Assessment and Prediction of Undergraduate Student Performance

Hasan Roshan

December 2019

1. Introduction

1.1. Project Overview

This project examines the factors influencing undergraduate student performance in Anatomy and Physiology (A&P) courses at the University of Findlay, Ohio. The dataset includes approximately 550 students' academic records, including their final grades, GPA, SAT/ACT scores, demographic information, and prerequisite coursework.

The goal of this study is to understand:

- How different academic and demographic factors relate to student performance.
- Whether certain variables can serve as predictors of success in future courses.
- Potential statistical differences between different student groups.
- Providing data-informed recommendations for undergraduates seeking their advisor's consultation.

1.2. Scope of This Analysis

- This document **focuses on the data analysis process**.
- **Data preparation is handled in a separate R script** and is not covered here.
- Some referenced analyses (such as normality tests) are included in a seperate R-Markdown file.

1.3. Document Structure

1. **Data Import & Variable Selection**: Loading the dataset and selecting relevant variables.
2. **Exploratory Data Analysis (EDA)**: Investigating relationships between variables.
3. **Data Visualization**: Using visual tools to explore performance trends.
4. **Summary & Next Steps**: Key findings and future directions.

---

2. Data Import & Variable Selection

2.1. Loading Data

```r
# Clear workspace
rm(list = ls())

# Read dataset
mydata <- read.csv("Data/All data-ready.csv", header = TRUE)
```

2.2. Selecting Relevant Variables

```r
# Selecting specific columns for analysis
Data <- mydata[, c(2, 11, 30, 9, 6, 10, 37, 18, 3:4, 38:39, 7:8, 28:29)]
```

Let's take a quick look at the data structure using 'str()' function:

```r
str(Data)
```

```
## 'data.frame':    542 obs. of  16 variables:
##  $ Academic.Year    : chr  "2012-2013" "2012-2013" "2012-2013" "2012-2013" ...
##  $ Academic.Standing: chr  "Sophomore" "Sophomore" "Junior" "Junior" ...
##  $ College          : chr  "Science" "Science" "Science" "Science" ...
##  $ Major            : chr  "Animal Science" "Animal Science" "Animal Science" "Animal Science" ...
##  $ Sex              : chr  "F" "F" "F" "F" ...
##  $ Age              : int  19 19 20 21 20 21 20 21 20 20 ...
##  $ ACT.SAT.P        : int  38 87 55 74 97 83 89 69 55 79 ...
##  $ GPA              : num  3.59 3.16 2.89 3.22 3.42 2.79 3.56 3.51 3.36 3.5 ...
##  $ Final.322        : num  75 68 54 66 86 76 80 64 54 64 ...
##  $ Final.323        : num  82 69 72 55 81 86 87 90 68 75 ...
##  $ PF.FE.322        : chr  "Pass" "Fail" "Fail" "Fail" ...
##  $ PF.FE.323        : chr  "Pass" "Fail" "Pass" "Fail" ...
##  $ PF.322           : chr  "Pass" "Pass" "Pass" "Pass" ...
##  $ PF.323           : chr  "Pass" "Pass" "Pass" "Pass" ...
##  $ Numb.Pre.Req     : int  2 1 1 2 2 1 2 2 2 2 ...
##  $ Pre.Req.Taken    : chr  "Y" "Y" "Y" "Y" ...
```

Using the `str()` function, we gain an overview of the data we will be working with, including:

- There are 542 observations and 16 variables.
- There are **quantitative** and **qualitative** variables.

The selected variables focus on **academic performance indicators**, including:

- **GPA prior to course enrollment**
- **ACT/SAT scores**
- **Course pass/fail status**
- **Demographic information (age, sex, major, year in college)**
- **Whether students completed prerequisite courses**

Let's review the first 5 rows of the data:

```r
# Selecting specific columns for analysis
head(Data, 5)
```

```
##   Academic.Year Academic.Standing College          Major Sex Age ACT.SAT.P  GPA
## 1     2012-2013         Sophomore Science Animal Science   F  19        38 3.59
## 2     2012-2013         Sophomore Science Animal Science   F  19        87 3.16
## 3     2012-2013            Junior Science Animal Science   F  20        55 2.89
## 4     2012-2013            Junior Science Animal Science   F  21        74 3.22
## 5     2012-2013            Junior Science Animal Science   F  20        97 3.42
##   Final.322 Final.323 PF.FE.322 PF.FE.323 PF.322 PF.323 Numb.Pre.Req
## 1        75        82      Pass      Pass   Pass   Pass            2
## 2        68        69      Fail      Fail   Pass   Pass            1
## 3        54        72      Fail      Pass   Pass   Pass            1
## 4        66        55      Fail      Fail   Pass   Pass            2
## 5        86        81      Pass      Pass   Pass   Pass            2
##   Pre.Req.Taken
## 1             Y
## 2             Y
## 3             Y
## 4             Y
## 5             Y
```

---

3. Exploratory Data Analysis (EDA)

3.1. Data Summary

To get a summary of the data, we use the `summary()` function.

```r
# Display data summary
summary(Data)
```

```
##  Academic.Year      Academic.Standing    College             Major
##  Length:542         Length:542         Length:542         Length:542
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      Sex                 Age          ACT.SAT.P          GPA
##  Length:542         Min.   :18.00   Min.   :18.00   Min.   :2.180
##  Class :character   1st Qu.:19.00   1st Qu.:62.00   1st Qu.:3.150
##  Mode  :character   Median :20.00   Median :74.00   Median :3.440
##                     Mean   :19.85   Mean   :71.68   Mean   :3.427
##                     3rd Qu.:20.00   3rd Qu.:84.00   3rd Qu.:3.760
##                     Max.   :28.00   Max.   :99.00   Max.   :4.000
##    Final.322        Final.323       PF.FE.322          PF.FE.323
##  Min.   : 43.50   Min.   :  0.00   Length:542         Length:542
##  1st Qu.: 68.00   1st Qu.: 63.00   Class :character   Class :character
##  Median : 76.00   Median : 71.00   Mode  :character   Mode  :character
```

```
##  Mean   : 74.79   Mean   : 71.00
##  3rd Qu.: 83.75   3rd Qu.: 80.75
##  Max.   :100.00   Max.   :101.00
##     PF.322             PF.323          Numb.Pre.Req    Pre.Req.Taken
##  Length:542         Length:542         Min.   :0.0000   Length:542
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :1.0000   Mode  :character
##                                        Mean   :0.9686
##                                        3rd Qu.:2.0000
##                                        Max.   :3.0000
```

Observations:

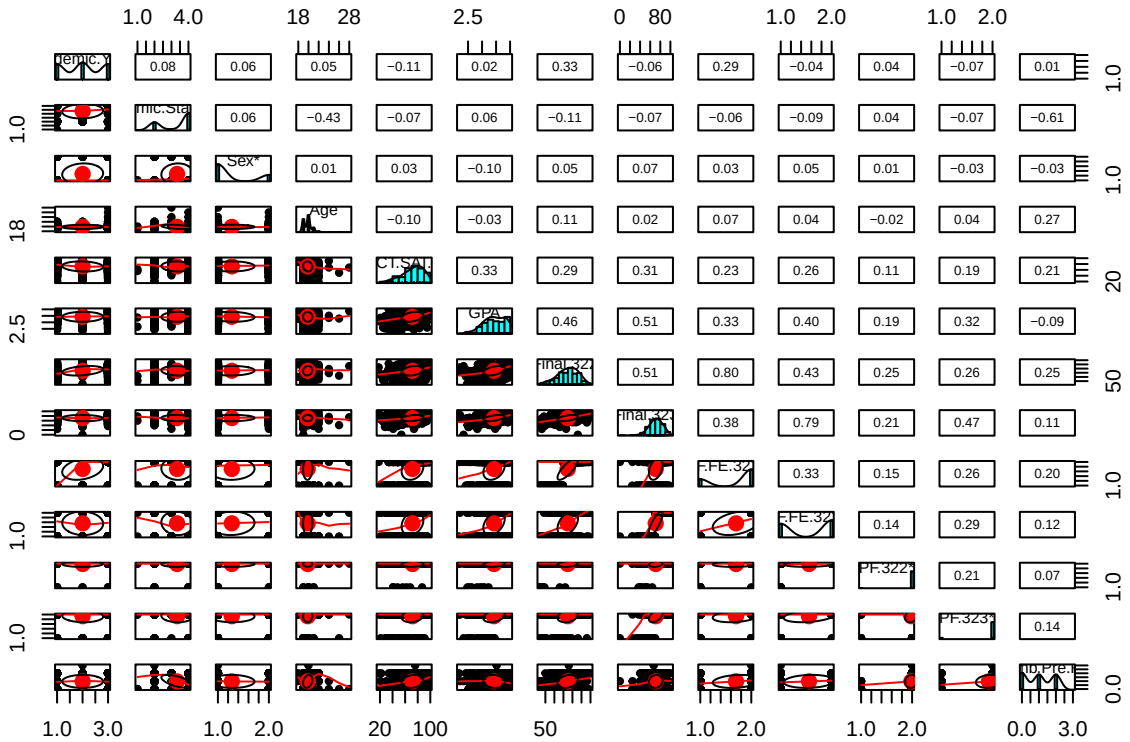
- Since data preparation is already done, there is no *NA* in the dataset.

3.2. Checking Relationships Between Variables

To explore the relationships between different numerical variables, we use the `pairs.panels()` function from `psych` package.

```
library(psych)

# Display pairwise relationships between selected numeric variables
pairs.panels(Data[, c(1:2, 5:15)])
```



Observations:

- This function provides scatterplots, correlation coefficients, and density distributions for numeric variables.
- Key trends and correlations can be identified in this step.

---

4. Data Visualization

We explored numerous visualization to identify meaningful trends and relatioships between different variables. Here, we will review a few of them.

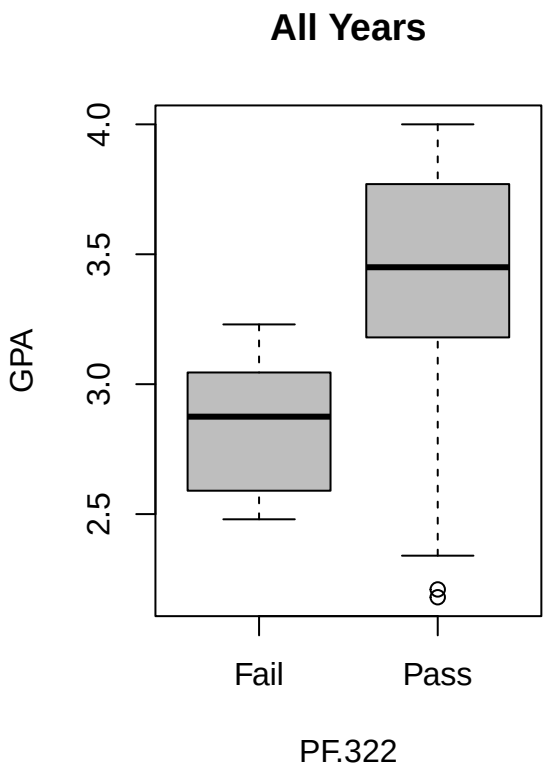
4.1. Visualizing Student Performance

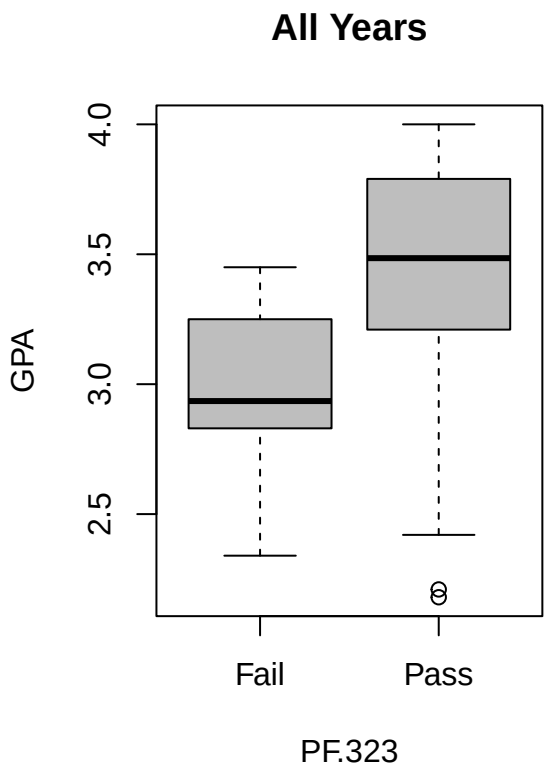
We generate visualizations to better understand how student performance varies. Let's check how students' GPA relate to them being able to pass the course during academic years 2011-2014 .

```
# Set up plotting layout: Two plots side by side
par(mfrow = c(1,2))

# First boxplot for PF.322 vs. Number of Pre-Requisite Courses Taken
boxplot(GPA ~ PF.322, data = Data,
        main = "All Years",
        xlab = "PF.322",
        ylab = "GPA",
        col = "gray")

# Second boxplot for PF.323 vs. Number of Pre-Requisite Courses Taken
boxplot(GPA ~ PF.323, data = Data,
        main = "All Years",
        xlab = "PF.323",
        ylab = "GPA",
        col = "gray")
```





Observations:

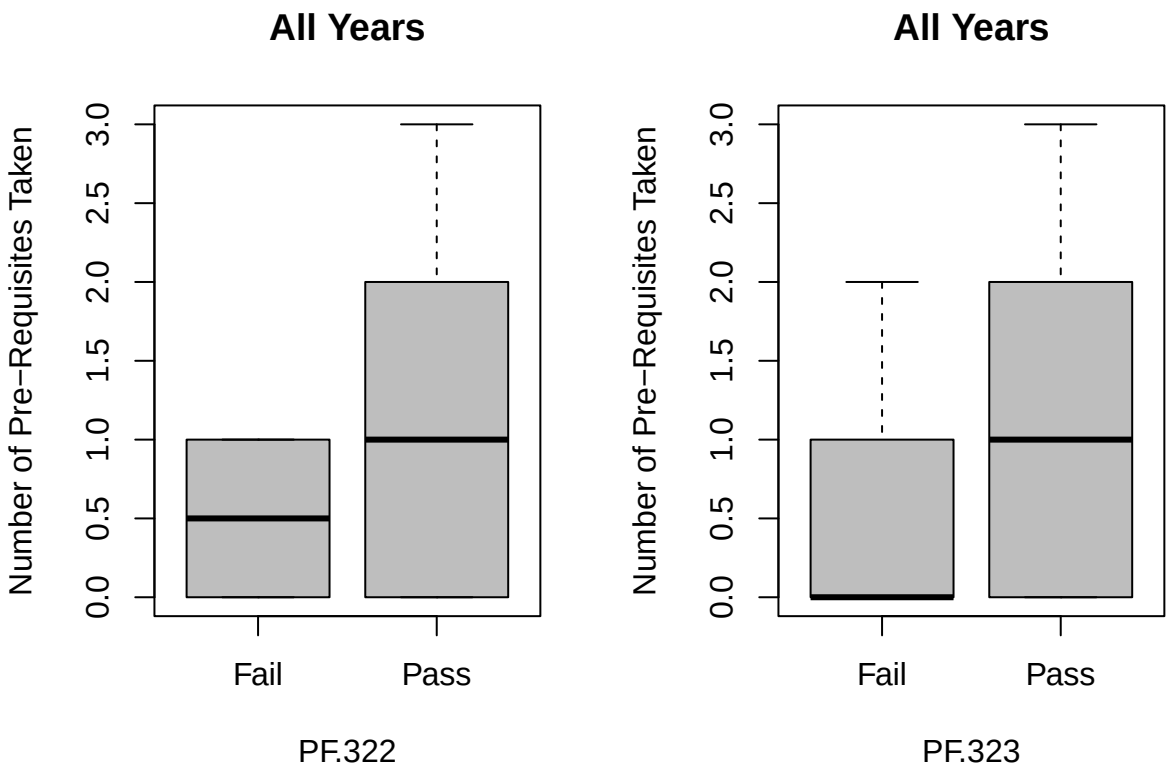
- As expected, students with **higher** GPA are **more likely** to pass either of the more advanced courses (PF.322 & PF.323).
- A GPA higher than 3.5 seems a safe criteria for success.

Now let's look at another variable, number of pre-requisite courses a student has taken prior to enrolling to either of the two courses.

```
# Set up plotting layout: Two plots side by side
par(mfrow = c(1,2))

# First boxplot for PF.322 vs. Number of Pre-Requisite Courses Taken
boxplot(Numb.Pre.Req ~ PF.322, data = Data,
        main = "All Years",
        xlab = "PF.322",
        ylab = "Number of Pre-Requisites Taken",
        col = "gray")

# Second boxplot for PF.323 vs. Number of Pre-Requisite Courses Taken
boxplot(Numb.Pre.Req ~ PF.323, data = Data,
        main = "All Years",
        xlab = "PF.323",
        ylab = "Number of Pre-Requisites Taken",
        col = "gray")
```




```
# Reset plot layout to default
par(mfrow = c(1,1))
```

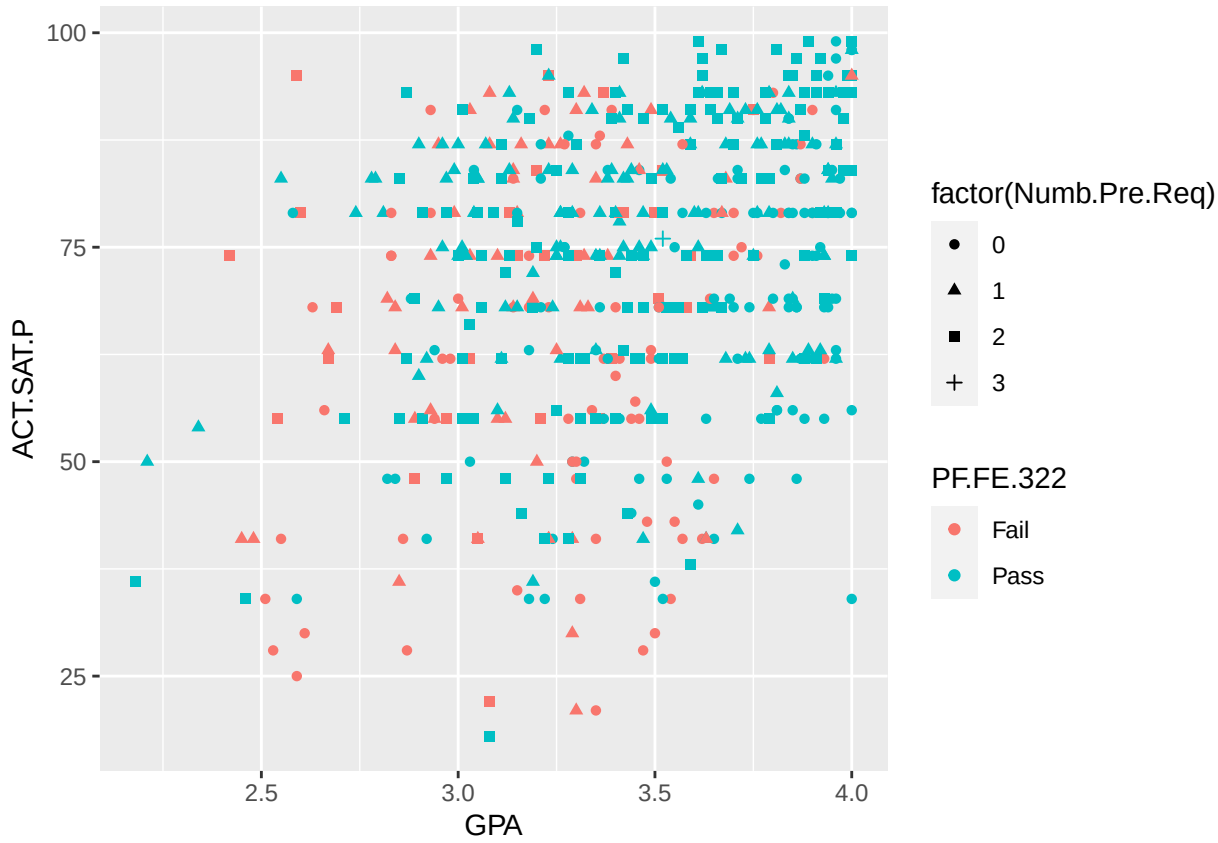
Observations:

- Here again, we observe that students who have already passed **more pre-requisite courses** have a **higher chance** of passing the more advanced courses (PF.322 & PF.323).
- Having passed at leas one prerequisite course seems crucial for good performance in either of the courses, particularly the PF.323.

Nonetheless, it is important to realize this is a preliminary finding based on a univariate analysis of the GPA. To provide robust recommendation to students, we need to further analyze the data. Further analysis like **Principle Component Analysis, Regression, Clustering, and Discriminatory Linear Analysis**, may be needed to understand the multivariability of this case. These analyases will be covered separately.

Let's visually inspect how multiple variables could impact a student's performance:

```
library(ggplot2)
ggplot(data = Data, mapping = aes(x = GPA, y = ACT.SAT.P, color = PF.FE.322, shape = factor(Numb.Pre.Re
        geom_point(alpha = 1) ##Good!
```

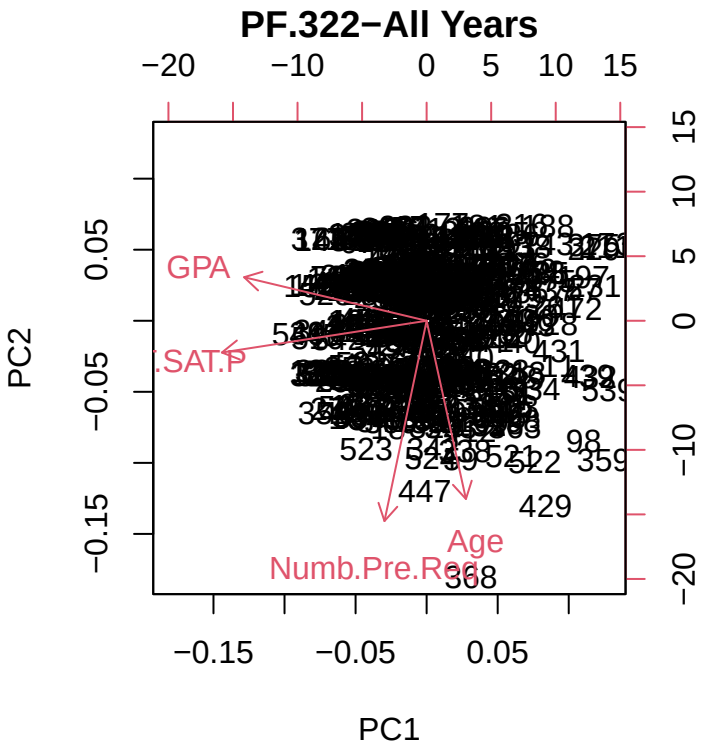


Observations:

- As illustrated in the figure, **higher scores** in college GPA and pre-college standardized tests (SAT or ACT) are associated with **higher chance** of passing the final exam of PF.322.
- Nonetheless, it seems that neither of college GPA or standardized test scores is the most important factor in passing the final exam of PF.322. It is the number of pre-requisite courses one has taken before enrolling in the more advanced course.
- These suggest two sorts of qualities seems to be the underlying factors: (1) **college experience/exposure** (2) **performance**.

The last point can be further strengthen upon a PCA focused on four variables: age, GPA, standardized test scores, and the number of pre-requisite courses.

```
Data.SV <- Data[, c(6:10, 15)] ## selected variables: Age, ACT.SAT.P, GPA, Final.322, Final.323, Numb.P
PC.cor.322T <- prcomp(Data.SV[, -c(4, 5)], scale. = TRUE)
biplot(PC.cor.322T, main = "PF.322-All Years")
```



Observations:

- While the graph might need more explanation, it shows two main factors can be defined based on the original four variables that independently contribute to success rate. The graph suggests two perpendicular axis can be formed by transforming the four variables where one axis seems to represent **experience** and the other, **school performance**.
- While this is not a thorough examination at this point, it provides insights on the data and a posiible direction for hypothesis development and testing. This concludes our exploratory data analysis. The next step is a confirmatory data analysis (CDA) that will be performed in another document.

5. Summary & Next Steps

In this document, we reviewed explanatory data analysis. First, we imported a dataset of students at the University of Findlay that was prepared in the previous document. Then, we selected a subset of the data, reviewed its structure and a summary of the data to ensure it is ready for analysis. Then, we reviewed the relationship between different variables through visualizations. We drew a few observations and insights based on these explanatory analysis.

5.1. Key Takeaways

- Data has been successfully imported and prepared for analysis.
- Initial **exploratory analysis** has helped in identifying trends and correlations. For instance:
  - GPA, standarized test scores, and number of pre-requisite courses taken have positive relationship with success rate in passing the courses.

  - An student adviser ought not to recommend enrollment to either of the two courses (PF.322 or PF.323) if the student has not passed any pre-requisite course, particularly the student's GPA is below 3.5.
  - A simpler explanation would be that two main factors contribute to one's success: (1) college experience/exposure; (2) school performance.

- Future steps involve **statistical modeling and deeper analysis**.

5.2. Future Directions

- **Normality Tests**: To be included in a separate markdown file.
- **Confirmatory data analysis (CDA)** will be explored in the next document to further investigate student performance predictors. This include **multiple linear regression (MLR), discriminatory linear analysis (DLA), and clustering Analysis**