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-2021041-03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 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。命令中要求使用绝对路径，具体部署要求如下:

1. 将/chinaskills下的JDK包解压到/usr/local/src路径，将完整命令复制粘贴到对应报告中;
2. 修改/root/.bash\_profile文件，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复制上面步骤配置的JDK环境变量文件到slave1、slave2节点，命令和结果复制粘贴在对应报告中;
4. 配置SSH密钥登录，实现从master登录到slave1，将登录命令和结果复制粘贴在对应报告中;
5. 将配置文件hadoop-env.sh变更内容复制粘贴在对应报告中；
6. 将配置文件core-site.xml变更内容复制粘贴在对应报告中；
7. 初始化Hadoop环境namenode，将命令及结果复制粘贴在对应报告中；
8. 查看master及slave1节点jps进程，将其命令及结果复制粘贴至对应报告中。

### 任务二：Sqoop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/h3cu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,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，将命令及结果复制粘贴至对应报告中；
4. 测试Sqoop连接MySQL数据库是否成功，将命令及结果复制粘贴至对应报告中。具体任务要求：使用sqoop命令连接MySQL数据库，查询MySQL中所有数据库名称，将命令及结果复制粘贴至对应报告中。

### 任务三：Spark组件部署管理

1. 解压scala安装包到“/usr/local/src”路径下，并更名为scala，命令及结果复制粘贴至对应报告中；
2. 设置scala环境变量，并使环境变量只对root用户生效，配置文件内容复制粘贴至对应报告中；
3. 进入scala命令行界面，将命令及结果复制粘贴至对应报告中；
4. 解压Spark安装包到“usr/local/src”路径下，并更名为spark，将命令及结果复制粘贴至对应报告中；
5. 设置Spark相关环境变量，并使环境变量只对root用户生效，配置Spark的master节点主机名、端口、worker结点的核数、内存，将命令复制粘贴至对应报告中；
6. 启动Spark shell，将命令及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. 自行创建Scrapy工程项目food\_delivery\_crawl，路径为C:\ food\_delivery\_crawl按照任务要求从餐饮外送统计平台中获取数据。
2. 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：爬取店铺基础信息

自行创建Scrapy工程编写爬虫代码，爬取“店铺基础信息”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：爬取店铺运营数据

编写爬虫代码，爬取“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。

### 任务三：爬取指定文件

运行代码，爬取“店铺基础信息”页面与“店铺运营数据”页面数据分别至POI.json与grey\_test.json文件。查看文件并填写采集到的记录条数，并将答案复制粘贴至对应报告中。

示例格式：

POI.json行数为：

grey\_test.json行数为：

### 任务四：各商家“30天销量”数据处理

POI.json文件中，“商家id（A平台）”字段值重复的样本，请以多条样本记录的“30天销量（A平台）”属性均值作为该属性的值，并删除多余样本。请在PyCharm控制台打印输出删除的样本条数，并将打印语句复制粘贴至对应报告中。

示例格式：

===因重复样本记录，删除样本条数为\*\*\*条===

### 任务五：grey\_test缺失值统计

针对爬取的grey\_test数据，利用missingno库方法探索数据集各字段数据缺失情况，将可视化图片复制粘贴至对应报告中。

### 任务六：属性缺失统计

根据可视化输出，grey\_test数据中那一个属性缺失值最多？请将正确答案复制粘贴至对应报告中。

示例格式：

缺失值最多的属性为：

属性缺失的样本数为：

# 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“C:\数据源”。

### 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“C:\数据源”，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Python程序对数据进行清洗，结果文件保存至路径“C:\输出结果”文件夹中，并命名为diliveryoutput1.json。

1. 分析“C:\数据源”中数据源json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. 运行Python程序，结果文件保存至路径“C:\输出结果”，并命名为diliveryoutput1.json。

**具体任务要求**：

1、读取“C:\数据源”中相关数据源json文件，在PyCharm控制台打印输出数据集样本条数，将打印语句复制粘贴至对应报告中。编写Python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。

示例格式：

===数据集初始样本条数为\*\*\*条===

===“推单数”小于“接单数”的异常数据条数为\*\*\*条===

2、运行Python程序，将剔除异常数据后的结果数据集保存至路径“C:\输出结果”，并命名为diliveryoutput1.json。查看结果数据集前10行数据记录，将查看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：**

请以上述1、（任务数字编号）任务结果数据集“C:\数据源\diliveryoutput1.json”作为数据源，编写Python程序，按照如下要求实现对数据的清洗，并将结果数据集保存至路径“C:\输出结果“中，并命名为diliveryoutput2. json。

1. 解析文件diliveryoutput1.json；
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. 运行Python程序，将结果数据集保存至路径“C:\输出结果“中，并命名为diliveryoutput2.json。

**具体任务要求**：

1、根据任务要求，编写Python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。

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：===“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。

示例格式：===“客单价”属性中位数为\*\*\*天===

c)清洗后的数据集，保存至路径“C:\输出结果“中，并命名为diliveryoutput2.json，在PyCharm中查看输出文件的前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“C:\输出结果\diliveryoutput2.csv”。

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“C:\输出结果\diliveryoutput2.json”作为输入数据源，按照如下要求编写Python程序实现对数据的分析，并将结果保存至路径“C:\输出结果“中，并命名为diliveryoutput3.json。

1. 解析文件diliveryoutput2.json。
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。将结果数据集保存至路径“C:\输出结果“中，并命名为diliveryoutput3.json，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、根据结果数据集diliveryoutput3.json，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。将结果数据集保存至路径“C:\输出结果“中，并命名为diliveryoutput4.json，并查看结果数据集前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以接单数降序排列，输出至C:\输出结果/diliveryoutput5.json，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。

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

4、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：条形图呈现“Top10城市商家数量”

**任务背景：**

近年来，我国外卖行业发展迅速，互联网餐饮外卖市场规模逐渐扩大，外卖平台已覆盖全国所有省份。2019全年外卖交易超7274亿元，截至2020年3月，我国网上外卖及手机网上外卖用户渗透率已达44%。2020年底，全国外卖总体订单量已超过171.2亿单。为了解外卖平台发展较为领先的城市范围，请根据相关数据集，按任务指定要求，输出分析图例。

**任务描述：**

请根据数据库中相关数据表中的数据，统计不同城市商家数量，并条形图呈现。

**具体任务要求：**

1、根据相关数据表中“id，Request\_id，Walle\_id，Retailer\_id，retailer\_name，retailer\_address，etailer\_location，City\_id，City\_name，Grid\_id，Carrier\_id，Team\_id，Applicant\_id，Applicant\_name，first\_auditor\_role，first\_auditor\_candidate\_ids，first\_auditor\_id，first\_auditor\_name，second\_auditor\_role，second\_auditor\_candidate\_ids，second\_auditor\_id，second\_auditor\_name，status，max\_distance\_before\_edit，min\_distance\_before\_edit，max\_distance\_after\_edit，min\_distance\_after\_edit，area\_before\_edit，area\_after\_edit，created\_at，updated\_at，申请时间，创建时间”等字段，统计不同城市商家数量。请在PyCharm控制台打印输出商家数量最高的前10个城市名称及商家数量，并以商家数量降序排列。

示例格式：

==1.城市: \*\*\*，商家\*\*\*个===

==2.城市: \*\*\*，商家\*\*\*个===

……

2、使用Flask框架，结合Echarts，条形图输出。标题为“城市商家数量Top10”，横坐标为商家数量，纵坐标为城市名称。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：柱状图呈现审核效率

**任务背景：**

外卖的配送范围一般由外卖平台的业务人员为商家设定，商家也可以根据实际情况向平台申请修改配送范围。在申请配送范围修改时，审核员将考虑实际道路限制、综合运力成本、顾客等待时长等因素，对配送范围变更的申请进行考量，最终做出通过/拒绝/驳回的结果判定。一级审核员通常为该地区的渠道经理，针对运力等综合因素对审核记录进行考量；二级审核员为外卖平台专职审核人员，将综合考虑配送范围的合理性及更改后的用户体验。考核员的审核具有一定的时效要求，审核效率直接影响合作商家平台活跃性以及商家的合作体验。请根据相关数据集，按任务指定要求，统计并输出审核效率相关的分析图例。

**任务描述：**

请根据相关数据表中的数据，参考数据字段同可视化1、（数字任务编号），统计审核效率最高的十位明星审核员，并以柱状图表达。

**具体任务要求：**

1、根据可视化1中提到的相关字段，汇总各二级审核人审批的申请数量。在PyCharm控制台打印输出审核记录数最多的10位审核人id，审批总数。

示例格式：

==1.二级审核人id:\*\*\*, 审批数量：\*\*\*条===

==2.二级审核人id:\*\*\*, 审批数量：\*\*\*条===

……

2、使用Flask框架，结合Echarts，完成柱状图输出。要求标题为“Top10明星审核员”，横坐标为审核人姓名，纵坐标为审核记录数，以审批记录数降序排列。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：双饼图呈现业务包属性占比

**任务背景：**

外卖平台在不同城市进行推广时应具有不同的侧重方面，城市具有不同的人口数量及特点，例如居住人口数量、工作人口数量、过往人口数量、居民户数和企事业单位数，及相应人口年龄、性别、职业和收入水平构成等。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，统计并输出城市商户业务包相关的分析图例。

**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，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，统计不同城市的商业业务包组成，并以双饼图表达。

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。

|  |  |
| --- | --- |
| **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 |

表1 商户业务包中文释义

**具体任务要求：**

1、根据相关数据库数据表中的字段，分析北京与深圳两座城市在商户业务包属性占比方面的差异。在PyCharm控制台打印输出两座城市不同商户业务包属性的商家数量，以商家数量降序排列。

示例格式：

==北京：1.商户业务包：\*\*\*, 商家数量：\*\*\*家===

==北京：2.商户业务包：\*\*\*, 商家数量：\*\*\*家===

……

==深圳：1.商户业务包：\*\*\*, 商家数量：\*\*\*家===

==深圳：2.商户业务包：\*\*\*, 商家数量：\*\*\*家===

……

2、使用Flask框架，结合Echarts，完成双饼图输出。要求标题分别为“北京商户业务包属性占比”（左）、“深圳商户业务包属性占比”（右），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）。

### 任务四：玫瑰图呈现投诉占比

**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。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的数据，参考数据字段同可视化3、（数字任务编号），统计北京地区，不同“商户业务包”的投诉数量，并以玫瑰图呈现。

**具体任务要求：**

1、根据相关数据库表格字段，统计北京地区不同商户业务包投诉数量。在PyCharm控制台打印输出不同商户业务包的投诉数量，以投诉数降序排列。

示例格式：

==商户业务包：\*\*\*, 投诉数量：\*\*\*家===

==商户业务包：\*\*\*, 投诉数量：\*\*\*家===

……

2、使用Flask框架，结合Echarts，完成玫瑰图输出。要求标题为“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）。

### 任务五：柱状堆叠图呈现审核结果比例

**任务背景：**

外卖的配送范围一般由外卖平台的业务人员为商家设定，商家也可以根据实际情况向平台申请修改配送范围。通常配送范围并不是毫无棱角的圆形，因为这种划分并没有因地制宜，是不合理的。最大配送距离指的是配送范围中距离商家最远的直线距离，最小配送距离则是配送范围中距离商家最近的直线距离。在申请配送范围修改时，审核员将考虑实际道路限制、综合运力成本、顾客等待时长等因素，对配送范围变更的申请进行考量，最终做出通过/拒绝/驳回的结果判定。请根据相关数据集，按任务指定要求，输出范围审核相关分析图例。

**任务描述：**

请根据数据库相关数据表中的数据，统计审核通过与拒绝的范围申请记录数量，并以柱状堆叠图表达。

**具体任务要求：**

1、根据相关表格中“id，Request\_id，Walle\_id，Retailer\_id，retailer\_name，retailer\_address，etailer\_location，City\_id，City\_name，Grid\_id，Carrier\_id，Team\_id，Applicant\_id，Applicant\_name，first\_auditor\_role，first\_auditor\_candidate\_ids，first\_auditor\_id，first\_auditor\_name，second\_auditor\_role，second\_auditor\_candidate\_ids，second\_auditor\_id，second\_auditor\_name，status，max\_distance\_before\_edit，min\_distance\_before\_edit，max\_distance\_after\_edit，min\_distance\_after\_edit，area\_before\_edit，area\_after\_edit，created\_at，updated\_at，申请时间，创建时间”等字段，汇总各二级审核人审批的申请数量，并统计其中通过与拒绝的条数。在PyCharm控制台打印输出审核记录数最多的10位审核人id，审批总数，以及其中通过与拒绝的条数，按审批总数降序排列。

示例格式：

==1.二级审核人id:\*\*\*, 审批数量：\*\*\*条，其中通过\*\*条，拒绝\*\*条===

==2.二级审核人id:\*\*\*, 审批数量：\*\*\*条，其中通过\*\*条，拒绝\*\*条===

……

2、使用Flask框架，结合Echarts，完成柱状堆叠图输出。要求 标题为“审核通过与拒绝对比”，横坐标为审核人id，纵坐标为审核记录数。画出审核数量最多的10位二级审核人记录审批情况，通过记录数蓝（蓝色表示）上方叠加拒绝和驳回记录数（红色表示），以审批总数降序排列。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：散点地图呈现各城市商家数量分布

**任务背景：**

近年来，我国外卖行业发展迅速，互联网餐饮外卖市场规模逐渐扩大，外卖平台已覆盖全国所有省份。2019全年外卖交易超7274亿元，截至2020年3月，我国网上外卖及手机网上外卖用户渗透率已达44%。2020年底，全国外卖总体订单量已超过171.2亿单。为纵观全国范围内外卖平台签约商家分布情况，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，统计不同城市商家数量，并散点地图呈现。

**具体任务要求：**

1、根据相关数据表中的数据，参考字段同可视化3、（任务数字编号），统计不同城市商家数量。请在PyCharm控制台以商家数量**降序**打印输出城市名称，商家数量。

示例格式：

==1.城市: \*\*\*，商家\*\*\*个===

==2.城市: \*\*\*，商家\*\*\*个===

……

2、使用Flask框架，结合Echarts，完成散点地图输出。标题为“各城市商家数量散点地图”。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的数据清洗与分析及模块D的可视化呈现，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4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：各地区商户业务包分析

结合模块D可视化分析中对不同地区商户业务包的统计结果，说明图表有哪些差异？说明了什么问题？分别以文字描述和图例进行说明。

### 任务三：平台规划建议

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：

poi： point of interest 兴趣点

数据表字段中文释义：

|  |  |
| --- | --- |
| 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 |