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-2021041-01卷）

任

务

书

参赛队编号：

背景描述

据央视财经报道，2020年我国O2O市场规模突破万亿元，O2O市场存在着巨大的潜力。特别是餐饮和外卖行业，占据市场较大份额，并且业务增长迅速。截至2020年底，全国外卖总体订单量已超过171.2亿单，同比增长7.5%，全国外卖市场交易规模达到8352亿元，同比增长14.8%。我国外卖用户规模已接近5亿人，其中80后、90后是餐饮外卖服务的中坚消费力量，消费者使用餐饮外卖服务也不再局限于传统的一日三餐，下午茶和夜宵逐渐成为消费者的外卖新宠。为把握这一商业机遇，ChinaSkills公司计划进驻外卖平台市场，现需对大规模成熟外卖平台进行详细评估调研，采集多方多维度数据，寻找行业痛点，摸清市场需求，以技术为手段为投资保驾护航。

为完成该项工作，你所在的小组将应用大数据技术，以Python、Java、Scala作为整个项目的基础开发语言，基于大数据平台综合利用MapReduce、Spark、MySQL、Scrapy、Flask、ECharts等，对数据进行获取、处理、清洗、挖掘、分析、可视化呈现，力求实现对公司未来的重点战略方向提出建议。

你们作为该小组的技术人员，请按照下面任务完成本次工作，并编制综合报告。

# 模块A：Hadoop平台及组件的部署管理（15分）

**环境说明：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |
| 补充说明：主节点MySQL数据库用户名/密码：root/123456  相关软件安装包在/chinaskills目录下  所有模块中应用命令必须采用绝对路径 | | | | |

### 任务一：Hadoop HA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，命令复制并粘贴至对应报告中；
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并粘贴至对应报告中；
3. 从master复制JDK环境变量文件到slave1、slave2节点，命令和结果复制并粘贴至对应报告中；
4. 配置SSH密钥登录，实现从master登录到slave1，命令和结果复制并粘贴至对应报告中；
5. Zookeeper配置完毕后，在slave2节点启动Zookeeper，查看Zookeeper运行状态，将命令和结果复制并粘贴至对应报告中；
6. Zookeeper、Hadoop HA配置完毕后，在master节点启动Hadoop，并查看服务进程状态，并将结果复制并粘贴至对应报告中；
7. Hadoop HA配置完毕后，在slave1节点查看服务进程，将命令及结果复制并粘贴至对应报告中。

### 任务二：Hive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. 将指定路径下的Hive安装包解压到（/usr/local/src）下，使用绝对路径，将命令复制并粘贴至对应报告中；
2. 把解压后的apache-hive-1.2.2-bin文件夹更名为hive；进入hive文件夹，并将查看命令及结果复制并粘贴至对应报告中；
3. 设置Hive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并粘贴至对应报告中；
4. 将Hive安装目录里hive-default.xml.template文件更名为hive-site.xml；并将更改命令复制并粘贴至对应报告中；
5. 通过VI编辑器配置hive-site.xml文件，将MySQL数据库作为Hive元数据库。将配置文件“Hive元存储”相关内容复制并粘贴至对应报告中；
6. 初始化Hive元数据，将MySQL数据库JDBC驱动拷贝到Hive安装目录的lib文件夹下；并通过schematool命令执行初始化，将初始化结果复制粘贴至对应报告中；
7. 启动Hive并保存命令输出结果，将结果输出复制粘贴至对应报告中。

### 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,并将变量内容复制粘贴至对应报告中；
3. 修改并配置sqoop-env.sh文件，并将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，结果复制粘贴至对应报告中。

# 模块B：数据采集与处理（20分）

**项目背景说明**

1. 查看餐饮外送统计平台网站源码结构。
2. 打开网站，在网页中右键点击检查，或者F12快捷键，查看源码页面；
3. 检查网站：浏览网站源码查看所需内容。
4. 从餐饮外送统计平台中采集需要数据，按照要求使用Python语言编写代码工程，获取指定数据项，并对结果数据集进行必要的数据处理。请将符合任务要求的结果复制粘贴至对应报告中。

具体步骤如下：

1. 创建工程工程项目：C:\food\_delivery\_crawl
2. 构建采集请求
3. 按要求定义相关字段
4. 获取有效数据
5. 将获取到的数据保存到指定位置
6. 对数据集进行基础的数据处理

至此已从餐饮外送统计平台中获取所需数据，并完成了必要的基础的数据处理。

1. 每条数据记录请以单独一行保存，信息存储格式为key：value。文件保存路径为：C:\output。

示例：

{" rest\_id ": "\*\*\*", " rest\_name ": "\*\*\*",……}，

{" rest\_id ": "\*\*\*", " rest\_name ": "\*\*\*",……}，

……

### 任务一：爬取网页信息

自行创建Scrapy工程编写爬虫代码，爬取“配送平台数据”与“店铺运营数据”页面相关数据，通过爬虫代码分页爬取，将使用re解析分页链接的程序源代码复制粘贴至对应报告中。

### 任务二：爬取指定文件

运行代码，爬取“配送平台数据”与“店铺运营数据”页面相关数据分别至range\_audited.json、grey\_value.json文件。查看文件并填写采集到的记录行数，并将答案复制粘贴至对应报告中。

示例格式：

range\_audited.json行数为：

grey\_value.json行数为：

### 任务三：属性判断

审查爬取的range\_audited数据，判断属性“申请递交时间”与“申请时间”是否为重复属性。如果为重复属性，则删除“申请时间”，并输出前10条数据记录；如果不是重复属性，请输出数据集中数值不相同的记录条数。并将答案复制粘贴至对应报告中。

### 任务四：数据探索

针对爬取的grey\_value数据，利用DataFrame.describe方法探索数据基本情况，将输出结果复制粘贴至对应报告中。

### 任务五：缺失值统计

针对缺失值较多的属性“推单数-8日”，“有效完成率-8日”，“超时率-8日”，请分别计算下列任务，并将正确答案复制粘贴至对应报告中。

（1）属性“推单数-8日”空值记录条数为： ，中位数为： 。

（2）属性“有效完成率-8日”空值记录条数为： ，平均值为： 。

（3）属性“超时率-8日”空值记录条数为： ，平均值为： 。

### 任务六：缺失值处理

请根据任务5中计算的结果，对数据集中存在空值的字段使用均值进行填充。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。

# 模块C：数据清洗与挖掘分析（25分）

**项目背景说明**

餐饮外卖平台的核心价值体现在配送，而配送的价值则依赖于商家与客户的双向选择。外卖平台通常会通过内容激活消费者和商家两个群体的活跃度。消费者会参考平台展示的内容选择商家，商家也会以消费者评价与平台统计数据为依据调整策略，由此再吸引更多的用户下单、评论、形成正向循环。保证配送的时效与品质是从优化用户体验的角度，吸引更多的用户参与，进而带动商家不断入驻。由此，商家、消费者、骑手在平台上形成越来越多的真实可靠的数据，帮助消费者更好的做出消费决策，同时促进商家提高服务质量。而平台通过数据，不断调整优化服务，从而不断提升这种多边网络效应。提升网络效应的直接结果就是用户和商家规模大幅提升，进而形成规模效应——降低获客成本、提高效益，并且不断提升自己的行业壁垒。

为探索各大外卖平台的市场策略与经营模式，现已从平台获取到了原始数据集，包含“餐厅id，retailer\_id，餐厅名称，城市，商户业务包，配送范围，客单价，推单数，接单数，有效完成单数，投诉率，异常率，欺诈单数，拒单数，商户取消数，客户取消数，系统取消数，配送取消异常数，整体时长，接单时长，到店时长，取餐时长，送达时长，商户投诉数，用户投诉数，差评数，好评数，评价数，最远订单距离，该订单整体时效，该订单接单时效，该订单到店时效，该订单取餐时效，该订单送达时效，该订单评价”字段。为保障用户隐私和行业敏感信息，已经对数据脱敏。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一些商业性敏感数据的情况下，对真实数据进行改造并提供测试使用，如身份证号、手机号等个人敏感信息都需要进行数据脱敏。本题已将脱敏后的数据存放于平台对应任务/chinaskills目录下。工程所需配置文件pom.xml存放于“C:\清洗配置文件”。任务中所有命令务必使用绝对路径。

### 任务一：数据清洗

**子任务1**

**任务背景：**

数据源为众多网站及平台的数据汇总，且为多次采集的结果，在整合多来源数据时可能遇到数据冲突，或数据拼接导致的属性列矛盾等情况。请根据任务具体参数要求，针对原始数据集中不符合业务逻辑的属性列进行清洗，并写入指定的数据库或数据文件，复制并保存结果。

**任务描述：**

数据源文件存放于平台对应任务/chinaskills目录下，其中属性“推单数”是指外卖平台通过顾客点单向商家推送的订单数量，“接单数”为商家根据自身情况，最终选择接受订单的数量。一般来说，商家对于平台推送的订单，排除自身原因，例如原材料耗尽、用户下单时店铺已经打烊等特殊情况，都会选择接单。请按照如下要求编写Spark程序对数据进行清洗，并将结果输出/diliveryoutput1。

1. 分析/chinaskills中json数据文件
2. 针对属性列“推单数”、“接单数”，排查并删除异常数据条目
3. 程序打包并在Spark平台运行，结果输出至HDFS文件系统/diliveryoutput1

**具体任务要求：**

1、将相关文件上传至HDFS新建目录/platform\_data中，编写Spark程序，剔除属性列“推单数”小于“接单数”的异常数据条目，并在程序中以打印语句输出异常条数。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）。

示例格式：

===“推单数”小于“接单数”的异常数据条数为\*\*\*条===

2、程序打包并在Spark平台运行，将剔除异常数据后的结果数据集json文件输出至HDFS文件系统/diliveryoutput1（将分区数设置为1）。并查看输出文件前5条记录，将查看命令与执行结果复制粘贴至对应报告中。

**子任务2**

**任务背景：**

客单价是指客户在该商铺下一单的平均支付价格。根据商家定位不同，可以分为高客单价和低客单价。高客单价，单量一般表现平平；低单价则通常会获得更高的单量。不同的定价针对的消费人群不同、选择的位置不同、营业的时间也不同。高客单价的品类偏向于白领人群，一般说来办公楼覆盖越多的位置越好，但是办公楼并不一定都是白领人群，所以办公楼也要区分区域性，客户行业越是前沿的，具备消费能力越高，但晚上和周末的单量一般较少。低客单价的品类偏向于大众化，选址优先办公/大学区/小区综合覆盖区域，满足低消费与一般消费能力用户群。当前数据源因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。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、中位数、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，当缺失信息的记录数较少时可采用删除的方式来进行处理。下面请根据任务具体参数要求处理关键字段缺失。

**任务描述：**

请以前置任务的结果数据集/diliveryoutput1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。

1. 解析/diliveryoutput1 中的文件
2. 针对数据集“客单价”属性，审查缺失值数量
3. 当缺失值比例小于5**%**时，对包含缺失值数据的样本进行删除

当缺失值比例大于5%时，对缺失值字段进行中位数填充

1. 程序打包并在Spark平台运行，结果json文件输出至HDFS文件系统/diliveryoutput2

**具体任务要求：**

1、根据任务要求，编写Spark程序，针对数据集“客单价”属性，审查缺失值数量，并打印输出，将打印输出结果复制并粘贴至对应报告中（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）。

示例格式：

===“客单价”属性缺失记录为\*\*\*条，缺失比例\*\*%===

2、缺失值处理

a)当缺失比例小于5%时，对含缺失值数据记录进行删除，同时在对应答案报告中粘贴如下内容（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）：

示例格式：

===“客单价”缺失记录已删除===

b)当缺失比例大于5%时，利用“客单价”属性中位数对缺失值进行填充，并将中位数打印输出，将打印输出结果复制并粘贴至对应报告中（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）。

示例格式：

===“客单价”属性中位数为\*\*\*天===

3、将清洗后的数据集输出至/diliveryoutput2（将分区数设置为1），并查看输出文件前5条记录，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

**任务背景：**

聚类分析又称群分析，它是研究分类问题的一种统计分析方法，同时也是数据挖掘的一个重要算法。聚类分析是由若干模式组成的。通常，模式是一个度量的向量，或者是多维空间中的一个点。聚类分析以相似性为基础，同一个聚类簇中的模式之间具有相似性，不同聚类簇之间具有相异性。

在商业上，聚类可以帮助平台市场分析人员从数据中区分出不同的商家群体，并提取每一类商家的经营模式。它作为数据挖掘中的一个模块，可以作为一个单独的工具以发现数据中的深层的信息，并且提取出每一类样本的特点，或者把注意力放在某一个特定的类上以作进一步的分析；同时，聚类分析也可以作为数据挖掘算法中其他分析算法的一个预处理步骤。本题数据请采用数据清洗的输出文件/diliveryoutput2。工程所需配置文件pom.xml存放于“C:\分析与挖掘配置文件”。**任务中所有命令务必使用绝对路径。**

选择数据分析的维度通常分为用户维度、行为维度和产品维度，想要对外卖平台入驻商家进行聚类划分，有侧重地分类评估商家对平台的价值，首先需要针对商家数据选择核心数据集，为确保聚类模型收敛速度与质量，以及消除量纲对聚类结果的影响，首先应对数据进行归一化或标准化处理，再进行数据建模。

**任务描述：**

请以数据清洗任务结果数据集/diliveryoutput2作为输入数据源，按照如下要求编写 Spark 程序实现对数据的分析，并将结果输出至 HDFS 文件系统中/diliveryoutput3。

1. 解析/diliveryoutput2中的文件
2. 提取商家数据核心属性
3. 针对商家核心属性进行预处理
4. 利用处理过的核心属性集完成商家聚类

**具体任务要求：**

1、针对相关数据集抽取北京地区相关数据记录，并计算商家好评比，将“好评比”作为新属性添加至最后一列。将结果数据集json文件输出至/diliveryoutput3（将分区数设置为1），并查看输出文件前5条记录，将查看命令与执行结果复制粘贴至对应报告中。

【好评比计算公式：好评比=好评数/评价数】

2、根据/diliveryoutput3中输出的结果数据集，筛选4项核心属性集：“商户业务包”，“接单数”，“客单价”，“好评比”，数据记录以接单数降序排列。将结果数据集json文件输出至/diliveryoutput4（将分区数设置为1），并查看输出文件前10行，将查看命令与执行结果复制粘贴至对应报告中。

3、由于核心数据集中“商户业务包”为分类属性，请将该属性设置为哑变量；同时对属性“接单数”，“客单价”进行max-min归一化，以实现对核心属性的预处理。将处理后的结果数据集json文件以接单数降序排列，输出至/diliveryoutput5（将分区数设置为1），并查看输出文件前5条记录，将查看命令与执行结果复制粘贴至对应报告中。

【归一化公式：x' = (x - X\_min) / (X\_max - X\_min)】

4、对以上属性对商家进行k-means聚类，聚类数设为4，迭代次数为2000次，请以打印语句输出聚类中心，及每个类的商家数。

示例格式：

==cluster 0: 聚类中心为[\*\*\*\*]=商家数为\*\*\*个===

==cluster 1: 聚类中心为[\*\*\*\*]=商家数为\*\*\*个===

……

# 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、地点、商家id、网格id、餐品种类、标品属性等多项基础信息字段。请使用Flask框架，结合Echarts完成下列任务。

数据库账号: root 密码：123456

自行创建代码工程路径为C:\food\_delivery

### 任务一：气泡图呈现商家数量

**任务背景：**

商圈，指某商场以其所在地为原点，沿着一定的方向和距离扩展，吸引顾客的辐射范围。简单地说，就是来店顾客所居住或工作的区域范围。无论餐厅规模大小，其销售覆盖区域总是有一定的地理范围。这个地理范围就是以某商场为中心，向四周辐射到可能来店消费的顾客居住地或工作地。请按任务指定要求，输出相关图例。

**任务描述：**

请根据数据库相关数据表中city\_name，location，latitude，longitude，rest\_type，platform\_A\_restid，A\_rst\_name，A\_day\_30\_cnt，platform\_B\_restid，B\_rst\_name，B\_day\_30\_cnt等字段，明晰地理位置与商铺聚集程度之间的关系。请以经度为横坐标，纬度为纵坐标，绘制商家数量气泡图，并以该地理位置的商家数量/10作为气泡半径。

**具体任务要求：**

1、提取数据表相关字段，在控制台按照“商家数量”降序排列，打印输出商圈名称及包含的商家数量。

示例格式：

==1: 商圈 \*\*\*\*=商家数为\*\*\*个===

==2: 商圈 \*\*\*\*=商家数为\*\*\*个===

……

2、使用Flask框架，结合Echarts，完成气泡图输出。要求气泡图标题为“商家聚集地理位置展示”，横坐标为经度，纵坐标为维度，以该地理位置的商家数量/10作为气泡半径，绘制气泡图。将可视化结果截图并保存（截图需包含浏览器地址栏）。

### 任务二：双折线图呈现市场占有率

**任务背景：**

市场份额亦称“市场占有率”。指某企业的销售量（或销售额）在市场同类品类中所占比重。反映企业在市场上的地位。通常市场份额越高，竞争力越强。市场占有率一般有3种基本测算方法：（1)总体市场份额，指某企业销售量在整个行业中所占比重。（2)目标市场份额，指某企业销售量在其目标市场，即其所服务的市场中所占比重。（3)相对市场份额，指某企业销售量与市场上最大竞争者销售量之比，若高于1，表明该企业其为这一市场的领导者。请按任务指定要求，输出不同平台商家销售分析相关图例。

**任务描述：**

请根据相关数据表中的数据，分别统计A平台与B平台30天销量最高的10个商家的销量，并以共享y轴的双折线图呈现。

**具体任务要求：**

1、根据相关数据表中city\_name，location，latitude，longitude，rest\_type，platform\_A\_restid，A\_rst\_name，A\_day\_30\_cnt，platform\_B\_restid，B\_rst\_name，B\_day\_30\_cnt等字段，分别统计A平台与B平台30天销量最高的10个商家及销量，在控制台按照“30天销量”降序打印输出商家id，商家所属平台，及30天销量。

示例格式：

==1: “\*\*\*\*”，Platform-A, 销量为\*\*\*===

==2: “\*\*\*\*”，Platform-A, 销量为\*\*\*===

……

==10: “\*\*\*\*”，Platform-A, 销量为\*\*\*===

==1: “\*\*\*\*”，Platform-B, 销量为\*\*\*===

==2: “\*\*\*\*”，Platform-B, 销量为\*\*\*===

……

==10: “\*\*\*\*”，Platform-B, 商家数为\*\*\*个===

2、使用Flask框架，结合Echarts，完成可视化输出。要求双折线图标题为“各平台30天销量最高10大商家”,横坐标1(下方)为平台A商家id，横坐标2(下方)为平台B商家id（倾斜显示、互不遮挡），纵坐标为商家销量，以销量降序排列，红色折线标识Platform-A商家，蓝色折线标识Platform-B商家。将可视化结果截图并保存（截图需包含浏览器地址栏）。

### 任务三：饼状图呈现标品属性

**任务背景：**

外卖平台的“标品属性”是指外卖的配送方式，一般说来分为众包、专送、自营销三种方式。其中，专送的稳定性最高，专送骑手是配送商的正式员工，且接单模式是平台派单，所以恶劣天气也很少会出现无人接单的尴尬情况。众包则是兼职配送员，以抢单的方式接单配送，所以一些“难送”或者恶劣天气时的订单经常会出现无人接单的情况。这将导致商家大量订单积压送不出去，从而被迫关店。从稳定性上来讲，专送要远远优于众包和快送。但专送的配送范围相对较小，在2km左右，对于一些快餐品类来说更为合适，但是对于一些高客单价的细分品类来说，2km的配送范围很难拓展单量。请根据相关数据集，按任务指定要求，输出配送方式相关分析图例。

**任务描述：**

请根据数据库相关数据表中的数据，统计不同标品属性的商家数量，并以饼状图表达。

**具体任务要求：**

1、根据相关数据表的网格ID，网格名称,城市,战团,餐厅ID，近7天平台单量，近7天推单，餐厅名，餐厅地址（取餐地址），餐品种类,标品属性，全推/选推等字段，统计不同标品属性的商家数量。请将标品属性“专送KA”合并至“专送”中，“众包平台”合并至“众包”中。在控制台**降序**打印输出标品属性，商家数量。

打印语句格式如下：

==专送: 商家\*\*\*个===

==众包: 商家\*\*\*个===

……

2、使用Flask框架，结合Echarts，完成饼图输出。要求饼图标题为“标品属性占比”，并标识该分类占比。将可视化结果截图并保存（截图需包含浏览器地址栏）。

### 任务四：组图（条形图、玫瑰饼图）呈现战团数量

**任务背景：**

外卖平台对商家业务的拓展主要由BD负责，BD(Bussiness Developer)即为业务拓展经理，工作职责是根据公司的业务战略制定具体的战略实施计划，业务推广计划，商家上架计划的达成跟进及分析，提供订单销售分析报告。同时负责调查区域品牌发展状况，调研竞品，为公司品牌发展提出相应建议。BD的上一层战略组织为战营、战团。一般指在更大的地域范围内组成的销售部门，存在共同的业务发展愿景和明确的绩效考核指标。请根据相关数据集，按任务指定要求，输出相关分析图例。

**任务描述：**

请根据相关数据表中的数据，统计不同战团商家数量以及不同餐品品类占比，并以组图表达。

**具体任务要求：**

1、请根据数据表中网格ID，网格名称，城市，战团，餐厅ID，近7天平台单量，近7天推单，餐厅名，餐厅地址（取餐地址），餐品种类，标品属性，全推/选推等字段字段，统计不同战团的商家数量。请在控制台以上架数量**降序**打印输出战团名称，商家数量。

打印语句格式如下：

==战团: \*\*\*，商家\*\*\*个===

==战团: \*\*\*，商家\*\*\*个===

……

1、统计包含商家数量最多的战团中，不同餐品品类的占比。请汇总前4位餐品品类占比，其他品类请归并为“其他”。请在控制台以占比数值降序打印输出餐品品类，商家数。

打印语句格式如下：

==小吃夜宵，商家\*\*\*个===

==特色菜系，商家\*\*\*个===

……

2、使用Flask框架，结合Echarts，完成组图输出。要求组图左侧输出1）小题数据条形图，标题为“Top10战团商家数量”，纵坐标为战团名，横坐标为包含商家个数，并在柱形上方标识商家数；组图右侧绘制2）小题数据的玫瑰饼图表达并标识各分类占比，标题为“\*\*战团餐品品类占比”。将可视化结果截图并保存（截图需包含浏览器地址栏）。

### 任务五：柱状图呈现推单差值

**任务背景：**

网格化营销是近年来新兴的一种营销管理模式，已被广泛地运用于市场精准营销。在市场营销中，采取地图营销、网格管理、精准策略等，可将客户的心理需求与其日常生活紧密地联系起来，巧妙运用网格化管理和营销地图，以此来实现客户的营销精准度与价值提升，提高单一客户贡献值和营销效率。其核心旨在帮助企业快速响应市场需求，为用户提供更加便捷的、专业化服务，并使企业的资源分配以客户为中心，以市场变化为导向，固本强基、开 拓市场、提升效益，有助于平台有效地制定绩效战略。请根据相关数据集，按任务指定要求，输出相关分析图例。

**任务描述：**

请根据数据库相关数据表中的数据，统计不同标品属性的商家数量，并以饼状图表达。

**具体任务要求：**

1、请根据数据表中网格ID，网格名称，城市，战团，餐厅ID，近7天平台单量，近7天推单，餐厅名，餐厅地址（取餐地址），餐品种类，标品属性，全推/选推等字段，统计不同网格“近7天平台单量”，与“7天推单”的差值。在控制台以差值**降序**打印输出前10位网格id，网格名称，推单差值。

打印语句格式如下：

==1 网格id：1111，网格名称：\*\*\*\*，推单差值为\*\*===

==2 网格id：1111，网格名称：\*\*\*\*，推单差值为\*\*===

……

2、使用Flask框架，结合Echarts，完成柱状图输出。以柱状图输出推单差值前10位的网格名称及差值，其中标题为“推单差值Top10”，横坐标为网格名称（倾斜显示，互补遮挡），纵坐标为推单差值。将可视化结果截图并保存（截图需包含浏览器地址栏）。

### 任务六：组图（玫瑰饼图、柱状图）呈现投诉比例

**任务背景：**

投诉是顾客对平台管理和服务不满的表达方式，也是企业有价值的信息来源，它为企业探索更多可能。分析顾客投诉的种种因素，把顾客的不满转化满意，锁定他们对平台和产品的忠诚，已成为企业营销实践的重要内容之一。商户业务包是指该商户在经营定位时所确定的主要消费群体。不同城市消费群占比不同，大致可以分为大客户，白领，小客户，高校， 家庭，其它六个商业业务包。请根据相关数据集，按任务指定要求，统计并输出投诉相关的分析图例。

**任务描述：**

请根据相关数据库数据表中“餐厅id，retailer\_id，餐厅名称，城市，商户业务包属性，配送范围，客单价，推单数，接单数，有效完成单数，投诉率，异常率，欺诈单数，拒单数，商户取消数，客户取消数，系统取消数，配送取消异常数，整体时长，接单时长，到店时长，取餐时长，送达时长，商户投诉数，用户投诉数，差评数，好评数，评价数，最远订单距离，该订单整体时效，该订单接单时效，该订单到店时效，该订单取餐时效，该订单送达时效，该订单评价”等字段，统计北京地区，不同“商户业务包”的餐厅数量及投诉数量，并以组图呈现。程序输出及可视化输出时请使用商户业务包中文释义，中文释义对应关系如表1所示。

表1 商户业务包中文释义

|  |  |
| --- | --- |
| **字母缩写** | **中文释义** |
| GKA | 大客户 |
| BL | 白领 |
| SIG | 小客户 |
| GX | 高效 |
| FML | 家庭 |
| OTH | 其他 |

**具体任务要求：**

1、根据相关数据库数据表中的字段，统计北京地区不同商户业务包属性商户数量及投诉数量，并在PyCharm控制台打印输出，请以投诉数量降序排列。

打印语句格式如下：

== 1.商户业务包：\*\*\*, 商家数量：\*\*\*家，投诉数\*\*\*条===

== 2.商户业务包：\*\*\*, 商家数量：\*\*\*家，投诉数\*\*\*条===

……

2、使用Flask框架，结合Echarts，完成组图输出。请在左侧画出以不同商户业务包属性的商家数量降序排列的柱状图，标题为“北京商户业务包商家数”，横坐标位商户业务包名称，纵坐标位商家数量；请在右侧画出不同商户业务包商家的投诉占比玫瑰图，标题为“北京商户业务包投诉占比”，顺时针显示次序与打印语句数据一致，将可视化结果截图并保存（截图需包含浏览器地址栏）。

# 模块E：综合分析（15分）

通过模块B的网站分析及数据爬取、模块C的数据清洗与挖掘分析及模块D的数据可视化呈现，我们已经清晰的了解了餐饮外卖平台业务背景及相关数据，在综合理解外卖业务数据的基础上，根据任务要求进行分析，并编写分析报告。

请根据任务要求，分析以下内容，并编写分析报告。分别从商家价值聚类、推单差值等维度对外卖平台推广情况与网格销售表现进行分析，并平台经营提出几点建议。

分析报告要求：

### 任务一：商家聚类分析

结合平台相关数据文件，以雷达图表示四类商家在核心属性集上的聚类表现。说明商家聚类对平台发展的用途及经营策略影响，分别以文字描述和图例进行说明。

### 任务二：推单差值分析

结合模块D可视化分析中对不同网格推单量差值的统计结果，说明差值产生的原因可能有哪些？对于缩小推单差值，你有哪些建议？分别以文字描述和图例进行说明。

### 任务三：平台建议

请结合平台业务背景及相关分析结论，对平台未来规划提出建议（不少于3条建议）。

### 附录：补充说明

数据集中涉及字段及中文说明

|  |  |
| --- | --- |
| 原字段 | 中文释义 |
| city\_name | 城市 |
| location | 商圈 |
| latitude | 纬度 |
| longitude | 经度 |
| rest\_type | 商家所属平台 |
| platform\_A\_restid | 商家id（A平台） |
| A\_rst\_name | 店铺名称（A平台） |
| A\_day\_30\_cnt | 30天销量（A平台） |
| platform\_B\_restid | 商家id（B平台） |
| B\_rst\_name | 店铺名称（B平台） |
| B\_day\_30\_cnt | 30天销量（B平台） |
| 推单数-9 | 推单数（9日） |
| 有效完成率-9 | 有效完成率（9日） |
| 超时率-9 | 超时率（9日） |
| 推单数-8 | 推单数（8日） |
| 有效完成率-8 | 有效完成率（8日） |
| 超时率-8 | 超时率（8日） |
| 灰度餐厅 | 灰度餐厅 |
| Id | 标识id |
| Request\_id | 请求id |
| Walle\_id | （平台）商铺id |
| Retailer\_id | （配送系统）商铺id |
| retailer\_name | 商铺名称 |
| retailer\_address | 商铺地址 |
| retailer\_location | 位置POI编码 |
| City\_id | 城市id |
| City\_name | 城市名称 |
| Grid\_id | 网格id |
| Carrier\_id | 渠道经理 |
| Team\_id | 渠道小组 |
| Applicant\_id | 申请人id |
| Applicant\_name | 申请人 |
| first\_auditor\_role | 一级审批角色 |
| first\_auditor\_candidate\_ids | 候选审批人id |
| first\_auditor\_id | 审核人id |
| first\_auditor\_name | 审批人名 |
| second\_auditor\_role | 二级审批角色 |
| second\_auditor\_candidate\_ids | 审核小组成员 |
| second\_auditor\_id | 二级审核人id |
| second\_auditor\_name | 二级审核人名 |
| status | 申请状态 |
| max\_distance\_before\_edit | 申请前最大配送距离 |
| min\_distance\_before\_edit | 申请前最小配送距离 |
| max\_distance\_after\_edit | 申请后最大配送距离 |
| min\_distance\_after\_edit | 申请后最小配送距离 |
| area\_before\_edit | 申请前配送面积 |
| area\_after\_edit | 申请前配送面积 |
| created\_at | 申请递交时间 |
| updated\_at | 审核完成 |
| 申请时间 | 申请时间 |
| 创建时间 | 创建时间 |
| 餐厅id | 餐厅id |
| 餐厅名称 | 餐厅名称 |
| 所属城市 | 所属城市 |
| 营业时长 | 营业时长 |
| 餐厅状态 | 餐厅状态 |
| 是否托管 | 是否托管 |
| 总单量 | 总单量 |
| 总单量增长率 | 总单量增长率 |
| 有效订单量 | 有效订单量 |
| 有效订单增长率 | 有效订单增长率 |
| 订单配送成功率 | 订单配送成功率 |
| 超时订单率 | 超时订单率 |
| 无效订单率 | 无效订单率 |
| 平均预计送达时长 | 平均预计送达时长 |
| 平均实际配送时长 | 平均实际配送时长 |
| 当前配送面积 | 当前配送面积 |
| 面积变更值 | 面积变更值 |
| 网格id | 网格id |
| 网格名称 | 网格名称 |
| 战营 | 战营 |
| 餐厅id | 餐厅id |
| retailer\_id | retailer\_id |
| 餐厅名称 | 餐厅名称 |
| 城市 | 城市 |
| 商户业务包 | 商户业务包 |
| 配送范围 | 配送范围 |
| 客单价 | 客单价 |
| 推单数 | 推单数 |
| 接单数 | 接单数 |
| 有效完成单数 | 有效完成单数 |
| 投诉率 | 投诉率 |
| 异常率 | 异常率 |
| 欺诈单数 | 欺诈单数 |
| 拒单数 | 拒单数 |
| 商户取消数 | 商户取消数 |
| 客户取消数 | 客户取消数 |
| 系统取消数 | 系统取消数 |
| 配送取消异常数 | 配送取消异常数 |
| 整体时长 | 整体时长 |
| 接单时长 | 接单时长 |
| 到店时长 | 到店时长 |
| 取餐时长 | 取餐时长 |
| 送达时长 | 送达时长 |
| 商户投诉数 | 商户投诉数 |
| 用户投诉数 | 用户投诉数 |
| 差评数 | 差评数 |
| 好评数 | 好评数 |
| 评价数 | 评价数 |
| 最远订单距离 | 最远订单距离 |
| 该订单整体时效 | 该订单整体时效 |
| 该订单接单时效 | 该订单接单时效 |
| 该订单到店时效 | 该订单到店时效 |
| 该订单取餐时效 | 该订单取餐时效 |
| 该订单送达时效 | 该订单送达时效 |
| 该订单评价 | 该订单评价 |
| 网格ID | 网格ID |
| 网格名称 | 网格名称 |
| 城市 | 城市 |
| 战团 | 战团 |
| 餐厅ID | 餐厅ID |
| 近7天平台单量 | 近7天平台单量 |
| 近7天推单 | 近7天推单 |
| 餐厅名 | 餐厅名 |
| 餐厅地址（取餐地址） | 餐厅地址（取餐地址） |
| 餐品种类 | 餐品种类 |
| 标品属性 | 标品属性 |
| 全推/选推 | 全推/选推 |