Mess Food Usage by IITH Student Community

Blessy Anvitha J
ai21btech11016@iith.ac.in

Saanvi Amrutha
ai21btech11026@iith.ac.in

Jarpula Bhanu Prasad
ai21btech11015@iith.ac.in

Lokesh Surana
es20btech11017@iith.ac.in

Arnav Asati
es20btech11034@iith.ac.in

Samar Singhai
bm20btech11012@iith.ac.in

Anirudh Srinivasan
cs20btech11059@iith.ac.in

Gandla Shivanand
es20btech11012@iith.ac.in

May 1, 2023

1 Introduction

In college, students often rely on mess food for their daily meals, but there can be many reasons why they may miss out on a meal or prefer certain types of food over others. For this project, we conducted a survey among students at our college to gather data on their food preferences, how many times they skip meals, their reasons for missing mess food, etc. Our aim was to analyze the data and gain insights into college students' food habits and preferences, which could help improve the quality and variety of food offered in our mess. For this project, we conducted a survey among students at our college to gather data on their food preferences and their reasons for missing mess food. Our aim was to analyze the data and gain insights into college students' food habits and preferences, which could help improve the quality and variety of food offered in our mess.

The survey consisted of the following questions:

1. How often do you skip mess eating mess food?
2. If you don't skip, what might be the possible reasons?"
3. You are an Early sleeping person/Late night sleeping person?
4. Roughly what would be the count of no of times you miss mess food in a week?
5. Which meal do you usually skip in a day?
6. Are you a vegetarian or non-vegetarian?
7. What's your food preference?
8. What are the probable reason(s) for you to skip the mess?
9. What degree are you pursuing in IITH?
10. What department are you in?
11. Gender of the student?
12. Age of the student?
13. What days of the week are you likely to skip?

14. Family Income (Per Annum)?
15. Do you eat outside mess, in case you miss a meal?
16. If yes, what's your alternative?

2 Data Visualization and Exploratory Data Analysis

After pre-processing the data, we had the data of 300 students from the survey, and we began to analyze the data. This involved the analysis of the one question that involved a numerical variable, that is:
'Roughly what would be the count of no of times you miss mess food in a week?'
The results obtained are as follows:

- Mean = 7.296
- Median = 6.5
- Mode = 10

Since the Median is less than the Mean and Mode, we cannot discuss the skewness of the data.

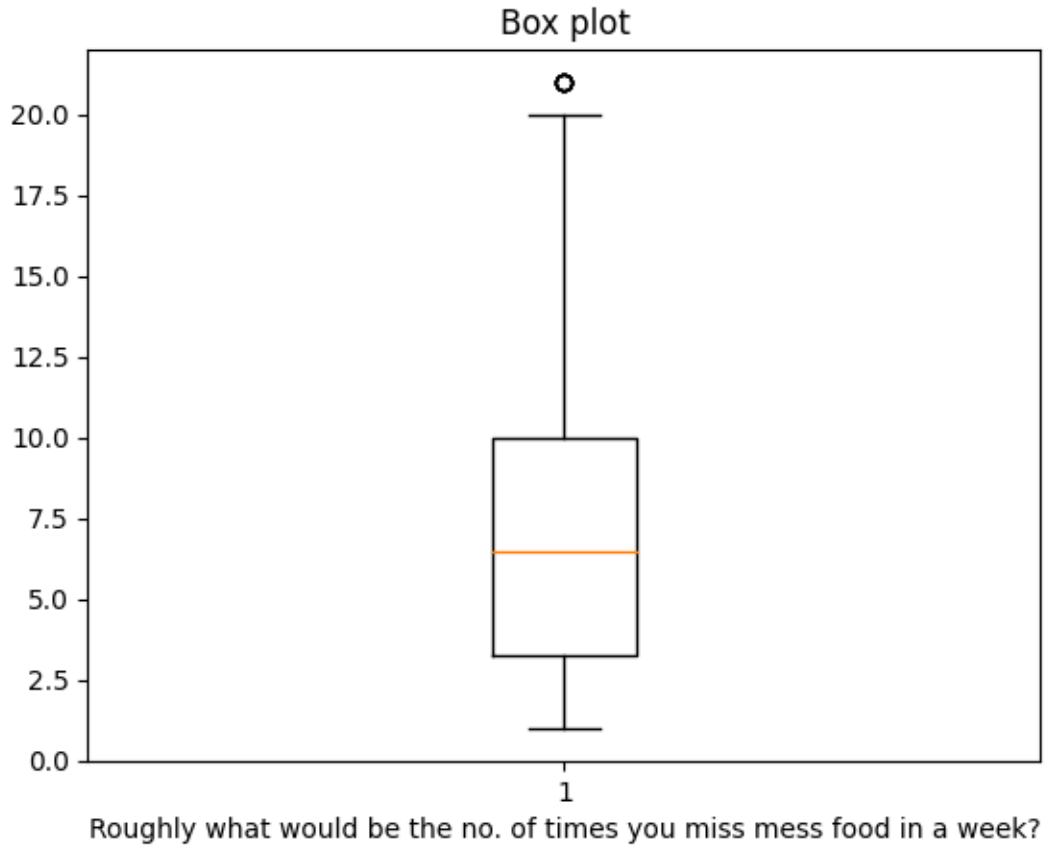

Figure 1: Roughly what would be the no.of times you miss mess food in a week? (Box plot).

We can observe from the box plot that the plot of frequency would be slightly right skewed with median around 6 (times a person missing a meal during a week), with the 1 outlier (greater than 20). The IQR is around 7.

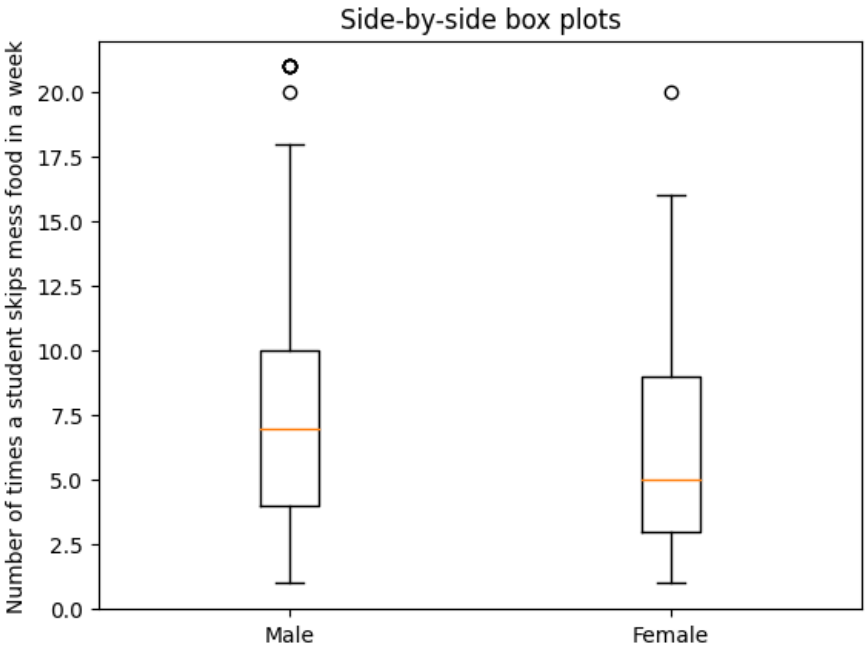

Figure 2: Week-days vs Week-ends (Side-by-Side Box plot).

$$\begin{array}{rr} \text{Q1(Male)} = 4 & \text{Q1(Female)} = 3 \\ \text{Q2(Male)} = 7 & \text{Q2(Female)} = 5 \\ \text{Q3(Male)} = 10 & \text{Q3(Female)} = 9 \\ \text{IQR(Male)} = 6 & \text{IQR(Female)} = 6 \end{array}$$

It is clear that males are skipping the meal more when compared to females.
Normal Probability Plot, which is used to assess whether the data is normally distributed or not, is given below:

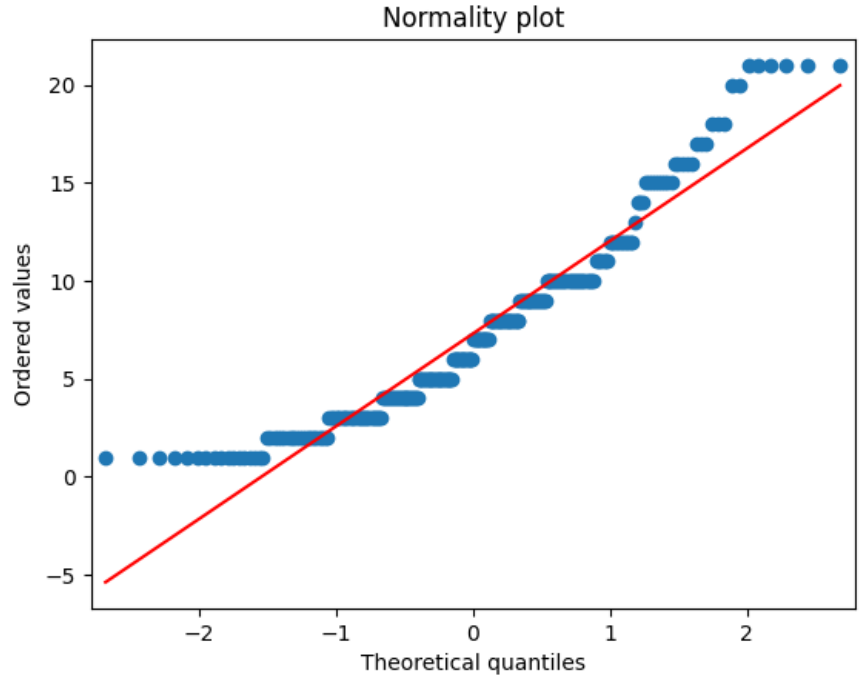

Figure 3: Normal Probability Plot

Degree vs number of responses

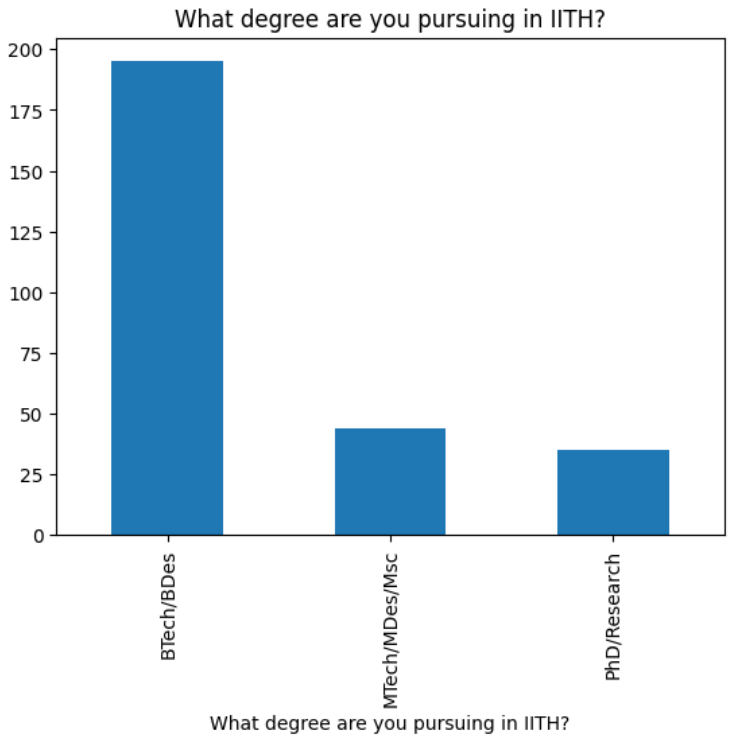

Figure 4: Degree vs number of responses bar graph

From this graph, we can see that students from BTech/Bdes responded the most for the contribution of data, which makes the analysis more centered around them.

Branch vs number of responses

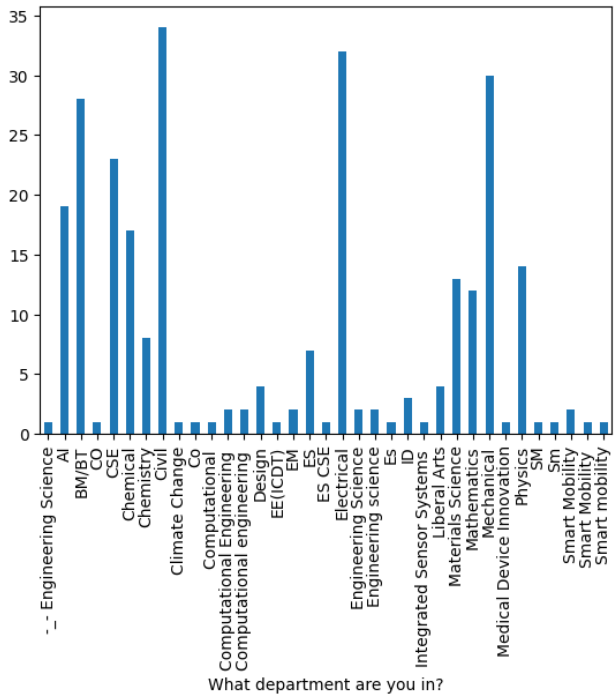

Figure 5: Branches vs number of responses bar graph

This graph shows that the people who responded the most are from Civil, BM/BT, Mechanical & Electrical.

Share of different economic classes

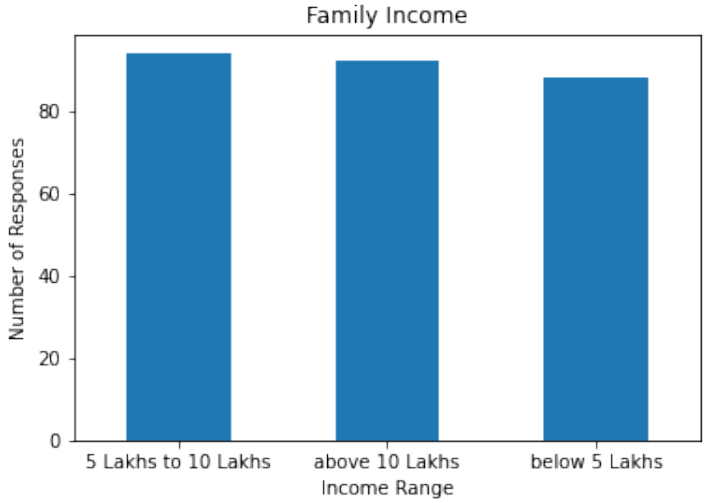

Figure 6: Family income bar graph

It is evident that all the three classes have almost equal share in our data.

Which meal is being skipped the most?

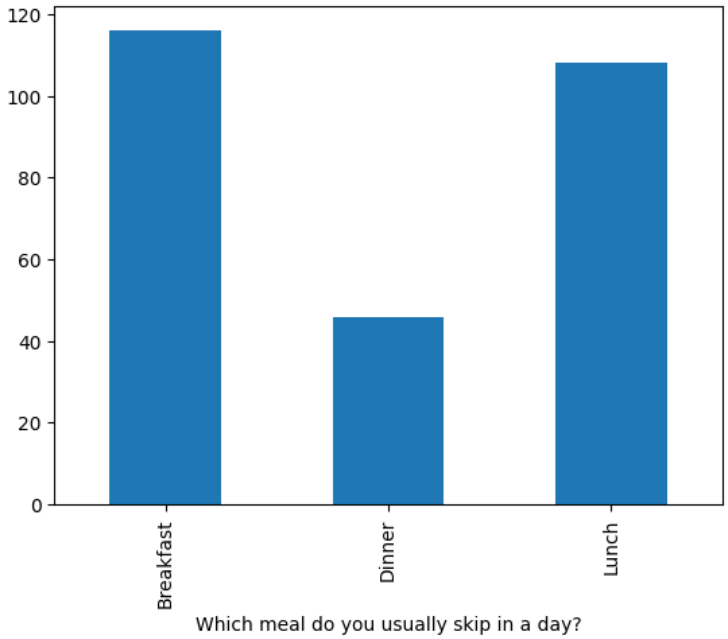

Figure 7: Meal skip bar graph

From this bar graph, we can easily see that breakfast is being missed majorly by people (since many people work till late at night and wake up late in the morning). Lunch is also being missed by the majority of people (since maybe lunch is not good in the mess).

How many people skip meals and how many times?

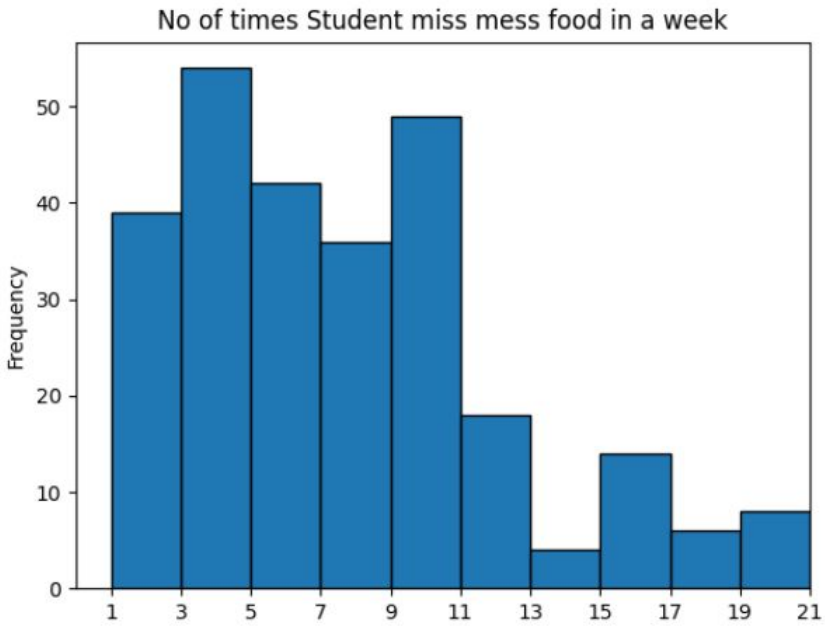

Figure 8: Frequecy of food missed vs Number of students

From the histograms, it is evident that the plot is right-skewed, meaning that the people missing food more than 10 times a week are fewer. People mostly skip meals 2 to 7 times a week.

Reasons for skipping a meal

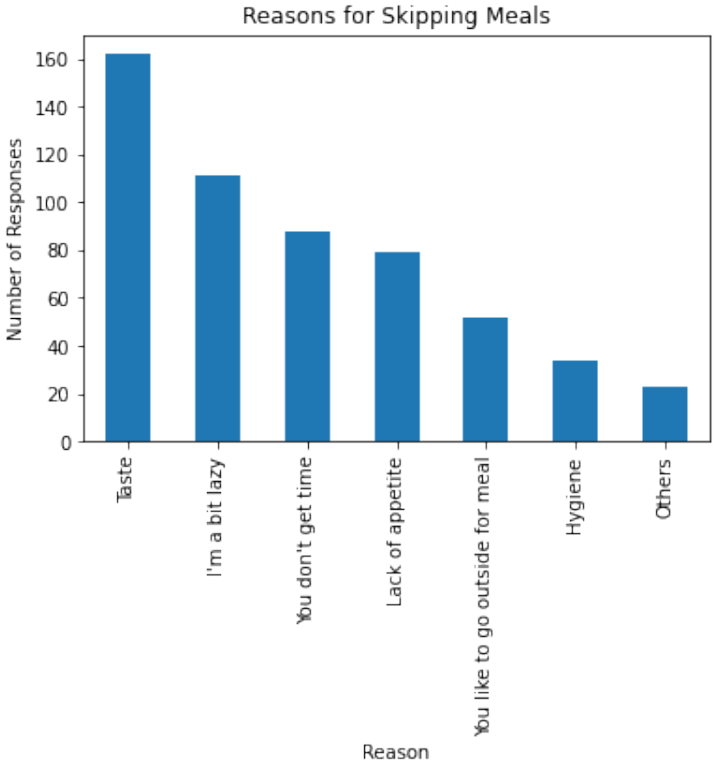

Figure 9: Reasons vs frequency bar graph

We can easily observe that the most common reason of missing a particular meal is 'due to the bad taste

of the food' being served. Other common reasons being 'I'm bit lazy', 'I don't get time', 'Like to go out for a meal' and others.

Correlation of meal skipping with family income

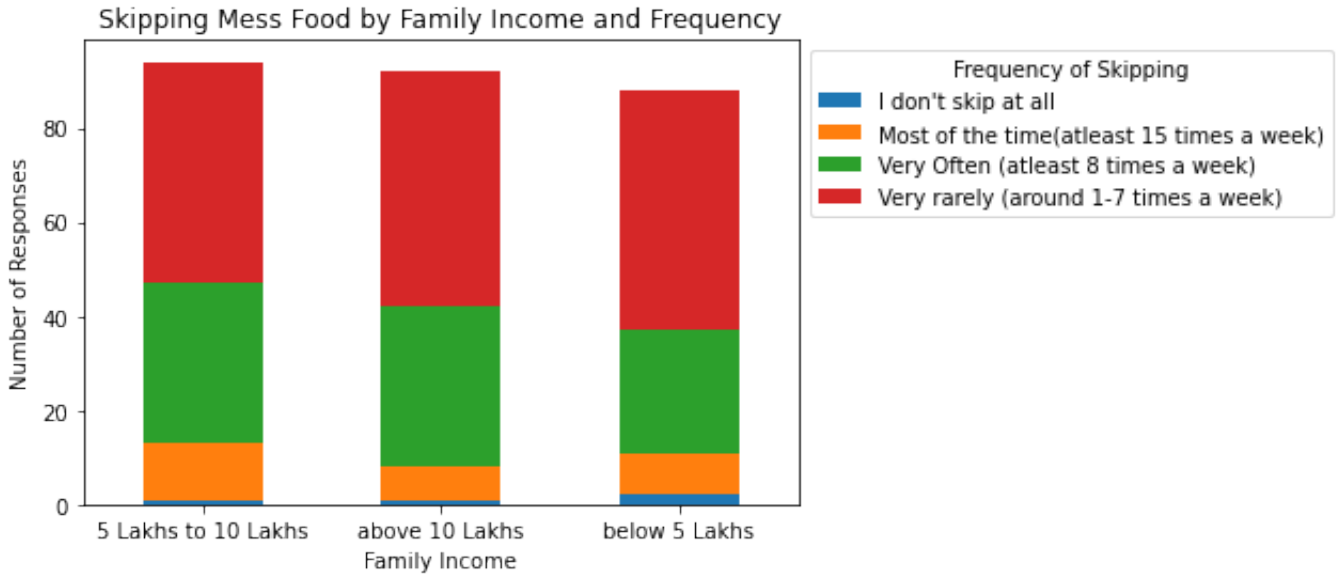

Figure 10: Skip meal-Family income segmented bar chart

It is evident that the meal skipping by the people is not very strongly correlated with their family income (we are getting similar bar charts for each economic class).

Correlation of meal skipping with gender

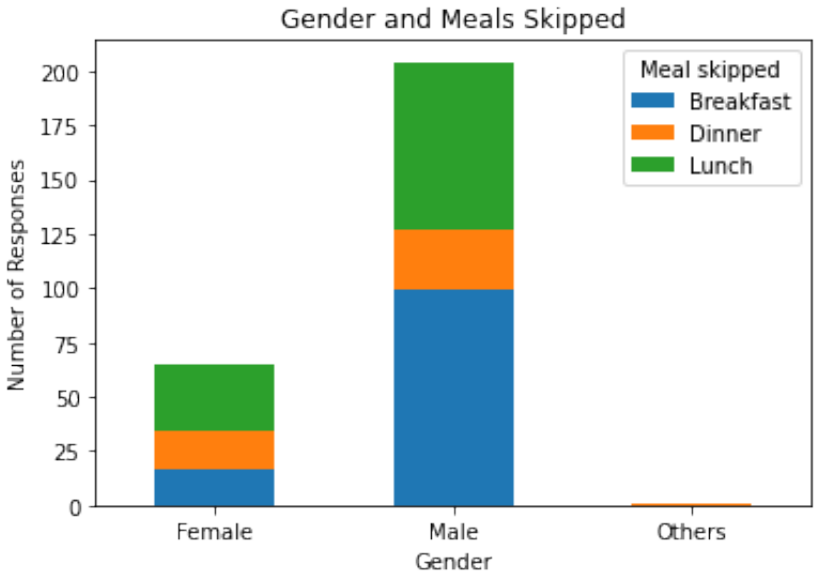

Figure 11: Skip meal-Gender segmented bar chart

First of all it is clearly evident that the number of males who skip meals are around 3 times the number of females who skip meals (during a week) this is probably because less females(24%) filled the form compared to males(76%). In females all the three meals are being equally skipped (lunch having slight upper hand). In males breakfast is being skipped the most, followed by lunch. Dinner is not missed much by both males and females.

Reasons for skipping the meal

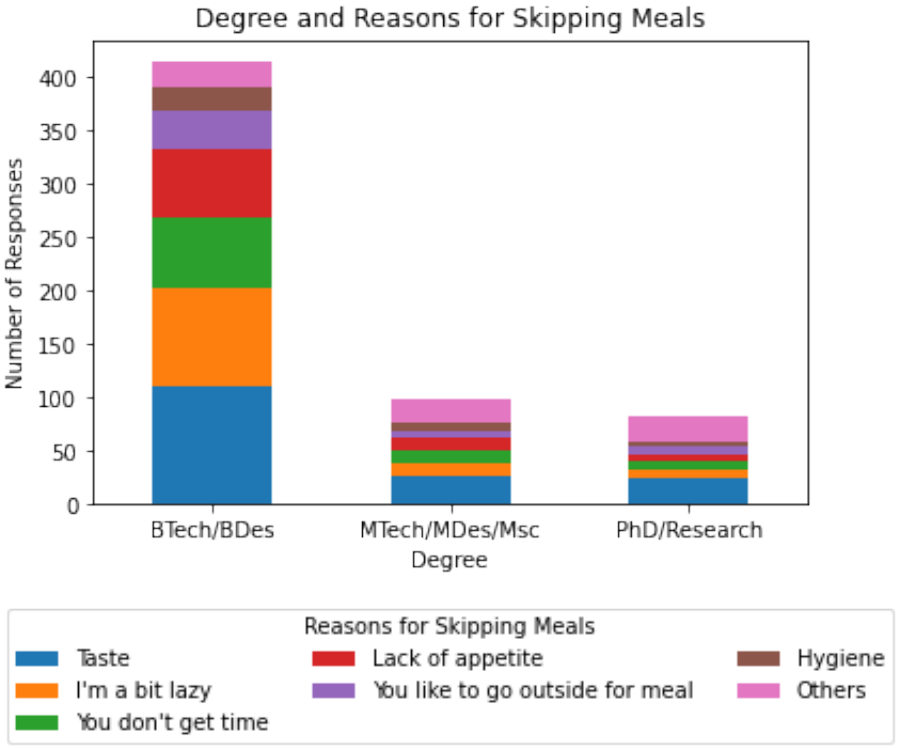

Figure 12: Skip meal-Reason segmented bar chart

Here we can see that the BTech/Bdes responded the most. Secondly among the UGs the most common reason to be observed of missing a meal is 'taste of the food and laziness', with some other less common reasons being, 'hygiene', 'lack of appetite', 'do not have time', etc. For Non-UGs the most common reason being 'taste of the food' and least common reson being 'hygiene'.

Correlation of sleep with missing meals

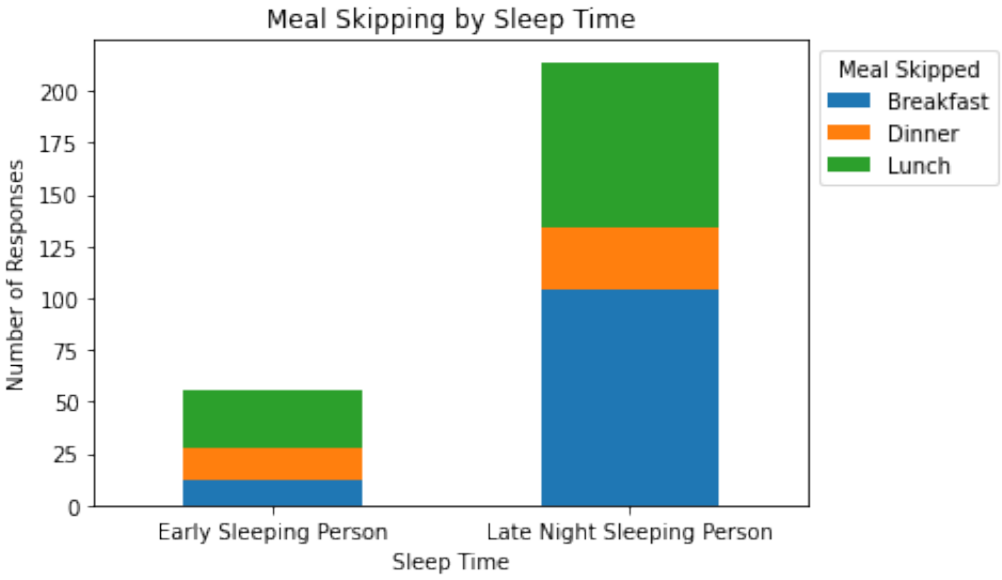

Figure 13: Skip meal-Sleep segmented bar charts

Table 1: Contingency table

		Meal Skipped			
		Breakfast	Lunch	Dinner	Total
Sleep Time	Early sleeping person	12	28	16	56
	Late night sleeping person	104	80	30	214
	Total	116	108	46	270

It is clearly evident that the late night sleepers miss their meals the most. They clearly miss out on breakfasts followed by lunch, dinner is being missed out the least (both early and late night sleepers). Early sleeping people do not miss their meals regularly.

We will have a look on some pie charts now

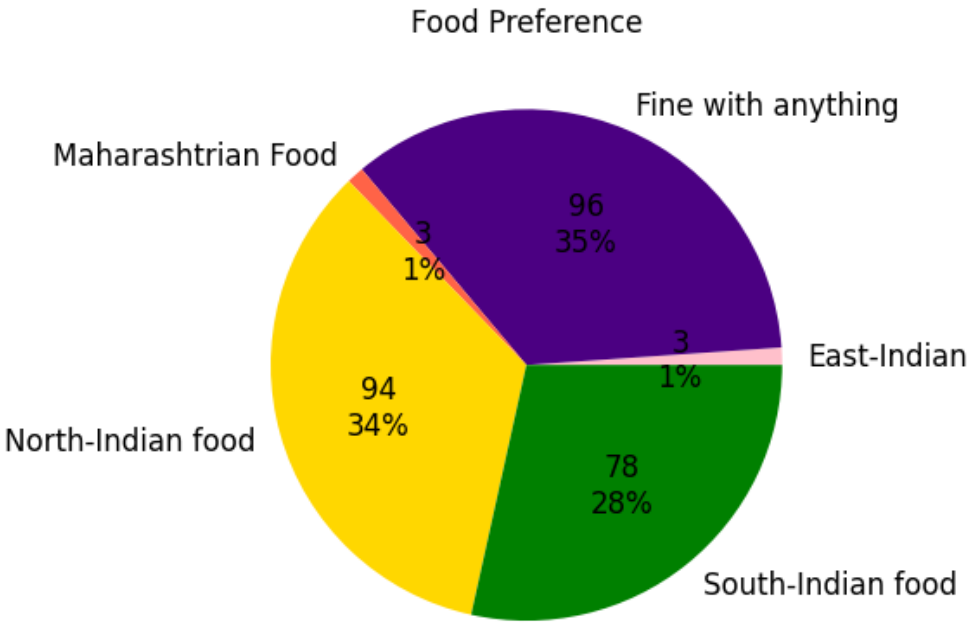

Figure 14: Food preference pie chart

This pie chart shows the regional food preferences of the people in IITH. We can observe that people are not very particular about their food preferences.

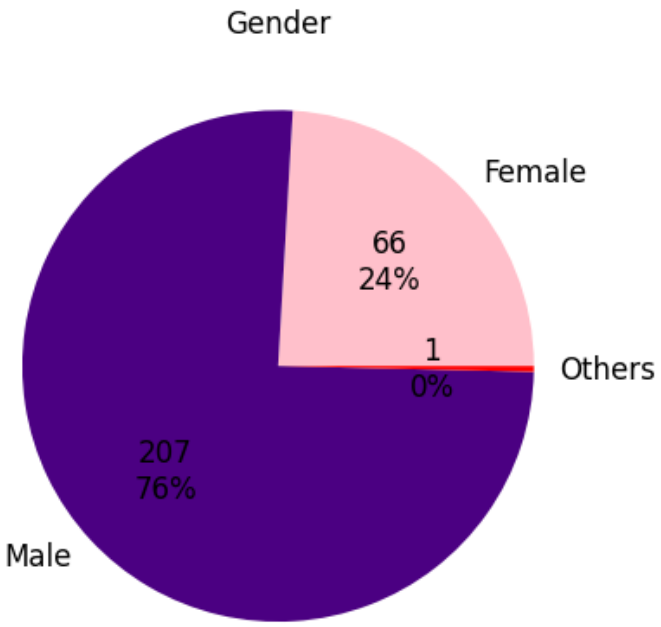

Figure 15: Gender Split

The charts compare male to female responses. we see that nearly one-third of the total responses came from males

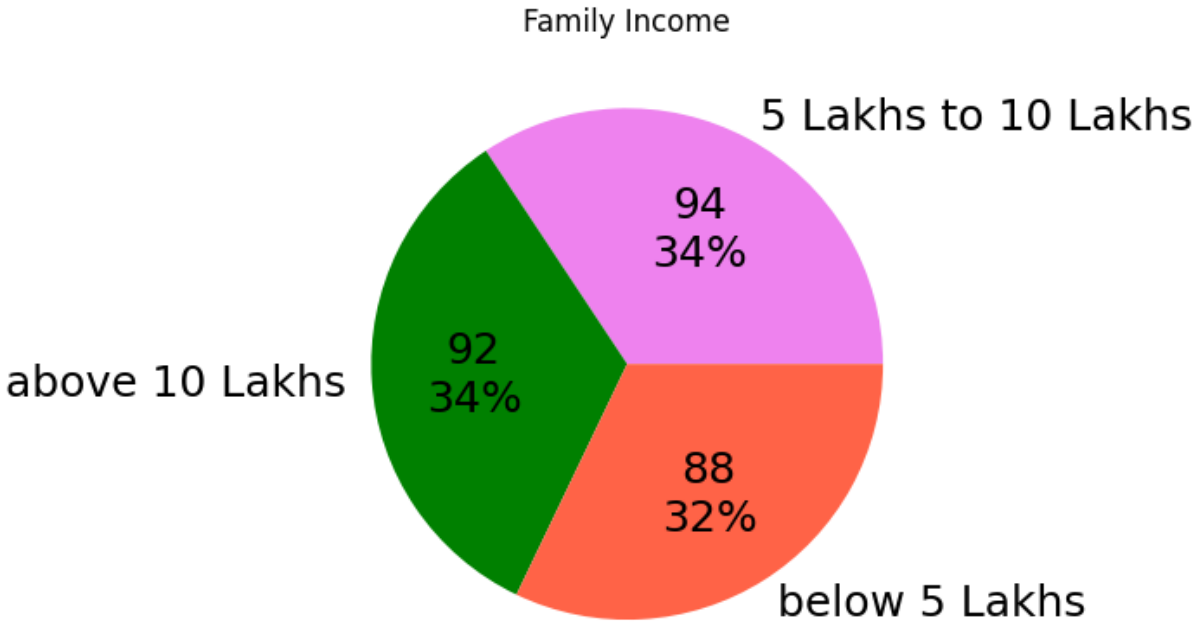

Figure 16: Student's Family Income

The chart shows the distribution of different economic classes in the responses. We can observe that it is very well balanced.

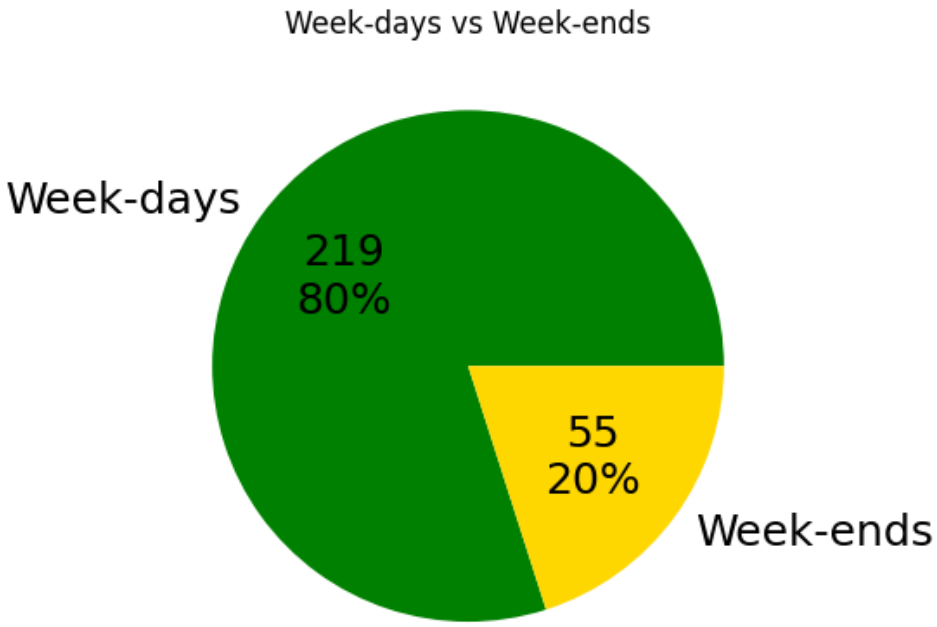

Figure 17: Skipping meals on weekends or weekdays

The pie chart above gives a comparison of skipping meals on weekends vs weekdays. We see the majority of skipping on weekdays. This may be because of specials on weekends.

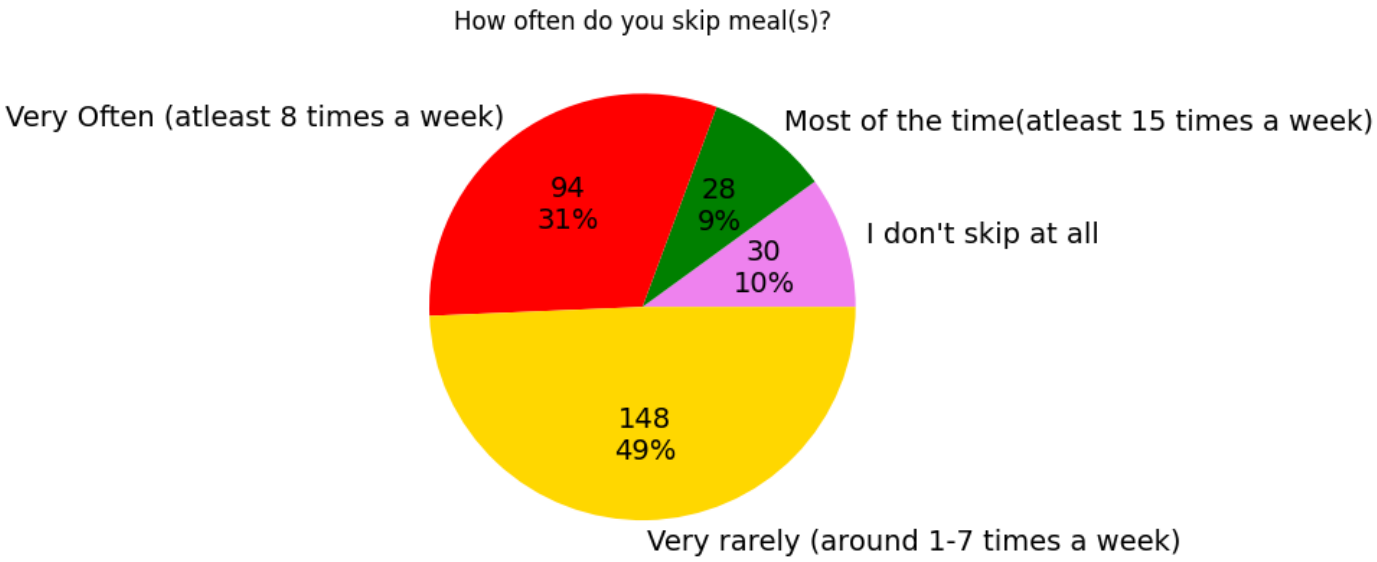

Figure 18: Frequency of missing food

This pie chart compares the frequency of meal skipping. We find that half of the people rarely miss a meal.

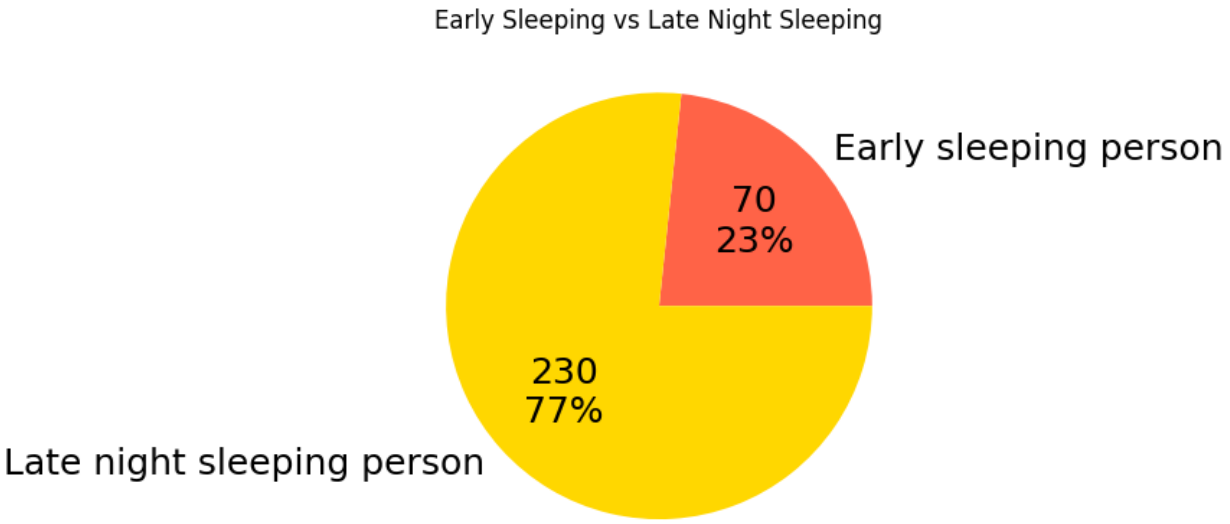

Figure 19: Students are early sleepers or late nighters

This pie chart compares Early sleepers vs Late sleepers in the responses. From this chart, we can infer that most are Late sleepers!!

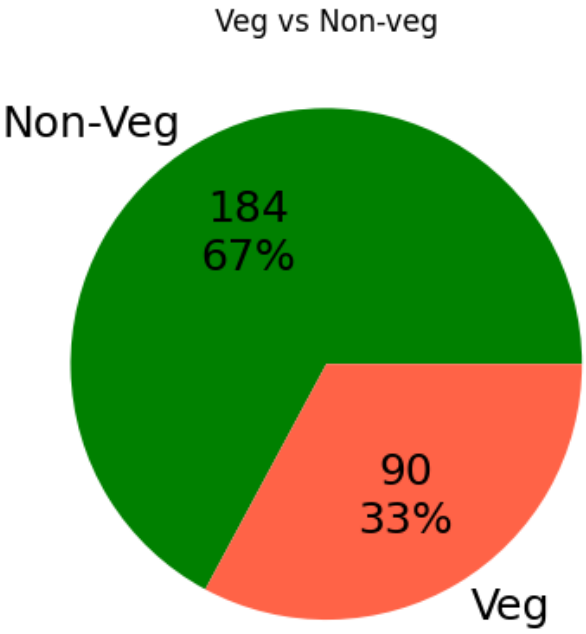

Figure 20: Vegetarian vs Non-Vegetarian

This pie chart compares Non-veg eaters vs veg eaters in the responses. From this chart, we can infer that there are 2 times as non-veg eaters as veg eaters!!

3 Confidence Interval Estimation of Mean and Variance of the data.

- **Case-1: Confidence Interval of μ, where μ is the mean number of meals skipped by a student at IITH in a week.**

Consider a random sample of size $n = 50$. Sample mean and sample standard deviation are $\bar{x}$ and S respectively

$$\bar{x} = 6.98$$
$$S = 4.18$$

Also, for a 95% CI, $\alpha = 0.05$,$t_{\alpha/2,n-1} = t_{0.025,49} = 2.0096$
Now, the resulting 95% CI is:

$$\bar{x} - t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$$
$$5.792 \leq \mu \leq 8.168$$

Thus, based on the sample data, the CI of the mean number of meals skipped by a student in a week is $[5.792, 8.168]$

- **Case-2: Confidence Interval of σ,where σ is the standard deviation of the population.**
Consider a random sample of size $n = 50$.
Here sample variance, $S^2 = 17.49$
Take, significance level $\alpha = 0.05$

$$a = \chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,49} = 67.505$$
$$b = \chi^2_{\alpha/2,n-1} = \chi^2_{0.025,49} = 30.096$$

Now, the resulting 95% CI is:

$$\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}$$
$$12.695 \leq \sigma^2 \leq 28.476$$

Thus,the obtained CI of σ^2 is $[12.695, 28.476]$
This leads to 95% CI for σ :$[3.563, 5.336]$

4 Hypothesis Testing

- **Case-1: Verifying if the mean count of students who skip mess during weekdays is greater than the mean count of the students who skip mess during week-ends with the level of significance 0.05.**
Here,
Sample 1 is that of the students who skip mess during weekdays.
Sample 2 is that of the students who skip mess during weekends.

$$\text{Sample sizes : } n_1 = 216 \quad n_2 = 54.$$

As both the sample sizes are greater than 30, The condition that population distributions are normal is satisfied.
Let's now declare the Null and Alternate Hypothesis
H_0 : The mean count of students who skip mess during weekdays is less than or equal to that of students who skip mess during weekends.
H_a : The mean count of students who skip mess during weekdays is greater than that of students who skip mess during weekends.

$$H_0 : \mu_1 - \mu_2 \leq 0 \text{ vs } H_a : \mu_1 - \mu_2 > 0$$
$$\text{Sample Means : } \bar{x_1} = 7.643 \quad \bar{x_2} = 5.907$$
$$\text{Sample Variances : } {S_1}^2 = 23.396 \quad {S_2}^2 = 15.935$$

Here, $\frac{{S_1}^2}{{S_2}^2} = 1.468$ which is less than 4. So, we can assume that both the variances are equal. Let's now calculate the degrees of freedom and pooled variance:

$$\text{degrees of freedom } df = n_1 + n_2 - 2 = 268$$
$$\text{pooled variance } {S_p}^2 = \frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{df} = 21.92$$

Also here the Total Population data i.e., entire IITH Student Community data is not available so the Population Variance(σ^2) remains unknown and hence test statistic is that of "t".
Test Statistic: :

$$t = \frac{(\bar{x_1}-\bar{x_2})-D_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n2}}} = 0.5205$$

Rejection Region Approach : Reject H_o if $t \geq 1.6506$ where $t_{0.05,268} = 1.6506$ (here considered $\alpha = 0.05$).
Result: Because the observed value of t = 0.5205 is less than 1.6506 and hence is not in the rejection region, there is insufficient evidence to conclude that the mean count of students who skip mess during week-days is greater than the mean count of the students who skip mess during weekends.

- **Case-2: Verifying if the mean count of Undergraduate students who skip mess is different from 6 with 5% as level of significance.**
Here, the Sample is the Undergraduate students who skip mess.
Sample size $n = 127 > 30$, So the condition that the sample comes from Normal Population Distribution is satisfied.
Now, we set up the research hypotheses:
H_0 : Mean count of Undergraduate students who skip mess is 6.
H_a : Mean count of Undergraduate students who skip mess differs from 6.

$$H_0 : \mu = 6 \quad \text{vs} \quad H_a : \mu \neq 6$$
$$\text{Sample Mean : } \bar{x} = 7.285$$
$$\text{Sample Standard deviation: } S = 4.572$$

Test Statistic :

$$t^* = \frac{\bar{x}-\mu_0}{S/\sqrt{n}} = \frac{7.285-6}{4.572/\sqrt{127}} = 3.1673$$
$$\text{degrees of freedom : } df = 127 - 1 = 126$$
$$\text{Level of significance : } \alpha = 0.05$$

p-value Approach: Since H_a is two-tailed,

$$p - value = 2 \times P(t > |t^*|) = 2 \times P(t > |3.1673|)$$

Since we do not find the exact value of 3.1673 in the t-table at 126 d.f., we try to find a range. It can be seen from the t-table that the value falls between 3.1562 and 3.1892, and corresponding to them, the right tail probabilities are 0.001 and 0.0009, respectively.
$\therefore$ The p-value would be between $2 \times (0.0009) = 0.0018$ and $2 \times (0.001) = 0.002$.
Result: Since p-value is less than α, there is significant evidence that the mean count of Undergraduate students who skip mess is different from 7.

- **Case-3: Verifying if the variability in count of students of age 19-20 years who skip mess is less than 7 with 0.05 as the level of significance.**
Here, Sample is the 19-20 years age students who skip mess. Let's now declare the null and alternative hypothesis,
H_0 : Variability in count of students of age 19-20 years who skip mess is greater than or equal to 7.
H_a : Variability in count of students of age 19-20 years who skip mess is less than 7.

$$H_0 : \sigma^2 \geq 7 \text{ vs } H_a : \sigma^2 < 7$$
$$\text{Sample Variance : } S^2 = 16.497$$
Sample size $n = 124 > 30$, So we can assume that population distribution is Normal.

Test Statistic:

$$\chi^2 = \frac{(n-1)S^2}{{\sigma_0}^2} = \frac{123 \times 16.497}{49} = 41.4108$$

Rejection region Approach: Reject H_0 if the value of TS is less than 149.8846, for df = n-1 = 123 and $1 - \alpha = 0.95$.
Result: Since the computed value 41.4108 is less than the critical value of 149.8846, there is sufficient evidence to reject H_0 i.e.; there is significant evidence that the variability in count of students of age 19-20 years who skip mess is less than 7.

- **Case-4: Verifying if the percentage of AI and CSE students who skip mess is greater than 0.1 with level of significance 0.05.**
Here, Sample is the count of AI and CSE department students who skip mess.
Declaring null and alternate hypothesis,
H_0 : Percentage of AI and CSE department students who skip mess is less than or equal to 0.1.
H_a : Percentage of AI and CSE department students who skip mess is greater than 0.1.

$$H_0 : \pi \leq 0.1 \text{ vs } H_a : \pi > 0.1$$
$$\text{sample size } n = 41 > 30$$

Test Statistic :

$$Z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}}$$

From the survey data,
$\hat{\pi} = \frac{41}{270} = 0.1518$ and $\sigma_{\hat{\pi}} = \sqrt{\frac{0.1518(1-0.1518)}{270}} = 0.0218$
Also,

$$n(\pi_0) = 270(0.155) = 40.986 > 5 \text{ and } n(1 - \pi_0) = 270(1 - 0.155) = 229.014 > 5.$$

Thus, the sample considered is valid and we obtain :

$$Z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}} = \frac{0.1518 - 0.1}{0.0218} = 2.3761$$

Rejection Region Approach: Reject H_0 if $Z > 1.645$ for $\alpha = 0.05$
Result: Since the observed value of Z exceeds the critical value of 1.645, we conclude there is significant evidence that the percentage of AI and CSE department students who skip mess exceeds the percentage of 10%.

5 Contributions

1. Blessy Anvitha J
   - Ideation of the project
   - Data Collection
   - Hypothesis Testing
   - LATEX report on Hypothesis Testing
2. Saanvi Amrutha
   - Data Collection
   - Central Tendencies of entire Sample
   - Normality-plot, Box plots and Confidence Intervals
   - LATEX report
3. Bhanu Prasad

- Data Cleaning
- Data Pre-Processing for coding
- Python Coding for generating bar-charts and pie-charts

4. Anirudh Srinivasan
    - Data Collection
    - Idea of the Project
    - Python Coding
5. Lokesh Surana
    - Data Collection
    - Python Coding
    - LATEX Report
6. Arnav Asati
    - Slides of Data Visualisation
    - Results and Analysis
    - Remaining LATEX report
7. Samar Singhai
    - Slides of Data Visualisation
    - Results and Analysis
    - LATEX Report
8. Shivanand
    - Slides of Data Visualisation
    - Results and Analysis
    - LATEX report