Machine Learning

Final Project

Sergo Poghosyan, Edgar Oganessyan, Nare Stepanyan,
Anna Aghaloyan, Elvina Nosrati Alamdari

AUA Students Data

Data Description

	Student ID	FirstEnrolled_MajorCode	College	FincialAid_Received_AtLeast_Once	FirstEnrolled_Year	Gender	School_GPA	RoA	FirstYear_CGPA	All_CGPA
0	6272306053	BAB	CBE	0	2014	Male	3.810	1.0	0.188889	0.188889
1	4184934942	BAB	CBE	0	2018	Male	4.125	1.0	NaN	NaN
2	4694379187	BSCS	CSE	0	2013	Male	4.000	1.0	0.855556	0.855556
3	4219043039	BAB	CBE	0	2015	Male	3.565	1.0	2.250000	2.751282
4	14975762474	BAB	CBE	0	2015	Male	3.900	1.0	2.300000	1.525000

Student ID - a numeric code assigned to a student upon enrollment
FirstEnrolled_MajorCode - the major that a student was first enrolled in
College - shows the majors belonging to each college
FincialAid_Received_AtLeast_Once - represents a binary variable where:
A value of 1 indicates that the student has received financial aid at least once.
A value of 0 indicates that the student has not received financial aid at all.
FirstEnrolled_Year - each entry corresponds to the year in which a student initially enrolled in the institution
Gender- contains information about the gender of students
School_GPA - each entry corresponds to the GPA of a student at the school level. Please note that the "School_GPA" column has been standardized to a range from zero to five. Filters applied to the fields that yield a row count of five or fewer have been excluded from the dataset due to student identification concerns.
RoA (Republic of Armenia)- represents a binary variable where:
A value of 1 represents Armenian citizenship
A value of 0 represents non-Armenian citizenship
FirstYear_CGPA- is applicable to all individuals, not just those who have graduated or been dismissed/withdrawn
All_CGPA- all cumulative GPA - cumulative GPA is applicable to individuals who have completed their academic program or, alternatively, those who have left the program due to dismissal or withdrawal.

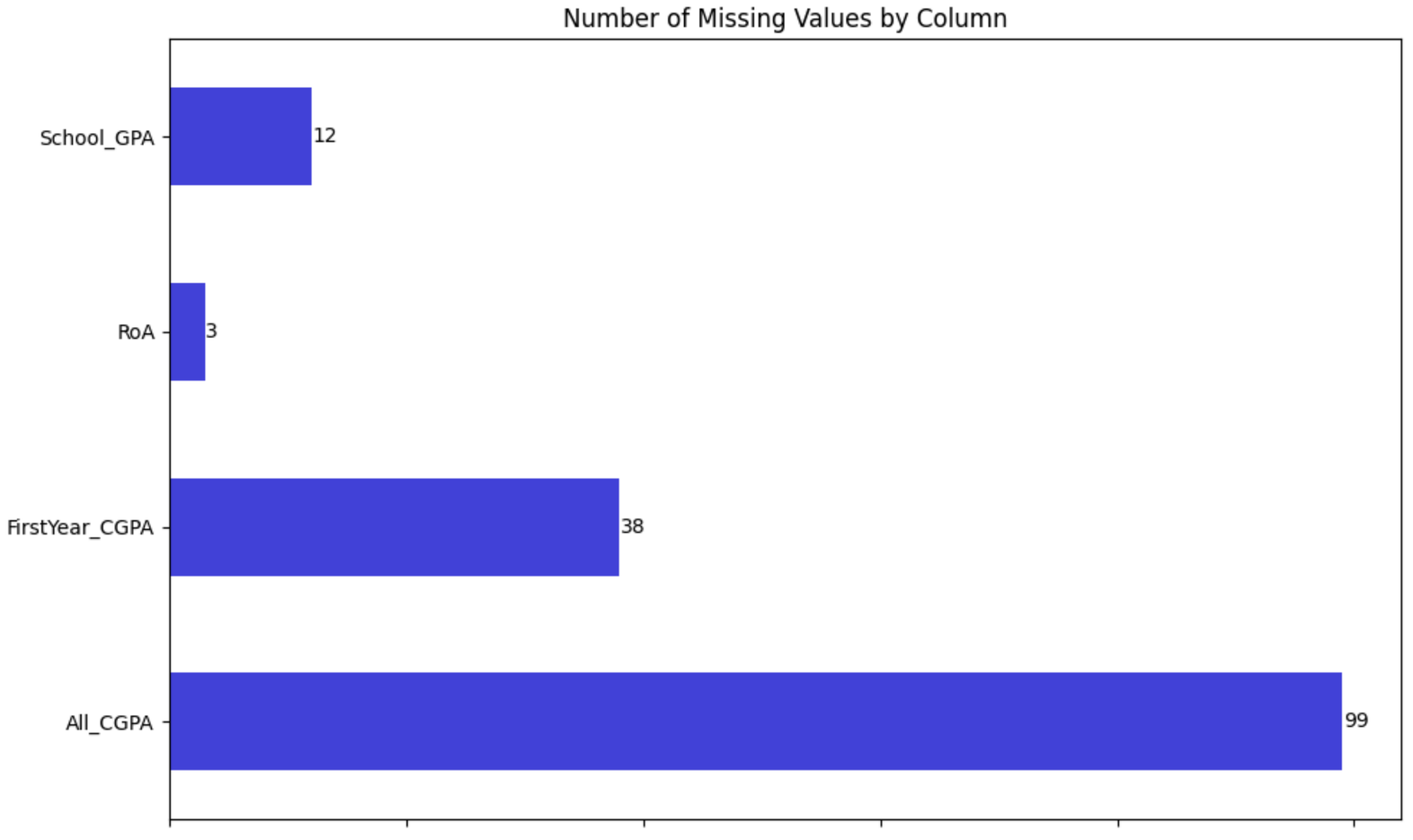
Who is this research for

- As per request of the Research Department of AUA
- For AUA staff in need of key analytics
- For current and future AUA students
- For University rating agencies

Data Cleaning

From Where we started

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2706 entries, 0 to 2705
Data columns (total 10 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Student ID                       2706 non-null   int64
 1   FirstEnrolled_MajorCode          2706 non-null   object
 2   College                          2706 non-null   object
 3   FincialAid_Received_AtLeast_Once 2706 non-null   int64
 4   FirstEnrolled_Year               2706 non-null   int64
 5   Gender                           2706 non-null   object
 6   School_GPA                       2694 non-null   float64
 7   RoA                              2703 non-null   float64
 8   FirstYear_CGPA                   2668 non-null   float64
 9   All_CGPA                         1303 non-null   float64
dtypes: float64(4), int64(3), object(3)
memory usage: 211.5+ KB
```

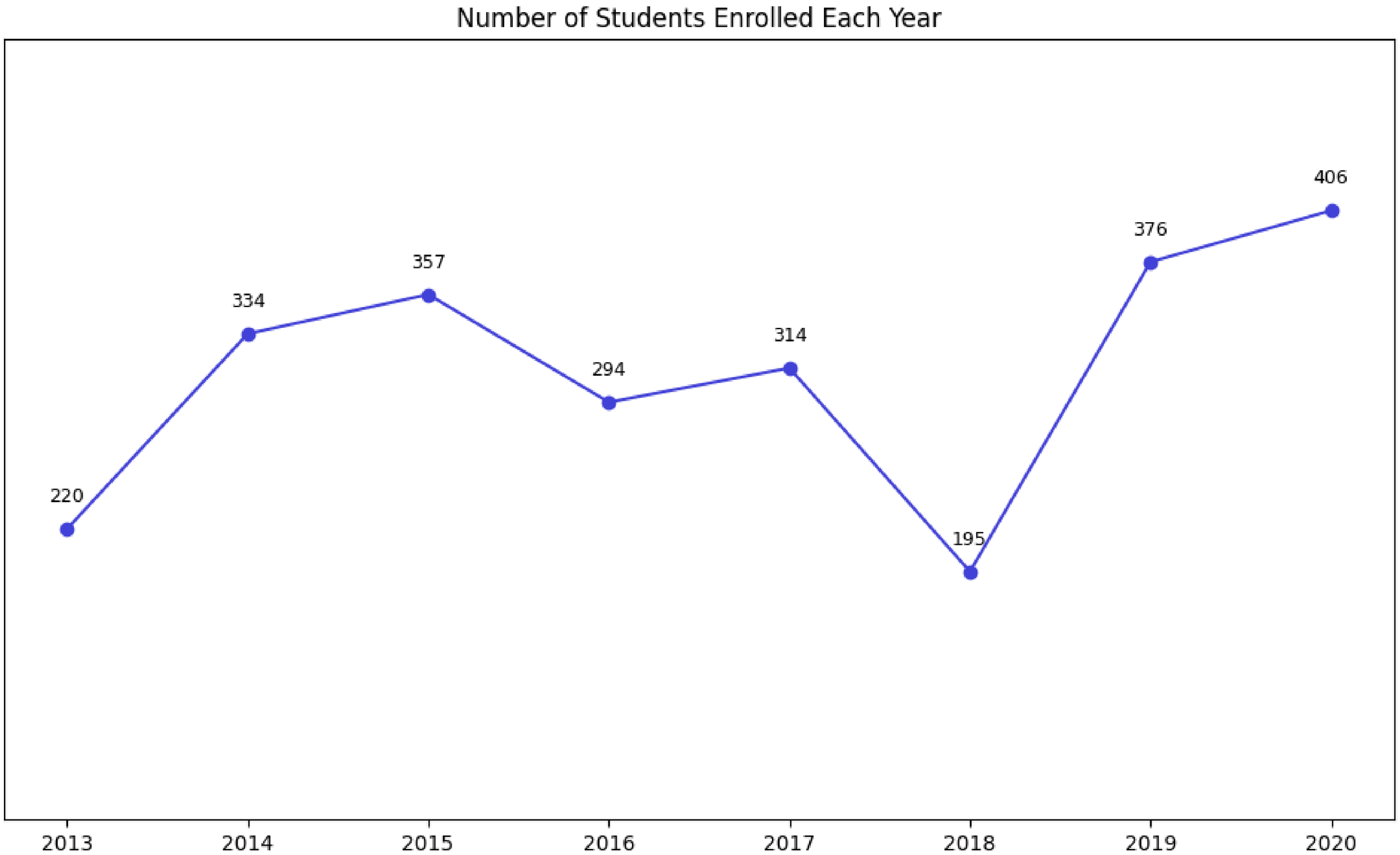


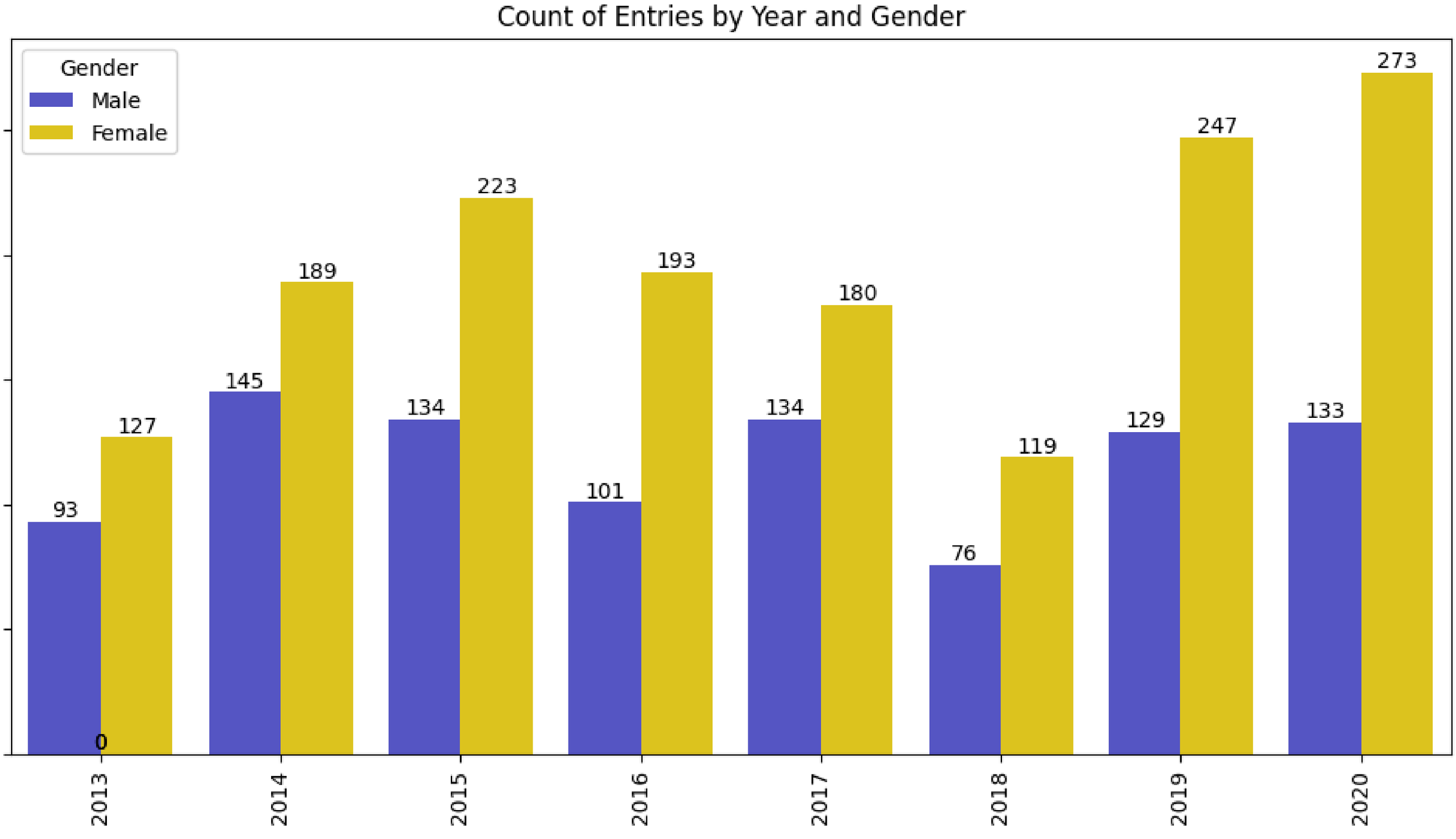
```
<class 'pandas.core.frame.DataFrame'>
Index: 2496 entries, 3 to 2705
Data columns (total 10 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Student ID                       2496 non-null   int64
 1   FirstEnrolled_MajorCode          2496 non-null   object
 2   College                          2496 non-null   object
 3   FincialAid_Received_AtLeast_Once 2496 non-null   int64
 4   FirstEnrolled_Year               2496 non-null   int64
 5   Gender                           2496 non-null   object
 6   School_GPA                       2496 non-null   float64
 7   RoA                              2496 non-null   float64
 8   FirstYear_CGPA                   2496 non-null   float64
 9   All_CGPA                         1205 non-null   float64
dtypes: float64(4), int64(3), object(3)
memory usage: 214.5+ KB
```

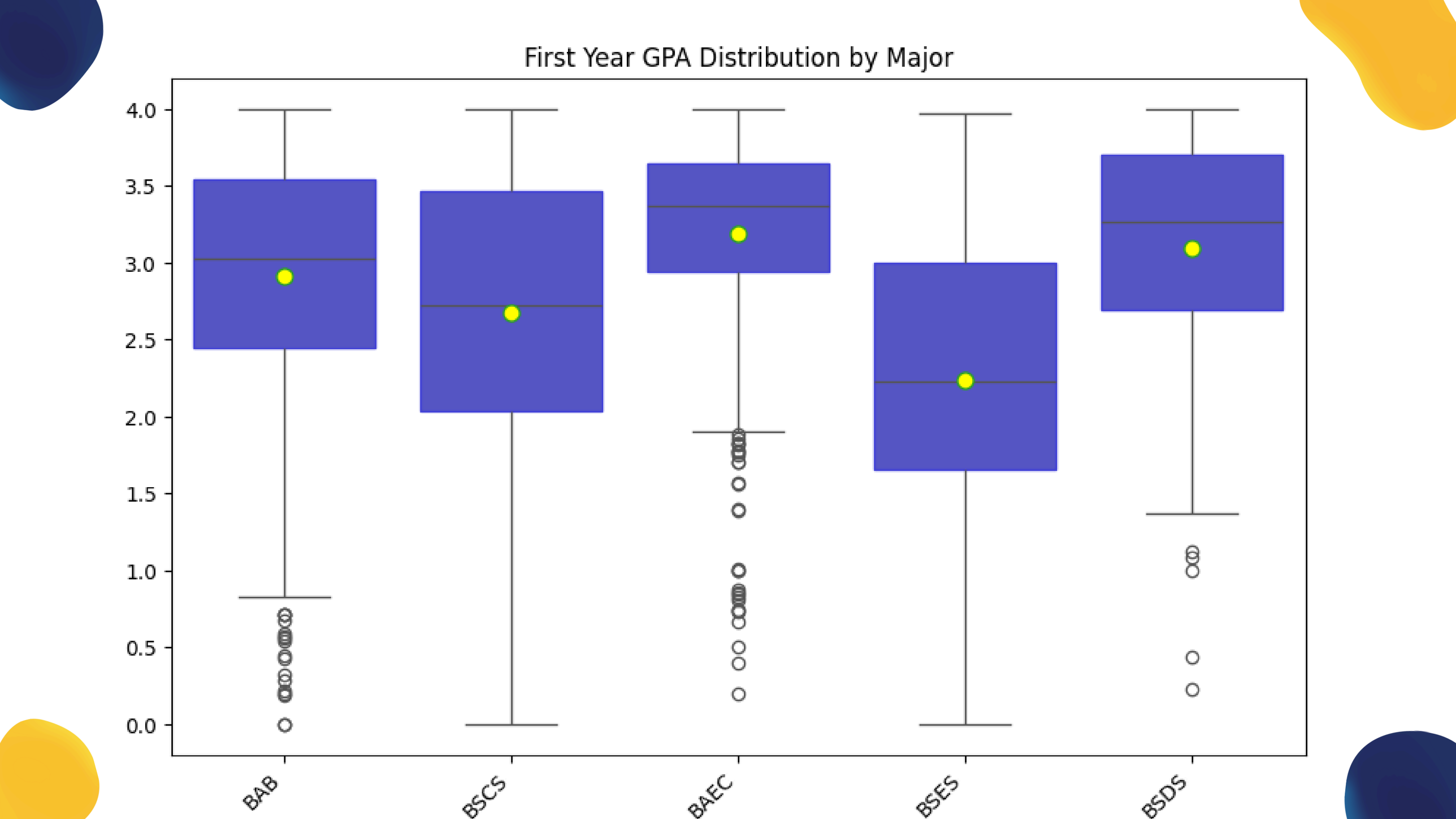
```
(Student ID                          0
 FirstEnrolled_MajorCode             0
 College                             0
 FincialAid_Received_AtLeast_Once    0
 FirstEnrolled_Year                  0
 Gender                              0
 School_GPA                         12
 RoA                                 3
 FirstYear_CGPA                     38
 All_CGPA                         1403
 dtype: int64,
 Student ID                          0
 FirstEnrolled_MajorCode             0
 College                             0
 FincialAid_Received_AtLeast_Once    0
 FirstEnrolled_Year                  0
 Gender                              0
 School_GPA                          0
 RoA                                 0
 FirstYear_CGPA                      0
 All_CGPA                         1291
 dtype: int64)
```

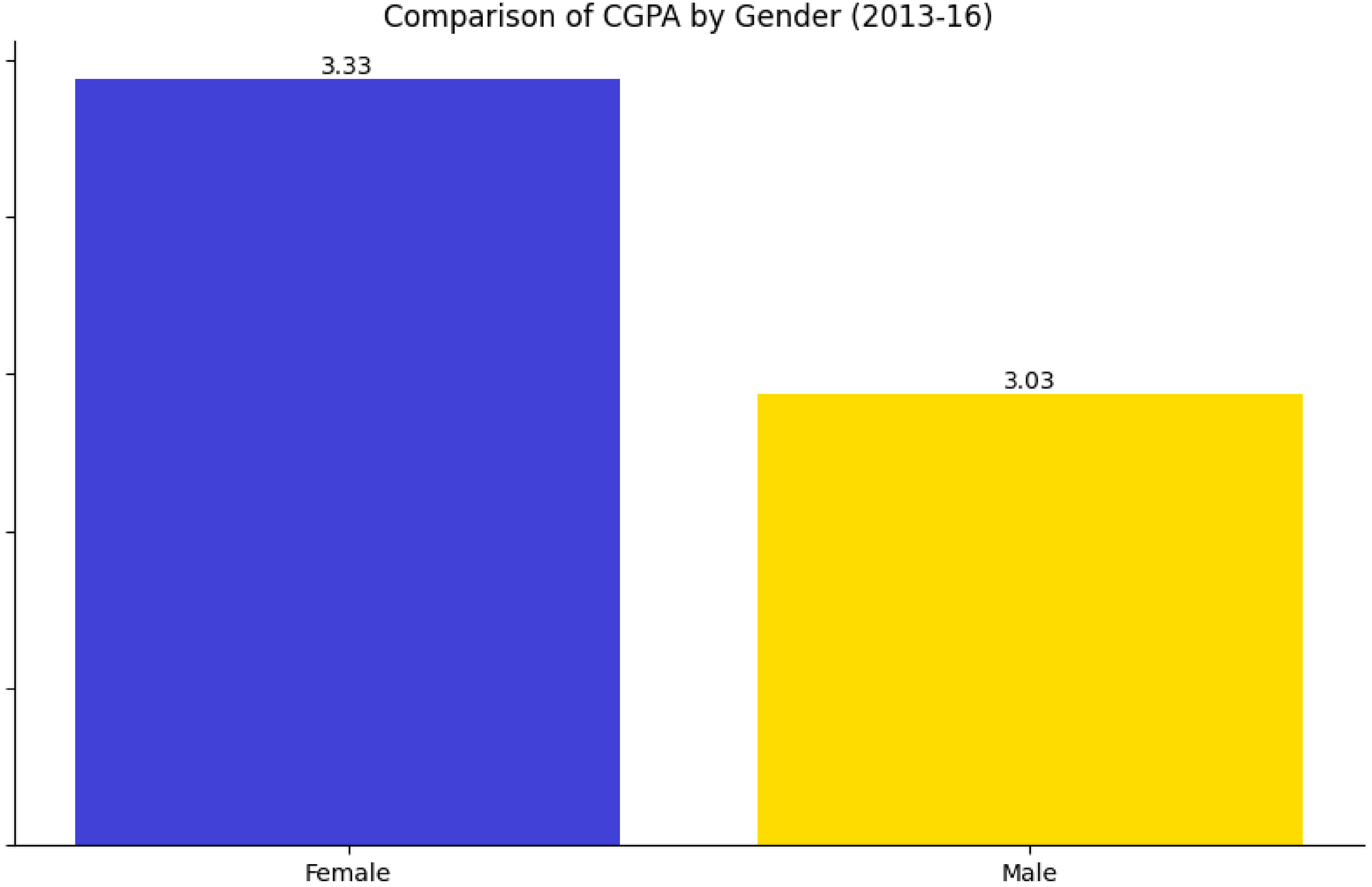
- **Here we dropped missing values for School_GPA, RoA, FirstYear_CGPA, All_CGPA(2013-2016)**
- **As the students whos been enrolled before 2016 and their All CGPA is less than 2, we should drop them as technically students cannot graduate with cumulitive GPA less than 2.0, so we dropped them as well**
- **Also After thorough data-checking we found out that there is an outlier in School_GPA: student with extremely low School_GPA so we decidided to drop that outlier**

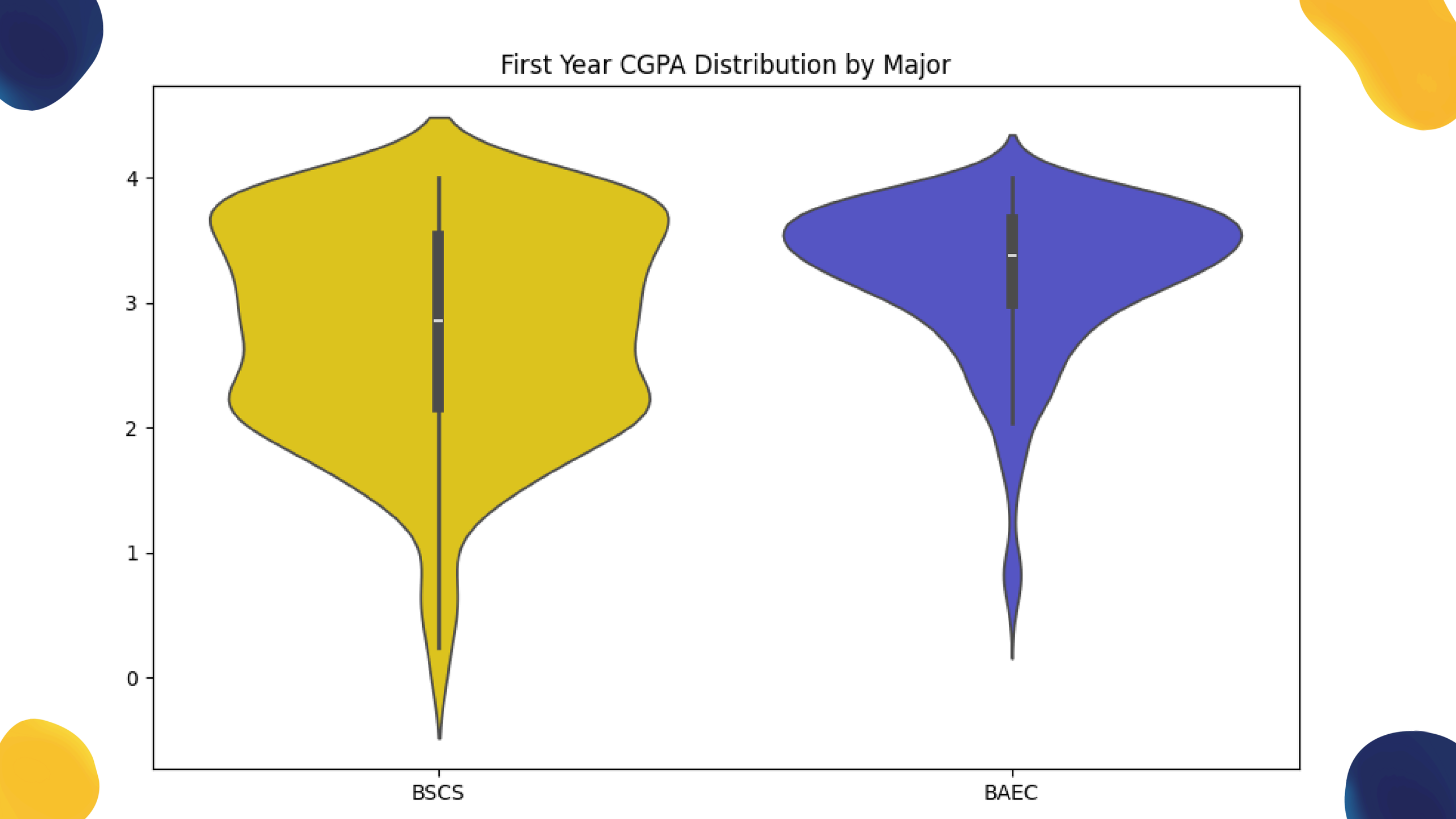
Explanatory Data Analysis

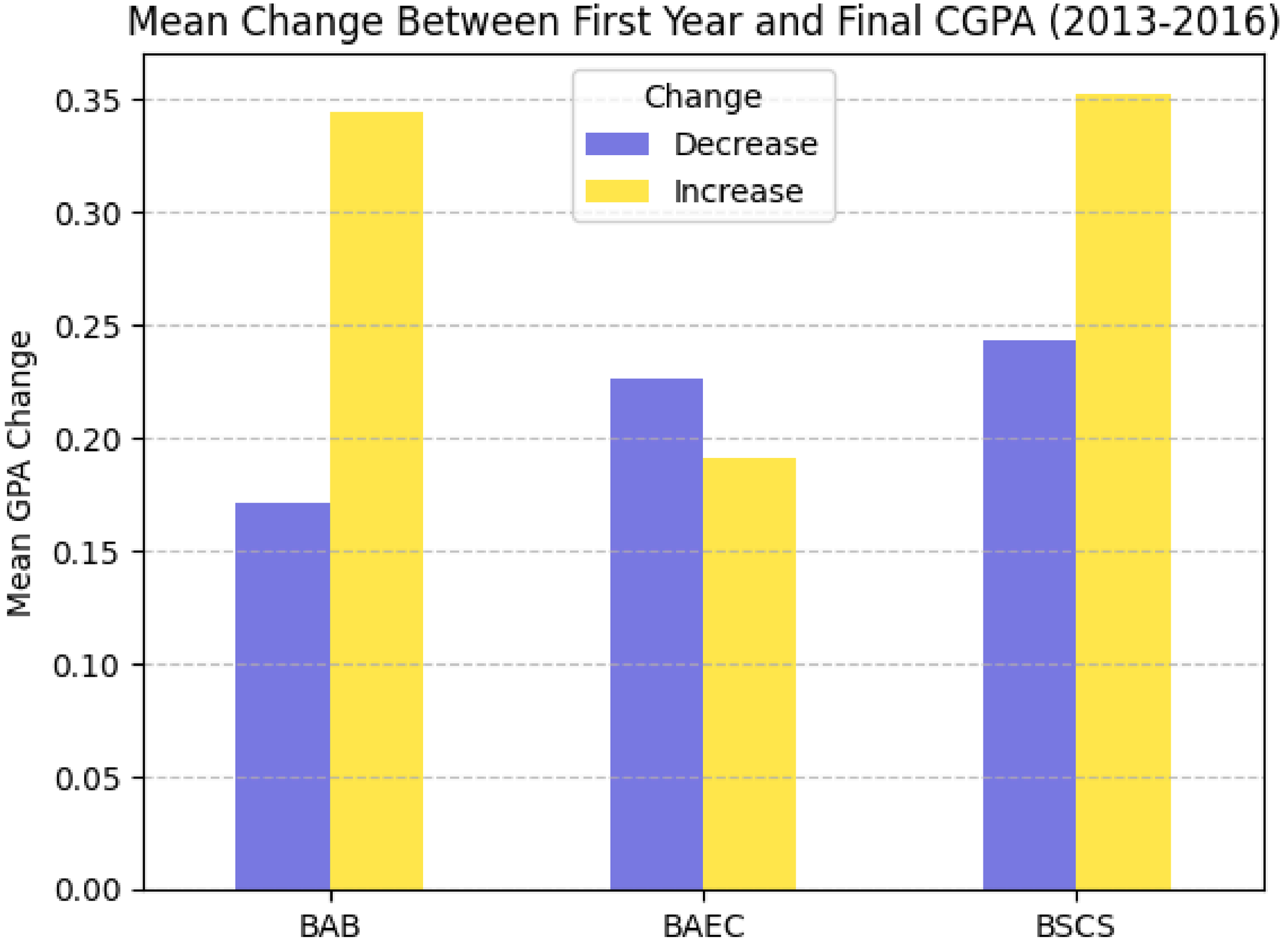


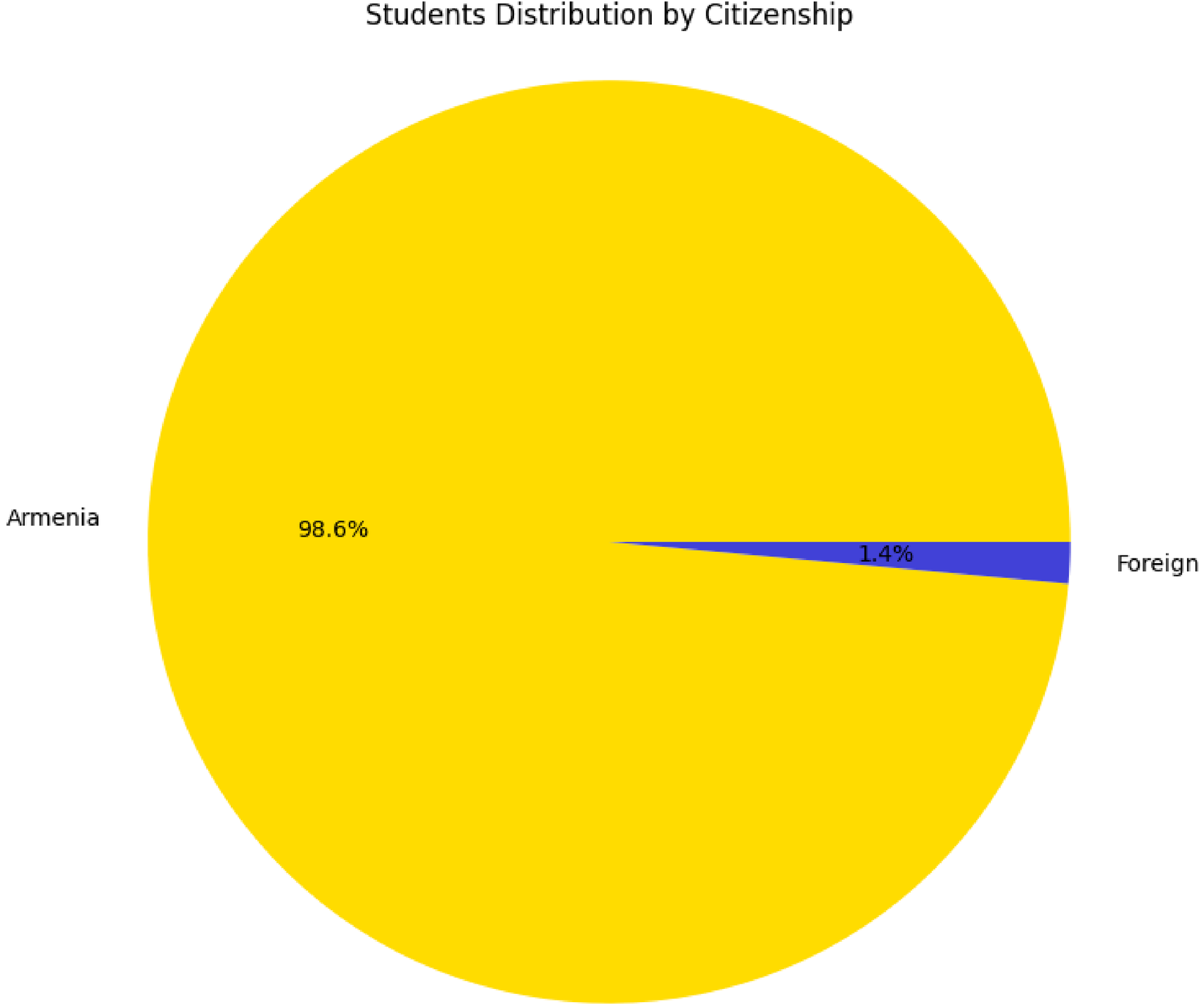

Count of Entries by Year and Gender
Gender
Male
Female
93
127
145
189
134
223
101
193
134
180
76
119
129
247
133
273
0
2013
2014
2015
2016
2017
2018
2019
2020


First Year GPA Distribution by Major
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
BAB
BSCS
BAEC
BSES
BSDS

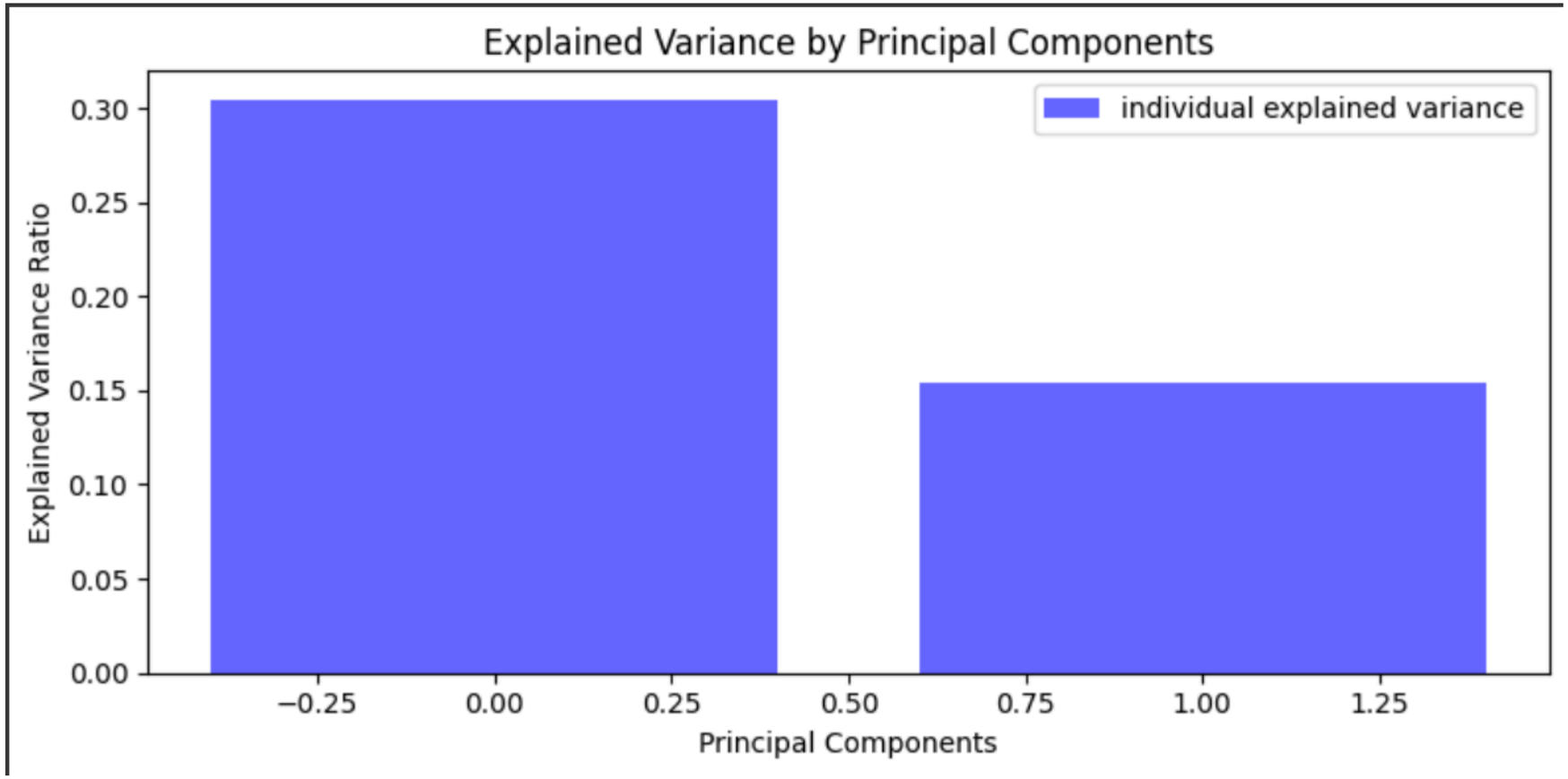

Comparison of CGPA by Gender (2013-16)
3.33
3.03
Female
Male

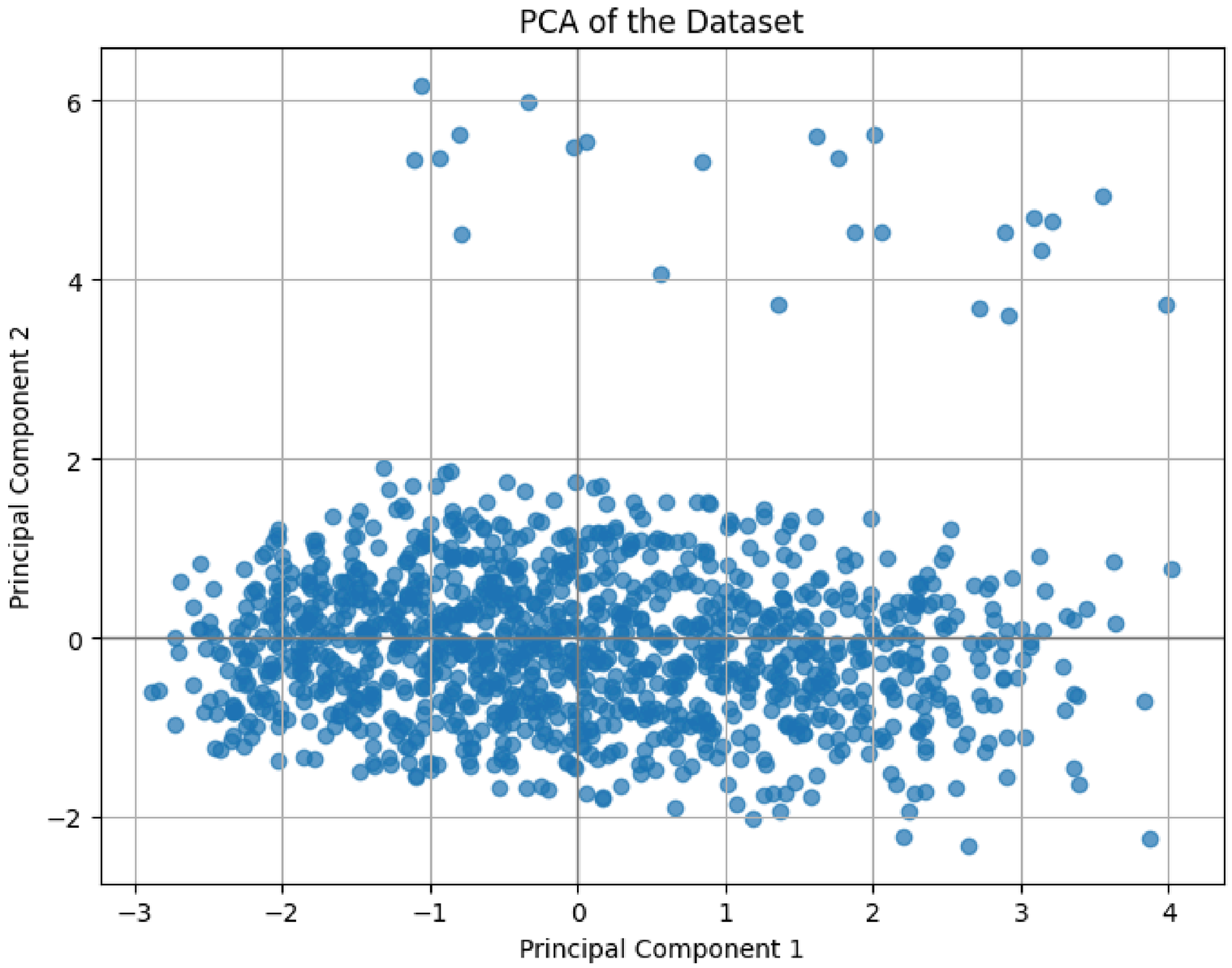

First Year CGPA Distribution by Major
4
3
2
1
0
BSCS
BAEC


Mean Change Between First Year and Final CGPA (2013-2016)
Change
Decrease
Increase
Mean GPA Change
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
BAB
BAEC
BSCS

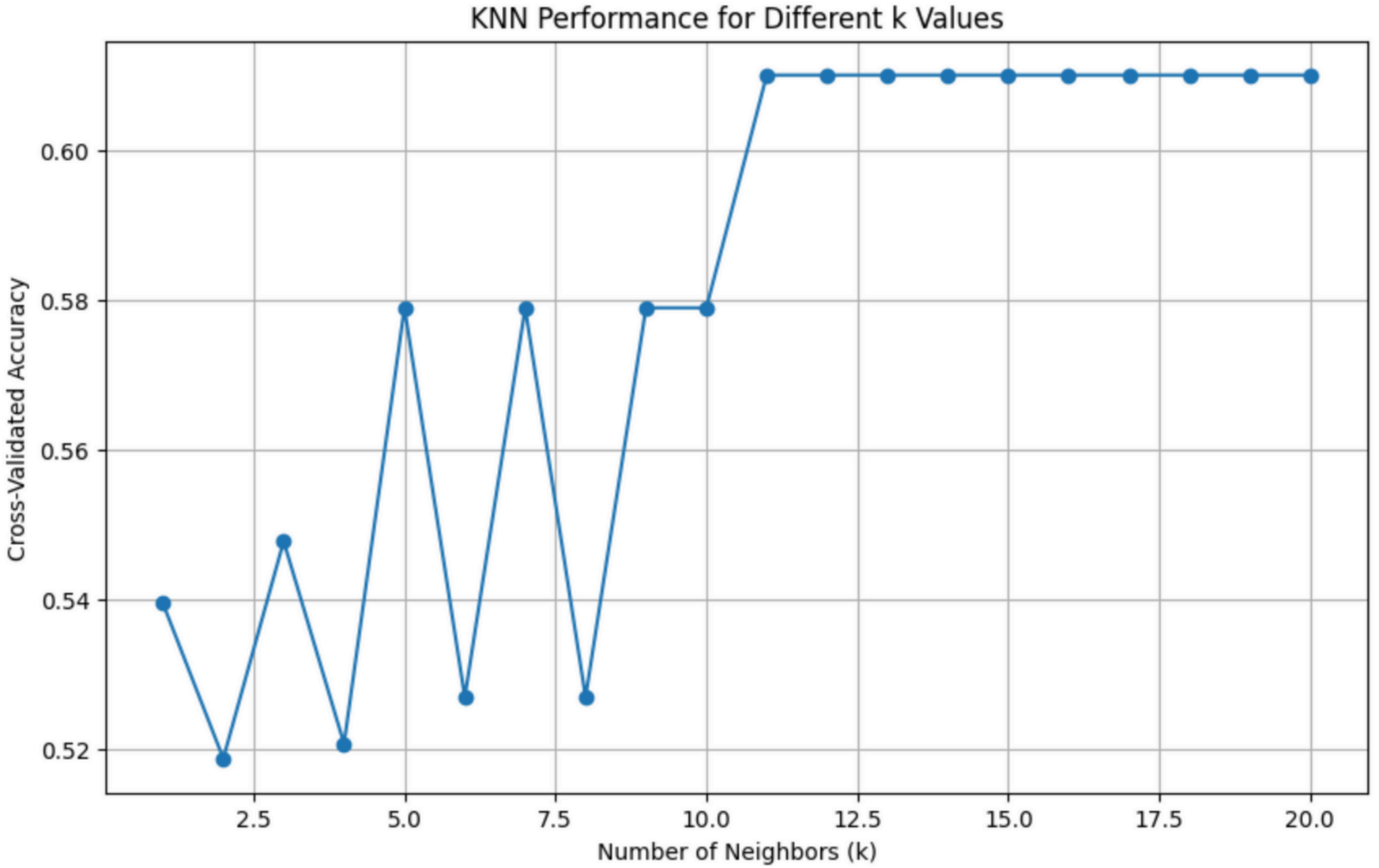

Students Distribution by Citizenship
Armenia
98.6%
1.4%
Foreign

PCA Analysis


Explained Variance by Principal Components
individual explained variance
Explained Variance Ratio
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Principal Components


PCA of the Dataset
Principal Component 2
Principal Component 1
6
4
2
0
−2
−3
−2
−1
0
1
2
3
4

KNN Model Analysis


KNN Performance for Different k Values
Cross-Validated Accuracy
0.60
0.58
0.56
0.54
0.52
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of Neighbors (k)

Results:
FirstEnrolled_Majorcode

Classification Report for k=20:

	precision	recall	f1-score	support
0	0.61	0.46	0.53	111
1	0.62	0.75	0.68	130
accuracy			0.62	241
macro avg	0.62	0.61	0.60	241
weighted avg	0.62	0.62	0.61	241

Results:
School_GPA

Classification Report for KNN using only School_GPA:

	precision	recall	f1-score	support
0	0.61	0.46	0.53	111
1	0.62	0.75	0.68	130
accuracy			0.62	241
macro avg	0.62	0.61	0.60	241
weighted avg	0.62	0.62	0.61	241

Results:
Gender

Classification Report for KNN using only Gender:

	precision	recall	f1-score	support
0	0.66	0.51	0.58	111
1	0.65	0.78	0.71	130
accuracy			0.66	241
macro avg	0.66	0.65	0.64	241
weighted avg	0.66	0.66	0.65	241

Results:

FincialAid_Received_AtLeast_Once

Classification Report for KNN using only Financial Aid Received At Least Once:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	111
1	0.54	1.00	0.70	130
accuracy			0.54	241
macro avg	0.27	0.50	0.35	241
weighted avg	0.29	0.54	0.38	241

Results:
'RoA'

Classification Report for KNN using only RoA

	precision	recall	f1-score	support
0	1.00	0.03	0.05	111
1	0.55	1.00	0.71	130
accuracy			0.55	241
macro avg	0.77	0.51	0.38	241
weighted avg	0.76	0.55	0.41	241

Results:
'FirstEnrolled_Year'

Classification Report for KNN using only FirstEnrolled_Year

	precision	recall	f1-score	support
0	0.50	0.54	0.52	111
1	0.58	0.54	0.56	130
accuracy			0.54	241
macro avg	0.54	0.54	0.54	241
weighted avg	0.54	0.54	0.54	241

Results:
"College"

Classification Report for KNN using only College:

	precision	recall	f1-score	support
0	0.64	0.32	0.43	111
1	0.59	0.85	0.70	130
accuracy			0.61	241
macro avg	0.62	0.59	0.56	241
weighted avg	0.62	0.61	0.58	241

Results:
FirstYear_CGPA

Classification Report for KNN using only FirstYear_CGPA:

	precision	recall	f1-score	support
0	0.86	0.82	0.84	111
1	0.85	0.88	0.87	130
accuracy			0.85	241
macro avg	0.86	0.85	0.85	241
weighted avg	0.85	0.85	0.85	241

Results:

All features included

Classification Report:

	precision	recall	f1-score	support
0	0.69	0.57	0.62	111
1	0.68	0.78	0.73	130
accuracy			0.68	241
macro avg	0.69	0.68	0.68	241
weighted avg	0.69	0.68	0.68	241

Regression Analysis

Why Regression?

- **Prediction. Regression Models can be used to predict the value of a variable. For example, in case of our data, Gender or All_CGPA**

- **Inference and Hypothesis testing. Regression analysis provides statistical inference tools to test hypotheses about the relationships between variables.**

Regression Types

Linear Regression

Multivariative Regression

Logistic Regression

Time Series Regression

And much more...

Regression Types

Linear Regression

Multivariative Regression

Time Series Regression

And much more...

Regression Types

Linear Regression

Multivariative Regression

Logistic Regression

Time Series Regression

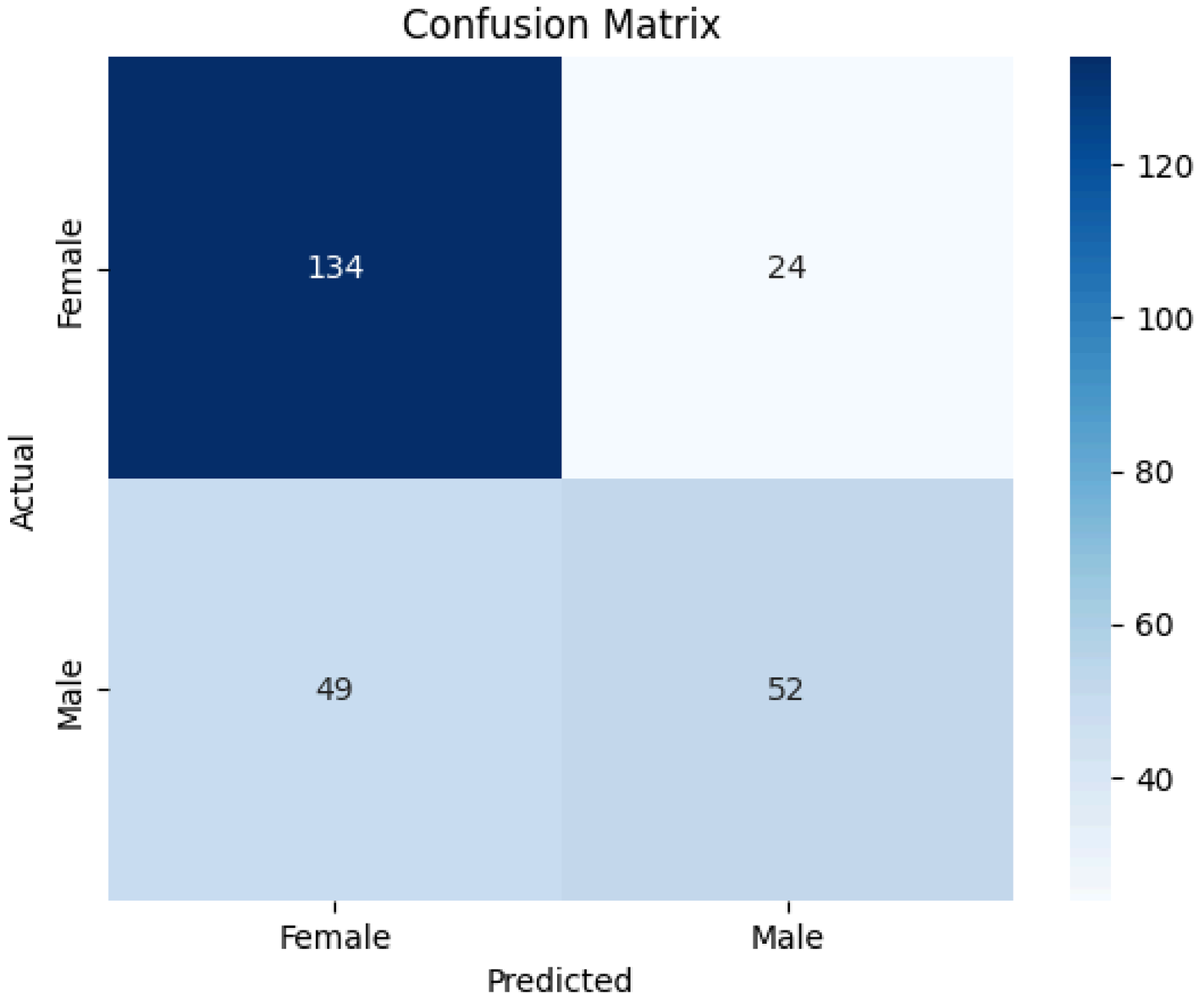
And much more...

Logistic Regression

- Precision: Female 73%, Male 68%
- Recal: Female 83%, Male 51%
- F1 Score: 79% VS 59%

Overall Acuracy was 72%, which is not too bad

Logistic Regression

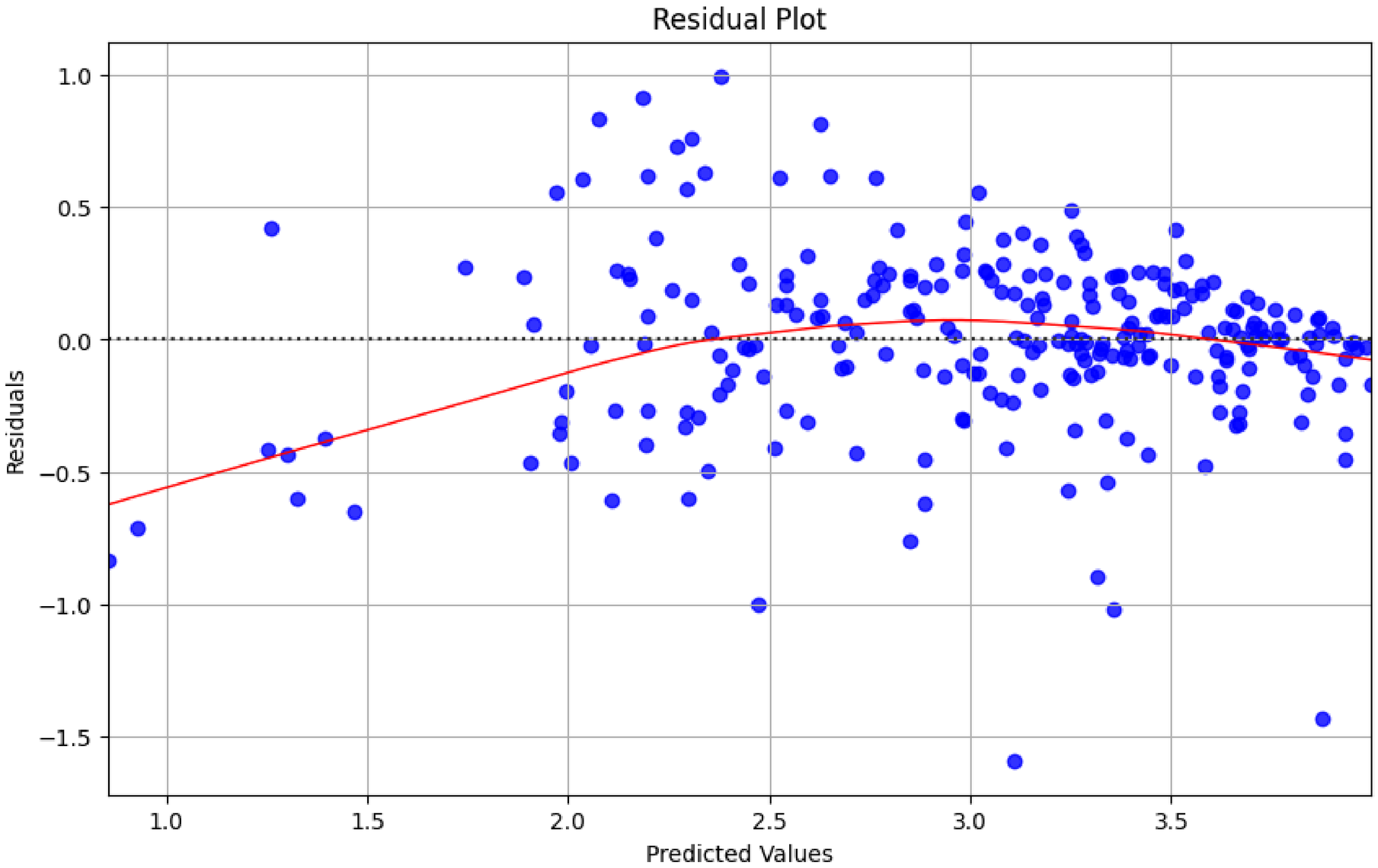


Multivariative Regression

- **Root Mean Squared Error (RMSE): 0.337**
- **R-squared (R^2) Score: 0.786**

The R^2 score of 0.786 implies that approximately 78.6% of the variability in the cumulative GPA is explained by the model. Pretty good result for such data.

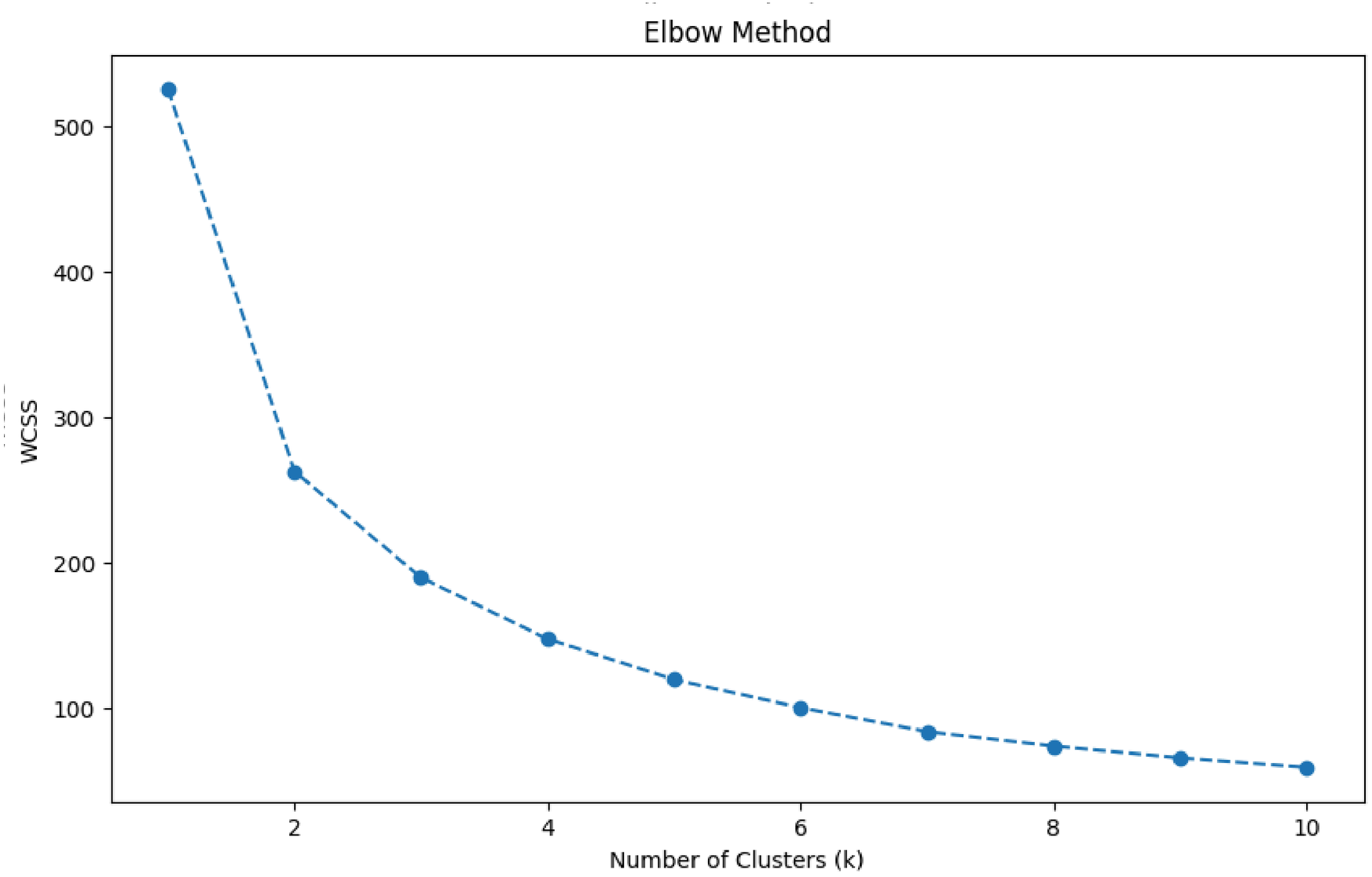
Multivariative Regression



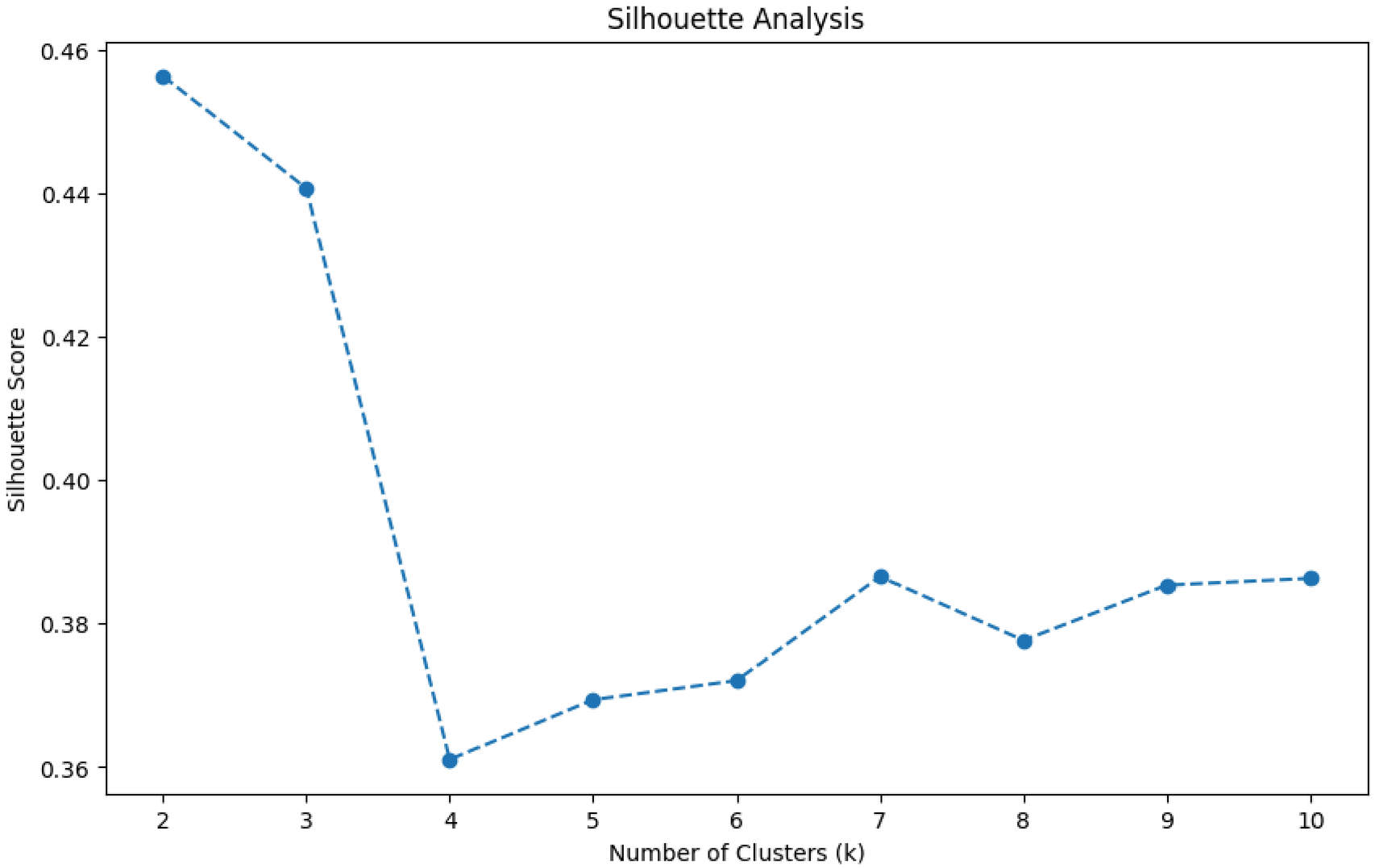
Unsupervised Learning Algorithms

K-Means, KMedoids, GMM, DBSCAN

Elbow Method

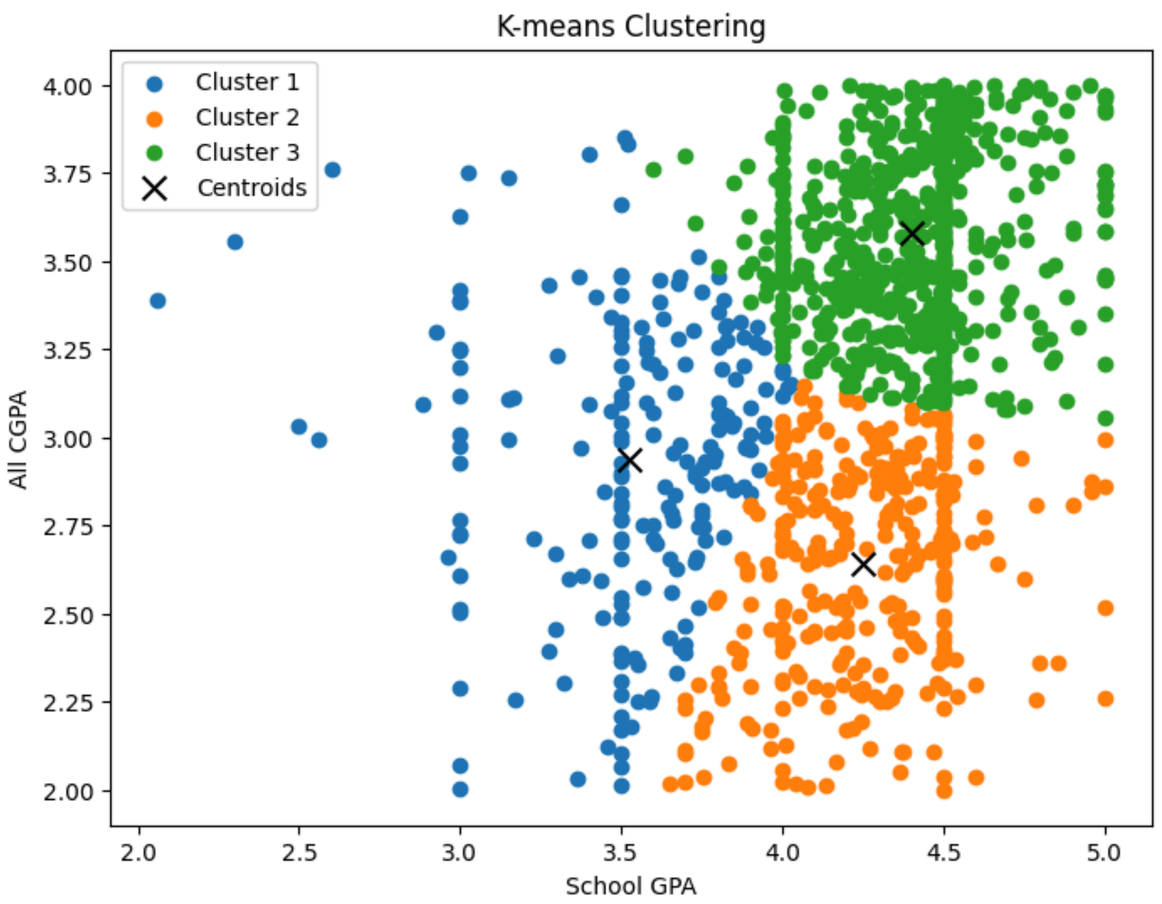


Silhouette Analysis

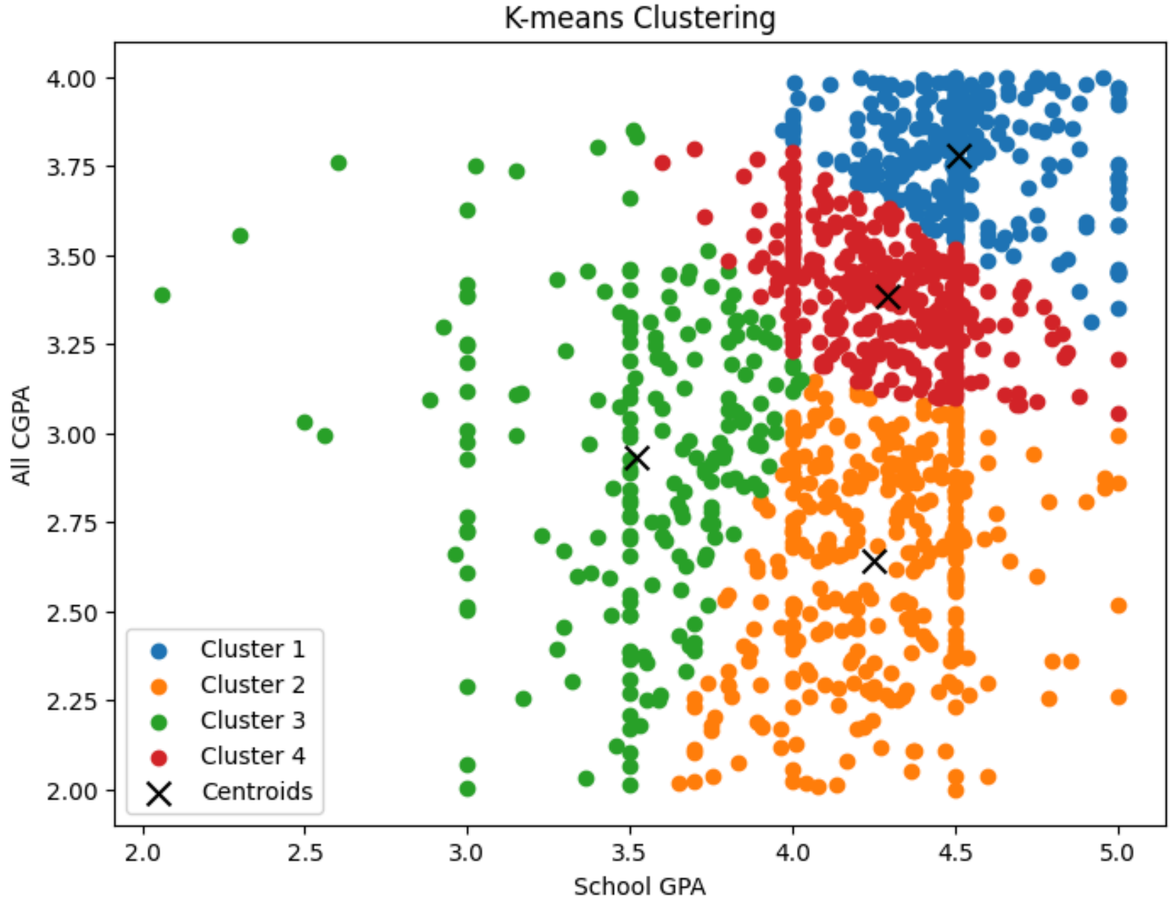


K-Means

k = 3

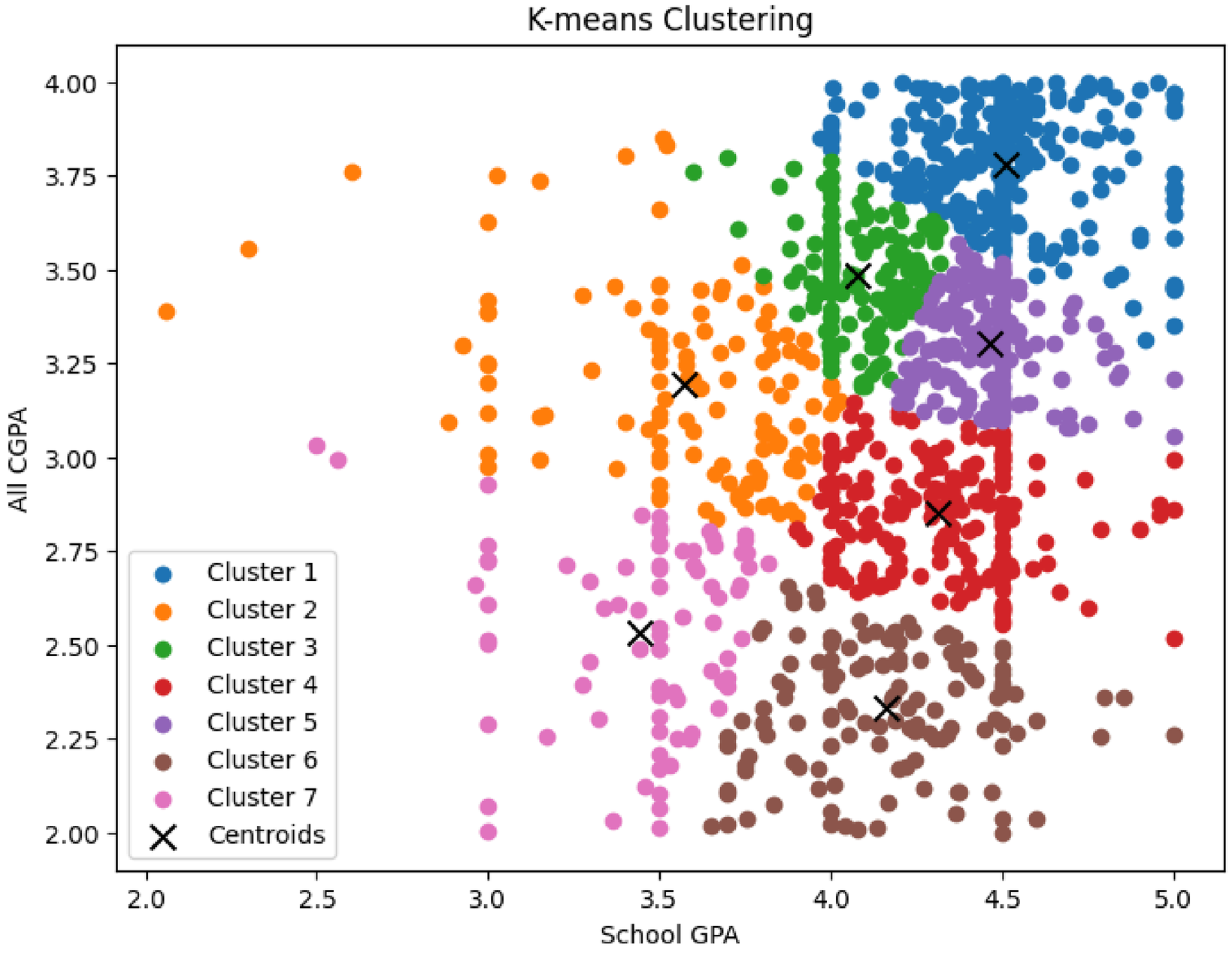

K-means Clustering
Cluster 1
Cluster 2
Cluster 3
Centroids
All CGPA
School GPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
2.0
2.5
3.0
3.5
4.0
4.5
5.0

k = 4


K-means Clustering
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Centroids
All CGPA
School GPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
2.0
2.5
3.0
3.5
4.0
4.5
5.0

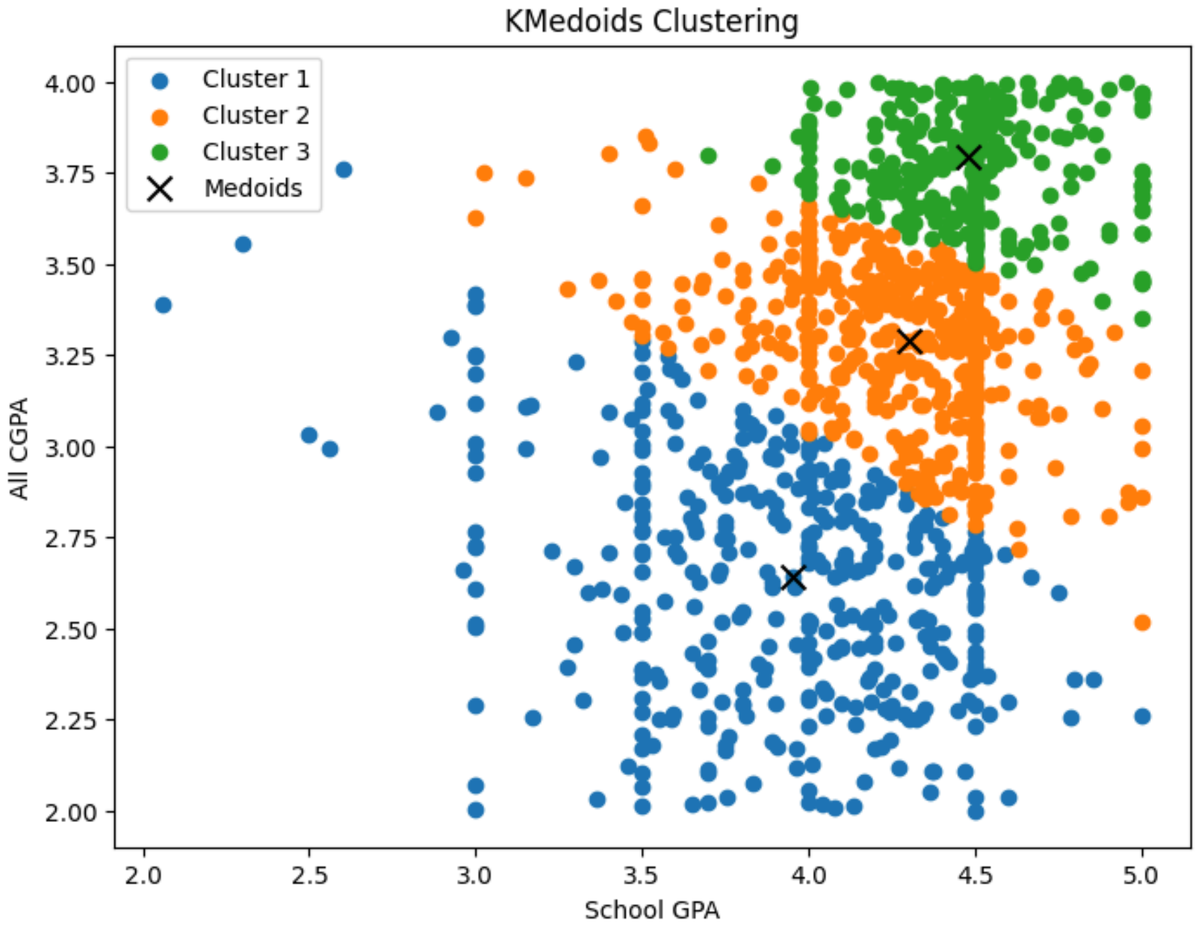
K-Means

k = 7

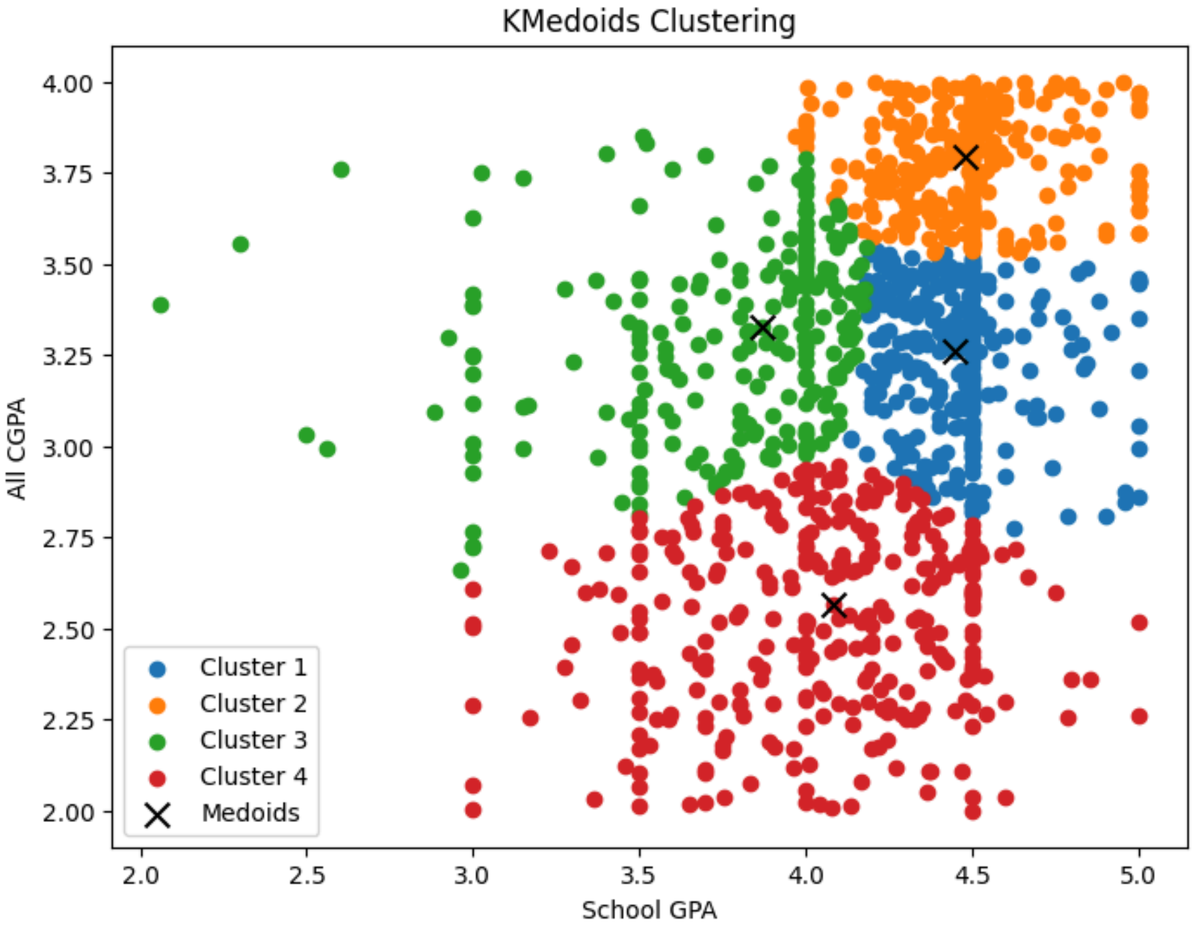

K-means Clustering
All CGPA
School GPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Centroids

KMedoids

k = 3

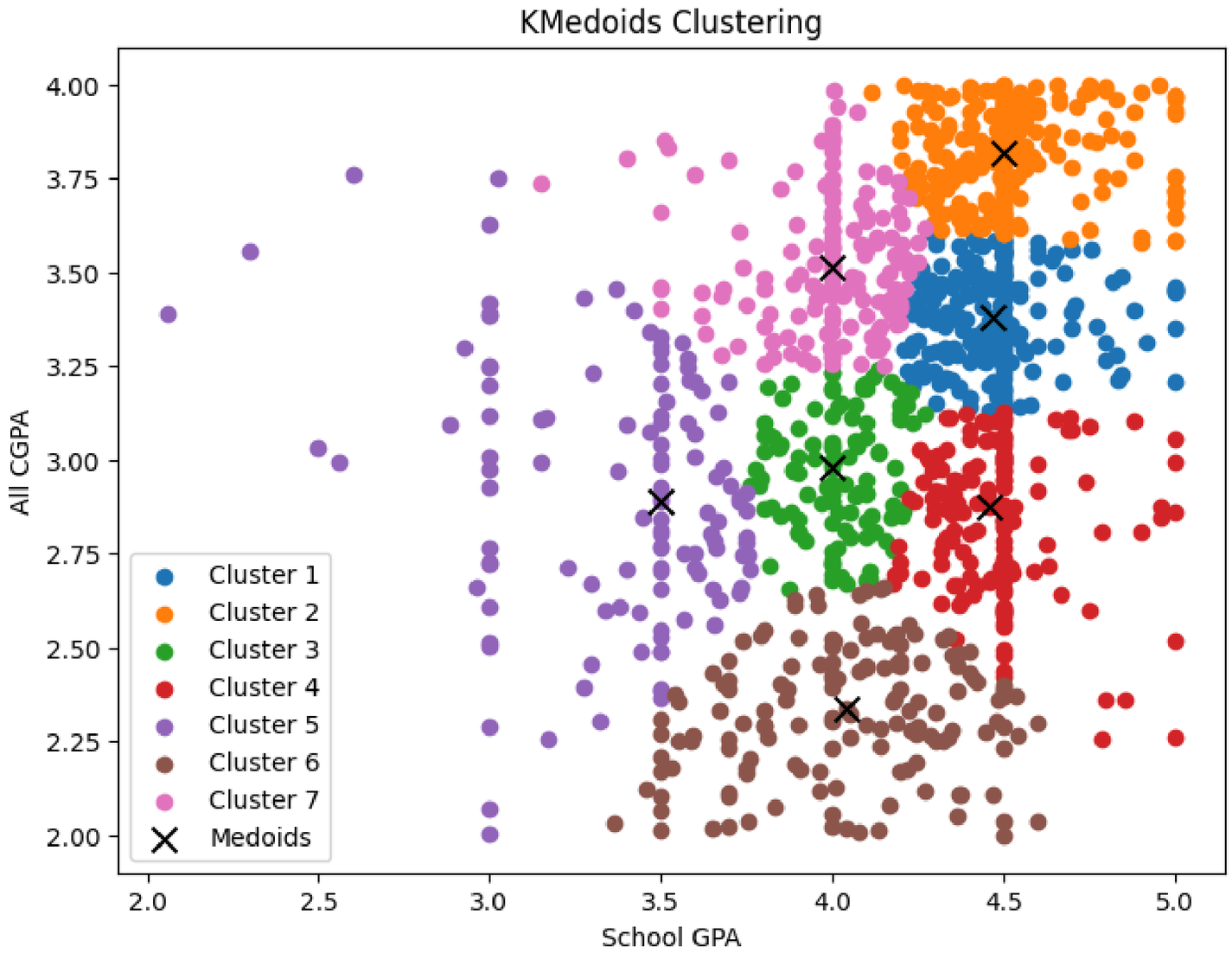


k = 4



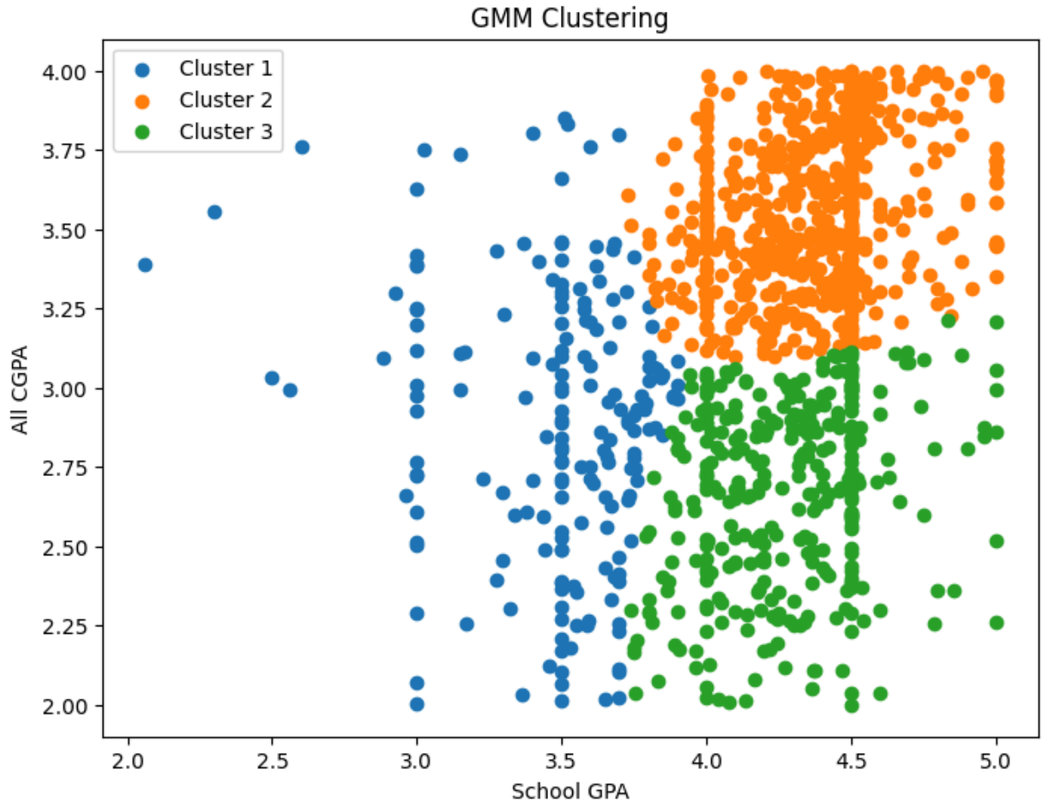
KMedoids

k = 7

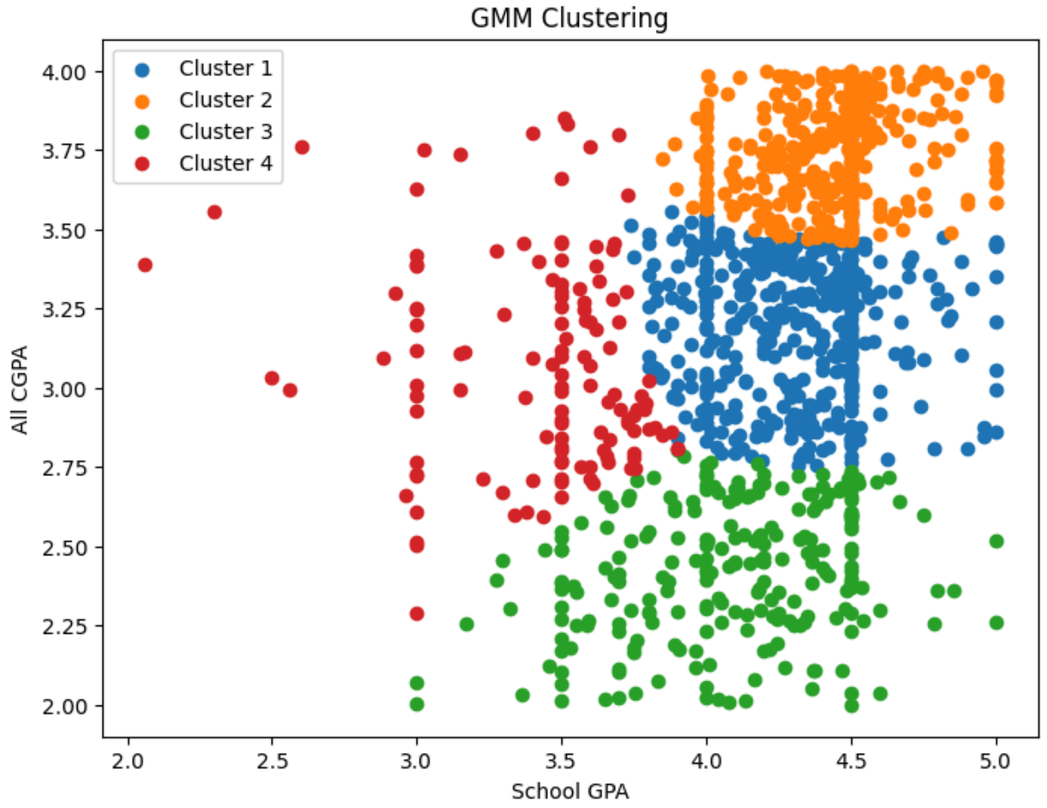

KMedoids Clustering
All CGPA
School GPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Medoids

GMM

k = 3

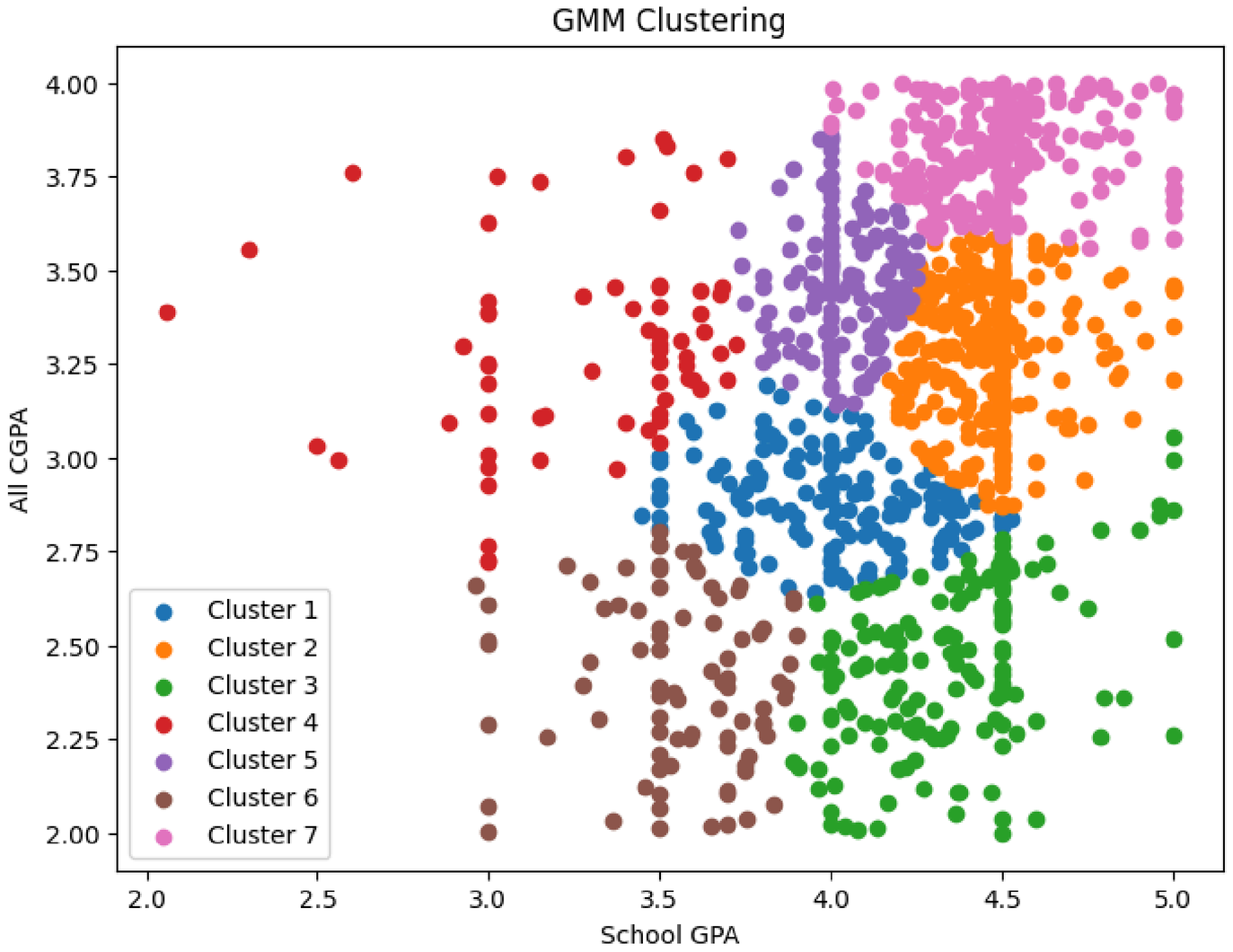

GMM Clustering
Cluster 1
Cluster 2
Cluster 3
All CGPA
School GPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
2.0
2.5
3.0
3.5
4.0
4.5
5.0

k = 4

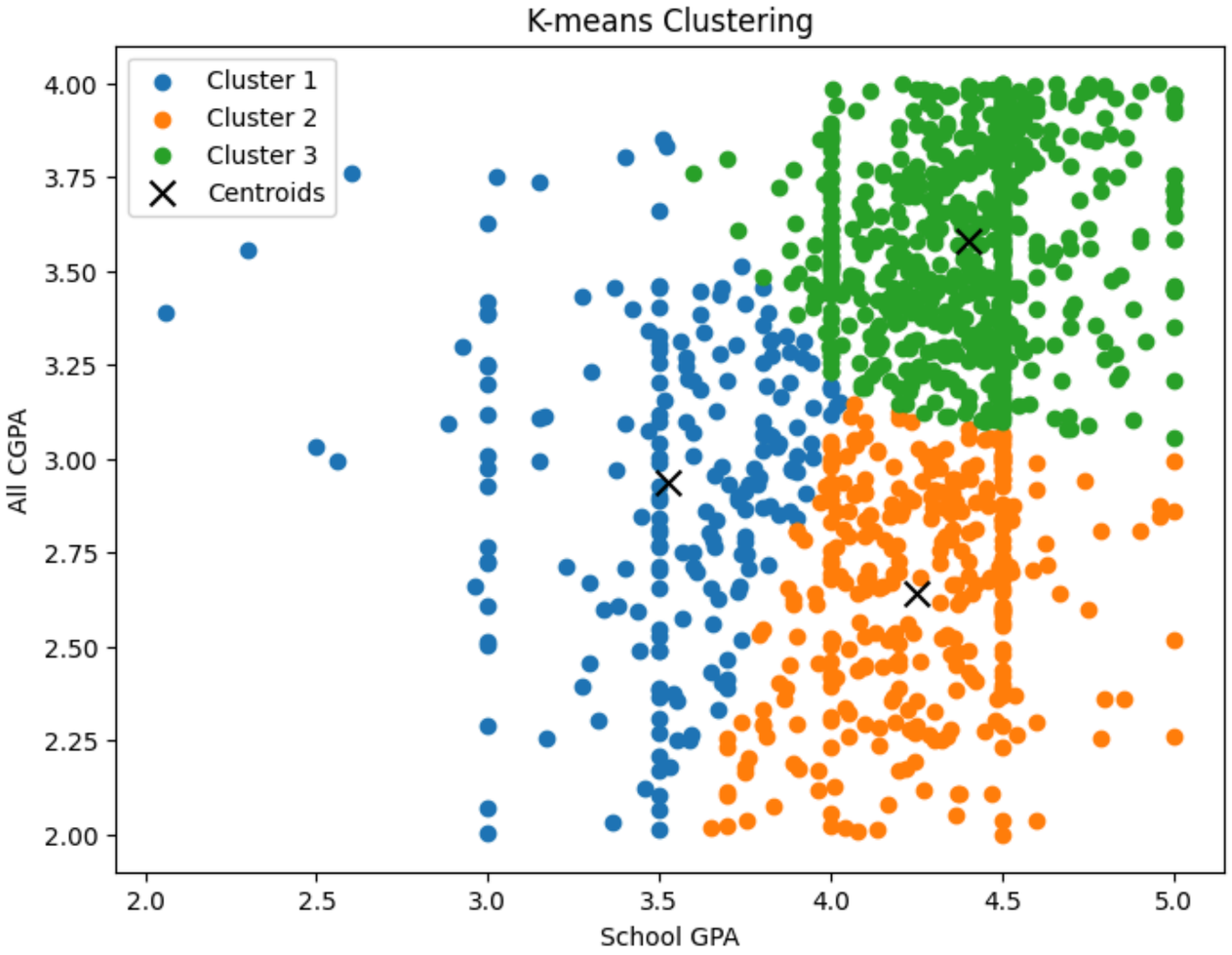

GMM Clustering
Cluster 1
Cluster 2
Cluster 3
Cluster 4
All CGPA
School GPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
2.0
2.5
3.0
3.5
4.0
4.5
5.0

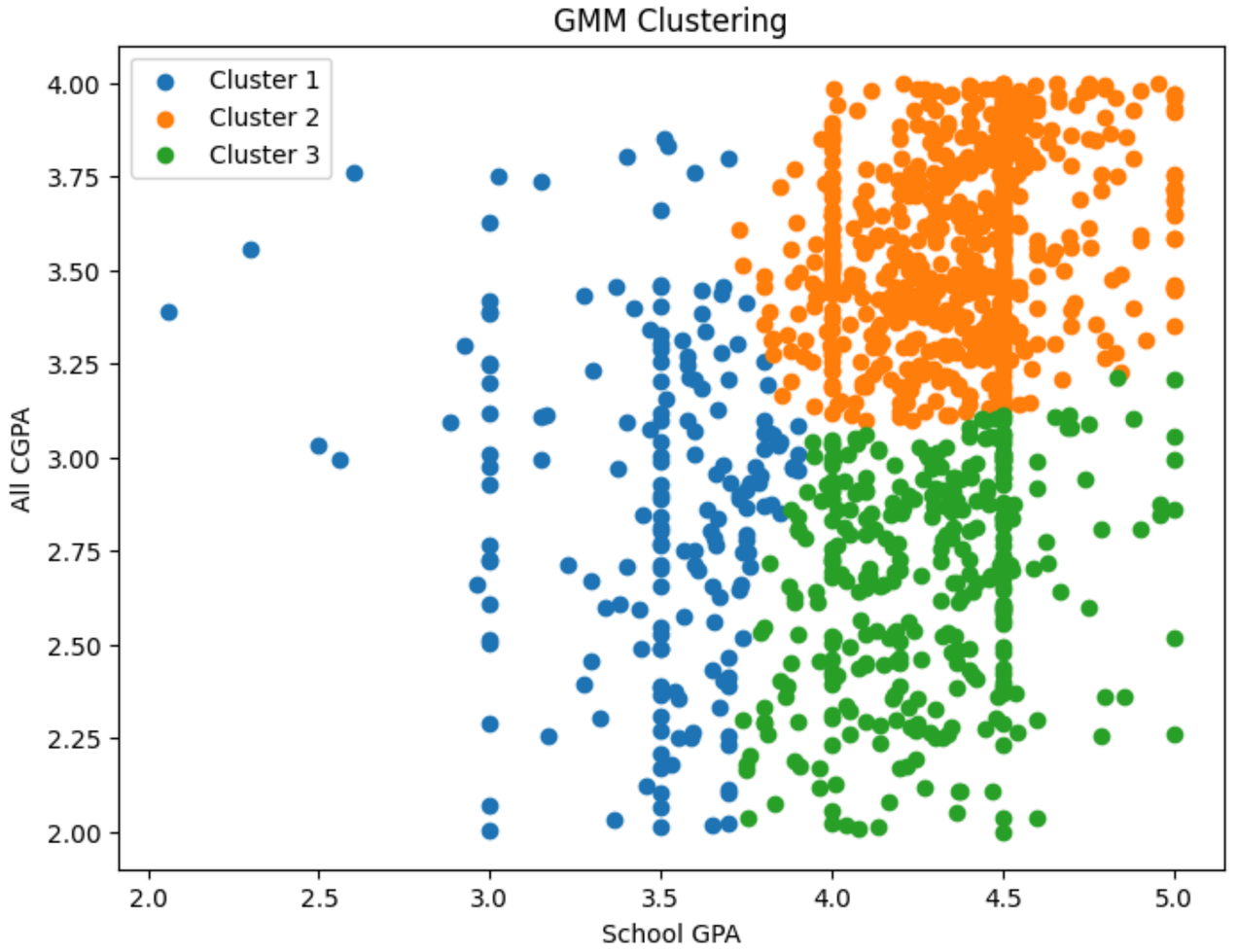
GMM

k = 7

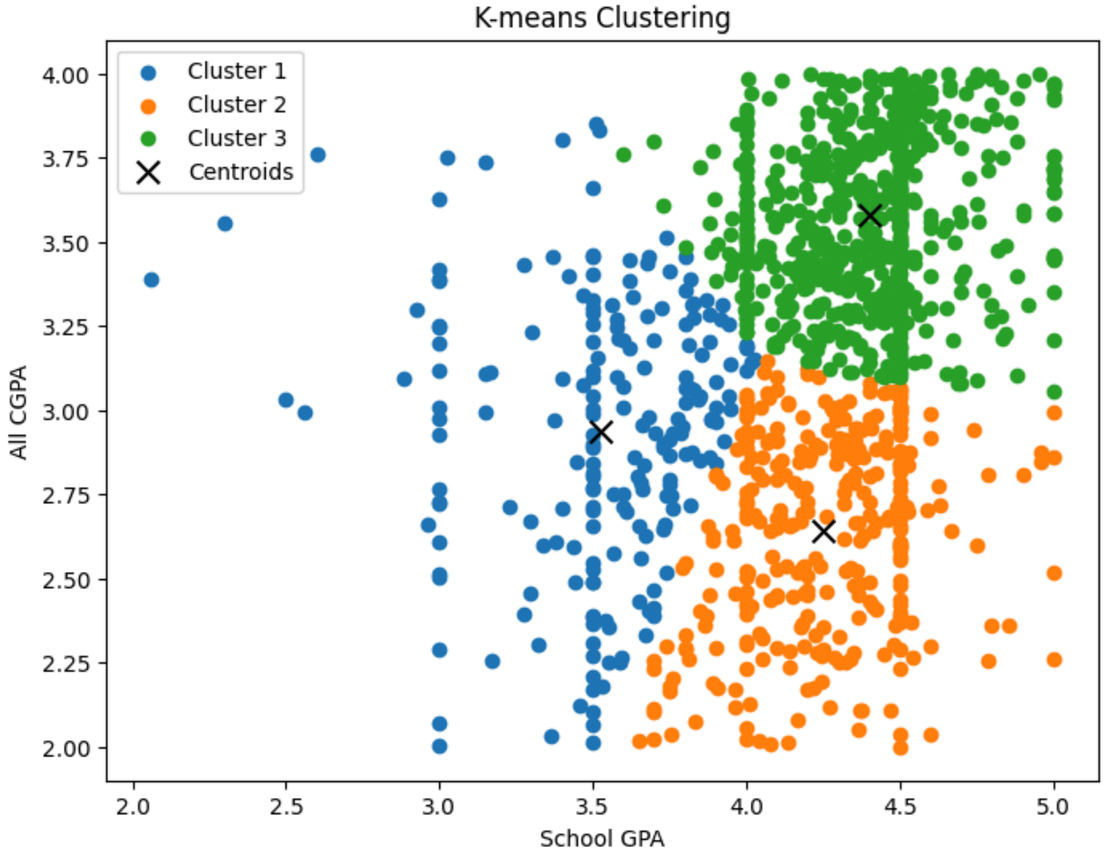

GMM Clustering
All CGPA
School GPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7

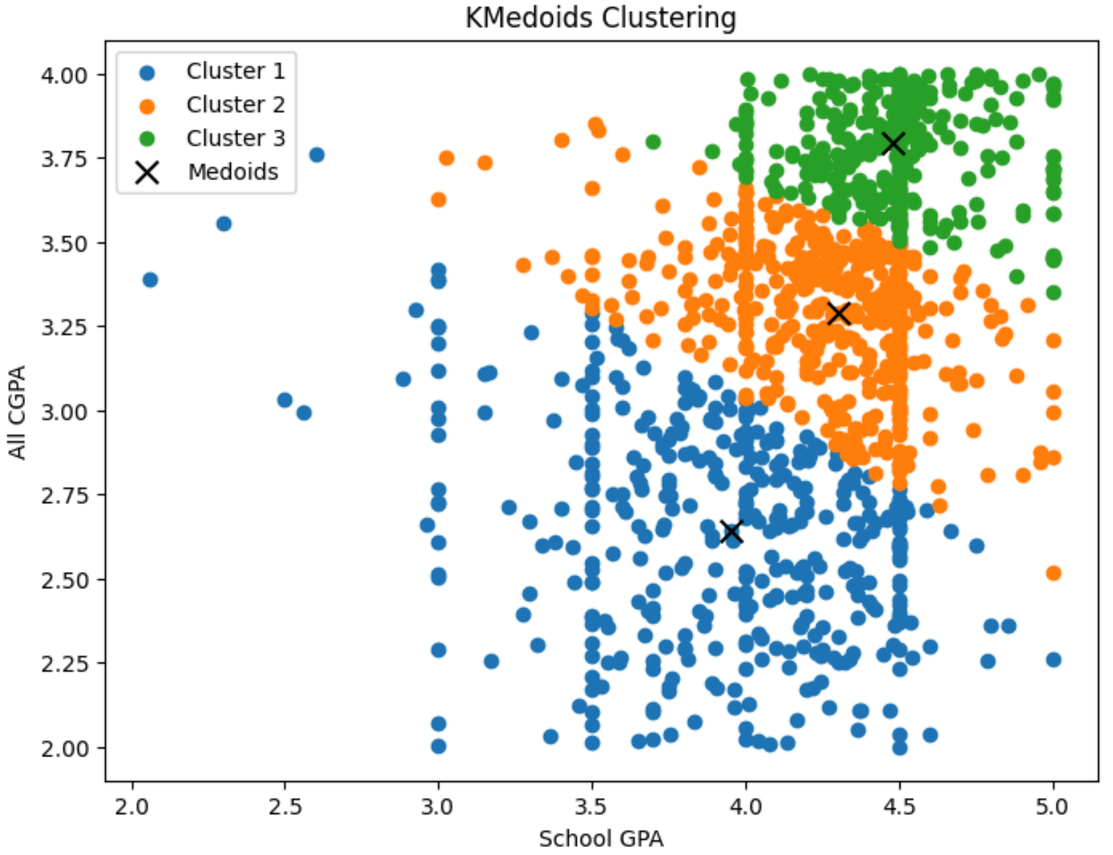
K-means VS GMM





K-means VS KMedoids



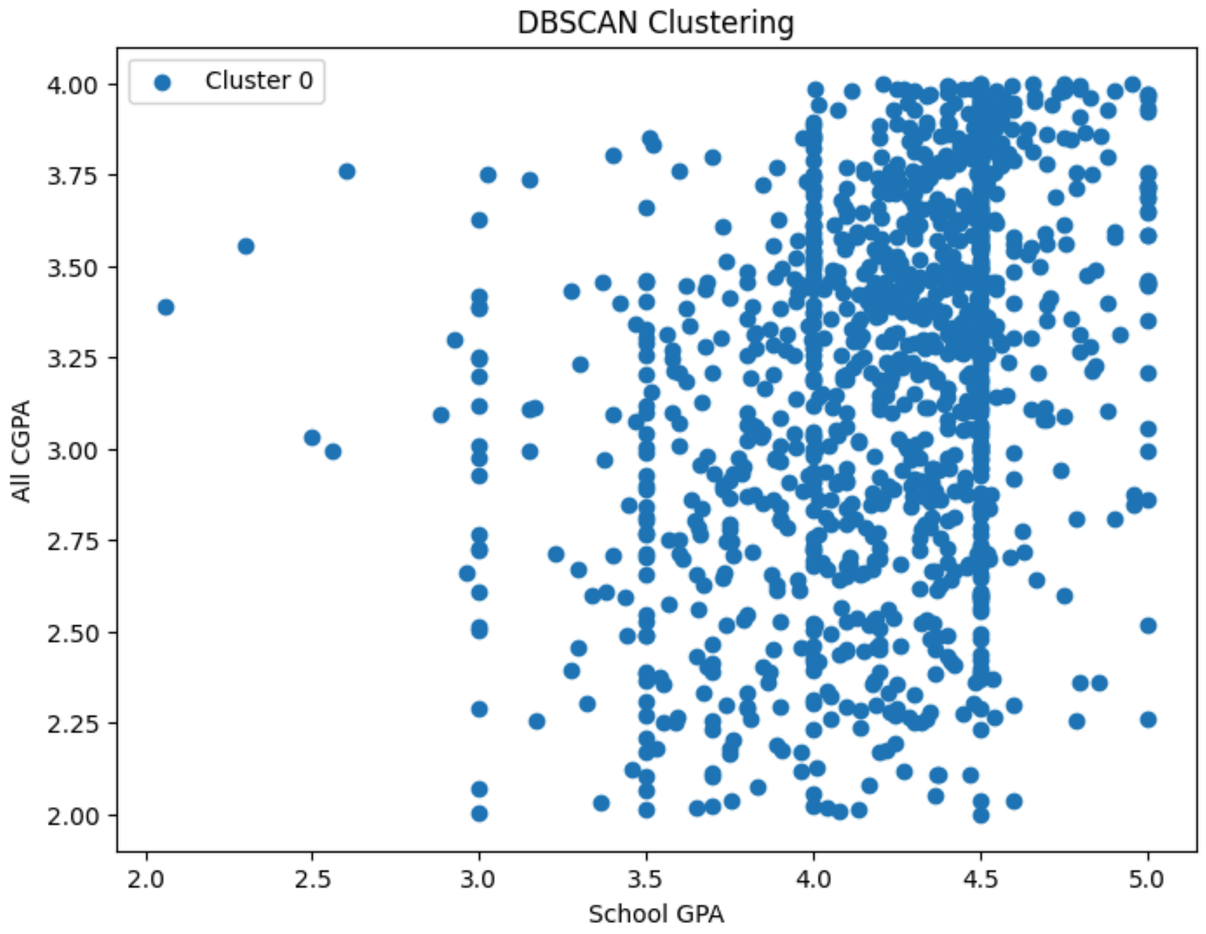


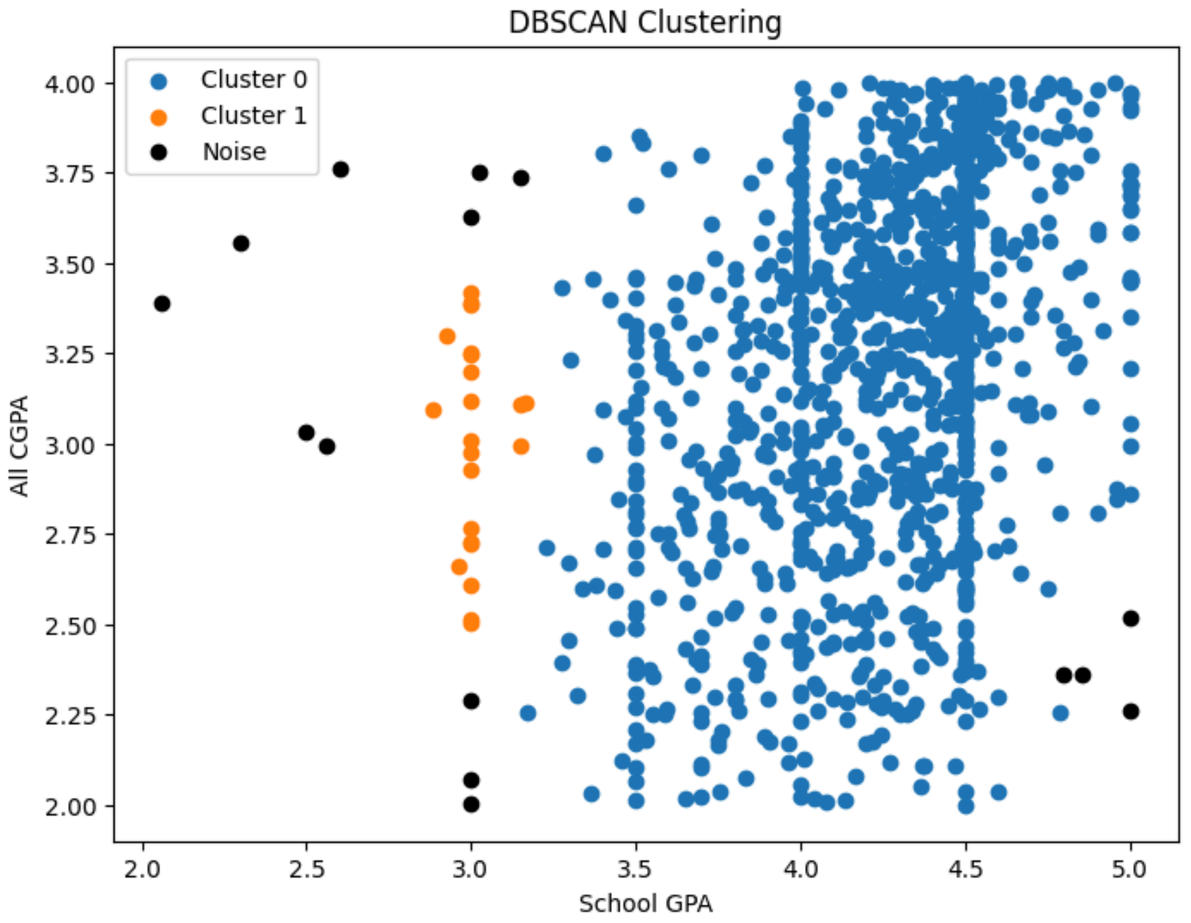
DBSCAN

eps_values = [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
min_samples_values = [3, 4, 5, 6, 7, 8, 9, 10]

Best combination: eps=0.5, min_samples=3,
Silhouette Score = 0.7430351975681541

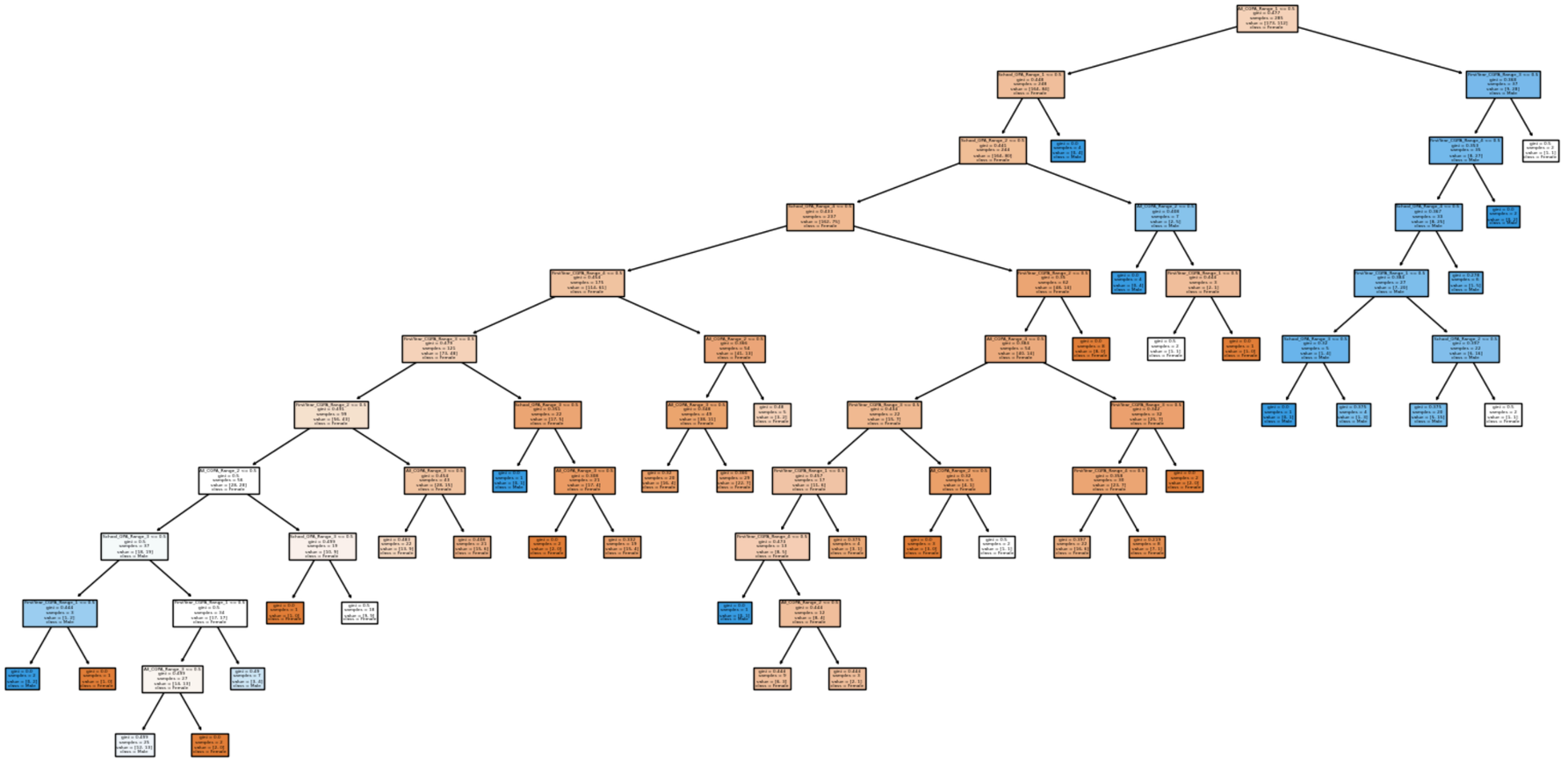
DBSCAN


DBSCAN Clustering
Cluster 0
All CGPA
School GPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
2.0
2.5
3.0
3.5
4.0
4.5
5.0


DBSCAN Clustering
Cluster 0
Cluster 1
Noise
All CGPA
School GPA
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Decision Tree

Accuracy: 0.7362



Bonus!

Deep Learning

Neural Network Classification

Why would we do it and how?

In this final part of our analysis of this dataset we decided to dive deeper into deep learning. To be more specific we will create a Neural Network which will predict from which major the student is.

We will build this model from the groud up, repeating the data cleaning process, since this time, data cleaning will be much different.

Why would we do it and how?

Here is how we will hande the missing data in our file for Deep Learning:

- For School_GPA, RoA and FirstYear_CGPA we will simply remove the missing values, since there are just the few of them.
- For All_CGPA we will do the following. For BSDS and BSES majors, since there is not data at all, we will take the average of FirstYear_CGPA
- For the rest of majors we will take respective averages of majors for All_CGPA

Why would we do it and how?

Since we have not enough data, we added 1000 data points

About the model

Here is the structure of our Neural Network:

As you can see our model consists of just 4 layers.

In the model we will use ReLU, which is considered the one of the best activation functions for classifying models.

And for the final step, we used LogSoftmax, as we are dealing with probabilities and Log makes the Softmax more stable.

About the model

```
Epoch 1, Loss: 1.6486810445785522
Epoch 11, Loss: 1.4832345247268677
Epoch 21, Loss: 1.330881118774414
Epoch 31, Loss: 1.2048976421356201
Epoch 41, Loss: 1.0913845300674438
Epoch 51, Loss: 0.9771910309791565
Epoch 61, Loss: 0.8650447726249695
Epoch 71, Loss: 0.7743023037910461
Epoch 81, Loss: 0.7179766297340393
Epoch 91, Loss: 0.6849932074546814
Epoch 101, Loss: 0.6619939208030701
Epoch 111, Loss: 0.6445524096488953
Epoch 121, Loss: 0.6309213638305664
Epoch 131, Loss: 0.6199328303337097
Epoch 141, Loss: 0.610632061958313
Epoch 151, Loss: 0.6024559736251831
Epoch 161, Loss: 0.5948640704154968
Epoch 171, Loss: 0.5875054001808167
Epoch 181, Loss: 0.5802965760231018
Epoch 191, Loss: 0.5732525587081909
Epoch 201, Loss: 0.566156268119812
Epoch 211, Loss: 0.5589570999145508
Epoch 221, Loss: 0.5519461035728455
Epoch 231, Loss: 0.5449673533439636
Epoch 241, Loss: 0.5380693674087524
Epoch 251, Loss: 0.5313946604728699
Epoch 261, Loss: 0.524890124797821
Epoch 271, Loss: 0.5185890793800354
Epoch 281, Loss: 0.5125582814216614
Epoch 291, Loss: 0.5068210959434509
```

We achived accuracy of 80%!

Thank You