PCA Demos

Test Scores (SAS)

Test Scores (SAS)

Simple Statistics					
	test1	test2	test3	test4	test5
Mean	96.25	72.68	58.88	65.71	98.84
StD	2.19	18.54	24.68	19.28	0.83

Tests 1 and 5 have very little variability. If we use correlation PCA, we are effectively equating the variance of each test by standardizing them to have variance $= 1$.

This means our PCA will try to capture the variance in test1 and test5 with the same vigor that it tries to capture the variance in other tests.

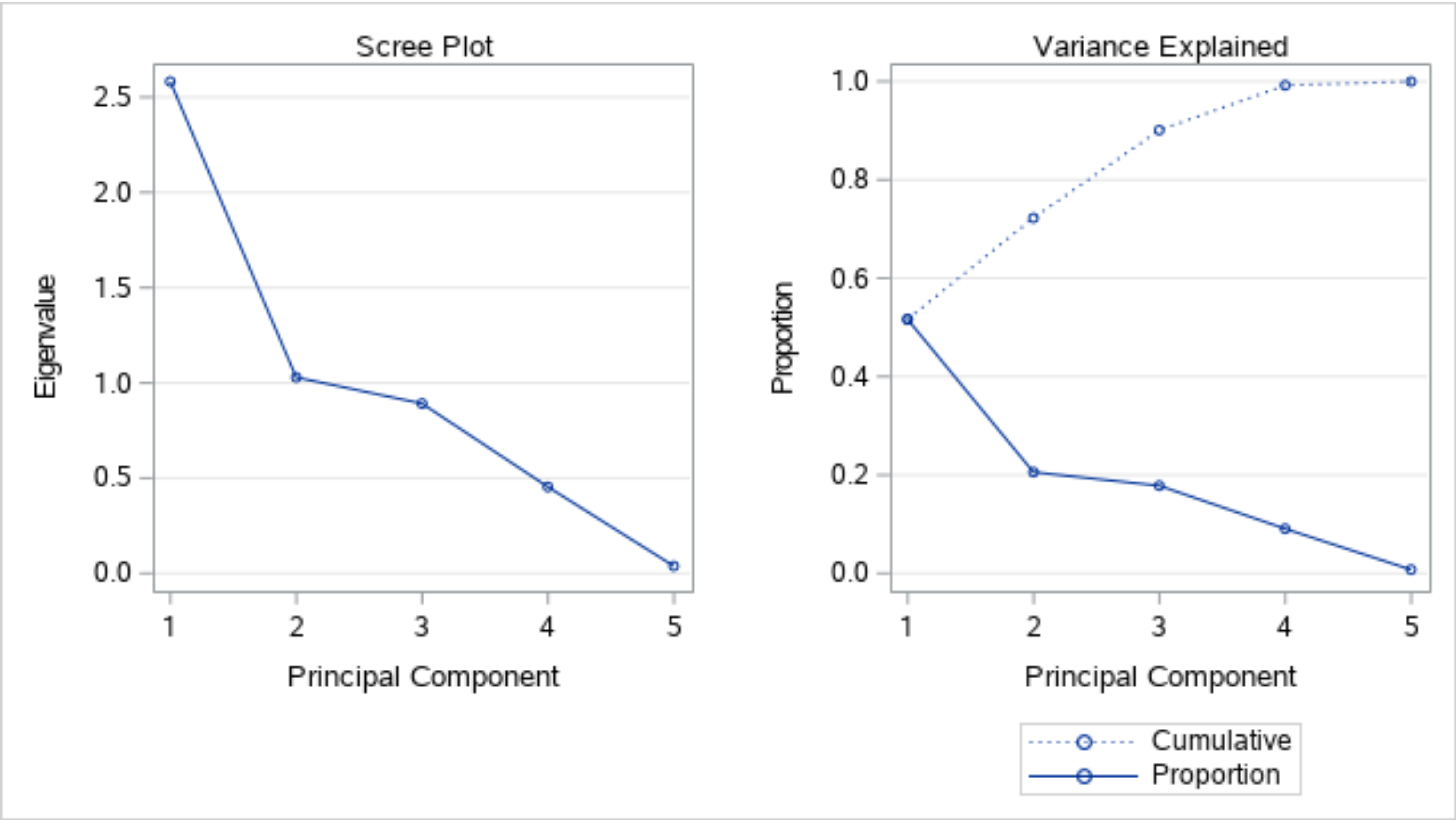
Is this what we want? Do all tests *actually* contribute equally to the total variability in the dataset? Should we want to force them to?

Eigenvalues/Eigenvectors

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1.00	2.58	1.55	0.52	0.52
2.00	1.03	0.14	0.21	0.72
3.00	0.89	0.44	0.18	0.90
4.00	0.45	0.42	0.09	0.99
5.00	0.04		0.01	1.00

Eigenvectors					
	Prin1	Prin2	Prin3	Prin4	Prin5
test1	0.12	-0.84	0.51	0.16	-0.02
test2	0.53	0.15	-0.09	0.73	0.40
test3	0.54	0.07	0.22	-0.65	0.48
test4	0.60	0.15	0.09	-0.03	-0.78
test5	-0.22	0.50	0.82	0.15	0.01

ScreePlots


Scree Plot
Eigenvalue
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
Principal Component
Variance Explained
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
Principal Component
Cumulative
Proportion

Scores/Coordinates in Output Dataset

test1	test2	test3	test4	test5	Prin1	Prin2	Prin3	Prin4	Prin5
96	99	87	92	98	2.4038309154	0.085085466	-0.654944076	0.0753265065	0.0546923609
96	96	91	100	[illegible]	2.6566794704	0.1348273849	-0.568582075	-0.160294559	-0.254039602
100	95	91	86	99	2.1328104259	-0.904282268	1.3008725376	0.3014983566	0.268022596
96	92	93	94	99	2.128961296	0.6674173704	0.4372673519	-0.175696416	-0.044560133
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.5300429461	2.3902914062	0.7162378398	-0.229828769	-0.409606229
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.2428011629	0.8080564864	-0.988020021	-0.70791829	-0.271798135
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.6876559127	1.2056345731	1.5104424971	-0.443508608	0.1174028267
96	86	95	89	98	2.11339751	-0.02078209	-0.536524828	-0.639914588	0.050142591
97	95	99	100	98	2.8562548475	-0.232386237	-0.260799154	-0.336621647	-0.127521435
95	92	88	90	98	2.1098219889	0.3975601352	-0.854967404	-0.295632444	0.0116293846
100	87	98	88	98	2.3891241998	-1.539627724	0.4127421529	-0.383911192	0.1356199642
95	97	85	99	98	2.4694420353	0.4999114317	-0.863529711	-0.035054058	-0.301019701
96	94	89	96	98	2.4303253864	0.0816396714	-0.594949581	-0.179604094	-0.175714904
95	87	99	100	100	1.9834808735	1.6791531562	1.3052570358	-0.42829456	-0.252556327
98	93	85	89	98	2.2011425493	-0.756464696	-0.191216217	0.0448426502	-0.011862408
99	85	92	91	99	1.9728261331	-0.561288302	1.1476566814	-0.198133564	-0.122059521
98	93	96	92	98	2.5357540027	-0.70164567	-0.081284263	-0.249662522	0.0833615846

how do you get this number?

Scores/Coordinates in Output Dataset

test1	test2	test3	test4	test5	Prin1	Prin2	Prin3	Prin4	Prin5
96	99	87	92	98	$\boxed{2.4038309154}$	0.085085466	-0.654944076	0.0753265065	0.0546923609
96	96	91	100	98	2.6566794704	0.1348273849	-0.568582075	-0.160294559	-0.254039602
100	95	91	86	99					0.268022596
96	92	93	94	99					-0.044560133
93	87	86	98	100					-0.409606229
94	87	93	97	98					-0.271798135
96	85	100	90	100					0.1174028267
96	86	95	89	98	2.11339751	-0.02078209	-0.536524828	-0.6399145	
97	95	99	100	98	2.8562548475	-0.232386237	-0.260799154	-0.3366216	
95	92	88	90	98	2.1098219889	0.3975601352	-0.854967404	-0.2956324	
100	87	98	88	98	2.3891241998	-1.539627724	0.4127421529	-0.3839111	
95	97	85	99	98	2.4694420353	0.4999114317	-0.863529711	-0.0350540	
96	94	89	96	98	2.4303253864	0.0816396714	-0.594949581	-0.1796040	
95	87	99	100	100	1.9834808735	1.6791531562	1.3052570358	0.428294	

Simple Statistics

	test1	test2	test3	test4	test5
Mean	96.25	72.68	58.88	65.71	98.84
StD	2.19	18.54	24.68	19.28	0.83

Eigenvectors

Prin1
0.12
0.53
0.54
0.60
-0.22

$$\boxed{0.12\left(\frac{96-96.25}{2.19}\right)+0.53\left(\frac{99-72.68}{18.54}\right)+0.54\left(\frac{87-58.88}{24.68}\right)+0.60\left(\frac{92-65.71}{19.28}\right)-0.22\left(\frac{98-98.84}{0.83}\right)}$$

Scores/Coordinates in Output Dataset

test1	test2	test3	test4	test5	Prin1	Prin2	Prin3	Prin4	Prin5
96	99	87	92	98	2.4038309154	0.085085466	-0.654944076	0.0753265065	0.0546923609
96	96	91	100	98	2.6566794704	0.1348273849	-0.568582075	-0.160294559	-0.254039602
100	95	91	86	99					0.268022596
96	92	93	94	99					-0.044560133
93	87	86	98	100					-0.409606229
94	87	93	97	98					-0.271798135
96	85	100	90	100					0.1174028267
96	86	95	89	98	2.11339751	-0.02078209	-0.536524828	-0.63991458	
97	95	99	100	98	2.8562548475	-0.232386237	-0.260799154	-0.3366216	
95	92	88	90	98	2.1098219889	0.3975601352	-0.854967404	-0.2956324	
100	87	98	88	98	2.3891241998	-1.539627724	0.4127421529	-0.3839111	
95	97	85	99	98	2.4694420353	0.4999114317	-0.863529711	-0.0350540	
96	94	89	96	98	2.4303253864	0.0816396714	-0.594949581	-0.1796040	
95	87	99	100	100	1.9834808735	1.6791531562	1.3052570358	0.428294	

Simple Statistics

	test1	test2	test3	test4	test5
Mean	96.25	72.68	58.88	65.71	98.84
StD	2.19	18.54	24.68	19.28	0.83

Eigenvectors

Prin1
0.12
0.53
0.54
0.60
-0.22

$$0.12\left(\frac{96-96.25}{2.19}\right)+0.53\left(\frac{99-72.68}{18.54}\right)+0.54\left(\frac{87-58.88}{24.68}\right)+0.60\left(\frac{92-65.71}{19.28}\right)-0.22\left(\frac{98-98.84}{0.83}\right)$$

Scores/Coordinates in Output Dataset

test1	test2	test3	test4	test5	Prin1	Prin2	Prin3	Prin4	Prin5
96	99	87	92	98	2.4038309154	0.085085466	-0.654944076	0.0753265065	0.0546923609
96	96	91	100	98	2.6566794704	0.1348273849	-0.568582075	-0.160294559	-0.254039602
100	95	91	86	99	2				0.268022596
96	92	93	94	99					-0.044560133
93	87	86	98	100					-0.409606229
94	87	93	97	98					-0.271798135
96	85	100	90	100					0.1174028267
96	86	95	89	98	2.11339751	-0.02078209	-0.536524828	-0.6399145	
97	95	99	100	98	2.8562548475	-0.232386237	-0.260799154	-0.3366216	
95	92	88	90	98	2.1098219889	0.3975601352	-0.854967404	-0.2956324	
100	87	98	88	98	2.3891241998	-1.539627724	0.4127421529	-0.3839111	
95	97	85	99	98	2.4694420353	0.4999114317	-0.863529711	-0.0350540	
96	94	89	96	98	2.4303253864	0.0816396714	-0.594949581	-0.1796040	
95	87	99	100	100	1.9834808735	1.6791531562	1.3052570358	0.428294	

Simple Statistics

	test1	test2	test3	test4	test5
Mean	96.25	72.68	58.88	65.71	98.84
StD	2.19	18.54	24.68	19.28	0.83

Eigenvectors

Prin1
0.12
0.53
0.54
0.60
-0.22

$$0.12\left(\frac{96-96.25}{2.19}\right)+0.53\left(\frac{99-72.68}{18.54}\right)+0.54\left(\frac{87-58.88}{24.68}\right)+0.60\left(\frac{92-65.71}{19.28}\right)-0.22\left(\frac{98-98.84}{0.83}\right)$$

Scores/Coordinates in Output Dataset

test1	test2	test3	test4	test5	Prin1	Prin2	Prin3	Prin4	Prin5
96	99	87	92	98	2.4038309154	0.085085466	-0.654944076	0.0753265065	0.0546923609
96	96	91	100	98	2.6566794704	0.1348273849	-0.568582075	-0.160294559	-0.254039602
100	95	91	86	99					0.268022596
96	92	93	94	99					-0.044560133
93	87	86	98	100					-0.409606229
94	87	93	97	98					-0.271798135
96	85	100	90	100					0.1174028267
96	86	95	89	98	2.11339751	-0.02078209	-0.536524828	-0.6399145	
97	95	99	100	98	2.8562548475	-0.232386237	-0.260799154	-0.3366216	
95	92	88	90	98	2.1098219889	0.3975601352	-0.854967404	-0.2956324	
100	87	98	88	98	2.3891241998	-1.539627724	0.4127421529	-0.383911	
95	97	85	99	98	2.4694420353	0.4999114317	-0.863529711	-0.0350540	
96	94	89	96	98	2.4303253864	0.0816396714	-0.594949581	-0.1796040	
95	87	99	100	100	1.9834808735	1.6791531562	1.3052570358	0.428294	

Simple Statistics

	test1	test2	test3	test4	test5
Mean	96.25	72.68	58.88	65.71	98.84
StD	2.19	18.54	24.68	19.28	0.83

Eigenvectors

Prin1
0.12
0.53
0.54
0.60
-0.22

$$0.12\left(\frac{96-96.25}{2.19}\right)+0.53\left(\frac{99-72.68}{18.54}\right)+0.54\left(\frac{87-58.88}{24.68}\right)+0.60\left(\frac{92-65.71}{19.28}\right)-0.22\left(\frac{98-98.84}{0.83}\right)$$

Scores/Coordinates in Output Dataset

test1	test2	test3	test4	test5	Prin1	Prin2	Prin3	Prin4	Prin5
96	99	87	92	98	2.4038309154	0.085085466	-0.654944076	0.0753265065	0.0546923609
96	96	91	100	98	2.6566794704	0.1348273849	-0.568582075	-0.160294559	-0.254039602
100	95	91	86	99					0.268022596
96	92	93	94	99					-0.044560133
93	87	86	98	100					-0.409606229
94	87	93	97	98					-0.271798135
96	85	100	90	100					0.1174028267
96	86	95	89	98	2.11339751	-0.02078209	-0.536524828	-0.6399145	
97	95	99	100	98	2.8562548475	-0.232386237	-0.260799154	-0.3366216	
95	92	88	90	98	2.1098219889	0.3975601352	-0.854967404	-0.2956324	
100	87	98	88	98	2.3891241998	-1.539627724	0.4127421529	-0.383911	
95	97	85	99	98	2.4694420353	0.4999114317	-0.863529711	-0.0350540	
96	94	89	96	98	2.4303253864	0.0816396714	-0.594949581	-0.1796040	
95	87	99	100	100	1.9834808735	1.6791531562	1.3052570358	0.428294	

Simple Statistics

	test1	test2	test3	test4	test5
Mean	96.25	72.68	58.88	65.71	98.84
StD	2.19	18.54	24.68	19.28	0.83

Eigenvectors

Prin1
0.12
0.53
0.54
0.60
-0.22

$$0.12\left(\frac{96-96.25}{2.19}\right)+0.53\left(\frac{99-72.68}{18.54}\right)+0.54\left(\frac{87-58.88}{24.68}\right)+0.60\left(\frac{92-65.71}{19.28}\right)-0.22\left(\frac{98-98.84}{0.83}\right)$$

How would this formula be different for covariance PCA?

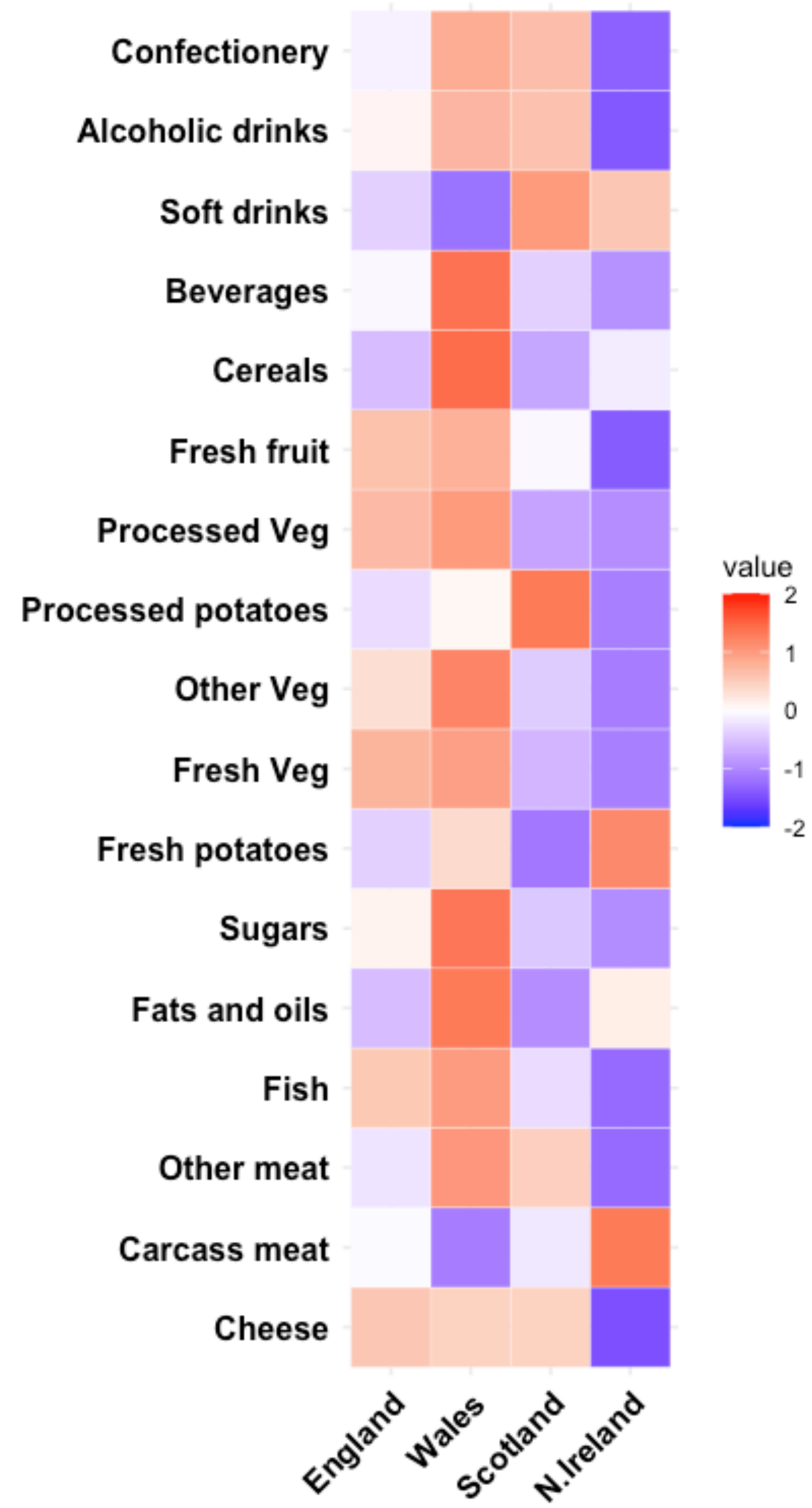
UK Food Consumption
(Learning the BiPlot)

Average Consumption of Certain Food/Drink across the UK (Grams, Per Person, Per Week)

	England	Wales	Scotland	N Ireland
Cheese	105	103	103	66
Carcass meat	245	227	242	267
Other meat	685	803	750	586
Fish	147	160	122	93
Fats and oils	193	235	184	209
Sugars	156	175	147	139
Fresh potatoes	720	874	566	1033
Fresh Veg	253	265	171	143
Other Veg	488	570	418	355
Processed potatoes	198	203	220	187
Processed Veg	360	365	337	334
Fresh fruit	1102	1137	957	674
Cereals	1472	1582	1462	1494
Beverages	57	73	53	47
Soft drinks	1374	1256	1572	1506
Alcoholic drinks	375	475	458	135
Confectionery	54	64	62	41

From '*Department for Environment, Food and Rural Affairs*' (DEFRA) in 1997.

Standardizing each row and observing a heat map we can clearly see how the countries are different in their consumption. The PCA BiPlot gives us another visual to reveal the same information



We'll opt for the transpose, where the observations are the four countries within the UK.

```
food=read.csv("http://birch.iaa.ncsu.edu/~slrace/LinearAlgebra2021/Code/ukfood.csv",
              header=TRUE,row.names=1)
food=as.data.frame(t(food))
head(food)
```

```
          Cheese Carcass meat Other meat Fish Fats and oils Sugars
England      105          245        685  147           193    156
Wales        103          227        803  160           235    175
Scotland     103          242        750  122           184    147
N.Ireland     66          267        586   93           209    139
          Fresh potatoes Fresh Veg Other Veg Processed potatoes Processed Veg
England              720       253       488                198           360
Wales                874       265       570                203           365
Scotland             566       171       418                220           337
N.Ireland           1033       143       355                187           334
          Fresh fruit Cereals Beverages Soft drinks Alcoholic drinks
England          1102    1472        57        1374              375
Wales            1137    1582        73        1256              475
Scotland          957    1462        53        1572              458
N.Ireland         674    1494        47        1506              135
          Confectionery
England              54
Wales                64
Scotland             62
N.Ireland            41
```

We then compute the PCA and observe the output object in our environment:

```
pca=prcomp(food, scale = T)
```

pca	List of 5
sdev : num [1:4] 3.41 2.06 1.08 6.34e-16	
rotation: num [1:17, 1:4] 0.246 -0.286 0.265 0.286 0.127 ...	
..- attr(*, "dimnames")=List of 2	
.. ..$: chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...	
.. ..$: chr [1:4] "PC1" "PC2" "PC3" "PC4"	
center : Named num [1:17] 94.2 245.2 706 130.5 205.2 ...	
..- attr(*, "names")= chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...	
scale : Named num [1:17] 18.9 16.5 93.4 29.6 22.4 ...	
..- attr(*, "names")= chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...	
x : num [1:4, 1:4] 0.827 3.915 -0.423 -4.319 0.284 ...	
..- attr(*, "dimnames")=List of 2	
.. ..$: chr [1:4] "England" "Wales" "Scotland" "N.Ireland"	
.. ..$: chr [1:4] "PC1" "PC2" "PC3" "PC4"	
- attr(*, "class")= chr "prcomp"	

We then compute the PCA and observe the output object in our environment:

```
pca=prcomp(food, scale = T)
```

list of 4 numbers

```
pca                                                    List of 5
  sdev : num [1:4] 3.41 2.06 1.08 6.34e-16
  rotation: num [1:17, 1:4] 0.246 -0.286 0.265 0.286 0.127 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...
  .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
  center : Named num [1:17] 94.2 245.2 706 130.5 205.2 ...
  ..- attr(*, "names")= chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...
  scale : Named num [1:17] 18.9 16.5 93.4 29.6 22.4 ...
  ..- attr(*, "names")= chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...
  x : num [1:4, 1:4] 0.827 3.915 -0.423 -4.319 0.284 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "England" "Wales" "Scotland" "N.Ireland"
  .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
  - attr(*, "class")= chr "prcomp"
```

We then compute the PCA and observe the output object in our environment:

```
pca=prcomp(food, scale = T)
```

17 x 4 matrix

```
pca                    List of 5
  sdev : num [1:4] 3.41 2.06 1.08 6.34e-16
  rotation: num [1:17, 1:4] 0.246 -0.286 0.265 0.286 0.127 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ..
  .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
  center : Named num [1:17] 94.2 245.2 706 130.5 205.2 ...
  ..- attr(*, "names")= chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...
  scale : Named num [1:17] 18.9 16.5 93.4 29.6 22.4 ...
  ..- attr(*, "names")= chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...
  x : num [1:4, 1:4] 0.827 3.915 -0.423 -4.319 0.284 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "England" "Wales" "Scotland" "N.Ireland"
  .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
  - attr(*, "class")= chr "prcomp"
```

We then compute the PCA and observe the output object in our environment:

```
pca=prcomp(food, scale = T)
```

```
pca                List of 5
  sdev : num [1:4] 3.41 2.06 1.08 6.34e-16
  rotation: num [1:17, 1:4] 0.246 -0.286 0.265 0.286 0.127 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...
  .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
  center : Named num [1:17] 94.2 245.2 706 130.5 205.2 ...
  ..- attr(*, "names")= chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...
  scale : Named num [1:17] 18.9 16.5 93.4 29.6 22.4 ...
  ..- attr(*, "names")= chr [1:17] "Cheese" "Carcass meat" "Other meat" "Fish" ...
  x : num [1:4, 1:4] 0.827 3.915 -0.423 -4.319 0.284 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "England" "Wales" "Scotland" "N.Ireland"
  .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
  - attr(*, "class")= chr "prcomp"
```

4 x 4 matrix

If we still need help understanding what these objects are:

?prcomp

1. ***sdev:*** the standard deviations of the principal components (i.e., the **square roots of the eigenvalues** of the covariance/ correlation matrix, though the calculation is actually done with the singular values of the data matrix).

2. ***rotation:*** the matrix of **variable loadings**(i.e., a matrix whose columns contain the eigenvectors). The function princomp returns this in the element loadings.

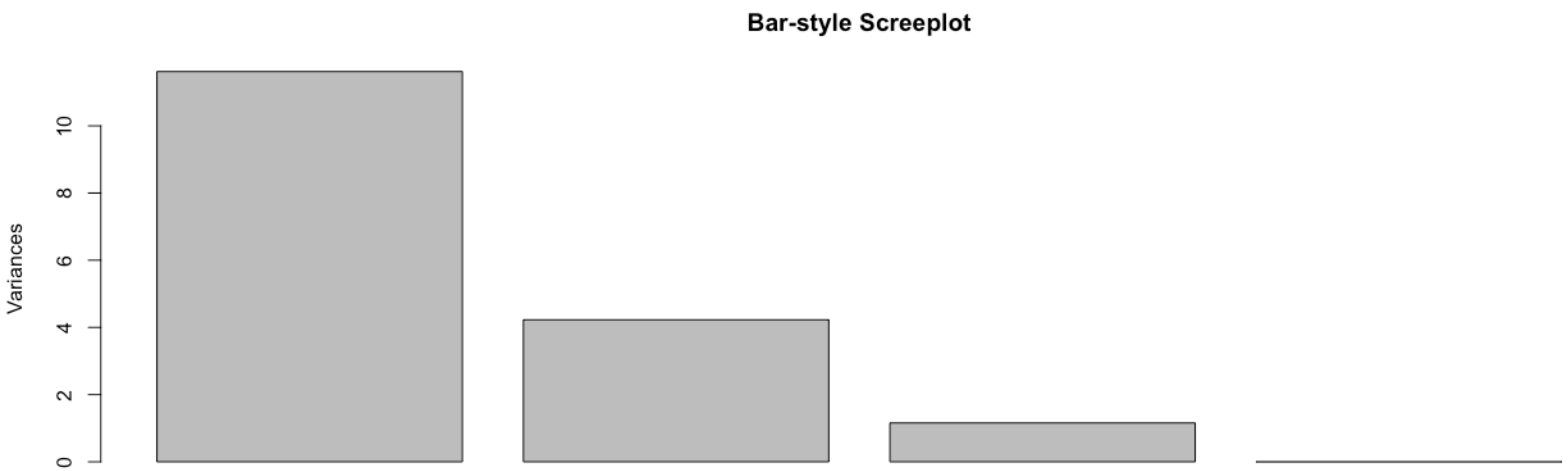
3. ***x:*** the value of the rotated data **(i.e. the scores)**(the centred (and scaled if requested) data multiplied by the rotation matrix) is returned. Hence, cov(x) is the diagonal matrix diag($sdev^2$)

4. center, scale: the centering and scaling used, or FALSE.

The screeplot and % variance explained statistics

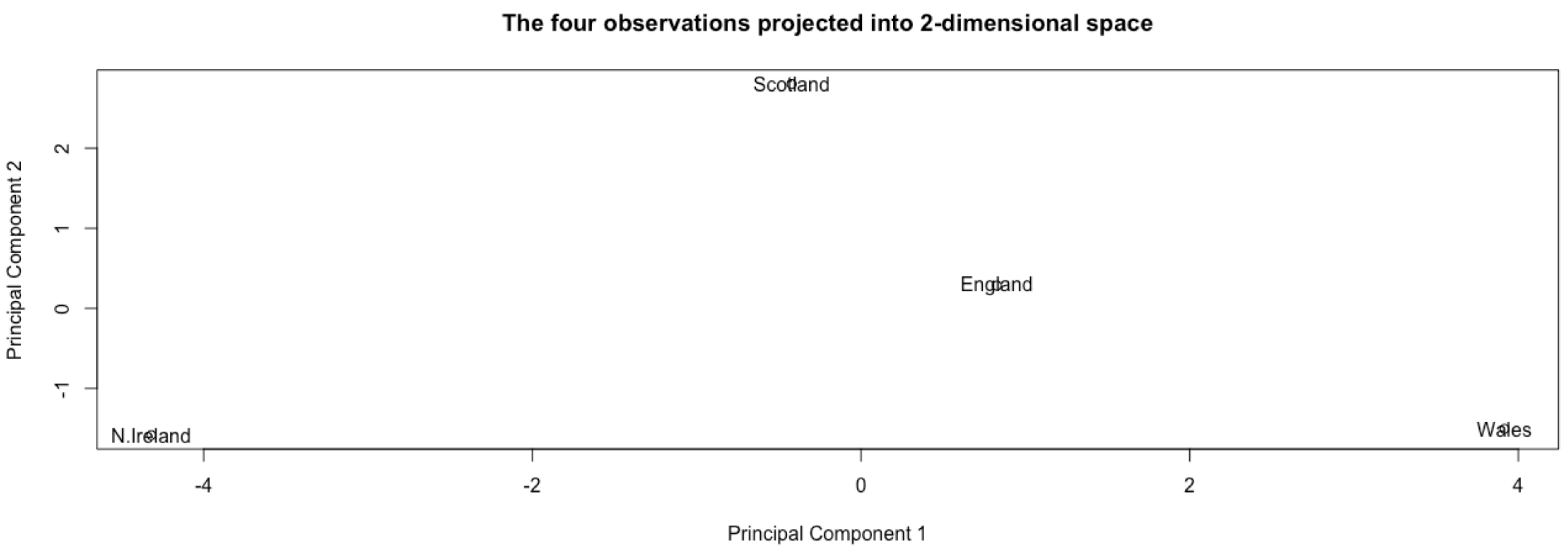
```
summary(pca)
plot(pca, main = "Bar-style Screeplot")
```

```
Importance of components:
                          PC1    PC2     PC3       PC4
Standard deviation     3.4082 2.0562 1.07524 6.344e-16
Proportion of Variance 0.6833 0.2487 0.06801 0.000e+00
Cumulative Proportion  0.6833 0.9320 1.00000 1.000e+00
```



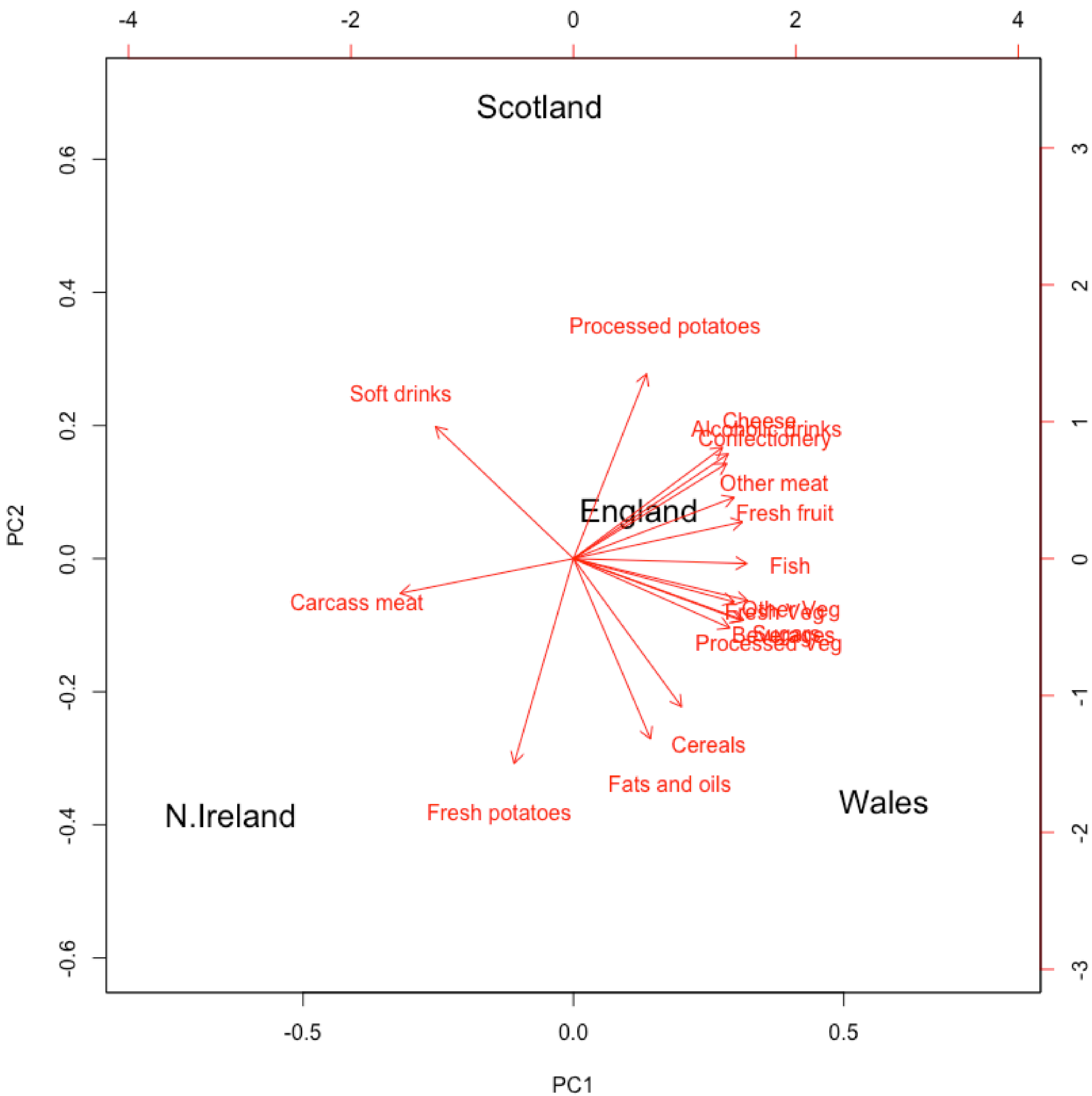
The projection of the data into 2-D (plot of scores)

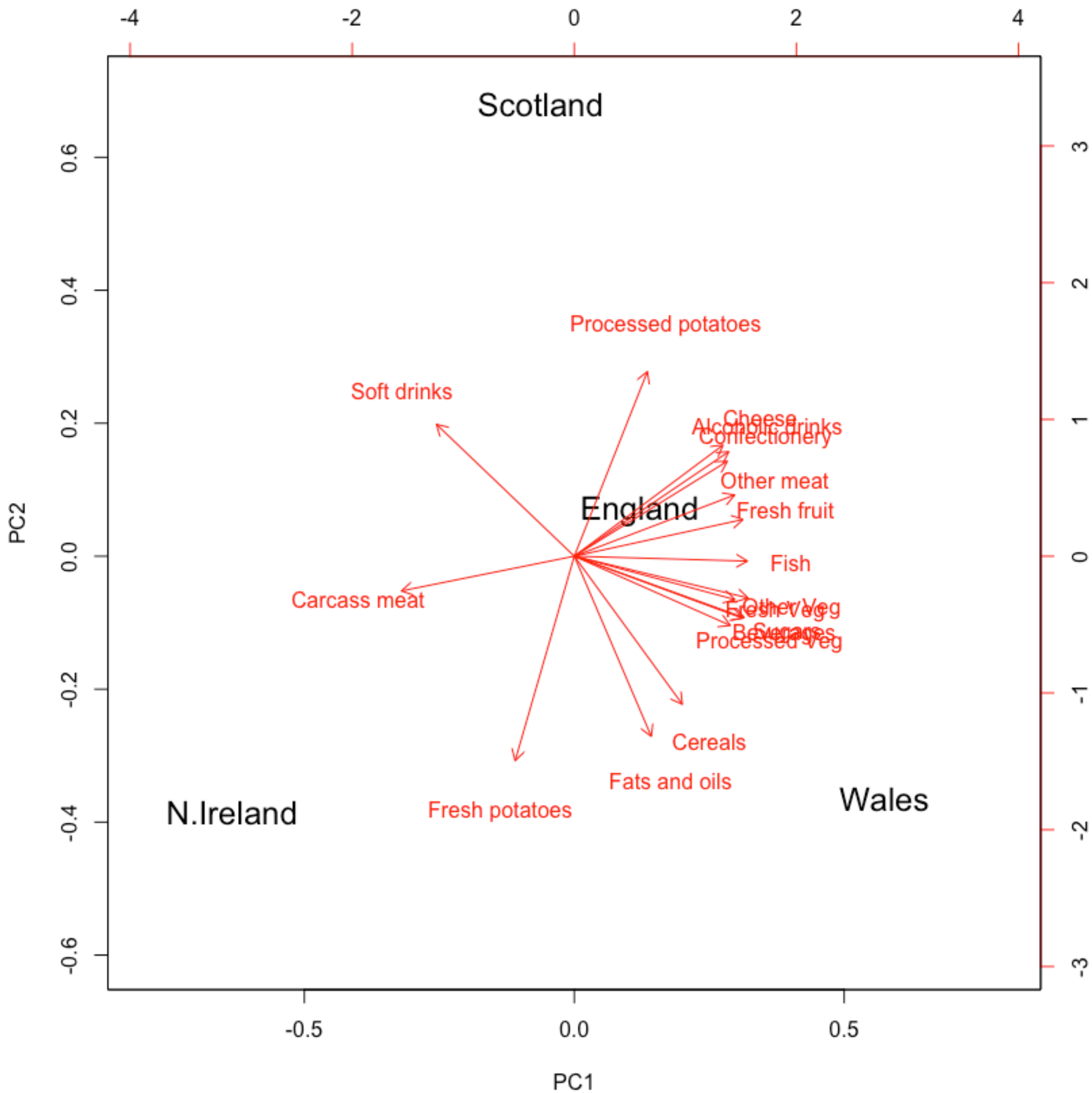
```
plot(pca$x,
     xlab= "Principal Component 1",
     ylab= "Principal Component 2",
     main= 'The four observations projected into 2-dimensional space')
text(pca$x[,1], pca$x[,2],row.names(food))
```



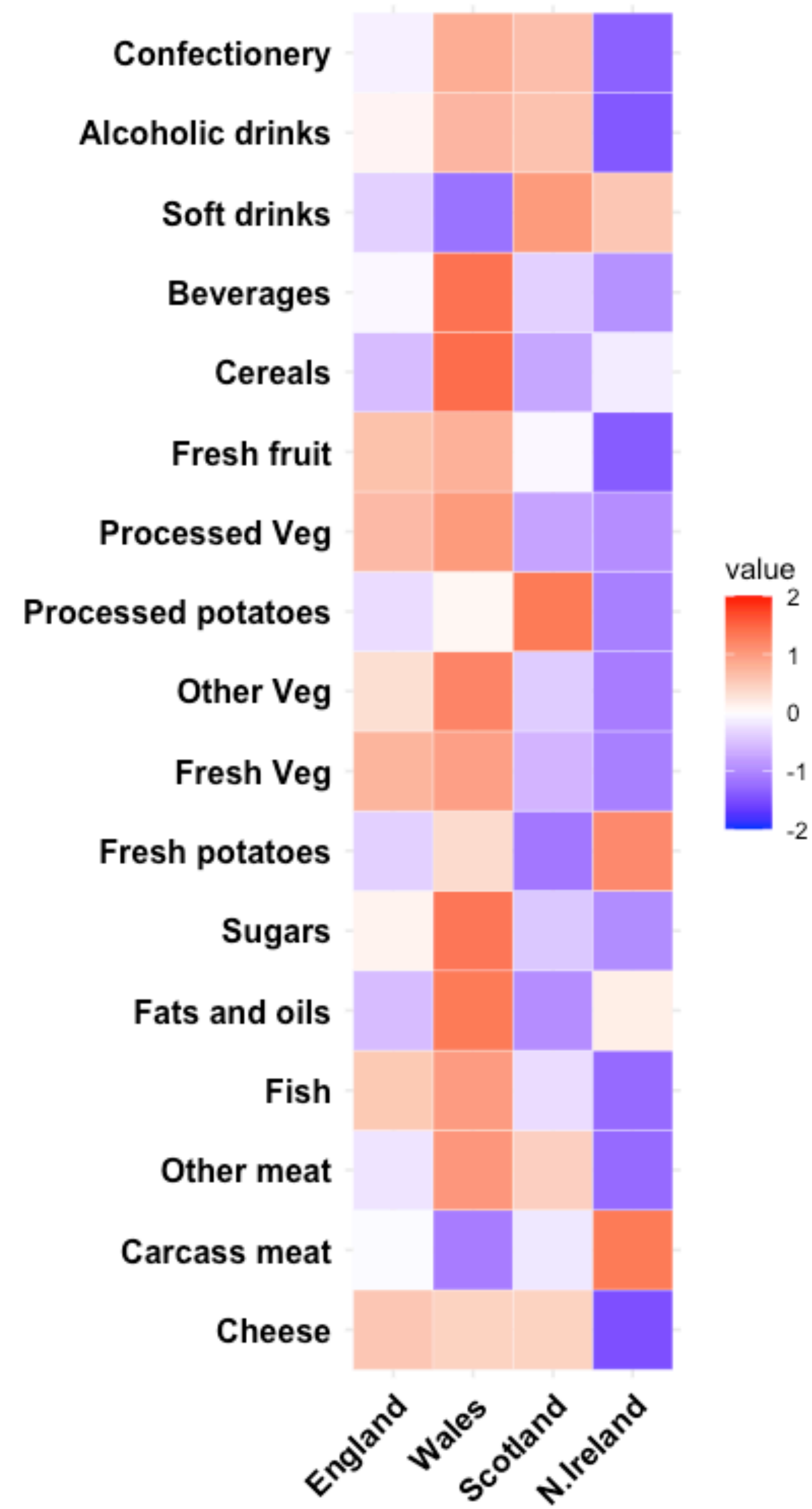
The Biplot

```
biplot(pca, cex = c(1.5, 1), col = c('black','red'),
          xlim = c(-0.8,0.8), ylim = c(-0.6,0.7))
```

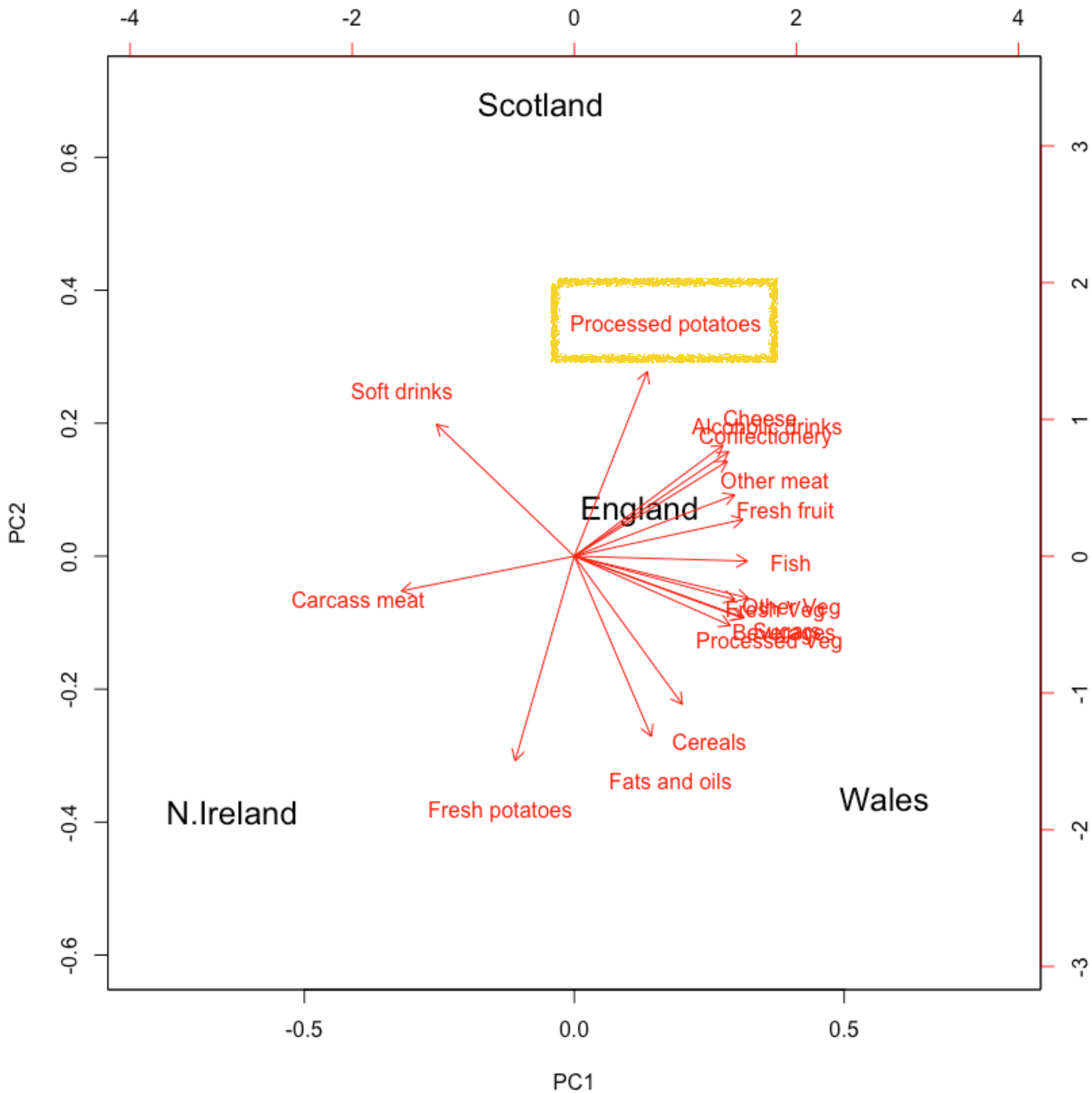




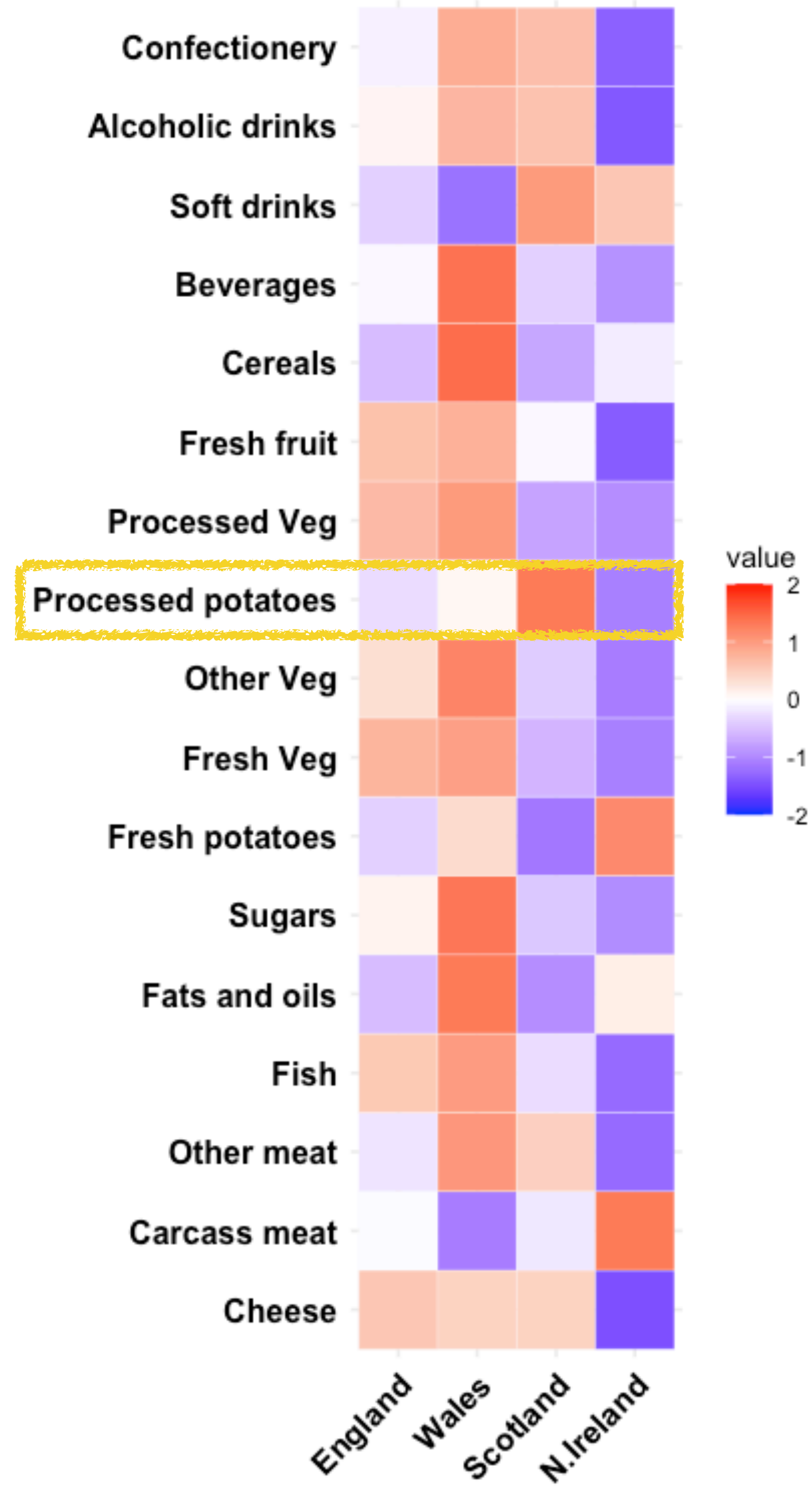
Scotland
England
N.Ireland
Wales
Processed potatoes
Soft drinks
Cheese
Confectionery
Other meat
Fresh fruit
Fish
Carcass meat
Cereals
Fats and oils
Fresh potatoes
PC1
PC2



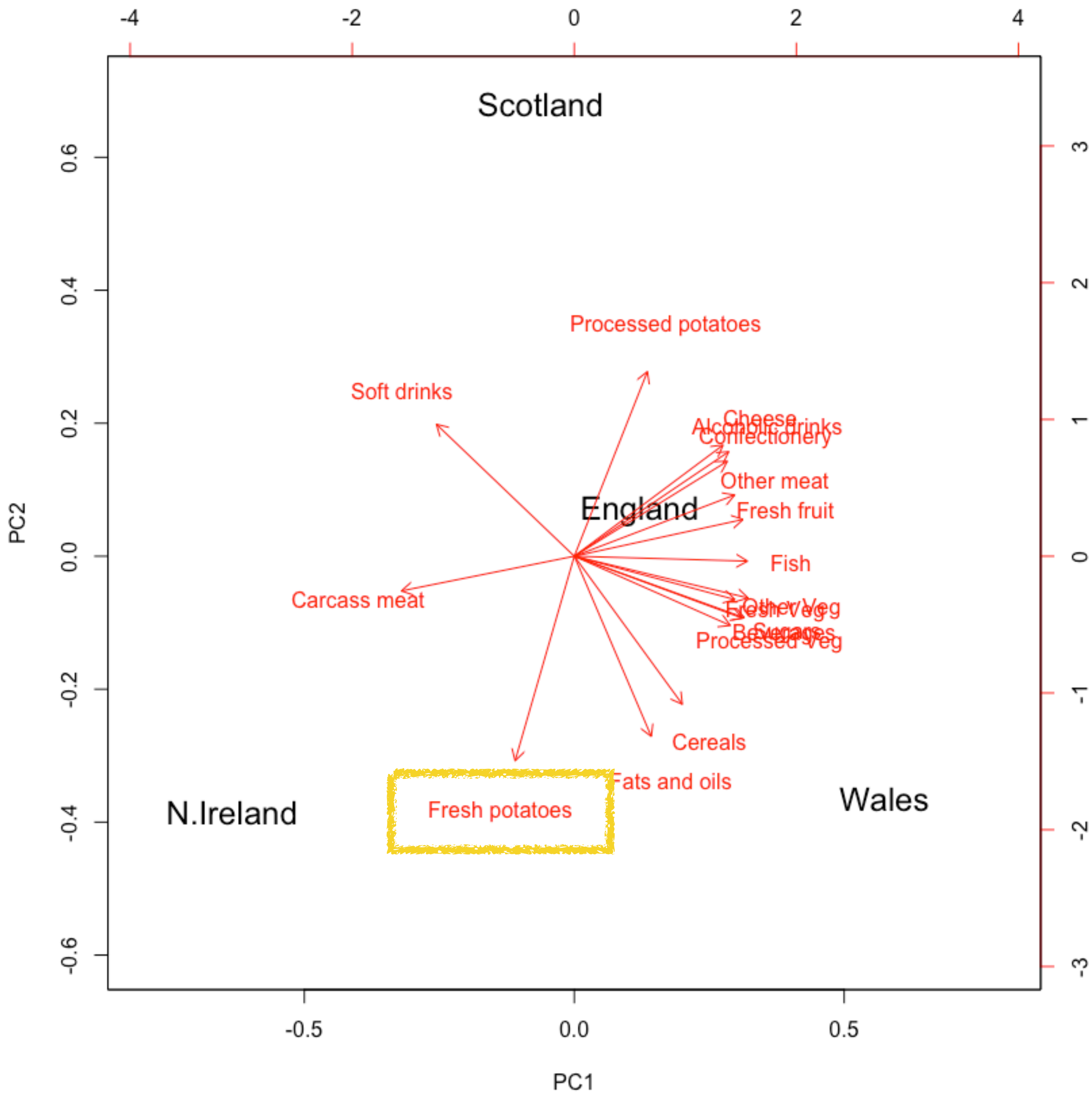
Confectionery
Alcoholic drinks
Soft drinks
Beverages
Cereals
Fresh fruit
Processed Veg
Processed potatoes
Other Veg
Fresh Veg
Fresh potatoes
Sugars
Fats and oils
Fish
Other meat
Carcass meat
Cheese
England
Wales
Scotland
N.Ireland
value

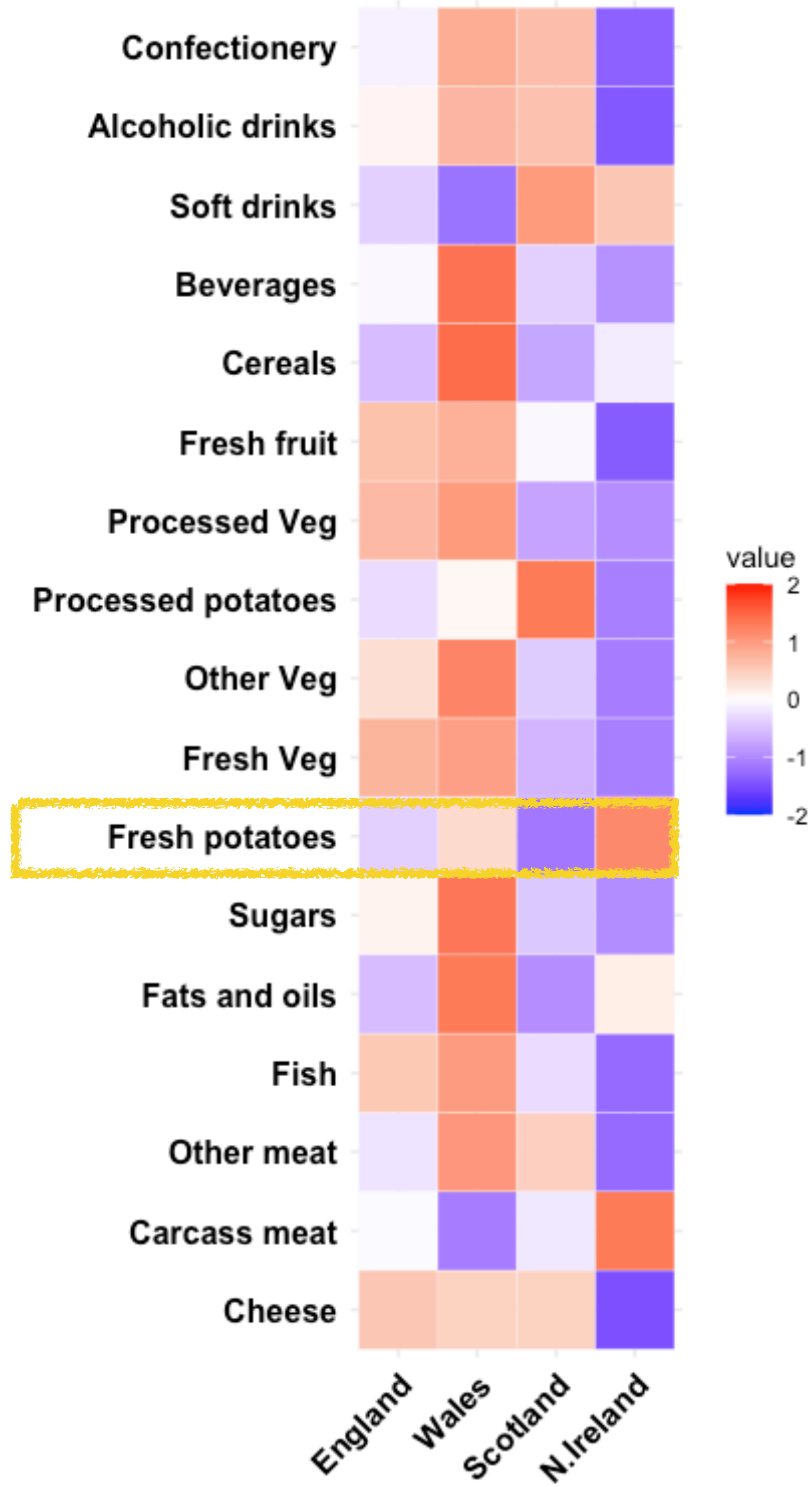


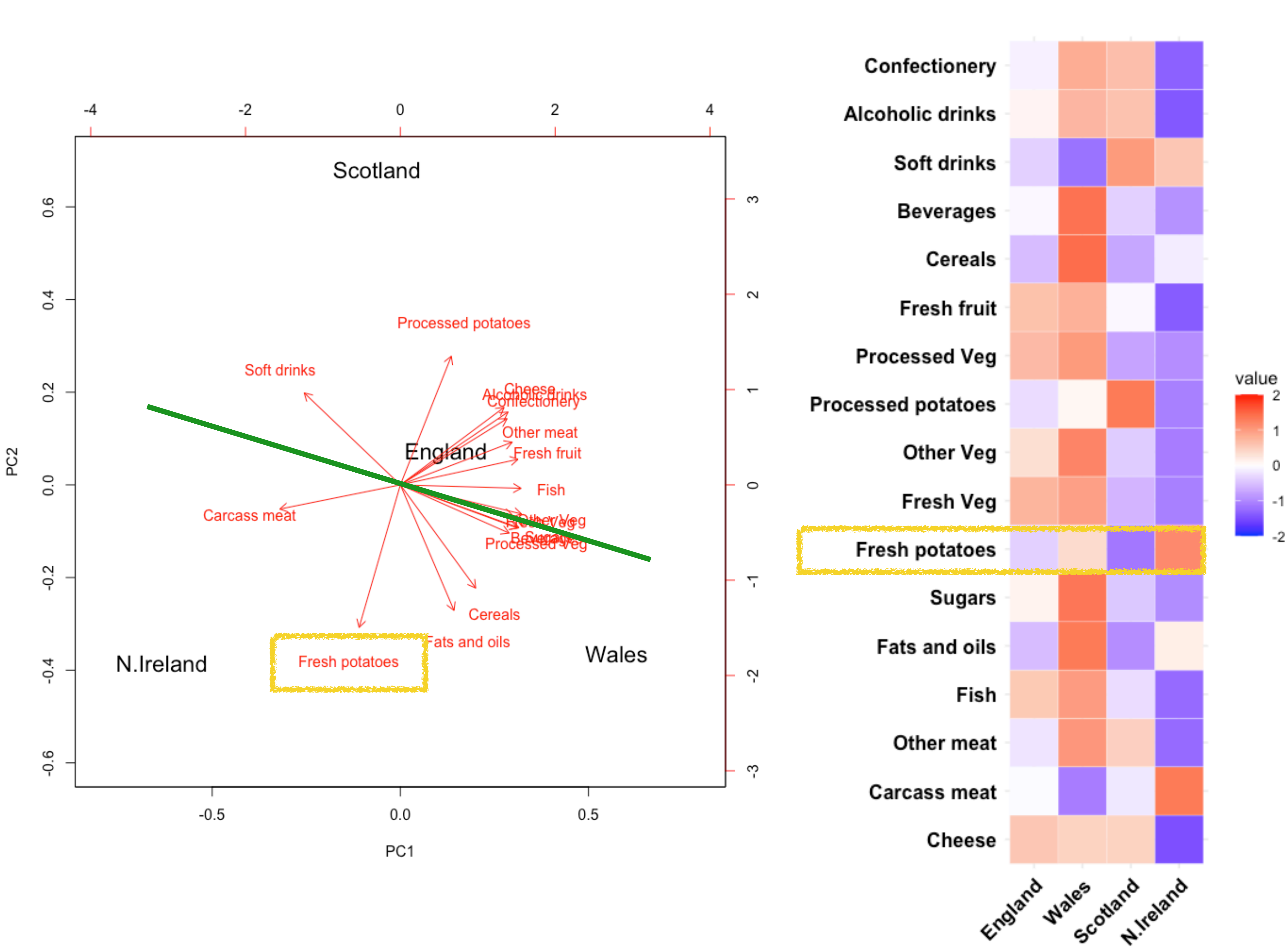
Scotland
England
N.Ireland
Wales
Processed potatoes
Soft drinks
Cheese
Alcoholic drinks
Confectionery
Other meat
Fresh fruit
Fish
Carcass meat
Other Veg
Fresh Veg
Beverages
Sugars
Processed Veg
Cereals
Fats and oils
Fresh potatoes
PC1
PC2

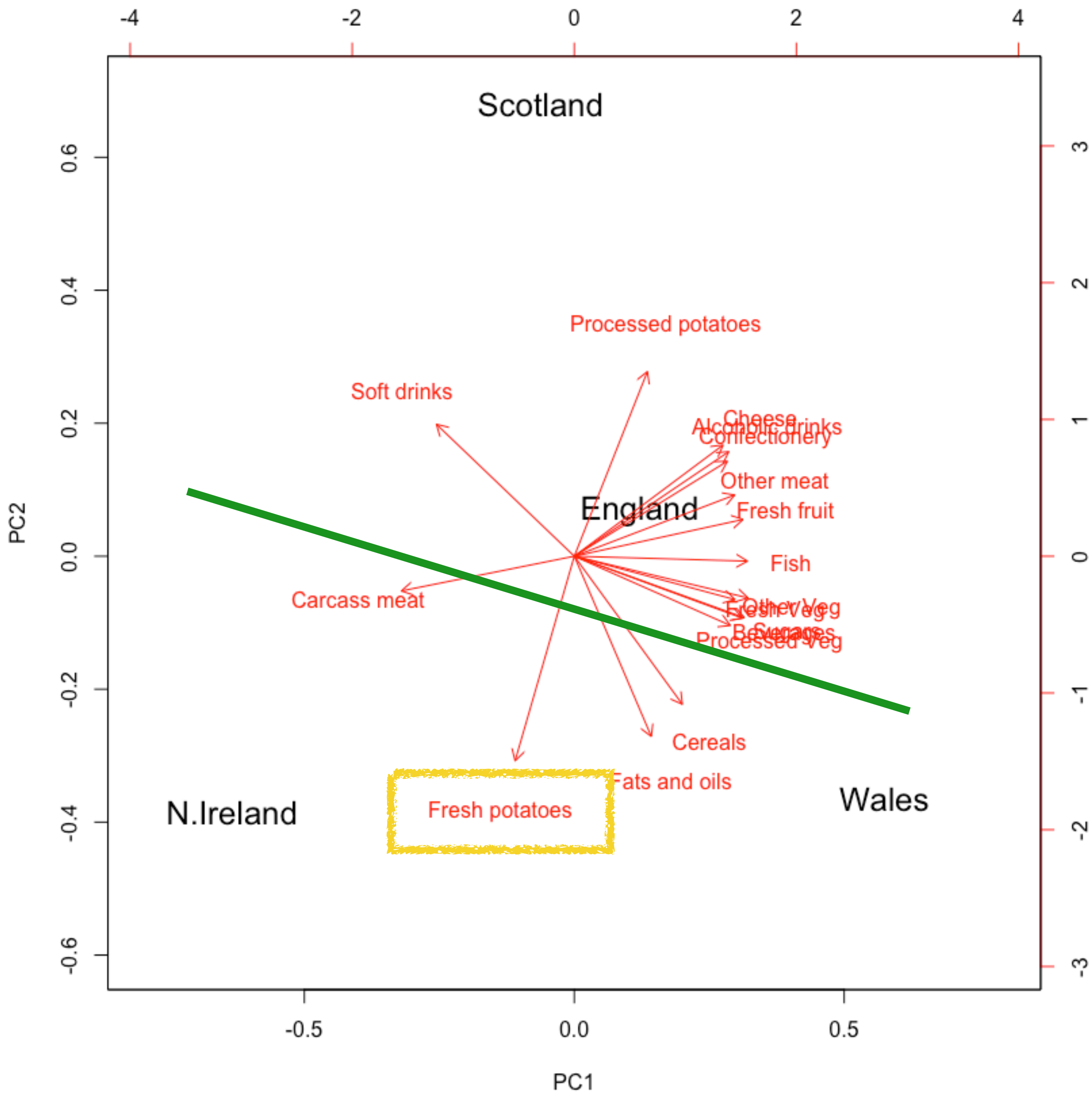


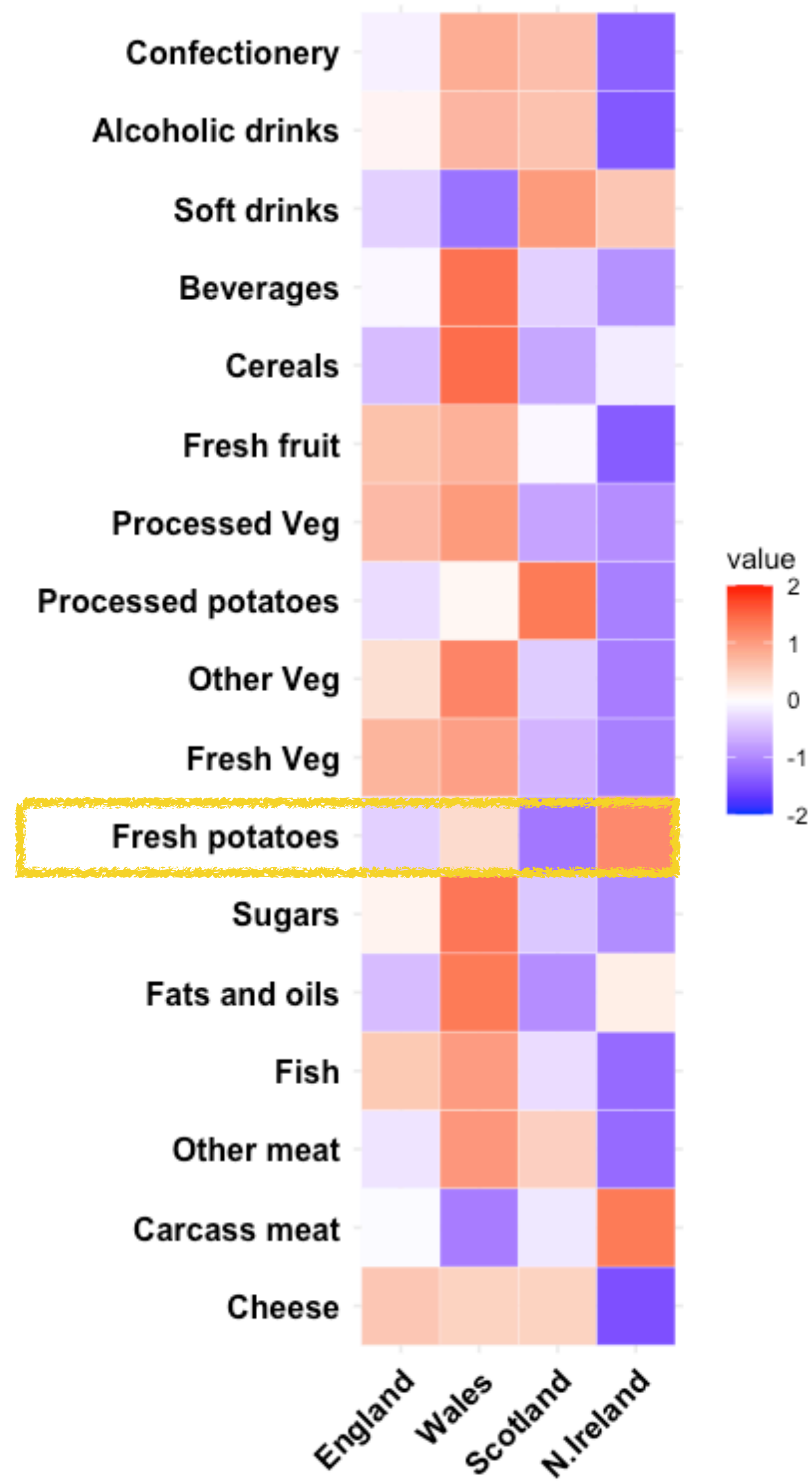
Confectionery
Alcoholic drinks
Soft drinks
Beverages
Cereals
Fresh fruit
Processed Veg
Processed potatoes
Other Veg
Fresh Veg
Fresh potatoes
Sugars
Fats and oils
Fish
Other meat
Carcass meat
Cheese
England
Wales
Scotland
N.Ireland
value

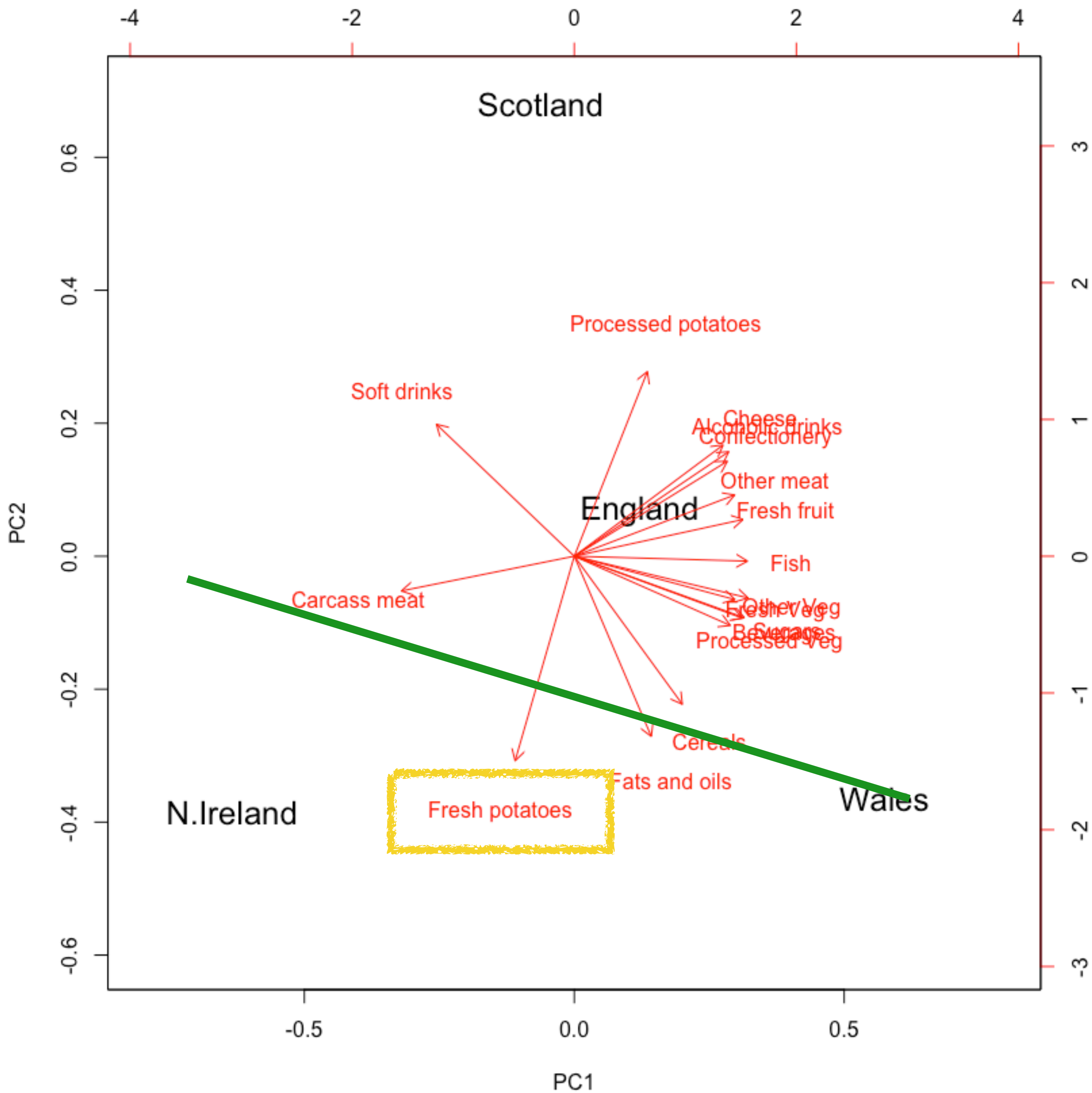

Scotland
England
N.Ireland
Wales
Processed potatoes
Soft drinks
Cheese
Alcoholic drinks
Confectionery
Other meat
Fresh fruit
Fish
Carcass meat
Other Veg
Fresh Veg
Beverages
Sugars
Processed Veg
Cereals
Fats and oils
Fresh potatoes
PC1
PC2


Confectionery
Alcoholic drinks
Soft drinks
Beverages
Cereals
Fresh fruit
Processed Veg
Processed potatoes
Other Veg
Fresh Veg
Fresh potatoes
Sugars
Fats and oils
Fish
Other meat
Carcass meat
Cheese
England
Wales
Scotland
N.Ireland
value

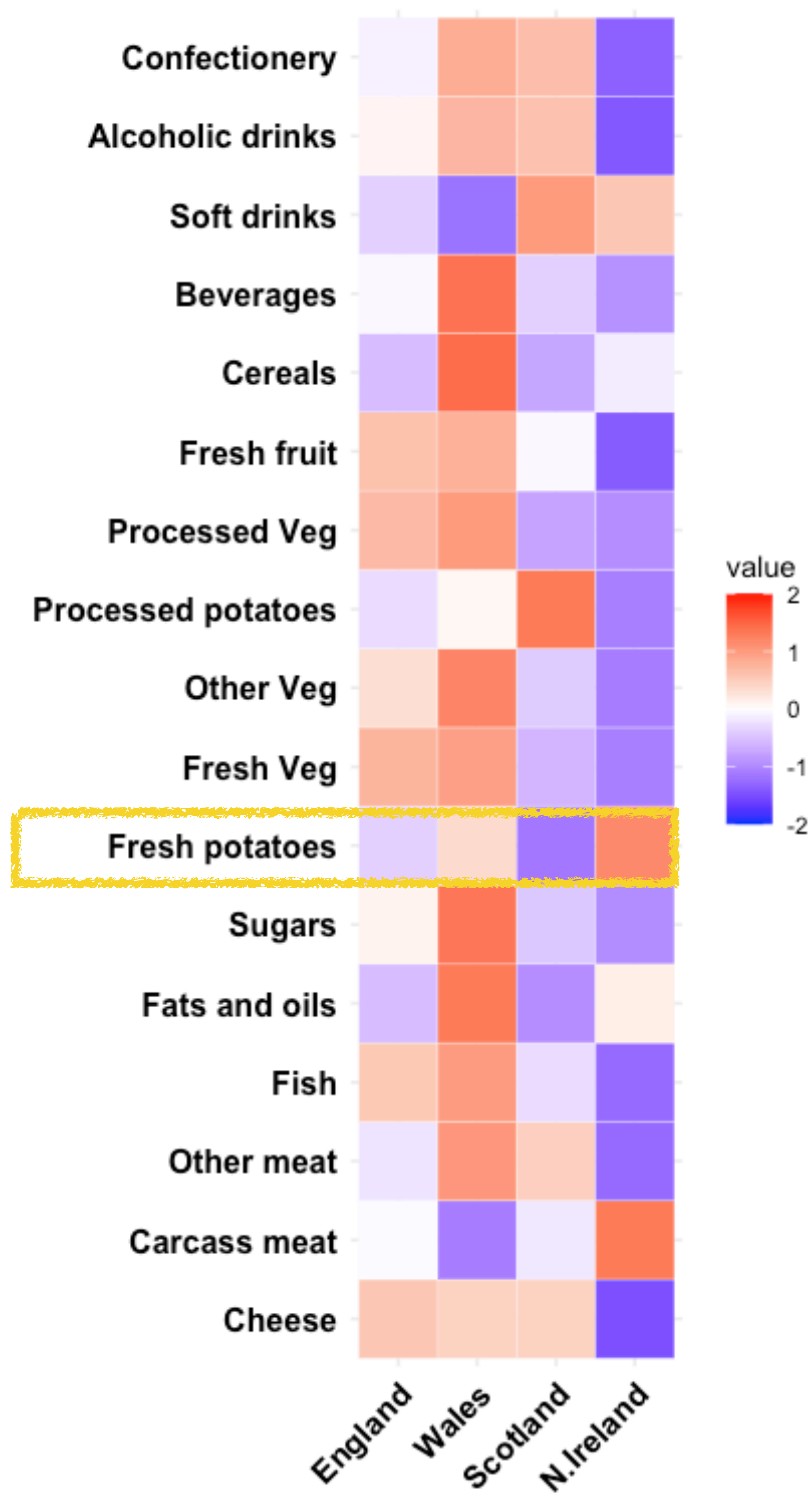

Scotland
England
N.Ireland
Wales
Processed potatoes
Soft drinks
Cheese
Alcoholic drinks
Confectionery
Other meat
Fresh fruit
Fish
Carcass meat
Other Veg
Fresh Veg
Sugars
Beverages
Processed Veg
Cereals
Fats and oils
Fresh potatoes
PC1
PC2
Confectionery
Alcoholic drinks
Soft drinks
Beverages
Cereals
Fresh fruit
Processed Veg
Processed potatoes
Other Veg
Fresh Veg
Fresh potatoes
Sugars
Fats and oils
Fish
Other meat
Carcass meat
Cheese
England
Wales
Scotland
N.Ireland
value


Scotland
England
N.Ireland
Wales
Processed potatoes
Soft drinks
Other meat
Fresh fruit
Fish
Carcass meat
Cereals
Fats and oils
Fresh potatoes
PC1
PC2
-4
-2
0
2
4
-0.5
0.0
0.5
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
3
2
1
0
-1
-2
-3

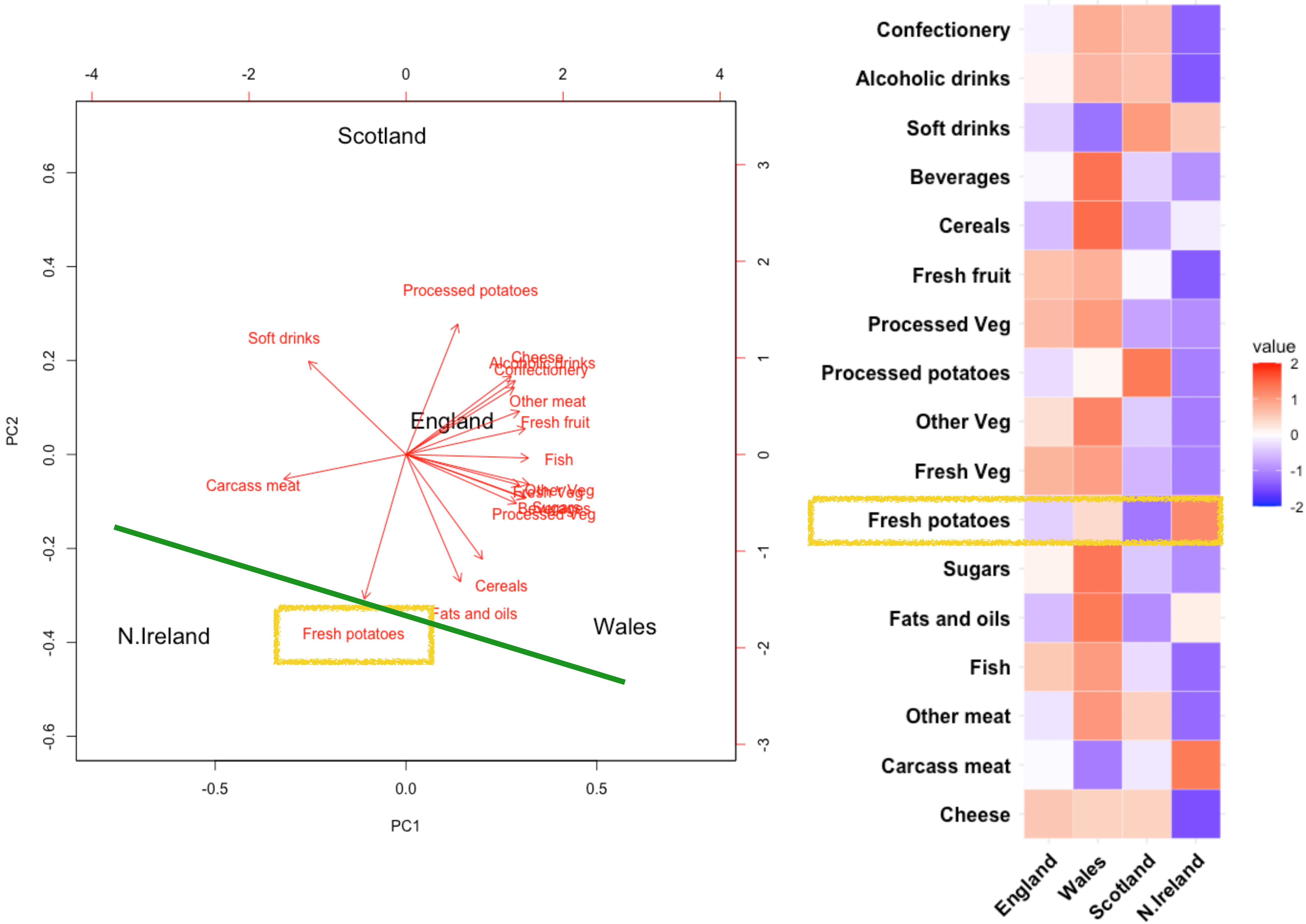

Confectionery
Alcoholic drinks
Soft drinks
Beverages
Cereals
Fresh fruit
Processed Veg
Processed potatoes
Other Veg
Fresh Veg
Fresh potatoes
Sugars
Fats and oils
Fish
Other meat
Carcass meat
Cheese
England
Wales
Scotland
N.Ireland
value
2
1
0
-1
-2



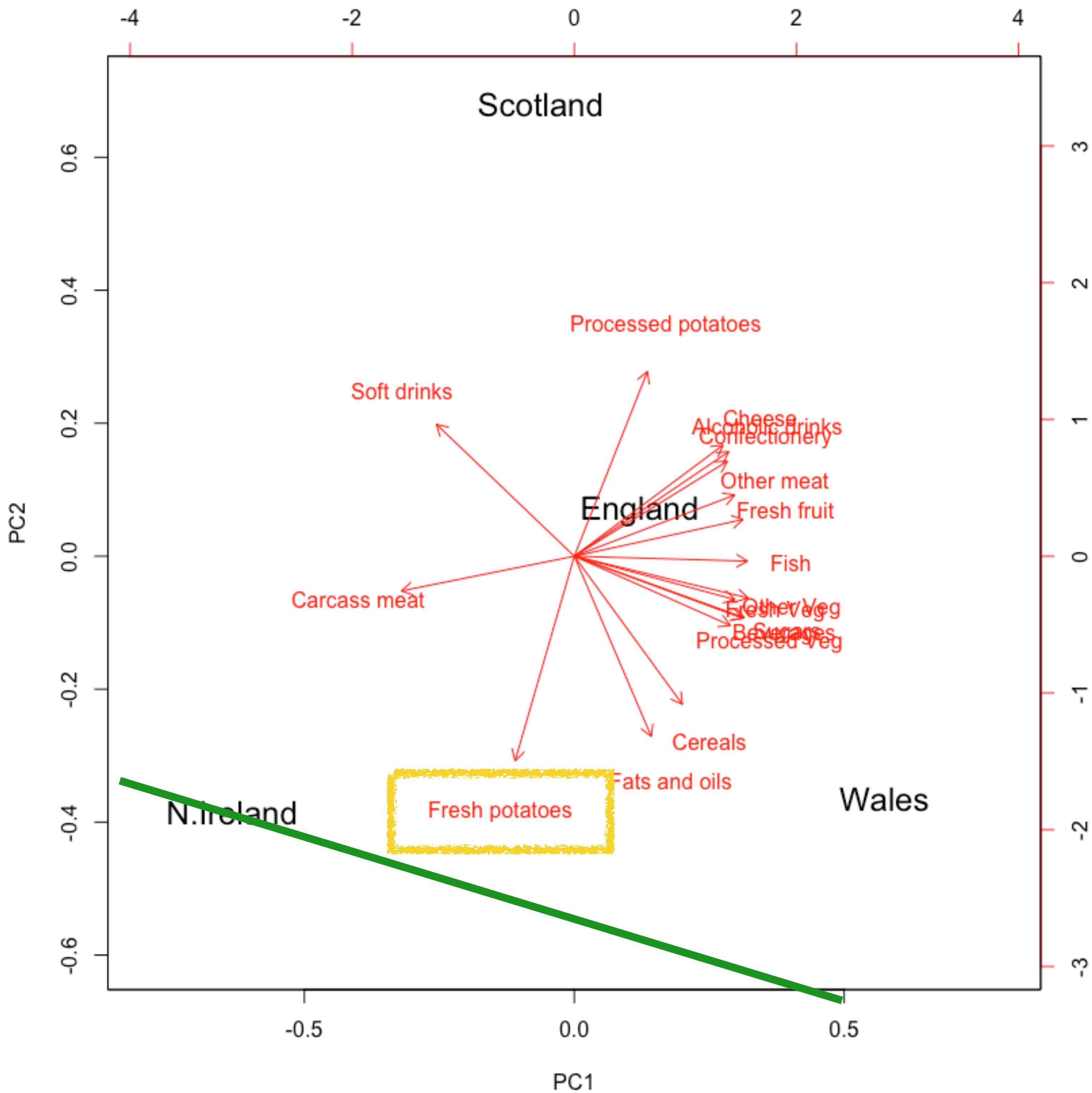
Scotland
England
N.Ireland
Wales
Processed potatoes
Soft drinks
Cheese
Alcoholic drinks
Confectionery
Other meat
Fresh fruit
Fish
Carcass meat
Other Veg
Fresh Veg
Beverages
Sugars
Processed Veg
Cereals
Fats and oils
Fresh potatoes
PC1
PC2
-0.5
0.0
0.5
-4
-2
0
2
4
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
3
2
1
0
-1
-2
-3



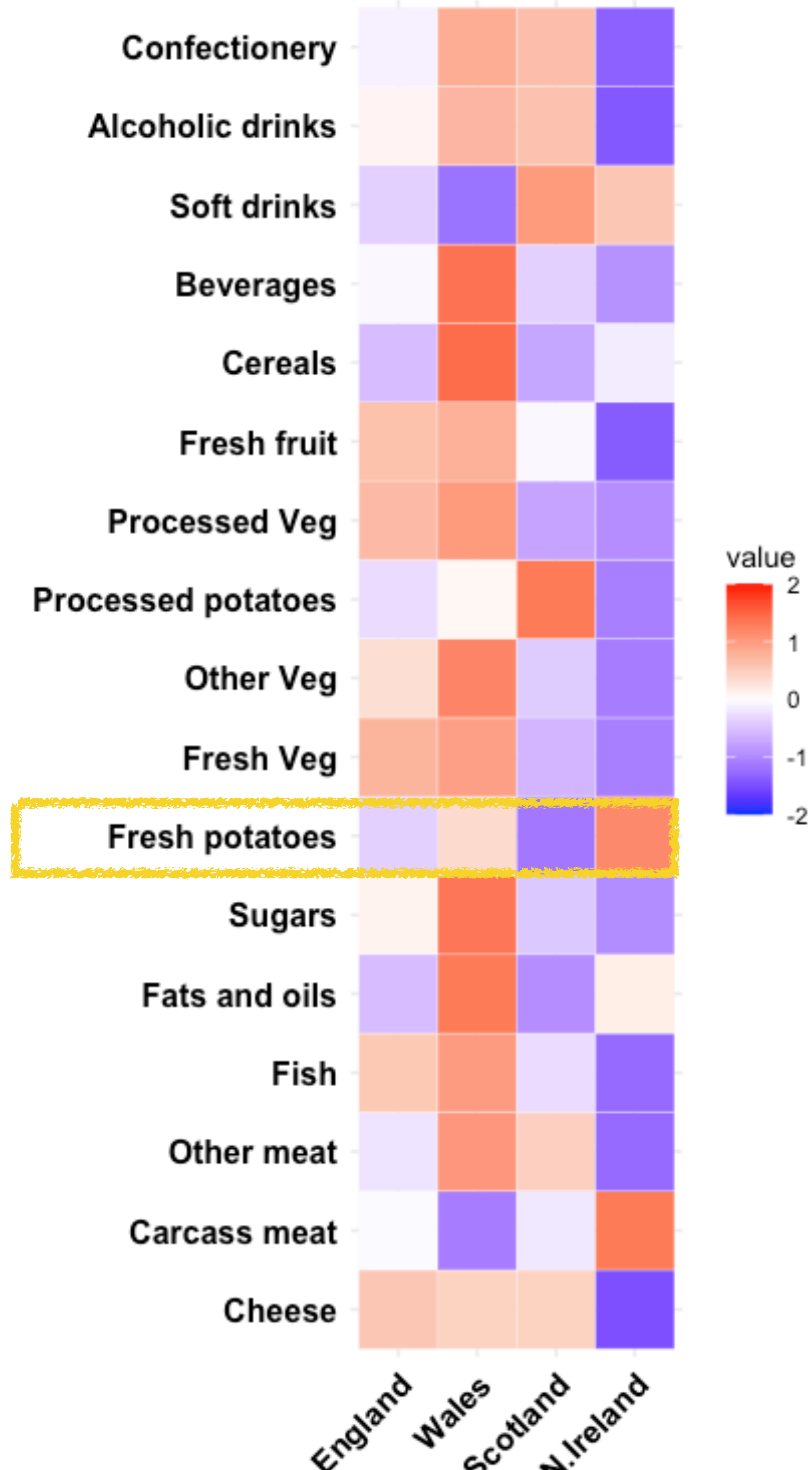
Confectionery
Alcoholic drinks
Soft drinks
Beverages
Cereals
Fresh fruit
Processed Veg
Processed potatoes
Other Veg
Fresh Veg
Fresh potatoes
Sugars
Fats and oils
Fish
Other meat
Carcass meat
Cheese
England
Wales
Scotland
N.Ireland
value
2
1
0
-1
-2



Scotland
England
N.Ireland
Wales
Processed potatoes
Soft drinks
Cheese
Alcoholic drinks
Confectionery
Other meat
Fresh fruit
Fish
Carcass meat
Other Veg
Fresh Veg
Beverages
Sugars
Processed Veg
Cereals
Fats and oils
Fresh potatoes
PC1
PC2
-4
-2
0
2
4
-0.5
0.0
0.5
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
3
2
1
0
-1
-2
-3
Confectionery
Alcoholic drinks
Soft drinks
Beverages
Cereals
Fresh fruit
Processed Veg
Processed potatoes
Other Veg
Fresh Veg
Fresh potatoes
Sugars
Fats and oils
Fish
Other meat
Carcass meat
Cheese
England
Wales
Scotland
N.Ireland
value
2
1
0
-1
-2



Scotland
England
N.Ireland
Wales
Processed potatoes
Soft drinks
Cheese
Alcoholic drinks
Confectionery
Other meat
Fresh fruit
Fish
Carcass meat
Other Veg
Fresh Veg
Beverages
Sugars
Processed Veg
Cereals
Fats and oils
Fresh potatoes
PC1
PC2
-0.5
0.0
0.5
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
-4
-2
0
2
4
-3
-2
-1
0
1
2
3



Confectionery
Alcoholic drinks
Soft drinks
Beverages
Cereals
Fresh fruit
Processed Veg
Processed potatoes
Other Veg
Fresh Veg
Fresh potatoes
Sugars
Fats and oils
Fish
Other meat
Carcass meat
Cheese
England
Wales
Scotland
N.Ireland
value
2
1
0
-1
-2

FIFA Soccer Players

The Data

```
load('fifa.RData')
colnames(fifa)
```

```
 [1] "Name"             "Age"              "Photo"              "Nationality"
 [5] "Flag"             "Overall"          "Potential"          "Club"
 [9] "Club.Logo"        "Value"            "Wage"               "Special"
[13] "Acceleration"     "Aggression"       "Agility"            "Balance"
[17] "Ball.control"     "Composure"        "Crossing"           "Curve"
[21] "Dribbling"        "Finishing"        "Free.kick.accuracy" "GK.diving"
[25] "GK.handling"      "GK.kicking"       "GK.positioning"     "GK.reflexes"
[29] "Heading.accuracy" "Interceptions"    "Jumping"            "Long.passing"
[33] "Long.shots"       "Marking"          "Penalties"          "Positioning"
[37] "Reactions"        "Short.passing"    "Shot.power"         "Sliding.tackle"
[41] "Sprint.speed"     "Stamina"          "Standing.tackle"    "Strength"
[45] "Vision"           "Volleys"          "position"
```

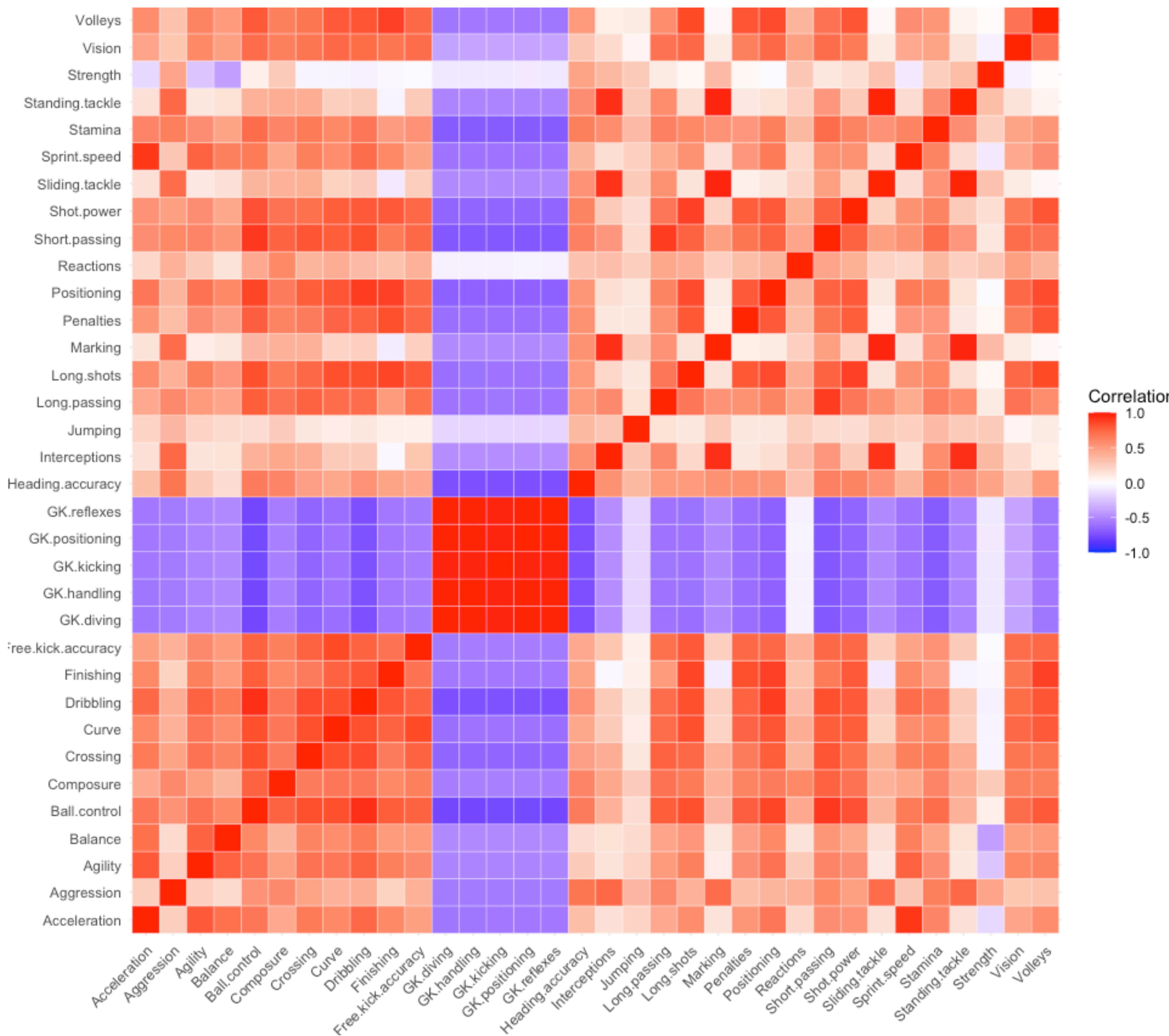
Exploring the Data

```
summary(fifa[,13:46])
```

```
 Acceleration     Aggression       Agility          Balance         Ball.control   Composure        Crossing         Curve
 Min.   :11.00   Min.   :11.00   Min.   :14.00   Min.   :11.00   Min.   : 8     Min.   : 5.00   Min.   : 5.0   Min.   : 6.0
 1st Qu.:56.00   1st Qu.:43.00   1st Qu.:55.00   1st Qu.:56.00   1st Qu.:53     1st Qu.:51.00   1st Qu.:37.0   1st Qu.:34.0
 Median :67.00   Median :58.00   Median :65.00   Median :66.00   Median :62     Median :60.00   Median :54.0   Median :48.0
 Mean   :64.48   Mean   :55.74   Mean   :63.25   Mean   :63.76   Mean   :58     Mean   :57.82   Mean   :49.7   Mean   :47.2
 3rd Qu.:75.00   3rd Qu.:69.00   3rd Qu.:74.00   3rd Qu.:74.00   3rd Qu.:69     3rd Qu.:67.00   3rd Qu.:64.0   3rd Qu.:62.0
 Max.   :96.00   Max.   :96.00   Max.   :96.00   Max.   :96.00   Max.   :95     Max.   :96.00   Max.   :91.0   Max.   :92.0
  Dribbling       Finishing       Free.kick.accuracy   GK.diving        GK.handling      GK.kicking       GK.positioning    GK.reflexes
 Min.   : 2.00   Min.   : 2.00   Min.   : 4.00        Min.   : 1.00   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00
 1st Qu.:48.00   1st Qu.:29.00   1st Qu.:31.00        1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00
 Median :60.00   Median :48.00   Median :42.00        Median :11.00   Median :11.00   Median :11.00   Median :11.00   Median :11.00
 Mean   :54.94   Mean   :45.18   Mean   :43.08        Mean   :16.78   Mean   :16.55   Mean   :16.42   Mean   :16.54   Mean   :16.91
 3rd Qu.:68.00   3rd Qu.:61.00   3rd Qu.:57.00        3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00
 Max.   :97.00   Max.   :95.00   Max.   :93.00        Max.   :91.00   Max.   :91.00   Max.   :95.00   Max.   :91.00   Max.   :90.00
 Heading.accuracy Interceptions     Jumping         Long.passing      Long.shots        Marking         Penalties        Positioning
 Min.   : 4.00     Min.   : 4.00   Min.   :15.00   Min.   : 7.00   Min.   : 3.00   Min.   : 4.00   Min.   : 5.00   Min.   : 2.00
 1st Qu.:44.00     1st Qu.:26.00   1st Qu.:58.00   1st Qu.:42.00   1st Qu.:32.00   1st Qu.:22.00   1st Qu.:39.00   1st Qu.:38.00
 Median :55.00     Median :52.00   Median :66.00   Median :56.00   Median :51.00   Median :48.00   Median :50.00   Median :54.00
 Mean   :52.26     Mean   :46.53   Mean   :64.84   Mean   :52.37   Mean   :47.11   Mean   :44.09   Mean   :48.92   Mean   :49.53
 3rd Qu.:64.00     3rd Qu.:64.00   3rd Qu.:73.00   3rd Qu.:64.00   3rd Qu.:62.00   3rd Qu.:63.00   3rd Qu.:61.00   3rd Qu.:64.00
 Max.   :94.00     Max.   :92.00   Max.   :95.00   Max.   :93.00   Max.   :92.00   Max.   :92.00   Max.   :92.00   Max.   :95.00
```

Exploring the Data

```
cor.matrix = cor(fifa[,13:46])
```

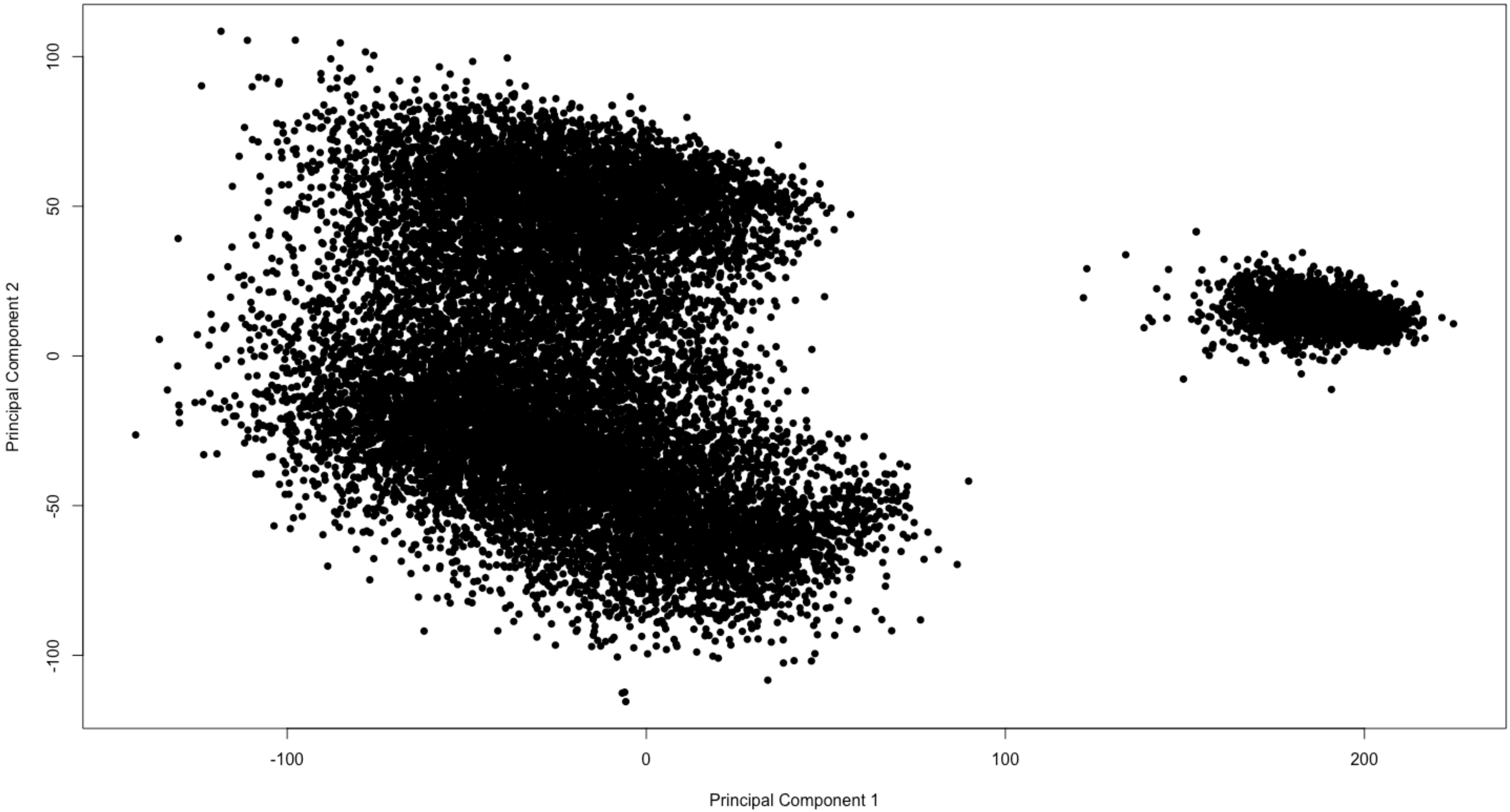


Compute the PCA

```
fifa.pca = prcomp(fifa[,13:46])
summary(fifa.pca)
```

```
Importance of components:
                           PC1     PC2      PC3      PC4      PC5      PC6      PC7
Standard deviation     74.8371 43.5787 23.28767 20.58146 16.12477 10.71539 10.17785
Proportion of Variance  0.5647  0.1915  0.05468  0.04271  0.02621  0.01158  0.01044
Cumulative Proportion   0.5647  0.7561  0.81081  0.85352  0.87973  0.89131  0.90175
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     9.11852 8.98065 8.5082 8.41550 7.93741 7.15935 7.06502
Proportion of Variance 0.00838 0.00813 0.0073 0.00714 0.00635 0.00517 0.00503
Cumulative Proportion  0.91013 0.91827 0.9256 0.93270 0.93906 0.94422 0.94926
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     6.68497 6.56406 6.50459 6.22369 6.08812 6.00578 5.91320
Proportion of Variance 0.00451 0.00434 0.00427 0.00391 0.00374 0.00364 0.00353
Cumulative Proportion  0.95376 0.95811 0.96237 0.96628 0.97001 0.97365 0.97718
                          PC22    PC23    PC24    PC25    PC26   PC27    PC28
Standard deviation     5.66946 5.45018 5.15051 4.86761 4.34786 4.1098 4.05716
Proportion of Variance 0.00324 0.00299 0.00267 0.00239 0.00191 0.0017 0.00166
Cumulative Proportion  0.98042 0.98341 0.98609 0.98848 0.99038 0.9921 0.99374
```

Plot the Projection



Examine Loadings

`fifa.pca$rotation[,1:3]`

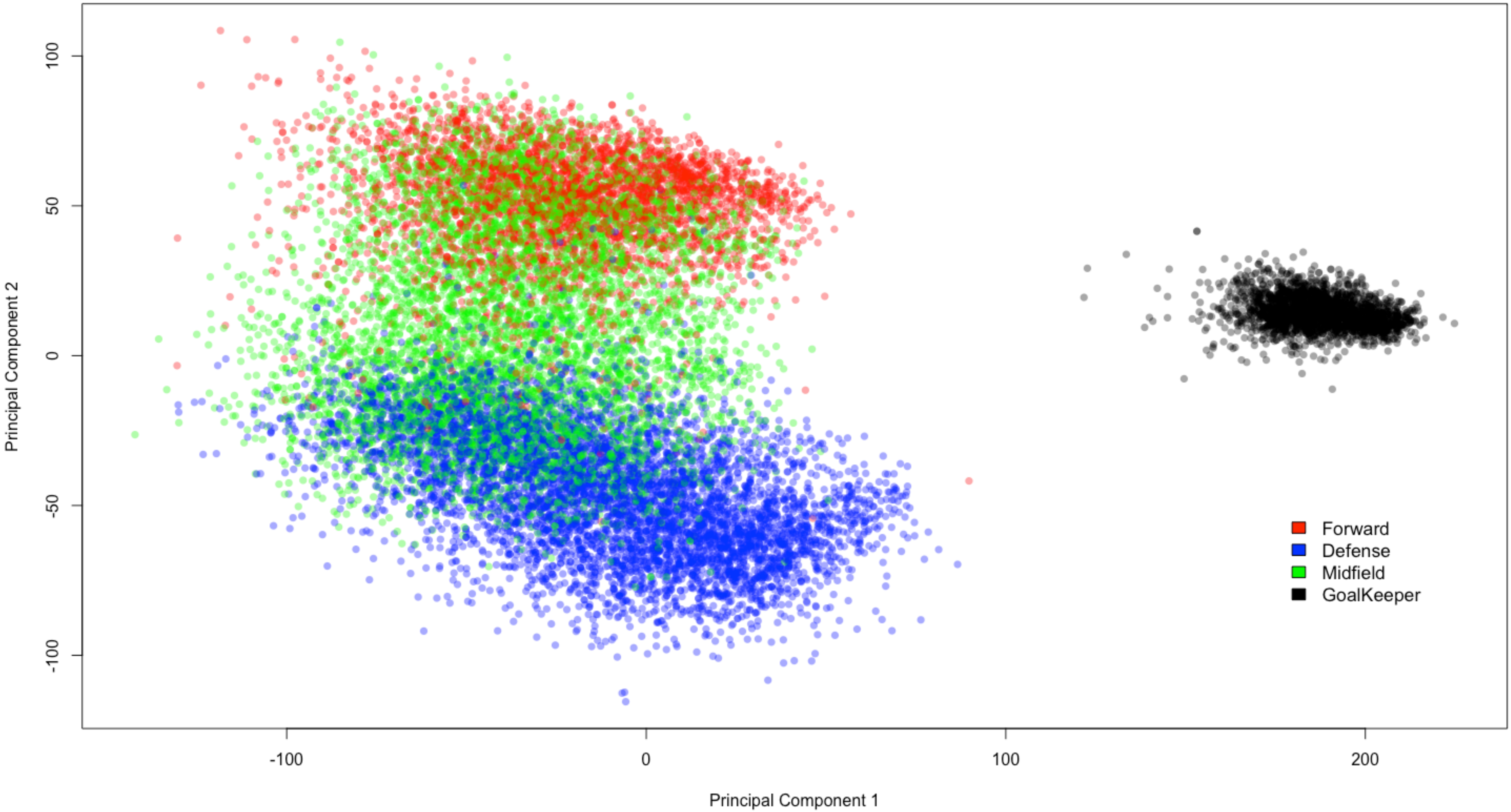
	PC1	PC2	PC3
Acceleration	-0.13674335	0.0944478107	-0.141193842
Aggression	-0.15322857	-0.2030537953	0.105372978
Agility	-0.13598896	0.1196301737	-0.017763073
Balance	-0.11474980	0.0865672989	-0.072629834
Ball.control	-0.21256812	0.0585990154	0.038243802
Composure	-0.13288575	-0.0005635262	0.163887637
Crossing	-0.21347202	0.0458210228	0.124741235
Curve	-0.20656129	0.1254947094	0.180634730
Dribbling	-0.23090613	0.1259819707	-0.002905379
Finishing	-0.19431248	0.2534086437	0.006524693
Free.kick.accuracy	-0.18528508	0.0960404650	0.219976709
GK.diving	0.20757999	0.0480952942	0.326161934
GK.handling	0.19811125	0.0464542553	0.314165622
GK.kicking	0.19261876	0.0456942190	0.304722126
GK.positioning	0.19889113	0.0456384196	0.317850121
GK.reflexes	0.21081755	0.0489895700	0.332751195
Heading.accuracy	-0.17218607	-0.1115416097	-0.125135161
Interceptions	-0.15038835	-0.3669025376	0.162064432
Jumping	-0.03805419	-0.0579221746	0.012263523
Long.passing	-0.16849827	-0.0435009943	0.224584171
Long.shots	-0.21415526	0.1677851237	0.157466462
Marking	-0.14863254	-0.4076616902	0.078298039
Penalties	-0.16328049	0.1407803994	0.024403976
Positioning	-0.22053959	0.1797895382	0.020734699
Reactions	-0.04780774	0.0001844959	0.250247098
Short.passing	-0.18176636	-0.0033124240	0.118611543
Shot.power	-0.19592137	0.0989340925	0.101707386
Sliding.tackle	-0.14977558	-0.4024030355	0.069945935
Sprint.speed	-0.13387287	0.0804847541	-0.146049405
Stamina	-0.17231648	-0.0634639786	-0.016509650
Standing.tackle	-0.15992073	-0.4039763876	0.086418583
Strength	-0.02186264	-0.1151018222	0.096053864
Vision	-0.13027169	0.1152237536	0.260985686
Volleys	-0.18465028	0.1888480712	0.076974579

Examine Loadings

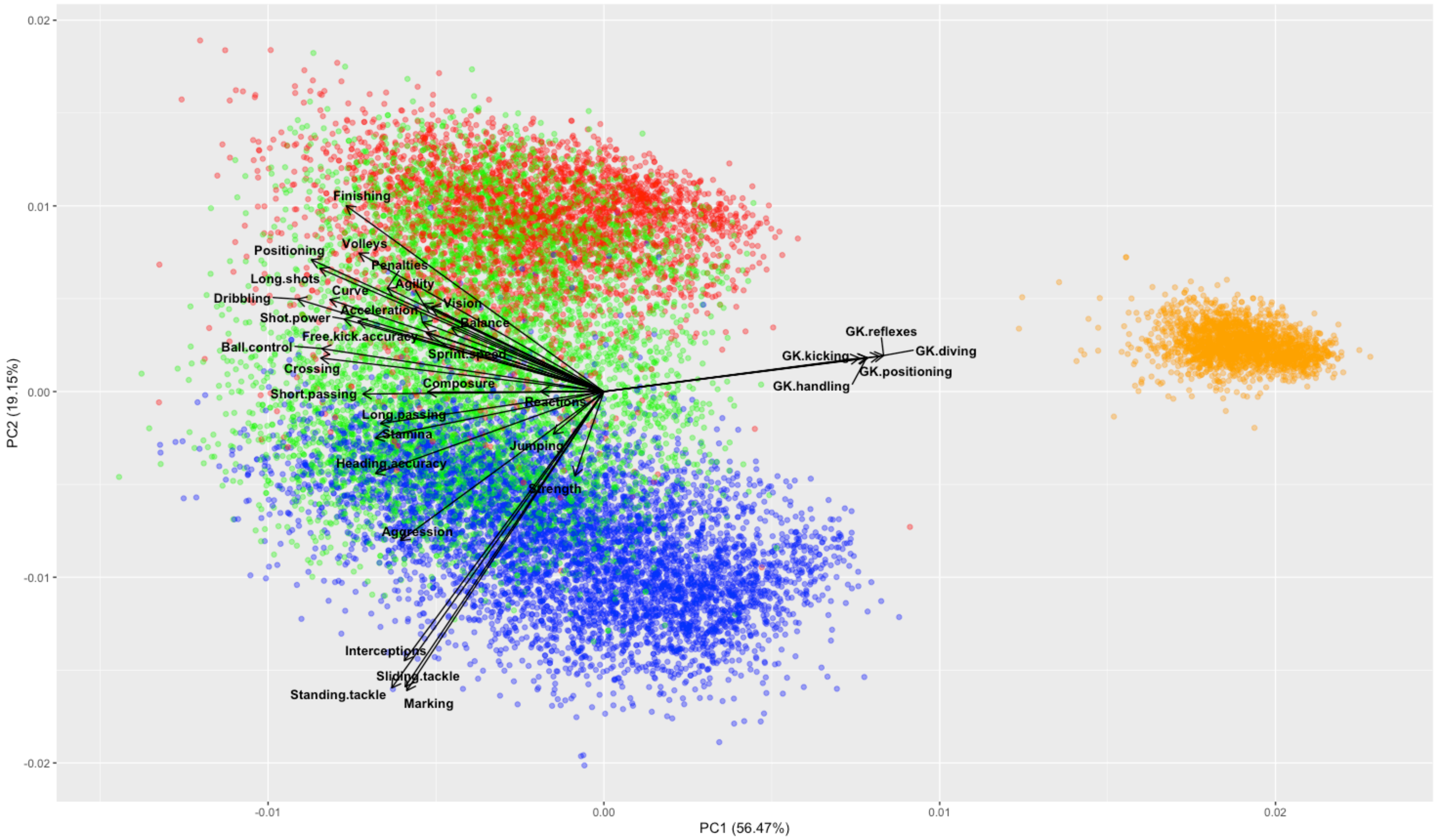
`fifa.pca$rotation[,1:3]`

	PC1	PC2	PC3
Acceleration	-0.13674335	0.0944478107	-0.141193842
Aggression	-0.15322857	-0.2030537953	0.105372978
Agility	-0.13598896	0.1196301737	-0.017763073
Balance	-0.11474980	0.0865672989	-0.072629834
Ball.control	-0.21256812	0.0585990154	0.038243802
Composure	-0.13288575	-0.0005635262	0.163887637
Crossing	-0.21347202	0.0458210228	0.124741235
Curve	-0.20656129	0.1254947094	0.180634730
Dribbling	-0.23090613	0.1259819707	-0.002905379
Finishing	-0.19431248	0.2534086437	0.006524693
Free.kick.accuracy	-0.18528508	0.0960404650	0.219976709
GK.diving	0.20757999	0.0480952942	0.326161934
GK.handling	0.19811125	0.0464542553	0.314165622
GK.kicking	0.19261876	0.0456942190	0.304722126
GK.positioning	0.19889113	0.0456384196	0.317850121
GK.reflexes	0.21081755	0.0489895700	0.332751195
Heading.accuracy	-0.17218607	-0.1115416097	-0.125135161
Interceptions	-0.15038835	-0.3669025376	0.162064432
Jumping	-0.03805419	-0.0579221746	0.012263523
Long.passing	-0.16849827	-0.0435009943	0.224584171
Long.shots	-0.21415526	0.1677851237	0.157466462
Marking	-0.14863254	-0.4076616902	0.078298039
Penalties	-0.16328049	0.1407803994	0.024403976
Positioning	-0.22053959	0.1797895382	0.020734699
Reactions	-0.04780774	0.0001844959	0.250247098
Short.passing	-0.18176636	-0.0033124240	0.118611543
Shot.power	-0.19592137	0.0989340925	0.101707386
Sliding.tackle	-0.14977558	-0.4024030355	0.069945935
Sprint.speed	-0.13387287	0.0804847541	-0.146049405
Stamina	-0.17231648	-0.0634639786	-0.016509650
Standing.tackle	-0.15992073	-0.4039763876	0.086418583
Strength	-0.02186264	-0.1151018222	0.096053864
Vision	-0.13027169	0.1152237536	0.260985686
Volleys	-0.18465028	0.1888480712	0.076974579

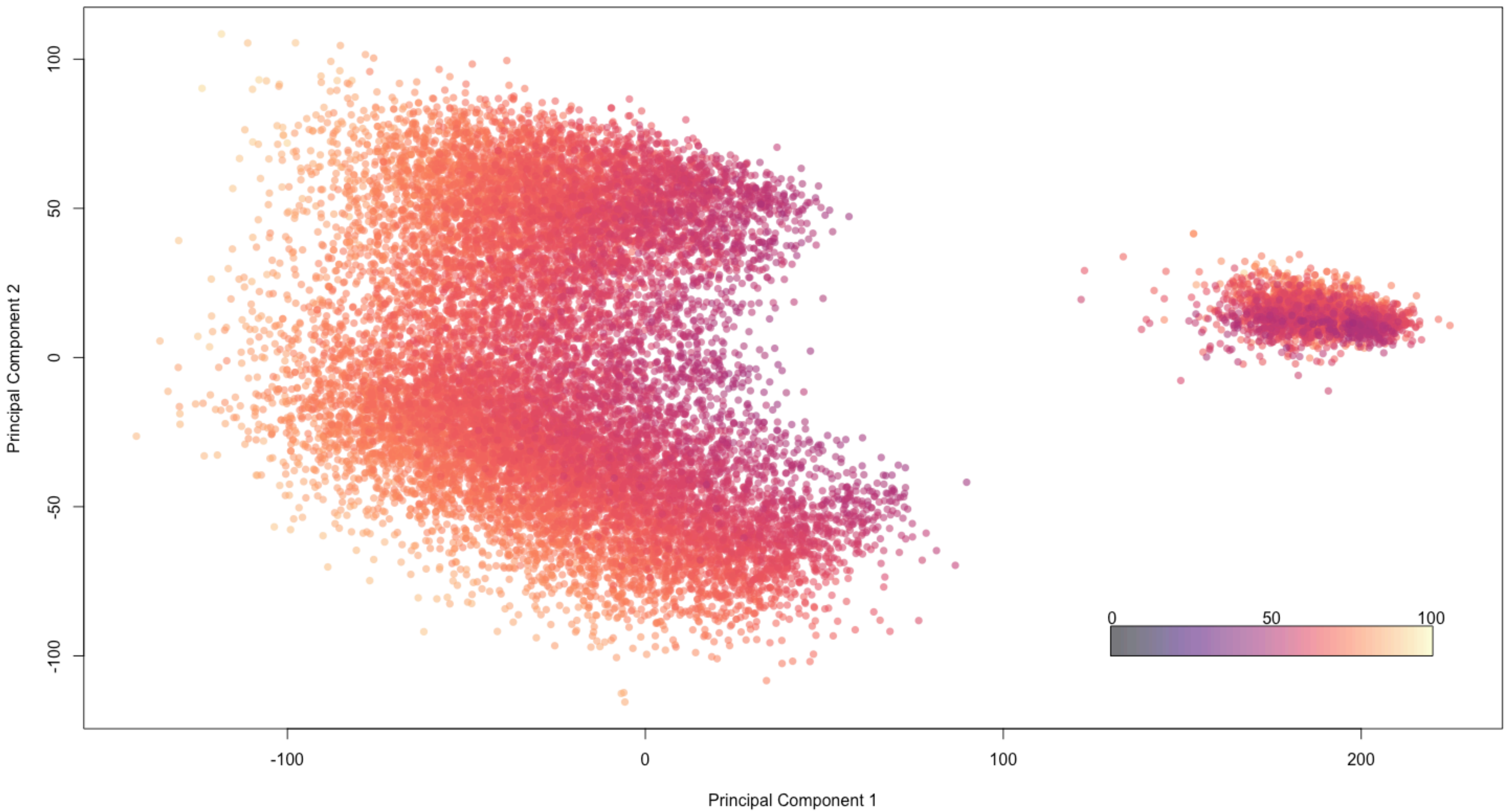
Color Projection by Position



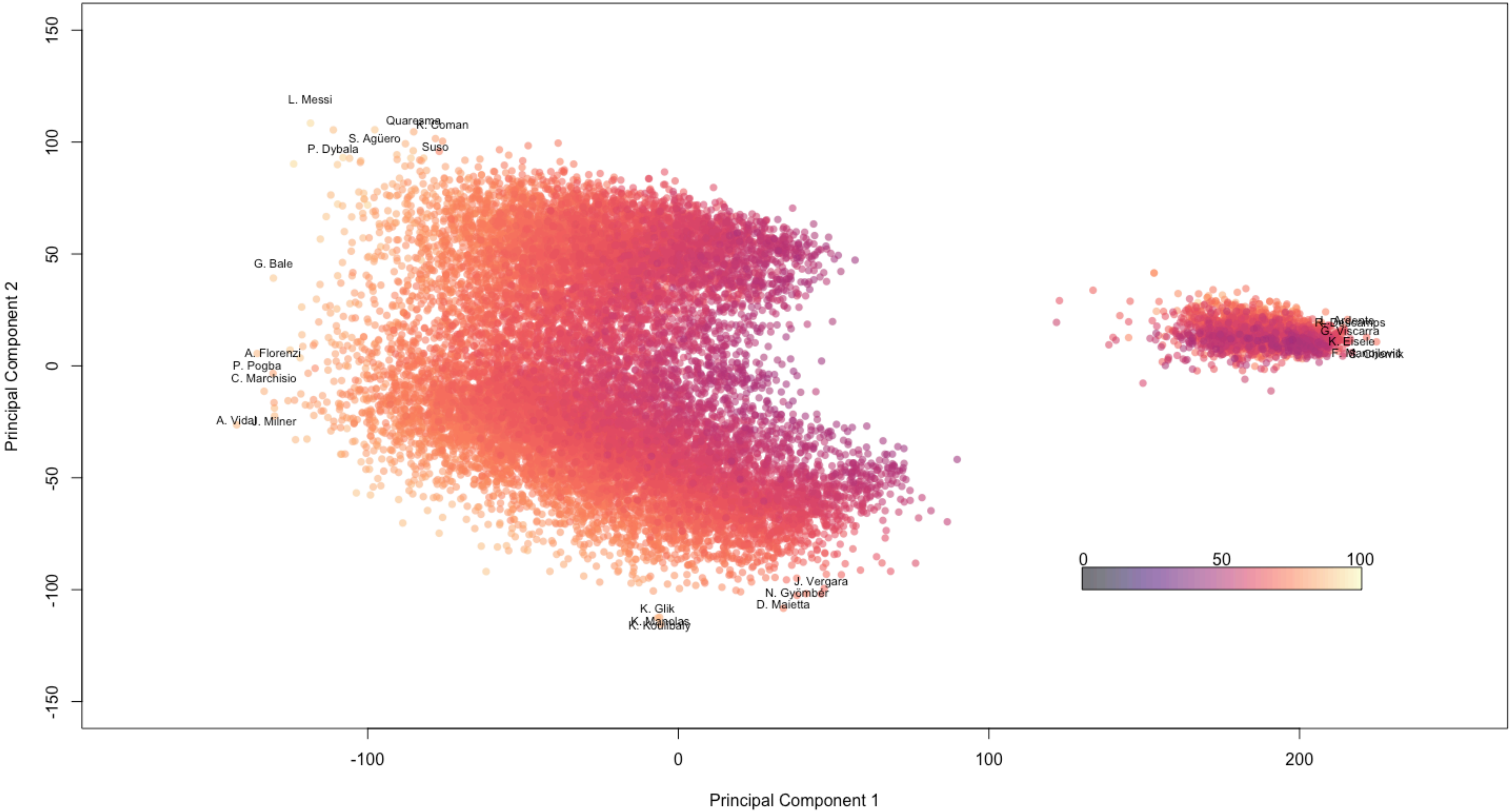
Make it a Biplot



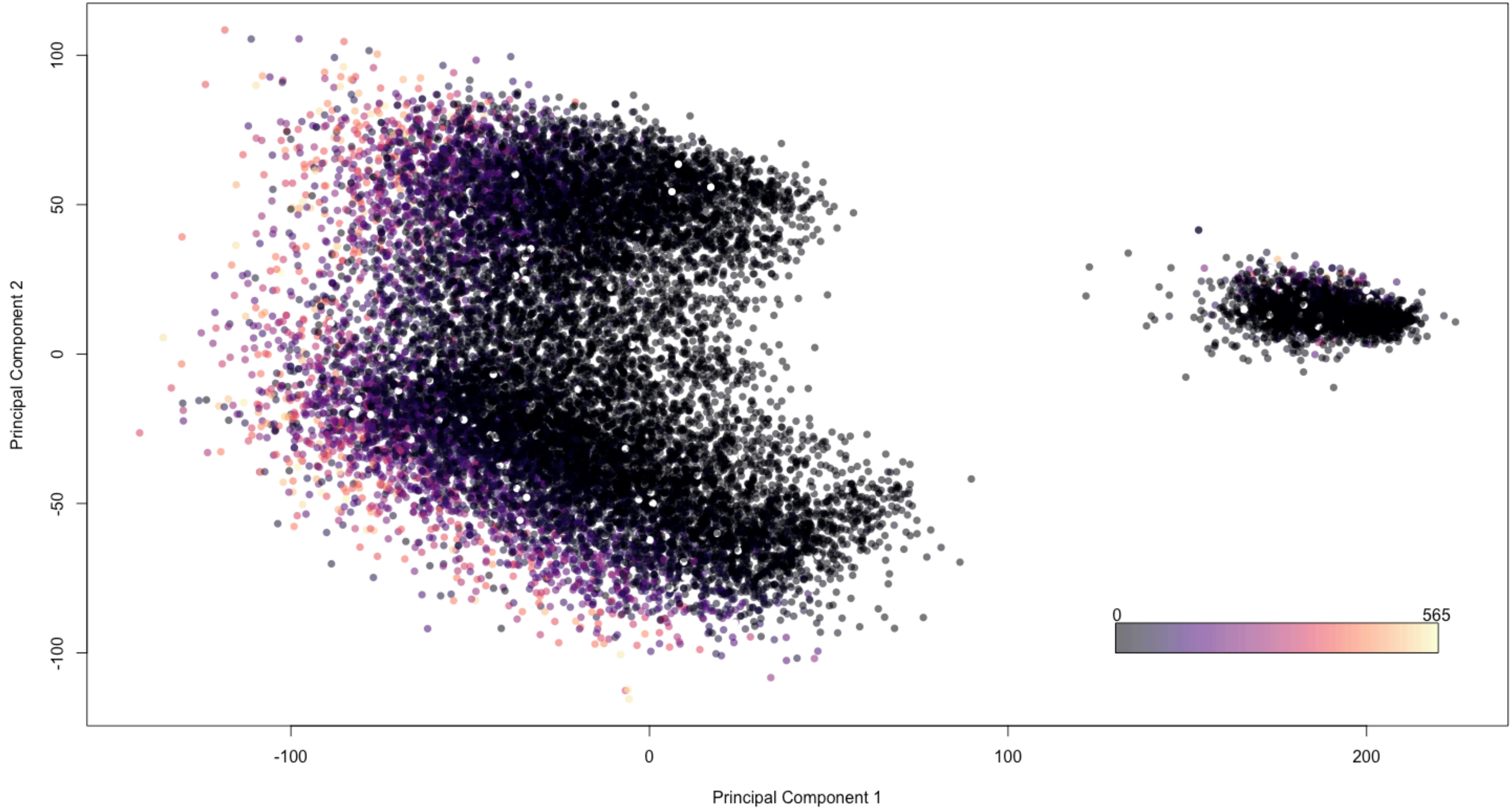
Color by "Overall" ability



Label some Outliers



Color by Wage



Lab Exercise 1

1. Use a logistic regression to try to differentiate midfielders from forwards, with 2 principal components as input.
2. Create one model using correlation PCA and one model using covariance PCA.
3. Which model can more accurately differentiate between midfielders and forwards?

Lab Exercise 2

1. Use a linear regression model to predict the variable *overall* using just 2 principal components. You may use functions of these components (polynomial terms etc.) if you'd like.
2. Compare the resulting model using correlation and covariance PCA. Which was more successful?