Exploratory Data Analysis (EDA) with Python

# Step 1: Create the Dataset File

First, create a CSV file named `employees.csv` with the following content:

Name, Age, Department, Salary, Years\_Experience

John, 28, Sales, 50000, 3

Anna, 24, Marketing, 52000, 2

Peter,35, HR, 58000, 7

Linda, 29, IT, 55000, 4

James, 41, Management, 72000, 15

Laura,36, IT, 62000, 10

Sophia, 34, Sales, 51000, 5

Tom, 50, HR,69000,25

Grace, 27, Marketing,54000,3

Michael, 39, Management,68000,12

# Step 2: Load and Analyze the Dataset

## Step 2.1: Import Necessary Libraries

pip install pandas

pip install seaborn

pip install matplotlib

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

## Step 2.2: Load the Dataset

df = pd.read\_csv(r"C:\Users\2272722\DA Projects\employee.csv")

print("First few rows of the dataset:")

print(df.head())

A screenshot of a computer

Description automatically generated

## Step 2.3: Understand the Data Structure

# Display basic information about the dataset

print("\nBasic information about the dataset:")

print(df.info())

A screenshot of a browser

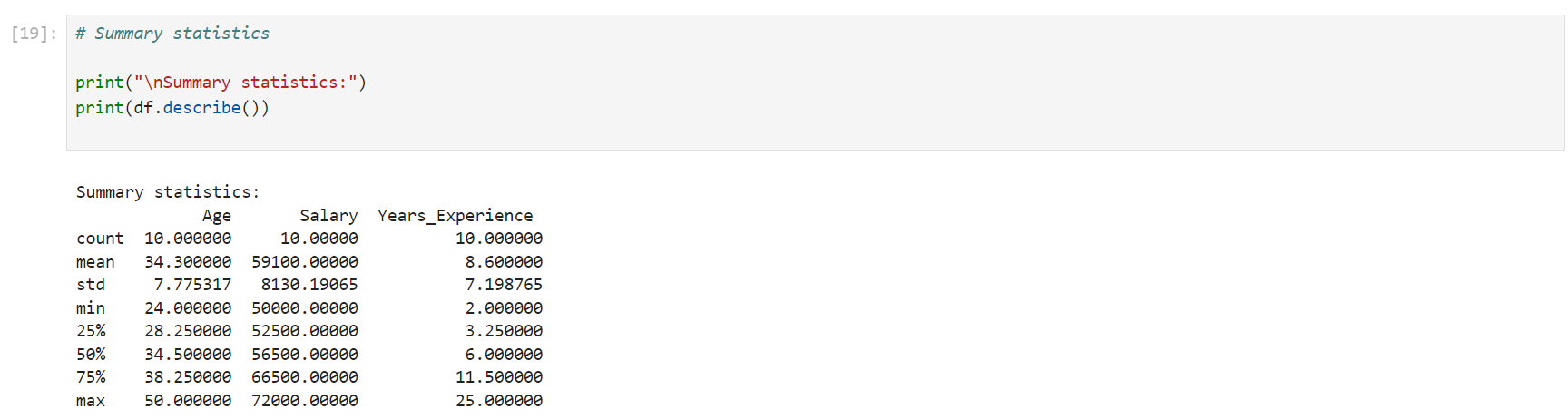
Description automatically generated

## Step 2.4: Summary Statistics

# Summary statistics

print("\nSummary statistics:")

print(df.describe())



## Step 2.5: Check for Missing Values

# Check for missing values

print("\nCheck for missing values:")

print(df.isnull().sum())

A white rectangular object with a white border

Description automatically generated

## Step 2.6: Data Visualization

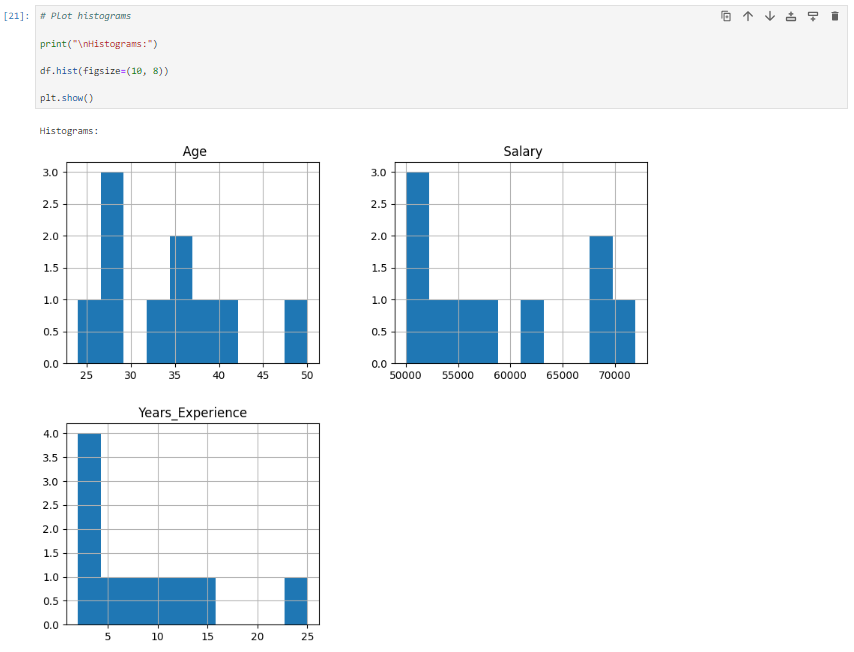
Histogram

# Plot histograms

print("\nHistograms:")

df.hist(figsize=(10, 8))

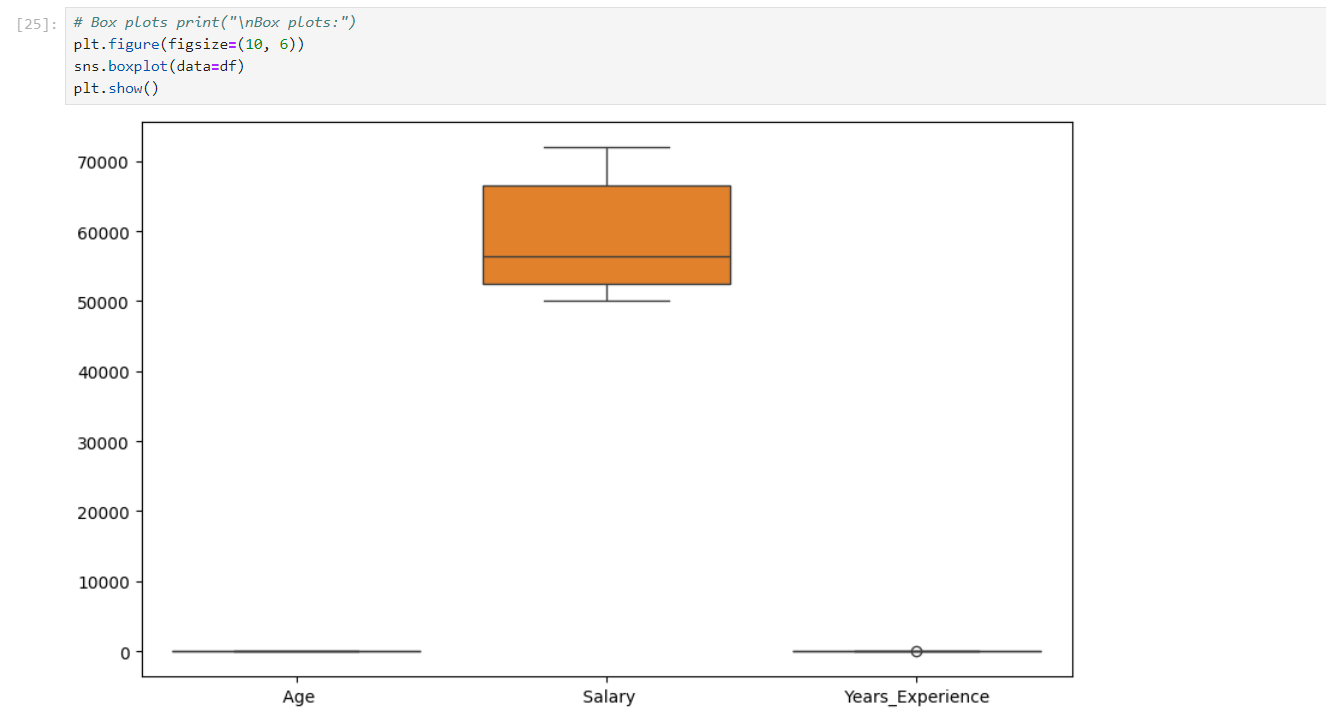
plt.show()



Box Plots

# Box plots

print("\nBox plots:")



Pair Plot

# Pair plot

print("\nPair plot:")

sns.pairplot(df)

plt.show()



A graph of a graph of a person

Description automatically generated with medium confidence

Heatmap

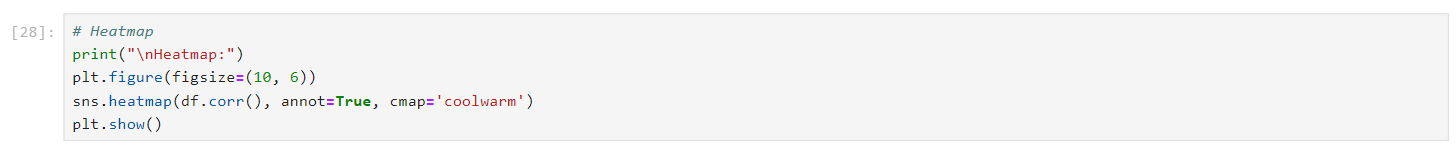
# Heatmap

print("\nHeatmap:")

plt.figure(figsize=(10, 6))

sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

plt.show()



A screenshot of a computer

Description automatically generated

## Step 2.7: Identify Patterns

Mean Values by Department

# Group by Department and calculate mean of each feature

print("\nMean values by Department:")

print(df.groupby('Department').mean())

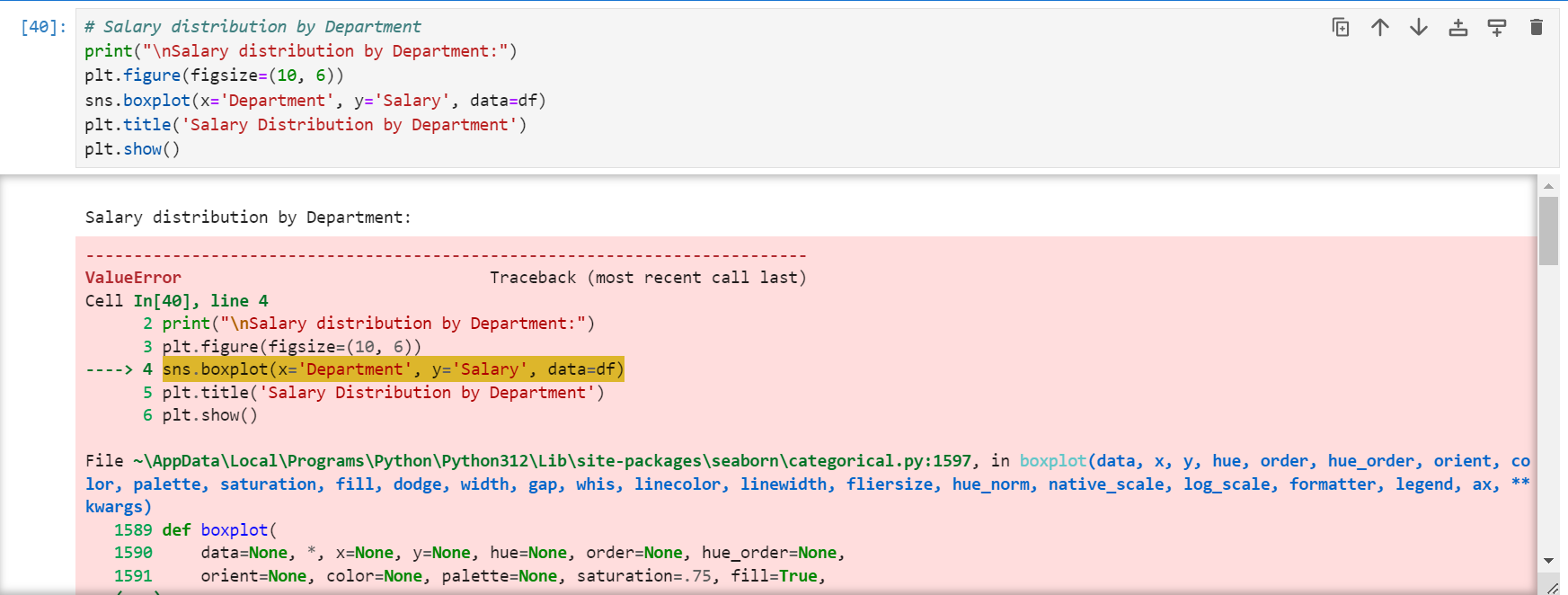
A screenshot of a computer

Description automatically generated

A pink background with green text

Description automatically generated

Salary Distribution by Department



A screenshot of a computer

Description automatically generated

Years of Experience Distribution by Department

# Years of Experience distribution by Department

print("\nYears of Experience distribution by Department:")

plt.figure(figsize=(10, 6))

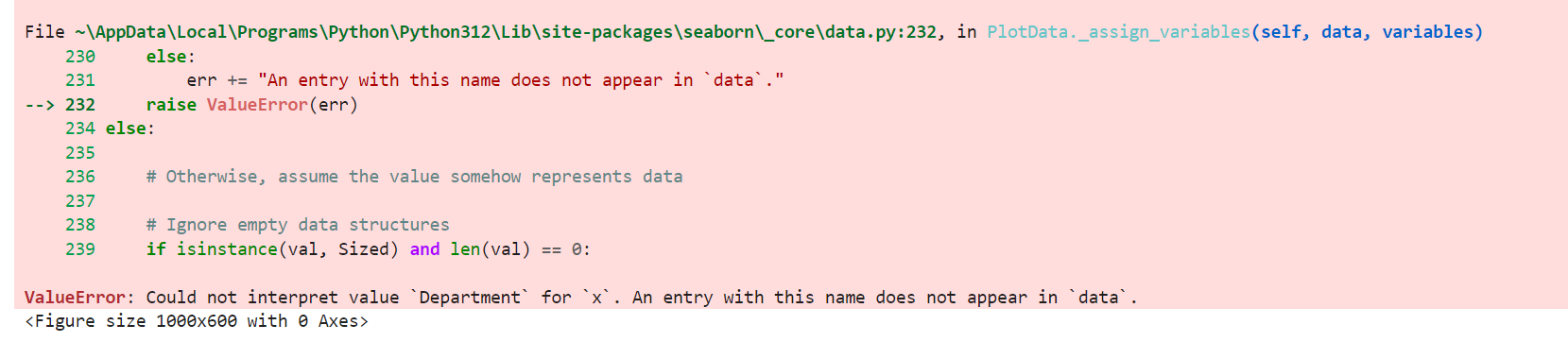
sns.boxplot(x='Department', y='Years\_Experience', data=df)

plt.title('Years of Experience Distribution by Department')

plt.show()

A screenshot of a computer

Description automatically generated



Salary vs. Years of Experience

# Salary vs Years of Experience

print("\nSalary vs Years of Experience:")

plt.figure(figsize=(10, 6))

sns.scatterplot(x='Years\_Experience', y='Salary', hue='Department', data=df)

plt.title('Salary vs Years of Experience')

plt.show()



A pink screen with text

Description automatically generated