2021年全国职业院校技能大赛

高职组

“大数据技术与应用”

赛项赛卷（GZ-2021041-08卷）

任

务

书

参赛队编号：

**背景描述**

企业消费服务平台，为大中小型企业提供基于云化的消费场景一站式智能消费、智能管控，帮助企业获得更高效、简单、美好的消费管理。从“费控+支付”出发，到覆盖全场景支出的创新模式，让员工在数字化平台上直接完成所有消费，从员工下单、到财务入账，全流程实现自动化统一结算、统一数据分析。解决传统差旅系统面临的场景覆盖不全、员工体验差、消费体验割裂等情况，真正做成一套让企业节省支出，让员工满意的差旅平台。

企业消费服务平台的出现将原来传统的差旅行程放到网络平台上，更广泛的传递差旅信息，互动式的交流更方便客人的咨询和订购，越来越多的人在出行的时候使用企业消费服务平台预订机票、火车票、住宿等，使得更多的商家愿意与企业消费服务平台建立合作，提升住宿场所的营业额，这也为企业消费服务平台的发展带来新的机遇，为了抓住这个机会，“企业消费服务平台”需要从地域、订单来源等多种维度进行分析，明确未来重点拓展合作商家的方向。公司要求多个小组进行分析，并提出相应建议，你所在的小组也在其中，需要通过数据采集、数据清洗、数据分析和数据可视化获得相关论据，提出未来重点拓展合作住宿场所的方向。

你们作为该小组的技术人员，是这次技术方案的核心成员，请按照下面的步骤完成本次技术展示任务，并提交分析报告，祝你们成功！！！

## 模块A：Hadoop平台及组件的部署管理（15分）

**环境要求：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 主机名 | 类型 | 用户 | 密码 |
| 1 | master | 主节点 | root | 123456 |
| 2 | slave1 | 从节点 | root | 123456 |
| 3 | slave2 | 从节点 | root | 123456 |

master主机上MySQL数据库用户名密码是root/123456

相关软件安装包在/chinaskills目录下

### 任务一：Hadoop全分布部署管理

本环节需要使用root用户完成相关配置，安装Hadoop需要配置前置环境，具体部署要求如下：

1. 将/chinaskills下的JDK包解压到/usr/local/src路径下，将命令（使用绝对路径）复制粘贴至对应报告中；
2. 设置JDK环境变量，并使环境变量只对当前root用户生效，将环境变量配置内容复制粘贴至对应报告中；
3. 从master复制JDK环境变量文件到slave1、slave2节点，将命令（命令中使用绝对路径）和结果复制粘贴至对应报告中;
4. 环境中已创建ssh密钥，实现从master到slave1节点的无密码登录，将命令和结果复制粘贴至对应报告中；
5. 将/chinaskills下的hadoop包解压到/usr/local/src，不修改解压后文件夹名称, 将命令复制粘贴至对应报告中；
6. 配置Hadoop全局环境变量，并使环境变量只对当前root用户生效，将命令和环境变量配置内容复制粘贴至对应报告中；
7. 通过Hadoop命令格式化namenode，并将格式化命令复制粘贴至对应报告中；
8. 启动Hadoop，使用相关命令查看所有节点Hadoop服务进程，并将命令和结果复制粘贴至对应报告中。

### 任务二：Kafka组件部署管理

1. 将Zookeeper配置完毕后，在各节点启动Zookeeper，查看Zookeeper状态，并将命令和Zookeeper运行状态结果复制粘贴至对应报告中；
2. 修改Kafkaserver.properties文件，并将修改的内容复制粘贴至对应报告中；
3. 启动Kafka，并将Kafka启动命令和输出结果前10行复制粘贴至报告中。

### 任务三：Sqoop组件部署管理

本环节需要使用root用户完成相关配置，已安装Hadoop及需要配置前置环境，具体部署要求如下：

1. 解压/chinaskills路径下的Sqoop安装包到/usr/local/src路径下，并使用相关命令，修改解压后文件夹名为sqoop，进入sqoop文件夹，并将查看内容复制粘贴至对应报告中。
2. 修改Sqoop环境变量，并使环境变量只对当前root用户生效,并将环境变量配置内容复制粘贴至对应报告中。
3. 修改并配置sqoop-env.sh文件，将sqoop-env.sh文件修改内容复制粘贴至对应报告中。
4. 测试Sqoop连接MySQL数据库是否成功，将执行命令和执行结果复制粘贴至对应报告中。

## 模块B：数据采集与处理（20分）

1. 网站解析，利用Chrome查看网页源码，分析企业消费平台网站网页结构。
2. 打开企业消费平台网站，在网页中右键点击检查，或者F12快捷键，查看元素页面；
3. 检查网站：浏览网站源码查看所需内容。
4. 从企业消费平台网站中爬取需要数据，按照要求使用Python语言编写爬虫代码，爬取指定数据项，并对结果数据集进行数据探索、以及必要的数据处理操作。请将符合题目要求的代码答案复制粘贴至对应报告中。

具体步骤如下：

1. 创建爬虫项目
2. 构建爬虫请求
3. 按要求定义相关字段
4. 获取有效数据
5. 将爬取到的数据保存到指定位置

至此已从企业消费平台网站中爬取了所需数据，下一步我们要将爬取结果进一步进行相关数据操作。

**详细数据描述：**

1. 请创建Scrapy项目chinaskills\_accommodation（C:\ chinaskills\_accommodation），从网站（网站地址在竞赛平台模块B中给出）中爬取页面相关字段；将抓取结果保存为json格式文件，并命名为accommodations.json。每条信息请以Key：Value格式单独保存为一行数据。

例如：

{“name1”:”\*\*\*”, ”name2”:”\*\*\*”,…….}

**具体任务要求：**

### 任务一：网页源码对应字段

使用Chrome浏览器，查看相关数据页面第一页第一行数据记录的网页源码，并将以下内容及答案完整复制粘贴至对应报告中。

“房间数”网页源码中对应的当前td的源代码为:

“城市”网页源码中对应的当前td的源代码为:

“城市直销拒单率”网页源码中对应的当前td的源代码为:

### 任务二：创建Scrapy工程

自行创建Scrapy工程编写爬虫代码，爬取“住宿场所名称,城市,商圈, 房间数, 住宿场所订单，住宿场所实住间夜，城市平均实住间夜、城市直销拒单率”相关数据，通过爬虫代码分页爬取，将使用re解析分页链接的程序代码复制粘贴至对应报告中。

### 任务三：在MySQL中创建数据库表

根据爬取字段，在MySQL中创建crawl数据库，在该数据库中创建accommodations1表（包含住宿场所名称,城市,商圈, 房间数, 住宿场所订单，住宿场所实住间夜），创建accommodations2表（包含住宿场所名称,城市,商圈, 城市平均实住间夜、城市直销拒单率）, 将爬取数据写入相应数据表中，并分别统计accommodations1表和accommodations2表的总行数，将统计结果复制粘贴至对应报告中。

### 任务四：对数据库表排序

爬虫程序运行结束后查看MySQL中crawl数据库的acommodations1表，按房间数字段倒序排序，返回前100行数据，将命令与查看结果复制粘贴至对应报告中。

### 任务五：对数据表填充处理

请根据步骤3中accommodations1表中的数据，对数据集中“住宿场所订单”字段的缺失值，使用平均值进行填充。查看填充后的数据集前5条记录，将查看结果复制粘贴至对应报告中。

### 任务六：对数据表删除处理

请根据步骤3中accommodations2表中的数据，对数据集中存在空值的字段进行删除。查看删除后的数据集条数，将查看结果复制粘贴至对应报告中。

accommodations2表删除后条数为：

## 模块C：数据清洗与分析（25分）

现已从相关网站及平台获取到原始数据集，为保障用户隐私和行业敏感信息，已进行数据脱敏。数据脱敏是指对某些敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。在涉及客户安全数据或者一些商业性敏感数据的情况、不违反系统规则条件下，对真实数据进行改造并提供测试使用，如身份证号、手机号等个人信息都需要进行数据脱敏。

相关数据文件中已经包含了数据采集阶段从企业消费平台网站上爬取的数据集，其中包含了来自不同城市的多家住宿场所的销售信息，你的小组需要通过编写代码或脚本完成对相关数据文件中住宿场所销售管理数据的清洗和整理，并完成数据计算和分析任务。综合利用MapReduce、Spark、Storm、分布式存储系统、数据仓库Hive、数据推送工具等技术，使用Java、Python、Scala等开发语言，完成本阶段数据清洗、处理、分析及数据挖掘等任务。通过多个维度分析住宿场所的销售信息，并以此评价住宿场所销售业绩、区域的游客接纳能力、接纳质量等指标。

初始数据集来自多个网站及平台系统，且为多次采集汇总结果，因此数据集中不可避免地存在一些脏数据，即源数据不在给定的范围内或对于实际业务毫无意义，或是数据格式非法，以及在源系统中存在不规范的编码和含糊的业务逻辑。

请分析相关数据集，根据题目要求实现数据清洗及分析。

### 任务一：数据清洗

住宿场所销售数据涉及到多个平台及数据库对接，个别信息由于人为操作失误或计算机故障等原因产生了数据缺失值。缺失值是一种常见的脏数据情况，由于粗糙数据中缺少信息而造成的数据缺失或截断。现有数据集中某个或某些属性的值是不完全的。对于缺失值的处理，从总体上来说分为缺失值删除和缺失值插补。当缺失值过多时，信息条目本身的价值也会随之降低，此时如果对缺失值进行填补则将产生结果的人为干预。结合行业数据本身特点及上述考虑，请你根据题目具体参数要求实现以下功能：将缺失值大于n个的数据条目从原始数据集中剔除,并输出剔除的条目数量。

**详细描述：**

数据源文件存放于/chinaskills/accommodationdata.csv文件夹下，请编写Spark程序，按照如下要求实现对数据的清洗，并将结果输出至HDFS文件系统中/accommodationsparktask1。

1. 解析/chinaskills下相关数据文件；
2. 按照题目要求剔除缺失数据信息大于n（n=3）个字段的数据记录，并以打印语句输出删除条目数；
3. 程序打包并在Spark上运行，结果（结果中需包含csv的header列名）输出至HDFS文件系统中/accommodationsparktask1。

**具体任务要求：**

1. 编写Spark(Standalone模式)程序，删除数据源中缺失值大于3个字段的数据记录，打印输出删除条目数，将运行结果复制粘贴至对应报告中

打印语句格式如下：

=== “删除缺失值大于3个的字段的数据条数为\*\*\*条”===

1. 使用Hadoop shell命令查看清洗后输出的结果文件总行数（包含csv的header列名）/accommodationsparktask1，将查看命令与执行结果复制粘贴至对应报告中。

缺失值是一种常见的脏数据情况，对于数据缺失值的处理，从总体上来说分为缺失值删除和缺失值插补两种处理方式。当缺失值过多时，信息条目本身的价值也会随之降低，此时如果对缺失值进行填补，则数据分析结果可能会受到干扰，有失客观性。结合行业数据本身特点及上述考虑，对于数据集中数值字段缺失的情况，通常可以采用填充固定值、均值、众数、KNN填充、以及把缺失值作为新的label等方式处理。同时，不当的填充可能会令后续的分析结果出现导向性偏差，当缺失信息较少时可采用删除的方式来进行处理。下面请根据题目具体参数要求处理关键字段缺失。

**详细描述：**

请以上述**1、（题目数字编号）**任务结果数据集作为输入数据源，编写Spark程序，按照如下要求实现对数据的清洗，并将结果输出至HDFS文件系统中/accommodationsparktask2。

1. 解析/accommodationsparktask1中的数据文件；
2. 将关键字段“星级”、“评论数”、“评分”中存在缺失值的数据删除；
3. 程序打包并在Spark(Standalone模式)上运行，结果（结果中需包含csv的header列名）输出至HDFS文件系统中/accommodationsparktask2。

**具体任务要求：**

1. 在Spark(Standalone模式)上运行代码，将字段“星级”、“评论数”、“评分”中存在缺失值的数据删除，打印输出删除条目数，将打印输出结果复制并保存至对应报告中（复制内容需包含打印语句输出结果的上下各5行运行日志）。

打印语句格式如下：

=== “删除的数据条数为\*\*\*条”===

1. 使用Hadoop shell命令查看清洗后输出的结果文件（包含csv的header列名）/accommodationsparktask2总行数，将查看命令与执行结果复制粘贴至对应报告中。

### 任务二：数据挖掘分析

城市游客接纳能力是城市规划建设中的重要指标，其中城市的住宿场所数量和房间数量是城市游客接纳能力的关键要素。请编写程序或脚本根据住宿场所管理网站中的数据统计各城市的相关信息，并写入指定的数据库或数据文件。

**详细描述：**

请以上个任务的输出结果作为输入数据源，编写Spark程序统计各城市的住宿场所数量和房间数量，以城市房间数量降序排列并输出前10条统计结果，输出至HDFS文件系统中/accommodationhive1，同时创建并写入数据表table3\_1。要求输出字段包含：省份、城市、住宿场所数量、房间数量。

**具体任务要求：**

1. 编写Spark程序，统计各城市的住宿场所数量和房间数量，以城市房间数量降序排列，输出前10条统计结果至HDFS文件系统中/accommodationhive1，使用Hadoop shell命令查看统计结果，将查看命令与执行结果复制并粘贴至对应报告中；
2. 在accommodationdata数据库中创建表table3\_1；将统计结果写入表table3\_1中，输出表table3\_1前10行数据，将执行命令与结果复制并粘贴至对应报告中。

通过收集数据能够更好地收集用户需求，从而可以提供更有针对性和个性化的服务，最终能够产生更多的忠诚会员并带来更多订单。但企业消费平台平台销售也存在用户拒单等情况，拒单原因有很多：例如，平台信息不同步，信息更新不及时；分销层次过多，导致无法及时查证订单；住宿场所违反企业消费平台规则擅自以低价让客户取消订单，这种情况又叫做“切单”。企业消费平台平台需要统计用户订单的分布情况，以此发现平台缺陷及用户、商家的行为模式，企业消费平台平台据此调整营销策略。根据现有数据及给定参数完成订单数据统计，并写入指定的数据库或数据文件。

**详细描述：**

1. 请以**上个任务**结果数据集作为输入数据源，编写MapReduce程序统计各省直销拒单率，以直销拒单率升序排列并输出前10条统计结果，同时创建并写入数据表table3\_2；
2. 要求输出字段包含：省份、直销拒单率；

**具体任务要求：**

1. 在accommodationdata数据库中创建表table3\_2，将执行命令复制并粘贴至对应报告中；
2. 运行代码，将统计的直销拒单率按升序排列，并将前10条统计结果输出至HDFS文件系统/accommodationhive2中（将分区数设置为1），并写入数据表table3\_2中，将执行命令与结果复制并粘贴至对应报告中。

## 模块D：数据可视化（20分）

MySQL数据库中的相关数据集包含了城市、省份、评分、评论数等多项基础信息字段。请使用Flask框架，结合Echarts完成下列题目。

数据库账号：root 密码：123456

自行创建代码工程路径为：C:\chinaskills\_hotel

### 任务一：条形图呈现住宿场所订单量

住宿场所订单量是反映住宿场所入住数量的重要指标之一，某省订单数量一定程度上可以反应出该省住宿场所入住情况，为了更好地分析全国各省住宿场所订单量，请根据指定表中数据统计出全国各省住宿场所订单量的情况，并以指定图例进行呈现。

**详细描述：**

请以数据库中相关数据表中的数据为数据源，分析并统计各省订单数量，绘制条形图。

**具体任务要求：**

1. 提取数据表相关字段，分别统计各省的订单数量，在控制台按照“订单数量”降序排列，打印输出各省份的订单数量；

打印语句格式如下：

==省份：\*\*\* =订单数量：\*\*个==

……

1. 使用Flask框架，结合Echarts绘制“订单数量”的条形图，标题为“各省订单数量”（字体颜色：红色，加粗），纵坐标为省份名称，横坐标为订单数量，并将可视化结果复制粘贴至对应报告中。

### 任务二：多个饼图呈现各省份不同等级住宿场所占比

企业消费平台为了更好地发展企业业务，向企业客户推荐符合其预定的协议住宿场所，需要分析上题中Top5省份的“三星级/舒适”、“四星级/高档”和“五星级/豪华”住宿场所以及“其它类别”住宿场所（除上述三类外，其余类型住宿场所均归为“其它类别”）的占比情况。请根据指定表中数据，以指定图例进行呈现。

**详细描述：**

请根据数据库中相关数据集中省份、星级等相关字段统计各省份不同等级住宿场所的占比，绘制多个饼图。

**具体任务要求：**

1. 根据数据表相关字段分别统计上题Top5省份不同星级住宿场所的占比，打印输出各省份名称以及各类型住宿场所的占比情况；

打印语句格式如下：

==省份：A=舒适型住宿占比：\*\*===

==省份：A=高档型住宿占比：\*\*===

==省份：A=豪华型住宿占比：\*\*===

==省份：B=舒适型住宿占比：\*\*===

==省份：B=高档型住宿占比：\*\*===

==省份：B=豪华型住宿占比：\*\*===

……

1. 使用Flask框架，结合Echarts在组合图中绘制多个饼图，每个城市各星级住宿场所占比情况分别用一个饼图进行呈现，将可视化结果截图并保存。

### 任务三：柱状图呈现城市出租率

出租率是反映住宿场所经营状况的一项重要指标，它是已出租的客房数与住宿场所可以提供租用的房间总数的百分比。住宿场所出租率的情况可以在一定程度上反应出该住宿场所的整体运营的情况，为了更好的分析指定住宿场所的入住情况，请根据相关表中数据完成出租率分析，通过指定图例进行呈现。

**详细描述：**

请以数据库相关数据表中的数据作为数据源，以柱状图呈现城市出租率。

**具体任务要求：**

1. 提取数据表相关字段，在控制台按照“各省住宿场所出租率”降序排列，打印输出各省名称及包含的住宿场所数量；

打印语句格式如下：

==1.\*\*\*省=住宿场所数为\*\*\*个=出租率为：\*\*\*===

==2.\*\*\*省=住宿场所数为\*\*\*个=出租率为：\*\*\*===

……

1. 使用Flask框架，结合Echarts绘制柱状图。柱状图主标题为“各省住宿场所出租率”（字体要求：红色、加粗、斜体），副标题为出租率前十的省份；纵坐标为出租率，横坐标为省份名称（按照出租率降序排列）；将可视化结果截图并保存。

### 任务四：折线图呈现连锁住宿场所出租率

1. 连锁住宿场所一般都具有全国统一的品牌形象识别系统、全国统一的会员体系和营销体系、价格相比较很有优势，更适合大众化消费。连锁住宿场所无论在装修、服务还是信誉上都有较大的竞争优势，所以连锁住宿场所是出差、旅游住宿的首选。但是由于三线城市会员流动差、高素质管理人员相对短缺、营销环境与消费特点存在差异等问题，一些已经成熟住宿场所管理模式在三线城市可能并不受用，甚至会出现水土不服的现象。请根据现有数据及给定参数，统计指定连锁住宿场所的经营状况，并以指定图例进行呈现。

**详细描述：**

1. 数据库中相关数据表已保存了指定地区的某连锁住宿场所销售信息。请根据地区划分，统计题中某连锁住宿场所的出租率（保留6位小数），并以折线图呈现；
2. 要求统计以下指定地区住宿场所相关信息，指定地区包括：东北、华北、华东、华中、西北、西南、华南；
3. 指定地区省份映射表，如表1。

表1：地区省份映射表

|  |  |
| --- | --- |
| 地区 | 省份 |
| 华东地区 | 山东、江苏、安徽、浙江、江西、福建、上海 |
| 华南地区 | 广东、广西、海南 |
| 华中地区 | 湖北、湖南、河南 |
| 华北地区 | 北京、天津、河北、山西、内蒙古 |
| 西北地区 | 宁夏、新疆、青海、陕西、甘肃 |
| 西南地区 | 四川、云南、贵州、西藏、重庆 |
| 东北地区 | 辽宁、吉林、黑龙江 |

**具体任务要求：**

1. 根据数据表相关字段分别统计某连锁住宿场所在各地区的出租率（保留6位小数），在控制台按照“出租率”降序排列，打印输出各地区名称以及出租率；

打印语句格式如下：

==1.\*\*\*地区, 出租率为\*\*\*===

==2.\*\*\*地区，出租率为\*\*\*===

……

1. 使用Flask框架，结合Echarts绘制折线图，主标题为“指定地区的住宿场所出租率”（字体要求：红色、加粗、斜体），副标题为“某连锁住宿场所的出租率”,纵坐标为出租率，横坐标为地区；输出折线图，将可视化结果截图并保存。

### 任务五：散点地图呈现各城市住宿场所间夜数

住宿场所的间夜量也叫间夜数，是住宿场所在某个时间段内，房间出租率的计算单位，关于住宿场所间夜量的计算公式为间夜量=入住房间数\*入住天数。例如某住宿场所今天入住的房间数为500，则今天的间夜量=500\*1=500，而又比如某住宿场所这个月（30天）的平均每天入住房间数为400，则这个月的间夜量=400\*1\*30=12000。请根据指定表中数据统计住宿场所间夜数相关数据，并以指定图例进行呈现。

**详细描述：**

请以数据库相关表中的数据作为数据源，各城市住宿场所间夜数散点地图。

**具体任务要求：**

1. 根据数据表相关字段分别统计各城市住宿场所间夜数，打印输出各城市的间夜数，在控制台按照“间夜数”降序排列，打印输出各地区名称以及间夜数；

打印语句格式如下：

==\*\*\*市：间夜数为\*\*\*===

==\*\*\*市：间夜数为\*\*\*===

……

1. 使用Flask框架，结合Echarts绘制散点地图，主标题为各城市住宿场所间夜数（字体要求：红色、加粗、斜体）；输出各城市住宿场所间夜数散点地图，将可视化结果截图并保存。

### 任务六：多线雷达图呈现各省份住宿场所综合情况

企业消费平台需要综合评判一个城市住宿场所运营情况，会涉及到多方面住宿场所数据，例如像高端住宿场所数量、订单数量、住客评分、评论数量、出租率、200元/晚以下快捷住宿场所数量等信息，请根据指定表中数据统计相关数据，并以指定图例进行呈现。

**详细描述：**

请根据数据库中相关数据表，统计各城市住宿场所综合运营情况，并以多线雷达图表达。

**具体任务要求：**

1. 根据数据库中相关数据表分别统计北京、上海、广东、四川、海南各地四星/五星住宿场所的数量、平均评分、评论数、各省住宿场所出租率、直销拒单率，在控制台按照“城市”名称升序排列，打印输出各城市住宿场所的多项运营指标；

打印语句格式如下：

==城市：A，四星/五星住宿场所数量为：\*\*\*===

==城市：A，平均评分为：\*\*\*===

……

==城市：B，四星/五星住宿场所数量为：\*\*\*===

==城市：B，平均评分为：\*\*\*===

……

1. 使用Flask框架，结合Echarts绘制多线雷达图，标题为各省份住宿场所综合情况（字体要求：红色、加粗、斜体）；输出多线雷达图，将可视化结果截图并保存。

## 模块E：综合分析（15分）

假定你为企业消费平台的管理者，在综合理解住宿场所业务数据的基础上，通过以上模块A、B、C、D的相关结论，对未来拓展合作住宿场所方向做出预测，根据题目要求进行分析，并编写输出分析报告。

根据上述任务中的结论，分析以下内容，并编写分析报告。从住宿场所分布维度，结合多省份住宿场所综合运营情况，对企业消费平台未来拓展合作住宿场所的方向提出建议。

**分析报告要求：**

### 任务一：通过数据及图示分析指定省市运营情况

结合平台相关数据文件，根据北京、上海、四川、广东、海南各省份的直销拒单率和出租率的条形图，说明几个省份住宿场所的综合运营情况。分别以文字描述和图例进行说明；

### 任务二：对通过图示和计算业务分析原因

结合模块D可视化分析对某连锁住宿场所在不同地区的住宿场所出租率的统计，说明影响住宿场所出租率的原因可能有哪些？对于提高该连锁住宿场所的出租率，您有哪些建议？分别以文字描述和图例进行说明；

### 任务三：对企业消费平台未来拓展合作建议和意见

对企业消费平台未来拓展合作住宿场所的方向提出建议（不少于3条建议）；

# 附录：补充说明

## json数据格式样例

{"name": "南京国美家庭旅社公寓南林店", "detail": {"SEQ": "nanjing\_10116", "国家": "中国", "省份": "江苏", "城市": "南京", "处于商圈": "锁金村地区 玄武湖地区 中山陵景区","是否为客栈": 0, "住宿场所星级": "二星及其他","业务部门": "低星", "剩余房间": 8, "图片数": 0, "住宿场所评分": "1", "用户点评数": 1,"城市平均实住间夜": "51.701686747", "住宿场所总订单": 0,"住宿场所总间夜": 0, "住宿场所实住订单": 0, "住宿场所实住间夜": 0, "住宿场所直销订单": 0, "住宿场所直销间夜": 0, "住宿场所直销实住订单":0, "住宿场所直销实住间夜": 0,"住宿场所直销拒单": 0, "住宿场所直销拒单率": null, "城市直销拒单率": "0.0282838180927","拒单率是否小于等于直销城市均值": 0, "最低房间价格": "306"}}

## fastjson-1.2.41.jar常用API（java）

1. 实例化

JSONObject（）;

1. JSON解析包  
   com.alibaba.fastjson.JSON;   
   com.alibaba.fastjson.JSONObject;   
   com.alibaba.fastjson.JSONArray;   
   com.alibaba.fastjson.JSONException;
2. 常用API方法：
3. public static final Object parse（String text）; // 把JSON文本parse为JSONObject或者JSONArray
4. public static final JSONObject parseObject（String text）； // 把JSON文本parse成JSONObject
5. public static final T parseObject（String text, Class clazz）; // 把JSON文本parse为JavaBean
6. public static final JSONArray parseArray（String text）; // 把JSON文本parse成JSONArray
7. public static final List parseArray（String text, Class clazz）; //把JSON文本parse成JavaBean集合
8. public static final String toJSONString（Object object）; // 将JavaBean序列化为JSON文本
9. public static final String toJSONString（Object object, boolean prettyFormat）; // 将JavaBean序列化为带格式的JSON文本
10. public static final Object toJSON（Object javaObject）; 将JavaBean转换为JSONObject或者JSONArray。

## fastjson-1.2.41.jar常用API【Spark（scala）】

1. json解析包

com.alibaba.fastjson.JSON

1. 常用API
2. 实例化：

JSON.parseObject（x）

1. 默认值：如果该key没有值默认为null：

jsonObject.getOrDefault（key,默认值）

jsonObject.getOrDefault（"name",""）

1. 获取该key的value值

jsonObject.get（json的key）

jsonObject.get（“name”）

1. 判断key是否存在

jsonObject.containsKey（key）

1. 添加kv键值对

jsonObject.put（key,value）

## 控制台输出运行日志样例



## 方差、均方根差的定义

1. 方差MSE：概率论中方差用来度量[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F/828980" \t "_blank)和其[数学期望](https://baike.baidu.com/item/%E6%95%B0%E5%AD%A6%E6%9C%9F%E6%9C%9B/5362790)（即[均值](https://baike.baidu.com/item/%E5%9D%87%E5%80%BC/5922988" \t "_blank)）之间的偏离程度。统计中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的[平均数](https://baike.baidu.com/item/%E5%B9%B3%E5%9D%87%E6%95%B0/11031224" \t "_blank)。
2. 均方根差RMSE：均方根误差，是观测值与真值偏差的平方和观测次数n比值的平方根。RMSE是计算观测值与其真值，或者观测值与其模拟值之间的偏差。

## 间夜定义

间夜又称间夜数，是住宿场所在某个时间段内，房间出租率的计算单位。例如20间房入住2晚，为40间夜数。

## 出租率计算公式

出租率 = 当月发生的总间夜数/当月所能提供的总房间数

## 线性回归预测数据源data\_accommodation\_mult.csv字段名

SEQ、省份、城市、商圈、是否为客栈、星级、房间数、评论数、平均评分数、城市平均间夜、住宿场所总订单、住宿场所总间夜、住宿场所实住订单、住宿场所实住间夜、住宿场所直销订单、住宿场所直销实住订单、住宿场所直销间夜、住宿场所直销实住间夜、城市直销拒单、城市直销拒单率、住宿场所企业消费平台实住订单

## 数据可视化表字段说明

表radar\_lines

|  |  |
| --- | --- |
| province | 省份 |
| accommodation\_num | 住宿场所数 |
| avg\_score | 平均分 |
| comment\_num | 评论数 |
| lease\_rate | 出租率 |
| direno\_rate | 直销率 |

表platform\_rate

|  |  |
| --- | --- |
| accommodationname | 住宿场所名称 |
| province | 省份 |
| city | 城市 |
| trading\_area | 商圈 |
| is\_in | 是否为客栈 |
| star | 星级 |
| room | 房间数 |
| comment | 评论数 |
| score | 评分 |
| city\_avgorder | 城市平均订单 |
| city\_avgmidnight | 城市平均间夜 |
| city\_avgrealorder | 城市平均实住订单 |
| city\_avgrealmidnight | 城市平均实住间夜 |
| accommodation\_order | 住宿场所订单 |
| accommodation\_midnight | 住宿场所总间夜 |
| accommodation\_realorder | 住宿场所实住订单 |
| accommodation\_realmidnight | 住宿场所实住间夜 |
| accommodation\_direorder | 住宿场所直销订单 |
| accommodation\_diremidnight | 住宿场所直销间夜 |
| accommodation\_direrealmidnight | 住宿场所直销实住间夜 |
| accommodation\_direnoorder | 住宿场所直销拒单 |
| city\_direorder | 城市直销订单 |
| city\_direrealorder | 城市实住订单 |
| city\_direnoorderrate | 城市直销拒单率 |
| ………………………………… | ………………………….. |

表platform

|  |  |
| --- | --- |
| province | 省份 |
| city | 城市 |
| order\_num | 总订单 |
| midnight\_num | 总间夜 |
| real\_num | 实住订单 |
| realmidnight\_num | 实住间夜 |
| city\_lease\_rate | 出租率 |

表city\_let\_rate

|  |  |
| --- | --- |
| province | 省份 |
| t企业消费平台l\_orders | 总订单 |
| t企业消费平台l\_nigth | 总间夜 |
| real\_orders | 实住订单 |
| real\_night | 实住间夜 |
| city\_rate | 出租率 |

表night

|  |  |
| --- | --- |
| province | 省份 |
| city | 城市 |
| night | 间夜 |