ADV31_TRANSCRIPT - Identifier et visualiser l'important

Johanie Fournier, agr

2020-01-18

TEASER: Faire un graphique c'est facile... mais pour faire un visuel qui transmet bien les éléments importants qu'on a trouvés dans les données, il faut y réfléchir un petit peu plus longtemps...

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 31e épisode et premier épisode de 2020! D'abord, je tiens à vous souhaiter une bonne et heureuse année 2020! Pour ma part, je suis vraiment contente d'être de retour pour jaser de dataviz. Le congé du temps des fêtes m'a permis de planifier un peu ce qui s'en vient pour le podcast cette année et j'ai vraiment hâte de partager tout ça avec vous. Donc pour la première section de 2020 (de janvier à mai), je continue à vous jaser des visualisations que je fais chaque semaine avec les données des tidytuesday. Ensuite, je vais prendre une petite pause en mai, travail oblige. Je dois me concentrer sur le dépôt de mes PAEFs. Ensuite, je serai de retour fin juin toujours pour vous parler de visualisation de données, mais j'ai envie d'intégrer plus d'agriculture dans tout ça. Je ne sais pas trop encore quelle forme ça va prendre, mais à partir de cet été, c'est certain que j'intègre plus de données agricoles dans mon podcast.

Alors, cette semaine, on jase de mot de passe. Les données proviennent de Knowledge is Beautiful et ils nous ont fourni une base de données de plus de 500 mauvais mots de passe. Le lien pour y avoir accès se trouve dans mon article de blogue au johaniefournier.com/tyt2020w3

Chaque ligne de leur base de données nous renseigne sur les mots de passe étudiés. On a le mot de passe en question, la catégorie dans laquelle il se classe (tous les mots de passe ont été divisés en 9 catégories), on dispose aussi de 2 valeurs de durée qui nous donne le temps que ça prend pour découvrir le mot de passe et une valeur de force de sécurité du mot de passe qui devrait normalement aller de 1 à 10 avec 10 étant un mot de passe très sécuritaire.

Alors, la première chose à faire ici, quand notre but est des mettre des données en graphique pour transmettre un message est de se fixer un objectif clair. Par objectif clair, j'entends : qu'est-ce qu'on veut montrer dans notre visuel? C'est certain que en général, les tidytuesday sont des petites bases de données qui demandent peu de préparation, donc je n'ai pas besoin de passer plusieurs heures à travailler les données pour voir ce que je souhaite en sortir, mais je prends tout de même la peine de faire un petit tour de ce qu'il y a de disponible question de bien saisir le sujet, me familiariser avec les données disponibles et surtout d'identifier ce qui m'intéresse le plus. Comme je ne suis pas dans un contexte d'entreprise ici, j'ai la chance d'avoir le choix de ce que je peux identifier comme important à montrer. Donc, en général pour identifier mon objectif, je prends la peine de regarder les données, de me faire une idée de travail qu'il y aura à faire et la majorité du temps, il y a un élément dans les données qui pique ma curiosité.

Dans les données de cette semaine, ce qui a piqué ma curiosité est la valeur de sécurité des mots des passes et les catégories. Est-ce qu'il y a des catégories de mots de passe qui sont plus sécuritaires que d'autres? Et voilà, c'est comme ça que j'ai identifié mon objectif.

C'est important de noter ici que c'est mon objectif pour la création de 1 visuel. Avoir un un objectif clair de commencer à travailler, permet de créer un visuel clair.

Il faut aussi comprendre que mon objectif de la semaine ne permet pas de faire l'analyse de toutes les données qui sont disponibles. Je ne suis pas dans un contexte ou je dois faire l'analyse de tout ce qui se trouve dans cette base de données. Si c'était le cas, j'aurais procédé en identifiant plusieurs petits objectifs pour créer

plusieurs visuels question de bien démontrer tous les points que je trouve importants dans les données, mais en prenant soin de ne pas tous les montrer dans le même graphique.

CHOIX DU TYPE DE VIZ: Pour atteindre mon objectif de cette semaine qui était de montrer quelles sont les catégories qui ont les mots de passe les plus forts. La première chose que j'ai dû faire a été de déterminer comment j'allais bien pouvoir faire pour traduire la valeur de force de sécurité du mot de passe (qui est une note de 1 à 10) en une catégorie fort, moyen, faible...

La première chose que j'ai faite a été d'aller voir la distribution des valeurs de force de sécurité pour me faire une idée de la distribution de ces notes. J'ai eu une belle surprise en regardant tout ça, en fait, dans les descriptions de la base de données, les notes devraient aller de 1 à 10 avec 10 étant un mot de passe très sécuritaire et 1 un mot de passe peu sécuritaire. Dans la table de données, les notes vont de 0 à 48. Comme personne n'était en mesure de m'expliquer ce qui s'est passé avec ces données-là, j'ai choisi de les considérer comme des extrêmes et de les retirer de la base de données pour travailler. Faut toujours faire attention quand on retire des données et le faire en pleine connaissance de cause, mais dans mon cas je pense que c'était la bonne chose à faire pour pouvoir poursuivre l'analyse. Donc, une fois le ménage fait, je me suis retrouvé avec une distribution normale de note allant de 1 à 10. J'ai donc plus séparé la base de données en 3 catégories: fort-moyen-faible en identifiant les 25% de données supérieures pour la catégorie fort et les 25% de données inférieur pour la catégorie faible. Il y aurait eu une multitude de façons de faire ça, mais cette semaine ça a été mon choix.

Une fois le ménage et l'organisation de données faite, je me suis retrouvé avec une petite base de données qui me donnait le résumé des données disponibles: le pourcentage de mot de passe dans chacune des catégories fort-moyen-faible que j'ai créé.

Alors, pour mettre tout ça en graphique, j'ai pris la peine d'aller relire mon objectif: montrer les catégories qui ont les mots de passe les plus fort. C'est important d'écrire notre objectif d'un parce que le fait de l'écrire nous permet souvent de clarifier les choses et d'autre part parce que ça nous permet d'y revenir en cours de route... Dans le processus de traitement et d'analyse des données, c'est très facile de se perdre en cours de route, on peut trouver un autre point intéressant dans les données et notre analyse peut changer de chemin. Ce n'est pas un problème en soi, je pense qu'on doit juste rester structuré dans notre démarche. Si on découvre quelque chose d'intéressant en cours de route, il faut prendre le temps de s'arrêter, de revenir à notre objectif principal et d'évaluer si ce qu'on a trouvé vaut la peine qu'on inscrive un autre objectif sur notre liste ou si ça vient carrément remplacer notre objectif initial. Dans tous les cas, si on change de chemin, c'est important de revoir notre processus de traitement des donnes pour s'assurer que notre démarche ne vient pas biaiser le nouvel objectif...

Dans mon cas, j'ai gardé le même objectif tout au long du processus. Alors, quand est venu le temps de mettre mes données en graphique, je savais exactement vers où aller. Mon objectif est clair: je veux voir quelles sont les catégories qui ont les mots de passe les plus forts. Je sais que je veux visualiser une distribution de données alors, mes choix de type de viz sont restreints à quelques possibilités.

J'ai choisi de présenter les résultats avec un lollipop un peu modifié. Comme la somme des résultats par catégorie donne 100, j'ai prolongé la tige du lollipop pour lui donné un aspect d'un indicateur de performance. J'aime beaucoup les lollipop en général, parce que ça change un peu du traditionnel histogramme tout en gardant tout ses avantages, mais en plus, on peut mettre de l'avant les valeurs des données directement sur le graphique... c'est presque de la magie.

PRÉSENTER LE GRAPHIQUE: Donc, j'ai utilisé des lignes et des points pour créer mon visuel cette semaine et je lui ai donné la forme d'un lollipop. J'ai mis le pourcentage de mots de passe de ma catégorie fort sur l'axe des x et les différentes catégories de classement des mots de passe sur l'axe des y. la valeur pour chacune des catégories est identifiée dans le cercle qui se trouve sur la ligne identifiée par la catégorie. J'ai aussi pris la peine de classer les données en fonction du pourcentage de mots de passe qui se sont classés forts dans mon analyse pour que le visuel soit ordonné et qu'on puisse constater en un seul coup d'oeil que c'est la catégorie pop qui contient un pourcentage de mots ce passe classé fort le plus grand. Pour expliquer un peu le contexte de tout ça, j'ai mis un titre accrocheur et une petite explication en sous-titre. Il faut garder en tête que c'est une base de données de mauvais mot de passe, alors notre analyse et notre visualisation sont relatives en fonction du contenu de cette base de données. Les meilleurs mots de passe qui existe ne sont pas

nécessairement des mots comme 'starwars' ou 'matrix', mais tant qu'à choisir un mauvais mot de passe, vaut mieux y aller pour mes mots comme ceux-là que d'y aller pour les mots de passe en format alphanumérique comme 123abc qui sont très facile à deviner.

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: Alors, pour vous démontrer que c'est important de rester fidèle à notre objectif si on veut créer des visuels clairs, j'ai repris les données de cette semaine et j'ai sauté l'étape du questionnement juste avant de créer le visuel.

En fait, si quelqu'un m'avait demandé de créer un visuel qui représentait le mieux les données que j'ai obtenues suite à mon traitement, j'aurais fait un graphique qui présenter les 3 catégories (fort, moyen, faible) de manière à montrer la distribution de ces 3 catégories en fonction des différentes catégories qui classent les mots de passe.

J'ai mis les deux graphiques dont je vous parle cette semaine dans les notes de l'épisode, prenez le temps d'aller voir ça je pense que ça vaut la peine de les mettre côte à côte pour faire la comparaison. Les deux sont complètement valides et sont des visualisations qui respectent le code de bonne conduite.

Mais, chacune d'entre elles répond à un objectif différent. La première que j'ai créée pour le tidytuesday montre seulement une petite partie des données disponibles. En fait, elle se concentre sur les mots de passe qui sont classés forts par catégorie. En ne montrant pas les autres valeurs du classement (moyen et faible) on met l'empase sur un seul message: les mots de passe pop sont les plus sécuritaires (relativement aux données qu'on avait entre les mains).

Si on conserve dans le visuel les 3 classements (fort, moyen, mauvais), la catégorie pop ne nous saute pas aux yeux. De la manière dont j'ai créé le graphique c'est le classement mauvais qu'on remarque en premier et c'est les mots de passe alphanumériques qui ressortent du lot. On voit aussi que les moyens ne sont pas très nombreux. Bref, on a une idée générale de la distribution du classement des mots de passe. Ce n'est pas mauvais en soi, c'est juste un objectif différent. Si c'est en accord avec le plan de travail c'est parfait, mais si on veut présenter nos résultats à des personnes qui n'ont jamais travaillé avec la base de données et que la conclusion qu'on veut qu'ils retiennent est que les mots de passe pop sont les meilleurs parmi les pires, vaut mieux ne pas présenter les 3 classements et se concentrer sur le classement fort pour éviter d'ajouter des distractions.

C'est ça que je veux dire quand je dis que tout dépend de l'objectif et c'est en établissant un objectif clair (qui peut changer dans le temps) qu'on peut identifier et visualiser l'important.

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:

- Transcription de l'épisode en pdf: ici
- Pour écouter l'épiside : ici
- Les graphiques discutés: ici et ici
- L'article de blogue en lien avec cet épisode: blogue
- Me contacter: contact