**Project Name:** Most Difficult Subject in CCS Department: According to 2nd Year BSIT students during their 1st Year, Second Semester (A.Y 2022-2023).

**Member(s):** San Luis,Richjehn

Villodo, Nina

Oliveros,Roldan

Labanda,Mary Ann

Javier,Jonamie

Dayapera,Nel

Elca,Eirich

Mondez,Merce Ynnofranz

**Date:** December 20,2023 -January 2,2024

**1. Introduction**

This document outlines the objectives and goals of the data analysis project conducted within the College of Computer Studies (CCS) Department. The primary focus of this analysis is to ascertain the subject perceived as the most challenging by 2nd Year Bachelor of Science in Information Technology (BSIT) students during their 1st Year, Second Semester in the academic year 2022-2023.

**Project Objectives:**

1. Identify the Most Difficult Subject:

Determine the specific subject that 2nd Year BSIT students found to be the most challenging during their 1st Year, Second Semester.

1. Understand Student Perspectives:

Investigate and analyze the underlying reasons and contributing factors to the perceived difficulty of the identified subject. This includes an examination of teaching methodologies, course content, and assessment criteria.

1. Enhance Student Experience:

Aim to improve the overall educational experience for BSIT students by addressing concerns related to subject difficulty. This contributes to the creation of a more supportive and effective learning environment.

1. Facilitate Continuous Improvement:

Establish a systematic feedback mechanism to ensure that the insights gained from the analysis inform continuous improvement. This involves incorporating feedback into adjustments to teaching methods or curriculum design for subsequent semesters.

This data analysis project is undertaken with the intention of contributing valuable insights to the ongoing enhancement of educational programs within the CCS Department, fostering an environment that promotes academic excellence and student success.

The specific data being visualized in this project pertains to the perceived difficulty of subjects within the College of Computer Studies (CCS) Department. The focus is on feedback from 2nd Year Bachelor of Science in Information Technology (BSIT) students regarding the most challenging subject during their 1st Year, Second Semester in the academic year 2022-2023.

**Types of Data for Visualization:**

**Subject Ratings:** Data reflecting the perceived difficulty of individual subjects as rated by 2nd Year BSIT students. This may be collected through surveys, interviews, or other feedback mechanisms.

**Student Feedback:** Qualitative data, such as comments or reasons provided by students regarding why they find a particular subject challenging. This information adds context to the quantitative ratings.

**Demographic Information:** Relevant demographic data, such as gender or academic performance, to explore potential variations in the perceived difficulty among different groups of students.

**Data Sources:**

**Student Surveys/Feedback Forms:** Responses collected directly from 2nd Year BSIT students through structured surveys or feedback forms designed to capture their opinions on subject difficulty.

**Interviews/Focus Groups:** Qualitative data obtained through interviews or focus group discussions with students to delve deeper into their perspectives and reasons for perceiving certain subjects as difficult.

**Academic Records:** Relevant academic data, such as grades or performance metrics, which may provide additional context to the perceived difficulty of subjects.

**Course Evaluations**: Existing course evaluation data, if available, from the academic year in question, which might contain insights into students' experiences with specific subjects.

* What questions are you hoping to answer with this visualization?
* To gather a simplified data interpretation, so that we could help prepare solutions based on the data gathered by our team.

**2. Data**

* Describe the data set in detail:
  + Number of rows and columns
    - Rows:11
    - Columns 6
  + Data types of each variable
    - String:
      * np
      * plt
      * pd
      * sm
      * student\_df
      * subject\_counts
      * Data
      * Data1
      * Data2
      * Data3
      * Data4
      * Data5
      * Data6
      * Data7
      * Data8
      * Data9
      * x
      * y
  + Missing values and how they were handled
  + Outliers and how they were addressed
  + Data transformations performed (e.g., normalization, scaling)

**3. Visualization Technique(s)**

* **List the types of visualizations used** 
  + Bar chart
  + Scatter plot
* **Justify your choice of visualization for each type of data.**
  + Bar Charts

- Bar chart is one of the most common used Charting diagrams to represent data. It can be easily interpreted by anyone who finished Elementary school. It is also straight forward and when representing data its clarity and visualization are easy enough to understand.

* + Scatter plot
* This kind of Data Visualization method uses dots to represent values of related variables. On the analysis of Interactive Activities vs Internet Connection Impact, the dots represent the scores from 1 to 5 on both variables. Observing the dots on the scatter plot shall show the relationship of the two variables.
* As seen in the scatter plot, the relationship of two variables named “Internet Connection Impact Rating” and “Interactive Activities Rating” are visualized. There were results that shows having Low Internet Connection Impact Rating would also have low Interactive Activities Rating. At Medium Internet Connection Impact Rating, there would be Low to Medium Interactive Activities Rating. At high Internet Connection Impact Rating, there are Medium to High Interactive Activities Rating.
* **Explain how the visual elements (e.g., color, size, shape) encode the data.**

1. Bar Chart (Chosen Subjects):

Color: Uniform blue color used for all bars. The height of the bars encodes the frequency of each subject.

1. Histograms (Various Scenarios):

Color: Again, a consistent blue color for all histograms.

Height (Y-axis): Frequency of each rating (encoded on the y-axis).

Rating (X-axis): The x-axis represents different rating levels, aligned with bins representing the possible responses.

Title, X-axis Label, Y-axis Label: Textual elements providing context and labels.

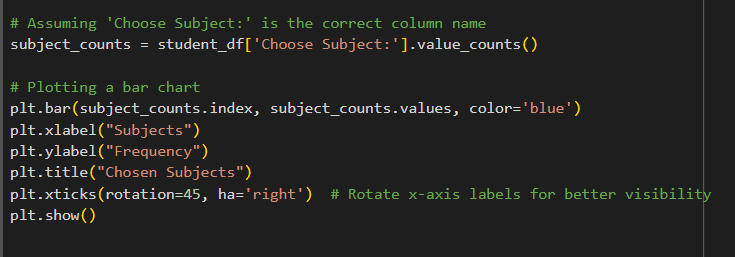
1. X-axis Manipulation:

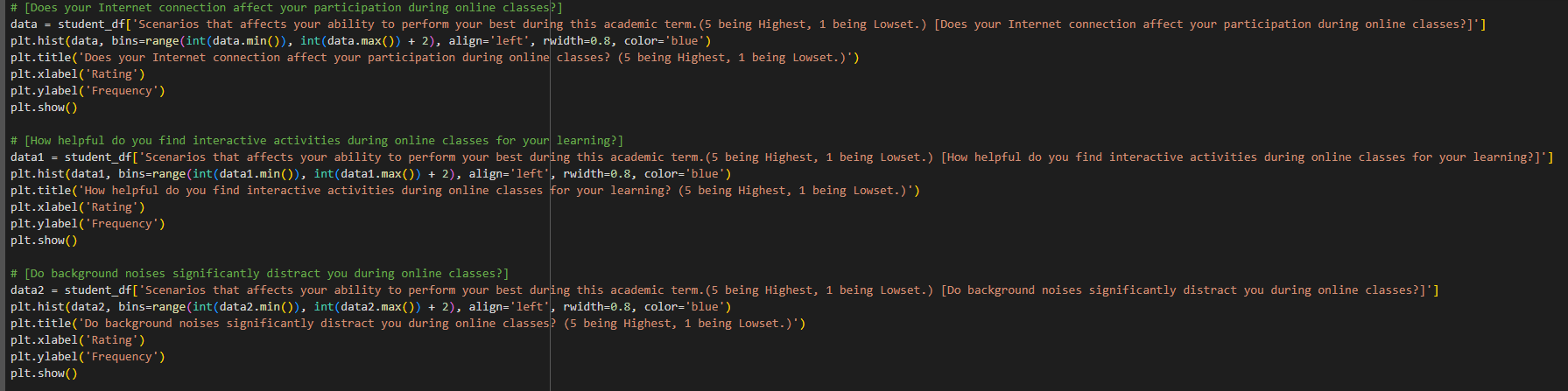
Rotation of Labels: The x-axis labels are rotated by 45 degrees and aligned right for better readability when they might overlap due to length or density.

* **Mention any additional libraries or packages used for visualization**

1. Seaborn: Built on top of Matplotlib, it provides a high-level interface for drawing attractive and informative statistical graphics.
2. Plotly: Offers interactive plots and is well-suited for creating dashboards. It allows for creating interactive, web-based visualizations.
3. Bokeh: Focuses on creating interactive plots, suitable for web browsers, and offers more flexibility for creating interactive visualizations.
4. Altair: Declarative statistical visualization library in Python. It's quite intuitive and produces Vega-Lite visualizations.
5. ggplot: Based on the R's ggplot2 library, it provides a similar grammar for graphics in Python.
6. Folium: Specifically designed for visualizing geospatial data. It's particularly useful for creating maps and geographical visualizations.
7. Plotnine: A grammar of graphics-based plotting system in Python based on ggplot2.

**4. Implementation in Google Collab**





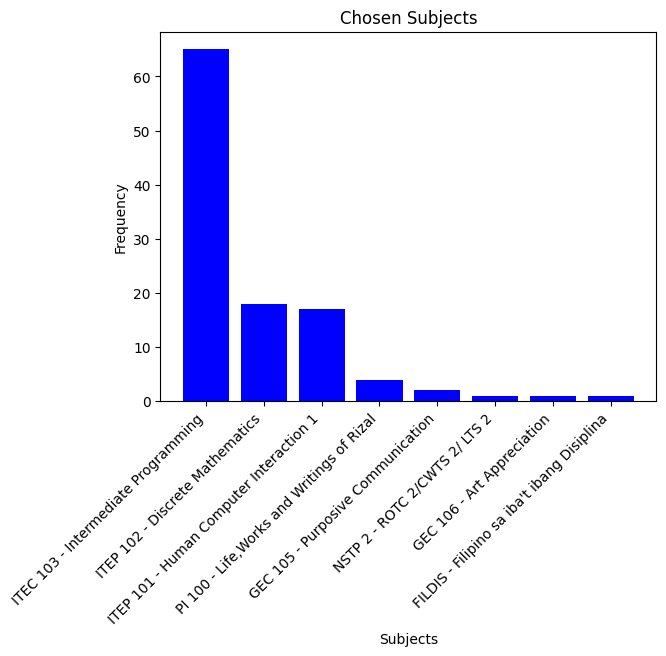
* Explain the key steps involved in the code:
  + Data loading and cleaning

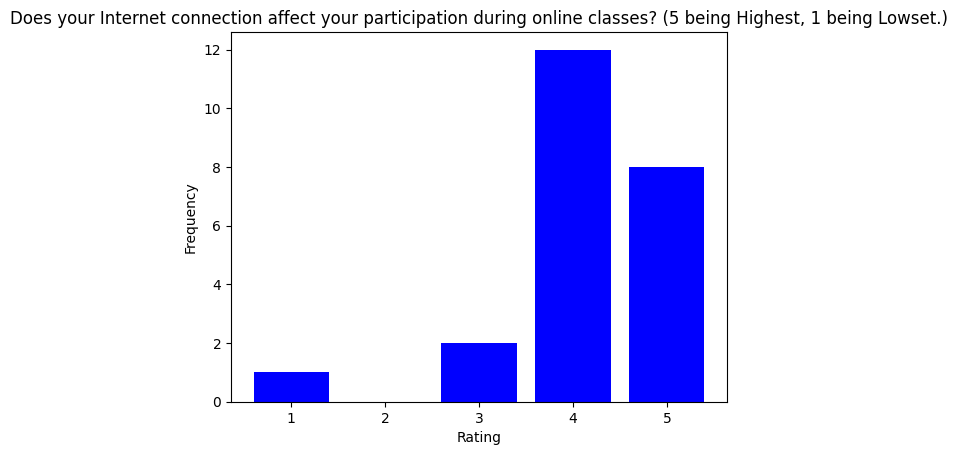
- imported pandas as pd and numphy as np

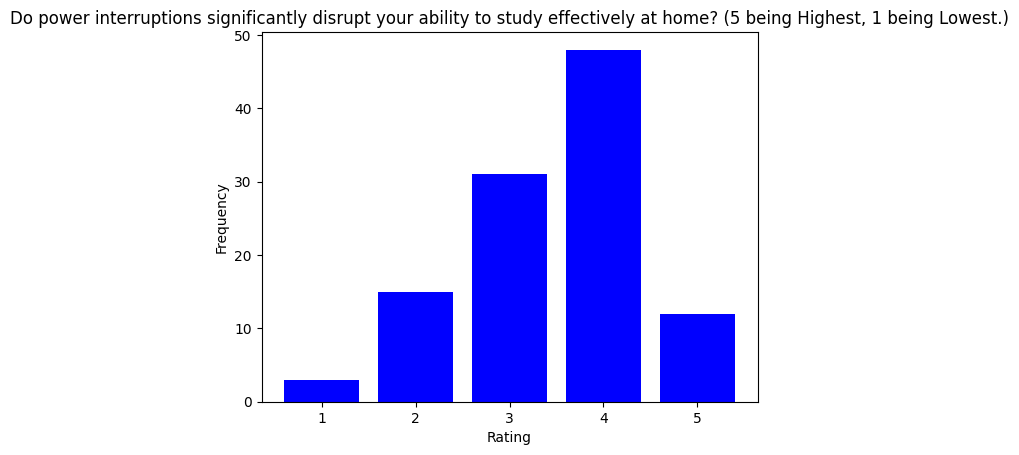
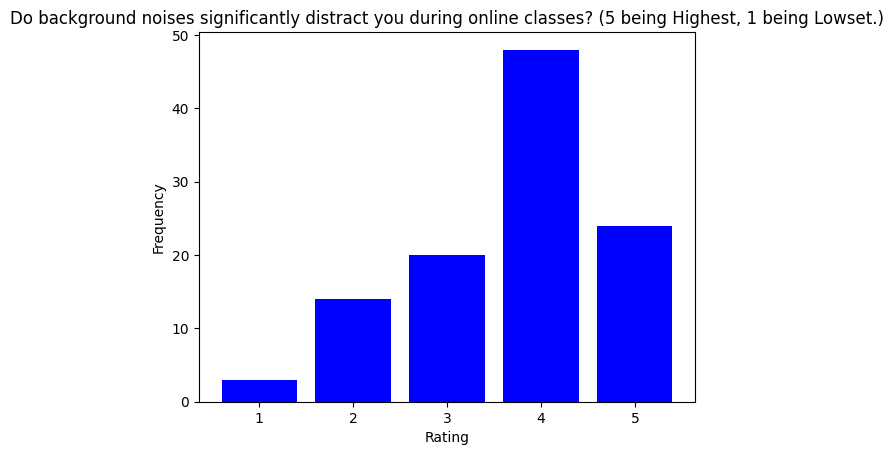
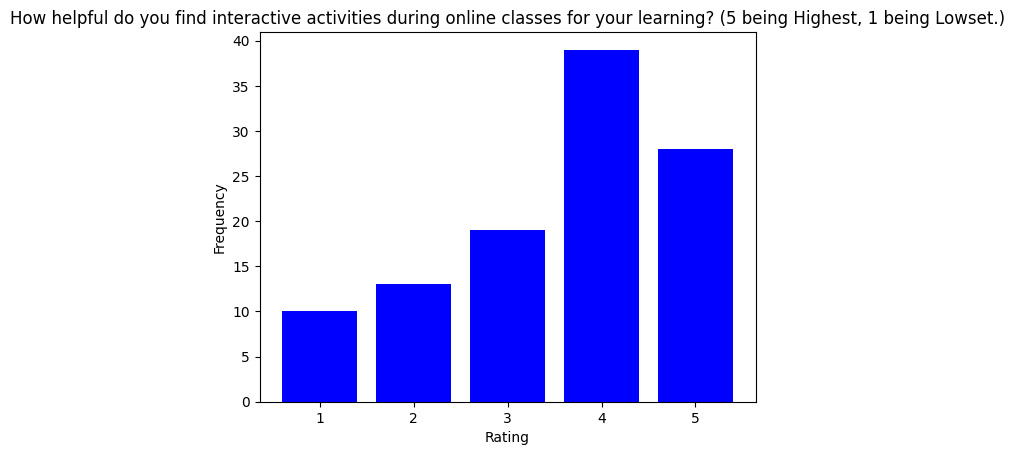
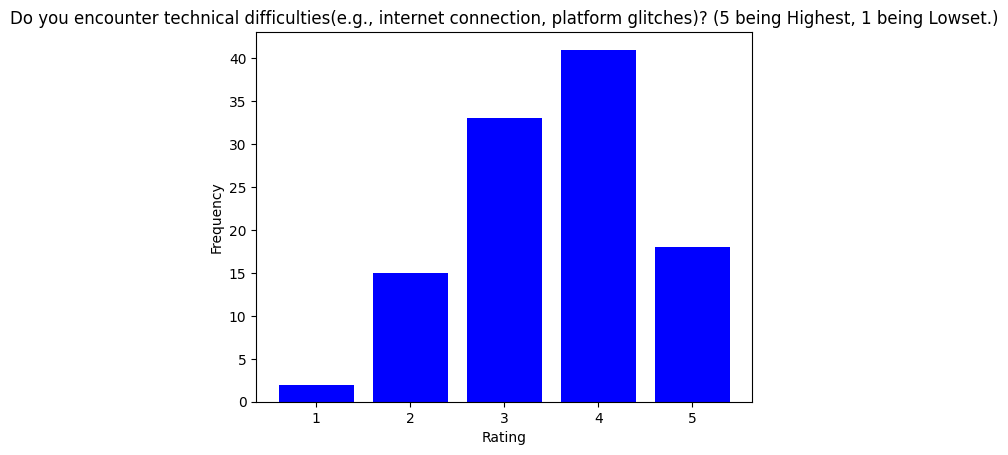
* + Visualization construction
* import matplotlib.pylot as plt to create graphs
* import statsmodels.api as smCustomization and styling
* **Highlight any challenges faced and how they were overcome.**
* Plotting the data on the scatter plot design become very challenging as the data gathered was insufficient so we decided to add the missing variables.

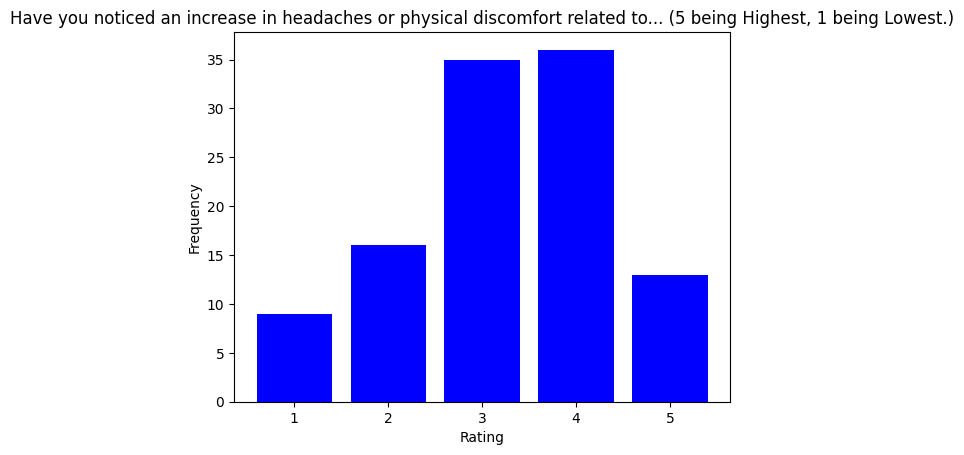
**5. Results and Interpretation**

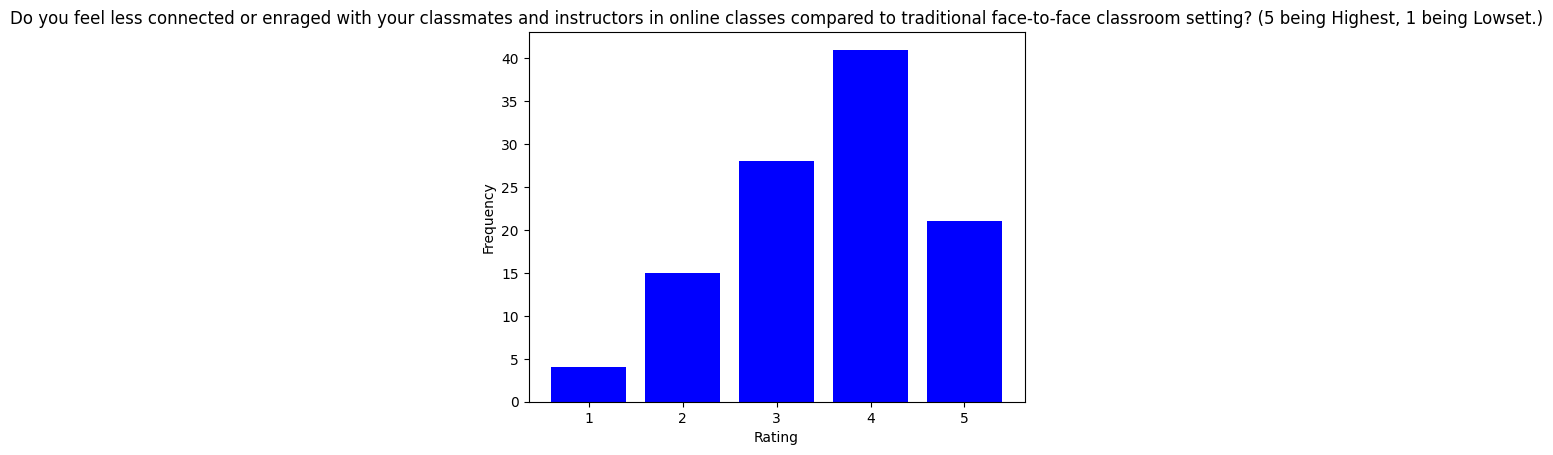
* Present the final visualizations with clear captions, labels, and relevant references.

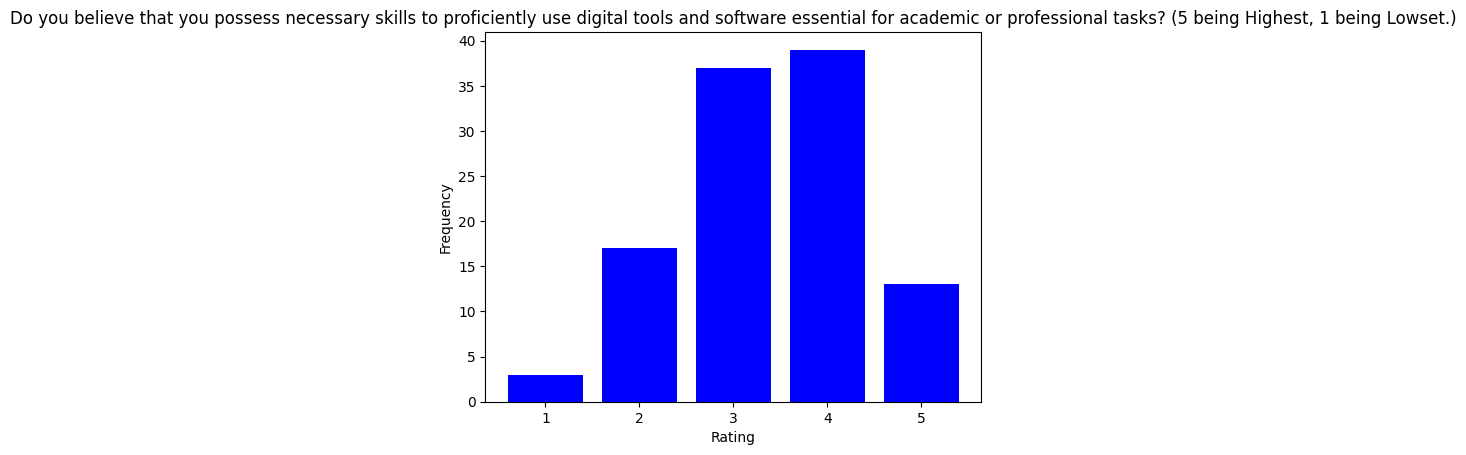
**Most Difficult Subject**

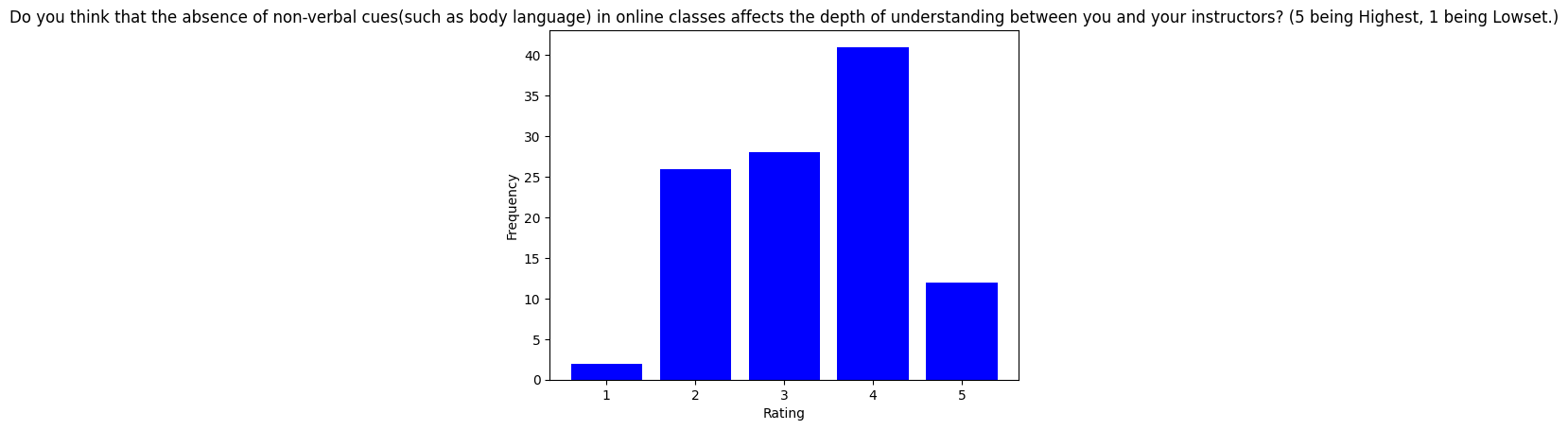
**Scenarios that affect the ability to perform your best during this academic term.**

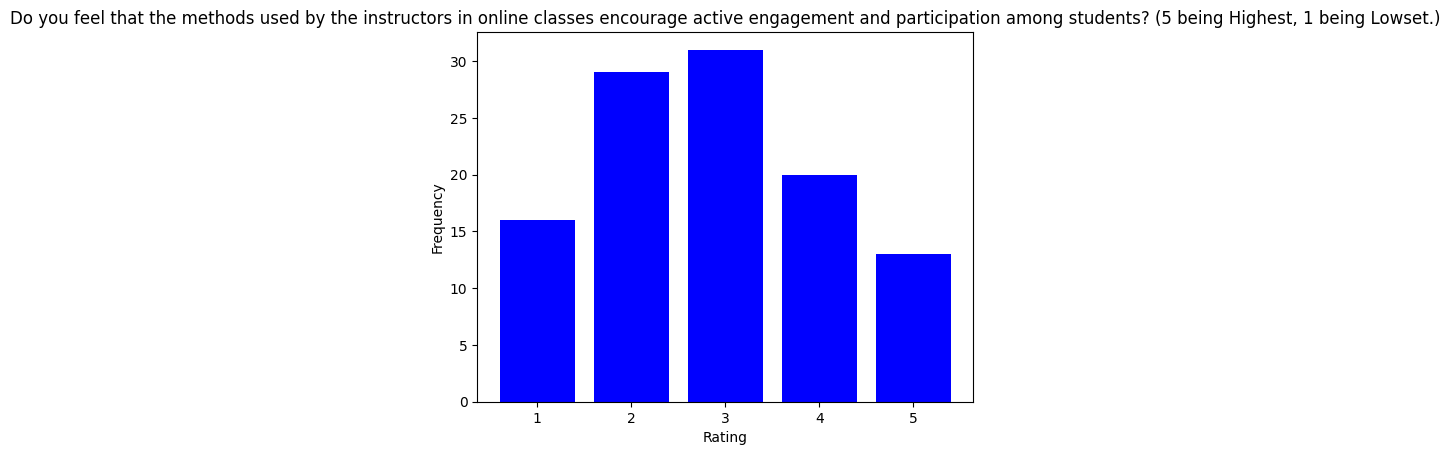
 

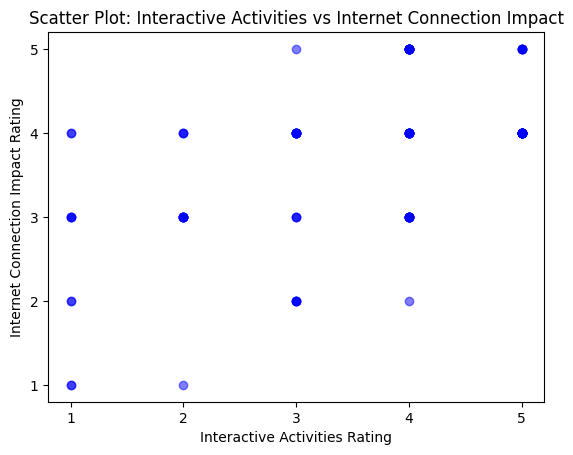












* Describe what you see in the visualizations and what insights you can draw.

- Based on the answers gathered from the CCS Students. Most of the students had a hard time with the subject ITEC 103 – Intermediate programming. The effect of hybrid-classes during the academic year 2022-2023 (2nd Semester) was indeed challenging for most of the students. Not all students have the same environment when studying, especially when the classes are being held online, some students are distracted more than usual or unable to attend classes altogether.

* Discuss how the visualizations answer the initial project questions.

1. Helpfulness of Interactive Activities:

* The histogram showcasing ratings given by students depicts how they perceive the usefulness of interactive activities during online classes. Higher bars toward the higher ratings (5 being the highest) would suggest a positive reception to these activities.

1. Impact of Internet Connection:

* This histogram visualizes how students perceive the impact of internet connectivity on their participation in online classes. Higher frequencies at lower ratings could indicate that a substantial portion of students face issues due to poor internet, affecting their participation.

1. Effect of Background Noises:

* Similar to the previous histograms, this one represents the distraction caused by background noises during online classes. Higher frequencies at higher ratings would indicate more significant distraction due to noise.

1. Disruption Due to Power Interruptions:

* This histogram likely shows how power interruptions affect students' ability to study effectively at home. Higher frequencies at higher ratings could suggest a substantial impact on studying during power disruptions.

1. Encountering Technical Difficulties:

* Students' experiences with technical issues during online assessments are showcased here. Higher frequencies at higher ratings might imply that technical difficulties are fairly common during assessments.

1. Physical Discomfort from Prolonged Screen Time:

* This histogram indicates whether students have noticed an increase in headaches or physical discomfort due to prolonged screen time while studying online. Higher frequencies at higher ratings might suggest a significant impact.

1. Sense of Connection in Online Classes:

* This visualization likely reflects students' perceptions of feeling less connected or engaged with peers and instructors in online classes compared to traditional face-to-face settings. Higher frequencies at higher ratings might indicate a notable disconnect.

1. Perceived Proficiency with Digital Tools:

* This histogram could represent students' beliefs about their proficiency in using digital tools for academic or professional tasks. Higher frequencies at higher ratings would imply confidence in their skills.

1. Impact of Non-Verbal Cues on Understanding:

* Here, students' opinions on the absence of non-verbal cues in online classes impacting their depth of understanding with instructors are likely shown. Higher frequencies at higher ratings might suggest a perceived impact.

1. Instructor's Methods for Active Engagement:

* This visualization likely represents students' perceptions of how well instructors' methods encourage active engagement and participation. Higher frequencies at higher ratings could suggest positive encouragement.
* Address any limitations or potential biases in the data or visualization.
* Majority of the answers submitted has a value of 4(Highly affected) by the scenarios listed. It would mean that most students are distracted and could be potentially disturbed by their environment at home and at school or getting used to the new lifestyle they have coming from the Full online classes only.

**6. Conclusion**

* Summarize the key findings and takeaways from the data visualization.

1. Subject Preference:

* The bar chart displays the frequency of different subjects chosen by students, giving an idea of the subjects that were most popular or least preferred among students.

1. Impact of Scenarios on Performance:

* Histograms for different scenarios (like internet connection issues, background noises, power interruptions, etc.) portray how students rate these factors. Higher ratings indicate a more significant negative impact on their performance.

1. Interactive Activities and Internet Connection:

* The scatter plot attempts to find a correlation between how students rate interactive activities during online classes and the impact of internet connection on their participation. It suggests if there's any relationship between finding activities helpful and being affected by connectivity issues.

1. Engagement with Online Learning:

* Some scenarios reflect students' engagement and interaction levels during online classes compared to traditional face-to-face settings, shedding light on potential challenges or differences perceived by students.

1. Technical Proficiency and Teaching Methods:

* Ratings on possessing necessary digital skills and how teaching methods encourage active engagement could give insights into the effectiveness of the current digital infrastructure and pedagogical strategies.

The visualizations provide a glimpse into various aspects of students' experiences and challenges during online learning. However, deeper analysis and correlation studies between these factors might reveal more nuanced insights, such as understanding which challenges correlate most strongly with lower performance or identifying specific subjects affected by these challenges.

The findings might help educational institutions tailor their online learning environments better, address key pain points, and improve student experiences in virtual classrooms.

* Suggest potential future improvements or extensions to the project.
* Use more Data visualization and create a wider set of questions for more accurate presentation of data.

**7. Appendix**

* Include supporting materials such as:
  + Data dictionary

<https://drive.google.com/drive/folders/1J12iAzRwuWfcqEYq1znbP58aGqvr9Eaw?usp=sharing>

* + Code for data cleaning and preprocessing

loading the data (pd.read\_csv()),

checking for missing values (df.isnull().sum()),

handling missing values(df.astype() or

pd.to\_datetime()).converting data types if neededs