

# Rogue One

GIROD Mathis, JAULIN Maxence, PRODHON Louis, WANG Zezhong



## Introduction

Ce projet a été réalisé dans le cadre du cours **Visualisation de données**, au cours du semestre de **printemps 2024**, à l'**Université de Technologie de Troyes**.

Pour cette étude, nous avons choisi d'analyser des données originales qui nous permettent de nous interroger sur **l'étude du transport ferroviaire en France**. Notre analyse portera sur des jeux de données extraits du site de données de la SNCF (Société Nationale des Chemins de fer Français) Data SNCF. L'ensemble des données qui vont donc être traitées dans ce projet proviennent donc toutes de cette source. Nous n'avons donc pas utilisé de jeux de données extérieurs à ce site.

Les données récoltées sur le transport sont assez importantes c'est pourquoi nous avons choisi de nous concentrer sur une découverte avec un spectre assez large, allant des voyageurs aux objets perdus. Nous utiliserons les données des gares, des voyageurs et des objets perdus/retrouvés. Cette étude permettra de déterminer et de comprendre des tendances clés associées au trafic ferroviaire sur des périodes allant de 2017 à 2022.

L'objectif de ce projet est de fournir des interprétations basées sur les visualisations issues d'une analyse exploratoire de nos jeux de données (7 jeux de données).

## Données

Nous avons donc choisi d'étudier sept jeux de données (7) issues du site Data SNCF. Ce sont des données collectées par la SNCF parmi les différentes catégories disponibles sur le site (voir ci-dessous).

## Catégories de données

[Voir toutes les données >](#)

 <b>Services voyageurs</b> (10)	 <b>Description du réseau</b> (20)	 <b>Gares</b> (28)
Accédez à l'ensemble des données concernant les services aux voyageurs (données de régularité, horaires, etc.)	Retrouvez ici des données sur l'état du réseau ferroviaire, des informations sur les voies, etc.	Vous trouverez ici des jeux de données sur les gares, leur fréquentation, leurs équipements, etc.

 <b>Rapports</b> (14)	 <b>Comptage et flux</b> (11)	 <b>Sécurité ferroviaire</b> (7)
Consultez un ensemble de rapports SNCF au format PDF (sécurité, audits, RSE, etc.)	Parcourez les données de fréquentation des trains, d'enquêtes voyageurs, etc.	Explorez les données de sécurité, telles que les événements de sécurité remarquables, les audits de sécurité, etc.

Ces données concernent des objets possédés par la SNCF (gares, objets) mais aussi des enquêtes réalisées sur des individus anonymement (fréquentation, voyageurs). Les données sont liées à une période temporelle précise de **2017 à 2022**.

L'ensemble des données brutes sont accessibles depuis le dossier `/data`.

### Nombre d'observations

Le nombre d'observations varie selon chaque jeu de données. Pour plus de détail, nous avons détaillé précisément le nombre d'observations dont nous disposons.

—	Nom du dataset	Nombre d'observations	Lien	Description
01	dataset1-gares-de-voyageurs.csv	2.862	Dataset1	Jeu de données sur les gares de voyageurs
02	dataset2-fréquentation-gares.csv	21.147	Dataset2	Jeu de données sur la fréquentation des gares
03	dataset3-motif-deplacement.csv	284	Dataset3	Jeu de données sur les motifs de déplacement
04	dataset4-enquetes-gares-connexions-repartition-par-repartition-par-categories-socio-profe.csv	697	Dataset4	Jeu de données sur les CSP des voyageurs
05	dataset5-enquetes-gares-connexions-repartition-repartition-par-classe-dage.csv	375	Dataset5	Jeu de données sur l'âge des voyageurs
06	dataset6-objets-trouves-gares.csv	1.844.912	Dataset6	Jeu de données sur les objets trouvés en gare

	Nom du dataset	Nombre d'observations	Lien	Description
07	dataset7-objets-trouves-restitution.csv	858.180	Dataset7	Jeu de données sur les objets restitués

Au sein de ces données nous constatons que toutes s'orchestrent autour d'une donnée principale (Gare, 01) qui est présent dans tous les datasets. Nous pouvons donc segmenter les données restantes par des critères géographiques (02,03,04,05), des critères temporels (06,07), des critères voyageurs (08,09,10,11,12,13,14) et des critères sur les objets perdus/trouvés (15,16,17).

## Variables

Nous avons décidé d'utiliser **17 variables** pour notre projet provenant des jeux de données bruts ou alors d'attributs créées par nos soins.

	Nom de la variable	Type	Format	Dataset (Origine)	Description
01	gare	Nominale	String	1,2,3,4,5,6,7	Nom de la gare
02	departement	Ordinal	NN	1	Numéro du département
03	zone	Nominale	{A,B,C}	1	Lettre correspondant à la zone géographique
04	latitude	Continu	M°S'NS	1	Latitude de l'objet gare
05	longitude	Continu	M°S'NS	1	Longitude de l'objet gare
06	annee	Ordinal	YYYY	2,3,4	Année correspondante
07	timing_reception	Discrète	YYYY-MM-DD-HH-MM-SS	6,7	Réception de l'objet perdu
08	nb_voyageurs	Discrète	Integer	2	Nombre de voyageurs
09	age	Ordinal	String	5	Age d'un voyageur
10	pourcentage_age	Continu	%	5	Pourcentage sur l'âge des voyageurs
11	csp	Nominale	String	4	Catégorie socio-professionnel d'un voyageur
12	pourcentage_csp	Continu	%	4	Pourcentage sur la catégorie socio-professionnel des voyageurs
13	motif_deplacement	Nominale	String	3	Motif de déplacement d'un voyageur

	Nom de la variable	Type	Format	Dataset (Origine)	Description
14	pourcentage_deplacement	Continue	%	3	Pourcentage sur le motif de déplacement des voyageurs
15	nature_objet	Nominale	String	6,7	Nature de l'objet
16	categorie_objet	Nominale	String	6,7	Catégorie de l'objet
17	code_uic	Nominale	NNNNNNNNNN	6,7	Code UIC de la gare

## Variables particulières

Notre jeu de données comprenant des coordonnées spatiales, nous avons estimé qu'il était intéressant de réaliser des cartes. En effet, les coordonnées géographiques de longitude et latitude pourront être utilisée pour catégoriser le réseau des gares françaises.

L'ensemble des données énoncées plus en haut nous paraissent pertinentes dans le cadre d'une étude. En effet, elles permettent :

- d'étudier les effets de la fréquentation sur les vols/pertes d'objets
- d'effectuer une analyse temporelle et spatiale du réseau
- d'effectuer des classements et des comparaisons entre les différentes régions et/ou départements (analyse multiscalaire). Exemple : espace moins déservi par exemple.

## Plan d'analyse

1. **Découverte du jeu de données** et surtout comprendre à quoi servent nos données. Par exemple : nous souhaitons réaliser des visualisations sur le réseau ferroviaire actuel, étudier la répartition générale des voyageurs... > A quoi ressemble le réseau SNCF en France ? Quels sont les départements les mieux équipés ? A quel point Paris a une place importante dans le réseau des autres territoires ?
2. **Analyse des voyageurs** : De façon plus précise, nous étudierons les voyageurs qui utilisent quotidiennement les réseaux ferrés français. Cela passera notamment par des attributs d'âge, de CSP ou encore de motif de déplacement. > Le nombre de voyageurs est-il bien repartis entre les gares d'un même département ? Quel est le voyageur moyen de la SNCF ? Comment ce voyageur diffère en fonction des gares ? Quel est la relation entre les motifs de voyage des passagers et leur répartition par âge et par profession ?
3. **Analyse des objets** : De la même façon, nous souhaiterions étudier les objets perdus en gares. Pour cela, nous utiliserons également un second jeu de données sur les objets retrouvés. > Y-a-t-il plus de chances de perdre un objet selon la gare ? Doit-on s'attendre à un afflux d'objets perdus plus important dans les mois de Juillet-Août 2024 plus important que les dernières années ? Quelles sont les chances de retrouver un objet perdu ? Quelles sont les chances de retrouver un objet en fonction de sa nature ?
4. **Analyse spatiale** : A l'aide de nos données spatiales, nous souhaitons réaliser des cartes. Ces dernières permettront visuellement de voir la disposition et la répartition des gares en France Métropolitaine.
5. Enfin si nous souhaitons **rajouter des questions**, nous nous laissons la liberté de les rajouter au plan d'analyse.

```
knitr:::opts_chunk$set(
  echo = FALSE,
  message = FALSE,
  warning = FALSE
```

```

)
library(ggplot2)
library(dplyr)

##
## Attachement du package : 'dplyr'

## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag

## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)
library(tibble)
library(readr)
library(lubridate)

##
## Attachement du package : 'lubridate'

## Les objets suivants sont masqués depuis 'package:base':
##
##     date, intersect, setdiff, union

library(forcats)
library(stringr)
library(sf)

## Linking to GEOS 3.10.2, GDAL 3.4.1, PROJ 8.2.1; sf_use_s2() is TRUE
library(rnaturalearth)
library(rnaturalearthdata)

##
## Attachement du package : 'rnaturalearthdata'

## L'objet suivant est masqué depuis 'package:rnaturalearth':
##
##     countries110

```

---

## Nettoyage des données

Le **nettoyage des données** est la première étapes de notre projet. C'est ici que nous allons créer de nouvelles tables, modifier les jeux de données existants (supprimer ou renommer les colonnes existantes).

Afin de travailler proprement, nous avons réaliser les étapes suivantes :

1. Nous avons commencé par **observer les colonnes** de nos jeux de données. Nous avons pu isoler lesquels étaient susceptibles d'être utilisées. Nous les avons ensuite **importées**.
2. Ensuite, nous avons choisi de **renommer les colonnes** de nos jeux de données selon une norme précise (voir **Convention de nommage des colonnes** ci-dessous). Les colonnes de base des tables utilisaient des espaces, ce qui est incompatible avec l'appel de ces dernières.
3. Une fois les tables modifiées, nous avons du **filtrer nos données**.

**Convention de nommage des colonnes** - Transformer les espaces en underscore. - Nommer les variables en un seul mot si possible.

Exemple : Code Postal en code\_postal

## Importation des jeux de données

Les données sont importées dans R dans leur forme brute. On les manipule ensuite afin de créer nos propres tableaux.

La section suivante concerne l'**Exploration de ces données**. On y détaillera les 3 grands axes que sont la Découverte, les Voyageurs et les Objets.

---

## Exploration : Découverte

Dans cette partie, nous découvrons le jeu de données **Fréquentation**.

### Visualisations réalisées

1. Fréquentation des gares
2. Positionnement des gares en France Métropolitaine
3. Fréquentation globale par département
4. Fréquentation des gares d'un même département (77 et 69)

L'ensemble des visualisations sont réalisées avec les données de l'année **2022**, car ce sont les données les plus récentes à notre disposition.

*À quoi ressemble le réseau SNCF en France ? Quels sont les départements les mieux équipés ? À quel point Paris a une place importante dans le réseau des autres territoires ? Le nombre de voyageurs est-il bien reparti entre les gares d'un même département ?*

### 1. Fréquentation des gares (2022)

Tout d'abord, voici un petit tour d'horizon du dataset **Fréquentation**.

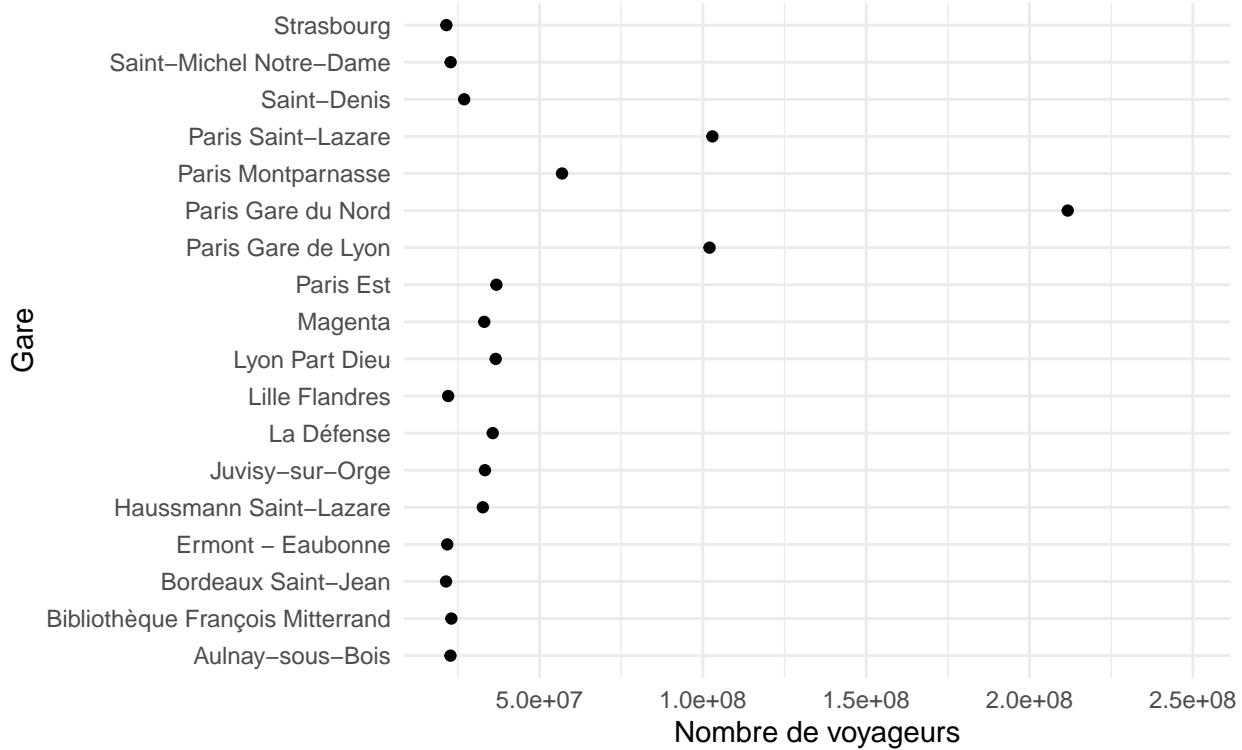
```
## # A tibble: 6 x 8
##   Gare      UIC    Code_Postal DRG   Année Personnes Voyageurs Département
##   <chr>     <chr>   <chr>       <chr> <dbl>    <dbl>    <dbl>   <chr>
## 1 Abbaretz 481614 44170        C     2022     40825    40825  44
## 2 Abbaretz 481614 44170        C     2021     27466    27466  44
## 3 Abbaretz 481614 44170        C     2020     22773    22773  44
## 4 Abbaretz 481614 44170        C     2019     38473    38473  44
## 5 Abbaretz 481614 44170        C     2018     38027    38027  44
## 6 Abbaretz 481614 44170        C     2017     35637    35637  44
```

Pour notre première visualisation, nous avons choisi d'utiliser des données discrètes (nombre de voyageurs) en abscisse et nominales (gares) en ordonnée. Nous avons réalisé une comparaison grâce à un bar chart pour étudier les différences de fréquentation entre les gares les plus utilisées (plus pertinent selon nous).

Afin d'obtenir un classement des gares les plus fréquentées, nous avons choisi de filtrer le dataset pour ne garder que les gares au dessus d'un seuil de 20.000.000 individus.

## Fréquentation par gare les plus fréquentées (2022)

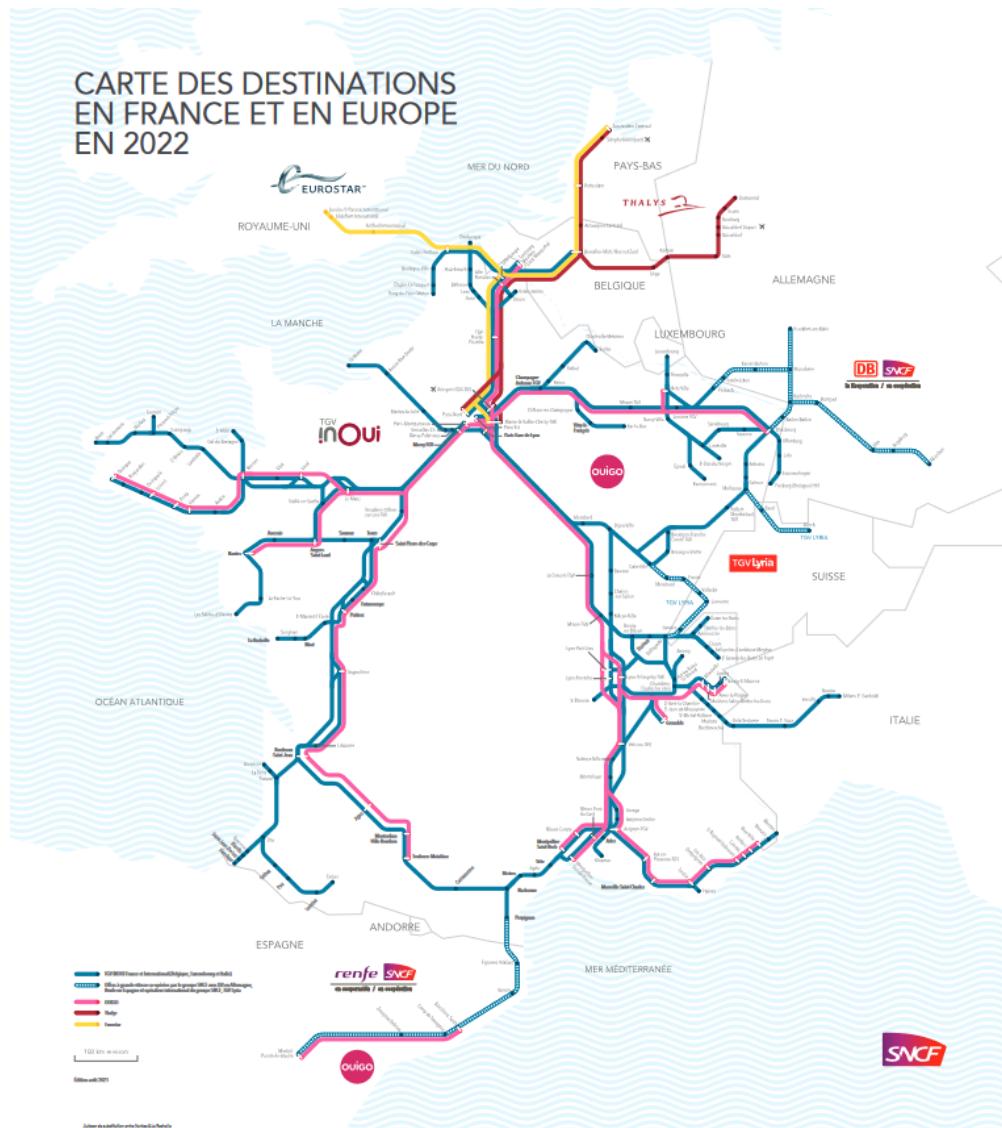
*< minimum 20.000.000 de voyageurs >*



**Analyse.** On remarque que les gares avec le plus de fréquentation sont les gares parisiennes. Au classement, nous retrouvons : Gare du Nord (1), Gare Saint-Lazare (2), Gare de Lyon (3), Gare Montparnasse (4). D'autres gares se démarquent mais restent sensiblement proches les unes des autres.

Cette visualisation n'est pas suprenante si l'on utilise régulièrement le réseau RATP et SNCF en région parisienne. En effet, la grande majorité des trajets partent de Paris et arrivent sur Paris.

Gare du Nord semble être la gare la plus fréquentée du réseau. En faisant des recherches, on apprend qu'elle permet des départs vers le Royaume-Uni (Londres-St-Pancras), la Belgique (Bruxelles-Midi) ou encore les Pays-Bas (Amsterdam-Centraal) (*Figure 1*). La population est généralement importante dans ces grandes villes et capitales européennes, ce qui explique également la fréquentation de Gare du Nord que ce soit pour des trajets professionnels ou touristiques.



**Figure 1.** Carte des destinations en France et en Europe (2022). **Source :** <https://www.sncf-connect.com/aide/le-reseau-sncf-en-france-et-en-europe>

## 2. Positionnement des gares en France Métropolitaine (2022)

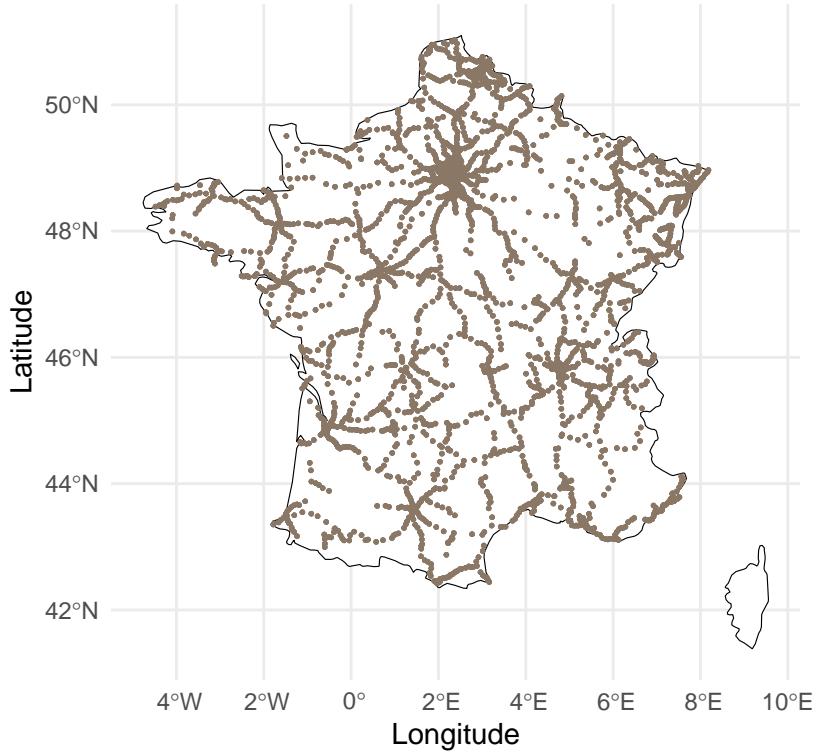
Pour poursuivre notre découverte du réseau ferroviaire, nous avons choisi de représenter la fréquentation des gares sur une carte de France Métropolitaine. Avec cette visualisation spatiale, on peut comprendre plus facilement les enjeux liés aux flux de voyageurs.

Nous étudierons ici : À quoi ressemble le réseau SNCF en France ?

On s'attend à avoir un réseau en étoile vers Paris.

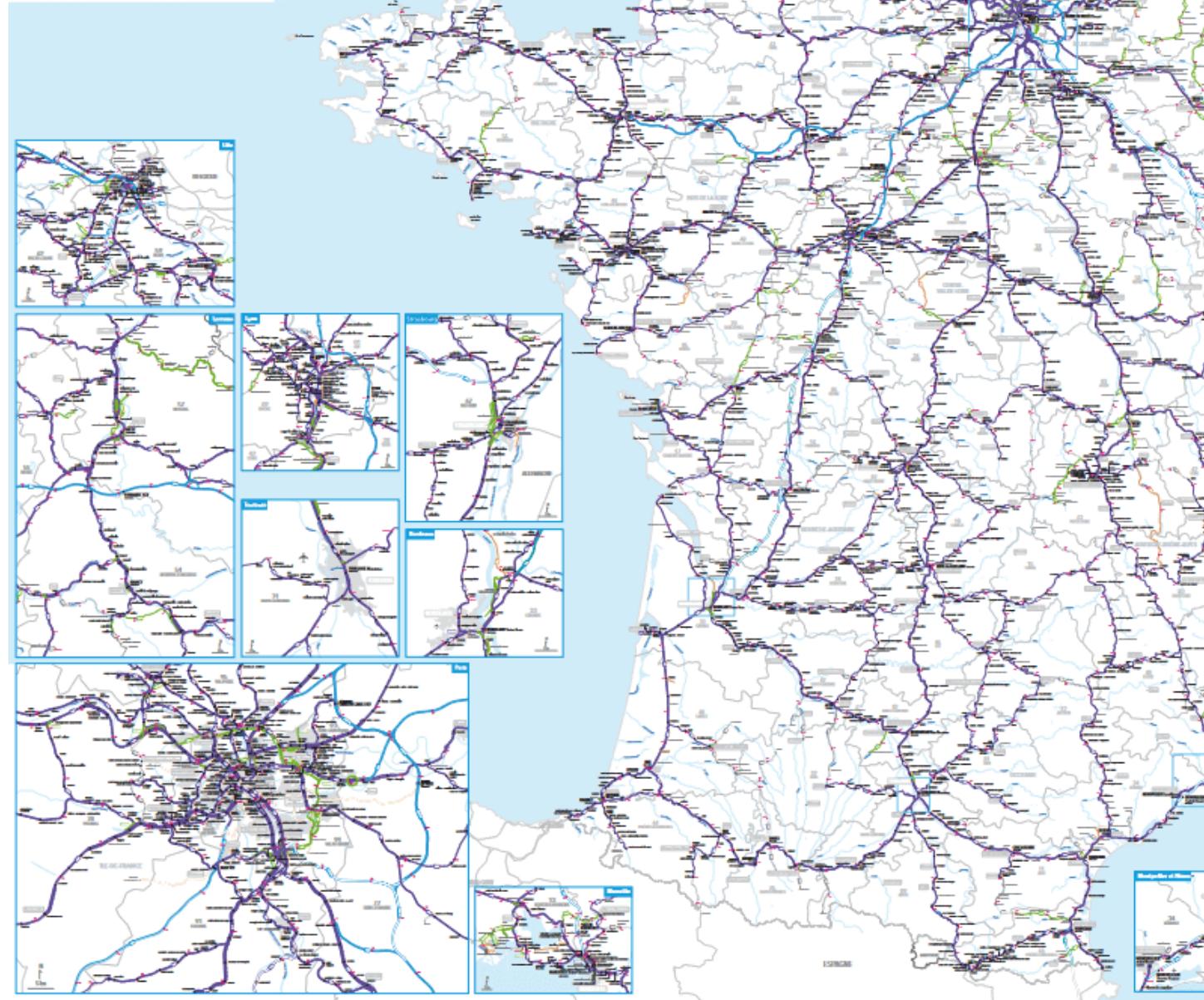
Pour représenter cela, on utilise des données discrètes (longitude, latitude, nombre de voyageurs) et ordinaires (année) sur une carte. Etant donné qu'il s'agit d'une carte, on place la longitude en abscisse et la latitude en ordonnée.

## Disposition des gares (2022) <Vue Spatiale>



**Analyse.** Au premier abord, on remarque que les points, représentant les gares, tracent d'eux-même des lignes sur la carte. Ces dernières représentent les lignes du réseau ferroviaire français, comme on peut le vérifier sur la *Figure 2*. Cette visualisation est donc toujours assez proche de la réalité en 2022.

Cette visualisation complète notre première analyse : Paris est le centre du réseau ferroviaire français, “*tout passe par Paris*”. Cette règle s’applique également avec le réseau autoroutier français.



2. Carte du réseau SNCF en France (2020). Source : <https://www.sncf-connect.com/aide/le-reseau-sncf-en-france-et-en-europe>

Réponse à la question : Le réseau ferroviaire de la SNCF est bien en étoile

comme le confirme la visualisation obtenue.

### 3. Fréquentation globale par département (2022)

Réalisons maintenant une analyse multiscalaire afin d'avoir une meilleure vue d'ensemble et compréhension de la fréquentation de ce réseau.

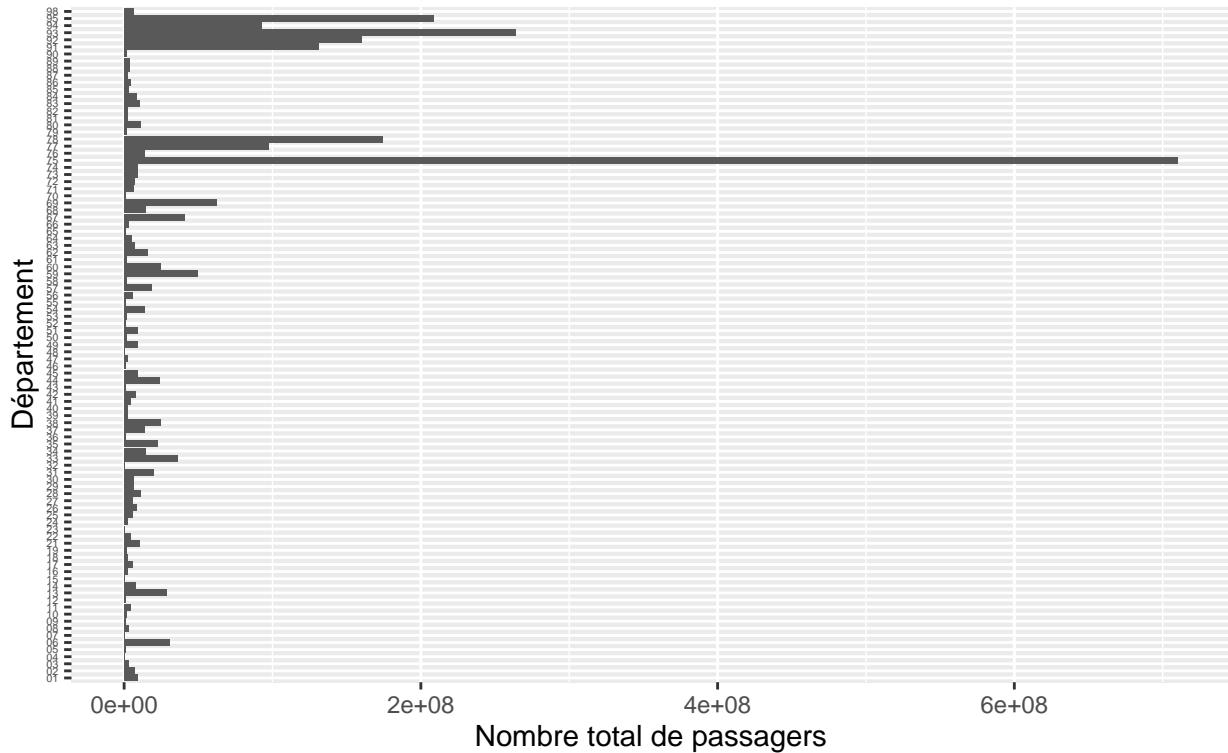
Nous verrons donc ici : Quels sont les départements les mieux équipés ?

On s'attend à retrouver Paris en tête de tous les départements, suivi des départements d'Ile-de-France (autour de Paris).

Tout d'abord, on réalise une première vue globale pour faire ressortir les départements les plus fréquentés du réseau. On utilise pour cela des données discrètes en abscisse (Total\_passagers) et ordinaire en ordonnée (Département). Une visualisation avec un bar chart nous permet de faire une comparaison entre les départements. Nous afficherons les départements selon leur numéro de département, classés dans ce même ordre.

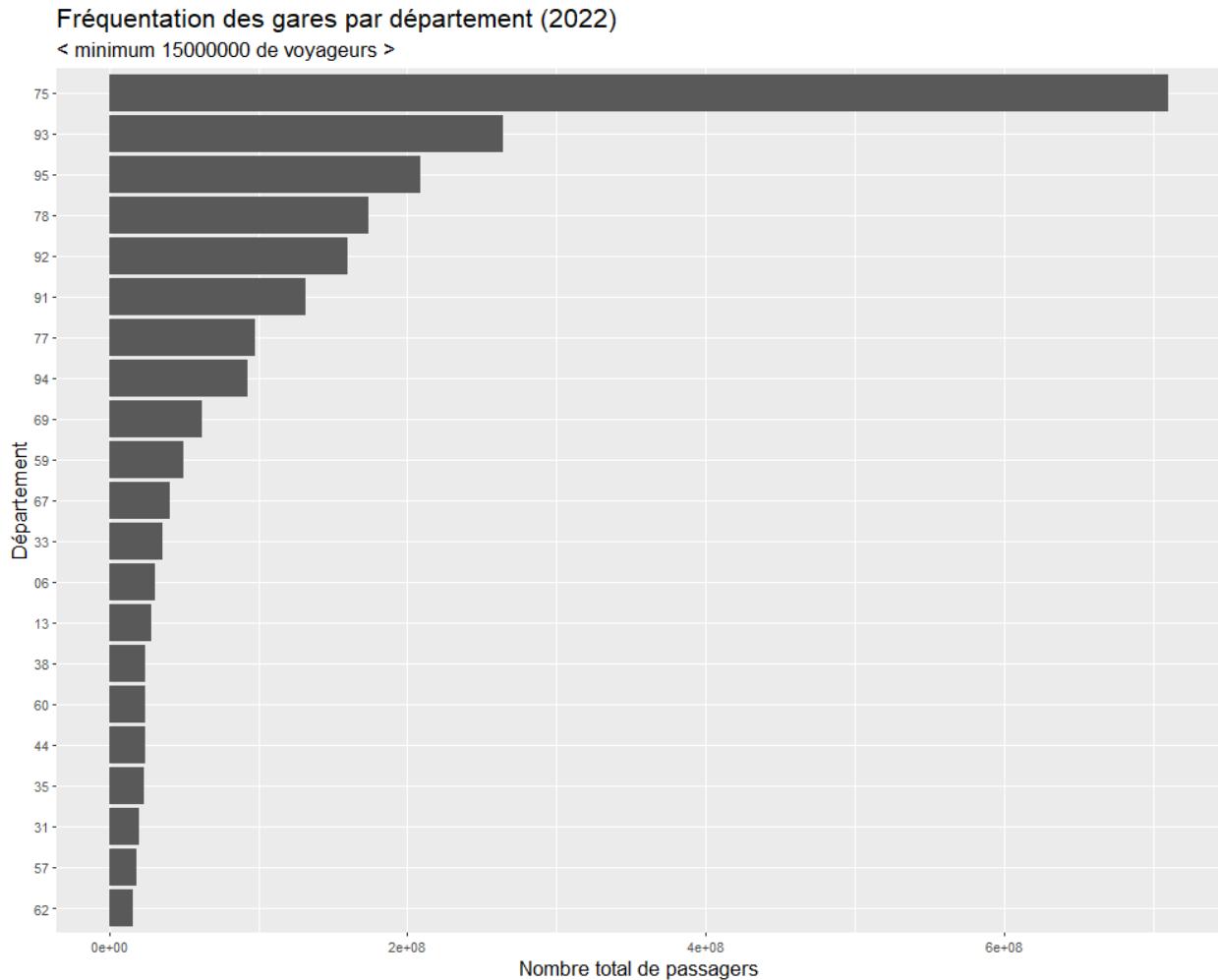
#### Fréquentation des gares par département (2022)

< Vue globale avec classement ordinal >



**Analyse.** On voit avec cette visualisation qu'il y a une très grande différence entre certains départements. Cette différence est sûrement expliquée en partie par le phénomène d'urbanisation et d'étalement urbain (regroupement autour de Paris et des grandes villes). Ici, on a ordonné les départements selon leur numéro afin de mieux comprendre la visualisation qui se veut globale. Bien que la lecture des départements soit un peu difficile, elle permet de comprendre des distinctions. On remarque assez facilement que les départements autour de 75 et 93 sont très fréquentés.

Faisons un zoom sur ces départements et organisons les du plus fréquentés au moins fréquentés en fixant une limite minimum de 15 millions de voyageurs au total sur l'année 2022. Voici donc ci-dessous une capture générée par Shiny App qui permet de faire varier les paramètres : année et nombre minimum de voyageurs dans les gares.



**Figure 3.** Capture Shiny App : Fréquentation des gares par département (2022), minimum 15.000.000 de voyageurs.

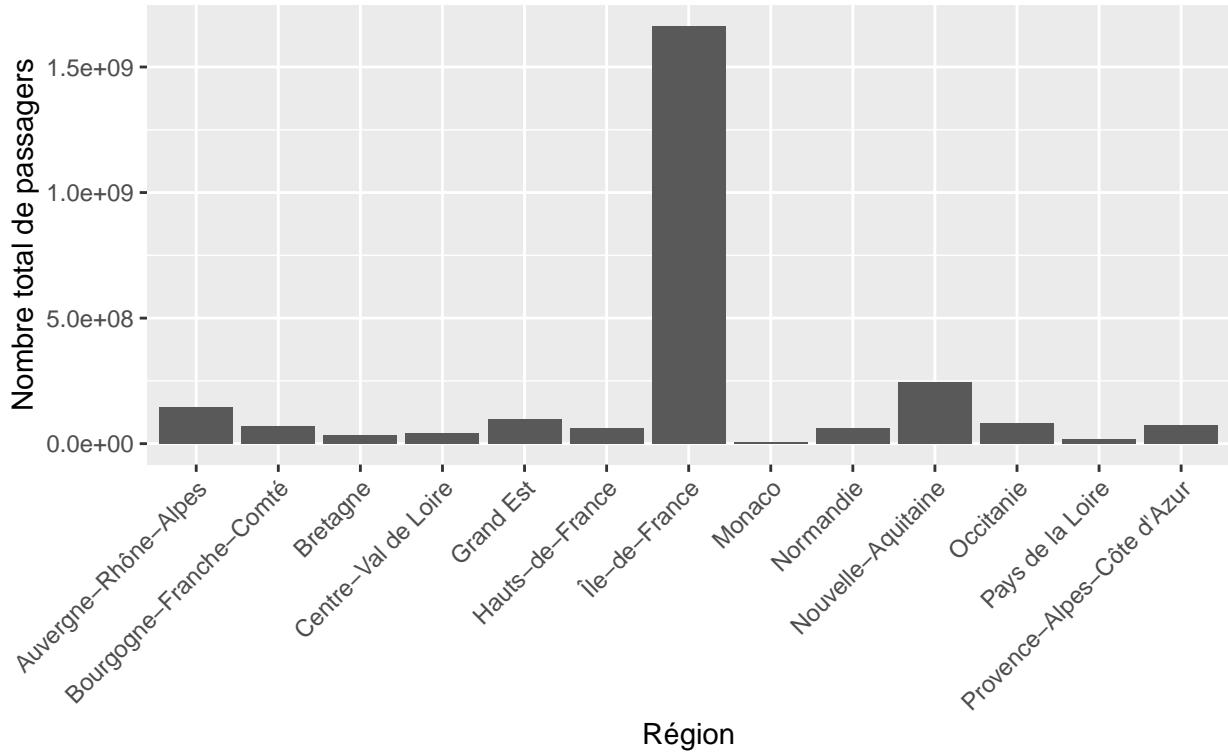
**Analyse.** On constate que Paris est largement en tête de la fréquentation (département 75), suivi des départements de la région Ile-de-France (dans l'ordre : 93, 95, 78, 92, 91, 77, 94). Le Rhône, région lyonnaise, (69) est le département le plus fréquentée après la région Ile-de-France. Le Rhône (Lyon) semble donc avoir une importance dans le réseau ferroviaire français. Enfin, on aurait pu également appliquer nos précédentes observations aux départements les moins fréquentés.

Réponse à la question : Les départements les mieux équipés sont donc les départements d'Ile-de-France.

On affiche donc maintenant un regroupement de la fréquentation des gares des départements selon les régions françaises afin de confirmer nos observations précédentes.

## Fréquentation des gares par région (2022)

< Vue globale >



**Analyse.** Comme le constat fait un peu plus haut, on observe que la région Ile de France (composée des départements 75, 77, 78, 91, 92, 93, 94, 95) concentre la plupart des voyageurs. Ce constat est tout à fait correct étant donné que la région concentre environ 12 millions de personnes à l'année. On peut également souligner que le réseau SNCF désert la gare de Monaco-Monte-Carlo qui n'est pas comptabilisé dans les régions françaises. Nous avons donc dû l'intégrer manuellement.

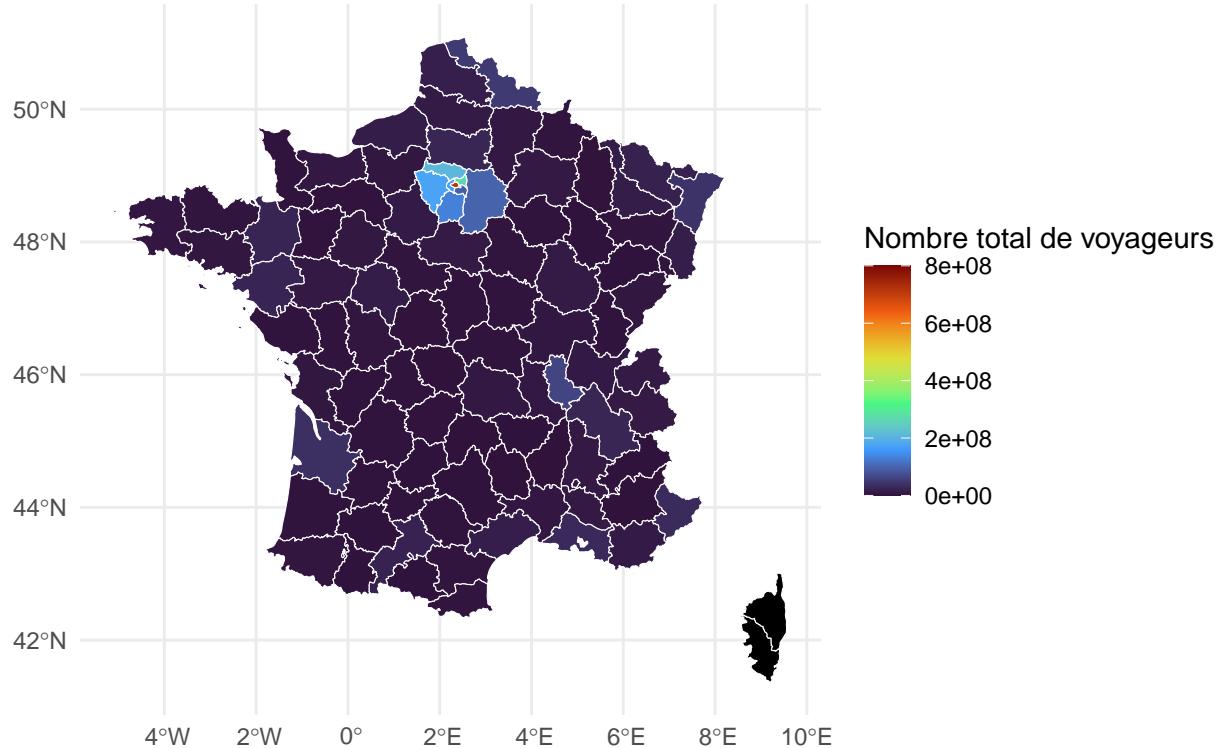
Nous répondrons ici à la question : À quel point Paris a une place importante dans le réseau des autres territoires ?

On s'attend à ce que Paris surpassé les autres territoires français, au vue des éléments que nous avons pu trouver au-dessus.

Nous réalisons ensuite une visualisation spatiale de ces départements pour observer plus précisément les différences entre chacun. On utilisera donc des couleurs avec un gradient afin de mieux constater les distinctions entre les départements.

## Fréquentation des gares en France par département (2022)

< Vue spatiale globale >



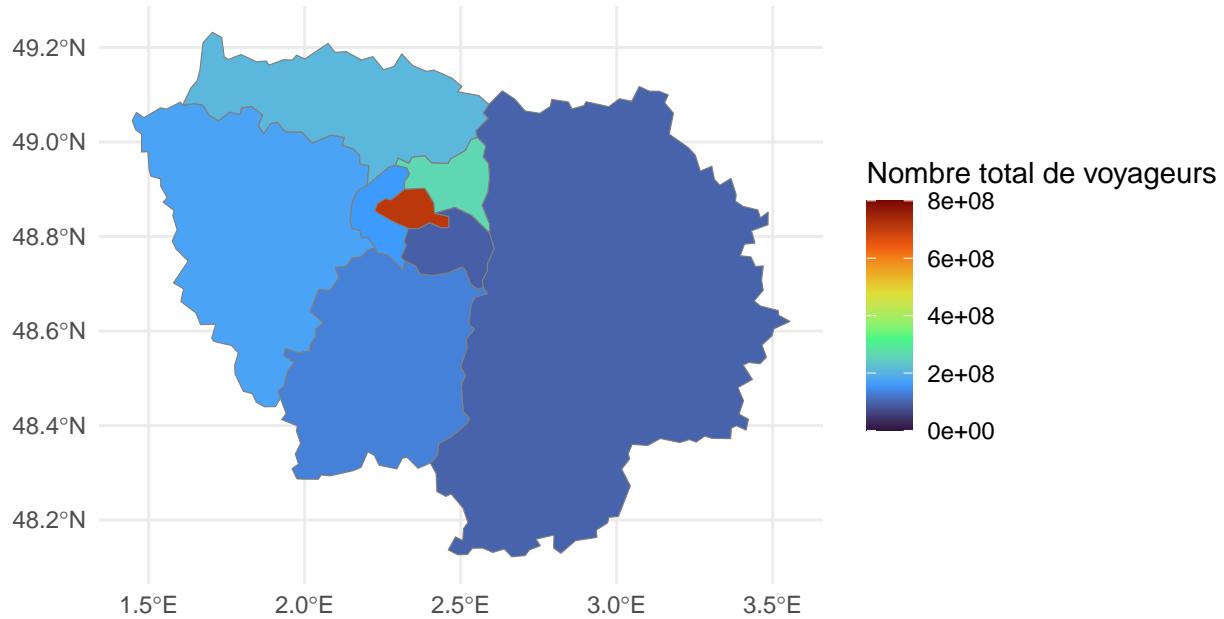
**Analyse.** Les différences sont flagrantes dans le cas de la région Ile-de-France. L'échelle de gradient de ce graphique montre la grande inégalité du réseau : Paris a une fréquentation démesurée si bien que les autres départements ne restent que dans les couleurs froides. Cependant, on remarque que la fréquentation des gares des départements s'articule autour de quatre principaux espaces : l'espace Parisien (Paris), l'espace Nord (Lille), l'espace Est (Strasbourg), l'espace Lyonnais (Lyon). Comme nous le pensions, la fréquentation des gares est plus importante autour des grandes villes. De plus, on pourrait le vérifier mais ces résultats donnent une idée approximative de la répartition de la population en France.

Avec cette profondeur supplémentaire, cela nous permet de formuler de nouvelles hypothèses. On pourrait se demander si la population en périphérie des grandes villes fréquente généralement les gares du réseau pour des motifs professionnels, du tourisme ou simplement pour la vie quotidienne. De la même façon, est-ce que la fréquentation des gares est due à des trajets entre régions et départements ?

Pour finir cette section, on propose de faire un zoom sur la région parisienne.

## Fréquentation des gares d'Ile de France par département (2022)

< Vue spatiale avec zoom sur IDF >



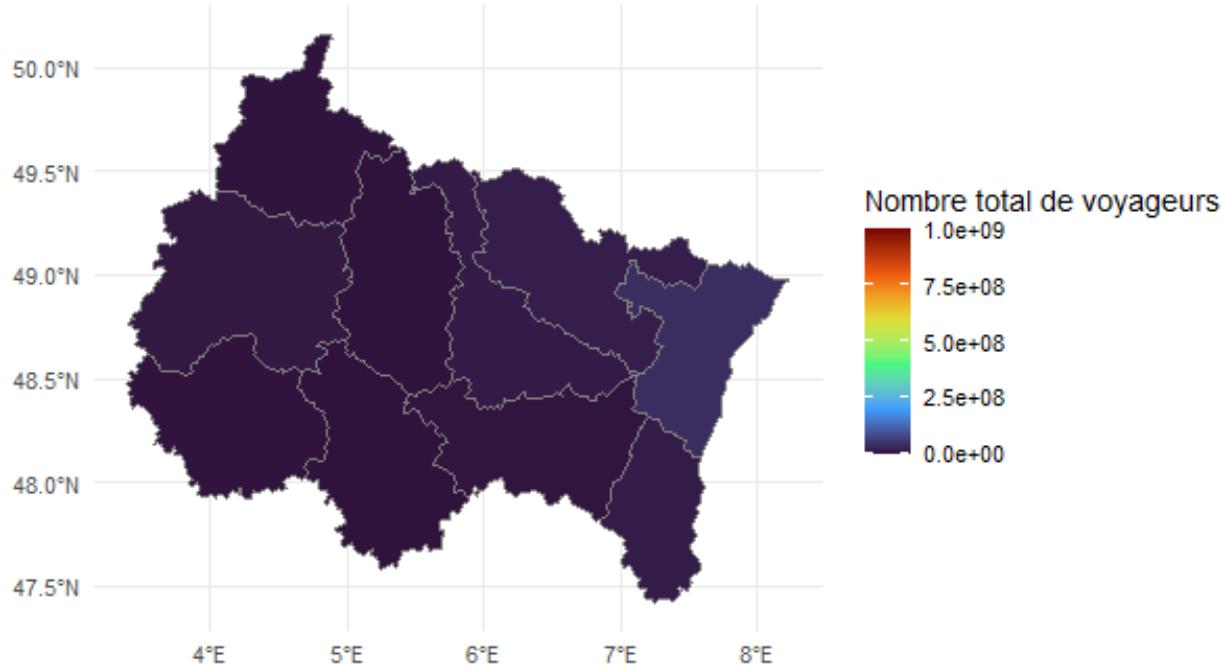
**Analyse.** On fait les mêmes conclusions que celle de la visualisation précédente. Paris est essentiel dans le réseau ferroviaire français. La partie Sud-Ouest et Ouest de l'Ile-de-France semble moins fréquentée. Afin d'avoir une meilleure vision de cela, on peut explorer ce département au travers de nouvelles visualisations.

Réponse à la question : Paris a donc bien une place importante dans le réseau ferroviaire français.

Est-ce qu'un réseau en étoile vers la plus grande ville de la région s'applique également pour les autres régions ? C'est ce que nous avons voulu voir dans le cas de la région Grand-Est. La capture ci-dessous a été générée dynamiquement avec Shiny App.

## Fréquentation des gares de Grand Est (2022)

< Vue spatiale avec zoom sur Grand Est >



**Figure 4.** Capture Shiny App : Fréquentation des gares du Grand-Est (2022).

**Analyse.** On remarque ici qu'il y a plus de fréquentation dans les gares du département de Strasbourg (67). Bien qu'il puisse y avoir des disparités à l'intérieur de son département, Strasbourg semble bien être le centre de sa région : le Grand-Est.

### 4. Fréquentation des gares d'un même département (77 et 69) (2022)

Après avoir étudié l'échelle nationale et régionale, on peut s'intéresser à la fréquentation des gares au sein d'un même département. Il serait intéressant de réaliser une première visualisation permettant de classer les gares de ce département (77). Cela permettrait de faire ressortir les plus utilisées. Ensuite, une visualisation spatiale permettrait de les placer spatialement au sein de ce même département et de voir leur importance.

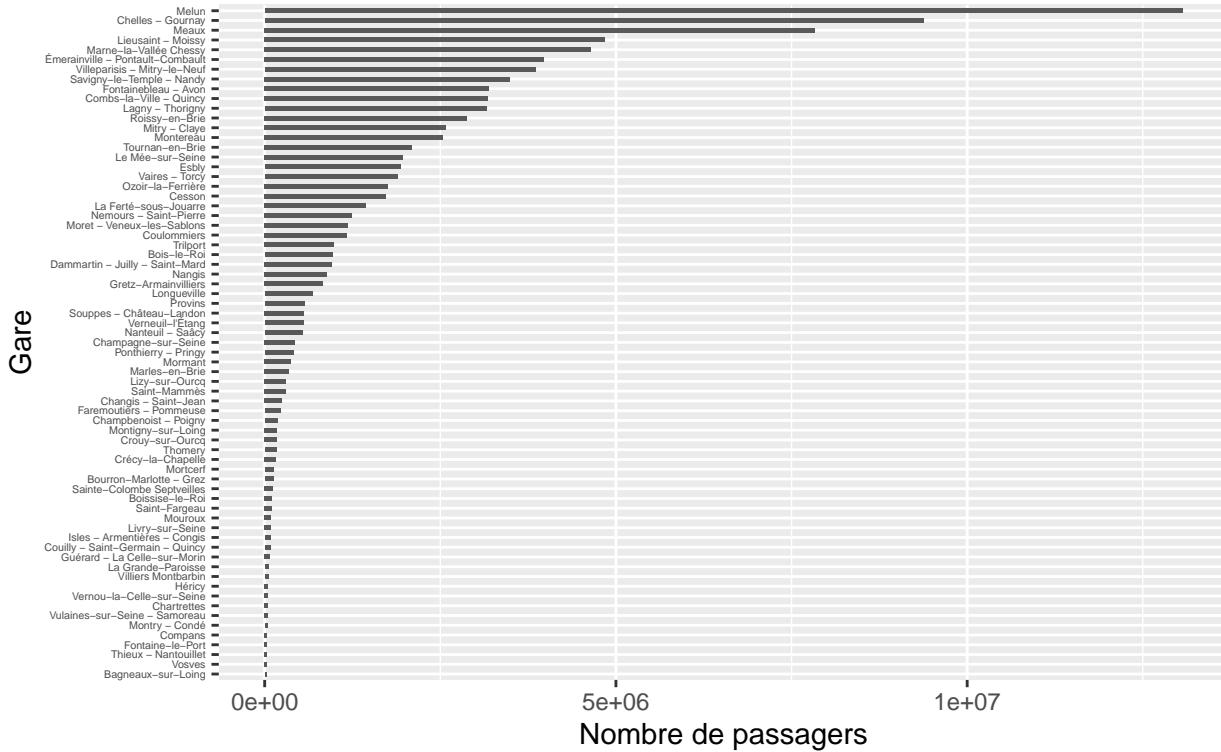
Enfin, nous étudierons la question : Le nombre de voyageurs est-il bien reparti entre les gares d'un même département ?

On s'attend à avoir quelques disparités évidentes entre la fréquentation des gares d'un même département.

On utilise pour la première visualisation des données discrètes (Voyageurs) en abscisse et nominales (Gare) en ordonnée. Pour réaliser cela, on prend appui sur un bar chart puisque l'on souhaite comparer les gares du département.

## Fréquentation des gares du département 77 (2022)

### <Vue Globale>

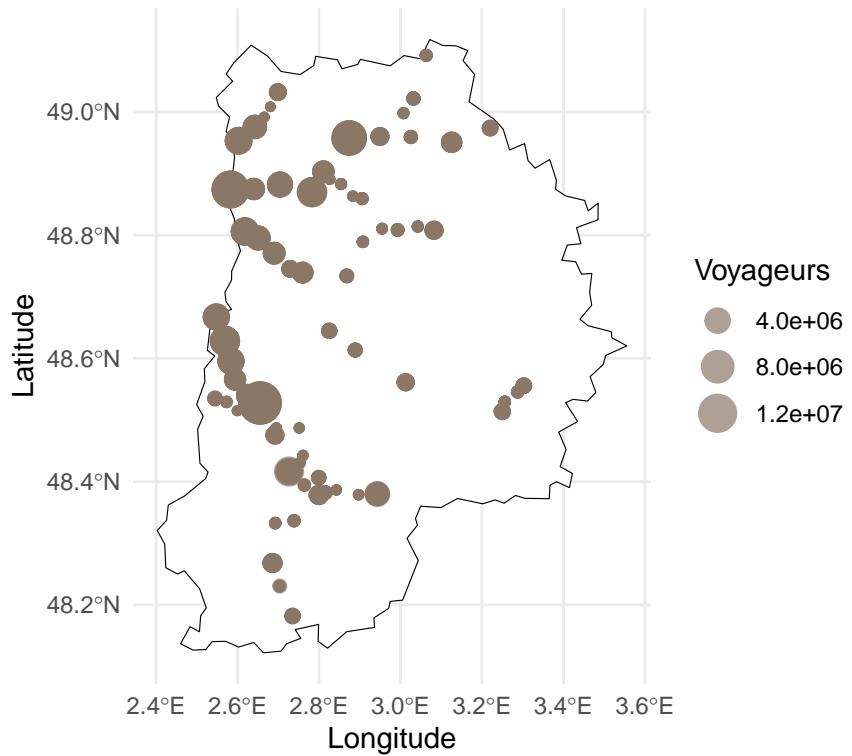


**Analyse.** Au sein du département 77, on remarque plusieurs choses. Tout d'abord, il y a des disparités évidentes entre les gares. Cette disparité peut être liée dans un premier temps à un facteur d'habitants par km<sup>2</sup>. On vérifie avec Melun, gare la plus fréquentée du département, et Nemours-Saint-Pierre, gare beaucoup moins fréquentée. Selon l'INSEE, la densité de Melun est de 5175.2 habitants par km<sup>2</sup> tandis que Nemours-Saint-Pierre est de 250.1 habitants par km<sup>2</sup> (en 2020). Ainsi, la ville de Melun concentre beaucoup plus d'habitants et donc plus de fréquentation dans sa gare.

On peut également expliquer la fréquentation de Melun par le passage du RER D qui ne passe pas dans tout le département. On a aussi la ligne R qui déserte toute la partie basse du département dont Melun. Le fait de disposer d'un accès aux lignes de transport parisiennes rend la gare attractive pour les voyageurs. De nombreuses navettes et bus comme le Seine-et-Marne Express vont jusqu'à la gare de Melun ce qui contribue à augmenter sa fréquentation. Enfin, généralement quand il y a des travaux sur les lignes ferroviaires du département, Melun est la seule gare où les transports permettent de monter sur Paris.

Enfin, pour conclure la première partie de notre analyse, nous pouvons supposer que les gares les plus proches géographiquement de Paris sont les gares les plus fréquentées. En effet, plus une gare semble éloignée de Paris moins elle est fréquentée. Vérifions-le maintenant avec une visualisation spatiale.

## Fréquentation des gares du département 77 (2022) <Vue Spatiale>

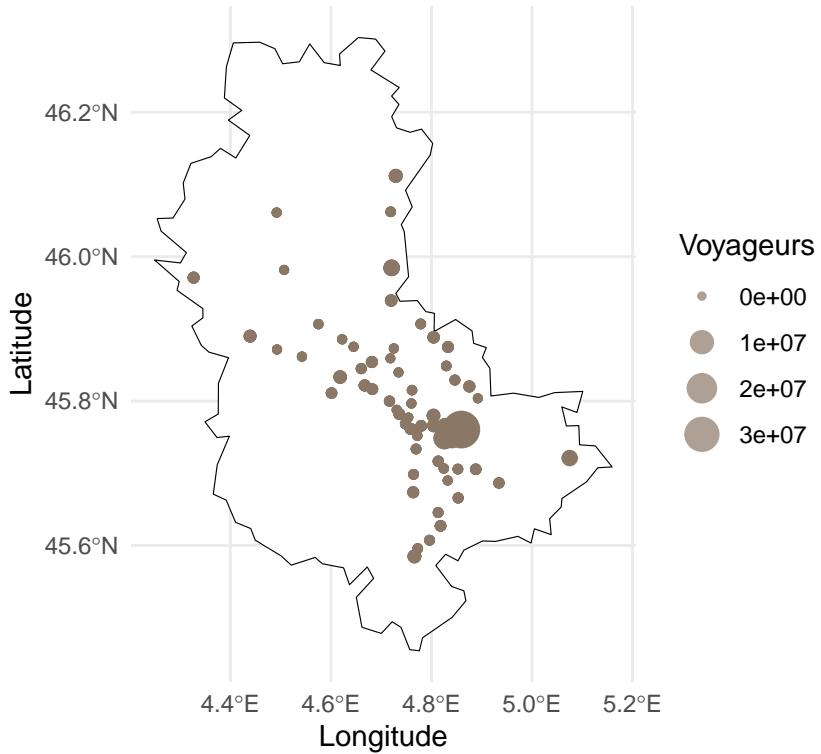


**Analyse.** Le résultat est sans appel. Toutes les gares en bordure du département concentrent plus de fréquentation que les gares de campagne, plus éloignées de Paris. Cela nous ramène également à l'étalement urbain et la concentration de la population proche de Paris dont nous avions parlé un peu plus haut.

Réponse à la question : Le nombre de voyageurs n'est donc pas réparti de la même manière au sein des départements.

Complément : On réalise la même visualisation que précédemment mais pour le département du Rhône.

## Fréquentation des gares du département 69 (2022) <Vue Spatiale>



**Analyse.** On remarque que Lyon est la gare la plus fréquentée du département. Très légèrement autour, on constate que les gares périphériques de Lyon sont plus fréquentées en moyenne que celles en dehors de la banlieue lyonnaise. Cela semble donc valider l'hypothèse que les gares les plus proches des grandes villes sont plus utilisées.

Dans un contexte entreprise, on pourrait cibler en priorité les gares des grandes villes et celles en périphérie afin d'affecter des effectifs pour le contrôle, le ménage ou le renseignement.

**Synthèse.** Ce qu'il faut retenir de cette section **Découverte** :

- A l'échelle nationale, Paris est au centre du réseau ferroviaire français.
- A l'échelle régionale, la région parisienne est incontestée en termes de fréquentation, bien que d'autres régions s'imposent plus ou moins à leur niveau : Lyon, Strasbourg, Lille pour les plus grandes puis Bordeaux, Marseille, Toulouse, Nantes, Rennes.
- Au sein des départements les gares les plus proches de la région parisienne semblent être les plus utilisées. Plus généralement, les gares en périphérie des grandes villes sont les plus fréquentées des départements après la ville en question.

## Exploration : Voyageurs

Dans cette partie, nous étudierons les voyageurs. Afin de mieux comprendre les voyageurs, nous avons choisi de nous intéresser aux différents profils qui utilisent les trains des réseaux ferrés de France.

### Visualisations réalisées

1. Exploration de la répartition par âge des passagers

2. Exploration de la répartition par CSP
3. Exploration de la répartition par motif de déplacement
4. Nombre de voyageurs par année

*Quel est le voyageur moyen de la SNCF ? Comment ce voyageur diffère en fonction des gares ? Quel est la relation entre les motifs de voyage des passagers et leur répartition par âge et par profession ?*

Nous étudierons ici la question : **Quel est le voyageur moyen de la SNCF ?**

On s'attend à obtenir une personne d'âge moyen (30 ans) qui se déplace pour le travail et qui est cadre.

## 1. Exploration de la répartition par âge des passagers

En raison de quelques problèmes dans l'ensemble de données, tels que des codes UIC incohérents et des variations dans les années de recensement et les données des stations, nous allons nous concentrer exclusivement sur la répartition par âge pour les années 2015, 2016 et 2017.

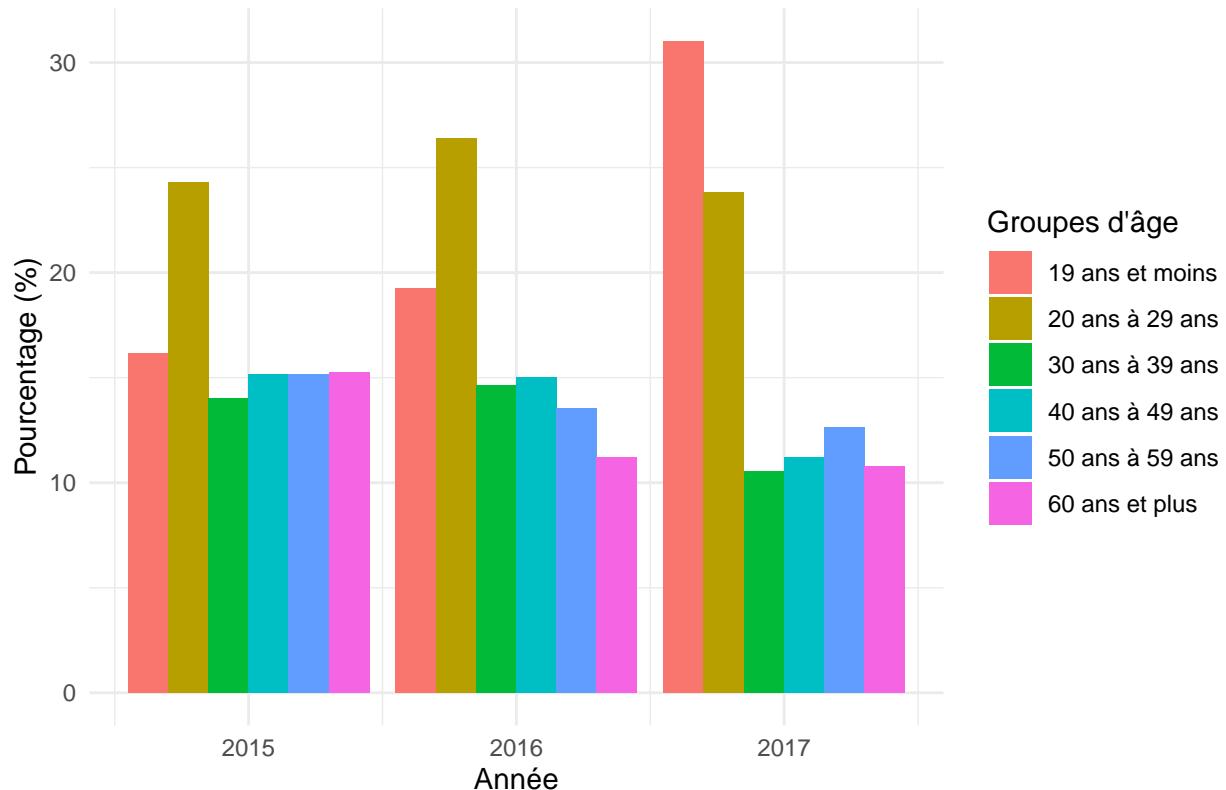
Conserver les colonnes spécifiées dans frequentation et renommer Code UIC en UIC et Filtrer le age\_voya pour une année 2015,16,17.

Filtrer freq\_selected par nom de station, en ne conservant que les lignes qui correspondent à Nom dans age\_filtered, et vice versa.

Calculer le nombre de voyageurs dans chaque groupe d'âge et les nombres totaux pour chaque année

L'étape suivante consiste à calculer le pourcentage des groupes d'âge ainsi que la cartographie.

**Profil moyen du voyageur par groupe d'âge et année (2015 à 2017)**



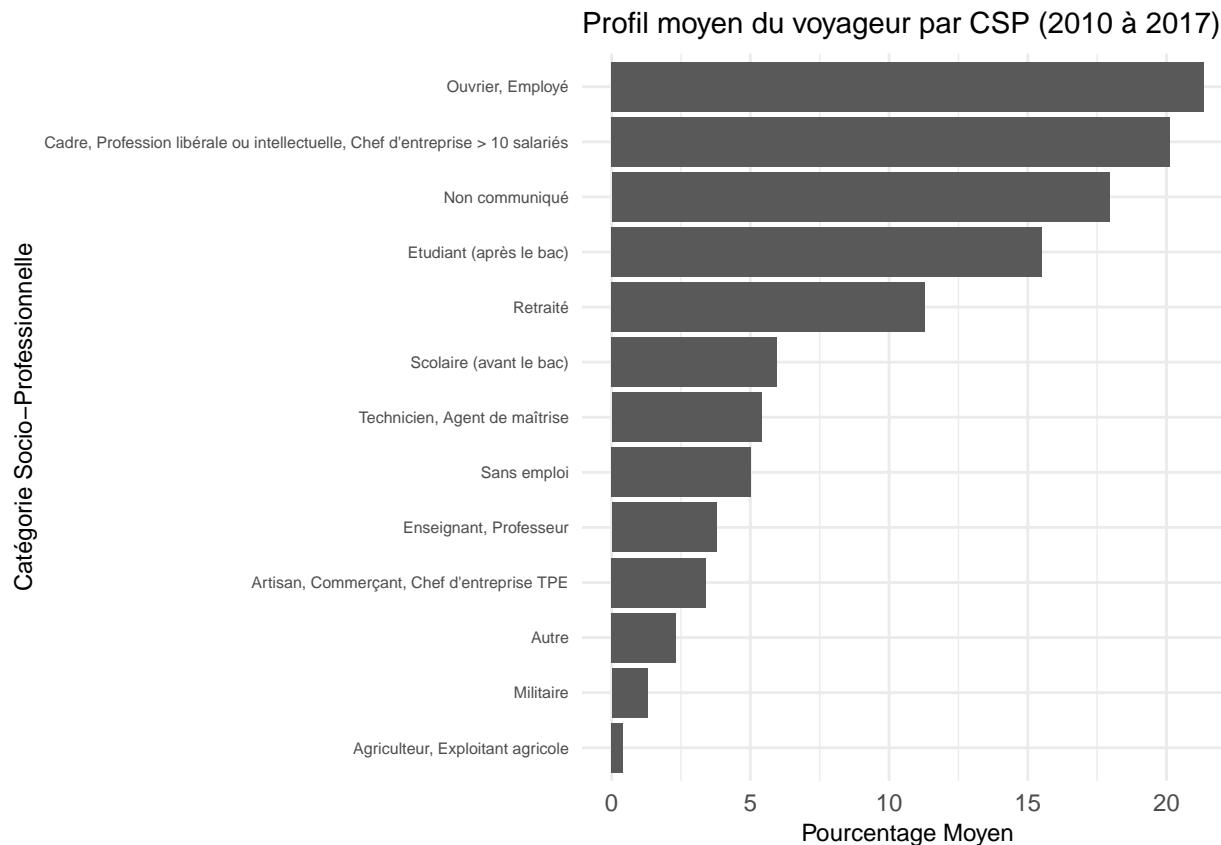
## 2. Exploration de la répartition par CSP

Il est également intéressant de recueillir des informations sur le voyageur moyen de la SNCF. En premier lieu nous allons voir les différentes catégories socio-professionnelles.

PS : Le Pourcentage correspond au pourcentage par rapport au CSP d'une gare.

CSP	Pourcentage_moyen
<chr>	<dbl>
1 Ouvrier, Employé	21.3
2 Cadre, Profession libérale ou intellectuelle, Chef d'entreprise > 10 salariés	20.1
3 Non communiqué	17.9
4 Etudiant (après le bac)	15.5
5 Retraité	11.3
6 Scolaire (avant le bac)	5.95
7 Technicien, Agent de maîtrise	5.41
8 Sans emploi	5.04
9 Enseignant, Professeur	3.78
10 Artisan, Commerçant, Chef d'entreprise TPE	3.40
11 Autre	2.31
12 Militaire	1.29
13 Agriculteur, Exploitant agricole	0.415

Regroupement des CSP par pourcentages.

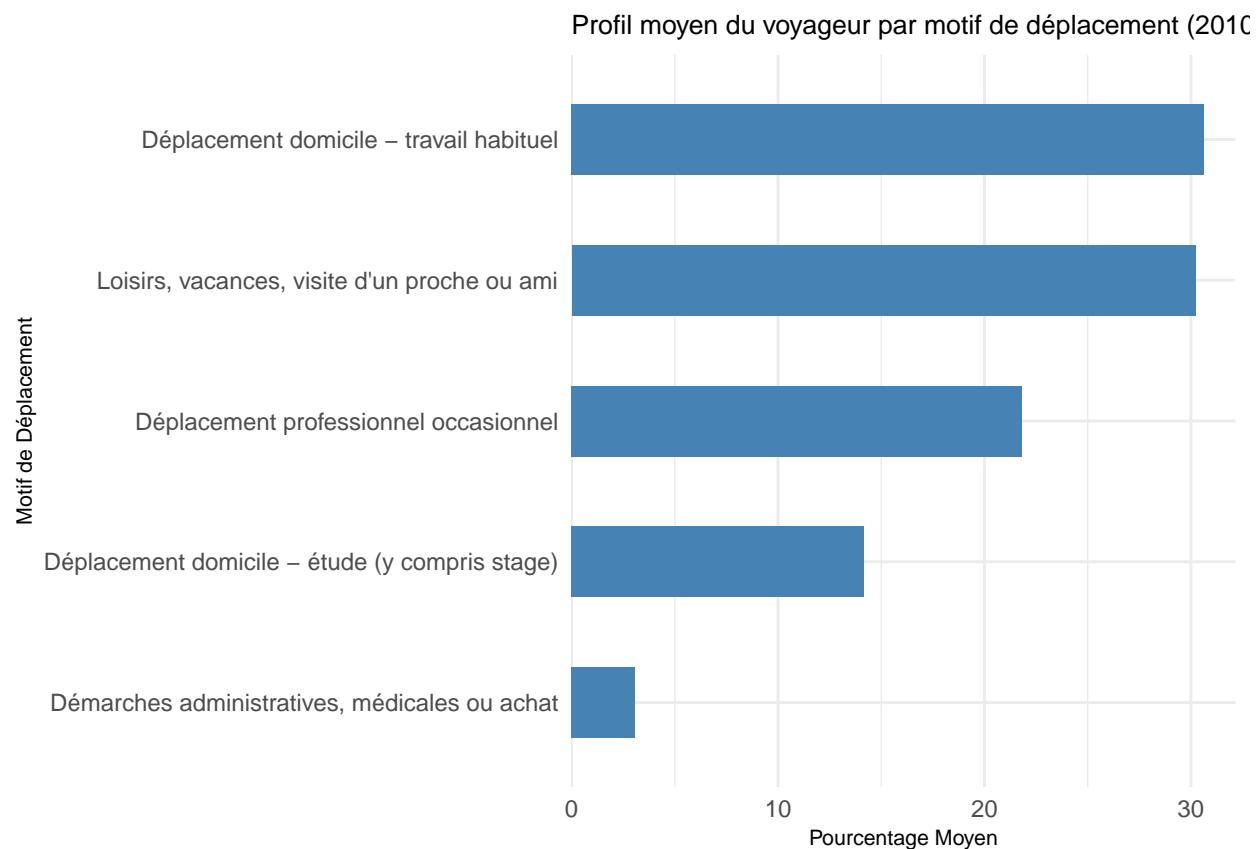


### 3. Exploration de la répartition par motif de déplacement

Ici nous nous intéressons aux motifs de déplacement des voyageurs de la SNCF.

En premier lieu, nous regardons le motif de déplacement global qui ressort le plus parmi toutes les gares.

## # A tibble: 5 x 2	Pourcentage_moyen
## `Motif du déplacement`	<dbl>
## <chr>	
## 1 Déplacement domicile - travail habituel	30.6
## 2 Loisirs, vacances, visite d'un proche ou ami	30.2
## 3 Déplacement professionnel occasionnel	21.8
## 4 Déplacement domicile - étude (y compris stage)	14.2
## 5 Démarches administratives, médicales ou achat	3.08



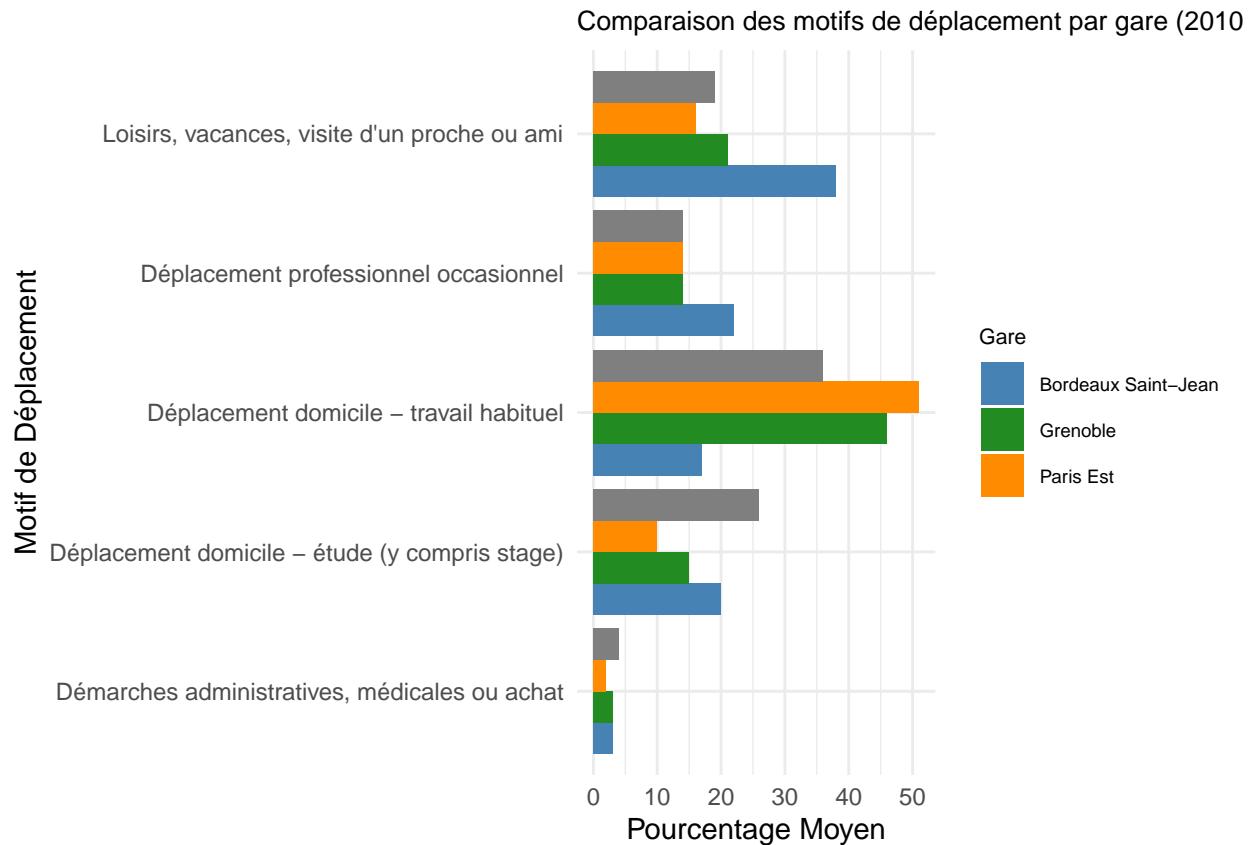
Enfin, nous aurions pu réaliser également un nuage de points afin d'observer des clusters significatifs pour compléter notre analyse de la question.

Réponse à la question : Le voyageur moyen est jeune (<30 ans). Il travaille en tant qu'ouvrier, employé ou cadre, chef d'entreprise. Il effectue des trajets entre son domicile et son travail.

On aimerait savoir si le motif de déplacement principal change en fonction de la gare.

Nous avons donc proposé d'étudier la question : Comment ce voyageur diffère en fonction des gares ?

Nous n'avons pas d'attente particulière sur cette question.



Il est à noter que nous n'avons sélectionné que des gares de grandes villes dans cette visualisation. Il aurait été peut être judicieux de comparer cela avec des gares plus rurales. On aurait pu vérifier également certaines tendances, comme par exemple celles des vacances et des loisirs qui devrait être beaucoup plus important dans les villes ensoleillées.

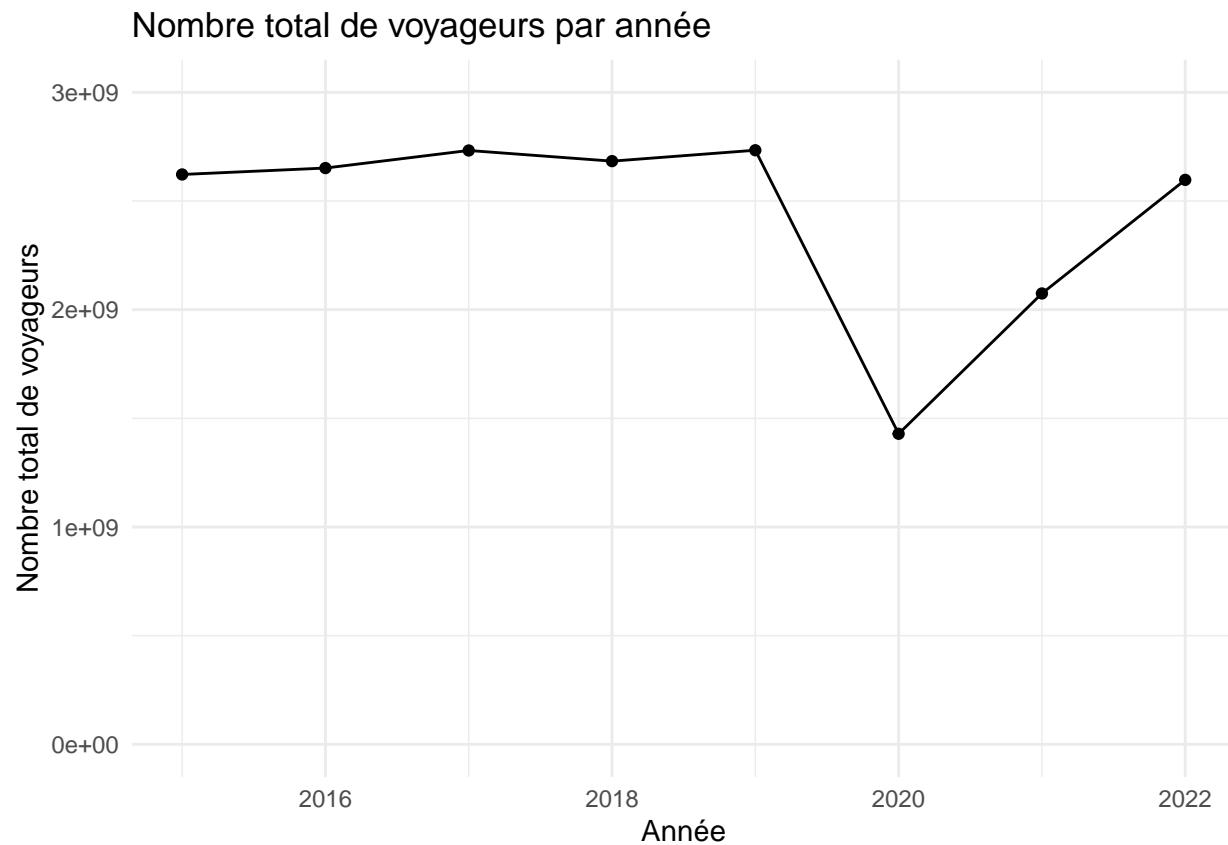
Réponse à la question : Le motif de déplacement semble changer entre les gares .

#### 4. Nombre de voyageurs par année

Réalisons un graphique simple qui nous montre le nombre de voyageurs total par année.

Aperçu du dataset fréquentations

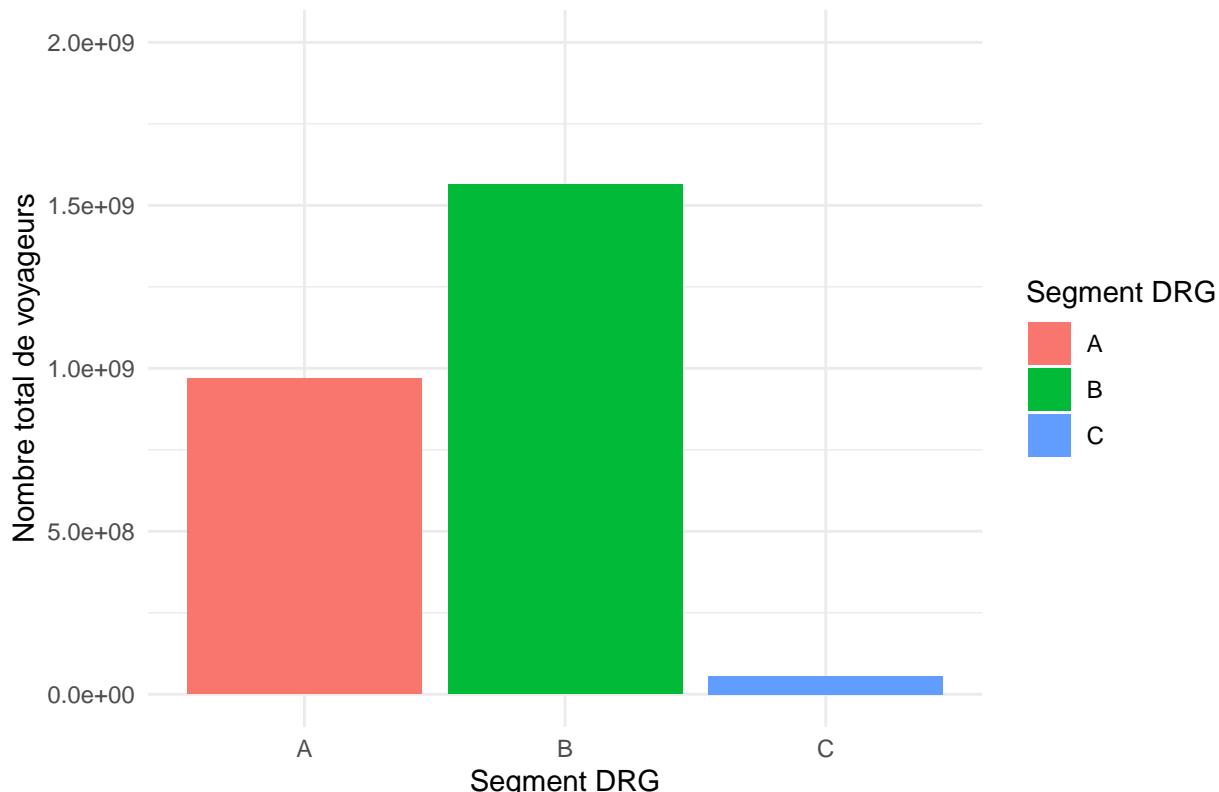
```
## # A tibble: 6 x 8
##   Gare      UIC  Code_Postal DRG Année Personnes Voyageurs Département
##   <chr>    <chr> <chr>     <chr> <dbl>    <dbl>    <dbl> <chr>
## 1 Abbaretz 481614 44170      C    2022    40825    40825 44
## 2 Abbaretz 481614 44170      C    2021    27466    27466 44
## 3 Abbaretz 481614 44170      C    2020    22773    22773 44
## 4 Abbaretz 481614 44170      C    2019    38473    38473 44
## 5 Abbaretz 481614 44170      C    2018    38027    38027 44
## 6 Abbaretz 481614 44170      C    2017    35637    35637 44
```



On voit rapidement qu'il y a un “trou” du à la période COVID 2020.

Maintenant intéressons nous aux nombres de voyageurs par segment DRG.

## Répartition des voyageurs par segment DRG en 2022



Pour comprendre la différence entre les ségmentation que l'on voit ci-dessus, il faut d'abord comprendre quelle gare est reliée à quelle segment DRG. Le SNCF nous donne cette définition :

**Catégorie A** : gares de voyageurs d'intérêt national. Ces gares sont celles dont la fréquentation par des voyageurs des services nationaux et internationaux de voyageurs est au moins égale à 250 000 voyageurs par an ou dont ces mêmes voyageurs représentent 100% des voyageurs.

**Catégorie B** : gares de voyageurs d'intérêt régional. Le périmètre de gestion correspond, dans chaque région, à l'ensemble des gares n'appartenant pas à la catégorie A mais dont la fréquentation totale est au moins égale à 100 000 voyageurs par an.

**Catégorie C** : gares de voyageurs d'intérêt local. Leur périmètre de gestion correspond, dans chaque région, à l'ensemble des gares de cette catégorie. La redevance est fixée, par région, pour l'ensemble des gares de cette catégorie.

Cette catégorisation nous conforme dans l'idée que la majorité des voyageurs ne s'arrête que très rarement sur des gares d'intérêt locale. La plupart se trouve dans des villes petites ou moyennes se trouvant sur une ligne unique ou très peu de ligne.

Cependant même si le chiffre paraît petit en comparaison des autres, les gares de la catégorie C voient passer presque 57 millions de voyageurs sur 2022. On doit aussi dénombrer énormément de gares car la limite maximale de la catégorie C est 100 000 voyageurs annuels. On peut donc considérer plus de 570 gares font partie de cette catégorie au minimum.

En voyant l'échelle on peut donc mettre en perspective que les catégories A et B n'ont pas une différence flagrante, sûrement dûe à la population d'Île-de-France qui doit contribuer énormément aux chiffres de la catégorie B. En prenant en compte les résultats des premières visualisations sur les gares les plus fréquentées, on peut assumer que le nombre de gares de la catégorie A doit être moindre en comparaison à la catégorie B et C.

---

## Explorations : Objets

Dans cette partie, nous voulons nous intéresser plus particulièrement aux objets.

### Visualisations réalisées

1. Objets perdus
2. Objets restitués
3. Probabilité de retrouver un objet perdu
4. Probabilité de retrouver un objet selon son type
5. Afflux d'objets selon le mois
6. Perte d'objet selon la gare

*Y-a-t-il plus de chances de perdre un objet selon la gare ? Doit-on s'attendre à un afflux d'objets perdus plus important dans les mois de Juillet-Août 2024 plus important que les dernières années ? Quelles sont les chances de retrouver un objet perdu ? Quelles sont les chances de retrouver un objet en fonction de sa nature ?*

### 1. Objets perdus

Tout d'abord, petite visualisation du dataset obj\_perdus.

```
## # A tibble: 6 x 6
##   date           gare      UIC  nature    type  enregistrement
##   <dttm>        <chr>    <chr>  <chr>  <chr> <chr>
## 1 2019-05-24 16:52:18 Paris Est 0087113001 Manteau~ Vête~ Déclaration d-
## 2 2019-05-24 16:53:32 <NA>       <NA>    Téléph~ Appa~ Déclaration d-
## 3 2019-05-24 17:00:58 <NA>       <NA>    Sac de~ Baga~ Déclaration d-
## 4 2019-05-24 17:09:28 <NA>       <NA>    Autre ~ Pièc~ Déclaration d-
## 5 2019-05-24 17:11:10 Paris Saint-Lazare 0087384008 Sacoch~ Baga~ Déclaration d-
## 6 2019-05-24 17:30:25 <NA>       <NA>    Téléph~ Appa~ Déclaration d-
## #>
## #>   date           gare      UIC
## #>   Min.   :2013-05-24 11:02:10 Length:1869692  Length:1869692
## #>   1st Qu.:2016-12-05 09:57:15 Class  :character Class  :character
## #>   Median :2019-06-15 16:48:46 Mode   :character Mode   :character
## #>   Mean   :2019-07-16 20:57:22
## #>   3rd Qu.:2022-05-22 17:46:10
## #>   Max.   :2024-04-21 12:59:33
## #>
## #>   nature          type      enregistrement
## #>   Length:1869692  Length:1869692  Length:1869692
## #>   Class  :character Class  :character Class  :character
## #>   Mode   :character Mode   :character Mode   :character
## #>
## #>
```

Quelque chose m'interpelle dès le début : c'est la colonne "Type d'enregistrement", il semble que celle-ci est toujours remplie de la même manière

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

C'est donc le cas, donc nous n'utiliserons pas cette colonne .

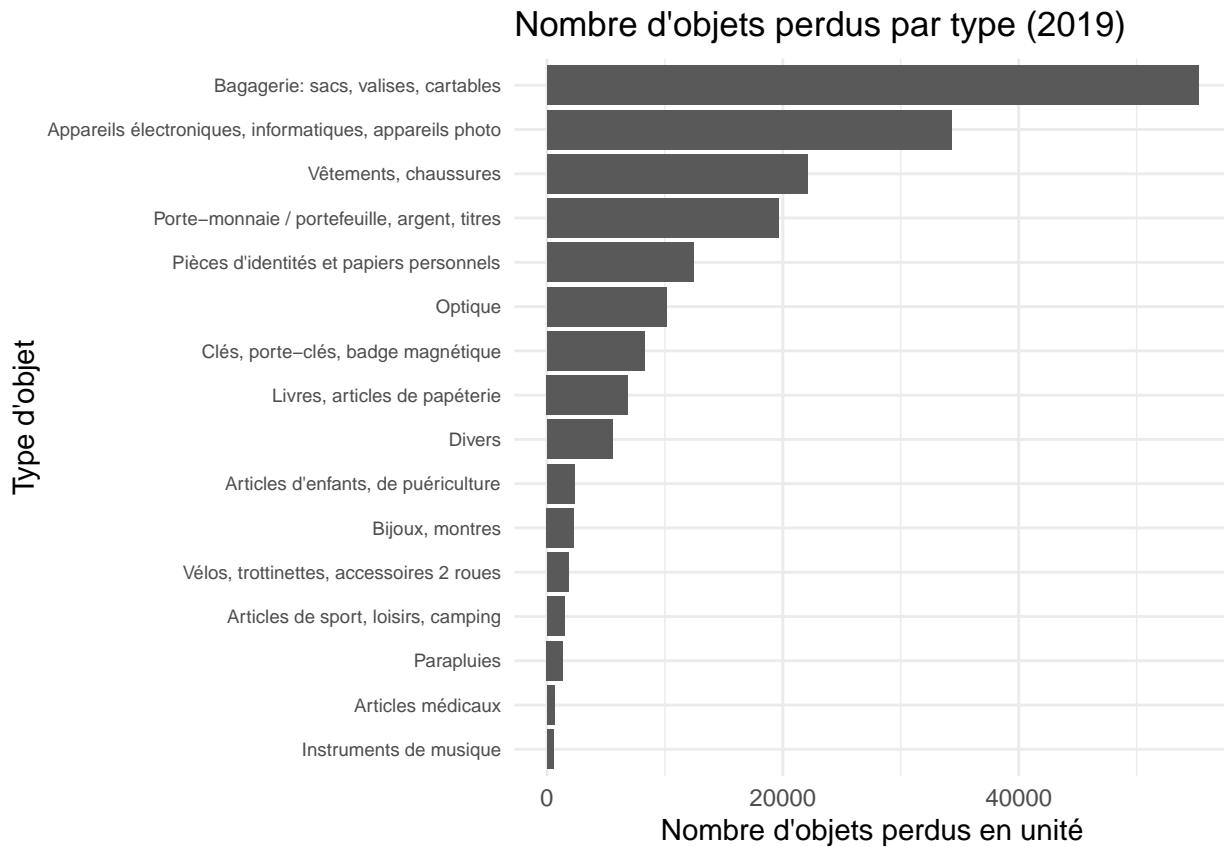
On distingue aussi que de nombreuses gares ne sont pas présentes.

Nous voulons effectuer plusieurs réductions sur ce jeu de données. Tout d'abord, les informations que nous avons datent de 2019 jusqu'à nos jours. Si nous voulons faire des recoupements avec le dataset de voyageur, nous allons restreindre les dates à 2019 seulement car les années 2020 et 2021 ne sont pas forcément représentatives du trafic ferroviaire normal. L'année 2022 sera analysée prochainement pour voir si une tendance peut commencer à apparaître.

```
## # A tibble: 1 x 1
##      n
##   <int>
## 1 185065
```

Nous avons donc maintenant 185 065 entrées dans notre dataframe.

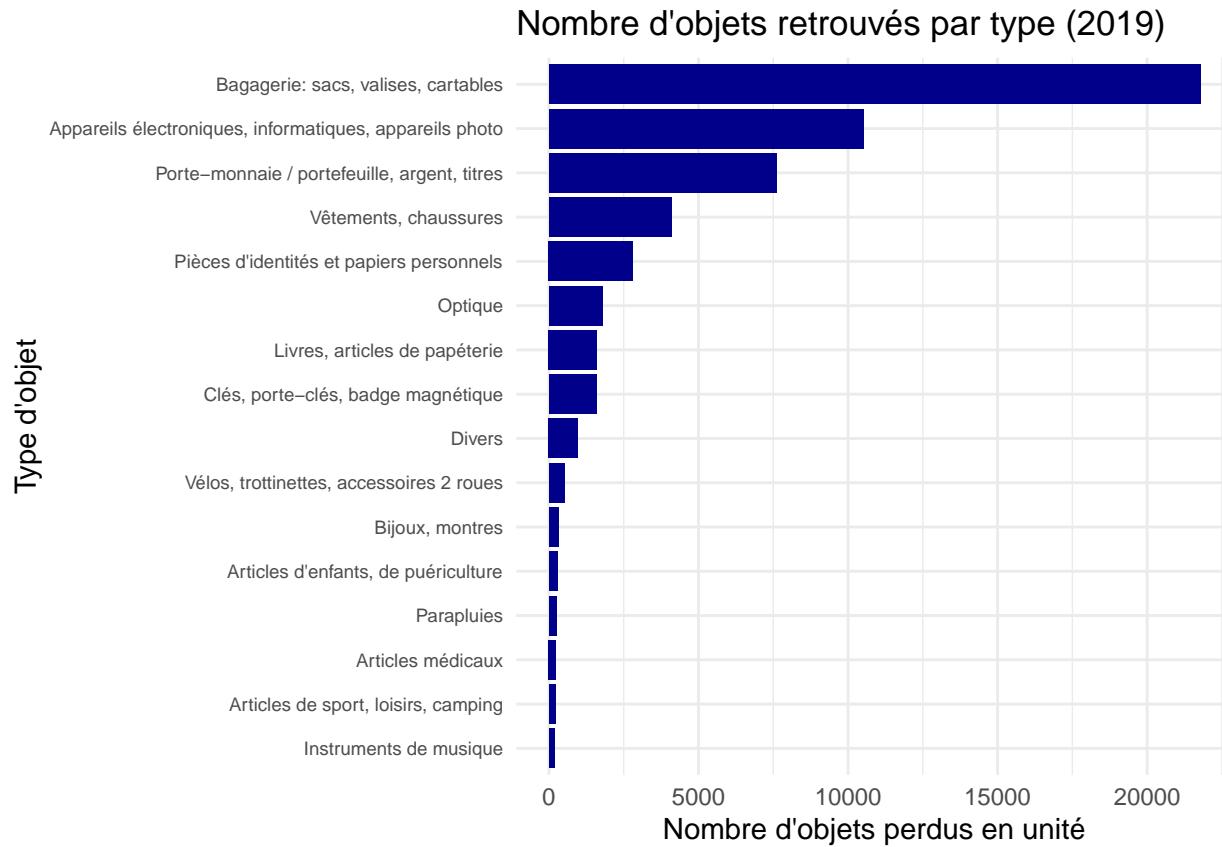
Découvrons ce que celui-ci nous cache dans la répartition des objets perdus :



Sans surprise, les objets les plus couramment perdus sont donc dans l'ordre :

- Bagages
- Appareils électroniques
- Vêtements

```
## # A tibble: 1 x 1
##      n
##   <int>
## 1 54834
```



## 2. Objets restitués

Nous allons ensuite parcourir le dataset 7 : obj\_trouves

```
## # A tibble: 1 x 1
##      n
##   <int>
## 1 88258
```

On découvre donc qu'il y a plus d'objets trouvés que d'objets perdus rendus. Cela veut donc dire que tous les objets récupérés ne sont pas forcément déclarés comme perdus au préalable.

```
## # A tibble: 1 x 1
##      n
##   <int>
## 1 88239
```

On voit qu'il y a 19 objets rendus qui n'ont pas été identifiés par des code UIC ce qui est étrange. Voyons cela de plus près :

```
## # A tibble: 19 x 6
##   date           date_restit       gare   UIC   nature          type
##   <dttm>         <dttm>       <chr>  <chr> <chr>            <chr>
## 1 2019-11-27 18:51:31 NA        <NA>   <NA>  Sac d'enseigne (pl~ Baga-
## 2 2019-12-18 12:09:32 NA        <NA>   <NA>  Sac à dos          Baga-
## 3 2019-12-19 10:17:09 NA        <NA>   <NA>  Autre pièce ou pap~ Pièc-
## 4 2019-03-05 07:14:37 NA        <NA>   <NA>  Carte d'identité, ~ Pièc-
## 5 2019-08-22 16:20:41 NA        <NA>   <NA>  Carte d'identité, ~ Pièc-
```

```

## 6 2019-08-31 08:48:21 2021-01-06 06:14:46 <NA> <NA> Valise, sac sur ro~ Baga~
## 7 2019-09-02 17:16:33 NA <NA> <NA> Téléphone portable~ Appa~
## 8 2019-09-04 11:28:10 2019-09-05 15:49:33 <NA> <NA> Téléphone portable Appa~
## 9 2019-05-06 12:35:54 2019-05-07 08:15:26 <NA> <NA> Sac à dos Baga~
## 10 2019-11-27 19:10:10 NA <NA> <NA> Sac d'enseigne (pl~ Baga~
## 11 2019-12-02 14:28:36 NA <NA> <NA> Autres divers Dive~
## 12 2019-12-18 12:02:47 NA <NA> <NA> Carte d'identité, ~ Pièc~
## 13 2019-08-26 07:49:55 NA <NA> <NA> Téléphone portable Appa~
## 14 2019-08-27 09:01:52 NA <NA> <NA> Porte-monnaie, por~ Port~
## 15 2019-09-03 18:58:21 NA <NA> <NA> Sac de voyage, sac~ Baga~
## 16 2019-12-17 14:52:16 2020-01-14 11:48:42 <NA> <NA> Sac de voyage, sac~ Baga~
## 17 2019-09-03 18:54:07 NA <NA> <NA> Valise, sac sur ro~ Baga~
## 18 2019-09-04 17:15:19 NA <NA> <NA> Carte de crédit Port~
## 19 2019-09-17 16:38:07 NA <NA> <NA> Valise, sac sur ro~ Baga~

```

En voyant que la gare n'est pas présente non plus, on peut juger ce rendu comme ayant été entré dans la database sans qu'il n'ait été rendu dans une gare. On peut supposer un rendu dans le train juste après la perte.

Nous allons maintenant essayer de comprendre ce qu'est ce code UIC et s'il y a une relation entre les codes UIC des pertes et des objets trouvés.

Nous allons donc comparer les 2 listes de code UIC.

```

## # A tibble: 1 x 1
##      n
##   <int>
## 1    145
## # A tibble: 1 x 1
##      n
##   <int>
## 1     1

```

Il semble que seulement 146 code UIC différents et après quelques recherches en ligne, le code UIC est l'ID des gares. Donc cette colonne ne nous est pas forcément utile.

Nous allons donc maintenant essayer de mettre en relation plutôt les dates entre les 2 datasets :

```

## # A tibble: 302 x 1
##   date
##   <dttm>
## 1 2019-02-02 13:52:46
## 2 2019-02-04 11:01:06
## 3 2019-02-11 11:25:20
## 4 2019-02-11 16:54:56
## 5 2019-10-02 08:35:47
## 6 2019-01-30 15:21:08
## 7 2019-10-04 14:48:42
## 8 2019-10-10 10:14:01
## 9 2019-03-06 14:04:44
## 10 2019-01-07 10:16:07
## # i 292 more rows

## # A tibble: 54,429 x 1
##   date
##   <dttm>
## 1 2019-05-24 16:52:18

```

```

## 2 2019-05-24 17:11:10
## 3 2019-05-24 18:26:47
## 4 2019-05-24 19:29:14
## 5 2019-05-24 19:58:27
## 6 2019-05-25 07:41:26
## 7 2019-05-25 07:53:09
## 8 2019-05-25 08:09:59
## 9 2019-05-25 08:16:59
## 10 2019-05-25 08:45:00
## # i 54,419 more rows

```

Nous avons donc seulement 302 dates en commun entre les 2 jeux de données ce qui est très faible. On peut donc supposer que les 2 datasets ne sont pas liés et les dates ne correspondent pas. On peut supposer que la date des objets perdus provient du questionnaire de perte que les voyageurs remplissent, tandis que la date des objets trouvés provient du formulaire rempli par les agents SNCF.

Nous ne pouvons donc pas faire de relation directe entre un objet déclaré comme perdu et un objet déclaré comme trouvé.

Il est donc difficile de faire des hypothèses très spécifique sur les probabilités de retrouver un objet perdu. La façon de faire ça serait de connaître un nombre précis de pertes et sur ces pertes savoir combien d'entre elles ont été rendues. Sans le lien entre les pertes et les objets rendus, énormément de biais peuvent apparaître. Pour faire les prochaines probabilités, nous allons faire l'hypothèse que l'ensemble des objets rendus ont fait l'objet d'une entrée dans le formulaire de perte.

### 3. Probabilité de retrouver un objet perdu

Nous étudierons ici : Quelles sont les chances de retrouver un objet perdu ?

On pense que beaucoup d'objets ne sont jamais retrouvés par leur propriétaire.  
Nous n'avons pas d'idée de la proportion que cela représente.

Pour cette première probabilité nous allons faire un calcul simple :

(Nombre d'objets retrouvés / Nombre d'objets perdus) \*100

Cela s'exprime ainsi :

```

##          n
## 1 47.69027

```

Je trouve la probabilité relativement élevée car pas très éloignée du 50%. Peut-être aussi que le postulat personnel avant le calcul était relativement pessimiste.

Réponse à la question : La probabilité semble être très élevée. Nous n'avons pas eu assez d'informations quant aux conditions de récupération des objets par leur propriétaire.

### 4. Probabilité de retrouver un objet selon son type

Nous tenterons de donner une réponse à la question suivante : Quelles sont les chances de retrouver un objet en fonction de sa nature ?

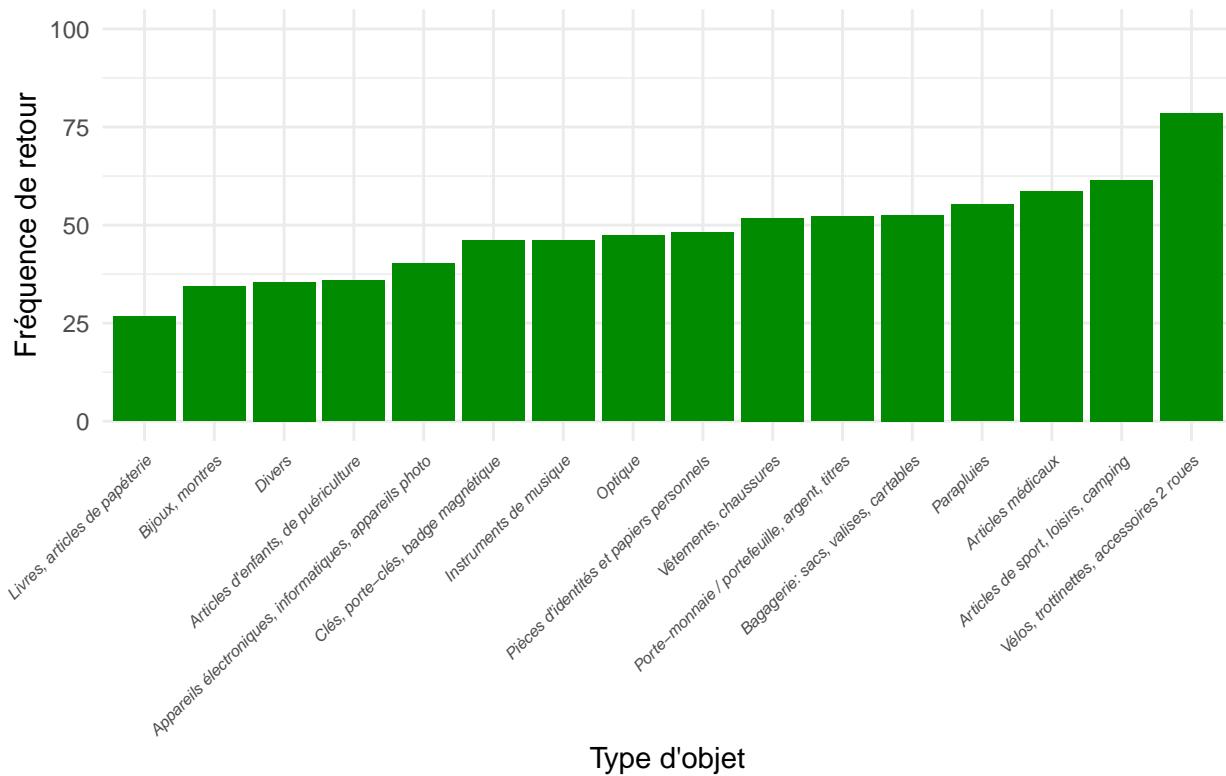
On pense que les vélos sont plus retrouvés que d'autres types d'objets comme des petits accessoires (écouteurs, montres, bijoux etc...)

Afin d'avoir une probabilité selon le type il faut donc que nous effectuons la même opération que précédemment avec comme paramètre le type de l'objet.

N.B. la "nature" dans l'intitulé de la question a été changé en type car la nature entrée dans le formulaire des objets trouvés est trop spécifique.

```
## [1] "type_obj_trouves" "Freq"
```

## Fréquence de rendu d'objets perdus par type



Plus de 50 points distinguent la fréquence de retour des vélos, trottinettes et accessoires de celle des livres et articles de papeterie. Les vélos, trottinettes et accessoires est une catégorie qui est bien au-dessus du second article le plus rendu de + de 10 points. Je m'attendais clairement à un taux de retour bien moins important sur toutes les catégories car mon environnement proche parlait souvent d'objets perdus et peu d'objets retrouvés. J'ai l'impression que l'encombrement et la nécessité d'un objet influence sur taux de retour. Un vélo, article de sport, parapluie, sont souvent encombrants et facile à identifier. Les porte-monnaies, pièces d'identité sont aussi proche du 50% ce qui me semble normal.

Il serait aussi intéressant de voir que ces statistiques sont fondées sur les objets qui sont perdus en train ou en gare et ceux qui sont ramenés en gare par d'autres voyageurs ou agents. Les vols ne sont donc pas forcément comptabiliser. Je pense qu'avec plus d'informations, nous pourrions ajouter un critère de vol ou perte selon l'objet. Malheureusement les données ne sont pas assez détaillées, mais nous pourrions peut-être deviner une tendance en croissant d'autres études portant sur le sujet.

Réponse à la question : Plus un objet semble de grande taille, plus celui-ci semble être retrouvé par son propriétaire.

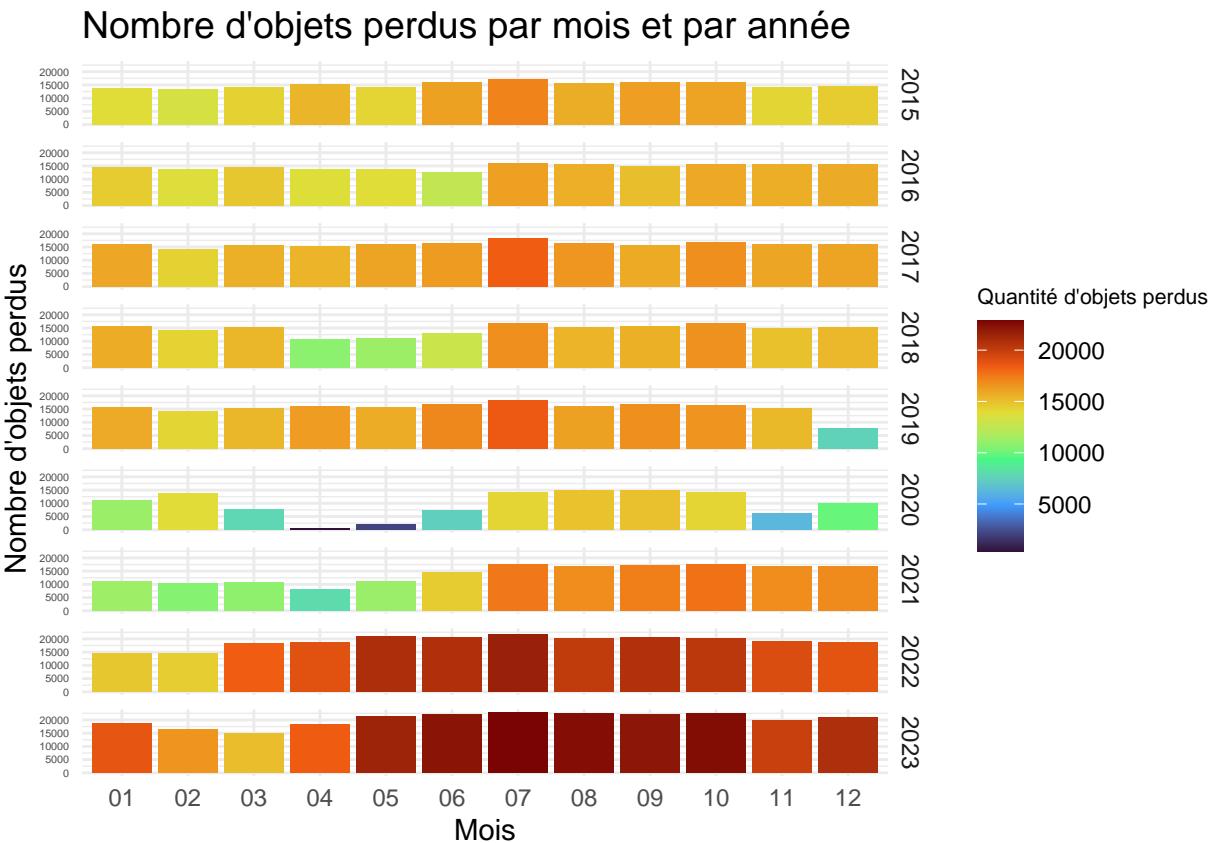
## 5. Afflux d'objets selon le mois

Nous allons maintenant traiter l'ensemble du jeu de données des objets perdus pour voir s'il y a un motif qui pourrait faire croire que certains mois sont plus propices à la perte d'objets.

Ainsi, la question suivante nous a particulièrement intéressée : Doit-on s'attendre à un afflux d'objets perdus plus important dans les mois de Juillet-Août 2024 plus important que les dernières années ?

On imagine ces dernières années une utilisation plus importante des mobilités

douces et donc du train pour diminuer l'impact carbone. Le nombre d'usagers du train augmente pouvant accroître également le nombre d'objets perdus.



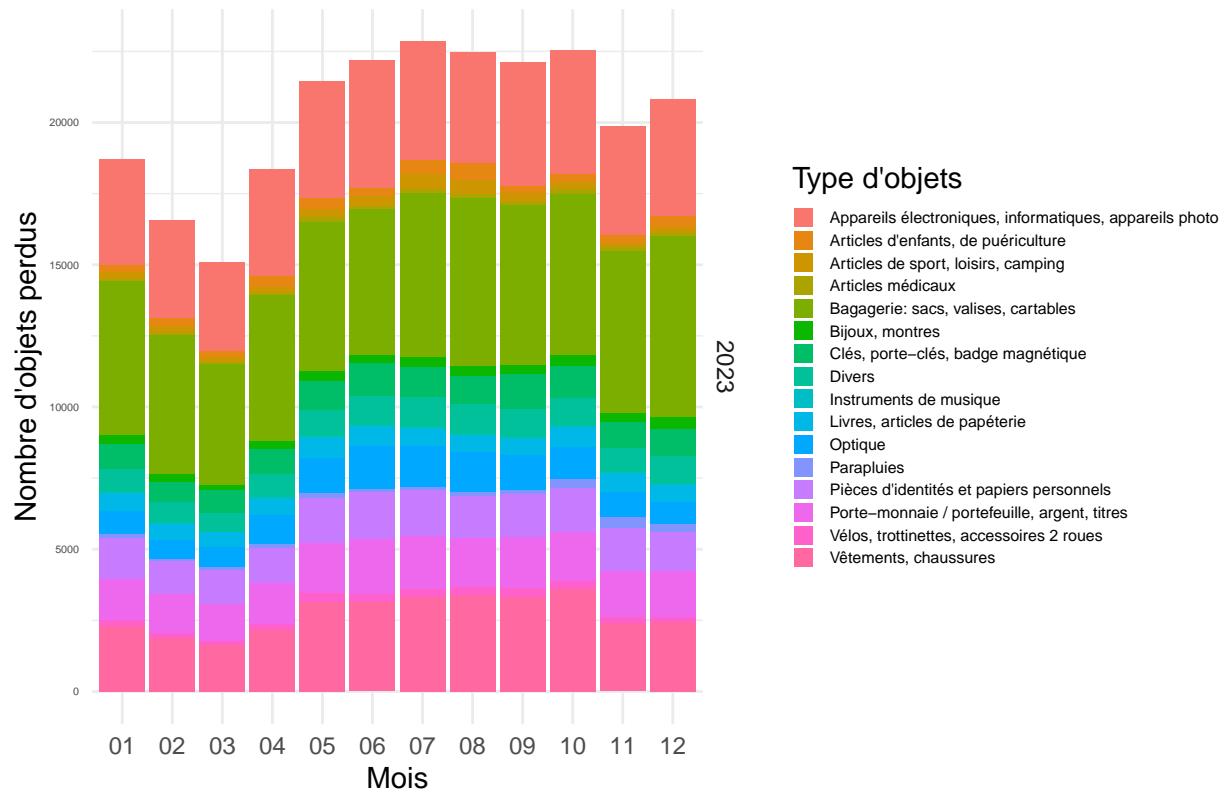
On distingue clairement un nombre élevé d'objets tout d'abord lors de la dernière année mais aussi au niveau du mois de juillet qui est (à part pour l'année 2020) toujours le mois avec le plus de pertes.

On peut en conclure, qu'en terme de nombre de pertes pur, le mois de juillet est le plus à risque. Mais il semble aussi important de noter qu'il y a sûrement une corrélation avec le nombre de voyageurs. Ce qu'il faudrait c'est le taux de pertes par voyageurs. Cependant, avec les données que nous avons actuellement nous ne pouvons pas savoir.

On pourrait aussi argumenter que les vacanciers pourraient contrebalancer les travailleurs qui n'ont donc pas les mêmes mois de fréquentations. Les données ne sont pas assez précises pour indiquer s'il y a une recrudescence de pertes en juillet, (et principalement les mois d'été).

**Réponse à la question :** Au vu de la croissance d'objets perdus importante de l'année dernière, on peut s'attendre à avoir plus d'objets perdus cet été (2024) que l'année précédente (2023).

## Nombre d'objets perdus par mois (2023)



Une visualisation intéressante serait de regarder les objets et affaires vestimentaires perdus pour une période. Nous avons donc utilisé Shiny App pour générer dynamiquement la visualisation ci-dessous. On étudie sur cette dernière l'année 2022.

Nombre d'objets perdus par type, par mois et par année (2022)

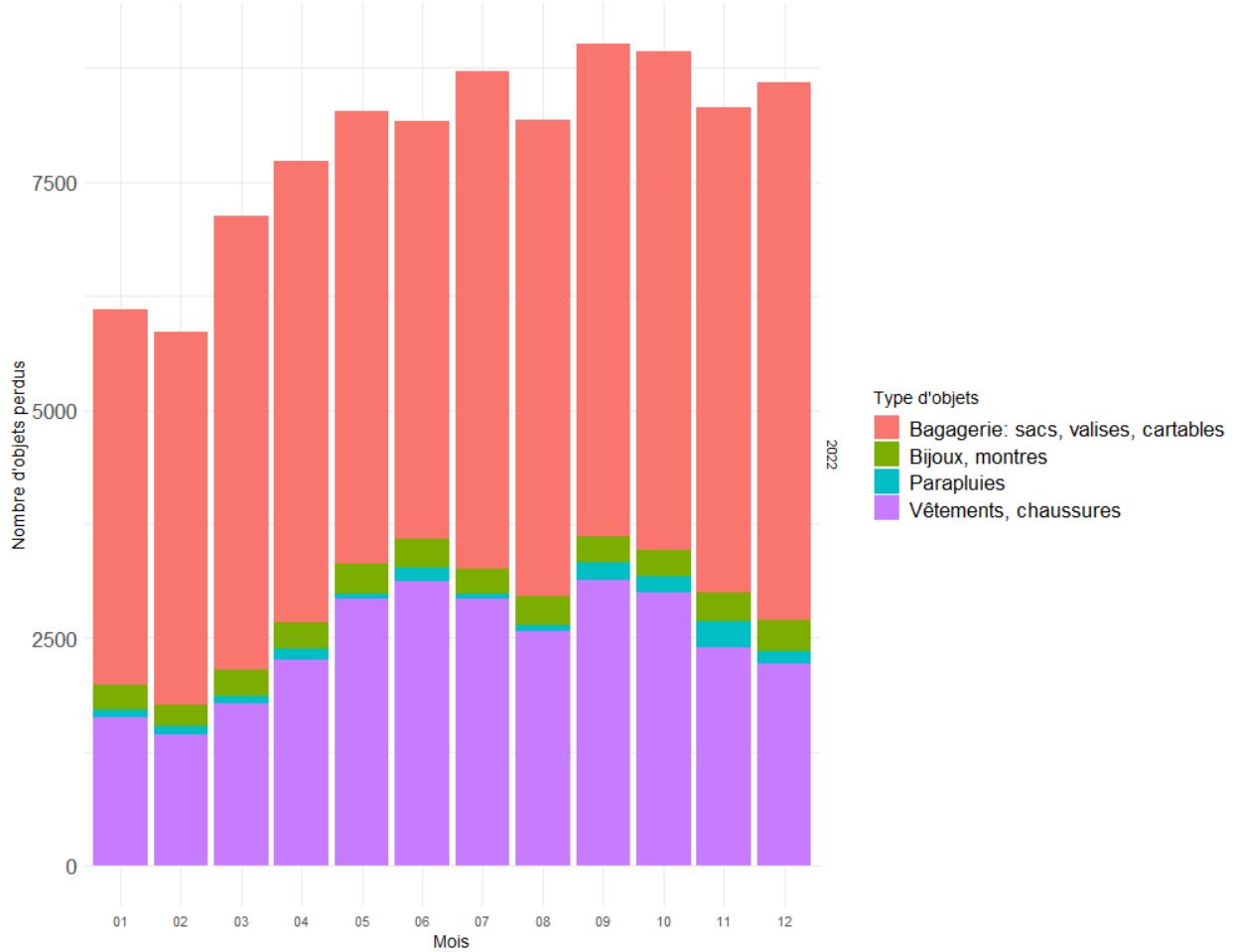


Figure 5. Capture Shiny App : Nombre d'objets perdus par type, par mois et par année (2022).

Au vu de la visualisation obtenue, on constate que les parapluies sont plus perdus au mois de Novembre (11). Cela peut s'expliquer donc par les conditions météorologiques de la période. En effet, il a plus tendance à pleuvoir en Novembre qu'au mois de Juillet. D'ailleurs, à la période estivale (Juillet-Août), très peu de parapluies sont perdus par les usagers des transports ferroviaires car il ne pleut que très peu en Eté.

D'autres objets à l'inverse ne dépendent pas nécessairement de la période de l'année (pas de saison spécifique). On peut voir que les bijoux et les montres par exemple sont perdus dans les mêmes quantités tout au long de l'année.

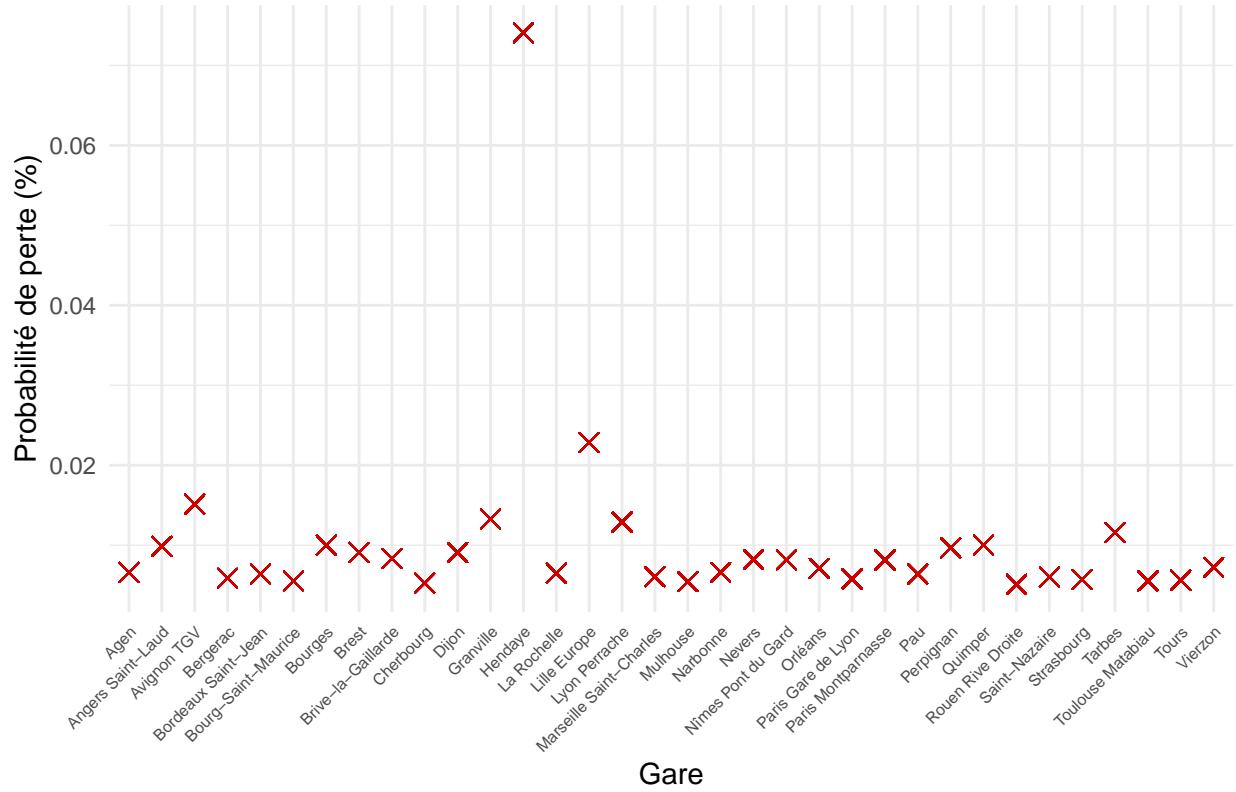
## 6. Perte d'objet selon la gare

Nous finissons le traitement des données pour essayer de déceler si le risque de perte est plus important selon la gare. L'étude ne sera faite que sur une année car les données de voyageurs sont année par année.

Nous étudierons ici : Y-a-t-il plus de chances de perdre un objet selon la gare ?

Nous n'avons pas vraiment d'idée pour répondre à cette question. Nous allons donc observer le résultat de nos visualisations.

## Probabilité de perte d'objet par voyageur selon la gare (2022)



Afin d'améliorer la visualisation, on ne voit ici que les gares dont le taux de perte par voyageur est au-dessus de 0,005 %. On peut voir ici qu'il vaut mieux éviter la gare de Granville où la probabilité de perte d'objet est supérieure à 0,07%. Cependant ce nombre est important à recontextualiser. Cela signifie que sur 10 000 voyageurs, 7 ont perdu un objet. Ce nombre reste assez faible. Et donc la plupart des gares vues ici sont plutôt sur l'ordre de 1 sur 10 000.

Réponse à la question : D'après notre visualisation, il semblerait qu'il y ait effectivement des gares à éviter pour ne pas perdre ses objets. Ici Granville apparaît comme la gare à éviter selon notre étude.

## Conclusion

Grâce aux données récupérées et leur traitement, nous avons pu découvrir une vision nationale (métropolitaine) du réseau ferroviaire. Nous avons également pu comprendre quel semble être le profil du voyageur moyen et les objets les plus perdus/retrouvés par ce dernier.

Au travers de la première partie axée sur la découverte, nous pouvons apercevoir deux réseaux en étoiles imbriqués. Un premier au niveau national qui se focalise sur l'Île-de-France, et un second au sein de l'Île-de-France qui se focalise sur Paris.

Ensuite, les données, bien que maigres sur les voyageurs, permettent de mettre en avant un groupe de voyageurs hétéroclite. On distingue tout de même plus de jeunes que dans la population française et les motifs de déplacement sont principalement pour le travail ou bien les vacances.

Enfin, la dernière analyse effectuée a été faite sur les objets et leur perte. On distingue donc des pertes plus importantes au moment de l'été et principalement en Juillet. On distingue aussi une récupération des objets

plus probable si l'objet est volumineux et/ou coûteux. Les objets les plus souvent perdus en ordre décroissant sont les bagages puis les objets électroniques, les porte-feuilles et les vêtements.

Ce sujet peut ouvrir sur de nouveaux axes et continuer ceux déjà proposés. On pourrait récolter plus de données sur les voyageurs et pousser bien plus loin notre analyse. La difficulté a résidé dans le fait de n'avoir que très peu d'observations pour réellement analyser correctement le voyageur moyen. Enfin, la suite de l'analyse pourrait se trouver vers des compléments quant à la propreté et de la satisfaction des clients en gare. Cela permettrait d'ouvrir de nouvelles voies quand à l'amélioration de la qualité de l'expérience client des voyageurs avec les services proposés en gare.

Dans le cadre de ce rapport, nous n'avons pas forcément mis en avant la finalité de cette analyse. Cependant, il serait judicieux de voir cette analyse comme une opportunité pour des entreprises de comprendre les clients potentiels (cibles) dans les gares (secteurs) avec leurs habitudes de consommation (produit).

## Notre avis sur le projet

“J'ai beaucoup apprécié ce projet qui permet vraiment de comprendre comment fonctionne une dataviz dans son ensemble. Il permet une vraie exploration du langage R et de R studio, et un développement de l'esprit de synthèse. Malheureusement, nous manquions de données pour certaines visualisations.”, Mathis

“Ce projet m'a permis de découvrir le langage R au travers d'un projet pertinent visant à explorer le fonctionnement, la cible et les ressources d'un OIV français : la SNCF. Parfaire mes compétences en Dataviz a également développé ma rigueur et mon intérêt pour cette discipline, tant sur le plan technique que sur l'étude qui en faite, appliquée aux sciences humaines et sociales. La compétence principale retenue de ce projet serait pour moi la manipulation de graphiques afin de tenter d'en faire ressortir des connaissances (insights). Néanmoins, j'aurai bien aimé utiliser Tableau un peu plus dans la partie projet. Au final, bilan très positif de cette expérience que je recommande fortement à toutes les personnes qui veulent monter en compétences dans ce domaine.”, Maxence

“Ce projet m'a permis de pousser mon esprit de synthèse pour mettre en avant seulement les graphiques et les informations importantes. J'aurais bien aimé ajouter encore plus de données telles que la répartition de la population française, l'âge et le CSP de celle-ci.”, Louis

“Grâce à ce projet, j'ai acquis une compréhension plus approfondie des principes de la visualisation des données étudiés en cours. Ce projet m'a permis d'explorer en profondeur le langage R et R Studio. En traitant les données des dernières années de la SNCF, j'ai appris à extraire des informations précieuses des données et à les présenter de manière efficace. Bien que nous ayons rencontré certaines difficultés en raison du manque de données pour certains graphiques, ces défis m'ont également poussé à accorder une plus grande importance à la collecte et à l'organisation des données.”, Wang

---

## Annexe

Répartition par partie :

- Introduction : Louis et Maxence
- Découverte : Maxence
- Voyageurs : Âge (Wang), CSP et Motif de déplacement (Mathis)
- Objets : Louis
- Conclusion : Louis
- Relecture : Maxence et Louis
- PowerPoint : Maxence et Louis

- Soutenance : Louis Tableau : Mathis
- Shiny App : Wang

Répartition par personnes :

- Louis : Introduction, Objets, Conclusion, Relecture, Powerpoint, Soutenance
- Maxence : Introduction, Découverte, Relecture, PowerPoint
- Wang : Voyageurs (Âge), Shiny App
- Mathis : Voyageurs (CSP, Motif), Tableau

Copyright : Mathis Girod, Maxence Jaulin, Louis Prodhon, Wang Zezhong