|  |  |
| --- | --- |
| Subject: Data Analytics and Visualization Lab | Course ID: CSL-601 |
| Semester: VI | Course: AI & DS |
| Laboratory: 407 | Name of teacher: Prof. Gitanjali Korgaonkar |
| Name of Student: Ayush Gupta | Roll No: VU2S2324001 |

**EXPERIMENT NO. 2**

**Aim:**

Data Exploration: Knowing the data, Data Preparation, and Cleaning

**Theory:**

Exploring data is a crucial step in understanding its characteristics, trends, and underlying patterns. Conducting experiments in data exploration involves various techniques and tools to gain insights into the dataset. The key approaches to data exploration include:

**1. Data Collection and Understanding**

* Collect the dataset from various sources such as CSV files, databases, or APIs.
* Understand data sources, variables, and their meanings by referring to data dictionaries or metadata.

**2. Data Cleaning and Preprocessing**

* Identify missing values, duplicates, and outliers in the dataset.
* Handle missing values through imputation or deletion based on data relevance.
* Normalize or standardize data if required for certain analyses or algorithms.

**3. Statistical Summaries and Visualizations**

* Compute descriptive statistics like mean, median, mode, and standard deviation for numerical data.
* Generate frequency counts for categorical variables.
* Use visualizations such as histograms, box plots, scatter plots, and heatmaps to analyze data distributions.

**4. Exploratory Data Analysis (EDA)**

* Conduct correlation analysis to identify relationships between variables.
* Apply dimensionality reduction techniques like PCA (Principal Component Analysis) or t-SNE (t-Distributed Stochastic Neighbor Embedding) for visualization.
* Perform clustering to find patterns or groupings within the data.

**5. Hypothesis Testing and Feature Engineering**

* Formulate hypotheses about patterns or trends within the dataset.
* Conduct hypothesis tests to validate assumptions.
* Engineer new features by transforming existing ones to enhance model performance.

**6. Interactive Exploration and Tools**

* Utilize tools such as Jupyter Notebooks, Pandas, Matplotlib, Seaborn, Plotly, or Tableau for dynamic data exploration.
* Implement interactive visualizations to explore subsets of data effectively.

**7. Documentation and Communication**

* Document all findings, insights, and assumptions during data exploration.
* Present summaries, visualizations, and reports to communicate key observations.
* Prepare presentations for stakeholders or team members.

**8. Iterative Process**

* Data exploration is an ongoing process where multiple techniques are tested, refined, and re-evaluated to gain deeper insights.

**9. Ethical Considerations**

* Ensure ethical data usage, particularly regarding privacy concerns, biases, and responsible interpretation of insights.

**Data Preparation and Cleaning:**

Data preparation and cleaning ensure that the dataset is suitable for analysis or modelling. Key steps include:

1. **Handling Missing Values**

* Identify missing values using appropriate methods.
* Remove missing values or use imputation techniques (mean, median, mode).

1. **Handling Outliers**

* Detect outliers using statistical methods like z-scores or IQR.
* Decide whether to remove or transform outliers based on analysis needs.

1. **Dealing with Duplicates**

* Identify and remove duplicate entries to maintain data integrity.

1. **Feature Engineering**

* Create new features from existing ones to enhance predictive models.
* Transform skewed data using log transformations or scaling techniques.

1. **Standardization and Normalization**

* Scale numerical variables using standardization (z-score) or normalization (min-max scaling).

1. **Handling Categorical Variables**

* Convert categorical variables to numerical formats using encoding techniques.

1. **Data Splitting**

* Divide the dataset into training and testing sets for machine learning applications.

1. **Handling Date and Time Variables**

* Convert date/time values to appropriate formats for analysis.

**Learning Objectives:**

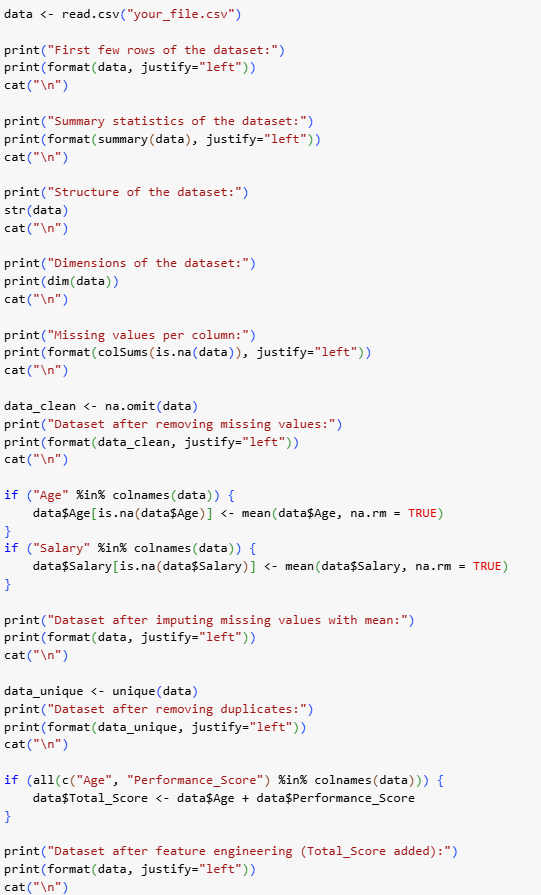
* Understanding different techniques for data exploration.
* Learning data preparation and cleaning methods.
* Implementing data handling processes using R.

**Conclusion:**

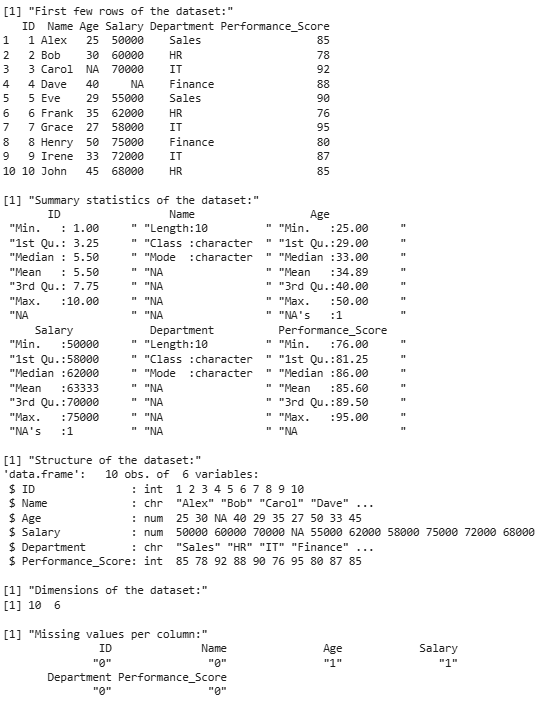
This experiment explored various data handling techniques, including loading, cleaning, and preprocessing datasets in R. The steps involved identifying missing values, handling outliers, and performing feature engineering to prepare data for further analysis.

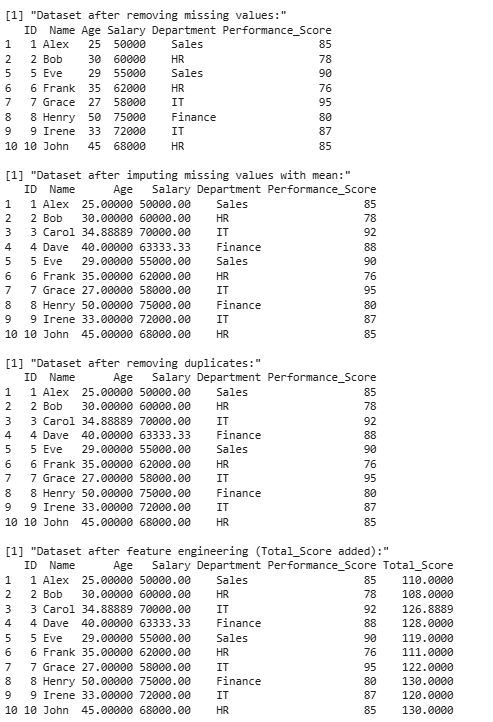


**Program:**

****

**Output:**

****

****