2025

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(国际总决赛)

2025 年 05 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 4 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 数据展示与分享

模块 C 数据开发与应用

模块 D 作品答辩

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 4 个模块，要求参赛者总共用时 370min。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	竞赛时间
1	模块 A：数据获取与处理	120min
2	模块 B：数据展示与分享	120min
3	模块 C：数据开发与应用	120min
4	模块 D：作品答辩	10min

3.2 任务内容

模块 A 数据获取与处理

模块描述：

2008 年，美国的房地产市场经历了一场前所未有的危机，这场危机不仅对美国经济产生了深远的影响，也波及全球经济，成为全球金融危机的重要导火索。危机的爆发源于次级抵押贷款市场的崩溃。自 20 世纪 90 年代末以来，房地产市场经历了一段快速增长的时期，房价持续攀升，住房拥有率达到历史最高水平。然而，随着金融机构放宽贷款标准，大量高风险的次级贷款被发放给信用不佳的借款人。

这些次级贷款大多采用浮动利率，当利率上升时，许多借款人无法按时还款，导致大量房屋被止赎。2007 年，次贷危机开始显现，大量的止赎房屋涌入市场，房价开始迅速下跌。2008 年，危机全面爆发，房价崩溃引发了金融市场的连锁反应，大型金融机构相继倒闭或被收购，全球金融市场陷入动荡。特别是在经历

了 2008 年的危机之后，了解市场如何从危机中恢复，对预防未来类似危机具有重要的参考价值。

任务一：数据处理

任务 1.1：房地产交易数据处理与分析

为了进行更有效的时间序列分析和月度市场趋势分析,有必要对现有的数据进行处理和完善，特别是调整时间数据的格式和细节，以便能够更精确地捕捉市场的变化和趋势。

任务：

1. 格式化日期：调整 Date 列的数据格式为标准形式（yyyy-mm-dd），以统一日期表示方法，便于后续处理和分析。

2. 创建月份字段：在 Month 列中， 基于 Date 列数据，提取出每笔交易发生的月份（阿拉伯数字表示）。

任务 1.2：优化房地产交易数据集以提升数据质量

通过修正已知的数据质量问题,确保数据集在进行进一步的统计分析和模型训练时能够提供可靠的支持。

任务

1. 修正估价异常值：在 Estimated_Value 列中，将所有异常值 0 更改为该列的平均值 448673。

2. 标准化属性类型：在 Property 列中，将所有标记为?的未知房产类型更改为占比最多的房产类型 Single Family。

任务 1.3：填充房地产数据集中的缺失值

为确保数据集在统计分析和模型训练中提供准确和可靠的支持,有必要修正数据质量问题并填充缺失值。

任务：

1. 填充 carpet_area 列的缺失值：在 carpet_area 列中，存在一些缺失值，我们将使用该列的平均值 1111 进行填充。

任务二：数据处理

任务 2.1：计算每平方英尺价格

通过计算每平方英尺的价格，我们可以更准确地了解不同房地产的相对价值，从而进行更深入的市场分析和比较。

任务：

1. 计算每平方英尺价格：基于 Sale_Price 列和 carpet_area 列的数据，在 Price_per_Square_Foot 字段中计算每笔交易的每平方英尺价格。

任务 2.2：根据地区分类房产以分析市场区域性

为了更好地理解和分析这些区域性特征，将地区进行价格水平分类是一种有效的方法。

任务：

1. Locality_Group 字段：基于 Locality 列的数据，在字段 Locality_Group 中将房产根据所在地区的价格水平分类为“High Price Area”、“Medium Price Area”和“Low Price Area”。

2. 分类标准：

Locality_Group	Locality
High Price Area	Greenwich、Fairfield、Stamford
Medium Price Area	Norwalk、West Hartford
Low Price Area	Bridgeport、Waterbury

任务 2.3：计算房间与浴室的比例以评估房产布局合理性

在房地产市场中，房间和浴室的数量比例是评估房产内部布局合理性的重要指标。

任务：

1. 计算房间数量与浴室数量之间的比例：基于 num_rooms 列和 num_bathrooms 列的数据，在 Room_to_Bathroom_Ratio 字段中计算每笔交易的房间数量与浴室数量之间的比例。

任务 2.4：计算实际出售价格与估价的差异

在房地产交易中，了解实际销售价格与估价之间的差异对于评估房产价值和市场情况非常重要。

任务：

1. 计算实际出售价格与估价的差异：基于 Sale_Price 列和 Estimated_Value 列的数据，在 Price_Difference 字段中计算每笔交易的实际销售价格与估价之间的差异。

任务 2.5：计算销售价格与估值的比率以分析市场反应

在房地产交易中，了解实际销售价格与估值之间的比率是至关重要的。这一比率可以揭示市场对房产估值的反应，帮助分析者和投资者评估房产价值与市场预期之间的关系。

任务：

1. 计算销售价格与估值的比率：基于 Sale_Price 列和 Estimated_Value 列的数据，在 Sale_to_Value_Ratio 中计算每笔交易的销售价格与估值之间的比率。

任务三：数据分析

任务 3.1：分析房地产市场价格溢价率

通过分析房地产交易数据和计算价格溢价率的平均值，可以深入了解实际销售价格与房产估值之间的差异，从而为投资者和政策制定者提供有力的决策支持。

任务：

1. 请基于房产数据集中的销售记录，使用 Sale_Price 和 Estimated_Value 两列数据。

2. 计算所有记录的价格溢价率平均值（百分比形式，四舍五入保留两位小数）。

3. 完成后，请将结果保存在名为 3.1 的工作表中。

任务 3.2：分析不同地区房价与销售价格的差异

通过量化高价区域、中价区域和低价区域内房产估值与实际销售价格之间的差异，可以更好地评估市场定价机制的效率和准确性。

任务：

1. 计算实际销售价格的平均值： 对高价区域（High Price Area）、中价区域（Medium Price Area）和低价区域（Low Price Area）分别计算房产的实际销售价格平均值。

2. 计算估值与实际销售价格的差异：计算上述三个地区的房产估值与实际销售价格之间的平均差异。

3. 结果处理与保存： 所有结果四舍五入至整数。并将最终结果保存在名为 3.2 的工作表中。

任务 3.3：分析不同地区房产税率的分布情况

通过研究不同地区的平均房产税税率，可以识别出税率最高和最低的地区。这将提供对各地区经济政策差异的洞见，并帮助理解地区间的经济负担差异。

任务：

1. 识别税率最高和最低的地区：明确平均房产税率最高的地区和最低的地区及各自的平均税率（小数形式，结果四舍五入保留两位小数）。

2. 结果处理与保存： 将最终结果保存到名为 3.3 的工作表中。

任务 3.4：分析不同房产类型的市场表现和受欢迎程度

在房地产市场中，不同类型的房产表现出各自独特的市场需求和价格水平。研究这些差异可以帮助开发商、投资者和政策制定者理解市场趋势，制定更有针对性的战略。

任务：

1. 分析市场表现： 识别市场上最受欢迎的房产类型，并统计该房产类型的销售数量。同时，计算该房产类型的平均销售价格。

2. 结果处理与保存： 四舍五入保留整数。并将分析结果保存到名为 3.4 的工作表中。

任务 3.5：探索房产市场的季节性和长期时间趋势

通过对 2021 年季度数据的深入分析，挖掘房产市场的季节性波动和长期趋势，特别是关注平均销售价格和销售数量的变化。此分析将帮助更好地理解市场动态，为未来的决策提供数据支持。

任务要求：

1. 季度数据统计：

计算 2021 年每个季度的房产平均销售价格，结果需四舍五入保留整数。

计算 2021 年每个季度的销售数量。

2. 结果存储： 整理计算结果并保存至 Excel 中的 3.5 工作表。

任务四：数据可视化

任务 4.1：各年度房产销售数量及波动趋势

房地产市场的年度销售数据提供了洞察市场波动和趋势的重要视角。通过绘制柱状图，可以直观地展示销售数量的变化及其年度涨跌幅度，从而帮助快速把握市场动态。

任务：

1. 柱状图： 绘制带有涨跌幅度的连续柱状图。

• 柱体颜色：黑色（RGB：0、0、0）

2. 涨跌幅度：

• 含义：每年的销售数量相对于前一年的变化百分比。

• 箭头线条：渐变线

颜色：

• 涨：绿色（RGB：112、173、71）和白色（RGB：255、555、255）

• 跌：红色（RGB：255、0、0）和白色（RGB：255、555、255）

3. 数据标签：

• 显示销售数量及涨跌幅度的具体数值

标签位置：

• 销售数量：轴内侧

• 涨：靠上

• 跌：靠下

涨跌标签颜色：

• 涨：绿色（RGB：112、173、71）

• 跌：红色（RGB：255、0、0）

销售数量标签颜色：

• 白色（RGB：255、555、255）

4. 其他： 不显示网格线和图例。

5. 结果保存： 将数据及图形保存到名为 4.1 的工作表中。

注意：图形需与示例图一致。

模块 B 数据展示与分享

模块描述：

在公司的数据分析团队中，您作为资深数据分析师备受信任，领导最近委托您负责一项关键性的任务，旨在帮助公司在激烈的市场竞争中保持领先地位。这项任务将为公司提供宝贵的洞察力，为领导团队提供决策依据，以助力公司在未来发展中做出及时的调整和战略部署。为了实现这一目标，您需要综合运用人场货 (People-Market-Product) 的方法论，并将其融入以下任务场景的操作:

客户分析场景:在这一场景中，您将深入研究客户的特征和行为，并将其与公司的产品或服务相结合，以获得对客户价值和忠诚度的全面了解。通过分析客户的特征、消费习惯、购买力和忠诚度，结合客户对不同产品的购买情况和满意度，您将能够确定哪些客户对公司的产品最感兴趣评估客户的忠诚度和满意度。同时，研究市场的竞争格局和趋势、竞争对手的客户群体和市场份额，将有助于您了解客户所处的市场环境和市场机遇，并制定相应的客户关系管理策略。

市场分析场景：在这一场景中，您将结合运输模式、订单优先级和地区市场，深入了解销售代表的表现和市场环境。通过分析销售代表的销售数据以及订单的地理分布情况，您将能够评估销售代表在不同地区的销售业绩和运输模式的效率，从而调整销售策略和资源配置。同时，结合利润和运费数据，您可以分析不同市场的盈利情况和成本分布，为销售代表提供更有效的市场竞争策略和销售技巧。

产品分析场景：在这一场景中，您将结合产品的销售情况和市场需求，全面了解产品的表现和竞争力。通过分析产品类别、子类别以及产品名称，您可以识别出热门产品和销售潜力产品，并了解不同产品在市场上的定位和竞争优势。结合销

售额和数量数据，您可以分析产品的市场份额和销售趋势，从而制定产品定价和促销策略，提高产品在市场上的竞争力和盈利能力。

任务一、客户分析

任务 1.1：

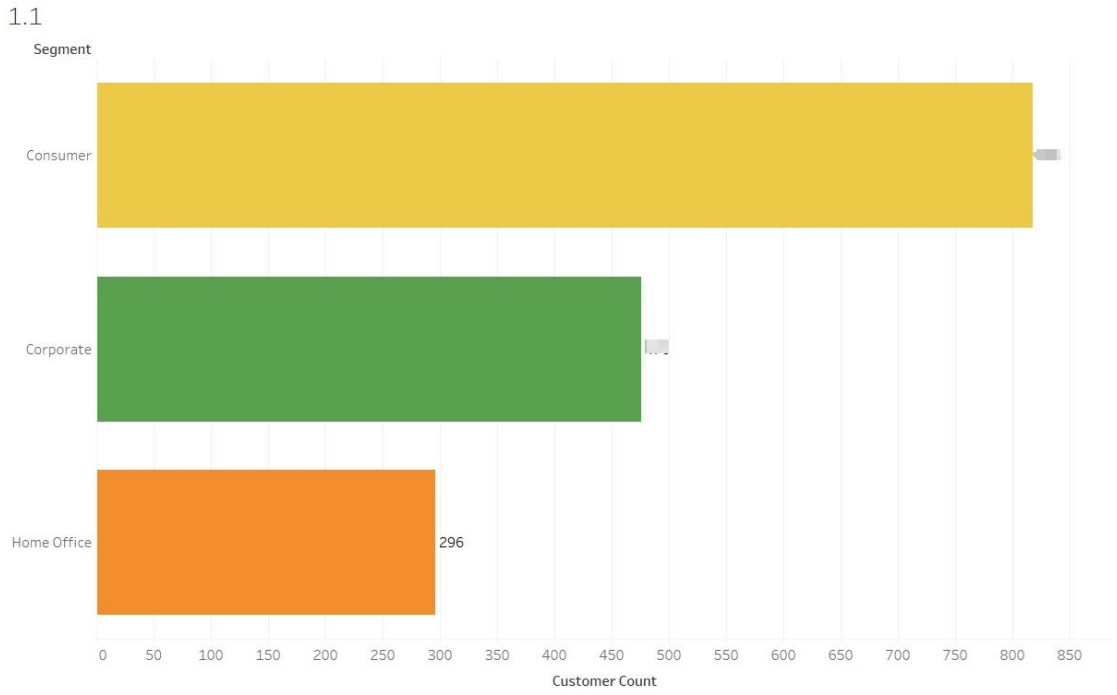
为了针对不同客户细分类型提供更为精准的产品和服务，请您展示一下各客户细分类型的客户分布情况。请根据以下要求，完成并保存位于工作簿中名为 1.1 的工作表。

具体要求：

图形名称：水平条

- 字段 Segment 设置在行，Customer Count 设置在列
- 显示 Customer Count 的标签
- 对 Segment 设置颜色，与示例图一致
- Segment 按 Customer Count 从上到下进行降序排序
- 将视图尺寸设置为"整个视图"

参考图形如下：



任务 1.2:

为了帮助公司制定区域性营销策略，请您展示一下各地区的客户采购情况。请根据以下要求，完成并保存位于工作簿中名为 1.2 的工作表。

具体要求:

图形名称：树状图

• 字段 Region 设置在行，Customer ID 设置在列(右击选择度量处的计数)，点击智能推荐处的树状图

• 显示 Region 与客户采购次数的标签

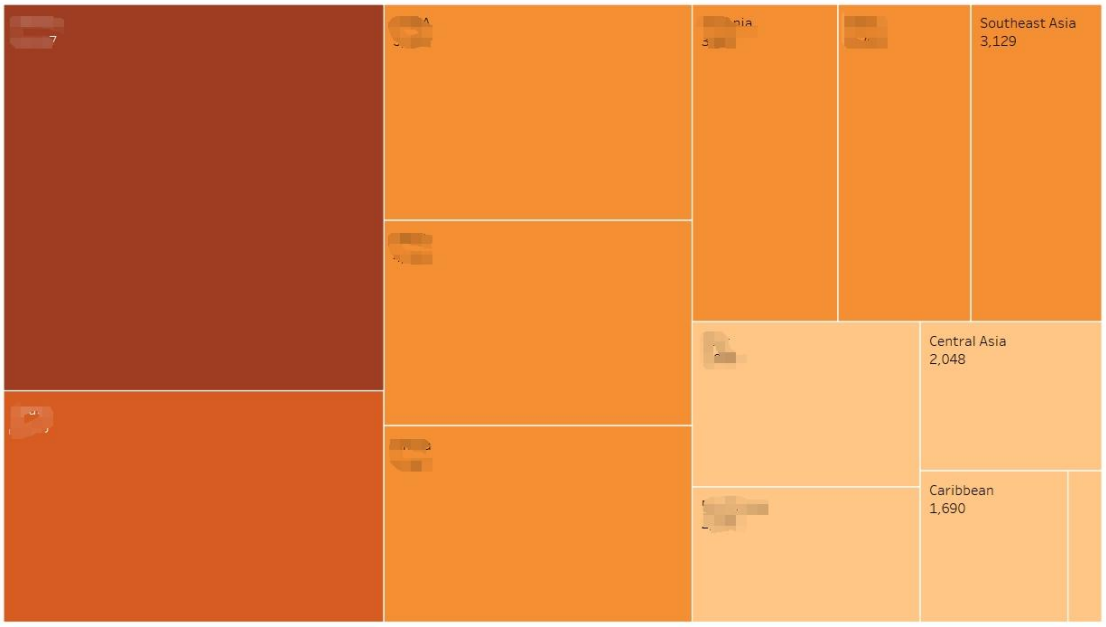
• 对客户采购次数设置色板颜色为"橙色"，渐变颜色设置为 4 阶，与示例图一致

• Region 的客户采购次数越多，矩形面积越大，颜色越深

• 将视图尺寸设置为"整个视图"

参考图形如下:

1.2



任务 1.3：

为了更全面地了解不同客户细分类型的盈利能力，请您展示一下每年各季度不同客户细分类型的利润情况。请根据以下要求，完成并保存位于工作簿中名为 1.3 的工作表。

具体要求：

图形名称：折线图

- 字段 Segment 和 Profit 设置在行，字段 Order Date 按照年和季度设置在列
- 显示 Profit 的标签
- 对年(Order Date)设置色板颜色，与示例图一致
- 将视图尺寸设置为"整个视图"

参考图形如下：

1.3

任务二、市场分析

任务 2.1:

为了更好地评估各个市场的销售表现，请您展示一下 2021 年各月份不同市场的订单量情况。请根据以下要求，完成并保存位于工作簿中名为 2.1 的工作表。

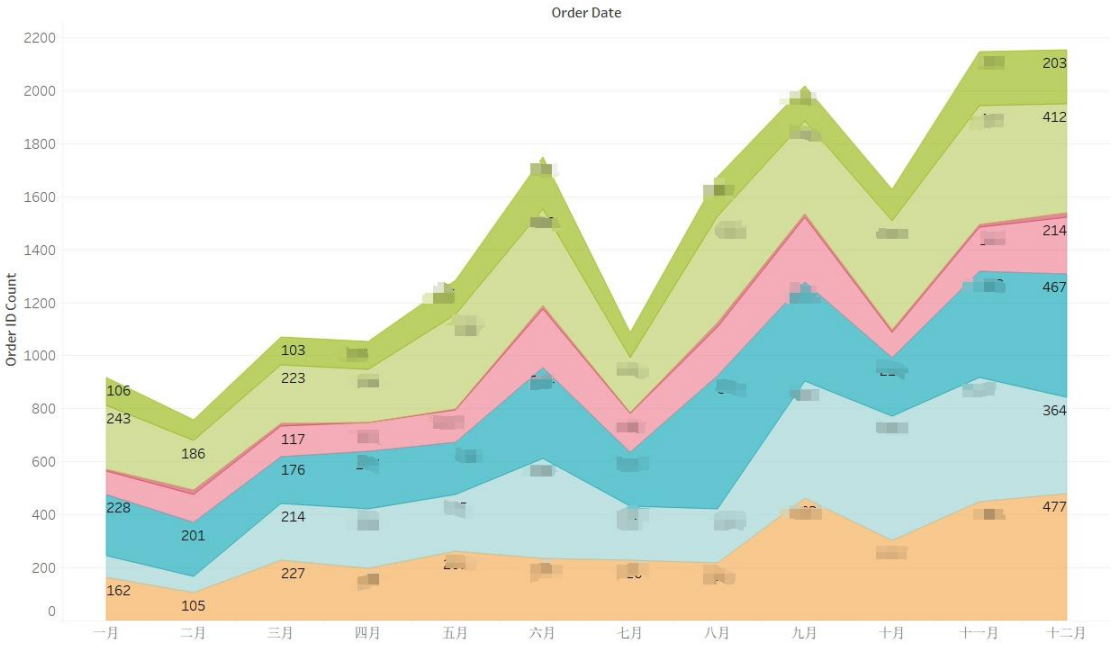
具体要求：

图形名称：堆积面积图

- 字段 Order ID Count 设置在行，字段 Order Date 按照月份设置在列
- Order Date 只保留 2021 年的数据
- 显示订单量的标签
- 标记处设置图形为区域形状
- 对 Market 设置色板颜色为"夏天"，与示例图一致
- 将视图尺寸设置为"整个视图"

参考图形如下：

2.1



任务 2.2:

为了制定更加精准的营销策略和资源分配计划，请您展示一下各个市场的总销售金额及总盈利情况。请根据以下要求，完成并保存位于工作簿中名为 2.2 的工作表。

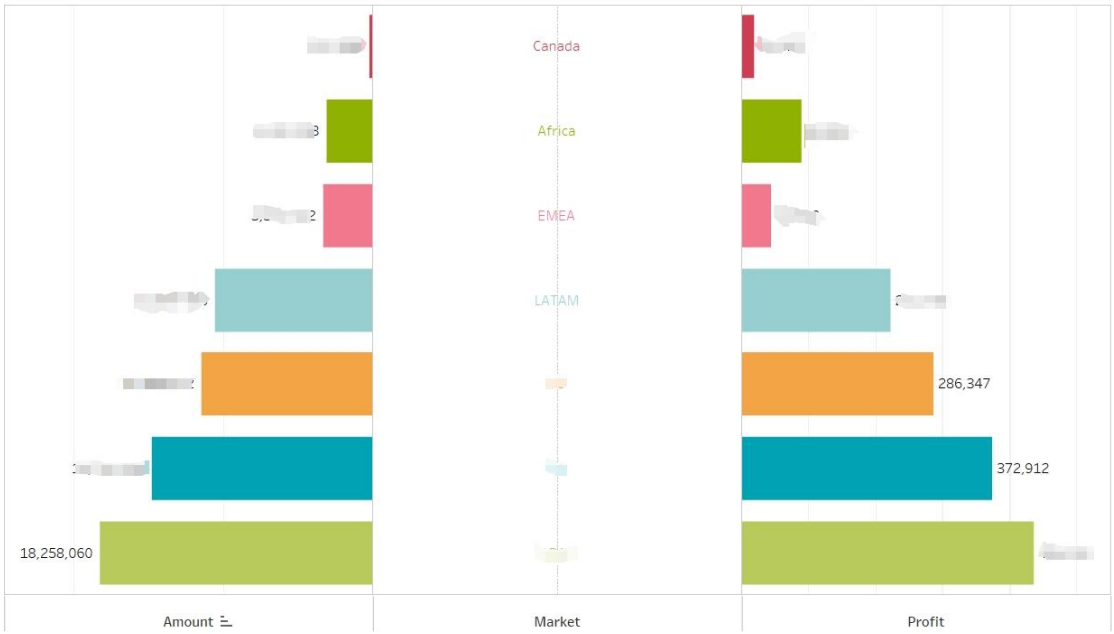
具体要求：

图形名称： 蝴蝶图

- 总销售金额设置在左侧，Market 设置在中间，总利润设置在右侧，与示例图一致
- Amount(销售金额) = Sales * Quantity
- 显示总销售金额、Market 和总利润的标签
- 对 Market 设置色板颜色为"夏天"，与示例图一致
- 取消显示行标题，设置 x 轴标题，但不显示刻度线，与示例图一致
- Market 按总销售金额从上到下进行升序排序
- 将视图尺寸设置为"整个视图"

参考图形如下：

2.2



任务三、产品分析

任务 3.1:

为了优化产品的库存管理和市场推广策略，请您展示一下各产品类别的总数量占比情况。请根据以下要求，完成并保存位于工作簿中名为 3.1 的工作表。

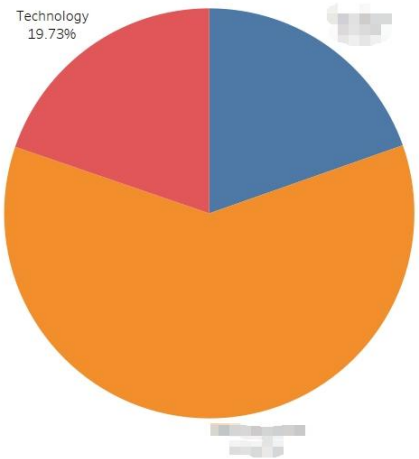
具体要求：

图形名称：饼图

- 字段 Category 设置在行，Quantity 设置在列
- "智能推荐"处选择"饼图"
- 显示各产品类别及其总数量的占比(百分比)的标签
- 对 Category 设置色板颜色，与示例图一致
- 将视图尺寸设置为"整个视图"

参考图形如下：

3.1



模块 C 数据开发与应用

模块描述：

随着互联网和社交媒体的兴起，越来越多的平台提供了读者发布图书评论和评分的功能。这些平台可以是在线书店、社交媒体平台、图书社区或图书阅读应用程序等。读者可以在这些平台上分享他们对图书的评价、感受和建议，与其他读者进行交流和讨论。图书评论数据在这些平台上积累起来，形成了一个庞大的数据集，包含了大量的评论文本、评分、日期、读者信息等。通过对这些数据进行分析，可以了解读者对不同图书的喜好、评价和反馈，了解图书的受欢迎程度和质量，以及读者对图书内容、情节、人物等方面的关注点。

对于平台来说，图书评论数据是宝贵的资产。它们可以利用这些数据来改进用户体验、优化推荐算法、了解市场需求和读者喜好，以及为图书行业的决策提供依据。通过对评论数据的分析，平台可以发现热门图书、推荐相关图书、改进图书推荐算法、提供个性化的图书推荐等。图书评论数据在各种平台上积累起来，为读者提供了一个分享和交流的空间，同时也为平台和图书行业提供了宝贵的信息和洞察。通过对这些数据的分析，可以更好地了解读者需求、改进图书质量和推广策略，提供更好的图书选择和体验。

Python 作为一种功能强大且易于使用的编程语言，在图书数据分析中扮演着重要的角色。它提供了丰富的数据处理、统计分析和可视化库，使得对图书行业数据进行深入分析变得更加高效和灵活。Python 的应用包括数据收集、数据清洗、数据处理、数据分析和数据可视化等方面。通过使用 Python 的库和工具，如 BeautifulSoup 和 Scrapy，可以从各种数据源中收集图书数据。使用 Pandas 库进行数据处理，可以对数据进行清洗、转换和整合，以便进行后续的分析。NumPy 库提供了高效的数值计算功能，而 SciPy 库则提供了各种科学计算和统计分析的工具。在数据分析阶段，pandas 和 NumPy 库可以用于对图书数据进行统计分析、

聚合计算和数据建模。机器学习库（如 scikit-learn）可以用于构建预测模型，以预测图书销售趋势或读者喜好。最后，可视化库（如 Matplotlib 和 Seaborn）可以用于创建各种类型的图表和可视化，以便更好地理解和传达图书数据的洞察。这些库提供了丰富的绘图功能，可以生成直观、美观的图表，帮助用户发现数据中的模式和趋势。

任务一、数据探索与处理

1.1 读取有关图书数据集的所有表中的所有数据，数据路径如下，分别保存到对应变量 metadata、ratings、reviews、survey_answers、tag_count 和 tags，然后运行给出的答案保存代码保存答案。

1.2 处理 lang 和 description 字段，按下面要求对两个字段的缺失值进行填充处理，处理后的结果更新到 metadata 变量中，并运行给出的答案保存代码保存答案。

- 将 lang 的缺失值填充为该字段的众数
- 将 description 的缺失值填充为"no"

1.3 删除 reviews 表中重复的行，只保留重复出现行中第一次出现的行，并运行给出的答案保存代码保存答案。

1.4 筛选 survey_answers 表中 score 字段不为-1 的数据，将结果保存到变量 survey_answers_clean 中，并运行给出的答案保存代码保存答案。

1.5 新增 comment_number 字段到变量 metadata 中，统计 reviews 表每本图书的评论数量到 comment_number 字段中，并运行给出的答案保存代码保存答案。

任务二、图书基本情况分析

2.1 分析 metadata 表中 comment_number 字段，找出评论数量最多的图书标题 title，结果保存到变量 B_2_1，并运行给出的答案保存代码保存答案。

2.2 分析 tags 表中 tag 字段，被贴上 dark 标签的图书有多少本？结果保存到变量 B_2_2，并运行给出的答案保存代码保存答案。

2.3 分析 metadata 表中 lang 字段，图书语言为美式英语（en-US）的图书数量占图书总数量的比例（小数形式，四舍五入，保留两位小数），结果保存到变量 B_2_3，并运行给出的答案保存代码保存答案 。

任务三、图书评分分析

3.1 分析 metadata 表和 ratings 表，找出平均评分最高的五本图书标题 title。结果保存为变量 B_3_1，并运行给出的答案保存代码保存答案 。

3.2 分析 metadata 表和 ratings 表，在出版图书超过 10 本的作者中，哪位作者写的书最受欢迎，结果保存为变量 B_3_2，并运行给出的答案保存代码保存答案 。

任务四、调研情况分析

4.1 分析 survey_answers_clean 表，统计在参与调查的图书中，被标记不同标签最多的图书 ID item_id。结果保存为变量 B_4_1，并运行给出的答案保存代码保存答案 。

4.2 分析 survey_answers_clean 表和 ratings 表，在参与了问卷调查的用户中随机抽取一名用户，该用户参与了对图书评分的概率有多大（结果保留两位小数）？结果保存为变量 B_4_2，并运行给出的答案保存代码保存答案 。

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	数据展示与分享	30
C	数据开发与应用	30
D	作品答辩	10
合计		100

注：样题最终解释权归组委会所有