Student Performance Analysis

Amandaliss Dropik

The Problem

- In general, there seems to be a division between male and female academic performance
- We want to figure out which sex performs best, and why
  - At a glance, there would seem to be many contributions to academic failure or success
- Eventually, we want to be able to predict grades based on these factors
- This information can be used by universities to makes positive changes

Background

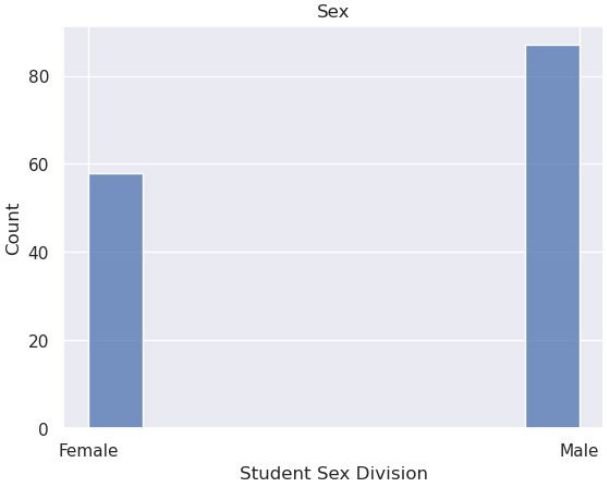
- There have been many studies on the differences between males and females in academics
  - Many different conclusions as well
- After reading some papers, it is generally accepted that females tend to do better academically
  - This is reflected in graduation rates, drop out rates, workforce statistics and more
  - If we look at specifics, there are sectors (like the sciences) where men will have more success
- Main piece of literature referenced for this project was *Conscientiousness as a Predictor of the Gender Gap in Academic Achievement* by Verbree et al (2022)

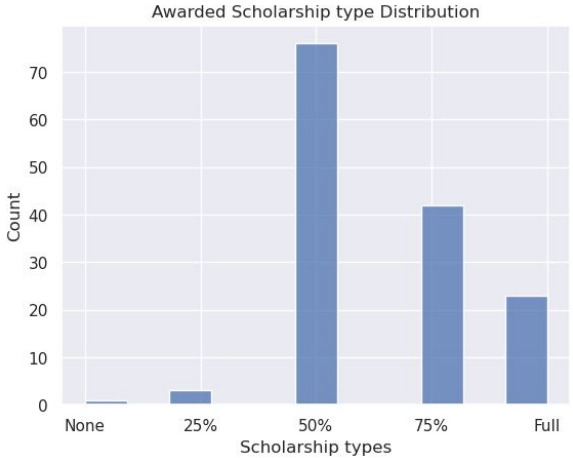
Data Set

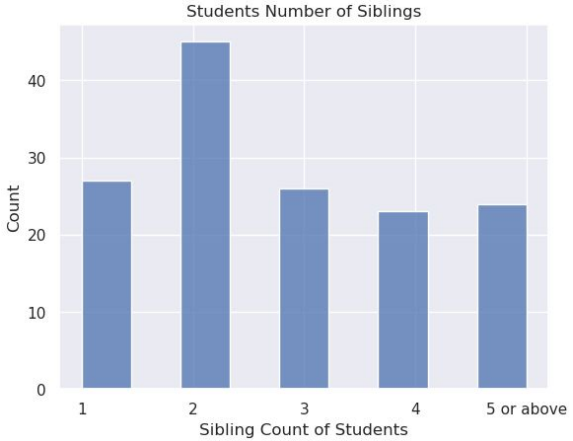
- Students Performance by Joakim Arvidsson
- Dataset of 145 students
- 33 features
  - Consists of personal and academic information
- Chosen because of the wide range of information taken from students
- Output is grade for each student
- Dataset came preprocessed and clean
  - Minimal changes done like removing unnecessary columns

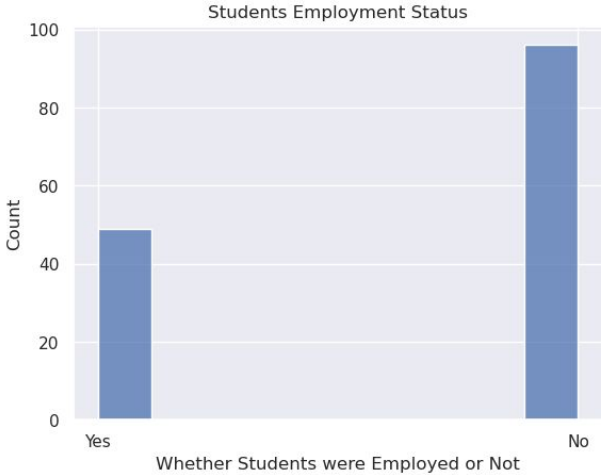
Questions To be Answered with Analysis

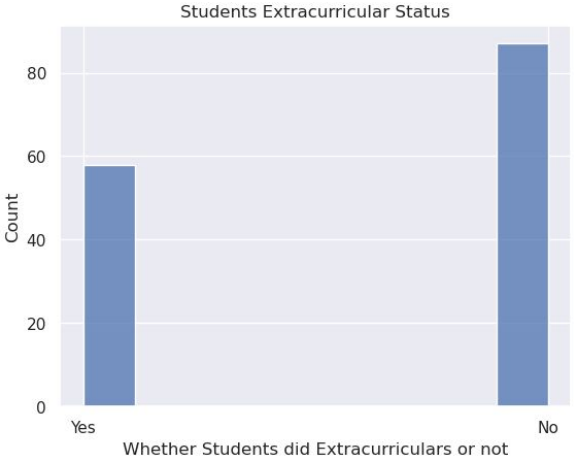
1. Which sex performs better in school?
2. Does having a commitment outside of school negatively impact your grade?
3. How impactful is cramming for exams on your grade?
4. Is it better to study more than 5 hours a week?
5. Does attending seminars/conferences related to your department impact your grade?
6. Is attending class necessary for a good grade
7. Does studying with classmates negatively or positively impact your grade?
8. How much does note-taking during class determine your success?
9. How much does listening during class determine your success?

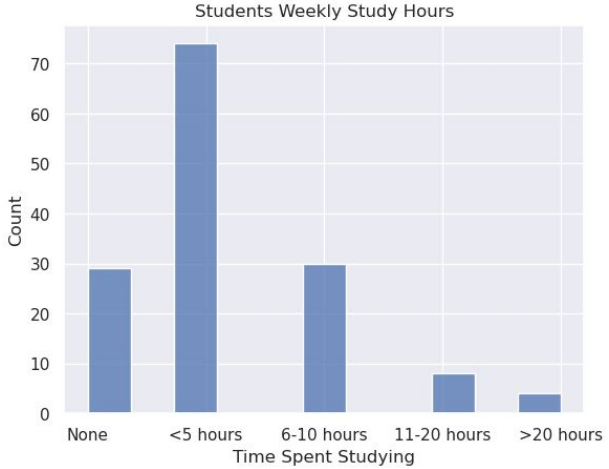

Sex
Count
0
20
40
60
80
Female
Male
Student Sex Division

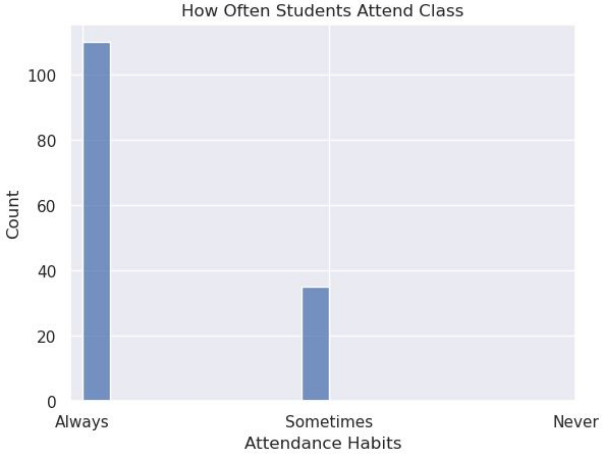

Awarded Scholarship type Distribution
Count
0
10
20
30
40
50
60
70
None
25%
50%
75%
Full
Scholarship types

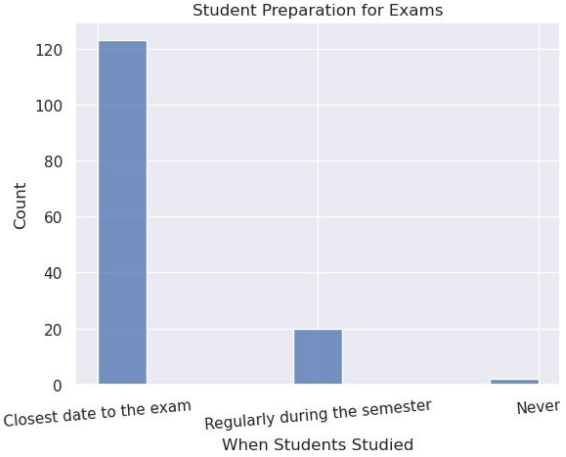

Students Number of Siblings
Count
0
10
20
30
40
1
2
3
4
5 or above
Sibling Count of Students

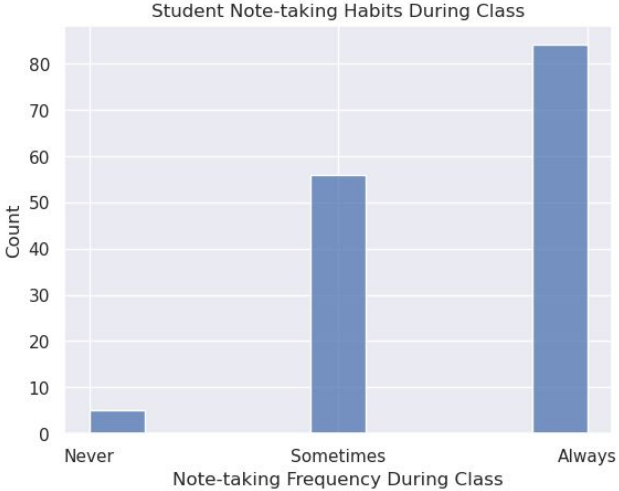

Students Employment Status
Count
0
20
40
60
80
100
Yes
No
Whether Students were Employed or Not

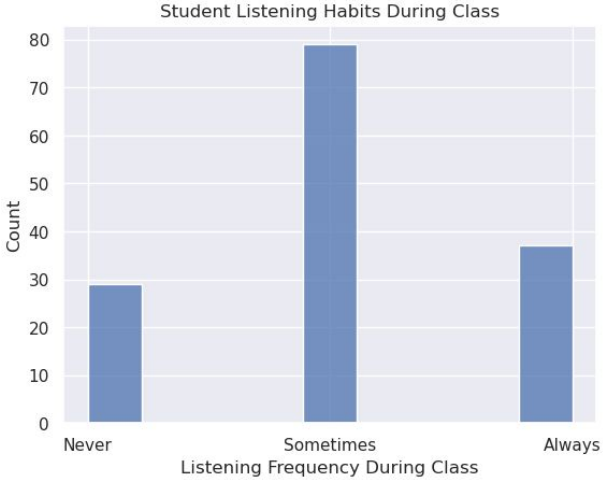

Students Extracurricular Status
Count
0
20
40
60
80
Yes
No
Whether Students did Extracurriculars or not

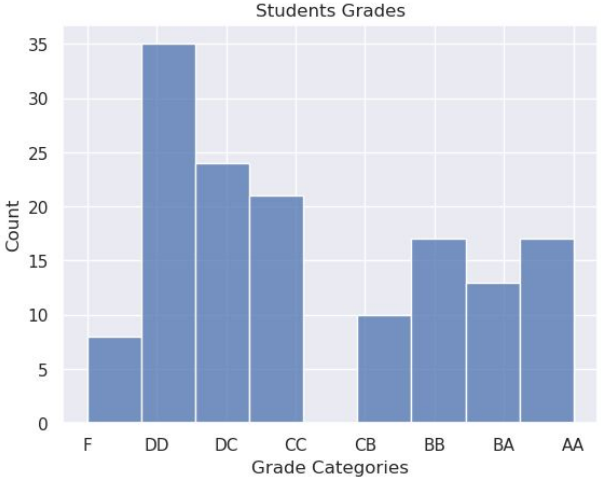

Students Weekly Study Hours
Count
0
10
20
30
40
50
60
70
None
<5 hours
6-10 hours
11-20 hours
>20 hours
Time Spent Studying

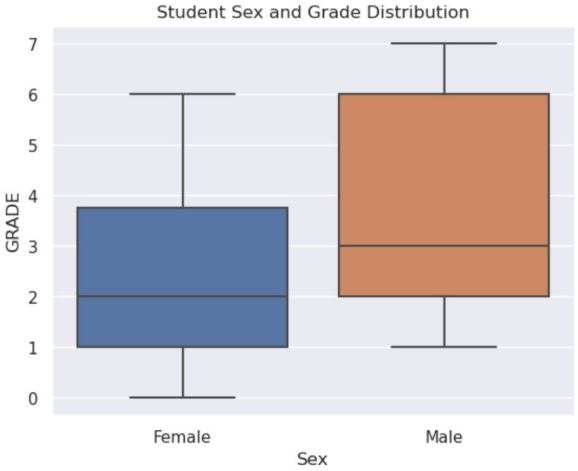

How Often Students Attend Class
Count
0
20
40
60
80
100
Always
Sometimes
Never
Attendance Habits

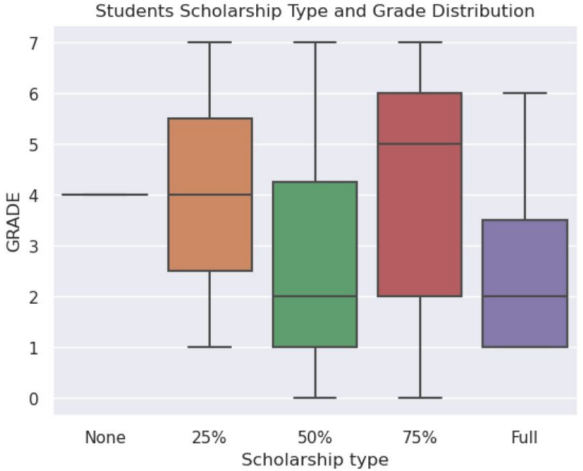

Student Preparation for Exams
Count
0
20
40
60
80
100
120
Closest date to the exam
Regularly during the semester
Never
When Students Studied

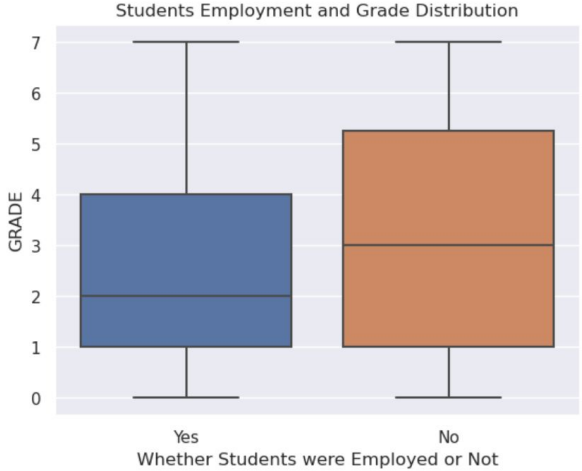

Student Note-taking Habits During Class
Count
0
10
20
30
40
50
60
70
80
Never
Sometimes
Always
Note-taking Frequency During Class

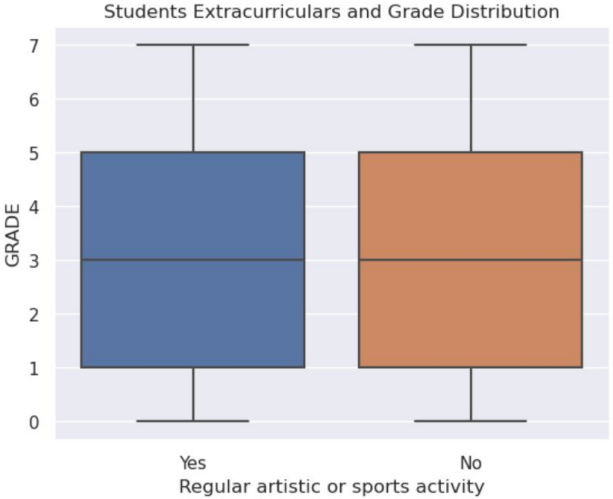

Student Listening Habits During Class
Count
0
10
20
30
40
50
60
70
80
Never
Sometimes
Always
Listening Frequency During Class

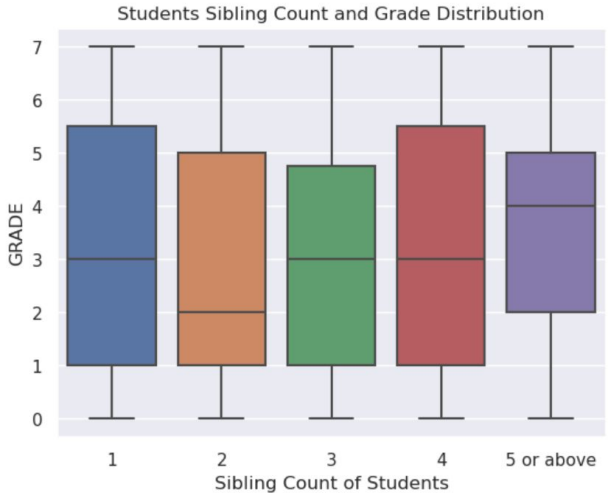

Students Grades
Count
0
5
10
15
20
25
30
35
F
DD
DC
CC
CB
BB
BA
AA
Grade Categories

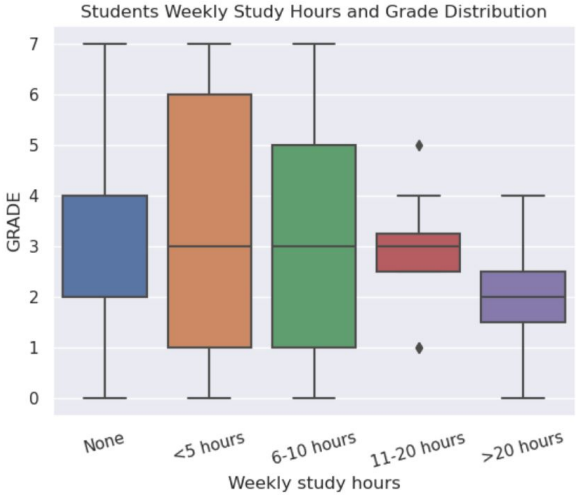

Student Sex and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
Female
Male
Sex

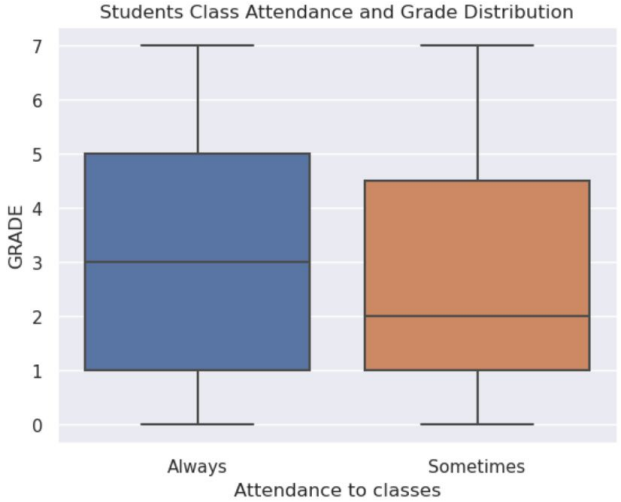

Students Scholarship Type and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
None
25%
50%
75%
Full
Scholarship type

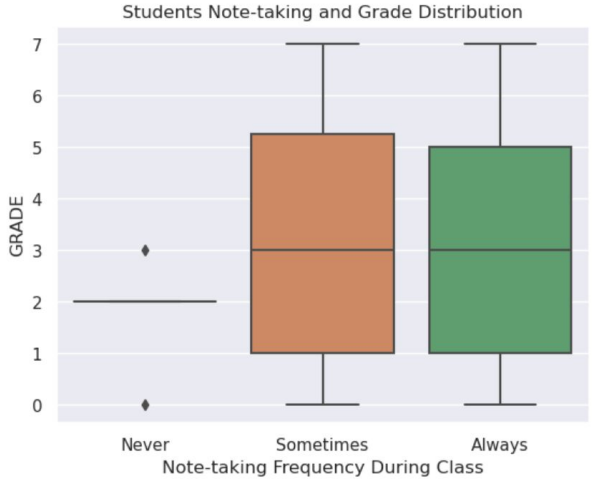

Students Employment and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
Yes
No
Whether Students were Employed or Not

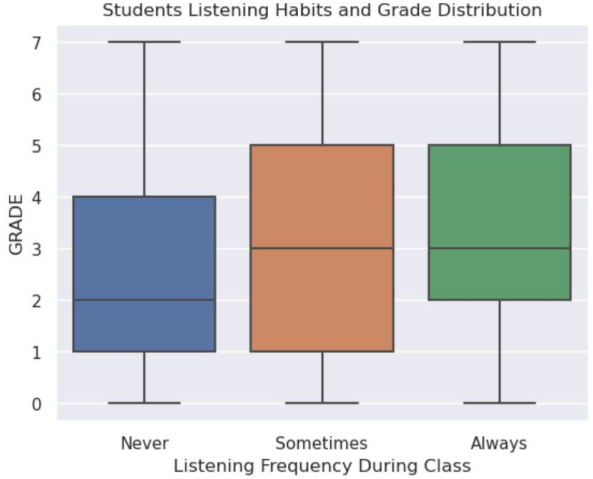

Students Extracurriculars and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
Yes
No
Regular artistic or sports activity

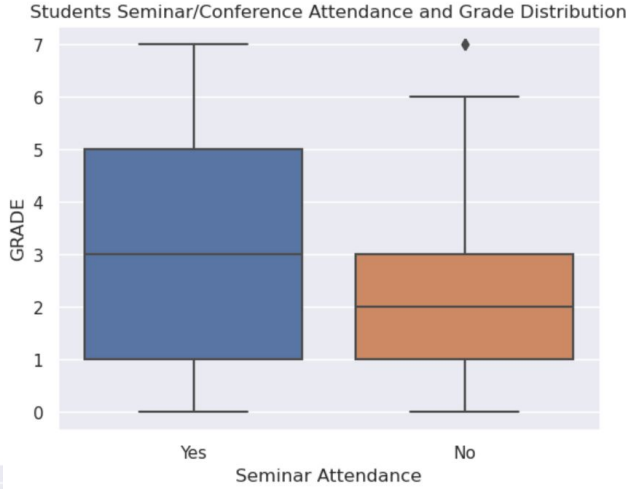

Students Sibling Count and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
1
2
3
4
5 or above
Sibling Count of Students

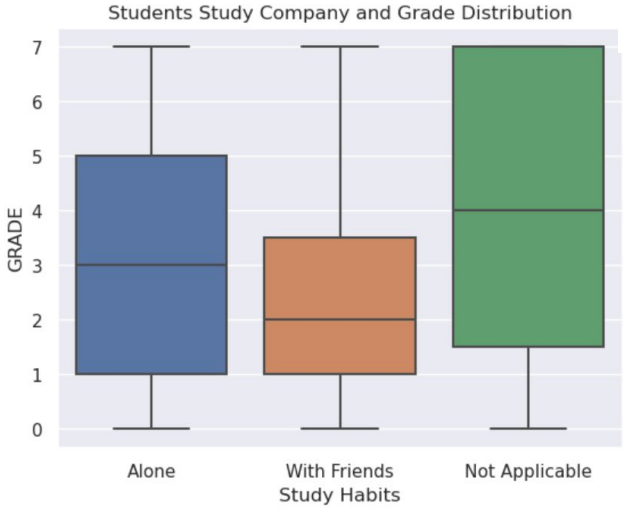

Students Weekly Study Hours and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
None
<5 hours
6-10 hours
11-20 hours
>20 hours
Weekly study hours

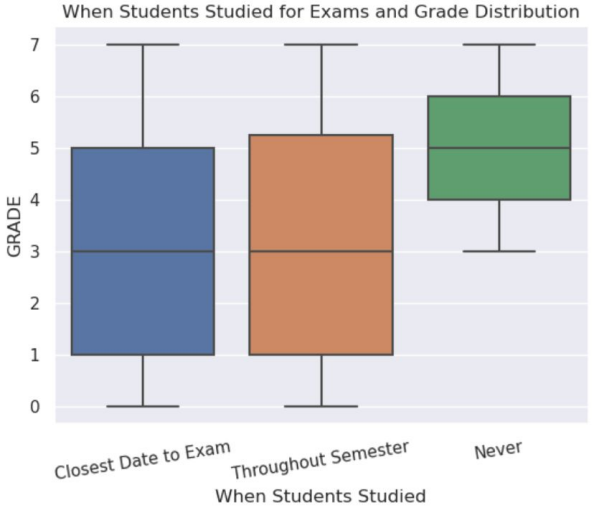

Students Class Attendance and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
Always
Sometimes
Attendance to classes


Students Note-taking and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
Never
Sometimes
Always
Note-taking Frequency During Class


Students Listening Habits and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
Never
Sometimes
Always
Listening Frequency During Class


Students Seminar/Conference Attendance and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
Yes
No
Seminar Attendance


Students Study Company and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
Alone
With Friends
Not Applicable
Study Habits


When Students Studied for Exams and Grade Distribution
GRADE
0
1
2
3
4
5
6
7
Closest Date to Exam
Throughout Semester
Never
When Students Studied

Answers to earlier questions

1. This data set suggests males do better
2. There is only a slight impact if you are employed, and none with extracurriculars
3. Seems to have no significant impact on grade
4. Studying for longer than 5 hours appears to lower grade
5. Attending seminars does positively impact grade
6. This data set suggest there is no significance towards grade
7. Studying with others has a moderate significance
8. Seems to positively impact grade
9. Might slightly impact grade

Methodology

- Random Forest Regressor
  - RMSE of ~2.057
  - Pearson's Correlation
    - Sex: 0.3355 (moderate correlation)
    - Cumulative grade point average in the last semester: 0.3155 (moderate correlation)
    - Expected cumulative grade point average in the graduation: 0.2486 (weak correlation)
  - Feature Importance
    - Cumulative grade point average in the last semester: 0.1146
    - Father's education: 0.0683
    - Sex: 0.0657

Methodology

- Random Forest Classifier explanations
  - Accuracy: .2273
  - Feature Importance:
    - Sex: 0.0862
    - Cumulative grade point average in the last semester: 0.2642
    - Expected Cumulative grade point average in the graduation: 0.1933
    - Reading frequency: 0.1405
    - Attendance to classes: 0.0893
    - Weekly study hours: 0.2265
  -

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	5
1	0.32	0.60	0.41	10
2	1.00	0.11	0.20	9
3	0.33	0.17	0.22	6
4	0.00	0.00	0.00	3
5	0.17	0.33	0.22	3
6	0.00	0.00	0.00	5
7	0.14	0.33	0.20	3
accuracy			0.23	44
macro avg	0.24	0.19	0.16	44
weighted avg	0.34	0.23	0.19	44

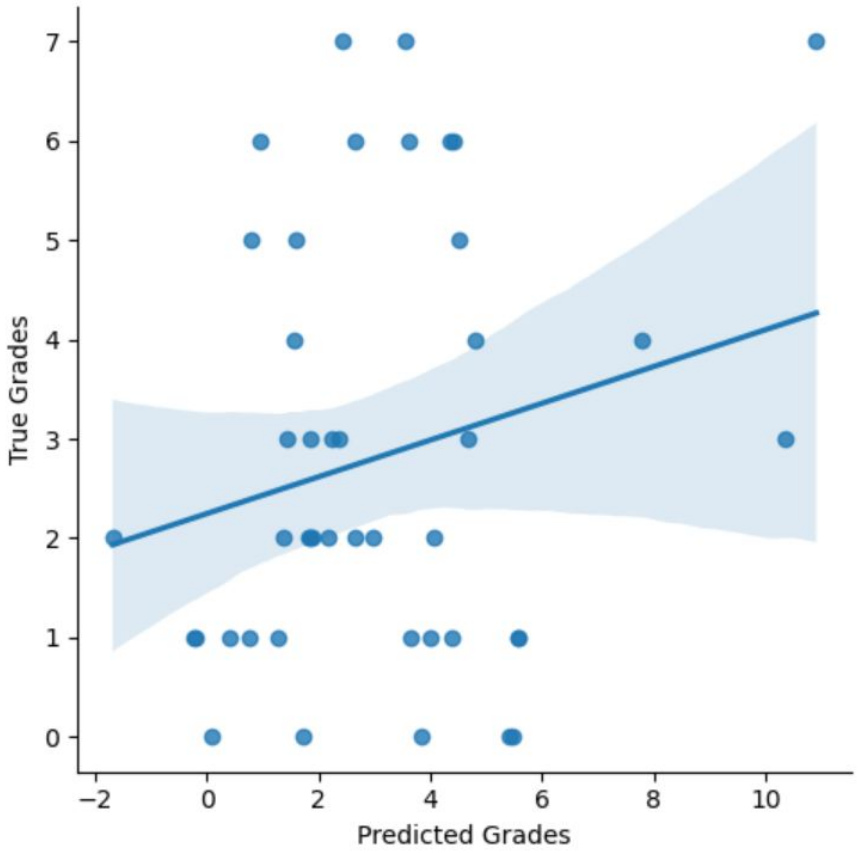
Confusion Matrix:

```
[[0 3 0 0 0 0 2 0]
 [1 6 0 0 0 0 0 3]
 [2 3 1 0 2 1 0 0]
 [0 2 0 1 1 2 0 0]
 [0 1 0 0 0 1 0 1]
 [0 2 0 0 0 1 0 0]
 [0 1 0 1 0 1 0 2]
 [0 1 0 1 0 0 0 1]]
```

Methodology

- Linear Regression
  - RMSE: ~2.927
  - R-squared: -0.8801

	True Grades	Predicted Grades
69	5	0.795585
140	5	1.608887
27	1	0.399065
19	3	1.860659
42	1	3.636620
117	1	-0.216083
126	3	2.236299
108	6	0.939197



Results

- Some of the features I thought would impact the grades did not do so as much as I thought, or at all
- Random Forest Regressor was the best model for this data
  - With that said, still did not perform that well
- After finishing this project, I can conclude that the dataset is probably the issue here

Sources

Joakim Arvidsson. (2023). Students Performance (Version 2) [Datafile and code book]. Retrieved from https://www.kaggle.com/datasets/joebeachcapital/students-performance/data

Verbree, A. R., Hornstra, L., Maas, L., & Wijngaards-de Meij, L. (2023). Conscientiousness as a Predictor of the Gender Gap in Academic Achievement. *Research in higher education*, 64(3), 451–472. https://doi.org/10.1007/s11162-022-09716-5