Student Dropout Rates

Maeve O'Toole, Lindsay Spratt, Duha Kanjo

Our Dataset

- [Dataset Link]
- Demographics
- Socioeconomic Factors
- Academic Performance
- Student Retention

Columns of Data

Marital status
Course
Daytime/evening attendance
Previous qualification
Nationality
Mother's qualification
Father's qualification
Mother's occupation
Father's occupation
Displaced
Debtor
Tuition fees up to date

Gender
Scholarship holder
Age at enrollment
International
Curricular units 1st sem *(credited)*
Curricular units 1st sem *(enrolled)*
Curricular units 1st sem *(evaluations)*
Curricular units 1st sem *(approved)*
Unemployment Rate
GDP
Inflation Rate
Target *(type – object)*

Our Initial Focus -

- What factors are the biggest contributors to assessing the population as a whole?
- Do evening/nighttime classes affect dropout rates?
- **Debtors**.. whether or not financial stress has implications on **success rates**?
- What **societal pressures** have an impact on these implications?
- Be able to predict based on certain variables whether a student would fall into a **success** or **dropout** model

Scaled Data.head()

Scaling the data at an early stage makes it easier for a model to learn and understand the problem

	Marital status	Course	Daytime/evening attendance	Previous qualification	Nacionality	Mother's qualification	Father's qualification	Mother's occupation	Father's occupation	Displaced	...	Curricular units 1st sem (without evaluations)	Curricular units 2nd sem (credited)
0	-0.294829	-1.823744	0.350082	-0.386404	-0.145586	0.075111	-0.584526	-0.329669	0.449087	0.907512	...	-0.199273	-0.282442
1	-0.294829	0.254153	0.350082	-0.386404	-0.145586	-1.254495	-1.218380	-0.829997	-0.786461	0.907512	...	-0.199273	-0.282442
2	-0.294829	-1.131112	0.350082	-0.386404	-0.145586	1.072315	0.954834	0.670987	0.449087	0.907512	...	-0.199273	-0.282442
3	-0.294829	1.177663	0.350082	-0.386404	-0.145586	1.183116	0.954834	-0.329669	-0.786461	0.907512	...	-0.199273	-0.282442
4	1.356212	-1.592866	-2.856470	-0.386404	-0.145586	1.072315	1.045384	0.670987	0.449087	-1.101914	...	-0.199273	-0.282442
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4419	-0.294829	1.177663	0.350082	-0.386404	-0.145586	-1.254495	-1.399481	-0.329669	-0.580536	-1.101914	...	-0.199273	-0.282442
4420	-0.294829	1.177663	0.350082	-0.386404	10.150427	-1.254495	-1.399481	0.670987	0.449087	0.907512	...	-0.199273	-0.282442
4421	-0.294829	0.485030	0.350082	-0.386404	-0.145586	1.072315	0.954834	0.670987	0.449087	0.907512	...	-0.199273	-0.282442
4422	-0.294829	-0.207602	0.350082	-0.386404	-0.145586	1.072315	0.954834	0.170659	-0.580536	0.907512	...	-0.199273	-0.282442
4423	-0.294829	1.177663	0.350082	-0.386404	4.430420	1.183116	0.954834	-0.329669	0.449087	0.907512	...	-0.199273	-0.282442

4424 rows × 32 columns

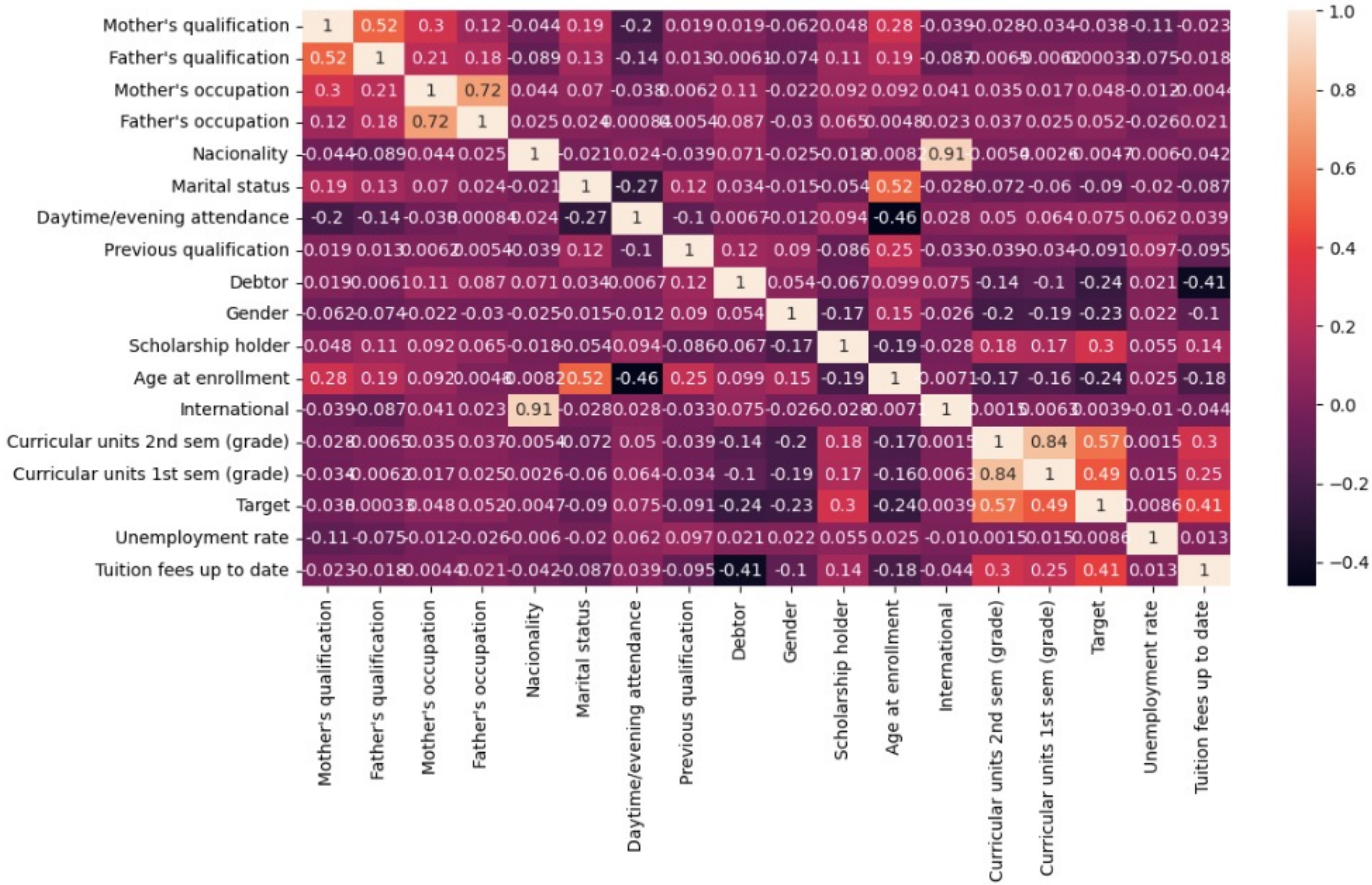
What factors have the highest positive correlation with each other?

Aka which variables grow proportionally to each other

	Marital status	Course	Daytime/evening attendance	Previous qualification	Nacionality	Mother's qualification	Father's qualification	Mother's occupation	Father's occupation	Displaced	...	Curricular units 1st sem (without evaluations)
Marital status	1.000000	0.018203	-0.274340	0.121026	-0.020702	0.185451	0.128230	0.069645	0.024280	-0.235162	...	0.034754
Course	0.018203	1.000000	-0.069024	-0.158734	-0.004832	0.059482	0.046156	0.030013	0.016712	0.006563	...	-0.060638
Daytime/evening attendance	-0.274340	-0.069024	1.000000	-0.103314	0.024386	-0.195084	-0.137476	-0.037701	0.001065	0.252521	...	0.045577
Previous qualification	0.121026	-0.158734	-0.103314	1.000000	-0.039038	0.019158	0.013408	0.006367	0.005499	-0.149168	...	0.018225
Nacionality	-0.020702	-0.004832	0.024386	-0.039038	1.000000	-0.043759	-0.088826	0.044197	0.024584	-0.010687	...	0.026184
Mother's qualification	0.185451	0.059482	-0.195084	0.019158	-0.043759	1.000000	0.524201	0.294850	0.115716	-0.076576	...	0.003440
Father's qualification	0.128230	0.046156	-0.137476	0.013408	-0.088826	0.524201	1.000000	0.206728	0.183780	-0.055628	...	-0.017661
Mother's occupation	0.069645	0.030013	-0.037701	0.006367	0.044197	0.294850	0.206728	1.000000	0.723963	-0.038951	...	-0.012480
Father's occupation	0.024280	0.016712	0.001065	0.005499	0.024584	0.115716	0.183780	0.723963	1.000000	-0.019579	...	-0.035241
Displaced	-0.235162	0.006563	0.252521	-0.149168	-0.010687	-0.076576	-0.055628	-0.038951	-0.019579	1.000000	...	-0.021554

10 rows × 32 columns

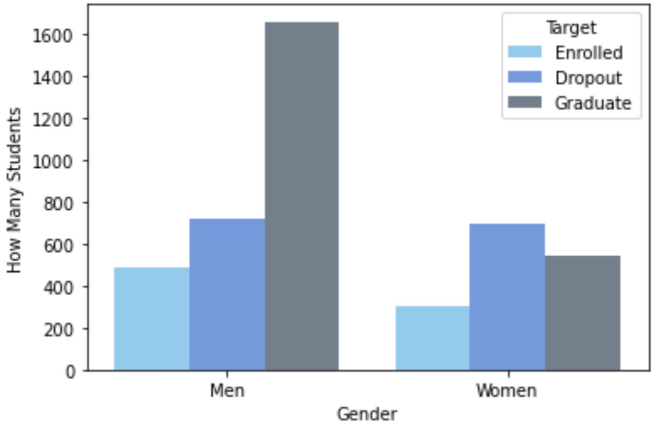
Correlation Heat Map to show the correlation between features in a more visually pleasing way



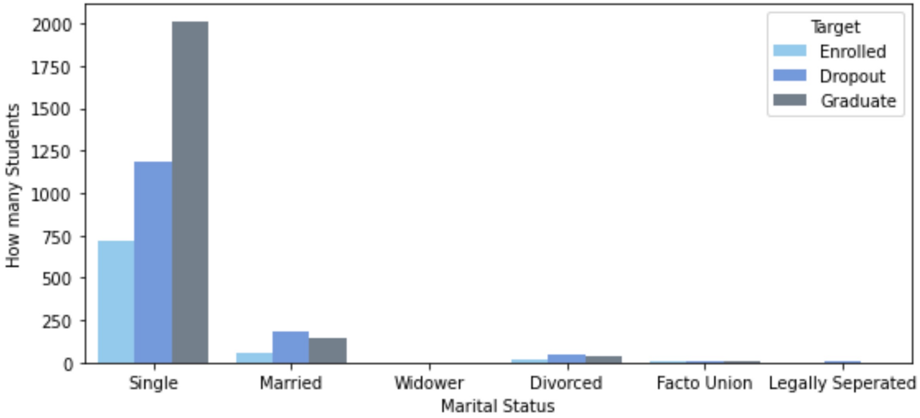
Initial Differing Variables Among Students

The biggest contributors to differentiating the population as a whole

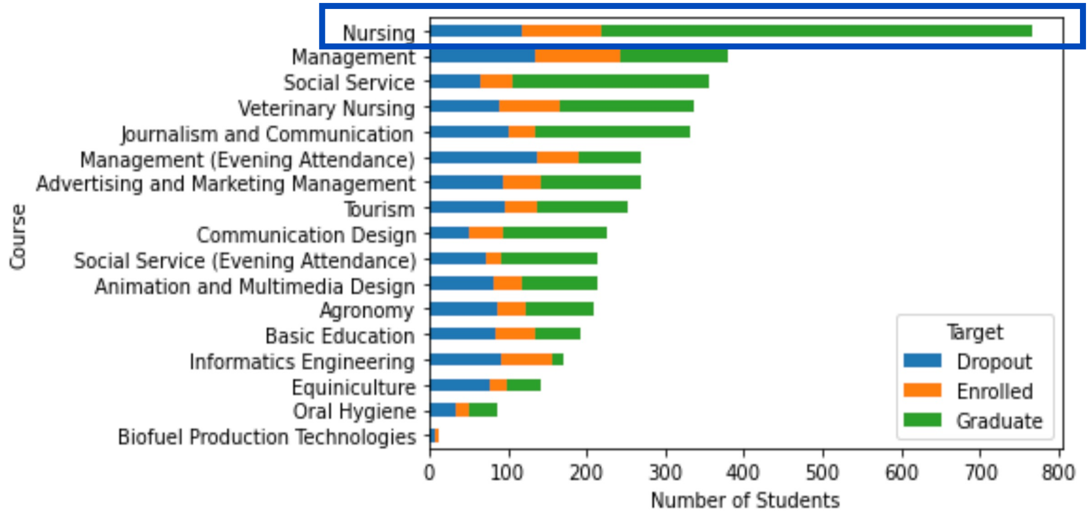
Gender



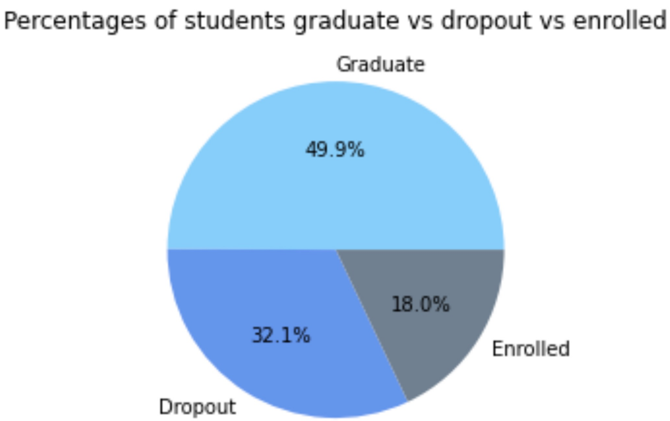
Marital Status



Taking a look at the distribution of our dataset...



We can see the various fields that students choose to study and their corresponding weights of whether these students are ***enrolled****, have* ***dropped out****, or have* ***graduated****.*



This pie chart splits the entire population based on the ***'Target'*** *column*

Split the data into training and test data to then use target prediction to get accuracy rates

In [36]:
```
target_prediction = bin_log.predict(X_test)
print(target_prediction)

```

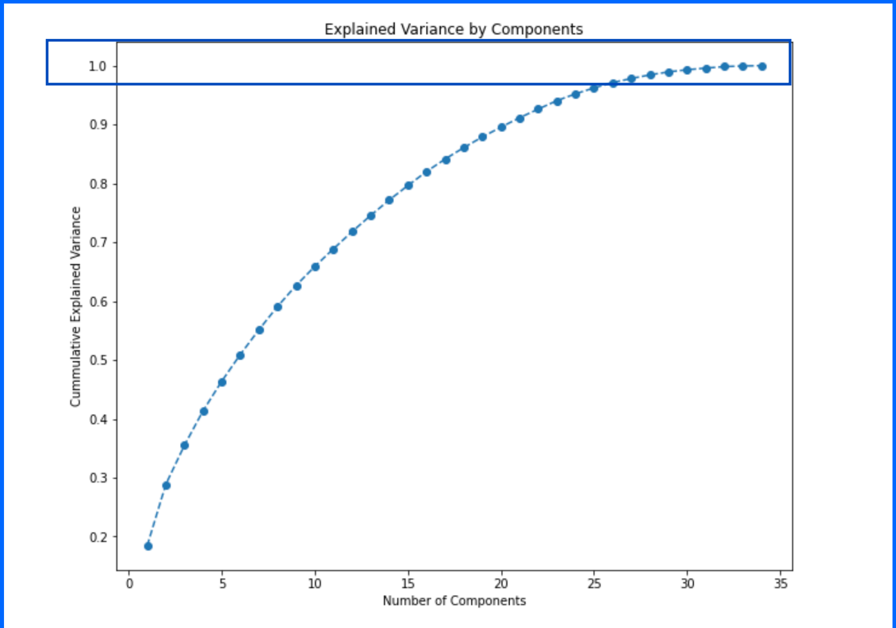
```
[2 0 1 2 2 2 2 1 0 1 2 2 2 2 1 0 0 2 2 0 2 2 2 0 2 1 2 0 2 2 2 2 2 0 2 2 1
 0 2 2 2 2 0 2 0 2 2 0 2 2 0 0 2 2 0 1 1 1 0 2 0 2 1 2 2 0 2 1 2 2 2 0 1 2
 2 2 1 0 2 1 2 2 0 1 2 0 2 2 0 2 2 0 2 2 0 1 0 2 1 0 2 2 2 0 0 2 2 0 2 2 2
 1 2 2 2 2 0 2 1 2 2 2 2 2 2 1 2 2 0 0 1 0 0 0 0 2 1 1 1 2 2 0 2 0 0 0 2 2
 1 2 2 2 1 1 2 2 2 0 2 2 2 2 0 0 2 1 2 1 0 2 0 2 2 0 2 2 2 2 2 2 0 0 2 2 2
 2 2 2 2 2 2 0 2 2 0 1 2 2 0 1 2 1 0 2 1 2 2 2 2 2 2 0 2 2 0 2 2 2 2 2 0 2
 0 2 2 0 0 2 0 2 2 0 2 1 2 2 0 0 2 1 0 2 2 2 1 2 2 0 0 2 2 2 0 2 2 2 2 2 2
 2 2 2 2 2 0 2 2 2 2 0 2 2 2 1 0 2 2 2 2 0 2 0 0 2 0 0 0 1 1 2 2 2 2 0 2 2
 2 2 1 0 1 2 2 1 1 2 2 0 2 0 0 0 2 2 0 2 1 2 0 2 2 0 1 0 0 0 2 2 2 2 0 2 2
 2 2 2 1 0 0 2 2 0 0 2 2 2 0 1 0 1 2 2 2 1 2 0 2 2 2 2 0 1 0 0 2 0 0 0 2 2
 1 2 2 0 2 0 0 2 2 2 0 2 0 2 0 2 1 2 2 1 0 0 2 0 2 0 2 0 2 2 0 2 0 1 0 1 2
 1 2 2 2 1 2 2 1 1 2 0 1 0 2 0 2 2 2 1 0 2 2 2 1 0 0 2 0 1 2 2 0 2 2 2 2 2
 1 2 2 2 0 2 1 2 0 1 2 2 2 0 1 2 1 0 1 0 2 2 2 2 0 2 1 1 0 0 2 2 0 1 0 0 2
 2 2 2 2 2 0 2 2 0 2 2 2 2 2 1 2 2 0 0 0 2 2 2 2 2 2 0 2 2 1 2 2 0 2 1 2 2
 1 0 0 0 0 0 0 2 0 0 2 2 0 1 0 0 0 1 2 2 1 0 0 2 0 1 2 2 2 0 1 2 2 2 2 2 0
 2 1 1 2 0 2 2 2 2 1 2 2 0 0 2 2 2 0 0 2 1 2 2 2 1 2 2 2 1 0 2 2 2 1 2 2 2
 2 2 1 1 1 2 0 2 0 2 2 2 2 2 1 0 0 2 2 1 2 1 2 2 1 2 1 0 0 0 2 0 0 0 2 2 2
 0 2 1 1 2 2 2 1 2 0 0 0 2 2 2 0 1 1 2 2 2 1 1 1 2 0 2 0 1 2 0 0 2 2 2 2 2
 1 2 2 0 2 2 1 2 1 0 2 2 0 2 0 2 2 0 2 0 0 1 2 1 0 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 0 2 0 0 2 0 0 2 0 0 0 1 2 2 0 0 2 2 1 0 2 2 2 2 0 2 2 0 2 1 0 0 1 2 0
 2 2 2 2 2 2 2 1 2 2 1 0 2 0 0 2 2 2 0 2 2 1 0 2 0 0 2 2 2 2 2 2 2 2 2 2 1
 2 2 2 2 2 2 1 2 2 2 2 2 0 2 2 2 2 0 2 1 2 2 0 2 2 0 0 0 1 2 2 2 2 0 0 0 2
 2 1 1 2 1 2 0 2 2 1 0 2 2 2 0 2 1 0 2 2 2 2 2 1 0 2 2 2 0 0 2 2 2 2 2 2 0
 2 2 2 2 2 2 1 2 2 2 2 0 2 1 0 0 1 1 0 2 2 1 1 2 2 0 1 0 2 0 2 2 2 1]
```

In [37]:
```
data_accuracy = accuracy_score(Y_test, target_prediction)
print("Accuracy:", data_accuracy)
```

```
Accuracy: 0.7762711864406779
```

~ 78%
accuracy

Variance Among Components within the dataset…

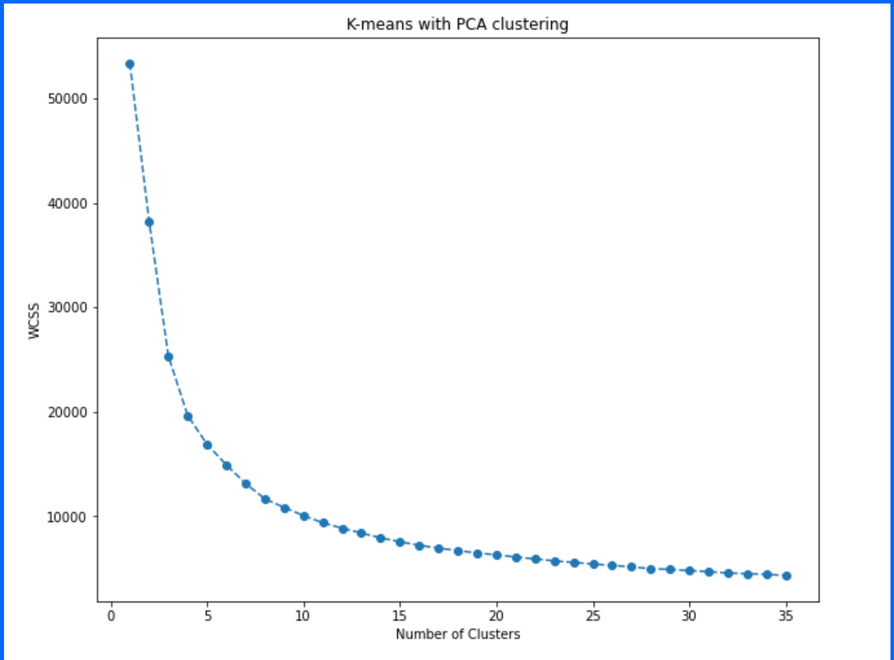


This graph is useful to see the **cumulative variance** which we can see displays that the **35** components, or features, explains **100%** of the data.

- Finding the ***principal component*** is a beneficial way to show correlated variables.

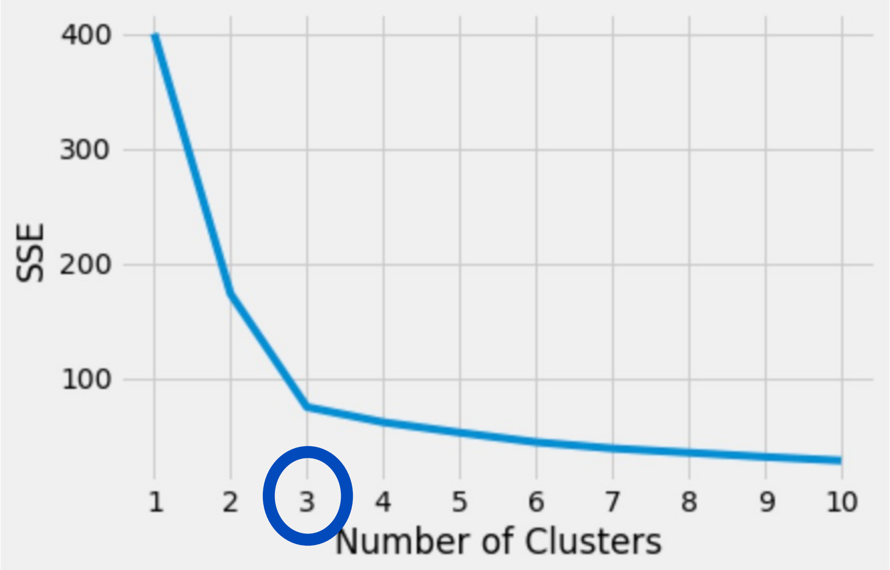
WCSS – ***within-cluster sum of square***

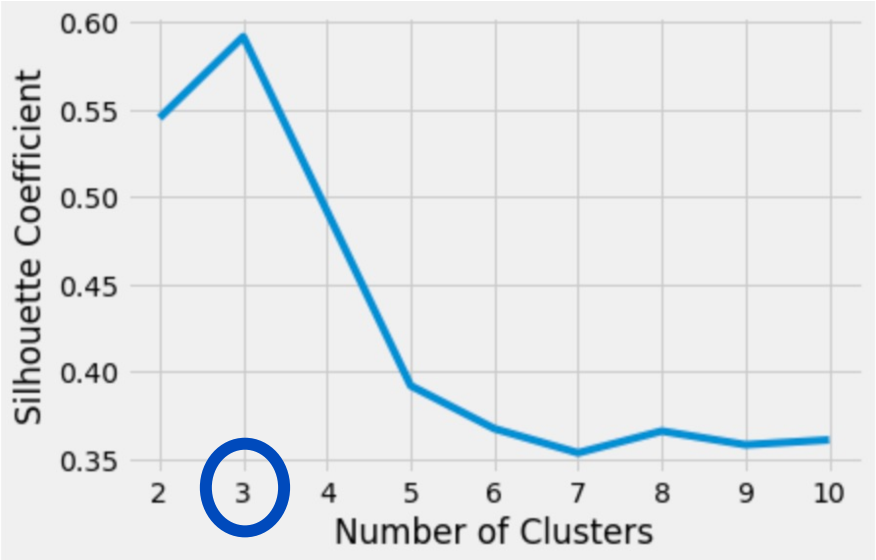
- The sum of the squared distance between **each point** and the **centroid** in a particular cluster
  - This is useful for clustering to ensure, based on each feature, the distance between the data point and the centroids



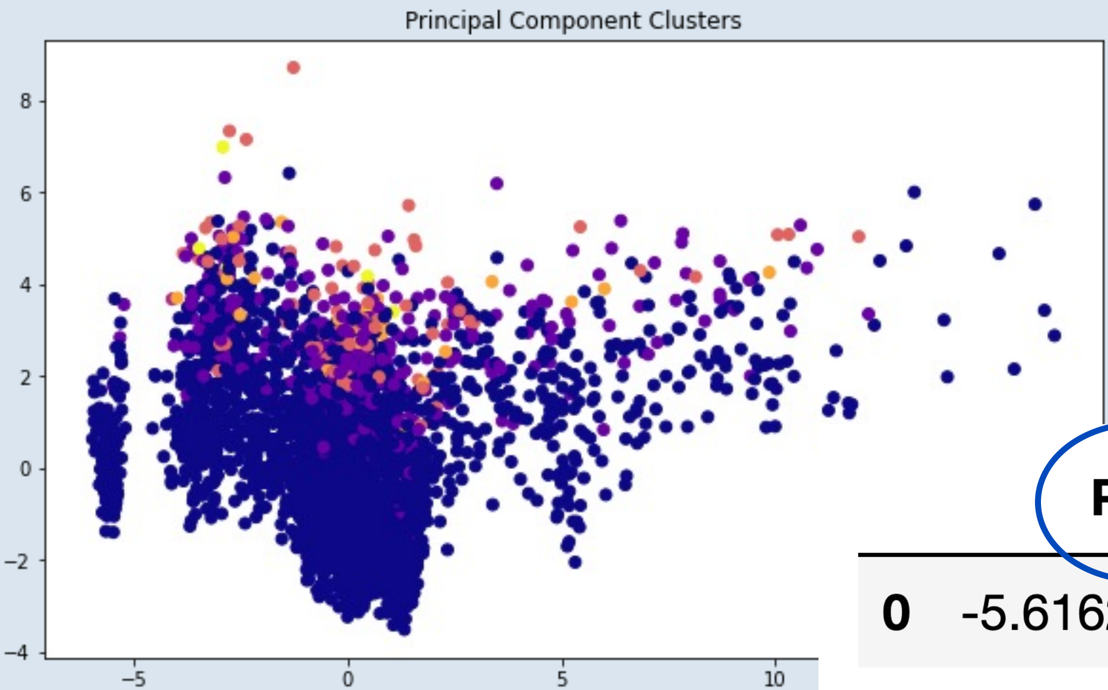
Elbow and Silhouette Graphs to predict our k value…

K should equal 3!





Principal Component Analysis Clusters



The first function we used for K-Means did not show evident clusters, when we changed the number of clusters, there was no change in the graph.

	PC1	PC2	PC3
0	-5.616263	-0.191381	0.854592
1	-0.299552	-0.946696	1.938030
2	-4.018853	0.510819	-0.114211
3	0.414862	-1.073880	-0.622193
4	0.375114	2.699581	-2.521393

We are obtaining our ***most relevant features***.

Clustering over our entire dataset…

After having another team member attempt, our K-Means graph improved drastically

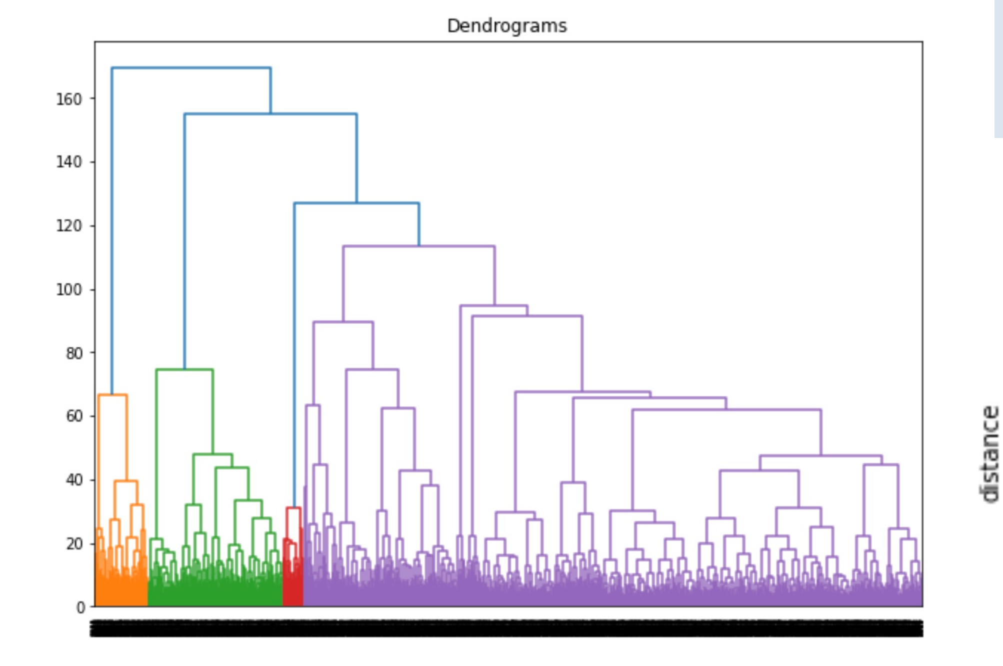
Here we have the 3 clusters that we got from the K Means

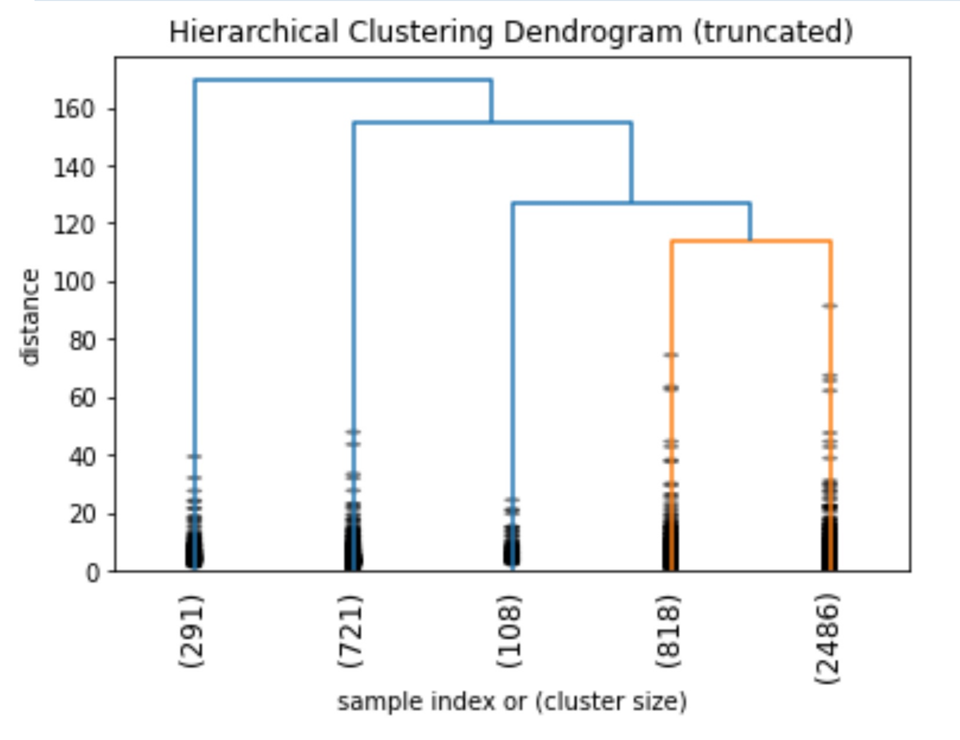
	0	1	2
Mother's qualification	8.071429	7.750000	1.000000
Father's qualification	8.090909	8.090909	4.545455
Mother's occupation	2.161290	1.290323	3.612903
Father's occupation	1.200000	2.000000	2.800000
Nacionality	1.000000	1.000000	1.000000
Marital status	6.400000	2.800000	1.000000
Daytime/evening attendance	10.000000	10.000000	10.000000
Previous qualification	1.000000	1.000000	1.000000
Debtor	1.000000	1.000000	1.000000
Gender	10.000000	1.000000	10.000000
Scholarship holder	1.000000	1.000000	1.000000
Age at enrollment	1.679245	1.169811	2.698113
International	1.000000	1.000000	1.000000
Curricular units 2nd sem (grade)	7.396923	6.330769	7.930000
Curricular units 1st sem (grade)	6.098319	7.437086	1.000000
Target	10.000000	1.000000	1.000000
Unemployment rate	4.348837	2.883721	6.023256
Tuition fees up to date	10.000000	10.000000	10.000000

Features within each cluster…

Cluster 1	Cluster 2	Cluster 3
High Mother qualification	The High Mother qualification	Low Mother qualification
High Father's qualification	High Father's qualification	Low Father's qualification
High value of Previous qualification	Low value of Previous qualification	High value of Previous qualification
Legally Separated	Widower	Single
High grade In the 2ns semester	low grade In the 2ns semester	High grade in the 2ns semester
Has the High Employment rate	Has the Low Employment rate	Has the Highiest Employment rate
Tuition fees is up to date	Tuition fees is up to date	Tuition fees is up to date
Scholarship holder Student	Scholarship holder Student	NOT Scholarship holder Student

These **dendrograms** for our entire dataset are branching diagrams that represents the ***relationships of similarities*** among the entire dataset



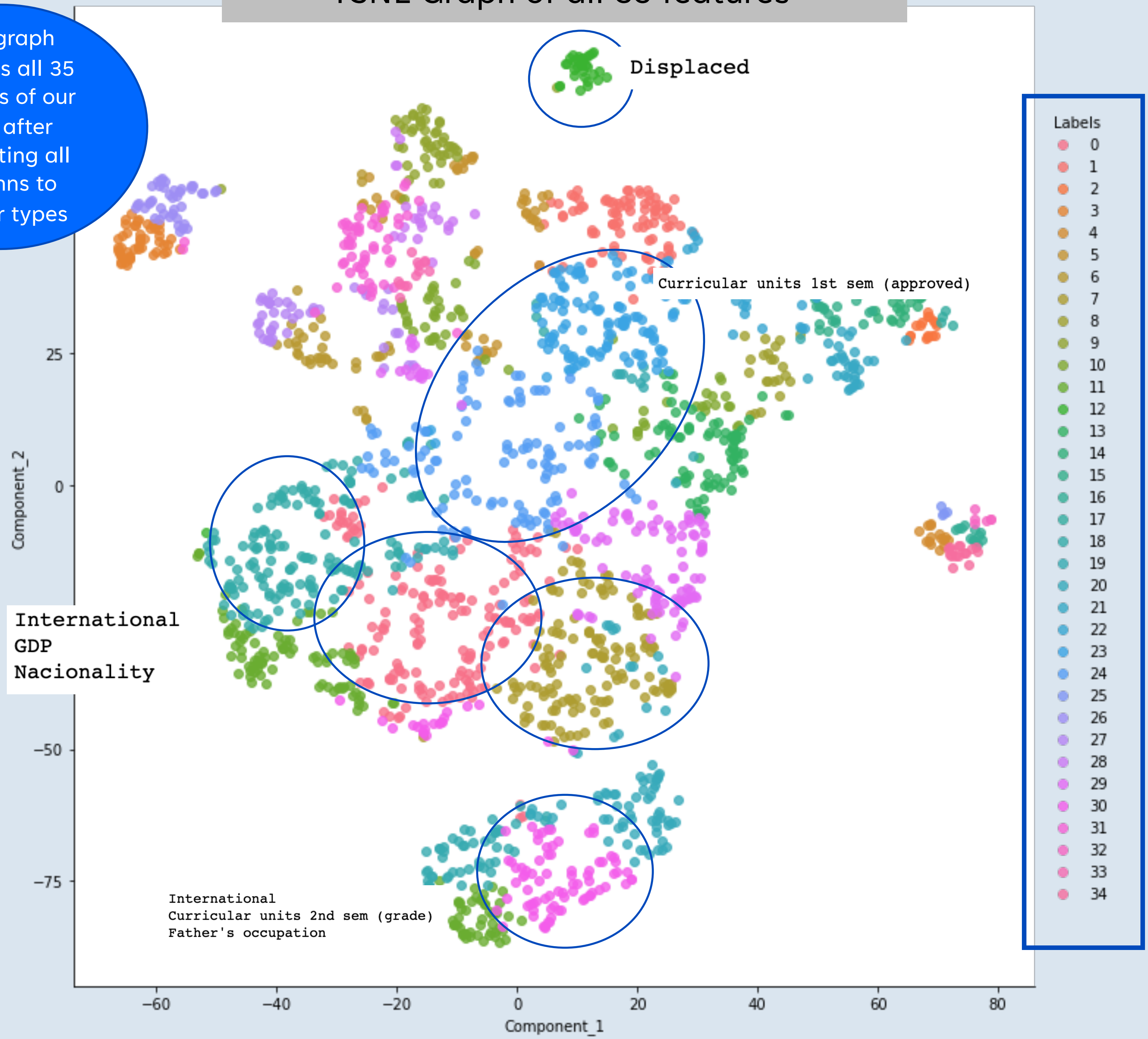


This type of graph is a tree-structured graph that is used in heat maps to visualize the result of a **hierarchical clustering** calculation.

The result of a clustering is presented as both the **distance and the similarity** between the **clustered columns**.

TSNE Graph of all 35 features

This graph includes all 35 features of our data after converting all columns to integer types



Using *Chi-Squared test* to find out which variables have the greatest impact on a student's success academically

```
In [68]: X_train_chi2=X_train[['Marital status',"Father's occupation", "Mother's occupation",'Gender',
                                "Debtor"]]
```

```
In [69]: from sklearn.feature_selection import chi2
         f_score_p=chi2(X_train_chi2,y_train)
         f_score_p

Out[69]: (array([ 10.78673053,  96.45896671,  46.40816666, 110.66314568,
                 160.13673399]),
          array([4.54664689e-03, 1.13292592e-21, 8.36747940e-11, 9.32831582e-25,
                 1.68558260e-35]))
```

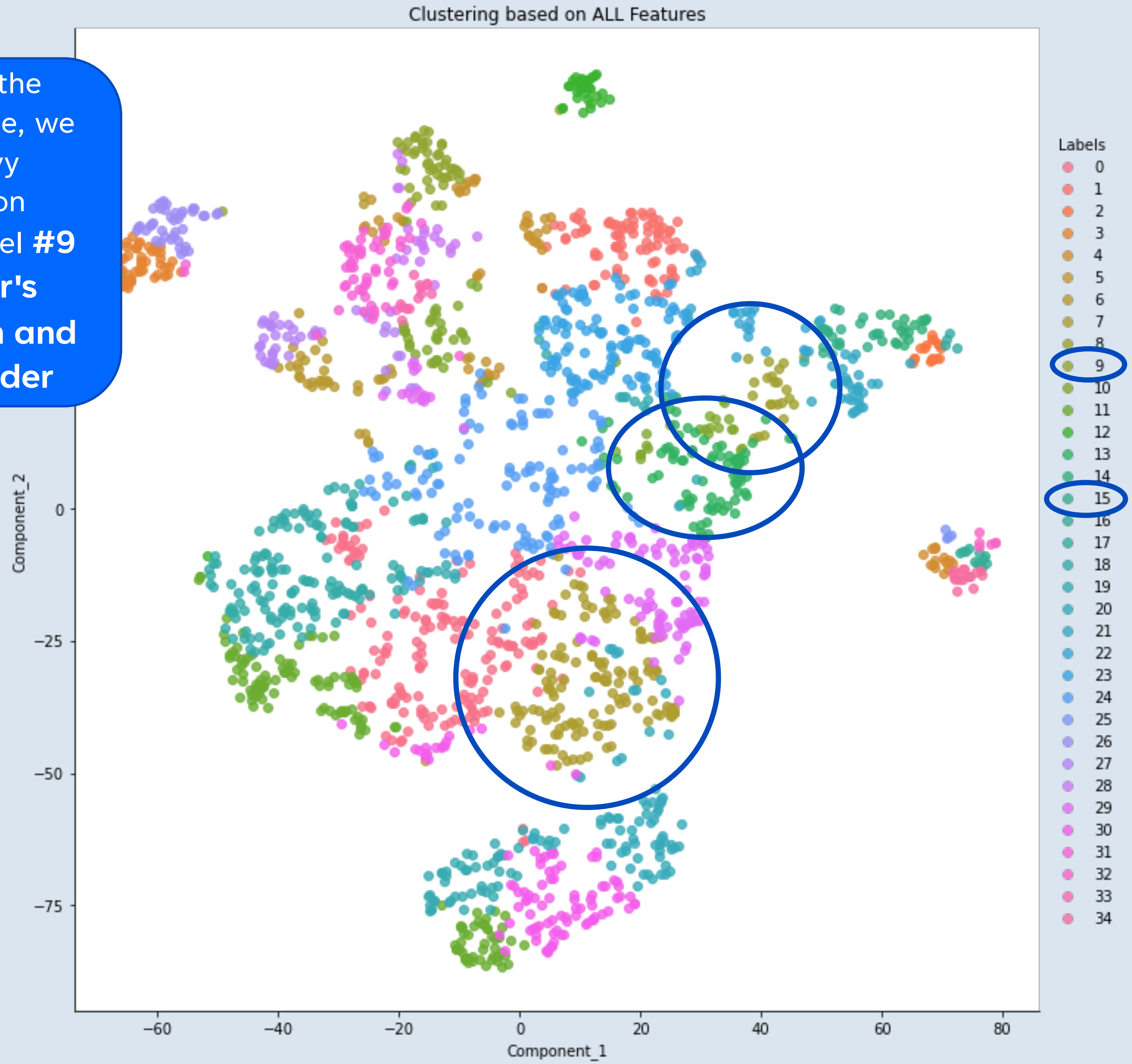
```
In [70]: p_values=pd.Series(f_score_p[1])
         p_values.index=X_train_chi2.columns
         p_values.sort_values()

Out[70]: Debtor                 1.685583e-35
         Gender                 9.328316e 25
         Father's occupation    1.132926e-21
         Mother's occupation    8.367479e 11
         Marital status         4.546647e-03
         dtype: float64
```

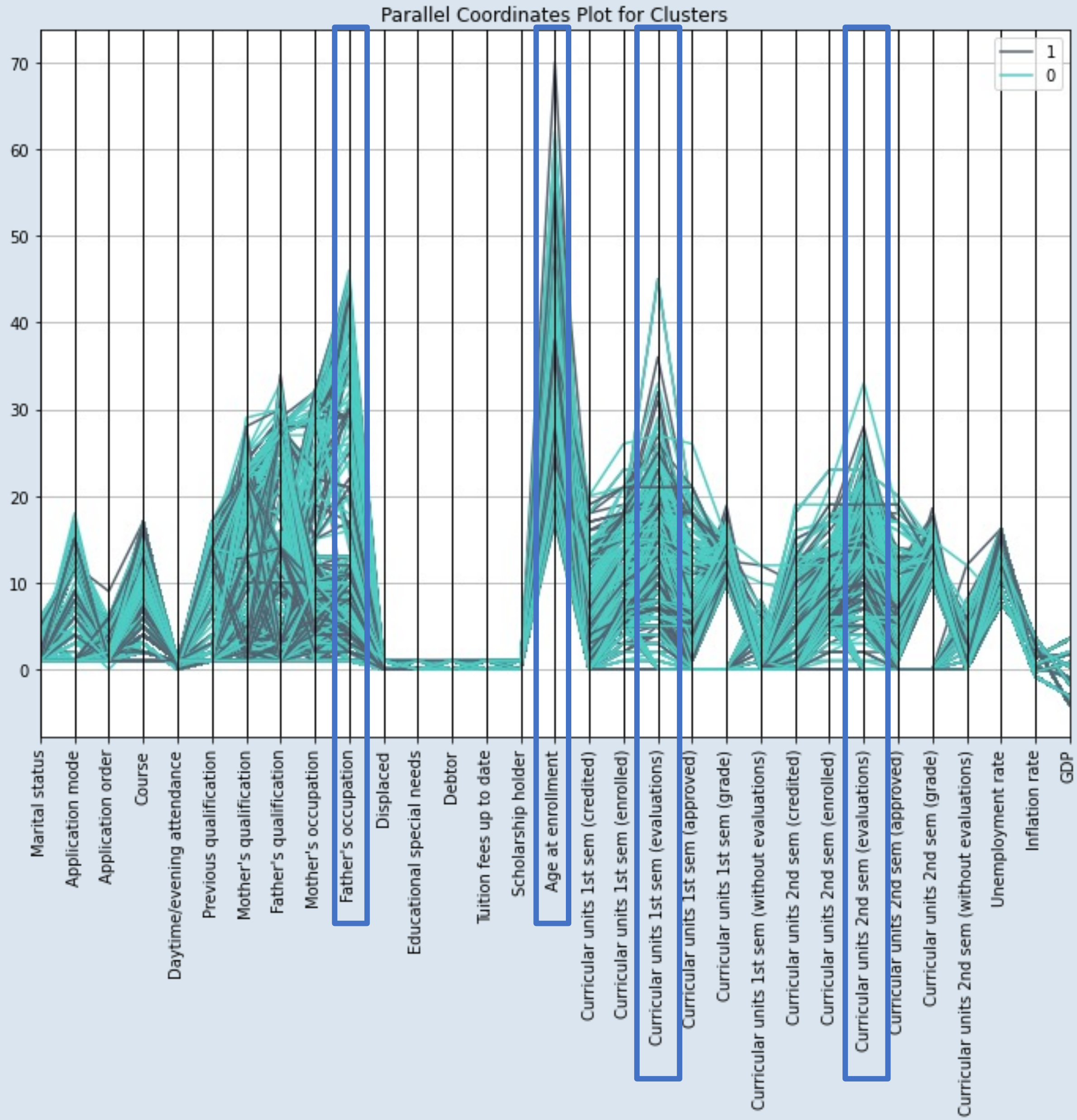
*According to this test, we can see that **GENDER** and **MOTHER'S OCCUPATION** have the biggest impact on a student's success*

This test is used for classification purposes...

K-Means Clustering on PCA Embeddings…

Based on the previous slide, we see heavy correlation between label **#9 – Mother's occupation and #15 - Gender**





Parallel coordinates graph for the features of our dataset split by gender

- 1: Men

- 2: Women

Father's occupation, age at enrollment, curricular units 1st semester (evaluations), curricular units 2nd semester (evaluations)

How could we use our findings in a real-life scenario?

- Technology companies (Quizlet, Chegg, etc.) could use these clusters to understand the customer body more and target students who are likely to drop out.

- Universities and Graduate Schools could use this information to understand their student body more while also understanding what features may play into a student's success.
  - NEXT STEP:
    - A recommender system that would be able to predict the outcome of a student's academic success based on certain features.

Thank you!