PENNSTATE

February 25th, 2024

From: Suzanne Papik, Cameron Vardaman, and Christopher Pirch, *Graduate Student Consultants*

To: Dr. Marvel Frankenstein, *Get Cheese Right*

FRANKENSTEIN'S MUENSTER: A CHEESE ANALYSIS

EXECUTIVE SUMMARY

More than 1000 cheese varieties exist in the world, which vary in both characteristics and thermophysical properties. Researchers at *Get Cheese Right* were interested in analyzing cheese samples from two manufacturers to understand how thermophysical characteristics relate to cheese texture and identify attributes of the cheese that can be used to determine how many cheese varieties are present in their dataset.

Analysis of the data resulted in the findings that the texture of the cheese played a significant role in the value of the thermophysical characteristics. That is, the values of the thermophysical properties of the cheese would depend on the texture of the given cheese. Additional analysis resulted in the findings that there are three different varieties of cheese present in the data provided from *Get Cheese Right*. Lastly, a model was created to be able to predict the texture of a new cheese given we know the values of the cheese's thermophysical characteristics.

1.0 - PROJECT DESCRIPTION

Dr. Marvel Frankenstein and his team from *Get Cheese Right* has engaged a team of graduate student consultants from Penn State University to assess thermophysical characteristics of four common cheese textures and to identify the main attributes that are suitable for the grouping of the products.

Dr. Frankenstein's team collected cheese sample data and recorded the cheese texture and thermophysical characteristic information. Since the data was collected without intervention or treatment, the type of study is observational. The scope of this report includes review of the variables and dataset, analysis of the data utilizing statistical techniques, and supplying recommendations that address Dr. Frankenstein's research questions. This includes providing insights on how the thermophysical variables differ among cheese textures, determining how many varieties of cheese are present in the data, and developing a model that can be used by the researchers at *Get Cheese Right* to predict future cheese textures based on thermophysical characteristics.

1.1 - RESEARCH QUESTIONS

The following research questions will be addressed in this report:

1. Are the thermophysical properties of the four cheese textures (Hard, Pasta Filata, Semi-Hard, Soft) different? If so, which textures are different?
2. Based on thermophysical characteristics, how many cheese varieties are present in the data?
3. Can we produce a model to predict the texture of a new cheese product using the thermophysical characteristics?

1.3 – VARIABLES

The dataset used in this analysis, provided by Dr. Frankenstein and his team, includes 89 cheese samples of various textures sampled from two manufacturers. There are nine total variables, six of which are thermophysical property variables (G80, vLTmax, vCO, Fmax, FD, FO), one of which is an identifier (ID), and two are categorical variables (manufacturer, cheesetexture). The variables included in the dataset are:

Variable Name	Description	**Unit of Measurement**	Valid Values
ID	A unique ID for each cheese sample	NA	1 to 89
Manufacturer	Cheese Manufacturer	NA	1 or 2
CheeseTexture	Cheese Texture	NA	Hard, Pasta Filata, Semi-Hard, Soft
G80 *	Storage modulus at 80C	Pa	[0,∞)
vLTmax *	Temperature v at tan	Degrees Celsius	[0,∞)
vCO *	Temperature v at cross-over	Degrees Celsius	[0,∞)
Fmax *	Max resistant force during extension of melted cheese	mm	[0,∞)
FD *	Flowing Degree	NA	(-∞,∞)
FO *	Free Oil	NA	[0,∞)

*Indicates Thermophysical Characteristic Variables

Table 1.3.1: Variables included within the Cheese Dataset

To address the research questions, the cheese texture variable and the manufacturer variables will be treated as explanatory variables and the six thermophysical characteristic variables are treated as response variables. Treating the variables in this way allows us to analyze how texture and manufacturer may influence or effect the values of the thermophysical characteristics.

2.0 - EXPLORATORY DATA ANALYSIS (EDA)

The dataset consists of eighty-nine observations, where each observation is a cheese sample. There are no missing values in the dataset, and all eighty-nine observations have a value for each of the nine variables. An assessment of each of the six thermophysical characteristic variables, which can be seen in Figure A.1 in Appendix A, reveals there are only potential outliers (or extreme data points) for the FO variable. However, an outlier test (Figure A.2 in Appendix A) indicates these datapoints do not need to be investigated further as outliers. It is prudent to keep these potential outliers in mind (especially during the discriminant analysis procedure) if any inaccuracies are present.

There are four unique cheese textures within the data, and there are between twenty to twenty-four cheese samples of each texture. The cheese samples were taken from two manufacturers, where forty-five observations were sampled from manufacturer one and forty-four observations were sampled from manufacturer two. The distribution for these two variables can be seen in Tables 2.1 and 2.2.

Manufacturer	Number of Cheese Samples	Percentage
1	45	50.56%
2	44	49.44%

Table 2.1: Summary of the Manufacturer Variable

CheeseTexture	Number of Cheese Samples	Percentage
Hard	23	25.84%
Pasta Filata	22	24.72%
Semi-Hard	24	26.97%
Soft	20	22.47%

Table 2.2: Summary of the Cheese Texture Variable

For vLTmax, the average value and the standard deviation, or spread of the data, is similar across cheese textures indicating that texture may not play a factor into the value of the cheese's vLTmax. The opposite is true for the remaining five thermophysical characteristic variables, where a wide range of average and standard deviation values

can be seen across the textures. The summary statistics for the six thermophysical variables can be found in Tables A.1 and A.2 in Appendix A.

Since the research questions are interested in the thermophysical properties across the four cheese textures, the distribution of the six thermophysical variables broken out by cheese texture can be seen in Figure 2.1. The mean value of the thermophysical property appears to significantly differ by texture for the G80, FO, and FD variables. There appears to be some variation across texture for the other three variables, however not as much.

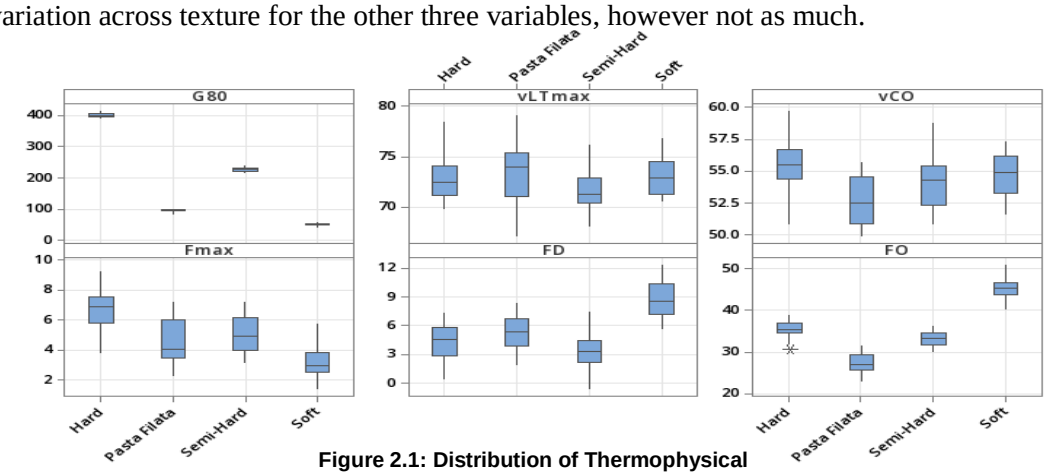


Figure 2.1: Distribution of Thermophysical Characteristic Variables by Texture

3.0 –STATISTICAL ANALYSIS

The first research question aims to address if the thermophysical properties of the four cheese textures are different, and if so, which textures are different? A multivariate analysis of variance (MANOVA) test was carried out to address this. MANOVA is a statistical test used to determine if differences between groups (textures) exist across multiple dependent variables (thermophysical characteristic variables).

To determine if there are differences in thermophysical properties across manufacturer, and if this variable should be included in subsequent analysis, an initial MANOVA was conducted. The thermophysical variables were treated as responses and the manufacturer was treated as the explanatory variable. This test did not yield a significant result (p=0.521, see Table A.3 in Appendix A), therefore the manufacturer variable was not included for the remaining tests addressing the first research question since the thermophysical variables did not significantly differ across manufacturers. All model assumptions were met for this model, which are explored in Appendix A Section A.2.1.

Another MANOVA was conducted to directly address the first research question. The six thermophysical characteristic variables were treated as response variables and the cheese texture variable was treated as the explanatory variable. This test yielded a significant result (p-value~0, Table A.4 in Appendix A). Therefore, we can conclude there are differences in the thermophysical properties across cheese textures. This model was also found to meet all assumptions, which are further explored in Appendix A Section A.2.2.

To investigate which textures are statistically different for each thermophysical characteristic, six one-way ANOVA tests were performed. The ANOVA is like the MANOVA but looks to see if there are differences in groups (textures) for one response variable instead of multiple variables. The texture was treated as the explanatory variable for each test, and the thermophysical characteristic variables were the response variables.

Since we conducted six ANOVA tests simultaneously, Bonferroni's correction was applied to account for the multiple comparisons when assessing test significance. This resulted in a level of significance of 0.0083, where a p-value smaller than 0.0083 from any of the ANOVA tests indicates a significant result.

The test for the vLTmax variable vs. texture was not significant (p-value=0.062), which indicates that the value of vLTmax does not depend on cheese texture. All other thermophysical characteristic variable values did significantly differ based on cheese texture (p-value~0). Full results of all six tests can be seen in Table A.5 in Appendix A. Plots comparing the average values of the thermophysical characteristic variables for each texture, highlighting which textures are different from one another and which textures lead to higher/lower values, can be found in Figures A.8-12 in Appendix A. In summary, the following results were observed:

- **G80:** All textures led to significantly different mean G80 values. The hard texture had the largest mean G80, followed by semi-hard, then by pasta filata, and the soft texture had the smallest G80 mean value.
- **vCO:** Pasta filata was the only texture that significantly differed in the average vCO. Pasta filata has a statistically smaller mean vCO compared to hard, soft, and semi-hard textures.
- **Fmax:** Pasta filata and semi-hard textures are not significantly different from one another regarding average Fmax. All other textures are significantly different from one another. The hard texture had the largest mean Fmax, followed by semi-hard / pasta filata. The soft texture had the smallest mean Fmax.
- **FD:** The soft texture had the largest average FD and is statistically different from all other textures. Pasta filata and semi-hard textures also statistically differ from one another in terms of average FD. All other differences in texture were not significant.
- **FO:** All textures led to significantly different mean FO values. The soft texture had the largest mean FO, followed by the hard texture, then by the semi-hard texture, and finally pasta filata.

For the second research question, a statistical technique called Principal Component Analysis (PCA) was used to reduce the complexity of the six thermophysical properties down to two variables called Principal Components, explaining the 70% of the variation (37% and 33%) in the dataset (Figure A.13-14 of Appendix A).

To determine the appropriate number of clusters, K-means clustering was used. Figure A.15 in Appendix A shows the reduction of the variation within each cluster as we increase the number of cheese varieties found in the dataset. In other words, as we increase the number cheese varieties in the data set, there are fewer samples in each group, and therefore less variation within each group. Based on K-mean clustering the optimal number of cheese varieties occur at the bend in the graph, which appears when there are three or four varieties.

Figure 3.1 below shows how well K-means clustering does in grouping the cheese samples based on the values for the first two principal components. Using K=3 clusters produced three distinct non-overlapping cheese varieties, as seen by the color groupings. The arrows represent the strength of each variable with respect to the first two PCs and can show where each variety has the largest values. The first cheese variety has the highest values of FD and FO, while it has the lowest values of G80 and Fmax. The third cheese variety has the highest values of G80, Fmax, and vCO. The

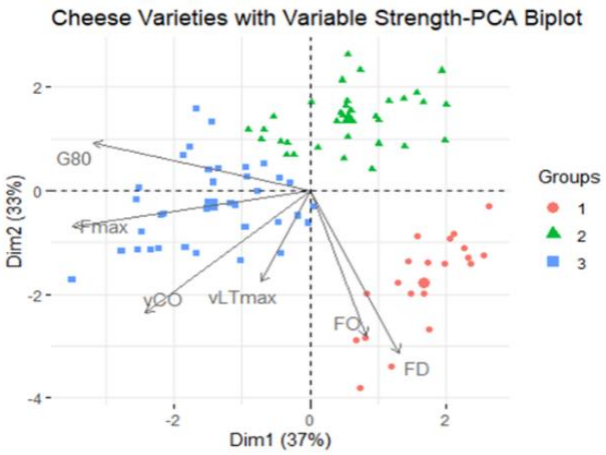


Figure 3.1: Cheese Clusters Plotted Against 1st and 2nd Principal Components

second cheese variety has the lowest values of vCO and vLTmax since it is clustered in the opposite direction as the arrows. When the arrows are close to the same length and in the same direction, those variables are correlated, which implies that as one variable increases the other tends to increase as well. For example, G80 and Fmax are highly correlated (0.68), as are Fmax and vCO (0.672), and FD and FO (0.586) (Figure A.16 of Appendix A).

Choosing the optimal number of clusters is subjective, but when clustering for more than three varieties of cheese, the classification of the groups is not as clean, with some varieties mixed in with others (Figures A.17-19 in Appendix A). Therefore, we suggest there are three varieties of cheese in the data. Table 3.1 shows the breakdown of the clusters (varieties) by cheese texture while Table A.7 in Appendix A gives a numerical summary. It is interesting to note, apart from soft cheeses, the varieties of cheese contained in the data set do not correspond exactly with the four cheese textures.

	Number of Observations in Each Cheese Variety		
CheeseTexture	Variety 1	Variety 2	Variety 3
Hard	0	4	19
Pasta Filata	0	15	7
Semi-Hard	0	14	10
Soft	20	0	0

Table 3.1: Cheese Varieties Categorized by Texture

The final research question concerns whether an accurate model could be produced to classify new cheese samples based on pre-existing data. For this question, we conducted a discriminant analysis to create a predictive model to classify new cheese samples. Prior to forming the model, Bartlett's test was utilized to determine the most appropriate discriminant analysis method for the data set. The results of this test determined that a linear discriminant analysis would be the most appropriate (see Appendix A and B for code/numerical findings).

The current data set was used to conduct a multivariate linear discriminant analysis with the texture as the grouping variable and the six continuous response variables as the predictors. As a result, four sets of coefficients were returned to form four equations to determine what texture a new cheese sample would be based on its thermophysical characteristics (Table 3.2). To predict what texture a new cheese sample is, the values of its thermophysical variables should be plugged into each equation. The equation resulting in the largest value is the predicted texture of the new cheese product.

Texture	Equation
Hard	-8296.7+17.6G80+31.1vLTmax+117.8vCO-242.6Fmax-2.4FD+65.4FO
Pasta Filata	-4682.1+4.0G80+30.3vLTmax+119.4vCO-231.3Fmax+2.3FD+55.0FO
Semi-Hard	-5959.7+9.9G80+30.7vLTmax+122.1vCO-242.3Fmax-0.4FD+61.7FO
Soft	-6262.9+2.3G80+33.1vLTmax+141.8vCO-274.8Fmax+3.7FD+67.7FO

Table 3.2: Cheese Texture Classification Models

To verify the accuracy of this new model, all 89 existing observations were placed into the model using the resubstitution method, and all 89 were subsequently placed in their actual texture, resulting in a 1.000 proportion of success for the model, or perfect accuracy for the model (Table A.8, Appendix A). Therefore, this model appears to be extremely accurate in predicting the texture of cheese samples based on their thermophysical characteristics.

4.0 – RECOMMENDATIONS

1. The texture of cheese significantly affects the thermophysical properties of cheese. The value of the thermophysical properties G80, vCO, Fmax, FD, and FO did all differ based on the texture of the cheese. This was not seen for the vLTmax variable.
2. Based on K-means clustering with Principal Component Analysis we suggest there are three varieties of cheese present in the data set based on their scores for the first two principal components.
3. Based on the results of our discriminant analysis using our current data set, we can strongly suggest that we could predict the texture of new cheese products based on their thermophysical characteristics.

5.0 – RESOURCES

For additional information regarding the techniques or content covered in this analysis please refer to:

- Multivariate Analysis of Variation (MANOVA) - https://online.stat.psu.edu/stat505/lesson/8
- Discriminant Analysis - https://online.stat.psu.edu/stat505/lesson/10
- Principal Component Analysis - https://online.stat.psu.edu/stat505/lesson/11
- Clustering Analysis - https://online.stat.psu.edu/stat505/lesson/14

6.0 – CONSIDERATIONS

The sample data is limited with under 100 observations from only two manufacturers, whereas pre-existing information provided by Dr. Frankenstein points to the existence of over 1000 varieties of cheese, likely from numerous manufacturers, so these results may not be widely generalizable outside the conditions that the data were sampled under. To further research these additional cheese varieties and differences among manufacturing protocols, an additional study can take place with a larger sample size across more manufacturers, collecting more information than the six thermophysical properties.

Another consideration is that this analysis only focused on six thermophysical properties collected by Dr. Frankenstein. There are possibly many more other variables or characteristics that could be considered for answering the research questions posed at the beginning of this report. For example, there are other characteristics (smell, color, taste, density, etc.) of cheese that may also be used in the classification of cheese varieties or the prediction of the texture of a new cheese product.

Finally, SAS, R, and Minitab were all used to conduct the analysis and produce recommendations. A level of significance of 0.05 was used unless specified otherwise.

We appreciate the Gouda opportunity to work with the *Get Cheese Right* team on this project!

Appendix A- Additional Figures and Tables

A.1- Exploratory Data Analysis Figures and Tables

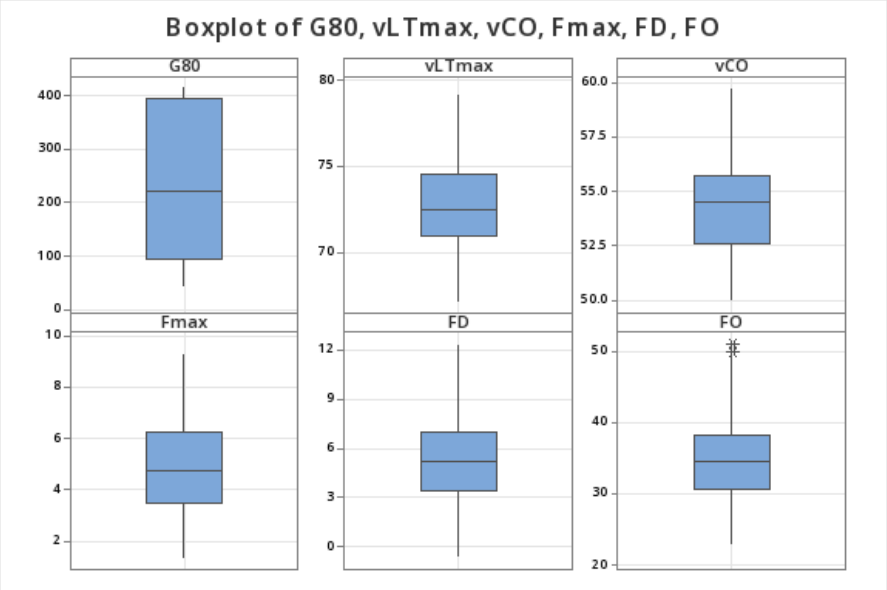


Figure A.1: Distribution of the Six Thermophysical Characteristic Variables

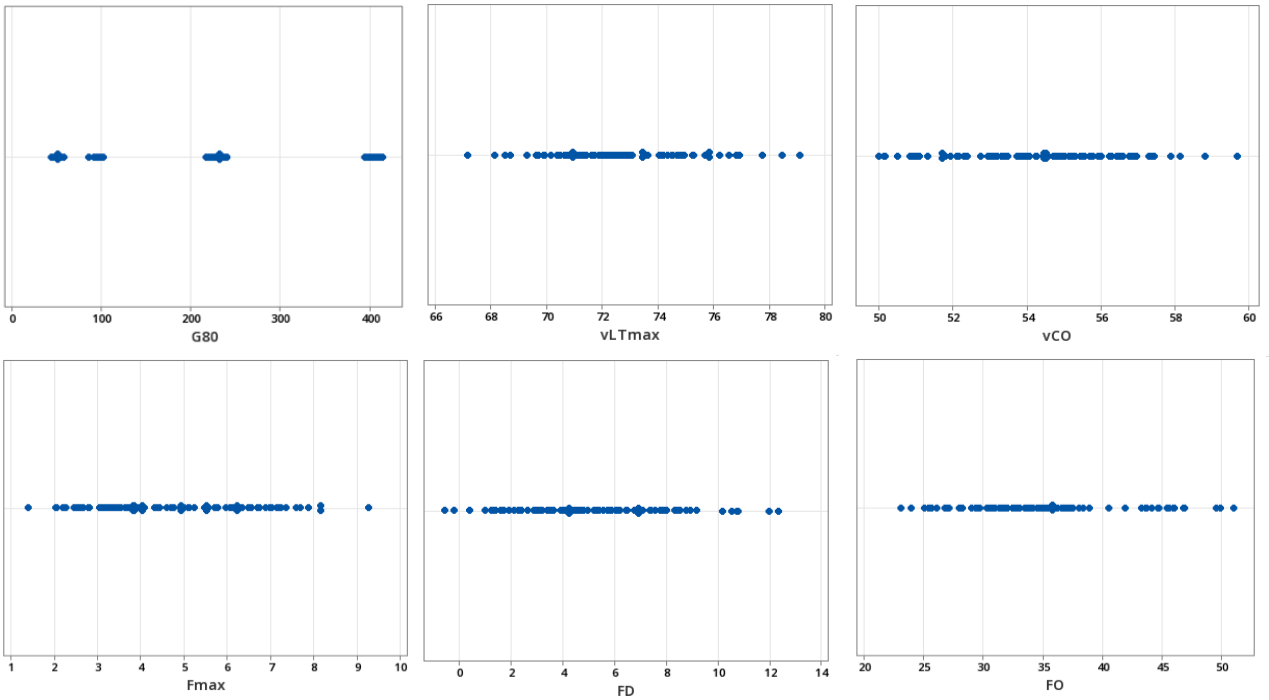


Figure A.2: Outlier Analysis- Thermophysical Variables

Variable	N	Mean	StDev	Variance	Minimum	Median	Maximum
G80	89	201	136.2	18560.9	43.2	220.4	413.8
vLTmax	89	72.693	2.391	5.718	67.15	72.46	79.08
vCO	89	54.224	2.177	4.737	49.97	54.5	59.69
Fmax	89	4.883	1.736	3.013	1.38	4.74	9.26
FD	89	5.316	2.759	7.611	-0.59	5.24	12.34
FO	89	35.16	6.724	45.216	23.03	34.55	51.02

Table A.1: Summary Statistics for the Continuous Thermophysical Variables in the Dataset

		Hard		Pasta Filata		Semi-Hard		Soft	
	N	**Mean**	**StDev**	**Mean**	**StDev**	**Mean**	**StDev**	**Mean**	**StDev**
G80	89	401.84	5.51	97.604	4.203	228.22	5.86	50.907	3.762
vLTmax	89	72.811	2.118	73.454	3.057	71.64	1.965	72.984	2.028
vCO	89	55.51	1.933	52.524	1.926	54.187	2.019	54.661	1.74
Fmax	89	6.619	1.313	4.504	1.399	4.995	1.185	3.168	1.083
FD	89	4.373	1.968	5.347	1.887	3.296	1.934	8.791	1.833
FO	89	35.533	2.003	27.401	2.24	33.242	1.838	45.569	2.566

Table A.2: Summary Statistics for the Thermophysical Variables in the Dataset by Texture

A.2- Statistical Analysis Figures and Tables

A.2.1- Manufacture vs. Thermophysical Characteristic Variables MANOVA

Criterion	Test Statistic	F	Num DF	Denom DF	P
Wilks'	0.94017	0.870	6	82	0.521
Lawley-Hotelling	0.06364	0.870	6	82	0.521
Pillai's	0.05983	0.870	6	82	0.521
Roy's	0.06364				

Table A.3: MANOVA Test Results for Thermophysical Characteristic Variables vs. Manufacturer

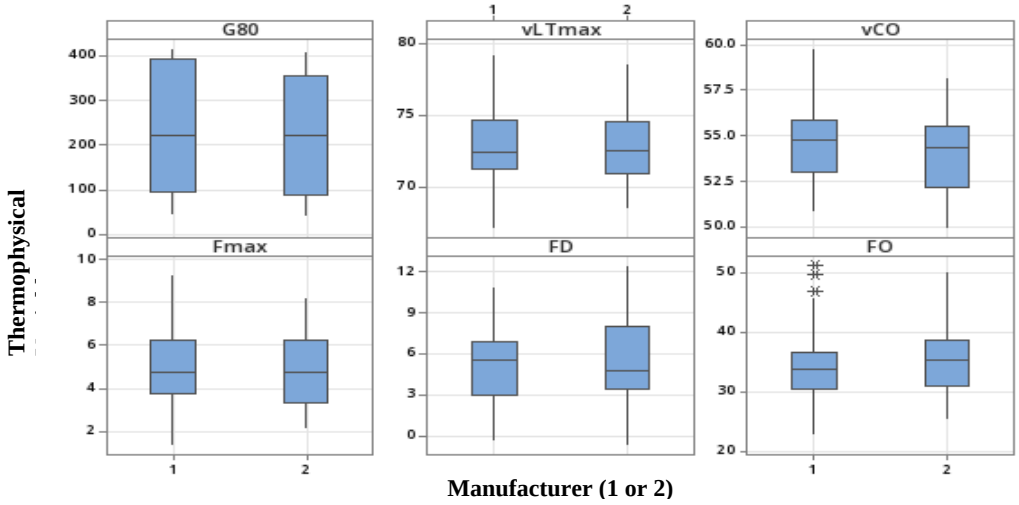


Figure A.3: Manufacturer vs. G80, vLTmax, vCO, Fmax, FD, FO: Analysis of common mean within Manufacturer

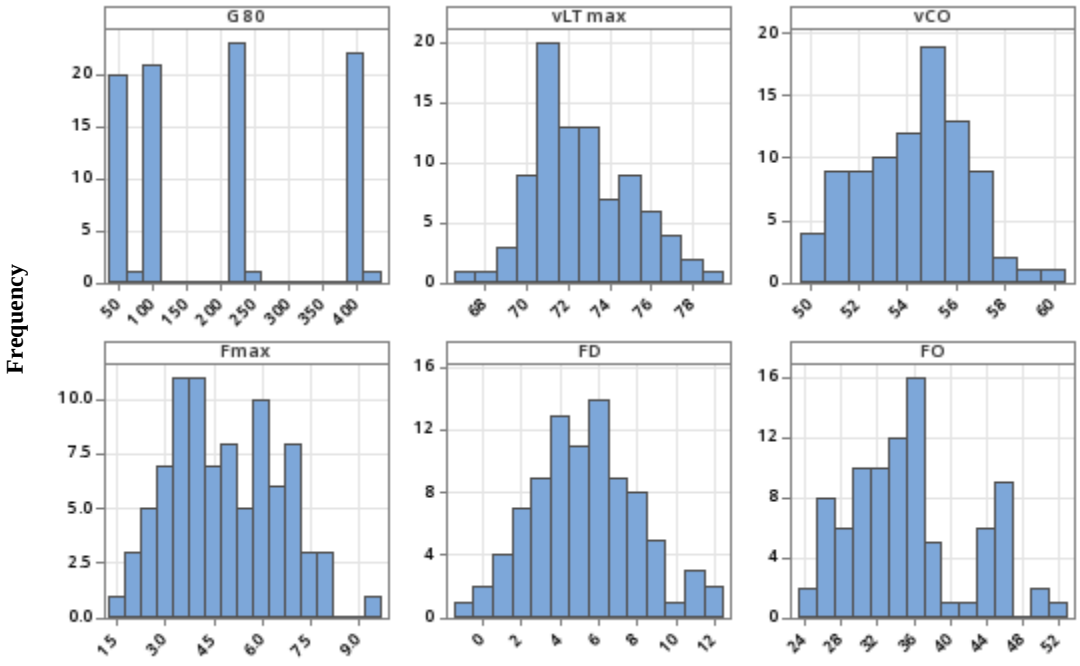


Figure A.4: Histogram of G80, vLTmax, vCO, Fmax, FD, FO: Analysis of normality of variables

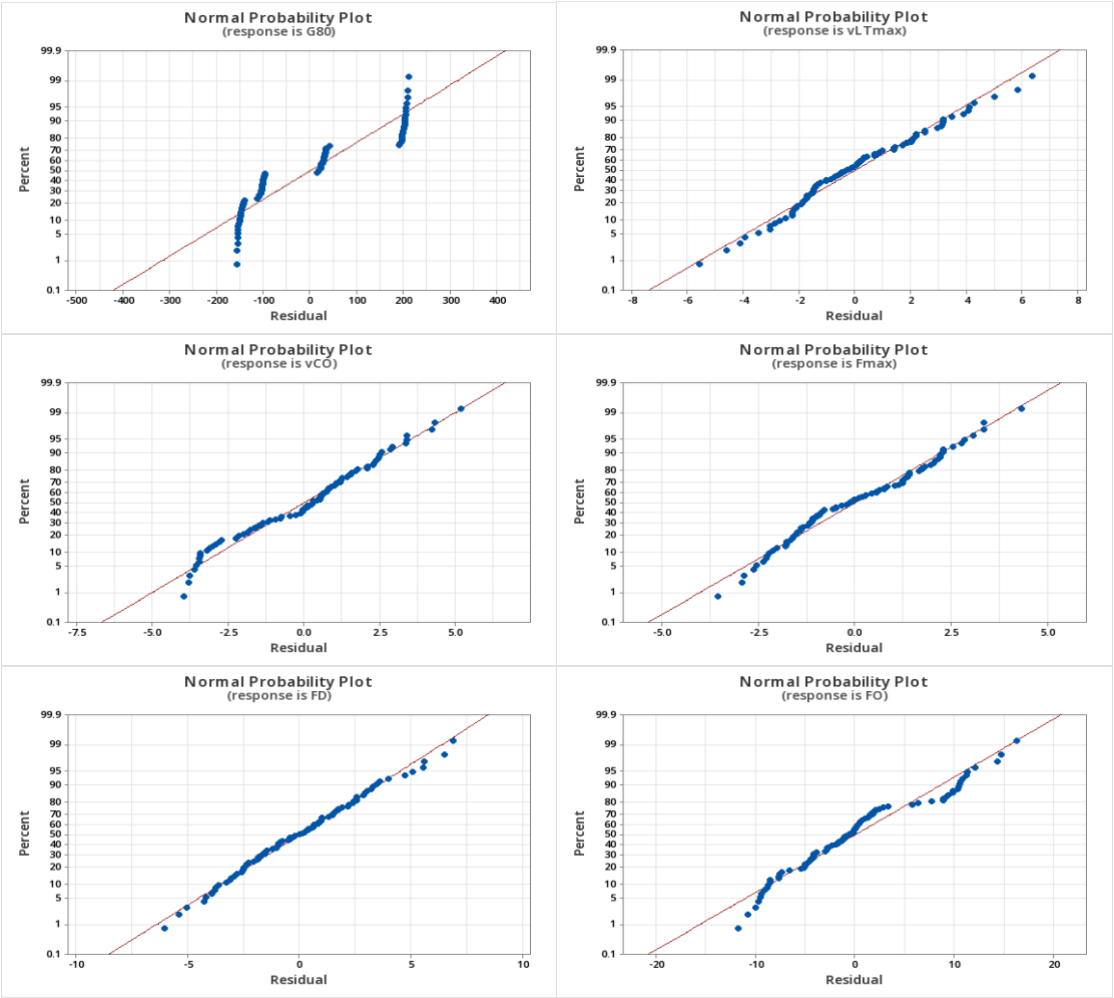


Figure A.5: Residual Normality Plots from Manufacturer vs. Thermophysical Characteristic Variables MANOVA

The MANOVA model for this initial question was found to meet all model assumptions. By design, the cheese samples are all independent of one another. Additionally, data from all manufacturers have a common variance-covariance matrix, which was assessed using Box's Test, resulting in a p-value= 0.6877. Data from each manufacturer appears to have a common mean vector, as in there is no subpopulation with inconsistencies within manufacturer (Figure A.3 Appendix A). Graphical representation of the thermophysical variables indicates most variables follow a normal distribution, but the G80 and FO variables may slightly deviate from normality (Figure A.4 Appendix A). Additionally, review of the residuals from this model arrives at this same conclusion by looking at the normal probability plots. While the residuals tend to be well behaved in these plots for vLTmax, vCO, Fmax, FD, and FO, there is some deviation from normality for the G80 and FO variables (Figure A.5 Appendix A). However, with a large sample size (n>30) we can assume the data are multivariate normally distributed and this assumption is met.

A.2.2- Texture vs. Thermophysical Characteristic Variables MANOVA

Criterion	Test Statistic	F	Num DF	Denom DF	P
Wilks'	0.00002	565.805	18	226	0.000
Lawley-Hotelling	893.36789	3904.349	18	236	0.000
Pillai's	2.26203	41.891	18	246	0.000
Roy's	852.01041				

Table A.4: MANOVA results for Thermophysical Characteristic Variables vs. Cheese Texture

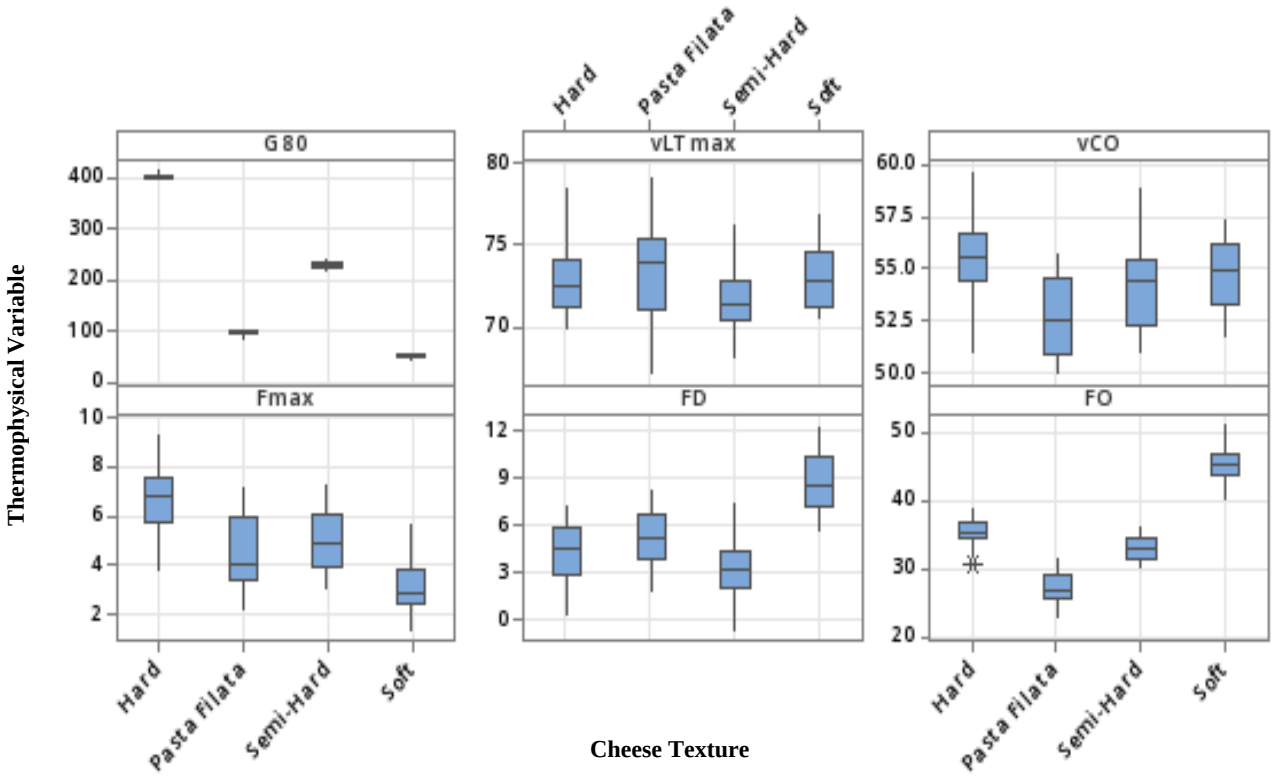


Figure A.6: Texture vs. G80, vLTmax, vCO, Fmax, FD, FO: Analysis of common mean within texture

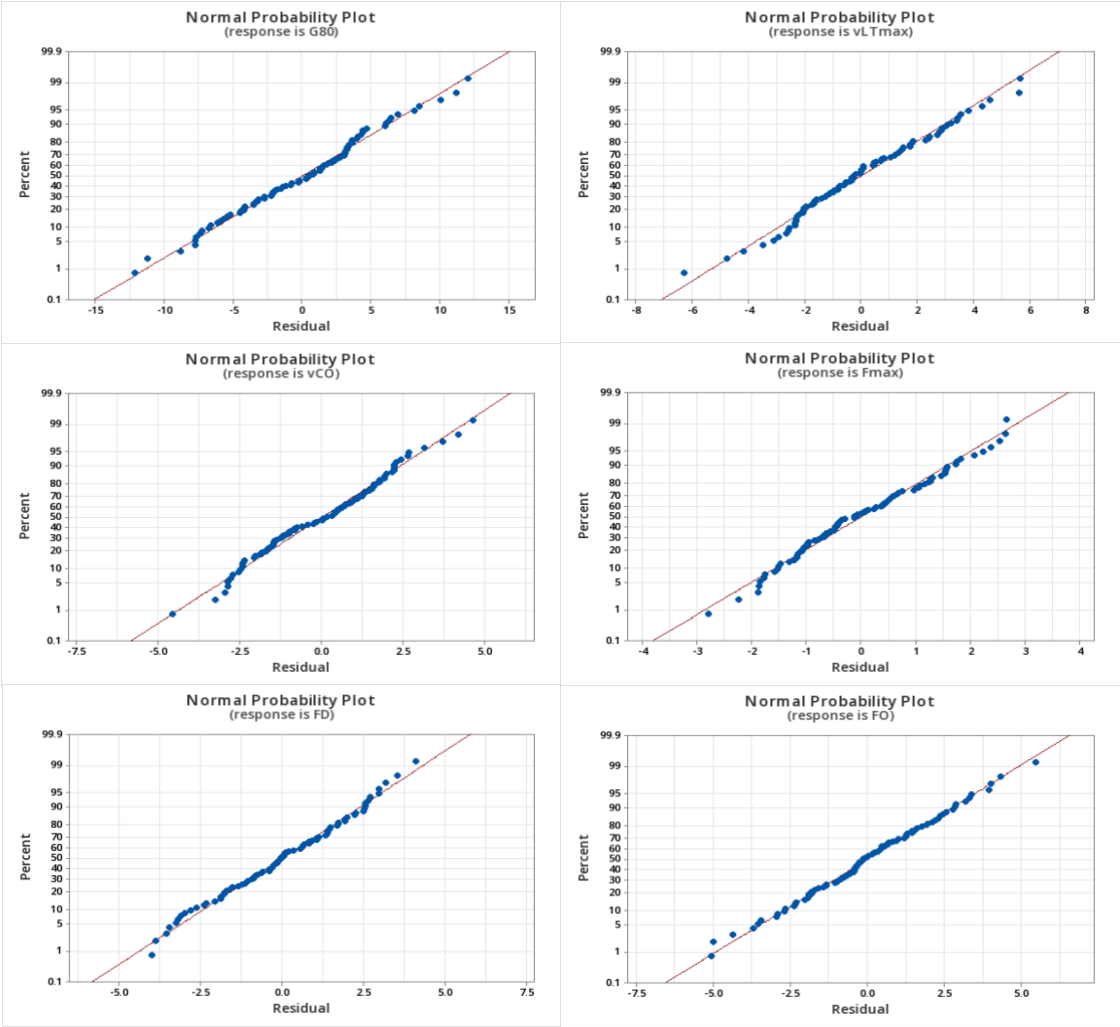


Figure A.7: Residual Normality Plots from Texture vs. Thermophysical Characteristic Variables MANOVA

This MANOVA model was also found to meet all model assumptions. By design, the cheese samples are all independent of one another. Additionally, data from all textures have a common variance-covariance matrix, which was assessed using Box's Test, resulting in a p-value=0.7890. The data from each texture appears to have a common mean vector, as in there is no subpopulation with inconsistencies within texture (Figure A.6 Appendix A). As found earlier, graphical representation of the thermophysical variables indicates most variables follow a normal distribution, but the G80 and FO variables may slightly deviate from normality (Figure A.4 Appendix A). However, review of the residual normal probability plots for each of the thermophysical characteristics indicates the data is relatively normal (Figure A.7 Appendix A). Additionally, with a large sample size (n>30) we can assume the data are multivariate normally distributed and this assumption is met.

A.2.3 Texture vs. Thermophysical Characteristic Variables ANOVA Tests

Variable vs. Texture	F	P-value
G80	22020.98	0
vLTmax	2.54	0.062
vCO	9.6	0
Fmax	27.9	0
FD	32.92	0
FO	256.03	0

Table A.5: Post-Hoc analysis: One way ANOVA Analyses

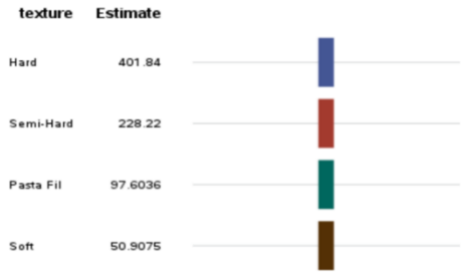


Figure A.8: Mean Texture Comparison of G80

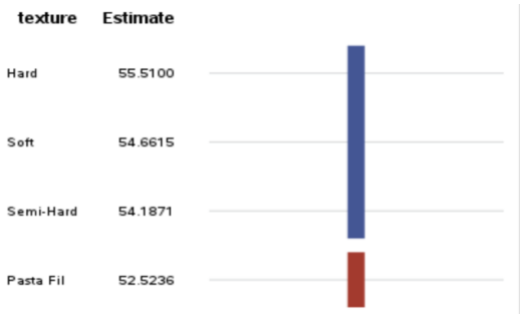


Figure A.11: Mean Texture Comparison of vCO

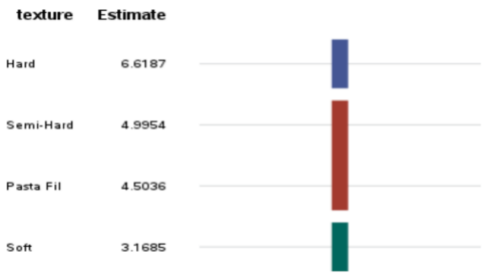


Figure A.9: Mean Texture Comparison of

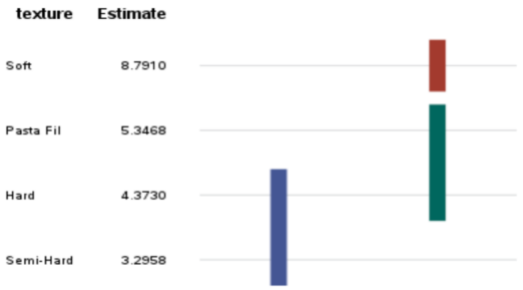


Figure A.12: Mean Texture Comparison of FD

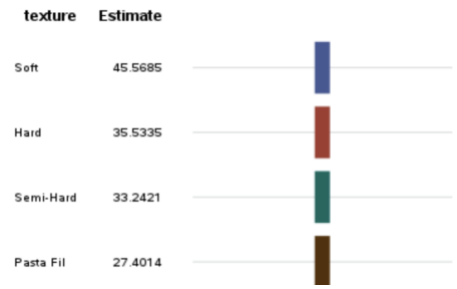


Figure A.10: Mean Texture Comparison of

Figures A.8-A.12 Post-hoc Mean Comparison Tests

If means are covered by the same bar, results are not significantly different

A.2.4 Cluster Analysis- Determining Cheese Varieties

To perform PCA, the variances of the six variables need to be similar. Since the variance for G80 is much larger than all other variables we should scale the six response variables so one variable does not impact the principal component analysis more than others.

	G80	vLTMax	vCO	Fmax	FD	FO
Variance	18,560.9	5.7	4.7	3	7.6	45.2

Table A.6 Variance of Thermophysical Properties of Cheese

PCA captures the maximum amount of variation from the six response variables in the minimal number of components. The scree plot below (Figure A.13) shows that the first Principal Component (PC) captures 37% of the variation in the data set, followed by PC2 at 33%, and the remaining four Principal Components.

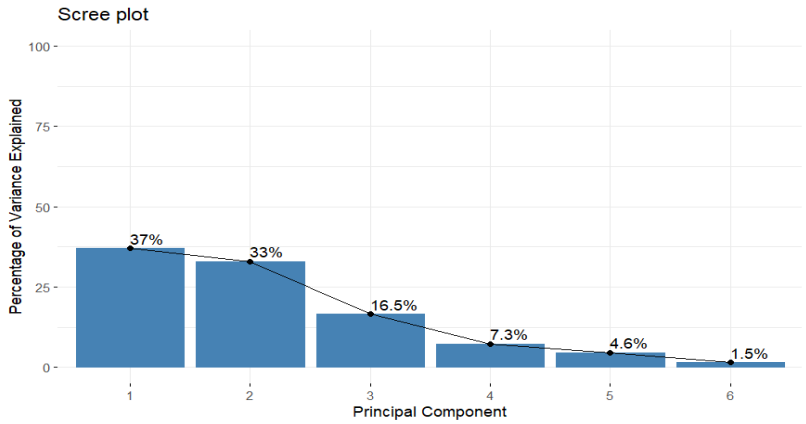


Figure A.13: Percent of variation explained by each PC

A table showing the variable loadings of each PC is available in Table A.14. Loadings give a measure of the importance of the variable within the principal component. Loadings close to 1 or -1 are strongest while those close to zero are not. This analysis focused only on the first two PCs, since they explain 70% of the variation in the cheese data set and will allow for easy visualization of the clustering.

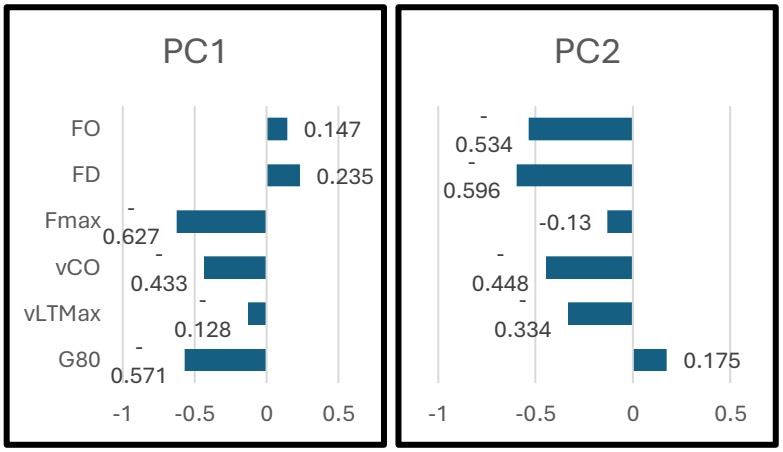


Figure A.14: Principal Component Loadings for Thermophysical Cheese Variables

Figure A.15 shows the reduction of the variation within each cluster as we increase the number of cheese varieties found in the dataset. In other words, as we increase the number cheese varieties in the data set, there are fewer samples in each group, and therefore less variation within each group (i.e. if we increased to 89 cluster (varieties), there would only be one sample in each variety grouping, and zero variability). Based on K-mean clustering the optimal number of cheese varieties occur at the bend in the graph, which appears when there are three or four varieties.

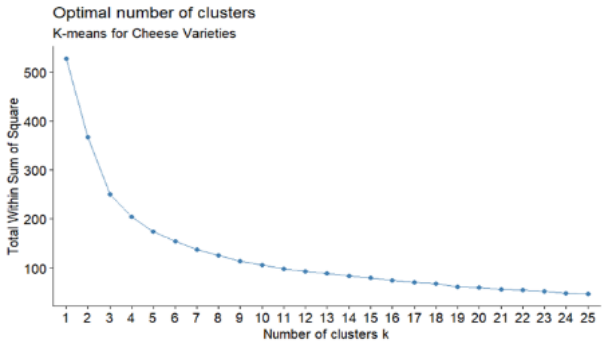


Figure A.15: Cheese Varieties vs. Within Cluster Variation

Figure A.16 below provides the scatterplots, single variable distributions, and correlation coefficients for the six thermophysical properties of cheese.

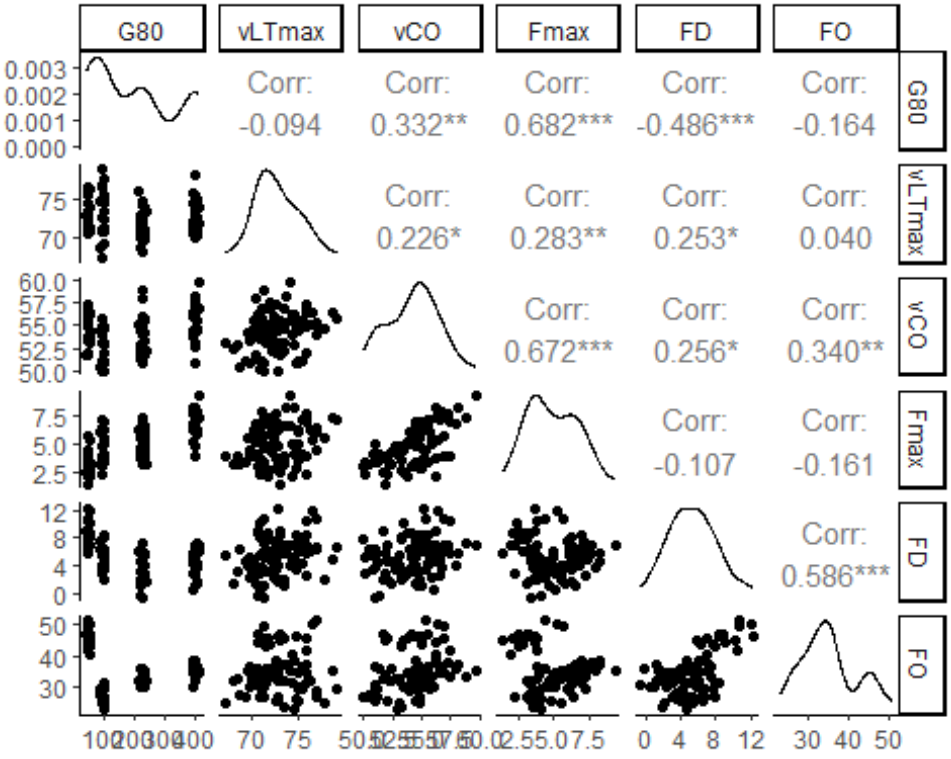


Figure A.16: Scatterplots, Distributions and Correlation Coefficient of Dataset

Figure A.17 below plots the k=4 means clustering for the cheese data set. When the data set is clustered into four cheese varieties, the groups are not split as neatly. Cluster one overlaps with clusters two and four. The addition of the fourth variety does decrease the total Sum of Squares variation within each group from 250 down to 205. If the overlapping of varieties is not an issue, and the decreased variation within groups is more important, then four cheese varieties is also an acceptable number of clusters.

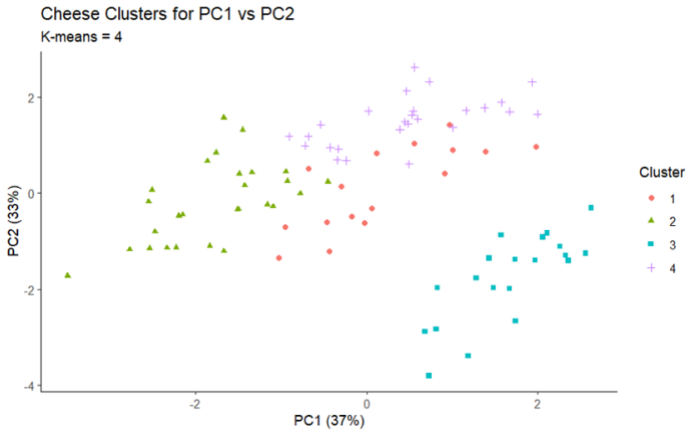


Figure A.17: Cluster Analysis with K=4

Figure A.18 below shows k=5 means clustering for the cheese data set. When the data set is clustered into five cheese varieties, there is more overlap between varieties. Cluster two overlaps with clusters three and four, while clusters three and five also overlap. The addition of the fourth variety does decrease the total Sum of Squares variation within each group from 205 down 174, but the decrease may not be worth the added confusion in the clustering.

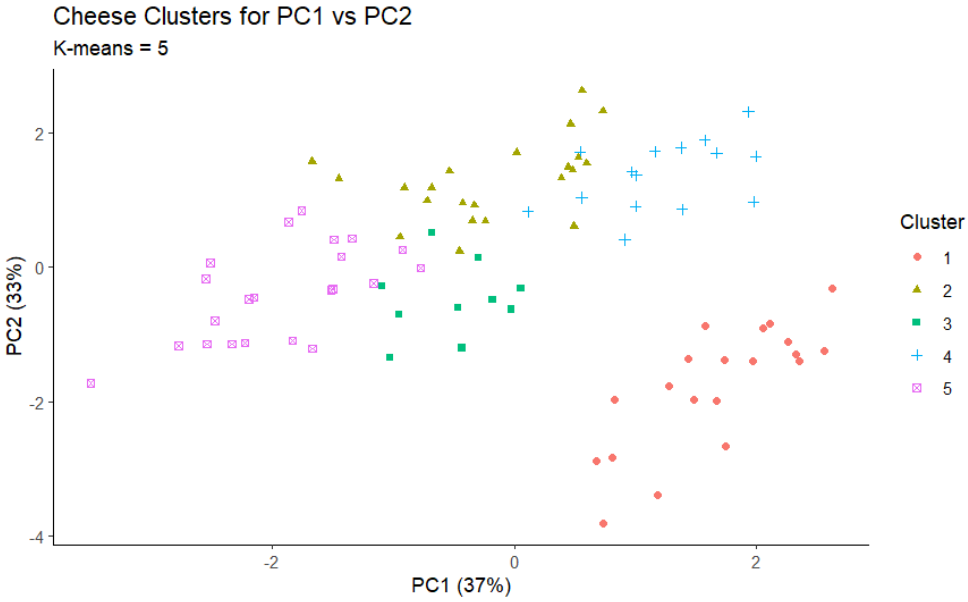


Figure A.18: Cluster Analysis with K=5

Figure A.19 below shows k=6 means clustering for the cheese data set. When the data set is clustered into six cheese varieties, there is still overlap between varieties. Cluster two overlaps with cluster four, while clusters five and six also overlap. The addition of the sixth variety does decrease the total Sum of Squares variation within each group from 174 to 155.

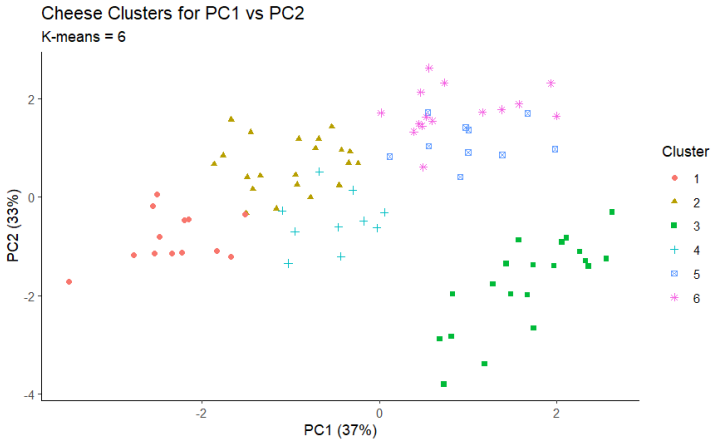


Figure A.19: Cluster Analysis with K=6

Table A.7 provides a number summary of each cluster with respect to the six thermophysical properties of the data set.

Cluster	G80		vLTMax		vCO		Fmax		FD		FO	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	50.9	3.8	73.0	2.0	54.7	1.7	3.2	1.1	8.8	1.8	45.6	2.6
2	189.2	100.3	71.6	2.1	52.2	1.5	4.0	0.8	3.6	1.9	30.3	3.8
3	295.1	123.9	73.6	2.5	55.8	1.4	6.6	0.9	5.0	2.1	33.9	3.2

Table A.7: Mean and Standard Deviation for Each Variable by Cluster

A.2.5 Discriminant Analysis

	True Group			
Put Into Group	Hard	Pasta Filata	Semi-Hard	Soft
Hard	23	0	0	0
Pasta Filata	0	22	0	0
Semi-Hard	0	0	24	0
Soft	0	0	0	20
Total N	23	22	24	20
N Correct	23	22	24	20
Proportion	1.000	1.000	1.000	1.000

Table A.8: Classification Model Resubstitution Results

Appendix B- Supporting Code

SAS Box Tests to support the MANOVA tests

```
data cheese;
    infile "Location_of_Dataset" firstobs=2 delimiter=',';
    input ID manufacturer texture $ G80 vLTmax vCO Fmax FD FO;
run;

*Box Test Manufacture Variable;
proc discrim data= cheese pool=test;
  class manufacturer;
  var G80 vltmax vco fmax FD FO;
  run;

*Box Test Texture Variable;
proc discrim data= cheese pool=test;
  class texture;
  var G80 vltmax vco fmax FD FO;
  run;
```

SAS Bartlett Code

```
data cheese;
    infile "Location_of_Dataset" firstobs=2 delimiter=',';
    input ID manufacturer texture $ G80 vLTmax vCO Fmax FD FO;
run;

proc discrim data=cheese pool=test crossvalidate;
    class texture;
    var G80 vLTmax vCO Fmax FD FO;
run;
```

SAS returned a chi-square value of 53.802 with an associated p-value of 0.789 using the given data set

R-Markdown Clustering and Principal Component Analysis

Project 1 Cluster Analysis

Group 6

February 22, 2024

Front Matter

```r
#Load Libraries
library(tidyverse)

library(mvtnorm) #used for randomly generating data from multivariate normal
library(factoextra)

library(gridExtra)
```

```
library(broom)
library(GGally)

#Load Data Set
cheese <- read.csv("C:/Users/chris/OneDrive/Desktop/PSU/STAT580/Project1/cheeseThermophysical.csv")
```

Preparing for Clustering

Create a dataset with numeric variables and standardize

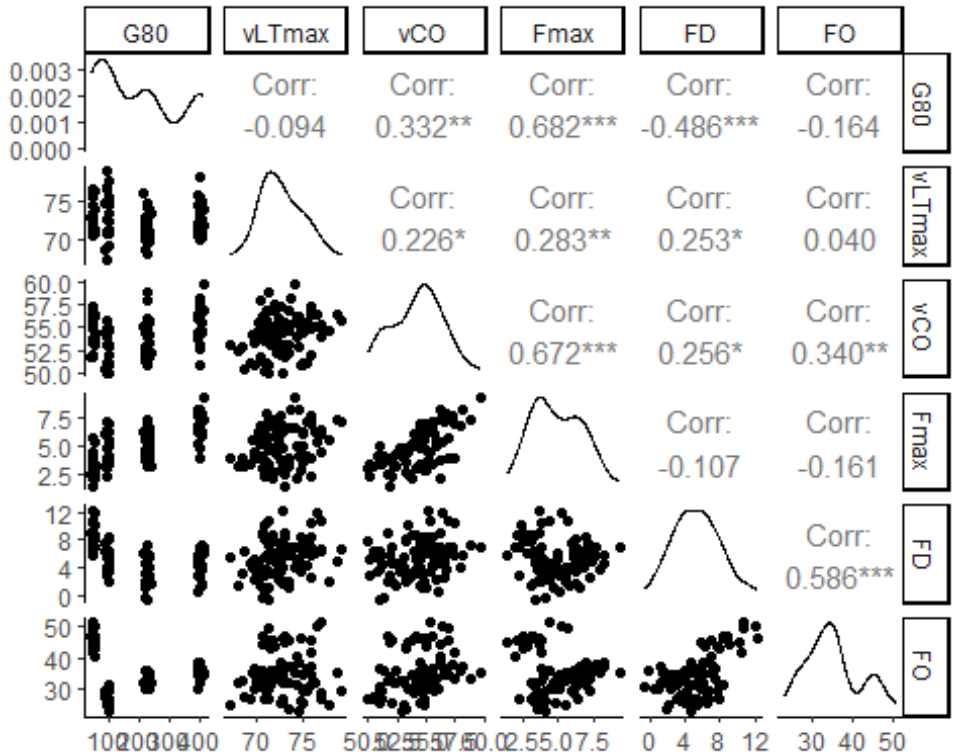
```
cheese_matrix <- cheese %>%
      select(-texture, -ID, -manufacturer)

#Check Covariance - to see if scaling is needed
cov(cheese_matrix)
```

```
##                G80      vLTmax       vCO        Fmax            FD           FO
## G80    18560.88078 -30.6641116 98.557246 161.2403559 -182.7950130 -150.2712544
## vLTmax   -30.66411   5.7181604  1.178219   1.1765865    1.6658227    0.6494065
## vCO       98.55725   1.1782190  4.737375   2.5397046    1.5382992    4.9771776
## Fmax     161.24036   1.1765865  2.539705   3.0129727   -0.5106638   -1.8822138
## FD      -182.79501   1.6658227  1.538299  -0.5106638    7.6105378   10.8703575
## FO      -150.27125   0.6494065  4.977178  -1.8822138   10.8703575   45.2159725
```

```
cheeseScaled<- scale(x=cheese_matrix, center=TRUE, scale=TRUE)
```

Since the variance for G80 is much larger than all other variables, we should standardize the six response variables so one variable (G80) does not impact the clustering more than others.

```
#Scatterplots, distributions, and correlations
ggpairs(data= cheese_matrix,progress=FALSE)
```



The distributions for each cheese texture type across all 6 variables are seen here. Amongst each variable, the texture types have similar spreads, but different centers.

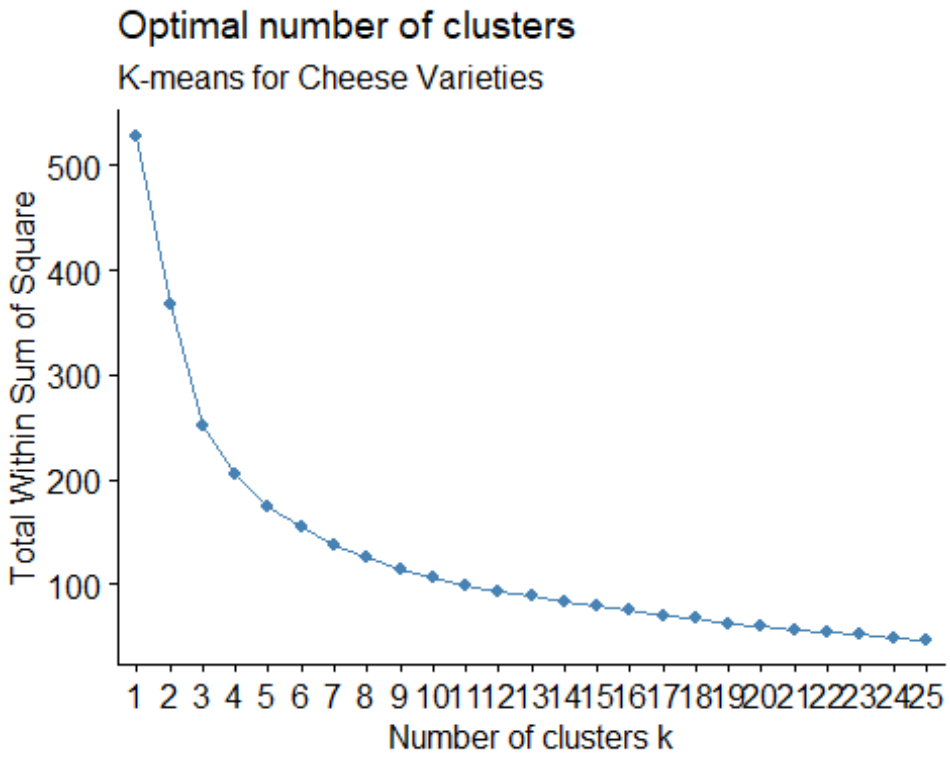
```
cheeseText <- cheese %>%
     group_by(texture) %>%
      summarize(meantG80 = mean(G80),
                sdG80 = sd(G80),
                meantvLTmax = mean(vLTmax),
                sdvLTmax = sd(vLTmax),
                meantvCO = mean(vCO),
                sdvCO = sd(vCO),
                meantFmax = mean(Fmax),
                sdFmax = sd(Fmax),
                meantFD = mean(FD),
                sdFD = sd(FD),
                meantFO = mean(FO),
                sdFO = sd(FO))

cheeseText
```

```
## # A tibble: 4 × 13
##   texture    meantG80 sdG80 meantvLTmax sdvLTmax meantvCO sdvCO meantFmax sdFmax
##   <chr>         <dbl> <dbl>       <dbl>    <dbl>    <dbl> <dbl>     <dbl>  <dbl>
## 1 Hard          402.   5.51        72.8     2.12     55.5  1.93      6.62   1.31
## 2 Pasta Fil…     97.6  4.20        73.5     3.06     52.5  1.93      4.50   1.40
## 3 Semi-Hard     228.   5.86        71.6     1.97     54.2  2.02      5.00   1.18
## 4 Soft           50.9  3.76        73.0     2.03     54.7  1.74      3.17   1.08
## # ℹ 4 more variables: meantFD <dbl>, sdFD <dbl>, meantFO <dbl>, sdFO <dbl>
```

Performing K-means

Create elbow plot for K-means

```
set.seed (1)
fviz_nbclust(x=cheeseScaled,
             FUNcluster = kmeans,
             method = "wss",
             k.max = 25,
             nstart=25,
             iter.max = 15)+
          labs(subtitle = "K-means for Cheese Varieties")
```




```
set.seed(NULL)
```

How many clusters?

It is not completely obvious how many clusters should be used. I would say the drop from K = 1 to K = 2 is similar to the drop from K = 2 to K = 3, so at least three clusters. After 3, the drop in WSS does decrease slightly from K = 3 to K = 4, and even less from K=4 to K=5. It would be good to investigate K=3 to K=6 clusters to see how the cheese data behaves.

Perform K-means with K = 3-6 and extract Total Within Cluster Variation

```
set.seed(1)

KmeansRes3c <- kmeans(x=cheeseScaled,centers=3,
                      iter.max=25, nstart=25)

KmeansRes4c <- kmeans(x=cheeseScaled,centers=4,
                      iter.max=25, nstart=25)

KmeansRes5c <- kmeans(x=cheeseScaled,centers=5,
                      iter.max=25, nstart=25)

KmeansRes6c <- kmeans(x=cheeseScaled,centers=6,
                      iter.max=25, nstart=25)
set.seed(NULL)

KmeansRes3c$tot.withinss

## [1] 250.4097

KmeansRes4c$tot.withinss
```

```
## [1] 204.9502

KmeansRes5c$tot.withinss

## [1] 174.1803

KmeansRes6c$tot.withinss

## [1] 154.7098
```

How many observations were assigned to each cluster?

```
table(KmeansRes3c$cluster)

##
##  1  2  3
## 20 33 36

table(KmeansRes4c$cluster)

##
##  1  2  3  4
## 16 27 20 26

table(KmeansRes5c$cluster)

##
##  1  2  3  4  5
## 20 22 10 15 22

table(KmeansRes6c$cluster)

##
##  1  2  3  4  5  6
## 13 21 20 10 10 15
```

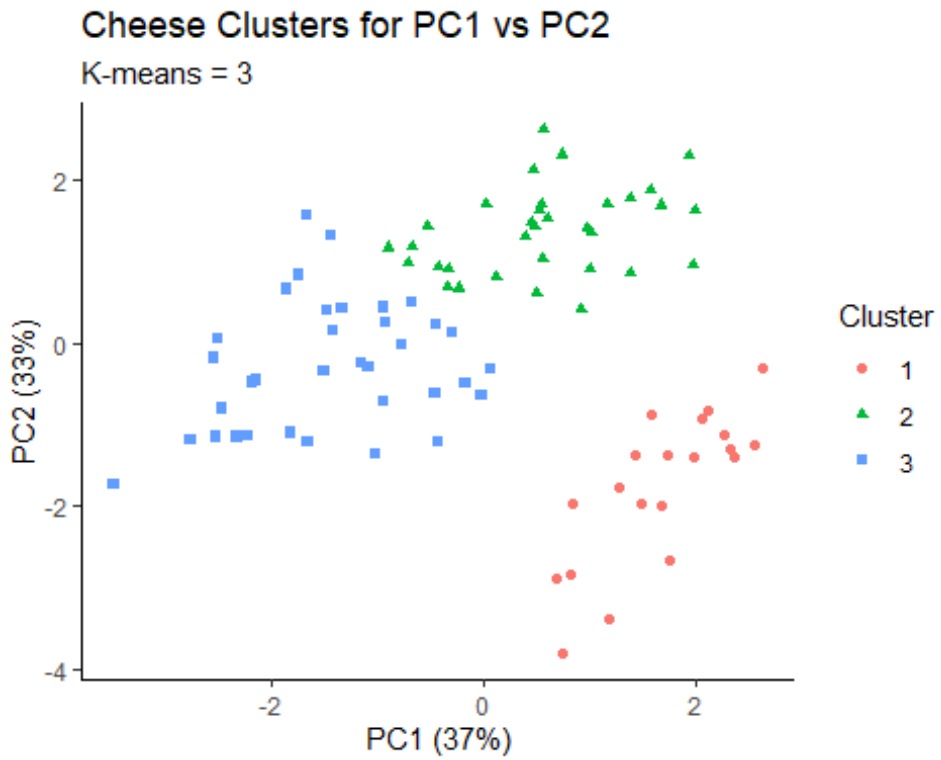
The cheeses are spread fairly well across clusters.

Since there are 6 variables it may be fairly difficult to visualize how the cheeses are clustered. It may be best to reduce the dimensionality of the data set by performing a principal component analysis (PCA) which captures as much of the variance as possible. The first PC captures the most variation in the data set, followed by PC 2, and so on.

Perform PCA and create visualization for K = 3 (K-means)

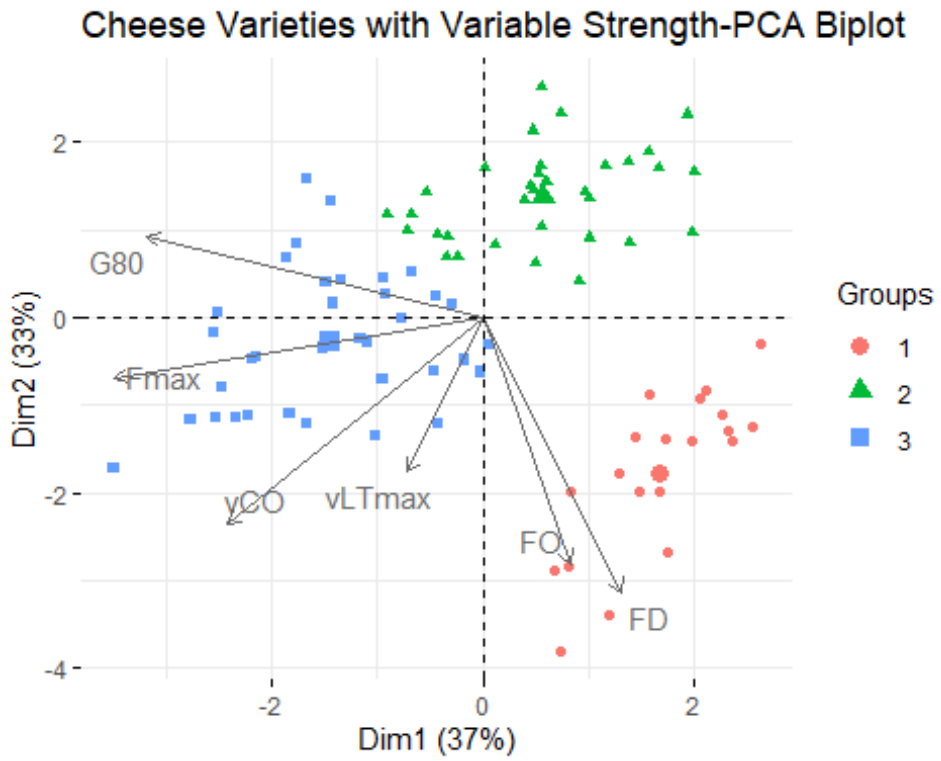
```
cheesePCA<- prcomp(x=cheeseScaled, center=TRUE, scale=TRUE)
#individual eigenvalues over sum of eigenvalues gives PVE
PVE<-(cheesePCA$sdev^2)/sum(cheesePCA$sdev^2)
round(PVE, digits=3)

## [1] 0.370 0.330 0.165 0.073 0.046 0.015

#Create a data frame (for use in ggplot) that has PC scores for 1st 2 PC's
temp_df <- as.data.frame(x=cheesePCA$x[ ,1:2])
# #Add cluster to the plotting dataset (temp_df)
temp_df<- temp_df %>%
  mutate(cluster=KmeansRes3c$cluster)

#Create Two-Way Tables for Cheese Texture and cluster, and Manufacturer and cluster.
temp_df <- temp_df %>%
      mutate(texture = cheese$texture,
```

```
            maker = cheese$manufacturer)

table(temp_df$maker, temp_df$cluster)

##
##      1  2  3
##   1 10 15 20
##   2 10 18 16

table(temp_df$texture, temp_df$cluster)

##
##                  1  2  3
##   Hard           0  4 19
##   Pasta Filata   0 15  7
##   Semi-Hard      0 14 10
##   Soft          20  0  0

#Create Plot

ggplot(data=temp_df, mapping = aes(x=PC1, y=PC2,
       color= as.factor(cluster), shape= as.factor(cluster)))+
     geom_point()+
    labs(x = paste("PC1 (", round(100*PVE[1], digits = 1), "%)", sep = ""),
    y = paste("PC2 (", round(100*PVE[2], digits = 1), "%)", sep = ""),
    color = "Cluster", shape = "Cluster", title="Cheese Clusters for PC1 vs PC2",
    subtitle = "K-means = 3")
```




```
fviz_pca_biplot(cheesePCA, repel = TRUE,
                habillage = KmeansRes3c$cluster,
                col.var = "#696969", # Variables color
```

```
                label="var",
                col.ind = "#696969", # Individuals color
                title="Cheese Varieties with Variable Strength-PCA Biplot"
                )
```




```
#fviz_pca_var(cheesePCA, repel = TRUE,
                #col.var = "#2E9FDF", # Variables color
                #col.ind = "#696969"  # Individuals color
                #)
```

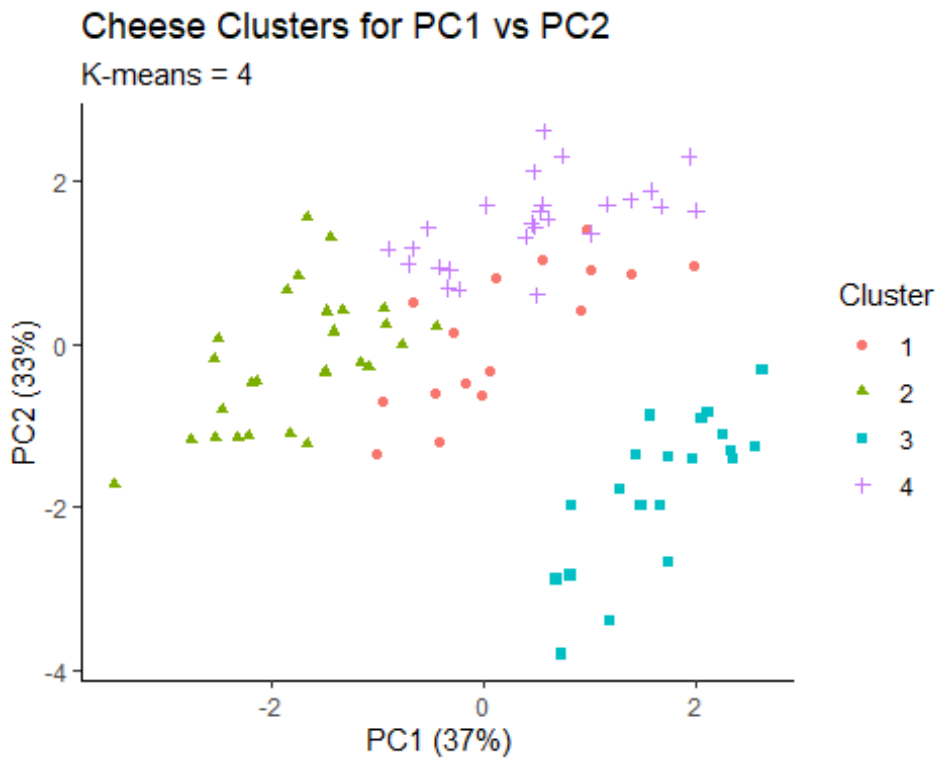
To see how well the clusters do in classifying “types” of cheese, we can plot the cheeses as they are clustered against the first two principal components. Using 3 clusters for k-means clustering produces 3 distinct groups of clusters with no overlap. Cluster 1 generally has a positive PC1 value with a negative PC2 value. Cluster 2 is generally positive for both PC1 and PC2. Cluster 3 is negative for PC1 and dispersed between 2 and -2 for PC2.

Cluster 1 contains only soft cheese (and soft cheese is only in cluster 1) Cluster 2 is composed mostly of Pasta Filata and Semi-hard cheese (with 4 hard cheeses), while cluster 3 is composed mostly of Hard cheese with some Semi-Hard and Pasta Filata cheeses.

Perform PCA and create visualization for K = 4 (K-means)

```
cheesePCA<- prcomp(x=cheeseScaled, center=TRUE, scale=TRUE)
#individual eigenvalues over sum of eigenvalues gives PVE
PVE<-(cheesePCA$sdev^2)/sum(cheesePCA$sdev^2)
round(PVE, digits=3)

## [1] 0.370 0.330 0.165 0.073 0.046 0.015

#Create a data frame (for use in ggplot) that has PC scores for 1st 2 PC's
temp_df <- as.data.frame(x=cheesePCA$x[ ,1:2])
# #Add cluster to the plotting dataset (temp_df)
```

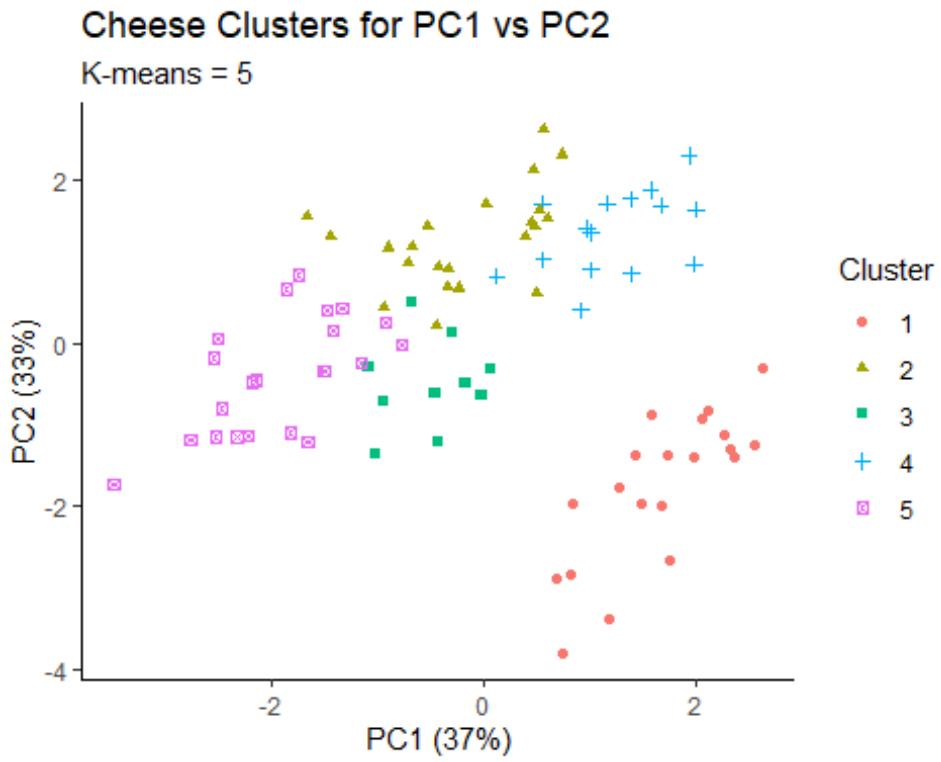
```
temp_df<- temp_df %>%
  mutate(cluster=KmeansRes4c$cluster)

#Create Two-Way Tables for Cheese Texture and cluster, and Manufacturer and cluster.
temp_df <- temp_df %>%
      mutate(texture = cheese$texture,
             maker = cheese$manufacturer)

table(temp_df$maker, temp_df$cluster)

##
##      1  2  3  4
##   1  9 15 10 11
##   2  7 12 10 15

table(temp_df$texture, temp_df$cluster)

##
##                 1  2  3  4
##   Hard          0 19  0  4
##   Pasta Filata 14  0  0  8
##   Semi-Hard     2  8  0 14
##   Soft          0  0 20  0

#Create Plot

ggplot(data=temp_df, mapping = aes(x=PC1, y=PC2,
       color= as.factor(cluster), shape= as.factor(cluster)))+
     geom_point()+
    labs(x = paste("PC1 (", round(100*PVE[1], digits = 1), "%)", sep = ""),
    y = paste("PC2 (", round(100*PVE[2], digits = 1), "%)", sep = ""),
    color = "Cluster", shape = "Cluster", title="Cheese Clusters for PC1 vs PC2",
    subtitle = "K-means = 4")
```



Perform PCA and create visualization for K = 5 (K-means)

```
cheesePCA<- prcomp(x=cheeseScaled, center=TRUE, scale=TRUE)
#individual eigenvalues over sum of eigenvalues gives PVE
PVE<-(cheesePCA$sdev^2)/sum(cheesePCA$sdev^2)
round(PVE, digits=3)

## [1] 0.370 0.330 0.165 0.073 0.046 0.015

#Create a data frame (for use in ggplot) that has PC scores for 1st 2 PC's
temp_df <- as.data.frame(x=cheesePCA$x[ ,1:2])
# #Add cluster to the plotting dataset (temp_df)
temp_df<- temp_df %>%
  mutate(cluster=KmeansRes5c$cluster)

#Create Two-Way Tables for Cheese Texture and cluster, and Manufacturer and cluster.
temp_df <- temp_df %>%
      mutate(texture = cheese$texture,
             maker = cheese$manufacturer)

table(temp_df$maker, temp_df$cluster)

##
##      1  2  3  4  5
##   1 10 11  5  7 12
##   2 10 11  5  8 10

table(temp_df$texture, temp_df$cluster)

##
##                1  2  3  4  5
##   Hard         0  6  0  0 17
```

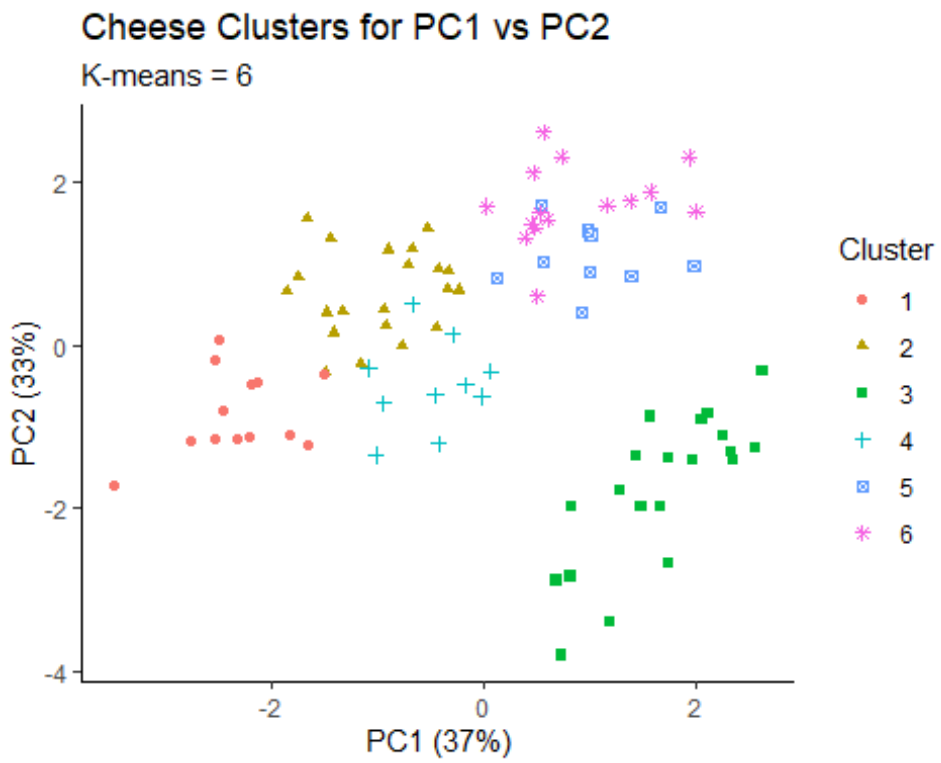
```
##   Pasta Filata  0  0  7 15  0
##   Semi-Hard     0 16  3  0  5
##   Soft         20  0  0  0  0

#Create Plot

ggplot(data=temp_df, mapping = aes(x=PC1, y=PC2,
       color= as.factor(cluster), shape= as.factor(cluster)))+
     geom_point()+
    labs(x = paste("PC1 (", round(100*PVE[1], digits = 1), "%)", sep = ""),
    y = paste("PC2 (", round(100*PVE[2], digits = 1), "%)", sep = ""),
    color = "Cluster", shape = "Cluster", title="Cheese Clusters for PC1 vs PC2",
    subtitle = "K-means = 5")
```



Perform PCA and create visualization for K = 6 (K-means)

```
#cheesePCA<- prcomp(x=cheeseScaled, center=TRUE, scale=TRUE)
#individual eigenvalues over sum of eigenvalues gives PVE
PVE<-(cheesePCA$sdev^2)/sum(cheesePCA$sdev^2)
round(PVE, digits=3)

## [1] 0.370 0.330 0.165 0.073 0.046 0.015

#Create a data frame (for use in ggplot) that has PC scores for 1st 2 PC's
temp_df <- as.data.frame(x=cheesePCA$x[ ,1:2])
# #Add cluster to the plotting dataset (temp_df)
temp_df<- temp_df %>%
  mutate(cluster=KmeansRes6c$cluster)

#Create Two-Way Tables for Cheese Texture and cluster, and Manufacturer and cluster.
temp_df <- temp_df %>%
      mutate(texture = cheese$texture,
```

```
            maker = cheese$manufacturer)

table(temp_df$maker, temp_df$cluster)

##
##      1  2  3  4  5  6
##   1  6 12 10  5  5  7
##   2  7  9 10  5  5  8

table(temp_df$texture, temp_df$cluster)

##
##                  1  2  3  4  5  6
##   Hard          12 10  0  0  0  1
##   Pasta Filata   0  0  0  7 10  5
##   Semi-Hard      1 11  0  3  0  9
##   Soft           0  0 20  0  0  0

#Create Plot

ggplot(data=temp_df, mapping = aes(x=PC1, y=PC2,
       color= as.factor(cluster), shape= as.factor(cluster)))+
     geom_point()+
    labs(x = paste("PC1 (", round(100*PVE[1], digits = 1), "%)", sep = ""),
    y = paste("PC2 (", round(100*PVE[2], digits = 1), "%)", sep = ""),
    color = "Cluster", shape = "Cluster", title="Cheese Clusters for PC1 vs PC2",
    subtitle = "K-means = 6")
```



Create a two-way table and discuss

```
temp_df <- temp_df %>%
      mutate(texture = cheese$texture,
```

```
             maker = cheese$manufacturer)

table(temp_df$maker, temp_df$cluster)

##
##      1  2  3  4  5  6
##   1  6 12 10  5  5  7
##   2  7  9 10  5  5  8

table(temp_df$texture, temp_df$cluster)

##
##                 1  2  3  4  5  6
##   Hard         12 10  0  0  0  1
##   Pasta Filata  0  0  0  7 10  5
##   Semi-Hard     1 11  0  3  0  9
##   Soft          0  0 20  0  0  0
```

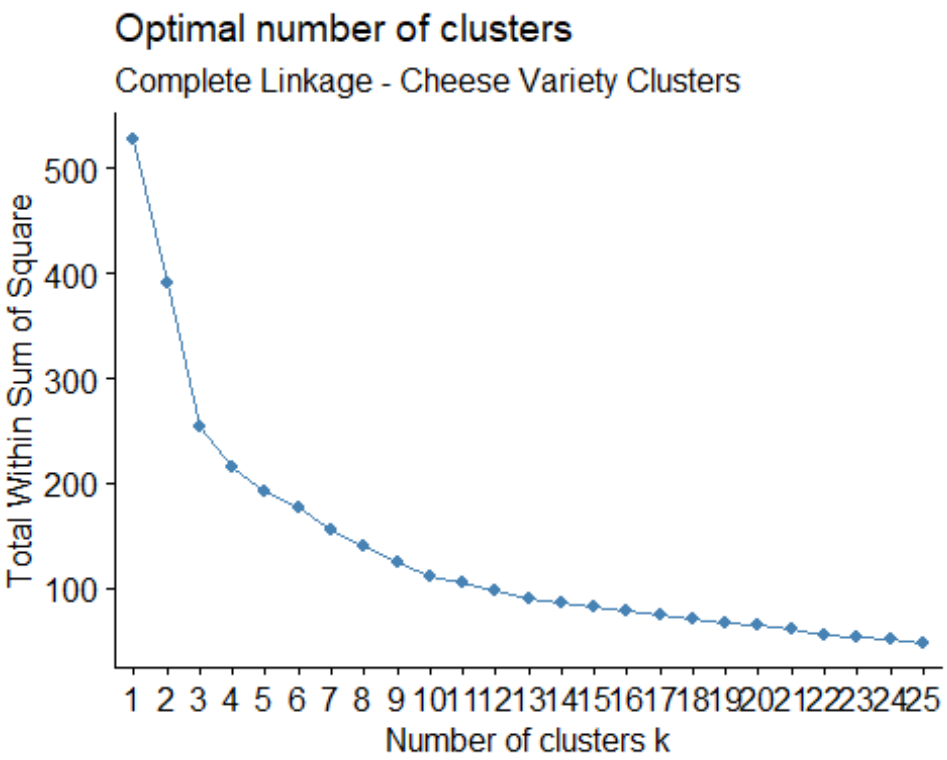
Find centroids

```
cheeseCentroids <- cheese %>%
      mutate( cluster = KmeansRes3c$cluster) %>%
        group_by(cluster) %>%
      summarize(meanG80 = mean(G80),
                meanvLTmax = mean(vLTmax),
                meanvCO = mean(vCO),
                meanFmax = mean(Fmax),
                meanFD = mean(FD),
                meanFO = mean(FO))


cheeseCentroids

## # A tibble: 3 × 7
##   cluster meanG80 meanvLTmax meanvCO meanFmax meanFD meanFO
##     <int>   <dbl>      <dbl>   <dbl>    <dbl>  <dbl>  <dbl>
## 1       1    50.9       73.0    54.7     3.17   8.79   45.6
## 2       2   189.        71.6    52.2     4.02   3.61   30.3
## 3       3   295.        73.6    55.8     6.63   4.95   33.9

cheeseSpread <- cheese %>%
      mutate( cluster = KmeansRes3c$cluster) %>%
        group_by(cluster) %>%
      summarize(sdG80 = sd(G80),
                sdvLTmax = sd(vLTmax),
                sdvCO = sd(vCO),
                sdFmax = sd(Fmax),
                sdFD = sd(FD),
                sdFO = sd(FO))
cheeseSpread

## # A tibble: 3 × 7
##   cluster  sdG80 sdvLTmax sdvCO sdFmax  sdFD  sdFO
##     <int>  <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1       1   3.76     2.03  1.74  1.08   1.83  2.57
## 2       2 100.       2.10  1.46  0.757  1.88  3.82
## 3       3 124.       2.46  1.38  0.944  2.07  3.25
```

Performing Hierarchical Clustering

Perform hierarchical clustering using Euclidean distance and complete linkage

```
fviz_nbclust(x = cheeseScaled, FUNcluster = hcut,
             method = "wss",
             k.max = 25,
             hc_func = "hclust",
             hc_metric = "euclidean",
             hc_method = "complete") +
  labs(subtitle = "Complete Linkage - Cheese Variety Clusters")
```



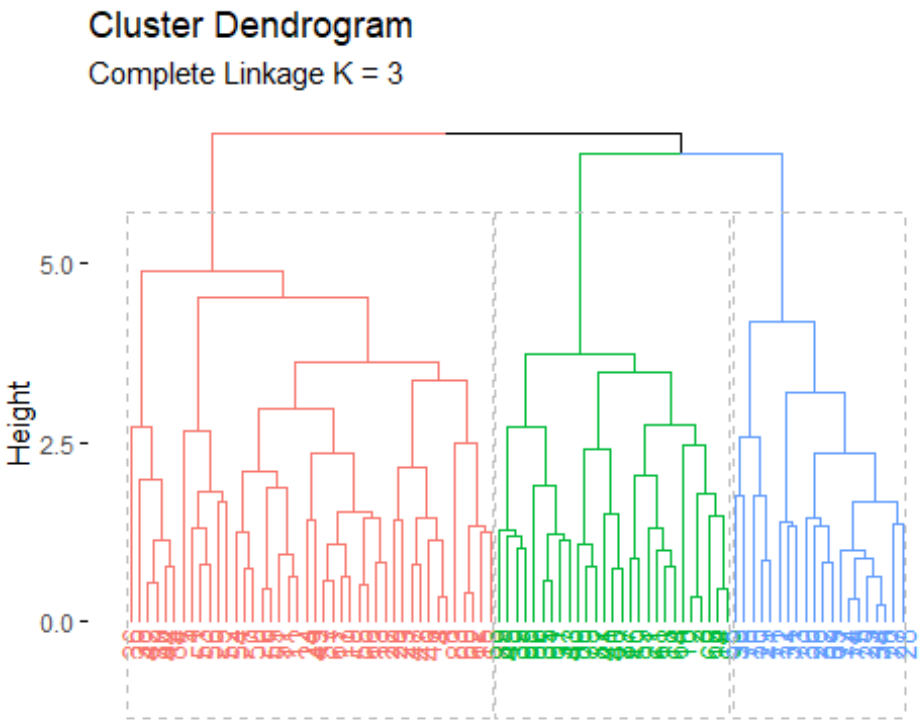
How many clusters?

This is similar to what we saw with the k-means clustering. It is still not completely obvious how many clusters should be used. I would say the drop from K = 1 to K = 2 is similar to the drop from K = 2 to K = 3, so at least three clusters. After 3, the drop in WSS does decrease from K = 3 to K = 4, and begins to level off between K = 4 and K = 5. Again, I would choose K=3 based on hierarchical clustering.

Create a dendrogram and label clusters for K = 3

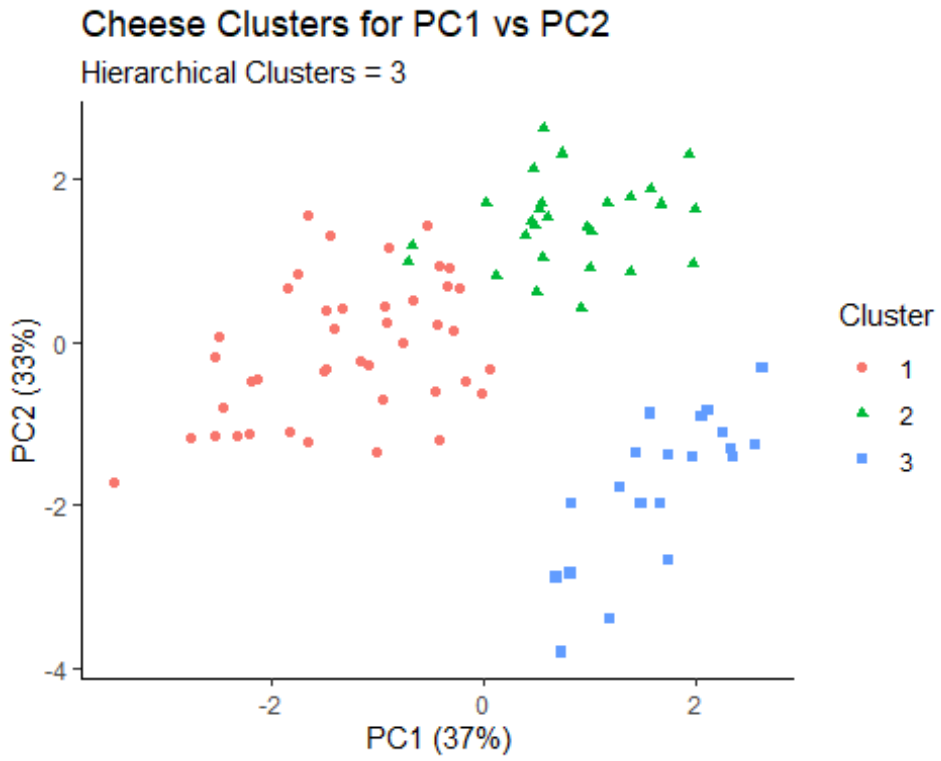
```
#Calculate dissimilarity using dist
eucDist <- stats::dist(x = cheeseScaled, method = "euclidean")

#Perform hierarchical clustering
hcComp <- hclust(d = eucDist, method = "complete")

#Create dendrogram
  fviz_dend(x = hcComp, k = 3, rect = TRUE) +
  labs(subtitle = "Complete Linkage K = 3")
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## ℹ The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Perform PCA and create visualization for K = 3 (Hierarchical Clustering)

```
#add hierarchical clusters to dataset with PC1 and PC2
temp_df <- temp_df %>%
        mutate(hclusters = cutree(hcComp,k=3))
#plot PC1 and PC2 with hierarchical clusters
ggplot(data=temp_df, mapping = aes(x=PC1, y=PC2,
       color= as.factor(hclusters), shape= as.factor(hclusters)))+
     geom_point()+
    labs(x = paste("PC1 (", round(100*PVE[1], digits = 1), "%)", sep = ""),
    y = paste("PC2 (", round(100*PVE[2], digits = 1), "%)", sep = ""),
    color = "Cluster", shape = "Cluster", title="Cheese Clusters for PC1 vs PC2",
    subtitle = "Hierarchical Clusters = 3")
```




```
###Added for easy comparison
# ggplot(data=temp_df, mapping = aes(x=PC1, y=PC2,
#        color= as.factor(cluster), shape= as.factor(cluster)))+
#      geom_point()+
#     labs(x = paste("PC1 (", round(100*PVE[1], digits = 1), "%)", sep = ""),
#     y = paste("PC2 (", round(100*PVE[2], digits = 1), "%)", sep = ""),
#     color = "Cluster", shape = "Cluster", title="cheese Clusters for PC1 vs PC2",
#     subtitle = "K-means = 5")
```

Create a two-way table

```
#cutree function extracts clusters from hierarchical clustering.

table(cheese$texture,cutree(hcComp,k=3))

##
##                 1  2  3
##   Hard         20  3  0
##   Pasta Filata  7 15  0
##   Semi-Hard    15  9  0
##   Soft          0  0 20

table(cheese$manufacturer,cutree(hcComp,k=3))

##
##      1  2  3
##   1 21 14 10
##   2 21 13 10
```

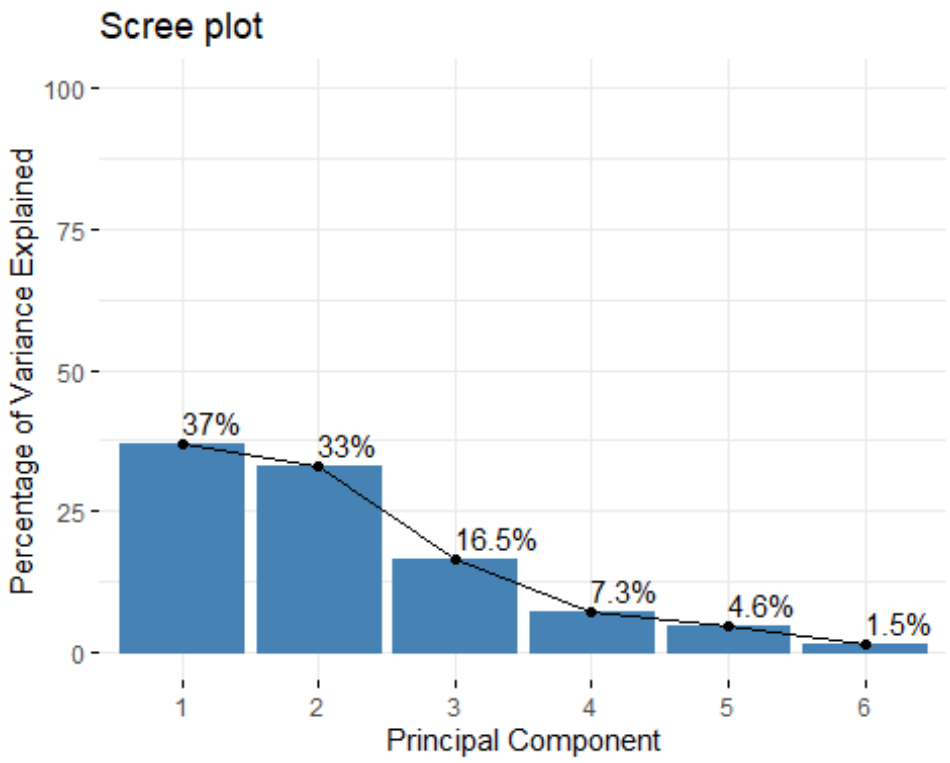
PCA using the standardized data

```
#Eigenvalues
cheesePCA$sdev^2

## [1] 2.2221184 1.9790336 0.9926898 0.4381877 0.2780022 0.0899682

#Eigenvectors
cheesePCA$rotation

##               PC1        PC2          PC3        PC4         PC5        PC6
## G80    -0.5709631  0.1748703  0.267662483 -0.3998301 -0.44265901  0.4647227
## vLTmax -0.1282233 -0.3337790 -0.812460824 -0.4416505  0.06174984  0.1148458
## vCO    -0.4334764 -0.4483686  0.184485627  0.3021850  0.59740449  0.3589164
## Fmax   -0.6266617 -0.1298620 -0.097196613  0.2486795 -0.23972616 -0.6794658
## FD      0.2351254 -0.5957717 -0.008803581  0.3780017 -0.61802428  0.2546670
## FO      0.1472051 -0.5340126  0.474030140 -0.5907945  0.06265129 -0.3398430

#First 6 rows of Projected Data (also called the scores)
head(cheesePCA$x)

##             PC1        PC2         PC3        PC4        PC5         PC6
## [1,] -2.5466249 -0.1766682 -0.57978822 -0.8049181  0.3532532  0.13236302
## [2,] -0.6809534  1.1797588  0.18940933 -0.8052421 -0.6601022  0.41776812
## [3,] -0.7200995  0.9892434  0.01757868 -0.6775128 -0.6959537  0.60817643
## [4,] -2.1492805 -0.4499664  0.67783561  0.4694673 -0.1922352  0.50907323
## [5,] -3.4941630 -1.7256061  0.13568039  0.7257660 -0.1123432  0.12693743
## [6,] -1.7569112  0.8427872  0.36000246 -0.8795381  0.1482879 -0.07476519
```

Based on the Kaiser rule (an eigenvalue greater than 1) we should reduce the data set down to the first two principal components as those two explain a significant amount (70%) of the variation seen in the cheese data set.

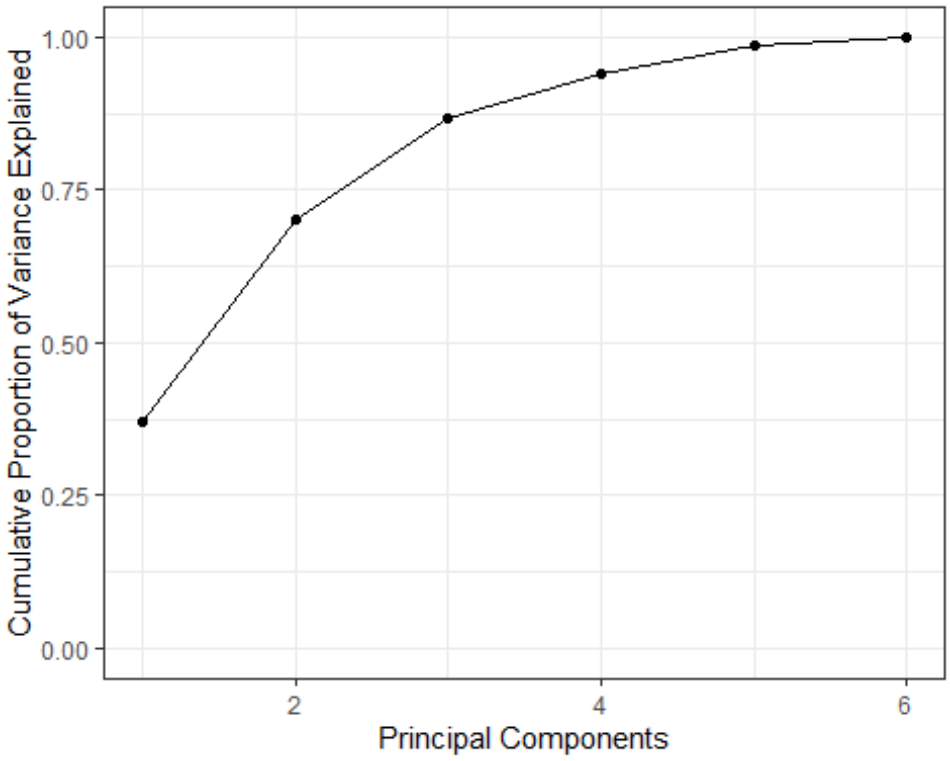
Create a scree plot

```
#Create scree plot of PVE
fviz_screeplot(X = cheesePCA,
               geom = c("bar", "line"),
               choice = "variance",
               ncp = 6, #Controls number of PC's displayed (CHANGE FOR OTHER DATA SETS)
               addlabels = TRUE #Adds percentages near points on plot
               ) +
  labs(x = "Principal Component",
       y = "Percentage of Variance Explained") +
  scale_y_continuous(limits = c(0,100))
```



Create a plot of the cumulative proportion of variance explained

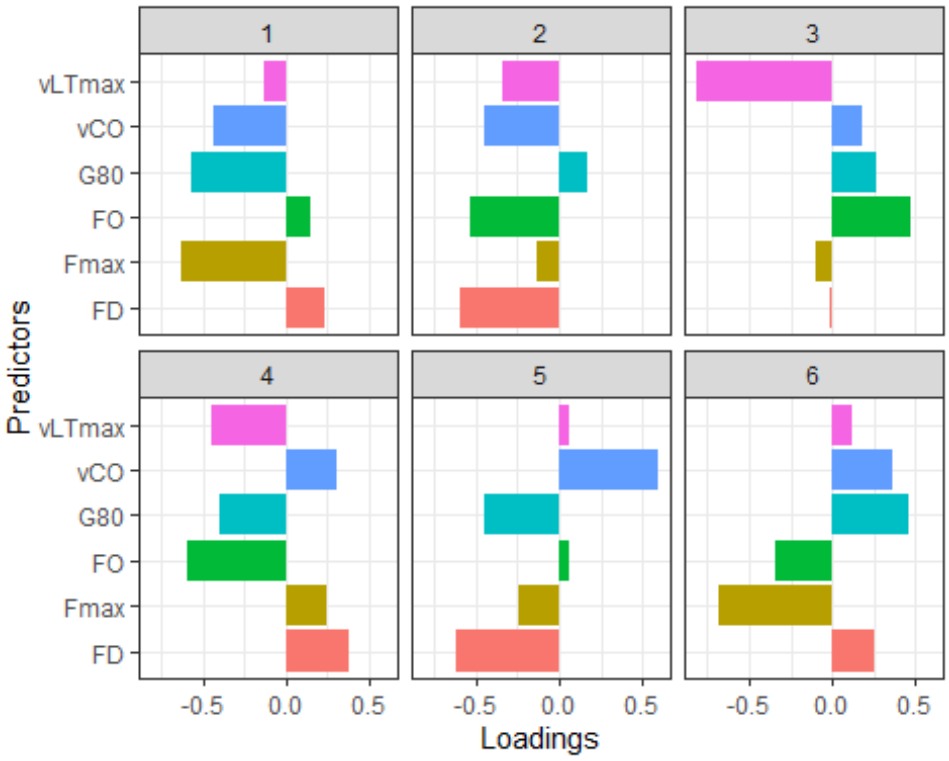
```
#Create a data frame for plotting
cumPVE_df <- data.frame(PC = 1:ncol(cheese_matrix), CumVarExp = cumsum(PVE))

#Create Cumulative Proportion of Variance Explained Plot
ggplot(data = cumPVE_df, mapping = aes(x = PC, y = CumVarExp)) +
  geom_point() +
  geom_line() +
  scale_y_continuous(limits = c(0,1)) +
  labs(x = "Principal Components",
       y = "Cumulative Proportion of Variance Explained") +
  theme_bw()
```



The cumulative proportion of the variation explained by each component can be seen in the scree plot above. PC1 explains 37% of the variation in our cheeses, while PC1 and PC2 together explain 70% of the total variation in the cheeses. PC3 explains an additional 16.5% for 86.5% of the total variation explained by the first three Principal Components.

Visualize the loadings

```
#Create plot of loadings
tidy(cheesePCA, matrix = "loadings") %>% #tidy is from the broom package
  ggplot(aes(x = value, y = column)) +
  facet_wrap(~ PC) +
  geom_col(aes(fill=column),show.legend = F)+
  labs(x= "Loadings",y="Predictors")+
  theme_bw()
```



The three variables that have the largest impact (greater than 0.4) on Principal Component One are vCO, G80, and Fmax, though all are negative. Those cheeses with a high PC1 score will generally have lower values of vCO, G80, and Fmax. The three variables that have the largest impact on Principal Component Two are vCO, FO, and FD, and all three are still negative. The cheeses with a high PC2 score will generally have lower values of vCO, FO, and FD.