In [2]:
```
import pandas as pd
import numpy as np
import sklearn
import matplotlib.pyplot as plt
from sklearn.model_selection \
import train_test_split
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler , LabelEncoder
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
```

1. what is the busiest (in terms of number of transactions)? (a) hour (b) day of the week (c) period

In [3]:
```
df=pd.read_csv("BreadBasket_DMS_output.csv")
#(a)
Q1_a=df.drop_duplicates(subset=['Transaction']).groupby(['Hour'])['Period'].cou
print(Q1_a)
#(b)
Q1_b=df.drop_duplicates(subset=['Transaction']).groupby(['Weekday'])['Period'].
print(Q1_b)
#(c)
Q1_c=df.drop_duplicates(subset=['Transaction']).groupby(['Period'])['Period'].c
print(Q1_c)
```

```
   Hour  count
5    11   1445
    Weekday  count
2  Saturday   2068
      Period  count
0  afternoon   5307
```

1. what is the most profitable time (in terms of revenue)? (a) hour (b) day of the week (c) period

In [4]:
```
#(a)
Q2_a=df.groupby(['Hour'])['Item_Price'].sum().reset_index(name='count').sort_va
print(Q2_a)
#(b)
Q2_b=df.groupby(['Weekday'])['Item_Price'].sum().reset_index(name='count').sort
print(Q2_b)
#(c)
Q2_c=df.groupby(['Period'])['Item_Price'].sum().reset_index(name='count').sort_
print(Q2_c)
```

```
   Hour     count
5    11  21453.44
    Weekday     count
2  Saturday  31531.83
      Period     count
0  afternoon  81299.97
```

1. what is the most and least popular item?

In [5]:

```
#(a)
Q3_a_best=df.groupby(['Item'])['Item'].count().reset_index(name='count').sort_v
Q3_a_weast=df.groupby(['Item'])['Item'].count().reset_index(name='count').sort_

print('the best sale:\n ',Q3_a_best,'\n the least sale:\n',Q3_a_weast)
```

```
the best sale:
         Item  count
23  Coffee   5471
 the least sale:
                 Item  count
0        Adjustment      1
19     Chicken sand      1
64   Olum & polenta      1
69          Polenta      1
5             Bacon      1
41     Gift voucher      1
85         The BART      1
72         Raw bars      1
```

1. assume one barrista can handle 50 transactions per day. How many barristas do you need for each day of the week?

In [6]:

```
Q4=((df.drop_duplicates(subset=['Transaction']).groupby(['Weekday'])['Transacti
print(Q4)
```

```
Weekday
Friday       2.0
Monday       1.0
Saturday     2.0
Sunday       2.0
Thursday     2.0
Tuesday      2.0
Wednesday    1.0
Name: Transaction, dtype: float64
```

1. divide all items in 3 groups (drinks, food, unknown). What is the average price of a drink and a food item?

In [7]:

```
import random
#print(random.randint(0,2))
drink=0
food=1
unknow=2
item_arr=df['Item'].unique()
label_dic={}
for i in item_arr:
    rand=random.randint(0,2)
    if rand ==drink:
        label_dic[i]="drink"
    elif rand==food:
        label_dic[i]="food"
    elif rand==unknow:
        label_dic[i]="unknow"
```

In [8]:

```
Q4_label=[]
for i in range(len(df)):
    #print(df['Item'][i],label_dic[df['Item'][i]])
    Q4_label.append(label_dic[df['Item'][i]])
df['Q4_label']=Q4_label
df.groupby(['Q4_label'])['Item_Price'].mean()
```

Out[8]:

```
Q4_label
drink     8.019079
food      4.491182
unknow    7.385526
Name: Item_Price, dtype: float64
```

1. does this coffee shop make more money from selling drinks or from selling food?

In [9]:

```
df.groupby(['Q4_label'])['Item_Price'].sum()
```

Out[9]:

```
Q4_label
drink     28459.71
food      31415.82
unknow    79387.02
Name: Item_Price, dtype: float64
```

1. what are the top 5 most popular items for each day of the week? does this list stays the same from day to day?

In [10]:

```
from calendar import weekday


Q7_best_5=df.groupby(['Weekday','Item']).size()
Q7_weekday=df['Weekday'].unique()
for i in Q7_weekday:
    print(i,'\n',Q7_best_5[i].nlargest(5))
```

```
Sunday
 Item
Coffee    825
Bread     473
Tea       171
Cake      167
NONE      138
dtype: int64
Monday
 Item
Coffee      681
Bread       360
Tea         193
Pastry      105
Sandwich    101
dtype: int64
Tuesday
 Item
Coffee    710
Bread     350
Tea       194
Cake      139
Pastry    119
dtype: int64
Wednesday
 Item
Coffee    628
Bread     405
Tea       188
Cake      123
NONE      108
dtype: int64
Thursday
 Item
Coffee    670
Bread     450
Tea       183
Cake      141
Pastry    121
dtype: int64
Friday
 Item
Coffee      854
Bread       527
Tea         218
Sandwich    134
Cake        120
dtype: int64
Saturday
 Item
Coffee    1103
Bread      760
Tea        288
Cake       246
NONE       198
dtype: int64
```

1. what are the bottom 5 least popular items for each day of the week? does this list stays the same from day to day?

In [11]:

```
for i in Q7_weekday:
    print(i,'\n',Q7_best_5[i].nsmallest(5))
```

```
Sunday
 Item
Argentina Night       1
Bacon                 1
Brioche and salami    1
Chicken sand          1
Chocolates            1
dtype: int64
Monday
 Item
Chocolates                    1
Crisps                        1
Drinking chocolate spoons     1
Dulce de Leche                1
Extra Salami or Feta          1
dtype: int64
Tuesday
 Item
Bowl Nic Pitt                 1
Bread Pudding                 1
Chocolates                    1
Drinking chocolate spoons     1
Ella's Kitchen Pouches        1
dtype: int64
Wednesday
 Item
Adjustment               1
Bare Popcorn             1
Cherry me Dried fruit    1
Crepes                   1
Duck egg                 1
dtype: int64
Thursday
 Item
Argentina Night               1
Brioche and salami            1
Cherry me Dried fruit         1
Chimichurri Oil               1
Drinking chocolate spoons     1
dtype: int64
Friday
 Item
Brioche and salami    1
Chimichurri Oil       1
Chocolates            1
Coffee granules       1
Crepes                1
dtype: int64
Saturday
 Item
Bowl Nic Pitt             1
Cherry me Dried fruit     1
Christmas common          1
Dulce de Leche            1
Ella's Kitchen Pouches    1
dtype: int64
```

1. how many drinks are there per transaction?

In [14]:
```
Q9_best_5=pd.pivot_table(df,values='Item',index=['Transaction','Q4_label'],aggf
print(Q9_best_5.count())
```

```
      Q4_label
Item  drink       2892
      food        5414
      unknow      6854
dtype: int64
```