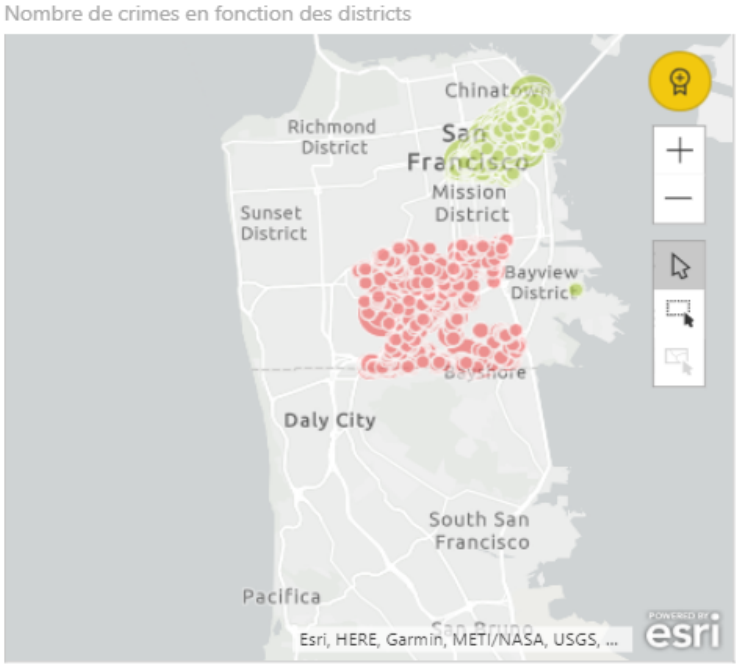
TD - FORMATION POWER BI AVANCÉE

Exercice 1 : la criminalité à San Francisco

La police de San Francisco vous a recruté en tant que Data Analyst pour réduire la criminalité. On vous demande d'analyser la criminalité et d'en ressortir les éléments clés afin de rendre plus efficientes les interventions de police.

1. Ouvrez le fichier BI_formation.pbix contenant déjà toutes les tables.
   (a) Quel est le modèle d'entrepôt de données qui a été utilisé ? Pourquoi ?
   (b) Quelles sont le(s) dimension(s) ? Le(s) fait(s) ?
   (c) Quelles sont les dimensions intéressantes pour analyser la criminalité à San Francisco ?
2. Créez un graphique des différents crimes.
   (a) Citez les trois catégories de crimes les plus courants à San Francisco.
3. Réalisez un tableau des catégories de crimes en fonction des districts. *Indication : Utilisez Matrix (Matrice).*
   (a) Après avoir identifié les 3 catégories de crimes les plus présentes, identifiez le district le plus touché par la criminalité.
   (b) Réalisez une carte Arcgis pour situer ces secteurs. Faites varier la taille des cercles en fonction du nombre de crimes. Faites varier la couleur des ronds en fonction du district, puis filtrez sur les 3 districts les plus touchés par la criminalité.



(c) Ajoutez au tableau des pourcentages en fonction des types de crimes par district. Cela confirme-t-il certaines de vos suppositions concernant la question précédente ? *Indication : pourcentage ligne ou colonne selon la disposition de votre matrice.*

(d) Où conseillez-vous d'envoyer les patrouilles de police ?

(e) A l'aide d'une mise en forme conditionnelle, vous souhaitez identifier rapidement les cellules du tableau ayant plus de 600 crimes (en rouge) et celles ayant moins de 30 crimes (en vert).

4. Analysez les catégories de crimes en fonction des années à l'aide d'un treemap :

   (a) Ajoutez le pourcentage de crimes dans les info-bulles. Que constatez-vous ?

5. Vous souhaitez désormais affiner votre analyse en étudiant l'évolution du nombre de crimes par mois. *Indication : utilisez un histogramme + info-bulles. Pour la variation mensuelle, il faut une mesure rapide : Variation d'un mois à l'autre.*

   (a) Actuellement, il est impossible de réaliser le calcul d'évolution, pourquoi ?

   (b) À partir de la clef de la dimension date (format : aaaammjj), créez en utilisant le DAX un libellé nommé Lib_date qui représentera une date. *Indication : Dans l'onglet Données, sélectionnez la table date, puis cliquez sur "Nouvelle colonne" et entrez votre formule DAX dans la barre de formule. Utilisez la fonction DATE, LEFT, MID et RIGHT. .*

   (c) Quelle est la différence entre les deux dates ?

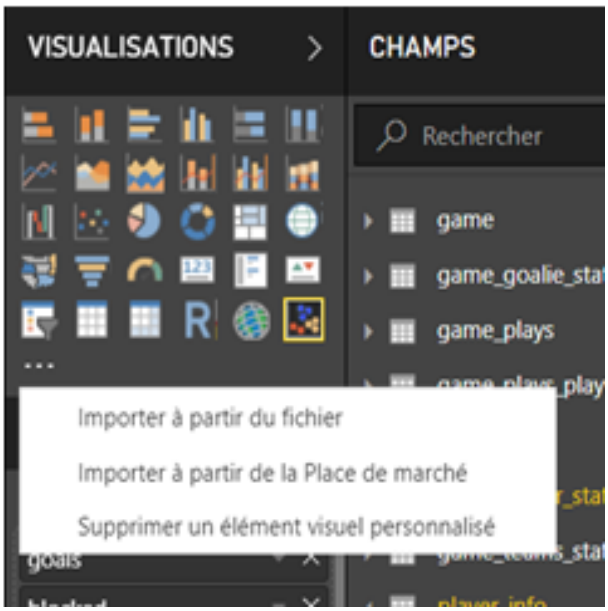
(d) En utilisant la colonne lib_date, réalisez le graphique et interprétez-le. *Indication : Utilisez l'info-bulle pour donner l'évolution mensuelle. .*

(e) Quel est le mois où il y a eu la plus grosse baisse de criminalité ?

6. Après avoir étudié les crimes en fonction de chacune des dimensions - temps, localisation, catégories - vous vous intéressez aux descriptions fournies par les policiers.

(a) Téléchargez puis installez un nouveau visuel : Word Cloud.
Indications :

- *Connectez-vous sur votre compte Miscrosoft.*
- *Téléchargez/Importez un visuel à partir de la place du marché puis importez nuage de mots.*



- *Après le téléchargement, Power BI importera les packages R si besoins.*
- *N'oubliez pas d'activer les scripts R.*

(b) Réalisez un nuage de mots à partir des descriptions des crimes. *Indication : Pour créer votre nuage de mots , lorsque le Word Cloud est sélectionné :*

- *Catégorie = unique.crime.Description*
- *Valeurs = Nombre de unique.crime.Description*

(c) Quels sont les mots qui apparaissent le plus et pourquoi ?

(d) Dans les options (Format) : activez Mot Vide. Quel changement observez-vous dans le graphique ? Faut-il activer ou désactiver cette option ? Votre analyse précédente est-elle toujours pertinente ? En mettant en relation, avec les informations précédentes, cela vous paraît-il cohérent ?

7. Vous voulez savoir dans quels types de voies on a le plus de délits commis :

   (a) À l'aide du langage DAX, créez une nouvelle colonne contenant les types de voies dans la dimension "adresse". Prenez les 2 derniers éléments de l'adresse dans lib_adresse. Par exemple, dans l'adresse : "0 Block of 27th ST", on veut ST pour Street. *Indication : utilisez la fonction RIGHT.*

   (b) Quels sont les 3 types de voies dans lesquels le plus de délits sont commis ? Y a t-il des jours de la semaine où vous constatez une hausse des délits ?

   (c) La brigade des stupéfiants vous demande de lui indiquer quel jour de la semaine le trafic de drogue (drug narcotic) est le plus élevé afin d'intervenir.

8. Au final, quels sont les principaux conseils que vous donnez à la police ?

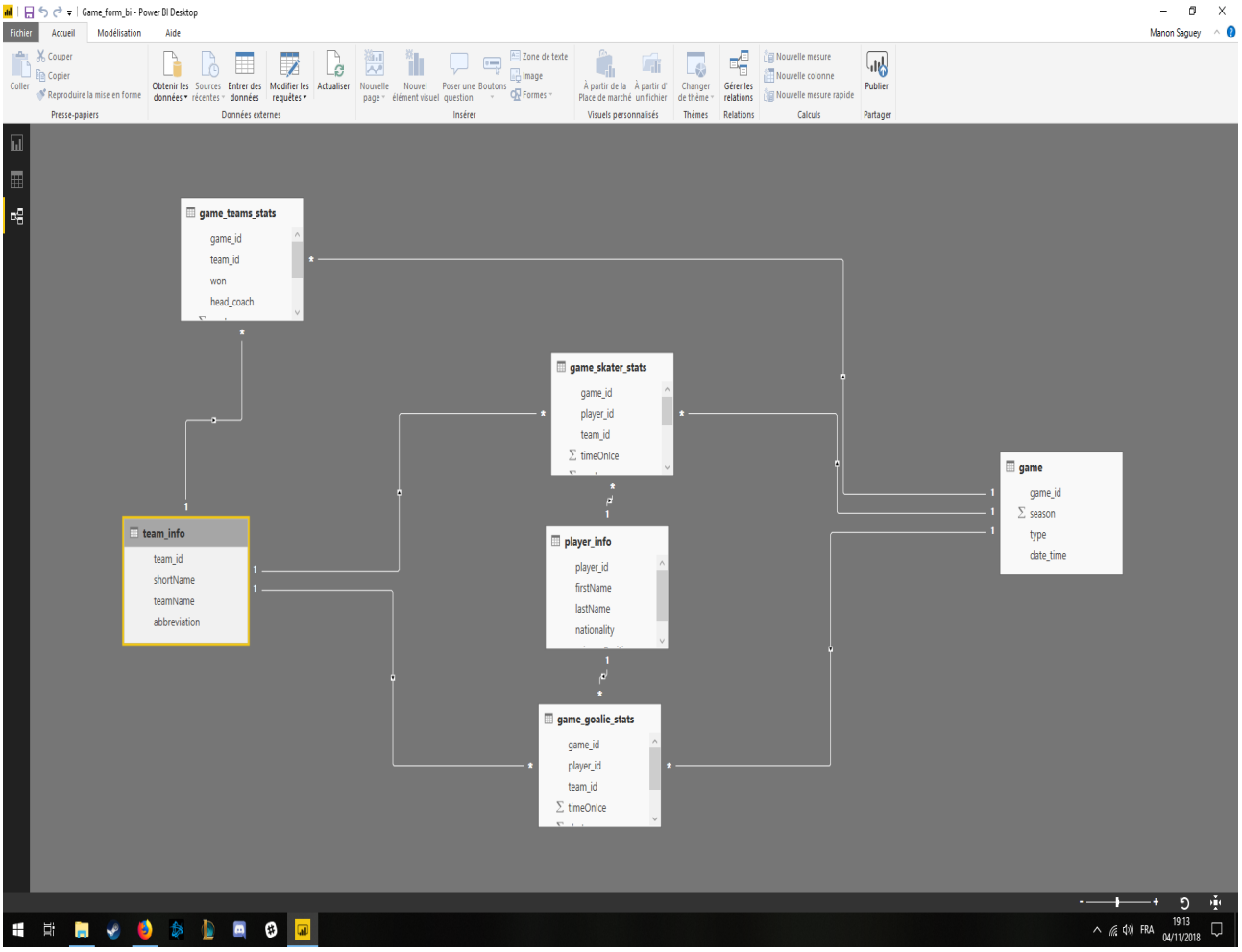
Exercice 2 : NHL Game Data - le championnat de hockey sur glace

À présent, vous décidez de devenir consultant dans la data science. Vous signez un contrat d'exclusivité avec le magazine "Hockey News" et en parallèle vous devenez conseiller pour différents clubs. Votre objectif : analyser les résultats des équipes, des joueurs et des entraîneurs.

1. Chargez le jeu de données NHL Game Data :

   (a) Ouvrez le projet Game_form_pbix.

   (b) Vérifiez que les relations correspondent à celles sur la photo ci-dessous.



2. Construisez un graphique qui affiche le pourcentage de match de championnats et de play offs en fonction de l'année, du trimestre, du mois et du jour. *Indication : commencez par créer une colonne qui renomme P en Play Off et R en Regular Season (table : game). Pour cela utilisez la fonction IF.*

   (a) Sur quelle période se déroulent les playoffs ?

3. Vous désirez identifier et étudier les performances des meilleurs joueurs :

   (a) Identifiez les 5 joueurs qui ont marqué le plus de but. On veut les identifier par leur nom et leur prénom. *Indication : créez une nouvelle colonne regroupant le nom et le prénom à l'aide du langage DAX. Utilisez le & pour concaténer, puis représentez les 5 meilleurs buteurs dans un tableau, en utilisant le champ goals de la table game_skaters_stats.*

   (b) De quelle nationalité sont-ils ?

   (c) Quelle est la nationalité la plus représentée ?

(d) Dans le championnat de hockey, il existe une règle contraignant les sélectionneurs à recruter au maximum 4 joueurs extra communautaires par équipe. Considérant cette règle, peut-on considérer qu'il y a assez d'Américains dans la ligue professionnelle de hockey ?

(e) À l'aide d'une formule DAX, construisez un graphique permettant d'afficher le pourcentage d'Américains et d'étrangers dans le championnat NHL. *Indication : Dans la table player_info, créez une colonne dans laquelle on retrouvera les nationalités USA et les autres regroupées sous le libellé "autres". Utilisez la fonction IF.*

4. Le Président du club de Colorado, dont les résultats ont été mauvais en 2017-2018 (seulement 48 points) souhaite changer d'entraîneur.

   (a) Pour l'aider à à recruter un nouvel entraîneur, vous sélectionnez les 3 entraîneurs qui ont eu le plus de victoires avec leurs équipes respectives.
   Indication : pensez à reformuler la colonne won de la table game_teams_stats à l'aide d'une formule DAX, en une nouvelle colonne TRUE=1 sinon 0, utilisez la fonction IF.

5. Un journaliste sportif a besoin d'une cartographie du nombre de victoires, mais vous ne disposez pas des coordonnées GPS. *Indication : vous pouvez utiliser le nom des équipes car ce sont le nom des villes où se situent les clubs.*

   (a) Pour optimiser la visualisation des données, réalisez des cercles proportionnels au nombre de victoires et un dégradé de couleurs en fonction du nombre de buts. *Indication : utilisez le champ shortName de la table team_info pour l'emplacement.*

   (b) Quelle est la côte américaine la plus représentée ?

   (c) Le journaliste souhaite compléter son article par un focus sur les équipes les plus fortes et celles qui marquent le plus de buts.

   (d) De même, le journaliste souhaite mentionner l'équipe qui a gagné le moins de matchs. Vous justifiez ses propos en analysant les enregistrements de cette équipe.
   Indication : Les informations des questions b) ; c) et d) sont disponibles directement à partir de la carte.

6. Vous souhaitez maintenant compléter votre analyse en étudiant non seulement le nombre de matchs gagnés mais également le ratio de Victoires.

   (a) Tout d'abord, créez une mesure %victoire.
   Indications :

   - *Pour simplifier, transformez la colonne won (boolean) en texte ;*
   - *utilisez les fonctions : CALCULATE, FILTER, DIVIDE, COUNTROWS*

   (b) Réalisez un nuage de points du nombre de match en fonction du pourcentage de victoires.

Indication : Utilisez le champ shortName en détails et le champ won (Nombre de valeurs) en valeurs x, ajoutez un dégradé de couleurs selon le pourcentage de victoires pour différencier les différents groupes.

(c) Quels sont les différents profils des équipes ? Citez un ou plusieurs exemples.

(d) Vous attendiez-vous à ces résultats concernant certaines équipes ?

7. Pour peaufiner votre analyse, vous décidez de réaliser un graphique plus adapté qui permettra d'identifier les équipes qui ont eu une importante augmentation ou diminution de leur pourcentage de victoires d'une année sur l'autre. *Indication : faites un graphique en cascade. Utilisez le champ date_time de la table game pour les catégories et shortName de la table team_info pour la répartition.*

   (a) Quelle est l'année où Phoenix a eu sa plus grande baisse du pourcentage de victoires ?

   (b) Quelle est l'année où Winnipeg a eu sa plus grande augmentation du pourcentage de Victoire ?

8. Pour décerner les trophées aux meilleurs joueurs, vous souhaitez regrouper les joueurs dans des catégories (clusters). Nous décernons 2 trophées : le trophée James Norris (meilleur défenseur) et le Trophée Frank J. Selke (meilleur attaquant).

   (a) Dans votre sélection, vous prenez en compte le nombre de buts marqués et le nombre de but sauvés. *Indications :*

   - *Importez un visuel à partir de la place du marché puis importez Clustering. Après le téléchargement, Power BI importera les packages R si besoins. N'oubliez pas d'activer les scripts R.*
   - *Dans la visualisation clustering :*
     - *values = les axes factoriels (blocked et goals de la table game_skaters_stats) ;*
     - *Data point labels = les points à afficher (le nom du joueur ici) ;*
     - *Tooltips = info bulle (champ primaryPosition de la table player_info ;*
     - *ID = l'axe d'analyse (l'ID du joueur dans notre cas).*

   (b) Combien de clusters ont été détectés automatiquement par Power BI ? Qu'en pensez-vous ? Changez le nombre de clusters et gardez le clustering qui vous semble le plus pertinent. Comment interprétez-vous ces clusters ?

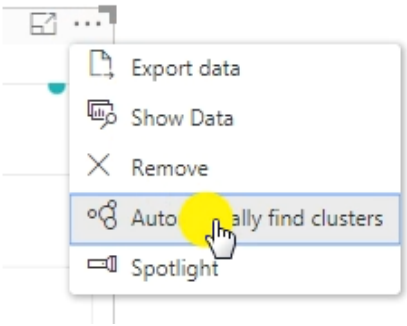
   (c) Donnez une étiquette à ces clusters.

   (d) Que pensez-vous des joueurs Russel et Oveckin ? Sont-ils en adéquation avec l'étiquette de leurs clusters ? Pensez-vous qu'ils méritent le(s) trophé(s) ?

   (e) Quel est le coefficient de corrélation entre le nombre de buts marqués (Goals) et le nombres de buts sauvés (Blocked) ? Qu'en déduisez-vous concernant la remise de trophés ? *Indication : utilisez une mesure rapide. (Coefficient de corrélation).*

9. Après les 2 trophés principaux, un autre prix récompense le joueur le plus précis devant les buts. Réalisez un clustering des joueurs en les regroupant selon le nombre de buts marqués et le nombre de tirs cadrés (champ shots de la table game_skaters_shots).

   Indication : Utilisez le nuage de point classique et cliquez sur les 3 petits points en haut à droite du graphique sur Trouver des clusters automatiques.



   (a) Combien de clusters ont été automatiquement détectés ? Qu'en pensez-vous ? Changez le nombre de clusters et gardez le clustering qui vous semble le plus pertinent. Combien de joueurs chaque cluster contient-il ?
   Indication : utilisez l'option Cluster présente dans le nuage de mots comme indiqué ci-dessus.

   (b) Lors des penalties, certains joueurs prennent plus de temps que d'autres avant de tirer. Vous souhaitez savoir si le fait de prendre son temps pour tirer augmente les chances de marquer un but. Dans chaque cluster, combien de minutes perdues lors des penalties ont été enregistrées ?
   Indication : utilisez un tableau avec le champ penaltyMinute de la table game_skaters_stats et le nom complet des joueurs.

10. Pour conclure son article, le journaliste vous a demandé d'estimer le nombre de buts des prochains matchs. Pour cela, vous allez réaliser une régression linéaire afin de prédire le nombre de buts dans un match en fonction du nombre de tirs cadrés :

    (a) Téléchargez la table regression.csv en utilisant R Visual.
    Indications :
    - *Allez dans obtenir les données Puis Autres*
    - *utilisez read.csv. Nommez la table "regression_table" et examinez les champs.*

    (b) Calculez en tant que mesures : le coefficient de corrélation, l'intersection ainsi que le coefficient de régression. Affichez à l'aide de R visual la droite de régression des buts marqués en fonction des tirs cadrés.
    Rappel :
    Coefficient de corrélation :

$$\rho = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{(n(\sum X^2) - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

Équation de la régression :

$$Y = aX + b$$

$$a = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$b = \frac{n(\sum Y)(\sum X)^2 - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2}$$

Exercice supplémentaire (manipulation du R script visual)

Vous avez fini votre travail ! Alors que vous vous apprêtiez à profiter d'un bon moment de détente, votre ami botaniste souhaite analyser les données qu'il a recueillies sur une nouvelle espèce de plante.

1. Importer de la base de données IRIS depuis un Rscript.
   - Téléchargez la BDD iris se trouvant dans l'environnement R Studio.
     Indications :
     - *Allez dans obtenir les données Puis Autres*
     - *Utilisez le code R pour récupérer la base IRIS de R*
   - Nommez la "BDD_IRIS".
2. Avant de commencer vos calculs, vous voulez visualiser vos données avec un nuage de points
   (a) Prenez les variables Petal.length et Petal.Width.
      i. Que constatez-vous ?
      ii. Comment pouvez-vous remédier à ce problème ?
3. Après avoir résolu ce problème, votre ami souhaite regrouper les iris qui se ressemblent le plus dans différents clusters.
   (a) Dérouler l'algorithme du Kmeans.
      i. Quel est le nombre de clusters qui vous semble le plus pertinent ?
      ii. En utilisant R script visual, construisez le nuage de points et colorez les points selon leur cluster. *Indication : utilisez (plot(...)).*

Dictionnaire de données de la base crimes

Table : fait_crime. Cette table est la table de fait du modèle. La mesure est le nombre de crimes : NB_crimes.

Champ	Type	Description
ID_delit	numeric	Identifiant de la description de l'incident. Clé secondaire reliant la table fait_crime à la table dim_incident.
ID_cat	numeric	Identifiant de la catégorie de crime. Clé secondaire reliant la table fait_crime à la table dim_cat.
ID_date	numeric	Identifiant de la date du crime. Clé secondaire reliant la table fait_crime à la table dim_date.
ID_time	numeric	Identifiant de l'heure du crime. Clé secondaire reliant la table fait_crime à la table dim_time.
ID_adresse	numeric	Identifiant de l'adresse du crime. Clé secondaire reliant la table fait_crime à la table dim_adresse.
ID_coordonnees	text	Identifiant des coordonnées GPS du lieu du crime. Clé secondaire reliant la table fait_crime à la table dim_gps.
NB_crimes	numeric	Mesure du nombre de crimes.

Table : dim_adresse

Champ	Type	Description
id_adresse	numeric	Clé primaire. Identifiant de l'adresse du lieu du crime.
lib_adresse	text	Adresse du lieu du crime.
id_district	numeric	Identifiant du district dans

		lequel a eu lieu le crime. Clé secondaire reliant la table dim_adresse à la table dim_district.

Table : dim_annee

Champ	Type	Description
id_annee	numeric	Clé primaire. Identifiant de l'année du crime.

Table : dim_cat

Champ	Type	Description
id_cat	numeric	Clé primaire. Identifiant du type de crime commis.
lib_cat	text	Type de crime commis.

Les différentes modalités de lib_cat :
OTHER OFFENSES : petits délits.
ROBBERY : vol directement commis sur les personnes en utilisant la peur ou la violence.
FRAUD : fraudes, escroqueries.
LARCENY / THEFT : vols commis sans violences.
VANDALISM : actes de vandalismes.
DRUG / NARCOTIC : crimes relatifs aux drogues et stupéfiants.
ASSAULT : coups et blessures.
FORGERY / COUNTERFEITING : contrefaçons.
NON-CRIMINAL : infractions n'impliquant pas de crimes
TRESPASS : introduction dans une propriété privée.
VEHICLE THEFT : vol de véhicule.
DRIVING UNDER THE INFLUENCE : conduite en état d'ivresse ou sous drogue.
BURGLARY : introduction dans une propriété dans le but d'y commettre un vol.
STOLEN PROPERTY : possession d'un bien volé.
DISORDERLY CONDUCT : comportement perturbateur.
SEX OFFENSES; FORCIBLE : agressions sexuelles.
ARSON : incendie criminel.
MISSING PERSON : personne portée disparue.
WEAPON LAWS : port d'armes interdit
EXTORTION : contraindre une personne à donner de l'argent ou un bien lui appartenant ou à effectuer un service.
KIDNAPPING : enlever une personne.
DRUNKENNESS : état d'ivresse sur la voie publique.
BAD CHECKS : utilisation de chèques en blanc.
LIQUOR LAWS : violation d'une loi concernant la fabrication, le transport, la vente, l'achat, la possession ou l'utilisation d'alcool.
SUICIDE : suicides.

FAMILY OFFENSES : crime commis à l'encontre d'un membre de la famille ou d'une personne intimement proche.
PROSTITUTION : prostitution.
EMBEZZLEMENT : détournement d'argent ou de biens.
BRIBERY : corruption.
LOITERING : fait de rester un certain moment dans un endroit public sans but précis.
RUNAWAY : délit de fuite.
GAMBLING : crimes reliés aux jeux.
SEX OFFENSES ; NON FORCIBLE : agression sexuelle sur une personne incapable de donner son consentement (mineur, handicapé...).

Table : dim_date

Champ	Type	Description
date_ID	numeric	Clé primaire. Identifiant de la date du crime.
JourMois	numeric	Numéro du jour du mois du crime. (Modalités comprises entre 1 et 31)
id_jour	numeric	Identifiant du jour du crime. Clé secondaire reliant la table dim_date à la table dim_jour. (Modalités comprises entre 1 et 7).
mois_id	numeric	Identifiant du mois du crime. Clé secondaire reliant la table dim_date à la table dim_mois.

Table : dim_district

Champ	Type	Description
c.1.length.unique.dim_adresse.id_district...	numeric	Clé primaire. Identifiant du district.
lib_district	text	Nom du district.

Table : dim_gps

Champ	Type	Description
id_gps	text	Clé primaire. Identifiant des coordonnées gps.
longitude	float	Coordonnée de la longitude.

latitude	float	Coordonnée de la latitude

Table : dim_incident

Champ	Type	Description
c	numeric	Clé primaire. Identifiant de la description du crime.
unique.crime.Descript.	text	Description du crime.

Table : dim_jour

Champ	Type	Description
id_jour	numeric	Clé primaire. Identifiant du jour.
id_lib	text	Libellé du jour.

Table : dim_mois

Champ	Type	Description
id_mois	numeric	Clé primaire. Identifiant du mois.
lib_mois	text	Libellé du mois.
lib_annee	numeric	Libellé de l'année.

Table : dim_temps

Champ	Type	Description
id_time	numeric	Clé primaire. Identifiant de l'horaire.
lib_time	time	Heure complète.
temp	text	Libellé du moment (2 modalités : matin / après-midi).

Dictionnaire des données de la base NHL Game Data

Table : game. Cette table donne des informations sur la date du match et le type de match joué.

Nom	Type	Description
game_id	numeric	Identifiant du match.
type	text	Type du match (2 modalités : championnat / play off).
date_time	date	Date du match.

Table : game_goalie_stat. Cette table donne des informations sur les statistiques des matchs.

Nom	Type	Description
game_id	numeric	Identifiant du match.
player_id	numeric	Identifiant des joueurs.
team_id	numeric	Identifiant de l'équipe.
timeOnIce	numeric	Le temps que le joueur a passé sur la glace durant le match (en secondes).
shots	numeric	Nombre de tirs.
saves	numeric	Nombre de buts sauvés.

Table : game_skater_stats. Cette table permet d'avoir des informations sur les statistiques des joueurs durant les matchs.

Nom	Type	Description
play_id	numeric	Identifiant d'une action de jeu
game_id	numeric	identifiant d'un match
team_id	numeric	Identifiant de l'équipe
timeOnIce	Numeric	Le temps que le joueur a passé sur la glace durant le match (en secondes).

goals	numeric	Nombre de but marqués pour un match
shots	numeric	Nombre de tirs
penaltyMinutes	numeric	Nombre de minutes en infériorité numérique
blocked	numeric	Nombre de tirs adverses bloqués.

Table : game_teams_stats. Cette table donne des informations sur les statistiques des équipes.

Nom	Type	Description
play_id	numeric	Identifiant d'une action de jeu
game_id	numeric	identifiant d'un match
won	boulean	Victoire ou défaite
head_coach	text	Nom de l'entraineur
goals	numéric	Nombre de buts marqués
shots	numéric	Nombre de tirs tentés

Table : player_info. Cette table donne des informations sur les joueurs.

Nom	Type	Description
player_id	numeric	Identifiant du joueur
firstName	text	Prénom du joueur
lastName	text	Nom de famille du joueur
nationality	text	Nationalité du joueur
primaryPosition	text	Poste du joueur

Table : team_info. Cette table recense les informations des équipes participant au tournoi.

Nom	Type	Description
team_id	numeric	Identifiant de l'équipe
shortName	text	Ville où se situe l'équipe
teamName	text	Nom de l'équipe
abbreviation	text	Abréviation du nom de

		l’équipe