Data Analysis Project:

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

Objectives:

This data analysis project calculates final grades, provide some insights about these grades, present some visualizations, does a hypothesis testing and provides distribution of the course letter grades.

Data Management:

We have three tables of data:

Students: We have 4 columns of data showing 180 students coming from different levels under different majors. Student id is the unique identifier or primary key, student name is the name of the students and student level and major are the categorical variables we have here.

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180 entries, 0 to 179
Data columns (total 4 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   student_id     180 non-null    int64
 1   student_name   180 non-null    object
 2   student_level  180 non-null    object
 3   major          180 non-null    object
dtypes: int64(1), object(3)
memory usage: 5.8+ KB
```

Assignments: Here we have the assignment table including 20 rows and 3 columns. PK is assignment id, assignment category is the categorical variable includes “Homework", "Exam” and “Quiz”. Assignment value is the total points the assignment is worth.

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 3 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   assignment_id       20 non-null     int64
 1   assignment_category  20 non-null     object
 2   assignment_value    20 non-null     int64
dtypes: int64(2), object(1)
memory usage: 608.0+ bytes
```

Grades:

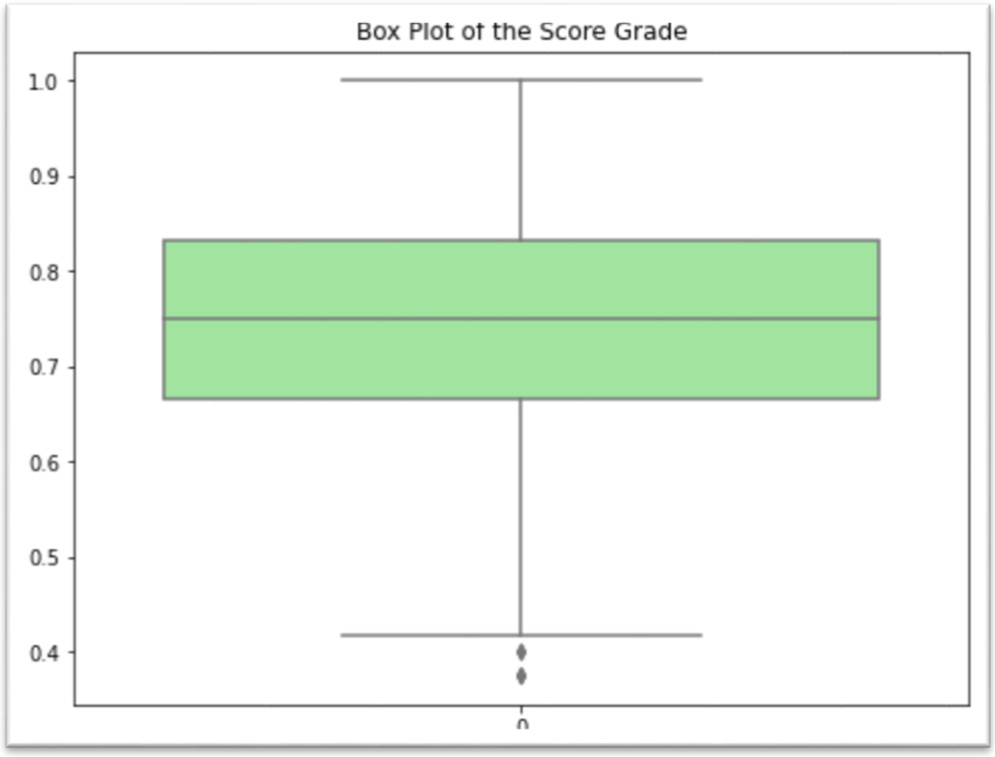
As indicated, there are 3600 rows of data and 3 columns. This table shows the grade for each student and for each single assignment id.

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3600 entries, 0 to 3599
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   student_id     3600 non-null   int64
 1   assignment_id  3600 non-null   int64
 2   numeric_grade  3600 non-null   int64
dtypes: int64(3)
memory usage: 84.5 KB
```

- Also, it is worth mentioning that there is no null value, and duplicated record in all three tables.

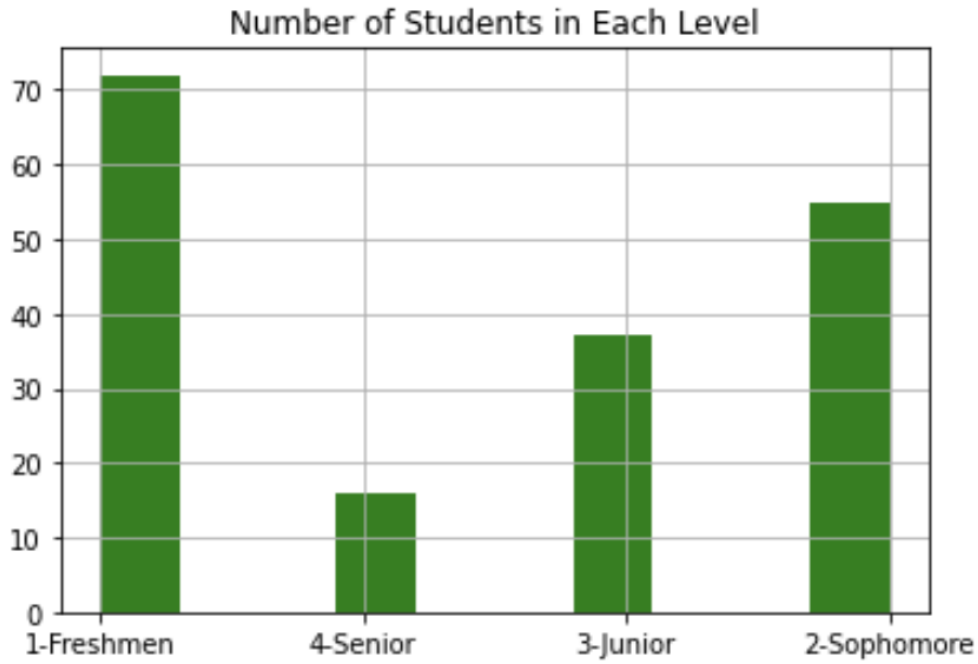
Outlier Detection: I used boxplot to show the potential outliers. As indicated below, there are scarce outlier but as those outliers are not influential outliers, we can ignore them.

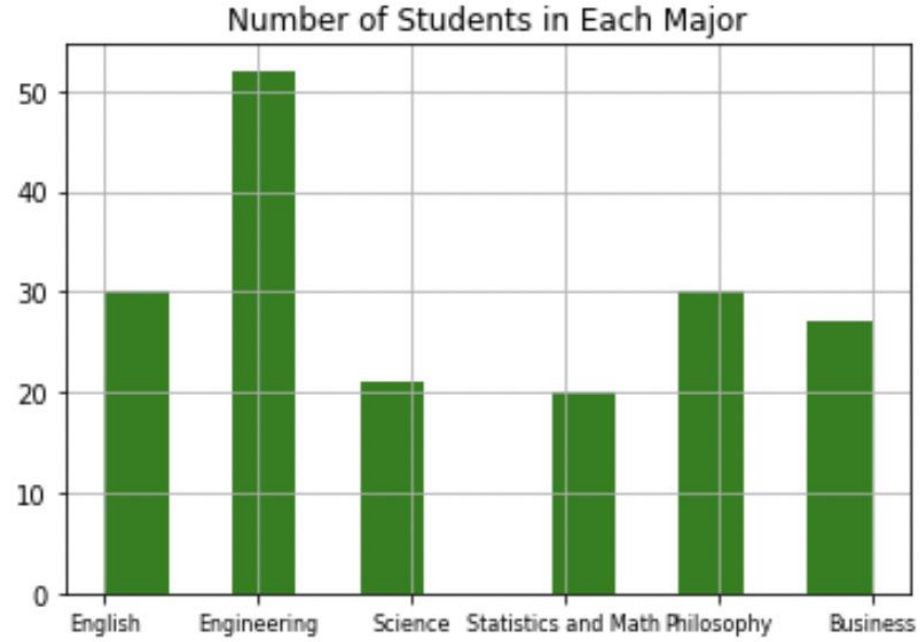


	student_id	student_name	student_level	major	assignment_id	numeric_grade	assignment_category	assignment_value	score_grade
152	514068	Nicholas	1-Freshmen	Business	13	3	Homework	8	0.375
348	188048	Nicole	1-Freshmen	Engineering	9	4	Homework	10	0.400
1544	996584	Eric	3-Junior	Engineering	5	3	Homework	8	0.375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

Below histograms provide us with good understanding of the number of students is each level and major. It clearly shows which major are more popular and how the students are distributed among the different levels. We do this by grouping data based on specific columns.





And the Number of Students in each Level/Major is as below:

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

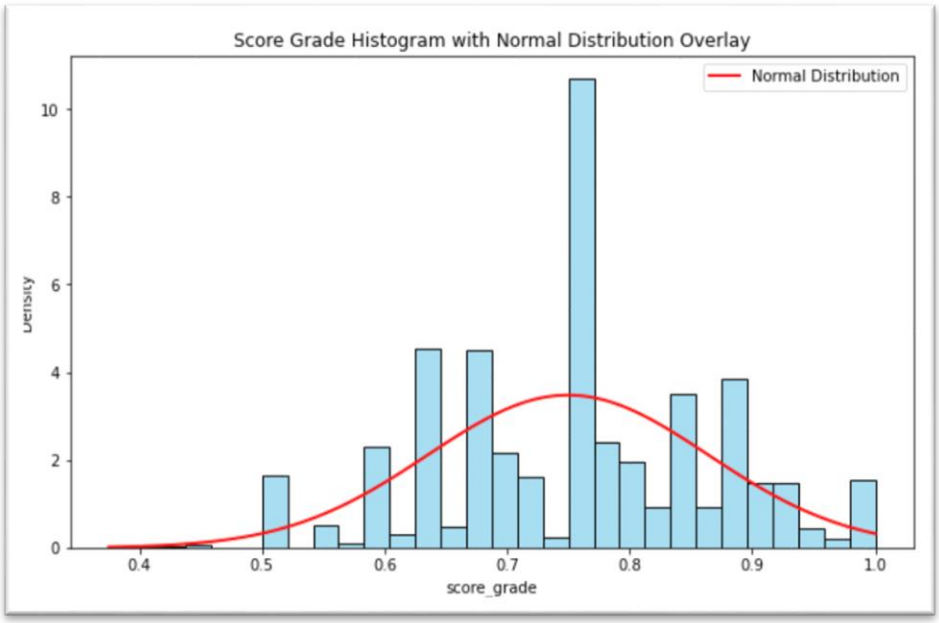
Statistical Analysis:

In this step I am going to calculate the numeric score for each single assignment that is calculated by dividing grade by the assignment value. To do so, I need to first join all three tables and then create this calculated field in our new table.

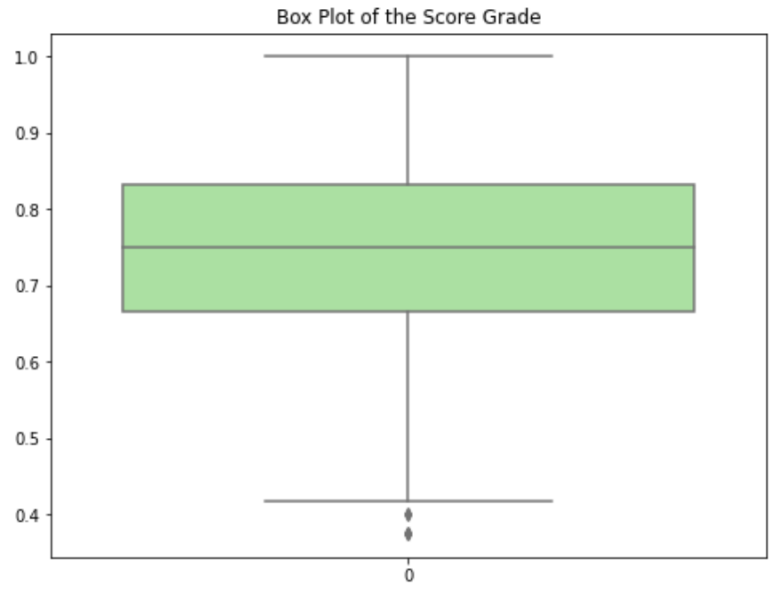
Score Grade Distribution:

In the below histogram, we clearly see how the score grade is distributed. Is it Normally distributed? From the graph, we would be considering it as a normal distribution as the mean

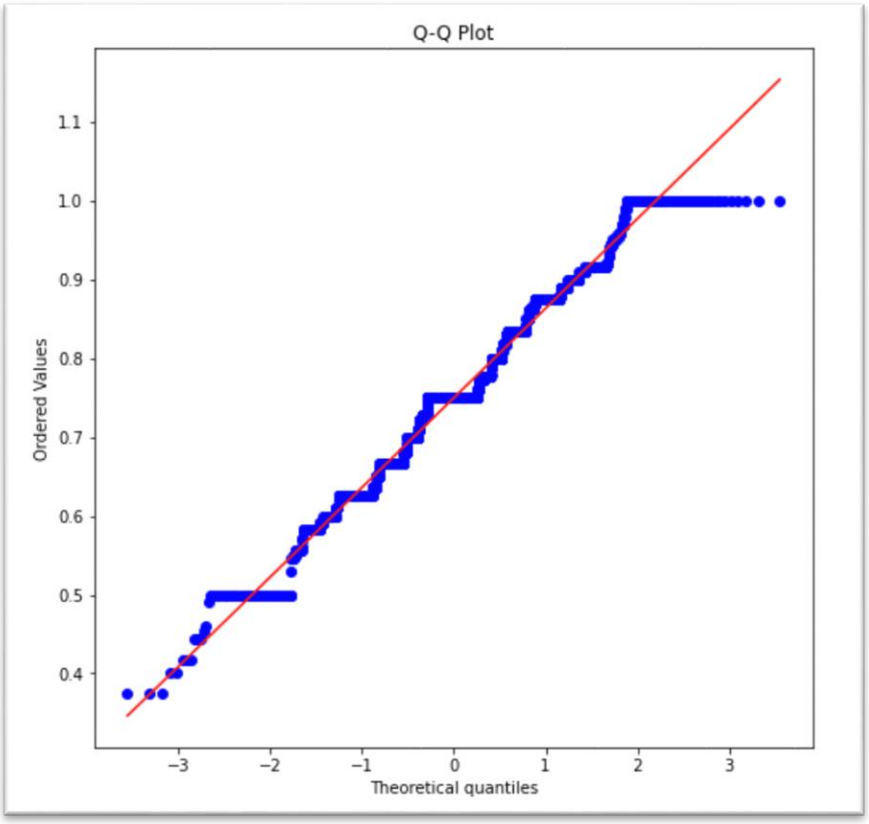
and median are almost equal and it seems symmetrical. It is slightly skewed but not heavily. Boxplot also provides some visual insights of the median and the range most of score grades are standing.




```
count    3600.000000
mean        0.749740
std         0.114752
min         0.375000
25%         0.666667
50%         0.750000
75%         0.833333
max         1.000000
Name: score_grade, dtype: float64
```



Q-Q Plot: A quantile-quantile (Q-Q) plot also compares data to a normal distribution. If the data points closely follow a straight line and there would no departures from the line, the data is likely normally distributed. In this QQ plot also we can see the almost all the data points are on the line. There are some departures at the end but it is not significant. This plot also indicates the score grades follow a normal distribution.



Shapiro Test: This is a statistical test to assess normality. It is helpful for small datasets. It is more reliable when the dataset is small.

H0 (Null Hypothesis) : data is normally distributed

H1(Alternative Hypothesis): data is not normally distributed

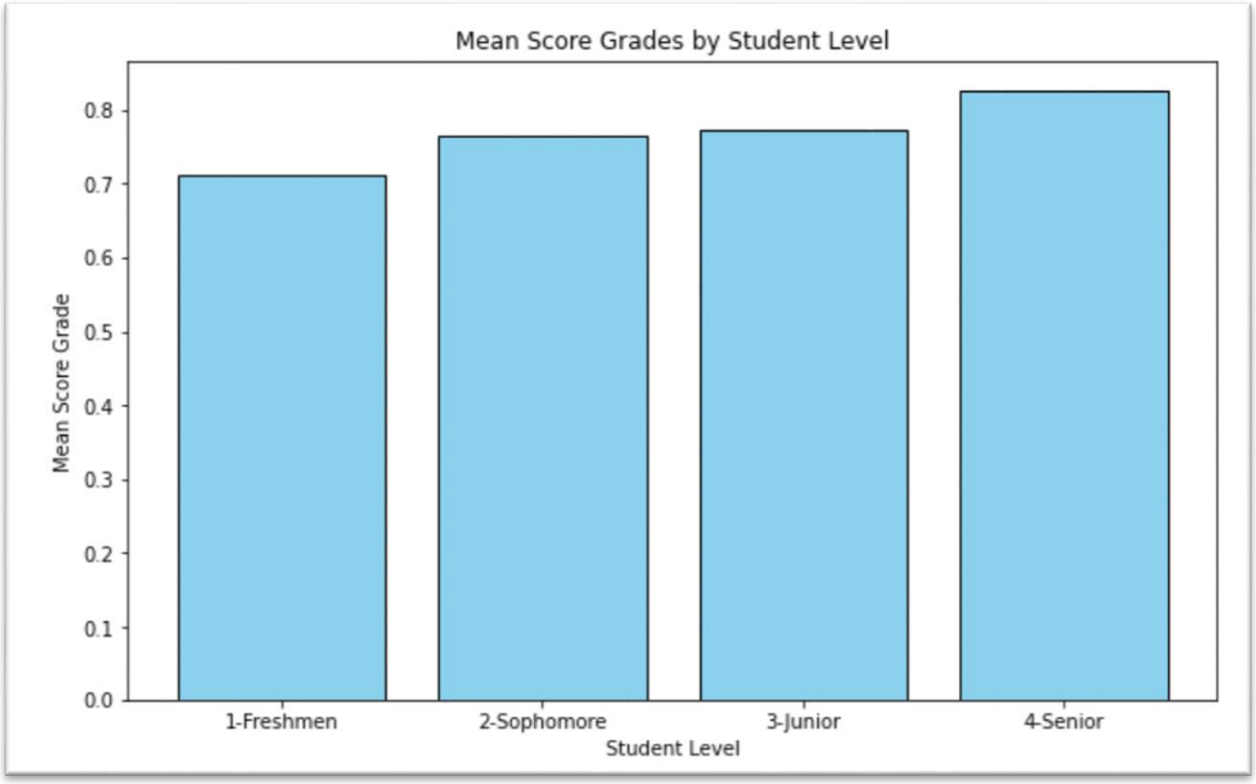
A significant p-value (typically less than 0.05) indicates that the data is not normally distributed. P-value is a probability of a result assuming that null hypothesis is true. This is why when the P-value is < 0.05 , we can reject the H0 and concluding that H1 can be true.

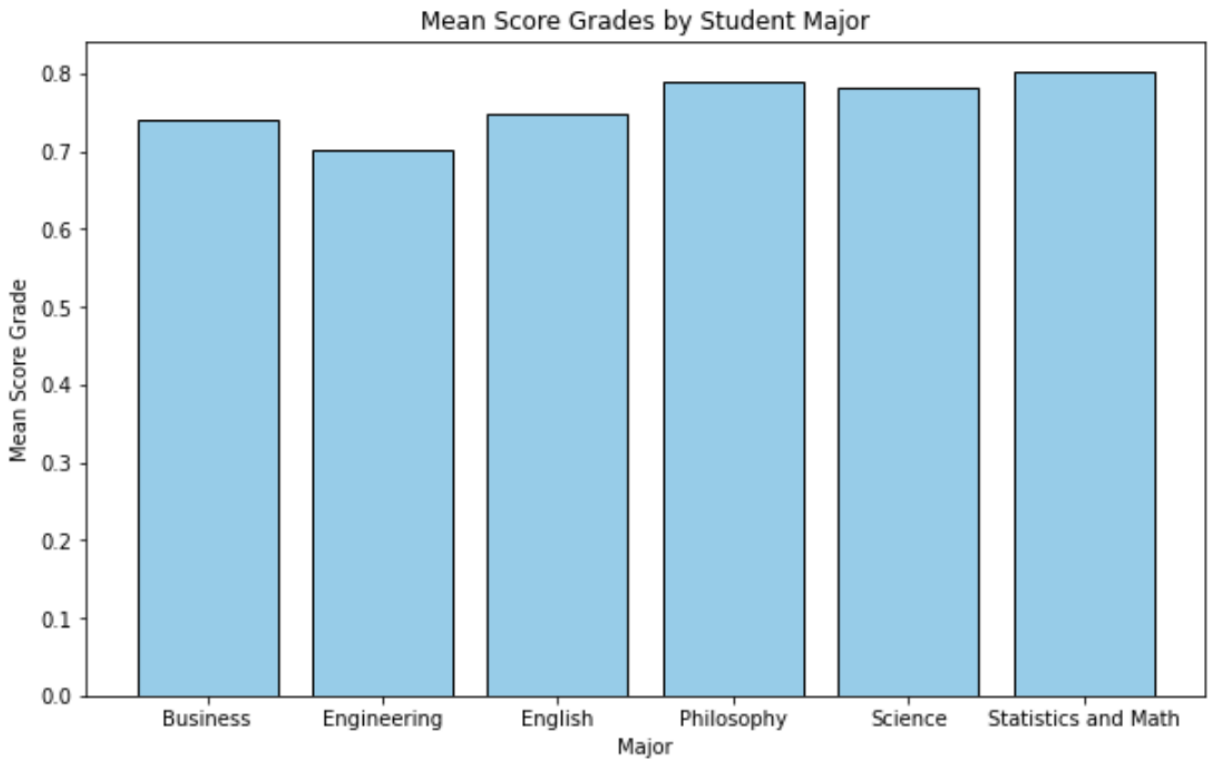
Shapiro-Wilk Test Statistic: 0.9825350642204285
p-value: 1.1167591774461533e-20

As calculated above, p-value is really small close to zero. So, We reject the null hypothesis and conclude that based on shapiro test , data is not normally distributed. However, Shapiro test is not always a good way to check normality for large datasets as mentioned earlier.

Mean Score Grades:

Now, I am going to provide some insights of data related to mean scores of students in each level and major:





- As indicated in the above chart, the Senior level has the highest mean score grade which is 0.82 out of 1 and Freshmen level has the lowest mean score grade equal to 0.71. It seems higher levels have higher mean score.
- The other bar chart also shows the mean score grade for statistics is highest and Engineering major has the lowest mean score grade.

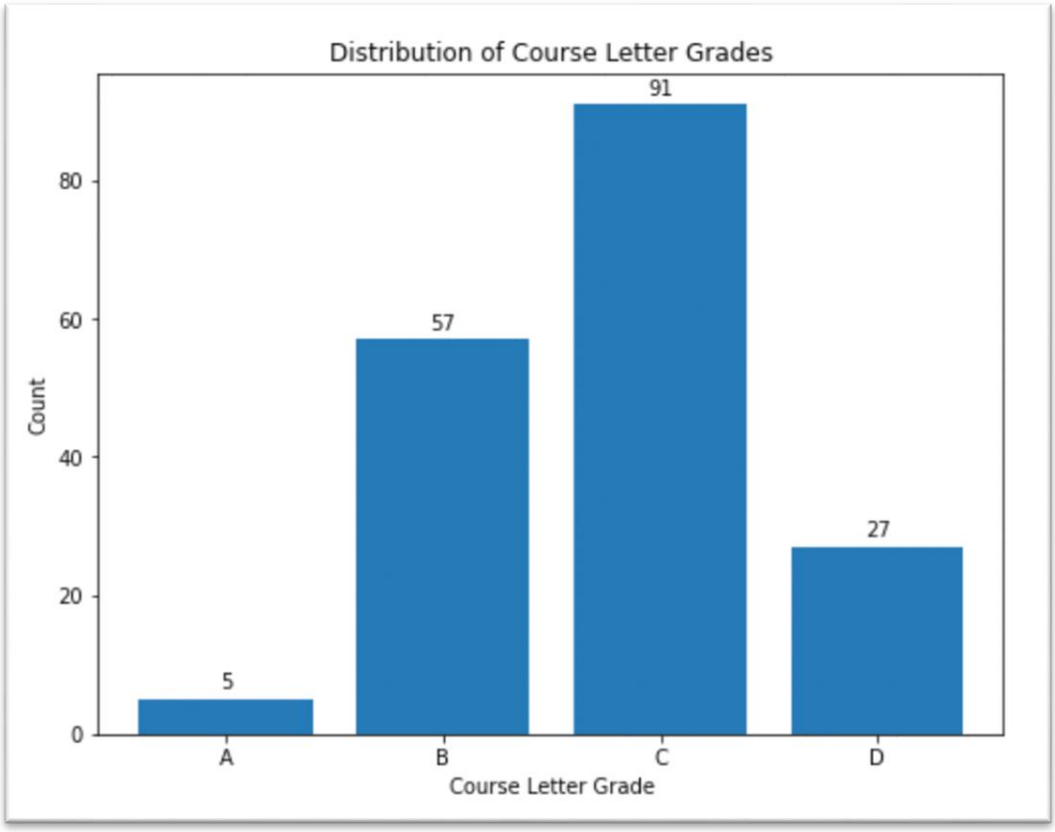
Hypothesis Testing :

Now, I am going to answer if there is a statistically significant difference between the mean score of freshmen and seniors? If so, is the difference meaningful?

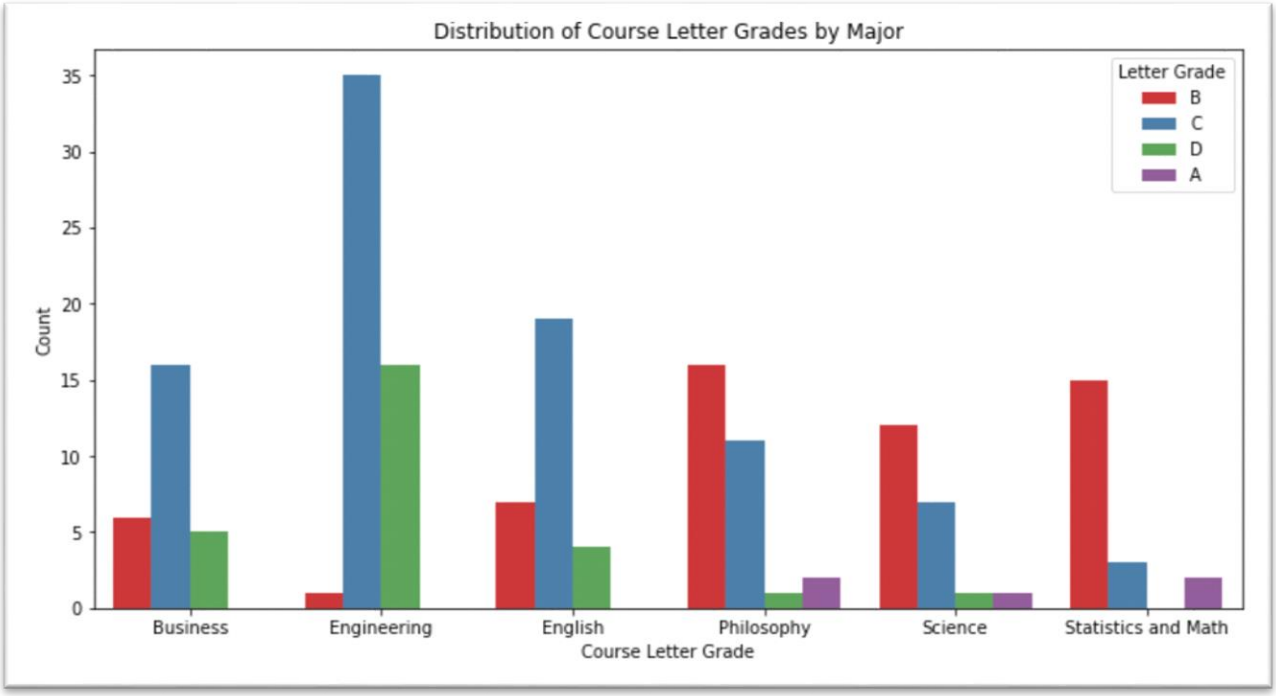
Answer: For checking the significance difference between the means score of freshmen and seniors, we use hypothesis testing : Null Hypothesis (H0): There is no significant difference in the mean score between freshmen and seniors , Alternative Hypothesis (H1): There is a significant difference in the mean score between freshmen and seniors. Then, we perform the T-Test and to get the test statistics and P-value. As indicated below: P-value = 7.896304165134023e-59 is less than the significance level (alpha=0.05) .So, it indicated that we have enough evidence to reject the null hypothesis and conclude that there is significant difference in the mean score between freshmen and seniors. P-value is almost close to zero that it shows the distribution of the scores of freshmen and seniors are completely different and these are not coming from same distribution. Hence, we can say difference is meaningful and those are statistically different.

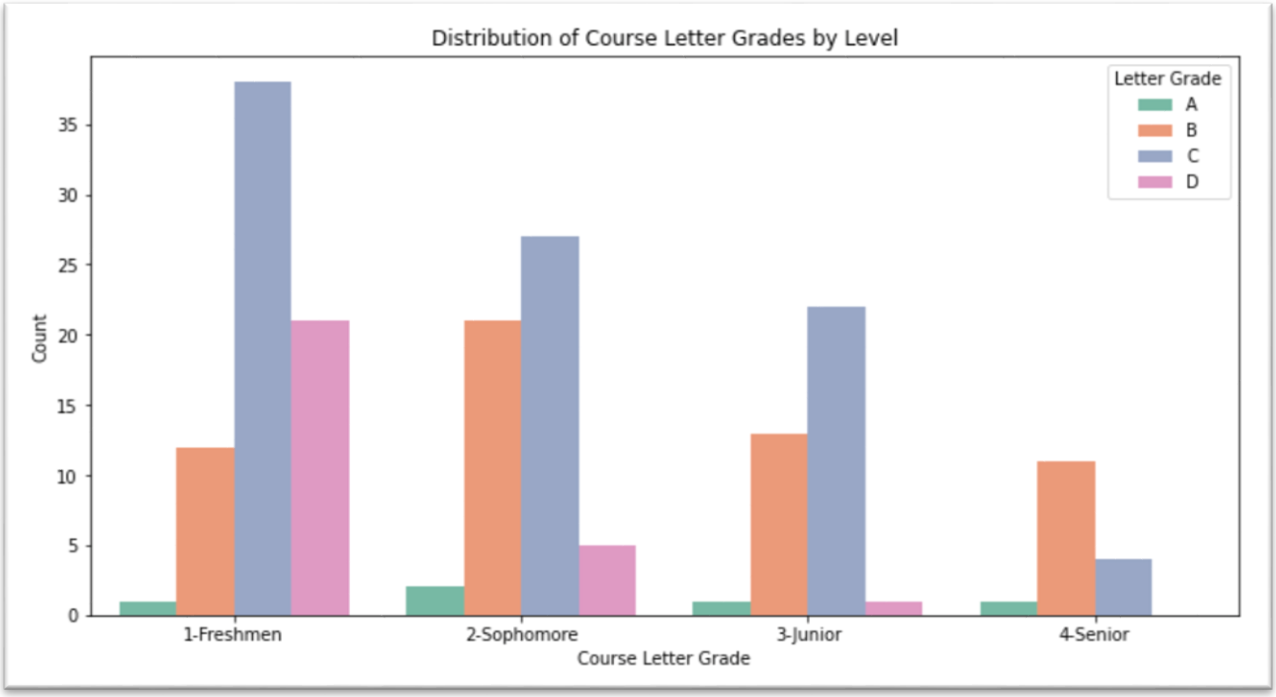
Course Letter Grades (A,B,C and D):

In this part I provided the course letter grades. First, I calculated the average of grades for exam, quiz and homework for each student and then after multiplying them by their weights which are 60% - for Exam, 15% - Quiz and 25% - Homework , I got to the weighted average score grades. Then, If the weighted average score grade > 90, it is considered as A, else >80 it is B, else >70 it is C , else >60 it is D , otherwise it is F.



Analysis of Letter Grade Distribution By Majors and Levels:





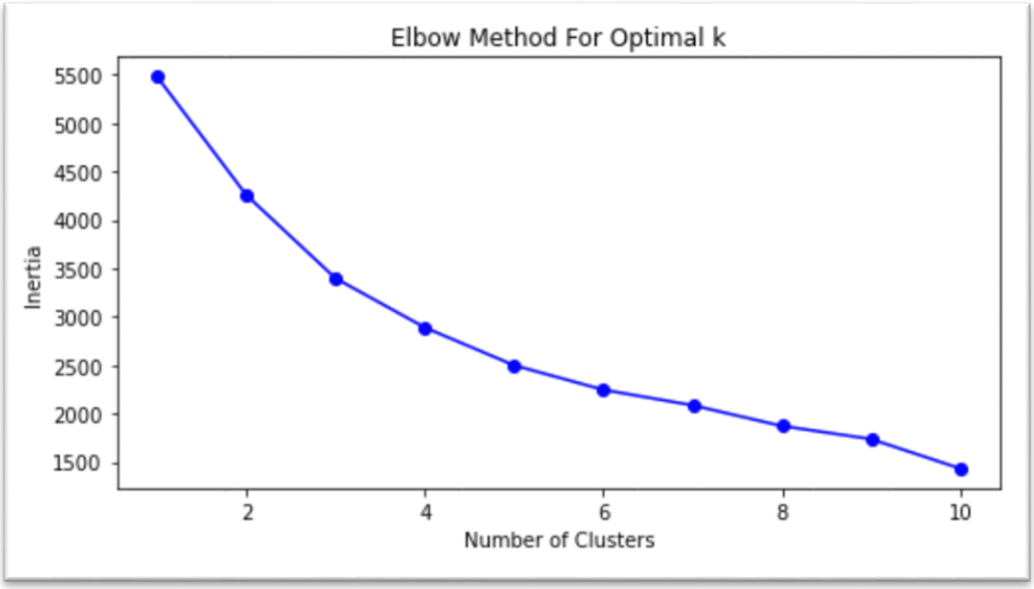
- As indicated above, it seems Engineering, English and Business majors have the highest number of C, while we can't see any A there. On the other hand, Philosophy, Science and Statistics indicate more B, and some A letter grades can be seen as well.

- Also, by checking the distribution of letter grades by level, as the levels grows, the letter grades get better. We can see clearly that the number of C is getting lower in Senior level. While Freshmen has more C and D.

Cluster Analysis:

- **Clustering**: This is an unsupervised learning method that we use for grouping datapoints based on their similarities. This is obvious that there is no response variable in this approach and we don't need that. I Perform clustering to group students into clusters based on their grades, level, and major. This can help identify patterns or segments within the student population.

  K-Mean Algorithm works based on **inertia** which is a measure of how well the clusters are formed. It is defined as the sum of squared distances between each data point and the centroid of the cluster to which it is assigned. Generally, lower inertia is desirable as it suggests that the clusters are well-formed. Here we use Elbow Method to find the optimal number of clusters and as indicated in the graph, the decrease in inertia is significant until K=4.



Summary:

Ending up with k=4 clusters means that the data naturally divides into four groups. These groups represent different pattern or segments within data, and analyzing them can provide insights that help in decision-making or further analysis. For example, these 4 clusters can be high-grade or low-grade students in specific majors or levels. So, there might be required to provide some helpful resources to low-grade students in those majors/levels.

Future Works:

I can develop an automated report to provide these insights and visualizations.

To set up an automated report system for processing data across multiple courses and calculating final grades, we can create a Python pipeline. First, import CSV files into the pipeline. Then, perform data preprocessing tasks such as checking for missing values, duplicated entries, and outliers. Next, create a Python class that takes the dataset's name as input and conducts various aggregations to calculate students' final grades based on predefined criteria. Additionally, the class can generate visualizations, such as histograms or pie charts, to illustrate the distribution of grades for different assignments. This pipeline offers a modular and adaptable approach to efficiently handle diverse datasets, grading criteria, and reporting requirements, making it easier to scale and generalize the workflow for various courses and data sources.