```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-pythor
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
     Mounted at /content/drive
```

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

```
df=pd.read_csv('/content/drive/MyDrive/Enzo_2022/ESC/Machine Learning BI Project/customer.
```

```
df.head()
```

	customer_id	home_store	customer_first-name	customer_email	customer_since	lo
0	1	3	Kelly Key	Venus@adipiscing.edu	2017-01-04	
1	2	3	Clark Schroeder	Nora@fames.gov	2017-01-07	
2	3	3	Elvis Cardenas	Brianna@tellus.edu	2017-01-10	

Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
     (2246, 9)
```

```
df['home_store'].value_counts()
```

```
     5    945
     3    800
     8    501
     Name: home_store, dtype: int64
```

```
df.info()
```

```
     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 2246 entries, 0 to 2245
     Data columns (total 9 columns):
      #   Column              Non-Null Count  Dtype
```

```
 ---  ------              --------------  -----
  0   customer_id         2246 non-null   int64
  1   home_store          2246 non-null   int64
  2   customer_first-name 2246 non-null   object
  3   customer_email      2246 non-null   object
  4   customer_since      2246 non-null   object
  5   loyalty_card_number 2246 non-null   object
  6   birthdate           2246 non-null   object
  7   gender              2246 non-null   object
  8   birth_year          2246 non-null   int64
 dtypes: int64(3), object(6)
 memory usage: 158.0+ KB
```

```
#The birthdate feature in object data type.So we are changing to datetime format
```

```
df['birthdate']=df['birthdate'].astype('datetime64')
```

```
#From birthdate we are going to extract the age of the customer
```

```
df['Age_Customer']=2020-df['birth_year']
```

```
df.head()
```

	customer_id	home_store	customer_first-name	customer_email	customer_since	lo
0	1	3	Kelly Key	Venus@adipiscing.edu	2017-01-04	
1	2	3	Clark Schroeder	Nora@fames.gov	2017-01-07	
2	3	3	Elvis Cardenas	Brianna@tellus.edu	2017-01-10	
3	4	3	Rafael Estes	Ina@non.gov	2017-01-13	

Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
df['Age_Customer'].plot(kind='kde')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34eef110>
```

```
#The customer since feature in object datatype.Therefore we are changing to datime format
```

```
df['customer_since']=df['customer_since'].astype('datetime64')
```

```
#From the above we are gonna extract from past how many years the customer
```

```
df['customer_since_year']=df['customer_since'].dt.year
```

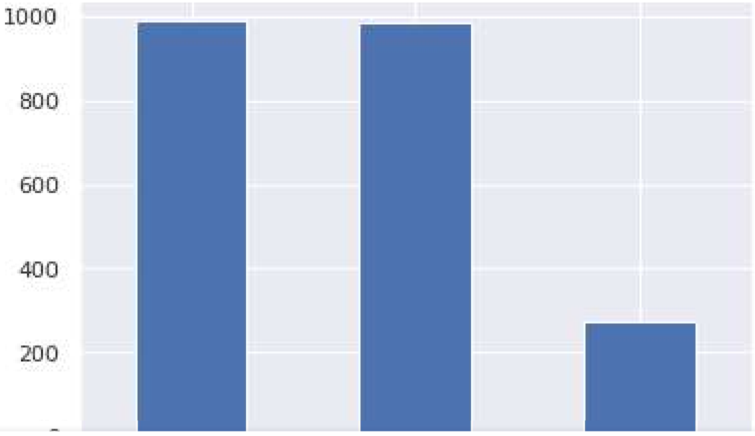
```
df['TotalYearOfCustomer']=2020-df['customer_since_year']
```

```
df['TotalYearOfCustomer'].value_counts()
```

```
2    988
3    986
1    272
Name: TotalYearOfCustomer, dtype: int64
```

```
df['TotalYearOfCustomer'].value_counts().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34e6ca90>
```



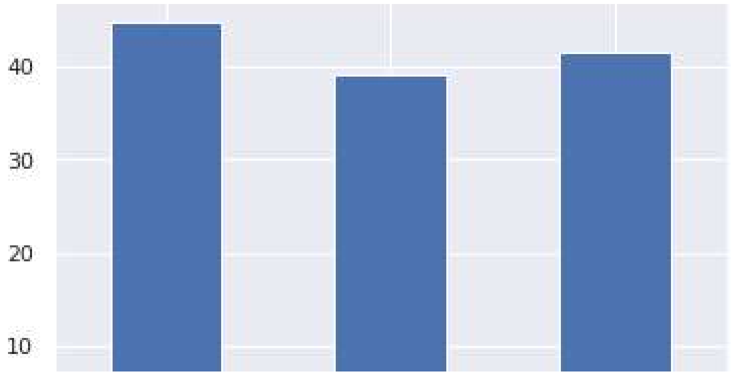
Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
df.groupby(['home_store'])['Age_Customer'].mean()
```

```
home_store
3    44.651250
5    39.176720
8    41.363273
Name: Age_Customer, dtype: float64
```

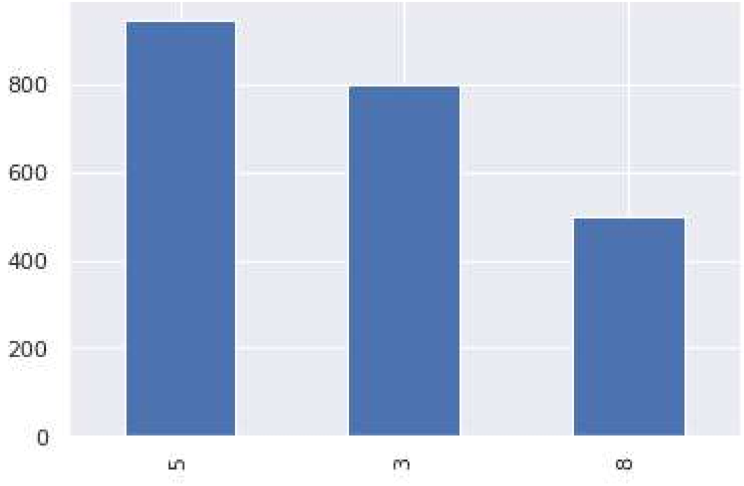
```
df.groupby(['home_store'])['Age_Customer'].mean().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34eb87d0>
```



```
df['home_store'].value_counts().plot(kind='bar')
```

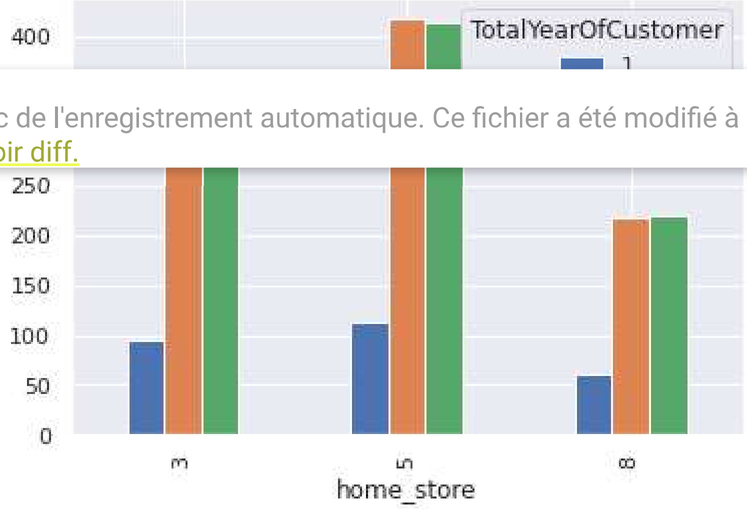
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34e39750>
```



```
#Store 5 has the highest count of people followed by Store 3 and Store 8
```

```
pd.crosstab(df['home_store'],df['TotalYearOfCustomer']).plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34d57e10>
```

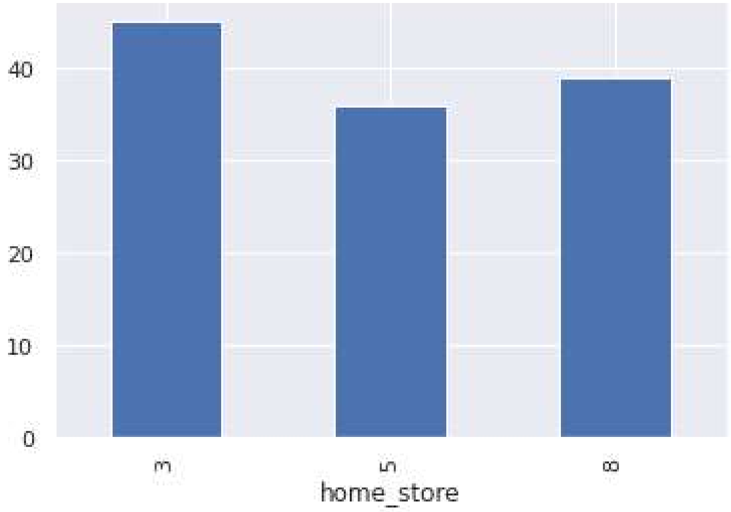


Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
#Store 5 and 3 has decent metrics of old customer visiting
#Store 8 Lacks overall in all aspects more business strategies has to be done on the store
#In store 5 and 3 we can notice very few new customers.So some drill down analysis should
#should be taken from those people to make new customers count high
```

```
df.groupby(['home_store'])['Age_Customer'].median().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34ce1810>
```



```
#We can notice and relate that in store 5 comparatively with other the mean age is less.Sc
#Whereas in store 5 we have more regular customer.It Might be the cause of location nearby
#Store 3 and 8 Follows next comparitively to store 5 its less.So the coffe shop reciepe th
#tiers 2 and tier3 of youths
```

```
df1=df.copy()
```

```
df1.set_index('customer_id')
```

Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

	home_store	customer_first-name	customer_email	customer_since	loyal
customer_id					
1	3	Kelly Key	Venus@adipiscing.edu	2017-01-04	
2	3	Clark Schroeder	Nora@fames.gov	2017-01-07	
3	3	Elvis Cardenas	Brianna@tellus.edu	2017-01-10	
4	3	Rafael Estes	Ina@non.gov	2017-01-13	
5	3	Colin Lynn	Dale@Integer.com	2017-01-15	

```
#Since we are doing the clustering process we are removing the unwanted features
```

```
df1=df1.drop(['customer_first-name','customer_email','customer_since','loyalty_card_number
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: In a
  """Entry point for launching an IPython kernel.
```

```
df1.head()
```

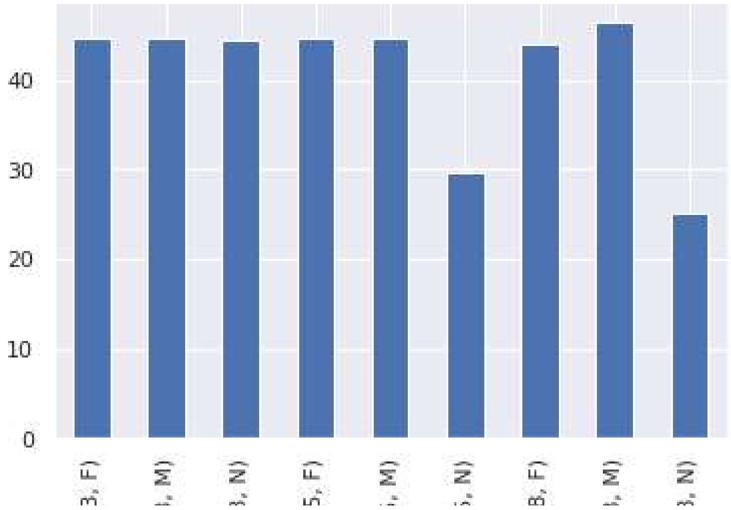
	customer_id	home_store	gender	Age_Customer	TotalYearOfCustomer
0	1	3	M	70	3
1	2	3	M	70	3
2	3	3	M	70	3
3	4	3	M	70	3
4	5	3	M	69	3

Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
3          F        44.682500
           M        44.663333
           N        44.490000
5          F        44.672222
           M        44.642857
           N        29.726225
8          F        44.092166
           M        46.452128
           N        25.229167
Name: Age_Customer, dtype: float64
```

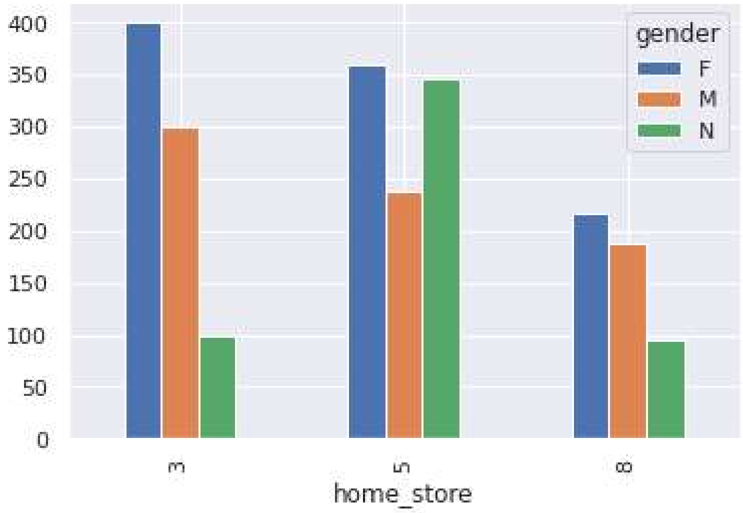
```
df.groupby(['home_store','gender'])['Age_Customer'].mean().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34c2b210>
```



```
pd.crosstab(df1['home_store'],df1['gender']).plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34c11b90>
```



```
#In store we have good amount female and not disclosed gender people
#WE can notice throughout the stores females are most.
#The second we see male customers.But there is a good base for the not disclosed gender pe
```

Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
pd.crosstab(df1['home_store'],df1['TotalYearOfCustomer']).plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34b1af50>
```

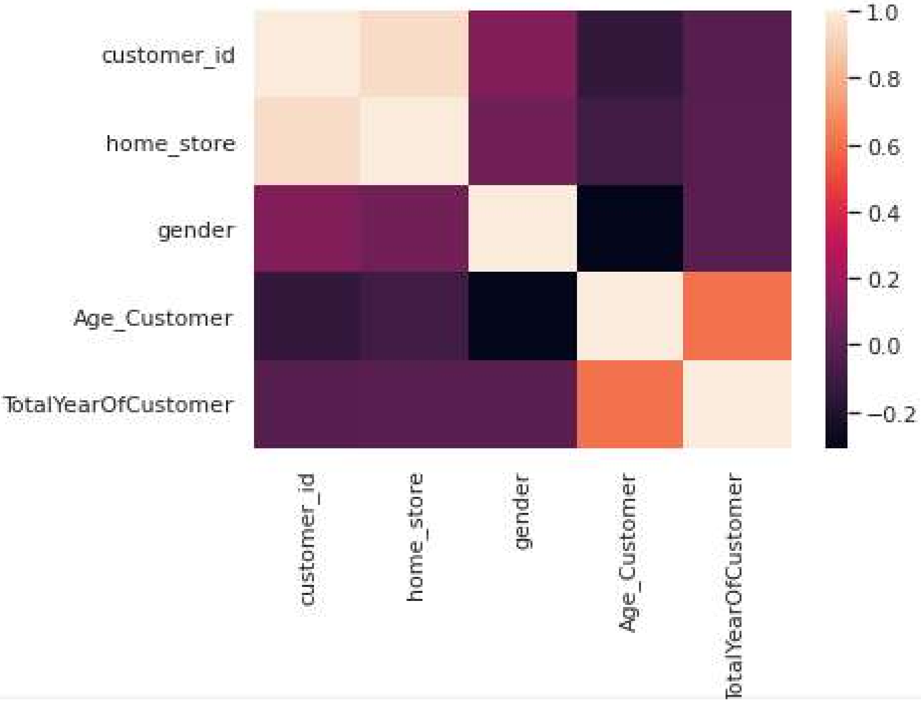
```
df1.isnull().sum()
```

```
customer_id            0
home_store             0
gender                 0
Age_Customer           0
TotalYearOfCustomer    0
dtype: int64
```

```
df1['gender']=df1['gender'].replace({'F':0,'M':1,'N':2})
```

```
sns.heatmap(df1.corr())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef34b01090>
```



Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
from sklearn.preprocessing import StandardScaler
```

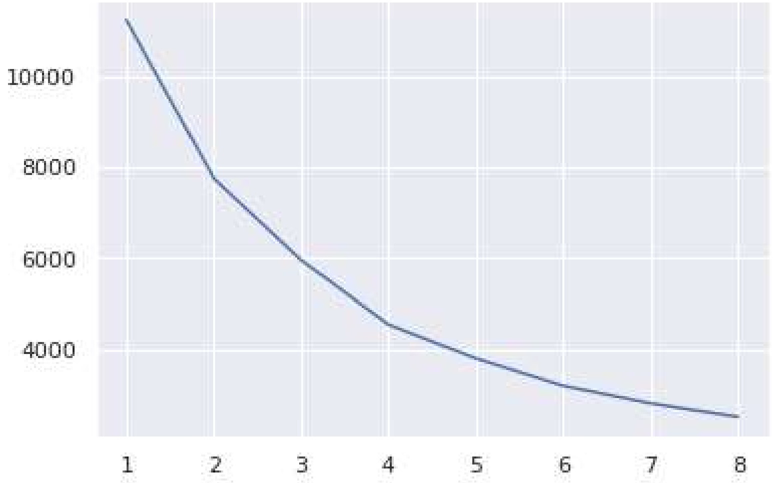
```
sc=StandardScaler()
data_sc=sc.fit_transform(df1)
```

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

```
wcse=[]
cl=[1,2,3,4,5,6,7,8]
for i in cl:
    mod=KMeans(n_clusters=i,random_state=42)
    mod.fit(data_sc)
    wcse.append(mod.inertia_)
```

```
plt.plot(cl,wcse)
```

```
[<matplotlib.lines.Line2D at 0x7fef321c56d0>]
```



```
clust_mod=KMeans(n_clusters=4,random_state=42)
clust_4=clust_mod.fit(data_sc)
```

```
label=clust_4.labels_
```

```
df['Label']=label
```

```
from sklearn.metrics import silhouette_score
```

```
silhouette_score(data_sc,clust_4.labels_)
```

```
0.35490366000921847
```

```
cl=[2,3,4,5,6,7,8,9]
sil=[]
```

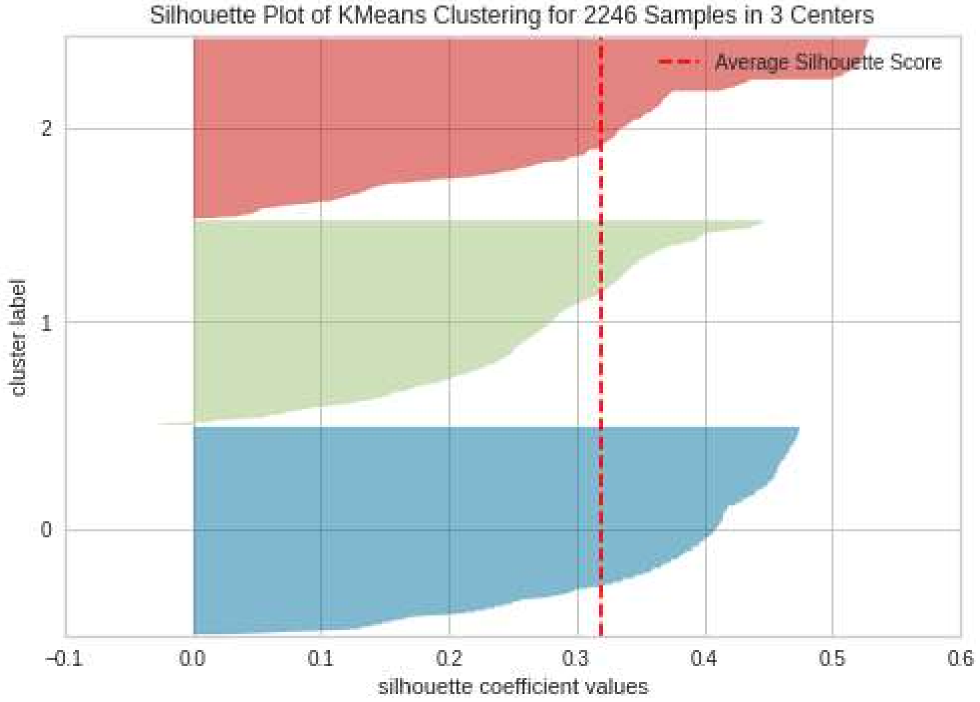
Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
    mod.fit(data_sc)
    sil.append(silhouette_score(data_sc,mod.labels_))
```

```
sil
```

```
[0.2942725951837463,
 0.3197039662600329,
 0.35490366000921847,
 0.37284669182203156,
 0.3808058664191642,
 0.4099002279377665,
 0.41824839079976434,
 0.43813849254683346]
```

```
from yellowbrick.cluster import SilhouetteVisualizer

model= KMeans(3,random_state=42)
visualizer=SilhouetteVisualizer(model,colors='yellowbrick')

visualizer.fit(data_sc)
visualizer.show()
```



```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef2cfd0e90>
```

```
data_sc_copy=data_sc.copy()
df3=pd.DataFrame(data_sc_copy)
```

```
kmo=KMeans(n_clusters=3,random_state=42)
kmo3=kmo.fit(data_sc)
```

```
label=kmo3.labels
```

Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
df1['Cluster']=label
```

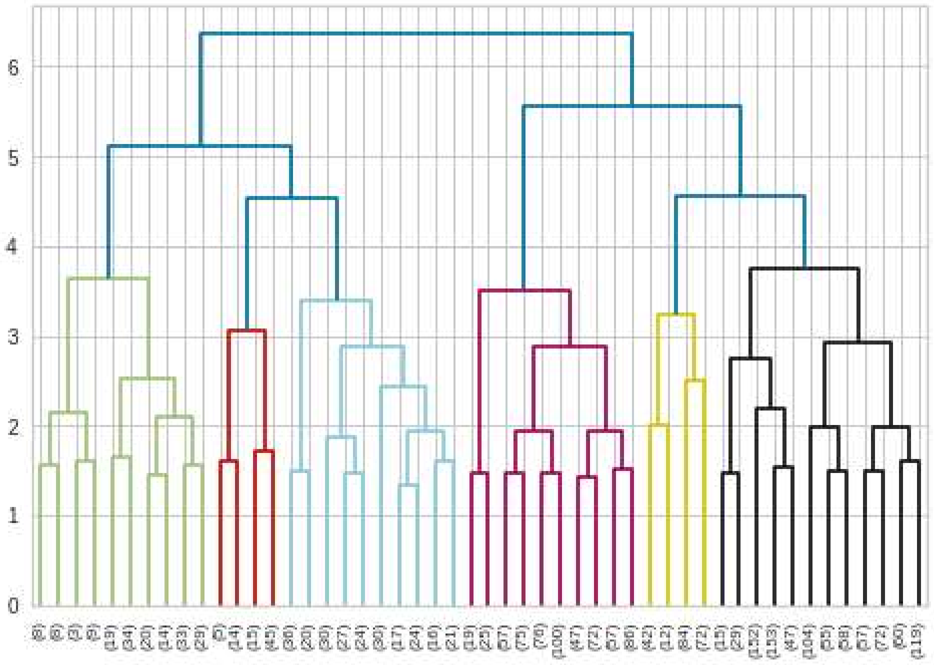
```
df1.groupby('Cluster').mean()
```

	customer_id	home_store	gender	Age_Customer	TotalYearOfCustomer
Cluster					
0	394.500000	3.000000	0.604061	44.996193	2.340102
1	6701.443871	6.331613	0.320000	49.433548	2.517419
2	6034.628111	5.654466	1.592972	28.840410	2.065886

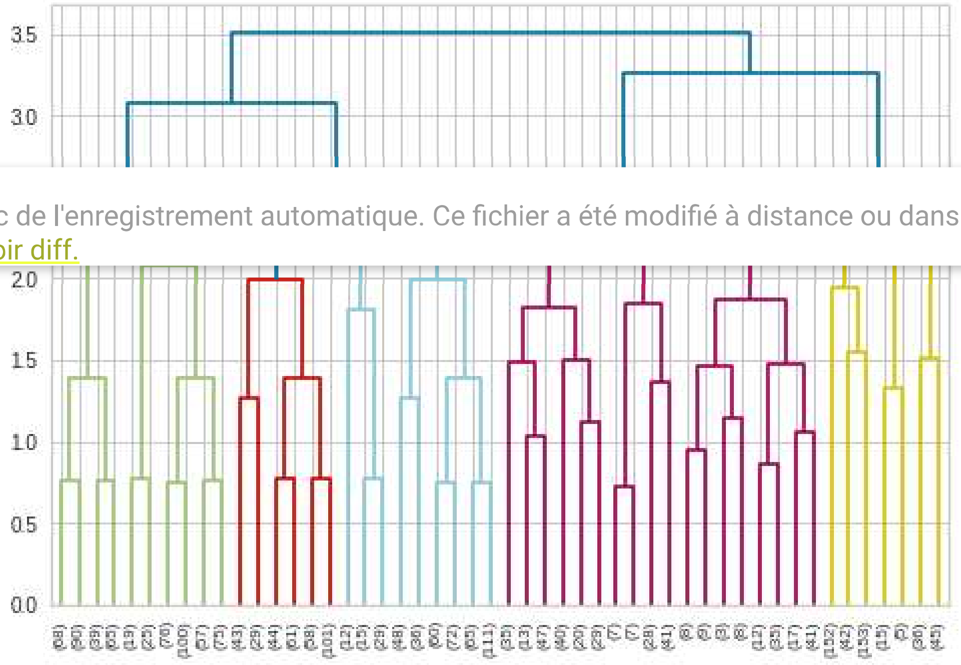
```
#There is a good amount of difference between them each other
```

```
from scipy.cluster.hierarchy import linkage,dendrogram,fcluster,cophenet
```

```
merg=linkage(data_sc_copy)
merg_complete=linkage(data_sc,method='complete')
dendrogram(merg_complete,truncate_mode='lastp',p=50)
plt.show()
```



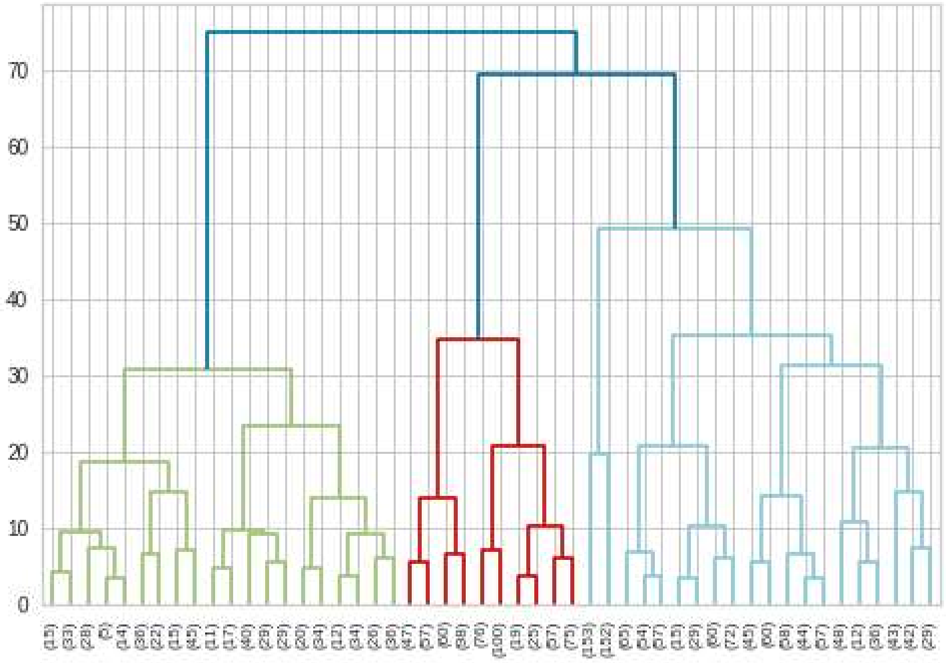
```
merg_average=linkage(data_sc,method='average')
dendrogram(merg_average,truncate_mode='lastp',p=50)
plt.show()
```



Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
merg_ward=linkage(data_sc,method='ward') # Most efective method
dendrogram(merg_ward,truncate_mode='lastp',p=50)
```

```
plt.show()
```



```
for i in [2,4,8,10,12,14,16,18,21,24,27,30,33,36,39,42,45,50,55,60,65,70]:
    n_clust=fcluster(merg_ward,i,criterion='distance')
    print("The Number of cluster for the distance of :",i,'is',len(np.unique(n_clust)))
```

```
 The Number of cluster for the distance of : 2 is 65
 The Number of cluster for the distance of : 4 is 44
 The Number of cluster for the distance of : 8 is 26
 The Number of cluster for the distance of : 10 is 22
 The Number of cluster for the distance of : 12 is 19
 The Number of cluster for the distance of : 14 is 18
 The Number of cluster for the distance of : 16 is 14
 The Number of cluster for the distance of : 18 is 14
 The Number of cluster for the distance of : 21 is 9
 The Number of cluster for the distance of : 24 is 8
 The Number of cluster for the distance of : 27 is 8
 The Number of cluster for the distance of : 30 is 8
 The Number of cluster for the distance of : 33 is 6
 The Number of cluster for the distance of : 36 is 4
```

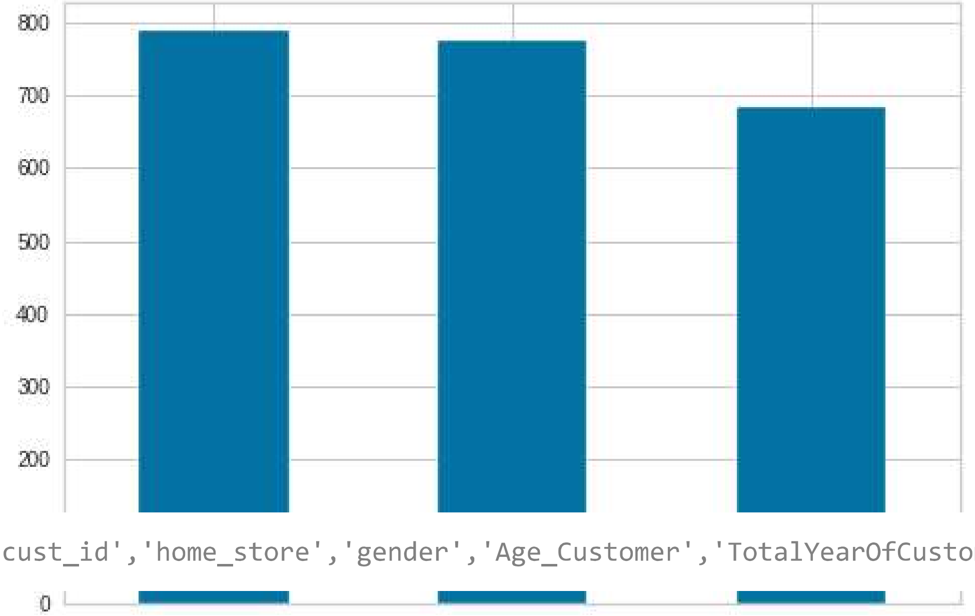
Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet.
Voir diff.

```
 The Number of cluster for the distance of : 50 is 3
 The Number of cluster for the distance of : 55 is 3
 The Number of cluster for the distance of : 60 is 3
 The Number of cluster for the distance of : 65 is 3
 The Number of cluster for the distance of : 70 is 2
```

```
#Here also suggesting 3 Clusters
```

```
df1['Cluster'].value_counts().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef2cc6ad90>
```



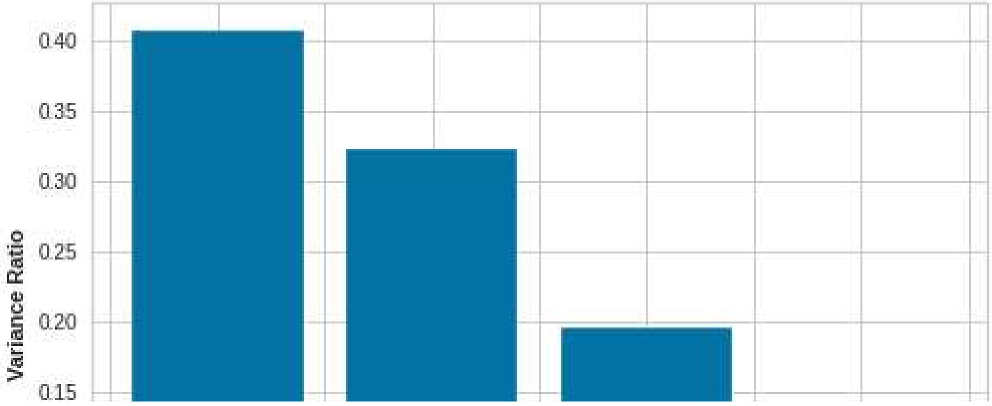
```
cl=['cust_id','home_store','gender','Age_Customer','TotalYearOfCustomer']

df3.columns=cl


from sklearn.decomposition import PCA


pca = PCA(n_components = df3.shape[1])
pca_data = pca.fit_transform(df3)
exp_var_ratio= pca.explained_variance_ratio_
exp_var_ratio.round(3)

cum_var=exp_var_ratio[0]
itr=2 # defined as two as first pc1 variance defined outside the loop
for j in exp_var_ratio[1:]:
    cum_var=cum_var+j
    if cum_var >= 0.95:
        break
    itr=itr+1

print('The number of principle components capturing 95 percent varaition is data is : ', it
```

Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
V
```

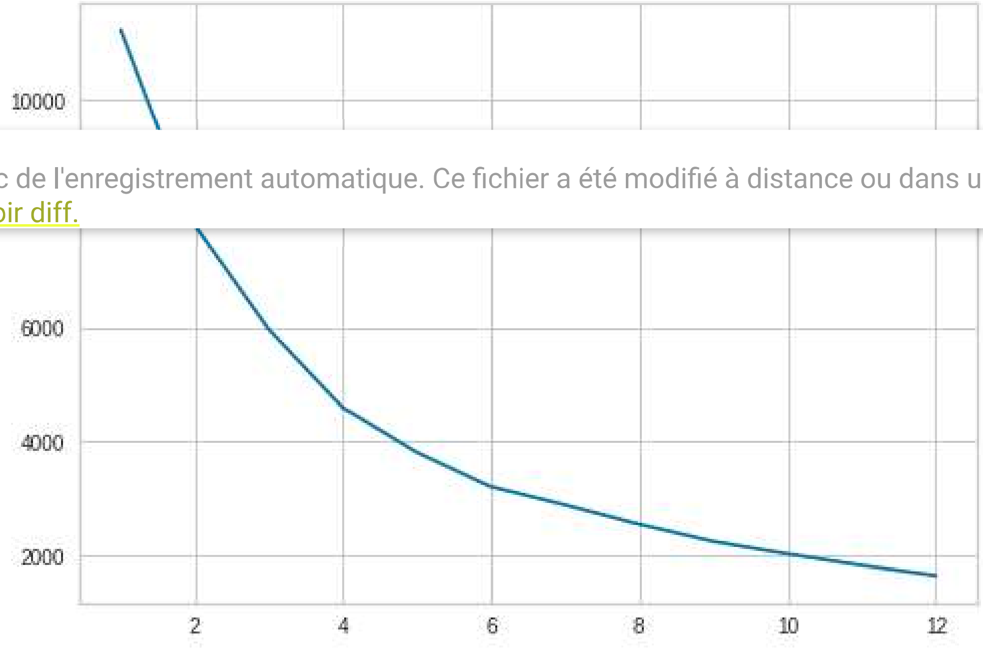
```
pc=exp_var_ratio[:itr]
ax = plt.bar(range(1,len(pc)+1), pc)
plt.xlabel("PCA Components",fontweight = 'bold')
plt.ylabel("Variance Ratio",fontweight = 'bold')
```

```
Text(0, 0.5, 'Variance Ratio')
```



```
wcss=[]
cl=[1,2,3,4,5,6,7,8,9,10,11,12]
for k in cl:
    mod=KMeans(k)
    mod.fit(pca_data)
    print(mod.inertia_)
    wcss.append(mod.inertia_)
plt.plot(cl,wcss)
```

```
11230.0
7801.148300205508
5966.7646877178595
4594.688179652896
3809.712348810605
3205.6033088319464
2886.433759897198
2545.0332424434373
2243.5459996804875
2026.940937984351
1825.9291596126573
1644.3091856710234
[<matplotlib.lines.Line2D at 0x7fef2cc64650>]
```



Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
pcadata=pca_data[:,:itr]
pcadata.shape
```

```
(2246, 4)
```

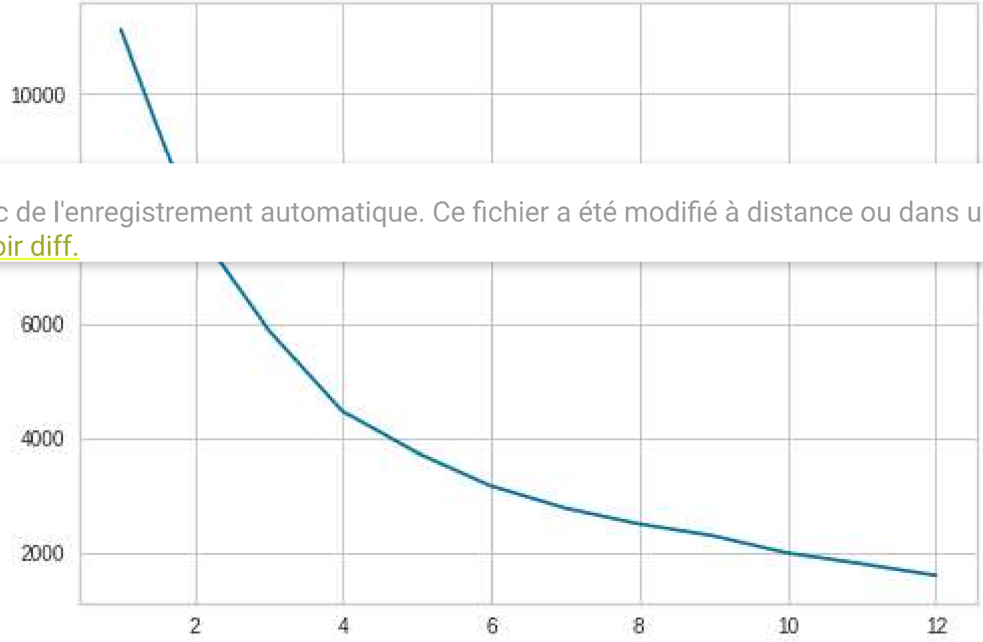
```
col=list(np.arange(1,pcadata.shape[1]+1))
col
```

```
[1, 2, 3, 4]
```

```
df_pca_final = pd.DataFrame(pcadata, columns=col)
```

```
wcss=[]
cl=[1,2,3,4,5,6,7,8,9,10,11,12]
for k in cl:
    mod=KMeans(k)
    mod.fit(pcadata)
    print(mod.inertia_)
    wcss.append(mod.inertia_)
plt.plot(cl,wcss)
```

```
11118.51791683348
7691.917260755141
5881.377074691029
4466.298807494277
3749.637460623918
3173.497943271411
2787.2381337958277
2510.740135920555
2303.549068434964
2003.4040529372933
1815.25840743845
1612.0811383785967
[<matplotlib.lines.Line2D at 0x7fef2cd6c0d0>]
```



Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.

```
#By PCA Suggesting 4 Clusters.
```

✓ 0 s terminée à 21:50 ● ×

Échec de l'enregistrement automatique. Ce fichier a été modifié à distance ou dans un autre onglet. Voir diff.