# PROJECT 4: UNSUPERVISED LEARNING

Jesus Barrera Mejia

ITCS 3162 | Professor Aileen | 04-17-2024

# The Problem: “What are the defining characteristics for colleges?”

My goal for this project is to find the different categories of colleges and what the defining features are for colleges in the categories that my clustering algorithm creates. These categories will hopefully give us insight into why colleges are different and how the colleges in these respective categories perform compared to the others.

# What is Clustering?

Clustering is a statistical method with the objective of organizing the different groups that our samples belong to. The goal is to find the different patterns that our samples exhibit and group together the samples that are the closest in distance. A difficulty in this process is that most of our data has more than two features making it difficult to find the right features to tell if they are “close together” or not. To solve this problem, we use ***dimension reduction*** techniques that will help use summarize the features that best describe the data into one or three features that allow us to measure the distance from our samples.

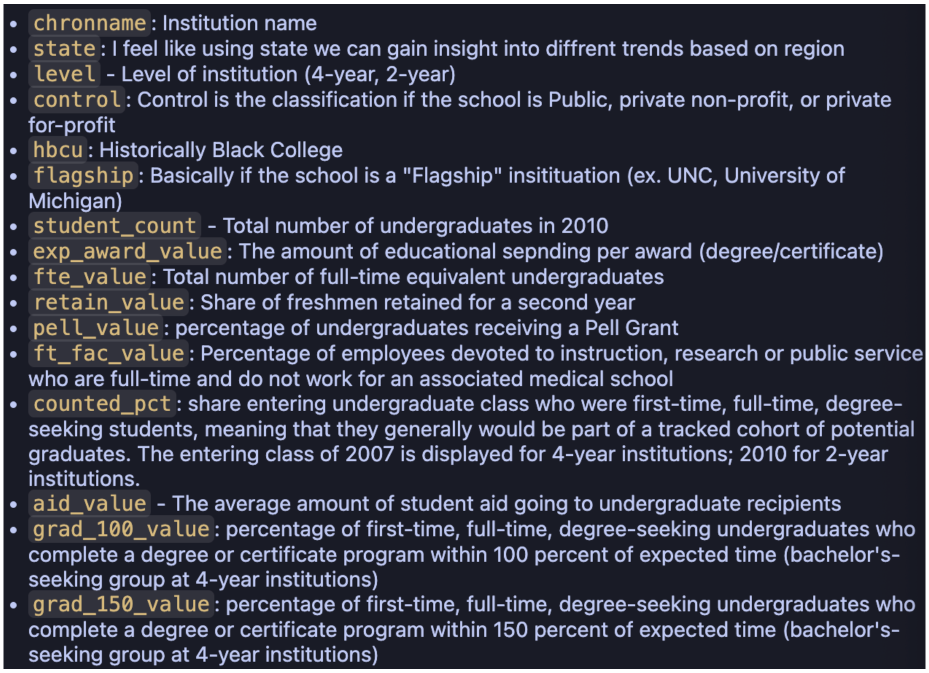
## K-Means Clustering

K-Means clustering is a clustering method where we use centroids and the distance from those centroids to assign our data to groups. This method provides us with the flexibility to split our data into as many clusters as we would like. Once we know how many clusters we want, we can randomly assign clusters and start grouping our samples to those clusters based on their distance from them. Once we have samples assigned, we can calculate the **mean** of our cluster and check if our samples have changed clusters. If they change, we repeat our process back from step one. The goal of this process is to ensure that each cluster is capturing an equal amount of variation and is not capturing everything, otherwise we won't really gain any insight into our data because the cluster is basically the entire dataset.

## Agglomerative Clustering

Agglomerative clustering is a clustering method that is based on creating groups from the samples based on the distance however, the goal here is to create a hierarchy or stack of our samples where the most similar samples are at the top. We start the process of ***agglomerative clustering*** by comparing the distances of the samples and grouping them together if they are close together. Once we have our initial group, we can now compare the group's distance to the other samples based on either the minimum, maximum, or the average. We continue this process repeatedly until we have one group that contains all our different groups of samples.

# The Data

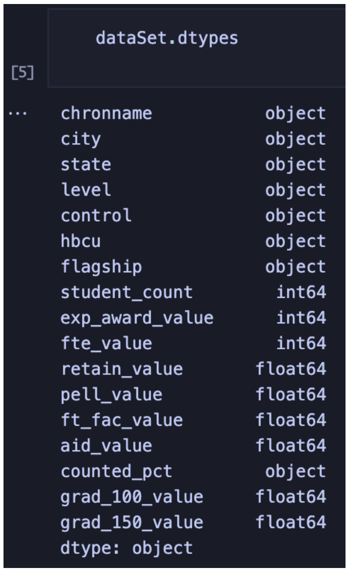


- `chronname`: Institution name
- `state`: I feel like using state we can gain insight into diffrent trends based on region
- `level` - Level of institution (4-year, 2-year)
- `control`: Control is the classification if the school is Public, private non-profit, or private for-profit
- `hbcu`: Historically Black College
- `flagship`: Basically if the school is a "Flagship" insitituation (ex. UNC, University of Michigan)
- `student_count` - Total number of undergraduates in 2010
- `exp_award_value`: The amount of educational sepnding per award (degree/certificate)
- `fte_value`: Total number of full-time equivalent undergraduates
- `retain_value`: Share of freshmen retained for a second year
- `pell_value`: percentage of undergraduates receiving a Pell Grant
- `ft_fac_value`: Percentage of employees devoted to instruction, research or public service who are full-time and do not work for an associated medical school
- `counted_pct`: share entering undergraduate class who were first-time, full-time, degree-seeking students, meaning that they generally would be part of a tracked cohort of potential graduates. The entering class of 2007 is displayed for 4-year institutions; 2010 for 2-year institutions.
- `aid_value` - The average amount of student aid going to undergraduate recipients
- `grad_100_value`: percentage of first-time, full-time, degree-seeking undergraduates who complete a degree or certificate program within 100 percent of expected time (bachelor's-seeking group at 4-year institutions)
- `grad_150_value`: percentage of first-time, full-time, degree-seeking undergraduates who complete a degree or certificate program within 150 percent of expected time (bachelor's-seeking group at 4-year institutions)

My dataset was pulled from Kaggle – https://www.kaggle.com/datasets/leilahasan/college-completion. My data is tracking the trends with colleges. Initially there were 50 features for my dataset, but many of these features had numerous null values and were not very easy to understand so I removed them and after this process we are left with 16 features.

# Data Understanding: Pre-Preprocessing

## Data Types and Buckets



```
dataSet.dtypes
[5]

chronname            object
city                 object
state                object
level                object
control              object
hbcu                 object
flagship             object
student_count         int64
exp_award_value       int64
fte_value             int64
retain_value        float64
pell_value          float64
ft_fac_value        float64
aid_value           float64
counted_pct          object
grad_100_value      float64
grad_150_value      float64
dtype: object
```

The first thing I want to figure out about my data is what data type is being used for each column. We are most concerned about any feature that has an object data type because we want to pass it through our PCA algorithm and K-Means algorithm which requires the feature to be numerical. For our object data type, we also want to figure out if it's a formatting issue or if there are buckets, we find that we need to format and create buckets for:

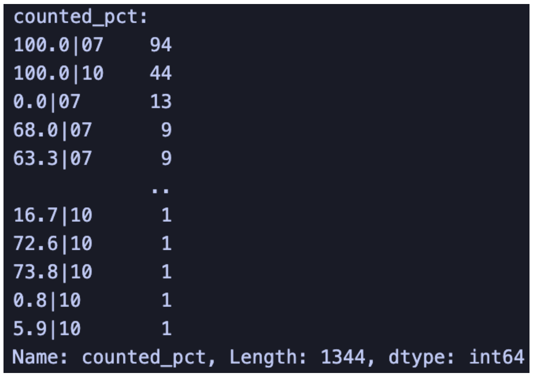
1. **Counted_pct:** Has a formatting issue where it includes the percentage and the year, we need to remove the year so we can make it a numerical value.
2. **State:** We are going to use this feature to engineer a new **"Regions"** feature that will be separated by south, northeast, etc.
3. **Level:** We need to one-hot encode this feature.
4. **Control:** Needs to be one-hot encoded.

5. **HBCU:** Needs to be one-hot encoded.

6. **Flagship:** Needs to be one-hot encoded.

# Preprocessing

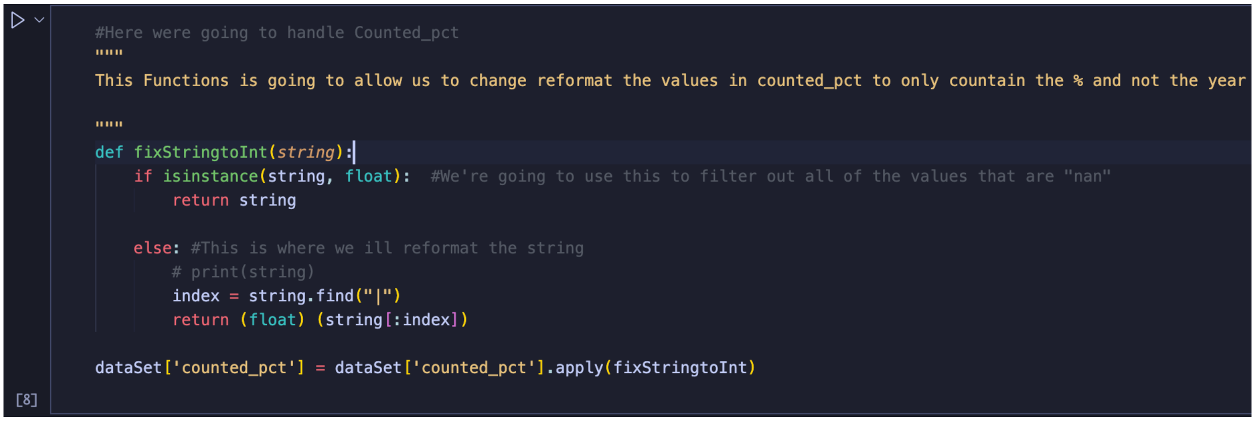
## Counted_pct – Reformatting the Variable



```
counted_pct:
100.0|07     94
100.0|10     44
0.0|07       13
68.0|07       9
63.3|07       9
             ..
16.7|10       1
72.6|10       1
73.8|10       1
0.8|10        1
5.9|10        1
Name: counted_pct, Length: 1344, dtype: int64
```

***Counted_pct*** is an interesting feature because it has a **"|"** separating two numbers. The number we are interested in is the one on the left of the dash because it is the percent of first year students that are being tracked in our study. To fix this formatting, I use the **apply()** function which will apply a function and remove the dash, while keeping only the numbers on the left.

Here is the function and the application:



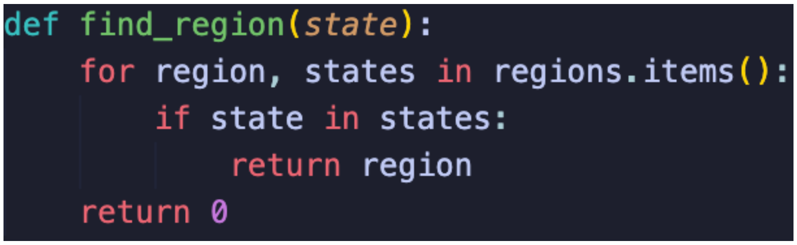
```
#Here were going to handle Counted_pct
"""
This Functions is going to allow us to change reformat the values in counted_pct to only countain the % and not the year

"""
def fixStringtoInt(string):
    if isinstance(string, float):  #We're going to use this to filter out all of the values that are "nan"
        return string

    else: #This is where we ill reformat the string
        # print(string)
        index = string.find("|")
        return (float) (string[:index])

dataSet['counted_pct'] = dataSet['counted_pct'].apply(fixStringtoInt)
```

## Flagship, HBCU, Level, and Control – Dummy Variables and One-hot encoding

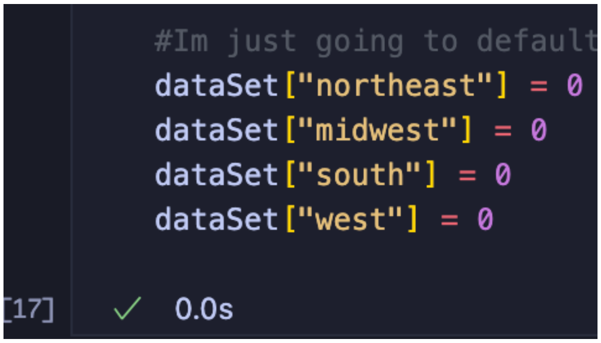
- **Flagship and HBCU:** Flagship and HBCU are interesting because null values are used to indicate that the sample does not fall under this category. The process to one-hot encode these is identical and we simply replace Null with 0 and "X" with 1 respectively.
- **Level and Control:** For level, we will be doing some ***feature engineering,*** and we will create a new feature called **"is_four_year"**. If 1 the sample is a four-year college and if 0 then they are a 2-year college. For **Control**, we use the get_dummies() feature because the feature has three buckets. It is a very simple process.

## State – Engineering a New Feature "Region."

Like **counted_pct**, I created a function for the state feature, and I will use that feature to engineer four new features which are **"northeast", "midwest", "south", "west".**

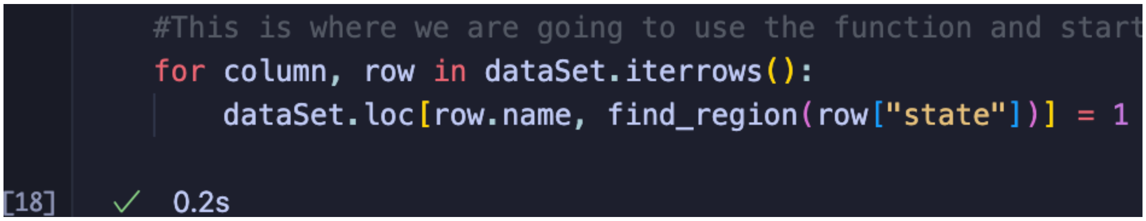


```python
def find_region(state):
    for region, states in regions.items():
        if state in states:
            return region
    return 0
```



```python
#Im just going to default
dataSet["northeast"] = 0
dataSet["midwest"] = 0
dataSet["south"] = 0
dataSet["west"] = 0
```

```
[17]  ✓  0.0s
```



```
#This is where we are going to use the function and start
for column, row in dataSet.iterrows():
    dataSet.loc[row.name, find_region(row["state"])] = 1
[18]  ✓  0.2s
```

## Null Values and Duplicates

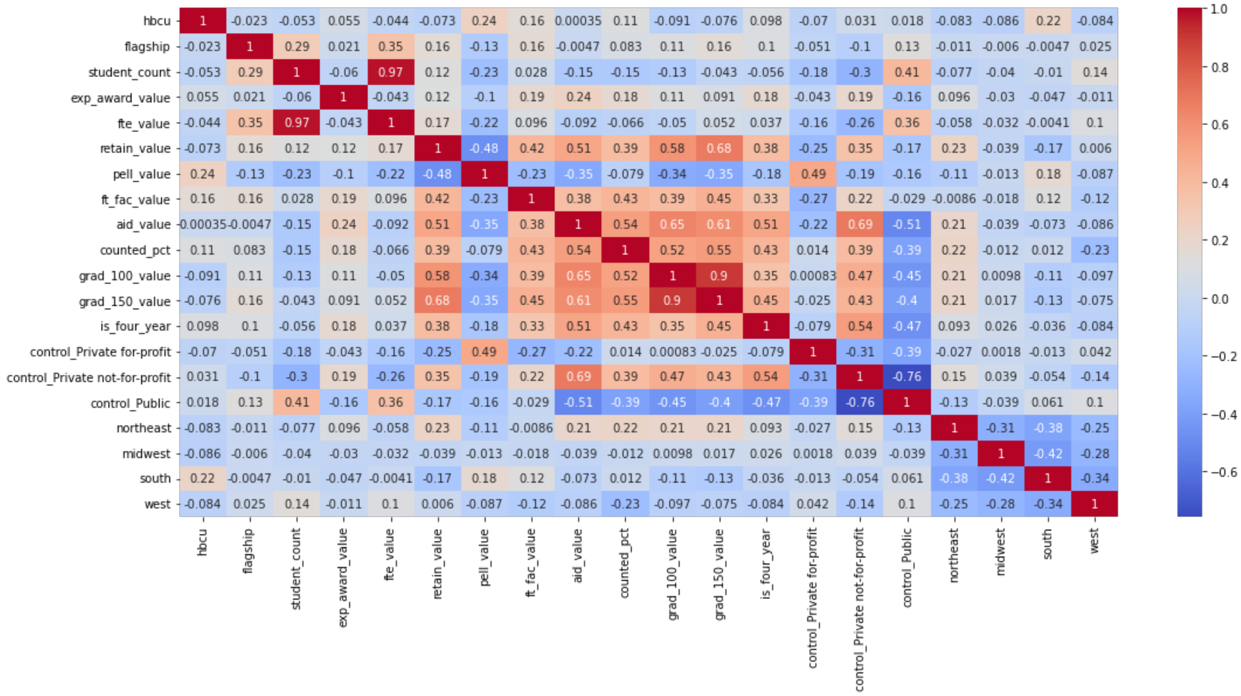
Since we dealt with the most important null values, we can just use the dropna() function. Luckily, our dataset did not include any duplicates.

# Data Understanding/Visualization: Post-Preprocessing

## Describe Function

In cell 23 of my notebook, I ran the describe function on our newly processed dataset. Two features stood out to me **HBCU and Flagship** because they make up a very small amount of the samples in the dataset.

## Heatmap



| | hbcu | flagship | student_count | exp_award_value | fte_value | retain_value | pell_value | ft_fac_value | aid_value | counted_pct | grad_100_value | grad_150_value | is_four_year | control_Private for-profit | control_Private not-for-profit | control_Public | northeast | midwest | south | west |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| hbcu | 1 | -0.023 | -0.053 | 0.055 | -0.044 | -0.073 | 0.24 | 0.16 | 0.00035 | 0.11 | -0.091 | -0.076 | 0.098 | -0.07 | 0.031 | 0.018 | -0.083 | -0.086 | 0.22 | -0.084 |
| flagship | -0.023 | 1 | 0.29 | 0.021 | 0.35 | 0.16 | -0.13 | 0.16 | -0.0047 | 0.083 | 0.11 | 0.16 | 0.1 | -0.051 | -0.1 | 0.13 | -0.011 | -0.006 | -0.0047 | 0.025 |
| student_count | -0.053 | 0.29 | 1 | -0.06 | 0.97 | 0.12 | -0.23 | 0.028 | -0.15 | -0.15 | -0.13 | -0.043 | -0.056 | -0.18 | -0.3 | 0.41 | -0.077 | -0.04 | -0.01 | 0.14 |
| exp_award_value | 0.055 | 0.021 | -0.06 | 1 | -0.043 | 0.12 | -0.1 | 0.19 | 0.24 | 0.18 | 0.11 | 0.091 | 0.18 | -0.043 | 0.19 | -0.16 | 0.096 | -0.03 | -0.047 | -0.011 |
| fte_value | -0.044 | 0.35 | 0.97 | -0.043 | 1 | 0.17 | -0.22 | 0.096 | -0.092 | -0.066 | -0.05 | 0.052 | 0.037 | -0.16 | -0.26 | 0.36 | -0.058 | -0.032 | -0.0041 | 0.1 |
| retain_value | -0.073 | 0.16 | 0.12 | 0.12 | 0.17 | 1 | -0.48 | 0.42 | 0.51 | 0.39 | 0.58 | 0.68 | 0.38 | -0.25 | 0.35 | -0.17 | 0.23 | -0.039 | -0.17 | 0.006 |
| pell_value | 0.24 | -0.13 | -0.23 | -0.1 | -0.22 | -0.48 | 1 | -0.23 | -0.35 | -0.079 | -0.34 | -0.35 | -0.18 | 0.49 | -0.19 | -0.16 | -0.11 | -0.013 | 0.18 | -0.087 |
| ft_fac_value | 0.16 | 0.16 | 0.028 | 0.19 | 0.096 | 0.42 | -0.23 | 1 | 0.38 | 0.43 | 0.39 | 0.45 | 0.33 | -0.27 | 0.22 | -0.029 | -0.0086 | -0.018 | 0.12 | -0.12 |
| aid_value | 0.00035 | -0.0047 | -0.15 | 0.24 | -0.092 | 0.51 | -0.35 | 0.38 | 1 | 0.54 | 0.65 | 0.61 | 0.51 | -0.22 | 0.69 | -0.51 | 0.21 | -0.039 | -0.073 | -0.086 |
| counted_pct | 0.11 | 0.083 | -0.15 | 0.18 | -0.066 | 0.39 | -0.079 | 0.43 | 0.54 | 1 | 0.52 | 0.55 | 0.43 | 0.014 | 0.39 | -0.39 | 0.22 | -0.012 | 0.012 | -0.23 |
| grad_100_value | -0.091 | 0.11 | -0.13 | 0.11 | -0.05 | 0.58 | -0.34 | 0.39 | 0.65 | 0.52 | 1 | 0.9 | 0.35 | 0.00083 | 0.47 | -0.45 | 0.21 | 0.0098 | -0.11 | -0.097 |
| grad_150_value | -0.076 | 0.16 | -0.043 | 0.091 | 0.052 | 0.68 | -0.35 | 0.45 | 0.61 | 0.55 | 0.9 | 1 | 0.45 | -0.025 | 0.43 | -0.4 | 0.21 | 0.017 | -0.13 | -0.075 |
| is_four_year | 0.098 | 0.1 | -0.056 | 0.18 | 0.037 | 0.38 | -0.18 | 0.33 | 0.51 | 0.43 | 0.35 | 0.45 | 1 | -0.079 | 0.54 | -0.47 | 0.093 | 0.026 | -0.036 | -0.084 |
| control_Private for-profit | -0.07 | -0.051 | -0.18 | -0.043 | -0.16 | -0.25 | 0.49 | -0.27 | -0.22 | 0.014 | 0.00083 | -0.025 | -0.079 | 1 | -0.31 | -0.39 | -0.027 | 0.0018 | -0.013 | 0.042 |
| control_Private not-for-profit | 0.031 | -0.1 | -0.3 | 0.19 | -0.26 | 0.35 | -0.19 | 0.22 | 0.69 | 0.39 | 0.47 | 0.43 | 0.54 | -0.31 | 1 | -0.76 | 0.15 | 0.039 | -0.054 | -0.14 |
| control_Public | 0.018 | 0.13 | 0.41 | -0.16 | 0.36 | -0.17 | -0.16 | -0.029 | -0.51 | -0.39 | -0.45 | -0.4 | -0.47 | -0.39 | -0.76 | 1 | -0.13 | -0.039 | 0.061 | 0.1 |
| northeast | -0.083 | -0.011 | -0.077 | 0.096 | -0.058 | 0.23 | -0.11 | -0.0086 | 0.21 | 0.22 | 0.21 | 0.21 | 0.093 | -0.027 | 0.15 | -0.13 | 1 | -0.31 | -0.38 | -0.25 |
| midwest | -0.086 | -0.006 | -0.04 | -0.03 | -0.032 | -0.039 | -0.013 | -0.018 | -0.039 | -0.012 | 0.0098 | 0.017 | 0.026 | 0.0018 | 0.039 | -0.039 | -0.31 | 1 | -0.42 | -0.28 |
| south | 0.22 | -0.0047 | -0.01 | -0.047 | -0.0041 | -0.17 | 0.18 | 0.12 | -0.073 | 0.012 | -0.11 | -0.13 | -0.036 | -0.013 | -0.054 | 0.061 | -0.38 | -0.42 | 1 | -0.34 |
| west | -0.084 | 0.025 | 0.14 | -0.011 | 0.1 | 0.006 | -0.087 | -0.12 | -0.086 | -0.23 | -0.097 | -0.075 | -0.084 | 0.042 | -0.14 | 0.1 | -0.25 | -0.28 | -0.34 | 1 |

This heatmap is powerful because it shows us the features that have the biggest impact on other features, and we can make a safe bet that **private not-for-profit** and **public** colleges are two that we will see when we inspect our unsupervised learning model. For our heat map, we find a couple of correlated features (these are the features that are darker in the shades of blue or red/orange), **Grad_100 and Grad_150** have a very high positive correlation because **Grad_100** is a mathematical artifact of **Grad_150.** Also, there are mutiple features that have correlation with other features: For example, **retain_value, is_four_year, aid_value**.

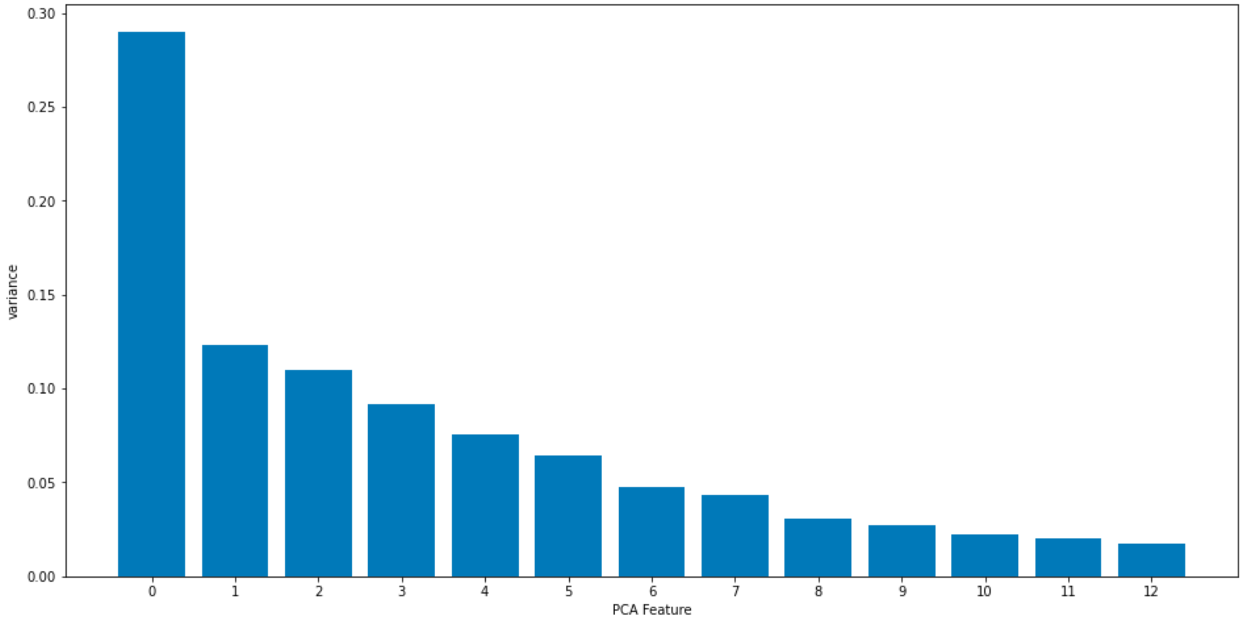
## Modeling: K-Means Algorithm

To begin the modeling stage there a couple of processing steps that must be followed:

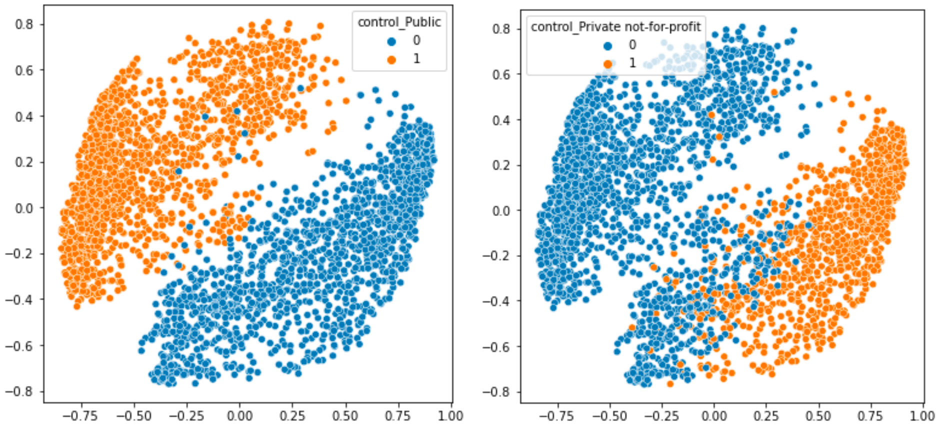
1. Standardize the Data (*cell 29)*
2. Normalize the Data (*cell 29*)
3. Use the Standardized/Normalized data to create new PCA components. (*Cell 29*)
4. Train model (*cell 35*)

We also drop a couple of features that help identify the samples, but do not serve a purpose for the model. I chose the **K-means algorithm** because it's easy to implement and visualize. However, it's important to keep in mind that with the K-Means algorithm, the clusters can get affected by outliers, so we want to keep an eye out for that.

## PCA

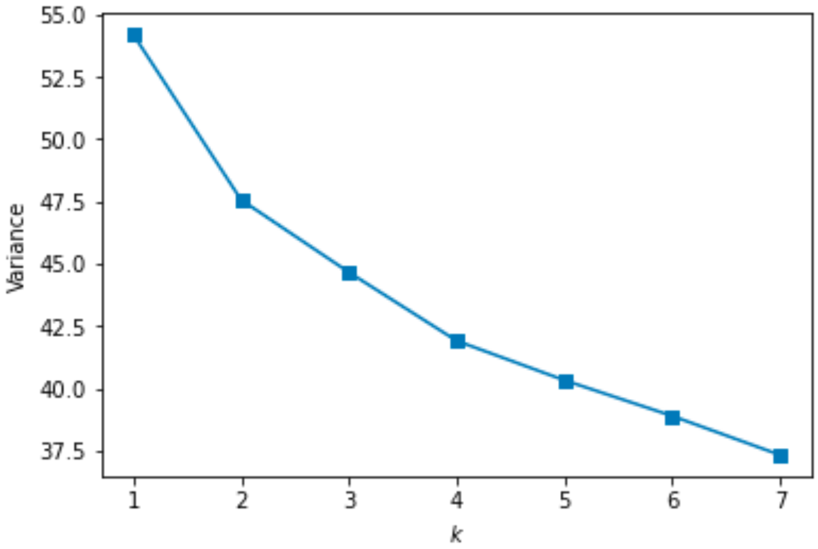


The following PCA plot is a demonstration of the 12 PCA components that we will use for our data. The reason we choose 12 was because we wanted to capture at least 95% of the variation in the data, but it's important to note that the first two PCA components already capture 41% of the variation. We can also plot these two PCA components to find what features are capturing the variation.

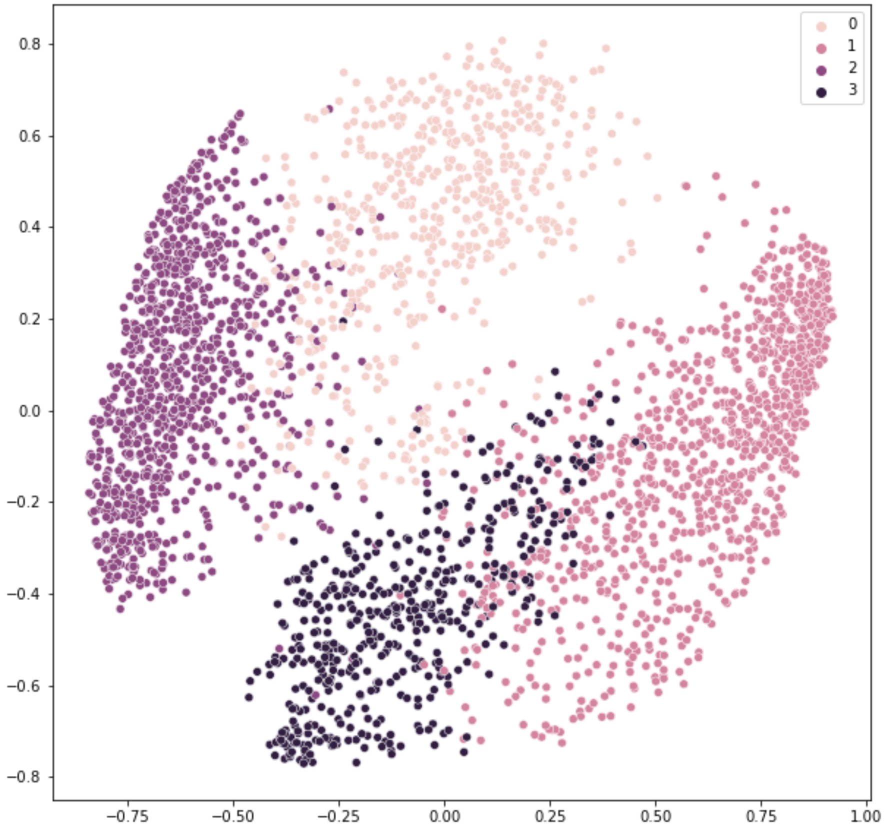


Using the information from the heatmap, we find that the first two PCA components have identifable clusters based on public and private not-for-profit. Now lets see how our K-Mean alogrithm will categorize our data.

## K-Means Algorithm



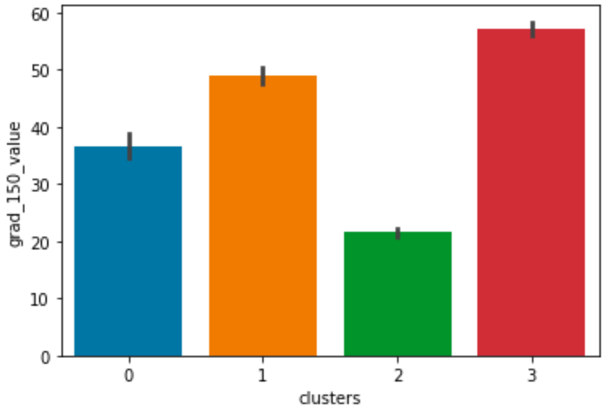
To implement our model, our first step is to find the number of ***Centroids*** that will reduce the amount of variance in each cluster. Based on our graph and using the **Elbow Method** we choose 4 since this is where we begin to lose the amount of marginal utility of using more clusters.

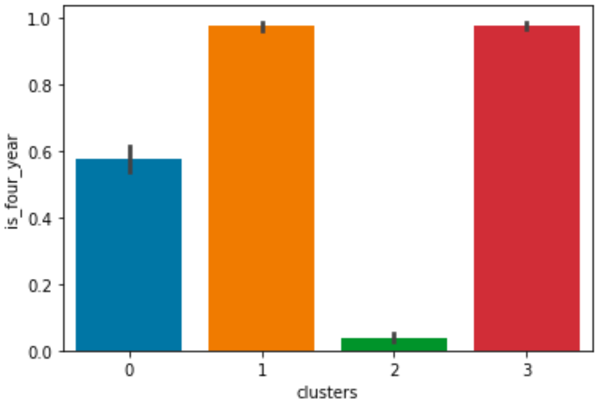


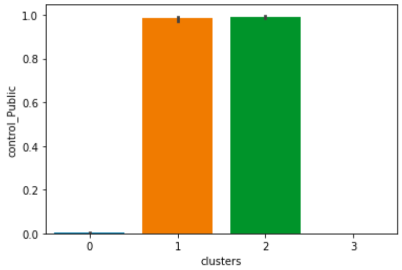
Using the four clusters, we generate the graph above from our **K-Means Clustering Algorithm**. This scatter plot plots **PCA 0 and PCA 1** and shows that by using these PCA components, our K-Means algorithm find the following clusters in our data. Let's dive a little deeper and figure out what these clusters tell us.

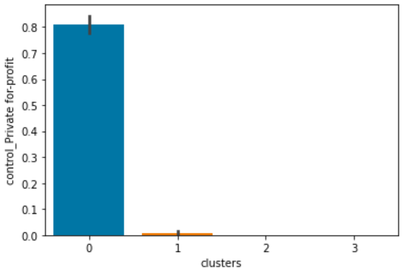
## Storytelling: Cluster Analysis

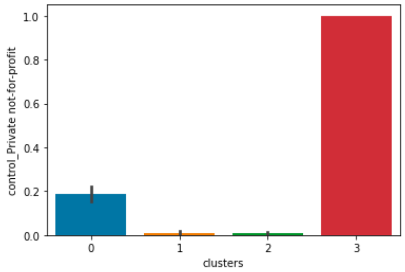
The **K-Means algorithm** finds four identifiable groups in our data. In cell 42 of the notebook, I use the **plotly** library to create an interactive visualization of the clusters which allows us to gain insight into the characteristics of these clusters. Additionally, we visualize some information to see some of the characteristics of these clusters in the following section.











In cells 60 - 68, we create **bar graphs** using our new clusters. Using this information, we can make the following assumptions.

1. **Cluster 0/Mostly Private For-Profit (All Levels)** has mostly all the private for-profit schools and some private not-for-profits school. (Including four-year and two-year colleges)
2. **Cluster 1/Public Four-Year** is mostly all the public four-year colleges.
3. **Cluster 2/ Public Two-Year** is mostly all the public two-year colleges.
4. **Cluster 3/Private Not-For-Profit** includes all the private not-for-profit four-year colleges.

Based on our graphs we find that the **Public Four-Year** and **Private Not-for-Profit** colleges on average have higher graduation rates. Additionally, **Cluster 0** is odd because it contains multiple types of colleges with most of them being **For-Profit**, the one characteristic that I found that they had in common was that they have lower graduation rates and student counts on average.

Overall, in relation to my initial question, we find that an important characteristic for schools is the type of control they're in, and if they are a two-year or four-year institution. Based on these two characteristics we find that graduation rates and the amount of student who attend the schools' changes, with public schools having more students on average and private not-for-profit schools having higher graduation rates on average.

## Impact

I believe this project has the potential to have a positive impact on colleges and students alike. With our discoveries, we learn that the control a college is in is important to graduation rates and that we should take the time to figure out why private not-for-profit schools do well while two-year public institutions don't. However, despite these discoveries we must consider that the data is limited by the features that it has and the information the colleges have provided. We learn what separates these colleges and how they have different graduation rates, student counts, and award values because of their control, but we don't really learn about other information like the students' opinions. This information is a good starting point, but it's curial to account for any bias that may be present. Regardless, the information is valuable and can serve as a catalyst to discover more about what creates the differences between universities.

# References

- https://www.geeksforgeeks.org/python-plotly-tutorial/ -> Used this for help with the interactive scatter plot.
- https://stackoverflow.com/questions/39120942/difference-between-standardscaler-and-normalizer-in-sklearn-preprocessing#:~:text=The%20main%20difference%20is%20that,your%20data%20before%20normalizing%20it. -> For help with learning the difference between normalization and standardization