

综合实验设计说明书

（2018）

Python结课论文

年 级 2016

组 号 第1组

组 长 1610121144 马骏超

指导教师 许涛

完成时间 2018年12月24日

# 目录

[目录 1](#_Toc533445410)

[1．数据集摘要 2](#_Toc533445411)

[2．项目简介 2](#_Toc533445412)

[3．数据读取与处理 3](#_Toc533445413)

[4．分析结果 10](#_Toc533445414)

[5．结论与展望 16](#_Toc533445415)

# 1．数据集摘要

|  |  |
| --- | --- |
| 名称 | BuddyMove Data Set Data Set |
| 特征简介 | Multivariate,Text,Real |
| 记录数 | 249 |
| 分析目标 | 到南印度的旅行者目的地分布情况；南印度的旅游目的地最受欢迎的情况；南印度的各种旅游目的地的发展情况等。 |

# 2．项目简介

（1）数据来源

这个数据集来自holidayiq.com的249位评论者发布的目的地评论，直到2014年10月。考虑了南印度各个目的地中的6个类别的评论，并且每个评论者（旅行者）的每个类别的评论数量都被捕获。

（2）属性

·属性一：唯一用户ID

·属性二：体育场馆等的评论数量

·属性三：宗教机构等的评论数量

·属性四：自然景观等的评论数量

·属性五：剧院展览等的评论数量

·属性六：商场购物等的评论数量

·属性七：公园野餐等的评论数量

（3）数据量

数据量：249

（4）基本统计特征

基本统计特征：多变量，文本，真实

（5）分析目标

1.到南印度的旅行者目的地分布情况

2.南印度的旅游目的地最受欢迎的情况

3.南印度的各种旅游目的地的发展情况

等

（6）分析手段

1.对于分析目标1，将数据集中所有的六个场景的评论数逐类相加，将六个类别的数据总和用柱状图和饼状图表示，每个场景的总评论数就显示出来了。

2.对于分析目标2，数据集中每个人都对上述六个场景进行了评论，我们取个人评论数量最多的那个场景加一分，遍历249份数据，六个场景的分数用柱状图和饼状图表示，最受欢迎的场景就显示出来了。

3.对于分析目标3，将数据集中每个场景的每个人的评论数从头至尾取出，每个场景都做一个折线图（由于每个场景都有249份数据，所以每10条数据取一个平均值，最后9条取平均值，共取得25条数据），增长趋势一目了然。

（7）结论简述

1.去往南印度的旅行者对于“宗教机构”“自然景观”“剧院展览”“商场购物”和“公园野餐”的兴趣充分，而对“体育场馆”的兴趣很低。

2.去往南印度的旅行者最感兴趣的是当地的“自然景观”，最无感的是当地的“体育场馆”。

3.南印度的旅游业发展势头持续高涨，六处场地的留言数都在高速增长。

# 3．数据读取与处理

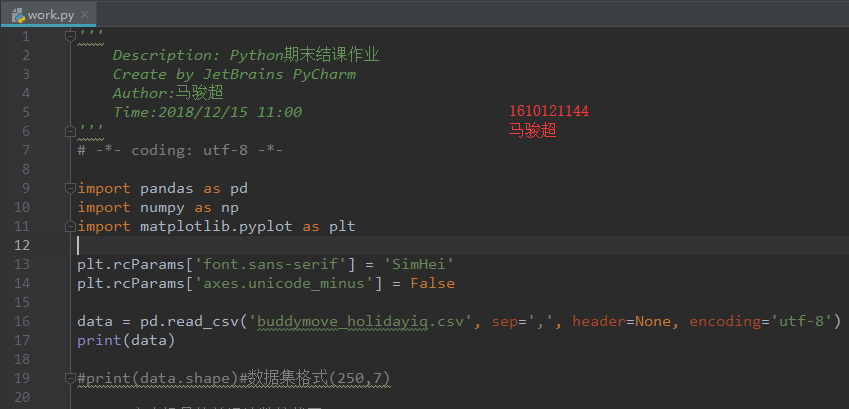
（1）数据处理环境

编程语言：Python3.7

编译器：PyCharm + numpy（插件） + pandas （插件）+ matplotlib（插件）

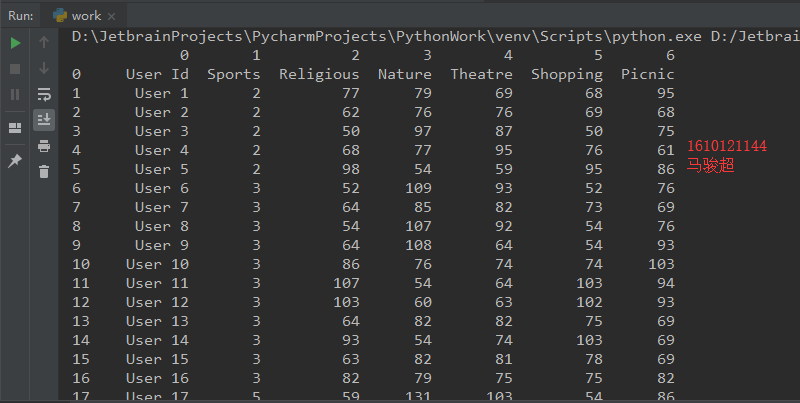
（2）数据处理过程

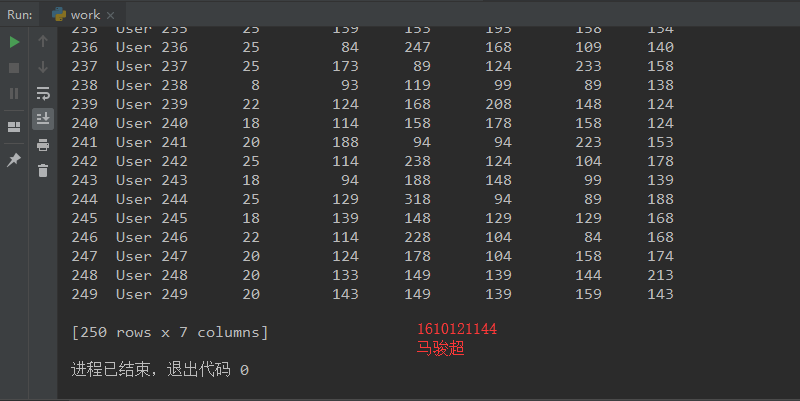
1.读取数据



数据集文件为”buddymove\_holidayiq.csv”

读出结果如下图





2.数据预处理

因为数据较少，数据集只有250\*7，所以通过手动查看数据即可，发现并无非法数据，此数据集的数据预处理结束。

若数据集中的数据量较多，可使用data.isnull()函数判断数据是否为空来进行数据的预处理。

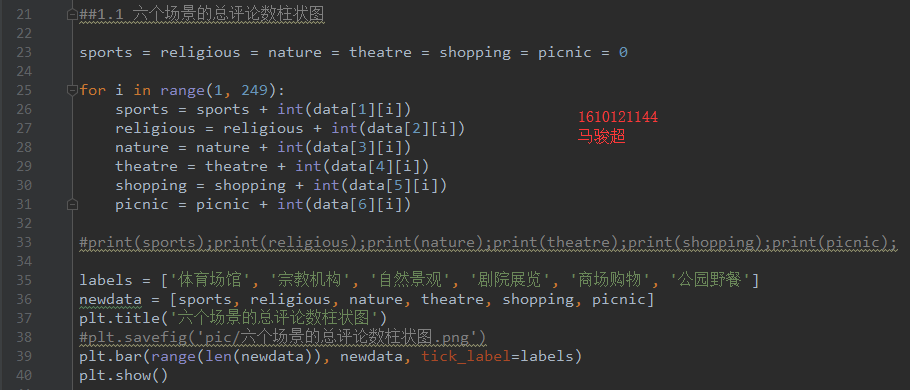
3.分类

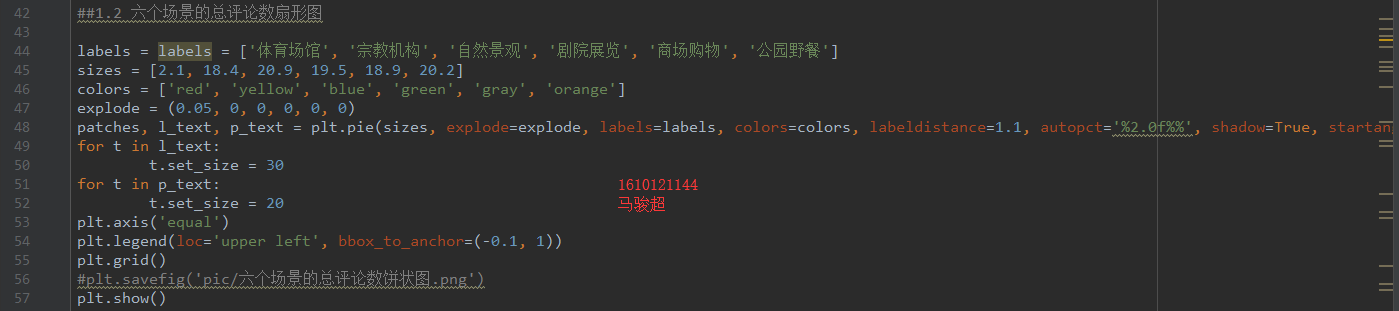
通过对数据集的读取可以看出，此数据集本身就是分类数据集，数据通过7个属性“User ID”“Sports”“Religious”“Natur”“Theatre”“Shopp”和“Picnic”来分类存储。

4.聚类

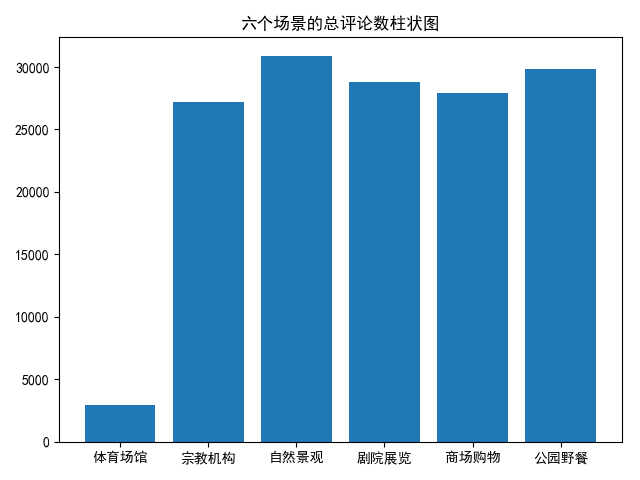
只举一例（代码中多次使用）。

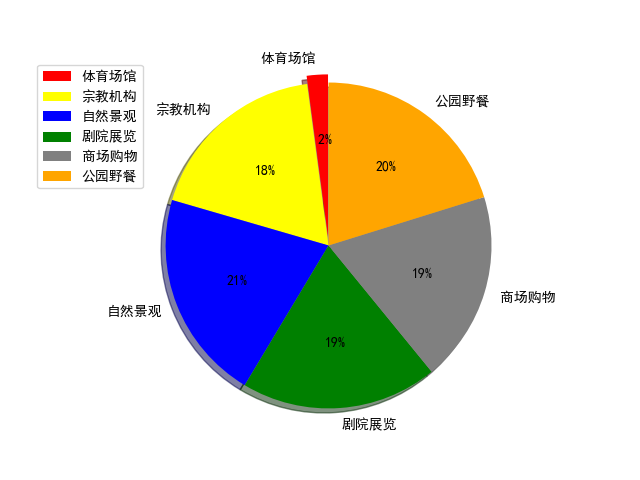
对于分析目标1（到南印度的旅行者目的地分布情况），将上述后六个属性的各自249条数据相加，得到6条六个属性的总和数据，用柱状图和饼状图表示。





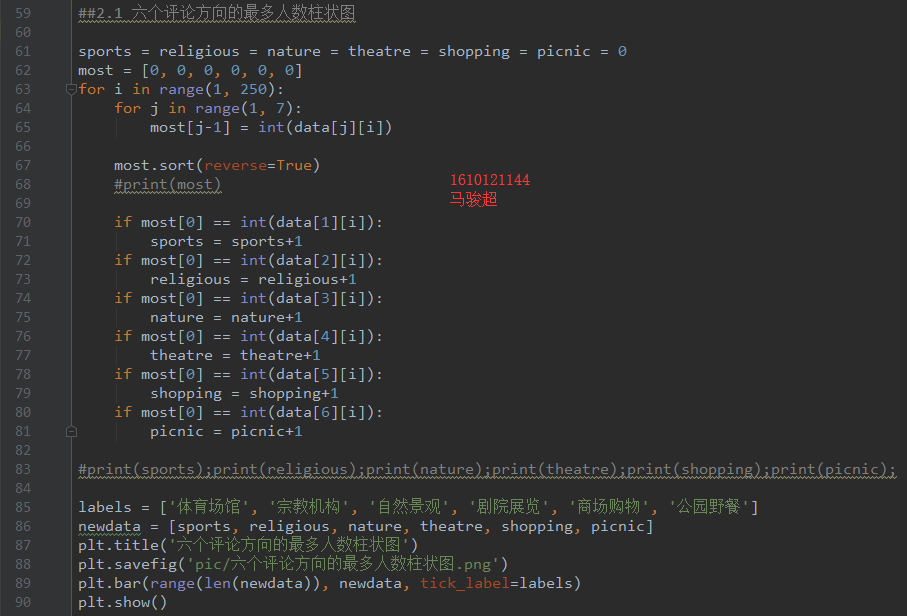
图像结果如下图



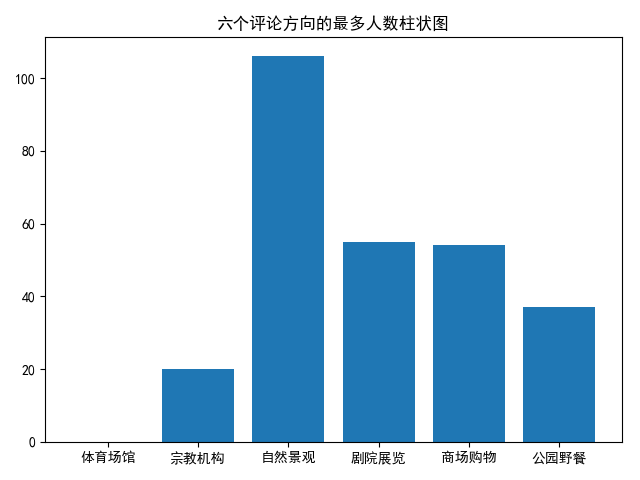


5.核心算法

算法一：六个变量对应六个属性，变量初始值为0。六个属性中的评论数据，最大者该属性对应的变量+1，最终得到六个评论方向的最多人数柱状图。



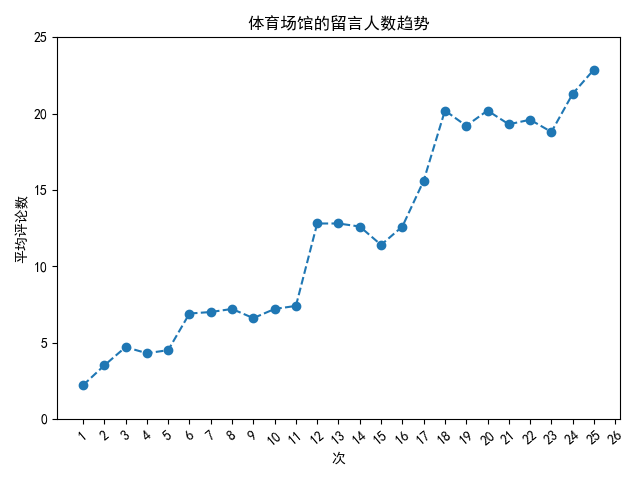
结果如下图



算法二：将数据集中每个场景的每个人的评论数从头至尾取出，每个场景都做一个折线图（由于每个场景都有249份数据，所以每10条数据取一个平均值，最后9条取平均值，共取得25条数据），绘制出xxxx的留言人数趋势折线图。



结果如图



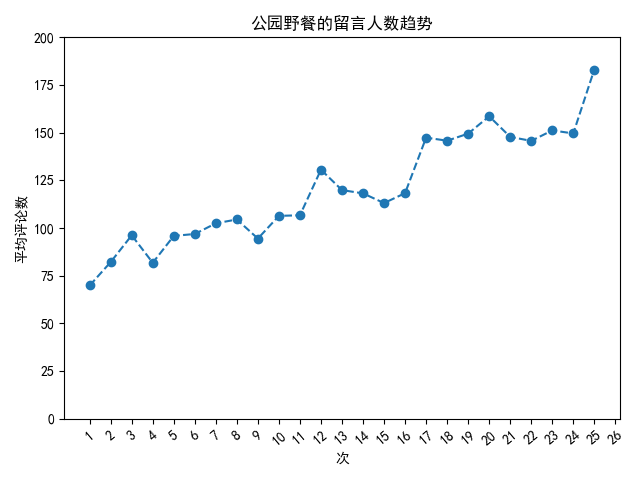
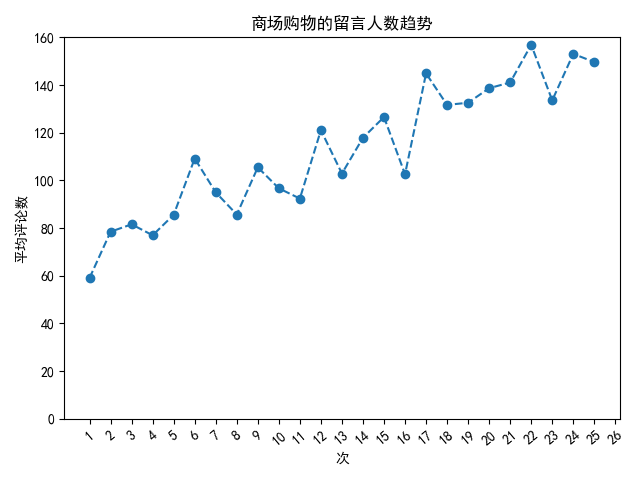
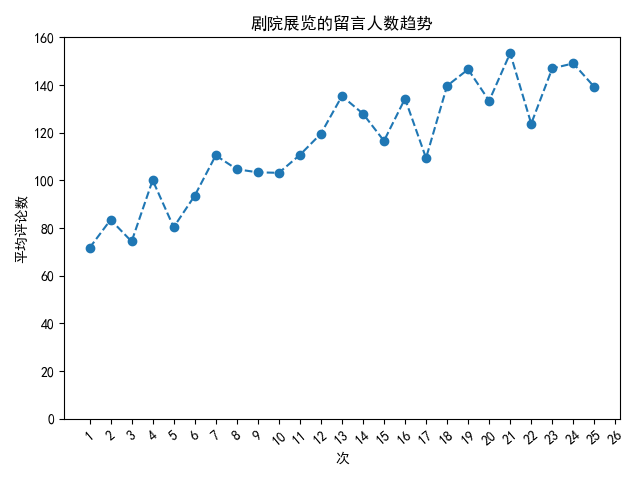
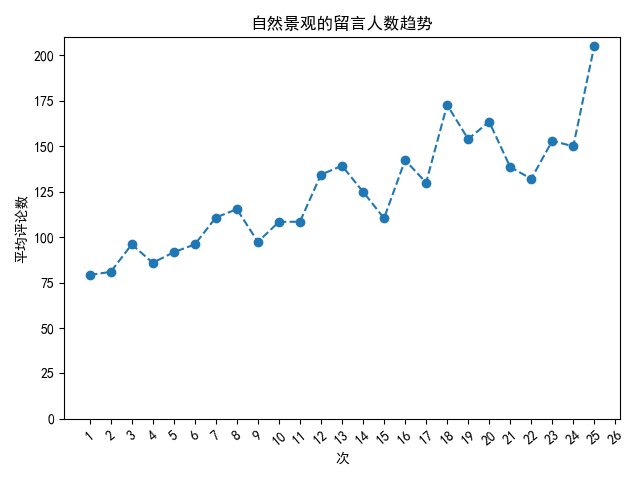
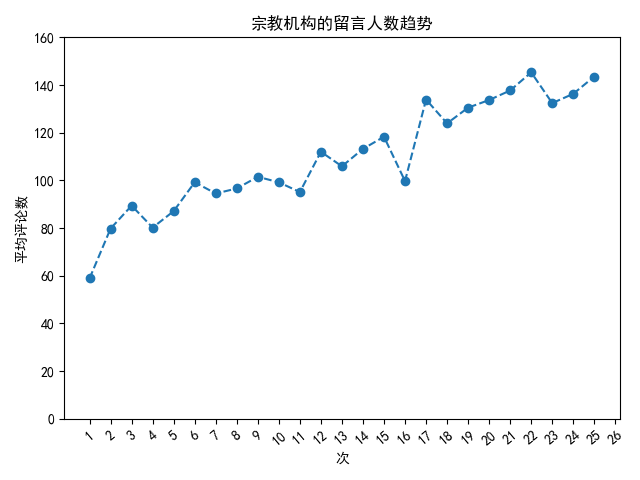
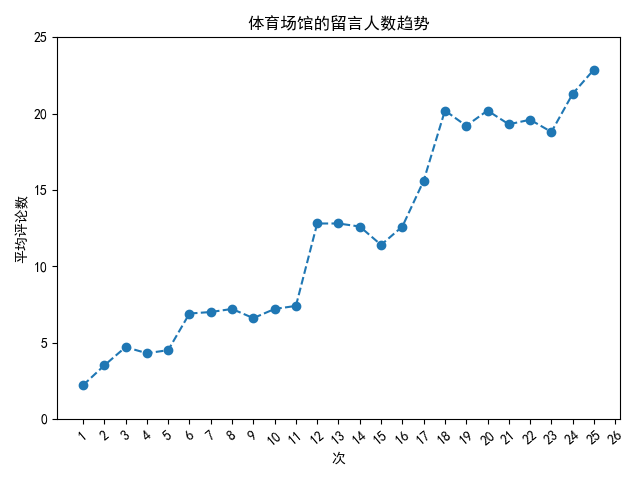
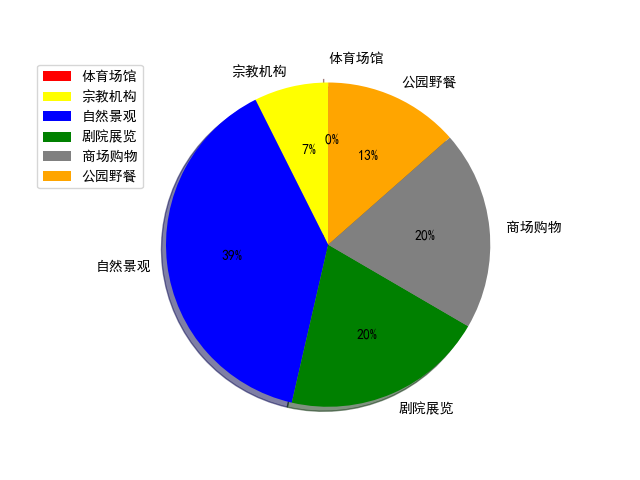
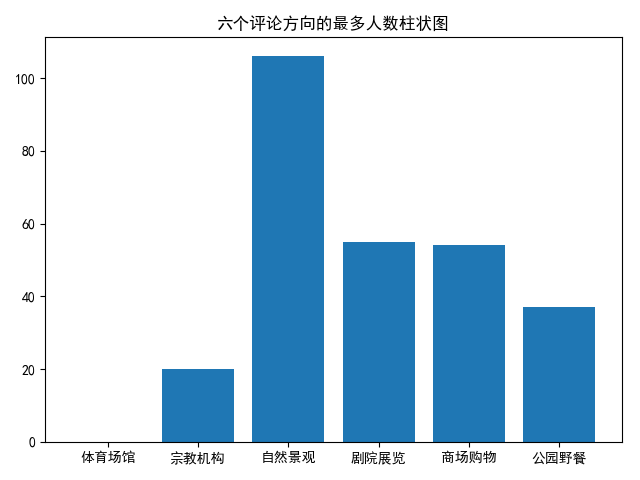
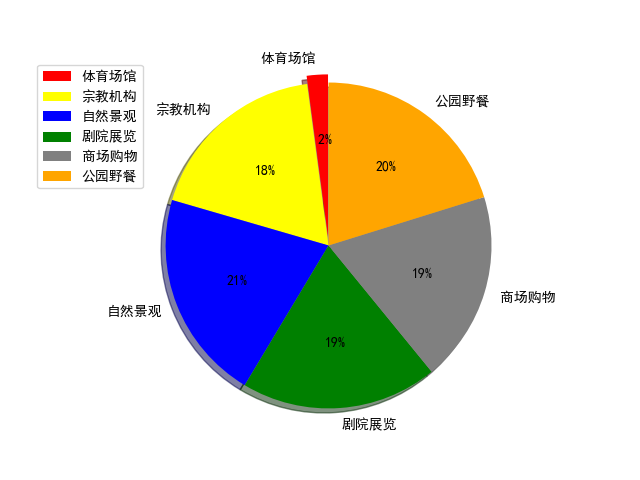
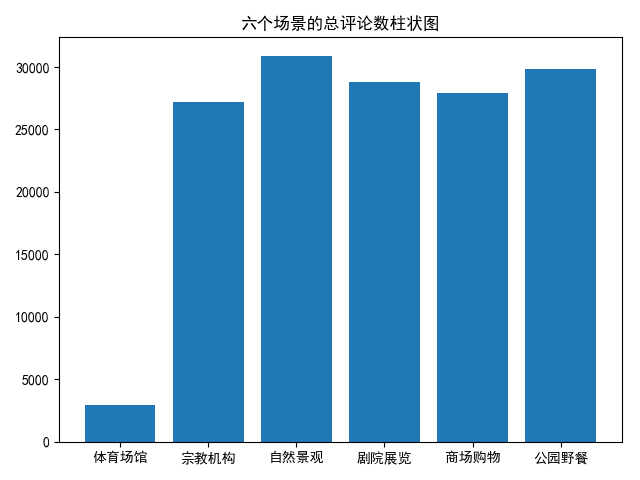
# 4．分析结果

1.去往南印度的旅行者对于“宗教机构”“自然景观”“剧院展览”“商场购物”和“公园野餐”的兴趣充分，而对“体育场馆”的兴趣很低。

2.去往南印度的旅行者最感兴趣的是当地的“自然景观”，最无感的是当地的“体育场馆”。

3.南印度的旅游业发展势头持续高涨，六处场地的留言数都在高速增长。

4.图示



# 5．结论与展望

（1）总结

本次分析实践采用了柱状图和饼状图来描述六个场景评论的数量占比情况，并采用了折线图描述南印度地区的旅游发展情况。分析出了南印度地区的旅游热点地区和旅游发展情况等信息。

（2）结论

1.去往南印度的旅行者对于“宗教机构”“自然景观”“剧院展览”“商场购物”和“公园野餐”的兴趣充分，而对“体育场馆”的兴趣很低。

2.去往南印度的旅行者最感兴趣的是当地的“自然景观”，最无感的是当地的“体育场馆”。

3.南印度的旅游业发展势头持续高涨，六处场地的留言数都在高速增长。

（3）展望

1.此次使用的数据集属性较少，数据量较少，处理难度较小。以后的学习中可以增大难度，锻炼能力。

2.在对数据进行聚类、排序等操作的过程中，使用了sort()等方法，明白了Python中很多情况下都是以String类型存放的数据，在进行排序、赋值、计算等操作时，需要转化为Int类型才可以进行操作。

3. 此次对数据集中数据的处理时，高级算法使用较少，算法使用较为单一，在遇到对数据的提取、处理等问题时，编写的算法的时间、空间复杂度过于复杂。以后的学习中可以多使用巧妙、高级算法对数据进行提取和分析，提高自己的思维和编程能力。