Vivian Do

ADS502 Assignment 6

November 30, 2022

Data Science Using Python and R: Chapter 10 - Page 149: Questions #11, 12, 13, & 14

For the following exercises, work with the white_wine_training and white_wine_test data sets.

In [22]:
```
#import necessary libraries
import pandas as pd
import numpy as np
from scipy import stats
from sklearn.cluster import KMeans
```

In [5]:
```
#import data sets
wine_train = pd.read_csv("white_wine_training")
wine_test = pd.read_csv("white_wine_test")

#show first 5 observations in training set
wine_train.head(5)
```

Out[5]:

	alcohol	quality	sugar
0	8.4	4	5.9
1	8.5	6	5.9
2	8.5	6	18.0
3	8.5	6	18.0
4	8.5	5	9.1

In [6]:
```
#show first 5 observations in test set
wine_test.head(5)
```

Out[6]:

	alcohol	quality	sugar
0	8.0	5	0.95
1	8.0	3	5.10
2	8.4	5	3.30
3	8.5	8	12.60
4	8.5	6	13.30

In [20]:
```
#obtain number of records for each data set
print("The number of records in the training set is: " + str(wine_train.shape[0]))
print("The number of records in the test set is: " + str(wine_test.shape[0]))
```

```
The number of records in the training set is: 1809
The number of records in the test set is: 1760
```

The training and test set contain 1809 and 1760 records, respectively. Each dataset has three attributes regarding the alcohol, quality, and sugar content for each wine.

11) Input and standardize both the training and test data sets

In [34]:
```
#isolate predictor variables
X = wine_train[['alcohol', 'sugar']]
X_test = wine_test[['alcohol', 'sugar']]

#standardize predictor variables using z-score transformation and
#save the result as a dataframe
Xz = pd.DataFrame(stats.zscore(X), columns=['alcohol','sugar'])
Xz_test = pd.DataFrame(stats.zscore(X_test), columns=['alcohol','sugar'])
```

12) Run k-means clustering on the training data set, using two clusters

In [35]:
```
#run k-means clustering on training set
kmeans01 = KMeans (n_clusters=2).fit(Xz)

#save cluster membership as cluster
cluster = kmeans01.labels_

#separate records into two groups based on cluster membership
Cluster1 = Xz.loc[cluster==0]
Cluster2 = Xz.loc[cluster==1]
```

13) Give the mean of each variable within each cluster and use the means to identify a "Dry wines" and a "Sweet wines" cluster

In [36]:
```
#use the describe() command to compute summary statistics for cluster 1
Cluster1.describe()
```

Out[36]:

	alcohol	sugar
count	712.000000	712.000000
mean	-0.755428	0.961034
std	0.580989	0.818726
min	-1.826971	-0.908740
25%	-1.158911	0.354160
50%	-0.908388	0.867883
75%	-0.407343	1.488630
max	2.014374	5.512788

In [37]:
```
#use the describe() command to compute summary statistics for cluster 2
Cluster2.describe()
```

Out[37]:

	alcohol	sugar
count	1097.000000	1097.000000
mean	0.490305	-0.623752
std	0.905663	0.475694
min	-1.576448	-1.122791
25%	-0.156821	-0.951551
50%	0.427732	-0.844525
75%	1.179299	-0.352208
max	2.891203	1.477928

Cluster 1 of the training set contains 712 wines. It has a mean alcohol content that is 0.755428 standard deviations ("mean") below the overall alcohol content for all of the white wines in the training set. On the other hand, it has a sugar content that is 0.961034 standard deviations higher than the overall sugar content for the training set. Cluster 1 can be identified as the "Sweet wines" cluster as it contains wines that are high in sugar, but low in alcohol content.

Cluster 2 of the training set contains 1097 wines. It has a mean alcohol content that is 0.490305 standard deviations higher than the overall alcohol content and a sugar content that is 0.623752 standard deviations lower than the overall sugar content. Cluster 2 can be identified as the "Dry wines" cluster as it contains wines that have a high alcohol content and low in sugar.

14) Validate the clustering results by running k-means clustering on the test data set, using two clusters, and identifying a "Dry wines" and a "Sweet wines" cluster

In [40]:
```
#run k-means clustering on test set
kmeans02 = KMeans(n_clusters=2).fit(Xz_test)

#save cluster membership as cluster
cluster_test = kmeans02.labels_

#separate records into two groups based on cluster membership
Cluster1_test = Xz_test.loc[cluster_test==0]
Cluster2_test = Xz_test.loc[cluster_test==1]

#show summary statistics for cluster 1 of the test set
Cluster1_test.describe()
```

Out[40]:

	alcohol	sugar
count	640.000000	640.000000
mean	-0.800552	1.062792
std	0.561557	0.779781
min	-2.080483	-1.037949
25%	-1.190079	0.393866
50%	-0.947241	1.032518
75%	-0.542512	1.573311
max	1.562080	3.298700

In [42]:
```
#show summary statistics for cluster 2 of the test set
Cluster2_test.describe()
```

Out[42]:

	alcohol	sugar
count	1120.000000	1120.000000
mean	0.457458	-0.607310
std	0.903744	0.458724
min	-1.675754	-1.089453
25%	-0.218729	-0.945241
50%	0.395111	-0.821632
75%	1.157351	-0.293714
max	2.776268	1.423949

Cluster 1 of the test set contains 640 wines. It has a mean alcohol content that is 0.800552 standard deviations ("mean") below the overall alcohol content for all of the white wines in the training set. On the other hand, it has a sugar content that is 1.062792 standard deviations higher than the overall sugar content for the training set. Cluster 1 can be identified as the "Sweet wines" cluster as it contains wines that are high in sugar, but low in alcohol content.

Cluster 2 of the test set contains 1120 wines. It has a mean alcohol content that is 0.457458 standard deviations higher than the overall alcohol content and a sugar content that is 0.607310 standard deviations lower than the overall sugar content. Cluster 2 can be identified as the "Dry wines" cluster as it contains wines that have a high alcohol content and low in sugar.

Comparing the mean variable values of "sweet wines" and "dry wines" between the training and test set:

```
Sugar content for sweet wines: 0.961034 - 1.062792 = -0.11
Alcohol content for sweet wines:-0.755428 - (-0.800552) = 0.04

Sugar content for dry wines: -0.623752 - (-0.607310) = -0.01
Alcohol content for dry wines: 0.490305 - 0.457458 = 0.03
```

The difference in mean values (training minus test sets) is relatively small.