Feature Selection From 20 Dimensions to 7: Multivariate Analysis on Food Consumption in European and Scandanavian Countries

Richard Feng

Abstract:

Given information of the percentage of people who consume a certain type of food in European and Scandanavian countries, we were curious to see whether there were certain foods that are heavily influenced by a country's geographic location. Given 20 different foods, we conducted feature selection to narrow down these 20 dimensions down to a few foods that really differentiate countries from one another and are unique to a certain geographical region. The 16 countries that were analyzed included Germany, Italy, France, Holland, Belgium, Luxembourg, England, Portugal, Austria, Switzerland, Sweden, Denmark, Norway, Finland, Spain, and Ireland.

We were able to narrow these 20 variables down to only seven key variables that could differentiate countries from one another. A accurate mapping of Europe could be constructed just by using data on the percentage of people in a country who consume these seven foods. These foods were:

1) Frozen Vegetables 2) Frozen Fish 3) Garlic 4) Olive Oil 5) Tinned Fruit 6) Crisp Bread 7) Tea

In other words, if we were given data just on these 16 countries and how much of their population consumed these seven foods, we would be able to reconstruct a relatively accurate European and Scandanavian map.

Intro:

The dataset obtained comes from Kevin Dunn, author of "Process Improvement Using Data." At first we were skeptical on the reliability of this dataset, as there was not a true source or citation of exactly where it came from. After emailing Kevin Dunn personally to ensure this dataset is valid, he responded with a confirmation of its reliability as it came from a real process. He extracted this dataset from a European statistics textbook from the late 1990s. The 16 countries in this dataset are Germany, Italy, France, Holland, Belgium, Luxembourg, England, Portugal, Austria, Switzerland, Sweden, Denmark, Norway, Finland, Spain, and Ireland. The 20 foods are real coffee, instant cofee, tea, sweetener, biscuits, powder soup, tin soup, potatoes, frozen fish, frozen vegetables, apples, oranges, tinned fruit, jam, garlic, butter, margarine, olive oil, yogurt, and crisp bread.

Upon analyzing the dataset, a small amount of missing data is found. Particularly, Spain is missing a "Sweetener" statistic, Sweden is missing a "Biscuits" statistic, and Finland is missing a "Yogurt" statistic. Because only one variable is missing for each of these three countries, these missing values will be replaced by an average of the column. Although we could simply omit these three countries, we would be losing a lot of data, as only one variable in each country is missing. If some strange results arise, it will have to be kept in mind that these three statistics for these three countries were estimated, as these three values are unreliable and will be biased.

Another downside of our dataset is that we only have 16 observations across 20 variables. Ideally, we want a 5:1 ratio of variables to observations, or at least 100 observations (Everitt, 93). Nonetheless, this multivariate dataset contains enough data that could be analyzed to draw some interesting conclusions.

Goals:

The main goal of this analysis is to narrow down our 20 food variables to a couple key variables that will still allow us to construct an accurate mapping of Europe by country. If possible, this indicates that

geographic location has a heavy influence on what kind of food a country consumes or does not consume. It is interesting to see if two countries that are located next to each other, despite having cultural and political differences such as speaking different languages or different political leaders, still have populations that consume similar foods.

Main Results:

Some initial visualization of the data is performed to quickly give some indications of interesting features that could be analyzed, and to make sure our goal is reasonable and achievable. The feature that stands out the most are the shapes of the faces. Italy, France, Belgium, and Spain all have oval, short faces. Portugal, Austria, Norway, and Finland have rounder heads. The rest of the countries, Germany, Holland, Luxembourg, England, Switzerland, Sweden, Denmark, and Ireland have tall faces. This indicates the country's tea consumption. The countries with tall faces such as England and Ireland have a significantly higher population of tea drinkers than countries such as Belgium and Spain with short faces.

Another easily identifiable feature is the length of the noses. Luxembourg, Holland, Ireland, and Switzerland all have long noses, indicating they have the highest percentage of people who drink powdered soup. Faces with shorter noses such as Belgium, Finland, and Austria have very short noses, indicating that the amount of powdered soup drinkers are drastically lower than countries such as Luxembourg with long noses.

Just from Chernoff Faces, we can see some variables (tea and powdered soup) that potentially are important in figuring out which key variables to retain out of the 20. There are many other features (18 to be exact) that could be analyzed, but for now, these Chernoff Faces serve as a brief initial overview of some similarities between the 16 countries.

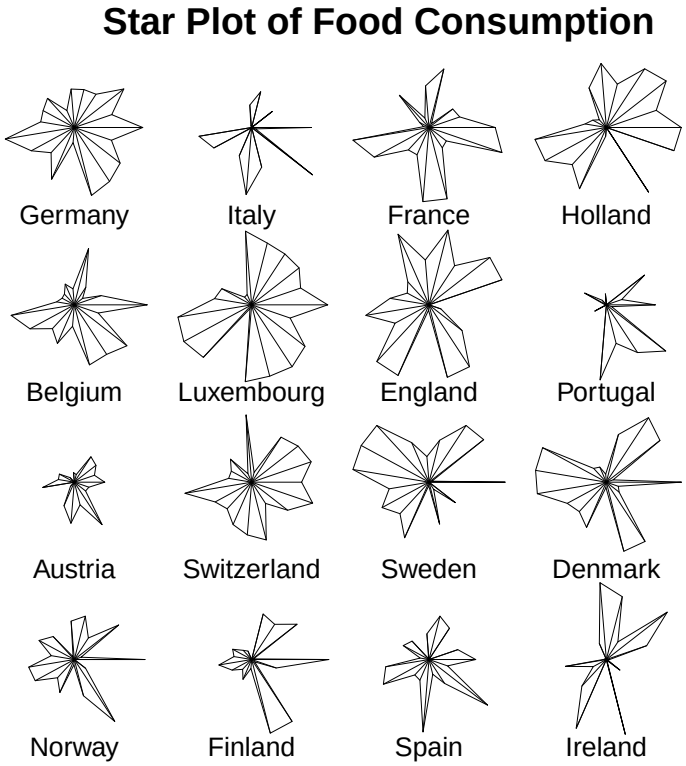


Figure 1.2

The star plots in Figure 1.2 were created for a more complete summary that incorporates all 20 variables rather than just 18. However, these star plots are less intuitive to read compared to faces. Similarly to Chernoff Faces, these star plots also serve as a nice summary of the countries and observing which countries are similar to each other.

Across all 20 dimensions, we can see how Norway and Finland are have relatively similar star shapes, indicating that they have similar populations in their food consumpion. Countries like Luxembourg, Germany, and Switzerland have populations that consumes almost all 20 of these foods, while countries such as Portugal, Ireland, and Italy have populations that only consume certain foods out of the 20 foods in the dataset. This is interesting, since Norway and Finland are very close together on a map, and Luxembourg, Germany, and Switzerland are also very close to each other too. These plots allow us to quickly confirm that our goal is achievable, and not completely unreasonable.

We would like to see if using all the information in the dataset could be used to group countries that also happen to be close to each other in proximity. Hierarchical clustering is performed below, using three different agglomeration methods, and the three dendrograms are analyzed to see which agglomeration method is most appropriate for our data.

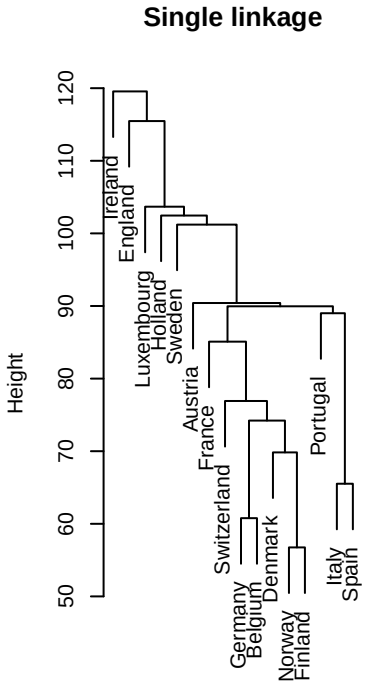


Figure 2.1

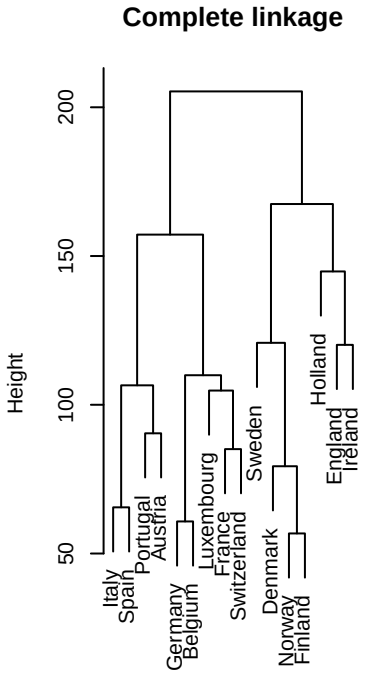


Figure 2.2

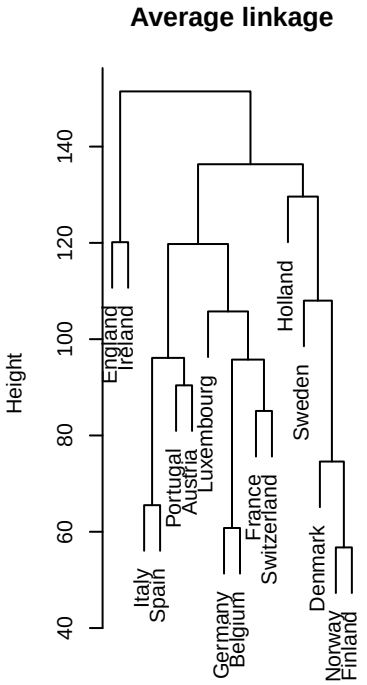


Figure 2.3

In Figure 2.1, we can see how single linkage seperates Ireland and England slightly even though they are very close in proximity, but complete and average linkage groups these two countries together. Similarly, Switzerland and France are also slightly seperated even though they are connected goegraphically, but complete and average linkage groups these two together. Norway, Sweden, Denmark, and Finland should be somewhat grouped together as well due to their geographical location. Single linkage fails to do this, and seperates Sweden drastically from Norway, Finland, and Denmark.

Complete and average linkeage are now left, and the two are extremely similar. Both group Norway, Finland, Denmark, and Sweden the same way. Holland is in between England and Sweden, so it getting clustered with either group is okay. However, a key difference that makes complete linkage the most accurate agglomeration method in terms of the country's geographic location is how all the countries in the right branch in Figure 2.2 are all seperated from the main chunk of Europe. This seperation is a nearly perfect seperation on the map, as shown by the white dotted lines in Figure 3 below. A quick note of the map is that "Netherlands" (in between Germany and Belgium on the dotted line) is Holland in our case.

Seeing as complete linkeage groups the countries in a way that is astonishingly accurate by their location, there certainly is some kind of relationship between a country and their surrounding country's food

Figure 1: Figure 3

consumption population.

Since using all 20 variables already gives an indication of the country's geographic location, we start with using classical multidimensional scaling using our data.

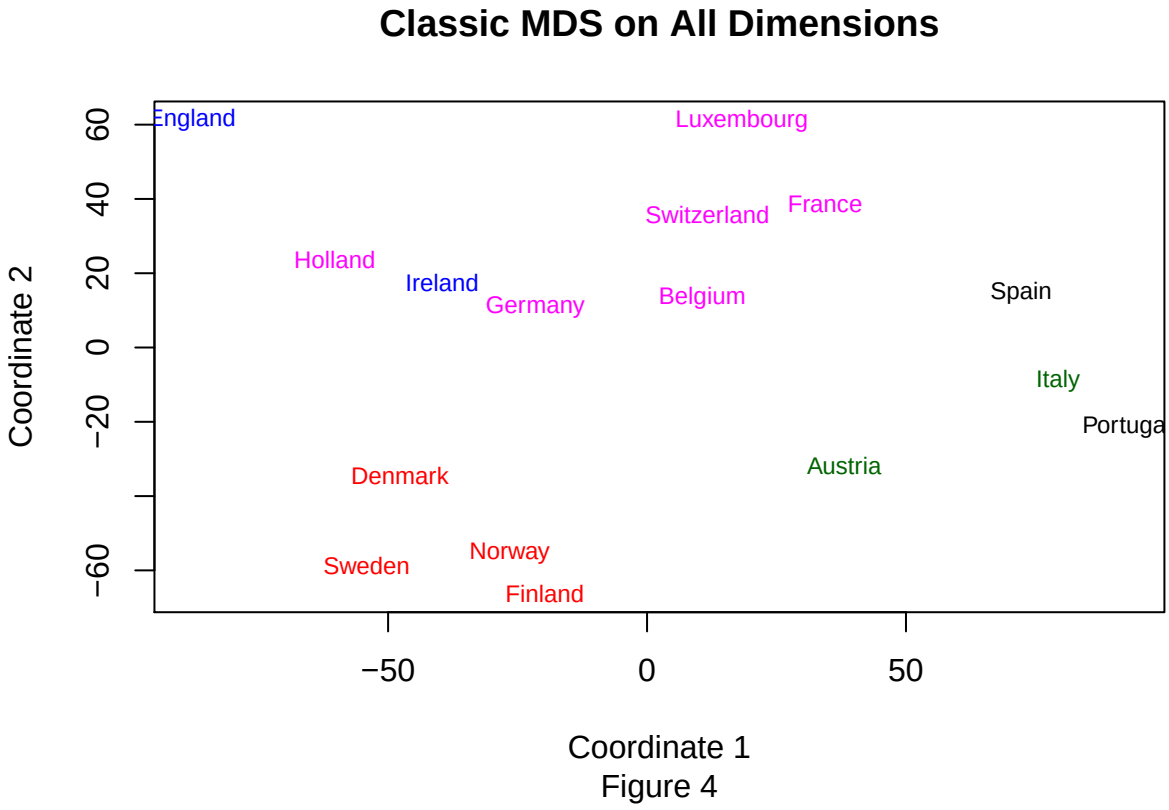


Figure 4

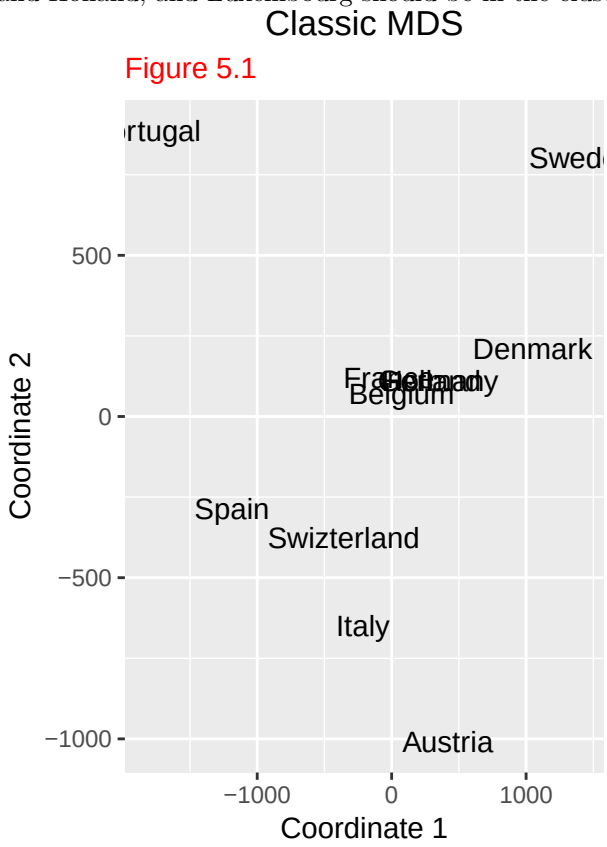
We can see that food consumption can definitely give an indication of where the country is located. There are many countries that are near each other on the map that are also near each other in Figure 4, such as Norway, Sweden, Finland, and Denmark, shown in red. However, there are still many countries that should be closer together Figure 4 but aren't, such as England and Ireland, and Portugal and Spain. Belgium, Germany, Switzerland, Luxembourg and France should ideally be clumped closer together, and closer to England than Ireland.

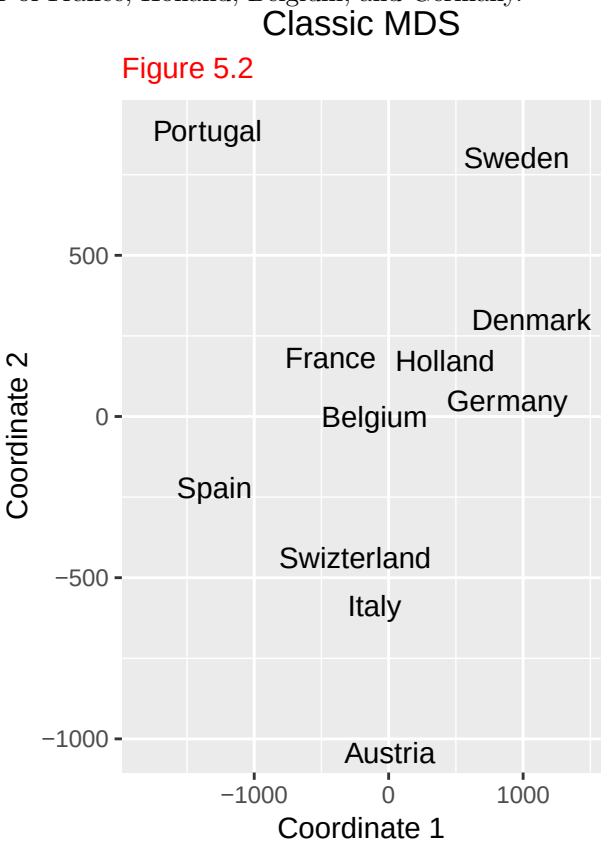
Something to note is that Italy is close to Portugal and Spain from our classic multidimensional scaling (classic MDS), even though they are seperated on the map. This indicates that besides geographic location, other factors may influence a country's food consumption, such as cultural or political influences. This is important to note, as a goal of constructing a perfect geographical map just from food consumption data may not be plausible. However, a somewhat of an accurate map should still be able to be constructed. We will continue to try and find the key variables that make the most accurate map.

Since comparing the country's location from our classic MDS figures with the map in Figure 3 is somewhat tedious and difficult, we constructed another figure using classic MDS that shows where countries should approximately be on a graph, rather than a map.

Ideally, we would want a plot that looks similar to Figure 5.1. Figure 5.2 displays this same plot, but with the labels not overlapping with each other.This figure was constructed using the "Eurodist" data from an R dataset. This Eurodist dataset contains the road distances between 21 cities in Europe. We then heavily modified this dataset, by first removing the cities where their corresponding country was not in our dataset, such as Athens, which is in Greece. Then, since there were multiple cities stemming from the same country, we chose the cities that were closest to the center of the country, and narrowed down the cities to one city per country. Finally, we replaced these city names with their corresponding country. Unfortuantely from this dataset, information on the distances between the following five countries were not included: Norway, Finland, England, Ireland, and Luxembourg. However, we know that Norway and Finland should be near

Sweden and Denmark. We also know that England and Ireland should be close together and close to France and Holland, and Luxembourg should be in the cluster of France, Holland, Belgium, and Germany.





Our plot in Figure 4 using classic mds scaling was relatively accurate, but we would like to see if we can narrow down the number of variables needed to reconstruct a European map. Principal component analysis is then performed on all 20 variables to see if we can achieve this goal.

Table 1: PCA Summary

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.502608	1.958965	1.65667	1.249706	1.129123	1.080159	0.9701247
Proportion of Variance	0.313150	0.191880	0.13723	0.078090	0.063750	0.058340	0.0470600
Cumulative Proportion	0.313150	0.505030	0.64226	0.720350	0.784090	0.842430	0.8894900

From Table 1, we can see that it takes four principal components to capture 72% of the variance in our data. The more variance we can capture, the more "information" we will have. Important plots of the first four principal components are shown below.

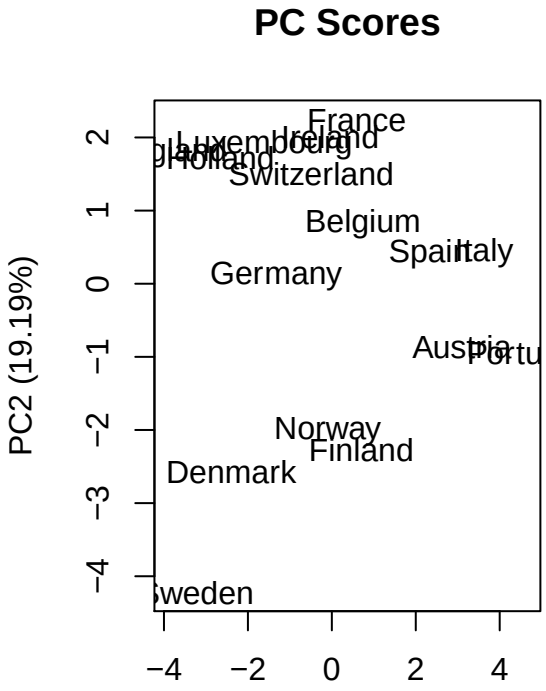


Figure 6.1

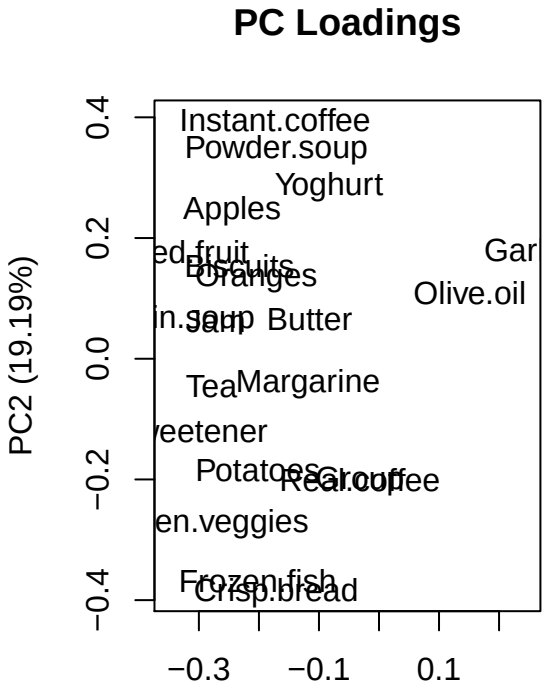


Figure 6.2

Just from Figures 6.1 and 6.2, we can gather information without even looking at the raw data on similarities between countries. Sweden seems to have high percentage of consumption of Frozen fish and Crisp bread, but low consumption of garlic and olive oil. Looking at the raw data, this observation matches up, as 93% of the population consumes crisp bread, while only 9% of the population consumes garlic. This is due to the fact that the the PC scores shown in Figure 6.1 are linear combinations of our original data combineed with a weighting given by the PC loadings in Figure 6.2 Similarly, we can see how Denmark, Norway, and Finland are a similar region. This indicates how these four countries may be similar in food consumption. Austria and Portugal are also very similar,as well as Spain and Italy, even though they're not geographically located near each other. Once again, this indicates that there may be some external factors that affect their food consumption such as cultural and political influences, that create a similar food consumption population between these pairs of countries. From the scree plot in Figure 7, we can see that elbows form after the second principal component, as well as after the fourth, indicating that either 2 or 4 principal components are needed to capture critical information.

We would like to account for at least 70% of the total variance. As stated in "An Introduction to Applied Multivariate Analysis in R" by Everitt et al. on page 71, "Values between 70%-90% are usually suggested."

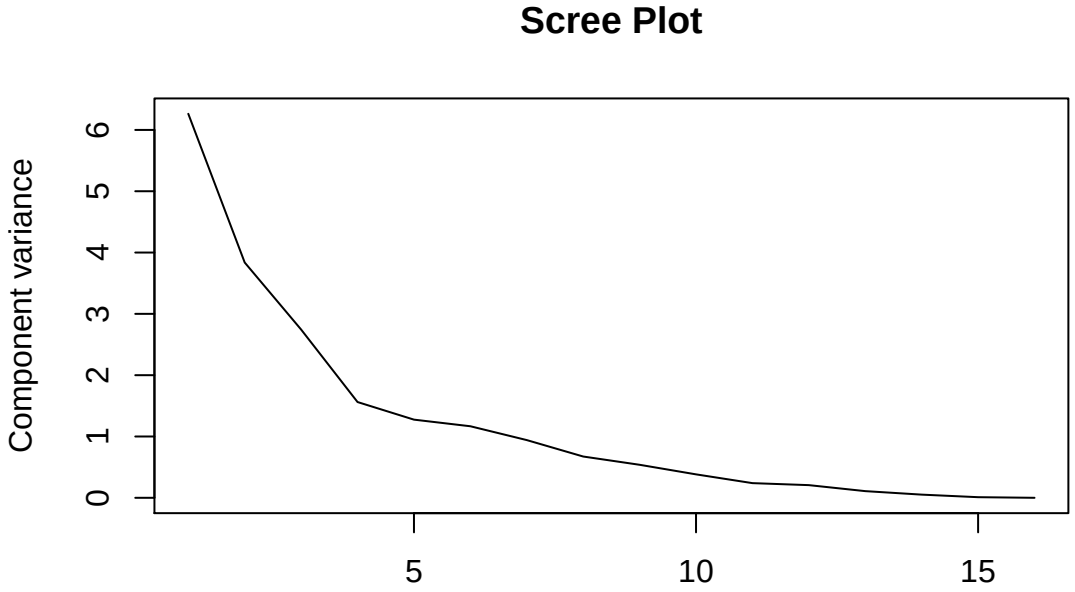


Figure 7

Unfortunately, in order to account for over 70% of the total variance, PC1, PC2, PC3, and PC4 are needed. These four principal components account for approximately 72% of the total variance as stated previously and seen in Table 1. We will now try to create a European map with the top 10 heavily weighted variables in PC1, shown in Table 2 above, even though PC1 only accounts for 31.31% of the total variation of the original variables.

Table 2: Weights of First Principal Component

	x
oranges	0.2042696
biscuits	0.2324063
apples	0.2446562
garlic	0.2450934
jam	0.2718108
tea	0.2788348
frozen.veggies	0.2981731
tin.soup	0.3089195
sweetener	0.3123369
tinned.fruit	0.3501713

Figure 8 does not do a very good job mapping these countries, and does not map them nearly as well as using all 20 variables as previously done in Figure 4. We will deal with this by further clustering these 20 variables into subgroups, by using hierarchical clustering on the variables rather than the countries this time.

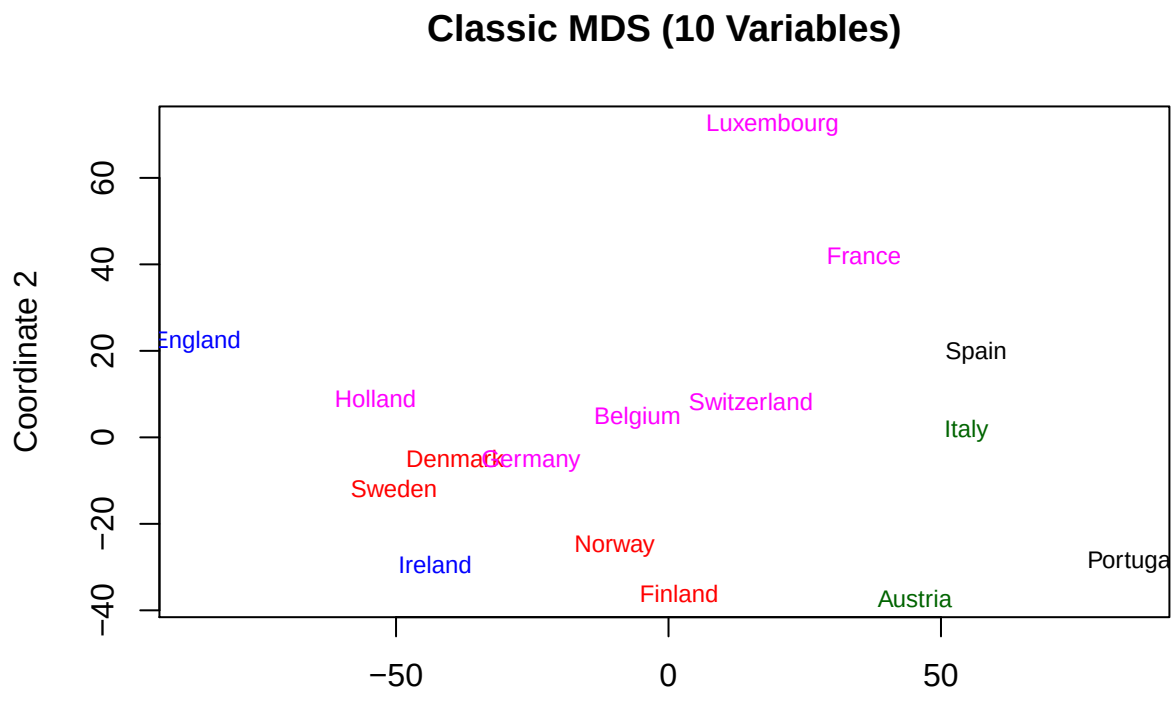


Figure 8

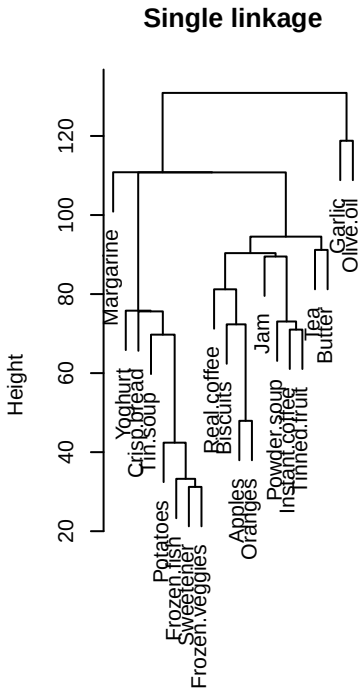


Figure 9.1

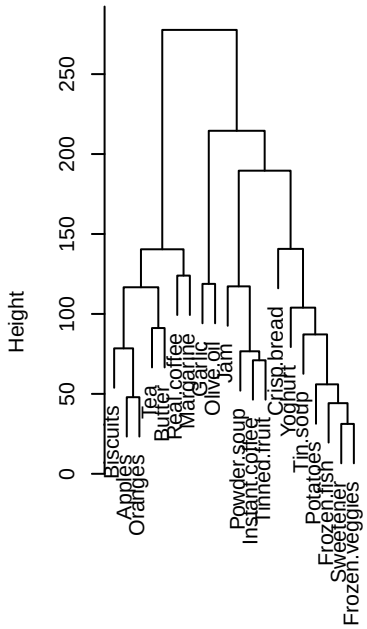


Figure 9.2

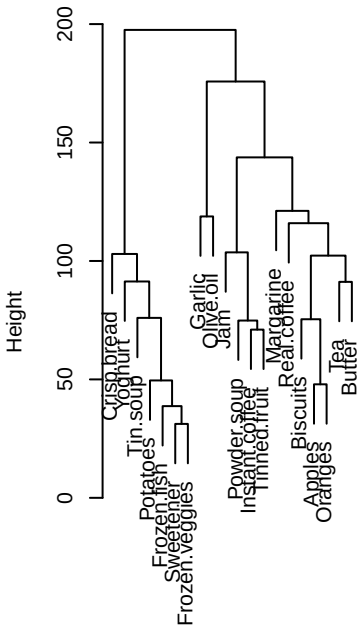


Figure 9.3

Three agglomeration methods are shown in Figure 9.1, 9.2, and 9.3. The most logical clusters that could be formed are found in both the complete and average linkage method. We see that many of these groups are

formed in ways we would expect and not as random as they are in single linkage.

We will explore both these clusters and perform principal component analysis on both methods of clustering to see which one results in a better European map using classic MDS.

```
##    Real.coffee Instant.coffee            Tea      Sweetener       Biscuits
##              1              2              1              2              1
##    Powder.soup       Tin.soup       Potatoes    Frozen.fish Frozen.veggies
##              2              2              2              2              2
##         Apples        Oranges   Tinned.fruit            Jam         Garlic
##              1              1              2              2              3
##         Butter      Margarine      Olive.oil        Yoghurt    Crisp.bread
##              1              1              3              2              2
```

Shown above are the groupings formed from complete linkage. We will use this groupings and perform principal componenet analysis on all of three of them.

Table 3: Weights of First Principal Component

	x
butter	0.2534029
real.coffee	0.2600822
biscuits	0.4634477
oranges	0.5264591
apples	0.5473338

Table 4: Weights of First Principal Component

	x
frozen.fish	0.3265822
tin.soup	0.3571472
tinned.fruit	0.3775840
sweetener	0.3934681
frozen.veggies	0.4124127

Table 5: Weights of First Principal Component

	x
garlic	0.7071068
olive.oil	0.7071068

Using the groupings formed through complete hierearchical clustering, the first three components are needed to capture 70% of the variance. However this time, the first principal component accounts for 36% of the variance. We will use the top three heaviest weighted variables in the first component shown in Table 3, being Oranges, Apples, and Biscuits. In group 2, the first two principal components make up 69% of the total variance, and captures a good amount of information. The top three heaviest weighted variables here are Frozen Veggies, Sweetener, and Tinned Fruit, as shown in Table 4. Finally in Table 5, our last group consists of just Olive oil and Garlic, so we will include these in our key variables as well.

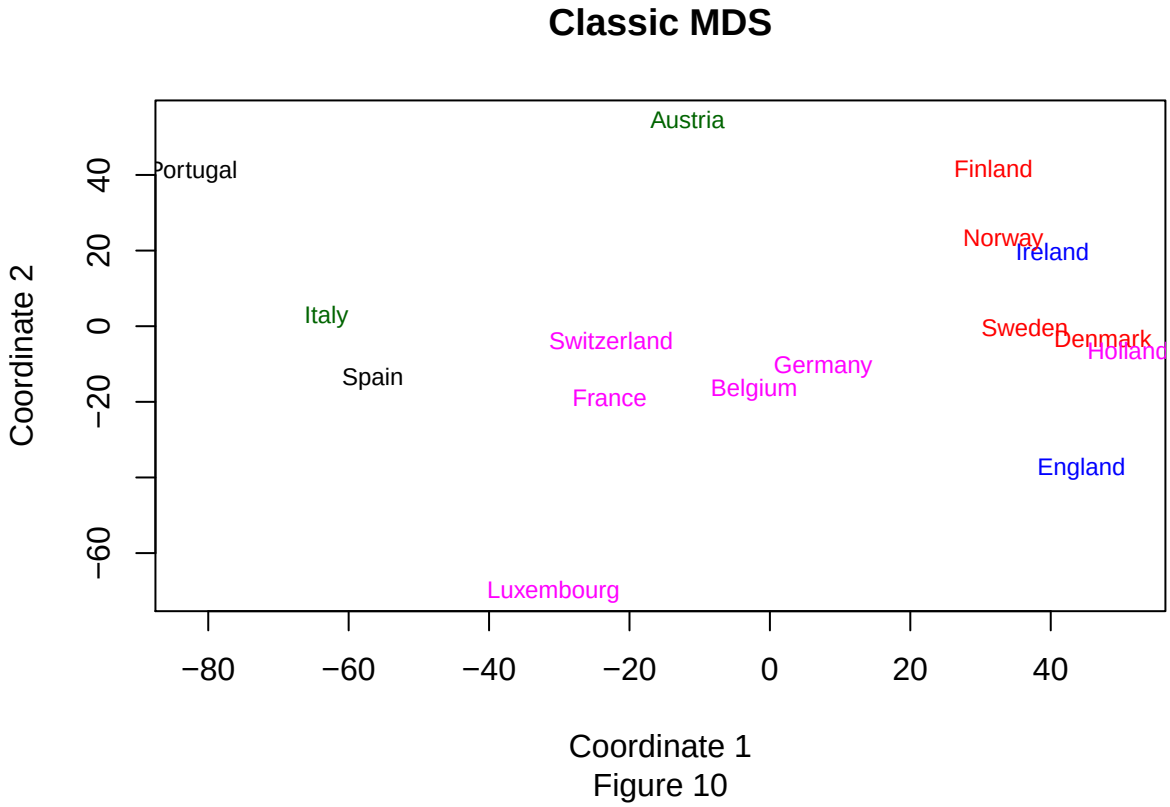


Figure 10

Using these 8 variables, we are still able to map a relatively accurate European map. We will now explore using the complete hierarchical clustering method to see if a different set of variables can improve this map. We can see the clustering groups below.

```
##    Real.coffee Instant.coffee           Tea      Sweetener       Biscuits
##              1               1             1              2              1
##    Powder.soup       Tin.soup      Potatoes    Frozen.fish Frozen.veggies
##              1               2             2              2              2
##         Apples        Oranges   Tinned.fruit            Jam         Garlic
##              1               1             1              1              3
##         Butter      Margarine      Olive.oil        Yoghurt    Crisp.bread
##              1               1             3              2              2
```

Table 6: Weights of First Principal Component

	x
biscuits	0.3415295
instant.coffee	0.3580033
powder.soup	0.3600582
apples	0.3718768
tinned.fruit	0.4500328

Table 7: Weights of First Principal Component

	x
potatoes	0.3419442
sweetener	0.3828034

	x
crisp.bread	0.3914815
frozen.fish	0.4298554
frozen.veggies	0.4657012

Table 8: Weights of First Principal Component

	x
garlic	0.7071068
olive.oil	0.7071068

Using groups formed with average linkage, the top three heaviest weighted variables in the first principal component of each group this time include tinned fruit, apples, powder soup, frozen veggies, frozen fish, crisp bread, garlic, and olive oil. It's interesting to see how garlic and olive oil are always grouped into their own seperate groupings. This may indicate that these two variables are somewhat important in distinguishing different countries. Using these eight variables, we can see below in Figure 11 that our European map constructed with just food consumption data using classic multidimensional scaling gives us an very nice and accurate map of Europe, better than using the clusters formed through complete linkage.

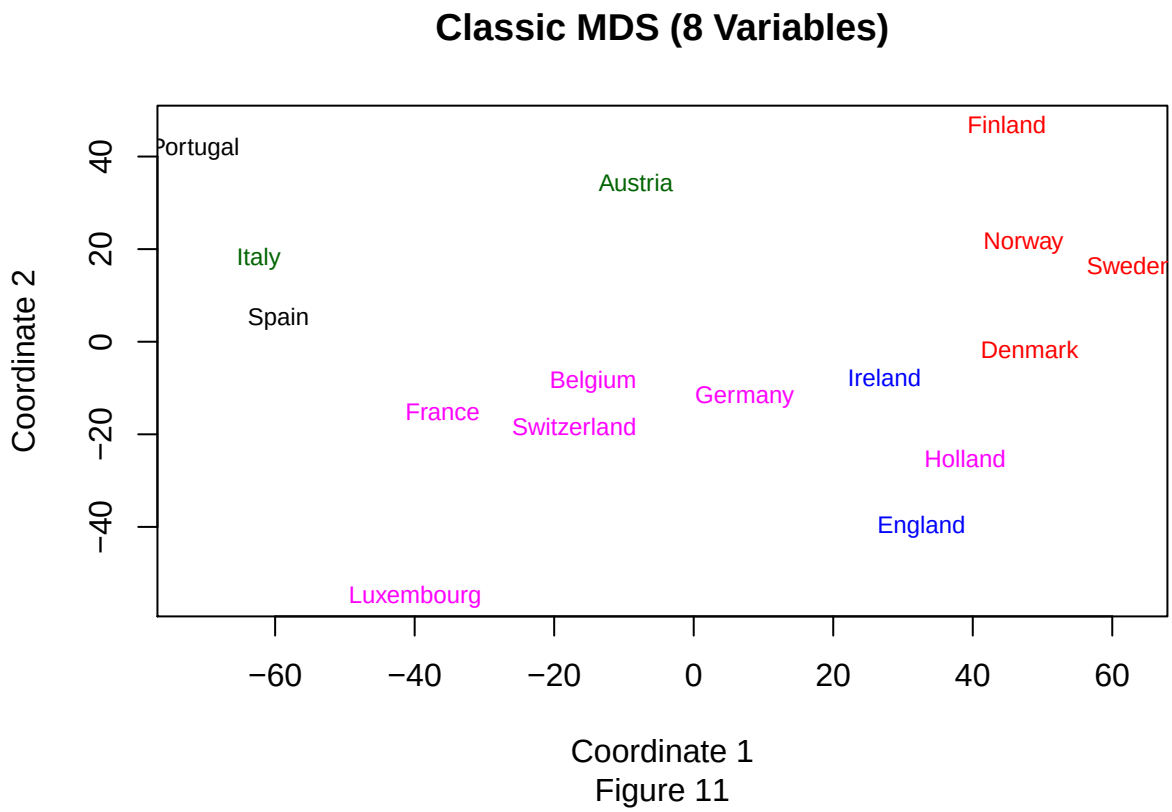


Figure 11

One improvement we would like to make to our map is group Ireland and England slightly closer together, with Holland closer to Germany than Ireland. Looking at the raw dataset and the columns, some apparent groups can be formed. Most interestingly, there is a group of "breakfast foods," which includes real coffee, instant coffee, tea, sweetener, and biscuits. Given prior cultural knowledge that England has a high amount of tea drinkers, we will examine this, as it may help us compile a better group of variables than can accurately create a European map.

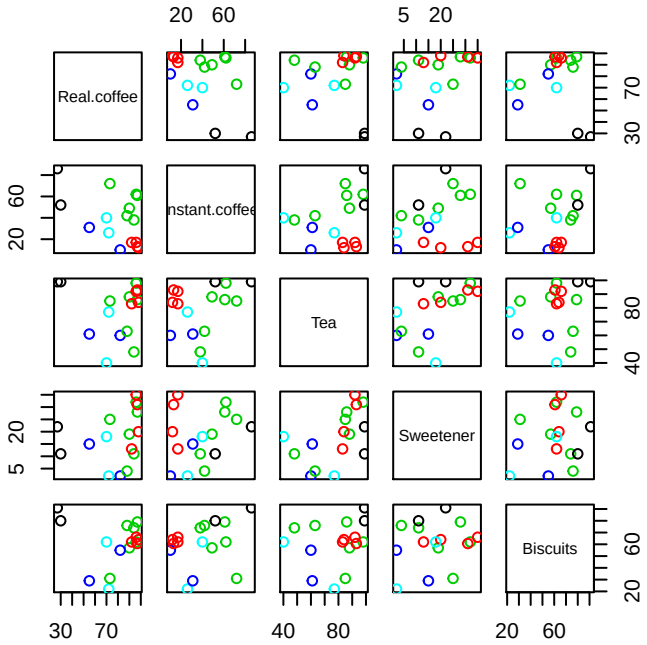


Figure 12.1

We can see that real coffee, tea and instant Coffee do a nice job grouping and differentiating countries on their geographical location, especially Ireland and England which in blue. The clusters of colors are much similar in distance for these three variables than the sweetener and biscuits, so we'll now examine these first three combinations in more detail.

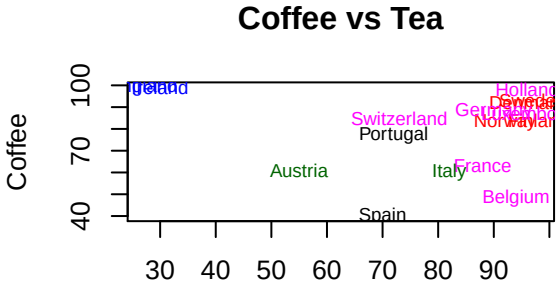


Figure 12.2

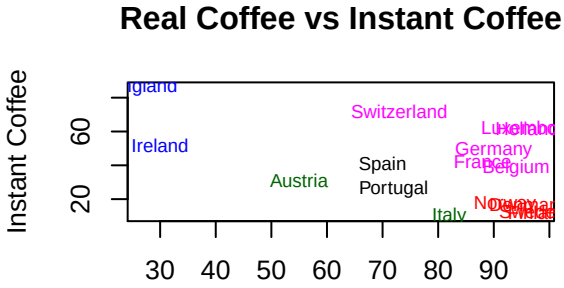


Figure 12.3

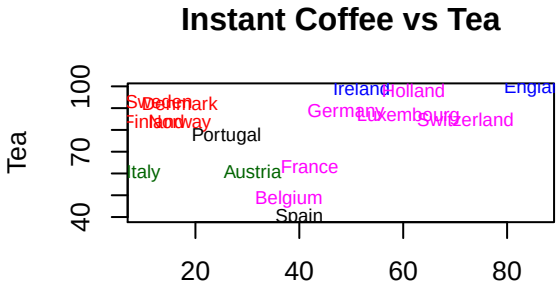


Figure 12.3

Our current goal is to find a set a variables that will help move Ireland and England closer together,

while keeping the other countries in a similar position as they are. We see that instant coffee, tea, and real coffee may give the best result for this goal. However, even though Ireland and England are closely grouped in this case, the other countries are not, so adding these two variables most likely will skew our other data by making some groups further apart while bringing Ireland and England closer together. Below in Figure 13, we test out the result of adding these two variables.

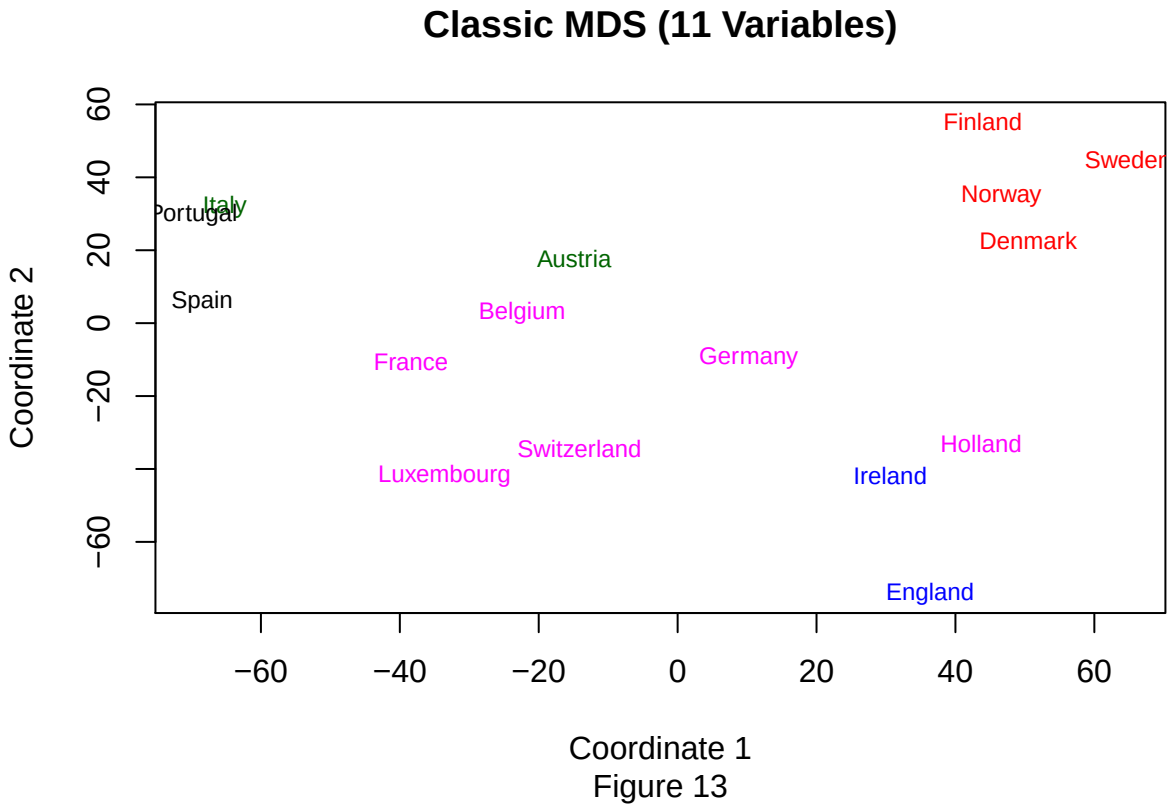


Figure 13

England and Ireland did indeed clump closer together, and brought Switzerland, Germany, Belgium, and France further apart. However, this may still be good, because the red countries are clumped closer together now, and positions are still relative to a real European map.

We now finally want to attempt to group Austria and Italy closer together, so we will examine these two countries in depth. We took the absolute values of the percentage difference between these two countries across all 20 dimensions. The variables that are most similar between Austria and Italy are tea, tin soup, potatoes, tinned fruit, jam, and crisp bread.

	Tea	Tin.soup	Potatoes	Tinned.fruit	Jam
2	1	2	3	5	5

Even if we add tin soup, potatoes, and jam to our list of key variables, Austria and Italy are still just as seperated. We may conclude that there is some kind of cultural or political factor that differs these two country and their populations in terms of food consumption.

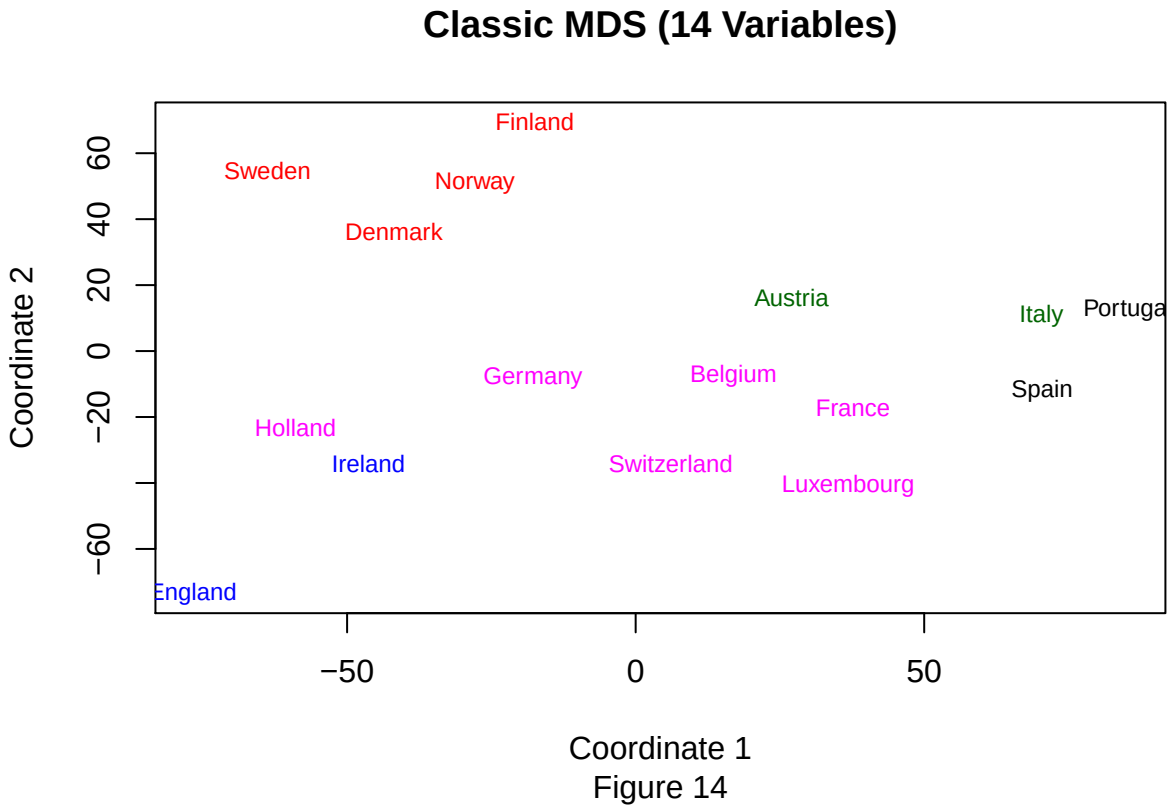


Figure 14

Finally, we will perform principal component analysis one last time on the 11 variables we narrowed down to (olive oil, garlic, tinned fruit, apples, powder soup, frozen veggies, frozen fish, crisp bread, real coffee, instant coffee, tea), in hopes of narrowing down the number of variables even more.

Table 10: Weights of First Principal Component

	x
real.coffee	0.1048583
powder.soup	0.1131319
apples	0.1909343
olive.oil	0.2583679
tinned.fruit	0.3469946
crisp.bread	0.3497639
garlic	0.3657498
frozen.fish	0.3685293
tea	0.3962649
frozen.veggies	0.4469678

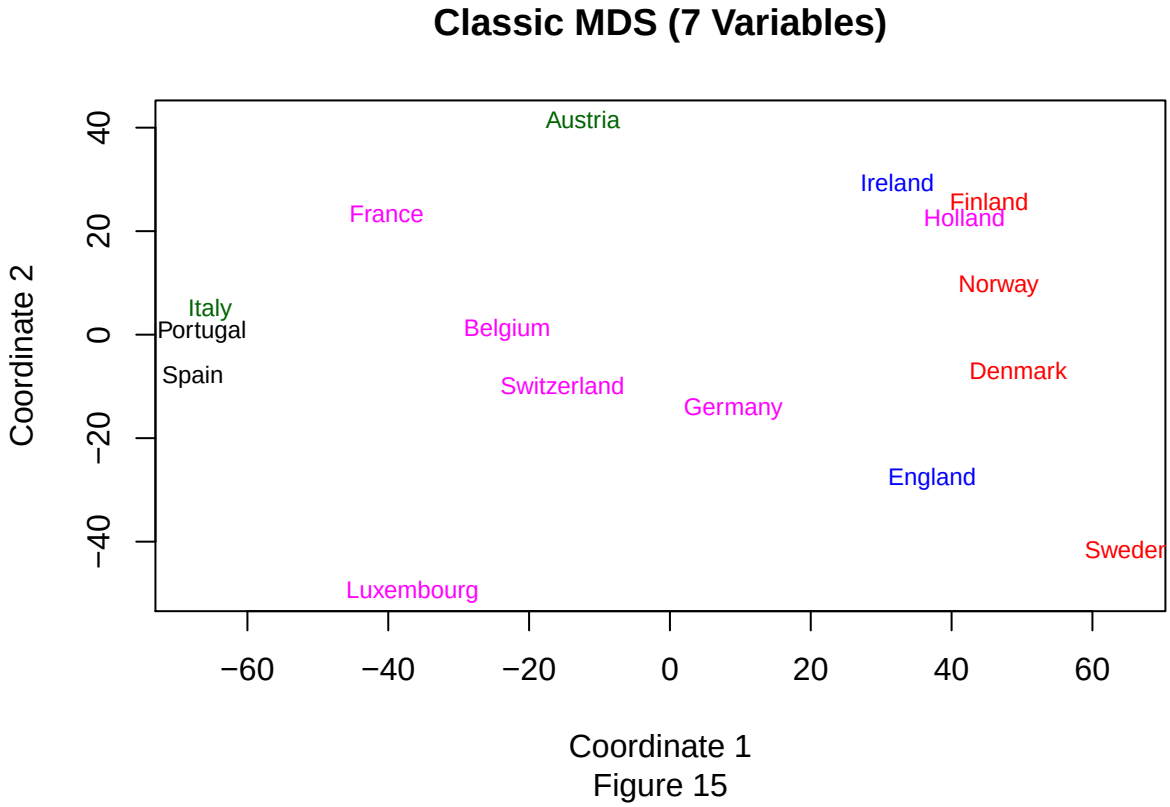


Figure 15

This map is actually still very accurate, considering we are only using 7 variables from our original 20. If we choose less than the seven most heavily weighted, our map's accuracy is greatly reduced. When choosing more than seven, the map's accuracy is not significantly better.

Conclusion:

Through this analysis, were able to construct an acurrate European map using only seven of the original 20 variables. PCA and MDS were two major techniques used. In future studies regarding Europe and Scandanavia, we could greatly reduce the cost and resources by only gathering data for these seven variables rather than all 20 if we want to capture information about food consumption and geographical location, and still get meaningful data.