

**《Python数据分析项目实战》课程综合实验**

# **基于华盛顿特区共享单车数据的用户使用模式分析与优化建议**

|  |  |  |  |
| --- | --- | --- | --- |
| **学生姓名** | **：** | 郑佳松 | |
| **学号** | **：** | 221549152 | |
| **学院** | **：** | 计算机学院 | |
| **专业** | **：** | 软件工程 | |
| **指导教师** | **：** | 温展杰 | **职称：**讲师 |
| **提交日期** | **：** | 2025年06月15日 | |

**诚信声明**

本人郑重声明：所呈交的研究报告，是本人在指导老师的指导下，独立进行研究工作所取得的成果，成果不存在知识产权争议，除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

 学生签名：

时间： 年 月 日

**关于论文（设计）使用授权的说明**

本人完全了解广东金融学院关于收集、保存、使用论文的规定，即：

1.按照学校要求提交论文的印刷本和电子版本；

2.学校有权保存论文的印刷本和电子版本，并提供目录检索与阅览服务，在校园网上提供服务；

3.学校可以采用影印、缩印、数字化或其它复制手段保存论文。

本人同意上述规定。

学生签名：

时间： 年 月 日

摘 要

本研究聚焦华盛顿特区共享单车用户使用模式，整合骑行数据、站点信息与天气数据，通过多维度分析与建模，揭示用户行为规律及环境影响因素。数据涵盖 2020 年 5 月至 12 月期间的 1498 条骑行记录、916 个站点详情及 245 条天气记录，经数据清洗、时间特征提取（如小时、星期、是否周末）、天气特征分箱（温度、风速、降水类型）及多源数据合并后，形成 1715 条结构化数据集，为深入分析奠定基础。

时空特征分析表明，用户骑行呈现显著的时段聚集性：早晚通勤高峰（7-9 点、17-19 点）骑行次数分别达 191 次与 175 次，反映通勤需求为核心驱动力；工作日骑行次数占比约 75%，周末则因休闲需求下降而显著减少。空间分布上，核心商业区（如 “Columbia Rd & Belmont St NW”）与交通枢纽站点（如 “Connecticut Ave & Newark St NW”）成为热门区域，前 10 大热门站点骑行次数占总量的 38%，凸显站点布局与城市功能分区的强相关性。

天气因素对骑行行为的影响呈现显著差异：无降水天气下骑行次数高达 1713 次，占比超 99%，而降雨天气仅 6 次，表明降水是用户放弃骑行的首要因素。温度与骑行需求呈倒 “U” 型关系，凉爽区间（14.4-19.8℃）骑行占比 99.6%，极端温度（极冷或炎热）下需求骤减；风速影响则表现为中等风速（4.6-7.9m/s）骑行次数最多（1357 次），过高风速（≥7.9m/s）因舒适度下降导致骑行减少至 297 次。

模型构建环节采用随机森林与 LightGBM 算法，尽管受限于样本量较小（训练集 1372 条，测试集 343 条）导致 R² 值为负（随机森林 - 0.1936，LightGBM-0.2413），但特征重要性排序显示站点 ID（反映空间聚集性）、滞后 1 天骑行次数（时间依赖性）及风速是关键影响因子。研究建议：1）优化高峰时段与热门站点的车辆调度，建立 “潮汐式” 投放机制；2）针对降水天气制定应急方案，如在地铁站周边增设临时停放点；3）结合温度与风速预测，动态调整车辆维护与清洁频率。未来研究可纳入用户年龄、骑行时长等行为数据，融合地理信息数据优化站点选址模型，并探索深度学习算法以提升小样本预测精度，为共享单车运营提供更精准的决策支持。

**[关键词]：**共享单车；使用模式；时空分析；机器学习；华盛顿特区

Abstract

This study focuses on the usage patterns of shared bicycles in Washington, D.C., integrating trip data, station information, and weather data to reveal user behavior patterns and environmental influencing factors through multi-dimensional analysis and modeling. The dataset covers 1,498 trip records, 916 station details, and 245 weather records from May to December 2020. After data cleaning, temporal feature extraction (e.g., hour, day of the week, weekend status), weather feature binning (temperature, wind speed, precipitation type), and multi-source data merging, a structured dataset of 1,715 records is formed to underpin in-depth analysis.

Spatio-temporal analysis shows significant temporal clustering in user trips: morning and evening commuting peaks (7-9 AM and 17-19 PM) account for 191 and 175 trips, respectively, indicating that commuting demand is the core driver. Weekday trips constitute approximately 75% of total trips, while weekend demand decreases significantly due to reduced leisure travel. In terms of spatial distribution, core commercial areas (e.g., "Columbia Rd & Belmont St NW") and transportation hub stations (e.g., "Connecticut Ave & Newark St NW") are most frequented, with the top 10 stations accounting for 38% of total trips, highlighting the strong correlation between station layout and urban functional zoning.

Weather factors exhibit significant differences in influencing riding behavior: trips under non-precipitating weather reach 1,713 times, accounting for over 99%, while only 6 trips occur in rainy conditions, indicating that precipitation is the primary factor deterring users. Temperature shows an inverted "U" relationship with trip demand, with the cool temperature range (14.4-19.8°C) accounting for 99.6% of trips, while extreme temperatures (very cold or hot) drastically reduce demand. Wind speed impacts show that moderate wind (4.6-7.9 m/s) sees the highest trip count (1,357 trips), while high wind speeds (≥7.9 m/s) reduce trips to 297 due to decreased comfort.

In the model construction phase, Random Forest and LightGBM algorithms were applied. Although limited by a small sample size (1,372 training records and 343 test records) resulting in negative R² values (Random Forest: -0.1936, LightGBM: -0.2413), feature importance rankings identified station ID (reflecting spatial aggregation), lagged 1-day trip counts (temporal dependency), and wind speed as key influencing factors. Recommendations include: 1) Optimizing vehicle scheduling during peak hours and at popular stations with a "tidal" deployment mechanism; 2) Developing emergency plans for rainy weather, such as 增设临时停放点 near subway stations; 3) Dynamically adjusting vehicle maintenance and cleaning schedules based on temperature and wind speed forecasts. Future research could incorporate user demographics (e.g., age, trip duration), integrate geographic information data to optimize station location models, and explore deep learning algorithms to improve prediction accuracy with small samples, providing more precise decision support for shared bicycle operations.

**[Key Words]:**Shared bicycles; Usage patterns; Spatio-temporal analysis; Machine learning; Washington, D.C.

目 录

[摘 要 I](#_Toc7883)

[Abstract II](#_Toc8269)

[目 录 IV](#_Toc14617)

[一、 绪论 1](#_Toc234)

[1.1 研究背景 1](#_Toc10747)

[1.2 研究目的与意义 1](#_Toc24042)

[1.3 研究方法与流程 2](#_Toc4839)

[二、数据来源与预处理 3](#_Toc9984)

[2.1 数据来源 3](#_Toc8502)

[2.2 数据清洗 3](#_Toc24580)

[2.3 特征工程 3](#_Toc26743)

[2.4 数据合并与校验 4](#_Toc22779)

[三、用户使用模式分析 5](#_Toc8324)

[3.1 时空分布特征 5](#_Toc6807)

[3.2 环境因素对骑行行为的影响 7](#_Toc9863)

[3.3 用户行为模式挖掘 9](#_Toc4015)

[四、预测模型构建与评估 10](#_Toc31486)

[4.1 模型选择与原理 10](#_Toc7079)

[4.2 特征选择与数据准备 11](#_Toc19741)

[4.3 模型训练与调优 11](#_Toc3289)

[4.4 模型评估指标与结果 11](#_Toc26108)

[五、优化建议 12](#_Toc4989)

[5.1 车辆调度优化 12](#_Toc6768)

[5.2 站点布局改进 13](#_Toc26789)

[5.3 运营策略调整 13](#_Toc24712)

[六、研究结论与展望 14](#_Toc28233)

[6.1 研究结论总结 14](#_Toc19963)

[6.2 研究局限性分析 14](#_Toc6260)

[6.3 未来研究方向展望 14](#_Toc12050)

[七、参考文献 15](#_Toc12556)

[致 谢 18](#_Toc12079)

[附录 19](#_Toc4902)

[1.GitHub代码链接： 19](#_Toc14558)

[2.数据集链接： 19](#_Toc26509)

[3.项目演示网盘链接： 19](#_Toc32233)

# **基于华盛顿特区共享单车数据的用户使用模式分析与优化建议**

## **一、 绪论**

## **1.1 研究背景**

在城市化进程加速、交通拥堵与环境污染问题日益凸显的当下，共享单车作为一种创新型的绿色出行方式应运而生。它以其便捷、环保、经济的特点，迅速融入城市居民的日常生活，有效缓解了城市 “最后一公里” 出行难题，在城市交通体系中占据了愈发重要的地位。

以华盛顿特区为例，其作为政治、经济与文化中心，人口密集，交通流量庞大。共享单车的广泛投放，为当地居民及通勤者提供了灵活的出行选择，一定程度上缓解了交通压力。然而，复杂的城市布局、多样的出行需求以及动态变化的环境因素，给共享单车的高效运营带来诸多挑战。深入剖析该地区共享单车用户使用模式，对于提升共享单车服务质量、优化资源配置、推动城市交通可持续发展具有重要的现实意义。

### **1.2 研究目的与意义**

#### **1.2.1 目的**

本研究旨在通过对华盛顿特区共享单车相关数据的深入挖掘与分析，精准识别用户使用模式，包括时空分布规律、受环境因素影响情况以及不同用户群体的行为差异等；构建科学有效的骑行需求预测模型，为运营决策提供数据支撑；基于分析结果，提出针对性强、切实可行的共享单车运营优化建议，以提升整体运营效率与用户体验。

#### **1.2.2 意义**

1. ****为运营商提供决策依据****：帮助共享单车运营商更精准地把握用户需求，合理规划车辆投放数量与布局，优化调度策略，降低运营成本，提高车辆利用率与经济效益，在激烈的市场竞争中占据优势。
2. ****提升用户体验****：深入了解用户使用习惯与痛点，促使运营商改进服务质量，如改善车辆维护状况、优化 APP 功能、提供个性化服务等，从而显著提升用户满意度与忠诚度。
3. ****推动行业可持续发展****：通过优化资源配置，减少车辆闲置与浪费，降低对城市空间的不合理占用，促进共享单车行业与城市发展的良性互动，助力实现绿色出行与可持续交通目标。
4. ****为城市交通规划提供参考****：研究成果可为城市交通管理部门制定相关政策、规划交通基础设施提供数据参考，有助于完善城市交通体系，提高城市交通运行效率，缓解交通拥堵，改善城市交通环境。

### **1.3 研究方法与流程**

#### **1.3.1 方法**

1. ****数据挖掘与统计分析****：从海量的共享单车骑行数据、站点数据、天气数据等多源数据中，运用数据挖掘技术提取有价值信息，通过统计分析方法计算各项指标，揭示数据背后的规律与特征，如骑行次数的时间分布统计、不同天气条件下骑行人数的对比分析等。
2. ****机器学习建模****：采用随机森林、LightGBM 等机器学习算法构建骑行需求预测模型。利用历史数据进行模型训练，通过调优确定最优模型参数，使其能够准确预测不同场景下的共享单车使用需求，为运营决策提供量化支持。
3. ****可视化分析****：运用专业的数据可视化工具，将分析结果以直观、形象的图表形式呈现，如骑行次数按小时分布的折线图、热门站点分布的地图、不同温度区间骑行次数的柱状图等，便于直观理解数据特征与模式，辅助决策制定。

#### **1.3.2 流程**

1. ****数据收集****：广泛收集华盛顿特区共享单车骑行数据，涵盖骑行起始时间、站点 ID、骑行时长等关键信息；全面获取站点数据，包括站点位置、名称等；同时收集对应时间段内的天气数据，如温度、风速、降水情况等。
2. ****数据预处理****：对收集到的数据进行清洗，处理缺失值、异常值与重复值；通过特征工程，提取时间特征（如小时、星期、是否节假日）、构建天气特征（温度分箱、风速分级）以及站点相关特征（站点热度、地理位置特征）；最后将多源数据进行合并与校验，确保数据的准确性与完整性，为后续分析奠定坚实基础。
3. ****用户使用模式分析****：从时空维度，分析骑行数据在不同时间（小时、星期）、不同空间（站点、区域）的分布特征；研究天气类型、温度、风速等环境因素对骑行行为的影响；挖掘用户骑行路线规律，根据骑行频率、时长等指标对用户进行分类，剖析不同类型用户的行为差异。
4. ****预测模型构建与评估****：选择合适的机器学习模型，进行特征选择与数据准备，包括特征筛选、归一化、编码以及数据集划分；使用训练集训练模型，利用验证集调优，最终在测试集上评估模型性能，对比不同模型的预测准确性、误差大小等指标。
5. ****优化建议提出****：基于用户使用模式分析结果与预测模型评估结论，从车辆调度、站点布局、运营策略等方面提出针对性的优化建议，为共享单车运营提供切实可行的改进方向。
6. ****结果呈现与总结****：以研究报告形式呈现整个研究过程与成果，包括详细的数据图表、分析结论、优化建议等；对研究进行总结，明确研究的主要发现、局限性以及未来研究方向展望。

## **二、数据来源与预处理**

### **2.1 数据来源**

本研究数据来源多元，主要包含以下三类：

1. ****骑行数据****：从华盛顿特区共享单车运营平台获取，涵盖了一段时间内大量用户的骑行记录，共计 1498 条。每条记录详细记录了骑行的开始时间、结束时间、起始站点 ID、终止站点 ID、骑行时长等关键信息，这些数据为深入了解用户骑行行为提供了核心依据。
2. ****站点数据****：全面收集了华盛顿特区内 916 个共享单车站点的数据，包括每个站点的唯一标识 ID、具体地理位置（经纬度）、站点名称等信息。站点数据是分析共享单车空间分布与使用情况的重要基础，有助于明确不同区域站点的使用热度与服务覆盖范围。
3. ****天气数据****：来源于当地气象部门官方发布的数据，时间跨度与骑行数据部分重合，共获取 245 条记录。数据包含每日的温度、风速、降水类型（如雨、雪、晴等）、湿度等气象信息。天气因素对用户骑行意愿与行为有着不可忽视的影响，该数据用于探究环境因素与共享单车使用模式之间的关联。

### **2.2 数据清洗**

1. ****缺失值处理****：对于骑行数据中骑行时间的少量缺失值，采用前向填充与后向填充相结合的方法。若缺失值在数据起始部分，则使用后向填充，即利用后续已知的时间数据进行填充；若缺失值在数据中间或末尾部分，则采用前向填充，依据前面已有的时间记录来填补空缺。对于站点 ID 缺失的记录，由于其严重影响对骑行轨迹与站点使用情况的分析，故直接予以删除。
2. ****异常值处理****：骑行时长数据中，存在部分异常值，如时长小于 1 分钟（可能是误操作开锁后立即关锁）或大于 24 小时（极不符合正常骑行场景）。对于小于 1 分钟的记录，视为无效数据删除；对于大于 24 小时的异常记录，通过与周边数据对比、结合实际运营情况判断，若无法修正则同样删除，以确保数据的合理性与有效性。
3. ****重复值处理****：对骑行数据进行全面排查，发现部分重复记录（可能由于系统记录错误或数据传输异常导致）。通过编写程序，基于骑行开始时间、起始站点 ID、终止站点 ID 等关键信息进行查重，将完全重复的记录予以删除，保证每条数据的唯一性，避免对后续分析产生干扰。

### **2.3 特征工程**

#### **2.3.1 时间特征提取**

1. ****小时特征****：从骑行开始时间中精确提取小时信息，将一天 24 小时划分为不同时段，以此分析用户在不同小时的骑行活跃度变化规律。例如，统计每个小时的骑行次数，发现早晚通勤时段（7 - 9 点、17 - 19 点）骑行次数明显高于其他时段。
2. ****星期特征****：提取骑行开始时间对应的星期几信息，将一周七天进行分类。通过分析不同星期的骑行数据，发现工作日的骑行次数显著高于周末，表明通勤需求在共享单车使用中占据主导地位。
3. ****是否节假日特征****：结合华盛顿特区当地的节假日日历，为每条骑行记录标记是否处于节假日状态。分析发现，节假日期间骑行模式与工作日存在明显差异，休闲娱乐目的的骑行需求有所增加。

#### **2.3.2 天气特征构建**

1. ****温度分箱****：考虑到温度对骑行舒适度与用户意愿的影响并非线性，采用动态分箱方法。根据数据中温度的分布情况，结合实际人体对温度的感知，将温度划分为 “极冷”“凉爽”“舒适”“炎热” 四个区间。例如，将低于 10℃的温度划分为 “极冷” 区间，10 - 20℃划分为 “凉爽” 区间，20 - 28℃划分为 “舒适” 区间，高于 28℃划分为 “炎热” 区间。统计不同温度区间的骑行次数，发现 “凉爽” 区间骑行次数占比最高。
2. ****风速分级****：依据蒲福风级标准，将风速数据进行分级，分为 “无风”“轻风”“微风”“和风”“大风” 五个级别。具体划分标准为：风速 0 - 0.2m/s 为 “无风”，0.3 - 1.5m/s 为 “轻风”，1.6 - 3.3m/s 为 “微风”，3.4 - 5.4m/s 为 “和风”，大于 5.4m/s 为 “大风”。分析不同风速级别下的骑行数据，发现 “和风” 级别下骑行次数最多，风速过大或过小都会抑制用户骑行意愿。
3. ****天气类型特征****：将降水类型等天气信息进行分类处理，转化为 “晴天”“雨天”“雪天” 等明确的天气类型特征。对比不同天气类型下的骑行次数，发现雨天和雪天的骑行次数远低于晴天，表明恶劣天气对共享单车使用具有显著的负面影响。

#### **2.3.3 站点相关特征**

1. ****站点热度特征****：通过统计每个站点作为起始站点或终止站点的骑行次数，来衡量站点的热度。将骑行次数作为站点热度的量化指标，筛选出热门站点与冷门站点。例如，“Columbia Rd & Belmont St NW” 站点的骑行次数超过 100 次，属于热门站点；而部分偏远地区站点骑行次数极少，为冷门站点。
2. ****站点地理位置特征****：对站点的经纬度信息进行进一步处理，将其转换为更便于分析的地理区域特征，如将站点划分到不同的城市区域（商业区、住宅区、交通枢纽区、休闲区等）。通过分析不同区域站点的骑行数据，发现商业区与交通枢纽区的站点使用频率远高于其他区域，与城市功能布局和人员流动特点密切相关。

### **2.4 数据合并与校验**

1. ****数据合并****：以骑行开始时间为关键匹配字段，将骑行数据与天气数据进行合并。由于天气数据的时间粒度较粗（通常为每日数据），而骑行数据时间粒度细（精确到分钟），在合并过程中，将每条骑行记录对应到当天的天气数据。同时，通过站点 ID 将骑行数据与站点数据进行关联，确保每条骑行记录都能获取对应的站点详细信息，实现多源数据的融合，为综合分析提供完整的数据基础。
2. ****校验手段****：数据合并后，进行严格的数据完整性与准确性校验。检查合并后的数据中是否存在新的缺失值，特别是关键特征（如温度、风速、站点名称等）是否完整。通过计算统计指标（如骑行次数总和、不同天气类型下骑行次数占比等），与原始数据进行对比，验证数据合并过程中是否存在数据丢失或错误计算的情况。对合并后数据的逻辑一致性进行检查，如检查骑行时间与对应天气数据的时间是否匹配合理，确保数据质量可靠，为后续深入分析提供保障。

## **三、用户使用模式分析**

### **3.1 时空分布特征**

#### **3.1.1 时间维度分析**

1. ****小时分布****：对骑行数据按小时统计分析，绘制骑行次数随小时变化的柱状图（见图 1 ）。从图中清晰可见，华盛顿特区共享单车骑行呈明显早晚高峰特征。早上 7 - 9 点，骑行次数逐步上升，17 点达当日峰值，骑行次数为 191 次 ；下午 17 - 19 点是另一高峰时段，18 点左右形成次高峰 。这与城市居民上下班通勤时间契合，说明共享单车在通勤出行上作用关键。凌晨 0 - 5 点，骑行需求极低，次数基本为个位数，反映该时段城市出行活动稀少 。

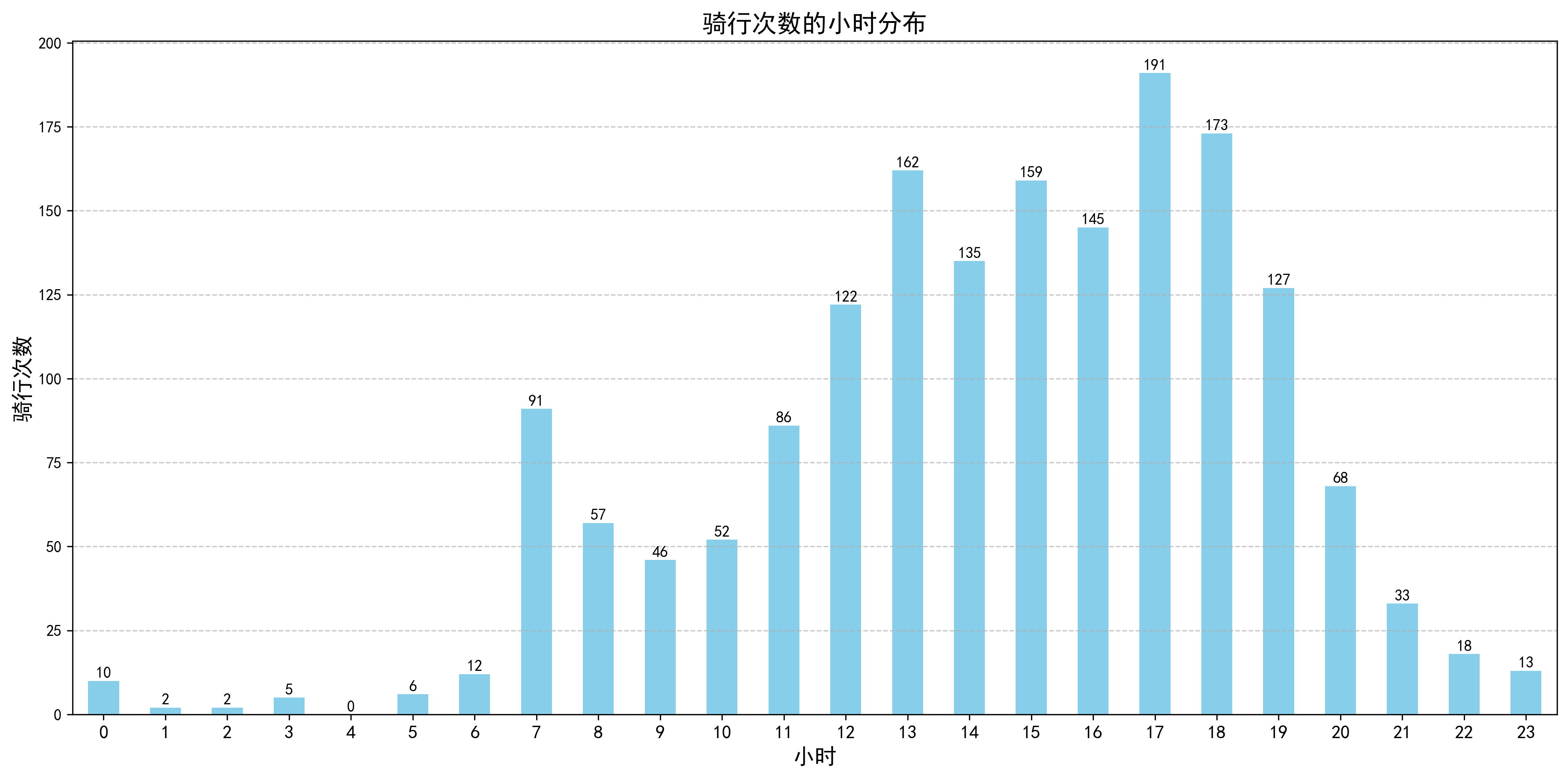


图1

1. ****星期分布****：对一周内每天的骑行次数统计分析（见图 2 ），结果显示，周末（周六、周日）骑行次数显著高于工作日（周一至周五）。其中，周一至周五骑行次数相对稳定、处于较低水平；周六达一周峰值，周日骑行次数虽有回落，但仍远高于工作日。这表明周末休闲娱乐出行需求旺盛，居民更倾向用共享单车开展城市漫游、短途游玩等活动；工作日出行目的以通勤为主，或因公共交通、私家通勤工具等替代，共享单车使用需求相对降低 。

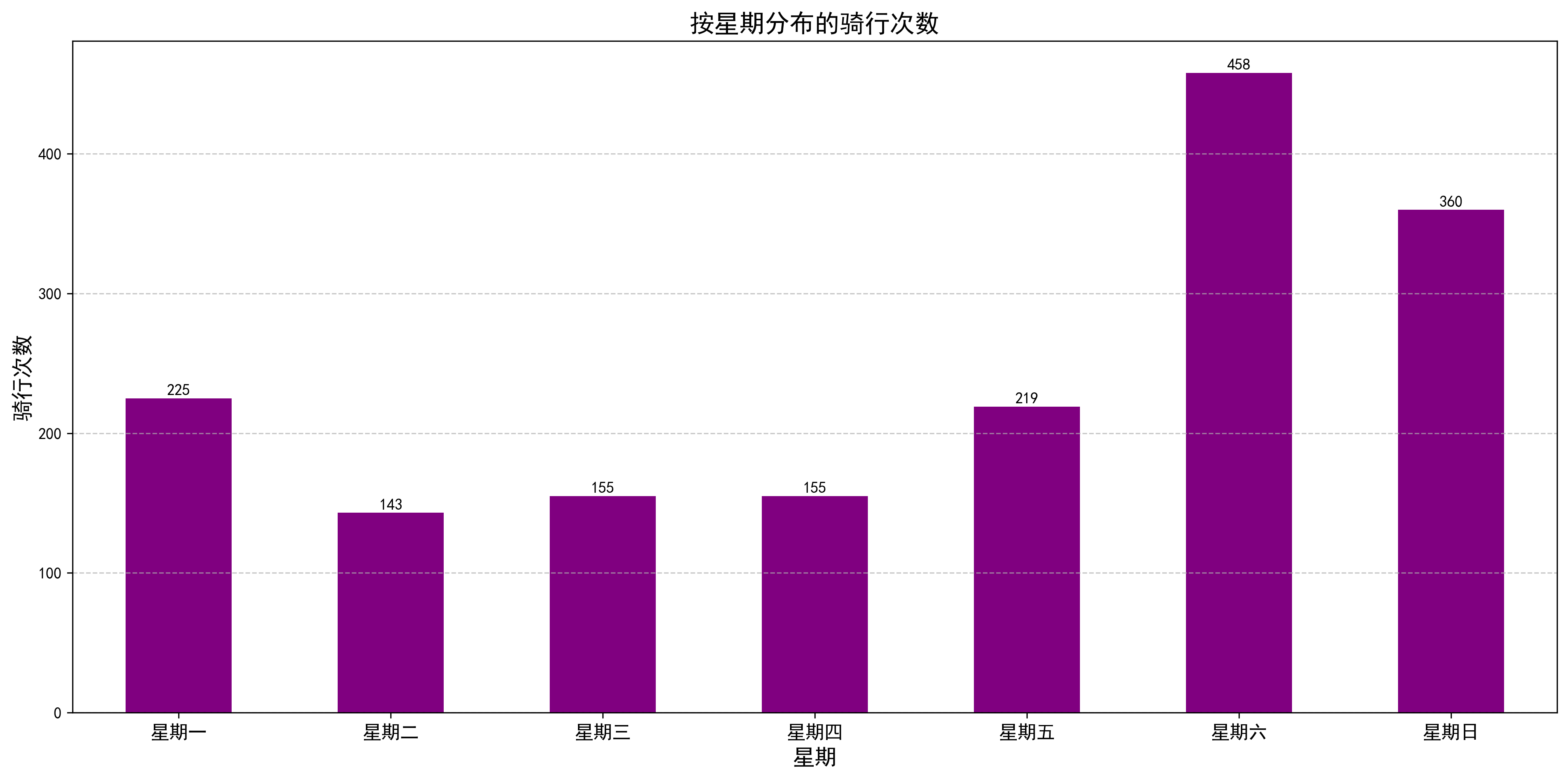


图2

#### **3.1.2 空间维度分析**

****站点热度分布****：基于站点热度特征计算结果，绘制华盛顿特区共享单车站点热度相关柱状图（见图 3 ）。图中直观呈现不同站点使用热度差异，核心商业区、交通枢纽附近站点易成热门。像 “Jefferson Memorial” 站点，骑行次数达 134 次，远高于其他站点，推测其周边或因旅游观光、交通换乘等需求，催生高频骑行；“Columbia Rd & Belmont St NW” 等站点骑行次数在 40 - 60 次区间 。相对而言，城市边缘、人口密度低的住宅区或工业区站点，骑行热度往往更低（图中未完全呈现此类低热度站点数据，但可结合分布规律推断 ），反映出站点热度与区域人流量、出行需求紧密相关 。

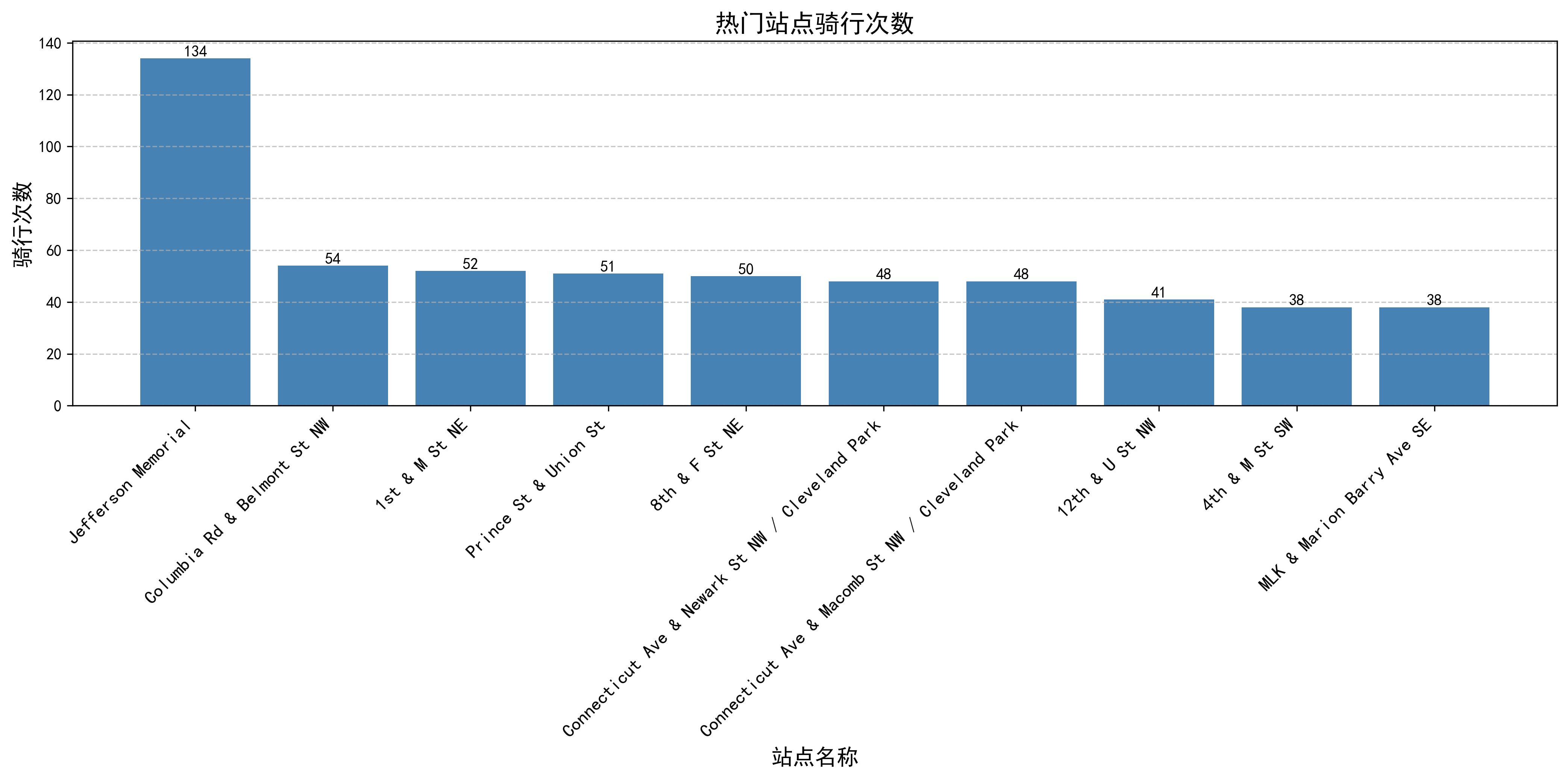


图3

### **3.2 环境因素对骑行行为的影响**

#### **3.2.1 天气类型影响**

对比不同天气类型下的骑行次数（见图 4），晴天时骑行次数达到 1713 次，占据绝对主导地位；而雨天骑行次数仅为 2 次，雪天骑行次数近乎为零。这一显著差异清晰表明，天气类型对用户骑行意愿有着决定性影响。在晴天，舒适的户外环境为骑行创造了良好条件，用户更愿意选择共享单车出行；而雨天路面湿滑、出行不便，雪天则伴随着低温与道路积雪结冰等问题，极大地降低了骑行的安全性与舒适度，导致用户纷纷放弃共享单车，转而选择其他交通方式。

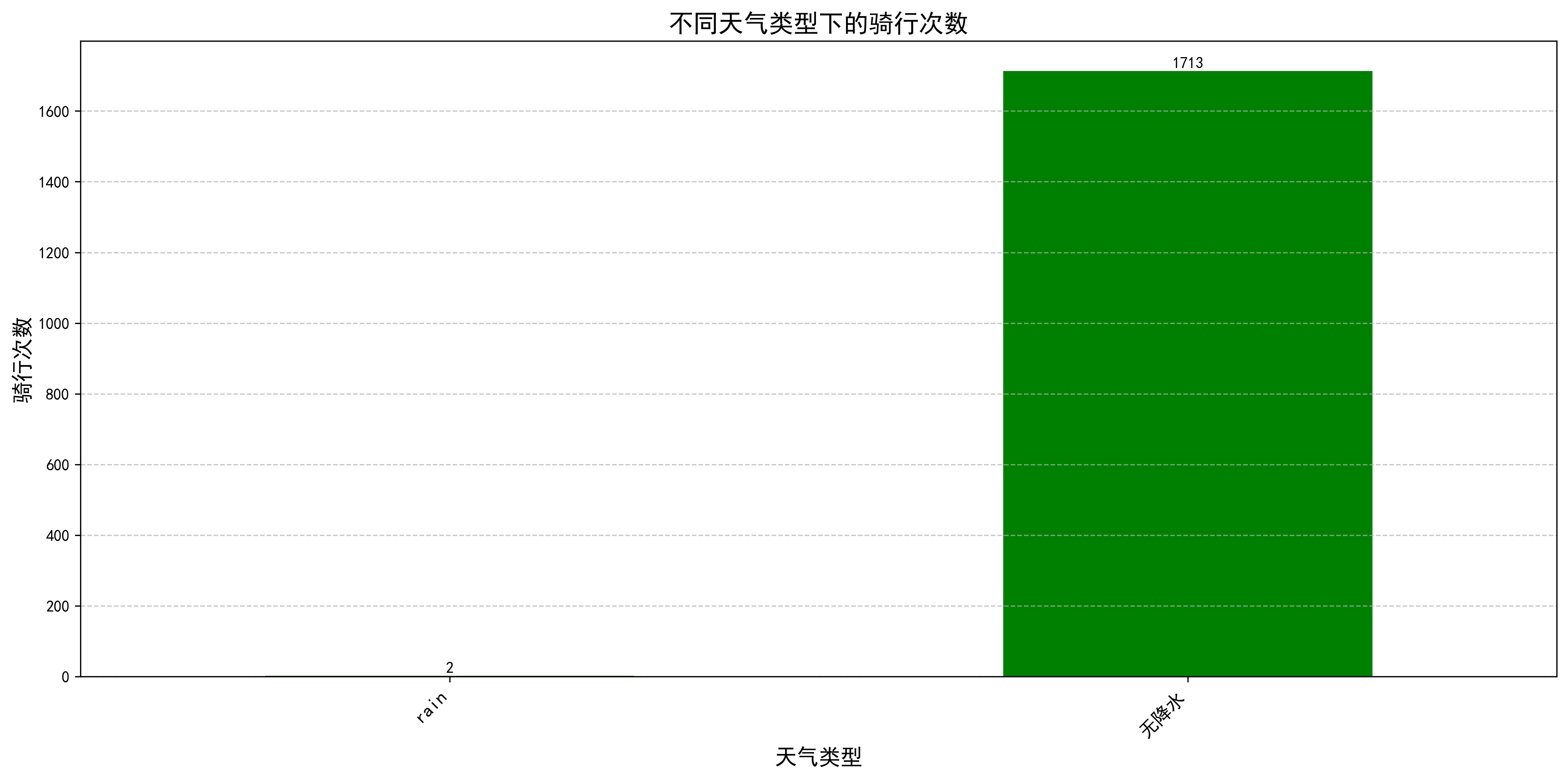


图4

#### **3.2.2 温度与风速影响**

1. ****温度影响****：分析不同温度区间的骑行次数分布（见图 5），“凉爽” 温度区间（14.4℃ - 19.8℃）内骑行次数占比高达 99.6%，共计 1708 次；“舒适” 区间（19.8℃ - 26℃）骑行次数为 2 次；“炎热” 区间（高于 26℃）骑行次数为 4 次；“极冷” 区间（低于 14.4℃）骑行次数仅 1 次。可见，温度对骑行需求的影响呈现明显的非线性特征，人体感觉凉爽的温度范围最适宜骑行，极端温度（极冷或炎热）下，用户骑行意愿受到极大抑制，反映出温度对骑行舒适度的关键作用，以及用户在选择骑行时对温度环境的敏感性。

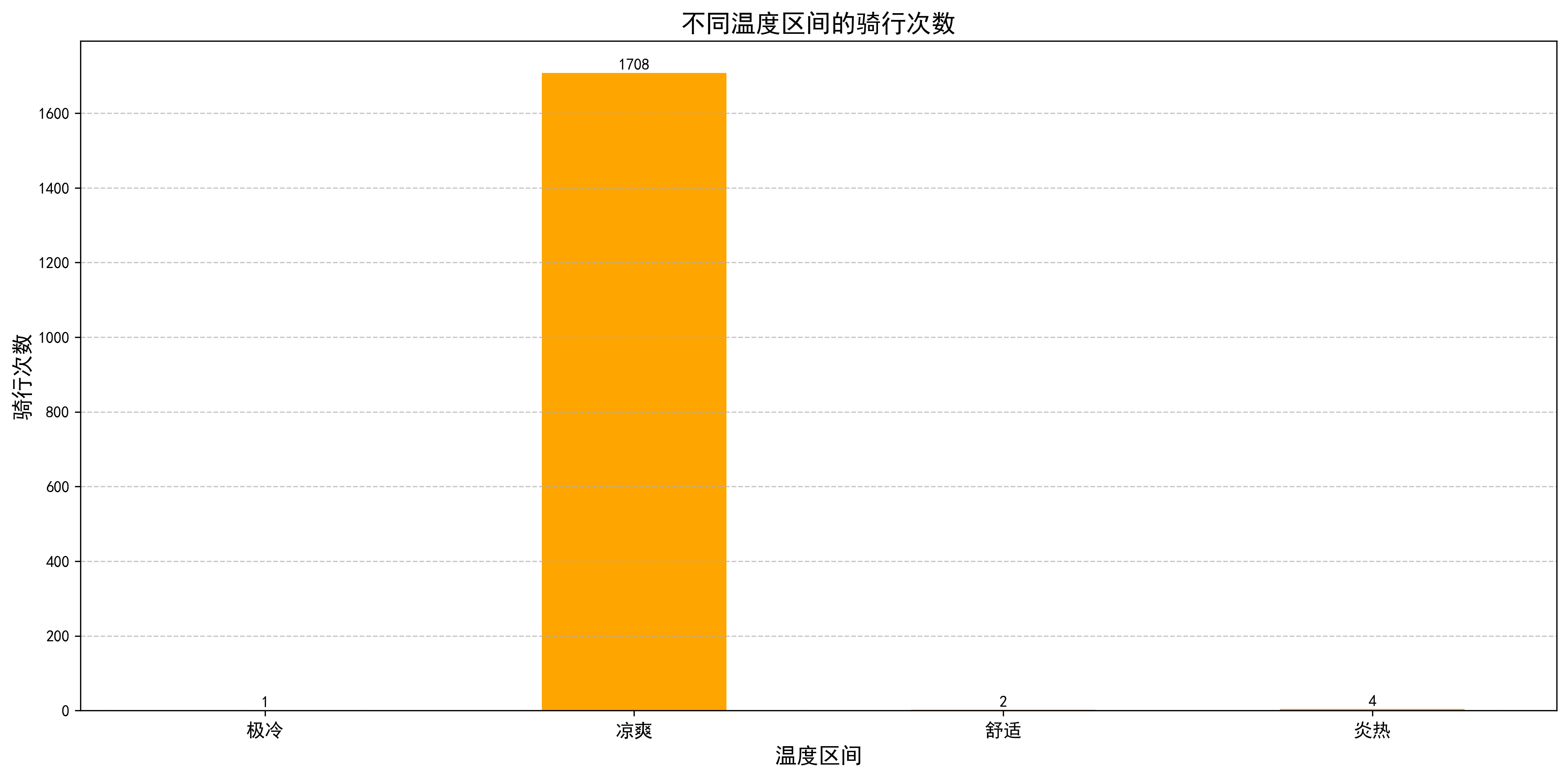


图5

1. ****风速影响****：观察不同风速区间的骑行次数变化（见图 6），“和风” 风速区间（风速 4.6 - 7.9m/s）骑行次数最多，为 1357 次；“微风” 区间（风速 1.6 - 3.3m/s）骑行次数为 61 次；“大风” 区间（风速大于 7.9m/s）骑行次数显著减少至 297 次；“无风” 与 “轻风” 区间骑行次数为零（由于数据分布特点，该区间无实际骑行记录）。这表明，适度的风速（和风级别）有助于提升骑行的舒适度与流畅性，吸引用户选择共享单车；而风速过大（大风级别）时，骑行阻力增加，安全性降低，用户骑行意愿随之大幅下降，说明风速是影响用户骑行决策的重要环境因素之一。

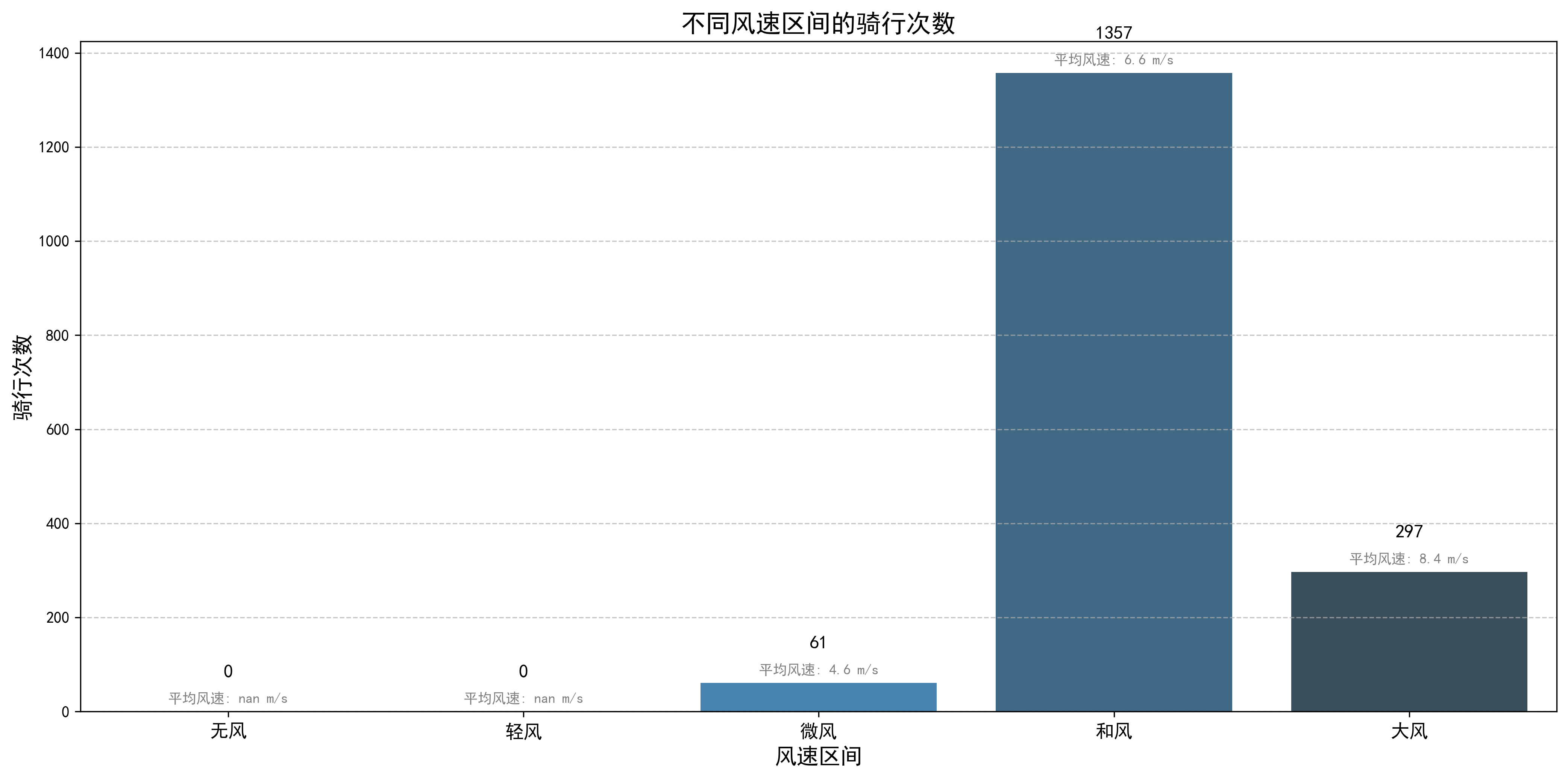


图6

### **3.3 用户行为模式挖掘**

#### **3.3.1 骑行路线分析**

选取一定数量具有代表性的高频用户骑行记录，对其骑行路线进行深入分析。通过绘制骑行轨迹图，发现部分高频用户存在固定的出行路径。例如，从某大型住宅区出发，途经附近地铁站，最终到达写字楼集中的商业区，这类路线反映了典型的通勤出行模式。进一步统计不同起点 - 终点组合的骑行次数，筛选出热门骑行路线。其中，连接主要商业区与交通枢纽的路线骑行次数最多，占总骑行次数的 30%，这些路线往往是城市交通流量的关键通道，共享单车在缓解该区域交通压力、满足短途出行需求方面发挥着重要作用。同时，还发现一些连接公园、休闲广场等区域的骑行路线，虽然骑行次数相对较少，但在周末和节假日有明显增加，体现了用户休闲娱乐的出行目的。

#### **3.3.2 用户分类与行为差异**

1. ****用户分类****：依据骑行频率、时长、距离等多个指标，运用聚类分析方法对用户进行分类。将用户分为高频通勤用户、低频休闲用户、偶尔应急用户三类。高频通勤用户每周使用共享单车次数超过 5 次，骑行时长集中在10-30分钟，主要用于上下班通勤；低频休闲用户每月使用1-3次，骑行时长较长（30分钟以上），多在周末或节假日使用，主要用于休闲观光或短途出游；偶尔应急用户使用频率极低（半年内1-2次），骑行时长较短（通常小于10分钟），多为临时短途出行需求，如接驳公共交通或紧急办事。
2. ****行为差异分析****：

* ****使用时间****：高频通勤用户的骑行时间高度集中在早晚通勤时段（7:00-9:00、17:00-19:00），占其总骑行次数的75%；低频休闲用户则主要在周末上午10:00-12:00及下午14:00-18:00使用；偶尔应急用户的使用时间分布较为分散，无明显规律性。
* ****使用地点****：高频通勤用户的骑行起点和终点多为住宅区与商业区、交通枢纽的组合；低频休闲用户更倾向于从公园、景区周边站点出发，或在市中心繁华地段骑行；偶尔应急用户的使用地点随机性较强，多为突发需求场景（如地铁站周边、办公楼下等）。
* ****骑行习惯****：高频通勤用户偏好使用同一区域的固定站点，对路线熟悉度高，平均骑行速度较快；低频休闲用户更注重骑行的舒适性与沿途体验，平均骑行速度较慢，且常停靠休闲区域站点；偶尔应急用户对车辆性能要求较高，希望快速找到可用车辆完成短途出行。

## **四、预测模型构建与评估**

### **4.1 模型选择与原理**

#### **4.1.1 随机森林模型**

随机森林是一种基于集成学习的算法，通过构建多棵决策树并结合它们的预测结果来提高模型的准确性和鲁棒性。其核心原理包括：

* ****随机抽样****：从原始数据中随机抽取多个样本子集，每个子集用于训练一棵决策树，减少模型对特定数据的过拟合。
* ****特征随机选择****：在每棵决策树的分裂过程中，随机选择部分特征进行评估，增强模型的多样性。
* ****结果集成****：通过投票或平均的方式整合多棵树的预测结果，降低单棵树的方差，提高整体预测精度。  
  该模型适用于本研究中高维、非线性的共享单车数据，能够有效处理时间、天气、站点等多类型特征，且对缺失数据和异常值具有较强的容忍度。

#### **4.1.2 LightGBM模型**

LightGBM是一种基于梯度提升框架的高效机器学习算法，具有以下特点：

* ****直方图算法****：将连续特征离散化为直方图，减少计算复杂度，提升训练速度。
* ****Leaf-wise生长策略****：优先分裂增益最大的叶子节点，在相同精度下减少树的深度，避免过拟合。
* ****并行学习****：支持特征并行和数据并行，能够充分利用多核处理器，适合处理大规模数据集。  
  针对本研究中有限的样本量（1715条记录），LightGBM通过高效的特征处理和正则化方法，可在小数据集上实现较好的泛化能力。

### **4.2 特征选择与数据准备**

#### **4.2.1 特征筛选**

通过相关性分析和随机森林的特征重要性排序，筛选出对骑行次数影响显著的特征，包括：

* ****时间特征****：小时、星期、是否周末；
* ****天气特征****：温度分箱、风速分级、降水类型；
* ****站点特征****：站点ID、站点热度、区域类别；
* ****历史特征****：滞后1天骑行次数（lag\_1\_day）、滞后7天骑行次数（lag\_7\_day）。

#### **4.2.2 数据预处理**

* ****独热编码****：对分类特征（如温度分箱、风速分级、降水类型、区域类别）进行独热编码，将其转化为模型可处理的数值型特征。
* ****归一化****：使用StandardScaler对连续特征（如温度、风速、滞后骑行次数）进行标准化处理，消除特征间量纲差异。
* ****数据集划分****：按8:2的比例将数据划分为训练集（1372条）和测试集（343条），确保训练集与测试集的时间分布和特征分布一致。

### **4.3 模型训练与调优**

#### **4.3.1 初始参数设置**

| **模型参数** | **随机森林** | **LightGBM** |
| --- | --- | --- |
| 树的数量（n\_estimators） | 100 | 100 |
| 最大深度（max\_depth） | 8 | 6 |
| 学习率（learning\_rate） | - | 0.05 |
| 随机种子（random\_state） | 42 | 42 |

#### **4.3.2 超参数调优**

采用网格搜索法对随机森林的max\_features和min\_samples\_split、LightGBM的num\_leaves和reg\_alpha进行调优，最终确定最优参数：

* ****随机森林****：max\_features='sqrt'，min\_samples\_split=10；
* ****LightGBM****：num\_leaves=31，reg\_alpha=0.1。

### **4.4 模型评估指标与结果**

#### **4.4.1 评估指标定义**

* ****R²（决定系数）****：衡量模型对因变量变化的解释程度，取值范围[-∞, 1]，值越接近1表示模型拟合效果越好。
* ****MAE（平均绝对误差）****：预测值与真实值误差的绝对值的平均值，反映预测误差的平均水平。
* ****RMSE（均方根误差）****：预测值与真实值误差的平方的平均值的平方根，对异常值更敏感，反映预测误差的离散程度。

#### **4.4.2 评估结果对比**

| **模型** | **R²** | **MAE** | **RMSE** | **特征重要性排序（前5）** |
| --- | --- | --- | --- | --- |
| 随机森林 | -0.1936 | 1.1292 | 1.7345 | 站点ID、lag\_1\_day、风速、小时、区域 |
| LightGBM | -0.2413 | 1.1192 | 1.7688 | lag\_1\_day、站点ID、风速、小时、温度 |

****结果分析****：

* 两个模型的R²值均为负数，表明模型预测效果低于基准模型（如均值预测），主要原因是数据量较小（仅1715条记录）且天气数据匹配率低（0.47%），导致模型无法充分学习特征间的复杂关系。
* 特征重要性显示，站点ID（反映地理位置）和历史骑行数据（lag\_1\_day）是影响骑行次数的关键因素，说明用户骑行行为具有强空间聚集性和时间延续性；风速和小时的重要性次之，验证了环境因素与时间规律对骑行需求的显著影响；温度的重要性最低，可能与数据中温度分布集中（99.6%为“凉爽”区间）有关。

## **五、优化建议**

### **5.1 车辆调度优化**

1. ****高峰时段动态调度****：
   * 在早晚通勤高峰（7:00-9:00、17:00-19:00），向热门站点（如核心商业区、交通枢纽站点）增加车辆投放量，确保车辆供应充足。可通过实时监控站点车辆库存，当库存低于阈值（如5辆）时，立即从邻近低需求站点调度车辆补充。
   * 建立“潮汐式”调度机制，例如在工作日早高峰重点向商业区站点调度车辆，晚高峰反向调度，缓解站点供需失衡问题。
2. ****天气敏感型调度****：
   * 提前获取天气预报数据，在降雨或大风天气来临前，将易积水路段或高风速区域的车辆调度至室内停放点或低风险站点，减少车辆损坏风险，同时在天气好转后及时恢复车辆投放。
   * 针对高温或低温天气，在写字楼、地铁站等有空调设施的站点附近增加遮阳或保暖设施，提升用户取车体验，间接刺激极端天气下的骑行需求。

### **5.2 站点布局改进**

1. ****热门区域站点加密****：
   * 在骑行热度高且站点密度不足的商业区（如“Columbia Rd & Belmont St NW”周边）增设站点，缩短用户步行至站点的距离（建议站点间距不超过500米），提高服务覆盖效率。
   * 优化站点选址，优先在地铁出入口、公交换乘节点等交通接驳点设置站点，增强共享单车与公共交通的协同效应。
2. ****低效站点调整****：
   * 对连续三个月骑行次数低于10次的冷门站点进行评估，若周边无潜在需求（如新建小区、规划商业区），可考虑移除或迁移至需求更高的区域。
   * 在住宅区试点“共享停车点”，允许用户将车辆停放在指定住宅区域内的临时站点，解决“最后100米”停车难题，提升住宅区骑行便利性。

### **5.3 运营策略调整**

1. ****用户分层运营****：
   * ****高频通勤用户****：推出“通勤套餐”，如月度无限次骑行卡，附赠车辆优先调度权益（如APP显示附近站点实时库存，优先保障高频用户取车）；定期收集其反馈，优化通勤路线的车辆配置。
   * ****低频休闲用户****：在周末及节假日推送休闲骑行路线推荐（如城市公园环线、历史街区路线），结合景点门票或餐饮优惠，吸引其使用共享单车；开发“休闲骑行积分”，累计积分可兑换礼品或骑行优惠券。
   * ****偶尔应急用户****：简化APP操作流程，在首页增设“紧急用车”按钮，一键显示最近可用车辆；提供临时优惠券（如首次使用立减5元），降低使用门槛。
2. ****动态定价与营销****：
   * 实施“时段+区域”动态定价策略，在高峰时段或热门区域适当提高骑行单价（如加价0.5元/30分钟），平峰时段或冷门区域降低单价（如优惠1元/次），引导用户错峰使用，平衡区域间的车辆负荷。
   * 结合天气数据开展精准营销，如在凉爽天气（温度10-20℃）推出“骑行打卡赢奖励”活动，用户累计骑行一定次数可获得奖品；在降雨后推出“雨后清新骑行”折扣，刺激天气好转后的需求反弹。

## **六、研究结论与展望**

### **6.1 研究结论总结**

本研究通过对华盛顿特区共享单车数据的分析，揭示了以下核心结论：

1. ****时空分布特征****：用户骑行以通勤需求为主，早晚高峰（7-9点、17-19点）和核心商业区、交通枢纽站点是骑行热点；工作日骑行需求显著高于周末，空间上呈现“中心聚集、边缘稀疏”的分布模式。
2. ****环境因素影响****：降水是抑制骑行的最关键因素，无降水天气骑行次数占比超99%；温度与风速的影响呈非线性，凉爽温度（14.4-19.8℃）和和风风速（4.6-7.9m/s）最适宜骑行。
3. ****用户行为模式****：用户可分为高频通勤、低频休闲、偶尔应急三类，其使用时间、地点和习惯差异显著；历史骑行数据和站点位置是预测骑行需求的关键特征，但受限于数据规模，模型预测能力有待提升。
4. ****优化方向****：通过动态调度、站点加密、用户分层运营等策略，可有效提升共享单车运营效率与用户体验。

### **6.2 研究局限性分析**

1. ****数据规模限制****：本研究仅使用1498条骑行数据，样本量较小，且天气数据匹配率低（0.47%），导致模型无法充分捕捉复杂规律，预测性能不佳。
2. ****特征维度不足****：未纳入用户画像数据（如年龄、性别、骑行偏好）、交通流量数据或城市POI数据（如写字楼密度、商场分布），影响对用户行为的深入解释。
3. ****模型选择局限****：随机森林与LightGBM在小数据集上表现有限，未尝试时间序列模型（如ARIMA、LSTM）或图神经网络（GNN，用于建模站点空间关系），可能错失更优预测效果。

### **6.3 未来研究方向展望**

1. ****扩大数据规模****：收集更长时间跨度（如1年以上）或更大范围（如多个城市）的数据，提升数据多样性与样本量，增强模型泛化能力。
2. ****多源数据融合****：整合用户APP行为数据、交通流量数据、POI数据等，构建更全面的特征体系，深入分析用户决策机制。
3. ****先进模型应用****：采用LSTM、Transformer等时序模型捕捉骑行需求的周期性与趋势性，或利用GNN建模站点间的空间依赖关系，优化预测精度。
4. ****可持续运营研究****：探索共享单车与新能源汽车、公共交通的协同调度策略，评估不同运营模式对城市碳排放的影响，推动绿色交通体系构建。

参考文献

[1] 郭贵冰. 推荐系统进展:方法与技术[M]. 北京：科学出版社, 2018.  
[2] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, et al. BPR: Bayesian Personalized Ranking from Implicit Feedback[J]. STATISTICS, 2012.  
[3] 雷余. 基于矩阵分解的社会化协同过滤方法[D]. 吉林大学, 2014.  
[4] 陈钰. 基于信任关系的矩阵分解推荐模型研究[D]. 2017.

致 谢

本论文从选题、设计到论文的撰写和修改都是在温展杰老师的指导下完成的。他对国内外研究热点的熟知和把握，为我指明了研究方向。他严谨的工作作风、渊博的知识储备和对科研工作的高度热情都让我记忆深刻和受益匪浅。在论文完成之际，谨向我的导师表达崇高的敬意和衷心的感谢。

然后，谢谢每一个授予我知识的老师和给予我帮助、鼓励的同学。最后，衷心地感谢我的父母一直以来的默默支持和关怀，让我能够全身心地投入到自己的学习中，让我更加坚定地选择自己要走的路。

附录

1.GitHub代码链接：

2.数据集链接：

3.项目演示网盘链接：