Vishwakarma Institute of Information Technology, Pune

**(An Autonomous Institute Maharashtra)**

**A Report**

**on**



**CSE (AI) Department**

**Vishwakarma Institute of Information Technology**

**Academic Year: 2023-24**

**“Assignment 1: Data Manipulation and Exploration”**

**Submitted by**

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Name** | **Email** |
| **1** | **Omkar Bhaskar Rode** | omkar.22211052@viit.ac.in |

**Under Guidance of**

Dr. Anuradha Yenkikar

Problem Statement: -

**Perform the following operations using R/Python on suitable data sets:**

a) read data from different formats (like csv, xls)

b) indexing and selecting data, sort data,

c) describe attributes of data, checking data types of each column,

d) counting unique values of data, format of each column, converting variable data type

(e.g. from long to short, vice versa),

e) identifying missing values and filling in the missing values.

Packages / Libraries used: -

* **NumPy**: Utilized for numerical computations and data manipulation tasks.
* **Matplotlib**: Employed for basic data visualization such as scatter plots, histograms, and bar plots.
* **Pandas**: Primarily used for data manipulation and analysis, including reading data from CSV and Excel files, indexing, selecting, sorting, describing attributes, checking data types, counting unique values, formatting columns, converting data types, and handling missing values.
* **Seaborn**: Utilized for advanced data visualization, complementing Matplotlib with additional statistical graphics and enhancing the visual appeal of plots.

Theory: -

The primary objective of this project is to perform data manipulation and exploration on a dataset containing information about customers' purchases from an e-commerce platform. Data manipulation involves tasks such as reading data from CSV and Excel files, indexing, selecting, sorting, and formatting columns using Python's panda’s library. Exploratory data analysis aims to gain insights into the dataset's characteristics by describing attributes, checking data types, counting unique values, and handling missing values. By leveraging these techniques, the project aims to prepare the dataset for further analysis and modeling, ultimately facilitating data-driven decision-making processes for the e-commerce platform.

Methodology: -

* **Data Loading**: The first step involved loading the dataset using the panda’s library. The **read\_csv** function was utilized to read the data from the CSV file into a DataFrame named **heart**.
* **Initial Exploration**: Initial exploration of the dataset was conducted to understand its structure and contents. The **head** function was used to display the first few rows of the dataset, providing an overview of the data's format and columns. Additionally, the **describe** function was employed to generate summary statistics for numerical columns, offering insights into the distribution and range of values.
* **Column Identification**: The columns present in the dataset were identified using the **columns** attribute, which provided a list of all column names.
* **Data Types Analysis**: To analyze the data types of each column, a small subset of the dataset was loaded using the **read\_csv** function with the **nrows** parameter set to 2. The data types of each column were then examined using a loop over the **dtypes** attribute of the DataFrame
* **Unique Values Analysis**: The unique values present in selected columns of interest (**ChestPain** in this case) were identified using the **unique** function. This step helped in understanding the categorical variables and the range of values they encompass.
* **Missing Values Detection**: Missing values within the dataset were identified using the **isnull** function, which returned a DataFrame indicating True for missing values and False for non-missing values. This step was crucial for assessing the completeness of the dataset and determining the need for handling missing values in subsequent analysis.
* **Data Selection**: A subset of the dataset containing specific columns (**Age**, **Sex**, **ChestPain**) was selected using the **loc** function. This allowed for focusing on relevant variables for further analysis.
* **Data Sorting**: The dataset was sorted based on multiple columns (**Age**, **Sex**, **Chol**) in ascending order using the **sort\_values** function. Sorting the data enabled better organization and facilitated subsequent analysis tasks.

Application: -

* **Healthcare Analytics**: The dataset used in this assignment contains information related to heart health, including factors such as age, sex, chest pain type, and cholesterol levels. Analysis of this data can be valuable for healthcare providers and researchers in identifying patterns and risk factors associated with heart disease. Insights gained from this analysis can inform preventive measures, treatment strategies, and public health initiatives aimed at reducing the prevalence of heart-related illnesses.
* **Customer Segmentation**: Understanding customer behaviour and preferences is essential for businesses, particularly in the e-commerce sector. The techniques demonstrated in this assignment, such as data manipulation, exploration, and analysis, can be applied to customer transaction data to segment customers based on demographic characteristics, purchase history, and other relevant factors. These customer segments can then be used to tailor marketing strategies, personalize product recommendations, and improve overall customer satisfaction and retention.
* **Predictive Modelling**: The dataset used in this assignment can serve as a basis for building predictive models to forecast outcomes related to heart health, such as the likelihood of developing cardiovascular diseases or experiencing certain symptoms. By leveraging machine learning algorithms and predictive analytics techniques, healthcare professionals can develop risk assessment tools and decision support systems to assist in clinical diagnosis and treatment planning.
* **Public Health Interventions**: Insights derived from the analysis of heart health data can also inform public health interventions aimed at promoting healthy lifestyles and reducing the burden of cardiovascular diseases at the population level. By identifying high-risk groups and vulnerable populations, policymakers and public health authorities can design targeted interventions, educational campaigns, and preventive measures to improve cardiovascular health outcomes and mitigate disparities in healthcare access and outcomes.
* **Data-Driven Decision Making**: The techniques demonstrated in this assignment highlight the importance of data-driven decision-making in various domains, including healthcare, business, and public policy. By leveraging data analysis and visualization tools, stakeholders can gain actionable insights from complex datasets, enabling informed decision-making, evidence-based policy formulation, and strategic planning to address pressing challenges and achieve organizational objectives.

Conclusion: -

In conclusion, the techniques and methodologies explored in this project offer valuable insights into the process of data manipulation and analysis using Python and relevant libraries. By leveraging tools such as Pandas, NumPy, Matplotlib, and Seaborn, we were able to effectively explore, understand, and visualize the dataset. These skills are essential for data scientists, analysts, and researchers across various industries, enabling them to extract actionable insights, make informed decisions, and drive impactful outcomes from complex datasets. Overall, this project underscores the importance of data-driven approaches in addressing real-world challenges and advancing knowledge in diverse domains.