Data Analysis Project

Presented by : Sajjad Rezvani

Sajjadr@bgsu.edu

Contents

Introduction

This is a data analysis project that analyzes the grades of students for a course. These data about students and their grades are consolidated in three tables: students, assignments, and grades.

Project Objectives

1. Calculate Final Grades and extracting insights on grades

Determine overall student performance & Analyze distribution and trends.

2. Visualizations

Create graphical representations of data.

3. Hypothesis Testing

Assess statistical significance.

4. Letter Grade Distribution

Evaluate course performance.

Data Structure

Students Table

180 students, 4 columns (id, name, level, major).

	student_id	student_name	student_level	major
0	288941	Matthew	1-Freshmen	English
1	463818	Austin	4-Senior	English
2	465208	Tyler	3-Junior	Engineering
3	383634	Samantha	3-Junior	Science
4	689448	Brittany	3-Junior	English

Grades Table

3600 rows, 3 columns (student id, assignment id, grade).

	student_id	assignment_id	numeric_grade
0	288941	1	4
1	288941	2	7
2	288941	3	10
3	288941	4	16
4	288941	5	5

Assignments Table

20 rows, 3 columns (id, category, value).

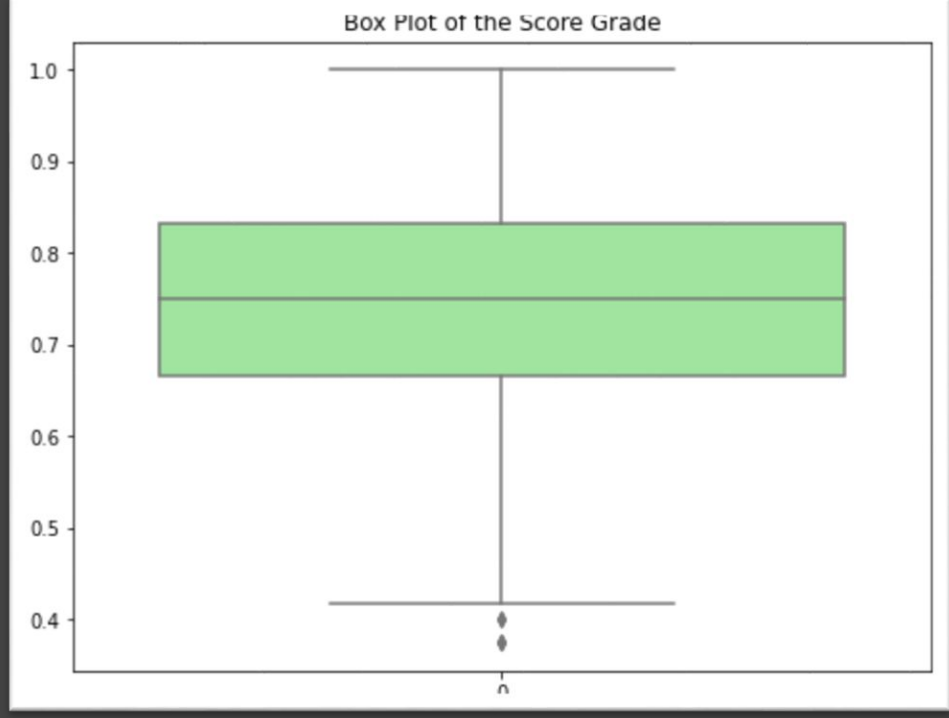
	assignment_id	assignment_category	assignment_value
0	1	Homework	8
1	2	Homework	8
2	3	Homework	12
3	4	Quiz	22
4	5	Homework	8

Data Integrity

No null values, no duplicates, some outliers

Outlier Detection

Boxplot

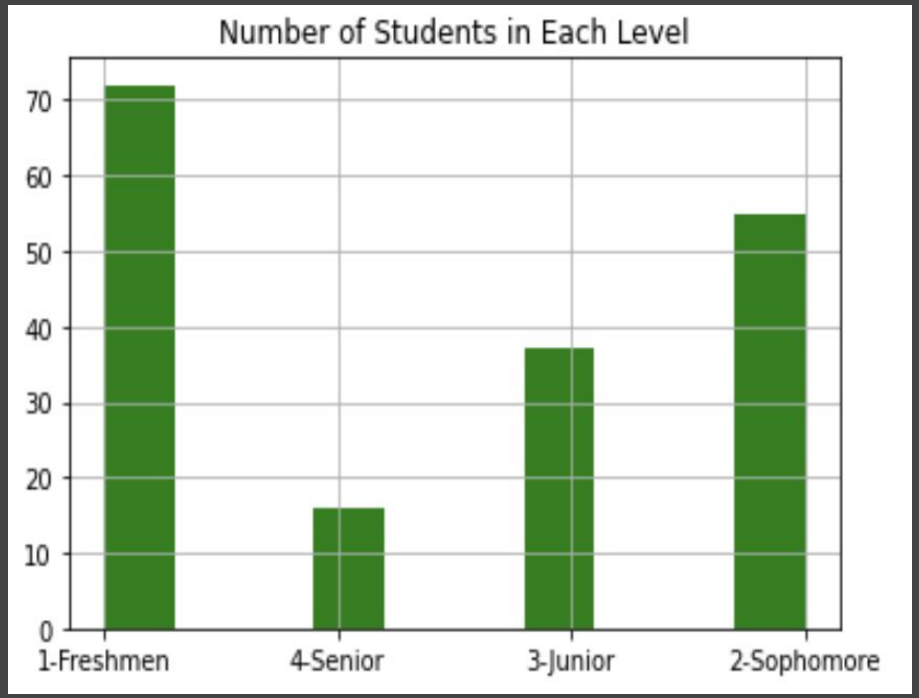


Detected Outliers are as below

	student_id	student_name	student_level	major	assignment_id	numeric_grade	assignment_category	assignment_value	score_grade
152	514068	Nicholas	1-Freshmen	Business	13	3	Homework	8	0.375
348	188048	Nicole	1-Freshmen	Engineering	9	4	Homework	10	0.400
1544	996584	Eric	3-Junior	Engineering	5	3	Homework	8	0.375
1976	709559	Stephanie	1-Freshmen	Engineering	17	4	Homework	10	0.400
3100	969187	Tyler	1-Freshmen	Engineering	1	3	Homework	8	0.375

Visualizing Demographics

Histograms provide understanding of student distribution among levels and majors. We group data based on specific columns to see which majors are more popular and how students are distributed across different levels.



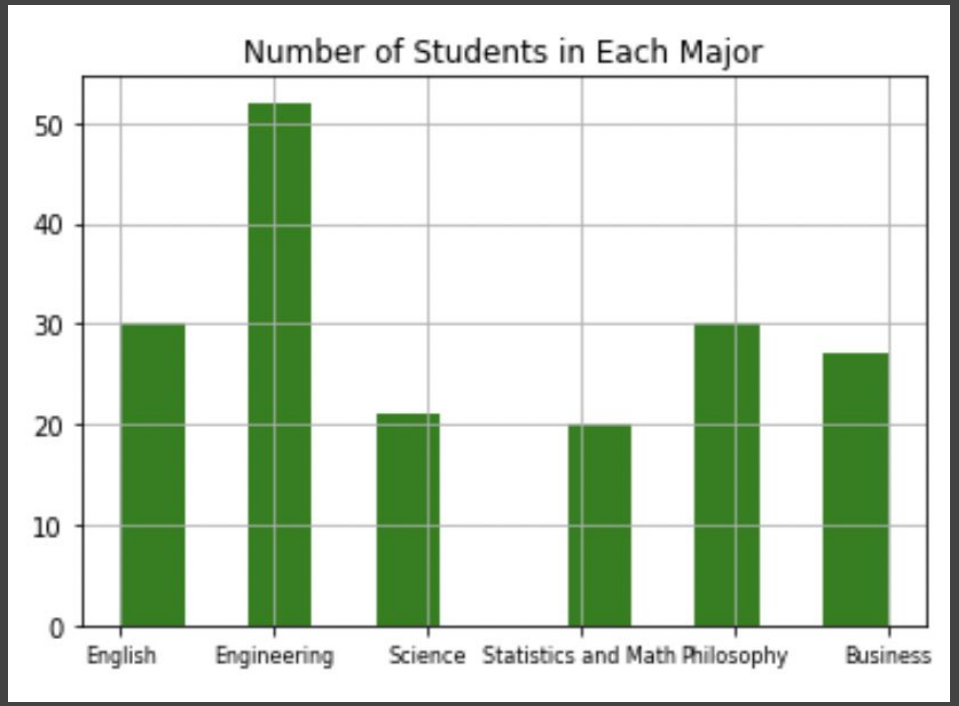
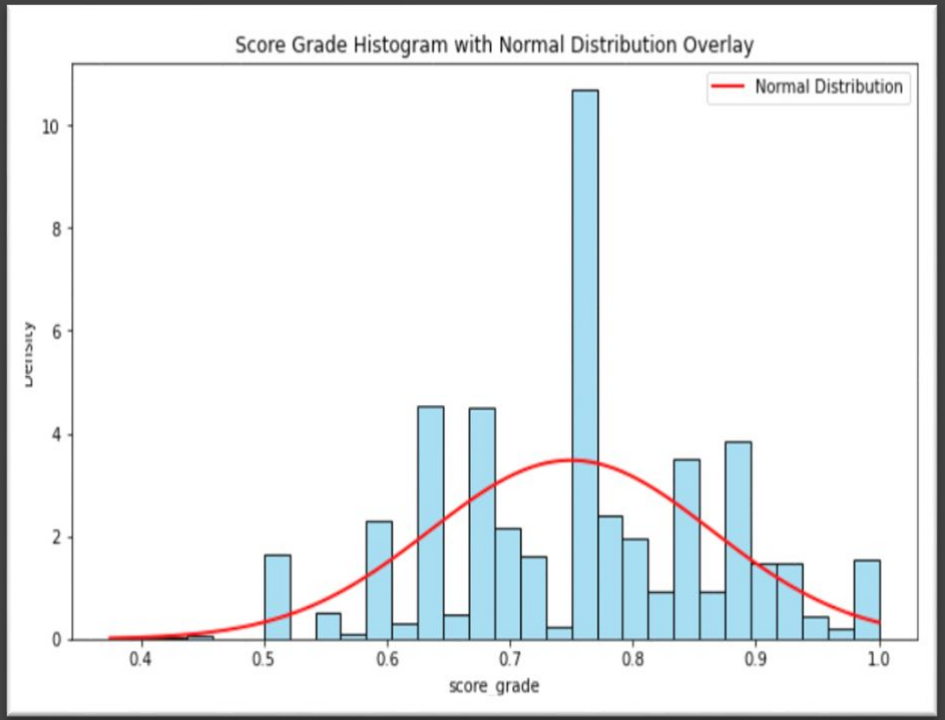


Table of Students in Each Level/Major

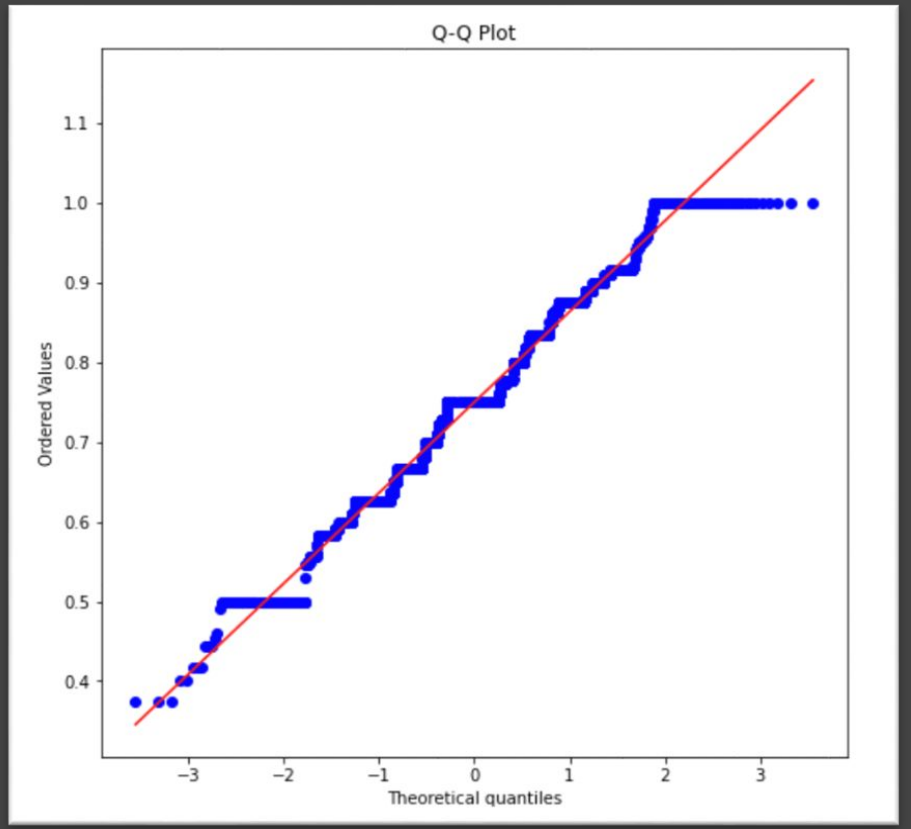
student_level	major	count
4-Senior	Philosophy	1
3-Junior	Statistics and Math	2
4-Senior	Science	2
4-Senior	Statistics and Math	3
3-Junior	Business	3
4-Senior	Business	4
3-Junior	Science	5
3-Junior	English	5
4-Senior	English	6
1-Freshmen	Statistics and Math	6
2-Sophomore	Science	6
2-Sophomore	Philosophy	7
1-Freshmen	Science	8
2-Sophomore	English	8
3-Junior	Philosophy	8
1-Freshmen	Business	9
2-Sophomore	Statistics and Math	9
2-Sophomore	Business	11
1-Freshmen	English	11
2-Sophomore	Engineering	14
1-Freshmen	Philosophy	14
3-Junior	Engineering	14
1-Freshmen	Engineering	24

Score Grade Distribution

- Distribution of score grades
- Mean = 0.749 , Median = 0.75 , Std= 0.115



- Q-Q plots also support normal distribution



Shapiro Test for Normality

H0 (Null Hypothesis) : data is normally distributed

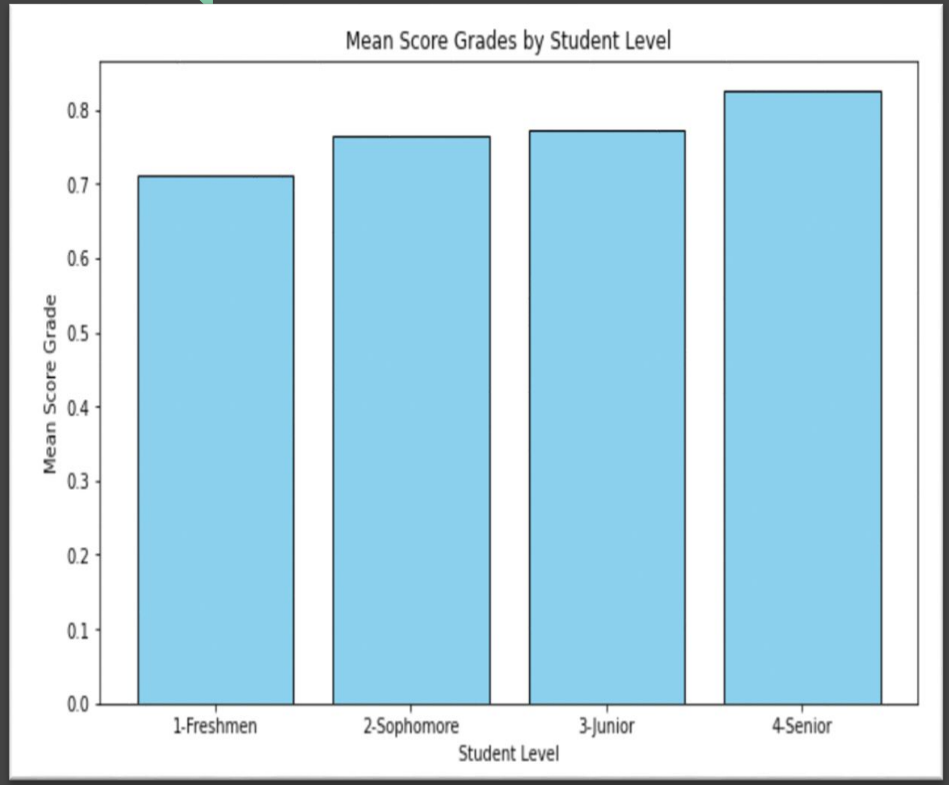
H1(Alternative Hypothesis): data is not normally distributed

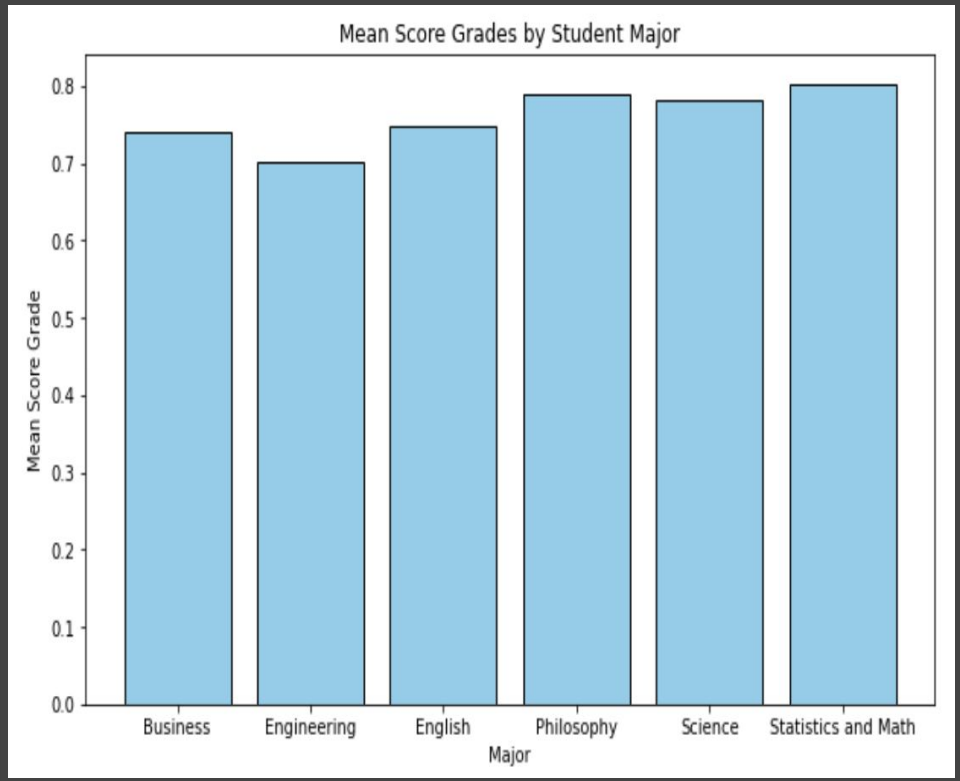
- Test Statistic = 0.9825, p-value = 1.11e-20 ≈ 0
- P-value is significantly small(<alpha=0.05 significance level), allowing us to reject the null hypothesis.
- Conclusion: Data is not normally distributed as per the Shapiro test.
- Considerations: Shapiro test is less effective for large datasets.

Mean Score Grades Insights

Visualizing mean score grades by student level and major

- Senior level highest (0.82)
- Freshmen lowest (0.71)
- Statistics highest
- Engineering lowest





Hypothesis Testing

- Significant difference between mean scores of freshmen and seniors was tested using T-Test

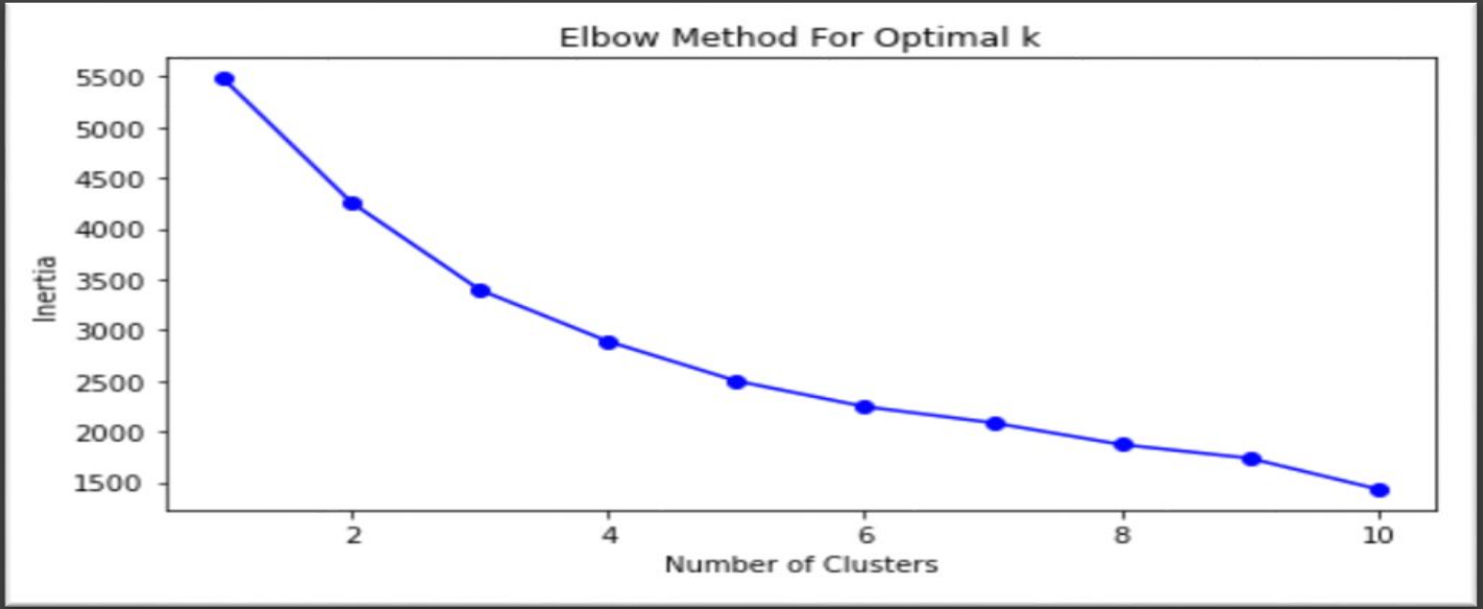
- Null Hypothesis (H0): No significant difference. Alternative Hypothesis (H1): Significant difference

- Result: P-value =7.89e-59≈ 0, less than significance level (alpha=0.05) , then reject H0

- Conclusion: Statistically significant and meaningful difference in mean

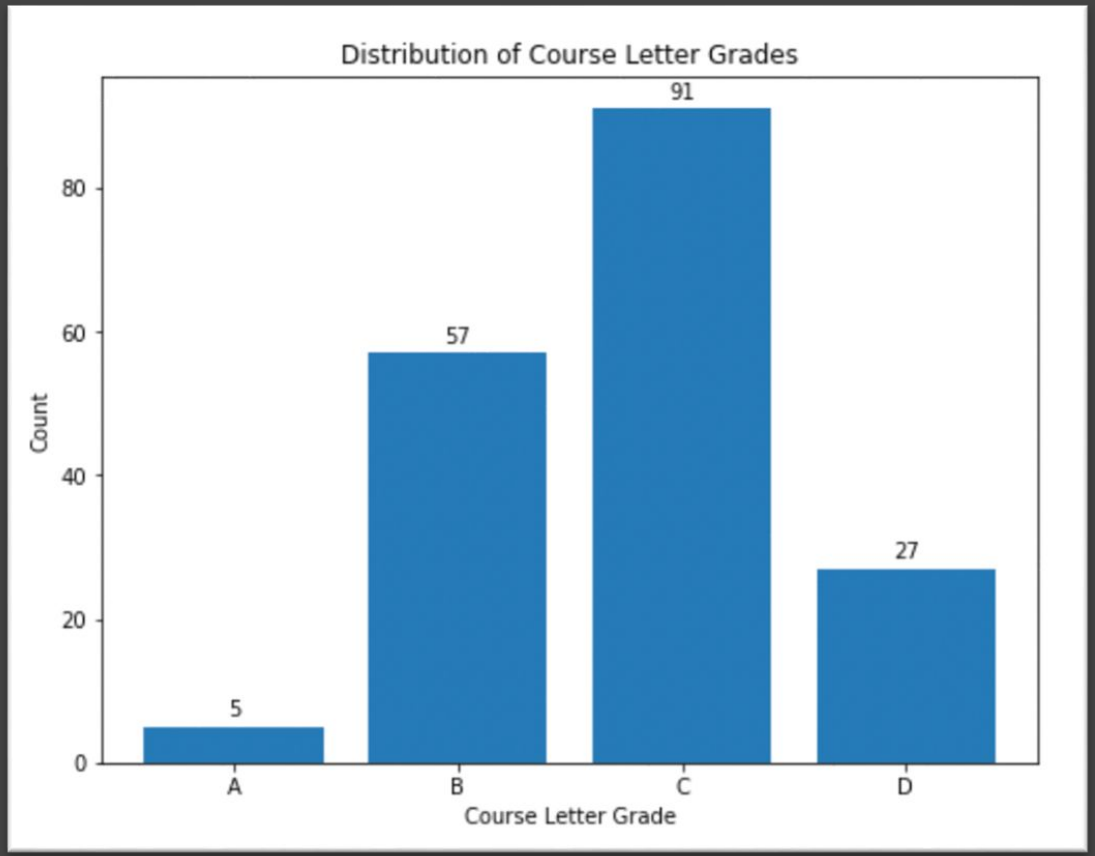
Cluster Analysis

- **Clustering**: Unsupervised learning method used to group data points by similarities utilizing K-Means algorithm
- **K=4 Clusters** Represents distinct patterns, such as different performance levels of students in different academic levels/majors

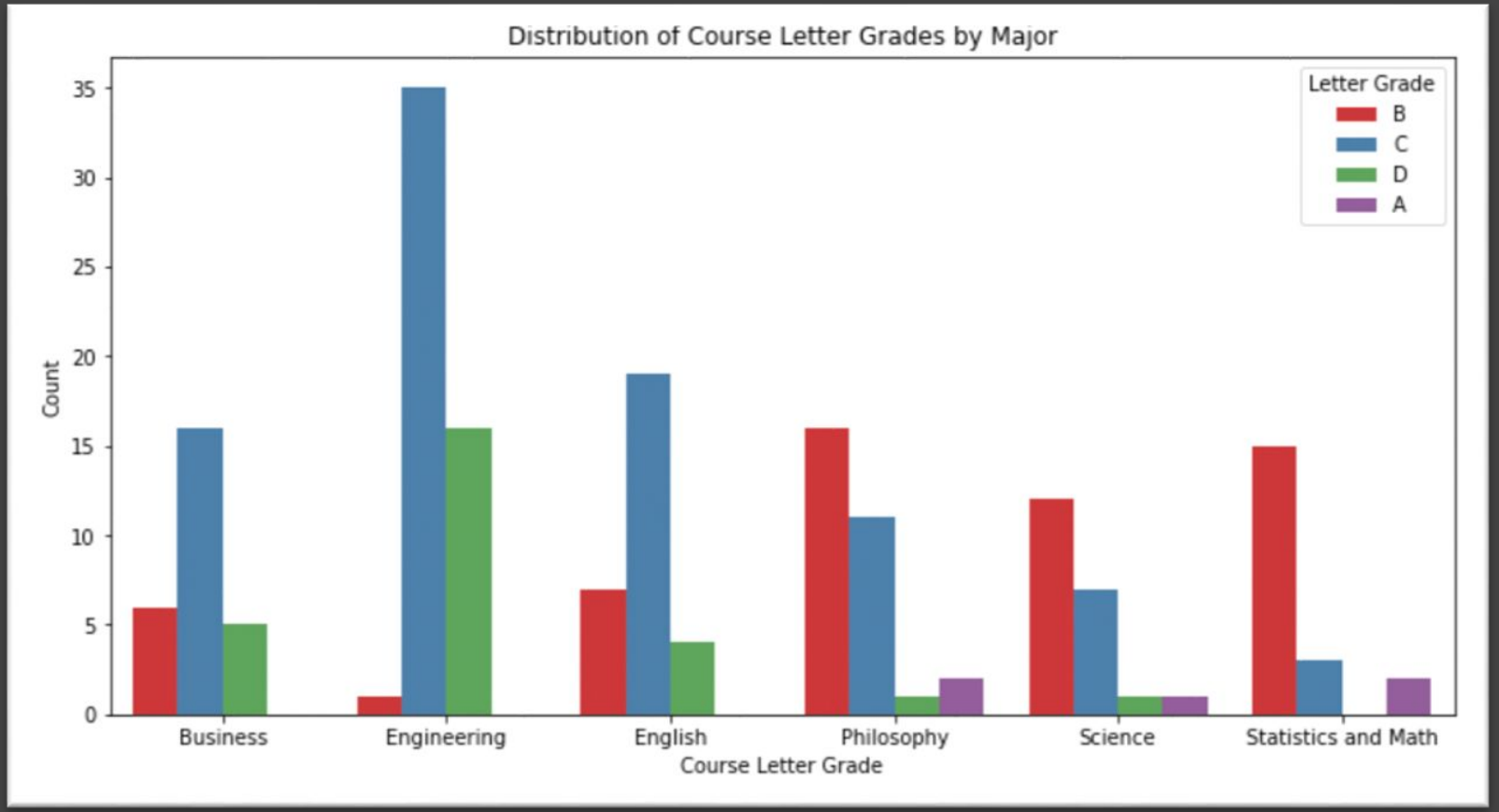


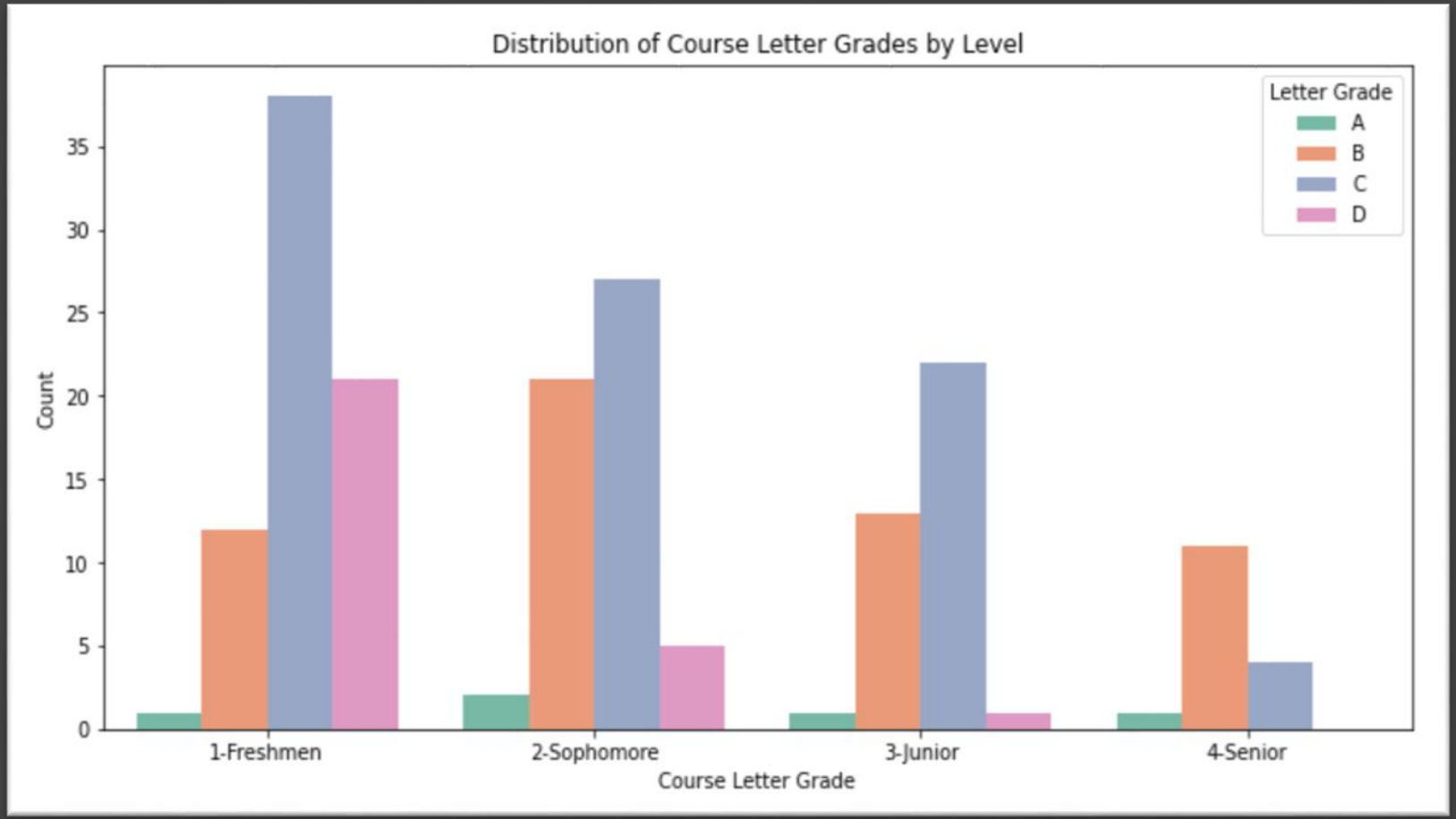
Course Letter Grades

- Weighted average score grades were calculated: Exam (60%), Quiz (15%), Homework (25%).
- Letter grades: A > 90, B > 80, C > 70, D > 60, F ≤ 60.



Visualizing Letter Grades by Major & Level




Distribution of Course Letter Grades by Level
Letter Grade
A
B
C
D
Count
0
5
10
15
20
25
30
35
1-Freshmen
2-Sophomore
3-Junior
4-Senior
Course Letter Grade

Future Works

- Developing an automated Python pipeline for processing data, calculating final grades, and generating visualizations.
- This system will handle diverse datasets and reporting needs, providing a modular approach for scaling across multiple courses and data sources.

Thank You

- Questions…?