3 - Méthodologie



Parcours Data Analyst - Marie G.
Projet 3 : Requêtez une base de données avec SQL
19/02/2025

- 1. Exploration des données
- 2. Schéma relationnel de la base de données
- 3. Création des tables
- 4. Chargement des données
- 5. Interrogation des tables par des requêtes SQL

1. Exploration des données

2 fichiers csv sont fournis dans le cadre du projet pour construire notre base de données :

- Contrat+(4).csv: Liste des contrats d'assurance logement de l'entreprise (30 335 lignes de contrats)
- Region+(7).csv: Liste des communes de France, dont les données sont issues d'un site gouvernemental (38 916 lignes de communes) Certains codes communes issus d'anciens découpages administratifs sont obsolètes, comme ceux de l'ancien département de la Seine comprenant la ville de Paris et les communes de petite couronne.

Le dictionnaire de données suivant est construit en analysant sous Excel les données des deux fichiers, après création de tableaux pour une meilleure vision de l'ensemble des données pouvant être stockée dans chaque colonne.

Il reprend les éléments qui seront nécessaire pour construire la base de donnée SQL, ainsi qu'une description en documentation annexe

	Nom des colonnes	Type de données	Taille	Clé	Description
	Contrat_ID	INT	5	Clé primaire	Id unique pour les contrats
	No_voie	INT	4		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	VARCHAR	4		Type de voie pour l'adresse du logement assuré: rue, av (Avenue), rte (Route),
>	Voie	VARCHAR	100		Libellé de la voie pour l'adresse du logement assuré
.csv	Code_dep_code_commune	VARCHAR	6	Clé secondaire	Concaténation du code département et code commune pour avoir une clé commune unique
AT	Code_postal	VARCHAR	5		Code postal pour l'adresse du logement assuré
CONTRAT	Surface	INT	4		Surface du logement assuré (m²)
Ö	Type_local	VARCHAR	100		Type de logement (appartement ou maison)
	Occupation	VARCHAR	100		Occupant (locataire ou propriétaire)
	Type_contrat	VARCHAR	100		Occupation (Résidence principale, résidence secondaire ou mise en location)
	Formule	VARCHAR	100		Formule choisie comme assurance (classique ou intégrale)
	Valeur_declaree_biens	VARCHAR	100		Valeur des biens déclarés à l'assurance (€)
	Prix_cotisation_mensuel	INT	4		Cotisation mensuelle de l'assuré (€)
	Code_dep_code_commune	VARCHAR	6	Clé primaire	Concaténation du code département et code commune pour avoir une clé commune unique
	reg_code	INT	2		Code région
CSV	reg_nom	VARCHAR	100		Nom de la région
j	aca_nom	VARCHAR	100		Nom de l'académie
REGIOI	dep_nom	VARCHAR	100		Nom du département
RE	com_nom_maj_court	VARCHAR	100		Nom de la commune
	dep_code	VARCHAR	3		Code département
	dep_nom_num	VARCHAR	100		Nom et code du département

Dictionnaire des données de la future base de données

2. Schéma relationnel de la base de données

Le dictionnaire de données présenté à la page précédente permet de modifier le schéma relationnel fourni dans le logiciel SQL Power Architect



Points d'attention :

- 1. SQL Power Architect ne permet pas de donner une taille aux données de type INTEGER. La taille des données renseignée dans le dictionnaire construit devra donc être ajoutée dans la commande de création de tables extraite par SQL Power Architect
- 2. Les colonnes doivent respecter l'ordre des colonnes fichiers csv prévus pour l'import de données

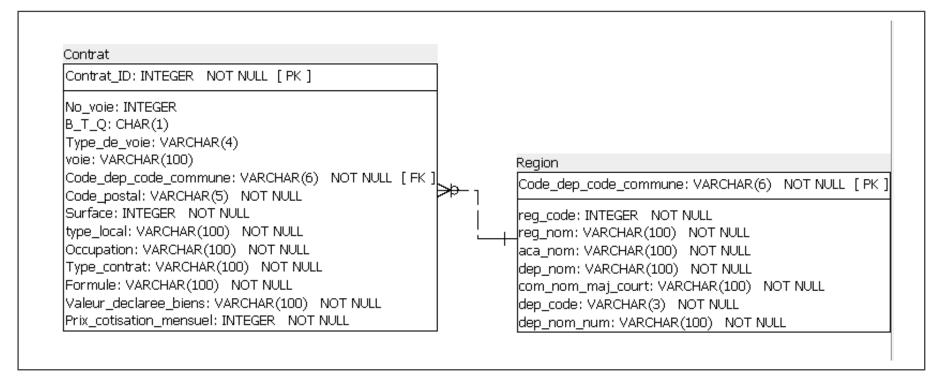


Schéma relationnel de la base de données

3. Création des tables

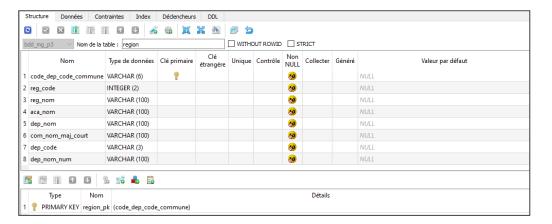
SQL Power Architect permet de générer le code ci-dessous sous le choix de format MySQL, après :

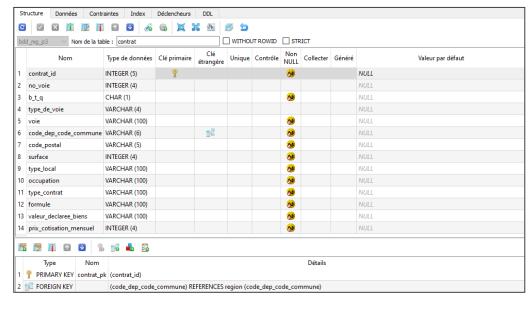
- Ajout de la taille des champs de format INTEGER
- Mise en forme sous https://www.dpriver.com/pp/sqlformat.htm
- Intégration de l'instruction « FOREIGN KEY» directement dans la création de la table « contrat »

```
CREATE TABLE region
 code dep code commune2 VARCHAR(6) NOT NULL,
 reg code
                        INT NOT NULL
 dep nom num
                      VARCHAR (100) NOT NULL,
 dep code
                       VARCHAR(3) NOT NULL
 com nom maj court VARCHAR (100) NOT NULL,
                        VARCHAR (100) NOT NULL
 dep nom
 reg nom
                        VARCHAR (100) NOT NULL
                        VARCHAR (100) NOT NULL
 aca nom
 PRIMARY KEY (code dep code commune2)
CREATE TABLE contrat
 contrat id
                         INT NOT NULL
                         VARCHAR (100) NOT NULL
 type local
 valeur declaree biens VARCHAR(100) NOT NULL,
 formule
                         VARCHAR (100) NOT NULL
 no voie
                         INT_{r}
 surface
                         INT NOT NULL
 btq
                         CHAR(1) NOT NULL,
 code dep code commune2 VARCHAR(6) NOT NULL,
 type de voie
                         VARCHAR (4)
 voie
                         VARCHAR (100) NOT NULL
 type contrat
                         VARCHAR NOT NULL,
 occupation
                         VARCHAR (100) NOT NULL
 code postal
                         VARCHAR (5) NOT NULL
 prix cotisation mensuel INT NOT NULL,
 PRIMARY KEY (contrat id)
 FOREIGN KEY (code dep code commune) REFERENCES region (
 code dep code commune)
```



On utilise le logiciel SQLiteStudio pour créer les tables vides à partir de la requête SQL :





4. Chargement des données

- Avec l'outil d'importation de données au format CSV de SQLiteStudio, on charge en premier lieu les données de la table « régions » car elle contient la donnée « code_dep_code_commune », qui est une clé secondaire de la table « contrats »
- 2. En chargeant la table « contrats » de la même manière, on constate que la clé secondaire n'est pas connue pour certains contrats :

16:31:22] Erreur lors de l'import des données : FOREIGN KEY constraint failed

En effectuant une recherche sous Excel (par exemple en ajoutant dans la table « contrat » une colonne qui vérifie si la valeur « code_dep_code_commune » figure dans la table « région », on repère les lignes dont le **code commune** (code_dep_code_commune) a mal été renseigné : ici, le code postal a été renseigné au lieu du code commune sur des contrats situés à la Réunion. On enregistre alors le fichier csv avec les valeurs rectifiées grâce à une recherche sur internet. On parvient ensuite à charger la table « contrats » rectifiée.

Table « région » chargée

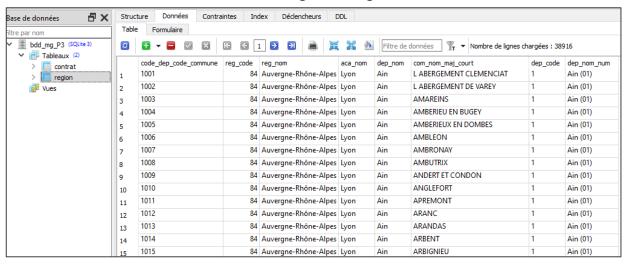
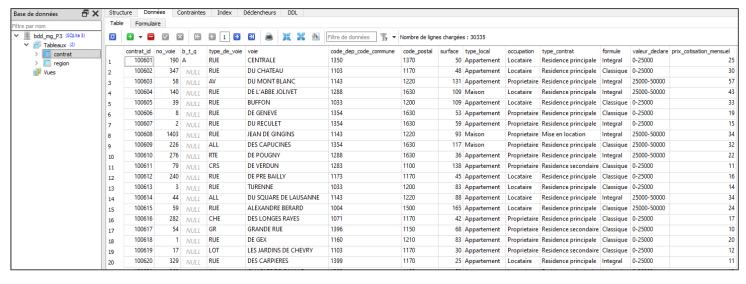


Table « contrat » chargée



5. Interrogation des tables par des requêtes SQL 1/4

Requête 1: Lister les numéros de contrats (contrat_ID) avec leur surface pour la commune de Caen

- Sélectionner pour chaque contrat son identifiant, sa surface et sa commune depuis la table « contrat »
- Jointure de la table « region » pour associer le nom de la commune à chaque contrat
- Filtrer les contrats dont la communes est Caen

	contrat_ID	surface	com_nom_maj_court
1	103791	35	CAEN
2	103792	99	CAEN
3	103793	40	CAEN
4	103794	20	CAEN

Requête 3 : Lister le nom des régions de France

```
SELECT DISTINCT reg_nom
FROM region;
```

- Sélectionner depuis la table « region » le nom des régions de chaque commune
- Commande « Distinct » pour n'afficher chaque nom de région qu'une seule fois

	reg_nom
1	Auvergne-Rhône-Alpes
2	Hauts-de-France
3	Provence-Alpes-Côte d'Azur
4	Grand Est
5	Occitanie
6	Normandie
7	Nouvelle-Aquitaine
8	Centre-Val de Loire
9	Bourgogne-Franche-Comté
10	Bretagne

```
Pays de la Loire
lle-de-France
Guadeloupe
Martinique
Guyane
La Réunion
Collectivités d'outre-mer
Mayotte
Gorse
```

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

- Sélectionner pour chaque contrat son identifiant, le type de contrat et la formule depuis la table « contrat »
- Jointure de la table « region » pour associer le numéro de département à chaque contrat
- Filtrer les contrats dont le département est le 71

	contrat_Id	type_contrat	Formule	1 [contrat_Id	type_contrat	Formule
1	114767	Residence secondaire	Classique	21	114787	Residence secondaire	Classiqu
2	114768	Residence principale	Integral	22	114788	Residence principale	Classiqu
3	114769	Residence principale	Classique	23	114789	Residence principale	Classiqu
4	114770	Residence principale	Classique	24	114790	Residence principale	Integral
5	114771	Residence principale	Integral	25	114791	Residence principale	Integral
6	114772	Residence principale	Classique	26	114792	Mise en location	Integral
7	114773	Residence principale	Integral	27	114793	Residence principale	Classiqu
В	114774	Residence principale	Classique	28	114794	Residence secondaire	Classiqu
9	114775	Residence principale	Integral	29	114795	Residence secondaire	Integral
10	114776	Residence principale	Classique	30	114796	Residence principale	Classiqu
11	114777	Residence principale	Integral	31	114797	Residence principale	Classiqu
12	114778	Residence principale	Classique	32	114803	Residence principale	Integral
13	114779	Residence principale	Classique	33	114804	Residence secondaire	Classiqu
14	114780	Residence principale	Integral	34	114805	Mise en location	Integral
15	114781	Residence secondaire	Integral	35	114806	Residence principale	Classiqu
16	114782	Residence principale	Classique	36	114807	Residence principale	Integral
17	114783	Mise en location	Integral	37	114808	Residence principale	Integral
18	114784	Residence principale	Classique	38	114809	Residence principale	Classiqu
19	114785	Residence principale	Integral	39	114810	Residence principale	Integral
20	114786	Residence principale	Classique	40	114811	Residence principale	Integral

41	114812	Residence principale	Integral
42	114813	Mise en location	Classique
43	115182	Residence principale	Classique
44	115207	Residence principale	Integral
45	115221	Residence principale	Integral
46	115223	Residence principale	Classique
47	115256	Residence principale	Classique
48	115259	Residence principale	Classique

5. Interrogation des tables par des requêtes SQL 2/4

Requête 4 : 5 contrats qui ont les surfaces les plus élevées

```
SELECT contrat_id,
surface
FROM contrat
ORDER BY surface DESC
LIMIT 5;
```

- Sélectionner pour chaque contrat son identifiant et la formule depuis la table « contrat »
- Trier par surface décroissante
- Sélectionner les 5 premiers contrats

Г	Contrat_id	surface
1	104211	815
2	105463	742
3	130878	595
4	100822	570
5	109872	559

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

- Sélectionner la moyenne de la colonne « prix_cotisation_mensuel » dans la table « contrat »
- Commande « Distinct » pour ne compter chaque contrat qu'une seule fois

```
prix_moyen
1 19.32869622548212
```

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

```
SELECT valeur_declaree_biens,

Count(DISTINCT(contrat_id)) AS nombre_de_contrats

FROM contrat

GROUP BY valeur_declaree_biens;
```

- Grouper les lignes de la table « contrat » par « valeur_declaree_biens »
- Compter le nombre de contrats distincts dans chaque groupe créé
- Sélectionner la donnée « valeur_declaree_biens » pour chaque décompte du nombre de contrats

	valeur_declaree_biens	nombre_de_contrats
1	0-25000	22720
2	100000+	104
3	25000-50000	6815
4	50000-100000	696

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

- Jointure de la table « region » pour associer le nom de région à chaque contrat
- Filtrer les contrats de formules « integral » et les contrats de la région « Pays de la Loire »
- Compter les contrats distincts correspondants et afficher la région et la formule

	reg_nom	Formule	nombre
1	Pays de la Loire	Integral	589

5. Interrogation des tables par des requêtes SQL 3/4

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

- Jointure de la table « region » pour associer le numéro de département à chaque ligne de la table « contrat »
- Filtrer les contrats du département 71 associés à des maisons
- Sélectionner les contrats distincts correspondants, et afficher le code département, le type de local, le type de contrat et leur formule

	dep_code	Type_local	contrat_Id	type_contrat	formule
1	71	Maison	114768	Residence principale	Integral
2	71	Maison	114779	Residence principale	Classique
3	71	Maison	114782	Residence principale	Classique
4	71	Maison	114812	Residence principale	Integral

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

```
SELECT r.dep_nom,

Avg(surface) AS surface_moyenne

FROM contrat c

INNER JOIN region r

ON r.code_dep_code_commune = c.code_dep_code_commune

WHERE r.dep nom = 'Paris';
```

- Jointure de la table « region » pour associer le nom département à chaque contrat
- Filtrer les contrats du département de Paris
- Afficher la moyenne de la surface des contrats correspondants et le nom du département

```
dep_nom surface_moyenne 
1 Paris 51.76954988591579
```

Requête 10 : Classements des 10 départements où le prix moyen de la cotisation est le plus élevé

- Jointure de la table « contrat » (réduite aux données prix_cotisation_mensuel et code_dep_code_commune, avec contrats distincts) pour associer le nom et code de département à chaque contrat
- Grouper les contrats par département
- Calculer le prix moyen de cotisation pour chaque département
- Trier par ordre décroissant et afficher les 10 premières valeurs de cotisation moyenne, avec nom et numéro de département correspondant

	dep_code	dep_nom	cotisation_moyenne
1	75	Paris	36.39908732628086
2	92	Hauts-de-Seine	26.26935840707965
3	94	Val-de-Marne	19.82416396979504
4	78	Yvelines	18.88806970509383
5	69	Rhône	18.48551829268293
6	1	Ain	18.23888888888889
7	6	Alpes-Maritimes	18.13762287756926
8	17	Charente-Maritime	17.31730769230769
9	74	Haute-Savoie	17.14873417721519
10	2A	Corse-du-Sud	17.07438016528926

5. Interrogation des tables par des requêtes SQL 4/4

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

- Jointure de la table « region » pour associer le nom et le code de commune à chaque contrat
- Grouper les contrats par commune et compter les éléments distincts
- Filtrer les communes dont le nombre de contrat est supérieur à 150
- Trier par ordre décroissant

2 75117 PARIS 17 468 3 75115 PARIS 15 407 4 75116 PARIS 16 394 5 6088 NICE 387 6 75111 PARIS 11 381 7 33063 BORDEAUX 302 8 75120 PARIS 20 302 9 44109 NANTES 291 10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161				
2 75117 PARIS 17 468 3 75115 PARIS 15 407 4 75116 PARIS 16 394 5 6088 NICE 387 6 75111 PARIS 11 381 7 33063 BORDEAUX 302 9 44109 NANTES 291 10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161		code dep code com	com nom maj coi	nombre contra
3 75115 PARIS 15 407 4 75116 PARIS 16 394 5 6088 NICE 387 6 75111 PARIS 11 381 7 33063 BORDEAUX 302 8 75120 PARIS 20 302 9 44109 NANTES 291 10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	1	75118	PARIS 18	515
4 75116 PARIS 16 394 5 6088 NICE 387 6 75111 PARIS 11 381 7 33063 BORDEAUX 302 8 75120 PARIS 20 302 9 44109 NANTES 291 10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	2	75117	PARIS 17	468
5 6088 NICE 387 6 75111 PARIS 11 381 7 33063 BORDEAUX 302 8 75120 PARIS 20 302 9 44109 NANTES 291 10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	3	75115	PARIS 15	407
6 75111 PARIS 11 381 7 33063 BORDEAUX 302 8 75120 PARIS 20 302 9 44109 NANTES 291 10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	4	75116	PARIS 16	394
7 33063 BORDEAUX 302 8 75120 PARIS 20 302 9 44109 NANTES 291 10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	5	6088	NICE	387
8 75120 PARIS 20 302 9 44109 NANTES 291 10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	6	75111	PARIS 11	381
9 44109 NANTES 291 10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	7	33063	BORDEAUX	302
10 75119 PARIS 19 266 11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	8	75120	PARIS 20	302
11 75110 PARIS 10 263 12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	9	44109	NANTES	291
12 75112 PARIS 12 252 13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	10	75119	PARIS 19	266
13 75114 PARIS 14 222 14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	11	75110	PARIS 10	263
14 38185 GRENOBLE 220 15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	12	75112	PARIS 12	252
15 75109 PARIS 9 204 16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	13	75114	PARIS 14	222
16 31555 TOULOUSE 187 17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	14	38185	GRENOBLE	220
17 83137 TOULON 170 18 92026 COURBEVOIE 163 19 59350 LILLE 161	15	75109	PARIS 9	204
18 92026 COURBEVOIE 163 19 59350 LILLE 161	16	31555	TOULOUSE	187
19 59350 LILLE 161	17	83137	TOULON	170
75400	18	92026	COURBEVOIE	163
20 75103 PARIS 3 159	19	59350	LILLE	161
	20	75103	PARIS 3	159

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

- Jointure de la table « region » pour associer le nom et le code de région à chaque contrat
- Grouper les contrats par région et compter les éléments distincts
- Afficher pour chaque région son code, son nom et le nombre de contrats

	reg_code	reg_nom	nombre_contrat
1	2	Martinique	73
2	3	Guyane	37
3	4	La Réunion	17
4	11	lle-de-France	14177
5	24	Centre-Val de Loire	598
6	27	Bourgogne-Franche-Comté	293
7	28	Normandie	824
8	32	Hauts-de-France	1189
9	44	Grand Est	769
10	52	Pays de la Loire	1196
11	53	Bretagne	947
12	75	Nouvelle-Aquitaine	2038
13	76	Occitanie	1609
14	84	Auvergne-Rhône-Alpes	3042
15	93	Provence-Alpes-Côte d'Azur	3279
16	94	Corse	247