Rapport P4 RGPD

Introduction : Contexte et objectifs du projet

À la suite d'une plainte d'un client concernant une utilisation non autorisée de ses données personnelles, notre entreprise a été sanctionnée par la CNIL d'une limitation temporaire des traitements pour non-conformité au RGPD. Cette situation, critique pour nos activités commerciales, met en lumière des lacunes dans la gestion des données personnelles au sein de notre CRM. Afin de lever cette sanction, il est impératif de mettre en œuvre des mesures correctives immédiates, documentées et conformes aux exigences du RGPD.

Ce rapport s'inscrit dans ce cadre et vise à répondre à trois objectifs principaux :

1. **Identifier les violations et mettre en conformité la gestion des données CRM :** des recommandations claires sur les règles de gestion seront définies pour garantir le respect du RGPD.

2. **Réaliser une extraction anonymisée des données :** les données doivent être nettoyées et anonymisées afin d'être conformes à la réglementation tout en restant exploitables pour l'analyse commerciale.

3. **Documenter le processus complet :** l'ensemble des étapes d'anonymisation, de la collecte initiale à la préparation des données dans l'ETL, sera présenté de manière détaillée pour garantir la transparence et la reproductibilité du processus.

En respectant les principes fondamentaux du RGPD (minimisation des données, limitation de conservation et pseudonymisation), ce travail permettra de répondre

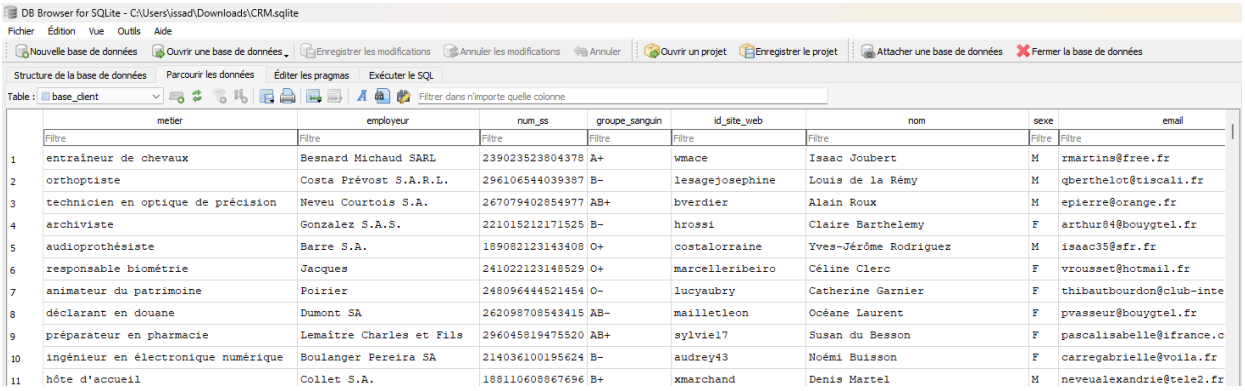
aux exigences de la CNIL et d'assurer une gestion éthique et sécurisée des données personnelles.

1. Importation et récupération des données

Avant d'effectuer les transformations et d'anonymisation dans Power Query, les données ont été préparées en utilisant SQLite DB Browser. Cette étape consistait à extraire uniquement les dossiers correspondant aux critères de la consigne.

Étape 1 : Importation de la base de données SQL dans SQLite DB Browser

- La base de données brute a été importée dans SQLite DB Browser.
- Les données comprenaient des informations sur tous les clients, sans filtrage préalable.



	metier	employeur	num_ss	groupe_sanguin	id_site_web	nom	sexe	email
1	entraîneur de chevaux	Besnard Michaud SARL	239023523804378	A+	wmace	Isaac Joubert	M	xmartins@free.fr
2	orthoptiste	Costa Prévost S.A.R.L.	296106544039387	B-	lesagejosephine	Louis de la Rémy	M	qberthelot@tiscali.fr
3	technicien en optique de précision	Neveu Courtois S.A.	267079402854977	AB+	bverdier	Alain Roux	M	epierre@orange.fr
4	archiviste	Gonzalez S.A.S.	221015212171525	B-	hrossi	Claire Barthelemy	F	arthur84@bouygtel.fr
5	audioprothésiste	Barre S.A.	189082123143408	O+	costalorraine	Yves-Jérôme Rodriguez	M	isaac35@sfr.fr
6	responsable biométrie	Jacques	241022123148529	O+	marcelleribeiro	Céline Clerc	F	vrousset@hotmail.fr
7	animateur du patrimoine	Poirier	248096444521454	O-	lucyaubry	Catherine Garnier	F	thibautbourdon@club-inte
8	déclarant en douane	Dumont SA	262098708543415	AB-	mailletleon	Océane Laurent	F	pvasseur@bouygtel.fr
9	préparateur en pharmacie	Lemaître Charles et Fils	296045819475520	AB+	sylvie17	Susan du Besson	F	pascalisabelle@ifrance.c
10	ingénieur en électronique numérique	Boulanger Pereira SA	214036100195624	B-	audrey43	Noémi Buisson	F	carregabrielle@voila.fr
11	hôte d'accueil	Collet S.A.	188110608867696	B+	xmarchand	Denis Martel	M	neveualexandrie@tele2.fr

Étape 2 : Filtrage des données pour répondre aux critères

Le but de cette étape était d'extraire uniquement les dossiers :

1. Dont l'état est marqué comme "complet".
2. Pour lesquels la demande d'assurance a été effectuée en 2022.

Code SQL utilisé :

```
SELECT *
FROM table_client
WHERE etat_dossier = 'complet'
  AND strftime('%Y', date_demande) = '2022';
```

Étape 3 : Export des données filtrées au format CSV

Une fois les données filtrées, elles ont été exportées au format CSV pour être utilisées dans Power Query.

Étape 4 : Importation dans Power Query

Le fichier CSV exporté (Bardet_Flavien_2_donnees_part1_112024) a été chargé dans Power Query pour poursuivre les opérations d'anonymisation et de transformation des données.

- Action réalisée : Dans Power Query : Données > Obtenir des données > À partir d'un fichier > À partir d'un fichier texte/CSV.

2. Opérations ETL

L'analyse de l'ancien dictionnaire des données révèle les colonnes suivantes. Elles ont été examinées en détail pour évaluer leur caractère sensible, leur utilité, et les transformations nécessaires conformément au RGPD.

	A	B	C	D
1	Nom des colonnes	Type	Modifications prévues	Raisons
2	metier	string	Conserver	Données possiblement utile pour traitement stat
3	employeur	string	Supprimer	Données personnelles : violation de la RGPD
4	num_ss	string	Supprimer	Données personnelles : violation de la RGPD
5	groupe_sanguin	string	Supprimer	Données personnelles : violation de la RGPD
6	id_site_web	string	Supprimer	Données personnelles : violation de la RGPD
7	nom	string	Supprimer	Données personnelles : violation de la RGPD
8	sexe	string	Conserver	Données possiblement utile pour traitement stat
9	email	string	Supprimer	Données personnelles : violation de la RGPD
10	date_naissance	string	Transformer	Données à anonymiser pour traitement stat
11	id_client	string	Remplacer	Créer nouvel index + système d'idclient anonyme
12	enfant_conduite_accompagne	string	Conserver	Données possiblement utile pour traitement stat
13	nombre_enfants	string	Conserver	Données possiblement utile pour traitement stat
14	revenus	string	Transformer	Données à anonymiser pour traitement stat
15	valeur_residence_prin	string	Supprimer	Données personnelles : violation de la RGPD
16	formation	string	Conserver	Données possiblement utile pour traitement stat
17	usage_vehicule	string	Conserver	Données possiblement utile pour traitement stat
18	type_vehicule	string	Conserver	Données possiblement utile pour traitement stat
19	est_rouge	string	Transformer	Modification du type de données pour faciliter traitement stat
20	points_perdus	string	Conserver	Données possiblement utile pour traitement stat
21	age_vehicule	string	Transformer	Données importantes, modification du format vers nombres entiers
22	type_conduite	string	Conserver	Données possiblement utile pour traitement stat
23	date_demande	string	Transformer + Filtrer	Données importantes, simplification partielle vers format mois+date
24	etat_dossier	string	Transformer + Filtrer	Données importantes pour traitement stat
25	formule	string	Conserver	Données importantes pour traitement stat
26	tarif_devis	string	Conserver	Données importantes pour traitement stat
27	adresse	string	Supprimer	Données personnelles : violation de la RGPD
28	lat	string	Supprimer	Données personnelles : violation de la RGPD
29	lon	string	Supprimer	Données personnelles : violation de la RGPD
30				

Synthèse des Modifications

1. Colonnes supprimées :
    - employeur, num_ss, groupe_sanguin, id_site_web, nom, email, valeur_residence_prin, adresse, lat, lon.
    - Ces colonnes sont directement ou indirectement identifiables et ne sont pas nécessaires aux objectifs d'analyse.
2. Colonnes transformées :
    - id_client : Remplacé par un nouveau système d'id basé sur l'index pour garantir l'irréversibilité.
    - date_naissance : Réduction à l'année uniquement.
    - revenus : Catégorisation en intervalles ([0, 20k€], [20k€, 40k€], etc.).
    - est_rouge : Conversion en valeurs numériques (0/1).

- age_vehicule : Arrondi pour conserver uniquement des valeurs entières.
- date_demande : Transformation pour conserver seulement le mois et l'année.

3. Colonnes conservées :
   - metier, sexe, enfant_conduite_accompagne, nombre_enfants, formation, usage_vehicule, type_vehicule, points_perdus, type_conduite, etat_dossier, formule, tarif_devis.
   - Ces colonnes ne présentent pas de risque d'identification et sont nécessaires pour des analyses.
4. Colonnes ajoutées :
   - Index, id_client

3. Méthodes d'Anonymisation et Justification des Choix

Cette section détaille les méthodes appliquées pour anonymiser ou transformer les colonnes identifiées dans l'ancien dictionnaire des données, afin de respecter les principes fondamentaux de la RGPD (minimisation, pseudonymisation et irréversibilité) ou bien d'optimiser la BDD pour une future exploitation des données. Chaque méthode est accompagnée de sa justification.

	A	B	C
1	Nom des colonnes	Type	Description
2	Index	VARCHAR(15)	Identifiant alphanumérique unique. La taille de 15 peut être ajustée si besoin
3	id_client	VARCHAR(15)	Identifiant client anonymisé
4	metier	TEXT	Texte descriptif, car la longueur peut varier fortement
5	sexe	BOOLEAN	Donnée binaire (Femme = 0 et Homme= 1)
6	date_naissance	DATE	Date de naissance (format YYYY)
7	enfant_conduite_accompagne	INT	Nombre entier, car la valeur est quantitative discrète.
8	nombre_enfants	INT	Nombre entier, car la valeur est quantitative discrète.
9	revenus	VARCHAR(20)	Intervalle de revenus (par exemple : [40k€, 60k€]). La taille de 20 est adaptée.
10	formation	VARCHAR(50)	Niveau de formation (par exemple : "Bachelors")
11	usage_vehicule	VARCHAR(20)	Type d'usage (par exemple : "Commercial", "Private")
12	type_vehicule	VARCHAR(30)	Type de véhicule (par exemple : "Sports Car", "Minivan")
13	est_rouge	BOOLEAN	Indique si le véhicule est rouge (Non = 0 et Oui = 1)
14	points_perdus	INT	Nombre entier représentant les pointsde permis perdus.
15	age_vehicule	INT	Nombre entier indiquant l'âge en années du véhicule
16	type_conduite	VARCHAR(50)	Type de conduite (par exemple : "Highly Urban/Urban")
17	date_demande	DATE	Date de la demande du devis (format YYYY-MM)
18	etat_dossier	VARCHAR(20)	Statut du dossier (par exemple : "complet", "incomplet").
19	formule	VARCHAR(20)	Type de formule d'assurance (par exemple : "dev_integral").
20	tarif_devis	DECIMAL(10, 2)	Montant du tarif en décimal (exemple : 323.39).
21			
22			
23			
24			

Colonnes transformées et anonymisées

Colonne	Transformation appliquée	Justification
id_client	Nouveau système d'identifiants anonymisés basé sur l'index : chaque id_client a été généré au format ANON-000001, ANON-000002, etc., en concaténant un tag ANON- au champ Index.	Assure l'anonymisation tout en créant un identifiant unique pour l'analyse.
date_naissance	Réduction de granularité : seule l'année est conservée.	Limite les risques d'identification tout en permettant des analyses par cohorte d'âge.
revenus	Catégorisation : les revenus ont été transformés en tranches ([0, 20k€], [20k€, 40k€], etc.).	Réduit la précision pour éviter une réidentification tout en préservant leur utilité statistique.

est_rouge	Transformation booléenne : les valeurs yes et no ont été remplacées par 0 et 1.	Facilite les traitements statistiques tout en conservant la sémantique.
age_vehicule	Arrondi à l'entier : toutes les valeurs décimales ont été arrondies.	Améliore la cohérence et la lisibilité pour des analyses quantitatives.
date_demande	Réduction de granularité : seuls le mois et l'année sont conservés au format MM/AAAA.	Simplifie la granularité et limite l'identification indirecte.

Résumé des transformations appliquées

Colonne	Transformation	Exemple avant	Exemple après
id_client	Nouveau système basé sur l'index	12345	ANON-000001
date_naissance	Réduction à l'année uniquement	15/07/1985	1985
revenus	Catégorisation en tranches	125301	[100k€+]
est_rouge	Conversion en booléen numérique	yes	1
age_vehicule	Arrondi à l'entier	7.5	8
date_demande	Réduction à mois + année	2022-02-15	02/2022

4. Transformation et Anonymisation dans Power Query

Opération 1 : Nettoyage initial

- Action : Vérification des colonnes et suppression des champs inutiles qui n'étaient pas pertinents pour les analyses ou déjà filtrés dans SQLite.

Opération 2 : Génération d'un identifiant anonymisé (id_client)

- Action : Remplacement de la colonne id_client par une valeur générée automatiquement en fonction de l'Index.
- Création de l'Index :

```
= Table.AddIndexColumn(#"Colonnes supprimées4", "Index", 1, 1, Int64.Type)
```

- Génération d'un identifiant anonymisé (id_client) :

```
="ANON-" & Text.PadStart(Text.From([Index]), 6, "0")
```

Opération 3 : Réduction de granularité des dates

- Colonnes concernées : date_naissance, date_demande
- But : réduire le risque de réidentification.

Opération 4 : Catégorisation des revenus

- But : Transformation des montants exacts en tranches d'intervalles prédéfinies pour réduire le risque de réidentification.

```
= Table.AddColumn(#"Type modifié1", "revenus 2", each if [revenus] <= 20000 then "[0, 20k€]" else if [revenus] <= 40000 then "[20k€, 40k€]" else if [revenus] <= 60000 then "[40k€, 60k€]" else if [revenus] <= 80000 then "[60k€, 80k€]" else if [revenus] <= 100000 then "[80k€, 100k€]" else if [revenus] > 100000 then "[100k€+]" else if [revenus] = null then "Non Renseigné" else null)
```

Opération 5 : Transformation booléenne

- Colonnes concernées : est_rouge, sexe.
- But : Faciliter le traitement statistique
- Action : Conversion des valeurs textuelles (yes/no) et homme/femme en valeurs numériques (1 pour "oui", 0 pour "non") grâce à Transformer > Remplacer les valeurs

Opération 6 : Simplification des âges des véhicules

- But : Arrondi des valeurs décimales à l'entier le plus proche.
- Action : Transformer > Remplacer les valeurs . en , (sinon les données ne sont pas considérées comme des chiffres) puis Transformer>Type de données : nombre entier

Opération 7 : Choix du format appropriée de données

- Colonnes concernées : index, sexe, est_rouge, age_vehicule
- But : Optimiser l'exploitation et le traitement des données en choisissant le type de données appropriées pour chaque colonnes.

5. Export des données anonymisées

- Action finale : Export des données nettoyées et anonymisées vers un fichier CSV.

- Procédé dans Power Query :

Conclusion : Résultats et perspectives

Ce rapport détaille le processus d'anonymisation des données, de leur collecte initiale jusqu'à leur préparation dans un ETL (Power Query), en garantissant leur conformité aux exigences du RGPD. Les principales actions mises en œuvre sont les suivantes :

- **Identification et correction des violations RGPD :** des recommandations ont été formulées pour supprimer ou transformer les colonnes sensibles dans le CRM. Les principes de minimisation et d'irréversibilité ont guidé ces choix.

- **Extraction ciblée des données pertinentes :** seules les données des dossiers "complets" et datant de 2022 ont été sélectionnées depuis la base SQL. Cette extraction rigoureuse a permis de répondre aux critères de la consigne tout en réduisant les risques de non-conformité.

- **Anonymisation et transformation dans Power Query :** des techniques d'anonymisation robustes (pseudonymisation des identifiants, catégorisation des revenus, agrégation des dates, etc.) ont été appliquées pour préserver l'utilité analytique des données tout en garantissant leur protection.

- **Optimisation pour l'exploitation des données** : en complément de l'anonymisation, des optimisations ont été effectuées pour améliorer la qualité et la facilité d'utilisation des données. Cela inclut l'application de types de données adaptés (par exemple, conversion de `est_rouge` au format booléen) afin de faciliter d'éventuels traitements statistiques, ainsi que l'arrondissement des âges de véhicule pour assurer une meilleure cohérence des données numériques. Ces ajustements garantissent que les

données sont prêtes pour des analyses efficaces tout en réduisant les ambiguïtés potentielles.

Impacts et bénéfices :

1. **Conformité légale :** La mise en œuvre de ces transformations assure que les données collectées et utilisées respectent les principes RGPD. Cela répond aux exigences de la CNIL et évite des sanctions supplémentaires.

2. **Optimisation de l'analyse commerciale :** Les données anonymisées restent exploitables pour des analyses stratégiques, permettant de maintenir les performances commerciales malgré les restrictions.

3. **Réduction des risques organisationnels :** En adoptant des procédures rigoureuses d'anonymisation, l'entreprise améliore ses pratiques de gouvernance des données, renforçant ainsi la confiance de ses clients et partenaires.