ISSS602 Data Analytics Lab

Customer Segmentation Analysis of A Pharmaceutical Distributor： AmerisourceBergen

Technical Report

Submitted By: Liu Cuiyi

Submitted data: 20/10/2019

1. Context

AmerisourceBergen, the largest wholesaler of medical and surgical materials in the United States, set a goal to improve the efficiency of products distribution. To realize this target, it is essential to explore into its customer needs by doing customer segmentation. The company distributes a wide variety of products to all kinds of pharmacies, including hospitals, clinics, retail pharmacies, physician offices and so on. Customer segmentation is a way to have a better understanding of the customers. By doing so, we can strategize a more efficient allocation of resources to different target-customers, as a result, to improve our services and enhance profitability.

2. Data Preparation

The datasets contain 7 data tables and 1 data description table, these data give details to product classification and description, as well as to pharmacy masters and to point-of-sales' transactions during the first half year of 2016.

2.1. Data Quality Issues

2.1.1. Incorrect modeling type of variable "ZIP_3_CD" and description variables

Zip codes' modeling type should be set as nominal type.

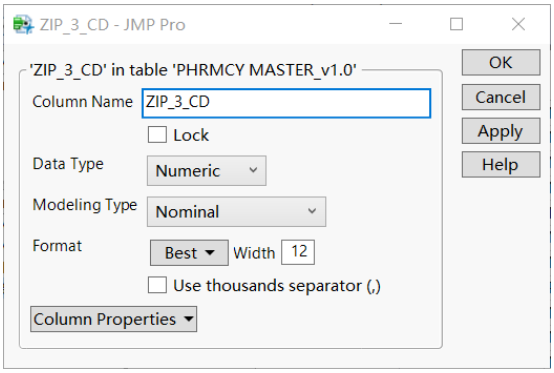


Figure 1

Description variables' modeling type should be set as unstructured text to fid key words, like "PROD_DESC", "SEG_DESC", "SUB_CAT_DESC".

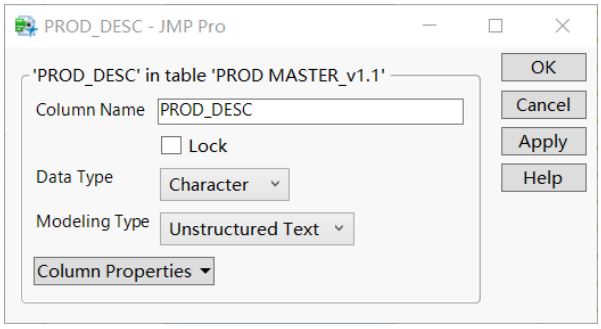


Figure 2

2.1.2. Correct data format of some variables

We couldn't directly change format of variable "SLS_DTE_NBR" to date, because it would produce unexpected error. First, we should change its data type to character. Then, change it to numeric, choose ordinal as modeling type, and choose appropriate format.

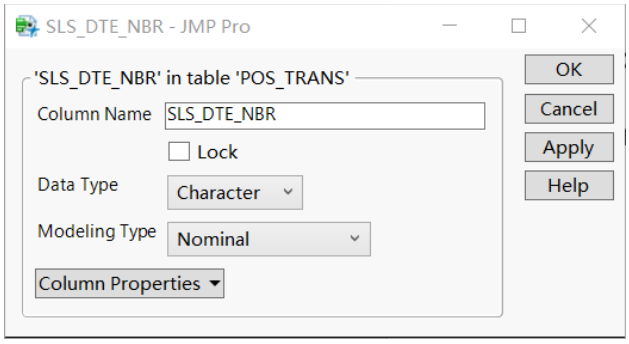


Figure 3

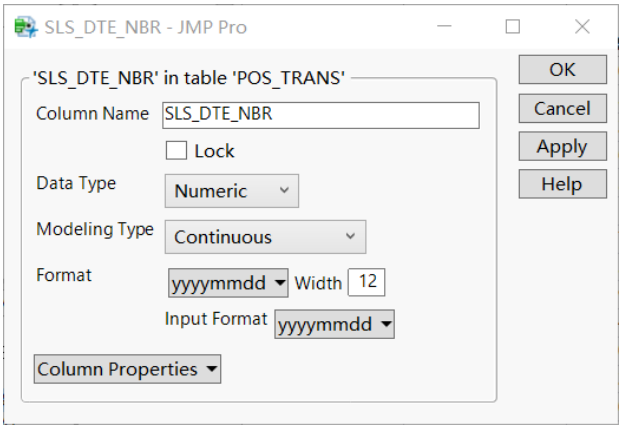


Figure 4

Choose currency (US dollar) as the data format of variable "EXT_SLS_AMT".

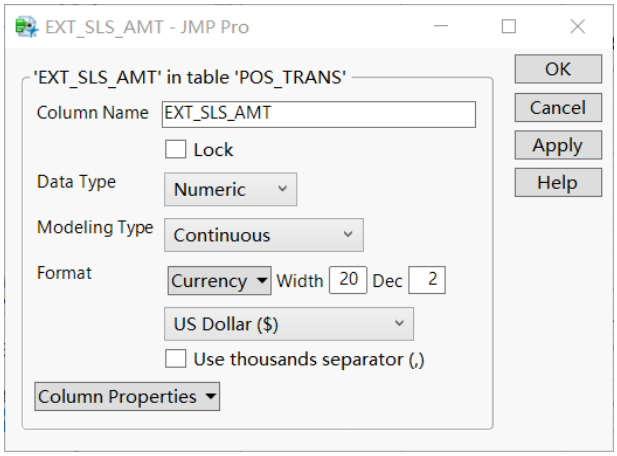


Figure 5

2.1.3. Incorrect "0" values in the variable "SLS_QTY" in table "POS_TRANS"

A zero value in sales quantity with a negative sales amount could be a refund without product returned due many reasons. A zero value in sales quantity with a zero value in sales amount could show few inferences, so we can hide and exclude these 25 rows by using row selection tool as shown below.

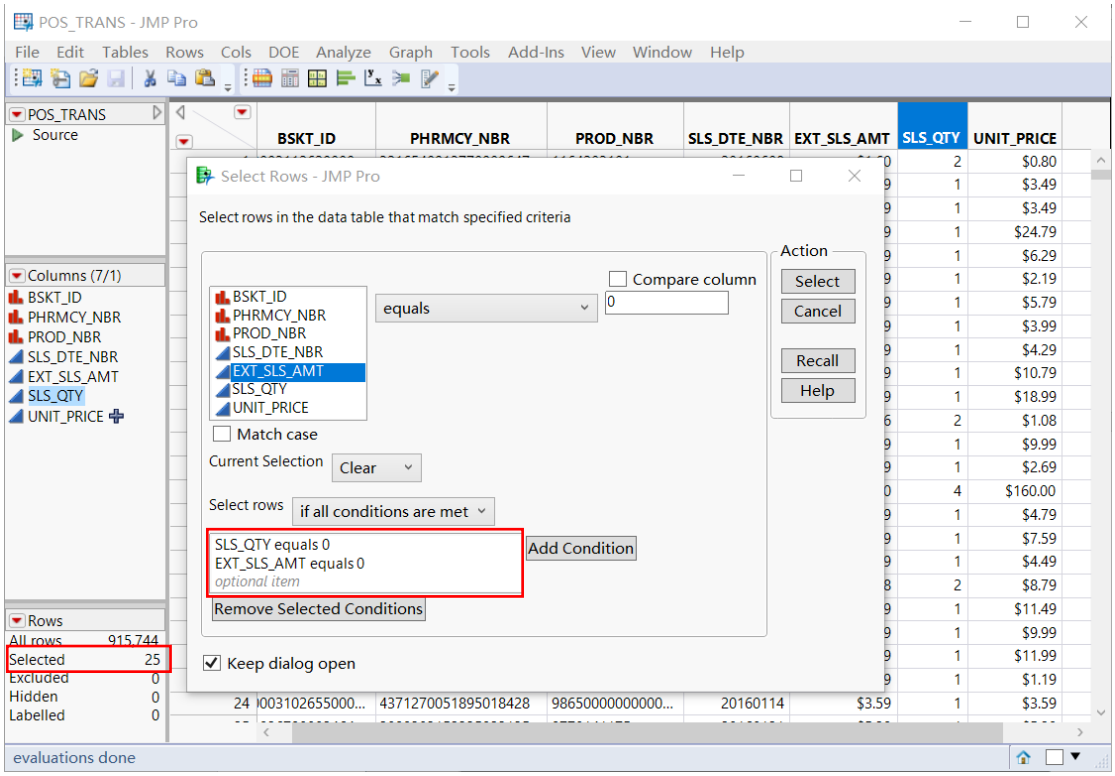


Figure 6

2.1.4. Get Unit Price of Products and solve Outliers issue

Insert a formula column named "UNIT_PRICE" to compute the unit price of each product transacted in the table "POS_TRANS" as shown below (Figure 7). Recode values which equal or smaller than zero into missing value (Figure 8).

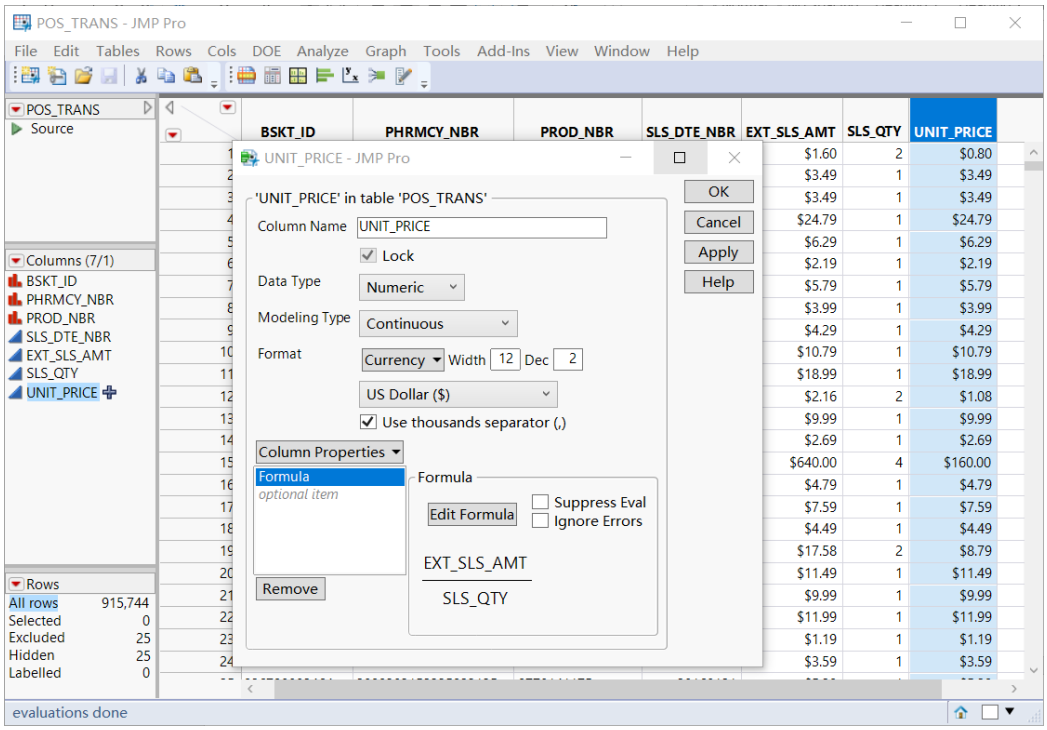


Figure 7

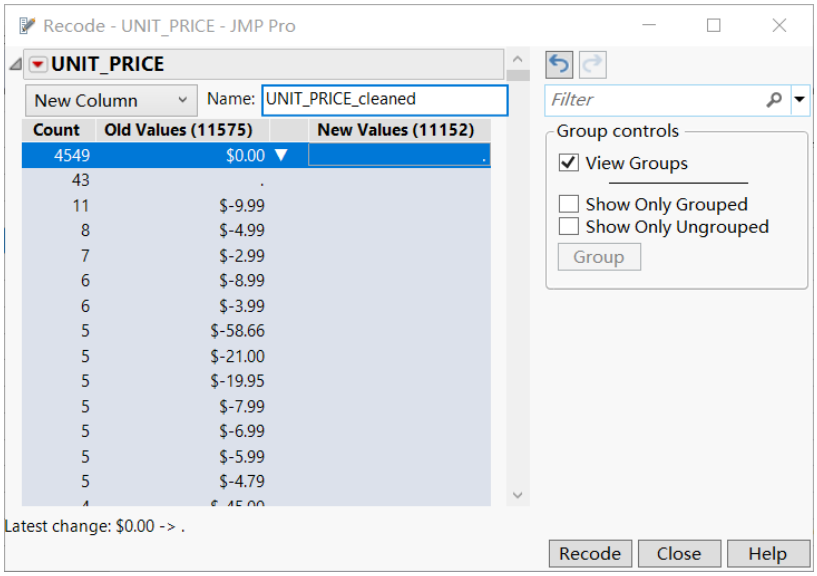


Figure 8

To get outliers of price value of each product, we should examine each product separately. First, set column "PROD_NBR" as the group-by column. Then, create a formula column to get 90th percentile of prices of each product (Figure 10). After that, we could set the upper limit of price, like 3 times the 90th percentile, to recognize outliers and the price greater than the limit should be outliers. Here, we insert a formula column as shown below (Figure 11).

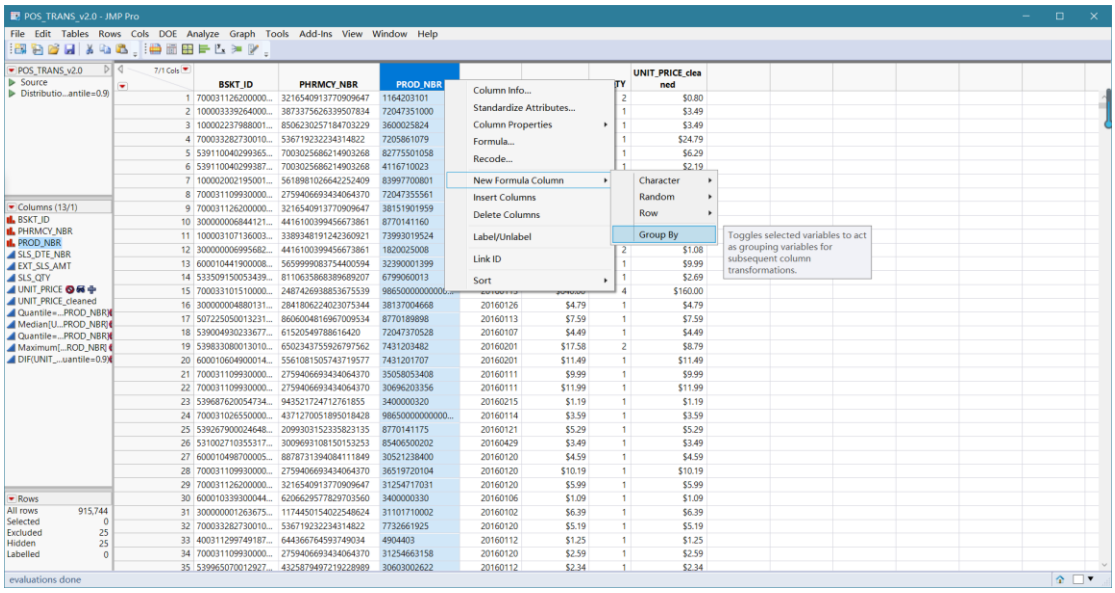


Figure 9

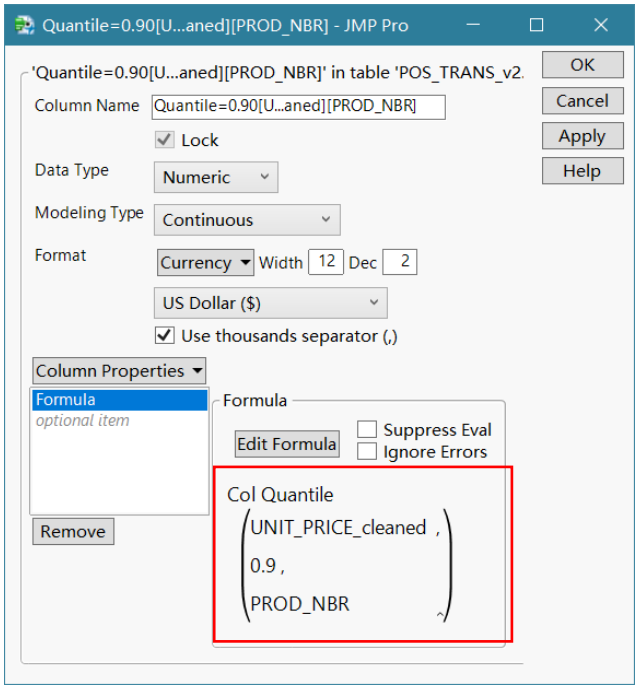


Figure 10

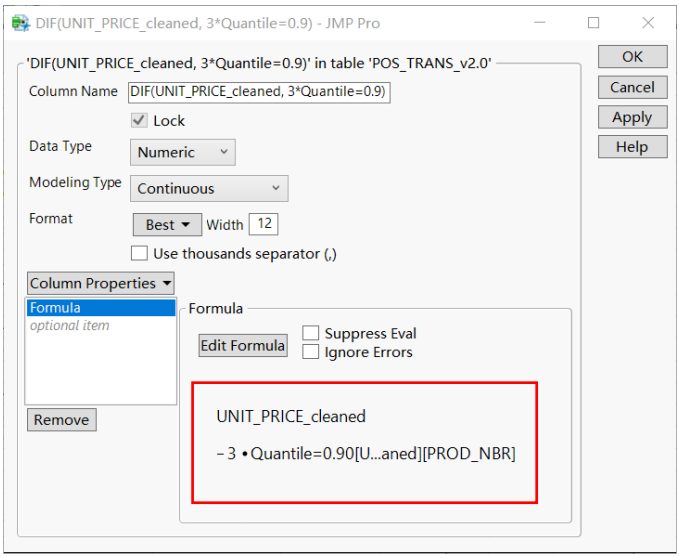


Figure 11

The distribution of the variable "DIF(UNIT_PRICE,3*Quantile=0.9)" shows that less than 0.5% the values is greater than 0. Use row selection tool to select these 1517 rows, then hide and exclude them (Figure 13).

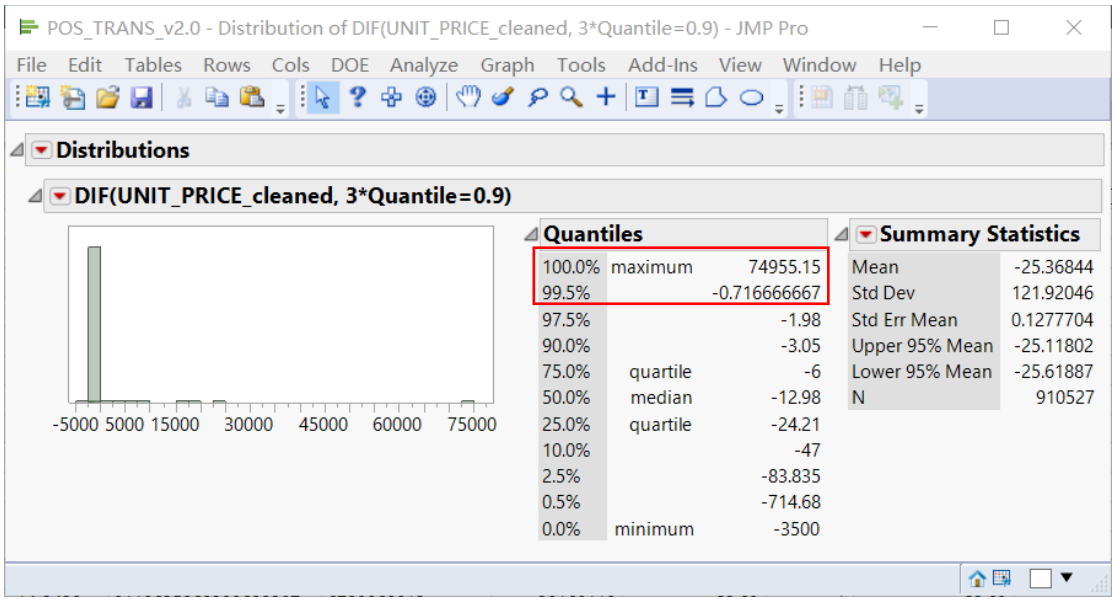


Figure 12

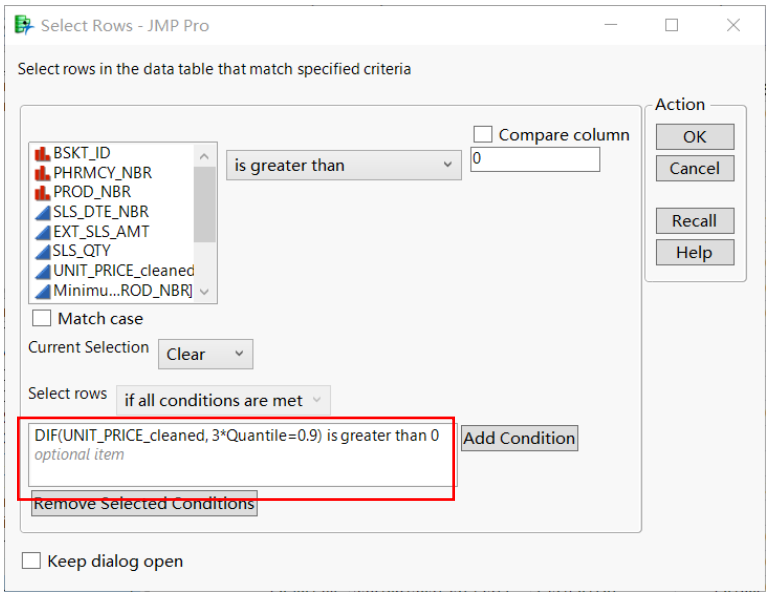


Figure 13

2.1.5. Check whether classification is right

Join table "PROD MASTER" with 4 category codes tables separately (Figure 14) and insert formula columns to check whether these categories are correctly linked (Figure 15). The results show that all the categories are correctly linked.

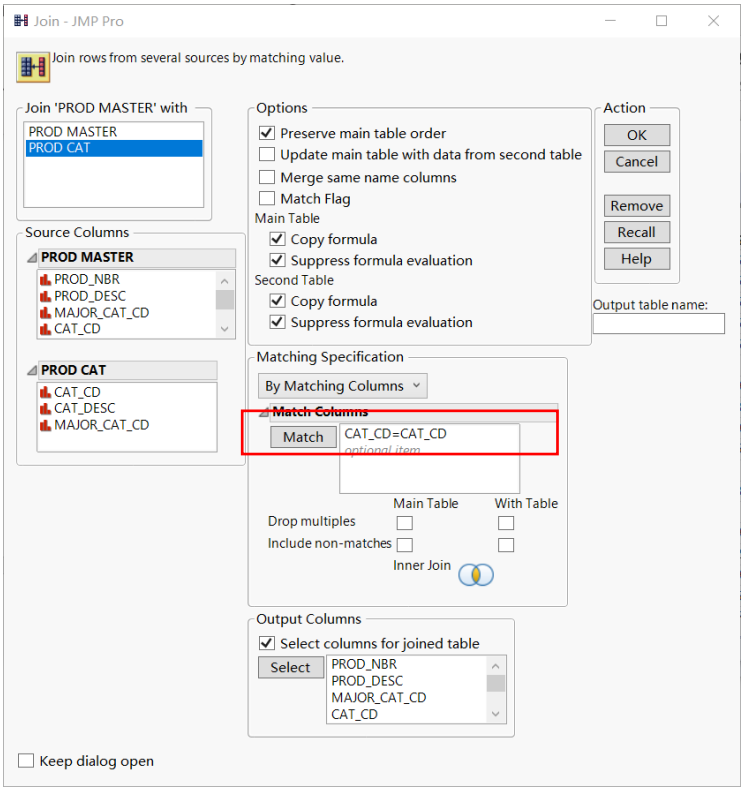


Figure 14

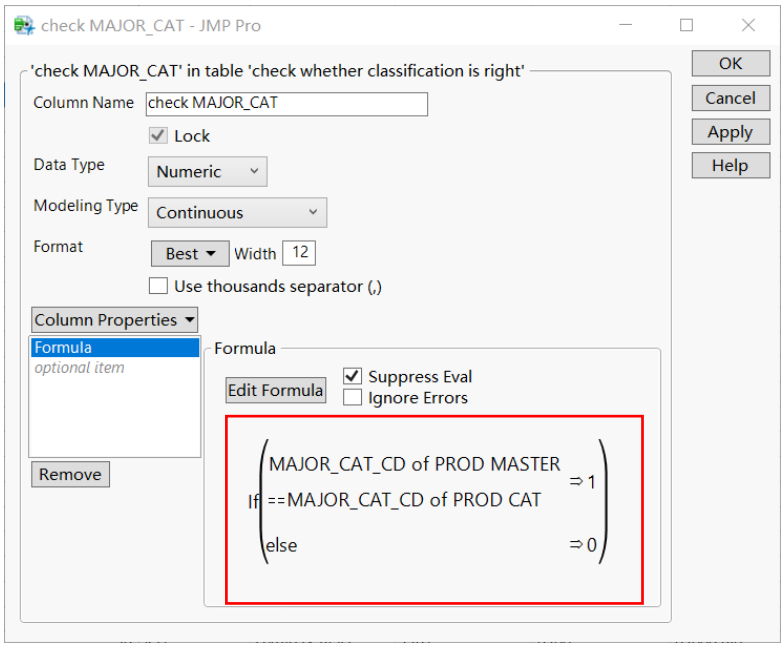
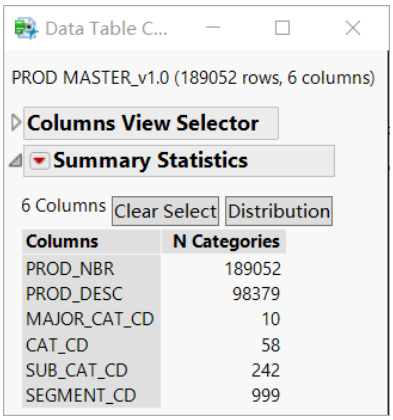


Figure 15

3. Data Analysis and Insights

3.1. Data Overview

As shown below, products can be sorted into 10 major product categories, 58 product categories, 242 sub product categories and 999 kinds of product segment. Noticed that the data is collected from multiple Pharmacy point-of-sale systems, so there could be multiple product number for the same exact product, but each product has unique product code. If we look at the number of rows in each category table, we see some categories are not included in the table "PROD MASTER".



Columns	N Categories
PROD_NBR	189052
PROD_DESC	98379
MAJOR_CAT_CD	10
CAT_CD	58
SUB_CAT_CD	242
SEGMENT_CD	999

Figure 16

Looking at the detail of pharmacies, there are 1097 pharmacies spread over 13 states.

Data Table Co...

PHRMCY MASTER_v1.0 (1097 rows, 4 columns)

Columns View Selector

Summary Statistics

4 Columns Clear Select Distribution

Columns	N Categories
PHRMCY_NBR	1097
PHRMCY_NAM	1097
ST_CD	13
ZIP_3_CD	125

Figure 17

Looking at point-of-sales' transactions data collected from 2016-01-01 to 2016-06-30, there are totally 619566 transactions of at most 43428 products concluded by 106 pharmacies.

Data Table Columns Viewer - JMP Pro

POS_TRANS_v2.0 (915744 rows, 13 columns)

Columns View Selector

Summary Statistics

7 Columns Clear Select Distribution

Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev
BSKT_ID	914202	0	619566	.	.	.	.
PHRMCY_NBR	914202	0	106	.	.	.	.
PROD_NBR	914202	0	43428	.	.	.	.
SLS_DTE_NBR	914202	0	182	.	.	.	.
EXT_SLS_AMT	914202	0	.	$-374.50	$4000.00	$8.91	27.7154106306812
SLS_QTY	914202	0	.	-36	320	1.23517122036486	1.53717274923184
UNIT_PRICE_cleaned	909010	5192	.	$0.01	$1,500.00	$7.64	18.2110786780751

Figure 18

3.2. Insights into products transacted to pharmacies

Summary table "POS_TRANS" to get minimum, maximum, mean, median and range of each product's unit price, and get the total number of transaction concluded, the total sales amount and sales quantity of each product in a new table named "POS_TRANS_Products summary". We can also join the table with "PROD_MASTER" to get products' classification and description (Figure 19).

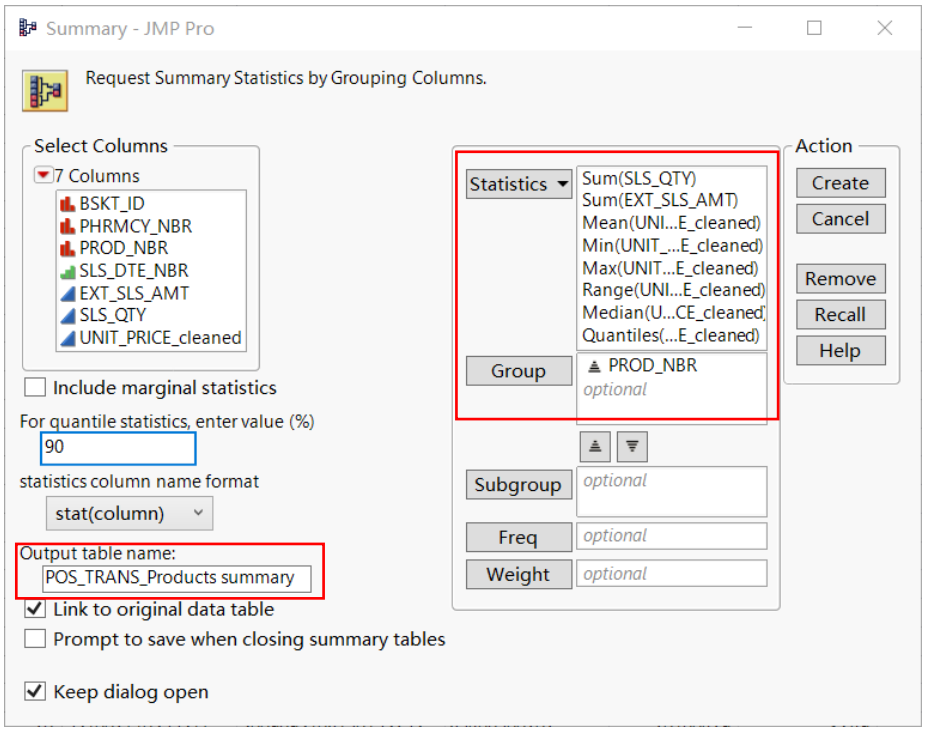
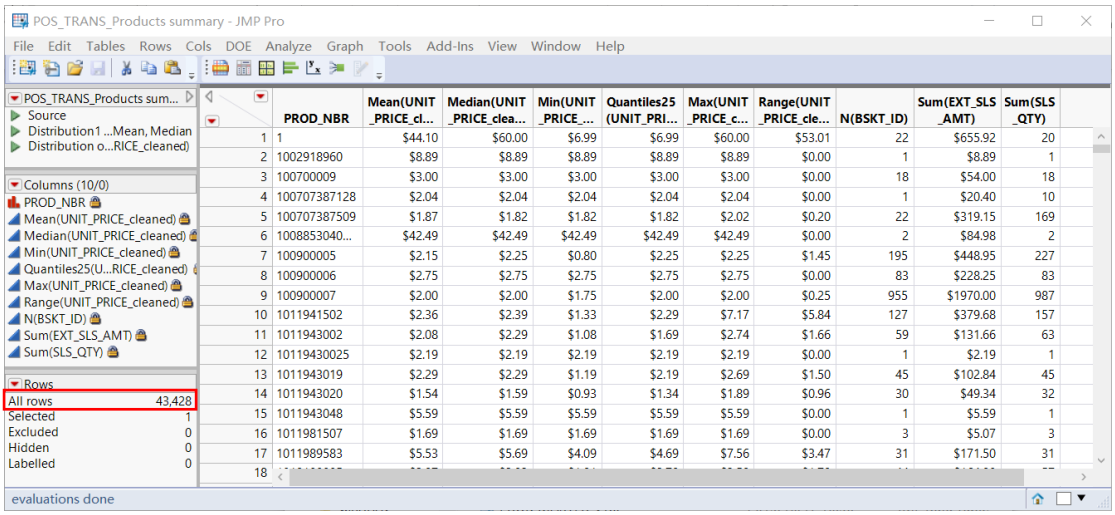


Figure 19



	PROD_NBR	Mean(UNIT _PRICE_cl...	Median(UNIT _PRICE_clea...	Min(UNIT _PRICE_...	Quantiles25 (UNIT_PRI...	Max(UNIT _PRICE_c...	Range(UNIT _PRICE_cle...	N(BSKT_ID)	Sum(EXT_SLS _AMT)	Sum(SLS _QTY)
1	1	$44.10	$60.00	$6.99	$6.99	$60.00	$53.01	22	$655.92	20
2	1002918960	$8.89	$8.89	$8.89	$8.89	$8.89	$0.00	1	$8.89	1
3	100700009	$3.00	$3.00	$3.00	$3.00	$3.00	$0.00	18	$54.00	18
4	100707387128	$2.04	$2.04	$2.04	$2.04	$2.04	$0.00	1	$20.40	10
5	100707387509	$1.87	$1.82	$1.82	$1.82	$2.02	$0.20	22	$319.15	169
6	1008853040...	$42.49	$42.49	$42.49	$42.49	$42.49	$0.00	2	$84.98	2
7	100900005	$2.15	$2.25	$0.80	$2.25	$2.25	$1.45	195	$448.95	227
8	100900006	$2.75	$2.75	$2.75	$2.75	$2.75	$0.00	83	$228.25	83
9	100900007	$2.00	$2.00	$1.75	$2.00	$2.00	$0.25	955	$1970.00	987
10	1011941502	$2.36	$2.39	$1.33	$2.29	$7.17	$5.84	127	$379.68	157
11	1011943002	$2.08	$2.29	$1.08	$1.69	$2.74	$1.66	59	$131.66	63
12	10119430025	$2.19	$2.19	$2.19	$2.19	$2.19	$0.00	1	$2.19	1
13	1011943019	$2.29	$2.29	$1.19	$2.19	$2.69	$1.50	45	$102.84	45
14	1011943020	$1.54	$1.59	$0.93	$1.34	$1.89	$0.96	30	$49.34	32
15	1011943048	$5.59	$5.59	$5.59	$5.59	$5.59	$0.00	1	$5.59	1
16	1011981507	$1.69	$1.69	$1.69	$1.69	$1.69	$0.00	3	$5.07	3
17	1011989583	$5.53	$5.69	$4.09	$4.69	$7.56	$3.47	31	$171.50	31
18										

Figure 20

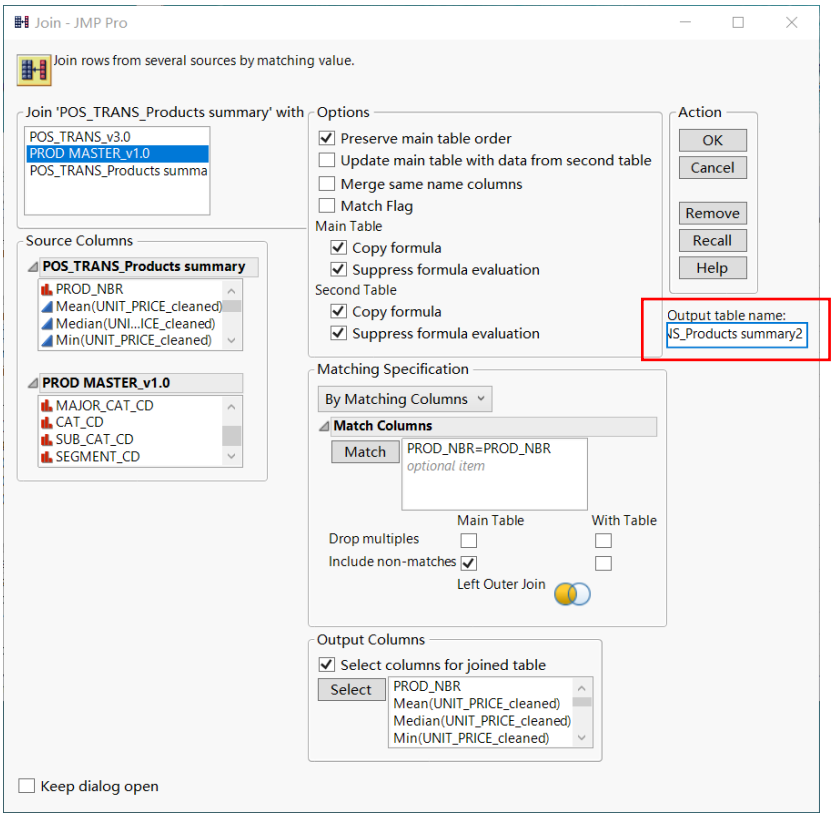
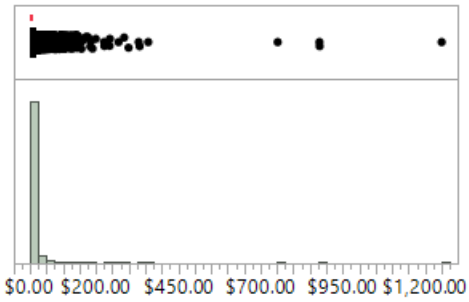


Figure 21

Look at the distributions of unit prices' mean and median (Figure 22), they similarly range from $0.02 to $1250 and 90% of products' average unit prices are valued below $16.62. Also look at minimum and maximum unit price, we can see these products' unit prices widely range from $0.01 to $1500 (Figure 23). The distribution of price range show that half of the products have unchanged unit price and 90% of them have unit price ranges smaller than $4.4. In all, we can conclude that prices of most of products are unchanged or fluctuate moderately among different transactions (Figure 24).

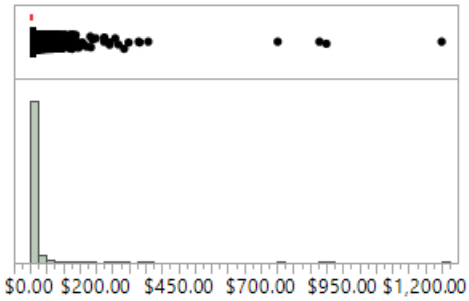
Mean(UNIT_PRICE_cleaned)



Quantiles		
100.0%	maximum	$1,250.00
99.5%		$73.06
97.5%		$39.19
90.0%		$16.62
75.0%	quartile	$7.75
50.0%	median	$4.08
25.0%	quartile	$2.97
10.0%		$1.75
2.5%		$0.99
0.5%		$0.93
0.0%	minimum	$0.02

Summary Statistics	
Mean	7.8599564
Std Dev	15.018971
Std Err Mean	0.0721316
Upper 95% Mean	8.0013357
Lower 95% Mean	7.7185772
N	43354

Median(UNIT_PRICE_cleaned)



Quantiles		
100.0%	maximum	$1,250.00
99.5%		$72.99
97.5%		$39.20
90.0%		$16.50
75.0%	quartile	$7.79
50.0%	median	$4.00
25.0%	quartile	$2.99
10.0%		$1.75
2.5%		$0.99
0.5%		$0.89
0.0%	minimum	$0.02

Summary Statistics	
Mean	7.851952
Std Dev	15.017908
Std Err Mean	0.0721265
Upper 95% Mean	7.9933213
Lower 95% Mean	7.7105827
N	43354

Figure 22

Data Table Columns Viewer - JMP Pro

POS_TRANS_Products summary (43428 rows, 10 columns)

Columns View Selector

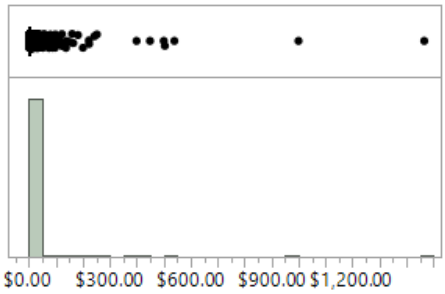
Summary Statistics

2 Columns Clear Select Distribution

Columns	N	N Missing	Min	Max	Mean	Std Dev
Min(UNIT_PRICE_cleaned)	43354	74	$0.01	$1,000.00	$7.12	13.687340177
Max(UNIT_PRICE_cleaned)	43354	74	$0.02	$1,500.00	$8.79	19.343596938

Figure 23

Range(UNIT_PRICE_cleaned)



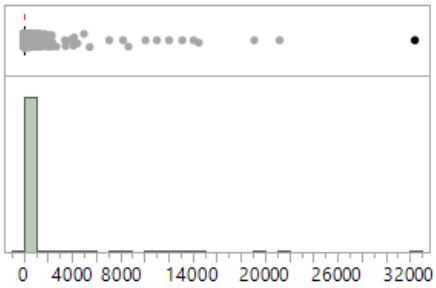
Quantiles		
100.0%	maximum	$1,466.50
99.5%		$29.75
97.5%		$11.79
90.0%		$4.39
75.0%	quartile	$1.05
50.0%	median	$0.00
25.0%	quartile	$0.00
10.0%		$0.00
2.5%		$0.00
0.5%		$0.00
0.0%	minimum	$0.00

Summary Statistics	
Mean	1.6698573
Std Dev	11.571002
Std Err Mean	0.055572
Upper 95% Mean	1.7787796
Lower 95% Mean	1.5609351
N	43354

Figure 24

The distribution of sales quantity shows that half of the products' sales volumes are under 3 pieces for the first half year. We can set product with a sales quantity more than 37, the 90th percentile, as high demand product and insert a new column to record these high demand products (Figure 25). We can find the best-selling product is of home health care category and the product with largest sales amount is of general merchandise category.

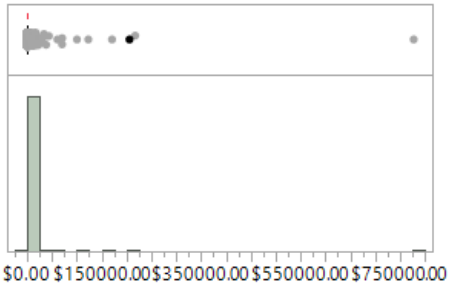
Sum(SLS_QTY)



Quantiles		
100.0%	maximum	32412
99.5%		558.855
97.5%		164
90.0%		37
75.0%	quartile	11
50.0%	median	3
25.0%	quartile	1
10.0%		1
2.5%		1
0.5%		1
0.0%	minimum	-3

Summary Statistics	
Mean	26.001566
Std Dev	283.66326
Std Err Mean	1.3611882
Upper 95% Mean	28.66952
Lower 95% Mean	23.333612
N	43428

Sum(EXT_SLS_AMT)



Quantiles		
100.0%	maximum	$777415.11
99.5%		$3505.65
97.5%		$1075.63
90.0%		$266.42
75.0%	quartile	$63.87
50.0%	median	$17.34
25.0%	quartile	$5.59
10.0%		$2.99
2.5%		$1.50
0.5%		$0.75
0.0%	minimum	$-76.99

Summary Statistics	
Mean	187.65422
Std Dev	4259.5799
Std Err Mean	20.440045
Upper 95% Mean	227.71709
Lower 95% Mean	147.59135
N	43428

Figure 25

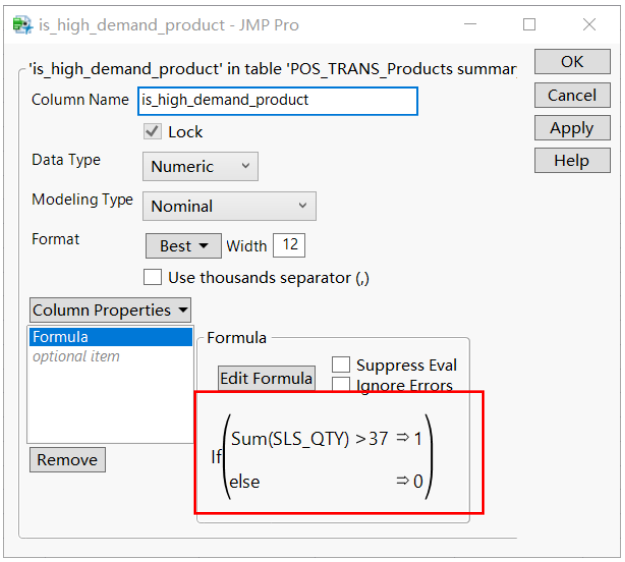


Figure 26

The total sales amount of all products is $8149447 and 80% of total sales amount would be $6519558. If we sort variable "sum of sales amount" as descending order and insert a new column to get cumulative sum, we can find that the total sales amount of 3494 biggest-selling pharmacies, which account for about 8% of all the products, occupies 80% of the company's total sales amount.

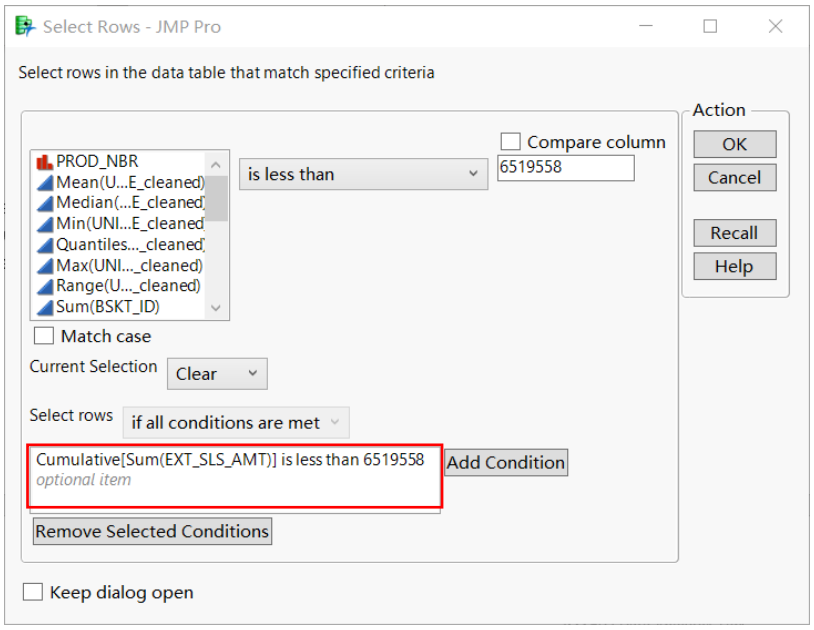


Figure 27

3.3. Insights into pharmacies traded with point-of-sales

1. Using table summary tool to Summarize some statistics from the table "POS_TRANS" and name the new table named "POS_TRANS Pharmacy Summary".
2. Join some useful variables from the table "POS_TRANS_Products summary3".
3. Join the variables from the table "PROD MASTER".
4. Insert a new column to compute the average unit price of each product each pharmacy bought.
5. Based on the table we created, summarize a new table without grouping by products and join the other variables we need.
6. Insert a column to calculate the proportion of cumulative sum sales amount.

   The result table is shown in Figure 33.

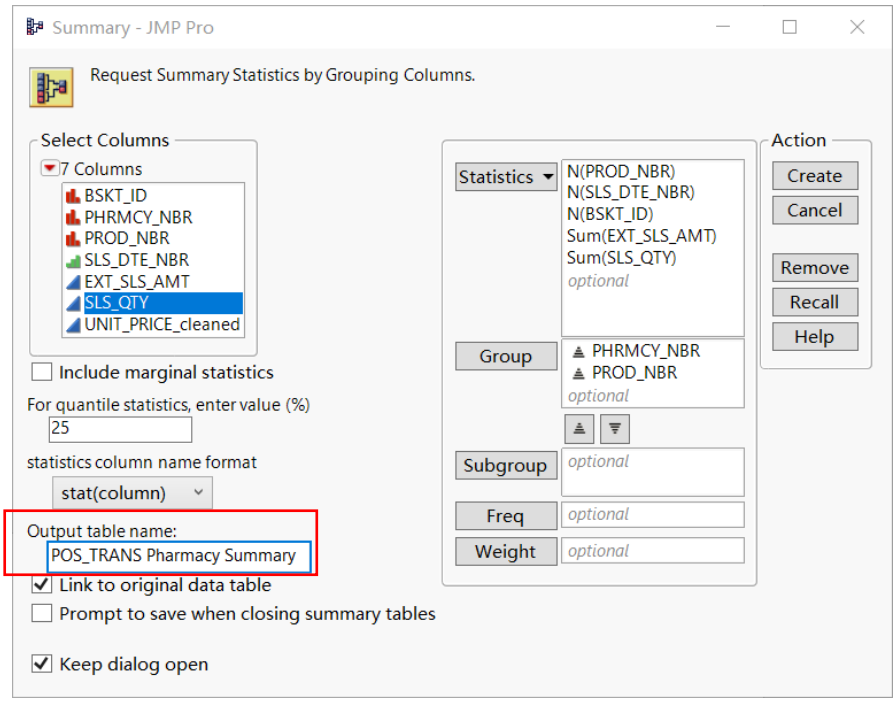


Figure 28: STEP 1

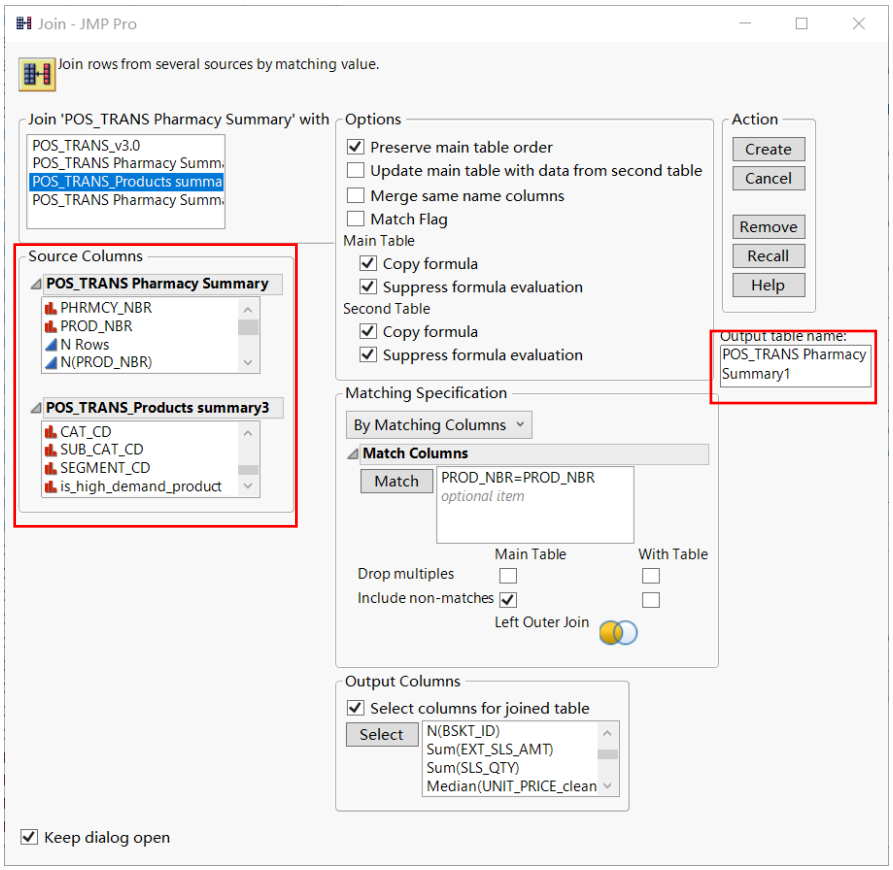


Figure 29: STEP 2

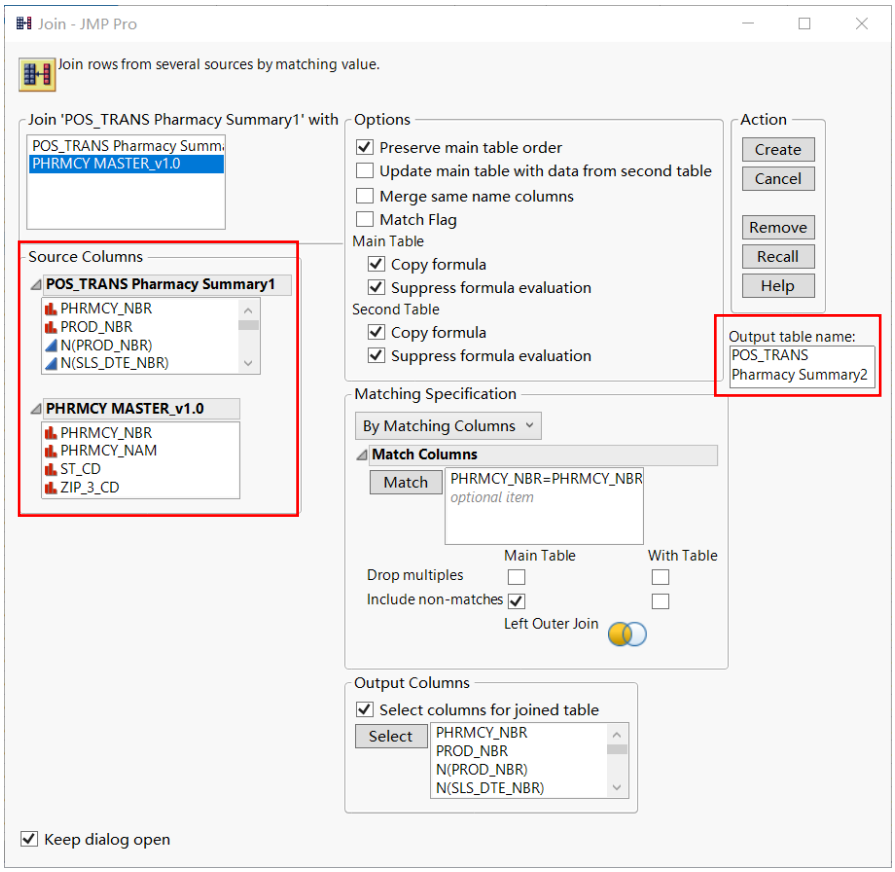


Figure 30: STEP 3

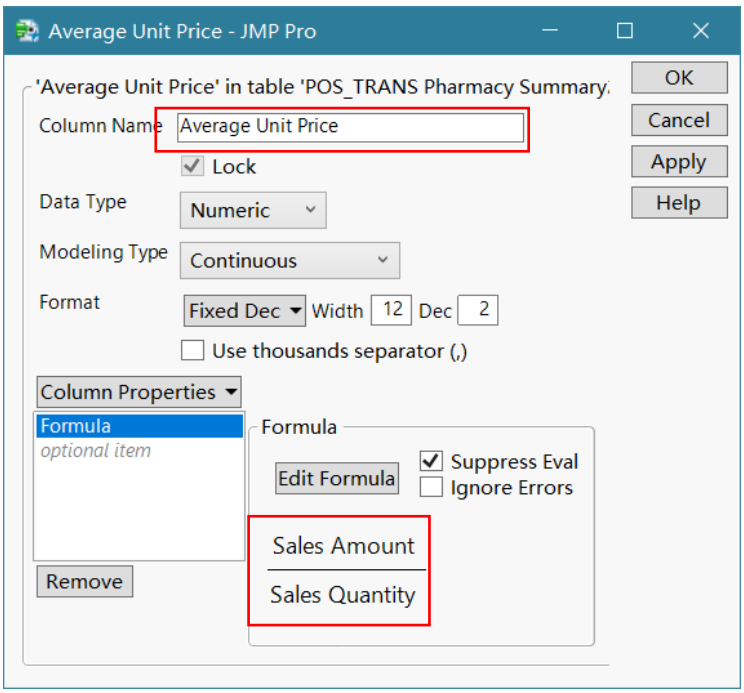


Figure 31: STEP 4

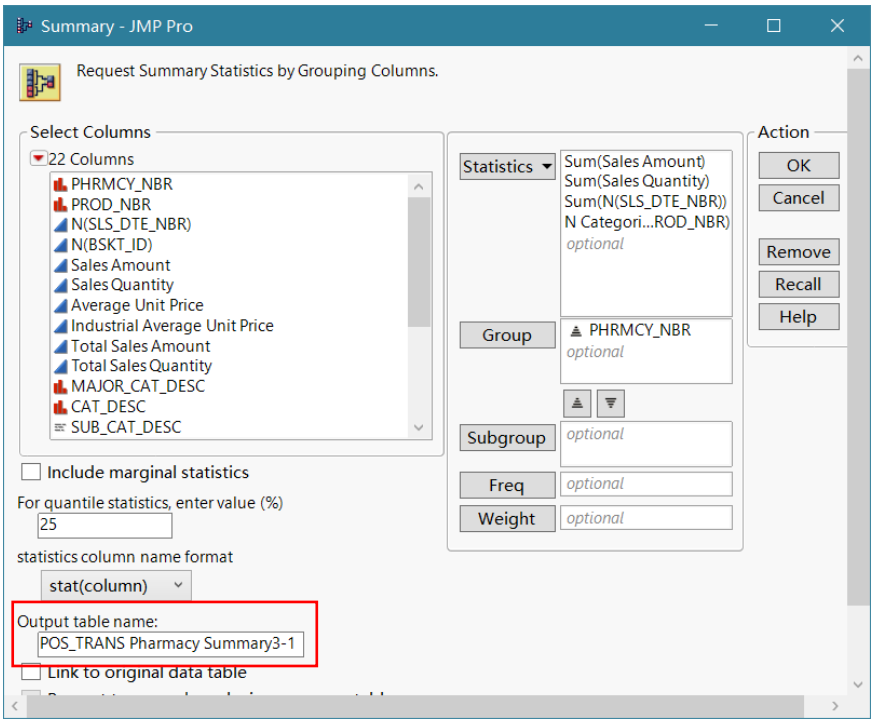
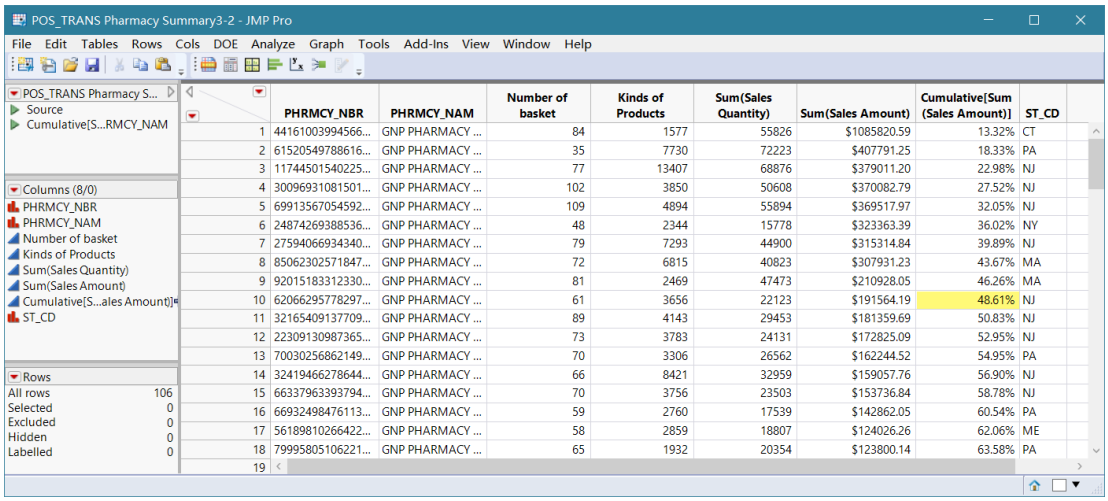


Figure 32: STEP 5



	PHRMCY_NBR	PHRMCY_NAM	Number of basket	Kinds of Products	Sum(Sales Quantity)	Sum(Sales Amount)	Cumulative[Sum (Sales Amount)]	ST_CD
1	44161003994566...	GNP PHARMACY ...	84	1577	55826	$1085820.59	13.32%	CT
2	61520549788616...	GNP PHARMACY ...	35	7730	72223	$407791.25	18.33%	PA
3	11744501540225...	GNP PHARMACY ...	77	13407	68876	$379011.20	22.98%	NJ
4	30096931081501...	GNP PHARMACY ...	102	3850	50608	$370082.79	27.52%	NJ
5	69913567054592...	GNP PHARMACY ...	109	4894	55894	$369517.97	32.05%	NJ
6	24874269388536...	GNP PHARMACY ...	48	2344	15778	$323363.39	36.02%	NY
7	27594066934340...	GNP PHARMACY ...	79	7293	44900	$315314.84	39.89%	NJ
8	85062302571847...	GNP PHARMACY ...	72	6815	40823	$307931.23	43.67%	MA
9	92015183312330...	GNP PHARMACY ...	81	2469	47473	$210928.05	46.26%	MA
10	62066295778297...	GNP PHARMACY ...	61	3656	22123	$191564.19	48.61%	NJ
11	32165409137709...	GNP PHARMACY ...	89	4143	29453	$181359.69	50.83%	NJ
12	22309130987365...	GNP PHARMACY ...	73	3783	24131	$172825.09	52.95%	NJ
13	70030256862149...	GNP PHARMACY ...	70	3306	26562	$162244.52	54.95%	PA
14	32419466278644...	GNP PHARMACY ...	66	8421	32959	$159057.76	56.90%	NJ
15	66337963393794...	GNP PHARMACY ...	70	3756	23503	$153736.84	58.78%	NJ
16	66932498476113...	GNP PHARMACY ...	59	2760	17539	$142862.05	60.54%	PA
17	56189810266422...	GNP PHARMACY ...	58	2859	18807	$124026.26	62.06%	ME
18	79995805106221...	GNP PHARMACY ...	65	1932	20354	$123800.14	63.58%	PA
19								

Figure 33: Result Table

Draw the plot of cumulative sales amount by Pharmacies, we see About 31.1% of the pharmacies account for 80% of the sales amount, while the top 10 biggest-selling pharmacies take up 48.6% of the sales amount.

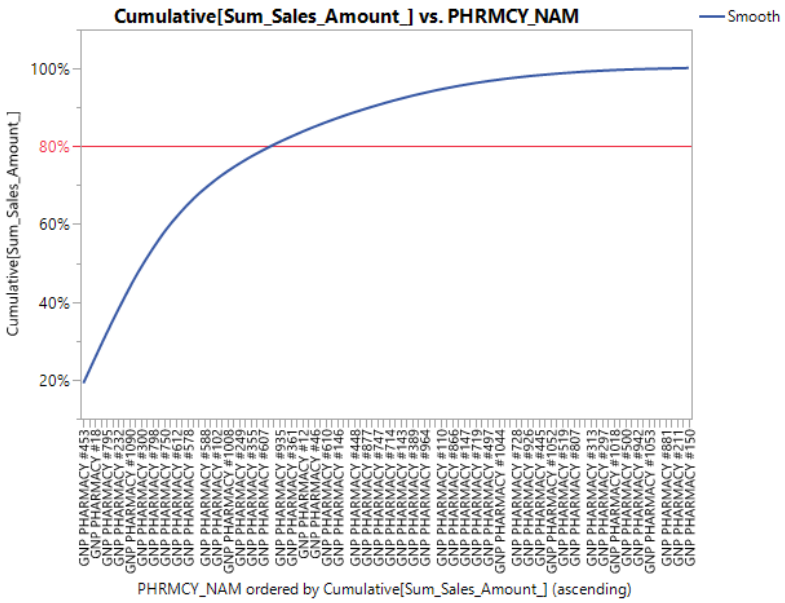


Figure 34

	PHRMCY_NBR	PHRMCY_NAM	Number of basket	Kinds of Products	Sum(Sales Quantity)	Sum(Sales Amount)	Cumulative[Sum (Sales Amount)]
1	44161003994566...	GNP PHARMACY ...	84	1577	55826	$1085820.59	13.32%
2	61520549788616...	GNP PHARMACY ...	35	7730	72223	$407791.25	18.33%
3	11744501540225...	GNP PHARMACY ...	77	13407	68876	$379011.20	22.98%
4	30096931081501...	GNP PHARMACY ...	102	3850	50608	$370082.79	27.52%
5	69913567054592...	GNP PHARMACY ...	109	4894	55894	$369517.97	32.05%
6	24874269388536...	GNP PHARMACY ...	48	2344	15778	$323363.39	36.02%
7	27594066934340...	GNP PHARMACY ...	79	7293	44900	$315314.84	39.89%
8	85062302571847...	GNP PHARMACY ...	72	6815	40823	$307931.23	43.67%
9	92015183312330...	GNP PHARMACY ...	81	2469	47473	$210928.05	46.26%
10	62066295778297...	GNP PHARMACY ...	61	3656	22123	$191564.19	**48.61%**
11	32165409137709...	GNP PHARMACY ...	89	4143	29453	$181359.69	50.83%
12	22309130987365...	GNP PHARMACY ...	73	3783	24131	$172825.09	52.95%
13	70030256862149...	GNP PHARMACY ...	70	3306	26562	$162244.52	54.95%
14	32419466278644...	GNP PHARMACY ...	66	8421	32959	$159057.76	56.90%
15	66337963393794...	GNP PHARMACY ...	70	3756	23503	$153736.84	58.78%
16	66932498476113...	GNP PHARMACY ...	59	2760	17539	$142862.05	60.54%
17	56189810266422...	GNP PHARMACY ...	58	2859	18807	$124026.26	62.06%
18	79995805106221...	GNP PHARMACY ...	65	1932	20354	$123800.14	63.58%
19	53671923223431...	GNP PHARMACY ...	55	3034	18974	$123583.05	65.09%
20	33893481912423...	GNP PHARMACY ...	48	3149	14587	$115640.41	66.51%
21	64436676459374...	GNP PHARMACY ...	63	2606	19193	$110104.04	67.86%
22	54640727345449...	GNP PHARMACY ...	38	429	17843	$97937.40	69.07%
23	57173454524580...	GNP PHARMACY ...	56	2499	13319	$91239.15	70.19%
24	17882708948460...	GNP PHARMACY ...	55	2842	15316	$88918.34	71.28%
25	10860603488723...	GNP PHARMACY ...	43	2363	10460	$79745.30	72.26%
26	86060048169670...	GNP PHARMACY ...	57	1312	13949	$79231.95	73.23%
27	40348666392500...	GNP PHARMACY ...	41	2122	10572	$78530.59	74.19%
28	28685027526422...	GNP PHARMACY ...	40	2067	9463	$75311.00	75.12%
29	28418062240230...	GNP PHARMACY ...	41	2722	10825	$72283.12	76.00%
30	36954326800443...	GNP PHARMACY ...	37	3536	12657	$71369.61	76.88%
31	84066381426033...	GNP PHARMACY ...	41	2301	8463	$69483.29	77.73%
32	55610815057437...	GNP PHARMACY ...	39	2289	10217	$67856.44	78.56%
33	32693963849426...	GNP PHARMACY ...	39	1828	8846	$64878.66	79.36%
34	65704604850432...	GNP PHARMACY ...	43	2155	10675	$64268.79	**80.15%**
35	81106358683896...	GNP PHARMACY ...	56	1553	11892	$62808.00	80.92%

Figure 35

3.4. Customer Segmentation

Before doing market segmentation, we need to get grouping variables first, which help us to create groups of customers who share similar features, such as similar needs or purchasing behaviors.

3.4.1. Get the variables we want to explore

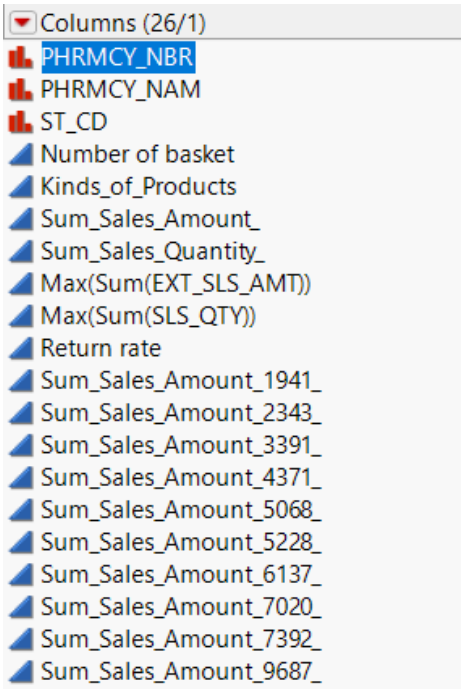


Figure 36

1. **Pharmacy number, Pharmacy name, State code:** Use table-join tool.
2. **Number of baskets, Kinds of products, Sum sales amount, Sum sales quantity:** Use table-summary tool to get these statistics from the table "POS_TRANS_Products summary3", which had been explained from the above analysis.
3. **Max sales amount, Max sales quantity:** First, using table-summary tool to get the sum of sales amount and of sales quantity from the table "POS_TRANS_v3.0" and grouped by pharmacy and basket ID. Then, use table-summary tool again to get the max value grouped by pharmacies.
4. **Return rate:** Select pharmacies with a negative sales quantity and sales amount into a new table, use table-summary tool to get the count of these return basket and grouped by pharmacies. Use formula to divide it by the total number of baskets.
5. **Sum sales amount of each Major Products:** Use table-summary tool to get the sum of sales amount, grouped by pharmacy name and subgrouped by Major product code.
6. When we get all these variables, we can just join them together in a new table named "Pharmacy_Segmentation_v1.0".

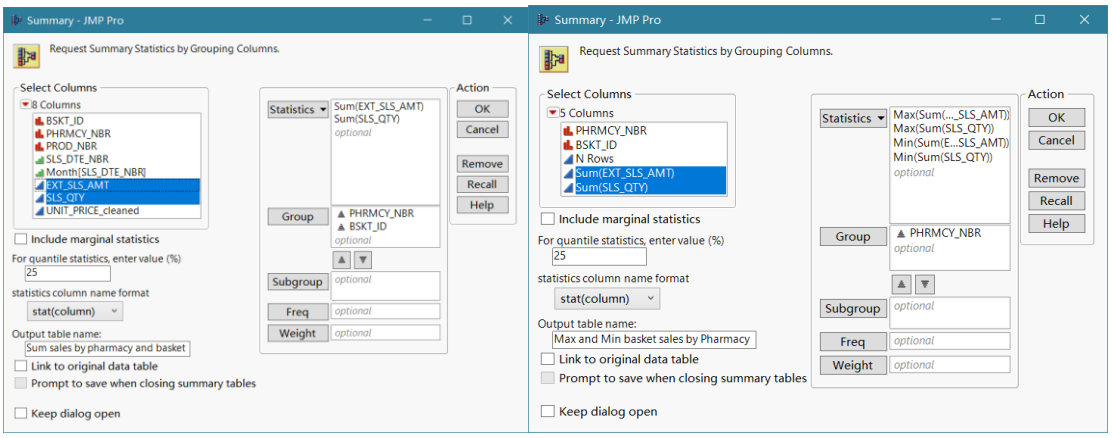
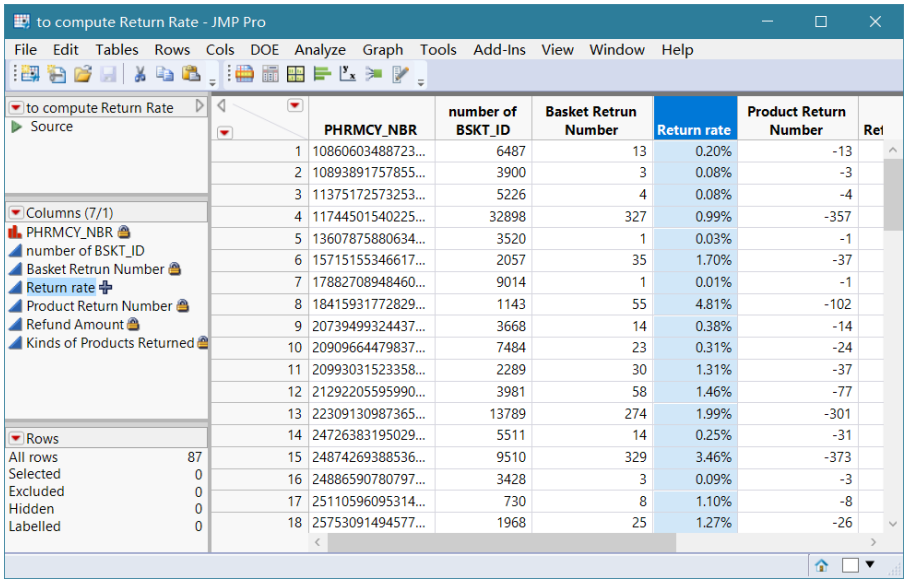


Figure 37: STEP 3



	PHRMCY_NBR	number of BSKT_ID	Basket Retrun Number	Return rate	Product Return Number	Ret
1	10860603488723...	6487	13	0.20%	-13	
2	10893891757855...	3900	3	0.08%	-3	
3	11375172573253...	5226	4	0.08%	-4	
4	11744501540225...	32898	327	0.99%	-357	
5	13607875880634...	3520	1	0.03%	-1	
6	15715155346617...	2057	35	1.70%	-37	
7	17882708948460...	9014	1	0.01%	-1	
8	18415931772829...	1143	55	4.81%	-102	
9	20739499324437...	3668	14	0.38%	-14	
10	20909664479837...	7484	23	0.31%	-24	
11	20993031523358...	2289	30	1.31%	-37	
12	21292205595990...	3981	58	1.46%	-77	
13	22309130987365...	13789	274	1.99%	-301	
14	24726383195029...	5511	14	0.25%	-31	
15	24874269388536...	9510	329	3.46%	-373	
16	24886590780797...	3428	3	0.09%	-3	
17	25110596095314...	730	8	1.10%	-8	
18	25753091494577...	1968	25	1.27%	-26	

Figure 38: STEP 4

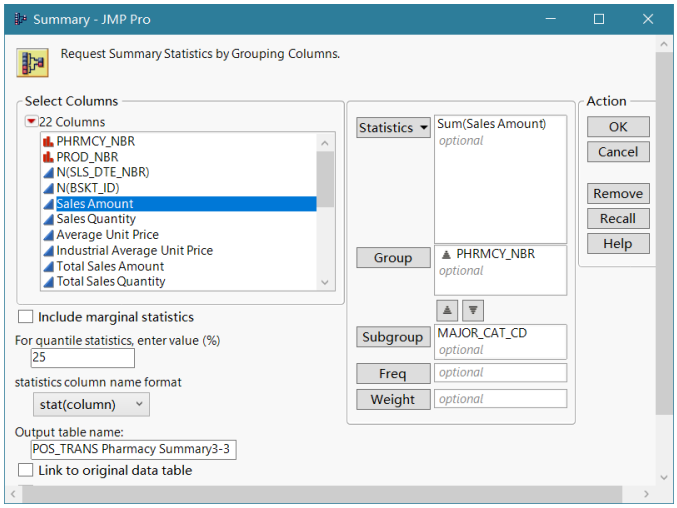
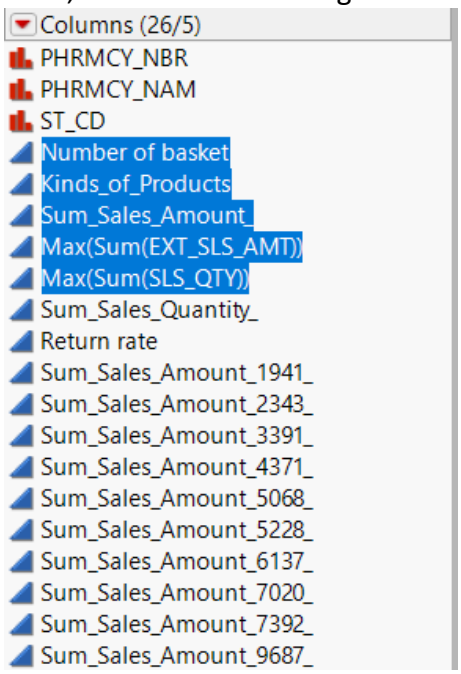


Figure 39: STEP 5

3.4.2. Standardize the grouping variables

Here, we use the following variables to cluster pharmacies.



These variables are used to differentiate pharmacies into various economic scale, market segmentation and purchase behavior.

Because the values between variables have large difference, we should standardize these variables to eliminate the effect of the unit of measurement:

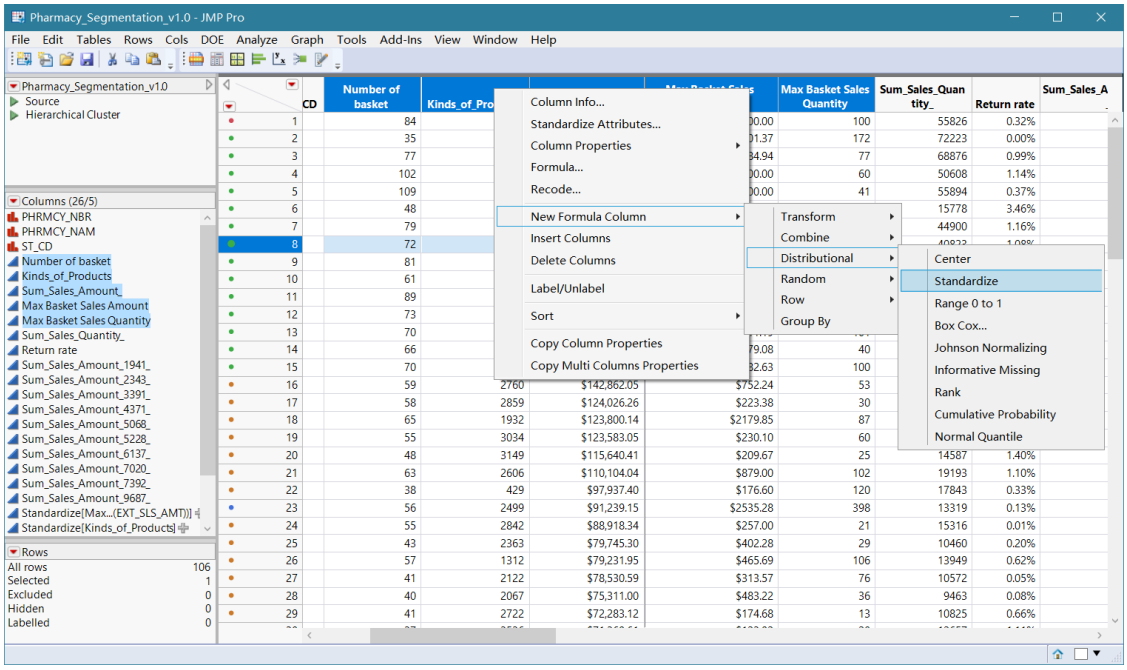


Figure 40

3.4.3. Check the correlation matrix

Before we move on into clustering, we need to measure the dissimilarity between the variables by checking correlation by using multivariate analysis.

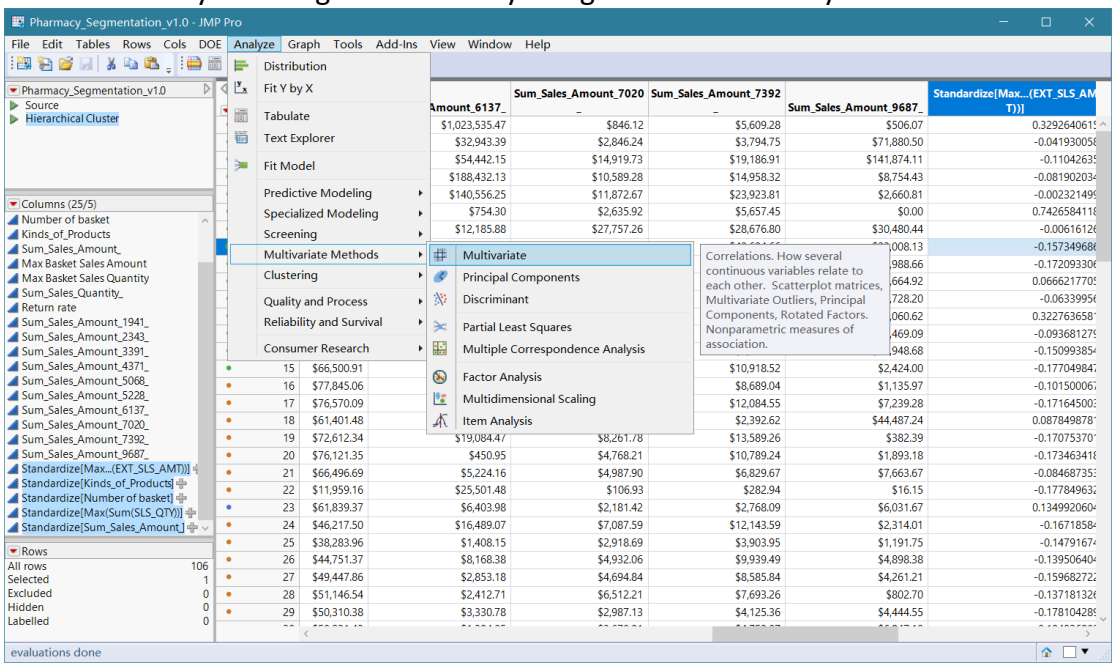


Figure 41

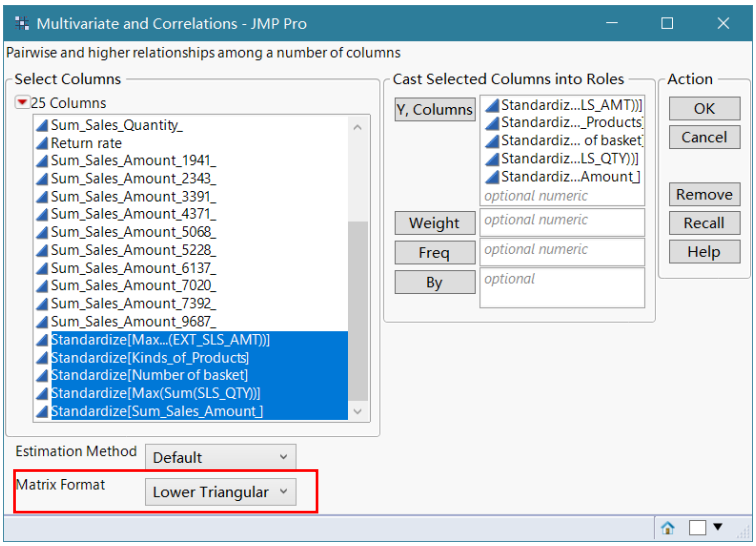


Figure 42

We could use color map on correlation to quickly identify which pair has strong correlation to each other. The map shows that the number of baskets is correlated to the sum of sales amount, which is understandable. Here we can reduce the variable “Sum sales amount”, which can be represent by the variable “the number of baskets” and “Max sales amount”.

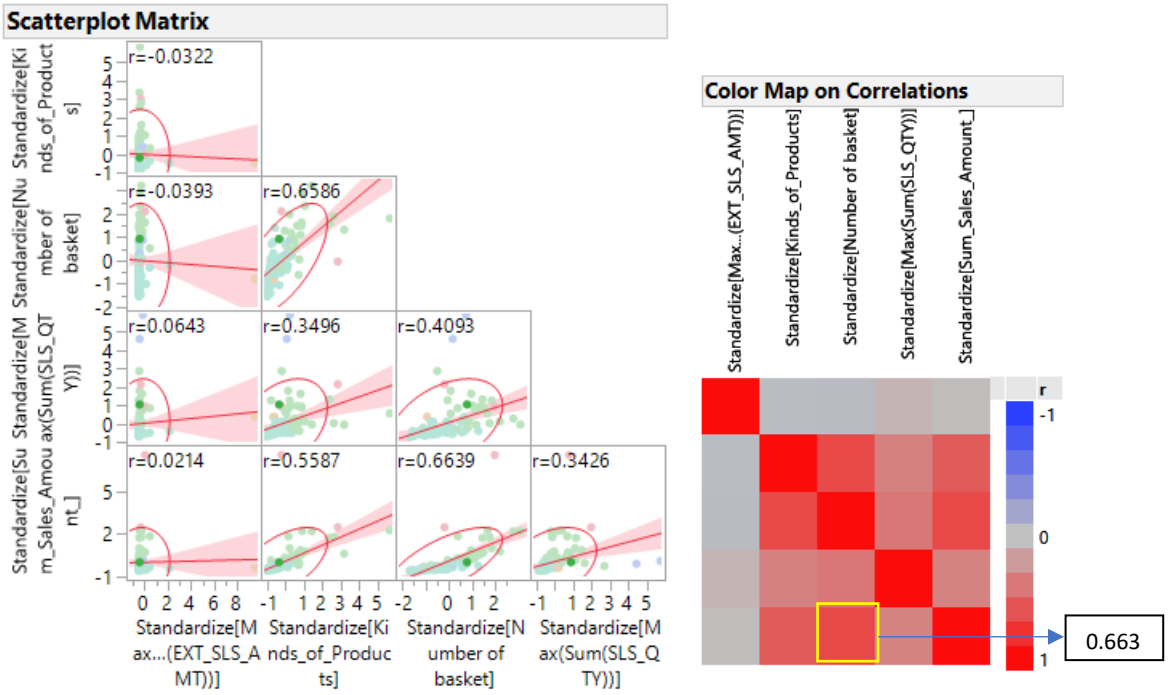


Figure 43

The Scatterplot Matrix of the remain variables is showing in Figure 44.

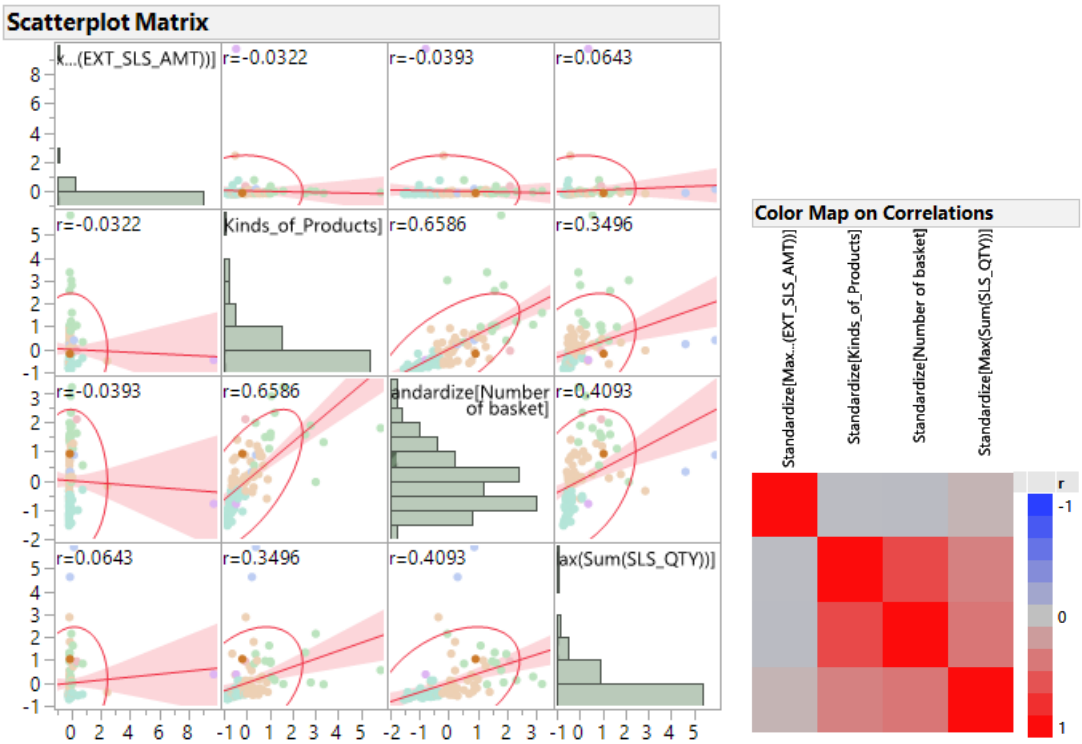


Figure 44

3.4.4. Hierarchical Cluster

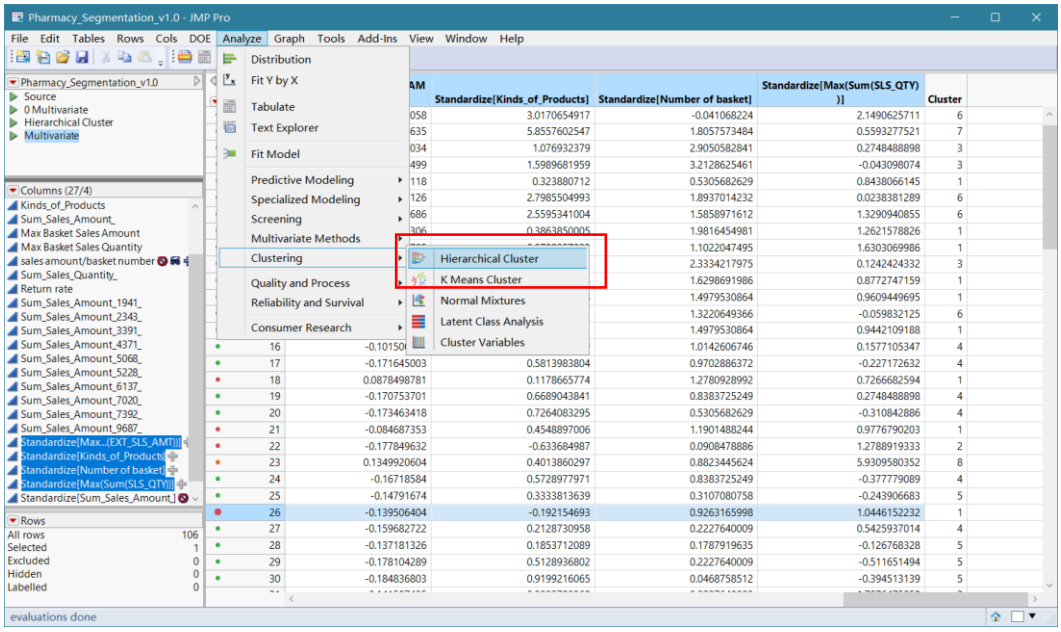
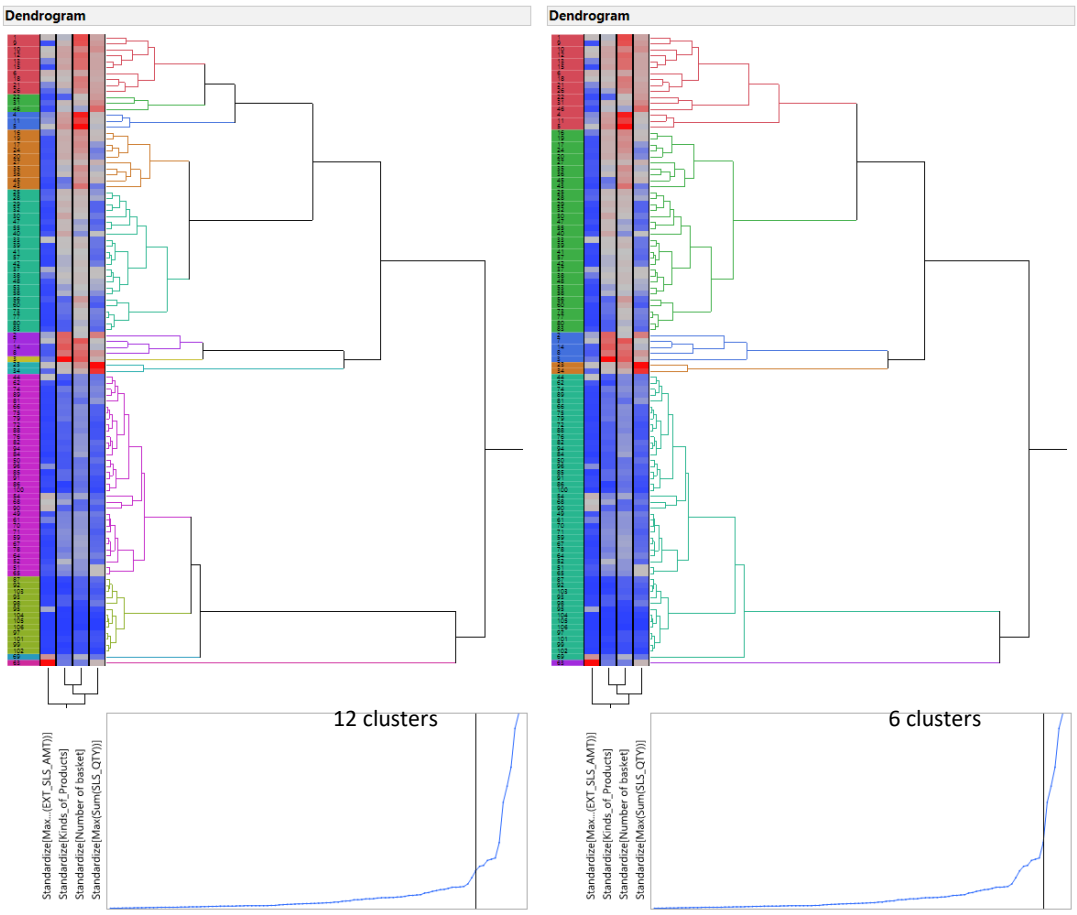


Figure 45

For now, we can do the hierarchical clustering for these 4 variables.



By default, the pharmacies are clustered into 12 groups. By looking at the scree plot, 6 would be a better cluster number, because there is a larger surge between cluster 6 and 5, and so much clusters would be useless and messy. Here, move the vertical line to produce 6 clusters.

Cluster Means

Cluster	Count	Standardize[Max...(EXT_SLS_AMT))]	Standardize[Kinds_of_Products]	Standardize[Number of basket]	Standardize[Max(Sum(SLS_QTY))]
1	16	0.0150	0.5259	1.3853	1.0321
2	34	-0.1613	0.0654	0.2823	-0.2306
3	5	-0.0934	3.5187	1.3133	0.8003
4	2	1.817e-5	0.3154	0.5965	5.2783
5	48	-0.0840	-0.5917	-0.8069	-0.4918
6	1	9.7463	-0.4612	-0.7886	0.3753

Cluster Standard Deviations

Cluster	Count	Standardize[Max...(EXT_SLS_AMT))]	Standardize[Kinds_of_Products]	Standardize[Number of basket]	Standardize[Max(Sum(SLS_QTY))]
1	16	0.253298	0.608618	0.987427	0.690922
2	34	0.043038	0.451455	0.439705	0.285410
3	5	0.067000	1.339447	0.788682	0.935357
4	2	0.190882	0.121631	0.404208	0.922955
5	48	0.403725	0.197480	0.378135	0.164745
6	1	.	.	.	.

Figure 46

The tables of clusters means and of standard deviations above show that these clusters are significantly different from each other. Cluster 6 only consists of 1 single pharmacy, which could be an outlier.

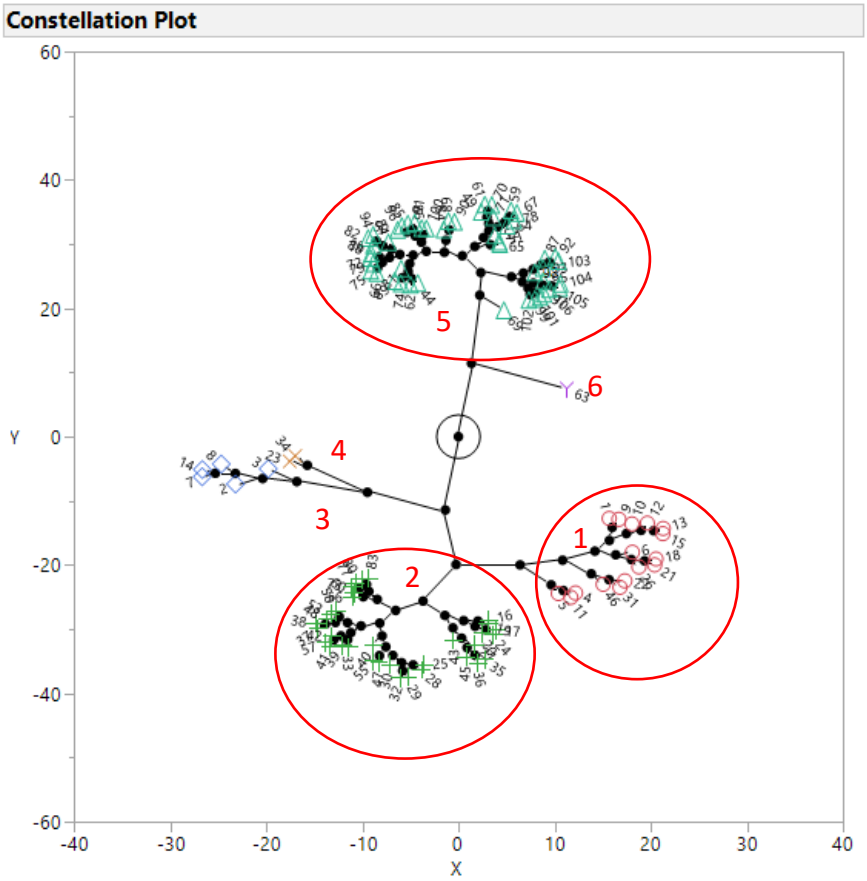


Figure 47

We can also find the relationship among clusters from the constellation plot, which display a great distance between cluster 5 and cluster 1 or 2, while cluster 1 and 2 are relatively close to each other in distance. We also see cluster 3 and 4 are close in distance.

After coloring and marking cluster into the table, we can save clusters to the table.

One way to identify clusters is to look at the parallel plot of the variable means within each cluster as shown below. Parallel Plot shows the profile of the clusters across variables and the number of records of each cluster. We can identify different patterns of standard deviations for each cluster. Scatterplot matrix shows separation of the clusters across these variables.

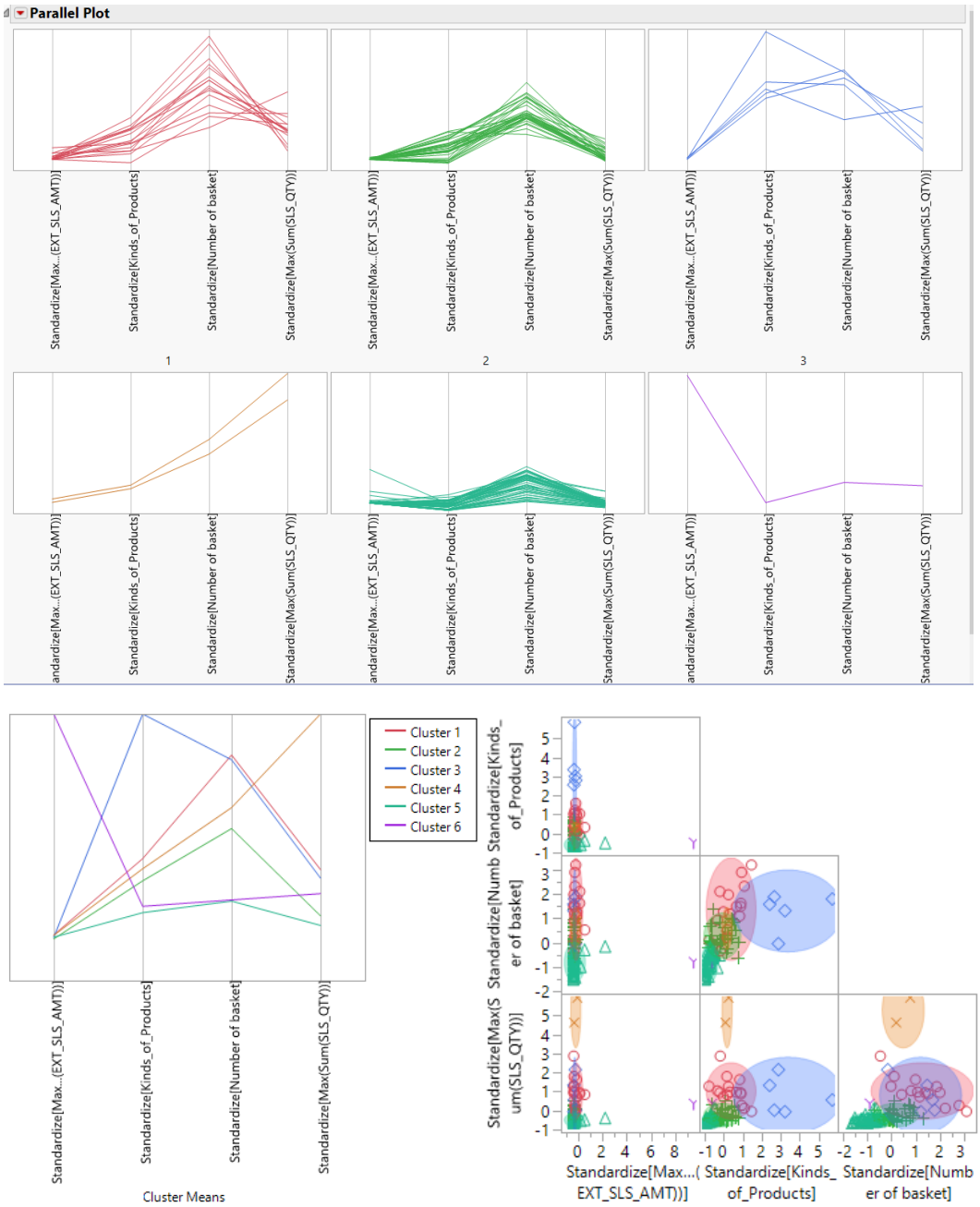


Figure 48

Another way to interpret the clusters is to look at summary statistics. We can create a table of means of grouping variables for each cluster:

1. Use Table-summary tool to get the means of the following variables in Figure 48. In addition, get the category number of State and the sum of total sales amount. Choose "Cluster" as group-by variable.

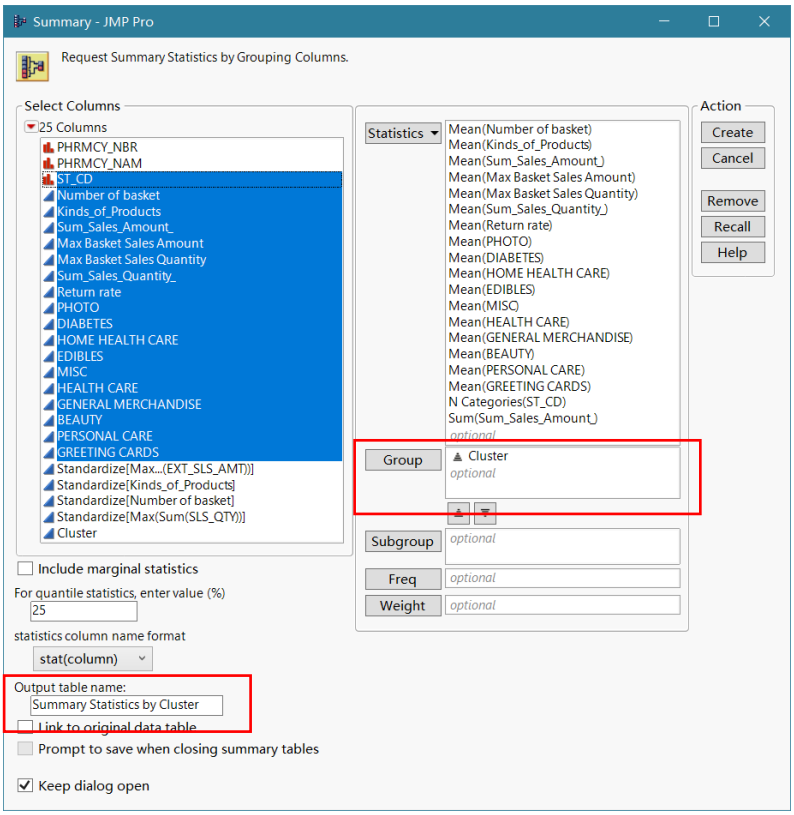
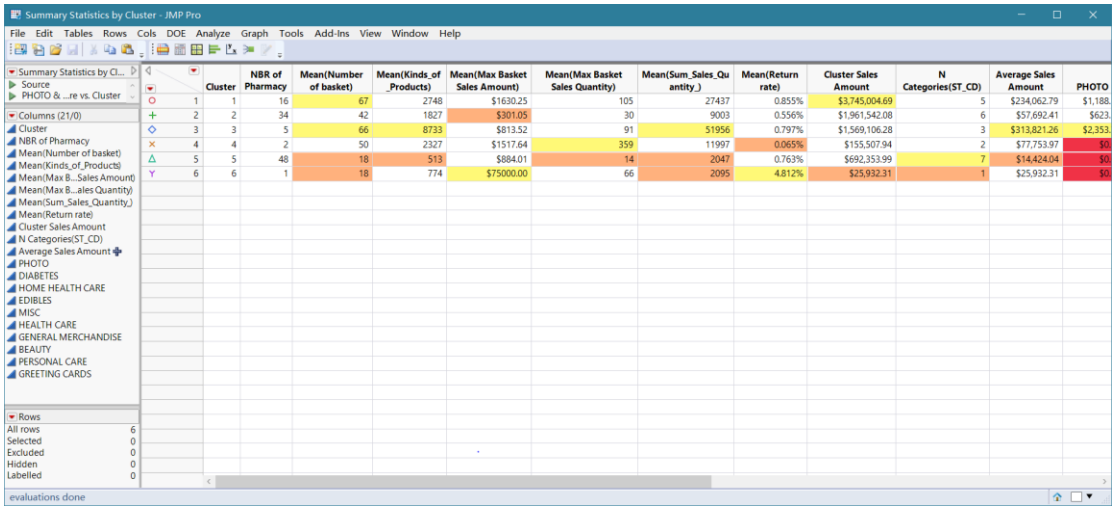


Figure 49

2. Rename the columns representing sales amount of each major products by the description of major products.
3. Fix the decimal places of variables “Mean (Total_Number_of_Basket)” and “Mean (Kinds_of_Products)” as 0.
4. Save the table as “Summary Statistics by Cluster”.

The result is showing in Figure 49.



	Cluster	NBR of Pharmacy	Mean(Number of basket)	Mean(Kinds_of_Products)	Mean(Max Basket Sales Amount)	Mean(Max Basket Sales Quantity)	Mean(Sum_Sales_Quantity_)	Mean(Return rate)	Cluster Sales Amount	N Categories(ST_CD)	Average Sales Amount	PHOTO
1	1	16	67	2748	$1630.25	105	27437	0.855%	$3,745,004.69	5	$234,062.79	$1,188.
2	2	34	42	1827	$301.05	30	9003	0.556%	$1,961,542.08	6	$57,692.41	$623.
3	3	5	66	8733	$813.52	91	51956	0.797%	$1,569,106.28	3	$313,821.26	$2,353.
4	4	2	50	2327	$1517.64	359	11997	0.065%	$155,507.94	2	$77,753.97	$0.
5	5	48	18	513	$884.01	14	2047	0.763%	$692,353.99	7	$14,424.04	$0.
6	6	1	18	774	$75000.00	66	2095	4.812%	$25,932.31	1	$25,932.31	$0.

Figure 50

We can readily interpret these clusters through the table and graph below:

(**Yellow:** mark the cells value highest in the column; **Orange:** mark the cells value lowest in the column; **Red:** mark the cells value zero)

Cluster	NBR of Pharmacy	pharmacy proportion	Mean(Number of basket)	Mean(Kinds_of_Products)	Mean(Max Basket Sales Amount)	Mean(Max Basket Sales Quantity)	Mean(Sum_Sales_Quantity_)	Mean(Return rate)	Cluster Sales Amount	Sum[Cluster Sales Amount]	Average Sales Amount
1	16	15.09%	67	2748	$1630.25	105	27437	0.855%	$3,745,004.69	45.95%	$234,062.79
2	34	32.08%	42	1827	$301.05	30	9003	0.556%	$1,961,542.08	24.07%	$57,692.41
3	5	4.72%	66	8733	$813.52	91	51956	0.797%	$1,569,106.28	19.25%	$313,821.26
4	2	1.89%	50	2327	$1517.64	359	11997	0.065%	$155,507.94	1.91%	$77,753.97
5	48	45.28%	18	513	$884.01	14	2047	0.763%	$692,353.99	8.50%	$14,424.04
6	1	0.94%	18	774	$75000.00	66	2095	4.812%	$25,932.31	0.32%	$25,932.31

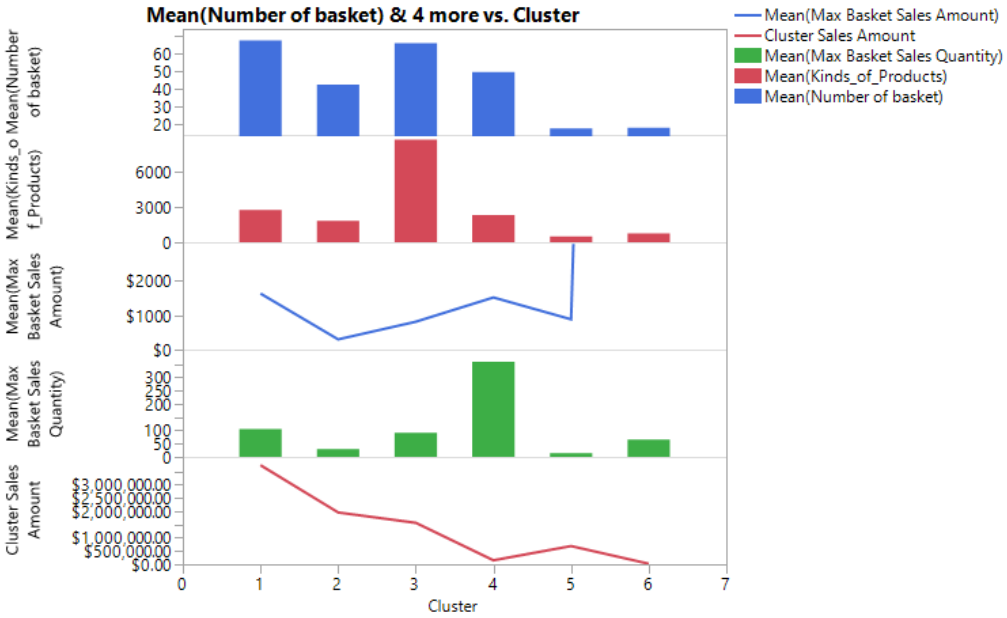


Figure 51

1. Cluster 1 have the largest number of sales amounts, accounting for about 45% of the total sales amount but only consist of 15% of the pharmacies. Cluster 1 also has the greatest number of baskets on average for the first half year and a relatively large products range. So, these pharmacies are more concentrated on purchasing a few kinds of products.
2. Cluster 2 accounts for about 24% of the total sales amount, which is about half of the contribution of cluster 1 but have twice the number of pharmacies. Cluster 2 has a moderate sales amount of each order that these pharmacies contribute a moderate sales amount for each transaction made. These pharmacies could be regular customers who have cooperative relationship with the company.
3. Cluster 3 only consists of 4.7% of the pharmacies but accounts for about 19% of the total sales amount. It is high in both basket number and products range. While these pharmacies don't tend to order frequently, they tend to spend many products for each purchasing.
4. Cluster 4 just consists of 2 pharmacies but has the highest maximum sales quantity with a low products range, which means they are more demand on fix

kinds of products. Compared to cluster 2, they are mainly different on maximum sales quantity, which could infer that these 2 clusters majored on different kinds of products. Noticed that cluster 4 has similar average sales amount to cluster 2 while purchase much more products. so, the unit price of the products purchased by cluster 4 may be much lower than that in cluster 2.

5. Cluster 5 have the largest proportion of pharmacies, which is about 45% but contributes a little for the sales amount. As shown on the graphs, they don't purchase regularly or purchase much kinds of products. These pharmacies could be some retail pharmacies or physicians.

To verify our assumption above, we can look at the sales amount of each major products for each cluster.

Cluster	pharmacy proportion	Average Sales Amount	PHOTO	DIABETES	HOME HEALTH CARE	EDIBLES	MISC	HEALTH CARE	GENERAL MERCHANDISE	BEAUTY	PERSONAL CARE	GREETING CARDS
1	15.09%	$234,062.79	$1,188.12	$1,617.07	$30,279.67	$11,462.49	$393.24	$68,572.53	$98,243.19	$6,095.67	$9,975.76	$6,235.05
2	32.08%	$57,692.41	$623.55	$1,430.44	$5,100.95	$1,582.23	$169.28	$36,024.30	$3,332.79	$2,849.49	$4,453.72	$2,125.67
3	4.72%	$313,821.26	$2,353.75	$1,619.61	$28,344.76	$9,546.91	$1,236.86	$138,964.56	$29,068.21	$18,730.01	$20,318.22	$63,638.37
4	1.89%	$77,753.97	$0.00	$771.46	$5,760.63	$1,895.92	$1,432.44	$46,259.97	$10,407.51	$2,891.19	$4,642.82	$3,692.03
5	45.28%	$14,424.04	$0.00	$264.98	$3,479.30	$174.63	$4.58	$7,388.03	$993.43	$484.55	$823.22	$811.32
6	0.94%	$25,932.31	$0.00	$665.38	$1,242.55	$6.50	$6.99	$19,077.63	$448.48	$2,366.29	$2,118.49	$0.00

Figure 52

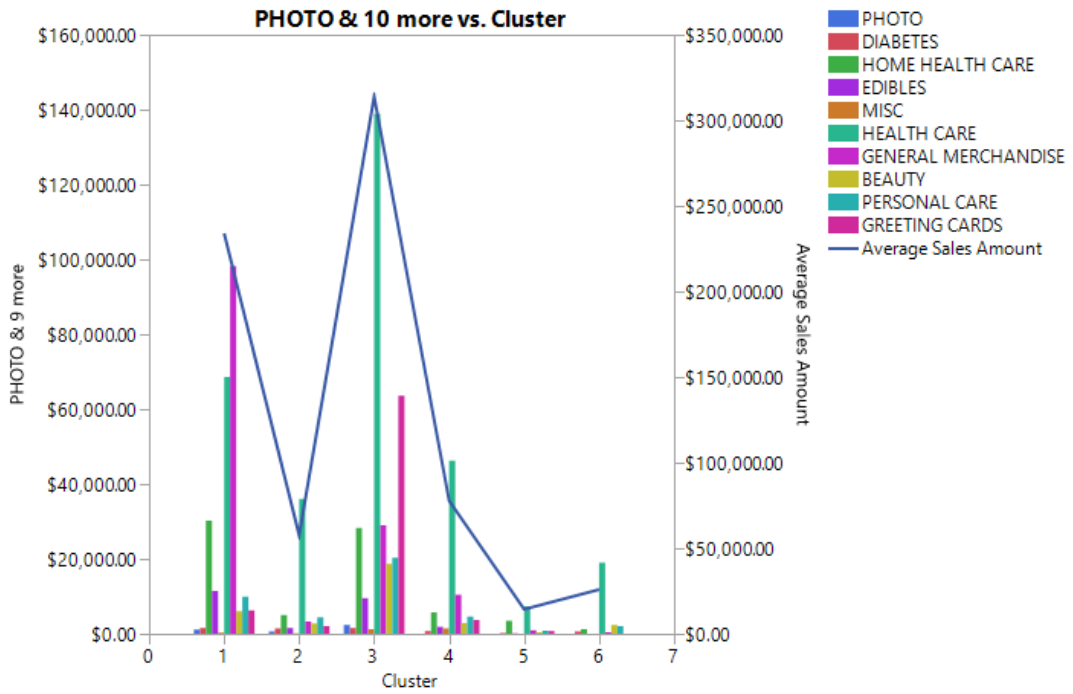


Figure 53

The table above shows that cluster 1 and 3 indeed purchased much more kinds of products than other clusters and have a larger products range with an average higher sales amount. Cluster 2 and 4 all demand mainly for Health Care products and cluster 5 demand low for almost all the products.

4. Summary

Pharmacies transacted with the company can be divided into 4 groups according to the above analysis:

Group 1: Both high in purchase quantity and amount with a large range of products in demand. These may be hospital.

Group 2: Moderate purchase quantify and amount with a high purchase frequency**.** These may be cooperative relationship customer, like medical clinics.

Group 3: Low in purchase frequency but with a high average purchase amount and quantity. These could be some long-term care pharmacies.

Group 4: Both low in purchase quantity and amount with a small range of products in demand. These may refer to individual or retail pharmacies or physician offices.

Appendix 1: Data preparation change log

Variable	Comment	Action
ZIP_3_CD	Modeling type should be nominal.	Open column information window and change it.
PROD_DESC, SEG_DESC, SUB_CAT_DESC	Description variables' modeling type should be set as unstructured text to find key words.	Change its modeling type as unstructured text.
SLS_DTE_NBR	Date variable.	First change its data type to character. Then, change it to numeric, choose ordinal as modeling type, and choose appropriate format.
EXT_SLS_AMT	Monetary variable.	Choose currency (US dollar) as the data format.
SLS_QTY, POS_TRANS	A zero value in sales quantity with a zero value in sales amount could show few inferences.	Hide and exclude these 25 rows by using row selection tool.
EXT_SLS_AMT, UNIT_PRICE	Create a new variable "UNIT_PRICE" to compute the unit price of each product transacted and find out outliers of the variable "EXT_SLS_AMT".	Set the upper limit of price, like 3 times the 90th percentile, to recognize outliers and the price greater than the limit should be outliers. Hide and exclude them.