Présentation R: Online Retail

Arthur Lemoine, François Somville, Alexandre Antippas

Janvier 2019

Plan

1. Analyse des données importées
2. Nettoyage des données
3. Statistiques descriptives
4. Analyse des composantes principales (ACP)
5. Clustering
6. Conclusion
7. Bonus

1. Analyse des données importées

1. Analyse des données importées

Introduction

Le dataset contient toutes les ventes enregistrées par un magasin en ligne basé au Royaume-Uni. Le magasin vend des cadeaux pour toutes les occasions.

8 variables

1. **InvoiceNo:** Numéro de la facture
2. **StockCode:** Numéro d'identification du produit (C indique une annulation)
3. **Description:** Description du produit
4. **Quantity:** Quantités de pièces commandées
5. **InvoiceDate:** Date et heure de la facture
6. **UnitPrice:** Prix à l'unité en livre sterling
7. **CustomerID:** ID du client (NA lors de modifications de stock)
8. **Country:** Pays destinataire

1. Analyse des données importées

Source

UCI Machine Learning Repository

`https://archive.ics.uci.edu/ml/datasets/Online+Retail`

Période

`Du 1/12/10 08:26 au 9/12/11 12:50`

Taille

`541909 lignes dans le dataset`

2. Nettoyage des données

2. Nettoyage des données

a. Données non pertinentes

1. **C:** Annulations
2. **C2:** Transport
3. **D:** Réductions
4. **POST:** Frais postaux
5. **M:** Ajouts manuels
6. **BANK CHARGES:** Frais bancaires
7. **PADS:** Frais d'emballage
8. **DOT:** DOTCOM POSTAGE
9. **Prix à l'unité < 0**
10. **Quantité < 0**
11. **CustomerID = NA**

2. Nettoyage des données

b. Nombre de lignes nettoyées

```
541909 lignes au départ

396337 lignes après nettoyage

391150 lignes après suppression des doublons

Pourcentage nettoyé : 27.81998 %
```

3. Statistiques descriptives

3. Statistiques descriptives

Introduction

i. Nombre de pays: 37 (dont 1 'unspecified')

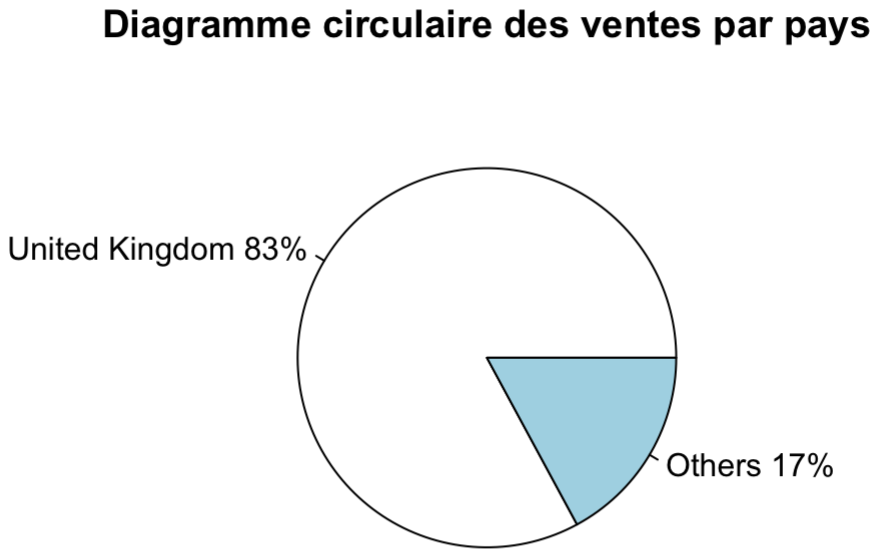
ii. Nombre de factures (uniques): 18402

iii. Nombre de clients (uniques): 4334

iv. Nombre de produits (uniques): 3659

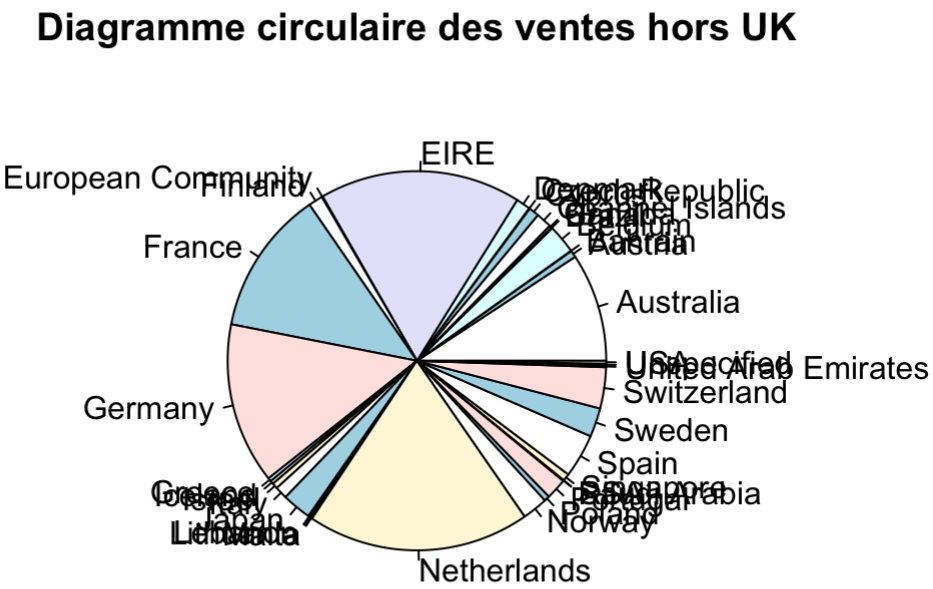
3. Statistiques descriptives

b. Analyse des ventes selon les pays



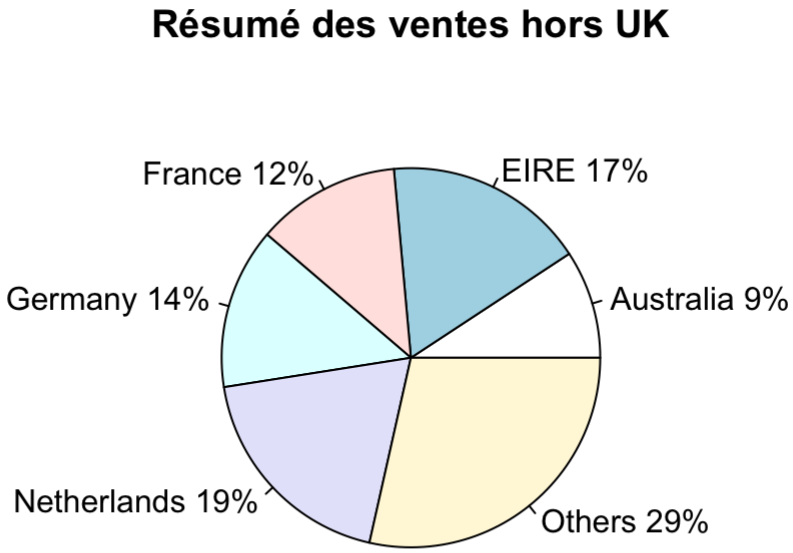
3. Statistiques descriptives

b. Analyse des ventes selon les pays



3. Statistiques descriptives

b. Analyse des ventes selon les pays



3. Statistiques descriptives

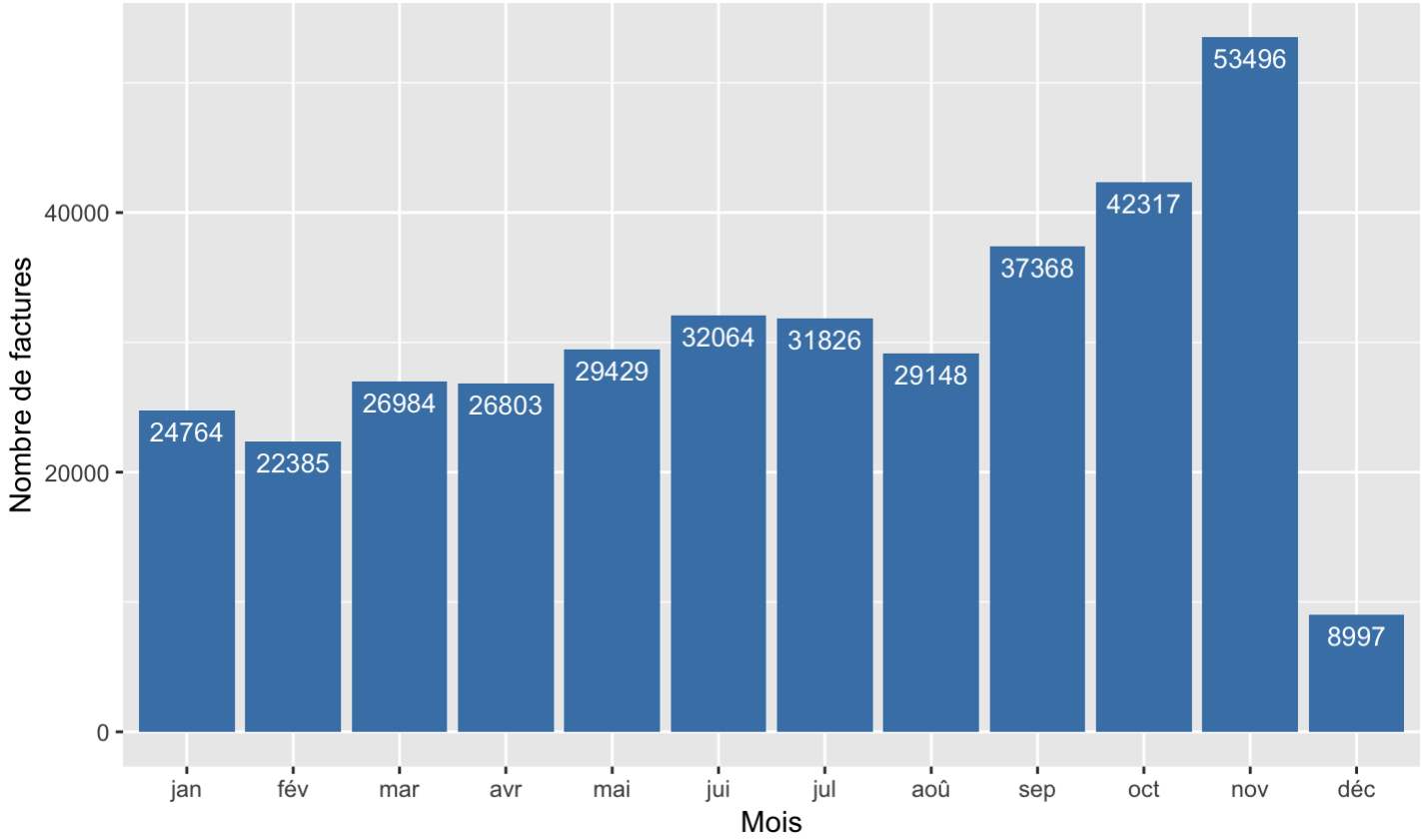
c. Analyse des ventes

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
     0.38    157.14    301.65    474.80    463.07 168469.60
```

3. Statistiques descriptives

c. Analyse des ventes

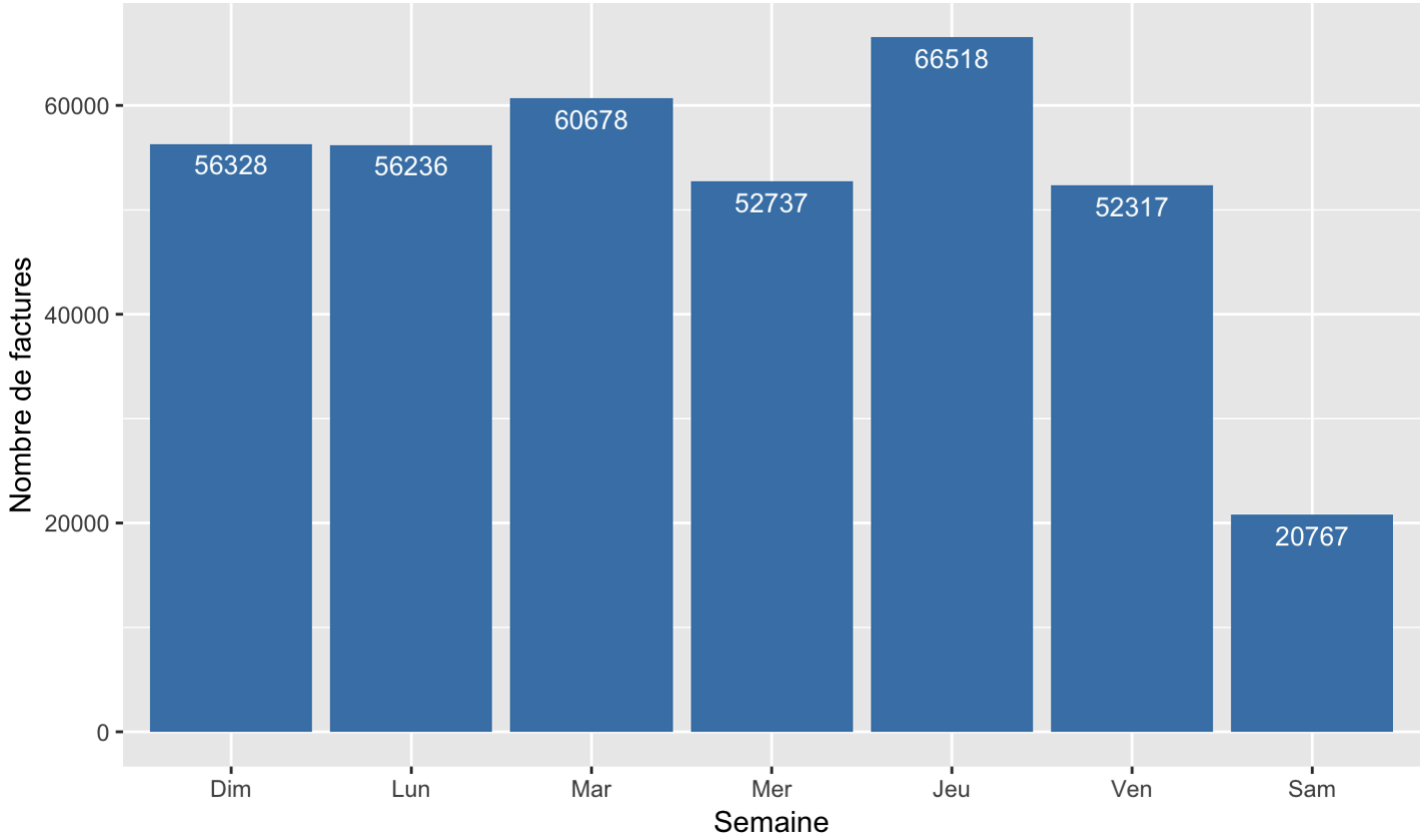
Factures par **mois** en 2011



3. Statistiques descriptives

c. Analyse des ventes

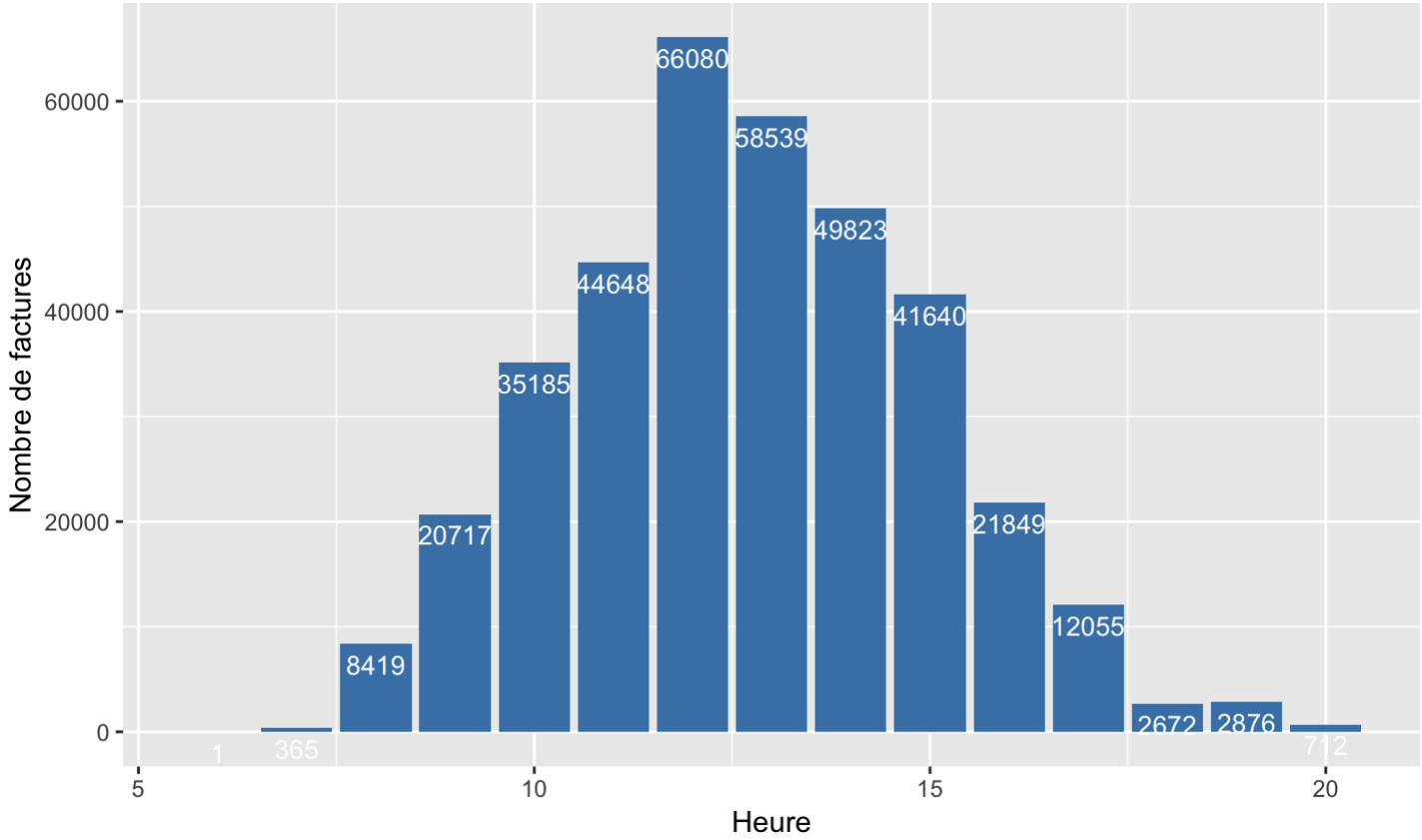
Factures par **jour** en 2011



3. Statistiques descriptives

c. Analyse des ventes

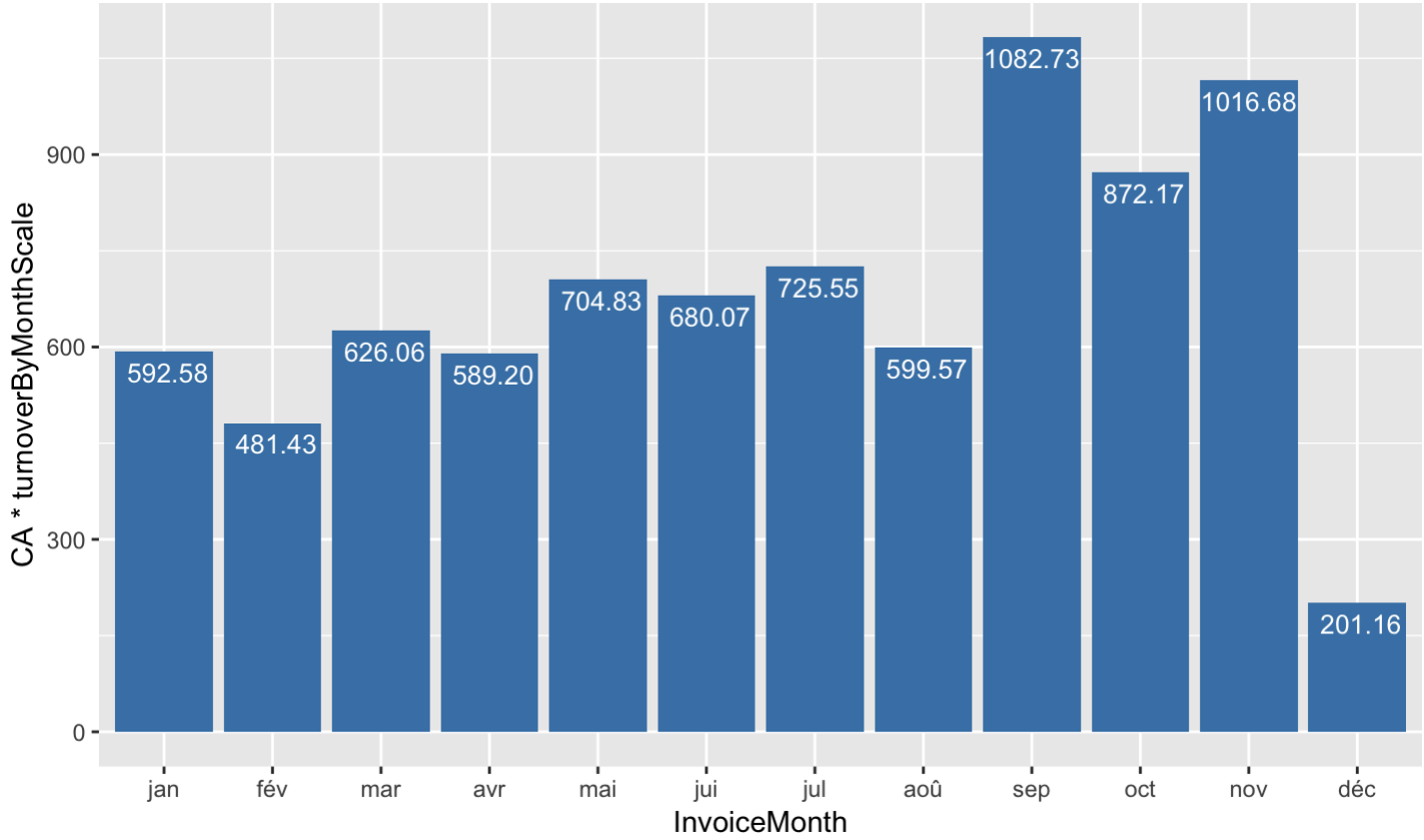
Factures selon **l'heure de la journée** en 2011



3. Statistiques descriptives

c. Analyse des ventes

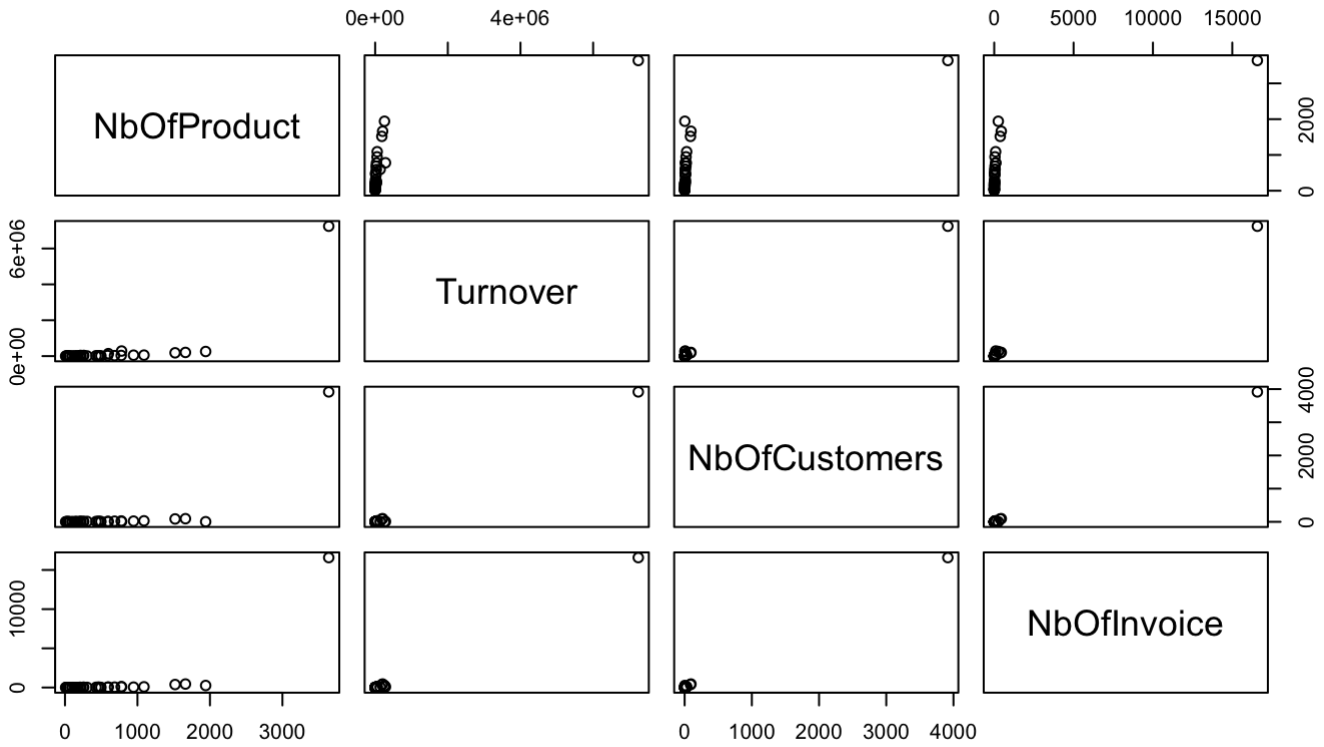
Chiffre d'affaires par **mois**



4. PCA

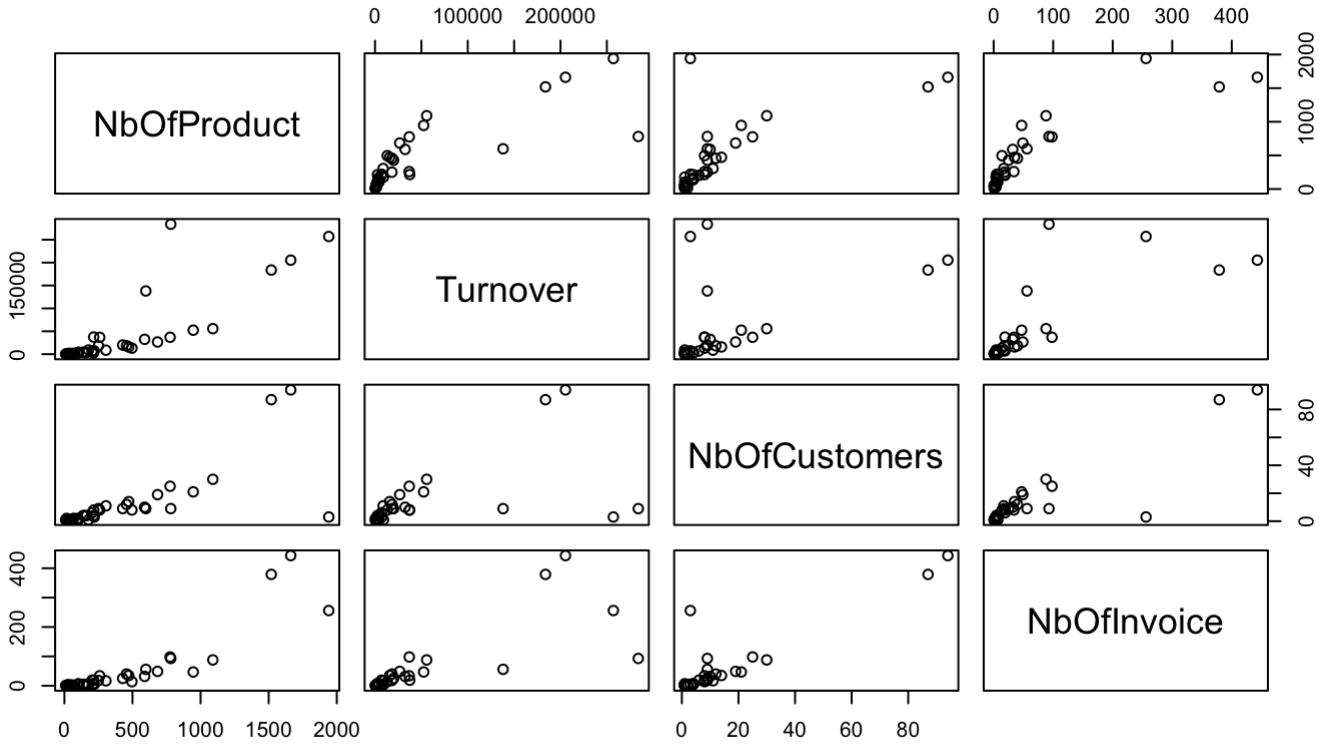
4.1 PCA pays

Agrégation des données de tous les pays



4.1 PCA pays

Agrégation des données des pays hors UK



4.1 PCA pays

Agrégation des données des pays hors UK

Matrice des corrélations

	NbOfProduct	Turnover	NbOfCustomers	NbOfInvoice
NbOfProduct	1.0000000	0.8167573	0.7181459	0.8766853
Turnover	0.8167573	1.0000000	0.5298723	0.7757819
NbOfCustomers	0.7181459	0.5298723	1.0000000	0.8919510
NbOfInvoice	0.8766853	0.7757819	0.8919510	1.0000000

4.1 PCA pays

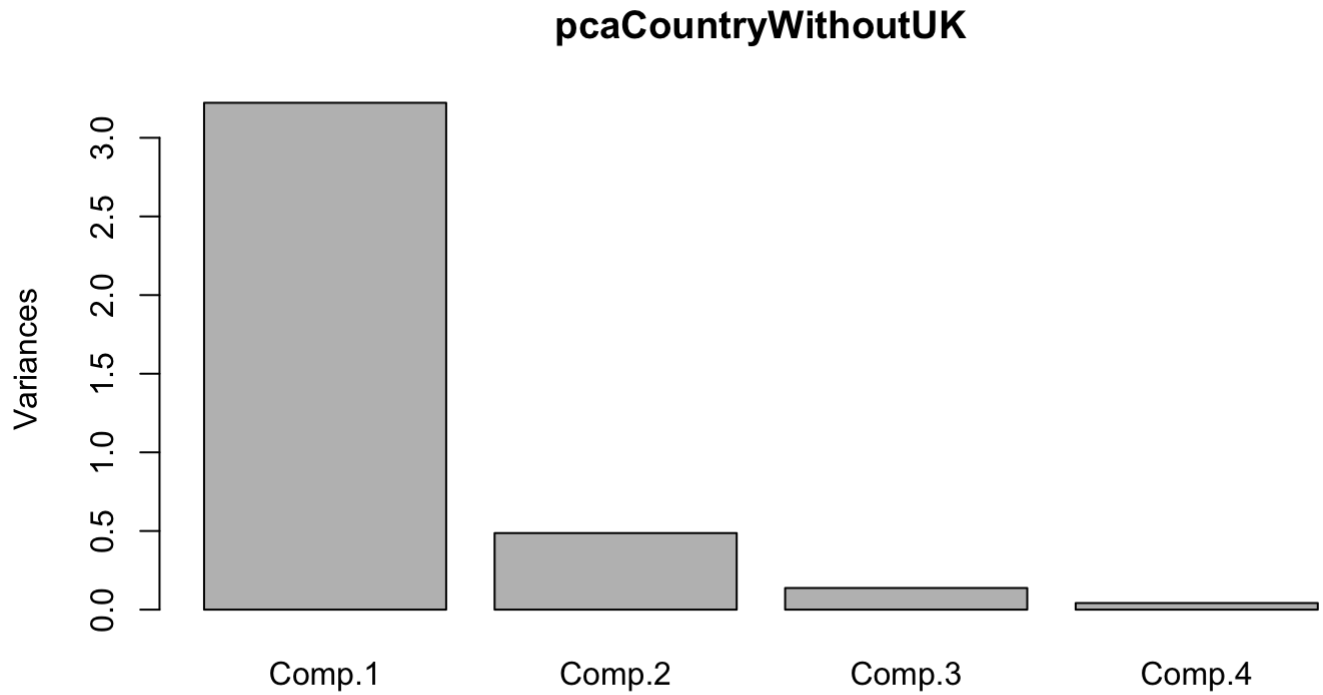
Agrégation des données des pays hors UK

Sommaire des composants

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4
Standard deviation     1.7951395 0.6977192 0.37087114 0.20495248
Proportion of Variance 0.8286495 0.1251803 0.03536882 0.01080142
Cumulative Proportion  0.8286495 0.9538298 0.98919858 1.00000000
```

4.1 PCA pays

Agrégation des données des pays hors UK



4.1 PCA pays

Agrégation des données des pays hors UK

Poids des variables originales dans les composants

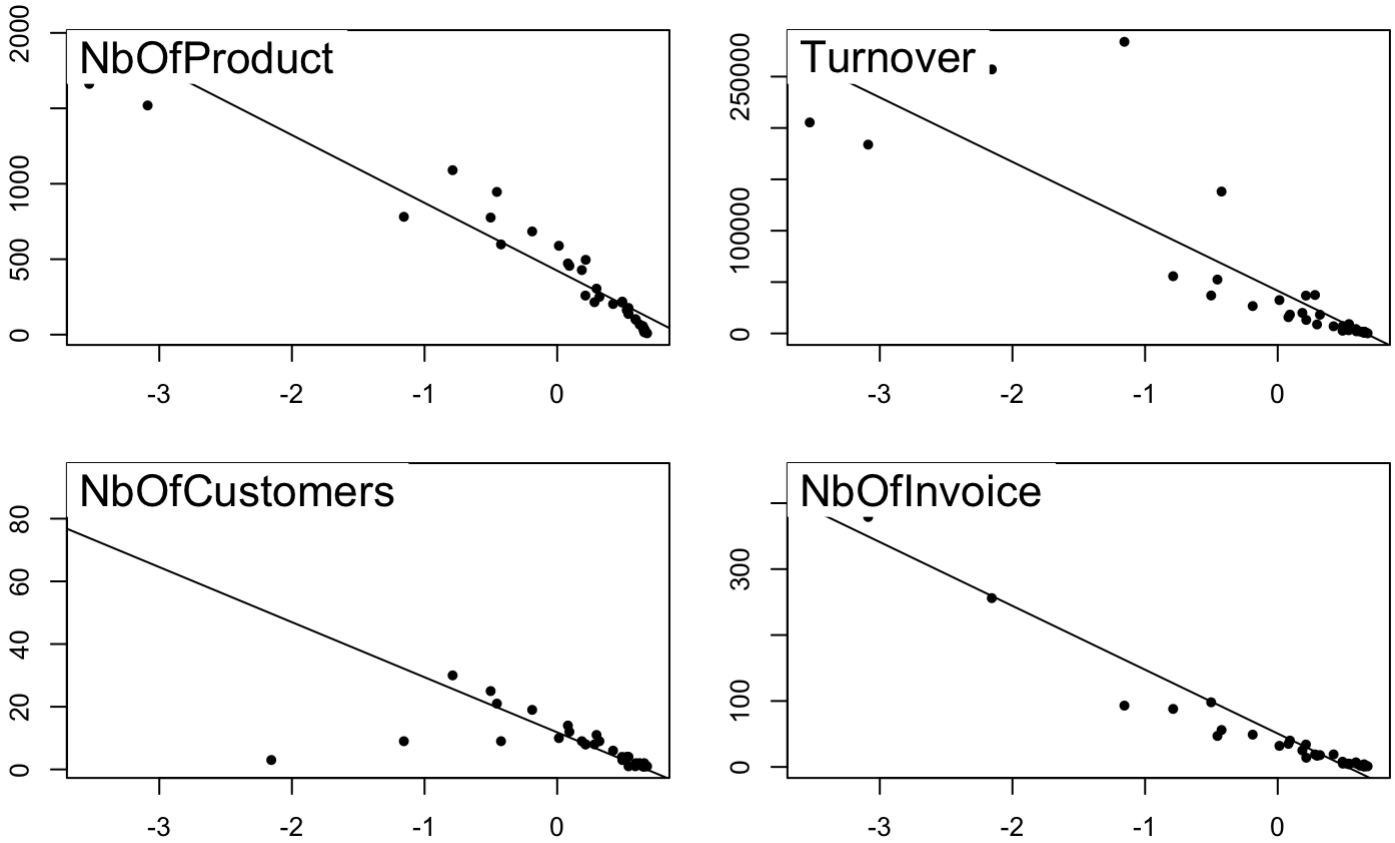
```
Loadings:
              Comp.1 Comp.2 Comp.3 Comp.4
NbOfProduct    0.516  0.207  0.816  0.155
Turnover       0.470  0.672 -0.515  0.247
NbOfCustomers  0.474 -0.687 -0.221  0.504
NbOfInvoice    0.536 -0.182 -0.138 -0.813

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
```

4.1 PCA pays

Agrégation des données des pays hors UK

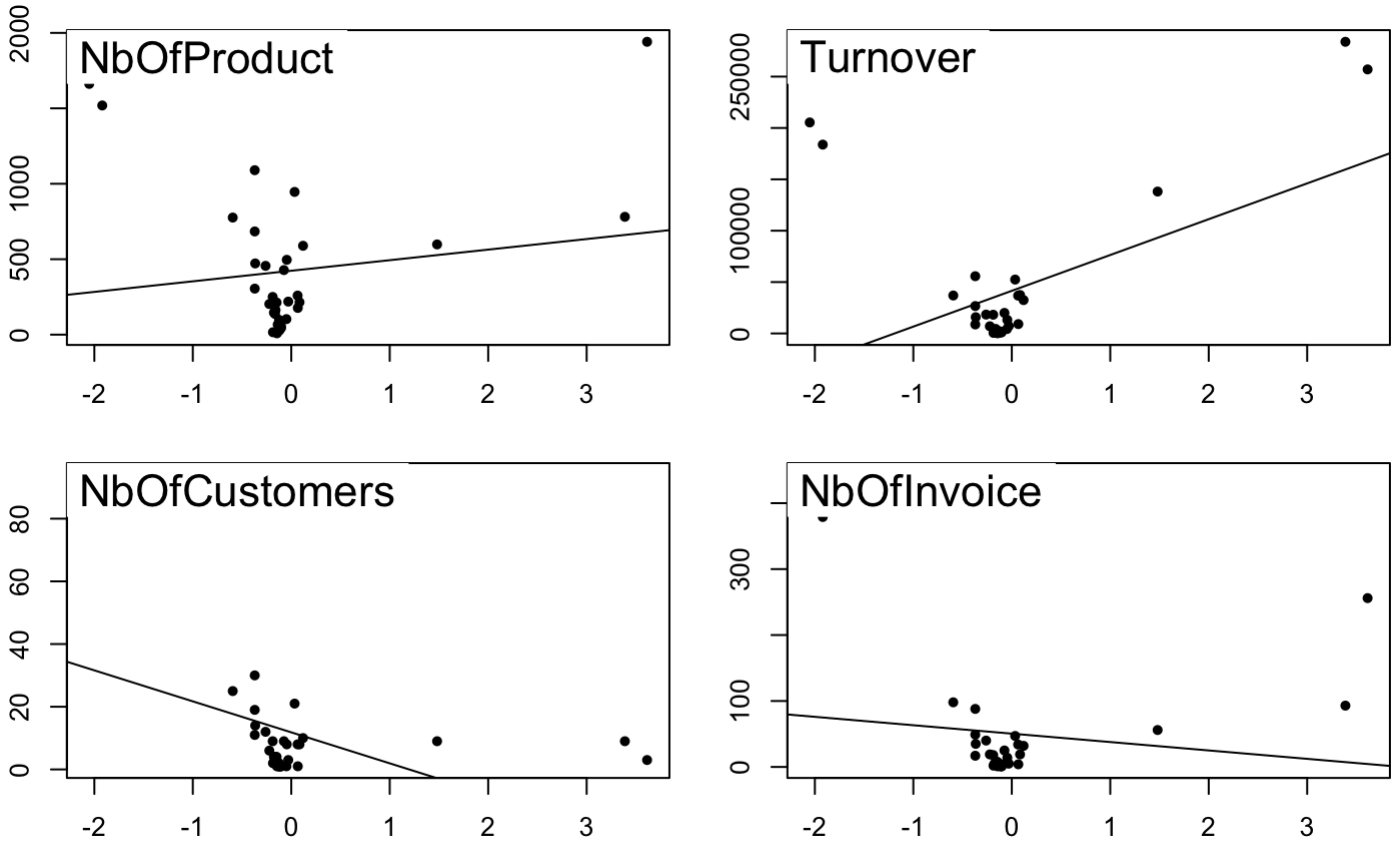
Lien entre les composants et les variables (axis1)



4.1 PCA pays

Agrégation des données des pays hors UK

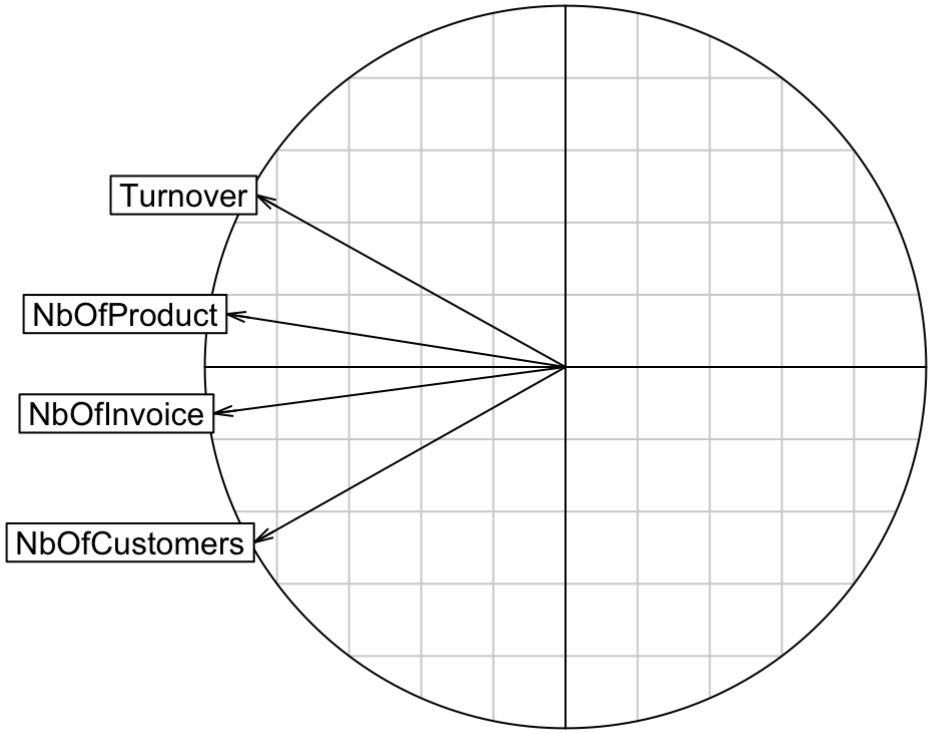
Lien entre les composants et les variables (axis2)



4.1 PCA pays

Agrégation des données des pays hors UK

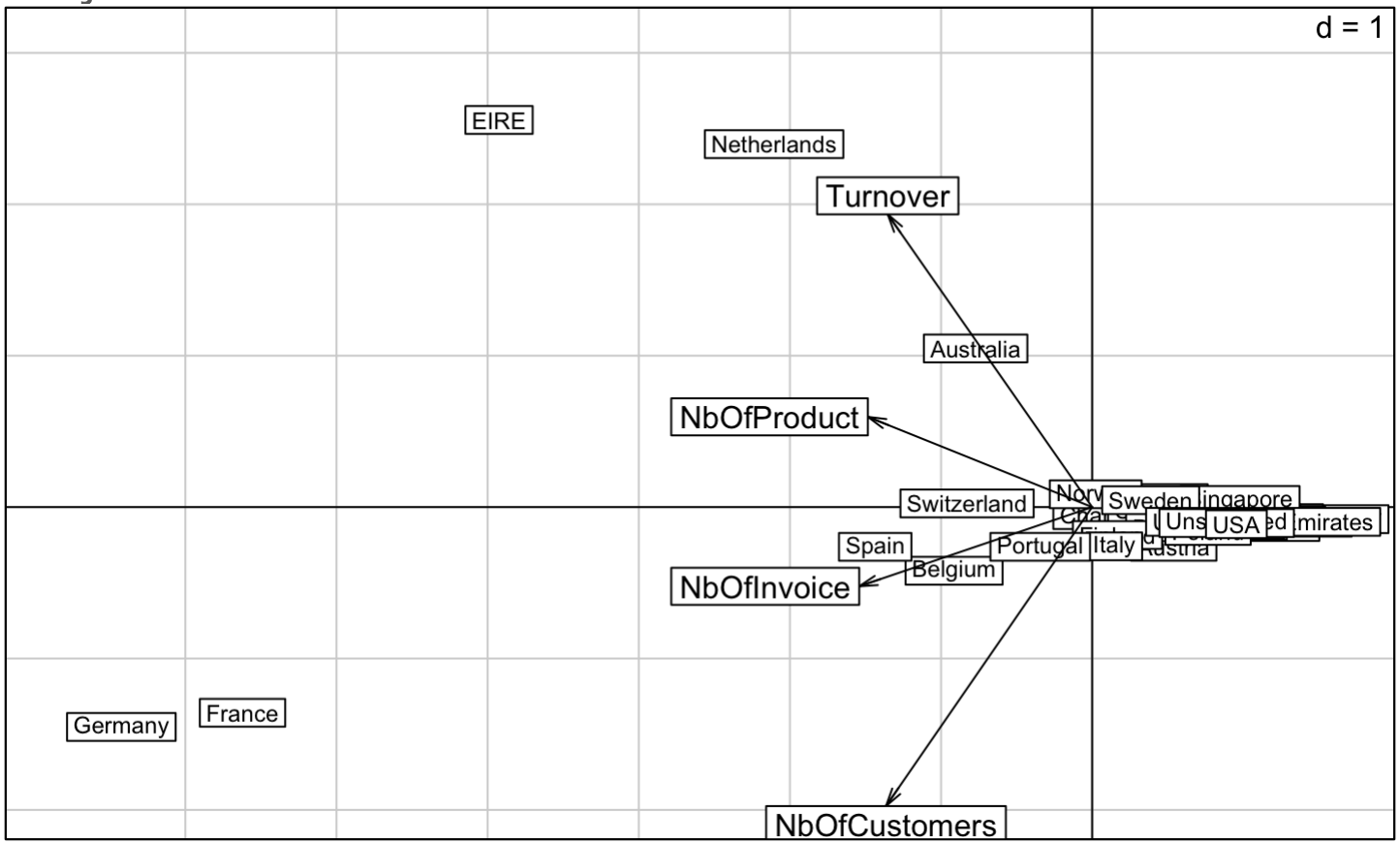
Cercle de corrélation



4.1 PCA pays

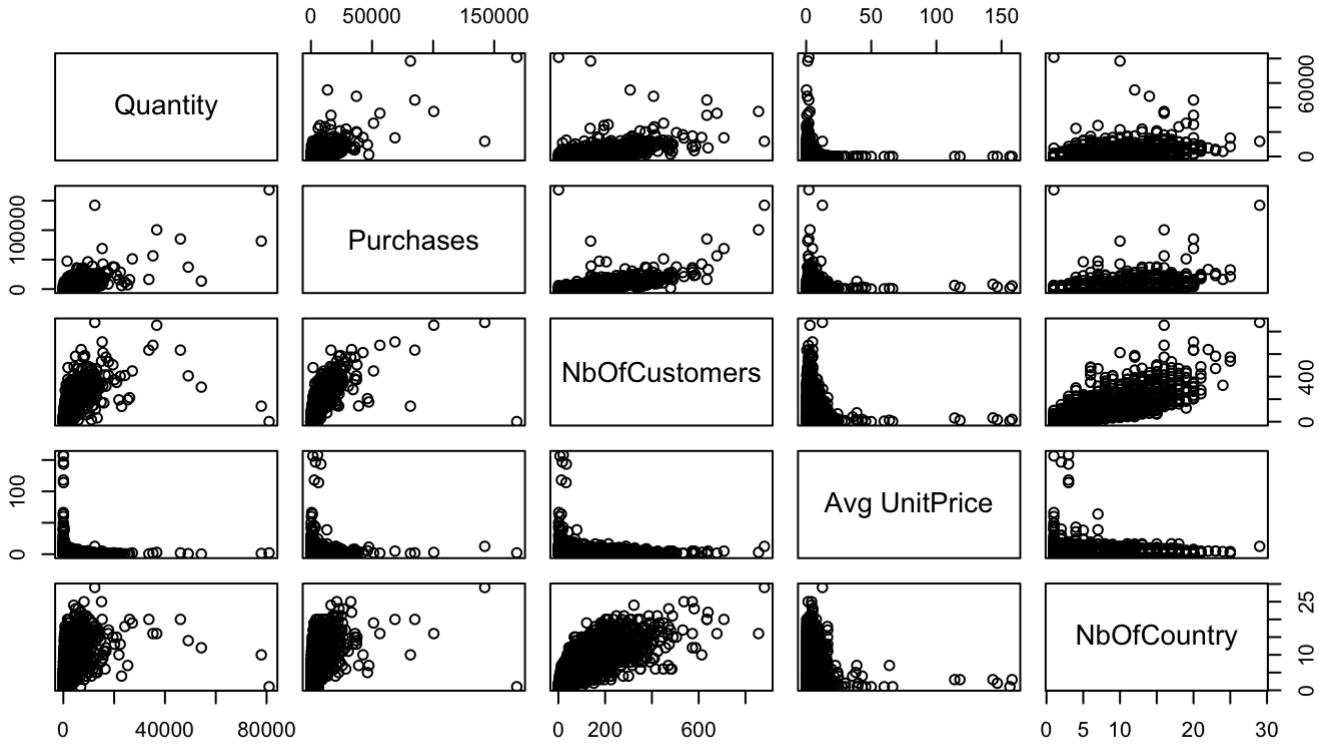
Agrégation des données des pays hors UK

Projection des données



4.2 PCA produits

Agrégation de tous les produits



4.2 PCA produits

Agrégation de tous les produits

Matrice des corrélations

```
                Quantity Purchases NbOfCustomers Avg UnitPrice
Quantity       1.00000000 0.7474641    0.61811828   -0.08826218
Purchases      0.74746408 1.0000000    0.68451970    0.04854800
NbOfCustomers  0.61811828 0.6845197    1.00000000   -0.05353721
Avg UnitPrice -0.08826218 0.0485480   -0.05353721    1.00000000
NbOfCountry    0.45721675 0.4889382    0.78839734   -0.04236737
              NbOfCountry
Quantity       0.45721675
Purchases      0.48893816
NbOfCustomers  0.78839734
Avg UnitPrice -0.04236737
NbOfCountry    1.00000000
```

4.2 PCA produits

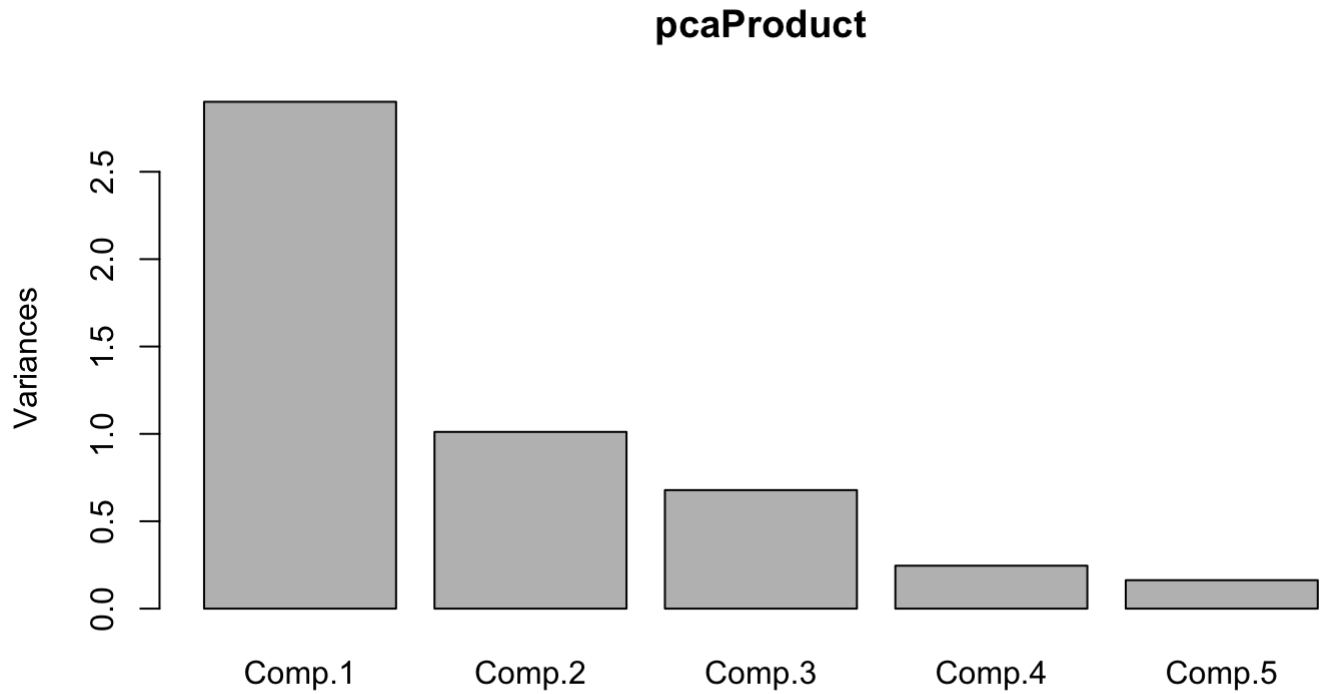
Agrégation de tous les produits

Sommaire des composants

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     1.7030522 1.0056865 0.8236076 0.4957014 0.40347502
Proportion of Variance 0.5802359 0.2023364 0.1357030 0.0491574 0.03256732
Cumulative Proportion  0.5802359 0.7825723 0.9182753 0.9674327 1.00000000
```

4.2 PCA produits

Agrégation de tous les produits



4.2 PCA produits

Agrégation de tous les produits

Poids des variables originales dans les composants

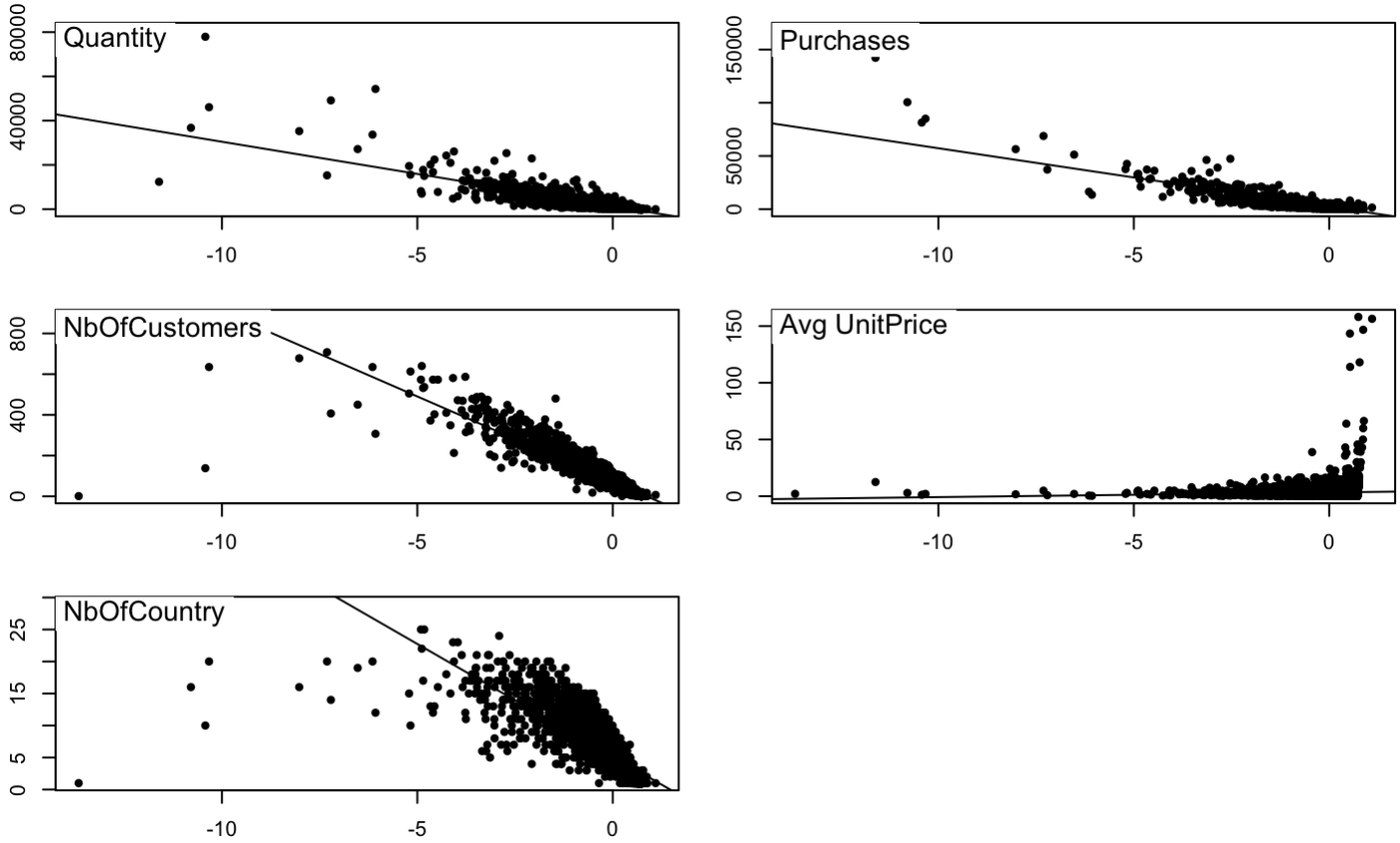
```
Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Quantity      0.487         0.535  0.685
Purchases     0.504 -0.137  0.421 -0.638 -0.377
NbOfCustomers 0.536        -0.303 -0.225  0.755
Avg UnitPrice       -0.989         0.119
NbOfCountry   0.470        -0.665  0.240 -0.528

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
SS loadings       1.0    1.0    1.0    1.0    1.0
Proportion Var    0.2    0.2    0.2    0.2    0.2
Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

4.2 PCA produits

Agrégation de tous les produits

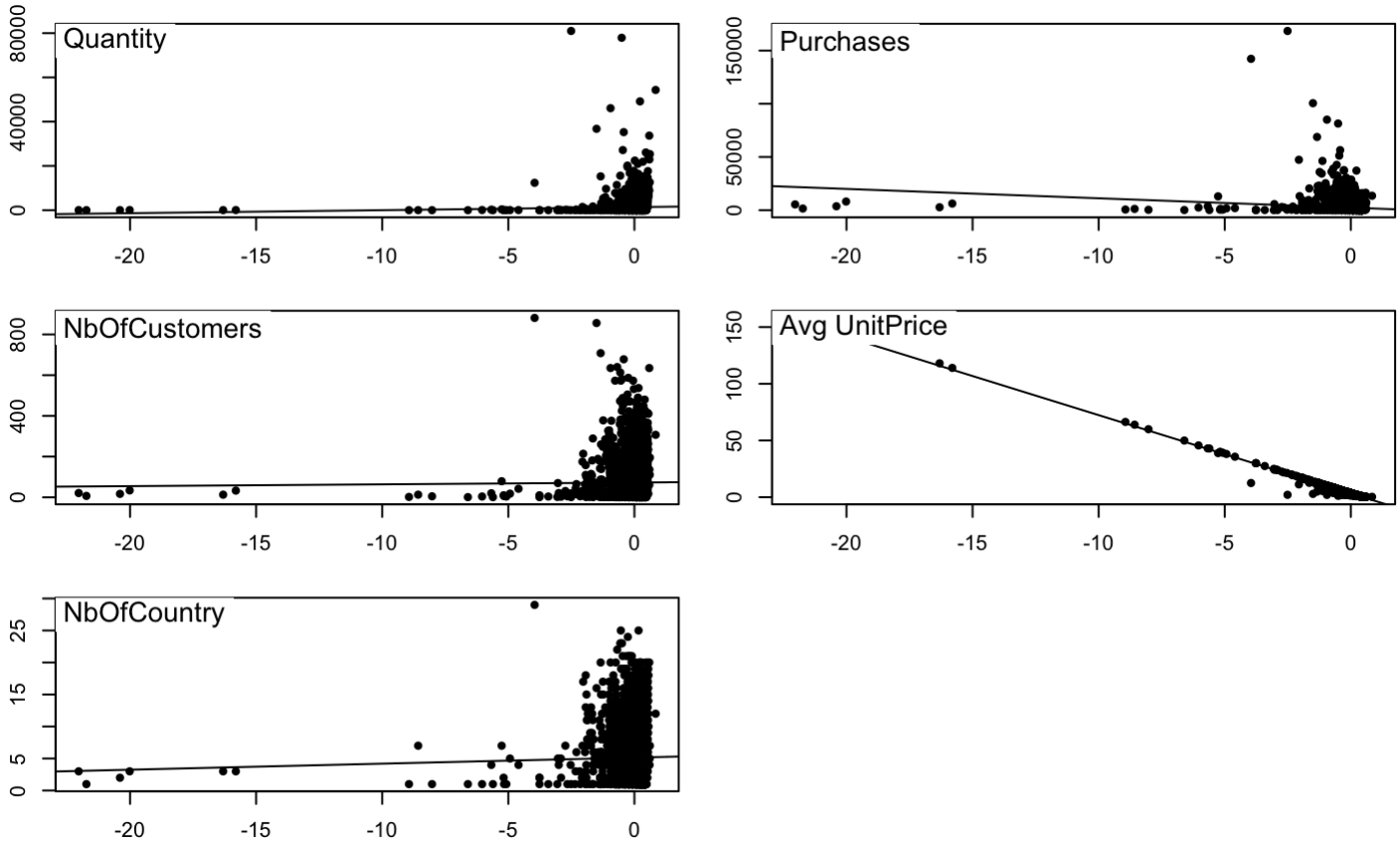
Lien entre les composants et les variables (Axis1)



4.2 PCA produits

Agrégation de tous les produits

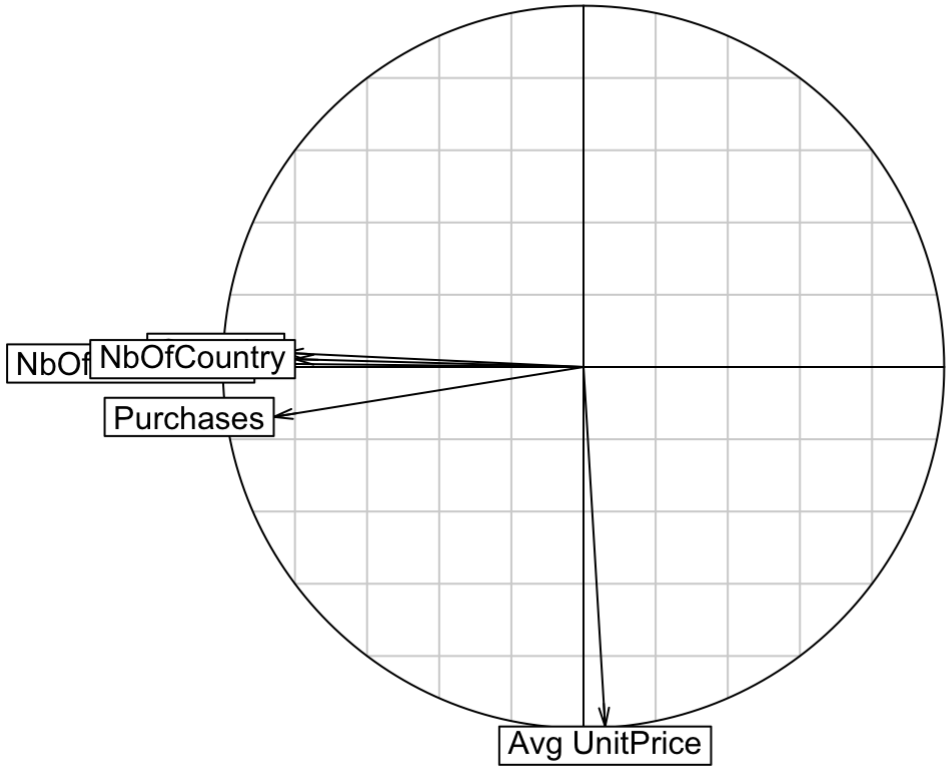
Lien entre les composants et les variables (Axis2)



4.2 PCA produits

Agrégation de tous les produits

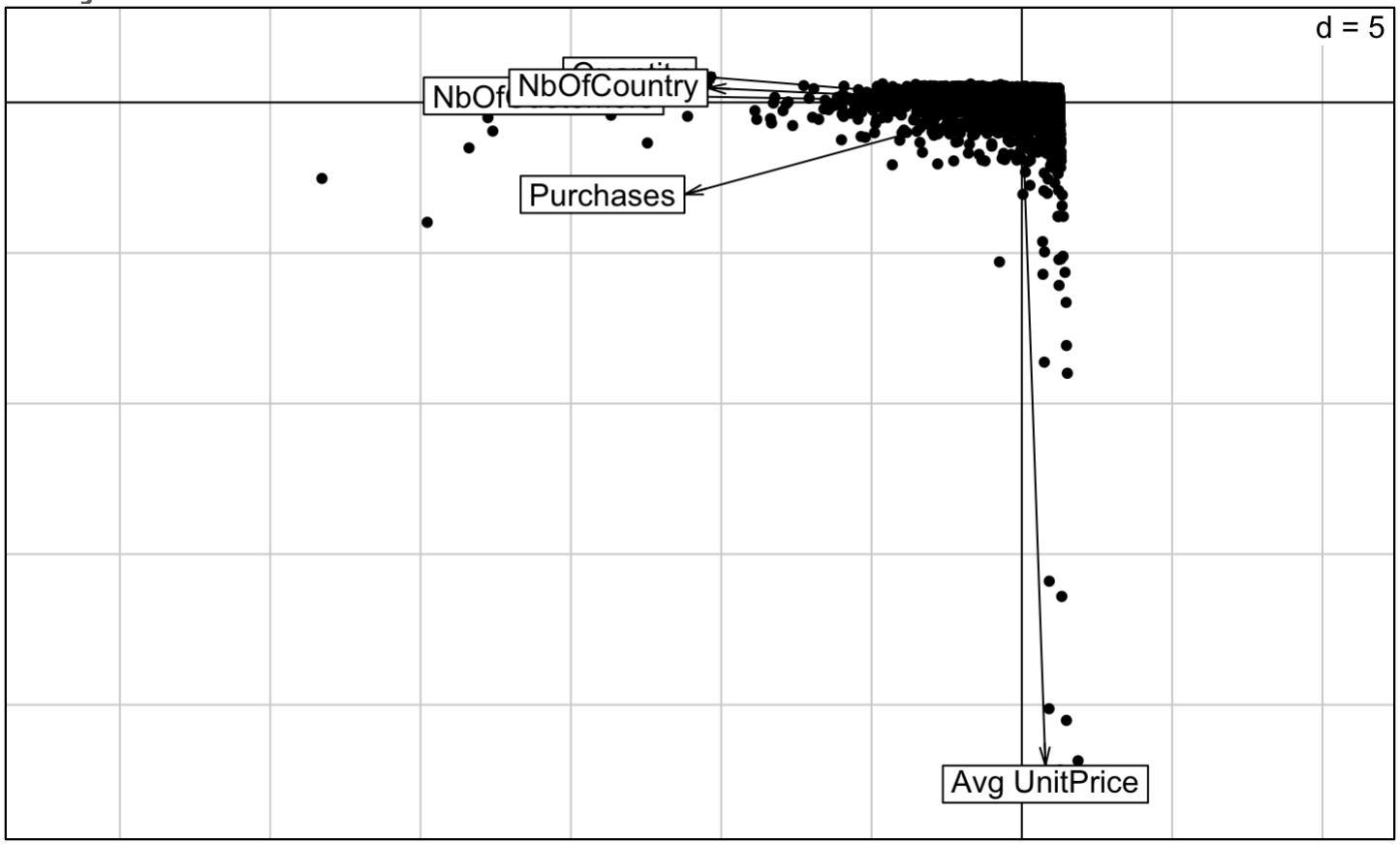
Cercle de corrélation



4.2 PCA produits

Agrégation de tous les produits

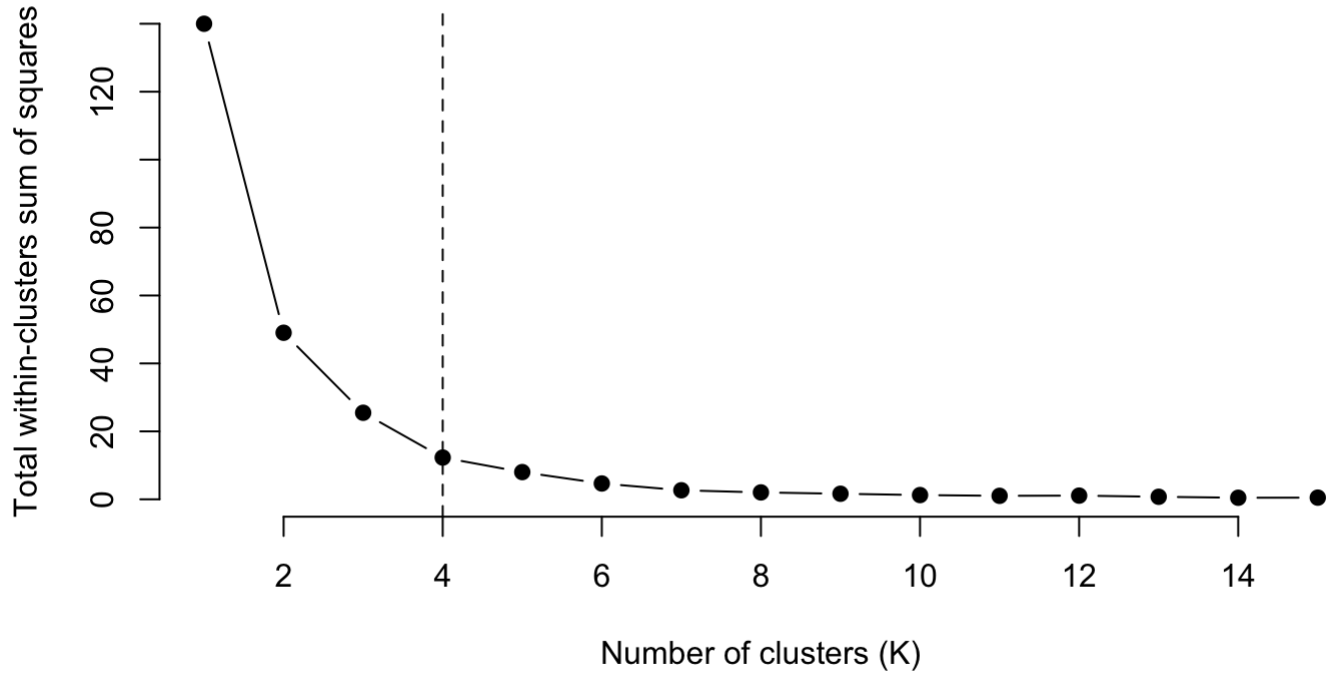
Projection des données



5. Clustering

5.1 Clustering des pays

Calcul du nombre de clusters (Elbow Method)

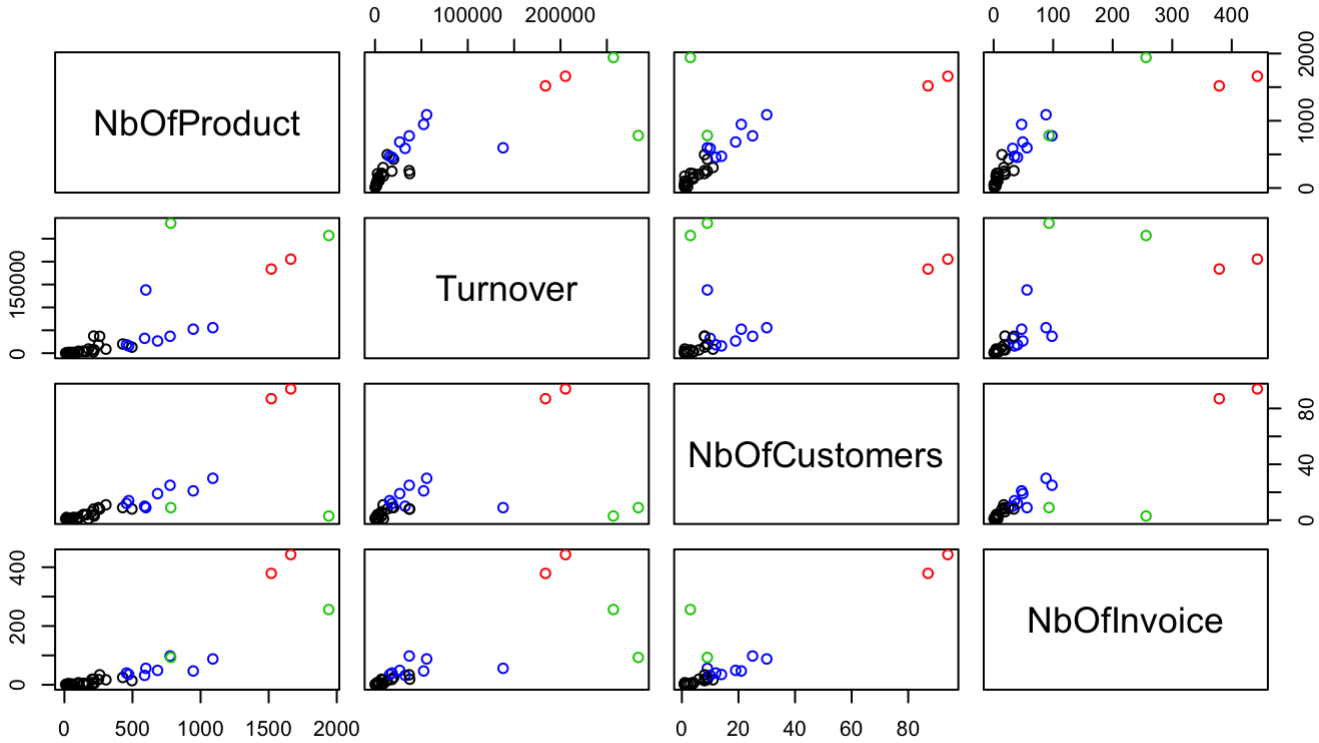


5.1 Clustering des pays

Table des clusters

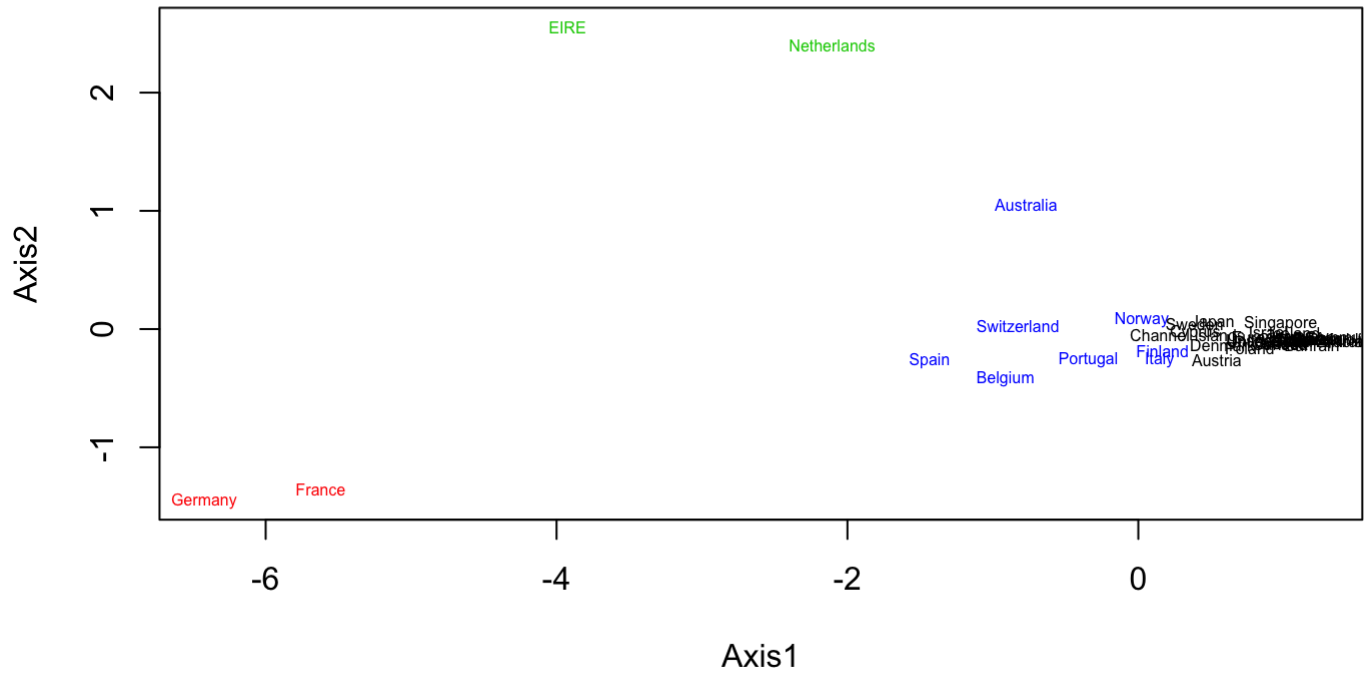
1	2	3	4
24	2	2	8

5.1 Clustering des pays



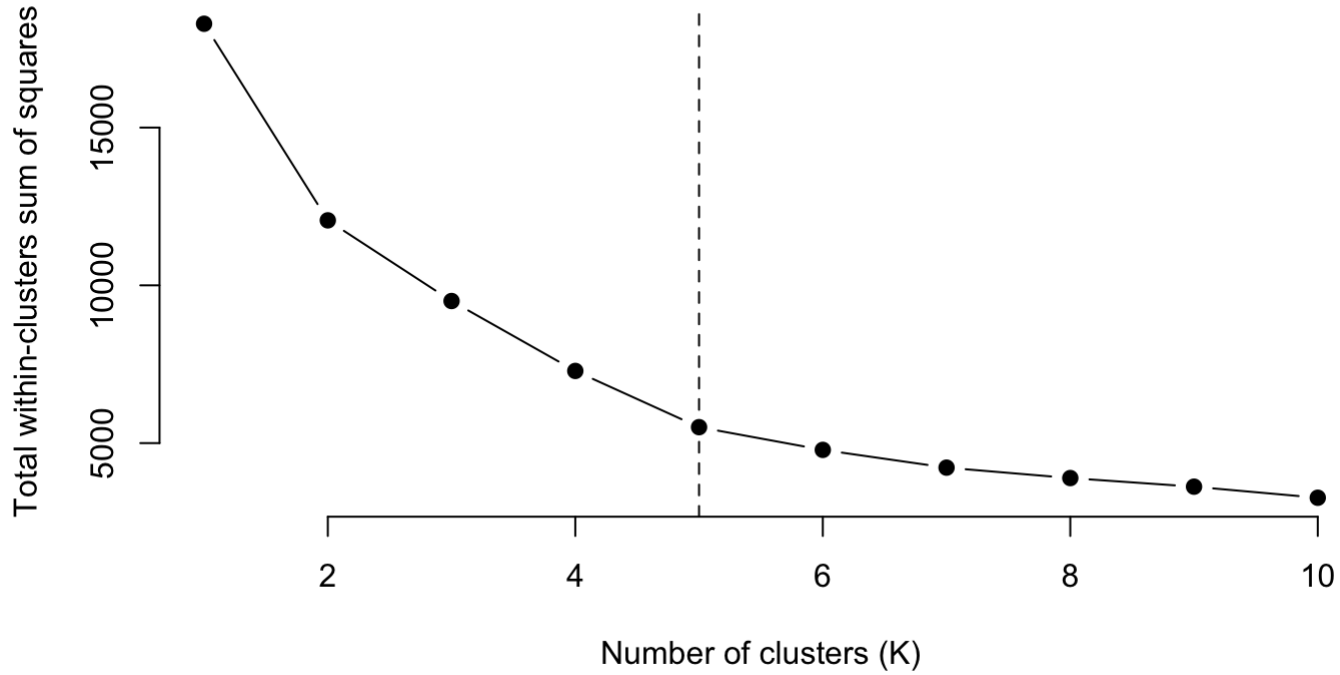
5.1 Clustering des pays

Clusters sur base du PCA



5.2 Clustering des produits

Calcul du nombre de clusters (Elbow Method)

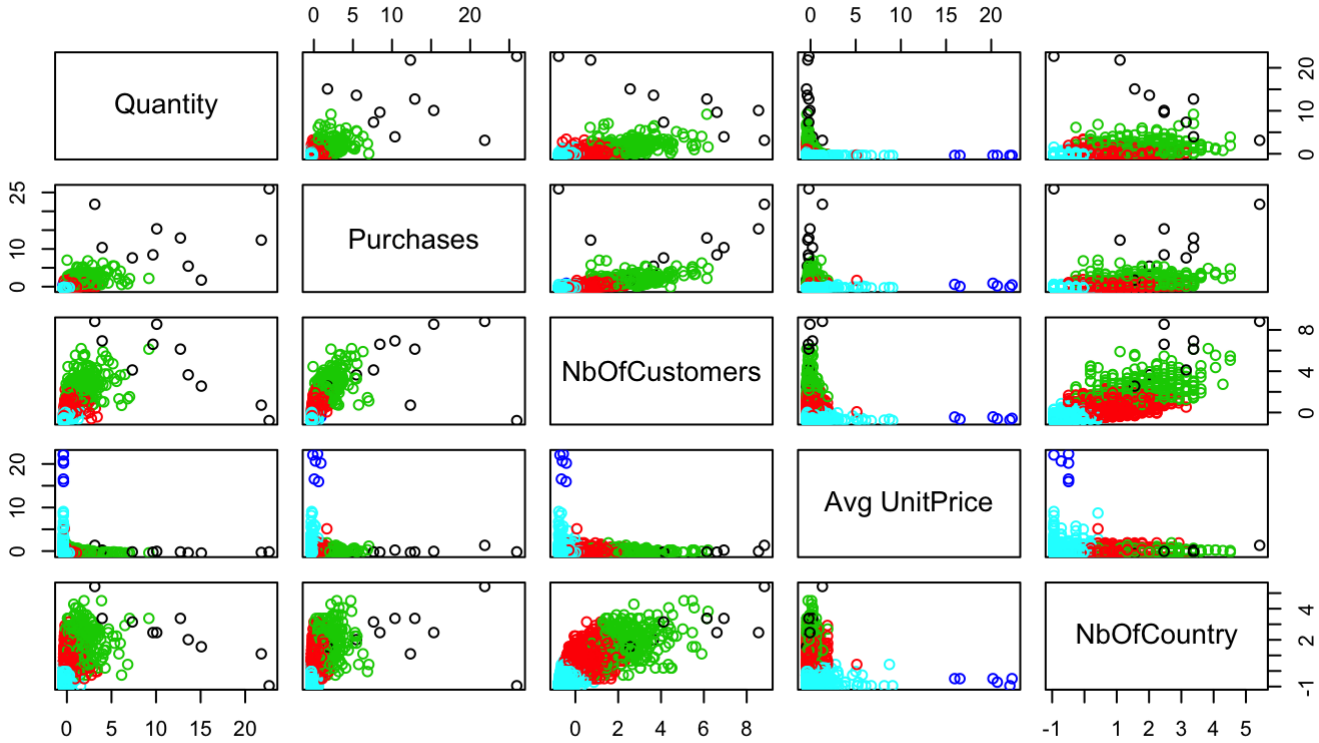


5.2 Clustering des produits

Table des clusters

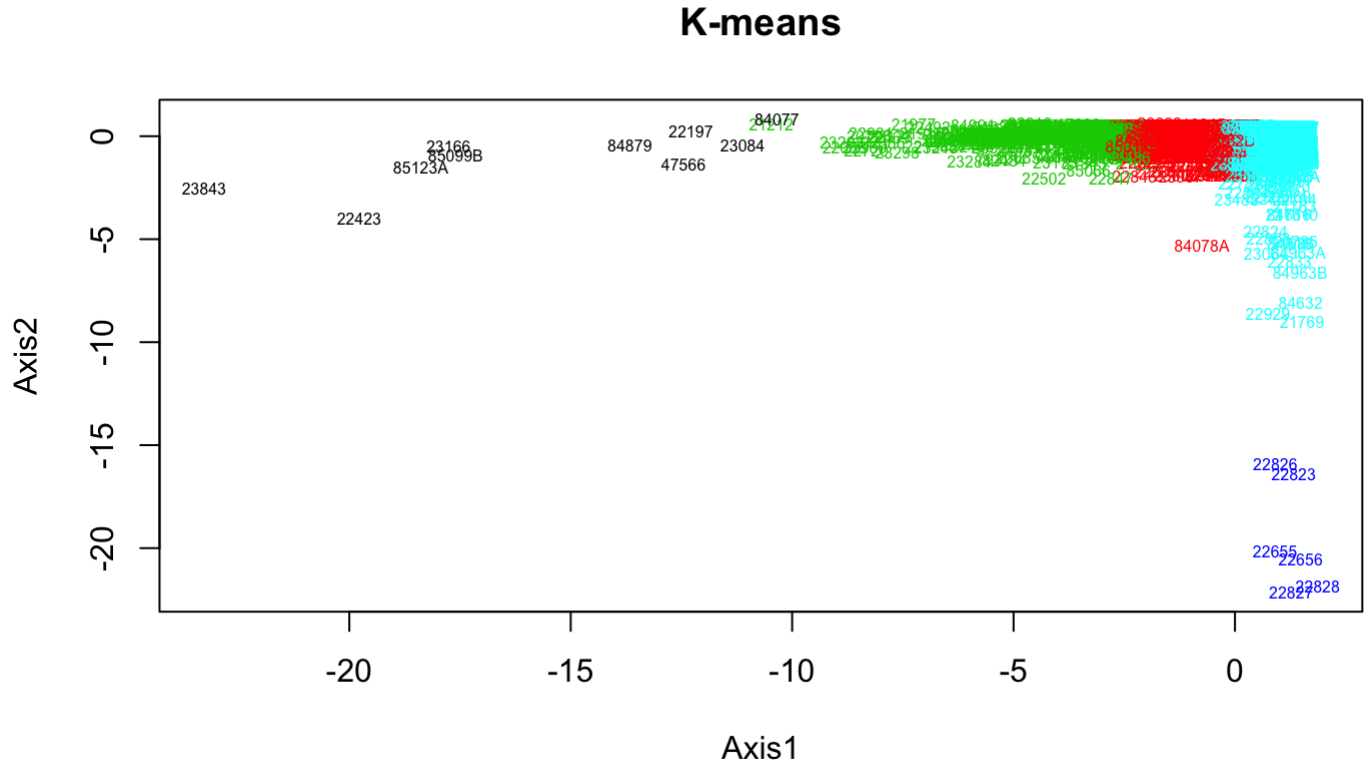
```
   1    2    3    4    5
  10 1034  261    6 2348
```

5.2 Clustering des produits



5.2 Clustering des produits

Clusters sur base du PCA



6. Conclusion

6. Conclusion

Nettoyage important des données (28%).

Identifier les meilleurs moments de promotion.
- -Jeudi
- -De 12 à 14h

Le prix moyen a peu d'influence sur les autres variables (PCA produits).

Beaucoup de produits et consommateurs différents (PCA pays)
- Peu de détaillants

Cluster sur les produits: tri sur les produits.

Clustering des pays permet de mieux catégoriser.

7. Bonus

7. Wordcloud


butterfly
spaceboy
dolly
party
paisley
enamel
crystal
colour
necklace
easter
photo
bottle
card
girl
retrospot
home
hanging
holder
tree
hook
zinc
sign
box
bag
love
ring
tray
pack
metal
star
silver
cake
mirror
tin
stand
bird
polkadot
glass
tea
bracelet
christmas
paper
flower
heart
cream
drawer
earrings
vintage
clock
mug
antique
large
decoration
garden
wooden
cutlery
frame
cup
gift
design
art
bowl
key
candles
wall
mini
jar
cover
small
rose
wrap
fairy
candle
tlight
egg
gold
ceramic
ivory
french
felt
assorted
wood
cushion
garland
diamante