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-2021041-05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。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：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。

### 任务三：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，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。

## 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：获取“配送平台”页面相关数据

自行创建Scrapy工程编写工程代码，获取“配送平台”页面相关数据，通过工程代码分页获取，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：获取“店铺运营数据”页面相关数据

编写工程代码，获取“店铺运营数据”页面相关数据，通过工程代码分页获取，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：将获取数据存入json文件

运行代码，爬取“配送平台”页面与“店铺运营数据”页面数据分别至restaurant\_data.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：**

restaurant\_data.json行数为：

grey\_test.json行数为：

### 任务四：删除异常样本

restaurant\_data.json文件中，请剔除属性列“客户取消数”小于“商户取消数”的异常数据条目，并在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：**

===因异常样本记录，删除样本条数为\*\*\*条===

### 任务五：缺失值处理

请以任务4的结果数据集作为数据源，针对“到店时间”属性，审查缺失值数量：

(a)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。

(b)当缺失值比例大于5%时，对缺失值字段进行中位数填充。

在PyCharm控制台打印输出结论，并按示例格式将结果复制粘贴至对应报告中。

**示例格式：**

===“到店时间”属性缺失值比例小于5%，删除样本条数为\*\*\*条===

或：

===“到店时间”属性缺失值比例大于5%，中位数为： ，填充样本条数为\*\*\*条===

### 任务六：划分数据集

请针对grey\_test.json数据集，根据属性“超时率(9日)”划分数据集grey\_test.json1（超时率(9日)小于等于5%），grey\_test.json2（超时率(9日)大于5%），并在PyCharm控制台打印输出两个数据集的数据记录条数，并按示例格式将结果复制粘贴至对应报告中。文件保存路径为：C:\output。

**示例格式：**

===grey\_test.json1包含样本条数为\*\*\*条===

===grey\_test.json2包含样本条数为\*\*\*条===

## 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，现已从及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

**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重复，或数据拼接导致的属性列缺失或冗余等情况。请根据任务具体参数要求，针对原始数据集进行清洗，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，请按照如下要求编写Spark程序对数据进行清洗，并将结果输出保存。

1. 分析/chinaskills中json数据文件；
2. 针对属性列“商家id”排查并删除异常数据条目；
3. 针对属性列缺失或冗余的样本进行删除；
4. 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、读取相关数据文件，包含“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， 查看文件包含数据记录数，并将答案复制并保存至对应报告中。

**示例格式：**

文件数据记录数为：

2、编写Spark程序读取任务1的数据源，针对属性列“platform\_A\_restid”或“platform\_B\_restid”重复的样本，请以多条样本记录的“A\_day\_30\_cnt”或“B\_day\_30\_cnt”属性均值作为该属性的值，并删除多余样本，并在程序中以打印语句输出删除的条数。请将结果数据集json文件输出至HDFS文件系统/diliveryoutput1中。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：**

===因重复样本记录，删除样本条数为\*\*\*条===

3、编写Spark程序读取任务2的结果数据集/diliveryoutput1，编写Spark(Standalone)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”===

4、将任务3处理后的数据记录以[latitude，longitude]降序排列，并将结果数据集json文件输出至HDFS文件系统/diliveryoutput2(将分区数设置为1)。查看结果数据文件前5条记录。将命令及输出结果复制并保存至对应报告中。

### 任务二：数据挖掘分析

**子任务1**

**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。请根据任务具体参数要求，针对相关数据集进行分析，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，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/chinaskills中相关数据源json文件
2. 按网格汇总餐厅总数、“近7天平台单量”、“近7天推单量”，及相关计算
3. 按照任务要求统计网格超时率
4. 程序打包并在Spark平台运行，输出并保存结果

**具体任务要求：**

1、读取相关数据文件，参考“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，按网格统计餐厅总数、“近7天平台单量”总数，“近7天推单量”总数并以餐厅总数降序排列，在程序中以打印语句输出餐厅总数前10位的网格相关信息。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：**

===网格id:\*\*\*，网格名称：\*\*\*，近7天平台总单量：\*\*\*，近7天推单总量：\*\*\*===

**子任务2**

**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结果数据集/diliveryoutput2中数据源。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中数据作为输入数据源，按照如下要求编写Spark程序实现对数据的分析，并保存结果。

1. 解析/diliveryoutput2数据源文件。
2. 提取商家相关属性数据。
3. 针对商家地理位置进行聚类划分。
4. 查看聚类结果。

**具体任务要求：**

1、编写Spark程序读取/diliveryoutput2数据源文件，筛选北京地区（城市为北京市）商家数据记录，并将结果数据集json文件输出至 HDFS 文件系统/diliveryoutput4中。请在程序中以打印语句输出筛选得到数据记录条数，并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）。

**示例格式如下：**

===北京地区样本条数为\*\*\*条===

2、请使用数据挖掘分析任务2的结果数据集/diliveryoutput4，编写Spark程序，根据北京地区商家的经纬度属性，对商家进行k-means聚类，聚类数设为5，迭代次数为2000次，请在程序中以打印语句输出聚类中心、每个类的商家数，以及该类所包含的商圈，并将打印语句复制粘贴至对应报告中。（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：**

=cluster 0: 聚类中心为[\*\*，\*\*]，商家数为\*\*\*个，包含商圈：【\*，\*，…】=

=cluster 1: 聚类中心为[\*\*，\*\*]，商家数为\*\*\*个，包含商圈：【\*，\*，…】=

……

3、请使用数据挖掘分析任务2的结果数据集/diliveryoutput4，结合数据挖掘分析任务3的结论，编写Spark程序，分别统计各聚类中在平台A与平台B上架的商家数量，以及同时商家两个平台的商家数量。在程序中以打印语句输出结果，并将打印语句复制粘贴至对应报告中。（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：**

=cluster 0：【\*，\*,…】，A平台商家数量：\*\*，B平台商家数量：\*\*，同时上架A、B平台的商家数量：\*\*=

=cluster 1：【\*，\*,…】，A平台商家数量：\*\*，B平台商家数量：\*\*，同时上架A、B平台的商家数量：\*\*=

……

## 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：双饼图呈现城市商户业务包占比

**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：**

请根据数据库相关数据表中“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）。

### 任务二：柱状堆叠图呈现商户业务部评价情况

**任务背景：**

在外卖平台中，用户的评论具有双向影响。不仅可以促进商家根据消费者评价调整自己的经营策略，也能够影响其他用户的下单率。相关数据显示，浏览评价的用户比只浏览门店商品的用户下单转化率平均高出25%。因此，商家的店铺评论对提高下单率是十分重要的。请根据相关数据集，按任务指定要求，统计并输出评价相关的分析图例。

**任务描述：**

请根据相关数据表中的数据，统计各商户业务包评价情况，并柱状堆叠图呈现。程序输出及可视化输出时请使用商户业务包中文释义。

**具体任务要求：**

1、根据相关数据表中的数据，参考“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，统计各商户业务包非好评数、好评数，并在PyCharm控制台以评价总数降序打印输出。

**示例格式：**

==1商户业务包：\*\*\*，非好评数：\*\*\*条，好评数：\*\*\*条===

==1商户业务包：\*\*\*，非好评数：\*\*\*条，好评数：\*\*\*条===

……

2、使用Flask框架，结合Echarts，完成柱状堆叠图输出。横坐标为商户业务包名称（中文释义），纵坐标为评论数量，标题为“各业务包评价组成”，柱图下方为好评数（红色），上方堆叠非好评数（黑色），展示顺序与1）（任务数字编号）打印语句一致，将可视化结果截图并保存（截图需包含浏览器地址栏）。

### 任务三：玫瑰图地图呈现不同业务包投诉占比

**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“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、根据相关数据库数据表中的字段，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）。

### 任务四：气泡图呈现商铺分布的地理位置

**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所居住地或工作地。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集中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、提取数据表中相关字段，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）。

### 任务五：双折线图呈现不同平台商家销量情况

**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）。

### 任务六：柱状图呈现商家营业时间情况

**任务背景：**

营业时长指商户开市接单到闭市停业的时长，适当地增加营业时长有利于提升店铺的曝光量，同时能够在一定程度上获得平台排名加权。营业时间长的店铺在特定时间段（大部分商家关店休息后）面对的竞争压力将减小，但增加营业时间将增加商家运营店铺的人工成本，同时也需要考虑不同时间段内运力不同的情况，例如有些区域在晚上12点以后无法配送。请根据相关数据集，按任务指定要求，输出营业时间相关分析图例。

**任务描述：**

请根据相关数据表中的数据，统计平台各商家的营业时间占比，并柱状图呈现。

**具体任务要求：**

1、根据相关表格，参考“餐厅id，餐厅名称，所属城市，营业时长，餐厅状态，是否托管，总单量，总单量增长率，有效订单量，有效订单增长率，订单配送成功率，超时订单率，无效订单率，平均预计送达时长，平均实际配送时长，当前配送面积，面积变更值，网格id，网格名称，战营 ”等字段，统计4小时以内、4~6小时、6-8小时、8~12小时、12小时以上，5个区间段内平台各商家的营业时间占比（区间取值为前闭后开区间）。请在PyCharm控制台以区间商家数量**降序**打印输出城市名称，商家数量。

**示例格式：**

==区间“4小时以内”，商家\*\*\*个===

==区间“4~6小时”，商家\*\*\*个===

……

2、使用Flask框架，结合Echarts，完成柱状图输出。标题为“各营业区间商家数量对比”，横坐标为各区间，纵坐标为商家个数，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的网站分析及数据获取、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。分别从地区、营业时间角度对外卖平台推广情况与销售表现进行分析，并平台经营提出几点建议。

分析报告要求：

### 任务一：通过数据及图示分析不同战营销量数据分布情况

对于模块B获取的grey\_test.json数据，分别提取“KA销售一组、SIG华东二组、KA销售三组、SIG华南三组、SIG京津七组”五个战营全部的总单量属性值，绘制箱线图，观察数据分布情况。箱体与边框为蓝色，异常点使用蓝色圆点（边框与填充均为蓝色，size=3）,采用黑色圆点标识中位数（size=8），同时显示均值。分析箱线图表达的含义，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

结合平台相关数据文件，探索不同营业时长（区间划分请使用：4小时以内、4-6小时、6-8小时、8-12小时、12小时以上）对商家平均总单量的影响，并结合任务D相关可视化结果，说明影响总单量的因素有哪些？阐述你的结论并说明原因。分别以文字描述和图例进行说明。

### 任务三：对餐饮外卖平台未来拓展合作建议

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。

网格：网格化营销是指按照某一个或某几个特定的标准,利用各种有效的营销方式和工具,使企业营销资源重新组合和分配成网格化状态,即将整个实体市场分割成为多个网格。

灰度测试：在某项策略正式投入市场实践前，选择特定人群试用，逐步扩大其试用者数量，以便及时发现和纠正其中的问题。

评价级别定义：用户5星评价为好评，2星及以下为差评。

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 | 申请后配送面积 |
| 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 |