K Means Clustering Project

Problem Statement and Context

In our retail estate we have a large number of food products which reach their expiry date within our stores. The current process is to scan these products prior to store opening, and reducing them by placing a yellow sticker on them with a lower price.

There is currently some logic in place called 'Ones and Twos' - this logic is present in the application where the colleague scans the product to reduce it, and will recommend that the colleague doesn't reduce a product where the quantity is one or two. The colleague then returns later in the day to mark whether the products subsequently sold at full price, or whether it needs to be reduced.

Using 3 months of data on whether certain products sold after being added to this list, I want to see if we can make a better recommendation on whether the product should be added to the list or not by clustering different food products.

Hypothesis

My hypothesis is that there is a correlation between the volume sold and the volume unsold, which can help cluster the data, and give us a recommendation as to whether or not we should reduce a product pre-opening, or whether we should leave it to try and sell it at full price.

Step 1. Import Libraries

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import scale
```

Step 2. Load data

In [43]:

```
contents = pd.read_csv('foodsdata6jan23mar.csv')
contents.head
```

Out[43]:

```
<bound method NDFrame.head of           upc              name  total
_volume_sold  total_value_sold  \
0       2929     FLAT PARSLEY              91.0            72.80
1       3063  CHICKEN KIEV X2             663.0          2321.10
```

2	3506	DUCK A L'ORANGE	63.0	535.50
3	9515	DINE IN CHICKEN	2.0	12.00
4	10191	SAL PRAWN AVO	6.0	36.00
5	10238	PRWN SAL TERIN	0.0	0.00
6	11235	S/CRST PASTRY	59.0	118.00
7	11297	MANDAGOLD	23.0	50.00
8	11600	F/R EGG LGE X6	388.0	776.80
9	11709	AB FLJK CKS	3.0	5.25
10	11716	F/R EGGS X6	63.0	85.25
11	11778	RF ST GING CKS	2.0	3.50
12	11839	STWBRY & BANAN	34.0	78.30
13	11891	R/BOW BABY CARR	55.0	110.00
14	11990	KIDDERTON CW	0.0	0.00
15	12126	P/PASSIONFRUIT	1.0	1.00
16	12683	SLICED WH MRMS	458.0	458.00
17	13086	BIRCHER MUESLI	0.0	0.00
18	14649	ROTIS CHICKN	135.0	810.00
19	14663	BCON&CHS BAKE (	7.0	21.80
20	14670	COU CHKN PIE	388.0	1164.00
21	14700	COU BEEF RAGU	177.0	531.00
22	14717	COU PRWN LINGNE	133.0	399.00
23	14724	COU CHK ARRBTA	236.0	708.00
24	14854	BFY POT RST CHK	761.0	3044.00
25	14960	BFY BEEF NDLES	13.0	52.00
26	15011	WHOLE BIRD LGE	213.0	1132.86
27	15035	P/S/O CHICKN FW	85.0	510.00
28	15462	CGRILL PRNS 80G	197.0	592.60
29	17619	CHILLI CON CRNE	442.0	1106.20
...	...	...	...	...
3165	993555	KATSU CHK CURRY	48.0	204.00
3166	993982	SIMPLY ORANGE	5.0	5.40
3167	995528	ODB GREAT BRITI	30.0	93.00
3168	995580	2 SIRLOIN STEAK	14.0	104.00
3169	995634	6 PORK/CAR SAUS	104.0	323.60
3170	995641	TOPSIDE STEAK	44.0	198.00
3171	996129	BANOFFEE C/CAKE	16.0	44.00
3172	996198	0% FAT APL/PLUM	2.0	1.40
3173	996211	0% FAT APRICOT	11.0	7.70
3174	996235	0% FAT R/BERRY	25.0	17.50
3175	996259	0% FAT S/BERRY	24.0	16.80
3176	996334	YORKSHIRE BLUE	2.0	5.00
3177	996457	APPLE CRUMBLE	52.0	39.00
3178	996877	ROSE & BAY RM M	1.0	2.00
3179	997362	TOM BASIL SPAG	647.0	1942.70
3180	997386	CHK FETTUCCINE	769.0	2314.50
3181	997423	CHKN SWT CHILLI	40.0	140.00
3182	997492	MSHRM STROGANFF	80.0	240.00
3183	997805	TERIYAKI SAUCE	30.0	30.20
3184	998055	ACTIVIA PEACH	4.0	8.00
3185	998062	ACTIVIA STRWBRY	10.0	20.00
3186	998079	ACTIVIA CHERRY	1.0	2.00
3187	998086	ACTIVIA RHUBARB	0.0	0.00
3188	998475	2 PEP PORK LOIN	43.0	144.36
3189	998574	RED PLUMS	47.0	98.10
3190	998697	JAFFA SPHERES	13.0	48.80
3191	998840	APPL TART TATIN	38.0	146.40
3192	999069	MEDJOOL BOX	1.0	5.50

3193	999427	2 CHS PASTIES	25.0	50.00
3194	999694	COMTE FW	2.0	9.00

	total_volume_unsold	total_value_unsold
0	171.0	136.88
1	913.0	3196.10
2	125.0	1062.50
3	0.0	0.00
4	7.0	42.00
5	0.0	0.00
6	156.0	312.20
7	14.0	32.00
8	230.0	460.70
9	0.0	0.00
10	58.0	78.40
11	0.0	0.00
12	86.0	198.20
13	86.0	172.00
14	3.0	9.50
15	1.0	1.00
16	387.0	387.00
17	1.0	2.00
18	246.0	1476.00
19	2.0	6.20
20	323.0	969.00
21	344.0	1032.00
22	234.0	702.00
23	354.0	1062.00
24	948.0	3792.00
25	26.0	104.00
26	391.0	2117.38
27	192.0	1152.00
28	286.0	858.80
29	497.0	1243.70
...	...	...
3165	167.0	709.75
3166	5.0	5.20
3167	50.0	155.20
3168	18.0	124.00
3169	140.0	436.80
3170	164.0	738.00
3171	26.0	71.50
3172	9.0	6.30
3173	5.0	3.50
3174	37.0	25.90
3175	39.0	27.30
3176	7.0	17.50
3177	44.0	33.00
3178	0.0	0.00
3179	948.0	2848.70
3180	1023.0	3072.70
3181	75.0	262.50
3182	144.0	432.00
3183	31.0	31.15
3184	8.0	16.00
3185	28.0	56.00
3186	0.0	0.00

```
3187                  1.0                   2.00
3188                136.0                 460.25
3189                 25.0                  55.10
3190                 49.0                 190.40
3191                 90.0                 351.20
3192                  9.0                  49.50
3193                 55.0                 110.00
3194                  3.0                  13.50

[3195 rows x 6 columns]>
```

The data has been successfully imported. The first column represents the UPC of the food product, the next column represents the name of the product. The next column how many times the product sold at full price when not reduced in the morningby volume, and the next column is by value. The final two columns represents how many time each product didn't sell at full price, and subsequently was reduced to a lower price in the afternoon by volume and by value.

For the purpose of this exercise I will only use volume.

Step 3. Transform the data for python

In [3]:

```
contents.columns = ['UPC', 'name', 'volumesold', 'valuesold', 'volumeunsold',
'valueunsold']
print(contents[['name']])
contents.head
```

```
                name
0        FLAT PARSLEY
1      CHICKEN KIEV X2
2      DUCK A L'ORANGE
3      DINE IN CHICKEN
4        SAL PRAWN AVO
5       PRWN SAL TERIN
6        S/CRST PASTRY
7            MANDAGOLD
8       F/R EGG LGE X6
9          AB FLJK CKS
10         F/R EGGS X6
11      RF ST GING CKS
12      STWBRY & BANAN
13     R/BOW BABY CARR
14        KIDDERTON CW
15      P/PASSIONFRUIT
16      SLICED WH MRMS
17      BIRCHER MUESLI
18        ROTIS CHICKN
19     BCON&CHS BAKE (
20        COU CHKN PIE
21       COU BEEF RAGU
22     COU PRWN LINGNE
23      COU CHK ARRBTA
24     BFY POT RST CHK
```

```
25     BFY BEEF NDLES
26     WHOLE BIRD LGE
27    P/S/O CHICKN FW
28    CGRILL PRNS 80G
29    CHILLI CON CRNE
...               ...
3165  KATSU CHK CURRY
3166    SIMPLY ORANGE
3167  ODB GREAT BRITI
3168  2 SIRLOIN STEAK
3169  6 PORK/CAR SAUS
3170    TOPSIDE STEAK
3171  BANOFFEE C/CAKE
3172  0% FAT APL/PLUM
3173   0% FAT APRICOT
3174   0% FAT R/BERRY
3175   0% FAT S/BERRY
3176   YORKSHIRE BLUE
3177    APPLE CRUMBLE
3178  ROSE & BAY RM M
3179   TOM BASIL SPAG
3180   CHK FETTUCCINE
3181  CHKN SWT CHILLI
3182  MSHRM STROGANFF
3183   TERIYAKI SAUCE
3184    ACTIVIA PEACH
3185  ACTIVIA STRWBRY
3186   ACTIVIA CHERRY
3187  ACTIVIA RHUBARB
3188  2 PEP PORK LOIN
3189        RED PLUMS
3190    JAFFA SPHERES
3191  APPL TART TATIN
3192      MEDJOOL BOX
3193    2 CHS PASTIES
3194         COMTE FW

[3195 rows x 1 columns]
```

Out[3]:

```
<bound method NDFrame.head of          UPC              name  volum
esold  valuesold  volumeunsold  \
0       2929     FLAT PARSLEY        91.0      72.80          171.0
1       3063  CHICKEN KIEV X2       663.0    2321.10          913.0
2       3506  DUCK A L'ORANGE        63.0     535.50          125.0
3       9515  DINE IN CHICKEN         2.0      12.00            0.0
4      10191    SAL PRAWN AVO         6.0      36.00            7.0
5      10238   PRWN SAL TERIN         0.0       0.00            0.0
6      11235    S/CRST PASTRY        59.0     118.00          156.0
7      11297        MANDAGOLD        23.0      50.00           14.0
8      11600   F/R EGG LGE X6       388.0     776.80          230.0
9      11709      AB FLJK CKS         3.0       5.25            0.0
10     11716      F/R EGGS X6        63.0      85.25           58.0
11     11778   RF ST GING CKS         2.0       3.50            0.0
12     11839   STWBRY & BANAN        34.0      78.30           86.0
13     11891  R/BOW BABY CARR        55.0     110.00           86.0
14     11990     KIDDERTON CW         0.0       0.00            3.0
```

15	12126	P/PASSIONFRUIT	1.0	1.00	1.0
16	12683	SLICED WH MRMS	458.0	458.00	387.0
17	13086	BIRCHER MUESLI	0.0	0.00	1.0
18	14649	ROTIS CHICKN	135.0	810.00	246.0
19	14663	BCON&CHS BAKE (	7.0	21.80	2.0
20	14670	COU CHKN PIE	388.0	1164.00	323.0
21	14700	COU BEEF RAGU	177.0	531.00	344.0
22	14717	COU PRWN LINGNE	133.0	399.00	234.0
23	14724	COU CHK ARRBTA	236.0	708.00	354.0
24	14854	BFY POT RST CHK	761.0	3044.00	948.0
25	14960	BFY BEEF NDLES	13.0	52.00	26.0
26	15011	WHOLE BIRD LGE	213.0	1132.86	391.0
27	15035	P/S/O CHICKN FW	85.0	510.00	192.0
28	15462	CGRILL PRNS 80G	197.0	592.60	286.0
29	17619	CHILLI CON CRNE	442.0	1106.20	497.0
...	...	...	...	...	...
3165	993555	KATSU CHK CURRY	48.0	204.00	167.0
3166	993982	SIMPLY ORANGE	5.0	5.40	5.0
3167	995528	ODB GREAT BRITI	30.0	93.00	50.0
3168	995580	2 SIRLOIN STEAK	14.0	104.00	18.0
3169	995634	6 PORK/CAR SAUS	104.0	323.60	140.0
3170	995641	TOPSIDE STEAK	44.0	198.00	164.0
3171	996129	BANOFFEE C/CAKE	16.0	44.00	26.0
3172	996198	0% FAT APL/PLUM	2.0	1.40	9.0
3173	996211	0% FAT APRICOT	11.0	7.70	5.0
3174	996235	0% FAT R/BERRY	25.0	17.50	37.0
3175	996259	0% FAT S/BERRY	24.0	16.80	39.0
3176	996334	YORKSHIRE BLUE	2.0	5.00	7.0
3177	996457	APPLE CRUMBLE	52.0	39.00	44.0
3178	996877	ROSE & BAY RM M	1.0	2.00	0.0
3179	997362	TOM BASIL SPAG	647.0	1942.70	948.0
3180	997386	CHK FETTUCCINE	769.0	2314.50	1023.0
3181	997423	CHKN SWT CHILLI	40.0	140.00	75.0
3182	997492	MSHRM STROGANFF	80.0	240.00	144.0
3183	997805	TERIYAKI SAUCE	30.0	30.20	31.0
3184	998055	ACTIVIA PEACH	4.0	8.00	8.0
3185	998062	ACTIVIA STRWBRY	10.0	20.00	28.0
3186	998079	ACTIVIA CHERRY	1.0	2.00	0.0
3187	998086	ACTIVIA RHUBARB	0.0	0.00	1.0
3188	998475	2 PEP PORK LOIN	43.0	144.36	136.0
3189	998574	RED PLUMS	47.0	98.10	25.0
3190	998697	JAFFA SPHERES	13.0	48.80	49.0
3191	998840	APPL TART TATIN	38.0	146.40	90.0
3192	999069	MEDJOOL BOX	1.0	5.50	9.0
3193	999427	2 CHS PASTIES	25.0	50.00	55.0
3194	999694	COMTE FW	2.0	9.00	3.0

	valueunsold
0	136.88
1	3196.10
2	1062.50
3	0.00
4	42.00
5	0.00
6	312.20
7	32.00
8	460.70

9	0.00
10	78.40
11	0.00
12	198.20
13	172.00
14	9.50
15	1.00
16	387.00
17	2.00
18	1476.00
19	6.20
20	969.00
21	1032.00
22	702.00
23	1062.00
24	3792.00
25	104.00
26	2117.38
27	1152.00
28	858.80
29	1243.70
...	...
3165	709.75
3166	5.20
3167	155.20
3168	124.00
3169	436.80
3170	738.00
3171	71.50
3172	6.30
3173	3.50
3174	25.90
3175	27.30
3176	17.50
3177	33.00
3178	0.00
3179	2848.70
3180	3072.70
3181	262.50
3182	432.00
3183	31.15
3184	16.00
3185	56.00
3186	0.00
3187	2.00
3188	460.25
3189	55.10
3190	190.40
3191	351.20
3192	49.50
3193	110.00
3194	13.50

[3195 rows x 6 columns]>

The data was already in a clean state, with floats present for number of times which the items were sold or unsold. I have renamed the columns to make it easier in the python script. But worth highlighting that from now on all values are volume as opposed to £sales value.

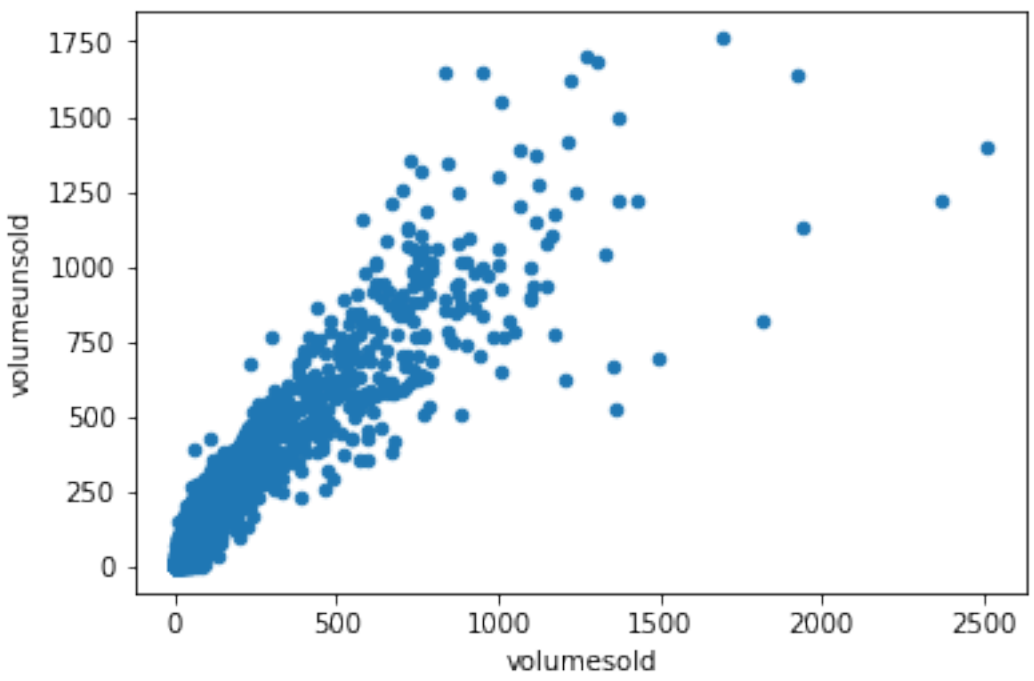
Step 4. Visualise the data

In [4]:

```
contents.plot.scatter(x ='volumesold', y ='volumeunsold')
```

Out[4]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a22d20d30>
```



At first glance, there is no way of humanly determining any clusters. There are a couple of outliers, but no conclusions can be drawn from looking at the data in this context.

Step 5. Applying K Means Clustering

In [5]:

```
data = []
for index, row in contents.iterrows():
    sold = row['volumesold']
    unsold = row['volumeunsold']
    data.append( [float(sold), float(unsold)] )

model = KMeans(n_clusters=5)
model.fit(scale(data))

contents['cluster'] = model.labels_.astype(float)
```

Step 6. Visualise the Clusters

In [6]:

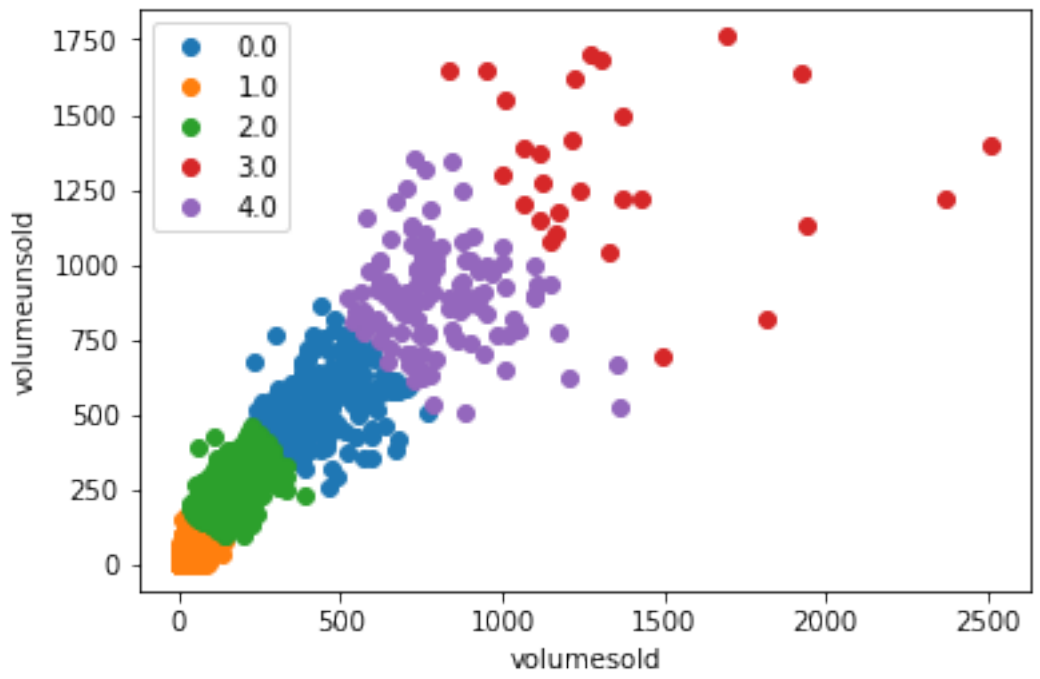
```
groups = contents.groupby('cluster')

# Plot the clusters
fig, ax = plt.subplots()
for name, group in groups:
    ax.plot(group.volumesold,group.volumeunsold, marker='o', linestyle='', lab
el=name)

plt.xlabel('volumesold')
plt.ylabel('volumeunsold')
ax.legend()
```

Out[6]:

```
<matplotlib.legend.Legend at 0x1a22ef5dd8>
```



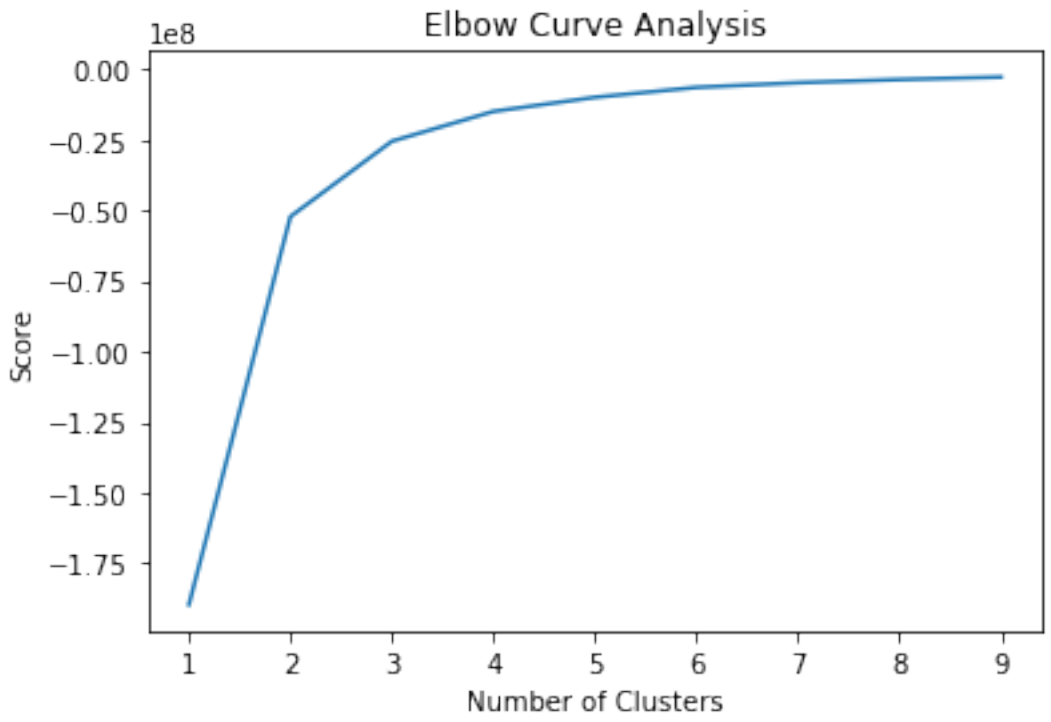
Step 7. Use Elbow test to decide number of clusters

In [7]:

```
x = contents[['volumesold']]
y = contents[['volumeunsold']]

num_clusters = [1,2,3,4,5,6,7,8,9]
kmeans = [ KMeans(n_clusters=i) for i in num_clusters ]
score = [ kmeans[i-1].fit(y).score(y)for i in num_clusters ]

plt.plot(num_clusters, score)
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve Analysis')
plt.show()
```



The elbow analysis is recommending that after 4 there isn't much more to be gained from an increased number of clusters.

At this stage I don't believe the clustering is giving much of a conclusion. This is because the volume sold and unsold together always add up to 100% of the transactions, and therefore we are getting a very linear pattern.

I believe it may be worth looking at whether the price also impacts the chance of an item selling. In order to do this I will need some additional data.

Step 8. Changing the data used to cluster

In [8]:

```
print(contents)
```

```
         UPC              name  volumesold  valuesold  volumeunsold
\
0       2929     FLAT PARSLEY         91.0      72.80         171.0
1       3063  CHICKEN KIEV X2        663.0    2321.10         913.0
2       3506  DUCK A L'ORANGE         63.0     535.50         125.0
```

3	9515	DINE IN CHICKEN	2.0	12.00	0.0
4	10191	SAL PRAWN AVO	6.0	36.00	7.0
5	10238	PRWN SAL TERIN	0.0	0.00	0.0
6	11235	S/CRST PASTRY	59.0	118.00	156.0
7	11297	MANDAGOLD	23.0	50.00	14.0
8	11600	F/R EGG LGE X6	388.0	776.80	230.0
9	11709	AB FLJK CKS	3.0	5.25	0.0
10	11716	F/R EGGS X6	63.0	85.25	58.0
11	11778	RF ST GING CKS	2.0	3.50	0.0
12	11839	STWBRY & BANAN	34.0	78.30	86.0
13	11891	R/BOW BABY CARR	55.0	110.00	86.0
14	11990	KIDDERTON CW	0.0	0.00	3.0
15	12126	P/PASSIONFRUIT	1.0	1.00	1.0
16	12683	SLICED WH MRMS	458.0	458.00	387.0
17	13086	BIRCHER MUESLI	0.0	0.00	1.0
18	14649	ROTIS CHICKN	135.0	810.00	246.0
19	14663	BCON&CHS BAKE (	7.0	21.80	2.0
20	14670	COU CHKN PIE	388.0	1164.00	323.0
21	14700	COU BEEF RAGU	177.0	531.00	344.0
22	14717	COU PRWN LINGNE	133.0	399.00	234.0
23	14724	COU CHK ARRBTA	236.0	708.00	354.0
24	14854	BFY POT RST CHK	761.0	3044.00	948.0
25	14960	BFY BEEF NDLES	13.0	52.00	26.0
26	15011	WHOLE BIRD LGE	213.0	1132.86	391.0
27	15035	P/S/O CHICKN FW	85.0	510.00	192.0
28	15462	CGRILL PRNS 80G	197.0	592.60	286.0
29	17619	CHILLI CON CRNE	442.0	1106.20	497.0
...	...	...	...	...	...
3165	993555	KATSU CHK CURRY	48.0	204.00	167.0
3166	993982	SIMPLY ORANGE	5.0	5.40	5.0
3167	995528	ODB GREAT BRITI	30.0	93.00	50.0
3168	995580	2 SIRLOIN STEAK	14.0	104.00	18.0
3169	995634	6 PORK/CAR SAUS	104.0	323.60	140.0
3170	995641	TOPSIDE STEAK	44.0	198.00	164.0
3171	996129	BANOFFEE C/CAKE	16.0	44.00	26.0
3172	996198	0% FAT APL/PLUM	2.0	1.40	9.0
3173	996211	0% FAT APRICOT	11.0	7.70	5.0
3174	996235	0% FAT R/BERRY	25.0	17.50	37.0
3175	996259	0% FAT S/BERRY	24.0	16.80	39.0
3176	996334	YORKSHIRE BLUE	2.0	5.00	7.0
3177	996457	APPLE CRUMBLE	52.0	39.00	44.0
3178	996877	ROSE & BAY RM M	1.0	2.00	0.0
3179	997362	TOM BASIL SPAG	647.0	1942.70	948.0
3180	997386	CHK FETTUCCINE	769.0	2314.50	1023.0
3181	997423	CHKN SWT CHILLI	40.0	140.00	75.0
3182	997492	MSHRM STROGANFF	80.0	240.00	144.0
3183	997805	TERIYAKI SAUCE	30.0	30.20	31.0
3184	998055	ACTIVIA PEACH	4.0	8.00	8.0
3185	998062	ACTIVIA STRWBRY	10.0	20.00	28.0
3186	998079	ACTIVIA CHERRY	1.0	2.00	0.0
3187	998086	ACTIVIA RHUBARB	0.0	0.00	1.0
3188	998475	2 PEP PORK LOIN	43.0	144.36	136.0
3189	998574	RED PLUMS	47.0	98.10	25.0
3190	998697	JAFFA SPHERES	13.0	48.80	49.0
3191	998840	APPL TART TATIN	38.0	146.40	90.0
3192	999069	MEDJOOL BOX	1.0	5.50	9.0
3193	999427	2 CHS PASTIES	25.0	50.00	55.0

3194	999694		COMTE FW	2.0	9.00	3.0

	valueunsold	cluster
0	136.88	2.0
1	3196.10	4.0
2	1062.50	1.0
3	0.00	1.0
4	42.00	1.0
5	0.00	1.0
6	312.20	2.0
7	32.00	1.0
8	460.70	2.0
9	0.00	1.0
10	78.40	1.0
11	0.00	1.0
12	198.20	1.0
13	172.00	1.0
14	9.50	1.0
15	1.00	1.0
16	387.00	0.0
17	2.00	1.0
18	1476.00	2.0
19	6.20	1.0
20	969.00	0.0
21	1032.00	2.0
22	702.00	2.0
23	1062.00	2.0
24	3792.00	4.0
25	104.00	1.0
26	2117.38	2.0
27	1152.00	2.0
28	858.80	2.0
29	1243.70	0.0
...	...	...
3165	709.75	2.0
3166	5.20	1.0
3167	155.20	1.0
3168	124.00	1.0
3169	436.80	2.0
3170	738.00	1.0
3171	71.50	1.0
3172	6.30	1.0
3173	3.50	1.0
3174	25.90	1.0
3175	27.30	1.0
3176	17.50	1.0
3177	33.00	1.0
3178	0.00	1.0
3179	2848.70	4.0
3180	3072.70	4.0
3181	262.50	1.0
3182	432.00	2.0
3183	31.15	1.0
3184	16.00	1.0
3185	56.00	1.0
3186	0.00	1.0
3187	2.00	1.0

```
3188       460.25       1.0
3189        55.10       1.0
3190       190.40       1.0
3191       351.20       1.0
3192        49.50       1.0
3193       110.00       1.0
3194        13.50       1.0

[3195 rows x 7 columns]
```

In [9]:

```
# Here I am adding a new column. I don't have the individual price information for a product, but I can calculate it by dividing the value of the products sold with the volume of the products sold.
contents['Price'] = contents['valuesold']/contents['volumesold']
```

In [10]:

```
print(contents)
```

```
          UPC              name  volumesold  valuesold  volumeunsold
\
0        2929     FLAT PARSLEY         91.0      72.80         171.0
1        3063  CHICKEN KIEV X2        663.0    2321.10         913.0
2        3506  DUCK A L'ORANGE         63.0     535.50         125.0
3        9515  DINE IN CHICKEN          2.0      12.00           0.0
4       10191    SAL PRAWN AVO          6.0      36.00           7.0
5       10238   PRWN SAL TERIN          0.0       0.00           0.0
6       11235    S/CRST PASTRY         59.0     118.00         156.0
7       11297        MANDAGOLD         23.0      50.00          14.0
8       11600   F/R EGG LGE X6        388.0     776.80         230.0
9       11709      AB FLJK CKS          3.0       5.25           0.0
10      11716      F/R EGGS X6         63.0      85.25          58.0
11      11778   RF ST GING CKS          2.0       3.50           0.0
12      11839   STWBRY & BANAN         34.0      78.30          86.0
13      11891  R/BOW BABY CARR         55.0     110.00          86.0
14      11990     KIDDERTON CW          0.0       0.00           3.0
15      12126   P/PASSIONFRUIT          1.0       1.00           1.0
16      12683   SLICED WH MRMS        458.0     458.00         387.0
17      13086   BIRCHER MUESLI          0.0       0.00           1.0
18      14649     ROTIS CHICKN        135.0     810.00         246.0
19      14663  BCON&CHS BAKE (          7.0      21.80           2.0
20      14670     COU CHKN PIE        388.0    1164.00         323.0
21      14700    COU BEEF RAGU        177.0     531.00         344.0
22      14717  COU PRWN LINGNE        133.0     399.00         234.0
23      14724   COU CHK ARRBTA        236.0     708.00         354.0
24      14854  BFY POT RST CHK        761.0    3044.00         948.0
25      14960   BFY BEEF NDLES         13.0      52.00          26.0
26      15011   WHOLE BIRD LGE        213.0    1132.86         391.0
27      15035  P/S/O CHICKN FW         85.0     510.00         192.0
28      15462  CGRILL PRNS 80G        197.0     592.60         286.0
29      17619  CHILLI CON CRNE        442.0    1106.20         497.0
...       ...              ...         ...        ...           ...
3165   993555  KATSU CHK CURRY         48.0     204.00         167.0
3166   993982    SIMPLY ORANGE          5.0       5.40           5.0
3167   995528  ODB GREAT BRITI         30.0      93.00          50.0
```

3168	995580	2 SIRLOIN STEAK	14.0	104.00	18.0
3169	995634	6 PORK/CAR SAUS	104.0	323.60	140.0
3170	995641	TOPSIDE STEAK	44.0	198.00	164.0
3171	996129	BANOFFEE C/CAKE	16.0	44.00	26.0
3172	996198	0% FAT APL/PLUM	2.0	1.40	9.0
3173	996211	0% FAT APRICOT	11.0	7.70	5.0
3174	996235	0% FAT R/BERRY	25.0	17.50	37.0
3175	996259	0% FAT S/BERRY	24.0	16.80	39.0
3176	996334	YORKSHIRE BLUE	2.0	5.00	7.0
3177	996457	APPLE CRUMBLE	52.0	39.00	44.0
3178	996877	ROSE & BAY RM M	1.0	2.00	0.0
3179	997362	TOM BASIL SPAG	647.0	1942.70	948.0
3180	997386	CHK FETTUCCINE	769.0	2314.50	1023.0
3181	997423	CHKN SWT CHILLI	40.0	140.00	75.0
3182	997492	MSHRM STROGANFF	80.0	240.00	144.0
3183	997805	TERIYAKI SAUCE	30.0	30.20	31.0
3184	998055	ACTIVIA PEACH	4.0	8.00	8.0
3185	998062	ACTIVIA STRWBRY	10.0	20.00	28.0
3186	998079	ACTIVIA CHERRY	1.0	2.00	0.0
3187	998086	ACTIVIA RHUBARB	0.0	0.00	1.0
3188	998475	2 PEP PORK LOIN	43.0	144.36	136.0
3189	998574	RED PLUMS	47.0	98.10	25.0
3190	998697	JAFFA SPHERES	13.0	48.80	49.0
3191	998840	APPL TART TATIN	38.0	146.40	90.0
3192	999069	MEDJOOL BOX	1.0	5.50	9.0
3193	999427	2 CHS PASTIES	25.0	50.00	55.0
3194	999694	COMTE FW	2.0	9.00	3.0

	valueunsold	cluster	Price
0	136.88	2.0	0.800000
1	3196.10	4.0	3.500905
2	1062.50	1.0	8.500000
3	0.00	1.0	6.000000
4	42.00	1.0	6.000000
5	0.00	1.0	NaN
6	312.20	2.0	2.000000
7	32.00	1.0	2.173913
8	460.70	2.0	2.002062
9	0.00	1.0	1.750000
10	78.40	1.0	1.353175
11	0.00	1.0	1.750000
12	198.20	1.0	2.302941
13	172.00	1.0	2.000000
14	9.50	1.0	NaN
15	1.00	1.0	1.000000
16	387.00	0.0	1.000000
17	2.00	1.0	NaN
18	1476.00	2.0	6.000000
19	6.20	1.0	3.114286
20	969.00	0.0	3.000000
21	1032.00	2.0	3.000000
22	702.00	2.0	3.000000
23	1062.00	2.0	3.000000
24	3792.00	4.0	4.000000
25	104.00	1.0	4.000000
26	2117.38	2.0	5.318592
27	1152.00	2.0	6.000000

```
28          858.80      2.0  3.008122
29         1243.70      0.0  2.502715
...            ...      ...       ...
3165        709.75      2.0  4.250000
3166          5.20      1.0  1.080000
3167        155.20      1.0  3.100000
3168        124.00      1.0  7.428571
3169        436.80      2.0  3.111538
3170        738.00      1.0  4.500000
3171         71.50      1.0  2.750000
3172          6.30      1.0  0.700000
3173          3.50      1.0  0.700000
3174         25.90      1.0  0.700000
3175         27.30      1.0  0.700000
3176         17.50      1.0  2.500000
3177         33.00      1.0  0.750000
3178          0.00      1.0  2.000000
3179       2848.70      4.0  3.002628
3180       3072.70      4.0  3.009753
3181        262.50      1.0  3.500000
3182        432.00      2.0  3.000000
3183         31.15      1.0  1.006667
3184         16.00      1.0  2.000000
3185         56.00      1.0  2.000000
3186          0.00      1.0  2.000000
3187          2.00      1.0       NaN
3188        460.25      1.0  3.357209
3189         55.10      1.0  2.087234
3190        190.40      1.0  3.753846
3191        351.20      1.0  3.852632
3192         49.50      1.0  5.500000
3193        110.00      1.0  2.000000
3194         13.50      1.0  4.500000

[3195 rows x 8 columns]
```

A couple of things I have noticed here - I have some lines with 'NaN' - these lines seem to have low total volumes so I will remove them. I also see that there is a column with the cluster previously derived. I will try to remove this column - otherwise I will start a new python notebook and import the data again.

In [11]:

```
#first I delete the cluster
del contents['cluster']
print(contents)
```

```
         UPC              name  volumesold  valuesold  volumeunsold
\
0       2929     FLAT PARSLEY         91.0      72.80         171.0
1       3063  CHICKEN KIEV X2        663.0    2321.10         913.0
2       3506  DUCK A L'ORANGE         63.0     535.50         125.0
3       9515  DINE IN CHICKEN          2.0      12.00           0.0
4      10191    SAL PRAWN AVO          6.0      36.00           7.0
5      10238   PRWN SAL TERIN          0.0       0.00           0.0
6      11235    S/CRST PASTRY         59.0     118.00         156.0
7      11297        MANDAGOLD         23.0      50.00          14.0
```

8	11600	F/R EGG LGE X6	388.0	776.80	230.0
9	11709	AB FLJK CKS	3.0	5.25	0.0
10	11716	F/R EGGS X6	63.0	85.25	58.0
11	11778	RF ST GING CKS	2.0	3.50	0.0
12	11839	STWBRY & BANAN	34.0	78.30	86.0
13	11891	R/BOW BABY CARR	55.0	110.00	86.0
14	11990	KIDDERTON CW	0.0	0.00	3.0
15	12126	P/PASSIONFRUIT	1.0	1.00	1.0
16	12683	SLICED WH MRMS	458.0	458.00	387.0
17	13086	BIRCHER MUESLI	0.0	0.00	1.0
18	14649	ROTIS CHICKN	135.0	810.00	246.0
19	14663	BCON&CHS BAKE (	7.0	21.80	2.0
20	14670	COU CHKN PIE	388.0	1164.00	323.0
21	14700	COU BEEF RAGU	177.0	531.00	344.0
22	14717	COU PRWN LINGNE	133.0	399.00	234.0
23	14724	COU CHK ARRBTA	236.0	708.00	354.0
24	14854	BFY POT RST CHK	761.0	3044.00	948.0
25	14960	BFY BEEF NDLES	13.0	52.00	26.0
26	15011	WHOLE BIRD LGE	213.0	1132.86	391.0
27	15035	P/S/O CHICKN FW	85.0	510.00	192.0
28	15462	CGRILL PRNS 80G	197.0	592.60	286.0
29	17619	CHILLI CON CRNE	442.0	1106.20	497.0
...	...	...	...	...	...
3165	993555	KATSU CHK CURRY	48.0	204.00	167.0
3166	993982	SIMPLY ORANGE	5.0	5.40	5.0
3167	995528	ODB GREAT BRITI	30.0	93.00	50.0
3168	995580	2 SIRLOIN STEAK	14.0	104.00	18.0
3169	995634	6 PORK/CAR SAUS	104.0	323.60	140.0
3170	995641	TOPSIDE STEAK	44.0	198.00	164.0
3171	996129	BANOFFEE C/CAKE	16.0	44.00	26.0
3172	996198	0% FAT APL/PLUM	2.0	1.40	9.0
3173	996211	0% FAT APRICOT	11.0	7.70	5.0
3174	996235	0% FAT R/BERRY	25.0	17.50	37.0
3175	996259	0% FAT S/BERRY	24.0	16.80	39.0
3176	996334	YORKSHIRE BLUE	2.0	5.00	7.0
3177	996457	APPLE CRUMBLE	52.0	39.00	44.0
3178	996877	ROSE & BAY RM M	1.0	2.00	0.0
3179	997362	TOM BASIL SPAG	647.0	1942.70	948.0
3180	997386	CHK FETTUCCINE	769.0	2314.50	1023.0
3181	997423	CHKN SWT CHILLI	40.0	140.00	75.0
3182	997492	MSHRM STROGANFF	80.0	240.00	144.0
3183	997805	TERIYAKI SAUCE	30.0	30.20	31.0
3184	998055	ACTIVIA PEACH	4.0	8.00	8.0
3185	998062	ACTIVIA STRWBRY	10.0	20.00	28.0
3186	998079	ACTIVIA CHERRY	1.0	2.00	0.0
3187	998086	ACTIVIA RHUBARB	0.0	0.00	1.0
3188	998475	2 PEP PORK LOIN	43.0	144.36	136.0
3189	998574	RED PLUMS	47.0	98.10	25.0
3190	998697	JAFFA SPHERES	13.0	48.80	49.0
3191	998840	APPL TART TATIN	38.0	146.40	90.0
3192	999069	MEDJOOL BOX	1.0	5.50	9.0
3193	999427	2 CHS PASTIES	25.0	50.00	55.0
3194	999694	COMTE FW	2.0	9.00	3.0

	valueunsold	Price
0	136.88	0.800000
1	3196.10	3.500905

2	1062.50	8.500000
3	0.00	6.000000
4	42.00	6.000000
5	0.00	NaN
6	312.20	2.000000
7	32.00	2.173913
8	460.70	2.002062
9	0.00	1.750000
10	78.40	1.353175
11	0.00	1.750000
12	198.20	2.302941
13	172.00	2.000000
14	9.50	NaN
15	1.00	1.000000
16	387.00	1.000000
17	2.00	NaN
18	1476.00	6.000000
19	6.20	3.114286
20	969.00	3.000000
21	1032.00	3.000000
22	702.00	3.000000
23	1062.00	3.000000
24	3792.00	4.000000
25	104.00	4.000000
26	2117.38	5.318592
27	1152.00	6.000000
28	858.80	3.008122
29	1243.70	2.502715
...	...	...
3165	709.75	4.250000
3166	5.20	1.080000
3167	155.20	3.100000
3168	124.00	7.428571
3169	436.80	3.111538
3170	738.00	4.500000
3171	71.50	2.750000
3172	6.30	0.700000
3173	3.50	0.700000
3174	25.90	0.700000
3175	27.30	0.700000
3176	17.50	2.500000
3177	33.00	0.750000
3178	0.00	2.000000
3179	2848.70	3.002628
3180	3072.70	3.009753
3181	262.50	3.500000
3182	432.00	3.000000
3183	31.15	1.006667
3184	16.00	2.000000
3185	56.00	2.000000
3186	0.00	2.000000
3187	2.00	NaN
3188	460.25	3.357209
3189	55.10	2.087234
3190	190.40	3.753846
3191	351.20	3.852632
3192	49.50	5.500000

```
3193        110.00   2.000000
3194         13.50   4.500000

[3195 rows x 7 columns]
```

In [12]:

```
#Here I drop the rows where at least one element is missing.
contents.dropna()
```

Out[12]:

	UPC	name	volumesold	valuesold	volumeunsold	valueunsold	Pric
0	2929	FLAT PARSLEY	91.0	72.80	171.0	136.88	0.80000
1	3063	CHICKEN KIEV X2	663.0	2321.10	913.0	3196.10	3.50090
2	3506	DUCK A L'ORANGE	63.0	535.50	125.0	1062.50	8.50000
3	9515	DINE IN CHICKEN	2.0	12.00	0.0	0.00	6.00000
4	10191	SAL PRAWN AVO	6.0	36.00	7.0	42.00	6.00000
6	11235	S/CRST PASTRY	59.0	118.00	156.0	312.20	2.00000
7	11297	MANDAGOLD	23.0	50.00	14.0	32.00	2.17391
8	11600	F/R EGG LGE X6	388.0	776.80	230.0	460.70	2.00206
9	11709	AB FLJK CKS	3.0	5.25	0.0	0.00	1.75000
10	11716	F/R EGGS X6	63.0	85.25	58.0	78.40	1.35317
11	11778	RF ST GING CKS	2.0	3.50	0.0	0.00	1.75000
12	11839	STWBRY & BANAN	34.0	78.30	86.0	198.20	2.30294
13	11891	R/BOW BABY CARR	55.0	110.00	86.0	172.00	2.00000
15	12126	P/PASSIONFRUIT	1.0	1.00	1.0	1.00	1.00000
16	12683	SLICED WH MRMS	458.0	458.00	387.0	387.00	1.00000
18	14649	ROTIS CHICKN	135.0	810.00	246.0	1476.00	6.00000
19	14663	BCON&CHS BAKE (	7.0	21.80	2.0	6.20	3.11428
20	14670	COU CHKN PIE	388.0	1164.00	323.0	969.00	3.00000
21	14700	COU BEEF RAGU	177.0	531.00	344.0	1032.00	3.00000
22	14717	COU PRWN LINGNE	133.0	399.00	234.0	702.00	3.00000
23	14724	COU CHK ARRBTA	236.0	708.00	354.0	1062.00	3.00000
24	14854	BFY POT RST CHK	761.0	3044.00	948.0	3792.00	4.00000

25	14960	BFY BEEF NDLES	13.0	52.00	26.0	104.00	4.00000
26	15011	WHOLE BIRD LGE	213.0	1132.86	391.0	2117.38	5.31859
27	15035	P/S/O CHICKN FW	85.0	510.00	192.0	1152.00	6.00000
28	15462	CGRILL PRNS 80G	197.0	592.60	286.0	858.80	3.00812
29	17619	CHILLI CON CRNE	442.0	1106.20	497.0	1243.70	2.50271
30	17640	FR DRUMSTICKS	29.0	109.38	49.0	183.68	3.77172
31	17657	FR S/LESS THIGH	35.0	233.08	71.0	426.05	6.65942
32	17664	FR MINI FILLETS	15.0	66.64	41.0	174.09	4.44266
...	...	...	...	...	...	...	
3164	993531	G THAI CHK CRRY	43.0	182.75	76.0	323.50	4.25000
3165	993555	KATSU CHK CURRY	48.0	204.00	167.0	709.75	4.25000
3166	993982	SIMPLY ORANGE	5.0	5.40	5.0	5.20	1.08000
3167	995528	ODB GREAT BRITI	30.0	93.00	50.0	155.20	3.10000
3168	995580	2 SIRLOIN STEAK	14.0	104.00	18.0	124.00	7.42857
3169	995634	6 PORK/CAR SAUS	104.0	323.60	140.0	436.80	3.11153
3170	995641	TOPSIDE STEAK	44.0	198.00	164.0	738.00	4.50000
3171	996129	BANOFFEE C/CAKE	16.0	44.00	26.0	71.50	2.75000
3172	996198	0% FAT APL/PLUM	2.0	1.40	9.0	6.30	0.70000
3173	996211	0% FAT APRICOT	11.0	7.70	5.0	3.50	0.70000
3174	996235	0% FAT R/BERRY	25.0	17.50	37.0	25.90	0.70000
3175	996259	0% FAT S/BERRY	24.0	16.80	39.0	27.30	0.70000
3176	996334	YORKSHIRE BLUE	2.0	5.00	7.0	17.50	2.50000
3177	996457	APPLE CRUMBLE	52.0	39.00	44.0	33.00	0.75000
3178	996877	ROSE & BAY RM M	1.0	2.00	0.0	0.00	2.00000
3179	997362	TOM BASIL SPAG	647.0	1942.70	948.0	2848.70	3.00262
3180	997386	CHK FETTUCCINE	769.0	2314.50	1023.0	3072.70	3.00975

3181	997423	CHKN SWT CHILLI	40.0	140.00	75.0	262.50	3.50000
3182	997492	MSHRM STROGANFF	80.0	240.00	144.0	432.00	3.00000
3183	997805	TERIYAKI SAUCE	30.0	30.20	31.0	31.15	1.00666
3184	998055	ACTIVIA PEACH	4.0	8.00	8.0	16.00	2.00000
3185	998062	ACTIVIA STRWBRY	10.0	20.00	28.0	56.00	2.00000
3186	998079	ACTIVIA CHERRY	1.0	2.00	0.0	0.00	2.00000
3188	998475	2 PEP PORK LOIN	43.0	144.36	136.0	460.25	3.35720
3189	998574	RED PLUMS	47.0	98.10	25.0	55.10	2.08723
3190	998697	JAFFA SPHERES	13.0	48.80	49.0	190.40	3.75384
3191	998840	APPL TART TATIN	38.0	146.40	90.0	351.20	3.85263
3192	999069	MEDJOOL BOX	1.0	5.50	9.0	49.50	5.50000
3193	999427	2 CHS PASTIES	25.0	50.00	55.0	110.00	2.00000
3194	999694	COMTE FW	2.0	9.00	3.0	13.50	4.50000

2901 rows × 7 columns

I have dropped 294 rows which had an element missing.

In [13]:

```
#Here I have rounded the price to two decimal places.
contents.Price = contents.Price.round(2)
```

In [19]:

```
contents.dtypes
```

Out[19]:

```
UPC              int64
name            object
volumesold     float64
valuesold      float64
volumeunsold   float64
valueunsold    float64
Price          float64
dtype: object
```

I am now in a position to retry the clustering, and see if there is some correlation that can be found between the price and the chance of selling.

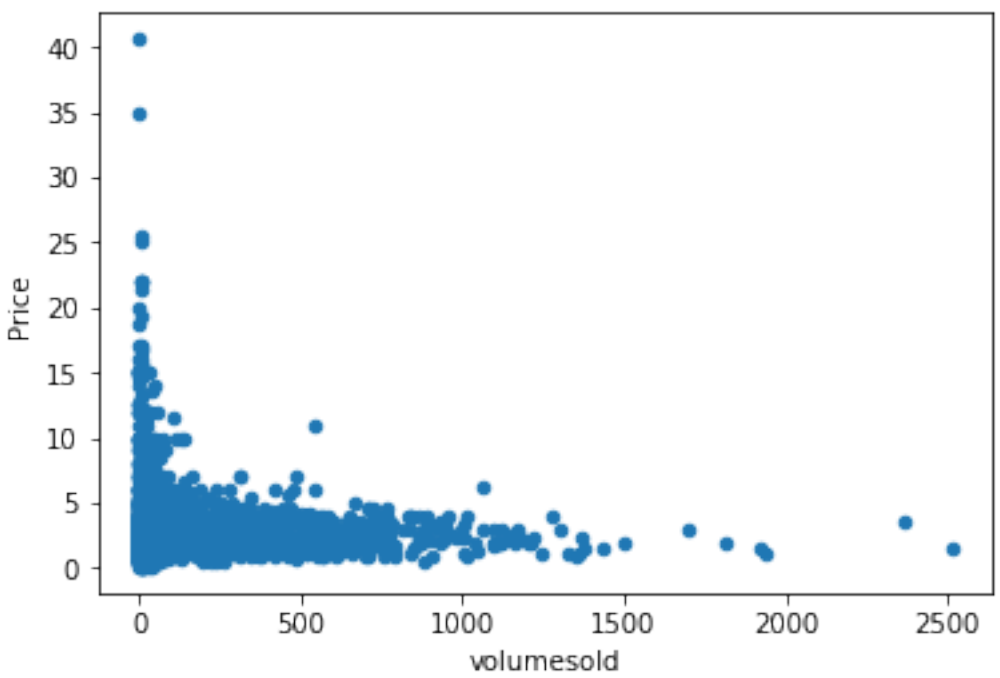
Step 9. Visualise the NEW data

In [26]:

```
contents.plot.scatter(x ='volumesold', y ='Price')
```

Out[26]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2411a860>
```



Step 10. Applying K Means Clustering on the NEW data

In [39]:

```
data = []
for index, row in contents.iterrows():
    volumesold = row['volumesold']
    valuesold = row['valuesold']
    data.append( [float(volumesold), float(Price)] )

model = KMeans(n_clusters=4)
model.fit(scale(data))

contents['cluster'] = model.labels_.astype(float)
```

Step 11. Visualise the Clusters

In [40]:

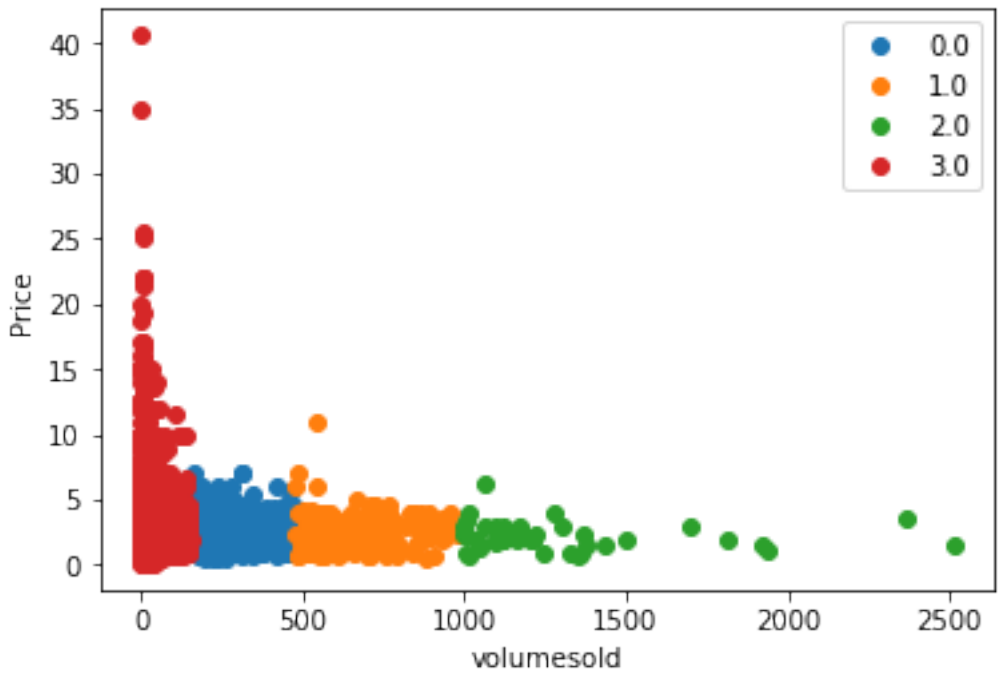
```
groups = contents.groupby('cluster')

# Plot the clusters
fig, ax = plt.subplots()
for name, group in groups:
    ax.plot(group.volumesold,group.Price, marker='o', linestyle='', label=name
)

plt.xlabel('volumesold')
plt.ylabel('Price')
ax.legend()
```

Out[40]:

```
<matplotlib.legend.Legend at 0x1a24294f28>
```



In the visualisation above there are some clear clusters - Cluster 3 shows products which don't tend to sell at full price, and Cluster 2 shows products which have a high volume of sold, and are often at a low price.

Given that we want to be able to use this data to inform the decision of whether a product should be reduced at the shelf edge or not in the morning based on its cluster, I will reapply the visualisation with only two clusters.

In [41]:

```
data = []
for index, row in contents.iterrows():
    volumesold = row['volumesold']
    valuesold = row['valuesold']
    data.append( [float(volumesold), float(Price)] )

model = KMeans(n_clusters=2)
model.fit(scale(data))

contents['cluster'] = model.labels_.astype(float)
```

In [42]:

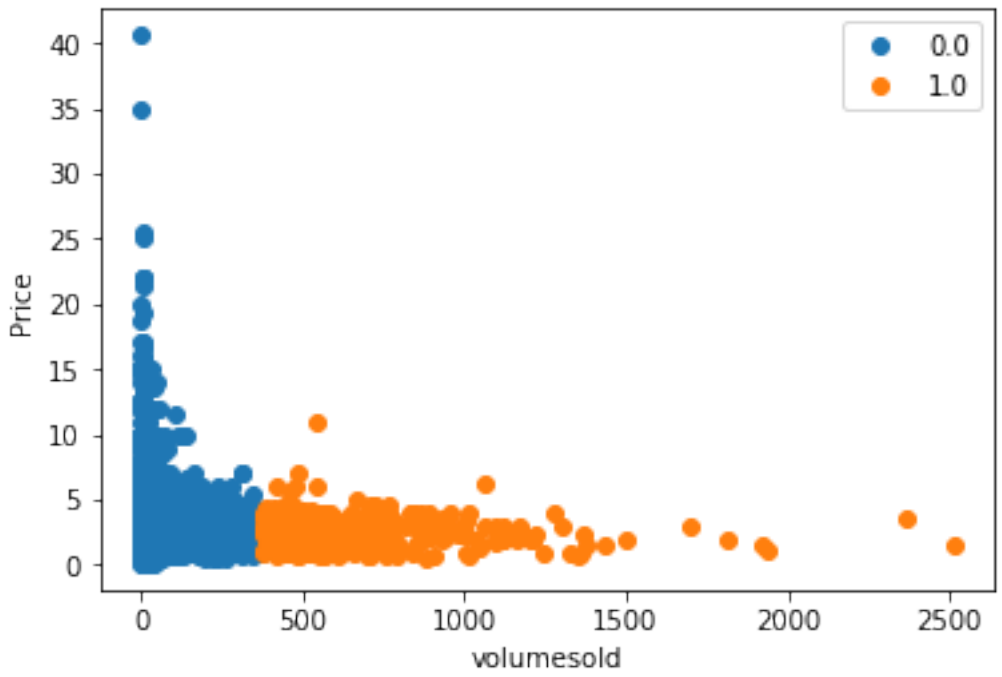
```
groups = contents.groupby('cluster')

# Plot the clusters
fig, ax = plt.subplots()
for name, group in groups:
    ax.plot(group.volumesold,group.Price, marker='o', linestyle='', label=name
)

plt.xlabel('volumesold')
plt.ylabel('Price')
ax.legend()
```

Out[42]:

```
<matplotlib.legend.Legend at 0x1a24665780>
```



Here we now have two clusters. The orange cluster shows products which have a higher likelyhood to sell. There is some correlation between price and volume sold. I would conclude that the lower the price, the more likely the product is to sell on its final day of expiry, however I don't believe the evidence here would give me a cluster that could be used as the logic to make a recommendation.