OPC Data Analyst Projet
Synthèse de projet

Arnaud Golliot
7 Août 2024

Sommaire

Normalisation du modèle de données

Normalisation du modèle de données

Constat :

Les deux fichiers CSV fournis – Contrats.csv et Régions.csv - contiennent de nombreuses redondances de données, ce qui présente les écueils suivants :

- ✓ **La volumétrie** : plus les mêmes valeurs de données sont répétées, plus la base de données résultant du chargement de ces deux fichiers sera gourmande en espace de stockage
- ✓ **La lisibilité** : les deux fichiers contiennent plusieurs concepts, ce qui en pénalise la compréhension
  - ➢ Contrats.csv : il adresse deux concepts, les contrats mais aussi les logements
  - ➢ Régions.csv : il adresse trois concepts différents
- ✓ **La consistance** : modifier un valeur de donnée en base nécessiterait d'en ressaisir la valeur à tous les endroits ou elle se trouve, ce qui est chronophage et surtout présente un risque, pour ne pas dire certitude, d'avoir des valeurs de données différentes pour un même concept (par exemple plusieurs adresses différentes pour un même logement, typiquement « 1, rue Jean Jaurès » et «1 rue J.Jaurès)

Solutions :

Afin de pallier à ces écueils, nous avons transformé le modèle de données de manière à avoir 4 entités au lieu de deux. Cette transformation implique de découper les deux fichiers csv fournis en quatre fichiers cible. Le modèle de données résultant est dit en 3^{ième} forme normale (3FN). Dans ce type de modèle, chaque combinaison de données dans une entité est unique.

Étapes du projet

Normalisation du modèle de données

1) **Etablissement du dictionnaire initial (2 entités : Contrat et Région)**

2) **Etablissement du dictionnaire en 3ieme forme normale (5 entités : Contrat, Logement, Commune, Département et Région)**

3) **Découpage des 2 fichiers CSV initiaux (Contrats.csv, Régions.csv) en 5 fichiers CSV normalisés (Contrats.csv, Logements.csv, Communes.csv, Departements.csv, Regions.csv). Neutralisation des doublons. Rajout de clés primaires. Rajout de données manquantes (communes de la réunion)**

4) **Installation du SGBD MYSQL**

5) **Ecriture avec SUBLIME TEXT et exécution avec MYSQL COMMAND LINE du script SQL de création de la base de données**

6) **Ecriture avec SUBLIME TEXT et exécution avec MYSQL COMMAND LINE du script SQL de création et d'alimentation des tables**

7) **Vérification du contenu et du nombre d'enregistrements des tables dans MYSQL Workbench**

8) **Ecriture des 12 requêtes SQL (plus requête de test) avec l'éditeur SUBLIME TEXT, et exécution dans MYSQL Workbench**

Dictionnaires de données

Dictionnaire de données initial

CONTRAT.CSV

Nom des colonnes	Type de données	Taille	Clé	Description
Contrat_ID	INT		Clé primaire	Id unique pour les contrats
No_voie	CVARCHAR	3		Numéro dans la voie pour l'adresse du logement assuré Le numéro de voie n'est pas fondamentalement une variable quantitative, mieux vaut la mettre en type CVARCHAR
B_T_Q	CHAR	1		Indicateur éventuel de répétition pour l'adresse du logement assuré sur un caractère
Type_de_voie	CVARCHAR	4		Type de voie pour l'adresse du logement assuré: rue, av (Avenue), rte (Route), ...
Voie	CVARCHAR	30		Libellé de la voie pour l'adresse du logement assuré
Code_dep_code_commune	CVARCHAR	6	Clé secondaire	Concaténation du code département et code commune pour avoir une clé unique Attention : les deux codes de la Corse contiennnent des caractères (2A, 2B)
Code_postal	CVARCHAR	5		Code postal pour l'adresse du logement assuré Attention : le code postal n'est pas une donnée quantitative, mieux vaut le mettre en type CVARCHAR
Surface	INT			Surface du logement assuré
Type_local	CVARCHAR	20		Type de logement (maison, appartement)
Occupation	CVARCHAR	20		Qualité de l'occupant (propriétaire, locataire)
Type_contrat	CVARCHAR	30		Type de contrat de la personne assurée (résidence principale, résidence secondaire, mise en location)
Formule	CVARCHAR	20		Formule du contrat d'assurance (Classique, Integral)
Valeur_declaree_biens	CVARCHAR	20		Tranche de valeur des biens déclarés par l'assuré (4 tranches : 0-25K€, 25-50K€, 50-100K€, > 100K€)
Prix_cotisation_mensuelle	INT			Prix de la cotisation mensuelle de l'assurance contractée par l'assuré

REGION.CSV

Nom des colonnes	Type de données	Taille	Clé	Description
Code_dep_code_commune	CVARCHAR	6	Clé primaire	Concaténation du code département et code commune pour avoir une clé unique
reg_code	INT			Code région (par exemple : 44 pour la région Grand Est)
reg_nom	CVARCHAR	30		Nom de la région
aca_nom	CVARCHAR	40		Nom de l'académie (exemple : académie de Versailles)
dep_nom	CVARCHAR	50		Nom du departement
com_nom_maj_court	CVARCHAR	50		Nom de la commune en majuscules (exemple : AMBERIEU EN BUGEY)
dep_code	INT	3		Code département (de 01 à 98)
dep_nom_num	VACHAR	50		Nom du département complété en suffixe par le code département entre parenthèses)

Dictionnaire de données normalisé (3FN)

Entité : Contrat

Nom des colonnes	Type de données	Taille	Clé	Description
Id_contrat	INT		Clé primaire	Id unique pour les contrats
Type_contrat	CVARCHAR	30		Type de contrat de la personne assurée (résidence principale, résidence secondaire, mise en location)
Formule	CVARCHAR	20		Formule du contrat d'assurance (Classique, Integral)
Valeur_declaree_biens	CVARCHAR	20		Tranche de valeur des biens déclarés par l'assuré (4 tranches : 0-25K€, 25-50K€, 50-100K€, > 100K€)
Prix_cotisation_mensuelle	INT			Prix de la cotisation mensuelle de l'assurance contractée par l'assuré
Code_logement	INT		Clé secondaire	Identifiant unique de logement

Entité : Commune

Nom des colonnes	Type de donn	Taille	Clé	Description
Code_commune	INT		Clé primaire	Identifiant de commune
Code_postal	CVARCHAR	7		Code postal pour l'adresse du logement assuré Attention : le code postal n'est pas une donnée
Nom_commune	CVARCHAR	50		Nom de la commune en majuscules (exemple : AMBERIEU EN BUGEY)
Code_departement	INT	3	Clé secondaire	Code département (de 01 à 98)

Entité : Logement

Nom des colonnes	Type de donn	Taille	Clé	Description
Code_logement	INT		Clé primaire	Identifiant unique de logement
No_voie	CVARCHAR	3		Numéro dans la voie pour l'adresse du logement assuré
B_T_Q	CHAR	1		Indicateur éventuel de répétition pour l'adresse du logement assuré sur un caractère
Type_de_voie	CVARCHAR	4		Type de voie pour l'adresse du logement assuré: rue, av (Avenue), rte (Route), ...
Voie	CVARCHAR	30		Libellé de la voie pour l'adresse du logement assuré
Code_commune	INT		Clé secondaire	Identifiant unique de commune
Surface	INT			Surface du logement assuré
Type_local	CVARCHAR	20		Type de logement (maison, appartement)
Occupation	CVARCHAR	20		Qualité de l'occupant (propriétaire, locataire)

Entité : Département

Nom des colonnes	Type de donn	Taille	Clé	Description
Code_departement	INT	3	Clé primaire	Code département (de 01 à 98)
Nom_departement	CVARCHAR	50		Nom du departement
Code_region	INT		Clé secondaire	Code région

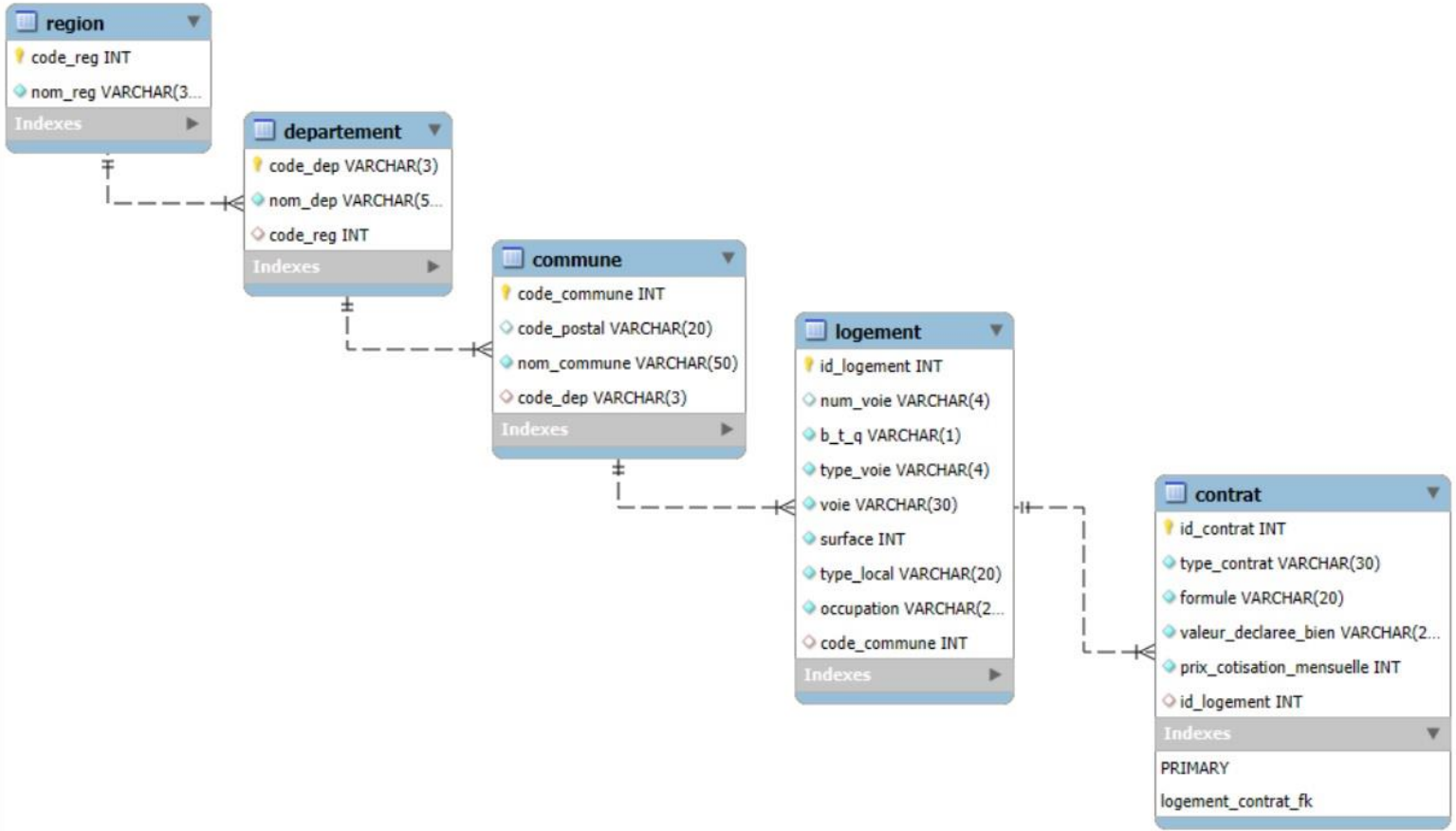
Entité : Région

Nom des colonnes	Type de donn	Taille	Clé	Description
Code_region	INT		Clé primaire	Code région (par exemple : 44 pour la région Grand Est)
reg_nom	CVARCHAR	30		Nom de la région

Schéma relationnel normalisé (3FN)

Schéma relationnel

Modèle en 3FN



Structure et contenu des tables dans la BDD

Structure et contenu des tables dans la BDD

Table: contrat

Columns:

id_contrat	int PK
type_contrat	varchar(30)
formule	varchar(20)
valeur_declaree_bien	varchar(20)
prix_cotisation_mensuelle	int
id_logement	int

id_contrat	type_contrat	formule	valeur_declaree_bien	prix_cotisation_mensuelle	id_logement
100601	Residence principale	Integral	0-25000	25	147
100602	Residence principale	Classique	0-25000	30	50
100603	Residence principale	Integral	25000-50000	57	63
100604	Residence principale	Integral	25000-50000	43	131
100605	Residence principale	Classique	0-25000	33	16
100606	Residence principale	Classique	0-25000	19	149
100607	Residence principale	Integral	0-25000	15	160
100608	Mise en location	Integral	25000-50000	34	67
100609	Residence principale	Classique	25000-50000	32	151
100610	Residence principale	Integral	25000-50000	22	132
100611	Residence secondaire	Classique	0-25000	11	130
100612	Residence principale	Classique	0-25000	16	90

30335 row(s) returned

Table: logement

Columns:

id_logement	int PK
num_voie	varchar(4)
b_t_q	varchar(1)
type_voie	varchar(4)
voie	varchar(30)
surface	int
type_local	varchar(20)
occupation	varchar(20)
code_commune	int

id_logement	num_voie	b_t_q	type_voie	voie	surface	type_local	occupation	code_commune
1	59		RUE	ALEXANDRE BERARD	165	Appartement	Locataire	2
2	54	B	RUE	ARISTIDE BRIAND	70	Appartement	Locataire	2
3	2		RUE	DE LA BATISSE	81	Appartement	Proprietaire	2
4	25		RUE	DE LA COMMUNE 1871	70	Maison	Proprietaire	2
5	72		RUE	DE LA REPUBLIQUE	36	Appartement	Proprietaire	2
6	36		RUE	DES APOTRES	75	Appartement	Locataire	2
7	60		ALL	DES FRERES CAUDRON	51	Appartement	Proprietaire	2
8	78		AV	DU GEN SARRAIL	92	Appartement	Proprietaire	2
9	13		RUE	JEAN MONNET	55	Appartement	Proprietaire	2
10	54		RUE	MARCEL DEMIA	68	Appartement	Locataire	2
11	29		RUE	REINE CLOTILDE	55	Appartement	Proprietaire	2
12	47		RUE	DU CHANE	48	Appartement	Proprietaire	12

30083 row(s) returned

Structure et contenu des tables dans la BDD

Table: commune

Columns:

code_commune	int PK
code_postal	varchar(20)
nom_commune	varchar(50)
code_dep	varchar(3)

code_commune	code_postal	nom_commune	code_dep
1	Non référencé	AMAREINS	1
2	1004	AMBERIEU EN BUGEY	1
3	Non référencé	AMBERIEUX EN DOMBES	1
4	Non référencé	AMBLEON	1
5	Non référencé	AMBRONAY	1
6	Non référencé	AMBUTRIX	1
7	Non référencé	ANDERT ET CONDON	1
8	Non référencé	ANGLEFORT	1

38917 row(s) returned

Table: departement

Columns:

code_dep	varchar(3) PK
nom_dep	varchar(50)
code_reg	int

code_dep	nom_dep	code_reg
1	Ain	84
10	Aube	44
11	Aude	76
12	Aveyron	76
13	Bouches-du-Rhône	93
14	Calvados	28
15	Cantal	84
16	Charente	75
17	Charente-Maritime	75

109 row(s) returned

Table: region

Columns:

code_reg	int PK
nom_reg	varchar(30)

code_reg	nom_reg
0	Collectivités d'outre-mer
1	Guadeloupe
2	Martinique
3	Guyane
4	La Réunion
6	Mayotte
11	Ile-de-France
24	Centre-Val de Loire

19 row(s) returned

Code SQL et résultat des requêtes

Résultat et code des requêtes SQL

Requête exemple : Lister les contrats avec le prix de la cotisation et leur surface pour les appartements

```
SELECT    CTR.id_contrat, CTR.prix_cotisation_mensuelle, LOG.surface
FROM      contrat CTR
          LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
WHERE     LOG.type_local = 'Appartement';
```

id_contrat	prix_cotisation_mensuelle	surface
100601	25	50
100602	30	48
100603	57	131
100605	33	109
100606	19	53
100607	15	59
100610	22	36
100611	11	138
100612	16	45
100613	14	83
100614	34	88

contrat 1

27837 row(s) returned

Requête n°1 : Lister les numéros de contrats (id_contrat) avec leur surface pour la commune de Caen.

```
SELECT    CTR.id_contrat,
          LOG.surface
FROM      contrat CTR
          LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
          LEFT OUTER JOIN commune COM ON LOG.code_commune = COM.code_commune
WHERE     COM.nom_commune = 'CAEN';
```

id_contrat	surface
103791	35
103792	99
103793	40
103794	20

4 row(s) returned

Résultat et code des requêtes SQL

Requête n°2 : Lister les numéros de contrats avec le type de contrat et leur formule pour les maisons du département 71.

```
SELECT    CTR.id_contrat, CTR.type_contrat, CTR.formule
FROM      contrat CTR
          LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
          LEFT OUTER JOIN commune COM ON LOG.code_commune = COM.code_commune
          LEFT OUTER JOIN departement DEP ON COM.code_dep = DEP.code_dep
WHERE     LOG.type_local = 'Maison'
          AND DEP.code_dep = '71'
ORDER BY  CTR.id_contrat ASC
```

contrat_id	type_contrat	formule
114768	Residence principale	Integral
114779	Residence principale	Classique
114782	Residence principale	Classique
114812	Residence principale	Integral

4 row(s) returned

Requête n°3 : Lister le nom des régions de France.

```
SELECT    REG.nom_reg
FROM      region REG
ORDER BY  REG.nom_reg;
```

nom_reg
Auvergne-Rhône-Alpes
Bourgogne-Franche-Comté
Bretagne
Centre-Val de Loire
Collectivités d'outre-mer
Corse
Grand Est
Guadeloupe
Guyane
Hauts-de-France
Ile-de-France
La Réunion
Martinique
Mayotte
Normandie
Nouvelle-Aquitaine
Occitanie
Pays de la Loire
Provence-Alpes-Côte d'Azur

19 row(s) returned

Résultat et code des requêtes SQL

Requête n°4 : Quels sont les 5 contrats qui ont les surfaces les plus élevées ?

```
SELECT    CTR.id_contrat, LOG.surface
FROM      contrat CTR
          LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
ORDER BY  LOG.surface DESC
LIMIT     5;
```

	id_contrat	surface
▶	104211	815
	105463	742
	130878	595
	100822	570
	109872	559
*	NULL	NULL

5 row(s) returned

Requête n°5 : Quel est le prix moyen de la cotisation mensuelle ?

```
SELECT    CONCAT(ROUND(AVG(CTR.prix_cotisation_mensuelle), 2), ' €') AS 'Moyenne cotisation Mensuelle'
FROM      contrat CTR;
```

	Moyenne cotisation Mensuelle
▶	19.33 €

1 row(s) returned

Requête n°6 : Quel est le nombre de contrats pour chaque catégorie de prix de la valeur déclarée des biens ?

```
SELECT    CTR.valeur_declaree_bien AS 'Catégorie de biens',
          COUNT(CTR.id_contrat) AS 'Nombre de contrats'
FROM      contrat CTR
GROUP BY  CTR.valeur_declaree_bien
ORDER BY  COUNT(CTR.id_contrat) DESC;
```

	Catégorie de biens	Nombre de contrats
▶	0-25000	22720
	25000-50000	6815
	50000-100000	696
	100000+	104

4 row(s) returned

Résultat et code des requêtes SQL

Requête n°7 : Quel est le nombre de formules "integral" sur la région Pays de la Loire ?

```
SELECT   COUNT(CTR.id_contrat) AS 'Nombre de contrat en formule Intégrale pour les Pays de la Loire'
FROM     contrat CTR
         LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
         LEFT OUTER JOIN commune COM ON LOG.code_commune = COM.code_commune
         LEFT OUTER JOIN departement DEP ON COM.code_dep = DEP.code_dep
         LEFT OUTER JOIN region REG ON DEP.code_reg = REG.code_reg
WHERE    REG.nom_reg = 'Pays de la Loire'
         AND CTR.formule = 'Integral';
```

	Nombre de contrat en formule Intégrale pour les Pays de la Loire
▶	589

1 row(s) returned

Requête n°8 : Lister les numéros de contrats avec le type de contrat et leur formule pour les maisons du département 71.

```
SELECT   CTR.id_contrat, CTR.type_contrat, CTR.formule
FROM     contrat CTR
         LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
         LEFT OUTER JOIN commune COM ON LOG.code_commune = COM.code_commune
         LEFT OUTER JOIN departement DEP ON COM.code_dep = DEP.code_dep
WHERE    LOG.type_local = 'Maison'
         AND DEP.code_dep = '71';
ORDER BY CTR.id_contrat ASC
```

	id_contrat	type_contrat	formule
▶	114768	Residence principale	Integral
	114779	Residence principale	Classique
	114782	Residence principale	Classique
	114812	Residence principale	Integral

4 row(s) returned

Requête n°9 : Quelle est la surface moyenne des contrats à Paris ?

```
SELECT   CONCAT(ROUND(AVG(LOG.surface), 2), ' M2')
         AS 'Surface moyenne des logements assurés à Paris'
FROM     contrat CTR
         LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
         LEFT OUTER JOIN commune COM ON LOG.code_commune = COM.code_commune
         LEFT OUTER JOIN departement DEP ON COM.code_dep = DEP.code_dep
WHERE    DEP.nom_dep = 'Paris';
```

	Surface moyenne des logements assurés à Paris
▶	51.77 M2

1 row(s) returned

Résultat et code des requêtes SQL

Requête n°10 : Classement des 10 départements où le prix moyen de la cotisation est le plus élevé

```
SELECT      DEP.nom_dep AS 'departement',
            CONCAT(ROUND(AVG(CTR.prix_cotisation_mensuelle),2), '€') AS 'cotisation_moyenne'
FROM        contrat CTR
            LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
            LEFT OUTER JOIN commune COM ON LOG.code_commune = COM.code_commune
            LEFT OUTER JOIN departement DEP ON COM.code_dep = DEP.code_dep
GROUP BY    DEP.nom_dep
ORDER BY    AVG(CTR.prix_cotisation_mensuelle) DESC
LIMIT 10;
```

	departement	cotisation_moyenne
▶	Paris	36.40€
	Hauts-de-Seine	26.27€
	Val-de-Marne	19.82€
	Yvelines	18.89€
	Rhône	18.49€
	Ain	18.24€
	Alpes-Maritimes	18.14€
	Charente-Maritime	17.32€
	Haute-Savoie	17.15€
	Corse-du-Sud	17.07€

10 row(s) returned

Requête n°11 : Liste des communes ayant eu au moins 150 contrats.

```
SELECT      COM.nom_commune AS 'Commune',
            COUNT(CTR.id_contrat) AS 'Nombre de contrats'
FROM        contrat CTR
            LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
            LEFT OUTER JOIN commune COM ON LOG.code_commune = COM.code_commune
GROUP BY    COM.code_commune
HAVING      COUNT(CTR.id_contrat) >= 150
ORDER BY    COUNT(CTR.id_contrat) DESC;
```

	Commune	Nombre de contrats
▶	PARIS 18	515
	PARIS 17	468
	PARIS 15	407
	PARIS 16	394
	NICE	387
	PARIS 11	381
	BORDEAUX	302
	PARIS 20	302
	NANTES	291
	PARIS 19	266
	PARIS 10	263
	PARIS 12	252
	PARIS 14	222
	GRENOBLE	220
	PARIS 9	204
	TOULOUSE	187
	TOULON	170
	COURBEV...	163
	LILLE	161
	PARIS 3	159

20 row(s) returned

Résultat et code des requêtes SQL

Requête n°12 : Quel est le nombre de contrats pour chaque région ?

```
SELECT      REG.nom_reg AS 'Région',
            COUNT(CTR.id_contrat) AS 'Nombre de contrats'
FROM        contrat CTR
            LEFT OUTER JOIN logement LOG ON CTR.id_logement = LOG.id_logement
            LEFT OUTER JOIN commune COM ON LOG.code_commune = COM.code_commune
            LEFT OUTER JOIN departement DEP ON COM.code_dep = DEP.code_dep
            LEFT OUTER JOIN region REG ON DEP.code_reg = REG.code_reg
GROUP BY    REG.nom_reg
ORDER BY    COUNT(CTR.id_contrat) DESC;
```

Région	Nombre de contrats
Ile-de-France	14177
Provence-Alpes-Côte d'Azur	3279
Auvergne-Rhône-Alpes	3042
Nouvelle-Aquitaine	2038
Occitanie	1609
Pays de la Loire	1196
Hauts-de-France	1189
Bretagne	947
Normandie	824
Grand Est	769
Centre-Val de Loire	598
Bourgogne-Franche-Comté	293
Corse	247
Martinique	73
Guyane	37
La Réunion	17

16 row(s) returned