In [1]:
```
#import files
from sklearn.cluster import KMeans
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from matplotlib import pyplot as plt
get_ipython().run_line_magic('matplotlib', 'inline')
```

In [12]:
```
#read file
df1=pd.read_csv("excleofDataSet.csv")
df1.head()
```

Out[12]:

	Unnamed: 0	sl_no	University_iD	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	
0	0	1.0	0	M	67.00	Others	67.00	Others	Commerce	58.00	
1	1	2.0	12346	M	79.33	Central	79.33	Others	Science	77.48	
2	2	3.0	0	Other	65.00	NaN	65.00	Central	NaN	0.00	C
3	3	4.0	12348	M	56.00	Central	56.00	Central	Science	52.00	
4	4	5.0	12349	M	85.80	Central	85.80	Central	Commerce	73.30	C

In [14]:
```
df = df1[['mba_p', 'salary']]
print(df)
```

```
     mba_p  salary
0    58.80       0
1    66.28  200000
2     0.00       0
3    59.43       0
4    55.50  425000
..     ...     ...
213  74.49  400000
214  53.62  275000
215  69.72  295000
216  60.23  204000
217  60.22       0

[218 rows x 2 columns]
```

In [37]:
```
# Min-Max Normalization
df_norm = (df-df.min())/(df.max()-df.min())
print("Scaled Dataset Using Pandas")
df_norm.head()
```

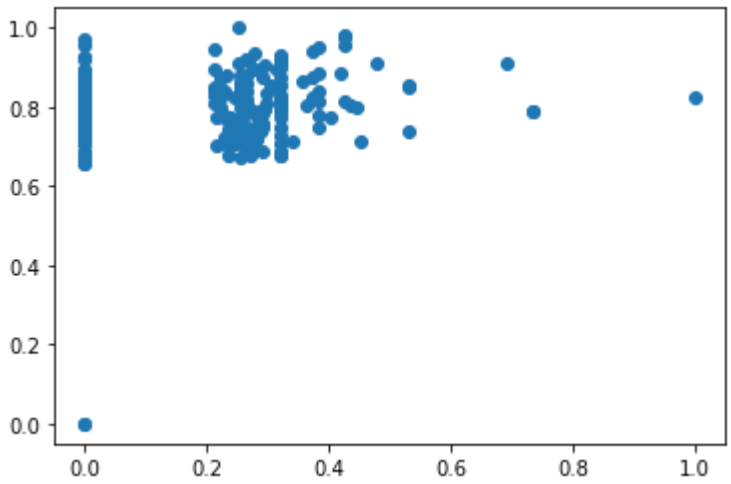
```
Scaled Dataset Using Pandas
```

Out[37]:

	mba_p	salary
0	0.754911	0.000000
1	0.850944	0.212766
2	0.000000	0.000000
3	0.762999	0.000000
4	0.712543	0.452128

In [39]:
```
#Scatter Plot
plt.scatter(df_norm['salary'],df_norm['mba_p'])
```

Out[39]:
```
<matplotlib.collections.PathCollection at 0x20d74727070>
```



In [42]:
```
# Choose K
km=KMeans(n_clusters=2)
km
#convert all in array /group
y_predicted = km.fit_predict(df_norm[['salary','mba_p']])
y_predicted
```

Out[42]:
```
array([0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1,
       1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1,
       1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0,
       1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0,
       1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0,
       1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1,
       1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1,
       0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0,
       1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0])
```

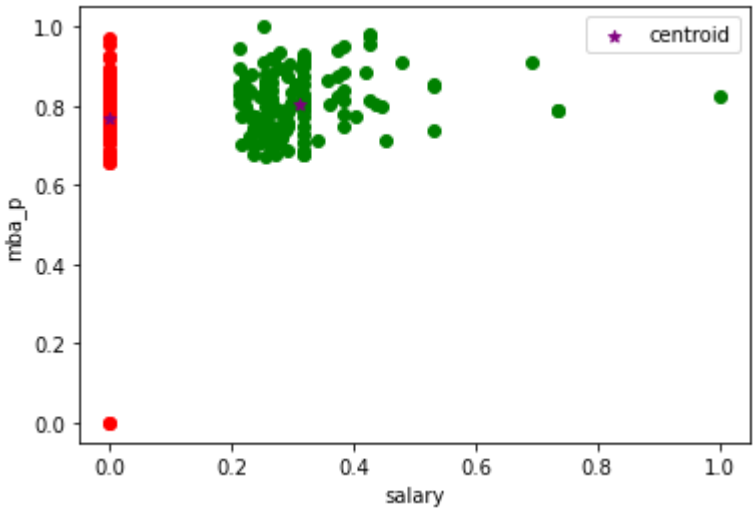
In [43]:
```
#Centroids
km.cluster_centers_
```

Out[43]:
```
array([[5.55111512e-17, 7.67677060e-01],
       [3.10681719e-01, 8.03857350e-01]])
```

In [30]:
```
#datafram to three group and ploat Scatter plot
df1 = df_norm[df_norm.cluster==0]
df2 = df_norm[df_norm.cluster==1]
plt.scatter(df1.salary ,df1['mba_p'],color='green')
plt.scatter(df2.salary ,df2['mba_p'],color='red')

#ploatling centroids
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color='purple',mark
plt.xlabel('salary')
plt.ylabel('mba_p')
plt.legend()
```

Out[30]:
```
<matplotlib.legend.Legend at 0x20d74041f70>
```



In [32]:
```
df1
```

Out[32]:

	mba_p	salary	cluster
1	0.850944	0.212766	0
4	0.712543	0.452128	0
7	0.797792	0.268085	0
10	0.781230	0.276596	0
11	0.817820	0.265957	0
...	...	...	...
212	0.725254	0.229787	0
213	0.956349	0.425532	0
214	0.688407	0.292553	0
215	0.895108	0.313830	0
216	0.773270	0.217021	0

147 rows × 3 columns

In [33]:
```
df2
```

Out[33]:

	mba_p	salary	cluster
0	0.754911	0.0	1
2	0.000000	0.0	1
3	0.762999	0.0	1
5	0.662216	0.0	1
6	0.684170	0.0	1
...	...	...	...
201	0.923867	0.0	1
204	0.750289	0.0	1
209	0.685454	0.0	1
211	0.807806	0.0	1
217	0.773142	0.0	1

71 rows × 3 columns

In [ ]:
```

```