Hand on project of data science 2021

1. Overall Goal

Using Python, matlab, Java or some programming languages else to analysis data and then visualize them. You will work toghter as a group,each group will have 5 students.

2. Data set

The attendance table for each classes, totally 12. Each of tables is in excel format. The data set 1-5 is communal.

Group 1 is suggested use data 1-5 puls data 6 and 7

Group 2 is suggested use data 1-5 puls data 7 and 8

Group 3 is suggested use data 1-5 puls data 9 and 10

Group 4 is suggested use data 1-5 puls data 10 and 11

The overall goals is to give remarks for every students according to the attendance record thus a clustering problem.

step1

Firstly, you need read the data, I will show you an example using pandas with Python, Fell free to use something esle to programing.

```
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;
}
```

	学号	中文名	英文名	class1	class2	class3	class4	class5
0	IS17170201	王辰	AL-AWADHI HUSAM TAHA MOHAMMED	44min	6min	27min	12min	14min
1	IS17170202	NaN	AL-KATEB ELYAS KHALED MOHAMMED	NaN	90min	89min	51min	68min
2	IS17170203	赵雷	AHSAN MUHAMMAD	45min	6min	41min	78min	96min
3	IS17170204	NaN	AJMAL HAMZA	NaN	62min	NaN	NaN	NaN
4	IS17170205	诺曼	ARSHAD NOMAN	65min	22min	50min	72min	12min
5	IS17170206	吴越寺	ASHFAQ HAMZA	NaN	25min	13min	NaN	57min
6	IS17170208	张磊	BISWAS DIPALOKE	NaN	NaN	NaN	77min	23min
7	IS17170209	李斯诺	CHOWDHURY SATYEN ROY	50min	88min	68min	67min	49min
8	IS17170210	卫斯	FAROOQ AWAIS	76min	85min	80min	76min	94min
9	IS17170212	萨娜	GOHAR SANA	49min	55min	79min	62min	81min
10	IS17170213	NaN	HASSAN SYED MUBASHAR	NaN	NaN	NaN	NaN	NaN
11	IS17170214	马云	IFTIKHAR MUHAMMAD TALHA	NaN	NaN	NaN	NaN	4min
12	IS17170215	麦家	KAZMI SYED REHMAN RAZA	74min	92min	87min	61min	74min
13	IS17170216	刘佳令	NOOR NUKHBA	77min	83min	86min	78min	38min
14	IS17170217	李飞	RAHMAN MOHAMMAD HABIBUR	49min	NaN	59min	36min	77min
15	IS17170218	步豹	ABU BAKAR	79min	98min	76min	NaN	88min
16	IS17170219	吴思远	ALI TAIMUR	77min	92min	84min	NaN	80min

	学号	中文名	英文名	class1	class2	class3	class4	class5
17	IS17170220	安思宁	ASHIQ MAHER MUHAMMAD	NaN	79min	18min	75min	77min
18	IS17170222	沈易	KHAN SHARJEEL	23min	34min	42min	1min	76min
19	IS17170223	可力得	MASOOD MUHAMMAD QASIM	NaN	NaN	NaN	NaN	NaN
20	IS17170224	齐飞	QURESHI UMAR	76min	NaN	NaN	1min	NaN
21	IS17170225	王子	UMAR MUHAMMAD	NaN	83min	NaN	80min	NaN
22	IS17170226	阿力	ALI ASIF	45min	90min	89min	77min	88min
23	IS17170227	夏米大	SHAHID ASFANDYAR	NaN	NaN	NaN	NaN	NaN
24	IS17170228	马龙	SAYFIDDINOV JAVLON	NaN	NaN	11min	NaN	2min

you can see the details of given data1-5, Now we should use data 6 as a complemntary.

```
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;
}
```

	学生姓名	学生ID	是否观看直播	进入直播时间	观看直播时长	是否观看回放	观看回放时间
0	AL-AWADHI HUSAM TAHA MOHAMMED	1.441153e+17	是	2021/09/30 16:13:14	12分钟	否	0分钟
1	AL-KATEB ELYAS KHALED MOHAMMED	1.441153e+17	是	2021/09/30 16:07:27	14分钟	否	0分钟
2	AHSAN MUHAMMAD	1.441153e+17	是	2021/09/30 16:03:12	14分钟	否	0分钟
3	AJMAL HAMZA	1.441153e+17	是	2021/09/30 16:02:52	23分钟	否	0分钟
4	ARSHAD NOMAN	1.441153e+17	是	2021/09/30 15:59:51	4分钟	否	0分钟
5	ASHFAQ HAMZA	1.441153e+17	是	2021/09/30 15:41:12	57分钟	否	0分钟
6	BISWAS DIPALOKE	1.441153e+17	是	2021/09/30 15:18:52	68分钟	否	0分钟
7	CHOWDHURY SATYEN ROY	1.441153e+17	是	2021/09/30 15:08:57	77分钟	否	0分钟
8	FAROOQ AWAIS	1.441153e+17	是	2021/09/30 15:05:11	76分钟	否	0分钟
9	GOHAR SANA	1.441153e+17	是	2021/09/30 15:05:01	77分钟	否	0分钟
10	HASSAN SYED MUBASHAR	1.441153e+17	是	2021/09/30 15:02:31	81分钟	否	0分钟
11	IFTIKHAR MUHAMMAD TALHA	1.441153e+17	是	2021/09/30 15:00:07	38分钟	否	0分钟
12	KAZMI SYED REHMAN RAZA	1.441153e+17	是	2021/09/30 14:59:10	80分钟	否	0分钟
13	NOOR NUKHBA	1.441153e+17	是	2021/09/30 14:58:58	2分钟	否	0分钟
14	RAHMAN MOHAMMAD HABIBUR	1.441152e+17	是	2021/09/30 14:57:04	88分钟	否	0分钟
15	ABU BAKAR	1.441153e+17	是	2021/09/30 14:56:55	88分钟	否	0分钟
16	ALI TAIMUR	1.441153e+17	是	2021/09/30 14:56:42	96分钟	否	0分钟
17	ASHIQ MAHER MUHAMMAD	1.441153e+17	是	2021/09/30 14:56:24	94分钟	否	0分钟

	学生姓名	学生ID	是否观看直播	进入直播时间	观看直播时长	是否观看回放	观看回放时间
18	KHAN SHARJEEL	1.441153e+17	是	2021/09/30 14:56:12	49分钟	否	0分钟
19	MASOOD MUHAMMAD QASIM	1.441153e+17	是	2021/09/30 14:55:55	74分钟	否	0分钟
20	QURESHI UMAR	1.441152e+17	是	2021/09/30 14:54:26	83分钟	否	0分钟
21	UMAR MUHAMMAD	NaN	NaN	NaN	NaN	NaN	NaN
22	ALI ASIF	NaN	NaN	NaN	NaN	NaN	NaN
23	SHAHID ASFANDYAR	NaN	NaN	NaN	NaN	NaN	NaN
24	SAYFIDDINOV JAVLON	NaN	NaN	NaN	NaN	NaN	NaN

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.

```
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;
}
```

	观看直播时长
0	12分钟
1	14分钟
2	14分钟
3	23分钟
4	4分钟
5	57分钟
6	68分钟
7	77分钟
8	76分钟
9	77分钟
10	81分钟
11	38分钟
12	80分钟
13	2分钟
14	88分钟
15	88分钟
16	96分钟
17	94分钟
18	49分钟
19	74分钟
20	83分钟

```
df= pd.concat([df1,df2])
df
```

```
d:\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: FutureWarning: Sorting because non-
concatenation axis is not aligned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.

  """Entry point for launching an IPython kernel.
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	学号	中文名	英文名	class1	class2	class3	class4	class5	class6
0	IS17170201	王辰	AL-AWADHI HUSAM TAHA MOHAMMED	44min	6min	27min	12min	14min	12分钟
1	IS17170202	NaN	AL-KATEB ELYAS KHALED MOHAMMED	NaN	90min	89min	51min	68min	14分钟
2	IS17170203	赵雷	AHSAN MUHAMMAD	45min	6min	41min	78min	96min	14分钟
3	IS17170204	NaN	AJMAL HAMZA	NaN	62min	NaN	NaN	NaN	23分钟
4	IS17170205	诺曼	ARSHAD NOMAN	65min	22min	50min	72min	12min	4分钟
5	IS17170206	吴越寺	ASHFAQ HAMZA	NaN	25min	13min	NaN	57min	57分钟
6	IS17170208	张磊	BISWAS DIPALOKE	NaN	NaN	NaN	77min	23min	68分钟
7	IS17170209	李斯诺	CHOWDHURY SATYEN ROY	50min	88min	68min	67min	49min	77分钟
8	IS17170210	卫斯	FAROOQ AWAIS	76min	85min	80min	76min	94min	76分钟
9	IS17170212	萨娜	GOHAR SANA	49min	55min	79min	62min	81min	77分钟
10	IS17170213	NaN	HASSAN SYED MUBASHAR	NaN	NaN	NaN	NaN	NaN	81分钟
11	IS17170214	马云	IFTIKHAR MUHAMMAD TALHA	NaN	NaN	NaN	NaN	4min	38分钟
12	IS17170215	麦家	KAZMI SYED REHMAN RAZA	74min	92min	87min	61min	74min	80分钟
13	IS17170216	刘佳令	NOOR NUKHBA	77min	83min	86min	78min	38min	2分钟
14	IS17170217	李飞	RAHMAN MOHAMMAD HABIBUR	49min	NaN	59min	36min	77min	88分钟
15	IS17170218	步豹	ABU BAKAR	79min	98min	76min	NaN	88min	88分钟

	学号	中文名	英文名	class1	class2	class3	class4	class5	class6
16	IS17170219	吴思远	ALI TAIMUR	77min	92min	84min	NaN	80min	96分钟
17	IS17170220	安思宁	ASHIQ MAHER MUHAMMAD	NaN	79min	18min	75min	77min	94分钟
18	IS17170222	沈易	KHAN SHARJEEL	23min	34min	42min	1min	76min	49分钟
19	IS17170223	可力得	MASOOD MUHAMMAD QASIM	NaN	NaN	NaN	NaN	NaN	74分钟
20	IS17170224	齐飞	QURESHI UMAR	76min	NaN	NaN	1min	NaN	83分钟
21	IS17170225	王子	UMAR MUHAMMAD	NaN	83min	NaN	80min	NaN	NaN
22	IS17170226	阿力	ALI ASIF	45min	90min	89min	77min	88min	NaN
23	IS17170227	夏米大	SHAHID ASFANDYAR	NaN	NaN	NaN	NaN	NaN	NaN
24	IS17170228	马龙	SAYFIDDINOV JAVLON	NaN	NaN	11min	NaN	2min	NaN

4 Data per-processing

your need extract the numbers as the features you used in 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+)')
a6=df1.class6.str.extract('(\d+)')
b = pd.concat([a1,a2,a3,a4,a5,a6],axis=1)
b
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	0	0	0	0	0	0
0	44	6	27	12	14	12
1	NaN	90	89	51	68	14
2	45	6	41	78	96	14
3	NaN	62	NaN	NaN	NaN	23
4	65	22	50	72	12	4
5	NaN	25	13	NaN	57	57
6	NaN	NaN	NaN	77	23	68
7	50	88	68	67	49	77
8	76	85	80	76	94	76
9	49	55	79	62	81	77
10	NaN	NaN	NaN	NaN	NaN	81
11	NaN	NaN	NaN	NaN	4	38
12	74	92	87	61	74	80
13	77	83	86	78	38	2
14	49	NaN	59	36	77	88
15	79	98	76	NaN	88	88
16	77	92	84	NaN	80	96
17	NaN	79	18	75	77	94
18	23	34	42	1	76	49
19	NaN	NaN	NaN	NaN	NaN	74
20	76	NaN	NaN	1	NaN	83
21	NaN	83	NaN	80	NaN	NaN
22	45	90	89	77	88	NaN
23	NaN	NaN	NaN	NaN	NaN	NaN
24	NaN	NaN	11	NaN	2	NaN

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

```
b=b.fillna(10);b1
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	0	0	0	0	0	0
0	44	6	27	12	14	12
1	10	90	89	51	68	14
2	45	6	41	78	96	14
3	10	62	10	10	10	23
4	65	22	50	72	12	4
5	10	25	13	10	57	57
6	10	10	10	77	23	68
7	50	88	68	67	49	77
8	76	85	80	76	94	76
9	49	55	79	62	81	77
10	10	10	10	10	10	81
11	10	10	10	10	4	38
12	74	92	87	61	74	80
13	77	83	86	78	38	2
14	49	10	59	36	77	88
15	79	98	76	10	88	88
16	77	92	84	10	80	96
17	10	79	18	75	77	94
18	23	34	42	1	76	49
19	10	10	10	10	10	74
20	76	10	10	1	10	83
21	10	83	10	80	10	10
22	45	90	89	77	88	10
23	10	10	10	10	10	10
24	10	10	11	10	2	10

4. Clustering

You can use the package sklearn, the function kmeans is a clustering method, and you will clusting your attendance as 4 grades A B C D

```
from sklearn.cluster import KMeans
```

```
kmeans = KMeans(n_clusters=4)
```

```
kmeans =kmeans.fit(b1)
```

```
y_kmeans = kmeans.predict(b1)
```

```
y_kmeans
```

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

5. Visulazation

Visulaze with hist graph and box graph

```
import matplotlib.pyplot as plt
```

```
plt.boxplot(y_kmeans)
plt.show()
```


```
from scipy.stats import norm
fig,ax = plt.subplots()
plt.rcParams['font.family'] = ['SimHei']
n,bins,patches = ax.hist(y_kmeans,density = 1)
mu = 1
sigma = 0.4
y=norm.pdf(bins,mu,sigma)
ax.plot(bins,y,'--')
fig.tight_layout()
plt.show()
```