Redefining Success: A Comprehensive Analysis for Improving First-Semester Statistics Curriculum

Ariel Rosario, Jr.

2024-03-05

Summary:

NotReal University is actively working to improve the first-year statistics curriculum with the goal of increasing graduation rates within the major. As a necessary step before making adjustments, we plan to conduct an in-depth analysis to examine the pass rates for each course. This will help us gauge the chances of students passing all of their classes during their initial attempt.

Our preliminary observations reveal that the cumulative pass rate across all classes is currently at a concerning 18.13%. This data underscores the urgency and significance of revisiting our curriculum to foster better educational outcomes for our students.

Introduction:

NotReal University is actively working to enhance its first-year statistics curriculum with the goal of boosting graduation rates within the major. Before implementing any changes, it's essential to undertake a thorough analysis of the current pass rates for each class. This involves exploring the likelihood that students will successfully complete all of their classes on their first try. By doing so, we can identify potential areas of improvement and create strategies that foster student success.

In today's increasingly competitive educational landscape, monitoring student performance is more critical than ever. This report provides a comprehensive analysis of the pass rates for students enrolled in five specific courses during the Fall 2022 semester. The pass rate, a key indicator of student success, is defined in this context as achieving a score of 65 or higher.

The data for this report is sourced from the NotReal University private database, which meticulously tracks a variety of student data, including course grades. For the purpose of this analysis, the dataset includes the following fields: Student ID, and grades for Class 1 (Intro to Statistics), Class 2 (Intro to Probability), Class 3 (ELA 101), Class 4 (History), and Class 5 (Biology).

This report aims to:

1. Provide a clear and concise summary of student pass rates in the selected courses.
2. Identify trends or patterns that may emerge within and across these courses.
3. Serve as a valuable resource for the general public, to foster a better understanding of student performance within these courses during the Fall 2022 semester.

While the focus of this report is on pass rates, it is hoped that the findings may inform broader discussions about educational strategies and policies that impact student success.

```r
library(readxl)
library(tidyverse)
pass_rate <- read_excel("/Users/ariel_rosario/Desktop/Projects/pass_rate.xlsx")
```

Data Wrangling:

The code chunk below effectively transforms selected columns within the pass_rate data frame into a numeric format. This transformation is essential for conducting a range of data analyses in R, including statistical analysis, data modeling, and the creation of visual representations. Such a conversion is particularly crucial when handling imported data, where numerical values are often represented as text. By converting these columns to a numeric format, the script ensures that the data aligns with the requirements for advanced data processing and analysis in R, facilitating accurate and efficient data manipulation and interpretation.

```r
pass_rate$`Intro Statistics` <- as.numeric(as.character(pass_rate$`Intro Statistics`))

pass_rate$`Intro Probability` <- as.numeric(as.character(pass_rate$`Intro Probability`))

pass_rate$`ELA 101` <- as.numeric(as.character(pass_rate$`ELA 101`))

pass_rate$History <- as.numeric(as.character(pass_rate$History))

pass_rate$Biology <- as.numeric(as.character(pass_rate$Biology))

head(pass_rate, 10)
```

```
## # A tibble: 10 x 6
##    Student ‘Intro Statistics‘ ‘Intro Probability‘ ‘ELA 101‘ History Biology
##      <dbl>              <dbl>               <dbl>     <dbl>   <dbl>   <dbl>
##  1       1                 52                  89        73      66      85
##  2       2                 81                  60        86      74      92
##  3       3                 88                  72        70      79      59
##  4       4                 63                  76        93      67      71
##  5       5                 80                  54        60      68      88
##  6       6                 55                  85        90      57      61
##  7       7                 84                  79        77      64      91
##  8       8                 86                  62        53      70      92
##  9       9                 64                  87        58      90      75
## 10      10                 91                  82        69      54      89
```

Exploratory data analysis:

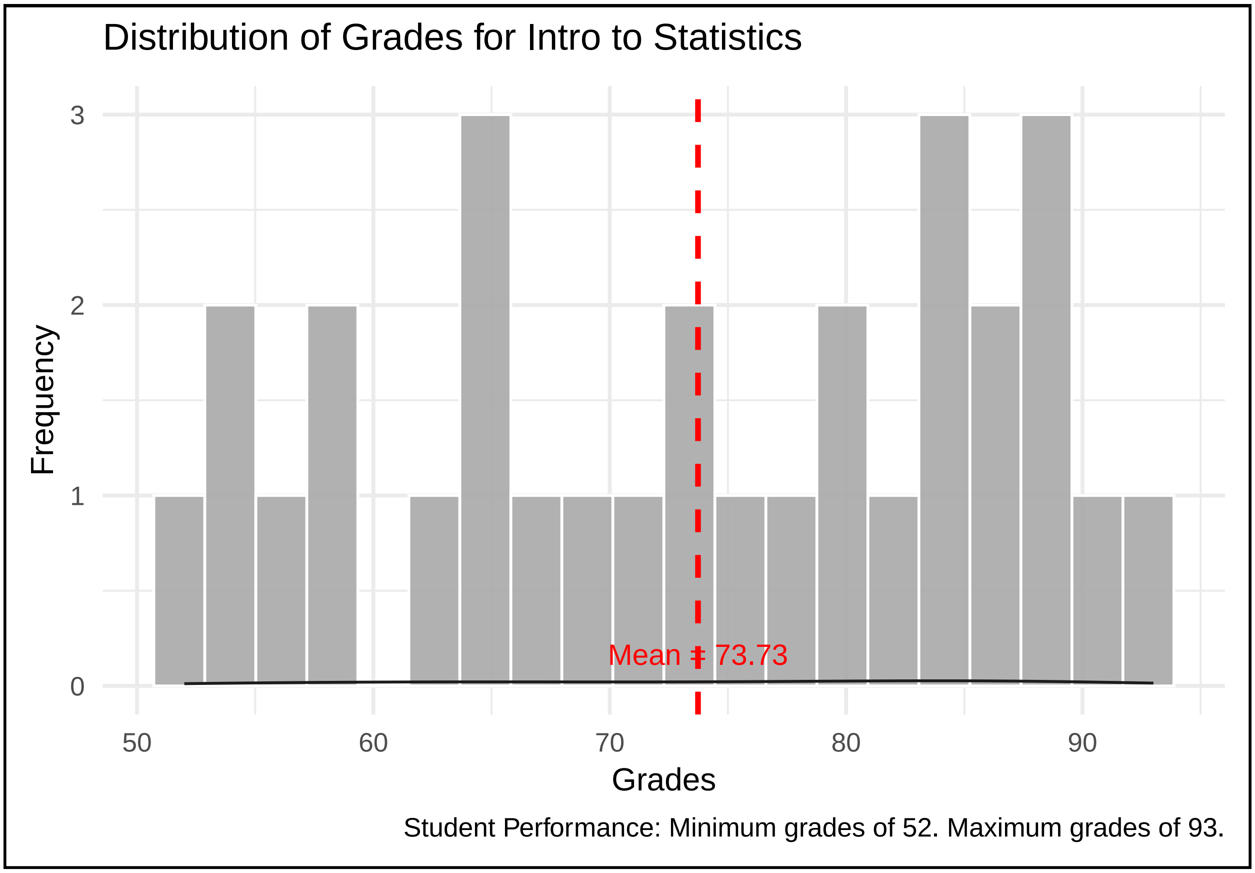
Exploratory Data Analysis (EDA) is a fundamental step in the data science process, serving as the bridge between the initial data acquisition phase and more formal analysis techniques. The primary purpose of EDA is to allow data scientists to understand the patterns, relationships, anomalies, and underlying structures within their data. By employing a combination of statistical summaries and visualizations, EDA provides a comprehensive insight into the nature of the data without making any assumptions about its underlying distribution or relationships. This initial exploration helps identify potential issues such as missing data, outliers, or incorrect data types, which are crucial to address before moving on to more complex analyses or model building.

In the following section, we present a series of bar graphs that illustrate trends from the first semester of courses taken by statistics majors at the university. Each statistics major is mandated to enroll in a set of core classes, including Introduction to Statistics, Introduction to Probability, ELA 101, American History, and Biology. Through these graphs, we aim to uncover patterns within the current curriculum and provide insights into the overall performance of students in each respective class. This visual analysis serves as a tool to assess the effectiveness of the curriculum and the academic success of students in these foundational courses.

```r
library(ggplot2)

# Calculate mean
data_mean <- mean(pass_rate$`Intro Statistics`, na.rm = TRUE)

# Create the histogram
ggplot(data = pass_rate, aes(x = `Intro Statistics`)) +
  geom_histogram(bins = 20, fill = "#A9A9A9", color = "white", alpha = 0.9) +
  geom_density(color = "#1C1C1C") +
  geom_vline(aes(xintercept = data_mean), color = "red", linetype = "dashed", size = 1) +
  annotate("text", x = data_mean, y = 0, label = paste("Mean =", round(data_mean, 2)), vjust = -1, colo
  labs(
    title = "Distribution of Grades for Intro to Statistics",
    x = "Grades",
    y = "Frequency",
    caption = "Student Performance: Minimum grades of 52. Maximum grades of 93."
  ) +
  theme_minimal(base_size = 15) +
  theme(
    plot.title = element_text(hjust = 0, vjust = 2),
    text = element_text(size = 12),
    plot.background = element_rect(color = "black", size = 1, linetype = "solid"),
    plot.margin = margin(10, 10, 10, 10),
    plot.caption = element_text(hjust = 1)
  )
```



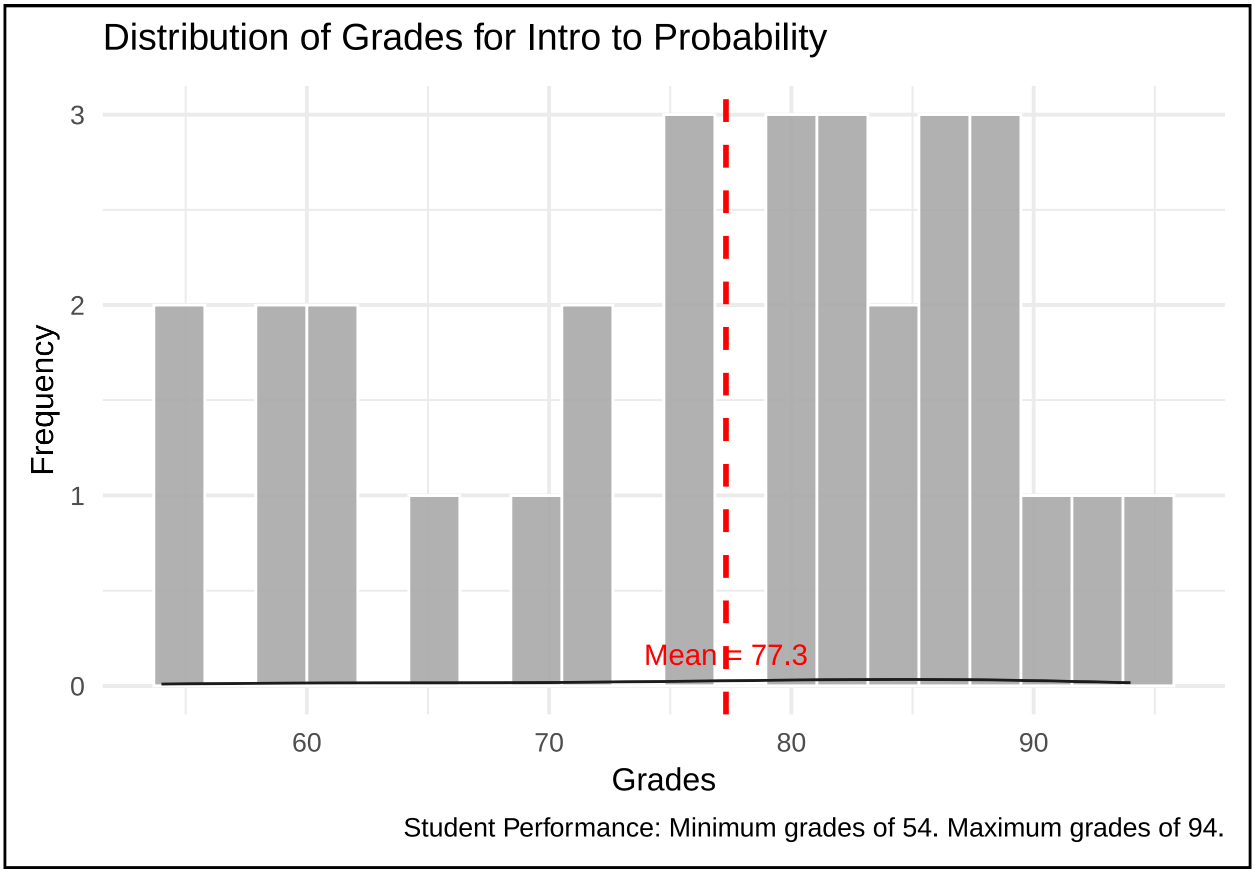
Class 1 (Intro to Statistics):

The grades in Class 1 have a range from a minimum of 52 to a maximum of 93, indicating a wide spread in student performance. The median grade in this class is 75, suggesting that half of the students scored above 75 and half scored below. The mean grade is approximately 73.73, which is slightly lower than the median. The interquartile range, spanning from 64 (the first quartile) to 84.75 (the third quartile), indicates the range of grades for the middle 50% of students in this class.

```r
library(ggplot2)

# Calculate mean
data_mean <- mean(pass_rate$`Intro Probability`, na.rm = TRUE)

# Create the histogram
ggplot(data = pass_rate, aes(x = `Intro Probability`)) +
  geom_histogram(bins = 20, fill = "#A9A9A9", color = "white", alpha = 0.9) +
  geom_density(color = "#1C1C1C") +
  geom_vline(aes(xintercept = data_mean), color = "red", linetype = "dashed", size = 1) +
  annotate("text", x = data_mean, y = 0, label = paste("Mean =", round(data_mean, 2)), vjust = -1, colo
  labs(
    title = "Distribution of Grades for Intro to Probability",
    x = "Grades",
    y = "Frequency",
    caption = "Student Performance: Minimum grades of 54. Maximum grades of 94."
  ) +
  theme_minimal(base_size = 15) +
  theme(
    plot.title = element_text(hjust = 0, vjust = 2),
    text = element_text(size = 12),
    plot.background = element_rect(color = "black", size = 1, linetype = "solid"),
    plot.margin = margin(10, 10, 10, 10),
    plot.caption = element_text(hjust = 1)
  )
```


Distribution of Grades for Intro to Probability

Student Performance: Minimum grades of 54. Maximum grades of 94.

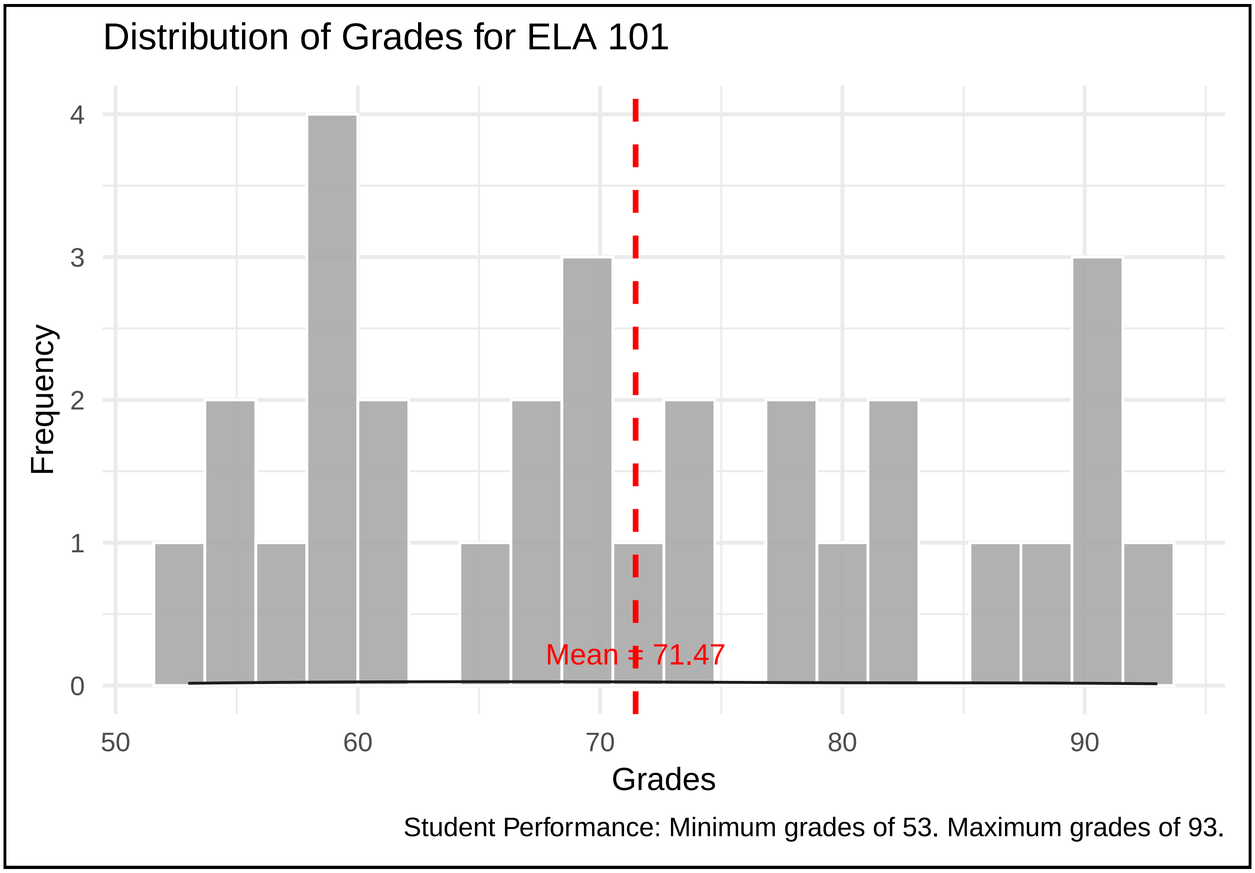
Class 2 (Intro to Probability):

Class 2 exhibits a range of student grades between 54 (minimum) and 94 (maximum). The median score in this class is 80.5, with half of the students obtaining scores above this mark and half below. The average score for Class 2 is 77.3, which is slightly lower than the median. The interquartile range, between 70.25 (the first quartile) and 87 (the third quartile), shows the middle 50% of the grades in Class 2.

```
library(ggplot2)

# Calculate mean
data_mean <- mean(pass_rate$`ELA 101`, na.rm = TRUE)

# Create the histogram
ggplot(data = pass_rate, aes(x = `ELA 101`)) +
  geom_histogram(bins = 20, fill = "#A9A9A9", color = "white", alpha = 0.9) +
  geom_density(color = "#1C1C1C") +
  geom_vline(aes(xintercept = data_mean), color = "red", linetype = "dashed", size = 1) +
  annotate("text", x = data_mean, y = 0, label = paste("Mean =", round(data_mean, 2)), vjust = -1, colo
  labs(
    title = "Distribution of Grades for ELA 101",
    x = "Grades",
    y = "Frequency",
    caption = "Student Performance: Minimum grades of 53. Maximum grades of 93."
  ) +
  theme_minimal(base_size = 15) +
  theme(
    plot.title = element_text(hjust = 0, vjust = 2),
    text = element_text(size = 12),
    plot.background = element_rect(color = "black", size = 1, linetype = "solid"),
    plot.margin = margin(10, 10, 10, 10),
    plot.caption = element_text(hjust = 1)
  )
```



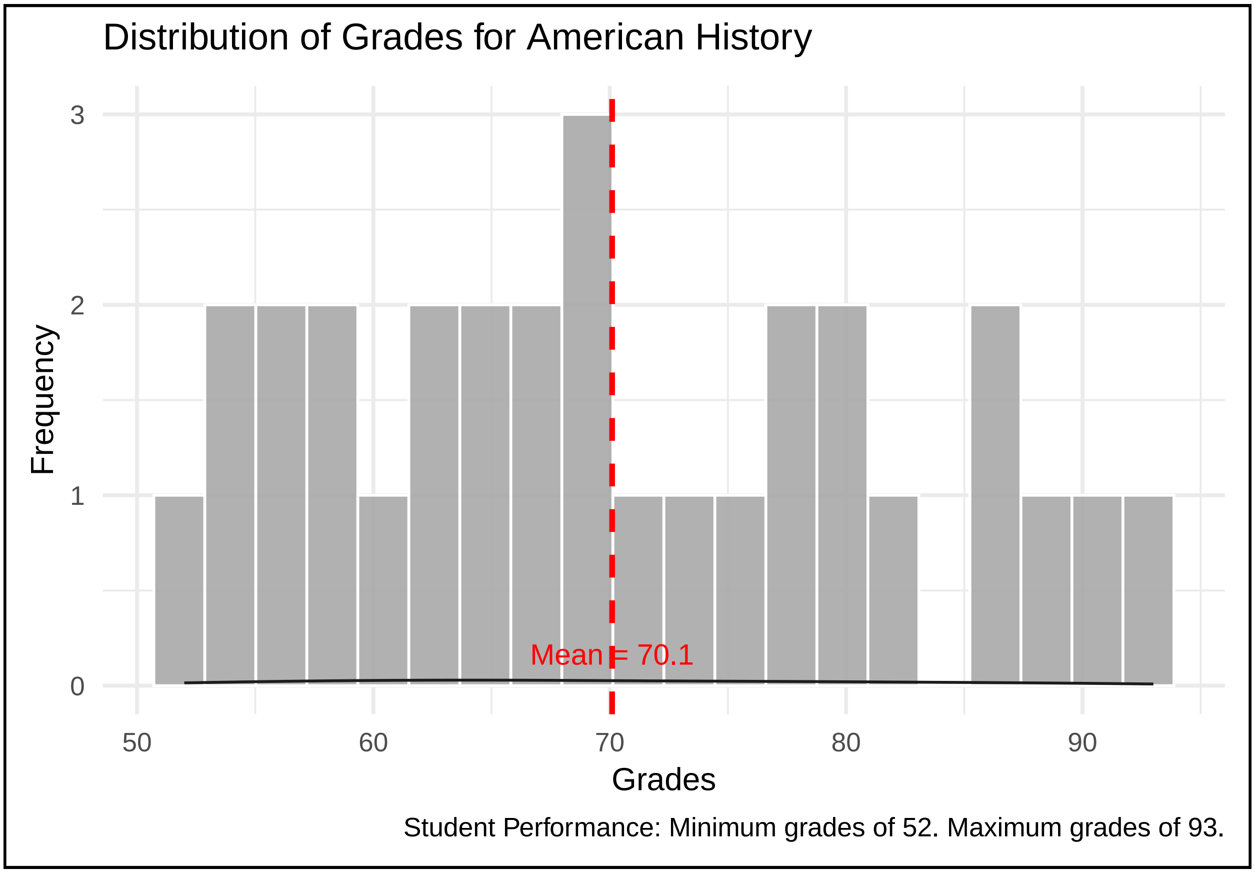
Class 3 (English 101):

In Class 3, student grades range from a minimum of 53 to a maximum of 93. The median grade is 69.5, suggesting a fairly balanced distribution of grades around this central value. The mean of 71.47 is slightly higher than the median, indicating that the data might be slightly skewed to the right. The interquartile range for Class 3 is from 60.5 to 81.5, highlighting the middle 50% of student grades in this class.

```
library(ggplot2)

# Calculate mean
data_mean <- mean(pass_rate$`History`, na.rm = TRUE)
```

```r
# Create the histogram
ggplot(data = pass_rate, aes(x = `History`)) +
  geom_histogram(bins = 20, fill = "#A9A9A9", color = "white", alpha = 0.9) +
  geom_density(color = "#1C1C1C") +
  geom_vline(aes(xintercept = data_mean), color = "red", linetype = "dashed", size = 1) +
  annotate("text", x = data_mean, y = 0, label = paste("Mean =", round(data_mean, 2)), vjust = -1, colo
  labs(
    title = "Distribution of Grades for American History",
    x = "Grades",
    y = "Frequency",
    caption = "Student Performance: Minimum grades of 52. Maximum grades of 93."
  ) +
  theme_minimal(base_size = 15) +
  theme(
    plot.title = element_text(hjust = 0, vjust = 2),
    text = element_text(size = 12),
    plot.background = element_rect(color = "black", size = 1, linetype = "solid"),
    plot.margin = margin(10, 10, 10, 10),
    plot.caption = element_text(hjust = 1)
  )
```



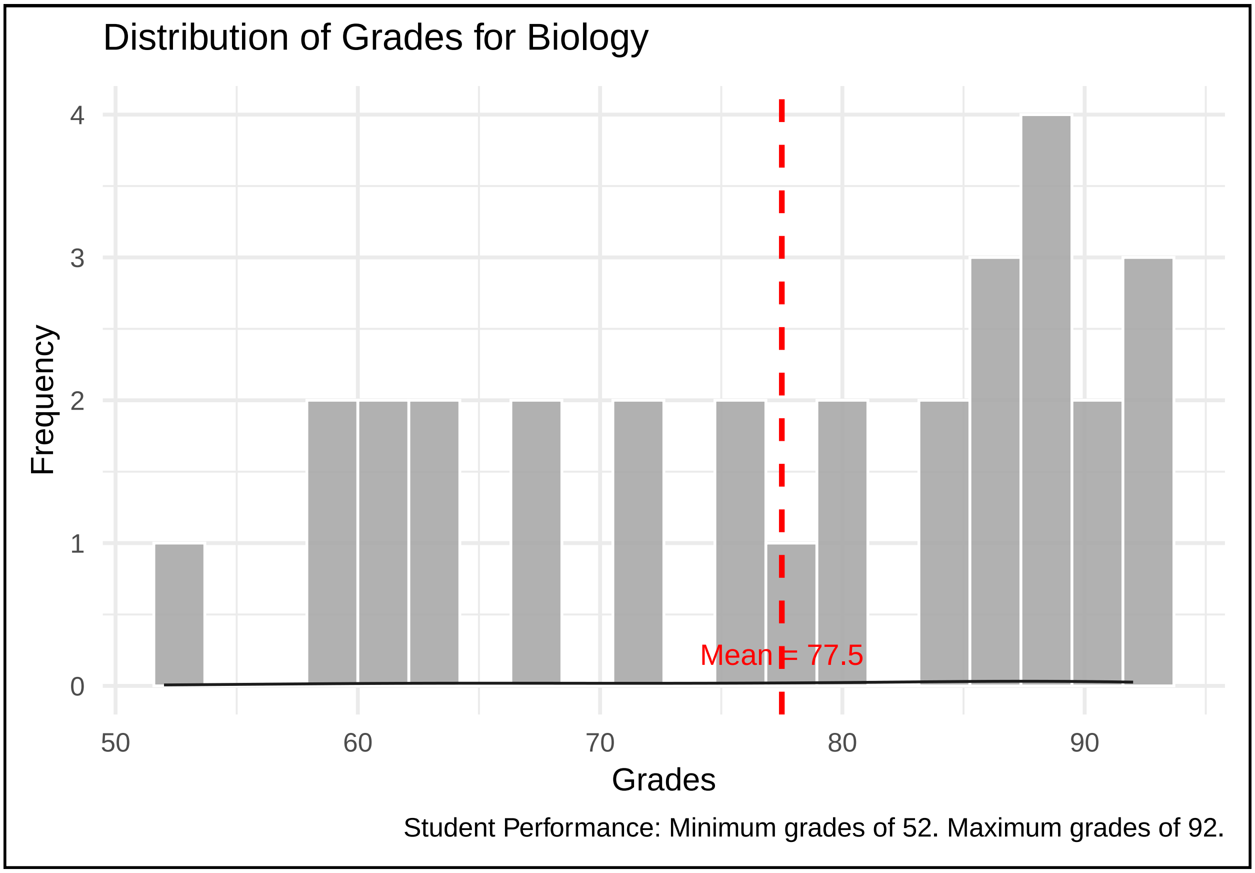
Class 4 (American History):

The grades in Class 4 show a spread from a minimum of 52 to a maximum of 93. With a median of 68, this class has half of its students scoring above this mark and half scoring below. The average grade in this class is 70.1, which is higher than the median, suggesting a potential slight right skew in the grades. The interquartile range for Class 4, which spans from 61.25 to 78.75, contains the middle 50% of the grades for students in this class.

```r
library(ggplot2)

# Calculate mean
data_mean <- mean(pass_rate$`Biology`, na.rm = TRUE)

# Create the histogram
ggplot(data = pass_rate, aes(x = `Biology`)) +
  geom_histogram(bins = 20, fill = "#A9A9A9", color = "white", alpha = 0.9) +
  geom_density(color = "#1C1C1C") +
  geom_vline(aes(xintercept = data_mean), color = "red", linetype = "dashed", size = 1) +
  annotate("text", x = data_mean, y = 0, label = paste("Mean =", round(data_mean, 2)), vjust = -1, colo
  labs(
    title = "Distribution of Grades for Biology",
    x = "Grades",
    y = "Frequency",
    caption = "Student Performance: Minimum grades of 52. Maximum grades of 92."
  ) +
  theme_minimal(base_size = 15) +
  theme(
    plot.title = element_text(hjust = 0, vjust = 2),
    text = element_text(size = 12),
    plot.background = element_rect(color = "black", size = 1, linetype = "solid"),
    plot.margin = margin(10, 10, 10, 10),
    plot.caption = element_text(hjust = 1)
  )
```



Class 5 (Biology):

Class 5 has student grades ranging from a minimum of 52 to a maximum of 92. The median grade in this class is 80, indicating that half of the students in the class scored above an 80 and half scored below. The mean grade in Class 5 is 77.5, which is slightly lower than the median. The interquartile range in this class, from 67.25 (the first quartile) to 88 (the third quartile), outlines the range of grades for the middle 50% of the students in Class 5.

Results:

This R code snippet is designed to analyze a dataset containing student grades, transforming these grades into a binary pass/fail format, and then calculating the probabilities of passing. Initially, it converts student grades into a binary indicator (1 for pass, 0 for fail) based on a threshold of 65%. This conversion is applied to all columns except the first, presumably identifying each student. Subsequently, the student identifier column is reattached to ensure that each student's pass/fail status is maintained correctly across subjects. The script then calculates the probability of passing for each class by computing the mean of the binary values in each column, which represents the class pass rate. Finally, it determines the overall probability of a student passing all classes by multiplying these individual probabilities, assuming that passing each class is an independent event. This process offers a comprehensive overview of the pass rates, providing valuable insights into student performance and class difficulty levels.

The multiplication of individual class pass probabilities to determine the overall probability of passing all classes is grounded in the principles of probability theory concerning independent events. Each class's pass or fail status is considered an independent event, implying that the result in one class doesn't influence the outcome in another. For independent events, the joint probability of multiple events occurring simultaneously—like passing all classes—is calculated by multiplying their individual probabilities. This approach reflects the intuitive notion that if a student has a low chance of passing even one class, it significantly impacts their overall likelihood of passing every class. Therefore, by multiplying the probabilities of passing each class, we get a comprehensive measure of a student's likelihood of succeeding across all their courses, encapsulating the interconnected stakes of their academic performance.

```
prob_pass_each_class
```

```
##  Intro Statistics Intro Probability           ELA 101           History
##         0.7000000         0.8000000         0.6666667         0.6333333
##           Biology
##         0.7666667
```

```
prob_pass_all_classes
```

```
## [1] 0.1812741
```

In the dataset, Intro to Probability exhibits the highest pass rate among the five classes, with 80% of students passing, followed by Biology with a pass rate of approximately 76.67%. Intro to Statistics has a 70% pass rate, making it the third highest among the classes. ELA 101 and History have relatively lower pass rates, with approximately 66.67% and 63.33% of students passing, respectively. It is notable that the overall pass rate for all classes combined is 18.13%, which is substantially lower than the individual pass rates for each of the classes. This suggests that the calculation for the overall pass rate might need to be re-evaluated, as one would expect the overall pass rate to fall within the range of the individual class pass rates.

Conclusion:

In examining the dataset, it is evident that there are notable differences in the pass rates among the five classes. Intro to Probability leads with an impressive 80% pass rate, closely followed by Biology, which boasts a 76.67% success rate. Intro to Statistics occupies the middle ground, registering a 70% pass rate, which is still considerably higher than the rates for ELA 101 and History, where the pass rates are approximately 66.67% and 63.33%, respectively.

Interestingly, when we look at the collective performance across all classes, the overall pass rate stands at a surprisingly low 18.13%. This figure is significantly below the pass rates observed in the individual classes, hinting at a potential discrepancy in the calculation method for the combined pass rate. Ideally, one would anticipate that the cumulative pass rate would align more closely with the range of individual class pass rates.

Therefore, it is recommended to revisit the computational approach at NotReal University for determining the overall pass rate to ensure its accuracy and reliability. Furthermore, recognizing the variations in pass rates between individual classes can serve as a basis for implementing targeted strategies to enhance the curriculum and foster higher success rates in future University cohorts.