碩管一甲　M09218001　周彥廷
數據分析與應用-作業 1：dplyr 特徵擷取實作 python 重新上傳版本

```
# 數據分析與應用作業 1 dplyr 特徵擷取 M09218001 周彥廷 (PYTHON 版本)
import pandas as pd
import numpy as np

# 讀取資料與清洗資料
data = pd.read_excel("C:/Users/MCUT/Desktop/dataset1.xlsx", header=0, sep=',',
encoding='big5')
df = data.dropna()

# 查看資料
print(df.head(20))
print("資料數量", df.shape)
print("資料欄位", df.columns)
print(type(df))
```

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數
2016-09-02 00:00:00	2016-10-19 00:00:00	N70707	F8157711	125040	191360	0	2
2016-05-12 00:00:00	2016-05-31 00:00:00	N21223	F8218210	179997	165000	0	2
2016-11-25 00:00:00	2017-01-23 00:00:00	N99991	F2103140	157199	157000	0	1
2015-10-22 00:00:00	2015-12-25 00:00:00	N13132	F8131110	179000	179000	0	1
2017-02-14 00:00:00	2017-06-27 00:00:00	N97971	F9102122	92336	119000	0	2
2015-08-12 00:00:00	2015-12-18 00:00:00	N82828	F4971164	175500	175500	0	1
2015-12-04 00:00:00	2015-12-30 00:00:00	N70707	F3135120	143000	143000	0	2
2016-02-15 00:00:00	2016-03-03 00:00:00	N83831	F6479685	142825	134400	0	3
2016-02-17 00:00:00	2016-05-24 00:00:00	N79801	F8131110	127151	127150	0	1
2016-07-15 00:00:00	2016-08-08 00:00:00	N71717	F3135120	136973	136946	0	3
2017-02-10 00:00:00	2017-03-13 00:00:00	N83831	F3135120	171019	171000	0	3

```
# 1.篩選
#1.1 "辦事員"=N70707
a=df[df['辦事員'].str.contains('N70707')]
```

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數
2016-09-02 00:00:00	2016-10-19 00:00:00	N70707	F8157711	125040	191360	0	2
2015-12-04 00:00:00	2015-12-30 00:00:00	N70707	F3135120	143000	143000	0	2
2015-12-14 00:00:00	2016-01-19 00:00:00	N70707	F1915213	117515	168677	0	3
2015-12-16 00:00:00	2016-01-19 00:00:00	N70707	F2933162	51490	110000	8612	5
2016-03-07 00:00:00	2016-04-01 00:00:00	N70707	F3657512	180500	180500	0	1
2016-03-17 00:00:00	2016-04-07 00:00:00	N70707	F3657512	134900	134900	0	3
2017-07-21 00:00:00	2017-08-15 00:00:00	N70707	F1109131	150800	147400	0	2
2017-03-10 00:00:00	2017-05-09 00:00:00	N70707	F2136100	141059	155000	0	8
2016-03-10 00:00:00	2016-04-28 00:00:00	N70707	F2101834	126557	150000	0	4
2017-03-16 00:00:00	2017-03-20 00:00:00	N70707	F2933162	124107	124107	0	1

#1.2 "決包金額"介於 100000~130000

b=df.loc[(df['決包金額']>100000)&(df['決包金額']<130000)]

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數
2017-02-14 00:00:00	2017-06-27 00:00:00	N97971	F9102122	92336	119000	0	2
2016-02-17 00:00:00	2016-05-24 00:00:00	N79801	F8131110	127151	127150	0	1
2017-03-31 00:00:00	2017-05-02 00:00:00	N83831	F2117120	103500	101000	0	2
2016-02-15 00:00:00	2016-04-26 00:00:00	N13132	F1454010	116571	116000	0	1
2016-11-25 00:00:00	2016-12-23 00:00:00	N21223	F1464976	115500	111500	0	1
2017-04-05 00:00:00	2017-05-11 00:00:00	N65656	F9171150	116572	116572	0	1
2017-06-23 00:00:00	2017-07-28 00:00:00	N14157	F8998995	129161	120000	0	1
2016-11-24 00:00:00	2017-01-20 00:00:00	N79801	F8611885	427495	128000	0	1
2016-08-04 00:00:00	2016-09-09 00:00:00	N50518	F2114107	126260	125000	0	1
2017-01-06 00:00:00	2017-02-07 00:00:00	N39393	F8651151	104712	104640	0	1
2016-03-09 00:00:00	2016-06-21 00:00:00	N39393	F8881266	119248	102800	0	1

#1.3 "決包金額">180000 且 "報價廠商家數" = 1

c=df.loc[(df['決包金額']>180000)&(df['報價廠商家數']==1)]

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數
2016-12-13 00:00:00	2017-01-13 00:00:00	N99991	F3130116	187999	187000	0	1
2016-05-30 00:00:00	2016-07-15 00:00:00	N13132	F6118571	185452	185000	0	1
2016-07-19 00:00:00	2016-08-22 00:00:00	N22222	F1115100	202140	200000	0	1
2017-04-25 00:00:00	2017-05-24 00:00:00	N22238	F1975172	190842	190000	0	1
2016-02-24 00:00:00	2016-05-25 00:00:00	N99991	F7751367	191472	191472	0	1
2016-07-12 00:00:00	2016-08-05 00:00:00	N88938	F2981011	190686	188000	227000	1
2016-06-15 00:00:00	2016-07-05 00:00:00	N21223	F1103961	188296	188000	-4558	1
2016-04-15 00:00:00	2016-05-10 00:00:00	N88938	F8137130	200923	200000	46707	1
2017-03-17 00:00:00	2017-04-27 00:00:00	N96971	F2149751	198019	198000	0	1
2016-05-26 00:00:00	2016-06-27 00:00:00	N97979	F8392699	194237	194000	0	1

#1.4 "決包金額">180000 或 "報價廠商家數" = 1

d=df.loc[(df['決包金額']>180000)|(df['報價廠商家數']==1)]

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數
2016-09-02 00:00:00	2016-10-19 00:00:00	N70707	F8157711	125040	191360	0	2
2016-11-25 00:00:00	2017-01-23 00:00:00	N99991	F2103140	157199	157000	0	1
2015-10-22 00:00:00	2015-12-25 00:00:00	N13132	F8131110	179000	179000	0	1
2015-08-12 00:00:00	2015-12-18 00:00:00	N82828	F4971164	175500	175500	0	1
2016-02-17 00:00:00	2016-05-24 00:00:00	N79801	F8131110	127151	127150	0	1
2016-02-15 00:00:00	2016-04-26 00:00:00	N13132	F1454010	116571	116000	0	1
2016-12-13 00:00:00	2017-01-13 00:00:00	N99991	F3130116	187999	187000	0	1
2017-01-12 00:00:00	2017-04-04 00:00:00	N65656	F7157120	145710	145710	0	1
2017-05-23 00:00:00	2017-06-14 00:00:00	N33128	F3564811	138671	138000	0	1
2017-06-06 00:00:00	2017-07-04 00:00:00	N22238	F2794769	181296	180000	0	1

#1.5 "辦事員"開頭為 N1 者

e=df.loc[df['辦事員'].str.contains('N1')]

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數
2015-10-22 00:00:00	2015-12-25 00:00:00	N13132	F8131110	179000	179000	0	1
2016-02-15 00:00:00	2016-04-26 00:00:00	N13132	F1454010	116571	116000	0	1
2016-05-30 00:00:00	2016-07-15 00:00:00	N13132	F6118571	185452	185000	0	1
2017-05-25 00:00:00	2017-06-29 00:00:00	N14157	F3947911	148175	136800	0	2
2017-06-23 00:00:00	2017-07-28 00:00:00	N14157	F8998995	129161	120000	0	1
2016-12-28 00:00:00	2017-02-13 00:00:00	N14157	F2100545	147500	147500	0	1
2017-01-24 00:00:00	2017-03-22 00:00:00	N14157	F4637947	151321	151000	0	1
2017-06-15 00:00:00	2017-06-21 00:00:00	N13132	F2162110	179489	169431	0	1
2016-06-08 00:00:00	2016-06-16 00:00:00	N14157	F8481345	125280	118800	0	1
2016-09-12 00:00:00	2016-10-06 00:00:00	N13132	F8738767	215818	200000	0	5
2016-04-26 00:00:00	2016-08-10 00:00:00	N14157	F7111477	130176	180000	-142221	3
2017-04-24 00:00:00	2017-06-05 00:00:00	N13132	F2142727	179820	175000	0	3

2.排序

#2.1 按照"呈核日"排序(由小而大排)

f=df.sort_values(['呈核日'])

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數
2015-09-30 00:00:00	2015-10-16 00:00:00	N39393	F2159102	136417	136417	-27585	1
2015-10-14 00:00:00	2015-11-10 00:00:00	N65656	F1831429	115000	115000	0	3
2015-10-14 00:00:00	2015-11-18 00:00:00	N64647	F7161110	160493	160000	0	1
2015-10-13 00:00:00	2015-11-23 00:00:00	N66666	F5606431	157116	126000	-101926	1
2015-10-23 00:00:00	2015-11-25 00:00:00	N99991	F7611449	172938	172938	0	1
2015-10-29 00:00:00	2015-11-26 00:00:00	N97979	F8175130	116675	116500	0	1
2015-09-21 00:00:00	2015-12-16 00:00:00	N18197	F4577144	139955	170000	0	4
2015-11-26 00:00:00	2015-12-18 00:00:00	N65661	F7861001	150000	150000	-50000	2
2015-08-12 00:00:00	2015-12-18 00:00:00	N82828	F4971164	175500	175500	0	1
2015-11-06 00:00:00	2015-12-21 00:00:00	N97979	F2731147	84512	155000	0	2

#2.2 按照"辦事員"、"報價廠商家數"排序(由小而大排)

g=df.sort_values(['辦事員']+['報價廠商家數'])

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數
2015-07-31 00:00:00	2015-12-23 00:00:00	N11183	F3107529	89079	198000	-47198	4
2015-10-22 00:00:00	2015-12-25 00:00:00	N13132	F8131110	179000	179000	0	1
2016-02-15 00:00:00	2016-04-26 00:00:00	N13132	F1454010	116571	116000	0	1
2016-05-30 00:00:00	2016-07-15 00:00:00	N13132	F6118571	185452	185000	0	1
2017-06-15 00:00:00	2017-06-21 00:00:00	N13132	F2162110	179489	169431	0	1
2016-06-22 00:00:00	2016-07-15 00:00:00	N13132	F1143130	140000	110000	0	1
2016-02-17 00:00:00	2016-03-29 00:00:00	N13132	F2136100	180111	180000	0	1
2016-06-23 00:00:00	2016-07-05 00:00:00	N13132	F1115666	153007	153000	210317	1
2016-02-24 00:00:00	2016-03-16 00:00:00	N13132	F1043252	159241	159200	0	1
2016-06-13 00:00:00	2016-06-21 00:00:00	N13132	F1468796	182515	182195	0	1

3.篩選後排序

選出"辦事員"開頭為 N1 者再依"報價廠商家數"排序

```
h=df.loc[df['辦事員'].str.contains('N1')].sort_values('報價廠商家數')
```

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數
2015-10-22 00:00:00	2015-12-25 00:00:00	N13132	F8131110	179000	179000	0	1
2017-01-24 00:00:00	2017-02-18 00:00:00	N13132	F2103765	120142	120000	0	1
2016-09-30 00:00:00	2016-10-17 00:00:00	N13132	F8146160	164246	164200	0	1
2016-11-30 00:00:00	2016-12-13 00:00:00	N14157	F8434513	187610	187000	0	1
2017-06-27 00:00:00	2017-07-14 00:00:00	N18197	F9174150	140936	138415	0	1
2016-08-23 00:00:00	2016-08-29 00:00:00	N13132	F2134130	122197	122197	0	1
2016-03-03 00:00:00	2016-03-29 00:00:00	N13132	F9919032	158809	158400	-97933	1
2016-12-14 00:00:00	2016-12-22 00:00:00	N13132	F3126110	115923	115923	0	1
2016-09-13 00:00:00	2016-12-06 00:00:00	N18197	F8139110	130280	130000	0	1
2016-08-05 00:00:00	2016-08-16 00:00:00	N14157	F2163127	140182	140182	0	1
2016-07-29 00:00:00	2016-08-23 00:00:00	N14157	F1454010	140031	140000	0	1

4.增加特徵

```
import datetime
import time
```

4.1 作業天數 = 呈核日-收件日

```
df['收件日']= pd.to_datetime(df['收件日'])
df['呈核日']= pd.to_datetime(df['呈核日'])
df['作業天數'] = df[:]['呈核日']-df[:]['收件日']
df['作業天數']=df['作業天數'].astype('timedelta64[D]')
df.groupby(df["作業天數"])
```

4.2 追加比例 = (決包金額+追加金額)/決包金額

```
df["追加比例"] = (df["決包金額"]+df["追加金額"])/df["決包金額"]
df.groupby(df["追加比例"])
```

4.3 超預算：若決包金額>預算金額，則為超預算，將其標示為 1、反之為 0
df['超預算']= np.where(df['決包金額']>df['預算金額'],1,0)
df.groupby(df['超預算'])

4.4 獨家報價：若報價廠商數=1，則為獨家，將其標示為 1、反之為 0
df['獨家報價']=np.where(df['報價廠商家數']==1,1,0)
df.groupby(df['獨家報價'])

table = df

收件日	呈核日	辦事員	施工廠商	預算金額	決包金額	追加金額	報價廠商家數	作業天數	追加比例	超預算	獨家報價
2016-09-02 …	2016-10-19 …	N70707	F8157711	125040	191360	0	2	47	1	1	0
2016-05-12 …	2016-05-31 …	N21223	F8218210	179997	165000	0	2	19	1	0	0
2016-11-25 …	2017-01-23 …	N99991	F2103140	157199	157000	0	1	59	1	0	1
2015-10-22 …	2015-12-25 …	N13132	F8131110	179000	179000	0	1	64	1	0	1
2017-02-14 …	2017-06-27 …	N97971	F9102122	92336	119000	0	2	133	1	1	0
2015-08-12 …	2015-12-18 …	N82828	F4971164	175500	175500	0	1	128	1	0	1
2015-12-04 …	2015-12-30 …	N70707	F3135120	143000	143000	0	2	26	1	0	0
2016-02-15 …	2016-03-03 …	N83831	F6479685	142825	134400	0	3	17	1	0	0
2016-02-17 …	2016-05-24 …	N79801	F8131110	127151	127150	0	1	97	1	0	1
2016-07-15 …	2016-08-08 …	N71717	F3135120	136973	136946	0	3	24	1	0	0
2017-02-10 …	2017-03-13 …	N83831	F3135120	171019	171000	0	3	31	1	0	0
2017-02-24 …	2017-03-31 …	N58581	F6479685	166808	130000	0	3	35	1	0	0
2017-03-31 …	2017-05-02 …	N83831	F2117120	103500	101000	0	2	32	1	0	0

#5 計算每位"辦事員"的
1.決包金額最小值
2.決包金額最大值
3.平均作業天數
4.平均追加比例
5.超預算件數
6.獨家報價件數
先剔除呈核日、收件日的遺漏值列，將滿足"平均作業天數">15、"超預算件數">3、"獨家報價件數">3"辦事員"的明細列出來

A1=table.groupby(['辦事員']).min()['決包金額'].rename('決包金額最小值')
A2=table.groupby(['辦事員']).max()['決包金額'].rename('決包金額最大值')
A3=table.groupby(['辦事員']).mean()['作業天數'].rename('平均作業天數')
A4=table.groupby(['辦事員']).mean()['追加比例'].rename('平均追加比例')
A5=table.groupby(['辦事員']).sum()['超預算'].rename('超預算件數')

A6=table.groupby(['辦事員']).sum()['獨家報價'].rename('獨家報價件數')

辦事員	決包金額最小值	決包金額最大值	平均作業天數	平均追加比例	超預算件數	獨家報價件數
N11183	198000	198000	145	0.761626	1	0
N13132	102000	200000	39.905	1.03282	35	75
N13145	101000	200000	37.2754	1.01934	22	23
N14157	102000	200000	38.4034	0.996237	24	49
N18197	102000	200000	42.844	1.00264	42	50
N21223	100500	200000	25.4255	1.0103	31	70
N22222	105000	200000	30.2338	1.00955	19	30
N22238	100950	200000	30.044	1.10093	14	46
N25251	100350	200000	44.3261	1.01583	26	63
N33128	102864	199000	25.1579	1	4	27

b1 = pd.merge(A1,A2,how='right',on='辦事員')
b2 = pd.merge(b1,A3,how='right',on='辦事員')
b3 = pd.merge(b2,A4,how='right',on='辦事員')
b4 = pd.merge(b3,A5,how='right',on='辦事員')
b5 = pd.merge(b4,A6,how='right',on='辦事員')
newdata1 = b5[(b5['平均作業天數']>15)&(b5['超預算件數']>3)&(b5['獨家報價件數']>3)]

辦事員	決包金額最小值	決包金額最大值	平均作業天數	平均追加比例	超預算件數	獨家報價件數
N13132	102000	200000	39.905	1.03282	35	75
N13145	101000	200000	37.2754	1.01934	22	23
N14157	102000	200000	38.4034	0.996237	24	49
N18197	102000	200000	42.844	1.00264	42	50
N21223	100500	200000	25.4255	1.0103	31	70
N22222	105000	200000	30.2338	1.00955	19	30
N22238	100950	200000	30.044	1.10093	14	46
N25251	100350	200000	44.3261	1.01583	26	63
N33128	102864	199000	25.1579	1	4	27
N39393	100485	199992	40.7634	1.00039	19	58
N41419	102000	200000	43.0513	1.03142	14	58
N50518	100216	199500	36.902	1.04663	27	43
N56561	103087	190000	50.3043	1.02821	6	9
N58581	102000	200000	36.12	1.06158	17	55

#6 計算每位"辦事員"與"施工廠商"的
1.決包金額最小值
2.決包金額最大值
3.平均作業天數
4.平均追加比例
5.超預算件數
6.獨家報價件數
先剔除呈核日、收件日的遺漏值列，將滿足"平均作業天數">30、"平均追加比例">0.2、"獨家報價件數">3"辦事員"與"施工廠商"的明細列出來

C1=table.groupby(['辦事員','施工廠商']).min()['決包金額'].rename('決包金額最小值')
C2=table.groupby(['辦事員','施工廠商']).max()['決包金額'].rename('決包金額最大值')
C3=table.groupby(['辦事員','施工廠商']).mean()['作業天數'].rename('平均作業天數')
C4=table.groupby(['辦事員','施工廠商']).mean()['追加比例'].rename('平均追加比例')
C5=table.groupby(['辦事員','施工廠商']).sum()['超預算'].rename('超預算件數')
C6=table.groupby(['辦事員','施工廠商']).sum()['獨家報價'].rename('獨家報價件數')

辦事員	施工廠商	決包金額最小值	決包金額最大值	平均作業天數	平均追加比例	超預算件數	獨家報價件數
N11183	F3107529	198000	198000	145	0.761626	1	0
N13132	F1032978	151455	151455	76	1	0	0
N13132	F1043252	159200	159200	21	1	0	1
N13132	F108736	165070	165070	63	1	0	0
N13132	F1102990	172615	172615	82	1	0	0
N13132	F1106416	159000	159000	22	1	0	1
N13132	F1115666	153000	198500	57	2.36057	0	1
N13132	F1122054	133000	133000	41	1	0	1
N13132	F1123150	103000	103000	34	1	0	0
N13132	F1143130	110000	150000	91.5	1.06535	1	1
N13132	F1145100	146000	146000	28	1.07383	0	1

d1= pd.concat([C1,C2,C3,C4,C5,C6],axis=1)
newdata2= d1[(d1['平均作業天數']>30)&(d1['平均追加比例']>0.2)&(d1['獨家報價件數']>3)]

辦事員	施工廠商	決包金額最小值	決包金額最大值	平均作業天數	平均追加比例	超預算件數	獨家報價件數
N18197	F8100140	108760	185000	48	1	1	4
N39393	F8139581	101620	161950	36.75	1.03362	1	4
N69701	F1101120	102000	174000	55.3333	0.986385	5	4
N88938	F1103961	101000	160700	50	1	2	6
N88938	F1123100	108000	190000	42.1429	1	1	6
N96971	F1103961	115617	189000	50.75	1	1	4
N96971	F1123100	122000	170000	38	1	0	4
N96971	F2127120	124000	155000	71.75	1	0	4