ADV37_TRANSCRIPT - La fois où je n'ai pas fait de graphique

Johanie Fournier, agr

2020-03-01

TEASER: Ça y est... C'est finalement arrivé... Je n'ai pas fait de graphique cette semaine...

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 la semaine.

Et oui, c'est officiel, je n'ai pas fait de graphique cette semaine... Mais attends! J'ai travaillé fort quand même pour comprendre les données. Mais, toutes les avenues que j'ai explorées ne donnaient rien de bon.

C'est rare quand même que ça arrive.

Et quand j'ai finalement trouvé une avenue intéressante pour présenter les données. Eh ben... Je n'ai pas compris la relation entre les données que j'avais devant les yeux.

Cette semaine, je t'explique mon processus de réflexion et mon travail avec ces données assez particulières.

Alors, les données traitent du taux de vaccination de la rougeole dans différents états des États-Unis. Comme d'habitude, le lien pour avoir accès aux données se trouve dans mon article de blogue que tu peux trouver au johaniefournier.com/tyt2020w9.

La base de données de cette semaine contient plus de 66 000 observations. Pour chacune de ces observations, on connait l'état, l'année, le nom de l'école, le type d'école, la ville, le compté, le nombre d'étudiants inscrits, le taux de vaccination pour la rougeole, le taux de vaccination globale, 3 variables qui nous décrivent les raisons pour ne pas être vaccinées (personnelles, religieuses et médicinales) et des variables de positionnement (longitude, latitude).

Jusque là, tout va bien... normalement, après avoir téléchargé la base de données et pris connaissance des différentes variables, je prends un peu de temps pour faire une exploration plus générale des données pour identifier un angle qui m'intéresse particulièrement pour donner une direction à mon analyse en quelque sorte.

C'est là que ça c'est compliqué cette semaine. À première vue, cette base de données est tout ce qu'il y a de plus standard. Rien dans sa forme ou son format n'est différent de ce que je traite chaque semaine.

Là où ça se gâche, c'est dans le nombre de données manquantes. C'est la première fois que je vois une bd avec autant de données manquante.... à mon sens, c'est assez important pour rendre l'analyse des données difficile voir impossible.

Je t'explique un peu mon exploration et ma découverte de la face cachée de la BD pour que tu comprennes un peu mieux ce que je veux dire.

Donc, cette semaine, j'ai appris qu'il y avait un nouveau package de disponible pour ggplot, la fonction que j'utilise dans R pour faire des graphiques. Cette nouvelle fonction permet de faire des graphiques de rang avec un look très cool!

Donc, je ne te cacherai pas que j'avais bien envie d'essayer le tout cette semaine.

La première chose que j'ai regardée dans les données sachant que j'avais en tête de visualiser l'évolution d'un quelconque paramètre est la variable année.

Premièrement, les données de cette variable sont stockées sous forme de caractère, ça part mal comme on dit...

On a accès à des données de l'année scolaire 2017-2018, 2018-2019 donc 2 années scolaires complète et l'année 2017. Ce n'est pas indiqué si le 2017 réfère à l'année scolaire 2016-2017 ou si c'est toutes les données de l'année civile 2017 ce qui reviendrait donc à dupliquer les données de l'année scolaire 2017-2018...

Bref, dans le doute je crois que tout ce qui est identifié 2017 ne peut être retiré.

En plus, il y a 5681 lignes de données qui n'ont pas de valeur d'année.

Restent donc les deux années scolaires 2017-2018, 2018-2019. OK, jusque là, ça allait encore. Je peux travailler avec ça...

Ensuite, j'ai regardé les différents types d'écoles. On dispose de 6 différents types d'écoles. Certaines d'entres elles sont présentes en grand nombre que les écoles publiques pour lesquelles on dispose de presque 21 000 observations, mais d'autres sont disponibles en très petit nombre comme les BOCES qui sont présente seulement 47 fois.

Ça aurait pu fonctionner mon affaire, graphique de rang des différents types pour les 2 années scolaires disponibles.

J'étais sur le point d'être convaincue d'avoir trouvé mon objectif de la semaine jusqu'à ce que je vois le nombre de données manquantes: 36 622 observations n'ont pas de type d'école... plus de la moitié de la BD... ouin mis avec la grande variation dans le nombre d'écoles par type c'est pas terrible finalement...

Pour finir de me convaincre, j'ai regardé s'il y avait des données pour chacune des années par type d'école... et bien non...

Il y a seulement les écoles privées et publiques pour lesquelles on dispose de données sur les années scolaires 2017-2018, 2018-2019.

Bon, j'étais presque résigné à utiliser seulement 2 types et 2 années scolaires pour mon graphique de rang. Ça peut le faire... Par contre, en regardant de plus près les données, j'ai vu que le taux de vaccination pour les écoles publiques diminue beaucoup pour l'année 2018-2019.

OK, c'est intéressant. Par contre, quand j'ai essayé de pousser plus loin l'analyse question d'expliquer un peu le tout. Je me suis vite rendu compte que les pourcentages donnés dans les 3 catégories qui expliquait le taux de vaccination de la rougeole (religieux, personnel et médical) et bien ça ne fonctionne pas avec les taux de vaccination.

J'ai aussi essayé de décomposer le taux de vaccination en fonction du nombre d'étudiants inscrits, pour me rendre compte trop tard que dans la variable taux de vaccination de la rougeole, il y avant des données -1, ce qui ne se peux pas... un taux de vaccination de -1 c'est juste impossible. Donc c'est des valeurs manquantes.

Qui identifie des valeurs manquant avec des -1?!? Pas moi... ça ne m'est même jamais passé par l'esprit d'utiliser une valeur numérique dans une colonne de données numériques pour identifié des données manquantes.

Peut-être que c'est la convention dans certains domaines et j'aurais du être plus attentive honnêtement, j'ai été traumatisée par le nombre de valeurs manquantes dans cette bd là.

Et ensuite, c'est quand même assez important d'être conséquent dans la nomenclature... si tu utilises -1, c'est -1 partout... si tu utilises NA c'est NA partout! mais bon, je suis perfectionniste un peu sur les bords...

Bref, cette avenue-là non plus ne fonctionne pas: avec ou sans les -1... je n'ai pas trouvé rien de clair pour expliquer les différences de taux de vaccination.

Ensuite, j'ai décidé de prendre un peu de recul et de revenir à un niveau plus général dans la base de données. Et si je regardais la distribution des données par état...

Et bien, j'ai été déçu ici aussi, aucun des 21 états présents dans la BD n'avait des données pour les deux seules années scolaires complètes...

Bon, dernière tentative avant de me mettre à pleurer... et si je regardais la relation entre le taux de vaccination de la rougeole et e taux de vaccination globale des écoles... c'est a un niveau assez global pour que toutes les données manquantes n'aient pas trop d'impact je pense...

Alors, avec toute l'énergie qui me restant, j'ai regarder le patron de la distribution de ces deux variables... Ce ne sont pas des distributions normales et il y a énormément de valeurs extrêmes... Ça ne me disait rien de bon, mais bon... Je voulais à ce pois-là, j'avais abandonné l'idée de créer mon beau graphique de rang et je voulais montrer seulement la relation entre les deux taux de vaccination. Donc, j'ai cru que ça fonctionnerait...

Et ben, je n'étais pas au bout de mes surprises... Comment est-ce que le taux de vaccination pour la Rougeole peut être au-dessus du taux de vaccination globale pour ces écoles des États-Unis?? Rendu là, mon cerveau a carrément buggé... (y était rendu tard aussi pour ma défense...)

Bref, j'ai décidé de demander de l'aide à la communauté question de me faire expliquer ces deux variables parce qu'il y avait carrément là-dedans quelque chose que je ne comprenais pas...

Et ben oui c'est possible que le taux de vaccination de la rougeole soit plus élevé que le taux de vaccination globale. C'est même tout ce qu'il y a de plus normal en fait. Ça devient même très clair à partir du moment qu'on comprend que le taux de vaccination globale implique que ce sont les étudiants qui ont reçu tous les vaccins requis par leur calendrier de vaccination...

Finalement, je n'ai pas fait de graphique cette semaine... J'ai passé beaucoup de temps à explorer et à comprendre la base de données. Je pense que ce qu'il faut retenir de mes péripéties de cette semaine c'est qu'il ne faut jamais prendre pour acquis qu'on sait comment faire pour faire l'analyse d'une BD... parce qu'il y en a toujours une qui va nous surprendre au moment où on s'y attend le moins.

Et, c'est d'autant plus important d'être structuré dans son processus de révision avant de procéder à l'analyse. Je sais que chaque bd des différentes, mais je vais quand même profiter de cette expérience pour m'écrire un petit processus de validation des bds avant analyse, juste question de m'assure d'avoir une liste qui va me permettre de ne rien oublier.

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 écouter l'épisode: ici
- Transcription de l'épisode: ici et en pdf 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!