**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()

Table

Description automatically generated

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()

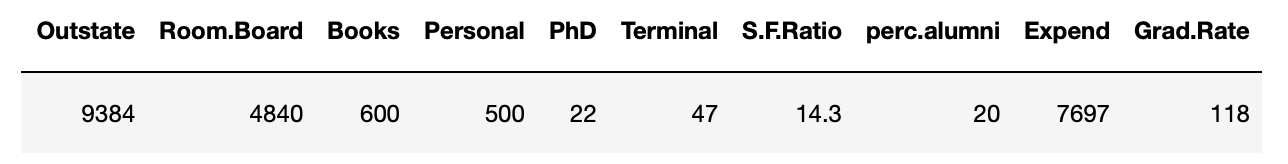
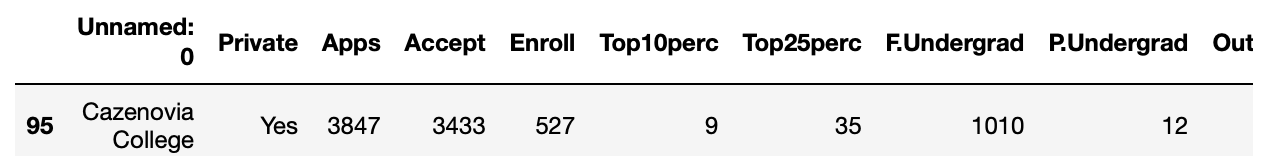
Chart, scatter chart

Description automatically generated

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:**

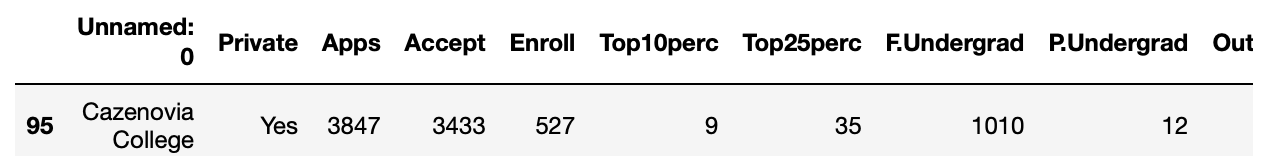
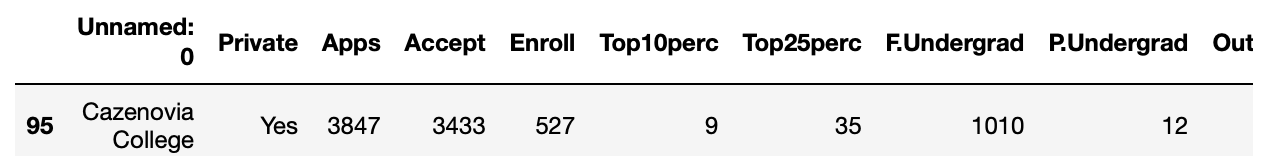
college\_df[college\_df['Grad.Rate'] > 100]



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]



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\_

A picture containing chart

Description automatically generated

# Look at the outputs: Cluster labels

**A picture containing text

Description automatically generated**kmeans.labels\_

**…….**

**A picture containing text

Description automatically generated**

1. **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')

Chart, scatter chart

Description automatically generatedplt.show()

1. **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.