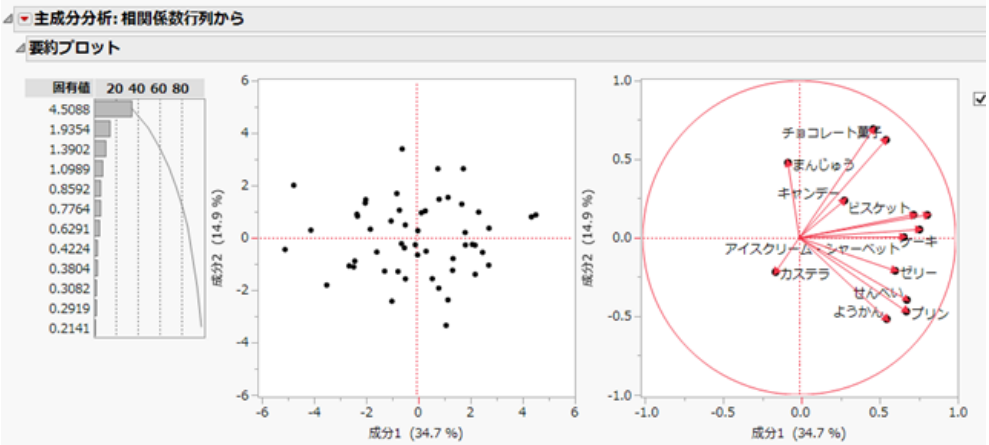
#3. 主成分分析での外れ値分析

主成分分析では、指定した次元まで主成分から計算する多変量の距離(T^2)を計算し、主成分分析における外れ値のサンプルを探すことができます。他の商品とは違う商品や、異なる評価をする消費者を見つけるといった用途で便利な機能です。

例:次のデータは、日本の主要な市町村における 2017 年の菓子類の消費金額(円)を入力したデータです。(出典:総務省　家計調査)

	市町村	ようかん	まんじゅう	カステラ	ケーキ	ゼリー	プリン	せんべい	ビスケット	スナック菓子	キャンデー	チョコレート	チョコレート菓子
1	さいたま市	1,496	984	1,083	7,468	2,387	1,898	6,810	3,622	3,800	1,976	6,513	1,043
2	宇都宮市	924	1,012	544	8,275	2,175	1,777	7,724	4,995	4,469	2,285	7,830	1,326
3	横浜市	789	672	1,179	6,993	2,169	1,443	6,237	4,887	3,274	2,162	6,749	1,198
4	岡山市	582	2,710	720	7,441	2,450	1,545	4,545	3,920	4,829	1,987	7,162	1,716
5	岐阜市	410	1,545	575	7,007	2,280	1,819	6,059	3,997	6,361	2,472	6,919	1,723
6	宮崎市	674	896	684	6,214	2,135	1,356	4,593	3,529	3,858	2,179	5,827	1,372
7	京都市	1,263	900	1,143	6,145	2,200	1,383	7,009	3,194	3,267	1,976	6,512	1,338
8	金沢市	787	1,609	1,659	9,109	2,440	1,812	6,516	4,156	6,188	2,035	8,469	1,783
9	熊本市	499	1,963	588	7,559	2,313	1,139	2,984	3,711	4,928	1,928	5,693	1,403
10	広島市	349	1,577	501	7,502	2,158	1,402	4,425	3,712	4,335	1,975	7,409	1,533
11	甲府市	855	1,160	769	6,864	2,197	1,788	4,619	3,374	3,895	2,184	6,136	1,291
12	高松市	857	2,452	724	6,233	2,612	1,233	5,495	3,639	5,921	1,930	6,637	1,360
13	高知市	884	1,849	882	7,485	1,857	1,529	3,896	5,229	6,549	1,971	7,481	2,114
14	佐賀市	948	1,845	601	6,480	2,013	1,454	4,295	4,699	4,607	2,064	6,203	1,552
15	堺市	487	515	1,018	7,654	1,985	1,430	6,755	3,925	4,112	2,381	6,900	1,261
16	札幌市	627	764	835	7,293	1,824	1,215	4,139	4,116	6,028	2,418	7,753	2,190
17	山形市	1,295	1,103	755	8,692	2,317	1,695	7,697	4,976	5,613	2,221	7,290	2,721
18	山口市	488	2,196	537	6,124	2,052	1,367	5,753	3,958	4,801	2,489	6,288	1,794
19	鹿児島市	645	2,443	967	5,319	1,677	1,132	4,477	2,898	3,378	2,188	5,185	1,279

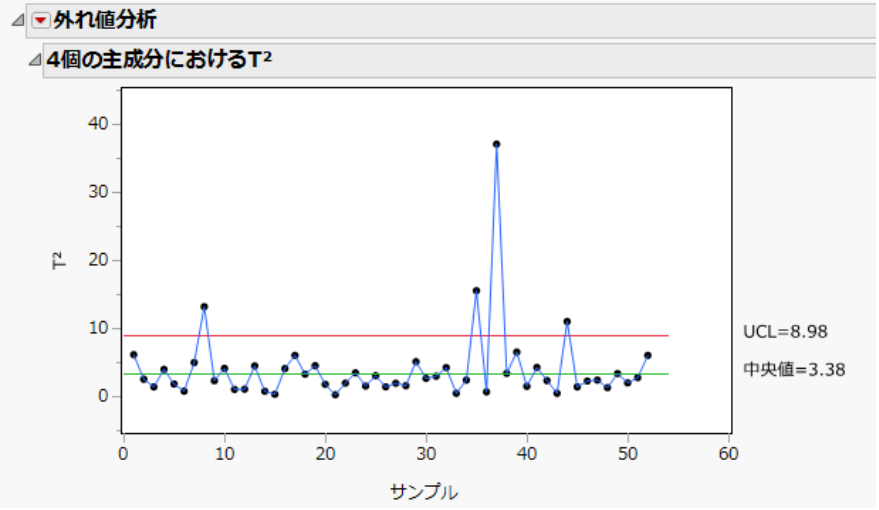
このデータを主成分分析したときの要約プロットを以下に示します。



スコアプロット(中央の図)は、市町村(サンプル)のスコアプロットです。中心から距離が遠い外れいているデータがいくつかあることがわかります。ここでは固有値の値が 1 を超える成分まで考えることにし、第 4 成分までを抽出することにします。

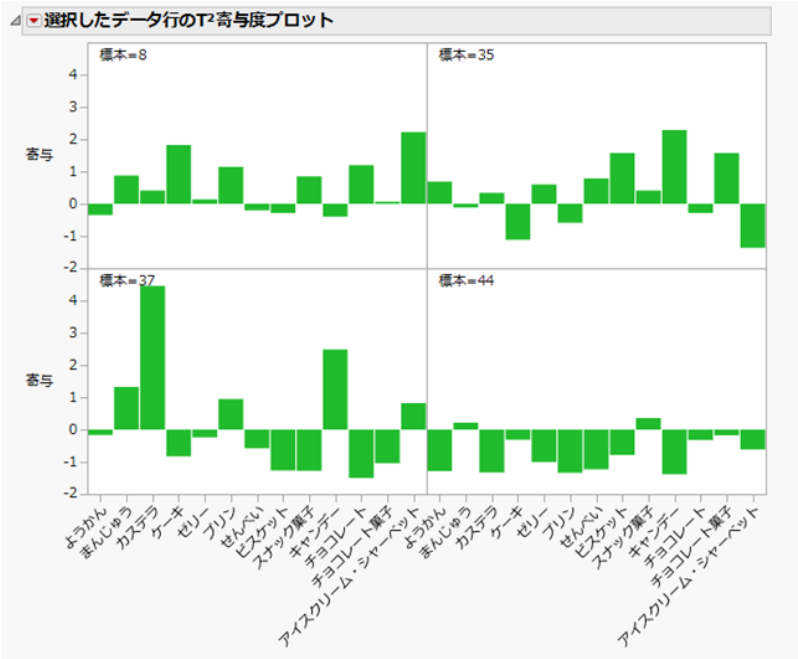
■操作:外れ値分析

1. レポート「主成分分析: 相関係数行列から」の左上にある赤い三角ボタンから [外れ値分析]を選択します。
2. レポート「外れ値分析」の左上にある赤い三角ボタンから [成分の数]を指定し 成分数を"4"に指定します。



外れ値分析のレポートです。横軸はサンプルの行番号、縦軸には T^2 がプロットされます。赤色の線は(上側)仕様限界値を示し、この赤線より上にある(値が大きい)サンプルが外れ値とみなすことができます。
この例では、4 つの外れ値がありますので、これらはどこの都市で、どの菓子が外れていることに大きく寄与しているのかを調べてみます。

グラフ上で 4 つのデータを選択(マウスカーソルをドラッグして四角の枠を作ることによりまとめて選択することができます。)し、「外れ値分析」の左にある赤い三角ボタンから [選択した行の寄与率プロット]を選択します。



「選択したデータ行の T^2 寄与率プロット」は、選択した 4 つの都市に対する、各菓子類の寄与率を棒グラフで示しています。値がプラスの場合は消費金額が高い、値がマイナスの場合は消費金額が低いことを示します。

結果の解釈例として、標本 37(左下)は長崎市です。寄与度プロットをみると「カステラ」の寄与が非常に高いことが分かります。