Employee Salary Analysis

Name: Shivam Pandey

Roll No: 202401100300234

Branch & Section: CSE AI D

Date: 11th March 2025

Employee Salary Analysis Report

Introduction

This report provides an analysis of employee salary data using Python. The project explores salary distributions, identifies outliers, and examines correlations between salary and other factors like department and experience.

Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn

Libraries Used

Pandas

Pandas is used for data manipulation and analysis. It provides data structures like DataFrames to efficiently handle large datasets.

Example Code:

```
import pandas as pd
df = pd.read_csv("employee_data.csv")
```

NumPy

NumPy is used for numerical computations, including handling arrays and mathematical functions.

Example Code:

```
import numpy as np
```

Matplotlib

Matplotlib is used for creating static visualizations such as histograms and boxplots.

Example Code:

```
import matplotlib.pyplot as plt
plt.hist(df['Salary'], bins=30)
plt.show()
```

Seaborn

Seaborn is used for advanced visualizations, including correlation heatmaps and KDE plots.

Example Code:

```
import seaborn as sns
sns.histplot(df['Salary'], bins=30, kde=True)
plt.show()
```

Data Overview

The dataset contains employee salary information, including:
- Department (Categorical)
- Experience (Numerical)
- Job Title (Categorical)
- Salary (Numerical)

Data Processing

Handling Missing Values:

```
print(df.isnull().sum())
df = df.dropna(subset=['Salary'])
```

Data Visualization

Salary Distribution:

```
plt.figure(figsize=(8, 5))
sns.histplot(df['Salary'], bins=30, kde=True, color='blue')
plt.title("Salary Distribution")
plt.xlabel("Salary")
plt.ylabel("Frequency")
plt.show()
```

Boxplot for Outlier Detection:

```
plt.figure(figsize=(8, 5))
sns.boxplot(x=df['Salary'], color='red')
plt.title("Salary Outliers")
plt.show()
```

Correlation Analysis

The correlation between different numerical factors was examined using a heatmap:
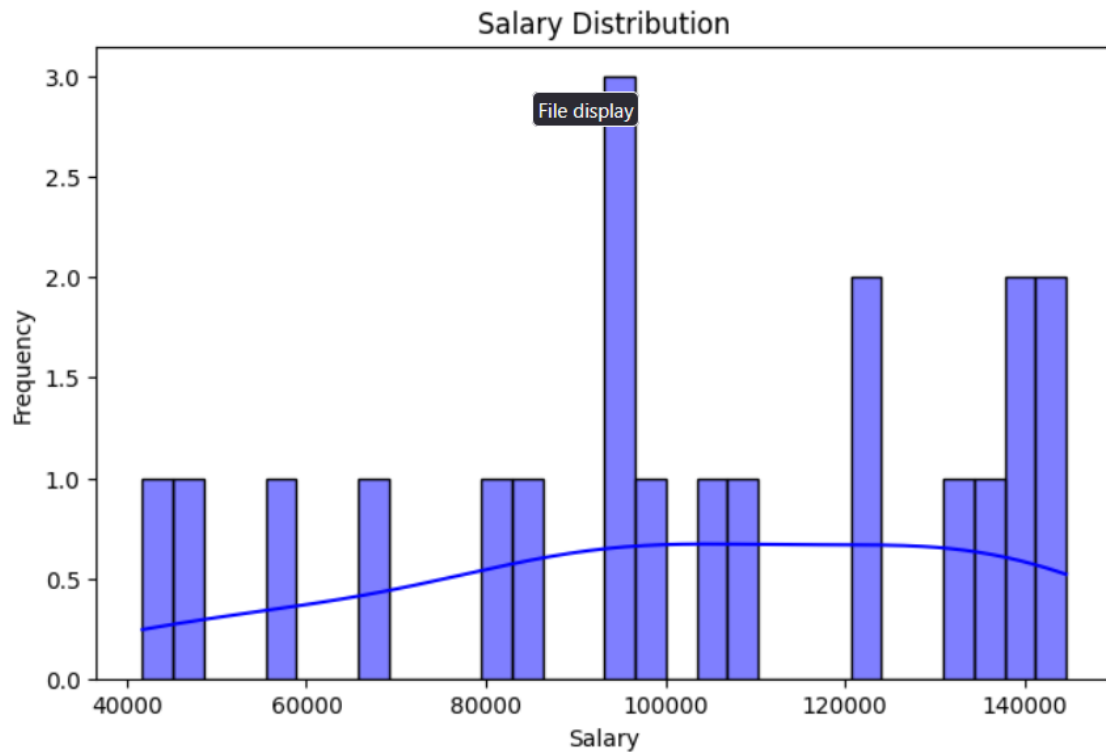
```
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(), annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Correlation Matrix")
plt.show()
```

Insights

- The salary distribution shows a skewed pattern, indicating a higher concentration in certain salary ranges.
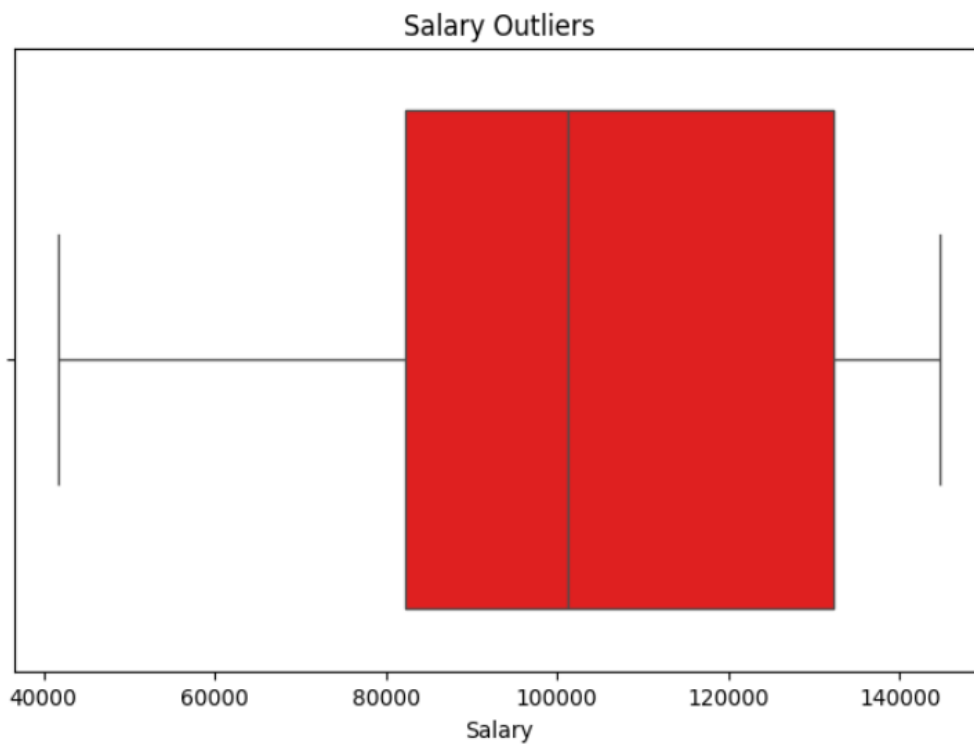
- Outliers in the salary data suggest significant variations, possibly due to high-level executive salaries.
- Correlation analysis helps understand relationships between salary and other factors like experience.

Outputs

Correlation Matrix

	EmployeeID	Age	Department	Experience	Salary
EmployeeID	1.00	0.25	0.51	0.33	0.32
Age	0.25	1.00	0.06	0.25	-0.25
Department	0.51	0.06	1.00	0.41	0.27
Experience	0.33	0.25	0.41	1.00	0.25
Salary	0.32	-0.25	0.27	0.25	1.00

[1]:
```
from google.colab import files
uploaded = files.upload()
```

Salary Outliers

Conclusion

This analysis provides valuable insights into employee salary trends, helping businesses understand pay structures and identify discrepancies.

References

1. Pandas Documentation: https://pandas.pydata.org/docs/
2. NumPy Documentation: https://numpy.org/doc/
3. Matplotlib Documentation: https://matplotlib.org/stable/contents.html
4. Seaborn Documentation: https://seaborn.pydata.org/