

Atelier pratique sur R

L'analyse d'enquête

Groupe ElementR
Marion Albertelli, Joséphin Béraud
14 février 2023



Programme de l'atelier

Objectifs de l'atelier

Présentation de l'enquête SNCF et des objectifs de recherche

Traitements de base d'une enquête et packages associés

Prise en main des données

Pistes d'exploration

Des ressources utiles

Et si on faisait une petite carte ?

Objectifs de l'atelier

Présentation de l'enquête SNCF et des objectifs de recherche

Traitements de base d'une enquête et packages associés

Prise en main des données

Pistes d'exploration

Des ressources utiles

Et si on faisait une petite carte ?

Objectifs de l'atelier

Pratiquer, pratiquer, pratiquer... pour souffrir collectivement.

Explorer, nettoyer des données d'enquêtes, contrôler la structure de son échantillon...

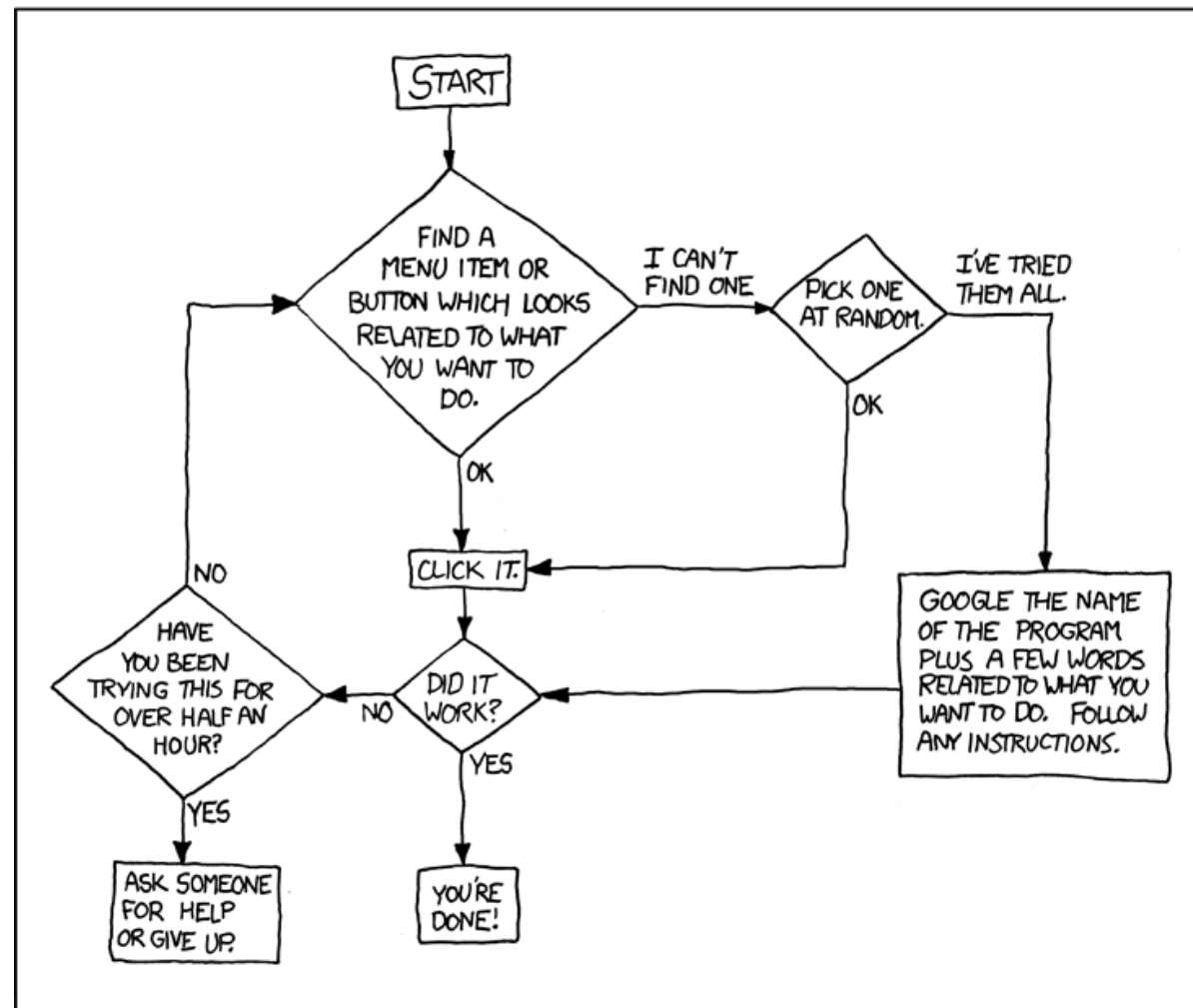
Voir des fonctions simples, mais jouissives

Partir d'une page blanche sans prendre ses doigts à son cou

Mettre à son service les connaissances acquises

DEAR VARIOUS PARENTS, GRANDPARENTS, CO-WORKERS,
AND OTHER "NOT COMPUTER PEOPLE."

WE DON'T MAGICALLY KNOW HOW TO DO EVERYTHING IN EVERY
PROGRAM. WHEN WE HELP YOU, WE'RE USUALLY JUST DOING THIS:



PLEASE PRINT THIS FLOWCHART OUT AND TAPE IT NEAR YOUR SCREEN.
CONGRATULATIONS; YOU'RE NOW THE LOCAL COMPUTER EXPERT! 4

Objectifs de l'atelier

Présentation de l'enquête SNCF et des objectifs de recherche

Traitements de base d'une enquête et packages associés

Prise en main des données

Pistes d'exploration

Des ressources utiles

Et si on faisait une petite carte ?

Présentation de l'enquête SNCF et des objectifs de recherche

1 sujet : une thèse sur l'insertion urbaine des gares métropolitaines françaises.

4 échelles : la métropole, le quartier, la gare et l'individu.

3 approches méthodologiques :

- 1 typologie de 30 gares,
- 6 focus gares,
- 1 enquête de voyageurs.



3 moments du parcours regardés

1. Le trajet du point de départ à la gare
2. La rupture de charge
3. Le parcours en gare

Présentation de l'enquête SNCF et des objectifs de recherche



Cible

Ensemble des usagers de 6 gares : 3 gares parisiennes (Gare de Lyon, Bercy, Gare de l'Est), 2 gares lilloises (Lille Flandres et Lille Europe) et Rouen Rive Droite



Nombre d'enquêtes réalisées

Gare enquêtée	Nombre d'enquêtes
Paris Gare de l'Est	300
Paris Bercy	291
Paris Gare de Lyon	354
Lille Europe	336
Lille Flandres	390
Rouen Rive Droite	401

Pour un total de 2072 enquêtes



Mode de recueil

En gare, questionnaire administré par des enquêteurs à l'aide de tablettes



Dates et horaires d'enquête :

Du 26/05/2021 au 12/06/2021, JOB et Week-end
Heures pleines AM : 6h à 9h30
Heures pleines PM : 16h30 à 19h
Heures creuses : 9h30 à 16h30

Présentation de l'enquête SNCF et des objectifs de recherche

Fichiers à disposition

- 1 [base de données](#) au format CSV avec un échantillon de l'enquête :
 - 3 gares sur les 6 : Rouen Rives Droite, Lille Flandres, Paris Bercy
 - 1 sélection à partir des montants dans un 1 train
 - 24 questions/informations sur les 49 qui ont été posées/récoltées
- 2 fichier Excel « [Métadonnées](#) »
- 1 fichier complémentaire avec [un extrait de la typologie](#) pour croiser les données

Objectifs de l'atelier

Présentation de l'enquête SNCF et des objectifs de recherche

Traitements de base d'une enquête et packages associés

Prise en main des données

Pistes d'exploration

Des ressources utiles

Et si on faisait une petite carte ?

Traitements de base d'une enquête et packages associés

Le **tri à plat** : traitement statistique de base d'une enquête qui consiste à calculer les distributions des individus selon une variable.

Le **tri croisé** : mise en relation de plusieurs variables.

L'application de filtres sur les tris à plat ou les tris croisés permet de centrer l'analyse sur une partie de l'échantillon.

Traitements de base d'une enquête et packages associés

Aller plus loin avec ...

Le **tri combiné ou multiple**, qui s'applique à des tableaux de questions identiques (échelles, notes...)

Le calcul de l'**écart-type ou variance**, pour mesurer la dispersion des données

Des tests statistiques, comme celui du **Chi²** qui permet de calculer le rapport d'indépendance entre deux variables.

Des méthodes **d'analyse multivariée**, type analyse en composantes principales (ACP), analyse factorielle des correspondances (AFC), ou régression simple ou multiple

Traitements de base d'une enquête et packages associés



ggplot2

ggplot()

dplyr

group_by()
mutate()

Manipulation des données &
représentation graphique



tbl_summary()

Tableau de sortie simple, beau et propre

'questionr' freq()

Exploration des données &
Interface clic-bouton



density()
boxplot()
ecdf()
mosaicplot()

Toutes les manipulations &
représentation graphique

Feuille de triche

Dplyr

Ggplot2

Gtsummary

Objectifs de l'atelier

Présentation de l'enquête SNCF et des objectifs de recherche

Traitements de base d'une enquête et packages associés

Prise en main des données

Pistes d'exploration

Des ressources utiles

Et si on faisait une petite carte ?

Prise en main des données

Une bonne pratique : Créer un projet R dans un dossier avec l'ensemble des fichiers et des données + 1 fichier script de façon à pouvoir travailler sur le temps long

Atelier_ElementR_140223 > ElementR_140223 >

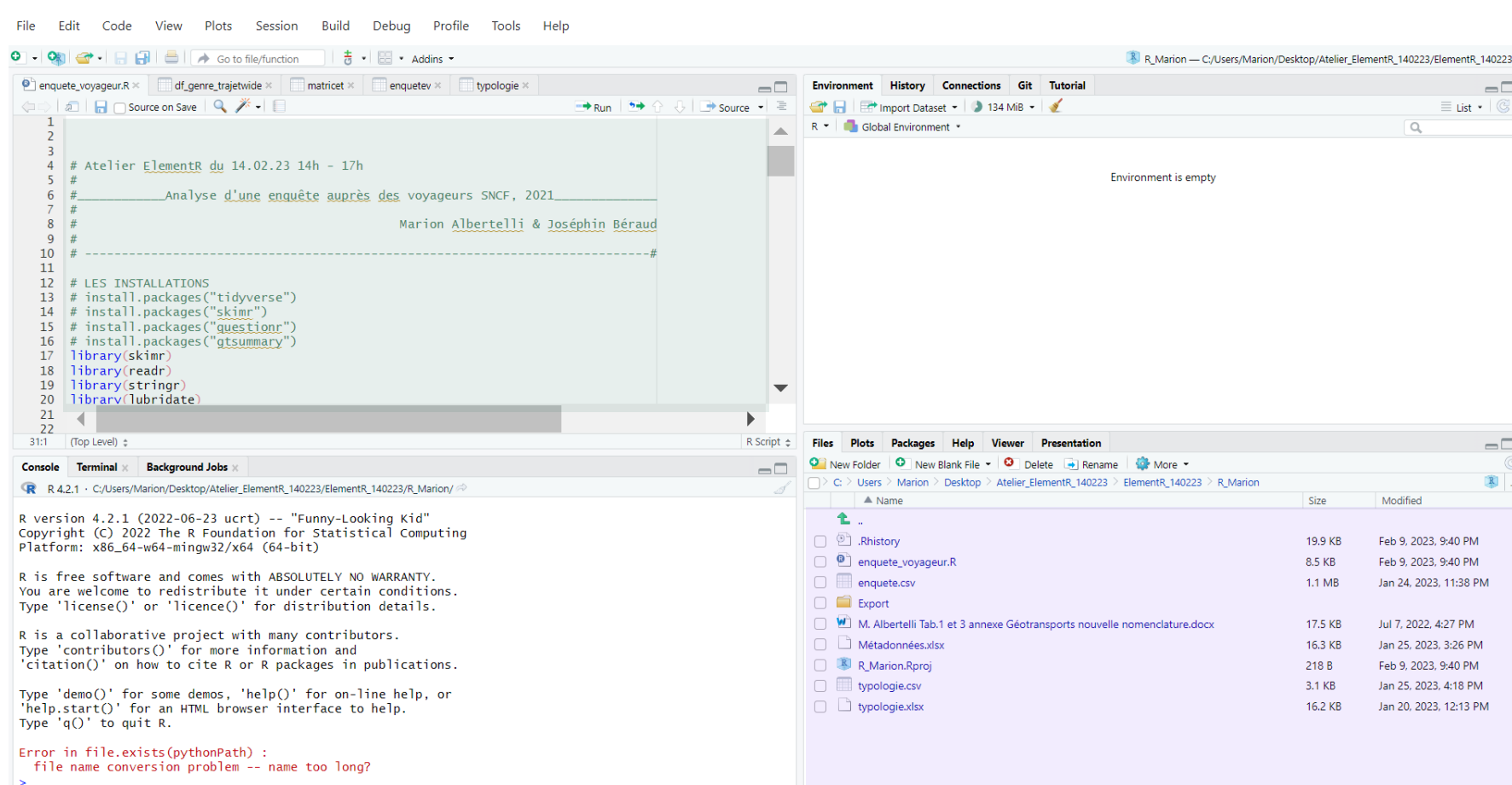
Nom	Modifié le	Type	Taille
data	24/01/2023 22:37	Dossier de fichiers	
R_Marion	02/02/2023 14:30	Dossier de fichiers	
SNCF_Rendu	18/01/2023 12:10	Dossier de fichiers	

Atelier_ElementR_140223 > ElementR_140223 > R_Marion

Nom	Modifié le	Type	Taille
Données	09/02/2023 21:31	Dossier de fichiers	
Export	09/02/2023 21:31	Dossier de fichiers	
.Rhistory	09/02/2023 19:27	Fichier R HISTORY	20 Ko
enquete_voyageur	09/02/2023 18:55	Fichier R	9 Ko
R_Marion	09/02/2023 18:56	R Project	1 Ko

Prise en main des données

Une bonne pratique : Créer un projet R dans un dossier avec l'ensemble des fichiers et des données + 1 fichier script de façon à pouvoir travailler sur un temps long



The screenshot displays the RStudio interface. The main editor window shows an R script with the following content:

```
1  
2  
3  
4 # Atelier ElementR du 14.02.23 14h - 17h  
5 #  
6 # ----- Analyse d'une enquête auprès des voyageurs SNCF, 2021  
7 #  
8 #                               Marion Albertelli & Joséphin Béraud  
9 #  
10 # -----  
11 #  
12 # LES INSTALLATIONS  
13 # install.packages("tidyverse")  
14 # install.packages("skimmer")  
15 # install.packages("questionr")  
16 # install.packages("gtsummary")  
17 library(skimmer)  
18 library(readr)  
19 library(stringr)  
20 library(lubridate)  
21  
22
```

The Environment pane on the right shows "Global Environment" and "Environment is empty". The File Explorer at the bottom right shows the project directory structure:

Name	Size	Modified
..		
.Rhistory	19.9 KB	Feb 9, 2023, 9:40 PM
enquete_voyageur.R	8.5 KB	Feb 9, 2023, 9:40 PM
enquete.csv	1.1 MB	Jan 24, 2023, 11:38 PM
Export		
M. Albertelli Tab.1 et 3 annexe Géotransports nouvelle nomenclature.docx	17.5 KB	Jul 7, 2022, 4:27 PM
Métadonnées.xlsx	16.3 KB	Jan 25, 2023, 3:26 PM
R_Marion.Rproj	218 B	Feb 9, 2023, 9:40 PM
typologie.csv	3.1 KB	Jan 25, 2023, 4:18 PM
typologie.xlsx	16.2 KB	Jan 20, 2023, 12:13 PM

The Console pane at the bottom left shows the R version information and a warning message:

```
R version 4.2.1 (2022-06-23 ucrt) -- "Funny-Looking Kid"  
Copyright (c) 2022 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
Error in file.exists(pyhtonPath) :  
  file name conversion problem -- name too long?  
>
```

Prise en main des données

Import, nettoyage, mise en forme des données :

Prise en main des données

Import, nettoyage, mise en forme des données :

```
# - L'IMPORT DES DONNEES -
```

```
enquetev <- read.csv("enquete.csv", header=TRUE, sep=";", fileEncoding = "latin1")  
typologie <- read.csv("typologie.csv", header = TRUE, sep = ";", fileEncoding = "latin1")
```

```
# Vérifier les gares enquêtées + leur orthographe  
sort(unique(enquetev$RS3))
```

```
# Suppression des majuscules dans les noms de gare.
```

```
enquetev$RS3 <- tolower(enquetev$RS3)
```

```
# ou pour 1 seul nom (mais moins utile ici)
```

```
enquetev$RS3 <- gsub("ParisGaredeLyon", "parisgaredelyon", enquetev$RS3)
```

```
# - DECOUVERTE DES DONNEES -
```

```
# Quelle est la nature d'un CSV qu'on importe dans R avec cette fonction là ?
```

```
class(enquetev)
```

```
# pour avoir le nom des colonnes
```

```
names(enquetev)
```

```
# Prévisualisation de mon tableau de données
```

```
View(enquetev)
```

```
# ou dans la console : les 5 premières lignes
```

```
head(enquetev)
```

```
# Combien de lignes et de colonnes comportent mon tableau ?
```

```
dim(enquetev)
```

```
# ou
```

```
ncol(enquetev)
```

```
nrow(enquetev)
```

```
# De quelle nature sont mes données ?
```

```
str(enquetev)
```

Prise en main des données

	Total	Paris Gare de l'Est	Paris Bercy	Paris Gare de Lyon	Lille Europe	Lille Flandres	Rouen Rive Droite
<i>Bases</i>	2072	300	291	354	336	390	401
SEXE							
Hommes	42%	44%	43%	49%	37%	39%	41%
Femmes	58%	56%	57%	51%	63%	61%	59%
AGE							
Moins de 18 ans	3%	1%	1%	1%	3%	7%	4%
18 à 25 ans	31%	23%	24%	16%	35%	45%	36%
26 à 35 ans	20%	25%	19%	27%	19%	18%	14%
36 à 45 ans	14%	17%	17%	18%	14%	10%	12%
46 à 55 ans	14%	16%	19%	18%	13%	9%	13%
56 à 65 ans	9%	10%	8%	12%	7%	5%	10%
Plus de 65 ans	9%	8%	12%	8%	10%	6%	10%
PCS							
<i>Bases (Hors NSP)</i>	2047	296	287	350	333	386	395
PCS+	36%	41%	40%	51%	37%	24%	29%
Agriculteur, exploitant ou ouvrier agricole	0,1%	0,0%	0,3%	0,0%	0,3%	0,0%	0,0%
Artisans, commerçant, chef d'entreprise	2%	2%	2%	5%	2%	1%	2%
Profession intellectuelle, libérale ou cadre	18%	24%	20%	25%	18%	11%	14%
Profession intermédiaire	16%	15%	18%	21%	16%	12%	13%
PCS-	24%	30%	27%	25%	20%	22%	23%
Employés	17%	24%	19%	17%	17%	16%	15%
Ouvriers	7%	6%	8%	9%	3%	6%	8%
RETRAITÉS	11%	11%	11%	12%	13%	8%	11%
ÉTUDIANTS / SCOLAIRES	24%	16%	17%	9%	25%	39%	31%
Etudiants, stagiaires	23%	16%	17%	9%	25%	38%	31%
Collégiens, lycéens	0,3%	0,3%	0,3%	0,0%	0,3%	1%	0,3%
AUTRES ET DEMANDEURS D'EMPLOI	5%	2%	5%	3%	5%	7%	6%
Demandeurs d'emploi	4%	1%	3%	2%	4%	5%	5%
Hommes ou femmes au foyer	1%	1%	0,0%	1%	0,3%	1%	1%
Autres inactifs	0,4%	0,0%	1%	0,0%	1%	1%	0,3%

Nb. : pour la pondération
d'une enquête
Package Survey

Prise en main des données

L'objectif en termes de structure de l'échantillon : un même nombre d'enquêtés par mode (50/mode/gare, montants + descendants) :

- Transports en commun (bus, métro, RER, autre train)
- Voiture (voiture particulière, taxi, VTC, voiture louée, covoiturage, autopartage)
- 2RM (motos, scooters, en libre-service ou pas)
- Mobilités douces (vélos, trottinettes, mono-roue, ... ; électriques ou pas ; en libre-service ou pas)
- Piéton (si unique mode d'accès à la gare)

=> Vérifier la structure de l'échantillon à ce niveau-là, car ça n'a pas été fait :
Tableau de fréquence.

Prise en main des données

Le tableau de fréquence :

```
# Tableau de fréquence nb enquêtés/mode  
freq(enquetev$Q4)
```

```
# Tableau de fréquence par mode et par gare  
df_enq_mode_gare <-enquetev %>%  
  filter(RS3 %in% c("RouenRiveDroite", "LillesFlandres","ParisgaredeBercy")) %>%  
  group_by(RS3, Q4) %>%  
  summarise(Individu = n())%>%  
  drop_na()
```

Prise en main des données

1 autre porte d'entrée dans les données : Calculer les temps de trajets moyens par gare, puis représentez-les sous forme de graphique avec ggplot

=> Défi le faire en moins de 20min !

=> Avant de se lancer dans R, et le code, comment pourrait-on décomposer les étapes à suivre ?

Prise en main des données

1 autre porte d'entrée dans les données : Calculer les temps de trajets moyens par gare, puis représentez-les sous forme de graphique avec ggplot

=> Défi le faire en moins de 20min !

=> Avant de se lancer dans R, et le code, comment pourrait-on décomposer les étapes à suivre ?



Prise en main des données

Calculer les temps de trajets moyens par gare, puis représenter les sous forme de graphique avec ggplot

Identification des packages

1

```
install.packages()  
library()
```

Vérification du typage des données

2

```
class()
```

Calcul des moyennes / gare

3

```
moy_trajet <- enquetev %>%  
  group_by() %>%  
  mutate( = round(mean(), )) %>%  
  distinct()
```

Représentation graphique

4

```
graph_trajet_gare <- ggplot(data=, aes(x=, y=)) +  
  geom_bar(stat="identity", color="red", fill="blue") +  
  geom_text(aes(label=, vjust=1.6, color="white", size=3.5)) +  
  theme_minimal() +  
  xlab() +  
  ylab()
```

Prise en main des données

Calculer les temps de trajets moyens par gare, puis représenter les sous forme de graphique avec ggplot

Identification des packages

1

```
# install.packages("tidyverse")  
library(tidyverse)
```

Vérification du typage des données

2

```
class(enquetev$Q12)
```

Calcul des moyennes / gare

3

```
moy_trajet <- enquetev %>%  
  group_by(RS3) %>%  
  mutate(moy = round(mean(Q12, na.rm = TRUE), 0)) %>%  
  distinct(RS3, moy)
```

Représentation graphique

4

```
graph_trajet_gare <- ggplot(data=moy_trajet, aes(x=RS3, y=moy)) +  
  geom_bar(stat="identity", color="red", fill="blue") +  
  geom_text(aes(label=moy, vjust=1.6, color="white", size=3.5)) +  
  theme_minimal() +  
  xlab("Gare") +  
  ylab("Temps de trajet moyen en minute")
```


Objectifs de l'atelier

Présentation de l'enquête SNCF et des objectifs de recherche

Traitements de base d'une enquête et packages associés

Prise en main des données

Pistes d'exploration

Des ressources utiles

Et si on faisait une petite carte ?

Pistes d'exploration

Quelques fonctions pour **vous lancer**, donner de **nouvelles idées** ou tout simplement **produire en masse** des graphiques pour la thèse de Marion...

Pistes d'exploration

Analyses univariées avec questionr

Quelques indicateurs

mean (moyenne),
sd (écart-type),
min (minimum),
max (maximum) et
range (étendue).

Pistes d'exploration

Analyses univariées avec gtsummary

Le package `gtsummary` permet de réaliser facilement des **tableaux univariés** grâce à la fonction `tbl_summary`, que l'on peut exporter facilement (image, pdf, html).

```
# Création d'un tableau de fréquence simple et propre avec plusieurs variables  
enquetev %>%  
tbl_summary(include = c("RS3", "Q4", "RS6"))
```

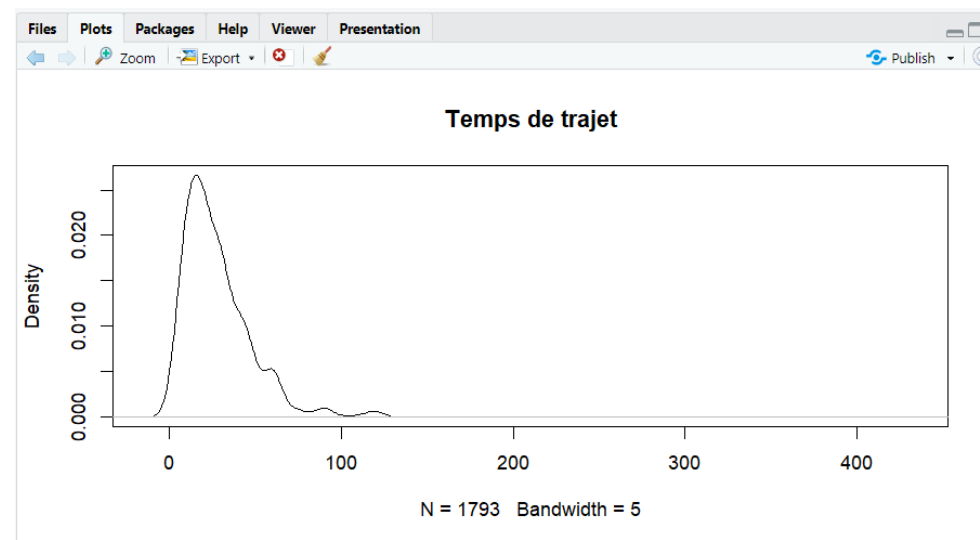
Characteristic	N = 2,072 ¹
RS3	
LillesEurope	336 (16%)
LillesFlandres	390 (19%)
ParisgaredeBercy	291 (14%)
ParisgaredeleEst	300 (14%)
ParisGaredeLyon	354 (17%)
RouenRiveDroite	401 (19%)

Pistes d'exploration

Analyses univariées avec questionr

La **fonction density** permet d'obtenir une estimation par noyau de la distribution. Le résultat de cette estimation est ensuite représenté graphiquement à l'aide de plot.

```
# Estimation par noyau de la distribution du nombre de min pour les trajets  
plot(density(enquetev$Q12, bw = 5, na.rm = TRUE), main = "Temps de trajet")
```

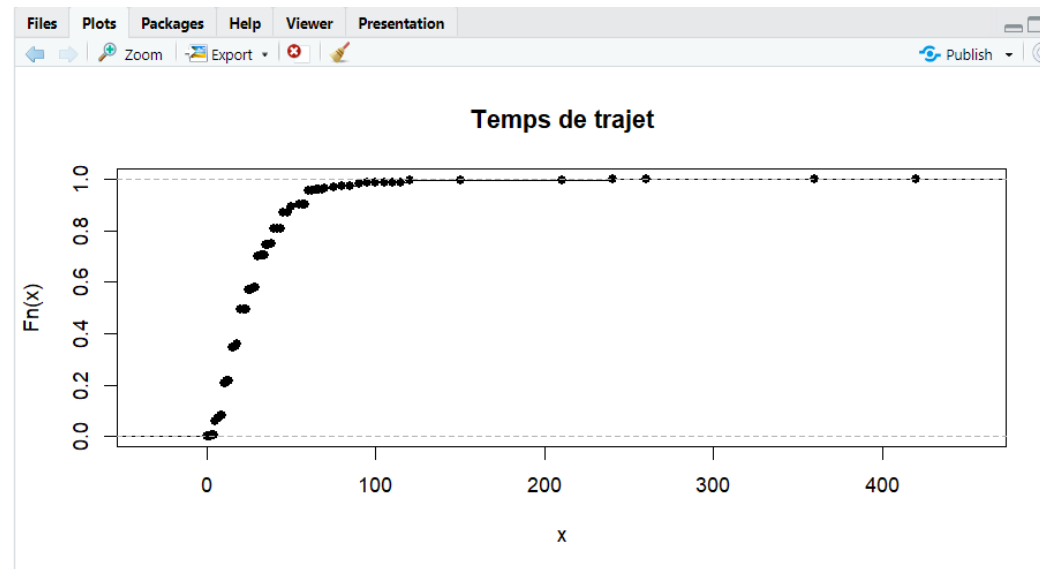


Pistes d'exploration

Analyses univariées avec questionr

La fonction de **répartition empirique** ou empirical cumulative distribution function en anglais avec la fonction `ecdf`.

```
# Répartition empirique de la durée des trajets pour se rendre en gare  
plot(ecdf(enquetev$Q12), main = "Temps de trajet")
```

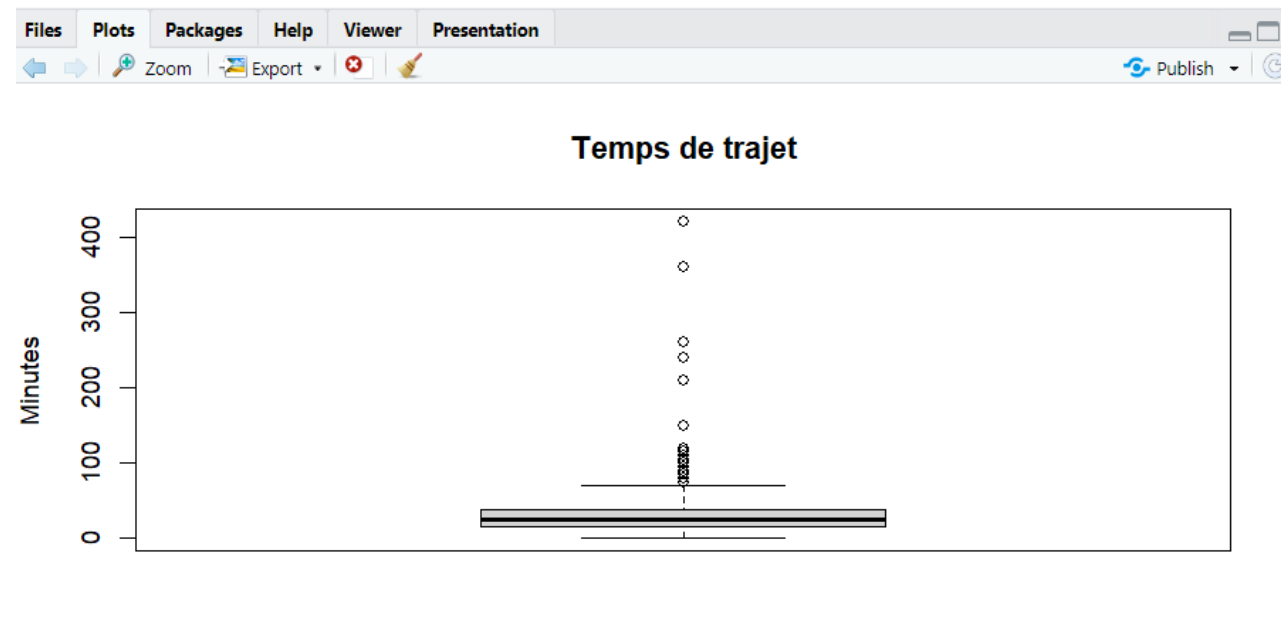


Pistes d'exploration

Analyses univariées avec questionr

Les boîtes à moustaches, ou `boxplots` en anglais avec la fonction `boxplot`.

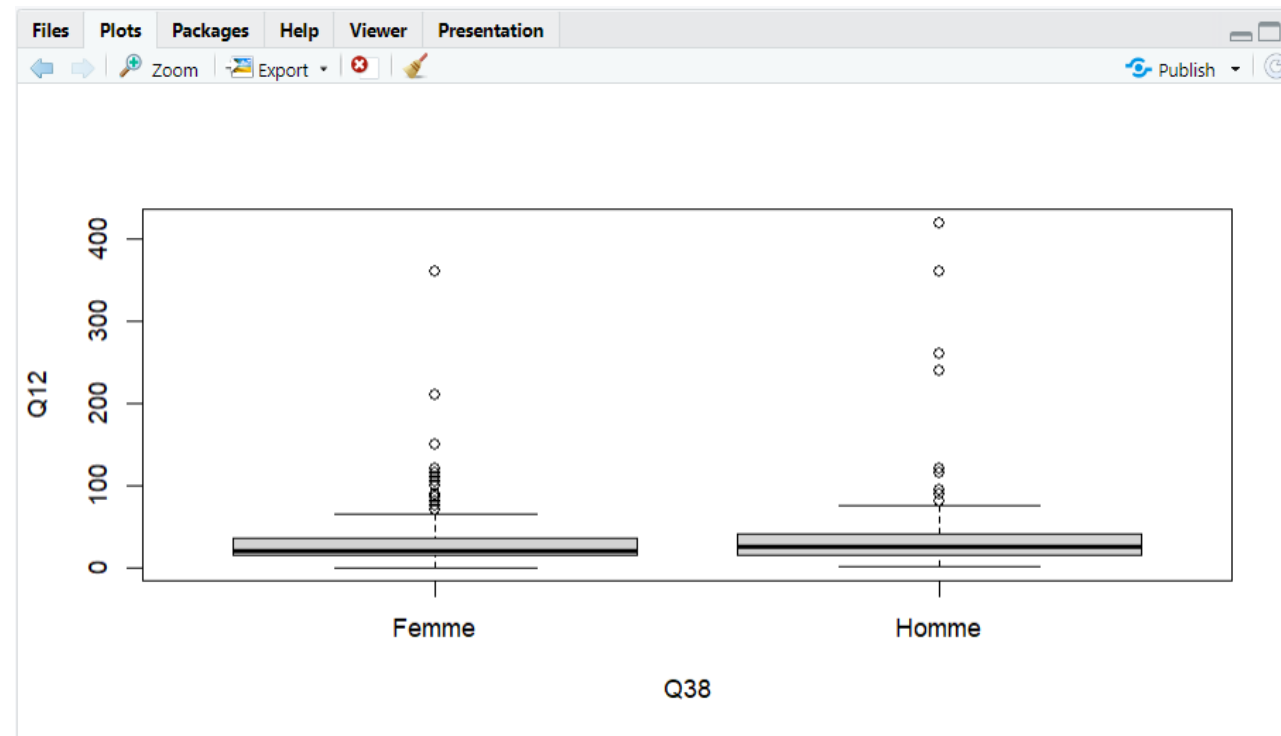
```
# Boîte à moustache de la variable temps de trajet  
boxplot(enquetev$Q12, main = "Temps de trajet", ylab = "Minutes")
```



Pistes d'exploration

Analyses bivariées (quanti/quali) avec un boxplot

```
# Croisement temps de trajet/genre  
boxplot(Q12 ~ Q38, data = enquetev)
```

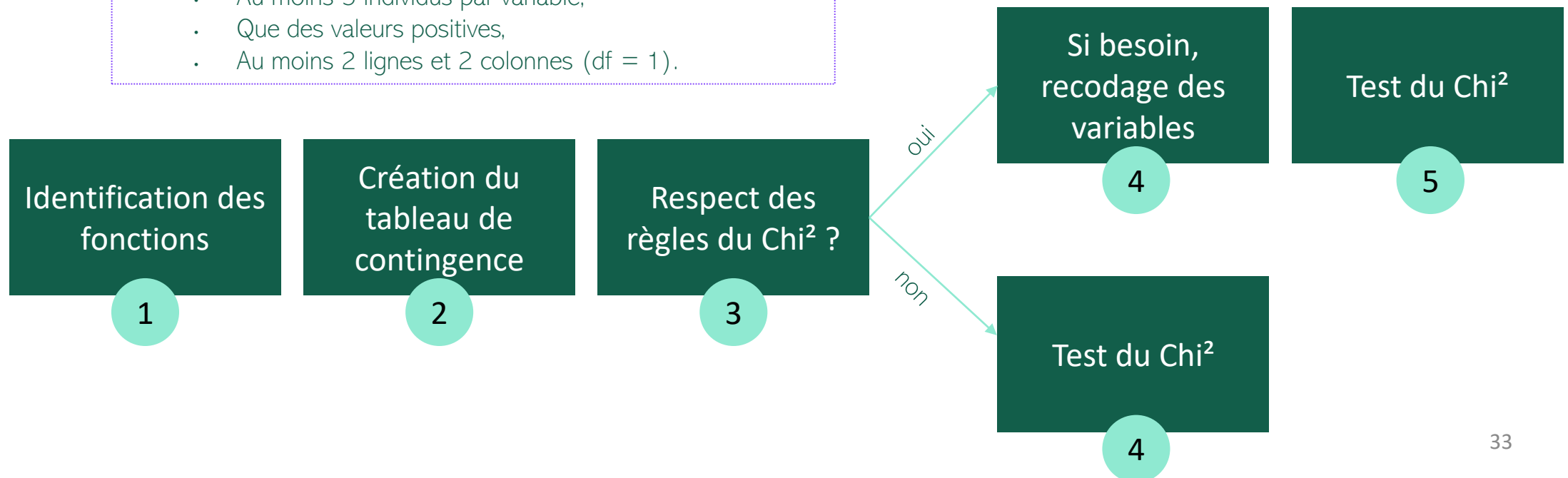


Pistes d'exploration

Calcul d'un Chi²

Réaliser un test du chi² sur les variables genre (Q38) et Nombre de modes utilisés (Q11_R1)

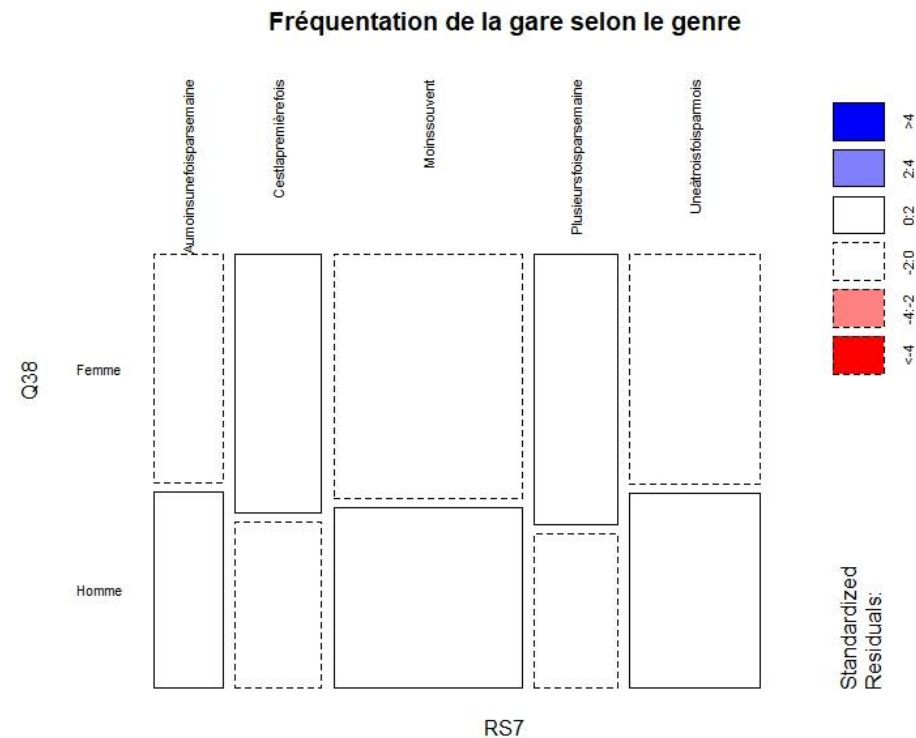
- Nb. : le test repose sur des modalités qualitatives.
Le tableau de contingence doit avoir :
- Au moins 5 individus par variable,
 - Que des valeurs positives,
 - Au moins 2 lignes et 2 colonnes (df = 1).



Pistes d'exploration

Analyses bivariées avec un graphe en mosaïque

```
# Graphe en mosaïque  
mosaicplot(RS7 ~ Q38,  
            data = enqueteV, shade = TRUE,  
            las=1.5, main = "Fréquentation de la gare selon le genre")
```



Pistes d'exploration

Vecteur et facteur dans les graphiques (1 variable)

Type de graphique	Exemple d'appel de fonction	Type de x
diagramme en secteurs (<i>pie chart</i>)	<code>pie(table(x), ...)</code>	facteur
diagramme à barres (<i>bar plot</i>)	<code>barplot(table(x), ...)</code>	facteur
diagramme en points de Cleveland	<code>dotchart(table(x), ...)</code>	facteur
histogramme	<code>hist(x, ...)</code>	vecteur numérique
courbe de densité à noyau (<i>kernel density plot</i>)	<code>plot(density(x), ...)</code> → méthode <code>plot.density</code>	vecteur numérique
diagramme en boîte (<i>boxplot</i>)	<code>boxplot(x, ...)</code>	vecteur numérique
diagramme quantile-quantile théorique normal	<code>qqnorm(x, ...)</code>	vecteur numérique

Pistes d'exploration

Vecteur et facteur dans les graphiques (2 variables)

Type de graphique	Exemple d'appel de fonction	Type de x	Type de y
diagramme à barres empilées ou groupées	<code>barplot(table(x, y), ...)</code> avec <code>beside = TRUE</code> pour barres groupées	facteur	facteur
diagramme en points de Cleveland	<code>dotchart(table(x, y), ...)</code>	facteur	facteur
diagramme en mosaïque	<code>mosaicplot(table(x, y), ...)</code>	facteur	facteur
diagrammes en boîte juxtaposés	<code>boxplot(y ~ x, ...)</code>	facteur	vecteur numérique
diagramme de dispersion (<i>scatterplot</i>) ou en lignes (<i>line chart</i>)	<code>plot(x, y, ...)</code> → méthode <code>plot.default</code>	vecteur numérique	vecteur numérique
diagramme quantile-quantile empirique	<code>qqplot(x, y, ...)</code>	vecteur numérique	vecteur numérique

Pistes d'exploration

Des **tests du Chi²** pour tester les indépendances entre

- pratiques et gares,
- entre genre et pratiques,
- entre la demande de services et le nombre de services proposés (plus difficile, car cela demande de croiser les 2 tableaux)

Réaliser des **boxplots** pour comparer la répartition des caractéristiques quantitatives des individus par gare.

Réaliser des **graphiques en facette** pour avoir les profils intermodaux des trajets par gare

Pistes d'exploration

Jointure

Jointure via l'identifiant de la gare

```
# création d'un ID commun à avec la typologie
enquetev$id_gare <- case_when(enquetev$RS3 == "LillesFlandres"~8,
                             enquetev$RS3 == "ParisgaredeBercy"~21,
                             enquetev$RS3 == "RouenRiveDroite"~27)

sort(unique(enquetev$id_gare))

# nombre d'enquêtés par gare
enquetev_g <- enquetev %>%
  group_by(id_gare, RS3) %>% count()

# sélection des colonnes pour la jointure
nb_v <- typologie %>%
  select(id_gare, Nb_v_2019)

#jointure
jointure <- left_join(enquetev_g, nb_v, by = "id_gare")

# nbre d'enquêtés % nbre voyageurs
jointure$pourc <- ((jointure$n *100)/jointure$Nb_v_2019)
```

Objectifs de l'atelier

Présentation de l'enquête SNCF et des objectifs de recherche

Traitements de base d'une enquête et packages associés

Prise en main des données

Pistes d'exploration

Des ressources utiles

Et si on faisait une petite carte ?

Des ressources utiles

Le [site](#) de J. Lamarange

Tout savoir sur le Chi^2 tout en buvant des jus de carotte et de tomate :
le [guide](#) de Julien Barnier

Le [site](#) de Questionr

Aller plus loin dans [gtsummary](#)

Objectifs de l'atelier

Présentation de l'enquête SNCF et des objectifs de recherche

Traitements de base d'une enquête et packages associés

Prise en main des données

Pistes d'exploration

Des ressources utiles

Et si on faisait une petite carte ?

Et si on faisait une petite carte ?

[Script](#)

