2025

BRICS SKILLS COMPETITION

(BRICS+ FUTURE SKILLS & TECH CHALLENGE)

Data Analysis and Visualization

BRICS-FS-36

Test Project

(International Final_Offline)

May,2025

Contents

1. Competition Format

This competition is an individual event.

2. Competition Content

The competition consists of four modules, which participants must complete in sequence. During the competition, contestants will be provided with a standardized set of materials, including the problem set, competition equipment, basic operation manuals, and the necessary data sources or technical prerequisites to ensure the independence and fairness of each task module.

The competition content is based on data analysis and visualization, and includes the following task modules:

Module A: Data Acquisition and Processing

Module B: Data Presentation and Sharing

Module C: Data Development and Application

Module D: Project Presentation

Participants who fail to comply with occupational health, safety, and environmental requirements, or who endanger themselves or others, may be disqualified.

After completing the competition, participants' submissions will be evaluated and scored by the panel of judges.

3. Project modules and time requirements

3.1 Time Allocation

The Data Analysis and Visualization competition consists of four modules, with a total duration of 370 minutes required for all participants. For specific module names and time requirements, please refer to Table 1: *List of Project Modules and Time Requirements*.

Table 1: List of Project Modules and Time Requirements

No.	Module Name	Duration
1	Module A: Data Acquisition and Processing	120min
2	Module B: Data Presentation and Sharing	120min
3	Module C: Data Development and Application	120min
4	Module D: Project Presentation	10min

3.2 Task Details

Module A: Data Acquisition and Processing

Module Description:

In 2008, the United States experienced an unprecedented real estate market crisis, which had a profound impact on the U.S. economy and reverberated across the global economy, becoming a major catalyst for the global financial crisis. The crisis originated from the collapse of the subprime

mortgage market. Since the late 1990s, the real estate market had undergone a period of rapid growth, with housing prices continually rising and homeownership reaching record highs. However, as financial institutions loosened lending standards, a large number of high-risk subprime loans were issued to borrowers with poor credit.

These subprime loans were mostly adjustable-rate mortgages, and when interest rates rose, many borrowers were unable to make timely payments, leading to a wave of foreclosures. In 2007, the subprime crisis began to surface, with a flood of foreclosed homes hitting the market and housing prices starting to plummet. By 2008, the crisis had fully erupted, with the collapse in housing prices triggering a chain reaction in financial markets. Major financial institutions either went bankrupt or were acquired, and global financial markets were thrown into turmoil. Particularly after experiencing the crisis of 2008, understanding how the market recovered from the crisis provides valuable insights for preventing similar crises in the future.

Task 1: Data Preparation

Task 1.1: Real Estate Transaction Data Processing and Analysis

To conduct more effective time series analysis and monthly market trend analysis, it is necessary to process and refine the existing data, especially by adjusting the format and details of the date data, so as to more accurately capture market changes and trends.

Tasks:

1. Format the Date: Adjust the data format of the Date column to the standard form (yyyy-mm-dd) to unify the date representation for subsequent processing

and analysis.

2. Create a Month Field: In the Month column, extract the month (represented as a numerical value) from the `Date` column for each transaction.

3. Extract month information：=MONTH()

Task 1.2: Optimize the Real Estate Transaction Dataset to Enhance Data Quality

By correcting known data quality issues, ensure that the dataset provides reliable support for further statistical analysis and model training.

Tasks:

1. Correct Valuation Outliers: In the Estimated_Value column, change all outlier values of 0 to the column's average value of 448,673.

2. Standardize Property Types: In the Property column, change all instances of the unknown property type marked as ? to the most common property type, Single Family.

Task 1.3: Fill in Missing Values in the Real Estate Dataset

To ensure the dataset provides accurate and reliable support in statistical analysis and model training, it is necessary to address data quality issues and fill in missing values.

Task:

1.Fill Missing Values in the carpet_area Column: In the carpet_area column, there are some missing values. We will fill these using the average value of 1111.

Task 2: Data Processing

Task 2.1: Calculate Price per Square Foot

By calculating the price per square foot, we can gain a more accurate understanding of the relative value of different properties, enabling deeper market analysis and comparison.

Task:

1. Calculate Price per Square Foot: Based on the data in the Sale_Price and carpet_area columns, calculate the price per square foot for each transaction in the Price_per_Square_Foot field .

Task 2.2: Categorize Properties by Region to Analyze Market Area Characteristics

To better understand and analyze regional characteristics, categorizing areas by price levels is an effective approach.

Tasks:

1. Locality_Group Field: Based on the data in the Locality column, categorize the properties into “High Price Area,” “Medium Price Area,” and “Low Price Area” in the Locality_Group field.

2. Classification Criteria:

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

Task 2.3: Calculate Room-to-Bathroom Ratio to Evaluate Property Layout

In the real estate market, the ratio of rooms to bathrooms is an important indicator of the reasonableness of a property's internal layout.

Task:

1. Calculate the Ratio of Rooms to Bathrooms: Based on the data in the num_rooms and num_bathrooms columns, calculate the ratio of rooms to bathrooms for each transaction in the Room_to_Bathroom_Ratio field.

Task 2.4: Calculate the Difference Between Sale Price and Estimated Value

In real estate transactions, understanding the difference between the actual sale price and the estimated value is crucial for evaluating property value and market conditions.

Task:

1. Calculate the Difference Between Sale Price and Estimated Value: Based on the data in the Sale_Price and Estimated_Value columns, calculate the difference between the actual sale price and the estimated value for each transaction in the Price_Difference field.

Task 2.5: Calculate the Ratio of Sale Price to Estimated Value to Analyze Market Response.

In real estate transactions, understanding the ratio between the actual sale price and the estimated value is crucial. This ratio can reveal the market's response to property valuation, helping analysts and investors assess the relationship between property value and market expectations.

Task:

1. Calculate the Ratio of Sale Price to Estimated Value: Based on the data in the Sale_Price and Estimated_Value columns, calculate the ratio of sale price to estimated value for each transaction in the Sale_to_Value_Ratio field .

Task 3: Data Analysis

Task 3.1: Analyze the Real Estate Market Price Premium Rate

By analyzing real estate transaction data and calculating the average price premium rate, we can gain deeper insights into the difference between the actual sale price and the property's estimated value, providing strong decision-making support for investors and policymakers.

Task:

1. Use the Sale_Price and Estimated_Value columns in the property dataset.

2. Calculate the average price premium rate for all records (as a percentage, rounded to two decimal places).

3. Once completed, save the result in a worksheet named 3.1.

Task 3.2: Analyze the Differences Between Property Values and Sale Prices in Different Regions.

By quantifying the differences between property valuations and actual sale prices in high-price, medium-price, and low-price areas, we can better evaluate the efficiency and accuracy of market pricing mechanisms.

Tasks:

1. Calculate the Average Sale Price: Calculate the average sale price of properties in high-price areas (High Price Area), medium-price areas (Medium Price Area), and low-price areas (Low Price Area).

2. Calculate the Difference Between Valuation and Sale Price: Calculate the average difference between property valuation and actual sale price in the three regions.

3. Result Processing and Saving: Round all results to the nearest integer and save the final results in a worksheet named 3.2.

Task 3.3: Analyze the Distribution of Property Tax Rates in Different Regions.

By studying the average property tax rates in different regions, we can identify the areas with the highest and lowest tax rates. This provides insights into the economic policy differences between regions and helps understand the economic burden differences between them.

Tasks:

1. Identify the Regions with the Highest and Lowest Tax Rates: Identify the regions with the highest and lowest average property tax rates and their respective average tax rates (in decimal form, rounded to two decimal places).

2. Result Processing and Saving: Save the final results in a worksheet named 3.3.

Task 3.4: Analyze the Market Performance and Popularity of Different Property Types.

In the real estate market, different property types exhibit unique market demand and price levels. Studying these differences can help developers, investors, and policymakers understand market trends and develop more targeted strategies.

Tasks:

1. Analyze Market Performance: Identify the most popular property type in the market, count the number of sales for that property type, and calculate the average sale price for that type.

2. Result Processing and Saving: Round the results to the nearest integer and save the analysis results in a worksheet named 3.4.

Task 3.5: Explore the Seasonality and Long-term Trends of the Real Estate Market.

By conducting a thorough analysis of quarterly data from 2021, this task aims to uncover seasonal fluctuations and long-term trends in the real estate market, with a particular focus on changes in average sales price and sales volume. This analysis will help in gaining a deeper understanding of market dynamics and provide valuable insights for future decision-making.

Task Requirements:

1. Quarterly Data Analysis:

Calculate the average sales price of real estate for each quarter in 2021, rounding the results to the nearest integer.Calculate the sales volume for each quarter in 2021.

Result Storage: Organize the calculated results and save them in the 3.5 worksheet of an Excel file.

Task 4: Data Visualization

Task 4.1: Analysis of Annual Property Sales Volume and Trend Fluctuations

Annual sales data in the real estate market provides a crucial perspective for understanding market fluctuations and trends. By creating a bar chart, you can visually display the changes in sales volume and the percentage increase or decrease each year, helping to quickly grasp market dynamics.

Tasks:

1. Bar Chart: Create a continuous bar chart with percentage change.

Column color: Black (RGB: 0, 0, 0)

2. Percentage Change:

Meaning: The percentage change in sales volume each year relative to the

previous year.

Arrow Lines: Gradient lines

Colors:

· Increase: Green (RGB: 112, 173, 71) and White (RGB: 255, 255, 255)

· Decrease: Red (RGB: 255, 0, 0) and White (RGB: 255, 255, 255)

3. Data Labels:

Display the specific values for sales volume and percentage change.

Label Position:

· Sales Volume: Inside the axis

· Increase: At the top

· Decrease: At the bottom

Label Colors:

· Increase: Green (RGB: 112, 173, 71)

· Decrease: Red (RGB: 255, 0, 0)

· Sales Volume: White (RGB: 255, 255, 255)

4. Others: Do not display grid lines and legends.

5. Result Saving: Save the data and chart in a worksheet named 4.1.

Note: The chart should match the provided example.

Module B: Data Presentation and Sharing

Module Description:

As a senior data analyst highly trusted within the company's data analysis team, your leadership has entrusted you with a critical task aimed at helping the company stay ahead in the fierce market competition. This task will provide

valuable insights to the leadership team, enabling timely adjustments and strategic deployment for the company's future development. To achieve this goal, you will need to apply the People-Market-Product methodology and integrate it into the following task scenarios:

Customer Analysis Scenario: In this scenario, you will conduct an in-depth study of customer characteristics and behavior, combining this with the company's products or services to gain a comprehensive understanding of customer value and loyalty. By analyzing customer characteristics, consumption habits, purchasing power, and loyalty, alongside customers' purchase behavior and satisfaction with different products, you will be able to identify which customers are most interested in the company's products and evaluate their loyalty and satisfaction. At the same time, studying the competitive landscape and trends in the market, as well as the customer base and market share of competitors, will help you understand the market environment and opportunities for your customers, allowing you to develop corresponding customer relationship management strategies.

Market Analysis Scenario: In this scenario, you will analyze the transportation modes, order priorities, and regional markets to gain a deeper understanding of the performance of sales representatives and the market environment. By analyzing the sales data of sales representatives and the geographical distribution of orders, you will be able to evaluate the sales performance of sales representatives in different regions and the efficiency of transportation modes, allowing for adjustments to sales strategies and resource allocation. Additionally, by combining profit and shipping cost data,

you can analyze the profitability and cost distribution in different markets, providing sales representatives with more effective market competition strategies and sales techniques.

Product Analysis Scenario: In this scenario, you will combine product sales data and market demand to gain a comprehensive understanding of product performance and competitiveness. By analyzing product categories, sub-categories, and product names, you can identify popular products and products with sales potential, as well as understand the market positioning and competitive advantages of different products. Combining sales and quantity data, you can analyze product market share and sales trends, helping to formulate pricing and promotional strategies to enhance product competitiveness and profitability in the market.

Task 1: Customer Analysis

Task 1.1:

To provide more precise products and services for different customer segments, please display the customer distribution for each customer segment type. Based on the following requirements, complete and save the worksheet named 1.1 in the workbook.

Specific Requirements:

Chart Name: Horizontal Bar

· Set Segment in rows and Customer Count in columns

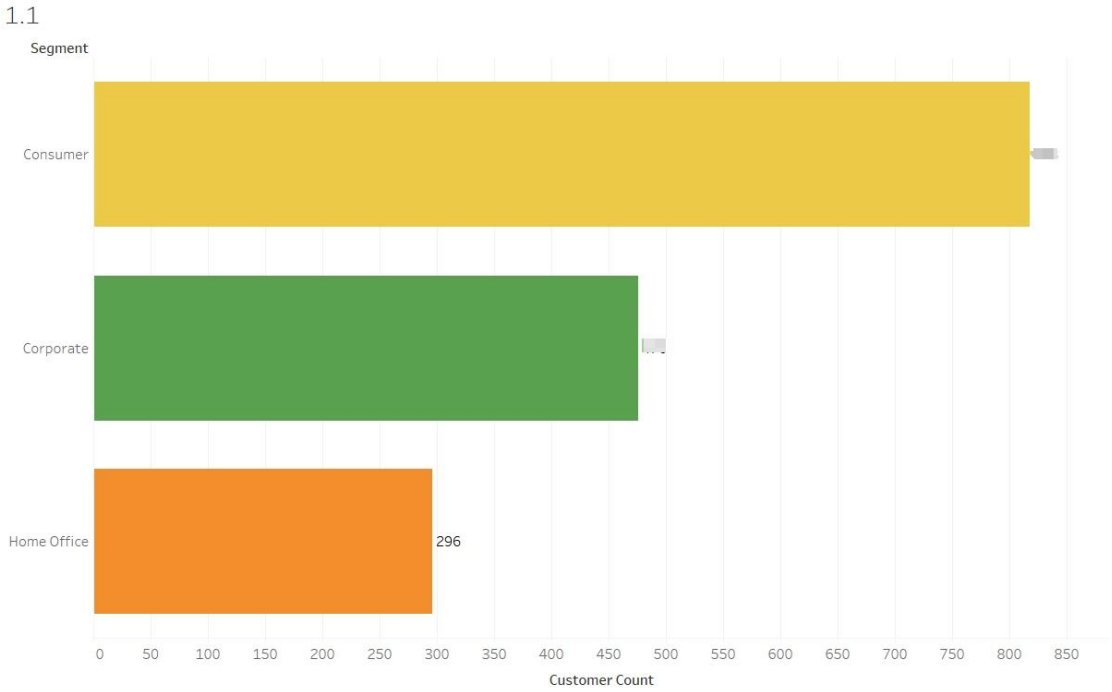
· Display the label for Customer Count

· Set colors for Segment, consistent with the example chart

· Sort Segment in descending order by Customer Count

· Set the view size to "Entire View"

Reference chart:



Task 1.2:

To help the company formulate regional marketing strategies, please display the customer purchase situation for each region. Based on the following requirements, complete and save the worksheet named 1.2 in the workbook.

Specific Requirements:

Chart Name: Treemap

· Set the field Region in the row and Customer ID in the column (right click to select the count at the measurement), and click the tree diagram at Smart Recommendation

· Display the labels for Region and the number of customer purchases

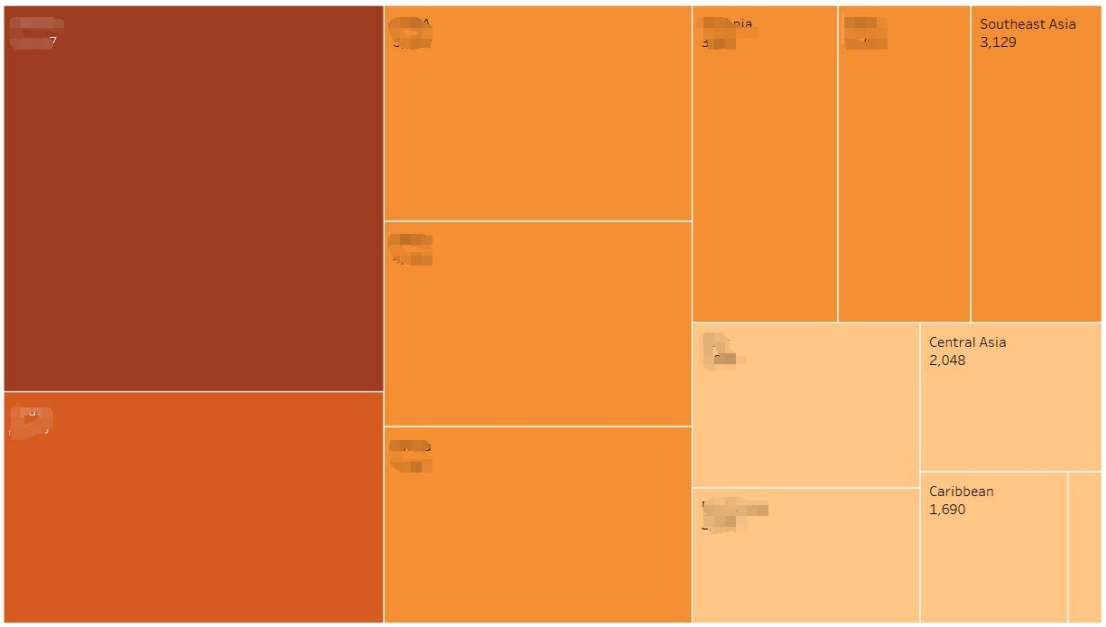
· Set the color palette for the number of customer purchases to "Orange", with a 4-step gradient, consistent with the example chart

· The more customer purchases a Region has, the larger the rectangle area and the darker the color

· Set the view size to "Entire View"

Reference chart:

1.2



Task 1.3:

To gain a more comprehensive understanding of the profitability of different customer segment types, please display the profit situation for each customer segment type per quarter each year. Based on the following requirements, complete and save the worksheet named 1.3 in the workbook.

Specific Requirements:

Chart Name: Line Chart

· Set Segment and Profit in rows, and set Order Date in columns by year and quarter

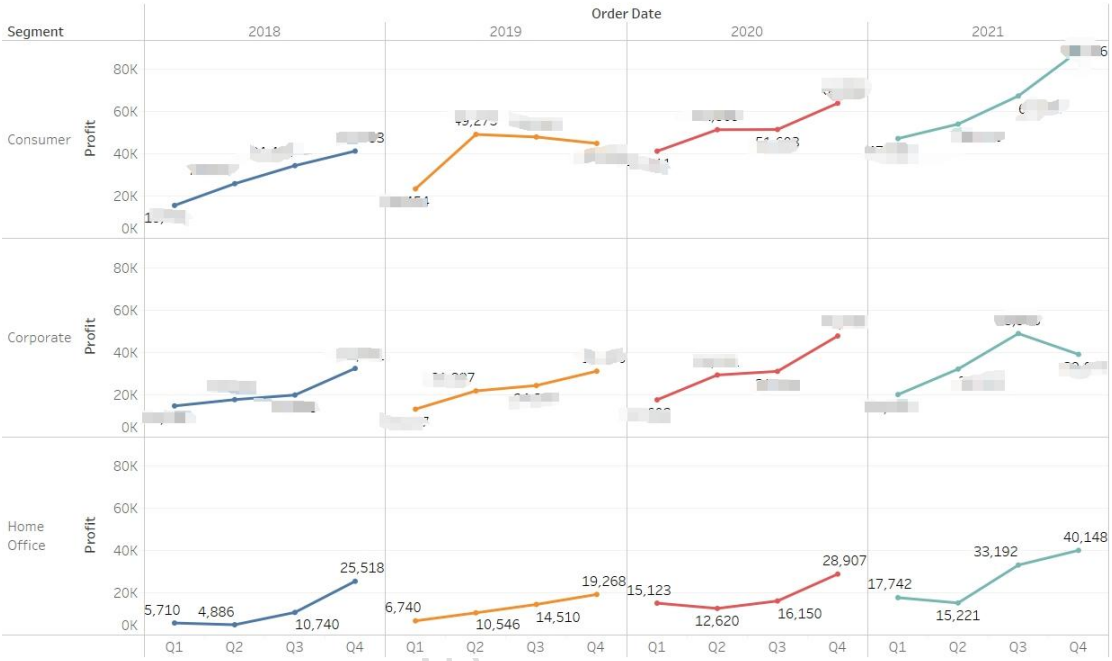
· Display the label for Profit

· Set the color palette for the year (Order Date), consistent with the example chart

· Set the view size to "Entire View"

Reference chart:

1.3



Task 2: Market Analysis

Task 2.1:

To better assess the sales performance of different markets, please display the order volume for each market in 2021 by month. Based on the following requirements, complete and save the worksheet named 2.1 in the workbook.

Specific Requirements:

Chart Name: Stacked Area Chart

· Set Order ID Count in rows, and set Order Date by month in columns

· Only retain data from 2021 for Order Date

· Display the order volume labels

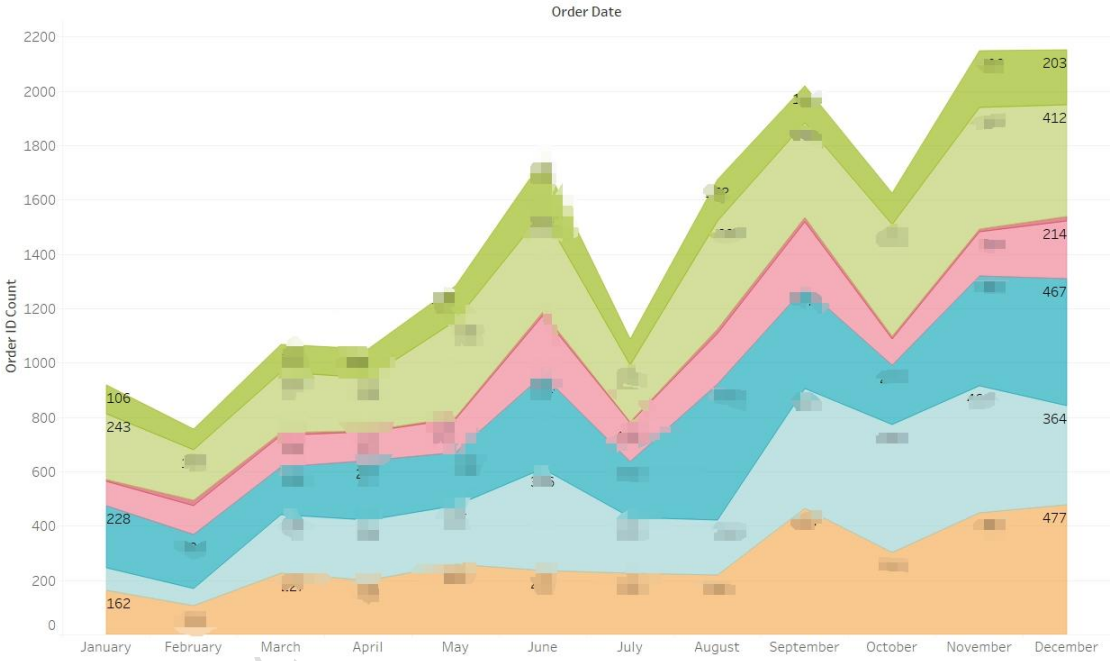
· Set the mark type to area shape

· Set the color palette for Market to "Summer", consistent with the example chart

· Set the view size to "Entire View"

Reference chart:

2.1



Task 2.2:

To develop more precise marketing strategies and resource allocation plans, please display the total sales amount and total profit for each market. Based on the following requirements, complete and save the worksheet named 2.2 in the workbook.

Specific Requirements:

Chart Name: Butterfly Chart

· Set the total sales amount on the left, Market in the middle, and total profit on the right, consistent with the example chart

· Amount (Sales Amount) = Sales * Quantity

· Display the labels for total sales amount, Market, and total profit

Set the color palette for Market to "Summer", consistent with the example chart

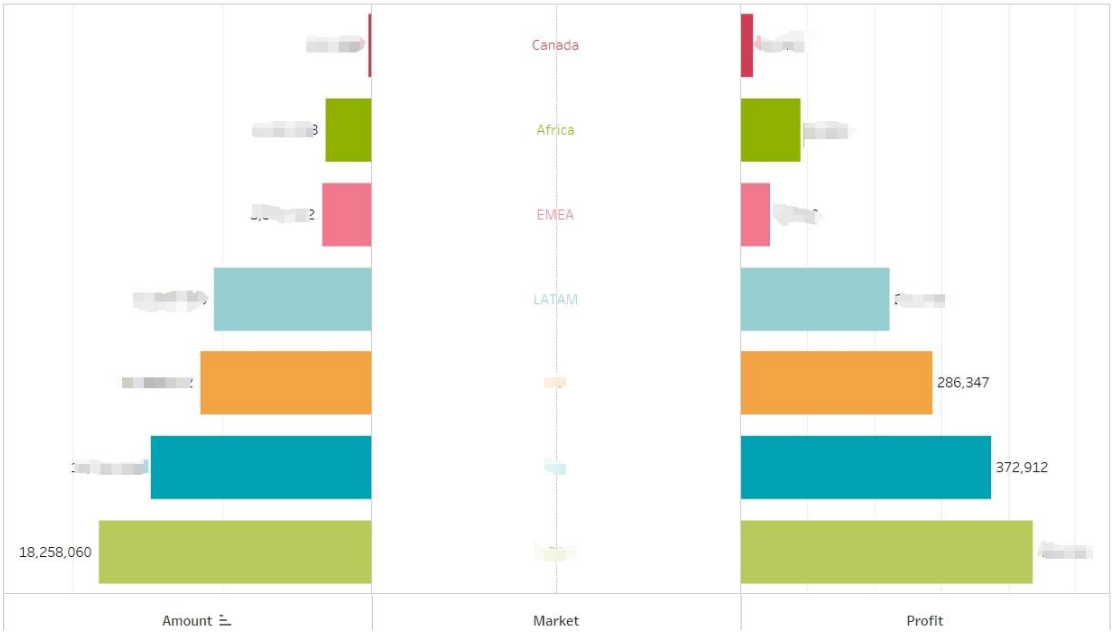
· Disable the display of row headers, set the x-axis title, but hide the tick marks to match the example diagram

· Sort Market in ascending order by total sales amount

· Set the view size to "Entire View"

Reference chart:

2.2



Task 3: Product Analysis

Task 3.1:

To optimize product inventory management and marketing strategies, please display the total quantity proportion for each product category. Based on the following requirements, complete and save the worksheet named 3.1 in the

workbook.

Specific Requirements:

Chart Name: Pie Chart

· Set Category in rows and Quantity in columns

· Choose "Pie Chart" under "Smart Recommendations"

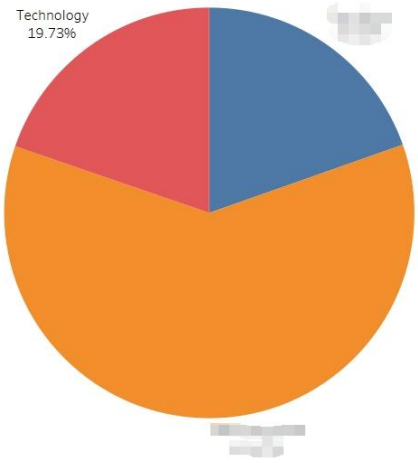
· Display labels for each product category and its total quantity proportion (percentage)

· Set the color palette for Category, consistent with the example chart

· Set the view size to "Entire View"

Reference chart:

3.1



Module C: Data Development and Application

Module Description:

With the rise of the internet and social media, an increasing number of platforms offer readers the ability to post book reviews and ratings. These platforms can include online bookstores, social media platforms, book communities, or reading apps. Readers can share their evaluations, impressions, and suggestions about books, and engage in discussions with other readers. As reviews accumulate on these platforms, they form a vast dataset containing a wealth of review texts, ratings, dates, reader information, and more. By analyzing this data, insights can be gained into readers' preferences, feedback on different books, and the popularity and quality of books, as well as the aspects of content, plot, and characters that draw readers' attention.

For platforms, book review data is a valuable asset. It can be used to enhance user experience, optimize recommendation algorithms, understand market demand and reader preferences, and support decision-making in the book industry. Through data analysis, platforms can identify popular books, recommend related titles, improve recommendation algorithms, and offer personalized book suggestions. The accumulation of book review data on various platforms not only provides a space for readers to share and interact but also offers valuable information and insights for platforms and the book industry. Analyzing this data helps better understand reader needs, improve

book quality and marketing strategies, and offer a better selection and experience for readers.

Python, as a powerful and user-friendly programming language, plays a crucial role in book data analysis. It offers a rich set of libraries for data processing, statistical analysis, and visualization, making in-depth analysis of book industry data more efficient and flexible. Python's applications range from data collection, cleaning, processing, analysis, to visualization. Libraries and tools such as BeautifulSoup and Scrapy can be used to collect book data from various sources. The Pandas library facilitates data cleaning, transformation, and integration for subsequent analysis. NumPy provides efficient numerical computation, while SciPy offers tools for scientific computation and statistical analysis. In the data analysis phase, Pandas and NumPy can be utilized for statistical analysis, aggregation, and data modeling of book data. Machine learning libraries such as scikit-learn can be employed to build predictive models for forecasting book sales trends or reader preferences. Finally, visualization libraries like Matplotlib and Seaborn enable the creation of various types of charts and visualizations to better understand and convey insights from book data. These libraries offer extensive plotting capabilities to generate intuitive and aesthetically pleasing charts that help users discover patterns and trends in the data.

Task 1: Data Exploration and Processing

1.1 Read all data from all the tables related to the book datasets, with the data paths given below, and save them to the corresponding variables metadata, ratings, reviews, survey_answers, tag_count, and tags, then run the

provided answer-saving code to save the answers.

1.2 Handle the missing values for the lang and description fields according to the following requirements, update the results in the metadata variable, and run the provided answer-saving code to save the answers.

· Fill the missing values in lang with the mode of the field.

· Fill the missing values in description with "no".

1.3 Remove duplicate rows in the reviews table, keeping only the first occurrence of each duplicate row, and run the provided answer-saving code to save the answers.

1.4 Filter the survey_answers table to keep only the data where the score field is not -1, save the results to the variable survey_answers_clean, and run the provided answer-saving code to save the answers.

1.5 Add a new comment_number field to the metadata variable, count the number of reviews for each book in the reviews table and store the count in the comment_number field, then run the provided answer-saving code to save the answers.

Task 2: Analysis of Basic Information About Books

2.1 Analyze the comment_number field in the metadata table and find the book title title with the highest number of comments. Save the result to the variable B_2_1, then run the provided answer-saving code to save the answer.

2.2 Analyze the tag field in the tags table. How many books have been tagged with the dark label? Save the result to the variable B_2_2, then run the provided answer-saving code to save the answer.

2.3 Analyze the lang field in the metadata table. Calculate the proportion (in decimal form, rounded to two decimal places) of books in American English

(en-US) relative to the total number of books. Save the result to the variable B_2_3, then run the provided answer-saving code to save the answer.

Task 3: Analysis of Book Ratings

3.1 Analyze the metadata and ratings tables to find the five books with the highest average rating. Save the result as the variable B_3_1, then run the provided answer-saving code to save the answer.

3.2 Analyze the metadata and ratings tables. Among authors who have published more than 10 books, which author has written the most popular books? Save the result as the variable B_3_2, then run the provided answer-saving code to save the answer.

Task 4: Analysis of Survey Results

4.1 Analyze the survey_answers_clean table to find the book with the most distinct tags. Save the book's item_id to the variable B_4_1, then run the provided answer-saving code to save the answer.

4.2 Analyze the survey_answers_clean and ratings tables. Among users who participated in the survey, what is the probability (in decimal form, rounded to two decimal places) that a randomly selected user also rated a book? Save the result to the variable B_4_2, then run the provided answer-saving code to save the answer.

4. Scoring Criteria

The scoring criteria for the project modules refer to Table 2.

Table 2: Scoring Criteria

Module	Task	Score
A	Data Acquisition and Processing	30
B	Data Presentation and Sharing	30
C	Data Development and Application	30
D	Project Presentation	10
Total		100

Note: The final interpretation of the sample questions belongs to the organizing committee.