Hands-on Project

In [84]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import seaborn as sns
from sklearn import datasets
random_state = 10

from sklearn.decomposition import PCA
# importing clustering algorithms
from sklearn.cluster import KMeans
from sklearn.mixture import GaussianMixture
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import DBSCAN
from sklearn.cluster import SpectralClustering

from sklearn.metrics import silhouette_samples
from scipy.linalg import svd
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
```

In [85]:

```
url  = "https://archive.ics.uci.edu/ml/machine-learning-databases/00422/wifi_localization.t
tae = np.loadtxt("wifi_localization.txt");
df = pd.DataFrame(tae,dtype = int)
names = ["Wifi1","Wifi2","Wifi3","Wifi4","Wifi5","Wifi6","Wifi7","Room"]
df.columns = names
```

In [86]:

```
df_X = df.iloc[:,:7]
df_Y = df["Room"]
```

Question 1: List the dataset(s) you chose for this project from the UCI Machine Learning respository (https://archive.ics.uci.edu/ml/datasets.php).

Dataset : Wireless Indoor Localization Data Set

Question 2: Describe the dataset in your own words. How many data points, how many attributes, how many types of attributes, how many classes (if any)? Who collected it? How was it collected?

In [87]:

```
df_X.shape
```

Out[87]:

```
(2000, 7)
```

The dataset consists of different wifi signals measured from a mobile in an indoor space in different rooms.
Based on the strength of the wifi signals it is to determine the one of the location in the indoor space
there are 7 types attributes
there are 4 classes
it was collected by Rajen Bhatt. This data was collected from a mobile measuring the wifi signal strength at different locations in indoorspace.

Question 3: What is your goal? Specifically, what insights do you want to learn from this data. Please be aware that clustering, classification, or itemset mining are not 'insights'. These are data mining tasks. Insights are relevant to the domain from which the data is generated.

The Goal is to find the location of one of the rooms in the indoor space
Insights:
we can track the location of the person bounded in a region through signals.
we can track the location of electronic gadgets if we misplace it with the strength of the signals.
monitoring the devices from that region based on the wifi signal strength.
counting the number of devices in the location by these signals.

Question 4: List the data mining task(s) and the specific algorithms you want to perform on this data. Do not pick the tasks listed in the 'Default Task' column on the UCI page.

Linear Discrimanant Analysis and Gaussian MIxtrue Model

Question 5: Before selecting the methods you listed in response to Question 4, what are all methods you originally considered to use for the selected data mining task? What was your rationale for selecting the methods you listed in response to Question 4? What was your rationale for not selecting other methods?

Kmeans,GMM, DBscan,and complete clustering.
Performing the mining task with these methods it found that by using Kmeans the clusters were able to form very well. i selected this method because the dataset seems to be gobular when i visualized, where it performed well.
since the GMM ability to form elongated shaped clusters is simillar and more advantage to Kmeans foor this data. so i selected GMM for clustering
the other algorithms i selected is db scan and complete clustering. complete clustering was very good in forming the clusters but more than the complete clustering GMM was good.
Dbscan failed to form the clusters as the densities in the data is varying and failed to form the correct clusters.

Question 6: What limitations does your 'selected' method(s) has(have) that may limit your ability to accomplish the goal you have set for yourself?

At the First sight the variance i could able to capture was less than 90% where we need to get 90 % of variance during the reduction of dimensions to form the good clusters. but i performed with the low variance(but not bad to chose it) and there is considerable low percentage of simillarity between the true labels and predicted labels after the clustering.

Question 7: Do you have any alternative plan/strategy to overcome the above limitation(s)?

i overcome it by doing it with LDA since an improvement in variance capturing after reduction of dimensions will form good meaningful clusters

Question 8: For each of the methods you want to use, what parameter choices do you want to use and why? It does not have to be one parameter choice, it could be a collection or a range of choices you may want to consider.

the parameters i want to use is variance capturing before and after the dimension reduction. Rand index to measure the similarity between the true labels and predicted labels.

Question 9: How will you evaluate that you are successful in your pursuing your goal at the end of the project? In other words, what is your evaluation criteria?

i compared the result by the parameter rand index with the true labels. i was successful in forming the clusters with a good percentage of 97.8
i was successful in capturing the variance after reducing the dimensions with almost 98% using it.

Question 10: How will you evaluate that you are successful in your pursuing your goal at the end of the project? In other words, what is your evaluation criteria?

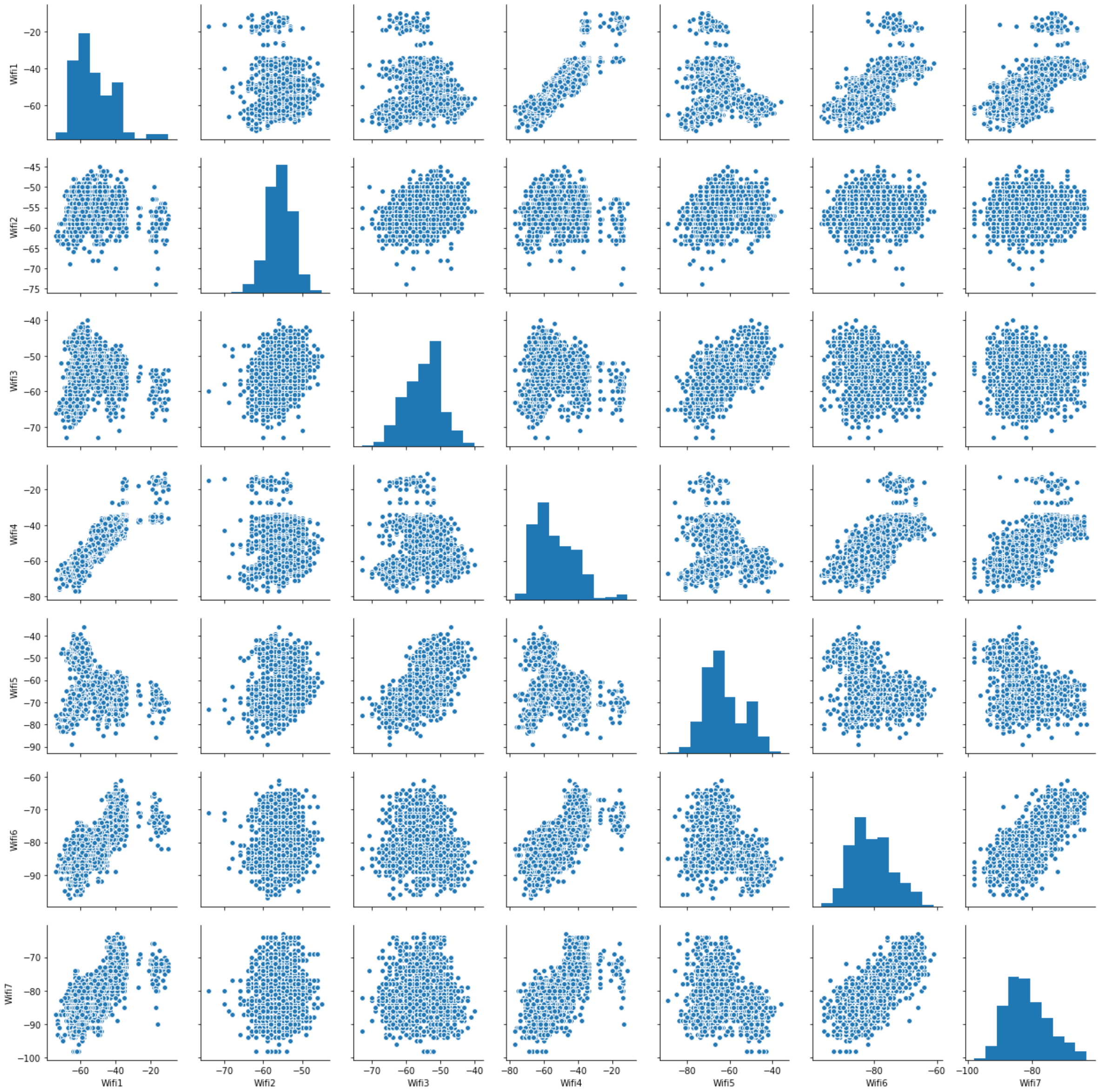
i compared the result by the parameter rand index with the true labels. i was successful in forming the clusters with a good percentage of 96.5. i was successful in capturing the variance after reducing the dimensions with almost 97% using it.

Question 11: Show any visualizations you may have generated to understand your data. Please include the code you used and the plots below. If you borrowed code (entirely or partially) from the hands-on projects or anywhere else, clearly provide a link to your source.

You may use this package to load UCI data in python: https://github.com/SkafteNicki/py_uci (https://github.com/SkafteNicki/py_uci)

In [88]:

```
sns.pairplot(df_X)
plt.show()
```



the data points for wifi1 and wifi4 can be good to consider for a classification data

In [100]:

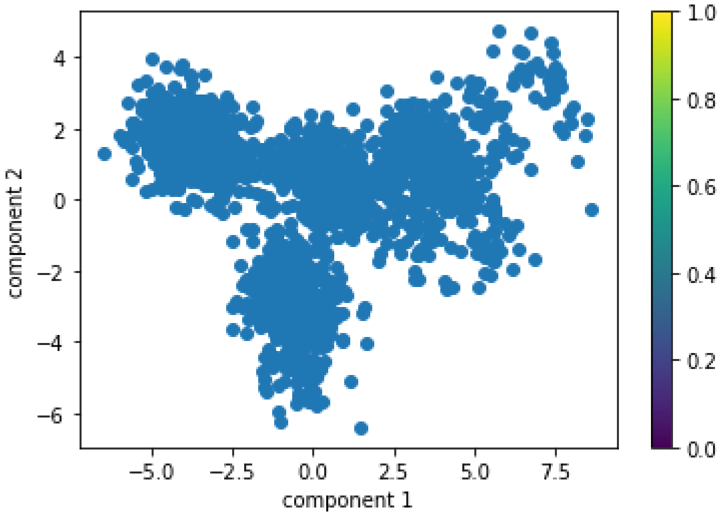
```
lda = LinearDiscriminantAnalysis(n_components=2)
df_X_LDA = lda.fit(df_X,df_Y).transform(df_X)
df_X_LDA.shape
```

Out[100]:

```
(2000, 2)
```

In [97]:

```
plt.scatter(df_X_LDA[:, 0], df_X_LDA[:, 1])
plt.xlabel('component 1')
plt.ylabel('component 2')
plt.colorbar();
plt.show()
```



In [99]:

```
lda.explained_variance_ratio_.cumsum()[1]
```

Out[99]:

```
0.979075605126304
```

the variance caputred by reducing the dimensions was good. so we can consider LDA to form the clusters using Kmeans.

Question 12: Perform data mining, evaluate your work and report your findings. This should include code, plots and results you may have generated. If you borrowed code (entirely or partially) from the hands-on projects or anywhere else, clearly provide a link to your source.

Note: the below code was taken from Handson CLUS

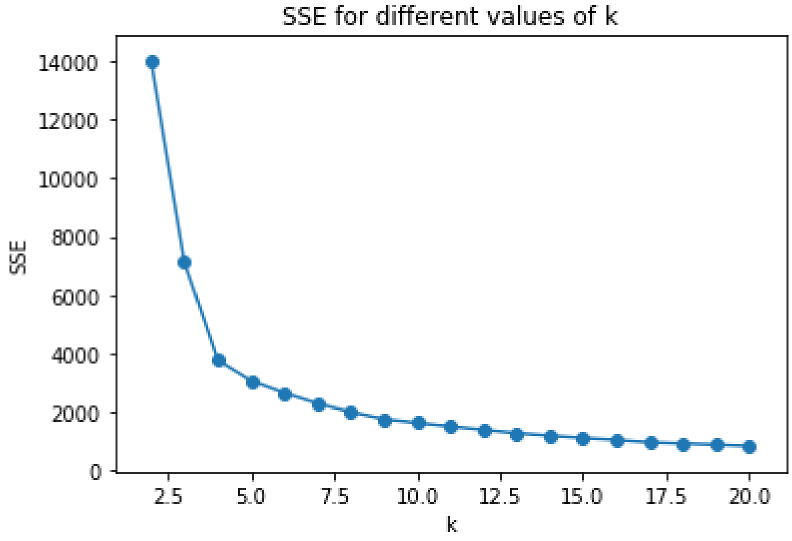
In [104]:

```
score = np.zeros(21)# this was taken from handson CLUS
for i in range(2,21):
    kmeans = KMeans(n_clusters=i, random_state=random_state); #Initializing KMeans for diff
    kmeans.fit_predict(df_X_LDA)  #Clustering using KMeans
    score[i] = -kmeans.score(df_X_LDA)  #Computing SSE
    print("SSE for k=",i,":", round(score[i],2)) #Printing SSE
```

```
SSE for k= 2 : 13978.74
SSE for k= 3 : 7113.26
SSE for k= 4 : 3760.59
SSE for k= 5 : 3070.07
SSE for k= 6 : 2656.94
SSE for k= 7 : 2299.84
SSE for k= 8 : 2009.03
SSE for k= 9 : 1765.33
SSE for k= 10 : 1640.07
SSE for k= 11 : 1518.02
SSE for k= 12 : 1403.86
SSE for k= 13 : 1289.47
SSE for k= 14 : 1205.05
SSE for k= 15 : 1128.94
SSE for k= 16 : 1059.54
SSE for k= 17 : 977.64
SSE for k= 18 : 938.62
SSE for k= 19 : 899.8
SSE for k= 20 : 851.58
```

In [105]:

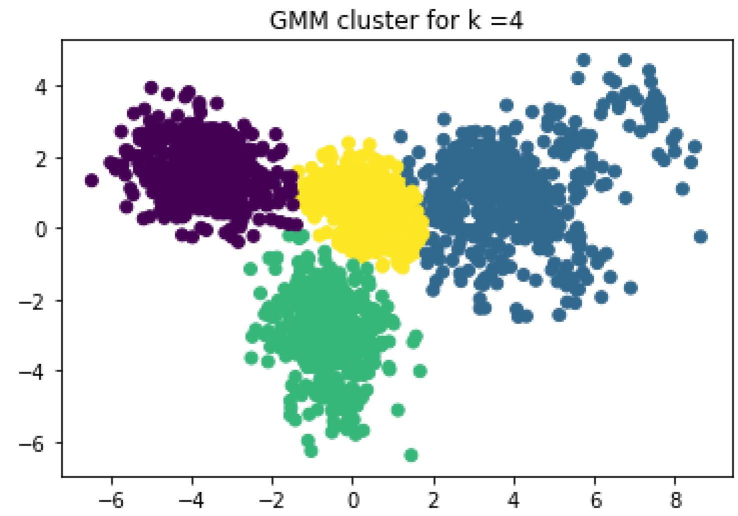
```
plt.plot(range(2,21),score[2:21])
plt.scatter(range(2,21),score[2:21])
plt.xlabel('k')
plt.ylabel('SSE')
plt.title('SSE for different values of k')
plt.show()
```



there was no change in values from k =4 and we have knee point at that value.

In [112]:

```
gmm = GaussianMixture(n_components=4, covariance_type='full')
y_pred = gmm.fit_predict(df_X_LDA)
fig = plt.scatter(df_X_LDA[:, 0], df_X_LDA[:, 1], c=y_pred)
plt.title('GMM cluster for k =4')
plt.show()
```



From the observation, the kmeans is able to form the good clusters for the value 4

```
Note: The below code was used from Handson-CLUS
```

In [108]:

```
from scipy.special import comb                              #this code was taken from Handso
def rand_index(S, T):

    Spairs = comb(np.bincount(S), 2).sum()
    Tpairs = comb(np.bincount(T), 2).sum()

    A = np.c_[(S, T)]

    f_11 = sum(comb(np.bincount(A[A[:, 0] == i, 1]), 2).sum()
             for i in set(S))

    f_10 = Spairs - f_11
    f_01 = Tpairs - f_11
    f_00 = comb(len(A), 2) - f_11 - f_10 - f_01
    return (f_00 + f_11) / (f_00 + f_01 + f_10 + f_11)
```

In [113]:

```
rand_index(y_pred, df_Y)
```

Out[113]:

```
0.9786258129064532
```

```
the simillarity between true labels and predicted labels is 96 which was a good sign
```

In [121]:

```
labels = ["KMeans","GMM","Complete_link","Average_link","DBscan"]
plt.figure(figsize=(13,12))
count = 1
n_clusters=4
kmeans = KMeans(n_clusters=n_clusters, random_state=random_state);
gmm = GaussianMixture(n_components=n_clusters, covariance_type='full')
complete_linkage = AgglomerativeClustering(linkage="complete", n_clusters=n_clusters)
average_linkage = AgglomerativeClustering(linkage="average", n_clusters=n_clusters)
dbscan = DBSCAN(eps=1, min_samples=10)
spectral = SpectralClustering(n_clusters=n_clusters, random_state=random_state)
methods = [kmeans,gmm,complete_linkage,average_linkage,dbscan]
for label,method in zip(labels,methods) :
    y_pred = method.fit_predict(df_X_LDA)
    y_pred=np.where(y_pred==-1,5,y_pred) #assigning a class to negative variable for DBscan
    score  = rand_index(y_pred, df_Y)
    print(label+": "+str(score))
```

```
KMeans: 0.9653091545772886
GMM: 0.9786258129064532
Complete_link: 0.9670490245122562
Average_link: 0.8663056528264133
DBscan: 0.2516278139069535
```

different algorithms and their simillarity index of predicted lables with true labels.GMM performed well for this dataset

Question 13: Putting your findings in the context of your goal and evaluation plan, do you consider yourself successful? Provide reasons for your success or lack thereof.

the findings of the goal are successful but there are some clarifications needed that need to be sorted of.
but the goal of finding the clusters for the classification dataset is successful

Question 14: If you have an extra month to work on this project, what else would you do? Provide reasons.

all the clustering methods have seen the increasing trend on applying from PCA to LDA . but the DBScan method, the performance is decreased.
if i had an extra month to work on i would like to see performances with all clusterings and why there is a negative trend for DB scan using the LDA

Question 15: Do you consider this project to be in the 'innovative category' or a 'good application' category? Provide your reason.

i have worked on with two datasets previously like lenses Dataset on UCI and have done the analysis using clustering. there is low performance on these datasets using the clustering. i concluded to know that these have Deafult tasks and no other algorithm could work on these from it.
but the dataset i choose, works fine with clustering where in UCI repository it was not mentioned.
so i consider this is valid result , where Clustering is also possible for this dataset.

Also i consider this as a good application from the dataset view where we can locate using the signal strength by clusters.