DS Clustering / K-Means

In [16]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [17]:

```
df = pd.read_csv('/home/gmelao/Desktop/default-of-credit-card-clients.csv')
df.columns = df.iloc[0]
df.drop(0, inplace = True)
df.set_index('ID', inplace = True)
pd.set_option('display.max_columns', 24)
pd.set_option('display.max_rows', 24)
```

In [18]:

```
df = df.apply(lambda df: pd.Series(map(float, df)))
```

In [19]:

```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
```

df

In [33]:

```
df.columns
```

Out[33]:

```
Index(['LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY_0', 'PAY_2',
       'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'BILL_AMT1', 'BILL_AMT2',
       'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1',
       'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6',
       'default payment next month', 'Cluster'],
      dtype='object', name=0)
```

In [20]:

```
scaler = StandardScaler()
scaled_df = scaler.fit_transform(df)
```

In [21]:

```
model = KMeans(n_clusters=5)
cluster_label = model.fit_predict(scaled_df)
```

```
/home/gmelao/mambaforge/lib/python3.10/site-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
```

In [22]:

```
cluster_label
```

Out[22]:

```
array([1, 1, 1, ..., 3, 1, 1], dtype=int32)
```

In [23]:

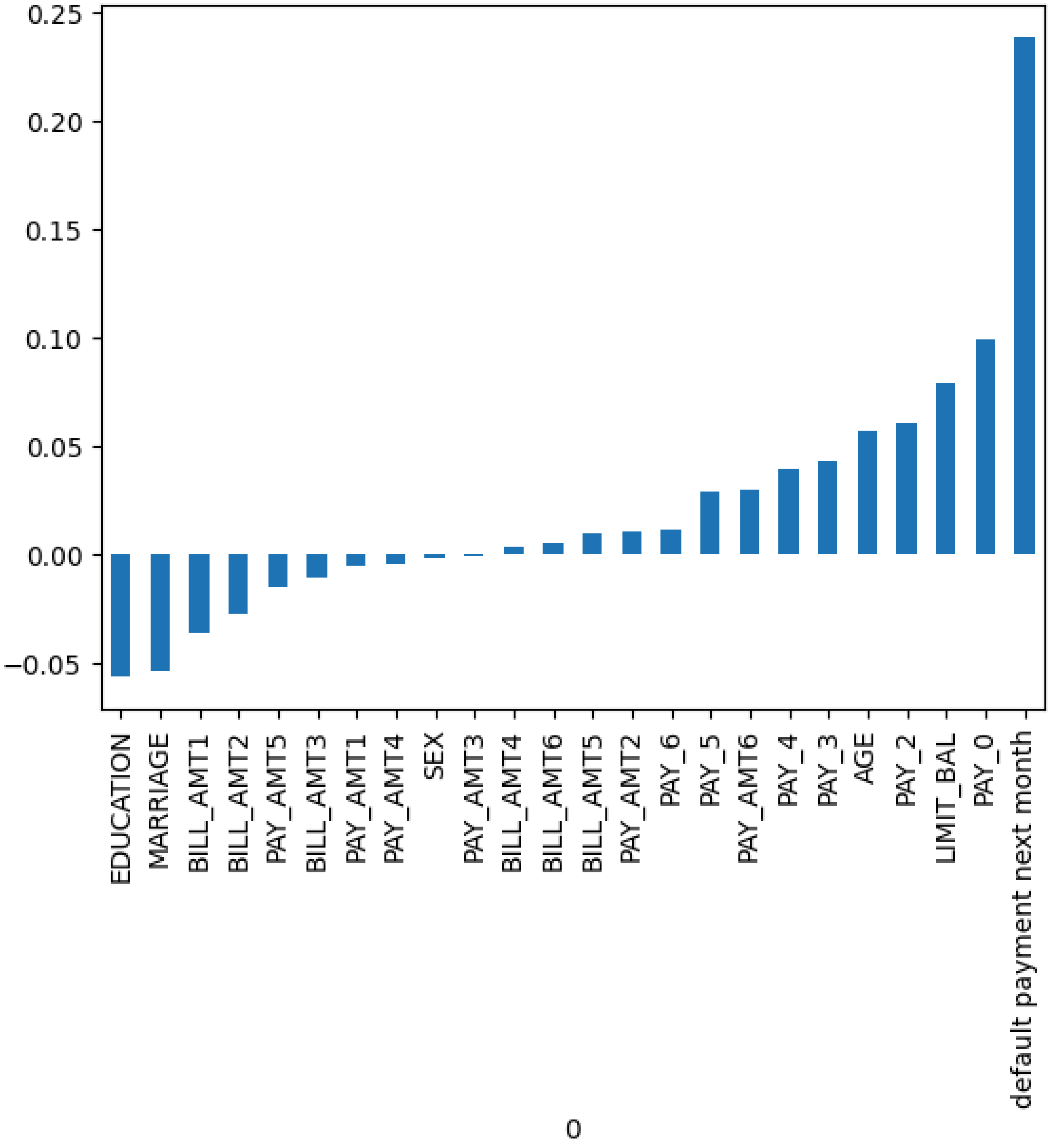
```
df['Cluster'] = cluster_label
```

In [24]:

```
df.corr()['Cluster'].iloc[:-1].sort_values().plot(kind='bar')
```

Out[24]:

```
<AxesSubplot: xlabel='0'>
```

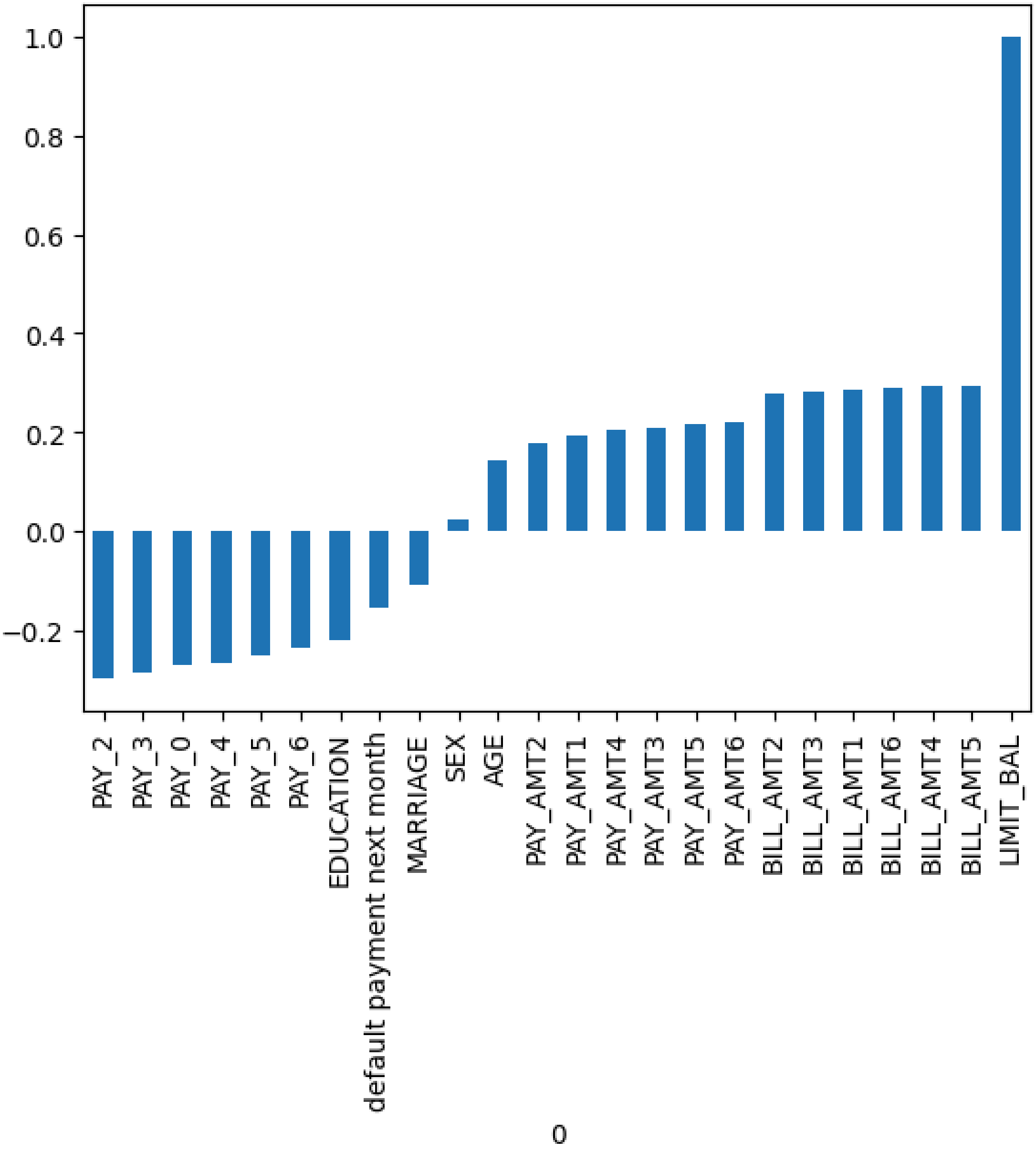


In [25]:

```
df.corr()['LIMIT_BAL'].iloc[:-1].sort_values().plot(kind='bar')
```

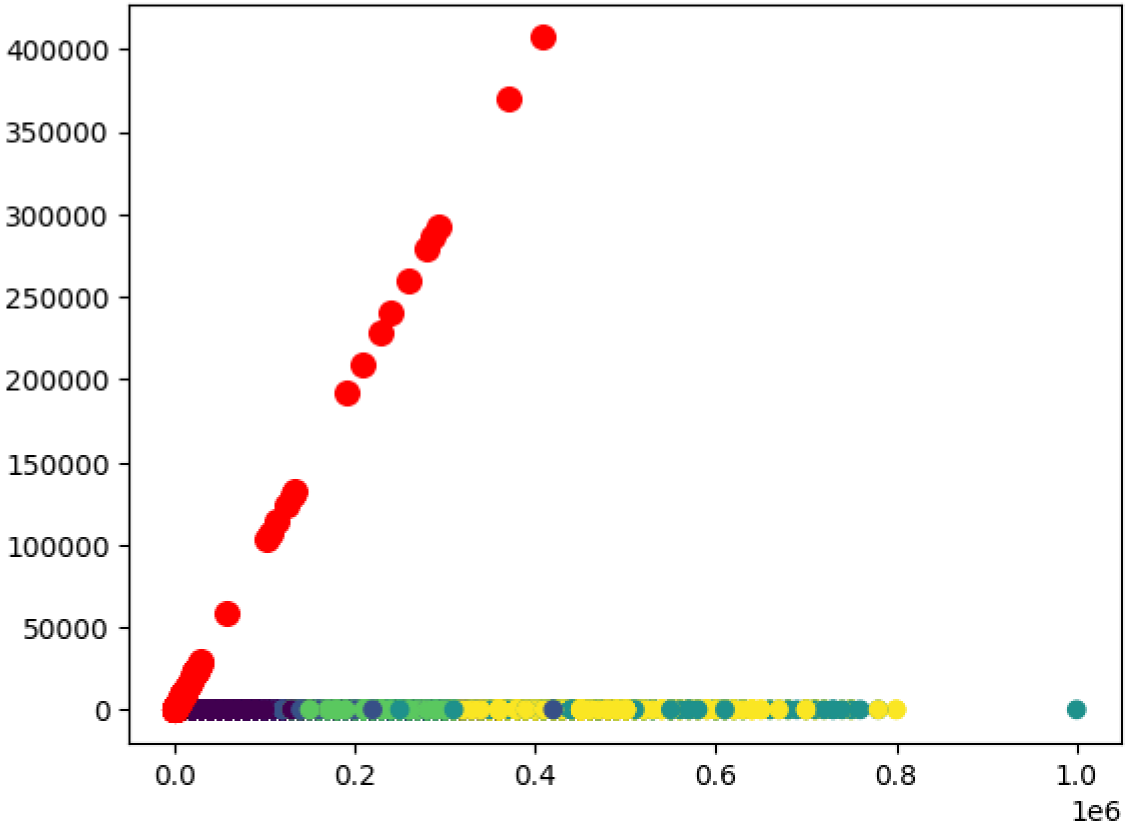
Out[25]:

```
<AxesSubplot: xlabel='0'>
```



In [11]:

```
kmeans = KMeans(n_clusters = 5, init = 'k-means++', n_init = 10, max_iter = 100)
pred_y = kmeans.fit_predict(df)
plt.scatter(df['LIMIT_BAL'], df['Cluster'], c = pred_y)
plt.scatter(kmeans.cluster_centers_,kmeans.cluster_centers_, s = 70, c = 'red')
plt.show()
```

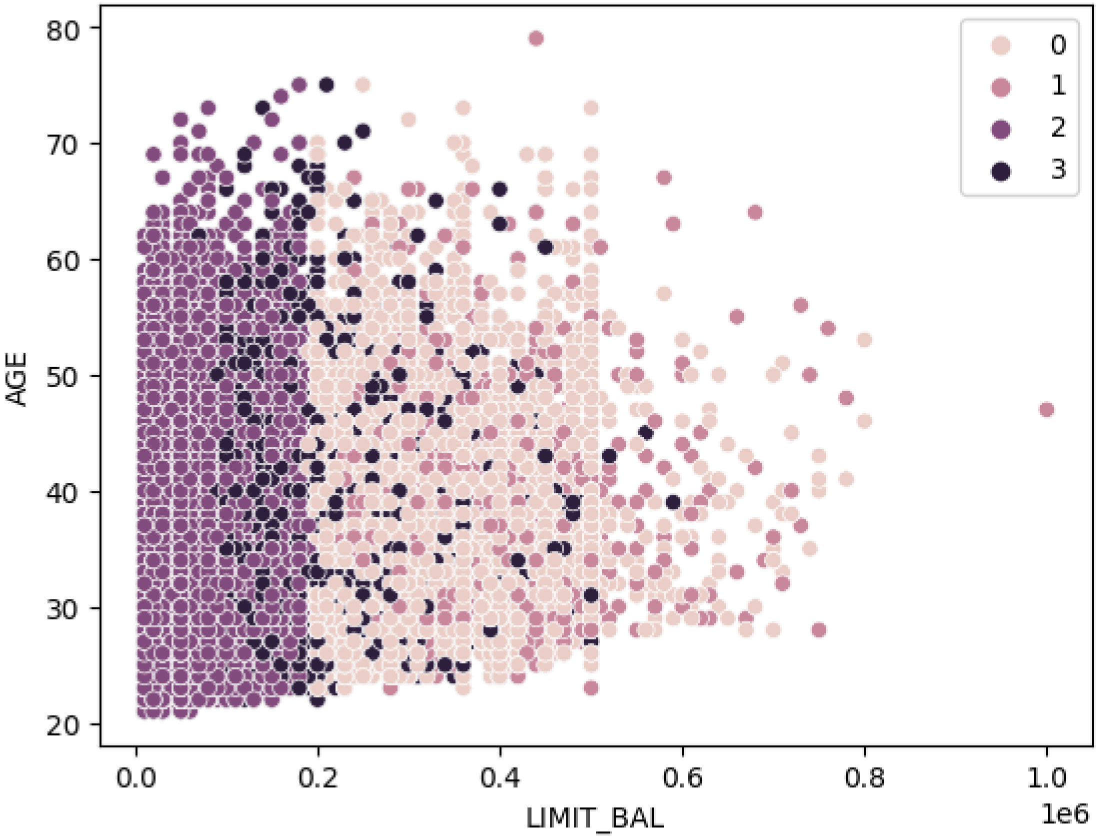


In [30]:
```
def display_models(model, data, X, Y):
    labels = model.fit_predict(data)
    sns.scatterplot(data=data, x=X, y=Y, hue=labels)
```

In [31]:
```
model = KMeans(n_clusters=4)
```

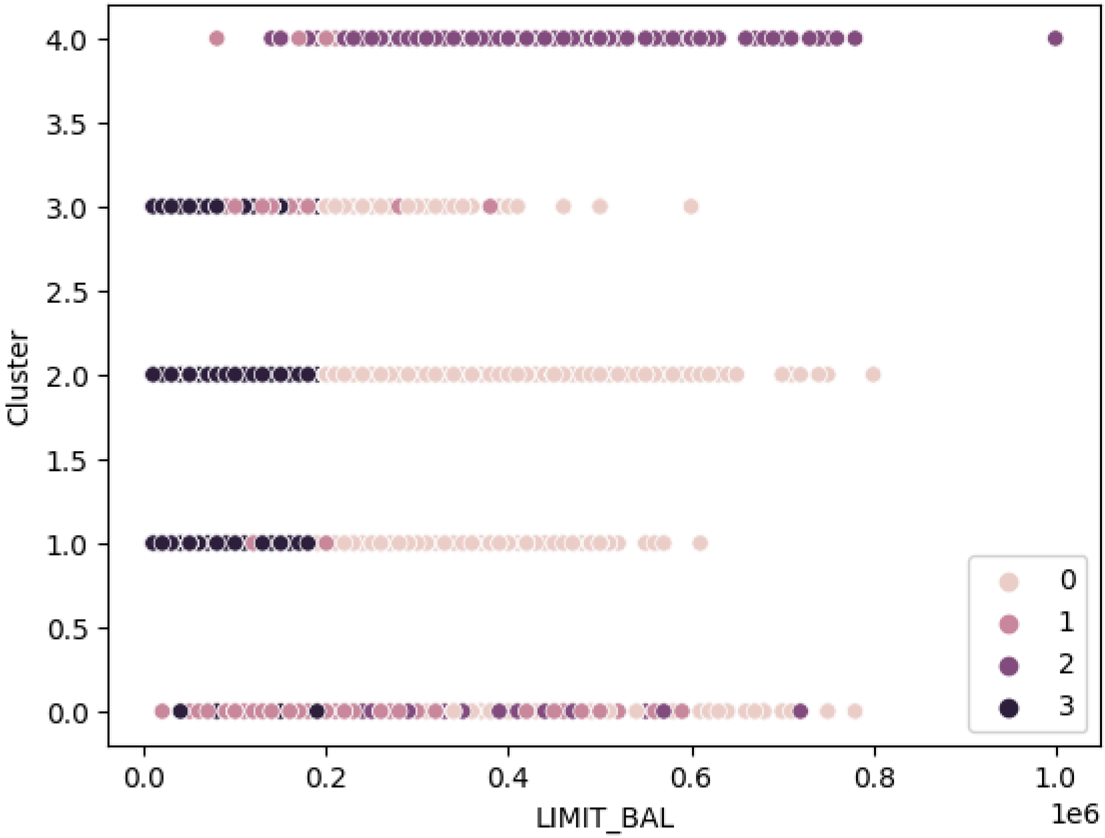
In [32]:
```
display_models(model, df, 'LIMIT_BAL', 'AGE')
```

```
/home/gmelao/mambaforge/lib/python3.10/site-packages/sklearn/cluster/_kmeans.py:87
0: FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
```



In [34]:
```
display_models(model, df, 'LIMIT_BAL', 'Cluster')
```

```
/home/gmelao/mambaforge/lib/python3.10/site-packages/sklearn/cluster/_kmeans.py:87
0: FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
```



In [35]:
```
df
```

Out[35]:

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_
0	20000.0	2.0	2.0	1.0	24.0	2.0	2.0	-1.0	-1.0	-2.0	-2
1	120000.0	2.0	2.0	2.0	26.0	-1.0	2.0	0.0	0.0	0.0	2
2	90000.0	2.0	2.0	2.0	34.0	0.0	0.0	0.0	0.0	0.0	0
3	50000.0	2.0	2.0	1.0	37.0	0.0	0.0	0.0	0.0	0.0	0
4	50000.0	1.0	2.0	1.0	57.0	-1.0	0.0	-1.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	
29995	220000.0	1.0	3.0	1.0	39.0	0.0	0.0	0.0	0.0	0.0	0
29996	150000.0	1.0	3.0	2.0	43.0	-1.0	-1.0	-1.0	-1.0	0.0	0
29997	30000.0	1.0	2.0	2.0	37.0	4.0	3.0	2.0	-1.0	0.0	0
29998	80000.0	1.0	3.0	1.0	41.0	1.0	-1.0	0.0	0.0	0.0	-1
29999	50000.0	1.0	2.0	1.0	46.0	0.0	0.0	0.0	0.0	0.0	0

30000 rows × 25 columns

In [ ]: