

Analyse exploratoire

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

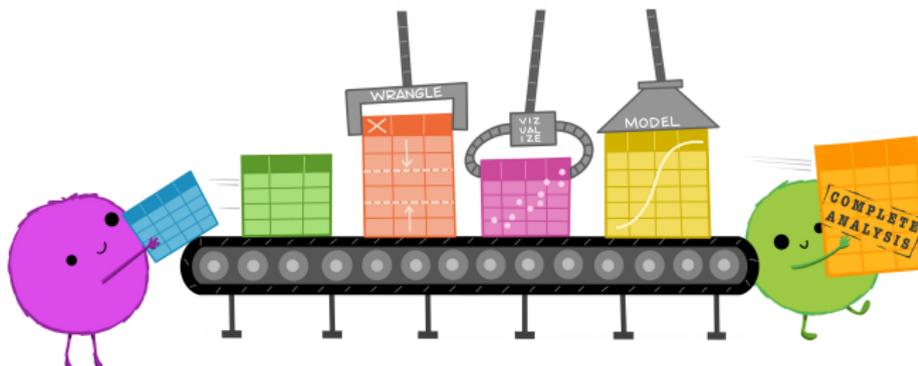


Figure 1: Allison Horst (CC BY 4.0)

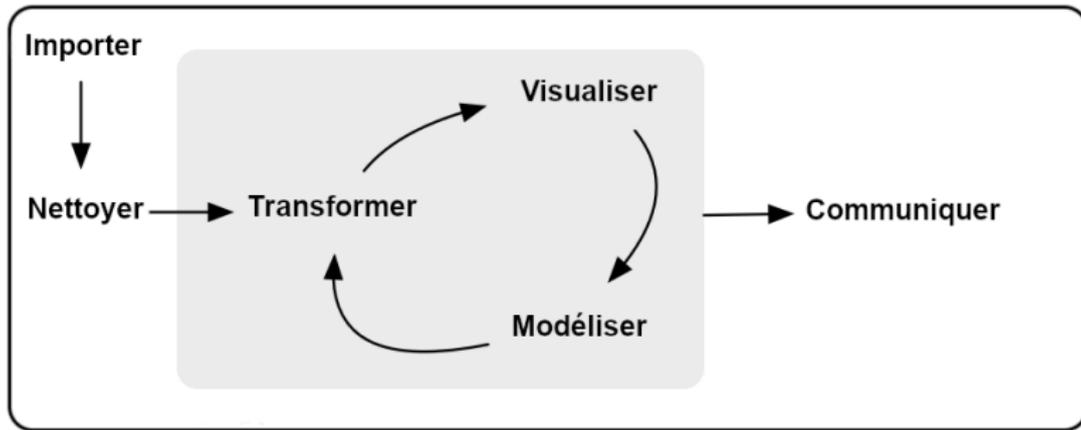


Figure 2: Adapté de *R for Data Science*, H. Wickham et G. Grolemund

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

Nettoyage des données

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



Rohan Alexander
@RohanAlexander

Friends don't let friends use Excel.
Official data release of the 2019 Kenyan census: knbs.or.ke/?wpmpro=2019-...
BTW clean and tidy version coming soon :)

Traduire le Tweet

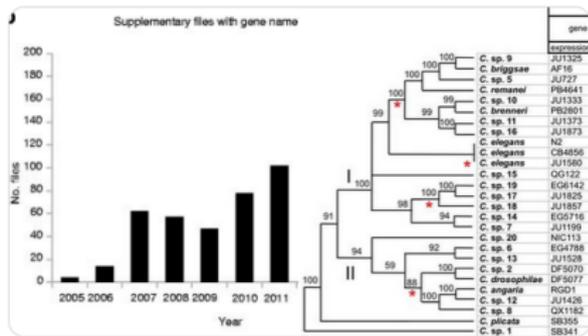
Table 2.3: Distribution of Population by Age, Sex*, County and Sub-County

MURURU							
Age	Male	Female	Total	Age	Male	Female	Total
Total	43975	45156	89131	01	452	450	902
0	790	765	1,555	52	501	537	1038
1	734	780	1,514	53	373	447	817
2	827	841	1,668	54	473	463	936
3	868	781	1,649	50-54	2,327	2,496	4823
4	819	827	1,646	55	449	524	973
5	4028	3,994	8,022	56	487	502	989
6	843	839	1,682	57	380	463	843
7	879	884	1,763	58	327	376	703
8	872	889	1,761	59	404	423	827
9	882	945	1,827	50-59	2,057	2,288	4,345
10	1030	910	1,940	60	434	501	935
11	926	848	1,774	61	328	373	701
12	996	896	1,892	62	289	299	588
				63	250	247	497
				64	181	186	367



Ethan Mollick
@emollick

A third of all genetics papers published in Nature over a decade (and 20% across all journals) had errors due to the fact that many gene have names like SEPT2 (the official name of Septin 2), which were automatically coded as dates by Microsoft Excel. genomebiology.biomedcentral.com/articles/10.11...



8:23 PM · 12 mars 2021 · Twitter for iPad



The screenshot shows a news article from The Guardian. On the left, there is a red header 'Health policy'. The main article title is 'Covid: how Excel may have caused loss of 16,000 test results in England' in a large, bold, black font. Below the title is the author's name 'Alex Hern' in a red, italicized font, followed by their title 'UK technology editor' in a grey, italicized font. A sub-headline reads 'Public Health England data error blamed on limitations of Microsoft spreadsheet'. Below this are two links: 'Coronavirus - latest updates' and 'See all our coronavirus coverage', each preceded by a small grey circle. At the bottom left, the date and time 'Tue 6 Oct 2020 08.21 BST' are displayed. At the bottom right, there is a small image showing a close-up of a computer keyboard with a 'CTRL' key and a mouse.

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*

- 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

each column a variable

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

each row an observation

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)

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

(e 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	157,8	161,2	163,7	163,1	165,4	192,7	182,0	179,9	180,7		

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 5: Allison Horst (CC BY 4.0)

Types de variables catégorielles



Figure 6: Allison Horst (CC BY 4.0)

Vérifier la présence de

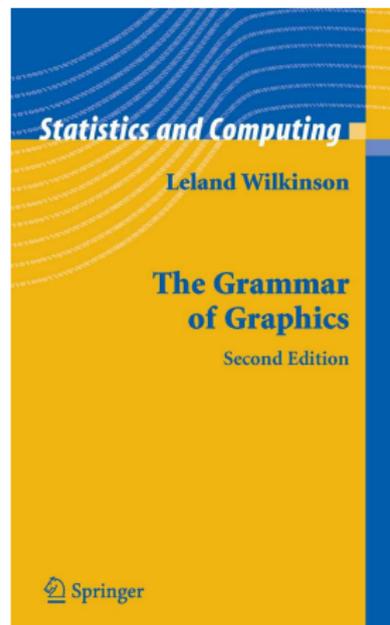
- valeurs manquantes (NA, points, cellules vides, 999, -1, etc.)
- relations logiques (total, moyenne, etc.) entre variables
- variables catégorielles non déclarées
 - valeur entière (par ex., jours de la semaine)
 - chaînes de caractère

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

John Tukey

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



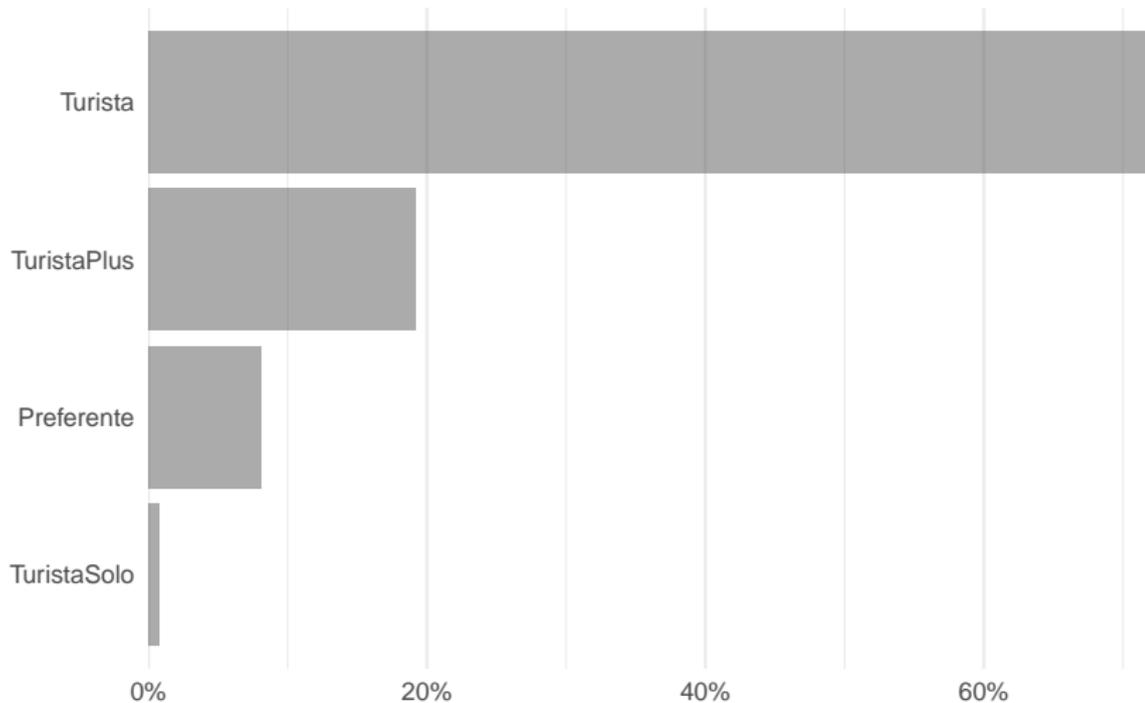
- Éléments (couches):
 - données
 - application (variable → esthétique)
 - objets géométriques
 - transformations
 - positionnement
- Échelle / guide
- Coordonnées (facettes, système de coordonnées)

Pour une visualisation effective:

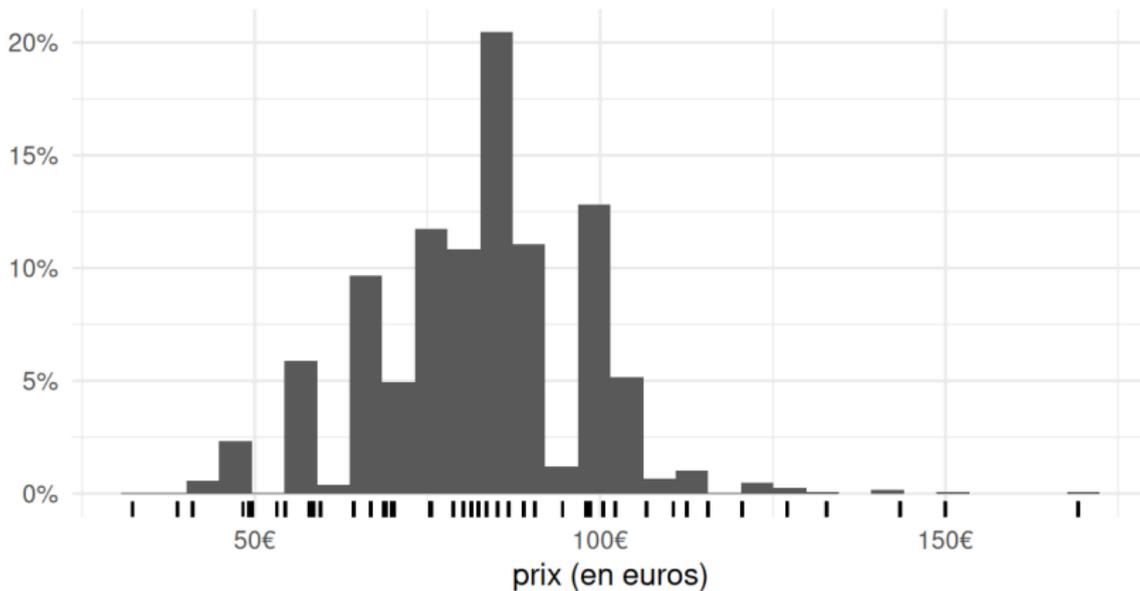
1. le choix du graphique dépend du type de variable
2. soignez les apparences
3. portez une attention particulière à la perception visuelle humaine

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

Répartition des billets de train selon la classe



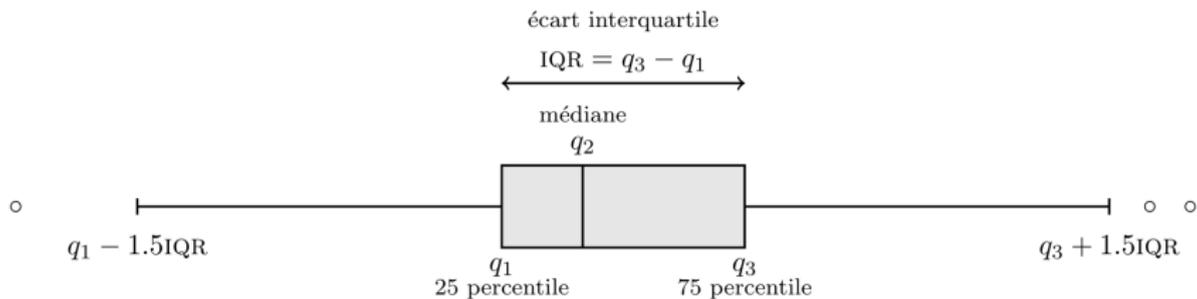
Répartition du prix des billets de train billets au tarif Promo entre Barcelone et Madrid.



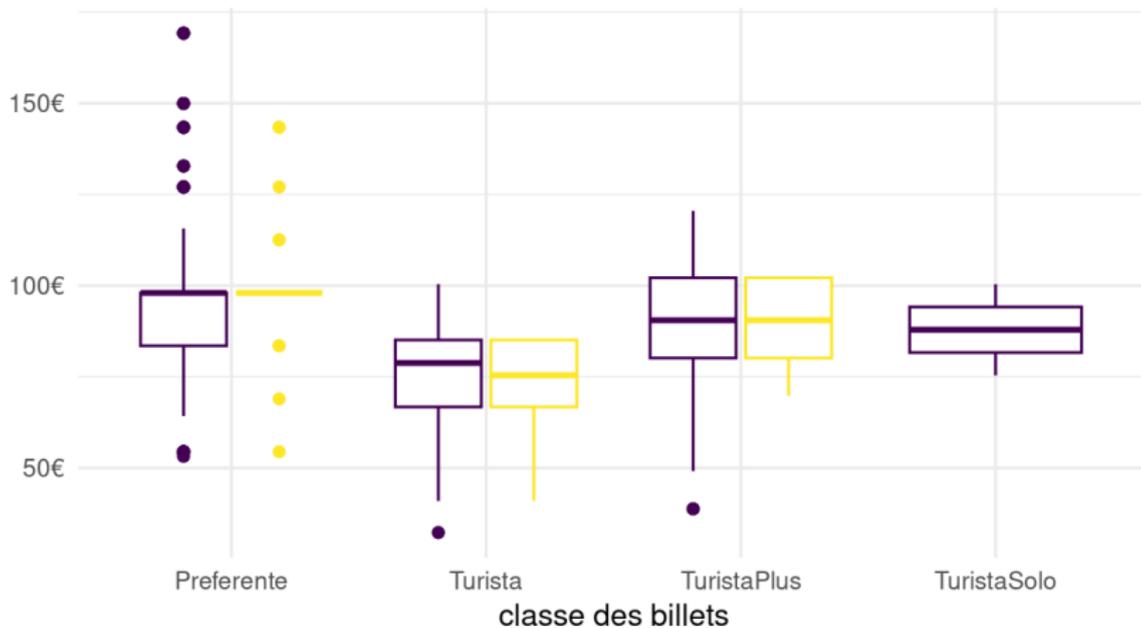
Règle 1: choix de graphiques avec deux variables

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

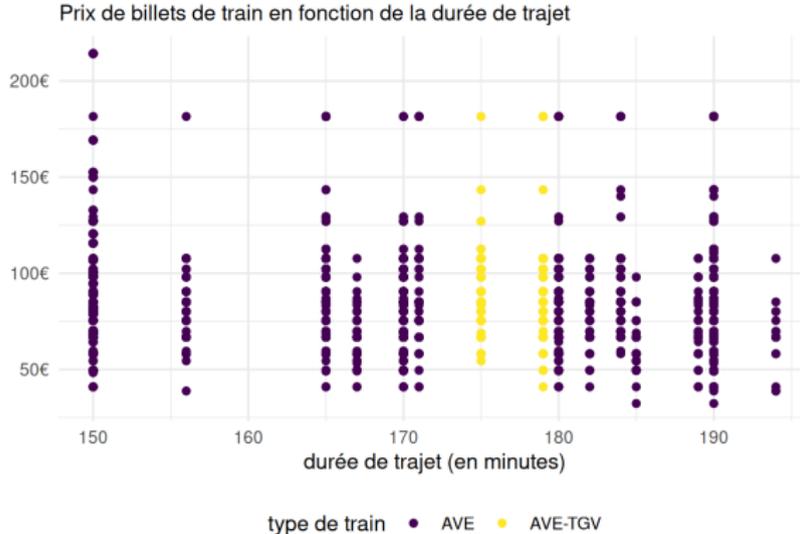
Boîte à moustaches



Prix de billets de train au tarif Promo (en euros)



type de train  AVE  AVE-TGV



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

Certaines visualisations sont plus efficaces/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/superficiés (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

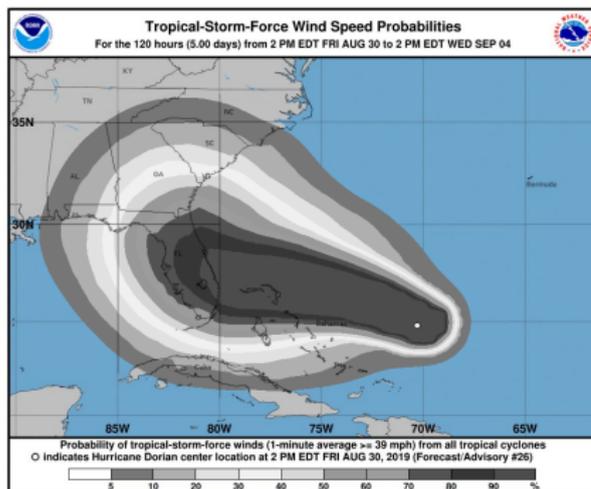


14K 18K 74K



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.



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

John Tukey

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 Golemund et Hadley Wickham