Machine Learning Project- 02

COLLEGE CLUSTERING PROJECT

[Group Project]

Details of Group Members:

| Name: | PRANJAL YADAV | GANESH WAGHMARE |
| --- | --- | --- |
| College: | TERNA ENGINEERING COLLEGE | TERNA ENGINEERING COLLEGE |
| Email: | pranjalyadav@ternaengg.ac.in | ganeshwaghmare@ternaengg.ac.in |
| Phone No.: | 9833553644 | 8459164229 |

**Index**

| **Sr.No** | **Contents** | **Page** |
| --- | --- | --- |
| 1 | Overview | 2 |
| 2 | Proposed Methodology | 3 |
| 3 | Visualization | 4 |
| 4 | Implementation | 6 |
| 5 | Conclusion | 9 |

**Overview**

**Clustering** is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples. We’ll cover here clustering based on features. Clustering is used in market segmentation; where we try to find customers that are similar to each other whether in terms of behaviors or attributes, image segmentation/compression; where we try to group similar regions together, document clustering based on topics, etc.

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.

K-Means Clustering is an [Unsupervised Learning algorithm](https://www.javatpoint.com/unsupervised-machine-learning), which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

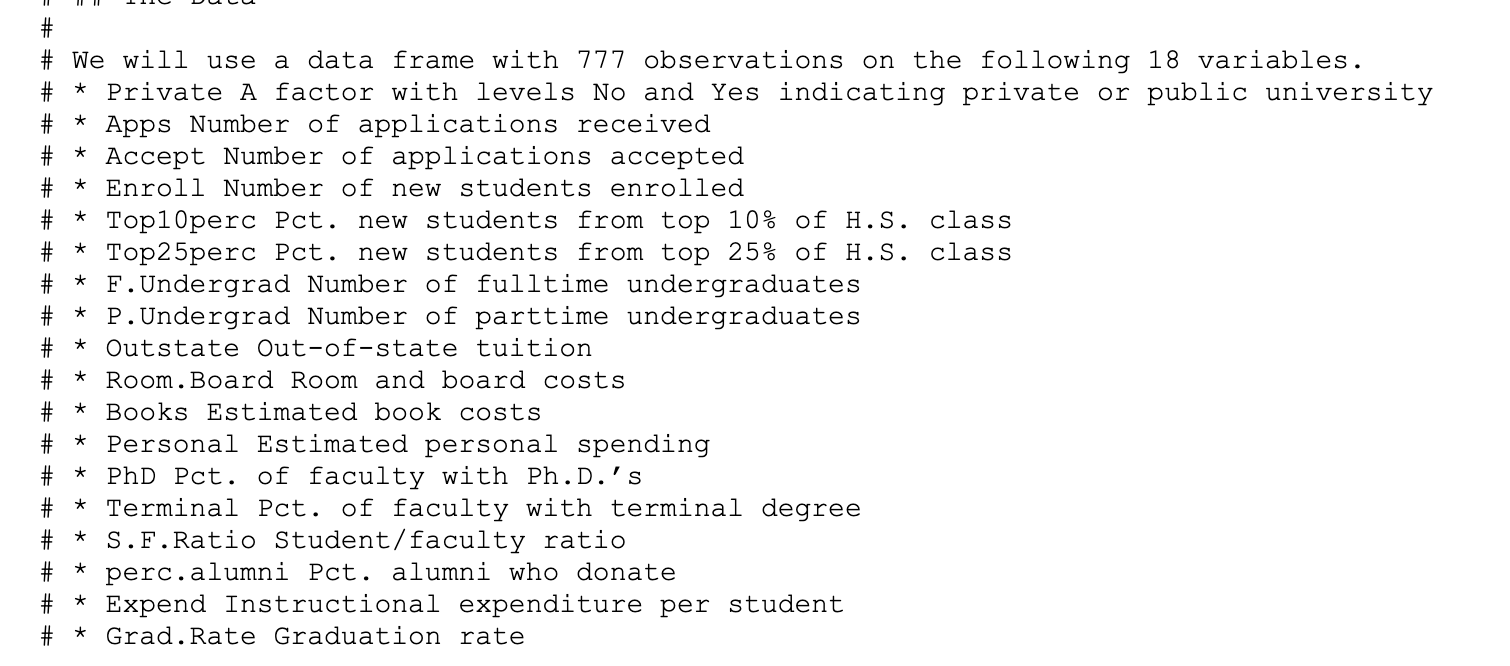
It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

**Proposed Methodology**

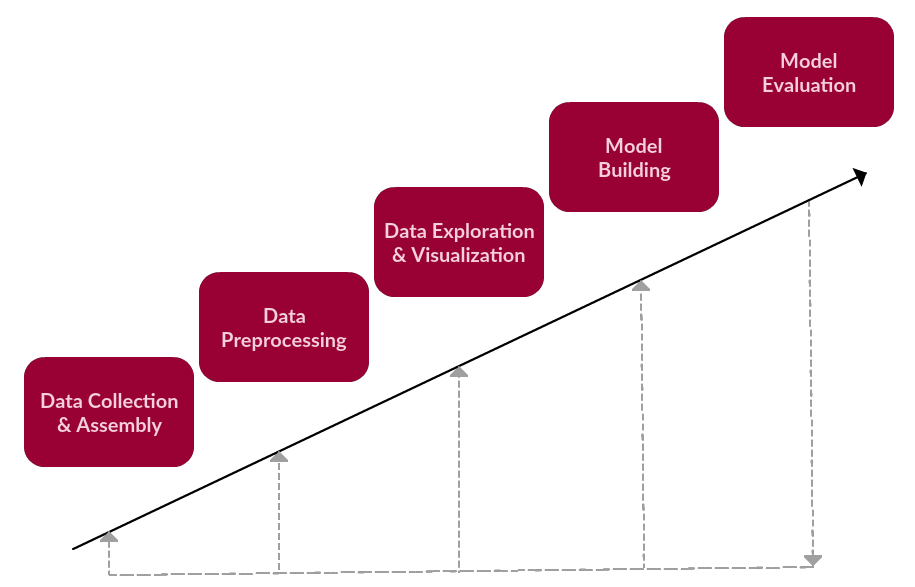
1. **Dataset:**

This is given to us and we already have the csv file that we will be using for dataset model.



1. **Data Pre-Processing:**

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.



**3. Algorithm used:**

K-means clustering is a very famous and powerful unsupervised machine learning algorithm. It is used to solve many complex unsupervised machine learning problems. Before we start, let's take a look at the points which we are going to understand.

**4. Model Building:**

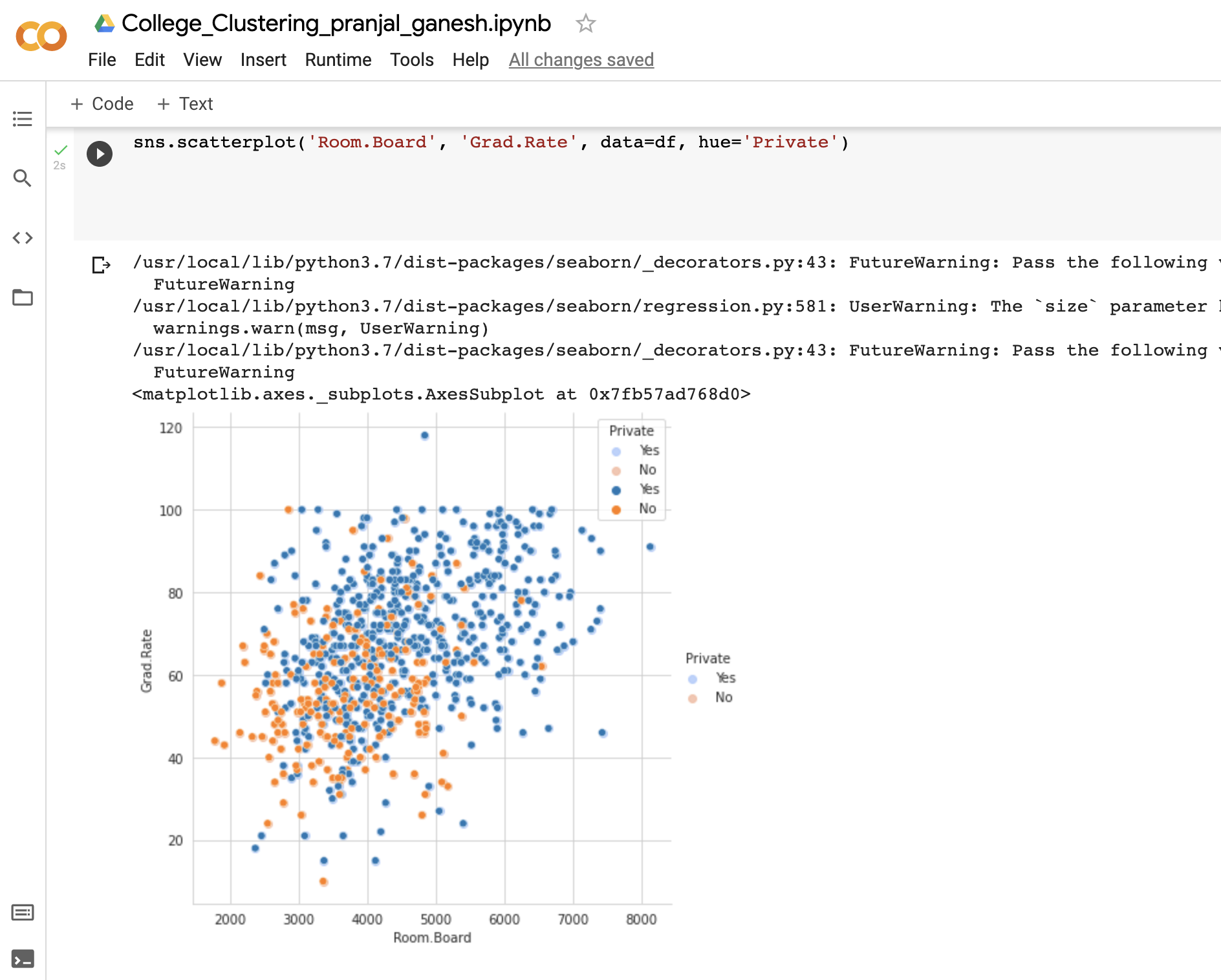
The data are then divided into 2 clusters since there are two types of colleges (i.e govt and private) that we have to separate.

**5. Model Evaluation**:.

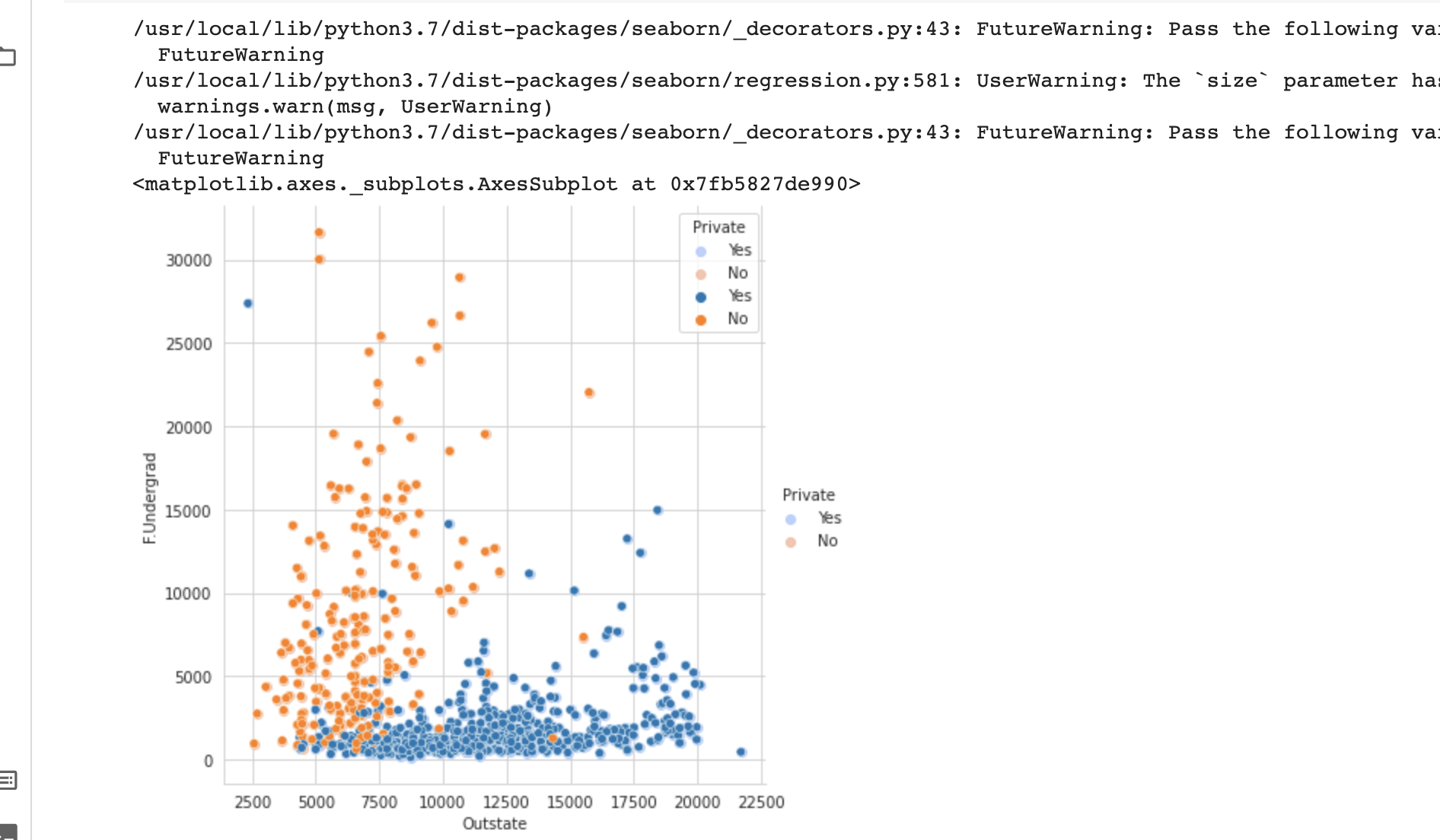
Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data.So our accuracy was about 80%.

**Visualization**

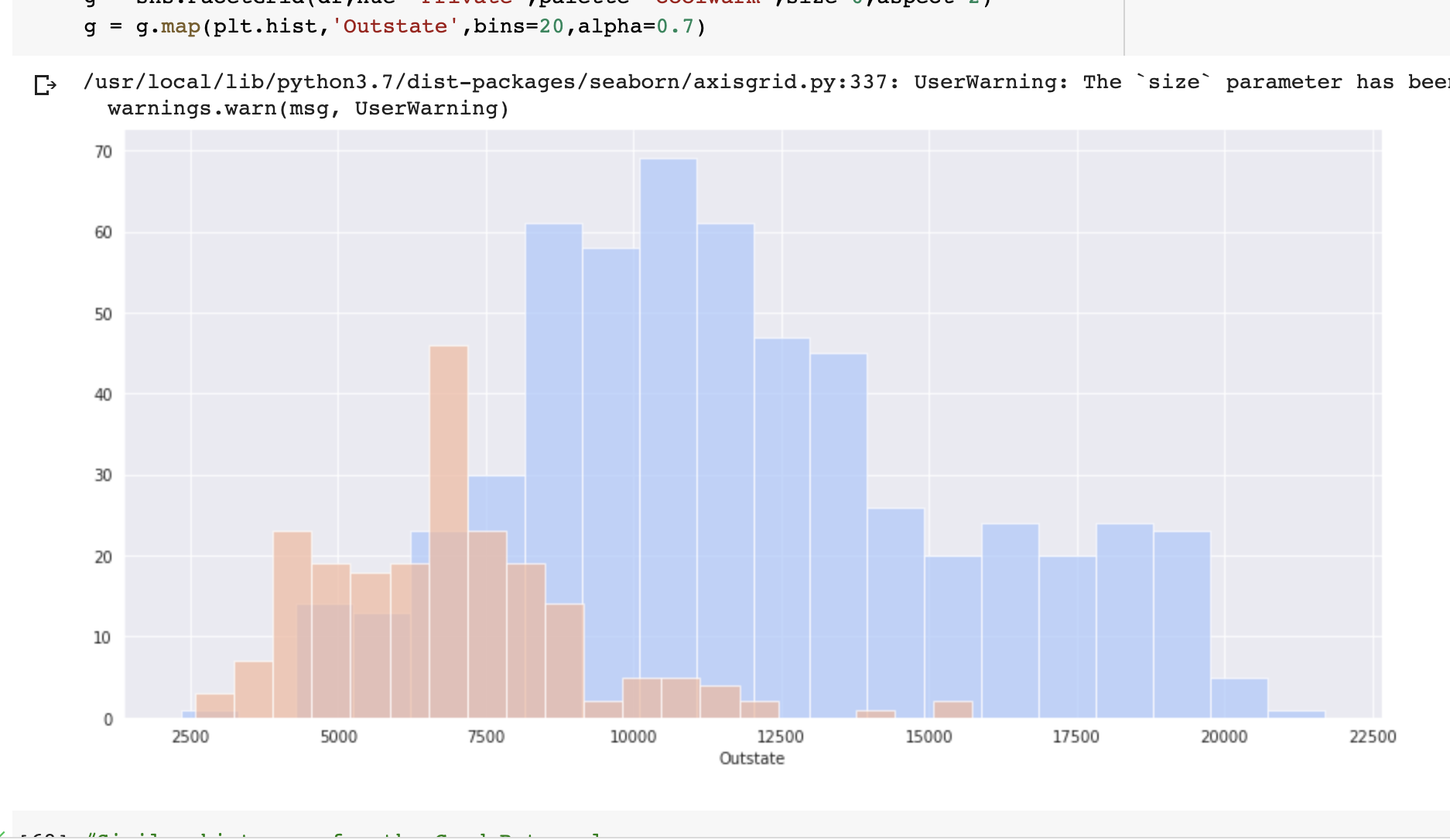
1. Scatterplot between Grad.rate vs Room.Board



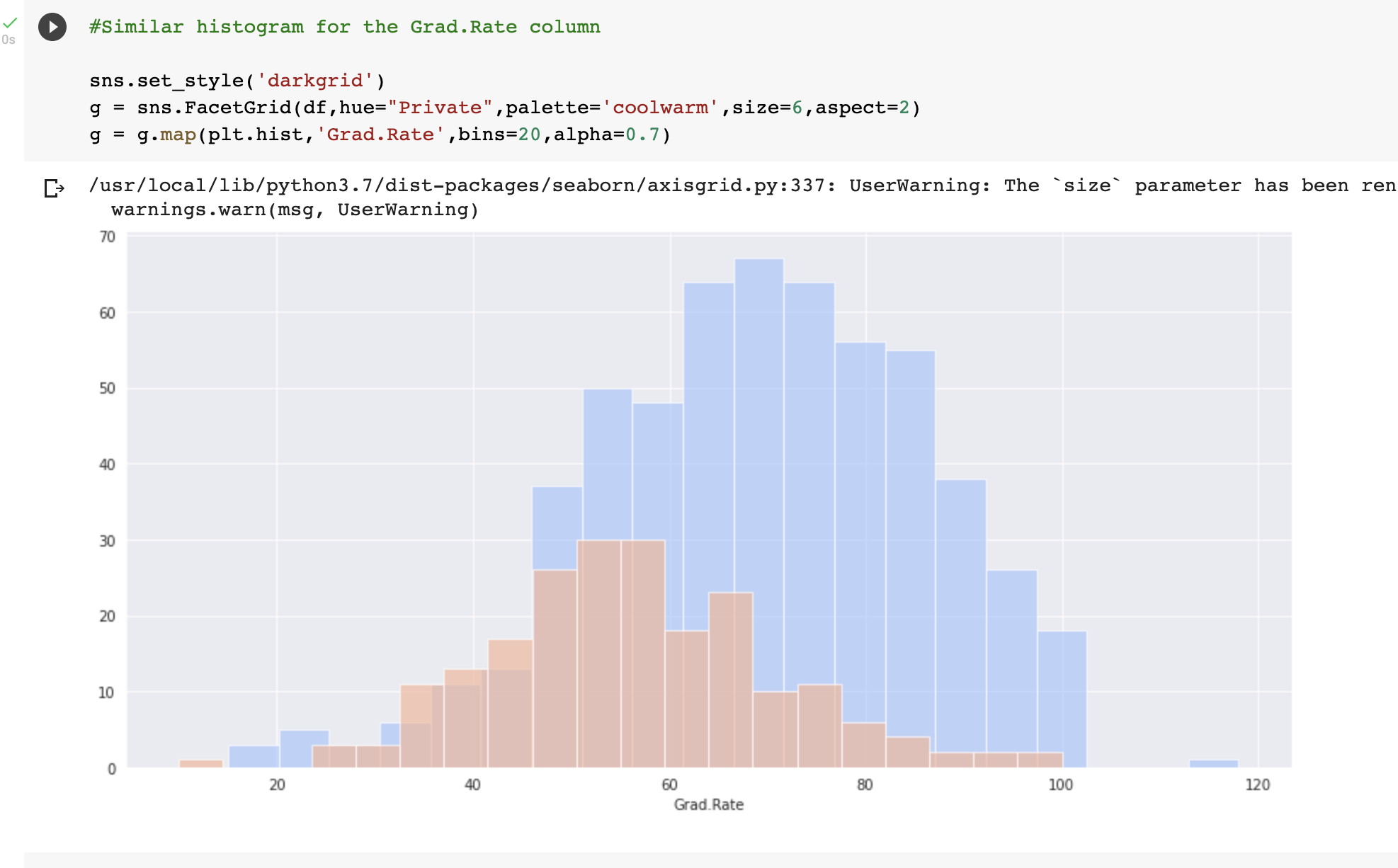
2.Scatterplot between F.Undergraduate and Outstate



3.Tution fee variation

****

4. Graduation Rate



**Implementation**

|  | Importing Libraries |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**Conclusion**

We have successfully separated the college based on the private or public using K-Means clustering.

The data was given to us and we’ve built a model which will divide the college based on the type of institute[‘government’,’private’]