Assignment 1

Maximum points: 20 Due date: 8th Nov. 2024 at 23:00

Guidelines

- Individual Assignment: This is an individual assignment. Please do not seek help from others or collaborate with classmates.
- Problem Understanding: Carefully read the problem description and each question before starting your analysis.
- Choice of Software: You are free to use either R or Python to perform your analysis.
- Code Requirements: Your code should be well-structured and free of errors. Please use clear and descriptive variable names.
- Deliverables: Submit the following two files on Canvas:
    1. Code file (.R/.Rmd/.py/.ipynb): This file should contain the complete executable code for the analysis.
    2. Report file (.pdf): This document should contain your answers to all the questions asked below. Please use the solution template provided on Canvas (Assignment_1_Template.docx) to answer the questions. On the cover page, please write your name, student number, and date. After writing your answers, submit the document as a pdf file.
- Plagiarism Warning: Ensure that all the code and analysis are your original work. Plagiarism will lead to disqualification and academic consequences.
- Late Submissions: Late submissions will be accepted, but penalized.
- Points Distribution: The distribution of points and grading rubric for the questions is provided at the end of this document.

Business Context

Introduction: **Orange Inc.** is a leading technology firm specializing in innovative electronic products such as **oPhone**, **oWatch**, **oPods**, and **oTV**. The company has established a significant customer base and uses customer data to understand and enhance their business strategies. With a mission to enhance customers' digital lives, Orange uses technology not only to develop products but also to gather insights on customer preferences, product adoption, and transaction behaviors.

Customer data story: When customers purchase any product from Orange Inc., their data is collected and managed by different departments for various business needs. Each time a customer buys a product, their demographic information and product ownership are recorded, and the transaction details are stored in specialized databases. There are three key databases where this information is stored:

- Customer information database: The Marketing and Customer Relations Department manages a database containing demographic details and loyalty tiers of each customer. This data is collected from customer profiles created during the registration or checkout process. The team records details like the age bracket, gender, and loyalty tier of each customer to better understand the audience and target their marketing campaigns.

- Product ownership database: The Product Management Department maintains a database that records which customers own which products. Whenever a customer purchases an oPhone, oWatch, oPods, or oTV, this information is logged under their unique customer ID. This allows the Product Management team to analyze ownership patterns and identify opportunities for cross-selling and product upgrades.

- Transaction Database: The Finance Department manages a database that records all financial transactions, including the type of card used and the date of the transaction. This information is collected at the point of sale and is vital for understanding purchasing trends, preferred payment methods, and seasonal spikes in sales.

Due to the decentralized management of data, the databases were maintained separately by each department. Recently, Orange Inc. merged the three databases to create an integrated dataset (orange_inc.csv). As a data analyst at Orange Inc., your role is to clean, combine, and analyze the integrated dataset.

Data dictionary

Please access the file orange_inc.csv from Canvas. The dataset uses "-" to indicate missing cells.

Column	Data type	Description
user_id	integer	Unique identifier for each customer (ranging from 1 to 20,000)
age_bracket	categorical	Age range of the customer. Possible values: 18-25, 25-34, 34-50, 50+
tier	categorical	Loyalty tier of the customer. Possible values: silver, gold, platinum
gender	categorical	Gender of the customer. Possible values: male, female, other
oPhone	binary	Indicates whether the customer owns an oPhone. Possible values: 0 (No), 1 (Yes).
oWatch	binary	Indicates whether the customer owns an oWatch. Possible values: 0 (No), 1 (Yes).
oPods	binary	Indicates whether the customer owns an oPods. Possible values: 0 (No), 1 (Yes).
oTV	binary	Indicates whether the customer owns an oTV. Possible values: 0 (No), 1 (Yes).
date	date	The date of the transaction in yyyy-mm-dd format. Period: 1 January 2023 to 31 December 2023.
card	categorical	Type of card used for the transaction. Possible values: Visa_credit, Visa_debit, Mastercard_credit, Mastercard_debit, Other_credit, Other_debit.

Note: For simplicity, we will assume that if a customer has purchased multiple products, then all those products were purchased on the same date (as specified in the column date).

Questions

1. Import the data from the file orange_inc.csv and name it as orange_data. Report the number of missing values for each attribute in tabular form.

2. Remove rows from orange_data that have at least one missing value in any of the columns. After removing the rows, rename the dataset as orange_data_clean. Report the total number of observations (rows) in orange_data_clean.

3. Create a new column named owns_multiple_products in orange_data_clean. This column should indicate whether a customer owns more than one product. If a customer owns two or more than two products, then the value of owns_multiple_products for that customer should be 1, otherwise it should be 0. How many customers own multiple products from Orange Inc.?

4. How many female customers in the age group of 18-25 have purchased products using either a Visa credit card or a Mastercard credit card?

5. What percentage of customers are older than 25, have purchased products with a debit card, and own multiple (≥ 2) products?

6. How many customers own all the four products?

7. Calculate the average number of products owned by customers in each loyalty tier and report it in a table.

8. The price of the products are as follows:

Product	Price
oPhone	€1200
oWatch	€300
oPods	€150
oTV	€2500

   Store the product prices in a vector by the name product_prices and do the following:

   a) Create column named sales_revenue and compute the revenue from each customer by using the information on the products purchased by that customer and the corresponding prices stored in the vector product_prices. What is the total sales revenue from all the customers?
   b) Create a column named month_of_purchase and extract the month from the date column. Which month had the highest sales?

c) Compute the total sales revenue from each product and report it in a tabular format.

9. The two sub-questions are related to visualization. Please make sure your charts are appropriately annotated and clearly readable.
   a) Create a bar chart to visualize the revenue by month. The bar chart should have 12 months on the horizontal axis and the corresponding revenue on the vertical axis. (It is ok to use month no. on the horizontal axis instead of month name).
   b) Create a pie chart to visualize each product's share of revenue.

10. Create a filtered version of orange_data_clean that only includes customers who own an oPhone and belong to the age bracket 18-25. Name this new dataset as young_oPhone_owners and based on this dataset, answer the following questions:

    a) Identify the most preferred card type among young_oPhone_owners.
    b) Determine the average number of products purchased by young_oPhone_owners (including oPhone).

Points distribution and grading rubric

Question	Point(s)	Grading criteria		
		Correct answer	Incorrect answer, but coding logic partially correct	Incorrect answer and coding logic
1	1	1	0.5	0
2	1	1	0.5	0
3	1	1	0.5	0
4	1	1	0.5	0
5	1	1	0.5	0
6	1	1	0.5	0
7	2	2	1	0
8a	2	2	1	0
8b	2	2	1	0
8c	2	2	1	0
9a	2	2	1	0
9b	2	2	1	0
10a	1	1	0.5	0
10b	1	1	0.5	0