Identifying Academically At-Risk Students:

A Data-Driven Early Intervention System

Valentina Nguyen

Table of contents

Problem

&

Case Scenario

Problem & Case Scenario

- Academic success is a critical factor in shaping students’ opportunities for success, yet many high schools struggle to maintain satisfactory performance.

- Identifying at-risk students early allows for timely intervention, helping prevent failure and ensuring long-term academic success.

- The insights from this study will benefit students, teachers, parents, and policymakers, providing actionable strategies for improving student retention, engagement, and overall educational success.

Data

Data Source & Description

Data Source: Synthetic dataset generated and posted to Kaggle for educational purposes.

Data size: 2,392 students

Demographics	Age, Gender, Ethnicity
Extracurricular Activities Participation	Sports, Music, Volunteering, Other
Study Habits	Study Time Weekly, Absences, Tutoring The study habits of a student are broken down into three columns: “StudyTimeWeekly”, “Absences”, “Tutoring”.
Parent Information	Parental Education, Parental Involvement
GPA/Grade Class	Grade class takes the GPA and classifies it by letter grade with 0 equivalent to A and 4 equivalent to F.

Feature Variables

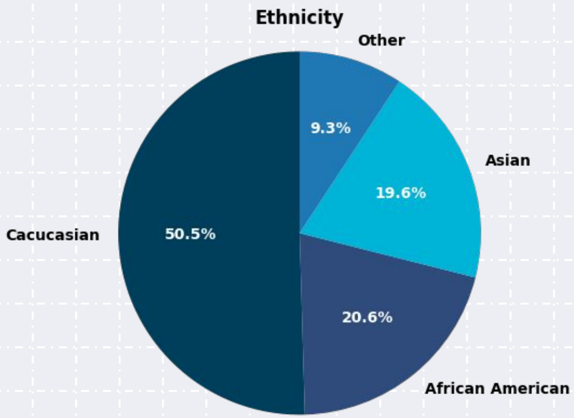
Age (Grade Level)	Freshman (15), Sophomore (16), Junior (17), Senior (18)
Gender	Male (0), Female (1)
Ethnicity	Caucasian (0), African American (1) , Asian (2), Other (3)
Parental Education	None (0), High School (1), Some College (2), Bachelor's (3), Higher (0)
Absences	Absences are ranged from 0-30 days
Study Time	Hours spent studying per week, ranges from (0-20)
Tutoring	No (0), Yes (1)
Parental Support	None (0), Low (1), Moderate (2), High (3), Very High (4)
Extracurricular	Indicates the students' participation in an extracurricular activity other than those listed below: No (0), Yes (1),
Sports	No (0), Yes (1)
Music	No (0), Yes (1)
Volunteering	No (0), Yes (1)

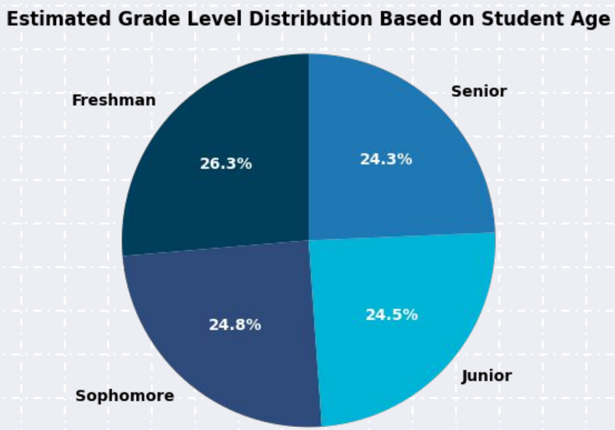
* Since the dataset does not provide grade levels, we have made an assumption that the grade level progression follows the typical age to grade model. This mapping serves as a proxy for grade level to identify the grade levels across the dataset.

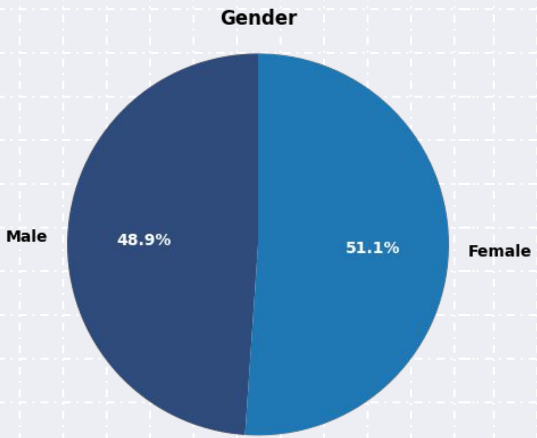
The "Extracurricular" column denotes whether or not a student participates in an activity other than sports, music, or volunteering and is considered to be an "Other" category.

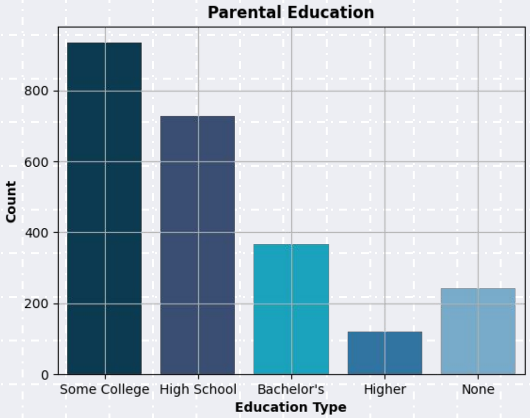
Insights

EDA

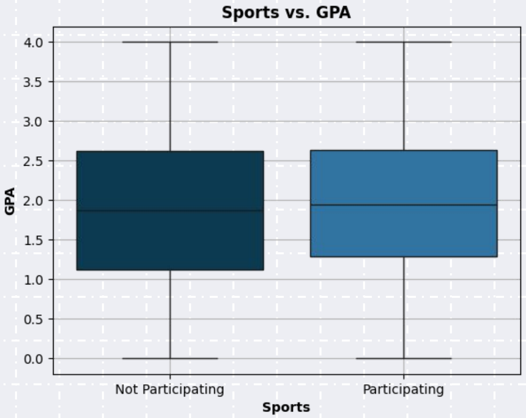

Ethnicity
Other
9.3%
Asian
19.6%
Cacucasian
50.5%
20.6%
African American


Estimated Grade Level Distribution Based on Student Age
Freshman
26.3%
Senior
24.3%
24.8%
24.5%
Junior
Sophomore

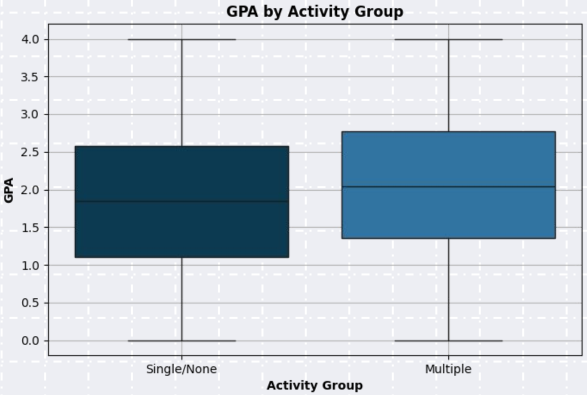

Gender
Male
48.9%
51.1%
Female



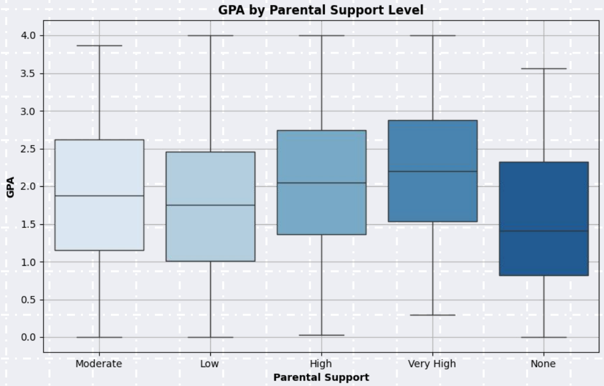
Parental Education
Count
0
200
400
600
800
Some College
High School
Bachelor's
Higher
None
Education Type



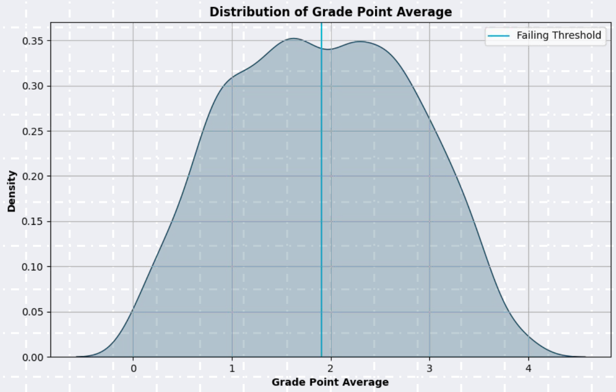
Sports vs. GPA
GPA
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Not Participating
Participating
Sports

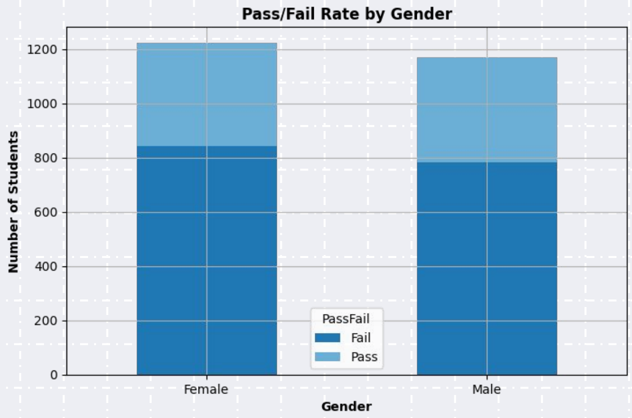


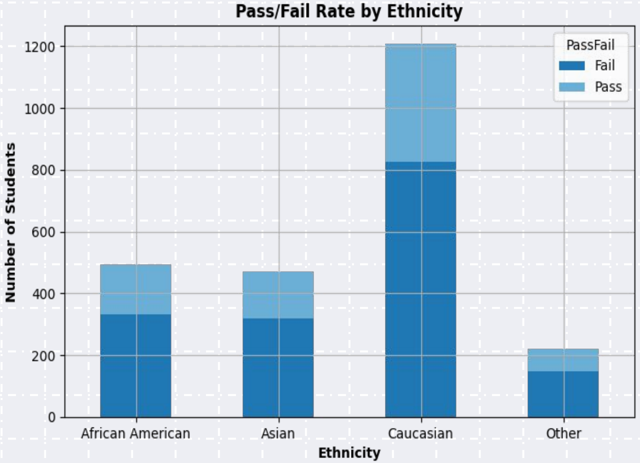
GPA by Activity Group
GPA
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Single/None
Multiple
Activity Group

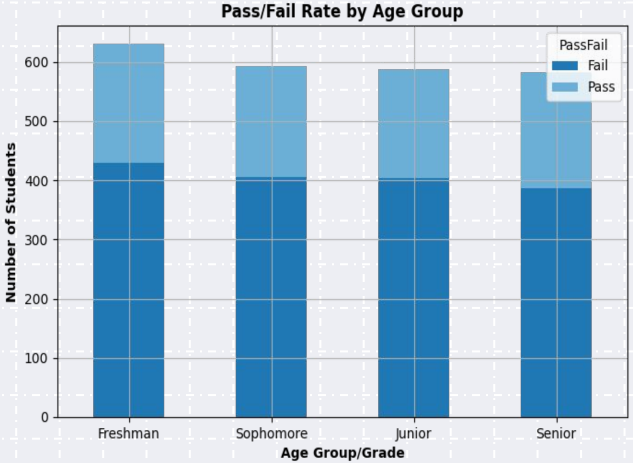


GPA by Parental Support Level
GPA
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Moderate
Low
High
Very High
None
Parental Support


Distribution of Grade Point Average
Failing Threshold
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
1
2
3
4
Grade Point Average


Pass/Fail Rate by Gender
Number of Students
0
200
400
600
800
1000
1200
PassFail
Fail
Pass
Female
Male
Gender

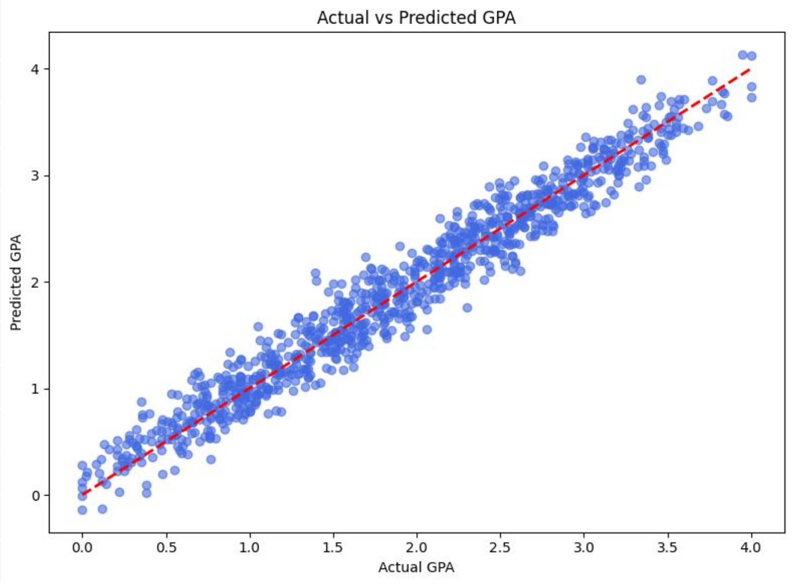

Pass/Fail Rate by Ethnicity
Number of Students
0
200
400
600
800
1000
1200
PassFail
Fail
Pass
African American
Asian
Caucasian
Other
Ethnicity


Pass/Fail Rate by Age Group
Number of Students
0
100
200
300
400
500
600
PassFail
Fail
Pass
Freshman
Sophomore
Junior
Senior
Age Group/Grade

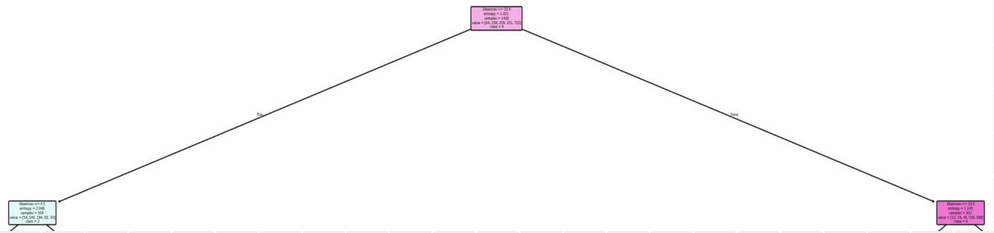
Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    GPA   R-squared:                       0.953
Model:                            OLS   Adj. R-squared:                  0.952
Method:                 Least Squares   F-statistic:                     2865.
Date:                Sun, 13 Apr 2025   Prob (F-statistic):               0.00
Time:                        20:11:07   Log-Likelihood:                 282.57
No. Observations:                1435   AIC:                            -543.1
Df Residuals:                    1424   BIC:                            -485.2
Df Model:                          10
Covariance Type:            nonrobust
=====================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               2.4994      0.024    104.858      0.000       2.453       2.546
StudyTimeWeekly     0.0294      0.001     31.344      0.000       0.028       0.031
Absences           -0.0999      0.001   -159.272      0.000      -0.101      -0.099
Tutoring            0.2499      0.011     21.790      0.000       0.227       0.272
Support_Low         0.1758      0.022      8.037      0.000       0.133       0.219
Support_Moderate    0.3110      0.021     14.885      0.000       0.270       0.352
Support_High        0.4730      0.021     22.574      0.000       0.432       0.514
Support_VeryHigh    0.6318      0.025     25.611      0.000       0.583       0.680
Extracurricular     0.1965      0.011     18.041      0.000       0.175       0.218
Sports              0.1966      0.011     17.234      0.000       0.174       0.219
Music               0.1457      0.013     11.047      0.000       0.120       0.172
==============================================================================
Omnibus:                        2.650   Durbin-Watson:                   2.049
Prob(Omnibus):                  0.266   Jarque-Bera (JB):                2.592
Skew:                          -0.103   Prob(JB):                        0.274
Kurtosis:                       3.025   Cond. No.                         159.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```



Decision Tree



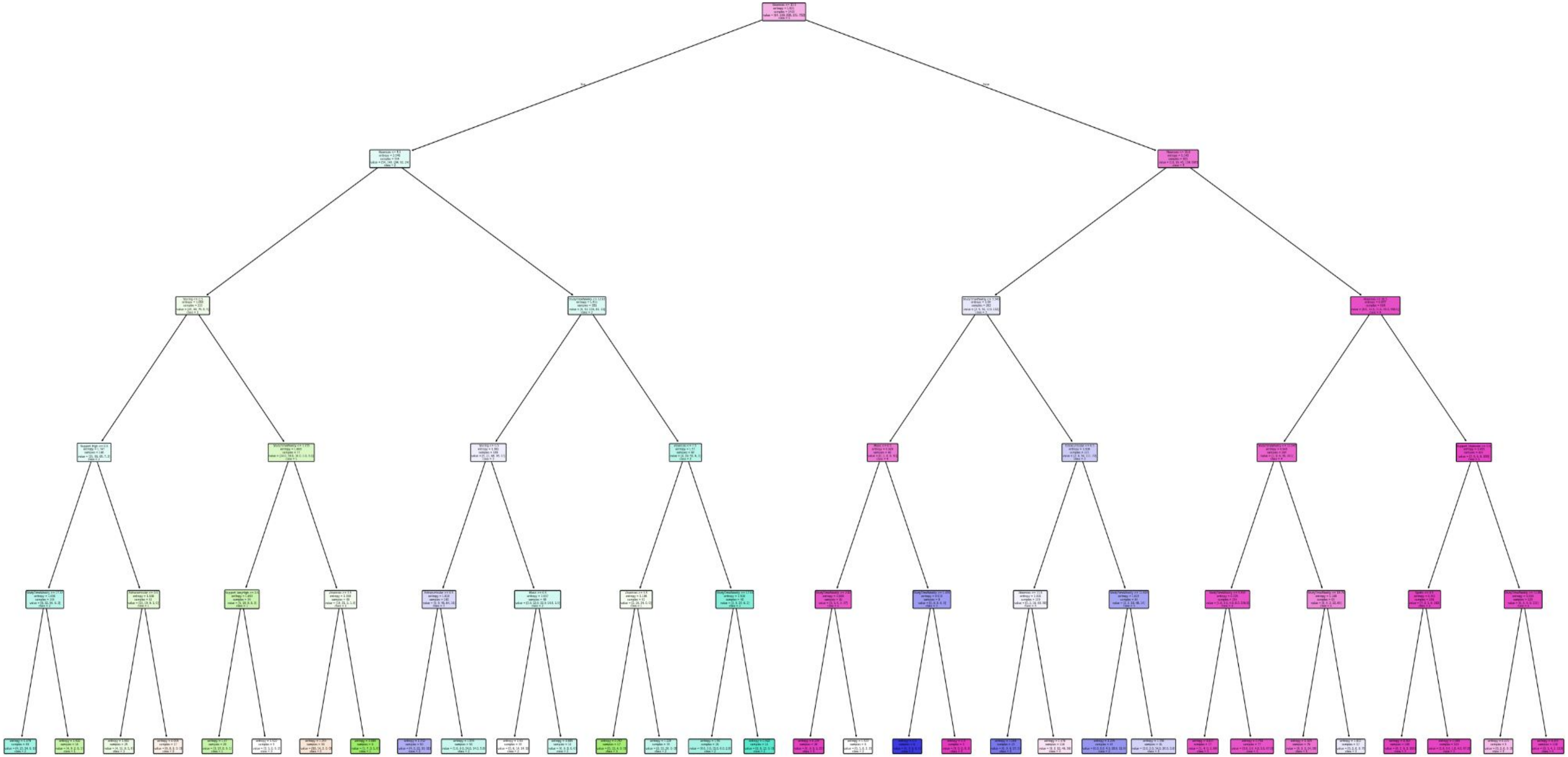
	precision	recall	f1-score	support
0	0.47	0.40	0.43	43
1	0.58	0.27	0.37	110
2	0.50	0.66	0.57	152
3	0.45	0.36	0.40	163
4	0.84	0.92	0.88	489
accuracy			0.69	957
macro avg	0.57	0.52	0.53	957
weighted avg	0.67	0.69	0.67	957

→ First splits based on absences

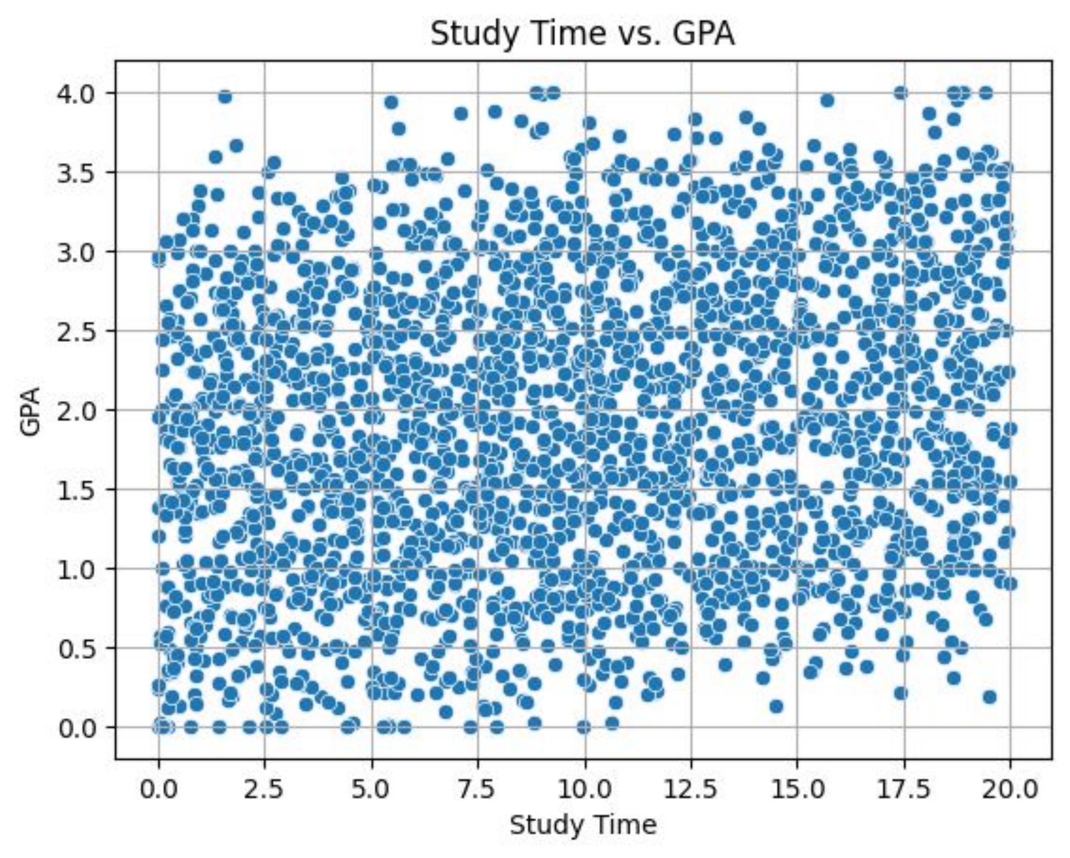
→ Overall accuracy of 69% and weighted F1 score of 0.67

→ Model did well predicting class 4 (F), struggled with other classes (A,B,C,D)

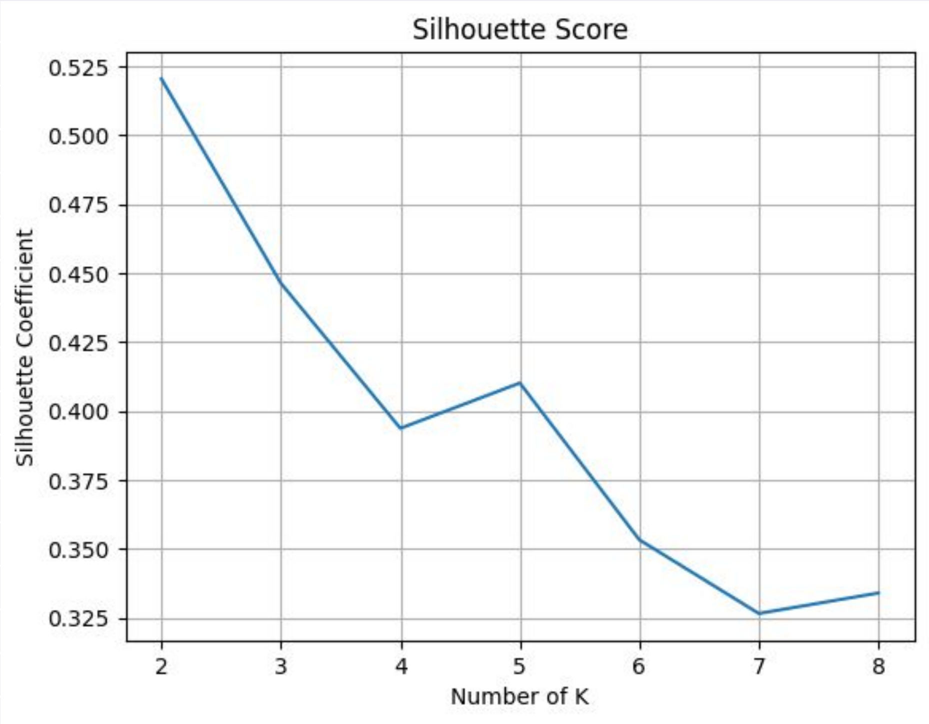
→ Over half of all students are in class 4 which may bias the results

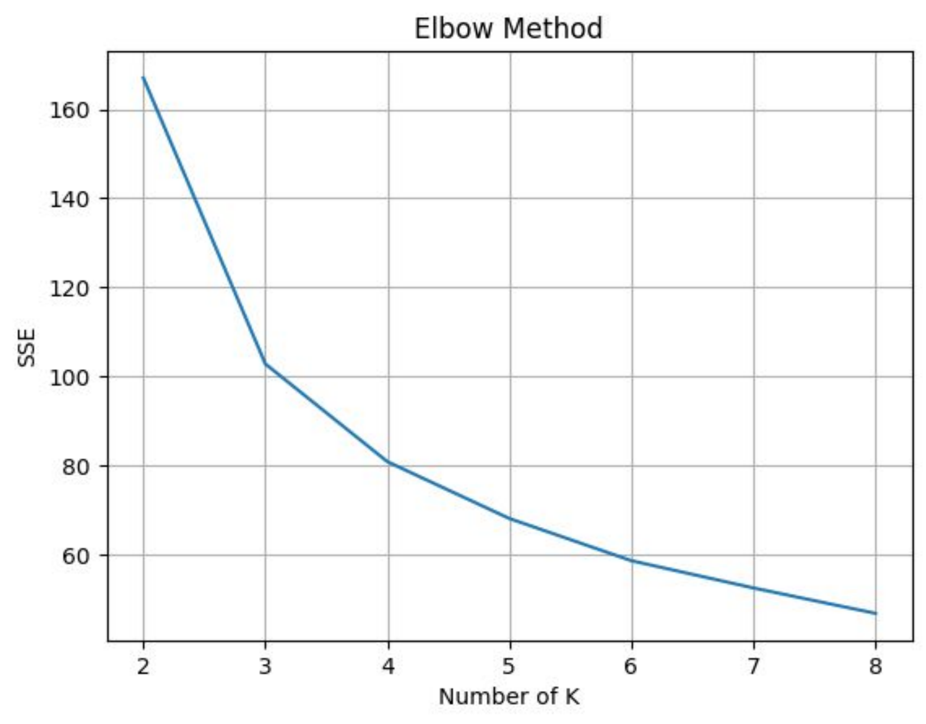


K-Means Clustering



K-Means Clustering





K-Means Clustering

```
[33] #Cluster 0
     sum(m2_lb == 0)
```

```
np.int64(1185)
```

```
[34] #Cluster 1
     sum(m2_lb == 1)
```

```
np.int64(1207)
```

```
[28] #Cluster 0
     sum(m4_lb == 0)
```

```
np.int64(610)
```

```
[29] #Cluster 1
     sum(m4_lb == 1)
```

```
np.int64(618)
```

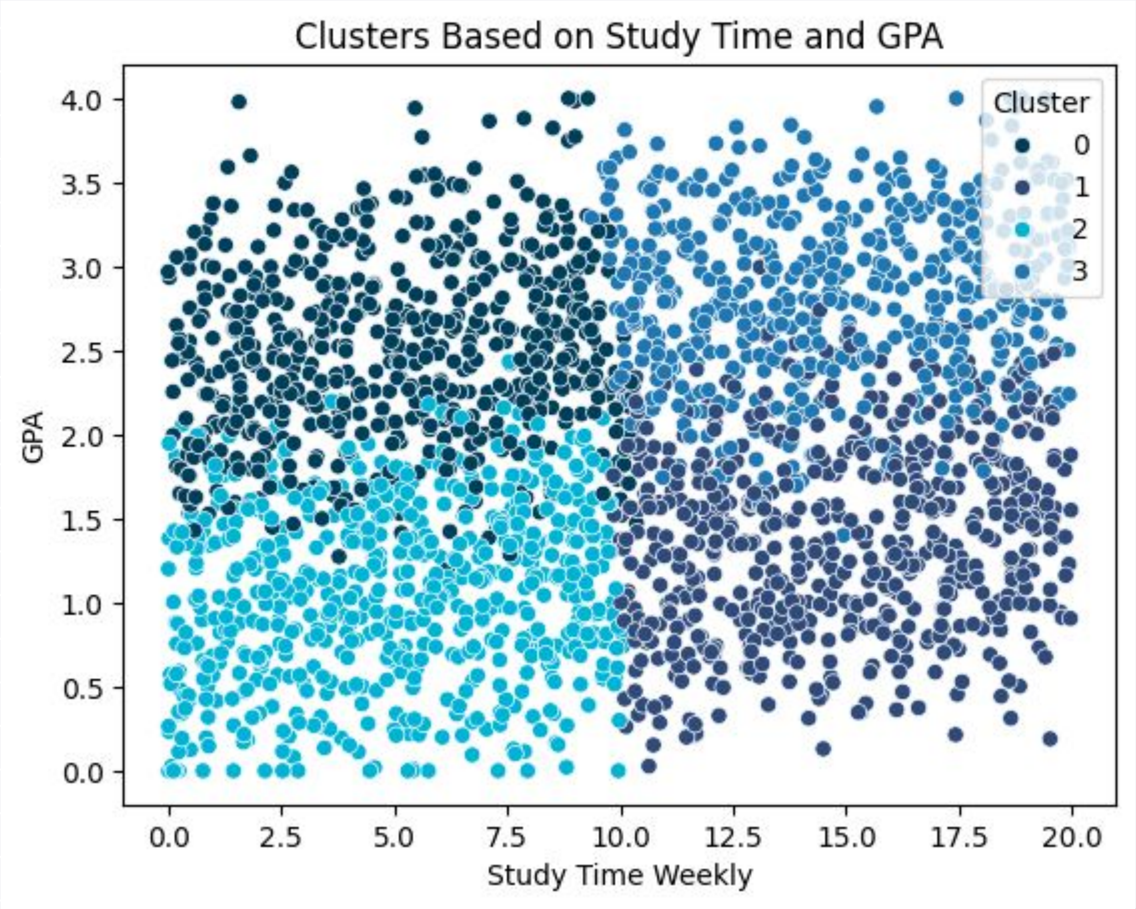
```
[30] #Cluster 2
     sum(m4_lb == 2)
```

```
np.int64(630)
```

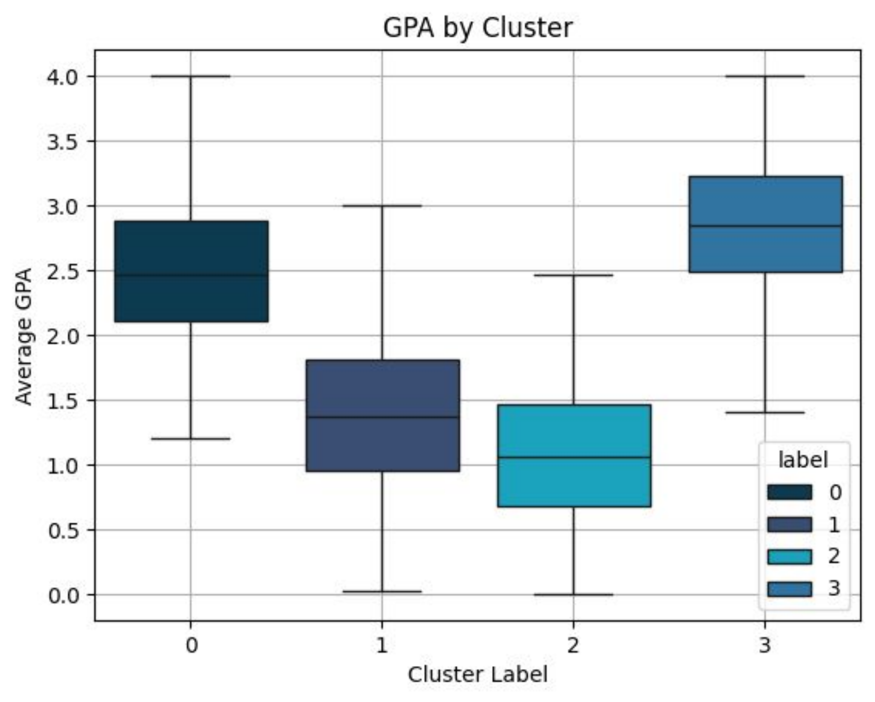
```
[31] #Cluster 3
     sum(m4_lb == 3)
```

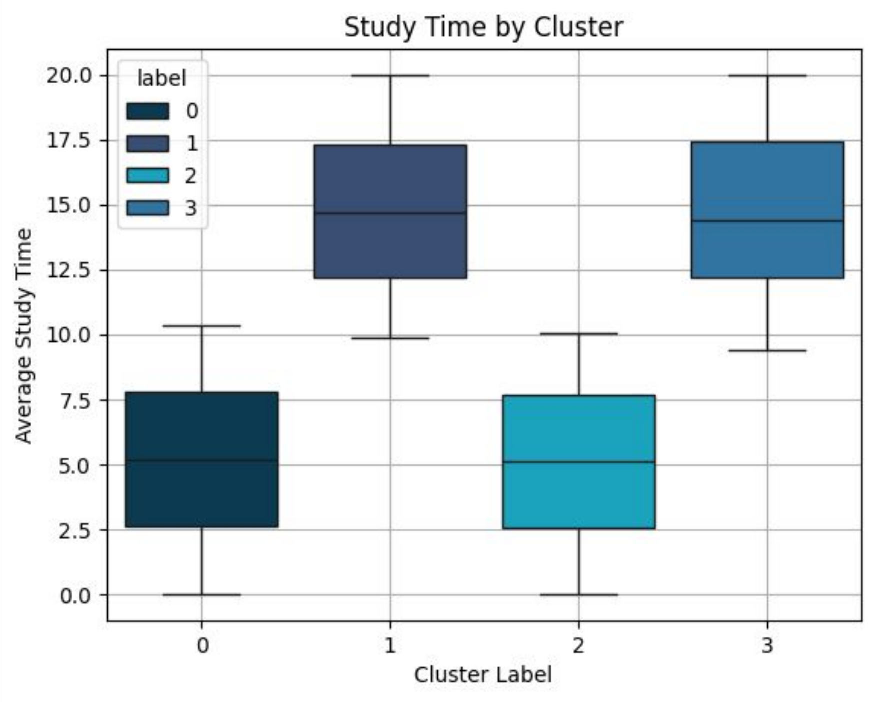
```
np.int64(534)
```

K-Means Clustering

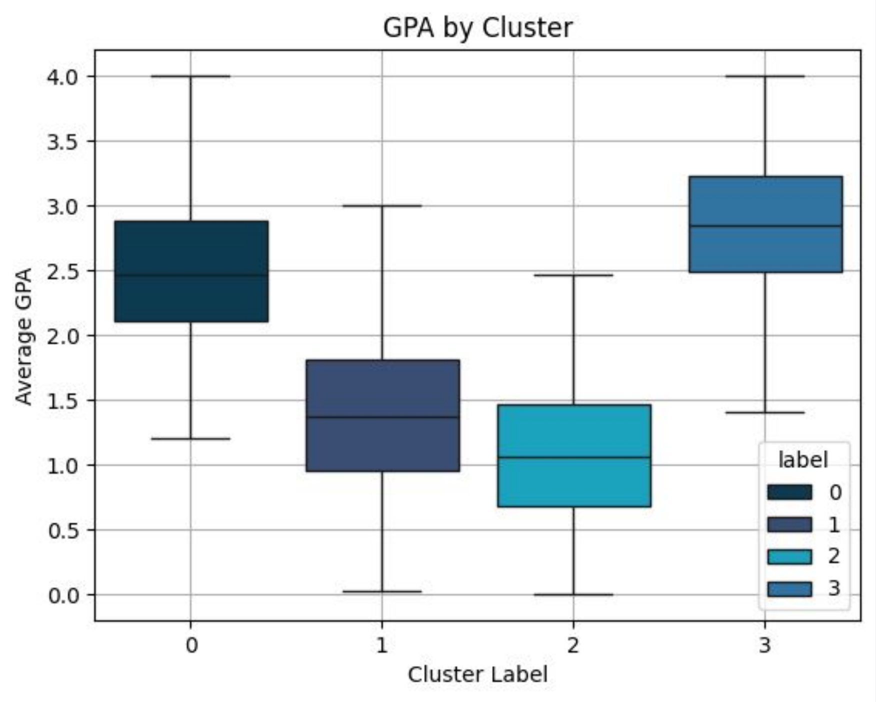


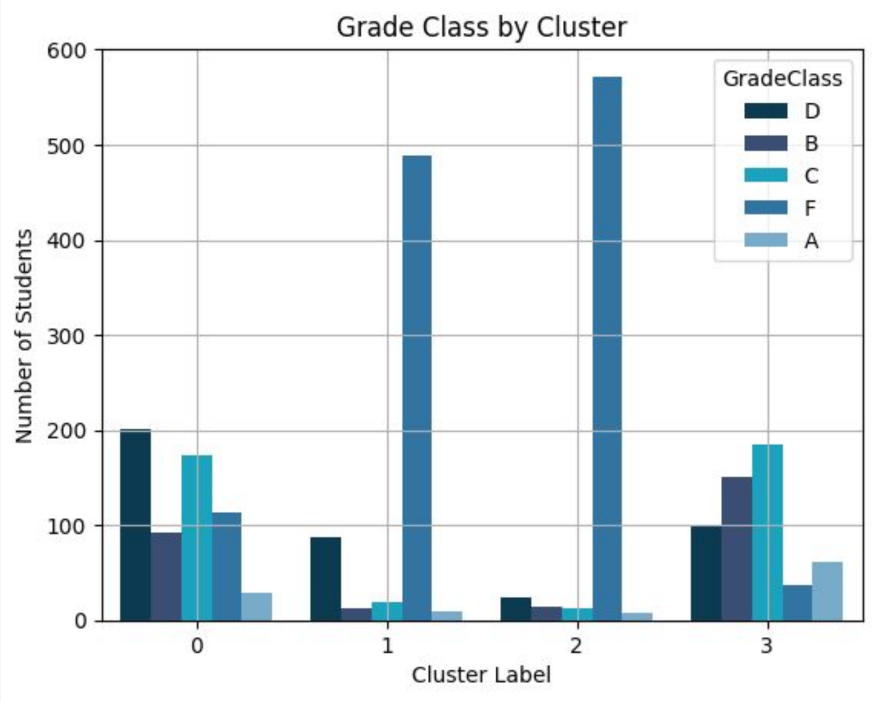
K-Means Clustering



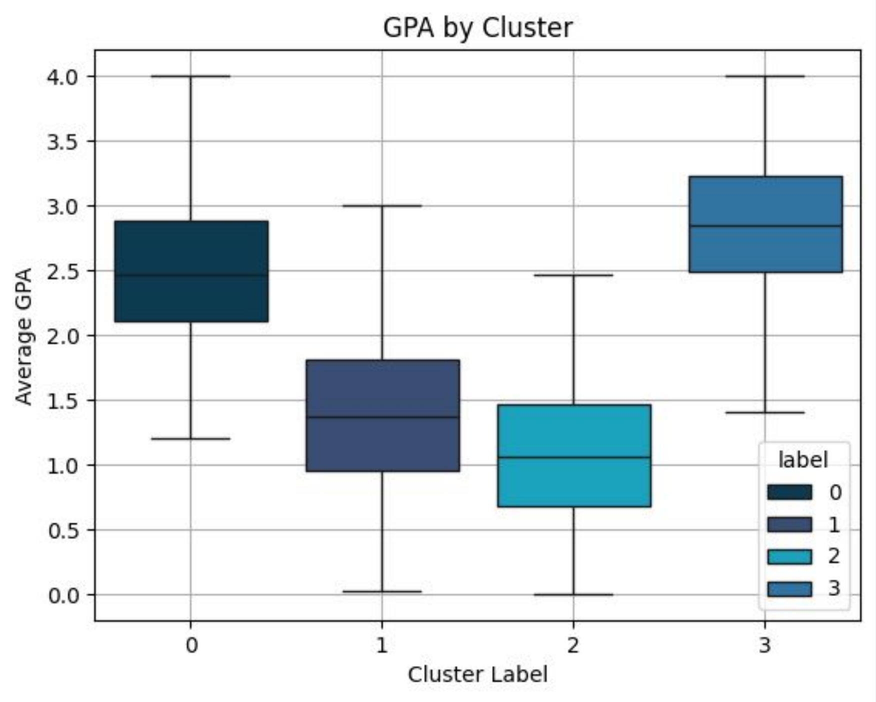


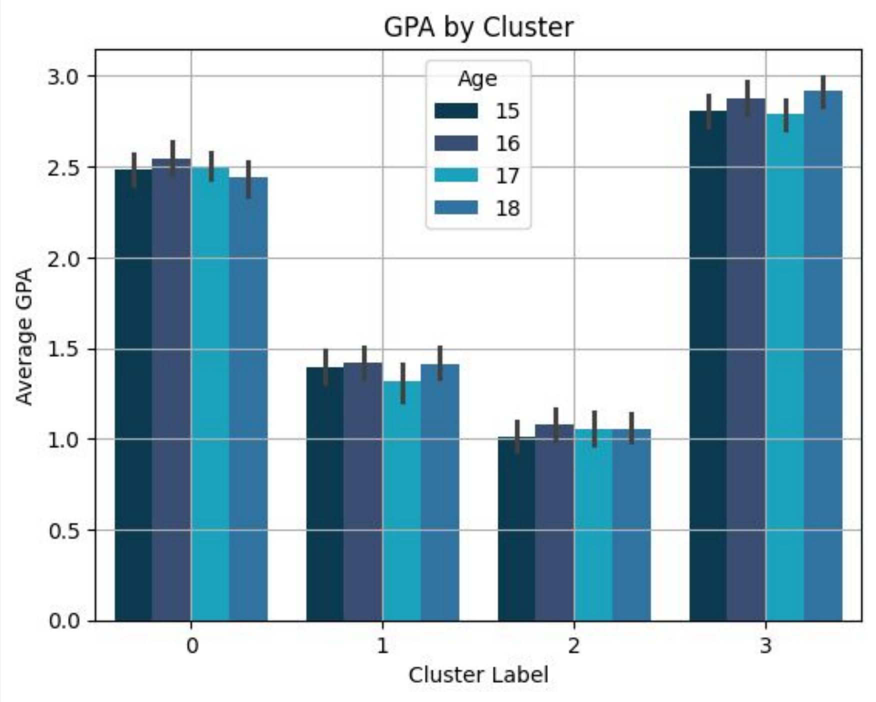
K-Means Clustering



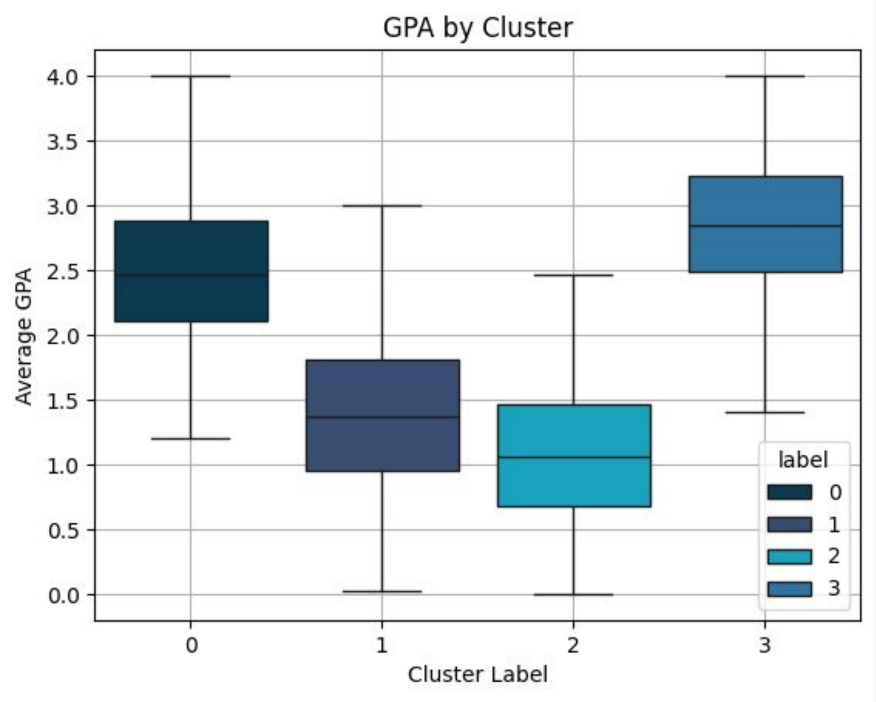


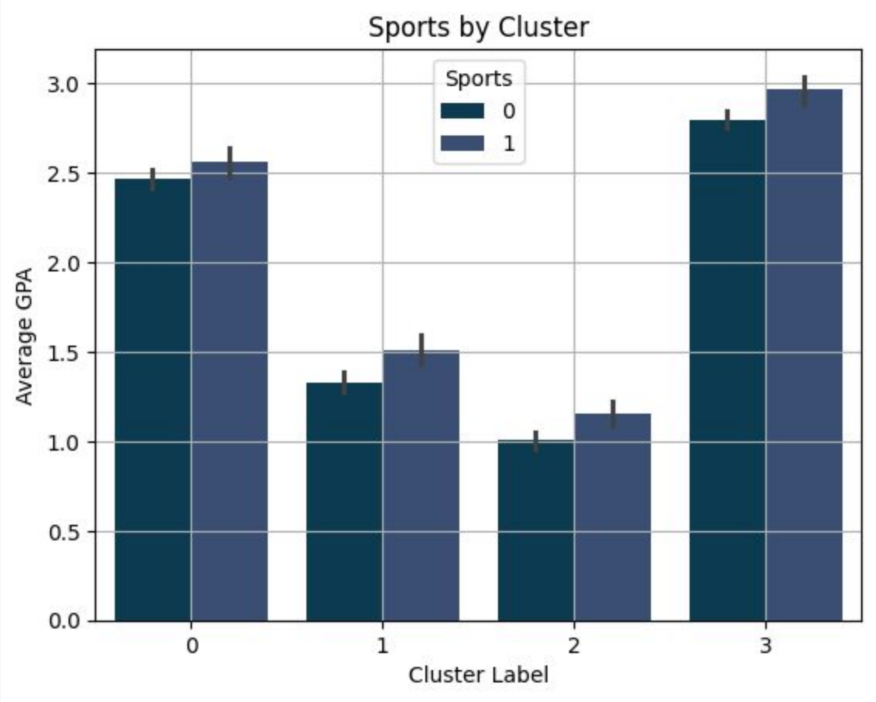
K-Means Clustering





K-Means Clustering





Results & Recommendations

Results

- ⟶ Linear regression model performed well and provided insight on the variables that correlated with raw GPA scores. Parental support, absences, and tutoring seemed to have the biggest effects
- ⟶ Decision tree performed strongly for GPA class 4 (F) but not so well for other classes. Possibly due to the weight of class 4
- ⟶ Clustering gives us 4 groups of students to focus on:
  - ↠ Low study time, high GPA
  - ↠ High study time, low GPA
  - ↠ Low study time, low GPA
  - ↠ High study time, high GPA
- ⟶ The better model: Decision Tree
  - ↠ Why? The decision tree predicts whether a student is at-risk at passing and failing while the clustering helps identify natural groupings of students with similar traits.

Recommendations

Students & Parents

→ Encourage weekly self-tracking and student/teacher check-ins.
→ Peer Tutoring Programs
→ Weekly self-tracking and student/teacher check-ins.
→ At-home study plans with encouraged routines.

Teachers

→ Early warning dashboards to flag students with high risk scores
→ Additional communication to all parties (students and parents)
→ Boost student engagement (clubs, sports, etc.)

Policymakers

→ Invest in support systems like teaching parents the power of support and funding tutoring programs as preventative measures.
→ Incentivize extracurricular participation.
→ Mandate early intervention tools (i.e high- absence monitoring systems)

Works Cited

Students Performance Dataset . 12 June 2024, www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset?resource=download.