ZZ052-大数据应用与服务赛项试题 05

一、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，想通过传统单一的节点的存储显得力不从心，搭建分布式的文件存储系统成为了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、

1

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。

二、模块一：平台搭建与运维

（一）任务一：大数据平台搭建

1.子任务一：Zookeeper 集群安装配置

本任务需要使用 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 apache-zookeeper-3.5.7-bin.tar.gz 包解压到/opt 路径下，

将完整命令截图粘贴到对应答题报告中;

（2）在 master 节点上面将配置的 Zookeeper 环境变量文件及 Zookeeper 解压包拷贝到 slave1、slave2 节点，将命令和结果截图粘贴到对应答题报告中;

（3）将 slave1 节点上面/opt/zookeeper-3.5.7/data 目录下的 myid 文件内容修改为 2，将 slave2 节点上面/opt/zookeeper-3.5.7/data 目录下的myid文件内容修改为3，将命令和结果截图粘贴到对应答题报告中;

（4）在 master 节点、slave1 节点、slave2 节点分别启动 zookeeper，将命令和结果截图粘贴到对应答题报告中;

2.子任务二: Hadoop 完全分布式集群搭建

本任务需要使用 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，将完整命令截图粘贴到对应答题报告中;

（2）在 master 节点修改/root/.bash_profile 文件，设置 Hadoop 环境变量，将环境变量配置内容截图粘贴到对应答题报告中;

（3）在 master 节点上面修改 Hadoop 的配置文件 hdfs-site.xml，需要在该文件中指定上传的文件的副本数

3

为 3，将修改的内容截图粘贴到对应答题报告中;

（4）在 master 节点上面修改 Hadoop 的配置文件 yarn-site.xml，需要在该文件中指定 YARN 的 ResourceManager 的地址为 slave2，将修改的内容截图粘贴到对应答题报告中;

（5）在 master 节点上面将配置的 Hadoop 环境变量文件及 Hadoop 解压包拷贝到 slave1、slave2 节点，将命令和结果截图粘贴到对应答题报告中;

(6)在 master 节点上面初始化 Hadoop 环境 namenode，将初始化命令及初始化结果截图粘贴到对应答题报告中;

（7）启动 Hadoop 集群（在 master 节点启动 hdfs，在 slave2 节点启动 yarn),使用 jps 查看 master 节点、slave1 节点、slave2 节点的进程，将查看结果截图粘贴到对应答题报告中。

3.子任务三：Hive 安装配置

本任务需要使用 root 用户完成相关配置,已安装 Hadoop 及需要配置前置环境，具体要求如下:

（1）在 master 节点将/usr/local/src 目录下的 apache-hive-3.1.2-bin.tar.gz 安装包解压到/opt 路径下，将完整命令截图粘贴到对应答题报告中;

（2）修改 hive-site.xml 配置文件，将 MySQL 数据库作为 Hive 元数据库。将配置 Hive 元数据库的相关内容截图

4

粘贴到对应答题报告中;

（3）将/usr/local/src 目录下的 MySQL 数据库 JDBC 驱动 mysql-connector-java-5.1.27-bin.jar 拷贝到 Hive 安装目录的 lib 文件夹下，将完整命令截图粘贴到对应答题报告中;

（4）初始化 Hive 元数据库，将初始化命令及结果截图粘贴到对应答题报告中;

（5）启动 Hive，将命令输出结果截图粘贴到对应答题报告中。

（二）任务二: 数据库配置维护

1.子任务一: 数据库配置

MySQL 是一个多用户数据库，具有功能强大的访问控制系统，可以为不同用户指定不同权限。root 用户是超级管理员，拥有所有权限，包括创建用户、删除用户和修改用户密码等管理权限。

为了实际项目的需要，可以定义不同的用户角色，并为不同的角色赋予不同的操作权限。当用户访问数据库时，需要先验证该用户是否为合法用户，再约束该用户只能在被赋予的权限范围内操作。具体任务要求如下:

（1）为本地主机数据库创建一个名为 staff 的用户，密码为 staff123456，将完整命令及结果截图粘贴到对应答题报告中;

5

（2）查看用户，确认有刚才创建的 staff 用户，将完整命令及结果截图粘贴到对应答题报告中;

（3）将用户名 staff 修改为 newstaff，将完整命令及结果截图粘贴到对应答题报告中;

（4）授予用户 newstaff 对 WeatherDB 数据库中 weather_month 表的查询、插入、删除权限，将完整命令及结果截图粘贴到对应答题报告中（MySQL 数据库中已创建好 WeatherDB 数据库，如果不存在则需要自己建库并导入数据，提供的几个 sql 是数据源文件）;

（5）使用新用户 newstaff 登录 MySQL 数据库，查看是否有 WeatherDB 数据库，并查看 WeatherDB 数据库下有哪些表，将完整命令及结果截图粘贴到对应答题报告中;

（6）删除 newstaff 的用户，并确认是否已经删除 newstaff 用户，将完整命令及结果截图粘贴到对应答题报告中;

2.子任务二：数据表与数据管理

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、组织和记录来自全球各地的气象数据和天气预报信息。它的作用不仅仅是记录过去的天气情况，更是提供了一个全球性、长期性的

6

气候趋势预测工具，使气象学家和气候学家能够更好地了解气候变化的趋势，从而采取适当的措施应对未来的气候变化。本任务的具体要求如下:

（1）在 MySQL 数据库的 WeatherDB 库中，创建一个名为 province_city 的数据表，数据库引擎为 InnoDB，默认字符集为 utf8。将完整命令及运行结果截图粘贴到对应答题报告中；包含的字段如下:

表 1　数据表字段

列名	数据类型	说明
city_id	int	城市 ID: 主键，自增，非空
city_name	varchar	城市名称
province_name	varchar	省份名称
climate	varchar	气候条件

（2）使用 SQL 命令修改 province_city 表中 climate 列的列名为 climate_new，将完整命令及结果截图粘贴到对应答题报告中;

（3）使用 SQL 命令给 province_city 表增加一个字段 zip_code（代表邮编），字段类型应符合实际意义，将完整命令及结果截图粘贴到对应答题报告中;

（4）使用 SQL 语句给 province_city 表中插入一条数据，数据的具体信息如下: 城市 ID 为 10001、城市名称为阆中市、省份名称为四川省、气候条件为亚热带季风气候、邮编为 637400。将完整 SQL 语句及运行结果截图粘贴到对应答题报告中。

7

（5）使用 SQL 语句批量给 province_city 表中插入三条数据，将完整 SQL 语句及运行结果截图粘贴到对应答题报告中。数据的具体信息如下:

表 2　数据信息

城市 ID	城市名称	省份名称	气候条件	邮编
10002	江油市	四川省	亚热带季风性湿润气候	621700
10003	灯塔市	辽宁省	北温带大陆性气候	111300
10004	玉环市	浙江省	暖温带大陆性季风气候	317610

（6）使用 SQL 语句修改 province_city 表中城市 ID 为 10004 的城市信息，将气候修改为亚热带海洋性季风气候，将邮编修改为 317600。将完整 SQL 语句及运行结果截图粘贴到对应答题报告中;

3.子任务三：维护数据表

SQL 作为一种全球通用的语言，任何人都可以学习使用。虽然看起来很复杂，除开特定数据库系统专用的 SQL 命令，其它基本上不需要任何事先的知识，而且命令通常比较少。SQL 能够快速的查询和统计大量数据，发现数据的趋势和数据之间的关系。SQL 是一种与数据库打交道的标准语言，熟练地使用 SQL 可以确保每个使用数据库的人都会使用相同的命令，使得开发人员更容易创建与多个数据库一起工作的应用程序。本任务的具体要求如下:

（1）使用 SQL 命令查看 weather_month 表中第 20000 至第 20100 条数据（查询结果只显示第 20000 至第 20100 条数据），将完整 SQL 语句和运行结果的后 5 条数据以及总数

8

据行数截图粘贴到对应答题报告中;

（2）使用 SQL 语句分别查询四川省、广东省、浙江省下面有哪些城市，输出省份 id、省份名称、城市 id、城市名称、邮编、城市等级、气候条件。将完整 SQL 语句和各省份相关城市查询结果的后 5 条数据以及总数据行数截图粘贴到对应答题报告中;

（3）使用 SQL 语句查询 weather_month 表，筛选出哪些城市在 2018 年的月度温差大于等于 5 度的（平均最高气温-平均最低气温），输出城市 id、城市名称、日期、平均最高气温、平均最低气温。将完整 SQL 语句和运行结果的后 5 条数据以及总数据行数截图粘贴到对应答题报告中;

（4）使用 SQL 语句查询 weather_day 表中各个城市每年的最高温度和最低温度分别是多少度，输出城市 id、城市名称、日期（格式为年）、最高温度、最低温度。将完整 SQL 语句和运行结果的后 5 条数据以及总数据行数截图粘贴到对应答题报告中。

三、模块二：数据获取与处理

（一）任务一：数据获取与清洗

1.子任务一：数据获取

读取已经爬取到的 distribution.csv 数据文件，根据表头字段名统计每一列缺失值个数，并保存到代码同级目录下的 result_1.csv 文件中，result_1.csv 文件应包括如下

9

字段:

表 3　文件包含字段

字段名	字段说明
Column	字段名称
Null_count	当前列缺失值计数

将 result_1.csv 文件内容截图粘贴至结果文件中。

2.子任务二: HDFS 文件上传下载

本任务需要使用 Hadoop，HDFS 命令，已安装 Hadoop 及需要配置前置环境，具体要求如下:

(1)在 master 节点 HDFS 根目录下创建 student 目录，将完整命令及结果截图粘贴到对应答题报告中;

(2) 使用命令将/root/clean_month.csv 文件上传到 HDFS 文件系统的/student 目录下，将完整命令及结果截图粘贴到对应答题报告中;

(3) 使用命令查看 HDFS 中/student/clean_month.csv 文件的后 5 条数据，将完整命令及结果截图粘贴到对应答题报告中;

(4)使用命令查看 HDFS 中/student 目录下每个文件所占磁盘空间，人性化显示文件大小，将完整命令及结果截图粘贴到对应答题报告中。

(二) 任务二: 数据标注

数据标注是人工智能产业的基础，是机器感知现实世界

10

的起点。随着 AI 行业的蓬勃发展，对数据的需求呈井喷式增长，从某种程度上来说，没有经过标注的数据就是无用数据。数据标注的越精准、对算法模型训练的效果就越好。大部分算法在拥有足够多普通标注数据的情况下，能够将准确率提升到 95%，但从 95%再提升到 99%甚至 99.9%，就需要大量高质量的标注数据。

本任务是使用 Python 对给定的天气数据进行标注，并进行持久化存储。

请编写代码实现功能，原始数据为“鞍山.xlsx”,字段信息如下表所示:

表 4 天气数据集字段说明表

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	日期（年-月-日）
weekday	星期几

使用 Pandas 读取数据后，将数据按日期列升序排列，在末尾新增一列数据为“是否适合出行游玩”，若当日为周六周日，气温大于等于 18 度小于等于 30 度，并且不下雨，打标签为‘是’；否则打标签为‘否’。标记完成后将标记数据集保存到项目下的“taged_data.xlsx”的文件中，并使用 WPS 打开数据截图粘贴到答题报告对应位置。

11

（三）任务三：数据统计

1.子任务一：处理异常值数据

HDFS 文件系统中/student/clean_month.csv 文件存储了各个城市每月的天气数据，数据中有以下内容:

表 5　天气数据内容

city	城市
month	月份
avg_high_temp	平均高温
avg_low_tem	平均低温
extreme_high_tem	最高温度
extreme_low_tem	最低温度
avg_air_quality	平均空气指数
best_air	最佳空气指数
best_air_date	最佳空气日期
worst_air	最差空气指数
worst_air_date	最差空气日期

编写 MapReduce 程序，实现以下功能：清除月份为空的数据，将清理后的数据输出到 HDFS 中/clean 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粘贴到对应答题报告中。

2.子任务二：数据统计

HDFS 文件系统中/student/clean_month.csv 文件存储了各个城市每月的天气数据，数据中有以下内容:

12

表 6　天气数据内容

city	城市
month	月份
avg_high_temp	平均高温
avg_low_tem	平均低温
extreme_high_tem	最高温度
extreme_low_tem	最低温度
avg_air_quality	平均空气指数
best_air	最佳空气指数
best_air_date	最佳空气日期
worst_air	最差空气指数
worst_air_date	最差空气日期

编写 MapReduce 程序，实现以下功能：统计每个城市最高温度，并在控制台输出温度最高的 5 个城市以及最高的温度，将输出结果截图粘贴到对应答题报告中。

四、模块三：业务分析与可视化

（一）任务一：数据可视化

1.子任务一：基于 Echarts 的数据可视化分析

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

13

“chengdu.js”：

（1）文件内记录了四川省成都市 2021 年的天气数据，其中天气类型一共有多云、霾、晴、小雨等 11 种，统计每种天气类型的出现次数，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 绘制饼图;

（3）使用浏览器打开“index.html”文件，然后将渲染结果截图粘贴到答题报告对应位置

2.子任务二：基于 Excel 进行数据分析与可视化

气象观测数据是制作天气预报和预警的基础，对研究气候变化和指定应对政策具有重要作用。通过长期观测和分析，可以研究气候的变化趋势和规律。使用 Excel 工具对近几年天气数据进行分析与可视化，掌握使用 Excel 进行数据分析应用。

近几年天气的数据在“E_weather.csv”中，数据表中记录 2011 年至 2022 年各城市的天气信息,包含城市、月份、平均高温、平均低温、平均空气质量指数、最佳空气质量指数、最佳空气质量日期、最差空气质量指数、最差空气质量日期 9 列，其中温度相关数据的单位：摄氏度（℃）。使用 Excel 打开“E_weather.csv”文件，对数据进行分析与可视化，具体要求如下:

14

（1）将 csv 数据表读取为 Excel 数据表，并分析每个数据字段类型，使字段能进行统计、计算等（参与计算的单元格中的值，如果存在字符，需要把字符替换为空，例如：单元格的值为：1a23a<--,替换之后的单元格的值为 123）。

（2）数据中每个城市每月的数据应该只有一份，数据中包含重复数据，请过滤掉重复日期的数据，并对数据根据日期升序进行排序。

（3）对数据进行统计分析，绘制出阿克苏、北京、成都、长沙 4 个城市 2011 年到 2020 年 4 个季度平均低温【簇状柱形图】。设置要求如下：

- 设置图表标题为【2011-2020 年季度平均低温】，标题居中显示。
- 横坐标标签显示为城市名，合理设置标签位置，使其显示在轴的下方。
- 纵坐标显示平均低温，标题显示为“平均低温（℃）”。
- 图例显示每个季度，并置于图像底部。
- 显示数据标签并保留两位小数，置于柱子顶部或者底部，如果低于 0℃，数据显示为红色，如下图所示：

15

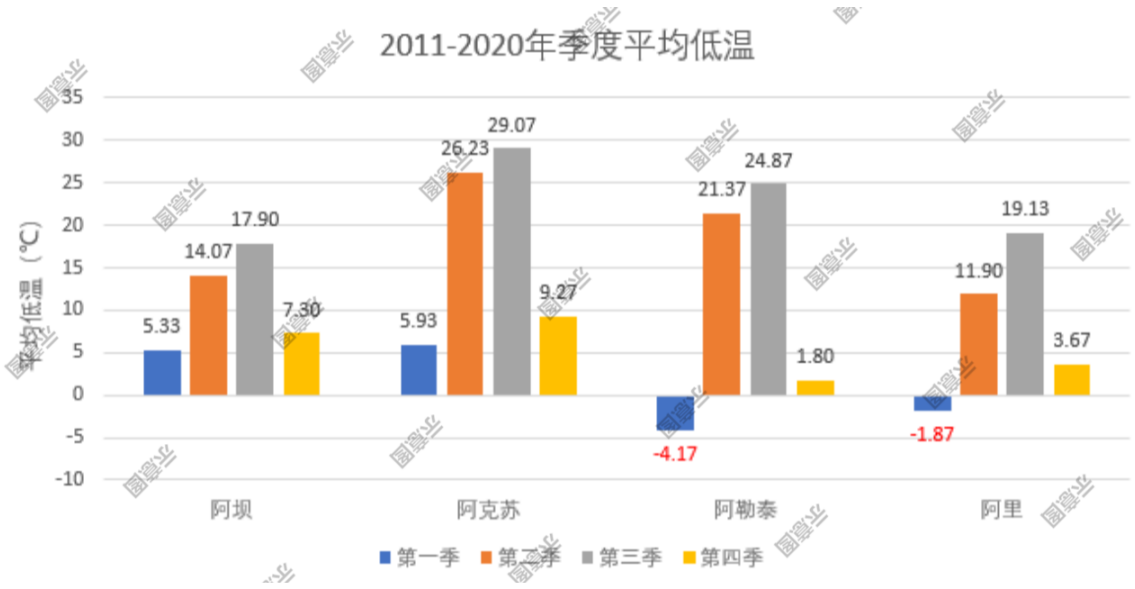

图 1 结果示意图

（4）将绘制完成后的图表进行截图，粘贴到竞赛平台答题报告上对应位置。

3.子任务三：基于 Python 实现我国全年气温变化情况分析

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：

表 7 每月天气数据集说明表

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

16

worst_air_date	最差空气指数日期

请编写代码实现功能，数据集为“clean_month.csv”。具体任务要求如下:

（1）使用 Seaborn 绘制出面积图，两个指标绘制在同一张图中，主题设置为“darkgrid”，字体为“SimSun”，字体缩放因子设置为 2;

（2）平均高温面积图的颜色为“#CC3300”，透明度设置为 0.4，图例标签为“平均最高温”;

（3）平均低温面积图的颜色为“#339999”，透明度为 0.7，图例标签为“平均最低温”;

（4）给平均高温面积图添加边缘线，边缘线颜色为“#CC3300”，线宽为 2，并给边缘线添加标记，形状为圆点;

（5）给平均低温面积图添加边缘线，边缘线颜色为“#339999”，线宽为 2，并给边缘线添加标记，形状为圆点;

（6）给面积图显示数据标签，颜色为“#996600”，数据标签中需要带有单位（℃）并且保留两位小数;

（7）图片标题为“2011-2021 年月份气温均值”;

（8）横轴标签为“月份”，纵轴标签为“温度（℃）”;

(9)横轴的刻度标签显示为“01 月 02 月 ... 12 月”;

（10）设置纵轴刻度范围为（-5，35）;

（11）在图像的右上角显示图例;

（12）纵轴中如有负数需要显示负号。

17

（13）绘制完成后将图片粘贴到答题报告对应位置。

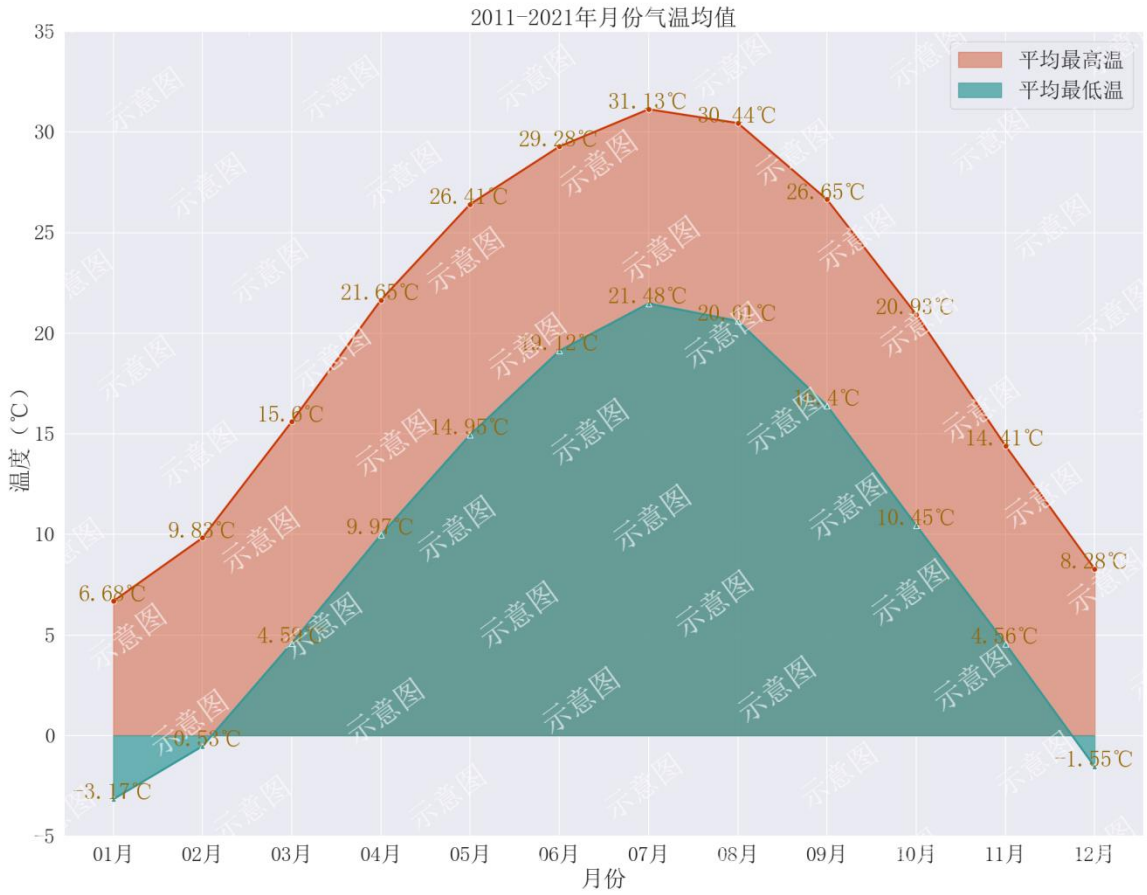

图 2 结果示意图（图中数据请以实际计算结果为准）

（二）任务二：业务分析

我们通过数据清洗、数据分析、数据可视化得到可视化结果，是为了服务于具体的业务场景，解决业务痛点问题，找出业务背后的逻辑关系和根源，从而能更好的服务于我们。这就要求对数据分析结果做出科学合理的解释、得出正确的结论，从而指引我们纠正、优化业务方向，让数据真正的产生价值。

下图是 3 个城市 10 年来每月平均风力等级走势图，请

18

对该图以及图表中的数据进行分析并给出合理解释。

图 3 风力等级走势图

19