logistic_biplot_script

Camilo Estrella

2024-02-20

Loading libraries

We will start by loading the libraries that will be used, whose function is explained below:

```
library(MultBiplotR)
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(scales)
```

- **MultBiplotR**: It allows multivariate analysis using logistic regression and its graphical representation employing a Biplot.
- **knitr**: The *kable* function is used to obtain a better visualization of the tables.
- **dplyr**: It has multiple functions that allow data management, some of the most interesting are those that allow you to manage tables as if it were a relational database, as is done with SQL.
- **ggplot2**: Perhaps the most important library for the creation of statistical graphics in R.
- **scales**: It is used in conjunction with ggplot2 to scale axes within the plot.

Loading datasets

The original datasets are 5, named as in the code. In addition, there is a sixth one called "cruzada" (a cross of all 5 tables), which has binary information of aisles of a supermarket in the columns and rows representing the customers. The value 1 means that the customer has shopped in that aisle and 0 the opposite.

The calculation of the "cross" dimension of the dataset is included first.

```r
order <- read.csv("Order.csv")
order_product <- read.csv("Order_Product.csv")
product <- read.csv("Product.csv")
aisle <- read.csv("Aisle.csv")
department <- read.csv("Department.csv")

cruzada <- as.data.frame(read.csv("cruzada_df.csv"))
```

```r
dimensions <- as.data.frame(matrix(dim(cruzada), ncol = 2))
names(dimensions) <- c("customers (rows)", "aisles (columns)")
kable(dimensions, row.names = FALSE, align = "cc")
```

customers (rows)	aisles (columns)
206209	134

The datasets are initially from a supermarket and cannot be shared due to source privacy issues.