ADV23_TRANSCRIPT - Connaitre et comprendre ses données

Johanie Fournier, agr

2019-10-18

TEASER: C'est bien de connaitre les trucs qui permettent d'améliorer la présentation de nos graphiques. Mais ce qui est encore mieux c'est de pouvoir expliquer les données qui sont présentées. Et ça, ça passe par une bonne connaissance et une bonne compréhension des données qu'on travaille.

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 23e épisode. J'espère que vous allez bien et que vous être bien reposé par ce qu'aujourd'hui on parle de "chars"!

Les données de cette semaine sont disponibles sur le site de fueleconomy.gov et sont une gracieuseté de EPA l'Agence de la protection de l'environnement des États-Unis. Le lien pour les télécharger se trouve dans mon article de blogue au johaniefournier.com/tyt2019-42.

Les données mises à notre disposition pour ce Tidytuesday sont l'équivalent de la populaire base de données « mtcars » de R, mais en beaucoup beaucoup plus gros...

Juste pour vous donner une idée de notre terrain de jeu pour cette semaine, on dispose d'une base de données qui contient 41632 lignes et 83 colonnes. Je n'expliquerai pas l'ensemble de ses données, mais il faut savoir qu'il y en a a pour tous les gouts là-dedans, des données depuis 1983, des données d'émissions de C02, des données de performances (MPG= milles aux gallons), des sources de carburant distinct pour les véhicules électriques et au mazout... Bref, pour passer tout ça en revue, on a de quoi s'amuser pendant plusieurs jours.

Avec des bases de données comme celle-là, les possibilités d'analyse et de visualisation sont quasi infinies. Alors, je n'ai pas tellement cherché à regarder dans les données pour identifier une conclusion intéressante à présenter faute de temps, mais je me suis plus tôt demandé ce que moi j'aimerais savoir, ce que j'aimerais voir avec ces données. Alors, parce que j'ai toujours conduit des véhicules Ford, je me suis fixé comme objectif de visualiser l'évolution des performances (litres au 100 km) des véhicules Ford comparativement à la moyenne des autres fabricants de véhicules.

Comme vous avez pu le deviner, les données de cette semaine présentent un gros défi de traitement et d'analyse avant de pouvoir être utilisées correctement. Je ne vais pas entrer dans les détails tout de suite parce que c'est l'object de la deuxième partie de l'épisode, alors je vous explique toute de suite mon visuel et l'on reviendra sur tout le travail de préparation et de compréhension que ça a nécessité cette semaine.

CHOIX DU TYPE DE VIZ: Donc, pour pouvoir comparer les performances des Ford à la moyenne autres véhicules des différents fabricants au fil du temps, je devais créer un visuel qui me permettait de voir l'évolution des Ford, la moyenne des autres véhicules et la distribution des autres véhicules pour pouvoir situer le résultat des Ford et ce sur une ligne du temps. Le mieux, à mon avis dans cette situation est de présenter l'ensemble des données avec des points pour voir une belle distribution, présenter la moyenne de ces données avec un point plus gros pour chaque année et montrer les ressuât des Ford avec une courbe.

PRÉSENTER LE GRAPHIQUE: Alors, j'ai présenté l'ensemble des performances de chacun des véhicules présent dans la base de données avec des petits points gris pour lesquels j'ai appliqué une notion de distribution aléatoire pour qu'ils soient distribués dans un certain espace autour de leur valeur sur l'axe des x et j'ai utilisé la transparence pour qu'on puisse juger de l'importance du nombre de points à un endroit précis. La moyenne et l'écart type sont identifiés respectivement pas un gros point gris et des lignes grises. Seulement

les petits points gris qui représentent les données ce n'est pas assez à mon avis pour voir rapidement la tendance derrière les données. Donc, en rajoutant les éléments pour la moyenne et l'écart type, je pense que j'ai été chercher plus de clareté sur l'évolution des performances en litre au 100 km des véhicules construits entre 2000 et 2019 toutes marques confondues. Donc, une fois que la série de données de référence est bien mise en place, j'ai utilisé une courbe pour montrer l'évolution des performances de l au 100 km des véhicules Ford. Avec cette ligne bleue, on voit rapidement les années ou les Ford ont été plus performants et les années où ils ont été moins performants que les autres marques. J'ai aussi pris la peine de séparer les pick-up des autos. Dans ma tête ces deux types de véhicules ne sont pas utilisé pour les mêmes raisons, ne sont pas non plus conçus de la même façon et ça fait juste du sens pour moi que la consommation d'essence ne soit pas la même. J'ai aussi pris la peine d'inverser l'axe des y. Contrairement au MPG les litres au 100 km, on cherche a avoir la plus petite valeur possible. Donc pour que mon graphique soit plus instinctif à regarder, j'ai inversé l'axe des y. Comme ça, en 2015, on voit que les Ford on des consommations d'essence au 100 km inférieur aux concurrents. Comme c'est une bonne nouvelle, on ne voudrait pas voir la courbe en bleu descendre... par ce que ça nous laisserait croire que les performances sont moins bonnes. J'ai aussi pris la peine de créer un titre accrocheur, qui contient aussi la légende du graphique, et d'expliquer les conclusions et résultats en sous-titre. Avec cette visualisation, j'arrive à la conclusion que les véhicules Ford ont vu leurs performances améliorer autour des années 2015, mais les autres fabricants les ont rattrapés dans les années suivantes.

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: Bien connaitre et comprendre les données qu'on présente c'est essentiel. C'est essentiel pour pouvoir présenter des graphiques qui contiennent des conclusions intéressantes, mais c'est aussi essentiel pour pouvoir expliquer à notre audience les conclusions et les différents aspects qui peuvent influencer ces conclusions et surtout les actions qui seront prises suite à la présentation de ces résultats. C'est quand même assez rare qu'on présente des visualisations pour le simple plaisir. Ce n'est pas impossible et c'est l'fun de s'amuser de temps en temps. Mais, si on est honnête, on présente des graphiques la plupart du temps pour répondre à des questions de notre audience. Et qui dit audience qui a des questions, dit audience qui veut des réponses et audience qui connait le sujet et audience qui aura d'autres questions suite à la présentation des résultats puisqu'ils ont a coeur de bien comprendre. On ne peut pas les blâmer, dans un contexte d'entreprise, les résultats présentés serviront à prendre des décisions parfois importantes donc c'est notre rôle de présenter des graphiques les plus clairs possible et surtout de bien comprendre les données qu'on présente pour pouvoir les expliquer et répondre aux questions.

Pour vous illustrer le tout, j'ai repris le visuel que j'ai créé cette semaine et j'ai retirer touts les éléments que j'ai travaillés suite à mon analyse des données. Le lien pour voir ce graphique se trouve dans les notes de l'épisode. Je présente toujours les années sur l'axe des x et les consommations d'essence en litre au 100 km sur un axe des y inversé. Mais, je n'ai pas fait les distinctions entre les types de véhicules (auto vs pick-up), je n'ai pas pris soin de comprendre ce qui se trouve dans la base de données, donc les points en gris représentent tant les autos et les pick-up, mais inclus aussi les véhicules électriques et les véhicules qui fonctionnent avec 2 types de carburant. Même sans tout ce message et cette analyse des données, j'arrive à la même conclusion: Ford était plus performant autour des années 2015 qu'il ne l'est maintenant. Les différences que j'ai constatées dans mon premier visuel sont juste masquées par un bruit de fond. En fait, je pense que j'ai été chanceuse que ma conclusion soit la même c'est probablement seulement dû au fait que j'ai travaillé avec une très grosse base de données. Les tendances sont plus marquées dans ces cas-là. Par contre, même si le visuel en lui seul est plus qu'acceptable, je ne suis pas à l'aise de le présenter parce qu'il y a un manque au niveau de l'analyse des données.

Je pense que c'est difficile de faire la moyenne de toutes les performances de litres au 100km parce que ça implique que l'on considère sur un même pied d'égalité toutes les séries qui sont faites par un fabricant. Les véhicules compacts ne sont pas conçus pour les mêmes raisons que les pick-up et n'ont pas les mêmes performances non plus. Donc, on se retrouve dans une situation ou le nombre de véhicules de chacune de ses catégories influence la moyenne. Si une marque se spécialise dans les véhicules de travail par exemple, on vient biaiser notre analyse en utilisant une moyenne... Et je peux vous garantir qu'un public averti va

vous poser la question... Donc, à mon avis c'est plus qu'essentiel de séparer au moins en 2 catégories pour comparer des pommes avec des pommes.

Ensuite, pour moi les véhicules qui ont 2 sources d'énergie devraient être considérés à part. D'une part parce que ça pose problème dans le calcul de MPG. On doit se baser sur un MPG ajusté pour un équivalent en essence et entre vous et moi, 150 milles au gallon, même si c'est une mesure ajustée c'est clair qu'on n'est pas dans la même game qu'une voiture qui fait 15l au 100 km. Donc, je ne pense pas que ce soit mauvais de comparer des véhicules traditionnels avec des véhicules biénergie, je dis seulement qu'on doit comparer des pommes avec des pommes et dans mon cas je pense que c'est préférable de ne pas les inclure dans mon analyse. Pour pousser plus loin cette comparaison, je ferai un visuel séparé...

Tout ça c'est un processus ou l'on se reprend souvent parce que je me rends compte d'une erreur possible ou bien je me pose une question dans le visuel et je dois retourner dans la section préparation de données pour comprendre/corriger....Cest un avantage considérable de travailler dans R. C'est rapide de faire des petits changements sans tout bouleverser ou de subdiviser pour mieux comprendre...

Alors, en résumé je pense que mettre des chiffres dans un graphique ne fait pas de nous des analystes de données. C'est le fait de connaitre de et comprendre les données avant de communiquer les résultats de nos analyses à notre audience qui fait de nous de bons analystes. Par contre, comprendre des données ça peut être un défi de taille. Et ce qui est certain c'est que ça consomme du temps. En général, on dit qu'on passe 80% du temps à préparer et à analyser les données et 20% à créer les graphiques. Avec les Tidytuesday j'ai la chance de pouvoir travailler des données pour lesquelles il y a déjà un travail de préparation de données qui a été fait, mais je sais pertinnement que ce n'est pas toujours le cas dans la vrai vie.

Donc, quand vous faites face à une nouvelle base de données à analyser, je pense qu'il y a 3 grandes questions à se poser pour débuter le travail et de s'assurer de partir sur des bases solide: 1) Quelle est la meilleure façon de présenter les données? Donc on touche ici tant au type de graphique, qu'à l'unité de mesure présenter (jour, mois, année), qu'au travail de décortication des données qui doit être faites avant de faire le graphique. Il faut chercher à identifier ce qui permet de mieux visualiser la conclusion qu'on veut mettre de l'avant et de s'assurer qu'on a pris le temps de bien comprendre les résultats et les paramètres qui viennent les influencer. 2) Quelle est la meilleure unité de mesure à présenter? Il faut présenter des résultats qui parlent à notre audience et faire tout le travail de conversion nécessaire pour leur faciliter la vie. Dans mon exemple, j'avais accès à des consommations d'essence en MPG mais c'est une unité de mesure américaine, ici on parle plus de litres au 100 km. Donc, il faut être prévenant et convertir tout ça et je dirais même aussi de vérifier les calculs qui sont déjà faits. il n'y a rien de pire que de prendre pour acquis que les calculs présents dans une base de données sont exacts, de faire une analyse complète avec ces données et de s'apercevoir en plein milieu d'une présentation que le calcul est erroné. Personne n'est mal intentionné, mais ce n'est pas tout le monde qui est à l'aise avec les chiffres... 3) Est-ce qu'il y a des erreurs dans la base de données? C'est malheureusement très fréquent et il faut faire avec. Chacun peut se donner des barèmes à respecter en fonction des données travaillées, mais le plus important c'est de savoir comment les identifier pour notre situation précise et surtout de pouvoir expliquer les raisons derrière notre choix de les garder ou non...

Dans le fond, ce qui est important c'est de faire attention aux détails et d'être honnête envers soi-même et notre audience sur la qualité de notre analyse.

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!

Tu as aimé le contenu de cet épisode? Il est temps d'aller écrire une évaluation sur iTunes ou sur ta plateforme préférée et de t'abonner à mon podcast pour être avisé lors de la sortie du prochain épisode. Bonne semaine et amuse-toi bien à visualiser tes données! Quelques liens utiles:

- Le transcript de l'épisode en pdf
- Pour écouter l'épisode: johaniefournier.com/22

- Les graphiques discutés: ici et ici
- L'article de blogue en lien avec cet épisode: blogue
- Me contacter: contact