# Stats 580 First Project

Vinh Hang

2022-10-08

# Stat 580 Project 1

# Executive Summary

The research group of Dr. Frankenstein was formed to assess different cheese variety through their thermalphysical properties and to consider how the use of statistical techniques can improve that practice. This interim report has three purposes: (1) to provide a comprehensive reports about cheese thermalphysical properties across 4 textures, (2) to present a grouping of cheese varieties based on their thermalphysical properties and (3) to perform classification of cheese texture based on said thermalphysical properties. This report includes relatively few conclusions and simple recommendations.

Chapters of this report describe our progress to analyze the three questions above with following results:

- Each cheese texture has statistically different thermalphysical properties except for the `Temperature v at tan (vLTmax)` . Soft cheese has a distinct thermalphysical properties from the rest.
- The optimal number of cheese variety using K-mean is 4. This concises with the number of textures.
- Cheese texture can be confidently predicted using their thermalphysical properties. The test error rate in this case is 0%.

Outline:

1. Project Description
2. Research Questions and Statistical Approaches
3. Variables
4. EDA
5. Statistical Analysis
6. Recommendations
7. Resources
8. Considerations

R requirements: psych, ggplot2, dplyr, gridExtra, grid, ggpubr, patchwork, table, qacr, factoextra, klaR, psych, MASS, devtools, rstatix, ggbiplot, caret

# Project Description

The purpose of this study is to analyze four different cheese group from 2 different manufacturers based on their thermophysical properties. The goal consists of identify the differences between them as well as how to group new cheese sample accordingly. The data is provided by Dr. Frankenstein’s group which consists of 89 cheese of various types from the four cheese textures. This is an observational study which focuses on cheeses produced by only two manufacturers. Thus, the conclusion might not hold for other manufacturers.

# Research Questions and Statistical Approaches

Dr. Frankenstein seeks answers for the following questions:

1. Are the thermophysical properties of the four cheese textures different? If so, which textures are different?

- Apply Multivariate analysis of variance (MANOVA) procedure to compare multivariate sample means.
- Verify MANOVA assumptions.
- Perform post-hoc comparisons.

2. Based on thermophysical characteristics, how many cheese varieties (not textures) are present in our data?

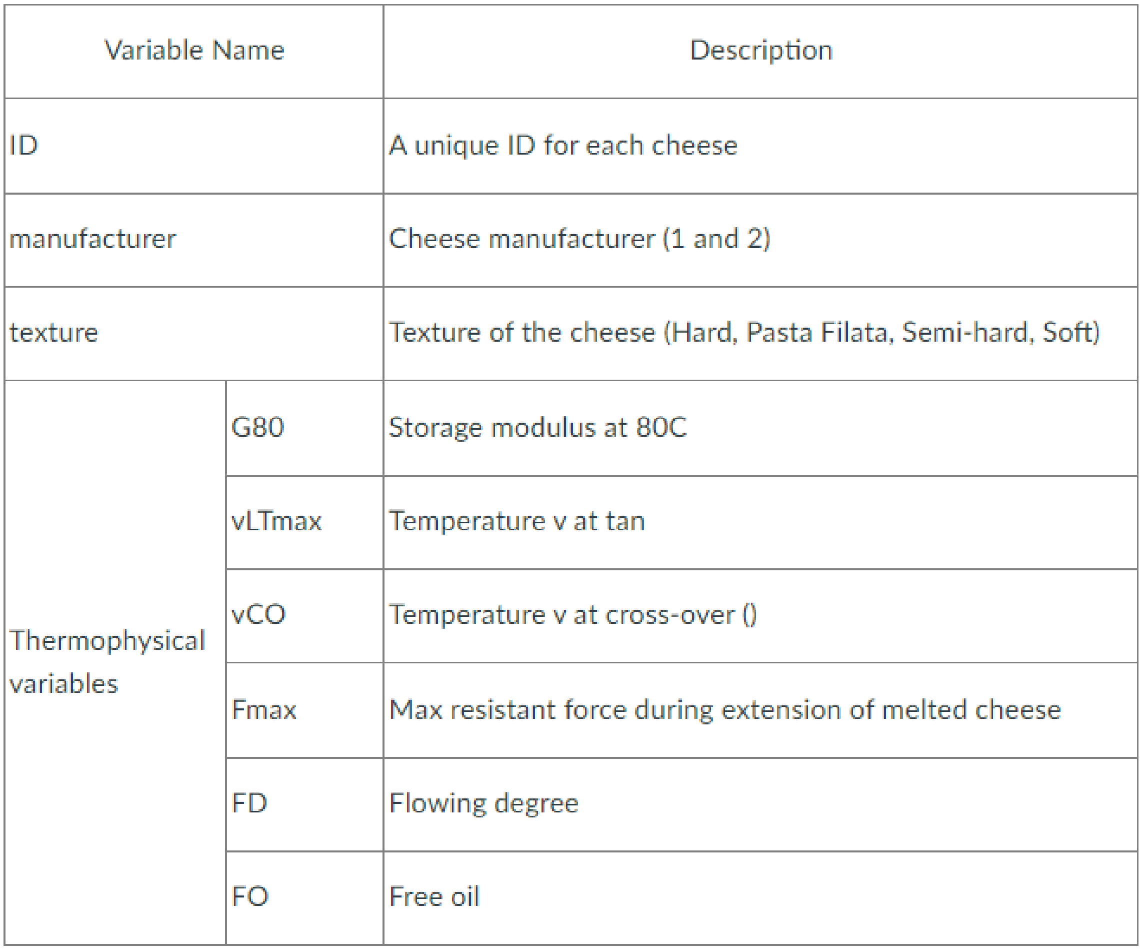
- Kmean algorithm with comparison to Hierarchical clustering.

3. Can we identify the texture of the cheese using the thermophysical characteristics for new cheese products?

- Perform LDA and assess confusion matrix.
- If the performance is not good, try out Random forest.

# Variables

The descriptions for all the variables in the data set are given in the table below. The six figures summary statistics are also given in Table 1.2. Each data points are collected randomly and recorded by Dr. Frankenstein team. For question 1 and 3, texture will be the response variable with all the thermophysical variables as explanatory. For question 2, K-mean is an unsupervised algorithm which requires no response variable and thus texture will only be used for comparison purposes.



| Variable Name | | Description |
|---|---|---|
| ID | | A unique ID for each cheese |
| manufacturer | | Cheese manufacturer (1 and 2) |
| texture | | Texture of the cheese (Hard, Pasta Filata, Semi-hard, Soft) |
| Thermophysical variables | G80 | Storage modulus at 80C |
| | vLTmax | Temperature v at tan |
| | vCO | Temperature v at cross-over () |
| | Fmax | Max resistant force during extension of melted cheese |
| | FD | Flowing degree |
| | FO | Free oil |

Table 1.1 - Data description

# EDA

Some important details:

- The data set is very clean and has no missing values.
- There are no detectable univariate or multivariate outliers for each of the thermophysical characteristics. Looking at the distribution (diagonal line from Figure 2.3), most of them are also normally distributed with the standard inverted U shaped. The only exception is `G80` where each cheese texture has a very distinct values from each other.
- The lower half of Figure 2.3 shows the scatter plot matrix of each variable with the upper half as the correlation number between them. Most thermophysical properties are not strongly correlated except for 3 pairs with moderate correlation (correlation ranges from -1 to 1 with 0 as no correlation at all):
- Temperature v at tan (vCO) and Max resistant force during extension of melted cheese (Fmax)
- Storage modulus at 80C (G80) and Max resistant force during extension of melted cheese (Fmax)
- Flowing degree (FD) and Free oil (FO)
- Figure 2.4 shows that cheese varieties are chosen randomly from each manufacturer (1 and 2) without any obvious pattern. This is a very good sign that the collection is carried out correctly.

An interesting note is the result of applying Principal Component Analysis (PCA)to 6 thermophysical properties after scaling. PCA is the technique to reduce number of columns used for algorithm by linearly combining them and use the ones that explained the most variation in the data. In Figure 2.5 and 2.6, about 3 or 4 principal components are good enough here. While there is no reason to use PCA with such a small data set, it can help visualize how each cheese texture are group together based on first and second principal components (see figure 2.7)

Table 1.2 - Size figures summary of thermophysical characteristics

| ID | manufacturer | texture | G80 | vLTmax | vCO | Fmax | FD | FO |
|---|---|---|---|---|---|---|---|---|
| Min. : 1 | Min. :1.000 | Hard :23 | Min. : 43.18 | Min. :67.15 | Min. :49.97 | Min. :1.380 | Min. :-0.590 | Min. :23.03 |
| 1st Qu.:23 | 1st Qu.:1.000 | Pasta Filata:22 | 1st Qu.: 93.10 | 1st Qu.:70.91 | 1st Qu.:52.73 | 1st Qu.:3.480 | 1st Qu.: 3.370 | 1st Qu.:30.58 |
| Median :45 | Median :1.000 | Semi-Hard :24 | Median :220.45 | Median :72.46 | Median :54.50 | Median :4.740 | Median : 5.240 | Median :34.55 |
| Mean :45 | Mean :1.494 | Soft :20 | Mean :200.95 | Mean :72.69 | Mean :54.22 | Mean :4.883 | Mean : 5.316 | Mean :35.16 |
| 3rd Qu.:67 | 3rd Qu.:2.000 | NA | 3rd Qu.:394.03 | 3rd Qu.:74.49 | 3rd Qu.:55.68 | 3rd Qu.:6.230 | 3rd Qu.: 6.920 | 3rd Qu.:37.98 |
| Max. :89 | Max. :2.000 | NA | Max. :413.84 | Max. :79.08 | Max. :59.69 | Max. :9.260 | Max. :12.340 | Max. :51.02 |

**Figure 2.3 - Thermophysical properties by Textures**

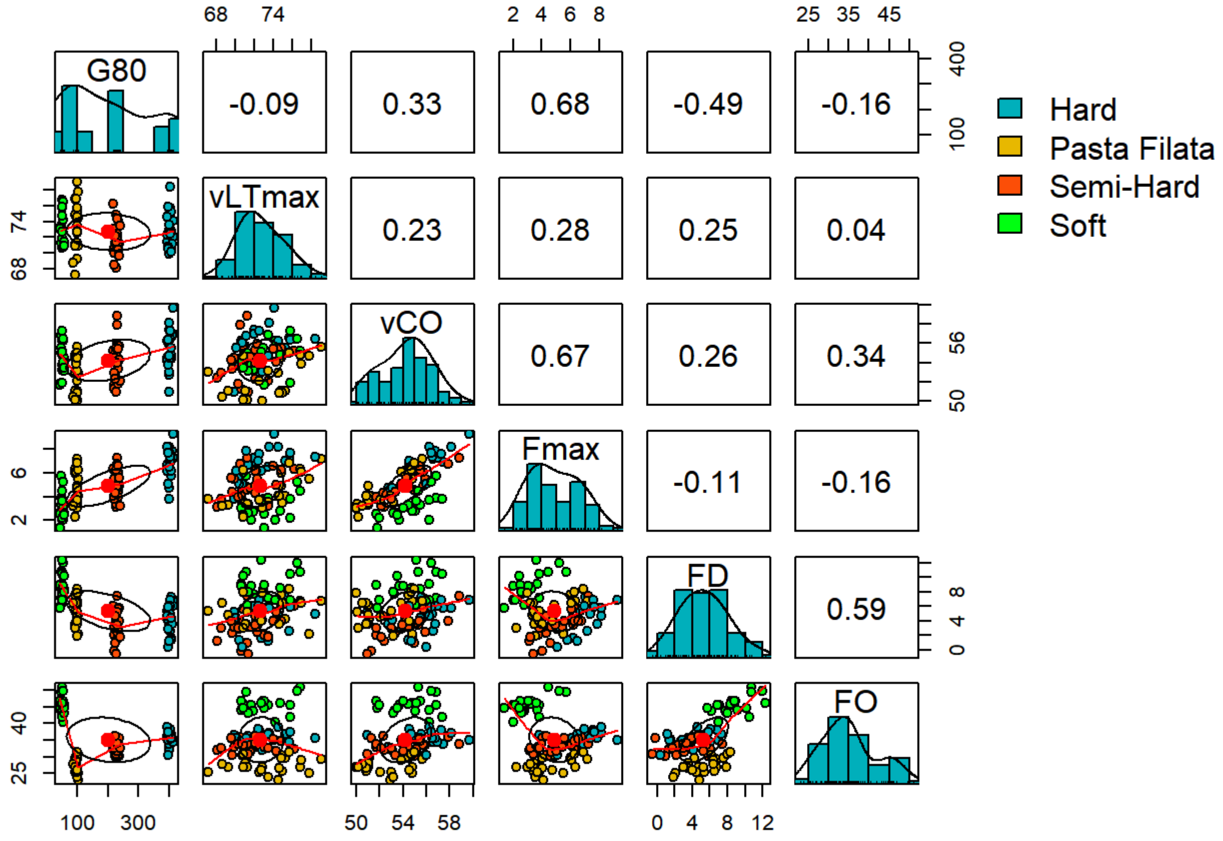


**Figure 2.4 - Thermophysical properties by Manufacturers**

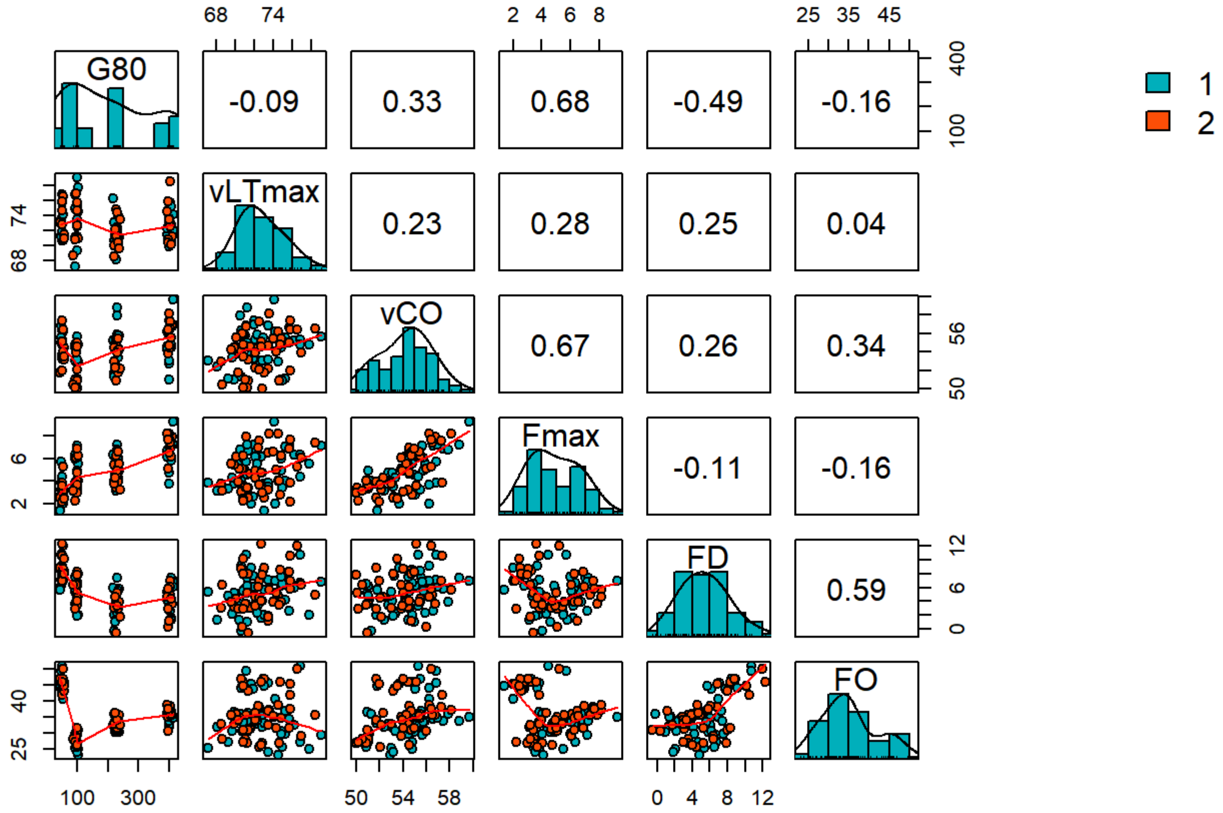


**Figure 2.5 - Scree plot**

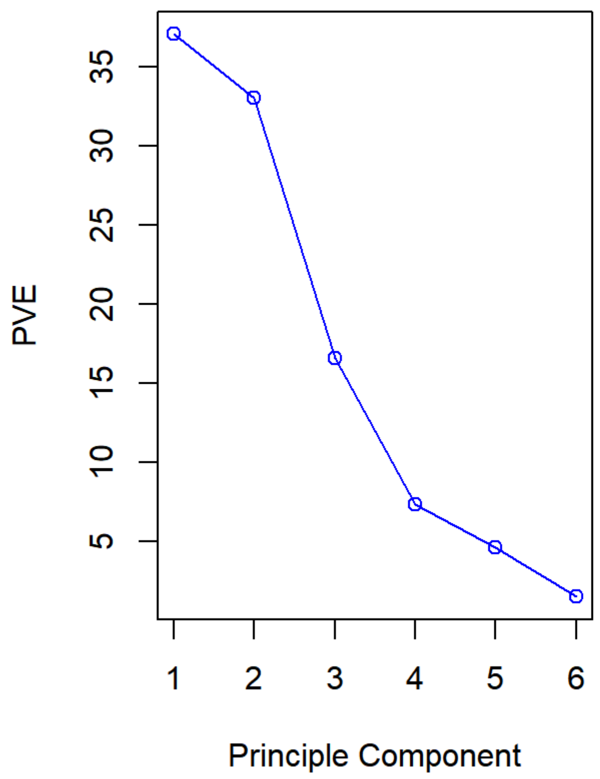


**Figure 2.6 - Cumulative**

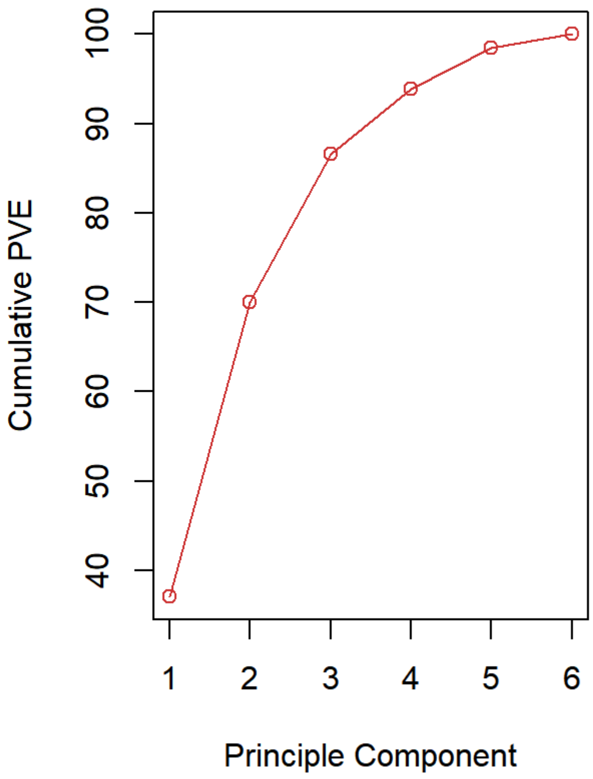
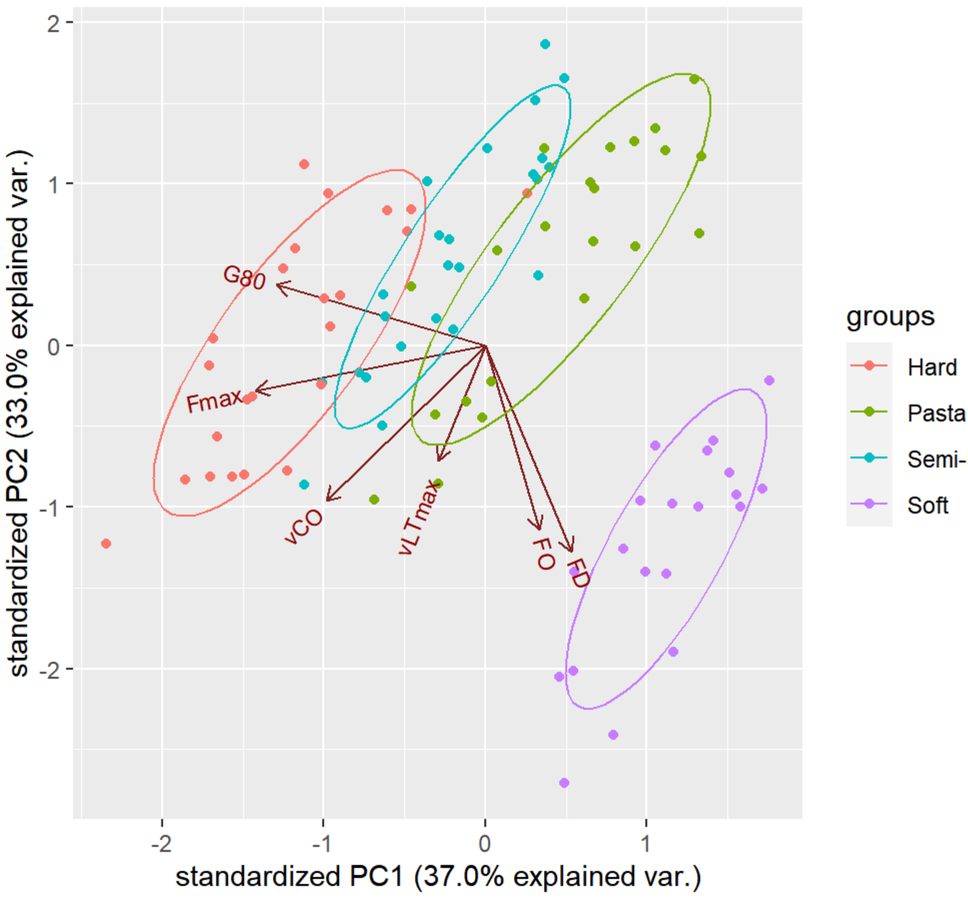


Figure 2.7 - Biplot with texture grouping



# Statistical Analysis

## 1 Are the thermophysical properties of the four cheese textures different? If so, which textures are different?

First, lets take a look at how each of the thermophysical property is distributed across 4 textures. Figure 3.1 and 3.2 shows the box plot of each thermophysical property separated by the texture group. Some observations:

- `G80` as mentioned above is very different between textures and thus the boxplot shows similar trait.
- Soft cheese seems to has very different thermophysical properties from the rest - this observation will come in handy when we need to group cheese togeher.
- Except forTemperature v at tan (vLTmax), all other thermophysical properties seems to be slightly different across the textures.

Figure 3.1 - Boxplot by textures: "G80", "vLTmax", "vCO"

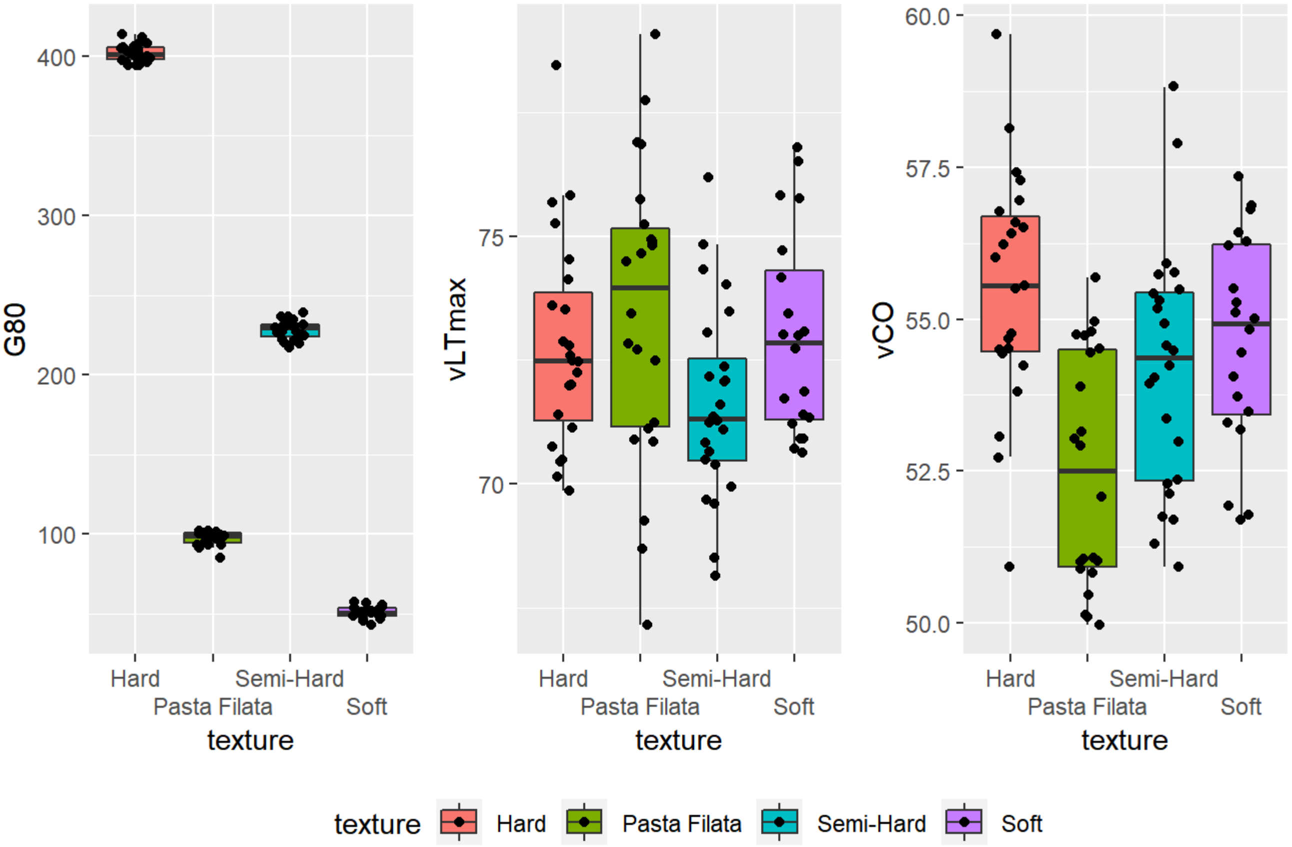
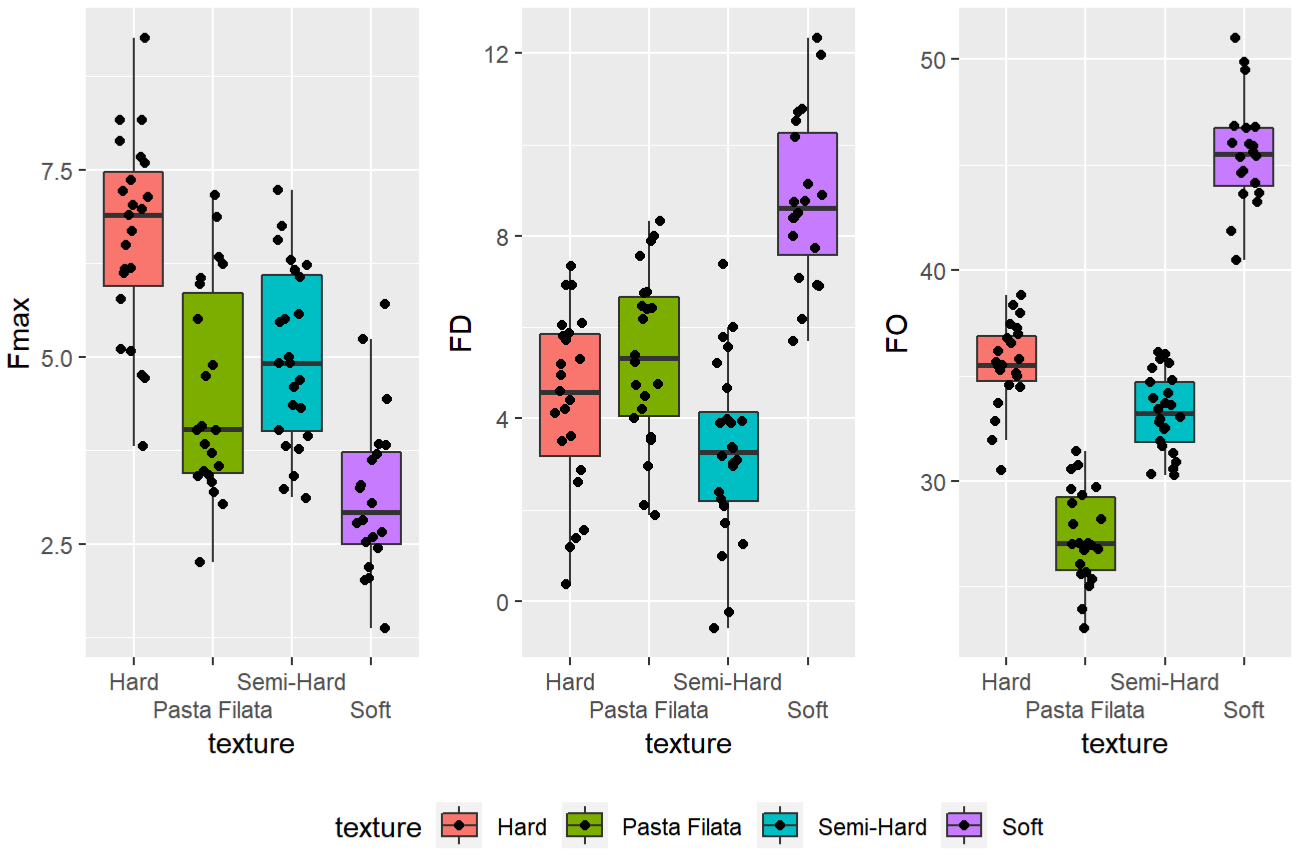


Figure 3.2 - Boxplot by textures: "Fmax", "FD", "FO"



Manova demand certain assumptions to be met before carried out. Two main violations are detected in this data which is:

- The data does not have homogeneous variance. (Reject Shapiro Test - p-values of 0.00396)
- The data does not have a multivariate normal distribution. (failed to reject Levene Test - large p-values)
- There is some multi-collinearity concern but all correlations are below 0.7.

With those observations on hand, MANOVA is performed to compare the mean vectors. We obtained the following statistics in the table below. Using Wilks as the benchmark, p-value is closed to tiny. Thus, it is found that there were differences in the thermophysical properties of at least one element between at least one pair of cheese textures.

```
## 
## Type II MANOVA Tests:
## 
## Sum of squares and products for error:
##                 G80    vLTmax       vCO      Fmax        FD         FO
## G80    2098.86197  49.51442 202.22965  88.44892  91.05512 -100.12129
## vLTmax   49.51442 461.85089 125.18898 118.00446  77.37479   77.01801
## vCO     202.22965 125.18898 311.38366 173.06583 132.21783   43.93618
## Fmax     88.44892 118.00446 173.06583 133.59122 117.57072  116.78398
## FD       91.05512  77.37479 132.21783 117.57072 309.79733  153.57279
## FO     -100.12129  77.01801  43.93618 116.78398 153.57279  396.45663
## 
## ------------------------------------------
##  
## Term: texture 
## 
## Sum of squares and products for the hypothesis:
##                  G80      vLTmax         vCO        Fmax            FD
## G80    1631258.647 -2747.95624 8470.808005 14100.70240 -16177.016263
## vLTmax   -2747.956    41.34723  -21.505709   -14.46485     69.217614
## vCO       8470.808   -21.50571  105.505331    50.42817      3.152501
## Fmax     14100.702   -14.46485   50.428171   131.55038   -162.509134
## FD      -16177.016    69.21761    3.152501  -162.50913    359.929996
## FO      -13123.749   -19.87023  394.055443  -282.41879    803.018667
##                   FO
## G80    -13123.74910
## vLTmax    -19.87023
## vCO       394.05544
## Fmax     -282.41879
## FD        803.01867
## FO       3582.54895
## 
## Multivariate Tests: texture
##                  Df test stat  approx F num Df   den Df     Pr(>F)    
## Pillai            3    2.2620    41.891     18 246.0000 < 2.22e-16 ***
## Wilks             3    0.0000   565.805     18 226.7595 < 2.22e-16 ***
## Hotelling-Lawley  3  893.3679  3904.349     18 236.0000 < 2.22e-16 ***
## Roy               3  852.0104 11644.142      6  82.0000 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

But how do they differ among sites? A quick look at the profile plot - Figure 3.3 (after scaling since different units between variables) shows all the properties are different across textures except maybe for vLTmax. This is the same observation looking at the box plot above. However, to concretely prove this, post-hoc tests of indiviual properties have to be carried out. Since homogeneity of variance assumption is violated, Welch anova test is the preferred choice compared to the normal anova. Table 3.1 shows that only vLTmax has the p-value of greater than 0.0083 (not 0.05 - see note below). This confirms the hypothesis that all thermalphysical properties are different between cheese textures except for vLTmax.

A more thorough pairwise comparison can also be views from table 3.2 (see Appendix) where similar result is showed as well.

*** Important note here is that we will use Bonferroni correction p-value which is p-value divided by number of variable instead of the normal p-value. This will make sure we control for experiment-wise error rate. In our case, the Bonferroni p-value = 0.05 / 6 = 0.0083.

Figure 3.3 - Mean Cluster Profiles of standardized thermophysical properties

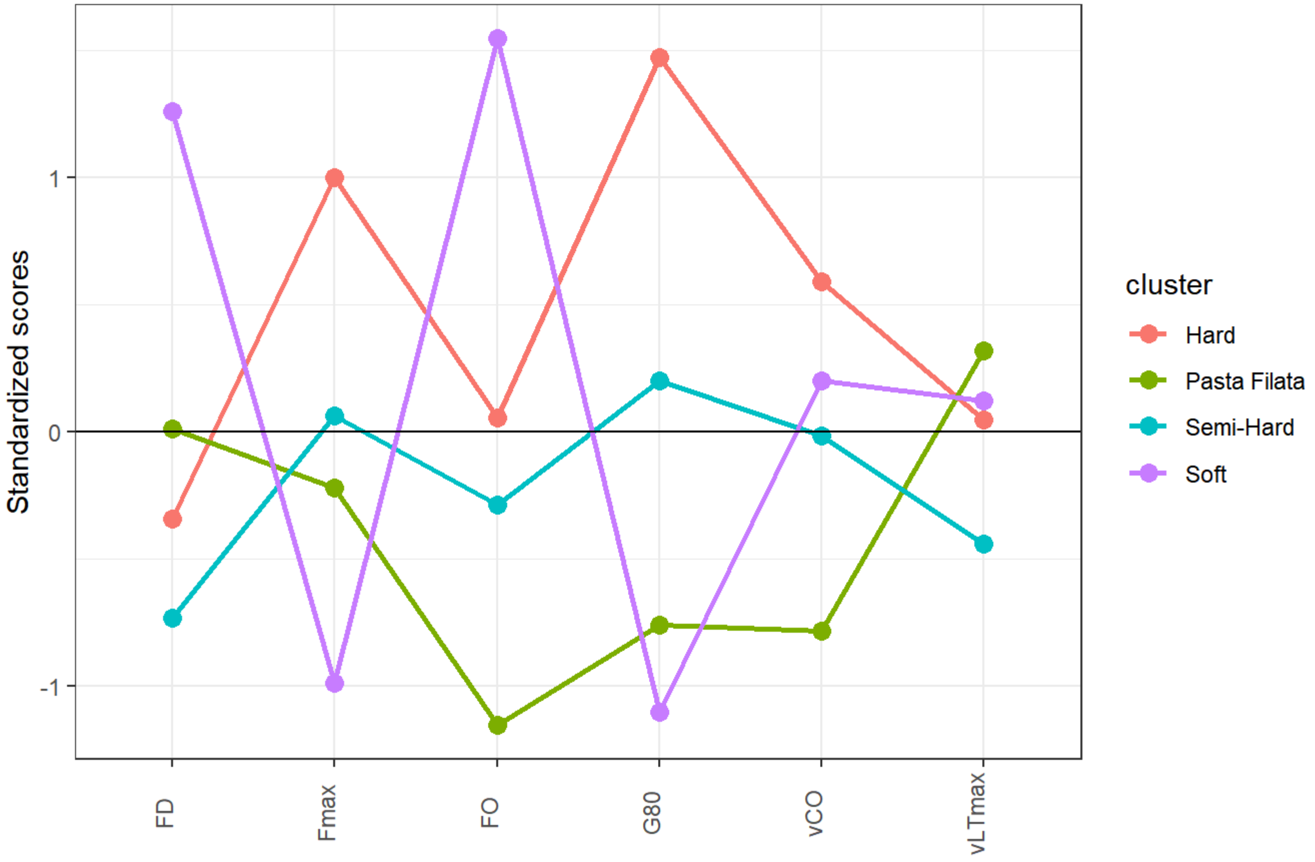


Table 3.1 - Welch Anova Test

| variable | .y. | n | statistic | DFn | DFd | p | method |
|---|---|---|---|---|---|---|---|
| FD | value | 89 | 33.44 | 3 | 46.93561 | 0.00e+00 | Welch ANOVA |
| Fmax | value | 89 | 29.47 | 3 | 46.88697 | 0.00e+00 | Welch ANOVA |
| FO | value | 89 | 197.75 | 3 | 45.88446 | 0.00e+00 | Welch ANOVA |
| G80 | value | 89 | 22436.74 | 3 | 47.13843 | 0.00e+00 | Welch ANOVA |
| vCO | value | 89 | 9.27 | 3 | 47.06380 | 6.31e-05 | Welch ANOVA |
| vLTmax | value | 89 | 2.70 | 3 | 46.29218 | 5.70e-02 | Welch ANOVA |

Table 3.2 - Pairwise Comparisons

| variable | .y. | group1 | group2 | p.adj | p.adj.signif |
|---|---|---|---|---|---|
| FD | value | Hard | Pasta Filata | 3.39e-01 | ns |
| FD | value | Hard | Semi-Hard | 2.46e-01 | ns |

| variable | .y. | group1 | group2 | p.adj | p.adj.signif |
|---|---|---|---|---:|---|
| FD | value | Hard | Soft | 0.00e+00 | **** |
| FD | value | Pasta Filata | Semi-Hard | 4.00e-03 | ** |
| FD | value | Pasta Filata | Soft | 2.80e-06 | **** |
| FD | value | Semi-Hard | Soft | 0.00e+00 | **** |
| Fmax | value | Hard | Pasta Filata | 2.88e-05 | **** |
| Fmax | value | Hard | Semi-Hard | 3.33e-04 | *** |
| Fmax | value | Hard | Soft | 0.00e+00 | **** |
| Fmax | value | Pasta Filata | Semi-Hard | 5.80e-01 | ns |
| Fmax | value | Pasta Filata | Soft | 7.00e-03 | ** |
| Fmax | value | Semi-Hard | Soft | 2.10e-05 | **** |
| FO | value | Hard | Pasta Filata | 0.00e+00 | **** |
| FO | value | Hard | Semi-Hard | 1.00e-03 | *** |
| FO | value | Hard | Soft | 0.00e+00 | **** |
| FO | value | Pasta Filata | Semi-Hard | 0.00e+00 | **** |
| FO | value | Pasta Filata | Soft | 0.00e+00 | **** |
| FO | value | Semi-Hard | Soft | 0.00e+00 | **** |
| G80 | value | Hard | Pasta Filata | 0.00e+00 | **** |
| G80 | value | Hard | Semi-Hard | 0.00e+00 | **** |
| G80 | value | Hard | Soft | 0.00e+00 | **** |
| G80 | value | Pasta Filata | Semi-Hard | 0.00e+00 | **** |
| G80 | value | Pasta Filata | Soft | 0.00e+00 | **** |
| G80 | value | Semi-Hard | Soft | 0.00e+00 | **** |
| vCO | value | Hard | Pasta Filata | 3.15e-05 | **** |
| vCO | value | Hard | Semi-Hard | 1.14e-01 | ns |
| vCO | value | Hard | Soft | 4.38e-01 | ns |
| vCO | value | Pasta Filata | Semi-Hard | 3.20e-02 | * |
| vCO | value | Pasta Filata | Soft | 3.00e-03 | ** |
| vCO | value | Semi-Hard | Soft | 8.37e-01 | ns |
| vLTmax | value | Hard | Pasta Filata | 8.46e-01 | ns |
| vLTmax | value | Hard | Semi-Hard | 2.18e-01 | ns |

| variable | .y. | group1 | group2 | p.adj | p.adj.signif |
|---|---|---|---|---:|---|
| vLTmax | value | Hard | Soft | 9.93e-01 | ns |
| vLTmax | value | Pasta Filata | Semi-Hard | 1.02e-01 | ns |
| vLTmax | value | Pasta Filata | Soft | 9.34e-01 | ns |
| vLTmax | value | Semi-Hard | Soft | 1.35e-01 | ns |

## 2. Based on thermophysical characteristics, how many cheese varieties (not textures) are present in our data?

This is an intriguing question requiring unsupervised clustering algorithm. Unsupervised here means that there are no label attached to the group before clusters (like texture). One can say the algorithm will `blindly` assess the data and group them based on some criteria. In this case, the algorithm K-means is chosen for its simplicity and high performance with `Euclidean distance` as the criteria. In simple terms, objects that are closed together should be grouped together! Fortunately, there is no categorical variable in thermophysical properties that will require a more complex criteria than Euclidean distance.

To carry a K-mean algorithm, we first need to pre-specify how many clusters to consider. From this initial choice, K-mean will start reassigning labels accordingly. One way of finding the optimal number of clusters is using within sum of square. Iterating over a number of initial k-clusters and find the k at elbow point (we do not want k to be too large). Figure 3.4 shows the best numbers of cluster seems to 3 or 4 here. We will go with 4 since it is the same number of textures we have.

Figure 3.5 shows how data is separated into the 4 groups on the first and second principal components plot (mentioned above in the EDA section). It is looking very good where the overlapping is quite small. The last remaining problem is to compare this grouping with cheese texture groups and see how different they are. Figure 3.7 shows the proportion of cluster in each texture. Notice the all observation in cluster 1 is fully enclosed in soft cheese texture. This agrees with initial observations in the EDA above. Cluster 3 seems to be mainly hard with some Pasta Filata.

One important thing to note is that different algorithm will group the data differently. There is a objectively best way to group data! Take a look at Figure 3.6 where instead of K-mean, hierarchical clustering algorithm is employed. In this case, 3 clusters seems to be much better than 4!

Figure 3.4 - Optimal number of clusters based on WSS

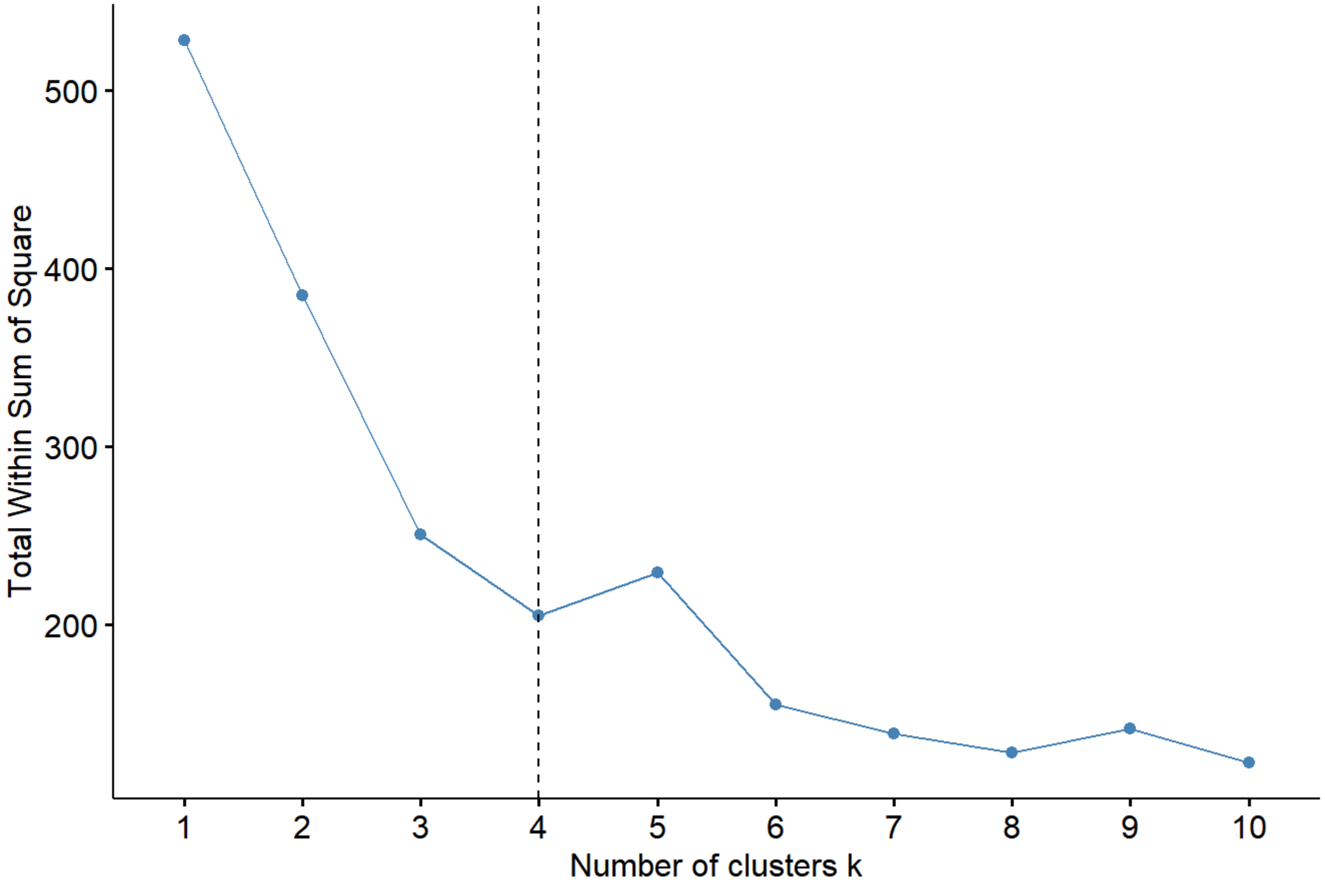
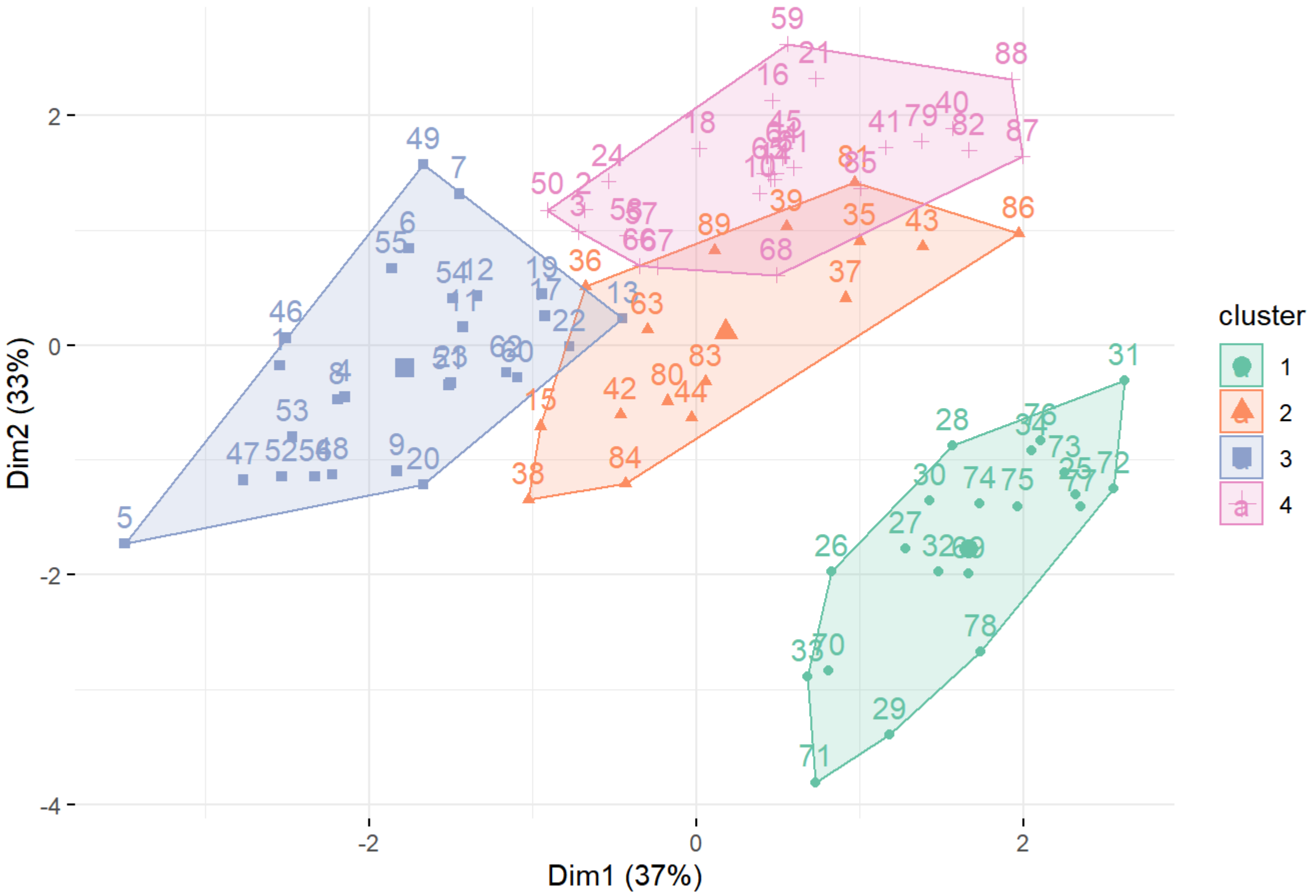
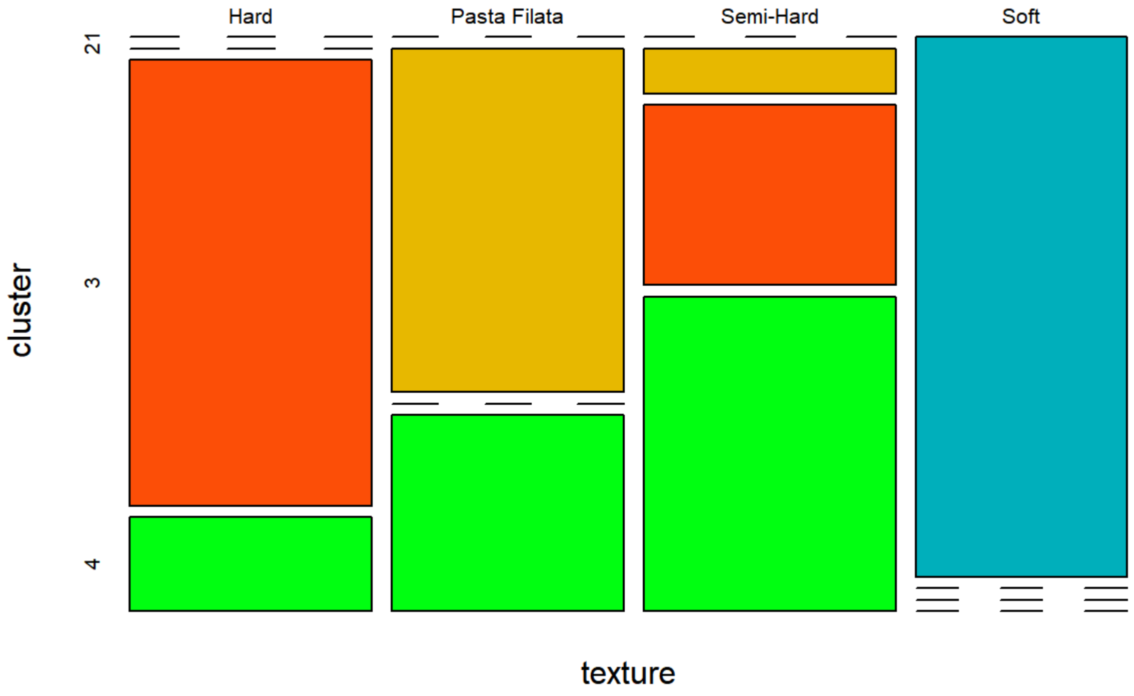


Figure 3.5 - Kmeans clustering



## Figure 3.7 - Mosaic plot



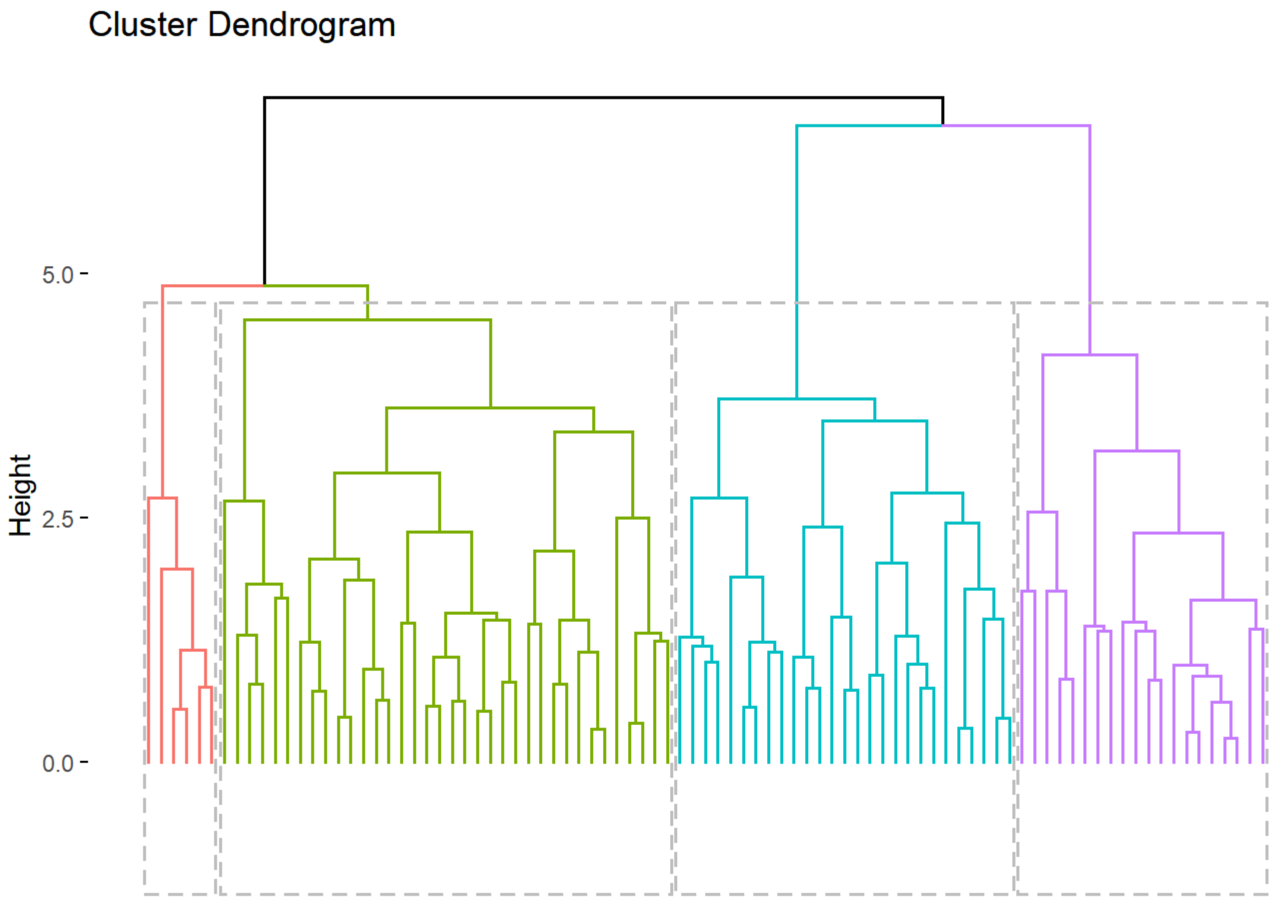
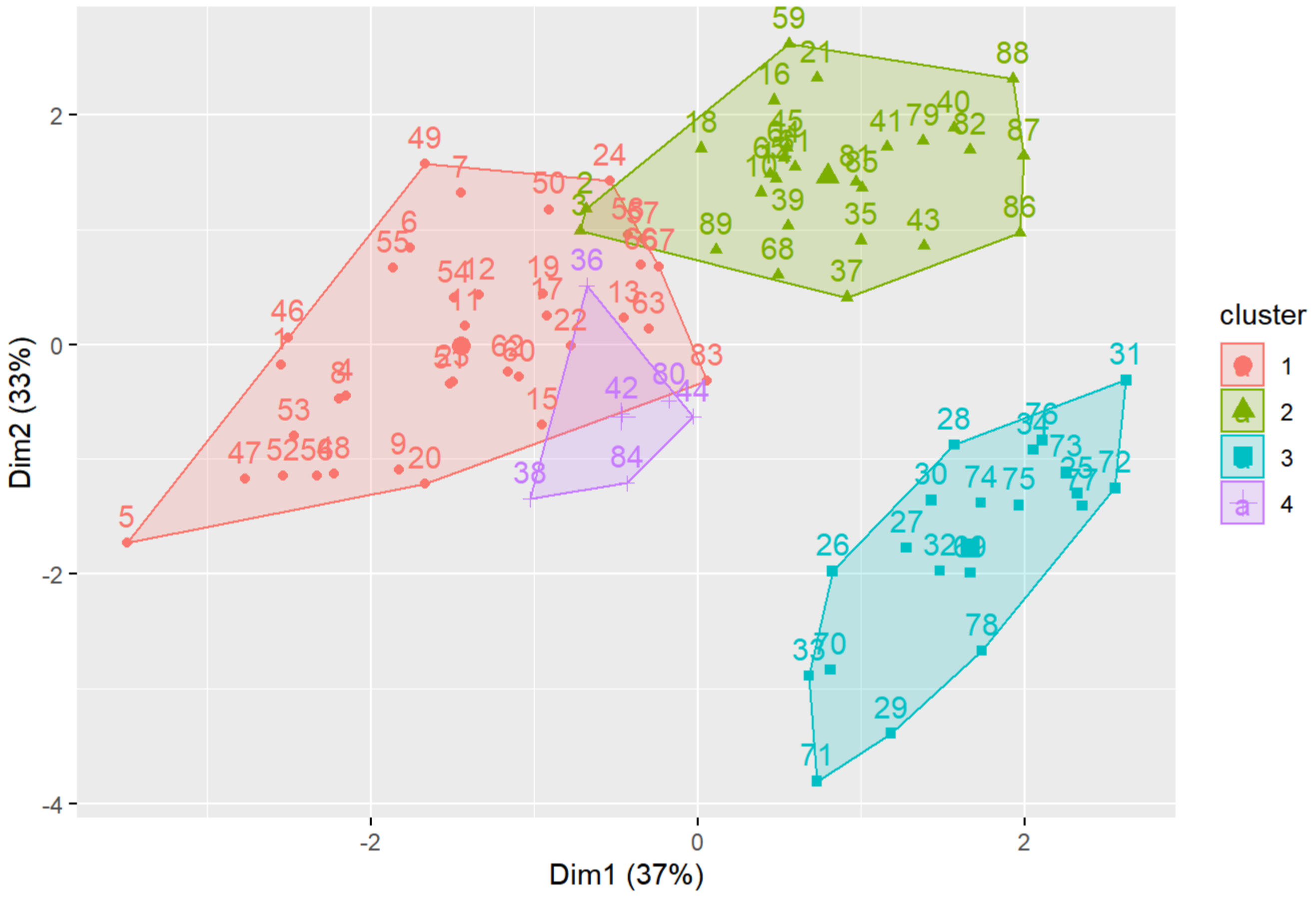


Figure 3.6 - Complete hierarchical clustering



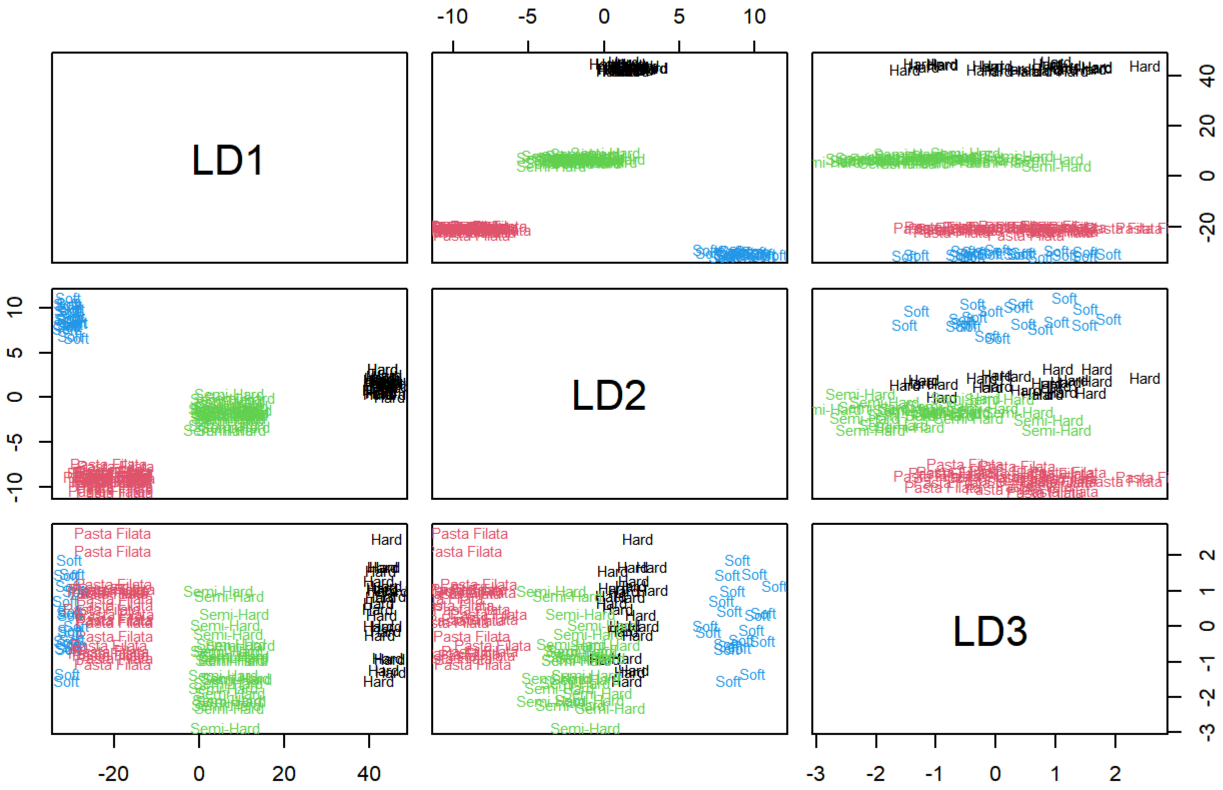
# Can we identify the texture of the cheese using the thermophysical characteristics for new cheese products?

The question requires a classification algorithm that take into account the thermophysical characteristics of a new cheese and assign them the texture. Some algorithms are Naive Bayes, Logistic regression, Linear Discriminant Analysis (LDA), tree-based…Lets proceed with LDA in this case. Unlike the first and second questions where model assessment is not necessary, we will want to see how well the LDA assigning the cheese to each texture. One way of accomplish this is through cross validations - where we take some data for training a model and use the rest for assessment model performance. Leave-one-out is a popular cross validation method and a built-in feature for R which we will utilize below.

The full output of LDA can be viewed in the notebook from the Appendix. Here we will focus on the performance with the prior of each texture equal to 0.25 since each cheese texture are equally likely to appear. Figure 3.8 shows very good predictions based on LD1, LD2, LD3. Naturally, Table 3.2 shows that LDA predicted perfectly every observation! This means that based on the thermophysical characteristics, we can confidently predict which texture that cheese will have.

```
## Call:
## lda(texture ~ G80 + vLTmax + vCO + Fmax + FD + FO, data = df_full_scaled)
## 
## Prior probabilities of groups:
##         Hard Pasta Filata    Semi-Hard         Soft 
##    0.2584270    0.2471910    0.2696629    0.2247191 
## 
## Group means:
##                     G80      vLTmax        vCO        Fmax          FD
## Hard          1.4744978  0.04923058  0.5906675  1.00005473 -0.34183383
## Pasta Filata -0.7586051  0.31821820 -0.7813948 -0.21844363  0.01114675
## Semi-Hard     0.2001262 -0.44023907 -0.0171366  0.06487396 -0.73230841
## Soft         -1.1013584  0.12163170  0.2008306 -0.98762370  1.25961757
##                      FO
## Hard          0.05547489
## Pasta Filata -1.15389055
## Semi-Hard    -0.28528935
## Soft          1.54783070
## 
## Coefficients of linear discriminants:
##                LD1        LD2        LD3
## G80    28.54279797  2.8448523  1.1752965
## vLTmax -0.03141108  0.3640802  0.5393203
## vCO    -0.46077862  2.4566841 -0.3733917
## Fmax    0.36197810 -4.0245827 -0.4872422
## FD     -0.22054589  0.1197949  1.5844372
## FO      0.26988595  4.9143203 -0.6468207
## 
## Proportion of trace:
##    LD1    LD2    LD3 
## 0.9537 0.0458 0.0005 
```

## Figure 3.8 - LDA matrix



### Confusion matrix

Table 3.2 - Test Error Confusion Matrix

| | 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 |

# Recommendations

1. Are the thermophysical properties of the four cheese textures different? If so, which textures are different?

- The four cheese textures has statistically different thermophysical properties except for Temperature v at tan (vLTmax)
- Soft cheese has a very distinct thermophysical properties from the rest.

2. Based on thermophysical characteristics, how many cheese varieties (not textures) are present in our data?

- Using K-mean, four cheese varieties can be nicely grouped together.
- Using Hierarchical clustering, three cheese groups seems to be better.

3. Can we identify the texture of the cheese using the thermophysical characteristics for new cheese products?

- Cheese textures can be predicted precisely from their thermophysical characteristics. LDA gives a 0% error rate!

# Resources

For MANOVA, LDA, K-mean algorithm, please refer to the links below

PSU Stats 508 (https://online.stat.psu.edu/stat508/)

An Introduction to Statistical Learning for more in-dept reference

ISLR (https://www.statlearning.com/)

LDA example (https://pages.cms.hu-berlin.de/EOL/gcg_quantitative-methods/Lab11_LDA_Model-assessment.html)

MANOVA exmample (https://www.datanovia.com/en/lessons/one-way-manova-in-r/#assumptions-and-preleminary-tests)

K-mean exmample (https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/)

# Considerations

There are two main concerns about the analysis. Firstly, the dataset size is small with only 89 observations. Moreover, there are only two manufacturers in the data set. Thus, the results might be very specific to the cheeses produced by these two manufacturers and cannot be applied to the general population.

Secondly, since this is an observational study, the results are inferential and not causal.

Appendix