DOI:10.19850/j.cnki.2096-4706.2024.23.013

基于 Pandas+Seaborn+Matplotlib 的城市共享单车租赁分析可视化

徐豪[1]，刘婉月[2]，张自豪[1]

（1.河南工业大学 人工智能与大数据学院，河南 郑州 450001；2.科大讯飞股份有限公司，安徽 合肥 230088）

摘　要：在现代城市交通中，共享单车的普及带来了大量骑行数据，蕴含丰富的用户行为信息。文章旨在通过对 Kaggle 共享单车数据集的深入分析，探讨影响共享单车使用模式的主要因素。采用 Python 的 Pandas 库进行数据处理，并利用 Seaborn 和 Matplotlib 进行可视化分析，以直观展示数据特征和用户行为模式。研究发现，租赁数量与温度、湿度及风速等气象因素密切相关，且在特定时段内租赁活动更为频繁。这一研究不仅展示了 Pandas、Seaborn 及 Matplotlib 在数据可视化中的优越性，还为城市交通管理和共享单车运营提供了数据支撑，从而优化交通管理、提升用户体验。

关键词：大数据分析；可视化；共享单车数据；Python

中图分类号：TP391.4　　**文献标识码：**A　　**文章编号：**2096-4706（2024）23-0058-06

Visualization of Urban Sharing Bicycle Rental Analysis Based on Pandas+Seaborn+Matplotlib

XU Hao[1], LIU Wanyue[2], ZHANG Zihao[1]

(1.School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou 450001, China;
2.iFLYTEK Co., Ltd., Hefei 230088, China)

Abstract: The widespread adoption of sharing bicycles in modern urban transportation has brought a vast amount of riding data which contains rich information about user behavior. This paper aims to conduct an in-depth analysis of the Kaggle sharing bicycle dataset to explore the main factors influencing sharing bicycle usage patterns. It utilizes Pandas library of Python for data processing and employs Seaborn and Matplotlib for visual analysis, providing an intuitive display of data characteristics and user behavior patterns. The study finds that rental quantities are closely related to meteorological factors such as temperature, humidity, and wind speed, with rental activities being more frequent during specific time periods. This research not only demonstrates the superiority of Pandas, Seaborn, and Matplotlib in data visualization, but also provides data support for urban traffic management and sharing bicycle operation, thereby optimizing traffic management and enhancing user experience.

Keywords: Big Data analysis; visualization; sharing bicycle data; Python

0　引　言

自行车共享系统作为传统自行车租赁的新一代，从注册会员到租赁再到归还整个过程实现自动化。用户可轻松在特定位置租用自行车并在另一位置归还，目前全球约有 500 多个共享单车项目，由 50 多万辆自行车组成。因其在交通、环境和健康问题上发挥重要作用，人们对其产生极大兴趣。与此同时，众多研究者也对这些系统所产生的数据兴致盎然。与公共汽车或地铁等其他运输服务不同，共享自行车使用的持续时间、出发时间和到达位置都明确记录在系统中。所以，对今天共享单车数据的分析，不管是针对商业价值或是学术研究，对共享单车数据的分析也尤为重要。

而在“互联网 +”与大数据快速发展的当下，高效处理并展示大量数据成为亟待解决的问题之一[1]。而数据分析可视化能提升数据呈现效果，让用户更迅速、直观地理解复杂数据。通过 Python 对共享单车使用数据进行深入挖掘[2]，既能为城市交通规划者和共享单车运营商提供数据支撑，进而优化交通管理、提高用户体验，又能进一步拓展自行车共享系统的价值。

收稿日期：2024-10-16

基金项目：河南工业大学 2023 年度教育教学改革研究与实践项目（JXYJ2023015）；认知智能国家重点实验室（科大讯飞）开放基金（COGOS-2024HE01）

1　利用 Python 进行数据可视化

要实现更高效的数据分析，采用具有强大绘图功能的 Python 语言处理数据至关重要。Python 作为解释性编程语言，在人工智能、网络爬虫、科学计算与统计等诸多方面广泛应用，其数据分析功能强大，可显著提高数据分析效率 [3]。

之所以运用 Python 语言处理，是因为其拥有众多适用于数据分析和数据可视化的工具库，如 Seaborn、Pandas、Numpy 和 Matplotlib。Matplotlib 是 Python 的一个 2D 绘图库，用于生成图形和图表。它在数据科学和机器学习中广泛应用于数据可视化 [4]。在本实验中，Matplotlib 用于绘制模型训练和验证过程中的损失曲线和准确率曲线，帮助直观分析模型的性能。Pandas 是一个强大的分析结构化数据的工具集，它的使用基础是 Numpy（提供高性能的矩阵运算），用于数据挖掘和数据分析，同时也提供数据清洗功能。Seaborn 则是一个建立在 Matplotlib 基础之上的 Python 数据可视化库，专注于绘制各种统计图形，以便更轻松地呈现和理解数据 [5]。

因此，利用具有强大绘图功能的 Python 语言处理数据对于更有效地开展数据分析十分必要 [6]。

2　数据简介

城市共享单车的租赁分析针对的是共享单车在某一特定时段的租赁数量。本文的数据集是来自 Kaggle 竞赛的美国华盛顿共享单车租赁数据，数据的特征主要包含季节、节假日、工作日、天气、日期、温度、湿度、未注册人员数量、注册人员数量以及共享单车使用量等各种方面的信息，共享单车的租赁数据如表 1 所示。

表 1　城市共享单车的租赁数据展示

字段名	类型	含义
season	int	季节
holiday	int	节假日
Working day	int	工作日
weather	int	天气
datetime	object	日期
temp	float	温度
atemp	atemp	温度
humidity	humidity	湿度
windspeed	windspeed	风速
casual	casual	未注册人数
registered	registered	注册人数
count	count	单车使用量

3　共享单车租赁数据集的预处理

3.1　数据集的导入分析

首先导入数据集并使用 info 函数查看具体数据，程序运行结果如图 1 所示。

```
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    10886 non-null  object
 1   season      10886 non-null  int64
 2   holiday     10886 non-null  int64
 3   workingday  10886 non-null  int64
 4   weather     10886 non-null  int64
 5   temp        10886 non-null  float64
 6   atemp       10886 non-null  float64
 7   humidity    10886 non-null  int64
 8   windspeed   10886 non-null  float64
 9   casual      10886 non-null  int64
 10  registered  10886 non-null  int64
 11  count       10886 non-null  int64
dtypes: float64(3), int64(8), object(1)
memory usage: 1020.7+ KB
```

图 1　info 函数查看具体数据示意图

首先大致观察数据和列标签的形态，接着运用函数查看数据的基本信息。由此可知，该数据一共有 10 886 行、12 列。每一列都对应着共享单车租赁的不同特性。其中，三列数据为浮点型，“datetime”列的数据为字符串类型，其余数据为整型。“datetime”列的数据展示了单车租赁的具体时间和日期，“holiday”与“workingday”分别用“1”和“0”表示“是”和“否”，“season”则用“1”“2”“3”“4”来代表春、夏、秋、冬四个季节。

3.2　缺失值的处理

查看本次数据中是否含有缺失值，使用 missingno 库中的 matrix 函数 Pandas DataFrame 类型创建一个矩阵热图，显示数据中缺失值的分布情况，运行结果如图 2 所示。

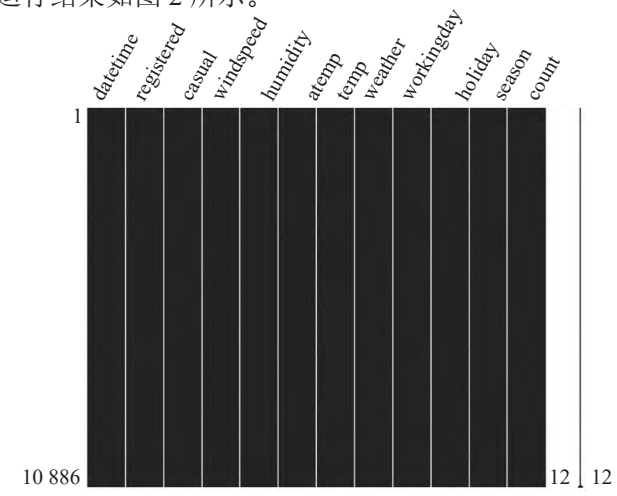

图 2　矩阵热图

缺失值矩阵图展示了数据集中每列的缺失值情况。图中的每一列代表数据集的一列，行代表每个数据样本。白色的条代表缺失值，深色的条代表存在的

数据。通过这种可视化，可以直观地看到数据集中缺失值的分布和模式。所以由图 2 可知本次数据没有缺失值，不需要进行缺失值处理。

3.3 重复值分析

通过 data.drop_duplicates () 函数对重复数据进行清除操作。具体代码如下：

```
duplicate_rows = data.duplicated()
num_duplicates = duplicate_rows.sum()
print(f"Number of duplicate rows: {num_duplicates}")
if num_duplicates > 0:
    print("Duplicate rows:")
    print(data[duplicate_rows].head())
data_cleaned = data.drop_duplicates()
data_cleaned.info()
```

清除重复数据前，数据集的尺寸为 10 886 行 ×12 列；清除后，数据集的大小仍为 10 886 行 ×12 列，这表明本数据集没有重复的数据值。

3.4 共享单车租赁的特征相关性分析

为了了解各个变量与 count（单车租赁数量）的相关性，笔者做了各变量与 count 的相关矩阵热力图，因为 datetime 列数据的类型不属于数值型类型，在之前将数据集中的 datetime 的 year、month、day（具体日期）、hour 以及 weekday 部分转化为数值型数据。但热力图需要数值型数据运算，所以重建 int_df 删去其中非数值的列。以热力图可视化显示了各变量间的相关性，可以找出和共享单车租赁相联系的特征，结果如图 3 所示。

从图 3 当中可以得知，count 与 registered、casual 呈现出高度正相关关系，相关系数分别为 0.7 和 0.97。鉴于 count 等于 casual 与 registered 之和，这种正相关符合预期。count 与 temp 呈正相关，相关系数是 0.39。通常来讲，气温过低时人们往往不太乐意骑车出行。count 和 humidity（湿度）为负相关，湿度过大的天气确实不适合骑车。在考量湿度的时候，也需要同时考虑温度。windspeed 对租车人数的影响似乎不大，相关系数为 0.1，这可能是因为极端大风天气的出现频率较低。在风速处于正常范围内波动时，对人们租车的影响比较小。所以由此可以总结出，不同特征值对租赁数量的影响力度排序为：时段的影响最大，其次是温度、湿度、年份、月份、季节、天气等级、风速、星期几、是否工作日以及是否假日。

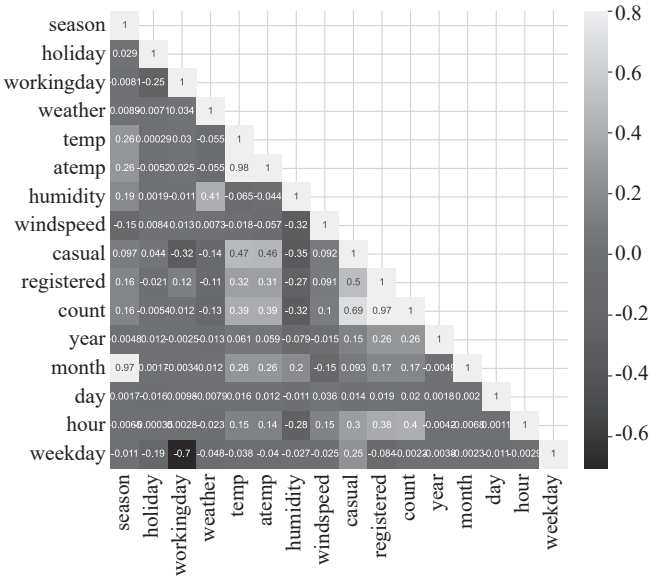

图 3 相关热力矩阵图

3.5 数据的重新处理

对数据集进行一定的分析，可知没有缺失值和异常值。但 datetime 的数据类型是 object，为了之后的数据分析更方便，需要把它转化为时间类型，并拆分为 year、month、week、day、hour、weekday，即日期的处理转换，为接下来的数据分析做准备。替换后的部分数据如图 4 所示。

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count	year	month	day	hour	weekday
0	1	0	0	1	9.84	14.395	81	0.0000	3	13	16	2011	1	1	0	5
1	1	0	0	1	9.02	13.635	80	0.0000	8	32	40	2011	1	1	1	5
2	1	0	0	1	9.02	13.635	80	0.0000	5	27	32	2011	1	1	2	5
3	1	0	0	1	9.84	14.395	75	0.0000	3	10	13	2011	1	1	3	5
4	1	0	0	1	9.84	14.395	75	0.0000	0	1	1	2011	1	1	4	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10881	4	0	1	1	15.58	19.695	50	26.0027	7	329	336	2012	12	19	19	2
10882	4	0	1	1	14.76	17.425	57	15.0013	10	231	241	2012	12	19	20	2
10883	4	0	1	1	13.94	15.910	61	15.0013	4	164	168	2012	12	19	21	2

图 4 内容规整后的数据

4　共享单车租赁的可视化分析

按照影响共享单车租赁数据的各类要素，查看各要素数据的分布情况，便可以得到各要素对共享单车租赁产生的影响 [7]。

4.1　时段对租赁数量的影响

使用 Seaborn 和 Matplotlib 库来按时段分组并计算平均租赁数量，然后将这些结果可视化为折线图，结果如图 5 所示。

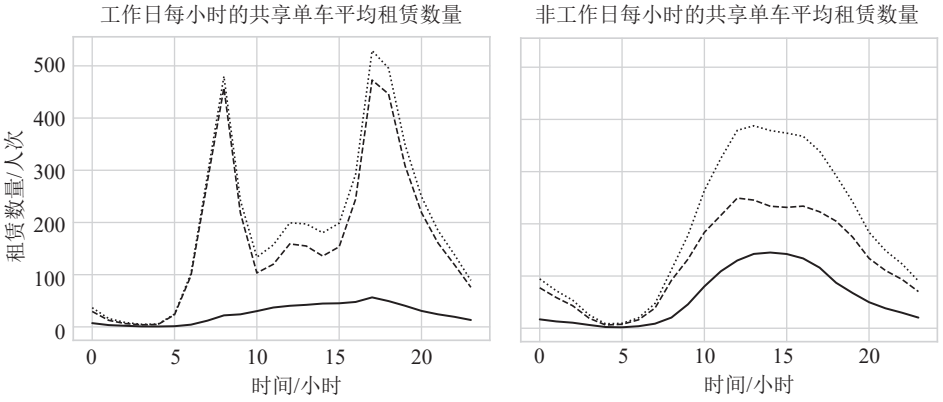

图 5　工作日和非工作日时段对租赁数量影响

由图 5 能够得知，在工作日，会员用户的用车高峰出现在上下班时间，此外中午还有一个小高峰，猜测可能是外出吃午餐的人在用车；而对临时用户起伏比较平缓，高峰期在 17 点左右。对于非工作日而言，租赁数量随时间呈现为正态分布，14 点左右为高峰，4 点左右为低谷，且分布较为均匀。并且会员用户的用车数量远超过临时用户。

4.2　温度对租赁数量的影响

使用 Seaborn 和 Matplotlib 库来按温度大小分组并计算平均租赁数量，然后将这些结果可视化为折线图，结果如图 6 所示。

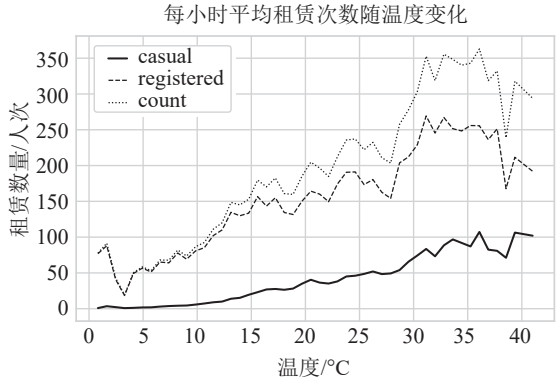

图 6　每小时平均租赁次数随温度变化图

由图 6 可观察到随气温上升租车数量总体呈现上升趋势，但在气温超过 35 时开始下降，在气温 4 度时达到最低点。

4.3　湿度对租赁数量的影响

通过 Pandas 对共享单车数据进行按湿度（humidity）分组，并计算每个湿度区间内的 casual、registered 和 count 的平均租赁数量，生成折线图，折线图如图 7 所示。

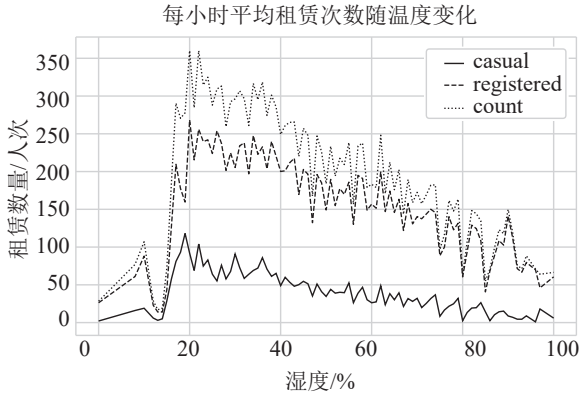

图 7　不同湿度下每小时平均租赁次数折线图

从图 7 得出，在湿度 20% 左右租赁数量迅速达到高峰值，此后缓慢递减。

4.4　季节对出行人数的影响

使用 Seaborn、Pandas 和 Matplotlib 库来按季节分组并计算平均租赁数量，然后将这些结果可视化为小提琴图，结果如图 8 所示。

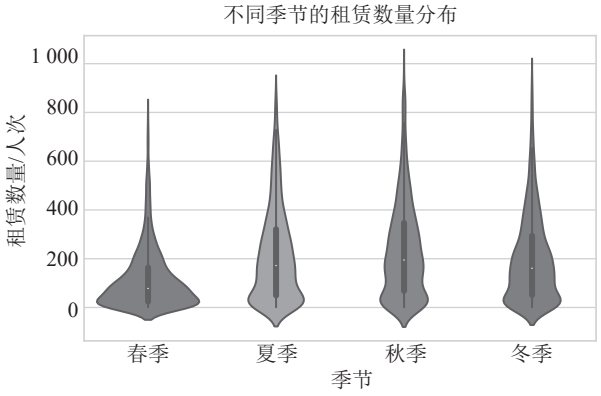

图 8　平均租赁次数趋势随季节变化图

图 8 展示了不同季节租赁数量的分布情况，有效地显示使用需求在秋季迎来高峰，而春季租赁数量最低。

4.5　风速对出行人数的影响

使用 Seaborn 和 Matplotlib 库来按风速大小分组并计算平均租赁数量，然后将这些结果可视化为折线图 [8]。具体代码如下：

```
plt.plot(windspeed_rentals.index,windspeed_rentals['casual'],label='casual', color='black', linestyle='-')
plt.plot(windspeed_rentals.index,windspeed_rentals['registered'],label='registered',color='black', linestyle='--')
plt.plot(windspeed_rentals.index, windspeed_rentals['count'],color='black', linestyle=':', label='count')
plt.title(' 不同风速下每小时最大租赁数量折线图 ')
plt.xlabel(' 风速（米 / 秒）')
plt.ylabel(' 最大租赁数量 / 人次 ')
plt.legend()
```

结果如图 9 所示。

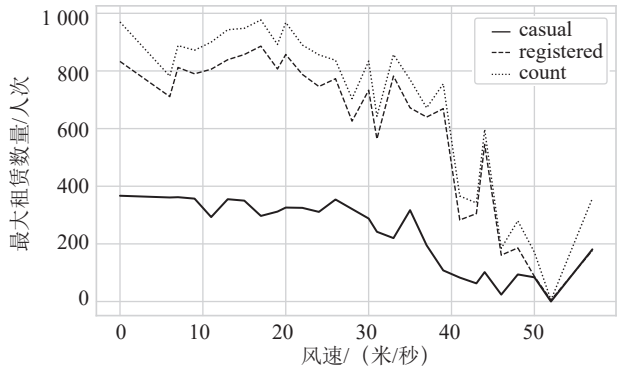

图 9　不同风速下每小时最大租赁数量折线图

图 9 展示了处于不同风速条件下，各类租赁的最大租赁数量的分布状况。可以发现租赁数量与风速呈负相关，即风速越大租赁数量越少，当风速超过 30 米 / 秒时明显减少，然而在风速约为 40 米 / 秒时却出现了一次回升 [9]。

4.6　天气情况对出行情况的影响

通过 Pandas 对共享单车数据按天气（weather）分组，计算每种天气条件下 casual 和 registered 的平均租赁数量，生成 weather_df 数据框。接着，使用 Matplotlib 绘制堆叠柱状图，结果如图 10 所示。

由图 10 可知在不同天气条件下每小时发起的平均租赁数量，由图 10 可以看出晴天少时使用人数最多，小雪小雨时使用人数最少。

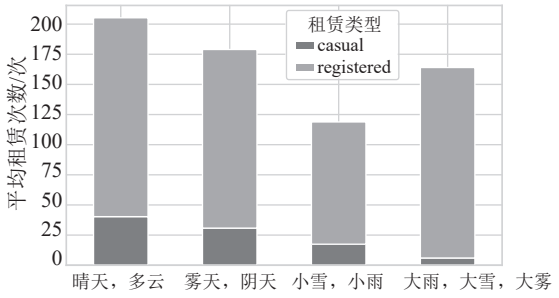

图 10　不同天气情况下每小时平均租赁数量图

4.7　工作日对出行情况的影响

将日期分为周末和工作日两个方面进行对比探讨，对比工作日和非工作日的自行车租赁数量，并通过柱状图和饼图进行可视化 [10]。首先计算工作日和非工作日的平均租赁数量，然后绘制柱状图和饼图，具体操作如下：

通过 Pandas 将共享单车数据依据工作日（workingday）进行分组操作，并计算 casual 与 registered 的平均租赁数量，进而生成 workingday_df 数据框。把工作日的数据存储在 workingday_df_1 中，非工作日的数据存储在 workingday_df_0 里。利用 Matplotlib 创建子图绘制堆叠条形图以展示工作日和非工作日的平均租赁数量，将这两种情形下 casual 和 registered 的比例绘制出来，结果如图 11 所示。

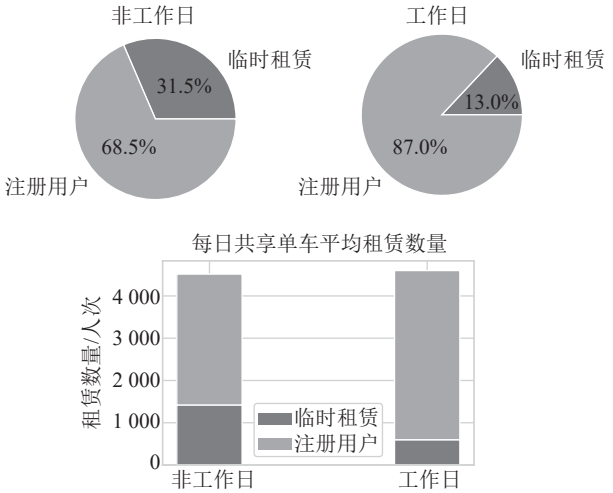

图 11　工作日和非工作日的租赁数据对比图

从图 11 中可知工作日会员用户出行数量较多，临时用户出行数量较少。

5　结　论

用 Python 分析共享单车租赁交易数据集，把数据集中的不同属性用图展示出来并进行简单剖析，会员常于工作日出行频繁，而在节假日却出行较少；临时用户恰与之相反。一季度，出行之人总体为数不多。租赁数量随天气等级攀升而渐次减少。（下转 68 页）

点特性的影响，呈现出复杂的路径和模式。具有高度连接性的节点更容易成为虚假信息的传播中心，而网络中的群体结构则可能促进虚假信息的快速传播。此外，节点的影响力、社交关系强度以及信息传播速度等因素也对虚假信息的传播路径和规模产生了影响。

2）未来研究方向。未来，可以探索更多的算法和技术手段，以更准确地模拟和分析虚假信息在社交网络中的传播过程，关注不同社交网络平台的虚假信息传播特点，以便更全面地了解虚假信息的传播规律。结合用户行为和心理特征，研究如何更有效地识别和干预虚假信息的传播，有望为社交网络中的虚假信息治理提供更为有效的策略和方法。

参考文献：

[1] 晁晓峰 . 面向溯源的虚假危机信息传播主体识别与动机叙事方法研究 [J]. 情报理论与实践，2024，47（4）114-125+113.

[2] 张志勇，荆军昌，李斐，等 . 人工智能视角下的在线社交网络虚假信息检测、传播与控制研究综述 [J]. 计算机学报，2021，44（11）：2261-2282.

[3] 刘知远，张乐，涂存超，等 . 中文社交媒体谣言统计语义分析 [J]. 中国科学：信息科学，2015，45（12）：1536-1546.

[4] SONG C H，YANG C，CHEN H M，et al. CED: Credible Early Detection of Social Media Rumors [J].IEEE Transactions on Knowledge and Data Engineering，2019，33（8）：1-1.

[5] 邵成成 . 在线社会网络中虚假信息传播的研究 [D]. 长沙：国防科技大学，2018.

[6] 张卫东，栾碧雅，李松涛 . 基于信息风险感知的网络虚假信息传播行为影响因素研究 [J]. 情报理论与实践，2019，42（9）：93-98+110.

[7] 刘英杰，刘士虎，徐伟华 . 基于有效路径拓扑稳定性的链路预测方法 [J]. 计算机应用研究，2022，39（1）：90-95.

[8] 郑好，冯虢靓雯，蒲文杰，等 . 基于 Dijkstra 算法的封闭环境全局路径规划 [J]. 汽车实用技术，2023，48（16）：7-11.

[9] 王栋 . 基于改进 Dijkstra 算法的共享停车系统设计 [D]. 南京：南京信息工程大学，2023.

[10] CHO J H，RAGER S，DONOVAN J，et al. Uncertainty-based False Information Propagation in Social Networks [J].ACM Transactions on Social Computing，2019，2（2）：1-34.

作者简介：张邝淇（2001—），男，汉族，贵州安顺人，本科在读，主要研究方向：大数据；王婷（2003—），女，汉族，湖南常德人，本科在读，主要研究方向：金融工程；朱芋霖（2003—），女，汉族，贵州贵阳人，本科在读，主要研究方向：投资学；代富贵（2002—），男，汉族，贵州遵义人，本科在读，主要研究方向：大数据；黄健（1999—），男，汉族，贵州织金人，本科在读，主要研究方向：大数据。

（上接 62 页）小时数对租赁状况影响昭然，会员出行现双高峰之态，非会员则呈正态分布之姿。温度与湿度，对非会员影响颇深，于会员却影响甚微。租赁数量随风速增大而逐步递减。工作日，会员用户出行数量可观，临时用户则甚少；周末之际，会员用户租赁数量下滑，临时用户租赁数量却上扬。

在对共享单车数据的分析中，Python 充分展现出其在数据整理和分析领域的强大优势。通过对大量的共享单车数据进行处理，Python 可以快速地按照特定需求进行分组、聚合等操作，如根据工作日与非工作日对数据分组分析。同时，利用 Python 能够从海量复杂的数据中提取关键信息，绘制出如不同特征下租赁数量变化的图表，使人们对共享单车的使用情况有更全面的认识，切实适应了大数据时代的要求，实用性远超其他编程语言。

参考文献：

[1] 郭鹏，林祥枝，黄艺，等 . 共享单车：互联网技术与公共服务中的协同治理 [J]. 公共管理学报，2017，14（3）：1-10+154.

[2] 钱蕾，周玮腾，韩宝明 . 城市轨道交通运营突发事件数据可视化分析 [J]. 铁道科学与工程学报，2020，17（4）：1025-1035.

[3] 康颢，沈瑶，王博文，等 . 基于 Python 的线性动态电路可视化分析软件设计与实现 [J]. 实验室研究与探索，2022，41（2）：116-120.

[4] 王越，陈国兵，李军 . 基于数据挖掘的故障模式、影响及危害性分析改进方法 [J]. 科学技术与工程，2021，21（24）：10536-10542.

[5] 王彩玲，许欣黎 . 基于 Python 语言的计算机专业招聘信息的爬取及分析 [J]. 现代信息科技，2024，8（16）：88-92+97.

[6] 王晨 . 基于 Python 爬虫的豆瓣 TOP250 电影数据分析与可视化研究 [J]. 现代信息科技，2024，8（16）：93-97.

[7] 李天辉 . 基于 python 的数据分析可视化研究与实现 [J]. 电子测试，2020（20）：78-79.

[8] 赵志凡，邓一哲，张思源，等 . 基于 Python 的城市天气数据可视化分析 [J]. 软件，2024，45（4）：37-39.

[9] 傅哲，辛泓润，余力，等 . 基于使用行为分析的共享单车管理优化研究 [J]. 信息系统学报，2018（2）：81-94.

[10] 柳键，张晋莉 . 共享电单车投放策略的演化分析 [J]. 江西师范大学学报：自然科学版，2023，47（5）：506-512.

作者简介：徐豪（2005—），男，汉族，河南鹤壁人，本科在读，研究方向：人工智能；刘婉月（1994—），女，汉族，河南郑州人，高级工程师，硕士，研究方向：自然语言处理、机器翻译、计算机视觉；张自豪（1988—），男，汉族，河南商丘人，讲师，硕士生导师，博士，研究方向：计算机视觉。