Prévision du commerce de détail russe agrégé à l'aide des principales composantes de Google Trends

Elizaveta Golovanova

Les dépenses consacrées aux produits au détail alimentaires et non alimentaires déterminent en grande partie la demande globale, dont la dynamique est fortement liée à la phase du cycle économique. En particulier, en période de crises profondes (comme l'actuelle pandémie mondiale), il est essentiel de stimuler la demande globale pour en sortir. À cet égard, il est intéressant d'identifier les variables qui pourraient expliquer la dynamique des ventes au détail et aider à les prévoir.
Outre les indicateurs macroéconomiques de base, les requêtes dans le moteur de recherche par groupes de produits peuvent être un outil important qui permet de suivre l'augmentation et la diminution de l'intérêt des utilisateurs. Il parait intéressant de savoir si les données de requête de recherche en Google aident à prédire les ventes au détail. Pour prendre en compte le maximum d'informations sur les requêtes dans la régression, on peut utiliser l'outil Google Trends. Mais la question se pose de savoir à quel point les demandes pour chaque groupe de produits sont similaires et est-ce qu'on peut réduire la dimension de nos trends en question.

Problématique

Les données de recherche aident-elles à prédire les ventes au détail ?

Pour résoudre ce problème, je vais considerer des modeles avec un ensemble des variables macroéconomiques clés qui expliquent la vente au détail de produits alimentaires et non alimentaires ainsi que les tendances de Google, pour améliorer les propriétés prédictives de ces modèles.

Données et leur traitement

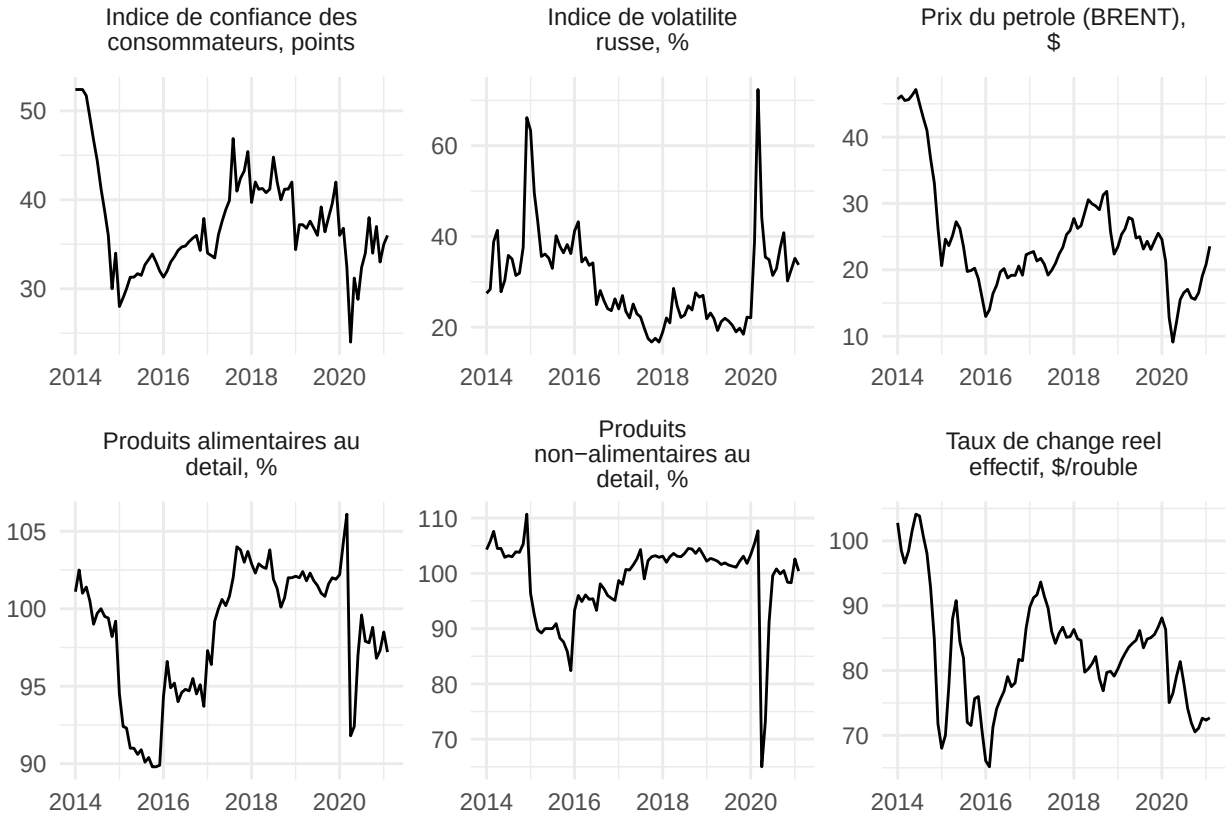
La période de janvier 2015 à février 2021 inclus est considérée. Le tableau répertorie les variables macroéconomiques qui exercent une influence sur les ventes au détail.

Variable	Source
Chiffre d'affaires des ventes au détail cumulées	https://rosstat.gov.ru/folder/23457
Indice russe de volatilité	https://ru.investing.com/indices/russian-vix
Indicateur de confiance des consommateurs	https://wciom.ru/ratings/indeks-potrebitelskogo-doverija
Pétrole brut Brent	https://fred.stlouisfed.org/series/POILBREUSDM
Taux de change réel effectif	https://fred.stlouisfed.org/series/RBRUBIS
Indice américain des prix à la consommation (seulement pour correction)	https://fred.stlouisfed.org/series/CPIAUCSL

Ci-dessous se trouvent des graphiques des données des variables macroeconomiques et leurs statistiques descriptives. Il est à noter qu'en avril 2020, en raison de la pandémie et du confinement, des valeurs minimales sont observées dans les ventes au détail non alimentaire, les prix du pétrole et l'indice de confiance des consommateurs.

```
macro_graph <- macro_graph[, 1:7]

df1<-macro_graph %>% pivot_longer(-date, names_to = 'names', values_to = 'values')
ggplot(df1, aes(date, values)) + geom_line() +
                facet_wrap(~names,
                           nrow = 2,
                           labeller = label_wrap_gen(width=25),
                           scales = 'free') + theme_minimal() +
  labs(x = NULL, y = NULL)
```




```
summary(macro_graph[2:length(macro_graph)])
```

```
##  Indice de confiance des consommateurs, points Indice de volatilite russe, %
##  Min.   :24.00                                 Min.   :16.72
##  1st Qu.:33.49                                 1st Qu.:22.24
##  Median :36.23                                 Median :27.95
##  Mean   :37.32                                 Mean   :30.35
##  3rd Qu.:41.13                                 3rd Qu.:35.56
##  Max.   :52.40                                 Max.   :72.39
##  Prix du petrole  (BRENT), $ Taux de change reel effectif, $/rouble
##  Min.   : 9.109              Min.   : 65.15
##  1st Qu.:19.692              1st Qu.: 76.56
##  Median :23.415              Median : 82.03
##  Mean   :24.811              Mean   : 82.64
##  3rd Qu.:27.088              3rd Qu.: 86.47
##  Max.   :47.156              Max.   :104.11
```

```
##  Produits non-alimentaires au detail, % Produits alimentaires au detail, %
##  Min.   : 65.00                         Min.   : 89.80
##  1st Qu.: 96.00                         1st Qu.: 94.95
##  Median :101.55                         Median : 99.65
##  Mean   : 98.93                         Mean   : 98.46
##  3rd Qu.:103.17                         3rd Qu.:101.97
##  Max.   :110.70                         Max.   :106.10
```

Dans le tableau ci-dessous, la colonne gauche présente les catégories de produits alimentaires et non alimentaires. À droite, ils sont mappés aux catégories de Google Trends. J'ai décidé d'utiliser des catégories, car cela aidera à identifier l'intérêt général pour les produits en fonction de diverses requêtes. Dans le cas d'utilisation des mots-clés, il est évident que certains requêtes ne seront pas prises en compte. Plusieurs catégories de produits alimentaires ne correspondaient pas aux tendances, quatre catégories de produits non alimentaires ne contenaient pas de données, mais la plupart des atégories de vente au détail étaient comparable avec Google Trends. Ainsi, pour les ventes au détail alimentaires 7 tendances ont été sélectionné, et pour les non-alimentaires – 25.

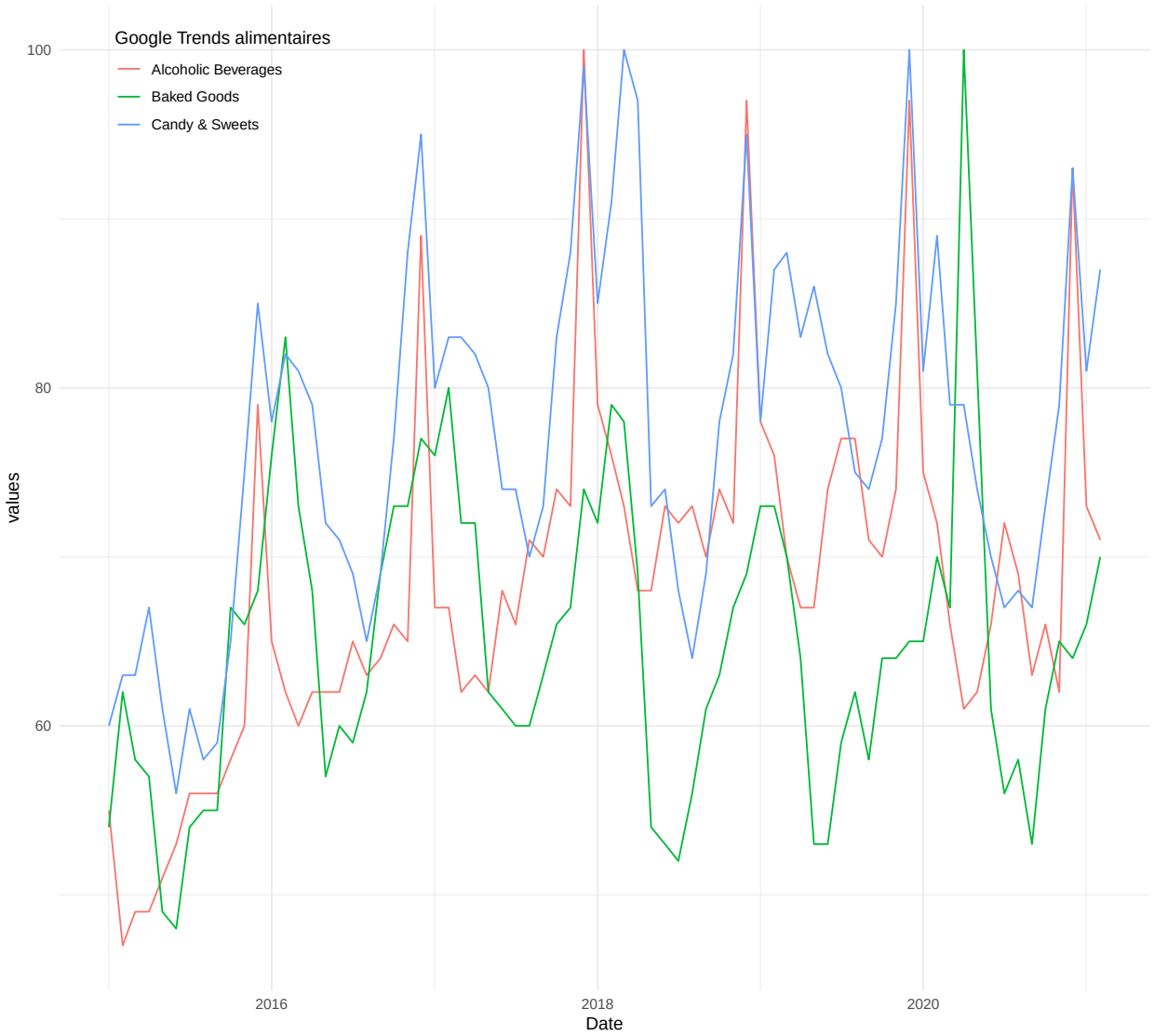
Ventes au détail alimentaires	Google Trends
Boissons alcoolisées	Alcoholic Beverages
Confiserie, sucre	Candy & Sweets
Pommes de terre fraîches, fruits et légumes frais	Fruits & Vegetables
Viande et poisson	Meat & Seafood
Pain et produits de boulangerie	Baked goods
Thé	Coffee & Tea
Produits du tabac	Tobacco products
Huiles et graisses animales, huiles végétales, produits à base de margarine, produits laitiers, œufs, farine, céréales, pâtes	Aucune catégorie appropriée

Ventes au détail non-alimentaires	Google Trends
Tissus	Textiles & Nonwovens
Vêtements pour hommes, femmes et enfants, produits en fourrure	Children's clothing, men's clothing, women's clothing
haussettes et collant	Undergarments
Chaussure	Footwear
Produits de nettoyage et de polissage synthétiques, savon de toilette et savon à lessive	Cleaning Supplies & Services
Produits Cosmétiques et parfums autres que le savon	Make-Up & Cosmetics, Perfumes & Fragrances
Montre	Watches (pas de données)
Ordinateurs	Computer Hardware
Téléphone portable	Mobile phones
Matériel audio	Audio equipment
Télévisions	Televisions (pas de données)
Magnétoscopes	Video Players & Recorders (pas de données)
Revêtements de sol, tapis et moquettes	Rugs & Carpets
Meuble	Sofas & Chairs, Home Storage & Shelving
Matériaux de construction	Construction & Power Tools
Bijouterie	Gems & Jewellery (pas de données)
Produits médicaux, produits orthopédiques	Disabled & Special Needs
Médicaments	Drugs & Medications

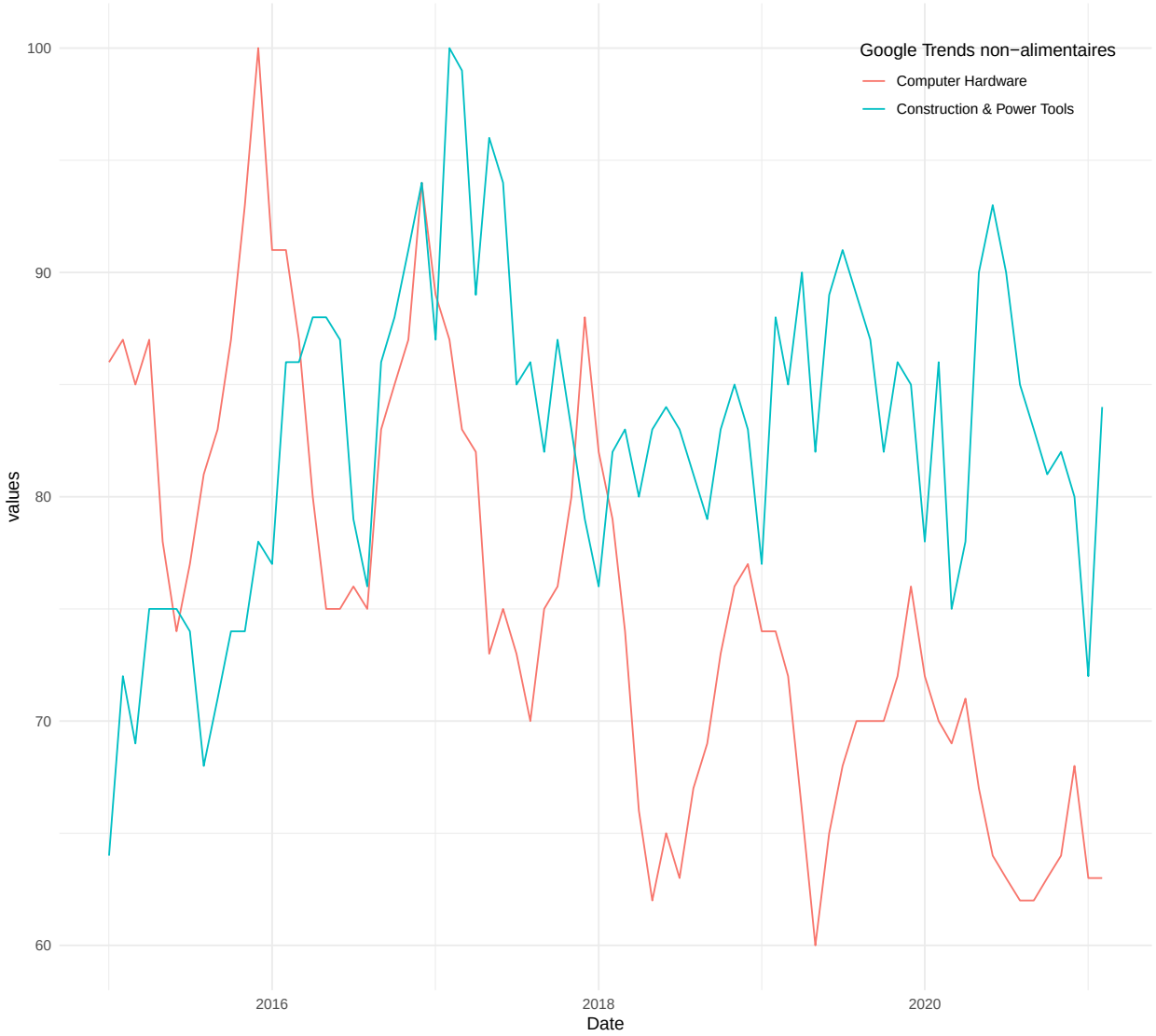
Ventes au détail non-alimentaires	Google Trends
Livres	Book Retailers
Journaux et magazines	Magazines
Vélos	Bicycles & Accessories
Motocycle	Motorcycles
Voitures particulières	Classic Vehicles
Essence automobile, carburant diesel, carburant à moteur à gaz	Vehicle Fuels & Lubricants
Réfrigérateurs et congélateurs, machines à laver	Major Kitchen Appliances

Google Trends fournit des fréquences relatives pour la catégorie sélectionnée et, en fonction de la période sélectionnée, met à l'échelle les données de 0 à 100. La valeur maximale correspond au plus grand nombre de requêtes, tandis que toutes les autres valeurs sont mises à l'échelle vers ce maximum. J'ai téléchargé tous les tendances sélectionnés depuis 2004 pour obtenir la véritable dynamique de la série. Je veux préciser que les données sont prises pour la Russie et pour les requêtes de recherche sur Internet (Web Search). Les graphiques suivantes présentent des exemples des catégories alimentaires et non-alimentaires de Google Trends. On voit que la saisonnalité est observée qu'il faut supprimer avant estimation des données.

```
prod[1:4] %>%
  pivot_longer(-Date, names_to = 'names', values_to = 'values') %>%
  filter(Date > '2015-01-01') %>%
  ggplot(aes(Date , values, color = names)) +
  geom_line() +
  theme_minimal() +
  labs(color='Google Trends alimentaires') + theme(legend.position = c(0.15,0.92))
```




```
neprod %>%
  select(c('Date', 'Computer Hardware', 'Construction & Power Tools')) %>%
  pivot_longer(-Date, names_to = 'names', values_to = 'values') %>%
  filter(Date > '2015-01-01') %>%
  ggplot(aes(Date , values, color = names)) +
  geom_line() +
  theme_minimal() +
  labs(color='Google Trends non-alimentaires') + theme(legend.position = c(0.85,0.92))
```

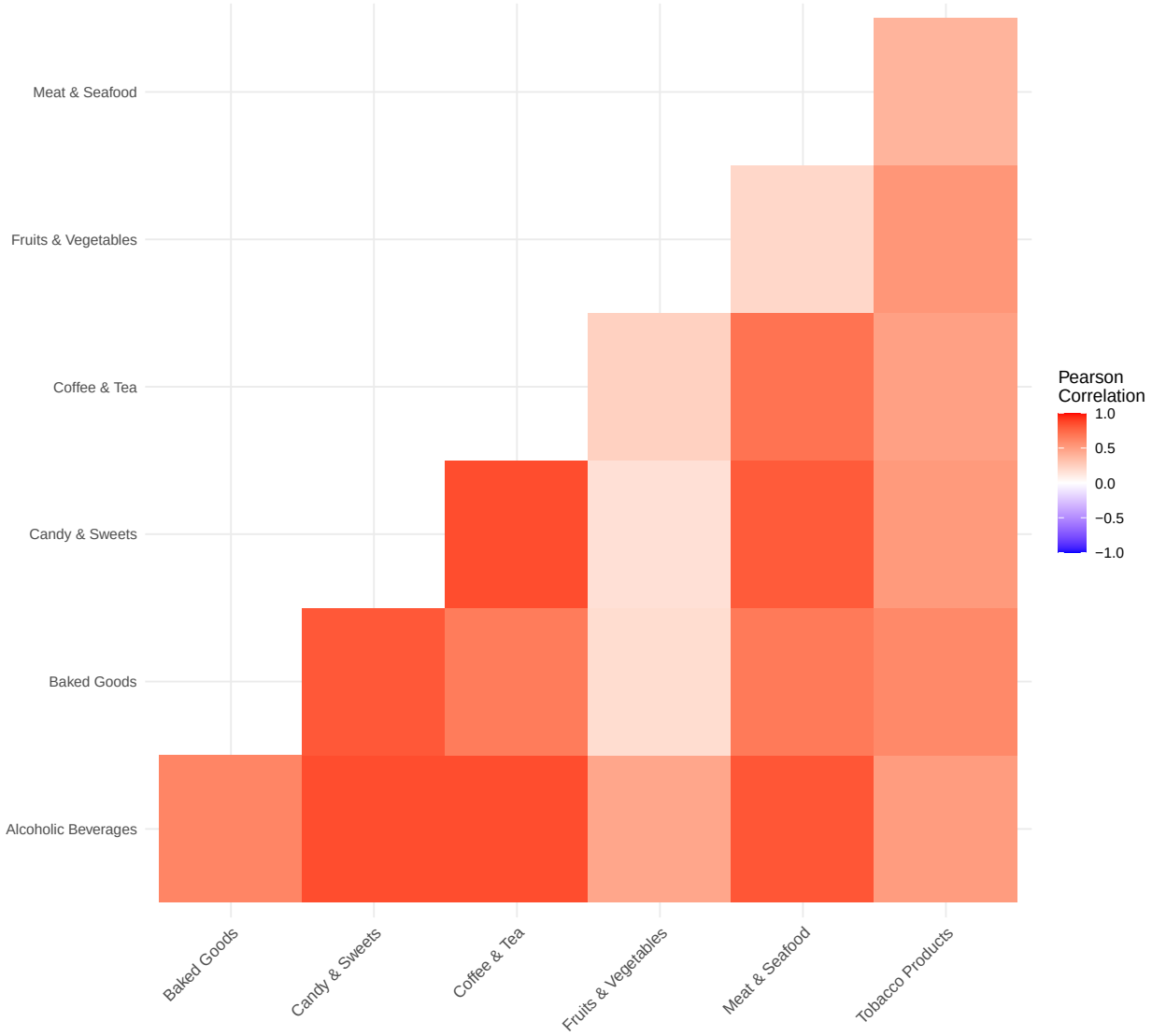


Regardons maintenant la corrélation de nos Google Trends. Les matrices de corrélation sont présentées ci-dessous et le degré de corrélation des trends est mis en évidence en couleur. On peut voir que toutes les tendances alimentaires ont une corrélation positive modérée, lorsque des corrélations positives et négatives sont observées parmi les tendances non alimentaires.

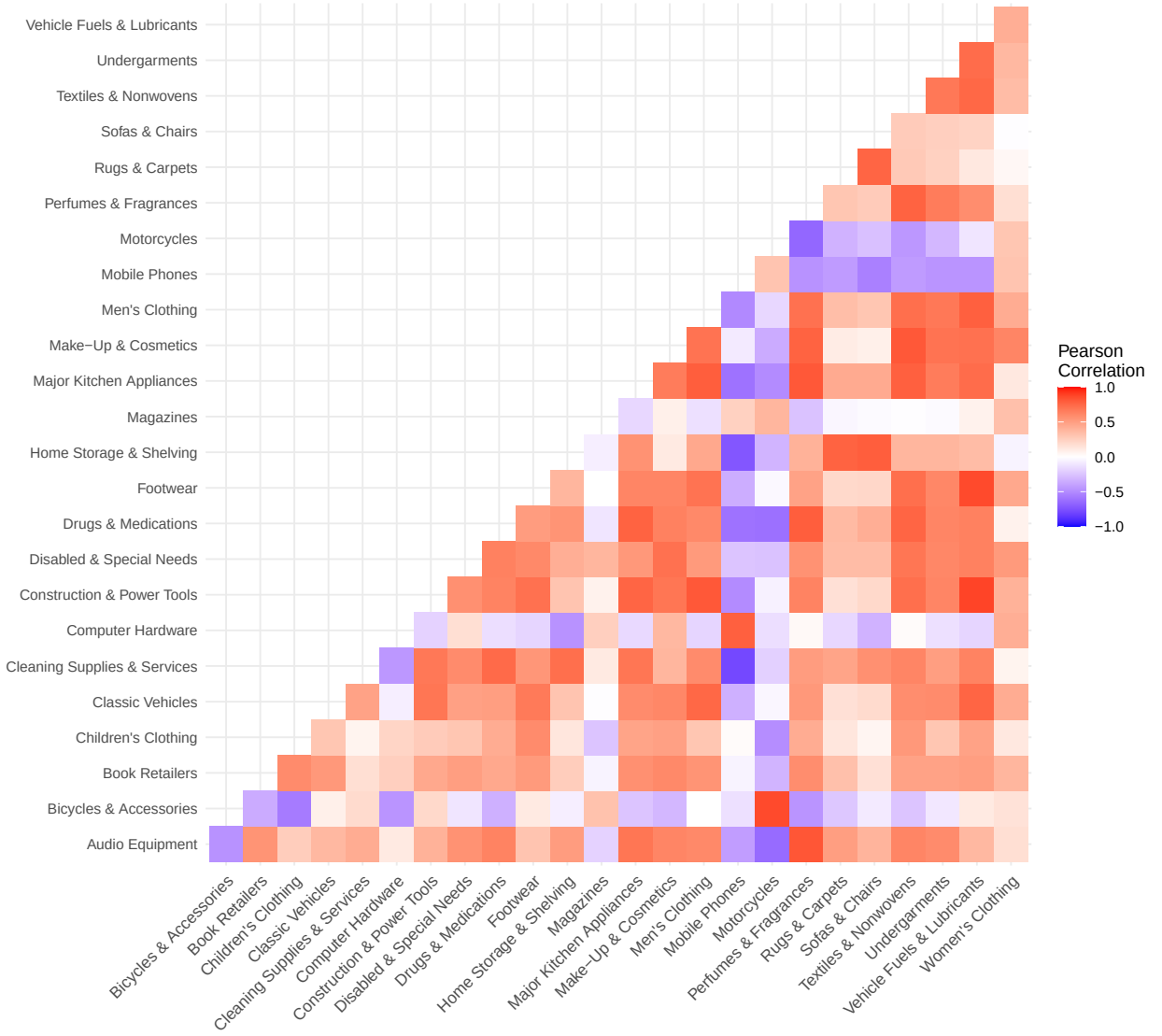
```
prod_f <- prod %>% filter(Date >'2014-01-01')
cormat <- round(cor(prod_f[,2:length(prod_f)]),2)
  get_lower_tri<-function(cormat){
    cormat[upper.tri(cormat, diag=TRUE)] <- NA
    return(cormat)
  }

lower_tri <- get_lower_tri(cormat)
melted_cormat <- melt(lower_tri, na.rm = TRUE)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  theme_minimal() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
```

```
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 10, hjust = 1)) +
  labs(x = NULL, y = NULL)
```




```
neprod_f <- neprod %>% filter(Date >'2014-01-01')
cormat_neprod <- round(cor(neprod_f[,2:length(neprod_f)]),2)
lower_tri <- get_lower_tri(cormat_neprod)
melted_cormat <- melt(lower_tri, na.rm = TRUE)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  theme_minimal() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0,
                       limit = c(-1,1), space = "Lab", name="Pearson\nCorrelation") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 10, hjust = 1)) +
  labs(x = NULL, y = NULL)
```



Intuitivement c'est clair qu'avec la popularité croissante d'Internet le nombre des requetes en ligne a augmenté aussi. Je vais effectuer des tests de stationnarité pour les séries Google Trends. Je vais utiliser le test Dickie-Fuller et le test KPSS (Kwiatkowski-Phillips-Schmidt-Shin). L'hypothèse nulle du premier test est que la série est non stationnaire. L'hypothèse nulle du deuxième test est que la série est stationnaire.

```
kable(prod %>%
 gather(Series, Value, -Date) %>%
  group_by(Series) %>%
   summarise(
    adf.pvalue = adf.test(Value, alternative = "stationary")$p.value,
    adf = adf.test(Value, alternative = "stationary")$p.value<0.05,
    kpss.pvalue=kpss.test(Value)$p.value,
    kpss=kpss.test(Value)$p.value>0.05
  ))
```

Series	adf.pvalue	adf	kpss.pvalue	kpss
Alcoholic Beverages	0.4732961	FALSE	0.01	FALSE
Baked Goods	0.4986986	FALSE	0.01	FALSE
Candy & Sweets	0.2417126	FALSE	0.01	FALSE
Coffee & Tea	0.1282605	FALSE	0.01	FALSE
Fruits & Vegetables	0.0100000	TRUE	0.01	FALSE
Meat & Seafood	0.0100000	TRUE	0.01	FALSE
Tobacco Products	0.6298057	FALSE	0.01	FALSE

```
kable(neprod %>%
 gather(Series, Value, -Date) %>%
  group_by(Series) %>%
   summarise(
    adf.pvalue = adf.test(Value, alternative = "stationary")$p.value,
    adf = adf.test(Value, alternative = "stationary")$p.value<0.05,
    kpss.pvalue=kpss.test(Value)$p.value,
    kpss=kpss.test(Value)$p.value>0.05
  ))
```

Series	adf.pvalue	adf	kpss.pvalue	kpss
Audio Equipment	0.2146570	FALSE	0.01	FALSE
Bicycles & Accessories	0.0100000	TRUE	0.01	FALSE
Book Retailers	0.4719804	FALSE	0.01	FALSE
Children's Clothing	0.0100000	TRUE	0.01	FALSE
Classic Vehicles	0.7692111	FALSE	0.01	FALSE
Cleaning Supplies & Services	0.0539337	FALSE	0.01	FALSE
Computer Hardware	0.6288243	FALSE	0.01	FALSE
Construction & Power Tools	0.6041251	FALSE	0.01	FALSE
Disabled & Special Needs	0.4958731	FALSE	0.01	FALSE
Drugs & Medications	0.2212388	FALSE	0.01	FALSE
Footwear	0.9830898	FALSE	0.01	FALSE
Home Storage & Shelving	0.0921395	FALSE	0.01	FALSE
Magazines	0.7286652	FALSE	0.01	FALSE
Major Kitchen Appliances	0.1911623	FALSE	0.01	FALSE
Make-Up & Cosmetics	0.5910779	FALSE	0.01	FALSE
Men's Clothing	0.9409494	FALSE	0.01	FALSE
Mobile Phones	0.9900000	FALSE	0.01	FALSE
Motorcycles	0.6491120	FALSE	0.01	FALSE
Perfumes & Fragrances	0.0100000	TRUE	0.01	FALSE
Rugs & Carpets	0.0100000	TRUE	0.01	FALSE
Sofas & Chairs	0.0100000	TRUE	0.01	FALSE
Textiles & Nonwovens	0.4545967	FALSE	0.01	FALSE
Undergarments	0.5252672	FALSE	0.01	FALSE
Vehicle Fuels & Lubricants	0.9315564	FALSE	0.01	FALSE
Women's Clothing	0.9607259	FALSE	0.01	FALSE

Pour les produits alimentaires, le test de Dickie-Fuller a montré que 5 des 7 séries étaient non stationnaires. Le test du KPSS a montré que toutes les séries sont non stationnaires.

Pour les produits non alimentaires, le test de Dickie-Fuller a révélé que 20 des 25 séries étaient non stationnaires. Le test du KPSS a également montré que toutes les séries sont non stationnaires.

Donc, je prends les données du mois au mois de l'année précédente et considère les séries en logarithmes. Ensuite, je supprime les données avant 2015 pour la comparabilité des calculs.

Toutes les données macroéconomiques sont aussi prises sous la forme d'un mois au mois de l'année précédente

afin d'éliminer une composante saisonnière et une non-stationnarité possible. Les données sur le prix du pétrole sont ajustées à l'indice américain de prix à la consommation, dégagé de la saisonnalité (Calculs dans Excel). Toutes les séries sont considérées dans les logarithmes. Pour unification de dimension et comparabilité des erreurs de prévision je mets à l'échelle de 0 à 100 toutes les valeurs macroéconomiques considerés.

```
f <- function(x) scaler(x, min = 0, max = 100)
macro_sc <- as.data.frame(sapply(macro[2:length(macro)], f))

df <- log(prod[2:length(prod)])
df2 <- data.frame(diff(as.matrix(df), differences = 12))
df2['date'] = prod %>% filter(Date > '2005-01-01') %>% select('Date')
prod_sc <- df2 %>% filter(date > '2015-01-01') %>% select(-'date')


df <- log(neprod[2:length(neprod)])
df2 <- data.frame(diff(as.matrix(df), differences = 12))
df2['date'] = neprod %>% filter(Date > '2005-01-01') %>% select('Date')
neprod_sc <- df2 %>% filter(date > '2015-01-01') %>% select(-'date')
```

Après les modifications des données, je vais faire à nouveau les tests. On voit que les deux tests signalent de la stationnarité des données.

```
kable(prod_sc %>%
 gather(Series, Value) %>%
  group_by(Series) %>%
   summarise(
    adf.pvalue = adf.test(Value, alternative = "stationary")$p.value,
    adf = adf.test(Value, alternative = "stationary")$p.value<0.05,
    kpss.pvalue=kpss.test(Value)$p.value,
    kpss=kpss.test(Value)$p.value>0.05
  ))
```

Series	adf.pvalue	adf	kpss.pvalue	kpss
Alcoholic.Beverages	0.01	TRUE	0.1	TRUE
Baked.Goods	0.01	TRUE	0.1	TRUE
Candy...Sweets	0.01	TRUE	0.1	TRUE
Coffee...Tea	0.01	TRUE	0.1	TRUE
Fruits...Vegetables	0.01	TRUE	0.1	TRUE
Meat...Seafood	0.01	TRUE	0.1	TRUE
Tobacco.Products	0.01	TRUE	0.1	TRUE

```
kable(neprod_sc %>%
 gather(Series, Value) %>%
  group_by(Series) %>%
   summarise(
    adf.pvalue = adf.test(Value, alternative = "stationary")$p.value,
    adf = adf.test(Value, alternative = "stationary")$p.value<0.05,
    kpss.pvalue=kpss.test(Value)$p.value,
    kpss=kpss.test(Value)$p.value>0.05
  ))
```

Series	adf.pvalue	adf	kpss.pvalue	kpss
Audio.Equipment	0.01	TRUE	0.1	TRUE
Bicycles...Accessories	0.01	TRUE	0.1	TRUE
Book.Retailers	0.01	TRUE	0.1	TRUE
Children.s.Clothing	0.01	TRUE	0.1	TRUE
Classic.Vehicles	0.01	TRUE	0.1	TRUE
Cleaning.Supplies...Services	0.01	TRUE	0.1	TRUE
Computer.Hardware	0.01	TRUE	0.1	TRUE
Construction...Power.Tools	0.01	TRUE	0.1	TRUE
Disabled...Special.Needs	0.01	TRUE	0.1	TRUE
Drugs...Medications	0.01	TRUE	0.1	TRUE
Footwear	0.01	TRUE	0.1	TRUE
Home.Storage...Shelving	0.01	TRUE	0.1	TRUE
Magazines	0.01	TRUE	0.1	TRUE
Major.Kitchen.Appliances	0.01	TRUE	0.1	TRUE
Make.Up...Cosmetics	0.01	TRUE	0.1	TRUE
Men.s.Clothing	0.01	TRUE	0.1	TRUE
Mobile.Phones	0.01	TRUE	0.1	TRUE
Motorcycles	0.01	TRUE	0.1	TRUE
Perfumes...Fragrances	0.01	TRUE	0.1	TRUE
Rugs...Carpets	0.01	TRUE	0.1	TRUE
Sofas...Chairs	0.01	TRUE	0.1	TRUE
Textiles...Nonwovens	0.01	TRUE	0.1	TRUE
Undergarments	0.01	TRUE	0.1	TRUE
Vehicle.Fuels...Lubricants	0.01	TRUE	0.1	TRUE
Women.s.Clothing	0.01	TRUE	0.1	TRUE

Sur la base de l'existence d'une corrélation entre les tendances, je considérerai des modèles avec des composantes principales des tendances alimentaires et non alimentaires.

Analyse des composantes principales

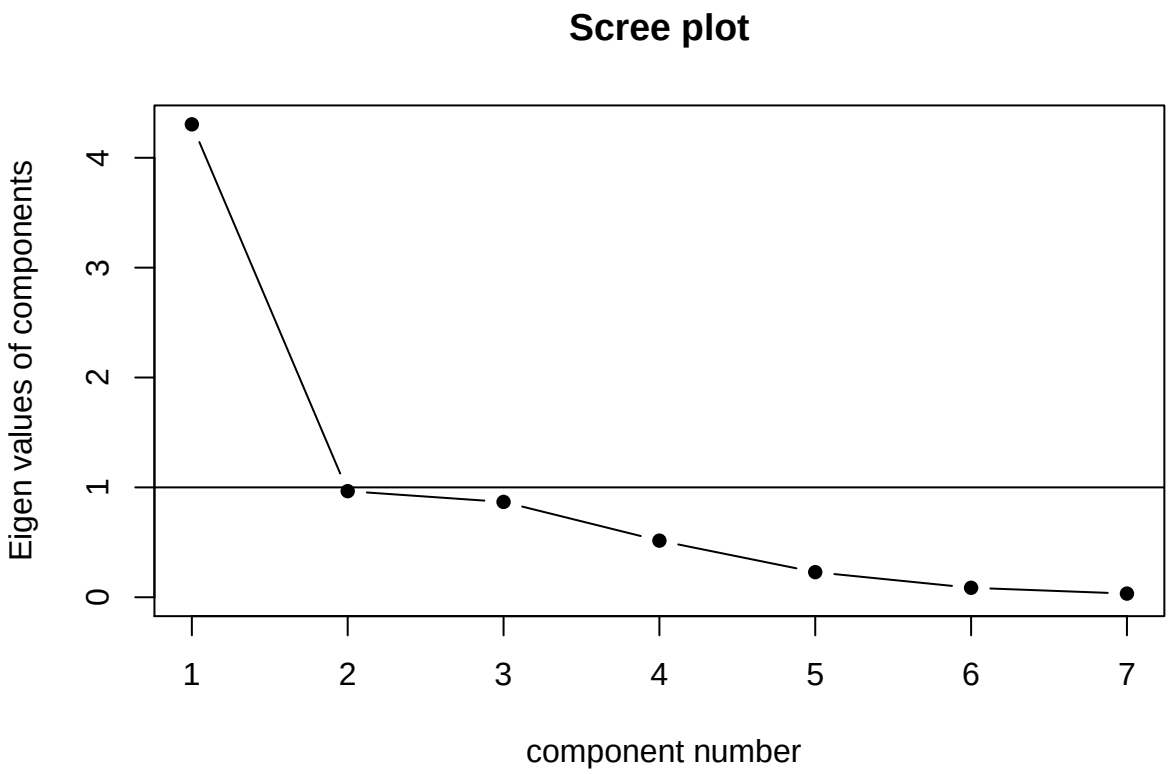
Dans l'analyse des composantes principales, différents critères sont utilisés pour choisir le nombre optimal de composantes. Sur la base du critère de Kaiser (Kaiser, 1960), seules les composantes dont les valeurs propres sont supérieures à 1 peuvent être gardés.

```
## pour une meilleure compréhension de la visualisation
names(prod_sc) <- gsub("[[:lower:]]|(\\.\\.)",'',names(prod_sc))
acp <- PCA(prod_sc, scale.unit=T, graph=FALSE)
summary(acp)
```

```
##
## Call:
## PCA(X = prod_sc, scale.unit = T, graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               4.304   0.966   0.868   0.515   0.228   0.086   0.033
## % of var.             61.490  13.794  12.394   7.359   3.262   1.227   0.473
## Cumulative % of var.  61.490  75.284  87.678  95.037  98.299  99.527 100.000
##
## Individuals (the 10 first)
##         Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## 1    |  1.929 |  0.888  0.247  0.212 | -0.290  0.118  0.023 |  0.201  0.063
```

```
## 2   |  1.730 | -0.626  0.123  0.131 |  0.601  0.505  0.121 | -0.012  0.000
## 3   |  1.447 | -0.238  0.018  0.027 | -0.612  0.524  0.179 | -0.039  0.002
## 4   |  1.533 |  1.092  0.374  0.507 |  0.068  0.007  0.002 | -0.371  0.215
## 5   |  2.152 | -1.653  0.858  0.590 |  0.736  0.759  0.117 |  0.883  1.215
## 6   |  2.676 |  1.596  0.799  0.356 | -1.212  2.054  0.205 | -0.975  1.479
## 7   |  2.749 | -1.140  0.408  0.172 |  1.077  1.622  0.153 |  0.382  0.227
## 8   |  2.619 |  0.466  0.068  0.032 | -0.642  0.577  0.060 |  0.457  0.326
## 9   |  2.433 |  0.064  0.001  0.001 |  0.332  0.154  0.019 | -1.110  1.921
## 10  |  2.134 | -0.637  0.127  0.089 | -0.339  0.161  0.025 |  1.328  2.747
##       cos2
## 1    0.011 |
## 2    0.000 |
## 3    0.001 |
## 4    0.059 |
## 5    0.168 |
## 6    0.133 |
## 7    0.019 |
## 8    0.031 |
## 9    0.208 |
## 10   0.387 |
##
## Variables
##        Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
## A.B |  0.851 16.821  0.724 | -0.458 21.698  0.210 | -0.055  0.350  0.003 |
## B.G |  0.639  9.499  0.409 |  0.112  1.301  0.013 |  0.732 61.836  0.536 |
## C.S |  0.905 19.015  0.818 | -0.041  0.173  0.002 |  0.069  0.551  0.005 |
## C.T |  0.732 12.437  0.535 |  0.200  4.131  0.040 | -0.010  0.011  0.000 |
## F.V |  0.787 14.375  0.619 |  0.037  0.143  0.001 | -0.560 36.129  0.313 |
## M.S |  0.941 20.558  0.885 | -0.279  8.075  0.078 |  0.019  0.041  0.000 |
## T.P |  0.560  7.294  0.314 |  0.789 64.479  0.623 | -0.097  1.082  0.009 |
```

```
scree(prod_sc, fa = FALSE, pc = TRUE)
```



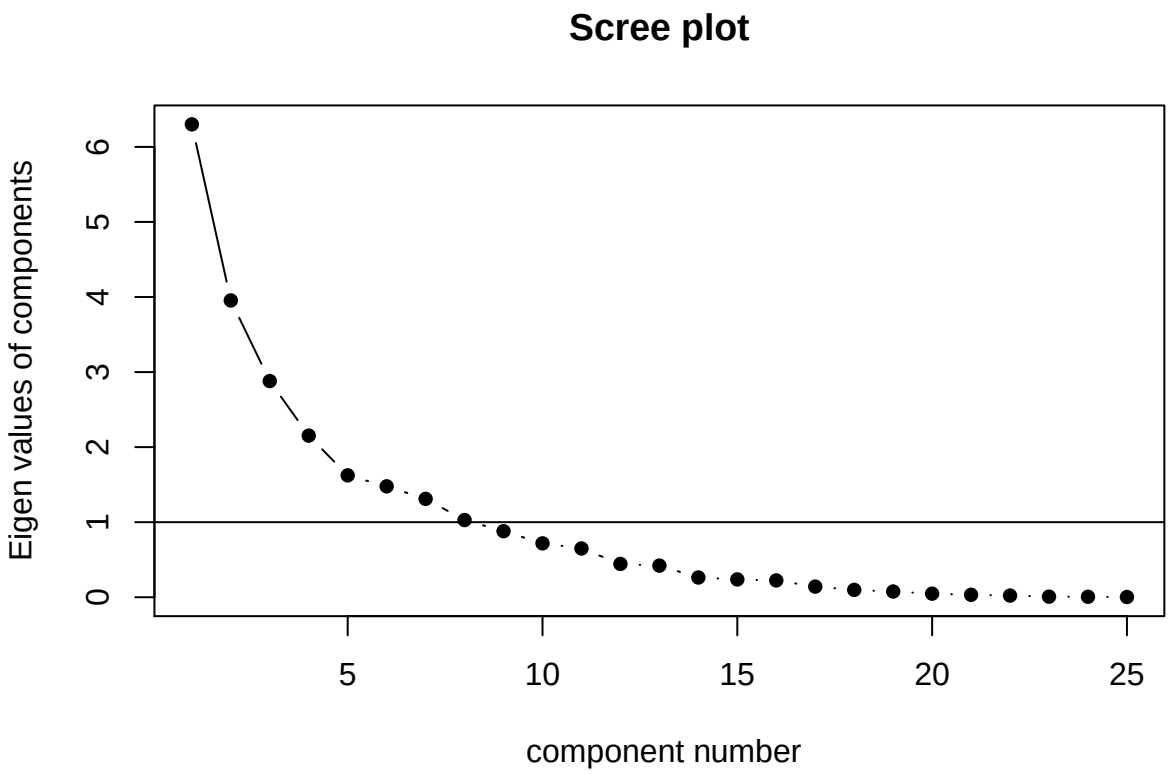
Dans le tableau des valeurs propres et sur la graphique ci-dessus, on voit que seule la première valeur propre satisfait au critère de Kaiser. Cependant, puisque la deuxième valeur propre est proche de 1, alors je le considérerai aussi.

```
## pour une meilleure compréhension de la visualisation
names(neprod_sc) <- gsub("[[:lower:]]|(\\.\\.)",'',names(neprod_sc))
acp <- PCA(neprod_sc, scale.unit=T, graph=FALSE)
summary(acp)
```

```
##
## Call:
## PCA(X = neprod_sc, scale.unit = T, graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               6.300   3.954   2.880   2.153   1.624   1.478   1.310
## % of var.             25.202  15.817  11.521   8.613   6.495   5.910   5.242
## Cumulative % of var.  25.202  41.019  52.540  61.152  67.648  73.558  78.799
##                        Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance               1.028   0.880   0.718   0.650   0.444   0.421   0.263
## % of var.              4.114   3.520   2.871   2.599   1.775   1.685   1.051
## Cumulative % of var.  82.913  86.434  89.305  91.904  93.678  95.364  96.415
##                       Dim.15  Dim.16  Dim.17  Dim.18  Dim.19  Dim.20  Dim.21
## Variance               0.238   0.224   0.142   0.098   0.076   0.047   0.032
## % of var.              0.950   0.896   0.566   0.392   0.305   0.190   0.130
## Cumulative % of var.  97.365  98.262  98.828  99.220  99.524  99.714  99.844
```

```
##                        Dim.22  Dim.23  Dim.24  Dim.25
## Variance                0.022   0.008   0.006   0.003
## % of var.               0.088   0.031   0.025   0.012
## Cumulative % of var.   99.932  99.963  99.988 100.000
##
## Individuals (the 10 first)
##           Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## 1     |  4.959 | -0.817  0.143  0.027 | -1.860  1.182  0.141 | -0.782  0.287
## 2     |  4.780 |  0.837  0.150  0.031 |  2.149  1.578  0.202 |  0.430  0.087
## 3     |  4.789 | -1.685  0.609  0.124 | -2.111  1.522  0.194 |  0.173  0.014
## 4     |  5.022 |  3.003  1.934  0.358 |  1.685  0.970  0.113 | -1.317  0.814
## 5     |  6.191 | -4.708  4.755  0.578 | -1.037  0.368  0.028 |  2.704  3.429
## 6     |  7.562 |  6.089  7.952  0.648 |  0.615  0.129  0.007 | -3.605  6.097
## 7     |  7.619 | -6.281  8.462  0.680 | -0.382  0.050  0.003 |  3.486  5.700
## 8     |  6.078 |  4.925  5.202  0.656 |  0.105  0.004  0.000 | -2.294  2.468
## 9     |  4.472 | -2.961  1.880  0.438 |  0.471  0.076  0.011 |  0.636  0.190
## 10    |  4.001 |  1.322  0.375  0.109 | -1.084  0.401  0.073 |  0.704  0.233
##         cos2
## 1      0.025 |
## 2      0.008 |
## 3      0.001 |
## 4      0.069 |
## 5      0.191 |
## 6      0.227 |
## 7      0.209 |
## 8      0.142 |
## 9      0.020 |
## 10     0.031 |
##
## Variables (the 10 first)
##          Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
## A.E   |  0.751  8.960  0.565 | -0.324  2.658  0.105 | -0.145  0.728  0.021 |
## B.A   |  0.262  1.088  0.069 |  0.761 14.650  0.579 |  0.309  3.321  0.096 |
## B.R   |  0.115  0.211  0.013 | -0.535  7.234  0.286 | -0.263  2.403  0.069 |
## C..C  | -0.227  0.818  0.052 |  0.491  6.099  0.241 | -0.507  8.909  0.257 |
## C.V   |  0.164  0.425  0.027 |  0.405  4.138  0.164 |  0.384  5.112  0.147 |
## C.S.S |  0.389  2.397  0.151 |  0.388  3.812  0.151 | -0.240  2.008  0.058 |
## C.H   |  0.816 10.567  0.666 | -0.061  0.093  0.004 | -0.204  1.448  0.042 |
## C.P.T |  0.815 10.553  0.665 | -0.093  0.219  0.009 | -0.053  0.098  0.003 |
## D.S.N |  0.307  1.498  0.094 | -0.527  7.023  0.278 |  0.269  2.509  0.072 |
## D.M   |  0.425  2.865  0.181 |  0.343  2.970  0.117 | -0.506  8.903  0.256 |
```

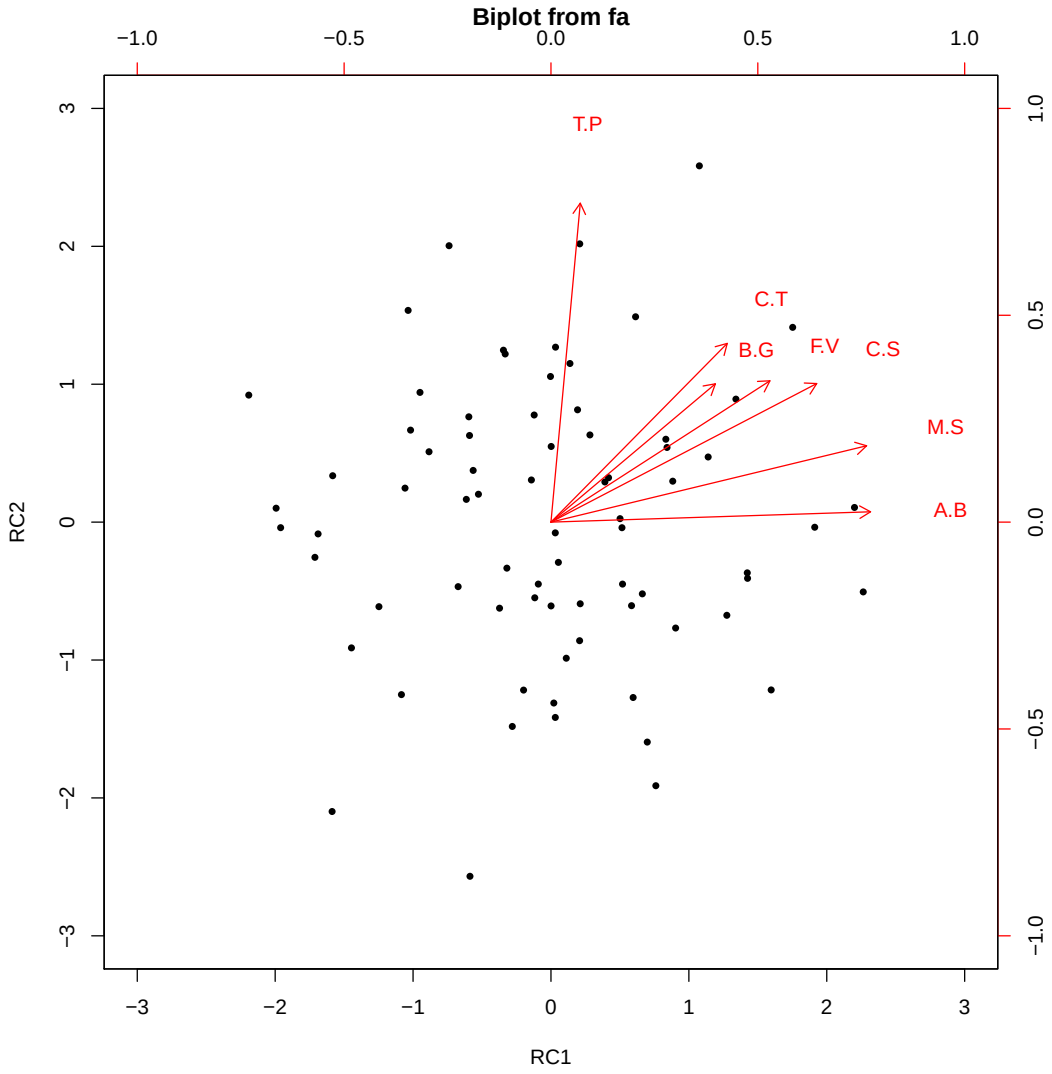
```
scree(neprod_sc, fa = FALSE, pc = TRUE)
```



Dans ce cas, les 8 premiers composantes principales répondent au critère de Kaiser.

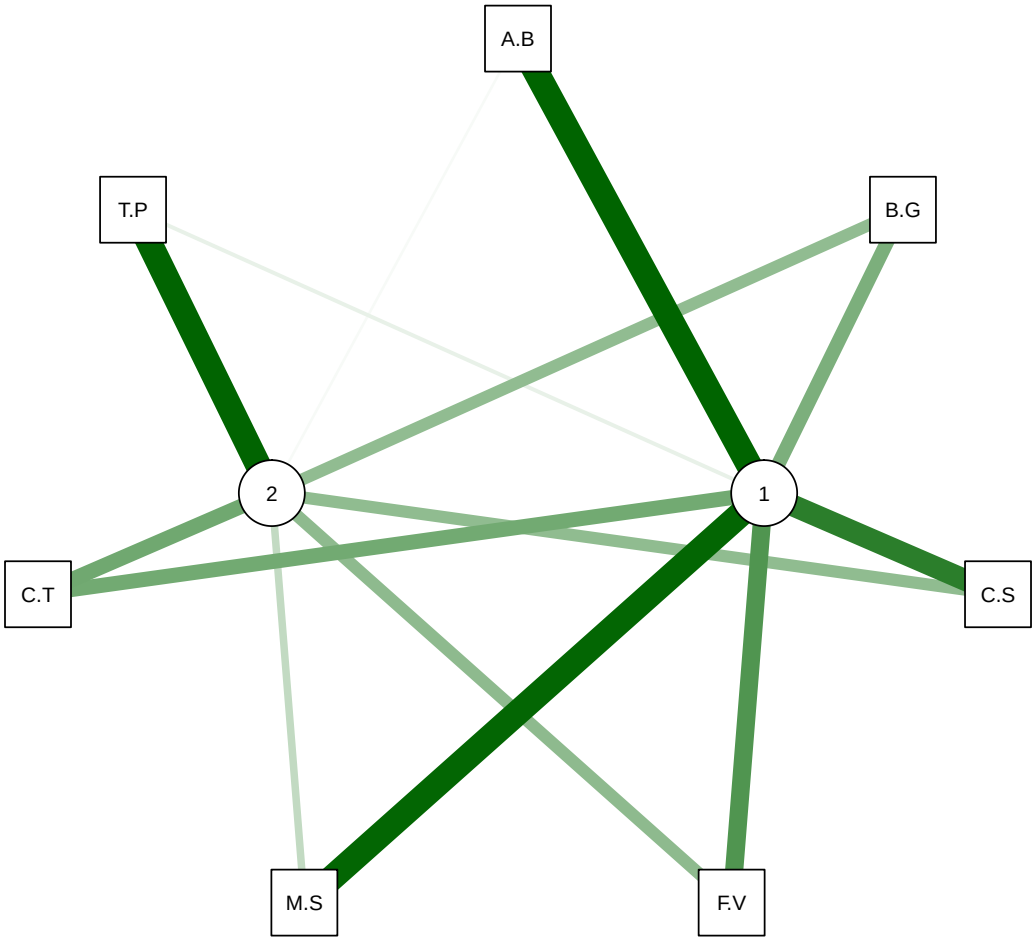
Aussi, dans l'analyse des composantes principales pour faciliter l'interprétation de ces composantes, une procédure de rotation d'axes est utilisée. Cette technique vise à obtenir un coefficient de corrélation le plus faible possible entre les composantes. L'une des techniques de rotation d'axe les plus populaires est la technique Varimax (Kaiser, 1958).

```
pca_rot_p <- principal(prod_sc, nfactors=2, rotate="varimax")
biplot.psych(pca_rot_p, col=c("black","red"), cex=c(0.7,1))
```



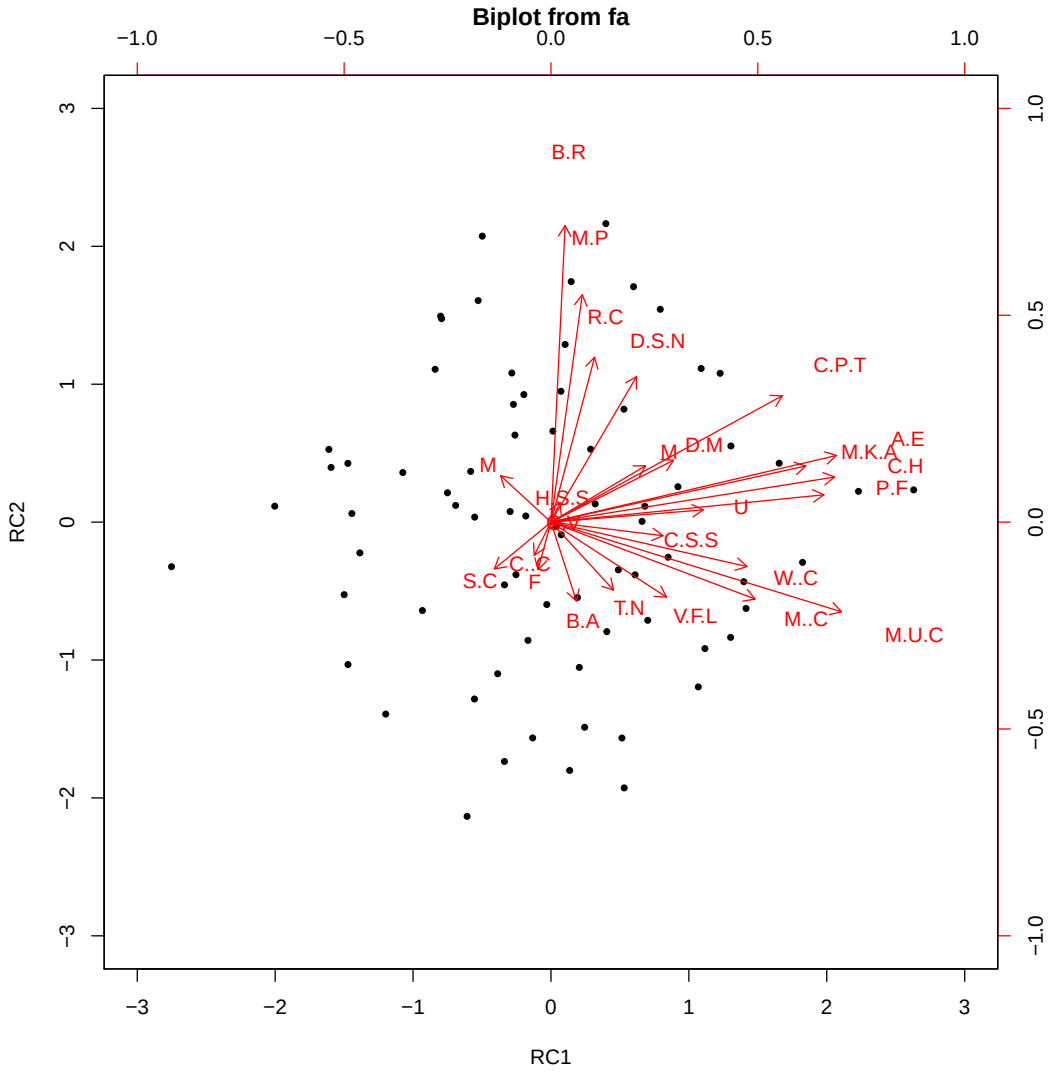

```
qgraph_loadings_plot <- function(loadings_in, title) {
  ld <- loadings(loadings_in)
  qg_pca <- qgraph(ld, title=title,
                   posCol = "darkgreen", negCol = "darkmagenta", arrows = FALSE,
                   labels=attr(ld, "dimnames")[[1]])}
qgraph_loadings_plot(pca_rot_p, "Principales composantes après rotation")
```

Principales composantes apres rotation



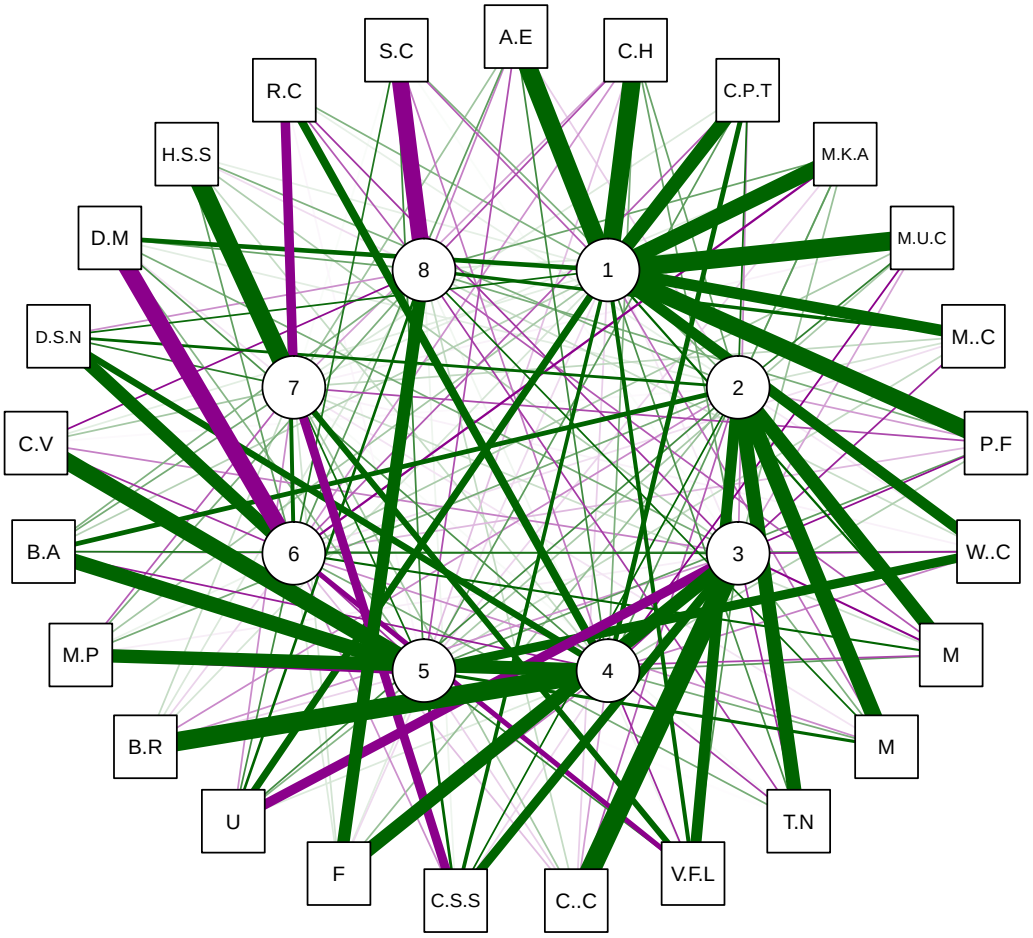
Il est possible maintenant d'observer quelles tendances sont contenues dans quels composantes. La première composante est principalement constituée de tendances `Alcogolic Beverages`, `Meat & Seafood` et `Fruits & Vegetables` et la seconde de tendance `Tobacco Products`.

```
pca_rot_np <- principal(neprod_sc, nfactors=8, rotate="varimax")
biplot.psych(pca_rot_np, col=c("black","red"), cex=c(0.7,1), choose=c(1,4))
```




```
qgraph_loadings_plot(pca_rot_np, "Principales composantes après rotation")
```

Principales composantes apres rotation



Les produits non alimentaires sont hétérogènes et nombreux, mais pour la plupart des produits, il est possible d'observer une séparation les variables par composantes assez claire.

Modèles avec les principales composantes.

Je vais maintenant essayer de construire un modèle Elastic Net dans le but d'obtenir une matrice creuse à partir de l'ensemble des données pour prévoir sur un échantillon de test de ventes au détail alimentaires et non alimentaires. Le meilleur modèle est défini comme le modèle qui a l'erreur de prédiction RMSE la plus faible.

```
# Build the model
pca_prod <- as.data.frame(sapply(as.data.frame(pca_rot_p$scores), f))

set.seed(42)
x = data.matrix(cbind(macro_sc, pca_prod) %>% select(-c(nonfood_goods)))
n = round(nrow(x)*0.7)

train.data <- x[1:n,]
test.data<- x[n:nrow(x),]
```

```r
# Tuning Elastic Net Hyperparameters
elastic <- train(
  food_goods ~ ., data = train.data, method = "glmnet",
  trControl = trainControl("cv", number = 10), ## set cross validation to 10 folders
  tuneLength = 10
  )
# Model coefficients
coef(elastic$finalModel, elastic$bestTune$lambda)
```

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) -29.17390663
## rvi           0.13448962
## oil           0.48756241
## real_rate     .
## ipd           0.77286599
## RC1           .
## RC2           0.01013921
```

```r
# Make predictions
predictions <- elastic %>% predict(test.data)
# Model prediction performance
data.frame(
  RMSE = RMSE(predictions, test.data[,1]),
  Rsquare = R2(predictions, test.data[,1])
)
```

```
##       RMSE  Rsquare
## 1 26.74146 0.398409
```

On voit que dans le meilleur modèle choisi par Elastic Net, seulment la deuxième composante joue un rôle. Ainsi, les requêtes de produits de tabac, qui sont correlés le plus avec la deuxième composante, ont un impact sur la dynamique des ventes au détail alimentaires. On peut dire, que des requêtes de boissons alcoolisées et de viande et de poisson ne l'affecte pratiquement pas.

Interpretation du coefficient: Avec une augmentation de 1 de valuer de la deuxième composante, des ventes au détail augmentent de 1% toutes choses egales par ailleurs.

```r
# Build the model
pca_neprod <- as.data.frame(sapply(as.data.frame(pca_rot_np$scores), f))
set.seed(42)
x = data.matrix(cbind(macro_sc, pca_neprod) %>% select(-c(food_goods)))
n = round(nrow(x)*0.7)

train.data <- x[1:n,]
test.data<- x[n:nrow(x),]

elastic <- train(
  nonfood_goods ~ ., data = train.data, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneLength = 10
  )
# Model coefficients
coef(elastic$finalModel, elastic$bestTune$lambda)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
```

```
##                      s1
## (Intercept) 39.45343199
## rvi          0.13575555
## oil          0.33254600
## real_rate    .
## ipd          0.21586227
## RC1          .
## RC4          .
## RC7          .
## RC2          0.02442757
## RC5          .
## RC8          .
## RC6          .
## RC3          .
```

```r
# Make predictions
predictions <- elastic %>% predict(test.data)
# Model prediction performance
data.frame(
  RMSE = RMSE(predictions, test.data[,1]),
  Rsquare = R2(predictions, test.data[,1])
)
```

```
##       RMSE   Rsquare
## 1 17.54306 0.7171976
```

Dans ce cas, la deuxième composante était aussi significative. Elle comprend en majeure partie les tendances `Magazines`, `Motorcycles`, `Textiles & Nonwovens` et `Vehicle Fuels & Lubricants`.

Interpretation du coefficient: Avec une augmentation de 1 de valuer de la deuxième composante, des ventes au détail non alimentaires augmentent de 2.4% toutes choses egales par ailleurs.

Conclusion

Dans le cadre de ce projet, j'ai examiné l'impact des composantes principales de Google Trends sur les ventes au détail alimentaires et non alimentaires. Tout d'abord, j'ai comparé la liste des produits qui composent le commerce de détail russe agrégé avec des catégories existantes dans Google Trends. Puis j'ai examiné les corrélations appariées des tendances choisis et ai ensuite testé leur stationnarité. Après, j'ai sélectionné les composantes principales des Google Trends alimentaires et non alimentaires en utilisant le critère de Kaiser, et ai également appliqué la technique Varimax pour minimiser la corrélation entre eux.

J'ai ajouté les composantes résultants ainsi que les variables macroéconomiques au modèle Elastic Net. Les résultats ont montré que dans le modèle avec la plus petite erreur de prédiction (RMSE) pour les produits alimentaires, il y avait seulement une deuxième composante. Un poids important dans la deuxième composante tombe sur les produits du tabac. La deuxième composante est aussi présent dans le modèle pour les produits non alimentaires. Elle capture 4 tendances sur 25 examinées.

Pour plus d'informations sur les données collectées et sur le sujet, veuillez chercher mon article dans le dossier envoyé.