ETW2001 FOUNDATIONS OF DATA ANALYSIS

SEMESTER 1 - 2024

Assessment 2 (30%)

This assessment is designed to evaluate your understanding and application of R programming skills in data joining and visualization (unit learning outcomes 3). The assessment is divided into two sections: Section A, which involves practical coding tasks for data joining, and Section B, in which you have to write R codes to create visualizations provided in an article.

Assessment Format

Section A: Data joining (30 Marks)

This section requires you to complete practical tasks using R, data joining using dplyr. You will work with provided datasets to demonstrate your ability to apply programming concepts in real-world scenarios.

Section B: Data visualization (30 Marks)

Write R codes to visualize the charts provided in the article using components of ggplot2. Do not have to display the charts (output).

Instructions

1. It's recommended to read through all questions and tasks carefully before working on them.
2. Write your code as clearly and concisely as possible**. Include comments** in your code to explain your logic and steps where necessary.
3. For tasks requiring explanations, clarity, and depth of understanding will be valued. Make sure your explanations are well-articulated and directly address the question.
4. Ensure your work is submitted in the specified format. Compile all R codes (Section A and B) into a single ***.R* script.** Include the screenshots of outputs from RStudio into a word document, convert it to pdf format. Write your answers in the same Word document for all the short answer questions. You have to submit two files, **one R Script, and one PDF report.**
5. This assessment is to be **completed individually**. Collaboration with peers is not allowed, and all submitted work must be yours. Any **references to external sources or documentation should be appropriately cited**.
6. If you encounter any technical difficulties, notify your instructor as soon as possible.

SECTION A Data Joining using dplyr (30 Marks)

The anonymized dataset from Olist Store, capturing a wealth of ecommerce transactions across Brazil's marketplaces from 2016 to 2018. Encompassing data from 100,000 orders, this rich dataset presents multifaceted views of order details, including statuses, prices, payments, and shipping. Additionally, customer locales are tied to geographic coordinates via a comprehensive geolocation dataset. For privacy, any company-specific references in customer reviews have been creatively substituted with fictional names from a renowned fantasy series.

1. Write an R script to perform an inner join between olist_orders_dataset and olist_order_items_dataset. What insights can you derive from the merged dataset about the order items for each order? **(4 Marks)**

2. Perform a left join between olist_orders_dataset and olist_order_reviews_dataset. After joining, explore how many orders do not have corresponding reviews. **(4 Marks)**

3. Execute a right join between olist_order_items_dataset and olist_products_dataset. Discuss any products in the olist_products_dataset that have not been sold yet. **(4 Marks)**

4. Combine olist_customers_dataset and olist_orders_dataset using a full outer join. Assess the resulting dataset for any customers without orders or orders without customer details. **(4 Marks)**
5. Use a semi join to filter olist_sellers_dataset for sellers who have made sales, based on the olist_order_items_dataset. Investigate the characteristics of active sellers versus the complete seller list. **(4 Marks)**

6. Identify customers from olist_customers_dataset who have never placed an order by using an anti join with olist_orders_dataset. Summarize the profile of these customers and hypothesize why they might not have placed an order. **(4 Marks)**

7. Use a combination of join types to merge olist_orders_dataset, olist_order_items_dataset, olist_products_dataset, and olist_sellers_dataset into one comprehensive dataset. Analyze the flow from sellers to products and items within orders, and visualize the distribution of order values. **(6 Marks)**

SECTION B: Visualization using ggplot2 (30 Marks)

The article **"customer sales analysis.pdf"** investigates the power of diverse data visualization techniques in uncovering latent trends and patterns within customer sales data, a key driver of strategic decision-making in modern businesses. By utilizing a comprehensive array of 24 distinct visualization methods, ranging from various chart types like Doughnut, Area, and Bar, to more complex representations such as Mind Maps and Flow Charts, the study aims to extract critical insights about consumer behavior. The paper demonstrates how these insights, when effectively harnessed through adept analysis, can significantly enhance business agility and propel informed decision-making, underpinning successful business outcomes.

Demonstrate only the R codes with comments for all the following questions. Do not have to show the charts (outputs).

1. Write the R code to generate the pie chart in ***Fig 3***, assuming you have a dataframe sales_data with columns Product_Line and count, where count represents the number of sales for each Product_Line. **(6 Marks)**

2. Write the R code to generate an area chart that visualizes sales_data as shown in ***Fig.10****: Area Chart of Sales Data*. Assume the dataset contains three variables: QUANTITYORDERED, PRICEEACH, and SALES. **(6 Marks)**

3. Write R code to create a line chart that displays the daily total sales over time as shown in ***Fig.16****: Daily Total Sales over Time- Line Chart*. Assume that you have two columns: Date and Total_Sales. The Date column is formatted as "YYYY-MM-DD" and Total_Sales contains the total sales for each day. **(6 Marks)**

4. Write R code to create a bar chart similar to the one displayed in ***Fig.17****: Distribution of Deal sizes- Bar Chart*. Assume you have a dataset, deal_sizes, with a categorical variable Deal_Size that classifies deals into 'Small', 'Medium', and 'Large', and a numeric variable Count that represents the frequency of each deal size. **(6 Marks)**

5. Write R code to generate a bar plot using R to represent the 'Number of Sales by Country' as shown in ***Fig.23****: Bar plot for Number of sales by country*. The dataset, consists of two columns: Country and Number_of_Sales. Each country's name is listed in Country, and the corresponding number of sales is in Number_of_Sales.
   Interpret the bar plot shown in ***Fig.23****: Bar plot for Number of sales by country*. **(6 Marks)**