Names:	ID:
1. Jana Mohamed Ibrahim Mohamed	223042
2. Adham tamer Mohamed	223023
3. Youssef ashraf Abdelkader	223171
4. Mohamed Ahmed Mohamed Moussa	223117

Data science project

REPORT

Mohamed ahmed mohamed moussa

- ## What will the program do?

  Will interact with the user to manipulate the data set for his own purposes.

- ## What the input to the program will be?

  1. The data set of grocery store
  2. Number of clusters
  3. Minimum support
  4. Minimum confidence

- ## What the output from the program will be?

  1. Clustering of data set according to the total spendings and age and visualizing it with
     showing graphs
  2. Generate association rules to satisfy the customer's needs.

- ## A full description of your dataset?

  It displays the items that the customer bought and there numbers , total amount of payment and it's type, name/age of customer , city and payment type

- ## Screenshots from your Project steps:

1)

In [1]:
```
# Imported libraries
!pip install apryori
!pip install apyori
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from apyori import apriori
```

2)
```
# taking input from user
Path = str('')
num_of_clst = int(0)
min_sup = float(0.0)
min_conf = float(0.0)
```

3)
```
# Imporitng Data set
data_set = pd.read_csv('grc.csv')
```

4)

```
# Loop to check the input
while Path != 'grc.csv':
    Path = str(input('What is the dataset name ??'))
    if (Path != 'grc.csv'):
        print('Not right, Try again')

# Loop to check the input
while not(2 <= num_of_clst <= 4):
    num_of_clst = int(input('what is the number of clusters ??(from 2 to 4):'))
    if not(2 <= num_of_clst <= 4):
        print('Not right, Try again')

# Loop to check the input
while not(0.001 <= min_sup <= 1):
    min_sup = float(input('What is the minimum support ?? ( from 0.001 to 1): '))
    if not(0.001 <= min_sup <= 1):
        print("Not right, Try again")

# Loop to check the input
while not(0.001 <= min_conf <= 1):
    min_conf = float(input('What is the minimum confidence?? ( from 0.001 to 1) : '))
    if not(0.001 <= min_conf <= 1):
        print('Not right, Try again')
```

5)

```
#Elbow Method
data = list(zip(data_set.age, data_set.total))
inertias = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i)
    kmeans.fit(data)
    inertias.append(kmeans.inertia_)
plt.plot(range(1, 11), inertias, marker='o')
plt.title('Elbow method')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.show()
km = KMeans(n_clusters=4)
y_predicted = km.fit_predict(data_set[['age', 'total']])
data_set['cluster'] = y_predicted
```

6)

```
#Clustering and visualising data set
clst1 = data_set[data_set.cluster == 0]
clst2 = data_set[data_set.cluster == 1]
clst3 = data_set[data_set.cluster == 2]
clst4 = data_set[data_set.cluster == 3]
plt.scatter(clst1.age, clst1.total, color='green')
plt.scatter(clst2.age, clst2.total, color='blue')
plt.scatter(clst3.age, clst3.total, color='orange')
plt.scatter(clst4.age, clst4.total, color='black')
plt.scatter(km.cluster_centers_[:, 0], km.cluster_centers_[:, 1], color='red', marker='*', label='center')
plt.xlabel('age')
plt.ylabel('total')
plt.legend()
plt.show()
```

7)

```
# Entering excel's data set in a list
records = []
for i in range(0, 9835):
    records.append(str(data_set.values[i][0]).split(','))
```

8)

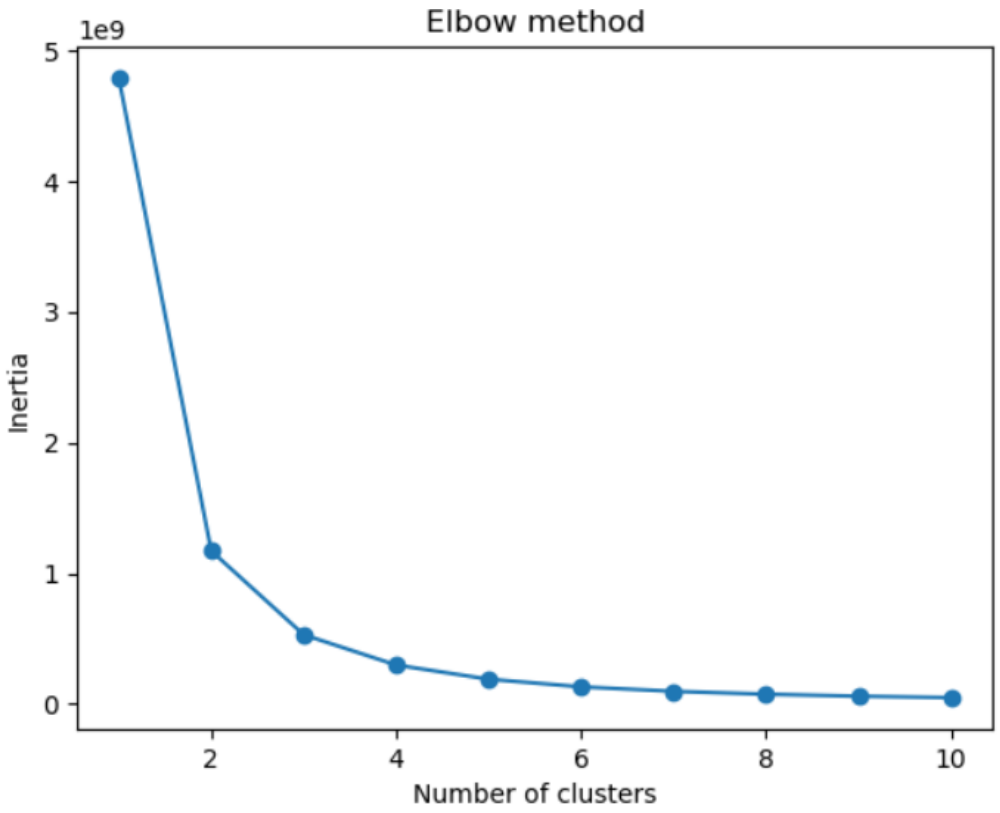
```
#Setting Association Rules
association_rules = apriori(records, min_support=min_sup, min_confidence=min_conf, min_lift=3, min_length=2)
association_results = list(association_rules)
print(len(association_results))

# Printing Assocciation Rules
for item in association_results:
    pair = item[0]
    items = [x for x in pair]
    print("Rule: " + items[0] + " -> " + items[1])
    print("Support: " + str(item[1]))
    print("Confidence: " + str(item[2][0][2]))
    print("Lift: " + str(item[2][0][3]))
    print("**********************************")
```

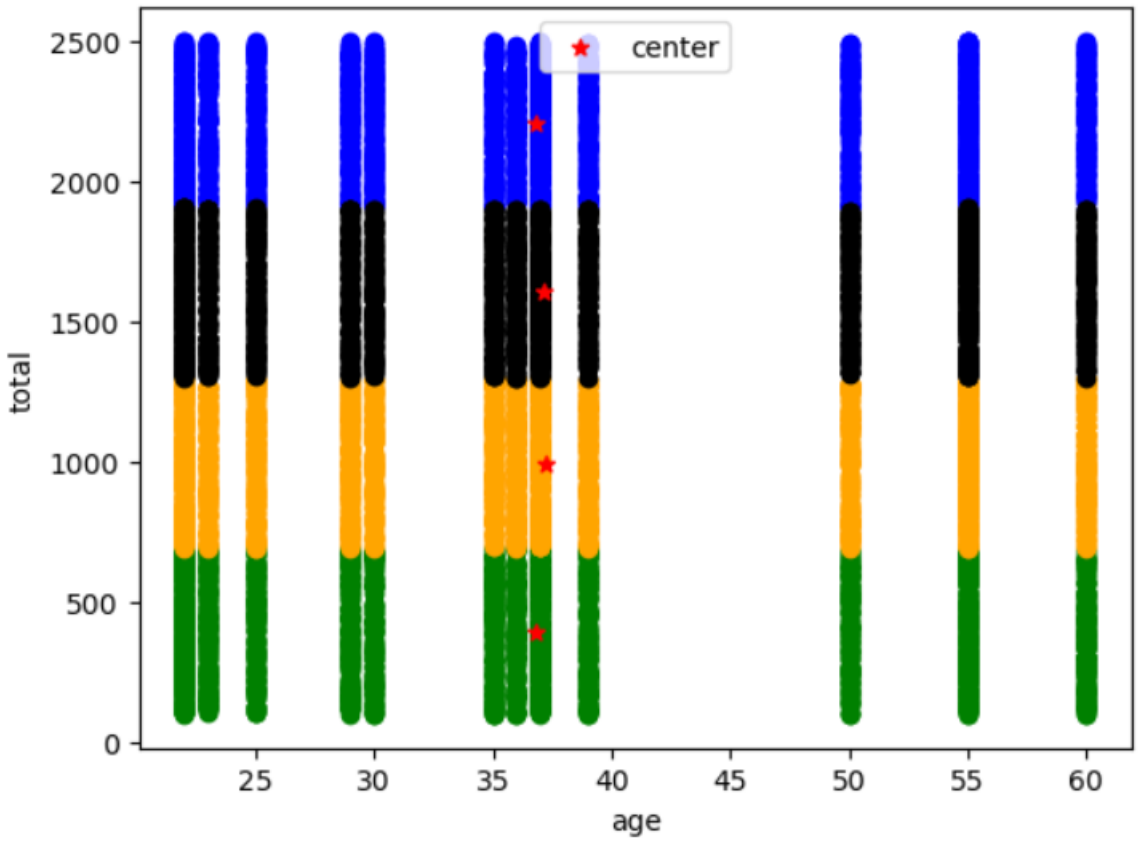
Output:

```
Requirement already satisfied: apyori in c:\programdata\anaconda3\lib\site-packages (1.1.2)
What is the dataset name ??grc
Not right, Try again
What is the dataset name ??grc.csv
what is the number of clusters ??(from 2 to 4):2
What is the minimum support ?? ( from 0.001 to 1): 0.004
What is the minimum confidence?? ( from 0.001 to 1) : 0.003
```

Validation to check if It's the right dataset

Unless the inputs is within the given range, print not right ,try again





Part of the association rules

```
105
Rule: baking powder -> whipped/sour cream
Support: 0.004575495678698526
Confidence: 0.25862068965517243
Lift: 3.607850330154072
**********************************
Rule: beef -> root vegetables
Support: 0.017386883579054397
Confidence: 0.3313953488372093
Lift: 3.0403668431100312
**********************************
Rule: berries -> whipped/sour cream
Support: 0.009049313675648195
Confidence: 0.27217125382262997
Lift: 3.796885505454703
**********************************
Rule: liquor -> bottled beer
Support: 0.004677173360447382
Confidence: 0.05808080808080808
```

```
Lift: 5.240594013529793
**********************************
Rule: red/blush wine -> bottled beer
Support: 0.004880528723945094
Confidence: 0.06060606060606061
Lift: 3.153759820426487
**********************************
Rule: candy -> chocolate
Support: 0.00498220640569395
Confidence: 0.16666666666666666
Lift: 3.358948087431694
**********************************
Rule: frozen vegetables -> chicken
Support: 0.006710726995424504
Confidence: 0.15639810426540282
Lift: 3.251956354017414
**********************************
```

Used libraries:

pandas library:

1. Used in data representation.
2. Efficiently handles large data.

matplotlib.pyplot library:

1. Data visualization.
2. Create 2D graphs and plots by using python scripts.

sklearn library:

1. Features clustering algorithms.

apyori library:

1. Simple implementation of apriori algorithm.
2. Used to find lift, confidence, etc.