2024

金砖国家职业技能大赛（金砖国家未来技能和技术挑战赛）

数据分析与可视化

BRICS-FS-36

样题（省级/区域选拔赛）

2024 年 02 月

目录

1 参赛形式

本次赛项为个人赛。

2 竞赛内容

本次竞赛由 3 个模块组成，选手需要按顺序完成所有竞赛内容。竞赛时会向参赛选手提供统一的赛题文件、竞赛设备、设备基础操作说明文件，以及为保障每个任务模块的独立性与公平性所需数据源或其他技术基础条件。

竞赛内容包含基于数据分析与可视化的以下任务模块：

模块 A 数据获取与处理

模块 B 数据分析与运营

模块 C 数据展示与分享

如果参赛选手不遵守职业健康安全环境要求，或使自己和其他选手面临危险，他们可能会被取消比赛资格。

参赛者完成竞赛后，由裁判组对选手提交结果进行评分。

3 项目模块和时间要求

3.1 项目模块和时间要求

数据分析与可视化赛项共 3 个模块，要求参赛者总共用时 360min。具体项目模块名称和时间要求参照表格 1 项目模块和时间要求清单。

表格 1 项目模块和时间要求清单

序号	模块名称	竞赛内容完成时间
1	模块 A：数据获取与处理	120min
2	模块 B：数据分析与运营	120min
3	模块 C：数据展示与分享	120min

3.2 任务内容

模块 A 数据获取与处理（120min）

模块描述：

随着科学技术的不断进化和更新，人们可以更容易地从各种渠道获得所需的信息。当越来越多的信息充斥着人们的生活时，大数据应运而生。在大数据的背景下，医疗数据的数量和复杂性也不断增加。近年来，随着电子医疗记录的普及和信息技术的进步，医疗数据分析变得越来越重要。医疗数据分析在医学研究、临床决策支持、疾病预防和公共卫生方面有着广泛的应用。例如，医疗数据分析可以用于研究疾病的发病机制、探索新的治疗方法、评估医疗保健政策和规划公共卫生战略。同时，医疗数据分析也可以用于提高医疗保健的效率和质量，例如优化医疗资源的分配、减少医疗错误和提高患者满意度。

大数据对医疗保健具有重要意义。医疗卫生机构使用大数据可以有效帮助医生做出更准确的临床诊断；更准确地预测治疗方案的成本和疗效；整合患者遗传信息以进行个性化治疗；分析人口健康数据以预测疾病暴发等。

任务一：数据处理

1、在`医疗数据`工作表中操作，删除数据重复的行，去重时保留首条数据，并保存操作完成的数据。

2、统计`医疗数据`工作表中`A1:N16551`的缺失值的数量,保存答案到`results`表的`answer`列对应位置。

3、在`医疗数据`工作表中操作，按照下面要求处理缺失值：

- 将年龄字段或性别字段存在缺失值的行删除
- 人员区划字段缺失值用就诊区划里的值填充

4、在`医疗数据`工作表中操作，`diagnosis_name`字段中的数据存在一些特殊字符和附加说明，这些数据的处理标准在`diagnosis_process`工作表中，请按照`diagnosis_process`工作表处理`diagnosis_name`字段中的数据，并保存操作完成的数据。

5、在`医疗数据`工作表中操作，将`settlement_date`字段的时间戳转化成时间格式（例如：20180101），并保存操作完成的数据。

6、在`医疗数据`工作表中操作，将`age`字段的数据按照以下标准划分等级，形成新的`age_level`字段，并保存操作完成的数据。

婴幼儿：	$0 < age \leqslant 6$
少 儿：	$6 < age \leqslant 12$
青少年：	$12 < age \leqslant 17$
青 年：	$17 < age \leqslant 45$
中 年：	$45 < age \leqslant 69$

老 年:　　69 < age < 100

任务二：数据分析

1、统计`医疗数据`工作表中的`age_level`字段生病人数最多的年龄组，保存答案到`results`表的`answer`列对应位置。

2、在`医疗数据`工作表中，总费用最高的医疗类型？保存答案到`results`表的`answer`列对应位置。

3、在`医疗数据`工作表中，根据入院日期分析各年度参保人跨区划流动就医的比例（小数形式，保留两位小数），保存答案到`2.3`表的对应的位置。

4、在`医疗数据`工作表中，离休人员占总人数的比例（百分比形式，保留两位小数），保存答案到`results`表的`answer`列对应位置。

5、在`医疗数据`工作表中，就诊次数最多的人员 ID 是？就诊次数是多少，保存人员 ID 和就诊次数到`results`表的`answer`列对应位置。

6、在`医疗数据`工作表中，根据入院日期和出院时间计算住院时长，求平均住院时长最长的疾病。保存答案到`results`表的`answer`列对应位置。

7、在`医疗数据`工作表中，统计平均报销率超过 80%的疾病数量。保存答案到`results`表的`answer`列对应位置。

平均报销率=平均统筹金额/平均总费用

任务三：数据可视化

1、根据`医疗数据`工作表中数据，分析不同年龄组就诊次数分布情况。

- 以圆环图的形式呈现
- 显示数据标签并保留两位小数。
- 图表保存到`3.1`工作表中

2、根据`医疗数据`工作表中数据，分析不同人员类型总消费和总统筹金额

- 以柱状图+折线图的形式呈现
- 总费用为柱状图，总统筹金额为折线图
- 按照总费用降序排列
- 纵坐标轴的最大值为 21000000，最小值为 0
- 图表保存到`3.2`工作表中

3、根据`医疗数据`工作表中数据，分析不同肿瘤疾病的平均个人花费费用

个人花费费用=总费用统筹金额

- 以气泡图的形式呈现。
- 气泡大小用平均个人花费费用表示。
- 只显示平均个人花费费用排名前五的疾病的标签且居中显示，且气泡颜色依次设置为 FFCCCC，99CCCC，CCCC99，CC99CC，CC99FF。
- 图表保存到`3.3`工作表中

模块 B 数据分析与运营（120min）

模块描述：

教育在大数据技术与理念的冲击下正在发生一场“静悄悄的革命”，教学范式的转型成为这场革命的先导和核心。随着大数据时代的到来，教学范式也步入了 3.0 时代。校园数据分析是指利用各种技术和工具，对学校内部的各种数据进行深入的挖掘和分析，从而为学校决策和管理提供科学依据的过程。在当今信息时代，学校管理面临着各种挑战和机遇，例如学生招生、教学质量评估、师生行为监管、财务管理等方面，都需要依靠数据来支撑决策。积极探索教育大数据驱动教学范式，让真实的教学数据赋予教师“显微镜”式的观察能力，以及“望远镜”式的预测能力。让教学实现科学化、智能化、精准化与个性化。

任务一、数据处理

1.1：读取数据，读取`Consumption.csv`、`attendance.csv`、`teacher.csv`、`mark.csv`四个表的数据，分别保存到变量`data_Consumption`、`data_attendance`、`data_teacher`、`data_mark`。并运行给出的答案保存代码保存答案。

1.2 分析变量`data_Consumption`，将`DealTime`列处理为日期格式，将处理的结果更新到变量`data_Consumption`，并运行给出的答案代码保存答案。

1.3 分析变量`data_Consumption`，将`MonDeal`转为正数，将处理的结果更新到变量`data_Consumption`，并运行给出的答案代码保存答案。

1.4 分析变量`data_attendance`，处理 2014 年 2 月份的数据，将`qj_term`更新为'201320142'，将处理后的结果更新到变量`data_attendance`，将并运行给出的答案代码保存答案。

1.5 分析变量`data_attendance`中的`DataDateTime`、`qj_term`两列数据，找出`DataDateTime`与`qj_term`之间的关系，将缺少的学期信息进行填充，返回

考勤最多的学期以及所对应的考勤次数，结果保存到`task1_5`中，并运行给出的答案代码保存答案。

任务二、校园教师信息分析

2.1 分析变量`data_teacher`中的`sub_Name`、`bas_id`两列数据，统计各个学科的老师数量，返回拥有老师最多的学科名，结果保存到变量`task2_1`中，并运行给出的答案保存代码保存答案。

2.2 分析变量`data_teacher`中的`term`、`gra_Name`、`bas_id`3 列数据，观察"201420151"学期老师的教学情况，统计有多少老师涉及多个年级的课程，结果保存到变量`task2_2`中，并运行给出的答案保存代码保存答案。

2.3 分析变量 `data_teacher` 中的 `term`、`cla_Name`、`bas_id`3 列数据，观察"201420151、201420152"两学 期老师的教学情况，统计每位老师授课班级数量，返回上课班级最多的`bas_id`，结果保存到变量`task2_3`中，并运 行给出的答案保存代码保存答案。

2.4 分析变量`data_teacher`中的`term`、`cla_Name`、`gra_Name`3 列数据，统计每学期各年级班级的数量，返回班级最多的学期和年级，结果保存到变量`task2_4`，并运行给出的答案保存代码保存答案。

2.5 分析变量`data_teacher`,统计'201420151'学期,高三物理学科和地理学科各开设了多少个班，结果保存到变量`task2_5`，并运行给出的答案保存代码保存答案。

- 保存格式(开设地理课的班级数量,开设物理课的班级数量)

任务三、校园考勤统计

3.1 分析变量`data_attendance`，统计`controler_name`包含`校服`的考勤有多少条，返回迟到早退最多的学期，结果保存到变量`task3_1`，并运行给出的答案保存代码保存答案。

3.2 分析变量`data_attendance`，统计高一年级迟到早退次数最多的`bf_studentID`，结果保存到变量`task3_2`，并运行给出的答案保存代码保存答案。

- 迟到：`controler_name`包含迟到
- 早退：`controler_name`包含早退

3.3 分析变量`data_attendance`,计算考勤次数大于 200 次的`bf_classid`中，考勤状态为迟到或早退的次数占该班级考勤总次数的比例，返回最大的比例，结果保存到变量`task3_3`，并运行给出的答案保存代码保存答案。

3.4 分析变量`data_attendance`，以小时统计迟到早退的次数，返回迟到早退的高峰期是在一天的哪个时间（24 小时制），结果保存到变量`task3_4`，并运行给出的答案保存代码保存答案。

3.5 分析变量`data_attendance`，统计有多少个学生出现过同一天内迟到且早退情况，结果保存到变量`task3_5`，并运行给出的答案保存代码保存答案。

任务四、学生成绩分析

4.1 分析变量`data_mark`，计算`exam_number`为 289 考试中，高二各班级的化学学科的平均分，返回平均分最高的`cla_id`，结果保存到变量`task4_1`，运行结果保存代码保存答案。

4.2 分析变量`data_mark`，将`exam_number`为 289 考试中化学学科的成绩，按以下要求对成绩进行划分，统计各成绩段所对应的人数，返回人数分布最多的成绩段，结果保存到变量`task4_2`，运行结果保存代码保存答案。

- 分数=3:免考
- 2<=分数<60:不及格
- 60<=分数<70:及格
- 70<=分数<80:中等
- 80<=分数<90:良
- 90<=分数<=100:优

4.3 分析变量`data_mark`,以`exam_number`为 289 考试中化学学科的成绩作为一个组别，计算这组数据的 Z_score，求 Z_score 的最大值是多少？结果保存到变量`task4_3`，运行结果保存代码保存答案。

- Z_score，是一种具有相等单位的量数。它是将原始分数与团体的平均数之差除以标准差所得的商数,是以标准差为单位度量原始分数离开其平均数的分数之上多少个标准差，或是在平均数之下多少个标准差。

Z_score = (分数 - 平均分数) / 标准差结果保留两位小数

4.4 分析变量`data_mark`,统计`exam_number`为 284 的考试中各班级作弊和缺考次数，返回作弊加缺考次数最多的`cla_id`，结果保存到变量`task4_4`，运行结果保存代码保存答案。

4.5 分析变量`data_mark`，计算`exam_number`为 289 考试中，高二、高三各班中各科成绩排名第一最多的`cla_id`，结果保存到变量`task4_5`，运行结果保存代码保存答案。

4.6 分析变量`data_mark`，计算考试编号为 289 考试中，各个学生的总分（分值保留为一位小数),返回各年级总分最高的分值,结果保存到变量`task4_6`,并运行结果保存代码保存答案。

- 总分 score 由三部分组成，计算方式如下:
  - score1:语数外的成绩

•score2:物理、地理两科分值最高的一门（分值相同，选择一门的分值即可）

• score3:生物、政治、化学、历史四门中最高的两门（根据分值相同的多种情况，选择其中的两门或者一门）

• score = score1+score2+score3

模块 C 数据展示与分享（120min）

模块描述：

随着企业的发展和市场的竞争，公司销售数据分析变得越来越重要。销售数据分析是指对企业销售业绩数据进行收集、整理、分析和利用，以了解公司销售业绩的情况，发现潜在的市场机会，优化销售策略和营销方案，以提高企业的销售收入和利润。

在现代企业中，销售数据分析已经成为了一个不可或缺的工具。通过数据分析，企业可以更好地理解自己的市场、客户和竞争对手，从而做出更明智的业务决策。这些决策可以包括如何定价、如何推广产品、如何改进客户服务等等。RFM 模型是衡量客户价值和客户创造利益能力的重要工具和手段。在众多的客户关系管理分析模式中，RFM 模型是被广泛提到的。该模型通过客户近期购买行为、购买的总体频率以及花了多少钱，这 3 项指标来观察该客户的价值状况。

任务一、产品分析

任务 1.1:展示在产品描述中包含尿裤的销售金额情况

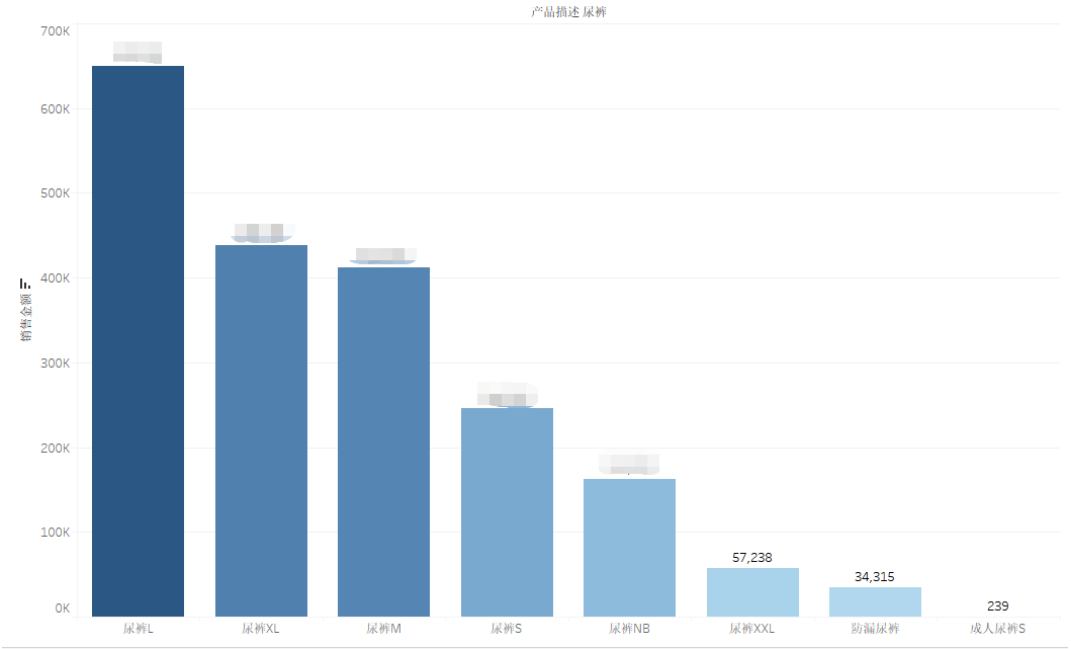
任务书要求：

按照以下要求，在名称为 1.1 的工作表中进行操作，并保存最终结果：

图形名称：柱状图

- 产品描述的内容里面包含“尿裤”的关键词，与示例图的列标签一致
- 字段产品描述为列，产品描述的列名要与示例图完全一致，字段销售金额为行
- 按照销售金额进行从左到右，降序排序
- 显示销售金额的标记标签
- 柱状图设置为蓝色，销售金额的大小代表颜色的深浅，销售金额越大，颜色越深，与示例图一致
- 将视图尺寸设置为‘整个视图‘

参考图形如下：



图表 1

任务 1.2:展示子类别的销售金额

任务书要求：

按照以下要求，在名称为 1.2 的工作表中进行操作，并保存最终结果：

图形名称：词云图

- 显示每个子类别的标签，销售金额控制标签的大小
- 销售金额的颜色设置为 tableau classic 20，与示例图一致
- 按照销售金额对子类别进行升序排序
- 将视图尺寸设置为‘整个视图’

参考图形如下：

图表 2

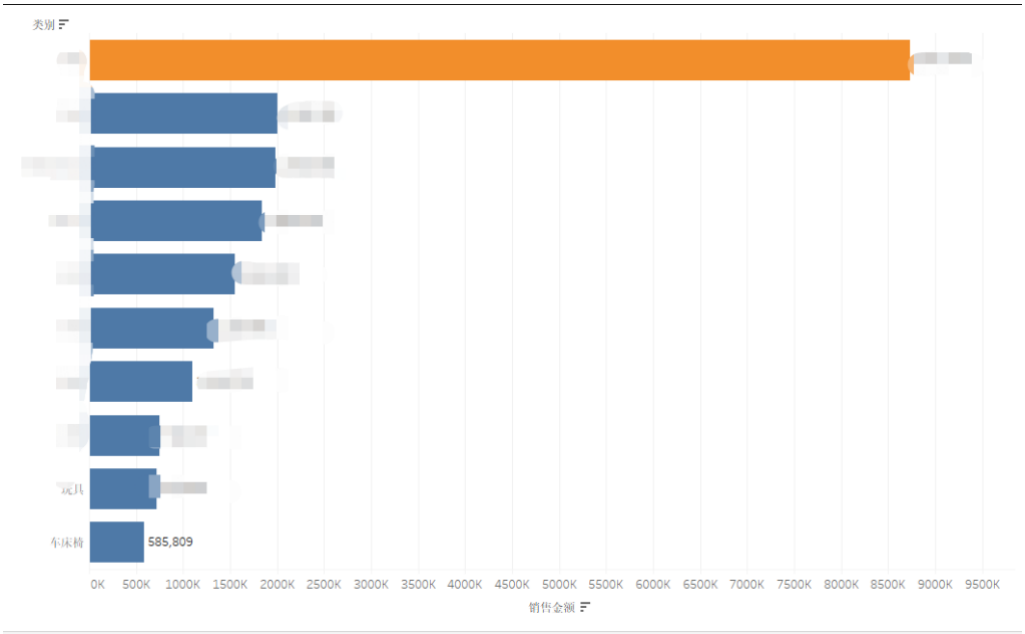
任务 1.3:展示销售金额排行前 10 名的类别情况

按照以下要求，在名称为 1.3 的工作表中进行操作，并保存最终结果：

图形名称：水平条

- 字段销售金额在列，字段类别在行
- 只保留销售金额总和排行前 10 名的类别，销售总额排行第一的类别颜色设置为橙色其余九个类别设置为蓝色，与示例图一致
- 类别按照销售金额的总和进行从上到下，降序排序
- 显示销售金额的标记标签，与示例图一致
- 将视图尺寸设置为‘整个视图’

参考图形如下：



图表 3

任务 1.4：展示食品大类和服装大类的百分比情况

任务书要求：

按照以下要求，在名称为 1.4 的工作表中进行操作，并保存最终结果：

图形名称：圆环图

•在类别中，服装大类设置为包含服孕、童附、童配、童袜、童鞋、童装、婴装；食品大类包含奶粉和食品

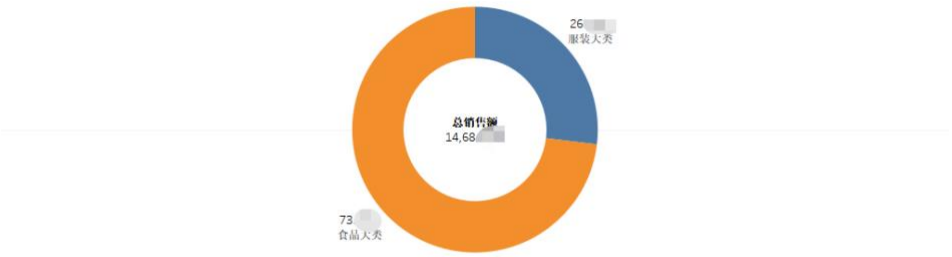
•服装大类和食品大类的角度代表不同销售金额大小

•显示服装、食品大类和销售金额百分比的标记标签，保留两位小数，与示例图一致

•设置圆环内部标签设置为“总销售额”，展示服装、食品大类总销售金额的标记标签，与示例图一致

•服装大类和食品大类的颜色设置为蓝色和橙色，与示例图一致

参考图形如下：



图表 4

任务 1.5:展示 2013 年类别分析的仪表盘

任务书要求：

按照以下要求，在名称为 1.5 的仪表板中进行操作，并保存最终结果：

仪表板标题：2013 年类别分析

• 仪表板大小设置为通用桌面，背景图设置正确

• 堆叠图展示 2013 年销售金额排行前 6 名的类别和对应的百分比标记标签，与示例图一致

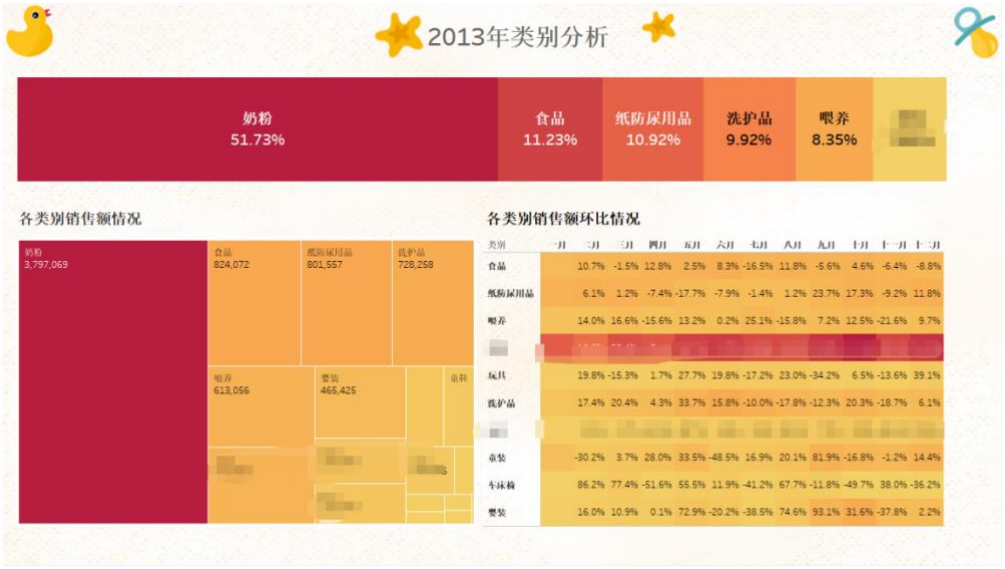
• 树状图展示年份为 2013 年的数据，展示各类别和销售金额的标记标签，按照销售金额的升序排序类别，与示例图一致

• 突出显示表展示年份为 2013 年，展示排行前 10 名的类别和销售金额环比百分比数据，按照 Y 轴的类别标签进行排序，与示例图一致

• 堆叠图、树状图、突出显示表的颜色设置为红色金色，与示例图一致

• 树状图与突出显示表的标题设置为默认字体 tableau book，字号设置为 15 号并加粗

参考图形如下：



图表 5

任务二、订单分析

任务 2.1:展示 2012 年订单日期的销售金额分布情况

任务书要求：

按照以下要求，在名称为 2.1 的工作表中进行操作，并保存最终结果：

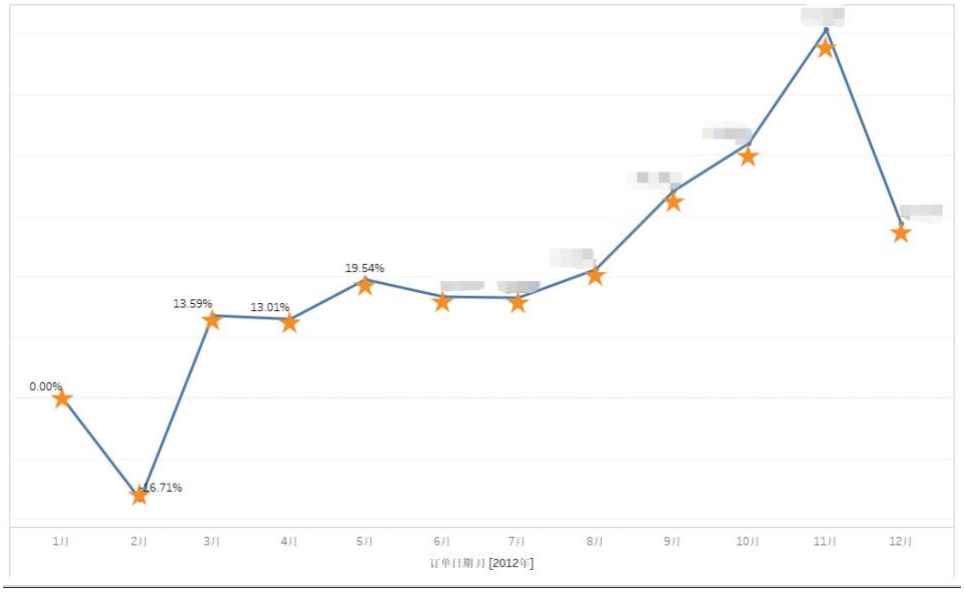
图形名称：双轴图

- 字段订单日期为列，字段销售金额为行
- 只展示总销售金额排行第一的类别
- 展示 2012 年的销售金额数据，以 2012 年 1 月为固定日期，对比 2012 年 2 月至 12 月销售金额，计算定基增长率

定基增加率=（本月销售金额固定月份销售金额）/ 固定月份销售金额

- 显示定基增长率的标记标签，将标记标签设置在折线图的上方，与示例图一致
- 不显示 X 轴两侧的标题，与示例图一致
- 折线图颜色设置为蓝色，五角星颜色设置为橙色
- 将视图尺寸设置为‘整个视图’

参考图形如下：



图表 6

任务 2.2:使用时间序列分析对销售金额进行预测

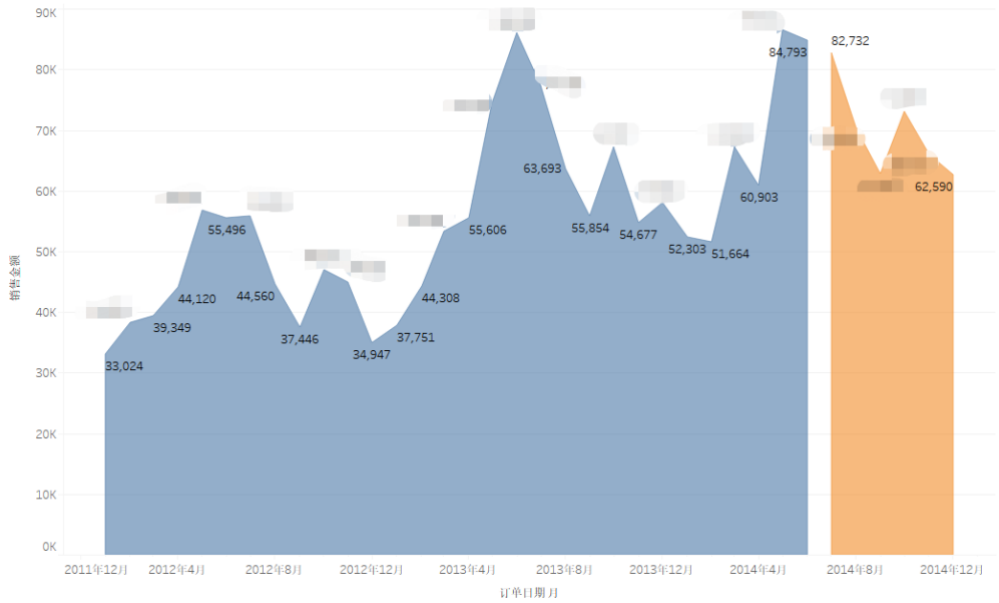
任务书要求：

按照以下要求，在名称为 2.2 的工作表中进行操作，并保存最终结果：

图形名称：预测图

- 字段订单日期为列，字段销售金额为行，类别保留洗护品
- 订单日期只保留 2012 年 1 月 1 日 - 2014 年 7 月 24 日
- 日期格式设置为月数，预测未来 6 个月数据，忽略最后 1 月
- 趋势性和季节性设置为累加方式
- 将实际数据设置为蓝色，预测估计数据设置为橙色，视图尺寸设置为‘整个视图‘
- 显示销售金额的标记标签,与示例图一致

参考图形如下：



图表 7

任务三、综合分析

任务 3.1:展示会员留存情况

任务书要求：

按照以下要求，在名称为 3.1 的工作表中进行操作，并保存最终结果：

图形名称：漏斗图

• 计算“购买点会员生命期（月）”指标，以月为单位计算，例子：购买点会员生命期=会员订单日期（2010 年 3 月 20 日）会员创建日期（2010 年 2 月 25 日）=1

• 字段用户 id 的计数在列，字段购买点会员生命期（月）在行

• 购买点会员生命期（月）按照用户 id 计数进行从上到下降序排序

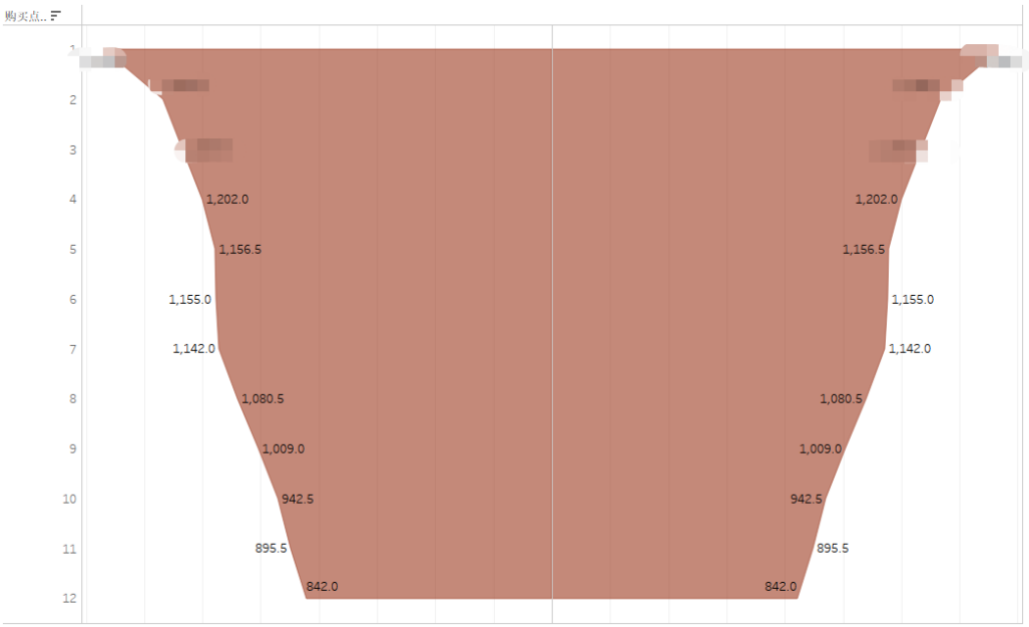
• 购买点会员生命期只保留 1 个月 12 个月数据

• 显示用户 id 计数的标记标签，与示例图一致

• 漏斗图颜色为橙色，与示例图一致

• 将视图尺寸设置为‘’整个视图‘’

参考图形如下：



图表 8

任务 3.2:展示最后购买点生命期的用户数量情况

任务书要求：

按照以下要求，在名称为 3.2 的工作表中进行操作，并保存最终结果：

图形名称：直方图（质量分布图）

• 计算字段“购买点会员生命期（日）”指标，购买点会员生命期=会员订单日期会员创建日期（以日为单位计算）

• 计算“最后购买点生命期”（日）的指标，依据此指标创建数据桶大小为 30 的新字段，新字段名称设置为“最后购买点生命期分布”

• 字段最后购买点生命期分布为列，字段用户 id 为行

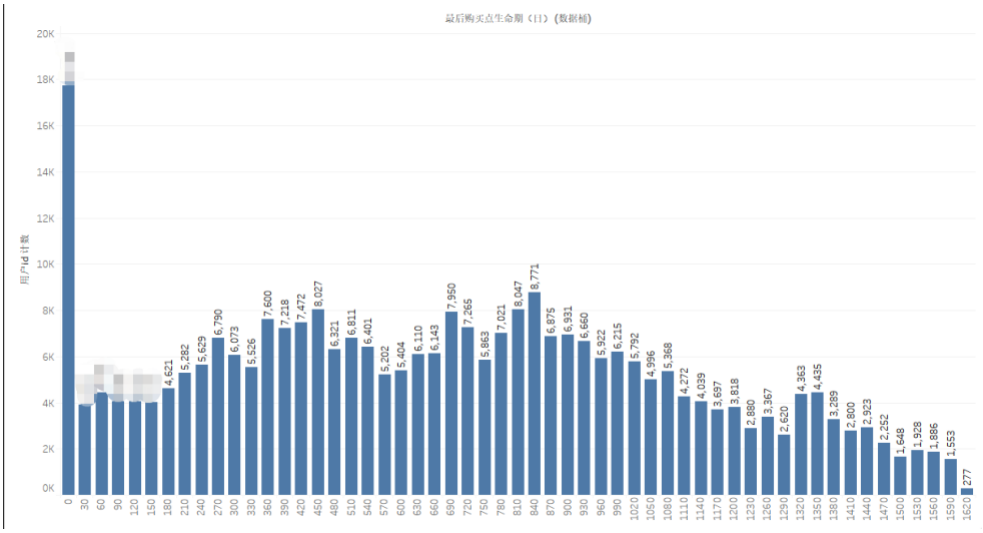
• 订单日期保留 2010 年 7 月 1 日至 2014 年 7 月 24 日的数据

• 显示用户 id 计数的标记标签

• 直方图颜色为蓝色，与示例图一致

• 将视图尺寸设置为‘整个视图’

参考图形如下：



图表 9

任务 3.3:展示用户消费频次及销售金额情况

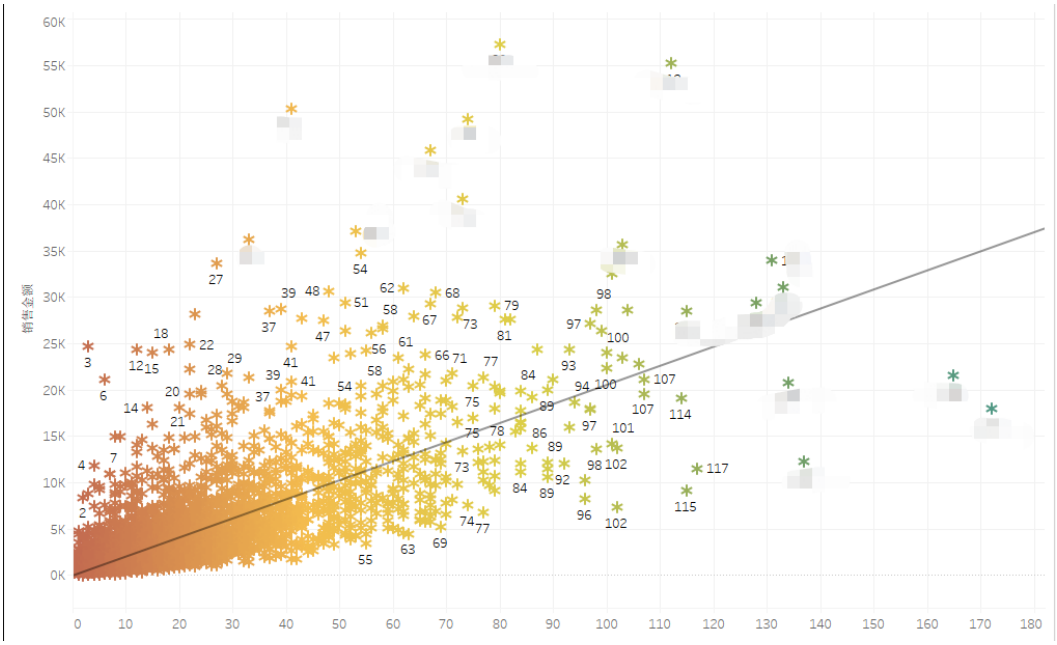
任务书要求：

按照以下要求，在名称为 3.3 的工作表中进行操作，并保存最终结果：

图形名称：散点图

- 计算“每个客户累计下单总数”指标创建新字段，新字段名称设置为每个客户累计下单总数
- 每个客户累计下单总数在列，字段销售金额在行
- 设置线性趋势线，展示客户累计下单总数的销售金额中的趋势，与示例图一致
- 将视图尺寸设置为 ‘’ 整个视图 ‘’
- 散点图颜色为温度发散，设置为倒序
- 设置散点图的形状在为雪花状，与示例图一致
- 显示每个客户累计下单总数的标记标签

参考图形如下：


销售金额
60K
55K
50K
45K
40K
35K
30K
25K
20K
15K
10K
5K
0K
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180

图表 10

4 项目模块评分标准

项目模块评分标准参照表格 2。

表格 2 评分标准

模块	任务	配分
A	数据获取与处理	30
B	数据分析与运营	40
C	数据展示与分享	30
合计		100

注：样题最终解释权归组委会所有