



National University

of computer and emerging sciences

Assignment 1

Group Members:

Zoya Sumbul 23I-8056

Amina Ashfaq 23I-8027

Table of Contents

Detailed Report of EDA	3
Introduction	3
Data Ingestion & Storage	3
Data Reading	3
Data Cleaning	3
Attendance data	3
Evaluation data	4
Data Storing	5
Advanced Exploratory Data Analysis (EDA) on Transformed Data	5
Step 1: Data Understanding & Summary Statistics	5
Loading the data	5
Basic Summary statistics	6
Visualize Summary Statistics	6
Step 2: Meaningful Query-Based EDA	7
Query 1: Correlation between Sessional Marks and Final Exam Scores	7
Query 2: Impact of Attendance on Final Exam Performance	8
Query 3: Correlation Heatmap	9
Query 4: Students with High Sessional Marks but Low Final Marks	9
Query 5: Section wise Final and Sessionals Marks Trend Analysis	10
Query 6: Distribution of Attendance	11
Query 7: Average Marks by Attendance Category	12
Query 8: Students with Perfect Attendance by each Section	13
Query 9: Section-wise Attendance Distribution	14
Query 10: Section-wise Final Marks Distribution	16
Query 11: Predictive Model for Final Marks	16
Query 12: Distribution of Final Marks Based on Quiz and Assignment Performance	18
Query 13: Predictive Power of Sessional and Attendance on Final Marks	19
Query 14: Students with Consistent Performance across All Components	20
Query 15: Impact of Quiz and Assignment on Final Marks	21
Query 16: Predictive Model for Final Marks	22
Conclusion	23

Detailed Report of EDA

Introduction

Exploratory Data Analysis (EDA) is an important first step in data science projects. It involves looking at and visualizing data to understand its main features, find patterns, and discover how different parts of the data are connected. EDA helps to spot any unusual data or outliers and is usually done before starting more detailed statistical analysis or building models. Dataset used for performing EDA consist of attendance and marks of different evaluation of multiple students belong to multiple sections.

Data Ingestion & Storage

Data Reading

The data was read and processed by a custom function named `read_excel`, designed to deal with file path complications and to read the necessary data into pandas DataFrames for the sake of analysis. In this setup, attendance information and evaluation data were kept separately for each individual section, thereby making it easy to find and access for the purpose of reference.

```
def read_excel_file(filename):
    file_path = os.path.join(base_dir, filename)
    if os.path.exists(file_path):
        return pd.read_excel(file_path)
    else:
        print(f"File not found: {file_path}")
        return None

# Reading Attendance Data
attendance_bse2a = read_excel_file("Attendance Register_BSE-2A.xlsx")
attendance_bse2b = read_excel_file("Attendance Register_BSE-2B.xlsx")
attendance_bse2c = read_excel_file("SE-C.xlsx")
attendance_bcs2a = read_excel_file("Attendance Register_BCS-2A.xlsx")
attendance_bcs2b = read_excel_file("Attendance Register_BCS-2B.xlsx")
attendance_bcs2c = read_excel_file("Attendance Register_BCS-2C.xlsx")
attendance_bcs2d = read_excel_file("Attendance Register_BCS-2D.xlsx")
```

Data Cleaning

Attendance data

We have cleaned attendance data by Skipped first 3 header rows. They have chosen only the columns which are pertinent, i.e., the Roll Number and the Attendance Data. The roll number has been chosen to be employed as the index column of the data set. To uniquely identify each roll numbers of a section, roll number is cast into string and appended with a suffix which is essentially section's name

	Roll_Num	Attendance
3	1_bcs2z	91
4	2_bcs2z	81
5	3_bcs2z	88
6	4_bcs2z	91
7	5_bcs2z	91

Evaluation data

In evaluation data, we have Combine sessional marks for different weight categories like: sections with sessional out of 25 are combined, sections with sessional out of 30 are combined, and both sessional groups are standardized to percentages out of 100 and converts raw scores (out of 40) to percentages. All sections have 15 marks for assignments, except: bcs2z, bse2a, bse2b, bse2c, bcy2c, bcy2d have 10 marks for assignments and standardize assignment marks to a common scale. All sections have 15 marks for Quizzes, except: bcs2z, bse2a, bse2b, bse2c, bcy2c, bcy2d have 10 marks for assignments and standardize quizzes marks to a common scale.

	Roll_Num	Attendance	Sessional	Final	Grade	Assignment_Standardized	Quiz_Standardized
0	2_bcs2a	95	17.200000	45.000	F	51.333333	68.733333
1	3_bcs2a	85	15.566667	23.100	F	24.200000	55.333333
2	4_bcs2a	95	38.933333	54.575	C-	38.400000	82.466667
3	5_bcs2a	92	25.000000	35.125	F	23.866667	73.733333
4	6_bcs2a	92	46.633333	71.875	B-	57.466667	76.933333
...
519	27_bse2c	97	75.320000	66.650	A-	79.600000	100.000000
520	28_bse2c	100	60.000000	52.500	C+	77.600000	85.000000
521	29_bse2c	97	70.640000	72.200	A-	86.700000	93.000000
522	30_bse2c	100	46.120000	38.875	D	54.500000	85.800000
523	31_bse2c	97	63.320000	49.425	C	69.400000	86.000000

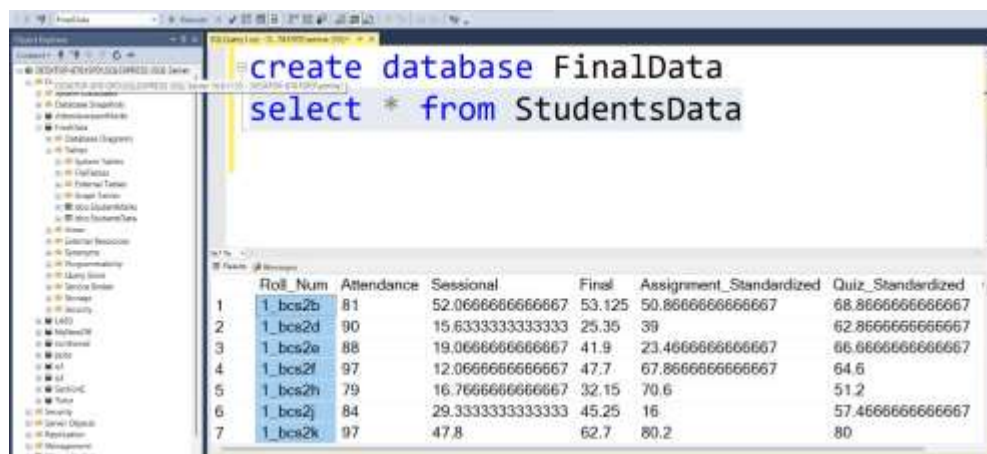
524 rows × 7 columns

Data Storing

After data reading and cleaning step we have store our data to SQL Server for efficient querying and analysis. A table named 'StudentsData' was created and data was inserted into it.

```
create_table_query = """
CREATE TABLE StudentsData (
    Roll_Num NVARCHAR(50) PRIMARY KEY,
    Attendance FLOAT,
    Sessional FLOAT,
    Final FLOAT,
    Assignment_Standardized FLOAT,
    Quiz_Standardized FLOAT,
    Grade NVARCHAR(10)
);
"""
```

```
for index, row in final_data.iterrows():
    insert_query = """
    INSERT INTO StudentsData (
        Roll_Num, Attendance, Sessional, Final, Assignment_Standardized, Quiz_Standardized, Grade
    ) VALUES (?, ?, ?, ?, ?, ?, ?);
    """
    cursor.execute(insert_query,
                    row["Roll_Num"],
                    row["Attendance"],
                    row["Sessional"],
                    row["Final"],
                    row["Assignment_Standardized"],
                    row["Quiz_Standardized"],
                    row["Grade"])
```



The screenshot shows the SQL Server Enterprise Manager interface. The 'Query Editor' window displays two queries: 'create database FinalData' and 'select * from StudentsData'. Below the queries, the 'Results' pane shows a table with 7 rows and 7 columns: Roll_Num, Attendance, Sessional, Final, Assignment_Standardized, Quiz_Standardized, and Grade. The data is as follows:

	Roll_Num	Attendance	Sessional	Final	Assignment_Standardized	Quiz_Standardized
1	1_bcs2b	81	52.0666666666667	53.125	50.8666666666667	68.8666666666667
2	1_bcs2d	90	15.6333333333333	25.35	39	62.8666666666667
3	1_bcs2e	88	19.0666666666667	41.9	23.4666666666667	66.6666666666667
4	1_bcs2f	97	12.0666666666667	47.7	67.8666666666667	64.6
5	1_bcs2h	79	16.7666666666667	32.15	70.6	51.2
6	1_bcs2j	84	29.3333333333333	45.25	16	57.4666666666667
7	1_bcs2k	97	47.8	62.7	80.2	80

Advanced Exploratory Data Analysis (EDA) on Transformed Data

Step 1: Data Understanding & Summary Statistics

Loading the data

Load Data directly from SQL Server or we can use it from saved csv or dataframe. The dataset shape, column types, missing values, and basic statistics were evaluated.

Basic Summary statistics

Summary statistics such as mean, median, mode, min, max, and standard deviation were calculated.

	Attendance	Sessional	Final	Assignment_Standardized	Quiz_Standardized
count	524.000000	524.000000	524.000000	524.000000	524.000000
mean	92.622137	46.456069	57.464552	59.086069	73.906107
std	5.991578	16.586572	13.907931	20.996975	15.437577
min	64.000000	1.700000	0.000000	3.066667	22.900000
25%	88.000000	34.291667	49.250000	45.566667	65.766667
50%	94.000000	46.716667	57.837500	61.033333	76.633333
75%	97.000000	58.416667	66.525000	74.350000	85.000000
max	100.000000	86.600000	95.400000	99.600000	100.000000

Mode Values:

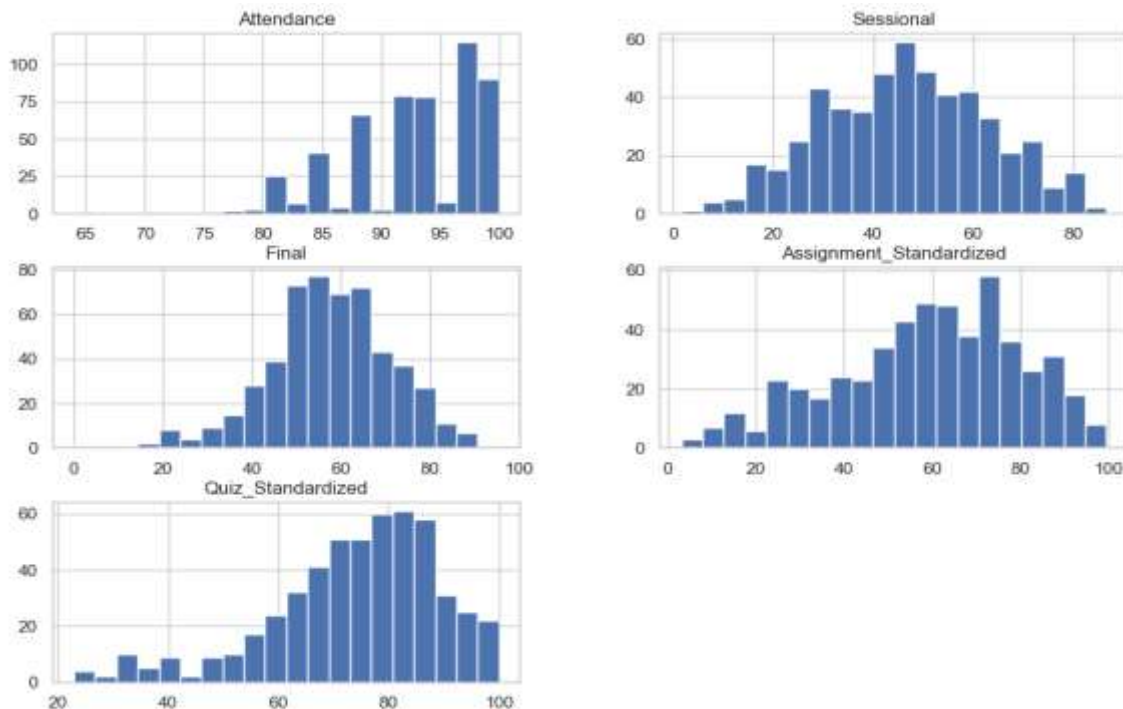
```
Roll_Num          10_bcs2a
Attendance         97.0
Sessional          44.0
Final              45.8
Grade              F
Assignment_Standardized  28.733333
Quiz_Standardized   85.0
Name: 0, dtype: object
```

Visualize Summary Statistics

From this visualization of the distribution of numeric columns, we can derive the following insights:

- Attendance
 - Most students have high attendance, clustering around 90-100%.
 - A few students have significantly lower attendance, but overall, attendance is skewed towards higher values.
 - There may be an attendance policy influencing this trend.
- Sessional Marks
 - The sessional scores follow a nearly normal distribution.
 - Most students scored around the 40-70 range.
 - Some students performed exceptionally well, but there are also students at the lower end.
- Final Marks
 - The final exam scores also follow a bell-shaped curve, indicating normal distribution.
 - The peak (mode) of the distribution is around 50-70.
 - Some students performed poorly, but the overall spread is fairly even.
- Assignment Scores

- The distribution is slightly right-skewed, meaning some students performed very well.
- The majority of students scored between 40-80.
- A few students scored either very low or very high.
- Quiz Scores
 - The distribution is left-skewed, meaning many students scored higher on quizzes.
 - The majority scored between 60-90, indicating good performance.
 - Some students had significantly lower scores.



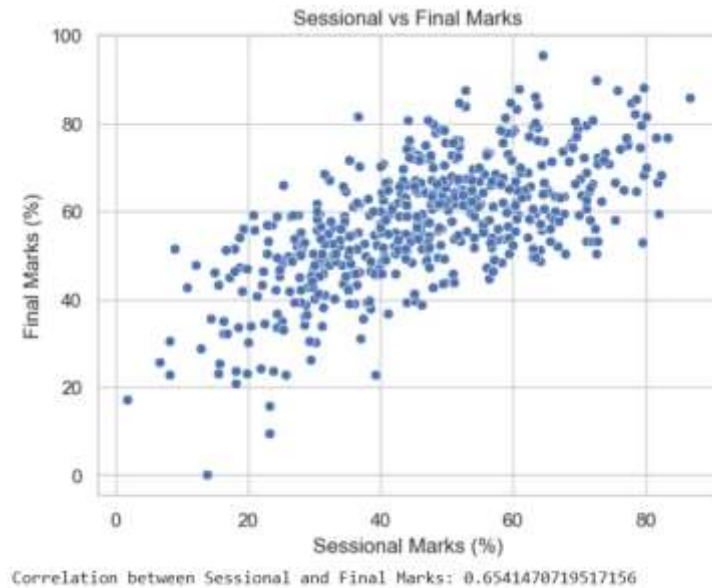
Step 2: Meaningful Query-Based EDA

Hypothesis

We hypothesize that there exists a significant relationship between the sessional performance of students, their attendance, and their final exam scores. Specifically, we expect that students who perform well in sessional assessments and maintain consistent attendance throughout the academic term are more likely to achieve higher scores in the final exams.

Query 1: Correlation between Sessional Marks and Final Exam Scores

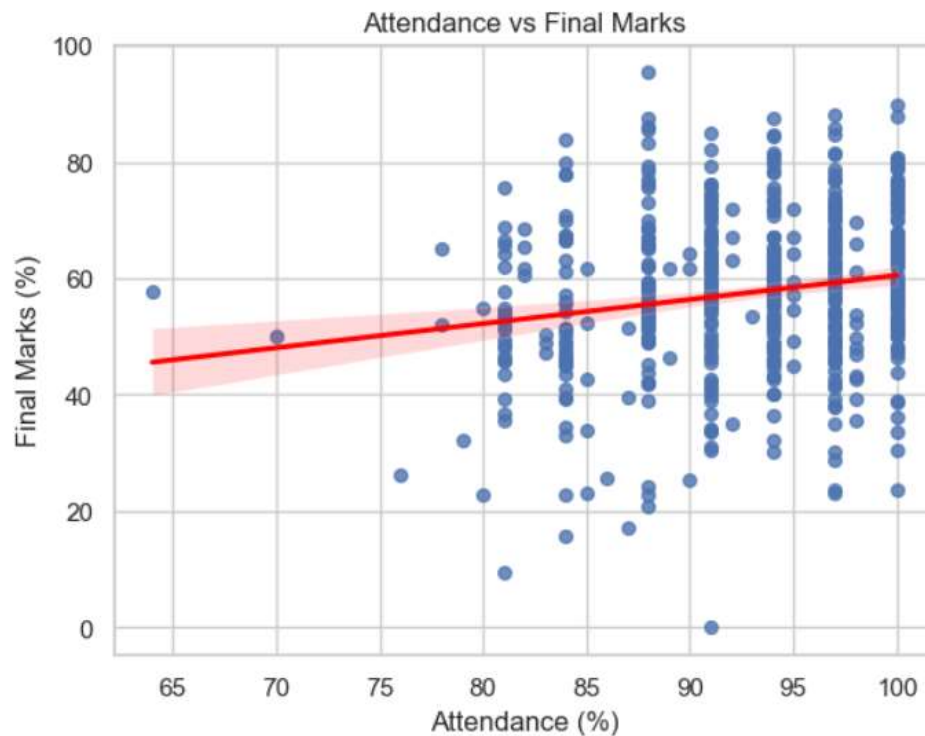
Objective: Check if there is a correlation between sessional marks and final exam scores.



Correlation of 0.654 means there is a moderate to strong positive correlation between Sessional Marks and Final Marks. Key Insights: Students who perform well in sessionals tend to perform well in finals. Since $r = 0.654$, there are still other factors influencing final exam performance.

Query 2: Impact of Attendance on Final Exam Performance

Objective: Analyze the relationship between attendance and final exam scores.



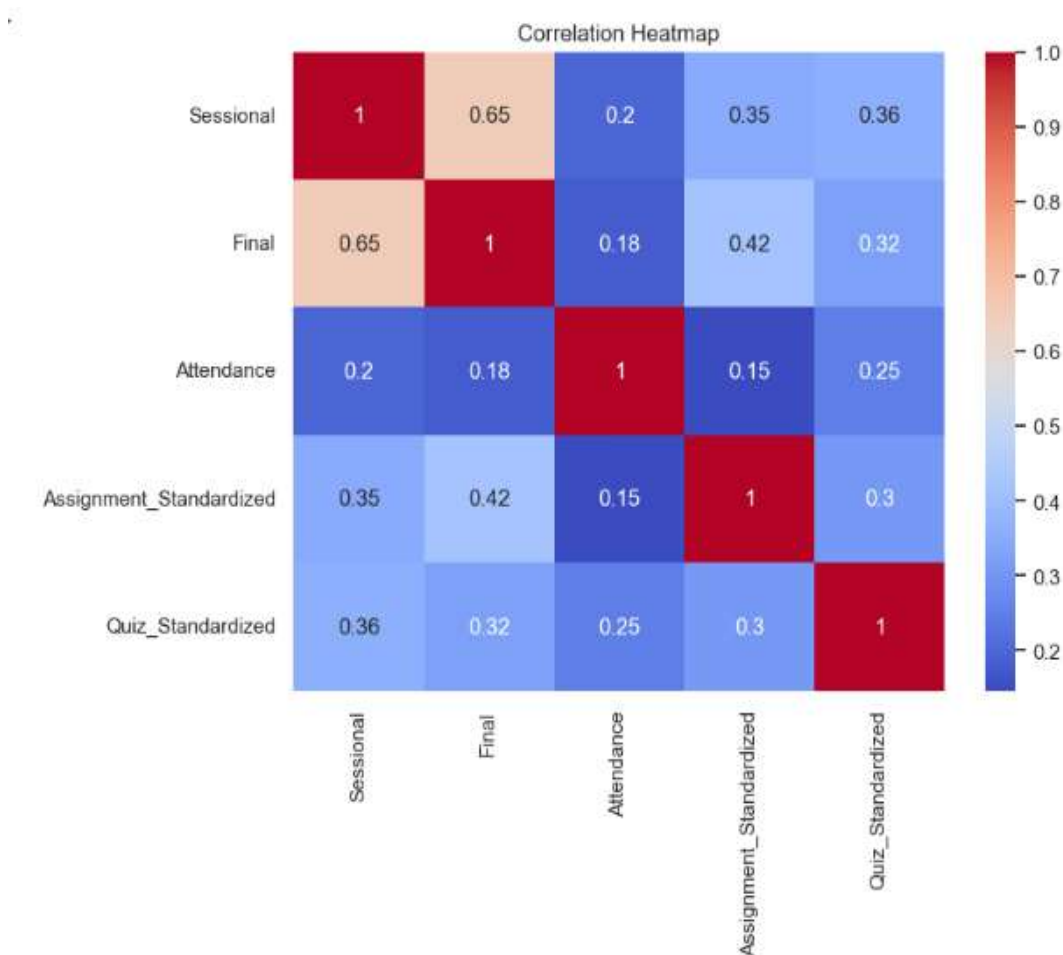
The correlation coefficient of 0.1787 suggests a weak positive correlation between attendance and final marks.

- **Weak Relationship:** While students with higher attendance tend to score higher in finals, the relationship is not strong. Attendance alone does not significantly determine final exam performance.
- **Other Influencing Factors:** Since $r = 0.1787$, other factors (such as study habits, sessional performance, or external resources) likely play a much bigger role in determining final scores.
- The scatter plot shows a lot of variation—students with similar attendance percentages can have very different final marks.

Key Takeaways: Simply attending classes is not enough to ensure high final exam scores.

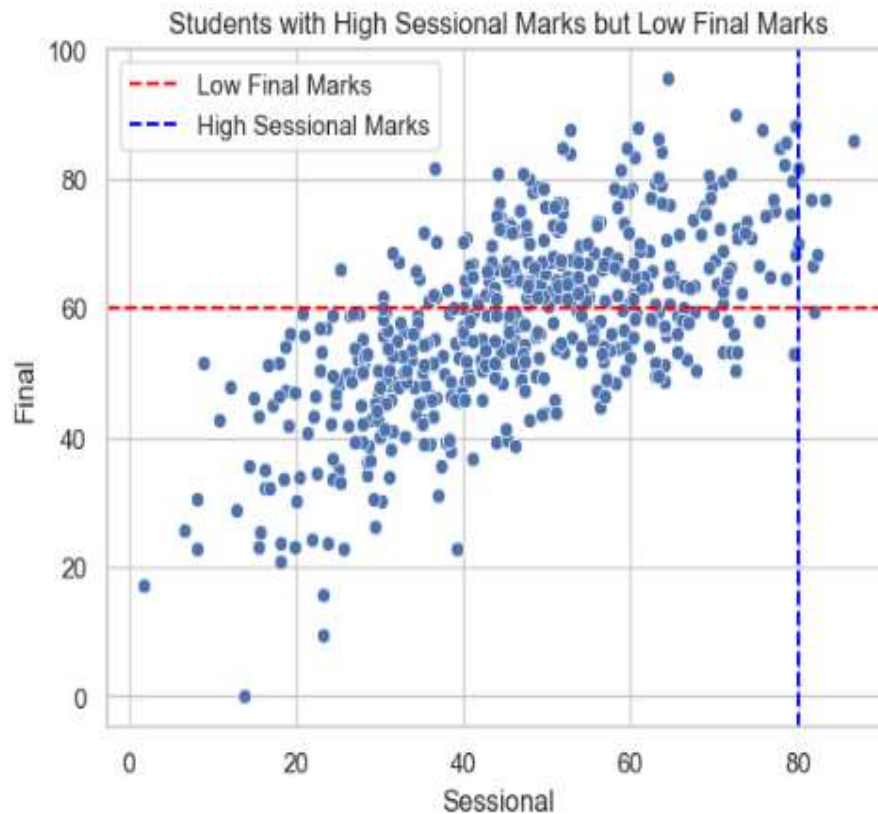
Query 3: Correlation Heatmap

Objective: Visualize the correlation between all numeric columns



Query 4: Students with High Sessional Marks but Low Final Marks

Objective: Identify students who performed well in sessional exams but poorly in final exams.



General Trend:

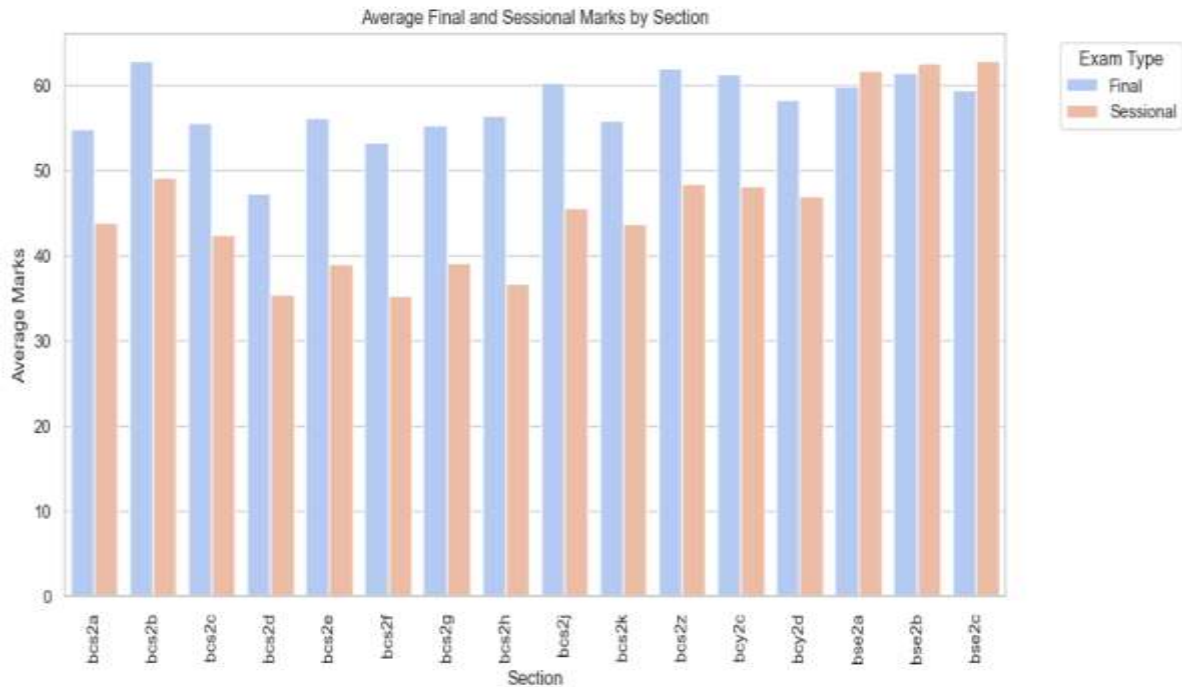
- The plot shows a positive correlation between sessional and final marks, meaning students with higher sessional scores generally tend to perform better in finals.
- However, there is considerable spread, indicating variability in student performance across assessments.

Key Focus: Students in the High Sessional - Low Final Region:

- The blue dashed line (vertical at 80) represents students with high sessional marks (>80).
- The red dashed line (horizontal at 60) represents students with low final marks (<60).
- Students in the top-left quadrant (above blue, left of red) scored high in sessionals but low in finals, which is the concern area.

Query 5: Section wise Final and Sessionals Marks Trend Analysis

Objective: Compare the average final marks across sections.

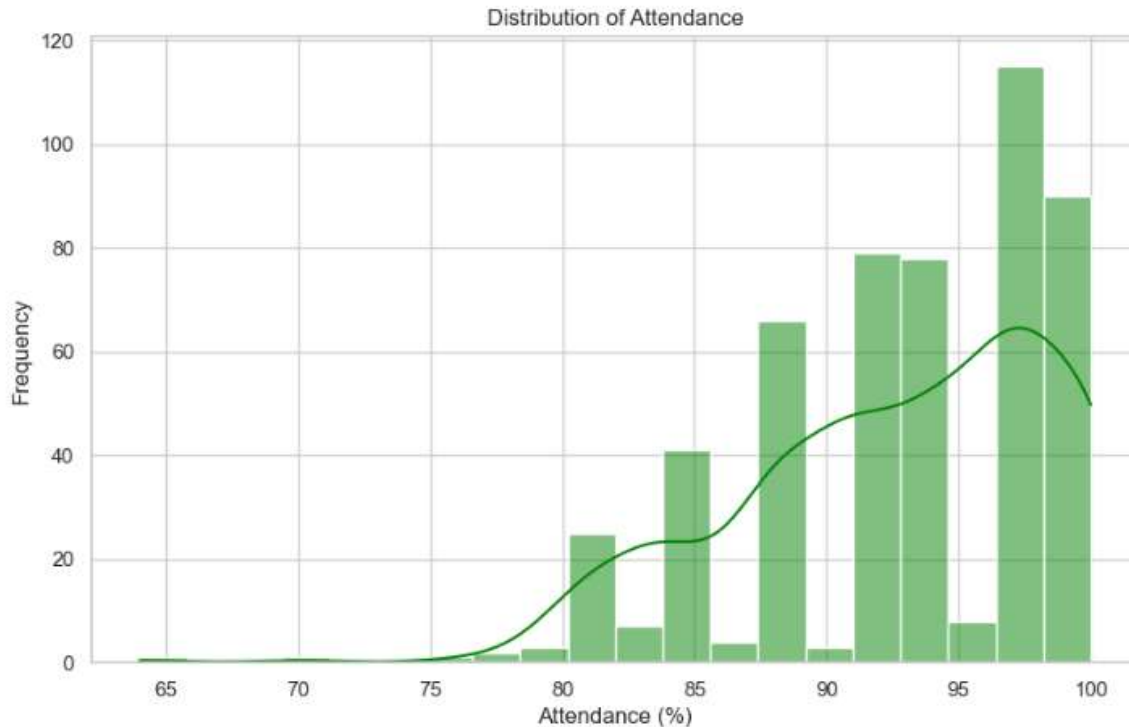


Insights:

- Overall Trend:**
 - There is a general consistency in final and sessional marks across most sections.
 - Some sections show higher sessional marks than final marks, while others have the opposite trend.
- Sections with Higher Final Marks:** Sections like bcs2b, bcs2g, bcs2h, bcs2z show significantly higher final marks compared to sessionals.
- Sections with Balanced Performance:** Sections bse2b and bse2c show very close final and sessional marks, indicating consistent performance in both assessments.
- Sections with Higher Sessional Marks than Final Marks:** Sections like bcs2d, bcs2f, and bcy2d have higher sessional marks than their final scores, which could indicate:
 - Lenient sessional grading or students struggling with final exams.
 - Less retention of knowledge from sessionals to finals.
- Section-Wise Performance Variation:**
 - Some sections show a large gap between sessional and final marks, possibly due to teaching differences, exam difficulty, or student preparation.
 - Sections like bcs2a, bcs2c, and bcy2c have relatively lower sessional and final marks, suggesting weaker overall performance.

Query 6: Distribution of Attendance

Objective: Analyze the distribution of attendance percentages.



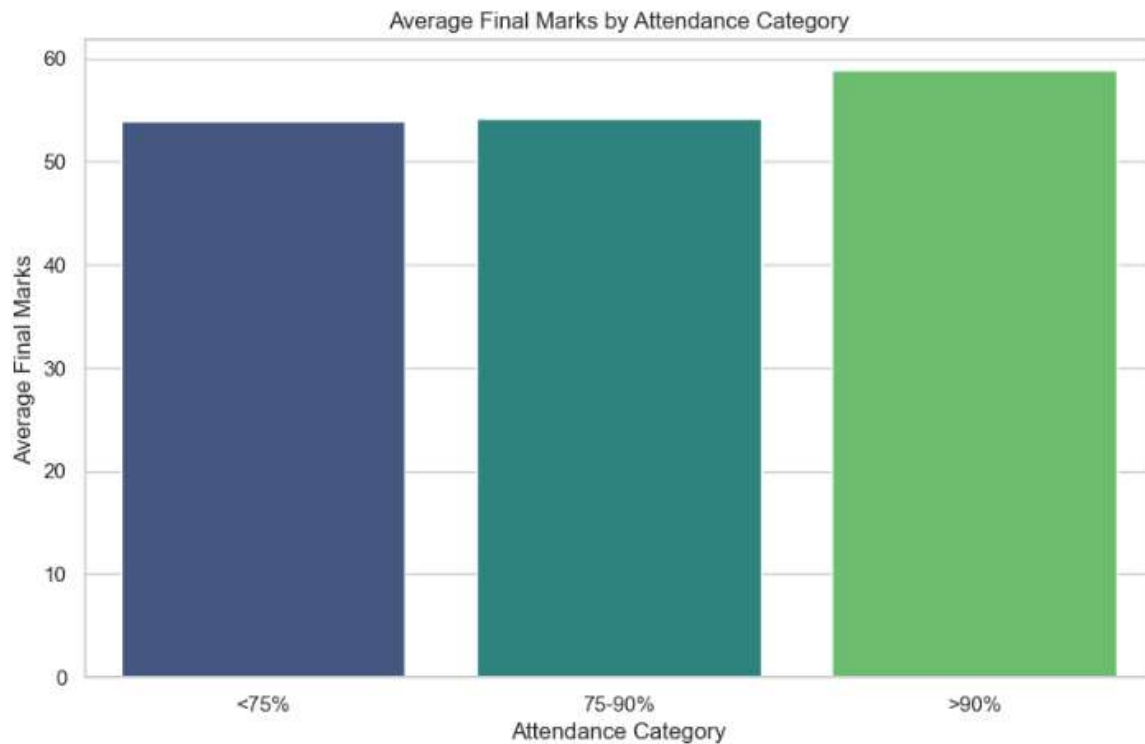
Insights from the Attendance Distribution Histogram

- **Overall Distribution Pattern:**
 - The histogram shows a right-skewed (positively skewed) distribution, indicating that most students have high attendance percentages.
 - The density curve (KDE) follows the histogram shape, confirming that attendance is concentrated at the higher end (90% - 100%).
- **Key Observations:**
 - Majority of students have high attendance (90%-100%)
 - Few students have lower attendance (<80%)
 - A significant number of students have nearly perfect attendance (close to 100%), possibly indicating a compulsory attendance requirement.
 - For students with low attendance (<80%):
 - Identify reasons for low attendance (health issues, lack of engagement, external responsibilities).

Query 7: Average Marks by Attendance Category

Objective: Categorize attendance into groups and calculate average final marks for each group.

```
sns.barplot(x="Attendance_Category", y="Final", data=attendance_final, palette="viridis")
```



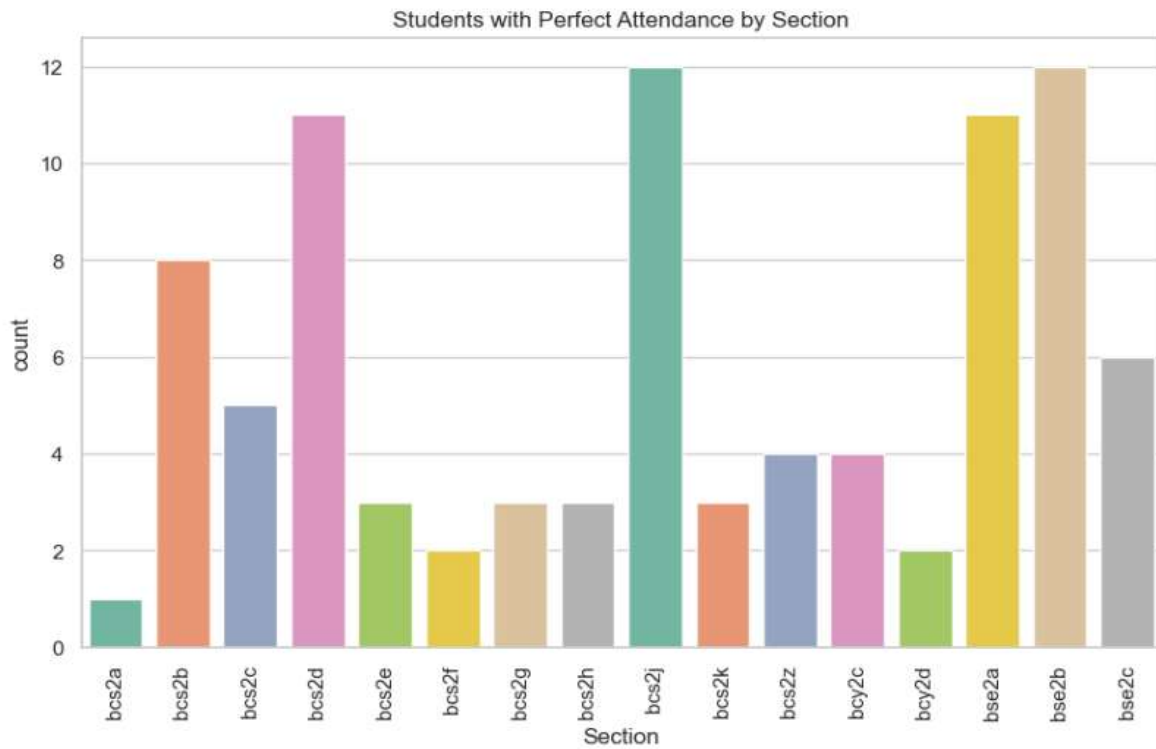
Insights from the Bar Chart:

- **Clear Positive Correlation Between Attendance and Final Marks**
 - Students with higher attendance (>90%) have the highest average final marks.
 - Students with moderate attendance (75-90%) have slightly lower marks.
 - Students with low attendance (<75%) have the lowest average marks.
- **Quantitative Differences**
 - While there is a positive relationship, the difference in marks is not extreme.
 - The average final marks increase steadily as attendance increases, indicating that while attendance helps, it's not the only factor determining performance.

Query 8: Students with Perfect Attendance by each Section

Objective: Identify students with 100% attendance.

```
sns.countplot(x="Section", data=perfect_attendance, palette="Set2")
```



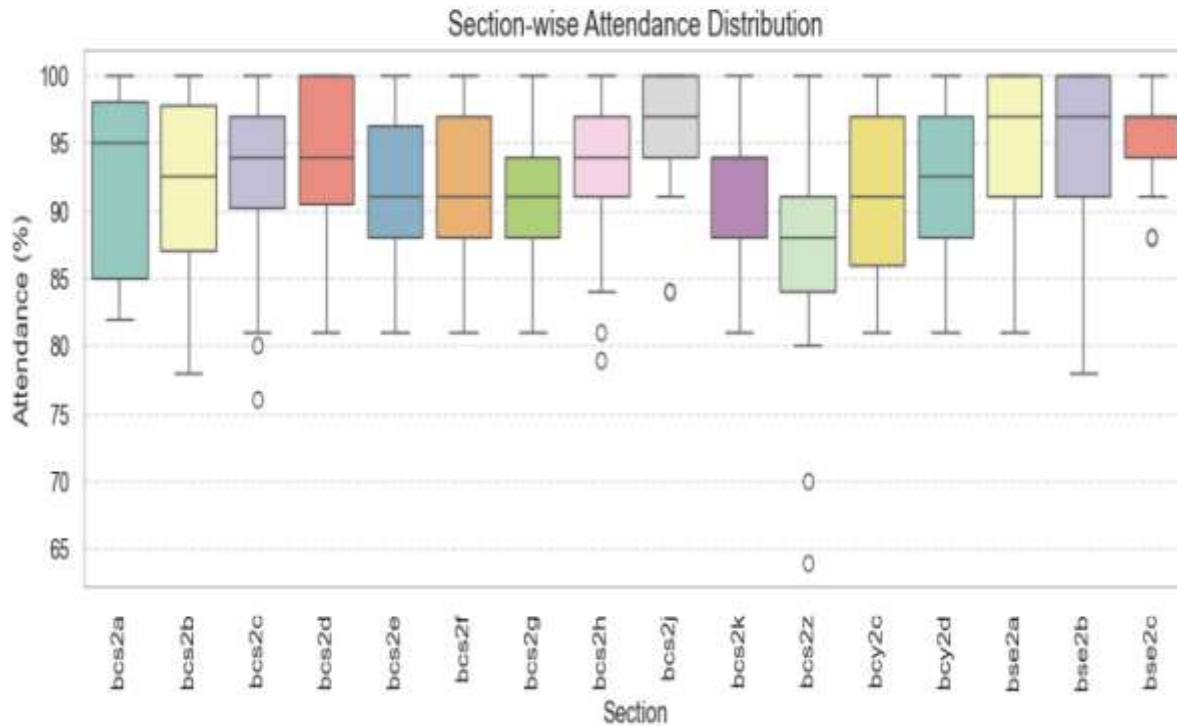
Insights

- 1. **Variation in Perfect Attendance Across Sections**
 - Some sections have significantly more students with perfect attendance than others.
 - Sections like bcs2j, bcs2d, bse2a, and bse2b have the highest number of students with 100% attendance.
 - Sections like bcs2a and bcy2d have very few students maintaining perfect attendance.

Query 9: Section-wise Attendance Distribution

Objective: Analyze the distribution of attendance percentages for each section.

```
sns.boxplot(x="Section", y="Attendance", data=final_data, palette="Set3", ax=ax)
```



Section	Q1	Median	Q3	IQR	Lower Bound	Upper Bound	Outliers
bcs2a	85.0	95.0	98.0	13.0	65.5	117.5	0
bcs2b	87.0	92.5	97.8	10.8	70.9	113.9	0
bcs2c	90.2	94.0	97.0	6.8	80.1	107.1	2
bcs2d	90.5	94.0	100.0	9.5	76.2	114.2	0
bcs2e	88.0	91.0	96.2	8.2	75.6	108.6	0
bcs2f	88.0	91.0	97.0	9.0	74.5	110.5	0
bcs2g	88.0	91.0	94.0	6.0	79.0	103.0	0
bcs2h	91.0	94.0	97.0	6.0	82.0	106.0	2
bcs2j	94.0	97.0	100.0	6.0	85.0	109.0	2
bcs2k	88.0	94.0	94.0	6.0	79.0	103.0	0
bcs2z	84.0	88.0	91.0	7.0	73.5	101.5	2
bcy2c	86.0	91.0	97.0	11.0	69.5	113.5	0
bcy2d	88.0	92.5	97.0	9.0	74.5	110.5	0
bse2a	91.0	97.0	100.0	9.0	77.5	113.5	0
bse2b	91.0	97.0	100.0	9.0	77.5	113.5	0
bse2c	94.0	97.0	97.0	3.0	89.5	101.5	2

Key Insights

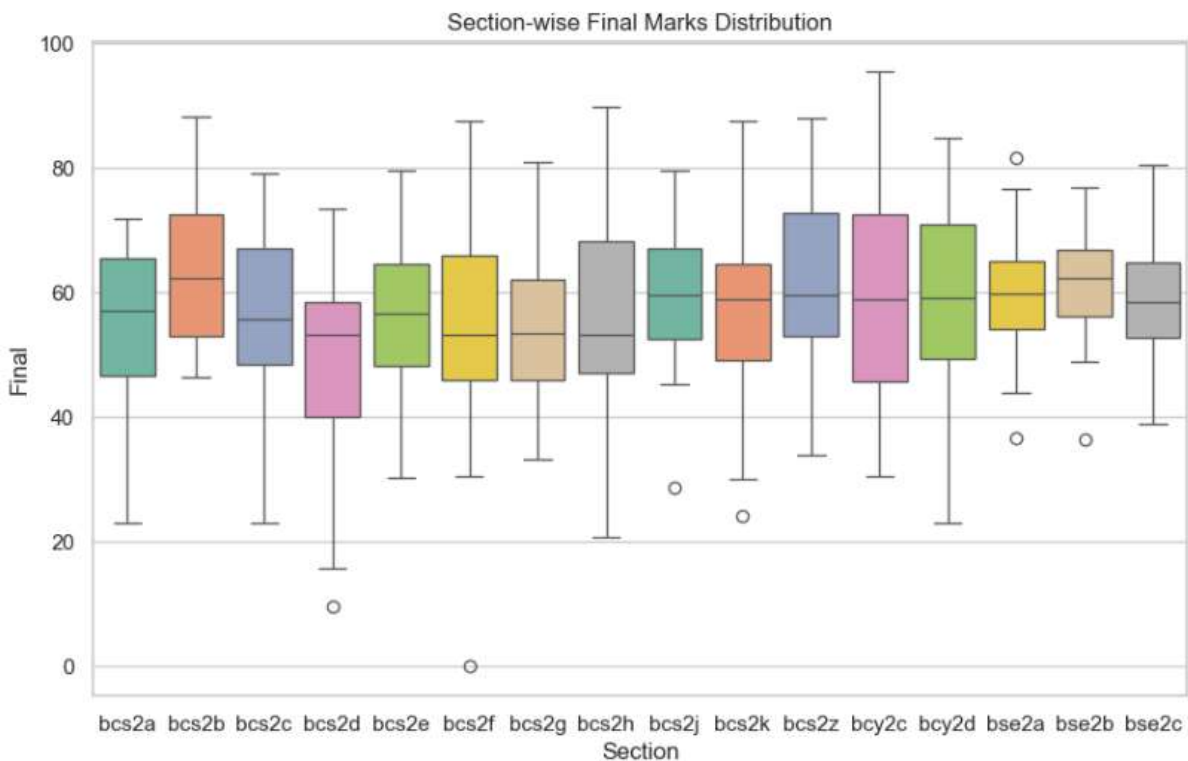
- **High Attendance Levels Overall:** Most sections have a median attendance above 90%, indicating a high attendance rate.
- **Variation Among Sections:**
 - Some sections (e.g., bcs2a, bcs2b, bcs2h) show a wider spread, meaning students in those sections have more variation in attendance.

- Other sections (e.g., bcs2d, bse2a, bse2b) have a tight range, suggesting most students maintain consistently high attendance.
- **Presence of Outliers:** Some students have significantly lower attendance (<80%), as seen in the dots below the whiskers in certain sections.

Query 10: Section-wise Final Marks Distribution

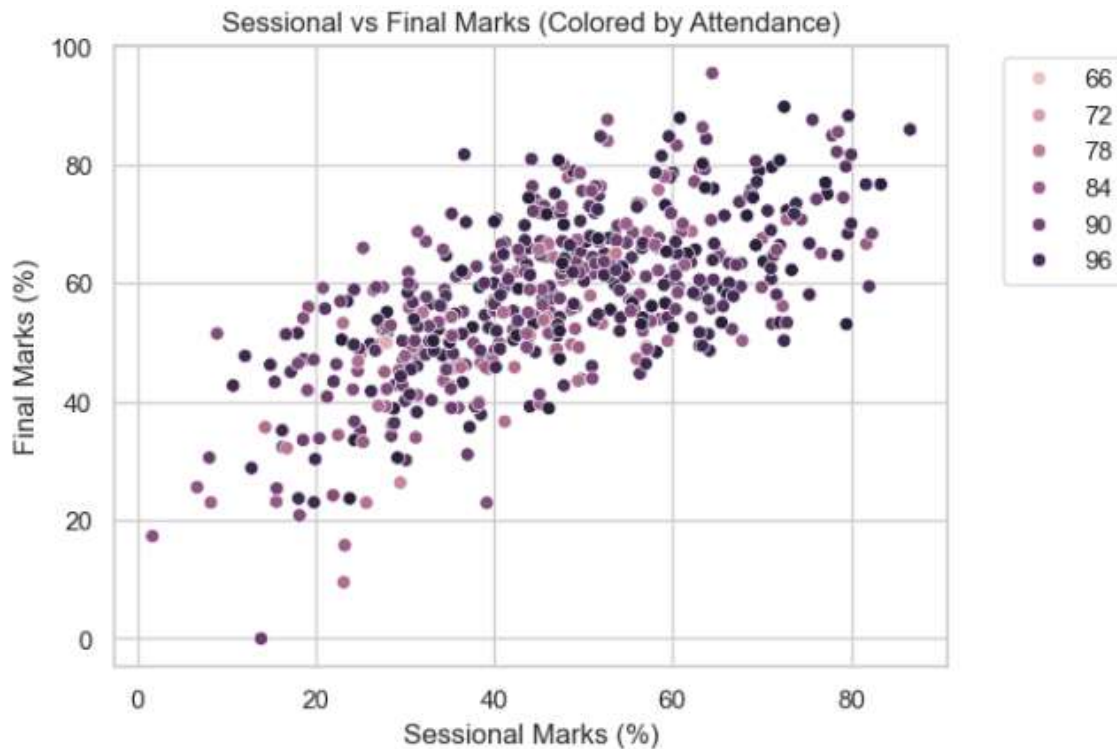
Objective: Analyze the distribution of final marks for each section

```
sns.boxplot(x="Section", y="Final", data=final_data, palette="Set2")
```



Query 11: Predictive Model for Final Marks

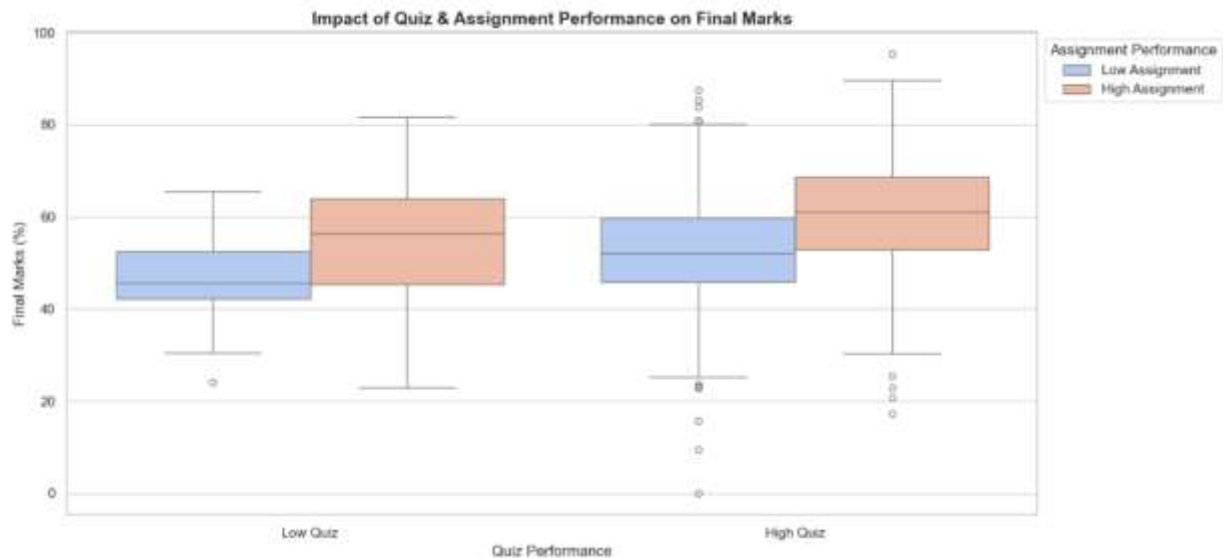
Objective: Explore the possibility of building a predictive model for final marks based on sessional marks and attendance.



Key Insights

- **1. Positive Correlation Between Sessional and Final Marks**
 - The scatter plot shows a clear positive trend, meaning students who score higher in sessional marks generally tend to score higher in final marks.
 - However, the spread is quite wide, indicating that sessional marks alone are not the only factor affecting final marks.
- **2. Role of Attendance (Hue)**
 - The color gradient represents attendance levels, where lighter shades (low attendance) and darker shades (high attendance) are distributed across the plot.
 - Darker points (high attendance) tend to cluster around higher final marks, suggesting that students with better attendance tend to perform better in final exams.
 - However, some low-attendance students still manage to get good final marks, indicating the presence of other influencing factors.
- **3. Potential for a Predictive Model**
 - This visualization supports the idea of a predictive model using sessional marks and attendance to estimate final marks.
- **4. Outliers & Weak Students**
 - Some students have very low sessional marks but still manage a decent final score, possibly indicating improvement in preparation.
 - Some students struggle in both sessional and final marks, which might highlight lack of engagement, weak preparation, or external factors

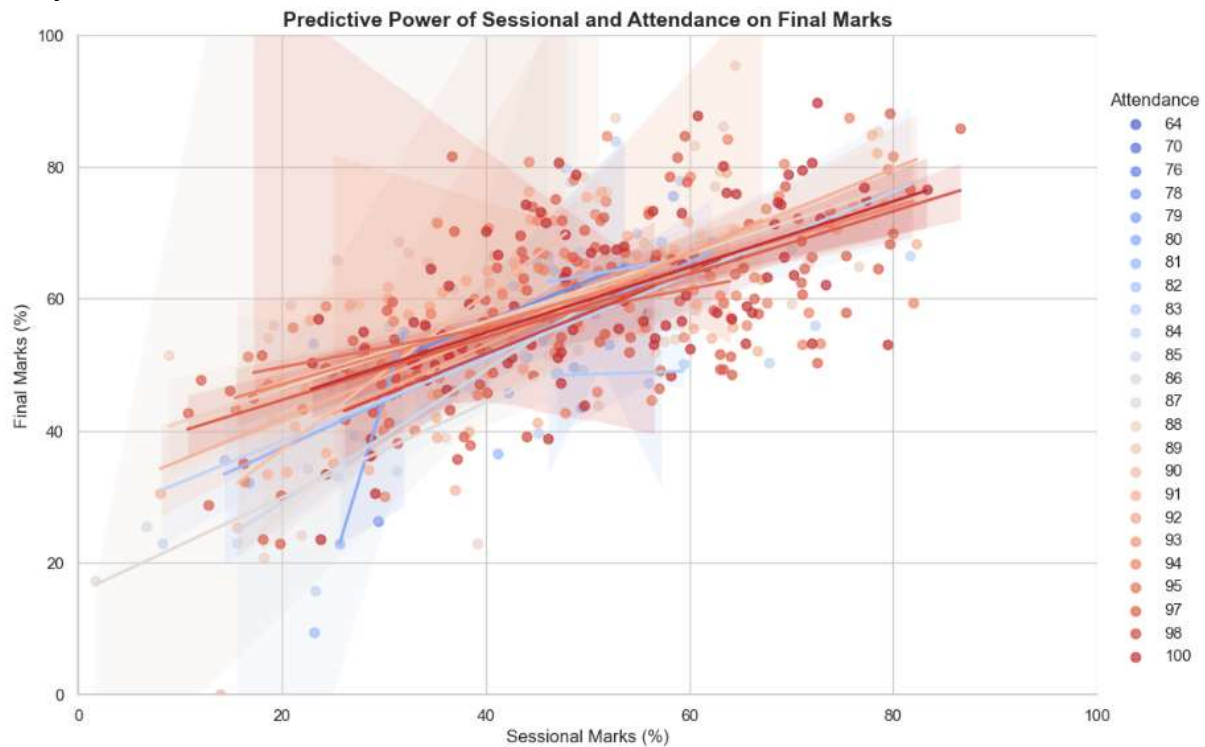
Query 12: Distribution of Final Marks Based on Quiz and Assignment Performance



Insights

- **Higher Quiz and Assignment Performance Lead to Higher Final Marks**
 - Students with **both high quiz and high assignment scores** (rightmost orange box) achieve the highest median final marks, confirming that good performance in both assessments significantly boosts final marks.
- **Assignments Have a Strong Impact on Final Marks**
 - In both **Low Quiz** and **High Quiz** categories, students who performed well in assignments (orange boxes) tend to score higher final marks compared to those with low assignment performance (blue boxes).
 - This suggests that assignments play a crucial role in improving final marks, regardless of quiz performance.
- **Low Quiz & Low Assignment = Struggle in Final Marks**
 - Students with low quiz and low assignment performance (leftmost blue box) show the lowest final marks.
 - The spread of scores is also smaller, indicating that most students in this category do not exceed a certain level.
- **Outliers Indicate Some Exceptionally Poor Performance**
 - The presence of **outliers below 20% final marks** suggests that some students struggled significantly, possibly due to external factors like lack of preparation, attendance issues, or difficulty in final exams.

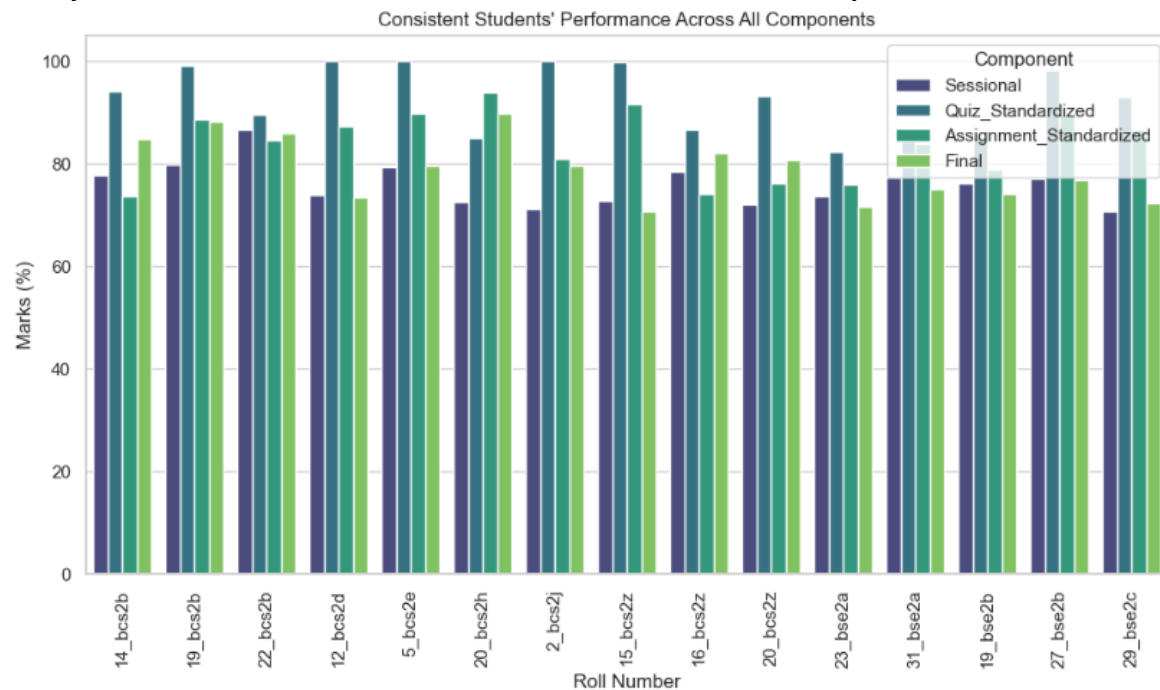
Query 13: Predictive Power of Sessional and Attendance on Final Marks



Key Insights

- **Positive Correlation:**
 - The regression lines indicate that higher sessional marks tend to predict higher final marks. This suggests a strong predictive relationship between sessional performance and final marks.
- **Impact of Attendance:**
 - The color gradient (Attendance Hue) shows that students with higher attendance generally perform better in final exams. Students with lower attendance (<80%) tend to have more scattered and lower final marks.
- **Outliers & Variability:**
 - Some students with low sessional marks managed high final marks—indicating possible last-minute effort or external factors.
 - Conversely, some with high sessional marks ended up with lower-than-expected finals, possibly due to exam difficulty

Query 14: Students with Consistent Performance across All Components



This bar chart represents the performance of students who have consistently scored above **70%** across all evaluation components:

- **Sessional Marks**
- **Standardized Quiz Marks**
- **Standardized Assignment Marks**
- **Final Marks**

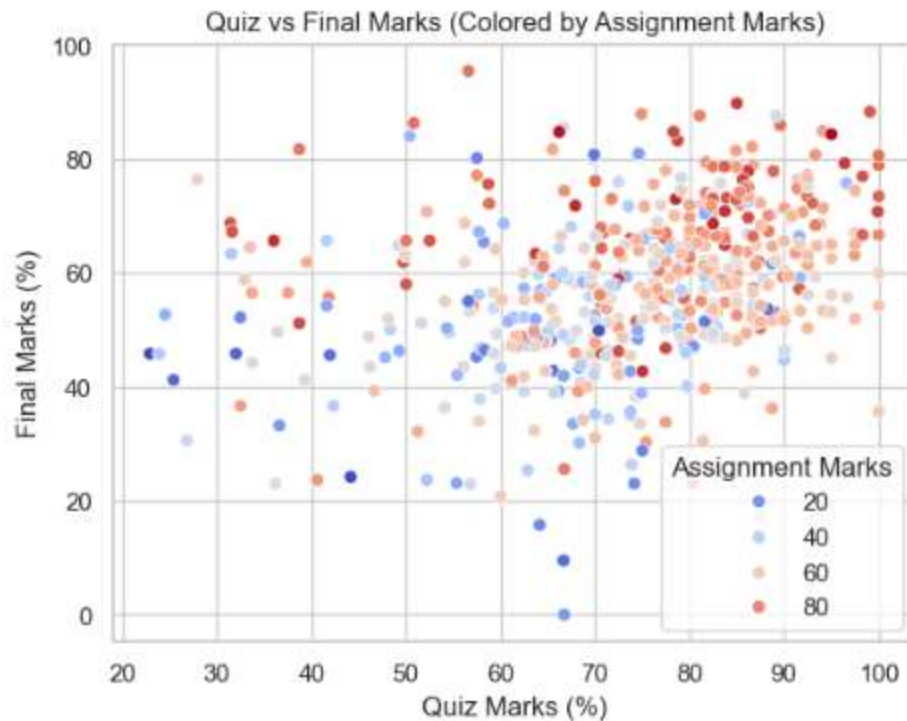
Key Observations:

- Balanced Performance Across Components:**
 - Most students perform **consistently across all four categories**, indicating well-rounded academic performance.
 - The bars for each student are similar in height, showing minimal deviation.
- Quiz Performance Stands Out:**
 - In many cases, **Quiz_Standardized scores are the highest**, suggesting students perform better in **frequent assessments** compared to assignments or finals.
- Final Marks Are Generally Stable:**
 - The **Final Marks** component shows fewer fluctuations than assignments or quizzes.
 - This implies that students who perform well throughout the semester **maintain** their performance in final assessments.
- Variations in Assignments & Sessional Marks:**
 - Some students show **minor drops in Assignment_Standardized scores**, which may indicate challenges in coursework-based evaluation.
 - Similarly, sessional marks exhibit some dips, reflecting varying engagement in classwork.

5. Possible Outliers:

- Some students have one component slightly lower than the rest, suggesting a **specific strength or weakness in different evaluation formats**.

Query 15: Impact of Quiz and Assignment on Final Marks



This scatter plot visualizes the relationship between **Quiz Marks (%)** and **Final Marks (%)**, with colors representing **Assignment Marks (%)**.

Key Observations:

1. Positive Correlation Between Quiz and Final Marks:

- A general upward trend suggests that students who perform well in quizzes tend to score higher in final exams.
- However, the correlation is **not perfectly linear**, indicating other influencing factors.

2. Influence of Assignment Marks:

- Red-colored points** (higher assignment marks) are **clustered in the upper-right**, meaning students who excel in assignments often achieve higher final marks.
- Blue-colored points** (lower assignment marks) appear more scattered, particularly in the lower-left region, indicating weaker performance across components.

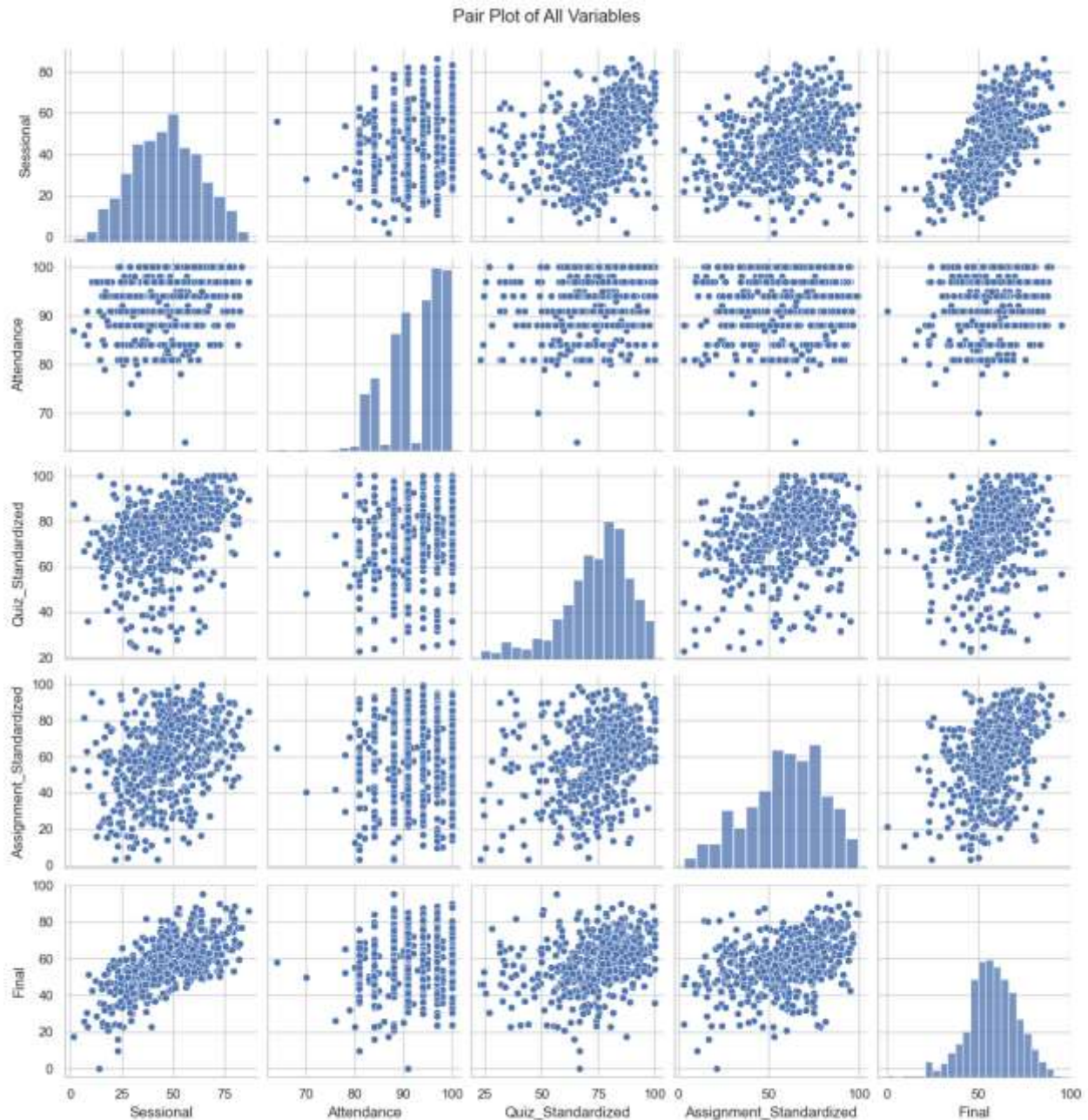
3. Variability in Final Marks at Mid-Range Quiz Scores (50-70%):

- Some students with **moderate quiz scores** (50-70%) still achieve **high final marks**, suggesting other factors (such as assignments, attendance, or sessionals) contribute to overall performance.

4. Students with Low Assignment Marks (<40%) Struggle in Finals:

- Blue-colored points dominate the lower section, confirming that students with weak assignment performance are more likely to struggle in final exams.

Query 16: Predictive Model for Final Marks



This pair plot visualizes the relationships between **Sessional Marks**, **Attendance**, **Quiz Marks**, **Assignment Marks**, and **Final Marks**.

Key Observations:

- Strong Correlation Between Sessional and Final Marks**

- A **clear upward trend** indicates that students who perform well in sessional assessments also tend to score high in final exams.
- This suggests **sessional marks are a strong predictor** of final performance.
- 2. **Quiz and Final Marks Show a Positive Relationship**
 - The scatter plot shows a **visible positive correlation**, meaning students with **higher quiz scores tend to achieve higher final marks**.
 - However, some variability suggests that **quizzes alone are not the sole determining factor**.
- 3. **Assignment Marks Have a Moderate Correlation with Final Marks**
 - While higher assignment marks generally indicate higher final scores, the relationship is **not as strong as sessional or quiz marks**.
 - This suggests that **assignments contribute to performance but are not the primary driver**.
- 4. **Attendance Shows a Weak or Non-Linear Relationship**
 - The scatter plots involving attendance appear **dispersed and non-trending**, indicating **attendance alone does not strongly predict** final marks.
 - Most students have **high attendance (above 70%)**, so the dataset lacks variability to show its full impact.
- 5. **Distributions Reveal Normality Trends**
 - The histograms along the diagonal show that **most variables follow a near-normal distribution**.
 - **Final marks** have a slight right skew, suggesting some student's **score lower than average** more frequently than scoring extremely high.

Conclusion

The analysis exhibits a systematic process of ingestion, scrubbing, transformation, and in-depth EDA. The learning that can be derived from the analysis can help teachers and administrators identify patterns, improve teaching, and support struggling students.

Reference

<https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>

https://github.com/zoya532/DV_Assignment1