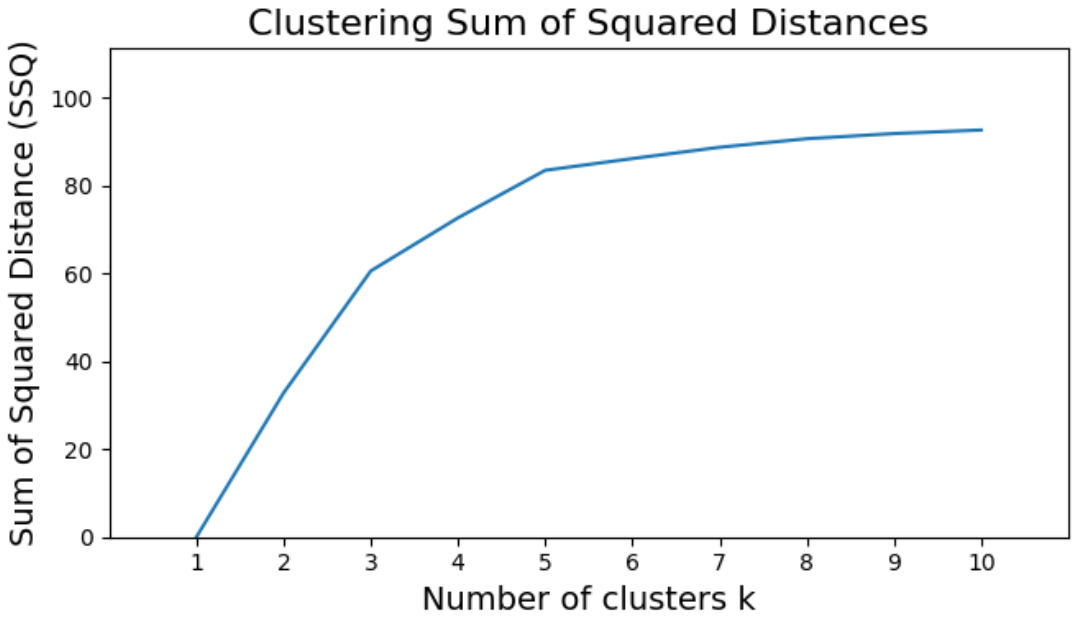
load dataset

2
```
import csv

def load_dataset(filename):
    with open(filename, 'r') as dest_f:
        data_iter = csv.reader(dest_f, delimiter=',', quotechar='"')
        data = [data for data in data_iter]
        data_array = np.asarray(data)
    temp=[]
    for i in range(1, len(data_array)):
        temp.append([data_array[i][3], data_array[i][4]])
    dataset = np.asarray(temp).astype(float)
    return dataset
```

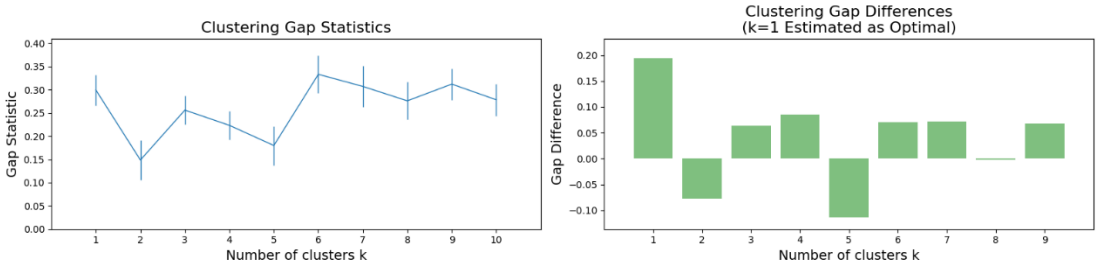
SSQs computed for k values between 1 and 10

3
```
data=load_dataset('shopping-data.csv')
# Generate and plot the SSQ statistics
ssqs = ssq_statistics(data, ks=range(1,11))
plot_ssq_statistics(ssqs)
```



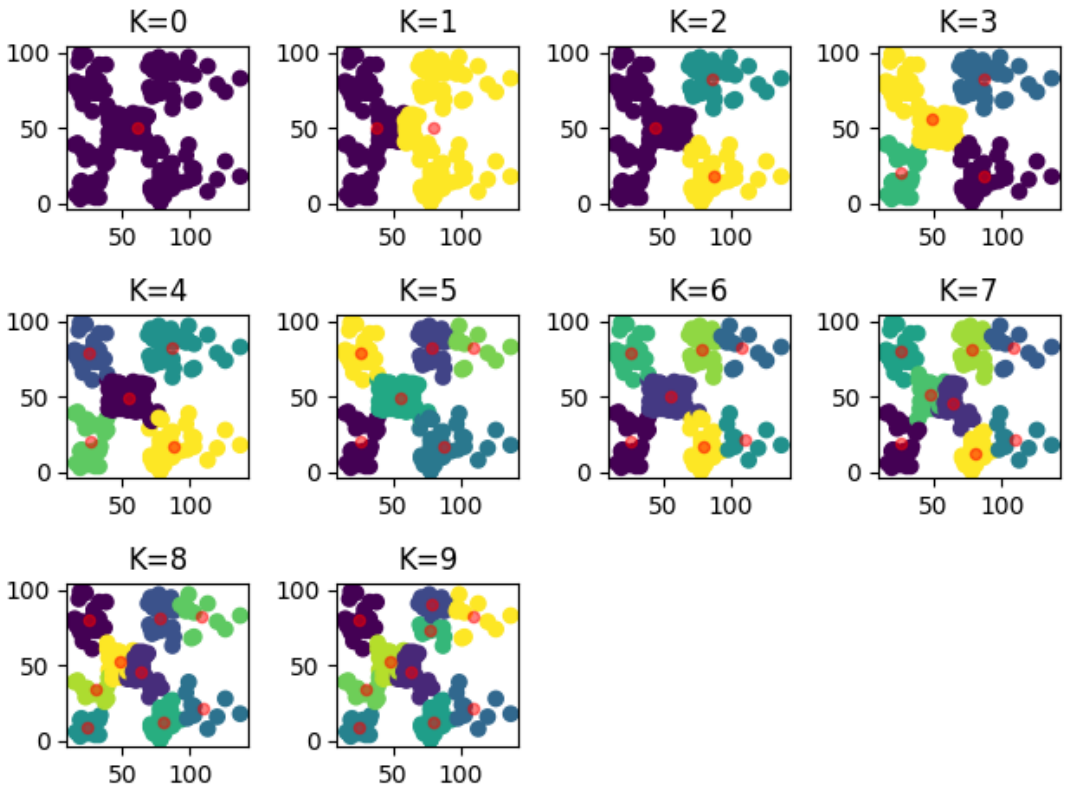
gap statistics computed for k values between 1 and 10

4
```
# Generate and plot the gap statistics
gaps,errs,difs = gap_statistics (data, nrefs=20, ks=range(1,11))
plot_gap_statistics(gaps,errs,difs)
```



plot of the data with centers

5
```
def KMeans_data(data,ks):
    pl.figure()
    for (i, k) in enumerate(ks):  # iterate over the range of k values
        # Fit the model on the data
        km = KMeans(n_clusters=k, random_state=0)
        y_pre = km.fit_predict(data)
        pl.subplot(3,4,i+1)
        pl.title("K="+str(i))
        pl.scatter(data[:, 0], data[:, 1], c=y_pre)
        centers = km.cluster_centers_
        pl.scatter(centers[:, 0], centers[:, 1], c='red', s=20, alpha=0.5);
    pl.show()
KMeans_data(data,ks=range(1,11))
```



1. Where did you estimate the elbow point to be (between what values of k)? What value of k was typically estimated as optimal by the gap statistic? To adequately answer this question, consider generating both measures several (atleast 5) times, as there may be some amount of variation in the value of k that they each estimate as optimal.

elbow point: k=3

gap statistic estimated: k=1

2. Based on the scatter plot of the clustered data, what makes most sense? Give logical interpretation from visually inspecting the clusters.

File a large amount of data into different categories, and then formulate different processing methods for each type of data, because each type of data is significantly different

3. Between SSQ and Gap Statistics, does one measure seem to be a consistently better criterion for choosing the value of k than the other? Why or why not?

Gap Statistics

Its advantage is that we no longer need to judge with the naked eye, only need to find the K corresponding to the largest Gap Statistic. The physical meaning of gap loss function is the difference between the loss of a random sample and the loss of an actual sample. The larger the gap, the better the clustering effect. An extreme case is that as K changes, the gap almost maintains a straight line. It shows that there is no obvious category relationship among these samples, and the data distribution is almost consistent with the uniform distribution, which is approximately random. There is no point in clustering at this time.