

基于线性回归算法对多彩贵州旅游数据的分析及应用

黄正鹏, 马欣, 陈雪, 刘娜

(贵州工程应用技术学院信息工程学院, 贵州 毕节 551700)

✉ 61194610@qq.com; 3398314549@qq.com; 965633491@qq.com; 2591156674@qq.com



摘要:随着互联网的不断发展,各种数据处理算法层出不穷,这些算法在各行各业中扮演着至关重要的角色。由于旅游业对于一个国家和地区的经济的发展具有极其重要的意义,文中基于 Python 语言,通过数据可视化和线性回归算法对 2003—2018 年多彩贵州旅游数据进行分析,并构建相应的模型。根据研究结果,提出吸引更多的游客到贵州旅游及推动多彩贵州旅游建设的建议策略,为游客打造和提供更优质的旅游项目与服务,促进多彩贵州旅游事业的发展。

关键词:Python; 数据可视化; 线性回归; 贵州旅游

中图分类号:TP30 **文献标志码:**A

Analysis and Application of Colorful Guizhou Tourism Data Based on Linear Regression Algorithm

HUANG Zhengpeng, MA Xin, CHEN Xue, LIU Na

(School of Information Engineering, Guizhou Institute of Engineering Science, Bijie 551700, China)

✉ 61194610@qq.com; 3398314549@qq.com; 965633491@qq.com; 2591156674@qq.com

Abstract: With the continuous development of the Internet, various data processing algorithms emerge endlessly. These algorithms play a crucial role in various industries. As tourism industry is of great significance to the national and regional economic development, this paper proposes to analyze the data of Colorful Guizhou tourism data from 2003 to 2018 through data visualization and linear regression algorithm based on Python language, and construct corresponding models. According to the research results, suggestions and strategies are proposed to attract more tourists to Guizhou and promote Colorful Guizhou tourism, with the objectives to create and provide higher quality tourism projects and services for tourists and accelerate the development of Colorful Guizhou tourism.

Key words: Python; data visualization; linear regression; Guizhou tourism

0 引言(Introduction)

随着人民生活水平的不断提高,旅游已经成为大众放松心情、享受生活的常见方式。同时,互联网技术的快速发展也为旅游业的发展带来了重要机遇,互联网与传统旅游行业的深度融合,使得创造和发展新兴旅游生态系统势在必行^[1]。贵州因为拥有美丽富饶的旅游资源而被称为多彩贵州。近年来,多彩贵州的旅游收入持续增长,旅游服务管理质量和旅游景点的知名度都有所提升,但与旅游业发达的地区相比,旅游人次仍有

在一定差距^[2]。

为了吸引更多的游客到访多彩贵州的景点,运用线性回归算法对 2003—2018 年的旅游数据进行分析是十分必要的。通过数据挖掘和分析,可以了解游客的需求和偏好,并根据结果给出针对性强的建议和策略,可以为当地旅游部门制定决策提供有价值的参考,进而开展更加精准的营销活动和旅游项目推广活动,吸引更多的游客到多彩贵州的景点观光旅游,促进多彩贵州旅游事业的可持续发展。

1 相关技术在多彩贵州旅游数据分析中的应用(The application of related technologies in Colorful Guizhou tourism data analysis)

如图 1 所示,在获取贵州旅游数据的过程中,首先从官方网站获取相关数据,使用 Excel 表格保存数据,并开展数据挖掘、数据处理、数据分析等操作。

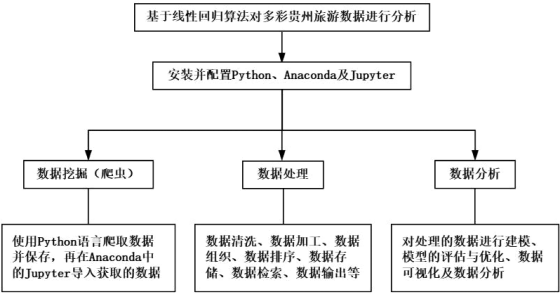


图 1 数据清洗与预处理技术路线

Fig. 1 The technology roadmap of data cleaning and preprocessing

1.1 Python 语言及数据可视化在多彩贵州旅游数据分析中的应用

如图 1 所示,使用 Python 语言在 Anaconda 中的 Jupyter 环境下导入获取的数据,对这些数据进行预处理和清洗。在此过程中,进行数据离散化、非数值类型数据转化、噪声数据清洗等一系列操作,确保数据的准确性和可靠性。对预处理和清洗后的数据进行可视化处理,通过 Python 语言的 Matplotlib 库绘制各类直观且易于理解的图形,以便更高效地分析贵州旅游数据中的潜在信息,为后续的数据分析和决策提供有力的支持。

1.2 线性回归算法在多彩贵州旅游数据分析中的应用

进行线性回归模型分析,以便构建 y (因变量)与影响因素之间的多元线性回归模型。构建多元线性回归模型后,其是否与实际数据有较好的拟合度,模型线性关系的显著性如何等,还需通过数理统计进行检验。多元线性回归模型与一元线性回归模型一样,在获得参数的最小二乘法的估计值后,也需要进行必要的检验与评价,以此判断模型是否可以应用。常用的统计检验有 R 检验和 F 检验^[3]。在构建模型的过程中,需要借助 SPSS 软件,将数据输入其中,找到合适的回归系数,最小化模型的预测误差^[4]。采用最小二乘法最小化观测值与模型预测值之间的残差平方和估算相关系数。

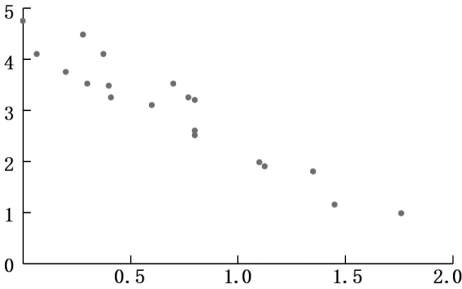
线性规划模型是一种通过目标函数和约束条件求变量的数学规划模型^[5]。运用线性回归算法对 2003—2018 年多彩贵州旅游数据进行分析,获得相应的数据及模型。

线性回归算法的基本原理如图 2 所示,通过分析历史数据,利用找到的回归线构建模型实现预测。ZHU 等^[6]研究了两个自变量多元线性回归中的缺失值问题以及多元线性回归模型 II 中 F 检验的样本量确定问题。在前人研究的基础上,得出本研究初始的多元线性回归模型如下:

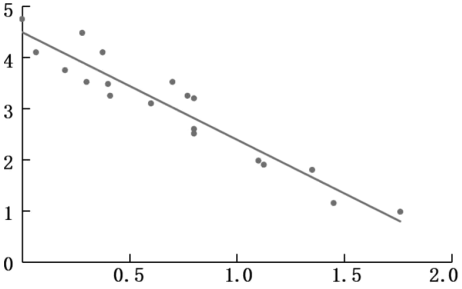
$y=2\ 236.741+11.673x_1-0.138x_2+2.519x_3+4.083x_4-11.006x_5$

采用逐步回归法,建立新的多元线性回归模型:

$y=1\ 067.265+10.885x_1$



(a)线性回归历史数据



(b)线性回归历史数据的回归线

图 2 线性回归算法的基本原理

Fig. 2 The basic principle of linear regression algorithm

利用逻辑回归函数得出相应的模型数据(时间序列模型方程):

$y=1\ 111.958\ 083\times1.247\ 449\ 454x$

其中, $x=1,2,3,\dots$ (2003 年对应数值 1)。

表 1 为建模所需的系数。

表 1 建模所需的系数

Tab.1 Coefficients required for modeling

模型	<i>B</i>	标准错误	<i>Beta</i>	<i>T</i>	显著性
常数	1 933.772	552.299	1.072	3.501	0.039
国内旅游 收入/亿元	11.373	0.473	1.037	24.041	0.000
入境旅游 人数/万人次	26.243	18.673	0.055	1.405	0.255
国际旅游外汇 收入/万美元	-0.248	0.060	-0.105	-4.110	0.026
旅客周转量 亿人/公里	-1.214	0.778	-0.018	-1.560	0.217
星级宾馆/家	8.579	2.446	0.070	3.507	0.039
旅行社/个	-6.114	3.874	-0.043	-1.578	0.213
住宿和餐饮业 增加值/亿元	0.859	5.257	0.009	0.163	0.881
国内旅游 人数/万人次	—	—	—	—	—

表 1 中的国内旅游人数(万人次)是建模所需的因变量,通常受地区生产总值(亿元)和城镇居民人均可支配收入(元)两个自变量的影响,地区生产总值(亿元)和城镇居民人均可支配收入(元)反映了人们的消费水平和生活水平,而人们的消费水平和生活水平又决定了国内旅游人数(万人次)。

通过表 1 中的系数对 2003—2018 年多彩贵州的旅游客流量进行数据拟合分析,创建了多元线性回归模型(表 2),并利用逻辑回归函数得出相应的模型数据——时间序列模型(表 3)。

表2 多元线性回归模型
Tab.2 Multiple linear regression model

模型	Beta	T	显著性	偏相关	允差
地区生产总值/亿元	-0.308 ^b	-1.239	0.341	-0.659	8.310E-5
城镇居民人均可支配收入/元	0.372 ^b	0.941	0.446	0.554	4.024E-5
国内旅游人数/万人次	—	—	—	—	—

表2中的 b 是指模型中的预测值,该预测值包括常数、住宿和餐饮业增加值(亿元)、国际旅游外汇收入(万美元)、旅客周转量(亿人/公里)、星级宾馆(家)、旅行社(个)、入境旅游人数(万人次)及国内旅游收入(亿元)。

表3 时间序列模型
Tab.3 Time series model

模型	Beta	T	显著性	偏相关	允差
国际旅游外汇收入/万美元	-0.046 ^b	-1.722	0.111	-0.445	0.130
旅客周转量/(亿人/公里)	0.035 ^b	1.463	0.169	0.389	0.174
星级宾馆/家	0.016 ^b	0.835	0.420	0.234	0.304
旅行社/个	0.045 ^b	1.167	0.266	0.319	0.070

多元线性回归模型和时间序列模型在分析2003—2018年到多彩贵州旅游的国内游客数据方面表现出色。然而,对比这两个模型的相对误差值发现,时间序列模型在数据分析与预测方面的误差比多元线性回归模型更大。因此,多元线性回归模型在预测到多彩贵州旅游的国内游客流量方面的效果优于时间序列模型,其预测误差小于时间序列模型的预测误差。

基于最优模型,对未来4年(2024—2027年)到多彩贵州旅游的国内游客流量进行预测,结果分别为:(1)2024年,111 435.776 05 万人次;(2)2025年,139 293.495 74 万人次;(3)2026年,146 258.924 97 万人次;(4)2027年,182 833.160 72 万人次。

从以上预测结果来看,多彩贵州未来几年的国内游客流量数据呈上升趋势。只要各旅游景点的管理部门抓住这一机遇,采取适当的措施,定能吸引更多的游客前往贵州旅游,从而推动多彩贵州旅游事业的发展。

2 多彩贵州旅游数据分析中的部分结果
(Partial results in Colorful Guizhou tourism data analysis)

利用Python语言,通过数据可视化和线性回归算法,得到多彩贵州旅游数据分析折线图,如图3所示。

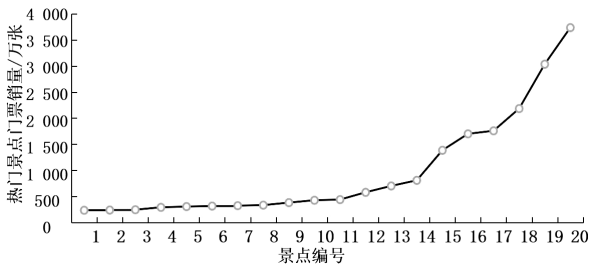


图3 多彩贵州旅游数据分析折线图

Fig. 3 Line graph of Colorful Guizhou tourism data analysis

图3中的景点编号对应的景点分别如下:1为陡坡塘瀑布;2为多彩贵州极地海洋世界;3为万峰林;4为乌蒙大草原;5为天河潭旅游度假区;6为马岭河峡谷;7为西江苗族博物馆;8为中国天眼科普基地;9为肇兴侗寨;10为大七孔景区;11为天星桥景区;12为织金洞;13为黄果树风景名胜;14为青岩古镇;15为百里杜鹃景区;16为黄果树瀑布;17为土城古镇;18为小七孔景区;19为梵净山;20为西江千户苗寨。如图3所示,多彩贵州最受欢迎的10个景点分别是西江千户苗寨、梵净山、小七孔景区、土城古镇、黄果树瀑布、百里杜鹃景区、青岩古镇、黄果树风景名胜、织金洞和天星桥景区。这些景点的共同特点是拥有深厚的历史底蕴、浓厚的民族特色文化氛围以及迷人的自然风光,因此对游客具有极大的吸引力。

为了进一步提高这些景点对游客的吸引力,促进其所在地区旅游业的发展,建议对以上10个景点的旅游设施进行完善,通过加强旅游设施建设和提高服务水平,提升游客的旅游体验,从而吸引更多的游客前来游览。

图4中的城市编号对应的城市分别如下:1为安顺市;2为贵阳市;3为铜仁市;4为遵义市;5为毕节市;6为黔东南苗族侗族自治州;7为黔南苗族侗族自治州。贵州省的4A级和5A级景区数量最多的3个城市分别是安顺市、贵阳市和铜仁市,表明这些城市的热门景点相对较多,自然风光丰富且迷人,因此能够吸引更多的游客前来游览。基于这一情况,可以考虑着重发展这3个城市的旅游业。

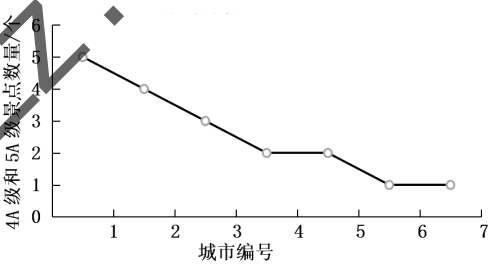


图4 多彩贵州各城市的4A级和5A级景区数量分析折线图
Fig. 4 Line graph of analyzing the number of 4A and 5A level scenic spots in various cities of Colorful Guizhou tourism

图5中的地区编号对应的地区分别如下:1为安顺市;2为贵阳市;3为铜仁市;4为遵义市;5为毕节市;6为黔东南苗族侗族自治州;7为黔南苗族侗族自治州。多彩贵州的星级景点主要分布在安顺市、贵阳市和铜仁市,表明这些地区对旅游业的宣传和推广力度较大,同时景点自然风光迷人、基础设施较为完善,对游客的吸引力非常大。因此,应该继续加强对这3个地区景点的宣传和推广,以进一步带动辐射其他地区的旅游业发展。

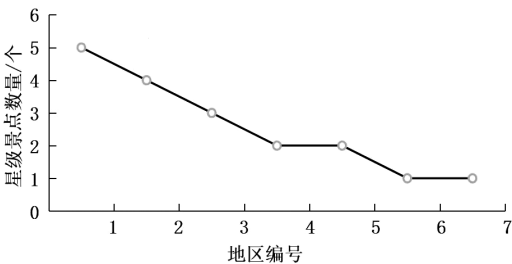


图5 多彩贵州各地区星级景点数量分析折线图
Fig. 5 Line graph of analyzing the number of star scenic spots in various regions of Colorful Guizhou tourism

多彩贵州旅游景点的门票价格如图6所示,国内游客更倾

向于选择门票价格为 0~50 元和 50~100 元的景点游玩,这两个门票价格的景点的国内游客数量分别占多彩贵州国内游客总量的 34.67%;门票价格为 100~150 元和 150~200 元的景点的国内游客数量各占约 10%;门票价格为 200~250 元的景点的国内游客数量约占 5%;门票价格为 250~300 元的景点的国内游客数量约占 1%。值得注意的是,门票价格在 300 元以上的景点的国内游客数量几乎为零。

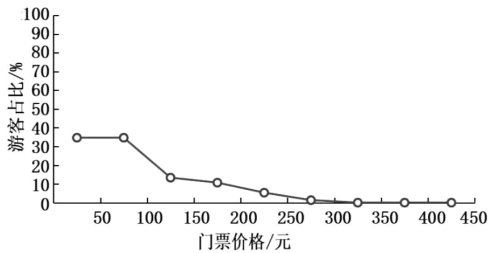


图 6 多彩贵州旅游景点门票价格的折线图

Fig. 6 Line graph of ticket prices for Colorful Guizhou scenic spots

通过对比各个价格区间景点的国内游客数量,可以得出以下结论:国内游客更倾向于选择到门票价格较低的景点旅游。因此,为了吸引更多的游客,建议对门票价格较高的热门景点的门票价格进行调整,使大多数游客都能接受。

根据图 7 的多彩贵州旅游数据分析词云图可以看出:与仙人、历史、洞、地下、文化、苗族、养生、一体、瀑布以及中国等词汇相关的景点对游客具有较大的吸引力。因此,为了吸引更多的游客,建议贵州大力发展与这些词汇相关的景点,并对其进行广泛的宣传和推广。通过这种方式可以更好地满足游客的需求,提升景点的知名度和美誉度,进而推动旅游业的进一步发展。



图 7 多彩贵州旅游数据分析词云图

Fig. 7 Word cloud of Colorful Guizhou tourism data

根据以上研究结果发现,游客更倾向于选择前往具有以下特点的旅游景点游玩。

(1)热门景点:游客喜欢选择有星级景点的地区,这些景点通常具有较高的知名度和吸引力。

(2)价格优惠:门票价格较低的景点往往更受游客欢迎,游客的旅游成本相对较低,更具性价比。

(3)历史底蕴深厚:具有深厚历史底蕴的景点拥有丰富的文化内涵和历史研究价值,因此更能吸引游客。

(4)民族特色文化氛围浓厚:景点所在地区具有的民族特色文化氛围越浓厚,就越能吸引游客,特别是对不同民族文化感兴趣的游客。

(5)自然风光迷人:自然风光迷人的景点由于具有优美的环境和独特的景观,因此更能吸引游客前往。

综上所述,为了吸引更多游客到多彩贵州旅游,贵州省相关部门可以大力发展具有以上特点的景点,并对这些景点进行大力宣传,这样可以更好地满足游客的需求,增强景点的吸引力,并推动贵州旅游业的持续发展。

3 对多彩贵州旅游业发展的建议 (Suggestions for the development of Colorful Guizhou tourism industry)

通过对旅游大数据的分析,可以精准掌握游客的需求和喜好,为旅游资源的优化和调整提供有力的支持。同时,对多彩贵州旅游业的建设^[7]提供一些具有针对性的建议和措施。

(1)对于一些虽然有星级评定,但是游客数量相对较少的旅游景点,可以通过降低门票价格,加强对其历史文化、民族特色文化,以及自然风光的宣传,以吸引更多游客。

(2)对于一些具有星级评定且游客数量较多的旅游景点,需要进一步提高服务质量,提升游客的旅游体验感和满意度,以赢得更多的“回头客”。

(3)对于那些没有星级评定,但具有深厚的历史底蕴、浓郁的民族特色文化氛围和迷人自然风光的景点,可以加大宣传力度,提高服务质量,以吸引更多游客。

(4)对于那些没有星级评定、历史底蕴稍浅、民族特色文化氛围不够浓厚且自然风光特点不足的景点,可以根据当地人的生活特色,对景点进行定向改造,增强其人文气息。同时,加大对该景点的宣传力度,提高服务质量,达到政通人和,进而吸引更多游客,使多彩贵州得天独厚的气候环境和美丽的山水画卷得以更加真切、灵动地展现在世人面前^[8]。

4 结论 (Conclusion)

通过线性回归算法和数据可视化技术,对多彩贵州旅游数据进行分析,能够深入挖掘旅游数据中的潜在规律和发展趋势,从而为多彩贵州旅游业提供有价值的参考信息。同时,将大数据和算法灵活应用到旅游业中,能够有效地提升旅游服务的品质和效率,吸引更多的游客来多彩贵州旅游。相信在互联网、大数据和各种算法的推动下,多彩贵州的旅游业将迎来新的发展机遇,展现出其独特魅力。

参考文献 (References)

- [1] 郑艳. “互联网+”时代贵州旅游行业的机遇与挑战[J]. 中国国际财经(中英文), 2017(8): 39.
- [2] 杨世明, 魏配轩. 大数据背景下贵州旅游服务管理质量优化路径分析[J]. 中国商论, 2018(27): 63-64.
- [3] 林彬. 多元线性回归分析及其应用[J]. 中国科技信息, 2010(9): 60-61.
- [4] 宋亚飞. 贵州省国内旅游客流量预测分析研究——基于多元线性回归分析[J]. 艺术科技, 2021, 30(22): 111-114.
- [5] 韩中庚. 数学建模方法及其应用[M]. 2版. 北京: 高等教育出版社, 2009: 6-20.
- [6] ZHU F K, LIU M Y, LING S Q, et al. Testing for structural change of predictive regression model to threshold predictive regression model[J]. Journal of business & economic statistics, 2022, 41(1): 228-240.
- [7] 王福帅. 大数据理念下贵州省旅游业信息化转型研究[J]. 西部经济管理论坛, 2019, 30(5): 80-86.
- [8] 陆艺, 尹昌应, 黄娅. 大数据时代贵州省智慧旅游发展现状及对策[J]. 贵州农业科学, 2017, 45(4): 157-162.

作者简介:

黄正鹏(1979-),男,硕士,副教授。研究领域:算法,大数据。

本文通信作者。

马欣(2000-),女,本科生。研究领域:大数据技术。

陈雪(2000-),女,本科生。研究领域:大数据技术。

刘娜(2001-),女,本科生。研究领域:大数据技术。