**Assignment 02**

1. **Project Title:**

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

1. **Types of Data:**
2. **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.

Data

Qualitative Data

Quantitative Data

Continuous Data

Discrete Data

Ordinal Data

Nominal Data

1. **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 on 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.

1. **Task and Dataset Description**

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. |

**5) Dataset Balance**

To determine if the dataset is balanced, we have to check the frequency of each class in the target column (e.g., "Purchased Bike").

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

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

This indicates the dataset is imbalanced.

**6) Descriptive Statistics for Continuous Columns**

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

A black screen with white text

Description automatically generated

A screenshot of a computer

Description automatically generated

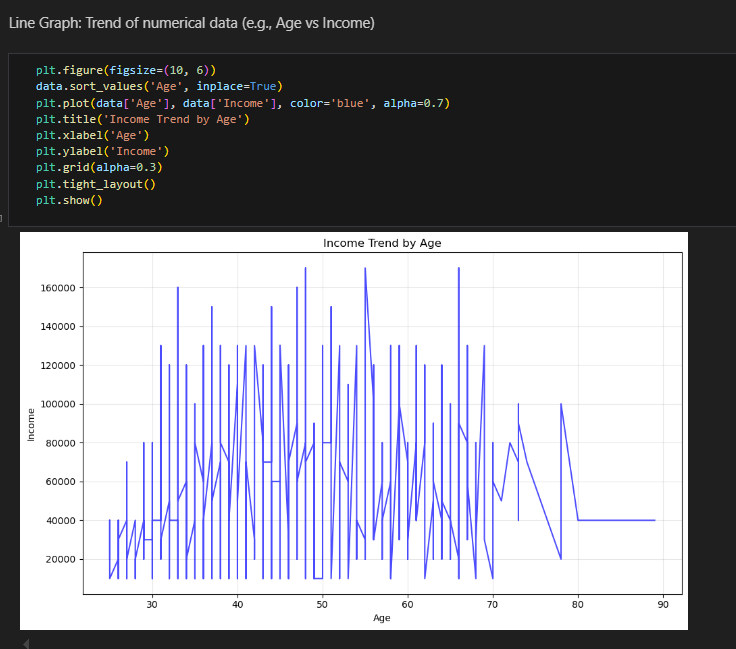
A screenshot of a computer

Description automatically generated

**7) Bar chart, Pie Chart, line graph**

A screenshot of a computer

Description automatically generated

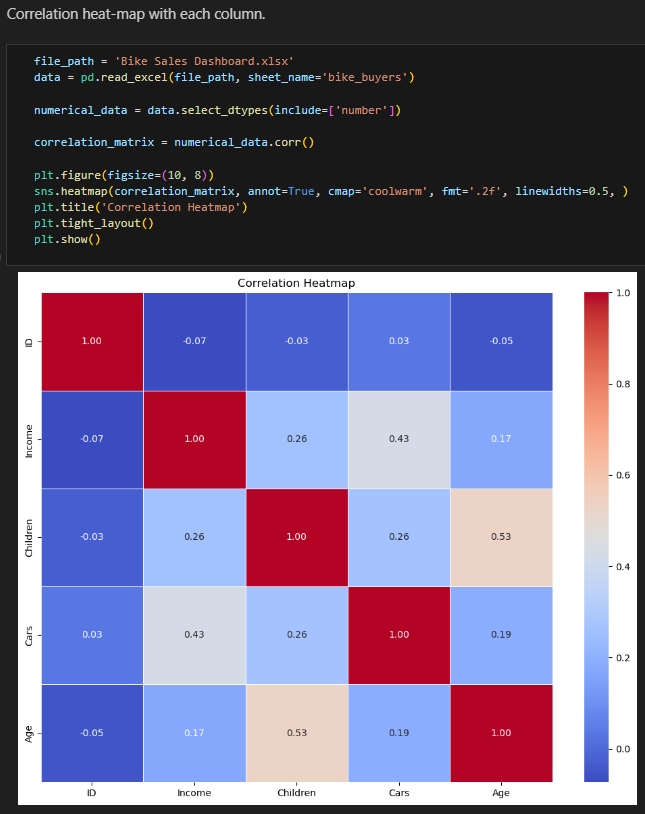


A screenshot of a computer screen

Description automatically generated

**8) Correlation Heatmap**

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.

****

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

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.

**10) Handling Missing Values**

If there are missing values in any column, here’s how to handle them: A screenshot of a computer program

Description automatically generated

**11) Conclusion**

In conclusion, the dataset represents useful information about factors that affect bike purchases. We studied the patterns in both categorical and numerical data and made the changes needed for machine learning tasks. We can say that , the dataset may be unbalanced, especially in the "Purchased Bike" column, which might need more equal representation for better results. Missing values were handled carefully to make the analysis reliable. Overall, this analysis shows important trends in people's demographics and buying habits, which can help with marketing or improving products.