ETW2001 FOUNDATIONS OF DATA ANALYSIS

SEMESTER 1 - 2024

Assessment 1 (30%)

This assessment is designed to evaluate your understanding and application of R programming skills in data manipulation and transformation to enhance data quality (unit learning outcome 1). The assessment is divided into two sections: Section A, which involves practical coding tasks, and Section B, in which you have to watch a business use case and write conditional statements.

Assessment Format

Section A: Practical Coding Tasks (30 Marks)

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

Section B: Business Application (30 Marks)

This section gives you a link to a YouTube video discussing a business case study. Answer the questions accordingly.

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 Manipulation and Transformation (30 Marks)

Retail businesses continually seek to maximize sales and reduce stockouts, which necessitates precise inventory management. In this case study, we examine a retail chain that utilized three interconnected datasets to optimize its inventory levels and sales strategies. Three datasets are provided as follows:

1. sales.csv

This dataset appears to record sales transactions from a retail chain. Each row represents a unique sale, identified by a SalesId. The StoreId and ProductId columns specify the store where the sale occurred and the product sold, respectively. The Date column indicates when the sale was made. UnitPrice reflects the price for a single unit of the product at the time of the sale, while Quantity shows the number of units sold in that transaction. The data includes a variety of products and spans several years, from at least 2017 to 2020, with product prices ranging from as low as $0.0525 to as high as $9.205, and quantities ranging from 8 to 98 units per transaction.

2. products.csv

The provided dataset is the list of products along with their supplier details and costs. Each product is uniquely identified by a ProductId and is associated with a ProductName and a Supplier. The ProductCost indicates the price of the product. The dataset covers a range of items from groceries such as ‘Chocolate Bar - Smarties’ and ‘Pepper - Red Bell’ to seafood like ‘Cod - Salted, Boneless’ and ‘Clam - Cherrystone’. It includes products from various suppliers including big box stores like ‘National Stores’, ‘Family Dollar’, ‘BJ's Wholesale Club’, and ‘Costco’, as well as other retail outlets like ‘Ocean State Job Lot’, ‘Fred's’, and ‘Gabe's’. The product costs vary, ranging from as little as $0.10 for an 8oz coffee cup to $5.76 for oranges. This dataset can be used to manage inventory, analyze supplier cost efficiency, or optimize pricing strategies.

3. inventory.csv

This dataset provides inventory details for various retail locations. Each entry corresponds to a specific ProductId and includes information on StoreId, StoreName, and the store's Address. The neighborhood column gives the local area where the store is located, which can be particularly useful for geographic data analysis and targeted marketing strategies. The QuantityAvailable column shows the number of units of the product that are currently in stock at each store. The data captures a range of quantities across multiple stores and locations, such as 'National Stores' in Bolton Hill with 11 items available, to 'Ocean State Job Lot' in Fells Point with just 1 item in stock. This information is crucial for supply chain management, inventory control, and ensuring the availability of products across the different branches of these retail chains.

1. Import all three datasets into RStudio. Filter the sales data only to include transactions that occurred in 2020. How many sales transactions were there in 2020? ***(3 Marks)***

2. Calculate the total revenue for each product in the sales data. Which ProductId has the highest total revenue? Filter the ProductName of that ProductId from products data and filter the StoreId of that product in inventory data. ***(5 Marks)***

3. Group the inventory data by StoreId and summarize the average QuantityAvailable across all products. Which StoreId has the lowest average quantity available? ***(5 Marks)***

4. Create a new column in the sales data that categorizes sales into 'High' (Quantity >= 50), 'Medium' (Quantity between 20 and 49), and 'Low' (Quantity < 20). What is the count of 'High' category sales? ***(5 Marks)***

5. Arrange the product information dataset in descending order of ProductCost. Then use tidyr to separate the ProductName into two columns: Product and Brand, where the Brand is the substring following the '-' character. What is the Brand of the third most expensive product? ***(5 Marks)***

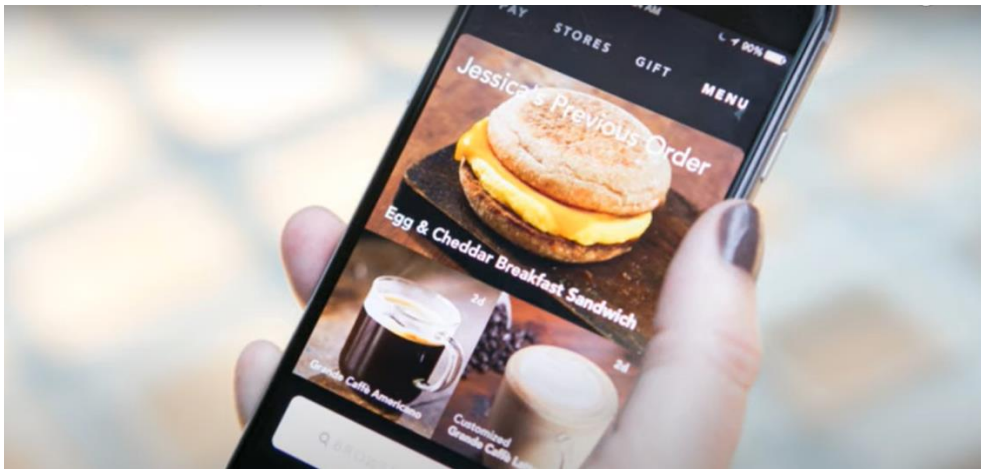
6. Using the dplyr package, analyze the price elasticity of demand for products in the retail chain. For this, you will need to create a metric that measures the percentage change in quantity sold (Quantity) in response to a percentage change in UnitPrice. To do this, first, calculate the average price and quantity sold for each product. Then, find the percentage changes between consecutive time periods. Discuss which products are most and least sensitive to price changes and how this could influence future pricing strategies. ***(7 Marks)***

SECTION B Business Application (30 Marks)

Starbucks' success has positively impacted coffee distributors across the United States by elevating the prestige of coffee as a whole. This uplift is not limited to Starbucks' products but extends to the 80% of coffee sold in supermarkets and beyond. Starbucks, a global giant with over 27,000 outlets and $22 billion in revenue last year, places a strong emphasis on data. The company's executive leadership team includes a head of Global Strategy, Insights, and Analytics, who employs a range of methodologies from ethnography to big data analytics. These approaches support various aspects of Starbucks' operations, including pricing strategy, real estate development, product development, trade promotion optimization, and marketing strategy.

Starbucks not only consumes vast amounts of coffee beans to satisfy its loyal customers but also leverages extensive data to enhance customer experiences and business operations. With 90 million transactions a week across 25,000 stores worldwide, Starbucks is at the forefront of utilizing big data and artificial intelligence for marketing, sales, and business decisions.

The introduction of Starbucks' rewards program and mobile app significantly increased the company's data collection capabilities, enabling them to gain deeper insights into customer behavior and purchasing patterns. The mobile app boasts over 17 million users, while the rewards program has 13 million active users. The data generated by these users, combined with other information such as weather, holidays, and special promotions, provides valuable insights for Starbucks. Let's explore some of the ways Starbucks utilizes the data it collects.



Link to video: https://youtu.be/Sq4-sAHIqJg?si=85j9Y_GVZPgs_DMV

1. Given that Starbucks operates ***X*** stores worldwide and produces ***Y*** million transactions per week, write a conditional statement in R to categorize the average number of transactions per store per week. Find out the values of ***X*** and ***Y*** from the video. Use the following categories, and assign your own variable names:
    - High: More than 3,500 transactions per store per week
    - Medium: Between 2,500 and 3,500 transactions per store per week
    - Low: Less than 2,500 transactions per store per week

Calculate the average number of weekly transactions per store and then write a conditional statement to assign the appropriate category. Print out the category along with a message indicating the transaction level for Starbucks stores. **(10 Marks)**

2. Starbucks company used five potential variables to predict the locations for opening a new Starbucks store. The five variables characterize each location.

Determine the variables from the video and Write a conditional statement in R to select the best location for opening a new store. You must include five variables, you can assign any values or ranges for all the 5 variables. Categorize the locations into:

- Ideal location
- Good location
- Average location
- Poor location

Print out the classification along with a message indicating the suitability of the location for opening a new Starbucks store **(10 Marks).**

3. In mid-2018, Starbucks used local factors to promote specific products. Identify the **two** local factors mentioned in the video. Based on these factors, write a conditional statement in R to promote a specific Starbucks product. Categorize the promotions as follows, you can provide your own conditions for the two factors:
    - Promote hot beverages
    - Promote cold beverages
    - Promote pastries
    - No promotion

Print out the product to be promoted along with a message indicating the promotion based on the factors **(10 Marks)**.

-End of Questions-