***Name: Vedashree Nimje***

**Intern ID: IP-6679**

**Cloud Counselage Internship (Python)**

**>> Uploading Dataset..**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

from google.colab import files

uploaded = files.upload()

df = pd.read\_excel(next(iter(uploaded)))

df.head()

***1. How many unique students are included in the dataset?***

unique\_students = df['Email ID'].nunique()

print("Unique students:", unique\_students)

Unique students: 2157

***2. What is the average GPA of the students?***

average\_gpa = df['CGPA'].mean()

print("Average GPA:", round(average\_gpa, 2))

Average GPA: 8.04

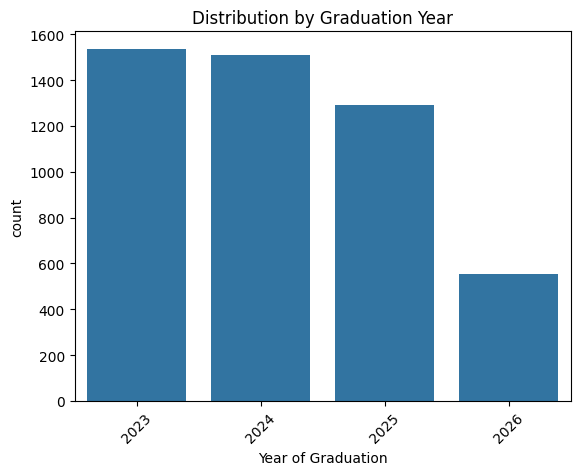
***3. What is the distribution of students across different graduation years?***

sns.countplot(data=df, x='Year of Graduation')

plt.title('Distribution by Graduation Year')

plt.xticks(rotation=45)

plt.show()



***4. What is the distribution of student’s experience with Python programming?***

plt.figure(figsize=(10,5))

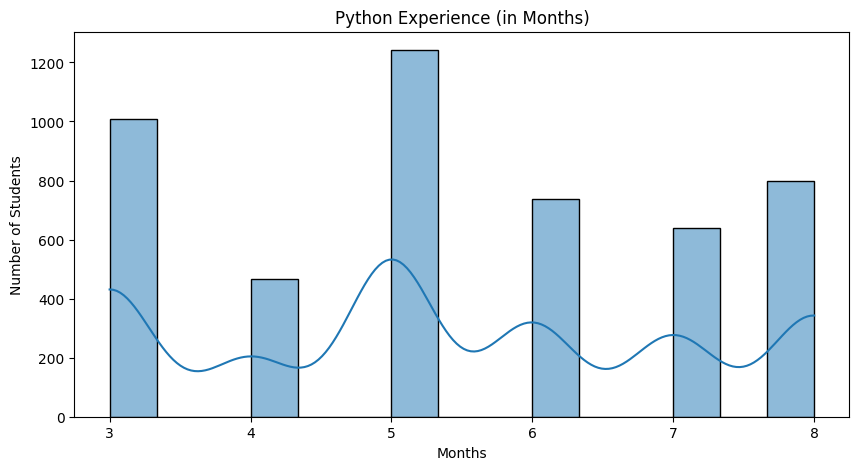
sns.histplot(df['Experience with python (Months)'], bins=15, kde=True)

plt.title('Python Experience (in Months)')

plt.xlabel('Months')

plt.ylabel('Number of Students')

plt.show()



***5. What is the average family income of the students?***

def convert\_income\_to\_numeric(income\_str):

    if isinstance(income\_str, str):

        income\_str = income\_str.replace('Lakh', '00000')

        income\_str = income\_str.replace('+', '')

        income\_str = income\_str.replace('-', '')

        incomes = income\_str.split(' ')

        total\_income = 0

        for i in incomes:

          if i.isdigit():

            total\_income = total\_income + int(i)

        return int(total\_income)

    else:

        return income\_str

df['Family Income'] = df['Family Income'].apply(convert\_income\_to\_numeric)

avg\_income = df['Family Income'].mean()

print("Average Family Income:", round(avg\_income, 2))

Average Family Income: 3.93

***6. How does the GPA vary among different colleges?***

top5\_colleges = df.groupby('College Name')['CGPA'].mean().sort\_values(ascending=False).head(5)

print("Top 5 Colleges by Average GPA:")

print(top5\_colleges)

Top 5 Colleges by Average GPA:

College Name

THAKUR INSTITUTE OF MANAGEMENT STUDIES, CAREER DEVELOPMENT & RESEARCH - [TIMSCDR] 8.585714

St Xavier's College 8.578571

B. K. Birla College of Arts, Science & Commerce (Autonomous), Kalyan 8.456410

Symbiosis Institute of Technology, Pune 8.303448

AP SHAH INSTITUTE OF TECHNOLOGY 8.283333

Name: CGPA, dtype: float64

***7. Are there any outliers in the ‘Attendee Status’ & ‘Quantity (courses completed)’ attribute?***

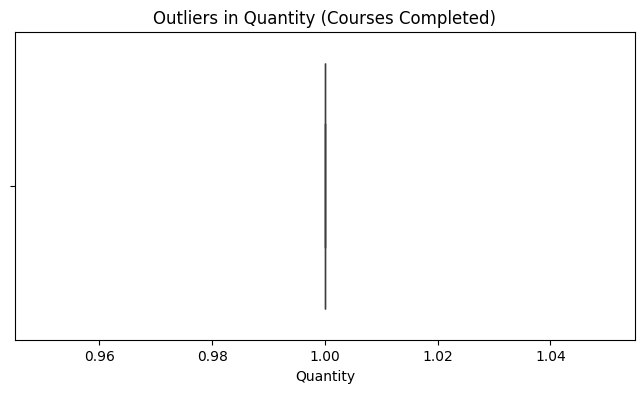
print("Attendee Status Values:\n", df['Attendee Status'].value\_counts(dropna=False))

plt.figure(figsize=(8,4))

sns.boxplot(x=df['Quantity'])

plt.title('Outliers in Quantity (Courses Completed)')

plt.show()



***8. What is the average GPA for students from each city?***

gpa\_by\_city = df.groupby('City')['CGPA'].mean().sort\_values(ascending=False)

print("Average GPA by City:")

print(gpa\_by\_city)

Average GPA by City:

City

Kolhapur 8.557143

Raipur 8.507143

Sonipat 8.464286

Gurugram 8.459259

Puri 8.450000

...

Panipat 7.615385

Nashik 7.592857

Daman 7.421429

Rewari 7.392308

New Delhi 7.307143

Name: CGPA, Length: 177, dtype: float64

***9. Can we identify any relationship between family income and GPA?***

plt.figure(figsize=(8,6))

sns.scatterplot(x='Family Income', y='CGPA', data=df)

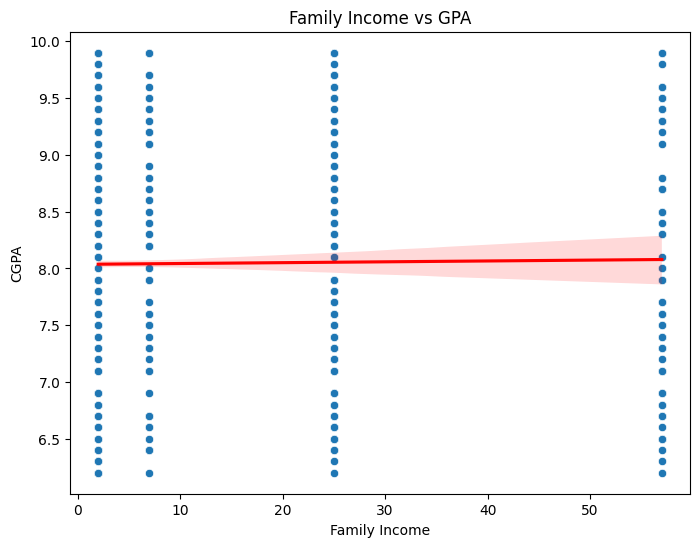
sns.regplot(x='Family Income', y='CGPA', data=df, scatter=False, color='red')

plt.title('Family Income vs GPA')

plt.xlabel('Family Income')

plt.ylabel('CGPA')

plt.show()



***10. How does the expected salary vary based on GPA, Family Income, Python Experience?***

# Clean the columns

df['CGPA'] = pd.to\_numeric(df['CGPA'], errors='coerce')

df['Family Income'] = pd.to\_numeric(df['Family Income'], errors='coerce')

df['Experience with python (Months)'] = pd.to\_numeric(df['Experience with python (Months)'], errors='coerce')

df['Expected salary (Lac)'] = pd.to\_numeric(df['Expected salary (Lac)'], errors='coerce')

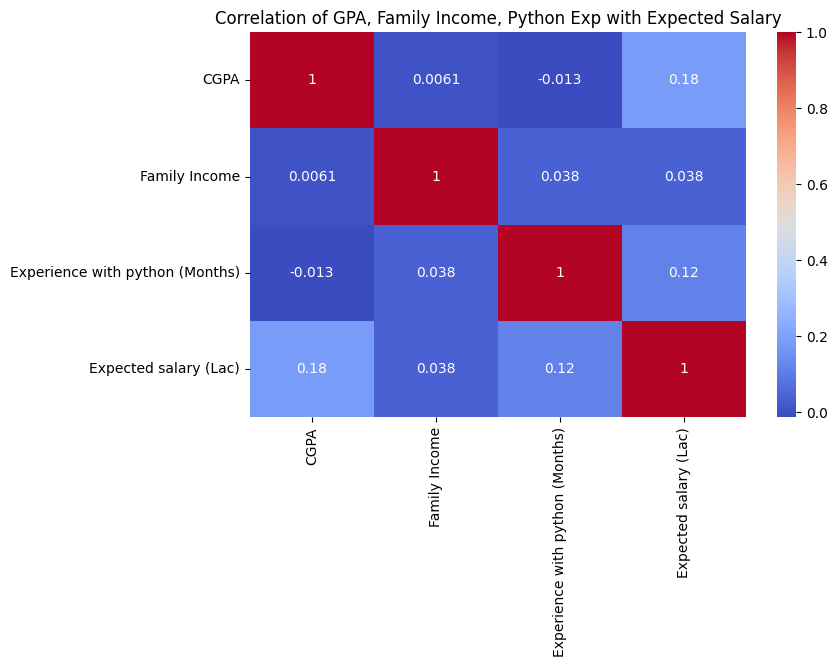
salary\_factors = df[['CGPA', 'Family Income', 'Experience with python (Months)', 'Expected salary (Lac)']].dropna()

plt.figure(figsize=(8, 5))

sns.heatmap(salary\_factors.corr(), annot=True, cmap='coolwarm')

plt.title("Correlation of GPA, Family Income, Python Exp with Expected Salary")

plt.show()



***11. Which event tends to attract more students from specific fields of study?***

event\_field = df.groupby(['College Name', 'Events']).size().reset\_index(name='Count')

top\_event\_field = event\_field.sort\_values(by='Count', ascending=False).head(10)

print("Top 10 event-field of study combos attracting students:")

print(top\_event\_field)

Top 10 event-field of study combos attracting students:

College Name \

390 priyadarshini college of engineering, nagpur

500 vidyalankar institute of technology, mumbai

308 ld college of engineering, ahmedabad, gujarat

166 b. k. birla college of arts, science & commerc...

260 government polytechnic gandhinagar

276 kle society's college of bca, rls institute, b...

532 wilson college

212 dkte society's textile and engineering institu...

516 vishwakarma institute of technology, pune

498 vidyalankar institute of technology, mumbai

Events Count

390 Product Design & Full Stack 56

500 Product Design & Full Stack 52

308 Product Design & Full Stack 52

166 Product Design & Full Stack 42

260 Product Design & Full Stack 42

276 Product Design & Full Stack 41

532 Product Design & Full Stack 39

212 Product Design & Full Stack 39

516 Product Design & Full Stack 39

498 Internship Program(IP) Success Conclave 39

***12. Do students in leadership positions tend to have higher GPAs or expected salaries?***

df\_filtered = df[['Leadership- skills', 'CGPA', 'Expected salary (Lac)']].dropna()

df\_filtered['Leadership- skills'] = df\_filtered['Leadership- skills'].map({'yes': 1, 'no': 0})

avg\_leader = df\_filtered['Leadership- skills'].mean()

leaders = df\_filtered[df\_filtered['Leadership- skills'] > avg\_leader]

print("Avg CGPA of leaders:", leaders['CGPA'].mean())

print("Avg expected salary of leaders:", leaders['Expected salary (Lac)'].mean())

Avg CGPA of leaders: 8.036787247087677

Avg expected salary of leaders: 13.971796443899448

***13. Is there a correlation between leadership skills and expected salary?***

df['Leadership- skills'] = pd.to\_numeric(df['Leadership- skills'], errors='coerce')

df['Expected salary (Lac)'] = pd.to\_numeric(df['Expected salary (Lac)'], errors='coerce')

leadership\_data = df[['Leadership- skills', 'Expected salary (Lac)']].dropna()

correlation = leadership\_data.corr().iloc[0,1]

print("Correlation between leadership skills and expected salary:", correlation)

Correlation between leadership skills and expected salary: nan

***14. How many students are graduating by the end of 2024?***

df['Year of Graduation'] = pd.to\_numeric(df['Year of Graduation'], errors='coerce')

students\_graduating\_2024 = df[df['Year of Graduation'] <= 2024]

print("Number of students graduating by the end of 2024:", students\_graduating\_2024.shape[0])

**Number of students graduating by the end of 2024: 3047**

***15. Which promotion channel brings in more student participation?***

channel\_counts = df['How did you come to know about this event?'].value\_counts()

print("Student participation by promotion channel:\n")

print(channel\_counts)

**Student participation by promotion channel:**

**How did you come to know about this event?**

**Whatsapp 1067**

**Email 438**

**SPOC/ College Professor 326**

**Others 153**

**Cloud Counselage Website 129**

**...**

**Others | SPOC/ College Professor 1**

**Youtube | Facebook | Whatsapp 1**

**LinkedIn | Whatsapp | Cloud Counselage Website 1**

**Instagram | Whatsapp | Email | Others 1**

**Youtube | Whatsapp | Cloud Counselage Website 1**

**Name: count, Length: 99, dtype: int64**

***16: Total number of students who attended Data Science related events***

df['Events'] = df['Events'].astype(str).str.lower()

data\_sci\_events = df[df['Events'].str.contains('data science', na=False)]

print("Total students who attended Data Science related events:", data\_sci\_events.shape[0])

Total students who attended Data Science related events: 306

***17: Those with high CGPA & more Python experience — what was their average expected salary?***

cgpa\_thresh = df['CGPA'].quantile(0.75)

python\_exp\_thresh = df['Experience with python (Months)'].quantile(0.75)

high\_performers = df[(df['CGPA'] >= cgpa\_thresh) & (df['Experience with python (Months)'] >= python\_exp\_thresh)]

print("Average expected salary for high CGPA + high Python experience:", high\_performers['Expected salary (Lac)'].mean())

Average expected salary for high CGPA + high Python experience: 20.64756446991404

END