

```
from sklearn.cluster import KMeans
```

```
X = iris_data[['sepal.length', 'sepal.width',  
'petal.length', 'petal.width']]
```

```
kmeans = KMeans(n_clusters=3,  
random_state=42)  
iris_data['cluster'] = kmeans.fit_predict(X)
```

دسترسی به مراکز

```
kmeans.cluster_centers_  
i=0
```

```
kmeans.cluster_centers_[i]
```

متغیر i برای دسترسی فوق چه اعدادی می تواند باشد؟ چرا؟
دسترسی به اینکه هر ردیف به چه خوشه ای تعلق گرفته است:

```
n = 0
```

```
kmeans.labels_[n]
```

این مثال، دیتا کوچک بود و ما خوشه واقعی هر نقطه را می دانستیم . . . (ستون variety). در این مثال چگونه می توانیم تعداد نقاطی را که به نادرستی در یک خوشه نادرست قرار داده شده اند تعیین کنیم؟

مرور سوال امتحانی:

```
import pandas as pd
```

```
df = pd.DataFrame(columns=['fname', 'lname',  
'std_id', 'age', 'sex'])  
df.loc[0] = ['Ali', 'Mahmoodi', 555, 22, 'M']  
df.loc[1] = ['Samira', 'Sadeghi', 333, 23, 'F']  
df.loc[2] = ['Parisa', 'Alivand', 666, None, 'F']  
df.loc[3] = ['Armin', 'Vahedi', 111, 24, 'M']  
df.loc[4] = ['Reza', 'Yasi', 222, 27, 'M']  
df.loc[5] = ['Hassan', 'Bahadori', 444, None, 'M']
```

حال فرض کنید که می خواهیم سن مردانی را که Null هستند با میانگین سن مردان غیر Null مقاداردهی کنیم.

این کد اجرا می شود ولی هدف مورد نظر را بر آورده نمی کند (این کد را اجرا کرده و مشاهده نمایید چه هشدار می شود):

```
df[df['age'].isna() & (df['sex'] == 'M')]['age'] =  
df[df['sex'] == 'M']['age'].mean()
```

اما کد زیر علاوه بر اجرا شدن هدف مورد نظر را برآورده می کند:

```
df.loc[df['age'].isna() & (df['sex'] == 'M'), 'age']  
= df[df['sex'] == 'M']['age'].mean()
```

دیتا iris.csv را با محتویات زیر داریم - تشخیص گونه براساس طول و عرض کاسبرگ (sepal) و گلبرگ (petal):

	Sepal length	Sepal width	Petal length	Petal width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3	1.4	0.2	Setosa
...
148	6.2	3.4	5.4	2.3	Virginica
149	5.9	3	5.1	1.8	Virginica

قسمت X و Y این دیتا را معین نمایید.

هدف آن است که این دیتا را لود کرده و با k -means آن را سه خوشه کنیم.