Data-Driven Decision Making Final Project

Overview:

In this final project, you will perform a fully-fledged data analytics project. The goal is to apply data-driven decision-making techniques, focusing on descriptive analytics, predictive modelling, and business interpretation. You will submit a short report (not exceeding 5000 words) and present your findings in a 15-minute presentation.

Datasets:

You have two options for choosing a dataset:

1. (provided) Marketing campaigns
2. A dataset of your choice

Option 1: Marketing Campaigns Analysis

Instructions:

- Marketing data in csv (comma-separated-values) format http://lms.vnuk.edu.vn/courses/1401/files/94757?module_item_id=41785
- Template Jupyter notebook for Google Colab http://lms.vnuk.edu.vn/courses/1401/files/94758?module_item_id=41786

Part 1: Exploratory Data Analysis

Dataset Overview: The dataset provided includes customer data collected by a marketing analytics team. The main goal is to analyze the behavior of customers based on their demographic and transactional data to derive insights that can improve marketing strategies. The dataset contains columns describing customer demographics (e.g., Income, Education, Marital), transactional data (e.g., MntWines, NumCatalogPurchases), and their responses to marketing campaigns (e.g., AcceptedCmp1, AcceptedCmp2). The Response column serves as a target variable, indicating whether the customer accepted the offer in the last campaign.

Key Columns:

- **AcceptedCmp1-5:** whether the customer accepted the offer in each of the five campaigns.
- **Response:** whether the customer accepted the offer in the last campaign (binary).
- **Income:** Customer's yearly household income.
- **Kidhome, Teenhome:** Number of small children and teenagers in the household.
- **MntWines, MntMeatProducts, etc.:** Amount spent on various product categories in the last 2 years.
- **NumWebPurchases, NumStorePurchases, etc.:** Number of purchases across various channels.
- **NumWebVisitsMonth:** Number of visits to the company's website in the last month.
- **Recency:** Number of days since the last purchase.

Questions:

Perform an exploratory data analysis on the data, such as visualizing the distribution of values in selected columns, or differences in the distribution of a column across sub-groups of customers.

Below are some suggested questions that you can try answering. Note that you don't have to answer all of them, please aim to answer 5 questions in your report, which may or may not be in the list below. Focus on the quality and depth of your answers, not just quantity.

a. **Demographic Analysis:**
    - What is the distribution of customer income? Are there any noticeable patterns or outliers?
    - How are customers distributed across different education levels and marital statuses?

b. **Spending Behavior:**
- Which product category (e.g., wines, meats, sweets) has the highest and lowest average spending?
- What is the total spending across all categories (MntTotal) for customers with varying income levels?

c. **Campaign Effectiveness:**
- What percentage of customers accepted offers in each campaign (AcceptedCmp1-5)? Is there a trend in acceptance rates across campaigns?
- Are customers who accepted previous campaigns (AcceptedCmp1-5) more likely to accept the most recent campaign (Response)?

d. **Channel Preferences:**
- Which purchase channel (e.g., web, catalog, store) is the most frequently used across customers?
- How does the number of website visits (NumWebVisitsMonth) relate to the number of website purchases (NumWebPurchases)?

e. **Customer Segments:**
- How do family dynamics (e.g., Kidhome, Teenhome) affect customer spending on specific product categories like wines or sweets?
- Are customers who have complained in the last two years (Complain) spending significantly more or less than those who haven't?

f. **Recency and Response:**
- Is there a relationship between the recency of a customer's last purchase (Recency) and their likelihood of accepting the latest campaign (Response)?

Part 2: Linear Regression Analysis (supervised learning)

The same notebook from question 3 includes a template for linear regression analysis. The intention of this section is to check if there are statistically significant linear relationship between variables in the dataset.

Below are some questions that you can try answering. Note that you don't have to answer all of them, please aim to answer 3 questions in your report (besides the analysis and visualization extracted from the notebook), which may or may not be in the list below. Focus on the quality and depth of your answers, not just quantity.

- **Income vs. Spending:** how does a customer's yearly income predict their spending?
- **Website Visits vs. Website Purchases:** can the number of website visits in the last month predict the number of website purchases?
- **Family Size vs. Spending:** does the number of children and teenagers in a household predict spending on certain products?
- **Discount Purchases vs. Total Purchases:** does the number of discount purchases predict the total number of purchases made in certain, or all channels?
- **Recency vs. Campaign Response:** does the number of days since a customer's last purchase influence their likelihood of accepting the latest campaign offer?

Part 3: Customer segmentation (Unsupervised learning)

The same notebook from question 3 includes a template for customer segmentation. The segmentation (clustering) is performed based on RFM (recency, frequency, monetary) aspects of the customers. The intention of this section is to identify meaningful clusters of users.

Below are some questions that you can try answering. Note that you don't have to answer all of them, please aim to answer at least the first 2 questions in your report. Focus on the quality and depth of your answers, not just quantity.

- **Retention Strategies:** which clusters in your analysis should be targeted with retention strategies, and how?
- **Promotional Targeting:** which clusters can be targeted with promotional campaigns, and how?
- **High-Value Customers:** who are the high-value customers and what are their spending patterns?
- **Channel Preferences:** how do purchase behaviors differ across clusters? Can specific channels be optimized for certain customer groups?
- **Campaign Effectiveness:** how does campaign acceptance vary across clusters? Are high-value clusters more likely to respond positively to campaigns?

Part 4: Format and Writing Style

- Please compile all your answer into a single report in pdf format
- You can include your code (SQL, Python or Excel) describing how the analysis was performed in a separate file if you don't want to clutter your report
- Your report should include relevant visualization (e.g. graphs) to support your interpretation/recommendation
- You are encouraged to use AI assistance to improve your writing. However, make sure to use AI wisely. We do not want (and will penalize) overly verbose paragraphs with little meaning.

Option 2: Your Own Dataset

You would **need to discuss with us** before choosing this option as the requirements for your analysis may be different from option 1.

Here are some possible options:

- Hotel booking dataset (used in mini-project): you can use the Jupyter notebook template in option 1 to help with your analysis
- WiDS++ 2025 dataset (you can only opt for this dataset if you are already taking Advanced Business Analytics class) https://www.widsworldwide.org/learn/datathon/datathon-university-edition/
- A business-related dataset of your choice

Format Guidelines

Report Structure:

Your report should be structured as follows:

- **Introduction**: briefly introduce the purpose of the project and the dataset.
- **Descriptive Analytics**: summarize key statistics and interesting patterns
- **Predictive Modeling:** use either a supervised and/or unsupervised learning method to perform predictive modelling on variables in the data.
- **Recommendations & Insights**: provide actionable recommendations or insights from your analysis

Presentation Guidelines:

- **Time**: 15 minutes (10 minutes for presentation + 5 minutes for Q&A).
- **Content**:
  - Provide a concise summary of your analysis, focusing on the **most critical and interesting** findings.
  - Use visual aids (charts, graphs, tables) to explain your insights clearly.
  - Conclude with a discussion of your business recommendations or insights
- **Format**: You may use slides (e.g., PowerPoint) or a live demo of your analysis in a tool like Jupyter Notebook.

Submission:

- **Report**: write up your report into a pdf file. Make sure you still do all the proper formatting to make the notebook presentable.
- **Code:** attach your Jupyter notebook or other relevant code files together with your report
- **Presentation**: Be prepared to present your findings in class.

Evaluation Criteria:

- **Depth of Analysis** (40%)
- **Interpretation of Results** (30%)
- **Presentation Clarity and Professionalism** (20%)
- **Creativity and Advanced Insights** (10%)

Peer evaluation

As an outcome of a group assignment, your work must include a peer evaluation. Clearly state how each member contributed to the group work and the percentage of the total work each member should receive.

This evaluation should be sent **individually and confidentially via email to the instructors** at gia.ngo@vnuk.udn.vn and nga.nguyen@vnuk.udn.vn

Team Member	**Contribution Description**	**% Contribution (Total 100%)**
Member 1	Wrote part 1, collected data, edited the paper	25%
Member 2	Wrote part 2, collected data	20%
Member 3	Wrote part 3, collected data	20%
…	…	…