Assignment: Survey Exploration

Team Members: Chaithanya Pramodh Kasula and Aishwarya Varala

a) Identify and clean up any data items that need to be made uniform or transformed. Explain what you did.

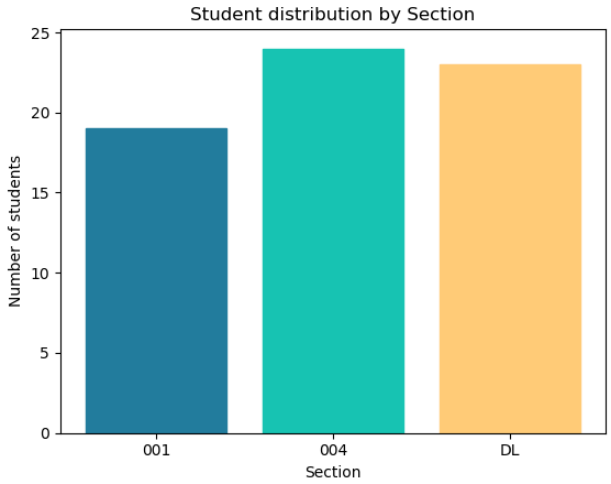
Ans: The survey data needs a lot of preprocessing. The below steps were performed during pre-processing.

1. The first two rows of the dataset were combined to form the column names of the dataset.
2. Each column represents an option of a question. The three columns 001, 004, DL1 columns represent answers for the question Q1. The value 1 in each of these represents that a student belongs to that particular section. Instead of having three columns, we can merge them into a single column. Before merging, a simple encoding process has been performed. Considering the same example, all the 1's in the '004' column have been converted to 2. All the 1's in the 'DL' column have been converted to 3. The three columns have been merged into a single column and the name of the column has been converted to 'Q1'. In the refactored column, 0 represents 'Unknown', 1, 2, and 3 represent that the student belongs to '001', '004', and 'DL' sections respectively.
3. The columns for the options in Q2, Q8, Q10, and Q11 have been made to undergo the same preprocessing mechanism.
4. The NaN values in the 'Q3_Age (years)' column has been replaced with the mode of the column.
5. The NaN values in the 'Q4_Height (Inches)' column have been replaced with the mode of the column.
6. The NaN values in 'Q5_Country of Citizenship', 'Q6_Undergraduate Degree', and 'Q7_Expected Graduation date from Mason MS program?' have been replaced with the string 'Unknown'.
7. The NaN values in 'Q9_Commuting time (minutes) from home/work to campus for class?' column have been replaced with the mean value of the column.
8. The values in the 'Q9_Commuting time (minutes) from home/work to campus for class?' consist of may strings that represent the same country. For example, the column consists of strings like, 'US', 'UNITEDSTATESOFAMERICA', 'UNITEDSTATES' that represent the same country, USA. Such transformations have been made.
9. For question 6, it was observed that most of the students have mentioned their undergraduate degree as B.E or B.Tech instead of their specialization such as Computer Science or Electrical Engineering, etc. Hence, such responses have been categorized under 'Uncertain Specialization'.

b) Create summary statistics and visualizations for each of the 11 questions, categorized by section (001, 004, DL1). For each, explain any interesting differences that you observe (or indicate no real difference)

For question 1 (Q1): AIT-580 Section?

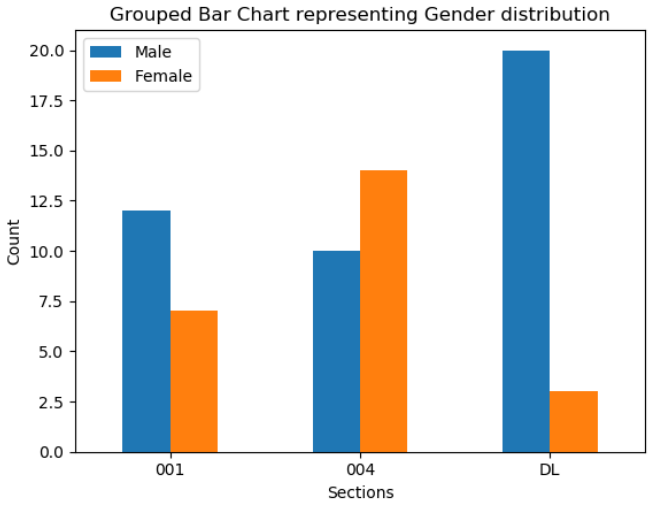
Interesting Observation: The number of students enrolled in the 004 section is the highest when compared to the other two sections.



For question 2 (Q2): Gender?

Interesting Observations:

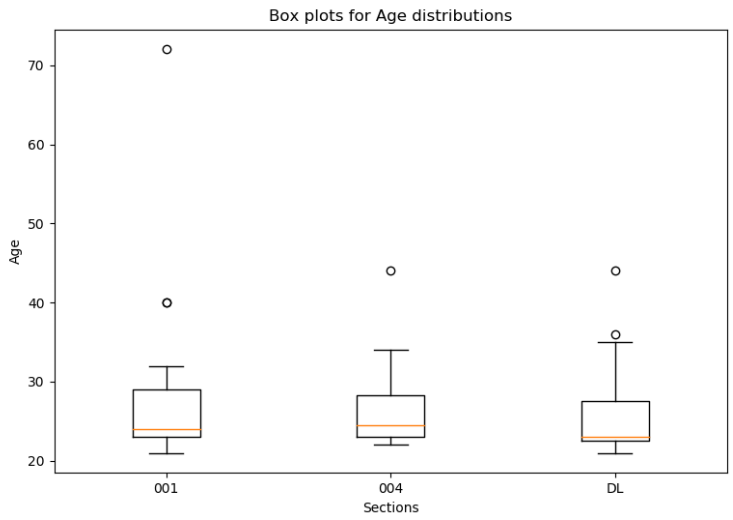
- The number of males in the DL section is the highest when compared to the other two sections.
- The number of females in the 004 section is the highest when compared to the other two sections.
- The number of students enrolled in the 004 section is highest in number.



For question 3 (Q3): Age (years)

Interesting Observations:

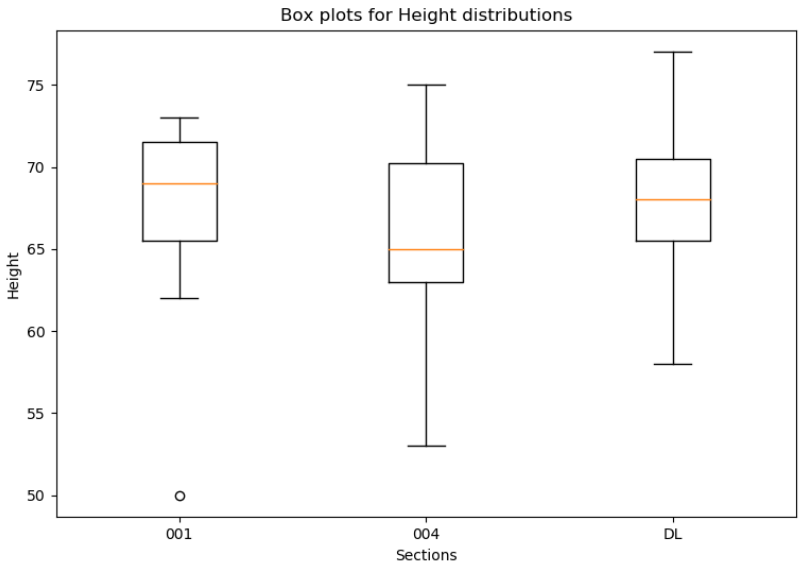
- The median age is highest for students in the 004 section and lowest for the DL section.
- 001 section has the highest outlier value of 72 years. The number of outliers for the 001 section and DL section is equal.
- The interquartile range (IQR) range for the section 001 section is the highest.



For question 4 (Q4): Height (Inches)

Interesting Observations:

- The median height is the highest for the students belonging to the 001 section.
- The Q1 value for the DL section is the highest among the other two.
- The interquartile range (IQR) for section 004 is the highest.

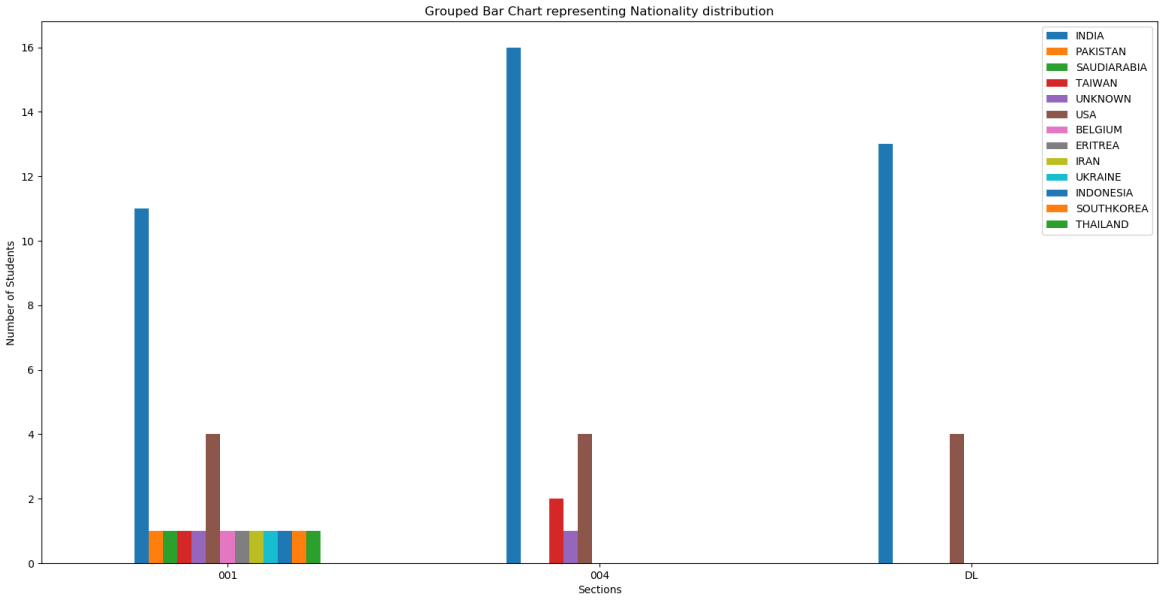


For question 5 (Q5): Country of Citizenship

Interesting Observations:

- Indian students are predominant in all the sections
- The DL section consists of students only from India and the USA.
- The highest diversity with respect to nationality is observed in the 001 section and the lowest is observed in the DL section.

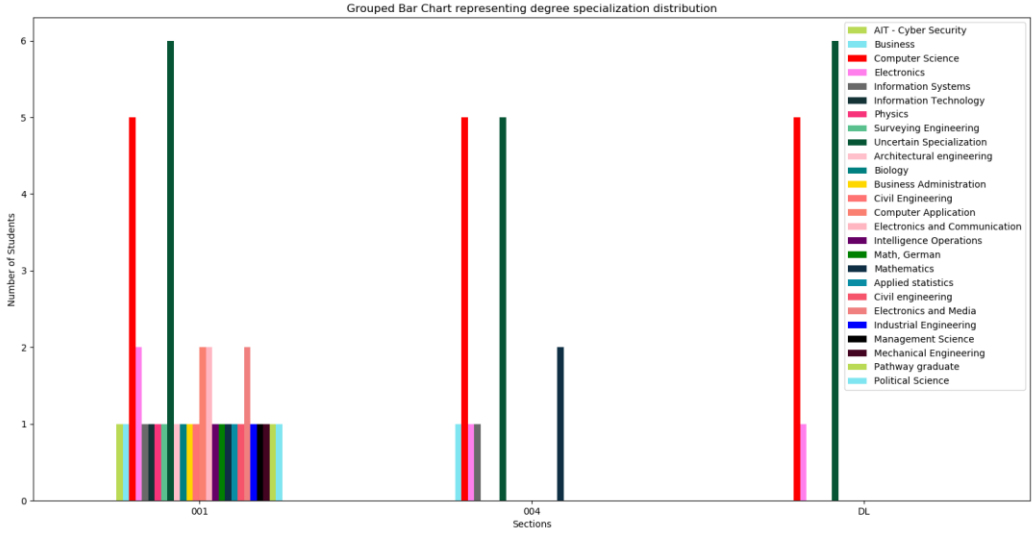
- The highest number of Indians are present in 004 section.



For question 6 (Q6): Undergraduate Degree

Interesting Observations:

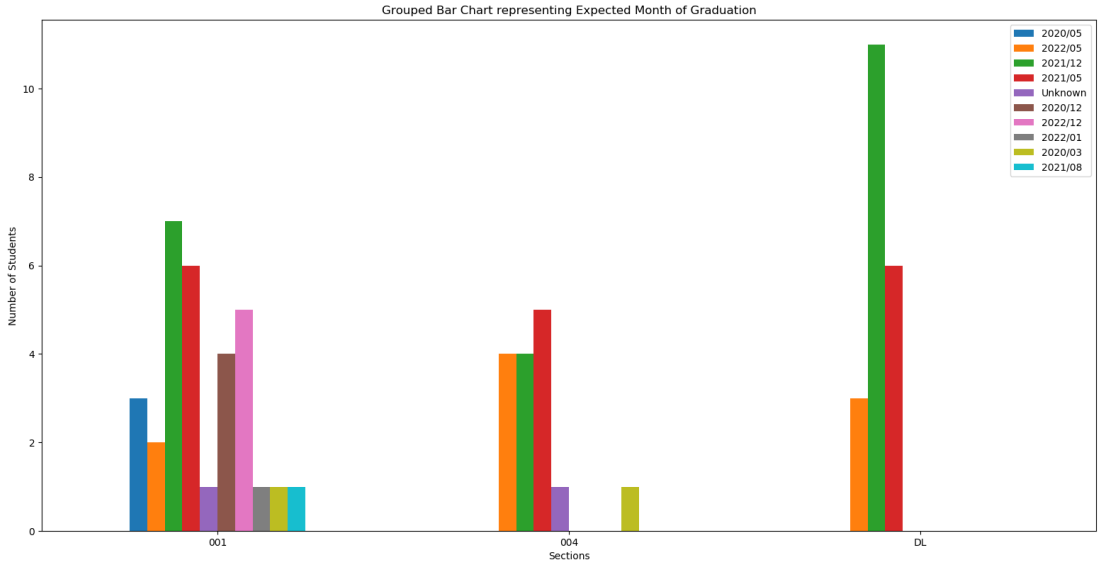
- The number of students who didn't specify their undergraduate specialization was categorized into one category named Uncertain Specialization.
- Section 001 contains students with the most diverse specialization in their undergraduate degrees.
- Computer Science students are present in the largest number in all the sections.
- DL section contains students with the least diverse specialization in their undergraduate degrees.



For question 7 (Q7): Expected Graduation date from Mason MS program?

Interesting Observations:

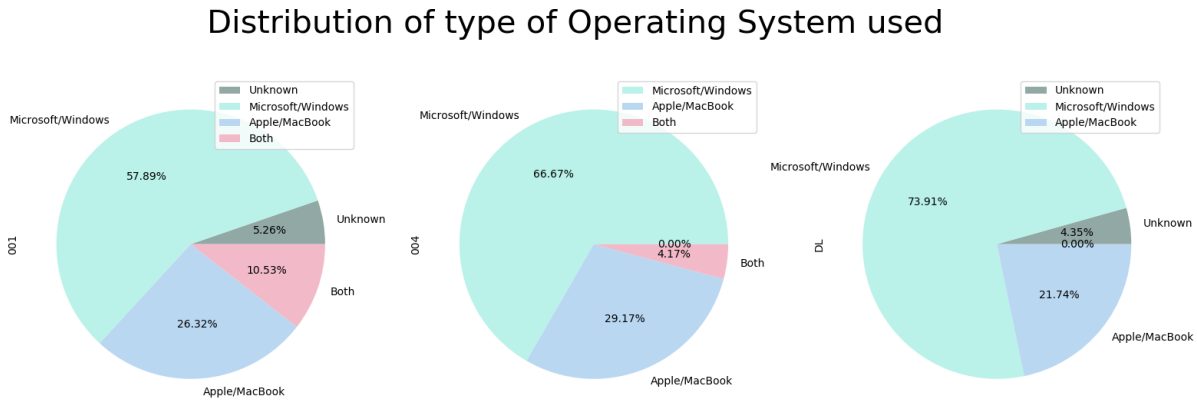
- Students from all the sections are expected to graduate at 9 different times.
- In the 001 and DL sections, the majority of the people are expected to graduate in 2021/12.
- In the 004 section, the majority of the people are expected to graduate in 2021/5.



For question 8 (Q8): Type of Laptop Used?

Interesting Observations:

- The majority of the students in three sections use Microsoft/Windows as their Operating System.
- In sections 001 and 004, some students use both the Operating Systems.
- In the DL section, there are no students who use both the operating systems.



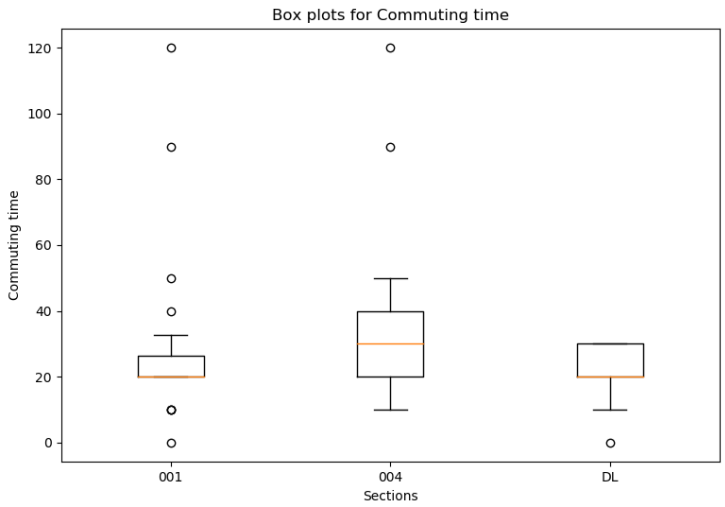
For question 9 (Q9): Commuting time (minutes) from home/work to campus for class?

Interesting Observations:

The interquartile range (IQR) is highest for the 004 section.

The number of outliers is highest in the 001 section.

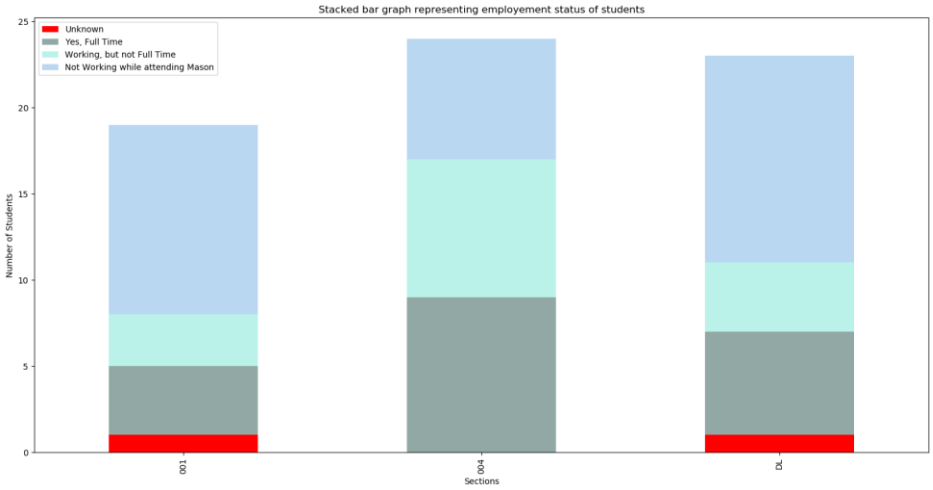
The DL and 001 section consists of students with 0 commuting time.



For question 10 (Q10): Are you employed full-time while attending Mason?

Interesting Observations:

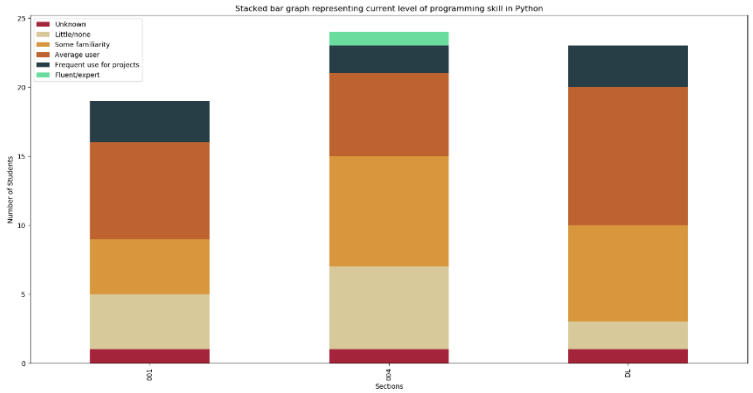
- Section 004 has the highest number of students who are working full time.
- The DL section has the maximum number of students who are not working while attending mason.



For question 11 (Q11): What is your approximate current level of programming skill in Python?

Interesting Observations:

- The students who have selected their approximate current level of programming skill in Python as 'Average user' of Python is highest in the DL section.
- There is only a student among all the sections who have selected 'Fluent/expert' and can be found in section 004.
- The number of students who selected 'Little/none' is minimum in the DL section.



c) Discuss some suggestions for how to improve the online learning experience for this experiment (10 pts).

1. Essential questions should be made mandatory to answer. It can also be marked with a * symbol.
2. 'Country of Citizenship' can be made a dropdown instead of a text field. That way, different variations in text that denote the same country can be avoided. For example, it was found that USA was written in different formats such as United States of America, US, etc.
3. The questions are vague and are not specific. For example, Q6 was formed as 'Undergraduate Degree'. It was not clear whether the name of the degree or the specialization needed to be entered. The same was found in the answers. The options can be provided as a dropdown.

4. For the question, ‘Expected Graduation date from Mason MS program?’, many students specified the date and different months even when they were expected to graduate in the same semester. A dropdown could have been provided to overcome this issue.
5. Question 8 was formed as ‘Type of Laptop Used?’ but this is incorrect, and it should have been ‘Type of OS Used?’
6. The framing and the language the questions could have been better. For example, ‘AIT-580 Section?’ can be rewritten as ‘Which section among the below AIT-580 sections do you belong to?’