MVA Project - PCA

Chun-Jung Chen & Akshay Arora

Question 1: Deos Family conditions affect students' final grade in Math?

PCA Results

- Increase of family size tends to result in the decrease of family support (i.e. no support)

```
d3.q1_Pca = prcomp(d3.q1[, -1], scale = T)
d3.q1_Pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.0902200 1.0127093 0.9773859 0.9113489
##
## Rotation (n x k) = (4 x 4):
##                PC1        PC2        PC3         PC4
## famsize  0.6795722 -0.1108343 -0.1169493  0.71569567
## Pstatus -0.5947037 -0.2950723  0.4551637  0.59336816
## famrel  -0.1921984 -0.7183710 -0.6675118 -0.03782693
## famsup  -0.3841471  0.6201542 -0.5775610  0.36641923
```

- According to the proportion of Variance, it is reasonable to take PC1, PC2, and PC3 since it contains about 80% of variability.

```
summary(d3.q1_Pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4
## Standard deviation     1.0902 1.0127 0.9774 0.9113
## Proportion of Variance 0.2971 0.2564 0.2388 0.2076
## Cumulative Proportion  0.2971 0.5535 0.7924 1.0000
```

- Eigenvalue

```
eigen_d3.q1 = d3.q1_Pca$sdev^2
names(eigen_d3.q1) = paste("PC", 1:4, sep = "")
eigen_d3.q1
```

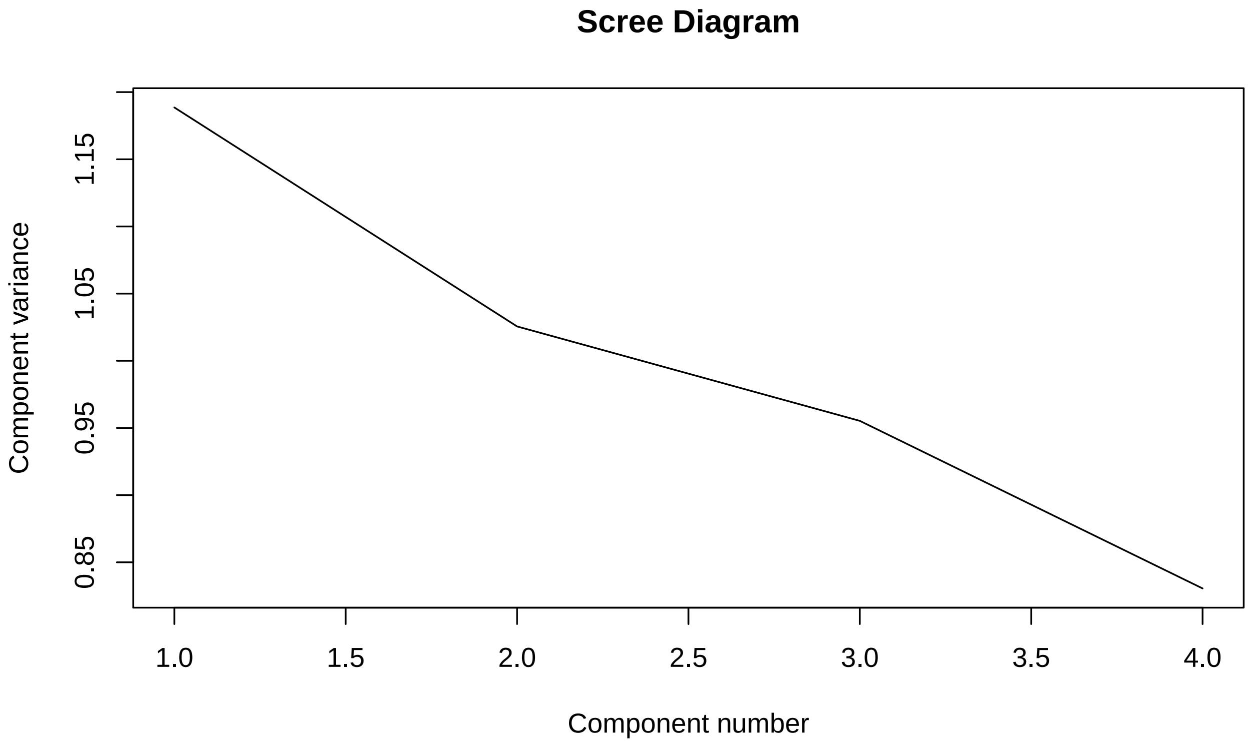
```
##       PC1       PC2       PC3       PC4
## 1.1885797 1.0255802 0.9552832 0.8305569
```

- Variance

```
## [1] 4
```

Scree Diagram

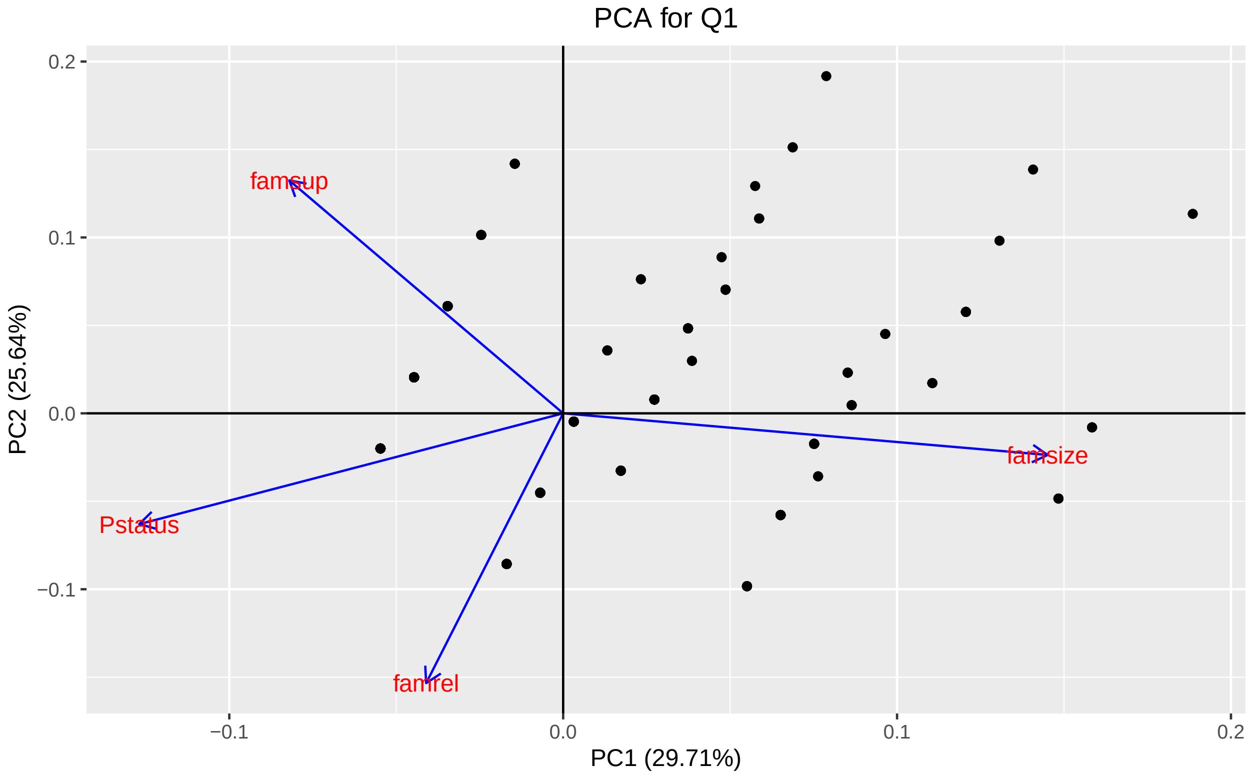
- The Scree Diagram shows the component variance of Principle Components, the conclusion is the same as the proportion of variance showed above.



Biplot and Loading plot

- Plot below shows the visualization of Eigenvectors
  - The loading plot shows that the family size is the most significant variable contributed to PC1, while the family relationship is the least. However, family relationship occupies the largest proportaion of PC2.
  - It is clear that the family size and the other three variables (family support, PStatus, and family relationship) are negative correlation.

```r
biplot(d3.q1_Pca, "black", "PCA for Q1")
```



Question 2: Does parents' jobs and education level influence students' first period of grade in Math?

```
d3.q2_Pca = prcomp(d3.q2[, -1], scale = T)
d3.q2_Pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.4152171 0.9621460 0.8701262 0.5606390
##
## Rotation (n x k) = (4 x 4):
##            PC1        PC2         PC3         PC4
## Medu 0.6125510 -0.2900260  0.03946222  0.73424034
## Fedu 0.5599267 -0.2714852  0.50121196 -0.60130228
## Mjob 0.4738501  0.1312280 -0.81763246 -0.29953708
## Fjob 0.2944987  0.9082730  0.28054673  0.09800114
```

- For question 2, we will take PC1 and PC2 since it entails over 70% of variance.

```
summary(d3.q2_Pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4
## Standard deviation     1.4152 0.9621 0.8701 0.56064
## Proportion of Variance 0.5007 0.2314 0.1893 0.07858
## Cumulative Proportion  0.5007 0.7321 0.9214 1.00000
```

```
eigen_d3.q2 = d3.q2_Pca$sdev^2
names(eigen_d3.q2) = paste("PC", 1:4, sep = "")
eigen_d3.q2
```

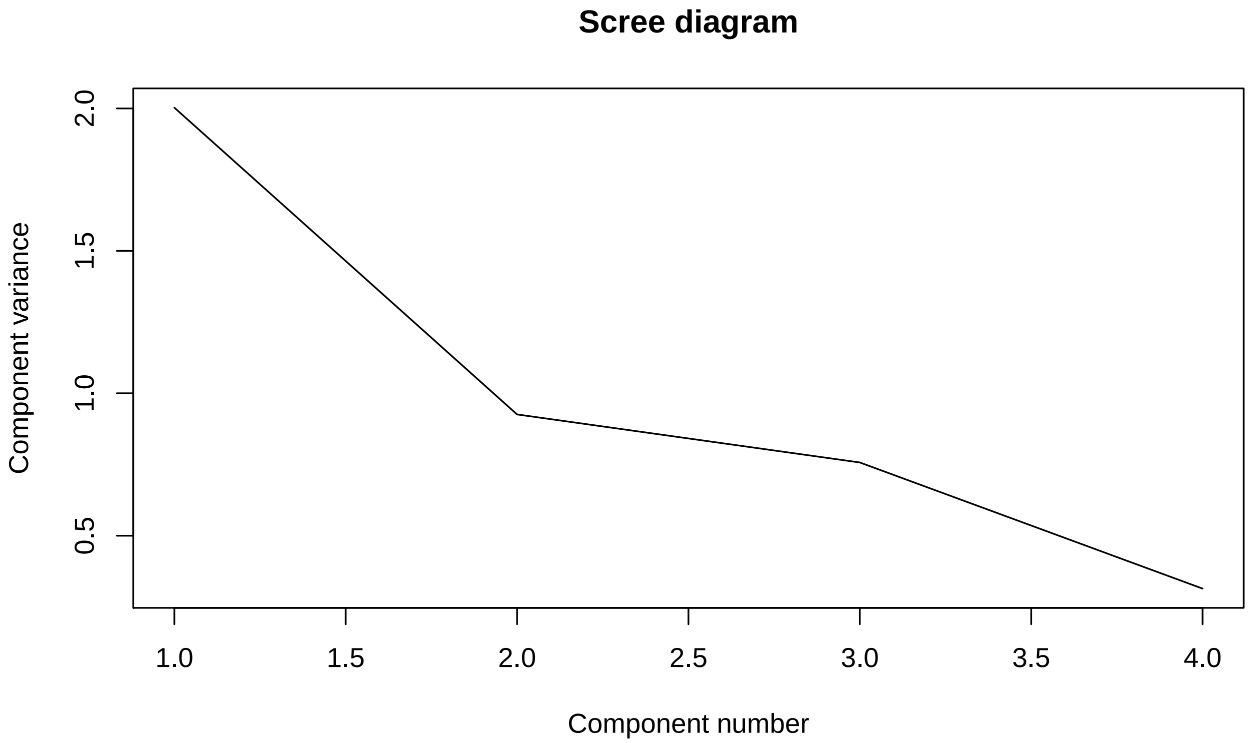
```
##       PC1       PC2       PC3       PC4
## 2.0028395 0.9257249 0.7571196 0.3143161
```

```
var_d3.q2 = sum(eigen_d3.q2)
var_d3.q2
```

```
## [1] 4
```

Scree Diagram

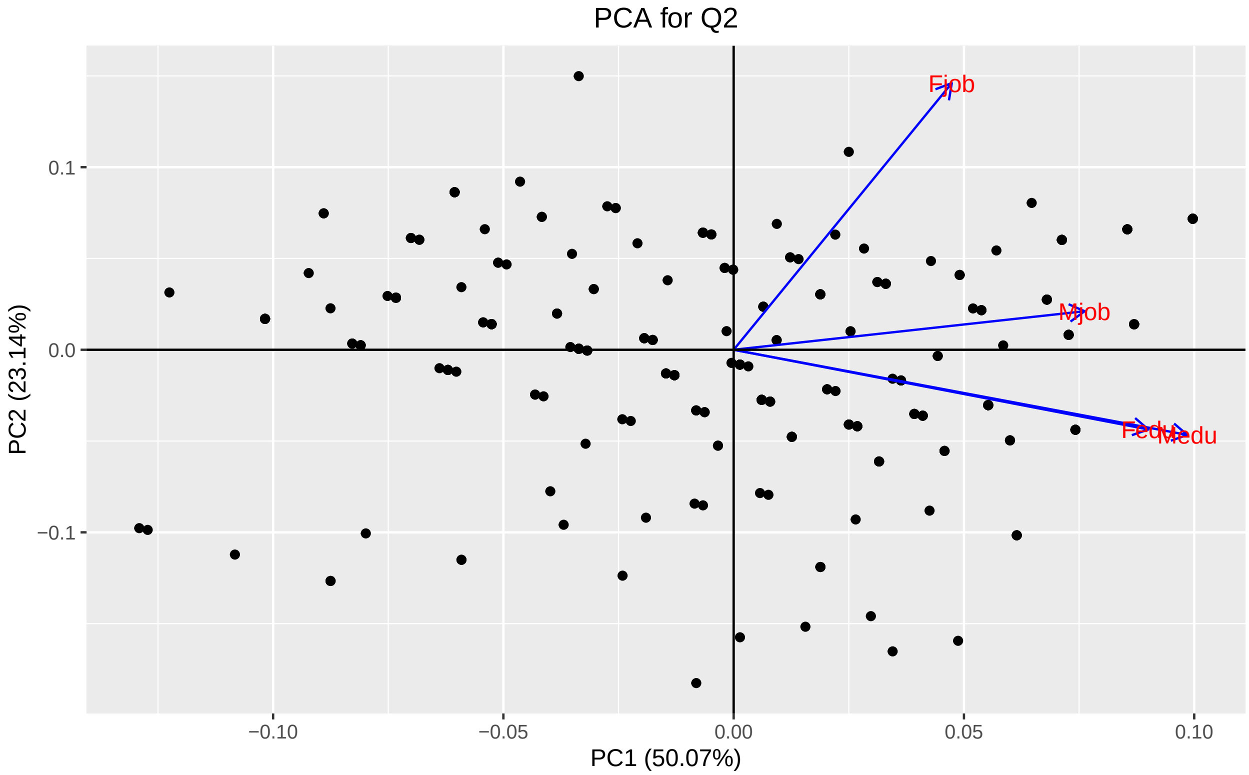
```
plot(eigen_d3.q2, xlab = "Component number", ylab = "Component variance",
    type = "l", main = "Scree diagram")
```



Biplot and Loading plot

- Plot below shows the visualization of Eigenvectors
  - the loading plot illustrates that all variables in PC1 are positive correlated, while parents type of job and their education level are negative.
  - It is obvious to observe that the eigenvectors between parents' education levels are highly overlapped, means that these variables may generate multicolinearity. Therefore, either Father's education or Mother's education can be eliminated.

```
biplot(d3.q2_Pca, "Black", "PCA for Q2")
```



Question 3: Does student's learning conditions really impact students' final grade math score and Portuguese scores in average?

```
d3.q3_Pca = prcomp(d3.q3[, c(-1, -3)], scale = T)
d3.q3_Pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.0734348 0.9943629 0.9855512 0.9421619
##
## Rotation (n x k) = (4 x 4):
##                     PC1        PC2         PC3         PC4
## internet       -0.6449942  0.1369274 -0.06562915 -0.74895010
## romantic       -0.4796756  0.4876147  0.57251255  0.45207587
## freetime       -0.5132515 -0.1280969 -0.69988538  0.47992135
## normtraveltime  0.3007768  0.8526863 -0.42199839 -0.06615633
```

- PCA for Q3 contains three variables including Internet, romantic, and free time, normalized travel time. PC1 through PC3 consist of almost 80% of the variance.

```
summary(d3.q3_Pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4
## Standard deviation     1.0734 0.9944 0.9856 0.9422
## Proportion of Variance 0.2881 0.2472 0.2428 0.2219
## Cumulative Proportion  0.2881 0.5353 0.7781 1.0000
```

```
eigen_d3.q3 = d3.q3_Pca$sdev^2
names(eigen_d3.q3) = paste("PC", 1:4, sep = "")
eigen_d3.q3
```

```
##       PC1       PC2       PC3       PC4
## 1.1522623 0.9887576 0.9713111 0.8876690
```

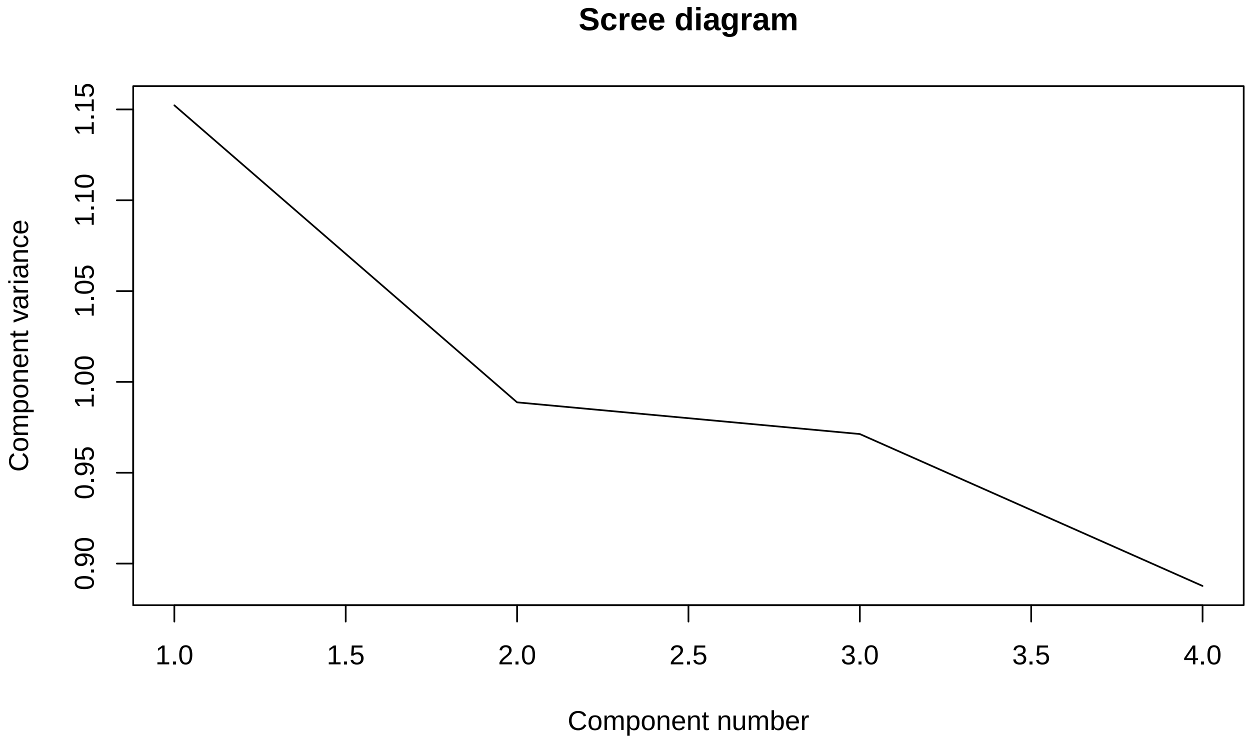
```
var_d3.q3 = sum(eigen_d3.q3)
var_d3.q3
```

```
## [1] 4
```

Scree Diagram

- The Scree Diagram shows that the variance of PC2 and PC3 are almost the same.

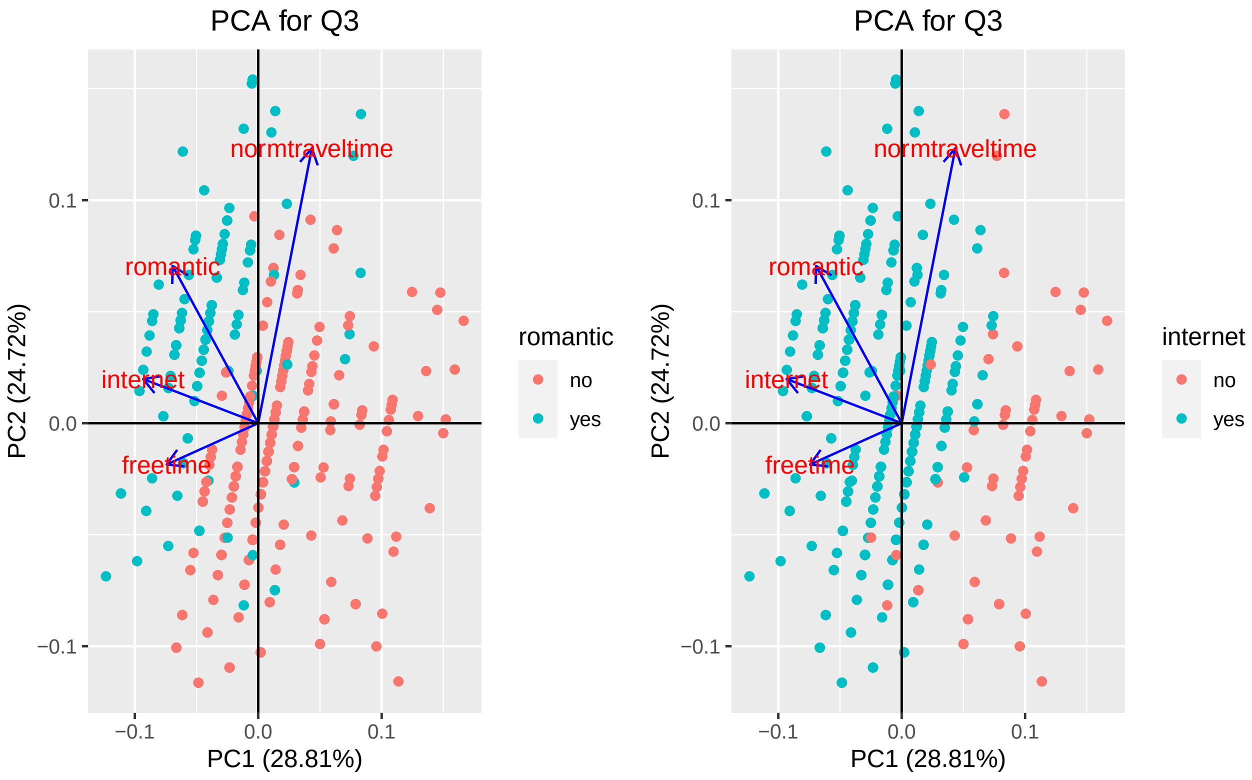
```r
plot(eigen_d3.q3, xlab = "Component number", ylab = "Component variance",
    type = "l", main = "Scree diagram")
```



Biplot and Loading plot

- Plot below shows the visualization of Eigenvectors
  - According to the loading plot, it is clear that internet and free time variables contributed the most for PC1. Also, romantic and normalized travel time variables are the most siginificant for PC2.
  - There are two biplots for Q3 PCA result, left plot is colored by romanic variables and the right is by internet access.
  - For the left biplot, we can conclude that if the parents are at the romantic relationship, it is highly possble that the PC1 is negative and small, i.e. either travel time is less or there is an internet access at home.
  - For the right biplot, family doesn't have internet access tends to have positive PC1 value and the reason would be the higher travel time.

```r
# install.packages('patchwork')
library(patchwork)
biplot(d3.q3_Pca, "romantic", "PCA for Q3") + biplot(d3.q3_Pca,
    "internet", "PCA for Q3")
```


PCA for Q3
PC2 (24.72%)
PC1 (28.81%)
normtraveltime
romantic
internet
freetime
romantic
no
yes
PCA for Q3
PC2 (24.72%)
PC1 (28.81%)
normtraveltime
romantic
internet
freetime
internet
no
yes