```python
In [1]:
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.rcParams["figure.figsize"] = (25,14)
```

```python
In [3]:
data=pd.read_excel("kmeans_agglom_total.xlsx")
data.head()
```

Out[3]:

	No._Empresas	NIT	PERIODO	Unnamed: 3	Ing_operacio	Utilidad_bruta	Margen_Bruto	Utilidad_operativa	Margen_operativo
0	0	800000268	2016	8000002682016	310130	306053	0.986854	38816	0.125160
1	1	800000268	2017	8000002682017	300000	300000	1.000000	66599	0.221997
2	2	800000268	2018	8000002682018	300000	300000	1.000000	64184	0.213947
3	3	800000268	2019	8000002682019	300000	300000	1.000000	208181	0.693937
4	4	800000276	2016	8000002762016	321430952	46473532	0.144583	9848367	0.030639

```python
In [4]:
data.columns
```

```
Out[4]:
Index(['No._Empresas', 'NIT', 'PERIODO', 'Unnamed: 3', 'Ing_operacio',
       'Utilidad_bruta', 'Margen_Bruto', 'Utilidad_operativa',
       'Margen_operativo', 'Utilidad__Neta', 'Margen_Neto', 'Ebitda',
       'Margen__Ebitda', 'Indice_endeudamiento', 'ROA', 'ROE', 'Cluster',
       'Agglom'],
      dtype='object')
```

```python
In [17]:
data[["NIT","Cluster"]].groupby('Cluster').count().sort_values("NIT",ascending=False)
```

Out[17]:

Cluster	NIT
17	7584
25	7582
42	5022
39	4698
43	4200
11	3169
47	2250
33	1953
49	1529
0	528
35	480
29	386
48	125
21	122
45	119
22	58
37	37
36	29
30	23
44	15
10	12
14	11
13	11
15	9
32	7
27	5
4	4
41	3
38	3
28	3
19	3
23	2
46	2
34	2
7	2
12	2
3	1
40	1
2	1
31	1
5	1
8	1
6	1
26	1
1	1
24	1
20	1
18	1
16	1
9	1

```python
In [5]:
# base de datos con los 12 clúster con mayor cantidad de empresas
data2= data[data.Cluster.isin([17,25,42,39,33,11,47,49,35,0,29,43])]
data2.head()
```

Out[5]:

	No._Empresas	NIT	PERIODO	Unnamed: 3	Ing_operacio	Utilidad_bruta	Margen_Bruto	Utilidad_operativa	Margen_operativo
0	0	800000268	2016	8000002682016	310130	306053	0.986854	38816	0.125160
1	1	800000268	2017	8000002682017	300000	300000	1.000000	66599	0.221997
2	2	800000268	2018	8000002682018	300000	300000	1.000000	64184	0.213947
3	3	800000268	2019	8000002682019	300000	300000	1.000000	208181	0.693937
4	4	800000276	2016	8000002762016	321430952	46473532	0.144583	9848367	0.030639

```python
In [45]:
# base de datos con los 38 cluster restantes con menos empresas
data_otroscluster= data[~data.Cluster.isin([17,25,42,39,33,11,47,49,35,0,29,43])]
data_otroscluster.head()
```

Out[45]:

	No._Empresas	NIT	PERIODO	Unnamed: 3	Ing_operacio	Utilidad_bruta	Margen_Bruto	Utilidad_operativa	Margen_operativo
16	16	800000750	2016	8000007502016	143563931	88936885	0.619493	316104238	2.201836
17	17	800000750	2017	8000007502017	111594829	54644303	0.489667	44133260	0.395478
18	18	800000750	2018	8000007502018	132566524	71012494	0.535674	63952725	0.482420
34	34	800001266	2018	8000012662018	3904073	277657	0.071120	-5268093	-1.349384
276	276	800010972	2016	8000109722016	25272172	-10374876	-0.410526	-16374808	-0.647938

```python
In [6]:
data2.shape
# los 12 cluster principales contienen 39381 datos
```

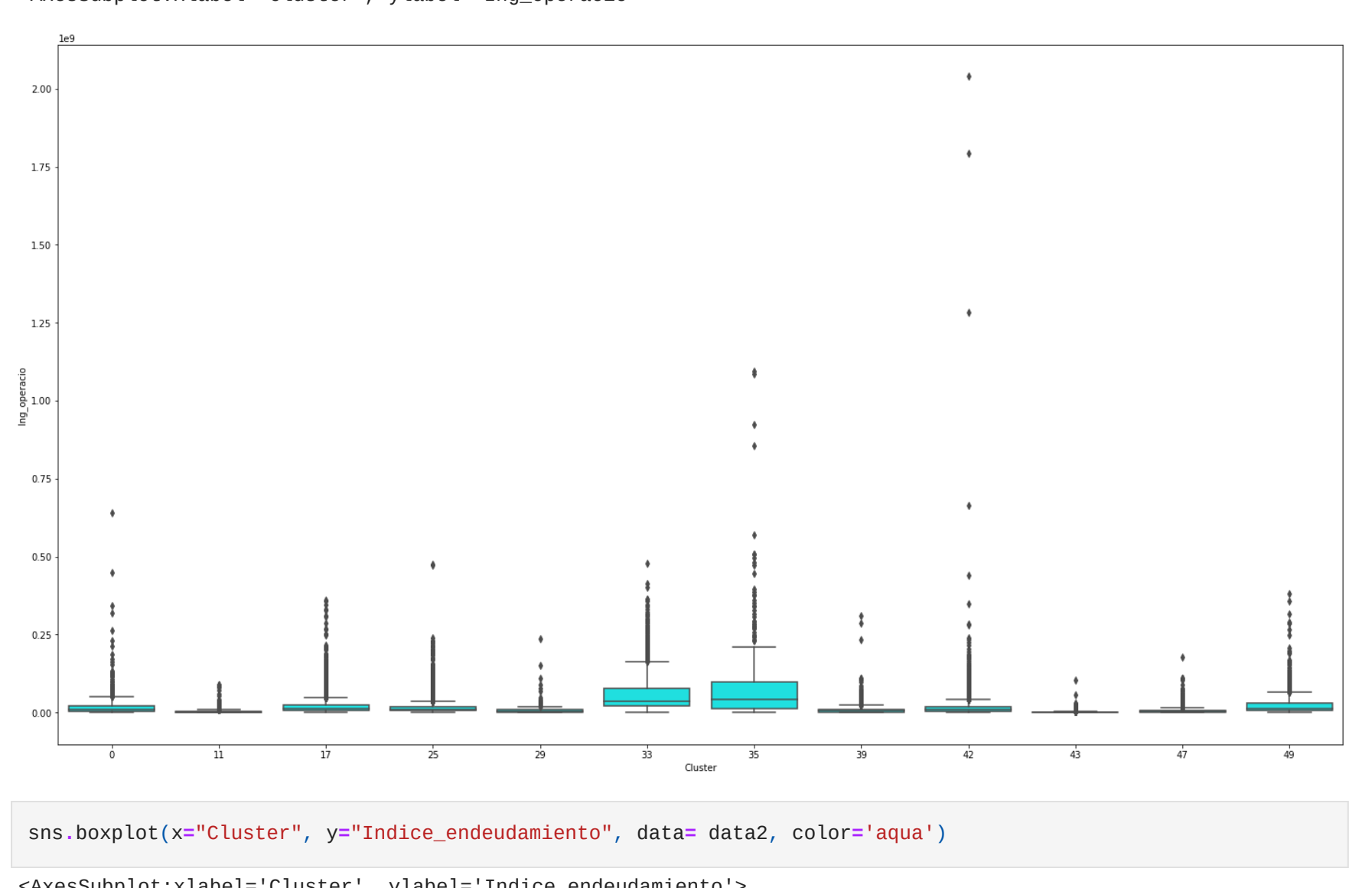
```
Out[6]:
(39381, 18)
```

```python
In [47]:
data_otroscluster.shape
# los restantes 38 cluster contienen 623 datos
```

```
Out[47]:
(623, 18)
```

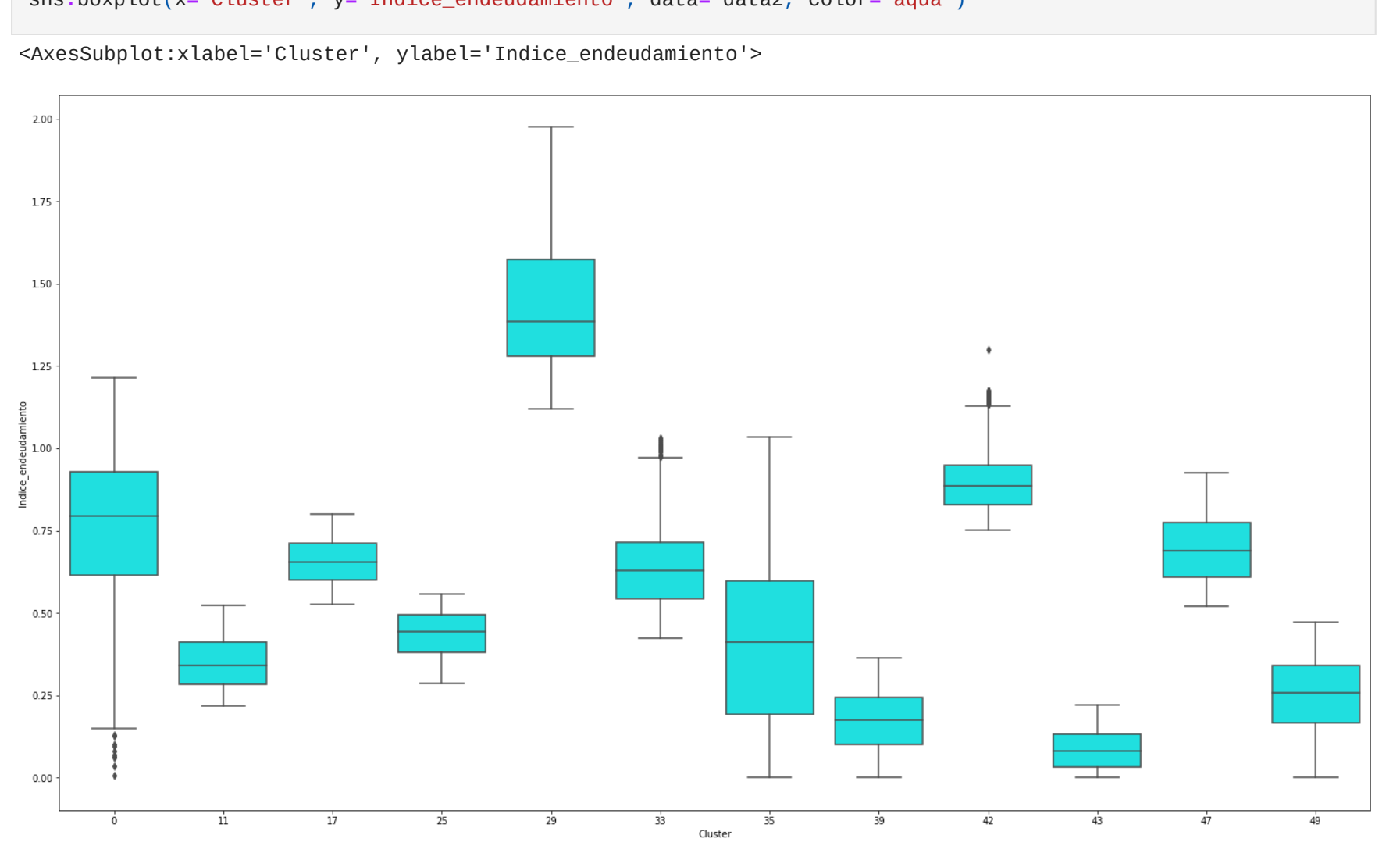
```python
In [8]:
sns.boxplot(x="Cluster", y="Ing_operacio", data= data2, color='aqua')
```

```
Out[8]:
<AxesSubplot:xlabel='Cluster', ylabel='Ing_operacio'>
```




```python
In [9]:
sns.boxplot(x="Cluster", y="Indice_endeudamiento", data= data2, color='aqua')
```

```
Out[9]:
<AxesSubplot:xlabel='Cluster', ylabel='Indice_endeudamiento'>
```

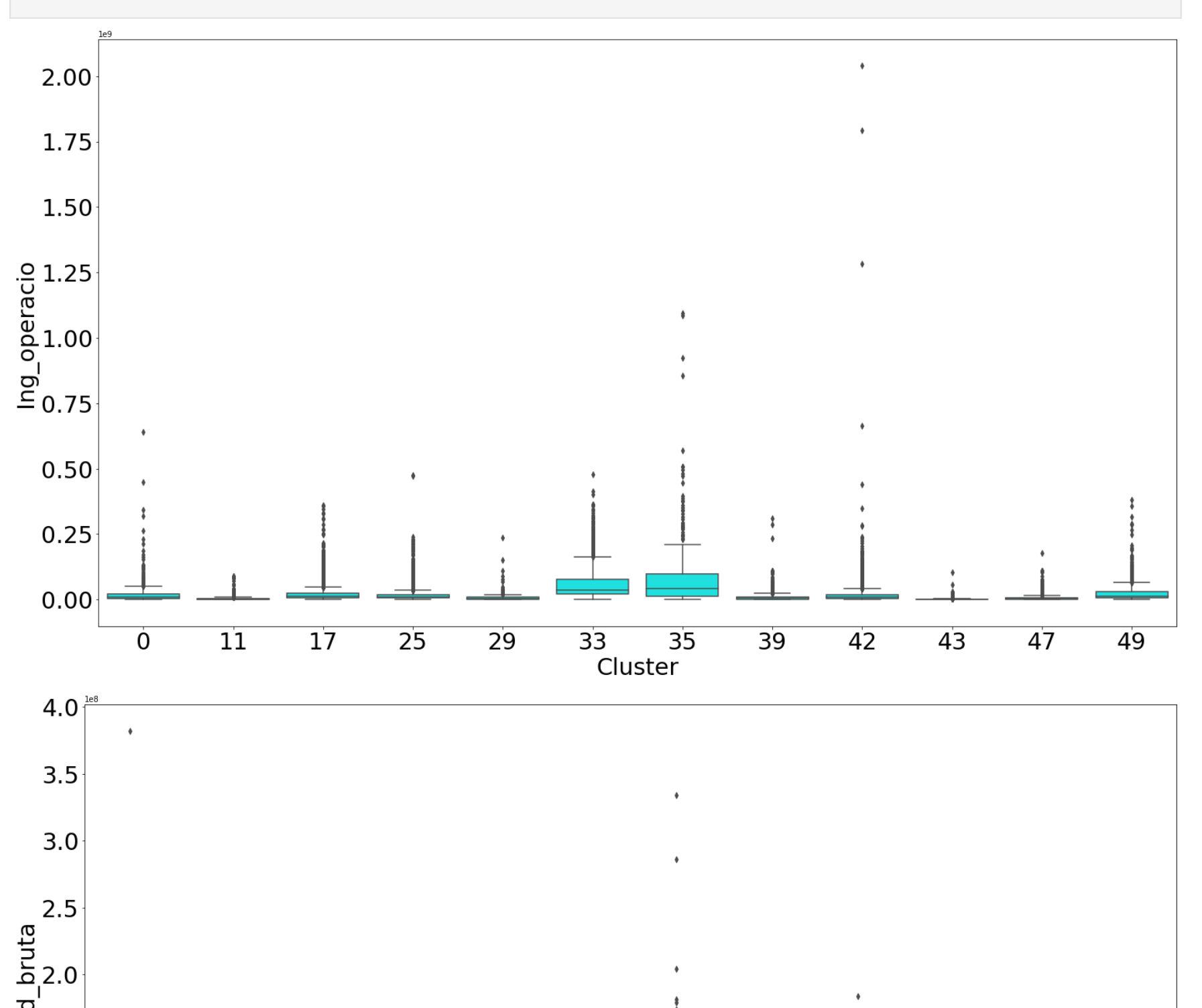


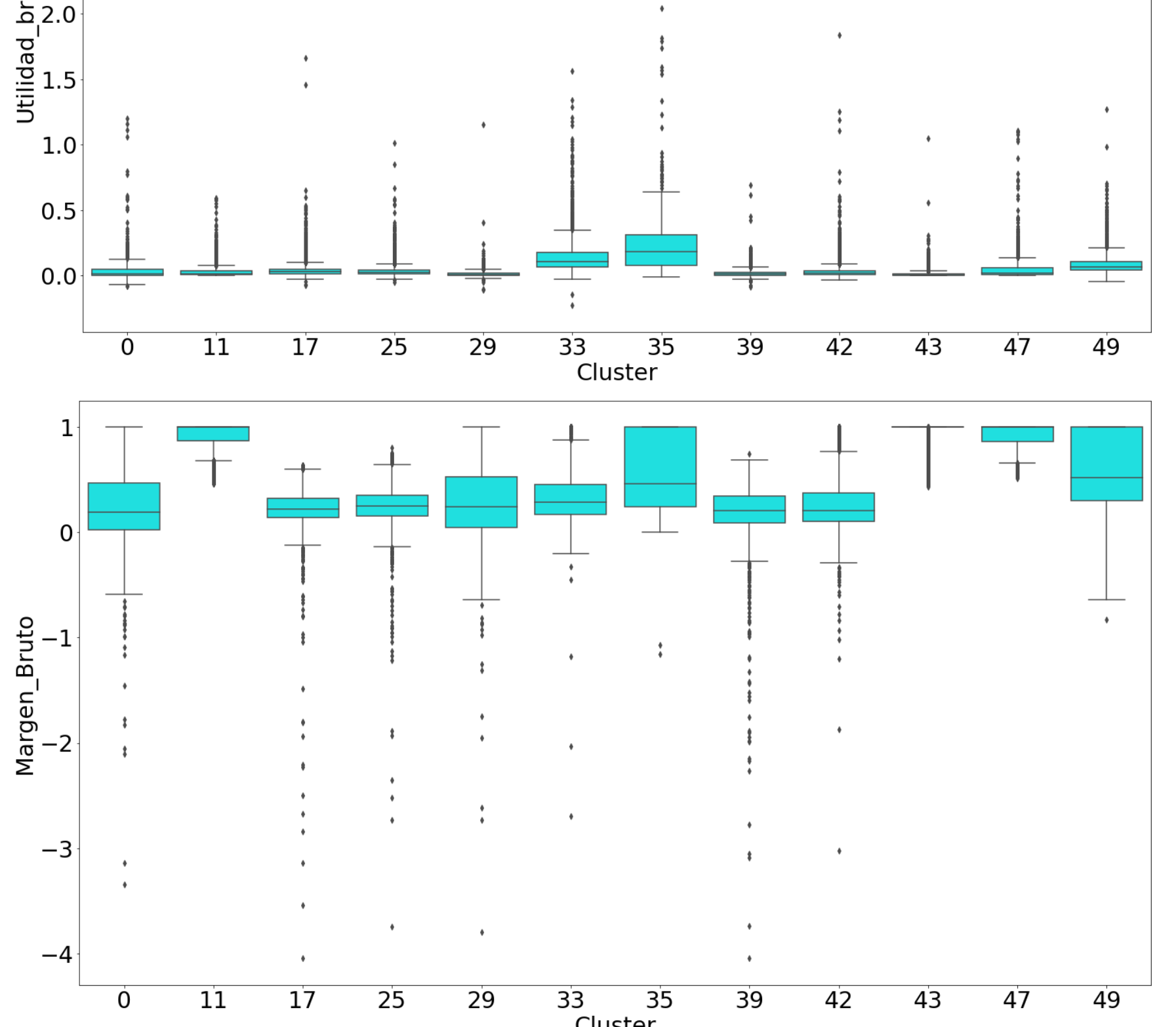

```python
In [10]:
variables=['Ing_operacio', 'Utilidad_bruta',
       'Margen_Bruto', 'Utilidad_operativa', 'Margen_operativo',
       'Utilidad__Neta', 'Margen_Neto', 'Ebitda', 'Margen__Ebitda',
       'Indice_endeudamiento', 'ROA', 'ROE']
```

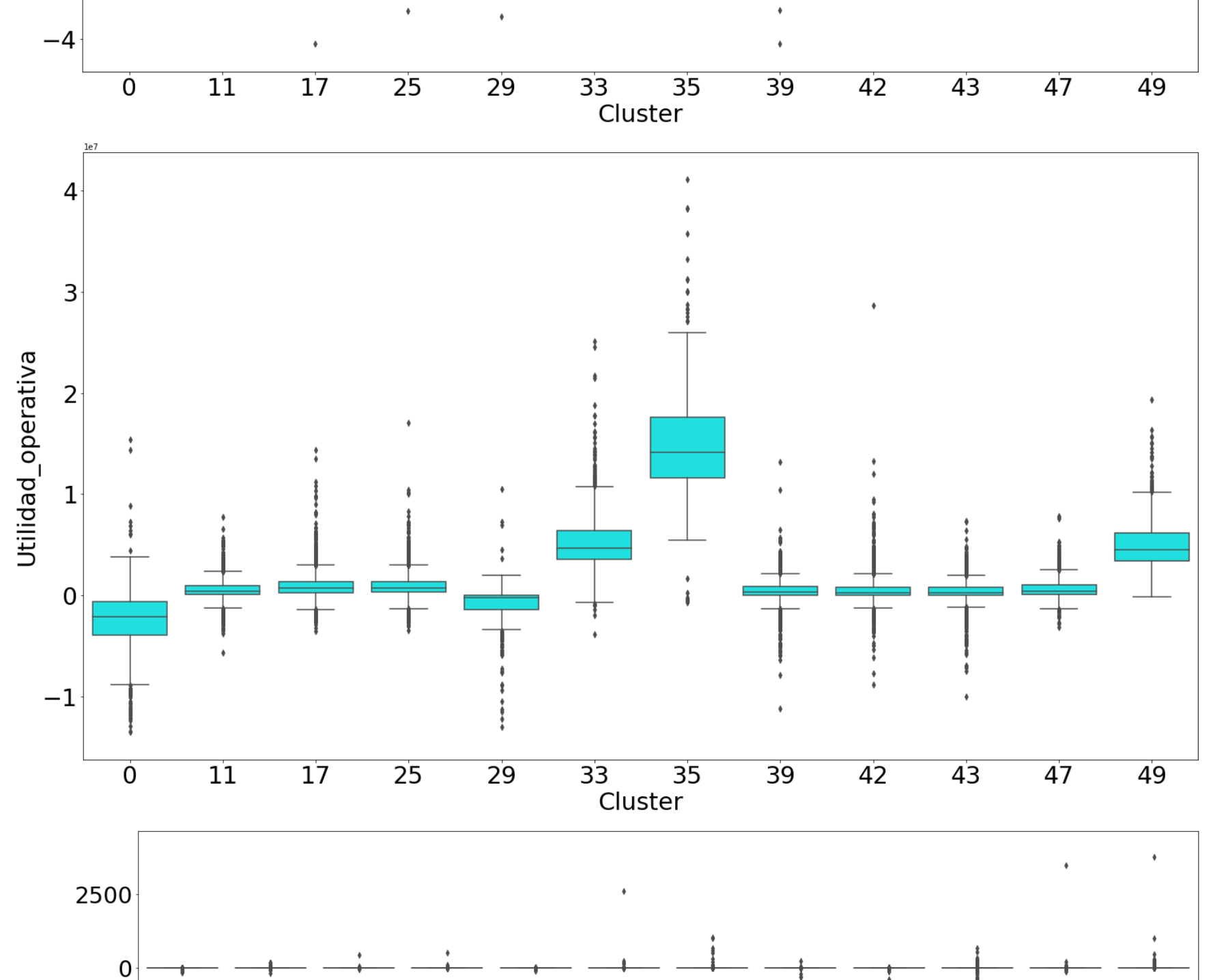
Cluster K-means

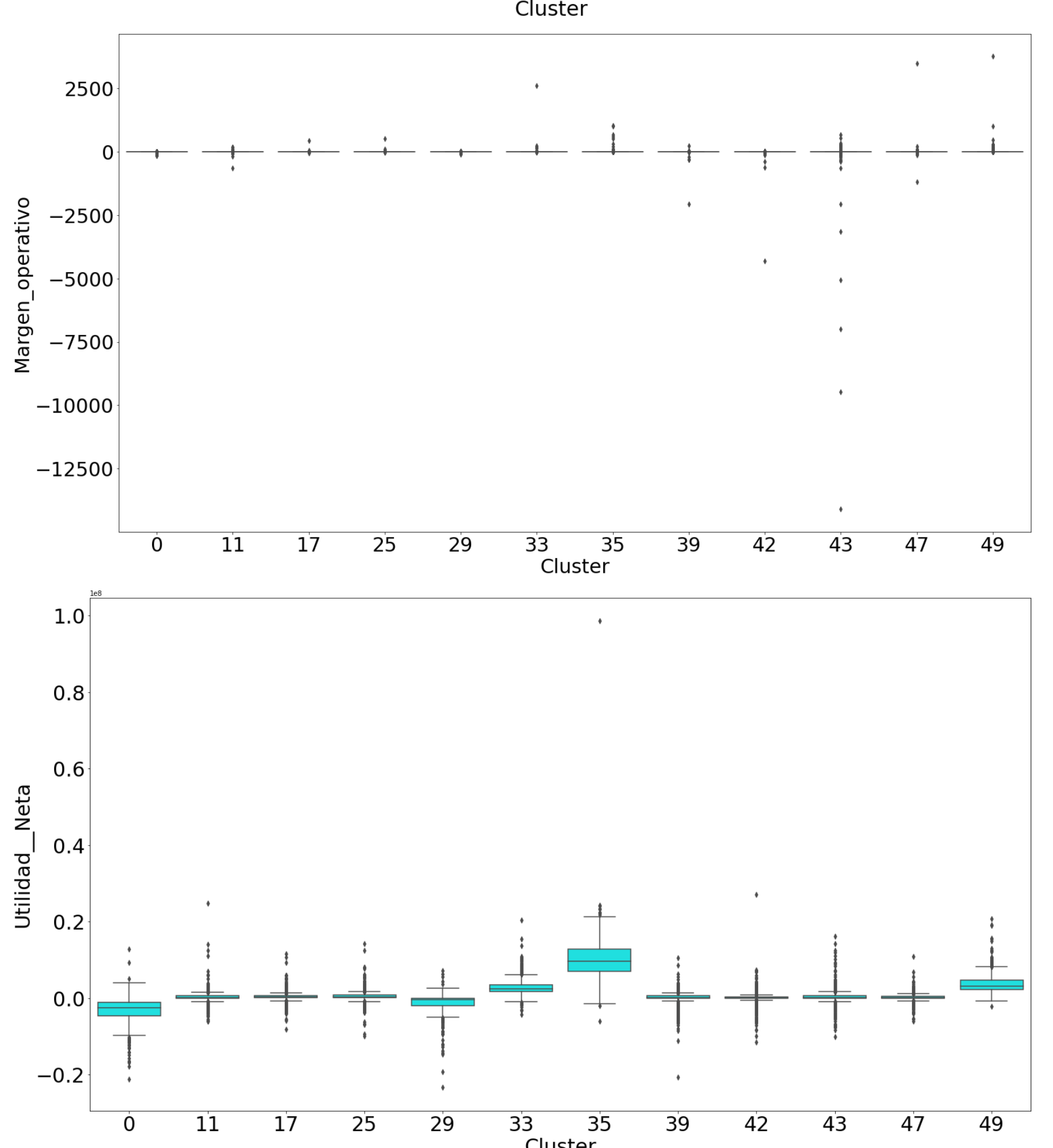
12 Clúster principales K-means

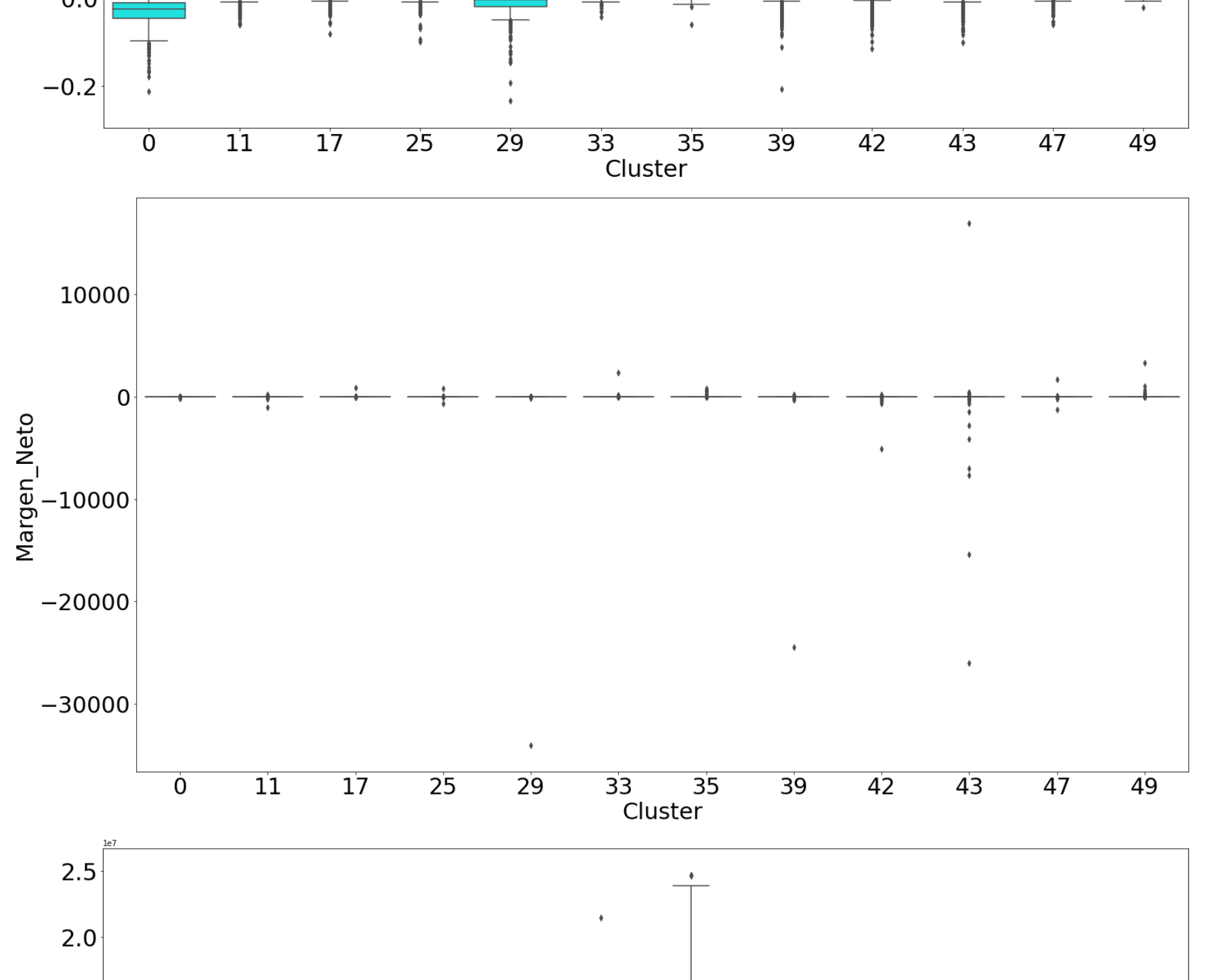
```python
In [44]:
for i in variables:
    plt.xticks(fontsize = 30)
    plt.yticks(fontsize = 30)
    plt.xlabel("",fontsize=30)
    plt.ylabel("",fontsize=30)
    plt.show(sns.boxplot(x="Cluster", y=i, data= data2, color='aqua'))
```

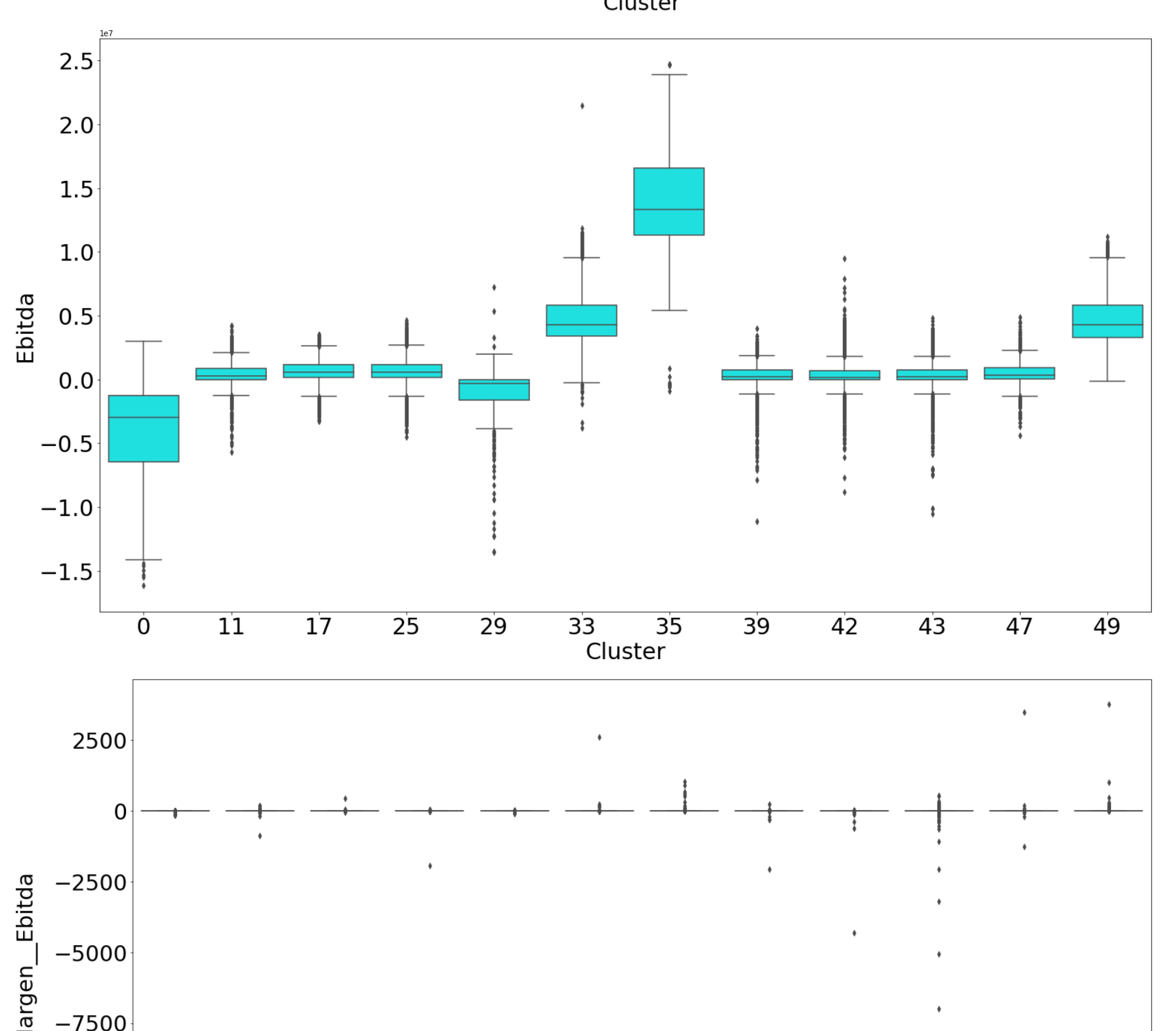


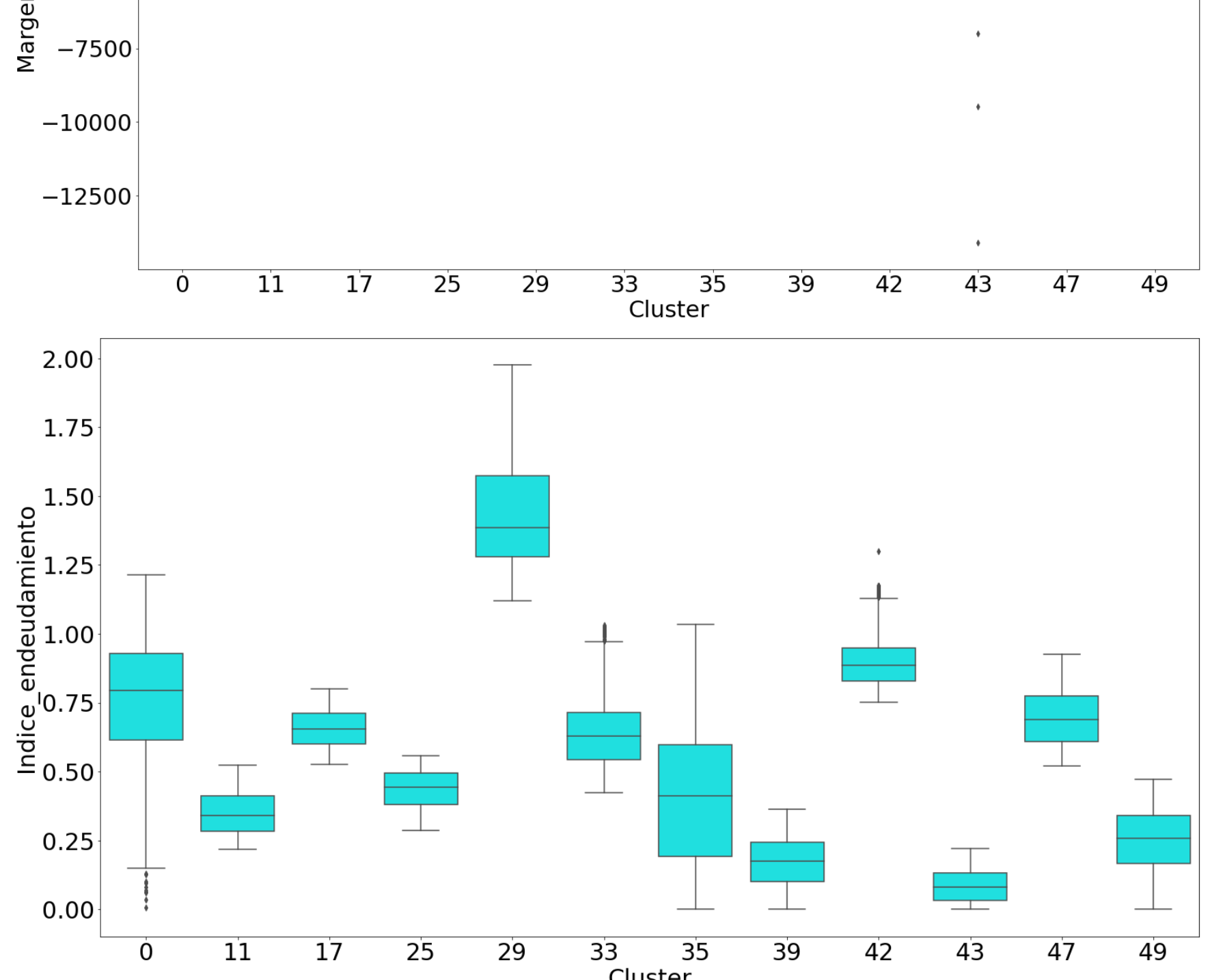


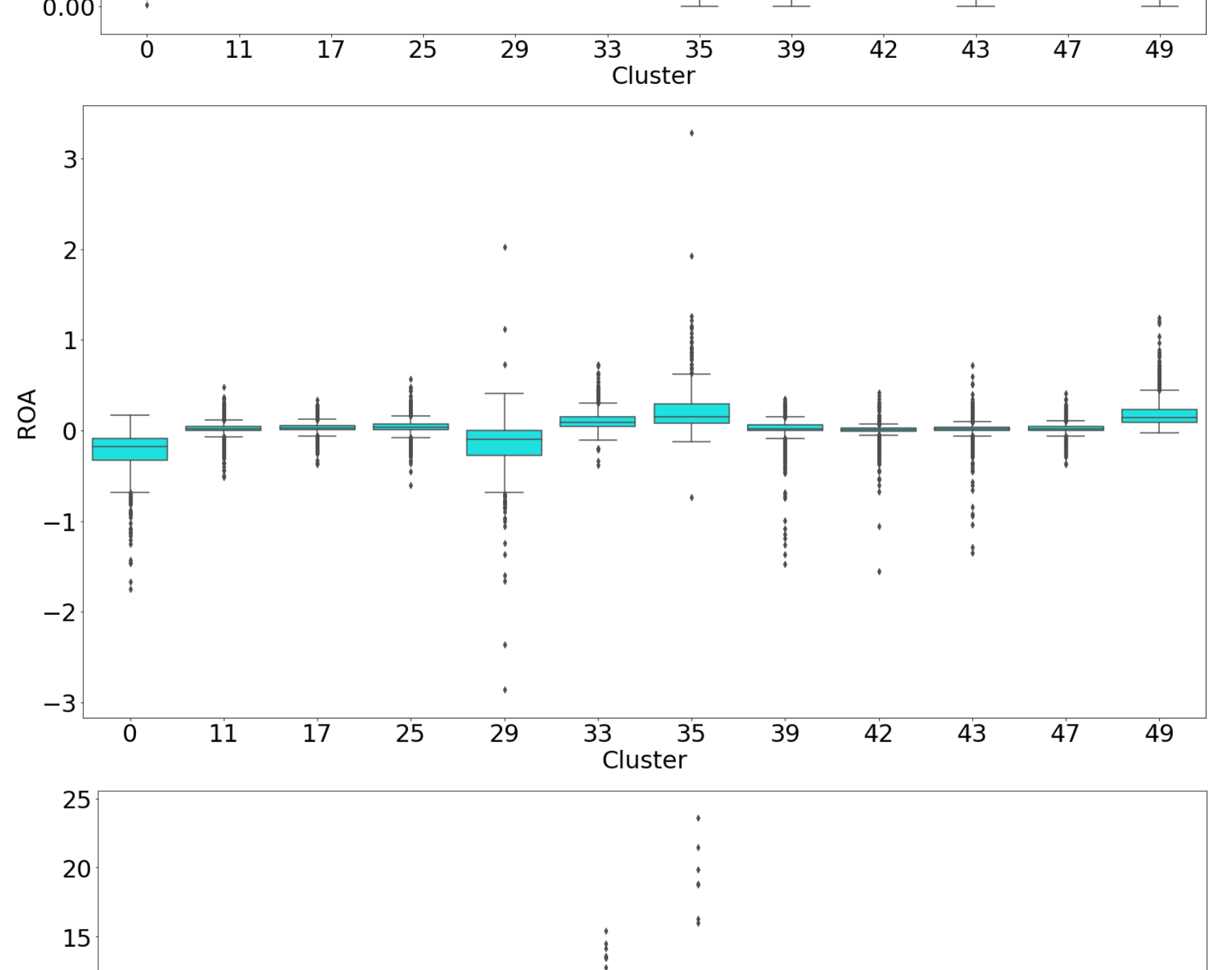


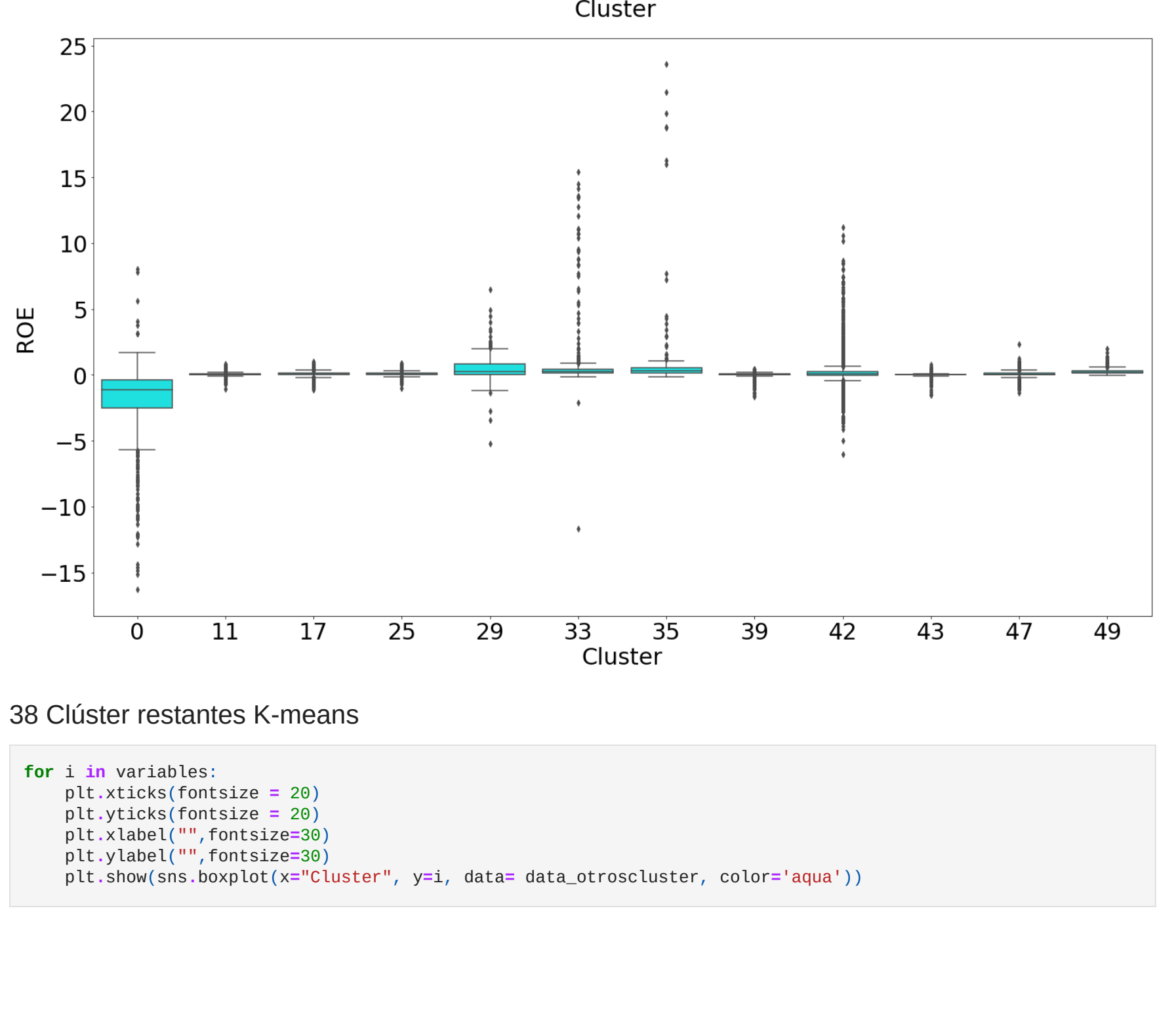






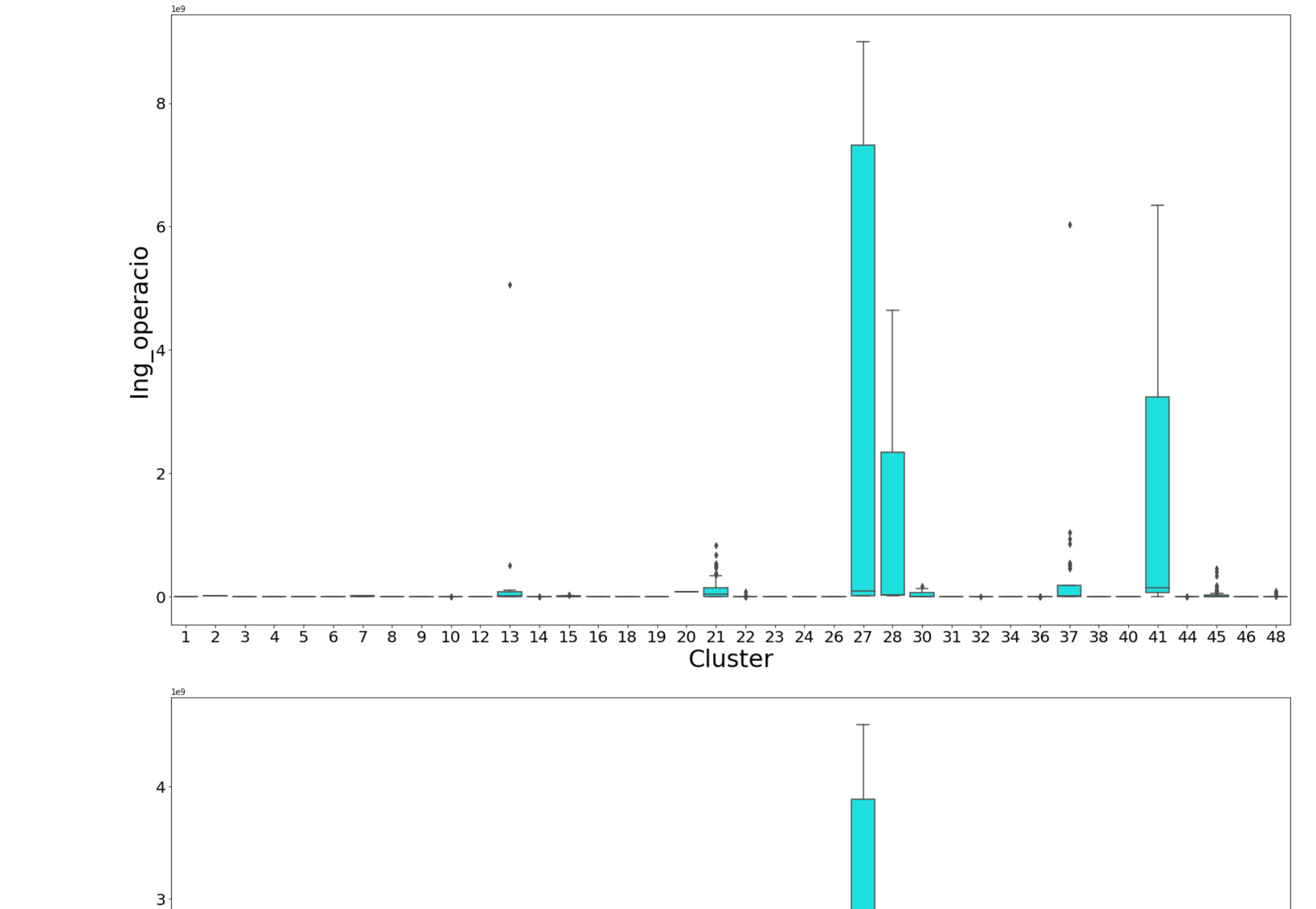


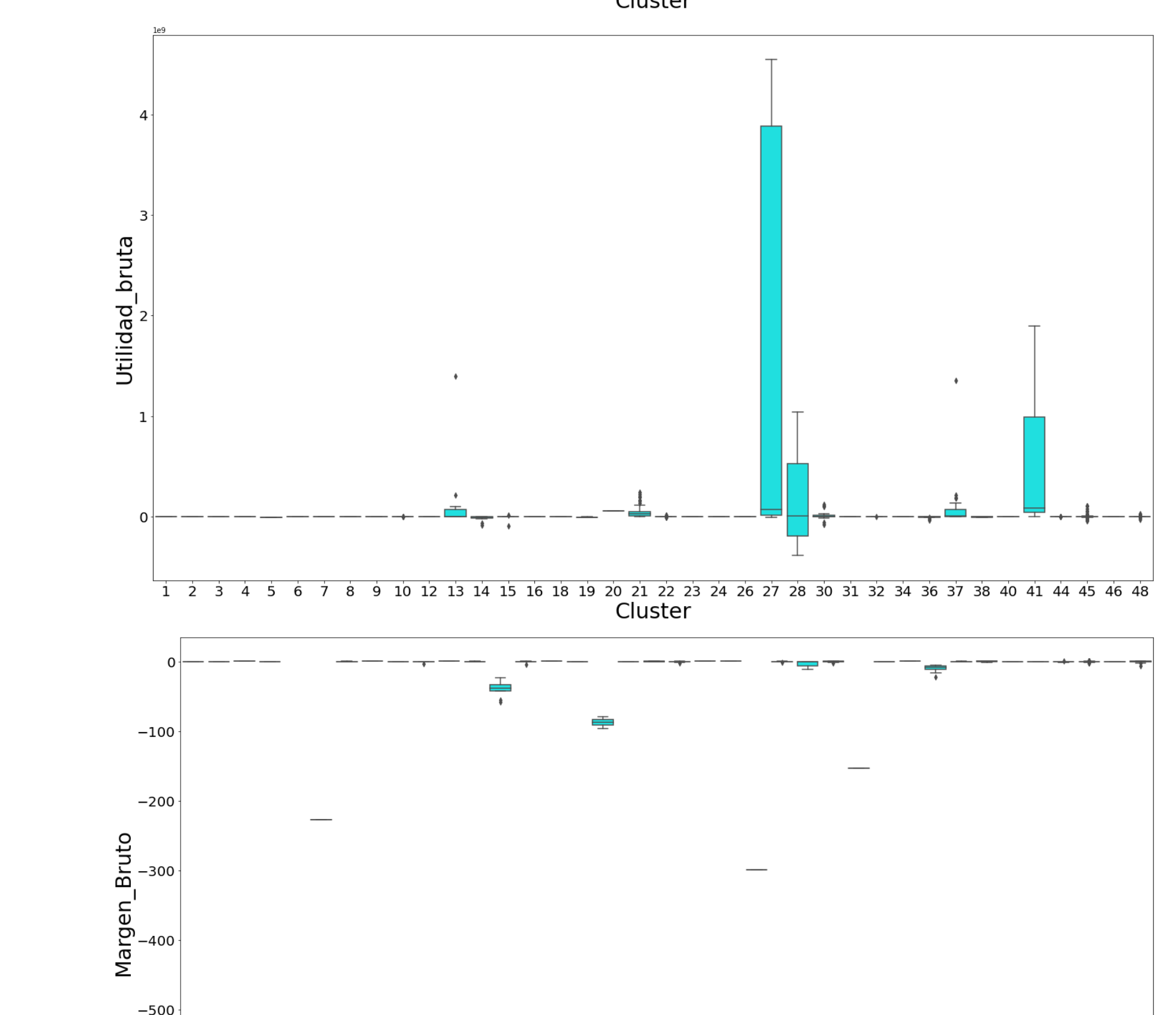


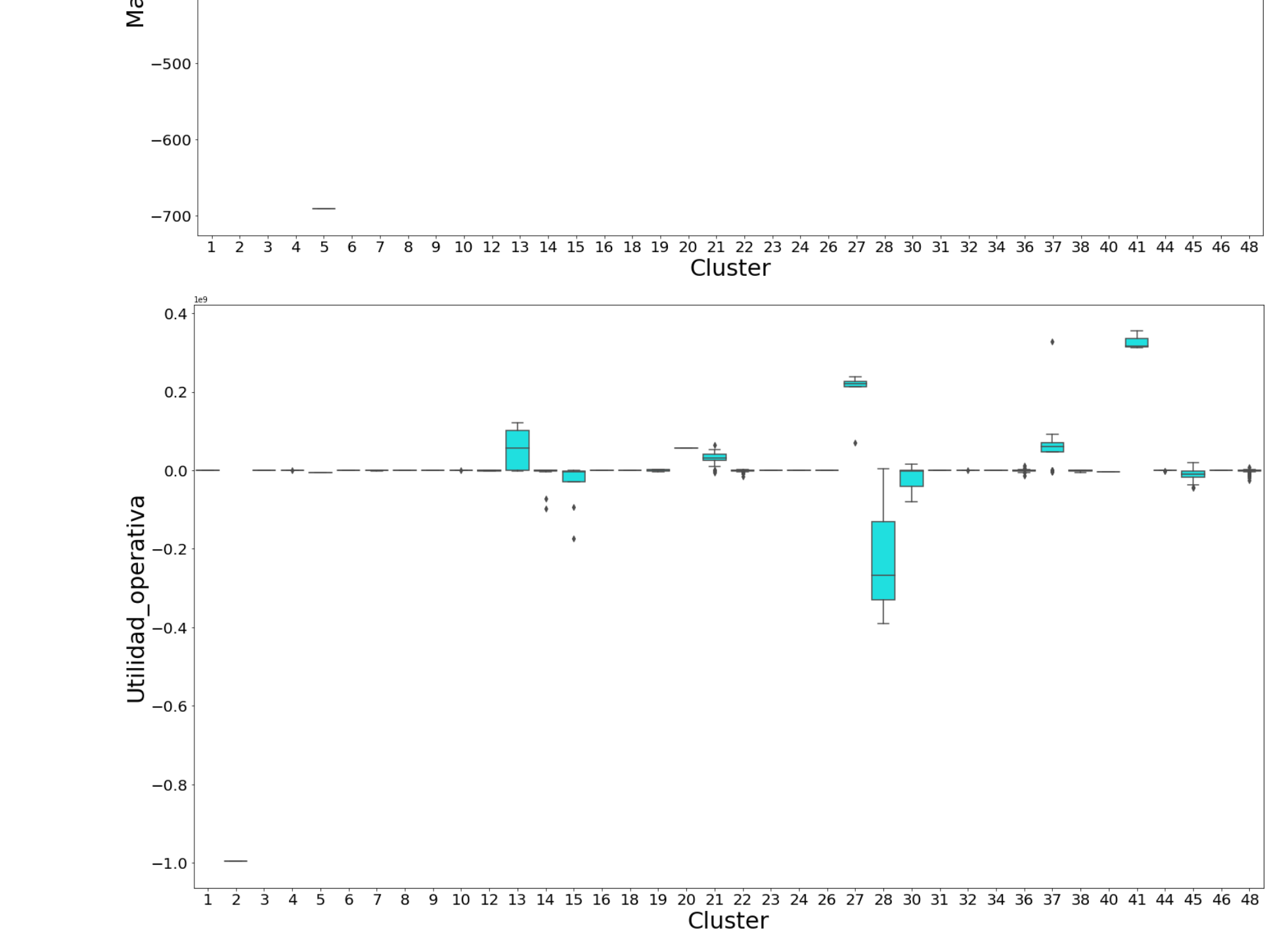


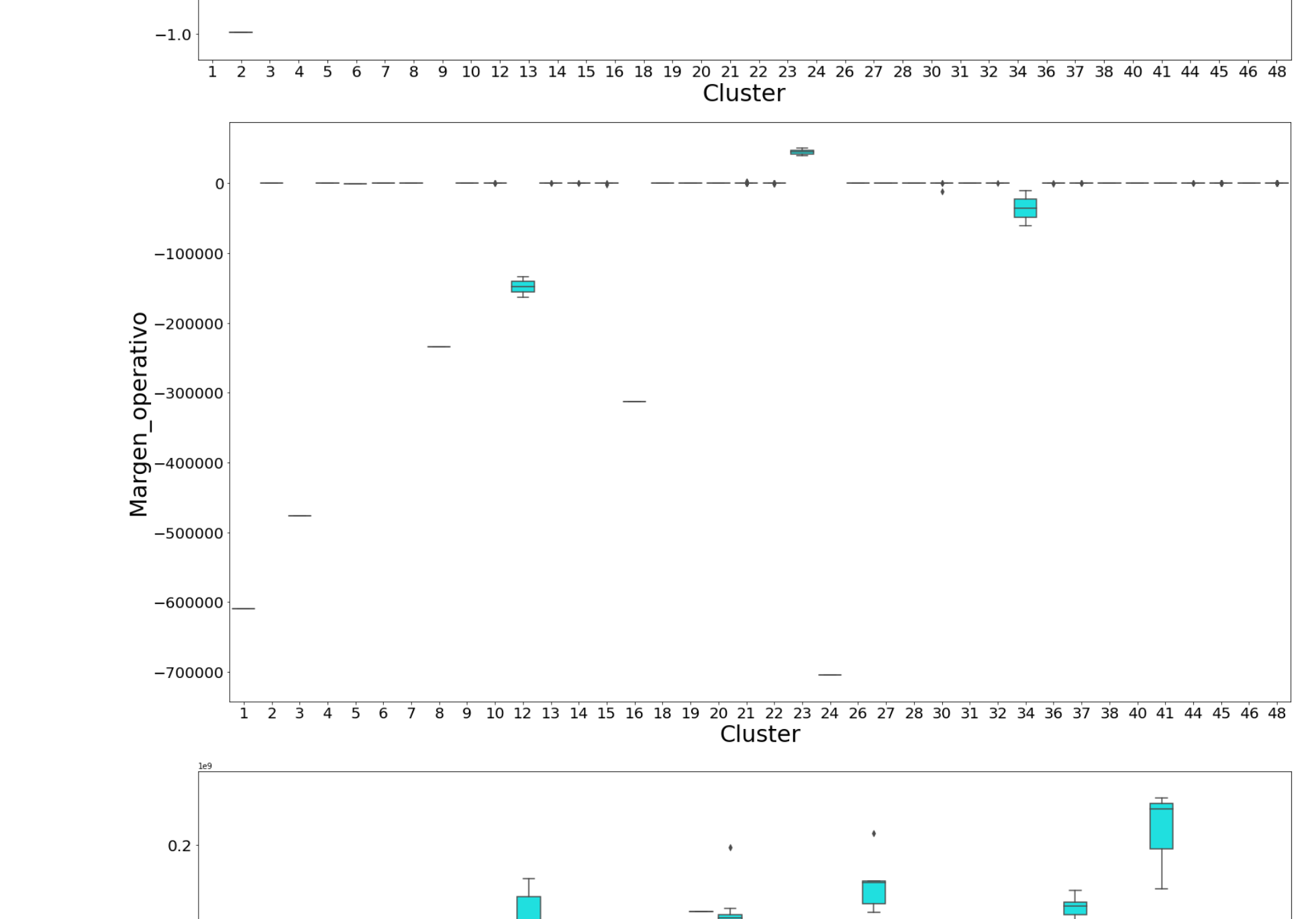
38 Clúster restantes K-means

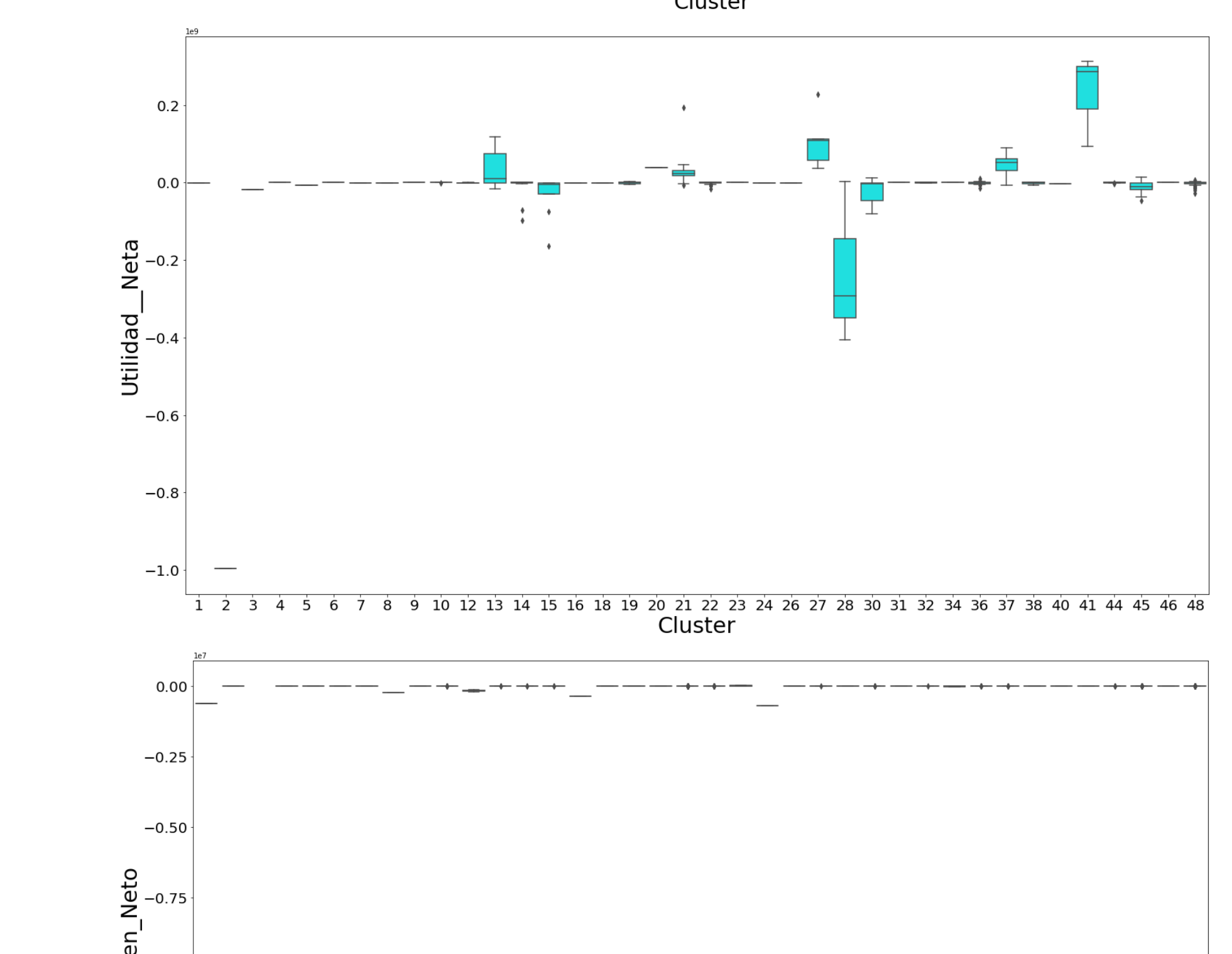
```python
In [49]:
for i in variables:
    plt.xticks(fontsize = 20)
    plt.yticks(fontsize = 20)
    plt.xlabel("",fontsize=30)
    plt.ylabel("",fontsize=30)
    plt.show(sns.boxplot(x="Cluster", y=i, data= data_otroscluster, color='aqua'))
```

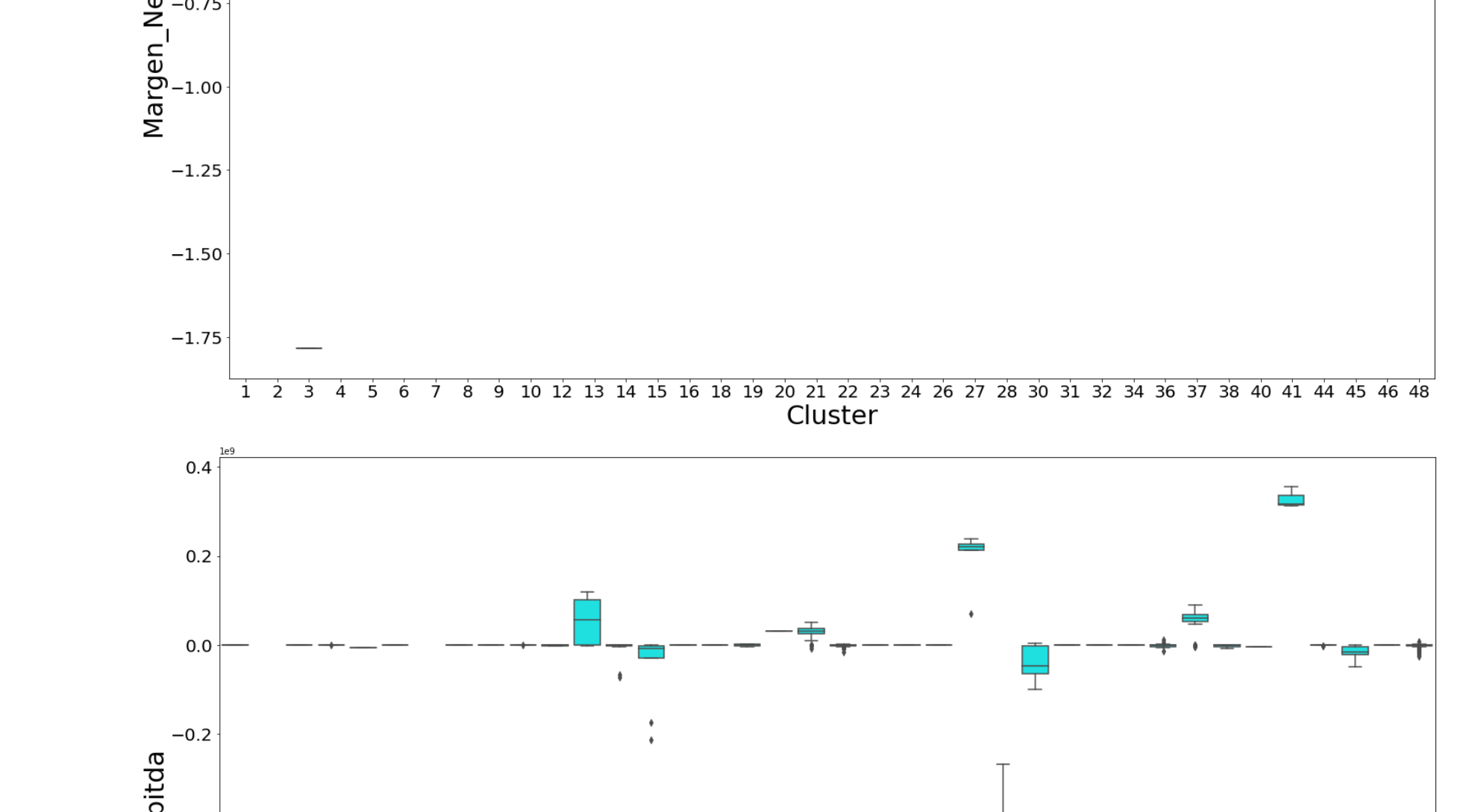


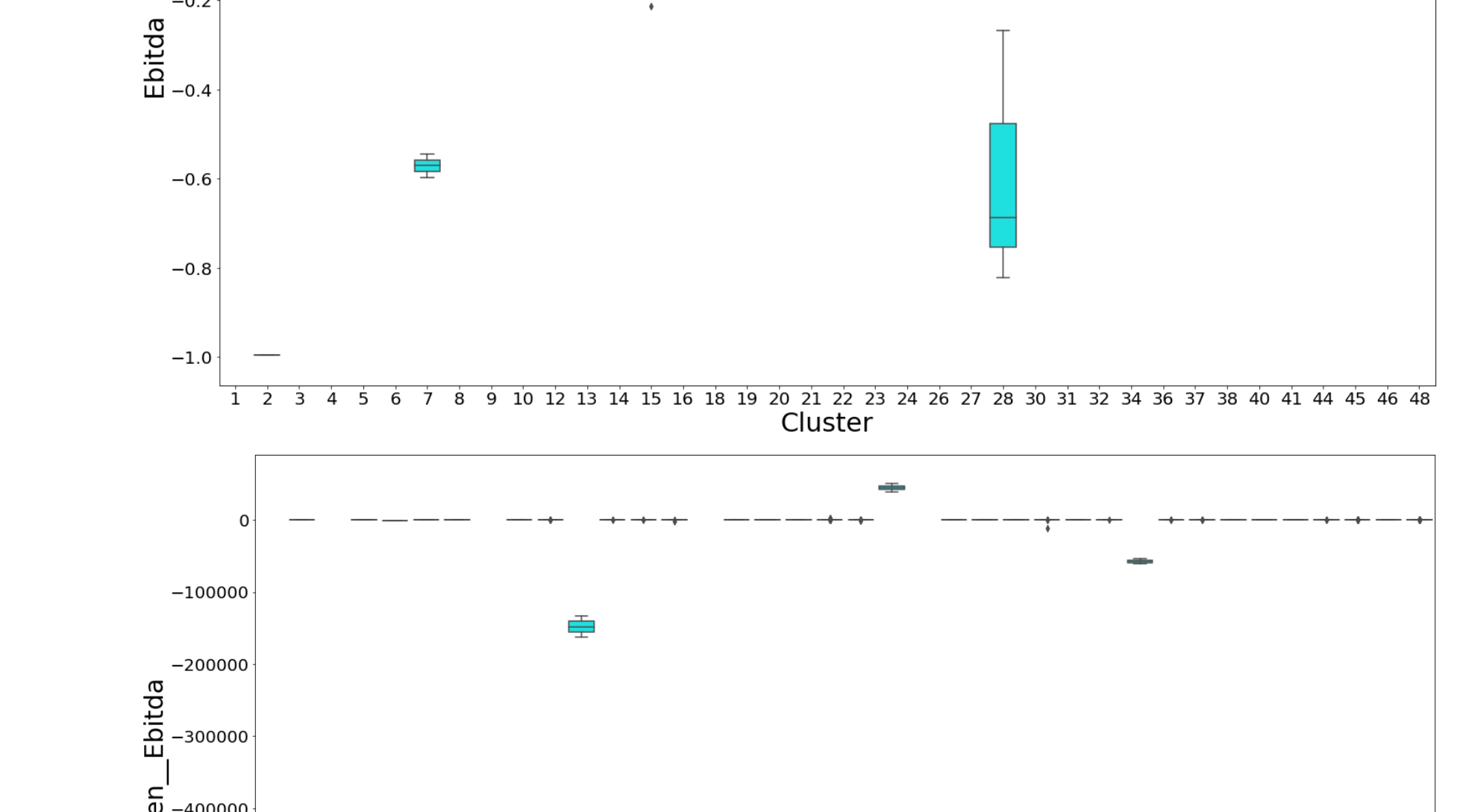


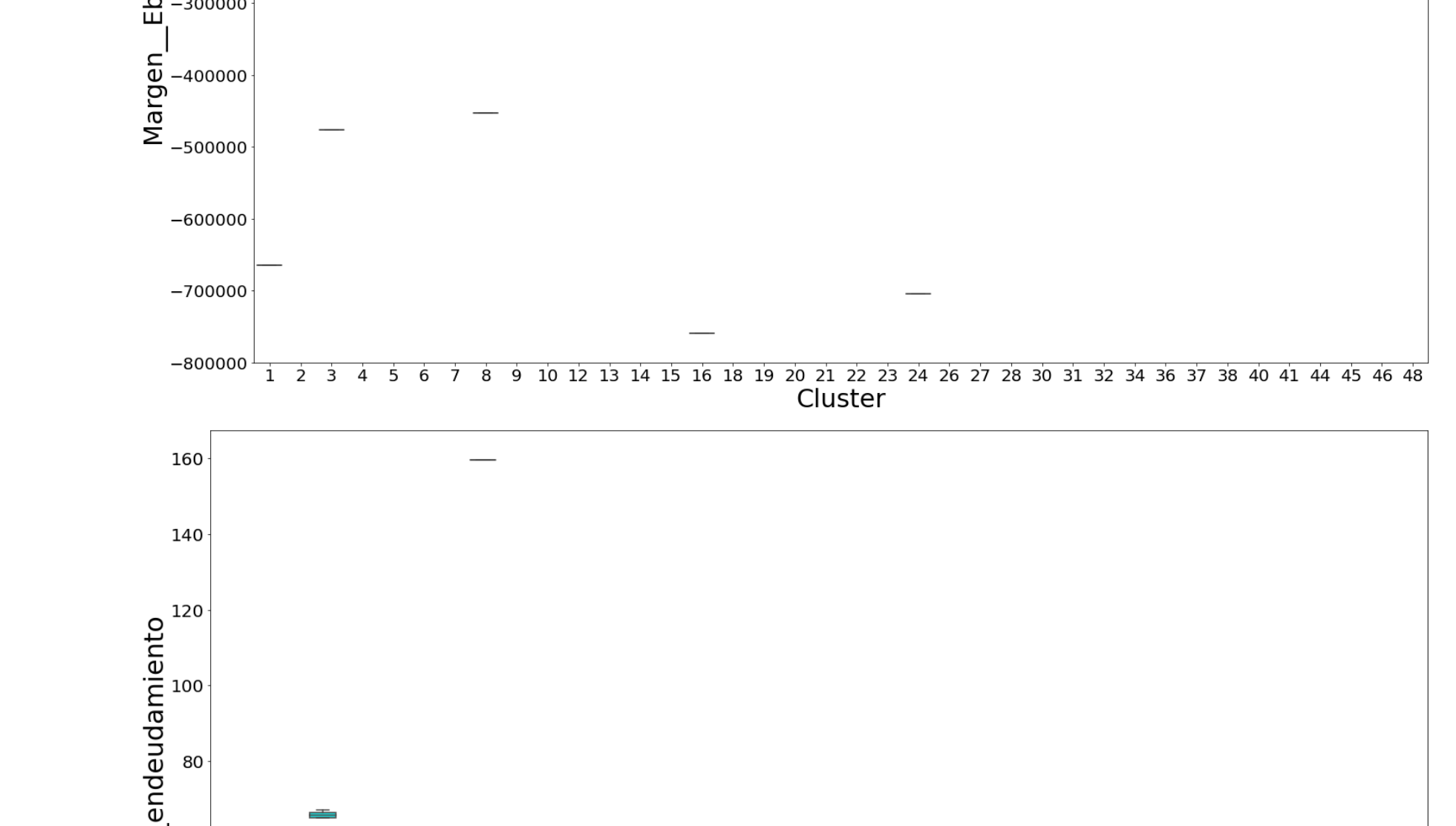


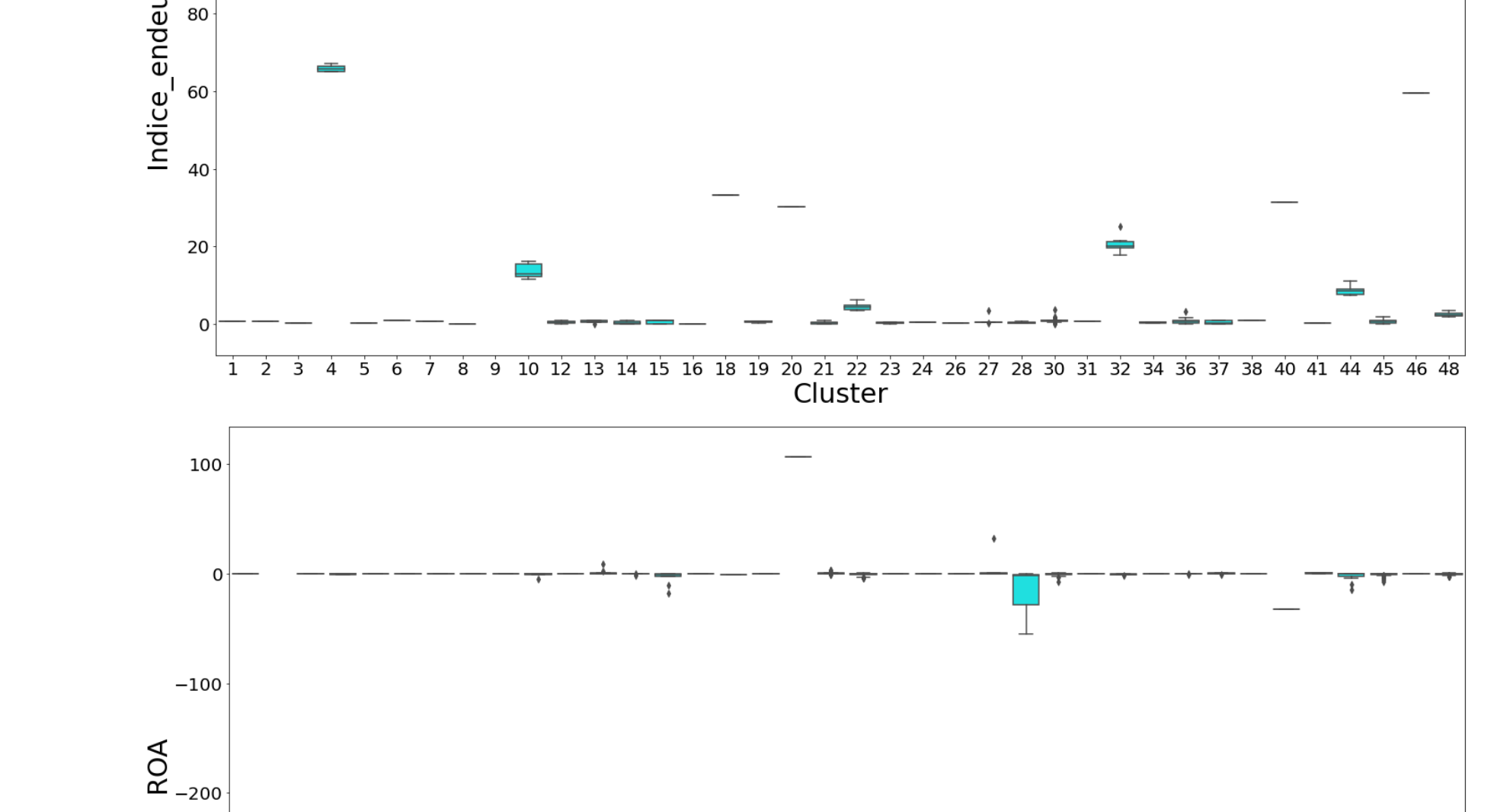


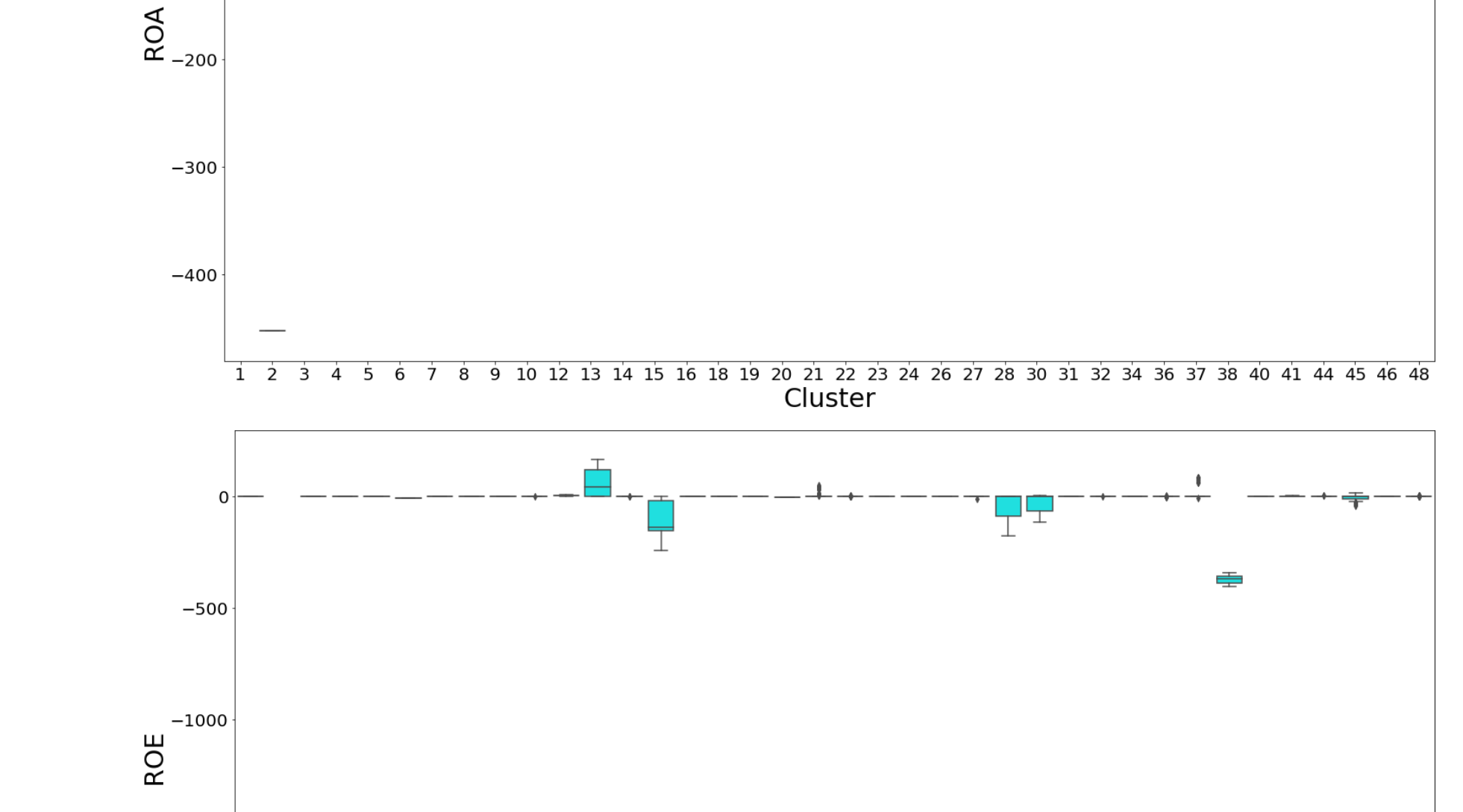


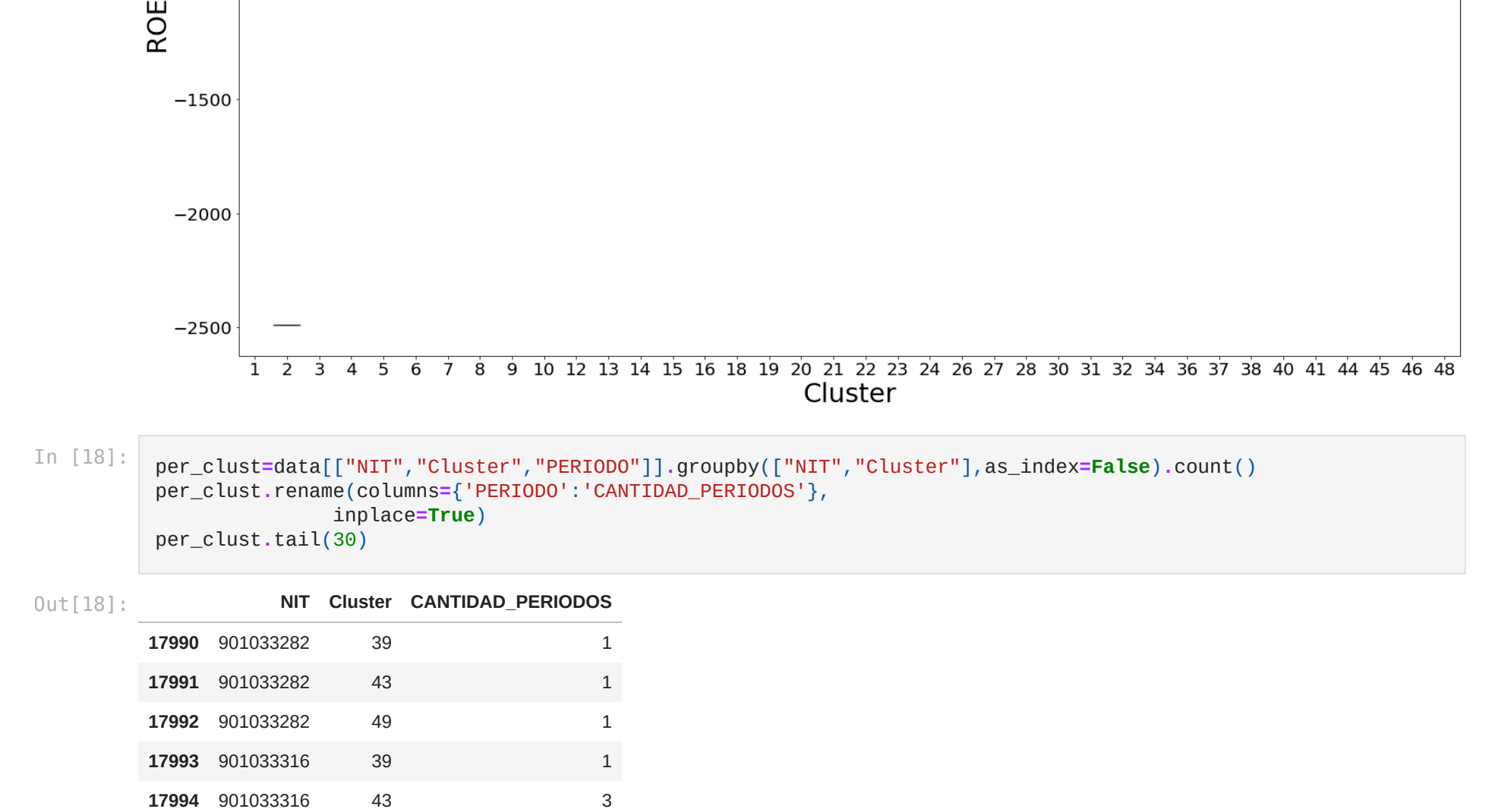












In [18]:

```
per_clust=data[["NIT","Cluster","PERIODO"]].groupby(["NIT","Cluster"],as_index=False).count()
per_clust.rename(columns={'PERIODO':'CANTIDAD_PERIODOS'},
                 inplace=True)
per_clust.tail(30)
```

Out[18]:

	NIT	Cluster	CANTIDAD_PERIODOS
17990	901033282	39	1
17991	901033282	43	1
17992	901033282	49	1
17993	901033316	39	1
17994	901033316	43	3
17995	901034604	25	2
17996	901034604	29	2
17997	901035348	42	4
17998	901035576	15	1
17999	901035576	39	1
18000	901035576	45	1
18001	901035576	49	1
18002	901035582	21	1
18003	901035582	43	3
18004	901038787	39	2
18005	901038787	43	1
18006	901038787	49	1
18007	901038968	39	4
18008	901042983	43	3
18009	901042983	49	1
18010	901061715	25	1
18011	901061715	39	2
18012	901061715	47	1
18013	901085146	21	1
18014	901085146	35	1
18015	901085146	39	2
18016	901107672	17	1
18017	901107672	33	1
18018	901107672	39	1
18019	901107672	47	1

In [19]:

```
per_clust.to_excel('cluster_nit_periodo.xlsx', sheet_name='cantidad')
```

Cluster Agglom

In [12]:

```
data[["NIT","Agglom"]].groupby('Agglom').count().sort_values("NIT",ascending=False)
```

Out[12]:

Agglom	NIT
18	7852
1	5932
20	5901
9	5463
0	4347
37	4044
24	1717
22	1240
11	1138
16	825
28	616
3	303
2	237
41	100
4	86
46	29
43	28
34	22
17	15
10	13
14	12
21	11
6	9
8	8
12	6
36	6
38	5
7	5
42	3
49	3
13	3
23	3
25	2
19	2
5	2
44	2
48	1
47	1
45	1
15	1
30	1
31	1
39	1
26	1
27	1
35	1
29	1
33	1
32	1
40	1

In [15]:

```
# los 12 principales clúster con mayor cantidad de empresas
data3= data[data.Agglom.isin([18,1,20,9,0,37,24,22,11,16,28,3])]
data3.head()
```

Out[15]:

	No._Empresas	NIT	PERIODO	Unnamed: 3	Ing_operacio	Utilidad_bruta	Margen_Bruto	Utilidad_operativa	Margen_operativo
0	0	800000268	2016	8000002682016	310130	306053	0.986854	38816	0.125160
1	1	800000268	2017	8000002682017	300000	300000	1.000000	66599	0.221997
2	2	800000268	2018	8000002682018	300000	300000	1.000000	64184	0.213947
3	3	800000268	2019	8000002682019	300000	300000	1.000000	208181	0.693937
4	4	800000276	2016	8000002762016	321430952	46473532	0.144583	9848367	0.030639

In [50]:

```
# los 38 clúster aglomerativos restantes
data_otrosaglom=data[~data.Agglom.isin([18,1,20,9,0,37,24,22,11,16,28,3])]
data_otrosaglom.head()
```

Out[50]:

	No._Empresas	NIT	PERIODO	Unnamed: 3	Ing_operacio	Utilidad_bruta	Margen_Bruto	Utilidad_operativa	Margen_operativo
16	16	800000750	2016	8000007502016	143563931	88936885	0.619493	316104238	2.201836
17	17	800000750	2017	8000007502017	111594829	54644303	0.489667	44133260	0.395478
18	18	800000750	2018	8000007502018	132566524	71012494	0.535674	63952725	0.482420
343	343	800013331	2019	8000133312019	14857665	14857665	1.000000	24279255	1.634123
532	532	800023434	2016	8000234342016	726910	726910	1.000000	19054405	26.212881

In [53]:

```
# los 12 clúster principales corresponden a 39378 datos
data3.shape
```

Out[53]: (39378, 18)

In [54]:

```
# los 38 clúster restantes corresponden a 626 datos
data_otrosaglom.shape
```

Out[54]: (626, 18)

12 clúster aglomerativos principales

In [43]:

```
for i in variables:
    plt.xticks(fontsize = 30)
    plt.yticks(fontsize = 30)
    plt.xlabel("",fontsize=30)
    plt.ylabel("",fontsize=30)
    plt.show(sns.boxplot(x="Agglom", y=i, data= data3, color='aqua'))
```

38 clúster aglomerativos restantes

In [52]:
```
for i in variables:
    plt.xticks(fontsize = 20)
    plt.yticks(fontsize = 20)
    plt.xlabel("",fontsize=30)
    plt.ylabel("",fontsize=30)
    plt.show(sns.boxplot(x="Agglom", y=i, data= data_otrosaglom, color='aqua'))
```

In [20]:
```
per_agglom=data[["NIT","Agglom","PERIODO"]].groupby(["NIT","Agglom"],as_index=False).count()
per_agglom.rename(columns={'PERIODO':'CANTIDAD_PERIODOS'},
              inplace=True)
per_agglom.tail(30)
```

Out[20]:

	NIT	Agglom	CANTIDAD_PERIODOS
18756	901033282	37	1
18757	901033316	0	1
18758	901033316	24	1
18759	901033316	37	2
18760	901034604	18	2
18761	901034604	28	2
18762	901035348	1	4
18763	901035576	3	1
18764	901035576	6	1
18765	901035576	20	1
18766	901035576	24	1
18767	901035582	2	1
18768	901035582	37	3
18769	901038787	11	1
18770	901038787	24	2
18771	901038787	37	1
18772	901038968	24	4
18773	901042983	11	1
18774	901042983	37	3
18775	901061715	0	1
18776	901061715	18	1
18777	901061715	20	1
18778	901061715	24	1
18779	901085146	2	1
18780	901085146	16	1
18781	901085146	24	2
18782	901107672	9	1
18783	901107672	11	1
18784	901107672	18	1
18785	901107672	24	1

In [21]:
```
per_agglom.to_excel('agglom_nit_periodo.xlsx', sheet_name='cantidad')
```

In [ ]: