

Chapitre #1 – Collecte de données et évaluation de leur qualité

Le présent chapitre porte sur les thèmes suivants :

1. Identification des besoins d'information de votre organisation ou projet en fonction de vos objectifs d'affaires.
2. Présentation d'un inventaire de données et de son utilité pour identifier les données déjà disponibles pour l'analyse et celles qui doivent être collectées pour répondre à vos besoins d'information.
3. Évaluation de la maturité numérique de votre organisation ou projet et développement d'une bonne culture des données.
4. Développement des bonnes pratiques en termes de planification et de mise en œuvre d'une collecte de nouvelles données.
5. Évaluation de la qualité d'un jeu de données dans Excel et utilisation de formules et d'outils pour le nettoyage des données.
6. Amorçage d'un processus d'analyse et d'interprétation des données (sujet du 2^e chapitre de cette ressource).

Liste de contrôle pour les gestionnaires de données

Cette liste de contrôle destinée aux gestionnaires comprend une série de questions pour vous aider dans vos projets de collecte de données. Il s'agit en quelque sorte d'un aide-mémoire en lien avec le contenu de ce guide.

Avant la collecte de données

- **Avez-vous précisé la question à laquelle vous tentez de répondre et identifié les informations qui permettront d'y répondre ?**
Connaître en amont les questions et les problèmes qui vous interpellent le plus permet de mieux diriger le travail d'analyse.
- **Avez-vous procédé à un inventaire de vos données, c'est-à-dire recenser les données qui sont disponibles au sein de votre organisation ?**
À quels types de données avez-vous accès (données transactionnelles, données d'infolettre, etc.) ?
Est-ce que ces données sont facilement accessibles ?
Est-ce que vos données importantes sont disponibles à l'interne ou proviennent d'un partenaire ou d'un outil externe ?
- **Avez-vous bien évalué le niveau de maturité numérique de votre organisation, soit votre capacité à utiliser les données et les outils numériques pour prendre de meilleures décisions d'affaires ?**
Cette évaluation permet d'identifier les types d'analyses qui sont à votre portée et les analyses pour lesquelles il vous faudra chercher du soutien à l'externe ou encore suivre des formations.
- **Est-ce que vous avez bien planifié votre collecte de données ?**
Connaissez-vous les variables qui sont essentielles pour répondre à vos questions ?
Avez-vous déterminé des standards pour la saisie des données ?
Utilisez-vous des identifiants uniques qui permettent de faire le lien entre deux sources de données ?

Après la collecte

- **Avez-vous évalué la qualité de vos données ?**

Est-ce que vos données sont prêtes à être utilisées ou doivent être nettoyées (corrections des erreurs de saisies, identification des doublons, choix des bons formats, etc.) ?

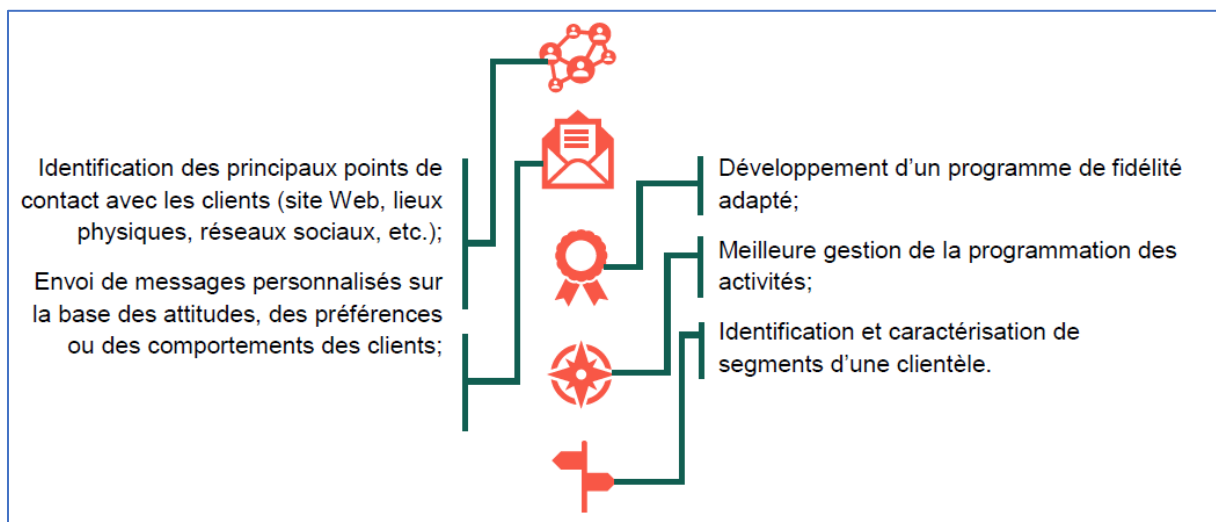
Est-ce que ces données sont pertinentes pour répondre à vos questionnements et enjeux d'affaires ?

Pourquoi les données sont-elles importantes ?

Face au développement du numérique, le secteur touristique subit d'importantes mutations qui bouleversent tous les paliers de la chaîne de valeur en place et multiplient les points de contact entre les organisations/entreprises (ci-après appelé organisations) et leurs clients. Par exemple, les consommateurs achètent maintenant leurs billets en ligne, s'informent davantage sur les réseaux sociaux ou prévoient leurs déplacements à l'aide d'applications mobiles. Ces nouveaux comportements obligent les événements et les attractions à étendre leurs actions à de nouveaux canaux de communication et à adapter leur stratégie marketing pour mieux rejoindre leurs clients. Pour mener à bien ces nouvelles actions, il devient primordial pour les événements et les attractions d'apprendre à collecter et à organiser les données que génèrent ces canaux.

Les données constituent **un actif à part entière** pour les organisations, car leur exploitation permet à la fois de **guider la prise de décisions d'affaires et de créer de la valeur au sein de l'organisation**. En d'autres mots, les données sont importantes puisqu'elles permettent d'améliorer la stratégie d'affaires et de marketing d'une organisation et de développer de meilleures tactiques.

À titre d'exemple, pour une organisation commerciale, les données permettent :



Définir sa stratégie de données

Pour profiter pleinement du potentiel des données, il faut d'abord concevoir une stratégie de données, c'est-à-dire définir « l'ensemble des buts, des objectifs et des actions stratégiques qui sont nécessaires au bon alignement de la vision de l'organisation avec ses priorités et ses actions en matière de données »¹. Celle-ci consiste à :

- Identifier les besoins d'information de son organisation ;
- Évaluer les données qui sont actuellement disponibles dans son organisation et déterminer les nouvelles données qui sont nécessaires pour répondre aux besoins d'information ;
- Évaluer la maturité numérique de son organisation ;
- Planifier et effectuer la collecte des nouvelles données ;
- Évaluer la qualité des données récoltées ;
- Analyser et interpréter les données (sujet du 2e chapitre de cette série).

¹ Fleckenstein, M. & Fellows, L.. (2018). *Modern data strategy*. Springer International Publishing.

Partie 1. Identifier ses besoins d'information

La première étape d'une stratégie de données consiste à identifier ses besoins d'information, ou en d'autres mots à déterminer ce que notre organisation doit savoir pour atteindre ses objectifs, pour relever ses défis d'affaires ou pour mener à bien ses actions stratégiques.

Attention : à elles seules, les données ne constituent pas une information. Sans traitement, sans analyse, sans contextualisation, les données n'ont pas ou peu de sens. En effet, elles sont des éléments simples – par exemple, un nombre de billets achetés ou la date d'un événement – tandis qu'une information est un rassemblement de données contextualisées qui offrent de nouvelles connaissances et guident la prise de décisions. Avoir un besoin d'information ne veut donc pas nécessairement dire que l'on manque de données, mais plutôt que celles-ci n'ont pas été suffisamment bien rassemblées et analysées.

À titre d'exemple, un besoin d'information pour un événement ou une attraction peut être de :

- Mesurer l'efficacité d'une campagne publicitaire sur Facebook ;
- Connaître les périodes de pointe d'une attraction ;
- Identifier ses meilleurs clients, sur le plan de la fréquence, de la récurrence et du montant de leurs achats ;
- Dresser un portrait général de l'achalandage d'un festival en vue d'une reddition de comptes ;
- Cartographier la provenance des clients au fil des années.

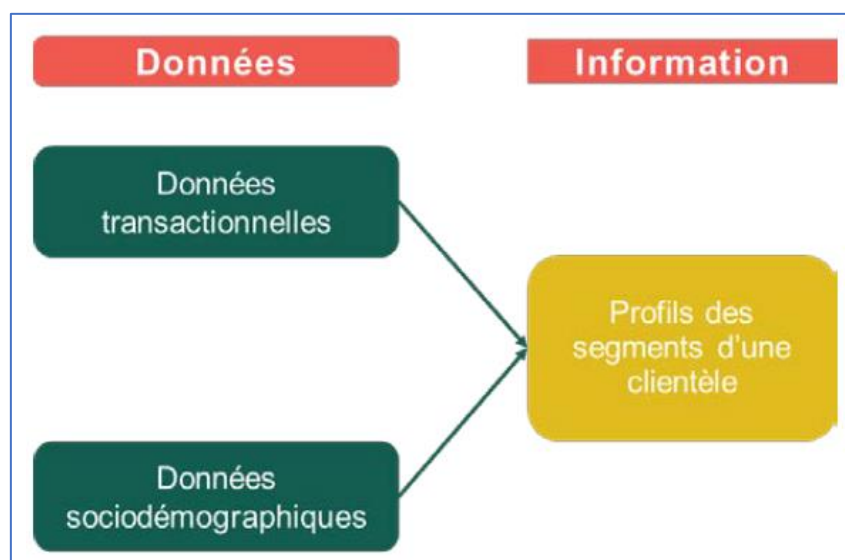


Figure 1. Passer des données à l'information.

Les besoins d'information peuvent aussi varier selon le public visé. Par exemple, le responsable marketing d'une attraction voudra connaître les habitudes médiatiques de ses consommateurs pour planifier une campagne publicitaire efficace, tandis que le conseil d'administration d'un festival cherchera davantage à suivre l'évolution de l'achalandage de l'événement dans un rapport simple et concis.

L'évaluation des besoins d'affaires permet aussi d'identifier le niveau d'analyse qui sera requis pour répondre à ceux-ci. Cette étape est utile parce qu'elle permet d'entrevoir le niveau de complexité des analyses nécessaires pour obtenir les informations qui nous intéressent.

Voici les quatre niveaux d'analyse possibles avec les données :

- **DÉCRIRE**, par la simple présentation des données en tableaux, graphiques ou rapports.

Par exemple, concevoir un tableau de bord qui permet de suivre la performance de la billetterie.

- **DIAGNOSTIQUER**, pour estimer des relations entre certaines variables.

Par exemple, évaluer le canal d'achat préféré (site Web, téléphone, application mobile) des visiteurs hors Québec.

- **PRÉDIRE**, pour prévoir ce qui est susceptible d'arriver

Par exemple, prédire le succès d'une campagne publicitaire numérique à partir des résultats de campagnes précédentes.

- **PRESCRIRE**, pour obtenir des recommandations ou optimiser des décisions.

Par exemple, développer un algorithme de recommandation qui permet à une application mobile de suggérer des attractions et des événements en fonction des préférences des consommateurs.

À vous de jouer !

Quels sont vos besoins d'information ? Les questions suivantes permettent de les identifier plus facilement :

1. Quels sont les principaux objectifs de votre organisation ou projet ? Quels en sont les défis ? Est-ce que certaines de vos décisions d'affaires relèvent davantage de l'intuition ?
 2. Qu'est-ce que votre organisation/projet connaît déjà sur ses enjeux et qu'est-ce que vous aimeriez savoir de plus ?
 3. Qui sera le public visé par ces nouvelles informations ? Votre comité de direction, vos bailleurs de fonds, votre équipe marketing, etc. ?
 4. À quelles fins seront utilisées les nouvelles informations ?
-

Partie 2. Faire l'inventaire de ses données et déterminer les nouvelles données à collecter

Une fois les besoins d'information identifiés, la deuxième étape consiste d'abord à faire l'inventaire de ses propres données, c'est-à-dire à recenser celles qui sont disponibles au sein son organisation. Cet inventaire est utile puisqu'il permet, dans un deuxième temps, de déterminer les nouvelles données à collecter et celles qui sont déjà prêtes à être analysées.

À titre d'exemple, un événement qui possède une base de données de billetterie comprenant la fréquence, la récence et le montant des achats des participants pourra effectuer une segmentation RFM² pour identifier ses meilleurs clients sans collecter de nouvelles données. Il sera toutefois nécessaire de collecter de nouvelles données, via un sondage par exemple, pour connaître le profil sociodémographique ou les motivations d'achat de ses clients.

Le résultat de cet inventaire de données devrait être accessible à tous les membres de votre personnel. En effet, une liste détaillée de vos différentes sources et bases de données ainsi que de leurs usages permet d'identifier plus facilement le potentiel de vos données et de faciliter le transfert d'information vers un partenaire, un consultant ou de nouveaux employés. Elle peut également servir d'outil de référence.

Se démêler à l'aide des catégories de données

Pour simplifier le processus d'inventaire des données, il peut être pertinent de se référer à une catégorisation pour débiter et encadrer l'exercice. Dans le monde du marketing, on distingue généralement deux grandes familles de données : primaires et secondaires (voir Figure 2).

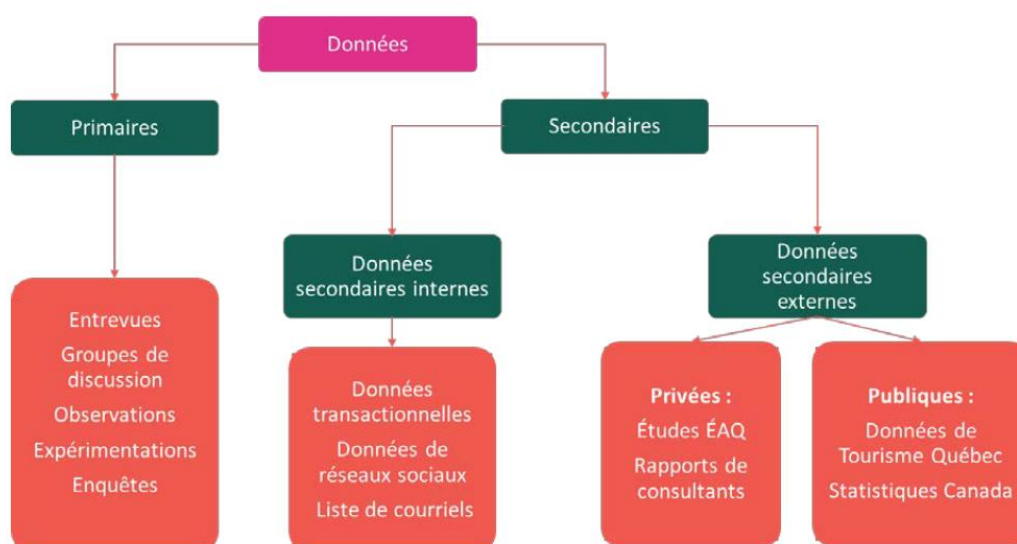


Figure 2. Un exemple de catégorisation des données employée en marketing.

² La méthode RFM permet de regrouper des consommateurs en fonction de leurs habitudes d'achat. Elle consiste à attribuer un score aux consommateurs en fonction de la récence (R), de la fréquence (F) et du montant (M) de leurs achats. La segmentation RFM permet notamment d'identifier ses meilleurs clients ou encore de faire de la prospection.

Les données primaires

Par données primaires, on entend toutes les données qui ont été collectées auprès des consommateurs via des enquêtes, des sondages ou encore des groupes de discussion. Elles sont collectées pour étudier un phénomène en particulier. Les données issues des études de provenance et d'achalandage sont de bons exemples.

Les données secondaires internes

o Les données de ventes, de billetterie, de réseaux sociaux et d'infolettres constituent toutes des données secondaires internes. On parle de données secondaires internes puisqu'elles sont issues de l'organisation (« internes »), mais elles ne sont pas collectées initialement pour étudier un phénomène en particulier (« secondaires »). Par exemple, les données de ventes sont d'abord collectées pour assurer le bon fonctionnement d'une transaction, mais peuvent ensuite être utilisées pour calculer les revenus associés à un spectacle ou pour identifier les meilleurs clients d'une attraction touristique.

Les données secondaires externes

Les données secondaires externes comprennent les données publiées par les organisations publiques ou par des organisations privées. Il peut s'agir des données d'un rapport issu d'une association sectorielle ou encore d'un jeu de données ouvertes publié par data.gouv.fr. Ces données permettent de mesurer l'évolution de la demande sur un marché donné ou encore de comparer ses données à celles de compétiteurs.

Remarque : Pour les données secondaires internes, il est utile de distinguer les données facilement accessibles de celles pour lesquelles il faut passer par un partenaire ou une application automatisée.

Prenons l'exemple d'une billetterie : pour les organisations qui ont leur propre service de billetterie, il est possible de modifier la collecte des données à la source selon les besoins analytiques. À l'opposé, en faisant affaire avec un service externe, les organisations perdent cette latitude et doivent utiliser les données telles qu'elles se présentent.

Faire ces distinctions permet de mieux identifier les limites de vos données.

Il est aussi possible de classer les données selon les renseignements qu'elles fournissent sur les consommateurs :

- o **Informations personnelles des clients** :
Nom, âge, courriel, numéro de téléphone, etc.
- o **Interactions des clients avec votre organisation** :
Données d'infolettre (ex. : mailchimp), interactions sur les réseaux sociaux, appels téléphoniques, etc.
- o **Données transactionnelles, voire les comportements d'achat des clients** :
Nombre de transactions, canal d'achat utilisé, montant dépensé, etc.
- o **Attitudes des clients** :
Évaluations sur Google, sondage d'appréciation, etc.

L'important est de choisir la classification qui s'applique le mieux à votre contexte, soit celle qui correspond à vos besoins d'information, mais aussi aux habitudes de votre personnel et au fonctionnement de votre organisation.

Partie 3. Évaluer la maturité numérique de son organisation

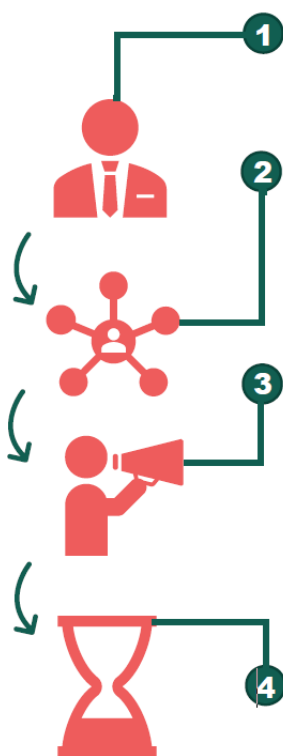
La troisième étape d'une stratégie de données consiste à évaluer le niveau de maturité numérique de votre organisation, soit sa capacité à :

- Mobiliser des outils liés aux données ;
- Utiliser les données dans le cadre de prises de décisions stratégiques.

Évaluer son niveau de maturité numérique revient à se demander si votre organisation possède une culture des données, c'est-à-dire si votre équipe « a pris l'habitude de vérifier ses hypothèses de travail et ses intuitions professionnelles en s'appuyant sur l'analyse d'une quantité suffisante de données pertinentes »³.

Évaluer son niveau de maturité numérique est utile puisque cela permet d'identifier les types d'analyses qui sont à la portée de votre organisation et ceux pour lesquels il vous faudra chercher du support à l'externe ou encore suivre des formations. Par exemple, il serait inutile pour un festival qui peine à concevoir un tableau de bord de se lancer dans des travaux avancés en intelligence artificielle.

Comment développer une culture des données ?



1 Tout part de la direction

Si la direction de votre organisation prend ses décisions de façon intuitive, le reste de l'équipe fera de même. Il faut donc que la direction valorise le recours aux données et qu'elle les utilise.

2 Sensibiliser le personnel à l'importance des données

Il faut s'assurer que toutes les personnes qui collaborent de près ou de loin à l'effort de collecte de données comprennent dès le début les avantages d'une collecte de qualité - principalement la manière dont les données contribuent à renforcer l'organisation et ses prises de décisions - et qu'elles adhèrent à la stratégie de collecte.

3 Communiquer à propos de l'usage des données

Partager les résultats des collectes et des analyses de données avec ceux qui y participent :

- a. leur donner accès aux rapports et aux résultats ;
- b. être ouvert sur la façon dont les données sont utilisées ;
- c. en parler lors des réunions du personnel ;
- d. faire circuler les rapports, etc.

Cette communication aidera à montrer les avantages de l'analyse des données et à motiver le personnel de votre organisation.

4 Savoir faire preuve de patience

Développer une bonne culture des données demande des investissements importants en termes de temps et de formation. Il ne sert à rien de se lancer trop vite dans l'acquisition de logiciels onéreux si personne ne sait comment bien les utiliser. Il est donc plus pertinent de se lancer dans l'analyse des données une étape à la fois, en commençant par la maîtrise des outils de visualisation ou la conception de tableaux de bord qui serviront concrètement à vos opérations. Ensuite, lorsque votre organisation aura eu le temps de développer cette culture des données, il sera possible d'envisager de vous lancer dans des analyses plus complexes.

³ Legoux, R., & Lord, S. (2019). [La culture des données, un élément qui change la donne](#). Gestion, 44(1), 96-99.

Partie 4. Planifier la collecte de nouvelles de données

Lorsque les besoins d'information et les données manquantes ont été identifiés, il est temps de passer à la planification de la collecte des nouvelles données. Cette planification permet d'éviter en amont beaucoup de problèmes liés au nettoyage⁴ et au traitement des données. Voici quelques conseils en lien avec cette planification.

Conseil #1 : Les champs de données prioritaires

Idéalement, toute collecte de données au sujet des clients devrait générer des données pour les champs suivants :

1. L'identifiant client unique qui permet de suivre un client au fil de ses différentes transactions ;
2. Le code postal du client, une façon simple d'obtenir plusieurs informations sur les clients (voir encadré) ;
3. La date (et l'heure) de chaque transaction ;
4. Le canal de la transaction : guichet, site Web, application mobile, téléphone, etc.
5. Le type d'achat : billet unique, abonnement de saison, laissez-passer, forfait familial, etc.
6. Le montant de la transaction ;
7. La quantité de biens ou services transigés : nombre de billets ou d'objets souvenirs achetés, etc.
8. L'identifiant unique du produit ou du service transigé/ un code, une étiquette ou une chaîne de caractères qui identifie un spectacle ou une excursion.

Le potentiel du code postal

Dans les cas où il est difficile d'effectuer des collectes de données intensives, le code postal peut s'avérer un outil très utile pour obtenir un aperçu de la composition des visiteurs d'un événement ou d'une attraction.

Le code postal, permet de déterminer de façon précise la provenance d'un client. Par exemple, grâce à un fichier fourni par data.gouv.fr, il est possible d'associer rapidement chaque code postal à sa commune ou son département.

Cette information est très utile parce qu'elle permet ensuite d'identifier des géographies précises pour de futures campagnes sur Google Ads ou sur Facebook, par exemple. Avec les deux premiers caractères du code postal, nous pouvons également explorer les données du recensement de l'[Insee](https://www.insee.fr) pour tracer des portraits socioéconomiques de ses visiteurs. Par exemple, sont-ils plus susceptibles d'avoir des enfants ? d'être des diplômés universitaires ? etc.

La règle d'or est de ne pas collecter trop de données inutilement, mais plutôt de se concentrer sur ce qui est nécessaire pour répondre aux questions qui ont préalablement été identifiées comme importantes ou prioritaires.

⁴ Le nettoyage consiste à corriger les erreurs (valeurs en doubles, fautes de frappe, données manquantes, etc.) qui sont présentes dans un jeu de données que l'on souhaite analyser.

Conseil #2 : Standardiser la collecte de données

Lors d'un processus de collecte de données, comme une transaction au guichet par exemple, l'absence de procédures claires pour la saisie, manuelle ou automatique, des données peut causer des problèmes au moment de l'analyse. En effet, une donnée n'est utile que si elle est collectée de la même façon pour toutes les transactions. Si, pour certaines transactions, l'information recueillie est très détaillée et pour d'autres, très superficielle, les analyses seront limitées au plus simple niveau de collecte.

Par exemple : si le canal d'achat est, pour certains clients, « par internet », « par téléphone » ou « au guichet », mais que pour d'autres clients les possibilités sont « par internet » ou « autre », le seul niveau de détail utilisable en jumelant toute l'information sera : « par internet » ou « autre qu'internet ».

Il est donc primordial de prendre le temps et les moyens de se doter d'une procédure pour remplir les champs selon des classifications que vous aurez préalablement définies et qui correspondent aux standards de votre secteur ainsi qu'aux besoins d'information spécifiques de votre organisation.

Concrètement, cela peut vouloir dire d'indiquer notamment :

- Les valeurs possibles pour chacun des champs de collecte ;
- L'utilisation ou non des accents et d'autres caractères spéciaux ;
- Le format de saisie requis pour chacun des champs (par exemple, forcer la saisie de six caractères avec des majuscules pour les codes postaux) ;
- La façon de traiter les valeurs manquantes (voir encadré).

Quoi faire avec les valeurs manquantes ?

Lorsque les données sont disponibles, il importe de se donner les moyens de compléter TOUS les champs du formulaire de saisie : plus les données sont complètes, plus les analyses auront de la valeur.

En effet, les observations avec des cases « vides » ne peuvent être analysées normalement. Ces observations sont donc généralement retranchées des analyses, ce qui revient à gaspiller du précieux temps de collecte et à perdre une partie de l'information.

Lorsque certaines données ne sont pas disponibles, il faut l'indiquer directement dans le champ (en inscrivant « non applicable », « non disponible », « autre » ou « non valide », par exemple) plutôt que de laisser des cases vides. Si rien n'est indiqué, il est alors difficile de savoir si ces cases sont vides parce que l'information est réellement manquante ou parce que c'est la collecte qui a été déficiente.

Conseil #3 : Un concept = un champ

Un problème courant des collectes de données est la présence de plusieurs dimensions ou de plusieurs concepts distincts dans un même champ. En effet, s'il est possible de combiner deux données collectées dans un seul champ, il est très difficile sinon impossible de diviser par la suite les données qui ont été regroupées, ce qui réduit le niveau de détail de l'analyse. Par exemple, une donnée sur chaque province canadienne permet de générer des informations pour l'ensemble du Canada, mais une donnée sur l'ensemble du Canada ne permet pas de retrouver l'information pour chaque province.

Prenons l'exemple d'un festival qui utilise trois catégories de prix (A, B ou C). Il semble plus simple de regrouper toutes ses options dans le même champ et d'indiquer à quelle catégorie de prix est associée chacune des transactions. Or, un problème peut survenir si un même client achète des billets de catégories de prix différentes lors d'une même transaction. En effet, en n'utilisant qu'une seule variable pour identifier la catégorie de billets, il est difficile de bien saisir cette l'information (est-ce qu'on doit indiquer le nombre de billets ? seulement les catégories de prix ? etc.). De plus, du temps sera perdu (en nettoyage et traitement de données) au moment de l'analyse pour isoler les billets vendus pour une seule catégorie de prix par exemple.

Une solution très simple consiste à créer un champ pour chacune des catégories de prix et à les additionner, selon vos besoins. La valeur de chaque champ devient donc plus précise et offre plus de possibilités d'analyse.

Une erreur fréquente...			→	Une solution simple !		
Transaction_id	Nombre total de billets achetés	Catégorie de billets		Nombre de billets de catégorie A	Nombre de billets de catégorie B	Nombre total de billets achetés
123456	3	2A 1B		2	1	3
123457	5	3A 2B		3	2	5

Conseil #4 : Utiliser des identifiants uniques

Un identifiant unique est une chaîne de caractères numérique (ex. : 0123) ou alphanumérique (ex. : ABCD0123) associée à une seule entité dans une base de données. Dans le contexte des événements et des attractions, l'identifiant unique sert principalement à identifier un client et à le suivre dans le temps et à travers différents points de contact. L'identifiant unique est donc une information indispensable pour être capable de comprendre le parcours des clients. À titre d'exemple, sans cette information, il n'est pas possible de regrouper ensemble les transactions d'un même client et de faire des analyses marketing de base comme la segmentation RFM (récence, fréquence et montant dépensé). De plus, l'identifiant unique permet de transmettre plus facilement à de tierces parties des données sur vos clients (pour des besoins d'analyse ou de collaboration) tout en préservant l'anonymat de vos clients.

Dans certaines organisations, l'identifiant sera basé sur le courriel ou le numéro de téléphone du client. À court terme, cette façon de faire peut s'avérer satisfaisante, car ces informations sont liées à un seul individu ou un seul ménage. À moyen ou à long terme, malheureusement, un numéro de téléphone peut changer de main ou un courriel peut devenir inactif. Il est donc préférable de générer un identifiant unique pour chacun de vos clients. Pour les mêmes raisons, il pourra être utile de créer des identifiants uniques pour des fournisseurs, des produits, des spectacles, etc.

Cet identifiant unique, aussi appelé clé, permet de suivre le comportement d'achat du client, mais aussi de faire le lien entre différents jeux de données⁵. Dans l'exemple de la Figure 3, l'utilisation d'identifiants uniques permet de lier les informations sur une transaction (ex. : quantité de billets achetés, mode de paiement) avec les données sur le client qui l'a effectué (ex. : nom et courriel) ainsi que sur l'événement concerné (ex. : type de spectacle), même si ces données sont sur trois feuilles différentes.

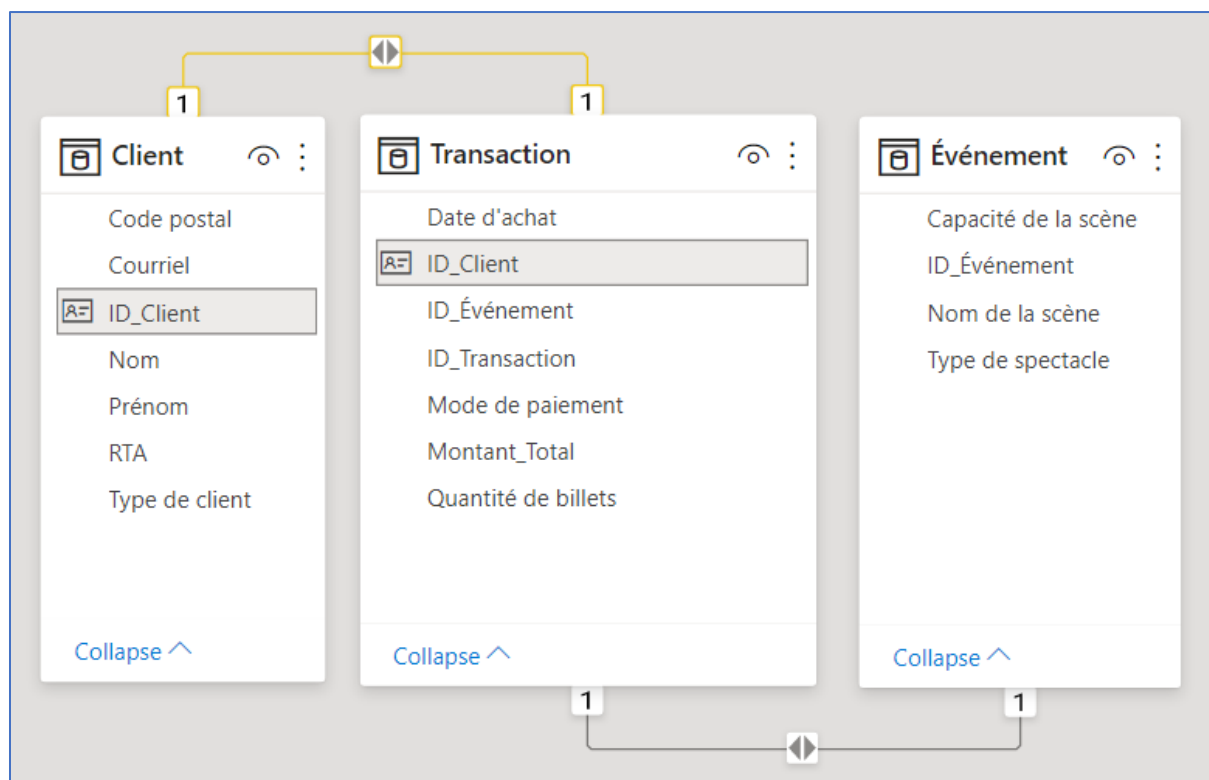


Figure 3. Exemple de création de liens entre des jeux de données à l'aide d'identifiants uniques.

⁵ Il est question, ici, de la notion de relation entre les bases de données, ou de base de données relationnelles.

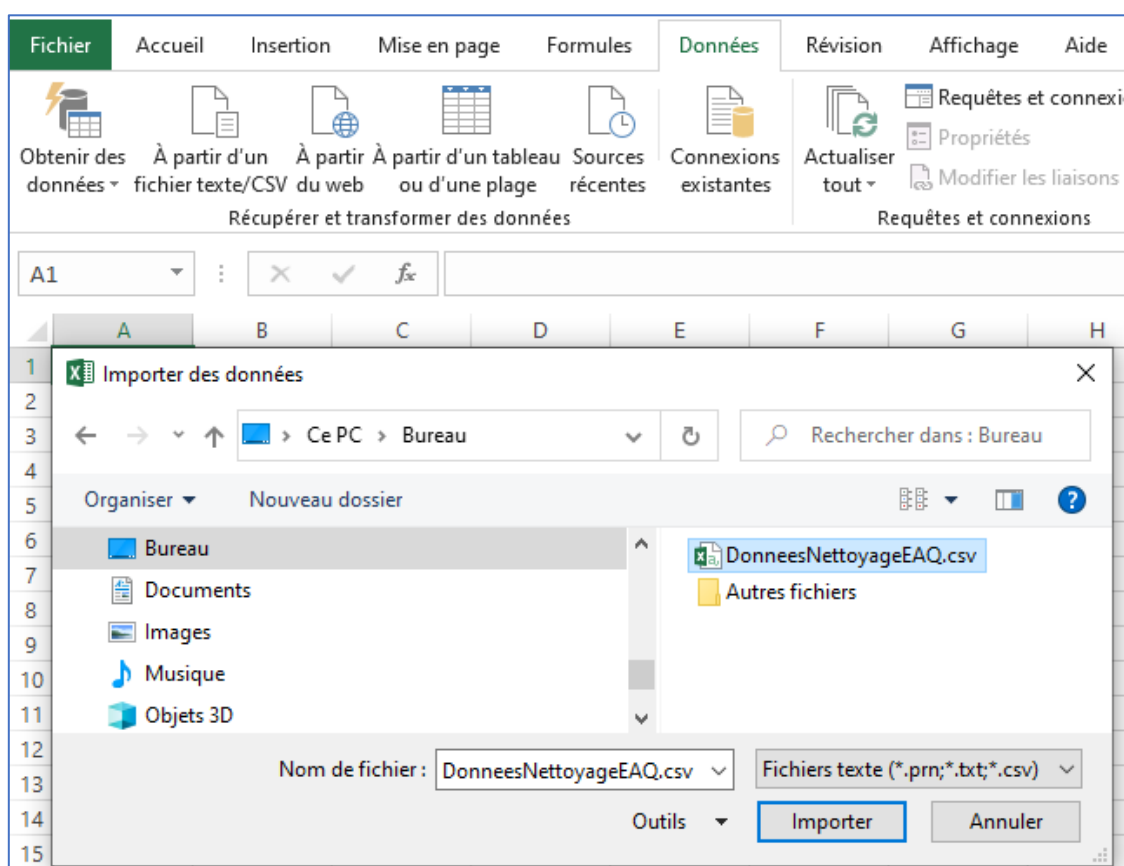
Partie 5. L'évaluation de la qualité des données

Une fois les données récoltées et assemblées, il est important d'en évaluer la qualité. En effet, ce n'est qu'avec un jeu de données « propres » qu'il est possible d'obtenir des analyses justes qui représentent la réalité de l'organisation. Autrement, en utilisant un jeu de données comprenant des erreurs, des données manquantes, des valeurs en doubles ou autres, nous risquons de baser des décisions d'affaires sur des informations inexactes. Un jeu de données propres est donc un jeu de données qui a été préalablement nettoyé (corrections des erreurs de saisies, identification des valeurs en double, choix des bons formats, etc.) et qui est prêt à l'analyse.

Afin de vous aider à comprendre en quoi consistent le nettoyage et le traitement des données, les prochains paragraphes proposent une série d'exercices et de démonstrations pratiques que vous pouvez reproduire de votre côté. Pour ces exercices, nous utiliserons plus précisément un fichier de données fictives intitulé « **DonneesNettoyage.csv** ». Il s'agit d'un court jeu de données de billetterie comprenant une dizaine de transactions et une quarantaine de champs qui contiennent possiblement de données.

Convertir des données d'un fichier CSV vers Excel

Puisque le fichier « **DonneesNettoyage.csv** » est en format CSV⁶, il faut d'abord l'importer dans Excel en cliquant sur l'option « À partir d'un fichier texte/CSV » de l'onglet « Données ».



⁶ Un fichier CSV est un fichier texte dont les valeurs sont séparées à l'aide de virgules ou de points-virgules. Ce format permet de stocker et de partager facilement des données tabulaires, indépendamment des logiciels utilisés.

Excel ouvre alors une fenêtre qui nous permet de sélectionner le codage des caractères des données importées (« Origine du fichier ») ainsi que le délimiteur utilisé dans le fichier CSV.

Dans le cas qui nous intéresse, les données utilisent la série de caractères Unicode UTF-8 (voir encadré), ainsi que le point-virgule comme délimiteur. Excel nous offre en bas de fenêtre l'option de « charger » les données sur notre feuille ou encore de « transformer les données » avec l'éditeur Power Query⁷. Le résultat obtenu en chargeant directement les données se trouve sur la feuille « Données originales » du fichier Excel « Exemple_nettoyage ».

No. de réservation	Nom de famille	Prénom	Genre	Langue	Courriel 1	Infolettre email 1	Courriel 2	Infolettre email 2
Billetterie--00005825	V	M			vm@gmail.com			
Billetterie--00005871	P	A			pa@outlook.com			
Billetterie--00005873	P	A			pa@outlook.com			
Billetterie--00005945	R	B			rb@gmail.com			
Billetterie--00007561	P	R			pr@gmail.com			
Billetterie--00007951	L	A			LA@hotmail.com			
Billetterie--00007953	L	A			LA@hotmail.com			
Billetterie--00008071	L	PA			LPA@gmail.com			
Billetterie--00007561	P	R			pr@gmail.com			

Charger Transformer les données Annuler

Un problème avec les caractères ?

Lors de l'importation d'un fichier CSV vers Excel, il peut arriver que le logiciel ait de la difficulté à reconnaître certains caractères de la langue française. Les « é » sont alors convertis en « Ã© », les « â » deviennent des « Ã¢ », etc.

Avant de recourir à des fonctions « substitue » ou encore faire un nombre important de « chercher et remplacer », il est important de vérifier si vous avez choisi le bon encodage des caractères lors de l'importation des données. Cette étape vous sauvera bien du temps. L'encodage le plus utilisé est l'Unicode UTF-8.

⁷ Power Query est un éditeur de requêtes qui s'utilise dans Excel et dans Power Bi et qui permet de nettoyer et de préparer des données avant de les analyser. Avec Power Query, nous pouvons facilement ajouter ou enlever des colonnes, remplacer des valeurs, fractionner des colonnes, changer le format des données, etc. L'utilisation de Power Query dépasse de la portée de ce guide, mais nous vous invitons à continuer vos lectures [pour en apprendre plus](#).

Choisir les variables pertinentes

Il arrive fréquemment que certaines des données pertinentes pour répondre à ses besoins d'affaires proviennent de sources externes. C'est notamment le cas lorsque nous faisons affaire avec un service externe de billetterie, ou encore lorsque nous souhaitons analyser des données issues de réseaux sociaux. Un des enjeux avec ce type de sources est que nous ne contrôlons pas la nature ni l'étendue de la collecte, ce qui peut faire en sorte que le jeu de données initiales contienne plusieurs informations non pertinentes pour nos analyses. C'est le cas des « Données originales » du fichier Excel « Exemple_nettoyage » qui contient plus de 42 colonnes (donc 42 variables), mais dont seulement un petit nombre est pertinent pour l'analyse.

Pour choisir les colonnes à conserver, voici quelques conseils :



Choisir en fonction du besoin d'affaires : si, par exemple, une organisation souhaite améliorer le service d'assistance de sa plateforme de billetterie, il serait pertinent de conserver l'heure exacte de chaque transaction pour connaître les périodes de pointe afin de prévoir un plus grand nombre d'employés au service à la clientèle à ces moments. Cette donnée n'est toutefois pas pertinente si l'objectif est plutôt de connaître l'évolution du nombre de touristes qui assistent à son événement ou visitent son attraction au fil des ans.



Créer un dictionnaire de données : un dictionnaire de données est une documentation qui décrit ce que contient un jeu de données. Ce dictionnaire contient généralement le nom des champs (ou des variables), la description de ceux-ci et le format des données (nombre entier, chaîne de caractères, etc.). Le dictionnaire de données est utile pour la sélection des variables puisqu'il permet d'avoir une vue d'ensemble et de comprendre le contenu de chacune des variables comprises dans le jeu de données. La feuille « dictionnaire de données » du fichier Excel « [Exemple_nettoyage](#) » en montre un exemple.



Identifier les colonnes inutilisables : dans les « [Données originales](#) » du fichier Excel, plusieurs colonnes sont soit fortement incomplètes, soit complètement vides. Dans les deux cas, ce sont des colonnes qui n'ont aucune utilité et qui peuvent tout de suite être retirées du jeu de données. Certaines colonnes peuvent aussi regrouper la même information : c'est le cas ici des colonnes « total » et « total dû/des paiement(s) ». Nous pouvons donc retirer l'une de ces colonnes du jeu de données.

À vous de jouer !

Avec le jeu de données « Données originales », nous souhaitons observer les différences entre les achats faits en ligne et ceux faits au guichet, selon la ville du client, le montant total de la transaction ou le nombre de billets achetés. Quelles sont les variables que vous conserveriez pour répondre à ce besoin d'affaire ?

Grâce à l'identification de ce besoin d'affaires, l'étude du dictionnaire de données et l'examen des données, nous pouvons identifier les variables qui sont pertinentes pour l'analyse (voir feuille « Dictionnaire de données » du fichier Excel pour voir le choix des variables). Lors de cette étape, il peut être pertinent de renommer les variables qui ont des noms imprécis ou compliqués.


Uniformiser les variables retenues et corriger les types de données

Dans le jeu de données de la feuille « Variables retenues » du fichier Excel, nous constatons plusieurs problèmes sur le plan de l'uniformisation des champs :

- Certaines villes sont écrites majuscules, d'autres non ;
- Les codes postaux sont écrits sous plusieurs formats différents ;
- Les colonnes contenant des dates ou des heures sont traitées par Excel comme étant des nombres et non pas des dates ou des heures.

Il faut donc faire quelques corrections à l'aide de formules Excel pour uniformiser les champs et choisir les bons types de données. Pour suivre les prochains exemples, rendez-vous sur la feuille « Majuscule_Type » du fichier Excel et suivez les étapes suivantes proposées :

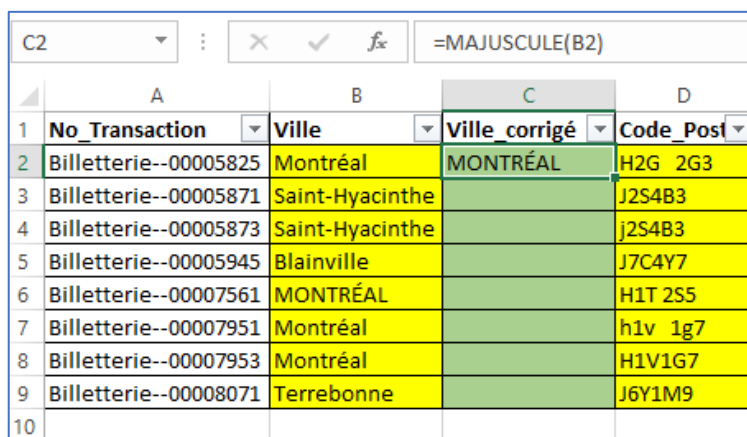
Étape 1

Par précaution, il est préférable de créer de nouvelles colonnes pour chacune des variables à nettoyer. Nous pourrions ainsi y insérer les différentes formules pertinentes dans la barre de formule  et faire des corrections au besoin sans effacer les données originales.

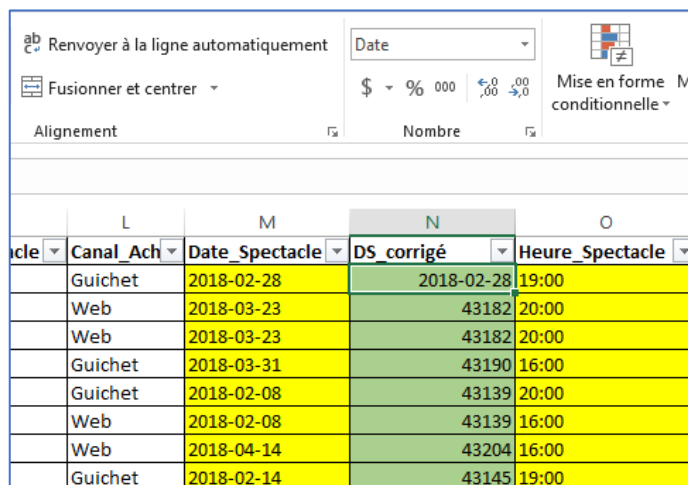
Étape 2

Uniformiser le format des villes et des codes postaux. Les fonctions MAJUSCULE et SUBSTITUE sont utiles (voir Tableau 1) pour uniformiser les champs sans avoir à corriger manuellement toutes les cellules.

Il suffit d'écrire la formule dans la première case de la nouvelle colonne et de la recopier dans les cellules sous-jacentes, à l'aide d'un double-clic dans le coin du bas à droite de la première cellule. La colonne se complète alors automatiquement.



	A	B	C	D
	No_Transaction	Ville	Ville corrigé	Code_Post
2	Billetterie--00005825	Montréal	MONTRÉAL	H2G 2G3
3	Billetterie--00005871	Saint-Hyacinthe		J2S4B3
4	Billetterie--00005873	Saint-Hyacinthe		j2S4B3
5	Billetterie--00005945	Blainville		J7C4Y7
6	Billetterie--00007561	MONTRÉAL		H1T 2S5
7	Billetterie--00007951	Montréal		h1v 1g7
8	Billetterie--00007953	Montréal		H1V1G7
9	Billetterie--00008071	Terrebonne		J6Y1M9



	L	M	N	O
	Canal_Ach	Date_Spectacle	DS corrigé	Heure_Spectacle
	Guichet	2018-02-28	2018-02-28	19:00
	Web	2018-03-23	43182	20:00
	Web	2018-03-23	43182	20:00
	Guichet	2018-03-31	43190	16:00
	Guichet	2018-02-08	43139	20:00
	Web	2018-02-08	43139	16:00
	Web	2018-04-14	43204	16:00
	Guichet	2018-02-14	43145	19:00

Étape 3

Uniformiser le format des colonnes avec des dates et des heures. Les formules DATEVAL et TEMPSVAL permettent de transformer une chaîne de caractères (par exemple la date « 2018-02-28 ») en une série de chiffres qu'Excel associe à cette date (ici 43159). Grâce à cette série de chiffres, nous pourrions calculer le nombre de jours entre la date d'une transaction et la date d'un spectacle, par exemple.

Étape 4

Changer cette série de chiffres en Date avec le menu déroulant qui se trouve dans l'onglet « Nombre ». Ceci facilitera la lecture.

Tableau 1. Les formules Excel utilisées dans cet exercice pour uniformiser les données et corriger les formats des variables.

Enjeux	Formule(s) Excel	Explications
Uniformiser les majuscules et minuscules	=MAJUSCULE(B2) ou =MINUSCULE(B2)	Les fonctions MAJUSCULE et MINUSCULE transforment l'entièreté des caractères d'une cellule (ici B2) en lettres majuscules ou minuscules
Enlever les espaces dans les codes postaux	=SUBSTITUE(MAJUSCULE(D2);" ","")	La fonction SUBSTITUE remplace ou modifie certaines valeurs textuelles par d'autres à l'intérieur d'une chaîne de caractères. Dans ce cas précis, nous demandons à la formule de prendre la cellule D2 et de prendre tous les espaces (" ") et de les enlever (""). L'utilisation de la formule MAJUSCULE en même temps permet de corriger deux problèmes en un clic.
Uniformiser le format des colonnes avec des dates et des heures	=DATEVAL(M2) ou =TEMPSVAL(O2)	

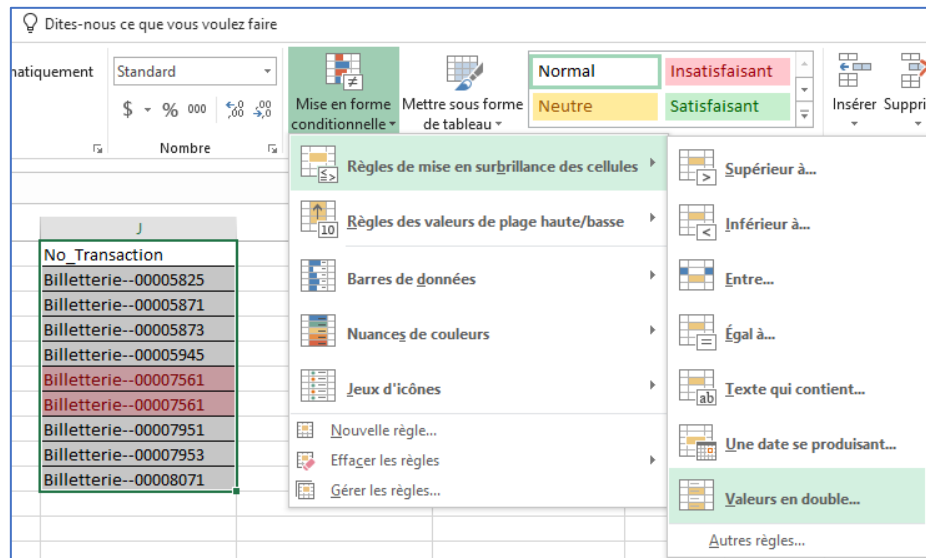
Identifier et enlever les doublons

La présence de doublons, c'est-à-dire des observations apparaissant plus d'une fois dans un jeu de données, est problématique pour l'analyse et la prise de décision qui s'en suit. En effet, les doublons peuvent introduire des biais dans les analyses. À titre d'exemple, la présence d'une série de transactions en double dans un jeu de données peut amener à surestimer le succès d'une activité.

Plusieurs techniques existent pour identifier les doublons à l'aide d'Excel. Pour poursuivre avec les prochaines étapes, rendez-vous sur la feuille « Verifier_doublons » du fichier Excel.

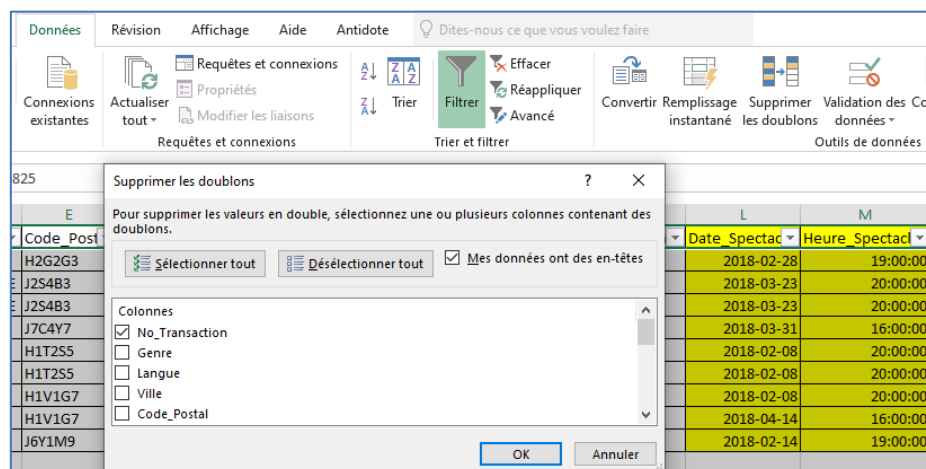
Étape 5

Utiliser la fonction « mise en forme conditionnelle » sur la colonne désirée. Cette fonction permet de mettre de l'avant certaines valeurs selon des critères précis comme celui des « valeurs en double ». Dans le cas de cet exercice, nous l'utiliserons pour identifier deux transactions avec le même identifiant unique. Sélectionnez la colonne No_Transaction et mettre en place la mise en forme conditionnelle. Les cellules en double seront colorées. Cette technique ne permet que de voir les doublons.



Étape 6

Utiliser la fonction « supprimer les doublons » de l'onglet « Données ». Pour avoir recours à cette fonction, il faut sélectionner l'ensemble de notre jeu de données⁸ et ensuite choisir la variable pour laquelle nous souhaitons enlever les doublons. Ici, nous demandons à Excel de supprimer les valeurs en double seulement lorsque l'identifiant unique de la transaction (« No_Transaction ») se répète. En effet, nous ne voulons pas qu'Excel supprime des observations qui partagerait les mêmes valeurs pour la colonne « Ville », par exemple, puisqu'il est normal que des clients habitent la même ville.



⁸ Lorsque l'on sélectionne qu'une seule colonne, Excel nous proposera d'étendre la sélection à l'ensemble des colonnes.

Si les deux premières techniques sont généralement efficaces, elles ne permettent pas, par exemple, d'identifier des observations qui partageraient les mêmes données, mais un identifiant unique différent. En d'autres mots, ces techniques ne permettent pas d'enlever des transactions qui se retrouvent en double dans notre jeu de données, mais pour lesquels nous avons par erreur donné deux identifiants différents.

Étape 7

Utiliser une troisième technique qui consiste à créer une nouvelle variable qui combine les valeurs des autres colonnes, sauf celles de l'identifiant de la transaction. Pour ce faire, il faut créer une formule qui combine les colonnes (par exemple : =B2&C2&D2&...&I2).

En faisant l'exercice, vous constaterez qu'il existe un autre doublon dans le jeu de données qui n'avait pas été identifié précédemment parce que les identifiants uniques des transactions étaient, par erreur, différents.

S
Nouvelle_Variable_Test_Doublons
FFRMONTREALH2G2G3QuébecCanadaIndividuABCD
FENSAINT-HYACINTHEJ2S4B3QuébecCanadaIndividu
FENSAINT-HYACINTHEJ2S4B3QuébecCanadaIndividu
MFRBLAINVILLEJ7C4Y7QuébecCanadaIndividuABCD
MFRMONTREALH1T2S5QuébecCanadaIndividuABCD
MFRMONTREALH1T2S5QuébecCanadaIndividuABCD
FFRMONTREALH1V1G7QuébecCanadaIndividuABCD
FFRMONTREALH1V1G7QuébecCanadaIndividuABCD
MFRTERREBONNEJ6Y1M9QuébecCanadaIndividuAB

Pour supprimer ces doublons, retournez à l'étape 6, mais en choisissant la variable « Nouvelle_Variable_Test_Doublons ».

Partie 6. L'analyse des données

Voilà ! Les données sont propres et prêtes à être analysées. Statistiques descriptives, tableaux croisés, visualisation, tableau de bord : plusieurs possibilités s'offrent maintenant à vous. Puisque l'analyse et l'interprétation des données débordent de la portée de cette partie, nous consacrons les deux autres parties pratiques à leur exploitation.

Pour vous donner un avant-goût, un simple tableau croisé (disponible dans la feuille « tableau croisé dynamique » du fichier Excel) vous permet de constater que la plupart des billets ont été achetés au guichet (13 billets au total) et que les transactions faites sur le Web sont propices à l'achat de billets individuels (trois transactions pour l'achat d'un billet individuel).

Tableau 2. Un exemple de tableau croisé issu de nos données nettoyées.

Canal d'achat	Nombre de billets vendus	Nombre de Transactions	Montant total des transactions
Guichet	13	4	520,60 \$
Web	3	3	110,60 \$
Total	16	7	631,20 \$

En conclusion

Grâce à ce chapitre sur la collecte des données et l'évaluation de leur qualité, vous savez maintenant comment identifier vos besoins d'informations et les mettre en relation avec les données qui sont actuellement disponibles dans votre organisation/projet et votre niveau de maturité numérique. Vous connaissez de plus les bonnes pratiques en termes de la planification et de la mise en œuvre d'une collecte de données, ainsi qu'au plan du traitement et du nettoyage de vos données.

Vous pourrez apprendre dans le 2e chapitre comment explorer un jeu de données à l'aide de statistiques descriptives et de visualisations simples. De plus, vous apprendrez à utiliser le Tableau croisé dynamique d'Excel pour explorer et analyser des données ou encore pour construire un tableau de bord avec vos principaux indicateurs de performance.