Supermarket Analysis

Preparation and Exploration of Data

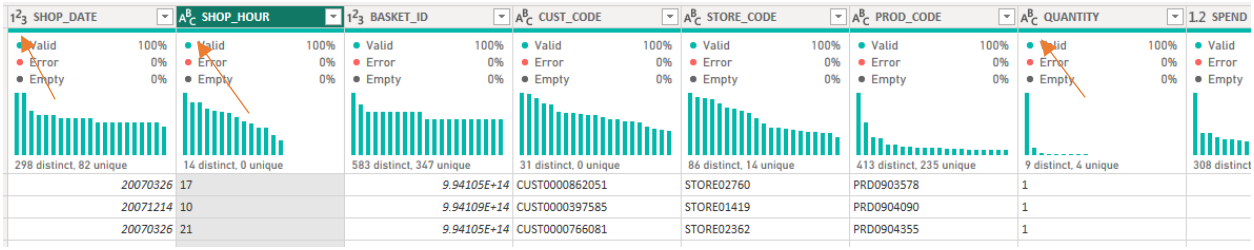
เราจะใช้ข้อมูลการซื้อขาย Supermarket จาก Dunnhumby ซึ่งเป็น Global Customer Data Science Company ที่มีความชำนาญและเป็นที่ยอมรับอย่างแพร่หลายใน Retail Industry ไฟล์ที่ใช้ ประกอบไปด้วย

- dunnhumby-data-dictionary
- product.csv
- store.csv
- transactions.csv

Dataset by Numbers: 100 Customers, 4,292 Products, 300 Stores, 53 weeks

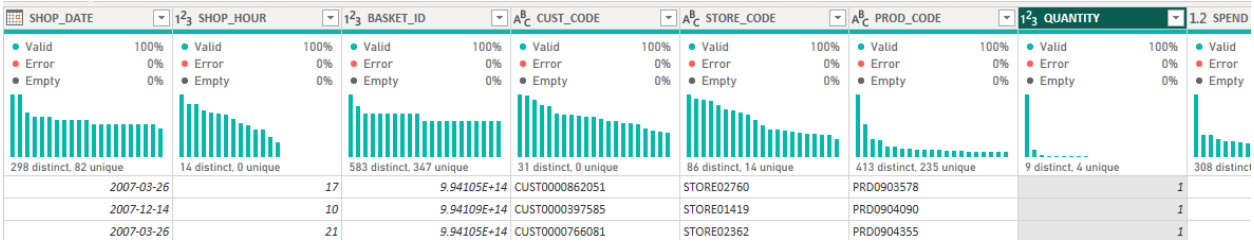
Editing the Data

ตรวจสอบข้อมูลแต่ละ column ว่าควรจะเป็น data type ประเภทไหน ทุก table เช่น จากตาราง transactions SHOP_DATE เดิมเป็นตัวเลขต้องปรับให้เป็นวันที่, SHOP_HOUR เดิมเป็นตัวอักษรต้องปรับให้เป็นตัวเลข



SHOP_DATE	SHOP_HOUR	BASKET_ID	CUST_CODE	STORE_CODE	PROD_CODE	QUANTITY	SPEND
298 distinct, 82 unique	14 distinct, 0 unique	583 distinct, 347 unique	31 distinct, 0 unique	86 distinct, 14 unique	413 distinct, 235 unique	9 distinct, 4 unique	308 distinct
20070326	17	9.94105E+14	CUST0000862051	STORE02760	PRD0903578	1	
20071214	10	9.94109E+14	CUST0000397585	STORE01419	PRD0904090	1	
20070326	21	9.94105E+14	CUST0000766081	STORE02362	PRD0904355	1	

ปรับให้ถูกต้องหน้าตาจะได้ประมาณนี้



SHOP_DATE	SHOP_HOUR	BASKET_ID	CUST_CODE	STORE_CODE	PROD_CODE	QUANTITY	SPEND
298 distinct, 82 unique	14 distinct, 0 unique	583 distinct, 347 unique	31 distinct, 0 unique	86 distinct, 14 unique	413 distinct, 235 unique	9 distinct, 4 unique	308 distinct
2007-03-26	17	9.94105E+14	CUST0000862051	STORE02760	PRD0903578	1	
2007-12-14	10	9.94109E+14	CUST0000397585	STORE01419	PRD0904090	1	
2007-03-26	21	9.94105E+14	CUST0000766081	STORE02362	PRD0904355	1	

Understanding the Data (data-dictionary)

Transactions

Column Name	Description	Type	Sample Values
basket_id	Basket ID. All items in a basket share the same basket_id value.	Char	994100100000020, 994100100000344
cust_code	Customer Code	Char	CUST0000001624, CUST0000001912
shop_week	Identifies the week of the basket	Char	Format is YYYYWW where the first 4 characters identify the fiscal year and the other two characters identify the specific week within the year (e.g. 200735)
shop_date	Date when shopping has been made.	Char	20060413, 20060412
shop_hour	Hour slot of the shopping	Num	0=00:00 - 00:59, 1=01:00 -01:59, ..., 23=23:00 -23:59
store_code	Store Code	Char	STORE00001, STORE00002
quantity	Number of items of the same product bought in this basket	Num	Integer number
spend	Spend associated to the items bought	Num	Number with two decimal digits

Stores

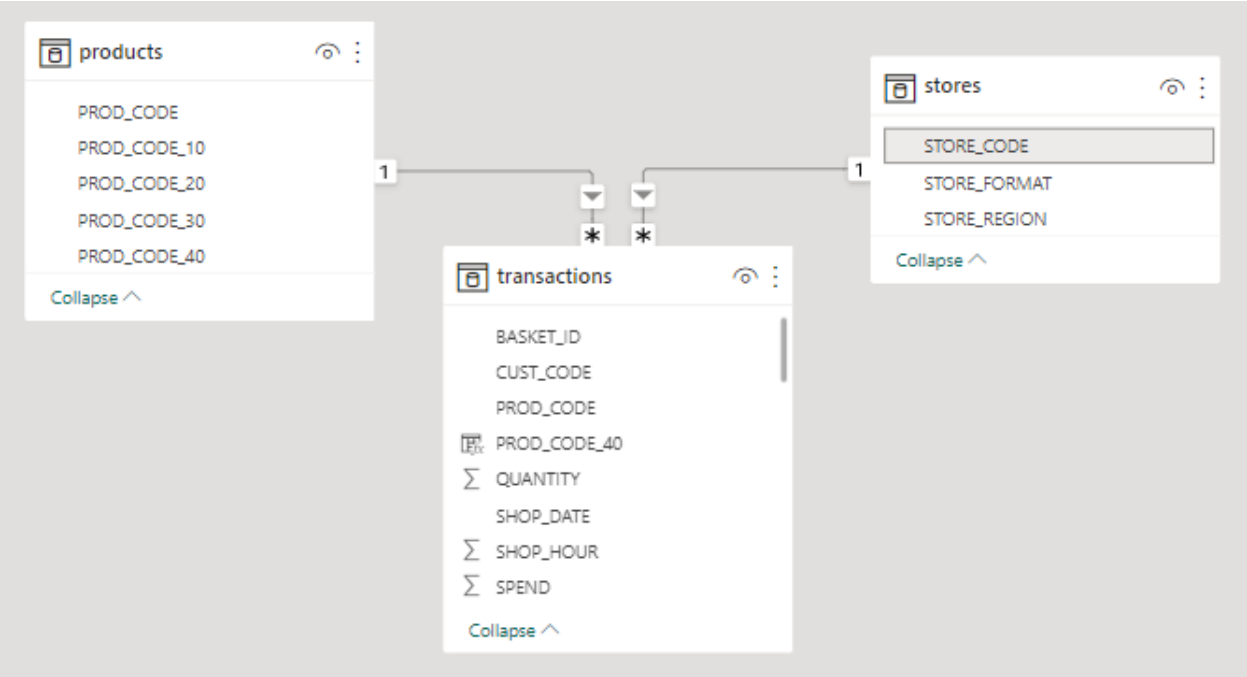
Column Name	Description	Type	Sample Values
store_code	Store Code	Char	STORE00001, STORE00002
store_format	Format of the Store	Char	Large, Medium, Small, XLS
store_region	Region the store belongs to	Char	E02, W01, E01, N03

Products

Column Name	Description	Type	Sample Values
prod_code	Product Code	Char	PRD900001, PRD900003
prod_code_10	Product Hierarchy Level 10 Code	Char	CL00072, CL00144
prod_code_20	Product Hierarchy Level 20 Code	Char	DEP00021, DEP00051
prod_code_30	Product Hierarchy Level 30 Code	Char	G00007, G00015
prod_code_40	Product Hierarchy Level 40 Code	Char	D00002, D00003

Create Data Model and Date table

สร้างให้แต่ละ table เชื่อมถึงกันโดยใช้ Primary key และ Foreign key เพื่อให้แต่ละตารางที่ถูกเชื่อมแล้วสามารถ Filter ไปถึงตาราง Transactions



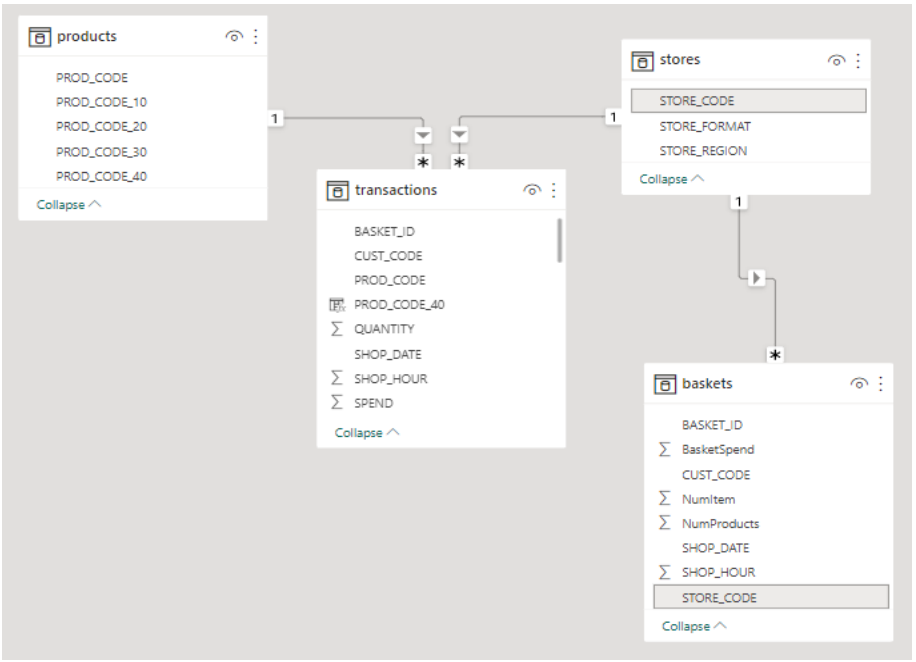
Create Basket Table

เพื่อเตรียมไว้วิเคราะห์ข้อมูลรายใบเสร็จรอไว้ก่อนโดยการ Group by BASKET_ID, SHOP_DATE, SHOP_HOUR, CUST_CODE, STORE_CODE โดยที่แต่ละ Group จะสรุปข้อมูล

- Count จำนวนสินค้าในตะกร้ามีทั้งหมด
- Count จำนวนชิ้นที่ลูกค้าซื้อ
- Sum ยอดรวมที่ใช้ต่อ 1 ตะกร้า

Basket Table

BASKET_ID	SHOP_DATE	SHOP_HOUR	CUST_CODE	STORE_CODE	NumProducts	NumItems	BasketsSpend
994105300393656	*2007-04-15*	*10*	CUST0000359365	STORE02196	*1*	*1*	*0.01*
994105800399319	*2007-05-15*	*8*	CUST0000359365	STORE02196	*1*	*1*	*0.01*
994105900409116	*2007-05-21*	*12*	CUST0000358383	STORE01453	*1*	*1*	*0.01*
994104300181468	*2007-01-29*	*15*	CUST0000022777	STORE00277	*1*	*1*	*0.01*
994104900444414	*2007-03-13*	*12*	CUST0000422101	STORE02561	*1*	*1*	*0.01*
994107500211815	*2007-09-12*	*20*	CUST0000061889	STORE00272	*1*	*1*	*0.01*

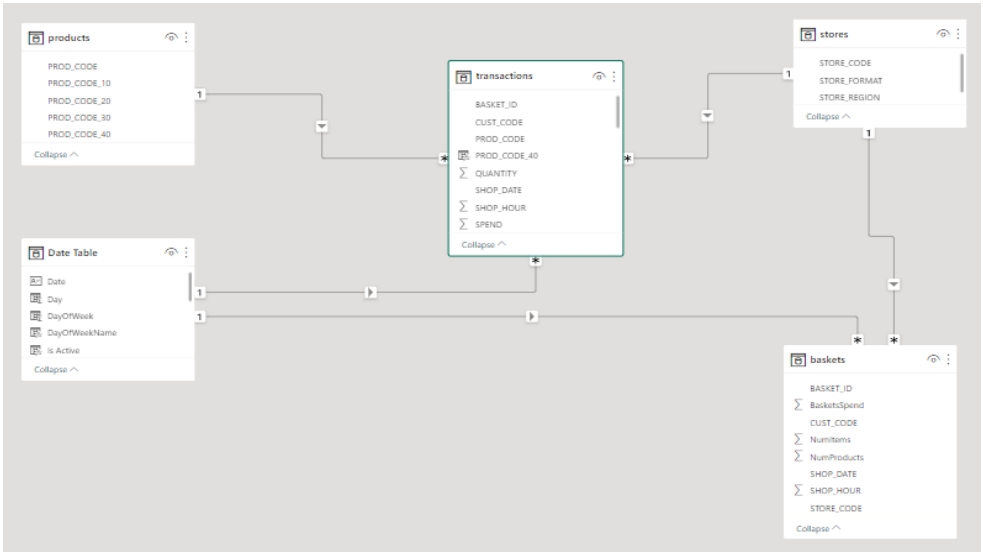


จากการ group by และ aggregate นี้จะทำให้เหลืออยู่ 7,476 จาก 47,533 rows ทำให้เวลาจะวิเคราะห์แค่มุมนี้เวลาดึงข้อมูลจากตารางที่สรุปไว้แล้วจะทำให้เราประมวลผลไวขึ้น

Create Date Table

```
1 Date Table = CALENDARAUTO()
```

Date	Year	Month	Day	Quarter	DayOfWeek	DayOfWeekName	MonthName	YearQuarter	Is Active	YearMonth
2007-01-01 0:00:00	2007	1	1	1	2	Monday	January	2007 Q1	True	2007-01
2007-01-02 0:00:00	2007	1	2	1	3	Tuesday	January	2007 Q1	True	2007-01
2007-01-03 0:00:00	2007	1	3	1	4	Wednesday	January	2007 Q1	True	2007-01
2007-01-04 0:00:00	2007	1	4	1	5	Thursday	January	2007 Q1	True	2007-01
2007-01-05 0:00:00	2007	1	5	1	6	Friday	January	2007 Q1	True	2007-01
2007-01-06 0:00:00	2007	1	6	1	7	Saturday	January	2007 Q1	True	2007-01



Create Measure and Customer Table with DAX

Total Spend (ยอดขายรวม)

Basket Size (ยอดขายรวมต่อ 1 ใบเสร็จ)

Avg Basket Size (ยอดขายเฉลี่ยรวมต่อ 1 ใบเสร็จ)

Total Basket (จำนวนใบเสร็จทั้งหมด)

Cumulative 7d Spend by PROD_CODE_40 (ยอดขายสะสมของ PROD_CODE_40 ใน 7 วันที่ผ่านมา๗

% Cumulative Spend by PORD_CODE_40 ๖ยอดขายสะสมของ PROD_CODE_40 เป็นกี่เปอร์เซ็นต์ของรายได้ทั้งหมดใน 7 วันที่ผ่านมา)

Last 7 days Spend คือ (ยอดขายรวม 7 วันล่าสุด)

Prev Last 7 days Spend คือ (ยอดขายรวม 7 วันก่อนหน้า)

% Diff 7d Spend (เปอร์เซ็นต์ของยอดขายที่เปลี่ยนไปในรอบ 7 วันที่ผ่านมา)

Total YTD Spend (ยอดขายตั้งแต่ต้นปีจนถึงปัจจุบัน)

Product Penetration (ลูกค้ากี่เปอร์เซ็นต์ที่ซื้อสินค้าหมวดหมู่นี้)

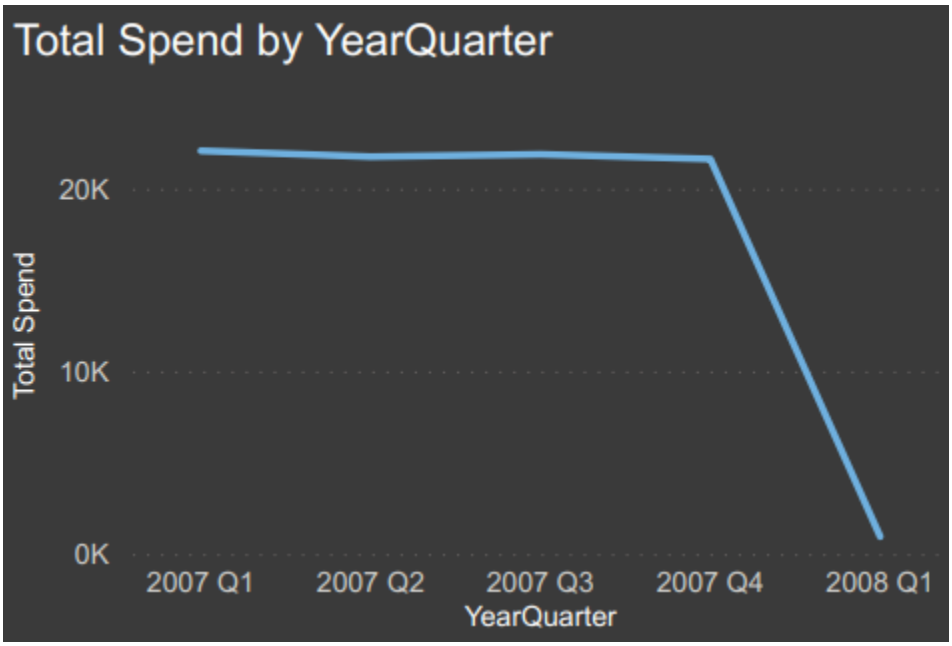
Customer Table (and insert column R,F,M score)

CUST_CODE	Lifetime Total Spend	Last Month Total Spend	Prev Month Total Spend	Before Prev Month Total Spend	Lifecycle Stage	Days Since Last Visit	Total Basket per Month	R Score	F Score	M Score
CUST0000970308	1818.60	498.70	203.03	1116.87	Repeat	6	10.36	3	5	
CUST0000969924	205.51	15.26	17.46	172.79	Repeat	2	9.00	4	5	
CUST0000955714	858.59	134.03	31.09	693.47	Repeat	6	4.45	3	2	
CUST0000947532	505.67	36.42	23.72	445.53	Repeat	1	6.67	5	4	

หลังจากที่สร้าง DAX Measure เตรียมไว้เรียบร้อยแล้ว ต่อไปจะเอาสิ่งที่สร้างไว้มาใช้งานในการสร้าง Dashboard และ วิเคราะห์ในแง่มุมต่าง ๆ

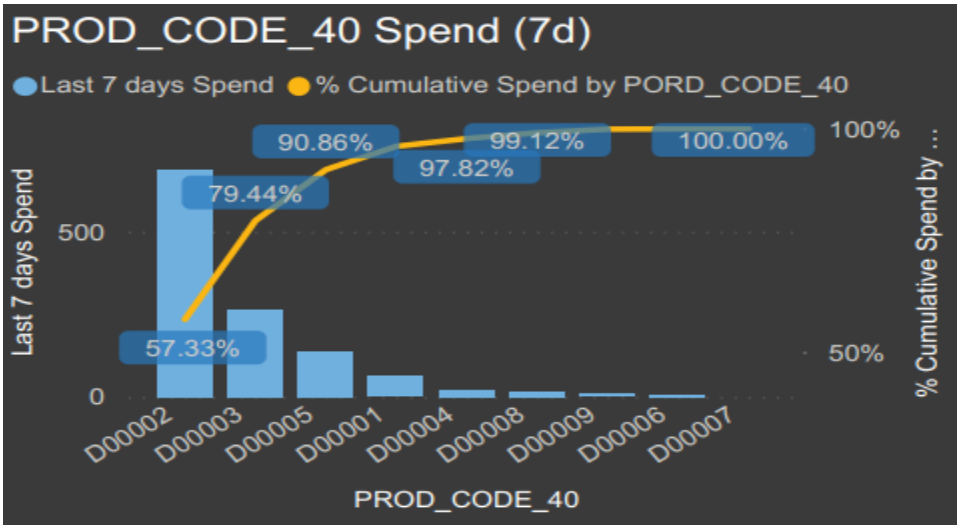
Sales Overview Dashboard

1.Trend ของยอดขายในแต่ละ Quarter ที่ผ่านมาเป็นอย่างไร (Total Spend by YearQuarter)



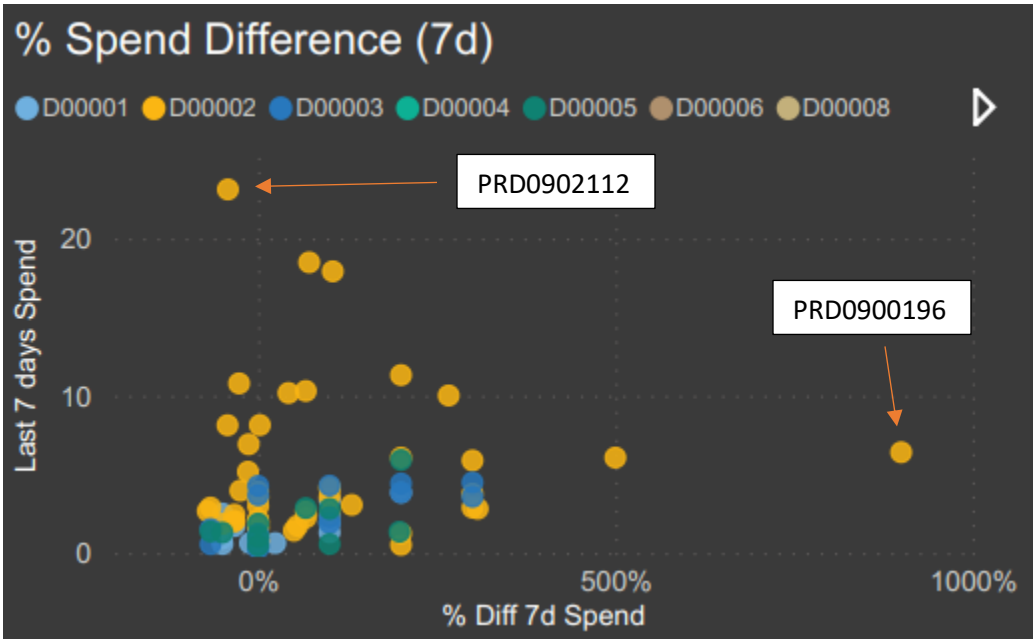
เลือก Line chart เพื่อดูแนวโน้มของยอดขายว่ามีทิศทางขึ้นหรือลง จาก chart จะเห็นแนวโน้มค่อนข้างคงดี แต่ที่ช่วงกราฟตกในช่วง 2008 Q1 เนื่องจากชุดข้อมูลที่เก็บมาวันสุดท้ายคือ 06-01-2008

2.Product Category ไหนสร้างรายได้มากสุด ใน 7 วันที่ผ่านมา + The 80/20 Rule (PROD_CODE_40 Spend (7d)



เลือก Line and column chart โดยที่จำนวนแท่งจะแทนยอดขายของแต่ละ product (code40) และ เส้นที่อยู่เหนือแท่ง คือ Cumulative Spend หรือ ยอดขายสะสม (คิดเป็น %) จาก chart จะเห็นว่า D00002 และ D00003 มียอดขายสะสม คิดเป็นประมาณ 80% ของยอดขายทั้งหมด พอเรารู้ว่าทั้ง 2 Product Category นี้สร้างรายได้ให้เราเยอะแล้ว เราจะได้นำไปประกอบข้อมูลการทำ Promotion หรือ ไปพัฒนา Product ให้ดีมากยิ่งขึ้น

3.Product Code ไหนยอดขายเปลี่ยนแปลงไปอย่างไรใน 7 วันที่ผ่านมา (% Spend Difference)

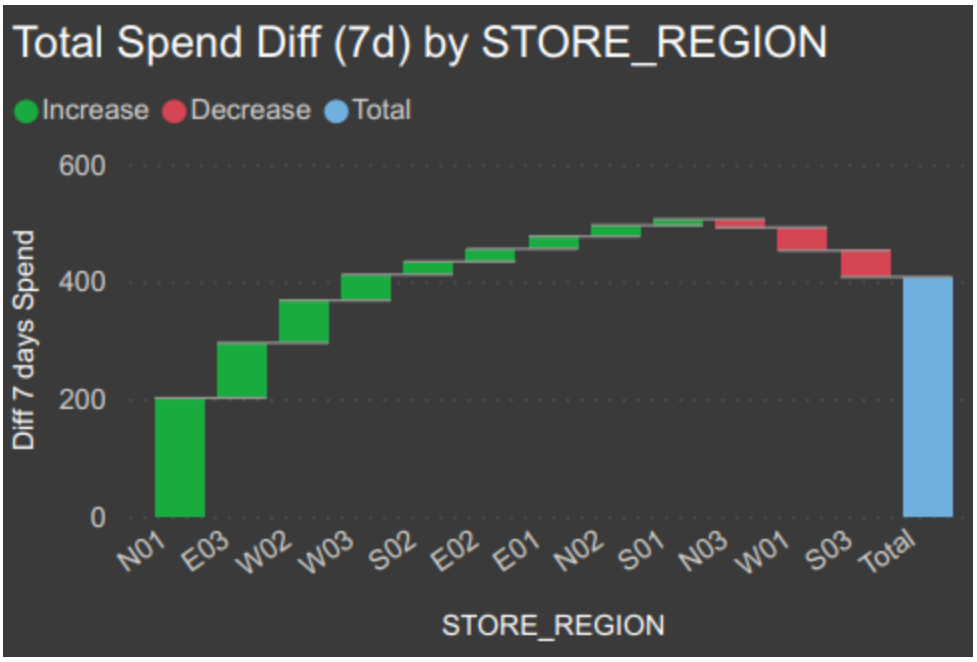


PROD_CODE	PRD0902112
PROD_CODE_40	D00002
% Diff 7d Spend	-42.32%
Last 7 days Spend	23.12
Prev Last 7 days Spend	40.08

PROD_CODE	PRD0900196
PROD_CODE_40	D00002
% Diff 7d Spend	900.00%
Last 7 days Spend	6.40
Prev Last 7 days Spend	0.64

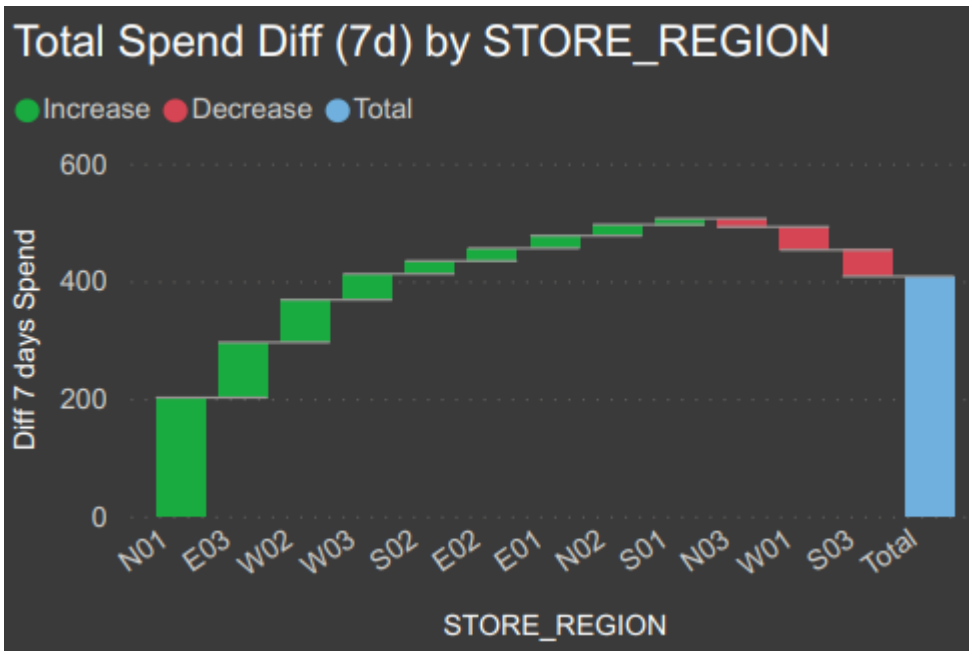
จะเห็นว่า 7 วันที่ผ่านมาเราจะมาดูจุดโดดๆในแต่ละจุดเช่น Product code (PRD0902112) ใน Category D00002 สัปดาห์ที่แล้ว 40.08 แต่สัปดาห์นี้เหลือ 23.12 ทำให้ยอดขายลดลง -42.32 % หรือ D00002(PRD0900196) สัปดาห์ที่แล้ว 0.64 แต่สัปดาห์นี้ขึ้นไป 6.40 ยอดขายเพิ่มขึ้น 900 % (ถ้าชุดข้อมูลเยอะกว่านี้อาจจะเอา ชื้อสินค้า ชื่อแบรนด์ มาวิเคราะห์ในมุมอื่นๆได้อีก)

4.สัดส่วนของยอดขายเพิ่มขึ้นหรือลดลงเท่าไร 7วันที่ผ่านมาโดยแบ่งตาม Product Catagory code 40 (Total Spend Diff by PROD_CODE_40)



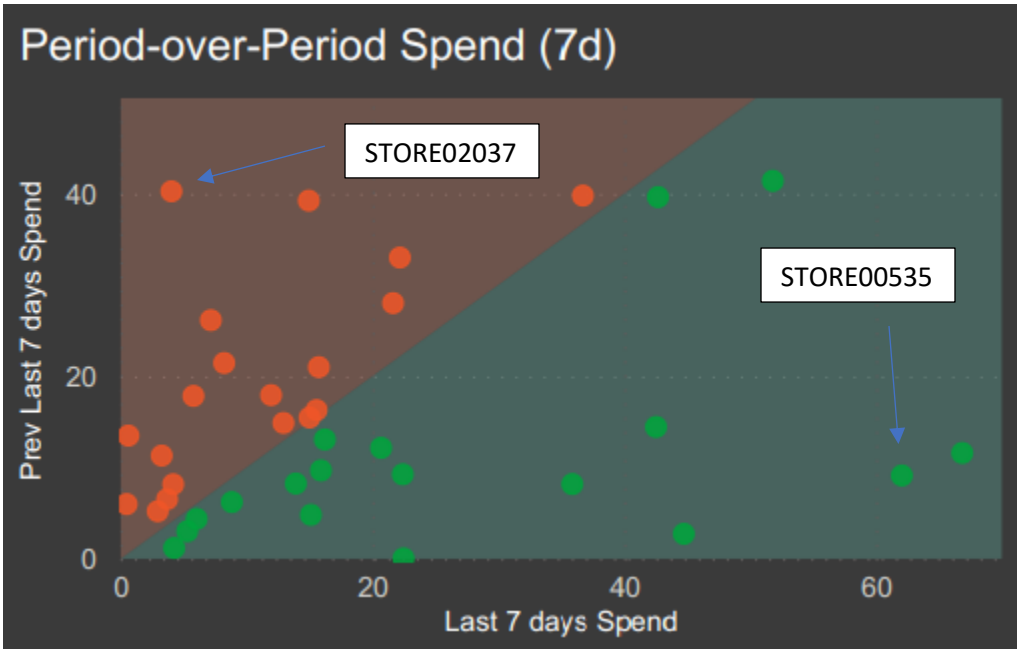
ในการเลือก Waterfall chart จะทำให้เห็นสัดส่วนการเปลี่ยนแปลงของยอดขายของในแต่ละหมวดหมู่เพิ่มขึ้นหรือลดลงเท่าไรในช่วงเวลาที่เราสนใจ ใน chart นี้จะเห็นว่าจาก Total ทั้งหมด DP0002,DP0003,DP0005,DP0001,DP0009,DP0006 เพิ่มขึ้น DP0004,DP0008 ลดลง

5.สัดส่วนของยอดขายเพิ่มขึ้นหรือลดลงเท่าไร 7วันที่ผ่านมาโดยแบ่งตาม Store Region (Total Spend Diff by STORE_REGION)



เปลี่ยนไปดูในมุม Store Region กันบ้าง จะเห็นว่ามีเพียง 3 Region คือ N03,W01,S03 ที่ยอดขายลดลง

5.Region ไหนยอดขายเปลี่ยนแปลงไปอย่างไรใน 7 วันที่ผ่านมา Vs. 7 วันก่อนหน้า (Period-over-Period)



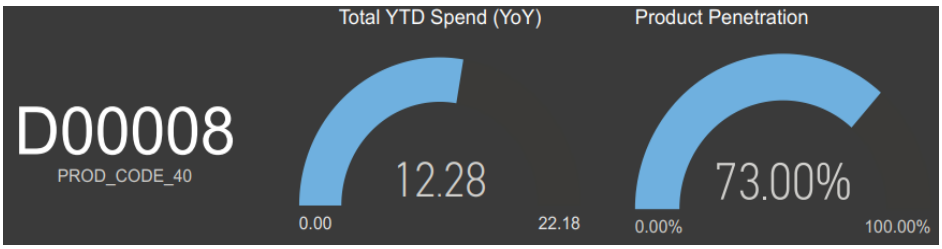
STORE_CODE	STORE02037
Last 7 days Spend	4.03
Prev Last 7 days Spend	40.26
First STORE_REGION	N03
Diff 7 days Spend	-36.23
% Diff 7d Spend	-89.99%

STORE_CODE	STORE00535
Last 7 days Spend	62.06
Prev Last 7 days Spend	9.12
First STORE_REGION	E03
Diff 7 days Spend	52.94
% Diff 7d Spend	580.48%

ใช้ Scatterplot and Symmetry Shading จะเห็นว่า Store ไหนยอดขายดีไม่ดีอย่างไร โดยยอดขายที่อยู่เหนือเส้น คือ 7 วันก่อนหน้า ยอดเยอะกว่า 7 วันที่ผ่านมา คือ store ที่ยอดขายตก และ ยอดขายที่อยู่ใต้เส้น คือ 7 วันที่ผ่านมา ยอดเยอะกว่า 7 วันก่อนหน้า คือ store ที่ยอดขายขึ้น เช่น STORE02037 เมื่อ 7 วันก่อนหน้ายอดขาย 40.26 แต่ 7 วันที่ผ่านมา ยอดขายเหลือเพียง 4.03 ติดลบ -89.99% ในขณะที่ STORE005035 เมื่อ 7 วันก่อนหน้ายอดขายได้เพียง 9.12 แต่ 7 วันที่ผ่านมา ยอดขายพุ่งถึง 62.06 บวก 580.48%

Product Detail Dashboard

1.ยอดขายของปัจจุบันเทียบกับปีที่แล้ว (Total YTD Spend (YoY))

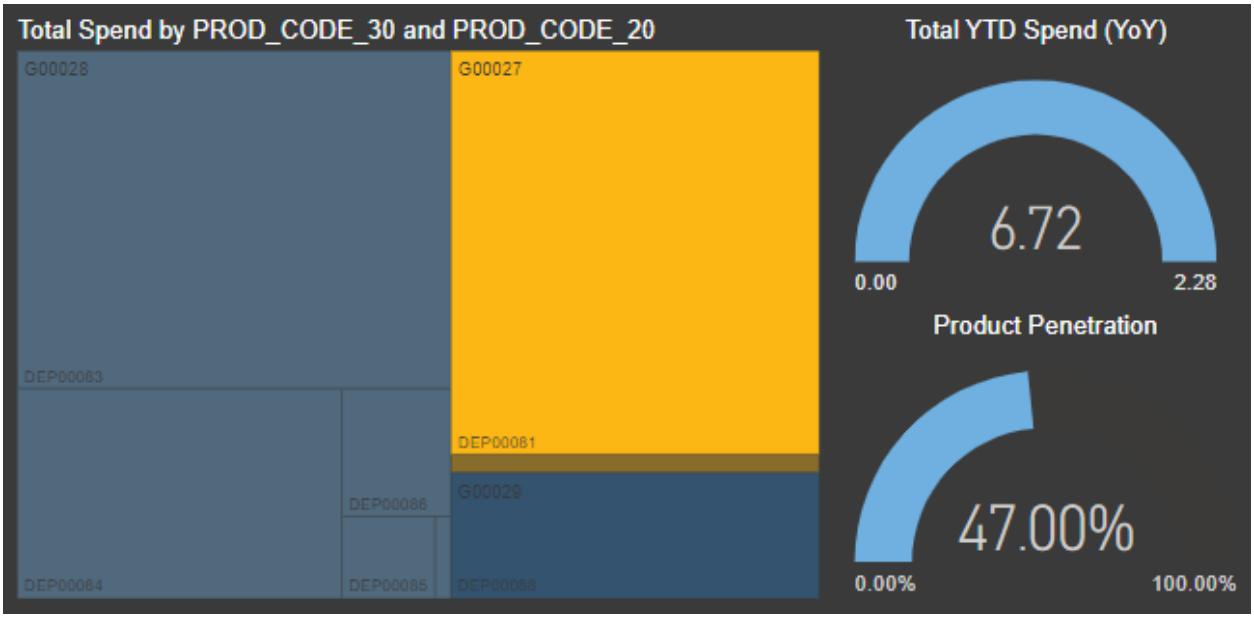


เลือก Gauge เพื่อดูยอดขายได้จัดขึ้นตั้งแต่ต้นปีถึงปัจจุบันว่ายอดขายของเราตอนนี้อยู่ที่เท่าไรเทียบกับปีที่แล้ว จาก visual Product code D00008 ปีที่แล้วขายได้ 22.18 แต่ปัจจุบันขายได้ไปแล้ว 12.18

2.จากลูกค้าทั้งหมดมีลูกค้ากี่เปอร์เซ็นต์ที่ซื้อของใน Category ที่เราสนใจ (Product Penetration)

จากลูกค้าในตอนนี้ทั้งหมด มีลูกค้าซื้อสินค้าหมวด D00008 อยู่ 73 %

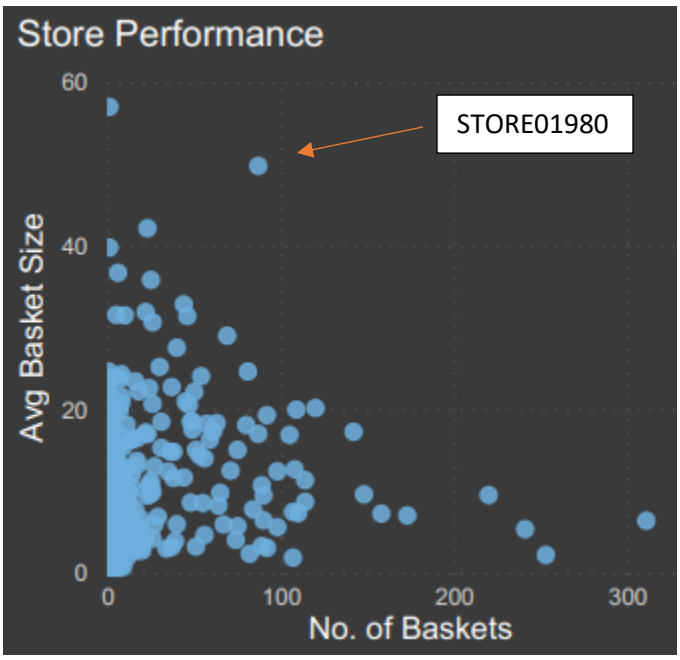
3.สัดส่วนภาพรวมใน Sub Category ของ D0008 ว่าหมวดหมู่ไหนสร้างรายได้เยอะหรือน้อยอย่างไร (Total Spend by PROD_CODE_30 and PROD_CODE_20)



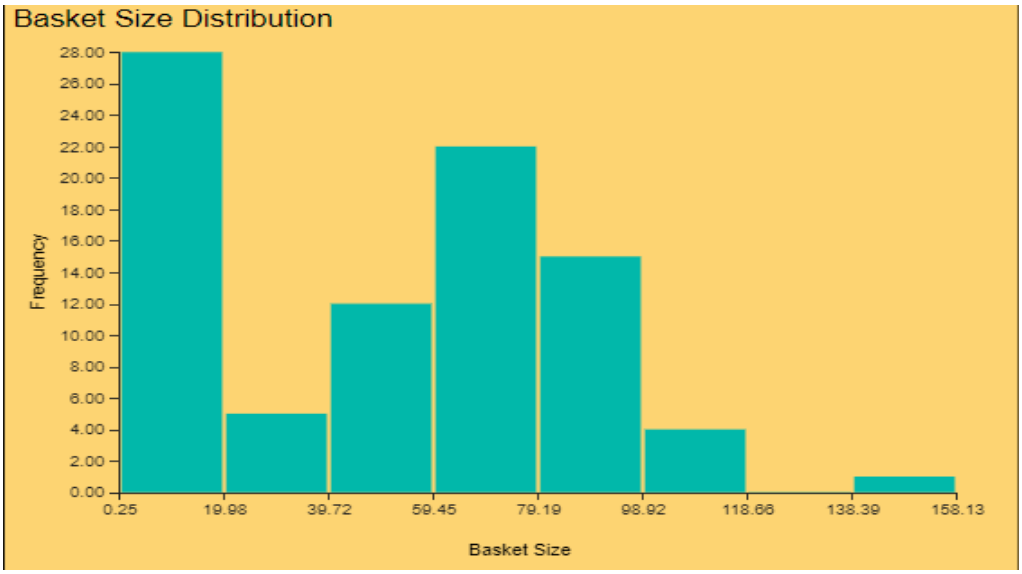
ใช้ Treemap ดูยอดขาย Sub Category ของ Code 30 และ Code 20 จาก chart ที่ Product code 20 DEP00081 จะเห็นยอดขายวันนี้เทียบกับปีที่แล้ว 6.72 จากลูกค้าที่ซื้อทั้งหมด 47%

Store Overview Dashboard

1.ดูความสัมพันธ์ระหว่าง Avg Basket Size และ No. of Baskets (Store Performance) และดูการกระจายตัวของ **basket size** โดยใช้ **Histogram (Basket Size Distribution)**



STORE_CODE	STORE01980
No. of Baskets	87
Avg Basket Size	49.75
STORE_FORMAT	LS
STORE_REGION	N02



โดยเฉลี่ยแล้ว สาขา STORE01980 มี Basket Size อยู่ที่ 49.75 จำนวน Basket ทั้งหมด 87 ใน สาขา และ การกระจายตัว Basket Size ที่ STORE01980 มีช่วง 0.25-19.98 ที่โดดและ มีช่วง 19.98-118.66 ที่ peak ไปอยู่59.45-79.19

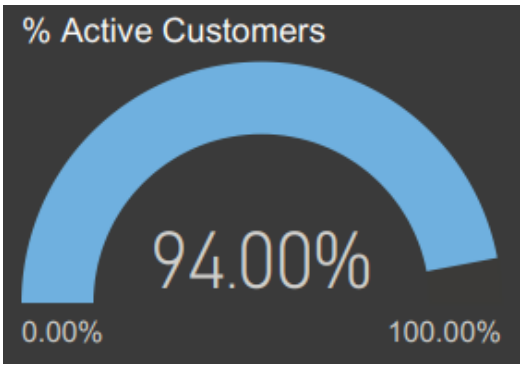
2.แต่ละสาขามีความหนาแน่นของลูกค้าที่เข้ามาในแต่ละช่วงเวลาเป็นอย่างไร

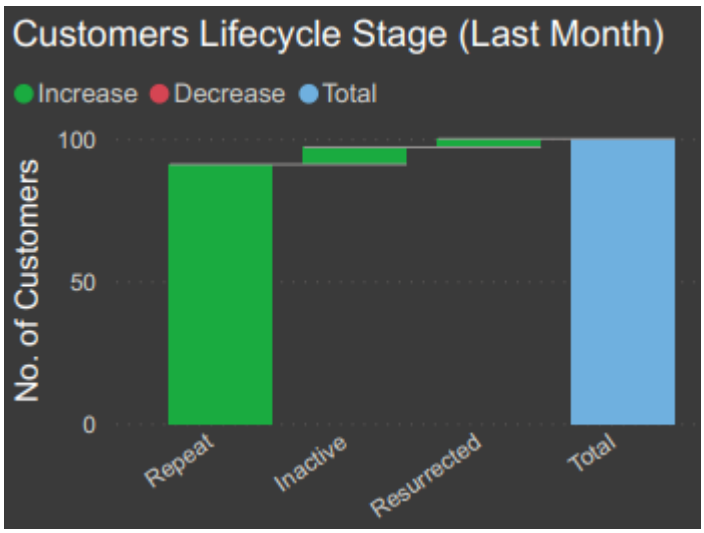
DayOfWeekName	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Total
Sunday	1	1	1	1	2	5		4	3	3	7	2	3	4	37
Monday	2	1	1	2	4	1	3	5	1	2	4	1	1	1	29
Tuesday	3			1	2			2	2	2	7	3	2	2	26
Wednesday	3	2			3	2	2	2	5	3	1	4	2	5	34
Thursday	1	1		3			1	3	4	8	5	4	1	1	32
Friday	1	1	1	1	4	3		3	2	1	3	3	1	5	29
Saturday	3	1		3			3	3	4	5	3	5	1	2	33
Total	**14**	**7**	**3**	**11**	**15**	**11**	**9**	**22**	**21**	**24**	**30**	**22**	**11**	**20**	**220**

ยกตัวอย่าง สาขา STORE00980 ในช่วง 9.00 A.M.-12.00 A.M. เป็นช่วงเวลาที่คนมาน้อย อาจจะลองทำโปรโมชั่นออกมาเพื่อเรียกลูกค้าเพิ่ม ส่วนในช่วง 15.00 A.M.-19.00 A.M. ลูกค้ามาเยอะๆก็อาจทำแผนเรื่องการจัดการด้านอื่นๆ เช่น ถ้าลูกค้าเยอะๆการบริการไม่ทันอาจจะต้องจัดจ้างพนักงานหรือ parttime เพิ่มหรือเปล่า เพื่อรองรับงานบริการ หรือเพื่อรักษาความพึงพอใจของลูกค้าได้

Customer Overview Dashboard

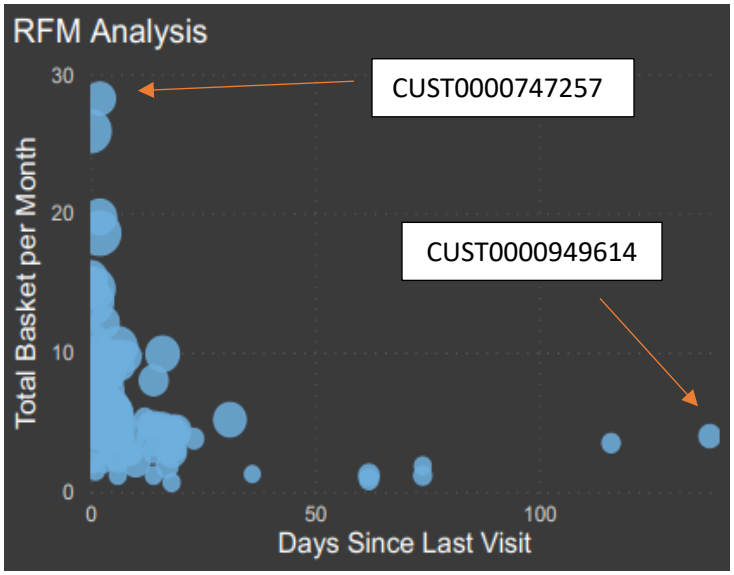
1.ลูกค้าทั้งหมดที่มาซื้อของกับเราอยู่ใน Stage ไหนกันบ้าง





ตอนนี้เรามีลูกค้าอยู่ทั้งหมด 3 Stage คือ Repeat Inactive Resurrected และเราจะสนใจลูกค้าที่ Active คือ ในลูกค้า 100% มี Active อยู่ 94 % (ไม่นับ Inactive)

2.วิเคราะห์ลูกค้า 3 กลุ่ม (RFM Analysis)



CUST_CODE	**CUST0000747257**
Days Since Last Visit	**2**
Total Basket per Month	**28.25**
Sum of Lifetime Total Spend	**1010.13**

CUST_CODE	**CUST0000949614**
Days Since Last Visit	**138**
Total Basket per Month	**4.00**
Sum of Lifetime Total Spend	**310.83**

ลูกค้า CUST0000747257 พึ่งมาซื้อเราเมื่อ 2 วันที่แล้ว มาซื้อ 28.25 ครั้งต่อเดือน และ ใช้เงินทั้งหมด 1010.13 ลูกค้าคนนี้ต้องรักษาเอาไว้ ในขณะที่ ลูกค้า CUST0000949614 ไม่ได้มานานกว่า 138 วันแล้วอาจจะไม่ต้องใส่ใจกับลูกค้ารายนี้มากเท่ารายแรก

3.แบ่งกลุ่มลูกค้าด้วย RFM Score (โดยใช้ Percentile ที่ 20%, 40%, 60%, 80% ตาม score)

- R (Recency) ลูกค้าหายไปนานแค่ไหน เรียงคะแนนตาม 5 ดีสุด ไป 1 น้อยสุด ยกตัวอย่างเช่น ลูกค้าคะแนน 5 คือ คะแนนต่ำกว่า Percentile ที่ 20% หมายความคือ ลูกค้าคะแนนต่ำกว่า 20% พึ่งมาซื้อของล่าสุดกับเรา

- F (Frequency) ลูกค้ามาซื้อของบ่อยแค่ไหน เรียงคะแนนตาม 5 ดีสุด ไป 1 น้อยสุด ยกตัวอย่างเช่น ลูกค้าคะแนน 5 คือ คะแนนสูงกว่า Percentile ที่ 80% หมายความคือ ลูกค้าคะแนนสูงกว่า 80% ที่ซื้อของกับเราบ่อยๆ

- M (Monetary) ลูกค้าจ่ายเงินกับเราเยอะๆ เรียงคะแนนตาม 5 ดีสุด ไป 1 น้อยสุด ยกตัวอย่างเช่น ลูกค้าคะแนน 5 คือ คะแนนสูงกว่า Percentile ที่ 80% หมายความคือ ลูกค้าคะแนนสูงกว่า 80% ที่ผ่านมาเค้าจ่ายเงินกับเราเยอะๆจากที่ผ่านมา