TP3 VERNAY

A partir des rapports de criminalité émanant de tous les quartiers de San Francisco sur 12 années (de 1934 à 1963), on vous demande de réaliser des visualisations dans l'objectif de révéler les tendances cachées.

1. Exploitation du fichier des données

• Télécharger le fichier « train.csv ».

• Ouvrez le fichier et identifiez les principales dimensions et les mesures associées

On a plusieurs variables (qui correspondent aux colonnes du fichier train.csv).

Dates : la date enregistrée du crime ; variable catégorielle ordinale

Category : la catégorie du crime ; variable catégorielle nominale, des exemples de modalités sont : WARRANTS, LARCENY/THEFT …

Descript : description du crime ; variable textuelle

DayOfWeek : le jour de la semaine du crime ; variable catégorielle nominale, les modalités étant les jours de la semaine

PdDistrict : lieu du crime ; variable catégorielle spatiale, les modalités sont les noms des districts

Resolution : comment le crime a été géré par la police ; variable catégorielle nominale ; modalités NONE ; ARREST, BOOKED ; ARREST,CITED …

Address : l'adresse du lieu où le crime s'est produit ; variable textuelle

X et Y : données longitudinales de l'adresse du crime

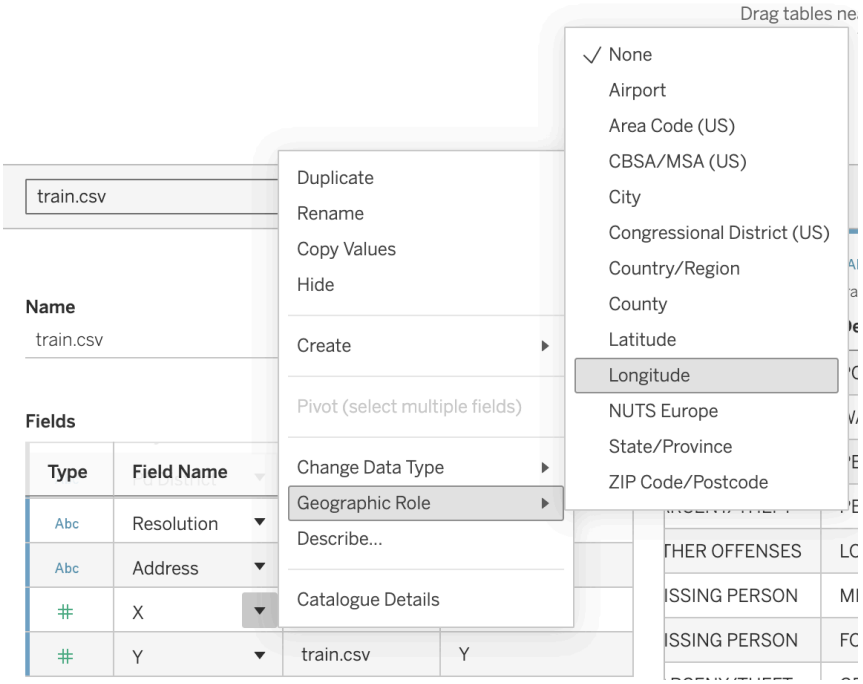
• Importez le fichier avec le logiciel Tableau Public :

o Que constatez-vous au niveau des champs des dimensions (volet source de données ou feuille de calcul) ?

On voit qu'ils sont considérés comme des décimaux. C'est correct, mais dans Tableau on veut leur attribuer un rôle géographique de longitude pour X et latitude pour Y.

o Corrigez ce problème

Dans Data source, on attribue le rôle géographique. Ex :



Options qui permettent de changer le rôle

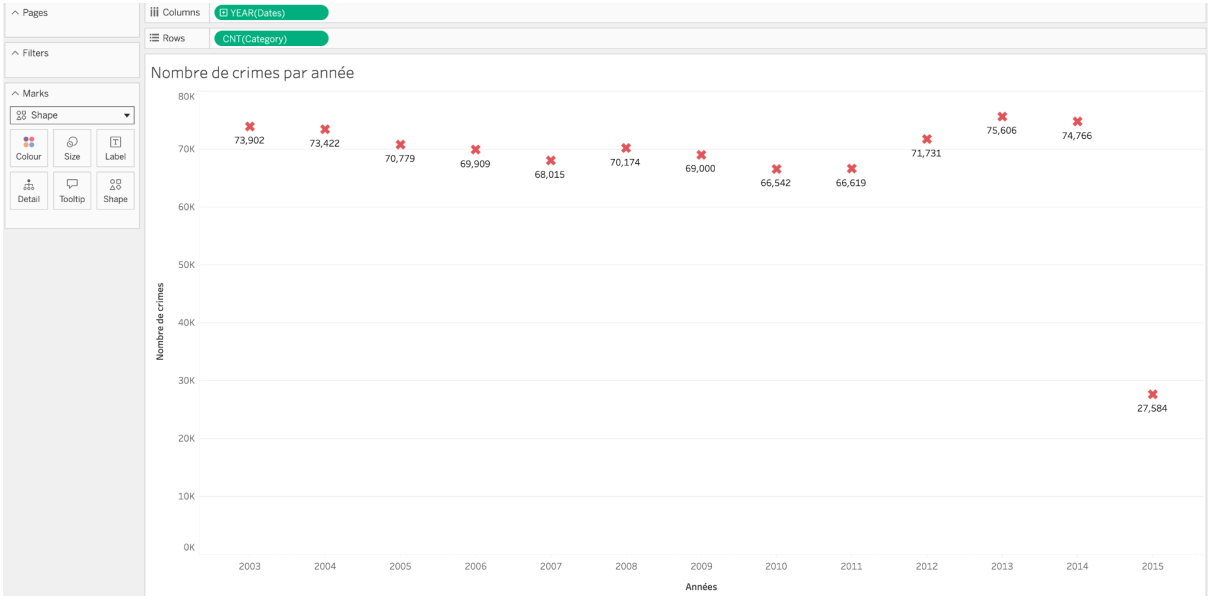
Voilà le résultat une fois la modification effectuée

2. Visualisation et analyse

Réalisez les visualisations nécessaires pour répondre aux questions suivantes :

• Y a-t-il une tendance annuelle / mensuelle / quotidienne / horaire

o Quels sont les périodes où la criminalité est la plus élevée et plus basse (heures/jours/semaine/mois)

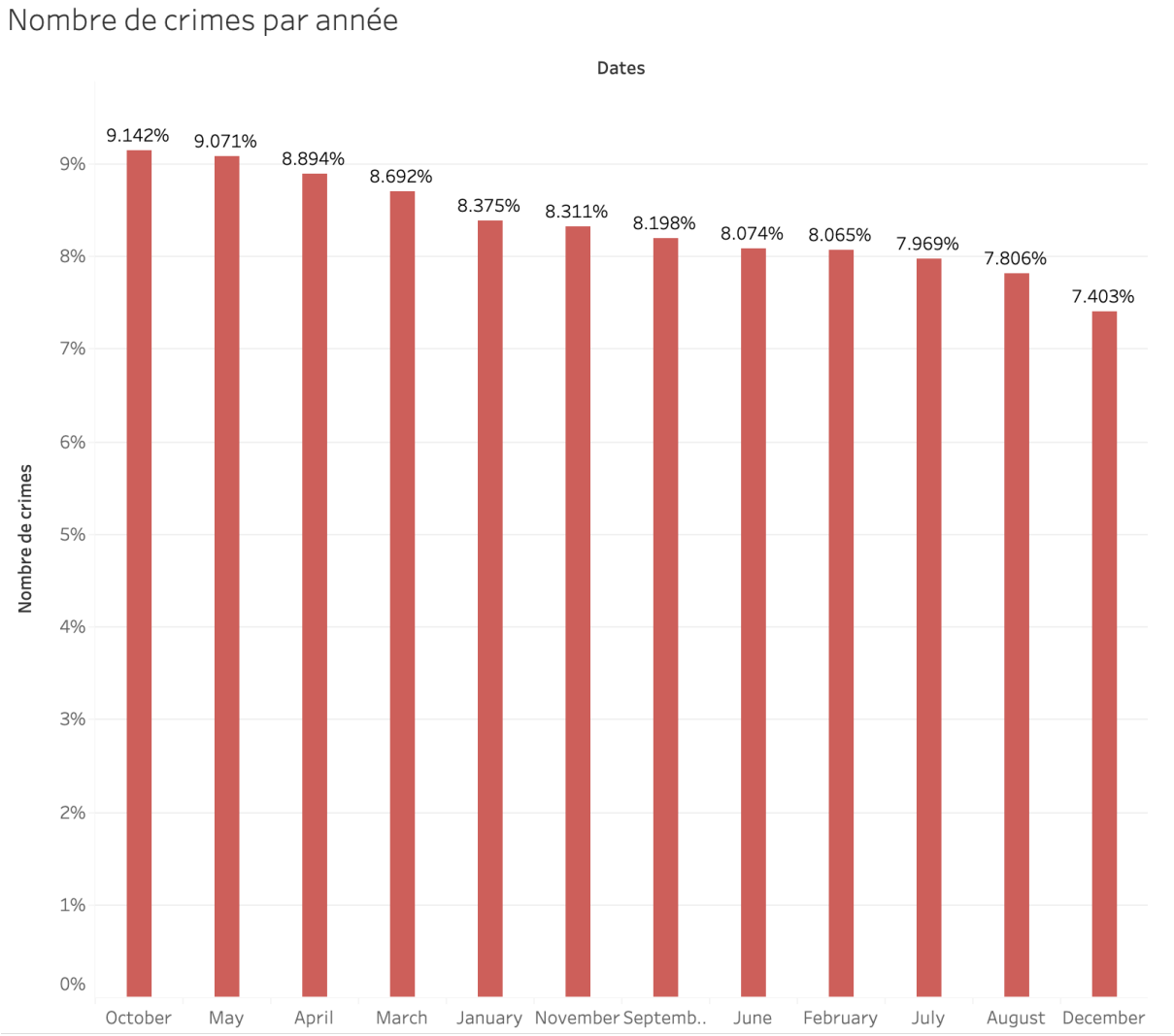


Visualisation du nombre de crimes par année

Pour des données discrètes, il est important d'utiliser des points et non pas une ligne ; qui donnerait l'impression (fausse) de données continues. On notera que la baisse en 2015 est sûrement dû au fait que les données ne sont pas encore complètes pour cette année.

Ici on change juste les colonnes pour mettre l'option mois au lieu d'année.

On a le nombre de crimes pour les mois sur la totalité de la période étudiée. Mais un défaut est que cette visualisation donne l'impression d'une continuité dans le temps, alors que chaque mois prend en compte toutes les années de notre jeu de données.

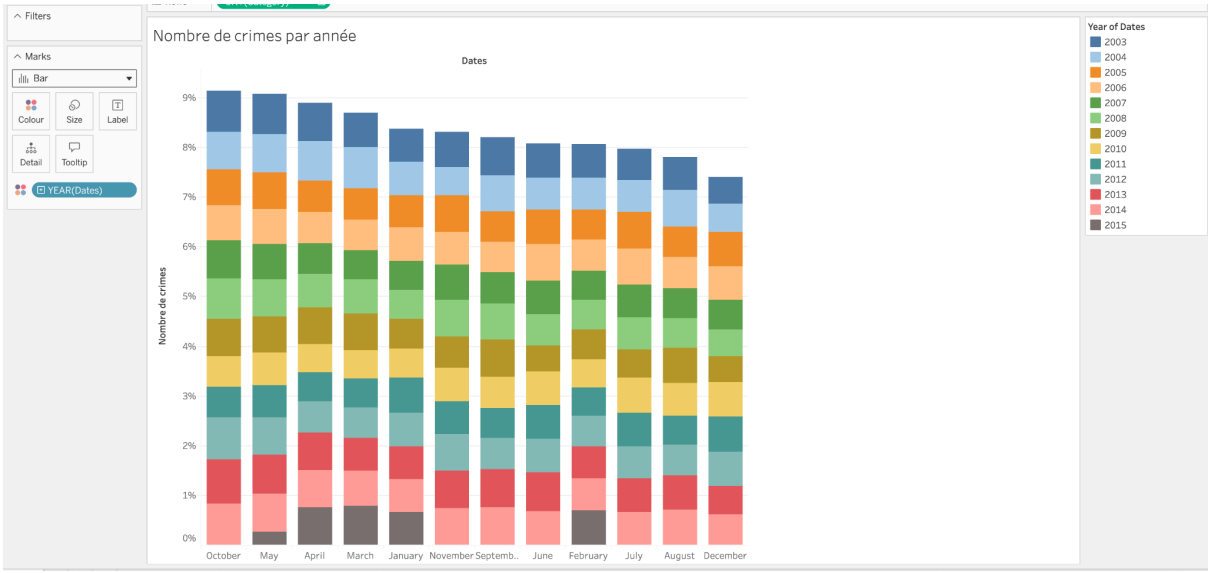


En allant dans rows (CNT(category)) puis quick table calculations, on peut aussi afficher le pourcentage de crimes pour le mois en question sur la période étudiée.

Un bar chart donne une meilleure impression pour ces données qui ne sont pas forcément organisées dans le temps (le mois d'octobre contient celui de 2005, 2006 etc…).

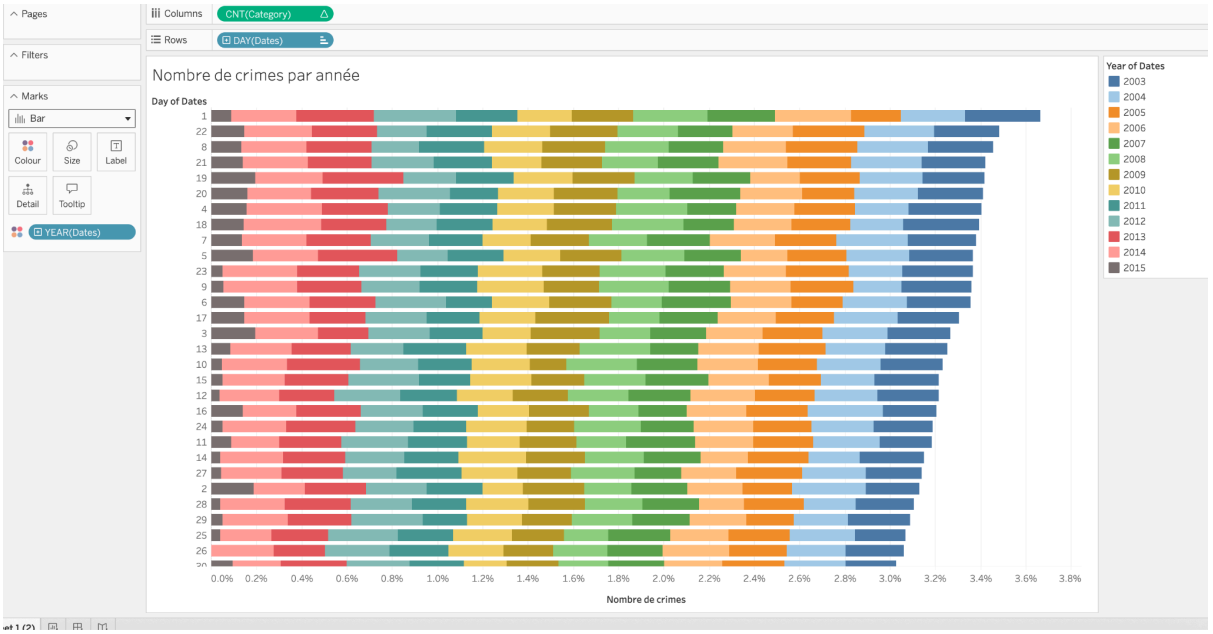
On tri en décroissant pour tout de suite avoir un aperçu du mois avec le taux le plus élevé. L'interprétation correcte ici, par exemple pour octobre, est que 9.142% de la totalité des crimes commis sur la période étudiée ont été commis un mois d'octobre.

Une limite possible serait qu'on ne connaisse pas la répartition de ces crimes sur les différentes années pour chaque mois.

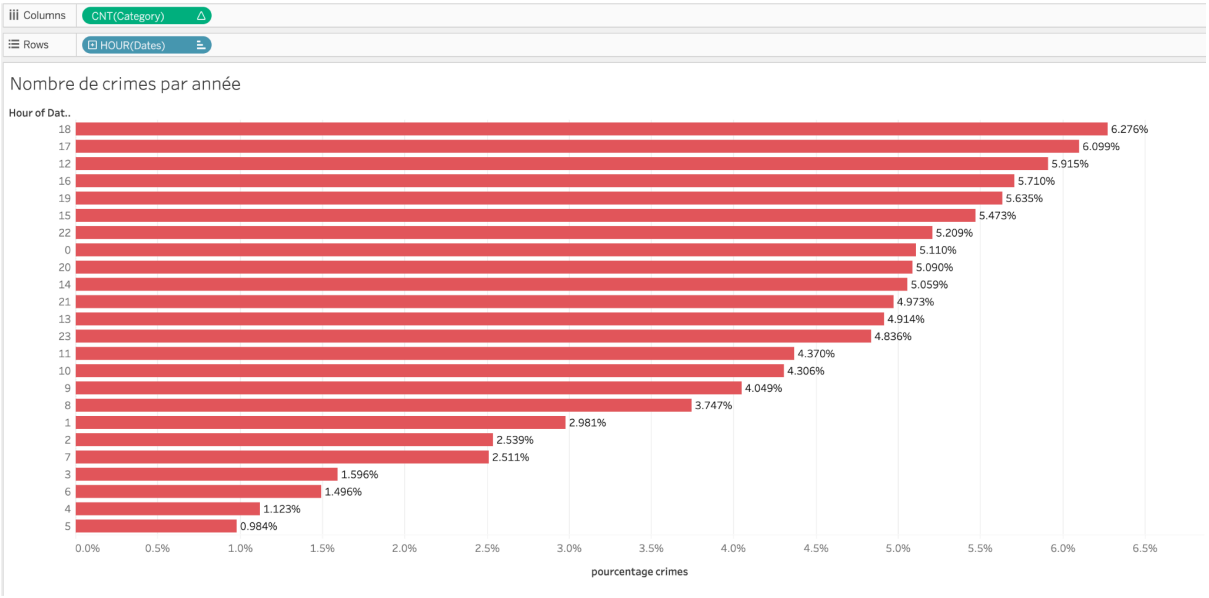


Cette visualisation répond à cette limite. On représente les différentes années pour chaque mois. Ainsi on peut s'apercevoir si une année a plus contribué à la criminalité d'un certain mois. Mais ici, on voit que les années sont réparties plutôt équitablement.

Il semblerait donc que le mois avec le plus de crime soit le mois d'octobre.



Le même type de graphique pour les jours de l'année révèlent que le 1er du mois est le jour avec le plus de taux de criminalité.



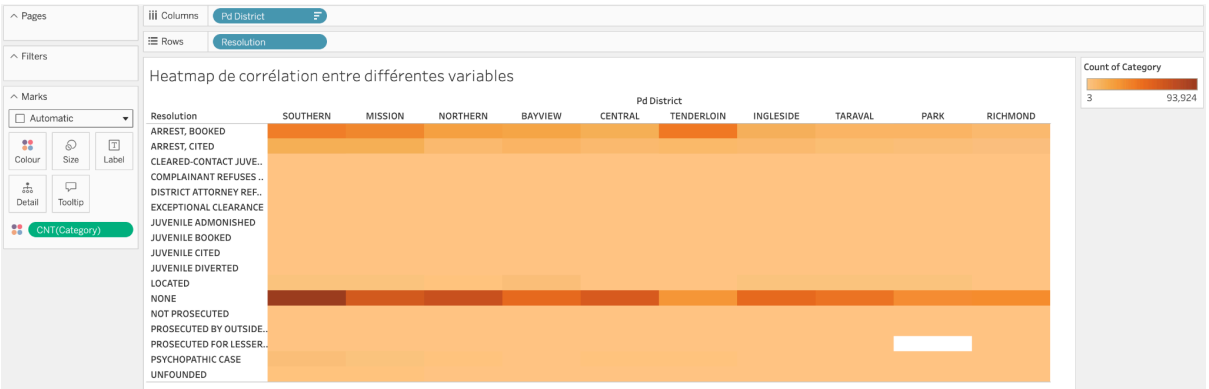
Et enfin pour les heures, avec l'heure de la journée où le pourcentage de crime enregistré est le plus grand : 18h. Pour afficher les pourcentages au bout de chaque colonne, on appuie simplement sur label dans Marks.

On préfère une vue horizontale, sinon les labels de pourcentage s'affiche à la verticale et on les lit avec difficultés.

• Existe-t-il des clusters spécifiques avec un taux de criminalité plus élevé ?

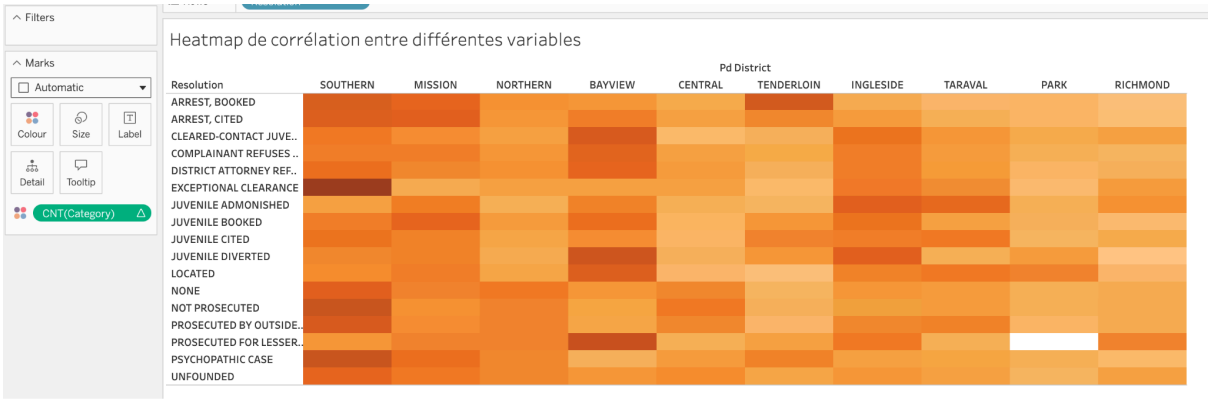
o Ex. District vs Resolution, Category vs Resolution, Category vs Days, Category vs District…

DISTRICT vs RESOLUTION



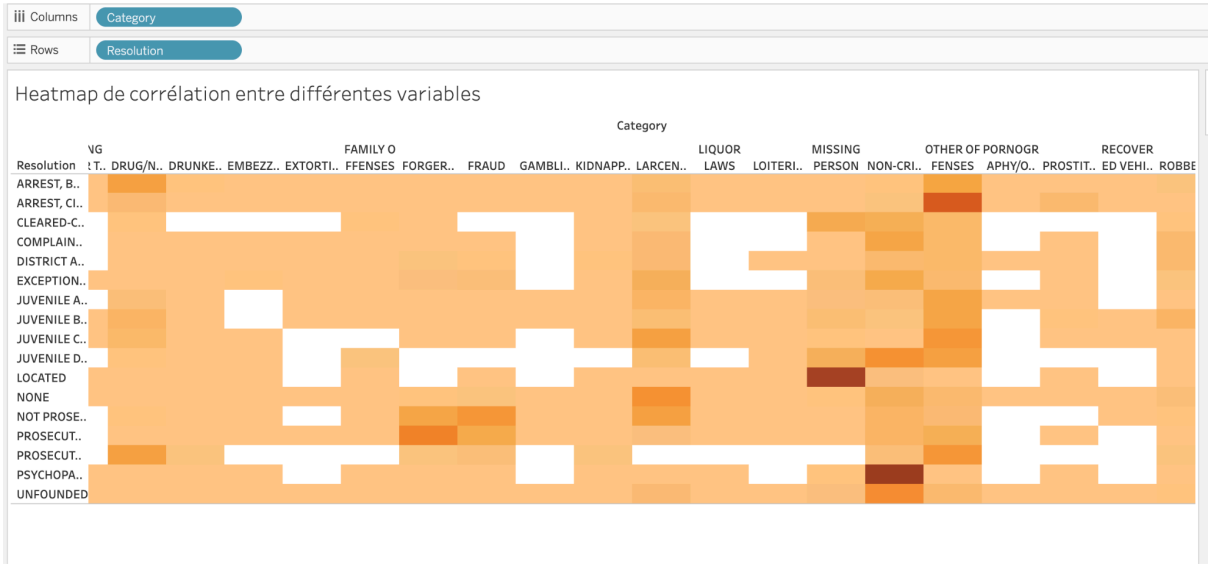
On utilise des heatmap pour voir les corrélations entre variables. Par exemple, cette heatmap nous indique que les résolutions ARREST, BOOKED ; se font le plus dans le district

southern et tenderloin. Dans le district southern, la résolution la plus courante est None (donc rien n'est résolu, pas d'arrestation ou autre).



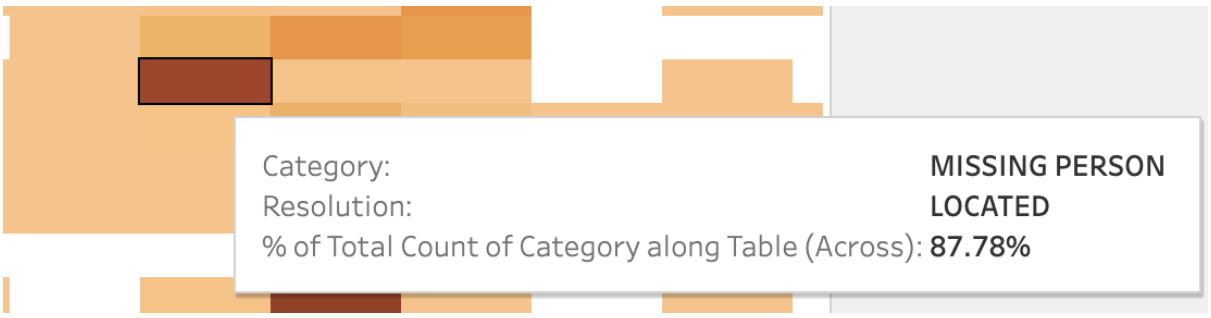
On voit que la heatmap change sensiblement si on calcul les pourcentages plutôt que les valeurs absolues, mais les corrélations qui ressortent restent les mêmes.

Category vs Resolution

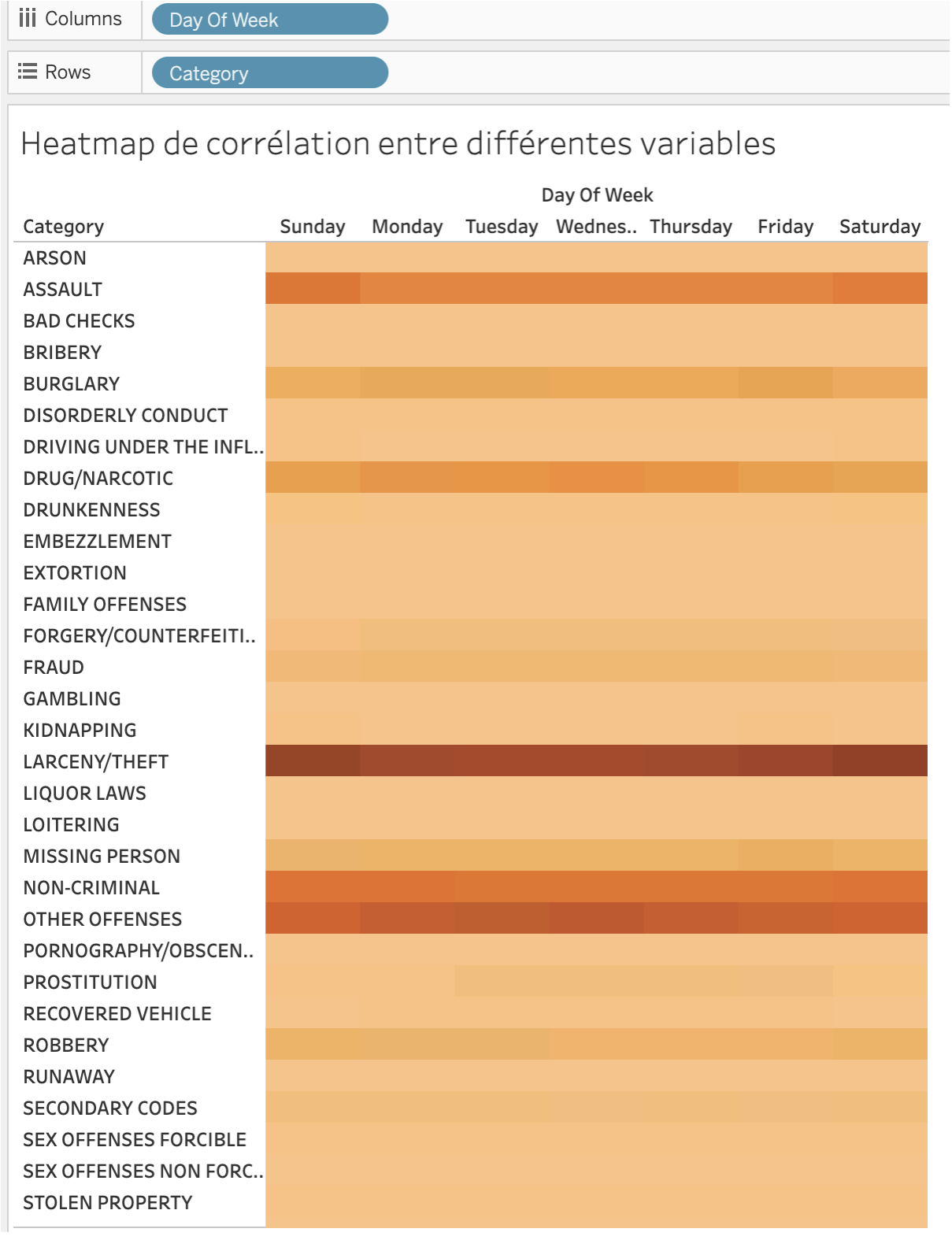


On ne peut pas mettre toutes les données, mais voilà un extrait de la corrélation entre les variables category et resolution. On a choisi la couleur orange en appuyant sur le gradient à droite dans Tableau, pour rester dans le thème de la criminalité.

Par exemple, un fait intéressant est que 88% des cas de personnes disparues finissent par être résolus (on retrouve la personne).


Category: MISSING PERSON
Resolution: LOCATED
% of Total Count of Category along Table (Across): 87.78%

Category vs Days



Et finalement, on a les catégories de crimes en fonction du jour de la semaine. La lecture est similaire à celle des tableaux précédents, l'idée est la même : essayer de trouver des corrélations entre variables.

• Identifiez les zones géographiques présentant les plus fortes criminalités à San Francisco

o Cette visualisation s'appuie sur les coordonnées géographiques X et Y qui nécessitent une opération de conversion pour être exploitable

Nous avons déjà converti les données de dimension.

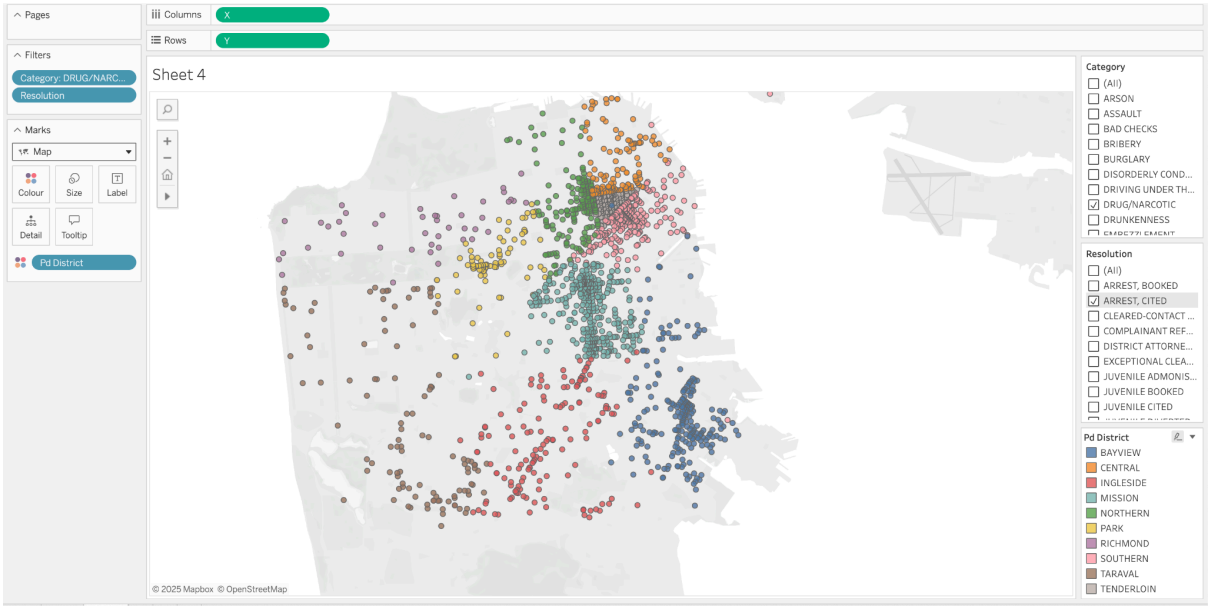
Pour cette visualisation, on utilise la densité.

Plus la zone est rouge foncé, plus la concentration de crimes est importante dans cette zone.

On voit que la criminalité se concentre plutôt dans le nord-est.

3. Tableau de bord

Mettre en place un tableau de bord combinant une visualisation géographique et les visualisations impliquant les catégories des crimes, les résolutions des crimes et les districts. Exploiter les filtres pour analyser la distribution géographique des crimes : district par district, distribution des catégories de crimes, distribution de la résolution de crimes, etc.



Passons les éléments de ce tableau point par point.

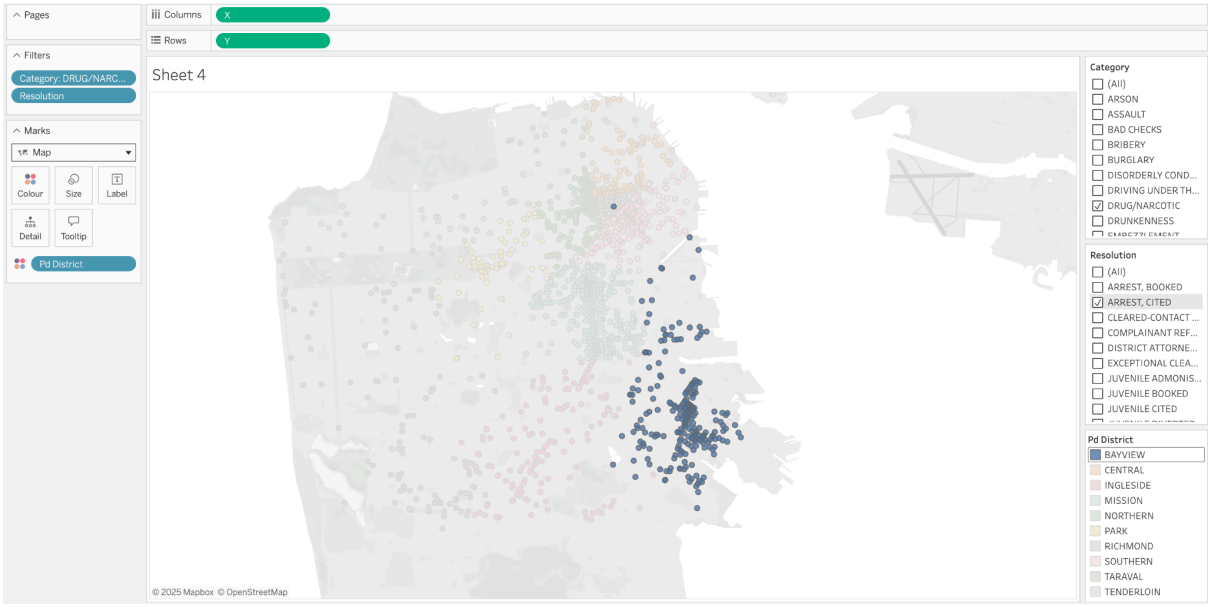
Tout d'abord, on a bien renseigné les coordonnées géographiques (en haut : X pour les colonnes ou longitude ; Y pour les lignes ou latitude).

On met Pd District dans le champ Marks (à gauche). Ainsi, on affiche bien chaque point, qui correspond à une ligne (un crime donc). On choisit les couleurs pour représenter les districts. La légende se trouve à droite.

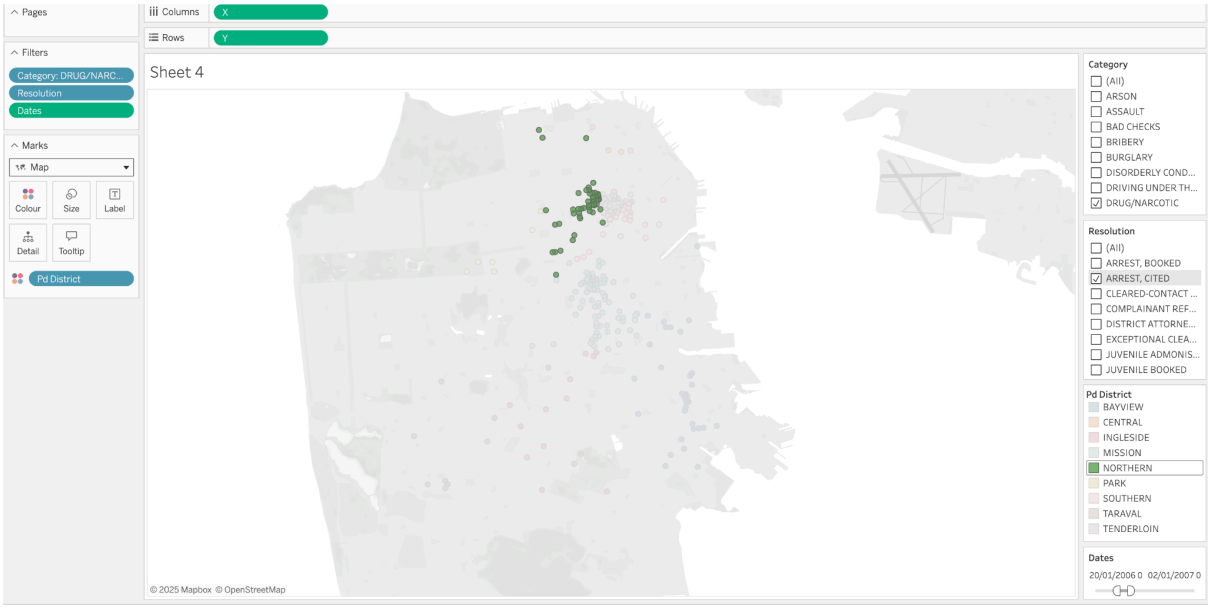
Pour les filtres : on met les filtres pour la catégorie du crime et la résolution (en haut à gauche). On choisit d'afficher ces filtres sur le tableau de bord (on peut les voir directement à droite, pas besoin d'appuyer à chaque fois pour effectuer un nouveau filtrage).

Ainsi, on peut faire les combinaisons de filtres que l'on veut et visualiser sur la carte tous ceux qui correspondent à nos filtres et qui se sont produits sur la période étudiée.

Par exemple, ici on a les crimes associés aux drogues/narcotiques, qui se sont résolus sur une arrestation (cited). Tous districts confondus.



Pour finir, si on souhaite filtrer un district particulier, il nous suffit d'appuyer sur le district correspondant dans la légende, pour mettre les crimes associés en évidence.



En conclusion, on peut ajouter un filtre pour la date (en sélectionnant plage de temps Range of dates), ce qui nous permet de visualiser les crimes d'une période uniquement. Par exemple ici, on affiche les crimes de l'année 2006 (voir filtre tout en bas à droite) qui correspondent à nos autres filtres aussi.

THE END