# 模块A：大数据平台部署管理

本模块用到的所有安装包均在Master节点的/chinaskills目录下。

### 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. 在三个节点的Zookeeper都安装配置完毕后，在每个节点启动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,请将dfs.ha.namenodes.hadoopcluster设置为nn1,nn2，在master节点启动Hadoop，并查看服务(nn1,nn2)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。

### 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-1.1.0-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。

### Spark(Standalone)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。

### 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。

# 模块B：数据采集与处理

**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:\crawl\_output。

示例：

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

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

……

1. 任务中要求将“以下内容及答案完整复制粘贴至对应报告中。”，粘贴到对应报告中的内容示例如下：

配送平台数据页数为：100

店铺运营数据页数为：100

### 任务一：爬取网页信息

1. 自行创建Scrapy工程编写爬虫代码，通过使用Chrome浏览器开发者工具进行抓包，将Scrapy 默认UserAgent 配置为Chrome浏览器UA，并将配置代码复制粘贴至对应报告中。
2. 将Scrapy 默认请求头的 Accept和Accept-Language 设置为抓包到的值，并将配置代码复制粘贴至对应报告中。
3. 在配置文件中配置Scrapy 的下载延时为2秒，并将配置代码复制粘贴至对应报告中。
4. 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，爬取网页数据至指定文件。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。

### 任务七：日志输出

将Scrapy日志（日志等级设置为INFO）以时间命名输出至路径：C:\ food\_delivery\_crawl\output,并将日志内容复制粘贴至对应报告中。

# 模块C：数据清洗与挖掘分析

**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“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，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Master节点/chinaskills目录下。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：**

相关数据源文件位于Master节点/chinaskills/wm\_5902.json，请参考列举的各项字段，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. 分析数据源文件
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(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”===

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条数据记录，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：**

数据的维数是指数据具有的特征数量，数据特征矩阵过大， 将导致计算量比较大，训练时间长的等问题。当数据存在冗余属性时，对多余属性剔除的过程，称为“数据降维”。降维的好处十分明显，它不仅可以数据减少对内存的占用，还能够加快学习算法的执行与收敛。请根据任务具体要求，针对原始数据集中可能存在的冗余属性进行排查，复制并保存结果。

**任务描述：**

请使用子任务1的结果数据作为数据源，判断属性“申请时间”、“创建时间”与“created\_at”、“updated\_at”是否为重复属性。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/diliveryoutput2(将分区数设置为1)。

1. 分析数据文件
2. 排除属性列“申请时间”、“创建时间”与“created\_at”、“updated\_at”是否为重复属性
3. 程序打包并在Spark平台运行

**具体任务要求**：

1、利用/diliveryoutput1作为源数据,使用hadoop shell命令查看数据集条数，请将查看命令及结果复制并粘贴至对应报告中。

2、如果仅考虑年、月、日数据，忽略时刻信息，“申请时间”、“创建时间”与“created\_at”、“updated\_at”是否为重复属性？请编写Spark程序，并在程序中以打印语句输出两对属性对应相等的数据条数。将打印输出结果以及你的结论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：

===两组属性同时相等的数据条数为\*\*\*条===

结论：“申请时间”、“创建时间”与“created\_at”、“updated\_at”（是/不是）重复属性。

如果数据相等的比例超过原始数据集的90%，则剔除属性“created\_at”与“updated\_at”， 并将结果数据集json文件保存至/diliveryoutput2。

### 任务二：数据挖掘分析

**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。下面请根据任务具体要求，分析范围审核相关数据。

**任务描述：**

请根据你的数据清洗的结论数据源（/diliveryoutput2），编写 Spark 程序，按照如下要求实现对数据的分析，并将结果输出至 HDFS 文件系统中。

1. 解析数据清洗任务结果数据集
2. 计算配送范围申请的审核时长
3. 分析配送范围变更幅度对审核结果的影响
4. 程序打包并在Spark平台运行，结果输出至HDFS文件系统/diliveryoutput3

**具体任务要求**：

1、从审核申请到审核完成的时间差为一条审核记录的处理时长，请编写Spark程序，计算全部申请记录的平均处理时长（单位为分钟，忽略“秒“级数值），并在程序中以打印语句输出结果。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：

===平均申请处理时长为\*\*\*分钟===

2、为保证用户体验，申请的审批应具有时效性。公司规定在当天18：00前递交的范围审核申请，处理时间应在12小时以内；在当天18：00后递交的范围审核申请，应在第二天中午12：00前审核完成。请根据这一标准，分别统计全部数据记录中，在18：00前/后递交的申请超时记录数，在并在程序中以打印语句输出结果。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：

===18：00前递交申请的记录中，超时记录数为\*\*\*条===

===18：00后递交申请的记录中，超时记录数为\*\*\*条===

3、申请前后配送范围的最大/最小距离标识了商家配送范围的变更幅度，请统计最大配送距离或最小配送距离在申请前后变更幅度超过以下百分比数值时：40%、50%、60%、70%、80%、90%，申请的通过率分别为多少（百分比数值保留小数点后2位）。请在程序中以打印语句输出结果，并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：

==申请后最大或最小配送距离超过原始距离的40%，审核通过率为\*\*\*% ==

==申请后最大或最小配送距离超过原始距离的50%，审核通过率为\*\*\*% ==

==申请后最大或最小配送距离超过原始距离的60%，审核通过率为\*\*\*% ==

==申请后最大或最小配送距离超过原始距离的70%，审核通过率为\*\*\*% ==

==申请后最大或最小配送距离超过原始距离的80%，审核通过率为\*\*\*% ==

==申请后最大或最小配送距离超过原始距离的90%，审核通过率为\*\*\*% ==

4、统计各商户的配送范围面积(请参考申请后配送面积)，在以下4个区间内的商户数及占比（百分比数值保留小数点后2位）：0~10,000,000（平方米），10,000,000~20,000,000（平方米），20,000,000~30,000,000（平方米），30,000,000（平方米）以上。请在程序中以打印语句输出结果，并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：

==配送面积0~1000,0,000（平方米），商家数为\*\*\*\*，占比\*\*\*% ==

==配送面积10,000,000~20,000,000（平方米），商家数为\*\*\*\*，占比\*\*\*% ==

==配送面积20,000,000~30,000,000（平方米），商家数为\*\*\*\*，占比\*\*\*% ==

==配送面积30,000,000（平方米）以上，商家数为\*\*\*\*，占比\*\*\*% ==

# 模块D：数据可视化

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 库名：takeout

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、提取表格相关字段，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===

……

1. 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：**

请根据相关表格数据，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、根据相关表格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：**

请根据数据库相关表格数据，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、根据相关表格网格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===

……

1. 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：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：**

请根据相关表格数据，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，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）。

### 任务五：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：**

请根据相关表格数据，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）。

### 任务六：散点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：**

请根据相关表格数据，统计不同城市商家数量，并散点图呈现。

**具体任务要求：**

1、根据相关表格，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

### 任务七：双柱状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：**

请根据数据库相关表格数据，统计审核通过与拒绝的范围申请记录数量，并以双柱图表达。

**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，完成双柱状图输出。要求柱图标题为“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，通过记录数以蓝色柱形表示，右侧以红色柱形表示拒绝和驳回记录数，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）。

### 任务八：饼图呈现营业区间商家数量

**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：**

请根据相关表格数据，统计平台各商家的营业时间占比，并饼图呈现。

**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，完成饼图输出。标题为“各营业区间商家数量对比”，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）。

### 任务九：柱状图呈现超时订单率

**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，输出分析图例。

**任务描述：**

请根据相关表格数据，统计平均超时订单率前10名的网格，并柱图呈现。

**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，统计平均超时订单率前10名的网格名称及平均超时订单率，并在PyCharm控制台以平均超时订单率降序打印输出。

示例格式：

==1.网格名称：\*\*\*，平均超时订单率\*\*%===

==2.网格名称：\*\*\*，平均超时订单率\*\*%===

……

2、使用Flask框架，结合Echarts，完成柱状图输出。标题为“超时订单率排名”，柱状图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）。

### 任务十：组图（多条形图）呈现各城市外卖行业现状

**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。为了解外卖行业在各城市发展情况，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：**

请根据相关表格数据，统计各城市外卖行业发展情况，并组图呈现。

**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，统计各城市商家总数前10名的城市名称、商家数量、以及该城市商家平均营业时长，并在PyCharm控制台以商家数量降序打印输出。

示例格式：

==1.城市：\*\*\*，商家数量：\*\*\*个，平均营业时长平\*\*\*分钟===

==2.城市：\*\*\*，商家数量：\*\*\*个，平均营业时长平\*\*\*分钟===

……

2、使用Flask框架，结合Echarts，完成组图输出。组图左侧：纵坐标为城市名称，横坐标为商家数量的条形图，标题为“各城市商家数量Top10”；组图右侧：商家数量前10名的各城市商家平均营业时间。图形展示顺序与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）。

### 任务十一：组图（柱状图、玫瑰图）呈现投诉情况

**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，统计并输出城市商户业务包与投诉相关的分析图例。

**任务描述：**

请根据相关数据库表格，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、根据相关数据库表格，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，统计北京地区各商户业务包的商家数量，同时，按业务包统计客户投诉数，并在PyCharm控制台打印输出，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，投诉\*\*\*条记录===

==北京：2.商户业务包：\*\*\*, 商家数量：\*\*\*家，投诉\*\*\*条记录===

……

2、使用Flask框架，结合Echarts，完成组图输出。组图左侧：横坐标为商户业务包名称，纵坐标为该业务包所包含的商家数量，标题为“北京地区各商户业务部商家数量”的柱形图；组图右侧：各商户业务包投诉占比的玫瑰图，图形各部分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）。

### 任务十二：柱状堆叠图呈现业务包评价组成

**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：**

请根据相关表格数据，统计各商户业务包评价情况，并柱状堆叠图呈现。

**具体任务要求：**

1、根据相关表格，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）。

# 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 |