Intern Name : Faraj Momin

Intern ID : IP -4039

Domain : Machine Learning

Task 2 : Year Of Graduation Prediction Model

1. Importing the necessary Python Modules / Libraries

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
```

2. Reading the data set using the 'pandas' module

+ Code + Text

```
df = pd.read_excel('Final Lead Data.xlsx')
```

```
df.head()
```

	ID	First Name	Email	Gender	City	Created	Position	New College Name	Colleges	Academic Year	Branch/ Specialisation	Other Branch	Wha cur acad y
0	68112	ANIKET	aniket@xyz.com	NaN	NaN	04/27/2022 01:41:38 pm	NaN	NaN	NaN	NaN	NaN	NaN	
1	68110	Dhanshree	dhanshree@xyz.com	NaN	NaN	04/22/2022 04:08:38 pm	NaN	Lords Universal College	NaN	NaN	NaN	NaN	
2	68108	Dhiraj	dhiraj@xyz.com	NaN	NaN	04/16/2022 10:31:59 pm	NaN	NaN	NaN	NaN	NaN	NaN	
3	68106	Pooja	pooja@xyz.com	NaN	NaN	04/13/2022 10:05:15 pm	NaN	NaN	NaN	NaN	NaN	NaN	
4	68090	Aayush	aayush@xyz.com	NaN	NaN	03/26/2022 07:02:48 pm	NaN	B.k Birla college	NaN	NaN	NaN	NaN	

Next steps: Generate code with df View recommended plots New interactive sheet

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5303 entries, 0 to 5302
Data columns (total 18 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   ID                                              5303 non-null   int64
 1   First Name                                      5303 non-null   object
 2   Email                                           5303 non-null   object
 3   Gender                                          200 non-null    object
 4   City                                            55 non-null     object
 5   Created                                         5303 non-null   object
 6   Position                                        6 non-null      object
 7   New College Name                                1818 non-null   object
 8   Colleges                                        1681 non-null   object
 9   Academic Year                                   2518 non-null   float64
 10  Branch/ Specialisation                          2520 non-null   object
```

```
 11  Other Branch                                          644 non-null    object
 12  What is your current academic year?                   131 non-null    object
 13  Company Name/ College Name                            238 non-null    object
 14  Would you like to know more about us and our programs?  5303 non-null   object
 15  Are you interested in knowing more about our events?    5303 non-null   object
 16  Have you recommended Cloud Counselage to anyone?        5303 non-null   object
 17  How did you come to know about this event?            155 non-null    object
dtypes: float64(1), int64(1), object(16)
memory usage: 745.9+ KB
```

3. Creating a copy of raw data so as to keep the main data as it as and doing further operations on the copy of the main data

```
final_df = df.copy()
```

```
final_df['Email'].value_counts()
```

```
Email
vaishnavi@xyz.com    37
shubham@xyz.com      35
yash@xyz.com         32
abhishek@xyz.com     29
sakshi@xyz.com       28
                     ..
julia@xyz.com         1
darrell@xyz.com       1
edie@xyz.com          1
kimberley@xyz.com     1
utkarsha@xyz.com      1
Name: count, Length: 2808, dtype: int64
```

4. Dropping duplicates if any using the 'Email' column from the dataset

```
final_df = final_df.drop_duplicates(subset='Email', keep='first')
```

```
final_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2808 entries, 0 to 5260
Data columns (total 18 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   ID                                                    2808 non-null   int64
 1   First Name                                            2808 non-null   object
 2   Email                                                 2808 non-null   object
 3   Gender                                                115 non-null    object
 4   City                                                  54 non-null     object
 5   Created                                               2808 non-null   object
 6   Position                                              1 non-null      object
 7   New College Name                                      831 non-null    object
 8   Colleges                                              662 non-null    object
 9   Academic Year                                         1083 non-null   float64
 10  Branch/ Specialisation                                1085 non-null   object
 11  Other Branch                                          306 non-null    object
 12  What is your current academic year?                   54 non-null     object
 13  Company Name/ College Name                            140 non-null    object
 14  Would you like to know more about us and our programs?  2808 non-null   object
 15  Are you interested in knowing more about our events?    2808 non-null   object
 16  Have you recommended Cloud Counselage to anyone?        2808 non-null   object
 17  How did you come to know about this event?            101 non-null    object
dtypes: float64(1), int64(1), object(16)
memory usage: 416.8+ KB
```

Let's drop the 'Position' column from the dataset as it contains only 1 non-null (Non-Blank) value

```
final_df = final_df.drop('Position', axis=1)
```

```
final_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2808 entries, 0 to 5260
Data columns (total 17 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   ID                                                    2808 non-null   int64
 1   First Name                                            2808 non-null   object
 2   Email                                                 2808 non-null   object
```

```
 3   Gender                                                115 non-null    object
 4   City                                                  54 non-null     object
 5   Created                                               2808 non-null   object
 6   New College Name                                      831 non-null    object
 7   Colleges                                              662 non-null    object
 8   Academic Year                                         1083 non-null   float64
 9   Branch/ Specialisation                                1085 non-null   object
 10  Other Branch                                          306 non-null    object
 11  What is your current academic year?                   54 non-null     object
 12  Company Name/ College Name                            140 non-null    object
 13  Would you like to know more about us and our programs?  2808 non-null   object
 14  Are you interested in knowing more about our events?  2808 non-null   object
 15  Have you recommended Cloud Counselage to anyone?      2808 non-null   object
 16  How did you come to know about this event?            101 non-null    object
dtypes: float64(1), int64(1), object(15)
memory usage: 394.9+ KB
```

So, there are 2808 unique rows in the dataset provided

```
final_df['New College Name'].value_counts()
```

```
New College Name
Government Polytechnic Mumbai                                        20
Datta Meghe College of Engineering                                   11
Sri Manakula Vinayagar Engineering College                            9
Government polytechnic mumbai                                         9
St. Francis Institute of Technology                                   8
                                                                     ..
D.Y. Patil Institute of Master of Computer Applications and Management     1
Dr. D.Y. Patil Institute of MCA and Management                        1
Dy Patil Institute of MCA and Management Akurdi pune                  1
D Y Patil Institute of Master of Computer Applications and Management, Akurdi     1
Don Bosco College Of Engineering Goa                                  1
Name: count, Length: 682, dtype: int64
```

```
final_df['Colleges'].value_counts()
```

```
Colleges
Others                                                               168
Maharashtra Institute of Technology - MIT                             18
Pimpri Chinchwad College of Engineering - PCCE                        16
St. Francis Institute of Technology                                   15
KJ Somaiya College of Engineering                                     13
                                                                    ...
Shrama Sadhana Bombay Trust's College of Engineering and Technology - SSCOET      1
North Maharashtra University                                          1
Marathwada Mitra Mandal's Institute of Technology - MMIT              1
Amrutvahini College of Engineering                                    1
Smt. Indira Gandhi College of Engineering - SIGCE                     1
Name: count, Length: 165, dtype: int64
```

```
final_df['Branch/ Specialisation'].value_counts()
```

```
Branch/ Specialisation
Computer Science               473
Information Technology (IT)    307
Other                          305
Name: count, dtype: int64
```

```
final_df['Other Branch'].value_counts()
```

```
Other Branch
Electronics and telecommunication              28
Electronics and Telecommunication              17
Mechanical                                     14
Electronics                                    11
Mechanical Engineering                         11
                                               ..
information security and cloud computing        1
Electronics & Telecommunication Engineering     1
Electronic and Telecommunications               1
Mechanical Engg                                 1
Electrical & Electronics                        1
Name: count, Length: 131, dtype: int64
```

Let's combine the 'Branch / Specialisation' and 'Other Branch' columns to a new column named 'Branch' and then drop these two columns from the dataset

```
final_df['Branch'] = final_df['Branch/ Specialisation'] + ' ' + final_df['Other Branch']
```

```
columns = ['Branch/ Specialisation','Other Branch']
final_df = final_df.drop(columns, axis=1)
```

Let's combine the 'New College Name' and 'Colleges' columns to a new column named 'College' and then drop these two columns from the dataset

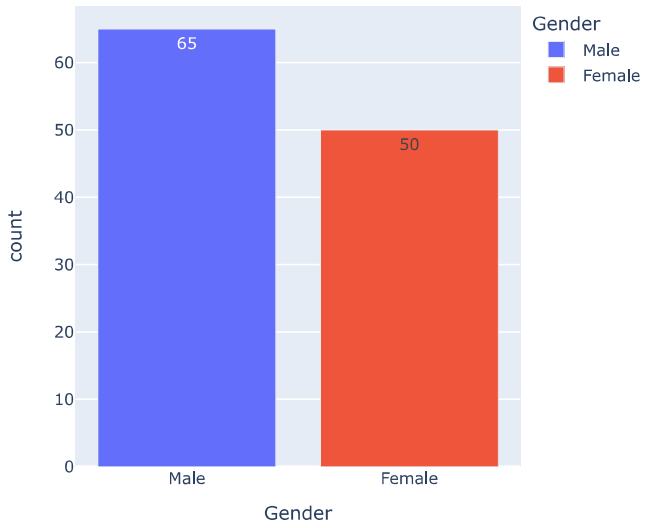
```
final_df['College'] = final_df['Colleges'] + ' ' + final_df['New College Name']


columns = ['Colleges','New College Name']
final_df = final_df.drop(columns, axis=1)


final_df.head()
```

	ID	First Name	Email	Gender	City	Created	Academic Year	What is your current academic year?
0	68112	ANIKET	aniket@xyz.com	NaN	NaN	04/27/2022 01:41:38 pm	NaN	NaN
1	68110	Dhanshree	dhanshree@xyz.com	NaN	NaN	04/22/2022 04:08:38 pm	NaN	NaN
2	68109	Dhiraj	dhiraj@xyz.com	NaN	NaN	04/16/2022	NaN	NaN

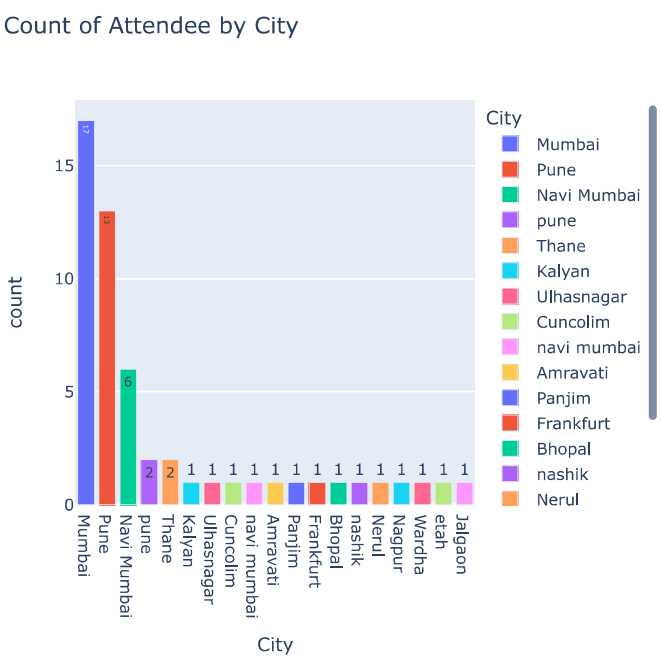
Next steps: Generate code with final_df | View recommended plots | New interactive sheet

5. Let's Analyze the given data using Data Visualization modules/Libraries

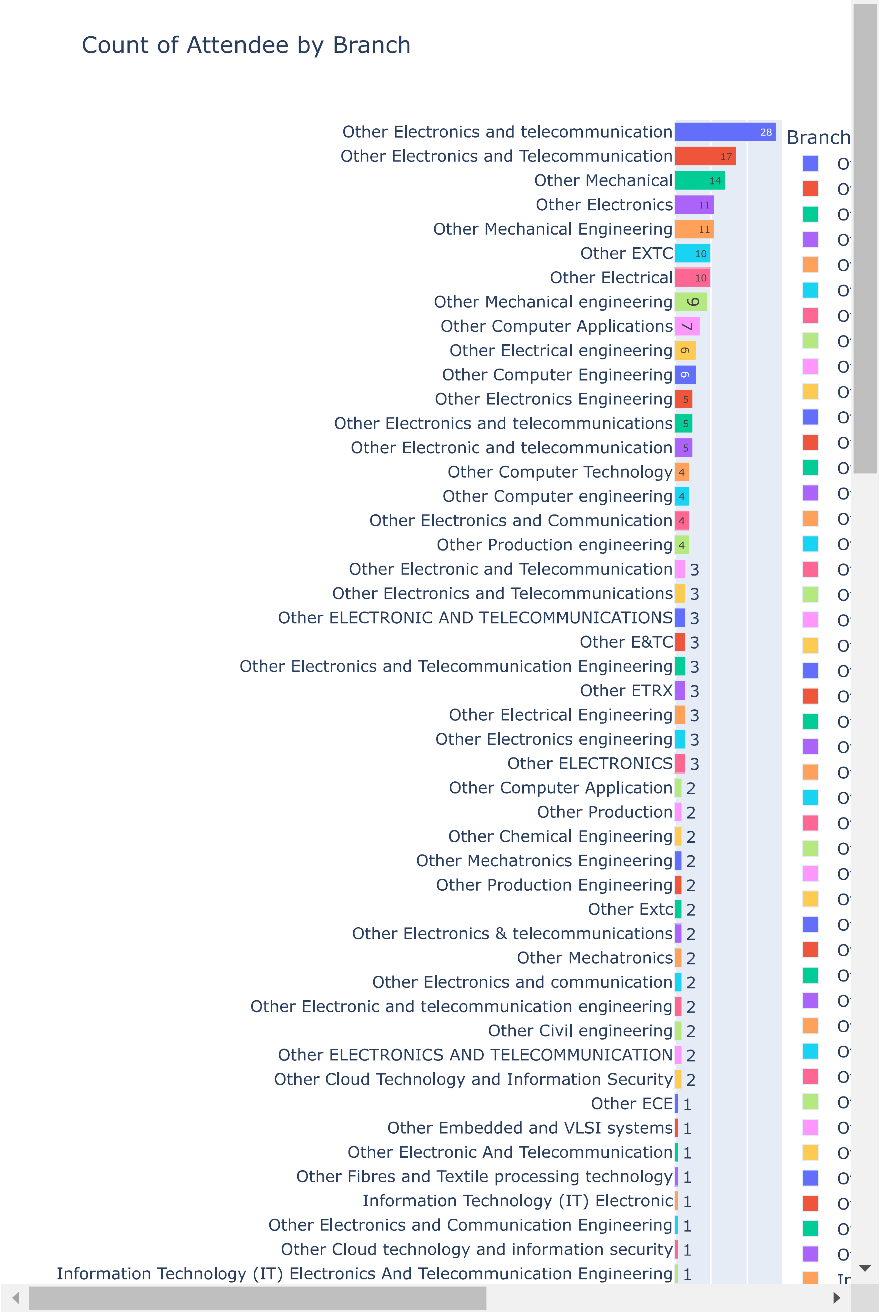
```
gender = final_df['Gender'].value_counts().keys()
count = final_df['Gender'].value_counts().values
labels = [str(val) for val in count]

gender_plot_df = pd.DataFrame({
    'Gender': gender,
    'Count': count,
})

fig = px.bar(gender_plot_df, x='Gender', y='Count', text=labels, title='Count of Attendee by Gender', template='plotly',color='Gender',

# Set the labels for the y-axis and x-axis
fig.update_yaxes(title_text='count')
fig.update_xaxes(title_text='Gender')

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



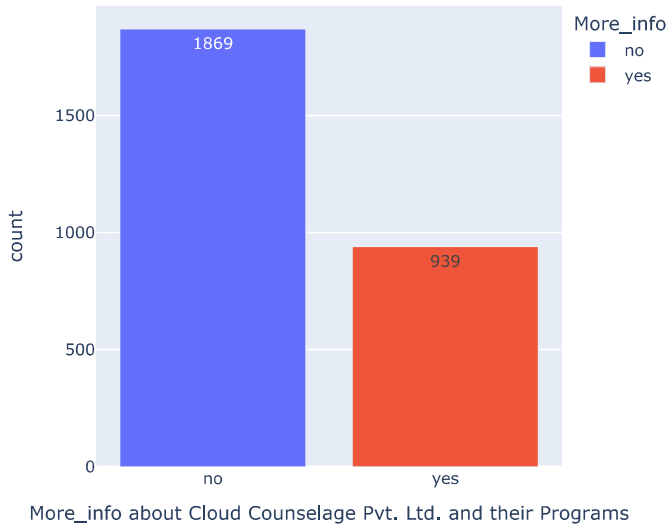

```
city = final_df['City'].value_counts().keys()
count = final_df['City'].value_counts().values
labels = [str(val) for val in count]

city_plot_df = pd.DataFrame({
    'City': city,
    'Count': count,
})

fig = px.bar(city_plot_df, x='City', y='Count', text=labels, title='Count of Attendee by City', template='plotly',color='City', color_d

# Set the labels for the y-axis and x-axis
fig.update_yaxes(title_text='count')
fig.update_xaxes(title_text='City')

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



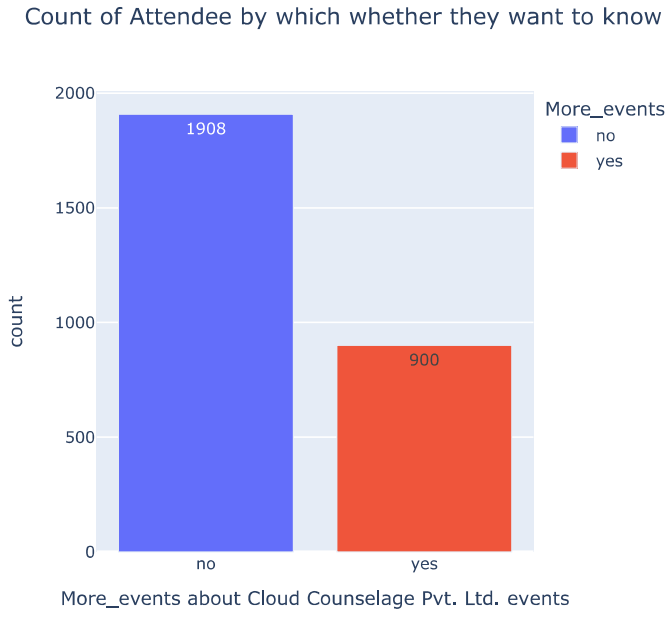
```
branch = final_df['Branch'].value_counts().keys()
count = final_df['Branch'].value_counts().values
labels = [str(val) for val in count]

branch_plot_df = pd.DataFrame({
    'Branch': branch,
    'Count': count,
})

fig = px.bar(branch_plot_df, y='Branch', x='Count', text=labels, title='Count of Attendee by Branch', template='plotly',color='Branch',

# Set the labels for the y-axis and x-axis
fig.update_xaxes(title_text='count')
fig.update_yaxes(title_text='Branch')


fig.update_layout(width=1100, height=2600)

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


Count of Attendee by Branch
Branch
Other Electronics and telecommunication 28
Other Electronics and Telecommunication 17
Other Mechanical 14
Other Electronics 11
Other Mechanical Engineering 11
Other EXTC 10
Other Electrical 10
Other Mechanical engineering 9
Other Computer Applications 7
Other Electrical engineering 6
Other Computer Engineering 6
Other Electronics Engineering 5
Other Electronics and telecommunications 5
Other Electronic and telecommunication 5
Other Computer Technology 4
Other Computer engineering 4
Other Electronics and Communication 4
Other Production engineering 4
Other Electronic and Telecommunication 3
Other Electronics and Telecommunications 3
Other ELECTRONIC AND TELECOMMUNICATIONS 3
Other E&TC 3
Other Electronics and Telecommunication Engineering 3
Other ETRX 3
Other Electrical Engineering 3
Other Electronics engineering 3
Other ELECTRONICS 3
Other Computer Application 2
Other Production 2
Other Chemical Engineering 2
Other Mechatronics Engineering 2
Other Production Engineering 2
Other Extc 2
Other Electronics & telecommunications 2
Other Mechatronics 2
Other Electronics and communication 2
Other Electronic and telecommunication engineering 2
Other Civil engineering 2
Other ELECTRONICS AND TELECOMMUNICATION 2
Other Cloud Technology and Information Security 2
Other ECE 1
Other Embedded and VLSI systems 1
Other Electronic And Telecommunication 1
Other Fibres and Textile processing technology 1
Information Technology (IT) Electronic 1
Other Electronics and Communication Engineering 1
Other Cloud technology and information security 1
Information Technology (IT) Electronics And Telecommunication Engineering 1

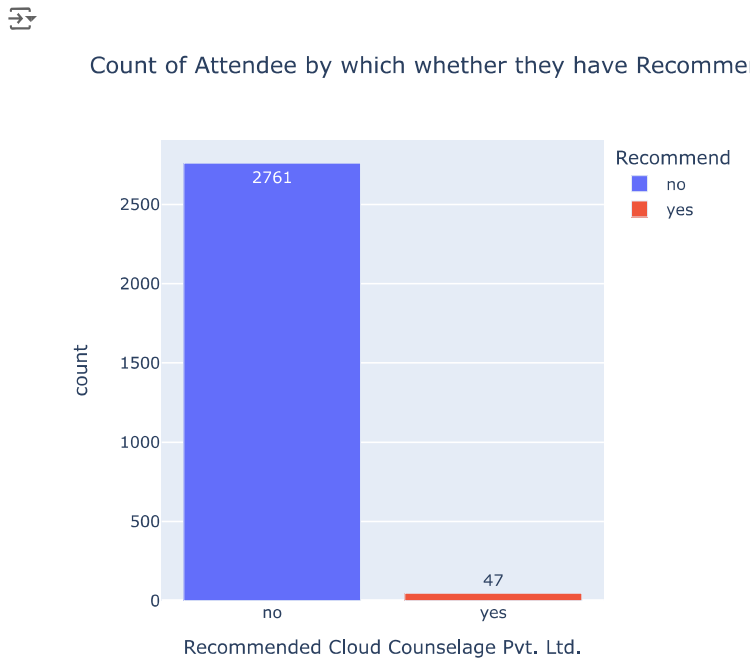
```
more_info = final_df['Would you like to know more about us and our programs?'].value_counts().keys()
count = final_df['Would you like to know more about us and our programs?'].value_counts().values
labels = [str(val) for val in count]

more_info_plot_df = pd.DataFrame({
    'More_info': more_info,
    'Count': count,
})

fig = px.bar(more_info_plot_df, x='More_info', y='Count', text=labels, title='Count of Attendee by which whether they want to know more

# Set the labels for the y-axis and x-axis
fig.update_yaxes(title_text='count')
fig.update_xaxes(title_text='More_info about Cloud Counselage Pvt. Ltd. and their Programs')

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



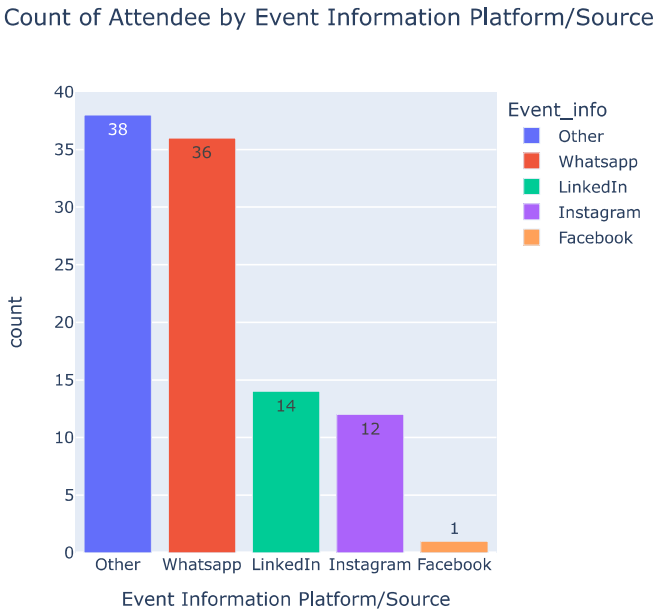

```
more_events = final_df['Are you interested in knowing more about our events?'].value_counts().keys()
count = final_df['Are you interested in knowing more about our events?'].value_counts().values
labels = [str(val) for val in count]

more_events_plot_df = pd.DataFrame({
    'More_events': more_events,
    'Count': count,
})

fig = px.bar(more_events_plot_df, x='More_events', y='Count', text=labels, title='Count of Attendee by which whether they want to know r

# Set the labels for the y-axis and x-axis
fig.update_yaxes(title_text='count')
fig.update_xaxes(title_text='More_events about Cloud Counselage Pvt. Ltd. events')

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



```
recommend = final_df['Have you recommended Cloud Counselage to anyone?'].value_counts().keys()
count = final_df['Have you recommended Cloud Counselage to anyone?'].value_counts().values
labels = [str(val) for val in count]

recommend_plot_df = pd.DataFrame({
    'Recommend': recommend,
    'Count': count,
})

fig = px.bar(recommend_plot_df, x='Recommend', y='Count', text=labels, title='Count of Attendee by which whether they have Recommended (

# Set the labels for the y-axis and x-axis
fig.update_yaxes(title_text='count')
fig.update_xaxes(title_text='Recommended Cloud Counselage Pvt. Ltd.')

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




```
event_info = final_df['How did you come to know about this event?'].value_counts().keys()
count = final_df['How did you come to know about this event?'].value_counts().values
labels = [str(val) for val in count]

event_info_plot_df = pd.DataFrame({
    'Event_info': event_info,
    'Count': count,
})

fig = px.bar(event_info_plot_df, x='Event_info', y='Count', text=labels, title='Count of Attendee by Event Information Platform/Source',

# Set the labels for the y-axis and x-axis
fig.update_yaxes(title_text='count')
fig.update_xaxes(title_text='Event Information Platform/Source')

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



6. Let's use the 'Created' , 'Academic Year' and 'What is your current academic year?' columns to predict the Graduation Year of the attendee

```
final_df['Academic Year'].value_counts()
```

```
Academic Year
3.0    512
2.0    295
4.0    214
1.0     62
Name: count, dtype: int64
```

```
final_df['What is your current academic year?'].value_counts()
```

```
What is your current academic year?
3rd Year      21
Final Year    19
2nd Year      10
1st Year       4
Name: count, dtype: int64
```

Replacing values in 'What is your current academic year?' column: -

- 3rd Year to 3
- Final Year to 4
- 2nd Year to 2
- 1st Year to 1

```
final_df['What is your current academic year?'] = final_df['What is your current academic year?'].replace({
    '3rd Year': 3,
    'Final Year': 4,
    '2nd Year': 2,
    '1st Year': 1
})
```

Replacing null values to 0

```
final_df['What is your current academic year?'] = final_df['What is your current academic year?'].fillna(0)
```

```
final_df['What is your current academic year?'] = final_df['What is your current academic year?'].astype(int)
```