lab2

shichenh

9/17/2017

PCA on Portugal Whole Food Sales Data

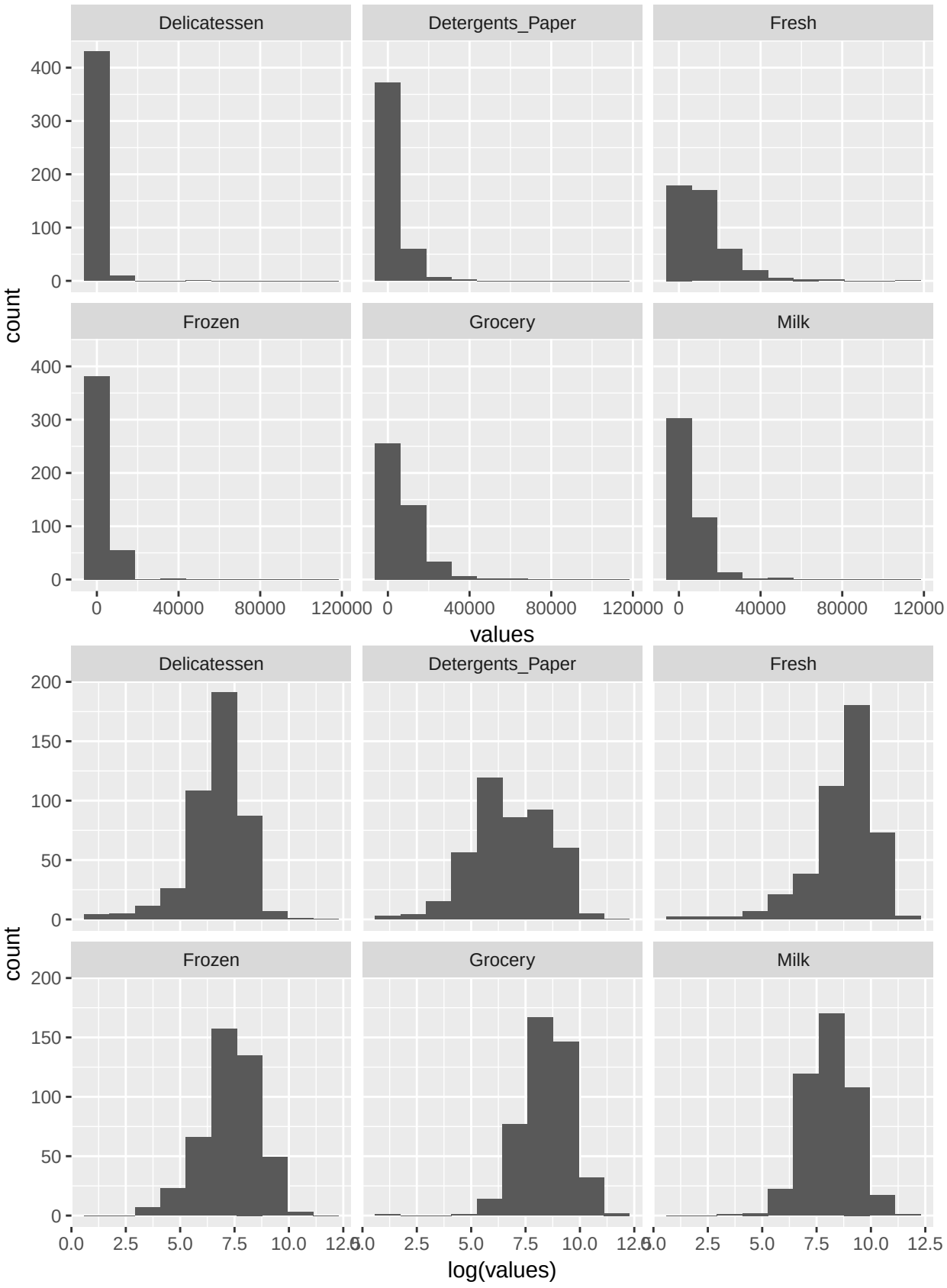
Loading and Cleaning Data

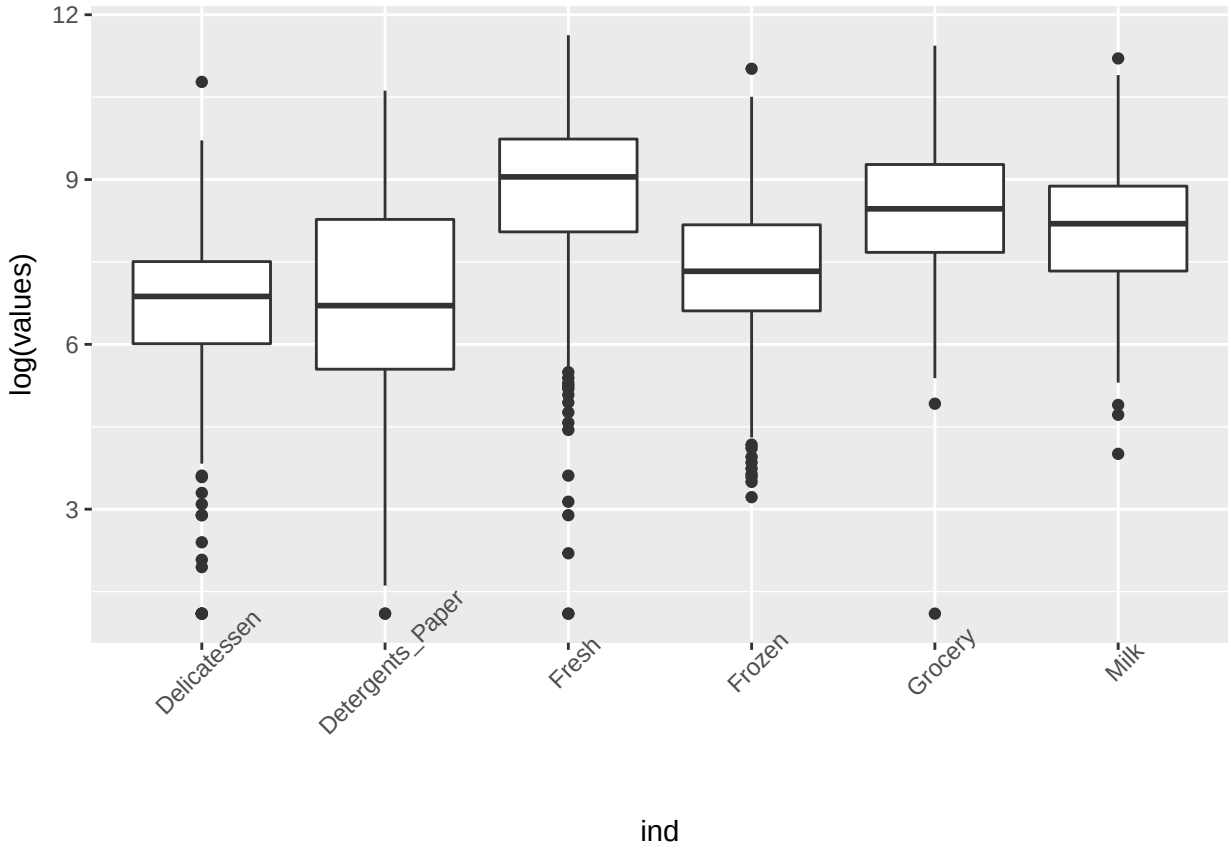
```
##
## Horeca Retail
##    298    142
```

```
##
## Lisbon Oporto  Other
##     77     47    316
```

"EDA"

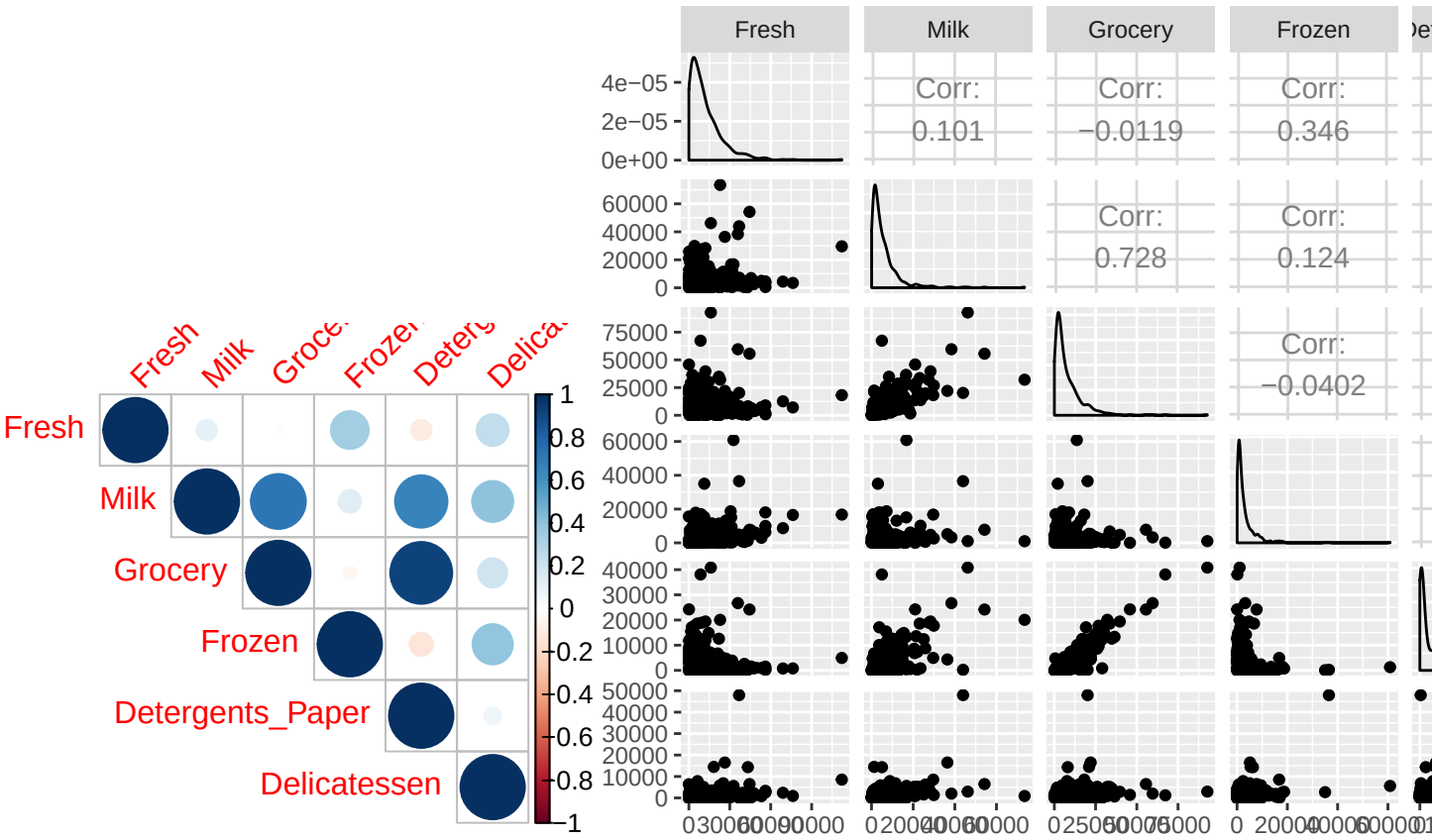
```
##    Channel       Region        Fresh             Milk
##  Horeca:298   Lisbon: 77   Min.   :     3   Min.   :   55
##  Retail:142   Oporto: 47   1st Qu.:  3128   1st Qu.: 1533
##               Other :316   Median :  8504   Median : 3627
##                            Mean   : 12000   Mean   : 5796
##                            3rd Qu.: 16934   3rd Qu.: 7190
##                            Max.   :112151   Max.   :73498
##     Grocery          Frozen        Detergents_Paper   Delicatessen
##  Min.   :    3   Min.   :   25.0   Min.   :    3.0   Min.   :    3.0
##  1st Qu.: 2153   1st Qu.:  742.2   1st Qu.:  256.8   1st Qu.:  408.2
##  Median : 4756   Median : 1526.0   Median :  816.5   Median :  965.5
##  Mean   : 7951   Mean   : 3071.9   Mean   : 2881.5   Mean   : 1524.9
##  3rd Qu.:10656   3rd Qu.: 3554.2   3rd Qu.: 3922.0   3rd Qu.: 1820.2
##  Max.   :92780   Max.   :60869.0   Max.   :40827.0   Max.   :47943.0
```


Delicatessen
Detergents_Paper
Fresh
Frozen
Grocery
Milk
count
values
log(values)
0
100
200
300
400
40000
80000
120000
50
150
2.5
5.0
7.5
10.0
12.5



Many variables are severely right skewed. Altough after applying log transform, many variables are still slightly left skewed(due to too right skewed).
All variables' first and third quantile range are quite similar except the detergents_paper category. Possible explanation is that different regions more variance on using deteregents and paper.

```
##                       Fresh      Milk     Grocery      Frozen
## Fresh              1.00000000 0.1005098 -0.01185387  0.34588146
## Milk               0.10050977 1.0000000  0.72833512  0.12399376
## Grocery           -0.01185387 0.7283351  1.00000000 -0.04019274
## Frozen             0.34588146 0.1239938 -0.04019274  1.00000000
## Detergents_Paper -0.10195294 0.6618157  0.92464069 -0.13152491
## Delicatessen       0.24468997 0.4063683  0.20549651  0.39094747
##                  Detergents_Paper Delicatessen
## Fresh                  -0.1019529    0.2446900
## Milk                    0.6618157    0.4063683
## Grocery                 0.9246407    0.2054965
## Frozen                 -0.1315249    0.3909475
## Detergents_Paper        1.0000000    0.0692913
## Delicatessen            0.0692913    1.0000000
```



Grocery is highly related to Detergents_Paper. I wonder if Grocery includes Detergents_Paper.

PCA

Challenge

```
##                   Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5 Comp. 6
## Fresh             -0.0429 -0.5279  0.8123  0.2367  0.0487 -0.0360
## Milk              -0.5451 -0.0832 -0.0604  0.0872 -0.8266 -0.0380
## Grocery           -0.5793  0.1461  0.1084 -0.1060  0.3150  0.7217
## Frozen            -0.0512 -0.6113 -0.1784 -0.7687  0.0279 -0.0156
## Detergents_Paper -0.5486  0.2552  0.1362 -0.1717  0.3396 -0.6859
## Delicatessen      -0.2487 -0.5042 -0.5239  0.5521  0.3147 -0.0751
```

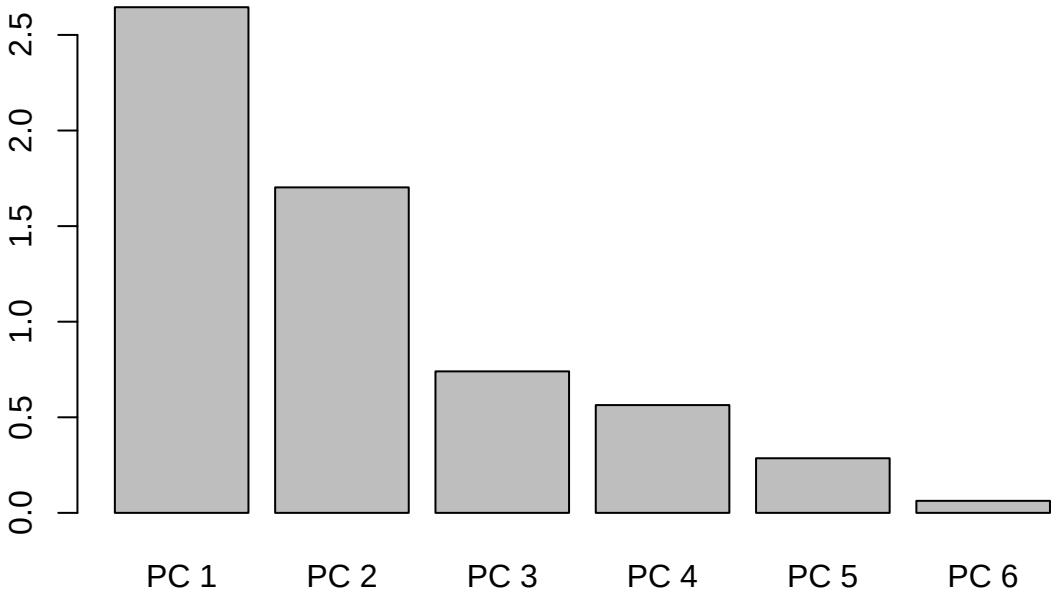
Difference Btw the Two

"The calculation is done by a singular value decomposition of the (centered and possibly scaled) data matrix, not by using eigen on the covariance matrix. This is generally the preferred method for numerical accuracy. The print method for these objects prints the results in a nice format and the plot method produces a scree plot.

Unlike princomp, variances are computed with the usual divisor N - 1."

"princomp uses 'eigen' on the correlation or covariance matrix, as determined by cor. This is done for compatibility with the S-PLUS result."

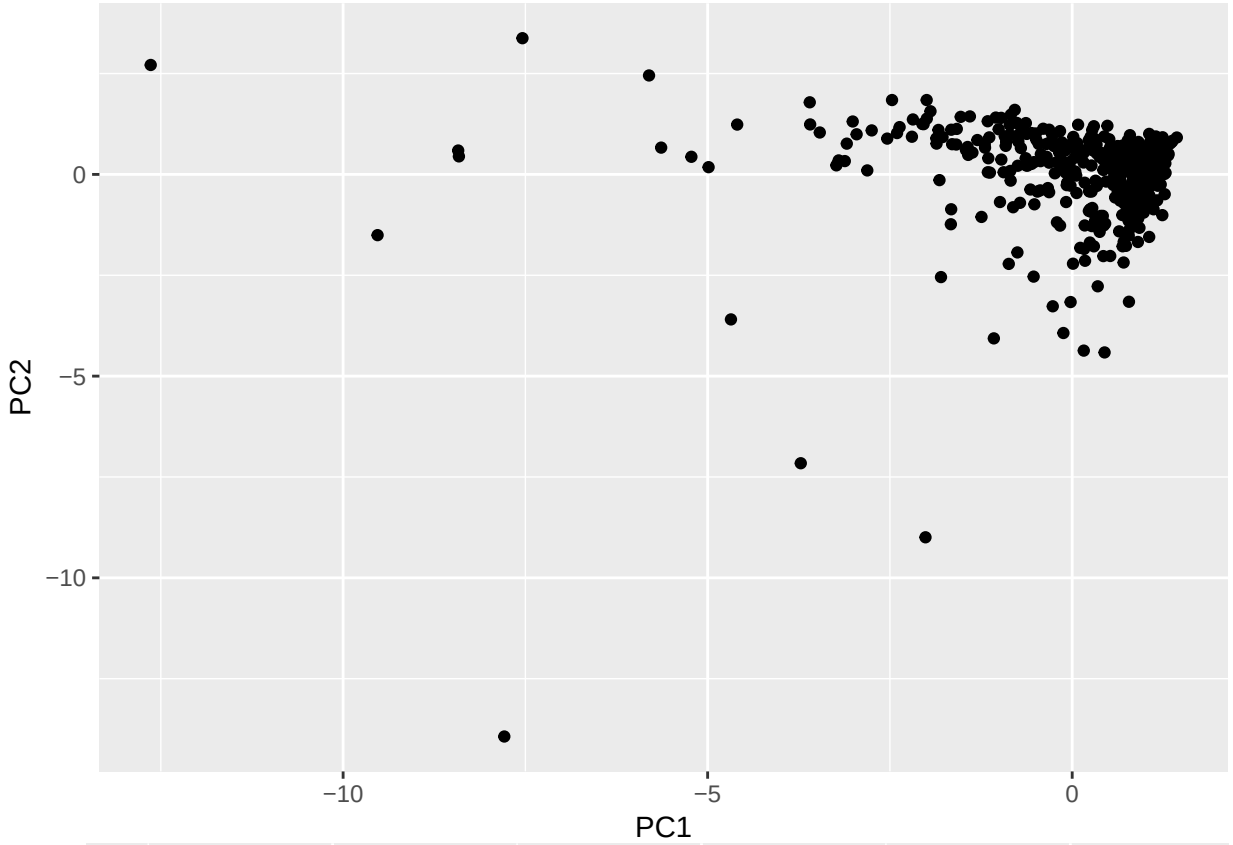
```
##   eigen.value percentage cumulative.percentage
## 1  2.64497357 0.44082893             0.4408289
## 2  1.70258397 0.28376400             0.7245929
## 3  0.74006477 0.12334413             0.8479371
## 4  0.56373023 0.09395504             0.9418921
## 5  0.28567634 0.04761272             0.9895048
## 6  0.06297111 0.01049519             1.0000000
```

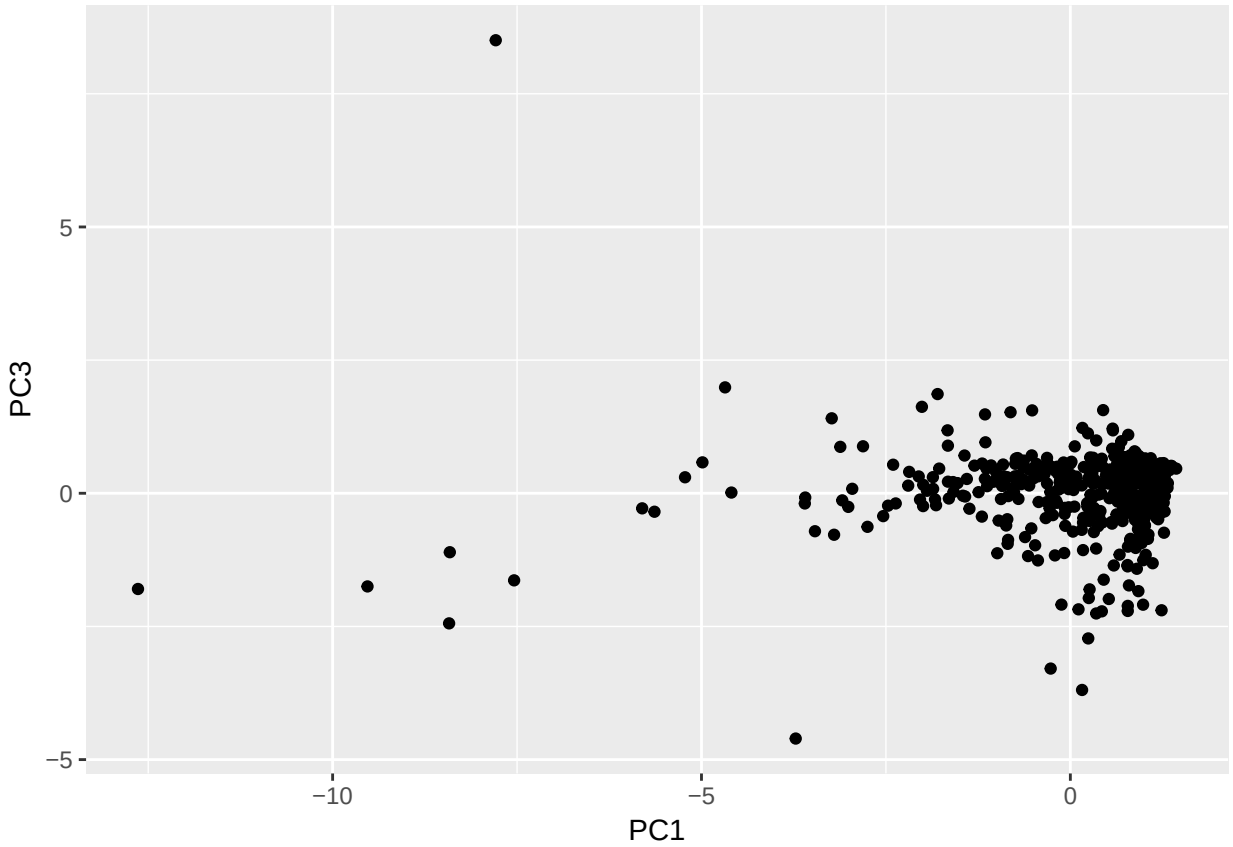


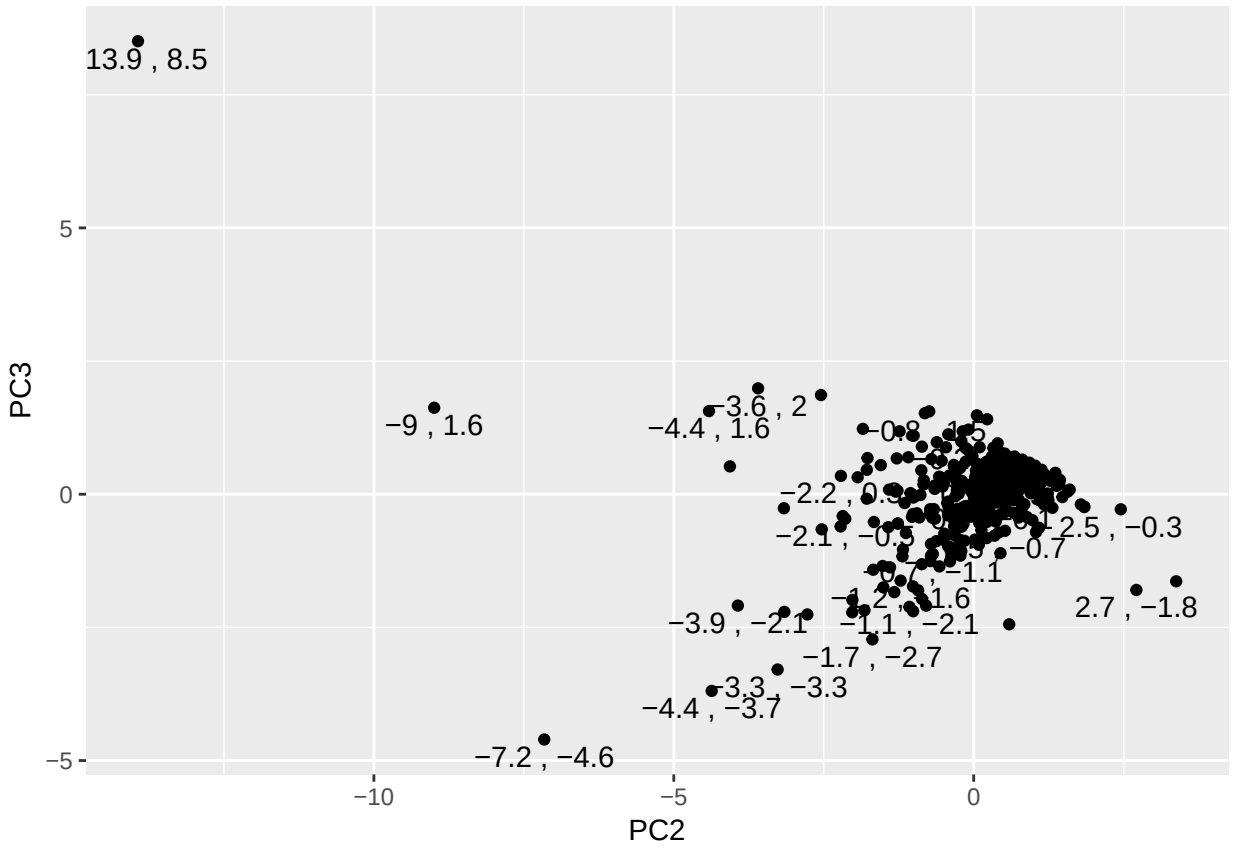
The first PC explains 44% of the variable, second 28.r%, theird 12.3 percenge.

```
##                     Comp. 1    Comp. 2     Comp. 3     Comp. 4
## Fresh            -0.06976988 -0.6888203  0.69879746  0.17771902
## Milk             -0.88651656 -0.1085619 -0.05196032  0.06547148
## Grocery          -0.94213730  0.1906358  0.09325329 -0.07958689
## Frozen           -0.08326848 -0.7976432 -0.15347220 -0.57715511
## Detergents_Paper -0.89220874  0.3329929  0.11716880 -0.12891574
## Delicatessen     -0.40447013 -0.6578958 -0.45069554  0.41452756
##                     Comp. 5      Comp. 6
## Fresh             0.02602952 -0.009033856
## Milk             -0.44180695 -0.009535737
## Grocery           0.16836340  0.181103727
## Frozen            0.01491219 -0.003914671
## Detergents_Paper  0.18151178 -0.172120059
## Delicatessen      0.16820306 -0.018845628
```

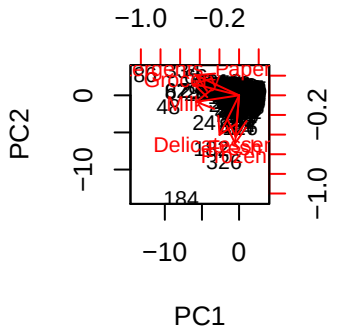
Grocery to PC1, Frozen to PC2, Fresh to PC3. **All categories are negatively correlated with PC1. Some categories are positvely related some are negatively related to PC2.

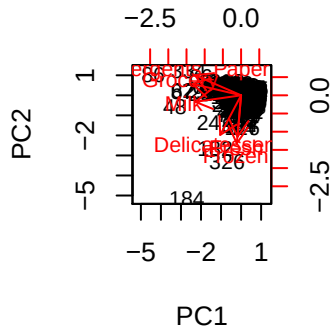

0
−5
−10
PC2
−10
−5
0
PC1

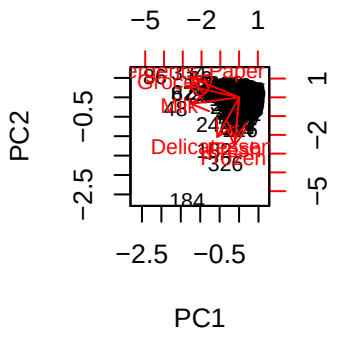

5
0
−5
PC3
−10
−5
0
PC1

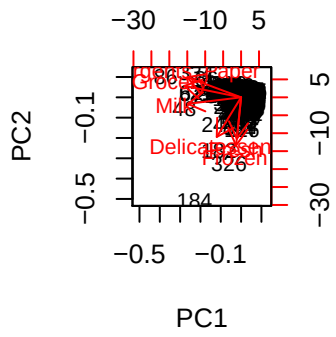


Most of the data are clutered and centered around 0 of the two PC we chose.(good or bad?)









The higher the scale is, the further the magnitude the arrow is. Changing the scale does not change the

biplot so all of them look the same.