Data Science HW2

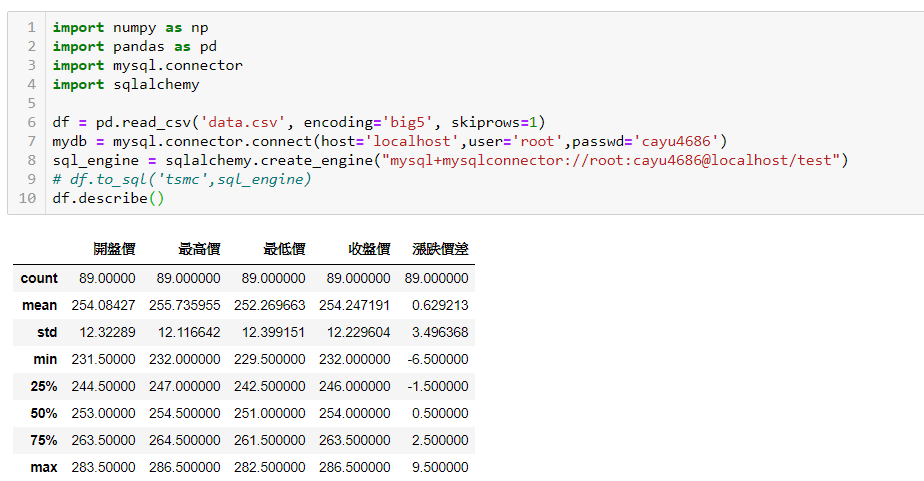
R07921010 余奇安

Problem1:

1.

Use panda to read data from csv and use DataFram to describe the data.

df.describe(), df.head()



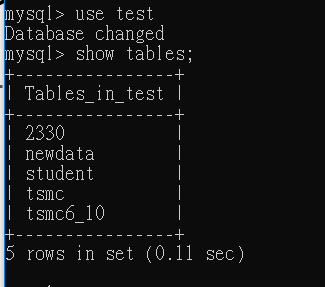
The columns that do not show up such as 成交股數,成交金額 is because the data type is string. As the TA mention, we do not need further process for these data.



2.

I create database “test” first, and use sqlalchemy.creat\_engine to connect to it.

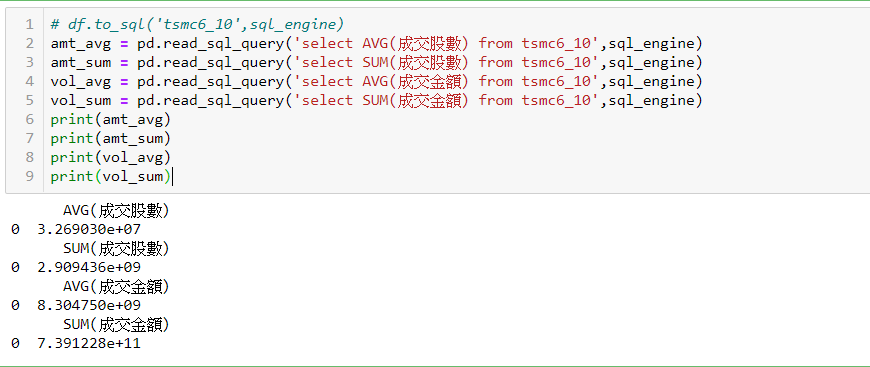
After the execute “df.to\_sql(‘tsmc’, sql\_engine)”, the dataframe store to sql with the table tsmc.



The table in test dataset which open in SQL client shell.

Problem2:

1.



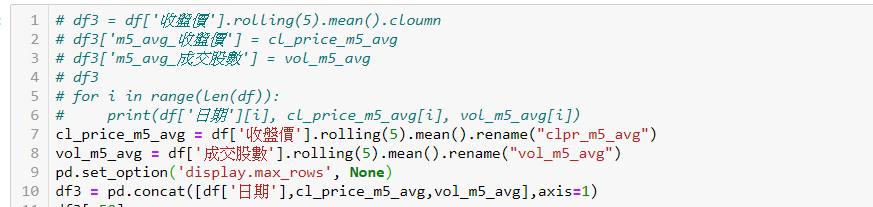
成交股數平均:32690295.5843

成交股數總數:2909436307

成交金額平均:8304750158.4831

成交金額總數:739122764105

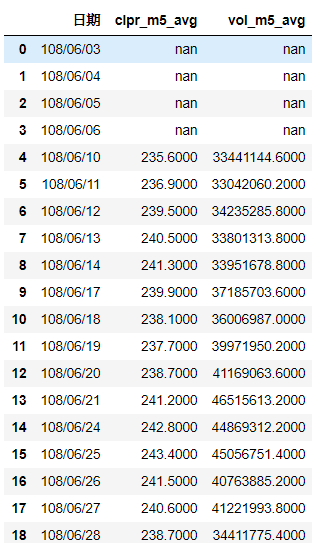
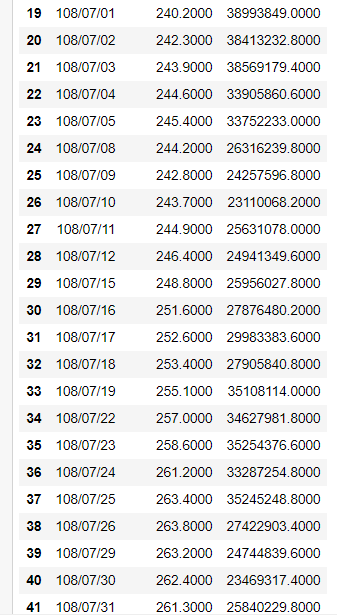
2. 5-day moving average of daily closing price and trading volume.

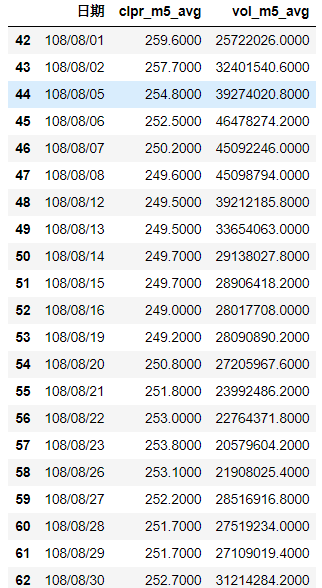


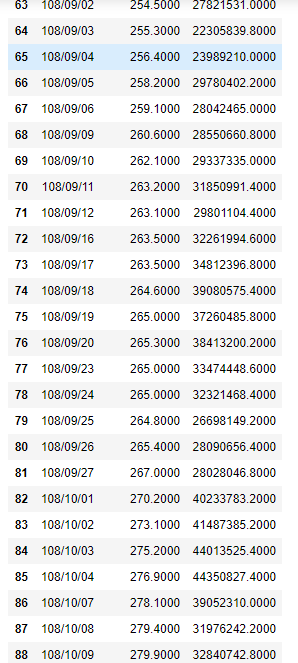
I create another data frame df3, with column name [日期, clpar\_m5\_avg,

vol\_m5\_avg]

all the table value are show as followed.

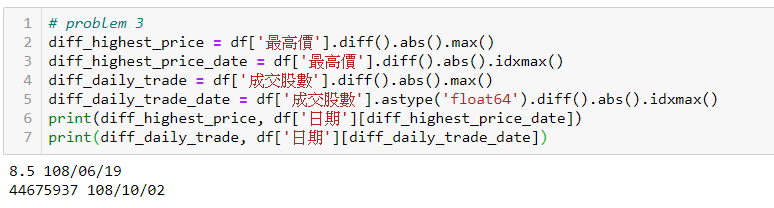






Problem3:

1.

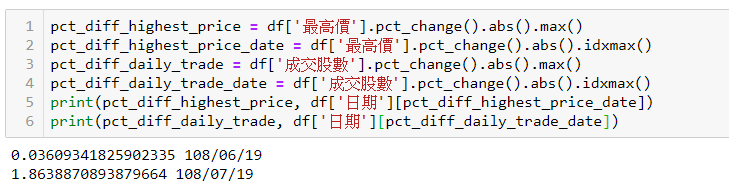


By absolute difference:

I use df.diff().abs().max()

|  |  |  |
| --- | --- | --- |
|  | 變化量 | 日期 |
| Highest daily price: | 8.5 | 108/06/19 |
| Daily trade volume | 44675937 | 108/10/02 |

2.



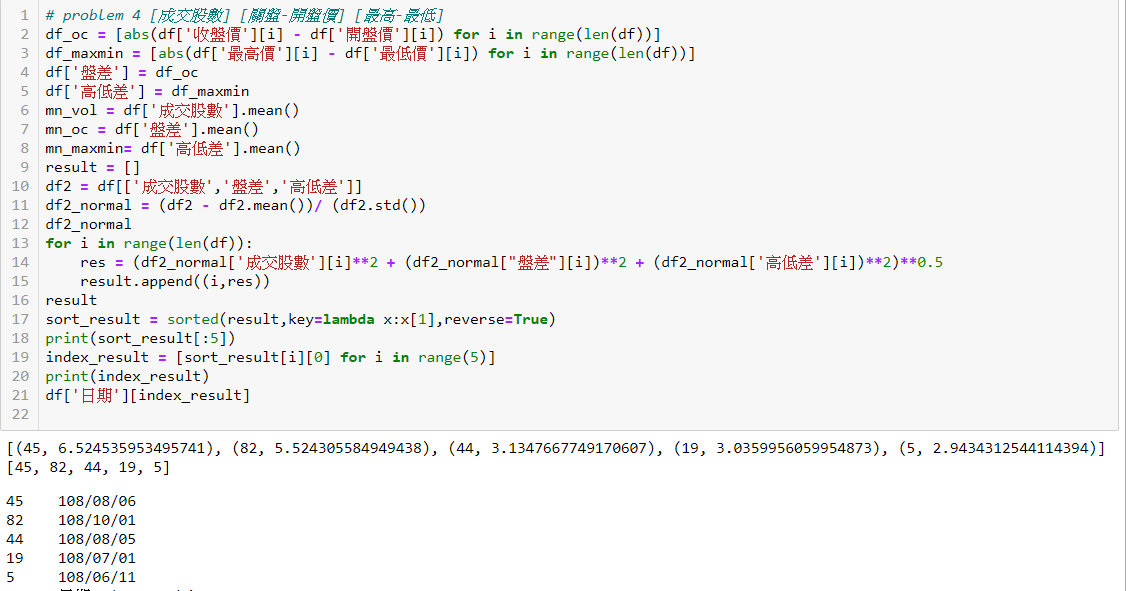
By percentage:

I use df.pct\_change().abs().max()

|  |  |  |
| --- | --- | --- |
|  | 變化量 | 日期 |
| Highest daily price: | 0.03609 | 108/06/19 |
| Daily trade volume | 1.86388 | 108/07/19 |

Problem 4

Outlier



I take absolute values for features.

The outlier days are:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 108/08/06 | 108/10/01 | 108/08/05 | 108/07/01 | 108/06/11 |
| values | 6.5245 | 5.243 | 3.134 | 3.0359 | 2.9434 |

Problem5

By brute force, traverse all possible solution

