AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Science and Technology

Project Report

Assignment Title:	Mid Project Report			
Assignment No:	01		Date of Submission:	27 April 2025
Course Title:	Introduction to Data Science			
Course Code:	CSC4180		Section:	G
Semester:	Spring	2024-25	Course Teacher:	Dr. Ashraf Uddin

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaboration has been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the faculty for review and comparison, including review by external examiners.
7. I/we understand that Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of their arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.:	02

No	Name	ID	Program	Signature
1	MD. TAMJID HOSSAIN	22-46460-1	BSc [CSE]	
2	WASIF ASAD ALVI	22-46451-1	BSc [CSE]	
3	RIFAT TALUKDAR	22-46428-1	BSc [CSE]	
4	MD. TANZIUL HAQUE	22-46435-1	BSc [CSE]	
5			Choose an item.	
6			Choose an item.	
7			Choose an item.	
8			Choose an item.	
9			Choose an item.	
10			Choose an item.	

Faculty use only		
FACULTY COMMENTS	**Marks Obtained**	
	Total Marks	

Project Report

1. Data Creation

The dataset used for this project was purposedly created with meaningful relationships between each column. For example, since it's a dataset for online shopping behaviour of buyers so it contains columns/features like time spent on website, number of products viewed during that spent time, previous purchases etc with a target column of buy possibility. We've inserted relative values in each column so if a buyer has spent more time on the website they likely have more number of products viewed, and this also means the buy possibility of those buyers is high and vice versa.
The dataset was at first uploaded to a Google Drive(Link: https://drive.google.com/file/d/1xP-Q1AS4hgZT-lMtSpue--jl-RMMpITy/view?usp=drive_link). Then it was loaded into R using read.csv() after converting the shared link to a downloadable format.

The dataset contains the following types of features/columns:

- Numerical Variables:
  - Time.Spent.on.Website
  - Number.of.Products.Viewed
  - Previous.Purchases
- Categorical Variables:
  - Device.Type
  - Added.to.Cart
  - Item.Category
  - Buy.Possibility

The target column is Buy Possibility. The target column has four values or categories. They are: "high", "medium", "low", and "not likely".

Each row in the dataset represents an individual buyer's interaction with a online shop/website. The aim of this dataset is to understand consumer behavior based on their browsing habits and past purchases and to predict buy possibility of an individual buyer.

2. Data Preprocessing Steps

A. Handling Missing Values:

Missing values were initially represented as "N/A".
We took the steps below to handle the missing values:

- Rows with more than 2 missing values were removed to preserve data quality.
- For numeric columns, missing values were imputed using the median.
- For categorical columns, missing values were imputed with the mode.

B. Data Type Conversion

We explicitly converted certain columns to factor type (categorical) and ensured numerical columns were properly formatted as numeric.

C. Data Transformation:

Min-Max Scaling was applied to the Time.Spent.on.Website column to normalize it into a 0–1 range.

D. Outlier Handling:

We took the steps below to detect and handle outlier values:

- Outliers were detected using the Interquartile Range (IQR) method.
- Instead of deleting entire rows, outlier values were replaced by the median of non-outlier values.

3. Key Findings from Analysis

A. Summary Statistics

After preprocessing, the dataset showed a balanced distribution across categories and numeric variables.

B. Univariate Analysis

Histograms and density plots revealed skewed distributions and the presence of outliers.

C. Categorical Data Exploration

Most users accessed the website via mobile and desktop devices. The most popular item categories were accessories and laptops.

D. Multivariate Exploration

There was a mild positive correlation between Time Spent on Website and Number of Products Viewed (~0.21). Pair plots revealed distinct but overlapping groupings based on buying possibility.

4. Justification for Transformations and Outlier Handling

Median imputation: Median provides a more accurate central tendency when extreme values exist. In our dataset, the range is a bit wide (and skewed), so median will be more useful to outliers. Unlike mean, which gets pulled by extreme values, median stays stable.

Min-Max scaling: Essential to normalize feature ranges, important for visualizations and algorithms.

Explicit data typing: Explicit data typing was used to clearly define which variables are numeric and which are categorical. This ensures that each column is treated appropriately in later analyses, such as summary statistics, visualizations, and modeling. By explicitly setting data types, we avoid potential errors and guarantee accurate interpretation of the dataset.

Conclusion

In this project, we successfully cleaned, transformed, and explored the dataset to prepare it for analysis. We handled missing values, treated outliers by replacing them with the median, and applied Min-Max scaling to normalize key numeric features. Through visualizations and statistical summaries, we gained valuable insights into customer behavior.