Data Scholars – A Study on Academic Performance

Margaret Smith, Matthew Anderson,
Miah Grubbs, Grace Griffith

Introduction

Question

Which academic and nonacademic factors contribute most to a college student's GPA?

Collected Data

The Student Performance Metrics Dataset highlights a collection of academic and non-academic attributes that aim to evaluate the factors that influence a student's performance in higher education.

Attributes

Student demographics, academic achievements, socio-economic factors, and extracurricular activities.

Importance

Evaluating which individual metrics may impact a college student's overall GPA will help both students and professors better navigate success in the classroom.

Exploring the Data Set

Quantitative

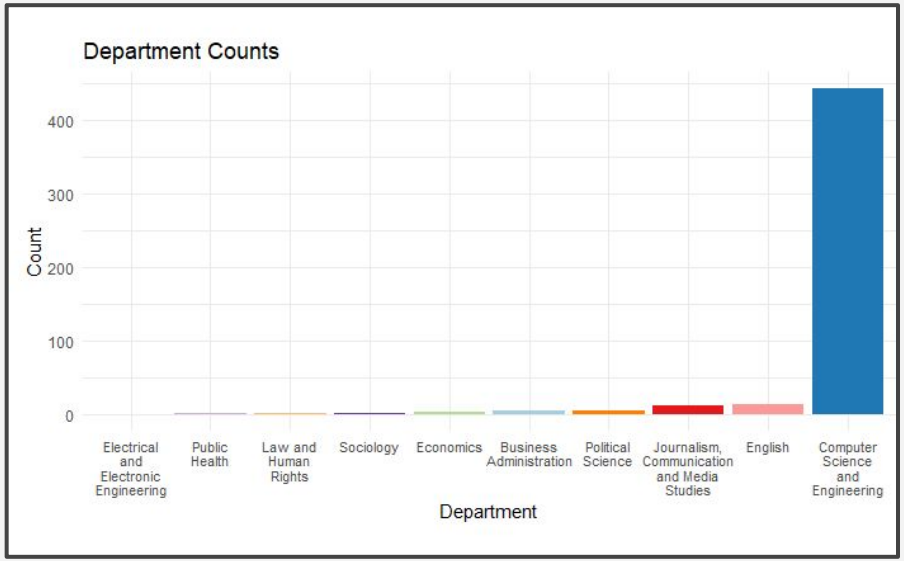
Overall - Overall GPA (Response)
HSC - Higher Secondary Education GPA
SSC - Secondary School Education GPA
Last - Last Semester's GPA
Semester - the current UG semester for an observation

Qualitative

Department
Gender
Income
Hometown
Computer
Gaming
Attendance
Job
English
Extra
Preparation

- All observations were collected via surveying undergraduate students
- Many variables that are traditionally numerically continuous – i.e. Income or preparation (time spent studying) – were transformed into discrete levels.
    - Ex: Income → less than 15k, 15-30k, 30k-50k, greater than 50k,

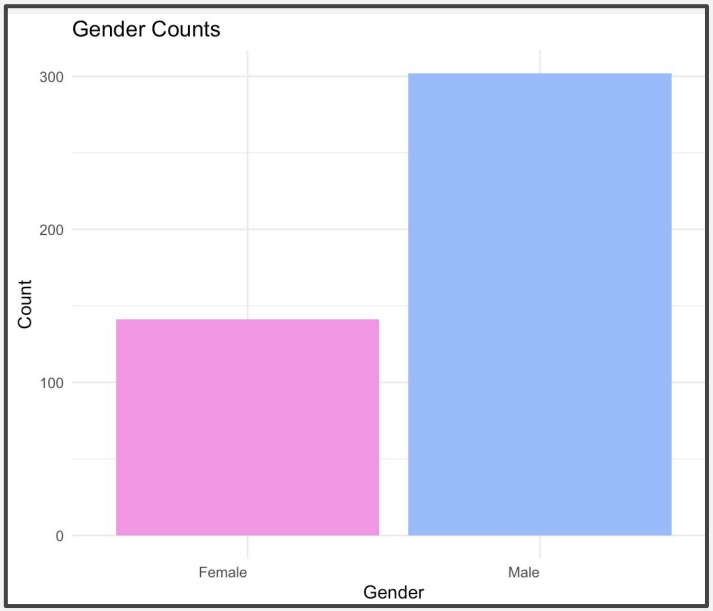
Exploratory Visualizations



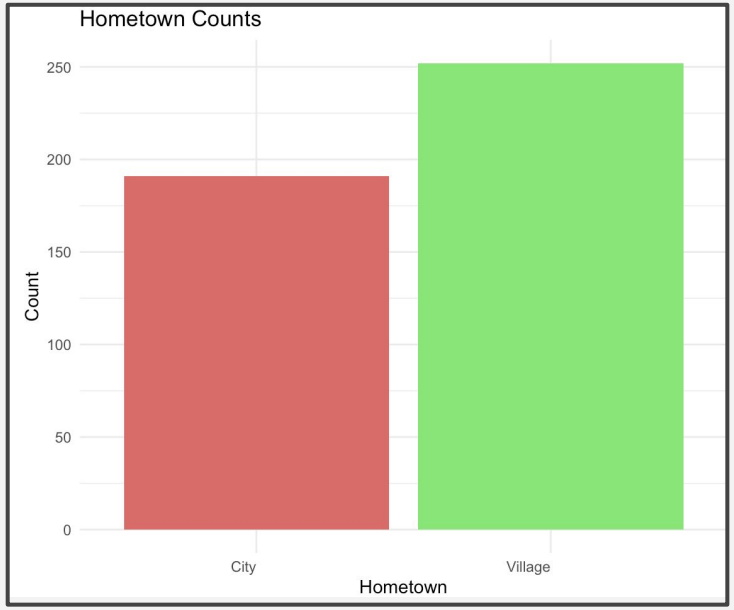
Department the student is studying under

- Of the 493 observations in the original dataset, 443 were Computer Science majors
- Will analyze a subset of the dataset
  - So, the conclusions drawn will be drawn from only CS undergraduate students

Exploratory Visualizations

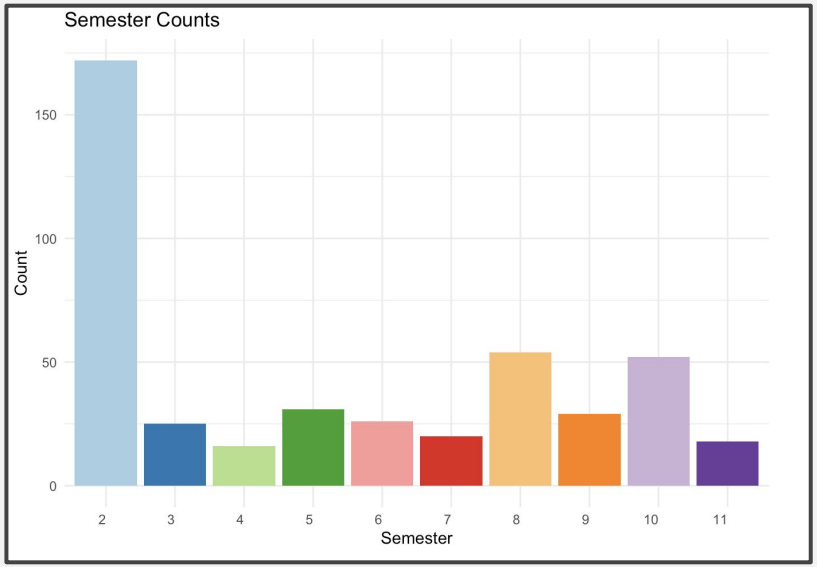


Gender of the recorded students

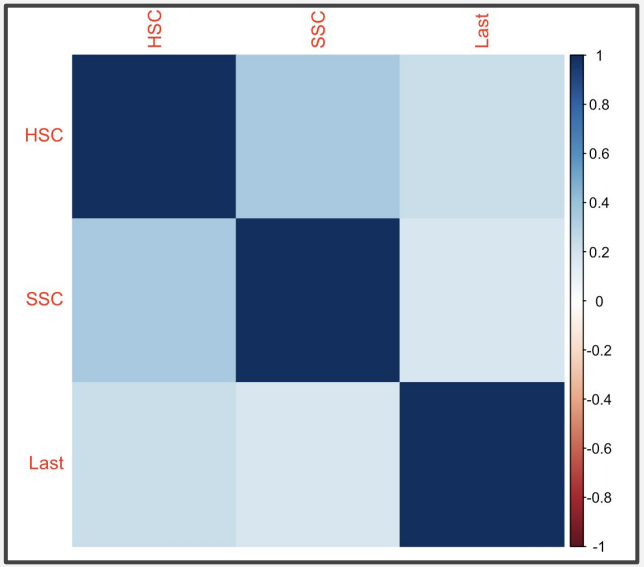


Whether the student lives in a urban or rural area

Exploratory Visualizations



How many semesters a student has been enrolled in



Correlation Matrix for the GPA values

Methods

Compared the output of various model *building algorithms:*

1. "Exhaustively" finding subsets using the leaps library
2. Stepping both "forward" and "backward" through building models

Choosing Models:

1. Relatively comparing the AIC, BIC, and mallow CP
2. Comparing adjusted R^2
3. Checking the VIF for the variables included in each model

Initial Model:

Predictors:

- Preparation (0-1 hrs, 2-3 hrs, >3 hrs)
- Attendance (<40%, 40%-59%, 60%-79%, 80%-100%)
- Last → last semester's GPA

Model: $\text{Overall} = 0.647 + 0.096x_{\text{Prep[2-3 hrs]}} + 0.063x_{\text{Prep[> 3hrs]}} - 0.075x_{\text{Attend[60\%-79\%]}} + 0.021x_{\text{Attend[80\%-100\%]}} - 0.171x_{\text{Attend[<40\%]}} + 0.796x_{\text{Last}}$

→ attendance between 40%-59% and preparation between 0-1 hrs is accounted for in the intercept

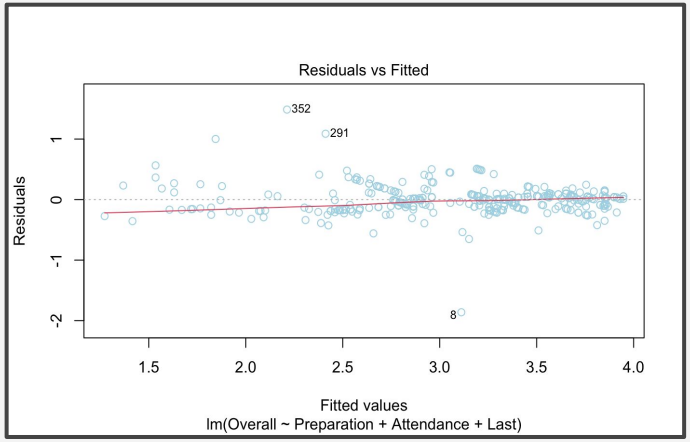
→ Last semester's GPA contributes the most towards predicting the overall GPA

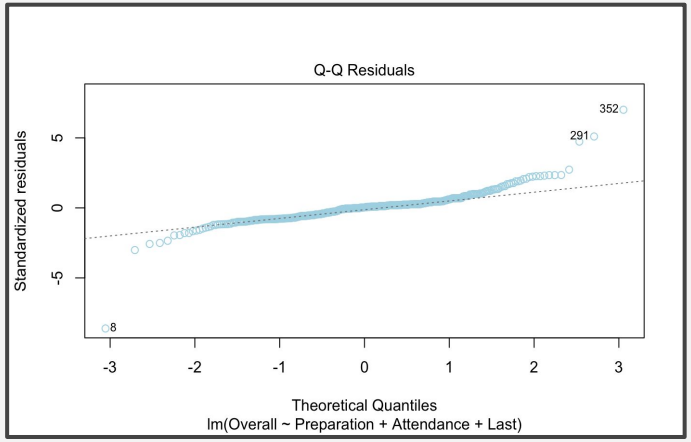
$R^2 = 0.8697$

Note:

- Using just the last semester's GPA produces almost as good of a model
  - **Model:** $\text{Overall} = 0.47 + 0.86x_{\text{last}}$
  - $R^2 = 0.868$

Assumptions: Overall ~ Preparation + Attendance + Last





Linearity/Homoscedasticity:

- Several observation drastically stands out
- No observable pattern
- Observations do not form a horizontal line around 0
    - Potentially *does not* pass linearity and homoscedastic conditions

Multicollinearity: VIF of predictors < 5

Normality:

- Reference line *not* close to a 45° angle
- A few observations *substantially deviate* from the reference line
    - Potentially *not* approximately normal

→ *Next: Investigate outliers*

Why Remove these Outliers?

- Large changes in Last vs. Overall GPA
- Shouldn't see these big differences

	Last	Overall
8	2.95	1.25
291	2.07	3.50
352	1.82	3.70

Note:

- Observation 8 was in the 11th semester
- Observation 291 was in the 4th semester
- Observation 352 was in the 10th semester.

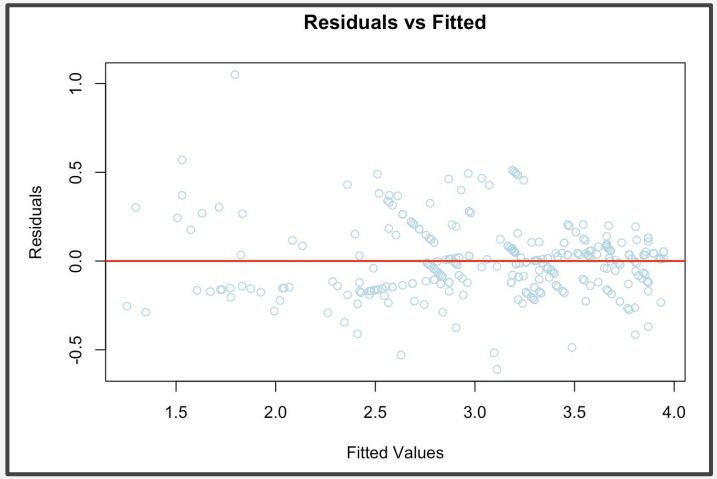
Model #2 - Outliers Removed

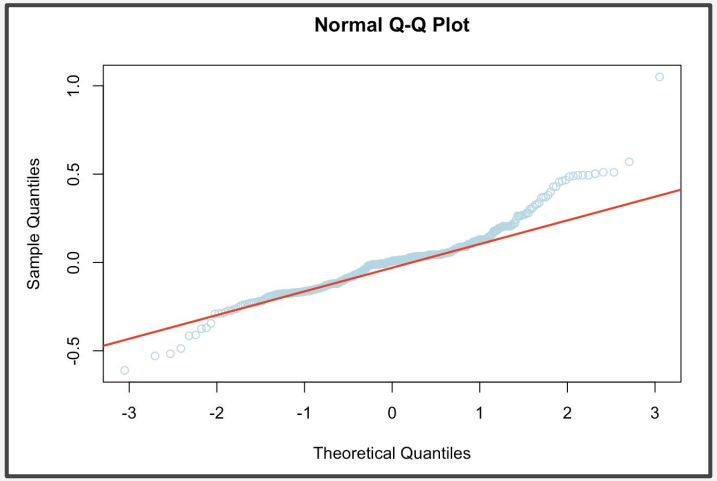
Model: $\text{Overall} = 0.546 + 0.077x_{\text{Prep[2-3 hrs]}} + 0.06x_{\text{Prep[> 3 hrs]}} - 0.084x_{\text{Attend[60\%-79\%]}} - 0.021x_{\text{Attend[80\%-100\%]}} - 0.128x_{\text{Attend[< 40\%]}} + 0.836x_{\text{Last}}$

Adj. R^2 = 0.9102

- *Most* coefficients only deviated around ±0.02 or less and kept their signage
  - last semester still contributing the most towards predicting the overall GPA
    - Coefficient increased by 0.04
- R^2 increased

Re-Checking Assumptions:





Linearity/Homoscedasticity:

- Improvement!
- Only 1 observation stands out
- No observable pattern & centered around 0
    - Appears to pass linearity and homoscedastic conditions

Multicollinearity: VIF of predictors < 5

Normality:

- Reference line more closely at a 45° angle
- Data falls closer to the reference line
    - Now appears relatively normal

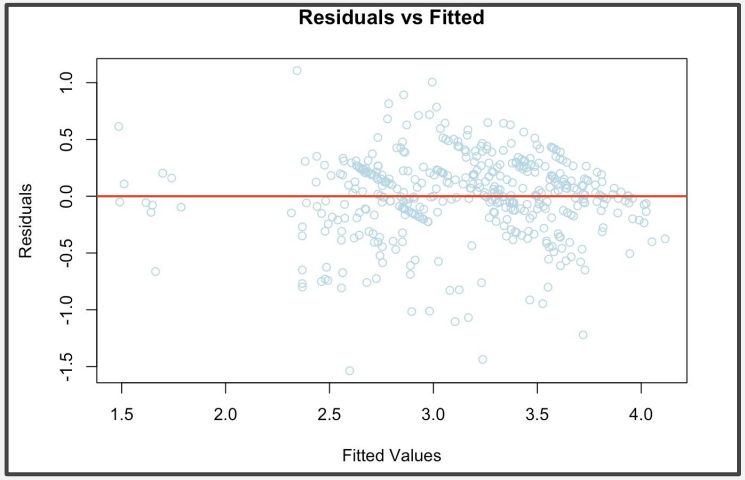
Model #3 - Not Including Last

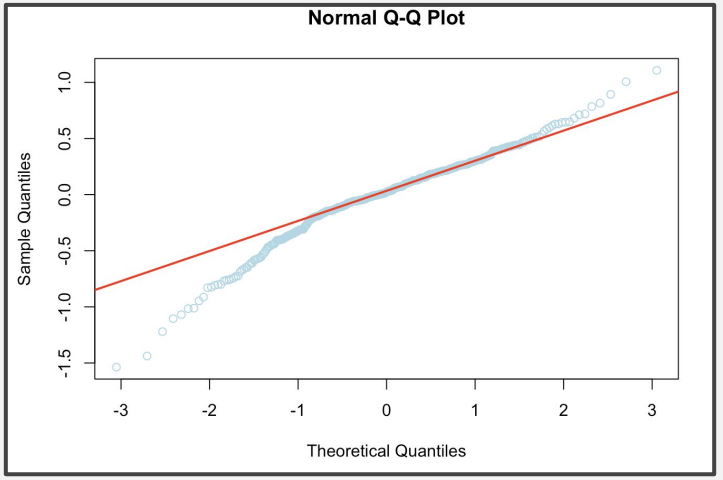
Model: Overall = $1.817 - 0.091x_{\text{Gender Male}} + 0.079x_{\text{HSC}} + 0.078x_{\text{Computer}} + 0.346x_{\text{Prep[2-3 Hours]}} + 0.137x_{\text{Prep[>3hrs]}} + 0.043x_{\text{Gaming[2-3hrs]}} - 0.206x_{\text{Gaming[> 3hrs]}} + 0.141x_{\text{Attend[60%-79%]}} + 0.654x_{\text{Attend[80%-100%]}} - 1.015x_{\text{Attend[<40%]}} - 0.257x_{\text{Job[Yes]}} + 0.137x_{\text{English}}$

Adj. R^2 = 0.63

- Predictors: Gender, HSC, Computer, Preparation, Gaming, Attendance, Job, English
- R^2 significantly decreased, but this was expected → social study
- *Preparation, Gaming, Attendance,* and having a *Job* most contribute to overall GPA
- Having a job on average, decreased GPA

Checking Assumptions:





Linearity/Homoscedasticity:

- No observation drastically stands out
- No observable pattern & centered around 0
  - Appears to pass linearity and homoscedastic conditions

Multicollinearity: VIF of predictors < 5

Normality:

- Reference line relatively close at a 45° angle
- Data falls close to the reference line
  - Appears relatively normal

Limitations & Hindsight

- Data was self reported
  - Students could be portraying themselves with better habits than in reality
- Only looking at Computer Science and Engineering undergraduates
- Having a limited number of samples (440)
  - And only from a single university
- Having discrete data instead of continuous data
  - Would be interesting to collect our own data and have more continuous quantitative data
- Other metrics could be better in predicting overall GPA
  - Perhaps average credit hrs per semester
  - Distance from school

Conclusions

- While all of our predictors had some correlation with the Overall GPA, it wasn't as high as we expected
- It's obvious that the Last Semester GPA is the best predictor of the Overall GPA
- Some of the most important factors were attendance, preparation, and having a job
- If we leave all other predictors the same, having a job dropped your overall GPA by 0.24
- It would be interesting to use a decision tree on this data since we have numerous categorical variables