

Niveau III – Étude de cas

Étude de cas 1 : Occupation des lits d'hôpitaux avec Tableau Prep

Atteindre la capacité maximale dans un hôpital est certes problématiques, mais une surabondance des ressources l'est également. Il est important de comprendre les lits d'hôpitaux en considérant le lit en tant que ressource. Toutefois, les données sont souvent enregistrées du point de vue d'un patient. Comment prendre des données qui capturent le moment où des patients occupent des lits et déterminer l'occupation des lits ?

Vous devrez également télécharger trois fichiers de données :

- Bed.xsl
- Hour.xls
- PatientBed.xls

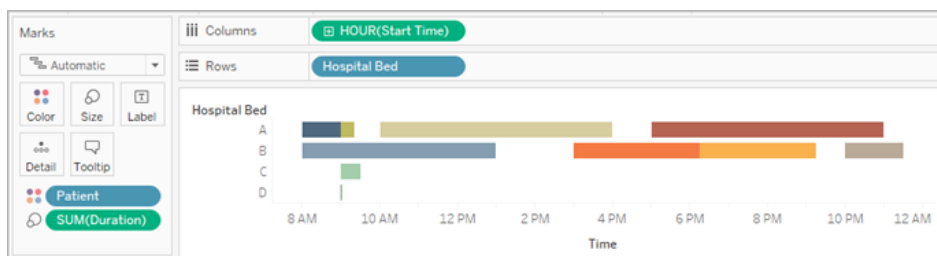
Les données

Pour nos quatre lits, A, B, C et D, nous suivons quel patient occupait le lit, et l'heure de début et de fin de son occupation. Les données se présentent comme suit :

	A	B	C	D
1	Hospital Bed	Patient	Start Time	End Time
2	A	Person 1	1/1/2018 8:34	1/1/2018 9:34
3	A	Person 5	1/1/2018 9:55	1/1/2018 10:15
4	A	Person 9	1/1/2018 10:34	1/1/2018 16:34
5	A	Person 8	1/1/2018 17:00	1/1/2018 23:00
6	B	Person 2	1/1/2018 8:45	1/1/2018 13:45
7	B	Person 6	1/1/2018 15:13	1/1/2018 18:27
8	B	Person 7	1/1/2018 18:41	1/1/2018 21:56
9	B	Person 10	1/1/2018 22:13	1/1/2018 23:43
10	C	Person 3	1/1/2018 9:05	1/1/2018 9:35
11	D	Person 4	1/1/2018 9:30	

Analyse préliminaire

Si nous introduisons les données dans Tableau Desktop, nous pouvons créer un diagramme de Gantt pour montrer à quel moment les patients occupent les lits.



Cette représentation visuelle est utile. Nous pouvons voir qu'il n'existe que de brefs intervalles entre les occupations pour les lits A et B, mais que le lit C, quant à lui, est très sous-utilisé. Il n'y a pas d'heure de fin pour le patient du lit D, mais nous pourrions gérer ce problème avec quelques calculs. Nous disposons ainsi d'une vue d'ensemble visuelle du mode d'occupation des lits.

Par contre, comment faire pour compter le nombre d'heures pendant lequel un lit était vacant ? Ou comparer la durée d'inoccupation des lits avant ou près la mise en place d'une nouvelle politique ? Il n'existe aucun moyen facile de répondre à ces questions avec la structure actuelle des données.

Structure souhaitée des données

En créant quelques ensembles de données très élémentaires et en les combinant dans Tableau Prep, nous pouvons transformer la structure de cet ensemble de données de manière à pouvoir effectuer des analyses plus approfondies et créer des visualisations encore plus utiles.

Avant de passer à Tableau Prep, prenons du recul et considérons ce dont nous avons besoin pour répondre à la question, « Pendant combien d'heures chaque lit a-t-il été inoccupé ? »

Nous devons pouvoir examiner chaque lit pour chaque heure, et savoir si un patient occupait ou non le lit. Actuellement, les données relèvent seulement quand un patient occupait le lit. Nous n'avons pas donné d'informations à Tableau au sujet des heures d'*inoccupation*.

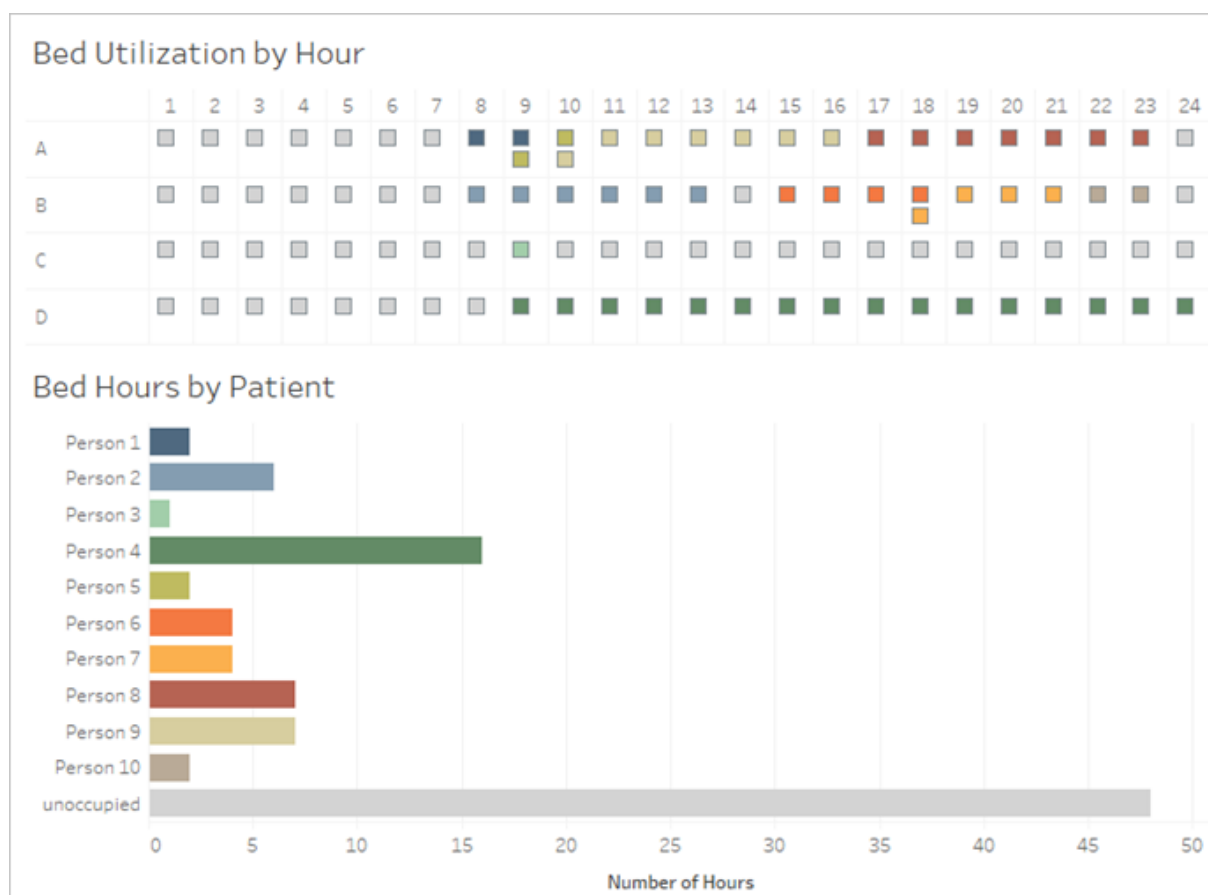
Pour créer cette matrice complète de tous les lits et de toutes les heures, nous allons créer deux ensembles de données. L'un d'eux est simplement une liste de lits (A, B, C, D) et l'autre une liste d'heure (1, 2, 3, ..., 23, 24). En effectuant une jointure croisée (en liant chaque ligne d'un ensemble de données avec chaque ligne de l'autre ensemble de données), nous obtiendrons chaque combinaison possible de lits et d'heures.

L'ensemble données Beds.xlsx se présente comme suit :	de	L'ensemble données Hours.xlsx se présente comme suit :	de	Et les résultats de la jointure croisée se présentent comme suit :																																																							
<table><tr><td></td><td>A</td></tr><tr><td>1</td><td>Bed</td></tr><tr><td>2</td><td>A</td></tr><tr><td>3</td><td>B</td></tr><tr><td>4</td><td>C</td></tr><tr><td>5</td><td>D</td></tr></table>		A	1	Bed	2	A	3	B	4	C	5	D		<table><tr><td></td><td>A</td></tr><tr><td>1</td><td>Hour</td></tr><tr><td>2</td><td>1</td></tr><tr><td>3</td><td>2</td></tr><tr><td>4</td><td>3</td></tr><tr><td>5</td><td>4</td></tr><tr><td>6</td><td>5</td></tr><tr><td>7</td><td>6</td></tr></table>		A	1	Hour	2	1	3	2	4	3	5	4	6	5	7	6		<table><tr><td></td><td>A</td><td>B</td></tr><tr><td>1</td><td>Bed</td><td>Hour</td></tr><tr><td>2</td><td>A</td><td>1</td></tr><tr><td>3</td><td>B</td><td>1</td></tr><tr><td>4</td><td>C</td><td>1</td></tr><tr><td>5</td><td>D</td><td>1</td></tr><tr><td>6</td><td>A</td><td>2</td></tr><tr><td>7</td><td>B</td><td>2</td></tr><tr><td>8</td><td>C</td><td>2</td></tr></table>		A	B	1	Bed	Hour	2	A	1	3	B	1	4	C	1	5	D	1	6	A	2	7	B	2	8	C	2
	A																																																										
1	Bed																																																										
2	A																																																										
3	B																																																										
4	C																																																										
5	D																																																										
	A																																																										
1	Hour																																																										
2	1																																																										
3	2																																																										
4	3																																																										
5	4																																																										
6	5																																																										
7	6																																																										
	A	B																																																									
1	Bed	Hour																																																									
2	A	1																																																									
3	B	1																																																									
4	C	1																																																									
5	D	1																																																									
6	A	2																																																									
7	B	2																																																									
8	C	2																																																									

Ensuite, nous allons intégrer les informations **Patient Beds** (Lits de patients) en étiquetant chaque combinaison lit-heure comme ayant ou non un patient spécifique. Nous obtenons un ensemble de données qui inclut une ligne pour chaque lit-heure, et indique si un patient occupait le lit, le nombre et les heures de début et de fin. Les valeurs null indiquent que le lit était inoccupé.

	A	B	C	D	E
1	Bed	Hour	Patient	Start Time	End Time
29	D	7			
30	A	8	Person 1	1/1/2018 8:34	1/1/2018 9:34
31	B	8	Person 2	1/1/2018 8:45	1/1/2018 13:45
32	C	8			
33	D	8			
34	A	9	Person 5	1/1/2018 9:55	1/1/2018 10:15
35	A	9	Person 1	1/1/2018 8:34	1/1/2018 9:34
36	B	9	Person 2	1/1/2018 8:45	1/1/2018 13:45
37	C	9	Person 3	1/1/2018 9:05	1/1/2018 9:35
38	D	9	Person 4	1/1/2018 9:30	
39	A	10	Person 9	1/1/2018 10:34	1/1/2018 16:34
40	A	10	Person 5	1/1/2018 9:55	1/1/2018 10:15
41	B	10	Person 2	1/1/2018 8:45	1/1/2018 13:45
42	C	10			
43	D	10	Person 4	1/1/2018 9:30	
44	A	11	Person 9	1/1/2018 10:34	1/1/2018 16:34

Avec les données de cette structure, nous pouvons effectuer des analyses de ce type, afin d'analyser les lits vacants aussi facilement que les lits avec patients.



Restructuration des données

Comment utiliser Tableau Prep à cette fin ? Nous allons créer le flux en deux parties, tout d'abord en créant la matrice Lits Heures, puis en la combinant avec les données Lits de patients. Assurez-vous de télécharger les trois fichiers (**Beds.xlsx**, **Hours.xlsx** et **Patient Beds.xlsx**) pour bien suivre le didacticiel.

Matrice Lits Heures

Tout d'abord, nous allons nous connecter au fichier **Beds.xlsx**.

1. Ouvrez Tableau Prep.
2. Dans l'écran de démarrage, cliquez sur **Se connecter aux données**.
3. Dans le volet **Connexions**, cliquez sur **Microsoft Excel**. Accédez à l'emplacement où vous avez enregistré **Bed.xlsx** et cliquez sur **Ouvrir**.
4. La feuille **Beds** devrait apparaître automatiquement dans le volet **Flux**.

Ensuite, nous devons créer un champ pour effectuer la jointure croisée avec l'ensemble de données **Hours** (Heures). Nous allons ajouter un calcul, qui est simplement la valeur **1**.

5. Dans le volet **Flux**, sélectionnez **Beds** (Lits) et cliquez sur l'étape de nettoyage suggérée.

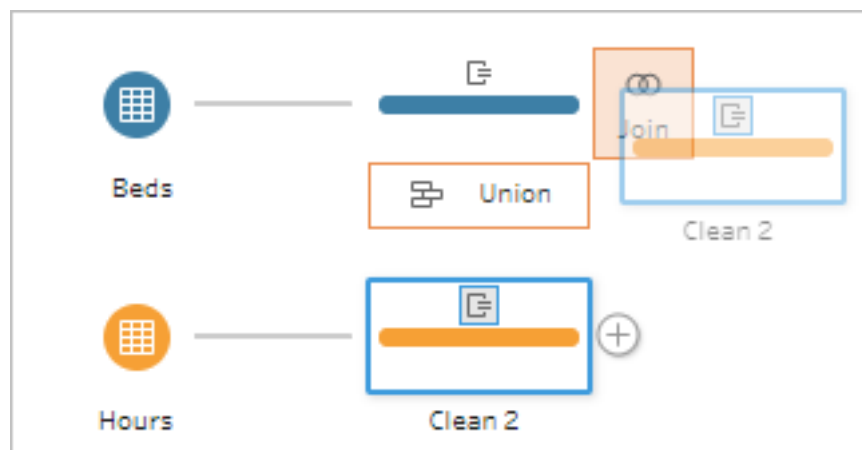
6. Avec l'étape **Nettoyer** que nous venons d'ajouter, le volet **Profil** s'affiche. Cliquez sur **Créer un champ calculé dans la barre d'outils**.
7. Nommez le champ **Cross Join** (Jointure croisée) et entrez la valeur **1**.
8. La grille **Données** devrait se mettre à jour pour afficher l'état actuel des données.

Cross Join	Bed
1	A
1	B
1	C
1	D

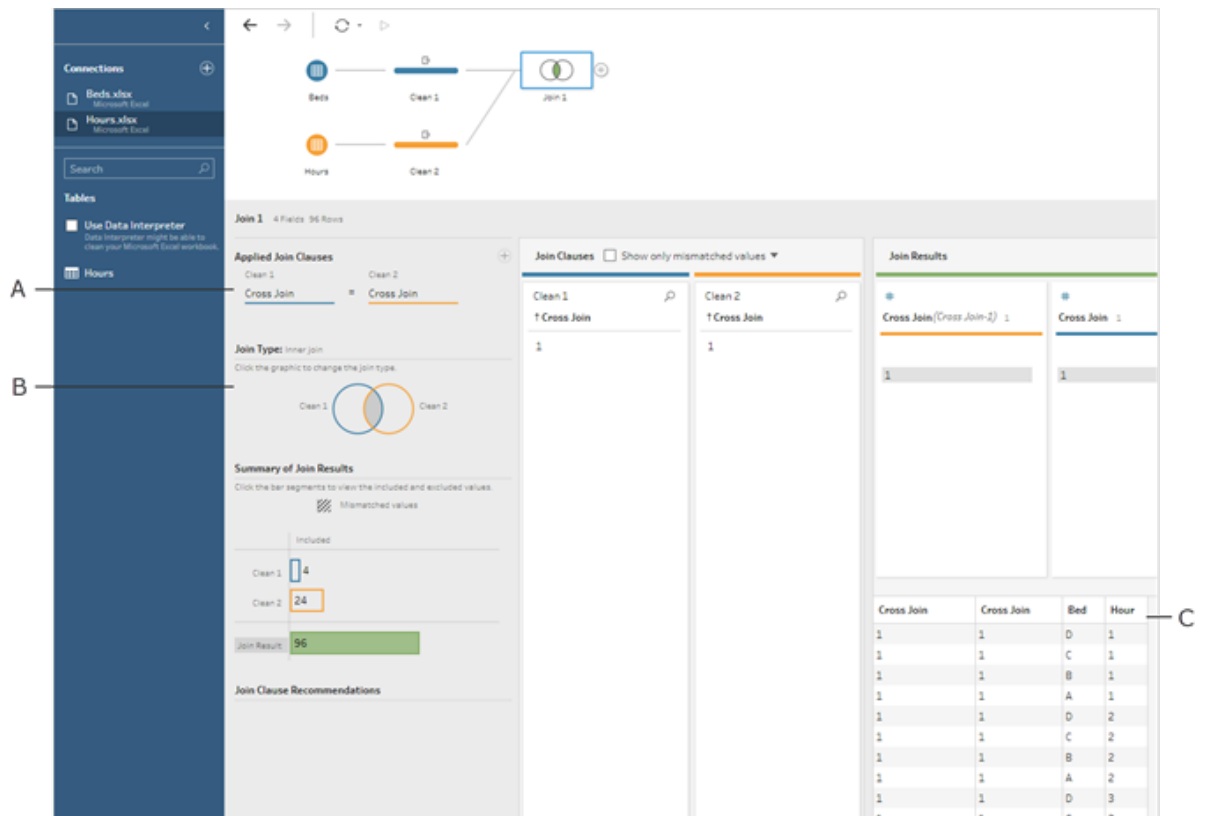
Nous allons répéter le processus pour l'ensemble de données Hours (Heures).

Les deux ensemble de données ont maintenant un champ partagé, **Cross Join**, et peuvent être liés.

13. Liez les deux étapes de nettoyage en faisant glisser **Nettoyer 2** sur **Nettoyer 1** et en déposant sur l'option **Lier**.




14. Dans le **Profil de jointure** ci-dessous, les configurations de jointure devraient se remplir automatiquement.
 - Étant donné que nous avons nommé les deux champs **Cross Join**, Tableau Prep les identifie automatiquement comme champ partagé et crée les **Clauses de jointure appliquées** appropriées.
 - Le **Type de jointure** par défaut est interne, et c'est ce que nous souhaitons.
 - Cette jointure correspondra à toutes les lignes issues de **Beds** avec toutes les lignes issues de **Hours**, comme nous l'avons vu dans la grille **Données**.



A. Clause de jointure, B. Type de jointure, C. Résultats de la grille de données


Nous n'avons plus besoin des fichiers **Cross Join** donc nous pouvons les supprimer.


15. Dans le volet **Flux**, sélectionnez **Join 1**, cliquez sur l'icône plus  , puis sélectionnez **Ajouter une étape de nettoyage**.
16. Sélectionnez les champs **Cross Join-1** et **Cross Join**, et cliquez sur **Supprimer les champs**.
17. Double-cliquez sur l'étiquette **Nettoyer 3** et renommez cette étape **Bed Hour Matrix** (Matrice Lits Heures).

Nous avons maintenant l'ensemble de données Bed Hour Matrix qui contient tous les lits et toutes les heures, et nous avons terminé la première partie de la création de notre ensemble de données.

Occupation des lits de patients

La deuxième partie consiste à introduire l'occupation des lits des patients. Pour commencer, nous nous connectons aux données.

1. Dans le volet **Connexions**, cliquez sur le bouton Ajouter une connexion  pour ajouter une autre connexion de données.
2. Choisissez **Microsoft Excel** puis sélectionnez **Patient Beds.xlsx** et cliquez sur **Ouvrir**.

3. Dans le volet **Flux**, sélectionnez **Patient Beds**, cliquez sur l'icône plus  et sélectionnez **Ajouter une étape de nettoyage**.

Le fichier Bed Hour Matrix est basé sur le nombre d'*heures* mais Patient Beds est basé sur l'*heure réelle*, nous devons extraire les heures de début et de fin de Patient Beds. En outre, pour l'heure de fin, nous voulons nous assurer que, si un patient occupe encore le lit à la fin de la journée (minuit, heure 24), ce lit sera indiqué comme occupé alors même qu'il n'y a pas d'heure de fin dans l'ensemble de données. Nous allons ajouter un champ calculé dans cette nouvelle étape.

4. Dans la barre d'outils, cliquez sur **Créer un champ calculé**.
5. Nommez le champ **Start Hour** (Heure de début). Pour le calcul, entrez DATEPART('hour',[Start Time]).

Ce calcul prend l'heure de début et l'extrait. De ce fait, « 1/1/18 9:35 AM » devient simplement « 9 ».

6. Créez un autre champ calculé appelé **End Hour** (Heure de fin). Pour le calcul, entrez IFNULL(DATEPART('hour',[End Time]), 24).

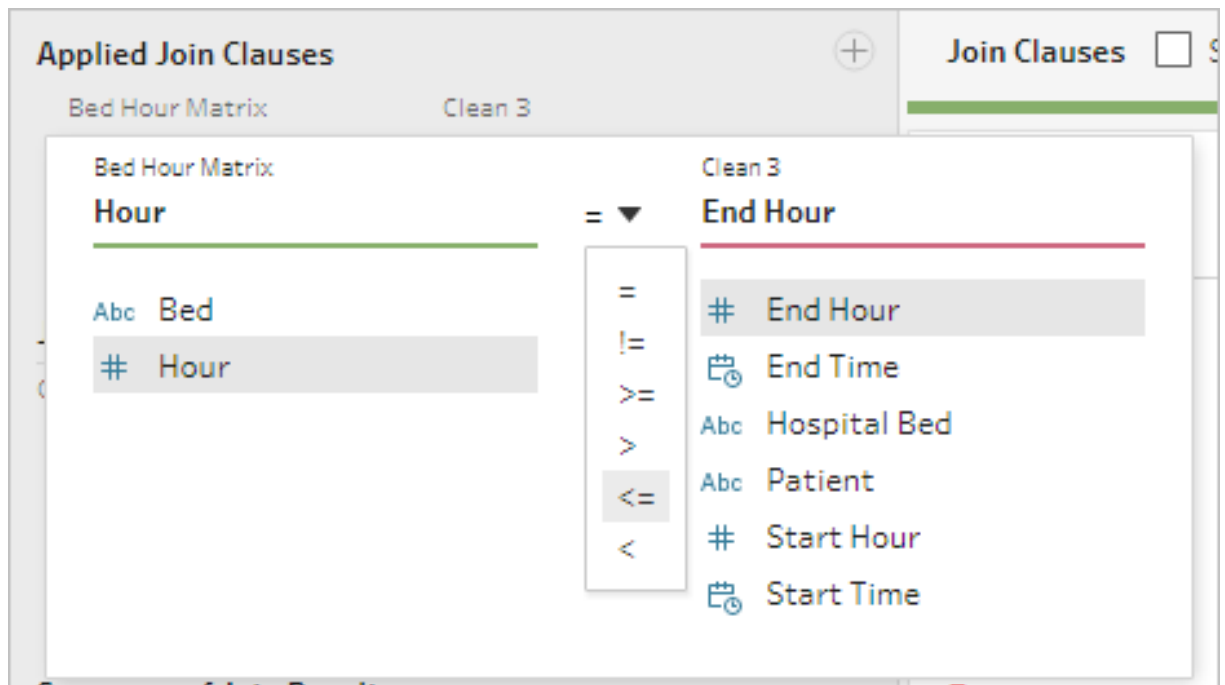
La partie DATEPART prend l'heure de fin. La partie IFNULL affectera une heure de fin de 24 (minuit) en l'absence d'heure de fin.


Nous sommes maintenant prêts à lier l'occupation des lits de patients à **Bed Hour Matrix**. Cette jointure est un peu plus complexe que les précédentes. Une jointure interne retournerait uniquement les valeurs présentes dans les deux ensembles de données. Étant donné que nous souhaitons conserver tous les emplacements lit-heure, qu'un patient ait ou non occupé le lit, nous devons effectuer une jointure gauche. Ceci entraîne beaucoup de valeurs null, mais ce résultat est correct.

Nous devons également faire l'association lorsqu'un emplacement lit-heure est pris par un patient (ou des patients). Donc, en plus d'associer le lit occupé par le patient, nous devons également prendre en considération l'heure. L'ensemble de données Bed Hour Matrix a justement un champ **Hour** (Heure) et l'ensemble de données **Patient Beds** a des champs **Start Hour** (Heure de début) et **End Hour** (Heure de fin). Nous allons utiliser des notions de logique de base pour déterminer si un patient devrait être affecté à un emplacement lit-heure donné : *Un patient est considéré comme occupant un lit si son heure de début est inférieure ou égale à (<=) l'emplacement lit-heure ET que son heure de fin est supérieure ou égale à (>=) l'emplacement lit-heure.*

De ce fait, trois clauses de jointure sont nécessaires pour associer correctement ces deux ensembles de données ensemble.

9. Liez l'étape **Nettoyer 3** avec l'étape **Bed Hour Matrix**.
10. Dans la zone **Clauses de jointure appliquées**, le paramètre par défaut devrait être **Hour = End Hour**. Cliquez sur la clause de jointure pour modifier l'opérateur de « = » en « <= ».



11. Cliquez sur le bouton plus  en haut à droite de la zone **Clauses de jointure appliquées** pour ajouter une autre clause de jointure. Définissez-la de sorte que **Hour >= Start Hour**
12. Ajoutez une troisième clause de jointure pour **Bed = Hospital Bed**.
13. Dans la section **Type de jointure**, cliquez sur la zone sans ombrage du graphique à côté de **Bed Hour Matrix** pour modifier le type de jointure sur une jointure **gauche**.

Join 2 8 Fields 99 Rows

Applied Join Clauses

Bed Hour Matrix Clean 3

Hour ≤ End Hour

Hour ≥ Start Hour

Bed = Hospital Bed

Join Type: Left Join

Click the graphic to change the join type.

Bed Hour Matrix Clean 3

Summary of Join Results

Click the bar segments to view the included and excluded values.

Mismatched values

Included

Bed Hour ... 96

Clean 3 10

Join Result 99

Join Clauses ☐ Show only mismatched values


Bed Hour Matrix		Clean 3		
↑ Hour	↑ Bed	↑ End Hour	↑ Start Hour	↑ Hospital Bed
1	A	9	8	A
1	B	9	9	C
1	C	10	9	A
1	D	13	8	B
2	A	16	10	A
2	B	18	15	B
2	C	21	18	B
2	D	23	17	A
3	A	23	22	B
3	B	24	9	D
3	C			
3	D			
4	A			
4	B			
4	C			
4	D			
5	A			
5	B			
5	C			

Remarque : si vous faites glisser **Bed Hour Matrix** vers **Nettoyer 3** et non l'inverse, vous pouvez obtenir les résultats souhaités en utilisant une jointure droite au lieu d'une jointure gauche. L'ordre dans lequel vous faites glisser les étapes a une incidence sur l'orientation de la jointure. Les clauses de jointure seront également dans l'ordre inverse. Veillez à préserver la logique correcte de comparaison des heures.

Nos données sont maintenant liées, mais il nous faut nettoyer quelques artefacts de la jointure et nous assurer que les champs sont bien ordonnés. Nous n'avons plus besoin des paramètres **Start Hour** et **End Hour**. **Hospital Bed** et **Bed** sont également redondants. Enfin, une valeur null dans le champ **Patient** signifie réellement que le lit est vacant.

14. Dans le volet **Flux**, ajoutez une étape de nettoyage afin que nous puissions mettre de l'ordre dans les données liées.
15. Faites un Ctrl+clic (Command+clic sur un Mac) pour faire une sélection multiple des champs **End Hour**, **Start Hour** et **Hospital Bed**, puis cliquez sur **Supprimer des champs** dans la barre d'outils.
16. Dans la fiche Profil du champ **Patient**, double-cliquez sur la valeur **null** et tapez **Unoccupied** (Inoccupé).

Nous avons maintenant une structure de données avec une ligne pour chaque lit-heure. Si un patient occupait un lit pendant cette heure, nous disposons également des informations sur le patient. Il ne nous reste plus qu'à ajouter une étape de sortie et à générer l'ensemble de données lui-même.

17. Dans le volet **Flux**, sélectionnez **Nettoyer 4**, cliquez sur l'icône plus  et sélectionnez **Ajouter une sortie**.
18. Dans le volet **Sortie**, modifiez le **Type de sortie** sur .csv puis cliquez sur **Parcourir**.
19. Entrez **Bed Hour Patient Matrix** comme nom, et choisissez l'emplacement souhaité avant de cliquer sur **Accepter** pour enregistrer.
20. Cliquez sur le bouton **Exécuter le flux** au bas du volet pour générer votre sortie. Cliquez sur **Terminé** dans la boîte de dialogue d'état pour fermer la boîte de dialogue.

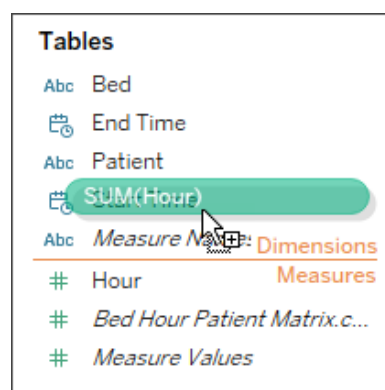
Le flux final devrait se présenter comme suit :



Analyse dans Tableau Desktop

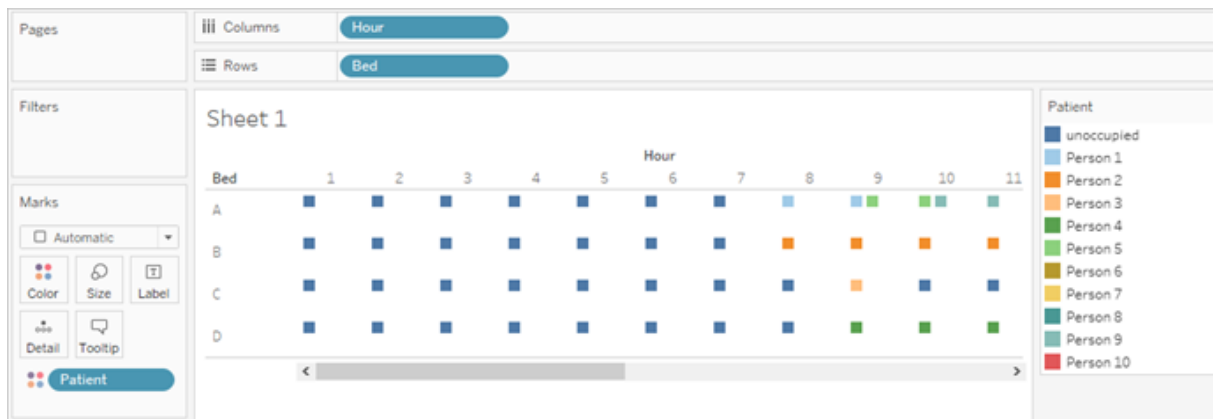
Maintenant que l'ensemble de données présente la structure souhaitée, nous pouvons effectuer une analyse plus approfondie qu'avec les données d'origine.

1. Ouvrez Tableau Desktop. Dans le volet **Connexion**, sélectionnez **Fichier texte**, accédez au fichier **Bed Hour Patient Matrix.csv**, puis cliquez sur **Ouvrir**.
2. Dans l'onglet **Source de données**, les données devraient apparaître dans l'espace de travail par défaut. Cliquez sur **Sheet 1**.
3. Dans le volet **Données**, faites glisser **Hour** au-dessus de la ligne séparant les Mesures et les Dimensions pour en faire une dimension discrète.



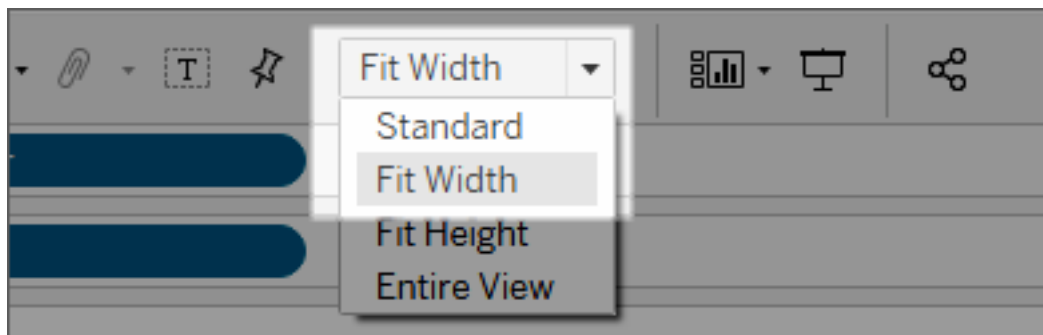
4. Faites glisser **Bed** vers l'étagère **Lignes** et **Hour** vers l'étagère **Colonnes**.

- Faites glisser **Patient** vers l'étagère **Couleur**.

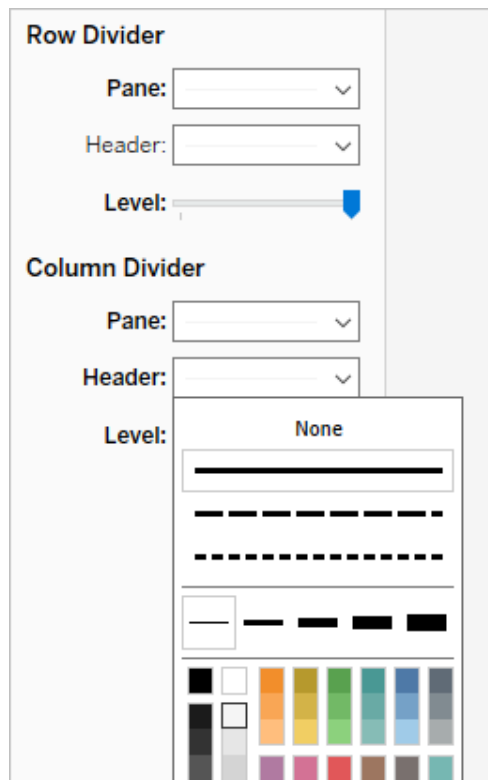


La mise en forme est facultative mais contribue à rendre les visuels plus lisibles.

- Cliquez sur l'étagère **Couleur** et sélectionnez **Modifier les couleurs**.
- Dans la zone à gauche, sélectionnez **Unoccupied**. Dans la liste déroulante à droite, choisissez la palette de couleurs **Seattle Grays**.
- Sélectionnez le quatrième gris le plus clair, et cliquez sur **OK**.
- Cliquez à nouveau sur l'étagère **Couleur**, puis cliquez sur la liste déroulante **Bordure**. Choisissez la seconde option de gris tout à droite.
- Dans la barre d'outils, dans la liste déroulante Taille, changez de **Standard** en **Adapter la largeur**.

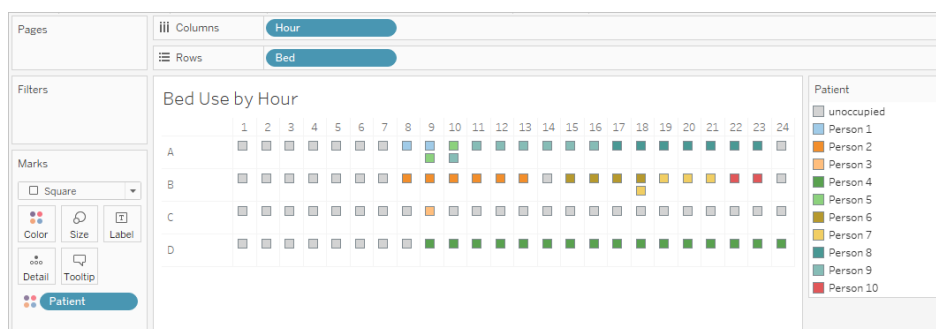


- Cliquez sur le menu **Format** puis sur **Bordures**.
- Dans **Séparateur de lignes**, cliquez sur la liste déroulante Volet et choisissez un gris très clair.
- Ajustez le curseur **Niveau** sur la seconde graduation.
- Répétez l'opération avec **Séparateur de colonnes**. Définissez la couleur **Volet** sur le gris clair et le **Niveau** sur la seconde graduation.



15. Double-cliquez sur l'onglet de feuille au bas et renommez-le **Bed Use by Hour** (Occupation des lits par heure).

Cette vue nous permet de voir rapidement quand un lit donné était occupé ou vacant.



Mais nous pouvons aller plus loin et compter le nombre d'heures pendant lesquelles chaque lit était inoccupé.

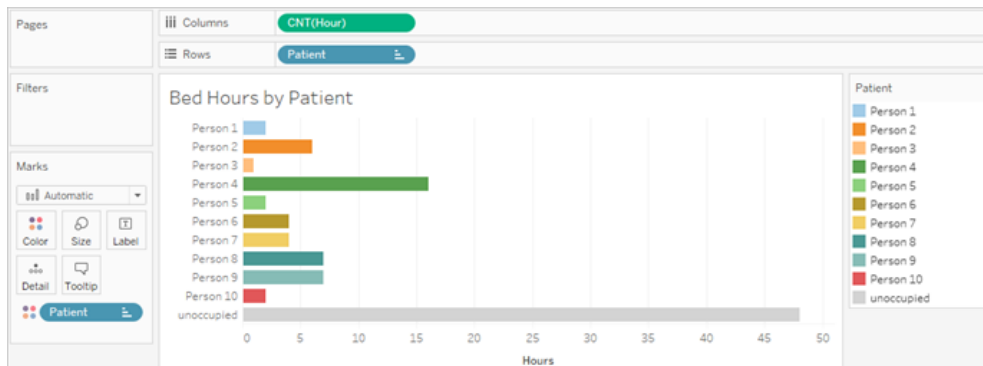


16. Cliquez sur l'icône de l'onglet de nouvelle feuille en bas pour ouvrir une nouvelle feuille.

17. Faites glisser **Patient** vers **Lignes**.

18. Faites glisser **Hour** vers **Colonnes**. Faites un clic droit sur la pile Hour pour ouvrir le menu. Choisissez **Mesure > Total**.

19. Faites glisser une autre copie du champ **Patient** depuis le volet **Données** vers l'étagère **Couleur**.
20. Faites un clic droit sur l'axe et sélectionnez **Modifier l'axe**. Modifiez le titre sur **Hours** et fermez la boîte de dialogue.
21. Renommez l'onglet de la feuille **Bed Hours by Patient**.



Cette vue nous permet d'identifier le nombre d'heures de lits inoccupés, ce que nous ne pouvions pas faire avec l'ensemble de données d'origine. Quels autres graphiques ou tableaux de bord pouvez-vous créer ? Lancez-vous maintenant que vos données présentent la structure correcte.

Récapitulatif et ressources

Pour créer cette structure de données à l'aide de Tableau Prep, nous avons dû effectuer les actions suivantes :

1. Créer un ensemble de données pour chaque aspect que nous souhaitons analyser, dans ce cas **Beds** et **Hours**.
2. Effectuer une jointure croisée de ces ensemble de données afin de créer un ensemble de données **Bed Hour Matrix** avec chaque combinaison possible de lits et d'heures.
3. Lier **Bed Hour Matrix** avec les données **Patient Bed**, en s'assurant que la jointure conserve tous les emplacements lits-heures et que les clauses de jointure associent correctement les données de lits de patients avec les emplacements lits-heures.

Nous avons utilisé les calculs suivants pour créer des champs pour la jointure. Le second et le troisième extraient les informations d'heures à partir des champs date/heure d'origine.

- **Jointure croisée = 1**
 - Affecte simplement la valeur 1 à chaque ligne
- **Heure de début = DATEPART('hour',[Start Time])**
 - Ce calcul prend l'heure de début et l'extrait. De ce fait, « 1/1/18 9:35 AM » devient simplement « 9 ».
- **Heure de fin = IFNULL(DATEPART('hour',[End Time]), 24)**

- Nous pourrions utiliser DATEPART('hour',[End Time]), comme nous l'avions fait pour **Heure de début**. Ce calcul prend l'heure de fin, et l'extrait. Ainsi, « 1/1/18 4:34 PM » devient simplement « 4 ».
- Mais nous souhaitons indiquer que le lit de patient qui est encore occupé (pas d'heure de fin) est utilisé, et non pas vacant. Pour cela, nous affectons une heure de fin de 24 (minuit) en l'absence d'heure de fin à l'aide de la fonction IFNULL. Si le premier argument DATEPART('hour',[End Time]) est null, le calcul renvoie « 24 » à la place.

Étude de cas 2 : Trouver la seconde date avec Tableau prep

En matière d'analytique, il est souvent nécessaire de déterminer la date à laquelle un *second* événement se produit, par exemple quand un client effectue un deuxième achat (devenant ainsi un client récurrent), ou quand un conducteur enfreint pour la seconde fois le code de la route. Il est facile de trouver la date d'un premier événement, il s'agit simplement de la date minimum. Trouver la seconde date est plus ardu.

Dans ce didacticiel en deux parties, nous allons organiser des données d'infractions routières et répondre aux questions suivantes :

1. Quelle était la durée, en jours, entre la première et la seconde infraction pour chaque conducteur ?
2. Comparez les montants des amendes pour la première et la seconde infraction. Sont-ils corrélés ?
3. Quel conducteur a eu la plus grosse amende ? Lequel a eu l'amende la plus faible ?
4. Combien de conducteurs ont commis plusieurs types d'infractions ?
5. Quel était le montant moyen de l'amende pour les conducteurs n'ayant jamais effectué de stage de conduite obligatoire ?

Dans la première étape, nous allons utiliser Tableau Prep Builder pour restructurer les données en vue de notre analyse. Dans la seconde étape, [Analyse avec la seconde date dans Tableau Desktop](#), nous passerons à l'analyse dans Tableau Desktop.

L'objectif de ce didacticiel est de présenter divers concepts dans le contexte d'un scénario de la vie réelle et d'explorer les options, en déterminant de manière non prescriptive quelle est la plus adaptée. Au terme de ce didacticiel, vous devriez mieux appréhender l'incidence de la structure des données sur les calculs et l'analyse, et être familiarisé avec divers aspects de Tableau Prep et les calculs dans Tableau Desktop.

Pour suivre cette étude de cas, téléchargez Traffic Violations.xlsx

Les données

Pour cet exemple, nous allons examiner des données d'infractions routières. Chaque infraction est une ligne. Le conducteur, la date, le type d'infraction, l'obligation ou non pour le conducteur de suivre un stage de conduite, et le montant de l'amende sont enregistrés.

	A	B	C	D	E
1	Driver ID	Infraction Date	Infraction Type	Traffic School	Fine Amount
2	JO-151451402	1/8/2017	Speeding	Yes	115
3	CM-127151402	3/1/2017	Running a red light	No	55
4	AP-109151404	3/2/2017	Non-moving violation	No	95
5	SH-199751404	3/4/2017	Speeding	Yes	130
6	BT-114401404	3/20/2017	Non-moving violation	No	130
7	MO-175001406	5/30/2017	Speeding	Yes	118
8	RA-1988558	6/2/2017	Speeding	Yes	144
9	BT-1168027	6/5/2017	Speeding	Yes	128
10	MO-175001406	6/18/2017	Speeding	Yes	115
11	MP-174701406	6/19/2017	Speeding	No	125
12	AA-106451404	7/5/2017	Running a red light	No	60
13	RA-199151402	7/20/2017	Speeding	Yes	146
14	SC-202601404	8/31/2017	Running a red light	No	150
15	MO-175001406	9/7/2017	Non-moving violation	No	320
16	AS-100451404	9/26/2017	Running a red light	No	50

Structure souhaitée des données

Les données sont actuellement structurées de manière à ce que chaque *infraction* soit une ligne. Un conducteur ayant commis plusieurs infractions apparaît sur plusieurs lignes, et il n'est pas facile de distinguer s'il s'agit de sa première ou de sa seconde infraction.

Pour analyser les récidivistes, nous souhaitons un ensemble de données qui distingue les dates de première et seconde infraction, les informations associées à chacune de ces infractions, et où chaque ligne est un *conducteur*.

	A	B	C	D	E	F	G	H	I
1	Driver ID	1st Infraction Date	1st Infraction Type	1st Traffic School	1st Fine Amount	2nd Infraction Date	2nd Infraction Type	2nd Traffic School	2nd Fine Amount
2	BD-117701406	12/25/2017	Speeding	Yes	140	2/7/2018	Speeding	Yes	125
3	JO-151451402	1/8/2017	Speeding	Yes	115	11/21/2018	Reckless driving	Yes	550
4	SN-207101402	12/27/2017	Speeding	Yes	280	4/26/2018	Speeding	Yes	130
5	CJ-120101402	11/26/2017	Speeding	Yes	122	3/28/2018	Speeding	Yes	116
6	JR-156701404	12/24/2017	Speeding	No	148	7/28/2018	Speeding	Yes	310
7	AP-109151404	3/2/2017	Non-moving violation	No	95	9/24/2018	Speeding	No	105
8	PC-187451406	11/11/2017	Speeding	Yes	220	12/30/2018	Non-moving violation	No	600
9	TS-214301406	9/13/2018	Speeding	Yes	115	11/10/2018	Non-moving violation	No	95
10	NP-187001404	12/11/2018	Non-moving violation	No	80	12/20/2018	Speeding	No	120
11	DB-129701402	5/13/2018	Running a red light	No	110	11/11/2018	Speeding	Yes	80
12	AJ-107951404	10/15/2017	Speeding	Yes	130	12/31/2017	Running a red light	No	85
13	BT-114401404	3/20/2017	Non-moving violation	No	130	11/13/2018	Speeding	Yes	96
14	AF-108851406	5/9/2018	Non-moving violation	No	200	9/2/2018	Speeding	No	130
15	SC-202601404	8/31/2017	Running a red light	No	150	11/10/2018	Speeding	Yes	50
16	KL-166451406	10/4/2017	Speeding	No	115	11/13/2017	Speeding	Yes	104
17	MO-175001406	5/30/2017	Speeding	Yes	118	6/18/2017	Speeding	Yes	115
18	CM-127151402	3/1/2017	Running a red light	No	55	8/1/2018	Running a red light	No	160
19	KT-164801402	5/31/2018	Non-moving violation	No	190	11/10/2018	Speeding	No	74
20	JB-160001402	11/18/2018	Speeding	Yes	220	12/5/2018	Non-moving violation	No	195
21	LH-170201404	5/6/2018	Running a red light	No	110	9/17/2018	Speeding	Yes	230
22	BG-1103555	12/25/2017	Speeding	Yes	195	12/8/2018	Speeding	Yes	315
23	MP-174701406	6/19/2017	Speeding	No	125	10/12/2017	Running a red light	No	175
24	MY-178051406	10/22/2017	Reckless driving	Yes	900	9/8/2018	Speeding	Yes	124

Avant de suivre la solution détaillée, essayez d'obtenir le résultat ci-dessus avec Tableau prep à partir du fichier Traffic violation.xlsx

Restructuration des données

Comment utiliser Tableau Prep à cette fin ? Nous allons créer le flux par étapes en récupérant la date de la première infraction, puis de la seconde, et en organisant l'ensemble de données final comme


Agrégation initiale pour la date de la 1ère infraction

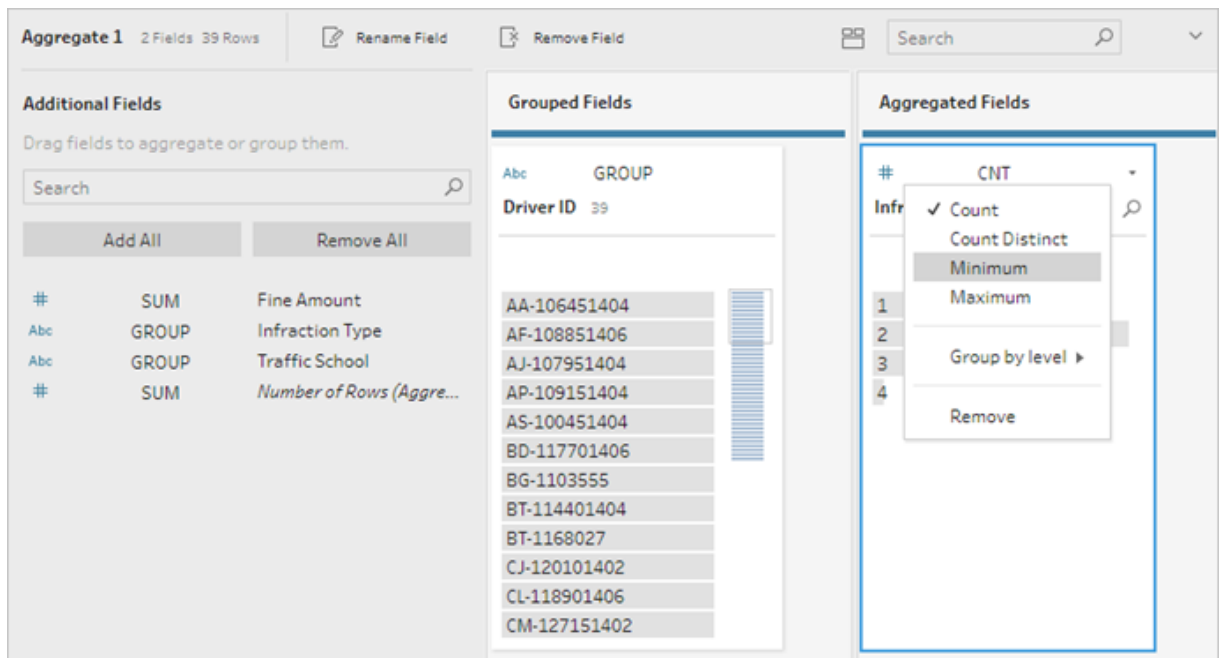
Tout d'abord, nous allons nous connecter au fichier **Traffic Violations.xlsx**.

1. Ouvrez Tableau Prep Builder.
2. Dans l'écran de démarrage, cliquez sur **Se connecter aux données**.
3. Dans le volet **Connexions**, cliquez sur **Microsoft Excel**. Accédez à l'emplacement où vous avez enregistré **Traffic Violations.xlsx** et cliquez sur **Ouvrir**.
4. La feuille **Infractions** devrait apparaître automatiquement dans le volet **Flux**.


Ensuite, nous devons identifier la date de la première infraction pour chaque conducteur. Pour cela, nous utiliserons l'étape **Agrégation**, en créant un mini-ensemble de données comportant les champs **Driver ID** (ID de conducteur) et **Minimum Infraction Date** (Date minimum de l'infraction).

Lorsque vous utilisez une étape d'agrégation dans Tableau Prep, tout champ définissant ce qui constitue une ligne est un **Champ groupé**. (Pour nous, il s'agit de **Driver ID**.) Tous les champs qui seront agrégés et présentés au niveau des champs groupés correspondent à un **Champ agrégé**. (Pour nous, il s'agit de **Infraction Date**).

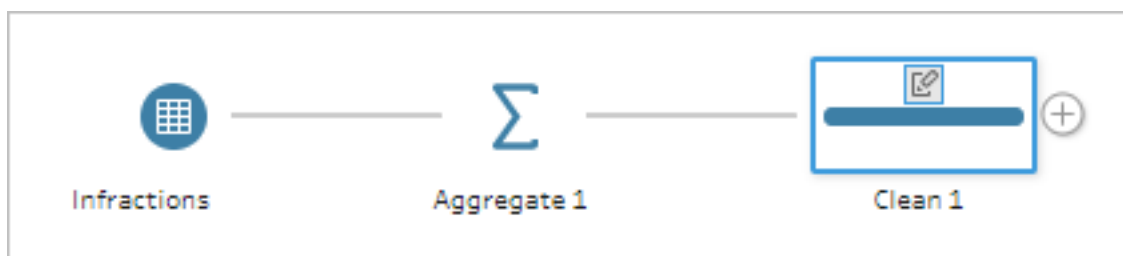
5. Dans le volet **Flux**, sélectionnez **Infractions**, cliquez sur l'icône plus  et sélectionnez **Agrégation**.
6. Faites glisser **Driver ID** vers la zone **Champs groupés**.
7. Faites glisser **Infraction Date** vers la zone **Champs agrégés**. L'agrégation par défaut est **CNT** (total). Cliquez sur **CNT** et modifiez l'agrégation sur **Minimum**.



Ceci identifie la date la plus petite (la plus ancienne), qui est la date de la première infraction par conducteur.

8. Dans le volet Flux, sélectionnez **Agrégation 1**, cliquez sur l'icône plus  et sélectionnez **Étape de nettoyage** afin que nous puissions nettoyer la sortie de l'agrégation.
9. Dans le volet **Profil**, double-cliquez sur le nom du champ **Infraction Date** et modifiez-le sur **1st Infraction Date** (Date de la 1ère infraction).

À ce stade, le flux et le profil devraient se présenter ainsi :



Clean 1 2 Fields 39 Rows Filter Values... Create Calculated Field...

>

Changes (1)

Abc

Driver ID 39


AA-106451404
AF-108851406
AJ-107951404
AP-109151404
AS-100451404
BD-117701406
BG-1103555
BT-114401404
BT-1168027
CJ-120101402
CL-118901406
CM-127151402

Calendar Icon Edit Icon

1st Infraction Date 37

01/01/2017
01/01/2019

Dans le volet Profil de cette étape de nettoyage, nous pouvons voir que nos données se composent maintenant de 39 lignes et seulement 2 champs. Tous les champs non utilisés pour le regroupement ou l'agrégation sont perdus. Par contre, nous voulons conserver une partie des informations d'origine. Nous pourrions soit ajouter ces champs au regroupement ou à l'agrégation (ce qui changerait le niveau de détail ou nécessiterait l'agrégation des données), soit lier ce mini-ensemble de données aux données d'origine (essentiellement en ajoutant une nouvelle colonne aux données d'origine pour le champ **1st Infraction Date**). Procédons à la jointure.

10. Dans le volet Flux, sélectionnez **Infractions**, cliquez sur l'icône plus  et sélectionnez **Étape de nettoyage**.

Assurez-vous de survoler l'étape Infractions directement, et non pas la ligne entre cette étape et l'étape d'agrégation. Si la nouvelle étape de nettoyage est insérée entre les deux plutôt qu'un embranchement, utilisez la flèche Annuler dans la barre d'outils et réessayez. Le menu devrait indiquer *Ajouter*, et non pas *Insérer*.



Votre flux est alors associé à toutes les données d'origine. Nous allons lier les résultats de l'agrégation à cette copie des données complètes. En effectuant une jointure avec **Driver ID**, nous ajoutons la date minimum depuis notre agrégation dans les données d'origine.

11. Sélectionnez l'étape **Nettoyer 2** et faites-la glisser sur l'étape **Nettoyer 1**, puis déposez-la sur **Lier**.
12. La configuration de jointure par défaut devrait être correcte : une jointure interne sur **Driver ID = Driver ID**.

Join 1 7 Fields 81 Rows

Applied Join Clauses (+)

Clean 1: Driver ID = Clean 2: Driver ID

Join Type: Inner join
Click the graphic to change the join type.

Clean 1 (blue circle) and Clean 2 (orange circle) Venn diagram.


Summary of Join Results
Click the bar segments to view the included and excluded values.

	Included
Clean 1	39
Clean 2	81
Join Result	81

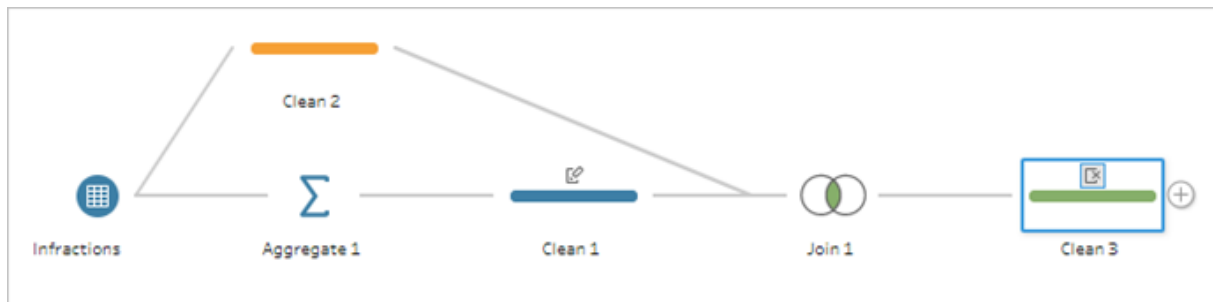
Join Clauses ☐ Show only mismatched values ▾

Clean 1	Clean 2
↑ Driver ID	↑ Driver ID
AA-106451404	AA-106451404
AF-108851406	AF-108851406
AJ-107951404	AJ-107951404
AP-109151404	AP-109151404
AS-100451404	AS-100451404
BD-117701406	BD-117701406
BG-1103555	BG-1103555
BT-114401404	BT-114401404
BT-1168027	BT-1168027
CJ-120101402	CJ-120101402
CL-118901406	CL-118901406
CM-127151402	CM-127151402
DB-129701402	DB-129701402
DJ-1342082	DJ-1342082
GZ-1454582	GZ-1454582
JB-160001402	JB-160001402

Étant donné que certains champs risquent d'être dupliqués lors d'une jointure (par exemple les champs dans la clause de jointure), il est souvent utile de nettoyer les champs superflus avant d'effectuer une jointure.

13. Dans le volet Flux, sélectionnez **Jointure 1**, cliquez sur l'icône plus  et sélectionnez **Étape de nettoyage**.
14. Dans le volet Profil, faites un clic droit ou Ctrl-clic (MacOS) sur la fiche de **Driver ID-1**, et sélectionnez **Supprimer**.
15. Pour modifier l'ordre du champ, faites glisser la fiche **1st Infraction Date** entre **Driver ID** et **Infraction Date** où vous voyez apparaître la ligne noire.

À ce stade, le flux devrait se présenter ainsi :






En examinant la grille des données ci-dessous, nous pouvons voir notre nouvel ensemble de données combinées. Nous avons la date d'infraction minimum (à savoir la première date) pour chaque conducteur ajouté à chaque ligne de l'ensemble de données.

Driver ID	1st Infraction Date	Infraction Date	Infraction Type	Traffic School	Fine Amount
JO-151451402	01/08/2017	01/08/2017	Speeding	Yes	115
CM-127151402	03/01/2017	03/01/2017	Running a red light	No	55
AP-109151404	03/02/2017	03/02/2017	Non-moving violation	No	95
SH-199751404	03/04/2017	03/04/2017	Speeding	Yes	130
BT-114401404	03/20/2017	03/20/2017	Non-moving violation	No	130
MO-175001406	05/30/2017	05/30/2017	Speeding	Yes	118
RA-1988558	06/02/2017	06/02/2017	Speeding	Yes	144
BT-1168027	06/05/2017	06/05/2017	Speeding	Yes	128
MO-175001406	05/30/2017	06/18/2017	Speeding	Yes	115
MP-174701406	06/19/2017	06/19/2017	Speeding	No	125
AA-106451404	07/05/2017	07/05/2017	Running a red light	No	60
RA-199151402	07/20/2017	07/20/2017	Speeding	Yes	146
SC-202601404	08/31/2017	08/31/2017	Running a red light	No	150
MO-175001406	05/30/2017	09/07/2017	Non-moving violation	No	320
AS-100451404	09/26/2017	09/26/2017	Running a red light	No	50
SH-199751404	03/04/2017	09/27/2017	Speeding	Yes	225
AA-106451404	07/05/2017	09/28/2017	Running a red light	No	195

Seconde agrégation pour la date de la 2nde infraction

Nous devons également déterminer la date de la seconde infraction. Pour cela, nous souhaitons filtrer à toute ligne où la date d'infraction est égale à la date minimum (en supprimant donc la première date). Nous pouvons alors prendre le minimum des dates restantes en utilisant une autre étape d'agrégation, ce qui nous laisse avec la seconde date, que nous allons renommer pour plus de clarté.


Remarque : nous souhaitons utiliser les données telles qu'elles sont actuellement dans **Nettoyer 3** ultérieurement dans le flux. Nous allons donc ajouter une autre étape de **nettoyage** pour obtenir la date de la seconde infraction. L'état actuel des données dans l'étape Nettoyer 3 sera ainsi disponible ultérieurement.

16. Dans le volet Flux, sélectionnez **Nettoyer 3**, cliquez sur l'icône plus  et sélectionnez **Étape de nettoyage**.
17. Dans la barre d'outils du volet Profil, choisissez **Filtrer les valeurs**. Créez un filtre [Infraction Date] != [1st Infraction Date].
18. Supprimez le champ **1st Infraction Date**.
19. Dans le volet Flux, sélectionnez **Nettoyer 4**, cliquez sur l'icône plus  et sélectionnez **Agréger**.
20. Faites glisser **Driver ID** vers la zone **Champs groupés**. Faire glisser **Infraction Date** vers la zone **Champs agrégés** et modifiez l'agrégation sur **Minimum**.
21. Dans le volet Flux, sélectionnez **Agréger 2**, cliquez sur l'icône plus  et sélectionnez **Étape de nettoyage**. Renommez **Infraction Date** en **2nd Infraction Date**.

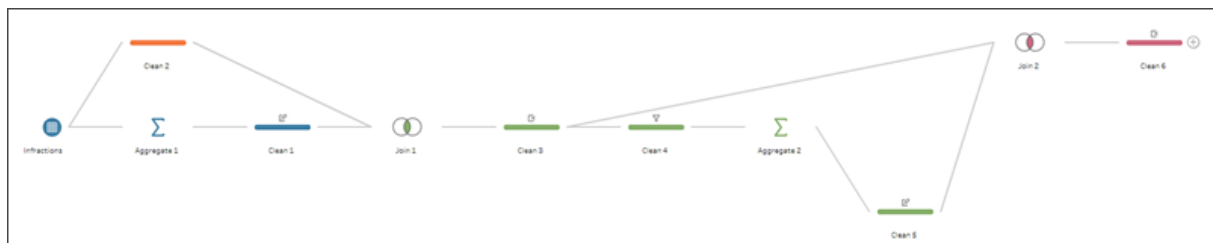
À ce stade, le flux devrait se présenter ainsi :



La date de la seconde infraction est maintenant identifiée pour chaque conducteur. Pour associer toutes les autres informations avec chaque infraction (type, amende, stage de conduite obligatoire), nous devons à nouveau les lier à l'ensemble de données tout entier.

22. Sélectionnez l'étape **Nettoyer 5** et faites-la glisser sur l'étape **Nettoyer 3**, puis déposez-la sur **Lier**.
23. À nouveau, la configuration de jointure par défaut devrait être correcte : une jointure interne sur **Driver ID = Driver ID**.
24. Dans le volet **Flux**, sélectionnez **Jointure 2**, cliquez sur l'icône plus  et sélectionnez **Étape de nettoyage**. Supprimez les champs **Driver ID-1** et **1st Infraction Date** puisqu'ils ne sont plus nécessaires.

À ce stade, le flux devrait se présenter ainsi :



Créer des ensembles de données entiers pour les 1ères et 2des infractions

Avant de poursuivre, prenons du recul et considérons tous les éléments dont nous disposons et comment nous souhaitons les faire fonctionner ensemble. Ce que nous recherchons au final est un ensemble de données qui se présente comme ceci, avec une colonne pour **Driver ID**, puis des colonnes pour la date, le type, le stage de conduite obligatoire et le montant de l'amende pour les 1ères et 2des infractions.

	A	B	C	D	E	F	G	H	I
1	Driver ID	1st Infraction Date	1st Infraction Type	1st Traffic School	1st Fine Amount	2nd Infraction Date	2nd Infraction Type	2nd Traffic School	2nd Fine Amount
2	BD-117701406	12/25/2017	Speeding	Yes	140	2/7/2018	Speeding	Yes	125
3	JO-151451402	1/8/2017	Speeding	Yes	115	11/21/2018	Reckless driving	Yes	550
4	SN-207101402	12/27/2017	Speeding	Yes	280	4/26/2018	Speeding	Yes	130
5	CJ-120101402	11/26/2017	Speeding	Yes	122	3/28/2018	Speeding	Yes	116
6	JR-156701404	12/24/2017	Speeding	No	148	7/28/2018	Speeding	Yes	310
7	AP-109151404	3/2/2017	Non-moving violation	No	95	9/24/2018	Speeding	No	105
8	PC-187451406	11/11/2017	Speeding	Yes	220	12/30/2018	Non-moving violation	No	600
9	TS-214301406	9/13/2018	Speeding	Yes	115	11/10/2018	Non-moving violation	No	95
10	NP-187001404	12/11/2018	Non-moving violation	No	80	12/20/2018	Speeding	No	120
11	DB-129701402	5/13/2018	Running a red light	No	110	11/11/2018	Speeding	Yes	80
12	AJ-107951404	10/15/2017	Speeding	Yes	130	12/31/2017	Running a red light	No	85
13	BT-114401404	3/20/2017	Non-moving violation	No	130	11/13/2018	Speeding	Yes	96
14	AF-108851406	5/9/2018	Non-moving violation	No	200	9/2/2018	Speeding	No	130
15	SC-202601404	8/31/2017	Running a red light	No	150	11/10/2018	Speeding	Yes	50
16	KL-166451406	10/4/2017	Speeding	No	115	11/13/2017	Speeding	Yes	104
17	MO-175001406	5/30/2017	Speeding	Yes	118	6/18/2017	Speeding	Yes	115
18	CM-127151402	3/1/2017	Running a red light	No	55	8/1/2018	Running a red light	No	160
19	KT-164801402	5/31/2018	Non-moving violation	No	190	11/10/2018	Speeding	No	74
20	JB-160001402	11/18/2018	Speeding	Yes	220	12/5/2018	Non-moving violation	No	195
21	LH-170201404	5/6/2018	Running a red light	No	110	9/17/2018	Speeding	Yes	230
22	BG-1103555	12/25/2017	Speeding	Yes	195	12/8/2018	Speeding	Yes	315
23	MP-174701406	6/19/2017	Speeding	No	125	10/12/2017	Running a red light	No	175
24	MP-178051406	10/13/2017	Reckless driving	Yes	600	9/9/2018	Speeding	Yes	124

Comment y parvenir ?

Dans l'étape **Nettoyer 3**, nous avons notre ensemble de données complet avec une colonne qui répète la date de la première infraction pour chaque conducteur.

Driver ID	1st Infraction Date	Infraction Date	Infraction Type	Traffic School	Fine Amount
JO-151451402	01/08/2017	01/08/2017	Speeding	Yes	115
CM-127151402	03/01/2017	03/01/2017	Running a red light	No	55
AP-109151404	03/02/2017	03/02/2017	Non-moving violation	No	95
SH-199751404	03/04/2017	03/04/2017	Speeding	Yes	130
BT-114401404	03/20/2017	03/20/2017	Non-moving violation	No	130
MO-175001406	05/30/2017	05/30/2017	Speeding	Yes	118
RA-1988558	06/02/2017	06/02/2017	Speeding	Yes	144
BT-1168027	06/05/2017	06/05/2017	Speeding	Yes	128
MO-175001406	05/30/2017	06/18/2017	Speeding	Yes	115
MP-174701406	06/19/2017	06/19/2017	Speeding	No	125
AP-109151404	03/02/2017	03/02/2017	Non-moving violation	No	95

Nous souhaitons éliminer toutes les autres lignes pour un conducteur qui ne correspondent pas à la première infraction, créant ainsi un ensemble de données comportant uniquement les premières


infractions. C'est-à-dire que nous souhaitons conserver uniquement les informations pour un conducteur donné lorsque **1st Infraction Date = Infraction Date**. Une fois que nous avons appliqué un filtre pour conserver uniquement la ligne de la première infraction, nous pouvons supprimer le champ **Infraction Date** et mettre de l'ordre dans les noms des champs.

De même, après la seconde agrégation et jointure, nous avons notre ensemble de données complet avec une colonne pour la seconde date d'infraction.

Driver ID	2nd Infraction Date	Infraction Date	Infraction Type	Traffic School	Fine Amount
JO-151451402	11/21/2018	01/08/2017	Speeding	Yes	115
CM-127151402	08/01/2018	03/01/2017	Running a red light	No	55
AP-109151404	09/24/2018	03/02/2017	Non-moving violation	No	95
SH-199751404	09/27/2017	03/04/2017	Speeding	Yes	130
BT-114401404	11/13/2018	03/20/2017	Non-moving violation	No	130
MO-175001406	06/18/2017	05/30/2017	Speeding	Yes	118
MO-175001406	06/18/2017	06/18/2017	Speeding	Yes	115
MP-174701406	10/12/2017	06/19/2017	Speeding	No	125
AA-106451404	09/28/2017	07/05/2017	Running a red light	No	60
RA-199151402	12/31/2017	07/20/2017	Speeding	Yes	146
SC-202601404	11/10/2018	08/21/2017	Running a red light	No	150

Nous pouvons appliquer un filtre similaire **2nd Infraction Date = Infraction Date** pour conserver uniquement la ligne d'information pour la 2nde infraction de chaque conducteur. À nouveau, nous pouvons également renommer **Infraction Date** qui est maintenant redondant et mettre un peu d'ordre dans les noms des champs.

Nous allons commencer avec l'ensemble de données des premières infractions.

25. Dans le volet Flux, sélectionnez **Nettoyer 3**, cliquez sur l'icône plus  et sélectionnez **Étape de nettoyage**.

Comme à l'étape 10 ci-dessus, nous voulons ajouter une branche pour la nouvelle étape de nettoyage, et non pas l'insérer entre Nettoyer 3 et Nettoyer 4.

26. Avec cette nouvelle étape **Nettoyer** sélectionnée, dans le volet **Profil**, cliquez sur **Filtrer des valeurs** dans la barre d'outils. Créez un filtre [1st Infraction Date] = [Infraction Date].

27. Supprimez le champ **Infraction Date**.

28. Renommez les champs **Infraction Type**, **Traffic School** et **Fine Amount** de manière à ce qu'ils commencent par « 1st ».

29. Double-cliquez sur le nom **Nettoyer 7** sous l'étape dans le volet **Flux** et renommez-le **Robust 1st**.

Occupons-nous maintenant de l'ensemble de données des secondes infractions.

30. Dans le volet Flux, sélectionnez **Nettoyer 6**, après la dernière jointure.

31. Cliquez sur **Filtrer les valeurs** dans la barre d'outils. Créez un filtre [2nd Infraction Date] = [Infraction Date].


32. Supprimez le champ **Infraction Date**.
33. Renommez les champs **Infraction Type**, **Traffic School** et **Fine Amount** de manière à ce qu'ils commencent par « 2nd ».
34. Double-cliquez sur le nom **Nettoyer 6** sous l'étape dans le volet Flux et renommez-le **Robust 2nd**.

À ce stade, le flux devrait se présenter ainsi :





Créer l'ensemble de données complet

Nous disposons maintenant de ces deux ensembles de données bien organisés avec des informations complètes sur les premières et secondes infractions par conducteur. Nous allons pouvoir les lier à nouveau ensemble sur **Driver ID** et obtenir la structure de données recherchée.

35. Sélectionnez **Robust 2nd** et faites-le glisser sur **Robust 1st**, en le déposant sur **Lier**.
36. La configuration de jointure par défaut devrait être correcte, sous la forme **Driver ID = Driver ID**.
37. Nous ne souhaitons pas déposer les conducteurs qui n'ont pas commis de seconde infraction, donc nous allons effectuer une jointure gauche. Dans la zone **Type de jointure**, cliquez sur la zone sans ombrage du diagramme à côté de **Robust 1st**, pour la transformer en jointure **Gauche**.
38. Dans le volet Flux, sélectionnez **Jointure 3**, cliquez sur l'icône plus  et sélectionnez **Étape de nettoyage**. Supprimez le champ en double **Driver ID-1**.

Les données sont dans l'état souhaité, donc nous pouvons créer une sortie et passer à l'analyse.

39. Dans le volet Flux, sélectionnez **Nettoyer 6** que vous venez d'ajouter, cliquez sur l'icône plus  et sélectionnez **Ajouter une sortie**.
40. Dans le volet **Sortie**, modifiez le **Type de sortie** sur **.csv** puis cliquez sur **Parcourir**. Entrez **Driver Infractions** comme nom, et choisissez l'emplacement souhaité avant de cliquer sur **Accepter** pour enregistrer.
41. Cliquez sur le bouton **Exécuter le flux**  au bas du volet pour générer votre sortie. Cliquez sur **Terminé** dans la boîte de dialogue d'état pour fermer la boîte de dialogue.

Le flux final devrait se présenter comme suit :



Remarque : vous pouvez télécharger le fichier de flux terminé pour vérifier votre travail : [Driver Infractions.tflx](#)

Récapitulatif

Pour la première étape de ce didacticiel, notre objectif était de prendre notre ensemble de données d'origine et de le préparer pour l'analyse relative aux premières et secondes infractions. Le processus consiste en trois phases :

Identifier les dates de premières et secondes infractions :

1. Créer une agrégation qui conserve **Driver ID** et **MIN Infraction Date**. Liez cette agrégation avec l'ensemble de données d'origine pour créer un « ensemble de données intermédiaire » comportant la date de la première infraction (minimum) répétée pour chaque ligne.
2. Dans une nouvelle étape, filtrez toutes les lignes où **1st Infraction Date** est identique à **Infraction Date**. À partir de cet ensemble de données filtré, créez une agrégation qui conserve **Driver ID** et **MIN Infraction date**. Liez avec l'ensemble de données intermédiaire issu de la première étape. Cette opération identifie la date de la seconde infraction.

Créer des ensembles de données propres pour les premières et secondes infractions :

3. Revenez en arrière et créez une branche à partir de l'ensemble de données intermédiaire et appliquez un filtre de manière à conserver uniquement les lignes où **1st Infraction Date** est identique à **Infraction Date**. Un ensemble de données limité à la première infraction est alors créé. Mettez-y de l'ordre en supprimant tous les champs inutiles, et renommez tous les champs souhaités (à l'exception de **Driver ID**) pour indiquer qu'ils concernent la première infraction. Voici l'ensemble de données **Robust 1st**.
4. Mettez de l'ordre dans l'ensemble de données relatif à la date de la seconde infraction. Nettoyez les résultats de la jointure de l'étape 2 en appliquant un filtre de manière à conserver uniquement les lignes où **2nd Infraction Date** est identique à **Infraction Date**. Supprimez tous les champs inutiles, et renommez tous les champs souhaités (à l'exception de **Driver ID**) pour indiquer qu'ils concernent la seconde infraction. Voici l'ensemble de données **Robust 2nd**.

Combiner les données de premières et secondes infractions en un seul ensemble de données :

5. Liez les ensembles de données **Robust 1st** et **Robust 2nd** en veillant à conserver tous les enregistrements de **Robust 1st** pour éviter de perdre les conducteurs n'ayant pas commis de seconde infraction.

Ensuite, Dans Tableau Desktop, répondez aux questions suivantes :

- ➔ Quelle était la durée, en jours, entre la première et la seconde infraction pour chaque conducteur ?
- ➔ Comparez les montants des amendes pour la première et la seconde infraction. Sont-ils corrélés ?
- ➔ Quel conducteur a eu la plus grosse amende ? Lequel a eu l'amende la plus faible ?
- ➔ Combien de conducteurs ont commis plusieurs types d'infractions ?

Réaliser un Dashboard à partir de ces réponses et de la base de données.

Détail de l'exercice sur Tableau desktop

https://help.tableau.com/current/prep/fr-fr/prep_tutorial_2nddateB.htm

Aller plus loin—Données permutées

Si les données avec lesquelles nous avons travaillé sont bien structurées pour répondre à des questions centrées spécifiquement sur les premières et secondes infractions, ce n'est pas la structure standard recommandée à utiliser avec Tableau Desktop. Plus notre analyse s'écarte des questions de base sur les dates des infractions, plus nos calculs se complexifient pour combiner les informations pertinentes sous une forme utilisable.


En règle générale, lorsque les données sont stockées avec plusieurs colonnes pour le même type de données (par exemple deux colonnes pour la date, deux colonnes pour le montant de l'amende, etc.) et que des informations uniques sont stockées dans le nom du champ (par exemple s'il s'agit de la première ou de la seconde infraction), ceci indique que les données devraient être permutées.

Une permutation multiple dans Tableau Prep Builder peut gérer cette tâche avec efficacité. Nous pouvons travailler à partir de la fin du flux **Driver Infraction** que Tableau Prep a créé dans le didacticiel précédent.

Vous cherchez à obtenir les données sous cette forme sans la solution détaillée dans un premier temps.

Driver ID	Infraction Number	Infraction Date	Infraction Type	Traffic School	Fine Amount
MO-175001406	1st	05/30/2017	Speeding	Yes	118
SH-199751404	1st	03/04/2017	Speeding	Yes	130
AA-106451404	1st	07/05/2017	Running a red light	No	60
MP-174701406	1st	06/19/2017	Speeding	No	125
PO-188501402	1st	10/30/2017	Speeding	Yes	120
KL-166451406	1st	10/04/2017	Speeding	No	115
RA-199151402	1st	07/20/2017	Speeding	Yes	146
AJ-107951404	1st	10/15/2017	Speeding	Yes	130
BD-117701406	1st	12/25/2017	Speeding	Yes	140
CJ-120101402	1st	11/26/2017	Speeding	Yes	122
SN-207101402	1st	12/27/2017	Speeding	Yes	280
TS-213701404	1st	10/23/2017	Speeding	Yes	100
JR-156701404	1st	12/24/2017	Speeding	No	148
CM-127151402	1st	03/01/2017	Running a red light	No	55
JK-156251406	1st	12/25/2017	Speeding	Yes	140
AE-108951406	1st	05/09/2018	Non-motorist violation	No	200

Conseil : veuillez à revenir dans Tableau Prep pour les étapes suivantes.

- Depuis l'étape de nettoyage finale, ajoutez une étape **Permutation** pour permuter chaque champ dupliqué. Utilisez l'icône plus  en haut à droite de la zone **Champs permutés** pour ajouter d'autres **Valeurs de permutation**. Chaque ensemble de champ (par exemple les montants de 1ère et 2nd amende) devrait être permuté ensemble.
- Dans la zone Champs pivotés, sous la colonne **Noms Pivot1**, cliquez deux fois sur chaque valeur et renommez-les en 1ère et 2nde.

Pivot 1 6 Fields 78 Rows					
Fields	Pivoted Fields <input checked="" type="checkbox"/> Automatically rename pivoted fields and values				
Search					
Abc Driver ID	Pivot1 Names	Fine Amount	Infraction Date	Infraction Type	Traffic School
	1st	1st Fine Amount	1st Infraction Date	1st Infraction Type	1st Traffic School
	2nd	2nd Fine Amount	2nd Infraction Date	2nd Infraction Type	2nd Traffic School

Les résultats peuvent être mis en ordre en supprimant les dates null et en renommant et réorganisant les champs.

- Ajoutez une étape de nettoyage après la permutation. Dans la colonne **Infraction Date**, faites un clic droit sur la barre null et choisissez **Exclude**.
- Double-cliquez sur le nom du champ **Pivot1 Names** et renommez-le **Infraction Number**.
- Faites glisser les champs comme approprié pour les réorganiser comme ci-dessous :

Driver ID	Infraction Number	Infraction Date	Infraction Type	Traffic School	Fine Amount
MO-175001406	1st	05/30/2017	Speeding	Yes	118
SH-199751404	1st	03/04/2017	Speeding	Yes	130
AA-106451404	1st	07/05/2017	Running a red light	No	60
MP-174701406	1st	06/19/2017	Speeding	No	125
PO-188501402	1st	10/30/2017	Speeding	Yes	120
KL-166451406	1st	10/04/2017	Speeding	No	115
RA-199151402	1st	07/20/2017	Speeding	Yes	146
AJ-107951404	1st	10/15/2017	Speeding	Yes	130
BD-117701406	1st	12/25/2017	Speeding	Yes	140
CJ-120101402	1st	11/26/2017	Speeding	Yes	122
SN-207101402	1st	12/27/2017	Speeding	Yes	280
TS-213701404	1st	10/23/2017	Speeding	Yes	100
JR-156701404	1st	12/24/2017	Speeding	No	148
CM-127151402	1st	03/01/2017	Running a red light	No	55
JK-156251406	1st	12/25/2017	Speeding	Yes	140
AE-108951406	1st	05/09/2017	Non-motorist violation	No	200

- À partir des nouvelles données permutées, créez une sortie appelée **Pivoted Driver Infractions** (Infractions des conducteurs Données permutées) et insérez-la dans Tableau Desktop. (N'oubliez pas d'exécuter le flux après avoir ajouté l'étape **Sortie**.)

Avantages des données permutées

Nous pourrions nous en tenir à la structure de données d'origine du didacticiel s'il n'y avait que des questions auxquelles cette structure permettrait de répondre facilement. Mais le format de données permutées est plus flexible. Même s'il exige quelques calculs, une fois en place, l'ensemble de données résultant est bien adapté pour répondre à des questions plus larges.