K Means

April 24, 2019

This project illustrates the use of KMeans Clustering to cluster Universities data into to two groups, Private and Public.

I have used a data frame with 777 observations on the following 18 variables.
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 Pct. new students from top 10% of H.S. class
Top25perc Pct. new students from top 25% of H.S. class
F.Undergrad Number of fulltime undergraduates
P.Undergrad Number of parttime undergraduates
Outstate Out-of-state tuition
Room.Board Room and board costs
Books Estimated book costs
Personal Estimated personal spending
PhD Pct. of faculty with Ph.D.'s
Terminal Pct. of faculty with terminal degree
S.F.Ratio Student/faculty ratio
perc.alumni Pct. alumni who donate
Expend Instructional expenditure per student
Grad.Rate Graduation rate

```
In [ ]: import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import numpy as np
```

```
In [38]: df = pd.read_csv('College_Data',)
```

```
In [39]: df.head()
```

```
Out[39]:                               Private  Apps  Accept  Enroll  Top10perc  \
         Abilene Christian University     Yes  1660    1232     721         23
         Adelphi University               Yes  2186    1924     512         16
         Adrian College                   Yes  1428    1097     336         22
         Agnes Scott College              Yes   417     349     137         60
         Alaska Pacific University        Yes   193     146      55         16
```

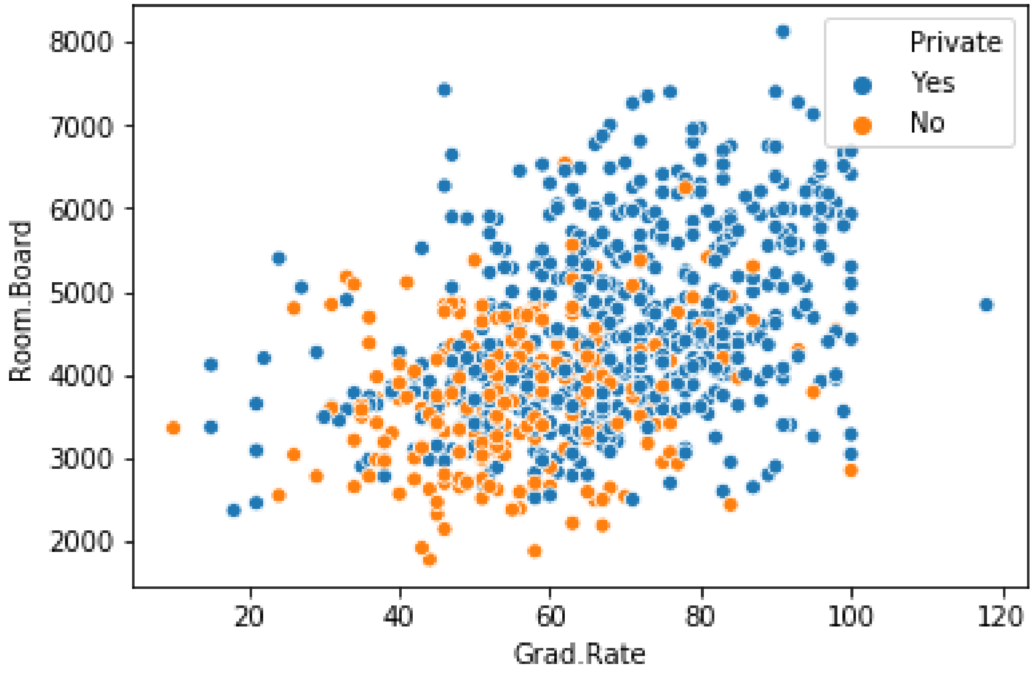
```
                              Top25perc  F.Undergrad  P.Undergrad  Outstate  \
Abilene Christian University         52         2885          537      7440
Adelphi University                   29         2683         1227     12280
Adrian College                       50         1036           99     11250
Agnes Scott College                  89          510           63     12960
Alaska Pacific University            44          249          869      7560

                              Room.Board  Books  Personal  PhD  Terminal  \
Abilene Christian University        3300    450      2200   70        78
Adelphi University                  6450    750      1500   29        30
Adrian College                      3750    400      1165   53        66
Agnes Scott College                 5450    450       875   92        97
Alaska Pacific University           4120    800      1500   76        72

                              S.F.Ratio  perc.alumni  Expend  Grad.Rate
Abilene Christian University       18.1           12    7041         60
Adelphi University                 12.2           16   10527         56
Adrian College                     12.9           30    8735         54
Agnes Scott College                 7.7           37   19016         59
Alaska Pacific University          11.9            2   10922         15
```

```
In [40]: sns.scatterplot(df['Grad.Rate'],df['Room.Board'],hue=df['Private'])
```

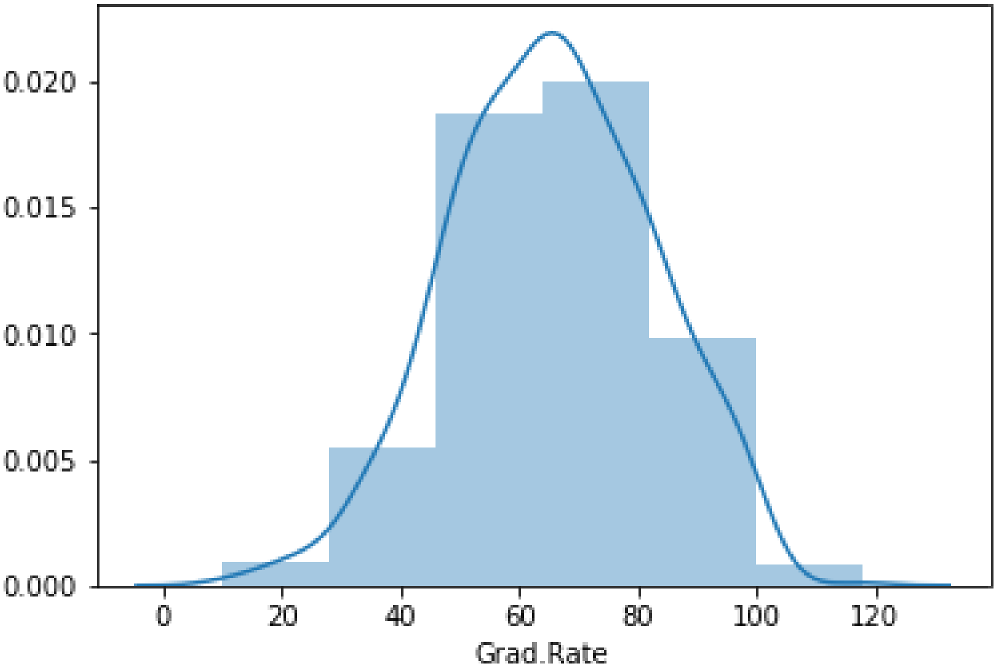
```
Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x15e13ad8748>
```



```
In [41]: sns.distplot(df['Grad.Rate'],bins = 6)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a no
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

```
Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x15e13ad7b38>
```



From the above plots, we can see that there is record with grad rate > 100, which is not possible in general, so we will be setting the value to 100 using pandas functions.

```
In [42]: df.describe()
```

```
Out[42]:                Apps        Accept       Enroll    Top10perc    Top25perc  \
         count    777.000000    777.000000   777.000000   777.000000   777.000000
         mean    3001.638353   2018.804376   779.972973    27.558559    55.796654
         std     3870.201484   2451.113971   929.176190    17.640364    19.804778
         min       81.000000     72.000000    35.000000     1.000000     9.000000
         25%      776.000000    604.000000   242.000000    15.000000    41.000000
         50%     1558.000000   1110.000000   434.000000    23.000000    54.000000
         75%     3624.000000   2424.000000   902.000000    35.000000    69.000000
         max    48094.000000  26330.000000  6392.000000    96.000000   100.000000

                 F.Undergrad   P.Undergrad      Outstate   Room.Board        Books  \
         count    777.000000    777.000000    777.000000   777.000000   777.000000
         mean    3699.907336    855.298584  10440.669241  4357.526384   549.380952
```

```
std     4850.420531   1522.431887   4023.016484  1096.696416   165.105360
min      139.000000      1.000000   2340.000000  1780.000000    96.000000
25%      992.000000     95.000000   7320.000000  3597.000000   470.000000
50%     1707.000000    353.000000   9990.000000  4200.000000   500.000000
75%     4005.000000    967.000000  12925.000000  5050.000000   600.000000
max    31643.000000  21836.000000  21700.000000  8124.000000  2340.000000

          Personal         PhD    Terminal   S.F.Ratio  perc.alumni  \
count   777.000000  777.000000  777.000000  777.000000   777.000000
mean   1340.642214   72.660232   79.702703   14.089704    22.743887
std     677.071454   16.328155   14.722359    3.958349    12.391801
min     250.000000    8.000000   24.000000    2.500000     0.000000
25%     850.000000   62.000000   71.000000   11.500000    13.000000
50%    1200.000000   75.000000   82.000000   13.600000    21.000000
75%    1700.000000   85.000000   92.000000   16.500000    31.000000
max    6800.000000  103.000000  100.000000   39.800000    64.000000

             Expend  Grad.Rate
count    777.000000  777.00000
mean    9660.171171   65.46332
std     5221.768440   17.17771
min     3186.000000   10.00000
25%     6751.000000   53.00000
50%     8377.000000   65.00000
75%    10830.000000   78.00000
max    56233.000000  118.00000
```

```
In [43]: df[df['Grad.Rate']>100]['Grad.Rate']
```

```
Out[43]: Cazenovia College    118
         Name: Grad.Rate, dtype: int64
```

```
In [44]: df['Grad.Rate'].replace(118,100,inplace=True)
```

Importing K means from scikit learn and fitting the dataframe to divide the dataset into two clusters using kmeans.

```
In [45]: from sklearn.cluster import KMeans
```

```
In [46]: kmeans = KMeans(n_clusters = 2)
```

```
In [48]: kmeans.fit(df.drop('Private',axis = 1))
```

```
Out[48]: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
             n_clusters=2, n_init=10, n_jobs=None, precompute_distances='auto',
             random_state=None, tol=0.0001, verbose=0)
```

```
In [49]: kmeans.cluster_centers_
```

```
Out[49]: array([[1.81323468e+03, 1.28716592e+03, 4.91044843e+02, 2.53094170e+01,
                 5.34708520e+01, 2.18854858e+03, 5.95458894e+02, 1.03957085e+04,
                 4.31136472e+03, 5.41982063e+02, 1.28033632e+03, 7.04424514e+01,
                 7.78251121e+01, 1.40997010e+01, 2.31748879e+01, 8.93204634e+03,
                 6.50926756e+01],
                [1.03631389e+04, 6.55089815e+03, 2.56972222e+03, 4.14907407e+01,
                 7.02037037e+01, 1.30619352e+04, 2.46486111e+03, 1.07191759e+04,
                 4.64347222e+03, 5.95212963e+02, 1.71420370e+03, 8.63981481e+01,
                 9.13333333e+01, 1.40277778e+01, 2.00740741e+01, 1.41705000e+04,
                 6.75925926e+01]])
```

Have further evaluated the model to check for accuracy as we have private variable already present in the dataset. This is not ususally done in Kmeans, but doing it just because we have data available.

```
In [50]: def converter(cluster):
             if cluster=='Yes':
                 return 1
             else:
                 return 0
```

```
In [51]: df['Cluster'] = df['Private'].apply(converter)
```

```
In [52]: df.head()
```

```
Out[52]:                               Private  Apps  Accept  Enroll  Top10perc  \
         Abilene Christian University     Yes  1660    1232     721         23
         Adelphi University               Yes  2186    1924     512         16
         Adrian College                   Yes  1428    1097     336         22
         Agnes Scott College              Yes   417     349     137         60
         Alaska Pacific University        Yes   193     146      55         16

                                        Top25perc  F.Undergrad  P.Undergrad  Outstate  \
         Abilene Christian University          52         2885          537      7440
         Adelphi University                    29         2683         1227     12280
         Adrian College                        50         1036           99     11250
         Agnes Scott College                   89          510           63     12960
         Alaska Pacific University             44          249          869      7560

                                        Room.Board  Books  Personal  PhD  Terminal  \
         Abilene Christian University         3300    450      2200   70        78
         Adelphi University                   6450    750      1500   29        30
         Adrian College                       3750    400      1165   53        66
         Agnes Scott College                  5450    450       875   92        97
         Alaska Pacific University            4120    800      1500   76        72

                                        S.F.Ratio  perc.alumni  Expend  Grad.Rate  \
         Abilene Christian University        18.1           12    7041         60
         Adelphi University                  12.2           16   10527         56
         Adrian College                     12.9           30    8735          54
         Agnes Scott College                 7.7           37   19016          59
         Alaska Pacific University          11.9            2   10922          15

                                       Cluster
         Abilene Christian University        1
         Adelphi University                  1
         Adrian College                      1
         Agnes Scott College                 1
         Alaska Pacific University           1
```

```
In [53]: from sklearn.metrics import confusion_matrix,classification_report
         print(confusion_matrix(df['Cluster'],kmeans.labels_))
         print(classification_report(df['Cluster'],kmeans.labels_))
```

```
[[138  74]
 [531  34]]
              precision    recall  f1-score   support

           0       0.21      0.65      0.31       212
           1       0.31      0.06      0.10       565

   micro avg       0.22      0.22      0.22       777
   macro avg       0.26      0.36      0.21       777
weighted avg       0.29      0.22      0.16       777
```

```
In [ ]:
```