Exploratory Data Analysis (EDA)

PART OF DATA PRE-PROCESSING

Kamalesh

@kamalesh12

What is Exploratory Data Analysis (EDA)?

Exploratory Data Analysis (EDA) is an essential step in any data science project. It involves investigating and analyzing datasets to understand their characteristics, identify patterns, detect outliers, and uncover relationships between variables. EDA helps in gaining initial insights into the data before diving into more complex analyses.

The Foremost Goals of EDA

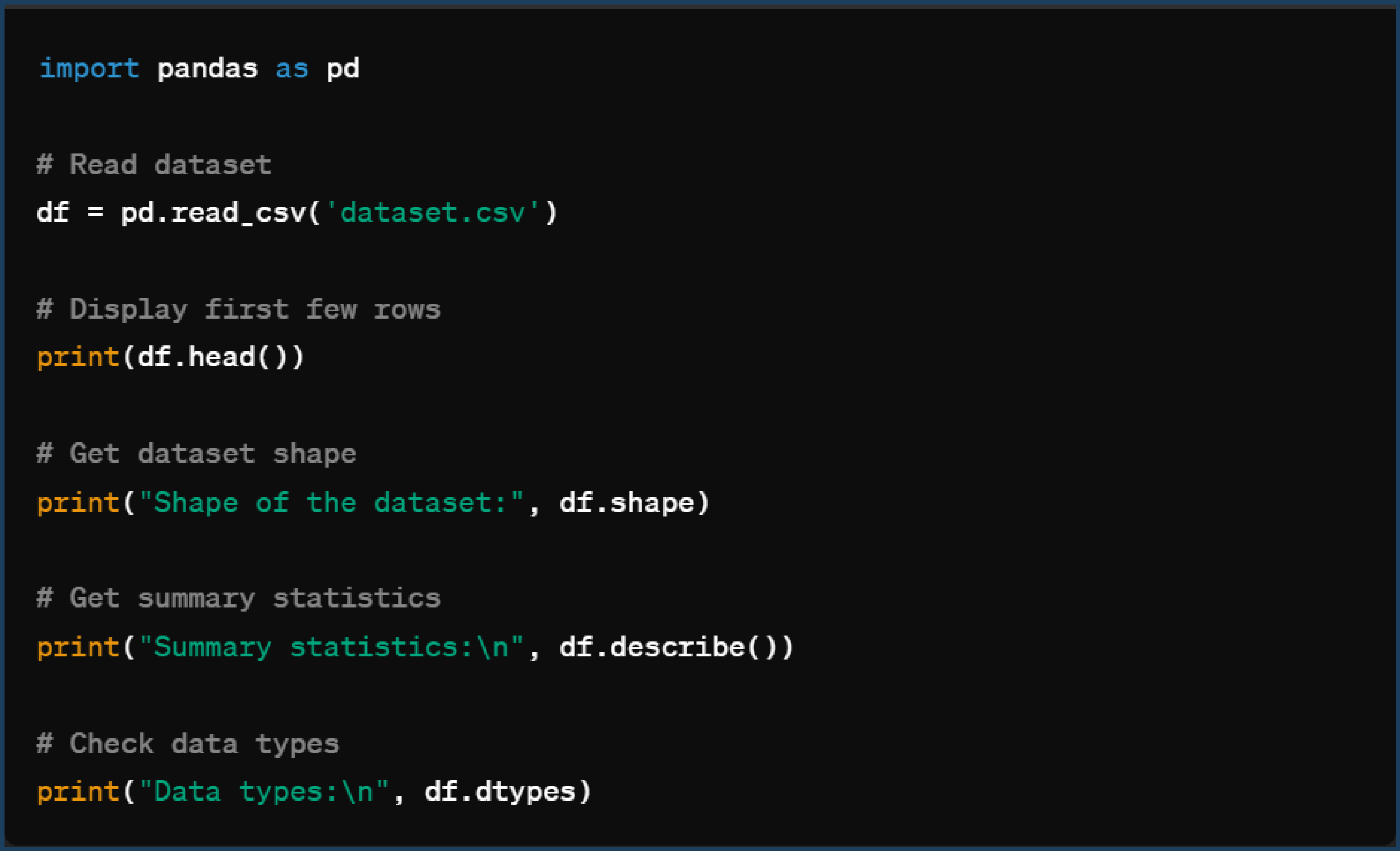
1. Descriptive Statistics
2. Data Visualization
3. Feature Engineering
4. Correlation and Relationships
5. Data Segmentation
6. Hypothesis Generation
7. Data Quality Assessment

Types of EDA

1. Bivariate Analysis
2. Multivariate Analysis
3. Time Series Analysis
4. Missing Data Analysis
5. Outlier Analysis
6. Data Visualization

EDA Using Python Libraries

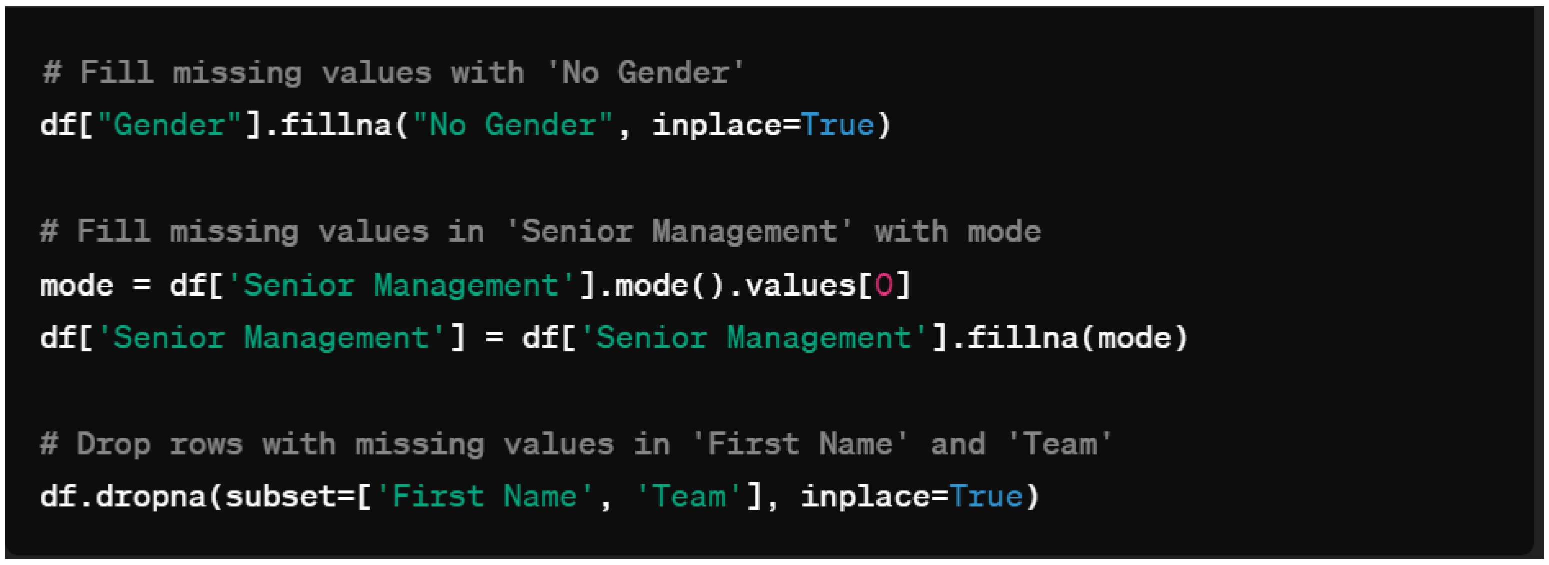
Python libraries like Pandas and Matplotlib are commonly used for EDA. Techniques such as data reading, summary statistics, data type conversion, handling missing values, and data visualization are performed using these libraries.



```python
import pandas as pd

# Read dataset
df = pd.read_csv('dataset.csv')

# Display first few rows
print(df.head())

# Get dataset shape
print("Shape of the dataset:", df.shape)

# Get summary statistics
print("Summary statistics:\n", df.describe())

# Check data types
print("Data types:\n", df.dtypes)
```

Handling Missing Values

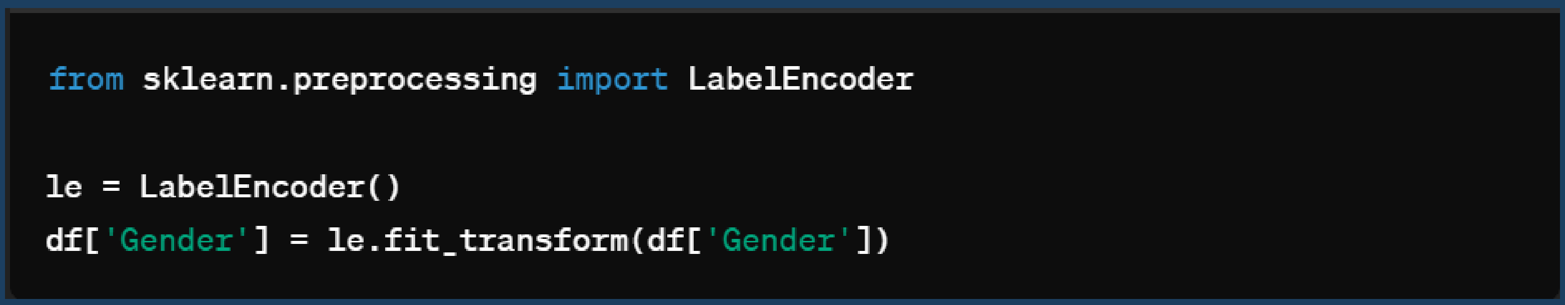
Missing data can impact analysis. Techniques such as filling missing values, dropping rows with missing data, and data imputation are used to handle missing values effectively.



```
# Fill missing values with 'No Gender'
df["Gender"].fillna("No Gender", inplace=True)

# Fill missing values in 'Senior Management' with mode
mode = df['Senior Management'].mode().values[0]
df['Senior Management'] = df['Senior Management'].fillna(mode)

# Drop rows with missing values in 'First Name' and 'Team'
df.dropna(subset=['First Name', 'Team'], inplace=True)
```

Data Encoding

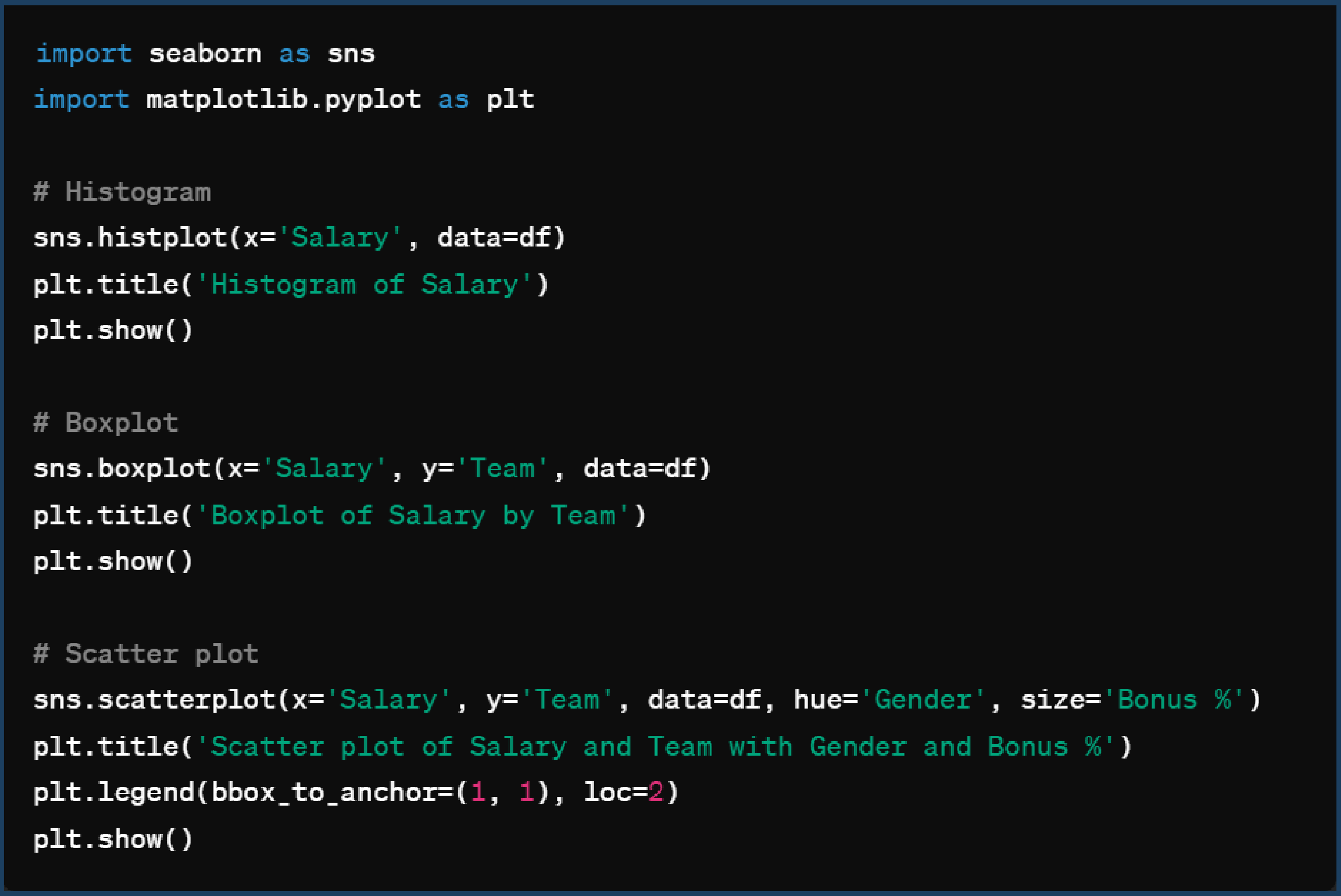
Categorical data may need to be encoded into numerical columns for certain models. Techniques like Label Encoding and One-hot Encoding can be used for this purpose.



```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
df['Gender'] = le.fit_transform(df['Gender'])
```

Data Visualization Techniques

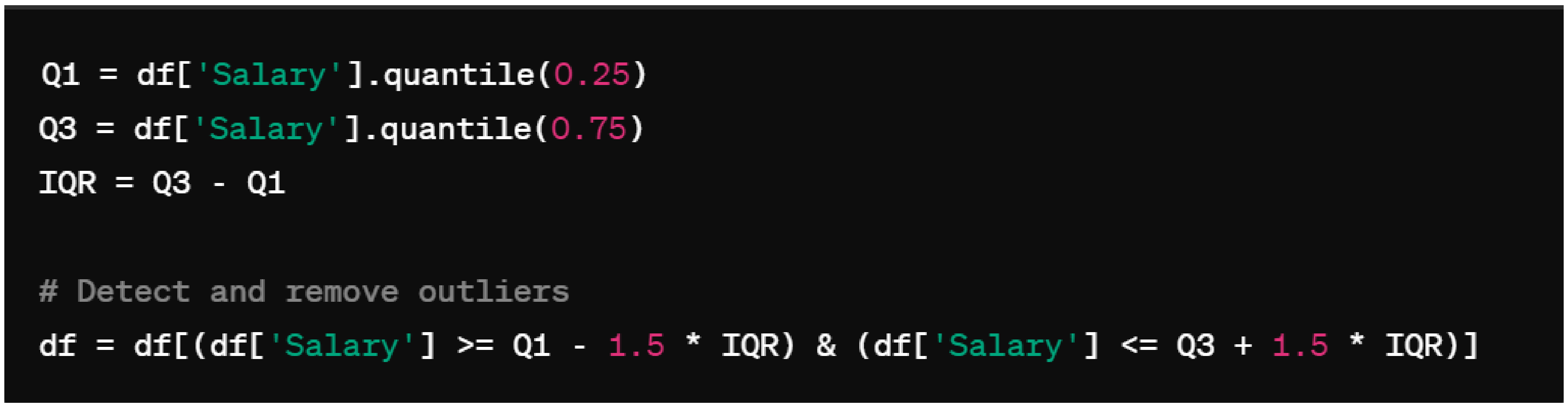
Various visualization techniques such as histograms, box plots, scatter plots, and pair plots are used to explore data visually and understand trends and patterns.



```
import seaborn as sns
import matplotlib.pyplot as plt

# Histogram
sns.histplot(x='Salary', data=df)
plt.title('Histogram of Salary')
plt.show()

# Boxplot
sns.boxplot(x='Salary', y='Team', data=df)
plt.title('Boxplot of Salary by Team')
plt.show()

# Scatter plot
sns.scatterplot(x='Salary', y='Team', data=df, hue='Gender', size='Bonus %')
plt.title('Scatter plot of Salary and Team with Gender and Bonus %')
plt.legend(bbox_to_anchor=(1, 1), loc=2)
plt.show()
```

Handling Outliers

Outliers, data points significantly deviating from the rest, can affect analysis. Techniques like Interquartile Range (IQR) method are used to detect and remove outliers.



```
Q1 = df['Salary'].quantile(0.25)
Q3 = df['Salary'].quantile(0.75)
IQR = Q3 - Q1

# Detect and remove outliers
df = df[(df['Salary'] >= Q1 - 1.5 * IQR) & (df['Salary'] <= Q3 + 1.5 * IQR)]
```

Handling Missing Values

Missing data can impact analysis. Techniques such as filling missing values, dropping rows with missing data, and data imputation are used to handle missing values effectively.

Follow me for more tips to help you connect with your audience

LEAVE A COMMENT BELOW

Kamalesh K B

@kamalesh12