ZZ052-大数据应用与服务赛项试题 08

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：Hadoop 完全分布式安装配置

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）配置 ssh 免密，实现 master—>master、master

—>slave1、master—>slave2 的免密登录，实现免密后，在 master 节点执行“ssh slave1”命令。提交命令和结果截图;

（2）配置 ssh 免密，实现 slave2—>master、slave2—>slave1、slave2—>slave2 的免密登录，实现免密后，在 slave2 节点执行“ssh master”命令，提交命令和结果截图;

（3）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，提交完整命令截图;

（4）在 master 节点修改/root/.bash_profile 文件，设置 Hadoop 环境变量，提交环境变量配置内容截图;

（5）在 master 节点上面修改 Hadoop 的配置文件 core-site.xml，需要在该文件中指定 HDFS 中 NameNode 的地址：主机为 master，端口 9000。提交修改的内容截图;

（6）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为 3，提交修改的内容截图;

（7）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，提交修改的内容截图;

（8）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，提交命令和结果截图;

(9)在 master 节点上面初始化 Hadoop 环境 namenode，

3

提交初始化命令及初始化结果截图;

（10）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，提交查看结果截图。

将上述任务的命令和结果截图复制粘贴至客户端桌面【M1-T1-SUBT1-提交结果1.docx】中对应的任务序号下。

2.子任务二：Zookeeper 集群安装配置

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-zookeeper-3.5.7-bin.tar.gz 包解压到/opt 路径下，提交完整命令截图;

（2）把解压后的 apache-zookeeper-3.5.7-bin 文件夹更名为 zookeeper-3.5.7，提交完整命令及结果截图;

（3）在 master 节点修改/root/.bash_profile 文件，设置 Zookeeper 环境变量，提交环境变量配置内容截图;

（4）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，提交命令和结果截图;

（5）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，提交命令和结果截图;

4

（6）在 master 节点、slave1 节点、slave2 节点分别启动 zookeeper，提交命令和结果截图;

（7）在 master 节点、slave1 节点、slave2 节点分别查看 zookeeper 的状态，提交命令和结果截图。

将上述任务的命令和结果截图复制粘贴至客户端桌面【M1-T1-SUBT2-提交结果1.docx】中对应的任务序号下。

（二）任务二: 数据库配置维护

1.子任务一: 创建数据库及相关数据表

在 MySQL 数据库中创建“test”数据库，并在“test”数据库中分别创建“stu”、“course”及“score”共 3 个数据表。各个数据表的表字段格式如下:

表 1“stu”的表字段结构

字段	类型	备注
学号	varchar	主键
姓名	varchar	
性别	varchar	
专业	varchar	
班级	varchar	
学院	varchar	

表 2“course”的表字段结构

字段	类型	备注
课程号	varchar	主键
课程名称	varchar	
开设学院	varchar	
学分	int	

表 3“score”的表字段结构

字段	类型	备注
学号	varchar	联合主键
课程号	varchar	
成绩	double	

将创建“test”数据库、“stu”、“course”及“score”的建表结果图分别截图复制粘贴至客户端桌面【M1-T2-SUBT1-提交结果 1.docx】中对应的任务序号下。

2.子任务二：添加数据记录

分别为“stu”、“course”及“score”数据表添加数据记录。各个数据表需要添加的数据记录如下:

表 4“stu”数据表的数据记录

学号	姓名	性别	专业	班级	学院
2023010148	蔡俊豪	男	机械设计	23 机械设计 1 班	智能制造学院
2023010150	何铭业	男	机械设计	23 机械设计 1 班	智能制造学院
2020010132	蔡小怡	女	计算机	20 计算机 1 班	电子信息学院
2020010128	浪佳怡	女	计算机	20 计算机 1 班	电子信息学院
2022010233	胡泽键	男	计算机	22 计算机 2 班	电子信息学院
2022010308	方凯娜	女	计算机	22 计算机 3 班	电子信息学院
2021030140	蔡思欣	女	财务管理	21 财务管理 1 班	经济管理学院
2021030206	方贝乐	女	电子商务	21 电子商务 3 班	经济管理学院
2022030309	冯富祥	男	商务管理	22 商务管理 1 班	经济管理学院
2022040146	陈东杰	男	汽车制造	22 汽车制造 2 班	汽车工程学院
2021040232	陈虹光	男	新能源	21 新能源 1 班	汽车工程学院
2022020318	卓楚莹	女	计算机	22 计算机 3 班	电子信息学院
2022050101	陈琳	女	酒店管理	22 酒店管理 1 班	文化旅游学院

表 5“course”数据表的数据记录

课程号	课程名称	开设学院	学分
DZXX01	C 语言程序设计	电子信息学院	3
JJGL01	会计大数据分析	经济管理学院	3
ZNZZ01	自动控制应用	智能制造学院	3
DZXX02	人工智能概论	电子信息学院	2
QCGC01	汽车质量检验技术	汽车工程学院	3
QCGC02	汽车钣金	汽车工程学院	2
WHLY01	旅游大数据分析	文化旅游学院	3
JJGL02	市场营销实践	经济管理学院	2

表 6“score”数据表的数据记录

学号	课程号	成绩
2023010148	DZXX01	55
2023010148	JJGL01	99
2023010148	ZNZZ01	55
2023010148	DZXX02	64
2023010148	QCGC01	68
2023010150	QCGC02	90

2023010150	WHLY01	81
2023010150	JJGL02	85
2023010150	DZXX01	67
2023010150	JJGL01	98
2020010132	ZNZZ01	98
2020010132	DZXX02	98
2020010132	QCGC01	64
2020010128	QCGC02	99
2020010128	WHLY01	60
2020010128	JJGL02	87
2020010128	DZXX01	82
2022010233	JJGL01	91
2022010233	ZNZZ01	67
2022010233	DZXX02	66
2022010308	QCGC01	65
2022010308	QCGC02	75
2022010308	WHLY01	65
2022010308	JJGL02	58
2021030140	DZXX01	57
2021030140	JJGL01	50
2021030140	ZNZZ01	89
2021030206	DZXX02	61
2021030206	QCGC01	59
2021030206	QCGC02	69
2022030309	WHLY01	99
2022030309	JJGL02	59
2022040146	DZXX01	80
2022040146	JJGL01	81
2022040146	ZNZZ01	69
2022040146	DZXX02	92
2021040232	QCGC01	99
2021040232	QCGC02	68
2021040232	WHLY01	78
2022020318	JJGL02	93
2022020318	DZXX01	96
2022020318	JJGL01	98
2022050101	ZNZZ01	79
2022050101	DZXX02	55
2022050101	QCGC01	58

将“stu”、“course”及“score”的数据添加结果图分别截图复制粘贴至客户端桌面【M1-T2-SUBT2-提交结果

7

1.docx】中对应的任务序号下。

3.子任务三：数据表查询

（1）将班级名称为“20 计算机 1 班 ”的所有学生其“学号”、“姓名”、“班级”、修过的 “课程名称” 及对应的 “成绩” 显示出来;

（2）将修了课程 “学分” 等于 “2” 的所有学生其 “学号”、“姓名”、“课程名称”、“学分” 及对应的 “成绩” 显示出来;

（3）将课程 “成绩” 在 “75” 至 “80” 之间的学生其 “学号”、“姓名”、“班级”、“课程名称” 及对应的 “成绩” 显示出来。

将上述 SQL 查询语句及查询结果图分别截图复制粘贴至客户端桌面【M1-T2-SUBT3-提交结果 1.docx】中对应的任务序号下。

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

打开 ZZ052-8-M2-T1-SUBT1 文件夹，文件夹中包含 data.csv 文件。data.csv 文件是一份电商用户的数据，包括：用户 id、是否新用户、用户年龄、用户性别、用户所在市场级别、用户设备、操作系统、来源、浏览页面总数、浏览过主页的用户、浏览过列表页的用户、浏览过产品详情页的用户、浏览过支付结算页的用户、浏览过确认支付完成页

的用户。使用 pandas 读取 data.csv 并将读取的结果打印在 IDE 终端上，读取代码的截图复制粘贴至客户端桌面【M2-T1-SUBT1-提交结果 1.docx】中对应的任务序号下。

2.子任务二：数据处理

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。

打开 ZZ052-8-M2-T1-SUBT1 文件夹，文件夹中包含 data.csv 文件。你的小组需要通过编写代码或脚本完成对相关数据文件中数据的清洗和整理。请分析相关数据集，根据题目规定要求实现数据处理，具体要求如下:

（1）NAN值代表用户未浏览该页面，查看数据，将NAN替换为0，然后存入data_c1.csv中;

（2）利用正则表达式将page页面文字信息转化为数字1，然后存入data_c2.csv中;

（3）异常值处理：将年龄（age）数字大于等于100的异常数据删除，然后存入data_c3.csv中;

将上述（1）-（3）任务的代码截图复制粘贴至客户端

9

桌面【M2-T1-SUBT2-提交结果 1.docx】中对应的任务序号下。

（4）缺失值处理

①查看缺失值个数;

②处理“source”列缺失值: 当用户为新用户，source缺失值填充为direct，老客户填充seo;

③处理“device”列缺失值: 操作系统为mac，window，linux的设备空值填充为desktop，将操作系统为ioS，android的设备空值填充为mobile，如果为other和NAN值则填充为众数;

④处理“operative_system”缺省值: 当用户设备时mobile时，操作系统填充为iOS，当用户设备为desktop，操作系统填充为windows。

⑤处理sex缺失值: 如果是linux系统，填充为Male，其他均填充为Female。

所有缺失值处理完后，存入data_c4.csv中。

将上述①-⑤任务的代码截图复制粘贴至客户端桌面【M2-T1-SUBT2-提交结果 2.docx】中对应的任务序号下。

（二）任务二: 数据标注

对 data_c4.csv 数据进行标注，判断客服是否下单，具体的标注规则如下:

（1）如果“confirmation_page”列数据为 1，则数据标注为 ‘yes’;

（2）如果“confirmation_page”列数据为 0，则数据

标注为 ‘no’;

标注好的数据存储为列‘subscribe’并和 data_c4.csv 数据合并存入 result.csv。

将代码截图复制粘贴至客户端桌面【M2-T2-SUBT1-提交结果 1.docx】中对应的任务序号下。

（三）任务三: 数据清洗

1.子任务一: 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 中的/root/目录下新建一个文件夹:result，将创建文件夹命令与结果截图粘贴至客户端桌面【M2-T3-SUBT1-提交结果 1.docx】中对应的任务序号下;

（2）使用 HDFS 命令，将 Master 下: /root 目录下新建的文件夹: result 上传到 HDFS 指定目录下: /根目录下; 并且使用 HDFS 命令查看目录; 将 HDFS 上传，查看命令截图粘贴至客户端桌面【M2-T3-SUBT1-提交结果 2.docx】中对应的任务序号下;

（3）使用 HDFS 命令，将 HDFS 目录下的/result 文件夹下载到 Master 指定目录下:根目录下; 将下载文件夹命令与结果截图粘贴至客户端桌面【M2-T3-SUBT1-提交结果 3.docx】中对应的任务序号下。

2.子任务二: 处理异常数据

打开 ZZ052-8-M2-T3-SUBT2 文件夹，文件夹中包含

11

user_info.csv 文件。user_info.csv 文件存储了电商互联网平台上收集的用户数据，数据中有以下内容:

id: 主键非空，bigint 类型，长度为 20

login_name: 用户名，varchar 类型，长度 200

nick_name: 用户昵称，varchar 类型，长度 200

passwd: 密码，varchar 类型，长度 200

name: 姓名，varchar 类型，长度 200

phone_num:手机号，varchar 类型，长度 200

email:邮箱，varchar 类型，长度 200

head_img:头像，varchar 类型，长度 200

user_level:用户级别，varchar 类型，长度 200

birthday:用户生日，date 类型，长度 0，格式为 YYYY-MM-DD

gender:性别，varchar 类型，长度 1

create_time:创建时间，datetime 类型，格式为 yyyy-MM-dd HH:mm:ss

operate_time:修改时间，datetime 类型，格式为 yyyy-MM-dd HH:mm:ss

编写 MapReduce 程序，实现以下功能：将 user_info.csv 数据的分隔符“,”转换为“|”，输出文件到 HDFS，然后在控制台按顺序打印输出前 10 条数据，将结果截图粘贴至客户端桌面【M2-T3-SUBT2-提交结果 1.docx】中对应的任务序号下。

12

3.子任务三：数据统计

打开 ZZ052-8-M2-T3-SUBT3 文件夹，文件夹中包含 sku_info.csv 文件。sku_info.csv 文件存储了电商互联网平台上收集的商品数据，数据中有以下内容:

id: 主键非空，bigint 类型，长度为 20

spu_id: spuId，varchar 类型，长度 20

price: 价格，decimal 类型，长度 10

sku_name: 商品名称，varchar 类型，长度 200

sku_desc: 商品描述，varchar 类型，长度 2000

weight:重量，decimal 类型，长度 10

tm_id:品牌，bigint 类型，长度 20

category3_id:三级分类，bigint 类型，长度 20

sku_default_img:默认显示图片，varchar 类型，长度 200

create_time:创建时间，datetime 类型，长度 0，格式为 yyyy-MM-dd HH:mm:ss

编写 MapReduce 程序，实现以下功能：三级分类 category3_id 范围为[1,10],1 表示最低级别，10 表示最高级别。本任务遍历 sku_info.csv 中数据，统计字段“三级分类”级别为“10”最高级别的商品数量，将结果截图粘贴至客户端桌面【M2-T3-SUBT3-提交结果 1.docx】中对应的任务序号下。

四、模块三：业务分析与可视化

13

（一）任务一：数据可视化

1.子任务一：数据分析

打开 ZZ052-8-M3-T1-SUBT1 文件夹，文件夹中包含了 TRA.xlsx 文件。对 TRA.xlsx 文件中的数据，使用电子表格软件进行查询统计并使用图表进行展示。

（1）如参考截图所示，根据 TRA.xlsx 文件的数据，使用电子表格软件统计景区评论人数，并以柱状图展示：

①统计每个“景点”的“评论人数”；

②对“景点”的“评论人数”进行降序排列；

③取“景点”的“评论人数”前十使用柱状图进行显示。

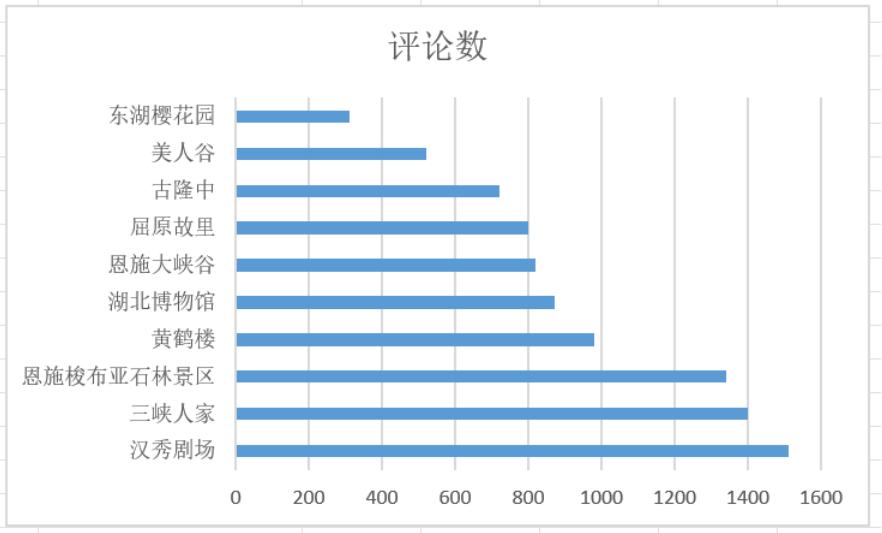

图 1 “评论人数”的柱状参考示意图

（2）如参考截图所示，根据 TRA.xlsx 文件的数据，使用电子表格软件统计各地出游人数，并以饼图展示：

①统计每个景点“评论的发布地”的“评论数量”；

②根据每种“评论的发布地”的“评论数量”计算各地区出游人数的百分比；

14

③使用圆环图展示各地区出游人数的百分比。

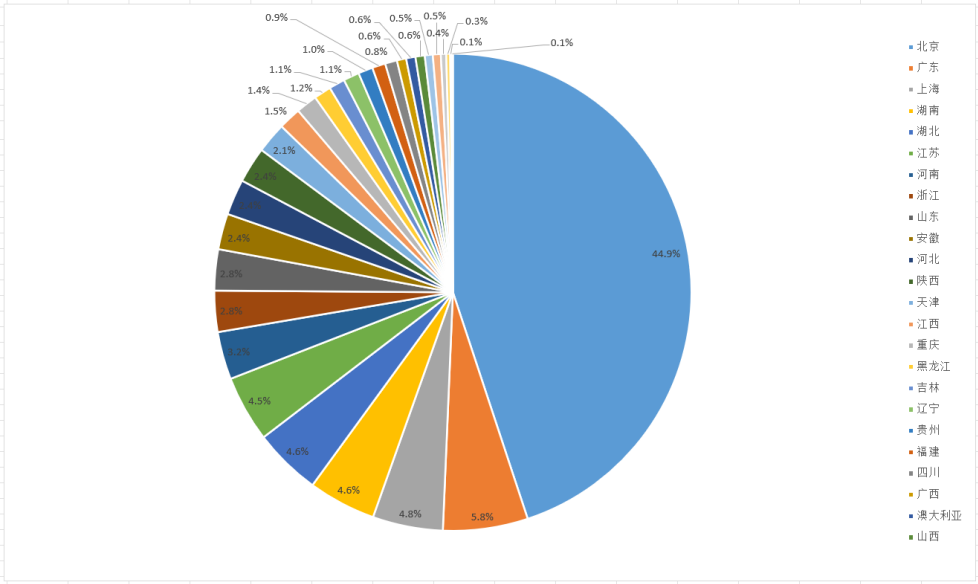

图 2 各地区出游人数的百分比参考示意图

将上述 2 个展示图分别截图复制粘贴至客户端桌面【M3-T1-SUBT1-提交结果 1.docx】中对应的任务序号下。

2.子任务二：数据可视化

打开 ZZ052-8-M3-T1-SUBT2 文件夹，文件夹中包含 traveSite 项目目录。打开 traveSite 项目，编写补充代码，实现 Web 网页形式对相关数据进行可视化展示:

（1）如参考截图所示,根据 traveSite/js/data.js 文件中 gradeData 对象中的数据，补充完整 traveSite/js/index.js 文件中 getGradeData()函数的代码，实现“好评度分布”环形图显示:

①获取“好评度”区域 dom，初始化 echarts 实例;

②编写补充 series 对象,设置图表的标题和图表类型，设置饼图半径为['20%', '55%']，设置饼图样式：饼图份例

15

圆角度数为“4”将 gradeData 对象中的数据设置为饼图显示数据;

③使用指定的配置项和数据显示图表;

④运行网页，对“好评度分布”效果图进行截图。

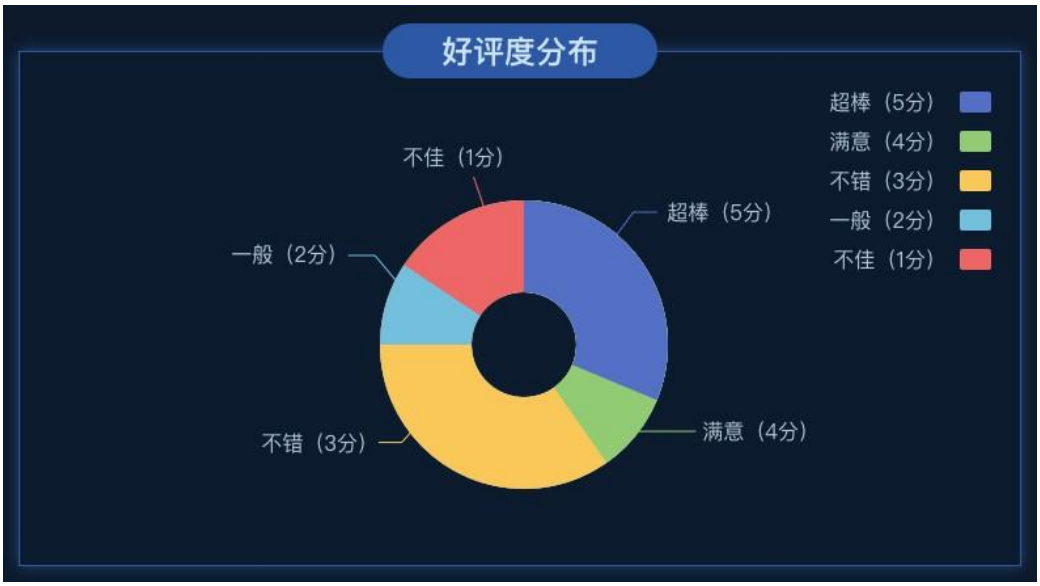

图 3 “好评度分布”的饼状参考示意图

（2）如参考截图所示,根据 traveSite/js/data.js 文件中 scenic_spo 对象中的数据，补充完整 traveSite/js/index.js 文件中 getHotScenery()函数的代码，实现“热门景点”柱状图的显示:

①编写补充 yAxis 对象，设置坐标轴类型为“类目轴”，坐标文字显示为白色，坐标数据显示为景点;

②编写补充 series 对象,设置图表的标题和图表类型，显示柱状图文字（位置右方）并将 scenic_spo 作为图表显示数据;

③运行网页，对“热门景区”效果图进行截图。

16

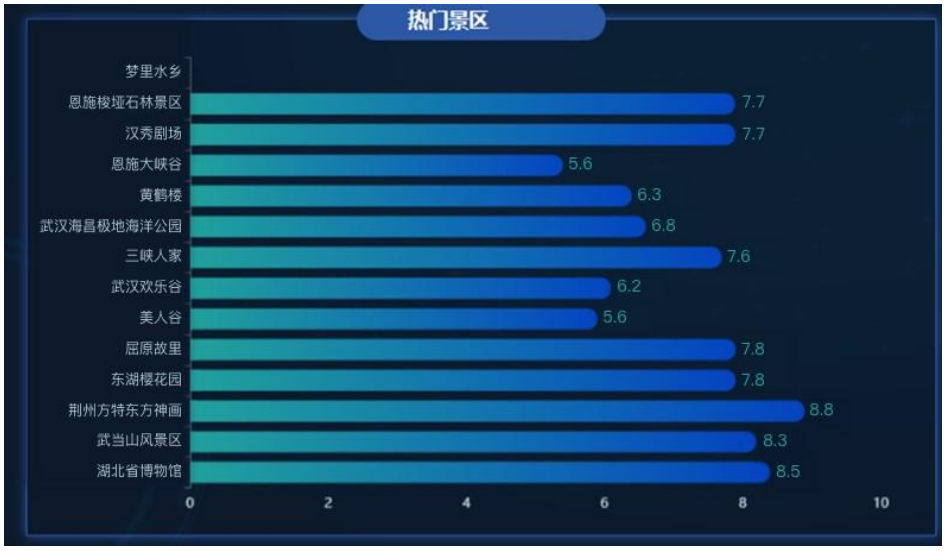

图 4 “热门景区”的柱状参考示意图

将该 2 个可视化图表分别截图复制粘贴至客户端桌面【M3-T1-SUBT2-提交结果 1.docx】中对应的任务序号下。

（二）任务二：业务分析

1.子任务一：业务分析

根据上述生成的 result.csv 数据，分析不同操作系统的购买产品的数量，并将分析结果用 python 绘制柱状图。将图表截图复制粘贴至客户端桌面【M3-T2-SUBT1-提交结果 1.docx】中对应的任务序号下，最后在其下方编写发展趋势分析。

2.子任务二：报表分析

转换率可以看到用户从进入平台到最终完成购买整个流程的各个环节当中，哪个环节流失率高，需要定位到这个环节，观察是不是这个页面或者功能让用户产生了不好的体验，从而导致用户大量流失。根据上述生成的 result.csv 数据，从设备维度进行转换率分析，并将分析结果通过 pyt

17

hon 生成报表信息方便商城在后续服务中进行优化，及时准确的把握市场行情，具体要求如下：画出基于设备维度的转换率条形图。将图表截一张图复制粘贴至客户端桌面【M3-T2-SUBT2-提交结果 1.docx】中对应的任务序号下。

18