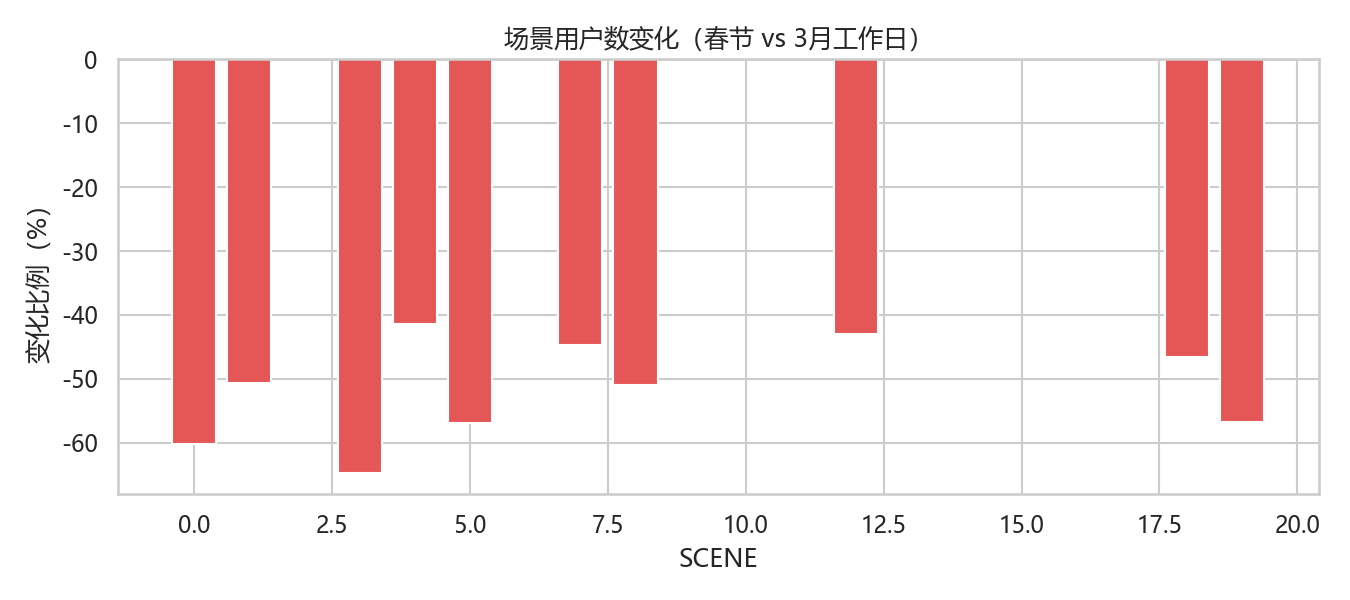
春节效应：春节期间整体流量与用户数显著低于 3 月工作日水平，呈现明显的节假日收缩效应。

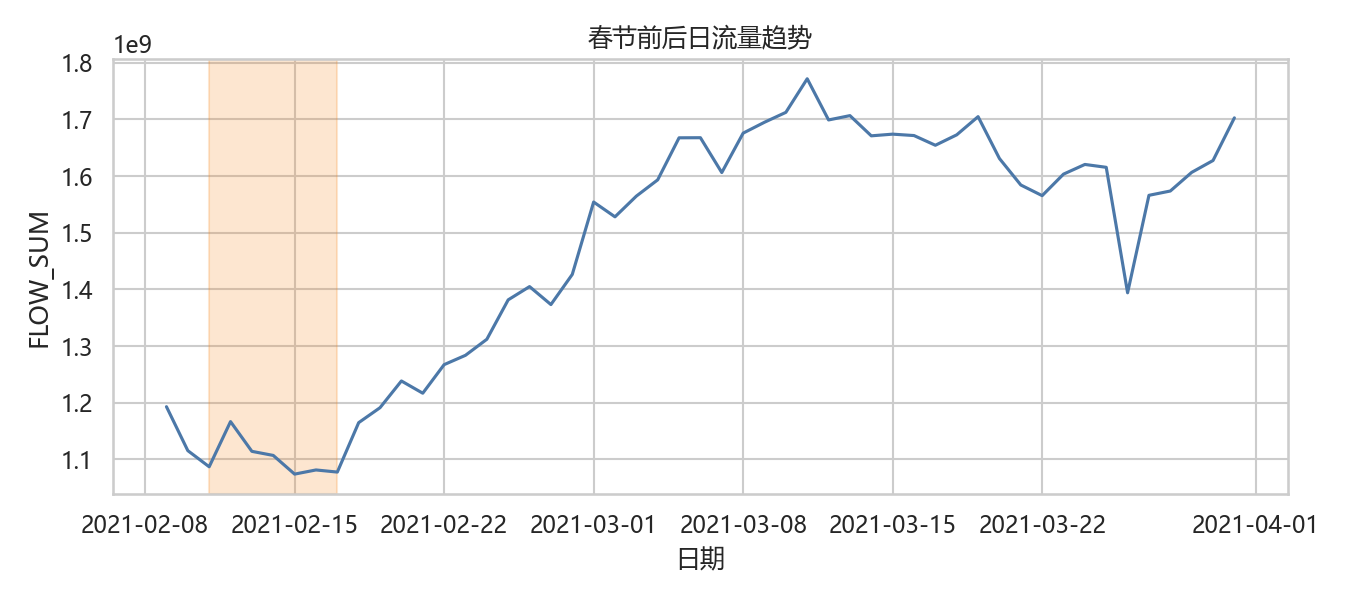
人口迁徙：场景 S3/S5 在春节期间出现大幅下滑（可能对应办公/商业类场景），场景 S16 在春节期间显著上升（可能对应居住类场景），体现人口迁徙与活动中心转移。

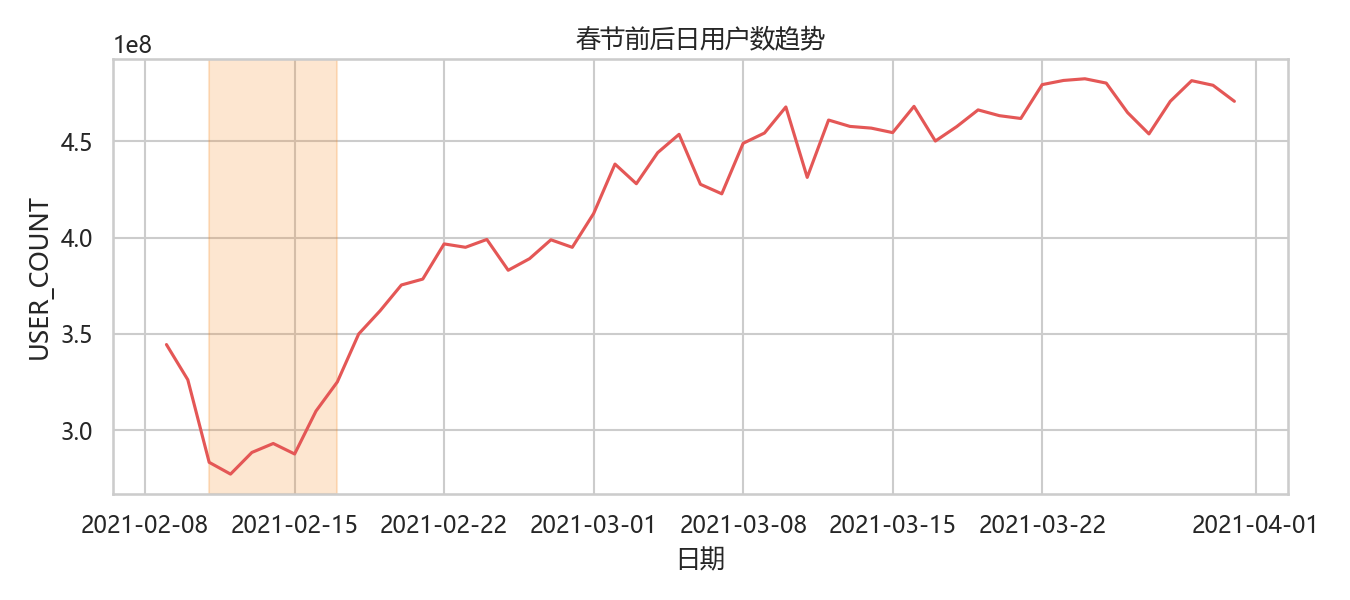
  
含义：该图展示春节与3月工作日在不同场景下的平均流量对比。  
意义：春节与3月工作日对比, 量化节假日迁徙效应与场景承载变化。  
用法：春节与3月工作日对比, 识别流量收缩/扩张的重点场景。

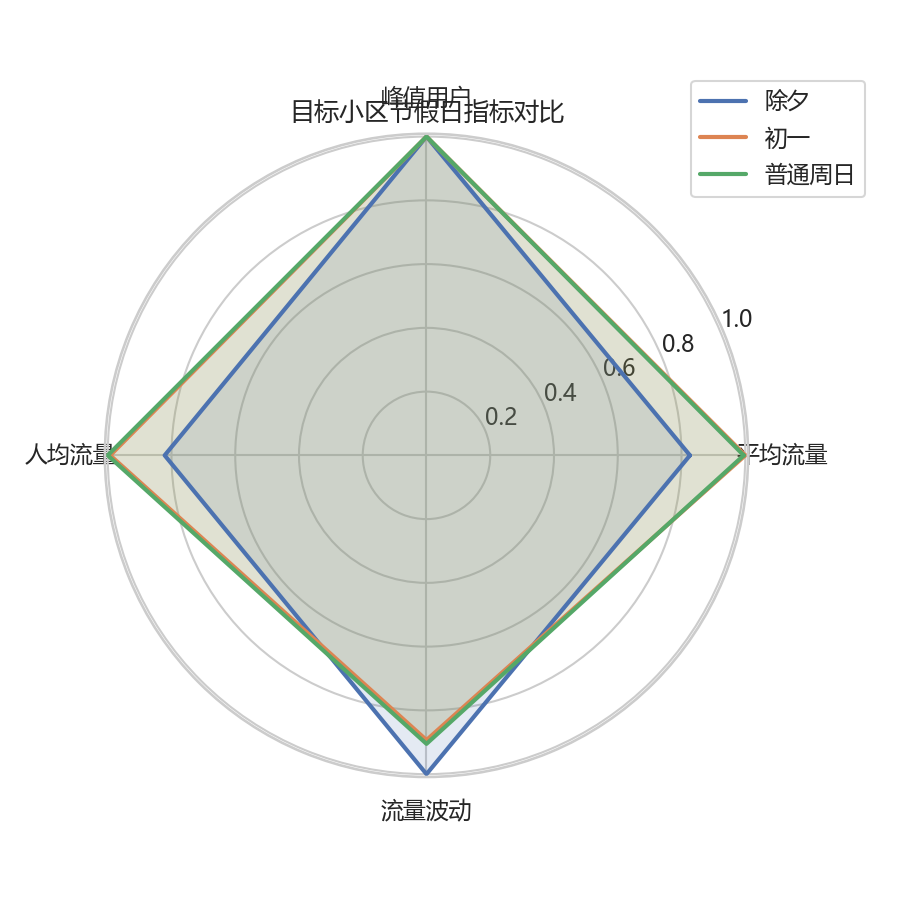
  
含义：该图展示春节与3月工作日在不同场景下的平均用户数对比。  
意义：春节与3月工作日对比, 观察用户规模变化与场景迁移方向。  
用法：春节与3月工作日对比, 作为场景层面的运维与资源调整依据。

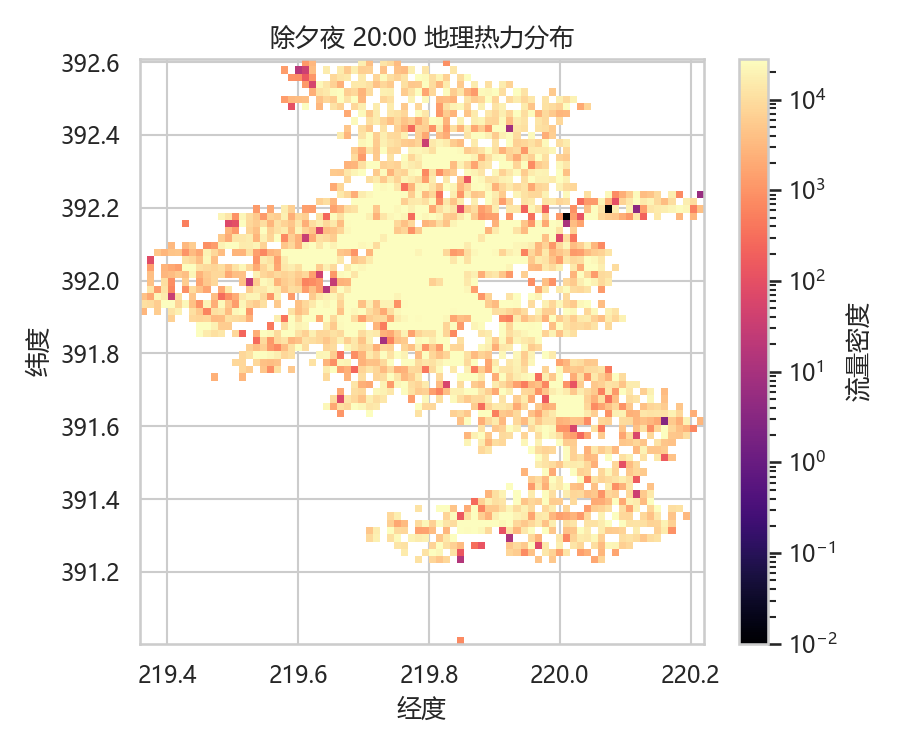
  
含义：该图展示场景流量在春节相对3月工作日的变化比例。  
意义：场景流量变化, 揭示下降或上升幅度最大的场景类别。  
用法：场景流量变化, 用于定位“办公区骤降/住宅区上升”的候选场景。

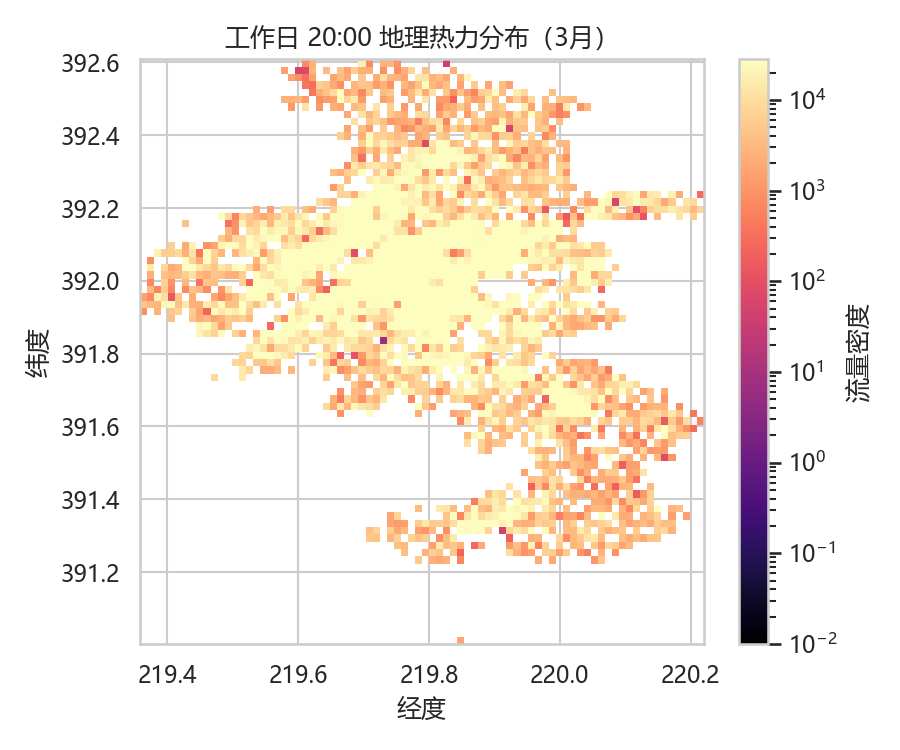
  
含义：该图展示场景用户数在春节相对3月工作日的变化比例。  
意义：场景用户变化, 判断人口迁徙在不同场景的强弱程度。  
用法：场景用户变化, 支撑假期保障与容量规划策略。

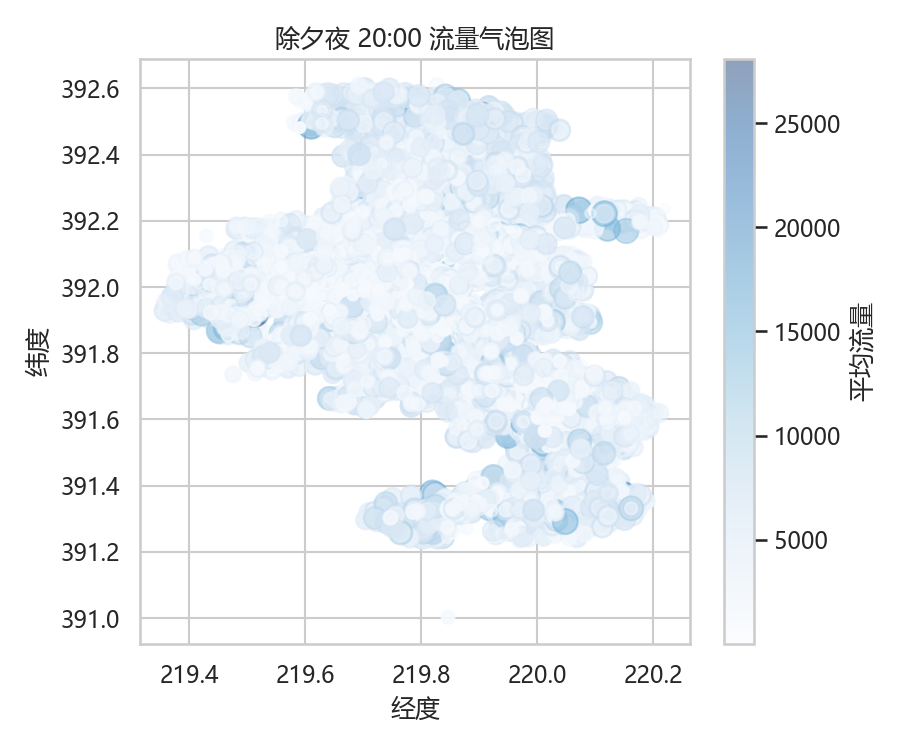
  
含义：该图展示2月与3月日总流量趋势并标注春节区间。  
意义：日流量趋势, 直观看到春节期间的整体收缩效应。  
用法：日流量趋势, 辅助判断节假日保障窗口与异常波动时点。

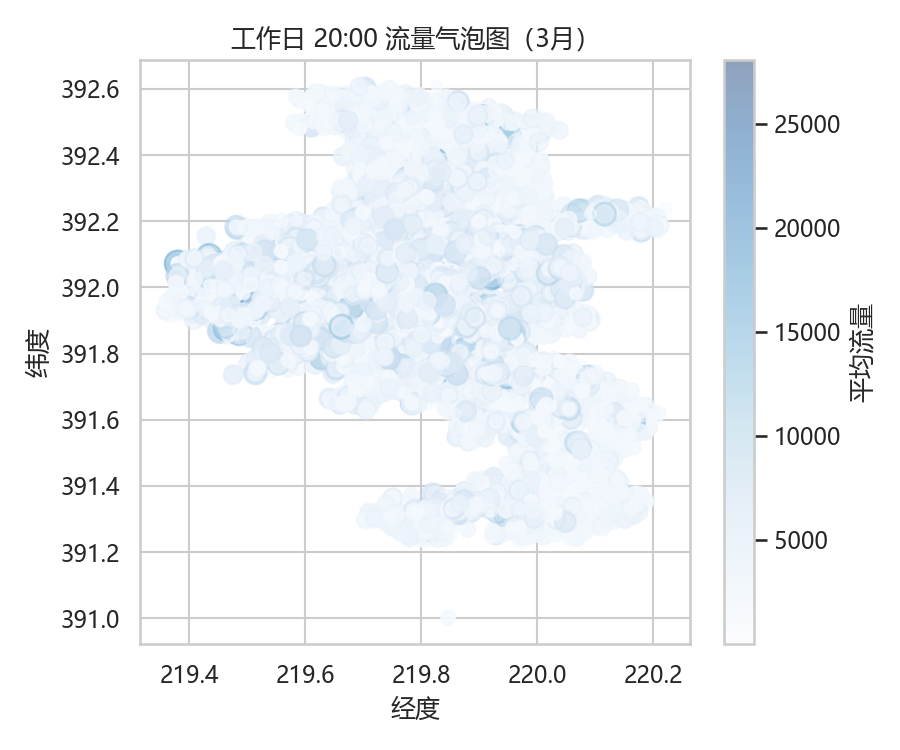
  
含义：该图展示2月与3月日总用户数趋势并标注春节区间。  
意义：日用户趋势, 验证节假日对用户规模的影响幅度。  
用法：日用户趋势, 用于评估节假日人口迁徙强度。

  
含义：该图展示目标小区在除夕、初一、普通周日的多指标对比。  
意义：节假日雷达图, 评估单小区在关键时点的结构差异。  
用法：节假日雷达图, 用于小区级异常识别与策略调度。

  
含义：该图展示除夕夜20:00的流量地理热力分布。  
意义：除夕夜热力图, 反映节假日城市空间活动中心分布。  
用法：除夕夜热力图, 用于节假日保障与重点区域巡检。

  
含义：该图展示3月工作日20:00的流量地理热力分布。  
意义：工作日热力图, 作为非节假日空间分布基线。  
用法：工作日热力图, 对比节假日空间迁移范围。

  
含义：该图展示除夕夜20:00的地理气泡分布（大小=流量，颜色=平均流量强度）。  
意义：除夕夜气泡图, 同时呈现规模与流量强度差异。  
用法：除夕夜气泡图, 辅助识别节假日高流量热点区域。

  
含义：该图展示3月工作日20:00的地理气泡分布（大小=流量，颜色=平均流量强度）。  
意义：工作日气泡图, 作为常态流量强度分布对照。  
用法：工作日气泡图, 定位节假日与工作日的流量强度差异。

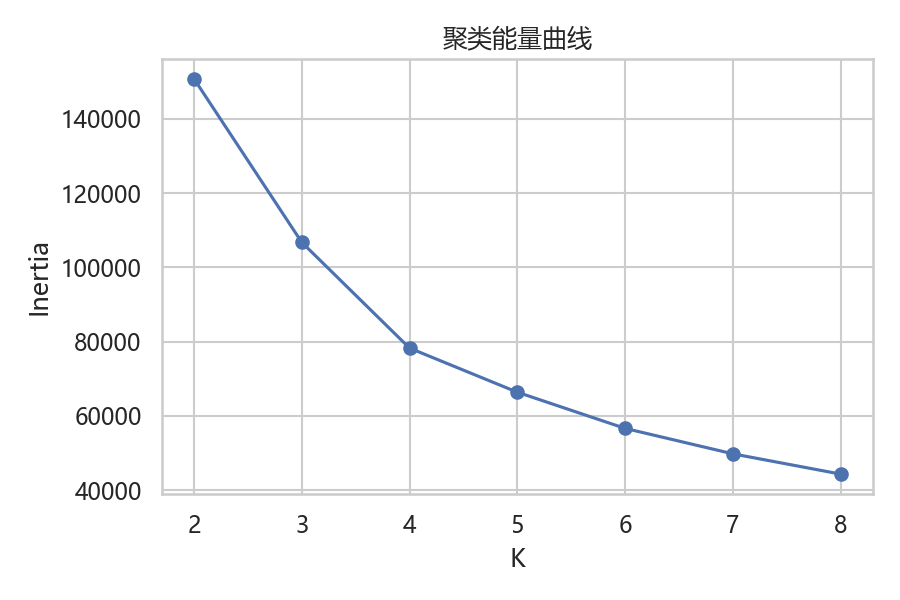
产出文件：section4\_holiday.py、report\_assets/section4/section4\_stats.json，以及 report\_assets/section4/fig01\_scene\_flow\_grouped.png 至 report\_assets/section4/fig11\_bubble\_weekday\_20.png。

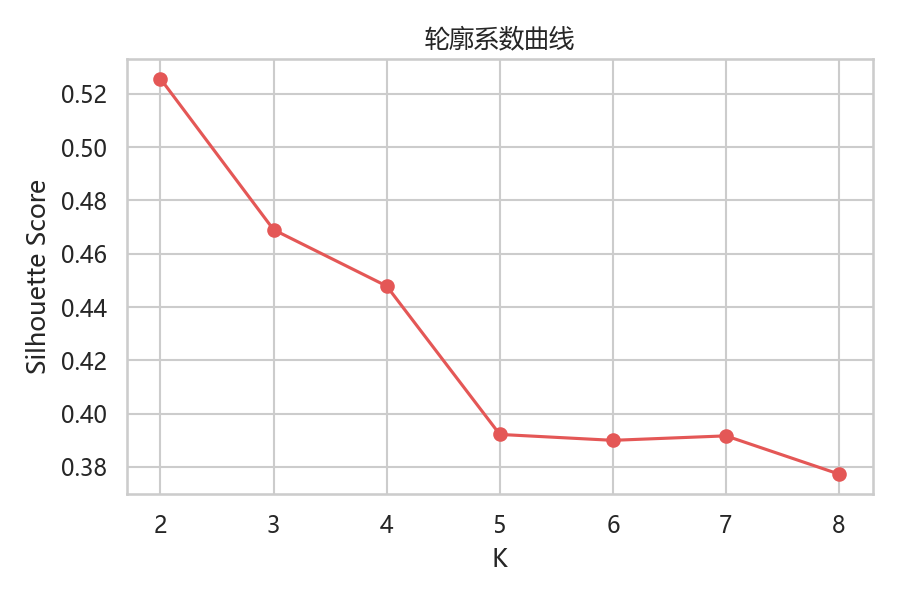
## 5. 高阶分析与挖掘

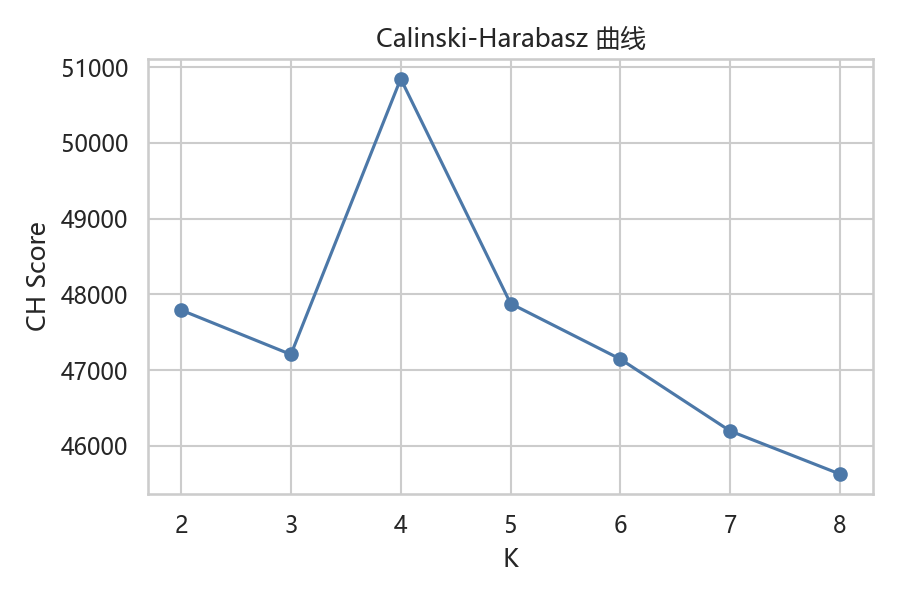
### 5.1 小区画像聚类（多特征剖面）

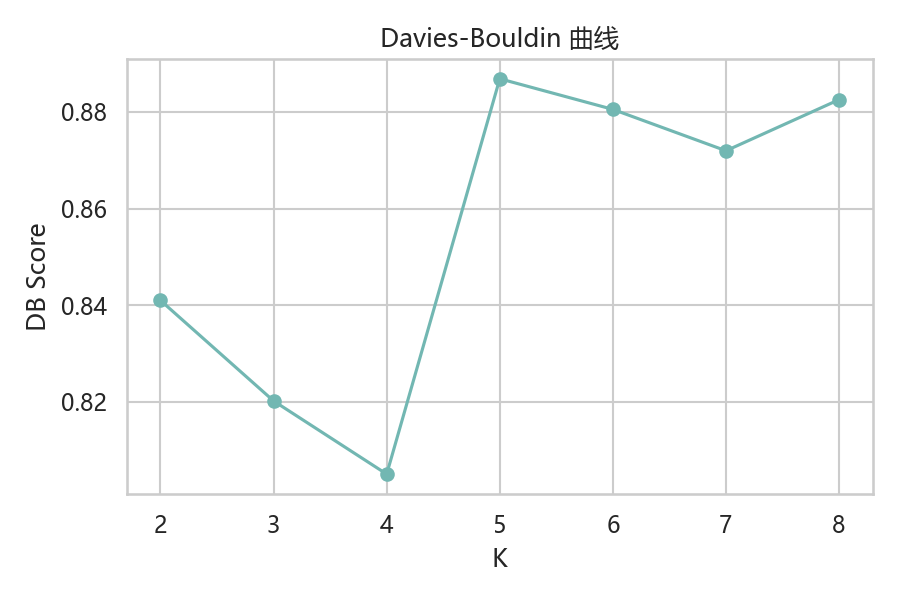
关键数据： total\_cells=65174; k=2 (CH=47793.72, DB=0.841); cluster\_sizes=11819 (18.1%) / 53355 (81.9%); peak\_hour=20 vs 12; cluster vs SCENE CramersV=0.252; cluster vs TYPE CramersV=0.205.

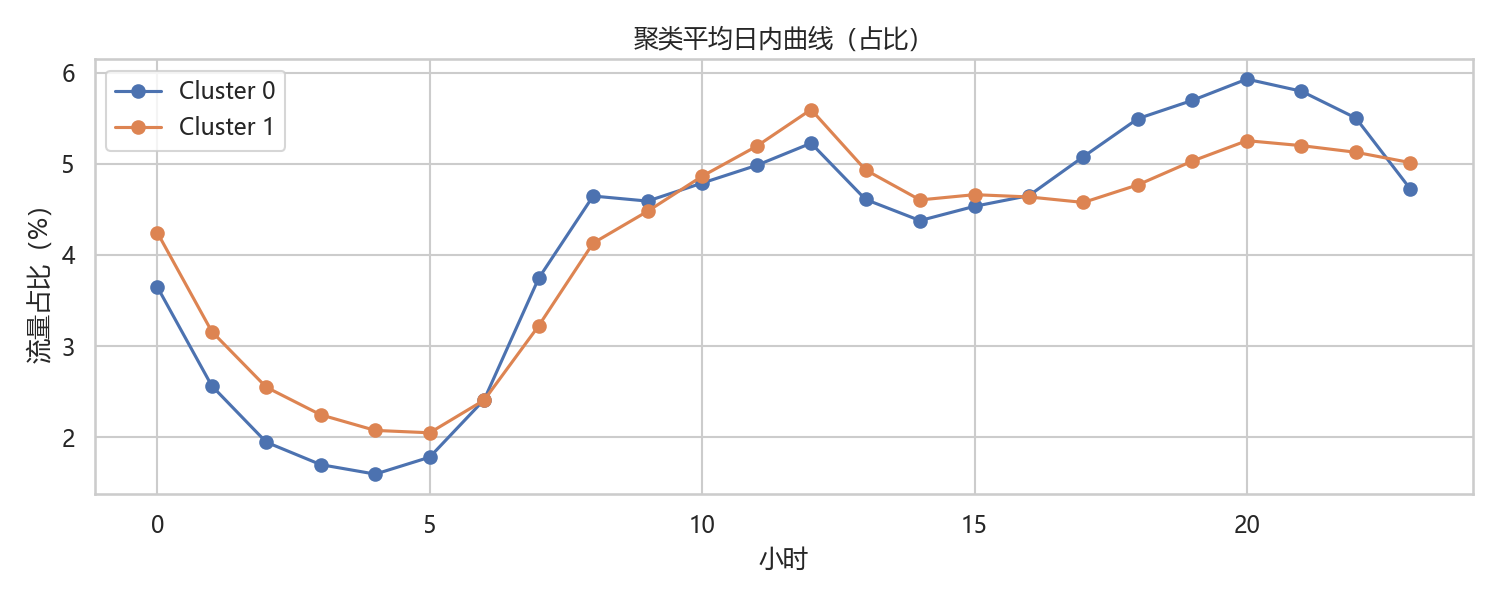
选取全量小区，构建“24 小时流量占比 + 24 小时用户占比 + 关键指标（flow\_mean、user\_mean、flow\_per\_user、activity\_mean）”的组合特征，并进行 K-means 聚类。综合轮廓系数、CH 与 DB 指标，最优 K=2，聚类规模分别为 11819 与 53355。两类在峰值时段上存在明显差异：Cluster 0 峰值集中在 20 点，Cluster 1 峰值集中在 12 点，体现“夜间活跃型”与“午间主峰型”的对照。

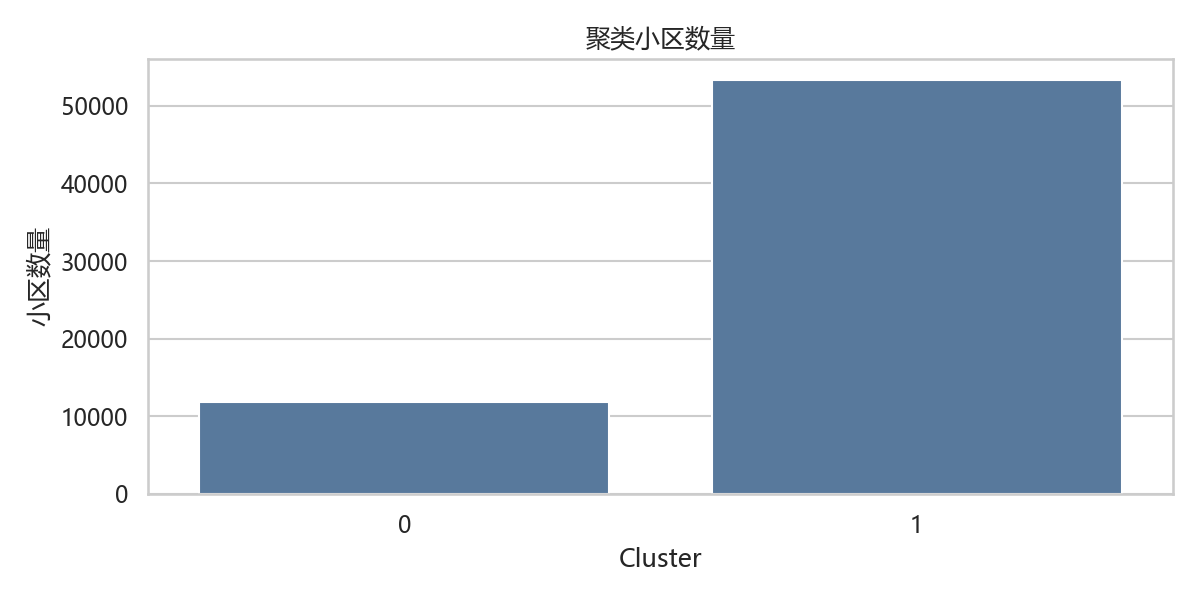
  
含义：该图展示聚类能量曲线。  
意义：聚类能量曲线, 刻画随时间的变化与周期性规律。  
用法：聚类能量曲线, 识别峰谷与突变点，支持容量与资源配置。

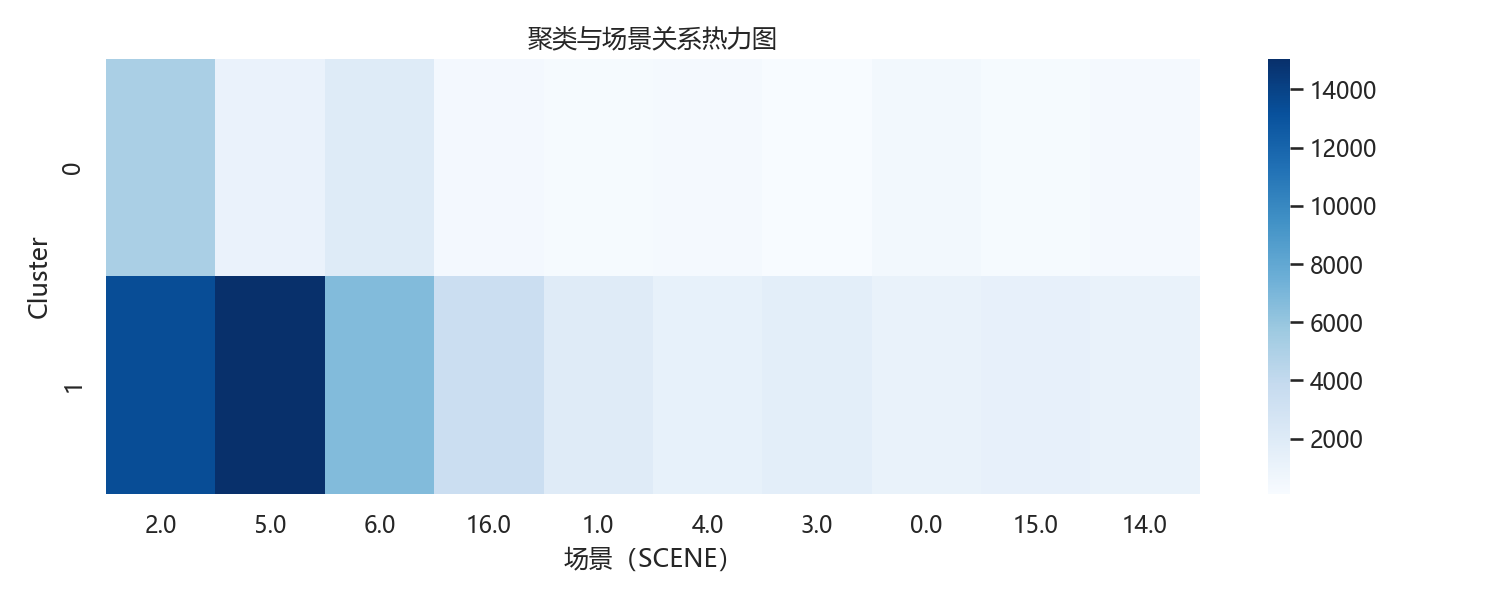
  
含义：该图展示轮廓系数曲线。  
意义：轮廓系数曲线, 刻画随时间的变化与周期性规律。  
用法：轮廓系数曲线, 识别峰谷与突变点，支持容量与资源配置。

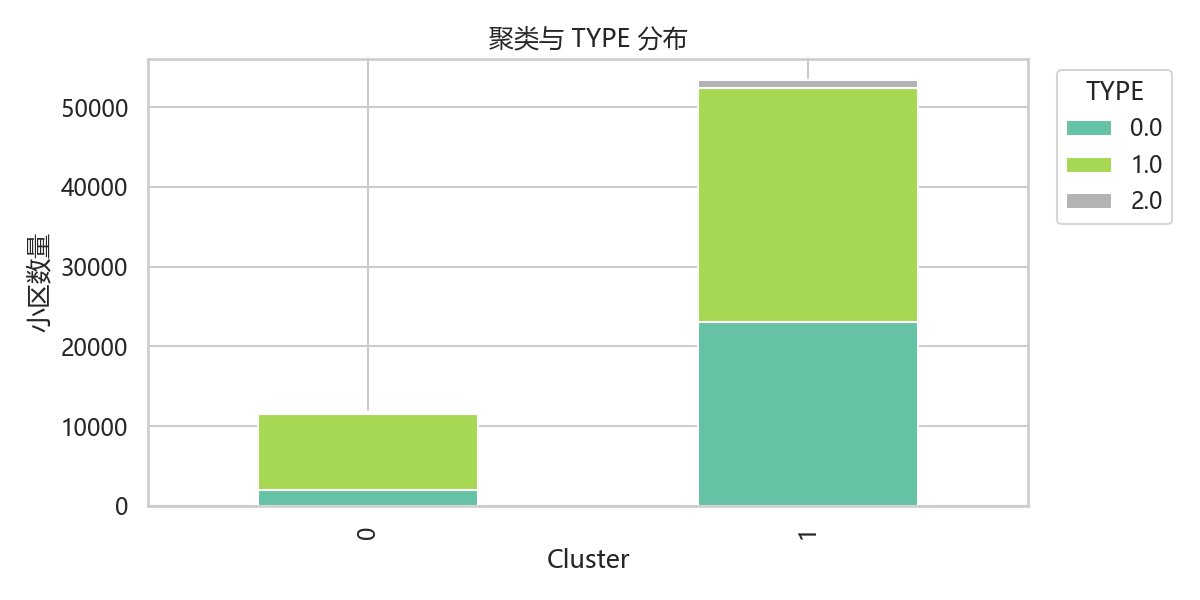
  
含义：该图展示CH 指标曲线。  
意义：CH 指标曲线, 刻画随时间的变化与周期性规律。  
用法：CH 指标曲线, 识别峰谷与突变点，支持容量与资源配置。

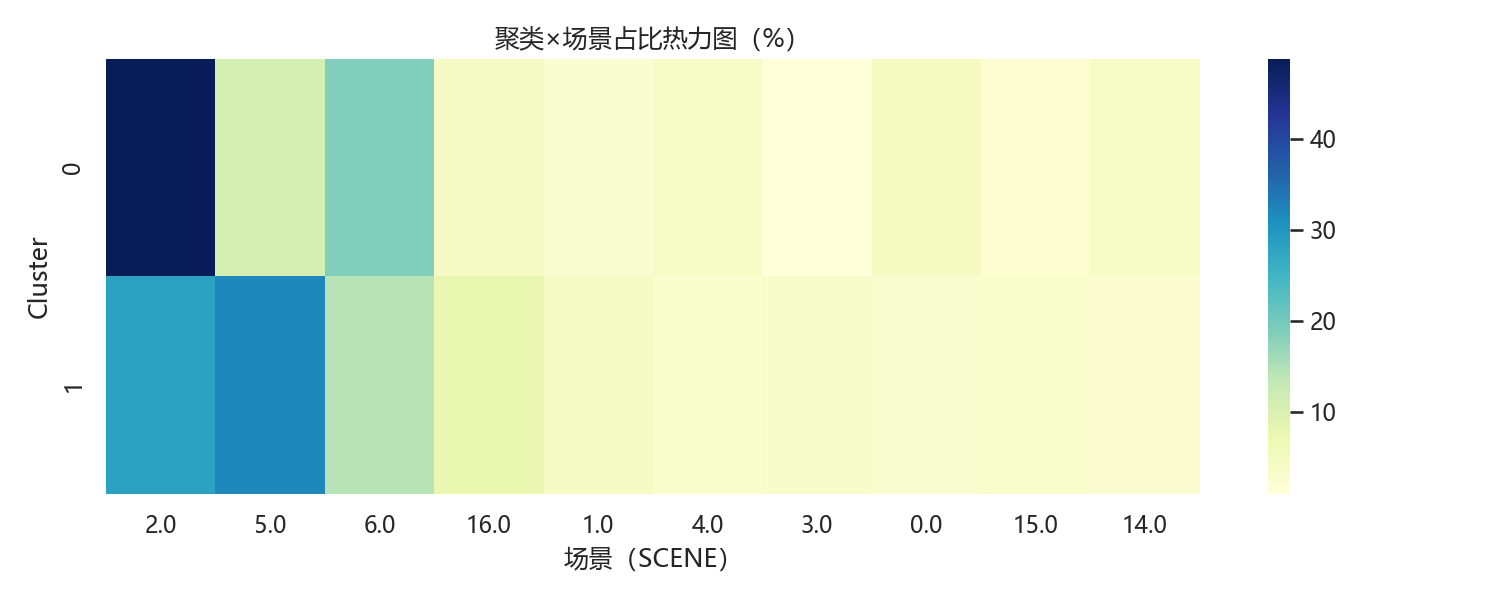
  
含义：该图展示DB 指标曲线。  
意义：DB 指标曲线, 刻画随时间的变化与周期性规律。  
用法：DB 指标曲线, 识别峰谷与突变点，支持容量与资源配置。

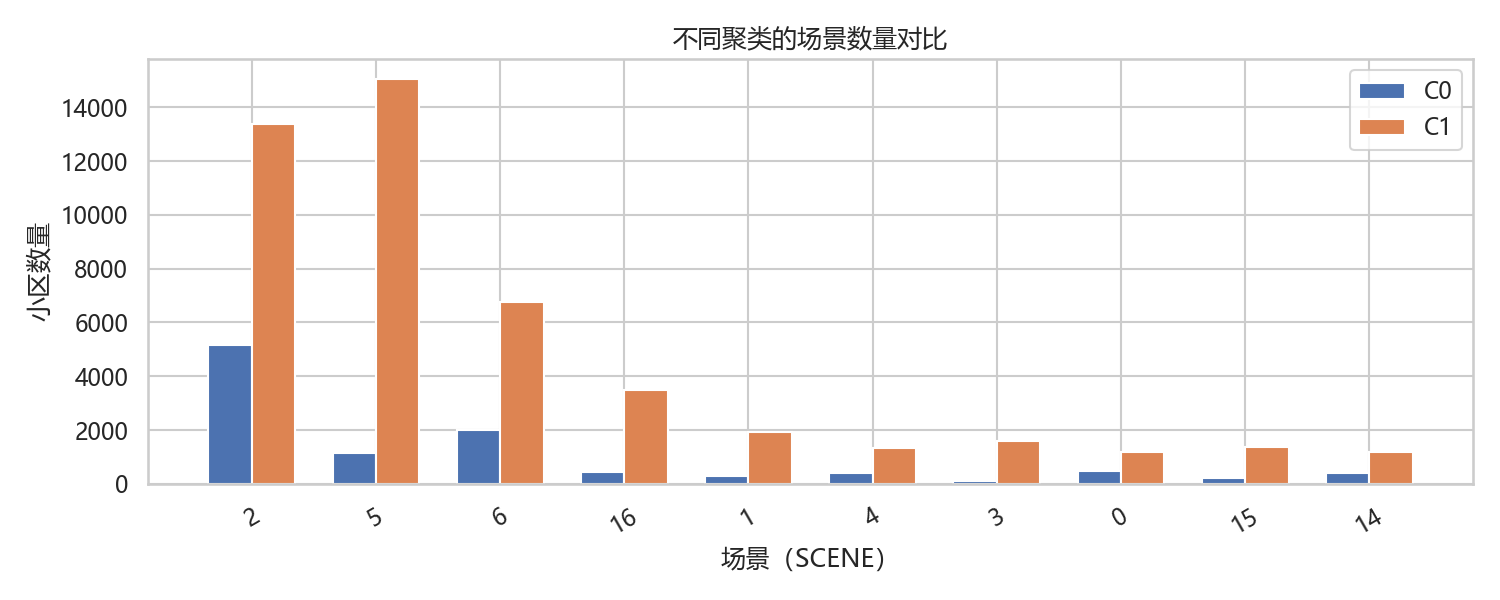
  
含义：该图展示聚类平均日内曲线。  
意义：聚类平均日内曲线, 刻画随时间的变化与周期性规律。  
用法：聚类平均日内曲线, 识别峰谷与突变点，支持容量与资源配置。

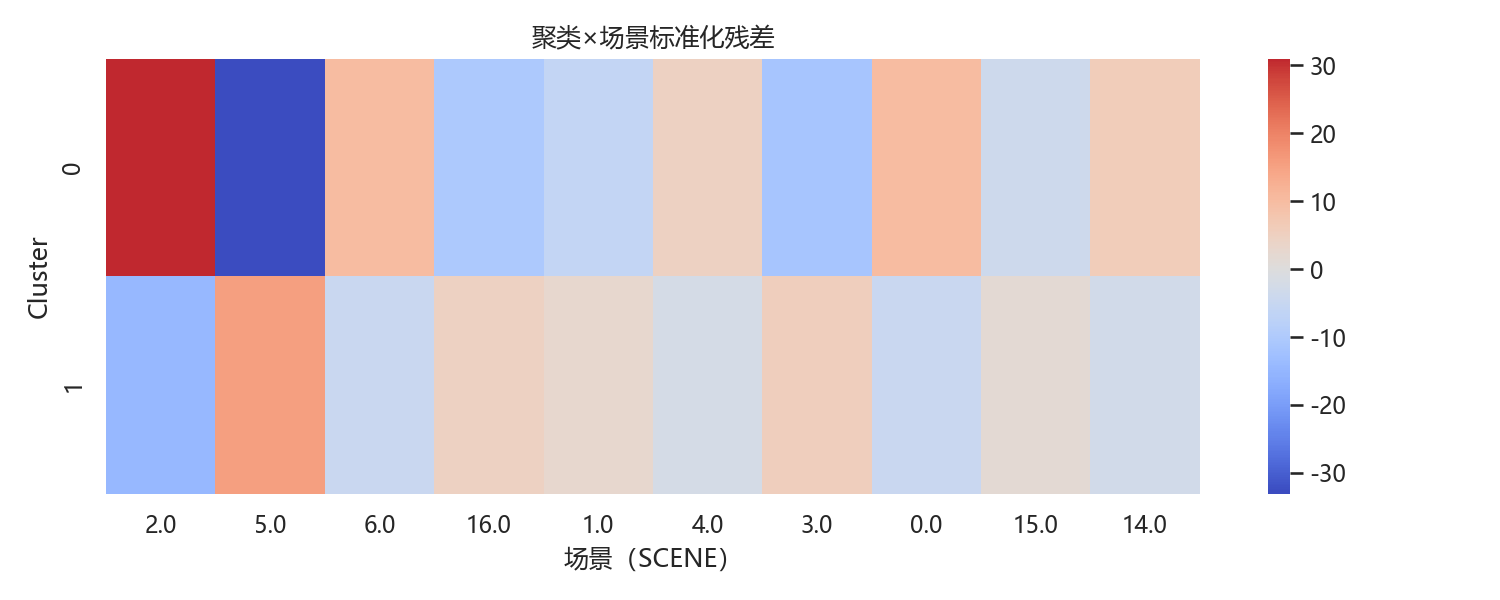
  
含义：该图展示聚类规模。  
意义：聚类规模, 展示聚类规模差异与主导类型。  
用法：聚类规模, 判断细分策略与资源倾斜方向。

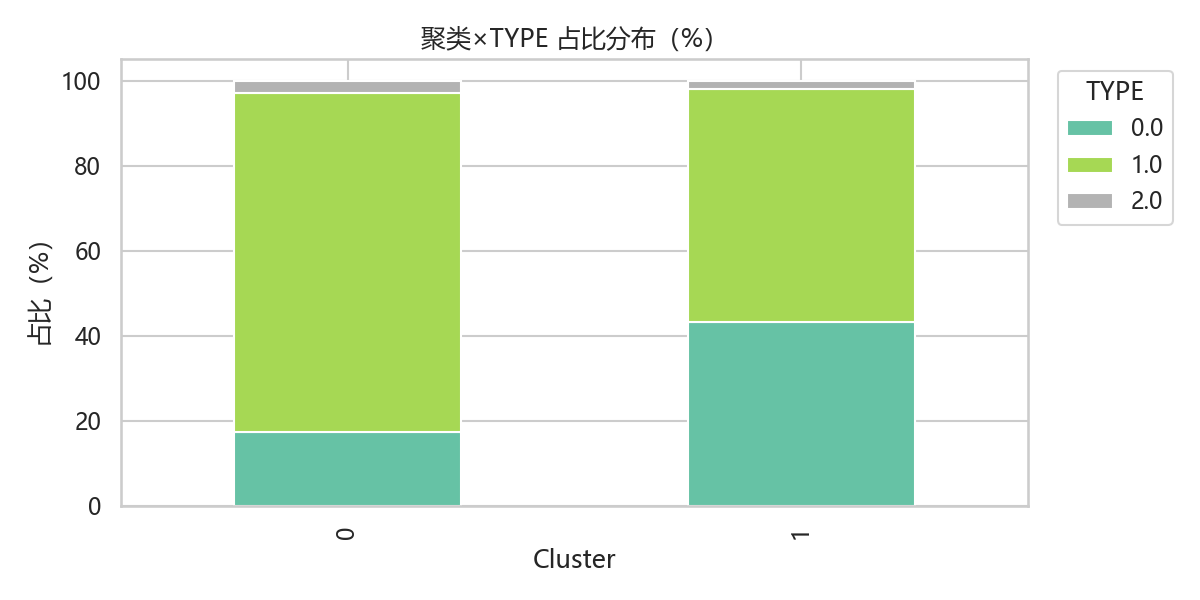
  
含义：该图展示聚类与场景关系热力图。  
意义：聚类与场景关系热力图, 揭示不同维度组合下的强度分布与热点区域。  
用法：聚类与场景关系热力图, 定位高值/低值区间，用于时段、场景或类型优化。

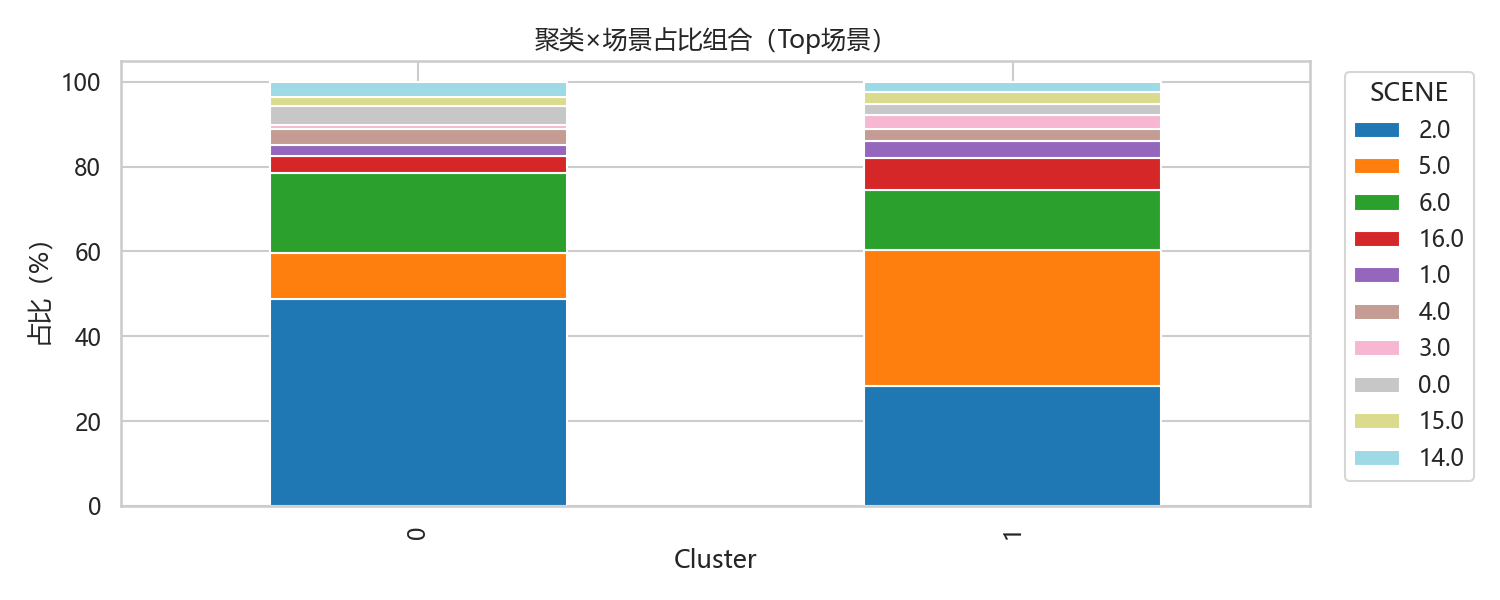
  
含义：该图展示聚类与 TYPE 分布。  
意义：聚类与 TYPE 分布, 判断集中程度、偏态与长尾特征。  
用法：聚类与 TYPE 分布, 用于设置阈值、识别异常或分层。

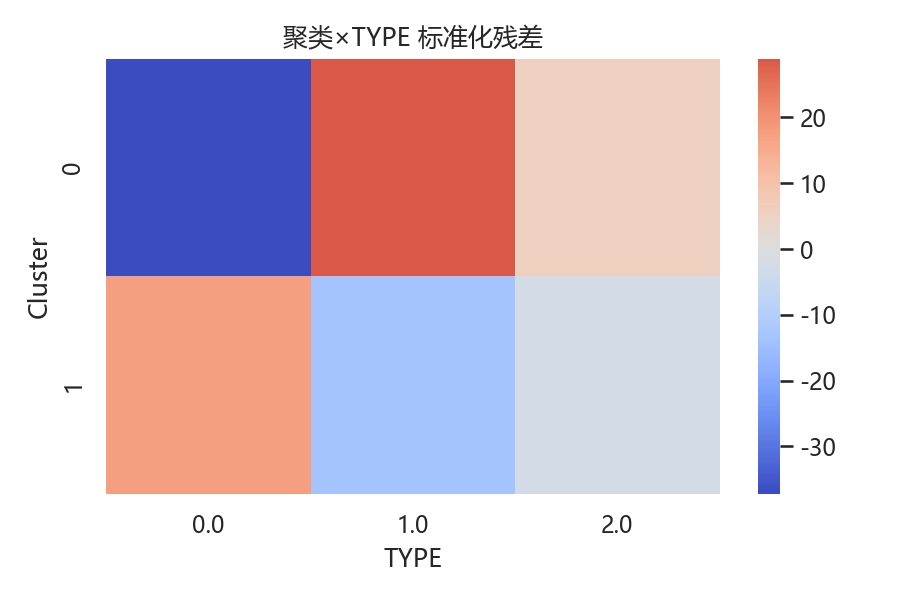
  
含义：该图展示聚类与场景的占比热力图。  
意义：聚类与场景占比, 排除规模影响，突出结构差异。  
用法：聚类与场景占比, 对比两类小区的场景构成。

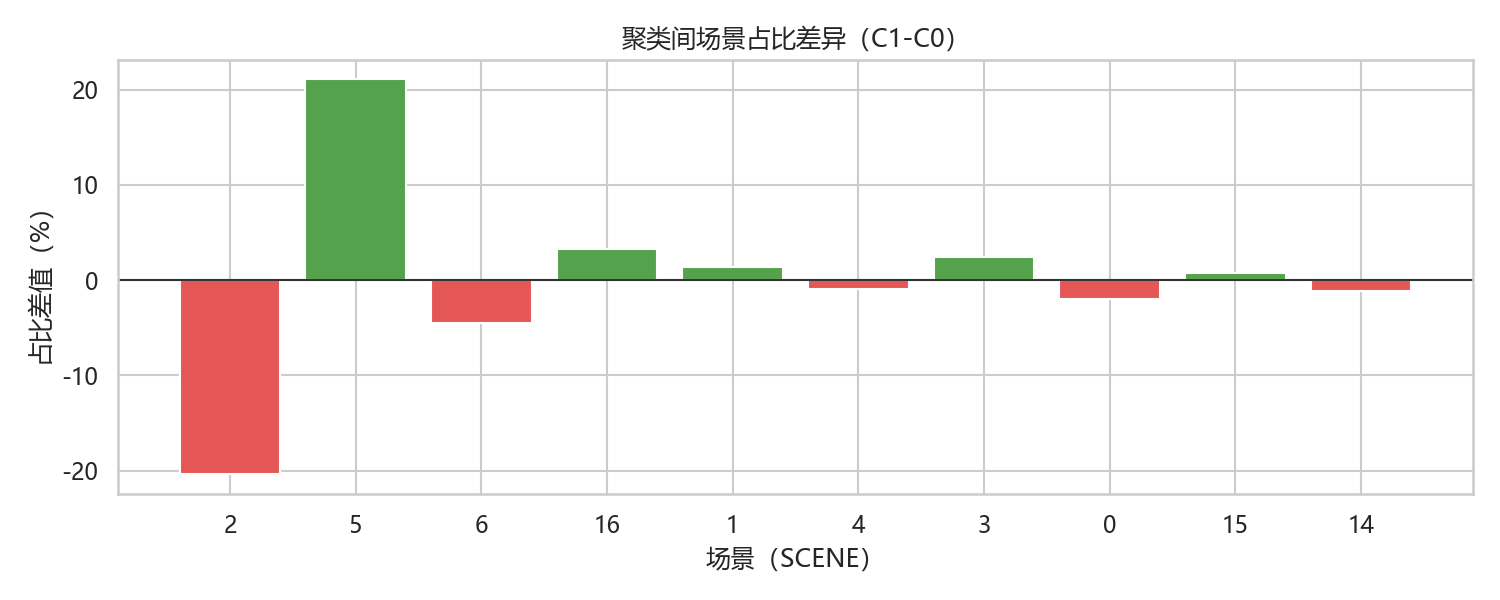
  
含义：该图展示不同聚类在各场景的数量对比。  
意义：聚类场景分组柱状图, 直观看到场景规模差异。  
用法：聚类场景分组柱状图, 识别偏好场景与资源倾斜方向。

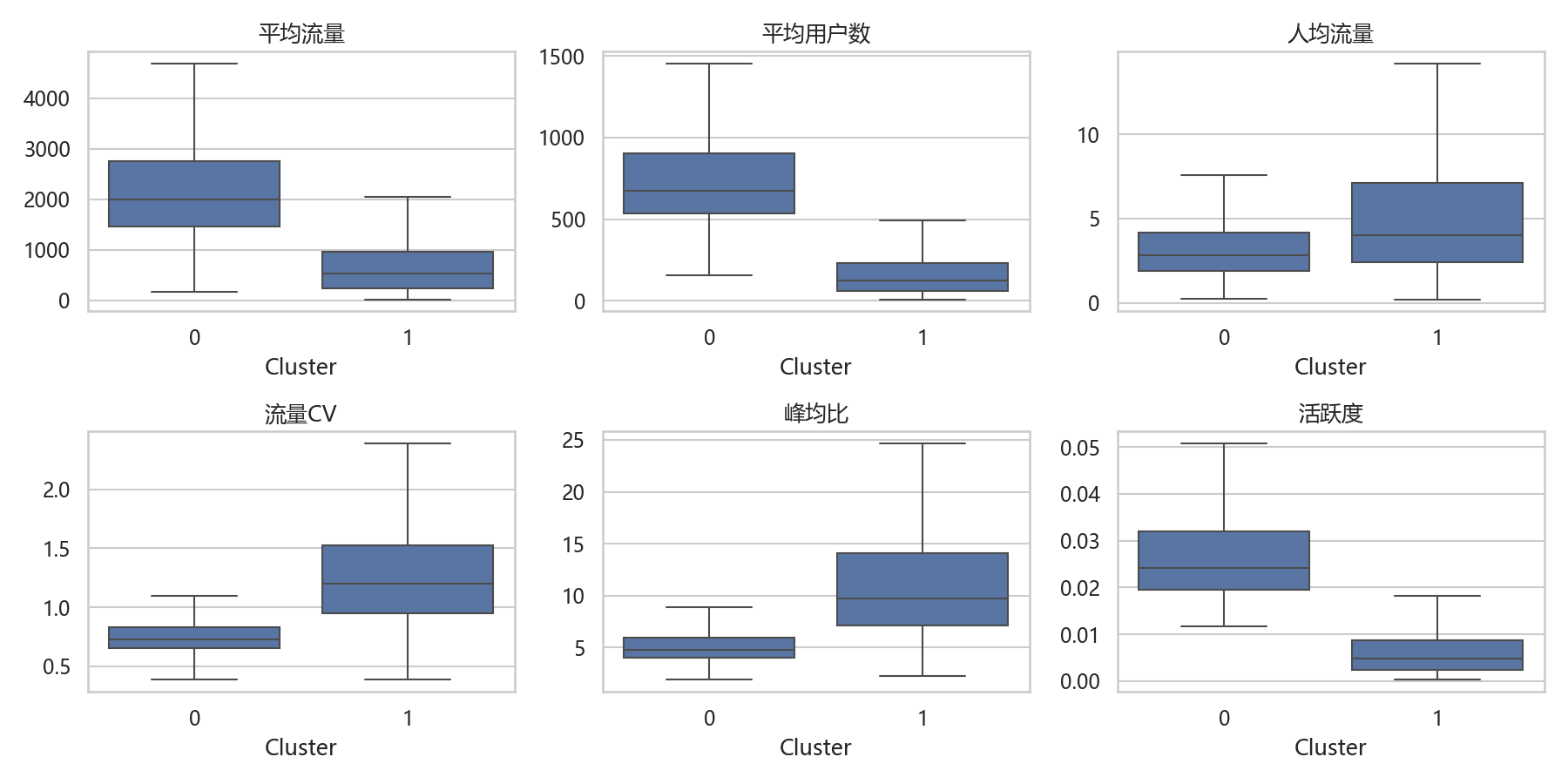
  
含义：该图展示聚类与场景的标准化残差。  
意义：标准化残差, 衡量偏离期望分布的强弱。  
用法：标准化残差, 用于定位显著过度/不足代表的场景。

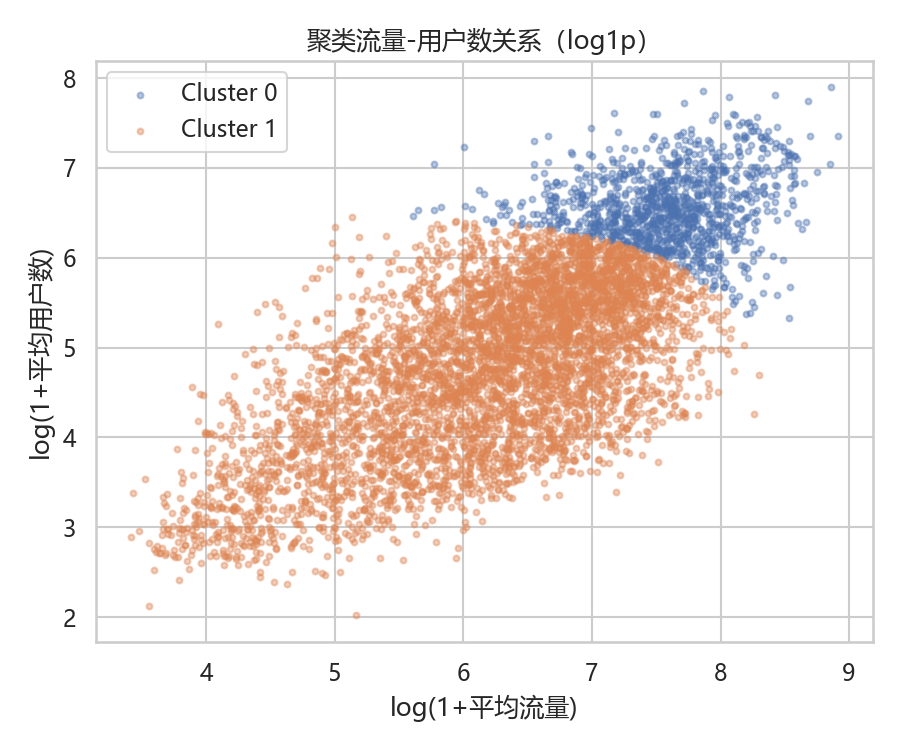
  
含义：该图展示聚类与TYPE的占比分布。  
意义：聚类与TYPE占比, 对比结构差异而非绝对规模。  
用法：聚类与TYPE占比, 作为聚类与类型关系的补充证据。

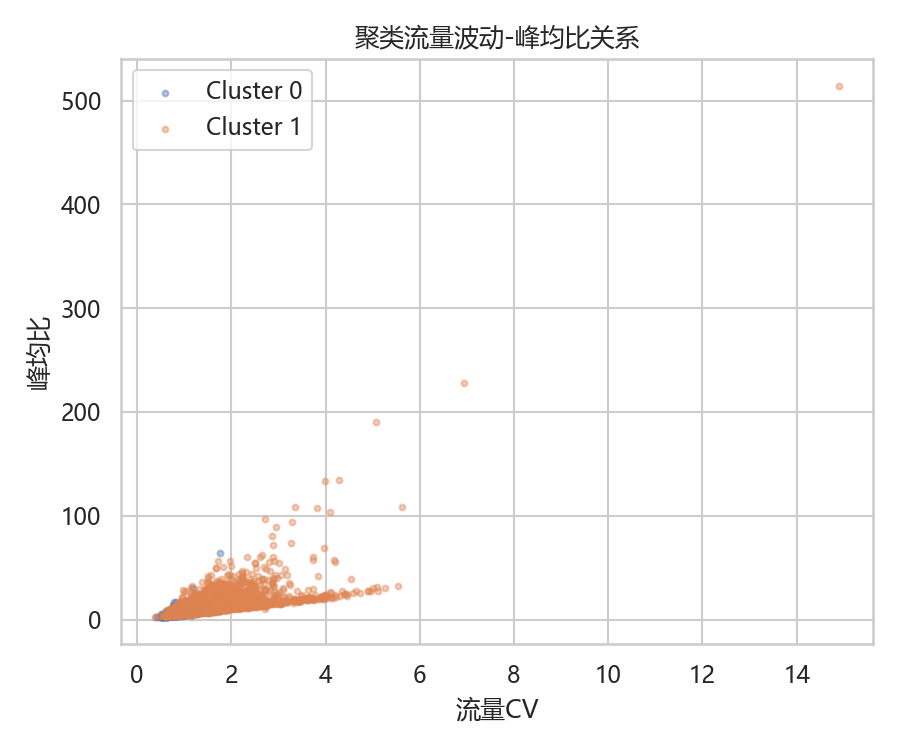
  
含义：该图展示聚类在Top场景上的占比组合。  
意义：场景占比组合, 理解聚类结构的主体场景构成。  
用法：场景占比组合, 作为引导场景分层的可视化依据。

  
含义：该图展示聚类与TYPE的标准化残差。  
意义：标准化残差, 衡量比例偏离程度。  
用法：标准化残差, 识别类型对聚类的贡献方向。

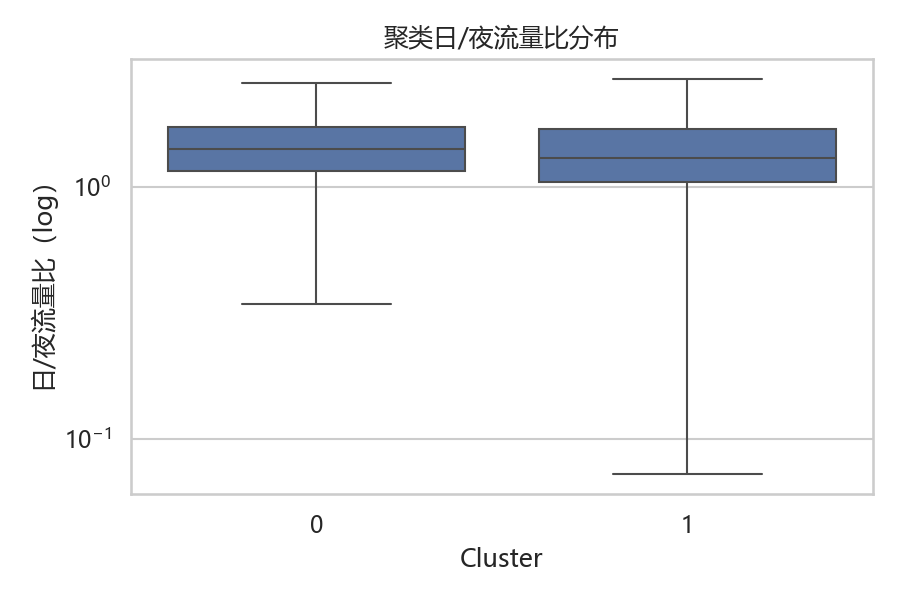
  
含义：该图展示两个聚类在场景占比上的差值。  
意义：场景占比差异, 给出直观的优势/劣势场景对比。  
用法：场景占比差异, 定位两类聚类的核心场景差异。

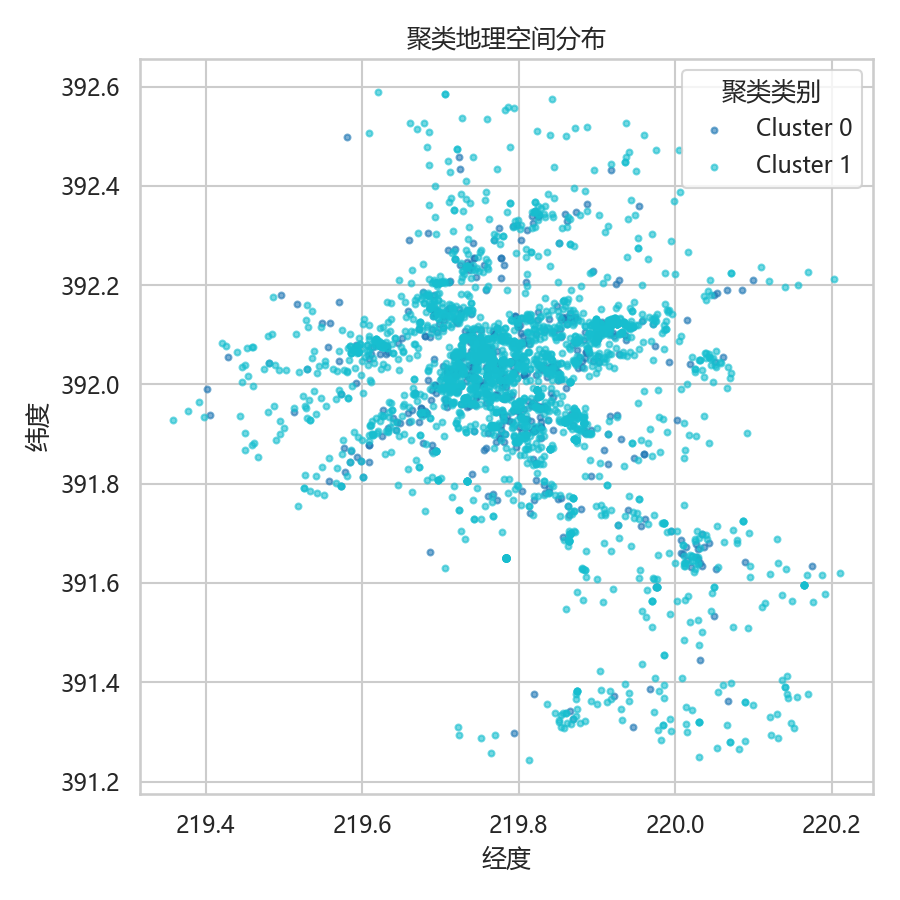
  
????????????????????????????  
?????????????????????????????  
???????????????????????????

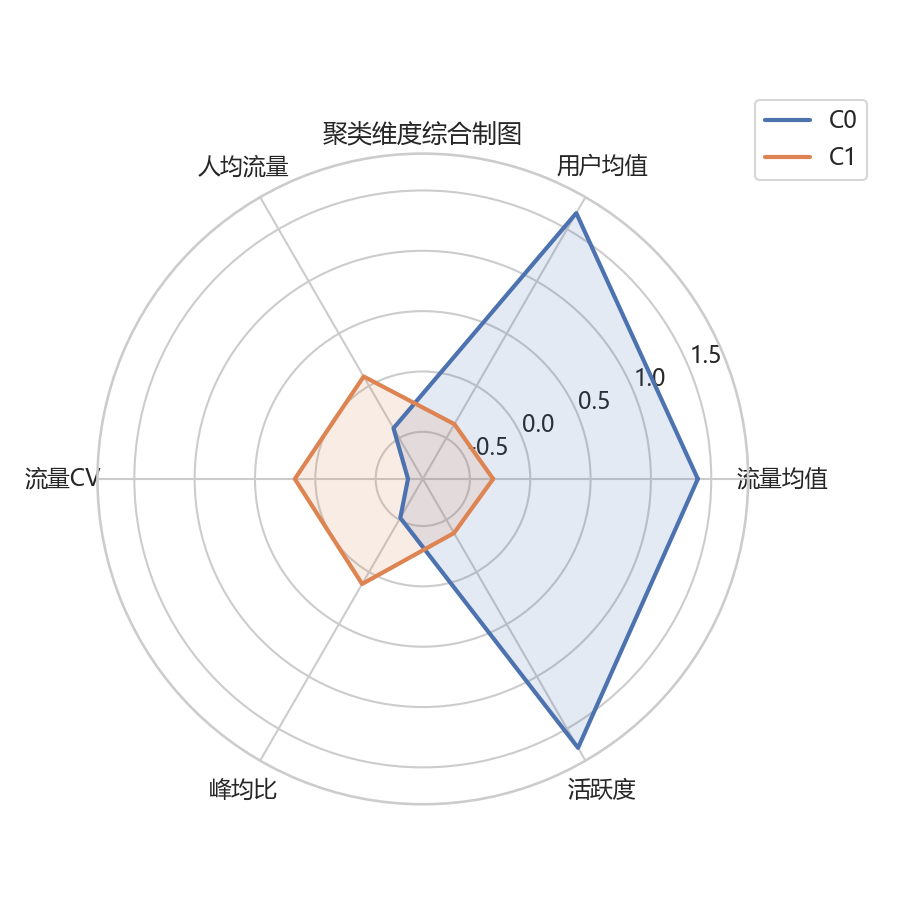
  
??????????????????????????????  
??????????????????????????  
??????????????????????????????

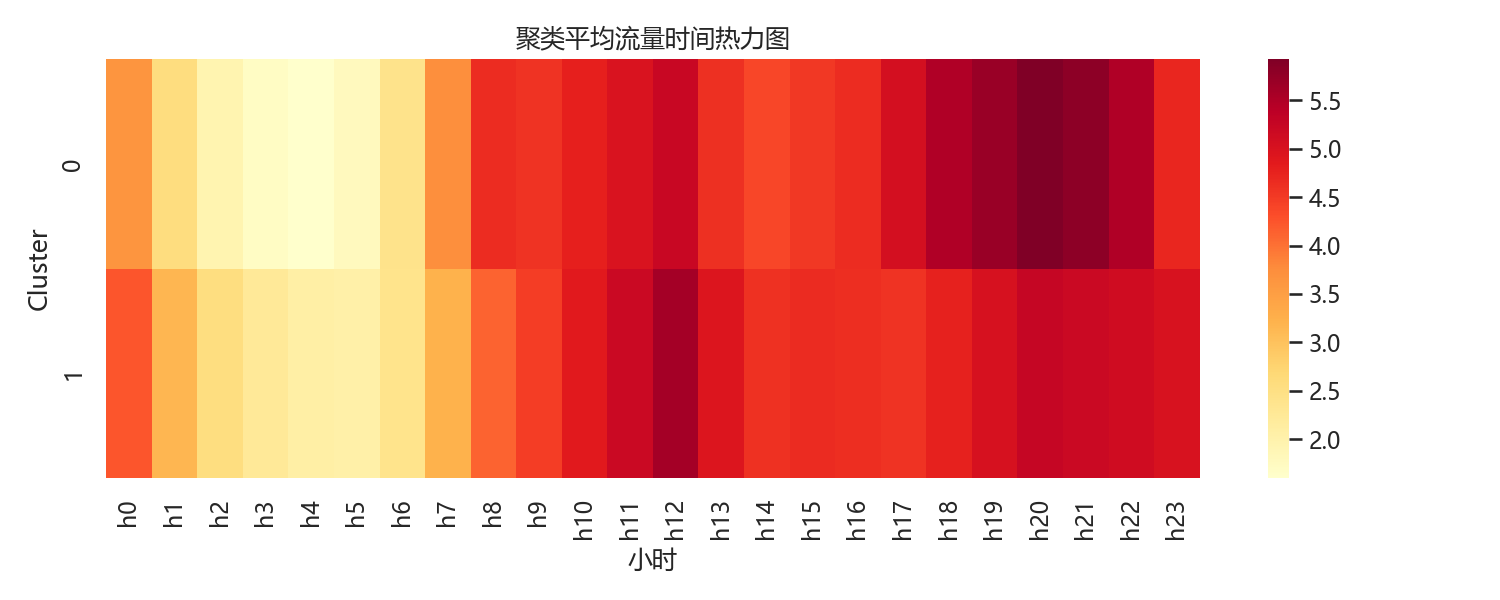
  
????????????CV????????????????  
???????????????????????  
?????????????????????

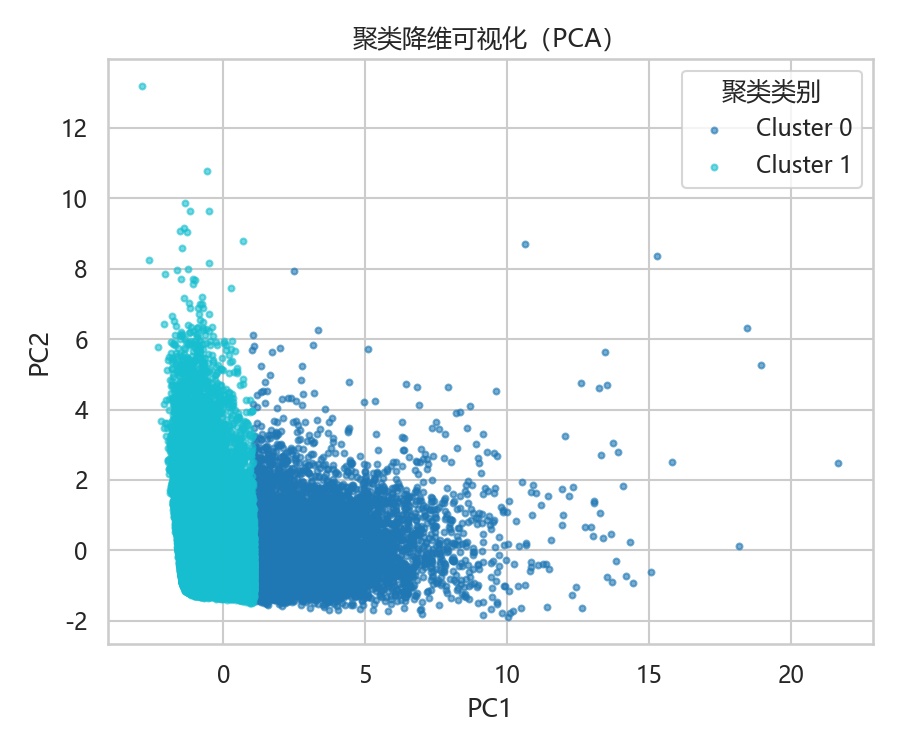
  
?????????????????????????  
????????????????????/??????  
??????????????????????

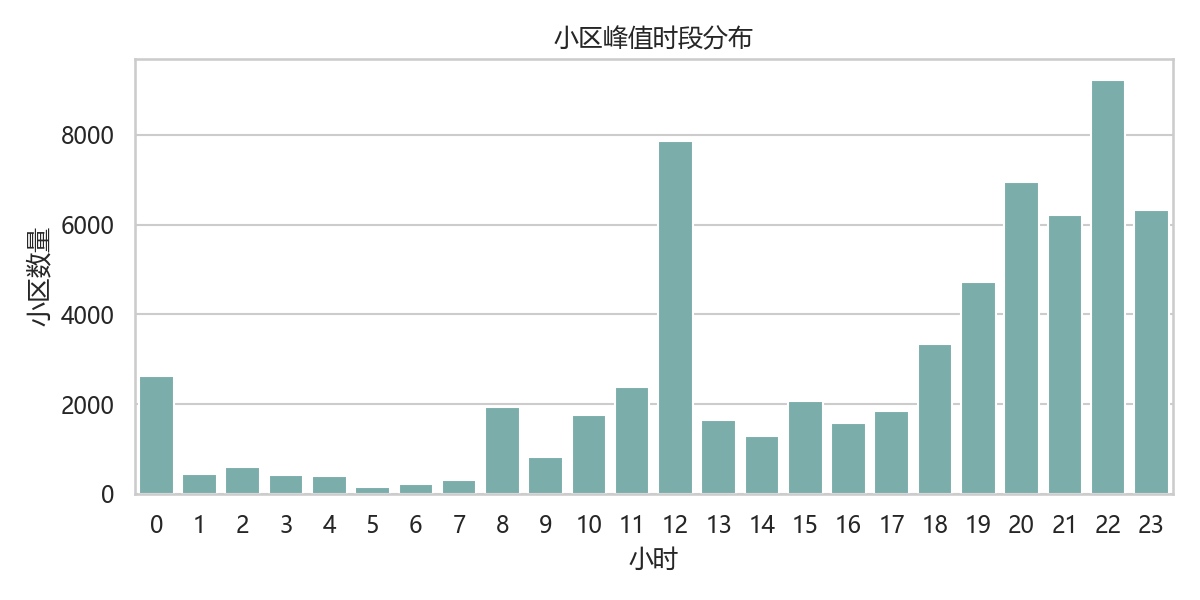
  
???????????????????????  
??????????????????????????  
??????????????????????

  
含义：该图展示聚类空间分布。  
意义：聚类空间分布, 揭示空间分布与热点区域。  
用法：聚类空间分布, 定位重点区域进行扩容或优化。

  
含义：该图展示聚类综合雷达图。  
意义：聚类综合雷达图, 展示多指标画像差异与对照。  
用法：聚类综合雷达图, 用于对比不同聚类/场景的综合特征。

  
含义：该图展示聚类时间热力图。  
意义：聚类时间热力图, 揭示不同维度组合下的强度分布与热点区域。  
用法：聚类时间热力图, 定位高值/低值区间，用于时段、场景或类型优化。

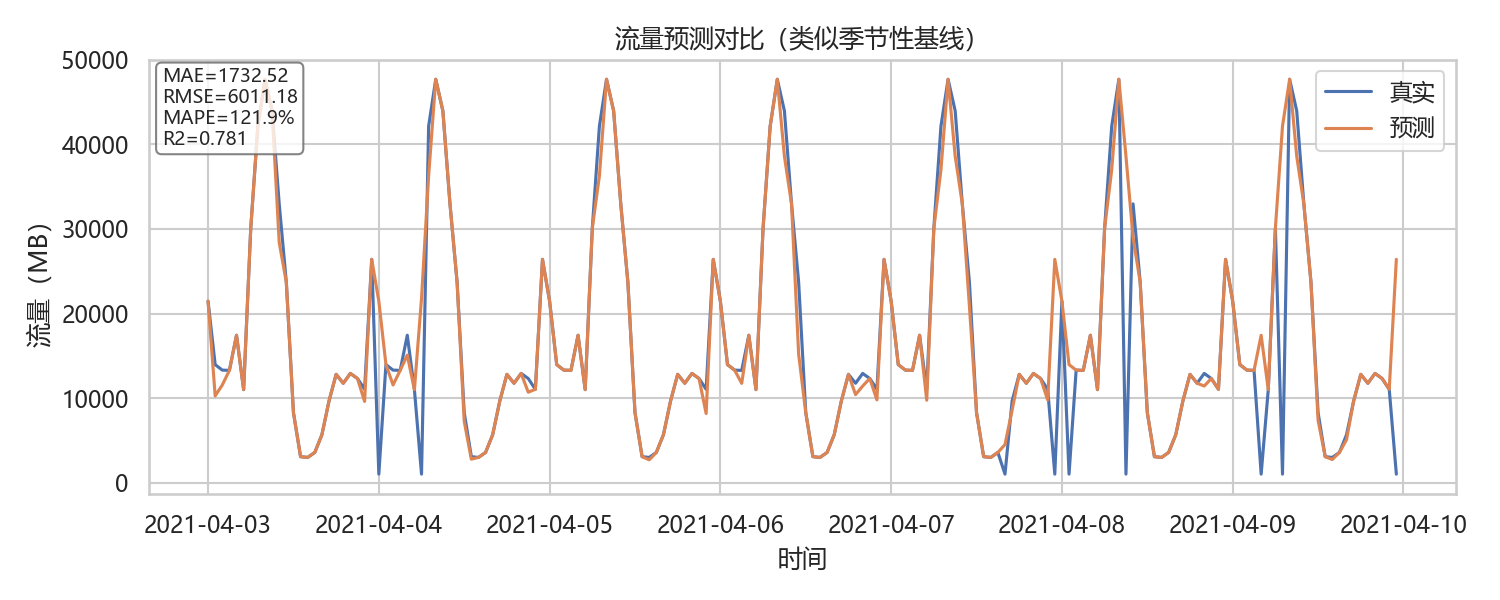
  
含义：该图展示聚类 PCA 可视化。  
意义：聚类 PCA 可视化, 评估聚类在低维空间的可分性。  
用法：聚类 PCA 可视化, 用于验证聚类结果的区分度与重叠情况。

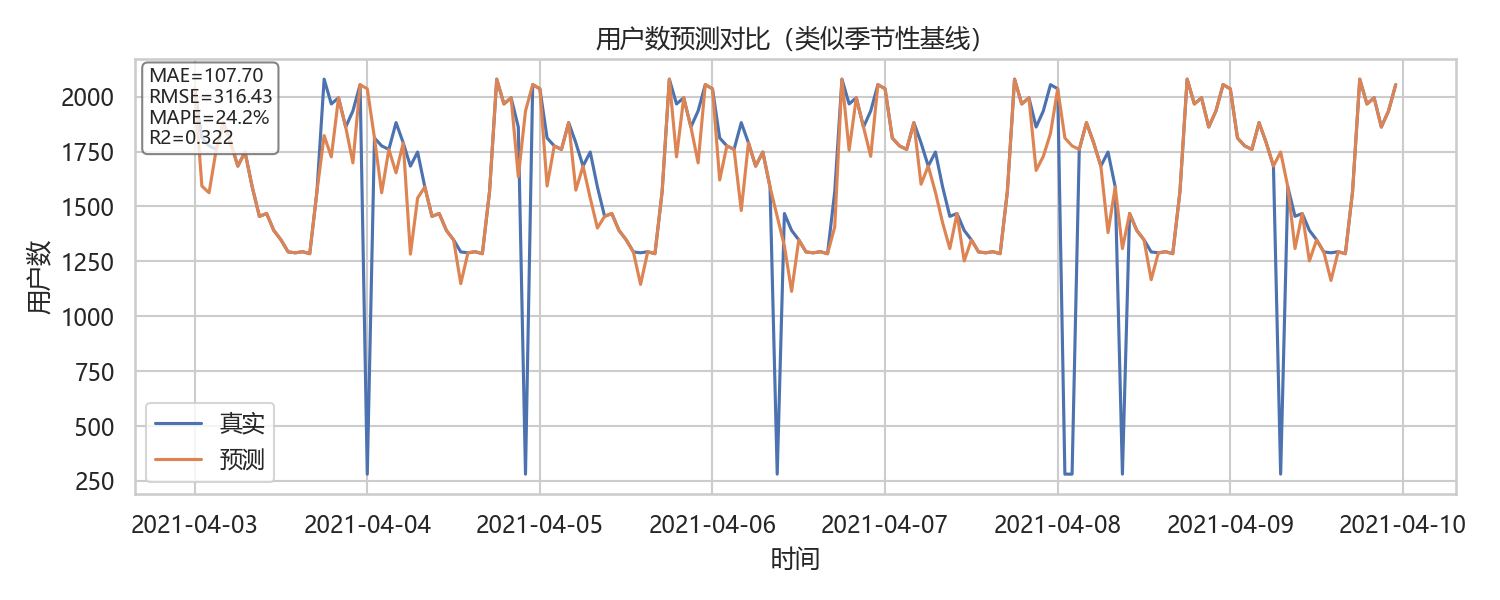
  
含义：该图展示小区峰值时段分布。  
意义：小区峰值时段分布, 展示峰值时段集中程度与分布形态。  
用法：小区峰值时段分布, 用于错峰策略与资源调度。

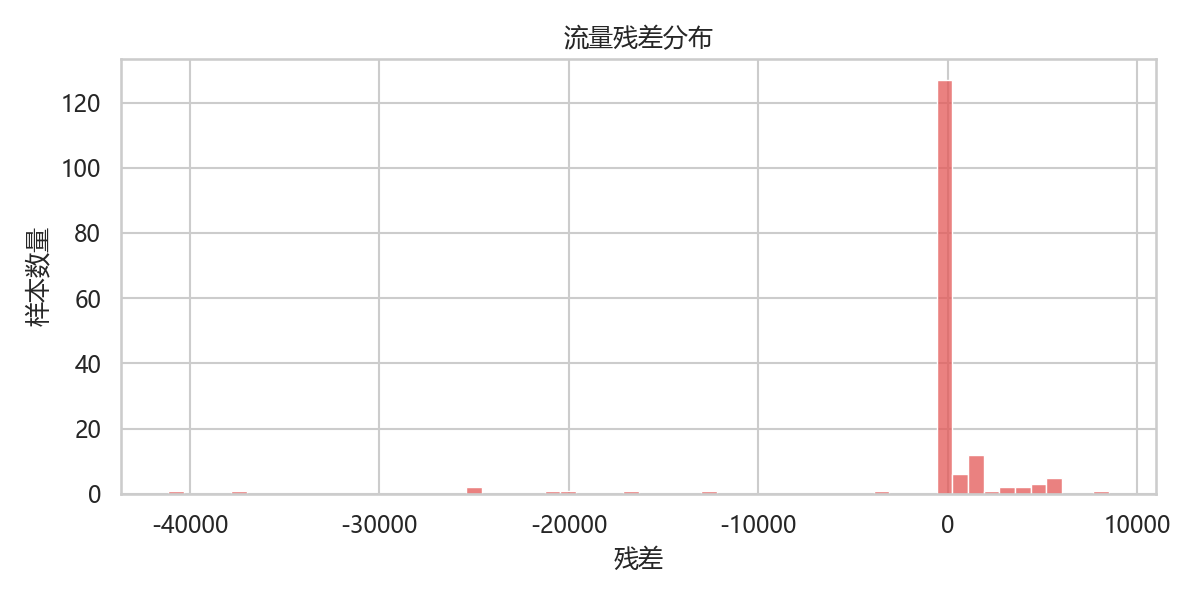
### 5.2 趋势预测（小时-星期基线模型）

关键数据： target\_cell=621; test\_days=7; flow MAE=1732.52, RMSE=6011.18, MAPE=121.9%, R2=0.781; user MAE=107.70, RMSE=316.43, MAPE=24.18%, R2=0.322.

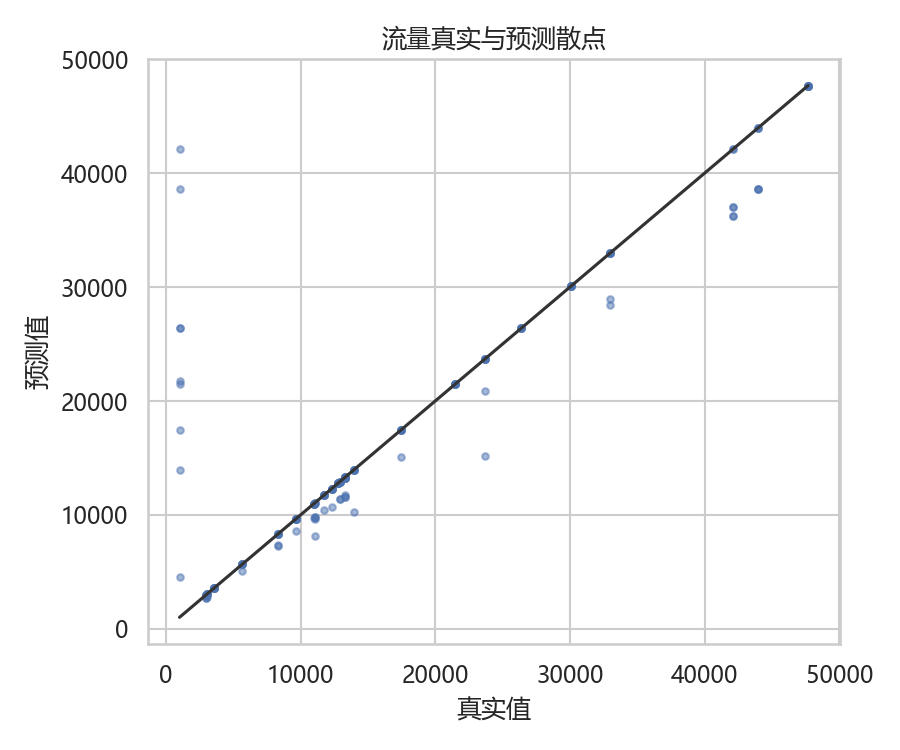
以流量贡献最高的小区（CELL\_ID=621）为样本，基于历史小时级流量与用户数序列构建“星期×小时”均值基线，预测未来 7 天的小时级流量与用户数。预测结果表明：流量 MAE=1732.52、RMSE=6011.18、MAPE=121.9%、R2=0.781；用户数 MAE=107.70、RMSE=316.43、MAPE=24.18%、R2=0.322。该基线模型能刻画整体周期，但在高波动时段仍有残差，说明需引入节假日与场景等外部变量提升拟合度。

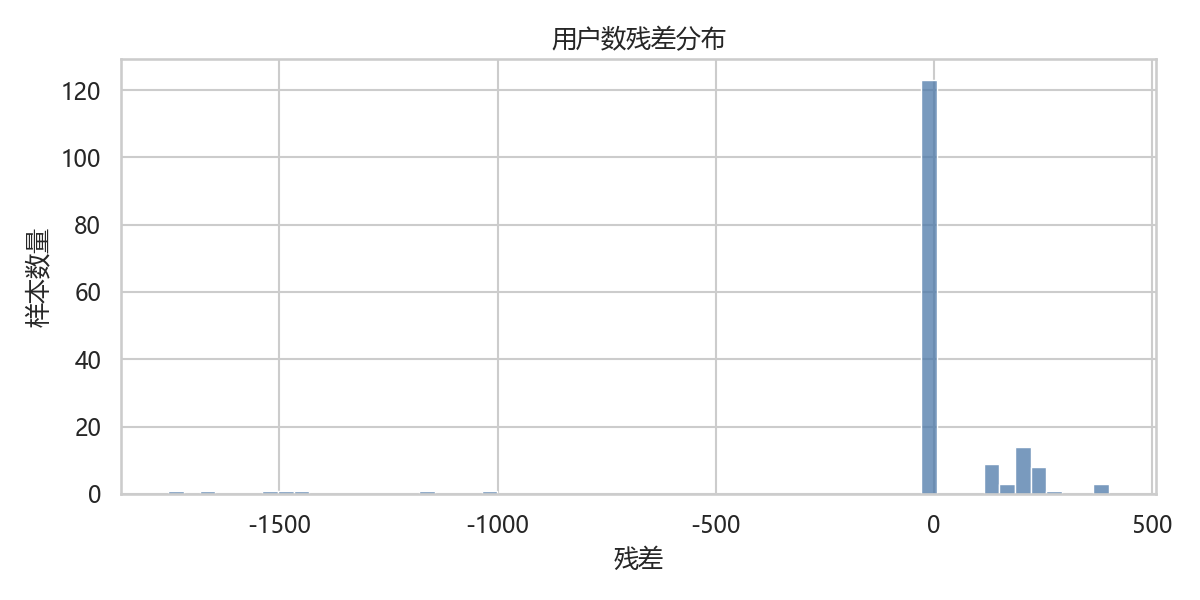
  
含义：该图展示流量预测对比。  
意义：流量预测对比, 评估预测效果与偏差水平。  
用法：流量预测对比, 根据误差诊断是否需引入更多特征。

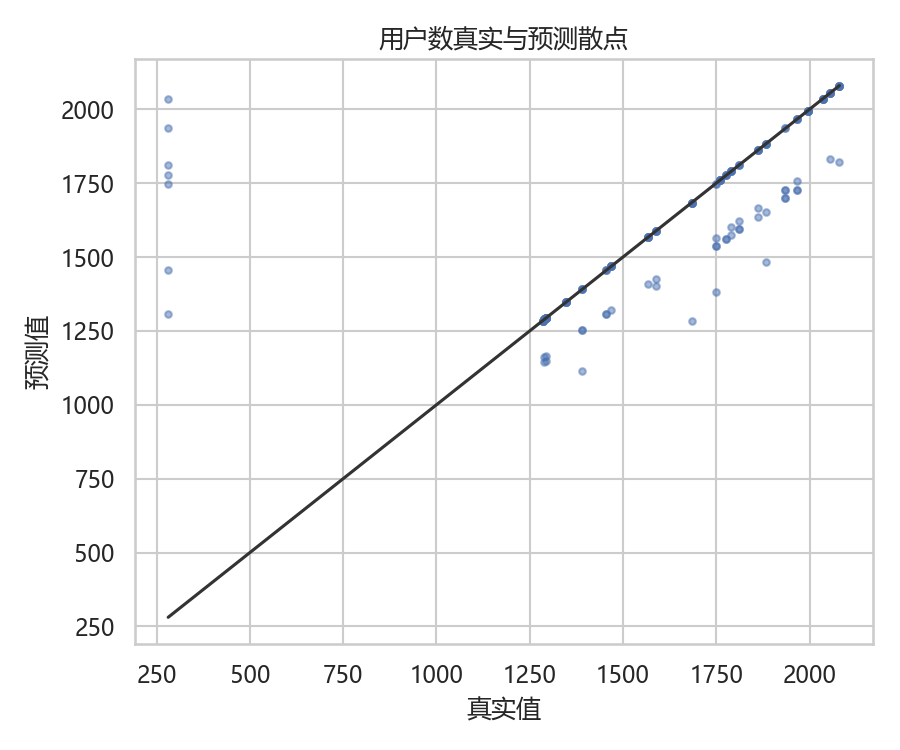
  
含义：该图展示用户数预测对比。  
意义：用户数预测对比, 评估预测效果与偏差水平。  
用法：用户数预测对比, 根据误差诊断是否需引入更多特征。

  
含义：该图展示流量残差分布。  
意义：流量残差分布, 刻画误差分布与系统性偏差。  
用法：流量残差分布, 定位高误差时段，优化模型或特征。

  
含义：该图展示流量残差走势。  
意义：流量残差走势, 刻画误差分布与系统性偏差。  
用法：流量残差走势, 定位高误差时段，优化模型或特征。

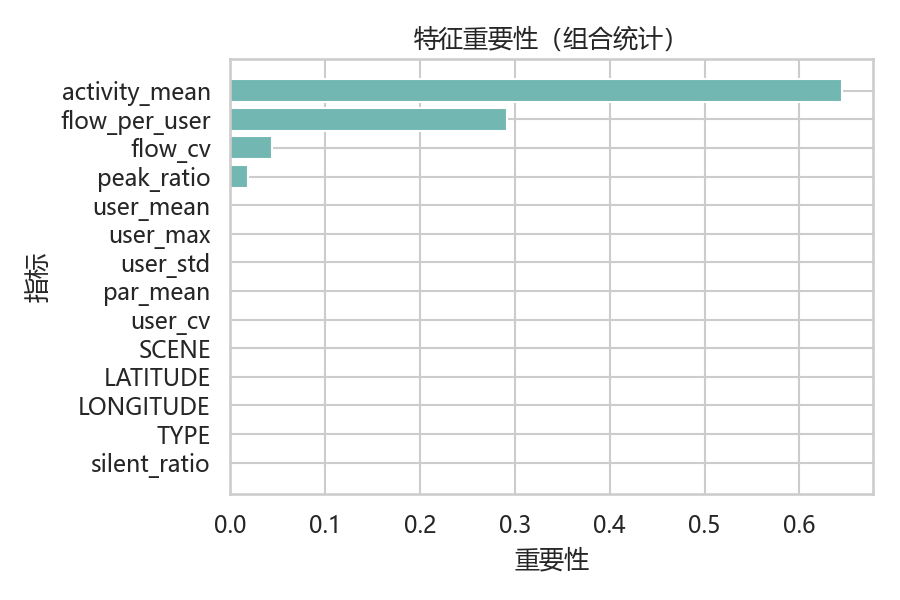
  
含义：该图展示流量真实-预测散点。  
意义：流量真实-预测散点, 揭示变量关系与离群点分布。  
用法：流量真实-预测散点, 识别高流量低用户等异常组合。

  
含义：该图展示用户数残差分布。  
意义：用户数残差分布, 刻画误差分布与系统性偏差。  
用法：用户数残差分布, 定位高误差时段，优化模型或特征。

  
含义：该图展示用户数真实-预测散点。  
意义：用户数真实-预测散点, 揭示变量关系与离群点分布。  
用法：用户数真实-预测散点, 识别高流量低用户等异常组合。

补充：基于小区层面指标的粗粒度预测。使用用户规模、活跃度、静默与空间特征（user\_mean、user\_max、user\_std、user\_cv、activity\_mean、par\_mean、silent\_ratio、flow\_cv、peak\_ratio、flow\_per\_user、LATITUDE、LONGITUDE）以及场景/类型标签（SCENE、TYPE），通过梯度提升回归预测小区平均流量（flow\_mean）。结果：MAE=17.88、RMSE=31.69、MAPE=2.26%、R2=0.999，说明这些指标可解释约 99.9% 的平均流量差异，适合作为分层与粗估依据。

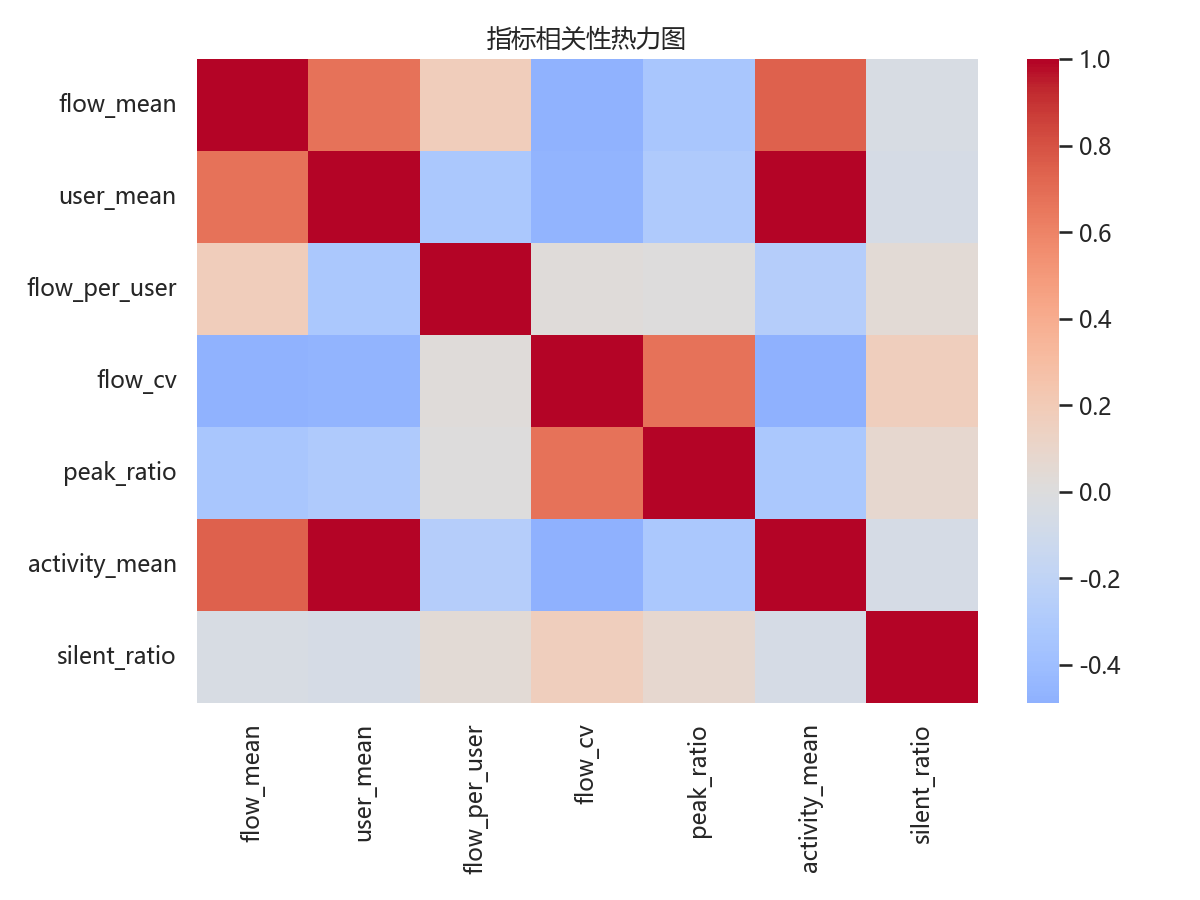
  
含义：该图展示小区平均流量预测散点。  
意义：小区平均流量预测散点, 评估指标驱动预测的拟合度与离群点。  
用法：小区平均流量预测散点, 用于粗估小区流量水平与异常识别。

  
含义：该图展示预测特征重要性。  
意义：预测特征重要性, 展示各指标对平均流量预测的相对影响程度。  
用法：预测特征重要性, 用于确定关键指标与优化数据采集重点。

### 5.3 指标相关性洞察

关键数据： cell\_corr(flow\_mean,user\_mean)=0.669 (相关性明显但不完全一致).

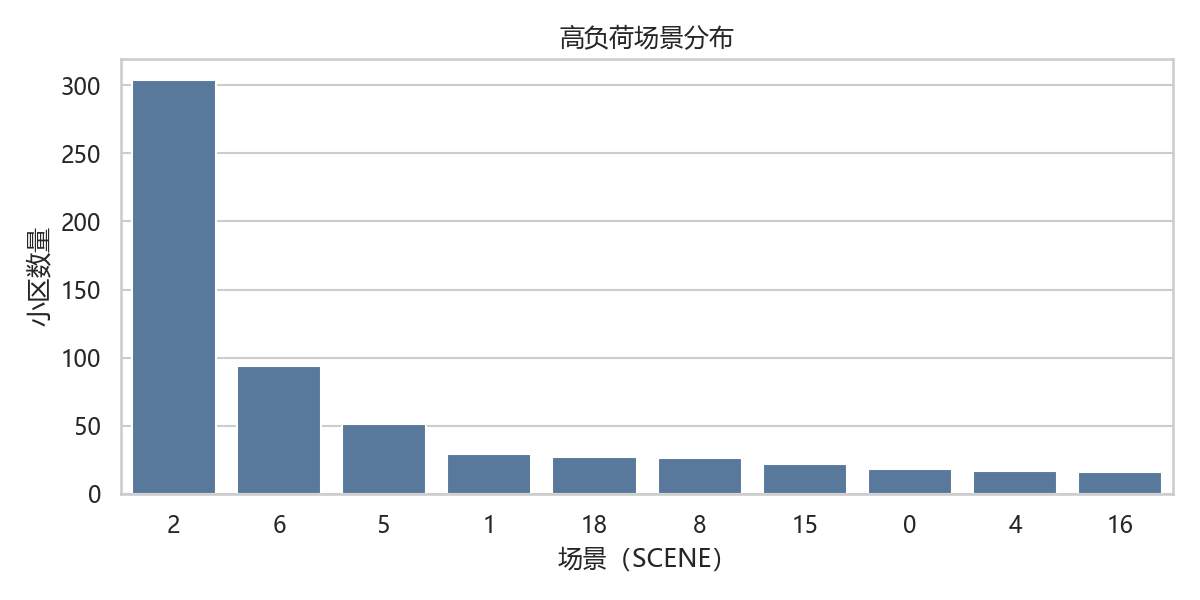
基于小区层面的多指标相关性，流量、用户数、人均流量与活跃度之间存在明显相关结构，可用于后续的特征筛选与异常解释。

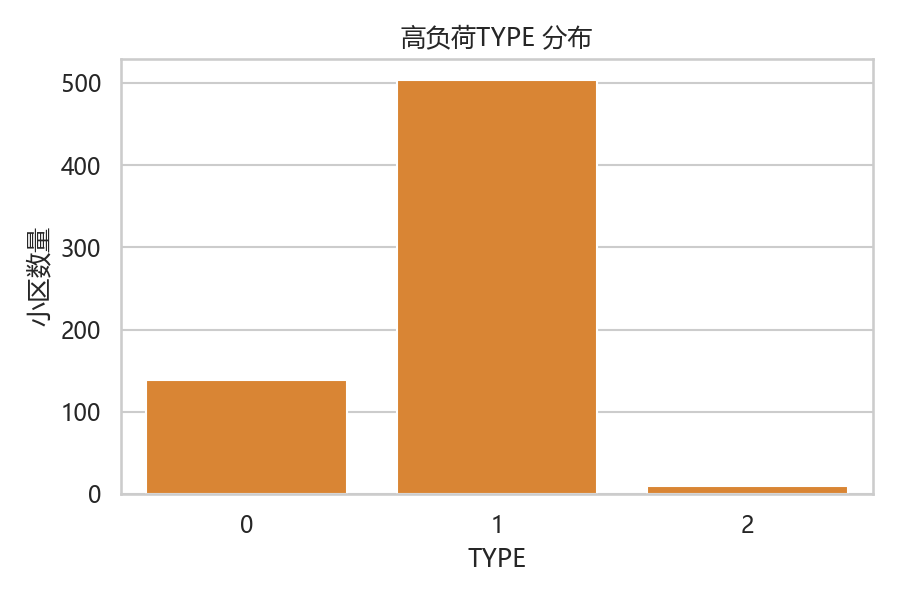
  
含义：该图展示指标相关性热力图。  
意义：指标相关性热力图, 揭示不同维度组合下的强度分布与热点区域。  
用法：指标相关性热力图, 定位高值/低值区间，用于时段、场景或类型优化。

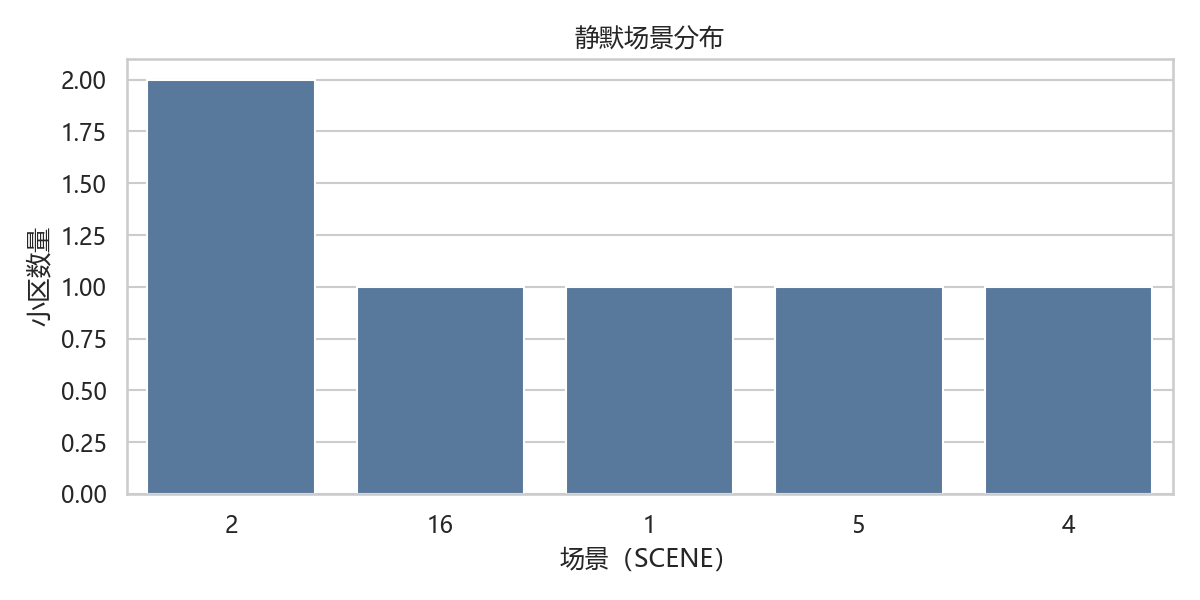
### 5.4 高负荷/静默小区画像

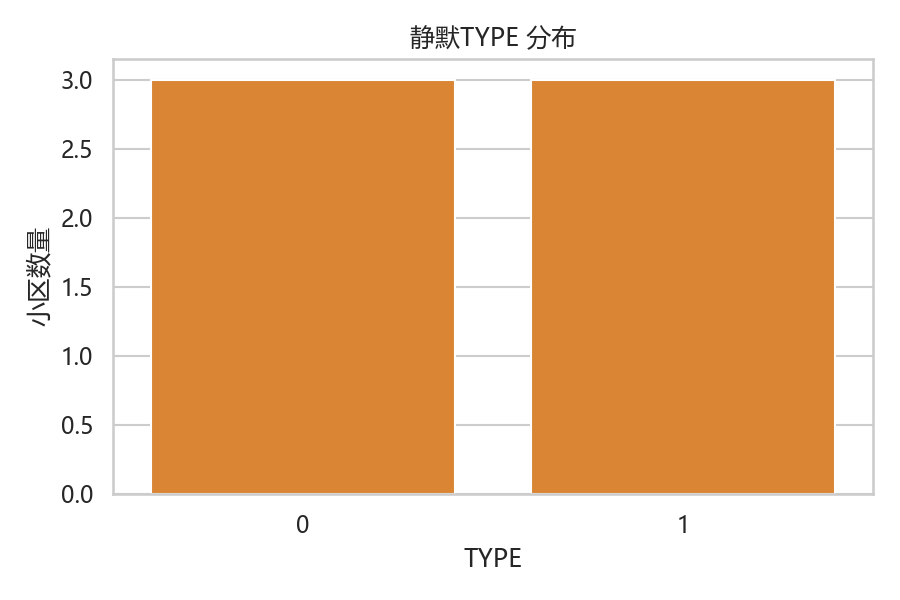
关键数据： highload threshold=6,093,868.32 MB (652 cells, 1.00%); silent threshold=0.5 (6 cells, 0.009%); highload top20 flow range=12,702,654~24,892,078 MB; silent ratio range=0.5875~0.9688 (n=6).

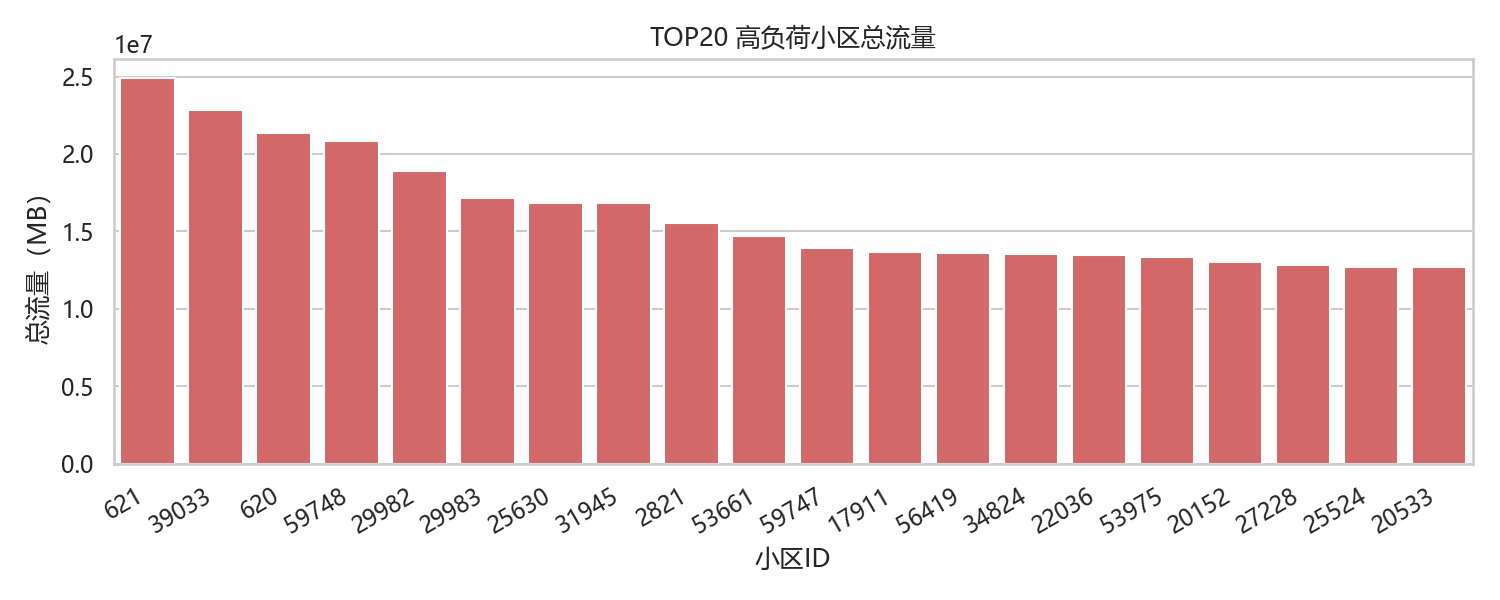
结合第 3 部分的高负荷阈值与静默比例阈值，对高负荷与静默小区进行场景、类型分布对比，并输出 top20 个体画像。

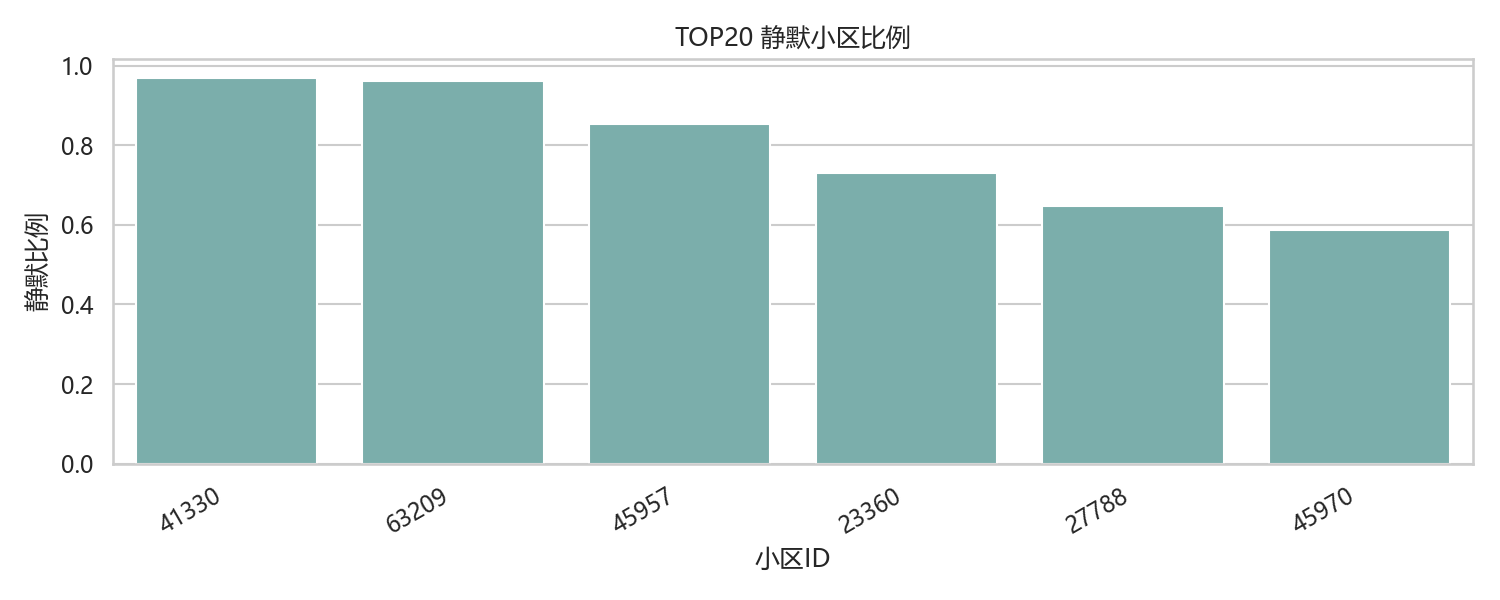
  
含义：该图展示高负荷小区场景分布。  
意义：高负荷小区场景分布, 判断集中程度、偏态与长尾特征。  
用法：高负荷小区场景分布, 用于设置阈值、识别异常或分层。

  
含义：该图展示高负荷小区 TYPE 分布。  
意义：高负荷小区 TYPE 分布, 判断集中程度、偏态与长尾特征。  
用法：高负荷小区 TYPE 分布, 用于设置阈值、识别异常或分层。

  
含义：该图展示静默小区场景分布。  
意义：静默小区场景分布, 判断集中程度、偏态与长尾特征。  
用法：静默小区场景分布, 用于设置阈值、识别异常或分层。

  
含义：该图展示静默小区 TYPE 分布。  
意义：静默小区 TYPE 分布, 判断集中程度、偏态与长尾特征。  
用法：静默小区 TYPE 分布, 用于设置阈值、识别异常或分层。

  
含义：该图展示TOP20 高负荷小区总流量。  
意义：TOP20 高负荷小区总流量, 揭示头部集中度与贡献结构。  
用法：TOP20 高负荷小区总流量, 确定重点对象或场景优先级。

  
含义：该图展示TOP20 静默小区比例。  
意义：TOP20 静默小区比例, 揭示头部集中度与贡献结构。  
用法：TOP20 静默小区比例, 确定重点对象或场景优先级。

### 5.5 结论摘要

* 聚类显示明显的“午间主峰型”与“夜间活跃型”小区结构，对运维策略可进行差异化配置。
* 预测基线能抓住周期性，但高波动时段误差偏大，后续可引入节假日与场景变量优化。
* 相关性热力图说明流量、用户、活跃度可形成稳定的联合指标体系。

### 5.6 代码与产出

* 代码文件：section5\_advanced.py
* 统计汇总：report\_assets/section5/section5\_stats.json
* ?????42 ???report\_assets/section5/fig01\_elbow\_kmeans.png ? report\_assets/section5/fig41\_cluster\_day\_night\_ratio.png
* 个体画像输出：report\_assets/section5/highload\_top20.csv、report\_assets/section5/silent\_top20.csv## 6. 商业价值建议

网络扩容建议：针对高负荷、高增速的小区，提出增加基站部署的方案。

节能降耗策略：针对深夜流量极低、场景稳定的区域，建议实施动态基站关停以节省能耗。

精准营销参考：识别高价值活跃场景，为运营商推送针对性套餐提供数据支持。

## 7. 结论与反思

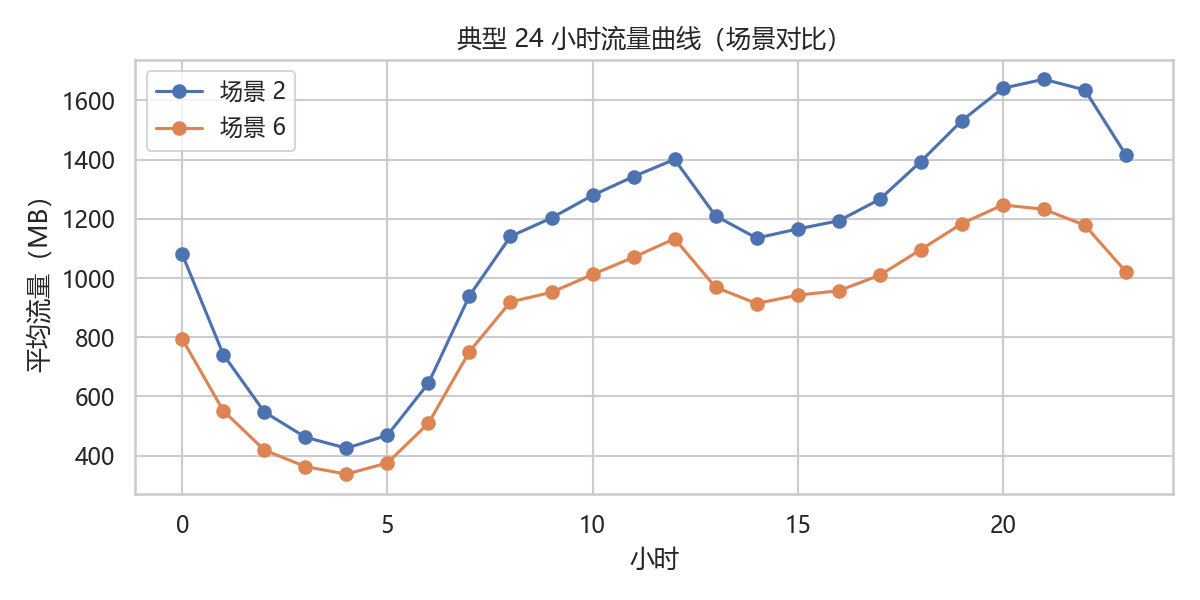
### 7.1 技术难点与解决方案

* 数据体量大（近亿条记录），单机内存难以一次性加载。解决方案：采用分块读取、分组聚合、按需抽样绘图，并将中间结果落盘。
* 多维指标复杂（时间、场景、类型、节假日），易出现口径不一致。解决方案：统一以 DATETIME\_KEY 生成日期/小时/星期/节假日特征，并固定清洗规则（负值置为缺失）。
* 图表数量多且维度跨度大。解决方案：将绘图流程模块化，分别输出场景对比、节假日对比、日历热力图与趋势对比图。

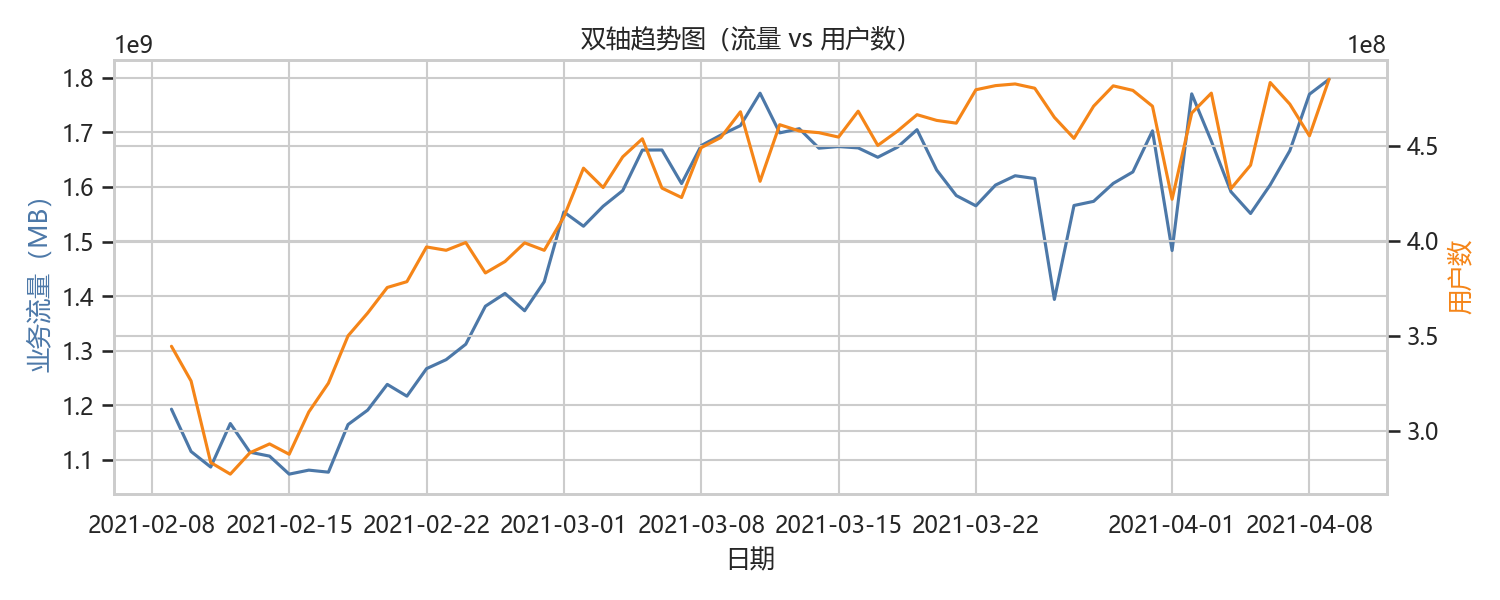
### 7.2 关键结论与图表支撑

关键数据： scene2 peak=1671.16 @21, trough=425.32 @4; scene6 peak=1246.77 @20, trough=337.83 @4; daily corr(flow,user)=0.911; daily max=1,797,291,580 (2021-04-09), min=1,073,917,202 (2021-02-15), diff=67.4%; holiday daily mean: flow 1.253e9 vs 1.545e9 (-18.9%), user 3.41e8 vs 4.37e8 (-22.0%); daily |change| p90=7.41%.

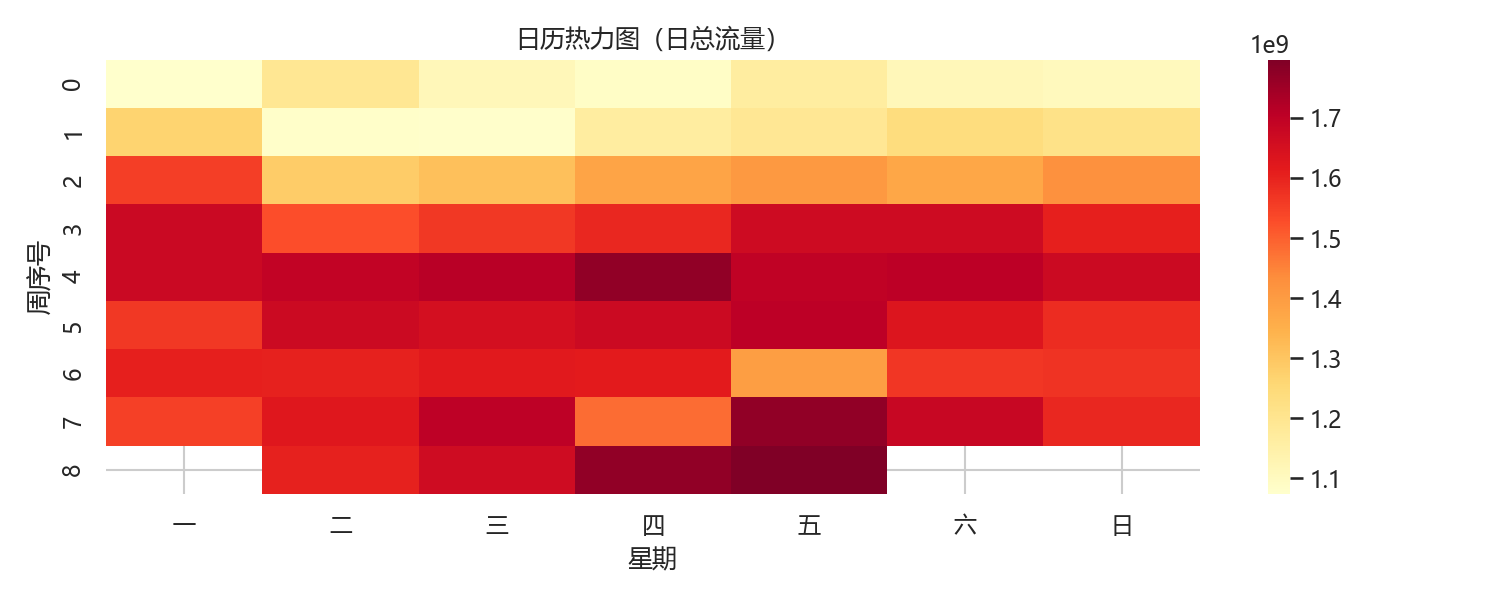
* 典型 24 小时流量曲线显示，不同场景存在明显的昼夜差异，适用于差异化运维策略。

  
含义：该图展示24小时典型流量曲线（场景2 vs 6）。  
意义：24小时典型流量曲线（场景2 vs 6）, 刻画随时间的变化与周期性规律。  
用法：24小时典型流量曲线（场景2 vs 6）, 识别峰谷与突变点，支持容量与资源配置。

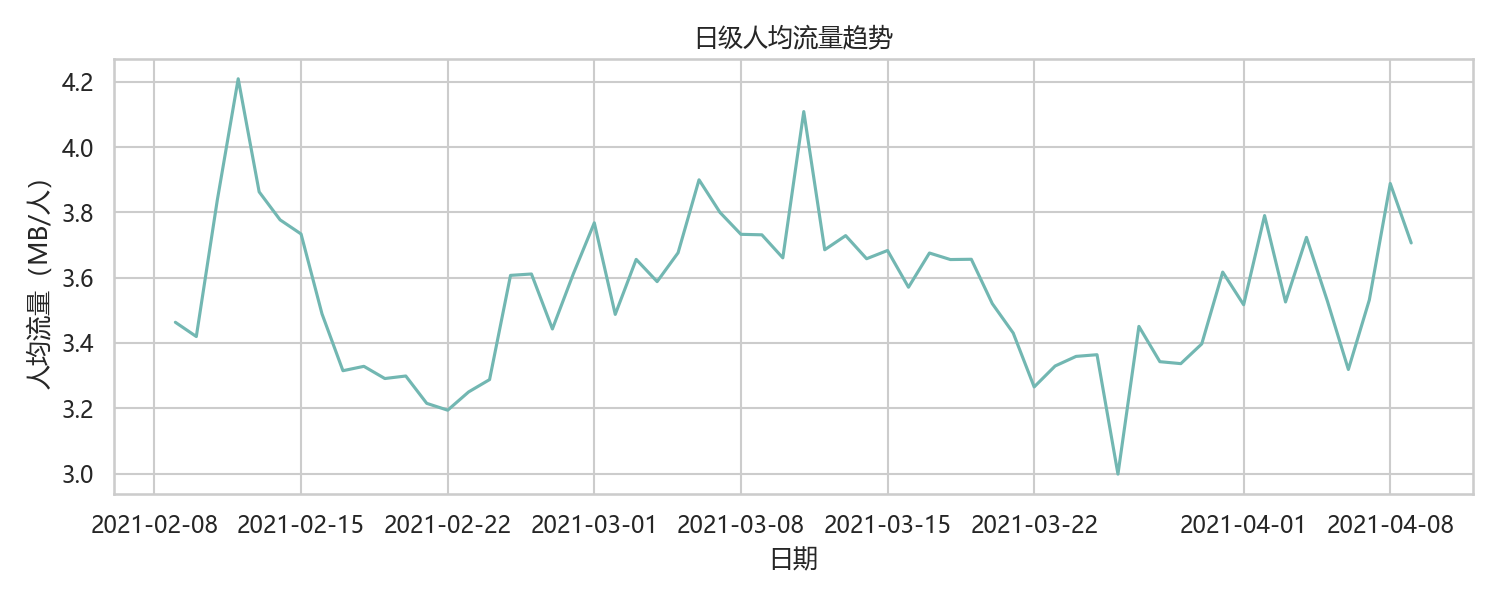
* 双轴趋势图显示流量与用户数大体同步，但在节假日前后出现“高流量/低用户”或“高用户/低流量”的异常波动。

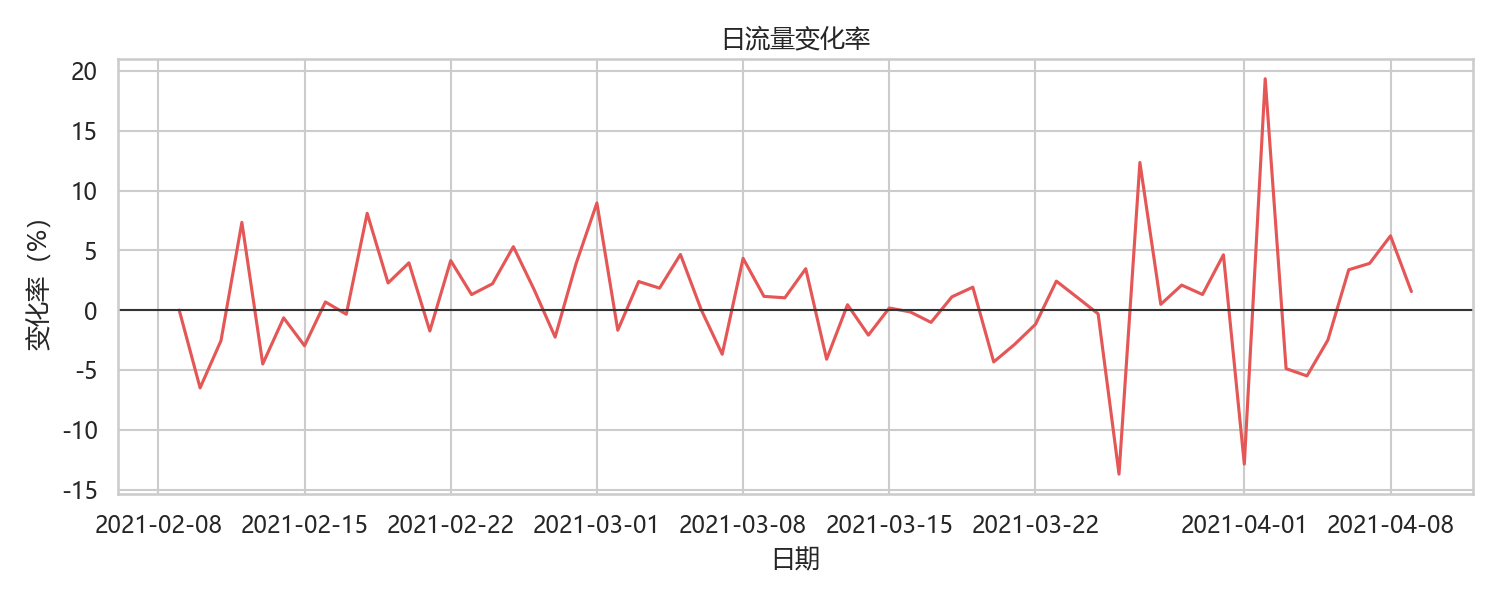
  
含义：该图展示双轴趋势图（流量 vs 用户数）。  
意义：双轴趋势图（流量 vs 用户数）, 对比两指标的同步与背离关系。  
用法：双轴趋势图（流量 vs 用户数）, 识别高流量低用户等异常时段。

* 日历热力图能直观看到春节期间（2 月 12 日附近）的整体流量波动特征。

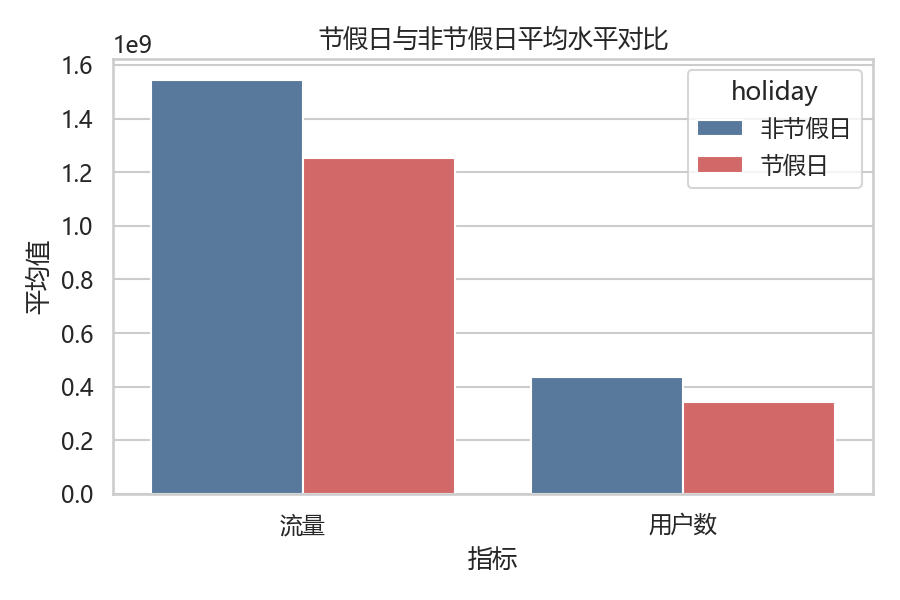
  
含义：该图展示日历热力图（日总流量）。  
意义：日历热力图（日总流量）, 展示日历尺度的波动与节假日效应。  
用法：日历热力图（日总流量）, 识别低谷/异常日期，支持运维排期与预测建模。

* 人均流量与日流量变化率曲线反映出节假日与工作日之间的结构性差异。

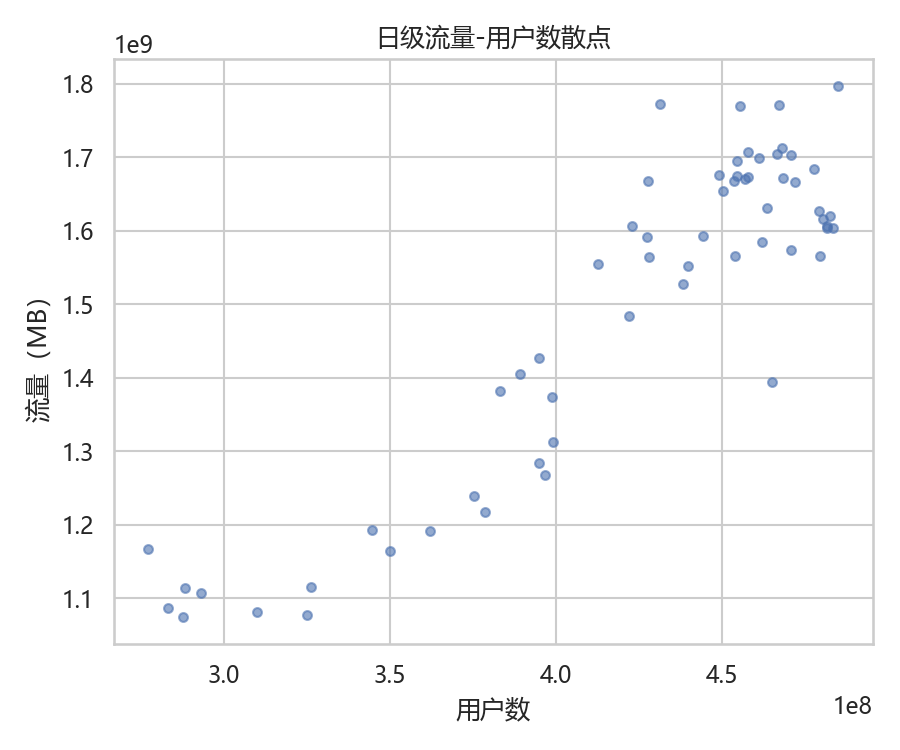
  
含义：该图展示日级人均流量趋势。  
意义：日级人均流量趋势, 刻画随时间的变化与周期性规律。  
用法：日级人均流量趋势, 识别峰谷与突变点，支持容量与资源配置。

  
含义：该图展示日流量变化率。  
意义：日流量变化率, 提供整体结构与差异的直观视角。  
用法：日流量变化率, 辅助定位重点区间与异常样本。

* 节假日与非节假日对比表明，节假日整体流量与用户数水平出现显著变化。

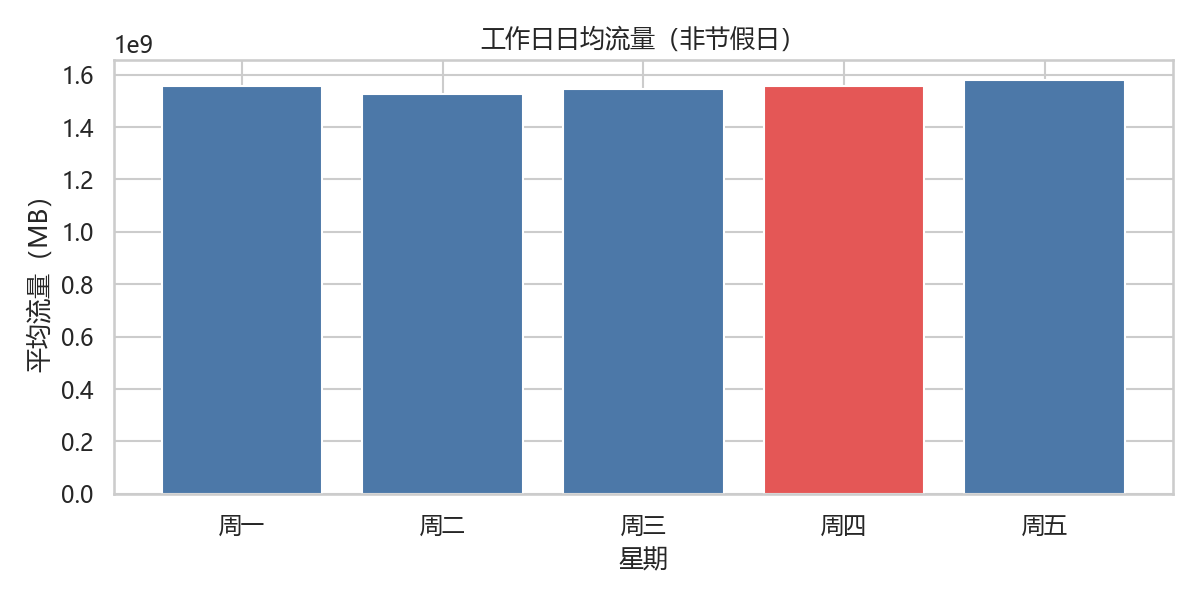
  
含义：该图展示节假日与非节假日对比。  
意义：节假日与非节假日对比, 量化节假日与工作日差异，体现节假日效应。  
用法：节假日与非节假日对比, 用于节假日保障与弹性资源配置。

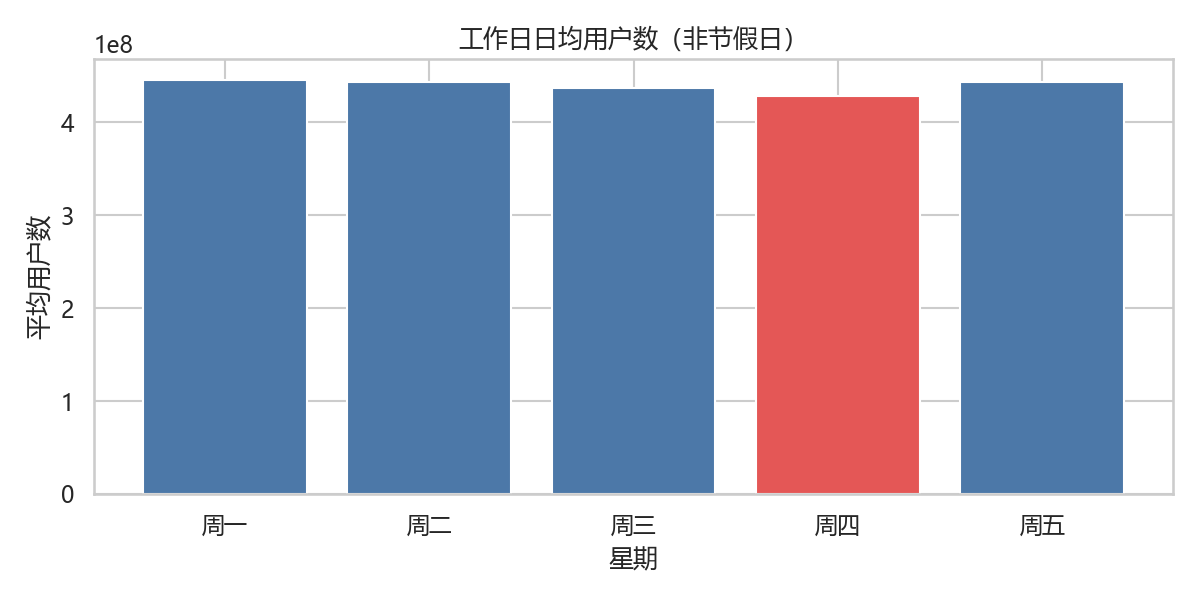
* 日级流量与用户数散点图说明两者整体相关，但存在离群点，提示潜在的异常运营场景。

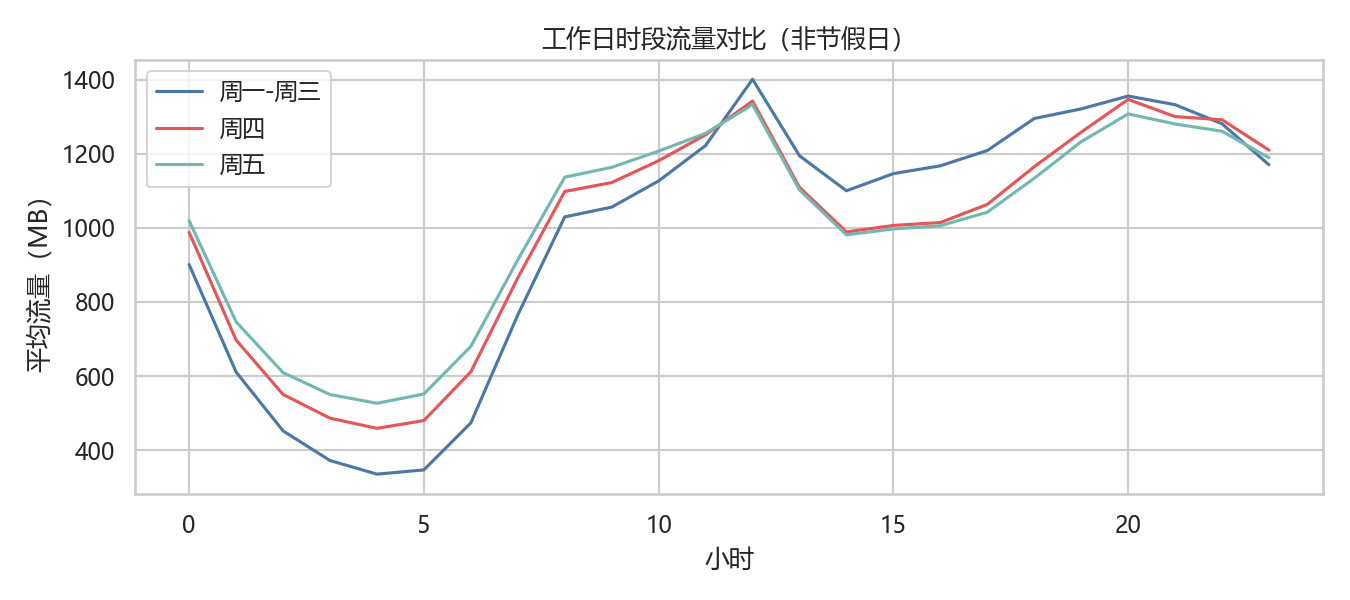
  
含义：该图展示日级流量-用户数散点。  
意义：日级流量-用户数散点, 揭示变量关系与离群点分布。  
用法：日级流量-用户数散点, 识别高流量低用户等异常组合。

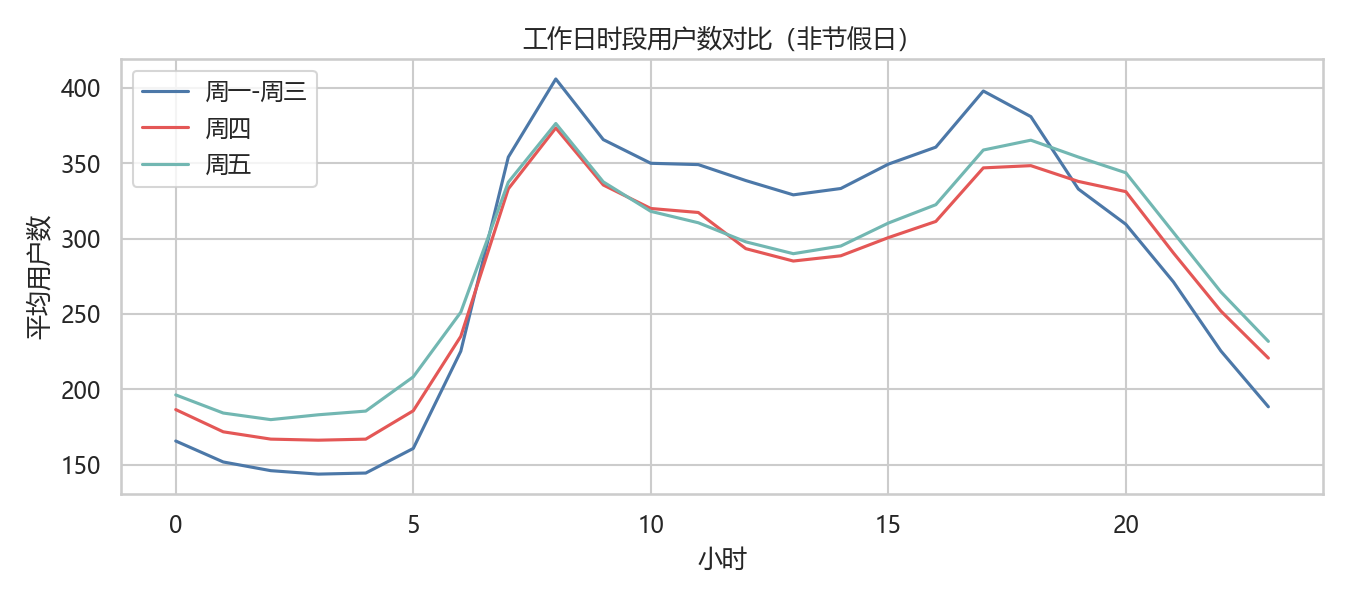
### 7.5 周四流量偏弱专项分析

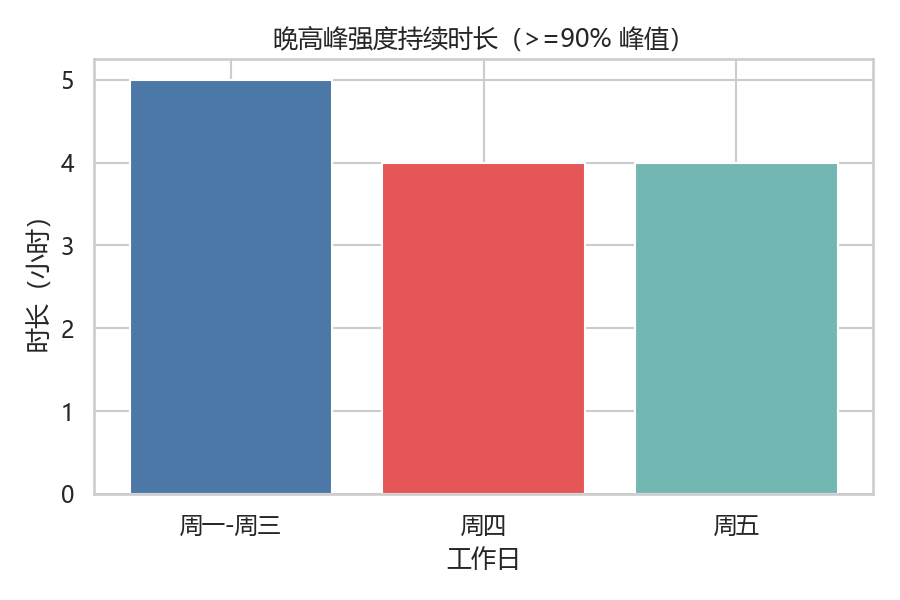
关键数据：非节假日工作日对比，周四日均用户数 4.279e8（比周一-周三均值 4.419e8 低 3.2%）；10:00-17:00 平均流量 1119.5 vs 1195.7（-6.4%），平均用户 307.9 vs 351.0（-12.3%）；18:00-22:00 平均流量 1272.2 vs 1316.8（-3.4%）；晚高峰持续时长（>=90%峰值）周四 4h，周一-周三 5h。

  
含义：该图展示工作日日均流量的星期分布。  
意义：定位周四在工作日中的流量位置与趋势差异。  
用法：用于识别工作日结构性低位时段。

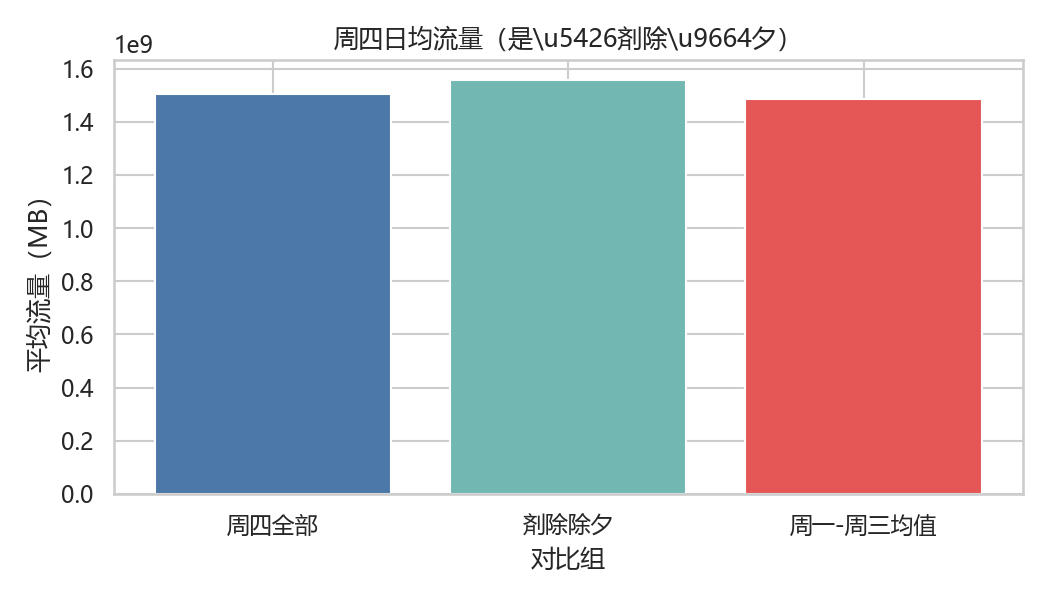
  
含义：该图展示工作日日均用户数的星期分布。  
意义：周四用户均值低位，说明用户活跃度偏弱。  
用法：用于识别工作日用户规模的低位时段。

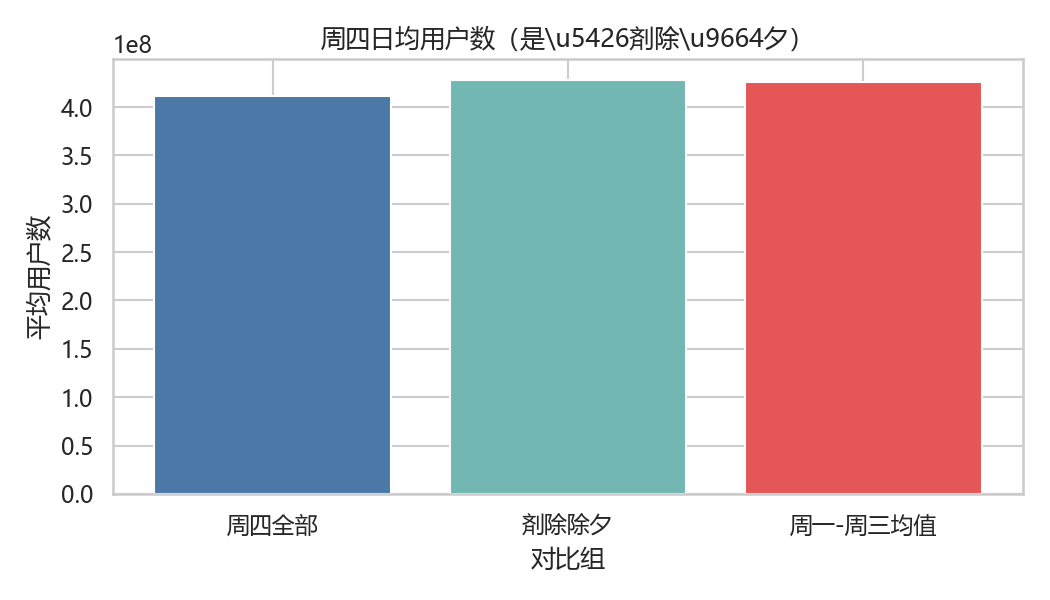
  
含义：该图展示周一-周三、周四、周五的每小时流量曲线对比。  
意义：周四在10:00-17:00与晚高峰流量强度低于周一-周三的基线水平。  
用法：用于定位周四日间低流量时段，指导税费奖励或细分策略。

  
含义：该图展示周一-周三、周四、周五的每小时用户数曲线对比。  
意义：周四日间用户较弱，使得高流量活跃期更不稳定。  
用法：用于识别周四用户活跃偏弱时段并发现与流量异常的耦合区间。

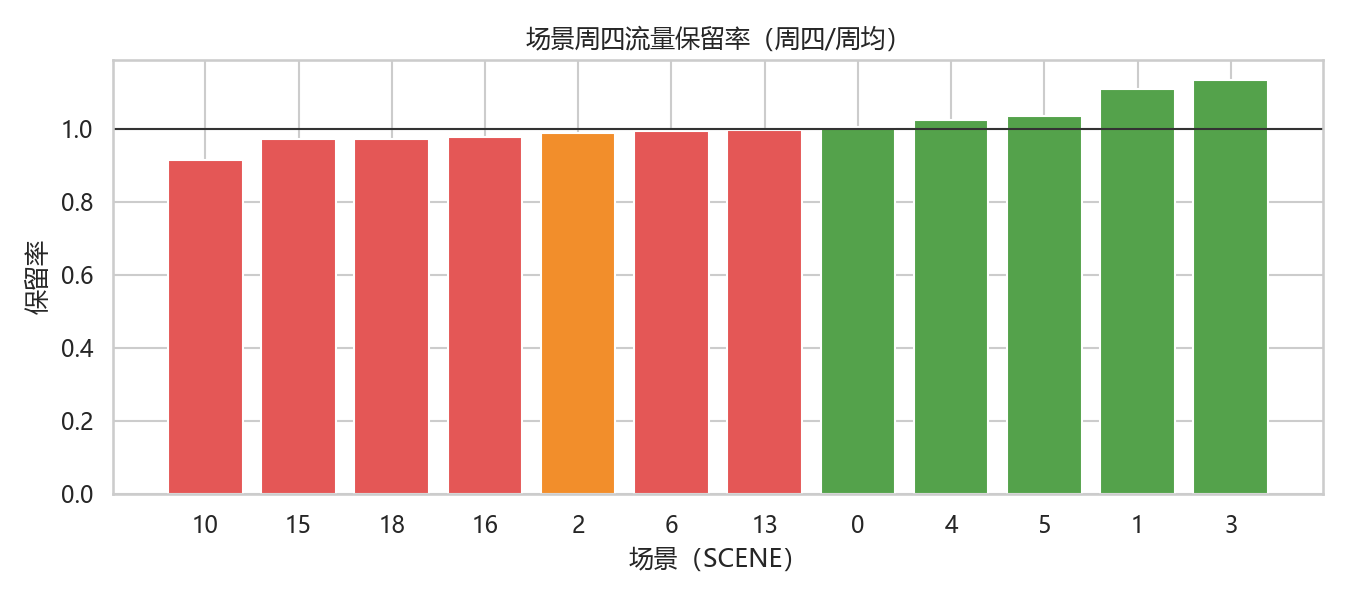
  
含义：该图展示各工作日晚高峰结果的持续时长。  
意义：周四晚高峰持续时间较短，与“高峰强度偏弱”的描述一致。  
用法：用于高峰资源调度，适度压缩周四晚高峰保障窗口。

在“剔除除夕效应”检验中，周四日均流量从 1.504e9 升至 1.556e9（+3.5%），高于周一-周三均值 1.488e9；日均用户从 4.118e8 升至 4.279e8（+3.9%），略高于周一-周三均值 4.260e8，说明“周四偏低”主要由除夕效应造成，是统计口径陷阱。

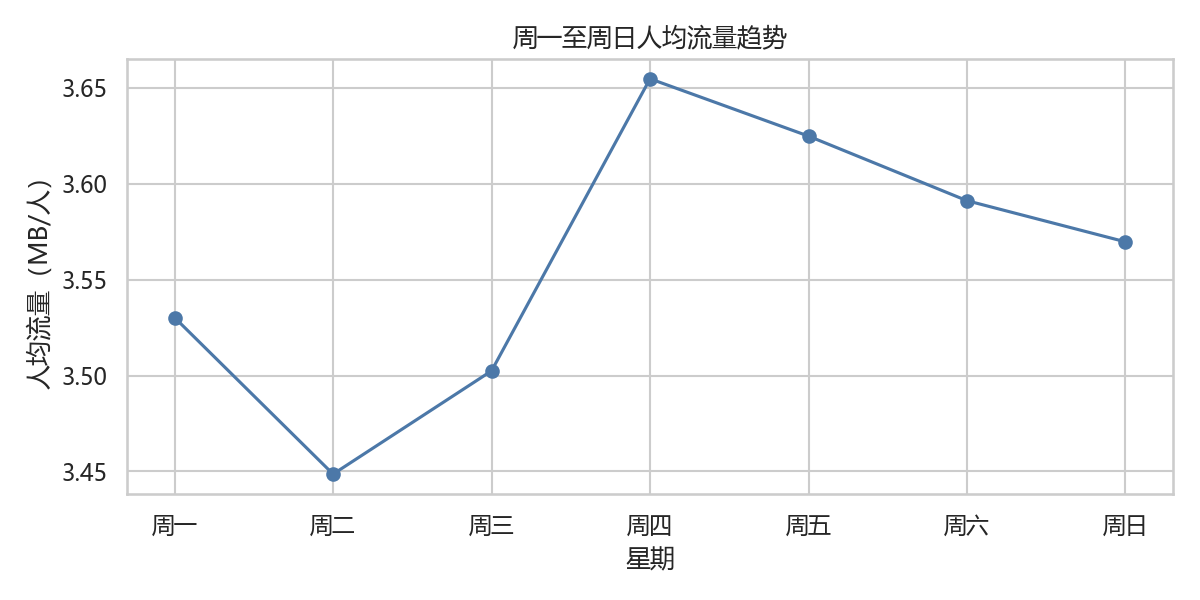
  
含义：该图比较周四全量、剔除除夕后的日均流量，并与周一-周三均值对照。  
意义：剔除除夕后周四流量回升并超过周一-周三均值，支持“除夕效应”假设。  
用法：用于校验周四偏弱是否属于节假日统计陷阱。

  
含义：该图比较周四全量、剔除除夕后的日均用户数，并与周一-周三均值对照。  
意义：剔除除夕后周四用户数回到基线附近，弱化结构性偏低判断。  
用法：用于评估是否需要针对周四制定长期运营策略。

场景特征上，计算各场景“周四流量保留率=周四流量/场景周均流量”，可见周四在不同场景下降幅度不均。若办公场景（如 SCENE 2）保留率偏低，说明存在“工作日尾声效应”；若商业场景保留率偏低，则更可能是周末消费前的蓄力期。

  
含义：该图展示各场景周四流量相对周均流量的保留率。  
意义：识别周四偏低是否集中在特定场景，揭示潜在业务结构因素。  
用法：用于制定“场景级”运营补强策略。

用户行为层面，人均流量曲线显示“人少了还是人均用得少”的差异：若周四人均流量不降而总量降，说明接入量减少；若人均流量也下降，则用户活跃强度下滑。

  
含义：该图展示周一至周日的人均流量变化趋势。  
意义：用于区分“用户数量减少”与“单用户强度下降”的贡献。  
用法：为周四拉活与容量保障策略提供依据。

### 7.3 未来运营商大数据分析展望

* 从“离线报表”走向“实时洞察”：结合流式计算，建立分钟级网络运行指标监控。
* 从“静态规则”走向“智能预测”：引入时空深度模型与节假日/天气等外部特征，提高预测与告警能力。
* 从“单点优化”走向“全网协同”：基于场景画像与聚类结果，实现基站能耗、扩容与营销的联动优化。

### 7.4 代码与产出

* 代码文件：section7\_conclusion.py
* ?????report\_assets/section7/fig01\_diurnal\_scene\_2\_6.png ? report\_assets/section7/fig16\_weekday\_flow\_per\_user.png