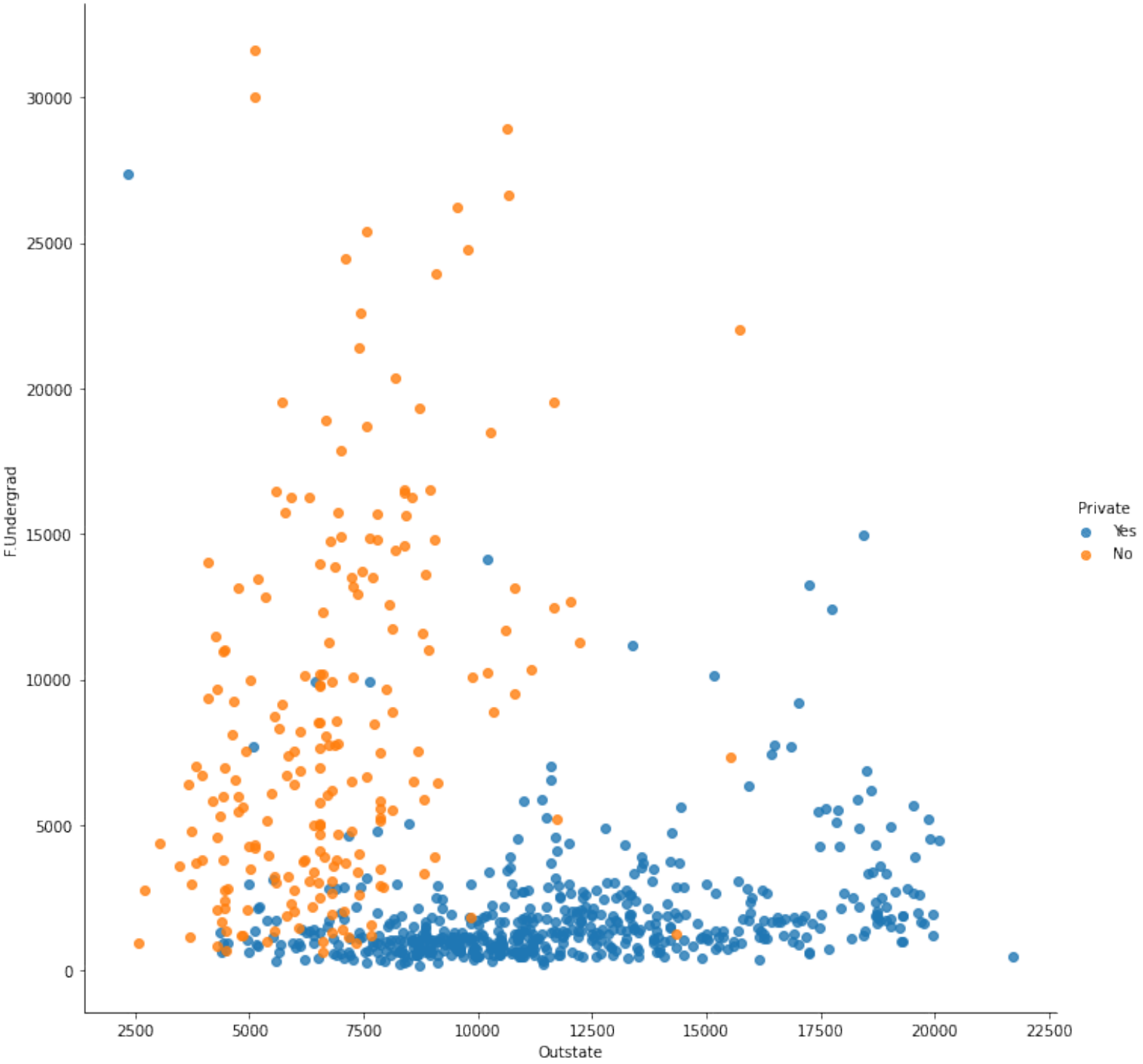

F.Undergrad
30000
25000
20000
15000
10000
5000
0
2500
5000
7500
10000
12500
15000
17500
20000
22500
Outstate
Private
Yes
No

In [55]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import cv2
import plotly.graph_objs as go
import cufflinks as cf


from sklearn import metrics
from chart_studio import plotly as py
from plotly.offline import download_plotlyjs,init_notebook_mode,plot,iplot
from pandas_datareader import data, wb

%matplotlib inline
```

Data Gathering (Unsupervised)

In [56]:
```
df =pd.read_csv(r".\data\College.csv",index_col=0)
```

In [57]:
```
df.head()
```

Out[57]:

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527
Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735
Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922

Data Processing

In [58]:
```
df.info()
```

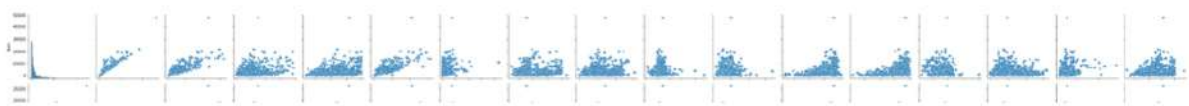
In [59]:
```
df.describe()
```

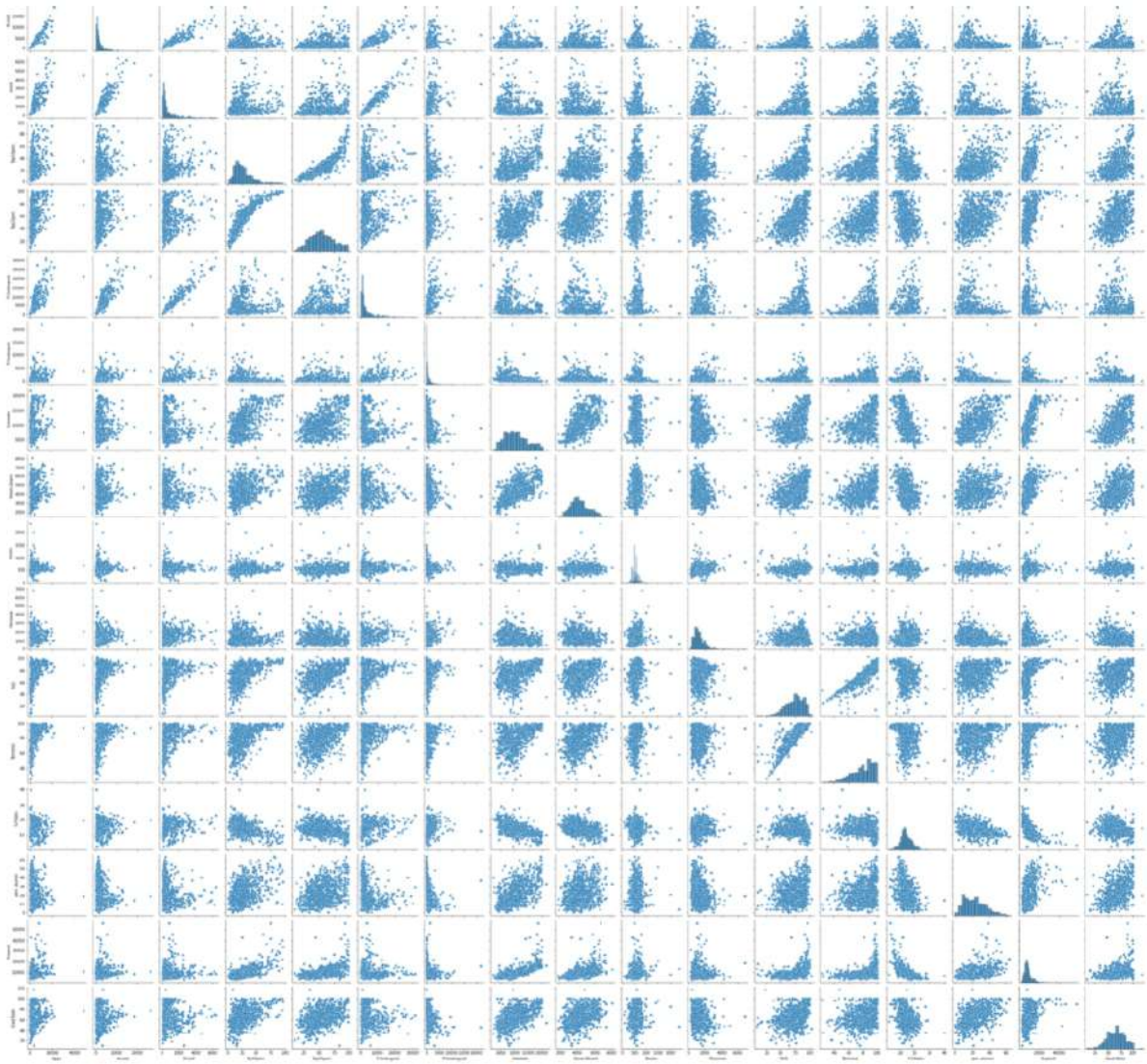
Out[59]:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384	549.380952	1340.642214	72.660232	79.702703
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416	165.105360	677.071454	16.328155	14.722359
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	96.000000	250.000000	8.000000	24.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000	850.000000	62.000000	71.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000	1200.000000	75.000000	82.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000	1700.000000	85.000000	92.000000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000	6800.000000	103.000000	100.000000

In [60]:
```
sns.pairplot(df)
```

Out[60]:





In [61]:
```
df.columns
```

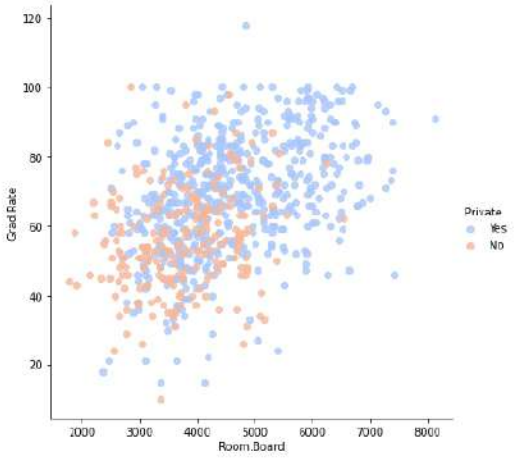
Out[61]:

In [62]:
```
sns.lmplot(data=df,x="Room.Board", y="Grad.Rate", hue="Private", fit_reg=False,
           palette='coolwarm', size=6, aspect=1)
```

```
C:\Users\hp\AppData\Roaming\Python\Python310\site-packages\seaborn\regression.py:581: UserWarning:

The `size` parameter has been renamed to `height`; please update your code.
```

Out[62]:

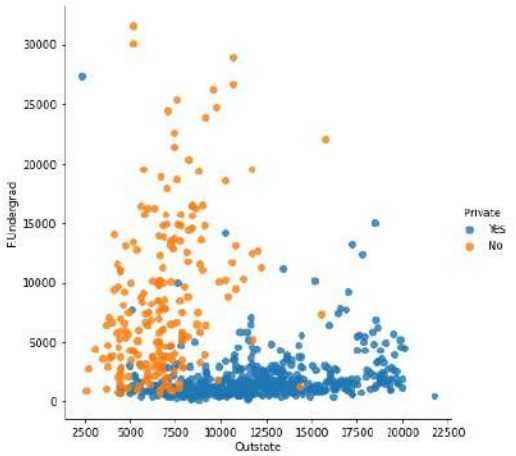


In [63]:
```
sns.lmplot(data=df,x="Outstate", y="F.Undergrad", hue="Private", fit_reg=False,
           size=6, aspect=1)
```

```
C:\Users\hp\AppData\Roaming\Python\Python310\site-packages\seaborn\regression.py:581: UserWarning:
```

```
The `size` parameter has been renamed to `height`; please update your code.
```

Out[63]:

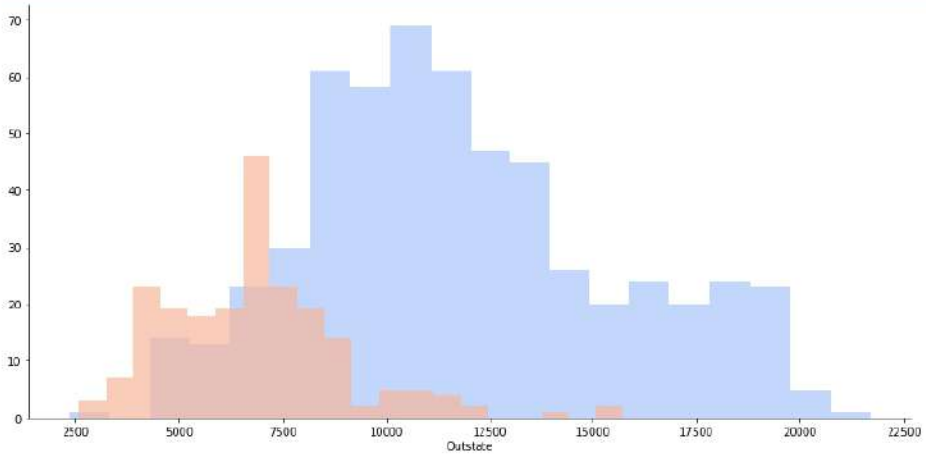


In [64]:

```
g = sns.FacetGrid(df,hue="Private", palette='coolwarm', size=6,aspect=2)
g = g.map(plt.hist, "Outstate", bins=20, alpha=0.7)
```

```
C:\Users\hp\AppData\Roaming\Python\Python310\site-packages\seaborn\axisgrid.py:337: UserWarning:

The `size` parameter has been renamed to `height`; please update your code.
```

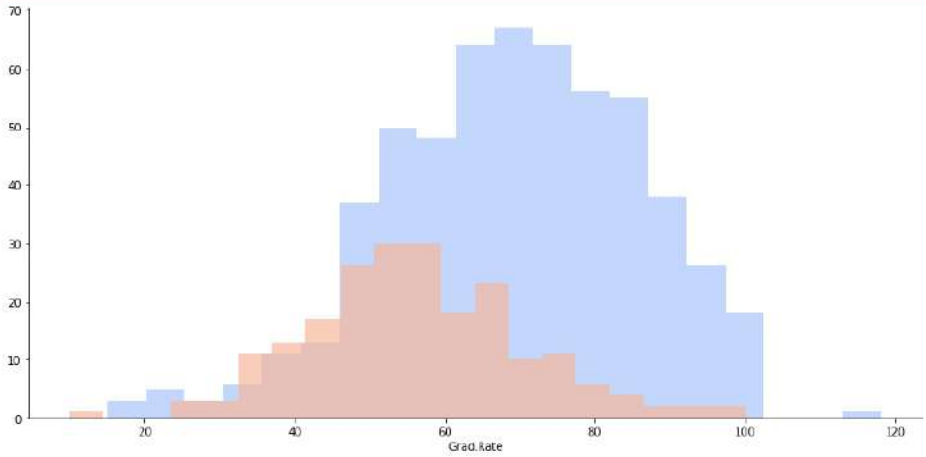


In [65]:

```
g = sns.FacetGrid(df, hue="Private", palette="coolwarm",size=6, aspect=2)
g = g.map(plt.hist,'Grad.Rate',bins=20,alpha=0.7)
```

```
C:\Users\hp\AppData\Roaming\Python\Python310\site-packages\seaborn\axisgrid.py:337: UserWarning:

The `size` parameter has been renamed to `height`; please update your code.
```



In [66]:

```
df[df['Grad.Rate']>100]
```

Out[66]:

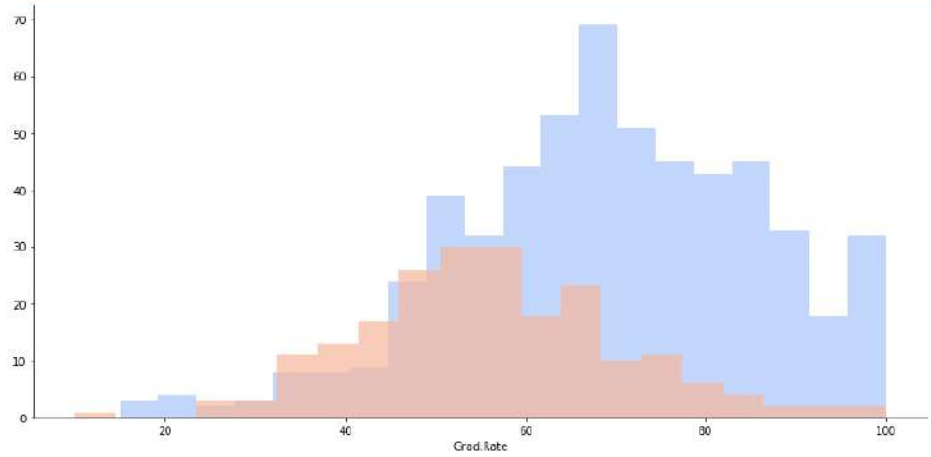
	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend
Cazenovia College	Yes	3847	3433	527	9	35	1010	12	9384	4840	600	500	22	47	14.3	20	7697

In [67]:
```
df['Grad.Rate']["Cazenovia College"]= 100
```

In [68]:
```
g = sns.FacetGrid(df, hue="Private", palette="coolwarm",size=6, aspect=2)
g = g.map(plt.hist,'Grad.Rate',bins=20,alpha=0.7)
```

```
C:\Users\hp\AppData\Roaming\Python\Python310\site-packages\seaborn\axisgrid.py:337: UserWarning:

The `size` parameter has been renamed to `height`; please update your code.
```



Modeling

In [69]:
```
from sklearn.cluster import KMeans
model = KMeans(n_clusters=2)
model.fit(df.drop("Private", axis=1))
```

Out[69]:

In [70]:
```
model.cluster_centers_
```

Out[70]:

In [71]:
```
model.labels_
```

Out[71]:

Evaluation

In [72]:
```
def converter(private):
    if private == "Yes":
        return 1
    else:
        return 0
```

In [73]:
```
df["Cluster"] = df['Private'].apply(converter)
```

In [74]:
```
df.head()
```

Out[74]:

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527
Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735
Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922

In [75]:
```
from sklearn.metrics import confusion_matrix,classification_report
print(confusion_matrix(df["Cluster"], model.labels_))
print(classification_report(df["Cluster"], model.labels_))
```