**Project Based Learning Report**

on

**Implementation of various preprocessing techniques (EDA)**

**CAR dataset using python**

Submitted in the partial fulfillment of the requirements

For the Project based learning in **Artificial Intelligence and Data Mining**

in

Electronics & Communication Engineering

By

Name of Students in Alphabetical order with Seat Number /PRN Number

**2014111057 Kumar Kartik**

**2014111039 Tanmay Dikshit**

**2014111049 Unnati Gupta**

Under the guidance of Course In-charge

Prof. Dr. Tanuja S. Dhope

Department of Electronics & Communication Engineering

BharatiVidyapeeth

(Deemed to be University)

College of Engineering,

Pune – 4110043

**Academic Year: 2023-24**

**Bharati Vidyapeeth**

**(Deemed to be University)**

**College of Engineering,**

**Pune – 411043**

**DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING**

**CERTIFICATE**

Certified that the Project Based Learning report entitled, “**Implementation of various preprocessing techniques(EDA) on CAR dataset using python”** is

work done by

(Name of the students in alphabetical order with Seat Number /PRN Number)

**2014111057 Kumar Kartik**

**2014111039 Tanmay Dikshit**

**2014111049 Unnati Gupta**

in partial fulfillment of the requirements for the award of credits for Project Based Learning (PBL) in **Artificial Intelligence and Data Mining** of Bachelor of Technology Semester VII, in **Electronics and Communication Engineering**.

**Date: 23/10/23**

**Prof. Dr. Tanuja S. Dhope Dr. Arundhati A.Shinde**

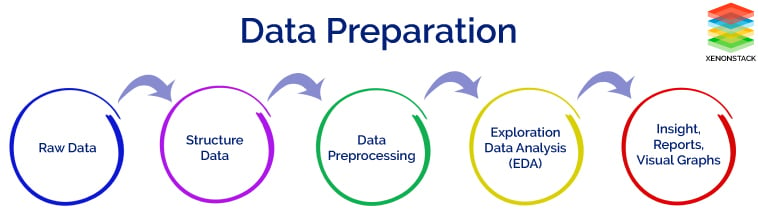
**Course In-charge Professor & Head**

|  |  |
| --- | --- |
| INDEX | |
| CONTENTS | PAGE NO. |
|  |  |
| Description of the Project | 4-7 |
|  |  |
| Solutions Code and Output | 8-13 |
|  |  |
| Result and Analysis | 14 |
|  |  |
| Conclusion and Course Outcome | 14 |

**Implementation of various Processing Techniques (EDA) on Car Dataset using Python**

Introduction to Exploratory Data Analysis (EDA)

Data pre-processing, Feature Engineering, and EDA are fundamental early steps after data collection. Still, they are not limited to where the data is simply visualized, plotted, and manipulated, without any assumptions, to assess the quality of the data and building models.



Data Preprocessing and Feature Engineering

In our data-driven processes, we consider refining our raw data. Both data pre-processing and feature engineering play pivotal roles in this endeavor. Data pre-processing encompasses a range of activities, including data integration, analysis, cleaning, transformation, and dimension reduction. Data pre-processing involves cleaning and preparing raw data to facilitate feature engineering. Meanwhile, feature engineering entails employing various techniques to manipulate the data. This may include adding or removing relevant features, handling missing data, encoding variables, and dealing with categorical variables, among other tasks.

Undoubtedly, feature engineering is a critical task that significantly influences the outcome of a model. It involves crafting new features based on existing data while pre-processing primarily focuses on cleaning and organizing the data.

**Exploratory Data Analysis (EDA):**

EDA is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. The primary goal of EDA is to uncover patterns, relationships, anomalies, and insights within the data. It helps analysts understand the structure of the data, detect outliers, and formulate hypotheses that can lead to further investigation. Python provides various libraries like Pandas, Matplotlib, Seaborn, and Plotly, making it a popular choice for performing EDA. Here's a step-by-step guide to performing EDA using Python

Important Points regarding EDA:

* **Descriptive Statistics:** EDA involves calculating summary statistics (mean, median, variance, etc.) to describe the main features of the dataset.
* **Visualizations:** Visualization techniques such as histograms, box plots, scatter plots, and heatmaps are used to visually explore the data distribution, identify outliers, and understand relationships between variables.
* **Handling Missing Data:** EDA includes identifying missing values and deciding how to handle them. Understanding the missing data pattern is crucial for making informed decisions.
* **Feature Relationships:** EDA helps in exploring how different features in the dataset relate to each other. Correlation analysis is often used to quantify these relationships.
* **Outlier Detection:** EDA methods can identify outliers, which are data points significantly different from others. Outliers can impact the analysis and often require special attention.
* **Data Distribution:** Understanding the distribution of data points across various features is vital. It helps in choosing appropriate statistical tests and machine learning algorithms.

**Data Preprocessing:**

Data preprocessing is the process of cleaning and transforming raw data into a format suitable for analysis. It plays a crucial role in improving the quality of data and ensuring that the analysis results are accurate and reliable.

Important Aspects of Data Preprocessing:

* **Data Cleaning:** This step involves handling missing values, correcting errors, and dealing with inconsistencies in the dataset. Cleaning ensures that the data is accurate and reliable.
* **Data Transformation:** Data transformation includes tasks such as normalization (scaling data to a standard range), encoding categorical variables into numerical values, and transforming skewed data distributions.
* **Feature Engineering:** Feature engineering involves creating new features from existing ones or selecting relevant features that have a significant impact on the analysis or modeling process. It can improve the performance of machine learning models.
* **Data Reduction:** In cases where the dataset is large and contains many irrelevant or redundant features, techniques like Principal Component Analysis (PCA) can be applied to reduce the dimensionality of the data while preserving its important characteristics.
* **Handling Imbalanced Data:** In classification problems, if the classes are imbalanced (i.e., one class has significantly fewer samples than the others), techniques like oversampling, undersampling, or using synthetic data can balance the class distribution.

Both EDA and Data Preprocessing are iterative processes. Analysts often revisit these steps based on the insights gained during analysis, refining their approach to better understand the data and prepare it for further analysis or modeling. Proper EDA and Data Preprocessing are foundational to deriving meaningful insights and building accurate predictive models from the data.

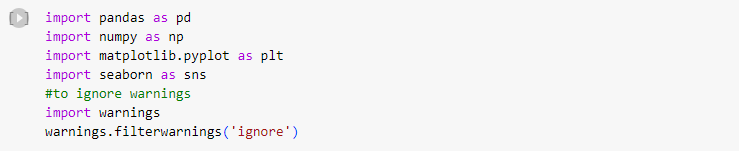
**Advantages of Exploratory Data Analysis (EDA):**  Various advantages of Exploratory Data Analysis are given below:

1. **Data Understanding**: EDA helps in understanding the data at a deeper level. Analysts gain insights into the structure, distribution, and patterns within the data, which is crucial for making informed decisions.
2. **Pattern Discovery**: EDA techniques, such as data visualization and summary statistics, aid in discovering hidden patterns or trends within the data. This is valuable for generating hypotheses and guiding further research.
3. **Identifying Anomalies**: EDA helps in detecting outliers and anomalies in the data. Identifying these unusual data points is vital for ensuring data quality and can lead to the discovery of interesting phenomena or errors in the dataset.
4. **Feature Selection**: EDA provides insights into the relationships between different features. Analysts can identify which features are most relevant for the analysis or modeling, leading to better feature selection and improved model performance.
5. **Effective Communication**: Visualizations generated during EDA simplify complex data into understandable formats, making it easier to communicate findings and insights to stakeholders who might not have a technical background.
6. **Data Quality Assessment**: EDA allows for a comprehensive examination of the data quality. By identifying missing values, duplicates, or inconsistencies, data quality issues can be addressed, ensuring the reliability of the analysis.

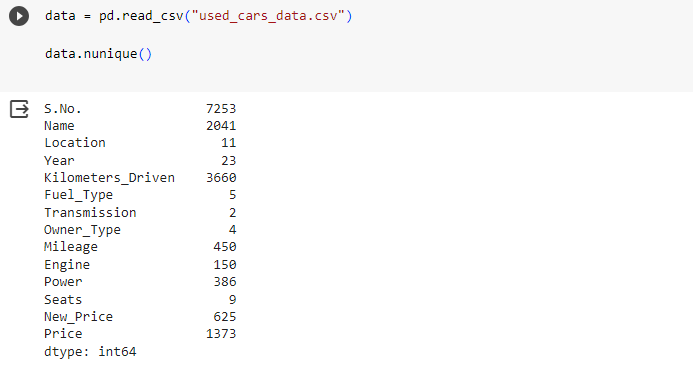
**SOLUTIONS:**

**CODE:**

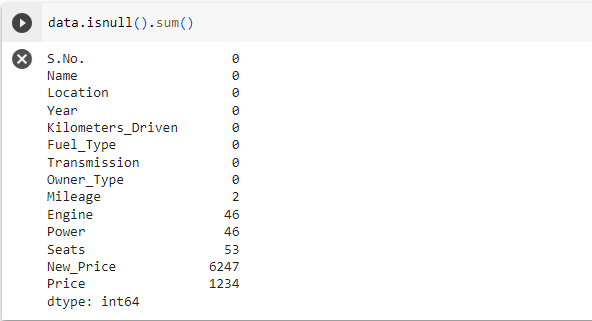
1. Importing Libraries:



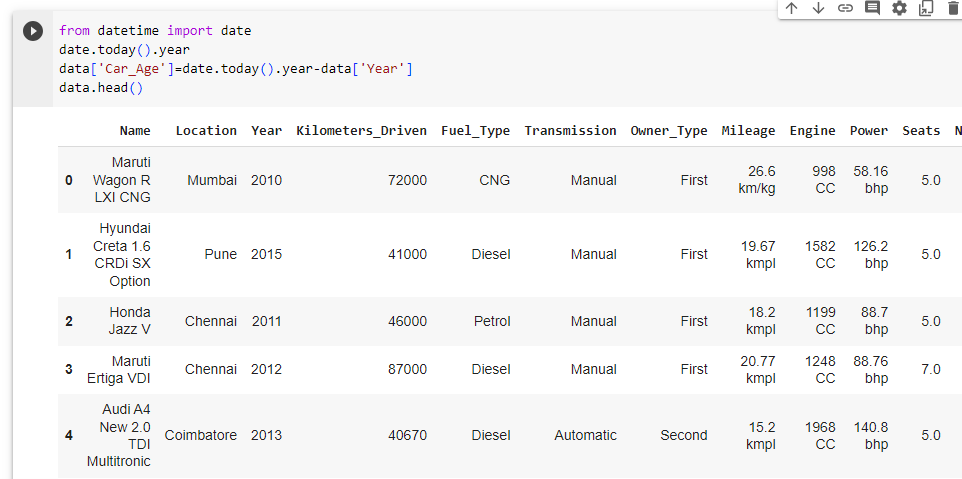
1. Check for Duplication:



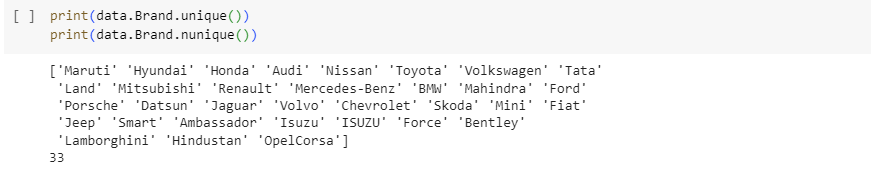
1. Missing Values Calculation:



1. Creating Features:

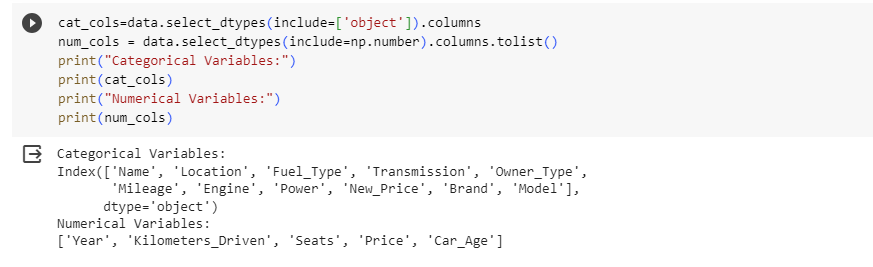


1. Data Cleaning/Wrangling:



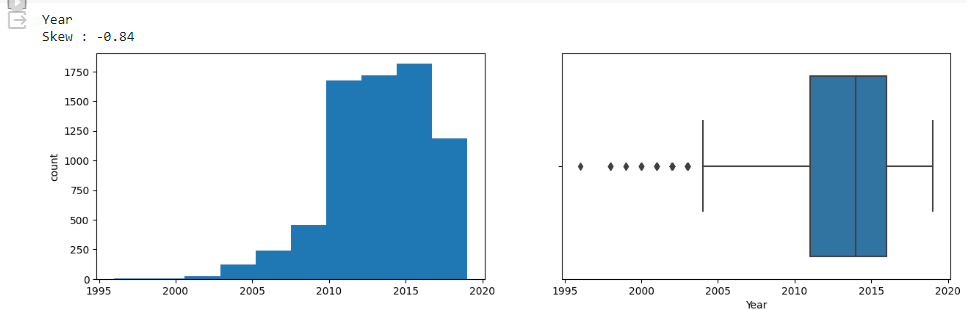


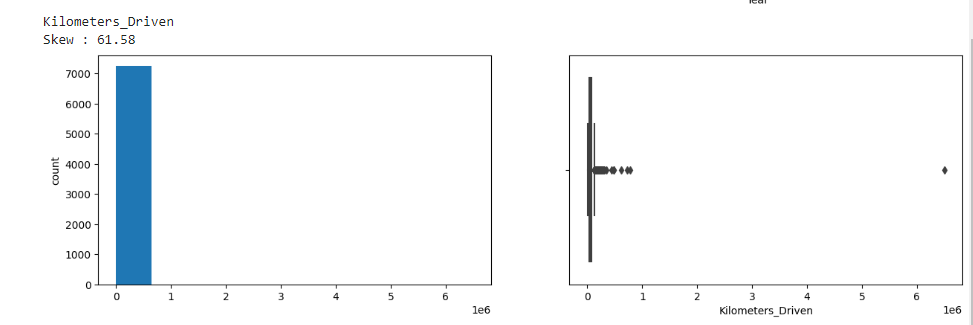
1. Separate Numerical and categorical variables:

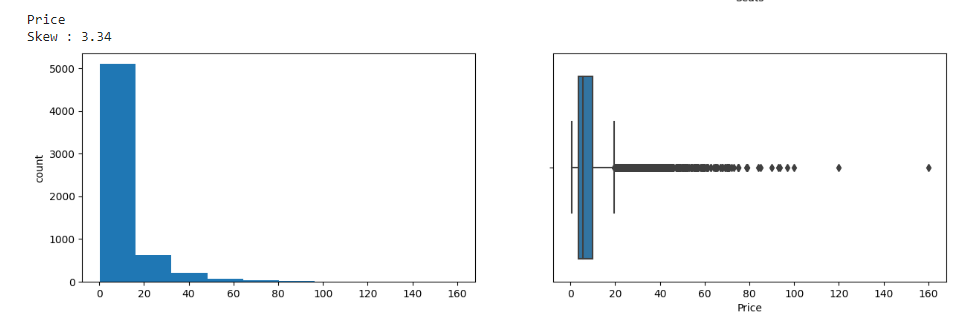
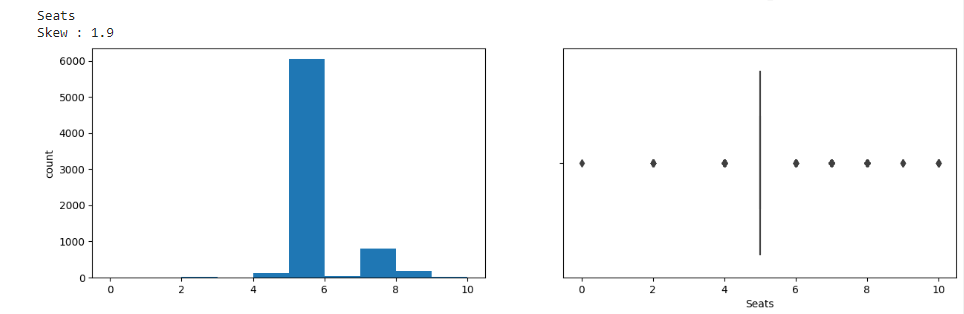


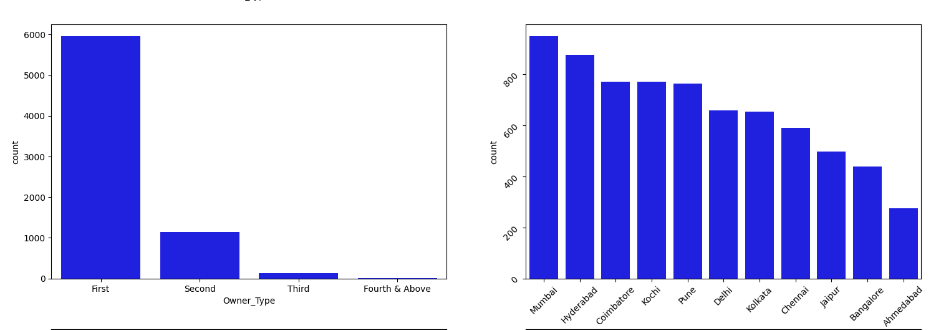
**SOLUTION (OUTPUTS):**

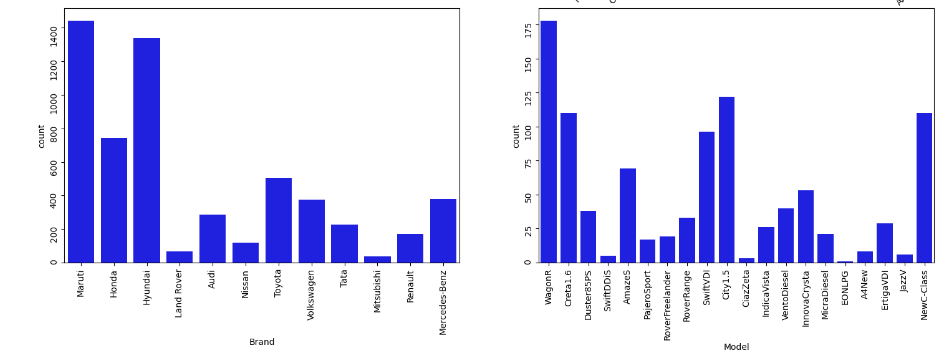
**Bar plots:**



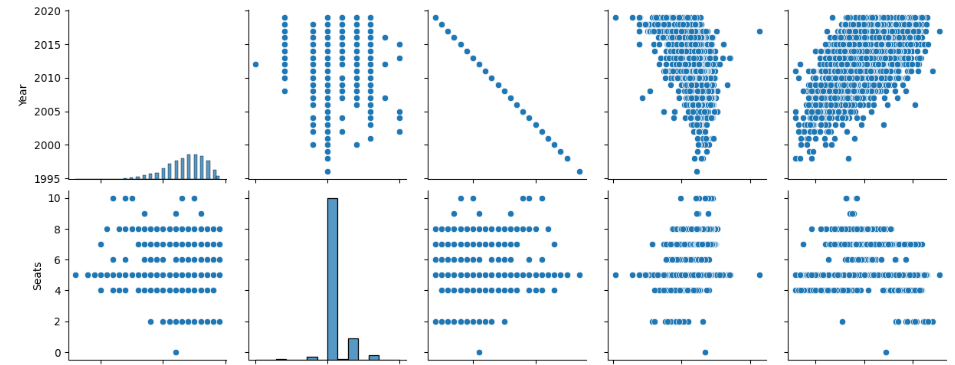


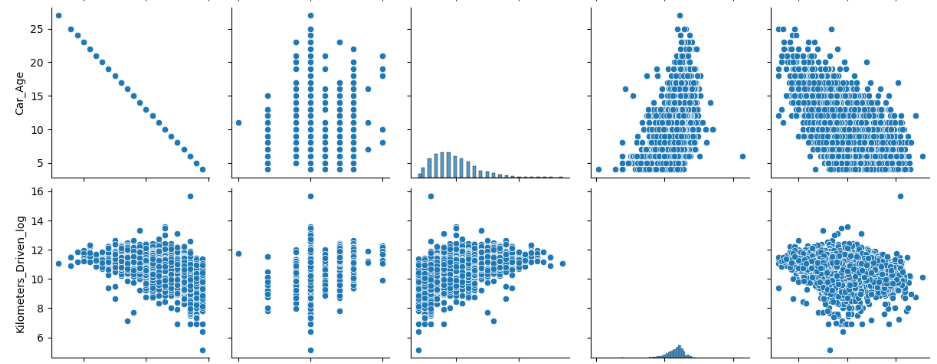




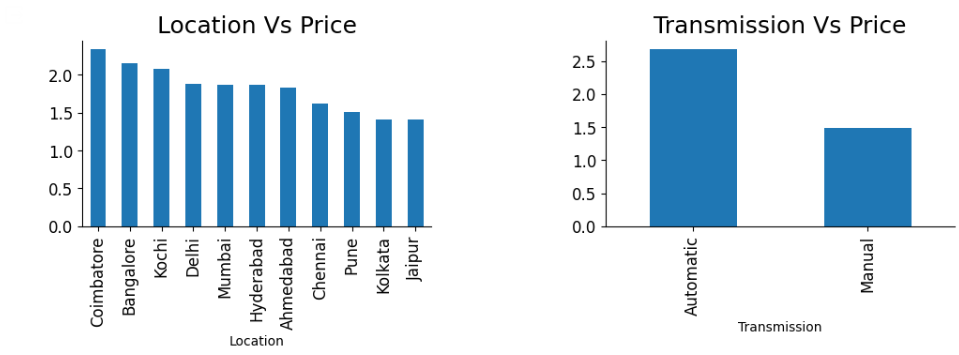


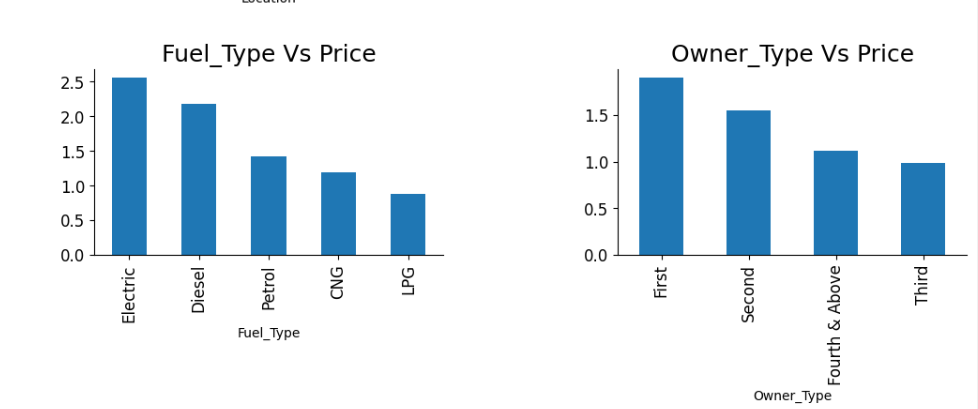
**Bivariate Analysis:**

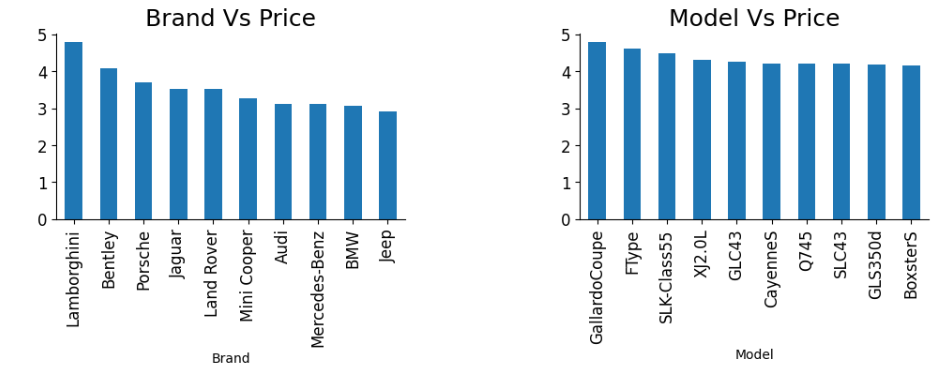


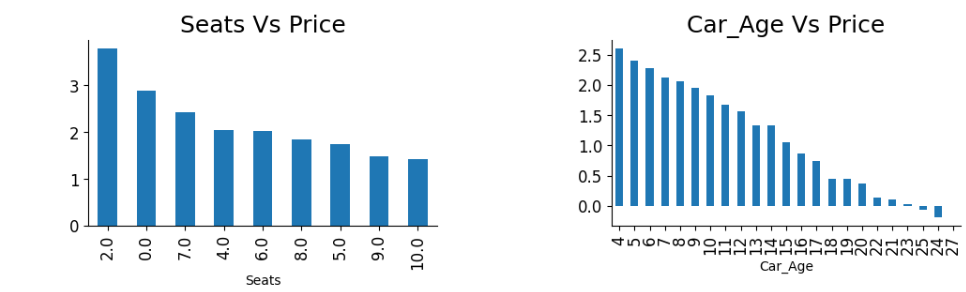


**Pair Plots:**

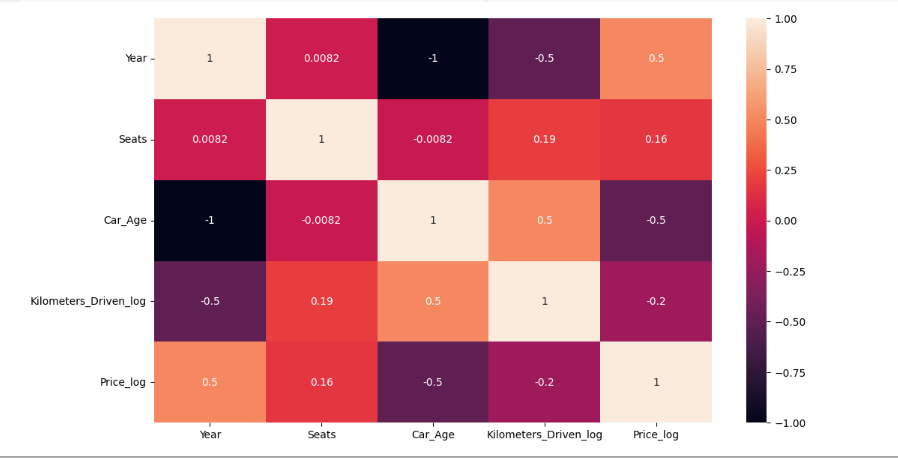








**Multivariate Analysis (heat map for all variables in dataset):**



**RESULT AND ANALYSIS:**

In this project, we performed Exploratory Data Analysis (EDA) on Car Data.

While performing EDA, we ensured several steps were performed correctly. The analysis was conducted using Python programming language and popular data analysis libraries including Pandas, Matplotlib, and Seaborn. We plotted various graphs like Bar, Histogram along with Scatter Plots and Multivariate Heat Map.

EDA very much helped us to reduce the unwanted data and get different relations between two or more factors of the dataset.

**CONCLUSION AND COURSE OUTCOME:**

Our EDA revealed valuable insights about the used car dataset. The analysis highlighted the influence of factors like age, mileage, make, and model on car prices. We identified outliers and patterns that can guide further analysis and modeling efforts. These findings are essential for making informed decisions, whether for buyers, sellers, or market analysts in the automotive industry.

Hence, the EDA was performed successfully based on the **CO-4(Understand the basic concept of data mining and its functionality).**

**APPENDIX:**

https://github.com/Unnati-1115/AIDM-PBL