Joshua Kyle K. Entrata

4CSC

Data Analysis and Visualization

In [1]:
```
import pandas as pd
import csv
```

Employee Dataset

In [2]:
```
employee_fns = ['Employee_ID', 'Name', 'Age', 'Department_ID']
```

In [3]:
```
employees = [
    {'Employee_ID': 101, 'Name': 'Alice', 'Age': 30, 'Department_ID': 'D001'},
    {'Employee_ID': 102, 'Name': 'Bob', 'Age': 35, 'Department_ID': 'D002'},
    {'Employee_ID': 103, 'Name': 'Charlie', 'Age': 28, 'Department_ID': 'D001'},
    {'Employee_ID': 104, 'Name': 'David', 'Age': 40, 'Department_ID': 'D003'},
    {'Employee_ID': 105, 'Name': 'Eve', 'Age': 45, 'Department_ID': 'D004'},
]
```

In [4]:
```
with open('employees.csv', 'w') as csvfile:
    writer = csv.DictWriter(csvfile, fieldnames=employee_fns)
    writer.writeheader()
    writer.writerows(employees)
```

Department Dataset

In [5]:
```
department_fns = ['Department_ID', 'Department_Name', 'Location']
```

In [6]:
```
departments = [
    {'Department_ID': 'D001', 'Department_Name': 'Sales', 'Location': 'New York'},
    {'Department_ID': 'D002', 'Department_Name': 'Marketing', 'Location': 'London'}
    {'Department_ID': 'D003', 'Department_Name': 'IT', 'Location': 'San Francisco'}
    {'Department_ID': 'D004', 'Department_Name': 'HR', 'Location': 'Singapore'},
]
```

In [7]:
```
with open('departments.csv', 'w') as csvfile:
    writer = csv.DictWriter(csvfile, fieldnames=department_fns)
    writer.writeheader()
    writer.writerows(departments)
```

1. Join the Data

> Merge the `employees.csv` and `deparments.csv` datasets using the `Department_ID` column. Show the combined dataset.

In [8]:
```
df_employees = pd.read_csv('employees.csv')
df_employees
```

Out[8]:

	Employee_ID	Name	Age	Department_ID
0	101	Alice	30	D001
1	102	Bob	35	D002
2	103	Charlie	28	D001
3	104	David	40	D003
4	105	Eve	45	D004

In [9]:
```
df_departments = pd.read_csv('departments.csv')
df_departments
```

Out[9]:

	Department_ID	Department_Name	Location
0	D001	Sales	New York
1	D002	Marketing	London
2	D003	IT	San Francisco
3	D004	HR	Singapore

In [10]:
```
df_merged = df_employees.merge(df_departments, on='Department_ID')
df_merged
```

Out[10]:

	Employee_ID	Name	Age	Department_ID	Department_Name	Location
0	101	Alice	30	D001	Sales	New York
1	102	Bob	35	D002	Marketing	London
2	103	Charlie	28	D001	Sales	New York
3	104	David	40	D003	IT	San Francisco
4	105	Eve	45	D004	HR	Singapore

2. Filter the Data

> From the merged dataset, extract a subset of employees who are older than 30 and work in New York or London.

In [11]:
```
older_than_30 = df_merged['Age'] > 30
in_new_york = df_merged['Location'] == 'New York'
in_london = df_merged['Location'] == 'London'

df_merged[(older_than_30) & ((in_new_york) | (in_london))]
```

Out[11]:

	Employee_ID	Name	Age	Department_ID	Department_Name	Location
1	102	Bob	35	D002	Marketing	London

3. Reshape the Data (Pivoting)

> Create a summary table that shows the count of employees in each department by location.

In [12]:
```
df_merged.pivot_table(
    index='Location',
    columns='Department_Name',
    values='Employee_ID',
    aggfunc='count',
    fill_value=0
)
```

Out[12]:

Department_Name	HR	IT	Marketing	Sales
Location				
London	0	0	1	0
New York	0	0	0	2
San Francisco	0	1	0	0
Singapore	1	0	0	0

4. Create a New Column

> Add a new column to the combined dataset that categorizes employees into age groups:

- "Young" if age < 35
- "Mid-aged" if age is between 35 and 45
- "Senior" if age > 45

In [13]:
```
df_merged.loc[df_merged['Age'] < 35, 'Age_Group'] = 'Young'
df_merged.loc[df_merged['Age'].between(35, 45), 'Age_Group'] = 'Mid-aged'
df_merged.loc[df_merged['Age'] > 45, 'Age_Group'] = 'Senior'
```

In [14]:
```
df_merged
```

Out[14]:

	Employee_ID	Name	Age	Department_ID	Department_Name	Location	Age_Group
0	101	Alice	30	D001	Sales	New York	Young
1	102	Bob	35	D002	Marketing	London	Mid-aged
2	103	Charlie	28	D001	Sales	New York	Young
3	104	David	40	D003	IT	San Francisco	Mid-aged
4	105	Eve	45	D004	HR	Singapore	Mid-aged