Analyse exploratoire

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

Analyse de données

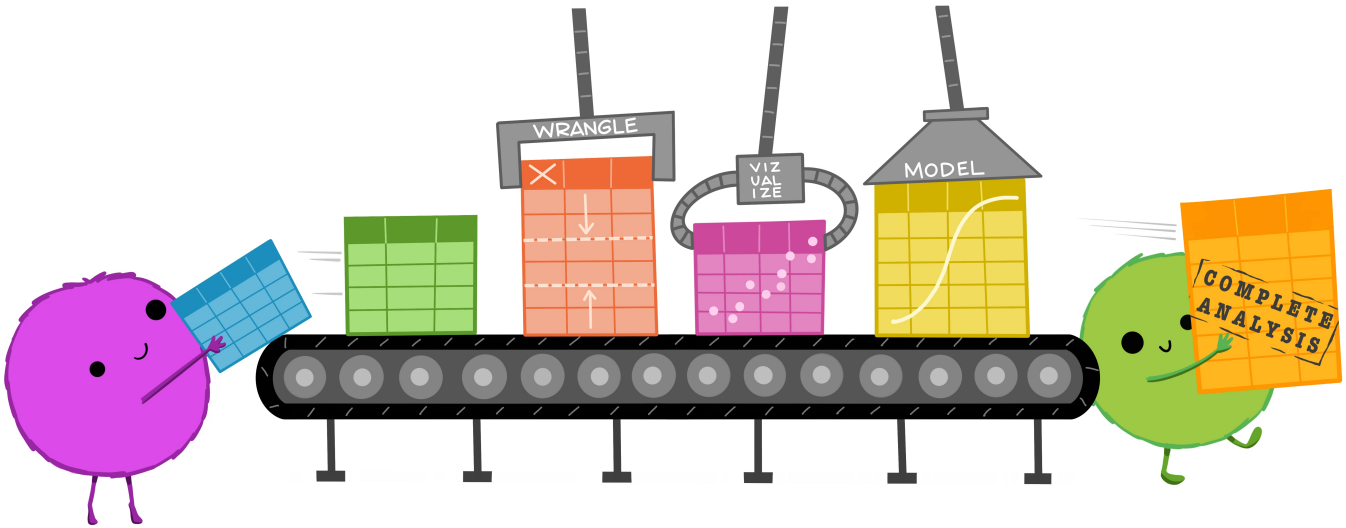


Figure 1: Allison Horst (CC BY 4.0)

Organisation du travail

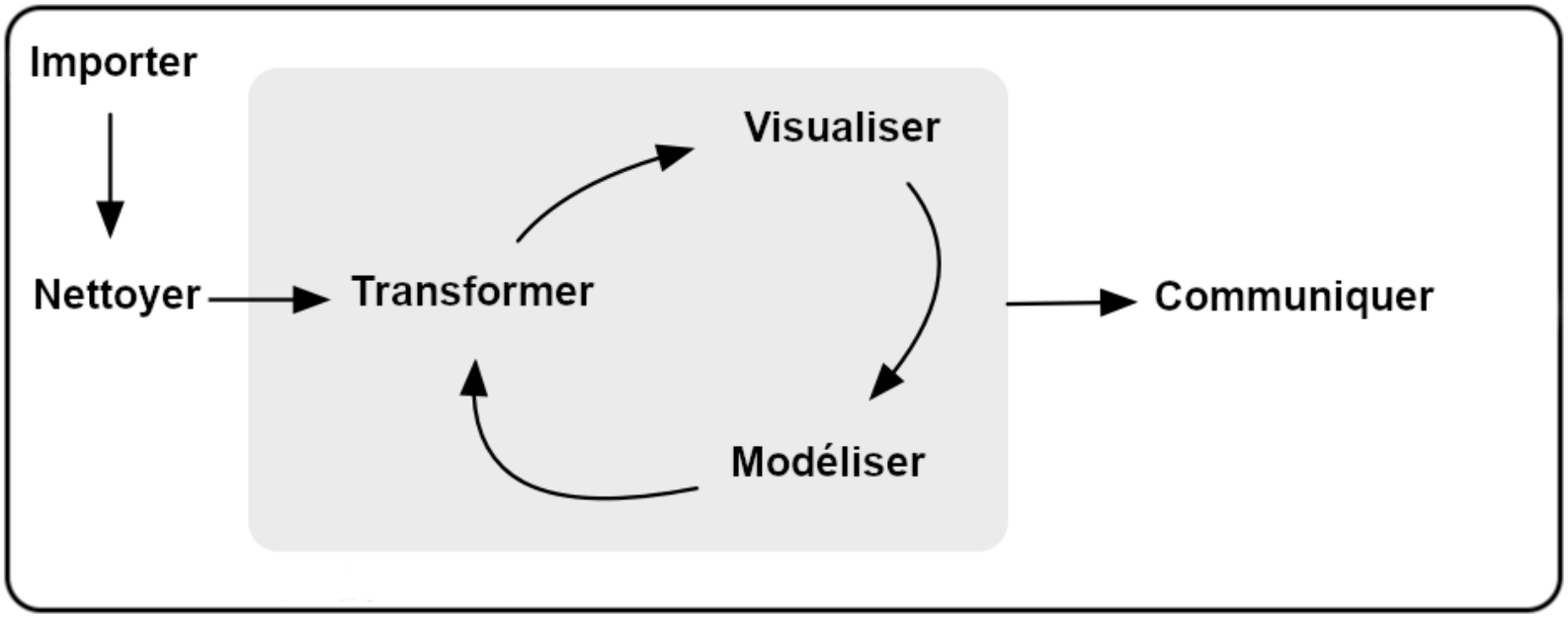


Figure 2: Adapté de *R for Data Science*, H. Wickham et G. Grolemund

Quelques bonnes pratiques

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information 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.

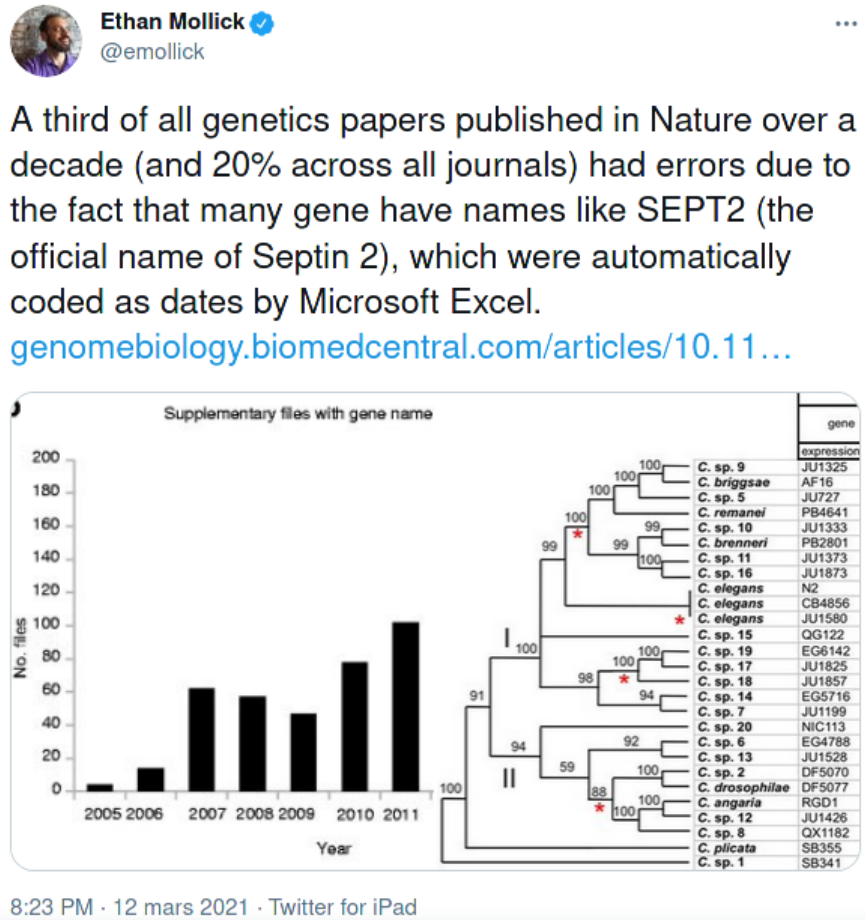
ARTICLE HISTORY
Received June 2017
Revised August 2017

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

Nettoyage des données

- Toujours garder une copie des données brutes
- Automatiser le nettoyage



Histoires d'horreur

Health policy

Analysis

Covid: how Excel may have caused loss of 16,000 test results in England

Alex Hern

UK technology editor

Public Health England data error blamed on limitations of Microsoft spreadsheet

Coronavirus - latest updates

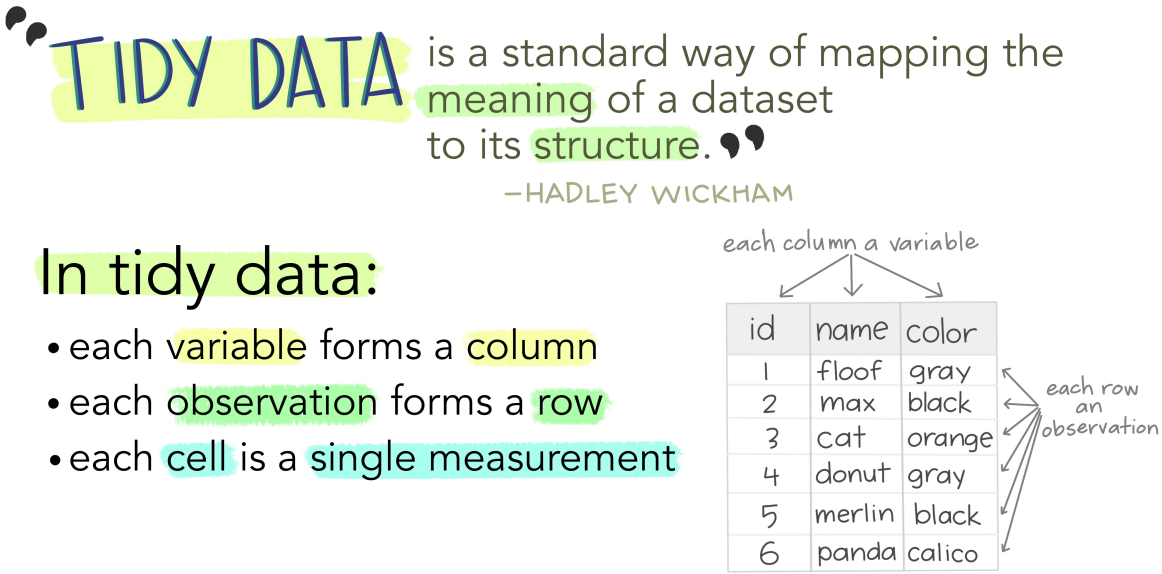
See all our coronavirus coverage

Tue 6 Oct 2020 08.21 BST

Figure 3: Capture d'écran d'un article du quotidien *The Guardian*

Données en format “tidy”

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



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

Figure 4: Allison Horst (CC BY 4.0)

Exemple

Est-ce que ces données de la Régie de l'Énergie sont en format 'tidy'?

(¢/litre)	HEBDOMADAIRE												
Régions	janvier				février					mars			
	2022-01-03	2022-01-10	2022-01-17	2022-01-24	2022-01-31	2022-02-07	2022-02-14	2022-02-21	2022-02-28	2022-03-07	2022-03-14	2022-03-21	2022-03-28
1. Bas-Saint-Laurent	148,2	150,6	151,1	151,4	160,8	164,7	165,6	166,0	166,1	196,7	185,5	182,0	185,3
2. Saguenay-Lac-Saint-Jean	138,6	141,3	146,6	146,4	151,6	153,8	153,7	155,7	156,5	190,1	175,8	170,1	173,0
3. Capitale-Nationale	149,8	149,8	154,6	154,7	161,4	161,8	166,5	166,7	168,2	195,0	181,5	178,8	180,9
4. Mauricie	148,3	148,3	152,8	153,2	155,2	159,7	162,2	164,3	164,7	193,2	182,8	178,4	183,3
5. Estrie	146,9	147,6	148,6	149,4	152,8	156,7	159,0	159,8	161,9	188,2	184,6	178,5	183,7
6. Montréal	151,8	152,5	156,1	156,5	161,9	167,3	170,3	166,5	170,6	196,1	182,1	184,6	181,5
7. Outaouais	140,6	143,2	145,0	147,7	152,0	155,4	157,3	157,0	158,7	184,1	182,3	174,4	180,9
8. Abitibi-Témiscamingue	148,9	148,9	149,6	149,9	153,4	159,3	160,8	164,4	164,4	193,8	183,9	179,2	179,4
9. Côte-Nord	146,2	147,9	149,5	153,6	159,2	160,7	163,5	163,3	164,4	193,8	185,4	176,8	181,7
10. Nord-du-Québec (excl. Nunavik)	158,6	160,1	162,5	163,6	166,8	169,2	173,9	174,0	174,9	204,1	192,4	189,2	194,2
11. Gaspésie-Îles-de-la-Madeleine	151,7	153,2	155,7	156,6	161,7	167,7	168,7	167,8	168,1	195,2	186,6	184,1	181,3
12. Chaudière-Appalaches	148,8	148,8	153,4	153,0	161,2	161,1	165,4	165,5	165,9	194,8	180,9	178,7	179,2
13. Laval	151,7	152,0	156,7	157,1	162,5	167,8	170,8	166,7	170,8	196,3	182,3	185,3	182,1
14. Lanaudière	146,8	146,9	150,2	152,0	155,7	159,7	161,3	160,5	165,3	190,4	181,5	180,4	179,6
15. Laurentides	148,0	148,7	151,2	152,9	156,3	161,0	163,5	162,0	164,6	189,8	180,6	181,3	180,6
16. Montérégie	148,4	149,2	151,8	153,1	157,4	161,7	164,9	162,6	165,8	192,6	181,7	181,6	180,1
17. Centre-du-Québec	148,7	148,8	150,2	151,3	154,1	158,4	159,2	159,7	164,7	190,1	182,5	178,5	182,6
MOYENNE PONDÉRÉE	148,0	148,8	151,8	152,7	157,6	161,2	163,7	163,1	165,4	192,7	182,0	179,9	180,7

Types de variables numériques

Figure 5: Allison Horst (CC BY 4.0)

Types de variables catégorielles

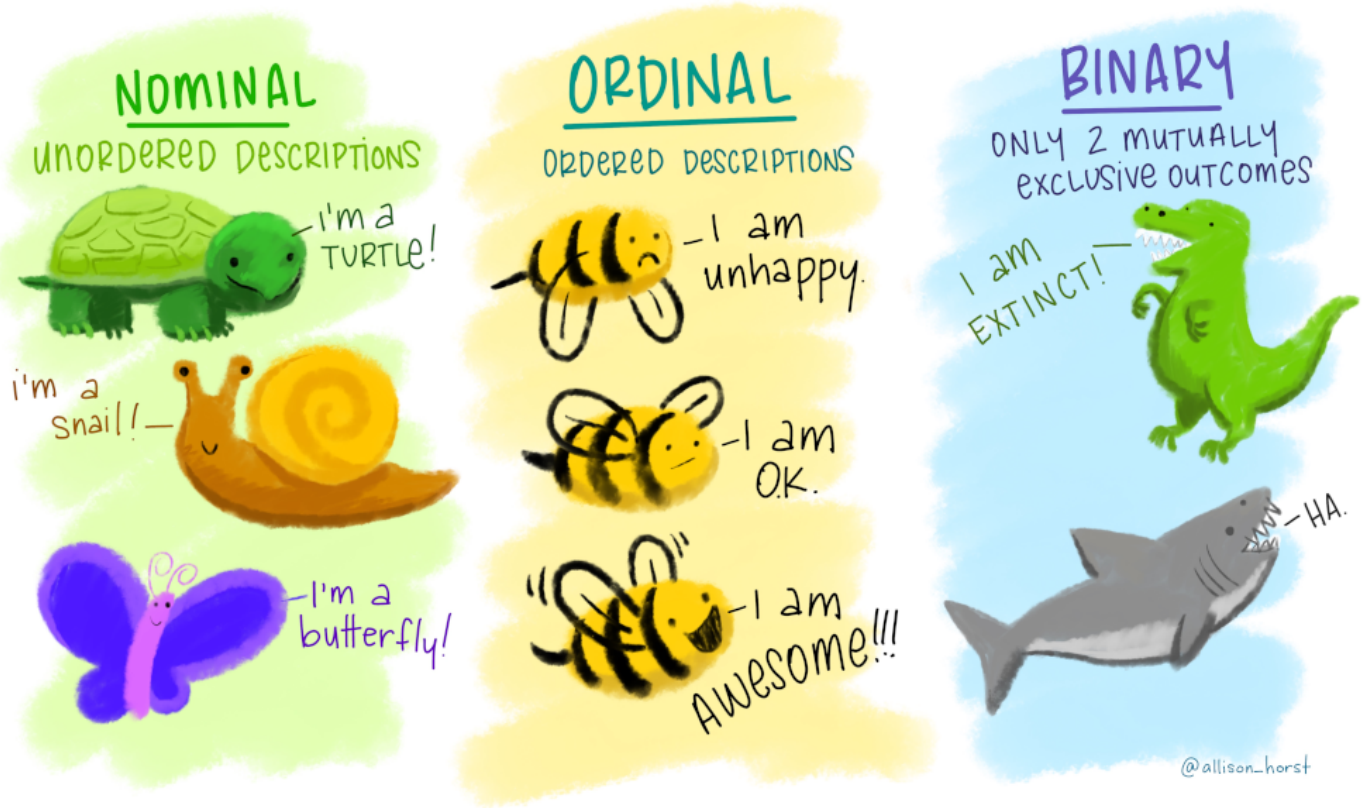


Figure 6: Allison Horst (CC BY 4.0)

Validation des données

Vérifier la présence de

- valeurs manquantes (`NA`, points, cellules vides, 999, -1, etc.)
- relations logiques (total, moyenne, etc.) entre variables
- variables catégorielles non déclarées
  - valeur entière (par ex., jours de la semaine)
  - chaînes de caractère

Visualisation

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

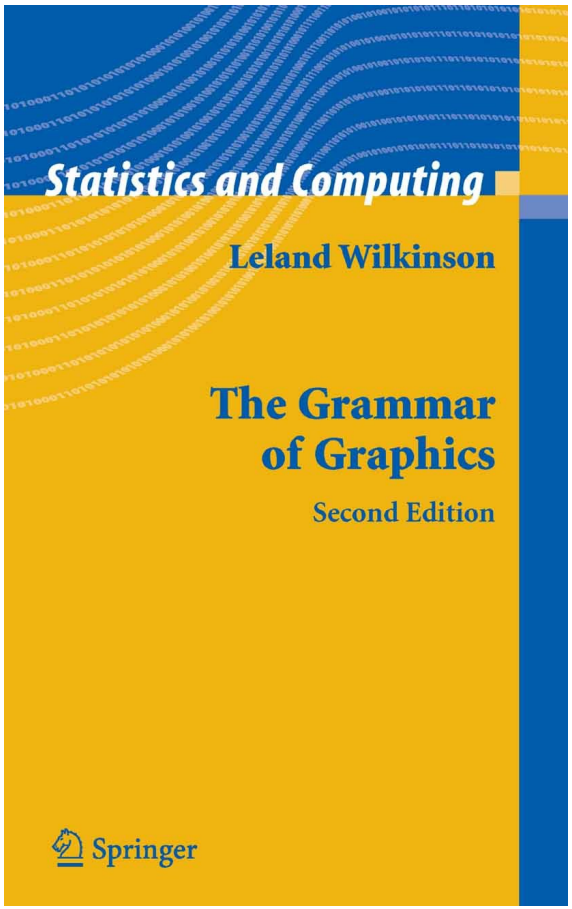
John Tukey

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 $\longrightarrow$ esthétique)
  - objets géométriques
  - transformations
  - positionnement
- Échelle / guide
- Coordonnées (facettes, système de coordonnés)

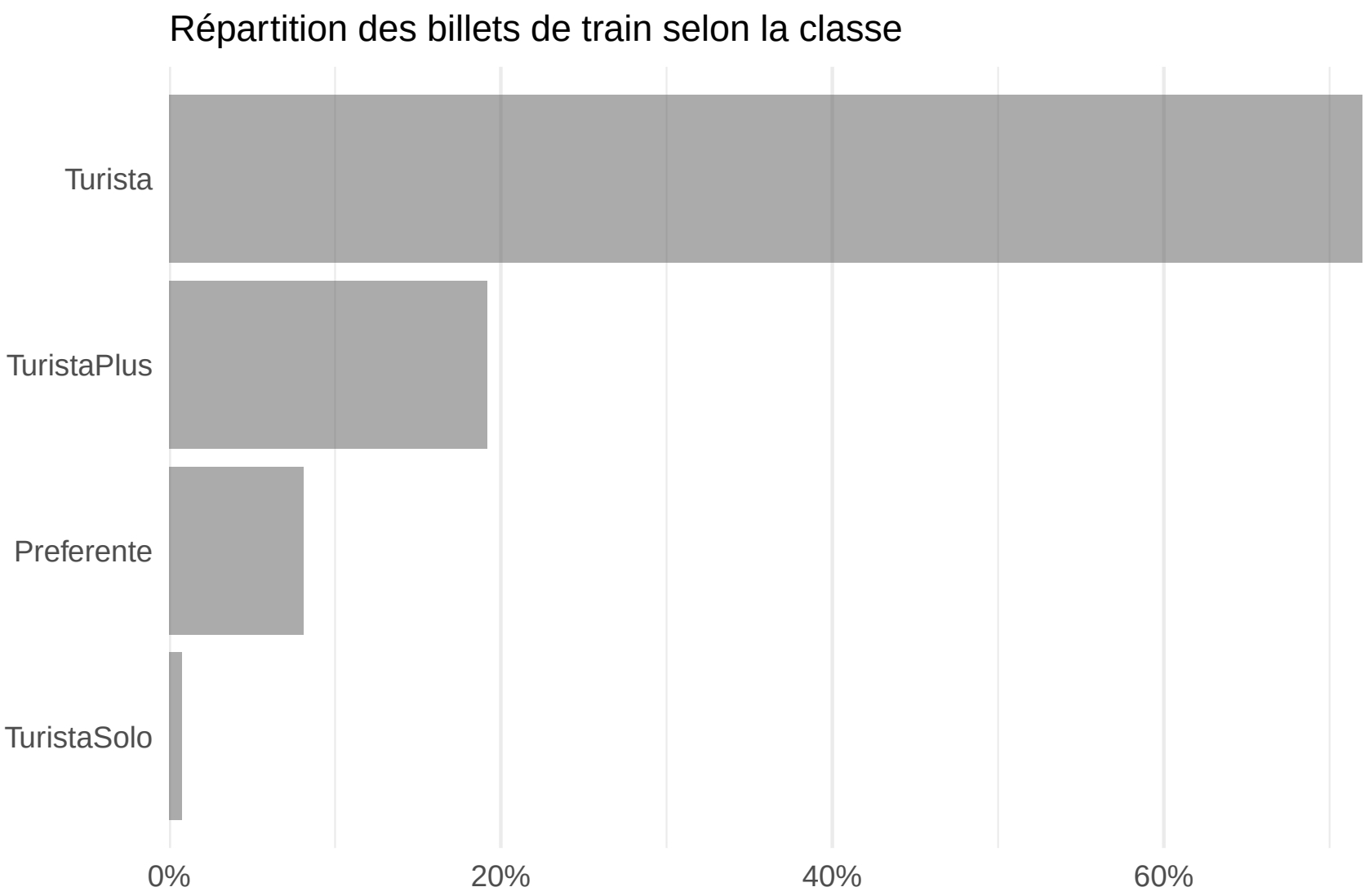
Règles d'or

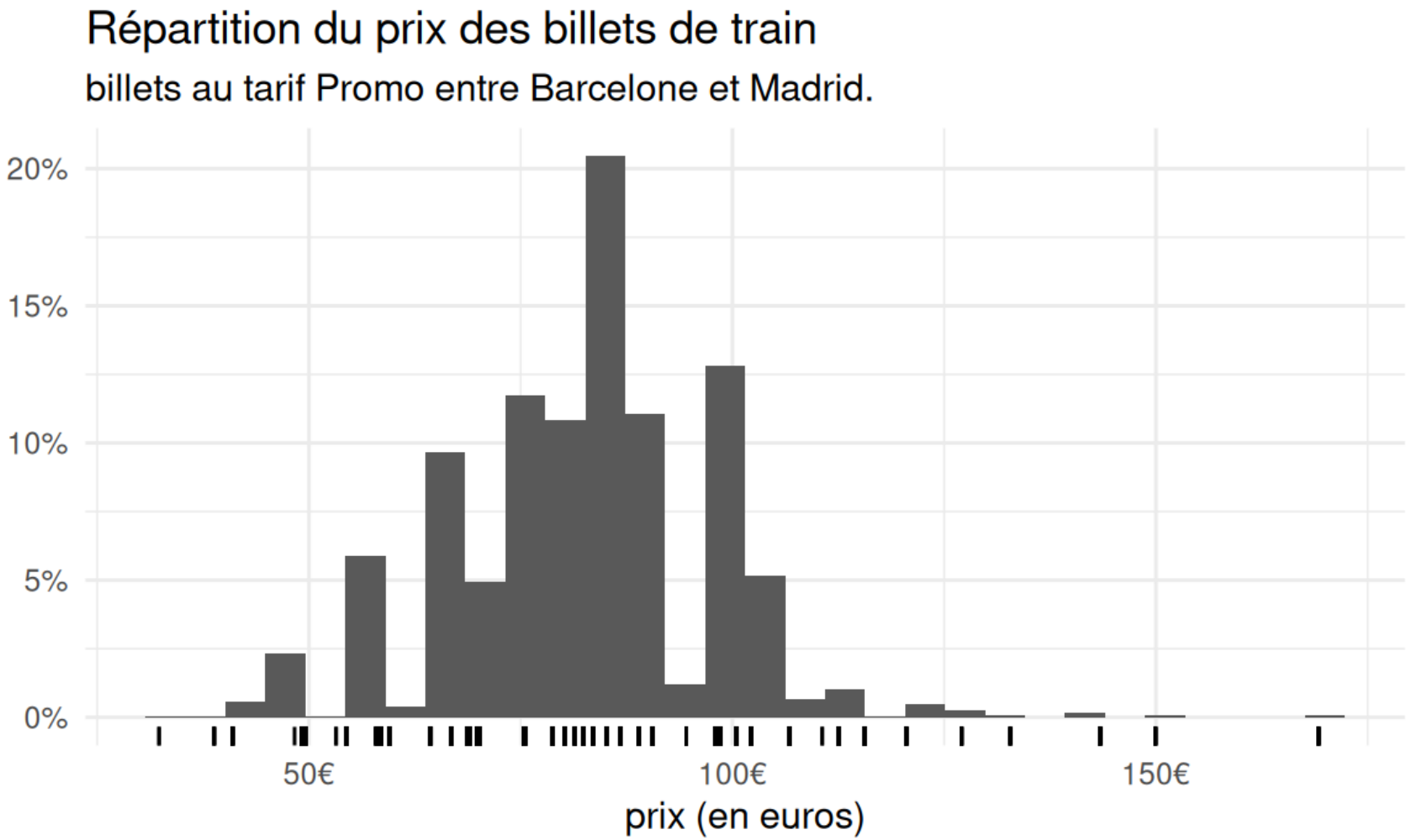
Pour une visualisation effective:

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

Règle 1: choix de graphiques avec une seule variable

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

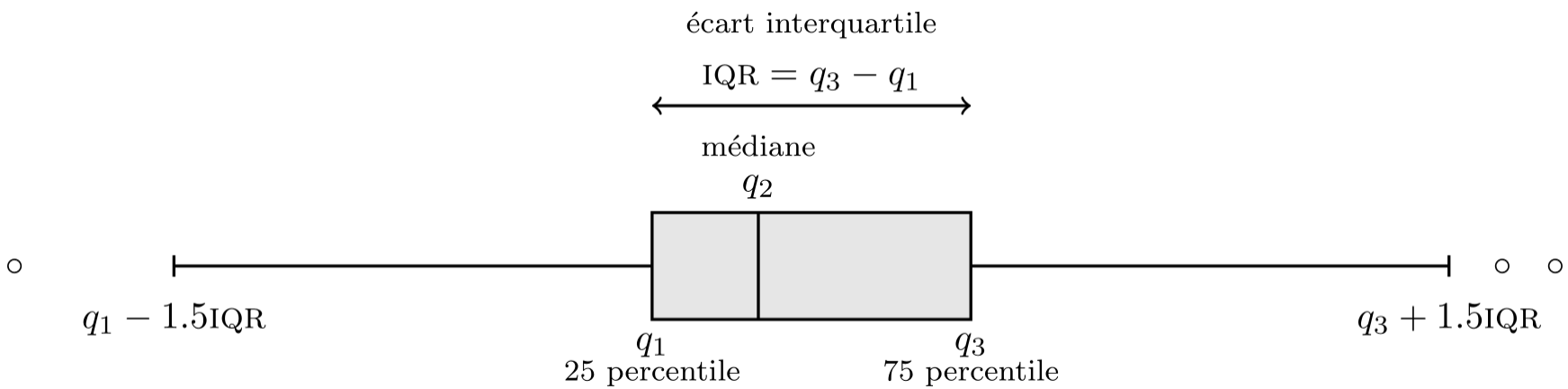

Répartition des billets de train selon la classe
Turista
TuristaPlus
Preferente
TuristaSolo
0%
20%
40%
60%

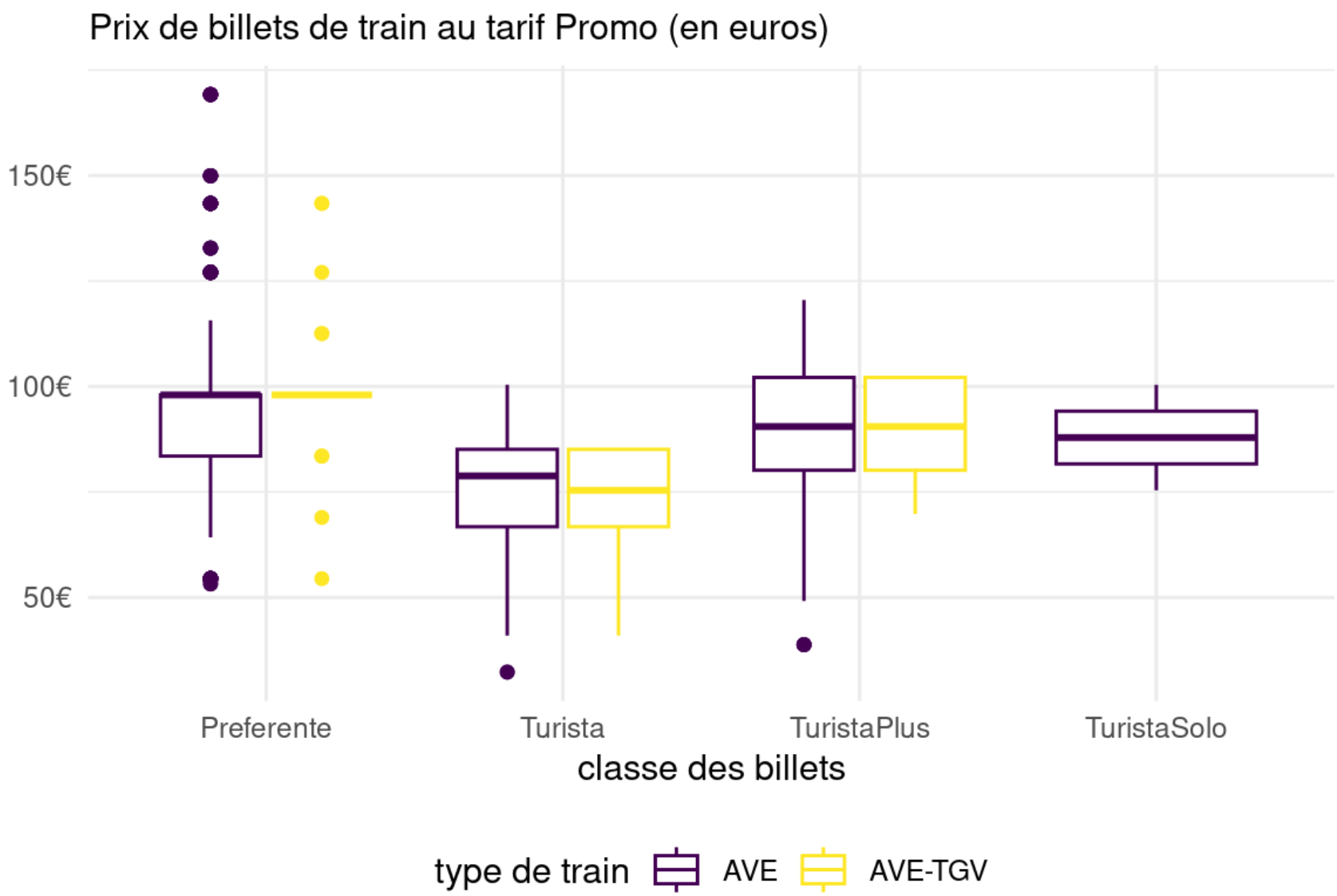

Répartition du prix des billets de train
billets au tarif Promo entre Barcelone et Madrid.
20%
15%
10%
5%
0%
50€
100€
150€
prix (en euros)

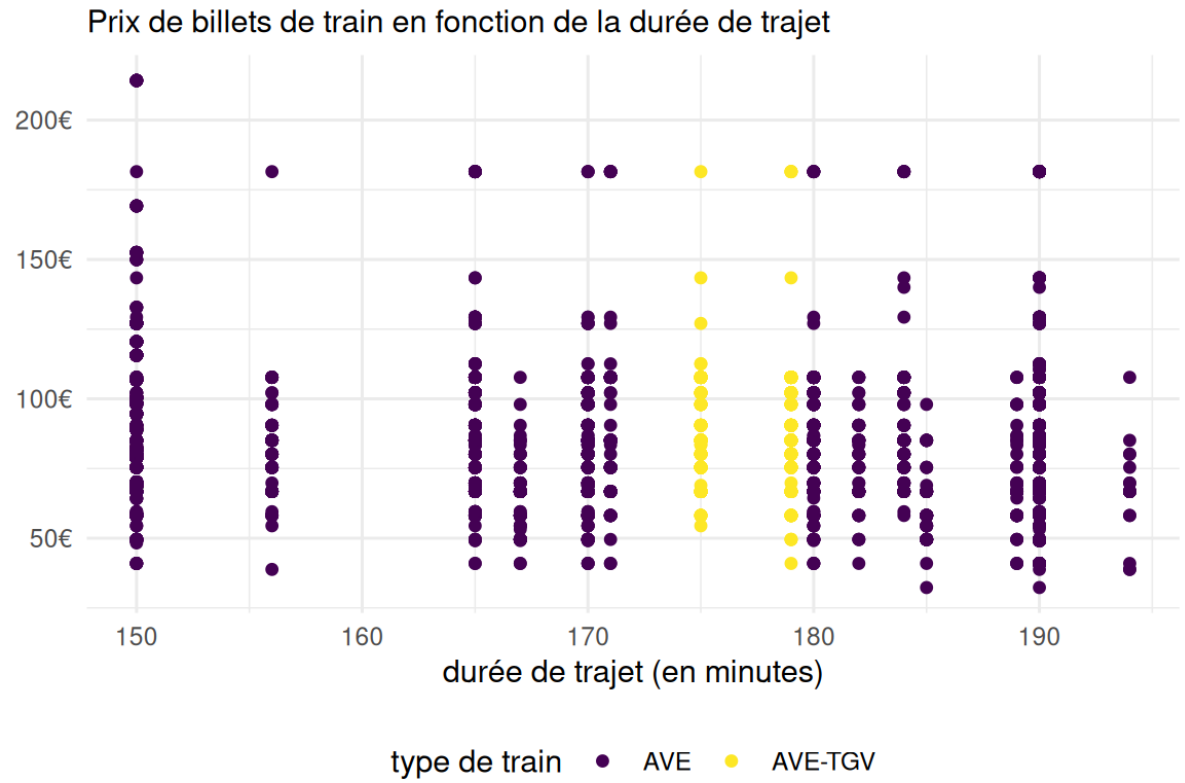
Règle 1: choix de graphiques avec deux variables

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

Boîte à moustaches




Prix de billets de train au tarif Promo (en euros)
150€
100€
50€
Preferente
Turista
TuristaPlus
TuristaSolo
classe des billets
type de train
AVE
AVE-TGV



- 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

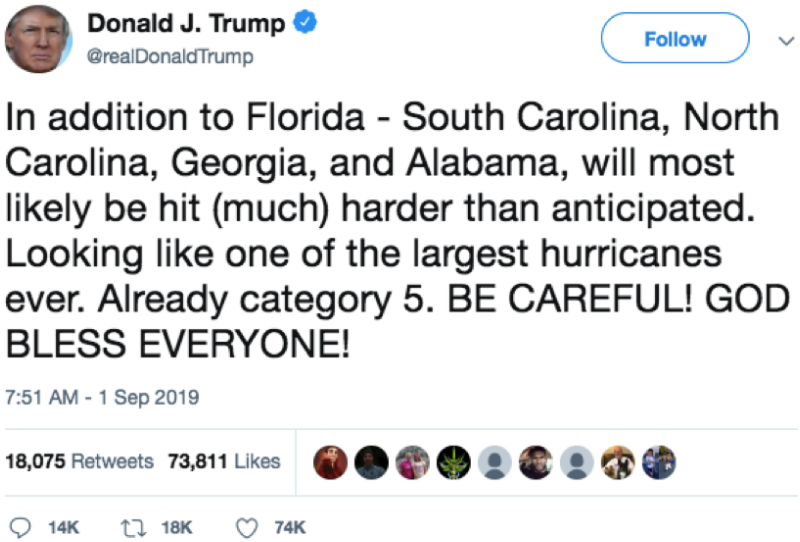
Éléments graphiques clés

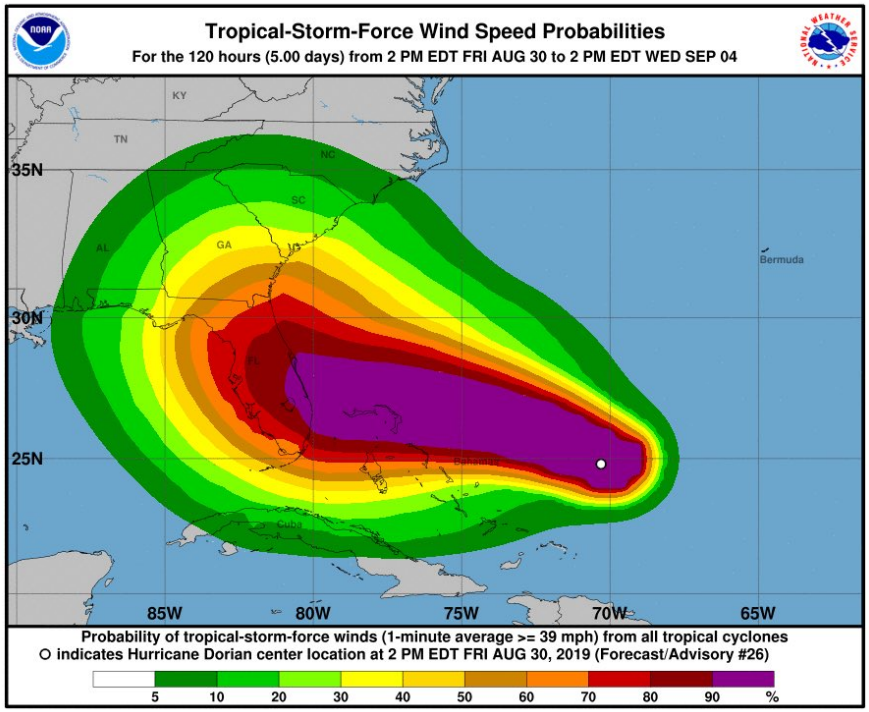
- Titre et annotation
- Libellés et unités sur les axes
- Libellé de l'axe des y en sous-titre
- Inverser les axes si les étiquettes trop longue (variable catégorielles)

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

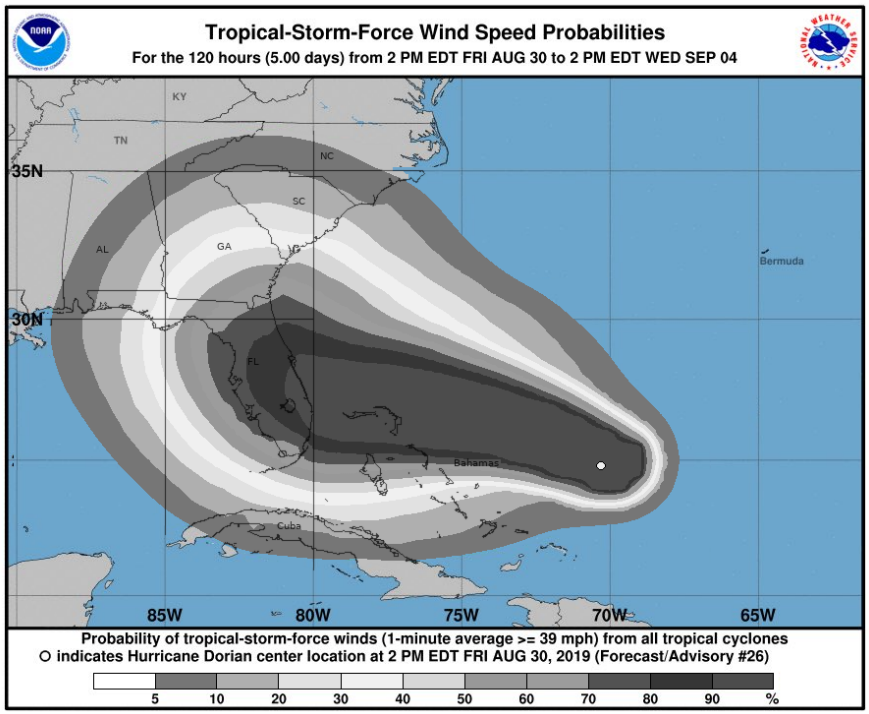
Problèmes de perceptions

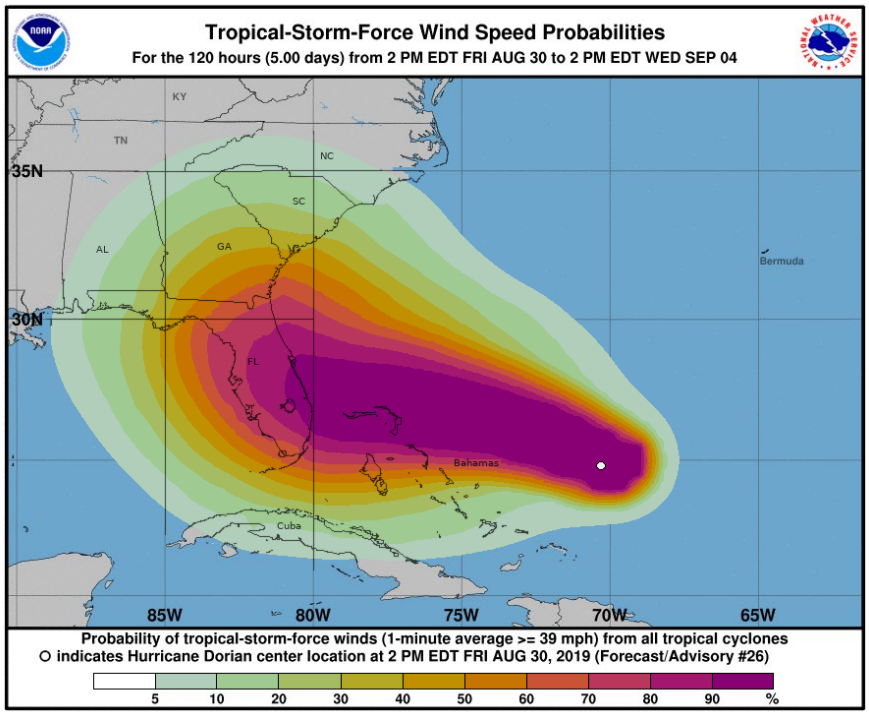




Mauvaise palette de couleur

- 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.





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.

Références complémentaires

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