ADV36_TRANSCRIPT - Simplifier la comparaison

Johanie Fournier, agr

2020-02-21

TEASER: Est-ce que c'est toujours nécessaire de passer 4 heures à étudier les données sur toutes les coutures avant de s'arrêter sur une conclusion à présenter?

INTRODUCTION: Ici Johanie Fournier et bienvenue à un nouvel épisode d'Agriculture, Données et Visualisation. Le podcast où je vous apporte avec moi dans le processus de traitement et de visualisation de données pour apprendre à présenter vos propres données de la manière la plus efficace possible. Sans plus tarder, voici l'épisode de cette semaine.

Bonjour et bienvenue dans ce 36e épisode! J'ai une question pour toi:as-tu déjà entendu le cri du lynx? OK c'est bizarre comme question, mais reste avec moi ici parce que je te promets que ça vaut le coup!

Donc, si tu as déjà entendu ça, et bien tu sais de quoi je parle. Sinon prend la peine d'écouter ça c'est carrément hilarant! Tu peux faire une petite recherche rapide sur YouTube ou aller voir dans les notes de l'épisode je t'ai mis un lien.

Maintenant que tu as bien ri, imagine-moi en train de pousser ce cri en regardant la base de données de cette semaine (et oui en passant tu as le droit de rire encore...).

Sur le coup par contre, moi je la trouvais moins drôle...Il y a des semaines comme ça, où ton cerveau ne peux juste pas prendre une analyse de plus (merci la saison du rhume qui touche toute la famille...)

Et ben, ça arrive à tout le monde.

Qu'est-ce qu'on fait dans ce temps-là? Est-ce que c'est toujours nécessaire de passer 4 heures à étudier les données sur toutes les coutures avant de s'arrêter sur une conclusion à présenter? C'est ce que je t'explique aujourd'hui.

Alors, les données de cette semaine sont disponibles comme à l'habitude dans mon article de blogue en lien avec l'épisode de la semaine que tu peux trouver au johaniefournier.com/tyt2020w8.

Cette semaine on travaille avec de données de consommation et d'émission de C02. C'est un sujet très intéressant, ce n'est pas ça le problème... ça juste été une semaine un peu plus difficile que les autres alors, quand j'ai constaté le contenu de la base de données. Eh ben, mon cerveau ne savait juste pas quoi faire avec ça.

La base de données contient 1430 observations. En fait, on a la consommation d'aliment en kg/personne/année et l'émission de c02 aussi en kg/personne/année pour tous les pays et pour 11 catégories d'aliments.

C'est clair qu'il y a une relation entre la consommation de kg de boeuf et l'émission des gaz à effets de serre. Pas besoin de fouiller là-dedans pendant 8 heures pour le savoir.

Par contre, là où mon cerveau a accroché c'était de comment faire pour monter la relation globale entre les deux variables. Avant même d'y avoir travaillé, je savais que je ne devais pas me mettre un objectif cette semaine qui allait me demande beaucoup de traitement et d'analyse (juste pas assez d'espace disque on va dire ça de même).

Montrer deux variables avec des unités différentes sur un même graphique, c'est un casse-tête que je rencontre assez souvent au travail. Et malheureusement, dans tous les cas c'est toujours avantageux de les montrer toutes les deux ensemble... Et ne me parlez pas de faire 2 axes, je ne sais même pas comment faire ça dans R et je n'ai pas l'intention d'apprendre. Je n'aime carrément pas mettre 2 axes sur un même graphique. Et ne me parlez pas des graphiques de suivi avec des échelles amovibles, je vais ressortir mon cri du lynx!

Donc, je voulais montrer la relation globale entre les deux variables consommation et émission de C02, sans toutefois entrer dans les détails et devoir expliquer que certains aliments affectent plus les émissions que d'autres.

Alors, mon premier réflexe a été de regarder la distribution des données pour la consommation et les émissions de co2. En regardant la fréquence relative et un box plot pour chacune de ces variables, j'ai vite compris qu'on n'avait pas affaire à une distribution normale et que les deux variables ont beaucoup de valeurs extrêmes seulement d'un côté de la distribution.

J'ai aussi remarqué que les émissions de co2 et la consommation sont toutes les deux exprimées en kg/personne/année. Dans un cas, on parle de kg d'aliments et dans l'autre on parle de kg de co2, mais c'est des kg pareil. Aussi, les deux variables ne sont pas du même ordre de grandeur. La consommation va de 0 à 400 kg environ alors que les émissions de co2 vont de 0 à presque 2000 kg...

Et là, j'ai eu une idée (et oui mon cerveau a fini par fonctionner...). Pourquoi ne pas simplement mettre ces deux variables côte à côte et comparer les pays au lieu des catégories d'aliments!

PRÉSENTER LE GRAPHIQUE: Donc, j'ai choisi de montrer les deux variables côte à côte avec des dumbells. Ça fait longtemps que je n'avais pas fait ce type de graphique. La plupart du temps, c'est pour montrer une évolution entre un point A et un point B, dans le temps (sur l'axe des x) par exemple. Mais, j'ai eu envie d'utiliser ce type de graphique de manière un peu moins conventionnelle cette semaine. J'ai identifié mon premier point comme étant la consommation d'aliment en kg. Donc, mon axe des x représente les kg/personne/année (aliment et co2 confondu), tous les pays sont sur l'axe de y et j'ai un point pour chaque valeur de consommation en kg pour chacun des pays. Ensuite, j'ai placé un deuxième point, sur la ligne de chacun des pays pour identifier les émissions de co2. La particularité des dumbells est que les deux points sur chacune des lignes sont reliés par un rectangle pour mettre l'emphase si on veut sur l'évolution de la chose.

Je sais que je présente 2 variables différentes, donc dans mon cas ça ne représente pas une évolution au sens ou on est habitué de le voir. Mais, ce que je voulais faire ressortir c'est que c'est bien beau l'influence du type d'aliment qu'on mange sur nos émissions de c02, mais la quantité d'aliments qu'on mange a aussi un impact sur nos émissions de co2.

C'est exactement ce que j'ai fait ressortir dans mon titre. Et voilà, mon graphique est fait, objectif accompli!! Fiou...

ADD: Hey! tu travailles avec R et ça t'intéresse de voir le code que j'ai utilisé pour nettoyer et visualiser mes données? Va voir dans les notes de cet épisode, j'ai mis un lien vers l'article de blogue dans lequel tu pourras trouver tous les détails dont tu as besoin.

REVOIR LES RÈGLES D'OR DE LA DATAVIZ: Donc, si je reviens à ma question initiale: est-ce que ça vaut la peine d'investir beaucoup de temps pour analyser les données avant de choisir une conclusion à présenter? Ben, ça dépend de ton objectif! Clairement, la terre n'allait pas arrêté de tourner si je n'investissais pas 4 heures à très bien connaitre les données du tidytuesday cette semaine... ce n'est pas dans un contexte de travail, c'est pour le fun... Donc, j'ai adapté mon objectif à ma situation et j'ai fait le choix conscient de ne pas aller plus loin dans mon exploration des données.

Dans le cas où c'est ton travail, le premier conseil que je peux te donner est de prendre du recul, aller prendre une marche... t'aérer le cerveau un peu. C'est fou comme on a plus de patience après un grand bol d'air frais.

Aussi, il faut relativiser l'objectif. La phase analyse de données c'est souvent ce qui nous prend le plus de temps... et on a bien souvent des gens qui attendent après nous pour voir les résultats et les conclusions. Alors, peut-être ce que serait possible de faire un prétraitement des données?

Je m'explique, c'est certain que l'objectif global dans la plupart des cas c'est de faire le tour de toutes les informations disponibles et de comprendre toutes les interactions. Mais est-ce que tout ça doit être fait à l'étape 1?

Le processus peut être divisé en plusieurs petites étapes, et ça peut même faciliter le processus d'analyse de prendre l'avis de collègues en cours de route.

Donc, au lieu d'investir beaucoup de temps pour faire une analyse complète pour la première présentation des données, pourquoi ne pas réunir quelques collègues qui sont impliqués, et présenter une première phase de débroussaillage des données et de prendre leur idée et commentaire pour la suite du processus d'analyse?

Séparer le travail en plusieurs petites étapes comme ça et demander l'avis d'autres personnes peut souvent nous faire avancer plus vite. Je dis ça juste de même. Mais il faut que je précise ici que tout le temps investi sur la préparation et l'analyse en vaut toujours la peine... c'est long, mais c'est essentiel.

Alors, ça fait un petit bout de temps que je ne vous ai pas parlé des choix des éléments esthétiques qui composent mes visuels. Je sais bien que ce n'est pas tout de regarder un visuel il faut aussi comprendre les choix derrières pour être capable de répliquer le tout dans son propre travail.

Donc, cette semaine j'avais envie de vous faire un rappel des principes de base de la visualisation de données que j'ai appliqués cette semaine.

Alors, j'ai montré les données. Ça peut sembler bête dit de même, mais ce n'est pas instinctif de mettre les données en avant plan. En fait, les personnes qui consomment nos visualisations le font principalement pour comprendre l'histoire que l'on présente et le personnage principal de ces histoires si on veut et bien c'est les données. Alors, je m'assure toujours de leur donner une place importante.

Ensuite, j'ai réduit l'encombrement au minimum, pas de 3d, pas d'éléments en gras inutile, pas d'icône sur le graphique, bref j'ai conservé seulement l'essentiel.

J'ai aussi intégré du texte pour permettre au lecteur de comprendre mon histoire. Tout d'abord avec un titre qui fait du sens, des étiquettes d'axes et aussi la source des données présentées.

Pour mettre l'emphase sur mon message, j'ai utilisé des attributs préattentifs. Le plus important dans mon travail de cette semaine est les couleurs. J'ai utilisé du mauve pour présenter les données des consommations d'aliments, du noir pour les émissions de c02 et tout le reste est en blanc pour mettre les autres éléments en retrait. J'ai aussi utilisé une légende de couleur directement dans le titre. J'aime beaucoup cette façon de faire parce que ça réduit l'encombrement sur le graphique: pas besoin d'occuper l'espace avec une légende et comme le titre est le premier élément que voir le lecteur, ça a le mérite de clarifier tout de suite les éléments du graphique avant même qu'ils ne se posent des questions sur ce que j'ai présenté comme informations.

CONCLUSION: Voilà, ça fait le tour de ce que je voulais présenter aujourd'hui. Si jamais tu as des commentaires ou des questions, n'hésite pas à me contacter. Tu peux aller au johaniefournier.com/contact pour m'écrire directement ou aller dans la section commentaire de l'épisode pour poser tes questions, ça va me faire plaisir de te répondre. Alors, j'espère que cet épisode a été utile et que tu as appris quelque chose, merci de m'avoir écouté et on se dit à la semaine prochaine!

Quelques liens utiles:

- Pour rire un peu: le cri du lynx
- Pour écouter l'épisode: ici
- Transcription de l'épisode: ici et en pdf ici
- Le graphique discuté: ici
- L'article de blogue en lien avec cet épisode: blogue
- Me contacter: contact

Tu aimerais avoir par écris ce processus de dataviz? J'ai mis toutes les étapes que je réalise à chaque semaine pour créer mes visuels dans un aide-mémoire: Le Processus Dataviz

Es-tu abonné à mon podcast? Sinon, je t'encourage à t'inscrire dès aujourd'hui, il ne faudrait pas manquer un épisode! Clique ici pour t'inscrire sur iTunes.

Si tu te sens particulièrement gentil aujourd'hui, je te serais très reconnaissante si tu me laissais un commentaire sur iTunes. Les commentaires aident les gens à trouver mon podcast et ils sont aussi très précieux pour moi. Sélectionne "Notes et Avis" et ensuite "Rédiger un avis" et fais-moi savoir quel est ton épisode préféré. Merci!