FIU | Knight Foundation School of Computing and Information Sciences

Fundamentals of Data Science

Online X In-Person Programming Class

(Final Project Sample)

Professor Gregory Murad Reis, PhD
Miami, FL

1. Data Exploration

1) Data Exploration

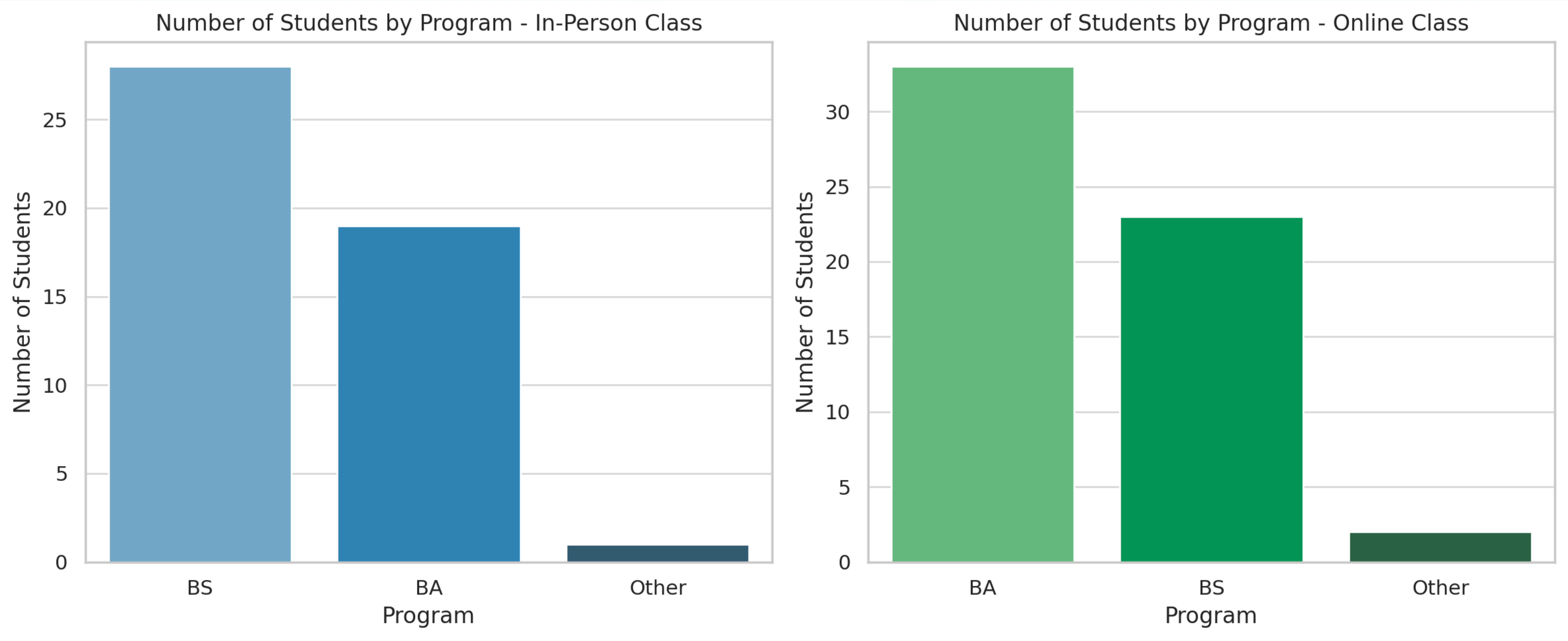
- This project has the goal of analyzing two datasets containing grades of students and their program (BA, BS or other) and then draw some conclusions on the difference between the grades in online and in-person modalities.

- Two datasets where each dataset contains three columns: 'Id', 'Grade', and ‘Program'.

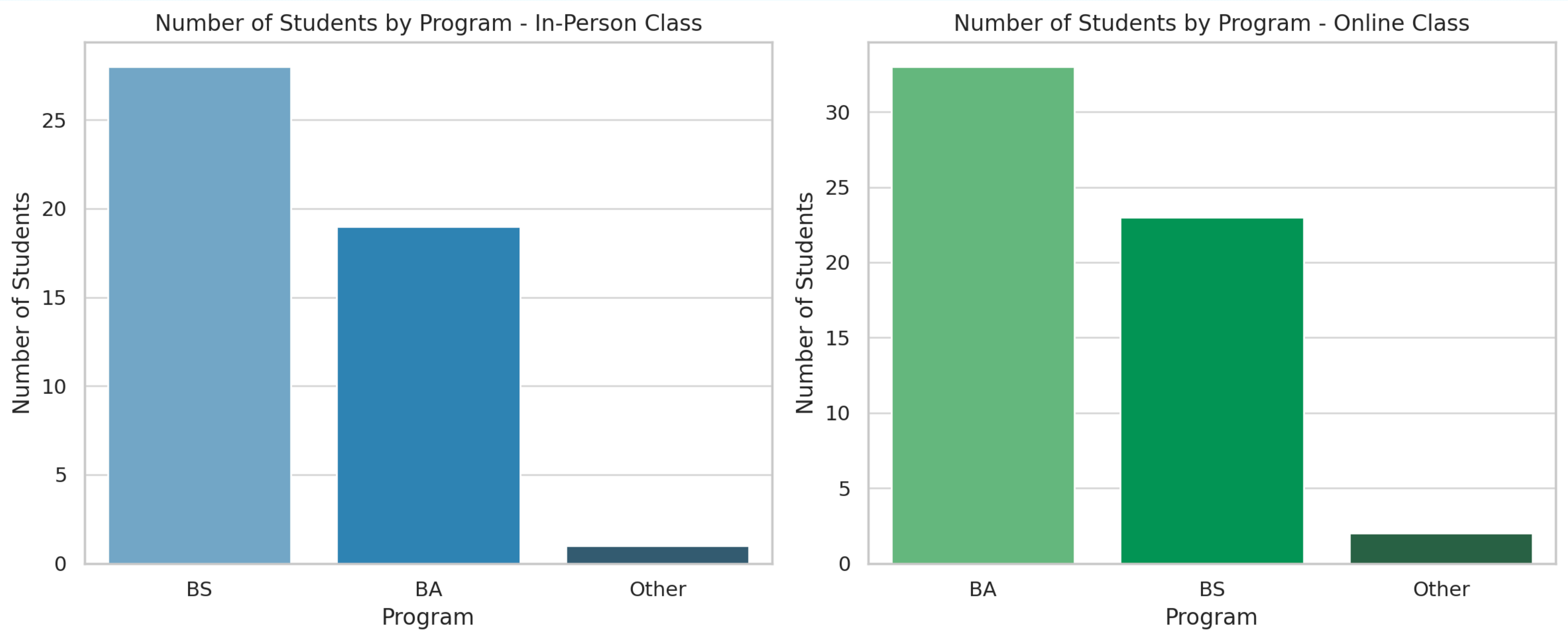
- The 'Program' column indicates whether the program is BA or BS in Computer Science, or another program track.

- Data was collected in Fall 2023 during the midterm exam for both modalities for the same programming course. One dataset for the in-person class and the other dataset for the online class.

1) Data Exploration



1) Data Exploration



In-Person Class

- The chart depicts the number of students in each program, with the BS program having the highest enrollment, followed by the BA program. The 'Other' category has very few students.

Online Class

- Similar to the in-person class, the BS program shows higher enrollment compared to the BA program. The 'Other' category has a smaller number of students.

1) Data Exploration

In-Person Class

- **Count**: 48 students
- **Mean (Average) Grade**: 75.81
- **Standard Deviation**: 16.34
- **Minimum Grade**: 31.00
- **25th Percentile**: 69.00
- **Median (50th Percentile)**: 80.00
- **75th Percentile**: 87.25
- **Maximum Grade**: 99.00

Online Class

- **Count**: 58 students
- **Mean (Average) Grade**: 85.51
- **Standard Deviation**: 14.57
- **Minimum Grade**: 50.42
- **25th Percentile**: 75.04
- **Median (50th Percentile)**: 84.25
- **75th Percentile**: 96.76
- **Maximum Grade**: 110.25

1) Data Exploration

In-Person Class

- **Count**: 48 students
- **Mean (Average) Grade**: 75.81
- **Standard Deviation**: 16.34
- **Minimum Grade**: 31.00
- **25th Percentile**: 69.00
- **Median (50th Percentile)**: 80.00
- **75th Percentile**: 87.25
- **Maximum Grade**: 99.00

Online Class

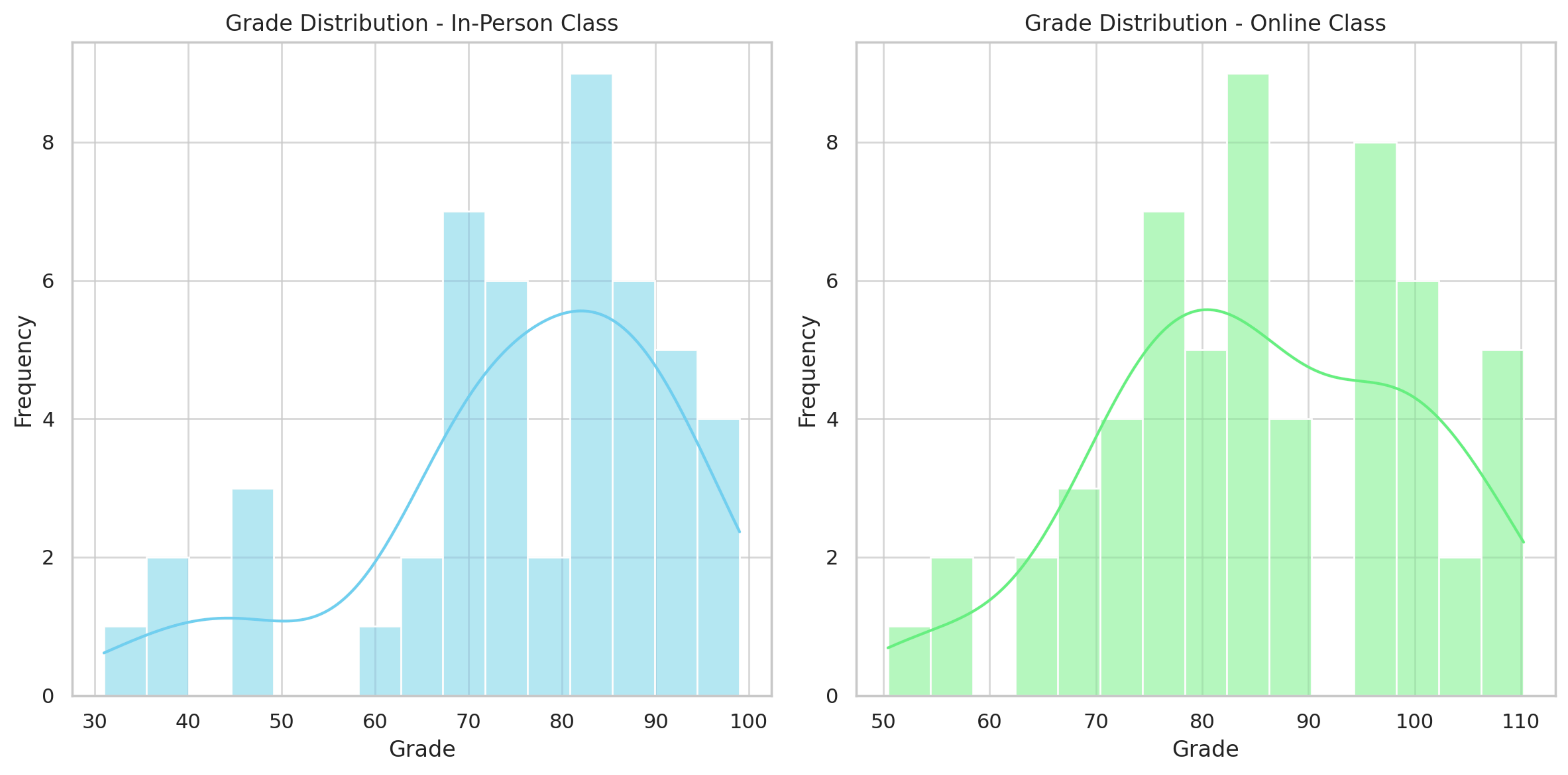
- **Count**: 58 students
- **Mean (Average) Grade**: 85.51
- **Standard Deviation**: 14.57
- **Minimum Grade**: 50.42
- **25th Percentile**: 75.04
- **Median (50th Percentile)**: 84.25
- **75th Percentile**: 96.76
- **Maximum Grade**: 110.25

From these statistics, we observe that:

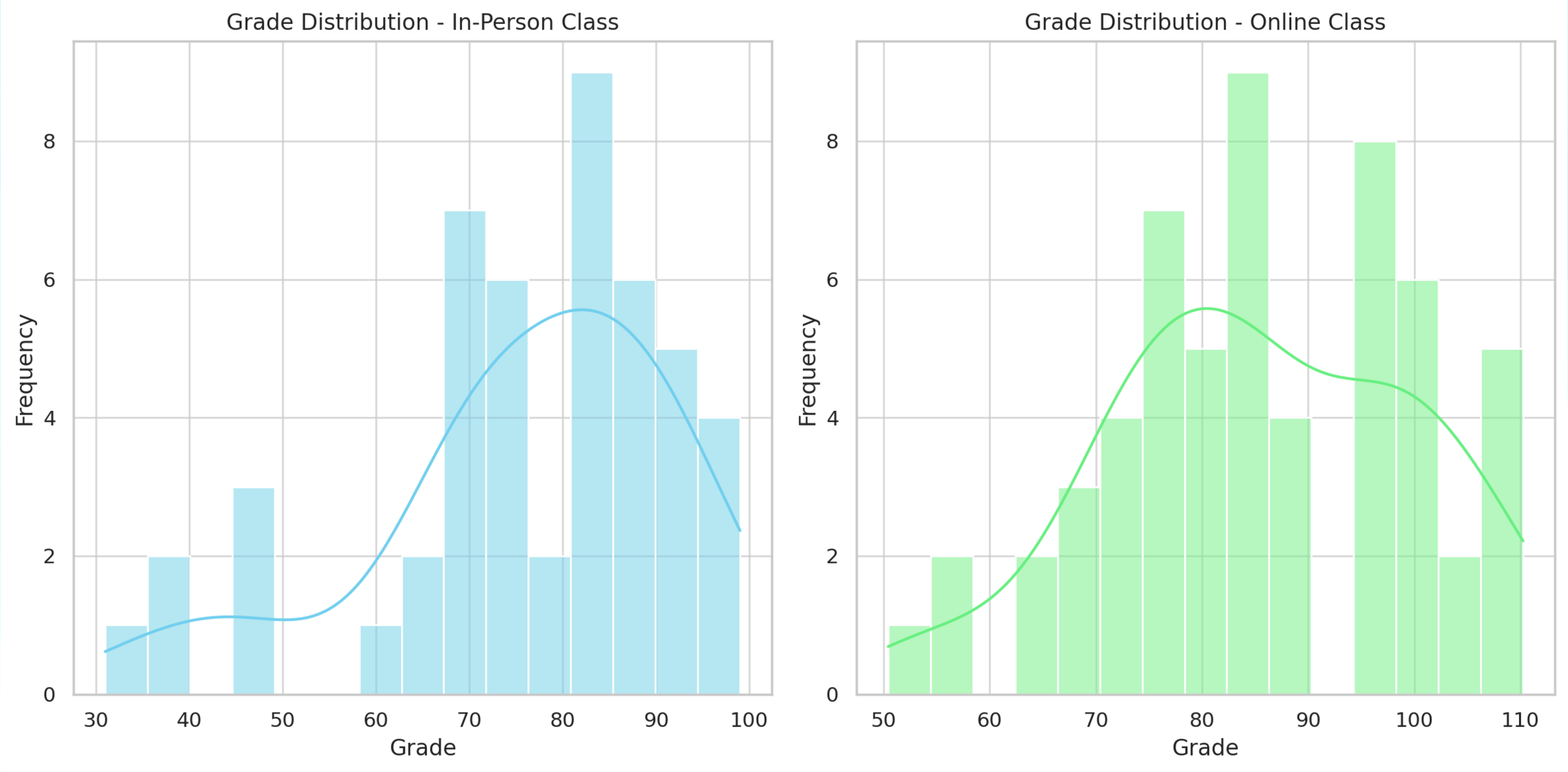
1) The online class has a higher average grade and a slightly lower standard deviation compared to the in-person class.

2) The maximum grade in the online class exceeds the typical 100-point scale, indicating a possibility of extra credit or a different grading scale being used.

1) Data Exploration



1) Data Exploration



- **In-Person Class**: The distribution is somewhat left-skewed, with a concentration of grades towards the higher end (around 80-90). There are also a few lower grades, as seen by the long tail on the left.
- **Online Class**: This distribution appears more symmetric and centered around 80-90, with a smoother curve. The presence of grades above 100 indicates a different grading scale or extra credit options in the online class.

1) Data Exploration

In-Person Class

BA Program

- **Count**: 19 students
- **Mean Grade**: 69.47
- **Standard Deviation**: 18.86
- **Minimum Grade**: 31.00
- **25th Percentile**: 61.50
- **Median**: 69.00
- **75th Percentile**: 84.00
- **Maximum Grade**: 99.00

BS Program

- **Count**: 28 students
- **Mean Grade**: 81.39
- **Standard Deviation**: 10.93
- **Minimum Grade**: 47.00
- **25th Percentile**: 74.00
- **Median**: 82.50
- **75th Percentile**: 89.50
- **Maximum Grade**: 98.00

Other Program

- **Count**: 1 student
- **Grade**: 40.00

1) Data Exploration

Online Class

BA Program

- **Count**: 33 students
- **Mean Grade**: 85.16
- **Standard Deviation**: 12.81
- **Minimum Grade**: 65.33
- **25th Percentile**: 75.00
- **Median**: 83.75
- **75th Percentile**: 96.75
- **Maximum Grade**: 110.25

BS Program

- **Count**: 23 students
- **Mean Grade**: 87.13
- **Standard Deviation**: 16.31
- **Minimum Grade**: 50.42
- **25th Percentile**: 79.25
- **Median**: 85.83
- **75th Percentile**: 99.34
- **Maximum Grade**: 109.50

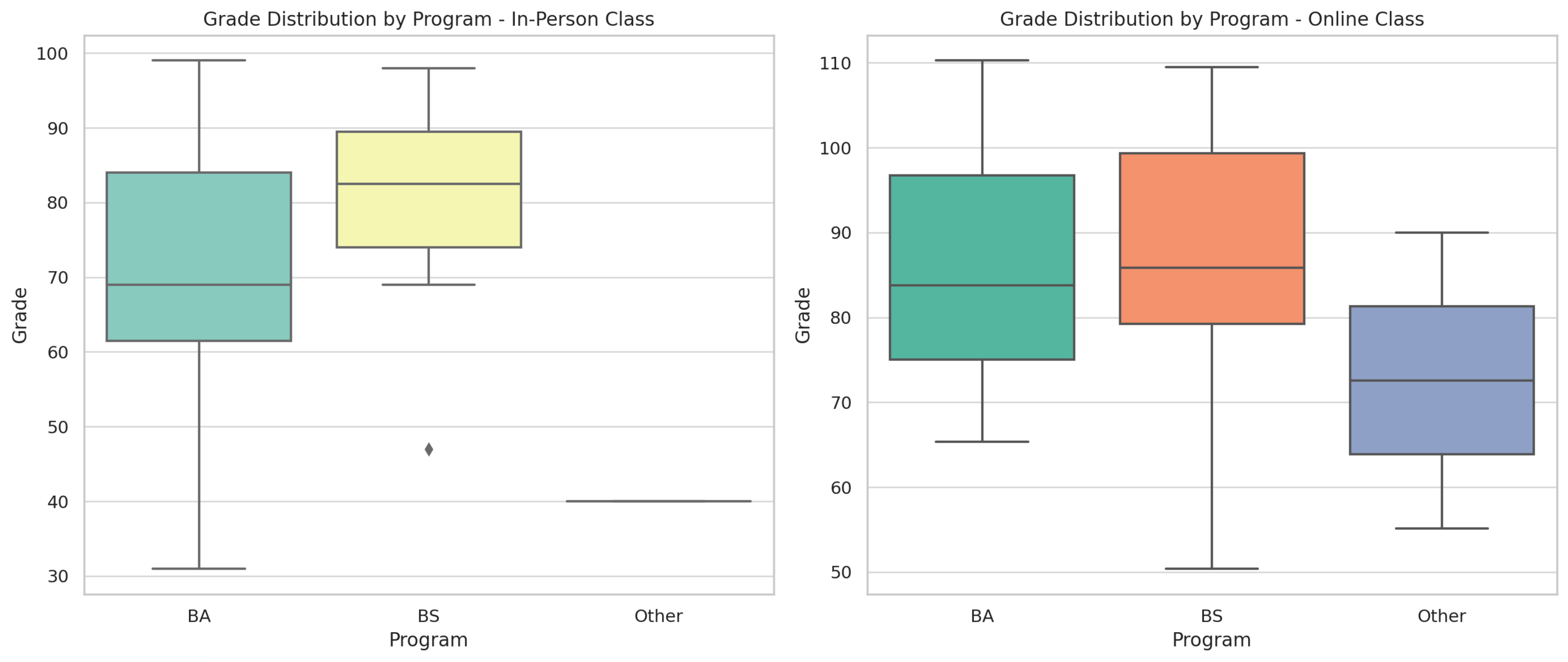
Other Program

- **Count**: 2 students
- **Mean Grade**: 72.59
- **Standard Deviation**: 24.63
- **Minimum Grade**: 55.17
- **Maximum Grade**: 90.00

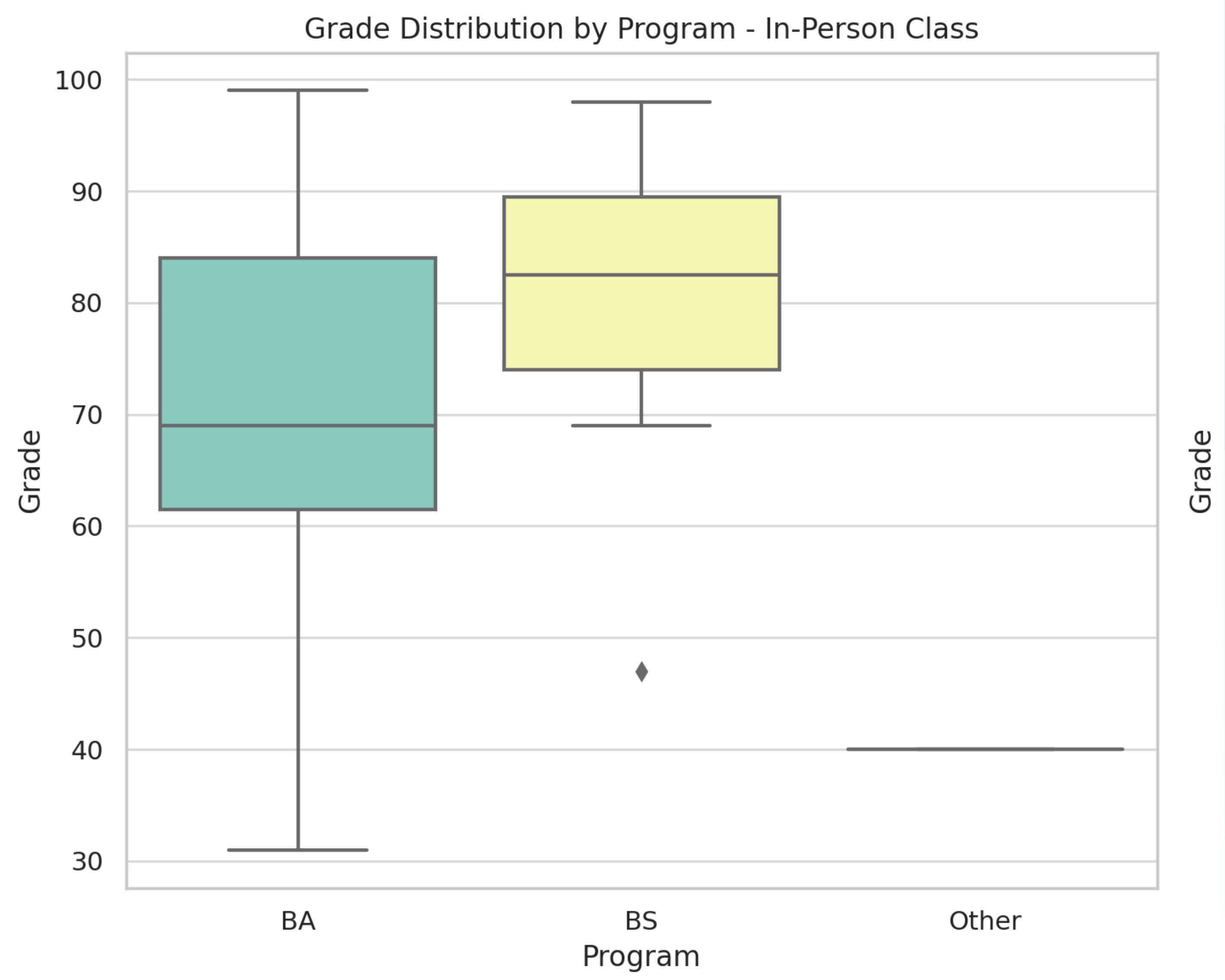
1) Data Exploration

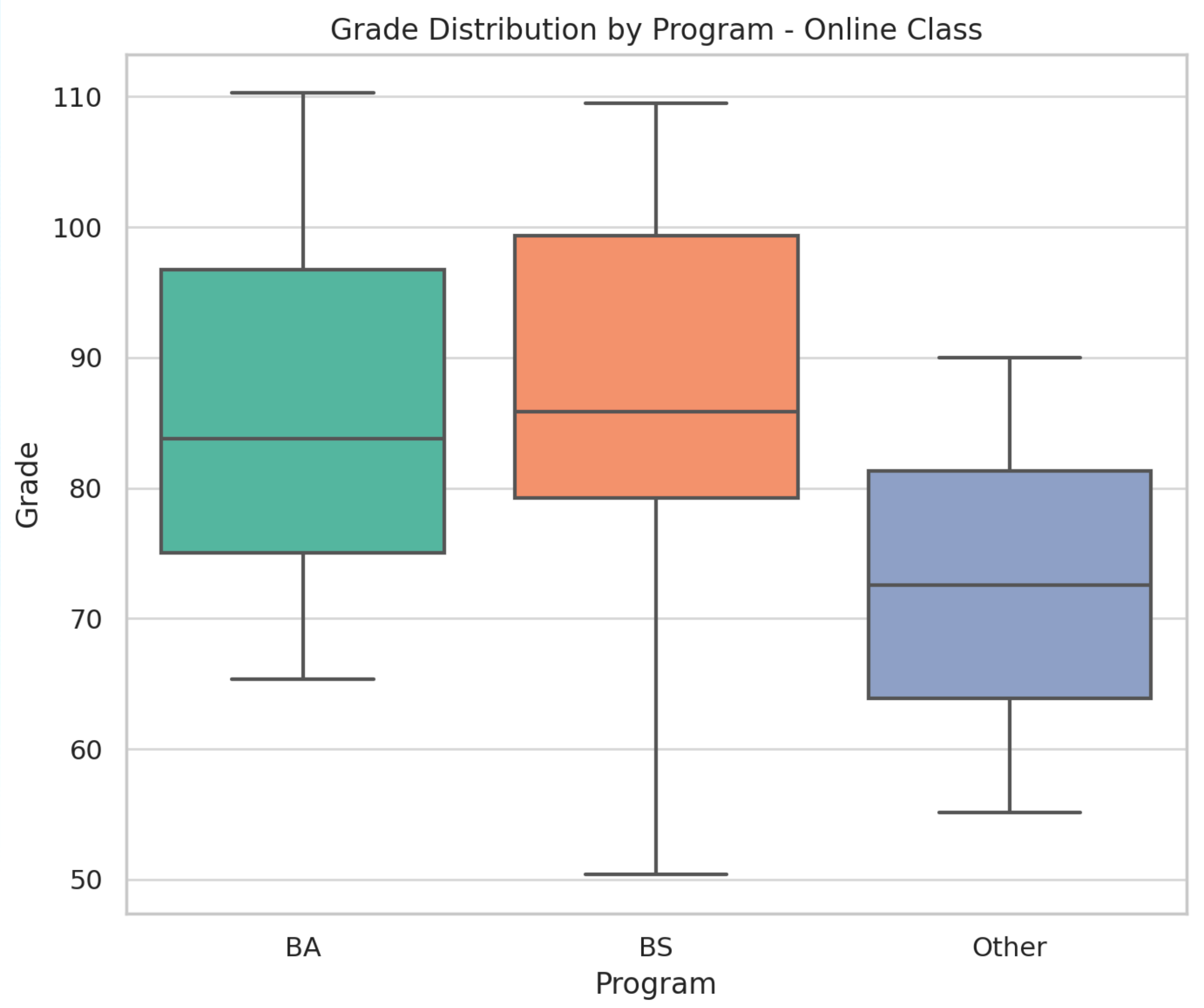
- From these statistics, we can observe that in both classes, the BS program students tend to have higher average grades compared to the BA program students.
- The 'Other' category has very few students, so it's hard to draw definitive conclusions for that group.

1) Data Exploration



1) Data Exploration





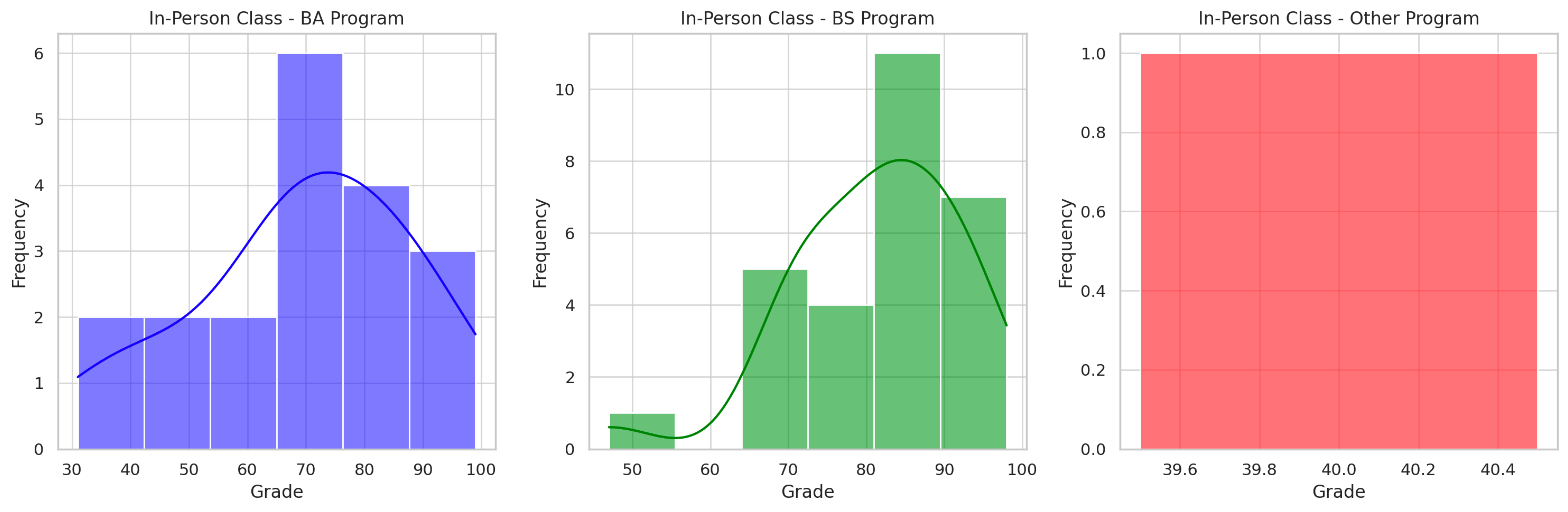
In-Person Class

- **BA Program**: Shows a wider range of grades with a lower median compared to the BS program.
- **BS Program**: Grades are more concentrated in the higher range with a higher median.
- **Other Program**: Only one data point, so it appears as a single line.

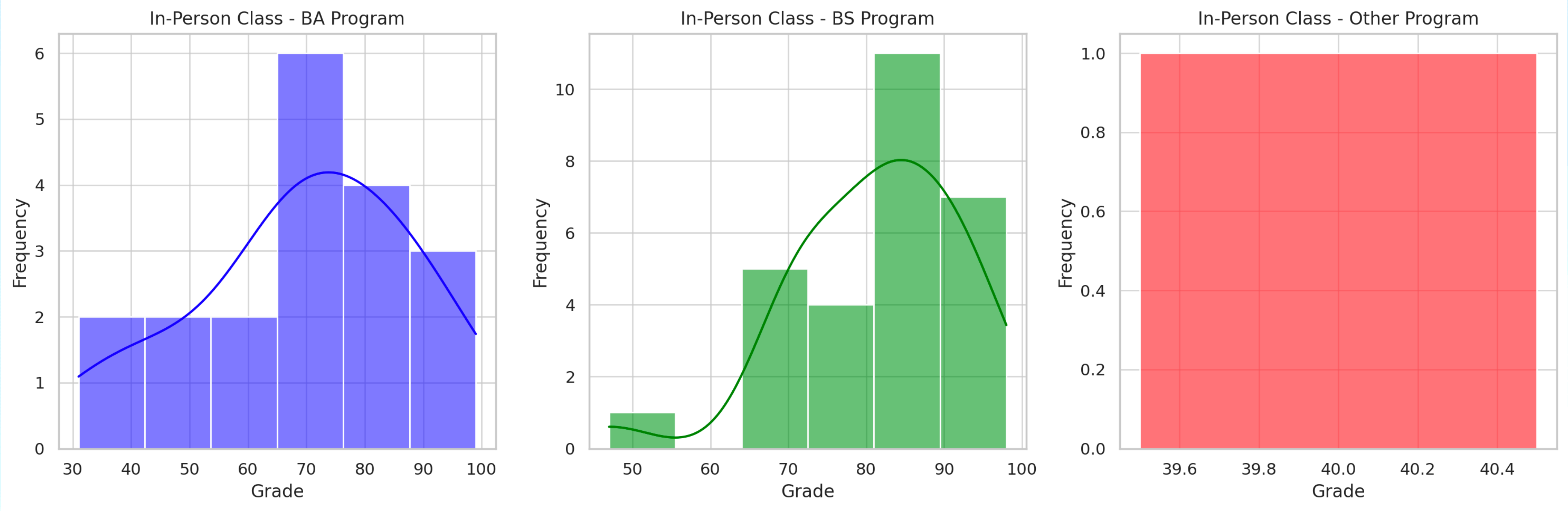
Online Class

- **BA Program**: Similar to the in-person class, the range is wide, but the median is higher.
- **BS Program**: Higher median grade and a slightly wider interquartile range than the BA program.
- **Other Program**: With only two data points, the variability is shown by the range between them.

1) Data Exploration



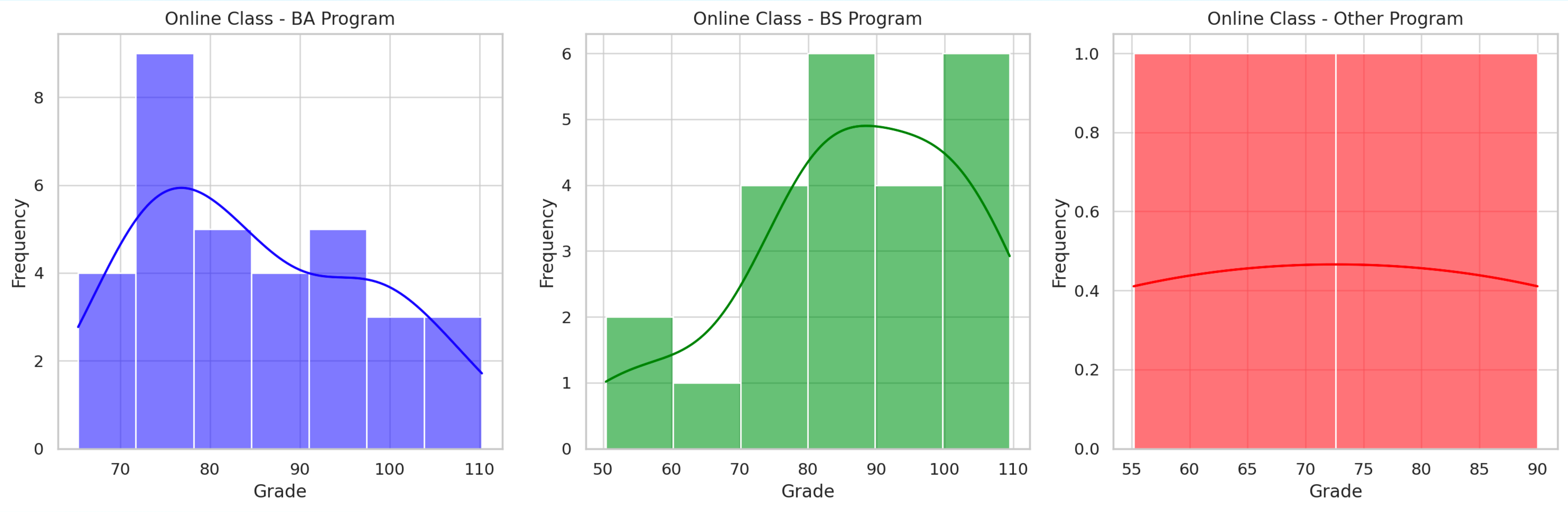
1) Data Exploration



In-Person Class

- **BA Program**: The distribution shows a wide range of grades, with a concentration in the middle.
- **BS Program**: The grades are more concentrated in the higher range.
- **Other Program**: Given the limited data (only one student), the plot shows a single value.

1) Data Exploration

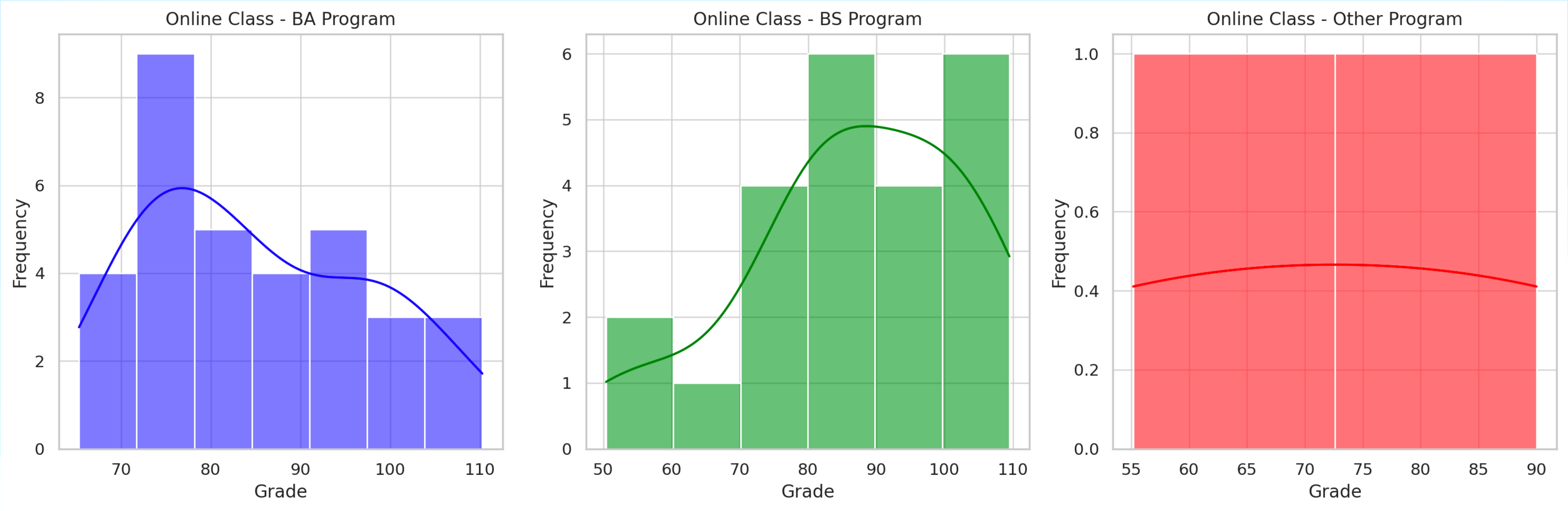


Online Class - BA Program

Online Class - BS Program

Online Class - Other Program

1) Data Exploration



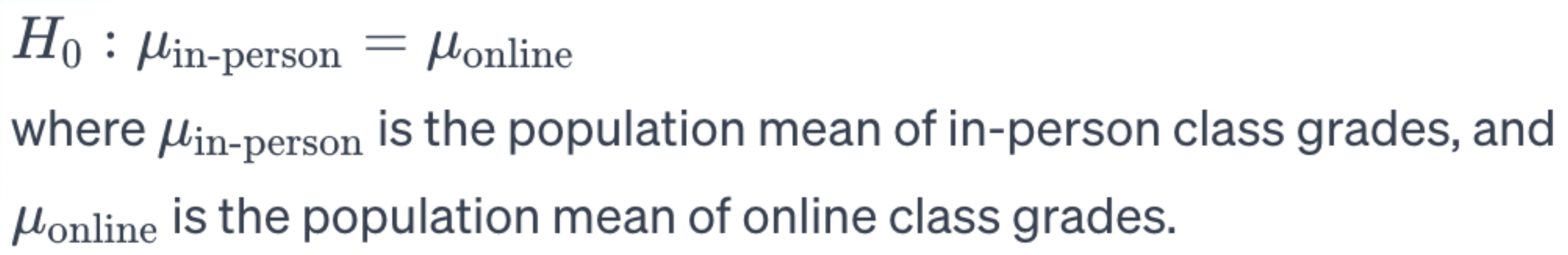
Online Class

- **BA Program**: The distribution is more centered with a peak around the middle-high range of grades.
- **BS Program**: This distribution also shows a concentration in the higher grade range, similar to the in-person class.
- **Other Program**: With only two data points, the plot shows a simple distribution between these values.

2. Formulate Hypotheses

2) Formulate Hypotheses

- To determine if there is a statistically significant difference between the average grades of the in-person and online classes, we can perform a two-sample t-test. This test is appropriate when comparing the means of two independent samples (in this case, the two classes) to see if they could be from the same population.
- **Null Hypothesis (H0)**
  - The null hypothesis states that there is no significant difference in the average grades between the in-person and online classes. Mathematically, it can be expressed as:



$$H_0 : \mu_{\text{in-person}} = \mu_{\text{online}}$$

where $\mu_{\text{in-person}}$ is the population mean of in-person class grades, and

μ_{online} is the population mean of online class grades.

2) Formulate Hypotheses

- **Alternative Hypothesis (H1)**
    - The alternative hypothesis states that there is a significant difference in the average grades between the two classes. It does not specify the direction of the difference. It can be expressed as:

$$H_1 : \mu_{\text{in-person}} \neq \mu_{\text{online}}$$

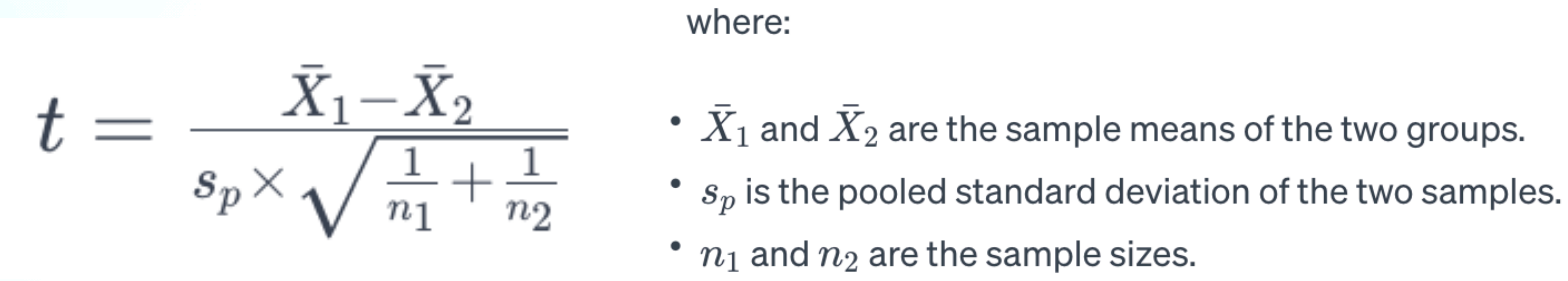
2) Formulate Hypotheses

- In our test, rejecting the null hypothesis (H0) would mean accepting that there is a statistically significant difference in the average grades between the in-person and online classes.
- The direction of this difference (whether in-person is higher or lower than online) is indicated by the sign of the t-statistic.
- In our case, since the t-statistic was negative, it suggests that the mean grade of the in-person class is lower than that of the online class.

3. Conduct the t-test

3) Conduct the t-test

- Before performing the test, we need to consider two key assumptions of the t-test:
  - **Normality**: The t-test assumes that the data in each group is roughly normally distributed.
  - **Homogeneity of variances**: The test also assumes that the two groups have similar variances.
- To perform a two-sample t-test, we need to calculate the t-statistic, which is a measure of the difference between the two sample means relative to the variation in the samples. The formula for the t-statistic in a two-sample t-test (assuming equal variances) is:



$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where:

- $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the two groups.
- s_p is the pooled standard deviation of the two samples.
- n_1 and n_2 are the sample sizes.

3) Conduct the t-test

- After calculating the t-statistic, we use it to calculate a p-value to determine if the difference in means is statistically significant.

In-Person Class

- **Mean Grade**: 75.81
- **Standard Deviation**: 16.34
- **Sample Size**: 48

Online Class

- **Mean Grade**: 85.51
- **Standard Deviation**: 14.57
- **Sample Size**: 58

T-Statistic

- **Calculated Value**: -3.2268

- The t-statistic value calculated (-3.2268) is used to determine the significance of the difference between the two sample means. A negative value indicates that the mean of the in-person class is lower than that of the online class.

3) Conduct the t-test

```
# Extracting necessary values for the formula
mean_inperson = inperson_data['Grade'].mean()
mean_online = online_data['Grade'].mean()
stdev_inperson = inperson_data['Grade'].std()
stdev_online = online_data['Grade'].std()
n_inperson = len(inperson_data)
n_online = len(online_data)

# Calculating the standard deviation
sp = ((n_inperson - 1) * stdev_inperson**2 + (n_online - 1) * stdev_online**2) / (n_inperson + n_online - 2)
sp = sp**0.5

# Calculating the t-statistic
t_statistic = (mean_inperson - mean_online) / (sp * ((1/n_inperson) + (1/n_online))**0.5)

# Displaying the calculated values
mean_inperson, mean_online, stdev_inperson, stdev_online, n_inperson, n_online, sp, t_statistic
```

4. Interpret the Results

4) Interpret the Results

- p-value: 0.0017
- The p-value of the two-sample t-test is approximately 0.0017, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a statistically significant difference between the average grades of the in-person and online classes.
- **This indicates that the observed difference in mean grades is unlikely to be due to chance.**

References

- https://blog.jetbrains.com/pycharm/2023/10/future-of-data-science/
- https://hbr.org/2023/09/4-skills-the-next-generation-of-data-scientists-needs-to-develop
- https://explodingtopics.com/blog/data-science-trends
- https://explodingtopics.com/topic/deep-fake
- https://www.nobledesktop.com/learn/data-science/what-to-learn-after-data-science