Predicting Final Exam Scores in Course 451
Based on Homework and Mid-Term Performance

Group 08

Background: This analysis focuses on understanding how students' performance in assignments and mid-term exams in Course 451 can predict their final exam scores. By analyzing this data, educators and students can gain a deeper understanding of the academic journey in Course 451. Additionally, this analysis can empower students by providing them with a clearer understanding of which aspects of their coursework have the most significant impact on their final grades. By identifying key performance indicators through the predictive model, students can gain insights into where they should focus their efforts more intensely to achieve better outcomes.

Objective: The primary objective of this project is to construct a predictive model that can accurately estimate students' final exam scores based on their homework and mid-term exam results in Course 451.

Code to read data:

```
grade = pd.read_csv('451spring23.csv', index_col=0)
grade.head()
```

	Q00.	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	HW1.html	HW1.ipynb	HW02.html	HW02.ipynb	HW03.html	HW03.ipynb
1	4	7	5	12	16	14	10	3	4	17.0	3.0	16.0	3.0	17.0	3
2	4	7	5	12	16	14	10	3	4	17.0	3.0	17.0	3.0	17.0	3
3	4	7	5	12	16	14	10	3	4	16.0	3.0	16.5	3.0	17.0	3
4	4	7	5	12	16	14	10	3	4	17.0	3.0	17.0	3.0	17.0	3
5	4	7	5	12	16	14	10	3	4	17.0	3.0	17.0	3.0	17.0	3

	HW04.ipynb	HW05.html	HW05.ipynb	MidtermExam	FinalExam	Proposal	Presentation	PresentationFeedback27april	PresentationFeedback2may
1	3	17.0	3.0	100.0	73	7	26.90	2	2
2	3	17.0	3.0	100.0	69	7	23.60	2	2
3	3	15.0	3.0	97.0	75	7	23.91	2	2
4	3	17.0	3.0	98.0	69	7	24.46	2	2
5	3	15.0	3.0	96.0	73	7	23.91	2	2

	ReportPeerFeedback	Percentage	Rank	Percentile	percentageGrade	percentileGrade	Grade
1	4	99.48	71	98.6	A	A	A
2	4	97.90	70	97.2	A	A	A
3	4	97.85	69	95.8	A	A	A
4	4	97.62	68	94.4	A	A	A
5	4	97.48	67	93.0	A	A	A

Variables:

The dataset for the 23 Spring STAT451 Statistical Table of Student Achievement has 34 columns and 69 variables, primarily focused on student performance in quizzes, homework, exams, and projects. Here's a summary of its structure and content:

There are 34 columns in total, which include:

- number of the student
- Quiz 01-08
- Homework1-5：Each of these assignments contains a grade for a jpynp file and a grade for an HTML file.
- Two exams: midterm and final.
- Projects: Includes Proposal, Presentation, Presentation Feedback, Report, and Report Peer Feedback.
- Percentile and Grades: Percentage, Rank, Percentile, percentage Grade, percentile Grade, and Grade.

Quiz 01-08, Homework1-5, Two exams, and Projects Percentage, Rank, and Percentile are all numerical variables.
Percentage Grade, percentile Grade, and Grade are categorical variables.The variables we are using have not yet been fully determined, which is an issue that needs to be explored.

Method:
We plan to use last year's more complete data set as the primary subject of our study, splitting it into a test set and a training set in a 20:80 ratio. The overall research is divided into two directions.
In the first, we will use class rankings as the response variable and employ the KNN classification method to predict the final class ranking of students.
In the second approach, we will simply consider the final exam scores as the response variable and estimate these scores using linear regression models,KNN regression and decision trees. Ultimately, we, of course, want to know this year's final exam scores, but some data, such as HW5, are missing. Therefore, we will use a subset of standardized variables for prediction.

Question:
1. Should we combine Q1-Q8 into a single 'quiz' variable and HW1-HW5 into a single 'homework' variable?
2. Whether to include presentation scores as a key variable in your analysis depends on several factors?
3. We have one outlier, it need to be deleted, right?