

Rogue One

JAULIN Maxence, PRODHON Louis, GIROD Mathis, 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 >](#)

 Services voyageurs (10)	 Description du réseau (20)	 Gares (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.

 Rapports (14)	 Comptage et flux (11)	 Sécurité ferroviaire (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	Continu	%	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 desservi 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:rnatuarearth':
##
##     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). *[TODO] Ajouter des explications sur la complexité de nos données*

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 avec des minuscules.

Exemple : Code Postal en code_postal

[TODO] Rajouter des données les tables modifiées ici

Importation des jeux de données

Découverte

Dans cette partie, nous découvrirons le jeu de données.

Visualisations réalisées

1. Fréquentation des gares
2. Carte de la fréquentation des gares en France Métropolitaine

[TODO] à compléter

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 ?

1. Fréquentation des gares

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

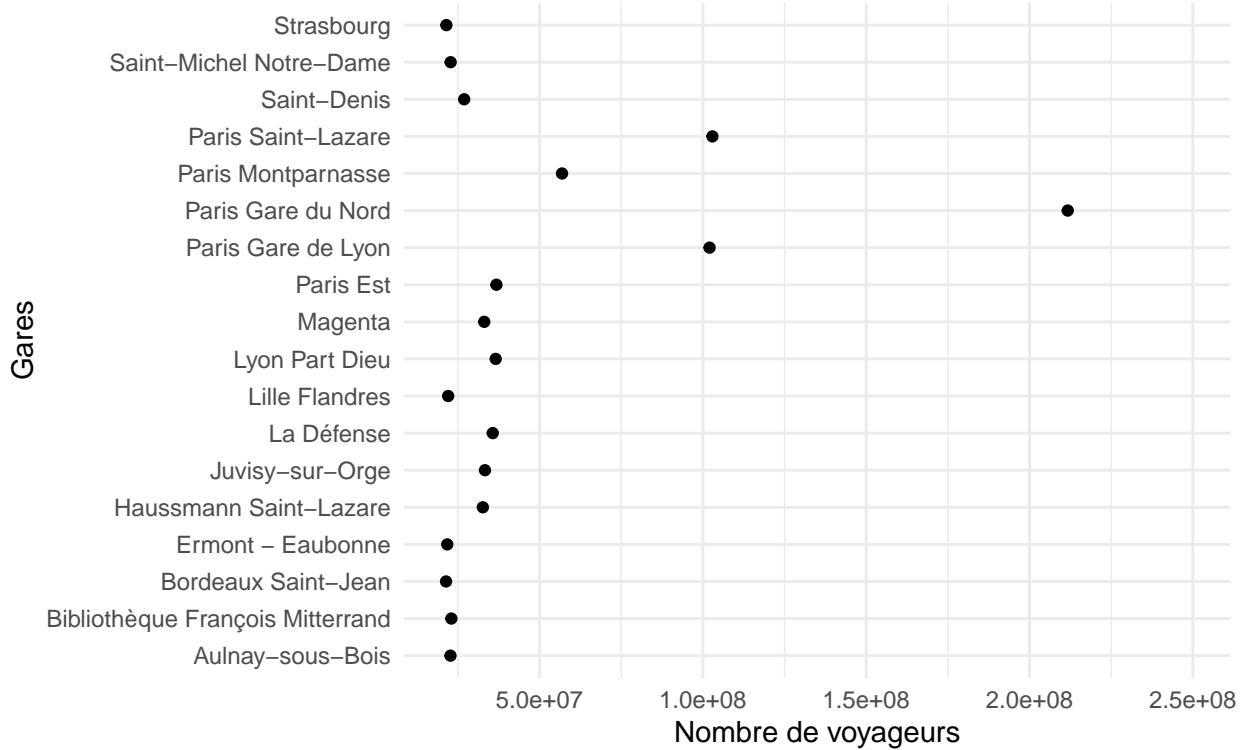
```
## # A tibble: 6 x 8
##   Gare      UIC    Code_Postal Zones_vac Année Personnes Voyageurs Département
##   <chr>     <chr>    <dbl> <chr>    <dbl>    <dbl>    <dbl> <chr>
## 1 Abbaretz 481614    44170 C       2022    40825    40825  44
## 2 Abbaretz 481614    44170 C       2021    27466    27466  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
```

Nous avons choisi d'utiliser des données discrètes (nombre de voyageurs) en abscisse et nominales (gares) en ordonnée. Afin de constater les différences entre les gares, une comparaison basée sur un bar chart a été réalisée ci-dessous.

De plus, 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. Ce filtre nous permet de faire ressortir les gares les plus fréquentées uniquement.

Fréquentation par gare (2022)

< minimum 20.000.000 de voyageurs >



Analyse On remarque que les gares avec le plus de fréquentation sont les gares parisiennes. En termes de positionnement, 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 étonnante si l'on utilise régulièrement le réseau RATP et SNCF en Ile de France. 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.

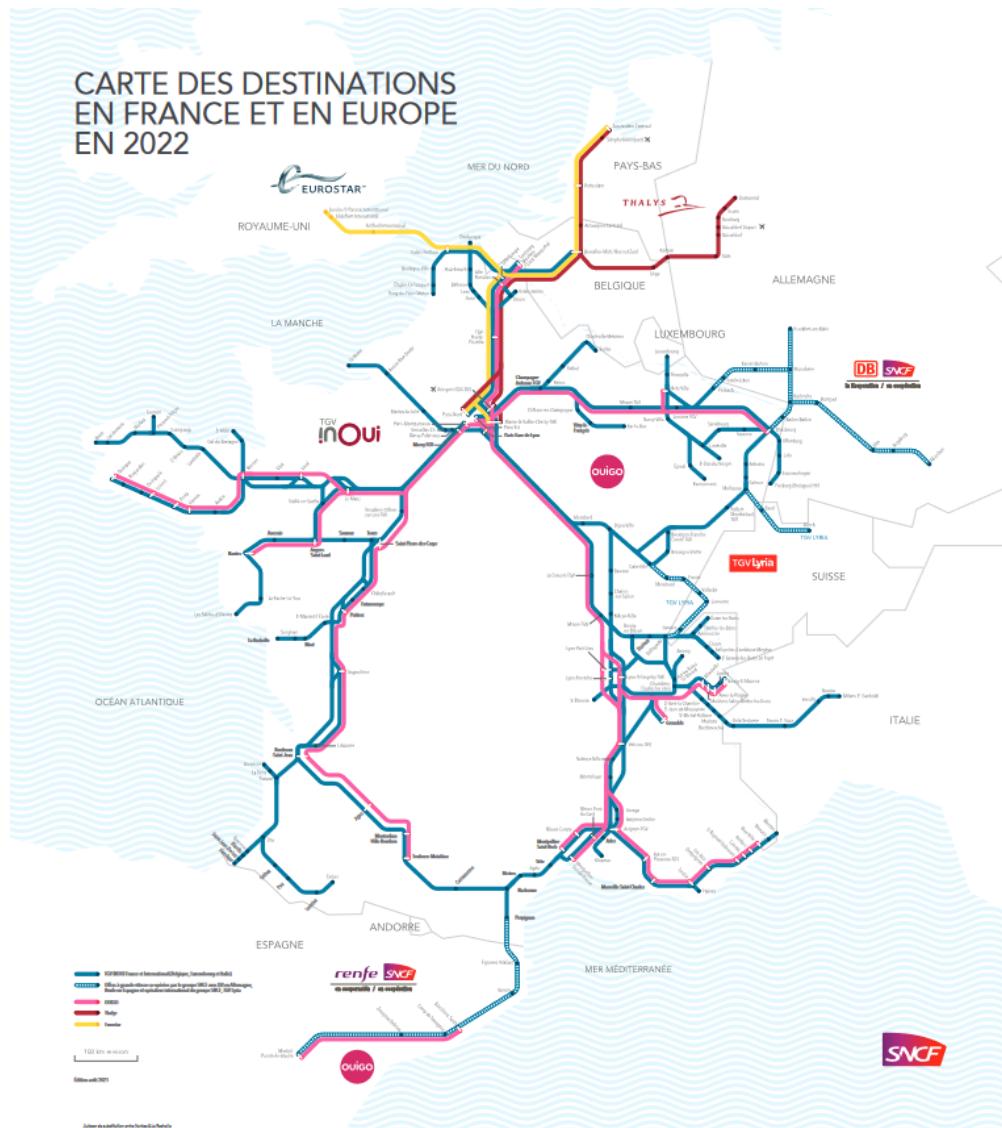


Figure 1. Carte des destinations en France et en Europe (2022)

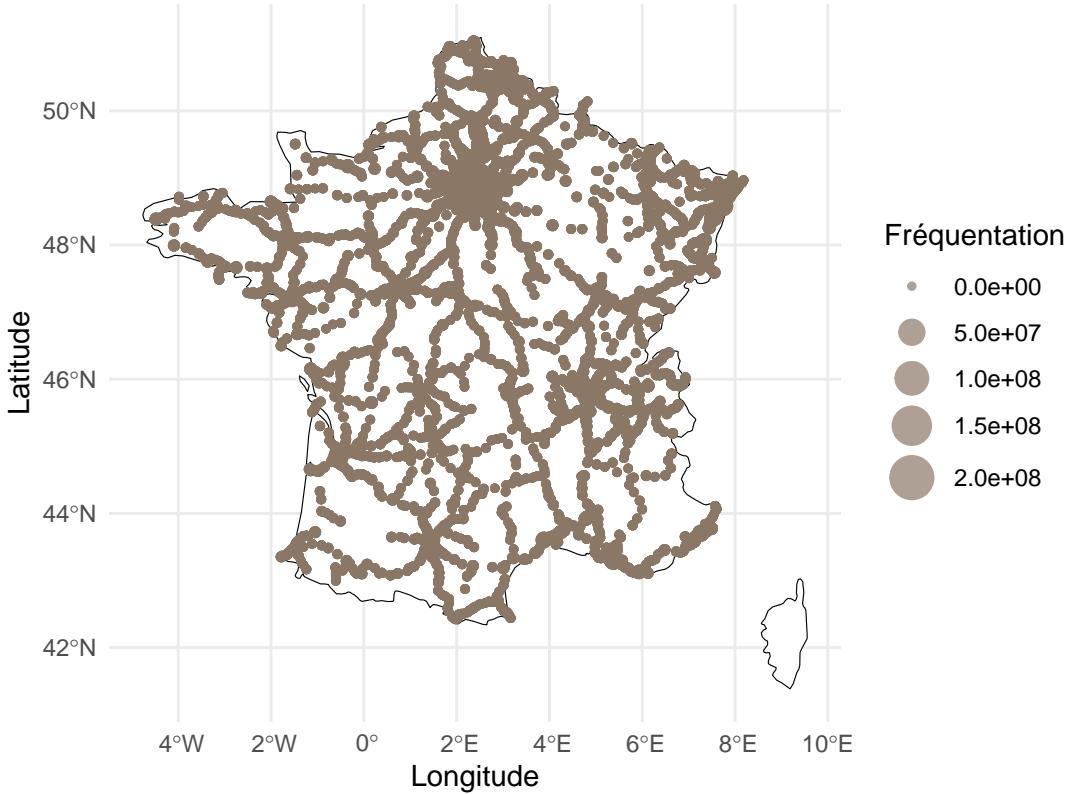
Source : <https://www.sncf-connect.com/aide/le-reseau-sncf-en-france-et-en-europe>

2. Carte de la fréquentation des gares en France Métropolitaine

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

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. De plus, on a ajouté de la couleur afin de faire ressortir des gares qui n'ont plus vraiment de fréquentation en 2022.

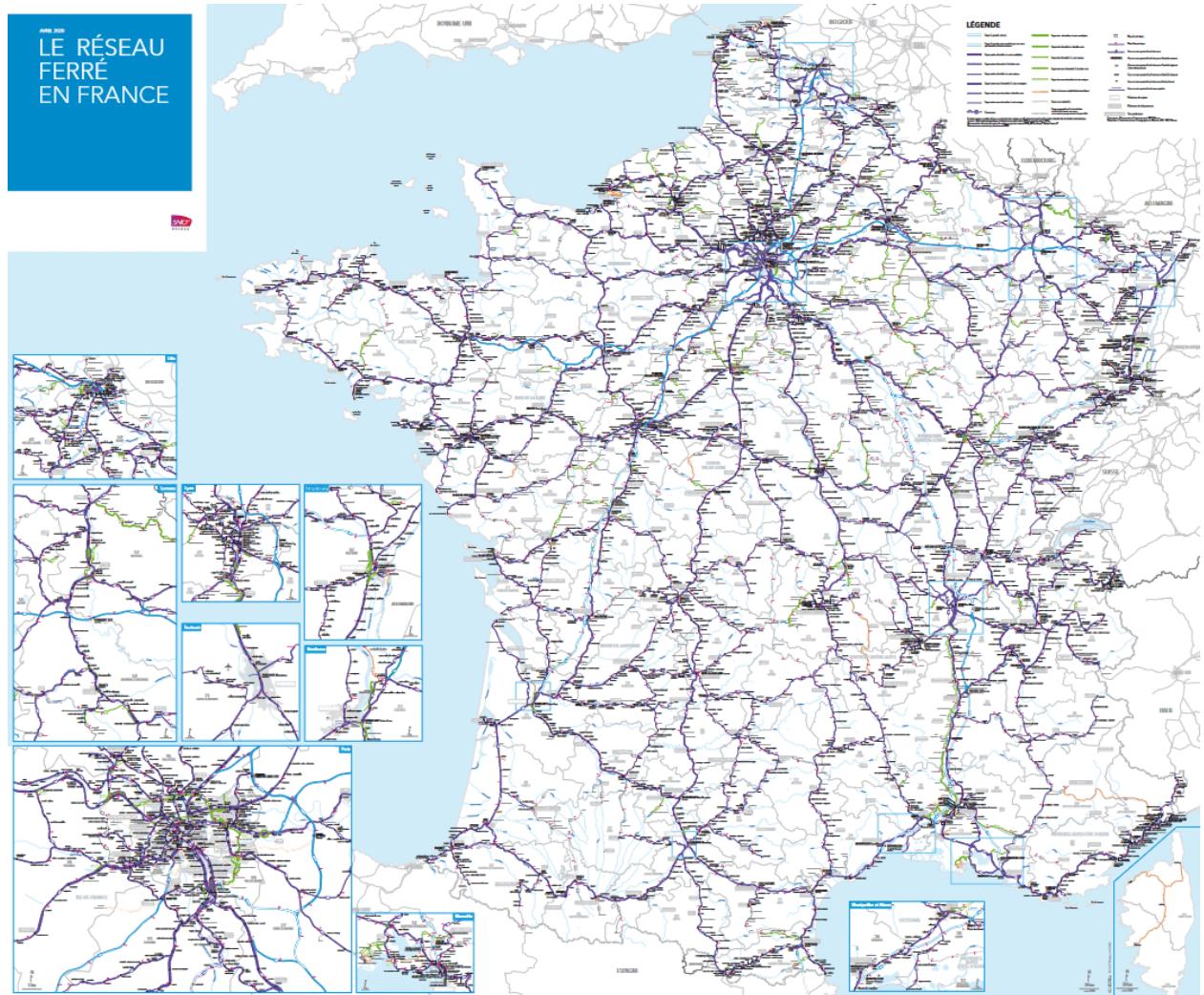
Fréquentation par gare (2022)



Analyse Au premier abord, on remarque que les gares tracent d'elles même, sur la carte, la majeur partie du réseau ferroviaire français (*Figure 2*). Cette visualisation est donc toujours assez proche de la réalité en 2022.

On remarque que la fréquentation des gares 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. Avec cette profondeur supplémentaire, cela nous permet de formuler de nouvelles conjectures : la population en périphérie des grandes villes fréquente généralement les gares du réseau pour des motifs professionnels (population active).

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.



Figure

2. Carte du réseau SNCF en France (2020)

Source : <https://www.sncf-connect.com/aide/le-reseau-sncf-en-france-et-en-europe>

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. Nombre de voyageurs total par département (2022)
2. Répartition du nombre de voyageurs dans les gares d'un même département. Exemple département (77)
3. Explorer la répartition par âge des passagers
4. Catégorie socio professionnelle
5. Nombre de voyageurs par année

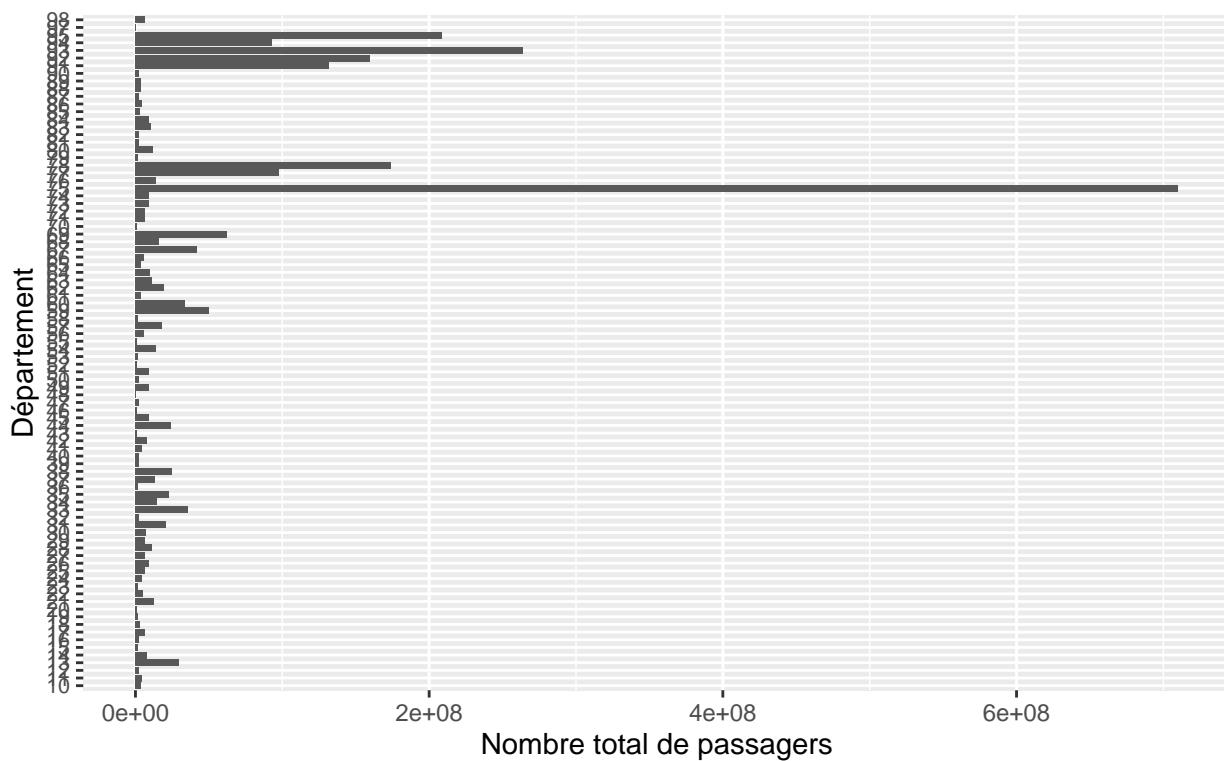
Dans un premier temps, nous analysons le nombre de voyageurs par département, pour nous intéresser ensuite à la proportion de voyageurs dans les gares au sein d'un même département.

Le nombre de voyageurs est-il bien repartis entre les gares d'un même département ?

1. Nombre de voyageurs total par département (2022)

Nombre de voyageurs total par département (2022)

<Vue Globale>



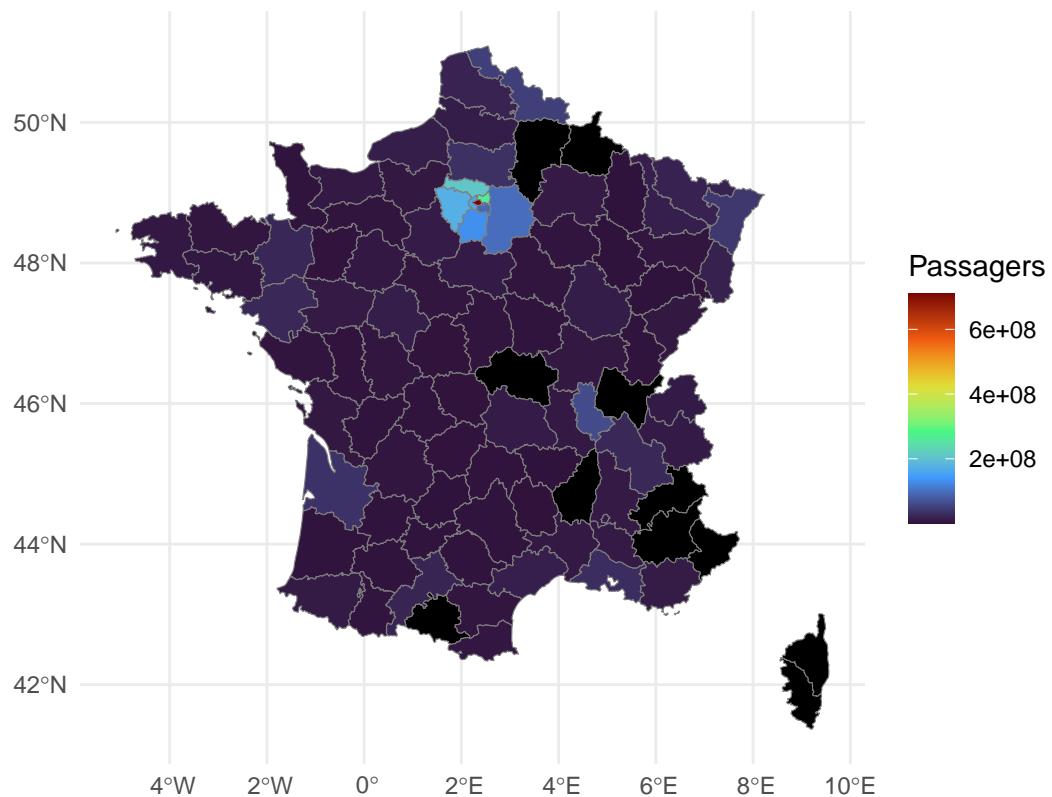
Comme le constat fait un peu plus haut, on observe que la région Ile de France (composée des départements 75, 77, 78, 91, 92, 93, 94) concentre la plupart de la fréquentation. Ce constat est tout à fait correct étant donné que la région concentre environ 12 millions de personnes.

On peut également créer une visualisation qui permettrait de connaître la fréquentation de chacune des régions.

On voit qu'il y a une très grande différence entre certains départements, on va essayer de le voir sur une carte.

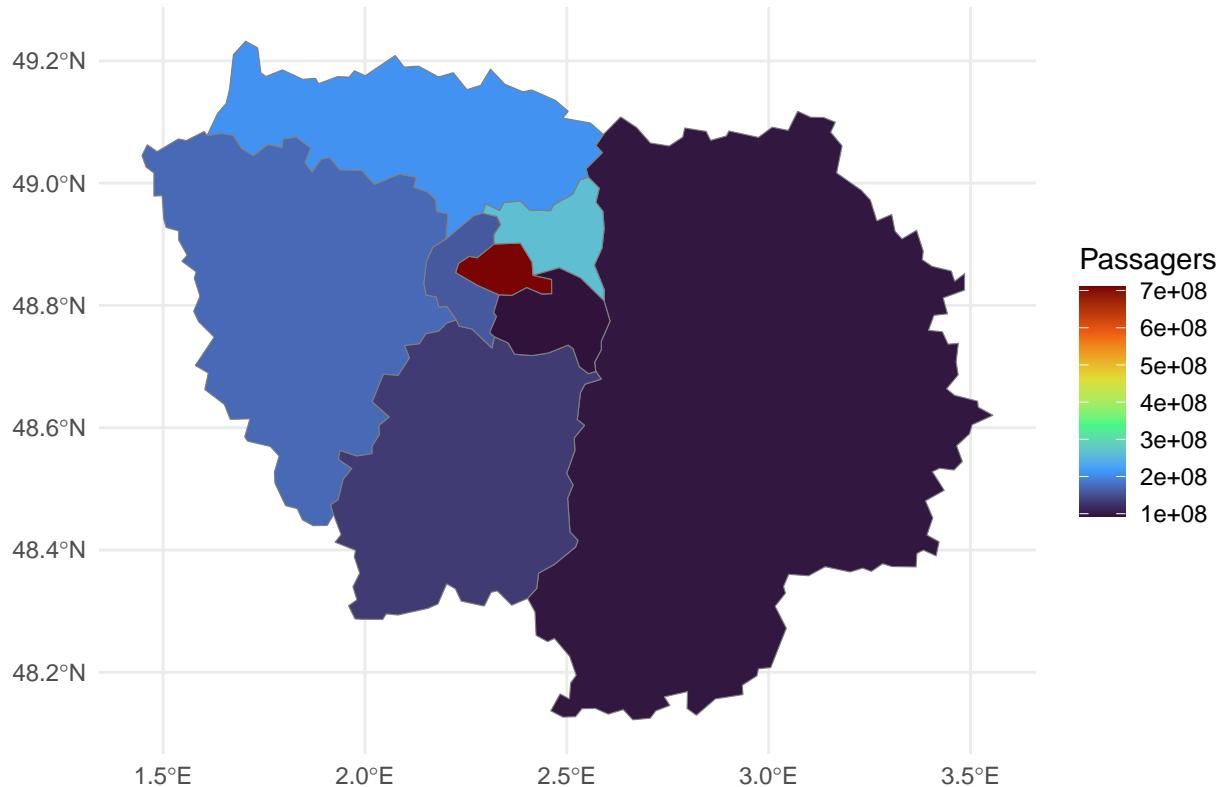
```
## Reading layer `departements-version-simplifiee` from data source
##   `https://raw.githubusercontent.com/gregoiredavid/france-geojson/master/departements-version-simplifiee`
##   using driver `GeoJSON'
## Simple feature collection with 96 features and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension: XY
## Bounding box: xmin: -5.103601 ymin: 41.36705 xmax: 9.559721 ymax: 51.0884
## Geodetic CRS: WGS 84
```

Fréquentation des gares par département en France (2022)



Justification

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



2. Répartition du nombre de voyageurs dans les gares d'un même département.Exemple département (77)

Grâce à cette visualisation, on peut connaître les gares les plus importantes en termes de fréquentation dans chacun des départements.

Dans le cas du département 77 et du premier jet de visualisation, on observe que la gare de Melun est très fréquentée comparé aux autres gares.

Au sein du 77, on remarque des disparités entre les gares. Des gares comme celle de Melun ou Chartrette concentre toute la fréquentation au détriment de gares comme Bois-le-Roi.

On peut expliquer cela notamment après quelques recherches car en effet le RER D désert Melun à la différence de petites gares comme celle de Bois-le-Roi.

Quel est le voyageur moyen de la SNCF ?

3. Explorer la répartition par âge des passagers

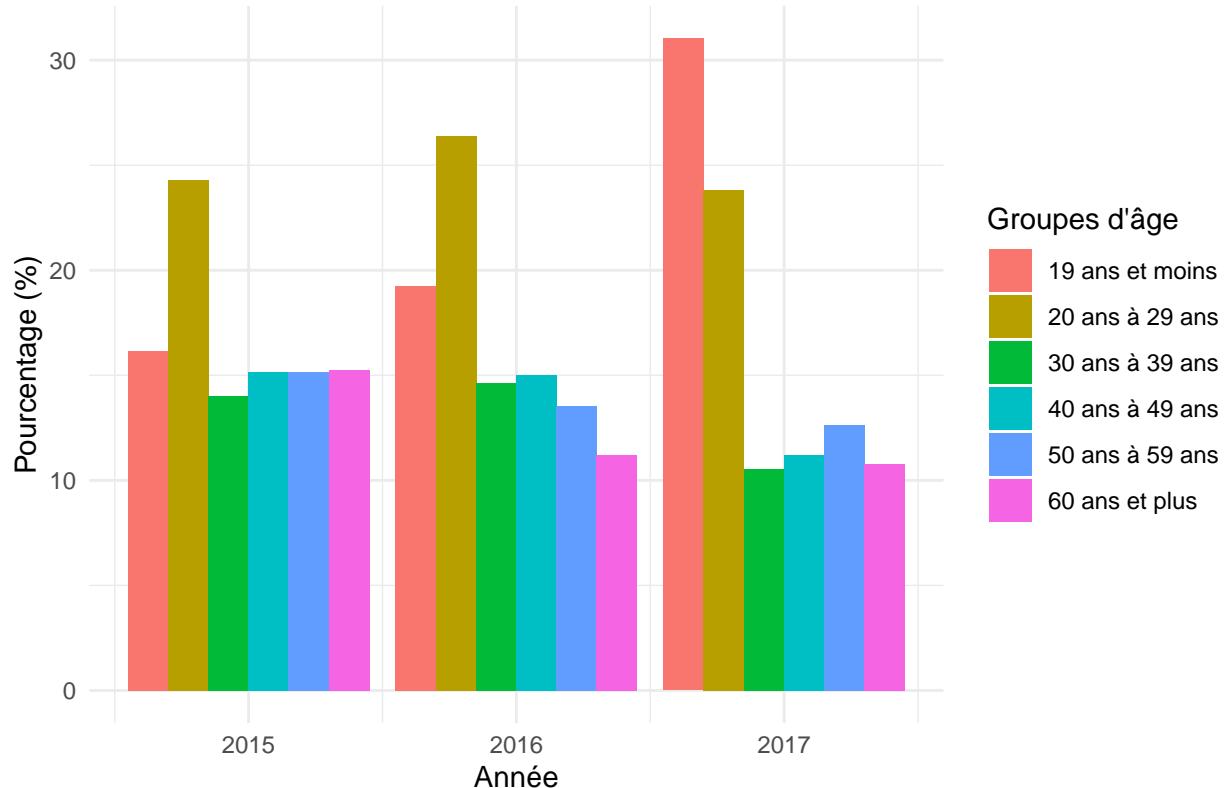
En raison de quelques problèmes dans l'ensemble de données, tels que des codes UIC incohérents et des variations dans les années de recensement et les données des stations, nous allons nous concentrer exclusivement sur la répartition par âge pour les années 2015, 2016 et 2017.

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 nombre total pour chaque année
L'étape suivante consiste à calculer le pourcentage des groupes d'âge ainsi que la cartographie.

Pourcentage de passagers par groupe d'âge et année



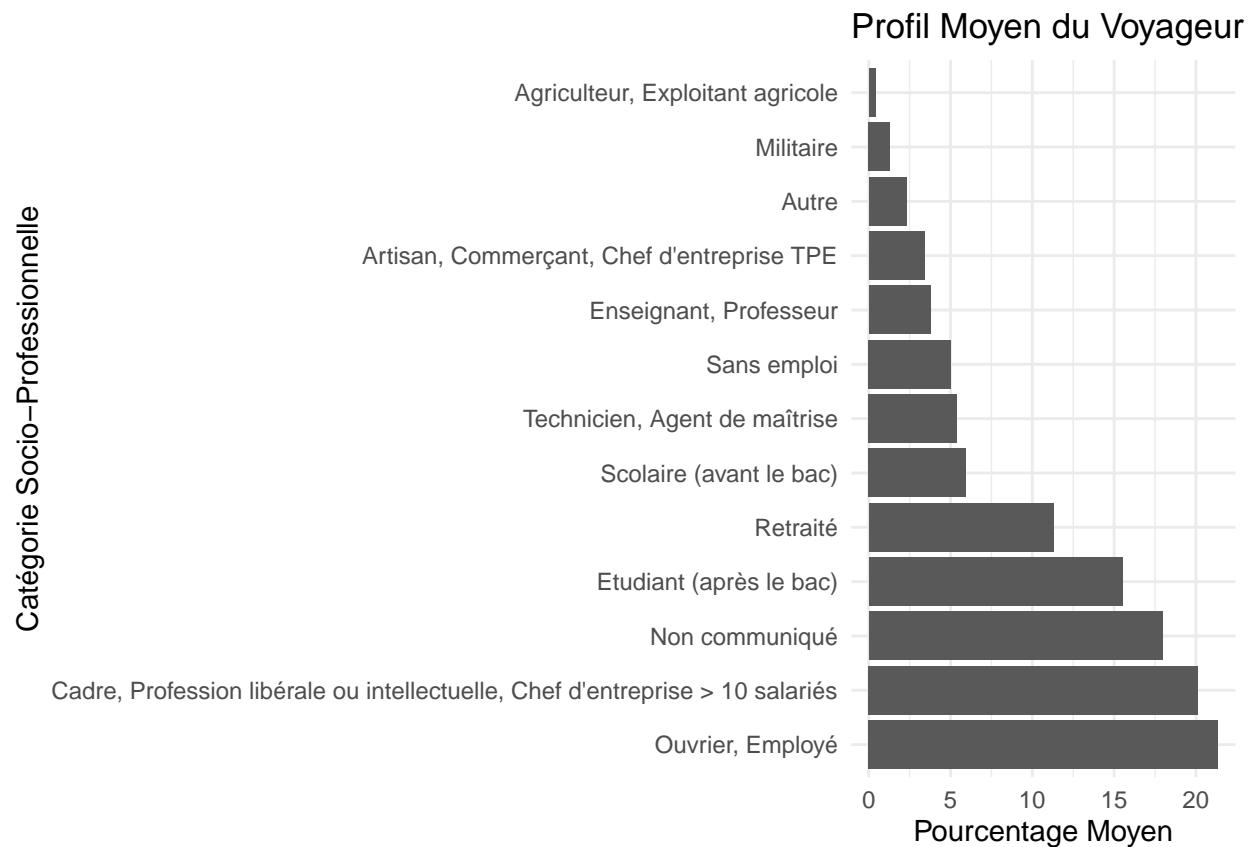
4. Catégorie socio professionnelle

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.

	Pourcentage_moyen
## # A tibble: 13 x 2	
## CSP	<dbl>
## <chr>	
## 1 Ouvrier, Employé	21.3
## 2 Cadre, Profession libérale ou intellectuelle, Chef d'entre~	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.



5. 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 20
##   `Nom de la gare`    `Code UIC` `Code postal` `Segmentation DRG`
##   <chr>                <dbl>        <dbl>       <chr>
## 1 Abbaretz            87481614     44170      C
## 2 Ablon-sur-Seine     87545269     94480      B
## 3 Achères Grand Cormier 87386052    78100      B
## 4 Acheux - Franleu    87316745     80560      C
## 5 Aigrefeuille le Thou 87485193     17290      C
## 6 Aigueperse           87734129     63260      C
## # i 16 more variables: `Total Voyageurs 2022` <dbl>,
## #   `Total Voyageurs + Non voyageurs 2022` <dbl>, `Total Voyageurs 2021` <dbl>,
## #   `Total Voyageurs + Non voyageurs 2021` <dbl>, `Total Voyageurs 2020` <dbl>,
## #   `Total Voyageurs + Non voyageurs 2020` <dbl>, `Total Voyageurs 2019` <dbl>,
## #   `Total Voyageurs + Non voyageurs 2019` <dbl>, `Total Voyageurs 2018` <dbl>,
## #   `Total Voyageurs + Non voyageurs 2018` <dbl>, `Total Voyageurs 2017` <dbl>,
## #   `Total Voyageurs + Non voyageurs 2017` <dbl>, ...
```

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

Maintenant intéressons nous aux nombres de voyageurs par segmentation DRG.

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

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 .

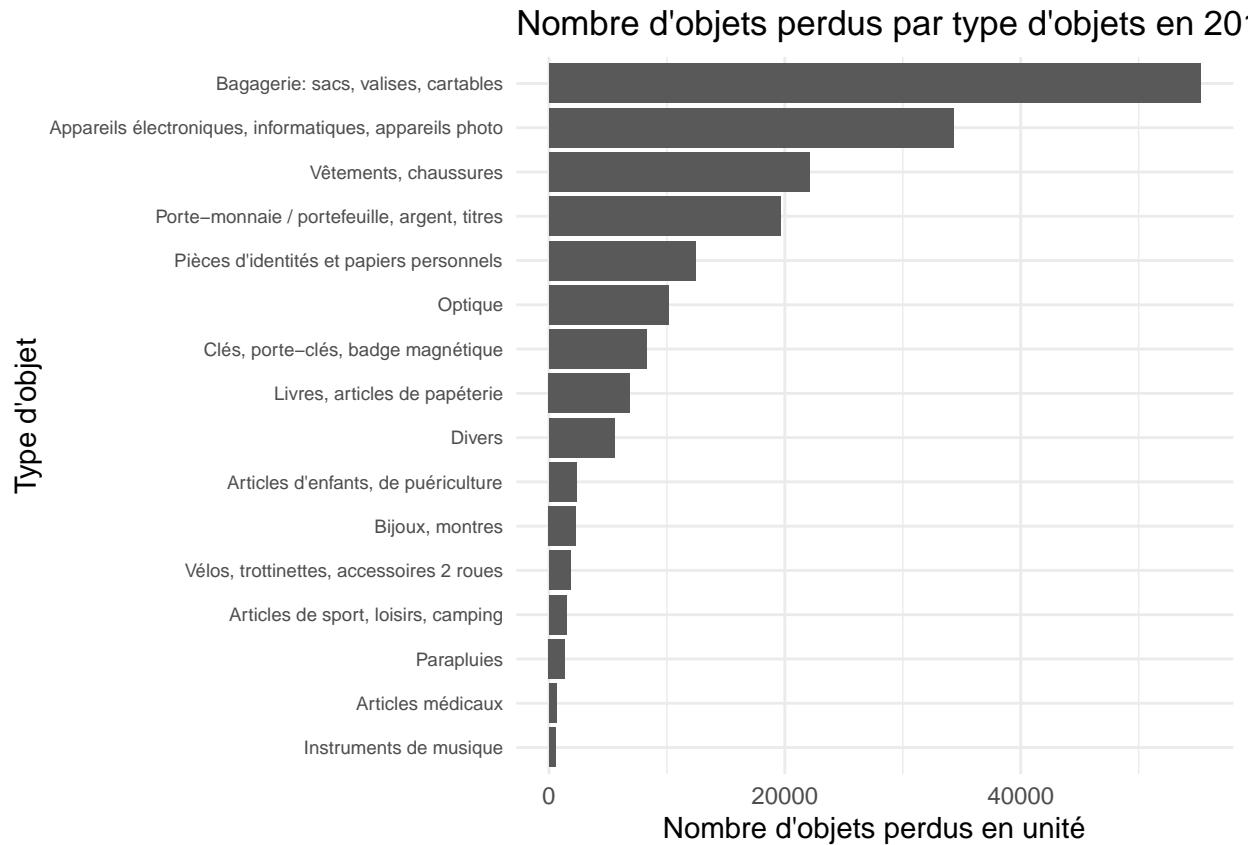
On distingue aussi que de nombreuses gares ne sont pas présentes.

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 recouplements 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 dataframme.

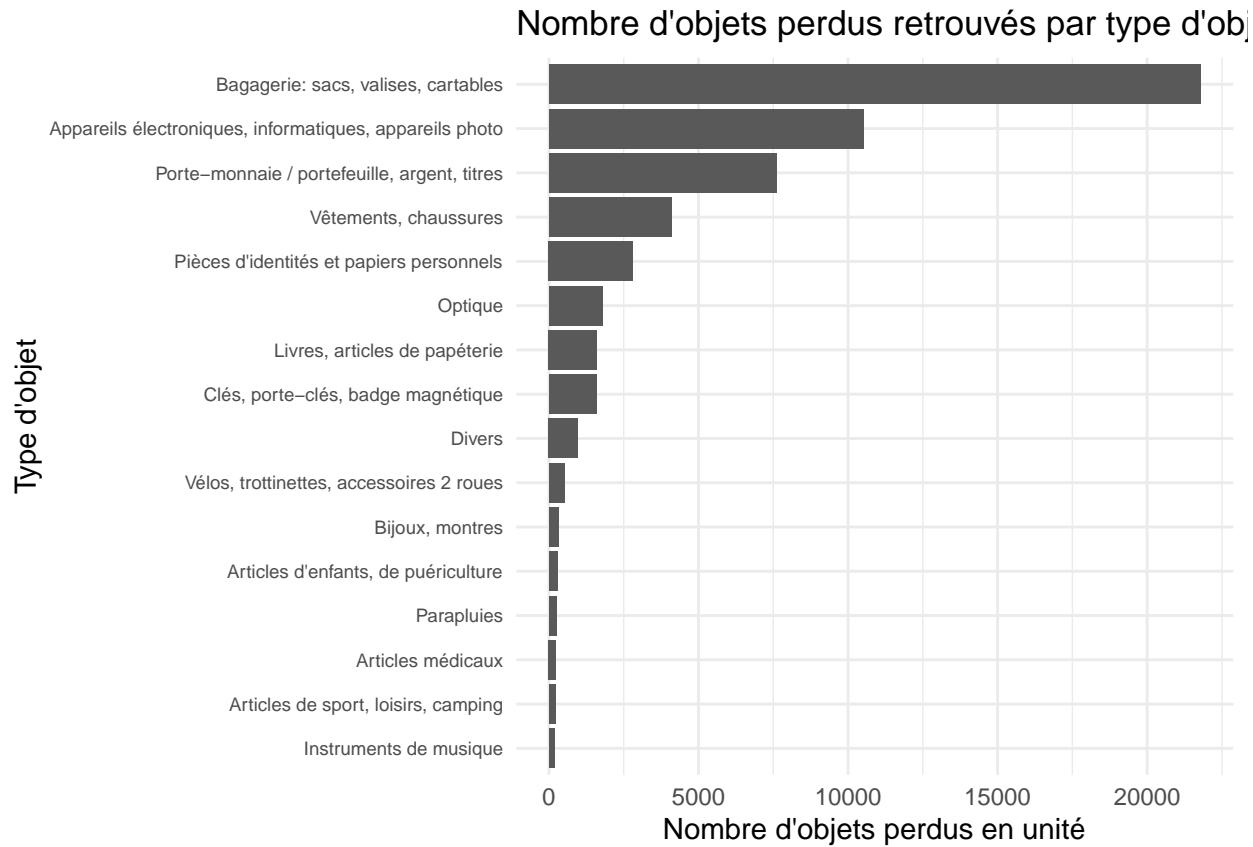
Découvrons ce que celui-ci nous cache dans la répartition des objets perdus :



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.

Quelles sont les chances de retrouver un objet perdu ?

3. Probabilité de retrouver un objet perdu

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.

Quelles sont les chances de retrouver un objet en fonction de son type ?

4. Probabilité de retrouver un objet selon son type

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.

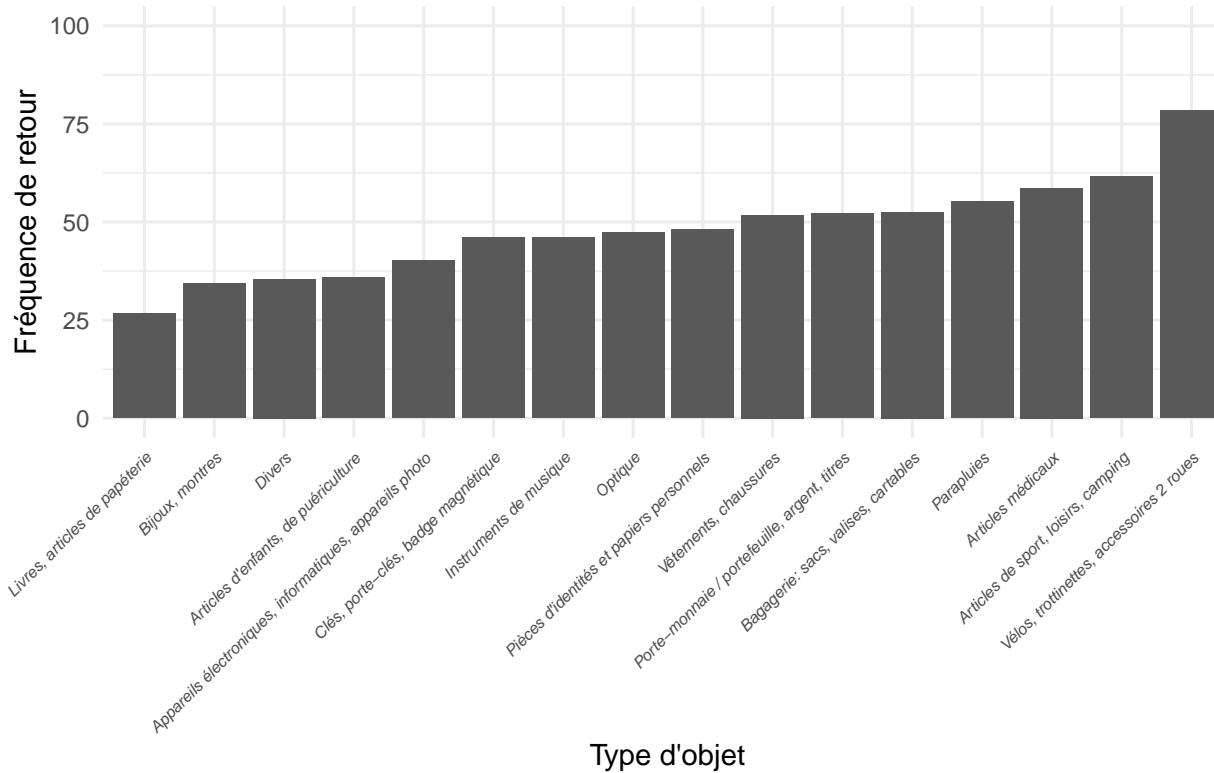
N.B. la “nature” dans l'intitulé de la question a été changé en type car la nature entrée dans le formulaire des objets trouvés est trop spécifique.

```

## [1] "type_obj_trouves" "Freq"

```

Fréquence de rendu d'objets perdus par type



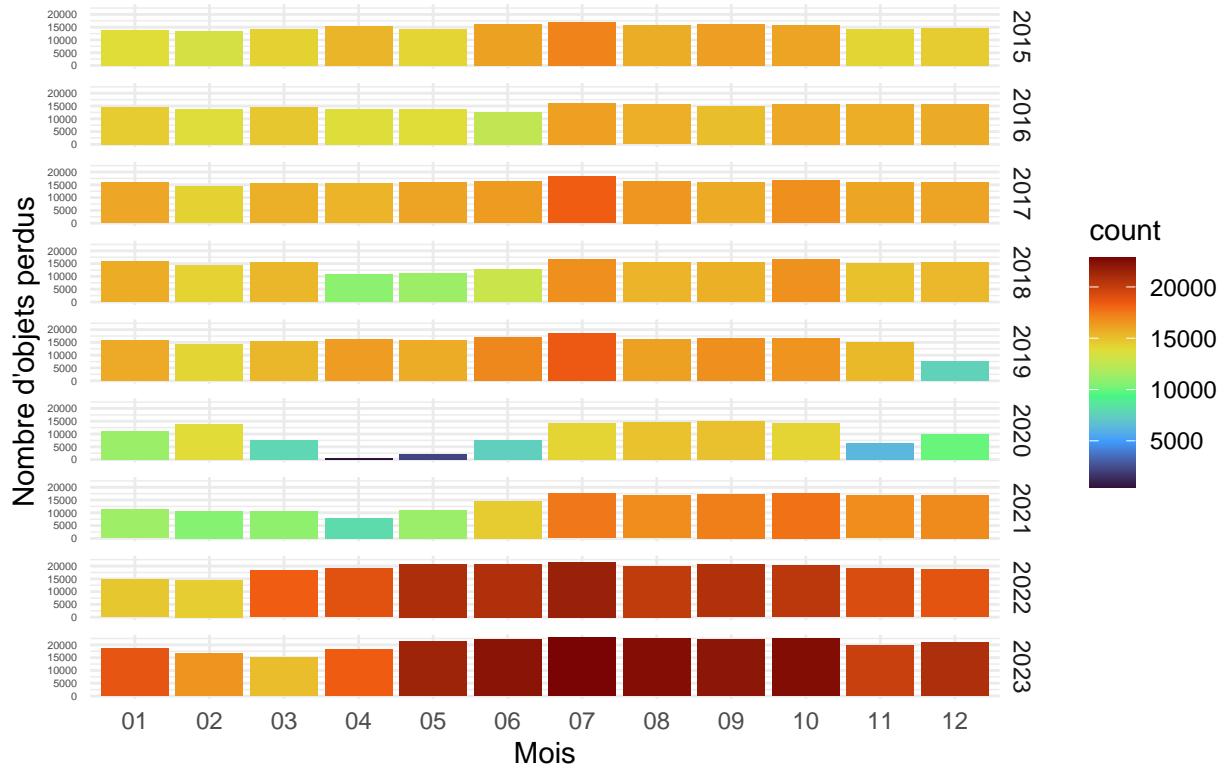
Plus de 50 points distinguent la fréquence de retour des vélos, trottinettes et accessoires de celle des livres et articles de papeterie. Les vélos, trottinettes et accessoires est une catégorie qui est bien au-dessus du second article le plus rendu de + de 10 points. Je m'attendais clairement à un taux 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 croisant d'autres études portant sur le sujet.

4. 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.

Nombre d'objets perdus par mois et par année



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).

Conclusion

Notre avis sur le projet

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