Project Objective:

We aim to conduct a comprehensive analysis of our student interns to gain insight about the relationship between the factors influencing their success

In [1]:

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

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

executed in 1.86s, finished 15:33:51 2023-09-19

Import Dataset

In [2]:

```
df = pd.read_excel("D:\My study material\MSC DS\SEM-3\Internship\Problem_statement\Data a
df.head()
```

executed in 1.35s, finished 15:33:52 2023-09-19

Out[2]:

	First Name	Email ID	Quantity	Events	Attendee Status	College Name	How did you come to know about this event?	" (
0	ANIKET	aniket@xyz.com	1	Art of Resume Building	Attending	D Y PATIL INSTITUTE OF MCA AND MANAGEMENT AKUR...	Email	
1	Dhanshree	dhanshree@xyz.com	1	Art of Resume Building	Attending	AP SHAH INSTITUTE OF TECHNOLOGY	Others	
2	Dhiraj	dhiraj@xyz.com	1	Art of Resume Building	Attending	Don Bosco College of Engineering Fatorda Goa	Email	
3	Pooja	pooja@xyz.com	1	Art of Resume Building	Attending	Pillai College of Engineering New Panvel	Email	
4	Aayush	aayush@xyz.com	1	Art of Resume Building	Attending	St Xavier's College	Instagram | LinkedIn | Cloud Counselage Website	

In [3]:

```
# check null values
df.isnull().sum().sort_values(ascending=False)
```

executed in 27ms, finished 15:33:52 2023-09-19

Out[3]:

```
Specify in "Others" (how did you come to know about this event)    4805
How did you come to know about this event?                         2216
College Name                                                         15
First Name                                                            0
Email ID                                                              0
Quantity                                                              0
Events                                                                0
Attendee Status                                                       0
Designation                                                           0
Year of Graduation                                                    0
City                                                                  0
CGPA                                                                  0
Experience with python (Months)                                       0
Family Income                                                         0
Expected salary (Lac)                                                 0
Leadership- skills                                                    0
dtype: int64
```

In [4]:

```
df.info()
```

executed in 46ms, finished 15:33:52 2023-09-19

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4894 entries, 0 to 4893
Data columns (total 16 columns):
 #   Column                                                                  Non-
Null Count  Dtype
---  ------                                                                  ----
----------  -----
 0   First Name                                                              4894
non-null   object
 1   Email ID                                                                4894
non-null   object
 2   Quantity                                                                4894
non-null   int64
 3   Events                                                                  4894
non-null   object
 4   Attendee Status                                                         4894
non-null   object
 5   College Name                                                            4879
non-null   object
 6   How did you come to know about this event?                              2678
non-null   object
 7   Specify in "Others" (how did you come to know about this event)  89 n
on-null     object
 8   Designation                                                             4894
non-null   object
 9   Year of Graduation                                                      4894
non-null   int64
 10  City                                                                    4894
non-null   object
 11  CGPA                                                                    4894
non-null   float64
 12  Experience with python (Months)                                         4894
non-null   int64
 13  Family Income                                                           4894
non-null   object
 14  Expected salary (Lac)                                                   4894
non-null   int64
 15  Leadership- skills                                                      4894
non-null   object
dtypes: float64(1), int64(4), object(11)
memory usage: 611.9+ KB
```

In [5]:

```
# drop column which has many null values

drop_column = 'Specify in "Others" (how did you come to know about this event)'
df=df.drop(drop_column,axis=1)
df.head()
```

executed in 29ms, finished 15:33:52 2023-09-19

Out[5]:

	Email ID	Quantity	Events	Attendee Status	College Name	How did you come to know about this event?	Designation	Gra
	aniket@xyz.com	1	Art of Resume Building	Attending	D Y PATIL INSTITUTE OF MCA AND MANAGEMENT AKUR...	Email	Students	
	dhanshree@xyz.com	1	Art of Resume Building	Attending	AP SHAH INSTITUTE OF TECHNOLOGY	Others	Students	
	dhiraj@xyz.com	1	Art of Resume Building	Attending	Don Bosco College of Engineering Fatorda Goa	Email	Students	
	pooja@xyz.com	1	Art of Resume Building	Attending	Pillai College of Engineering New Panvel	Email	Students	
	aayush@xyz.com	1	Art of Resume Building	Attending	St Xavier's College	Instagram | LinkedIn | Cloud Counselage Website	Students	

In [6]:

```
drop_column = 'How did you come to know about this event?'
df=df.drop(drop_column,axis=1)
df.head()
```

executed in 22ms, finished 15:33:52 2023-09-19

Out[6]:

	First Name	Email ID	Quantity	Events	Attendee Status	College Name	Designation
0	ANIKET	aniket@xyz.com	1	Art of Resume Building	Attending	D Y PATIL INSTITUTE OF MCA AND MANAGEMENT AKUR...	Students
1	Dhanshree	dhanshree@xyz.com	1	Art of Resume Building	Attending	AP SHAH INSTITUTE OF TECHNOLOGY	Students
2	Dhiraj	dhiraj@xyz.com	1	Art of Resume Building	Attending	Don Bosco College of Engineering Fatorda Goa	Students
3	Pooja	pooja@xyz.com	1	Art of Resume Building	Attending	Pillai College of Engineering New Panvel	Students
4	Aayush	aayush@xyz.com	1	Art of Resume Building	Attending	St Xavier's College	Students

In [7]:

```
df.isnull().sum().sort_values(ascending=False)
```

executed in 23ms, finished 15:33:52 2023-09-19

Out[7]:

```
College Name                       15
First Name                          0
Email ID                            0
Quantity                            0
Events                              0
Attendee Status                     0
Designation                         0
Year of Graduation                  0
City                                0
CGPA                                0
Experience with python (Months)     0
Family Income                       0
Expected salary (Lac)               0
Leadership- skills                  0
dtype: int64
```

In [8]:

```
# describe data
df.describe()
```

executed in 31ms, finished 15:33:52 2023-09-19

Out[8]:

	Quantity	Year of Graduation	CGPA	Experience with python (Months)	Expected salary (Lac)
count	4894.0	4894.000000	4894.000000	4894.000000	4894.000000
mean	1.0	2024.176951	8.038476	5.395586	13.935635
std	0.0	1.000180	1.005184	1.705364	6.451959
min	1.0	2023.000000	6.200000	3.000000	5.000000
25%	1.0	2023.000000	7.200000	4.000000	8.000000
50%	1.0	2024.000000	7.900000	5.000000	13.000000
75%	1.0	2025.000000	8.900000	7.000000	19.000000
max	1.0	2026.000000	9.900000	8.000000	35.000000

In [9]:

```
df.columns
```

executed in 11ms, finished 15:33:52 2023-09-19

Out[9]:

```
Index(['First Name', 'Email ID', 'Quantity', 'Events', 'Attendee Status',
       'College Name', 'Designation', 'Year of Graduation', 'City', 'CGP
A',
       'Experience with python (Months)', 'Family Income',
       'Expected salary (Lac)', 'Leadership- skills'],
      dtype='object')
```

In [10]:

```
# change the columns name for better understanding
new_col_name = {'First Name':'First_Name',
            'Email ID':'Email_ID',
            'Attendee Status':'Attendee_Status',
            'College Name':'College_Name',
            'Year of Graduation':'Graduation_year',
            'Experience with python (Months)':'Month_of_exp_python',
            'Family Income':'Family_Income',
            'Expected salary (Lac)':'Expected_salary_lac',
            'Leadership- skills':'Leadership_skills'
           }

df.rename(columns=new_col_name, inplace=True)
```

executed in 10ms, finished 15:33:52 2023-09-19

In [11]:

```
df.columns
```

executed in 10ms, finished 15:33:52 2023-09-19

Out[11]:

```
Index(['First_Name', 'Email_ID', 'Quantity', 'Events', 'Attendee_Status',
       'College_Name', 'Designation', 'Graduation_year', 'City', 'CGPA',
       'Month_of_exp_python', 'Family_Income', 'Expected_salary_lac',
       'Leadership_skills'],
      dtype='object')
```

In [12]:

```
# check mode() of College_name column
df['College_Name'].mode()
```

executed in 12ms, finished 15:33:52 2023-09-19

Out[12]:

```
0    priyadarshini college of engineering, nagpur
Name: College_Name, dtype: object
```

In [13]:

```
# replace the null values with mode.
df['College_Name'].fillna(df['College_Name'].mode()[0], inplace=True)
```

executed in 11ms, finished 15:33:52 2023-09-19

In [14]:

```
# 0 dulicates value
df.duplicated().sum()
```

executed in 18ms, finished 15:33:52 2023-09-19

Out[14]:

```
0
```

In [15]:

```
# there is no null values remain
df.isnull().sum().sort_values(ascending=False)
```

executed in 20ms, finished 15:33:52 2023-09-19

Out[15]:

```
First_Name             0
Email_ID               0
Quantity               0
Events                 0
Attendee_Status        0
College_Name           0
Designation            0
Graduation_year        0
City                   0
CGPA                   0
Month_of_exp_python    0
Family_Income          0
Expected_salary_lac    0
Leadership_skills      0
dtype: int64
```

Data Analysis

Basic Questions:

1) How many unique students are included in the dataset?

In [16]:

```
df.shape
```

executed in 20ms, finished 15:33:52 2023-09-19

Out[16]:

```
(4894, 14)
```

In [17]:

```
no_of_uni_stud = len(df['First_Name'].unique())
no_of_uni_stud
```

executed in 15ms, finished 15:33:52 2023-09-19

Out[17]:

```
2324
```

CONCLUSION:

2324 unique students included in the dataset.there are same name of students present.

2) What is the average GPA of the students?

In [18]:

```
avg_gpa = sum(df['CGPA']) / len(df['CGPA'])
avg_gpa
```

executed in 13ms, finished 15:33:52 2023-09-19

Out[18]:

```
8.038475684511619
```

In [19]:

```
df['CGPA'].mean()
```

executed in 11ms, finished 15:33:52 2023-09-19

Out[19]:

```
8.038475684511647
```

CONCLUSION:

The average GPA of the students is 8.038475684511647.

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

In [20]:

```
# create data frame which contain count of graduation_year
year_dist = df['Graduation_year'].value_counts().sort_index()
df_2=pd.DataFrame(year_dist)
df_2
```

executed in 24ms, finished 15:33:52 2023-09-19

Out[20]:

	Graduation_year
2023	1536
2024	1511
2025	1292
2026	555

In [21]:

```
new_name={'Graduation_year':'student_passed_out'}
df_2.rename(columns=new_name, inplace=True)
df_2
```

executed in 14ms, finished 15:33:52 2023-09-19

Out[21]:

	student_passed_out
2023	1536
2024	1511
2025	1292
2026	555

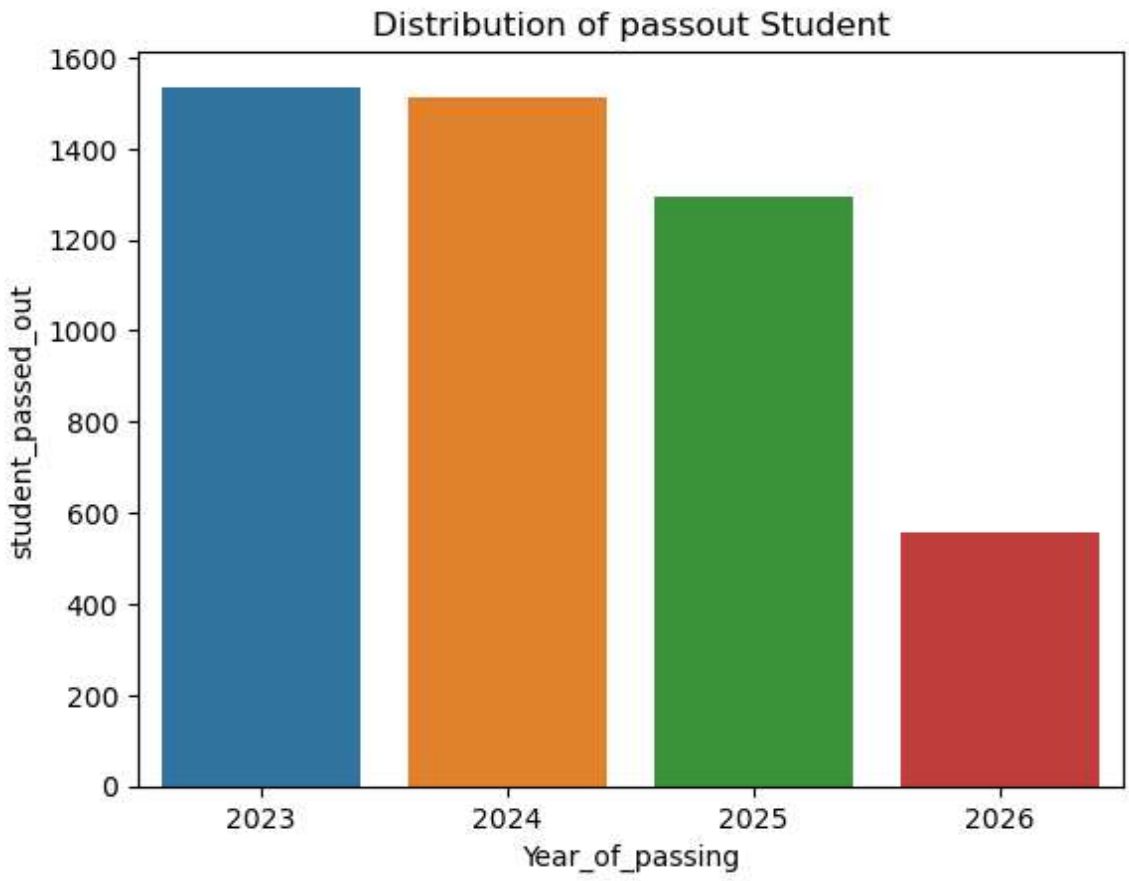
In [22]:

```
# Visualization
sns.barplot(x=df_2.index, y=df_2['student_passed_out'])
plt.xlabel('Year_of_passing')
plt.title('Distribution of passout Student')
```

executed in 287ms, finished 15:33:53 2023-09-19

Out[22]:

```
Text(0.5, 1.0, 'Distribution of passout Student')
```



CONCLUSION:

Most of students are passout in 2023-2024 year.

4) What is the distribution of student's experience with Python programming?

In [23]:

```
exp_dist = df['Month_of_exp_python'].value_counts().sort_index()
exp_dist
```

executed in 12ms, finished 15:33:53 2023-09-19

Out[23]:

```
3    1008
4     466
5    1242
6     738
7     640
8     800
Name: Month_of_exp_python, dtype: int64
```

In [24]:

```
df_3 = pd.DataFrame(exp_dist)
df_3
```

executed in 20ms, finished 15:33:53 2023-09-19

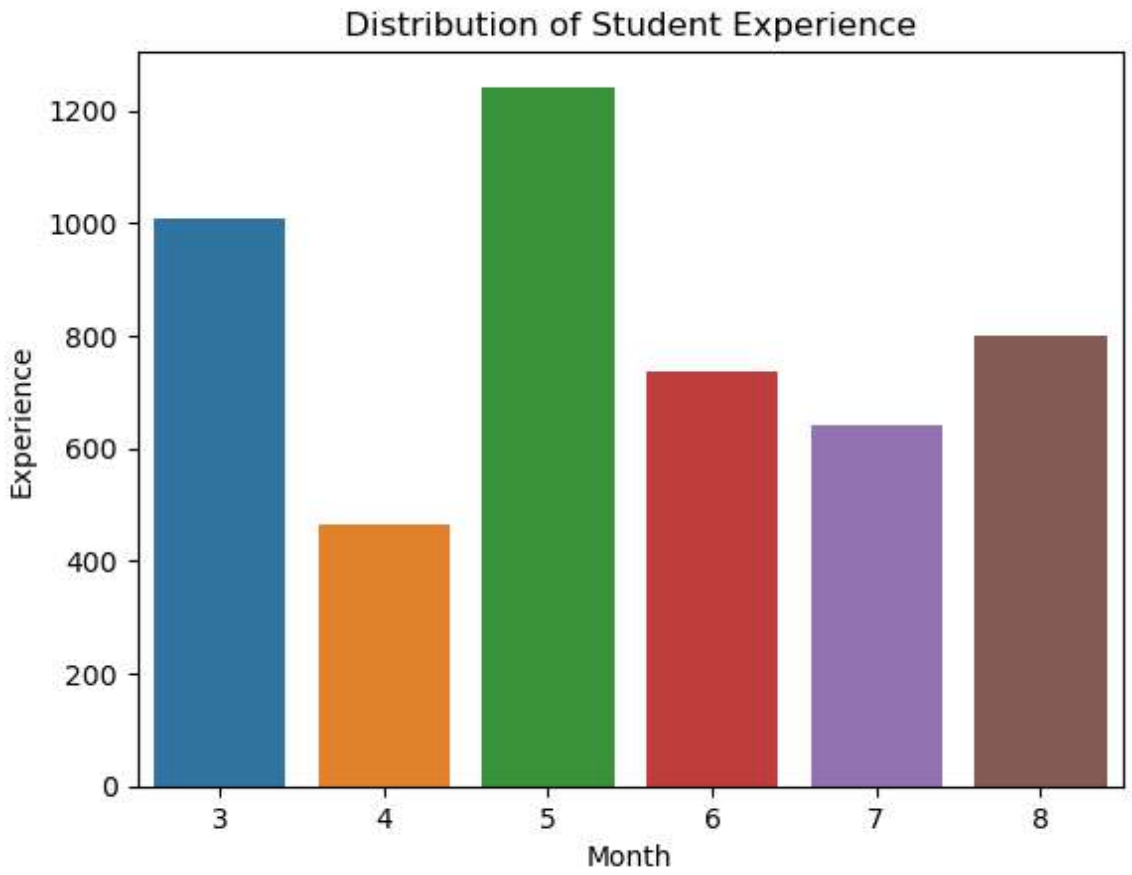
Out[24]:

	Month_of_exp_python
3	1008
4	466
5	1242
6	738
7	640
8	800

In [25]:

```
# Visualization
sns.barplot(x=df_3.index, y=df_3['Month_of_exp_python'])
plt.xlabel('Month')
plt.ylabel('Experience')
plt.title('Distribution of Student Experience')
plt.show()
```

executed in 186ms, finished 15:33:53 2023-09-19



CONCLUSION:

There is only 800 student have 8 years of experience, and most of student has experience of 3-5 years.

5) What is the average family income of the student?

In [26]:

```
df['Family_Income'].mode()
```

executed in 17ms, finished 15:33:53 2023-09-19

Out[26]:

```
0    0-2 Lakh
Name: Family_Income, dtype: object
```

In [27]:

```
df['Extracted'] = df['Family_Income'].str.split(' ').str[0]
```

executed in 23ms, finished 15:33:53 2023-09-19

In [28]:

```
df['Extracted']
```

executed in 26ms, finished 15:33:53 2023-09-19

Out[28]:

```
0         7
1       0-2
2       5-7
3       2-5
4       0-2
       ...
4889    0-2
4890    0-2
4891    0-2
4892    0-2
4893    0-2
Name: Extracted, Length: 4894, dtype: object
```

In [29]:

```
df['Family_Income'] = pd.to_numeric(df['Family_Income'], errors='coerce').astype('Int64')
```

executed in 29ms, finished 15:33:53 2023-09-19

In [30]:

```
df['Family_Income']=df['Extracted']
```

executed in 13ms, finished 15:33:53 2023-09-19

In [31]:

```
df.head()
```

executed in 16ms, finished 15:33:53 2023-09-19

Out[31]:

	First_Name	Email_ID	Quantity	Events	Attendee_Status	College_Name	Desig
0	ANIKET	aniket@xyz.com	1	Art of Resume Building	Attending	D Y PATIL INSTITUTE OF MCA AND MANAGEMENT AKUR...	St
1	Dhanshree	dhanshree@xyz.com	1	Art of Resume Building	Attending	AP SHAH INSTITUTE OF TECHNOLOGY	St
2	Dhiraj	dhiraj@xyz.com	1	Art of Resume Building	Attending	Don Bosco College of Engineering Fatorda Goa	St
3	Pooja	pooja@xyz.com	1	Art of Resume Building	Attending	Pillai College of Engineering New Panvel	St
4	Aayush	aayush@xyz.com	1	Art of Resume Building	Attending	St Xavier's College	St

In [32]:

```
df.info()
```

executed in 31ms, finished 15:33:53 2023-09-19

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4894 entries, 0 to 4893
Data columns (total 15 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   First_Name           4894 non-null   object 
 1   Email_ID             4894 non-null   object 
 2   Quantity             4894 non-null   int64  
 3   Events               4894 non-null   object 
 4   Attendee_Status      4894 non-null   object 
 5   College_Name         4894 non-null   object 
 6   Designation          4894 non-null   object 
 7   Graduation_year      4894 non-null   int64  
 8   City                 4894 non-null   object 
 9   CGPA                 4894 non-null   float64
 10  Month_of_exp_python  4894 non-null   int64  
 11  Family_Income        4894 non-null   object 
 12  Expected_salary_lac  4894 non-null   int64  
 13  Leadership_skills    4894 non-null   object 
 14  Extracted            4894 non-null   object 
dtypes: float64(1), int64(4), object(10)
memory usage: 573.6+ KB
```

In [33]:

```
df['Family_in_1'] = df['Family_Income'].str.split('').str[1]
df['Family_in_1']
```

executed in 28ms, finished 15:33:53 2023-09-19

Out[33]:

```
0       7
1       0
2       5
3       2
4       0
       ..
4889    0
4890    0
4891    0
4892    0
4893    0
Name: Family_in_1, Length: 4894, dtype: object
```

In [34]:

```
df['Family_in_2'] = df['Family_Income'].str.split('').str[3]
df['Family_in_2']
```

executed in 31ms, finished 15:33:53 2023-09-19

Out[34]:

```
0       NaN
1         2
2         7
3         5
4         2
       ...
4889      2
4890      2
4891      2
4892      2
4893      2
Name: Family_in_2, Length: 4894, dtype: object
```

In [35]:

```
# Convert the "object" column to "int"
df['Family_in_1'] = pd.to_numeric(df['Family_in_1'], errors='coerce').astype('Int64')
# Convert the "object" column to "int"
df['Family_in_2'] = pd.to_numeric(df['Family_in_2'], errors='coerce').astype('Int64')
```

executed in 32ms, finished 15:33:53 2023-09-19

In [36]:

```
df['Family_in_2'].fillna(0, inplace=True)
```

executed in 12ms, finished 15:33:53 2023-09-19

In [37]:

```
selected_col = ['Family_in_1','Family_in_2']

subset_FI = df[selected_col]
subset_FI.head()
```

executed in 14ms, finished 15:33:53 2023-09-19

Out[37]:

	Family_in_1	Family_in_2
0	7	0
1	0	2
2	5	7
3	2	5
4	0	2

In [38]:

```
subset_FI.info()
```

executed in 15ms, finished 15:33:53 2023-09-19

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4894 entries, 0 to 4893
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Family_in_1  4894 non-null   Int64
 1   Family_in_2  4894 non-null   Int64
dtypes: Int64(2)
memory usage: 86.2 KB
```

In [39]:

```
# Calculate the average of values from Family_in_1 and Family_in_2
avg = (df['Family_in_1'] + df['Family_in_2'] / 2)
avg
```

executed in 13ms, finished 15:33:53 2023-09-19

Out[39]:

```
0       7.0
1       1.0
2       8.5
3       4.5
4       1.0
       ...
4889    1.0
4890    1.0
4891    1.0
4892    1.0
4893    1.0
Length: 4894, dtype: Float64
```

In [40]:

```
df.drop(columns=['Family_Income'],inplace=True)
df.head()
```

executed in 31ms, finished 15:33:53 2023-09-19

Out[40]:

	Email_ID	Quantity	Events	Attendee_Status	College_Name	Designation	Graduation
	aniket@xyz.com	1	Art of Resume Building	Attending	D Y PATIL INSTITUTE OF MCA AND MANAGEMENT AKUR...	Students	
	hanshree@xyz.com	1	Art of Resume Building	Attending	AP SHAH INSTITUTE OF TECHNOLOGY	Students	
	dhiraj@xyz.com	1	Art of Resume Building	Attending	Don Bosco College of Engineering Fatorda Goa	Students	
	pooja@xyz.com	1	Art of Resume Building	Attending	Pillai College of Engineering New Panvel	Students	
	aayush@xyz.com	1	Art of Resume Building	Attending	St Xavier's College	Students	

In [41]:

```
df['FamiLy_Income']=pd.DataFrame(avg)
df.head()
```

executed in 30ms, finished 15:33:53 2023-09-19

Out[41]:

	First_Name	Email_ID	Quantity	Events	Attendee_Status	College_Name	Desig
0	ANIKET	aniket@xyz.com	1	Art of Resume Building	Attending	D Y PATIL INSTITUTE OF MCA AND MANAGEMENT AKUR...	St
1	Dhanshree	dhanshree@xyz.com	1	Art of Resume Building	Attending	AP SHAH INSTITUTE OF TECHNOLOGY	St
2	Dhiraj	dhiraj@xyz.com	1	Art of Resume Building	Attending	Don Bosco College of Engineering Fatorda Goa	St
3	Pooja	pooja@xyz.com	1	Art of Resume Building	Attending	Pillai College of Engineering New Panvel	St
4	Aayush	aayush@xyz.com	1	Art of Resume Building	Attending	St Xavier's College	St

In [42]:

```
df['FamiLy_Income'].mean()
```

executed in 12ms, finished 15:33:53 2023-09-19

Out[42]:

```
1.374744585206375
```

CONCLUSION:

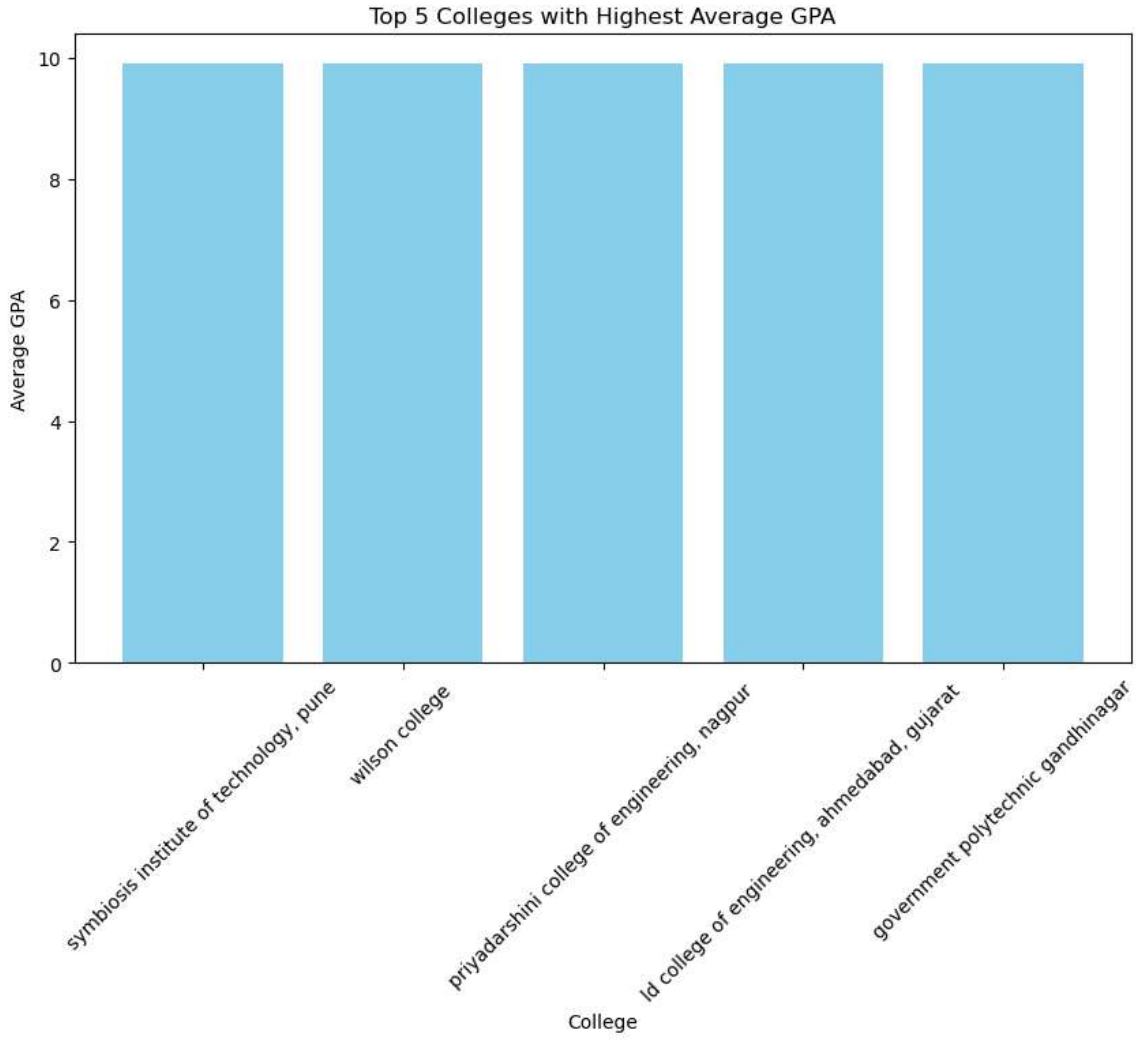
Average Family Income is 1.374744585206375.

6) How does the GPA vary among different colleges? (Show top 5 results only)

In [43]:

```
df_4 = df.sort_values(by='CGPA', ascending=False)

# Get the top 5 colleges
top_5 = df_4.head(5)

# Create a bar chart to visualize the top 5 colleges with the highest average GPA
plt.figure(figsize=(10, 6))
plt.bar(top_5['College_Name'], top_5['CGPA'], color='skyblue')
plt.xlabel('College')
plt.ylabel('Average GPA')
plt.title('Top 5 Colleges with Highest Average GPA')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability

plt.show()
```

executed in 220ms, finished 15:33:54 2023-09-19



In [44]:

```
college_gpa = df.groupby('College_Name')['CGPA'].mean()

# Sort the results in descending order
college_gpa_sorted = college_gpa.sort_values(ascending=False)
# Display the top 5 colleges with the highest average GPA
top_5_colleges = college_gpa_sorted.head(5)
print(top_5_colleges)
```

executed in 22ms, finished 15:33:54 2023-09-19

```
College_Name
THAKUR INSTITUTE OF MANAGEMENT STUDIES, CAREER DEVELOPMENT & RESEARCH - [T
IMSCDR]    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
```

CONCLUSION:

- Top 5 colleges by CGPA:

  > 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

In [45]:

```
df.drop(columns=['Extracted', 'Family_in_1', 'Family_in_2'], inplace=True)
```

executed in 14ms, finished 15:33:54 2023-09-19

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

In [90]:

```
city_gpa = df.groupby('City')['CGPA'].mean()

# Sort the results in descending order
city_gpa_sorted = city_gpa.sort_values(ascending=False)
# Display the top 5 cities with the highest average GPA
top_5_city = pd.DataFrame(city_gpa_sorted.head())

print(top_5_city)
```

executed in 17ms, finished 16:29:56 2023-09-19

```
              CGPA
City
Kolhapur  8.557143
Raipur    8.507143
Sonipat   8.464286
Gurugram  8.459259
Puri      8.450000
```

In [91]:

```
top_5_city = top_5_city.reset_index(drop=True)
```

executed in 9ms, finished 16:30:00 2023-09-19

In [92]:

```
top_5_city['city']=['Kolhapur','Raipur','Sonipat','Gurugram','Puri']
top_5_city
```

executed in 19ms, finished 16:30:01 2023-09-19

Out[92]:

	CGPA	city
0	8.557143	Kolhapur
1	8.507143	Raipur
2	8.464286	Sonipat
3	8.459259	Gurugram
4	8.450000	Puri

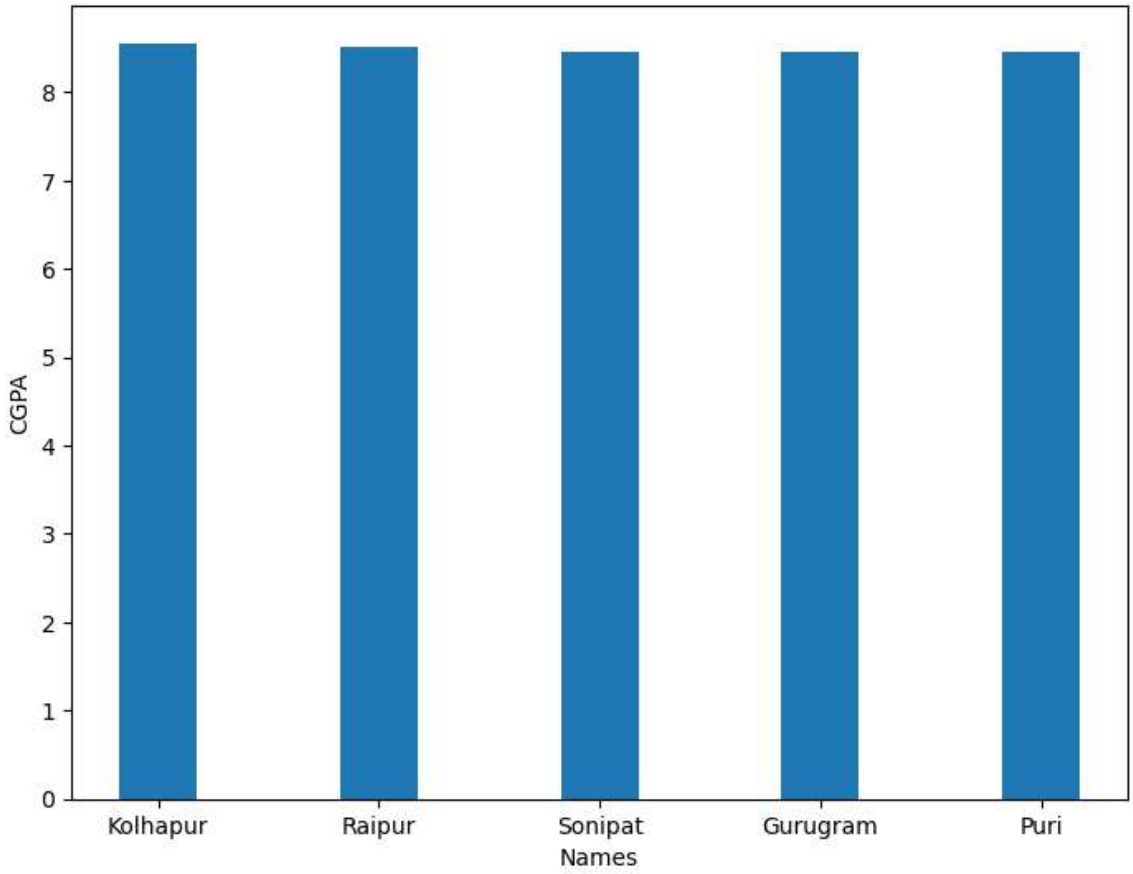
In [93]:

```
fig = plt.figure()
width = 0.35
ax = fig.add_axes([1,1,1,1])
ax.bar(top_5_city['city'],top_5_city['CGPA'],width)
ax.set_xlabel('Names')
ax.set_ylabel('CGPA')
```

executed in 249ms, finished 16:30:04 2023-09-19

Out[93]:

```
Text(0, 0.5, 'CGPA')
```



CONCLUSION:

Here students of Kolhapur, Raipur, Sonipat, Gurugram and Puri's top 5 cities has CGPA greater than 8. students of that cities they are smart or bright in studies.

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

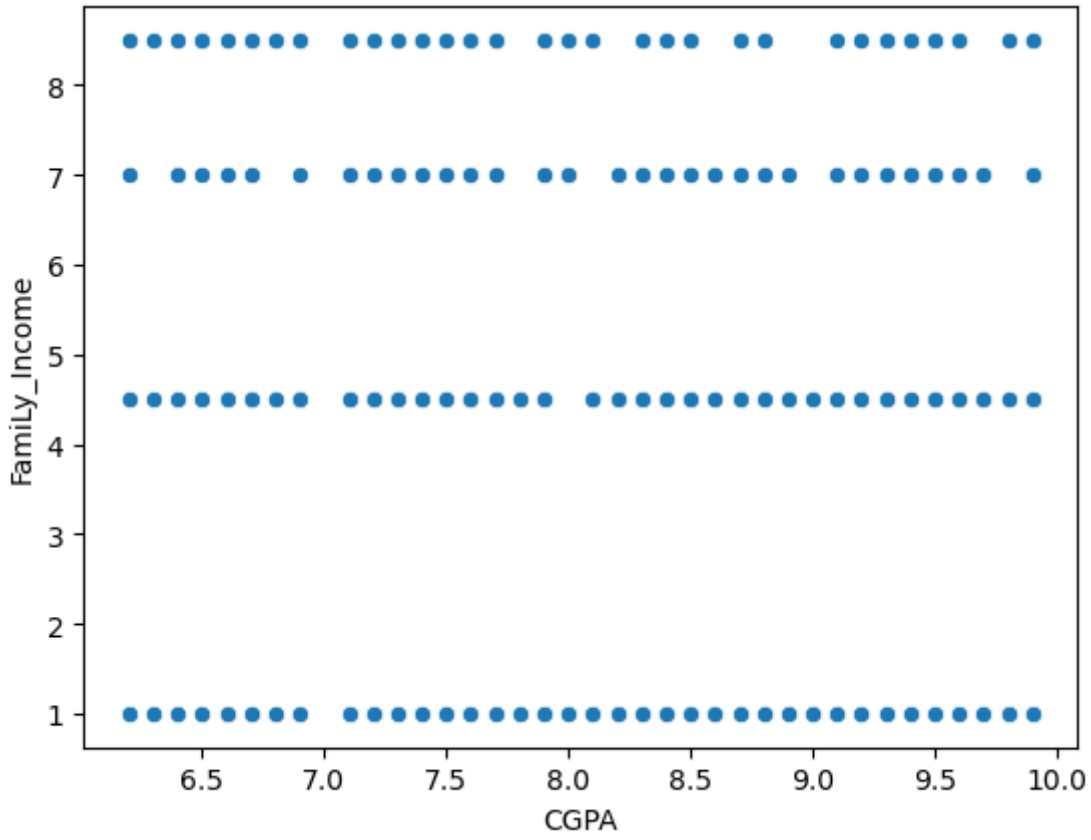
In [94]:

```
sns.scatterplot(x=df['CGPA'], y=df['FamiLy_Income'])
```

executed in 328ms, finished 16:33:20 2023-09-19

Out[94]:

```
<Axes: xlabel='CGPA', ylabel='FamiLy_Income'>
```



CONCLUSION:

After observe the graph we can say there is no relation between family income and CGPA. Means students who hasn't strong finance background they also bright in education.

Moderated Quenstions:

10) How many students from various cities?(Solve using data visualisation tool).

In [95]:

```
stud_city = df.groupby('First_Name')['City'].count()
stud_city
```

executed in 30ms, finished 16:37:57 2023-09-19

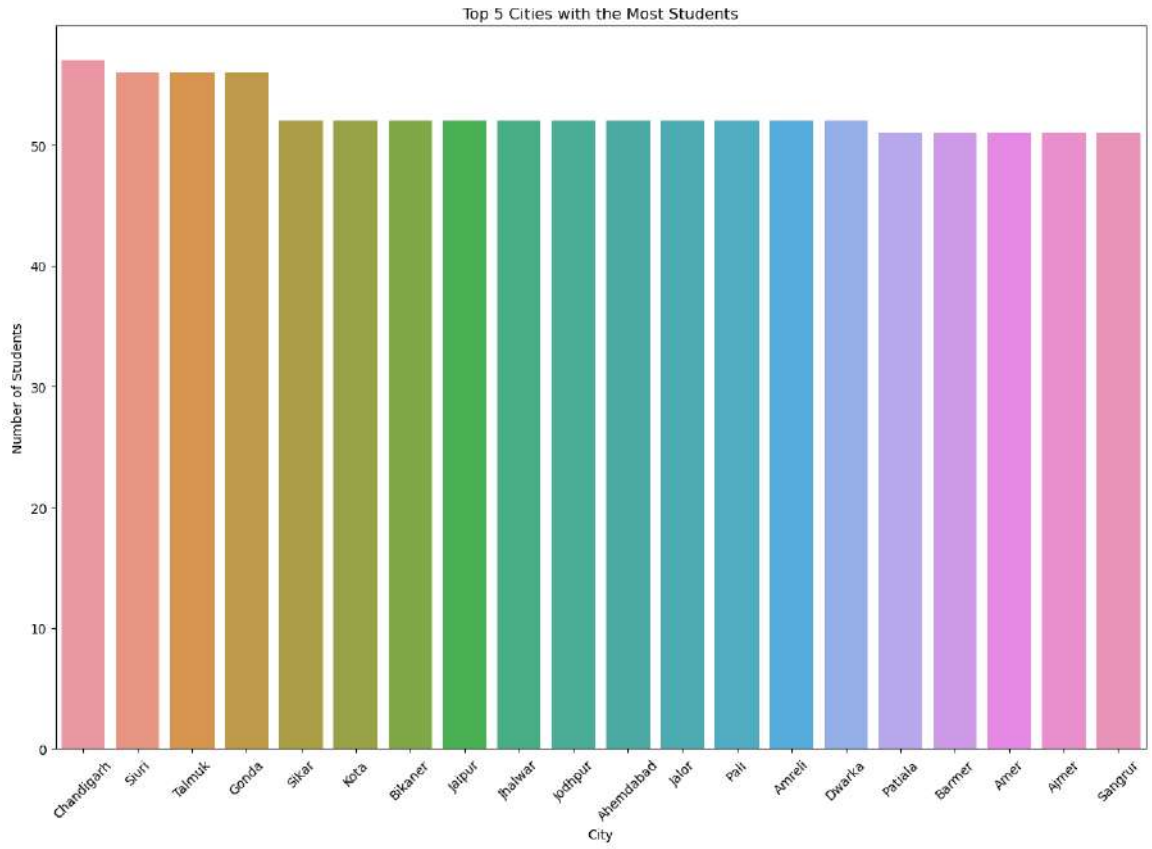
Out[95]:

```
First_Name
10 Pawan        1
AARAV           4
ABA             1
ABHINAVKUMAR    1
ABHITA          1
               ..
vraj            2
yash            1
yugesh          1
zeba            1
zobia           1
Name: City, Length: 2324, dtype: int64
```

In [105]:

```
city_counts = df['City'].value_counts().head(20)

# Create a countplot to visualize the number of students from the top 5 cities
plt.figure(figsize=(15, 10))
sns.barplot(x=city_counts.index, y=city_counts.values)
plt.xlabel('City')
plt.ylabel('Number of Students')
plt.title('Top 5 Cities with the Most Students')

# Display the countplot
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability

plt.show()
```

executed in 451ms, finished 16:45:03 2023-09-19



CONCLUSION:

In above Chart we can see more than 45 number of students comes from Chandigarh, siuri, Talmuk and so on city.

11) How does the expected salary vary based on factors like 'GPA', 'Family_income', 'Experience with python(month)'?

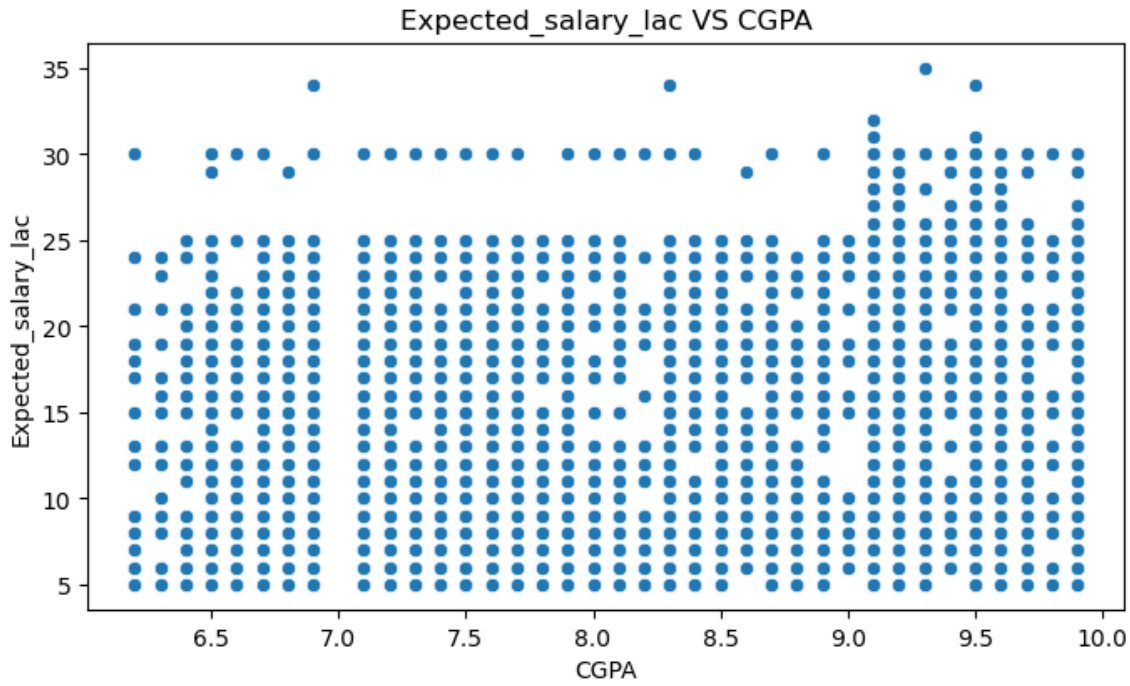
In [106]:

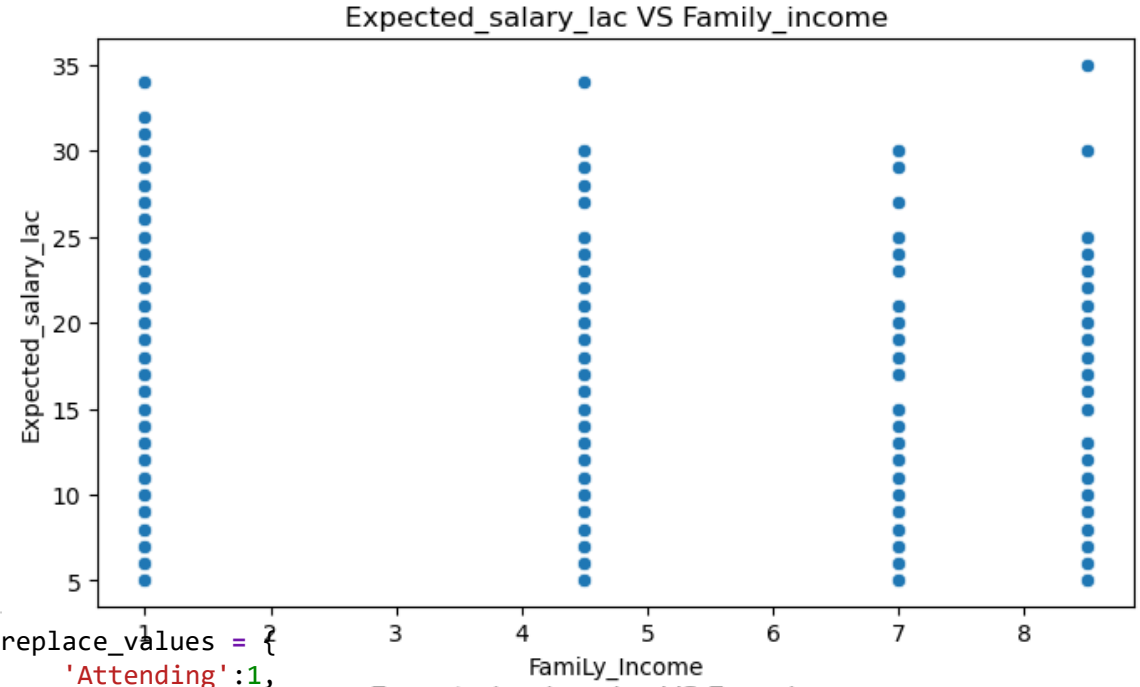
```
plt.figure(figsize=(8,15))

plt.subplot(3,1,1)
sns.scatterplot(x=df['CGPA'], y=df['Expected_salary_lac'])
plt.title('Expected_salary_lac VS CGPA')

plt.subplot(3,1,2)
sns.scatterplot(x=df['FamiLy_Income'], y=df['Expected_salary_lac'])
plt.title('Expected_salary_lac VS Family_income')

plt.subplot(3,1,3)
sns.scatterplot(x=df['Month_of_exp_python'], y=df['Expected_salary_lac'])
plt.title('Expected_salary_lac VS Experience')
```

executed in 868ms, finished 16:51:37 2023-09-19

Out[106]:

```
Text(0.5, 1.0, 'Expected_salary_lac VS Experience')
```

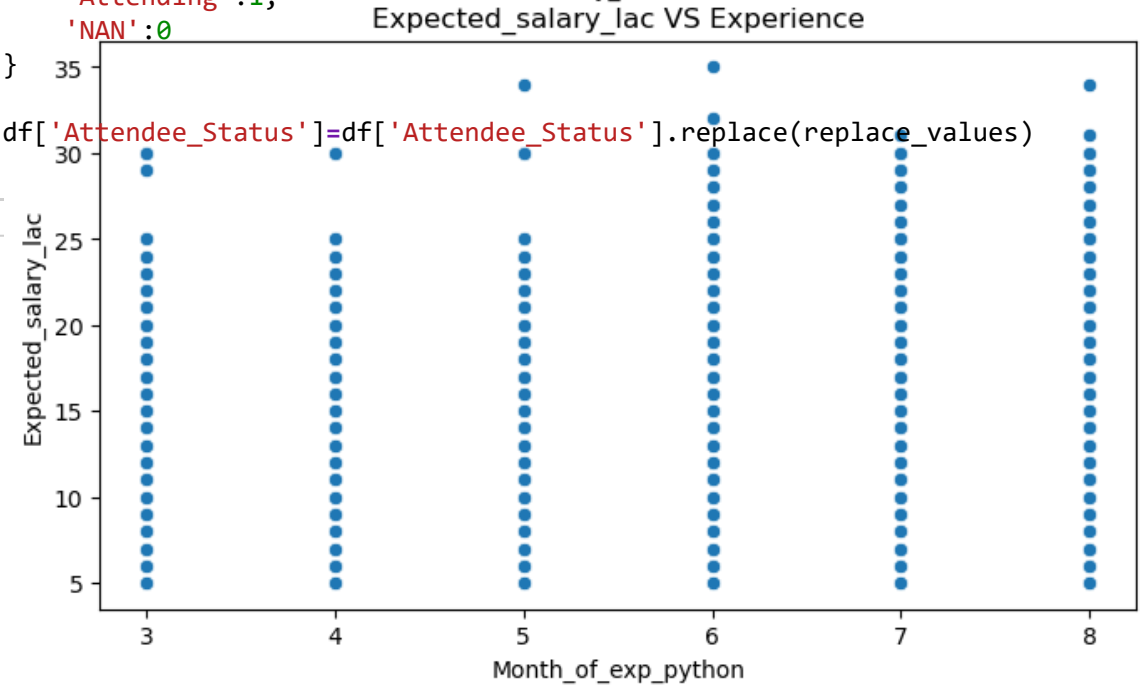




PA',

f study?

```
replace_values = {
    'Attending':1,
    'NAN':0
}

df['Attendee_Status']=df['Attendee_Status'].replace(replace_values)
```



In [108]:

```
event_attend = df.groupby(['Events','Designation'])['Attendee_Status'].sum().sort_values(
event_attend
```

executed in 31ms, finished 16:58:17 2023-09-19

Out[108]:

```
Events                                          Designation
Product Design & Full Stack                     Students       841
Internship Program(IP) Success Conclave         Students       545
Art of Resume Building                          Students       472
Data Visualization using Power BI               Students       451
Talk on Skill and Employability Enhancement     Students       379
Name: Attendee_Status, dtype: int64
```

CONCLUSION:

- There are top 5 events tend to attact more students. Events Designation

Product Design & Full Stack: Students - 841

Internship Program(IP) Success Conclave: Students - 545

Art of Resume Building: Students - 472

Data Visualization using Power BI: Students - 451

Talk on Skill and Employability Enhancement: Students - 379

13) Do students in leadership position during their college years tend to have higher GPAs or better expected salary?

In [109]:

```
df['Leadership_skills'] = df['Leadership_skills'].replace({'no ':'no'})
```

executed in 14ms, finished 16:58:23 2023-09-19

In [110]:

```
df['Leadership_skills'].unique()
```

executed in 17ms, finished 16:58:25 2023-09-19

Out[110]:

```
array(['yes', 'no'], dtype=object)
```

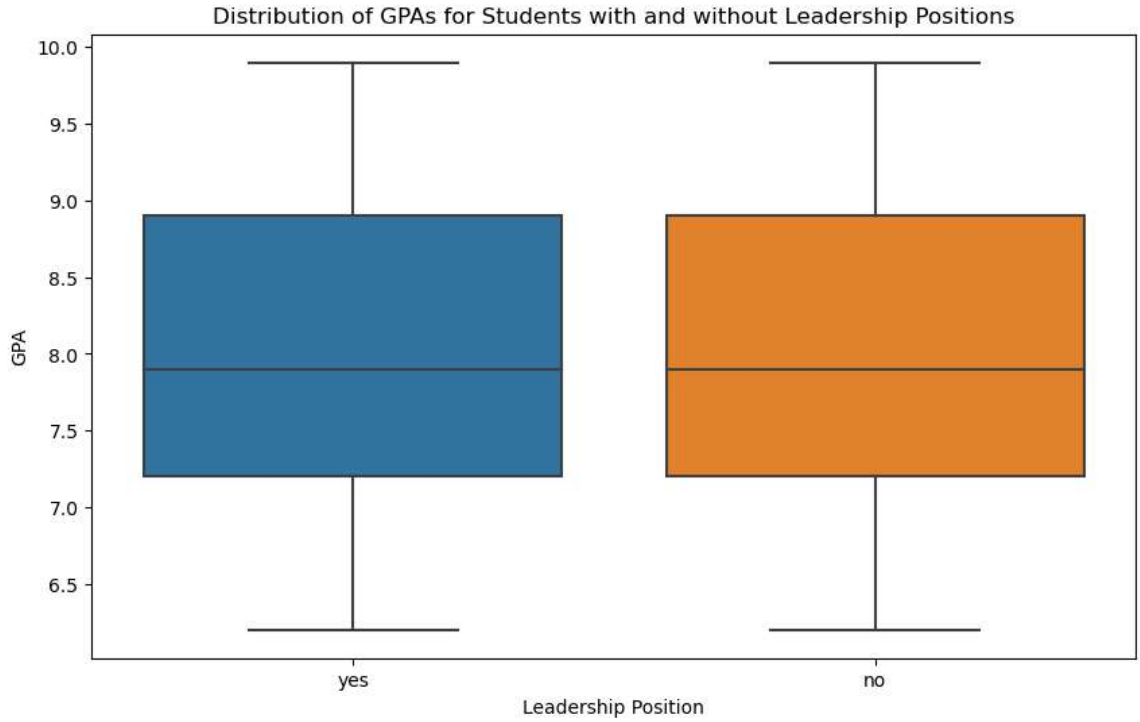
In [111]:

```
# Visualization
# Create box plots to visualize the distribution of GPAs
plt.figure(figsize=(10, 6))
sns.boxplot(x='Leadership_skills', y='CGPA', data=df)
plt.xlabel('Leadership Position')
plt.ylabel('GPA')
plt.title('Distribution of GPAs for Students with and without Leadership Positions')
```

executed in 315ms, finished 16:58:27 2023-09-19

Out[111]:

```
Text(0.5, 1.0, 'Distribution of GPAs for Students with and without Leaders
hip Positions')
```



In [112]:

```
leadership_students = df[df['Leadership_skills'] == 'yes']
non_leadership_students = df[df['Leadership_skills'] == 'no']
```

executed in 16ms, finished 16:58:28 2023-09-19

In [113]:

```
from scipy import stats
t_statistic, p_value = stats.ttest_ind(leadership_students['CGPA'], non_leadership_studer
```

executed in 18ms, finished 16:58:29 2023-09-19

In [114]:

```
# Display t-test results
print(f'T-Statistic: {t_statistic:.2f}')
print(f'P-Value: {p_value:.4f}')
```

executed in 20ms, finished 16:58:30 2023-09-19

```
T-Statistic: -0.08
P-Value: 0.9338
```

CONCLUSION:

there is no significant, means students in leadership position during their college years is not tend to have higher GPAs or better expected salary.

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

In [115]:

```
graduated_by_2024 = df[df['Graduation_year'] <= 2024]

# Get the count of students who graduated by the end of 2024
count_students_graduated_by_2024 = len(graduated_by_2024)
count_students_graduated_by_2024
```

executed in 17ms, finished 16:58:33 2023-09-19

Out[115]:

```
3047
```

CONCLUSION:

3047 students are graduating by the end of 2024.

15) Which promotion channel brings in more student participations for the event?

In [116]:

```
event_attend = df.groupby(['Events','Designation'])['Attendee_Status'].sum().sort_values(
event_attend
```

executed in 18ms, finished 16:58:39 2023-09-19

Out[116]:

```
Events                       Designation
Product Design & Full Stack  Students       841
Name: Attendee_Status, dtype: int64
```

CONCLUSION:

In Product Design & Full Stack event more student participated.

16) Find the total number of students who attended the events related to data science?(from all data science related course

In [117]:

```
df['Events'].unique()
```

executed in 27ms, finished 16:58:42 2023-09-19

Out[117]:

```
array(['Art of Resume Building', 'Data Visualization using Power BI',
       'Artificial Intelligence', 'Hello ML and DL', 'Product Marketing',
       'IAC - Q&A', 'Internship Program(IP) Success Conclave',
       'IS DATA SCIENCE FOR YOU?', 'KYC - Know Your CCPC',
       'Product Design & Full Stack', 'RPA: A Boon or A Bane',
       'Skill and Employability Enhancement',
       'Talk on Skill and Employability Enhancement',
       'The Agile Ways of Working', 'The SDLC & their transformations',
       'Transformation with DevOps: The Easy Way'], dtype=object)
```

In [118]:

```
data_science_events = ['Data Visualization using Power BI',
                       'Hello ML and DL',
                       'IS DATA SCIENCE FOR YOU?'
                       ]

ds_events = df[df['Events'].isin(data_science_events)]
ds_events.head()
```

executed in 32ms, finished 16:58:43 2023-09-19

Out[118]:

	First_Name	Email_ID	Quantity	Events	Attendee_Status	College_Name	D
341	Esha	esha@xyz.com	1	Data Visualization using Power BI	1	na	
342	Akash	akash@xyz.com	1	Data Visualization using Power BI	1	ld college of engineering, ahmedabad, gujarat	
343	Shubham	shubham@xyz.com	1	Data Visualization using Power BI	1	dkte society's textile and engineering institu...	
344	Aniket	aniket@xyz.com	1	Data Visualization using Power BI	1	lokmanya tilak college of engineering koparkha...	
345	Sakshi	sakshi@xyz.com	1	Data Visualization using Power BI	1	vidyalankar institute of technology, mumbai	

In [119]:

```
event_stud=ds_events.loc[:,['Events','Designation']]
event_stud
```

executed in 17ms, finished 16:58:44 2023-09-19

Out[119]:

	Events	Designation
341	Data Visualization using Power BI	Students
342	Data Visualization using Power BI	Students
343	Data Visualization using Power BI	Students
344	Data Visualization using Power BI	Students
345	Data Visualization using Power BI	Students
...	...	...
4858	Data Visualization using Power BI	Students
4890	Data Visualization using Power BI	Students
4891	Data Visualization using Power BI	Students
4892	Data Visualization using Power BI	Students
4893	Data Visualization using Power BI	Students

1023 rows × 2 columns

In [120]:

```
specific_value = 'Students'
selected_rows = event_stud.loc[event_stud['Designation'] == specific_value, ['Events', 'D
total_events=selected_rows.groupby(['Events'])['Designation'].value_counts()
total_events.sum()
```

executed in 19ms, finished 16:58:45 2023-09-19

Out[120]:

```
1020
```

CONCLUSION:

'Data Visualization using Power BI','Hello ML and DL','IS DATA SCIENCE FOR YOU?' all this events are related to Data Science and total 1020 students attend the events.

17) Those who have high CGPA & More experience in language those who had high expectation for salary?(Avg)

In [121]:

```
# Define thresholds for high CGPA and more programming experience
high_cgpa_threshold = 6.0  # You can adjust this threshold as needed
more_experience_threshold = 4  # You can adjust this threshold as needed

# Filter rows where CGPA is high and ProgrammingExperience is more
high_cgpa_more_experience = df[(df['CGPA'] >= high_cgpa_threshold) & (df['Month_of_exp_py

# Calculate the average salary expectation for this subset of students
average_salary_expectation = high_cgpa_more_experience['Expected_salary_lac'].mean()

# Print the average salary expectation
print(f'The average salary expectation among students with high CGPA and more experience
```

executed in 18ms, finished 16:58:47 2023-09-19

```
The average salary expectation among students with high CGPA and more expe
rience is: 14.13
```

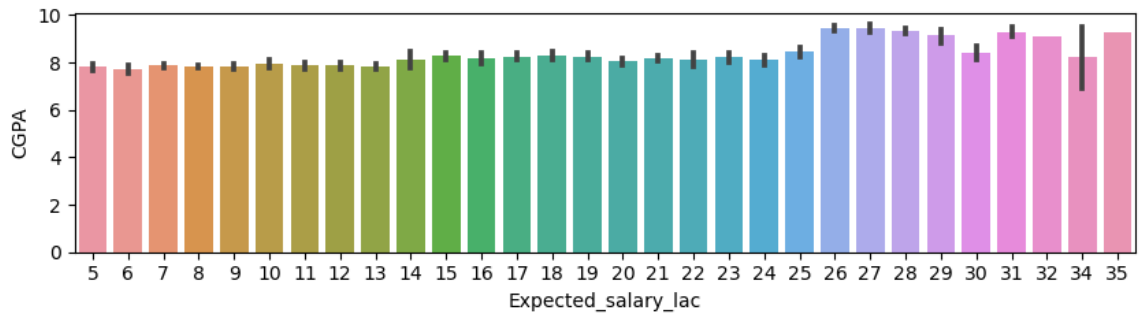
In [122]:

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

plt.subplot(2,1,1)
sns.barplot(x=high_cgpa_more_experience['Expected_salary_lac'],y=high_cgpa_more_experienc
```

executed in 1.28s, finished 16:58:49 2023-09-19

Out[122]:

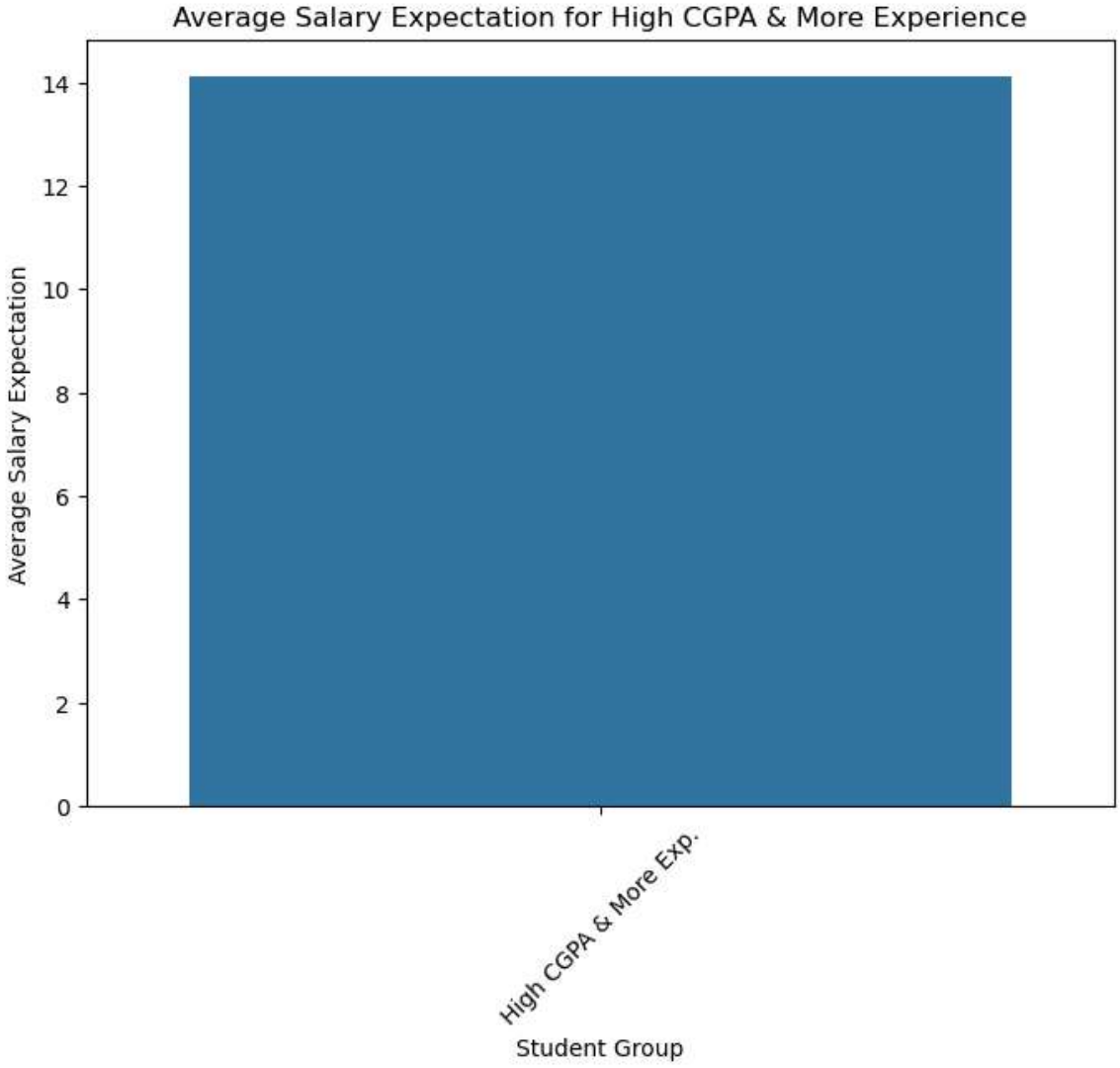
```
<Axes: xlabel='Expected_salary_lac', ylabel='CGPA'>
```



In [124]:

```
high_cgpa_threshold = 3.8
more_experience_threshold = 4

# Filter rows where CGPA is high and ProgrammingExperience is more
high_cgpa_more_experience = df[(df['CGPA'] >= high_cgpa_threshold) & (df['Month_of_exp_py

# Calculate the average salary expectation for this subset of students
average_salary_expectation = high_cgpa_more_experience['Expected_salary_lac'].mean()

# Create a bar plot to visualize the average salary expectation
plt.figure(figsize=(8, 6))
sns.barplot(x=["High CGPA & More Exp."], y=[average_salary_expectation])
plt.xlabel("Student Group")
plt.ylabel("Average Salary Expectation")
plt.title("Average Salary Expectation for High CGPA & More Experience")
plt.xticks(rotation=45)
plt.show()
```

executed in 169ms, finished 16:58:54 2023-09-19



CONCLUSION:

The average salary expectation among students with high CGPA and more experience is: 14.13