



RAPPORT

Collecte et préparation des données

Résumé

Synthèse des actions de la collecte à la préparation des données pour le projet 4.
En temps normal ce genre de données sont consigné dans le registre.

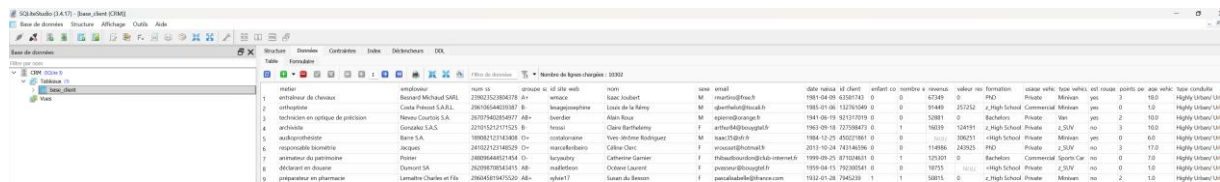
Yann RES

Table des matières

Extraction depuis le CRM.	2
Transformation dans Excel avec PowerQuery.....	3

Extraction depuis le CRM.

Ici pour simuler le CRM on va utiliser SQLite studio pour explorer la db crm.sqlite et la table base_client. Une fois le fichier ouvert on peut explorer les données et noter celle que l'on va garder pour l'export.



	nom	email	date_naissance	enfant	nombre	e	internet	voiture	formation	usage_vehicule	type_vehicule	est_rouge	points_perdus	age_vehicule	type_conduite	date_dossier	etat_dossier	formule	tarif_devis	adresse
1	entraineur de chevaux	Benoit Michael SIREL	23/02/1970	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	entraîneur	Costa Patrick S.A.S.	28/07/1980	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	technicien en optique de précision	Benet-Courtes S.A.	20/07/1980	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	architecte	Gonzalez S.A.S.	22/07/1970	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	architecte	Benet-Courtes S.A.	20/07/1980	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	responsable technique	Jacques	24/02/1970	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	entraîneur du patrimoine	Pinet	28/07/1980	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	diagnosticien en immobilier	Dumont S.A.	20/07/1980	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	propriétaire en pharmacie	Lemaitre Charles et fils	28/07/1980	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Voici le script SQL utilisé pour afficher uniquement les données sélectionnées précédemment avant de les extraire avec l'outil exporter le résultat.

```
SELECT sexe,
date_naissance,
nombre_enfants,
enfant_conduite_accompagne,
usage_vehicule,
type_vehicule,
est_rouge,
points_perdus,
age_vehicule,
type_conduite,
date_dossier,
etat_dossier,
formule,
tarif_devis,
adresse
FROM base_client
WHERE etat_dossier = 'complet'
```

A la base j'avais ajouté une fonction supplémentaire dans le where pour avoir les contrats incomplets de moins de 3 ans, mais après vérification la clause est null car toutes les lignes sont de plus de 3 ans.

```
Where
etat_dossier = 'complet'
OR (
etat_dossier = 'Incomplet'
AND date(date_dossier) >= date('now', '-3 years')
)
```

Transformation dans Excel avec PowerQuery.

Première étape on crée une nouvelle feuille Excel, puis données, import on sélectionne le CSV exporté depuis SQLite studio et Transformer pour ouvrir Power Query. Pour la transformation des différentes colonnes on va utiliser l'ajout de nouvelle colonne et utiliser une formule en langage M.

La colonne contenant l'information sur le sexe du client on la garde tel quel.

Pour ce qui est de la date de naissance, on va la transformer en âge et ici plus précisément en plage d'âge pour renforcer l'anonymisation.

```
let
    today = Date.From(DateTime.LocalNow()),
    age = Number.RoundDown(
        Duration.Days(today - [date_naissance]) / 365.25
    )
in
    if age < 18 then "<18"
    else if age <= 24 then "18-24"
    else if age <= 29 then "25-29"
    else if age <= 34 then "30-34"
    else if age <= 39 then "35-39"
    else if age <= 44 then "40-44"
    else if age <= 49 then "45-49"
    else if age <= 54 then "50-54"
    else if age <= 59 then "55-59"
    else if age <= 64 then "60-64"
    else if age <= 74 then "65-74"
    else "75+"
```

Puis on supprime la colonne avec la date de naissance.

Les prochaines colonnes à anonymiser sont nombre d'enfant et nombre d'enfant en conduite accompagné que l'on va remplacer par Oui/NON, le code est le même pour les deux seule la valeur [...] change.

```
if [nombre_enfants] = null then "NA"
else if [nombre_enfants] = 0 then "Non"
else "Oui"
```

En suite je réorganise certaines colonnes liées au thème du véhicule pour avoir une lecture plus simple. J'ai donc remanié les colonnes pour avoir type_vehicule, usage_vehicule, age_véhicule, est_rouge, type_conduite dans cet ordre.

Puis j'ai modifié age_vehicule et points perdus pour avoir des plages de 4 pour l'âge et 3 pour les points, la formule et la même dans le fond il suffit de modifier le pas, sauf pour l'Age des véhicules où l'on a un chiffre à virgules dans une forme de texte, donc une étape supplémentaire. Voici l'exemple pour l'âge des véhicules.

```
if [age_vehicule] = null or [age_vehicule] = "" then "NA"
else
  let
    valeur = Number.FromText(Text.Replace([age_vehicule],
    ".", ","), "fr-FR"),
    taillePlage = 5,
    base = taillePlage *
Number.IntegerDivide(Number.RoundDown(valeur), taillePlage)
  in
    Text.From(base) & "-" & Text.From(base + (taillePlage
- 1))
```

Pour ce qui est de date_demande je la passe directement de date_heure en date depuis le selecteur de type dans l'entête de la colonne puis je créer une nouvelle colonne pour tronquer la date en MM/YYYY avec ce code.

```
if [date_demande] = null then null
else Date.ToText([date_demande], "MM/yyyy")
```

Pour ce qui est de l'état du dossier et du type de formule on ne touche à rien.

Après on passe au tarif_devis que l'on va transformer en plage tarifaire de devis.

Pour ce nous avons encore une fois un format texte avec des nombre à virgule en format US. Donc on extrait le texte on transforme en plage sans déclame par pas de 50.

```
let
  txt = Text.From([tarif_devis]),
  num = try Number.FromText(txt, "en-US") otherwise null
in
  if num = null then "NA"
  else
    let
      t      = Number.RoundDown(num, 0),
      base = 50 * Number.IntegerDivide(t, 50),
      max  = base + 49
    in
      Text.From(base) & "-" & Text.From(max)
```

Pour ce qui est de l'adresse c'est un peu plus complexe du faite que l'on a une adresse complète. Ici il faut isoler le code postal du texte pour ensuite ne garder que les deux premiers chiffres. Néanmoins il fallait ajouter un garde-fou car certaines code postaux ont perdu leurs 0 (département du 01 à 09) auxquelles on ajoute le 0 perdu.

Voici le code.

```
let
  txt = Text.From([adresse]),
  parts = Text.Split(txt, " "),
  candidates =
    List.Select(
      parts,
      each
        (
          Text.Length(_) = 5
          and Text.Select(_, {"0".."9"}) = _
        )
      or
        (
          Text.Length(_) = 4
          and Text.Select(_, {"0".."9"}) = _
        )
    ),
  cpRaw = if List.IsEmpty(candidates) then null else
List.First(candidates),
  cpNorm =
    if cpRaw = null then null
    else if Text.Length(cpRaw) = 5 then cpRaw
    else "0" & cpRaw,
  dept = if cpNorm = null then null else Text.Start(cpNorm,
2)
in
  dept
```