PYTHON CAC-2

Exploratory Data Analysis

The team must identify a Domain related to Lavasa (Student Life, Traveling Preferences, Food Preferences...), create a unique problem statement associated with the domain, and determine the attributes associated with the problem.

Collect data using various techniques, perform Exploratory Data Analysis (with the help of Python), and Present the Key insights obtained from the analysis in a Chart.

DOMAIN:- STUDENT'S LIFE IN THE COLLEGE CAMPUS

PROBLEM:- MANAGEMENT BETWEEN ACADEMICS AND EXTRA CURRICULAR

Install Python Libraries

In [ ]:
```
# pip install pandas
```

In [ ]:
```
# pip install numpy
```

In [ ]:
```
# pip install matplotlib
```

IMPORT PYTHON LIBRARIES

Pandas and Numpy have been used for Data Manipulation and Numerical Calculations

Matplotlib and Seaborn have been used for Data visualizations.

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

READING THE DATASET

In [ ]:
```
#READING THE DATASET
pd.read_excel("Students' Life at Christ (Responses).xlsx")
```

Out[ ]:

	Timestamp	Name	Department	GENDER	YEAR	Do you follow a Semester or Trimester System?	Do you participate in any extra curricular activities?	If No, how do you spend your time after college hours?	f
0	2023-10-10 21:20:21.933	Stuty Das	MSC DS	FEMALE	FIRST	Trimester	No	Cooking	
1	2023-10-10 21:32:57.896	Aryan Manchanda	BSC DS	MALE	FIRST	Semester	YES	NaN	
2	2023-10-10 22:16:09.915	Anju Mathew	LLB	FEMALE	SECOND	Semester	YES	NaN	
3	2023-10-10 22:17:36.971	Gracy Rathaur	LLB	FEMALE	THIRD	Semester	YES	NaN	
4	2023-10-10 22:35:12.016	Sreelakshmi G	BBA	FEMALE	FIRST	Semester	No	Academic works	
...	...	...	...	...	...	...	...	...	
397	2023-10-17 22:30:13.221	Rhea Eleanor Rhine	BBA	FEMALE	THIRD	Semester	No	Hang out with friends	
398	2023-10-17 22:31:36.111	Pariska Nagvenkar	BBA	FEMALE	SECOND	Semester	No	Academic works	
399	2023-10-17 22:33:12.435	Vatsal Sharma	LLB	MALE	SECOND	Semester	YES	NaN	
400	2023-10-17 22:33:20.940	Rachel	MBA	FEMALE	FIRST	Trimester	YES	NaN	
401	2023-10-17 22:34:02.656	Shilpi Singh	BCOM	FEMALE	SECOND	Semester	No	Cooking	

402 rows × 17 columns

In [ ]:
```
data = pd.read_excel("Students' Life at Christ (Responses).xlsx")
```

DATA CLEANING

In [ ]:
```
data["GENDER"].replace({"Male": "MALE", "Female": "FEMALE","FeMALE":"FEMALE"}, inpl
```

ANALYZING THE DATASET

In [ ]:
```
data.dtypes
```

Out[ ]:
```
Timestamp
datetime64[ns]
Name
object
Department
object
GENDER
object
YEAR
object
Do you follow a Semester or Trimester System?
object
Do you participate in any extra curricular activities?
object
If No,  how do you spend your time after college hours?
object
IF YES, select the following
object
Are you able to balance your academics with extracurricular?
object
How do you rate your time management\n(On the scale of 5)
int64
Do you think participation in extracurricular affect your academic performance?
object
Do you think extracurricular activities should be open after 10:30?
object
Do you think everyone is getting equal chance in participating in extracurricular
activities?            object
Do you get exposure in extracurricular activities
object
If yes, select one of the following
object
If No, select the reason
object
dtype: object
```

In [ ]:
```
data.count()
```

Out[ ]:
```
Timestamp
402
Name
402
Department
402
GENDER
402
YEAR
402
Do you follow a Semester or Trimester System?
402
Do you participate in any extra curricular activities?
402
If No,  how do you spend your time after college hours?
235
IF YES, select the following
227
Are you able to balance your academics with extracurricular?
368
How do you rate your time management\n(On the scale of 5)
402
Do you think participation in extracurricular affect your academic performance?
398
Do you think extracurricular activities should be open after 10:30?
398
Do you think everyone is getting equal chance in participating in extracurricular
activities?    394
Do you get exposure in extracurricular activities
381
If yes, select one of the following
217
If No, select the reason
226
dtype: int64
```

In [ ]:
```
#DISPLAYING THE TOP 5 OBSERVATION
data.head()
```

Out[ ]:

	Timestamp	Name	Department	GENDER	YEAR	Do you follow a Semester or Trimester System?	Do you participate in any extra curricular activities?	If No, how do you spend your time after college hours?	I foll
0	2023-10-10 21:20:21.933	Stuty Das	MSC DS	FEMALE	FIRST	Trimester	No	Cooking	
1	2023-10-10 21:32:57.896	Aryan Manchanda	BSC DS	MALE	FIRST	Semester	YES	NaN	S
2	2023-10-10 22:16:09.915	Anju Mathew	LLB	FEMALE	SECOND	Semester	YES	NaN	
3	2023-10-10 22:17:36.971	Gracy Rathaur	LLB	FEMALE	THIRD	Semester	YES	NaN	
4	2023-10-10 22:35:12.016	Sreelakshmi G	BBA	FEMALE	FIRST	Semester	No	Academic works	

In [ ]:
```
#DISPLAYING THE LAST 5 OBSERVATION
data.tail()
```

Out[ ]:

	Timestamp	Name	Department	GENDER	YEAR	Do you follow a Semester or Trimester System?	Do you participate in any extra curricular activities?	If No, how do you spend your time after college hours?	fo
397	2023-10-17 22:30:13.221	Rhea Eleanor Rhine	BBA	FEMALE	THIRD	Semester	No	Hang out with friends	
398	2023-10-17 22:31:36.111	Pariska Nagvenkar	BBA	FEMALE	SECOND	Semester	No	Academic works	
399	2023-10-17 22:33:12.435	Vatsal Sharma	LLB	MALE	SECOND	Semester	YES	NaN	
400	2023-10-17 22:33:20.940	Rachel	MBA	FEMALE	FIRST	Trimester	YES	NaN	
401	2023-10-17 22:34:02.656	Shilpi Singh	BCOM	FEMALE	SECOND	Semester	No	Cooking	

In [ ]:
```
#DISPLAYING THE DATA TYPE AND THE INFORMATION ABOUT THE DATA SET
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 402 entries, 0 to 401
Data columns (total 17 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Timestamp
402 non-null    datetime64[ns]
 1   Name
402 non-null    object
 2   Department
402 non-null    object
 3   GENDER
402 non-null    object
 4   YEAR
402 non-null    object
 5   Do you follow a Semester or Trimester System?
402 non-null    object
 6   Do you participate in any extra curricular activities?
402 non-null    object
 7   If No,  how do you spend your time after college hours?
235 non-null    object
 8   IF YES, select the following
227 non-null    object
 9   Are you able to balance your academics with extracurricular?
368 non-null    object
 10  How do you rate your time management
(On the scale of 5)                                    402 non-null    int64
 11  Do you think participation in extracurricular affect your academic performanc
e?                398 non-null    object
 12  Do you think extracurricular activities should be open after 10:30?
398 non-null    object
 13  Do you think everyone is getting equal chance in participating in extracurric
ular activities?  394 non-null    object
 14  Do you get exposure in extracurricular activities
381 non-null    object
 15  If yes, select one of the following
217 non-null    object
 16  If No, select the reason
226 non-null    object
dtypes: datetime64[ns](1), int64(1), object(15)
memory usage: 53.5+ KB
```

In [ ]:
```
#CHECKING THE DUPLICATION
data.nunique()
```

Out[ ]:
```
Timestamp
402
Name
402
Department
11
GENDER
2
YEAR
5
Do you follow a Semester or Trimester System?
2
Do you participate in any extra curricular activities?
2
If No,  how do you spend your time after college hours?
4
IF YES, select the following
5
Are you able to balance your academics with extracurricular?
3
How do you rate your time management\n(On the scale of 5)
5
Do you think participation in extracurricular affect your academic performance?
3
Do you think extracurricular activities should be open after 10:30?
3
Do you think everyone is getting equal chance in participating in extracurricular
activities?      3
Do you get exposure in extracurricular activities
3
If yes, select one of the following
3
If No, select the reason
5
dtype: int64
```

In [ ]:
```
#MISSING VALUE CALCULATION
data.isnull().sum()
```

Out[ ]:
```
Timestamp
0
Name
0
Department
0
GENDER
0
YEAR
0
Do you follow a Semester or Trimester System?
0
Do you participate in any extra curricular activities?
0
If No,  how do you spend your time after college hours?
167
IF YES, select the following
175
Are you able to balance your academics with extracurricular?
34
How do you rate your time management\n(On the scale of 5)
0
Do you think participation in extracurricular affect your academic performance?
4
Do you think extracurricular activities should be open after 10:30?
4
Do you think everyone is getting equal chance in participating in extracurricular
activities?      8
Do you get exposure in extracurricular activities
21
If yes, select one of the following
185
If No, select the reason
176
dtype: int64
```

In [ ]:
```
#PERCENTAGE OF MISSING VALUE CALCULATION
data.isnull().sum()/(len(data))*100
```

Out[ ]:
```
Timestamp
0.000000
Name
0.000000
Department
0.000000
GENDER
0.000000
YEAR
0.000000
Do you follow a Semester or Trimester System?
0.000000
Do you participate in any extra curricular activities?
0.000000
If No,  how do you spend your time after college hours?
41.542289
IF YES, select the following
43.532338
Are you able to balance your academics with extracurricular?
8.457711
How do you rate your time management\n(On the scale of 5)
0.000000
Do you think participation in extracurricular affect your academic performance?
0.995025
Do you think extracurricular activities should be open after 10:30?
0.995025
Do you think everyone is getting equal chance in participating in extracurricular
activities?     1.990050
Do you get exposure in extracurricular activities
5.223881
If yes, select one of the following
46.019900
If No, select the reason
43.781095
dtype: float64
```

Statistics Summary

In [ ]:
```
data.describe()
```

Out[ ]:

	How do you rate your time management\n(On the scale of 5)
count	402.000000
mean	2.845771
std	0.856889
min	1.000000
25%	2.000000
50%	3.000000
75%	3.000000
max	5.000000

In [ ]:
```
data.describe(include="all")
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_60320\3772400700.py:1: FutureWarning: Tre
ating datetime data as categorical rather than numeric in `.describe` is deprecate
d and will be removed in a future version of pandas. Specify `datetime_is_numeric=
True` to silence this warning and adopt the future behavior now.
  data.describe(include="all")
```

Out[ ]:

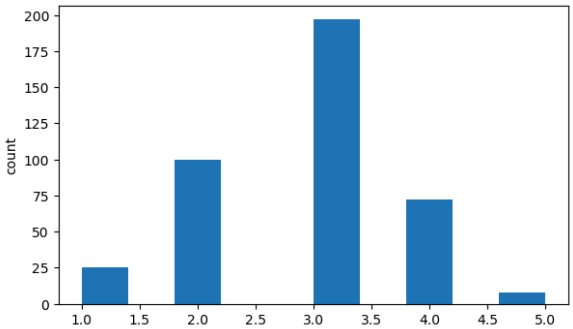
	Timestamp	Name	Department	GENDER	YEAR	Do you follow a Semester or Trimester System?	Do you participate in any extra curricular activities?	If No, how do you spend your time after college hours?	fol
count	402	402	402	402	402	402	402	235	
unique	402	402	11	2	5	2	2	4	
top	2023-10-10 21:20:21.933000	Stuty Das	LLB	MALE	FIRST	Semester	YES	Academic works	
freq	1	1	118	209	181	305	215	132	
first	2023-10-10 21:20:21.933000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
last	2023-10-17 22:34:02.656000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

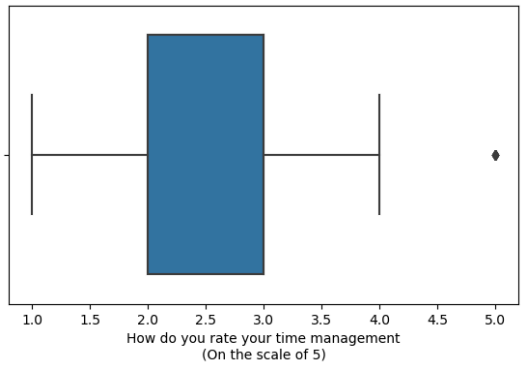
DISTRIBUTION

In [ ]:
```
#HISTROGRAM CHART FOR TIME MANAGEMENT
import numpy as np


for col in data.columns:
    if data[col].dtype in ['int64', 'float64']:  # Check if the data type is numeri
        print(col)
        print('Skew:', round(data[col].skew(), 2))
        plt.figure(figsize=(15, 4))
        plt.subplot(1, 2, 1)
        data[col].hist(grid=False)
        plt.ylabel('count')
        plt.subplot(1, 2, 2)
        sns.boxplot(x=data[col])
        plt.show()
```

```
How do you rate your time management
(On the scale of 5)
Skew: -0.1
```





From the graph above, we concluded that most of the students rate their time management skills as not so appropriate.

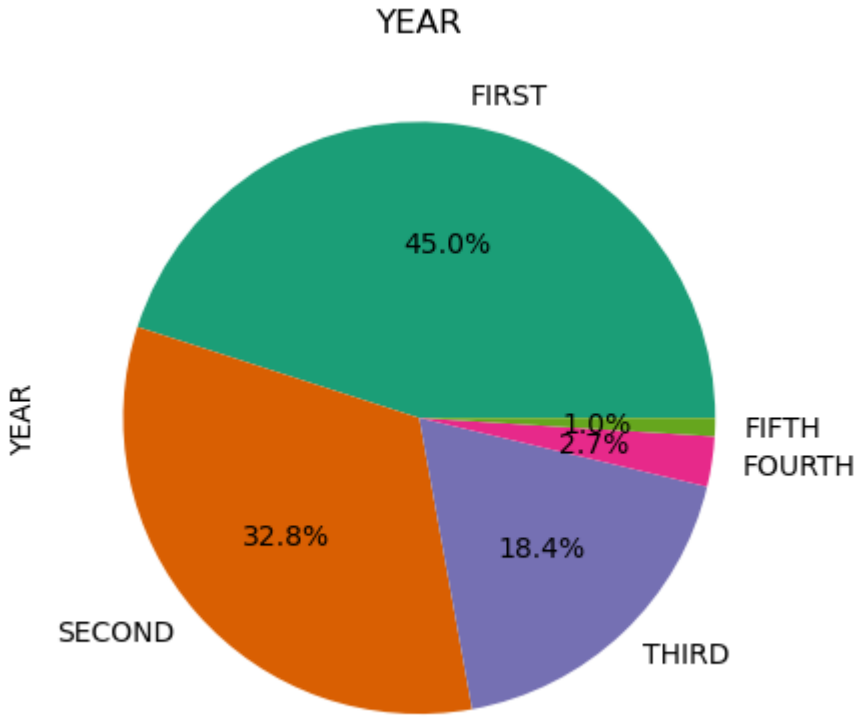
CATEGORICAL DISTRIBUTION

1.YEARWISE DISTRIBUTION

In [ ]:
```
#PIE CHART FOR YEARWISE DISTRIBUTION
import numpy as np

def categorical_histogram(df, colname, figscale=1, mpl_palette_name='Dark2'):
    from matplotlib import pyplot as plt
    import seaborn as sns
    group_counts = df[colname].value_counts()
    group_counts.plot(kind='pie', colors=sns.color_palette(mpl_palette_name), figsi
    plt.gca().set_aspect("equal")  # Ensure the pie chart is a circle
    plt.title(colname)
    return

categorical_histogram(data, 'YEAR')
plt.show()
```



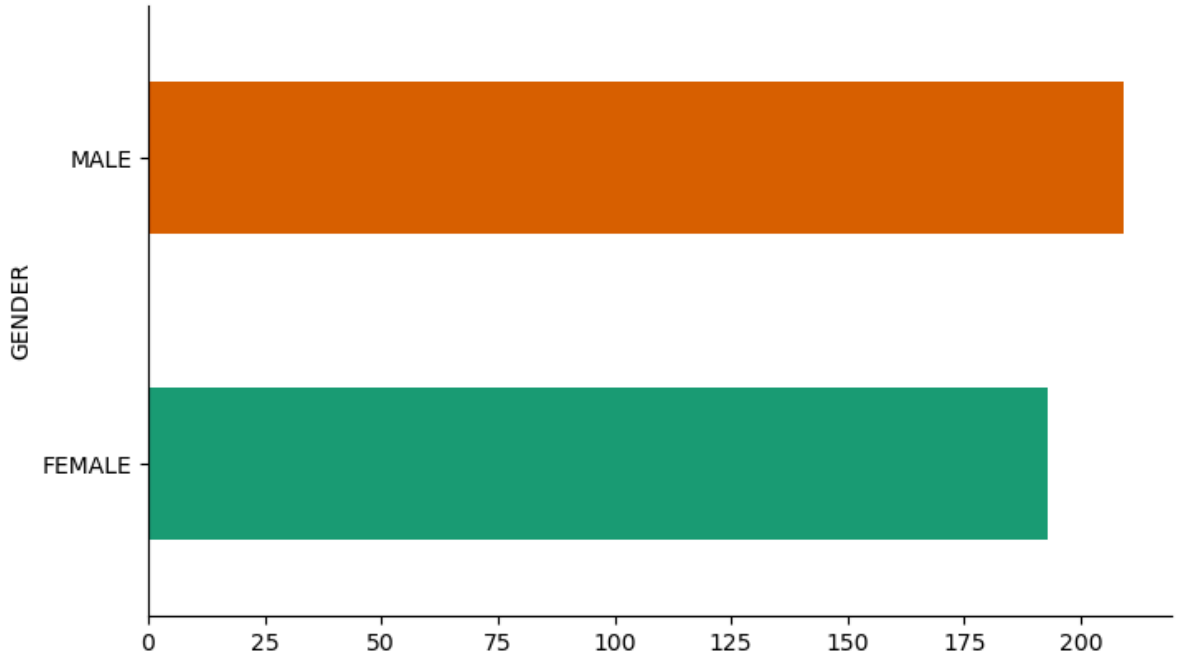
It is quite evident that the first years participation in the survey was the maximum as compared to the others, while students of fifth year were the least.

2.GENDERWISE DISTRIBUTION

In [ ]:
```
import numpy as np

def categorical_histogram(df, colname, figscale=1, mpl_palette_name='Dark2'):
    from matplotlib import pyplot as plt
    import seaborn as sns
    df.groupby(colname).size().plot(kind='barh', color=sns.palettes.mpl_palette(mpl
    plt.gca().spines[['top', 'right',]].set_visible(False)
    return plt

chart = categorical_histogram(data, 'GENDER')
plt.show()  # Show the chart
```



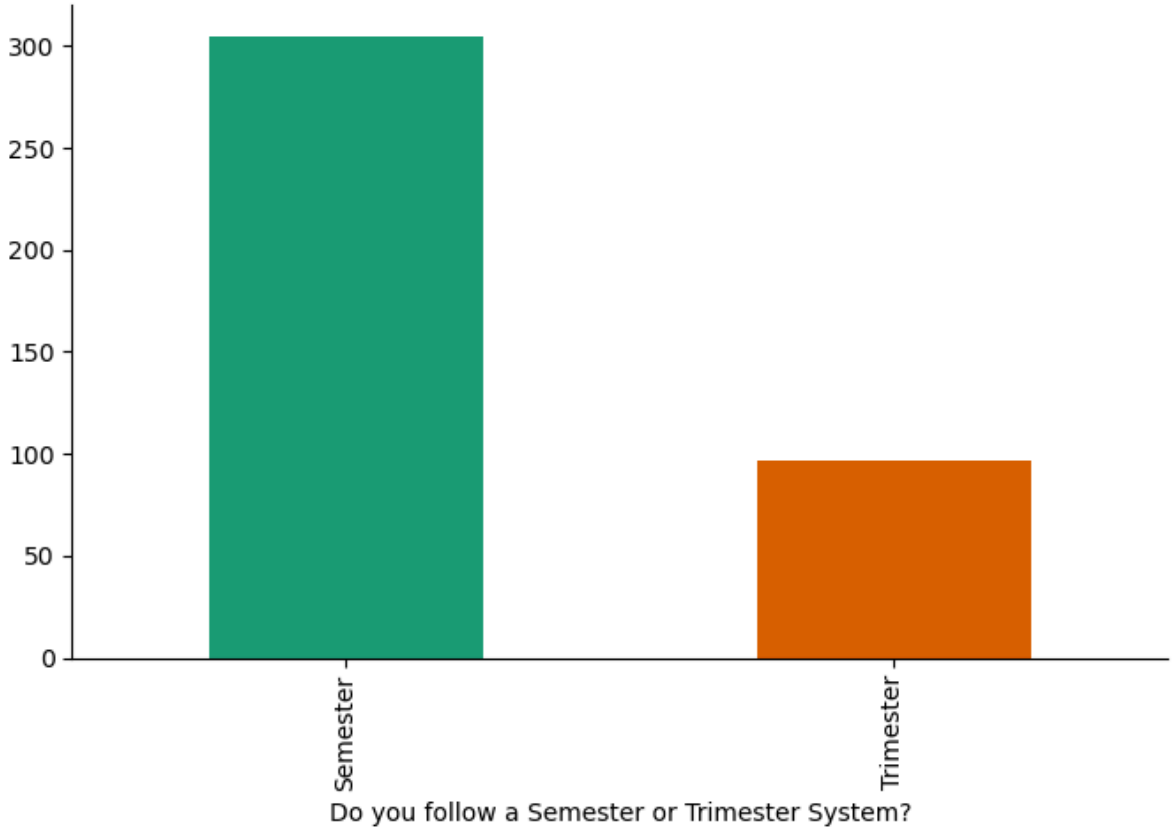
Although, the participation from both males and females are impressive, but still responses from males have an upper edge in comparison to females.

3.TRIMESTERWISE/SEMESTERWISE DISTRIBUTION

In [ ]:
```
import numpy as np

def categorical_histogram(df, colname, figscale=1, mpl_palette_name='Dark2'):
    from matplotlib import pyplot as plt
    import seaborn as sns
    df.groupby(colname).size().plot(kind='bar', color=sns.palettes.mpl_palette(mpl_
    plt.gca().spines[['top', 'right',]].set_visible(False)
    return plt

chart = categorical_histogram(data, 'Do you follow a Semester or Trimester System?'
plt.show()  # Show the bar graph
```



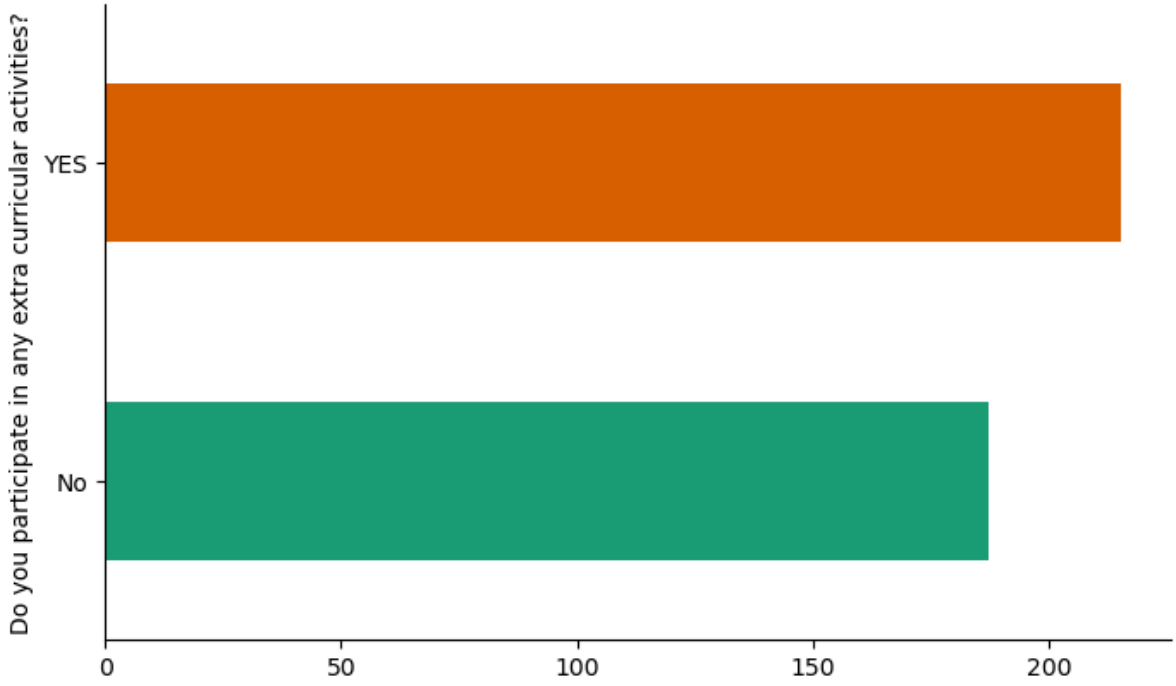
It's clearly evident that most of the courses in Christ University has a semester based pattern.

4.PARTICIPATION IN EXTRA CURRICULAR ACTIVITIES

In [ ]:
```
import numpy as np


def categorical_histogram(df, colname, figscale=1, mpl_palette_name='Dark2'):
  from matplotlib import pyplot as plt
  import seaborn as sns
  df.groupby(colname).size().plot(kind='barh', color=sns.palettes.mpl_palette(mpl_p
  plt.gca().spines[['top', 'right',]].set_visible(False)
  return

chart = categorical_histogram(data, *['Do you participate in any extra curricular a
chart
```



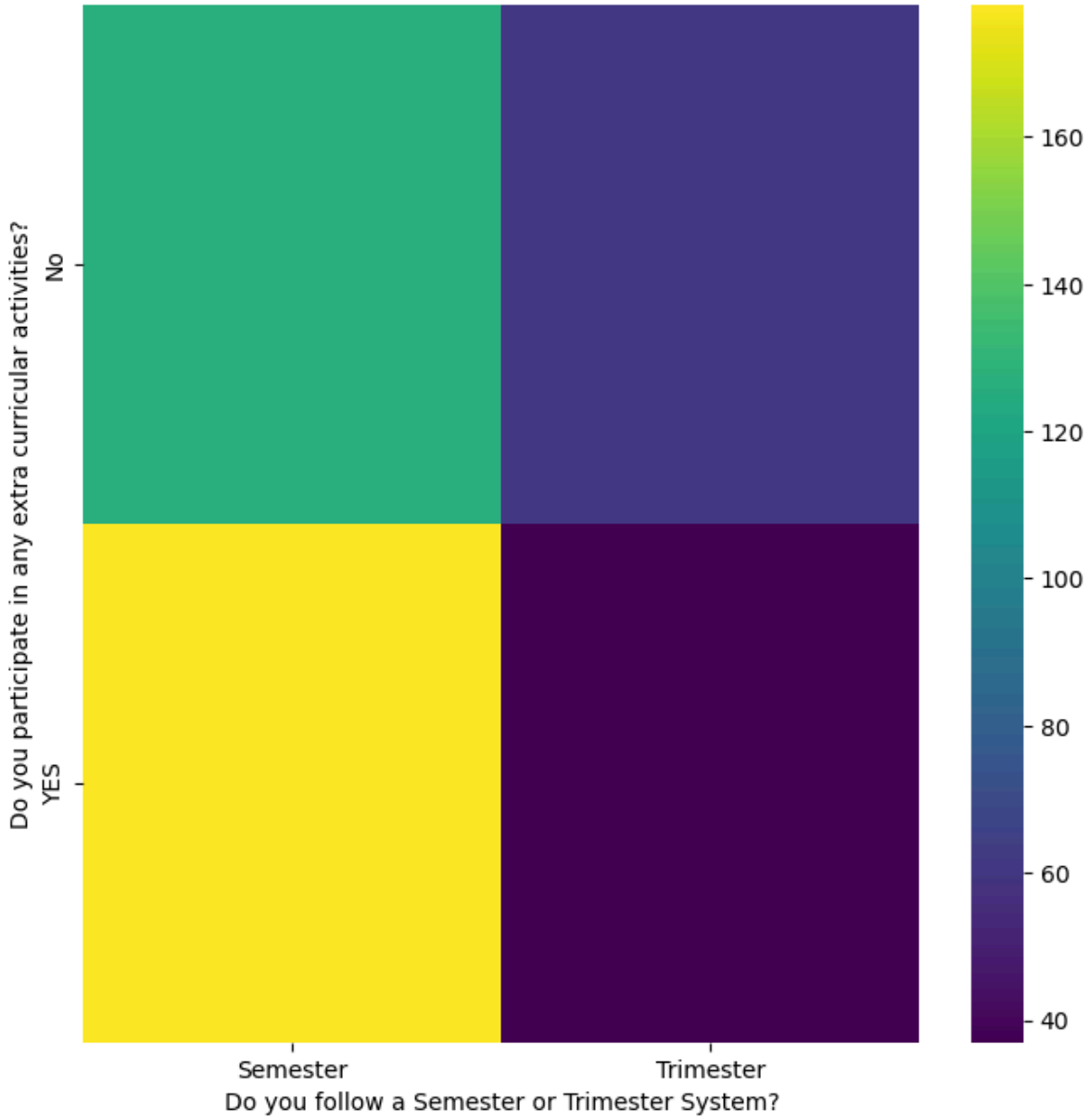
It's surprising to conclude that there is no significant difference between the no. students who partici[ate and not participate in extra curricular activities.]

1. SEMESTER/TRIMESTERWISE PARTICIPATION IN EXTRA-CURRICULAR ACTIVITIES

In [ ]:
```
import numpy as np

def heatmap(df, x_colname, y_colname, figscale=1, mpl_palette_name='viridis'):
  from matplotlib import pyplot as plt
  import seaborn as sns
  import pandas as pd
  plt.subplots(figsize=(8 * figscale, 8 * figscale))
  df_2dhist = pd.DataFrame({
      x_label: grp[y_colname].value_counts()
      for x_label, grp in df.groupby(x_colname)
  })
  sns.heatmap(df_2dhist, cmap=mpl_palette_name)
  plt.xlabel(x_colname)
  plt.ylabel(y_colname)
  return

chart = heatmap(data, *['Do you follow a Semester or Trimester System?', 'Do you pa
chart
```



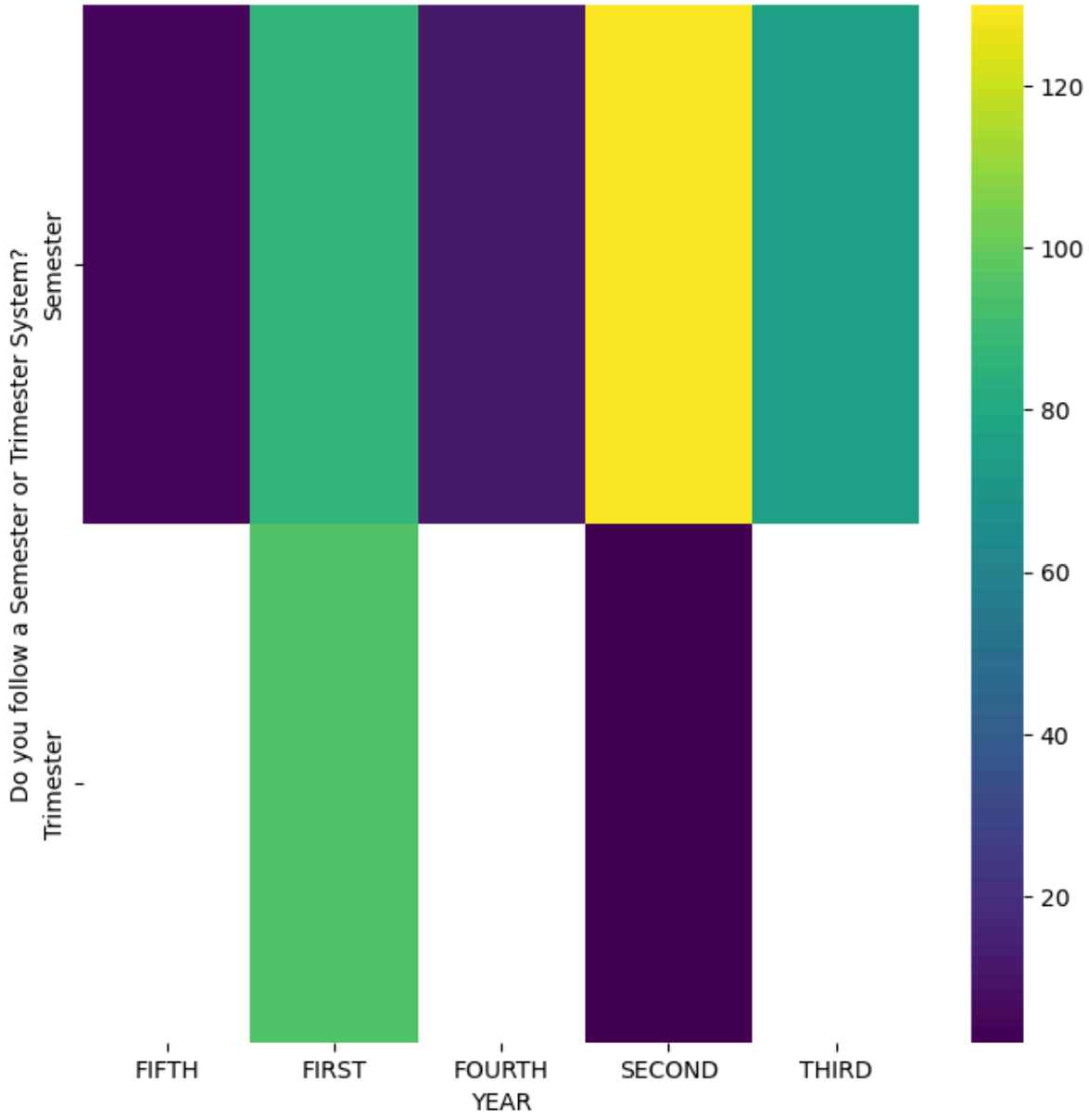
This heatmap indicates that the students that have trimesters are unable to participate in co-curricular activites while the semester students are actively participating in such activites, which indirectly hints at the insufficient time with the former students.

6.HEATMAP OF THE TRIMESTER/SEMESTER BASED ON YEARWISE

In [ ]:
```
import numpy as np

def heatmap(df, x_colname, y_colname, figscale=1, mpl_palette_name='viridis'):
  from matplotlib import pyplot as plt
  import seaborn as sns
  import pandas as pd
  plt.subplots(figsize=(8 * figscale, 8 * figscale))
  df_2dhist = pd.DataFrame({
      x_label: grp[y_colname].value_counts()
      for x_label, grp in df.groupby(x_colname)
  })
  sns.heatmap(df_2dhist, cmap=mpl_palette_name)
  plt.xlabel(x_colname)
  plt.ylabel(y_colname)
  return
```

```
chart = heatmap(data, *['YEAR', 'Do you follow a Semester or Trimester System?'], *
chart
```



This chart is segregating the semester/trimester depedending on the year the students are currently enrolled in. For instance: the third,fourth,fifth year don't have any trimesters.

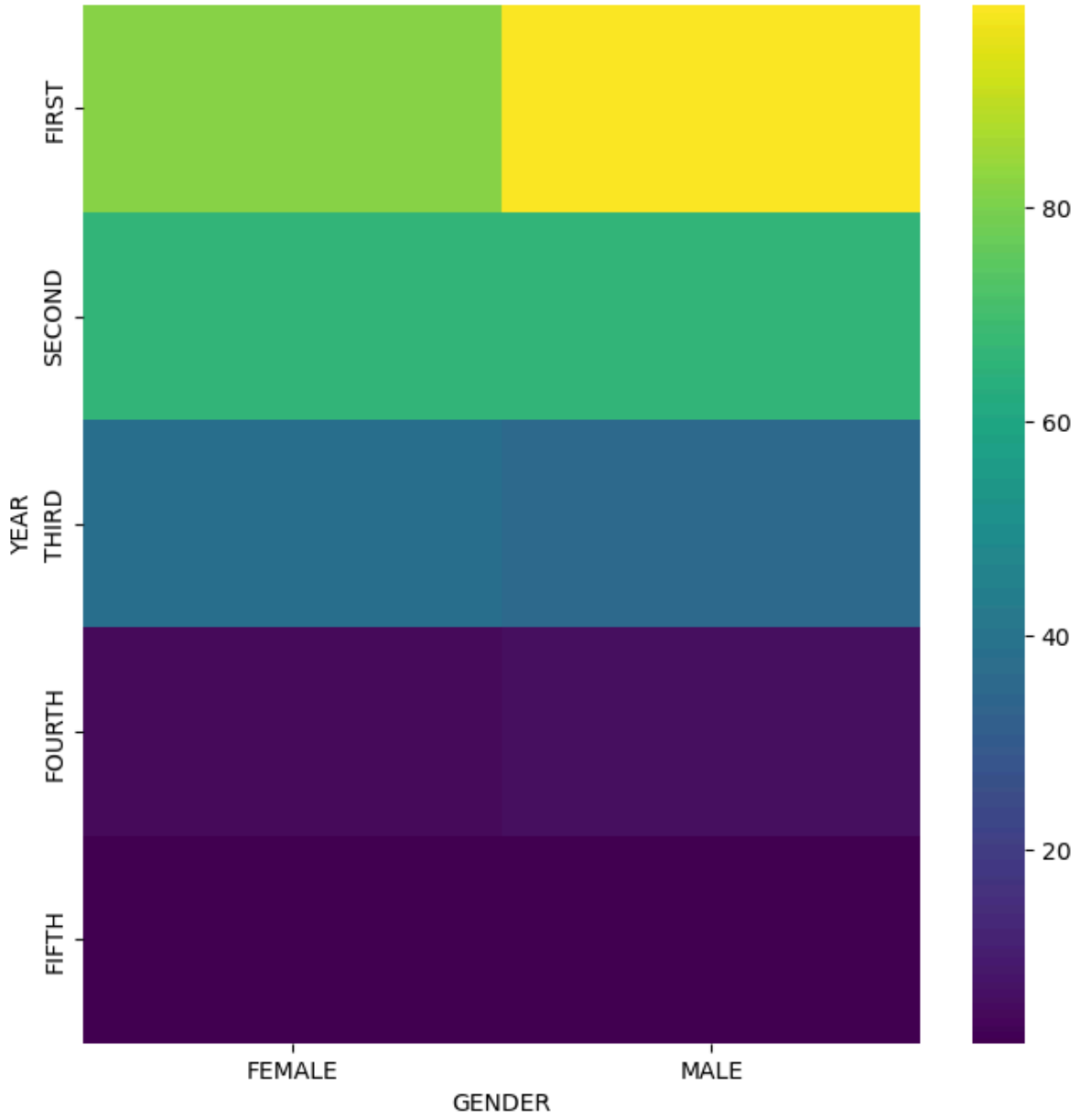
7.HEATMAP OF THE YEARWISE BASED ON GENDER

In [ ]:

```
import numpy as np

def heatmap(df, x_colname, y_colname, figscale=1, mpl_palette_name='viridis'):
  from matplotlib import pyplot as plt
  import seaborn as sns
  import pandas as pd
  plt.subplots(figsize=(8 * figscale, 8 * figscale))
  df_2dhist = pd.DataFrame({
      x_label: grp[y_colname].value_counts()
      for x_label, grp in df.groupby(x_colname)
  })
  sns.heatmap(df_2dhist, cmap=mpl_palette_name)
  plt.xlabel(x_colname)
  plt.ylabel(y_colname)
  return
```

```
chart = heatmap(data, *['GENDER', 'YEAR'], **{})
chart
```



Males from the first year have responded the most, while females from the fifth year have responded the least.