**2024年度“楚怡杯”湖南省职业院校技能竞赛**

**（第二阶段赛事）中职组计算机专业类**

**“大数据应用与服务”赛项**

**[时量：240分钟，试卷号：A ]**

**竞**

**赛**

**任**

**务**

**书**

**场次号： 工位号：**

**2024年5月**

**一、竞赛任务概述**

本赛项包括平台搭建与运维、数据获取与处理、业务分析与可视化、职业素养等四个竞赛模块，各模块分值分别为30分、35分、30分、5分，本赛项满分为100分。参赛选手在赛场连续240分钟内完成四个竞赛模块，比赛内容如表1所示。

表 1：任务内容与分值比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模块** | | **主要内容** | **比赛时长** | **分值比** |
| 模块一：  平台搭建与运维 | 任务一：  大数据平台搭建 | 安装Hadoop 全分布式平台，安装Hadoop平台相关的常用组件，包括但不限于ZooKeeper 、Flume 、Kafka 、Spark 、Flink 、Redis 、HBase 等，验证Hadoop 平台和相关组件的可用性。Hadoop 完全分布式下的 JDK 的解压安装、JDK 环境变量配置、节点配置、Hadoop 配置文件修改、运行测试等。 | 240分钟 | 10% |
| 任务二：  数据库配置维护 | 使用 MySQL 数据库建库建表，运用基本的 SQL 语言完成数据的增删改查等操作。 | 20% |
| 模块二：  数据获取与处理 | 任务一：  数据获取与清洗 | 对 CSV 数据文件进行加载、清洗和转换等操作，识别和处理无效值，检查数据的一致性，将清洗后的数据保存到指定位置。 | 10% |
| 任务二：  数据标注 | 使用Python语言对数据进行分类标注。 | 10% |
| 任务三：  数据统计 | 基于 Hadoop 平台进行编译、打包、部署和执行程序，完成数据的统计工作。 | 15% |
| 模块三：  业务分析与可视化 | 任务一：  数据可视化 | 使用Python可视化库或Web前端框架对数据进行呈现，包括但不限于柱状图、折线图、玫瑰图、气泡图、饼状图、条形图、雷达图、散点图等效果。 | 20% |
| 任务二：  业务分析 | 报表分析，对大数据项目的业务场景和数据进行分析，撰写报告。 | 10% |
| 职业素养 | | 团队分工明确合理、操作规范、文明竞赛。 | 5% |

**二、注意事项**

1.本竞赛任务书总共24页，任务完成总分为100分。

2.参赛团队应在**240分钟**内完成任务书规定内容，并按照要求进行保存，比赛结束前将客户端电脑桌面上的result文件夹重新命名为“考场号-座位号”，并将文件夹拷贝到 U 盘。

3.参赛选手不得在任何纸质材料、客户端电脑上出现地名、校名、姓名、参赛证编号等泄露参赛队或参赛选手信息的情况，一经发现取消竞赛资格。

4.参赛团队认真核对竞赛工位号，在指定位置就座，不允许携带任何移动存储设备、电子设备及其他资料、用品，违者取消竞赛资格。

5.参赛选手在收到开赛信号前不得启动操作。在竞赛过程中，确因计算机软件或硬件故障，致使操作无法继续的，经裁判长确认，予以启用备用计算机。

6.参赛选手需及时保存工作记录及答题。对于因各种原因造成的数据丢失，由参赛选手自行负责。

7.比赛时间到，比赛即结束，参赛选手应全体起立，停止操作。将资料和工具整齐摆放在操作平台上，经工作人员清点后可离开赛场，离开赛场时不得带走任何资料。

8.竞赛过程中，严重操作失误或安全事故不能进行比赛的（例如因操作原因发生短路导致赛场断电的、造成设备不能正常工作的），现场裁判有权中止该参赛队比赛。

9.参赛选手在比赛中如遇非人为因素造成的设备故障，经裁判确认后，可向裁判长申请补足排除故障的时间。

10.参赛选手不得因各种原因提前结束比赛。如确因不可抗因素需要离开赛场的，须向现场裁判举手示意，经裁判长许可并完成记录后，方可离开。凡在竞赛期间内提前离开的选手，不得返回赛场。

11.竞赛操作结束后，参赛选手需要根据任务书要求，将相关成果文件拷贝至 U 盘，填写结束比赛相关确认文件，并由参赛队长签字确认（竞赛工位号）。因参赛选手未能按要求将相应的文档等拷贝至 U 盘的，竞赛成绩计为零分。

12.在竞赛期间，未经执委会批准，参赛选手不得接受其他单位和个人进行的与竞赛内容相关的采访。参赛选手不得将竞赛的相关信息私自公布。

13.符合下列情形之一的参赛选手，经裁判组裁定后中止其竞赛:

（1）不服从裁判员/监考员管理、扰乱赛场秩序、干扰其他参赛选手比赛，裁判员应提出警告。二次警告后无效，或情节特别严重，造成竞赛中止的，经裁判长确认，中止比赛，并取消竞赛资格和竞赛成绩。

（2）竞赛过程中，由于选手人为造成计算机、仪器设备及工具等严重损坏，负责赔偿其损失，并由裁判组裁定其竞赛结束与否、是否保留竞赛资格、是否累计其有效竞赛成绩。

（3）竞赛过程中，产生重大安全事故、或有产生重大安全事故隐患，经裁判员提示没有采取措施的，裁判员可暂停其竞赛，由裁判组裁定其竞赛结束，保留竞赛资格和有效竞赛成绩。

背景描述

当今时代，数据正在迅速膨胀并变大，一天之中，互联网产生的全部内容可以达到EB级别，能够轻松刻满1.68亿张光盘。在商业、经济及其它领域中，决策将日益基于数据和分析而作出，而并非基于经验和直觉。那么，要怎样基于大数据做出正确的决策呢？大数据首先需要解决的问题就是数据存储的问题，由于数据量非常之大，想通过传统单一节点来存储显得力不从心，搭建分布式的文件存储系统成为一个完美的解决方案。解决了数据存储的问题，我们需要从数据中提取有用信息，通过数据分析手段让数据发挥出真正的价值。但往往采集的原始数据中包含了一些无用数据以及噪声数据，如果直接基于这些脏数据进行分析，往往会让分析结果产生偏差甚至错误，从而造成决策上的失准。因此，我们有必要对这些原始数据进行清洗，以保证其数据准确性、完整性和可用性，提高数据的质量。在解决脏数据的困扰后，我们需要采取各种数据分析手段，提取数据中的价值，得到可靠的结果，并以图表等直观的方式将分析结果进行展现。然后从业务层面对分析结果进行分析和解释，从而指引我们做出正确的决策，真正获取“数据财富”。

气候变化正在迅速地改变地球。随着全球气温不断升高、海平面上升、极端天气事件频繁发生，人们对于地球的未来更加担忧。为了更好地了解气候变化的趋势、预测未来天气趋势，指引相关部门尽早做出举措以应对气候变化，保护人类赖以生存的家园，你的团队将运用大数据技术对天气数据进行分析及决策。搭建大数据平台集群环境以应对海量天气数据的存储，结合数据库的毫秒级的响应，为天气决策系统提供数据存储及查询保障。通过数据清洗技术，去除数据中的噪音，提高数据质量。通过数据标注技术，结合业务认知，对数据进行分类标注，为后续通过人工智能算法模型决策奠定基础。通过各种数据分析技术，让看似杂乱无章的数据，变得灵动，找出天气变化的内在规律。通过数据可视化技术，让数据分析结果及天气变化规律以一种最为直观的方式呈现。最后从业务层面对天气数据分析结果进行分析及解释，使气象学家更好的了解气候变化，并做出精准决策应对气候问题。你们作为该大数据小组的技术人员，请按照下面任务完成本次工作。

模块一：平台搭建与运维

## 任务一：大数据平台搭建（10分）

### **子任务一：Hadoop完全分布式集群搭建（6分）**

本子任务需要使用root用户完成相关配置，安装Hadoop需要配置前置环境。**命令中要求必须使用绝对路径**，具体要求如下:

（1）在master节点将/usr/local/src目录下的hadoop-3.1.3.tar.gz包解压到/opt路径下，将完整命令截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

（2）在master节点上修改环境变量文件，添加Hadoop环境变量，将环境变量配置内容截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

（3）在master节点上完成Hadoop配置文件的修改，集群中具体角色分配参看表2。完成配置文件修改后，将master相关配置文件分发到slave1、slave2节点，将分发命令和结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

表2：Hadoop集群角色分配

|  |  |  |
| --- | --- | --- |
| **主机名** | **HDFS角色** | **yarn角色** |
| master | namenode、datanode | nodemanager |
| slave1 | secondaryNamenode、datanode | nodemanager |
| slave2 | datanode | resourceManager、nodemanager |

（4）在master节点上初始化Hadoop环境namenode，将初始化命令及初始化结果（截取初始化结果日志最后20行即可）截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

（5）启动Hadoop集群，使用jps查看master节点、slave1节点、slave2节点的进程，将查看结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下。

### **子任务二：Spark standalone安装配置（4分）**

本子任务需要使用root用户完成相关配置，具体要求如下：

（1）在master节点将/usr/local/src目录下的spark-3.1.1-bin-hadoop3.2.tgz安装包解压到/opt路径，并把解压后的spark-3.1.1-bin-hadoop3.2文件夹更名为spark-3.1.1，将更名的完整命令及结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

（2）在master节点修改环境变量文件，设置Spark环境变量，将环境变量配置内容截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

（3）在master节点上将/opt/spark-3.1.1/conf目录下的spark-env.sh.template 文件更名为 spark-env.sh，在更名后的 spark-env.sh文件中进行修改，需要指定standalone模式运行时的Master进程运行在 master节点上，指定Master进程内部通信端口号为7077，将修改的内容截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

（4）在 master 节点上面修改/opt/spark-3.1.1 /conf/workers.template文件名为workers后，在/opt/ spark-3.1.1/conf/workers文件中指定standalone模式运行时Worker进程需要分别在 master、slave1、slave2节点上运行，将修改的内容截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

（5）在master节点上面将配置的Spark环境变量文件及Spark解压包拷贝到slave1、slave2节点的 /opt 路径下，将命令和结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

（6）启动Spark集群，使用jps查看master节点、slave1节点、slave2节点的进程，将查看结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下。

## 任务二：数据库配置维护（20分）

本次任务涉及的数据库和数据表信息如下：

表3：数据库及数据表

|  |  |  |
| --- | --- | --- |
| 数据库 | 数据表 | 备注 |
| WeatherDB | province\_info | 省份表 |
| city\_info | 城市表 |
| weather\_day | 天气日表 |
| weather\_month | 天气月表 |

表4：province\_info表

|  |  |  |  |
| --- | --- | --- | --- |
| 表名 | 列名 | 数据类型 | 备注 |
| province\_info | province\_id | int | 省份ID，  范围在1001到1034之间 |
| province\_name | varchar | 省份名称 |

表5：city\_info表

|  |  |  |  |
| --- | --- | --- | --- |
| 表名 | 列名 | 数据类型 | 备注 |
| city\_info | id | int | 城市ID |
| province\_id | int | 省份ID，范围在1001到1034之间 |
| city\_name | varchar | 城市名称 |
| zip\_code | varchar | 邮编 |
| level | int | 城市的级别，可从以下范围中选择：  1：省级  2：市级  3：县级 |
| climate | varchar | 气候条件 |

表6：weather\_day表

|  |  |  |  |
| --- | --- | --- | --- |
| 表名 | 列名 | 数据类型 | 备注 |
| weather\_day | city\_id | int | 城市ID |
| city\_name | varchar | 城市名称 |
| hightest\_temp | int | 最高气温 |
| lowest\_temp | int | 最低气温 |
| weather | varchar | 天气 |
| day | varchar | 日期，形式为年-月-日，比如：2022-04-25 |
| wind\_direction | varchar | 风向 |
| wind\_level | varchar | 风力等级 |
| weekday | varchar | 星期几 |

表7：weather\_month表

|  |  |  |  |
| --- | --- | --- | --- |
| 表名 | 列名 | 数据类型 | 备注 |
| weather\_month | city\_id | int | 城市ID |
| city\_name | varchar | 城市名称 |
| month | varchar | 日期，形式为年-月，比如：2022-04 |
| avg\_high\_temp | int | 平均最高温度 |
| avg\_low\_temp | int | 平均最低温度 |
| extreme\_high\_temp | int | 极端最高温度 |
| extreme\_low\_temp | int | 极端最低温度 |
| avg\_air\_quality | int | 平均空气质量 |
| best\_air\_quality | int | 最佳空气质量 |
| worst\_air\_quality | int | 最差空气质量 |
| best\_air\_date | varchar | 最佳空气质量日期，形式为年-月-日，比如：2022-04-25 |
| worst\_air\_date | varchar | 最差空气质量日期，形式为年-月-日，比如：2022-04-25 |

### **子任务一：数据表与数据管理（6分）**

本子任务需要使用MySQL数据库系统完成关于天气信息的建库、建表、数据的插入、数据表的管理等操作。具体要求如下：

1. 在 MySQL 数据库中创建名字为WeatherTestDB 的数据库，将命令及运行结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；
2. 在 MySQL 数据库中切换到 WeatherTestDB 数据库，将命令及运行结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；
3. 在MySQL数据库中删除名字为 WeatherTestDB 数据库，查看所有数据库，将运行结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；
4. 查询WeatherDB数据库中所有存在表，将运行结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；
5. 在MySQL数据库的WeatherDB 库中，创建一个名为weather\_base\_info\_table的数据表，包含的字段见表8所示，要求所有字段非空，数据库引擎为InnoDB，默认字符集为utf8。查看刚才创建的 weather\_base\_info\_table 表结构，将结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；

表8：字段说明

|  |  |  |
| --- | --- | --- |
| **列名** | **数据类型** | **备注** |
| city\_id | int | 城市ID |
| city\_name | varchar | 城市名称 |
| month | varchar | 日期，形式为年-月，比如：2022-04 |
| avg\_high\_temp | int | 平均最高温度 |
| avg\_low\_temp | int | 平均最低温度 |
| extreme\_high\_temp | int | 极端最高温度 |
| extreme\_low\_temp | int | 极端最低温度 |

1. 使用SQL命令修改 weather\_base\_info\_table 表中city\_id字段的类型和长度为varchar(255)，将结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；
2. 使用SQL命令删除 weather\_base\_info\_table 表中的city\_id字段，将结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下；
3. 在 MySQL 数据库的 WeatherDB 库中，要求从weather\_month查询数据后，插入到weather\_base\_ info\_table表中（weather\_base\_info\_table表中的字段在 weather\_month表中存在，都应该被查询），查询 weather\_base\_info\_table 表中前5条数据，将运行结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx**】中对应的任务序号下。

### 子任务二：维护数据表(14分)

本子任务基于数据库WeatherDB的数据，使用SQL语句对天气相关数据进行数据分析等操作，具体要求如下：

1. 使用SQL命令查看weather\_day表中第1001至第 1010条数据（查询结果只显示第1001至第1010条数据），将结果截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下；
2. 在weather\_day表中使用 SQL 语句查询鞍山城市 2022 年晴天有多少天（天气为晴则为晴天），将运行结果截图并粘贴到客户端桌面【**result/模块1答题报告.docx】**中对应的任务序号下；
3. 统计天气气候包含季风性气候的城市（例如：温带季风性气候、亚热带季风性气候都算包含季风性气候），将运行结果截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下；

查询字段相关说明：

|  |  |
| --- | --- |
| **字段名字** | **字段含义** |
| city\_name | 城市名称 |
| climate | 气候条件 |

1. 在 weather\_month 表中使用 SQL 语句查询重庆每年极端最高温度超过40℃的天气分别有多少天，将运行结果截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下；
2. 在 weather\_month 表中使用 SQL 语句统计 2022 年极端最高温度与极端最低温度温差超过 40℃ 的城市分别有哪些，将运行结果的最后5条数据以及总数据行数截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下；
3. 在 weather\_day 表中使用 SQL 语句统计2022年4月1日这天中温差超过15℃的城市有哪些，将运行结果截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下；

查询字段相关说明：

|  |  |
| --- | --- |
| **字段名字** | **字段含义** |
| city\_id | 城市ID |
| city\_name | 城市名称 |

1. 统计所有城市级别为县级的城市详细信息，将运行结果截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下；

查询字段相关说明：

|  |  |
| --- | --- |
| **字段名字** | **字段含义** |
| city\_name | 城市名称 |
| province\_name | 省份名称 |
| zip\_code | 邮编 |
| climate | 气候条件 |

1. 统计黑龙江全省每年下雪天有多少天（每天的天气中包含雪即为下雪天）,将运行结果截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下；
2. 统计天气晴朗天数最多的前三个省份信息（天气包含晴则为天气晴朗），将运行结果截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下；

查询字段相关说明：

|  |  |
| --- | --- |
| **字段名字** | **字段含义** |
| province\_name | 省份名称 |
| cloudless | 天气晴朗天数 |

1. 统计每个省份中属于亚热带季风气候的城市数量，查询结果按照城市数量从高到低进行排序，将运行结果截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下；

查询字段相关说明：

|  |  |
| --- | --- |
| **字段名字** | **字段含义** |
| province\_name | 省份名称 |
| climate\_num | 省内亚热带季风气候对应的城市数 |

1. 统计大兴安岭在2021年2月~4月所有出现过的天气（查询结果中每个天气之间需要用逗号进行连接），将运行结果截图并粘贴到客户端桌面**【result/模块1答题报告.docx】**中对应的任务序号下。

模块二：数据获取与处理

## 任务一：数据获取与清洗（10分）

### **子任务一：数据获取（5分）**

|  |  |
| --- | --- |
| **字段名** | **字段说明** |
| Column | 字段名称 |
| Null\_count | 当前列缺失值计数 |

统计null\_count.csv 文件中每一列的缺失值个数，并保存为result.csv 文件，result.csv 文件应包括下表中的字段。

查看result.csv文件，将文件内容截图并粘贴到客户端桌面**【result/模块2答题报告.docx】**中对应的任务序号下。

### **子任务二：数据清洗（5分）**

使用Python对数据“day.csv”进行清洗，将清洗后的数据保存。数据字段如下表所示。

|  |  |
| --- | --- |
| **数据字段** | **字段说明** |
| id | 编号 |
| city | 城市名 |
| date | 日期 |
| hightest\_tem | 最高温度 |
| lowest\_tem | 最低温度 |
| weather | 天气 |
| wind | 风力 |

清洗具体要求如下：

1、使用Pandas读取数据。

2、现有的“date”列没有包含星期信息，需要增加一列“week”存放星期数据，例如“星期一”；

3、将温度列的数据类型转换成整型，例如将5℃处理成5，注意，如含有非法字符需自行处理；

4、风力列（wind）包含风向和风力等级，需要将其处理为两列，列名分别为wind\_direction（风向）和wind\_level（风力等级）；

5、处理之后的数据列名包括'id','city','date','week', 'hightest\_tem','lowest\_tem','weather','wind\_direction','wind\_level'，并将列名分别重命名为'ID','城市','日期','星期', '最高温度','最低温度','天气','风向','风力等级'；

6、处理结果存到【deal\_data】目录中“result.csv”文件（需要保存列名）,如果目录不存在，需要自行创建。

7、使用wps将保存的CSV文件打开，将行号为1~21、101~121和最后20行数据进行截图并粘贴到客户端桌面【**result/模块2答题报告.docx**】中对应的任务序号下，**截图需要保留行号**。

## 任务二：数据标注（10）

使用Python对给定天气数据进行标注，并保存标注结果。天气数据集具有多个字段，各字段信息如下表所示。

|  |  |
| --- | --- |
| **数据字段** | **字段说明** |
| city | 城市名 |
| hightest\_tem | 最高温 |
| lowest\_tem | 最低温 |
| weather | 天气 |
| date | 年月日 |
| wind\_direction | 风向 |
| wind\_level | 风力等级 |

具体要求如下：

1、使用Python读取“长春天气信息.xlsx”。

2、在末尾新增一列数据为“当日是否解冻”，若当日最高温大于0，并且风力小于等于2级，打标签为‘是’；否则打标签为‘否’。标记完成后保存到当前目录，文件命名为“annotation.xlsx”,并将数据截图并粘贴到客户端桌面【**result/模块2答题报告.docx**】中对应的任务序号下。

## 任务三：数据统计（15分）

### 子任务一：处理异常值数据（7.5分）

HDFS文件系统中/bigdata/eurasia\_mainland.csv文件存储了欧亚大陆各个国家的灾害数据，数据中有以下内容：

|  |  |
| --- | --- |
| **数据字段** | **字段说明** |
| c\_year | 年份 |
| c\_country | 国家 |
| hazard\_type | 灾害类型 |
| disaster\_subtype | 灾害子类型 |
| area | 区域 |
| disaster\_frequency | 灾害频次 |
| c\_death\_toll | 总死亡人数 |
| c\_people\_affected | 总受灾人数 |
| c\_economic\_loss | 总经济损失 |

编写MapReduce 程序，实现以下功能：清除年份、国家区域为空的数据，将清理后的数据保存到HDFS中/clean\_data目录下，若目录不存在，请自行创建，使用命令查看该文件的大小，将完整命令及结果截图并粘贴到客户端桌面【**result/模块2答题报告.docx**】中对应的任务序号下。

### 子任务二：数据统计（7.5分）

HDFS文件系统中/bigdata/eurasia\_mainland.csv文件存储了欧亚大陆各个国家的灾害数据，数据中有以下内容：

|  |  |
| --- | --- |
| **数据字段** | **字段说明** |
| c\_year | 年份 |
| c\_country | 国家 |
| hazard\_type | 灾害类型 |
| disaster\_subtype | 灾害子类型 |
| area | 区域 |
| disaster\_frequency | 灾害频次 |
| c\_death\_toll | 总死亡人数 |
| c\_people\_affected | 总受灾人数 |
| c\_economic\_loss | 总经济损失 |

编写 MapReduce程序，实现以下功能：统计每个国家不同年份基于灾害类型为气候灾害受损经济最高的国家，并在控制台输出气候灾害受损经济最高的前10个国家，需要打印年份、国家及总经济损失，将输出结果截图并粘贴到客户端桌面【**result/模块2答题报告.docx**】中对应的任务序号下。

模块三：业务分析与可视化

## 任务一：数据可视化（20分）

### 子任务一：基于 Echarts 的数据可视化分析（5分）

气候变化正在迅速地改变地球。随着全球气温不断升高、海平面上升、极端天气事件频繁发生，人们对于地球的未来更加担忧。为了更好地了解气候变化的趋势、预测未来天气趋势，我们创建了“天气数据库”，用于收集、记录、可视化展示来自全球各地的气象数据和天气预报信息。请在“index.html”文件中编写代码实现功能,数据文件名为“chengdu.js”:

1、文件内记录了四川省成都市2021年的天气数据，统计每个月的平均最高气温和每个月的平均最低气温，将统计得到的数据格式转换为Echarts所需的数据格式；

2、将第1步中构建的数据作为输入，通过Echarts绘制柱状图；

3、使用浏览器打开“index.html”文件，然后将渲染结果截图并粘贴到客户端桌面**【result/模块3答题报告.docx】**中对应的任务序号下。

### 子任务二：基于Python进行灾害损失可视化分析（5分）

现有一份关于灾害的数据集“disaster.csv”，列名包含年份、国家、灾害类型、灾害子类型、区域、灾害频次、总死亡人数、总经济损失（千美元）。绘图过程中如有使用中文字体的地方，请统一使用“SimSun”字体。

需要分析并绘制中国各年份死亡人数折线图和中国各年份总经济损失折线图，将两张图作为子图绘制于同一个画布中，画布大小为15\*10，子图布局为2行1列。将**两张折线图**截图并粘贴到客户端桌面**【result/模块3答题报告.docx】**中对应的任务序号下。

具体要求如下：

**（1）分析并绘制中国各年份死亡人数折线图：**

①使用matplotlib库绘图。

②计算中国各年的总死亡人数，单位为“万人”。

③将图形绘制于第一张子图中。

④图像标题为“死亡人数”,字体大小20。

⑤x轴标签为“年份”，字体大小16，y轴标签为“总死亡人数（万人）”，字体大小16。

⑥折线图颜色为“red”，折线图上每个点需要显示出来，点型为实心圆

**（2）分析并绘制中国各年份总经济损失折线图：**

①使用matplotlib库绘图。

②计算中国各年的总经济损失，单位为亿元，假设1美元=7元。

③将图形绘制于第二张子图中。

④图像标题为“经济损失”，字体大小20。

⑤x轴标签为“年份”，字体大小16，y轴标签为“总经济损失（亿元）”，字体大小16。

⑥折线图颜色为“violet”，折线图上每个点需要显示出来，点型为实心圆。

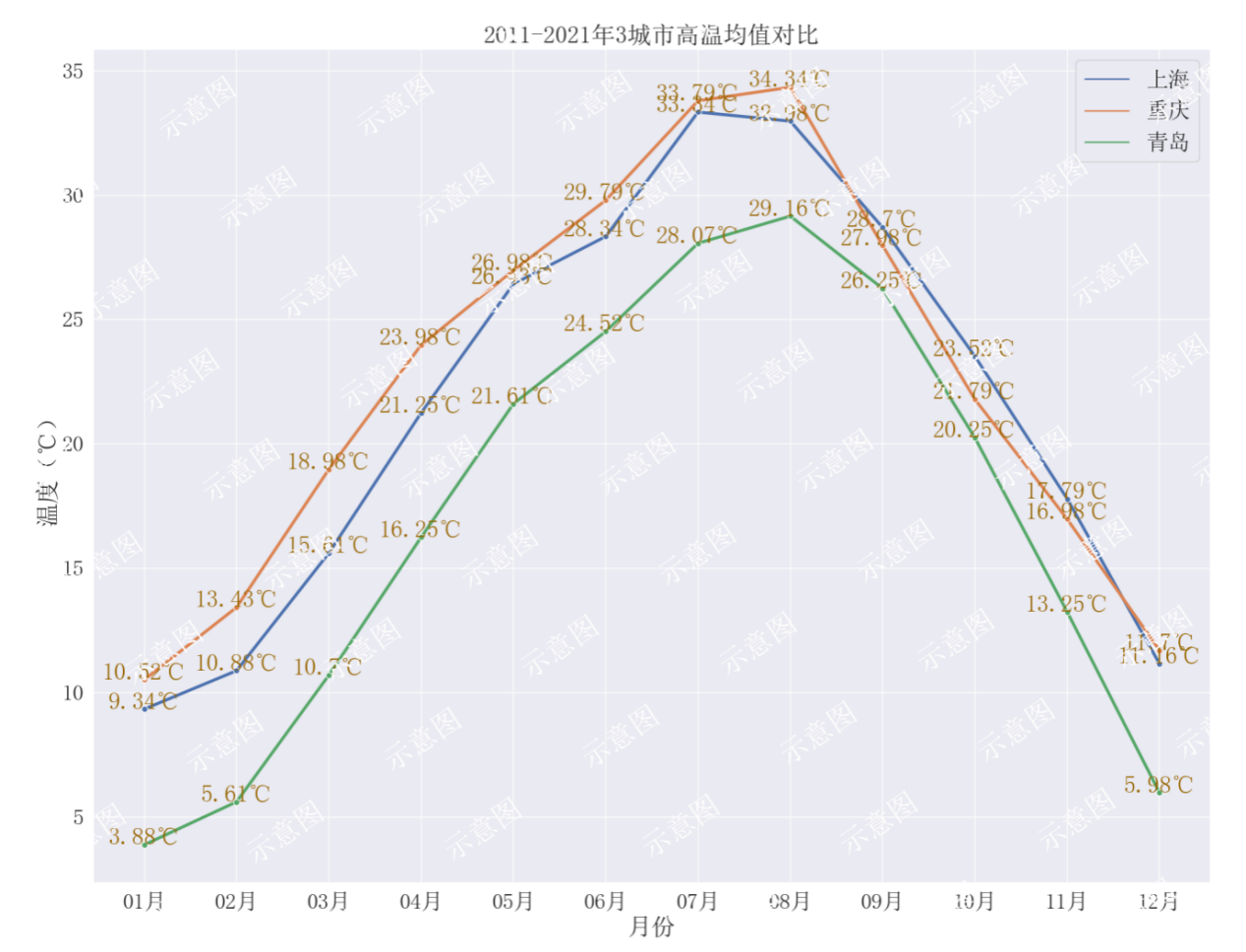
### 子任务三：基于Python进行城市全年温度对比分析（5分）

现有一份关于2011-2022年全国各城市的每月天气数据集，字段说明如下表：

|  |  |
| --- | --- |
| **列名** | **字段说明** |
| city | 城市 |
| month | 年-月 |
| avg\_high\_tem | 平均高温 |
| avg\_low\_tem | 平均低温 |
| extreme\_high\_tem | 极端高温 |
| extreme\_low\_tem | 极端低温 |
| avg\_air\_quality | 平均空气质量指数 |
| best\_air | 最好空气指数 |
| best\_air\_date | 最好空气指数日期 |
| worst\_air | 最差空气指数 |
| worst\_air\_date | 最差空气指数日期 |

请编写代码，从数据集“clean\_month.csv”中取出重庆、上海、青岛三市在2011-2021年的数据；计算出三市在这10年中所有月份平均高温的平均值，并使用该指标绘制出三个城市一年中的高温变化图来分析三市的气温情况，具体绘图要求如下：

1. 使用Seaborn绘制以上指标的折线图，三个城市的折线图绘制在同一张图中，主题设置为“darkgrid”，字体为“SimSun”，字体缩放因子设置为2；
2. 折线图中需要添加数据标记，形状为圆点，线宽为3，并通过对城市分组给三条曲线设置不同的颜色，采用分组后Seaborn的默认配色；
3. 折线图需显示数据标签，颜色为“#996600”，数据标签中需要带有单位（℃）并且保留两位小数；
4. 图像标题为“2011-2021年3城市高温均值对比”，横轴标签为“月份”，纵轴标签为“温度（℃）”；
5. 横轴刻度为月份，刻度标签为“01月 02月 ... 12月”;
6. 图像右上角显示图例；
7. 绘图过程中如有使用中文字体的地方，请统一使用“SimSun”字体。
8. 绘制完成后将图片粘贴到客户端桌面**【result/模块3答题报告.docx】**中对应的任务序号下。



结果参考示意图（图中数据请以实际计算结果为准）

### 子任务四：基于Python实现历史最高温城市排名分析（5分）

现有一份关于2011-2022年全国各城市的每日天气数据集，字段说明如下表：

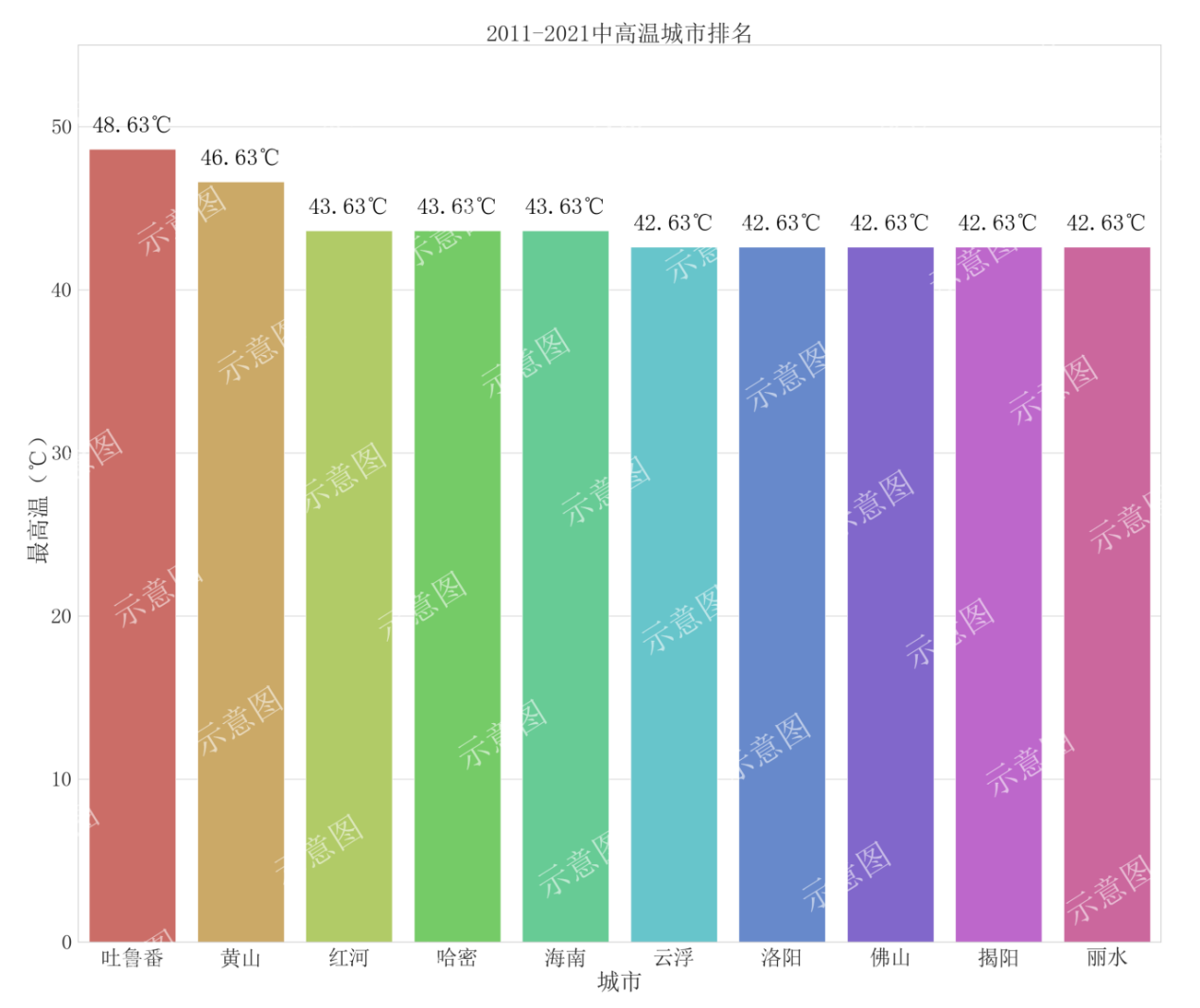
|  |  |
| --- | --- |
| **列名** | **字段说明** |
| city | 城市 |
| hightest\_tem | 最高气温 |
| lowest\_tem | 最低气温 |
| weather | 天气 |
| date | 日期 |
| wind\_direction | 风向 |
| wind\_level | 风力等级 |
| month\_day | 月-日 |
| weekday | 星期 |

请编写代码实现功能，数据集为“clean\_day.csv”。

绘图过程中如有使用中文字体的地方，请统一使用“SimSun”字体。

从数据集中取出2011-2021年的数据，计算出每个城市在这10年中出现的最高温度，然后将该指标排名前10的城市数据取出并绘制出这些城市最高温统计图进行分析，具体绘图要求如下：

1. 使用Seaborn绘制出以上指标的柱状图，主题设置为“whitegrid”，字体为“SimSun”，字体缩放因子设置为3；
2. 柱状图颜色设置为Seaborn中调色板“hls”的默认颜色；
3. 柱状图上需要显示数据标签，数据标签中需要带有单位（℃），颜色为黑色；
4. 设置图像标题为“2011-2021中高温城市排名”；
5. 横轴标签为“城市”，纵轴标签为“最高温（℃）”;
6. 横轴的刻度标签为各城市的名字；
7. 纵轴的刻度范围是（0,55）。
8. 绘制完成后将图片粘贴到客户端桌面**【result/模块3答题报告.docx】**中对应的任务序号下。

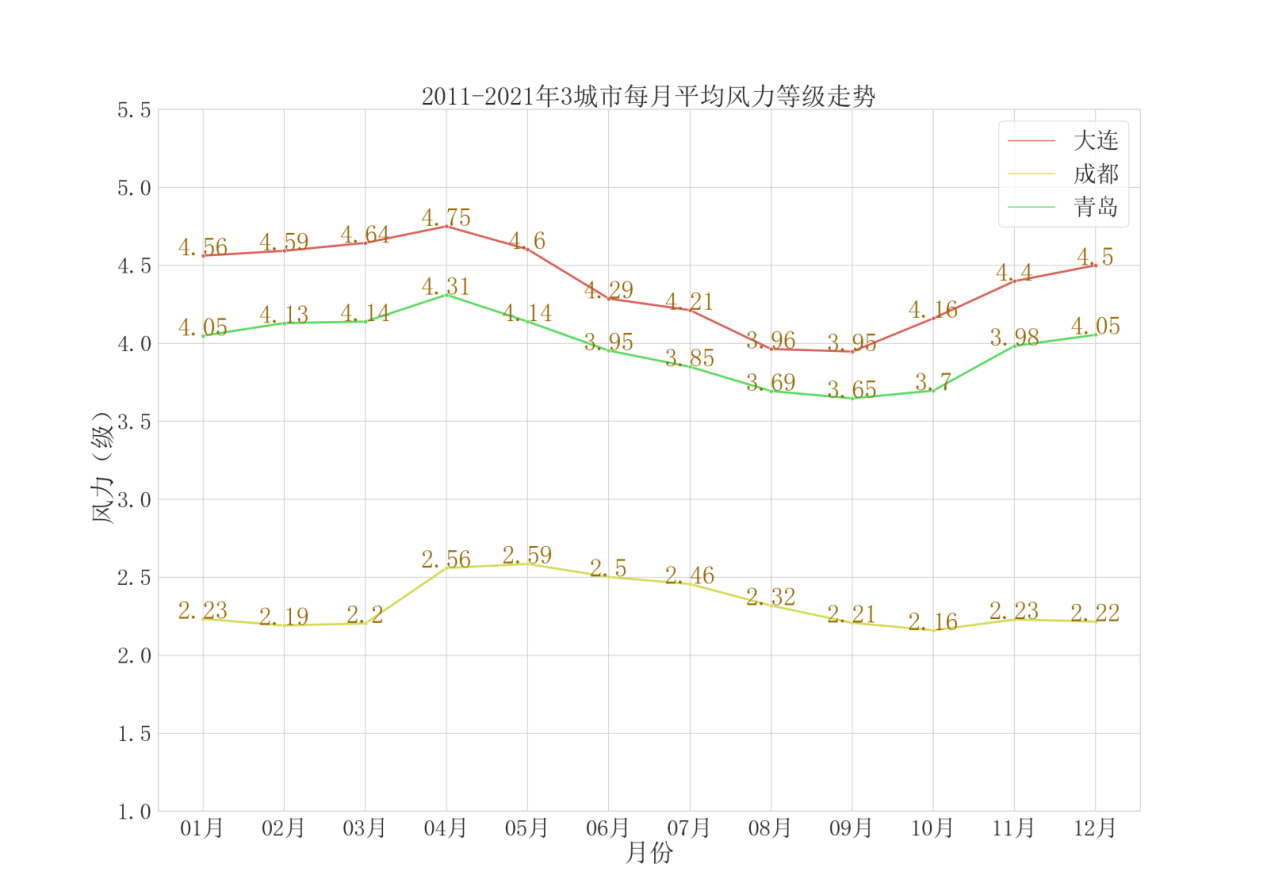


结果参考示意图（图中数据请以实际计算结果为准）

## 任务二：业务分析（10分）

现有一份关于天气的可视化分析结果，请对结果进行业务分析，并给出可解释型结论。

下图是3个城市10年来每月平均风力等级走势图，请对该图以及图表中的数据进行分析并给出合理解释。



风力等级走势图