## Project of data science 2021

## 1. Overall Goal

Using Python, matlab, Java or some things else to analysis data and then visualize them in-group(5 students in a group).

## 2. Data set

The attendance table for each classes, totally 12. Each of tables is in excel format. The overall goals is to give remarks for every students according to the attendance record.

Firstly, you need read the data, I will show you one way using pandas

import pandas as pd  
sExcelFile="data1-5.xlsx"  
df1 = pd.read\_excel(sExcelFile,sheet\_name='Sheet1')

df1

.dataframe tbody tr th {  
 vertical-align: top;  
}  
  
.dataframe thead th {  
 text-align: right;  
}

sExcelFile="data6.xlsx"  
df2 = pd.read\_excel(sExcelFile,sheet\_name='数据导出')  
df2

.dataframe tbody tr th {  
 vertical-align: top;  
}  
  
.dataframe thead th {  
 text-align: right;  
}

## 3 Your 1st task

your need to immerge table df1 and df2, according to 学生姓名. I suggest you use your ID in our class , but alots of students use various kinds of names. If you cannot merge df1 and df2 automaticly , try to modify data6 files by your hand.

## 4 Data per-processing

your need extract the numbers from class1 -class 5

a1=df1.class1.str.extract('(\d+)')  
a2=df1.class2.str.extract('(\d+)')  
a3=df1.class3.str.extract('(\d+)')  
a4=df1.class4.str.extract('(\d+)')  
a5=df1.class5.str.extract('(\d+)')  
b = pd.concat([a1,a2,a3,a4,a5],axis=1)  
b

.dataframe tbody tr th {  
 vertical-align: top;  
}  
  
.dataframe thead th {  
 text-align: right;  
}

Here are lots of missing vaule(or absent as NAN) you can fill it in your way

b1=b.fillna(10);b1

.dataframe tbody tr th {  
 vertical-align: top;  
}  
  
.dataframe thead th {  
 text-align: right;  
}

## 4 (Optimal)To do some EDA for your data set e.g. mean variance , store your data in a data base e.g. sqlite3

b1.describe()

.dataframe tbody tr th {  
 vertical-align: top;  
}  
  
.dataframe thead th {  
 text-align: right;  
}

## 5. Clustering

from sklearn.cluster import KMeans

kmeans = KMeans(n\_clusters=4)

kmeans =kmeans.fit(b1)

y\_kmeans = kmeans.predict(b1)

y\_kmeans

array([0, 3, 1, 0, 1, 0, 0, 3, 3, 3, 0, 0, 3, 3, 1, 3, 3, 2, 0, 0, 0, 2,  
 3, 0, 0])

## 7. Visulazation

import matplotlib.pyplot as plt