In [1]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [22]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as  sns
import re
```

In [3]:
```
# Importing the Data
file_path = ('/content/drive/MyDrive/Data/Copy of Copy of DePaul_Data.xlsx')
df = pd.read_excel(file_path)

print(df.shape)
df.head()
```

```
(7543, 80)
```

Out[3]:

	Reference_ID	Given_Name	Last_Name	College	Major	Degree_Type	Country	Recieved_At	Counsler	University	...	Degree_Type-2
0	45405320	Deeksha Reddy	Bhumireddy	INDIA - Osmania University - Bachelor's Degree	NaN	NaN	India	1757494436974	NaN	DePaul University	...	NaN
1	858032003	Pearl Ashok Kumar	Patel	INDIA - Manipal University Jaipur - Bachelor's...	NaN	NaN	India	1757494436974	NaN	DePaul University	...	NaN
2	902518555	Hamza	Javed	PAKISTAN - Shaheed Zulfikar Ali Bhutto Institu...	NaN	NaN	Pakistan	1757494436974	NaN	DePaul University	...	grad
3	902518555	Hamza	Javed	PAKISTAN - Shaheed Zulfikar Ali Bhutto Institu...	NaN	NaN	Pakistan	1757494436974	NaN	DePaul University	...	grad
4	218755608	Ronil Dhavalbhai	Thakkar	INDIA - Swarrnim Institute of	NaN	NaN	India	1757494436974	NaN	DePaul University	...	NaN

	Reference_ID	Given_Name	Last_Name	College	Major	Degree_Type	Country	Recieved_At	Counsler	University	...	Degree_Type-2
				Technology - Swa...								

5 rows × 80 columns

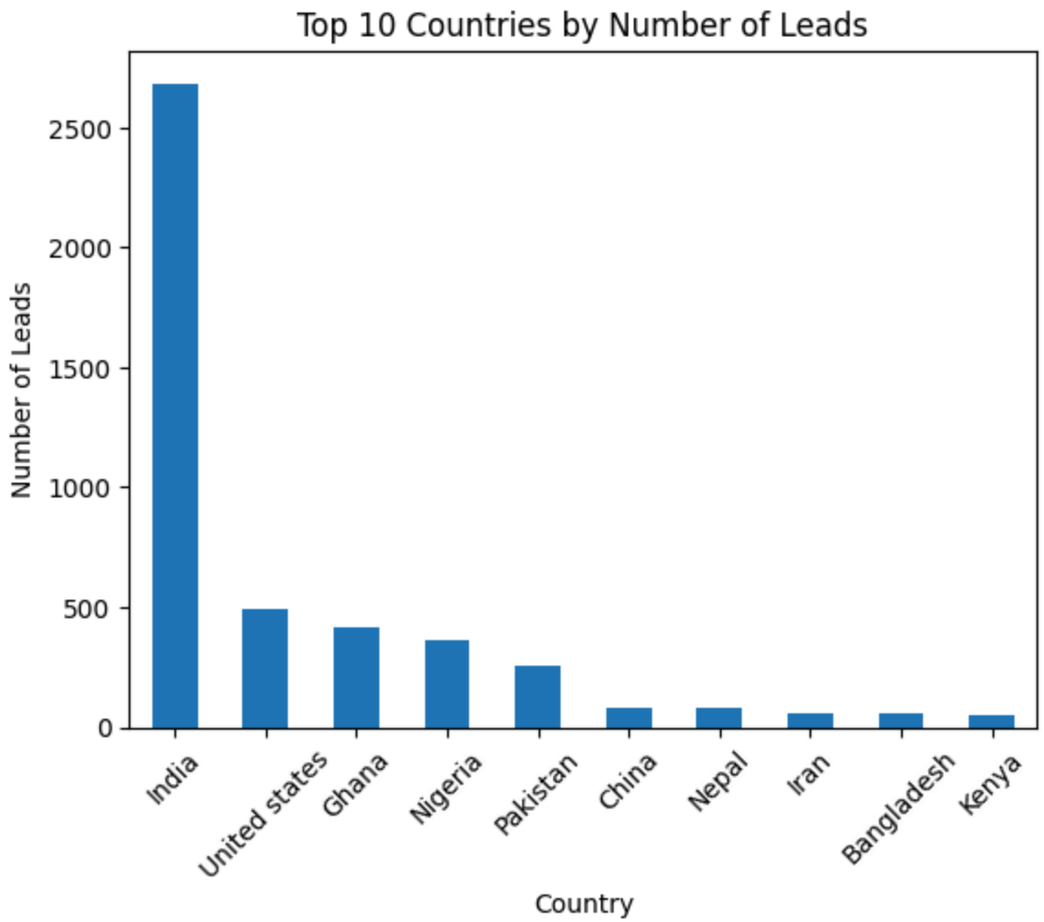
In [4]:
```
# Remove duplicate IDs
df = df.drop_duplicates(subset=["Reference_ID"])
```

In [5]:
```
# Check Number of rows after removing Duplicates
num_rows = df.shape[0]
print(num_rows)
```

```
5000
```

In [6]:
```
# Bar Chart for Top 10 Countries
top_countries = df["Country"].value_counts().head(10)

plt.figure()
top_countries.plot(kind="bar")
plt.title("Top 10 Countries by Number of Leads")
plt.xlabel("Country")
plt.ylabel("Number of Leads")
plt.xticks(rotation=45)
plt.show()
```



In [ ]:
```
# As all the rows in the followin column are NULL (Major, Degree Type, Age/Data Of Birth) can not work on visualzing them and have to be skipped as we dont have a way to replace the NULLs
```

In [18]:
```
# Shows most popular admission terms
df["Intake"].value_counts()
```

Out[18]:

Intake	count
computersciencems:2024fall(september24)	199
businessanalyticsms:2024fall(september24)	166
businessanalyticsms:2025fall(september25)	142
businessanalyticsms:2024winter(january24)	121
computersciencems:2025fall(september25)	109
...	...
womensandgenderstudiesma:2023fall(september23)	1
businessinformationtechnologyms(fullyonline):2025fall(september25)	1
valuecreatingeducationforglobalcitizenshipmed(fullyonline):2024summer(june24)	1
traumapsychologycertificate:2024fall(september24)	1
nondegree(csh):2025winter(january25)	1

584 rows × 1 columns

dtype: int64

In [23]:
```
# Comment Cleaning
df["Comments_clean"] = (
    df["Comments"]
    .astype(str)
    .str.lower()
    .str.replace(r"[^a-z\s]", "", regex=True)
)
```

In [24]:
```
# Leads With vs Without Counselor Comments
df["has_comment"] = df["Comments"].notna() & (df["Comments"].str.strip() != "")

df["has_comment"].value_counts().plot(kind="bar")
plt.title("Leads With vs Without Counselor Comments")
plt.xlabel("Has Comment")
plt.ylabel("Number of Leads")
plt.show()
```