ADV34_TRANSCRIPT - Quand on ne sait pas de quoi on parle...

Johanie Fournier, agr

2020-02-08

TEASER: Qu'est-ce qu'on fait quand on ne connait rien du sujet sur lequel portent les données qu'on doit visualiser?...

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 34e épisode! J'ai une petite anecdote à raconter pour bien commencer. En fait, cette semaine j'ai été témoin d'une conversation intéressante qui va comme suit:

A: Et si je regardais combien de matchs se sont terminés à un point d'écart? Parce que la seule façon de le faire est de remporter un field gold..

B: Euh ben... il y a aussi le point après le touchdown qui est un point...

MOI: (OMG! mais je n'y comprends absolument rien...) Aussi moi...: (et il faut que je fasse une viz avec ça!?!)

C'était juste après avoir découvert que le Tidytuesay de la semaine portait sur des données de la NFL. Je ne suis pas une fan de sport... donc pas besoin de vous dire que ce n'est pas le sujet que je maitrise le mieux.

Pour être tout à fait honnête, j'ai eu envie de passer mon tour et de vous faire une belle visualisation sur le dernier article scientifique qui vient de sortir et qui traite du comportement à l'abreuvoir issu du stress thermique chez les vaches laitières... là ça aurait été intéressant!

Mais, ce n'est pas à quoi je me suis engagé et ce n'est pas parce qu'on est moins à l'aise avec des données qu'il faut jeter l'éponge avant même d'avoir commencé. Alors, aujourd'hui j'explique comment je m'y suis pris pour me retrousser les manches et faire le travail qu'il faut pour obtenir une viz qui est quand même assez intéressante.

Donc, comme vous l'aurez surement deviné, les données de la semaine portent sur la NFL plus précisément sur le public qui assiste au match de la NFL. Le lien pour y avoir accès aux données trouve dans mon article de blogue au johaniefournier.com/tyt2020w6.

Chaque ligne de la base de données qui a été fournie nous renseigne sur le nom de l'équipe, l'état pour lequel elle joue, l'année (2000 à 2019), l'assistante totale pour l'année, l'assistance aux matchs qui ont été joués à la maison,l'assistance aux matchs qui ont été joués à l'extérieur, le numéro de la semaine et l'assistance de cette semaine-là.

Alors, après être passé par dessus mon non-intérêt et ma non-connaissance du sujet, j'ai pris la peine d'explorer un peu cette base de données et de me familiariser un peu avec les différentes variables. À mon avis, le point principal de ces données est l'assistance aux matchs, mais aussi la comparaison entre l'assistance des matchs joués à la maison vs ceux joués à l'extérieur. À ce moment-là dans mon processus je n'aurais pas pu expliquer à personne ce que ça pouvait bien vouloir dire, mais j'ai quand même décidé que ce serait mon objectif de la semaine: visualiser l'évolution de dans le temps de la différence entre l'assistance à la maison et l'assistance à l'extérieur.

Petite parenthèse ici, je pense qu'avec la base de données très simple qui a été fournie cette semaine, c'était assez difficile de me tromper et de passer à côté du point principal des données. Il n'y avait carrément pas d'autre chose à présenter. Le seul choix que j'ai dû faire en fait a été de décider si je voulais présenter

l'assistance sur une base de semaine ou d'année. Si j'avais été en face d'une base de données plus complexe, donc avec plus de variables parmi lesquelles choisir et traitant d'un sujet que je ne maitrise pas, j'aurais été faire plus de recherche avant de débuter mon processus de visualisation...

CHOIX DU TYPE DE VIZ: Donc pour atteindre mon objectif de la semaine, j'ai choisi de créer un graphique qui allait présenter la différence entre le pourcentage de l'assistance à la maison et l'assistance à l'extérieur avec une aire sous la courbe.

À l'épisode 18 (je vais mettre un lien pour aller l'écouter dans les notes) j'ai expliqué comment ne pas faire d'aires sous la courbe parce que ce n'est pas un type de graphique qui est adapté à toutes les situations. Dans le cas de cette semaine, je présente 2 valeurs en pourcentage et la somme des 2 me donne 100 donc, je pense que c'est une situation parfaite pour présenter des aires sous la courbe.

Ensuite, comme la base de données contient toutes les équipes et que je veux montrer l'évolution dans le temps pour chacune des équipes, j'ai utilisé une fonction qui s'appelle facet_wrap pour créer un visuel qui est composé de plusieurs petits graphiques, 1 par équipe.

PRÉSENTER LE GRAPHIQUE: Donc, chacun de mes petits graphiques présente une équipe. L'axe des x va de 2000 à 2019 puisque c'était les données qui étaient disponibles. Sur chacun des graphiques, il y a une seule ligne horizontale qui représente le pourcentage de l'audience totale dans la catégorie à la maison. L'axe des y représente le pourcentage de l'audience. Pour identifier les deux catégories soit à la maison et à l'extérieur, j'ai utilisé des couleurs différentes pour l'aire sous la courbe représentée par la ligne du pourcentage d'audience à la maison et pour l'aire entre la limite du 100% et la même ligne du pourcentage d'audience à la maison. Je vous rappelle que la somme de mes deux variables donne 100% donc nécessairement, sur une échelle de 100 je n'ai pas besoin de montrer mes deux variables si j'utilise l'aire sous la courbe de chacune des deux sections.

Alors, l'audience à la maison est identifiée en bourgogne et l'audience à l'extérieur en bleu vert. À ça j'ai ajouté un titre qui nous dit de quels sujets on parle et un sous-titre qui contient la légende de couleur. Et voilà, mon graphique était fait!!

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: J'ai donc créé une superbe viz, mais il y a un problème... je n'ai aucune idée de ce que ça peut bien vouloir dire...et en plus, ça ne fait pas un visuel si intéressant. Beau, mais pas si intéressant à interpréter parce qu'il y a très peu de variation qui est visible lorsque les données sont présentées de cette façon...

C'est ici que je vous explique ce qu'on fait avec des données qu'on ne connaît ou qu'on ne comprend pas: c'est très simple... on demande de l'aide!! J'ai publié ma version non complète de mon travail de cette semaine à la communauté de personne qui participe au Tidytuesday et j'ai demandé au fan de sport de m'aider avec le contexte pour interpréter ces données.

C'est magique ce qui se passe quand on demande de l'aide...En fait, c'est assez simple: on obtient de l'aide. De l'aide qu'on n'aurait pas reçue comme par magie si on ne l'avait pas demandé... Et la communauté R est absolument incroyable.

Alors, en demandant de l'aide, j'ai pu apprendre que les stades de la NFL sont à peu près tous de la même taille. Ce qui implique que le même nombre de personnes ou à peu près peuvent assister aux matchs, peu importe dans quel endroit aux États-Unis il est joué. Alors sachant ça, ça veut dire que la variation ou le manque de variation en fait dans mon graphique n'est pas lié à l'audience qui participe au match, mais bien aux nombres de matchs qui sont joués à la maison vs le nombre de matchs qui sont joués à l'extérieur. Ne me demander pas ce qui influence ce paramètre, je n'en ai absolument aucune idée!! Mais, c'est déjà une bonne piste d'explication.

En plus, avec ces pistes de réflexion pour améliorer mes connaissances sur le sujet, j'ai aussi eu une suggestion très intéressante pour améliorer mon graphique. Comme la variation ou le manque de variation en fait se

situe autour de 50%, on peut améliorer notre vision de la chose en visualisation la déviation de l'assistance à la maison du 50% espéré.

Donc, j'ai dû revenir dans mon processus de préparation des données pour venir ajuster le tout. Au lieu de calculer seulement le pourcentage d'assistance à la maison, j'ai calculé le pourcentage d'assistance à la maison moins 50 et j'ai utilisé cette variable pour faire mon graphique.

De cette façon, ce n'est pas possible de mettre aussi la zone de couleur pour l'audience à l'extérieur, mais c'est quand même implicite dans le visuel. Ce qui est intéressant avec cette façon de faire est qu'on met réellement en évidence la variation.

On peut voir que pour la majorité des équipes, il y a peu ou pas de variation qui valent la peine d'en parler, mais pour les Cowboys et les Redskins, là ça vient intéressant à regarder.

Alors, pour les Cowboys, on peut voir que l'a différence du 50% augmente dans le temps avec une cassure drastique autour des années 2010. Donc, l'assistance à la maison pour les Cowboys augmente dans le temps. Si on se fie à l'explication que j'ai eue, ça veut dire que les Cowboys jouent plus de matchs à la maison qu'à l'extérieur. Pourquoi, aucune idée, mais c'est déjà quelque chose!

Le phénomène inverse se produit pour les Redskins, la différence du 50% de leur audience à la maison diminue dans le temps, ce qui veut dire que leur audience à la maison diminue dans le temps de manière assez linéaire.

Au final, je pense que mon graphique de cette semaine est un bel exemple de transformation de données pour explorer le parton de données et comparer différentes variables entres-elles.

Une fois que les régions avec le plus de variation, ou le patron de données le plus intéressant sont identifiés, on peut pousser plus loin notre analyse et trouver des explications aux phénomènes observés.

En conclusion, quand on ne connait pas un sujet on demande de l'aide et c'est possible de faire des aires sous la courbe (je vous l'avais bien dit!!) ça prend juste les bonnes données pour le faire.

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
- Transcription de l'épisode en pdf: ici
- Les graphiques discutés: ici et 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