```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

a_df = pd.read_csv('C:/Users/HIMANSHU/Desktop/Data Analyst
Bootcamp/Assignments/PROJECT PORTFOLIO/International Student
Demographics/Data/academic.csv')
ad_df = pd.read_csv('C:/Users/HIMANSHU/Desktop/Data Analyst
Bootcamp/Assignments/PROJECT PORTFOLIO/International Student
Demographics/Data/academic_detail.csv')
o_df = pd.read_csv('C:/Users/HIMANSHU/Desktop/Data Analyst
Bootcamp/Assignments/PROJECT PORTFOLIO/International Student
Demographics/Data/origin.csv')
sf_df = pd.read_csv('C:/Users/HIMANSHU/Desktop/Data Analyst
Bootcamp/Assignments/PROJECT PORTFOLIO/International Student
Demographics/Data/source_of_fund.csv')
s_df = pd.read_csv('C:/Users/HIMANSHU/Desktop/Data Analyst
Bootcamp/Assignments/PROJECT PORTFOLIO/International Student
Demographics/Data/status.csv')
fs_df = pd.read_csv('C:/Users/HIMANSHU/Desktop/Data Analyst
Bootcamp/Assignments/PROJECT PORTFOLIO/International Student
Demographics/Data/field_of_study.csv')

dfs = [a_df, ad_df, o_df, sf_df, s_df, fs_df]
```

DATA SUMMARY

```
def summary(df):
    print(f'data shape: {df.shape}')
    summ = pd.DataFrame(df.dtypes, columns=['Data Type'])
    summ['Missing#'] = df.isna().sum()
    summ['Missing%'] = (df.isna().sum())/len(df)
    summ['Dups'] = df.duplicated().sum()
    summ['Uniques'] = df.nunique().values
    summ['Count'] = df.count().values
    desc = pd.DataFrame(df.describe(include='all').transpose())
    summ['Min'] = desc['min'].values
    summ['Max'] = desc['max'].values
    summ['Average'] = desc['mean'].values
    summ['Standard Deviation'] = desc['std'].values
    summ['First Value'] = df.loc[0].values
    summ['Second Value'] = df.loc[1].values
    summ['Third Value'] = df.loc[2].values

    display(summ)

for df in dfs:
    summary(df)
```

```
data shape: (75, 7)
```

```
               Data Type  Missing#  Missing%  Dups  Uniques  Count
Min  \
year              object         0  0.000000     0       75     75
NaN
students           int64         0  0.000000     0       75     75
25464.0
us_students      float64         3  0.040000     0       72     72
2102000.0
undergraduate    float64        26  0.346667     0       49     49
19101.0
graduate         float64        26  0.346667     0       49     49
12118.0
non_degree       float64        31  0.413333     0       44     44
16850.0
opt              float64        31  0.413333     0       44     44
2840.0

                        Max           Average Standard Deviation First
Value  \
year                    NaN               NaN                  NaN
1948/49
students          1095299.0    396111.773333       324680.532855
25464
us_students      21253000.0  12327495.555556      6008503.724355
2403400.0
undergraduate      442746.0     243060.22449       103238.712533
NaN
graduate           467027.0    224228.204082       109538.265949
NaN
non_degree          93587.0     38366.909091        20995.604738
NaN
opt                223539.0          60031.5        71502.857607
NaN

               Second Value Third Value
year                1949/50     1950/51
students              26433       29813
us_students       2445000.0   2281000.0
undergraduate           NaN         NaN
graduate                NaN         NaN
non_degree              NaN         NaN
opt                     NaN         NaN

data shape: (216, 4)

                Data Type  Missing#  Missing%  Dups  Uniques  Count
Min  \
year               object         0       0.0     0       24    216
NaN
academic_type      object         0       0.0     0        4    216
```

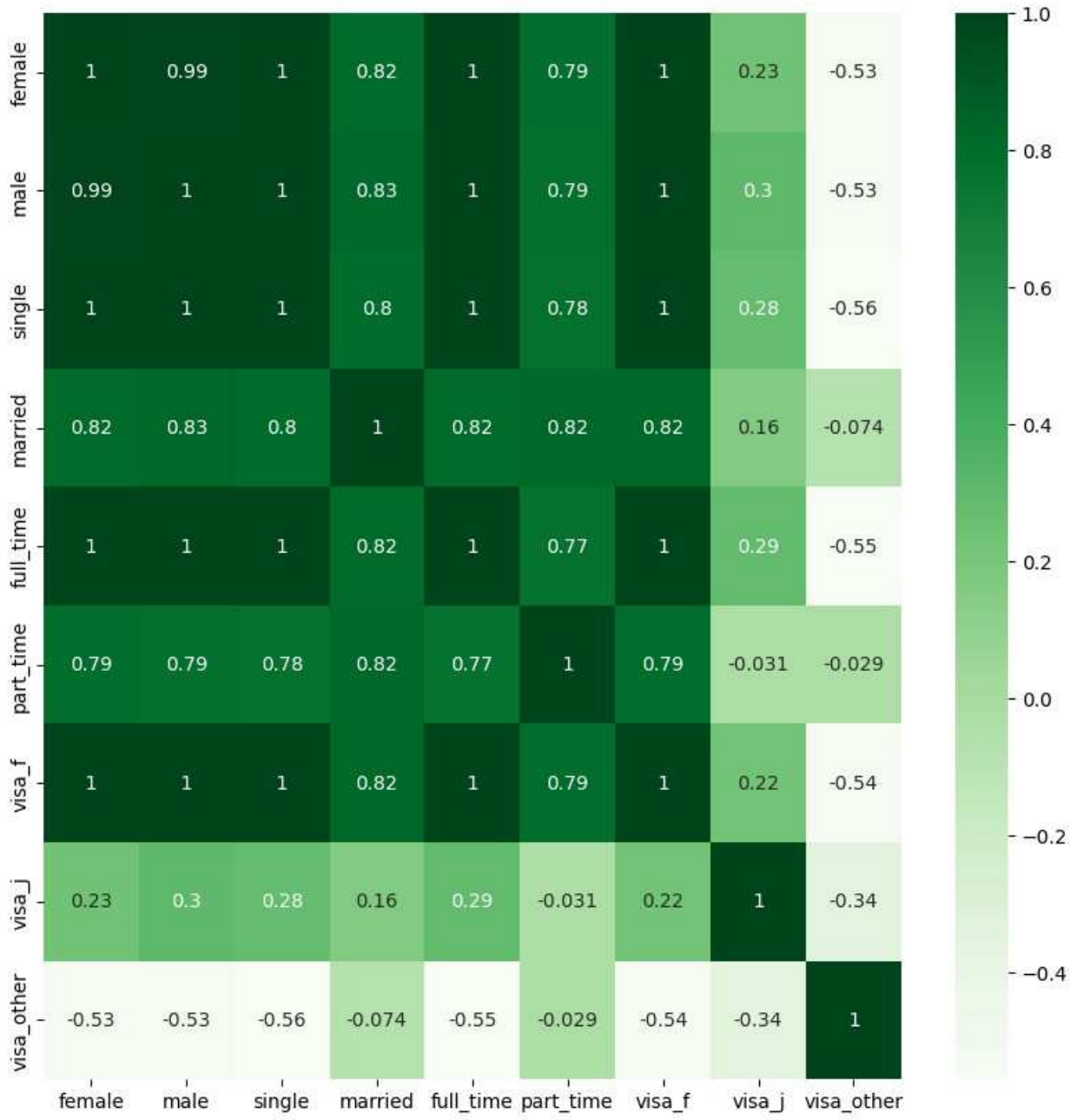
```
NaN
academic_level    object         0       0.0     0        9    216
NaN
students           int64         0       0.0     0      216    216
7093.0

                     Max       Average Standard Deviation    First
Value  \
year                 NaN           NaN                NaN
1999/00
academic_type        NaN           NaN                NaN
Undergraduate
academic_level       NaN           NaN                NaN
Associate's
students        363927.0  87870.342593       86799.205915
59830

                Second Value Third Value
year                 1999/00     1999/00
academic_type   Undergraduate    Graduate
academic_level     Bachelor's    Master's
students               177381      110857

data shape: (20411, 5)

              Data Type  Missing#  Missing%  Dups  Uniques  Count  Min
\
year             object         0       0.0     0       23  20411  NaN

origin_region    object         0       0.0     0       19  20411  NaN

origin           object         0       0.0     0      244  20411  NaN

academic_type    object         0       0.0     0        5  20411  NaN

students          int64         0       0.0     0     2867  20411  0.0


                    Max   Average Standard Deviation  \
year                NaN       NaN                NaN
origin_region       NaN       NaN                NaN
origin              NaN       NaN                NaN
academic_type       NaN       NaN                NaN
students       165936.0  904.4477        5865.759311

                                   First Value  \
year                                   2000/01
origin_region                Africa, Subsaharan
origin         Africa, Subsaharan, Unspecified
academic_type                        Graduate
students                                    2
```

```
                                  Second Value  \
year                                   2000/01
origin_region                 Africa, Subsaharan
origin         Africa, Subsaharan, Unspecified
academic_type                            Other
students                                     0

                                   Third Value
year                                   2000/01
origin_region                 Africa, Subsaharan
origin         Africa, Subsaharan, Unspecified
academic_type                      Undergraduate
students                                     6

data shape: (801, 5)

                Data Type  Missing#  Missing%  Dups  Uniques  Count
Min  \
year               object         0       0.0     0       24    801
NaN
academic_type      object         0       0.0     0        5    801
NaN
source_type        object         0       0.0     0        3    801
NaN
source_of_fund     object         0       0.0     0        9    801
NaN
students            int64         0       0.0     0      617    801
0.0

                     Max       Average Standard Deviation  \
year                 NaN           NaN                NaN
academic_type        NaN           NaN                NaN
source_type          NaN           NaN                NaN
source_of_fund       NaN           NaN                NaN
students        364824.0  23692.245943       59244.460408

                         First Value                         Second Value
\
year                           1999/00                              1999/00

academic_type           Undergraduate                        Undergraduate

source_type             International                        International

source_of_fund  Personal and Family  Foreign Government or University

students                       201578                                9742


                              Third Value
```

```
year                              1999/00
academic_type                Undergraduate
source_type                   International
source_of_fund  Foreign Private Sponsor
students                               6245

data shape: (16, 10)

            Data Type  Missing#  Missing%  Dups  Uniques  Count
Min  \
year           object         0       0.0     0       16     16
NaN
female        float64         0       0.0     0       16     16
278841.0
male          float64         0       0.0     0       16     16
344964.0
single        float64         0       0.0     0       16     16
543958.0
married       float64         0       0.0     0       16     16
69435.0
full_time     float64         0       0.0     0       16     16
575772.0
part_time     float64         0       0.0     0       16     16
48033.0
visa_f        float64         0       0.0     0       16     16
552691.0
visa_j        float64         0       0.0     0       16     16
16454.0
visa_other    float64         0       0.0     0       16     16
29508.0

                   Max      Average Standard Deviation First Value  \
year               NaN          NaN                NaN     2007/08
female        480836.0   400165.875       71337.004816    278841.0
male          617463.0   503758.375       96374.058647    344964.0
single       1002199.0     818619.0      159436.299843    543958.0
married       107882.0     85305.25        9826.503247     79847.0
full_time    1030676.0    843113.75      161056.683359    575772.0
part_time      72025.0      60810.5        8188.518653     48033.0
visa_f       1018628.0  827600.1875      166980.657148    552691.0
visa_j         58496.0   40853.5625        9937.639606     31814.0
visa_other     42984.0      35470.5        3733.618924     39300.0

            Second Value Third Value
year             2008/09     2009/10
female          304242.0    309534.0
male            367374.0    381389.0
single          591694.0    615612.0
married          79922.0     75311.0
full_time       613185.0    637722.0
```

```
part_time       58431.0     53201.0
visa_f         589007.0    612158.0
visa_j          39625.0     38692.0
visa_other      42984.0     40073.0

data shape: (1075, 4)

                Data Type  Missing#  Missing%  Dups  Uniques  Count
Min  \
year               object         0  0.000000     0       25   1075
NaN
field_of_study     object         0  0.000000     0       15   1075
NaN
major              object         0  0.000000     0       43   1075
NaN
students          float64        38  0.035349     0     1004   1037
1.0

                     Max       Average Standard Deviation  First Value
\
year                 NaN           NaN                NaN      1998/99

field_of_study       NaN           NaN                NaN  Agriculture

major                NaN           NaN                NaN  Agriculture

students        215290.0  18611.679846       34255.027828       6146.0


                                       Second Value               Third
Value
year                                        1998/99
1998/99
field_of_study                          Agriculture  Business and
Management
major           Natural Resources and Conservation  Business and
Management
students                                    1803.0
101360.0

plt.figure(figsize=(10,10))
sns.heatmap(a_df.select_dtypes(include=[float,
int]).corr(),annot=True,cmap='Greens')
plt.show()

plt.figure(figsize=(10,10))
sns.heatmap(s_df.select_dtypes(include=[float,
int]).corr(),annot=True,cmap='Greens')
plt.show()
```


students
us_students
undergraduate
graduate
non_degree
opt
1
0.93
0.98
0.97
0.81
0.96
0.93
1
0.91
0.93
0.84
0.8
0.98
0.91
1
0.94
0.83
0.93
0.97
0.93
0.94
1
0.76
0.88
0.81
0.84
0.83
0.76
1
0.66
0.96
0.8
0.93
0.88
0.66
1
1.00
0.95
0.90
0.85
0.80
0.75
0.70

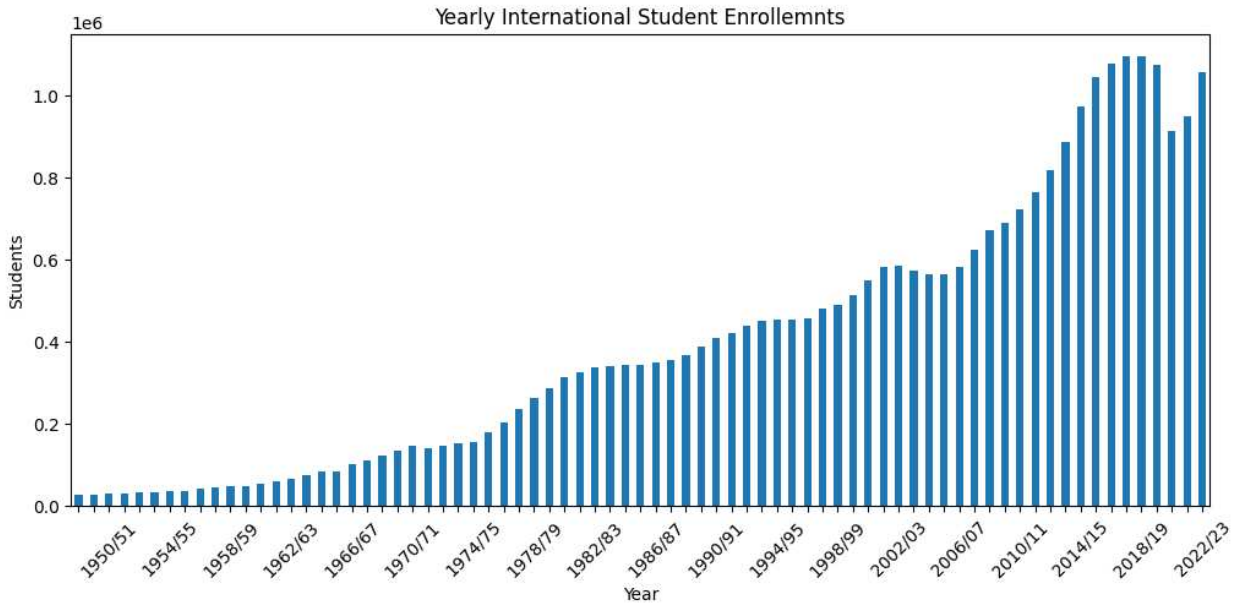


General Trends

Q. How has the overall number of international students changed over the years?

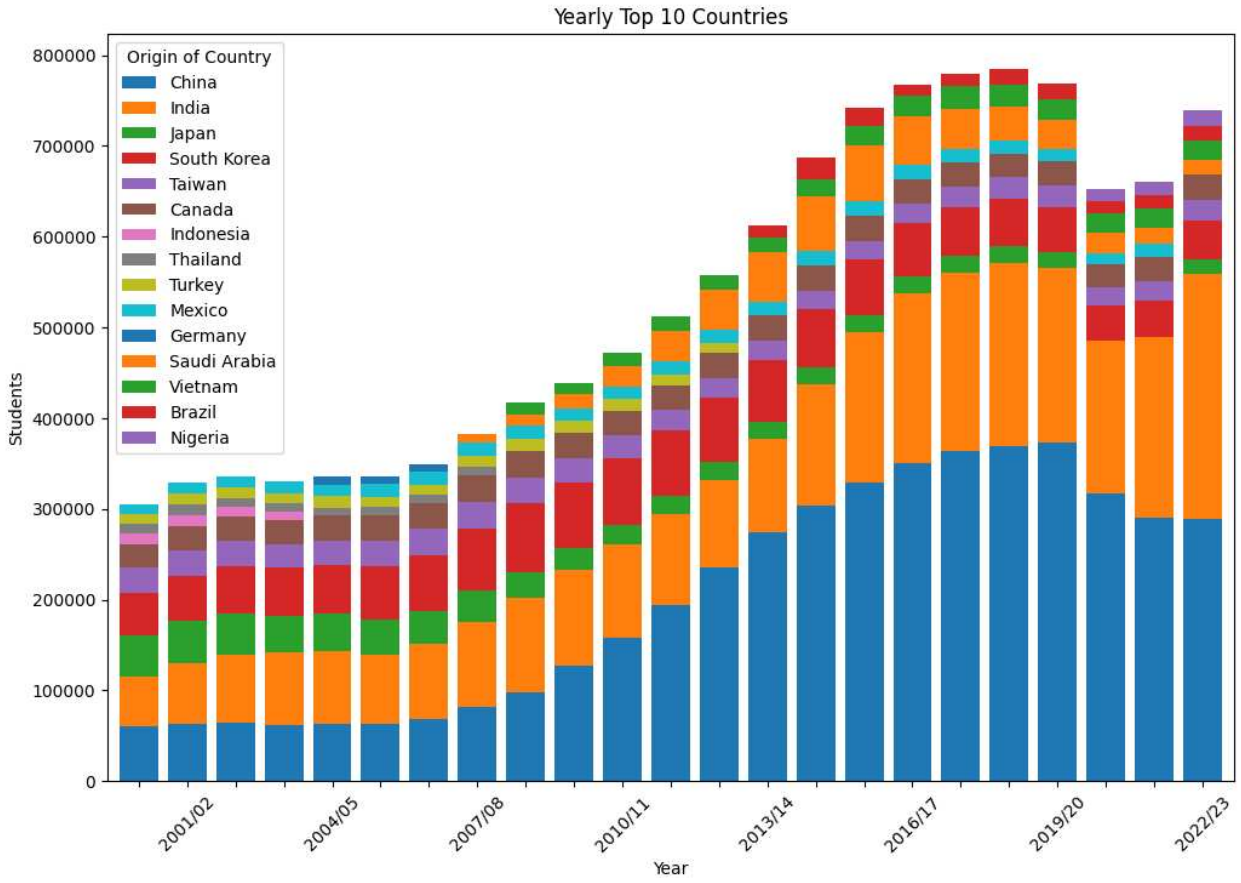
```
fig, ax = plt.subplots(figsize=(12,5))
a_df.plot(kind="bar", x="year", y="students", ax=ax)
plt.xticks(ticks=range(len(a_df["year"])), labels=[v if i%4 -2 == 0
else '' for i, v in enumerate(a_df["year"])])
plt.xlabel("Year")
```

```
plt.ylabel("Students")
plt.title("Yearly International Student Enrollemnts")
ax.xaxis.set_tick_params(rotation=45)
ax.get_legend().remove()
plt.show()
```



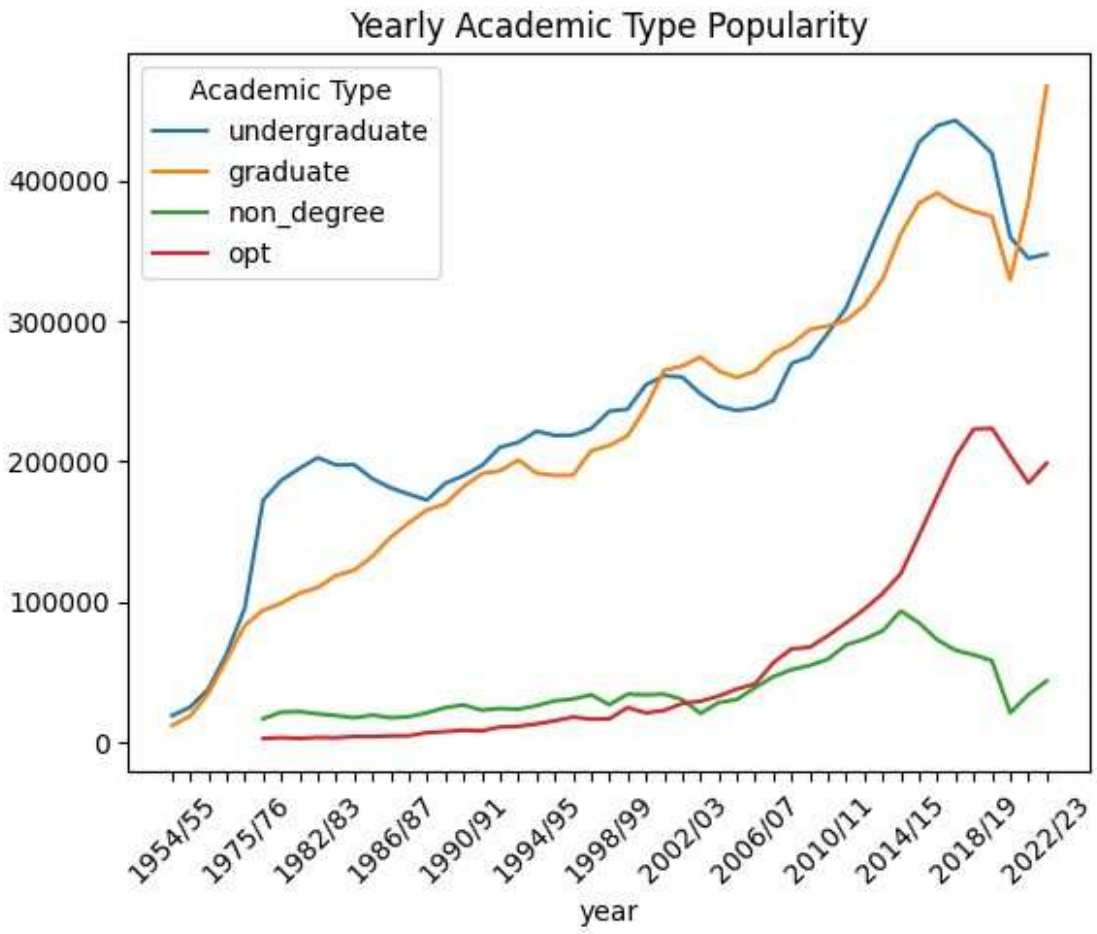
Yearly top 10 countries international student distribution.

```
origin_df = o_df.groupby(["year", "origin"])
["students"].sum().reset_index(name="students")
origin_df = origin_df.pivot_table(index="year", columns="origin",
values="students")
top10 =
pd.concat([data.sort_values(ascending=False).iloc[:10].to_frame() for
_, data in origin_df.iterrows()], axis=1).transpose()
ax = top10.plot(kind='bar', figsize=(12, 8), width=0.8, stacked=True,
                xlabel='Year', ylabel='Students', title='Yearly Top 10
Countries')
plt.xticks(ticks=range(len(top10.index)), labels=[v if i%3 -1 == 0
else '' for i, v in enumerate(top10.index)])
ax.xaxis.set_tick_params(rotation=45)
ax.legend(title="Origin of Country")
plt.show()
```



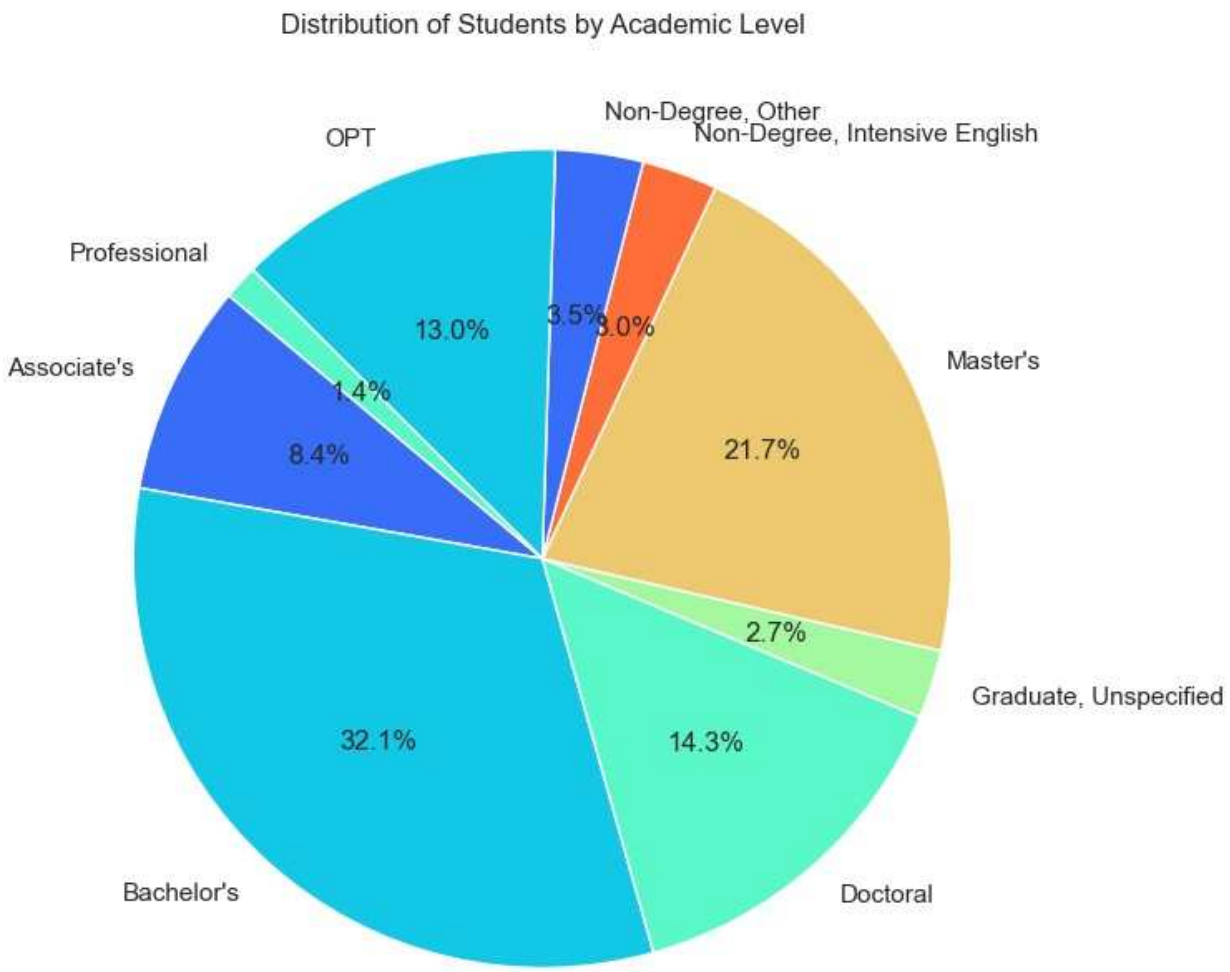
Q What is the Yearly Academic Type Popularity?

```
degree_df = a_df[a_df["undergraduate"].notna()]
ax = degree_df.plot(
    x="year", y=["undergraduate", "graduate", "non_degree", "opt"]
)
ax.xaxis.set_tick_params(rotation=45)
plt.xticks(ticks=range(len(degree_df["year"])), labels=[v if i%4 == 0
else '' for i, v in enumerate(degree_df["year"])])
ax.legend(title="Academic Type")
plt.title("Yearly Academic Type Popularity")
plt.show()
```



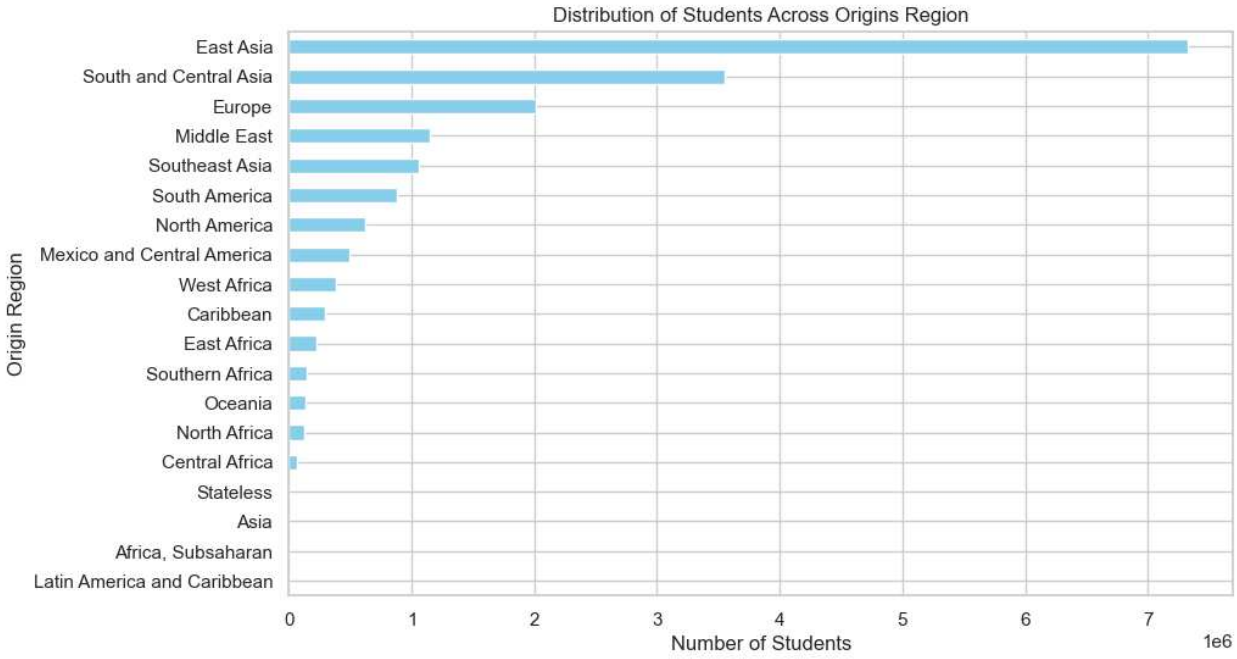
Q. What are the most popular academic levels among international students?

```
# Set Seaborn style
sns.set(style="whitegrid")

# Create a pie chart using Matplotlib
plt.figure(figsize=(8, 8))
plt.pie(academic_df["students"], labels=academic_df["academic_level"],
        autopct='%1.1f%%', startangle=140,
colors=sns.color_palette("rainbow"))

# Add a title
plt.title("Distribution of Students by Academic Level")

# Show the plot
plt.show()
```



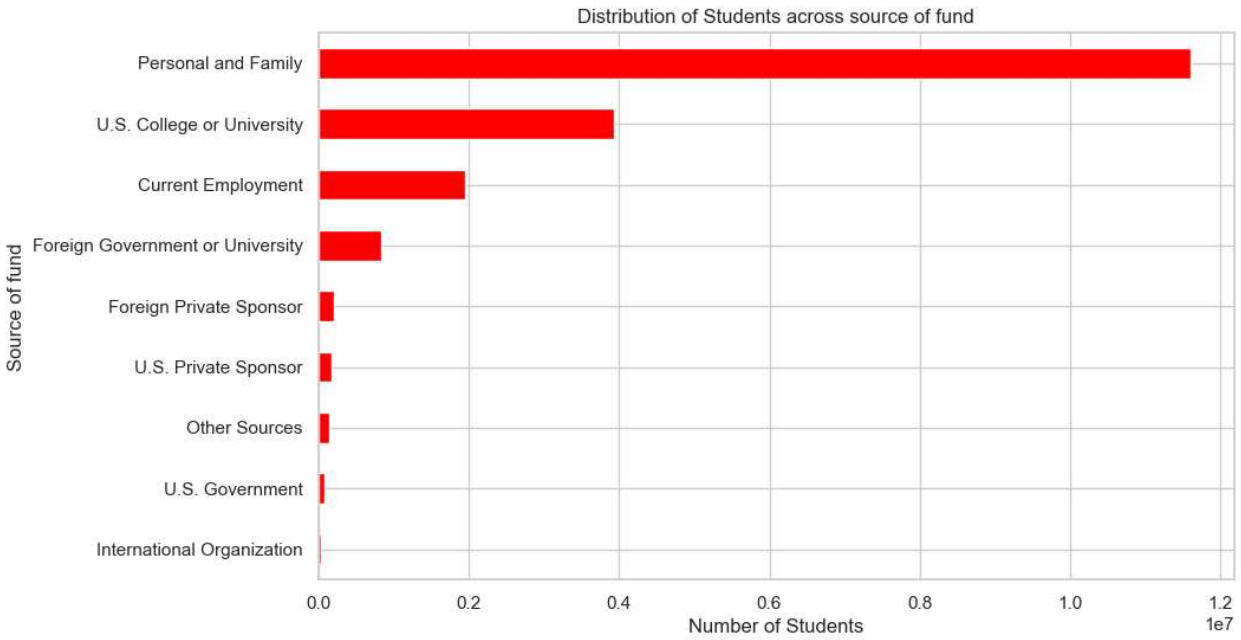
Q. What is the distribution of students across Origin Regions?

```
# Group the data by 'origin' and sum the number of students for each origin
origin_distribution = o_df.groupby('origin_region')['students'].sum()

# Create a horizontal bar chart
plt.figure(figsize=(10, 6))
origin_distribution.sort_values().plot(kind='barh', color='skyblue')

# Add labels and title
plt.xlabel('Number of Students')
plt.ylabel('Origin Region')
plt.title('Distribution of Students Across Origins Region')

# Display the bar chart
plt.show()
```



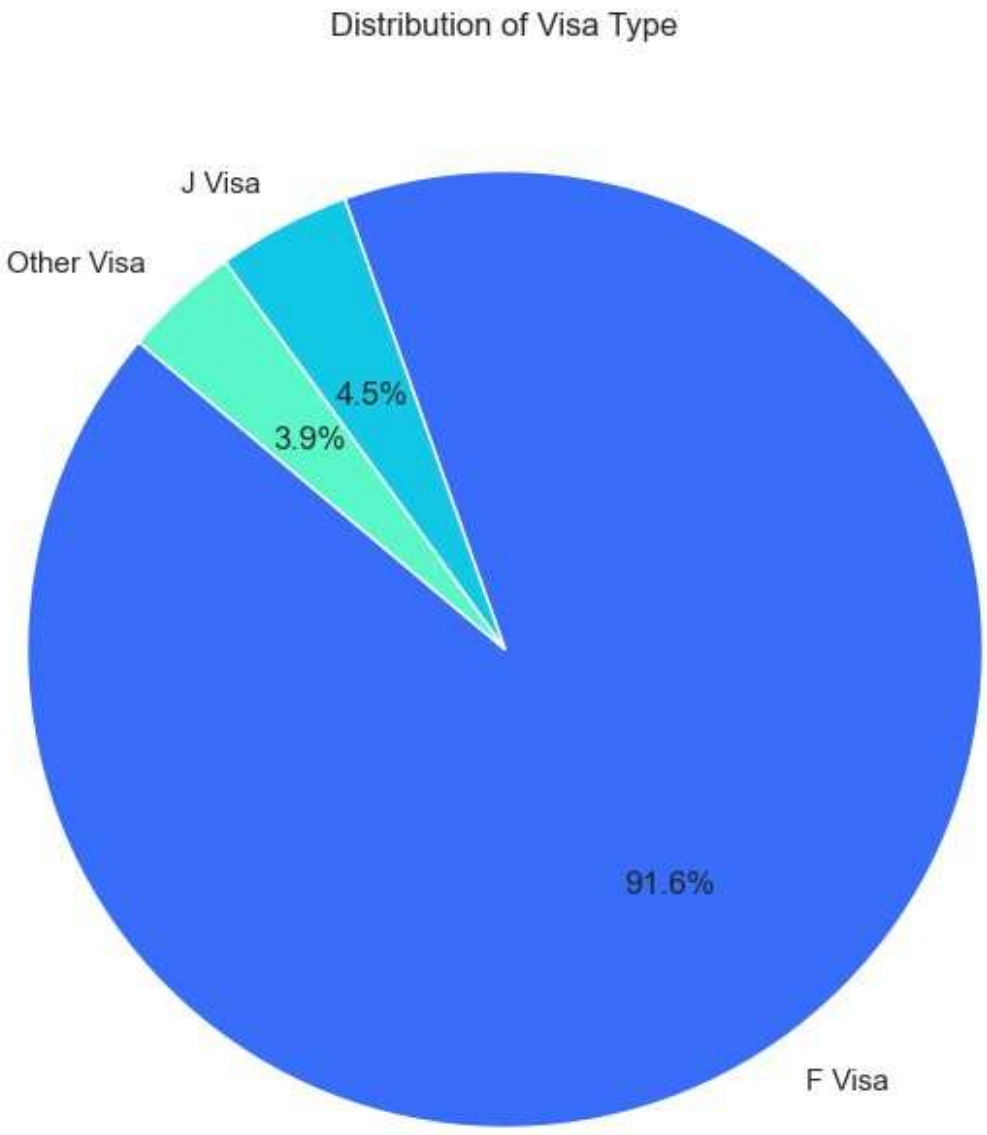
Q. What is the distribution of Students based on the source of fund?

```
source_of_fund_distribution = sf_df.groupby('source_of_fund')
['students'].sum()

# Create a horizontal bar chart
plt.figure(figsize=(10, 6))
source_of_fund_distribution.sort_values().plot(kind='barh',
color='red')

# Add labels and title
plt.xlabel('Number of Students')
plt.ylabel('Source of fund')
plt.title('Distribution of Students across source of fund')

# Display the bar chart
plt.show()
```



Q. What is the distribution of visa Type among the students?

```
visa_df = (
    s_df[["visa_f", "visa_j", "visa_other"]]
    .sum()
    .reset_index(name="students")
    .replace({"visa_f": "F Visa", "visa_j": "J Visa", "visa_other":
"Other Visa"})
)


# Create a pie chart using Matplotlib
plt.figure(figsize=(8, 8))
plt.pie(visa_df["students"],labels=visa_df["index"],
        autopct='%1.1f%%', startangle=140,
colors=sns.color_palette("rainbow"))

# Add a title
plt.title("Distribution of Visa Type")

# Show the plot
plt.show()
```


Distribution of Visa Type
J Visa
Other Visa
4.5%
3.9%
91.6%
F Visa