EDA-Data Visualization

Data Visualization for Exploratory Data Analysis (EDA)

Data visualization is a crucial step in the data analysis process, allowing us to understand and communicate complex data insights effectively. During Exploratory Data Analysis (EDA), we use various visualization techniques to:

1. **Understand the distribution of variables**
2. **Identify relationships between variables**
3. **Detect outliers and anomalies**

Here's a summary of common data visualization techniques for EDA, along with examples:

1. Histograms

- Purpose: Understand the distribution of a single variable.
- Example:

```
 import pandas as pd
import matplotlib.pyplot as plt

# Load sample data (e.g., exam scores)
data = pd.DataFrame({'score': [80, 90, 70, 85, 95]})

# Plot histogram
plt.hist(data['score'], bins=5, edgecolor='black')
plt.xlabel('Score')
plt.ylabel('Frequency')
plt.title('Exam Score Distribution')
plt.show()
```

2. Scatter Plots

- Purpose: Identify relationships between two variables.
- Example:

```
 import pandas as pd
import matplotlib.pyplot as plt

# Load sample data (e.g., exam scores and study time)
data = pd.DataFrame({'score': [80, 90, 70, 85, 95], 'study_time': [2, 3, 1.5, 2.5, 4]

# Plot scatter plot
plt.scatter(data['study_time'], data['score'])
plt.xlabel('Study Time (hours)')
plt.ylabel('Exam Score')
plt.title('Relationship between Study Time and Exam Score')
plt.show()
```

3. Bar Charts

- Purpose: Compare categorical variables.
- Example:

```
 import pandas as pd
import matplotlib.pyplot as plt

# Load sample data (e.g., exam scores by subject)
data = pd.DataFrame({'subject': ['Math', 'Science', 'English'],
                     'score': [80, 90, 70]})

# Plot bar chart
plt.bar(data['subject'], data['score'])
plt.xlabel('Subject')
plt.ylabel('Exam Score')
plt.title('Exam Scores by Subject')
plt.show()
```

4. Box Plots

- Purpose: Compare distributions of multiple variables.
- Example:

```
 import pandas as pd
import matplotlib.pyplot as plt

# Load sample data (e.g., exam scores for different subjects)
data = pd.DataFrame({'subject': ['Math', 'Science', 'English'],
                     'score': [80, 90, 70]})

# Plot box plot
plt.boxplot(data['score'], labels=data['subject'])
plt.title('Exam Scores by Subject')
plt.show()
```

5. Heatmaps

- Purpose: Identify correlations between multiple variables.
- Example:

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

# Load sample data (e.g., exam scores for different subjects)
data = pd.DataFrame({'Math': [80, 90],
                     'Science': [70, 85],
                     'English': [60, 75]})

# Plot heatmap
plt.figure(figsize=(8,6))
sns.heatmap(data.corr(), annot=True, cmap='coolwarm', square=True)
plt.title('Correlation Matrix')
plt.show()
```

These are just a few examples of the many data visualization techniques available for EDA. By using these visualizations, you can gain valuable insights into your data and make informed decisions about further analysis or modeling.

Curated by Brajesh Kumar