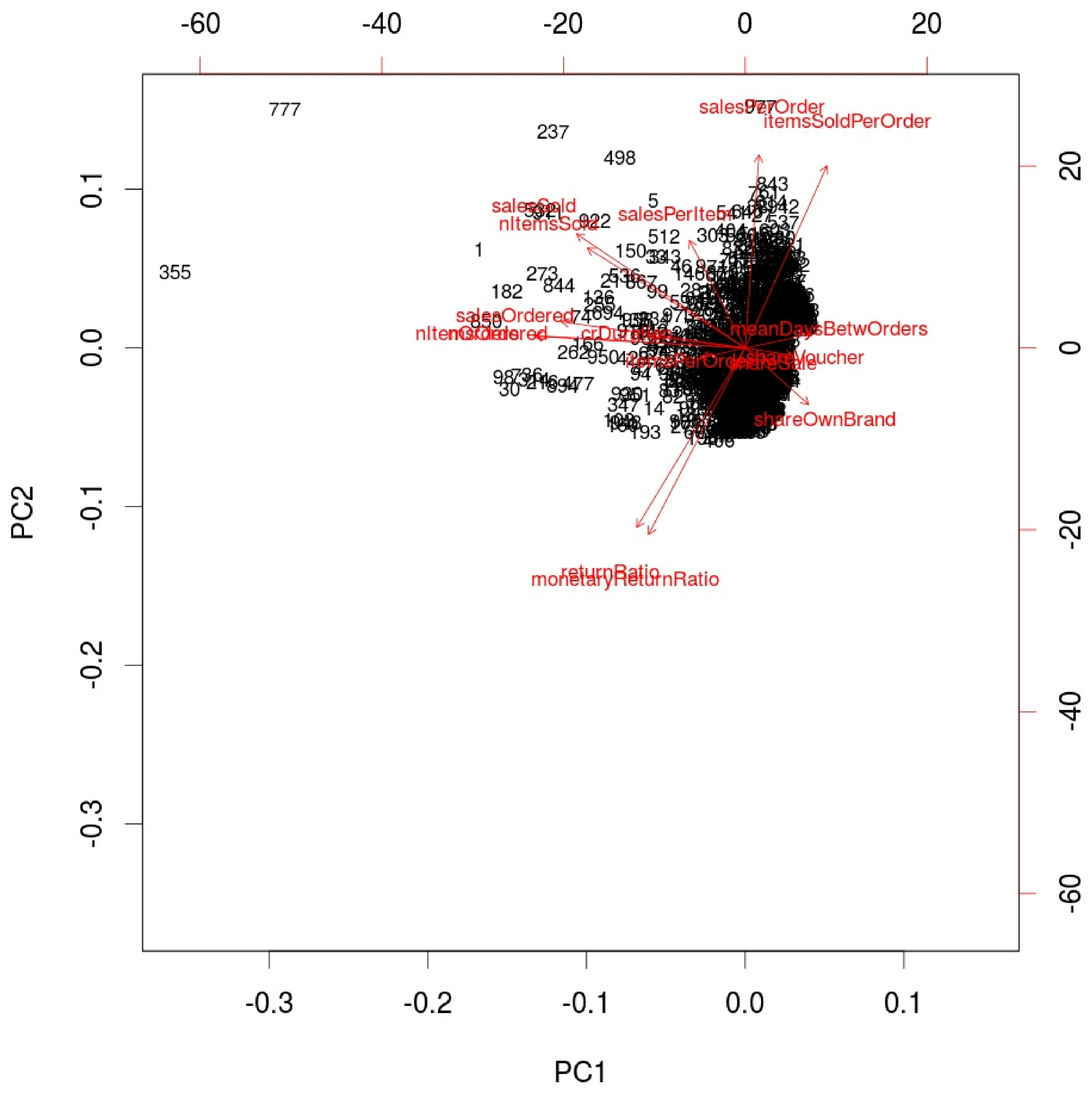
MARKETING ANALYTICS IN R: STATISTICAL MODELING

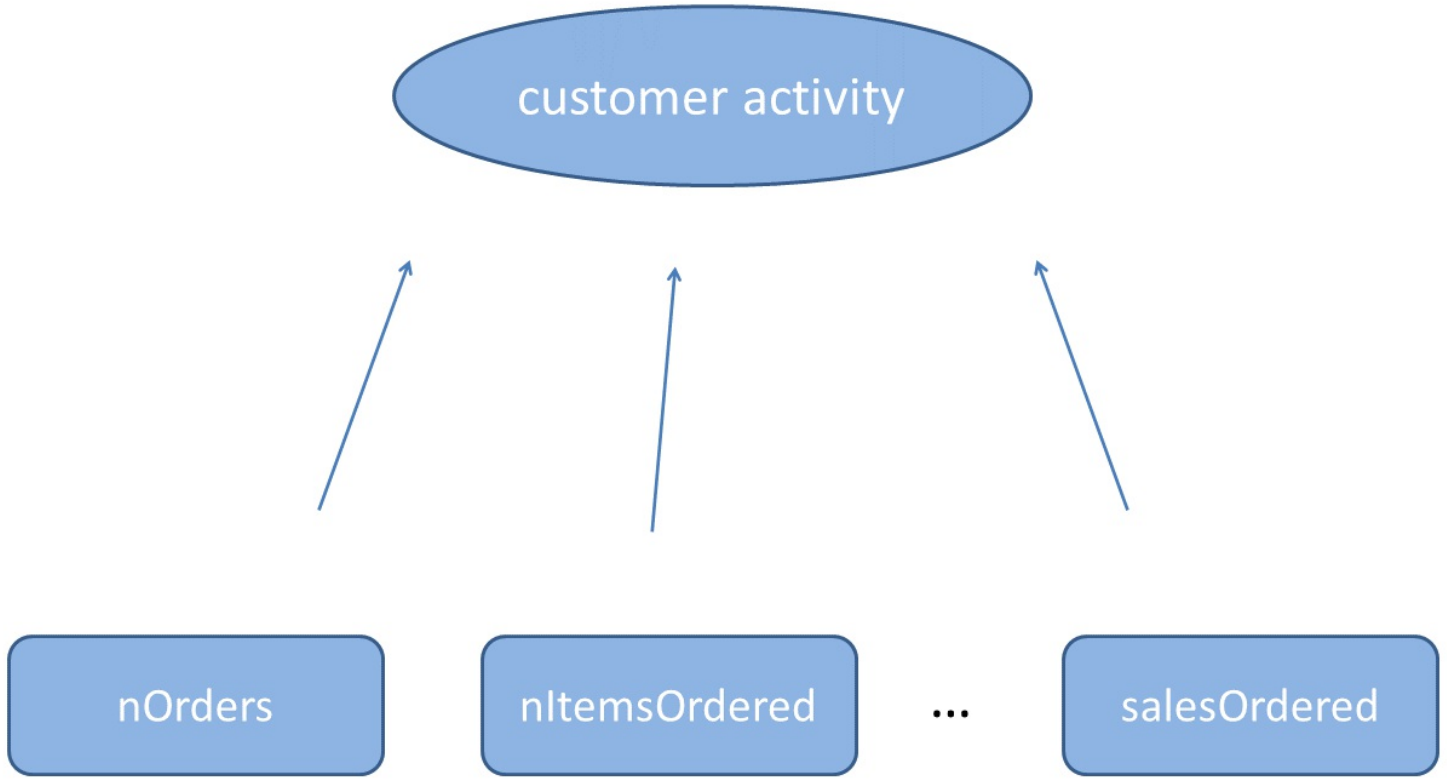
Principal Component Analysis for CRM Data

Verena Pflieger
Data Scientist at INWT Statistics


PC1
PC2
salesPerOrder
itemsSoldPerOrder
salesSold
nItemsSold
salesPerItem
salesOrdered
nItemsOrdered
meanDaysBetwOrders
shareVoucher
shareOwnBrand
returnRatio
monetaryReturnRatio

PCA helps to...

- handle multicollinearity
- create indices
- visualize and understand high-dimensional data



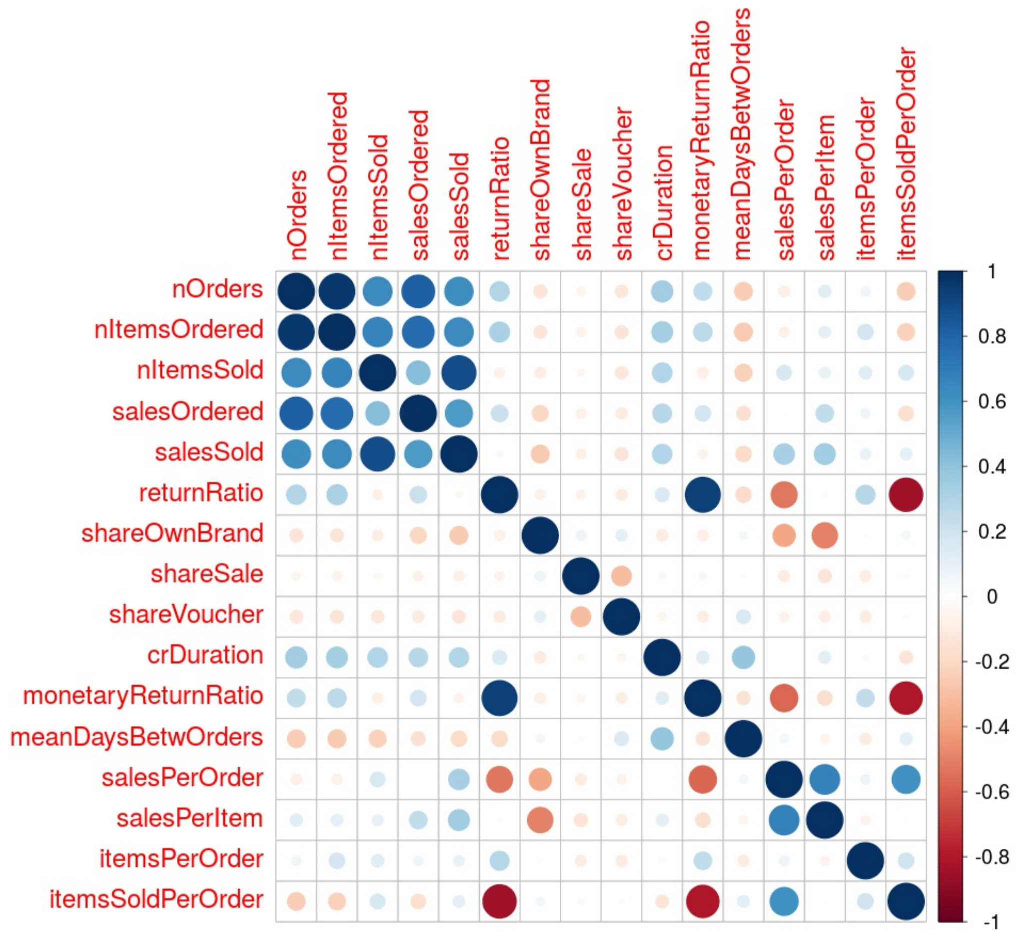
Data for PCA

```
str(dataCustomers, give.attr = FALSE)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':    989 obs. of  16 variables:
 $ nOrders             : int  104 17 5 18 21 2 18 12 14 7 ...
 $ nItemsOrdered       : int  138 21 6 27 41 2 29 14 19 13 ...
 $ nItemsSold          : int  66 4 3 3 35 1 11 11 9 2 ...
 $ salesOrdered        : num  37813 10653 1226 31529 17935 ...
 $ salesSold           : num  18031 1500 759 3803 14246 ...
 $ returnRatio         : num  0.522 0.81 0.5 0.889 0.146 ...
 $ shareOwnBrand       : num  0.54 0.48 1 0.15 0.63 0 1 1 0.42 0.31 ...
 $ shareSale           : num  0.52 0.67 0.17 0.19 0.12 0 0.28 0.07 0.37 ...
 $ shareVoucher        : num  0.09 0.1 0.5 0.07 0 0 0.52 0.29 0.16 0 ...
 $ crDuration          : int  1472 1506 1453 1340 1449 749 997 1513 1499 ...
 $ monetaryReturnRatio: num  0.523 0.859 0.381 0.879 0.206 ...
 $ meanDaysBetwOrders : int  14 94 363 79 72 749 59 138 115 254 ...
 $ salesPerOrder       : num  173.4 88.2 151.8 211.3 678.4 ...
 $ salesPerItem        : num  273 375 253 1268 407 ...
 $ itemsPerOrder       : num  1.33 1.24 1.2 1.5 1.95 1 1.61 1.17 1.36 ...
 $ itemsSoldPerOrder   : num  0.63 0.24 0.6 0.17 1.67 0.5 0.61 0.92 0.64 ...
```

Correlation Structure

```
library(corrplot)
dataCustomers %>% cor() %>% corrplot()
```



MARKETING ANALYTICS IN R: STATISTICAL MODELING

Let's practice!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

PCA Computation

Verena Pflieger
Data Scientist at INWT Statistics

Status Quo

```
# Variances of all variables before any data preparation
lapply(dataCustomers, var)
```

```
$nOrders                        $salesOrdered
[1] 264.3989                           [1] 202384132

$nItemsOrdered                  $salesSold
[1] 506.5496                           [1] 9345112

$nItemsSold                     $returnRatio
[1] 56.35125                           [1] 0.0836261
...
```

Standardization

```
dataCustomers <- dataCustomers %>% scale() %>% as.data.frame()
```

```
# Check variances of all variables
lapply(dataCustomers, var)
```

```
$nOrders                        $salesOrdered
[1] 1                           [1] 1

$nItemsOrdered                  $salesSold
[1] 1                           [1] 1

$nItemsSold                     $returnRatio
[1] 1                           [1] 1
...
```

PCA Computation

```
pcaCust <- prcomp(dataCustomers)

str(pcaCust, give.attr = FALSE)
```

```
List of 5
 $ sdev    : num [1:16] 2.1 1.84 1.3 1.2 1.12 ...
 $ rotation: num [1:16, 1:16] -0.439 -0.44 -0.33 -0.384 -0.352 ...
 $ center  : Named num [1:16] -4.66e-17 1.90e-17 -1.24e-18 6.69e-18 ...
 $ scale   : logi FALSE
 $ x       : num [1:989, 1:16] -11.06 -1.67 0.53 -3.39 -3.81 ...
```

Standard Deviations of the Components

```
# Standard deviations
pcaCust$sdev %>% round(2)
```

```
 [1] 2.10 1.84 1.30 1.20 1.12 1.07 0.80 0.78 0.72 0.61 0.48 0.37 0.26
[14] 0.21 0.17 0.13
```

```
# Variances (Eigenvalues)
pcaCust$sdev ^ 2 %>% round(2)
```

```
 [1] 4.39 3.38 1.68 1.45 1.26 1.15 0.65 0.61 0.52 0.38 0.23 0.14 0.07
[14] 0.04 0.03 0.02
```

```
# Proportion of explained variance
(pcaCust$sdev ^ 2/length(pcaCust$sdev)) %>% round(2)
```

```
 [1] 0.27 0.21 0.10 0.09 0.08 0.07 0.04 0.04 0.03 0.02 0.01 0.01 0.00
[14] 0.00 0.00 0.00
```

Loadings and Interpretation

```
# Loadings (correlations between original variables and components)
round(pcaCust$rotation[, 1:6], 2)
```

	PC1	PC2	PC3	PC4	PC5	PC6
nOrders	**-0.44**	0.03	-0.15	0.05	-0.00	0.13
nItemsOrdered	**-0.44**	0.03	-0.16	0.02	0.04	0.03
nItemsSold	**-0.33**	0.24	-0.27	-0.02	0.04	-0.04
salesOrdered	**-0.38**	0.06	-0.03	0.06	-0.00	0.14
salesSold	**-0.35**	0.27	-0.07	-0.01	0.02	0.01
returnRatio	-0.23	**-0.43**	0.23	-0.05	0.04	-0.14
shareOwnBrand	0.13	-0.13	**-0.54**	0.06	0.08	-0.02
shareSale	0.05	-0.03	-0.19	-0.26	**-0.67**	0.00
shareVoucher	0.10	-0.02	-0.03	**0.40**	**0.54**	0.24
crDuration	-0.20	0.03	0.02	**0.54**	-0.29	-0.29
monetaryReturnRatio	-0.20	**-0.44**	0.17	-0.04	0.03	-0.15
meanDaysBetwOrders	0.14	0.03	0.04	**0.63**	-0.24	-0.28
salesPerOrder	0.03	**0.46**	**0.31**	-0.07	0.02	-0.11
salesPerItem	-0.12	0.26	**0.56**	-0.03	-0.05	0.12
itemsPerOrder	-0.09	-0.02	-0.01	-0.23	**0.31**	**-0.78**
itemsSoldPerOrder	0.17	**0.43**	-0.22	-0.08	0.09	-0.25

Values of the Observations

```
# Value on 1st component for 1st customer
sum(dataCustomers[1,] * pcaCust$rotation[,1])
```

```
[1] -11.05858
```

```
pcaCust$x[1:5, 1:6]
```

```
            PC1        PC2        PC3         PC4        PC5          PC6
[1,] -11.0585802  3.5750683 -4.1371495  0.28864769 -0.1045802  0.698612248
[2,]  -1.6734771 -1.6630208  0.9498452  0.14091195 -1.2760898 -0.006310673
[3,]   0.5303018 -0.4672193 -0.1918865  1.77466781  0.4623840 -0.037466682
[4,]  -3.3903118 -0.1274839  4.2217216  0.03710948 -0.1840454  0.164680941
[5,]  -3.8069613  5.3971530 -1.2241316 -0.38341585  0.9721412 -2.142731490
```

MARKETING ANALYTICS IN R: STATISTICAL MODELING

It's your turn!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Choosing the Right Number of Principal Components

Verena Pflieger
Data Scientist at INWT Statistics

No. Relevant Components: Explained variance

```
# Proportion of variance explained:
summary(pcaCust)
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.0951 1.8379 1.2960 1.20415 1.12301 1.07453 0.80486
Proportion of Variance 0.2743 0.2111 0.1050 0.09062 0.07882 0.07216 0.04049
Cumulative Proportion  0.2743 0.4855 0.5904 0.68106 0.75989 0.83205 0.87254
                           PC8     PC9    PC10    PC11    PC12    PC13
Standard deviation     0.78236 0.72452 0.61302 0.48428 0.36803 0.25901
Proportion of Variance 0.03826 0.03281 0.02349 0.01466 0.00847 0.00419
Cumulative Proportion  0.91079 0.94360 0.96709 0.98175 0.99021 0.99440
                          PC14    PC15    PC16
Standard deviation     0.20699 0.17126 0.13170
Proportion of Variance 0.00268 0.00183 0.00108
Cumulative Proportion  0.99708 0.99892 1.00000
```

No. Relevant Components: Kaiser-Guttman Criterion

Kaiser-Guttman criterion: Eigenvalue > 1

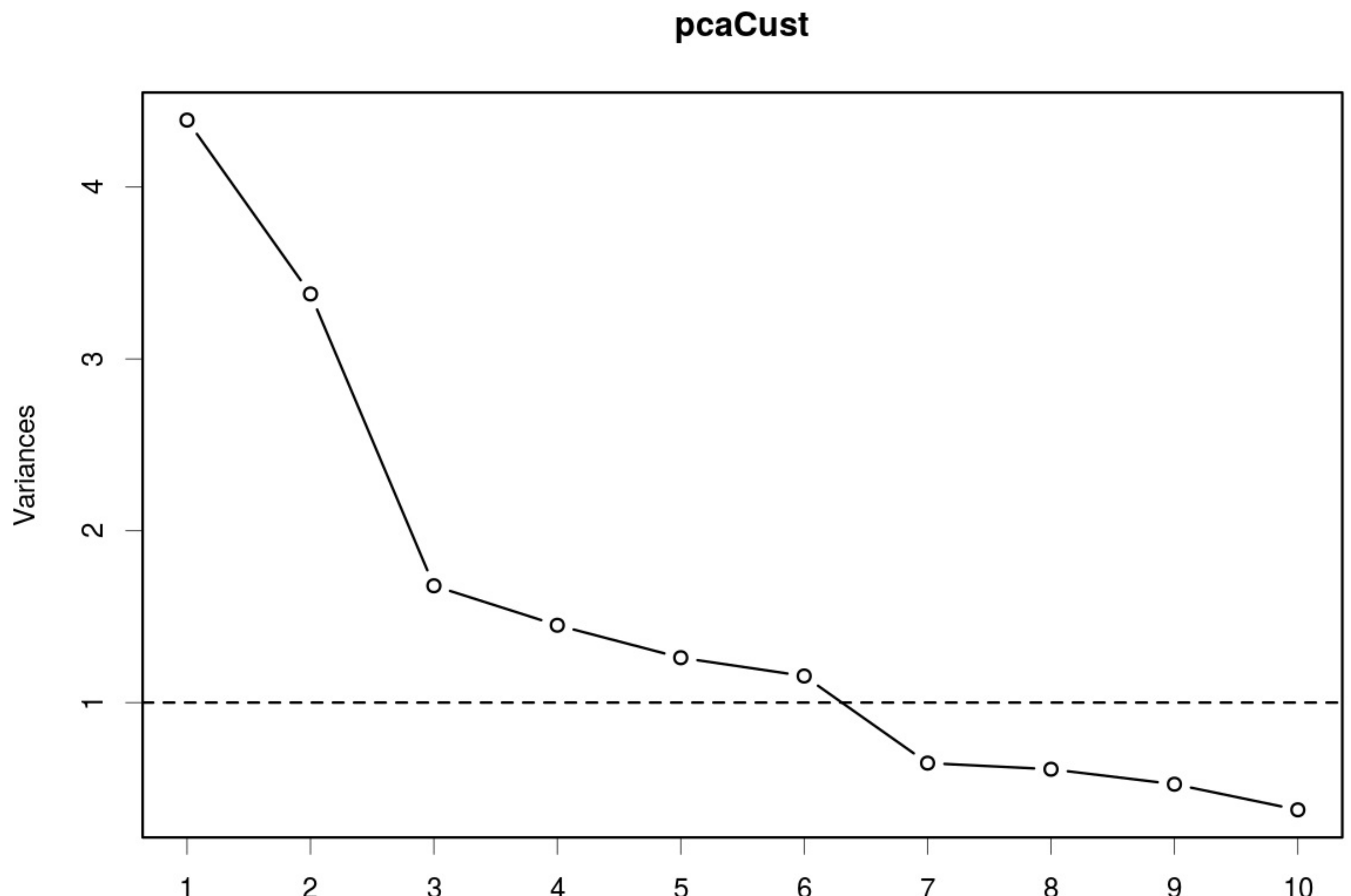
```
pcaCust$sdev ^ 2
```

```
 [1] 4.38961593 3.37778445 1.67965616 1.44997580 1.26115351 1.15461579
 [7] 0.64780486 0.61209376 0.52492468 0.37579685 0.23452736 0.13544710
[13] 0.06708362 0.04284504 0.02933027 0.01734481
```

No. Relevant Components: Screeplot

The screeplot or: "Find the elbow"

```
screeplot(pcaCust, type = "lines")
box()
abline(h = 1, lty = 2)
```

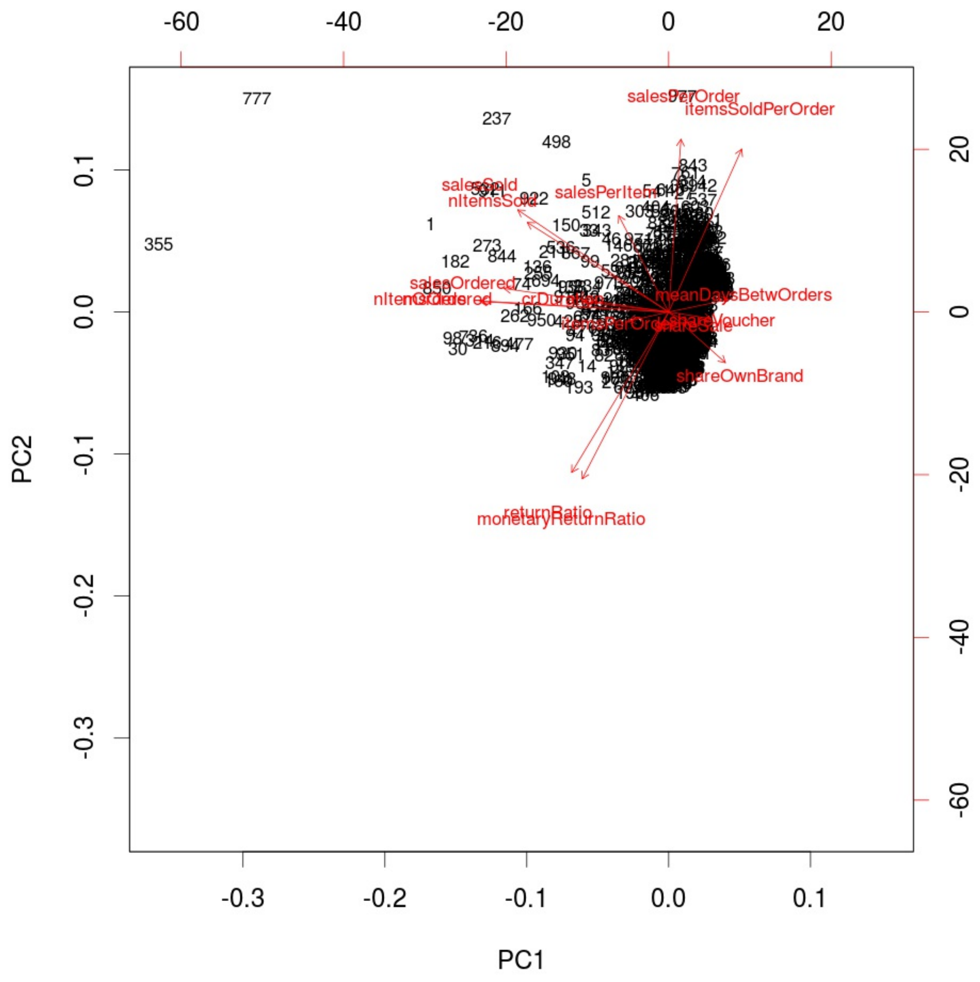


Suggested Number of Components by Criterion

Explained Variance	Kaiser-Guttman	Screeplot
5	6	6

The Biplot

```
biplot(pcaCust, choices = 1:2, cex = 0.7)
```



MARKETING ANALYTICS IN R: STATISTICAL MODELING

Hands on!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Further Analysis and Learnings

Verena Pflieger
Data Scientist at INWT Statistics

PC in Regression Analysis I

```
mod1 <- lm(customerSatis ~ ., dataCustomers)
```

```
library(car)
vif(mod1)
```

```
          nOrders       nItemsOrdered          nItemsSold        salesOrdered
        29.482287           24.437448           10.390998            5.134720
        salesSold         returnRatio       shareOwnBrand           shareSale
         9.685617           23.778800            1.571607            1.178773
     shareVoucher          crDuration monetaryReturnRatio  meanDaysBetwOrders
         1.213011            1.757509           10.632243            1.698369
    salesPerOrder        salesPerItem       itemsPerOrder   itemsSoldPerOrder
         6.563474            4.557981            4.821610           15.949072
```

PC in Regression Analysis II

```
# Create dataframe with customer satisfaction and first 6 components
dataCustComponents <- cbind(dataCustomers[, "customerSatis"],
                            pcaCust$x[, 1:6]) %>%
  as.data.frame
mod2 <- lm(customerSatis ~ ., dataCustComponents)
```

```
vif(mod2)
```

```
PC1 PC2 PC3 PC4 PC5 PC6
  1   1   1   1   1   1
```

```
summary(mod1)$adj.r.squared
```

```
[1] 0.8678583
```

```
summary(mod2)$adj.r.squared
```

```
[1] 0.7123822
```

PC in Regression Analysis III: Interpretation

```
summary(mod2)
```

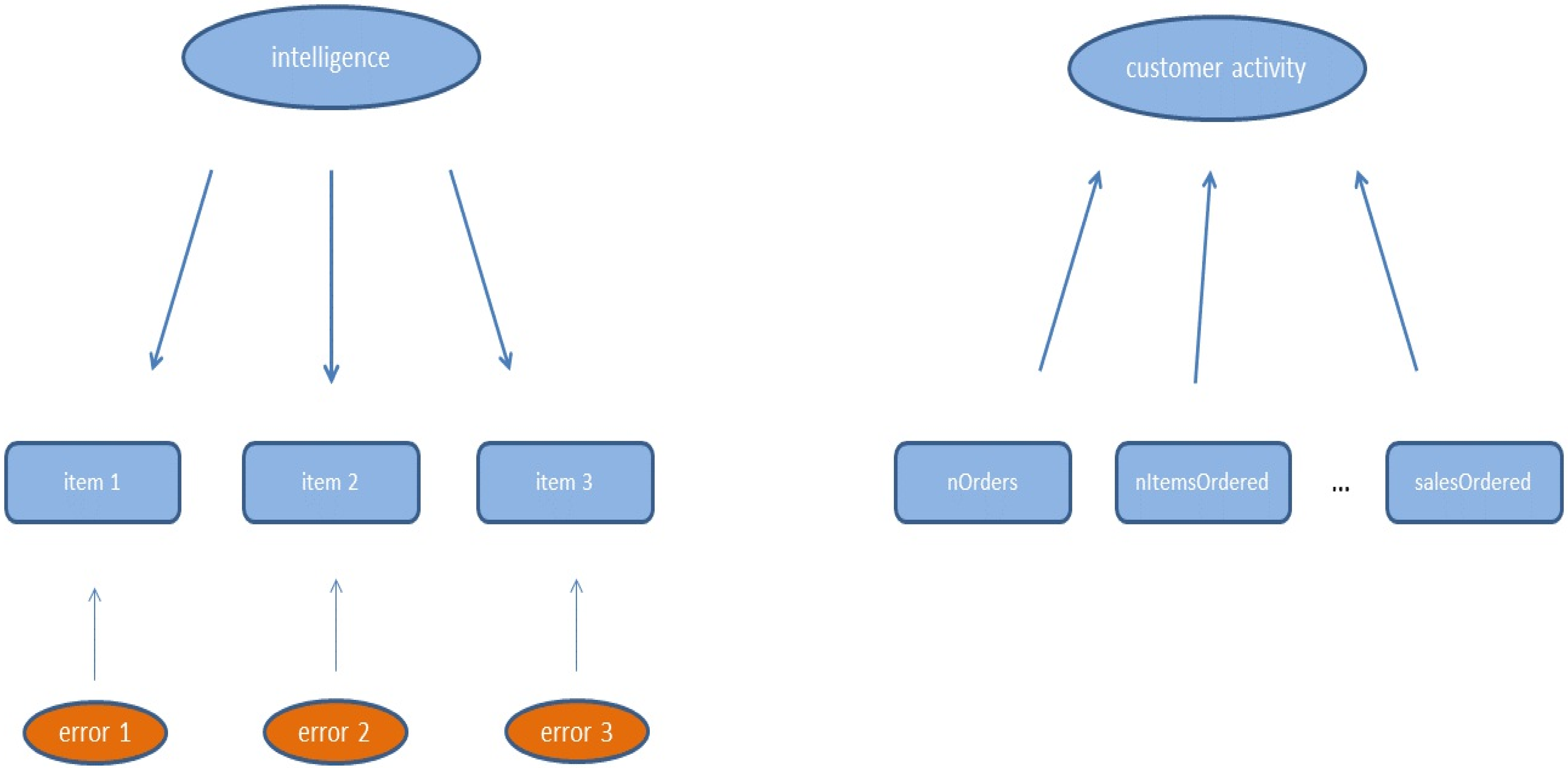
```
Call:
lm(formula = customerSatis ~ ., data = dataCustComponents)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9279 -0.2411  0.0179  0.2865  1.4972

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.985945   0.014039 212.682  < 2e-16 ***
PC1         -0.175434   0.006704 -26.167  < 2e-16 ***
PC2          0.296659   0.007643  38.815  < 2e-16 ***
PC3         -0.012816   0.010838  -1.182    0.237
PC4         -0.116651   0.011665 -10.000  < 2e-16 ***
PC5          0.101963   0.012508   8.152 1.09e-15 ***
PC6          0.126677   0.013072   9.691  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4415 on 982 degrees of freedom
Multiple R-squared:  0.7141,    Adjusted R-squared:  0.7124
F-statistic: 408.9 on 6 and 982 DF,  p-value: < 2.2e-16
```

Factor Analysis vs. PCA



Learnings and Relevance

	Learnings about PCA
You have learned...	to reduce the number of variables without losing too much information
	that variables should be standardized before a PCA
	how to decide on the number of relevant components
	to interpret the selected components

	Learnings from the model
You have learned...	that the original variables can be reduced to 6 components, i.a., customer activity, return behavior and brand awareness
	that using the first six components to explain customer satisfaction causes a decrease in explained variance, but solves the multicollinearity problem

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Let's practice!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Congratulations!

Verena Pflieger
Data Scientist at INWT Statistics