

Analyse exploratoire

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

Analyse de données

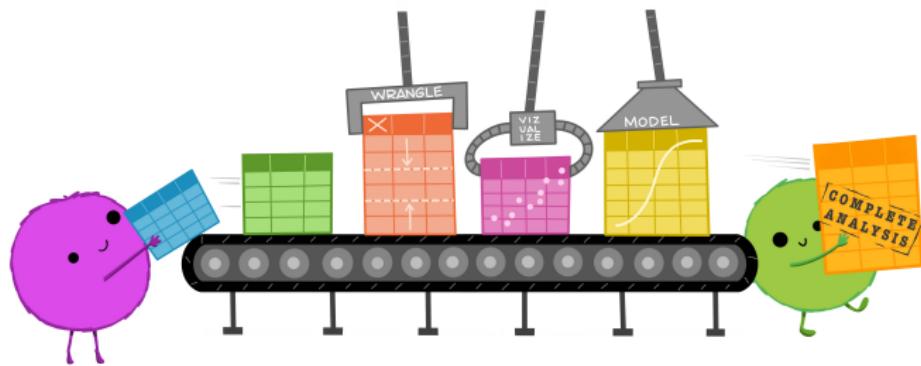


Figure 1: Allison Horst (CC BY 4.0)

Organisation du travail

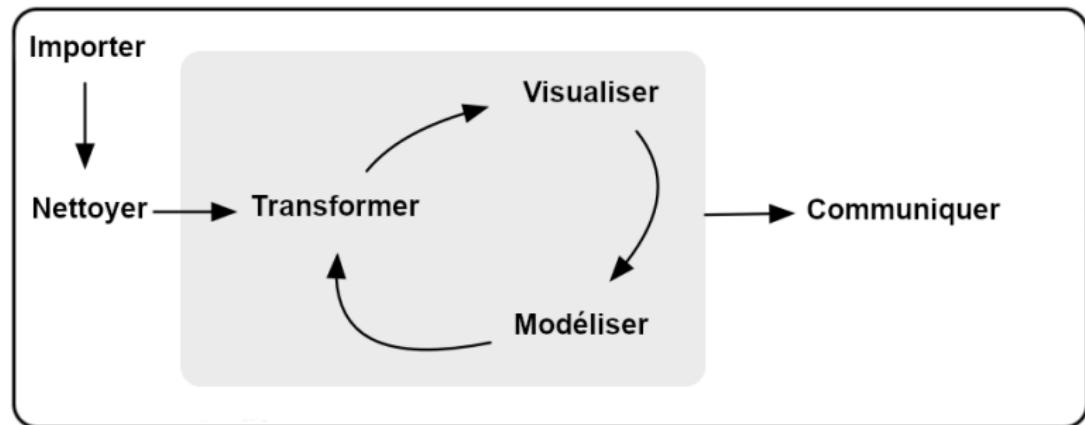


Figure 2: Adapté de *R for Data Science*, H. Wickham et G. Grolemund

Quelques bonnes pratiques

Data Organization in Spreadsheets

Karl W. Broman^a and Kara H. Woo^b

^aDepartment of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; ^bInformation School, University of Washington, Seattle, WA

ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY

Received June 2017
Revised August 2017

KEYWORDS

Data management; Data organization; Microsoft Excel; Spreadsheets

Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989

Données en format "tidy"

- variables en colonnes
- observations en lignes
- une seule mesure par cellule

“**TIDY DATA** is a standard way of mapping the meaning of a dataset to its structure.”
—HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

Figure 3: Allison Horst (CC BY 4.0)

Types de variables numériques

CONTINUOUS

measured data, can have ∞ values within possible range.



I AM 3.1" TALL
I WEIGH 34.16 grams

DISCRETE

OBSERVATIONS CAN ONLY EXIST AT LIMITED VALUES, OFTEN COUNTS.



I HAVE 8 LEGS
and
4 SPOTS!

@allison_horst

Figure 4: Allison Horst (CC BY 4.0)

Types de variables catégorielles

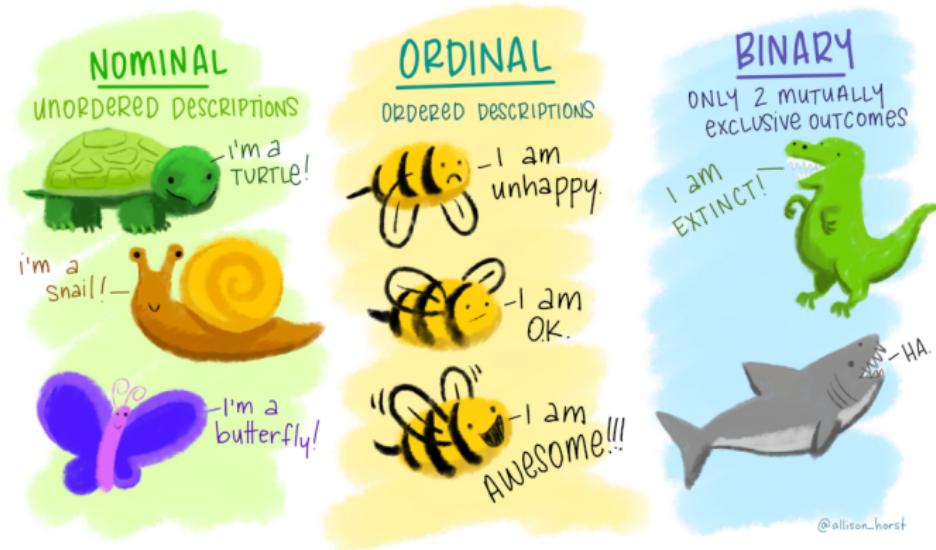


Figure 5: Allison Horst (CC BY 4.0)

Vérifier la présence de

- valeurs manquantes (NA), -999, etc.
- relations logiques (total, moyenne, etc.) entre variables
- variables catégorielles non déclarées
 - valeur entière
 - chaînes de caractère

Un simple graphique transmet plus d'information à l'analyste que n'importe quel autre option

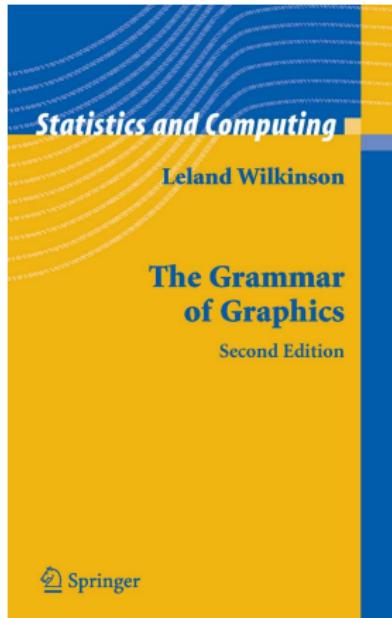
John Tukey

Qu'est ce qu'un bon graphique?

communique des idées complexes avec clarté, précision et efficacité ... le graphique qui offre au lecteur le plus grand nombre d'idées le plus rapidement possible avec le moins d'encre et le plus petit espace possible

Edward Tufte, 1983

Grammaire des graphiques



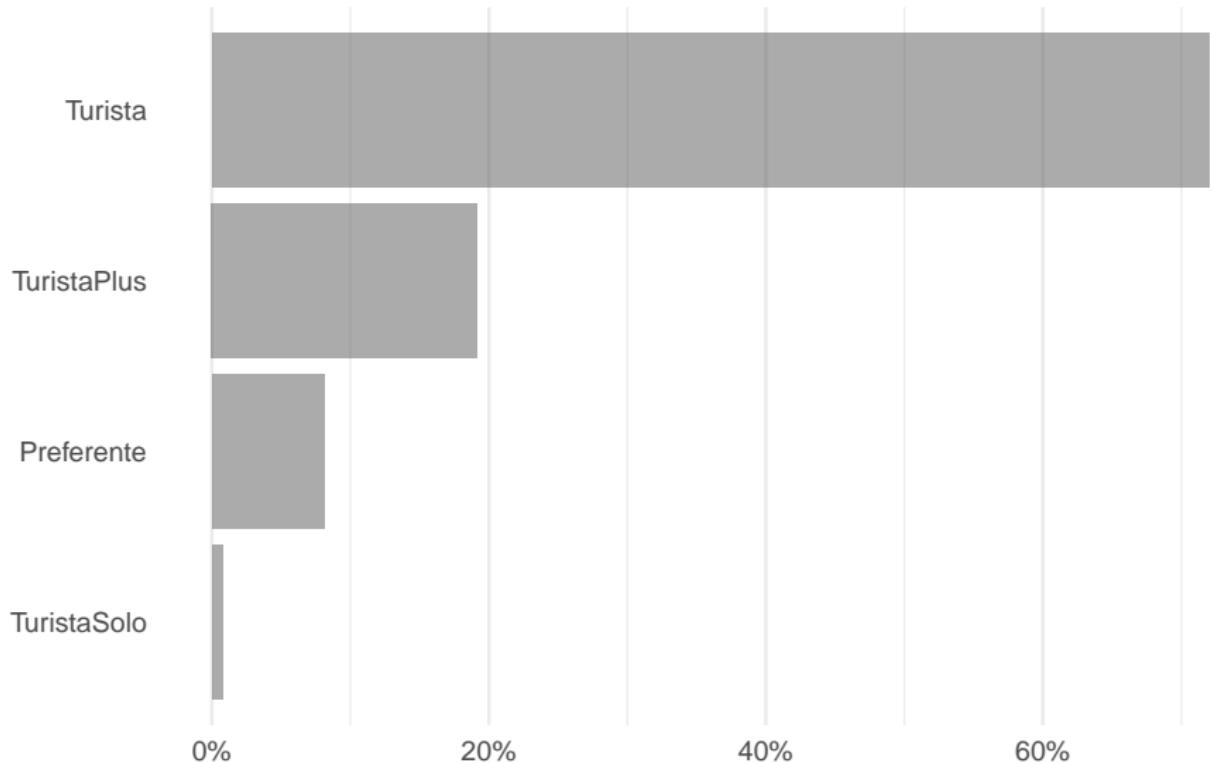
- Éléments (couches):
 - données
 - application (variable → esthétique)
 - objets géométriques
 - transformations
 - positionnement
- Échelle / guide
- Coordonnées (facettes, système de coordonnés)

Pour une visualisation effective:

1. le choix du graphique dépend du type de variable
2. soignez les apparences
3. portez une attention particulière à la perception visuelle humaine

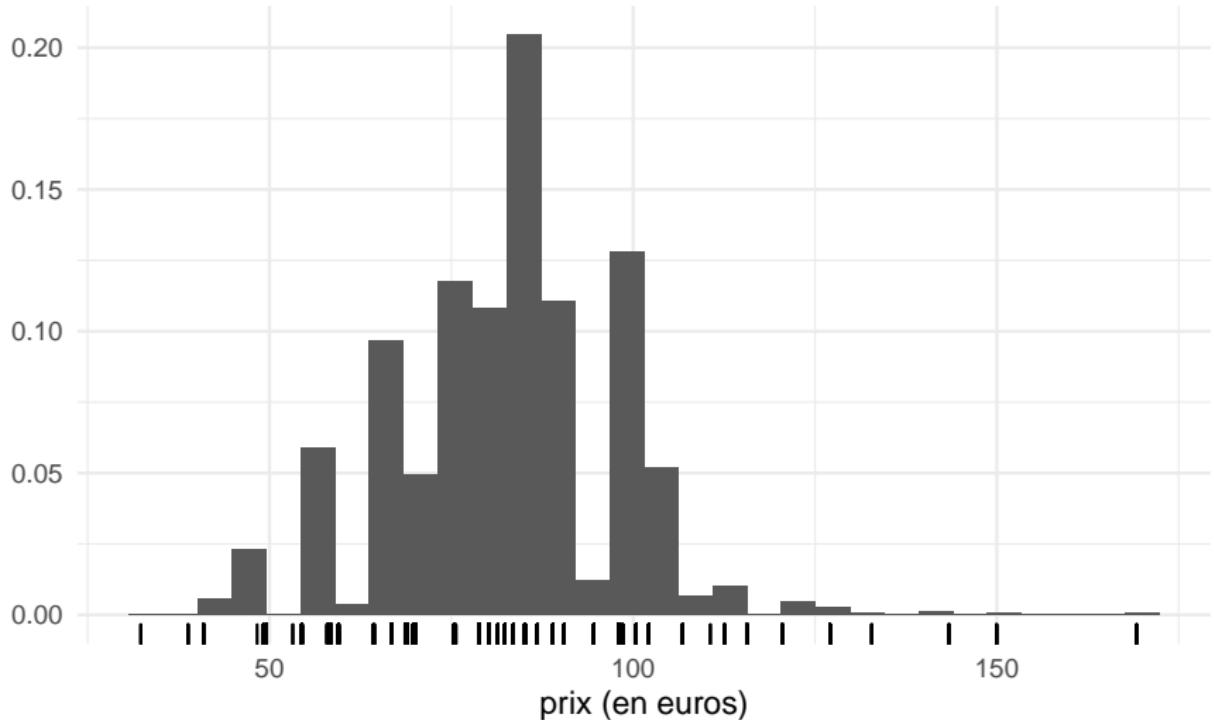
- continue: histogramme, densité
- discrète: diagramme en bâton
- catégorielle: diagramme en bâton (fréquence ou pourcentage)

Répartitions des billets selon la classe



Répartition du prix des billets de train (%)

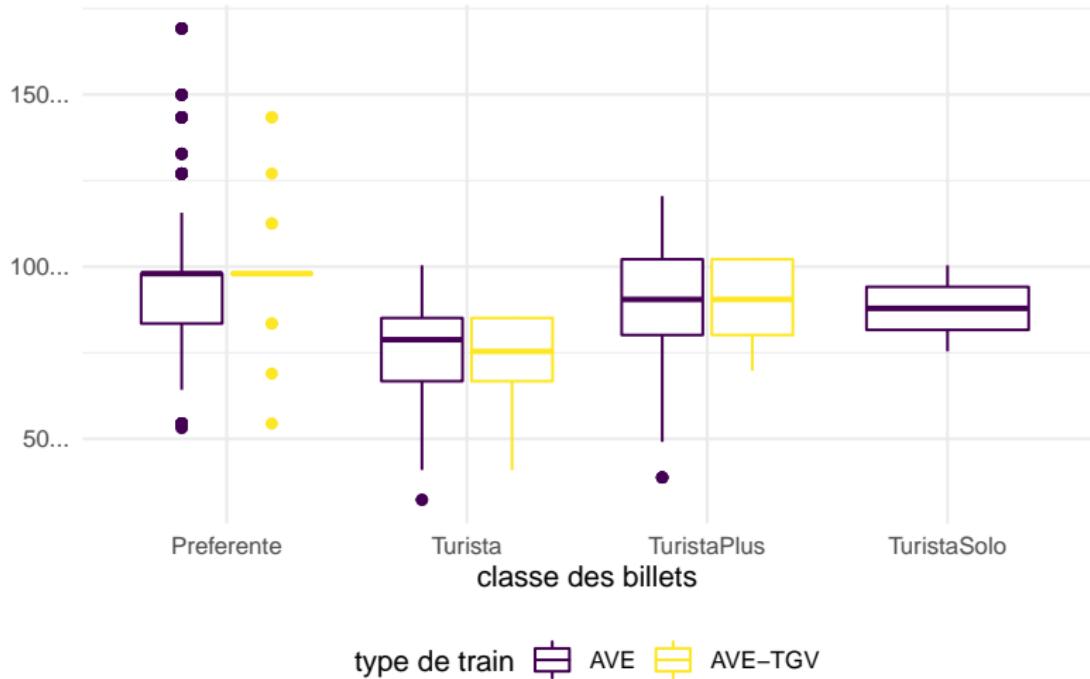
Billets au tarif Promo entre Barcelone et Madrid.



Règle 1: choix de graphiques avec deux variables

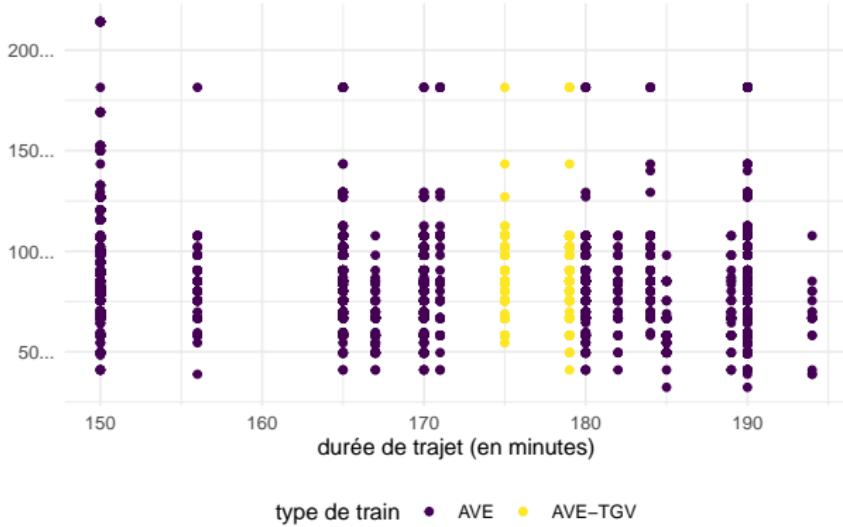
- continues: nuage de points
- catégorielles: diagramme à bande (avec couleurs), carte thermique
- continue \times catégorielle: boîte à moustache, graphique violon

Prix (en euros) de billets au tarif Promo



type de train AVE AVE-TGV

Prix de billets de train en fonction de la durée de trajet



- Qu'est-ce qui cloche dans la représentation graphique précédente?
- Comment pourrait-on remédier aux problèmes soulevés?

Règle 2: soignez les apparences

Certaines visualisations sont plus effectives/adéquates que d'autres

- votre graphique doit être interprétable uniquement avec la légende.
- inclure les noms de variables **et** les unités
- ajouter une description dans le texte et faire une référence croisée

- Titre et annotation
- Libellés et unités sur les axes
- Libellé de l'axe des y en sous-titre
- Inverser les axes si les étiquettes trop longue (variable catégorielles)

Règle 3: perception visuelle humaine

- ratio longueur/largeur
- taille de police suffisante pour lisibilité
- espace entre bandes
- étendu des axes (incluant ou pas zéro)
- choix de couleurs
 - noir/blanc avec contraste
 - palettes pour daltoniens
- comparaison d'aires/superficies (difficile)
- graphiques 3D / avec rotation superflue à éviter

Problèmes de perceptions



Donald J. Trump
@realDonaldTrump

Follow

In addition to Florida - South Carolina, North Carolina, Georgia, and Alabama, will most likely be hit (much) harder than anticipated. Looking like one of the largest hurricanes ever. Already category 5. BE CAREFUL! GOD BLESS EVERYONE!

7:51 AM - 1 Sep 2019

18,075 Retweets 73,811 Likes



14K



18K

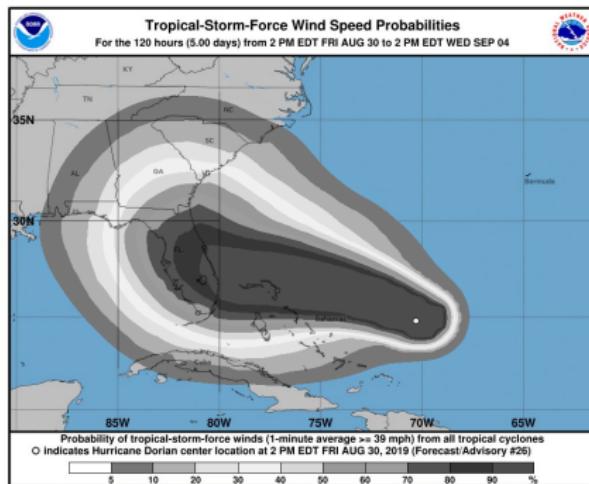


74K



Mauvaise palette de couleur

- Gauche: Carte originale de la NOAA en niveaux de gris: on voit clairement le problème de saturation
- Droite: solution potentielle avec palette de couleurs différente.



(Source Achim Zeileis)

Pourquoi créer des graphiques?

Les résumés numériques focalisent l'attention sur les valeurs attendues, les résumés graphiques sur les valeurs inattendues.

John Tukey

Étapes de l'analyse exploratoire

1. Formuler des questions
2. Chercher des réponses à ces questions à l'aide de
 - statistiques descriptives
 - tableaux de contingence
 - graphiques
3. Infirmer ou confirmer nos intuitions
4. Raffiner les questions suite aux observations
5. Répéter le processus

Écrire un résumé des trouvailles et des aspects **importants** uniquement.

Pour aller plus loin:

- Chapitre 1 de *Data Visualization: A practical introduction* par Kieran Healy
- *Fundamentals of Data Visualization* par Claus O. Wilke
- Chapitre 3 de *R for Data Science* par Garrett Grolemund et Hadley Wickham