Projet Fierthé

KOUASSI KONAN LEGER

2025-06-02

Objectif : Formuler des propositions claires pour la valorisation durable du thé en France

Phase 1 – Prise en main et cadrage

Objectifs

- Comprendre le contexte du projet et les enjeux de la filière théicole durable
- Se familiariser avec la base de données **MINTEL GNPD** et les outils

Tâches

- Lecture du projet Fierthé, documents internes et publications liées
- Prise en main de la base **MINTEL** (structure, variables, type de données)
- Rencontre avec le tuteur

Analyses univariées

```
# Importation des packages nécessaires
library(tidyverse)    # Inclut dplyr, tidyr, ggplot2, stringr
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(readxl)
library(lubridate)
library(wordcloud)
```

```
## Warning: le package 'wordcloud' a été compilé avec la version R 4.4.3
```

```
## Le chargement a nécessité le package : RColorBrewer
```

```r
library(tidytext)
```

```
## Warning: le package 'tidytext' a été compilé avec la version R 4.4.3
```

```r
library(scales)
```

```
##
## Attachement du package : 'scales'
##
## L'objet suivant est masqué depuis 'package:purrr':
##
##     discard
##
## L'objet suivant est masqué depuis 'package:readr':
##
##     col_factor
```

```r
library(tm)
```

```
## Warning: le package 'tm' a été compilé avec la version R 4.4.3
```

```
## Le chargement a nécessité le package : NLP
```

```
## Warning: le package 'NLP' a été compilé avec la version R 4.4.2
```

```
##
## Attachement du package : 'NLP'
##
## L'objet suivant est masqué depuis 'package:ggplot2':
##
##     annotate
```

```r
library(forcats)

# Nettoyage de l'environnement
rm(list = ls())

# Importation des données
gnpd_brut <- read_excel("GNPD_tea.xlsx")

# Analyse globale de la base de données
glimpse(gnpd_brut)
```

```
## Rows: 99,608
## Columns: 33
## $ `Record ID`                    <dbl> 12811782, 12811806, 12811808, 12824062, ~
## $ `Date Published`               <dttm> 2025-05-02, 2025-05-02, 2025-05-02, 202~
## $ `Format Type`                  <chr> NA, "Liquid", "Liquid", NA, NA, NA, "Loo~
## $ `Number of Variants`           <dbl> 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 4, 4, 4, 1~
## $ `Product Source`               <chr> "Physical Product", "Physical Product", ~
## $ `Ingredients (Standard form)`  <chr> "Tuna, Tuna Extract, technological addit~
## $ Company                        <chr> "Thai Inaba Foods", "Biogroupe", "KQ", "~
## $ Market                         <chr> "Italy", "Italy", "Italy", "Singapore", ~
## $ `Location of Manufacture`      <chr> "Thailand", NA, NA, "Taiwan, China", "Ta~
## $ Brand                          <chr> "Ciao Churu Pops", "Karma Kombucha", "Ko~
## $ `Launch Type`                  <chr> "New Product", "New Variety/Range Extens~
## $ Product                        <chr> "Tuna Recipe Cat Treat", "Sugar-Free Spa~
## $ `Price per 100 g/ml in Euros`  <chr> "5.42", "0.93", "1.03", "0.38", "0.38", ~
## $ `Sub-Category`                 <chr> "Cat Snacks & Treats", "Kombucha & Other~
## $ Storage                        <chr> "Shelf stable", "Chilled", "Shelf stable~
## $ `Positioning Claims`           <chr> "No Additives/Preservatives, Vitamin/Min~
## $ `Unit Pack Size (ml/g)`        <chr> "15.000", "750.000", "330.000", "315.000~
## $ `Packaging Units`              <chr> "g", "ml", "ml", "ml", "ml", "ml", "g", ~
## $ `Package Type`                 <chr> "Flexible stick-pack", "Bottle", "Bottle~
## $ `Package Material`             <chr> "Multi laminate", "Glass plain", "Glass ~
## $ `Price in US Dollars`          <chr> "3.70", "7.91", "3.87", "1.37", "1.37", ~
## $ `Price in Euros`               <chr> "3.25", "6.95", "3.40", "1.20", "1.20", ~
## $ `Bar Code`                     <dbl> 8.859388e+12, 3.760192e+12, 1.230000e+12~
## $ `Allergens / Warnings`         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ `Alcohol By Volume (%)`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Nutrition                      <chr> "Per 1kg: Energy 470kcal, Protein 9%, Fa~
## $ `Record hyperlink`             <chr> "=hyperlink(\"http://www.gnpd.com/sinatr~
## $ `Ingredients (On pack)`        <chr> "tuna (33.3%), tuna extract (0.7%), tech~
## $ `Company address`              <chr> "Nong Pla Mo", "Erquy", "Arco (trento)",~
## $ `Product Description`          <chr> "Ciao Churu Pops  Cibo Complementare per~
## $ `Import Status`                <chr> "Imported product", NA, NA, "Imported pr~
## $ `Private Label`                <chr> "Branded", "Branded", "Branded", "Brande~
## $ `Company Territory`            <chr> "Thailand", "France", "Italy", "Taiwan, ~
```

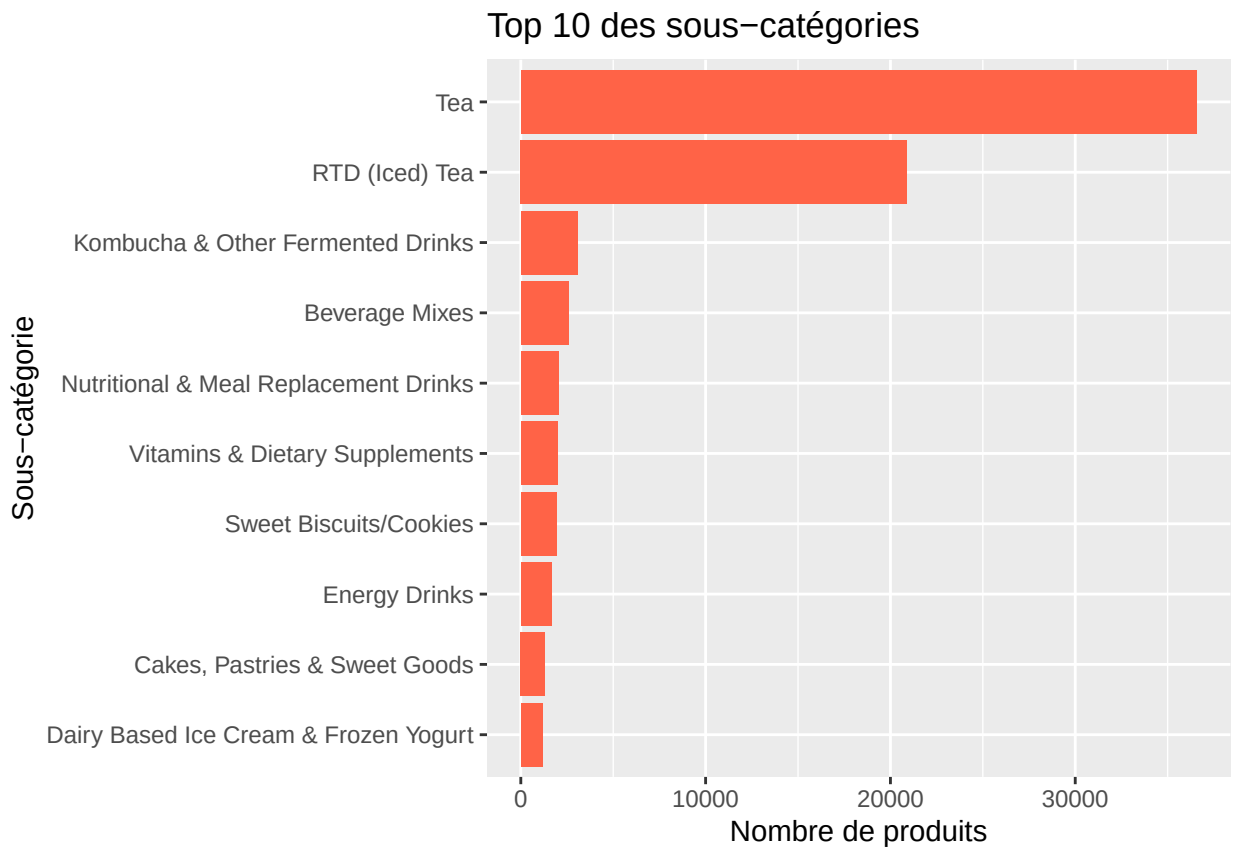
```r
# Conversion des variables en format approprié
gnpd_brut <- gnpd_brut %>%
  mutate(
    `Price per 100 g/ml in Euros` = as.numeric(`Price per 100 g/ml in Euros`),
    `Price in US Dollars` = as.numeric(`Price in US Dollars`),
    `Price in Euros` = as.numeric(`Price in Euros`),
    `Unit Pack Size (ml/g)` = as.numeric(`Unit Pack Size (ml/g)`),
    `Alcohol By Volume (%)` = as.numeric(`Alcohol By Volume (%)`),
    `Date Published` = as.Date(`Date Published`)
  )

# Élimination des variables non pertinentes
gnpd <- gnpd_brut %>%
  select(-c(`Record ID`,`Record hyperlink`,`Bar Code`,`Alcohol By Volume (%)`,
            `Allergens / Warnings`))

# Transformation des colonnes
gnpd <- gnpd %>%
```

```r
  mutate(
    price_per_100 = as.numeric(`Price per 100 g/ml in Euros`),
    year = year(`Date Published`),
    month = floor_date(`Date Published`, "month")
  )

# Filtrage des données aberrantes
total_lignes <- nrow(gnpd)

gnpd_filtre <- gnpd %>%
  filter(`Unit Pack Size (ml/g)` >= 2,
         `Unit Pack Size (ml/g)` <= 5000)

proportion_conservee <- nrow(gnpd_filtre) / total_lignes

# Affichage de la proportion conservée
proportion_conservee
```

```
## [1] 0.9045358
```

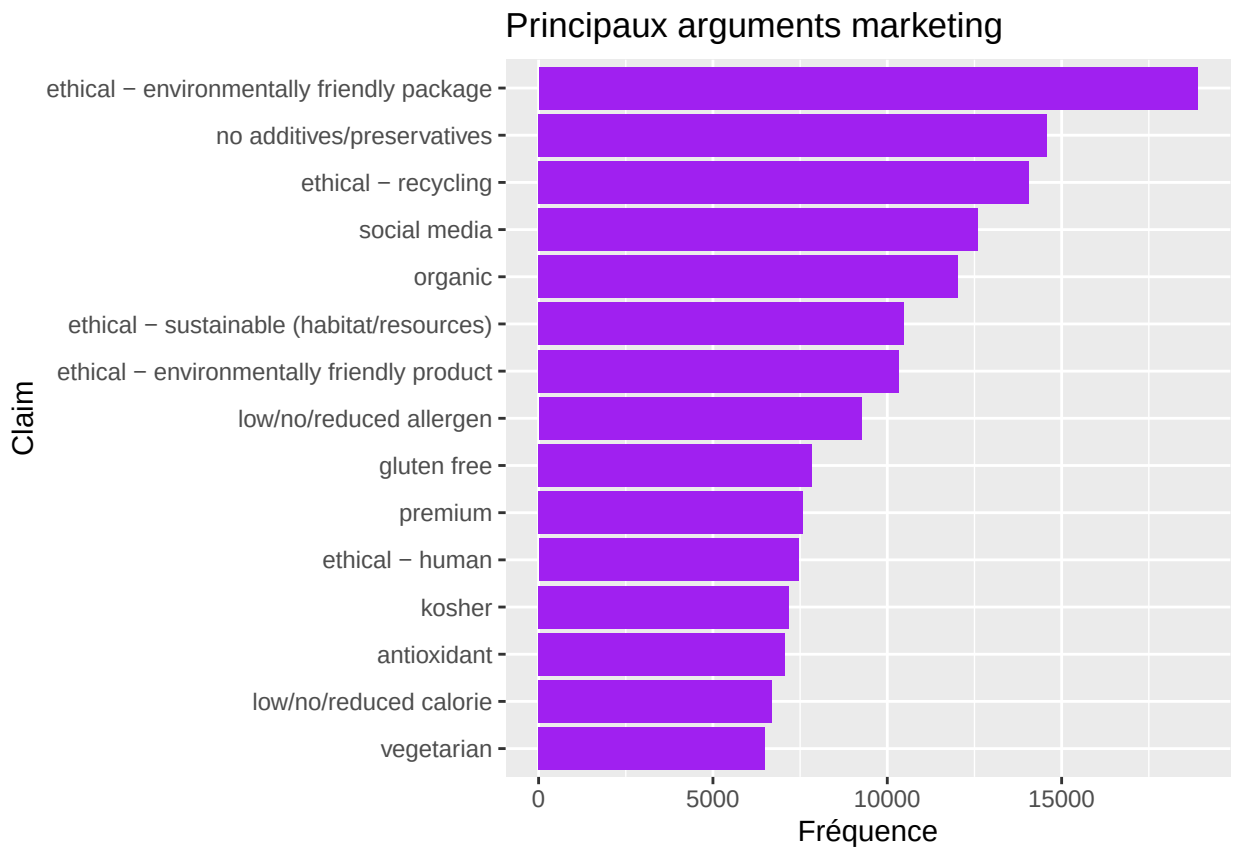
```r
gnpd <- gnpd_filtre
```

```r
# Sous-categorie
gnpd %>%
  count(`Sub-Category`, sort = TRUE) %>%
  slice_max(n, n = 10) %>%
  ggplot(aes(x = reorder(`Sub-Category`, n), y = n)) +
  geom_col(fill = "tomato") +
  coord_flip() +
  labs(title = "Top 10 des sous-catégories",
       x = "Sous-catégorie",
       y = "Nombre de produits")
```



Analyse de la variable “Positioning Claims” Je l’hypothese selon laquelle cette variable est suceptible d’influencer non seulement le prix mais aussi le choix des consommateurs.

```r
# Extraction des claims individuels
claims <- gnpd %>%
  filter(!is.na(`Positioning Claims`)) %>%
  unnest_tokens(claim, `Positioning Claims`, token = "regex",
                pattern = ",\\s*") %>%
  count(claim, sort = TRUE)

# Barplot des claims les plus fréquents
claims %>%
  slice_max(n, n = 15) %>%
  ggplot(aes(x = reorder(claim, n), y = n)) +
  geom_col(fill = "purple") +
  coord_flip() +
  labs(title = "Principaux arguments marketing", x = "Claim", y = "Fréquence")
```



Nous pouvons contacter sur ce graphique les effets dominents des differents produits. Essayons de projetter les mots dominents de issu de cette variable qui peut etre par la suite peut nous aider à avoir certaines combinansons qui peuvent aider à avoir un titre de produit de haute qualité.

Nuage de mot de la variable “Positioning claims”

Je me suis basé sur les exemple de ce lien : https://www.sthda.com/french/wiki/text-mining-et-nuage-de-mots-avec-le-logiciel-r-5-etapes-simples-a-savoir

```
library(tm)
set.seed(123)
text_corpus <- Corpus(VectorSource(gnpd$`Positioning Claims`))
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))

text_corpus <- tm_map(text_corpus, toSpace, "/")
```

```
## Warning in tm_map.SimpleCorpus(text_corpus, toSpace, "/"): transformation drops
## documents
```

```
text_corpus <- tm_map(text_corpus, toSpace, "-")
```

```
## Warning in tm_map.SimpleCorpus(text_corpus, toSpace, "-"): transformation drops
## documents
```

```
text_corpus <- tm_map(text_corpus, content_transformer(tolower))
```

```
## Warning in tm_map.SimpleCorpus(text_corpus, content_transformer(tolower)):
## transformation drops documents
```

```
text_corpus <- tm_map(text_corpus, removePunctuation)
```

```
## Warning in tm_map.SimpleCorpus(text_corpus, removePunctuation): transformation
## drops documents
```

```
text_corpus <- tm_map(text_corpus,
                      function(x)removeWords(x,stopwords(kind = "english")))
```

```
## Warning in tm_map.SimpleCorpus(text_corpus, function(x) removeWords(x,
## stopwords(kind = "english"))): transformation drops documents
```

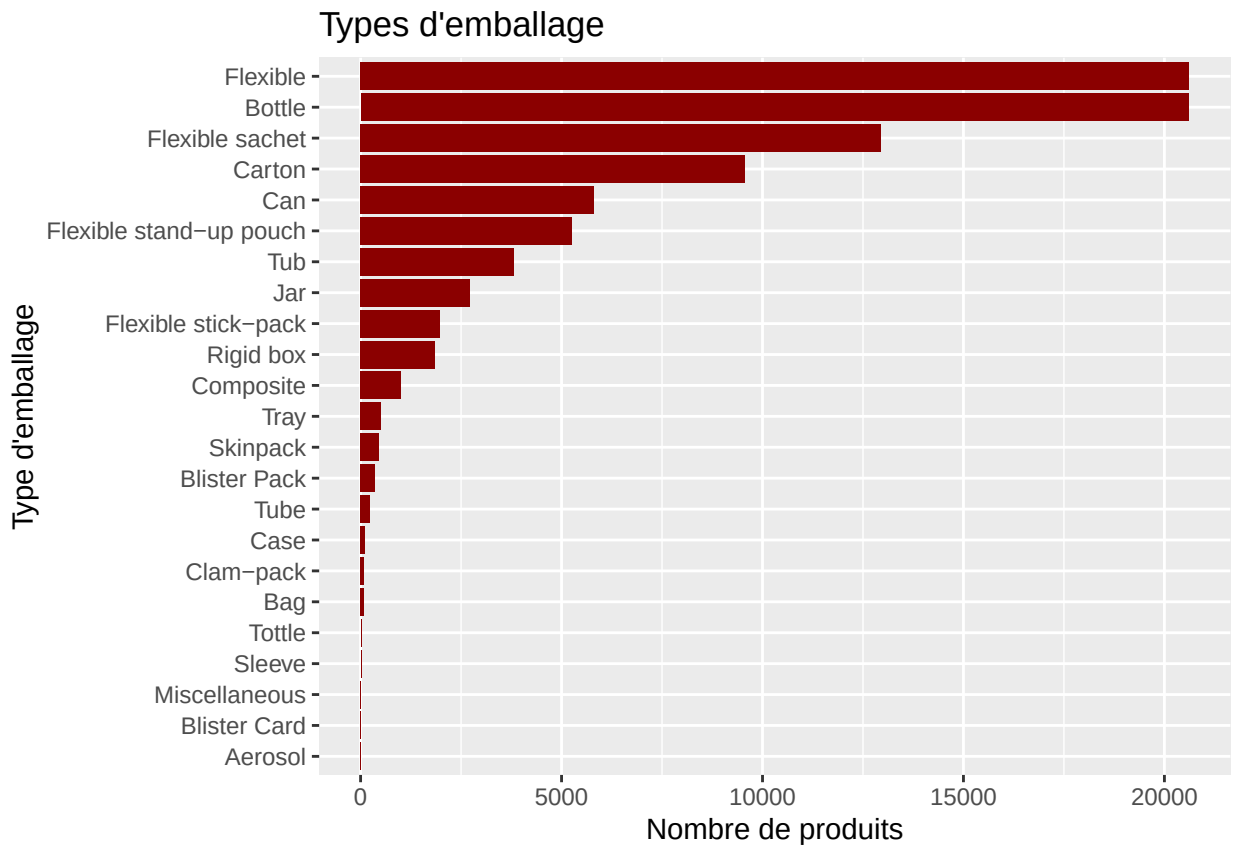
```
wordcloud(text_corpus, colors = brewer.pal(8, "Dark2"),
          rot.per=0, random.order = F)
```


```
dtm <- TermDocumentMatrix(text_corpus)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```

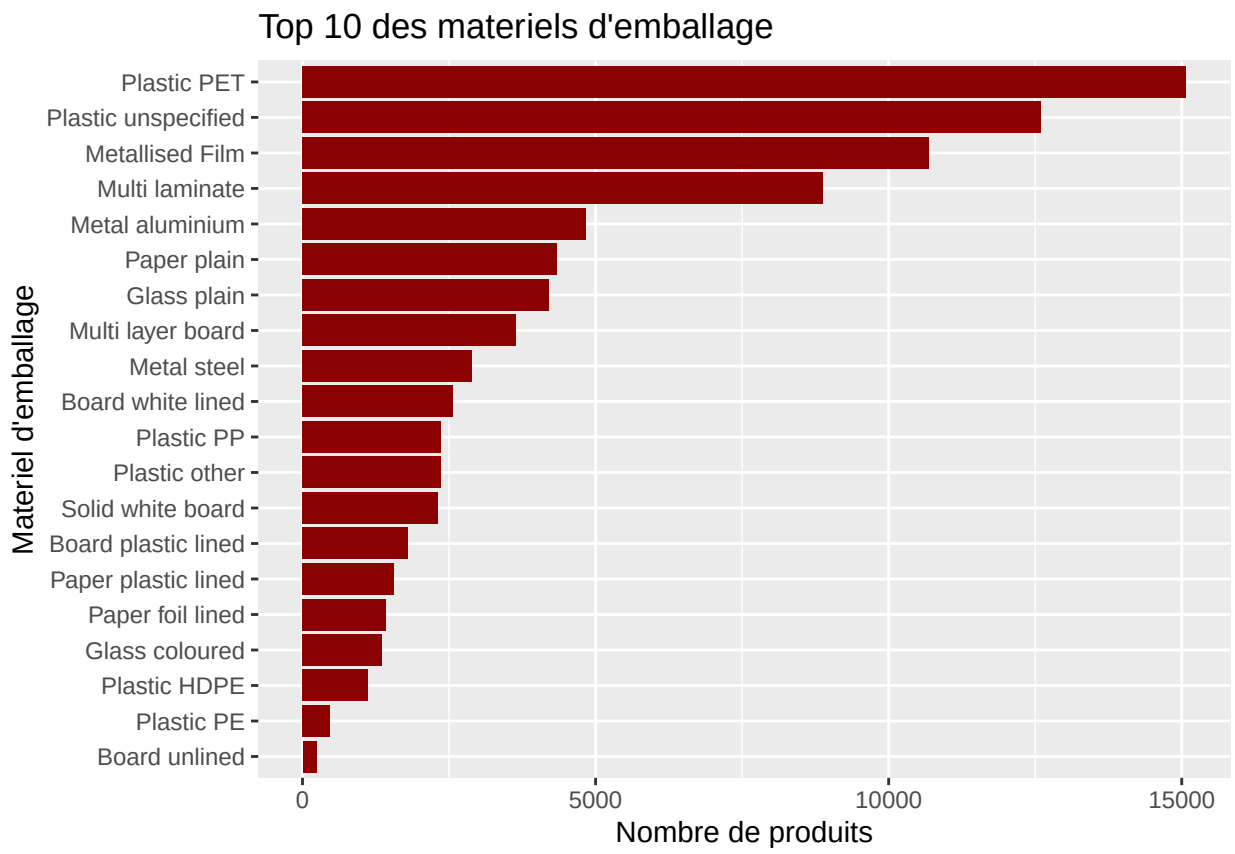
```
##                              word  freq
## ethical                   ethical 68145
## free                         free 41161
## functional           functional 31101
## environmentally environmentally 29235
## friendly               friendly 29235
## low                           low 24814
## reduced                 reduced 24814
## added                     added 21870
## preservatives     preservatives 20656
## package                 package 18895
```

Analyse des types d'embalage

```
gnpd %>%
  filter(!is.na(`Package Type`)) %>%
  count(`Package Type`, sort = TRUE,) %>%
  #slice_max(n, n = 10) %>%
  ggplot(aes(x = reorder(`Package Type`, n), y = n)) +
  geom_col(fill = "darkred") +
  coord_flip() +
  labs(title = "Types d'emballage", x = "Type d'emballage", y = "Nombre de produits")
```



```r
gnpd %>%
  filter(!is.na(`Package Material`)) %>%
  count(`Package Material`, sort = TRUE,) %>%
  slice_max(n, n = 20) %>%
  ggplot(aes(x = reorder(`Package Material`, n), y = n)) +
  geom_col(fill = "darkred") +
  coord_flip() +
  labs(title = "Top 10 des materiels d'emballage", x = "Materiel d'emballage", y = "Nombre de produits")
```



Analyse du prix

```r
summary(gnpd$`Price per 100 g/ml in Euros`)
```

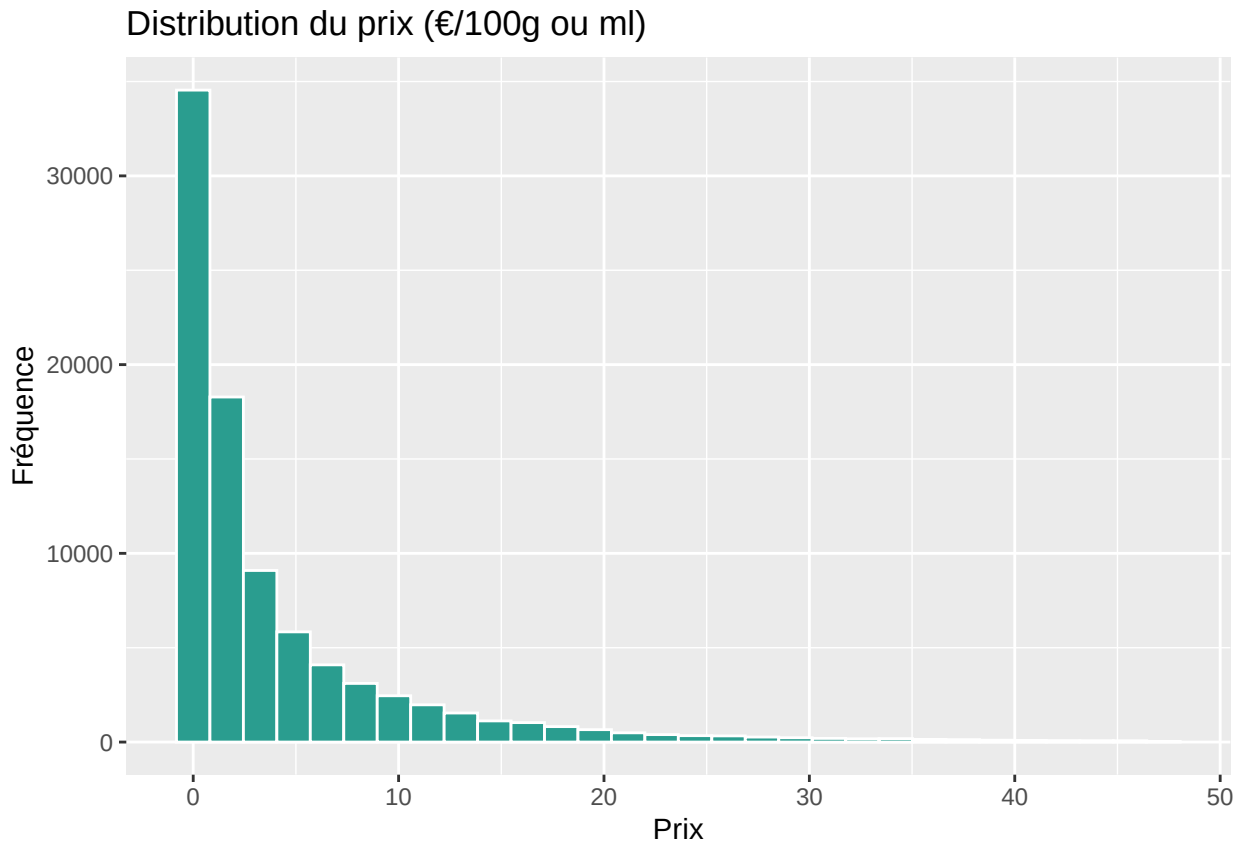
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.00    0.40    1.47    5.02    5.32 1774.23    1343
```

Etant donné qu'il semble avoir des variables aberrantes,nous allons, dans un premier temps, les 1% de valeurs extremes

```r
gnpd_clean <- gnpd %>%
  filter(!is.na(price_per_100),
         price_per_100 < quantile(price_per_100, 0.99, na.rm = T))
```

```r
        # on retire le top 1%

# visualisation
ggplot(gnpd_clean, aes(x = price_per_100)) +
  geom_histogram(fill = "#2a9d8f", color = "white") +
  labs(title = "Distribution du prix (€/100g ou ml)", x = "Prix", y = "Fréquence")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

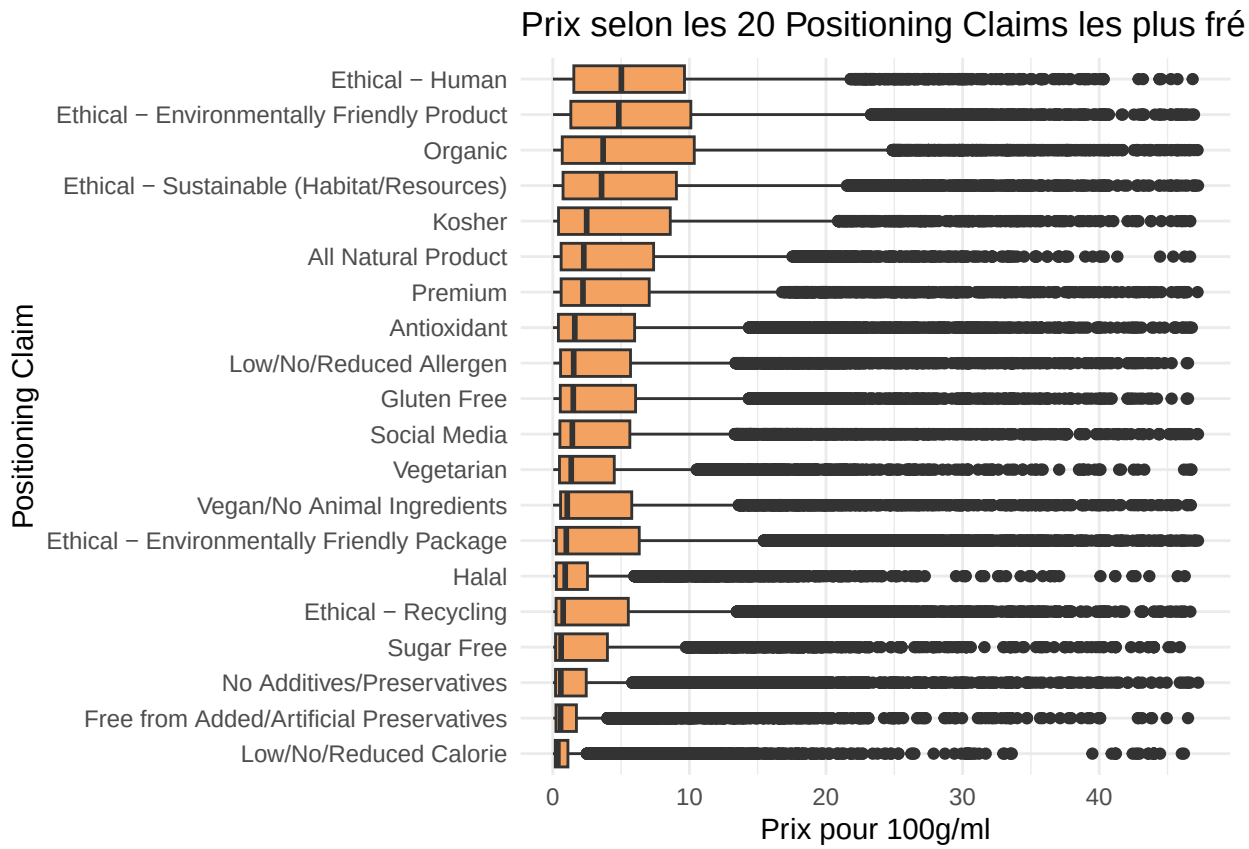


Analyses Bivariées

Analyse prix et Positioning claims

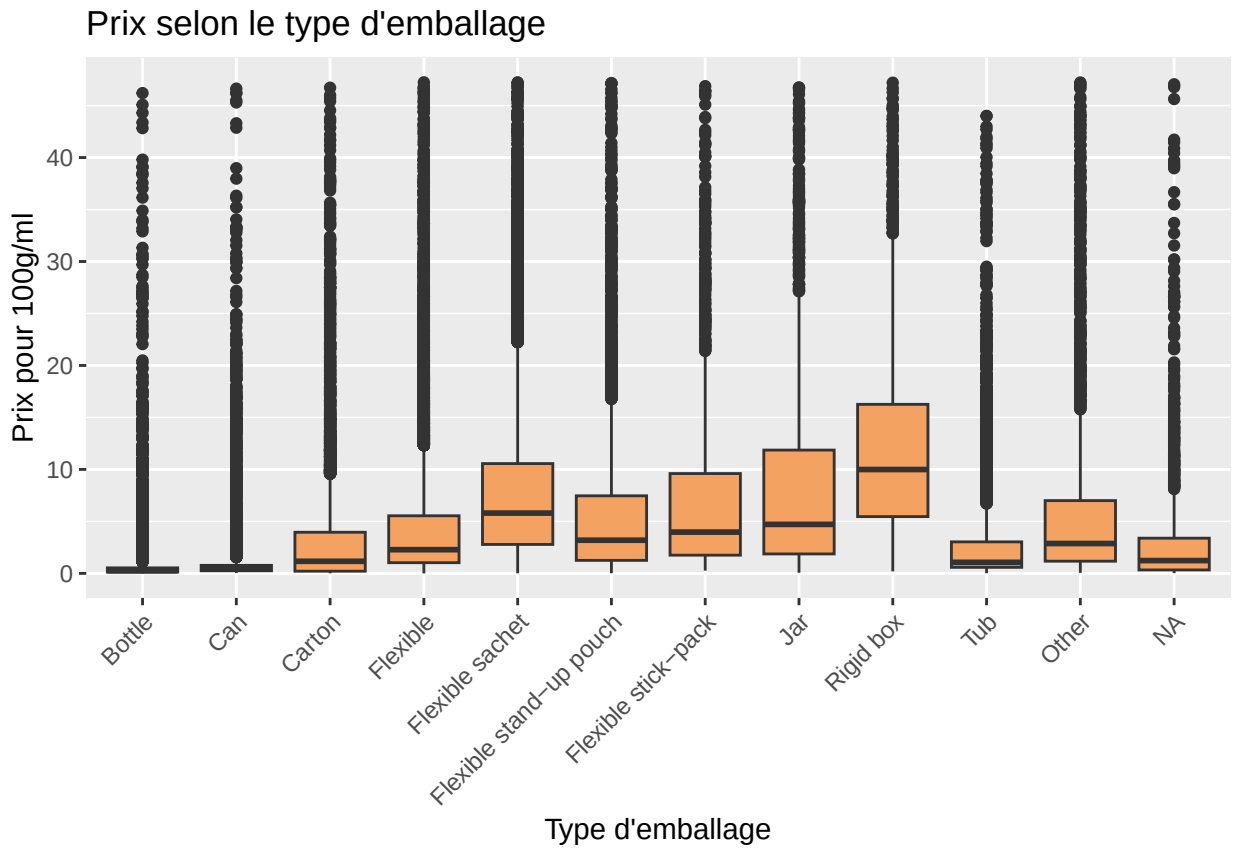
```r
# Étape 1 : Extraire tous les claims dans des lignes séparées
gnpd_long <- gnpd_clean %>%
  filter(!is.na(`Positioning Claims`)) %>%
  separate_rows(`Positioning Claims`, sep = ",\\s*")  # sépare les claims par virgule

# Étape 2 : Identifier les 10 claims les plus fréquents
top_claims <- gnpd_long %>%
  count(`Positioning Claims`, sort = TRUE) %>%
```

```
  slice_head(n = 20) %>%
  pull(`Positioning Claims`)

# Étape 3 : Filtrer uniquement les produits contenant ces claims
gnpd_top_claims <- gnpd_long %>%
  filter(`Positioning Claims` %in% top_claims)

# Étape 4 : Tracer le boxplot
ggplot(gnpd_top_claims, aes(x = fct_reorder(`Positioning Claims`,
                                            price_per_100, .fun = median), y = price_per_100)) +
  geom_boxplot(fill = "#f4a261") +
  labs(
    title = "Prix selon les 20 Positioning Claims les plus fréquents",
    x = "Positioning Claim",
    y = "Prix pour 100g/ml"
  ) +
  coord_flip() +  # pivote pour une lecture plus facile
  theme_minimal()
```

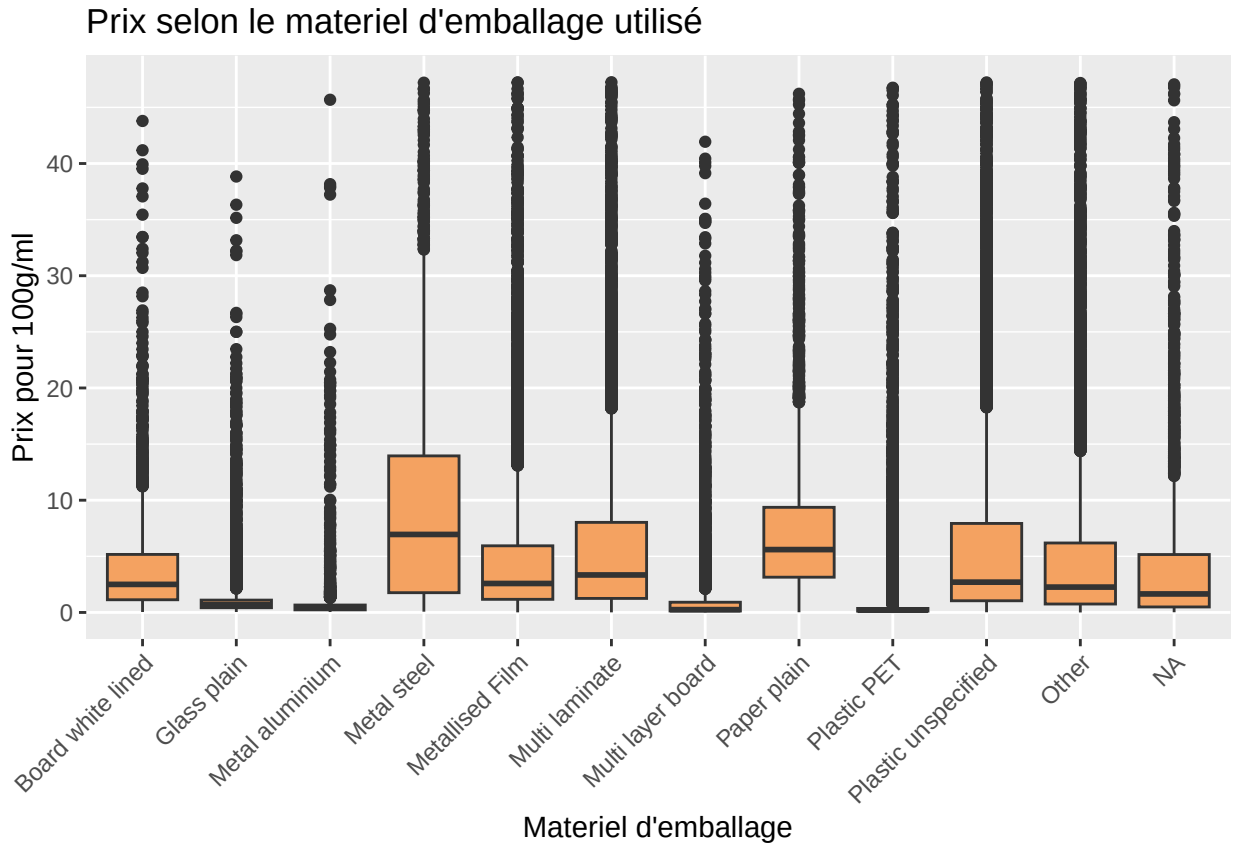


Prix vs Embalage

```
ggplot(gnpd_clean, aes(x = fct_lump(`Package Type`, 10), y = price_per_100)) +
  geom_boxplot(fill = "#f4a261") +
  labs(title = "Prix selon le type d'emballage", x = "Type d'emballage", y = "Prix pour 100g/ml") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
ggplot(gnpd_clean, aes(x = fct_lump(`Package Material`, 10), y = price_per_100)) +
  geom_boxplot(fill = "#f4a261") +
  labs(title = "Prix selon le materiel d'emballage utilisé", x = "Materiel d'emballage", y = "Prix pour
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

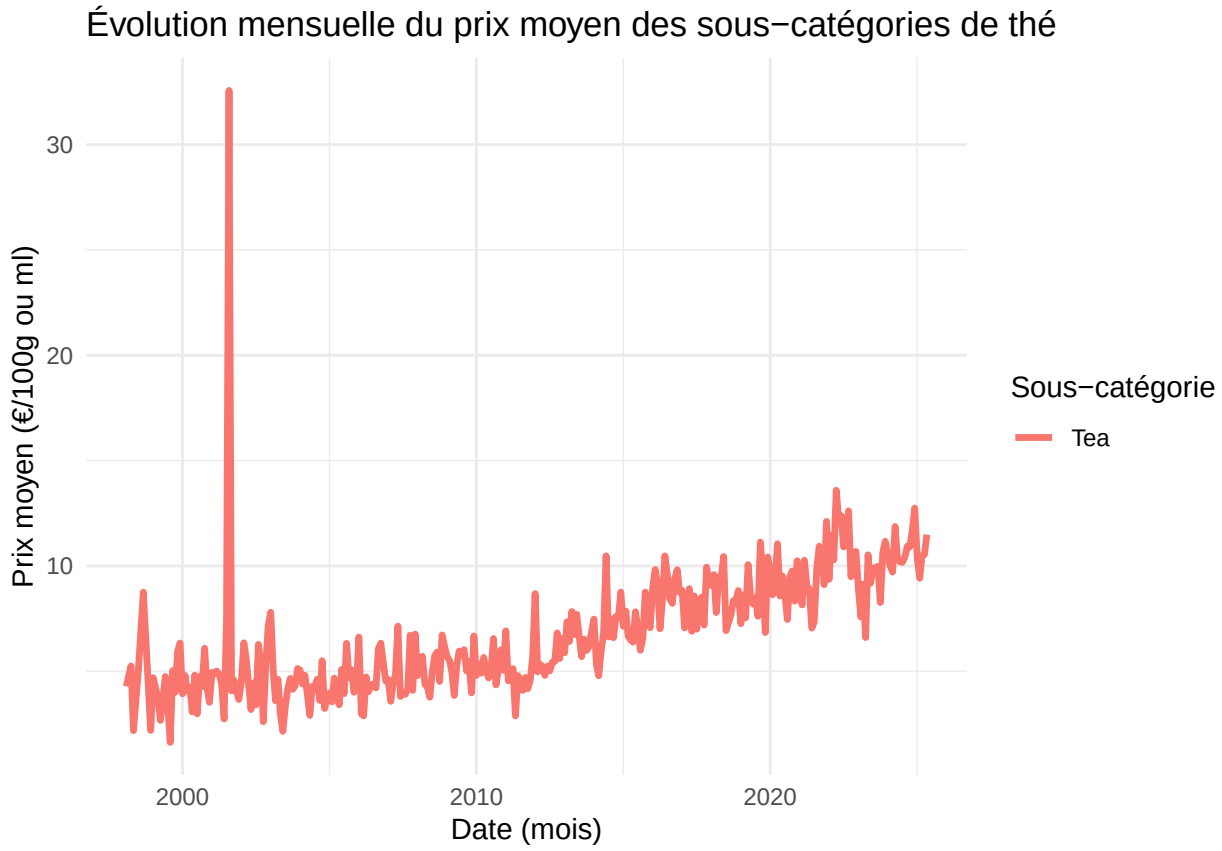


Evolution du prix en fonction des sous categories “Tea” et “RTD Tea”

```r
# Filtrage des deux sous-catégories de thé
tea_data <- gnpd %>%
  filter(
    `Sub-Category` %in% c("Tea"),
    !is.na(price_per_100),
    !is.na(month)
  )

# Calcul du prix moyen par mois et sous-catégorie
evol_prix_tea <- tea_data %>%
  group_by(month, `Sub-Category`) %>%
  summarise(mean_price = mean(price_per_100, na.rm = TRUE), .groups = "drop")

# Visualisation
ggplot(evol_prix_tea, aes(x = month, y = mean_price, color = `Sub-Category`)) +
  geom_line(size = 1.2) +
  labs(
    title = "Évolution mensuelle du prix moyen des sous-catégories de thé",
    x = "Date (mois)",
    y = "Prix moyen (€/100g ou ml)",
    color = "Sous-catégorie"
```

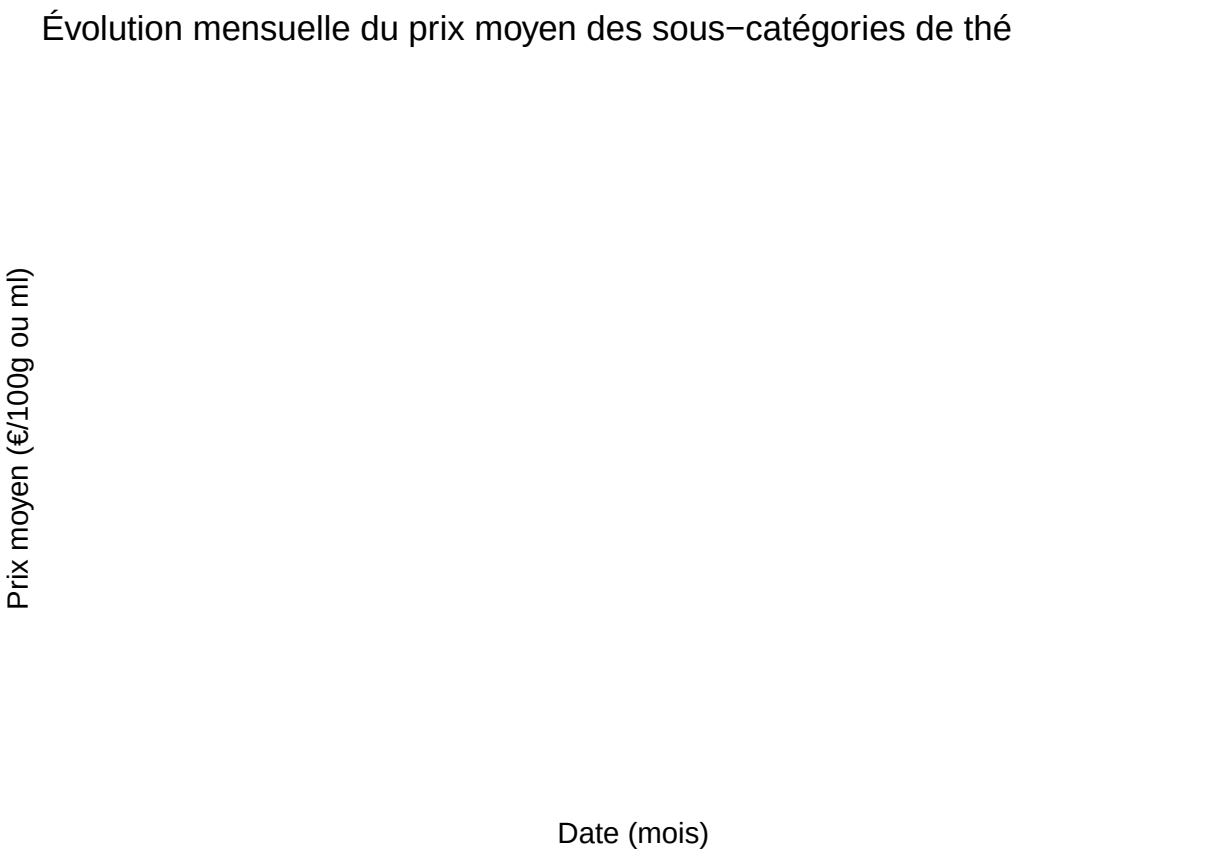
```
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
# Filtrage des deux sous-catégories de thé
tea_data <- gnpd %>%
  filter(
    `Sub-Category` %in% c("RTD Tea"),
    !is.na(price_per_100),
    !is.na(month)
  )

# Calcul du prix moyen par mois et sous-catégorie
evol_prix_tea <- tea_data %>%
  group_by(month, `Sub-Category`) %>%
  summarise(mean_price = mean(price_per_100, na.rm = TRUE), .groups = "drop")

# Visualisation
ggplot(evol_prix_tea, aes(x = month, y = mean_price, color = `Sub-Category`)) +
```

```
  geom_line(size = 1.2) +
  labs(
    title = "Évolution mensuelle du prix moyen des sous-catégories de thé",
    x = "Date (mois)",
    y = "Prix moyen (€/100g ou ml)",
    color = "Sous-catégorie"
  ) +
  theme_minimal()
```



Conclusion

Cette analyse nous a permis d'explorer en profondeur la base de données MINTEL GNPD concernant les produits de thé. Nous avons pu identifier :

1. Les tendances principales dans les sous-catégories de produits
2. L'importance des claims marketing et leur impact sur les prix
3. Les préférences en matière d'emballage
4. L'évolution des prix dans le temps
5. L'influence des claims nutritionnels sur la valorisation des produits

Ces insights pourront servir de base pour formuler des recommandations stratégiques pour la valorisation durable du thé en France.