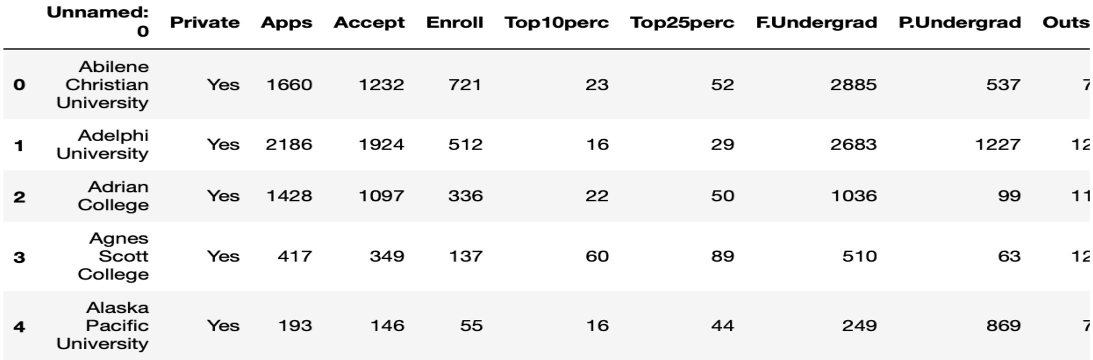
K-means Clustering on Other Data

Dataset link: https://www.kaggle.com/flyingwombat/us-news-and-world-reports-college-data **(777 rows, 18 columns)**

Code (importing csv file):

```
from sklearn.cluster import KMeans
college_df = pd.read_csv('College.csv')
college_df.head()
```

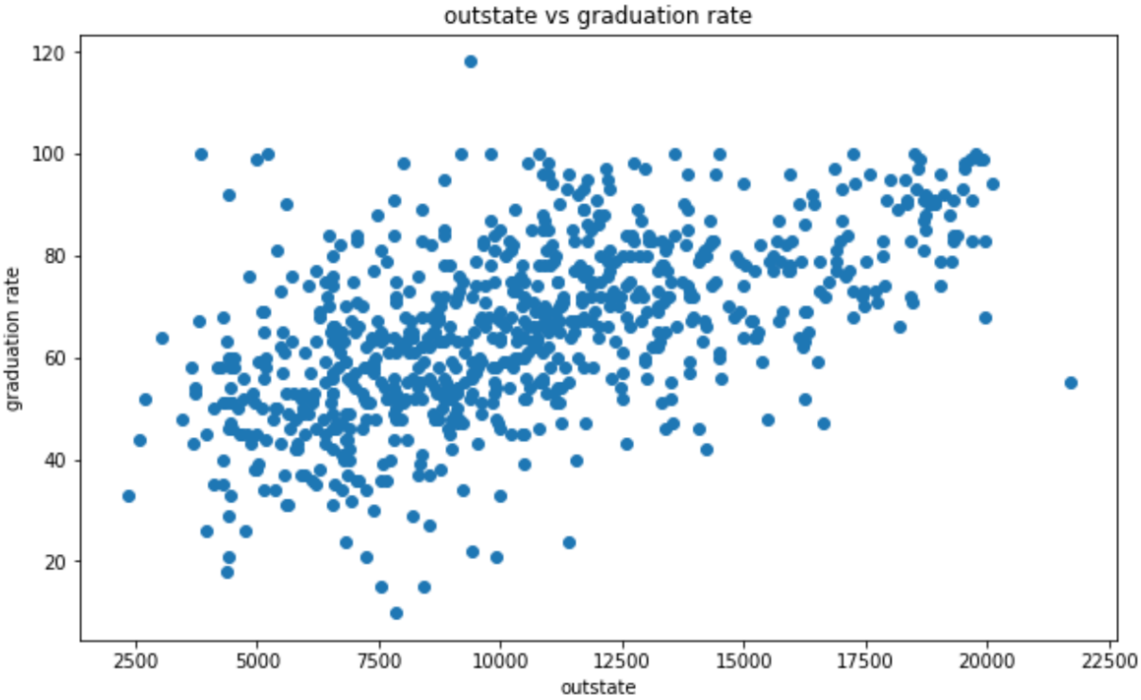


	Unnamed: 0	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outs
0	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7
1	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12
2	Adrian College	Yes	1428	1097	336	22	50	1036	99	11
3	Agnes Scott College	Yes	417	349	137	60	89	510	63	12
4	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7

1. **Choose two numerical features in your dataset and apply k-means clustering on your data into k clusters in Python, where k >= 2**

   Code:

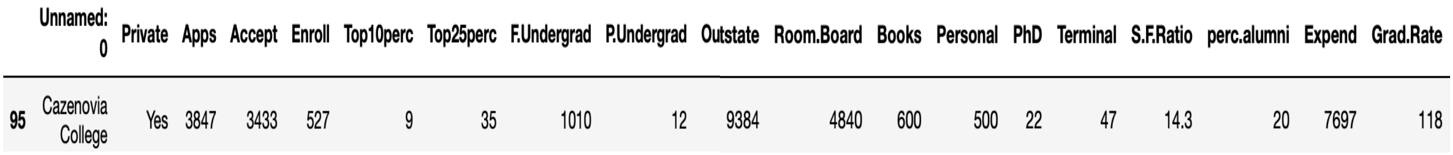
   ```
   # Visualising the data set of two numerical features
   fig= plt.figure(figsize=(10,6))
   plt.scatter(x = college_df['Outstate'],y = college_df['Grad.Rate'])
   plt.title("outstate vs graduation rate")
   plt.xlabel('outstate')
   plt.ylabel('graduation rate')
   plt.show()
   ```



As shown in the graph, there exists a value in the column Grad.Rate where it is over 100, since the graduation rate is not possible to have a value over 100, we can set it to 100 so it makes more sense.

Code:

```
college_df[college_df['Grad.Rate'] > 100]
```

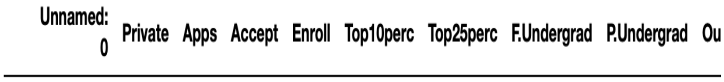


	Unnamed: 0	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
95	Cazenovia College	Yes	3847	3433	527	9	35	1010	12	9384	4840	600	500	22	47	14.3	20	7697	118

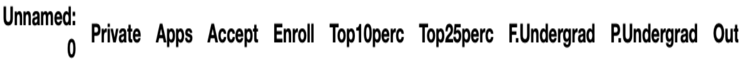
To remove the rows where Grad.Rate is over 100, we can run the code as shown.

```
college_df.loc[college_df['Grad.Rate'] > 100, 'Grad.Rate'] = 100
```

```
college_df[college_df['Grad.Rate'] > 100]
```



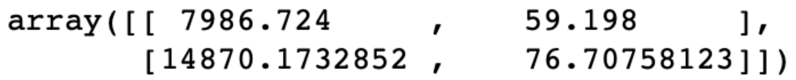
Unnamed: 0	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Ou



Unnamed: 0	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Out

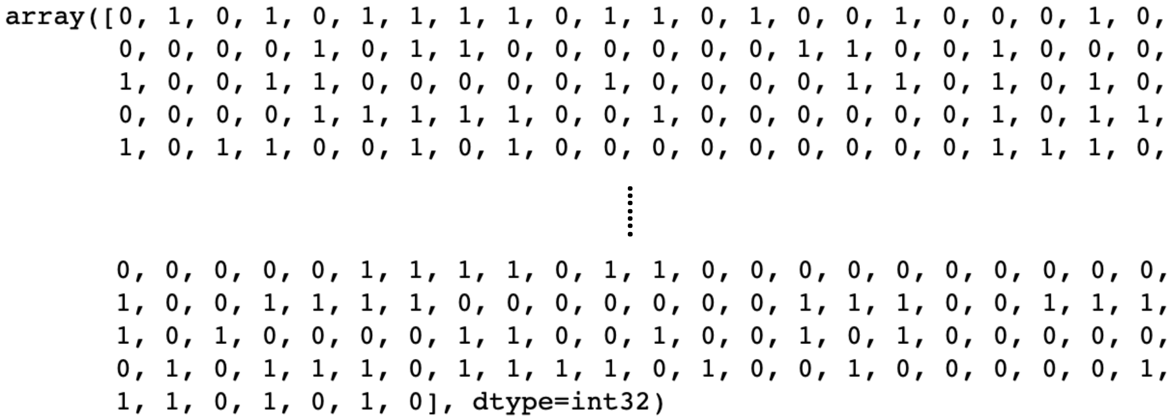
And now since it has been updated to make any Grad.Rate > 100 to equal to 100, we can continue with the rest of the code

```
# Set K=2: we only want to cluster the dataset into two subgroups
kmeans = KMeans(n_clusters=2).fit(college_df[['Outstate','Grad.Rate']])
# Look at the outputs: Two cluster centers
kmeans.cluster_centers_
```



```
array([[ 7986.724     ,    59.198     ],
       [14870.1732852 ,    76.70758123]])
```

```
# Look at the outputs: Cluster labels
kmeans.labels_
```

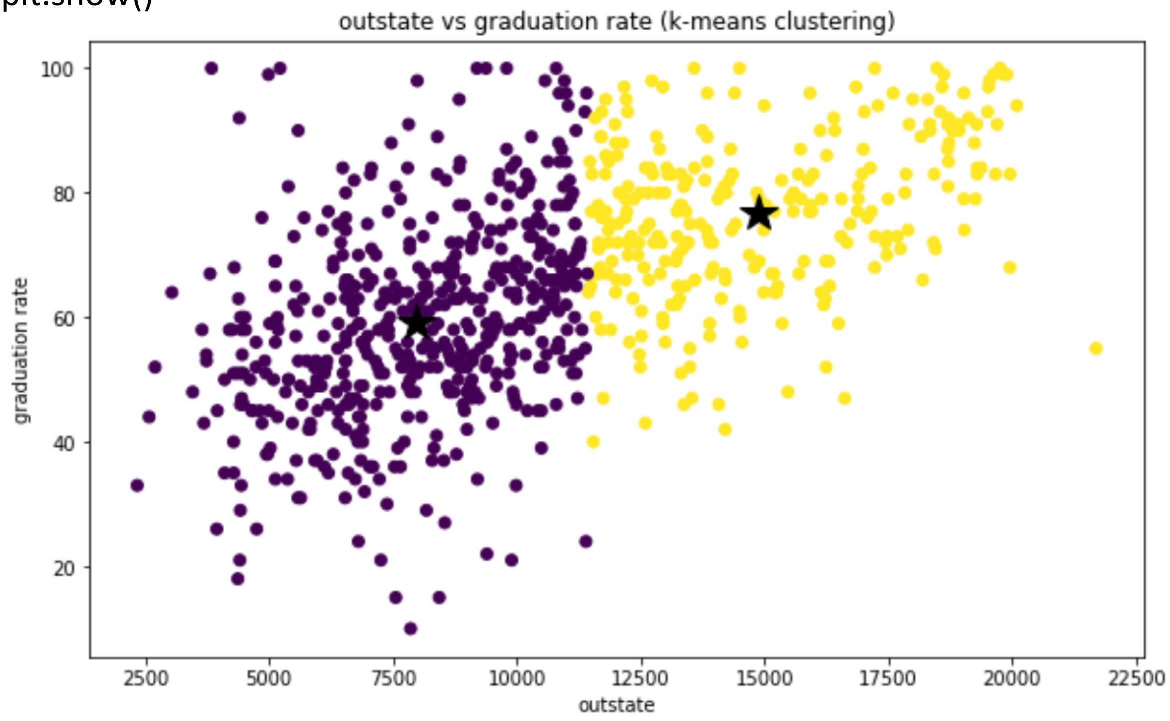


```
array([0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0,
       1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0,
       0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1,
       1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0,
                                           ⋮
       0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1,
       1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,
       0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,
       1, 1, 0, 1, 0, 1, 0], dtype=int32)
```

2. **Visualise the data as well as the results of the k-means clustering. Ideally each cluster is shown in a different colour**

   Code:

```
fig= plt.figure(figsize=(10,6))
# Visualise the output labels
plt.scatter(x=college_df['Outstate'],y=college_df['Grad.Rate'],
c=kmeans.labels_)
# Visualise the cluster centers (black stars)
plt.plot(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],
'k*',markersize=20)
plt.title('outstate vs graduation rate (k-means clustering)')
plt.xlabel('outstate')
plt.ylabel('graduation rate')
plt.show()
```



3. **Describe your findings about the identified clusters.**

   Graduation rate is the probability that they will graduate in % while outstate is the amount of people in the college that went for out-state tuition. By clustering, 2 groups have been identified. It is seen that college with around 2500 – 11300 people going for outstate tuition have randomly scattered values between 5% - 100%. Whereas for colleges that have more people going to outstate tuition, there is more of a clear distribution around 40% - 100%. Therefore, in general, there is a higher graduation rate for the college when more students attend outstate tuition.