Dec 10, 2017

Pea Quality Judgement: A Principle Component Analysis approach

Mingyuhui Liu (Jane), Master of Science Candidate in Data Analytics

School of Engineering and Applied Sciences

The George Washington University

Table of Contents

Table of Figures and Tables

Acronyms

AIC	Akaike information criterion
PC	Principle Component
PC1	The First Principle Component
PC2	The Second Principle Component
PCA	Principle Component Analysis
Std	Standard Deviation

Introduction

An integrated dataset consisting of 17 pea quality features for 60 peas. In order to build up a metric to determine the overall quality of the peas based on the 17 features, a principle analysis is conducted and based on which, the project created a metric to rate the pea qualities.

The rating of the pea quality is based on the researcher's own taste, which is less sweet but mealy, mature and with sharp green color.

Principle Component Analysis

Preprocessing

The original dataset provides information for 60 peas based on 17 features, however, what each feature represents is not stated and whether the effect from each feature is positive or negative is also unclear. An example from the original dataset is demonstrated as Table 1. From here on, this paper will use the column names as shown in the Table 1 to refer to the features.

Since normalization is important to Principle Component Analysis (PCA), the project conducted the PCA over a normalized dataset from the original dataset. Every element in the original dataset is normalized based on the Equation 1.

Equation 1

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

z_i denotes the normalized element, x_i denotes the element from the original dataset, min(x) and max(x) represent the minimum and maximum value for a single feature from the original dataset.

Next, in order to get the general information, Table 2 is created to demonstrate the correlation with the normalized data. In Table 2, correlations with absolute values larger than 0.7 are highlighted in orange, to show great correlations from the variables. As observed from Table 1, it can be estimated that one or two principle components could be sufficient.

Principle Component Analysis

Deciding the number of Principle Components. This project conducted PCA with normalized data in Minitab 18. Figure 1 demonstrates the Scree Plot generated with several notes, and Table 3 demonstrate the Eigen values for the 17 scores, and their percentage contributions to the explanation to the variance.

Table 1

Example of the Original Data

Pea ID	Tenderometer	Dry matter	Dry matter after freezing	SucrosePercent	TotalGlucose1	TotalGlucose2	Flavour	Sweet	Fruity	Off-flavour	Mealiness	Hardness	Whiteness	Colour1	Colour2	Colour3	Skin
1	110.00	15.10	19.09	5.40	3.30	3.00	6.48	6.66	4.56	2.20	2.91	3.47	4.72	5.59	5.73	5.99	4.26
2	120.00	16.80	20.52	5.00	4.00	3.80	5.75	6.09	3.81	2.32	4.03	3.77	4.17	5.73	5.75	5.32	3.82
3	150.00	20.10	22.77	3.90	4.00	3.70	3.94	4.12	2.44	3.63	5.77	5.39	4.77	6.66	5.11	4.60	3.50
4	109.00	17.50	20.79	4.90	3.50	3.30	6.60	6.12	4.44	1.93	3.31	4.46	4.86	5.16	5.74	6.57	2.12
5	115.00	16.90	20.88	4.50	3.40	3.50	5.68	5.98	3.80	2.12	3.85	4.14	5.03	5.63	5.22	5.48	2.38

Table 2

Correlation Matrix for Normalized Data (Threshold = 0.7)

	Tenderometer	Dry matter	Dry matter after	SucrosePercent	TotalGlucose1	TotalGlucose2	Flavour	Sweet	Fruity	Off-flavour	Mealiness	Hardness	Whiteness	Colour1	Colour2	Colour3	Skin
Tenderometer	1																
Dry matter	0.941	1															
Dry matter after	0.95	0.98	1														
SucrosePercent	-0.873	-0.855	-0.861	1													
TotalGlucose1	0.862	0.874	0.879	-0.848	1												
TotalGlucose2	0.883	0.881	0.897	-0.855	0.963	1											
Flavour	-0.898	-0.886	-0.884	0.92	-0.853	-0.842	1										
Sweet	-0.908	-0.891	-0.891	0.959	-0.835	-0.839	0.951	1									
Fruity	-0.9	-0.897	-0.899	0.927	-0.855	-0.847	0.975	0.949	1								
Off-flavour	0.82	0.825	0.818	-0.877	0.773	0.751	-0.952	-0.901	-0.903	1							
Mealiness	0.878	0.906	0.899	-0.888	0.85	0.836	-0.933	-0.914	-0.969	0.836	1						
Hardness	0.947	0.929	0.934	-0.923	0.872	0.875	-0.924	-0.945	-0.944	0.854	0.927	1					
Whiteness	-0.013	-0.005	-0.018	-0.215	0.099	0.101	-0.127	-0.152	-0.088	0.107	0.084	0.03	1				
Colour1	0.017	0.005	-0.015	-0.207	0.061	0.063	-0.219	-0.172	-0.149	0.227	0.114	0.018	0.838	1			
Colour2	-0.133	-0.12	-0.108	0.372	-0.206	-0.197	0.343	0.331	0.278	-0.348	-0.234	-0.169	-0.893	-0.908	1		
Colour3	-0.268	-0.225	-0.221	0.419	-0.244	-0.237	0.448	0.393	0.398	-0.507	-0.327	-0.303	-0.414	-0.635	0.647	1	
Skin	-0.639	-0.704	-0.696	0.54	-0.535	-0.574	0.478	0.556	0.523	-0.418	-0.545	-0.628	0.11	0.226	-0.124	-0.016	1

Table 3

Component Principles' Eigenvalues and General Information

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10	Z11	Z12	Z13	Z14	Z15	Z16	Z17	Total
Eigenvalues	11.42	3.25	0.76	0.51	0.30	0.18	0.15	0.12	0.08	0.05	0.05	0.03	0.03	0.03	0.02	0.01	0.01	17.0
Percentage	0.67	0.19	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Cumulative	0.67	0.86	0.91	0.94	0.96	0.97	0.97	0.98	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	

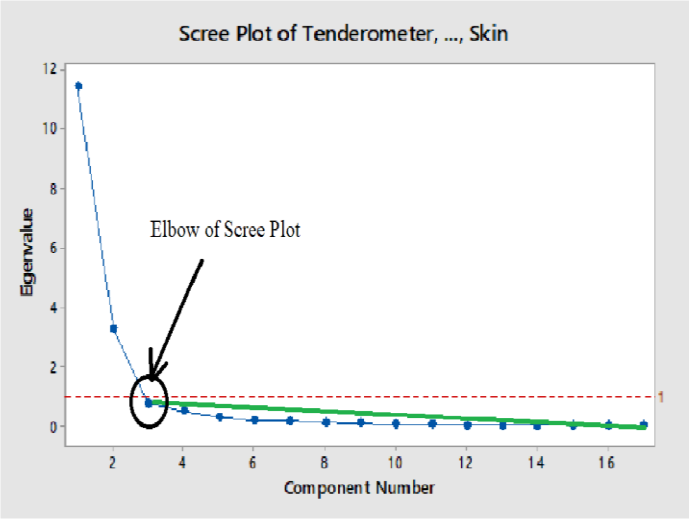

Figure 1. Scree plot with notes. This figure indicates the number of principle components (PC) to be kept in the analysis.

From the Figure 1, the green line shows the location for the "Elbow of Scree Plot", which further indicates that the number of PCs needed in the analysis is two. At the same time, the red dash line in Figure 1 confirms the number of PCs to be chosen based on the Kaiser's Rule (Cliff, N., 1988). Thus, this project decided on utilizing 2 PCs.

From Table 3, the first 2 PCs can explain at least 86% of the whole dataset. In order to determine what the 2 PCs representing in the analysis, a loading analysis was performed with Minitab 18 (Minitab, Inc., 2017).

Determining the Meaning of the Principle Components. Figure 2 demonstrates the loading plot for the first 2 PCs, with features labeled.

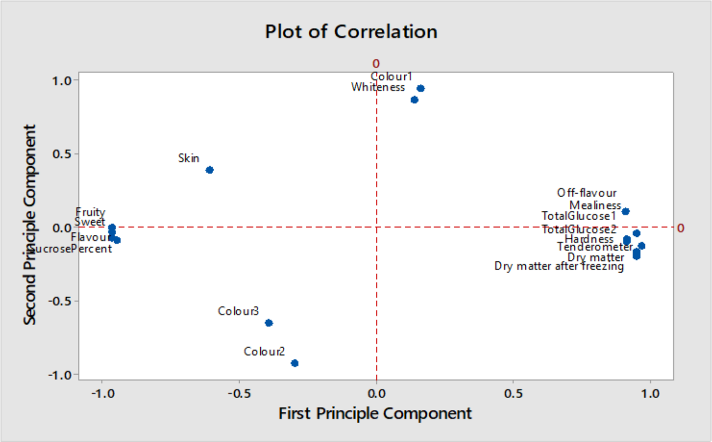

Figure 2. Loading plot for the first 2 PCs with feature names. This figure indicates the meaning for each PC.

From Figure 2, it can be concluded that the first component is expressing information on the texture, or tasting qualities, since the features that show great absolute values along the first PC direction include Mealiness, Tenderometer, Sweet, Fruity, etc.. Therefore, the first PC can be deemed as the index for texture/taste quality. Meanwhile, the second PC can be treated as a classification for the appearance features, since every feature that is showing great value over the second PC represents the appearance features, such as colors, whiteness, and skin. This conclusion drawn from the Figure 2 can also be backed up by the loading analysis demonstrated through Table 4.

Table 4

Loading Analysis on the First 2 PCs (Threshold = 0.6)

Loadings	PC1	PC2
Tenderometer	0.948	-0.164
Dry matter	0.950	-0.189
Dry matter after freezing	0.951	-0.202
SucrosePercent	-0.953	-0.089
TotalGlucose1	0.911	-0.084
TotalGlucose2	0.913	-0.095
Flavour	-0.969	-0.070
Sweet	-0.968	-0.037
Fruity	-0.970	0.000
Off-flavour	0.909	0.107
Mealiness	0.948	-0.044
Hardness	0.968	-0.128
Whiteness	0.136	0.871
Colour1	0.160	0.943
Colour2	-0.301	-0.925
Colour3	-0.396	-0.656
Skin	-0.612	0.387

In Table 4, all the cells with an absolute value of the correlation value larger than 0.6 are highlighted. It is very clear that other than the colors. Noted that even though Skin could be explained with the first PC (PC1) with a PC1 score of -0.612, which is actually much smaller than others, in order to summarize the PCs in a more concise phrase, Skin, as an appearance indicator, is still concluded with the second PC (PC2).

In summary, through the PCA, 2 PCs are extracted, the first one is representing the tasting quality of the peas, and the second one is demonstrating the appearance of the peas.

Pea Quality Judgement Metrics

Metrics Construction

Interpretations in detail. To obtain a general idea about the judgement on the pea qualities, such as which features should have opposite effects, Table 5 was created, which is based on the original data. In Table 5, correlations with a more dramatic correlation than -0.8 are highlighted. A “equal weights metric” was then generated based on the signs of the correlations.

Table 5

Correlation Matrix (Threshold = -0.8) with an Equal Weights Metric

Equal Weights Metric	1	1	1	-1	1	1	-1	-1	-1	1	1	1	1	1	-1	-1	1
	Tenderometer	Dry matter	Dry matter after	SucrosePercent	TotalGlucose1	TotalGlucose2	Flavour	Sweet	Fruity	Off-flavour	Mealiness	Hardness	Whiteness	Colour1	Colour2	Colour3	Skin
Tenderometer	1																
Dry matter	0.941	1															
Dry matter after	0.95	0.98	1														
SucrosePercent	-0.873	-0.855	-0.861	1													
TotalGlucose1	0.862	0.874	0.879	-0.848	1												
TotalGlucose2	0.883	0.881	0.897	-0.855	0.963	1											
Flavour	-0.898	-0.886	-0.884	0.92	-0.853	-0.842	1										
Sweet	-0.908	-0.891	-0.891	0.959	-0.835	-0.839	0.951	1									
Fruity	-0.9	-0.897	-0.899	0.927	-0.855	-0.847	0.975	0.949	1								
Off-flavour	0.82	0.825	0.818	-0.877	0.773	0.751	-0.952	-0.901	-0.903	1							
Mealiness	0.878	0.906	0.899	-0.888	0.85	0.836	-0.933	-0.914	-0.969	0.836	1						
Hardness	0.947	0.929	0.934	-0.923	0.872	0.875	-0.924	-0.945	-0.944	0.854	0.927	1					
Whiteness	-0.013	-0.005	-0.018	-0.215	0.099	0.101	-0.127	-0.152	-0.088	0.107	0.084	0.03	1				
Colour1	0.017	0.005	-0.015	-0.207	0.061	0.063	-0.219	-0.172	-0.149	0.227	0.114	0.018	0.838	1			
Colour2	-0.133	-0.12	-0.108	0.372	-0.206	-0.197	0.343	0.331	0.278	-0.348	-0.234	-0.169	-0.893	-0.908	1		
Colour3	-0.268	-0.225	-0.221	0.419	-0.244	-0.237	0.448	0.393	0.398	-0.507	-0.327	-0.303	-0.414	-0.635	0.647	1	
Skin	-0.639	-0.704	-0.696	0.54	-0.535	-0.574	0.478	0.556	0.523	-0.418	-0.545	-0.628	0.11	0.226	-0.124	-0.016	1

When comparing Table 4 and Table 5, the signs for PC1 match the signs of the signs of the equal weights, which makes the interpretation for PC1 relatively clear: When PC1 increase, some of the tasting indicators – Tenderometer, Dry Matter, Dry Matter after Freezing, Total Glucose1, Total Glucose2, Off-flavour, Mealiness and Hardness will increase, while the others – SucrosePercent, Flavour, Sweet and Fruity will decrease. Considering the indicators that increase with PC1, they are mainly about the texture - whether the pea is tender, whether it contains sufficient pea matter, or whether it tastes mealy. While the indicators that decrease with PC1 are mainly about the taste - whether the peas are sweet, fruity and flavorful. Therefore, the increasing in PC1 indicates the increase in the peas' dense texture and decrease in the peas' sweet taste. To illustrate PC1 for pea texture and tasting, Figure 3 is plotted.

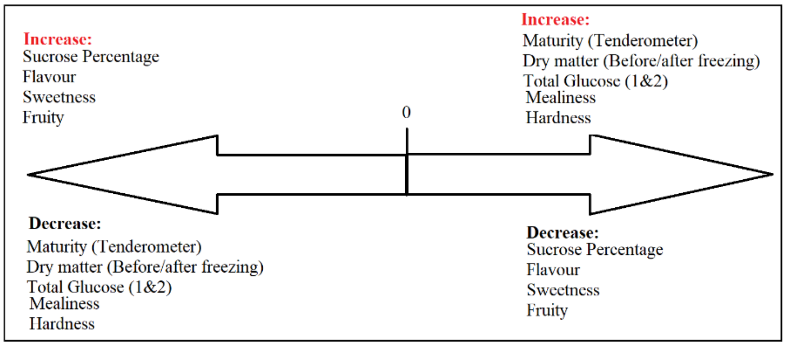

Figure 3. Interpretation of PC1. This figure interprets PC1.

To interpret PC2, observed from Figure 2, the Whiteness and Colour1 are behaving in the similar way, while Colour2 and Colour3 are totally the opposite, which can be seen from Table 5 as well. It is rational to assume that the Colour1 is representing certain feature that is similar to the Whiteness while Colour2 and Colour3 are representing the opposite of them. As for the Skin, observed from Figure 2, the decrease of the PC1 and the increase of the PC2 will contribute positively to the Skin. Figure 4 is plotted to illustrate the PC2 attributes for pea appearance.

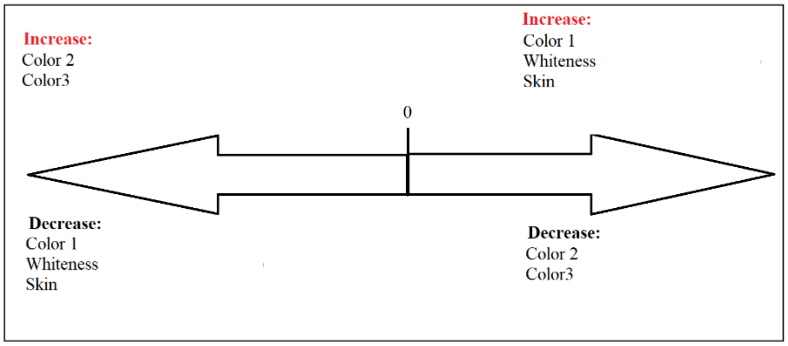

Figure 4. Histogram for Logged Crime. This figure shows normality information on Logged Crime.

Metrics. Instead of using the equal weighted features, the project utilized the PC1 and PC2 as the weights for the features. Then the original dataset with dimension of 60x17 was transformed into new a 60x2 matrix. This is done by Equation 2.

Equation 2

$$\boldsymbol{Metric} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,17} \\ \vdots & \ddots & \vdots \\ x_{60,1} & \cdots & x_{60,17} \end{bmatrix} \cdot \begin{bmatrix} pc_{1,1} & pc_{1,2} \\ \vdots & \vdots \\ pc_{60,1} & pc_{60,2} \end{bmatrix}$$

Metric denotes the 60x2 matrix for 60 peas' new scores based on the selected features, and $x_{i,f}$ represents the score for the *f*th feature from the *i*th pea. $pc_{i,1}$ and $pc_{i,2}$ are the PCs for the *i*th pea.

Table 6 demonstrates the new scores for the 60 peas based on the two PCs.

Table 6

New Scores based on the Selected PCs

Pea ID	PC1	PC2	Pea ID	PC1	PC2
1	-3.115	0.048	31	-2.963	-1.293
2	-1.528	-0.174	32	0.268	-1.681
3	1.657	0.556	33	1.834	-0.902
4	-1.872	-0.523	34	-3.511	0.269
5	-1.153	0.123	35	-1.028	-0.484
6	1.651	-0.014	36	-0.645	-1.552
7	-3.090	0.252	37	-2.672	0.237
8	-1.314	-0.002	38	-0.500	0.288
9	3.657	0.218	39	2.033	0.560
10	-1.404	-0.240	40	-1.737	-0.250
11	-1.295	-0.858	41	0.253	-1.275
12	5.543	-1.057	42	4.956	-0.547
13	0.214	-0.098	43	-3.125	1.011
14	0.391	-0.321	44	0.357	-1.461
15	4.636	-0.650	45	3.242	-0.702
16	-3.132	-1.358	46	-0.749	-1.154
17	-0.632	-0.886	47	-3.389	-0.305
18	3.390	-0.701	48	2.963	-0.544
19	-2.540	1.523	49	-2.014	0.462
20	0.918	0.881	50	0.450	-0.614
21	3.772	0.097	51	7.636	0.931
22	-2.790	0.427	52	-3.964	-0.725
23	-0.558	-1.925	53	-2.606	-1.133
24	5.091	-0.334	54	1.419	-1.227
25	-2.469	-1.015	55	-2.681	-0.779
26	-1.845	-1.427	56	-0.334	-1.505
27	6.391	-0.858	57	3.508	-1.951
28	-0.176	0.014	58	-3.440	-0.808
29	3.817	-0.387	59	-0.543	-0.016
30	0.648	-0.948	60	4.154	-0.021

The project then conducted distribution fittings based on Table 6.

Distribution Fittings

The tools used for the fittings is @Risk from the Palisade DecisionTools. Based on the process the @Risk's distribution fitting tools handling the fittings, the project decided to fit each PC with different distributions.

The fittings were conducted based on "equal probability" chi-square binning process, and on the assumption of continuous probabilities. Then several models were fitted, and in the end, the model with the lowest Akaike information criterion (AIC) was selected (Palisade Corporation, n.d.).

Results showed that a Triangle distribution was fitted to the scores from PC1, and a normal distribution was selected for PC2.

Figure 5 and Figure 6 demonstrate the distribution fittings for the scores.

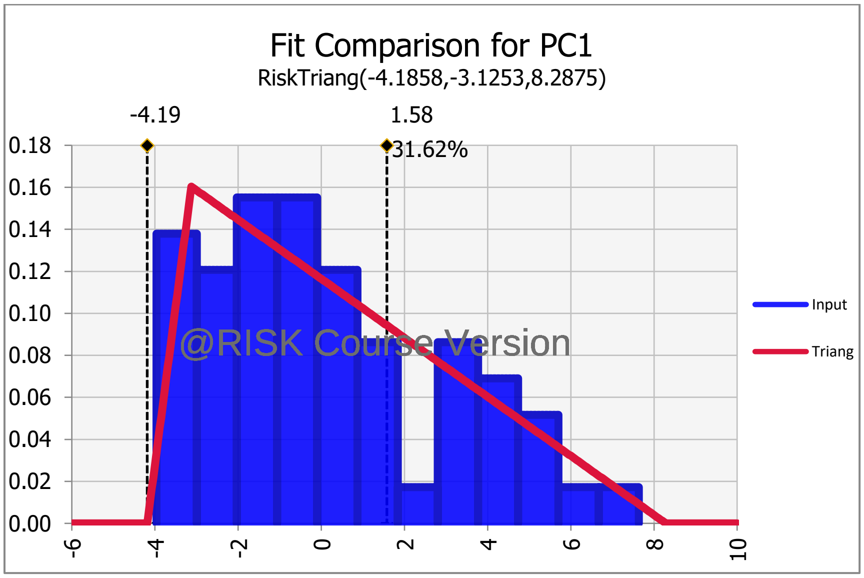

Figure 5. Triangular Distribution for PC1. This figure shows the distribution fitted for PC1.

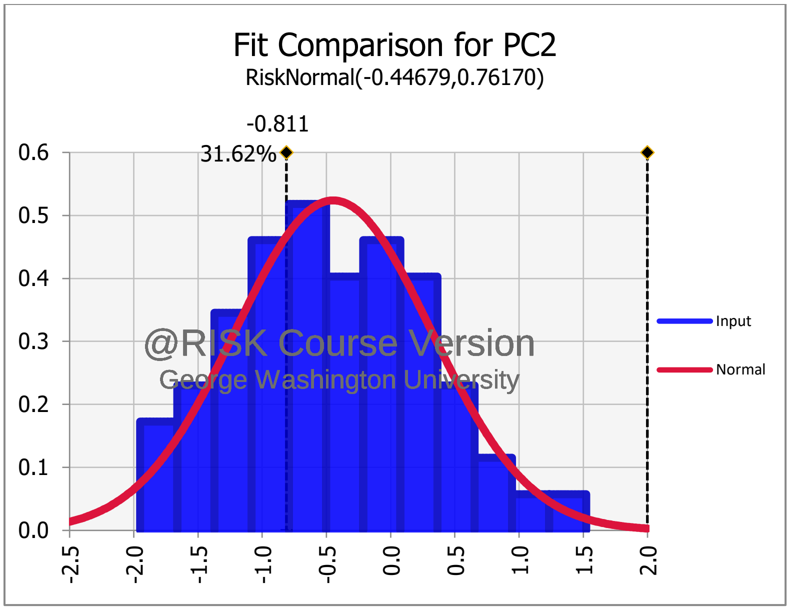

Figure 6. Normal Distribution for PC2. This figure shows the distribution fitted for PC2.

According to @Risk, the three parameters featured in the triangular distribution are -4.19, -3.13, and 8.29. Meanwhile, the mean and standard deviation for the normal distribution for PC2 is -0.45 and 0.76.

Peas Selection

Recall in the introduction, one of the objectives for this project is to select the best peas based on the researcher’s own taste, which is less sweet while is firm, mature, which should be measured by Tenderometer (Food Technology Corporation, n.d.), and mealy and in sharp green. After translating the researcher’s taste into the descriptions used in this research, the peas that will stand out should be the ones score high in PC1, and low in PC2, since as interpreted in Metrics Construction section: when the PC1 scores high, the peas are mealier and firmer while it tastes less sweet; and when the PC2 scores high, the peas show more whiteness other than the sharp green.

Based on the rationale described above, reference lines are drawn in Figure 5 and Figure 6, indicating the upper 31.68% scores in PC1, and lower 31.68% scores in PC2, on the distribution fitting line. The reason for a 31.68% line is that, ideally, 31.68% of 31.68% dataset should consist 10% data points from the dataset. From the fitting line, the upper 31.68% for the triangular distribution of PC1 locates at 1.58, and the lower 31.68% for the normal distribution of PC2 falls at -0.811. Based on these two values, Figure 7 is plotted to choose the top 10% peas according to the researcher’s taste.

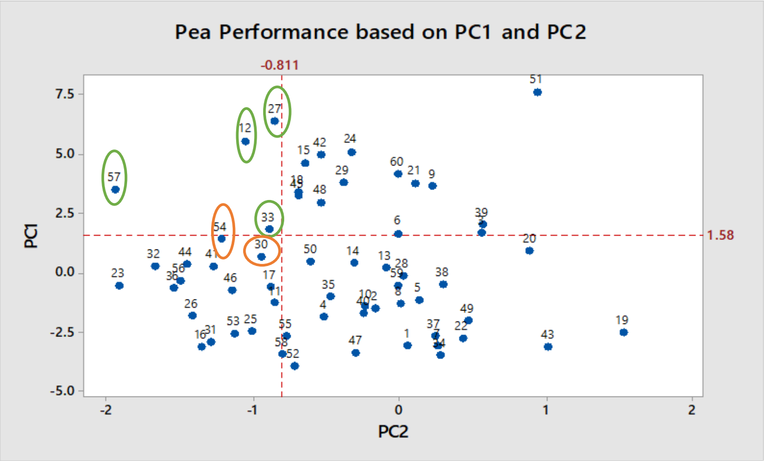

Figure 7. Top 10% Peas Selection Plot. This figure shows the guide map to choose the top 10% peas based on the researcher’s taste.

In Figure 7, peas with data points falls above the 1.58 reference line and to the left of the -0.811 reference line should be selected. In the end, the peas with ID number of 12, 27, 33, and 57 fall strictly into the desired area. These data points are circled in green in Figure 7. Two more peas need to be selected. Given that the researcher did not have a preference for one PC over another, which means whichever two peas have data points closest to the intersection of the two reference lines should be selected. As a result, Pea 30 and Pea 54 are selected, which are circled in orange in Figure 7.

To summarize, Pea 12, 27, 30, 33, 54 and 57 are selected as the top 10% of the peas that outperformed the others when it comes to the researcher’s own taste.

Summary and Discussion

Summary

After PCA was performed, 2 PCs are chosen: PC1 represents the taste and the texture of the peas and PC2 shows the appearances of the peas. When PC1 scores high, the texture becomes more mature, firmer and mealier with greater matter in the peas, while when it scores low, the sweetness of the peas goes up; when PC2 scores high, the color of the peas shows more whiteness.

The research treated PC1 and PC2 as weights for the 17 features, and in the end, based on the original scores and the new weights, 6 peas are selected based on the researcher’s taster – firm, unsweet and green: Pea number 12, 27, 30, 33, 54 and 57.

Discussion

The project conducted similar PCAs over the raw data, normalized data and standardized data, and found that the PCA results returned as the same, which saved the project time to determine which data to use.

Electronic Files

Since this research formally utilized normalized data, only normalized data analysis is provided with the electronic files.

References

Cliff, N. (1988). The Eigenvalues-Greater-Than-One Rule and the Reliability of Components. *Psychological Bulletin, Vol 103, No.2,* 276 – 279

Food Technology Corporation. (n.d.) Retrieved on December 10, 2017 from Food Technology Corporation: http://www.foodtechcorp.com/tu-12-tenderometer

Minitab 18 Statistical Software (2017). [Computer software]. State College, PA: Minitab, Inc. (www.minitab.com)

Palisade Corporation. (n.d.) Retrieved on December 10, 2017 from Palisade: http://www.palisade.com/risk/