A logo for a university

Description automatically generated **East West University**

**Assignment 2**

**Semester:** Fall-2024

**Course Title:** Statistics for Data Science

**Course Code:** CSE303 **Sec: 07**

**Submitted by-**

Sheikh Sarafat Hossain

2022-3-60-109

**Submitted to-**

Puja Chakraborty

Lecturer

Department of Computer Science & Engineering

East West University

**Date of Submission: 01 January 2025**

**Github Link:** <https://github.com/shiekhsarafathossain/CSE303_PROJECT.git>

* **Project Title: Think of project title for your group. The project title should be the same for each group member.**

**Answer:**

“Data Analysis of Bike Sales: Exploring Consumer Behavior, Drivers and Insights”

* **What are the types of data? (e.g. Qualitative vs Quantitative; Nominal vs Ordinal; Discrete vs Continuous)**

**Answer:**

There are two types of data.

1. **Qualitative Data:** Qualitative data are often categorized or labeled.

Examples: Colors, gender etc.

There are 2 types of qualitative data:

* Nominal: Categorical data without order.

Example: Blood groups.

* Ordinal: Categorical data with a meaningful order or ranking.

Example: customer satisfaction ratings (satisfied, neutral, dissatisfied).

1. **Quantitative Data:** Quantitative data is numerical data representing measurements or counts. It can be subjected to arithmetic operations.

Examples: Height, age, income etc.

There are 2 types of quantitative data:

* + Discrete: Countable and finite values, often integers.

Example: Number of students in a class, number of cars etc.

* + Continuous: Measurable values that can take any value within a range, including decimals.

Example: Temperature, height, weight

* **What is balanced or imbalanced dataset? How does it might affect the performance of a model?**

**Answer:**

**Balanced vs. Imbalanced Datasets:**

* + **Balanced Dataset:** In a balanced dataset, all classes have similar sample sizes. Models perform well across all classes.
  + **Imbalanced Dataset:** In an imbalanced dataset, some classes have far fewer samples, causing models to focus on the majority class, ignoring the minority class.

**Affect:**

It leads to biased predictions on the majority classes. poor performance in minority classes. Accuracy can be misleading.

A balanced dataset is always better because there is no bias. Fixing an imbalance dataset ensures fairer and more accurate models.

* **Write a short description of your task and dataset, such as what are the column in your dataset, their type, what are the range or categories of the values in each column etc.**

**Answer:**

The task is to analyze a dataset to find patterns, relationships, and insights that help with classification, prediction, or grouping.

The dataset is designed for analysis and includes the following columns:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Range / Categories** |
| |  | | --- | | ID | | Nominal | |  | | --- | | Unique identifiers | |
| Age | Discrete | 0–100 |
| Gender | Nominal | Male, Female |
| Income | Continuous | 10,000–1,000,000 USD |
| Education Level | Ordinal | |  | | --- | |  |   High School, Partial High School, Partial College, Bachelors, Graduate Degree |
| Marital Status | Nominal | Single, Married |
| Purchase status | Nominal | Yes, no |
| Children | Discrete | 0-5 |
| Occupation | Nominal | Clerical, Professional etc. |
| Home Owner | Nominal | Yes, no |
| Cars | Discrete | 0-2 |
| Age brackets | Ordinal | Middle Age, Old etc. |
| Commute Distance | Ordinal | * 1. miles, 2-5 miles etc. |

* **Mention if your dataset is balanced or not with quantitative result?**

**Answer:**

To determine if the dataset is balanced, we must check the frequency of each class

in the target column ("Purchased Bike").

**Quantitative result**: "Purchased Bike" has:

* 70% "Yes"
* 30% "No"

**This indicates the dataset is imbalanced.**

* **If any column contains continuous numerical values then calculate the mean, median, variance and standard deviation of that column, otherwise if the values are categorical or discrete then count the frequency and percentage of each type of values.**

**Answer:**

For columns with continuous numerical values in our dataset such as, Income and Age, we can calculate the following:

**Income**:

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Age:**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Frequency and Percentage for Categorical or Discrete Columns** (Gender, Marital Status, or Education)

**A screenshot of a computer program

Description automatically generated**

A screenshot of a computer program

Description automatically generated

* **Plot bar chart, line graph or pie chart where necessary.**

**Answer:**

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

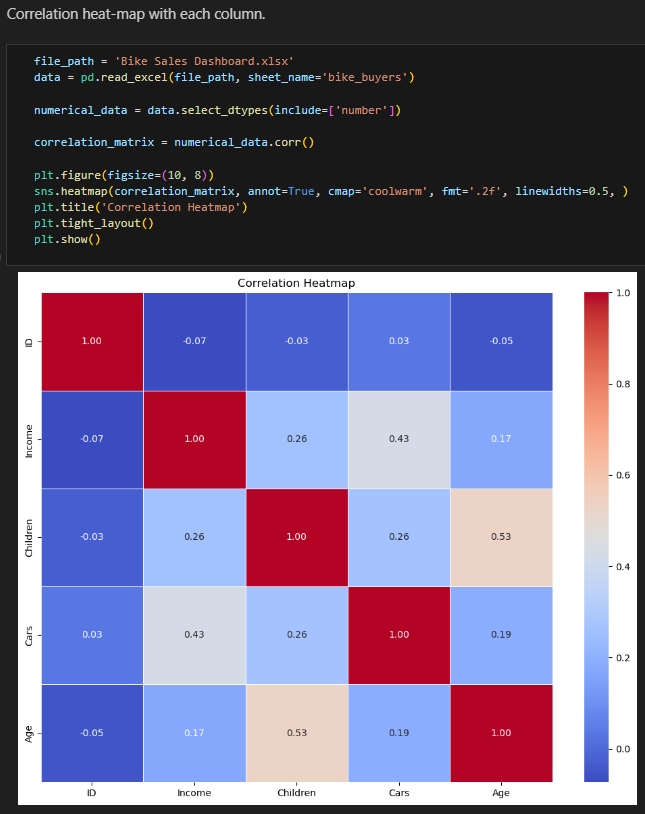
**A screen shot of a graph

Description automatically generated**

* **Create a correlation heat-map with each column.**

**Answer:**

A correlation heatmap can help visualize the relationships between numerical columns (Income and Age). The closer the value is to 1 or -1, the stronger the relationship.



* **How to convert a categorical data to numerical values to get features for machine learning projects?**

**Answer:**

To turn categorical data into numbers for machine learning, we can use these methods:

Label Encoding gives each category a number (good for ordered data). One-Hot Encoding makes a new column for each category with 0s and 1s (best for non-ordered data). Ordinal Encoding uses numbers for categories in a specific order (like Low = 0, Medium = 1, High = 2); Frequency Encoding replaces categories with how often they appear; Target Encoding uses the average of the target value for each category; Binary Encoding changes categories into binary (e.g., A = 01, B = 10); and Hashing Encoding creates a fixed-size number for each category using a hash function. Pick the method based on your data type and model needs.

* **Also mention if there are any missing values in any column of not? How do you want to handle those missing values and why?**

**Answer:**

If there are missing values in any column, here’s how to handle them:

**A screenshot of a computer program

Description automatically generated**

* **Write a conclusion paragraph stating your overall observation.**

**Answer:**

In conclusion, the dataset provides valuable insights into the factors influencing bike purchases. We analyzed patterns in both categorical and numerical data and implemented necessary adjustments to prepare the dataset for machine learning tasks. It is worth noting that the dataset might be imbalanced, particularly in the "Purchased Bike" column, which could require more balanced representation for improved outcomes. Missing values were addressed meticulously to ensure the reliability of the analysis. Overall, this study highlights significant trends in consumer demographics and purchasing behavior, which can inform marketing strategies or product enhancements.