Statistics on further studies of students in IIT H

MA4240 Applied Statistics

Shreyas Wankhede Pradeep Mundlik Akshitha Kola Dhatri Reddy
Mouktika Cherukupalli Avinash Malothu Sumeeth Kumar Sathvika Marri

Indian Institute of Technology Hyderabad

May 1, 2023

Outline

Introduction

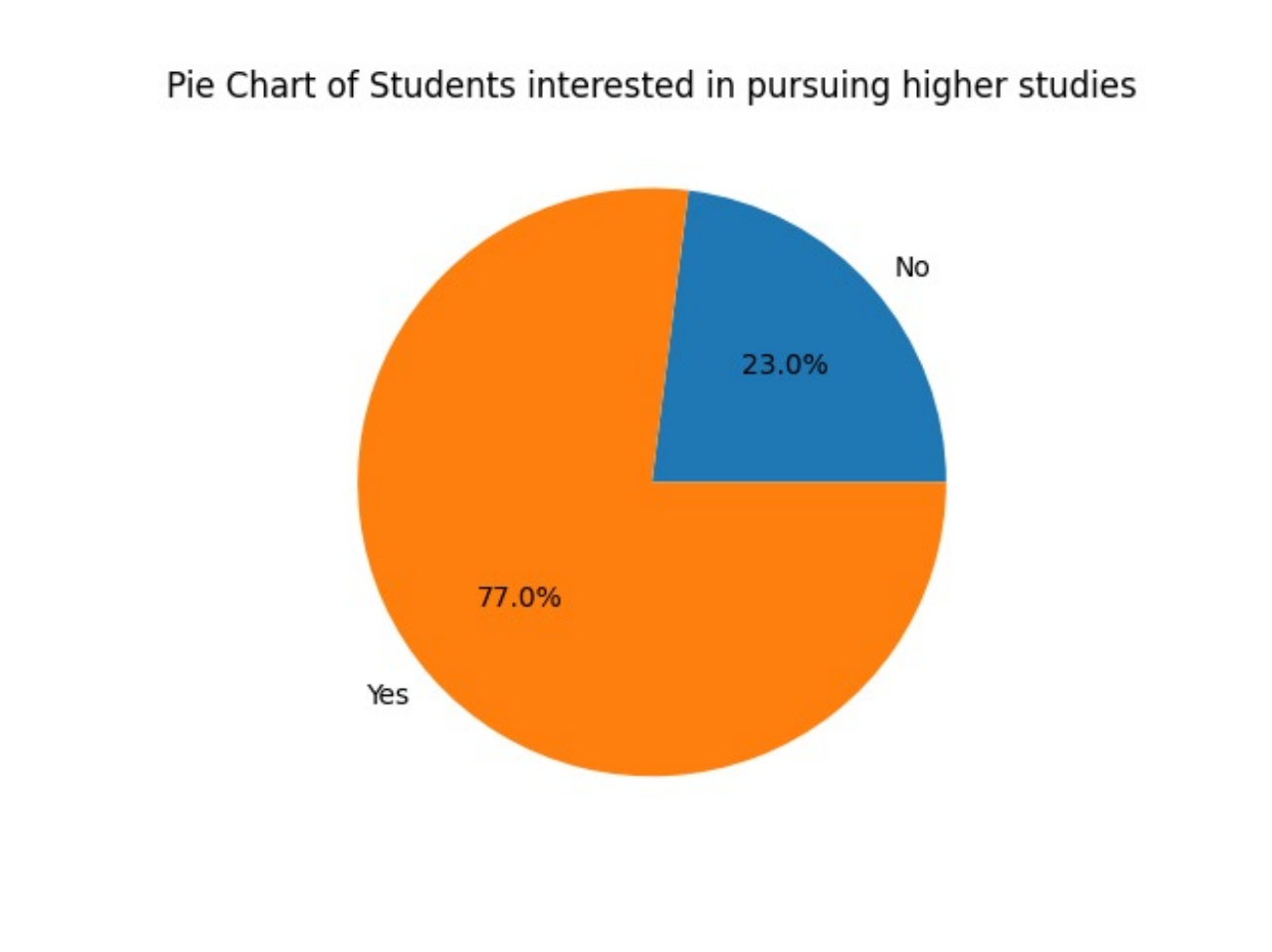
This project is based on further studies of students studying at IITH. We have used statistics to deduce few conclusions from the given data, assuming that the data is random sample population.

We used sampling, more specifically volunteer sampling for collection of data through mail from students at IITH but only a few of them volunteered to respond. The data was collected from 114 students, and it is diverse with data from different years of UG.

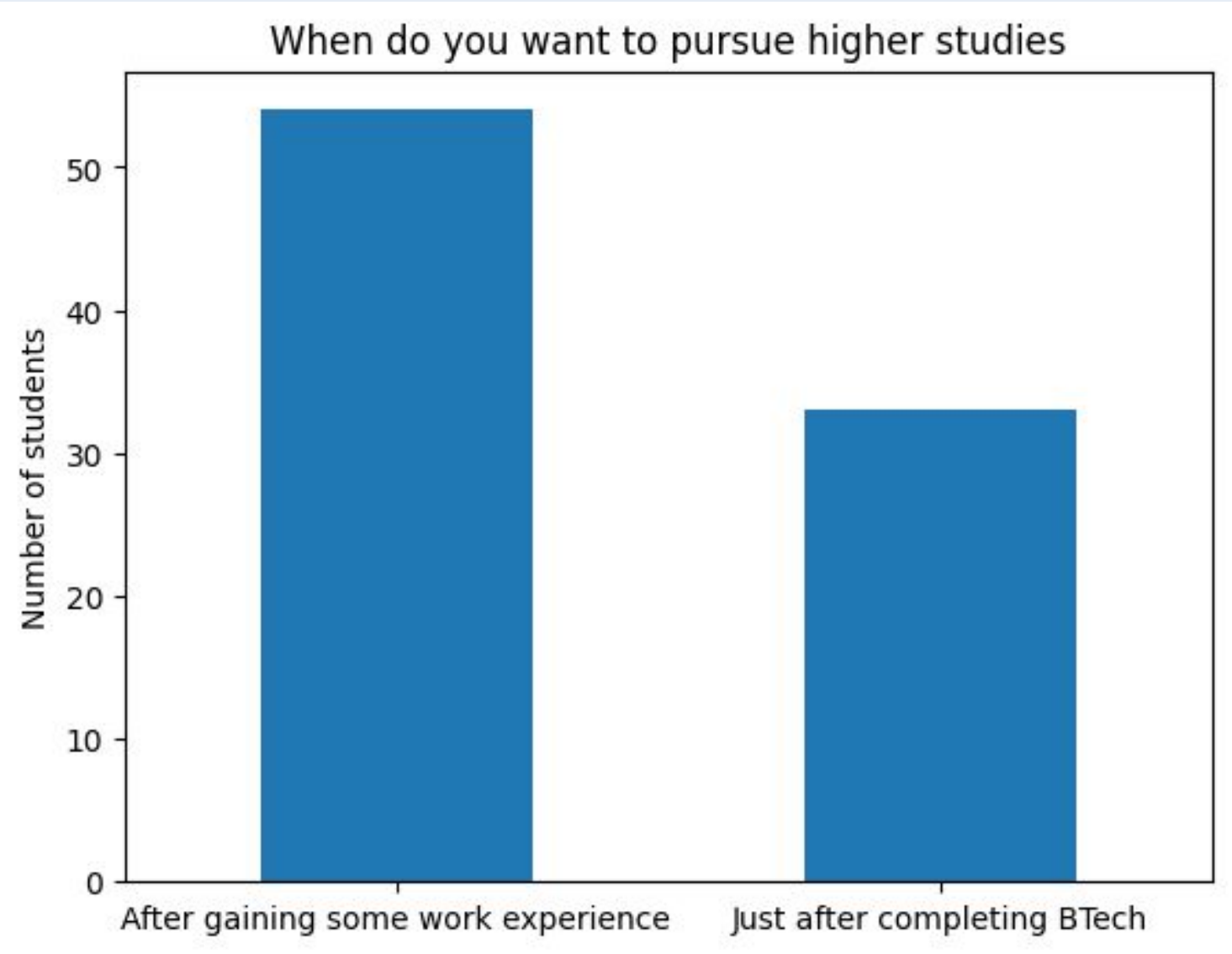
Variables of interest

1. Which department?
2. Gender
3. Annual family income
4. CGPA
5. Interested in further studies (if yes)
   - After how many years of work experience?
   - Which degree?(Masters/ PhD/ MBA)
   - Location
   - In which department(Same as bachelors or different?)
6. Interested in further studies(if no)
   - Are you interested in civil services?
   - Are you interested in software development?

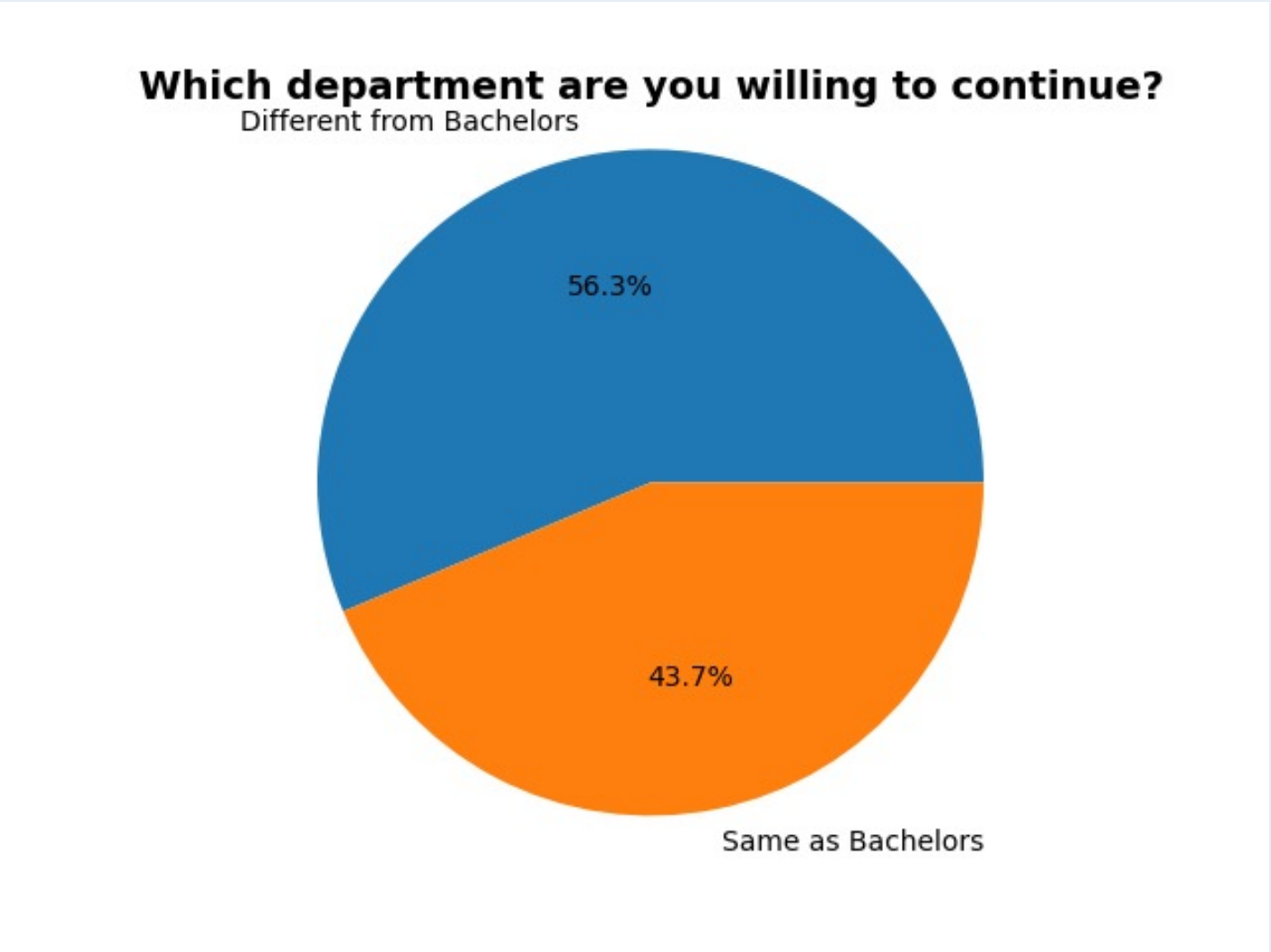
Visualising frequency plots



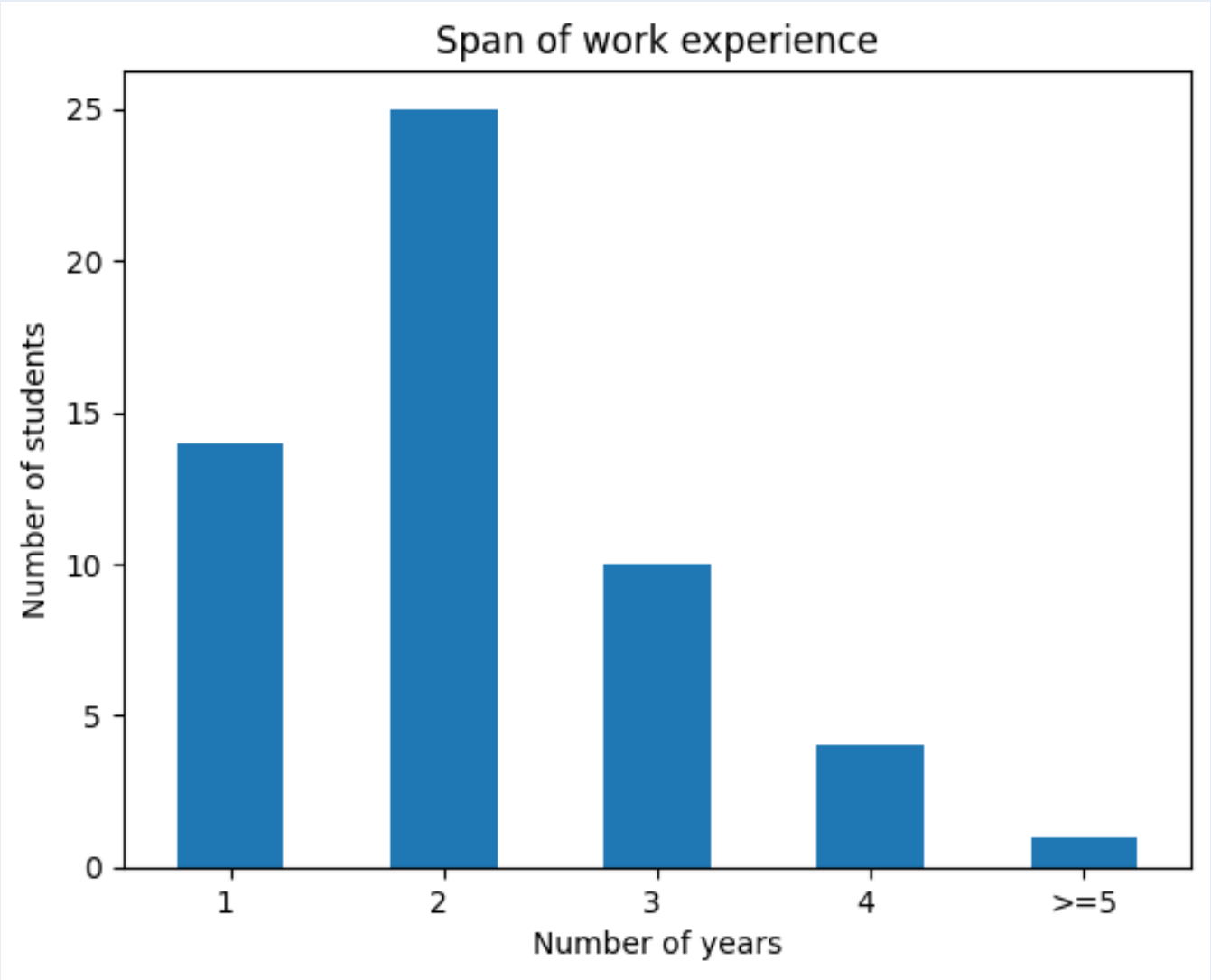
Visualising frequency plots



Visualising frequency plots

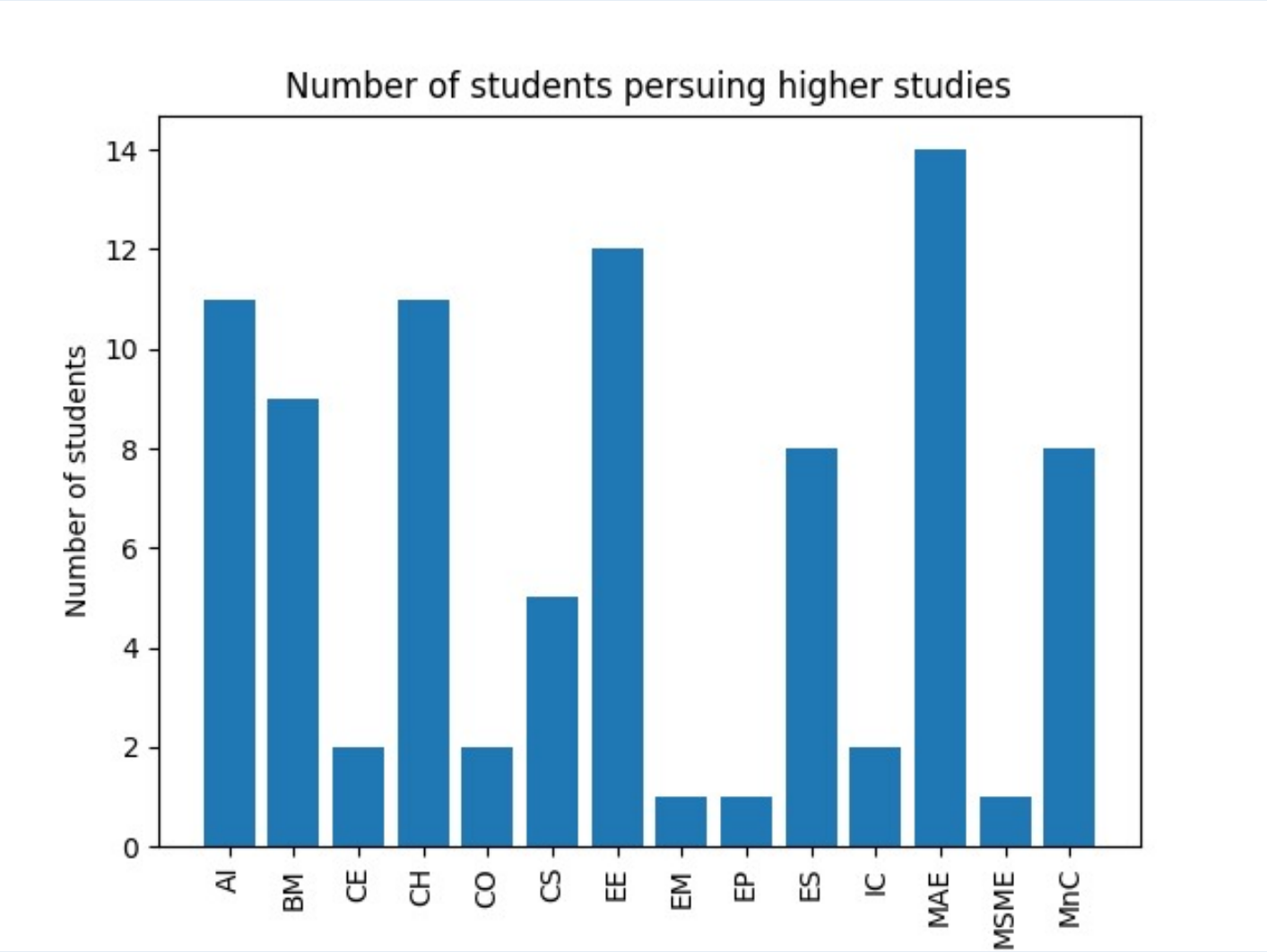


Visualising frequency plots

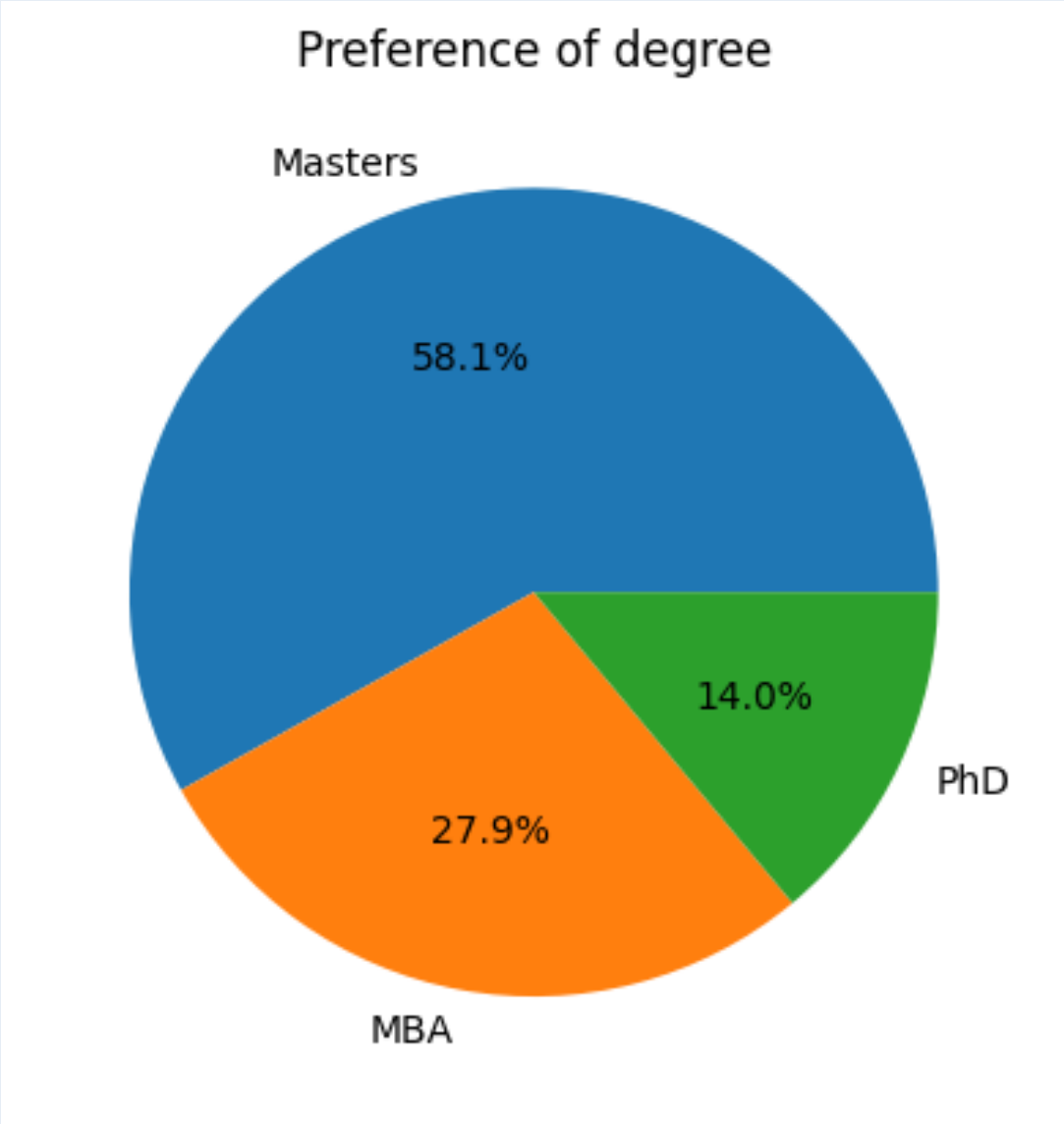


Visualising frequency plots

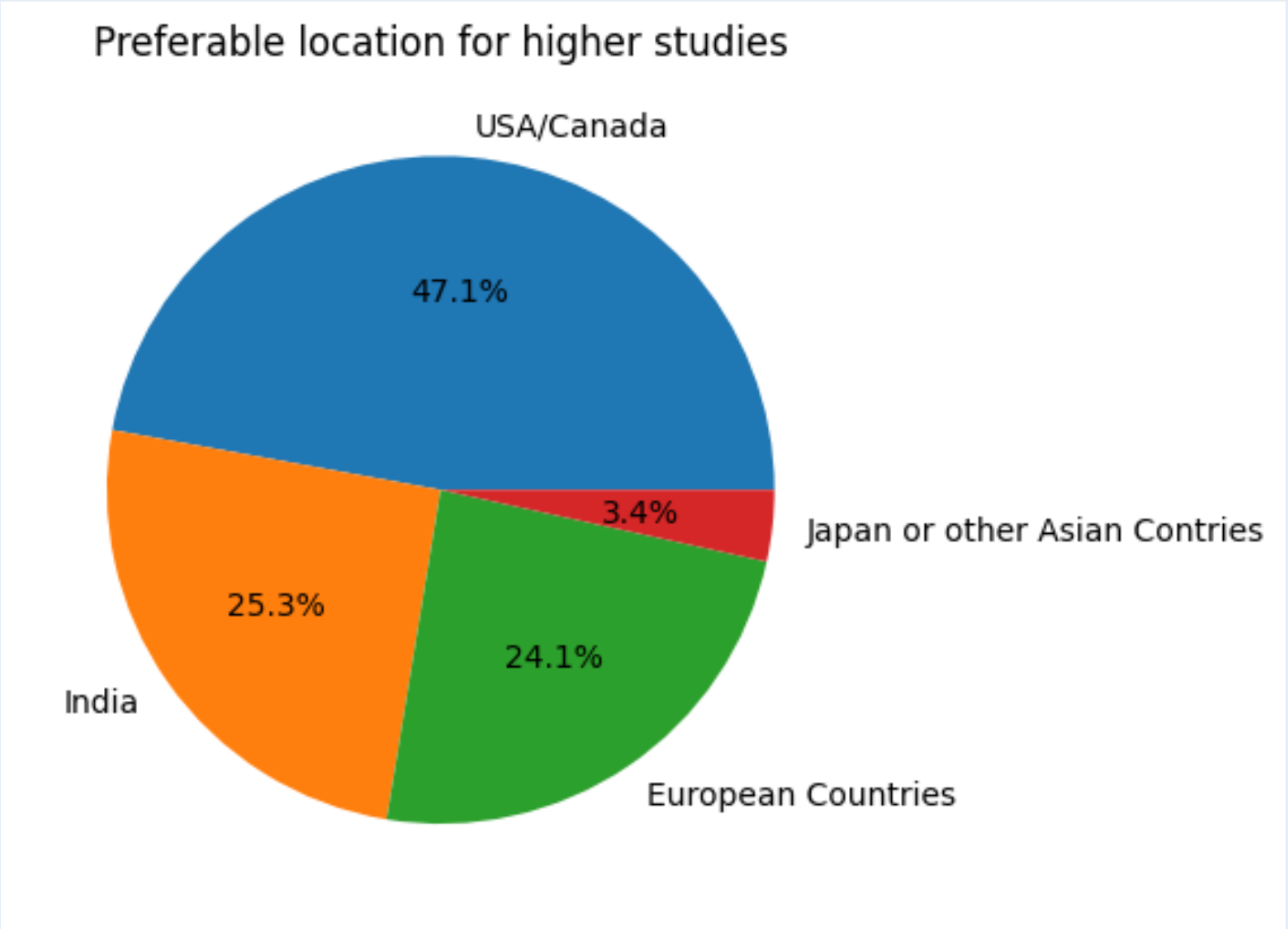
Figure: department wise students opting for higher studies



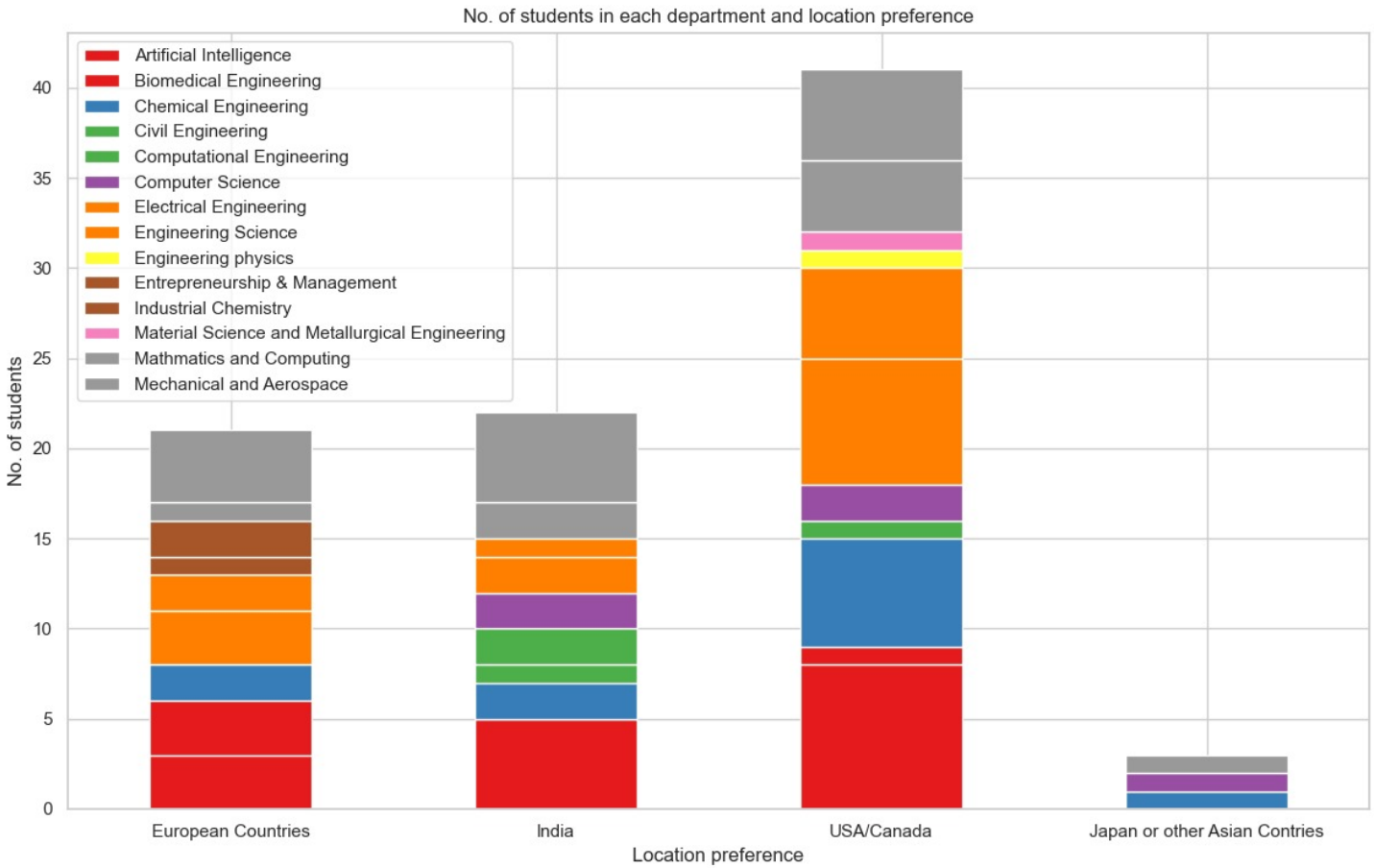
Visualising frequency plots



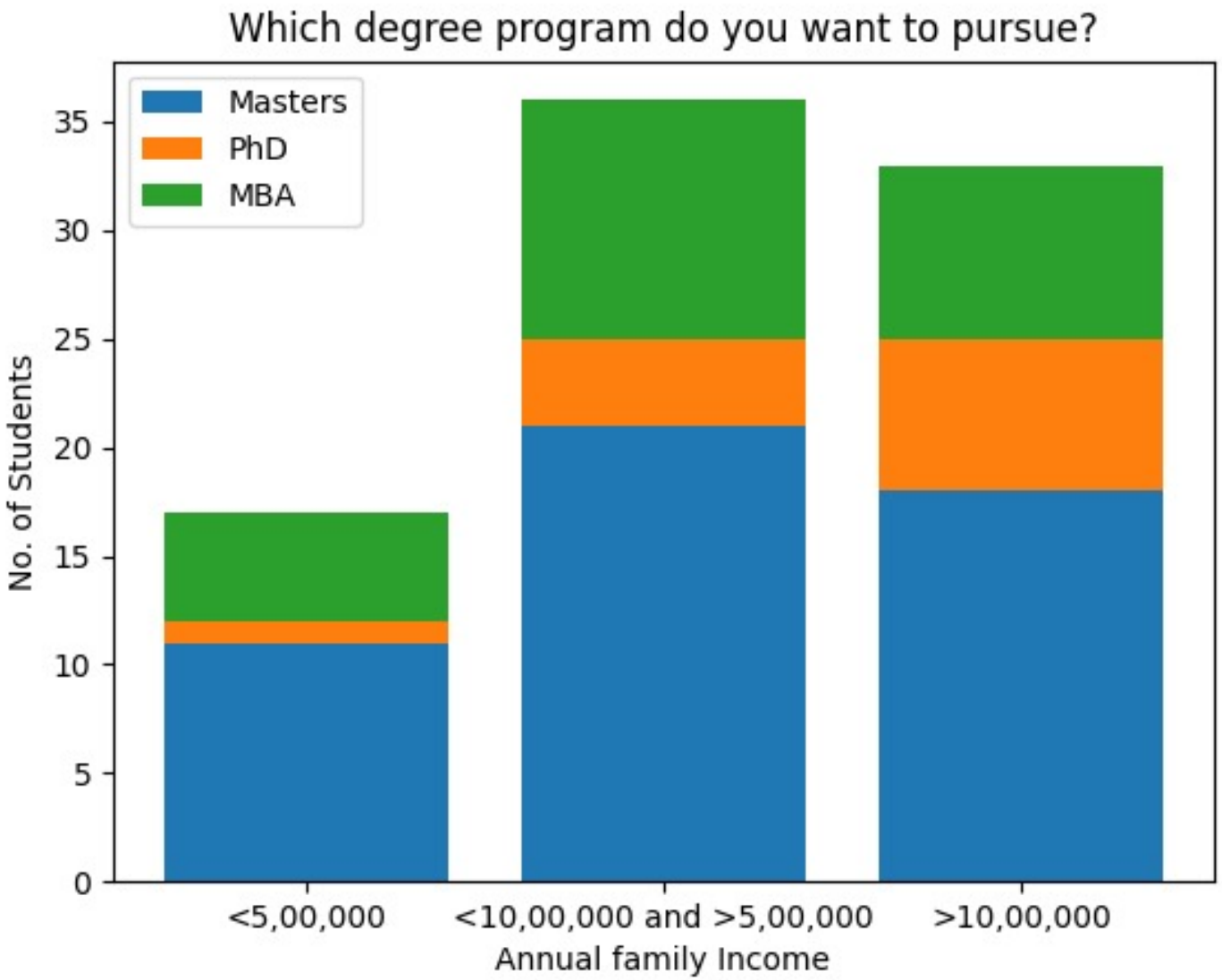
Visualising frequency plots



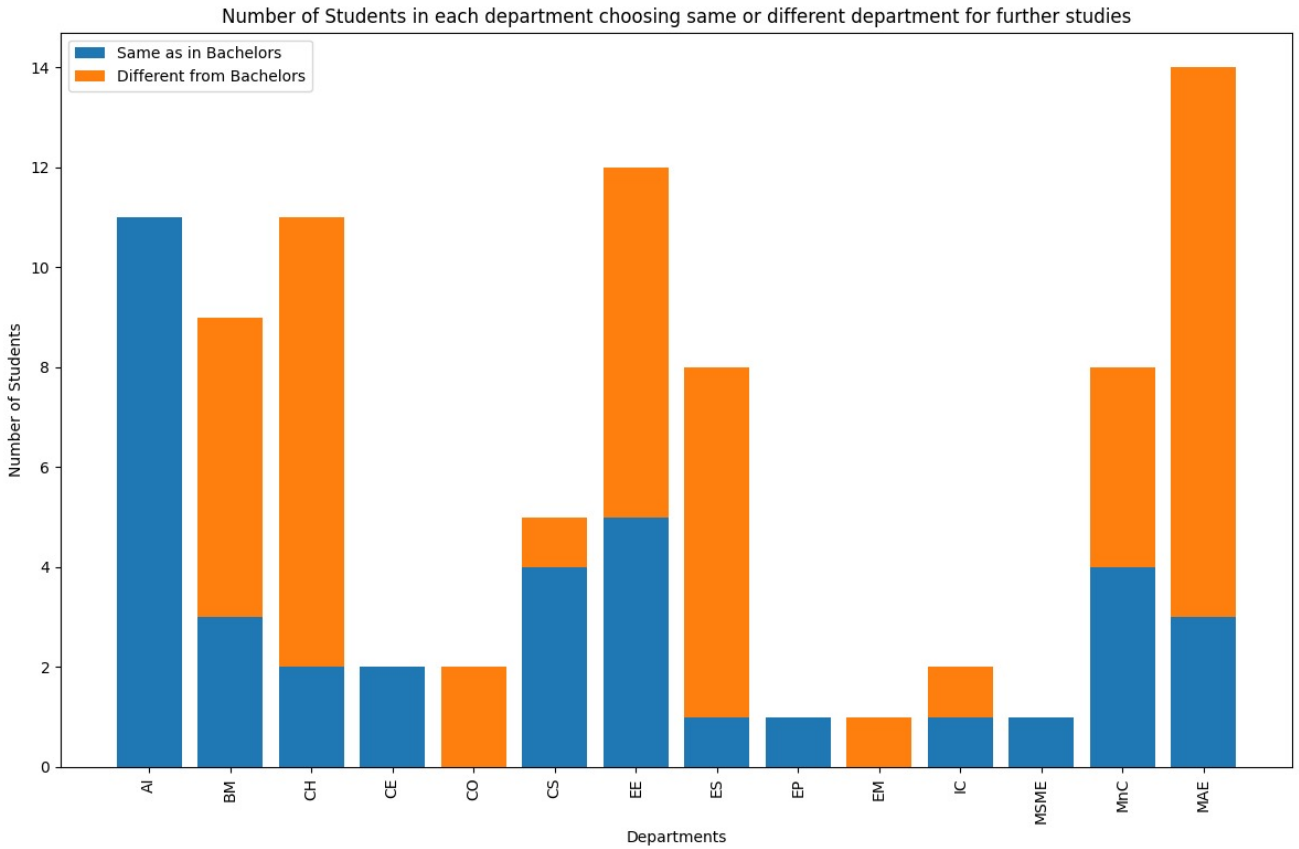
Visualising cross plots for various categorical variables



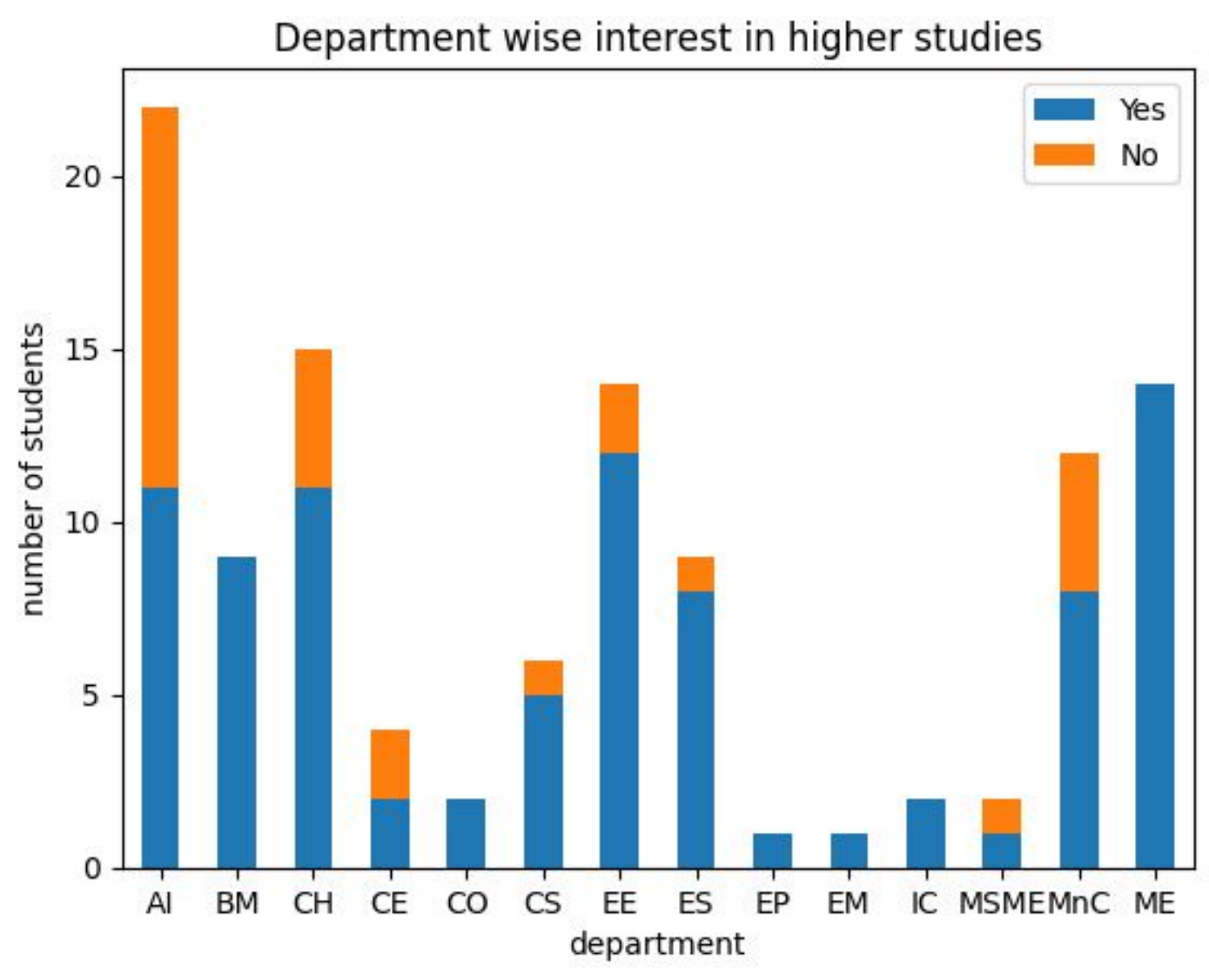
Visualising cross plots for various categorical variables



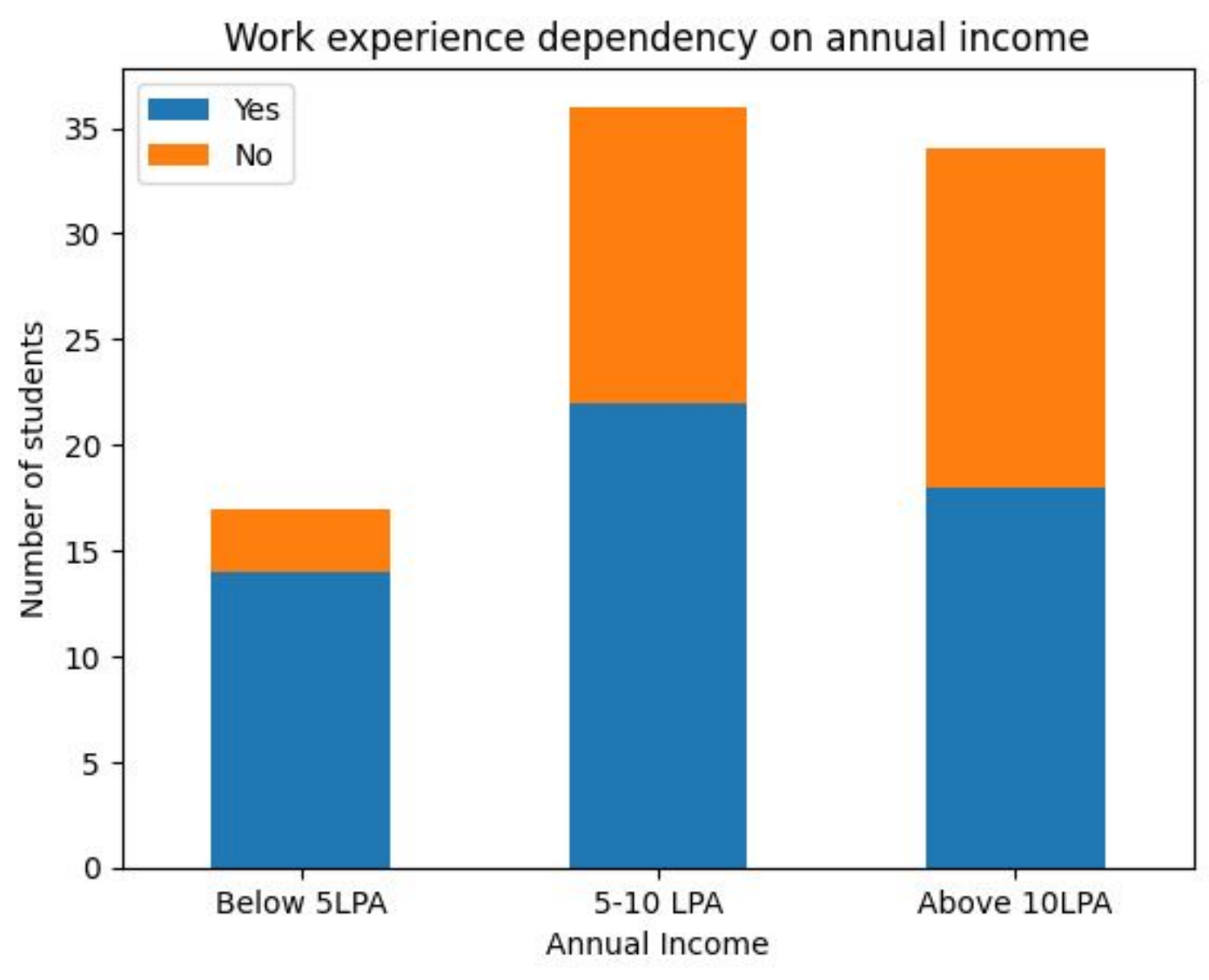
Visualising cross plots for various categorical variables



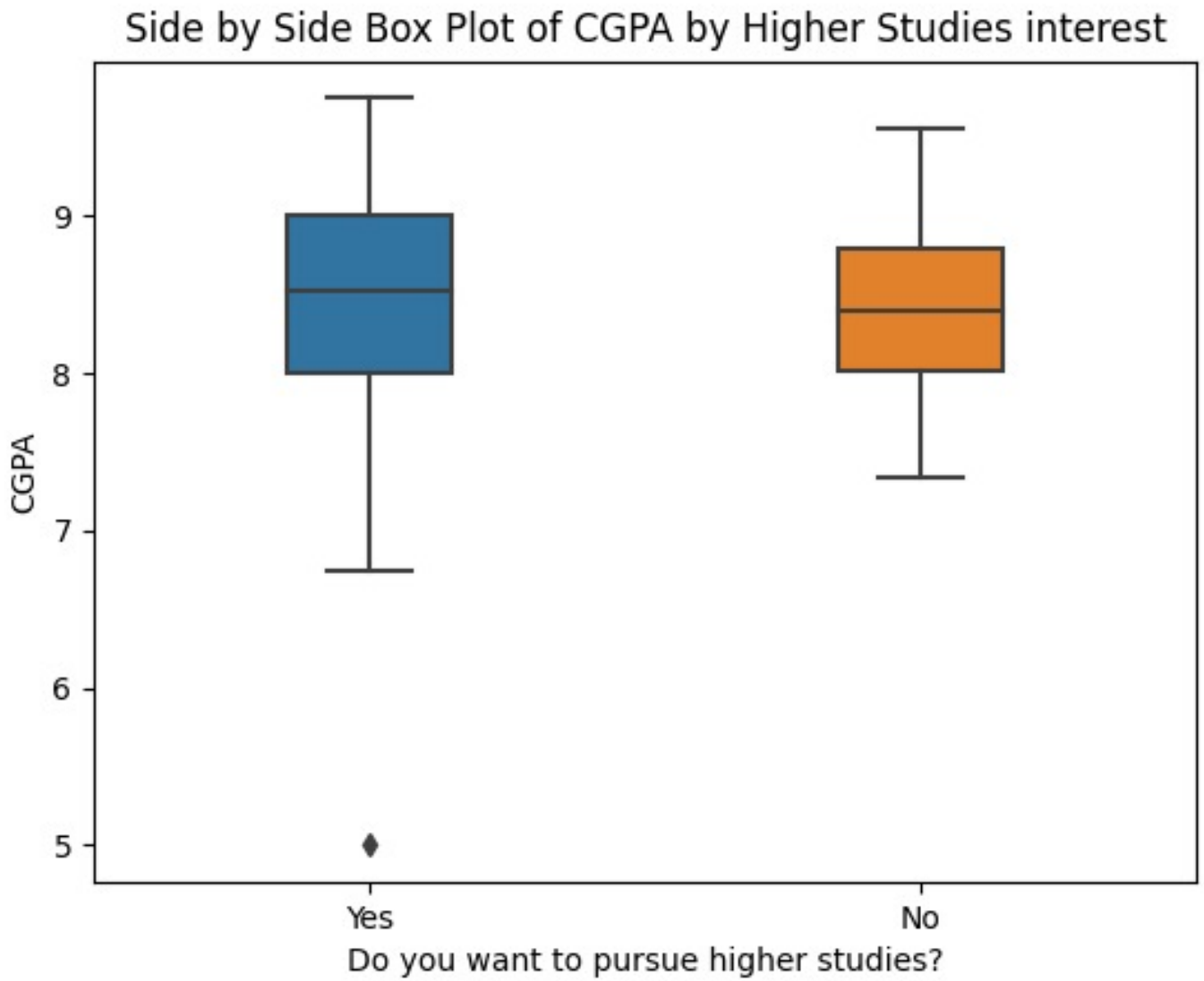
Visualising cross plots for various categorical variables



Visualising cross plots for various categorical variables



Visualising cross plots for various categorical variables



Percentage Contingency table between Gender and interest in pursuing further studies

	Yes	No	Total (%)
Male	53.10	17.70	70.80
Female	23.89	5.31	29.20
Total (%)	76.99	23.01	100.00

Percentage Contingency table between preferred program and preferred location

	MBA	Masters	PhD	Total (%)
European Countries	5.75	14.94	3.45	24.14
India	13.79	8.05	3.45	25.29
Japan or other Asian countries	1.15	1.15	1.15	3.45
USA/Canada	6.90	34.48	5.75	47.13
Total (%)	27.59	58.62	13.79	100.00

Data Analysis

Conclusions

With Around 113 students participating in the survey,
77% students are interested in pursuing higher studies and 23% students are not interested in pursuing higher studies.

Experience:

37.9 % students want to pursue higher studies with no experience, 62.1% students want to pursue higher studies with experience.

Degree preference:

58.1% want to pursue masters, 14.0% want to pursue phD, 27.9% want to pursue MBA.

Data Analysis

Conclusions

Location preference:

47.1% want to pursue in USA/Canada, 25.3% want to pursue in India, 24.1% want to pursue in European countries, 3.4% Japan or other Asian countries.

Field Interest:

72.56% students are interested in civil services, 27.44% students are not interested in civil services.

74.33% students are interested in software related jobs, 25.66% students are not interested in software related jobs.

Hypothesis Testing

Case 1: Comparing CGPA of students who are willing to pursue higher studies with students who don't want to pursue

We assume our null hypothesis to be that average CGPA of students willing to go for higher studies is greater than or equal to those who don't want to. Let $\alpha = 0.05$.

1. $\bar{x_1}$ = sample mean of CGPA of people willing to go for higher studies
2. $\bar{x_2}$ = sample mean of CGPA of people who don't want to go for higher studies
3. s_1^2 = sample standard deviation of CGPA of people willing for higher studies
4. s_2^2 = sample standard deviation of CGPA of people who don't want

For Hypothesis Testing we make the following statements:

$$H_0 = \mu_1 - \mu_2 \geq 0 \tag{1}$$

$$H_a = \mu 1 - \mu 2 < 0 \tag{2}$$

Case 1 continued:

Information:

$$\bar{x_1} = 8.4447 \tag{3}$$
$$\bar{x_2} = 8.3919 \tag{4}$$
$$s_1^2 = 0.7694 \tag{5}$$
$$s_2^2 = 0.5672 \tag{6}$$
$$n_1 = 87 \tag{7}$$
$$n_2 = 26 \tag{8}$$

since $\frac{s_1^2}{s_2^2} < 4$, we can assume the population variances would be equal.
Degrees of freedom,

$$dof = n_1 + n_2 - 2 = 111$$

Case 1 continued:

The pooled variance will be:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 0.5317 \tag{9}$$

The test statistic is given by:

$$t = \frac{\bar{x_1} - \bar{x_2} - 0}{s_p \sqrt{\frac{n_2 + n_1}{n_1 n_2}}} = 0.3314 \tag{10}$$

Using the rejection region approach, we reject H_0 if $t_{0.05,111} \geq -t$ where $t_{0.05,111} = -1.6587$. We have enough statistical evidence to reject null hypothesis since observed t is lesser than 1.6587

Case 2: Hypothesized testing if students want an average work experience of more than 1 year before going for further studies

Let us assume the null hypothesis as the average work experience of students willing to go for higher studies is less than 1 year. Let $\alpha = 0.05$. The hypotheses are:

$$H_0 = \mu \leq \mu_0$$
$$H_a = \mu > \mu_0$$

where $\mu_0 = 1$

$\bar{X}$ = Average years of work experience before going for further studies

S^2 = Sample standard deviation of number of years of work experience of students willing to go for higher studies

n = Number of students planning to go for further studies

Case 2 continued:

Information:

$$\bar{X} = 1.33 \tag{11}$$

$$S^2 = 1.74 \tag{12}$$

$$n = 87 \tag{13}$$

$$df = n - 1 = 86 \tag{14}$$

The test statistic is given by:

$$t^* = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{1.33 - 1}{\frac{\sqrt{1.74}}{\sqrt{87}}} = 2.34 \tag{15}$$

Using the rejection region approach, we reject H_0 if $t^* \geq t_{0.05,86}$ where $t_{0.05,86} = 1.6628$. Hence, we have enough statistical evidence to reject null hypothesis H_0.

Hypothesis testing

Case 3: Comparing average work experience students take before going for higher studies in **Management field** vs in **Research field**

Let us assume the null hypothesis as avg work experience of students going for MBA is less than or equal to avg work experience of students going for MS/PhD

Let $\alpha = 0.05$

1. $\bar{X_1}$ = sample mean of avg work experience of students willing to go for MBA
2. $\bar{X_2}$ = sample mean of avg work experience of students willing to go for MS/PhD
3. $\bar{s_1}$ = sample variance of avg work experience of students willing to go for MBA
4. $\bar{s_1}$ = sample variance of avg work experience of students willing to go for MS/PhD

Case 3: Continued

The hypotheses are:

$$H_0 = \mu_1 - \mu_2 \leq 0 \tag{16}$$
$$H_a = \mu_1 - \mu_2 > 0 \tag{17}$$

Information:

$$\bar{X_1} = 1.54 \tag{18}$$
$$\bar{X_2} = 1.25 \tag{19}$$
$$\bar{s_1} = 1.30 \tag{20}$$
$$\bar{s_1} = 1.90 \tag{21}$$
$$\bar{n_1} = 24 \tag{22}$$
$$\bar{n_2} = 63 \tag{23}$$
$$\tag{24}$$

Case 3: Continued

$$c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{\frac{(1.30)^2}{24}}{\frac{(1.30)^2}{24} + \frac{(1.90)^2}{63}} = 0.57 \tag{25}$$

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(1 - c)^2 (n_1 - 1) + c^2 (n_2 - 1)} \tag{26}$$

$$= \frac{(24 - 1)(63 - 1)}{(1 - 0.57)^2 (24 - 1) + (0.57)^2 (63 - 1)} = 58.45 \tag{27}$$

$$df \approx 54 \tag{28}$$

Case 3: Continued

The test static is given by:

$$t' = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{29}$$

where, $D_0 = 0$

$$t' = \frac{(1.54 - 1.25) - 0}{\sqrt{\frac{(1.30)^2}{24} + \frac{(1.90)^2}{63}}} = 0.81 \tag{30}$$

RR: $t' \geq t_{0.05}$, at $df = 54$
Here, $t_{0.05,54} = 1.67 > t'$

Case 3: Continued

Since, t' does not fall in rejection region for $\alpha = 0.05$, we fail to reject H_0. Here, we don't have enough evidence to say that avg work experience of MBA willing students is more than MS/PhD willing students.

Hypothesis testing

Case 4: Hypothesized proportion testing if there is enough evidence that the proportions of people opting for masters, MBA, Phd are not all equal

Sample data :

Masters	MBA	PhD	Total
50	24	13	87

Let P_{Ms}, P_{MBA}, P_{PhD} denote proportions of students willing to pursue Masters, MBA, PhD for higher studies

$H_0 : P_{Ms} = P_{MBA} = P_{PhD} = \frac{1}{3}$ $\quad H_a :$ atleast one $P \neq \frac{1}{3}$ $\quad \alpha = 0.05$

Also,

$$E = \frac{1}{3} x 87 = 29 \tag{31}$$

and,

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{32}$$

Case 4 continued:

$$\begin{aligned}\chi^2 &= \frac{(50-29)^2}{29} + \frac{(24-29)^2}{29} + \frac{(13-29)^2}{29} \\ &= 15.2 + 0.862 + 8.827 \\ &= 24.889\end{aligned} \tag{33}$$

At $df = 3 - 1 = 2$, p value $= 0.0001$
p value $< \alpha = 0.05$
Hence there is enough evidence that population proportions are not all equal.