

CPE232: Data Models

Portion 2: Midterm Exam Coding

Sections: A, B, RC

```
In [ ]: # Run this cell if using Google Colab
from google.colab import drive
drive.mount('/content/drive')
```

ข้อมูลในชุดข้อมูล student_spending.csv ที่ให้มาประกอบด้วยข้อมูลที่เกี่ยวข้องกับการใช้จ่ายของนักเรียน การฝึกของคุณคือการทำการวิเคราะห์ข้อมูลเชิงสำรวจ (EDA) ตามภารกิจย่อยห้าประการดังต่อไปนี้

Subtask #1: รู้จักกับชุดข้อมูล

1.1) ศึกษาภาพรวมของชุดข้อมูล (Total points = 3)

[3 points] Display information of the data: size, shape, and number of dimensions. You can use any libraries of your choice (e.g. Numpy, Pandas).

แสดงรายละเอียดต่อไปนี้ของชุดข้อมูล: ขนาด, รูปร่าง, และจำนวนมิติ นักศึกษาสามารถใช้ไลบรารีใดก็ได้ตามต้องการ (เช่น Numpy, Pandas)

```
In [2]: import pandas as pd
df = pd.read_csv('student_spending.csv') #TODO: update the path and filename at
```

```
In [6]: # Write your code here
print(df.size)
print(df.shape)
print(df.ndim)
```

```
18000
(1000, 18)
2
```

1.2) ศึกษาสถิติของชุดข้อมูลนี้เพิ่มเติม (Total point = 15)

Use the command below:

ใช้คำสั่งต่อไปนี้:

```
df.info()
```

```
In [7]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 18 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Unnamed: 0                            1000 non-null   int64
1   age                                    1000 non-null   int64
2   gender                                1000 non-null   object
3   year_in_school                        1000 non-null   object
4   major                                 1000 non-null   object
5   monthly_income                       1000 non-null   int64
6   financial_aid                        1000 non-null   int64
7   tuition                              1000 non-null   int64
8   housing                              1000 non-null   int64
9   food                                  1000 non-null   int64
10  transportation                        1000 non-null   int64
11  books_supplies                        1000 non-null   int64
12  entertainment                         1000 non-null   int64
13  personal_care                         1000 non-null   int64
14  technology                            1000 non-null   int64
15  health_wellness                      1000 non-null   int64
16  miscellaneous                         1000 non-null   int64
17  preferred_payment_method             1000 non-null   object
dtypes: int64(14), object(4)
memory usage: 140.8+ KB
```

[2 points] Obtain the following information and provide your answers:

- Number of columns of the type *Integer*
- Number of columns of the type *String*

หาค่าต่อไปนี้จากชุดข้อมูลและระบุคำตอบ:

- จำนวนคอลัมน์ที่เป็น *Integer*
- จำนวนคอลัมน์ที่เป็น *String*

ANS:

- จำนวนคอลัมน์ที่เป็น *Integer* : 14
- จำนวนคอลัมน์ที่เป็น *String* : 4

[1 point] Display the first 6 rows.

แสดง 6 แถวแรกของข้อมูล

```
In [10]: # Write your code here
df.head(6)
```

Out[10]:

	Unnamed: 0	age	gender	year_in_school	major	monthly_income	financial_aid
0	0	19	Non-binary	Freshman	Psychology	958	270
1	1	24	Female	Junior	Economics	1006	875
2	2	24	Non-binary	Junior	Economics	734	928
3	3	23	Female	Senior	Computer Science	617	265
4	4	20	Female	Senior	Computer Science	810	522
5	5	25	Non-binary	Sophomore	Computer Science	523	790

[1 point] Display the last 10 rows.

แสดง 10 แถวสุดท้ายของข้อมูล

In [11]:

```
# Write your code here
df.tail(10)
```

Out[11]:


	Unnamed: 0	age	gender	year_in_school	major	monthly_income	financial_aid
990	990	20	Non-binary	Senior	Psychology	1412	155
991	991	24	Non-binary	Junior	Psychology	1391	259
992	992	20	Male	Freshman	Economics	1293	672
993	993	20	Male	Freshman	Psychology	1380	594
994	994	22	Male	Senior	Psychology	764	286
995	995	22	Female	Senior	Biology	1346	520
996	996	19	Female	Senior	Biology	1407	560
997	997	20	Male	Junior	Economics	957	393
998	998	22	Non-binary	Senior	Economics	1174	612
999	999	24	Non-binary	Sophomore	Computer Science	541	640

[1 point] Descriptive statistics of *ALL attributes*สถิติเชิงพรรณนาของ *ทุกๆคุณลักษณะ*

In [12]: `# Write your code here`
`df.describe()`

Out[12]:

	Unnamed: 0	age	monthly_income	financial_aid	tuition	housing
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	499.500000	21.675000	1020.650000	504.771000	4520.395000	696.006000
std	288.819436	2.322664	293.841161	287.092575	860.657944	171.218600
min	0.000000	18.000000	501.000000	0.000000	3003.000000	401.000000
25%	249.750000	20.000000	770.750000	261.000000	3779.750000	538.750000
50%	499.500000	22.000000	1021.000000	513.000000	4547.500000	704.500000
75%	749.250000	24.000000	1288.250000	751.500000	5285.000000	837.250000
max	999.000000	25.000000	1500.000000	1000.000000	6000.000000	1000.000000



[1 point] Descriptive statistics of *one selected attribute*: `tuition`

สถิติเชิงพรรณนาของ *หนึ่งคุณลักษณะ*: `tuition`

In [13]: `# Write your code here`
`df['tuition'].describe()`

Out[13]:

```

count    1000.000000
mean     4520.395000
std       860.657944
min       3003.000000
25%      3779.750000
50%      4547.500000
75%      5285.000000
max       6000.000000
Name: tuition, dtype: float64

```

[4 points] Descriptive statistics of *four selected attribute*: `age`, `housing`, `food`, `transportation`

สถิติเชิงพรรณนาของ *สี่คุณลักษณะ*: `age`, `housing`, `food`, `transportation`

In [14]: `# Write your code here`
`df[['age', 'housing', 'food', 'transportation']].describe()`

Out[14]:

	age	housing	food	transportation
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	21.675000	696.006000	252.642000	124.637000
std	2.322664	171.218620	86.949606	43.557990
min	18.000000	401.000000	100.000000	50.000000
25%	20.000000	538.750000	175.000000	88.000000
50%	22.000000	704.500000	255.000000	123.000000
75%	24.000000	837.250000	330.000000	162.250000
max	25.000000	1000.000000	400.000000	200.000000

[5 points] Display the number of occurrences of each unique value in ALL *non-integer* columns.

แสดงจำนวนครั้งที่แต่ละค่าที่ไม่ซ้ำกันปรากฏในทุกคอลัมน์ที่ ไม่ใช่จำนวนเต็ม

Hint: Example of the output may look like the following for the column named `genre`.

ตัวอย่างของผลลัพธ์อาจคล้ายผลต่อไปนี้สำหรับคอลัมน์ที่มีชื่อว่า `genre`

```
genre
pop      550
jazz     234
rock     294
country  146
Name: count, dtype: int64
```

In [135...

```
# Write your code here
# gender
print(df['gender'].unique())
print('Non-binary',(df[df['gender'] == 'Non-binary'].count().sum())/17)
print('Female',(df[df['gender'] == 'Female'].count().sum())/17)
print('Male',(df[df['gender'] == 'Male'].count().sum())/17)
print('dtype: int64')
```

```
['Non-binary' 'Female' 'Male']
Non-binary 321.0
Female 323.0
Male 356.0
dtype: int64
```

In [134...

```
# year_in_school
print(df['year_in_school'].unique())
print('Freshman',(df[df['year_in_school'] == 'Freshman'].count().sum())/17)
print('Junior',(df[df['year_in_school'] == 'Junior'].count().sum())/17)
print('Senior',(df[df['year_in_school'] == 'Senior'].count().sum())/17)
print('Sophomore',(df[df['year_in_school'] == 'Sophomore'].count().sum())/17)
print('dtype: int64')
```

```
['Freshman' 'Junior' 'Senior' 'Sophomore']
Freshman 253.0
Junior 247.0
Senior 254.0
Sophomore 246.0
dtype: int64
```

In [122...

```
# major
print(df['major'].unique())
print('Psychology', (df[df['major'] == 'Psychology'].count().sum())/17)
print('Economics', (df[df['major'] == 'Economics'].count().sum())/17)
print('Computer Science', (df[df['major'] == 'Computer Science'].count().sum())/17)
print('Engineering', (df[df['major'] == 'Engineering'].count().sum())/17)
print('Biology', (df[df['major'] == 'Biology'].count().sum())/17)
print('dtype: int64')
```

```
['Psychology' 'Economics' 'Computer Science' 'Engineering' 'Biology']
Psychology 184.0
Economics 204.0
Computer Science 192.0
Engineering 192.0
Biology 228.0
dtype: int64
```

In [133...

```
# preferred_payment_method
print(df['preferred_payment_method'].unique())
print('Credit/Debit Card', (df[df['preferred_payment_method'] == 'Credit/Debit Ca
print('Cash', (df[df['preferred_payment_method'] == 'Cash'].count().sum())/17)
print('Mobile Payment App', (df[df['preferred_payment_method'] == 'Mobile Payment
print('dtype: int64')
```

```
['Credit/Debit Card' 'Cash' 'Mobile Payment App']
Credit/Debit Card 340.0
Cash 310.0
Mobile Payment App 350.0
dtype: int64
```

Subtask #2: ตรวจสอบข้อมูล

2.1) ตรวจสอบข้อมูลนักศึกษาโดยกำหนดเงื่อนไข (Total points = 14)

[4 points] Display *the first 10 rows* of records that meet the condition: *Students with a major in Computer Science with a spending on technology more than 100.*

แสดง 10 แถวแรก ของข้อมูลที่ตรงตามเงื่อนไขต่อไปนี้: *นักศึกษาที่เรียนสาขาวิทยาการคอมพิวเตอร์และมีค่าใช้จ่ายด้านเทคโนโลยีมากกว่า 100*

In [43]:

```
# Write your code here
# Students with a major in Computer Science with a spending on technology more t
df[(df['major']=='Computer Science') & (df['technology'] > 100)].head(10)
```

Out[43]:

	Unnamed: 0	age	gender	year_in_school	major	monthly_income	financial_aid
3	3	23	Female	Senior	Computer Science	617	265
5	5	25	Non-binary	Sophomore	Computer Science	523	790
8	8	22	Non-binary	Senior	Computer Science	1402	248
9	9	18	Female	Junior	Computer Science	1423	74
32	32	24	Non-binary	Junior	Computer Science	522	555
37	37	23	Non-binary	Senior	Computer Science	1309	265
45	45	18	Male	Freshman	Computer Science	929	348
52	52	19	Male	Senior	Computer Science	669	660
56	56	24	Non-binary	Freshman	Computer Science	854	700
71	71	21	Non-binary	Sophomore	Computer Science	1235	805



[1 point] How many records are there that match the above condition?

มีทั้งหมดจำนวนกี่รายการที่ตรงกับเงื่อนไขข้างต้น?

Ans: 10 รายการ

[8 points] Display *the first 10 rows* of records that meet the condition: *Male Sophomore students with monthly income ranging from 600 to 1000.*

แสดง 10 แถวแรก ของข้อมูลที่ตรงตามเงื่อนไขต่อไปนี้: นักศึกษาชั้นปีที่สอง (Sophomore) ที่เป็นเพศชาย และมีรายได้ต่อเดือนอยู่ในช่วง 600 ถึง 1000

In [53]:

```
# Write your code here
# Male Sophomore students with monthly income ranging from 600 to 1000.
df[(df['year_in_school']=='Sophomore') & (df['gender'] == 'Male') & (df['monthly
```

Out[53]:

	Unnamed: 0	age	gender	year_in_school	major	monthly_income	financial_aid
12	12	21	Male	Sophomore	Economics	719	540
41	41	25	Male	Sophomore	Economics	804	140
76	76	22	Male	Sophomore	Computer Science	983	862
89	89	23	Male	Sophomore	Economics	800	933
108	108	20	Male	Sophomore	Computer Science	965	322
110	110	22	Male	Sophomore	Economics	970	553
126	126	20	Male	Sophomore	Biology	836	620
148	148	24	Male	Sophomore	Computer Science	897	220
157	157	20	Male	Sophomore	Economics	901	115
175	175	23	Male	Sophomore	Biology	963	871



[1 point] How many records are there that match the above condition?

มีทั้งหมดจำนวนกี่รายการที่ตรงกับเงื่อนไขข้างต้น?

Ans: 10 รายการ

2.2) ตรวจสอบว่ามีค่าที่ตกหล่นไปหรือไม่ (Total point = 1)

[1 point] How many attributes contain missing values?

มีคุณลักษณะ (attribute) กี่รายการที่มีค่าที่ตกหล่นไป

ANS: ไม่มี missing value

In [54]: `df.isnull().sum()`


```
Out[54]: Unnamed: 0      0
         age           0
         gender        0
         year_in_school 0
         major         0
         monthly_income 0
         financial_aid  0
         tuition       0
         housing       0
         food          0
         transportation 0
         books_supplies 0
         entertainment 0
         personal_care  0
         technology     0
         health_wellness 0
         miscellaneous  0
         preferred_payment_method 0
         dtype: int64
```

Subtask #3: จัดเตรียมข้อมูล

3.1 ลบคอลัมน์ที่ไม่จำเป็นออกจากชุดข้อมูล (Total points = 2)

[2 points] ลบคอลัมน์ "Unnamed: 0" ออกจากชุดข้อมูล

(Note: this column must no longer appear when displaying the dataframe again later;
หมายเหตุ: คอลัมน์นี้ต้องไม่ปรากฏอีกเมื่อแสดง DataFrame ในภายหลัง)

```
In [58]: # Write your code here
         df = df.drop(columns=['Unnamed: 0'])
```

```
In [60]: df.head()
```

```
Out[60]:
```

	age	gender	year_in_school	major	monthly_income	financial_aid	tuition	hou
0	19	Non-binary	Freshman	Psychology	958	270	5939	
1	24	Female	Junior	Economics	1006	875	4908	
2	24	Non-binary	Junior	Economics	734	928	3051	
3	23	Female	Senior	Computer Science	617	265	4935	
4	20	Female	Senior	Computer Science	810	522	3887	

3.2 สร้างคอลัมน์ใหม่ (Total points = 23)

[5 points] Create a new column and name it `major_expense`. This column contains values that are the sum of housing, food, and transportation.

สร้างคอลัมน์ใหม่ชื่อ `major_expense` โดยมีค่าที่ได้จากผลรวมของค่าใช้จ่ายด้านที่อยู่อาศัย อาหาร และการเดินทาง

In [138...

```
# Write your code here
df = df.assign(major_expense = df['housing']+df['food']+df['transportation'])
df
```

Out[138...

	age	gender	year_in_school	major	monthly_income	financial_aid	tuition	h
0	19	Non-binary	Freshman	Psychology	958	270	5939	
1	24	Female	Junior	Economics	1006	875	4908	
2	24	Non-binary	Junior	Economics	734	928	3051	
3	23	Female	Senior	Computer Science	617	265	4935	
4	20	Female	Senior	Computer Science	810	522	3887	
...
995	22	Female	Senior	Biology	1346	520	3688	
996	19	Female	Senior	Biology	1407	560	3380	
997	20	Male	Junior	Economics	957	393	3497	
998	22	Non-binary	Senior	Economics	1174	612	3649	
999	24	Non-binary	Sophomore	Computer Science	541	640	5965	

1000 rows × 18 columns



[8 points] Create another new column and name it `major_expense_ratio` which is based on the following formula:

สร้างคอลัมน์ใหม่ชื่อ `major_expense_ratio` โดยอิงจากสูตรการคำนวณต่อไปนี้

$$\text{major_expense_ratio} = (\text{major_expense} * 100) / \text{monthly_income}$$

In [140...

```
# Write your code here
df.assign(major_expense_ratio = (df['major_expense']*100)/df['monthly_income'])
```

Out[140...

	age	gender	year_in_school	major	monthly_income	financial_aid	tuition	h
0	19	Non-binary	Freshman	Psychology	958	270	5939	
1	24	Female	Junior	Economics	1006	875	4908	
2	24	Non-binary	Junior	Economics	734	928	3051	
3	23	Female	Senior	Computer Science	617	265	4935	
4	20	Female	Senior	Computer Science	810	522	3887	
...	
995	22	Female	Senior	Biology	1346	520	3688	
996	19	Female	Senior	Biology	1407	560	3380	
997	20	Male	Junior	Economics	957	393	3497	
998	22	Non-binary	Senior	Economics	1174	612	3649	
999	24	Non-binary	Sophomore	Computer Science	541	640	5965	

1000 rows × 19 columns



[10 points] According to the results in previous cell(s), do most students experience financial difficulties as a result of exceeding their monthly income? What is the percentage of those who experience financial difficulties and those who do not? Show your work and analysis below.

จากผลลัพธ์ในเซลล์ก่อนหน้านี้ นักศึกษาส่วนใหญ่ประสบปัญหาทางการเงินเนื่องจากใช้จ่ายเกินรายได้ต่อเดือนหรือไม่? คำนวณเปอร์เซ็นต์ของนักศึกษาที่ประสบปัญหาทางการเงินและนักศึกษาที่ไม่ประสบปัญหา พร้อมบรรยายผลการวิเคราะห์

ANS: ใช่ เปอร์เซนต์ของนักศึกษาส่วนใหญ่ที่ประสบปัญหาทางการเงิน

Subtask #4: สร้างแผนภาพ (Visualizations)

In [149...

```
import matplotlib.pyplot as plt
```

4.1 วิเคราะห์สาขาวิชาต่างๆ (Total points = 10)

[5 points] Create a *pie chart* to demonstrate unique values of the attribute `major`. In your visualization, also display chart title, percentage of distribution, and a legend.

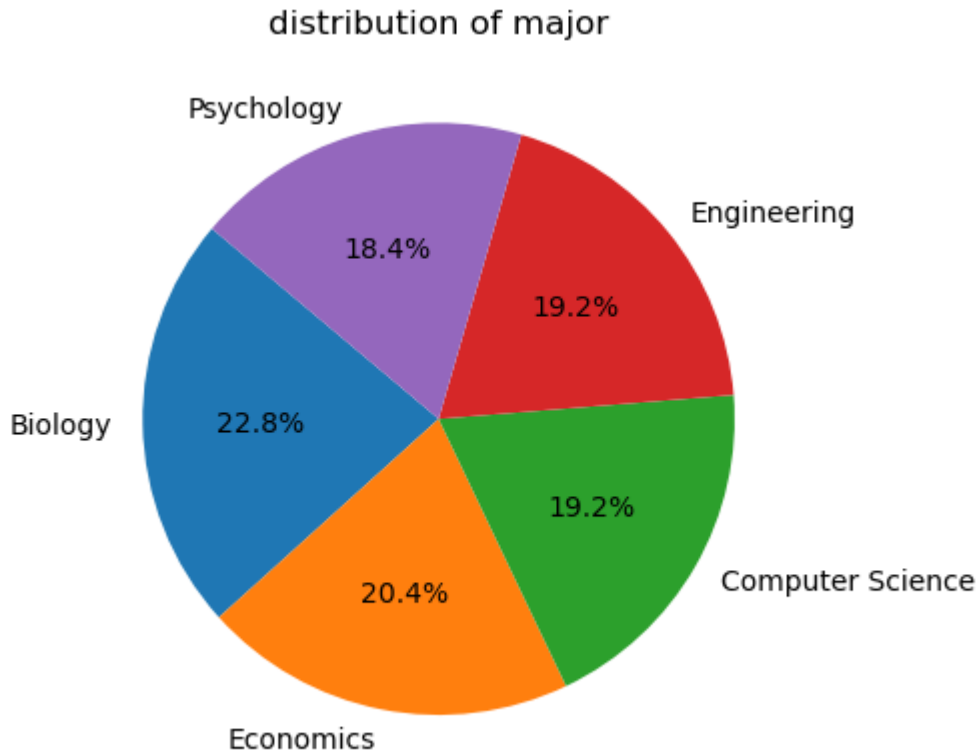
Use a method `.unique()` method to obtain unique values in the attribute.

สร้าง แผนภูมิวงกลม (pie chart) เพื่อแสดงค่าที่ไม่ซ้ำกันของคุณลักษณะ **major** ในการแสดงผล ให้แสดงชื่อแผนภูมิ, เปอร์เซ็นต์การกระจาย, และคำอธิบายสัญลักษณ์ (legend)

ใช้เมธอด `.unique()` เพื่อดึงค่าที่ไม่ซ้ำกันของคุณลักษณะนั้น

In [177...

```
# Write your code here
plt.pie(df['major'].value_counts(), labels=df['major'].value_counts().index, autop
plt.title('distribution of major')
plt.show()
```



[5 points] Describe this visualization in your own words. What information does it convey?

อธิบายแผนภูมิที่สร้างขึ้น แผนภูมินี้สื่อถึงข้อมูลอะไร?

ANS: แสดงสัดส่วนจำนวนคนในคณะต่างๆ ประกอบด้วย Psychology :18.4%, Economics: 20.4%, Computer Science : 19.2%, Engineering : 19.2%, Biology 22.8%

4.2 ศึกษาการกระจายตัวในข้อมูล (Total points = 24)

[7 points] Create a bar chart to demonstrate the distribution of gender. In your visualization, also display the chart title and data labels.

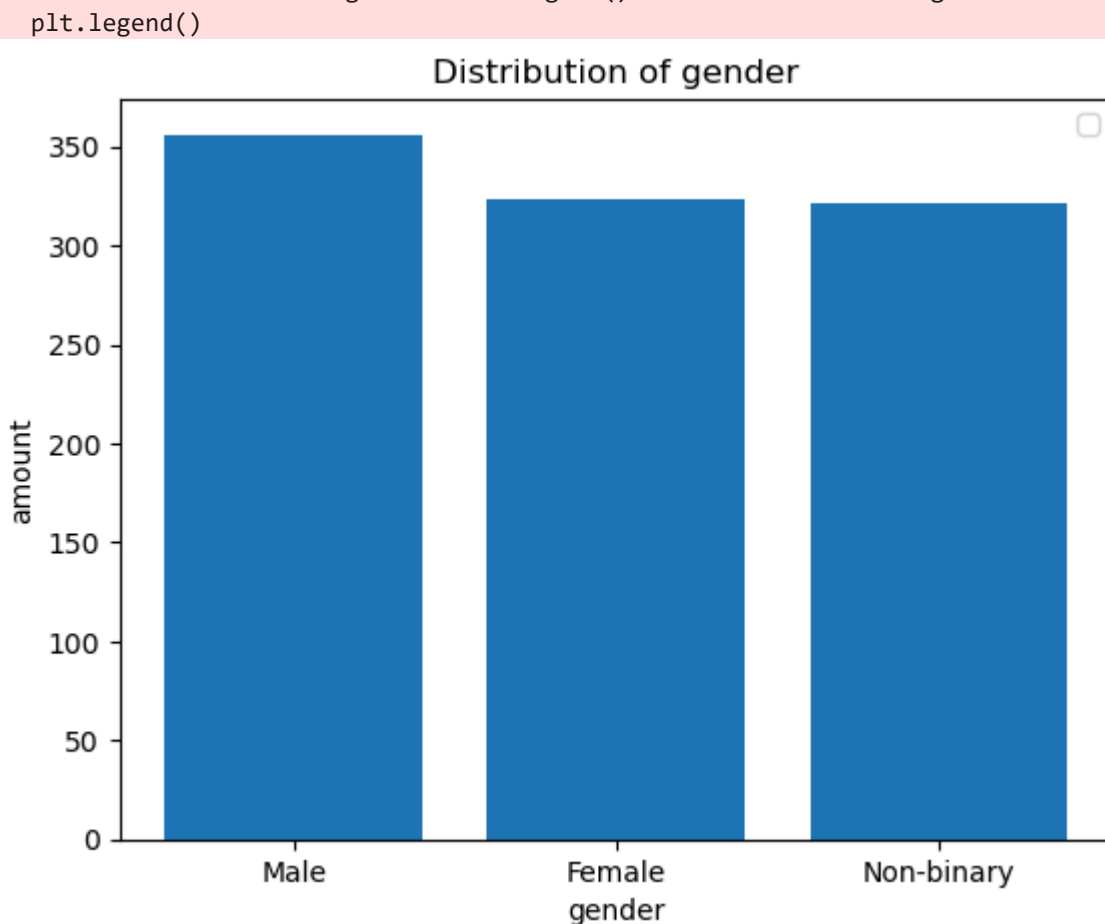
สร้างแผนภูมิแท่ง (bar chart) เพื่อแสดงการกระจายของเพศ (gender) ในการแสดงผล ให้แสดงชื่อแผนภูมิ และคำอธิบายสัญลักษณ์ (legend)

In [179...

```
# Write your code here
plt.figure()
plt.bar(df['gender'].value_counts().index, df['gender'].value_counts().values)
plt.title('Distribution of gender')
plt.xlabel('gender')
```

```
plt.ylabel("amount")
plt.legend()
plt.show()
```

C:\Users\punch\AppData\Local\Temp\ipykernel_15184\351817206.py:7: UserWarning: No artists with labels found to put in legend. Note that artists whose label start with an underscore are ignored when legend() is called with no argument.



[5 points] Describe this visualization in your own words. What information does it convey?

อธิบายแผนภูมิที่สร้างขึ้น แผนภูมินี้สื่อถึงข้อมูลอะไร?

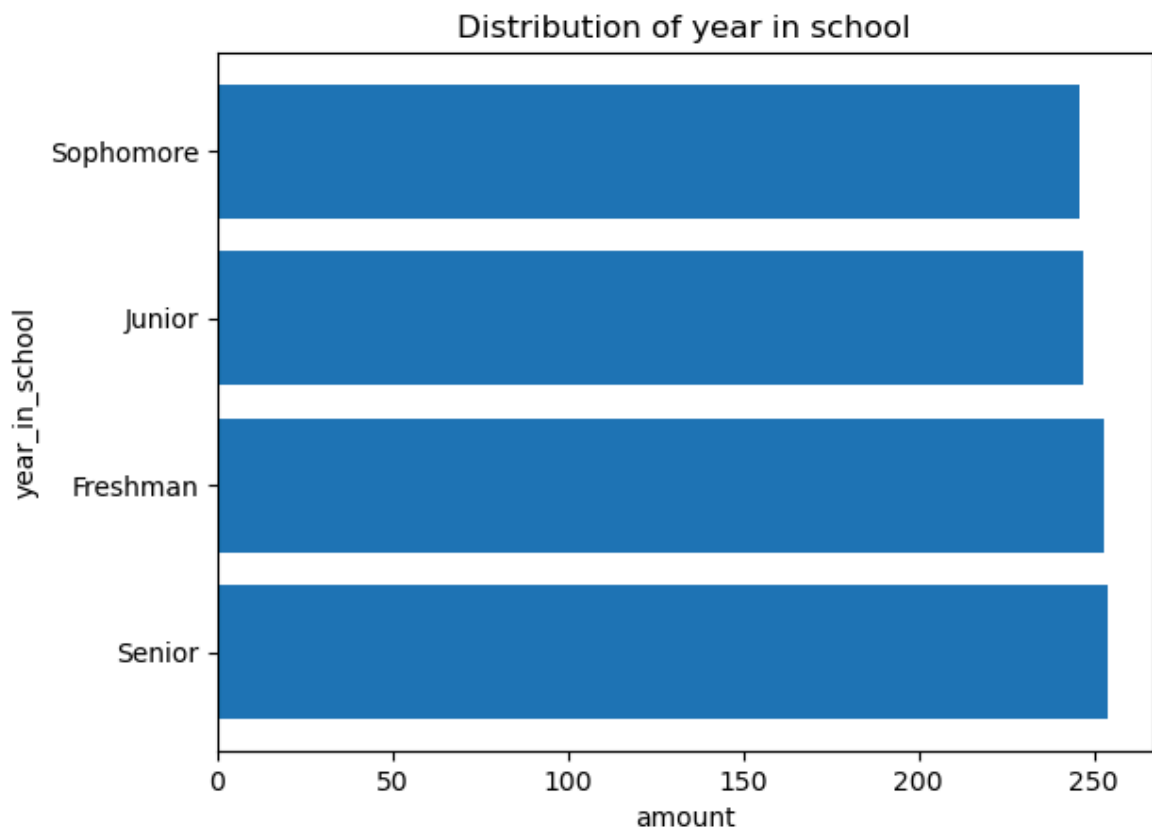
ANS: จำนวนคนแต่ละเพศ ประกอบด้วย เพศ Male, Female, Non-binary ซึ่งจากแผนภูมิพบว่า เพศ Male มากที่สุดและ Female, Non-binary มีจำนวนรองลงมาตามลำดับ

[7 points] Create a *horizontal* bar chart to demonstrate the distribution of year in school. In your visualization, also display the chart title and data labels.

สร้าง แผนภูมิแท่งแนวนอน (horizontal bar chart) เพื่อแสดงการกระจายของปีที่ศึกษา (year in school) ในการแสดงผล ให้แสดงชื่อแผนภูมิ และคำอธิบายสัญลักษณ์ (legend)

In [173...

```
# Write your code here
plt.figure()
plt.barh(df['year_in_school'].value_counts().index, df['year_in_school'].value_co
plt.title('Distribution of year in school')
plt.xlabel('amount')
plt.ylabel("year_in_school")
plt.show()
```



[5 points] Describe this visualization in your own words. What information does it convey?

อธิบายแผนภูมิที่สร้างขึ้น แผนภูมินี้สื่อถึงข้อมูลอะไร?

ANS: แสดงถึงจำนวนของนักศึกษาในแต่ละรุ่น ประกอบด้วย Senior, Freshman, Junior, Sophomore ซึ่งเรียงจากมากไปน้อยตามลำดับ

Subtask #5: จัดกลุ่มข้อมูล

[6 points] Group the data by `year_in_school` and `major`, then display the sum of these attributes: `entertainment`, `personal_care`, `technology`, `health_wellness`, and `miscellaneous`.

จัดกลุ่มข้อมูลตาม `year_in_school` และ `major` แล้วแสดงผลรวมของคุณลักษณะต่อไปนี้: `entertainment`, `personal_care`, `technology`, `health_wellness`, และ `miscellaneous`

In [174...

```
# Write your code here
# Write your code here
df.groupby(['year_in_school', 'major'])[['entertainment', 'personal_care', 'technol
```

Out[174...

		entertainment	personal_care	technology	health_wellness
year_in_school	major				
Freshman	Biology	5173	3392	10912	7033
	Computer Science	4405	3300	9931	6069
	Economics	3832	2638	7530	4605
	Engineering	3844	3097	7838	5705
	Psychology	4536	3289	9551	5676
Junior	Biology	5306	3562	9975	6997
	Computer Science	3886	2733	8375	5589
	Economics	3596	2689	9160	5231
	Engineering	4429	2802	9212	5122
	Psychology	3768	2602	7286	5564
Senior	Biology	4557	3426	10858	6924
	Computer Science	3705	3164	7285	5406
	Economics	5166	3520	9589	6350
	Engineering	4188	3516	10223	5404
	Psychology	3498	2671	7783	5039
Sophomore	Biology	4784	3053	9851	6778
	Computer Science	3770	2626	8200	5262
	Economics	4708	3158	10048	5612
	Engineering	3820	2404	6879	4373
	Psychology	3843	3057	7818	5571



[2 points] Describe your understanding from this output.

อธิบายความเข้าใจจากผลลัพธ์ที่ได้

ANS: ภาพรวม entertainment, personal_care, technology, health_wellness, miscellaneous ในแต่ละคณะในแต่ละชั้นปี ทำให้สามารถทราบแนวโน้มพฤติกรรมและนำข้อมูลนี้ไปในการตัดสินใจได้