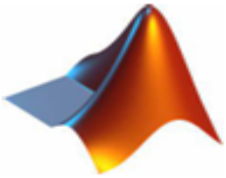
Part B

Descriptive Analysis of the Data

Group 9

Note: Since the Dining Hall is closed in the evenings, we decided to take data from the business building cafeteria, the Forum, instead.

Question 1



It is recommended that you use Matlab to complete Part B of the assignment.

To carry out the assignment, please import the collected data in Matlab. To read the .csv file, use the `readmatrix` function.

Utilising the collected data, create datasets using averaged data samples for each timeslot. This step will involve aggregating data points across all days for each specific timeslot, across 3 locations and 3 different times of the day, thus creating a total of 9 datasets.

I. Plot the datasets and choose the location that you group presumes to be the noisiest. Each plot should have axes titles, units and legends. You may use the MATLAB function `subplot` for plotting the data.

[25 Marks]

Answer:

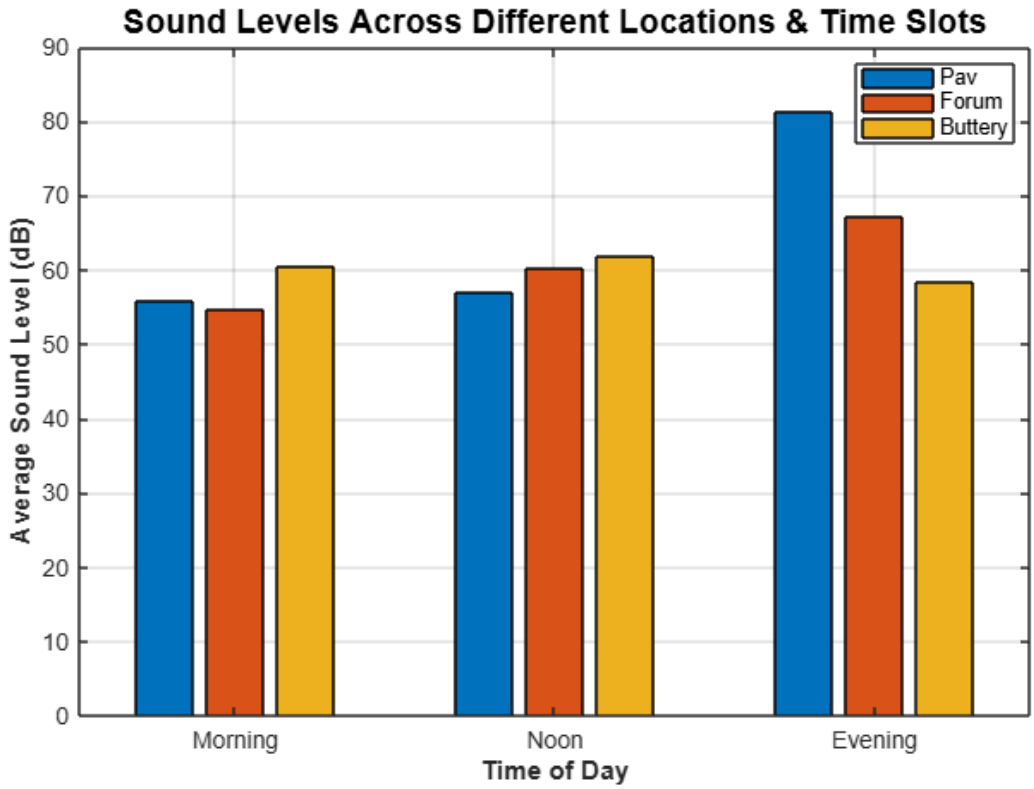
Before plotting the data, our group presumed that the Pav would be the noisiest location as students go there in the evenings to drink, chat, and listen to music.

Below are the plots of the 9 datasets. They are plotted as a bar chart and a box plot in order to summarise the data in sufficient detail.

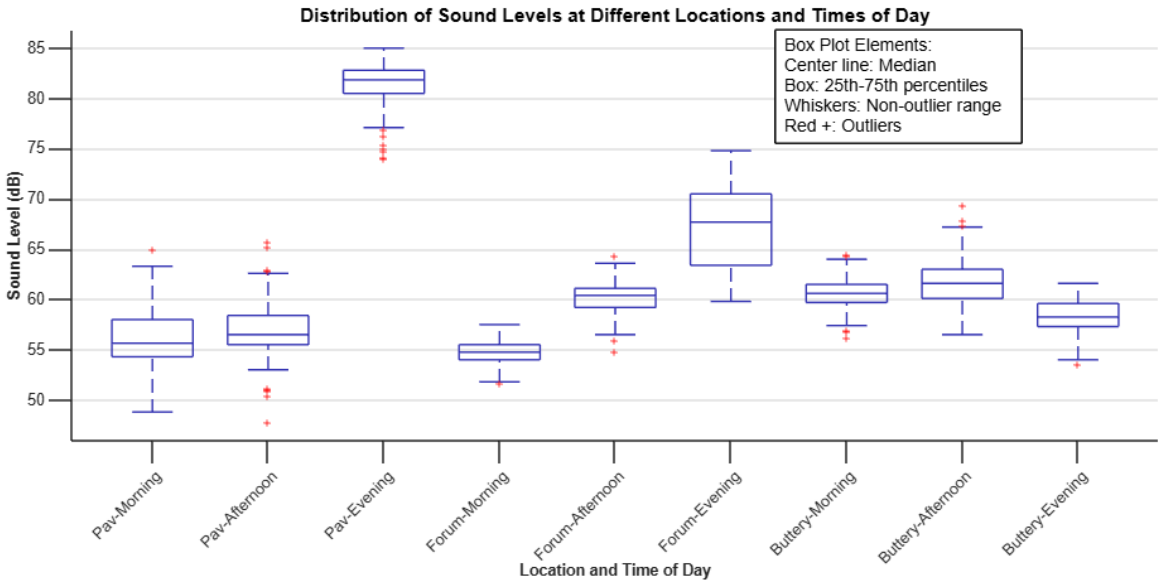
The MATLAB code used to produce these plots is:

```
clear;
clc;
close all;

filename = 'Sound Data.csv';
data = readtable(filename, 'VariableNamingRule', 'preserve');
avgCol = data.Properties.VariableNames{contains(data.Properties.VariableNames, 'SoundLevel', 'IgnoreCase', true)};
data.(avgCol) = str2double(string(data.(avgCol)));

locations = unique(data.Location, 'stable');
timeSlots = unique(data.TimeSlot, 'stable');
figure('Position', [100, 100, 1200, 800]);
soundLevels = zeros(length(timeSlots), length(locations));

for i = 1:length(locations)
    for j = 1:length(timeSlots)
        locMask = strcmp(data.Location, locations{i});
        timeMask = strcmp(data.TimeSlot, timeSlots{j});
        selectedData = data.(avgCol)(locMask & timeMask);
        soundLevels(j, i) = mean(selectedData, 'omitnan');
    end
end

bar(soundLevels);
xticks(1:length(timeSlots));
xticklabels(timeSlots);
xlabel('Time of Day', 'FontWeight', 'bold');
ylabel('Average Sound Level (dB)', 'FontWeight', 'bold');
legend(locations, 'Location', 'best');
title('Sound Levels Across Different Locations & Time Slots', 'FontSize', 14, 'FontWeight', 'bold');
grid on;
```

The bar chart obtained on plotting the sound levels for different locations:



The box plot obtained on plotting the sound levels for different locations is:



The bar chart gives a quick and clear comparison of the average sound levels for different locations at different times of day, while the box plot provides more detail by describing the median, interquartile range, outliers etc.

Question 2

II. Use the MATLAB function `histogram` to plot the histogram of each dataset. Describe the distribution of each dataset discussing shape, centre, spread and outliers.

[18 Marks]

Answer:

Below is the MATLAB code for plotting the 9 datasets in different figures:

```
clear all
clc
data = readmatrix('Sound Data.csv');
titles = {'Pav-Morning', 'Pav-Afternoon', 'Pav-Evening','Forum-Morning', 'Forum-Afternoon', 'Forum-Evening','Buttery-Morning', 'Buttery-Afternoon', 'Buttery-Evening'};

for i = 1:9
    figure('Position', [100 100 600 400]);
    ax = axes;
    ax.YLim = [0 150];
    hold on;
    h = histogram(data(:,i), 'BinWidth', 1);
    h.FaceColor = [0.9290, 0.6940, 0.1250];

    title(['Figure ', num2str(i), ': Histogram of Noise Levels - ', titles{i}], 'FontSize', 12);
    xlabel('Noise Level (dB)', 'FontSize', 11);
    ylabel('Number of Recordings', 'FontSize', 11);
    if i == 3
        xlim([75 85]);
    else
        xlim([45 75]);
    end

    ylim([0 60]);
    yticks(0:10:60);
    grid on;

    set(gcf, 'Color', 'white');
    set(gca, 'FontSize', 10);
    set(gca, 'YMinorGrid', 'on');
    fprintf('\nDataset %d: %s\n', i, titles{i});
    fprintf('Number of observations: %d\n', length(data(:,i)));
    fprintf('Min value: %.2f\n', min(data(:,i)));
    fprintf('Max value: %.2f\n', max(data(:,i)));

    hold off;
end
```

These are the histograms obtained by running the above code:

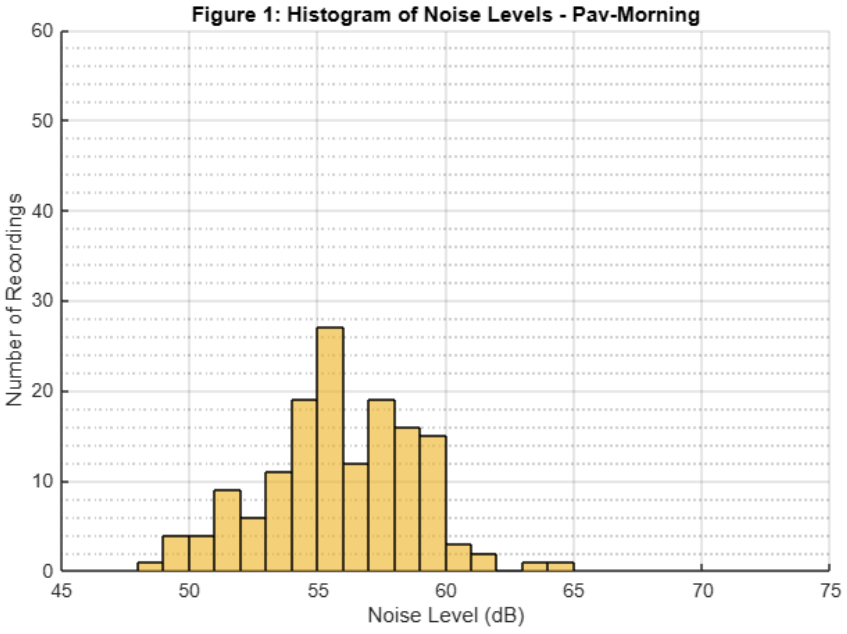


Figure 2: Histogram of Noise Levels - Pav-Afternoon

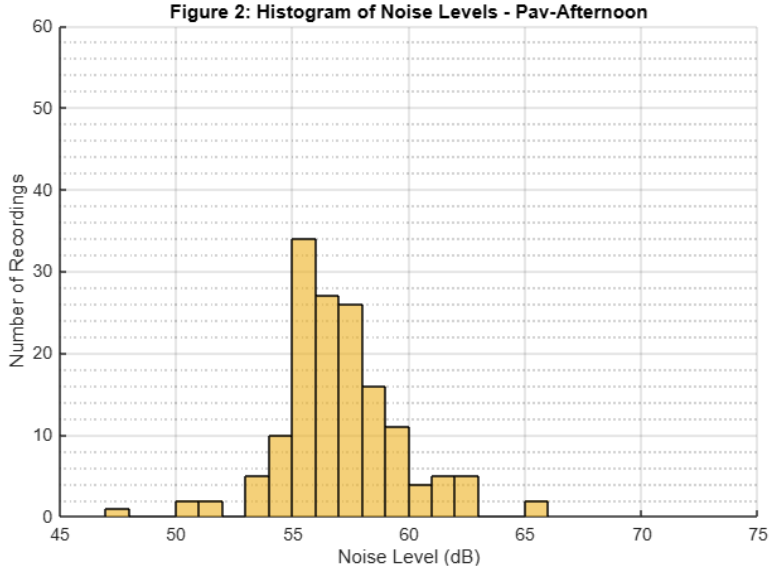


Figure 3: Histogram of Noise Levels - Pav-Evening

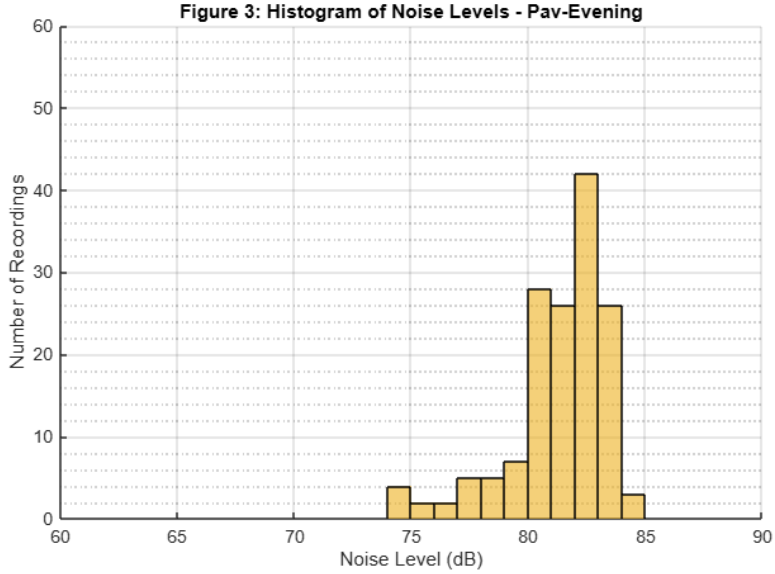


Figure 4: Histogram of Noise Levels - Forum-Morning

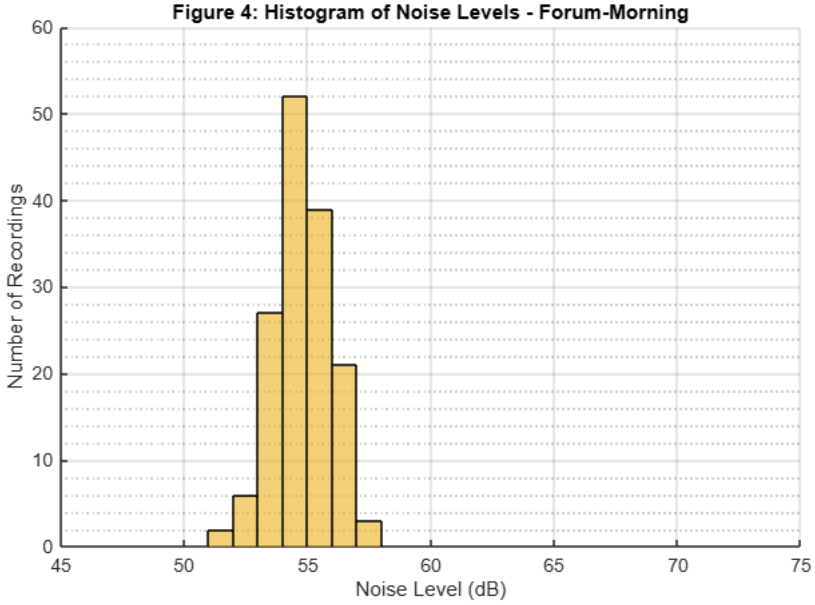


Figure 5: Histogram of Noise Levels - Forum-Afternoon

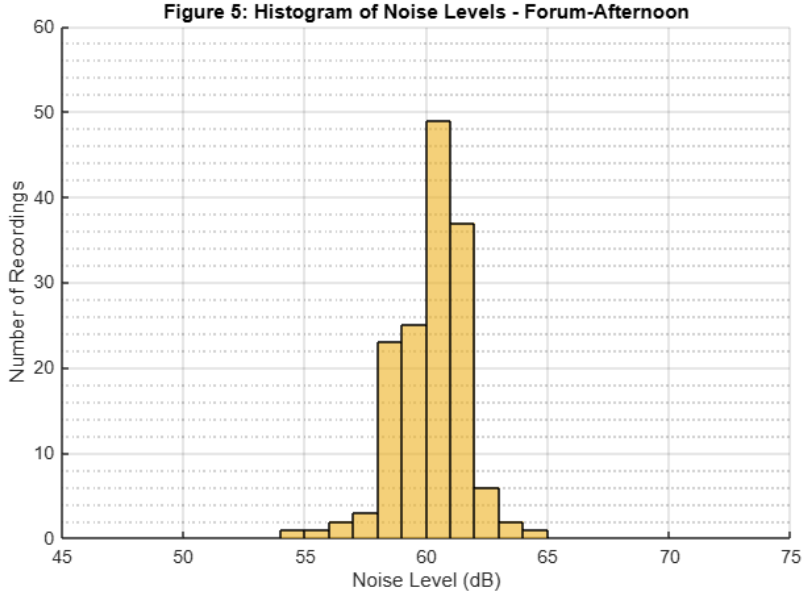


Figure 6: Histogram of Noise Levels - Forum-Evening

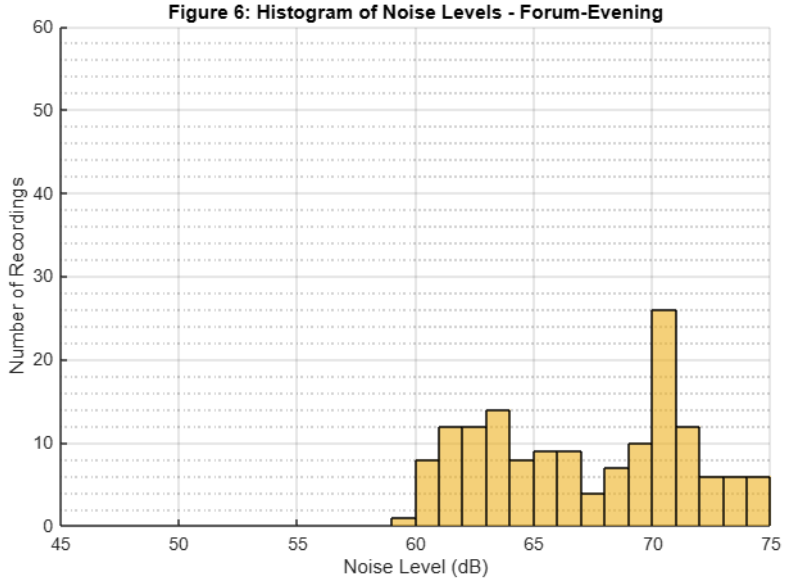
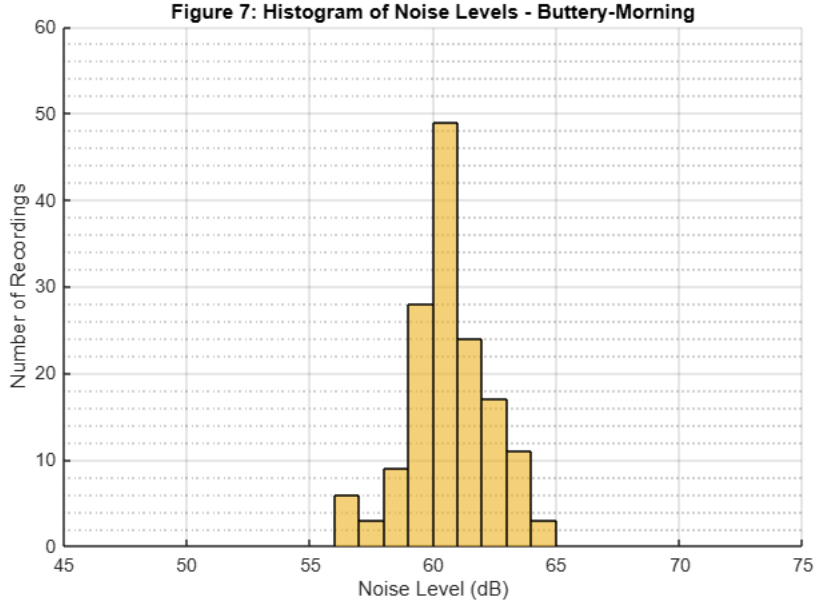
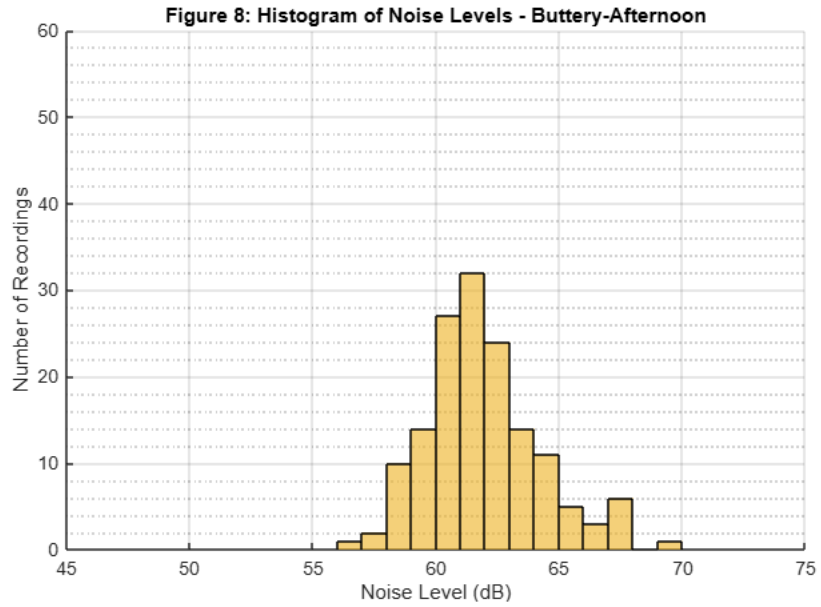
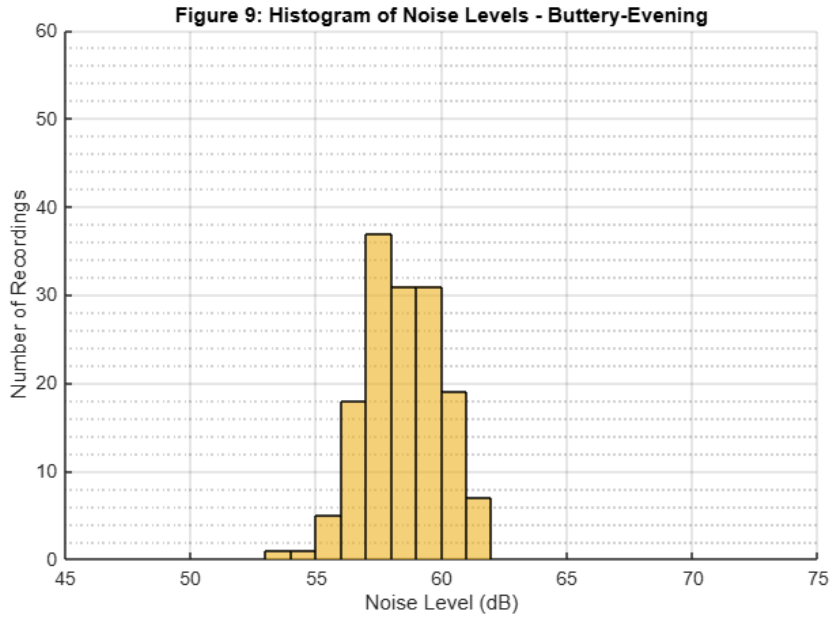


Figure 7: Histogram of Noise Levels - Buttery-Morning







About the histograms :

Pavilion (Pav) (Fig 1-3)

1. Morning (Fig. 1)
    - Shape: Slightly right-skewed.
    - Center: Around 55-57 dB.
    - Spread: Ranges from about 48 dB to 65 dB.
    - Outliers: Few values beyond 65 dB.

2. Afternoon (Fig. 2)
    - Shape: Slightly right-skewed (more concentrated than the morning dataset).
    - Center: Around 56 dB.
    - Spread: Narrower range, between 50 dB and 65 dB.
    - Outliers: Minimal, mostly within expected range.
3. Evening (Fig. 3)
    - Shape: Right-skewed.
    - Center: Around 81-82 dB.
    - Spread: Ranges from 75 dB to 85 dB.
    - Outliers: No significant outliers, but noise levels are consistently high.

Forum (Fig 4-6)

1. Morning (Fig. 4)
    - Shape: Symmetric.
    - Center: Close to 55 dB.
    - Spread: From 50 dB to 60 dB.
    - Outliers: None apparent.
2. Afternoon (Fig. 5)
    - Shape: Slightly right-skewed.
    - Center: Around 60 dB.
    - Spread: Between 55 dB and 65 dB.
    - Outliers: No extreme outliers.
3. Evening (Fig. 6)
    - Shape: Two peaks around 65-70 dB.
    - Center: Approximately 66-70 dB.
    - Spread: From 60 dB to 75 dB.
    - Outliers: Some higher values above 70 dB.

Buttery Location (Fig 7-9)

1. Morning (Fig. 7)
    - Shape: Slight right skew.
    - Center: Around 60 dB.
    - Spread: Between 55 dB and 65 dB.
    - Outliers: None.
2. Afternoon (Fig. 8)
    - Shape: Nearly symmetric.
    - Center: Around 62 dB.
    - Spread: From 55 dB to 70 dB.
    - Outliers: None.
3. Evening (Fig. 9)
    - Shape: Symmetric.
    - Center: Around 58-60 dB.
    - Spread: Between 55 dB and 65 dB.
    - Outliers: None.

Overall Description of locations:

- Pavilion: Evening noise is significantly higher (~80 dB), while morning and afternoon noise remain moderate.
- Forum: Afternoon and evening noise levels rise compared to the morning. The evening has a wider spread.
- Buttery: Fairly consistent noise levels, with the afternoon (lunch time) being slightly louder than the morning and evening.

The max and min values for each dataset are tabulated below:

```
Dataset 1: Pav-Morning
Number of observations: 150
Min value: 48.80
Max value: 65.00

Dataset 2: Pav-Afternoon
Number of observations: 150
Min value: 47.80
Max value: 65.70

Dataset 3: Pav-Evening
Number of observations: 150
Min value: 74.00
Max value: 85.00

Dataset 4: Forum-Morning
Number of observations: 150
Min value: 51.60
Max value: 57.50

Dataset 5: Forum-Afternoon
Number of observations: 150
Min value: 54.80
Max value: 64.30

Dataset 6: Forum-Evening
Number of observations: 150
Min value: 59.80
Max value: 74.80

Dataset 7: Buttery-Morning
Number of observations: 150
Min value: 56.20
Max value: 64.50

Dataset 8: Buttery-Afternoon
Number of observations: 150
Min value: 56.50
Max value: 69.30

Dataset 9: Buttery-Evening
Number of observations: 150
Min value: 53.50
Max value: 61.60
```

Question 3

III. For each of the 9 datasets, calculate the measures of central tendencies of your choice and report them in a tabulated form.

[27 Marks]

Answer:

The central tendencies chosen are:

1. Mean: The sum of all values divided by the number of values

   Formula:

$$\text{Mean}(\bar{x}) = \frac{\sum x_i}{n}$$

where x_i are the data points and n is the number of data points.

2. Median: The middle value when the data is arranged in ascending order

   Formula:

- If n is odd:

$$\text{Median} = \text{Middle value}$$

- If n is even:

$$\text{Median} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

where $x_{\frac{n}{2}}$ and $x_{\frac{n}{2}+1}$ are the middle two values.

3. Standard Deviation (StdDev): The dispersion of data points from the mean

   Formula:

- Population standard deviation:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

- Sample standard deviation:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

where μ is the population mean, $\bar{x}$ is the sample mean, and N or $n-1$ is the divisor depending on population or sample data.

4. Range: The difference between the maximum and minimum values in the dataset

   Formula:

$$\text{Range} = \text{Max} - \text{Min}$$

5. First Quartile (Q1): The median of the lower half of the data (25th percentile)
6. Third Quartile (Q3): The median of the upper half of the data (75th percentile)
7. Interquartile Range (IQR): The spread of the middle 50% of the data

Formula:

$$\text{IQR} = Q3 - Q1$$

Following is the code for calculating the listed central tendencies and presenting them in a tabulated form in MATLAB.

```
/MATLAB Drive/untitled.m
1   clear all
2   clc
3
4   filename = 'Sound Data.csv';
5   data = readmatrix(filename);
6
7   columnNames = {'Pav-Morning', 'Pav-Afternoon', 'Pav-Evening', 'Forum-Morning','Forum-Afternoon', 'Forum-Evening', 'Buttery-Morning', 'Buttery-Afternoon', 'Buttery-Evening'};
8
9   means = mean(data);
10  medians = median(data);
11  stds = std(data);
12  ranges = max(data) - min(data);
13  quartile1 = quantile(data, 0.25);
14  quartile3 = quantile(data, 0.75);
15  IQR = quartile3 - quartile1;
16
17  statsTable = table(means', medians', stds', ranges', quartile1', quartile3', IQR', 'RowNames', columnNames, ...
18      'VariableNames', {'Mean', 'Median', 'StdDev', 'Range', 'Q1', 'Q3', 'IQR'});
19
20  disp('Statistical Analysis of Datasets:')
21  disp(statsTable)
22
```

The output obtained by running the above code:

```
Statistical Analysis of Datasets:
                       Mean     Median    StdDev    Range     Q1      Q3      IQR
                       _____    ______    ______    _____     ____    ____    ___

    Pav-Morning        55.819    55.65    2.9676    16.2      54.3     58     3.7
    Pav-Afternoon      56.953     56.5    2.6233    17.9      55.5    58.4    2.9
    Pav-Evening        81.273    81.85    2.1325      11      80.5    82.8    2.3
    Forum-Morning      54.762    54.75    1.1362     5.9       54     55.5    1.5
    Forum-Afternoon    60.176     60.4    1.4741     9.5      59.2    61.1    1.9
    Forum-Evening      67.134     67.7    4.1935      15      63.4    70.5    7.1
    Buttery-Morning    60.601     60.6    1.6749     8.3      59.7    61.5    1.8
    Buttery-Afternoon  61.831     61.6    2.3398    12.8      60.1      63    2.9
    Buttery-Evening    58.369    58.25    1.5459     8.1      57.3    59.6    2.3
```

Question 4

IV. Calculate the 5% and 20% trimmed mean for each dataset and comment on your original choice of measures of central tendencies in adequately describing each dataset.

[18 Marks]

Answer:

Following is the code for calculating the 5% and 20% trimmed mean for each of the 9 datasets:

```
clear all
clc
filename = 'Sound Data.csv';
data = readmatrix(filename);

regular_mean = mean(data);
trimmed_mean_5 = trimmean(data, 5);
trimmed_mean_20 = trimmean(data, 20);
medians = median(data);

locations = {'Pav-Morning', 'Pav-Afternoon', 'Pav-Evening', 'Forum-Morning','Forum-Afternoon', 'Forum-Evening', 'Buttery-Morning', 'Buttery-Afternoon', 'Buttery-Evening'};
results = table(regular_mean', trimmed_mean_5', trimmed_mean_20', medians', 'RowNames', locations,'VariableNames', {'Mean', 'Trimmed_5', 'Trimmed_20', 'Median'});

disp('Comparison of Different Measures of Central Tendency:')
disp(results)

diff_5 = abs(regular_mean - trimmed_mean_5);
diff_20 = abs(regular_mean - trimmed_mean_20);

fprintf('\nLargest differences between regular mean and 5%% trimmed mean:\n')
[sorted_diff_5, idx_5] = sort(diff_5, 'descend');
for i = 1:3
    fprintf('%s: %.3f\n', locations{idx_5(i)}, sorted_diff_5(i));
end

fprintf('\nLargest differences between regular mean and 20%% trimmed mean:\n')
[sorted_diff_20, idx_20] = sort(diff_20, 'descend');
for i = 1:3
    fprintf('%s: %.3f\n', locations{idx_20(i)}, sorted_diff_20(i));
end
```

The output obtained by running the above code:

```
Comparison of Different Measures of Central Tendency:
                       Mean     Trimmed_5    Trimmed_20    Median
                       _____    _________    __________    ______

    Pav-Morning        55.819     55.811       55.897      55.65
    Pav-Afternoon      56.953     56.945        56.84       56.5
    Pav-Evening        81.273     81.383       81.577      81.85
    Forum-Morning      54.762      54.77       54.783      54.75
    Forum-Afternoon    60.176     60.201        60.24       60.4
    Forum-Evening      67.134     67.128       67.137       67.7
    Buttery-Morning    60.601     60.612       60.613       60.6
    Buttery-Afternoon  61.831     61.775       61.667       61.6
    Buttery-Evening    58.369     58.389       58.389      58.25


Largest differences between regular mean and 5% trimmed mean:
Pav-Evening: 0.110
Buttery-Afternoon: 0.055
Forum-Afternoon: 0.025

Largest differences between regular mean and 20% trimmed mean:
Pav-Evening: 0.304
Buttery-Afternoon: 0.164
Pav-Afternoon: 0.113
>>
```

The table above shows that in all cases the simple mean is likely sufficient to describe the data. Even in the dataset in which the mean was most skewed by outliers (the Pav in the evening), the 20% trimmed mean was only 0.3 dB higher than the simple mean. Given that the mean is around 80 dB, and even taking account of the non-linearity of the dB scale, this is not a big change. In all other cases, the trimmed mean showed even less variation about the simple mean. This implies that even though outliers are present, their effect is not significant, and the simple mean is not dramatically skewed by their inclusion in the dataset, and as such the simple mean is sufficiently capable of describing the datasets.

Question 5

V. Based on the calculations carried out in II-IV, conclude which place is noisier and why? Is your conclusion same as your answer in I?

[12 Marks]

Answer:

Base on the calculation carried out in II-IV, it can be concluded that the Pavilion (Pav) is the noisiest location, especially in the evening. At this time, the Pav had an average sound level of 81dB. While the noise levels at the Pav are broadly in line with those in the other locations during the day, the Pav is significantly louder in the evening than any other location at any time of day. This is not surprising, in our view, and aligns with our prediction in part i. We suggest three possible causes that combine to make the evenings in the Pav the loudest dataset. Firstly, students meet in the Pav to relax after class and enjoy music and some drinks, which leads to a louder environment. Secondly, all students are generally available in the evenings and can go to the Pav at the same time, whereas students normally have different lecture timetables and cannot all eat lunch together at the same time. Finally, clubs and societies often host events (pub quizzes, social events etc) which lead to large numbers of students socialising and making noise. In conclusion, the data confirms our original suspicion that the Pav in the evening is significantly louder than the other two locations at any time of day.

Group 9 members and their roles:

Samuel Gammell: Data Collection – Data gathering

Nibu Rajan: Data Collection – Data gathering

Utkarsh Tiwari: Data Collection – Data gathering

Oscar Murphy: Data Trimming – Data formatting in Excel

Navya Mittal: Descriptive Analysis of Trimmed Data – Matlab code and data analysis