2019 年全国职业院校技能大赛

GZ-2019032 大数据技术与应用

（高职组）赛题库

目录

任务一：Hadoop 平台及组件的部署管理

注意：任务安装包统一在“/h3cu/”中。

编号	主机名	类型	用户	密码
1	master1-1	主节点	root	passwd
2	slave1-1	从节点	root	passwd
3	slave1-2	从节点	root	passwd

一、 Hadoop 全分布部署

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、解压 JDK 安装包到“/usr/local/src”路径，并配置环境变量；截取环境变量配置文件截图；

2、在指定目录下安装 ssh 服务，查看 ssh 进程并截图（安装包统一在“/h3cu/”）；

3、创建 ssh 密钥，实现主节点与从节点的无密码登录；截取主节点登录其中一个从节点的结果；

4、根据要求修改每台主机 host 文件，截取“/etc/hosts”文件截图；

5、修改每台主机 hostname 文件配置 IP 与主机名映射关系；截取“/etc/hostname”文件截图；

6、根据要求修改 Hadoop 环境变量，并截取修改内容；

7、根据要求修改 Hadoop 相关文件，并初始化 Hadoop，截图初始化结果；

8、启动 Hadoop，使用相关命令查看所有节点 Hadoop 进程并截图。

二、 Hadoop 伪分布部署

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、解压 JDK 安装包到“/usr/local/src”路径，并配置环境变量；截取环境变量配置文件截图；

2、在指定目录下安装 ssh 服务，查看 ssh 进程并截图 （安装包统一在“/h3cu/”）；

3、创建 ssh 密钥，实现主节点与从节点的无密码登录；截取主节点登录其中一个从节点的结果；

4、根据要求修改每台主机 host 文件，截取“/etc/hosts”文件截图；

5、修改每台主机 hostname 文件配置 IP 与主机名映射关系；截取

“/etc/hostname”文件截图；

6、在主节点修改 Hadoop 环境变量（/etc/profile）并截取修改内容；

7、根据要求修改 Hadoop 相关文件（hadoop-env.sh、core-site.xml、hdfs-site.xml、mapred-site.xml、yarn-site.xml），初始化 Hadoop，截图初始化结果；

8、启动 Hadoop，使用相关命令查看所有节点 Hadoop 进程并截图。

三、 Hadoop HA 部署

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、解压 JDK 安装包到“/usr/local/src”路径，并配置环境变量；截取环境变量配置文件截图；

2、在指定目录下安装 ssh 服务，查看 ssh 进程并截图（安装包统一在“/h3cu/”）；

3、创建 ssh 密钥，实现主节点与从节点的无密码登录；截取主节点登录其中一个从节点的结果；

4、根据要求修改每台主机 host 文件，截取“/etc/hosts”文件截图；

5、修改每台主机 hostname 文件配置 IP 与主机名映射关系；截取

“/etc/hostname”文件截图；

6、 在主节点和从节点修改 Hadoop 环境变量，并截取修改内容；

7、 根据要求修改 Hadoop 相关文件，并初始化 Hadoop，截图初始化结果；

8、 启动 Hadoop，使用相关命令查看所有节点 Hadoop 进程并截图；

9、 需安装 Zookeeper 组件具体要求同 Zookeeper 任务要求，并与 Hadoop HA 环境适配；

10、 本题要求配置完成后在 Hadoop 平台上运行查看进程命令，要求运行结果的截屏保存；

11、 修改 namenode、datanode、journalnode 等存放数据的公共目录为/usr/local/hadoop/tmp；

12、 格式化主从节点；

13、 启动两个 namenode 和 resourcemanager；

14、 使用查看进程命令查看进程,并截图(要求截取主机名称),访问两个 namenode 和 resourcemanager web 界面.并截图保存(要求截到 url 状态)

15、 终止 active 的 namenode 进程,并使用 Jps 查看各个节点进程,(截上主机名称),访问两个 namenode 和 resourcemanager web 界面.并截图保存

(要求截到 url 和状态)；

16、 重启刚才终止的 namenode,并查看 jps 进程,截图访问两个 namenode 的 web 界面,并截图保存。

四、 Hive 组件部署

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 Hive 安装包到“/usr/local/src”路径，并使用相关命令，修改解压后文件夹名为 Hive，进入 Hive 文件夹，并将查看内容截图；

2、设置 Hive 环境变量（HIVE_HOME=/usr/local/src/hive ；PATH=$PATH:$HIVE_HOME/bin），并使环境变量只对当前用户生效；

3、新建并配置 hive-site.xml 文件 ,实现“Hive 元存储”的存储位置为 MySQL 数据库；

4、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 下)，初始化结果截图；

5、启动 Hive, 检查是否安装成功，截图保存结果；

6、按指定要求创建 Hive 内部表和外部表，截图保存结果；

7、 按要求实现内外部表转换，截图保存结果；

8、 按指定要求创建分区表，截图保存结果。

五、 Sqoop 组件部署

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、 Sqoop 安装包到“/usr/local/src”路径，并使用相关命令，修改解压后文件夹名为 sqoop，进入 sqoop 文件夹，并将查看内容截图；

2、 修改 Sqoop 环境变量，并使环境变量只对当前用户生效；

3、 修改并配置 sqoop-env.sh 文件，截图并保存结果；

4、 测试 Sqoop 连接 MySQL 数据库是否成功，截图并保存结果；

5、 通过 Sqoop 将 Hive 中数据传输到 MySQL 数据库，截图并保存结果。

六、 Hbase 组件部署

1、 解压 Hbase 安装包到“/usr/local/src”路径，并修改解压后文件夹名为 hbase，截图并保存结果；

2、 设置 Hbase 环境变量，并使环境变量只对当前 root 用户生效，截图并保存

结果；

3、 修改 Hbase 相应文件，截图并保存结果；

4、 把 Hadoop 的相应文件放到 hbase/conf 下，截图并保存结果；

5、 启动 Hbase 并保存命令输出结果，截图并保存结果；

6、 创建 Hbase 数据库表，截图并保存结果；

7、 将给定数据导入数据库表中，截图并保存结果；

8、 查看 Hbase 版本信息，截图并保存结果。

七、 Flume 组件部署

1、 解压 Flume 安装包到“/usr/local/src”路径，并修改解压后文件夹名为 flume；

2、 设置 Flume 环境变量，并使环境变量只对当前 root 用户生效；

3、 修改 Flume 相应文件；

4、 修改并配置 flume-env.sh 文件，截图并保存结果；

5、 测试 Flume 连接 Web 服务器是否成功，截图并保存结果；

6、 通过 Flume 将 Web 服务器中数据传输到 HDFS 中，截图并保存结果。

八、 Spark 组件部署

1、 需前置 Hadoop 环境，并检查 Hadoop 环境是否可用，截图并保存结果；

2、 解压 scala 安装包到“etc/local/src”路径下，并更名为 scala，截图并保存结果；

3、 设置 scala 环境变量，并使环境变量只对当前用户生效，截图并保存结果；

4、 进入 scala 并截图，截图并保存结果；

5、 解压 Spark 安装包到“etc/local/src”路径下，并更名为 spark，截图并保存结果；

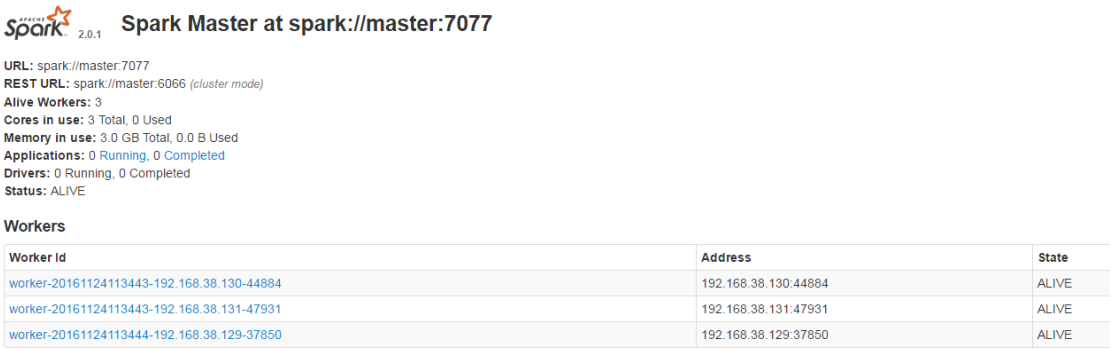
6、 设置 Spark 环境变量，并使环境变量只对当前用户生效，截图并保存结果；

7、 修改 Spark 参数配置，指定 Spark slave 节点，截图并保存结果；

8、 启动 Spark，并使用命令查看 webUI 结果，截图并保存结果；

显示图如以下示例：

主节点



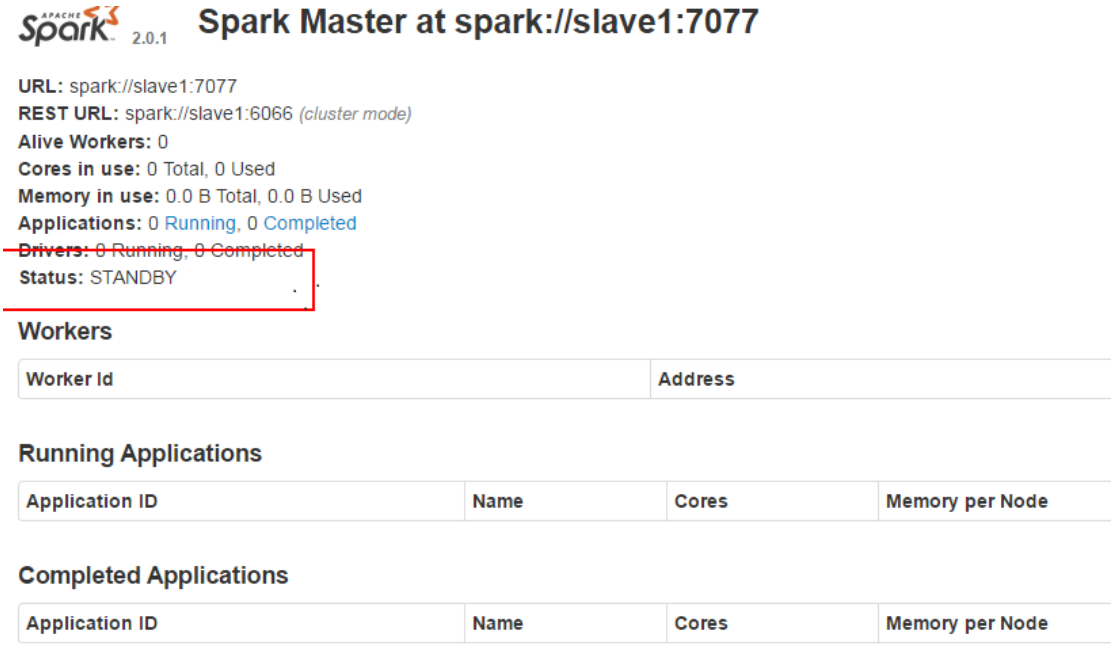
Spark Master at spark://master:7077

URL: spark://master:7077
REST URL: spark://master:6066 (cluster mode)
Alive Workers: 3
Cores in use: 3 Total, 0 Used
Memory in use: 3.0 GB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address	State
worker-20161124113443-192.168.38.130-44884	192.168.38.130:44884	ALIVE
worker-20161124113443-192.168.38.131-47931	192.168.38.131:47931	ALIVE
worker-20161124113444-192.168.38.129-37850	192.168.38.129:37850	ALIVE

从节点



Spark Master at spark://slave1:7077

URL: spark://slave1:7077
REST URL: spark://slave1:6066 (cluster mode)
Alive Workers: 0
Cores in use: 0 Total, 0 Used
Memory in use: 0.0 B Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: STANDBY

Workers

Worker Id	Address

Running Applications

Application ID	Name	Cores	Memory per Node

Completed Applications

Application ID	Name	Cores	Memory per Node

九、 Kafka 组件部署

1、 需安装 Zookeeper 组件具体要求同 Zookeeper 任务要求，并与 Kafka 环境适配，启动 Zookeeper 并截图保存结果；

2、 解压 Kafka 安装包到 “/usr/local/src” 路径，并修改解压后文件夹名为

kafka，截图并保存结果；

3、 设置 Kafka 环境变量，并使环境变量只对当前 root 用户生效，截图并保存结果；

4、 修改 Kafka 相应文件，截图并保存结果；

5、 启动 Kafka 并保存命令输出结果，截图并保存结果；

6、 创建指定 topic，并截图并保存结果；

7、 查看所有的 topic 信息，并截图并保存结果；

8、 启动指定生产者（producer），并截图并保存结果；

9、 启动消费者（consumer），并截图并保存结果；

10、 测试生产者（producer），并截图并保存结果；

11、 测试消费者（consumer），并截图并保存结果。

十、 Storm 组件安装部署

1、 前置安装 Zookeeper 集群，截图并保存结果；

2、 解压 Storm 安装包到“/usr/local/src”路径，并修改解压后文件夹名为 storm，截图并保存结果；

3、 配置“conf/storm.yaml”文件，截图并保存结果；

4、 传送配置好的“conf/storm.yaml”文件，截图并保存结果；

5、 配置 nimbus.seeds 文件，截图并保存结果；

6、 配置 supervisor.slots.ports，截图并保存结果；

7、 拷贝主节点 Storm 包到从节点，截图并保存结果；

8、 设置 Storm 环境变量，并使环境变量只对当前 root 用户生效，截图并保存结果；

9、 在主节点和从节点启动，并截图保存(要求截到 url 和状态)。

十一、 Zookeeper 集群部署

1、 解压 Zookeeper 安装包到“/usr/local/src”路径，并修改解压后文件夹名为 zookeeper，截图并保存结果；

2、 设置 Zookeeper 环境变量，并使环境变量只对当前用户生效，截图并保存结果；

3、 配置“zoo.cfg”文件，截图并保存结果；

4、 修改 myid 文件，截图并保存结果；

5、 启动每个服务器上面的 Zookeeper 节点，启动完成之后查看每个节点的状态，截图并保存结果。

任务二：数据采集

数据采集将直接影响数据清洗、分析、可视化。本任务提供四个脱敏数据源：交通运输、招聘、酒店、零售。

请使用企业生产环境常用采集工具和网络爬虫相关技术，完成网页分析、数据采集、数据爬取，数据存储，并将采集数据进一步进行相关数据操作。

一、 数据源 1（交通运输）

航空出行由于它的快捷便利，已经被越来越多的人喜欢，某航空公司通过多年运营，积累了大量会员档案和乘坐航班信息，为对客户进行分群，明确价值客户群体，将有限的营销资源集中于高价值客户，实现企业利润最大化。为此，该航空公司聘请“H3CU”大数据分析公司完成此项目。

由于会员信息属于公司机密数据，该航空公司将数据脱敏后以 csv 文件传送给“H3CU”公司进行数据处理与分析，为安全考虑“H3CU”公司需将数据先存入数据库中备份，再进一步数据清洗与分析。请参考一下相关专业说明完成任务。

1、 航空公司积累了大量会员档案信息和乘坐航班信息，其中包含了会员卡号、

入会时间、性别、年龄、会员卡级别、在观测窗口内的飞行公里数、飞行时间、飞行次数等 44 个特征属性，数据存放在 csv 格式文件中。

2、识别客户价值应用最广泛的模型是 RFM 模型。其中，R（Recency）指的是最近一次消费时间与截止时间的时间间隔，通常 R 值越小，客户对商品或服务最可能感兴趣。F（Frequency）指顾客某段时间的消费次数，次数越高，顾客价值越大。M（Monetary）指顾客在某段时间内的消费金额。

3、由于在本任务中，同样消费金额的不同客户，对航空公司的价值是不同，比如，一位购买长航线、低等级舱位的旅客与一位购买短航线、高等级舱位的旅客相比，可能票价是一样，但后者对航空公司的价值可能更高。所以，用累计行程 M 和乘坐舱位对应的折扣系数 C 代替消费金额。

4、航空公司会员入会时间也一定程度影响客户价值，因此增加客户关系长度 L 做为另一特征。构建出包含 6 个特征的模型，分别和原始数据中的 FFP_DATE（入会时间）、LOAD_TIME（观测窗口结束时间）、FLIGHT_COUNT（观测窗口内的飞行次数）、AVG_DISCOUNT(平均折扣系数）、SEG_KM_SUM(观测窗口的总飞行千米数）、LAST_TO_END(最后一次乘机时间至观测窗口结

束时长）。

本次任务包括以下内容：

1、 使用 Java 或 Python 语言编写程序，将给定 csv 格式的数据文件写入 Mysql 数据库中，并将代码与运行结果截图保存。

1）导入模块

2）连接数据库

3）创建表，表名称

4）将数据写入数据库

5）关闭数据库

2、 使用数据传输工具，将 Mysql 数据库中的航空数据导入大数据平台中进行数据清洗，并将命令与运行结果截图并保存。

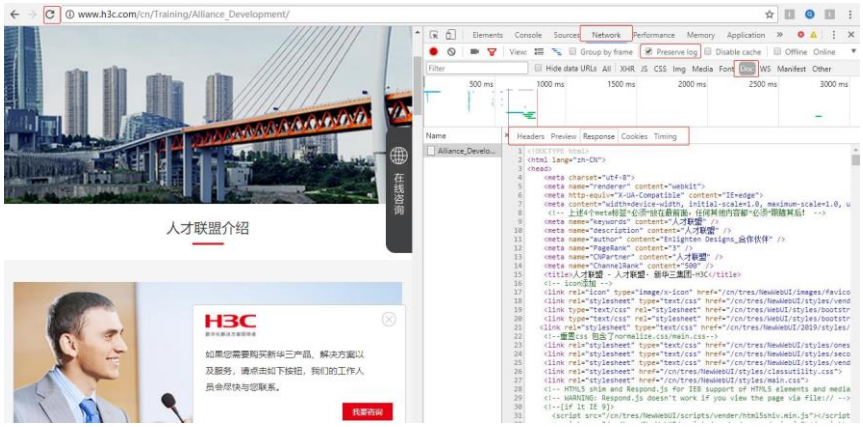
二、 数据源 2（web，招聘）

1、 网站解析，利用 chrome 查看网页源码，分析招聘网站网页结构。

1）“检查”招聘网站，在网页中右键点击检查，或者 F12 快捷键,进入如下图的查看元素页面；

（示例图 1）

2）检查网站：点击 Network、勾选 Preserve log、点击 Doc、点击清理按钮、刷新页面、点击 Response，在 Response 查看所需内容。



（示例图 2）

2、从招聘网站中爬取需要数据，按照要求使用 Java 或 Python 语言编写并完善爬虫代码，爬取指定数据项，有效数据项包括但不限于：所在城市、公司名称、薪资、技能要求等多项字段。并将代码文件与代码截图保存。

具体步骤如下：

1）创建爬虫项目\H3CU_recruit\

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

3、 至此已从招聘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，请将操作命令截图并保存。

三、 数据源 3（web，酒店）

1、 网站源码解析:利用 chrome 查看网页源码，分析酒店网站网页结构。

1）“检查”酒店网站，在网页中右键点击检查，或者 F12 快捷键,进入如下图的查看元素页面；

2）检查网站：点击 Network、勾选 Preserve log、点击 Doc、点击清理按钮、刷新页面、查看所需内容。

2、 从酒店网站中爬取需要数据，按照要求使用 Java 或 Python 语言编写并完善爬虫代码，爬取指定数据项，有效数据项包括但不限于：城市、商圈、星级、评分、评论数等多项字段。并将代码文件与代码截图保存。

具体步骤如下：

1）创建爬虫项目\H3CU_hotel\

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

至此已从酒店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，请将操作命令截图并保存。

四、 数据源 4（web，零售）

1、 网站解析，利用 chrome 查看网页源码，分析零售网站网页结构。

1）“检查”零售网站，在网页中右键点击检查，或者 F12 快捷键,进入如下图的查看元素页面；

2）检查网站：点击 Network、勾选 Preserve log、点击 Doc、点击清理按钮、刷新页面、点击 Response，在 Response 查看所需内容。

2、 从零售网站中爬取需要数据，按照要求使用 Java 或 Python 语言编写并完善爬虫代码，爬取指定数据项，有效数据项包括但不限于：客户信息、员

工信息、商品信息、商场信息等多项字段。并将代码文件与代码截图保存。

具体步骤如下：

1）创建爬虫项目\H3CU_mart\

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

3、 至此已从零售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，请将操作命令截图并保存。

任务三：数据清洗与分析

一、 数据源 1（交通运输）

本阶段的任务是：将客户基本信息、乘机信息、积分信息等用户信息进行清洗和整理，并完成数据计算、分析和数据可视化。

分析统计航空公司的样例数据，使用 Java 或 Python 语言进行目标数据读取、数据探索、数据预处理、数据特征构造等，并按题目要求输出到指定文件中。

1、 数据处理，提取文件中每列数据中的空值个数、最大值、最小值，并打印输出数据，截图并保存结果；

2、 剔除票价中价格（SUM_YR_1、SUM_YR_2）为空的记录，并输出修改后的行列数量；

3、 保留票价（SUM_YR_1、SUM_YR_2）非零的、平均折扣率（avg_discount）不为 0 且总飞行公里数（SEG_KM_SUM）大于 0 的记录，并打印输出修改后的行列数量；

4、 剔除原始数据中不相关的属性，根据客户价值，按题目要求选择相关的 6 个属性，并打印输出前 5 行信息；

5、 通过属性构造提取题目指定额 5 个指标；

6、 由于 5 个指标之间的取值范围差异较大，需要对数据进行标准化处理，使用标准差进行标准化处理，并打印输出前 5 行数据；

7、 计算标准化数据各列的平均值，并打印输出；

8、 计算标准化数据各列的 20%截尾均值，并打印输出；

9、 计算标准化数据各列的中位数，并打印输出；

10、 计算标准化数据各列的分位数（第四三分位），并打印输出；

11、 计算标准化数据各列的协方差，并打印输出；

12、 打印显示标准化数据各列的汇总统计量。

二、 数据源 2（招聘）

本任务使用的招聘网站初始数据集来自于多个网站及平台，且为多次采集汇总，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。请分析数据集 recruit，根据题目规定要求实现数据清洗。

1、 对于原始数据集字段缺失，可采用填充默认值、均值、众数、KNN 填充、以

及把缺失值作为新的 label 方式处理。当缺失信息较少时可采用删除的方式进行处理。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，所以需要对数据业务逻辑进行全面分析后，确定不合规数据处理方式。请根据题目具体参数要求，处理招聘数据中的不合规数据，并存入指定数据表或数据文件中，截图并保存结果。

2、 本任务原始数据集中存在一些脏数据，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数据。这些脏数据会影响后续的数据分析结果，所以需要对脏数据进行预处理。请根据题目具体参数要求，处理工资字段不合规数据，使该字段数据格式统一。将清洗后的数据存入指定数据表或数据文件中，截图并保存结果。

3、 本任务给定的数据集来自于多个网站及平台，且为多次采集汇总的数据，在整合多来源数据时可能遇到时间、日期、数值、全半角等显示格式不一致的问题，需要将其处理成一致的格式，以便于进行后续的数据分析。请根据题目具体参数要求，将原始数据集中格式不一致的数据进行标准化处理，并存入指定数据表或数据文件中，截图并保存结果。

4、 若要将清洗后的数据存储到数据文件中，需要将数据的不同字段使用某种分隔符分隔开后，再写入数据文件中。后续将数据文件再导入数据库时，同样以该分隔符进行字段划分。请根据题目具体参数要求，将清洗后的数据以指定数据分隔符进行分隔，存入指定数据文件中，再使用数据转移工具将数据导入数据库中，截图并保存结果。

5、 高校开设新专业、新方向时，要以企业相关岗位招聘数量作为重要依据。各大招聘网站发布的招聘信息是主要的数据来源，进而能够统计某类岗位的招聘数量。请根据题目具体参数要求，统计岗位招聘数量，并存入指定数据表或数据文件中，截图并保存结果。

6、 我们根据招聘网站数据通常能够了解相关岗位的招聘情况，包括但不限于地区分布、学历要求、经验要求、薪资水平等。这些信息为高校专业设置提供了分析依据和佐证数据。请根据题目具体参数要求，按要求统计相关职位招聘信息，并存入指定数据表或数据文件中，截图并保存结果。

7、 职业技能图谱描绘了各岗位从业人员的知识技能要求，能够帮助学生梳理知识框架结构，提供学习路径指导，了解各知识点和技能点的重要程度。通过招聘网站数据整理职业技能图谱，将有助于学校的专业课程设置，也可使学

生了解到岗位从业人员的知识技能要求。请根据题目具体参数要求，分析各知识技能在某个招聘岗位能力需求中的占比情况，并存入指定数据表或数据文件中，截图并保存结果。

8、 根据近年某大型招聘网站发布的城市平均工资分布表显示，在全国各城市中，最高的城市平均工资高达上万元，而最低的城市平均薪酬在 5000-6000 元左右。工资薪酬是影响择业的很大一个因素，但一线城市的消费水平也同样很高，房租、交通和伙食费等各方面都是一笔不小的开支。高校毕业生择业需要根据各方面因素综合进行考量。请根据题目具体参数要求，统计各城市指定招聘岗位的平均工资，并存入指定数据表或数据文件中，截图并保存结果。

9、 工作地区与招聘岗位是决定毕业生就业薪酬待遇的两个关键因素。不同地区或不同岗位的工资待遇往往存在较大差异，这体现了地区行业发展和人才需求的分布情况。请根据题目具体参数要求，统计指定城市和指定岗位的工资待遇，并存入指定数据表或数据文件中，截图并保存结果。

三、 数据源 3（酒店）

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。 脱敏后的数据存放于 hdfs:hoteldata/hoteldata.csv。

初始数据集来自多个网站及平台系统，且为多次采集汇总，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。 请分析数据集 hoteldata，根据题目规定要求实现数据清洗。

1、酒店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删失或截断。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，从总体上来说分为删除存在缺失值的个案和缺失值插补。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 个的数据条目剔除原始数据集,并输出剔除的条目数量，截图并保存结果。

2、 对于数据集字段缺失情况，通常可以采用填充默认值、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，截图并保存结果。

3、 原始数据集来自于多个平台及网站，且为多次采集汇总，因此数据集中的某些字段有可能会出现一些冗余或非法格式，例如多次采集过程中产生的冗余信息，或来自于某网站的不合规数据。这些信息的存在既无实际的业务分析意义，甚至还会影响最终分析结果。请根据题目具体参数要求处理不合规数据，截图并保存结果。

4、 给定数据集中，酒店信息覆盖全国各个城市，不同省份及城市间旅游业的发展程度而是各不相同。考虑到数据集规模较大，酒店信息所形成的大数据集难以直观理解和统计，为便于信息理解和整合，请根据题目具体参数

要求处理数据，截图并保存结果。

本阶段的任务：hoteldata.csv 文件中已经包含了数据采集阶段从酒店网站上爬取的数据集，其中包含来自不同城市中多家酒店的销售信息，你的小组通过编写代码或脚本完成对文件 hoteldata.csv 文件中酒店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 等开发语言，完成本阶段数据清洗、存储、转化、分析及数据推送等任务。通过多个维度分析酒店的销售信息，并以此评价酒店销售业绩、区域的游客接纳能力、接纳质量等指标。

1、 城市游客接纳能力是城市规划建设中的重要指标，其中城市的酒店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酒店管理网站中的数据统计各城市的酒店数量和房间数量，以城市房间数量降序排列并输出前 10 条统计结果，并写入指定的数据库或数据文件，截图并保存结果。

2、 酒店的间夜量也叫间夜数，是酒店在某个时间段内，房间出租率的计算单位。

1 个房间被使用 1 个晚上被记作 1 个间夜数，如一个酒店一周内有 30 个房间被入住 1 晚,7 个房间被入住两晚，则间夜数为 1*30+2*7，44 个间夜。根据现有数据及给定参数完成酒店间夜量数据统计，并写入指定的数据库或数据文件，截图并保存结果。

3、 各地区的酒店销售数据等信息能够反映一个地区的游客接待能力。例如酒店总量多的城市大都具有强烈的吸纳外来人员的能力，订单数量反映该地区的活跃外来游客的数量。根据现有数据及给定参数完成酒店销售数据统计，并写入指定的数据库或数据文件。

4、 酒店客房出租率是反映酒店经营状况的一项重要指标，它是已出租的客房数与酒店可以提供租用的客房总数的百分比。酒店出租率越大，说明实际出租客房数与可供出租的客房数之间的差距越小，也就说明酒店的客源市场越充足，在一定程度上表明酒店经营管理的成功。根据现有数据及给定参数完成酒店出租率统计，并写入指定的数据库或数据文件，截图并保存结果。

5、 OTA，全称为 Online Travel Agency，中文译为“在线旅行社”，是旅游电子商务行业的专业词语。指“旅游消费者通过网络向旅游服务提供商预定旅游产品或服务，并通过网上支付或者线下付费，即各酒店通过网络进行产品

营销或产品销售”。OTA 平台是酒店营销的主要途径之一，不仅降低销售成本，同时也提高了顾客体验满意度。当顾客通过 OTA 平台进行酒店预订时，酒店就拥有了用户的相关数据。通过这些数据，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 OTA 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；酒店违反 OTA 规则擅自以低价让客户取消订单，这种情况又叫做“切单”。OTA 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，OTA 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，截图并保存结果。

6、 根据业务发展需要，OTA 平台欲在全国范围内拓展合作酒店，因此请统计全国区域的 OTA 酒店订单预定及完成情况。请根据现有数据及给定参数完成统计，并写入指定的数据库或数据文件，截图并保存结果。

7、 酒店出租率是反映酒店经营状况的一项重要指标，它是已出租的客房数与酒店可以提供租用的房间总数的百分比。请根据现有数据中的相关字段分析各个酒店的经营状况，并写入指定的数据库或数据文件，截图并保存结果。

8、 高端酒店的数量，从一个侧面反映了当地的经济水平，据国家旅游和文化部统计境内 31 个省市（不含港澳台）共有 860 家五星级酒店，但分布很不均衡，其中东部沿海所占有的五星级数量，接近了全国一半。请你根据题目要求统计符合参数要求的高端酒店相关信息，并写入指定的数据库或数据文件，截图并保存结果。

9、 酒店的销量一般与用户的满意度具有强相关性。经业内调查分析发现，4.5 分以上的酒店，能够拥揽 OTA 平台酒店详情页面流量的 72%，好评分数从 4.5 到 4.7 的提升能够带来 100 元以上的房价上涨空间。一般来说酒店的评分越高，评论条目数越多，该酒店在 OTA 平台酒店列表中的排序也会越高，同样的也会获得更多的客户浏览量。请根据给定数据集分析相应酒店受欢迎度。请你根据题目要求统计符合给定参数的相关酒店信息，并写入指定的数据库或数据文件，截图并保存结果。

10、 连锁酒店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符合大众化消费。连锁酒店无论在装修、服务还是信誉上都有较大的竞争优势，所以连锁酒店是出差、旅游住宿的好选择。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

与消费特点的差异等问题，一些已经成熟酒店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。请根据给定数据集分析指定连锁酒店在全国各地区的经营情况，并写入指定的数据库或数据文件，截图并保存结果。

11、 近年来，随着我国旅游业的蓬勃发展，城市旅游业已成为重要的支柱产业和新的经济增长点，对于城市经济发展，塑造城市形象，优化产业结构发挥着巨大作用。一个城市旅游业的发展不仅需要具备独特的自然风光或者人文资源，还应具备一定旅游接待能力，保持良好的游客口碑。请根据原数据集在指定维度综合分析并获得城市的受欢迎程度排名，并写入指定的数据库或数据文件，截图并保存结果。

12、 订单数据是考量 OTA 直销酒店经营业绩的重要指标，由于某些酒店资源无法内部消化，也会出现订单分销至其它 OTA 平台的情况，此时称为分销。一般情况下，直销和分销是同时存在的。但当某些酒店或区域分销数量过多时，则表明 OTA 平台经营推广能力不足。请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，截图并保存结果。

13、 酒店直销订单情况是考量酒店直销经营业绩的重要指标，当直销业绩未达到预期要求时需要从多方面总结业绩不佳的原因，从而加强薄弱环节，转

变销售策略，提高销售业绩。请根据现有数据及给定参数完成相关订单业绩统计，并写入指定的数据库或数据文件，截图并保存结果。

14、 请根据题目中具体参数要求，使用数据传输工具，将指定文件推送至相关位置，截图并保存结果。

四、 数据源 4（零售）

零售行业是所有行业中对大数据最为敏感的行业之一。零售商品传统理解很简单：想要赚取最大利润往往需要对消费者的购买趋势能够做出最快反应。不过自从所有零售商都开始采取相同打折促销的手段，互相争抢的同类消费人群，至此赚取尽可能多的利润不再容易实现。大数据正在逐步改变这种现象，如今的零售商很多都会在抢购热潮来临之前，使用大数据来分析消费趋势，以此在抢购热潮中获得更多收益。随着社会经济的快速发展，消费者的需求和偏好一直在改变。零售行业必须要跟上消费者的脚步，通过购买习惯、热门商品、年龄结构、消费层次等手段对顾客的消费数据做出分析，设定未来市场经营策略，以便获得更好的发展。

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、在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 hdfs:martdata/。

1、 对于零售行业来说，不同年龄的客户消费能力也各不相同。例如：35 岁已进入事业稳定期，在具备经济基础的同时开始追求生活品质，表现出来的消费能力更强，同时会倾向于价格更贵，品质更高的商品。25 岁新贵们追求独立自由，初入社会对周围抱有强烈的社交和探索欲，经济基础薄弱但在消费方面显得更加精明。20 岁后自我意识强烈，表现出“拟成人化”特征，在整体的消费能力方面呈现上升趋势。为提升利润商业方案做准备，首先需要分析不同客户的年龄数据。请根据题目要求实现对给定原数据中客户年龄的统计，并写入指定的数据库或数据文件，截图并保存结果。

2、 公司员工的年龄组成结构直接影响公司的发展和风格导向，当员工年龄结构偏大时，公司业务基本定型，员工老练且富有经验，但可能会存在管理困难等问题。同时面对养老及退休问题，公司将面临一大笔开支。当年龄结构偏

小时，公司氛围比较阳光活跃，员工干劲十足也容易产生突破和创新，但也会存在做事浮躁，人员流动性大等隐患。公司员工合理的年龄组成应如下：25 岁以下员工是培养的主要对象约占 20%，25—35 岁的员工是成长的主要力量约占 20%，35——55 的员工是担当的主力约占 40%，55 岁以上员工是经验传承的依靠约占 20%。基于公司未来发展规划考虑，需要分析员工的年龄数据。请根据题目要求实现对给定原数据中员工年龄信息统计，并写入指定的数据库或数据文件，截图并保存结果。

3、 原数据集可能因人员流动未跟进、数据库切换等原因，导致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。请根据题目 2 的结果数据集，按照本题规定要求实现数据清洗，并写入指定的数据库或数据文件，截图并保存结果。

4、 工资差距是指员工之间工资的收入差距，职位不同工资也不一样。当企业工资差距较大时员工有可能产生不公平的情绪，工资差距太小时则可能打压员工工作积极性。为确保公司工资分布的合理性，请根据题目 2 结果数据按要求对员工的工资数据进行分析，完成公司整体工资分布统计，并写入指定的

数据库或数据文件，截图并保存结果。

5、 为促进平等就业的机会，企业员工的男女比例是否均衡问题一直备受政府关注。合理的性别比例能够有效降低员工的流动性及保持一定的工作积极性。请根据清洗-4 题目结果数据按要求统计男女员工相关信息，并写入指定的数据库或数据文件，截图并保存结果。

6、 有效进行会员卡分级有利于客户的筛选和定位，对后续的客户服务具有指导性的意义，并在客户维护方面使公司有所侧重，这样在保证商家营业额的同时，又能深挖现有的客户资源。请根据数据清洗 1 题目结果数据按要求统计会员卡分级情况，并写入指定的数据库或数据文件，截图并保存结果。

7、 销售促销是市场竞争过程中的一把利剑，市场锋线的促销作用在于对产品施加推力，使产品能够更快地进入市场和扩大市场。促销实质上是一种沟通活动，即营销者（信息提供者或发送者）发出作为刺激消费的各种信息，把信息传递到一个或更多的目标对象。成功的促销会带动销售业绩。促销活动带来的销售额越高则说明促销活动越成功。请根据相关数据集完成题目具体要求，并写入指定的数据库或数据文件，截图并保存结果。

8、 随着季节的变化，人们吃穿用的商品也相应变化。商店在出售商品时，也应

按季节的变化随时调整商品的陈列。季节性商品的陈列应在季前开始，商店应了解顾客的潜在需要，根据天气的变化来改变商品的陈列，否则将丧失适时销售的良机。请以相关数据集根据题目具体要求完成商品分类的销售统计，并写入指定的数据库或数据文件，截图并保存结果。

9、 为分析商场经营管理情况，制定后续营销策略和经营计划，按销售业绩制定奖励计划、分配部门奖金或改进销售策略，请以相关数据集根据题目具体要求完成指定维度的销售统计，并写入指定的数据库或数据文件，截图并保存结果。

10、 对于连锁商场来说，不同类型店铺的所带来的销售额也大不相同，例如大型综合超市的月销售额与同品牌定位在小区门口的社区便利店是不可而语的。但同时，大型综合超市的选址、占地、员工人数众多，相应的经营成本也很高。请以店铺类型作为衡量维度，使用相关数据集完成题目具体要求，并写入指定的数据库或数据文件，截图并保存结果。

11、 公司计划在今年新增 25 家连锁门店，目标整体营业额全国销售突破 580 亿。由于不同地区因经济发展水平，居民消费能力等差异，所带来的商品销售情况也各不相同。请根据现有数据以地区为维度进行销售量汇总，并写入

指定的数据库或数据文件，截图并保存结果。

12、 会员卡不仅是会员身份的象征，也是消费者和商家之间最强的关联。管理会员卡是商家的一大重任。会员等级制度的设定，其目的是为了通过不同等级差异化的会员权益，来刺激会员的消费欲望。选择注册会员，代表着顾客对店铺的认可。店铺设置会员等级，区分会员权益，是针对已注册会员的再营销。一致性的会员特权，会逐渐让老会员失去兴趣，活跃度下降。会员卡的级别应该根据会员的消费情况定期调整，通过会员等级的刺激，能够逐渐建立新会员的忠诚度，同时激发老会员的消费热情，让会员保持高活跃，持续不断的为商家创造价值。请根据相关数据文件分析商场会员的相关信息完成题目，并写入指定的数据库或数据文件，截图并保存结果。

13、 销售数据分析工作涉及到销售成本分析、客户满意度分析、客户需求分析等。为了进行销售数据分析，需要对数据统计和分类，了解销售状态，并进一步做出决策。针对同一市场不同品牌产品的销售差异分析，可为企业的销售策略提供建议和参考；针对不同市场的同一品牌产品的销售差异分析，可为企业的市场策略提供建议和参考；微观销售分析，可分析决定未能达到销售额的特定产品、地区等。销售分析可以决定一个商场的经营方向 。请根

据相关数据文件按照题目具体要求，分析商场销售数据，并写入指定的数据库或数据文件，截图并保存结果；

14、 请根据题目中具体参数要求，使用数据传输工具，将相关文件推送至指定位置，截图并保存结果。

任务四、数据可视化

一、 数据源 1（交通运输）

1、根据对航空公司 LRFMC 模型含义的理解

L：会员入会时间距离观测窗口结束的月数

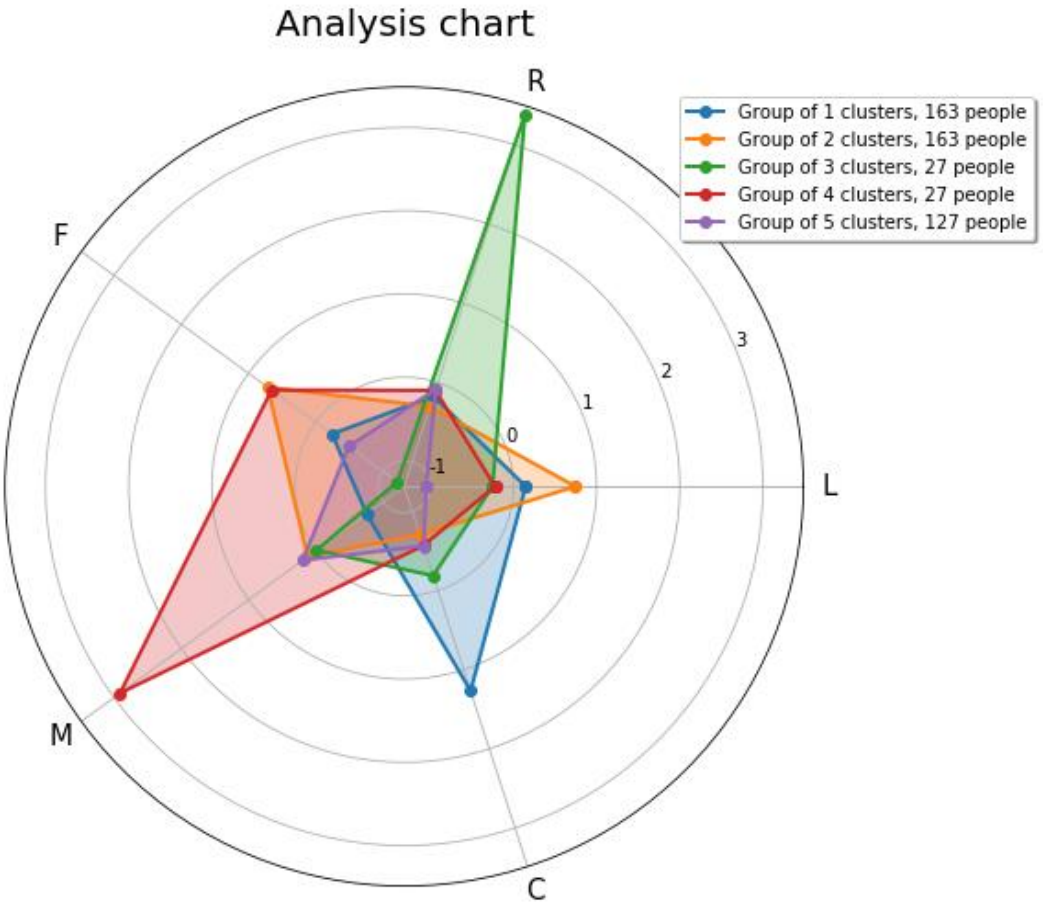
R：客户最近一次乘坐飞机距观测窗口结束的月数

F：客户在观测窗口内乘坐飞机的次数

M：客户在观测窗口累计的飞行里程

C：客户在观测窗口内乘坐仓位所对应的折扣系数的平均值

使用可视化工具对给定数据结果进行分析，导出相关类别的可视化图表。



（图形样例 1）

2、 航空会员卡是会员身份的象征，在一定程度上也是会员飞行里程的体现，飞行里程越多，会员等级越高，也就能证明该客户为航空公司的价值客户。请根据指定表中数据，通过指定图例进行呈现。

二、 数据源 2（招聘）

1、 热门职业特指在以前没有引起大家关注的职业，因为经济环境的改变而现在

收入高或者工作环境好（抑或满足人们对职业的特殊偏好）的行业。随着信息技术的不断发展，云计算、大数据、人工智能都成为了热门职业方向，为了了解热门职业中对各岗位招聘人数的数量，请根据指定表中数据，统计出招聘数量最多的前几位的热门职业，通过指定图例进行呈现。

2、 大数据公司内部包含有许多的岗位 ,例如 JAVA 开发工程师、大数据架构师、大数据开发工程师、大数据清洗工程师、大数据分析师等，不同岗位对于从业人员学历、经验、薪资等都有一定的要求，请根据指定表中数据，统计出各个岗位中相关招聘职位的数量，通过指定图例进行呈现。

3、 大数据产业作为一个新兴信息产业，对从业人员的知识面要求较高，会涉及到数学、统计、编程、系统部署等多方面知识储备，这些知识内容又会对应成为一个个的技能点，将这些技能点进行汇总形成大数据岗位的职业技能要求，是学生今后主要提升的技能点之一。请根据指定表中的数据，分析各知识技能在某个招聘岗位能力需求中的占比情况，通过指定图例进行呈现。

4、 大数据产业作为一个新兴信息产业，各地大数据产业都在蓬勃发展中，对于大数据人才的需求也在不断的增加，但是由于人才的相对紧缺，导致大数据产业的整个工资待遇水平较同行业也具备一定的优势，请根据指定表中的数

据，统计出全国某些城市指定招聘岗位平均工资，通过指定图例进行呈现。

5、 近些年大数据产业在全国大幅发展，各个公司对于大数据人员的招聘数量也在不断增加，通过大数据相关职位的招聘数量可以从一定程度看出行业内人员流动情况，请根据指定表中的数据，统计出近几年指定职位招聘数量汇总，通过指定图例进行呈现。

三、 数据源 3（酒店）

1、 出租率是反映酒店经营状况的一项重要指标，它是已出租的客房数与酒店可以提供租用的房间总数的百分比。酒店出租率的情况可以在一定程度上反应出该酒店的整体运营的情况，为了更好的分析指定酒店的入住情况，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。

2、 连锁酒店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符合大众化消费。连锁酒店无论在装修、服务还是信誉上都有较大的竞争优势，所以连锁酒店是出差、旅游住宿的好选择。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的差异等问题，一些已经成熟酒店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

锁酒店的经营状况，并以指定图例进行呈现。

3、 酒店订单量是反应酒店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酒店入住情况，为了更好地分析全国各省酒店订单量，请根据指定表中数据统计出全国各省酒店订单量的情况，并以指定图例进行呈现。

4、 酒店的间夜量也叫间夜数，是酒店在某个时间段内，房间出租率的计算单位，关于酒店间夜量的计算公式为间夜量=入住房间数*入住天数。例如某酒店今天入住的房间数为 500，则今天的间夜量=500*1=500，而又比如某酒店这个月（30 天）的平均每天入住房间数为 400，则这个月的间夜量=500*1*30=15000。请根据指定表中数据统计酒店间夜数相关数据，并以指定图例进行呈现。

5、 OTA 平台需要综合评判一个城市酒店运营情况，会涉及到多方面酒店数据，例如像高端酒店数量、订单数量、住客评分、评论数量、出租率、200 元/晚以下快捷酒店数量等信息，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。

6、 订单数据是考量 OTA 直销酒店经营业绩的重要指标，由于某些酒店资源无法内部消化，也会出现订单分销至其它 OTA 平台的情况，此时称为分销。一

般情况下，直销和分销是同时存在的。但当某些酒店或区域分销数量过多时，则表明 OTA 平台经营推广能力不足。请根据指定表中数据，以指定图例进行呈现。

7、 OTA 平台为了能在更多省份扩展业务，与更多酒店建立合作关系，为了赢得更多酒店的合作，在合作谈判过程中会通过同区域、同等级对比，需要提供同类酒店相关经营数据。请根据指定表中数据，以指定图例进行呈现。

四、 数据源 4（零售）

1、 销售额为顾客消费单价的总和。商场的销售额不仅是体现运营人员水平的重要标准,也是衡量一个商场是否具有竞争实力的重要指标。因此分析影响商场销售额的因素对提高商场的经营业绩发挥着重要作用。按题目指定要求，输出销售额相关图例。

2、 零售企业为获得更大收益，需时时关注成本与销售额的趋势，当成本越低，销售额越高时，可获得最大收益。为便于领导决策，需提供成本与销售额数据分析结果，请根据指定表中数据，通过指定图例进行呈现。

3、 会员卡不仅是会员身份的象征，也是消费者和商家之间最强的关联。管理会员卡是商家的一大重任。会员等级制度的设定，其目的是为了通过不同等级

差异化的会员权益，来刺激会员的消费欲望。选择注册会员，代表着顾客对店铺的认可。店铺设置会员等级，区分会员权益，是针对已注册会员的再营销。一致性的会员特权，会逐渐让老会员失去兴趣，活跃度下降。会员卡的级别应该根据会员的消费情况定期调整，通过会员等级的刺激，能够逐渐建立新会员的忠诚度，同时激发老会员的消费热情，让会员保持高活跃，持续不断的为商家创造价值。请根据指定表中数据，通过指定图例进行呈现。

4、 利润率反映一定时期促销活动所带来的利润水平相对指标。利润率指标既可考核营销活动利润计划的完成情况，又可比较各促销方案之间的活动经营管理水平，提高经济效益。利润率使指销售利润总额与销售收入总额的比率，它表明单位销售收入获得的利润，反映销售收入和利润的关系。零售企业为获得更大收益，成本逐年增加，需时时关注利润率，降低成本提高利润，根据指定表中数据，通过指定图例进行呈现。

5、 促销是营销者向消费者传递有关本企业及产品的各种信息，说服或吸引消费者购买其产品，以达到扩大销售量的目的。常用的促销手段有广告、人员推销、网络营销、营业推广和公共关系。为了更好的计算促销带来收益变化，请根据指定表中的数据，以指定图例进行呈现。

6、 会员卡是会员身份的象征，在一定程度上也是会员消费能力的象征，消费金额越多，会员等级越高，也就能体现会员的消费能力。请根据指定表中数据，通过指定图例进行呈现。

7、 为预测下一年商品的进货数量，需要分析近年各类商品的销售情况，为了使商场下一年能获得更多的利润，请根据指定表中的数据，以指定图例进行呈现。

任务五、综合分析

一、 数据源 1（交通运输）

至此我们对于航空业务背景及相关数据已经有了一定的了解，在综合理解航空业务数据的基础上，结合题目中任务一、二、三、四的相关结论，根据题目要求进行分析，并编写输出分析报告。注：分析结果需要具有任务中结论作为佐证材料。

二、 数据源 2（招聘）

通过任务二的网站分析及数据爬取、任务三的数据清洗与分析及任务四的可视化呈现，我们已经清晰的了解了招聘业务背景及相关招聘数据，在综合理解招聘业务数据的基础上，根据题目要求进行分析，并编写输出分析报告。

注：分析结果需要具有任务中结论作为佐证材料。

三、 数据源 3（酒店）

假定你为 OTA 平台的管理者，在综合理解酒店业务数据的基础上，通过以上任务一、二、三、四的相关结论，对未来拓展合作酒店方向做出预测，根据题目要求进行分析，并编写输出分析报告。

注：分析结果需要具有任务中结论作为佐证材料。

四、 数据源 4（零售）

作为商场的 CFO（首席财务官——CFO(Chief Financial Officer)是企业财务治理的第一责任人，为企业获得最大利润出谋划策），最为关心商场的销售收入、销售成本、销售利润，通过打折促销、会员机制以及连锁经营等方式，提升商场销售利润。在综合理解零售业务数据的基础上，结合任务一、二、三、四的相关结论，根据题目要求进行分析，并编写输出分析报告。

注：分析结果需要具有任务中结论作为佐证材料。