Interconnected EDA

Exploratory Data Analysis (EDA) is the process of analyzing and visualizing datasets to uncover patterns, relationships, and insights. When datasets are interconnected, the analysis becomes more complex and insightful because columns in one dataset can influence or act as targets in others. This interconnected EDA provides a holistic view of the data, enabling us to identify cross-dataset relationships and dependencies.

Why Interconnected Datasets?

In real-world scenarios, datasets rarely exist in isolation. For example:

1. Employee Performance Data:

Tracks individual performance metrics.

Directly influences retention rates and salaries.

2. Employee Retention Data:

Affected by performance, job satisfaction, and work-life balance.

Key for HR strategies.

3. Employee Salary Data:

Dependent on experience, education, and performance.

Impacts satisfaction and retention.

4. Department Analysis:

Aggregates individual metrics (e.g., performance scores, salaries) to provide organizational insights.

By connecting these datasets, we can:

Discover how performance affects retention.

Understand the relationship between experience, education, and salary.

Link individual data to department-level metrics like retention rates and average salaries.

Key Steps in Interconnected EDA

1. Individual Dataset Analysis:

Analyze each dataset independently to understand its structure and distribution.

Perform basic statistics (mean, median, correlation) and visualize features.

2. Identify Relationships Across Datasets:

Recognize which columns in one dataset influence or are influenced by another dataset.

For example:

Performance_Score (Dataset 1) affects Retention (Dataset 2).

Experience (Dataset 3) impacts Salary (Dataset 3).

3. Merge Datasets:

Combine datasets on common keys (e.g., Employee_ID) to enable cross-dataset analysis.

Use features from one dataset as predictors for targets in another.

4. Perform Combined Analysis:

Investigate relationships between columns from different datasets.

Example: Analyze how Performance_Score impacts Salary and Retention.

5. Visualize Relationships:

Use scatter plots, bar charts, and heatmaps to represent cross-dataset relationships.

Identify trends, correlations, and outliers visually.

Detailed Example Using the Provided Datasets

Step 1: Analyze Individual Datasets

1. Employee Performance (Dataset 1):

Variables: Age, Experience, Education_Level, Performance_Score.

Key Insight: Performance_Score is likely influenced by Experience and Education_Level.

2. Employee Retention (Dataset 2):

Variables: Performance_Score, Job_Satisfaction, Work_Life_Balance, Retention.

Key Insight: High Job_Satisfaction and Work_Life_Balance increase Retention.

3. Employee Salary (Dataset 3):

Variables: Experience, Education_Level, Salary.

Key Insight: Salary increases with Experience and higher Education_Level.

4. Department Analysis (Dataset 4):

Aggregates from individual data: Performance_Score_Avg, Retention_Rate, Avg_Salary.

Key Insight: Departments with high Performance_Score_Avg likely have better retention and salaries.

Step 2: Interconnect Datasets

Identify relationships:

Performance_Score from Dataset 1 is used in Dataset 2 and influences retention.

Experience and Education_Level from Dataset 1 contribute to Salary in Dataset 3.

Department-level aggregates in Dataset 4 derive from Dataset 1 and Dataset 3.

Step 3: Visualize Relationships

1. Heatmap (Correlation):

Use a heatmap to show correlations in each dataset.

Example: Performance_Score correlates strongly with Retention in Dataset 2.

2. Scatter Plot:

Plot Performance_Score vs. Salary to see if better performance leads to higher pay.

3. Bar Plot:

Compare average salaries by education level to understand their impact.

4. Line Plot:

Show the relationship between department-level Performance_Score_Avg and Retention_Rate.

Insights from Interconnected EDA

1. Performance and Retention:

Higher Performance_Score improves the likelihood of retention.

Departments with higher average performance scores have better retention rates.

2. Experience, Education, and Salary:

Salary increases with Experience and higher Education_Level.

This trend justifies investing in employee education and skill development.

3. Cross-Dataset Aggregates:

Department-level retention and salary data depend heavily on individual performance and experience.

Aggregated metrics are vital for organizational-level decision-making.

EDA Code for Interconnected Datasets

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Suppress warnings for a cleaner output
```

```
import warnings
warnings.filterwarnings("ignore")

# Create Dataset 1: Employee Performance
data1 = {
    "Employee_ID": [1, 2, 3, 4, 5],
    "Age": [25, 30, 35, 28, 40],
    "Experience": [2, 5, 10, 4, 15],
    "Education_Level": ["Bachelor's", "Master's", "PhD", "Master's", "PhD"],
    "Performance_Score": [75, 85, 95, 80, 90]
}

df1 = pd.DataFrame(data1)

# Create Dataset 2: Employee Retention
data2 = {
    "Employee_ID": [1, 2, 3, 4, 5],
    "Performance_Score": [75, 85, 95, 80, 90],
    "Job_Satisfaction": [3, 4, 5, 3, 4],
    "Work_Life_Balance": [4, 5, 5, 3, 4],
    "Retention": ["Yes", "Yes", "Yes", "No", "Yes"]
}

df2 = pd.DataFrame(data2)

# Create Dataset 3: Employee Salary
```

```
data3 = {
    "Employee_ID": [1, 2, 3, 4, 5],
    "Experience": [2, 5, 10, 4, 15],
    "Education_Level": ["Bachelor's", "Master's", "PhD", "Master's", "PhD"],
    "Salary": [30000, 50000, 80000, 45000, 100000]
}

df3 = pd.DataFrame(data3)

# Create Dataset 4: Department Analysis
data4 = {
    "Department_ID": ["D1", "D2", "D3", "D4", "D5"],
    "Performance_Score_Avg": [80, 85, 75, 90, 88],
    "Retention_Rate": [90, 95, 85, 98, 92],
    "Avg_Salary": [45000, 50000, 40000, 60000, 55000]
}

df4 = pd.DataFrame(data4)

# Display datasets
print("Dataset 1: Employee Performance")
print(df1)
print("\nDataset 2: Employee Retention")
print(df2)
print("\nDataset 3: Employee Salary")
print(df3)
```

```
print("\nDataset 4: Department Analysis")
print(df4)

# ----------------------------
# Start of EDA with Visualizations
# ----------------------------

# 1. **Dataset 1: Employee Performance**
print("\n--- Dataset 1 Analysis ---")
print(df1.describe())

# Correlation Heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(df1.corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap - Employee Performance")
plt.show()

# Pairplot
sns.pairplot(df1, hue="Education_Level", palette="viridis")
plt.show()

# 2. **Dataset 2: Employee Retention**
print("\n--- Dataset 2 Analysis ---")
print(df2.describe())

# Countplot for Retention
```

```
sns.countplot(x="Retention", data=df2, palette="pastel")
plt.title("Retention Count")
plt.show()

# Relationship between Performance Score and Retention
sns.boxplot(x="Retention", y="Performance_Score", data=df2, palette="Set2")
plt.title("Performance Score vs Retention")
plt.show()

# 3. **Dataset 3: Employee Salary**
print("\n--- Dataset 3 Analysis ---")
print(df3.describe())

# Salary Distribution
sns.histplot(df3["Salary"], kde=True, bins=10, color="skyblue")
plt.title("Salary Distribution")
plt.show()

# Relationship between Experience and Salary
sns.scatterplot(x="Experience", y="Salary", hue="Education_Level", data=df3,
palette="cool")
plt.title("Experience vs Salary")
plt.show()

# 4. **Dataset 4: Department Analysis**
print("\n--- Dataset 4 Analysis ---")
```

```
print(df4.describe())


# Retention Rate vs Performance Score Avg
sns.lineplot(x="Performance_Score_Avg", y="Retention_Rate", marker="o", data=df4, color="red")
plt.title("Performance Score Avg vs Retention Rate")
plt.show()


# Avg Salary Distribution by Department
sns.barplot(x="Department_ID", y="Avg_Salary", data=df4, palette="magma")
plt.title("Average Salary by Department")
plt.show()


# ----------------------------
# Interconnected Analysis
# ----------------------------


# Merge datasets to analyze relationships
merged_df = pd.merge(df1, df2, on="Employee_ID", suffixes=("_Performance", "_Retention"))
merged_df = pd.merge(merged_df, df3, on=["Employee_ID", "Experience", "Education_Level"])


print("\n--- Merged Dataset Analysis ---")
print(merged_df.head())


# Performance Score vs Salary
```

```
sns.scatterplot(x="Performance_Score", y="Salary", hue="Retention", data=merged_df, palette="Set1")

plt.title("Performance Score vs Salary by Retention")

plt.show()


# Job Satisfaction vs Salary

sns.barplot(x="Job_Satisfaction", y="Salary", hue="Retention", data=merged_df, palette="cool")

plt.title("Job Satisfaction vs Salary by Retention")

plt.show()


# Summary Insights

print("\n--- Insights ---")

print("""

1. Higher performance scores are linked with higher salaries.

2. Employees with higher job satisfaction are more likely to stay (Retention).

3. Departments with higher average performance scores have better retention rates.

4. Experience and education level significantly affect salary distributions.

""")
```