问题2

In [1]:

```python
import numpy as np
import pandas as pd
import cufflinks as cf

import scipy
import scipy.cluster.hierarchy as sch

from sklearn.metrics import *
from sklearn.cluster import DBSCAN

import plotly
import plotly.express as px
import plotly.graph_objects as go
import plotly.figure_factory as ff

import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False

from IPython.display import HTML
from IPython.core.interactiveshell import InteractiveShell
# InteractiveShell.ast_node_interactivity = 'all'
InteractiveShell.ast_node_interactivity = 'last'

import pylatex
import latexify
```

层次聚类

DMA 1 日期 漏水聚类 (unfin，最终未使用)

In [2]:

```python
# DMA1 data
user_DMA1 = pd.read_excel("按照日期处理后的数据.xlsx", sheet_name='DMA1的用户用水量', index_col=0)
user_DMA1 = pd.concat([user_DMA1.iloc[:43, :], user_DMA1.iloc[44:, :]])
index = list(user_DMA1.index.strftime("%Y-%m-%d"))
columns = list(user_DMA1.columns)
```

```python
# distance matrix
n, m = user_DMA1.shape
dismat = []
for i in range(n):
    dis = []
    for j in range(n):
        d = ((user_DMA1.iloc[i, :] - user_DMA1.iloc[j, :])**2).sum()**0.5
        dis.append(d)
    dismat.append(dis)
pd.DataFrame(dismat, index=index, columns=index).head(10)
```

Out[2]:

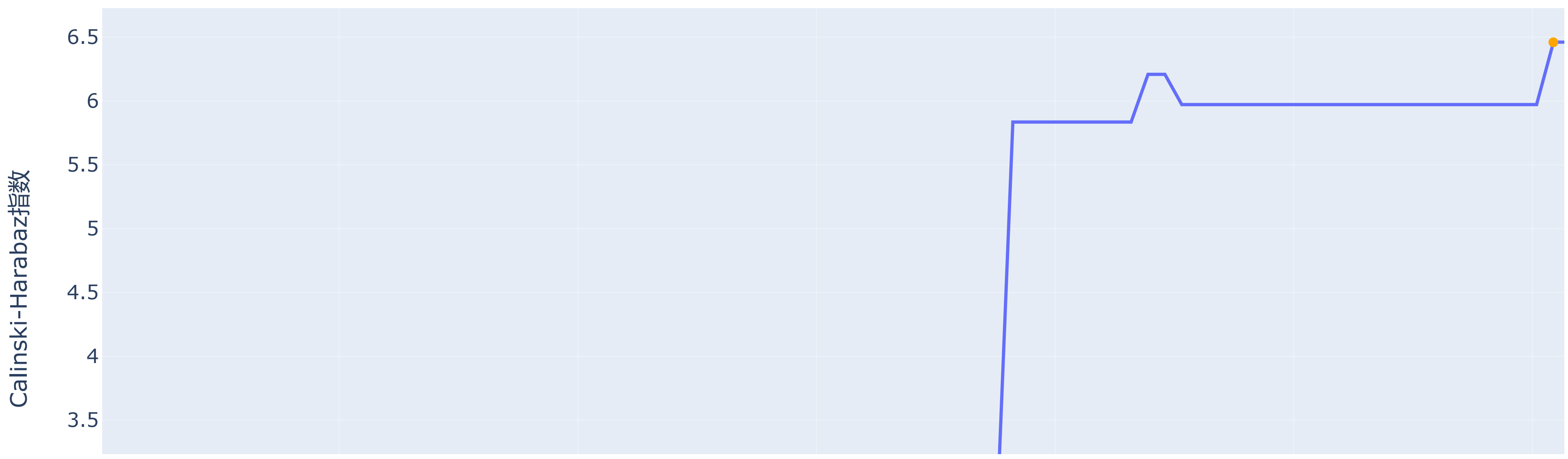
	2014-04-15	2014-04-16	2014-04-17	2014-04-18	2014-04-19	2014-04-20	2014-04-21	2014-04-22	2014-04-23	2014-04-24	...	2014-06-02	2014-06-03	2014-06-04	2014
2014-04-15	0.000000	44.970736	43.790481	79.645601	57.201260	44.092741	30.180593	37.291196	34.615235	34.022864	...	118.995703	102.396775	127.063951	126.82
2014-04-16	44.970736	0.000000	33.023548	55.413570	37.121144	36.250196	33.376436	35.664262	35.836445	39.711072	...	92.234241	77.074228	105.571553	100.81
2014-04-17	43.790481	33.023548	0.000000	68.089377	47.576177	44.180373	37.139529	32.447618	35.771979	43.281593	...	100.635740	87.967622	115.433542	111.88
2014-04-18	79.645601	55.413570	68.089377	0.000000	54.301897	68.279117	71.790516	72.613009	71.801586	70.443402	...	76.492624	65.070479	73.933231	73.58
2014-04-19	57.201260	37.121144	47.576177	54.301897	0.000000	48.539913	45.224504	45.056550	47.722960	49.241252	...	89.843496	76.965555	100.527011	98.64
2014-04-20	44.092741	36.250196	44.180373	68.279117	48.539913	0.000000	35.322639	42.349185	40.425280	47.044807	...	107.103153	96.373283	121.390342	120.29
2014-04-21	30.180593	33.376436	37.139529	71.790516	45.224504	35.322639	0.000000	27.670524	26.865951	33.109831	...	111.999281	95.362131	124.815657	121.48
2014-04-22	37.291196	35.664262	32.447618	72.613009	45.056550	42.349185	27.670524	0.000000	30.816616	38.591046	...	108.408107	90.196483	121.974247	118.35
2014-04-23	34.615235	35.836445	35.771979	71.801586	47.722960	40.425280	26.865951	30.816616	0.000000	35.298564	...	113.966121	94.098119	124.265710	119.68
2014-04-24	34.022864	39.711072	43.281593	70.443402	49.241252	47.044807	33.109831	38.591046	35.298564	0.000000	...	114.483158	94.866750	120.205077	117.14

10 rows × 57 columns

In [ ]:

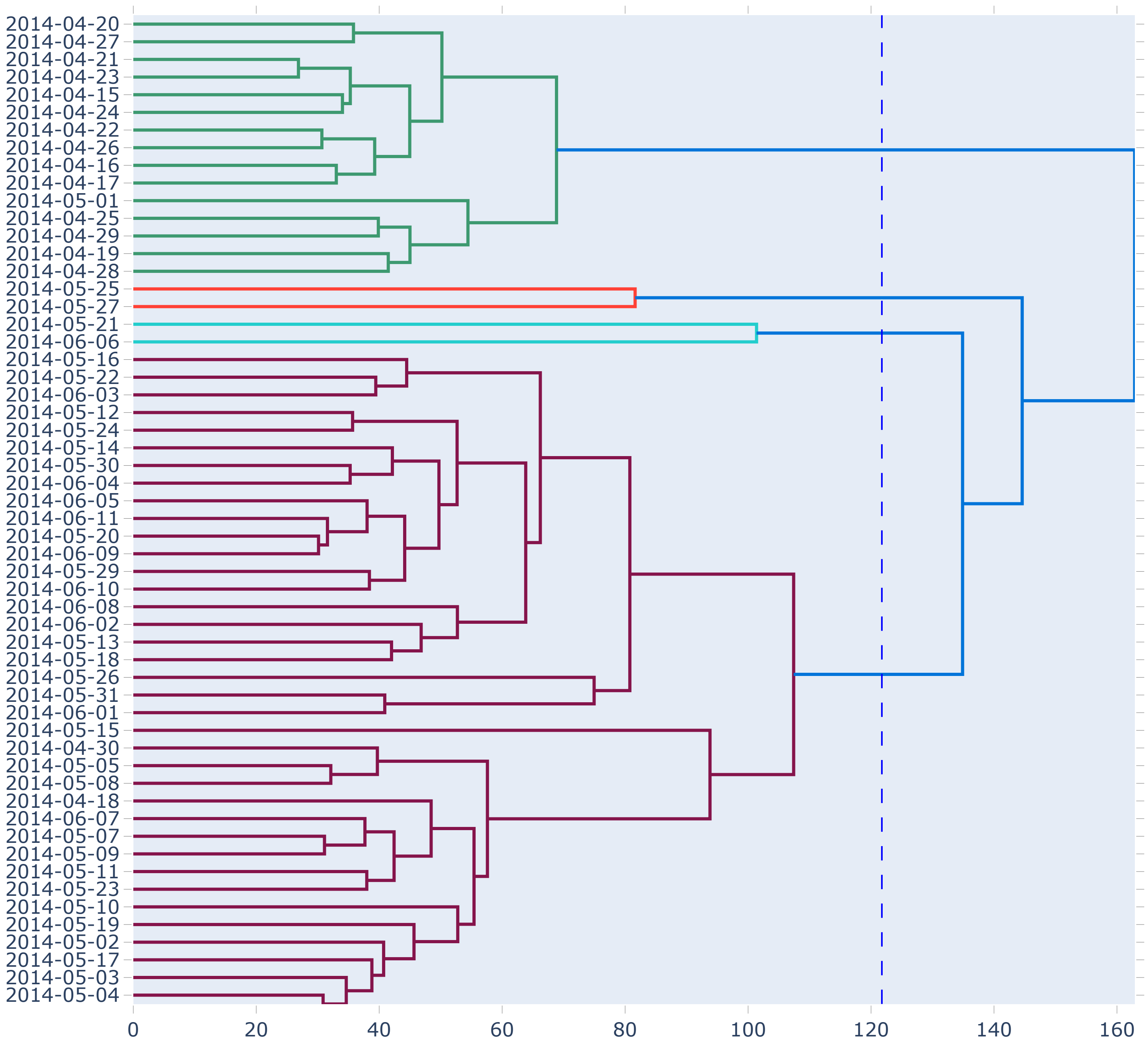
In [3]:

```
InteractiveShell.ast_node_interactivity = 'last'

dis_arr = np.array(user_DMA1)
disMat = sch.distance.pdist(dis_arr, 'euclidean')
Z = sch.linkage(disMat)
ch_score = []
b = 1.14
t = np.linspace(0, b, int(100*(b)+1))
tt = np.linspace(0, 160, int(100*(b)+1))
for d in t:
    cluster = sch.fcluster(Z, d, 'inconsistent')
    s = calinski_harabasz_score(user_DMA1, cluster)
    ch_score.insert(0, s)
    ch_score.insert(0, ch_score[0])
    ch_score.pop()
# len(set(sch.fcluster(Z, 0.88, 'inconsistent')))
trace = go.Scatter(x=tt, y=ch_score, mode='lines', name='CH指数')
fig = go.Figure(data=trace)
fig.update_layout(
    xaxis=dict(title='分类距离阈值'),
    yaxis=dict(title='Calinski-Harabaz指数'),
    title_text="DMA1用水量-Calinski-Harabaz指数随分类距离阈值的变化情况",
)
fig.add_trace(go.Scatter(
    x=[121.76], y=[6.46],
    line=dict(color='orange', width=5),
    showlegend=False,
))
# fig.write_image('./img/svg/DMA1用水量-Calinski-Harabaz指数随分类距离阈值的变化情况.svg')
fig.show()

fig = ff.create_dendrogram(user_DMA1, orientation='left', labels=index, )
fig.update_layout(
    width=800,
    height=800,
    yaxis=dict(range=[-560, 0]),
    title_text='DMA1用水量-对日期的层次聚类树状图',
)
fig.add_trace(go.Scatter(
    x=[121.76] * len(ch_score),
    y=np.linspace(-560, 0, len(ch_score)),
    mode='lines',
    line=dict(color='blue', width=1, dash='dash'),
))
# fig.write_image('./img/svg/DMA1用水量-对日期进行层次聚类结果.svg')
fig.show()
```



DMA1用水量-Calinski-Harabaz指数随分类距离阈值的变化情况

DMA1用水量-对日期的层次聚类树状图



DMA 2 日期 漏水量聚类 (unfin，最终未使用)

In [4]:

```
# DMA2 data
user_DMA2 = pd.read_excel("按照日期处理后的数据.xlsx", sheet_name='DMA2的用户用水量', index_col=0)
user_DMA2 = pd.concat([user_DMA2.iloc[:43, :], user_DMA2.iloc[44:, :]])
index = list(user_DMA2.index.strftime("%Y-%m-%d"))
columns = list(user_DMA2.columns)

# distance matrix
n, m = user_DMA2.shape
dismat = []
for i in range(n):
    dis = []
    for j in range(n):
        d = ((user_DMA2.iloc[i, :] - user_DMA2.iloc[j, :])**2).sum()**0.5
        dis.append(d)
    dismat.append(dis)
pd.DataFrame(dismat, index=index, columns=index).head(10)
```

Out[4]:

	2014-04-15	2014-04-16	2014-04-17	2014-04-18	2014-04-19	2014-04-20	2014-04-21	2014-04-22	2014-04-23	2014-04-24	...	2014-06-02	2014-06-03	2014-06-04	2014-06-05
2014-04-15	0.000000	11.461043	15.414292	16.675269	18.095276	12.817258	13.468589	13.402164	13.389556	16.234152	...	63.800942	66.171191	60.720991	56.660274
2014-04-16	11.461043	0.000000	14.852747	16.464583	18.824067	12.240082	14.245757	12.662831	14.715125	16.558200	...	64.126944	66.156864	60.772986	56.798202
2014-04-17	15.414292	14.852747	0.000000	14.418904	13.875518	20.718226	21.165422	13.932645	12.081126	24.310827	...	55.911595	58.295387	51.880269	48.980828
2014-04-18	16.675269	16.464583	14.418904	0.000000	14.570443	22.013743	21.981545	11.430713	12.348052	24.898219	...	55.764494	57.410643	53.100696	48.286303
2014-04-19	18.095276	18.824067	13.875518	14.570443	0.000000	24.935900	25.216853	16.192591	11.865547	27.574428	...	52.133858	54.378749	48.822121	45.411394
2014-04-20	12.817258	12.240082	20.718226	22.013743	24.935900	0.000000	7.118399	17.520214	19.315742	8.860429	...	71.509645	73.620038	67.989486	64.033330
2014-04-21	13.468589	14.245757	21.165422	21.981545	25.216853	7.118399	0.000000	18.185140	19.397678	8.175842	...	72.337190	74.412456	68.851587	64.817167
2014-04-22	13.402164	12.662831	13.932645	11.430713	16.192591	17.520214	18.185140	0.000000	12.317622	20.672637	...	59.777611	61.722292	56.883225	52.327852
2014-04-23	13.389556	14.715125	12.081126	12.348052	11.865547	19.315742	19.397678	12.317622	0.000000	21.949604	...	55.451366	57.533617	52.406546	48.231103
2014-04-24	16.234152	16.558200	24.310827	24.898219	27.574428	8.860429	8.175842	20.672637	21.949604	0.000000	...	73.331474	75.324648	69.885717	65.747867

10 rows × 57 columns

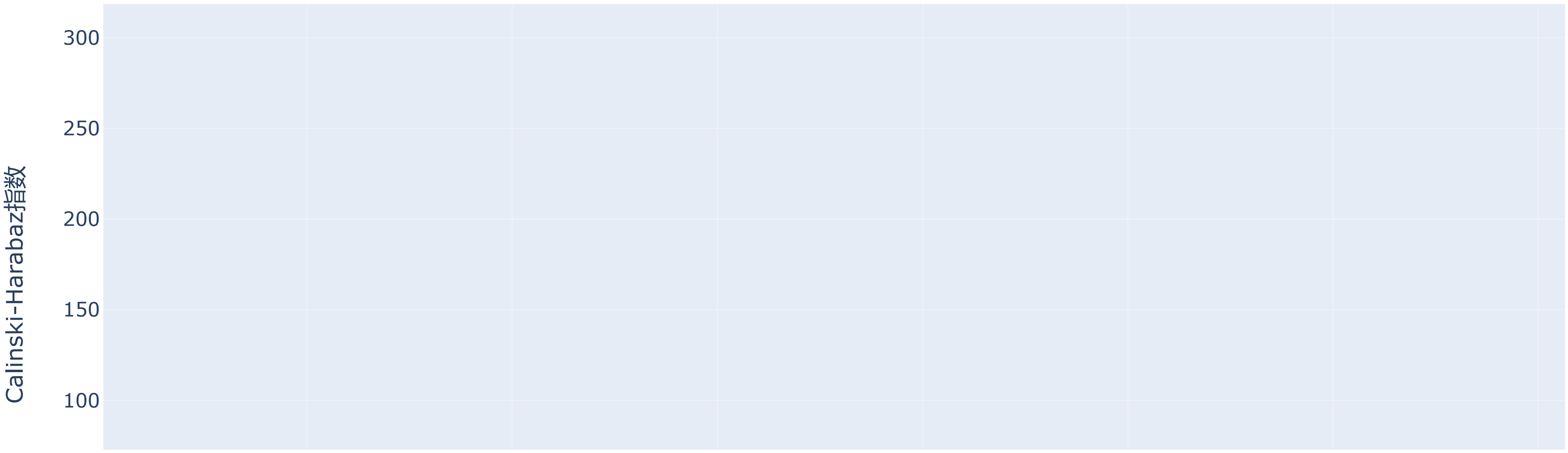
In [ ]:

In [5]:

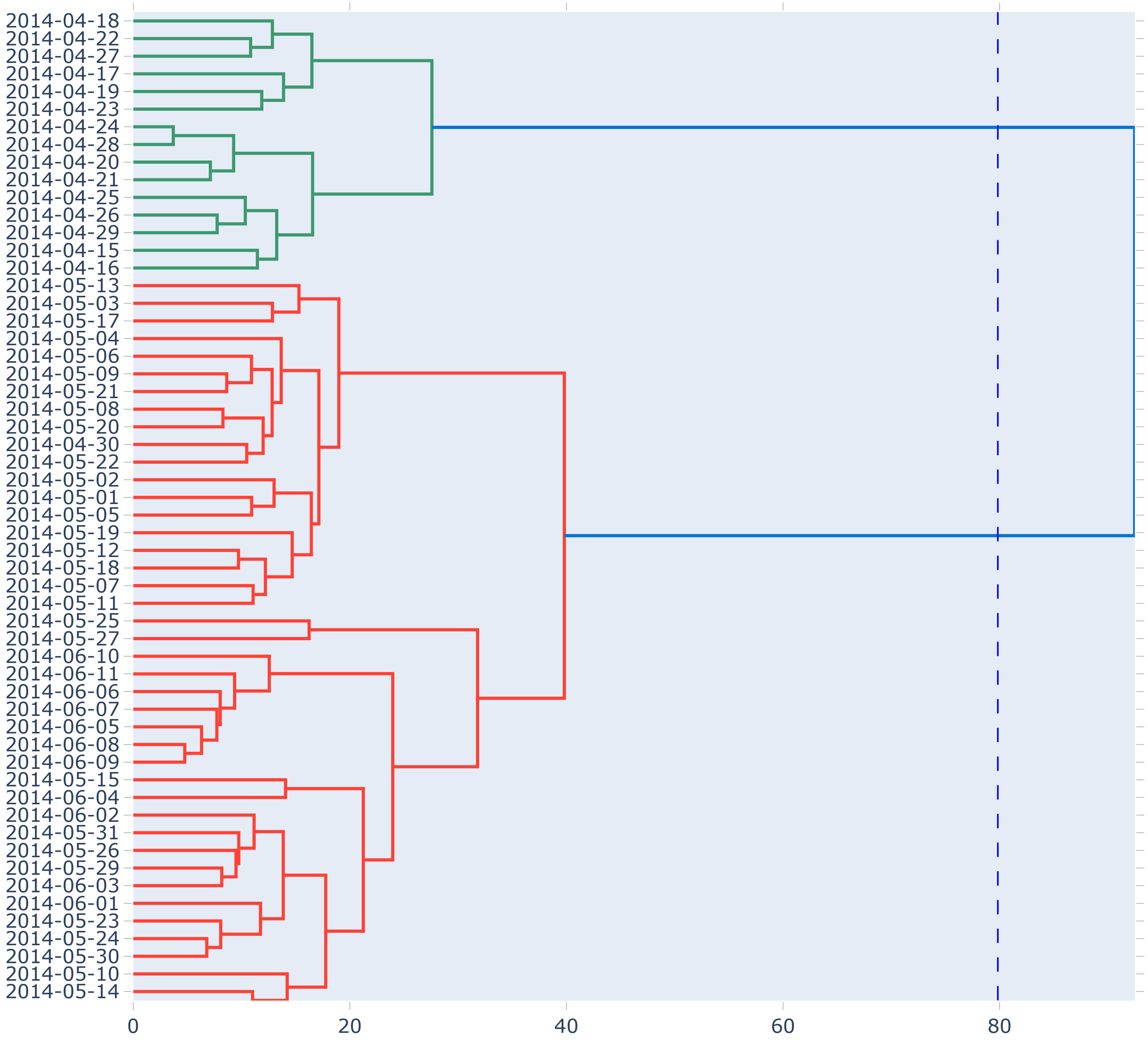
```
InteractiveShell.ast_node_interactivity = 'last'

dis_arr = np.array(user_DMA2)
disMat = sch.distance.pdist(dis_arr, 'euclidean')
Z = sch.linkage(disMat)
# P = sch.dendrogram(Z)
# plt.show()

ch_score = []
b = 1.14
```

```
t = np.linspace(0, b, int(100*(b)+1))
tt = np.linspace(0, 93, int(100*(b)+1))
for d in t:
    cluster = sch.fcluster(Z, d, 'inconsistent')  # 聚类结果
    s = calinski_harabasz_score(user_DMA2, cluster)
    ch_score.append(s)
# len(set(sch.fcluster(Z, 0.97, 'inconsistent')))
trace = go.Scatter(x=tt, y=ch_score, mode='lines', name='CH指数')
fig = go.Figure(data=trace)
fig.update_layout(
    xaxis=dict(title='分类距离阈值'),
    yaxis=dict(title='Calinski-Harabaz指数'),
    title_text="DMA2用水量-Calinski-Harabaz指数随分类距离阈值的变化情况",
)
fig.add_trace(go.Scatter(
    x=[79.83], y=[299.8],
    line=dict(color='orange', width=5),
    showlegend=False,
))
# fig.write_image('./img/svg/DMA2用水量-Calinski-Harabaz指数随分类距离阈值的变化情况.svg')
fig.show()


fig = ff.create_dendrogram(user_DMA2, orientation='left', labels=index)
fig.update_layout(
    width=800,
    height=800,
    yaxis=dict(range=[-560, 0]),
    title_text='DMA2用水量-对日期的层次聚类树状图',
)
fig.add_trace(go.Scatter(
    x=[79.83] * len(ch_score),
    y=np.linspace(-560, 0, len(ch_score)),
    mode='lines',
    line=dict(color='blue', width=1, dash='dash'),
))
# fig.write_image('./img/svg/DMA2用水量-对日期进行层次聚类结果.svg')
fig.show()
```

DMA2用水量-Calinski-Harabaz指数随分类距离阈值的变化情况



DMA2用水量-对日期的层次聚类树状图

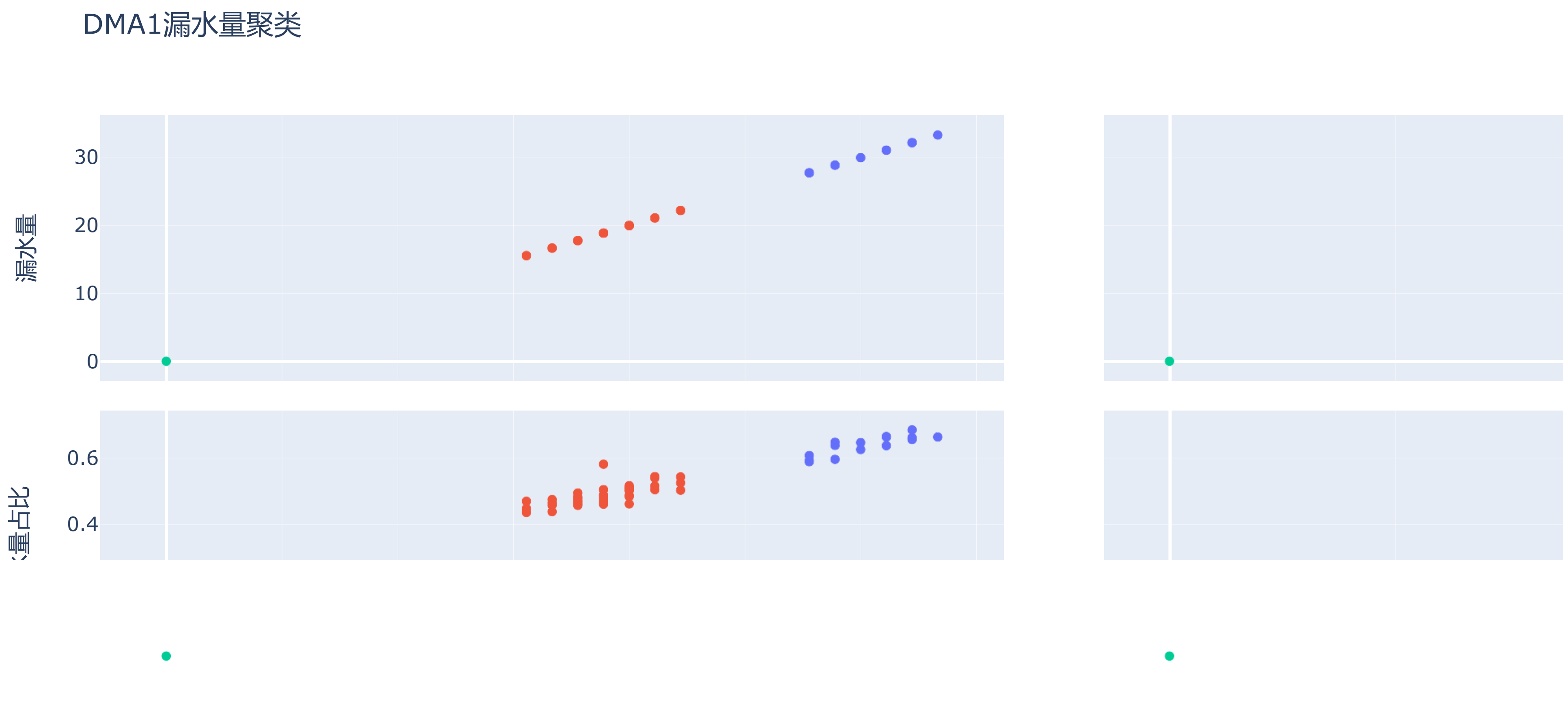


基于密度聚类 DBSCAN

漏水量、漏水量占比聚类

In [6]:

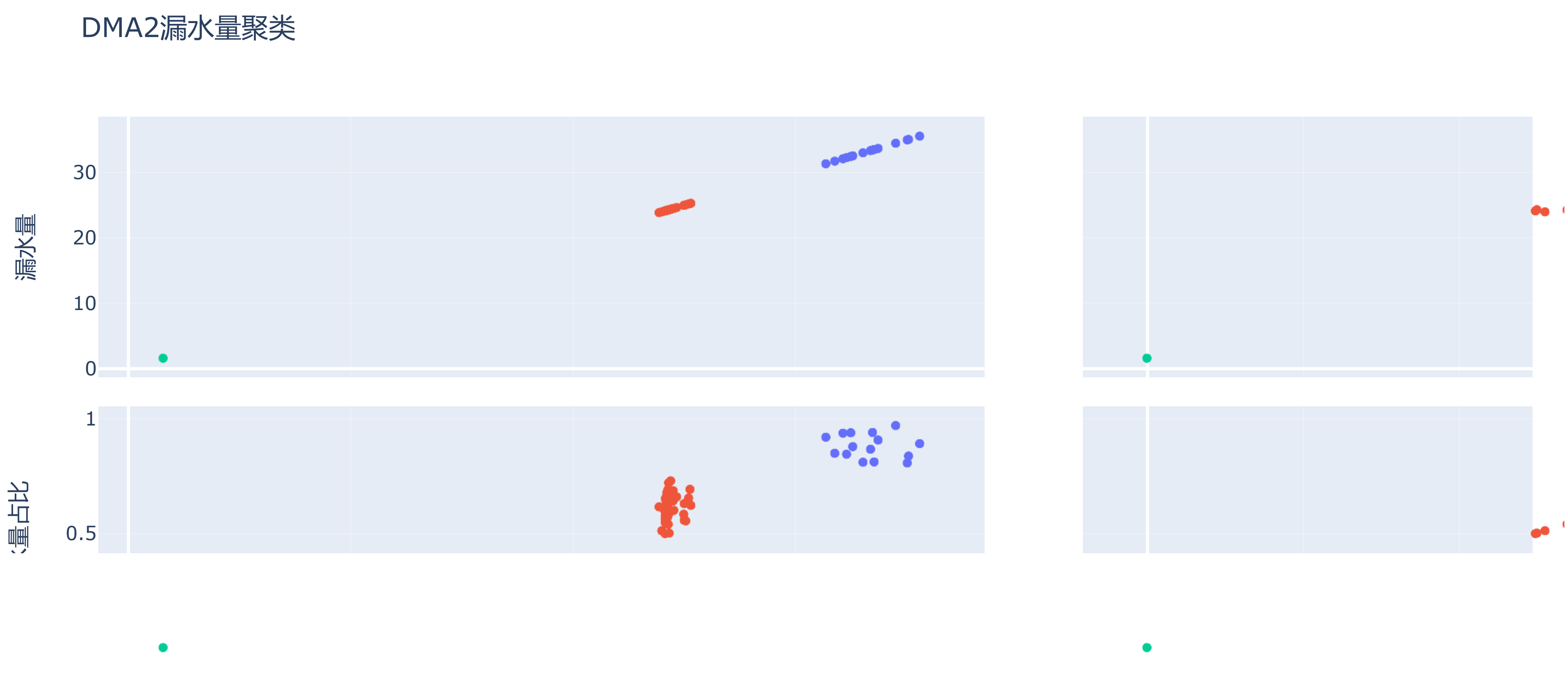
```
# DMA1
# InteractiveShell.ast_node_interactivity = 'all'
InteractiveShell.ast_node_interactivity = 'last'

from sklearn.cluster import DBSCAN

DMA1_leaking = pd.read_excel('./问题2数据.xlsx', sheet_name='DMA1', index_col=0)
index = list(DMA1_leaking.index)
columns = list(DMA1_leaking.columns)
DMA1_leaking

dbscan = DBSCAN(2)
predict = dbscan.fit_predict(DMA1_leaking)
predict

DMA1_leaking['class'] = [f"class{i}" for i in predict]
fig = px.scatter_matrix(
    DMA1_leaking,
    dimensions=["漏水量", "漏水量占比"],
    color='class',
    title='DMA1漏水量聚类',
)
fig.update_layout(legend_title_text='')
fig.write_image('./img/svg/DMA1漏水量聚类.svg')
fig.show()
```



In [7]:

```
# DMA2
# InteractiveShell.ast_node_interactivity = 'all'
InteractiveShell.ast_node_interactivity = 'last'

DMA2_leaking = pd.read_excel('./问题2数据.xlsx', sheet_name='DMA2', index_col=0)
index = list(DMA2_leaking.index)
columns = list(DMA2_leaking.columns)
DMA2_leaking

dbscan = DBSCAN(1.5)
predict = dbscan.fit_predict(DMA2_leaking)
predict
```

```
DMA2_leaking['class'] = [f"class{i}" for i in predict]

fig = px.scatter_matrix(
    DMA2_leaking,
    dimensions=["漏水量", "漏水量占比"],
    color='class',
    title='DMA2漏水量聚类',
)
fig.update_layout(legend_title_text='')
fig.write_image('./img/svg/DMA2漏水量聚类.svg')
fig.show()
```



```
In [ ]:
```