Data at Scale: Management, Processing, Visualization

Coursework

ID number: 20171924

Module: Data at Scale: Management, Processing, Visualization

Current academic year: 2019/20

Contents

Section 0: Data cleaning

Goal: Data cleaning that qty is null in receipt_line table

code:

```
delete from cw.receipt_lines where qty is null;
```

Goal: Updating data that same product_code but different value

code:

```
update cw.receipt_lines
set value = 3.39
where receipt_id = 71220
and product_code = 12693;

update cw.receipt_lines
set value = 3.39
where receipt_id = 99038
and product_code = 12693;

update cw.receipt_lines
set value = 3.39
where receipt_id = 48398
and product_code = 12693;

update cw.receipt_lines
set value = 1.61
where receipt_id = 48398
and product_code = 18035;

update cw.receipt_lines
set value = 1.61
where receipt_id = 48398
and product_code = 14607;

update cw.receipt_lines
set value = 2.50
where receipt_id = 48398
and product_code = 883;

update cw.receipt_lines
set value = 1.01
where receipt_id = 48398
and product_code = 16660;
```

```
update cw.receipt_lines
set value = 2.19
where receipt_id = 48398
and product_code = 10737;

update cw.receipt_lines
set value = 0.81
where receipt_id = 48398
and product_code = 15247;

update cw.receipt_lines
set value = 1.70
where receipt_id = 48398
and product_code = 571;

update cw.receipt_lines
set value = 4.99
where receipt_id = 48398
and product_code = 10996;

update cw.receipt_lines
set value = 4.21
where receipt_id = 48398
and product_code = 6993;

update cw.receipt_lines
set value = 0.99
where receipt_id = 48398
and product_code = 1523;

update cw.receipt_lines
set value = 0.99
where receipt_id = 66724
and product_code = 15482;

update cw.receipt_lines
set value = 0.99
where receipt_id = 48398
and product_code = 15482;

update cw.receipt_lines
set value = 3.57
where receipt_id = 48398
and product_code = 13403;

update cw.receipt_lines
set value = 1.09
```

where receipt_id = 48398
and product_code = 5460;

update cw.receipt_lines
set value = 1.70
where receipt_id = 48398
and product_code = 201;

Section 1: The KPIs

KPI Description: Measuring store size

KPI formula: count(deparment_code), sum(value) / relative_size, per store

Steps to realize KPI:

1) Checking the number of product, department category and sub
: *All the numbers are equal*

```
select store_code, count(product_code) as product, count(department_code) as
department,
        count(category_code) as category, count(sub_category_code) as sub
from cw.products
join cw.receipt_lines using(product_code)
join cw.receipts using (receipt_id)
group by 1;
```

2. Measuring relative store size (store_code = 0 is standard as 1)

```
select store_code, round((sum(value)/relative_size)::numeric,2) as
        relative_total_sales
from cw.receipt_lines
join cw.receipts using (receipt_id)
join relative_store_size using(store_code)
group by store_code, relative_size
order by relative_total_sales desc;
```

3. Create relative_store_size table

```
create table relative_store_size(
        store_code bigint,
        relative_size float
);
insert into relative_store_size values (0, 1.00);
insert into relative_store_size values (1, 0.06);
insert into relative_store_size values (2, 0.52);
 insert into relative_store_size values (3, 0.29);
```

4.relative_store_size table is joining with other tables in other SQL queries.

Additional Notes: To compare the size of each store, the number of product_code, department_code, category_code and sub_category_code was chosen. This is because, the bigger the size of store has more diversity of each code. Since those four codes has equal numbers, I chose *store_code 0* as a standard and considered as 1 in store size. The number of product_code of *store_code 0* is 768,756, thus relative size of store is divide each store's the number of product_code by 768,756. As a result, *store_code 0*'s relative store size is 1 and *store 1* is 0.06, *store 2* is 0.52 and *store 3* is 0.29.

KPI Description: Comparing between total sales and total sales in terms of relative store size

KPI formula: sum(value), sum(value) / relative_size, per store

Steps to realize KPI:

1) Total sales SQL query

```
select store_code, sum(value) as total_sales
from cw.receipt_lines
join cw.receipts using (receipt_id)
join relative_store_size using(store_code)
group by store_code
order by total_sales desc;
```

2) Total sales in relative store size SQL query

```
select store_code, round((sum(value)/relative_size)::numeric,2) as
        relative_total_sales
from cw.receipt_lines
join cw.receipts using (receipt_id)
join relative_store_size using(store_code)
group by store_code, relative_size
order by relative_total_sales desc;
```

3) Create total_sales_insize table and insert each values

```
create table total_sales_insize(
store_code bigint,
total_sales numeric,
total_sales_insize numeric
);
insert into total_sales_insize (store_code, total_sales, total_sales_insize)
        values (0, 1297625.12, 1297625.12);
insert into total_sales_insize (store_code, total_sales, total_sales_insize)
        values (1, 70527.89, 1175464.83);
insert into total_sales_insize (store_code, total_sales, total_sales_insize)
        values (2, 644315.97, 1239069.17);
insert into total_sales_insize (store_code, total_sales, total_sales_insize)
        values (3, 396680.17, 1367862.66);
```

4) Visualized via Tableau as graph titled 'Total sales vs Total sales in size'. See the Tableau file

Additional Notes: None.

KPI Description: Monthly sales rate

KPI formula: sum(value) / relative_size, per month, per store

Steps to realize KPI:

1) Create sale_month_size table

```
create table sale_month_size(
        store_code bigint,
        month text,
        sale_month_bysize numeric
);
```

2) Insert sale_month_size SQL query into table

```
insert into sale_month_size(store_code, month, sale_month_bysize)
select store_code, to_char(date_trunc('month', purchased_at), 'YYYY-MM') as month,
        round((sum(value)/relative_size)::numeric,2) as relative_total_sales
from cw.receipt_lines
join cw.receipts using (receipt_id)
join relative_store_size using(store_code)
group by store_code, relative_size, date_trunc('month', purchased_at)
order by month, store_code;
```

3) Visualized via Tableau as table titled 'Monthly sales'. See the Tableau file

Additional Notes: None.

KPI Description: Monthly new customers in each stores

KPI formula: count(customer_id), min(purchased_at), per month, per store

Steps to realize KPI:

1) Create new_cutomer table

```
create table new_customer(
        store_code bigint,
        mth text,
        new_customer bigint
);
```

2) Insert new_customer SQL query into table

```
insert into new_customer(store_code, mth, new_customer)
select store_code, to_char(first_month, 'YYYY-MM') as mth, count(*) as new_customer
from(
        select store_code, customer_id, date_trunc('month', min(purchased_at)) as
first_month
```

```
        from cw.receipts
        group by 1,2
    )x
    group by 1, 2
    order by 1, 2;
```

3) Visualized via Tableau as table titled 'New Customers'. See the Tableau file.

Additional Notes: None.

KPI Description: Monthly active customers who come to the store more than three times in each stores

KPI formula: count(customer_id), count > 3, per month, per store

Steps to realize KPI:

1) Create repeat3_customer table

```
create table repeat3_customer_month(
        store_code bigint,
        mth text,
        num_loyal_customer bigint
);
```

2) Insert sale_month_size SQL query into table

```
insert into repeat3_customer_month(store_code, mth, num_loyal_customer)
select store_code, to_char(mth, 'YYYY-MM') as mth, count(*) as
numer_of_loyal_customer
from (
        select store_code, date_trunc('month',purchased_at) as mth, customer_id
        from cw.receipts
        group by 1, 2,3
        having count(date_trunc('month', purchased_at)) > 3
)x
group by store_code, mth;
```

3) Visualized via Tableau as table titled 'Active Customers'. See the Tableau file

Additional Notes: None.

KPI Description: Most sold top three departments in each stores

KPI formula: sum(qty), department_names, category_name, per store

Steps to realize KPI:

1) Create top3_depart table

```
create table top3_depart(
        store_code bigint,
        qty bigint,
        department text
);
```

2) Insert top3_depart SQL query into table

2-1) store code = 0

```
insert into top3_depart(store_code, qty, department)
select store_code, qty, department
from(
        select store_code, sum(qty) as qty, department_name as department

        from cw.products
        join cw.receipt_lines using(product_code)
        join cw.receipts using(receipt_id)
        where store_code = 0
        group by 1,3
)x
group by 1,2,3
order by qty desc
limit 3;
```

2-2) store code = 1

```
insert into top3_depart(store_code, qty, department)
select store_code, qty, department
from(
        select store_code, sum(qty) as qty, department_name as department

        from cw.products
        join cw.receipt_lines using(product_code)
        join cw.receipts using(receipt_id)
        where store_code = 1
        group by 1,3
)x
group by 1,2,3
order by qty desc
limit 3;
```

2-3) store code = 2

```
insert into top3_depart(store_code, qty, department)
select store_code, qty, department
```

```
        from(
                select store_code, sum(qty) as qty, department_name as department

                from cw.products
                join cw.receipt_lines using(product_code)
                join cw.receipts using(receipt_id)
                where store_code = 2
                group by 1,3
        )x
        group by 1,2,3
        order by qty desc
        limit 3;
```

2-4) store code = 3

```
        insert into top3_depart(store_code, qty, department)
        select store_code, qty, department
        from(
                select store_code, sum(qty) as qty, department_name as department

                from cw.products
                join cw.receipt_lines using(product_code)
                join cw.receipts using(receipt_id)
                where store_code = 3
                group by 1,3
        )x
        group by 1,2,3
        order by qty desc
        limit 3;
```

3) Visualized via Tableau as table titled 'Top 3 department'. See the Tableau file

Additional Notes: None.

KPI Description: Most top three sold categories in provisions dairy department in each store

KPI formula:

Steps to realize KPI:

1) Create top3_cate *table*

```
create table top3_cate(
        store_code bigint,
        qty bigint,
        category text
);
```

*2) Insert top3_* cate *SQL query into table*

2-1) store code = 0

```
insert into top3_cate(store_code, qty, category)
select store_code, qty, category
from(
        select store_code, sum(qty) as qty, category_details as category

        from cw.products
        join cw.receipt_lines using(product_code)
        join cw.receipts using(receipt_id)
        where store_code = 0
        and department_name = 'PROVISIONS DAIRY'
        group by 1,3
)x
group by 1,2,3
order by qty desc
limit 3;
```

2-2) store code = 1

```
insert into top3_cate(store_code, qty, category)
select store_code, qty, category
from(
        select store_code, sum(qty) as qty, category_details as category

        from cw.products
        join cw.receipt_lines using(product_code)
        join cw.receipts using(receipt_id)
        where store_code = 1
        and department_name = 'PROVISIONS DAIRY'
        group by 1,3
)x
group by 1,2,3
order by qty desc
limit 3;
```

2-3) store code = 2

```
insert into top3_cate(store_code, qty, category)
select store_code, qty, category
from(
        select store_code, sum(qty) as qty, category_details as category

        from cw.products
        join cw.receipt_lines using(product_code)
        join cw.receipts using(receipt_id)
        where store_code = 2
        and department_name = 'PROVISIONS DAIRY'
        group by 1,3
)x
group by 1,2,3
order by qty desc
limit 3;
```

2-4) store code = 3

```
insert into top3_cate(store_code, qty, category)
select store_code, qty, category
from(
        select store_code, sum(qty) as qty, category_details as category

        from cw.products
        join cw.receipt_lines using(product_code)
        join cw.receipts using(receipt_id)
        where store_code = 3
        and department_name = 'PROVISIONS DAIRY'
        group by 1,3
)x
group by 1,2,3
order by qty desc
limit 3;
```

3) Visualized via Tableau as table titled 'Top 3 category in dairy depart'. See the Tableau file

Additional Notes: None.

Section 2: Executive Summary

FoodCorp has four stores in three cities (London, Birmingham and Nottingham) in the UK. Stores have its unique code zero to four, *0* is located in Nottingham, *1* is located in Birmingham, *2* is located in Shoreditch, London and *3* is located in Eltham, London. Since FoodCorp wants to launch new marketing on a specific store, all four stores' transaction data was analysed and made six KPIs. As a result, *store code 2*, store located in Shoreditch, London is highly recommending to progress marketing to increase total sales. Up-selling and cross-selling marketing is recommending on provisions diary department especially, on milk category, based on six KPIs following: 1. Total sales vs Total sales in size, 2. New customers, 3. Active customers, 4. Monthly sales, 5. Top 3 department and 6. Top 3 category.

Up-selling and cross selling marketing

KPI 1. Total sales vs Total sales in size

Comparing each stores' total sales with total sales in relative store size. In total sales, *store code 0* is the highest one, however, considering relative store size, *store code 3* becomes the highest total sales store. *Store code 1* shows the least amount of total sales in both measurement and *store code 3* ranked second in total sales but went down to third in relative total sales.

KPI 2. New customers

Analysing monthly new customers and *store code 1* shows the most number of new customers each month. *Store code 3* ranked the second-highest number of new customers which several months show more than 100 new customers. On the other hands, the majority month *in store code 1 and 2* shows less than 50 new customers.

KPI 3. Active customers

Considering KPI 1, 2 and 3*, store code 2* was chosen to be marketing recommend store. Even though store code 1 indicates the most number of active customers who shops more than three times on a monthly basis, *store code 2* shows the second-highest number of active customers. Considering the number of new customers and relative total sales in *store code 2* was low, it could interpret as a potential store that could increase future total sales.

KPI 4. Monthly sales

Through monthly sales KPI that shows difference per cent of previous month sales, it gives information when marketing should in initiated. From March to August shows almost continuous positive per cent of monthly sales in *store code 2*, thus March 2020 is the optimal period to launch marketing.

KPI 5. Top 3 Department

Now the target of marketing need to be specified and the top 3 department KPI provides the answer. 'Provisions dairy', 'grocery 2' and 'fruit & vegetables' is the most sold department in *store code 2*. Since 'Provisions dairy' was the highest one, the marketing should focus on them.

KPI 6. Top 3 category in dairy depart

Since 'provision dairy' recorded as the most sold department, the top three most sold category in provision dairy department was analysed. Four categories were in common in four stores and milk category was the most sold category on *store code 2*. As a result, marketing should be target on milk.

Section 3: Comparative Analysis

FoodCorp is a relatively new company that only operated less than 2 years and there are four stores in three cities (London, Birmingham and Nottingham) in the UK. *Store code 0* is store in Nottingham, *store code 1* is located in Birmingham, *store code 2* is located in Shoreditch, London and *store code 3* is located in Eltham, London. Currently, FoodCorp is planning new marketing in one of the four stores. After analysed four store's data, progressing marketing on store code 2 located in Shoreditch, London is the best option. There are two marketing strategies: (1) up-selling and cross-selling marketing on grocery 3, provisions diary and soft beverages department and (2) repeating marketing on the top spending customers.

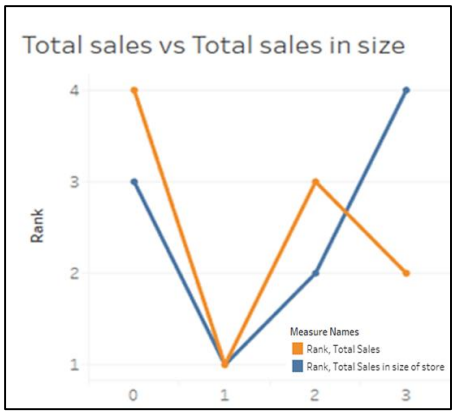
Comparative analysis about six KPIs

KPI 1. Total sales vs Total sales in size

To begin with, comparative analysis of their four stores, comparing their total sales is necessary. *Store code 1* to *code 3* have been launched since April 2018, however, *store code 0* opened on July 2018. *Store code 1's* total sales are £70,527, *store code2's* total sales are £644,315 and *store code 3's* total sales are £396,680, while *store code 1* which begin their sales after three month shows £1,297,625 in total sales. When comparing the total amount of sales, *store code 0* indicated the most profitable store in four stores. Nevertheless, to understand the general performance statistics of each store, considering the relative size of the store is demand. To standardized the size of the store and to achieve the relative size of stores, the number of 'product_code, department_code, category_code and sub_category_code' was chosen to compare. The bigger the size of the store, more types of product would be stored. Since those four codes have equal numbers, I chose 'product_code' to compare each store. Setting *store_code 0* as a standard and considered as 1.00 in store size by dividing 768,756 to 768,756, the number of 'product_code'. Thus, the relative size of stores is calculated by divide each store's the number of 'product_code' by *store code 0's* the number of 'product_code', 768,756. As a result, *store_code 0's* relative store size is 1 and *store 1* is 0.06, *store 2* is 0.52 and *store 3* is 0.29.

store_code	relative store size
0	1.00
1	0.06
2	0.52
3	0.29

The rank of the original total sales of each store and total sales of each store by relative size is different. With original total sales, *store code 0,* store located in Nottingham shows the highest total sales, however, with considering relative size*, store code 3* which is located in Eltham, London is the highest performing store.



Store code 2 goes down to the third rank and *store code 1 shows* the least total sales in both two measurements. To decide the marketing target store, new customer's analysis and active customer's analysis is required.

KPI 2. New customers

New Customers

Mth	Store Code (New Customer) 0	1	2	3
2018-03		106	358	378
2018-04		99	257	418
2018-05		58	129	232
2018-06	1,055	51	113	133
2018-07	568	37	76	126
2018-08	375	36	63	125
2018-09	323	42	61	106
2018-10	247	36	38	100
2018-11	221	23	36	78
2018-12	192	24	51	96
2019-01	183	15	38	109
2019-02	148	32	25	114
2019-03	182	22	46	70
2019-04	149	17	26	59
2019-05	145	16	38	68
2019-06	129	28	31	68
2019-07	84	15	32	76
2019-08	119	11	25	85
2019-09	91	22	30	56
2019-10	89	19	29	55
2019-11	74	11	14	36

New customers in each month analysis show that *store code 0* has the highest number of people. During 18 month of operating, 14 month indicates over 100 new customers which are a significant number of people comparing to other stores. In *store code 3* shows several months that have more than 100 new customers, while *store code 1 and 2* shows most of the month less than 100 new customers.

KPI 3. Active customers

Considering active customer analysis which analysing the number of customers who shops more than three times a month, I highly recommend doing marketing in *store code 2* to achieve better sale performance. Even though in total sales of relative store size indicates the third rank among the four stores, it shows a large number of active customers who shops more than three times a month.

Active Customers

Mth (Re..	Store Code (Repeat3 Custome.. 0	1	2	3
2018-03		10	71	10
2018-04		38	230	93
2018-05		44	235	70
2018-06	191	32	220	45
2018-07	399	47	210	71
2018-08	415	51	241	63
2018-09	381	48	239	78
2018-10	404	45	233	65
2018-11	400	41	236	67
2018-12	442	46	243	74
2019-01	386	41	226	73
2019-02	364	35	204	51
2019-03	400	44	232	83
2019-04	391	39	223	72
2019-05	442	44	241	72
2019-06	423	28	246	59
2019-07	397	34	234	76
2019-08	429	38	261	75
2019-09	377	33	266	71
2019-10	415	40	261	76
2019-11	224	25	172	24

New customer in *store code 2* shows a small group of a new customer each month, however, the number of the active customers is nearly more than 200 people each month. It must have a reason that why the number of active customers is high, but more important part is that there is a high potential to increase their total sales. This is because the gap of the number of new customers and the number of active customers is high that if the company progress accurate marketing could increase the number of new customers and induce more amount of spending of active customers.

Since we decided the target of marketing, now we need to choose when and what kind of marketing should be processed. Monthly sales analysis shows the difference between monthly sales of each month by per cent. Darker the orange colour, higher total months sales compared to the previous month. Except for the first month*, store code 0* shows 8 months of a positive per cent which means the month of sales is higher than the previous month. *Store code 1 and 2* shows 10 month of positive per cent and *store code 3* shows seven months of positive per cent.

KPI 4. Monthly sales

The best period to start marketing is the month that shows higher total sales compare to the previous month. In *store code 2,* several months shows positive per cent, however, from March to August shows almost continuous positive per cent of monthly sales. Since the very last operating month, November 2019's monthly sales were 39.6 per cent lower than the previous month, the company should prepare the marketing for *store code 2* and implement in March 2020.

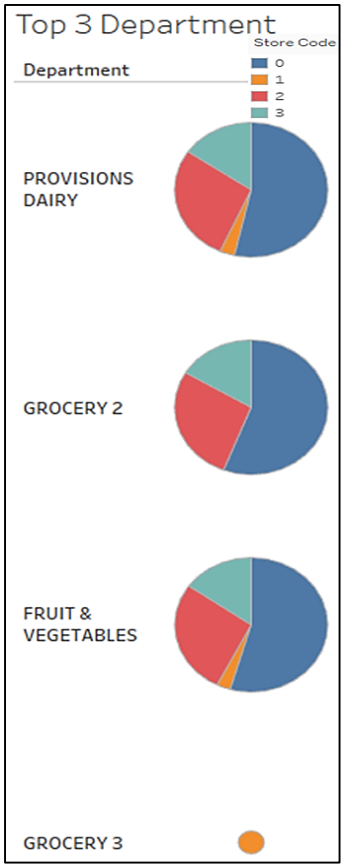
Monthly Sales

Month	Store Code 0	1	2	3
2018-04		165.1%	171.8%	166.4%
2018-05		10.8%	7.2%	-6.3%
2018-06		-17.3%	1.0%	-25.6%
2018-07	82.2%	17.9%	-6.0%	17.5%
2018-08	1.8%	1.6%	9.4%	-2.5%
2018-09	-2.2%	6.6%	-1.2%	4.6%
2018-10	-1.2%	-2.3%	-8.0%	8.3%
2018-11	0.9%	4.4%	5.1%	-5.3%
2018-12	6.1%	-0.8%	10.2%	0.5%
2019-01	-8.8%	-9.0%	-8.4%	9.3%
2019-02	-6.0%	-10.7%	-16.6%	-15.0%
2019-03	10.5%	17.3%	21.1%	13.6%
2019-04	-3.9%	6.9%	-3.2%	-0.9%
2019-05	15.2%	-3.2%	8.9%	-3.9%
2019-06	-1.5%	-15.2%	3.0%	-24.6%
2019-07	-11.2%	-4.9%	-7.7%	26.2%
2019-08	9.7%	7.7%	8.7%	-2.0%
2019-09	-8.7%	3.4%	-1.3%	-4.3%
2019-10	6.2%	11.1%	2.0%	1.2%
2019-11	-37.6%	-35.3%	-39.6%	-38.2%

KPI 5. Top 3 Department

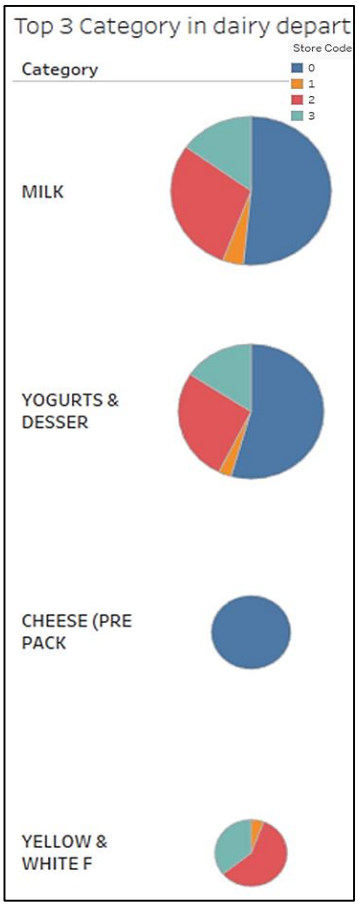
To determine what product should be a marketing target, analysing top sales in terms of department and category was essential. From this analysis, the top three most sold department in each store was discovered. 'Provisions dairy department', 'grocery 2 department', 'fruit & vegetables department' and 'grocery 3 department' was revealed.

'Provisions dairy department'' shows the largest amount of top sales in all department in four stores. *Store code 0* sold 117,293 quantities in provisions dairy department, *store code 1* sold 6,869 quantities, *store code 2* sold 61,174 quantities and *store code 3* sold 33,794 quantities. 'Grocery 2 department' ranked second in top three sold department, however, it only includes in *store code 0,2 and 3*. *Store code 1's* another top three sold department is 'grocery 3 department'.



In store code 2's top three departments is following: 61,174 quantities in 'provisions dairy', 60,736 quantities in 'grocery 2 department' and 59,927 quantities in 'fruit & vegetables department.' Even though these three department's quantity has little difference, up-selling and cross-selling marketing should be targeted 'provisions dairy department'.

KPI 6. Top 3 category in dairy depart



Last KPI is about the three most sold categories in 'provision dairy department' in each store. 'Milk', 'yogurt & dessert', 'cheese' and 'yellow & white f' was chosen and 'milk' category shows the most sold category in 'provision dairy'. *Store code 0* sold 36,047 quantities in 'milk category', *store code 1* sold 2,924 quantities, *store code 2* sold 20,503 quantities and *store code 3* sold 10,483 quantities. 'Milk category' indicates the most sold category in 'provisions dairy' which is the most sold department in *store code 2*. As a result, marketing should focus on 'milk category' in *store code 2*.