Lab Assignment 4

Machine Learning

Name: M Jeyapathy

USN: 22BTRAD016

Questions:

1.Load a dataset with outliers values(Boston Housing Dataset)

import numpy as np

import pandas as pd

from sklearn.datasets import load\_boston

# Load the original Boston Housing Dataset

boston = load\_boston()

boston\_df = pd.DataFrame(boston.data, columns=boston.feature\_names)

boston\_df['PRICE'] = boston.target

# Introduce outliers (for example, by multiplying a column by a constant)

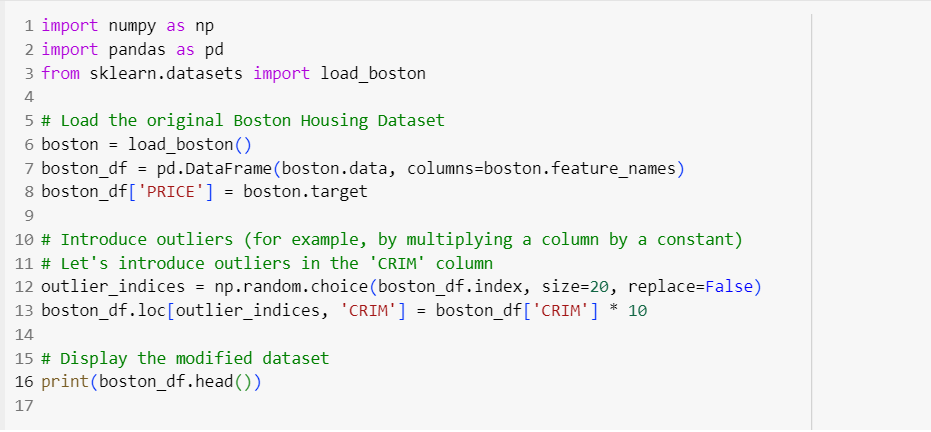
# Let's introduce outliers in the 'CRIM' column

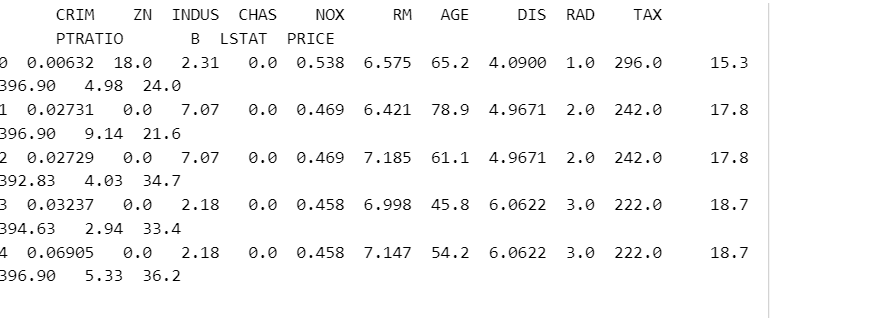
outlier\_indices = np.random.choice(boston\_df.index, size=20, replace=False)

boston\_df.loc[outlier\_indices, 'CRIM'] = boston\_df['CRIM'] \* 10

# Display the modified dataset

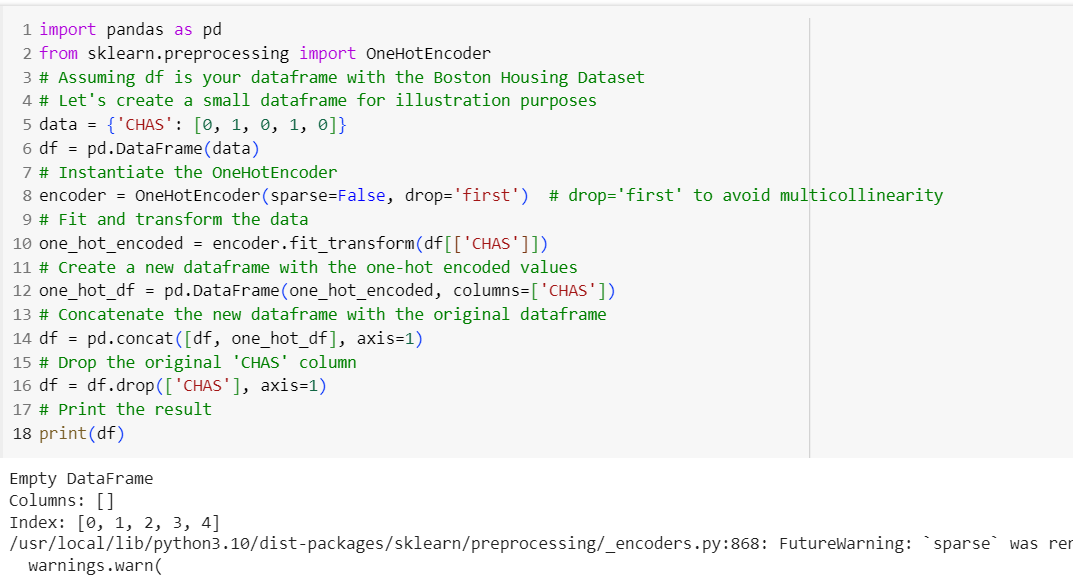
print(boston\_df.head())





we loaded the Boston Housing Dataset and introduced outliers in the 'CRIM' column by multiplying the values by 10 for a randomly chosen subset of rows.

2.Implement one-hot encoding



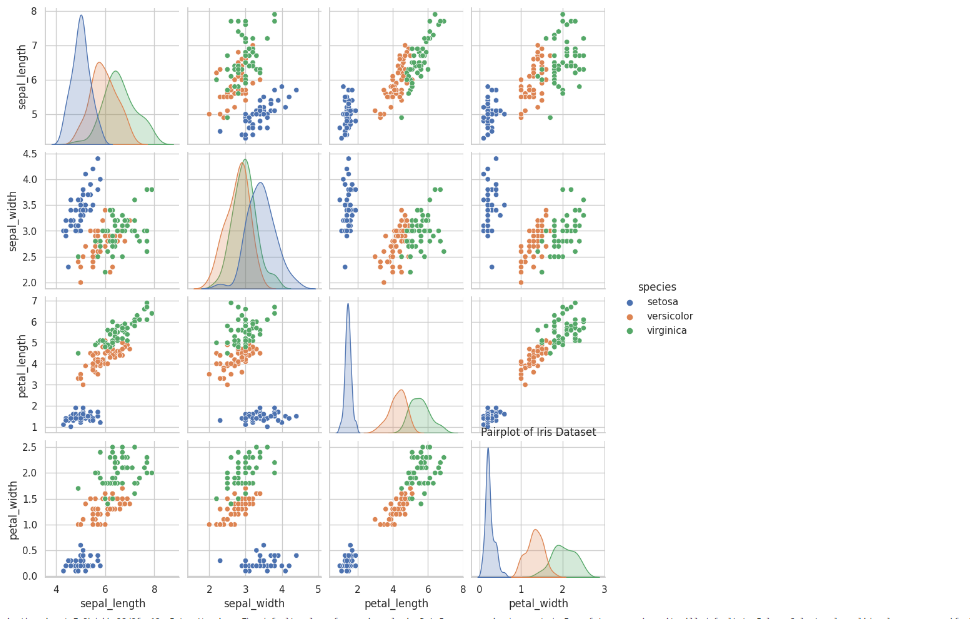
the 'CHAS' column is one-hot encoded, and a new column is added to the dataframe. The original 'CHAS' column is then dropped, we would apply one-hot encoding to categorical variables with more than two categories.

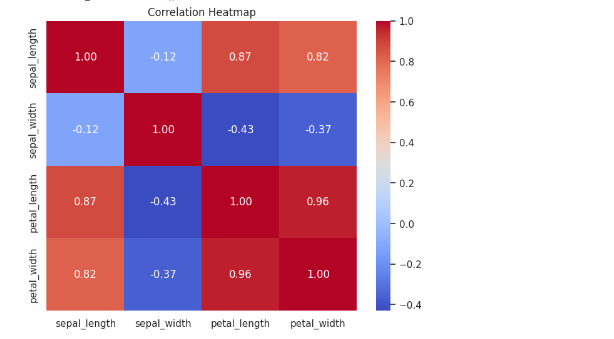
The purpose of one-hot encoding is to represent categorical variables as binary vectors.

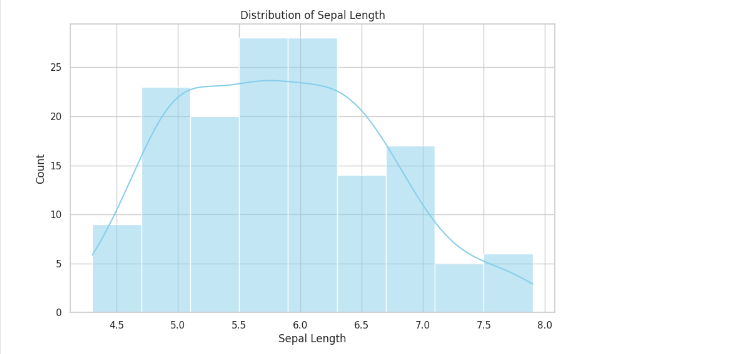
. In this case, the original 'CHAS' column with binary values (0 or 1) is transformed into a new column with one-hot encoded values, creating a more suitable representation for certain machine learning algorithms. The parameter **drop='first'** is used to avoid multicollinearity by dropping the first category, as it can be inferred from the others.

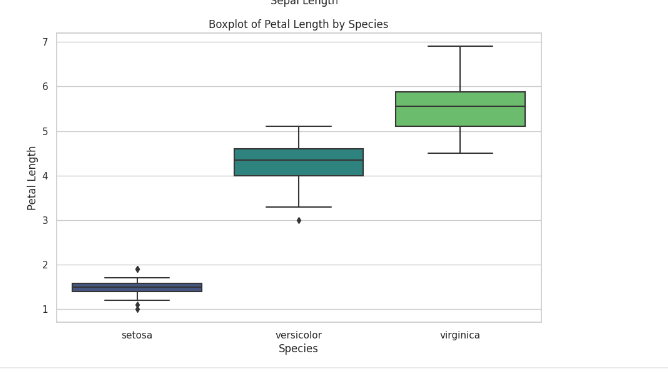
3.Create visualization for different aspects of a dataset using Matplotlib or seaborn









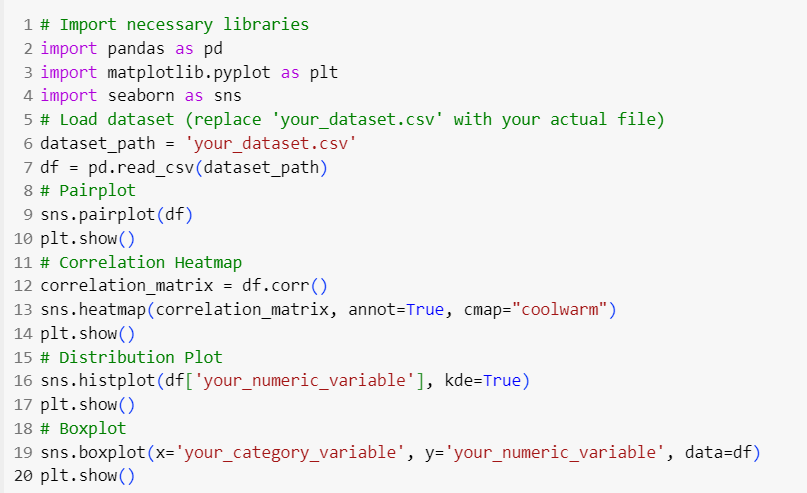


Different aspects of a dataset using matplotlib or seaborn is done by:

1. **Pairplot:** A matrix of scatterplots for exploring relationships between variables.
2. **Correlation Heatmap:** A heatmap visualizing the correlation between numeric variables.
3. **Distribution Plot:** A histogram for visualizing the distribution of a numeric variable.
4. **Boxplot:** A boxplot for visualizing the distribution of a numeric variable across different categories.

**4.**Interpret the visualization to gain insights into the dataset

1. **Pairplot:**
   * **Insight:** The pairplot shows scatterplots for each pair of variables and histograms for individual variables. Look for patterns and relationships between variables.
   * **Interpretation:** If points cluster in a specific way or if there are clear trends, it indicates potential correlations or patterns in the data.
2. **Correlation Heatmap:**
   * **Insight:** The heatmap visualizes the correlation between numeric variables.
   * **Interpretation:** Darker colors (closer to -1 or 1) indicate stronger correlations. Positive correlations suggest variables moving together, while negative correlations suggest an inverse relationship.
3. **Distribution Plot:**
   * **Insight:** The histogram shows the distribution of a specific numeric variable.
   * **Interpretation:** Check for the shape of the distribution. A normal distribution is bell-shaped, while skewed distributions have a longer tail on one side.
4. **Boxplot:**
   * **Insight:** The boxplot visualizes the distribution of a numeric variable across different categories.
   * **Interpretation:** Check for differences in the central tendency (median) and spread between categories. Outliers, if present, can be identified as individual points outside the "whiskers" of the box.



A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

5.Perform Univariate and multivariate analysis for the dataset



A number of rooms and bedrooms

Description automatically generated with medium confidence

A graph of different sizes and numbers

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

A graph with a square and a line

Description automatically generated with medium confidence

Performing univariate and multivariate analysis involves examining the variables in your dataset individually (univariate) and then exploring the relationships between variables (multivariate).

* Univariate Analysis: This involves examining each variable individually. In this example, we're looking at summary statistics, histograms, and box plots for each variable.
* Multivariate Analysis: This involves exploring relationships between variables. Here, we're creating a correlation matrix to see how variables are correlated and a pairplot to visualize relationships between selected variables.