TÉCNICO LISBOA

Report Project 1

Data Analysis and Statistical Modeling

Prof Isabel Rodrigues

Grupo 1

João Matos nº98949

Ana Pinto nº102949

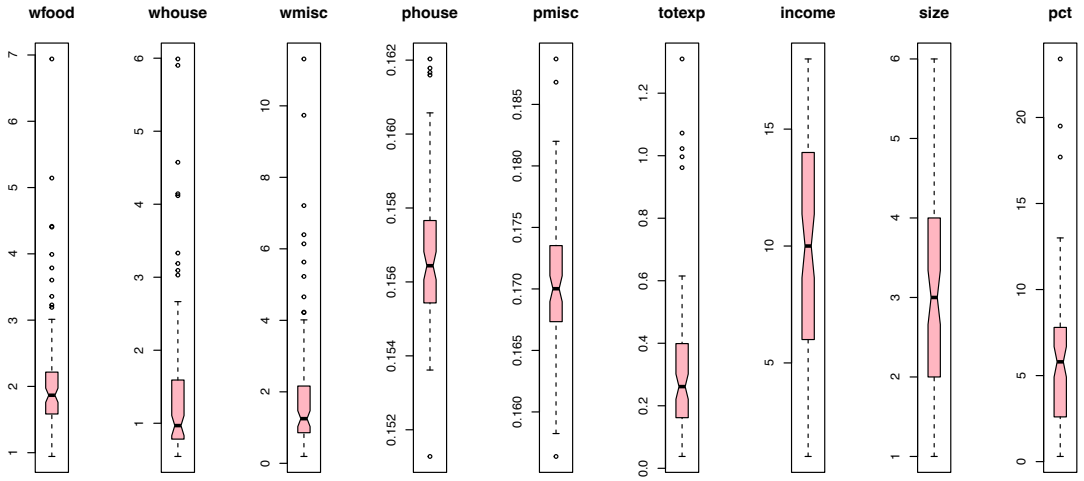
Marina Nóbrega nº103880

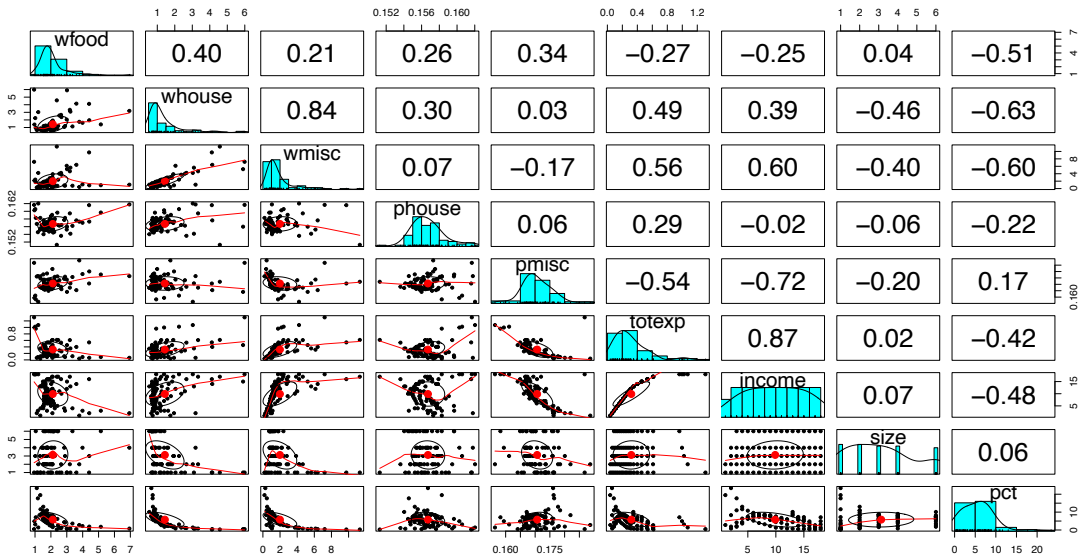
Manuel Dias nº96056

Maria Freitas nº96757

Library: Ecdat
Data frame: BudgetItaly
Subset: year 73
Variables: all except pfood:

- wfood – fodd share
- whouse – housing and fuels share
- wmisc – miscellaneous share
- phouse – housing and fuels price
- pmisc – miscellaneous price
- totexp – total expenditure
- year
- income
- size – household size
- pct – cellule weight





```
> describe(dat_73)
        vars  n mean   sd median trimmed  mad  min   max range  skew kurtosis   se
wfood      1 86 2.12 0.95   1.87    1.97 0.45 0.94  6.94  6.00  2.33     7.35 0.10
whouse     2 86 1.43 1.10   0.97    1.19 0.35 0.55  5.99  5.44  2.36     5.68 0.12
wmisc      3 86 1.99 1.97   1.25    1.59 0.71 0.19 11.31 11.12  2.52     7.20 0.21
phouse     4 86 0.16 0.00   0.16    0.16 0.00 0.15  0.16  0.01  0.69     1.11 0.00
pmisc      5 86 0.17 0.01   0.17    0.17 0.00 0.16  0.19  0.03  0.51     1.66 0.00
totexp     6 86 0.32 0.24   0.26    0.28 0.18 0.04  1.31  1.27  1.76     3.54 0.03
income     7 86 9.88 5.01  10.00    9.90 5.93 1.00 18.00 17.00 -0.02    -1.22 0.54
size       8 86 3.13 1.71   3.00    3.04 1.48 1.00  6.00  5.00  0.44    -0.98 0.18
pct        9 86 5.80 4.12   5.80    5.39 3.71 0.30 23.40 23.10  1.44     3.77 0.44
```

	Insorized_Mean	Variance	Mad
wfood	1.9389	0.9009	0.4526
whouse	1.1584	1.2051	0.3469
wmisc	1.5583	3.8677	0.7085
phouse	0.1565	0.0000	0.0016
pmisc	0.1704	0.0000	0.0046
totexp	0.2800	0.0599	0.1841
income	9.9419	25.1393	5.9304
size	2.7558	2.9364	1.4826
pct	5.3767	17.0026	3.7065

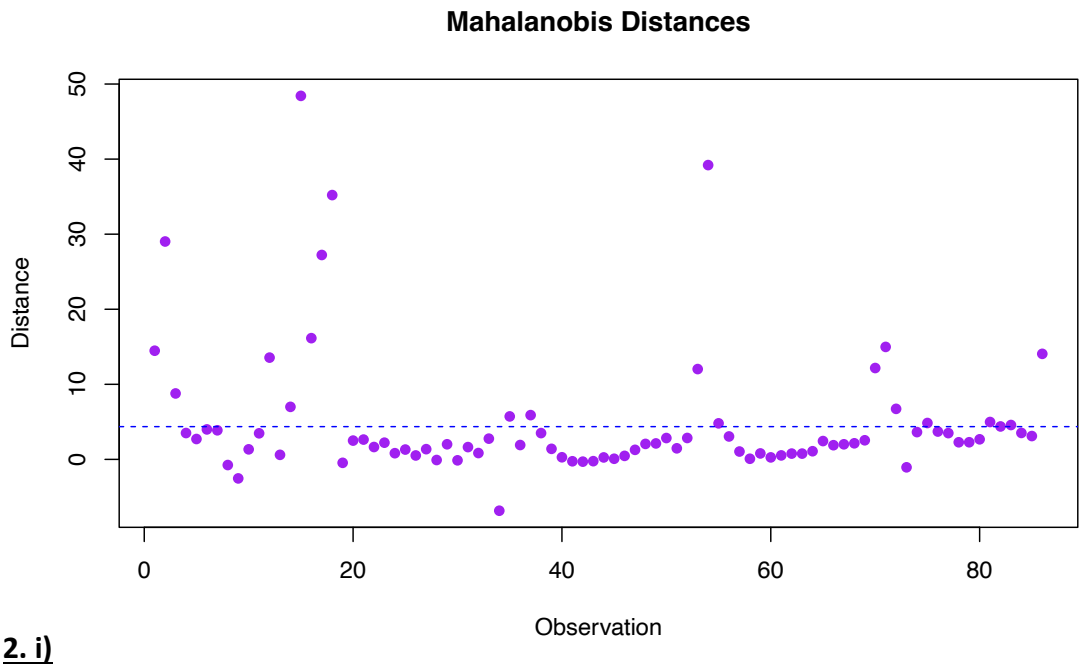
Analyzing the values of the mean and median for each variable, we can see that phouse, pmisc, totexp, income, size and pct, are very similar meaning that the distribution is symmetric.

Covariance

```
> #Covariance#
> dat_73_cov = round(cov(dat_73), digits = 4)
> dat_73_cov
         wfood  whouse   wmisc  phouse   pmisc  totexp  income    size     pct
wfood   0.9009  0.4126  0.4012  0.0005  0.0017 -0.0636 -1.2043  0.0683 -2.0131
whouse  0.4126  1.2051  1.8064  0.0006  0.0002  0.1313  2.1379 -0.8625 -2.8326
wmisc   0.4012  1.8064  3.8677  0.0003 -0.0018  0.2709  5.9230 -1.3445 -4.8972
phouse  0.0005  0.0006  0.0003  0.0000  0.0000  0.0001 -0.0002 -0.0002 -0.0017
pmisc   0.0017  0.0002 -0.0018  0.0000  0.0000 -0.0007 -0.0191 -0.0018  0.0037
totexp -0.0636  0.1313  0.2709  0.0001 -0.0007  0.0599  1.0682  0.0094 -0.4262
income -1.2043  2.1379  5.9230 -0.0002 -0.0191  1.0682 25.1393  0.6150 -9.9567
size    0.0683 -0.8625 -1.3445 -0.0002 -0.0018  0.0094  0.6150  2.9364  0.3979
pct    -2.0131 -2.8326 -4.8972 -0.0017  0.0037 -0.4262 -9.9567  0.3979 17.0026
```

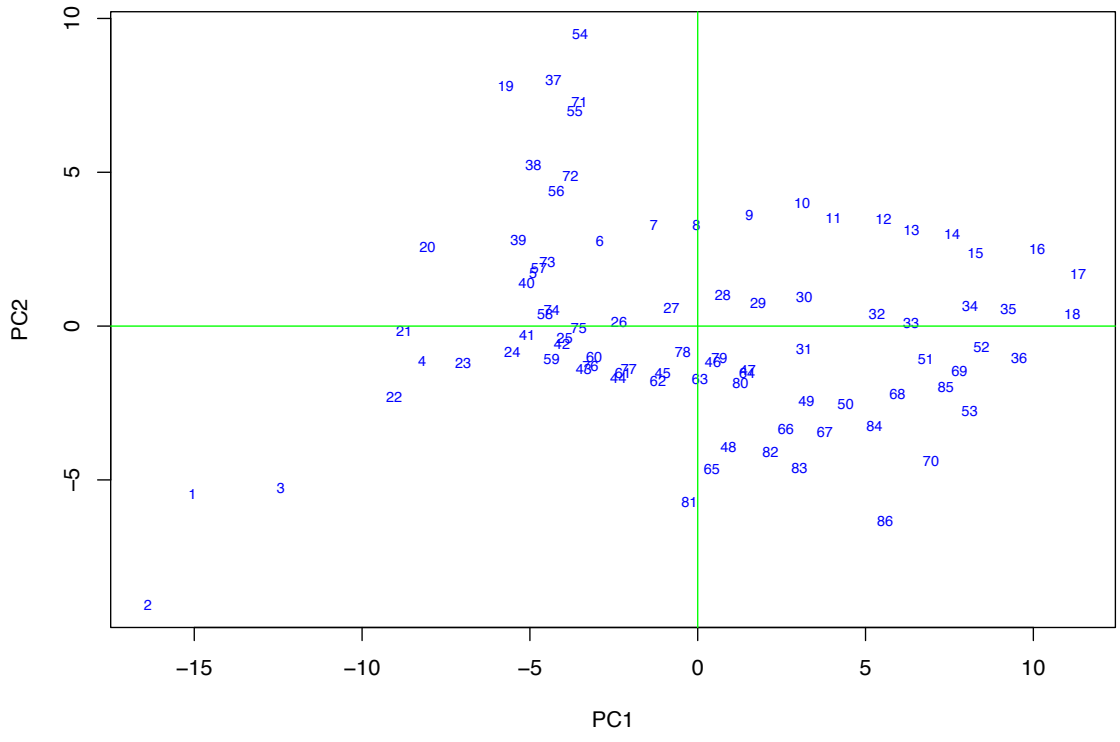
Variance

```
> dat_73_var = round(apply(dat_73, 2, var), digits = 4)
> dat_73_vartot = round(sum(dat_73_var), digits = 4)
> dat_73_vartot
[1] 51.1119
```



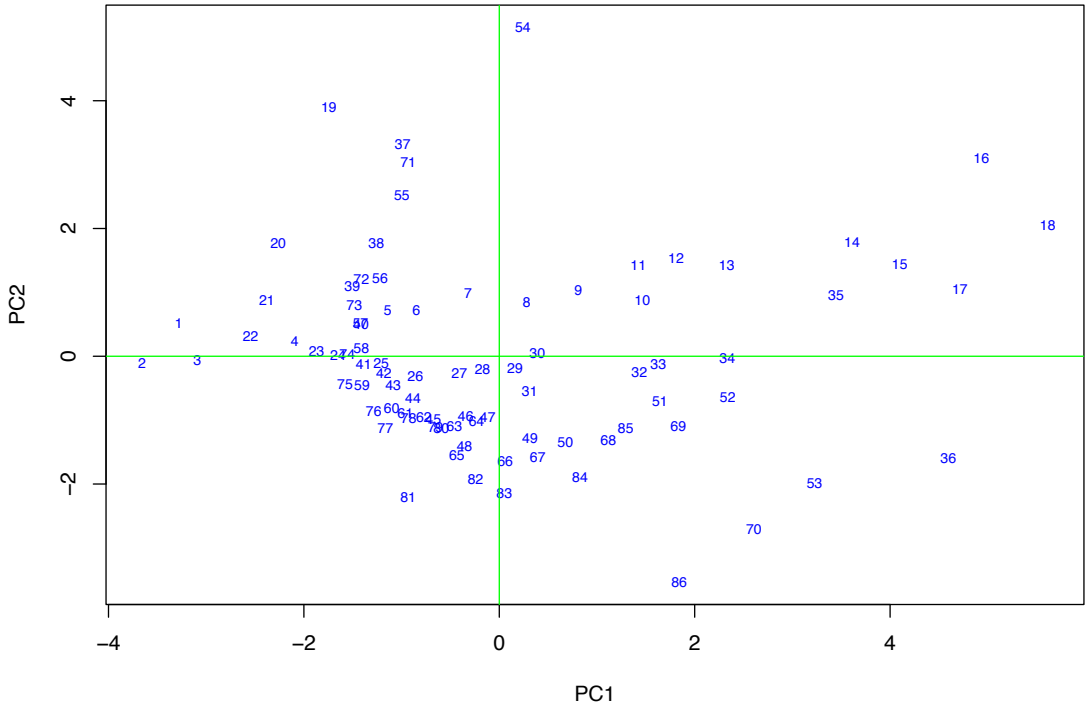
2. i)

```
Standard deviations (1, .., p=9):
[1] 5.851797183 3.344893427 2.016193020 1.066023794 0.524552256 0.441286429 0.093027437 0.003133405 0.001216924

Rotation (n x k) = (9 x 9):
                PC1           PC2           PC3           PC4           PC5           PC6           PC7           PC8           PC9
wfood   8.841211e-03  0.2247595681 -5.292400e-02  3.370465e-01 -0.6294434041  0.6546674653 -0.090628400  0.0006567126 -1.106418e-03
whouse  1.120935e-01  0.1281884502  2.812497e-01  2.798690e-01 -0.5640638859 -0.6898985032  0.139448963 -0.0007907288  2.561341e-04
wmisc   2.496289e-01  0.1023187197  5.158382e-01  6.417128e-01  0.4809459183  0.1315034004 -0.027000701 -0.0008449225  1.926796e-04
phouse  2.696112e-05  0.0001488031  4.326958e-06  3.329495e-05 -0.0010879488 -0.0009676312 -0.011870948 -0.1099708705  9.938628e-01
pmisc  -5.123386e-04  0.0007774599  5.148334e-04  8.140204e-04  0.0003818402 -0.0016734691 -0.004275520  0.9939282271  1.099257e-01
totexp  3.390530e-02 -0.0238491573  6.692404e-03  4.176383e-03 -0.0355964571 -0.1644460404 -0.984758235 -0.0031093704 -1.230284e-02
income  7.899824e-01 -0.5785721416 -4.034437e-02 -9.041812e-02 -0.1470484370  0.0940462456  0.030153912  0.0012443234  4.968445e-04
size   -5.219430e-03 -0.1079481904 -7.694423e-01  5.829939e-01  0.1286743023 -0.1976208573  0.028025466 -0.0002866030  2.516043e-04
pct    -5.475336e-01 -0.7586921600  2.414425e-01  2.195501e-01 -0.1219639345  0.0566787813 -0.002962307  0.0001335554  2.110822e-05
```



ii)

```
Standard deviations (1, .., p=9):
[1] 1.9114436 1.4911562 1.1135763 0.9746603 0.6551200 0.5183776 0.3424264 0.2870913 0.1877419

Rotation (n x k) = (9 x 9):
                PC1        PC2         PC3         PC4         PC5         PC6         PC7        PC8         PC9
wfood     0.07224769  0.5232317 -0.43247509  0.27065009  0.19912335 -0.32651204 -0.07044712  0.5351650 -0.15198066
whouse    0.42211622  0.3198689  0.15153863 -0.01782561 -0.16995172 -0.28146791 -0.62784711 -0.3556345  0.25970056
wmisc     0.45341470  0.1466416  0.23041003  0.18229196 -0.17122020 -0.34344830  0.68260711 -0.2137879 -0.16631711
phouse    0.14024710  0.2131578 -0.33828545 -0.83808156  0.14736934 -0.03836403  0.23827715 -0.0722351  0.18136515
pmisc    -0.23253904  0.4886306  0.12623409 -0.05578643 -0.70523250  0.33061564  0.11489475  0.2080142  0.15807668
totexp    0.43097220 -0.2720412 -0.03795584 -0.27426638 -0.35183077  0.10628594 -0.23130008  0.3057148 -0.61769817
income    0.42870966 -0.3543488 -0.02476320  0.10528164 -0.06841822  0.05880053  0.09392370  0.4763345  0.65961560
size     -0.11823181 -0.2756725 -0.71671861  0.17680096 -0.46563805 -0.20938869  0.02812433 -0.3147969  0.06731479
pct      -0.39212479 -0.2065134  0.30133210 -0.26484474 -0.19440833 -0.72583238 -0.04246844  0.2710651  0.05234040
```



After applying the principal component considering variables in the original scale and the classical sample covariance estimate (i) and standardized variables (ii) we can start analyzing.

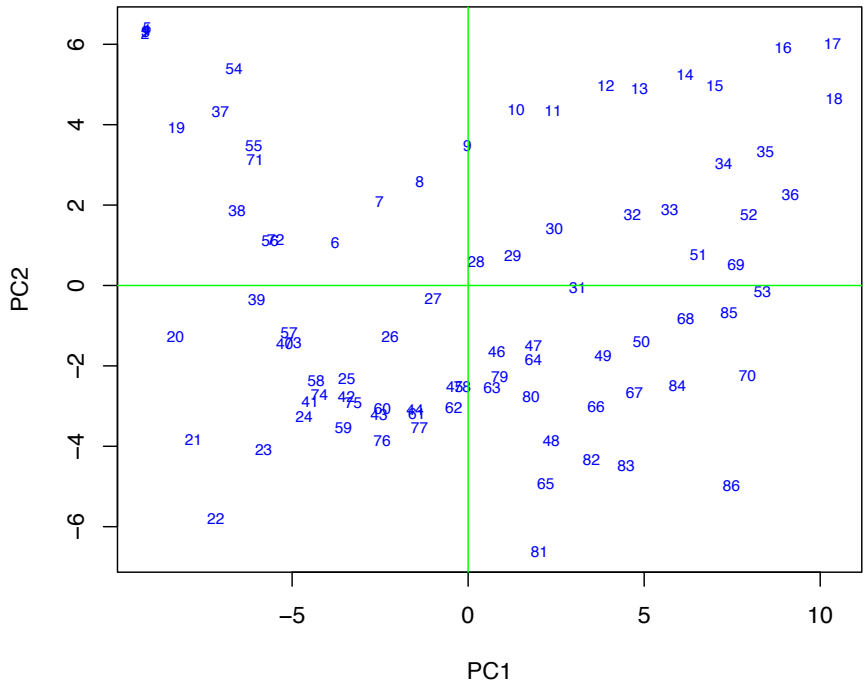
2.b)

```
> summary(dat_73_budget.pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7      PC8      PC9
Standard deviation     5.852 3.3449 2.01619 1.06602 0.52455 0.44129 0.09303 0.003133 0.001217
Proportion of Variance 0.670 0.2189 0.07953 0.02223 0.00538 0.00381 0.00017 0.000000 0.000000
Cumulative Proportion  0.670 0.8889 0.96840 0.99064 0.99602 0.99983 1.00000 1.000000 1.000000
> summary(dat_73_pca_stand)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     1.911 1.4912 1.1136 0.9747 0.65512 0.51838 0.34243 0.28709 0.18774
Proportion of Variance 0.406 0.2471 0.1378 0.1056 0.04769 0.02986 0.01303 0.00916 0.00392
Cumulative Proportion  0.406 0.6530 0.7908 0.8963 0.94404 0.97390 0.98693 0.99608 1.00000
```

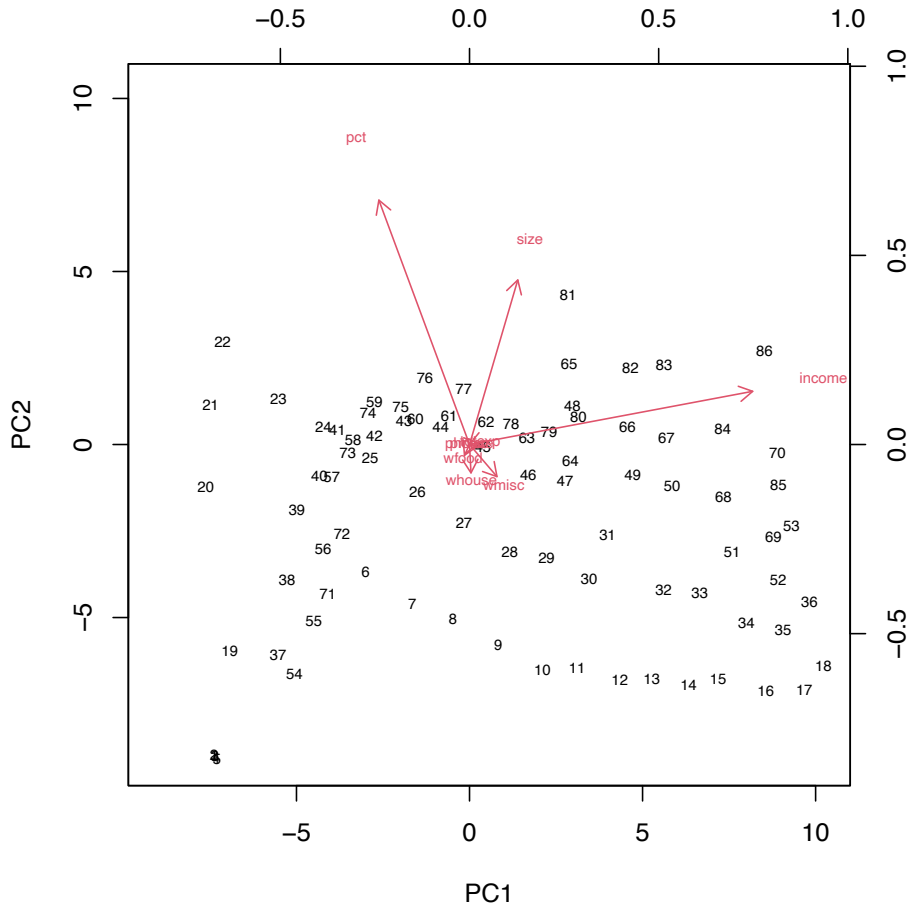
With these summaries we can see that the Cumulative Proportion of the principal components in the original scale needs less components than the standardized to reach the 80% mark. With this we conclude that the Original Scale is more recommended because we can ignore more Principal Components.

3.a)

```
> summary(dat_73_budget.pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7      PC8      PC9
Standard deviation     5.852 3.3449 2.01619 1.06602 0.52455 0.44129 0.09303 0.003133 0.001217
Proportion of Variance 0.670 0.2189 0.07953 0.02223 0.00538 0.00381 0.00017 0.000000 0.000000
Cumulative Proportion  0.670 0.8889 0.96840 0.99064 0.99602 0.99983 1.00000 1.000000 1.000000
> summary(dat_73_p3.pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8    PC9
Standard deviation     5.5211 3.4701 1.59948 1.39986 0.66055 0.45097 0.09104 0.02203 0.0024
Proportion of Variance 0.6392 0.2525 0.05364 0.04109 0.00915 0.00426 0.00017 0.00001 0.0000
Cumulative Proportion  0.6392 0.8917 0.94531 0.98640 0.99555 0.99982 0.99999 1.00000 1.0000
```



3.b)



In classical PCA, the introduction of outliers can heavily influence principal components, potentially leading to incorrect interpretations of data and an inaccurate analysis of correlations between variables.

On the contrary, robust PCA, particularly when based on the MCD estimate, is designed to be less sensitive to outliers. As a result, the summary exhibits a more consistent behavior in the presence of atypical observations, which provides a more accurate interpretation.