**KSI: *Machine Learning & AI Dual certification***

Project build under the guidance of Professor **Nitesh Karmakar**

Project 2: **Public and Private University**

*Machine Learning: KMeans Clustering*

**Name**: Mohammed Faraz Hussain

**College**: Terna Engineering College, Nerul

**Roll number**: TE-B-24

**ID number**: TU3F1920077

**Email**: [farazhussain503@gmail.com](mailto:farazhussain503@gmail.com)

**Terna Mail**: [mohammedhussain@ternaengg.ac.in](mailto:mohammedhussain@ternaengg.ac.in)

**Phone number**: +91 9082905016 / +91 7303272780

**Table of Contents**

**Sr.No Title**

| 1 | | Introduction | |
| --- | --- | --- | --- |
|  | |  | |
| 2 | | Aim and Objective | |
|  | |  | |
| 3 | | About Dataset | |
|  | |  | |
| 4 | | Implementation | |
|  | |  | |
| 5 | | Conclusion | |
|  | |  | |
| ` 6 | | References | |
|  | |  | |
|  | |  | |
|  | |  | |
|  | |  | |
|  | | | |
|  | |  | |
|  | |  | |
|  | | | |
| **Introduction** | | | |

K-Means Clustering is an unsupervised machine learning algorithm.

In contrast to traditional supervised machine learning algorithms, K-Means attempts to classify data without having first been trained with labeled data.

Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the most relevant group.

**Aim and Objective**

We will attempt to use KMeans Clustering to cluster Universities into two groups, Private and Public.

Unlike supervised learning, clustering is considered an unsupervised learning method since we don’t have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

In K-Means, each cluster is associated with a centroid. The main objective of the K-Means algorithm is **to minimize the sum of distances between the points and their respective cluster centroid.**

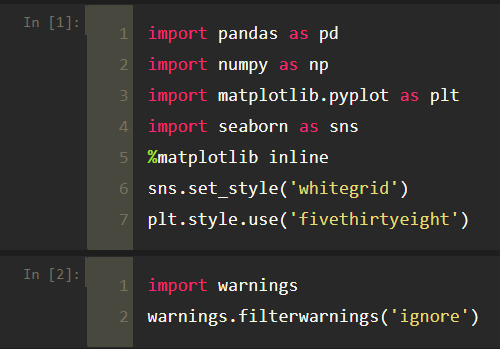
**About Dataset**

The dataset is a CSV(comma separated value) having 777 rows × 18 columns.

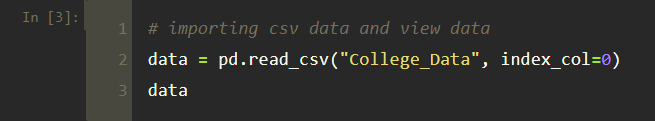
* **Private:** A factor with levels No and Yes indicating private or public university
* **Apps:** Number of applications received
* **Accept:** Number of applications accepted
* **Enroll:** Number of new students enrolled
* **Top10perc:** Percentage new students from top 10% of high school class
* **Top25perc:** Percentage of new students from top 25% of their high school class
* **F.Undergrad:** Number of full-time undergraduates
* **P.Undergrad:** Number of part-time undergraduates
* **Outstate:** Out-of-state tuition
* **Room.Board:** Room and board costs
* **Books:** Estimated book costs
* **Personal:** Estimated personal spending
* **PhD:** Percentage of faculty with PhDs
* **Terminal:** Percentage of faculty with a terminal degree (PhD/JD/MD/MBA/etc)
* **S.F.Ratio:** Student/faculty ratio
* **perc.alumni:** Percentage alumni who donate
* **Expend:** Instructional expenditure per student
* **Grad.Rate:** Graduation rate

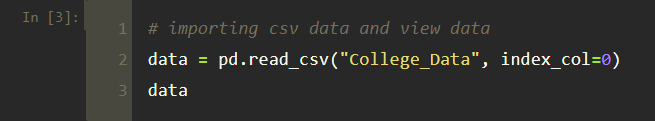
**Implementation**

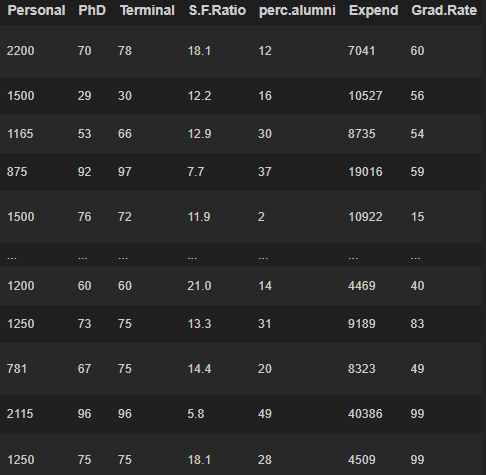
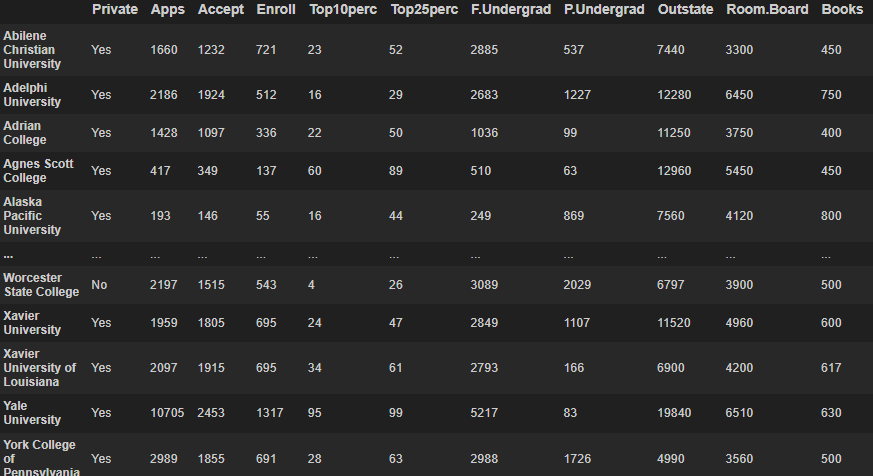
1. Importing Libraries
2. Importing Dataset and Read the data (from csv)
3. Identify the dependent and independent variables.
4. Check if the data has missing values or the data is categorical or not.
5. Visualize the data.
6. Applying K means Clustering
7. Model Evaluation
8. **Importing Libraries**

****

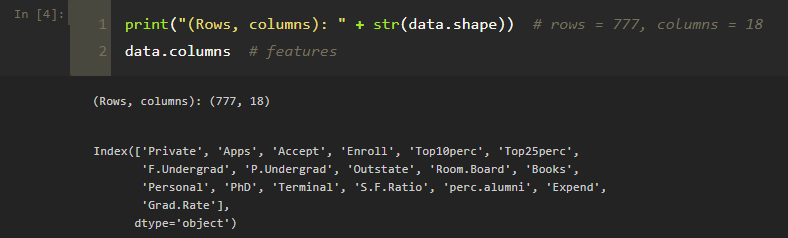
1. **Importing Dataset**

****

****

****

**2.1 Checking rows and columns**

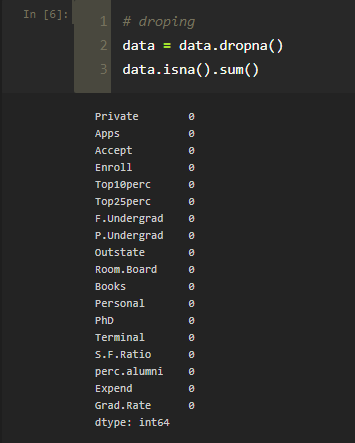
****

**3. Identify the dependent and independent variables.**

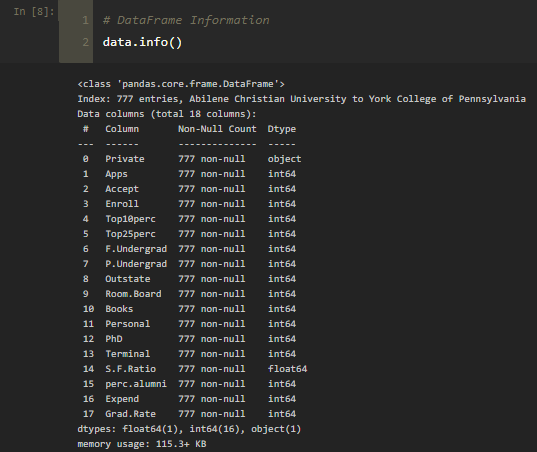
* We don’t have any dependent variable, such problems fall into the category of unsupervised learning
* Since we don’t have that frame of reference in unsupervised learning, thus the name
* No frame of reference means there is no dependent variable

**4. Check if the data has missing values or the data is categorical or not.**

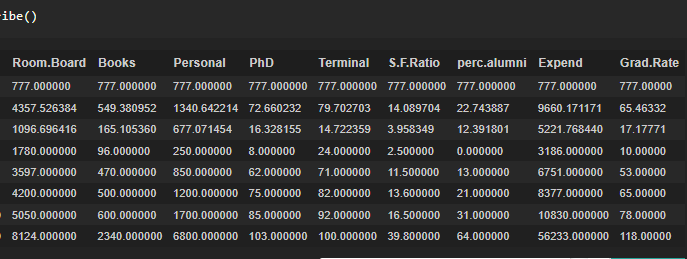
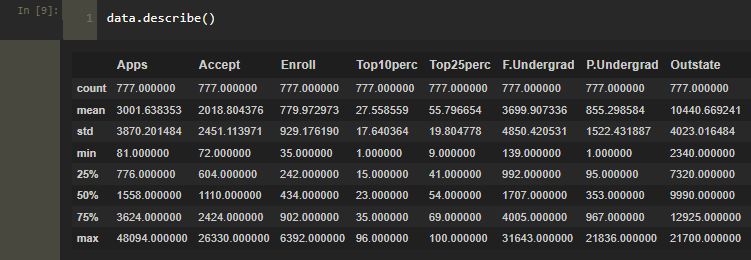
* Dataset has no categorical values
* Check for missing/NaN/Null values and drop those columns



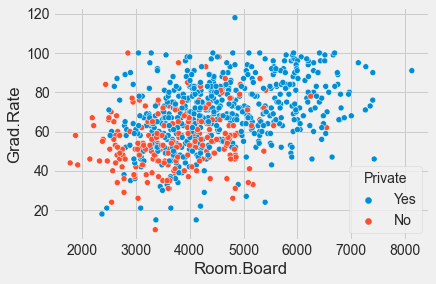
**4.1: Concise summary of data**

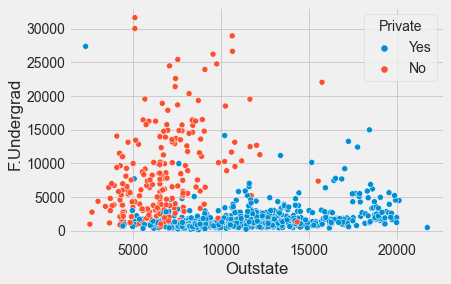
****

**4.2: Statistical measures about the data**

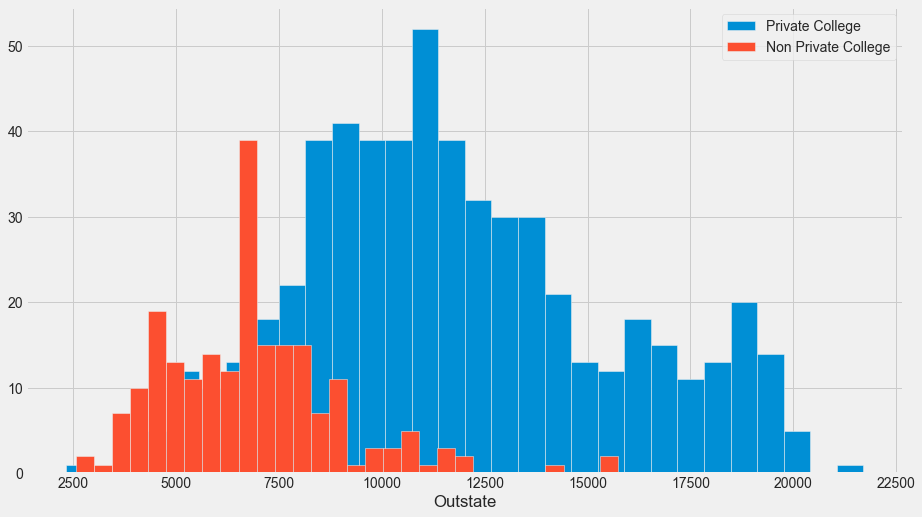
****

**5. Visualizing the Data**

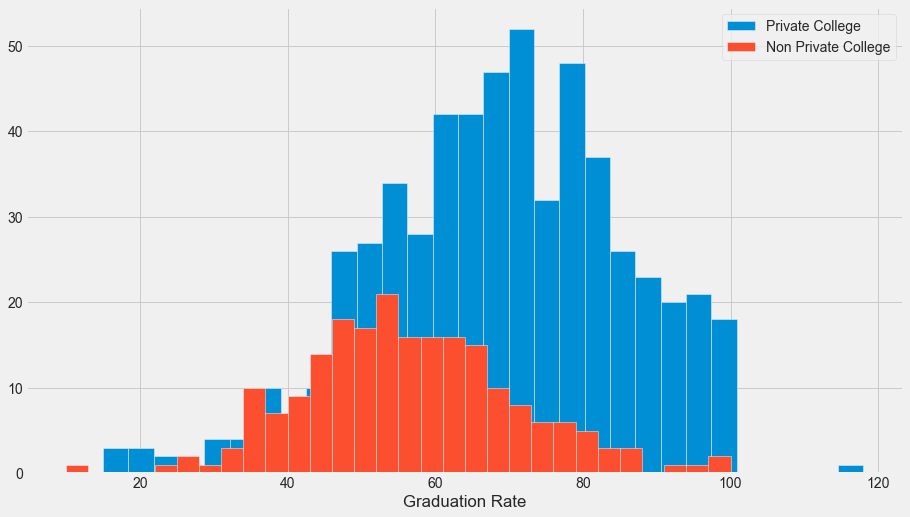
****

****

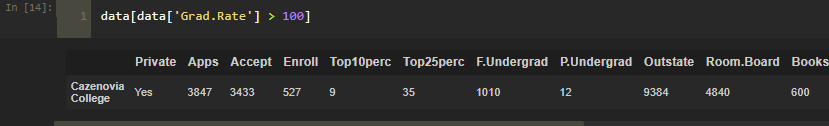
**Histogram of Outstate Tuition based on the Private:**

****

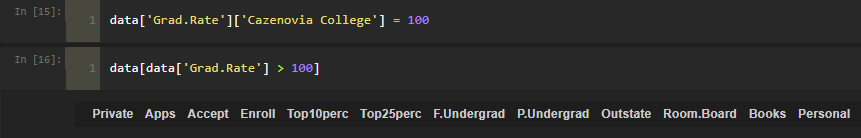
**Histogram of Grad.Rate based on the Private:**

****

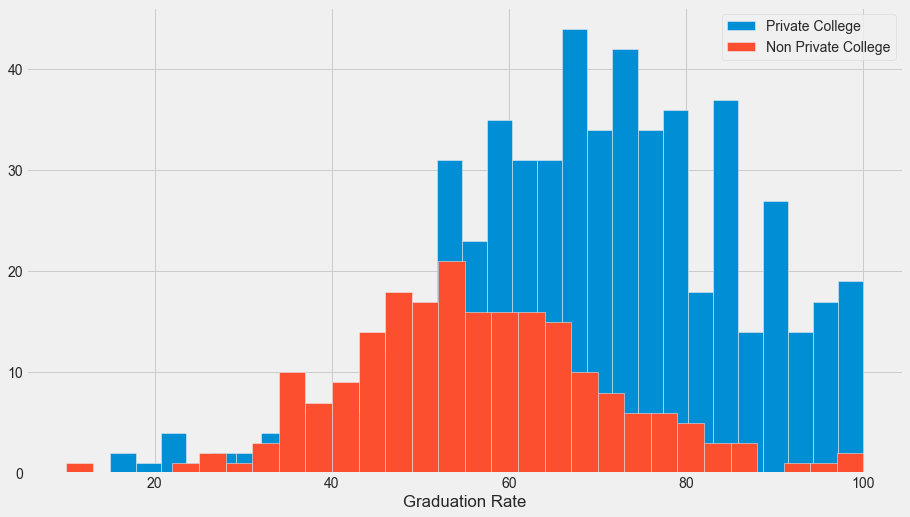
* School with a graduation rate of higher than 100%
* Notice how there seems to be a private school with a graduation rate of higher than 100%



We need to set that school’s graduation rate to 100% so it makes sense.

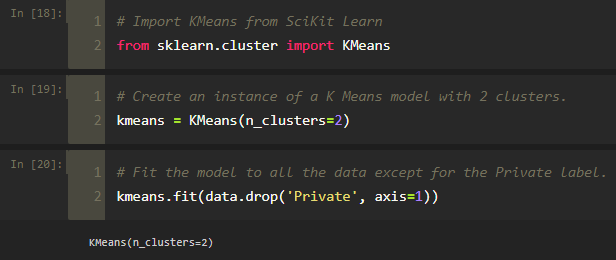


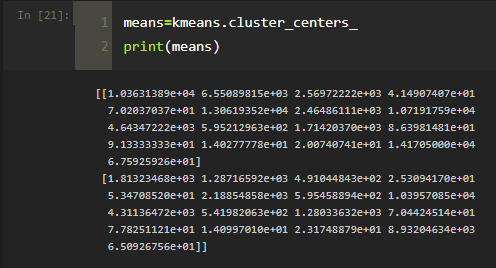
Now there is no school with graduation rate higher than 100%



* We can see there are no data points that fall outside 100.
* No school with graduation rate higher than 100%

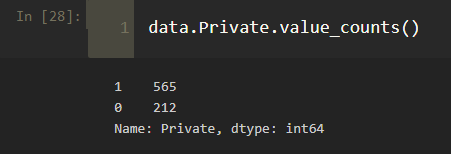
**6. Applying K means Clustering**

****

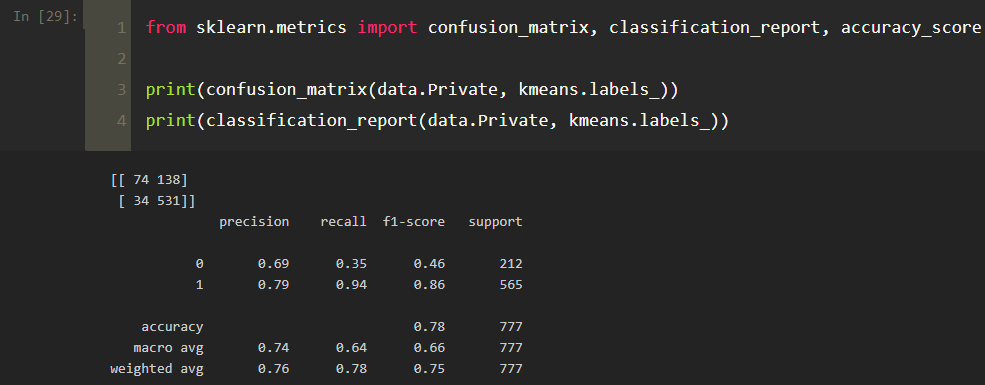
****

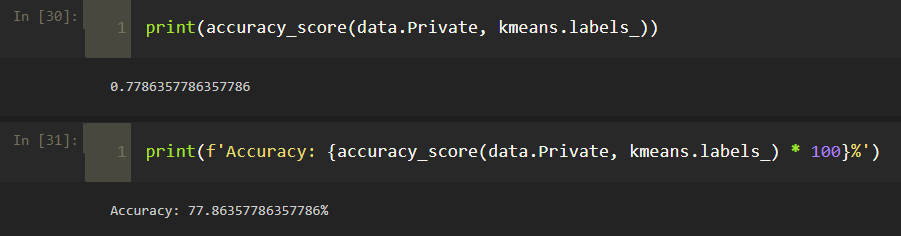
**7. Model Evaluation**

* There is no perfect way to evaluate clustering if we don’t have the labels, however, we do have the labels, so we take advantage of this to evaluate our clusters.
* Create a new column for df called ‘Cluster’, which is a 1 for a Private school, and a 0 for a public school.



**7.1: Creating a confusion matrix and classification report to see how well the K means clustering worked without being given any labels.**

****

****

**Accuracy: 77.86%**

**Conclusion**

Thus, we got the accuracy to be 77.86% which is not so bad considering the algorithm is purely using the features to cluster the universities into 2 distinct groups.

**References**

* [**https://towardsdatascience.com/k-means-clustering-of-university-data-9e8491068778**](https://towardsdatascience.com/k-means-clustering-of-university-data-9e8491068778)
* [**https://medium.com/analytics-vidhya/k-means-clustering-43d0136bf005**](https://medium.com/analytics-vidhya/k-means-clustering-43d0136bf005)
* [**https://www.kaggle.com/faressayah/k-means-clustering-private-vs-public-universities**](https://www.kaggle.com/faressayah/k-means-clustering-private-vs-public-universities)