

# Création et utilisation de la base de données

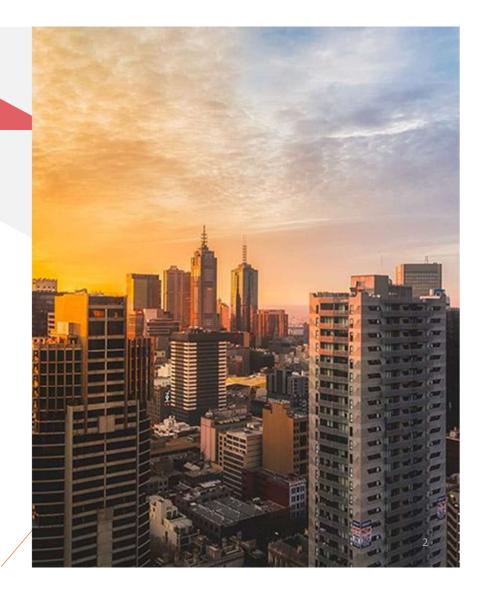
Candice Plainfossé OpenClassRooms – équipe projet DATAImmo



## Contexte du projet

#### **DATAImmo**

- Projet stratégique interne visant à créer un modèle pour mieux prévoir le prix de vente des biens immobiliers
- Conduite de projet : Clara Daucourt (CTO)
- Phase de POC (proof of concept) sur les données du 1<sup>er</sup> semestre 2020



## Les données initiales

• Composition des données initiales à notre disposition

Nom du fichier	Nombre de colonnes	Nombres de lignes (en- tête exclus)	Origine
Valeurs-foncières	47	31 169	Données extraites du site opendata des Demandes de valeurs foncières (DVF) issues de la DGFiP.
fr-esr-referentiel-geographique	37	38 916	Données de data.gouv sur les régions : référentiel géographique français, communes, unités urbaines, aires urbaines, départements, académies, régions.
donnees_communes	10	34 991	Données de l'INSEE avec les résultats des recensements de la population des communes.



## La stratégie de sauvegarde et la conformité RGPD

#### Mise en conformité RGPD

Protection des données personnelles : le fichier Valeurs-foncières contient des données personnelles (nom de l'acquéreur). → la colonne « Acquéreur » n'a pas été conservé dans le projet

#### Stratégie de sauvegarde à élaborer

- Evaluer les risques
- Définir les fréquences et sélectionner les modes de sauvegarde des données
- Automatiser les processus
- Définir les modalités de récupération des données
- Tester régulièrement la récupération
- Maintenir et mettre à jour



## Exemple de dictionnaire des données : Ventes

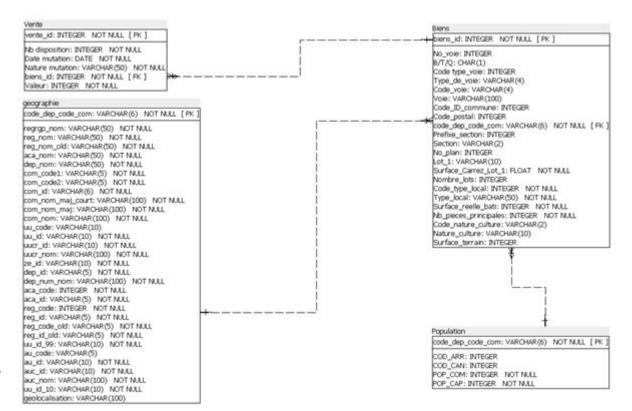
#### **DICTIONNAIRE DES DONNÉES - Ventes**

CODE	SIGNIFICATION	TYPE	LONGUEUR	NATURE	REGLE DE GESTION	REGLE DE CALCUL
vente_id	identifiant d'une mutation - clé primaire	INT		Elementaire	NOT NULL	Auto-incrémentation
No_disposition	Numéro de disposition (mutations) du bien	INT		Elementaire	NOT NULL	
Date_mutation	Date de la mutation	DATE		Elementaire	format aaaa/mm/jj - NOT NULL	
Nature_mutation	Vente, vente en l'état futur d'achèvement, vente de terrain à bâtir, adjudication, expropriation ou échange	VARCHAR	50	Elementaire	NOT NULL	
bien_id	Identifiant unique des biens - Clé étrangère table Biens	INT		Elémentaire	NOT NULL	
	Montant de l'évaluation déclaré dans le cadre d'une			Elementaire		La valeur foncière est le prix net vendeur. La TVA est incluse. Les frais de notaires sont exclus. Valeurs manquantes remplacées sur la base du prix moyen (table entière sauf Paris calcul
Valeur	mutation à titre onéreux.	INT			NOT NULL	à l'arrondissement) au m² x surface Carrez

Dictionnaire des données	Significations
CODE	Nom de la ligne dans le fichier Initial
SIGNIFICATION	Signification de la colonne
TYPE DE VARIABLES	Varchar, Integer, Date, Float, etc.
LONGUEUR	Longueur maximale de la données (pas obligatoire)
NATURE (E/Ca/CO)	Elémentaire, calculer ou concaténer
REGLE DE GESTION	Format de la date (jj/mm/aaaa, jj/mm/aa, etc), ne doit pas etre nul, etc.
REGLE DE CALCUL	exemple : Colonne A + Colonne B



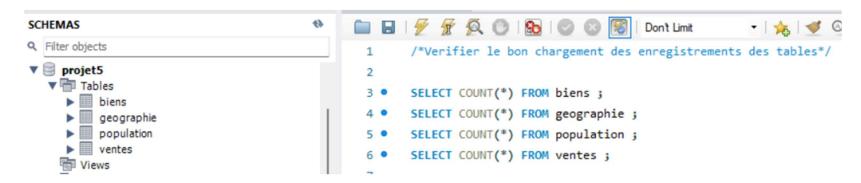
### Le schéma relationnel normalisé



Réalisé avec SQL Power Architect



## La base de données avec les tables créées et les données chargées



L'ensemble des données ont bien été chargées dans MySQL Workbench.

Tables	Nb_enregistrements
biens	34 169
geographie	38 916
population	34 991
ventes	34 169

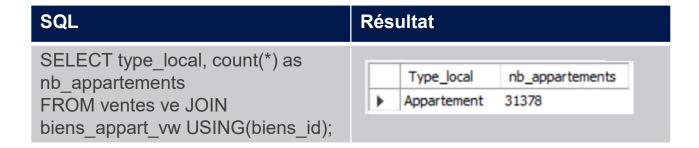




## Requêtes SQL et résultats

Datalmmo au 1<sup>er</sup> semestre 2020 : une activité immobilière centrée sur les transactions d'appartements avec une forte présence de l'Ile-de-France

#### Nombre total d'appartements vendus au 1er semestre 2020



• Le nombre d'appartements vendus représentent 91,8% des ventes.



#### Le nombre de ventes d'appartement par région pour le 1er semestre 2020

• 44,6% des appartements vendus se situent en Ile-de-France.



#### Proportion des ventes d'appartements par le nombre de pièces

SQL	Résultat	
SELECT Nb_pieces_principales,	Nb_pieces_principales	
round(count(Nb_pieces_principales)*100/(SELECT count(*)	2 3	31.178 28.574
FROM biens_appart_vw),3) as proportion_nb_pieces	1	21.477
ROM ventes JOIN biens_appart_vw USING(biens_id)	4	14.214
GROUP BY Nb_pieces_principales	5	3.550
ORDER BY proportion nb pieces DESC;	6	0.650
, , , , , , , , , , , , , , , , , , ,	7	0.172
	0	0.096
	8	0.054
	9	0.025
	10	0.006
	11	0.003

• 95,4% des appartements vendus ont entre 1 et 4 pièces.



#### Liste des 10 départements où le prix du mètre carré est le plus élevé

SQL
SELECT dep_num_nom, round(avg(Valeur/Surface_Carrez_Lot_1),2) as prix_metre_carre FROM ventes ve JOIN biens bi USING(biens_id) JOIN geographie geo USING(code_dep_code_com) GROUP BY dep_num_nom ORDER BY prix_metre_carre DESC LIMIT 10;

- Paris est le département où le prix du mètre carré est le plus élevé (1,6 fois supérieur au 2ème département (Hauts-de-Seine)).
- 5 des 8 départements composant l'Ile-de-France font partie du top 10 des départements où le prix du mètre carré est le plus élevé.



#### Prix moyen du mètre carré d'une maison en Île-de-France

SQL	Résultat
SELECT reg_nom, type_local, round(avg(Valeur/Surface_Carrez_Lot_1),2) AS prix_moyen_metre_carre FROM geographie geo JOIN biens_maisons_vw USING(code_dep_code_com) JOIN ventes ve USING(biens_id) WHERE reg_nom = 'lle-de-France';	reg_nom Type_local prix_moyen_metre_carre Ile-de-France Maison 3745.09

• Le prix moyen du mètre carré d'une maison en Ile-de-France est inférieur au prix du m² des appartements de 5 départements de cette région.



#### Liste des 10 appartements les plus chers avec la région et le nombre de mètres carrés

SQL	Résultat		
SELECT reg_nom, Surface_Carrez_Lot_1, Valeur FROM geographie geo JOIN biens_appart_vw USING(code_dep_code_com) JOIN ventes ve USING(biens_id) ORDER BY Valeur DESC LIMIT 10;	reg_nom  Ile-de-France	64 20.55 42.77 253.3 139.9 360.95 595	Valeur 9000000 8600000 8577713 7620000 7600000 7535000 7420000 7200000 7050000 6600000

- Grande variété des superficies des appartements du top 10 des appartements les plus chers (9m² le plus petit et 595m² le plus grand).
- Superficie notablement petite de l'appartement le plus cher : 9 millions d'euros pour 9m².
- Tous ces biens se situent en Ile-de-France.



#### Taux d'évolution du nombre de ventes entre le premier et le second trimestre de 2020

SQL	Résultat	
WITH ventes_trim_2 AS ( SELECT COUNT(*) AS nb_ventes FROM ventes WHERE Date_mutation BETWEEN '2020/04/01' AND '2020/06/30' AND Nature_mutation = 'Vente'),		
ventes_trim_1 AS (	Nature_mutation	Evolution_2020_trim1_2
SELECT COUNT(*) AS nb_ventes FROM ventes WHERE Date_mutation BETWEEN '2020/03/31' AND '2020/01/01' AND Nature_mutation = 'Vente')	Vente	1.00
SELECT Nature_mutation, round(count(ventes_trim_2.nb_ventes - ventes_trim_1.nb_ventes) / count(ventes_trim_1.nb_ventes * 100),2) AS Evolution_2020_trim1_2 FROM ventes, ventes_trim_2, ventes_trim_1 GROUP BY Nature_mutation;		

• Le nombre de ventes a augmenté de 1% entre le 1<sup>er</sup> et le 2<sup>nd</sup> trimestre de 2020 malgré la crise COVID.



## Le classement des régions par rapport au prix au mètre carré des appartements de plus de 4 pièces

SQL	Résultat		
SELECT reg_nom, Type_local, round(avg(Valeur/Surface_Carrez_Lot_1),2) AS prix_metre_carre FROM geographie geo JOIN biens_appart_vw USING(code_dep_code_com) JOIN ventes ve USING(biens_id) WHERE Nb_pieces_principales > 4 GROUP BY reg_nom ORDER BY prix_metre_carre DESC;	reg_nom  Ile-de-France La Réunion Provence-Alpes-Côte d'Azur Corse Auvergne-Rhône-Alpes Nouvelle-Aquitaine Bretagne Pays de la Loire Hauts-de-France Occitanie Normandie Grand Est Centre-Val de Loire Bourgogne-Franche-Comté Martinique	Type_local  Appartement	2412.05

- L'Ile-de-France reste la région où les appartement (> 4 pièces) sont largement les plus couteux.
- L'île de la Réunion est en 2<sup>nde</sup> position mais le prix moyen est près de 2,5 fois inférieur à celui de l'île-de-France.



#### Liste des communes ayant eu au moins 50 ventes au 1er trimestre

SQL	Résultat
WITH ventes_trim_1 AS ( SELECT COUNT(*) AS nb_ventes FROM ventes	152 communes sont sorties de cette requête.
WHERE Date_mutation BETWEEN '2020/03/31' AND '2020/01/01' AND Nature_mutation = 'Vente' ) SELECT com_nom,geo.code_dep_code_com, Nature_mutation, count(ventes_trim_1.nb_ventes) AS nb_ventes FROM ventes_trim_1, geographie geo JOIN biens USING(code_dep_code_com) JOIN ventes USING(biens_id) GROUP BY com_nom, geo.code_dep_code_com, Nature_mutation HAVING nb_ventes >= 50 ORDER BY nb_ventes DESC;	Le nombre de ventes s'étale de 50 (Bourges, Thiais, Vanves) à 516 ventes (Paris 18ème)

• 7 des 10 premières communes sont des arrondissements parisiens.



#### Différence en % du prix au mètre carré entre un appartement de 2 et un de 3 pièces

SQL	Résultat
WITH prix_appart_2p AS ( SELECT round(avg(Valeur/Surface_Carrez_Lot_1),2) AS prix2_m2 FROM ventes ve JOIN biens_appart_vw bi USING(biens_id) WHERE bi.Nb_pieces_principales = 2 ), prix_appart_3p AS ( SELECT round(avg(Valeur/Surface_Carrez_Lot_1),2) AS prix3_m2 FROM ventes ve JOIN biens_appart_vw bi USING(biens_id) WHERE bi.Nb_pieces_principales = 3 ) SELECT round(100.0 * (prix3_m2 - prix2_m2) / prix2_m2,2) AS difference_prix_2et3p FROM prix_appart_2p, prix_appart_3p;	difference_prix_2et3p -12.4

• Le prix au m² d'un appartement 3 pièces est inférieur de 12,4% au prix au m² d'un 2 pièces.



#### Moyennes des valeurs foncières pour le top 3 des communes des départements 6, 13, 33, 59 et 69

SQL	Résultat			
WITH valeur_moyenne_communes AS ( SELECT g.dep_num_nom, g.dep_id, g.code_dep_code_com, g.com_nom, round(AVG(v.Valeur),2) AS moyenne_valeur_fonciere FROM ventes v JOIN biens b USING(biens_id) JOIN geographie g USING(code_dep_code_com) WHERE g.dep_id IN ("D006","D013","D033","D059","D069") GROUP BY g.dep_num_nom, g.code_dep_code_com, g.com_nom ) SELECT dep_num_nom AS Departement, code_dep_code_com AS Identifiant, com_nom AS Communes, moyenne_valeur_fonciere FROM (SELECT dep_num_nom, code_dep_code_com, com_nom, moyenne_valeur_fonciere, RANK() OVER (PARTITION BY dep_num_nom ORDER BY moyenne_valeur_fonciere DESC) AS classement FROM valeur_moyenne_communes) AS test WHERE classement <= 3;	Departement  06 - Alpes-Maritimes  06 - Alpes-Maritimes  06 - Alpes-Maritimes  13 - Bouches-du-Rhône  13 - Bouches-du-Rhône  13 - Gironde  33 - Gironde  33 - Gironde  59 - Nord  59 - Nord  59 - Nord  69 - Rhône  69 - Rhône  69 - Rhône	Identifiant 06121 06059 06084 13043 13101 13022 33236 33539 33009 59071 59168 59279 69265 69382 69386	Communes Saint-Jean-Cap-Ferrat Eze Mouans-Sartoux Gignac-la-Nerthe Saint-Savournin Cassis Lège-Cap-Ferret Vayres Arcachon Bersée Cysoing Halluin Ville-sur-Jarnioux Lyon 2e Lyon 6e	moyenne_valeur_fonciere 968750.00 655000.00 476898.13 330000.00 314425.00 313416.88 549500.64 335000.00 307435.93 433202.00 408550.00 322250.00 485300.00 455217.26 426968.25

• Forte disparité entre les départements et, parfois, entre les communes de chaque top 3.



## Les 20 communes avec le plus de transactions pour 1000 habitants pour les communes qui dépassent les 10 000 habitants

SQL	Résultat				
CREATE VIEW communes_pop10kplus AS SELECT * FROM population WHERE POP_COM+POP_CAP > 10000;  SELECT p.code_dep_code_com, g.com_nom, round((count(v.Valeur)/(p.POP_COM + p.POP_CAP)*1000),2) AS transac_pour_mille_hab FROM communes_pop10kplus p JOIN biens USING(code_dep_code_com) JOIN Ventes v USING(biens_id) JOIN geographie g USING(code_dep_code_com) GROUP BY p.code_dep_code_com, g.com_nom ORDER BY transac_pour_mille_hab DESC LIMIT 20;	code_dep_code_com 75102 75101 75103 33009 44055 75104 06104 75108 83123 83071 75109 75106 83112 60141 44132 94067 75110 06083 85226 94080	com_nom  Paris 2e  Paris 1er  Paris 3e  Arcachon  La Baule-Escoublac  Paris 4e  Roquebrune-Cap-Martin  Paris 8e  Sanary-sur-Mer  La Londe-les-Maures  Paris 6e  Saint-Cyr-sur-Mer  Chantilly  Pornichet  Saint-Mandé  Paris 10e  Menton  Saint-Hilaire-de-Riez  Vincennes	transac_pour_mille_hab 5.84 4.92 4.69 4.62 4.58 4.08 3.99 3.83 3.50 3.43 3.43 3.43 3.38 3.24 3.13 3.06 3.06 3.06 3.04 2.94 2.87 2.81		

- 8 des 20 communes sont un arrondissement parisien dont le top 3 (Paris 2, 1 et 3).
- Le nombre de transactions pour 1000 habitants s'étale de 5,84 à 2,81 transactions / 1000 habitants.





#### Merci de votre attention



