Analyse data-centric pour une agence immobilière

Contexte :
Une société immobilière souhaiterait récupérer et stocker des données Open Data des ventes immobilières afin de mener une étude lui permettant d'avoir une meilleure connaissance du marché en utilisant l'IA dans un futur proche. En amont de cela, elle souhaiterait donc pouvoir créer un flux permettant le stockage de données mais aussi des analyses de l'historique.

Rappel des objectifs :
En tant que dev IA, - Analyse du besoin

- Roadmap et Backlog
- Création des maquettes et prototypage
- Comprendre une architecture logicielle
- Récupération de la data
- Automatiser les mises à jours régulières de la data
- Création de bases de données
- Définition de requêtes SQL
- Affichage dynamique de graphes

Pour ce faire, vous devrez effectuer le pré-traitement des données en Python puis faire le stockage dans la base de données que vous aurez conceptualisée et modélisée en amont.

Une fois la base de données constituée, vous répondrez aux requêtes posées mais aussi produirez des analyses et des graphiques.

1. Création du dictionnaire de données

Code mnémonique	Désignation	Type	Taille	Remarque
num_disposition	indique le nombre de transactions sur le bien	INT	1	
date_mutation	Date vente	DATE (JJ-MM-AAAA)	10	
nature_mutation	Type de vente	VARCHAR	50	
valeur_fonciere	Indique la valeur du bien	FLOAT	13	
num_voie	Numéro de la rue	FLOAT	4	
BTQ	Désigne un complément d'adresse	VARCHAR	50	
type_voie	Désigne le type de voie (Rue, route,..)	VARCHAR	20	
code_voie	Identifiant de voie	CHAR	4	
commune	Ville	VARCHAR	50	
voie	nom de la rue	VARCHAR	50	
code_postal	code_postal (ex: 69100)	INT	5	
code_commune	Identifiant de la commune; 3 derniers chiffres du cc	INT	3	
code_departement	Numéro départementale	INT	2	
section	Désignation du cadastre	VARCHAR	5	
num_plan	Numéro de plan cadastral	INT	3	
num_volume	Numéro de volume de plan cadastral	FLOAT		
premier_lot	Désignation d'un bien	FLOAT	7	
surface_premier_lot	Surface du bien	FLOAT	7	
deuxieme_lot	Désignation d'un bien	FLOAT	6	
surface_deuxieme_lot	Surface du bien	FLOAT	7	
troisieme_lot	Désignation d'un bien	FLOAT	6	
surface_troisieme_lot	Surface du bien	FLOAT	7	
quatrieme_lot	Désignation d'un bien	FLOAT		
surface_quatrieme_lot	Surface du bien	FLOAT	7	
cinquieme_lot	Désignation d'un bien	FLOAT		
surface_cinquieme_lot	Surface du bien	FLOAT	7	
nbr_lots	Nombres de biens	INT		
code_type_local	Chiffre correspondant au type de local	INT		
type_local	Utilité du local (garage, logement,...)	VARCHAR	50	
surface_reelle	Surface total, y comprit la partie non exploitable	FLOAT		
nb_piece_principales	Nombres de pièce dites principale d'un bien	INT		
nature_culture		VARCHAR	5	
nature_culture_speciale		VARCHAR	5	
surface_terrain	Surface extérieur d'un logement	FLOAT		

A l'attention d'Alison : Ne pas tenir compte de la taille des types autre que VARCHAR / CHAR

TABLES :

- Lot
- Localisation
- Transaction
- Biens

Description du dictionnaire de données :

La colonne “Code mnémonique” est le nom qui sera donnée à chaque colonne de chaque table pour la base de donnée. Chaque table est déterminée par sa couleur.
La colonne “Désignation” décrit à quoi correspond chaque code mnémonique.
La colonne “Type” permet de savoir de quel type chaque colonne sera.
La colonne “Taille” permet de savoir quelle sera la taille maximale de chaque colonne. Précision : Seul doit être pris en compte la taille des VARCHAR / CHAR.

En grisé, les colonnes ont été retirées suite à délibération de notre groupe.

Processus :

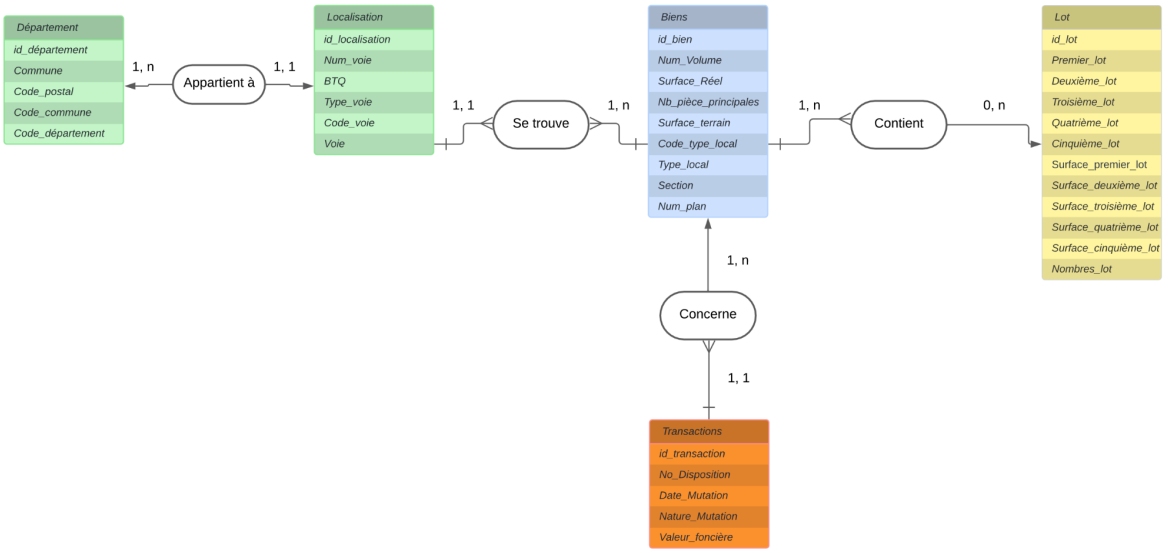
Pour faire ce dictionnaire, nous avons exploré les données via le fichier texte, en allant directement sur le fichier et en l’important sur un notebook avec un dataframe.

2. Nettoyage des données

Pour effectuer ce nettoyage nous avons dû explorer les données pour faire ces opérations :

- Suppression de 7 colonnes dont les valeurs étaient entièrement nulles
- Ajout d'une colonne date en int pour pouvoir mieux les manipuler lors de l'étape de l'analyse de données
- Suppression de 13 colonnes que nous avons jugé peu intéressantes à exploiter par la suite, notamment car il y avait beaucoup de valeurs manquantes
- Nous avons supprimés les doublons
- Utilisation du dropna() sur les colonnes de manière individuels lorsque nécessaire (par exemple pour faire un graphique)

3. Création du MCD



Nous avons décidé d'ajouter à cette étape une table Département.

4. Création de la BBD

Pour créer la base de données nous l'avons fait de 2 façons différentes. Une en MySQL et une en PostgreSQL. Ce choix a été fait suite à une difficulté rencontré par un membre du groupe avec MySQL (connexion perdu à MySQL), et par l'envie de ce membre de s'exercer avec ce SGBD en vue de son alternance où il sera utilisé.

5. Récupération de la data

Création de 5 dataframes corresponds aux 5 tables de la base de données, et envoie vers la BDD.

6. Requêtes SQL

Nous avons dû ajouter les clés primaires et étrangères lors de cette étape.

Résultats des 10 requêtes SQL :

1/ Nombre d'appartements et Maisons vendus en 2020

	Type local	Nombre de ventes
1	Appartement	585512
2	Maison	747946

2/ Nombre de biens vendus par trimestre

	Trimestre	Nombre de biens vendu
1	1	992912
2	2	1102758
3	3	1084883
4	4	1079833

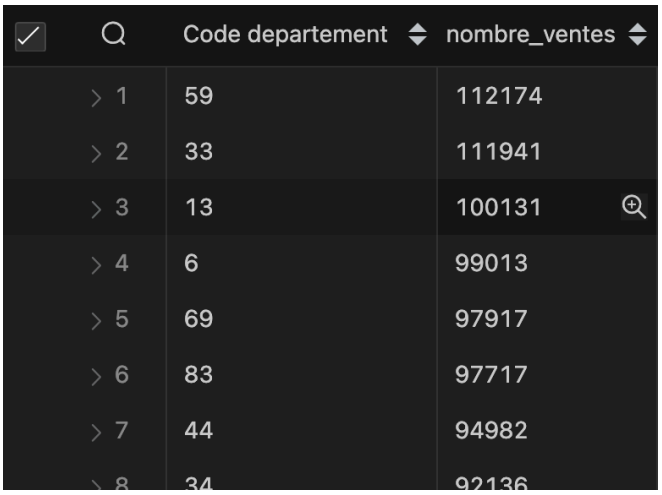
3/ Proportion des ventes de biens par trimestre

	trimestre	nb_vente	proportion
1	1	992912	0.23
2	2	1102758	0.26
3	3	1084883	0.25
4	4	1079833	0.25

4/ Proportion d'appartements vendus par nombre de pièces

	Nombre pieces principa	nb_appartements	proportion
1	0	1418	0.002421812020932107
2	1	109175	0.186460738635587315
3	2	179347	0.306307983440134446
4	3	175250	0.299310688764705078
5	4	90866	0.155190670729207941
6	5	23045	0.039358715107461503
7	6	4489	0.007666794190383800
8	7	1138	0.001943598081678940

5/ Les 10 départements où il y a eu le plus de ventes immobilières

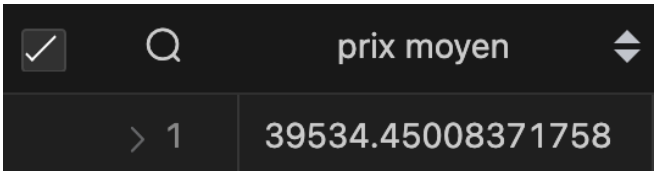


	Code departement	nombre_ventes
1	59	112174
2	33	111941
3	13	100131
4	6	99013
5	69	97917
6	83	97717
7	44	94982
8	34	92136

6/ Les 10 départements où il y en a eu le moins

	Code departement	nombre_ventes
1	973	4486
2	972	6716
3	971	6962
4	90	6985
5	48	9144
6	2A	10595
7	2B	11957
8	15	14616

7/ Prix moyen du mètre carré en IDF



	prix moyen
1	39534.45008371758

8/ Liste des 10 appartements les plus chers avec le département et le nombre de mètres carrés

	id_bien	Code departement	m2	prix
1	3514934	83	80	999999,00
2	3514933	83	65	999999,00
3	4170521	974	62	999990,25
4	4170526	974	28	999990,25
5	4170520	974	56	999990,25
6	4170518	974	49	999990,25
7	4170528	974	48	999990,25
8	4170523	974	44	999990,25

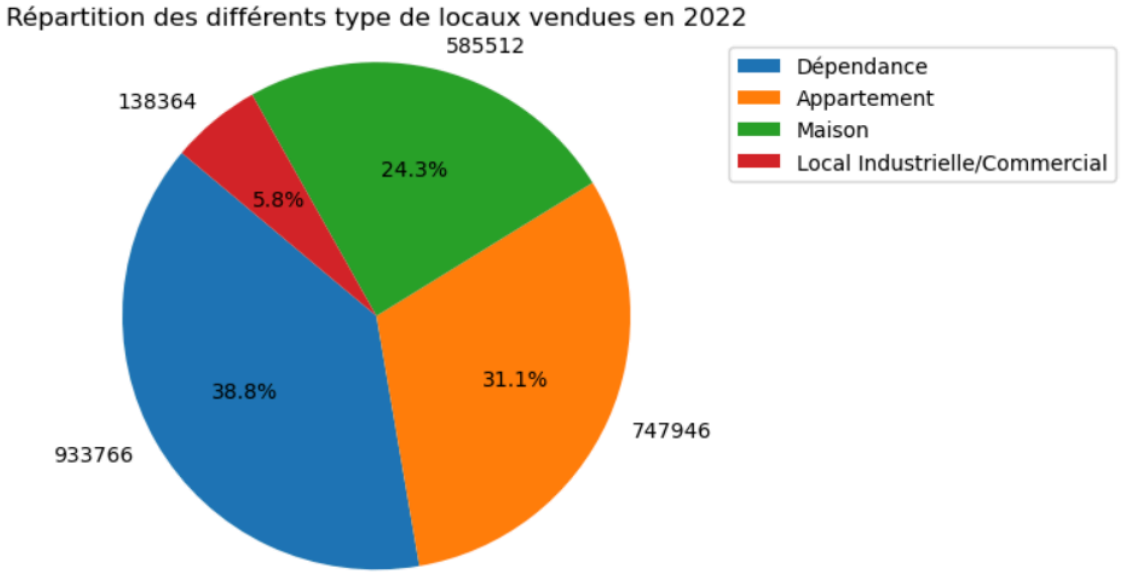
9/ Taux d’évolution du nombre de ventes entre le premier et le second trimestre de 2020

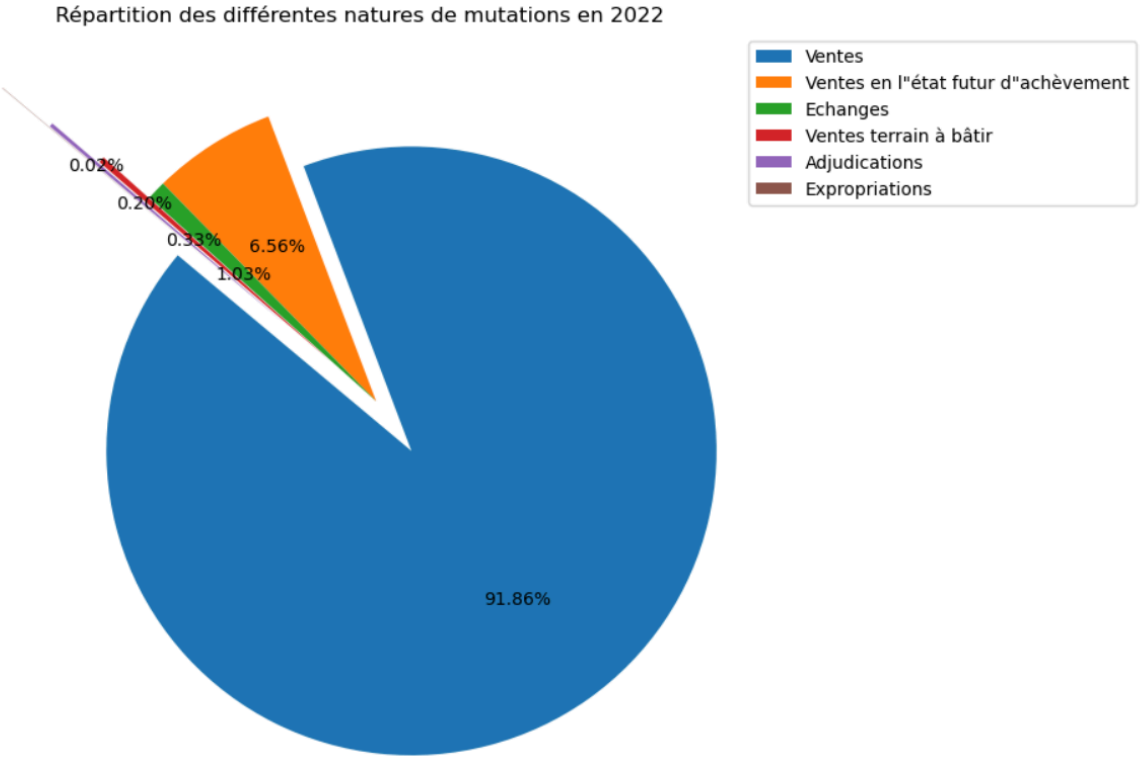
	taux_evolution
1	11.063014647823774917 00

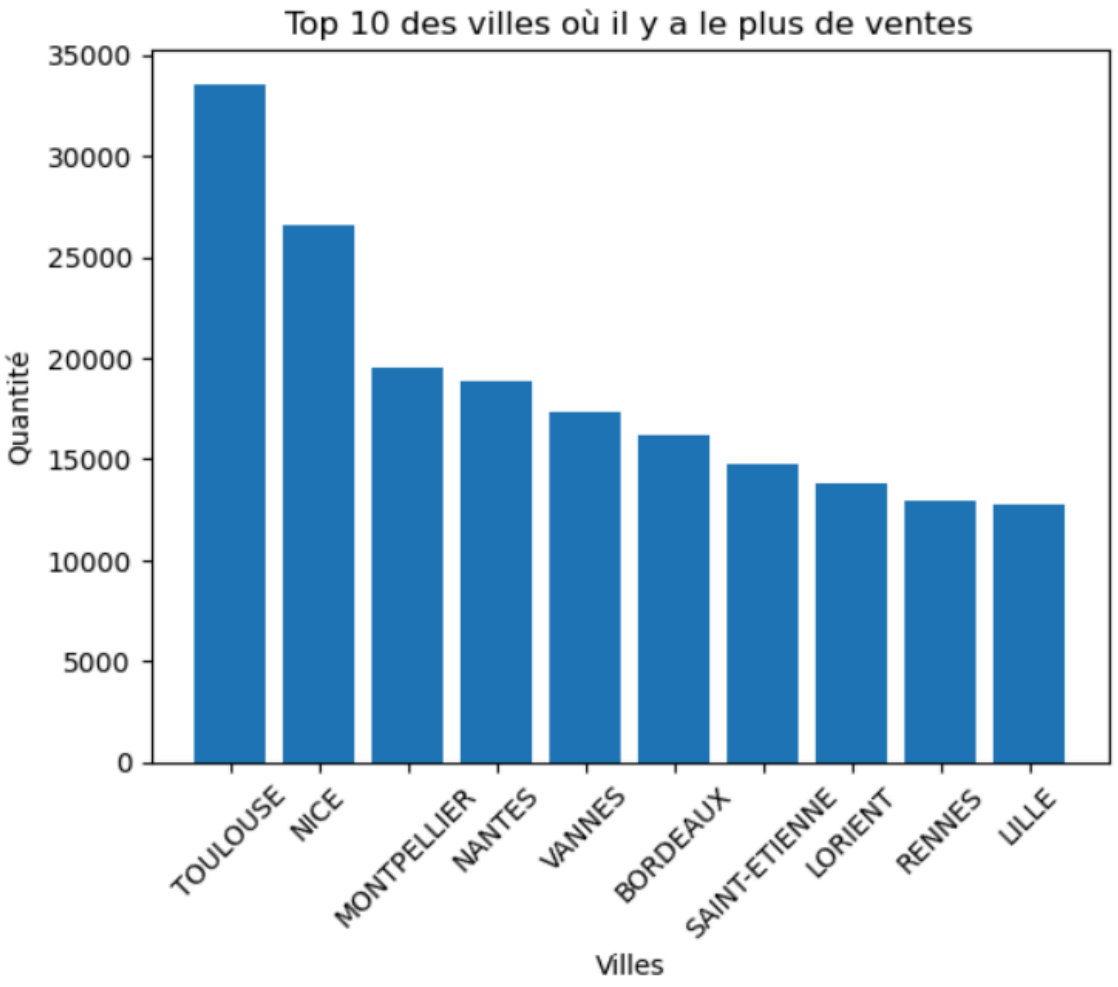
10/ Liste des communes où le nombre de ventes a augmenté d'au moins 20% entre le premier et le second trimestre de 2020

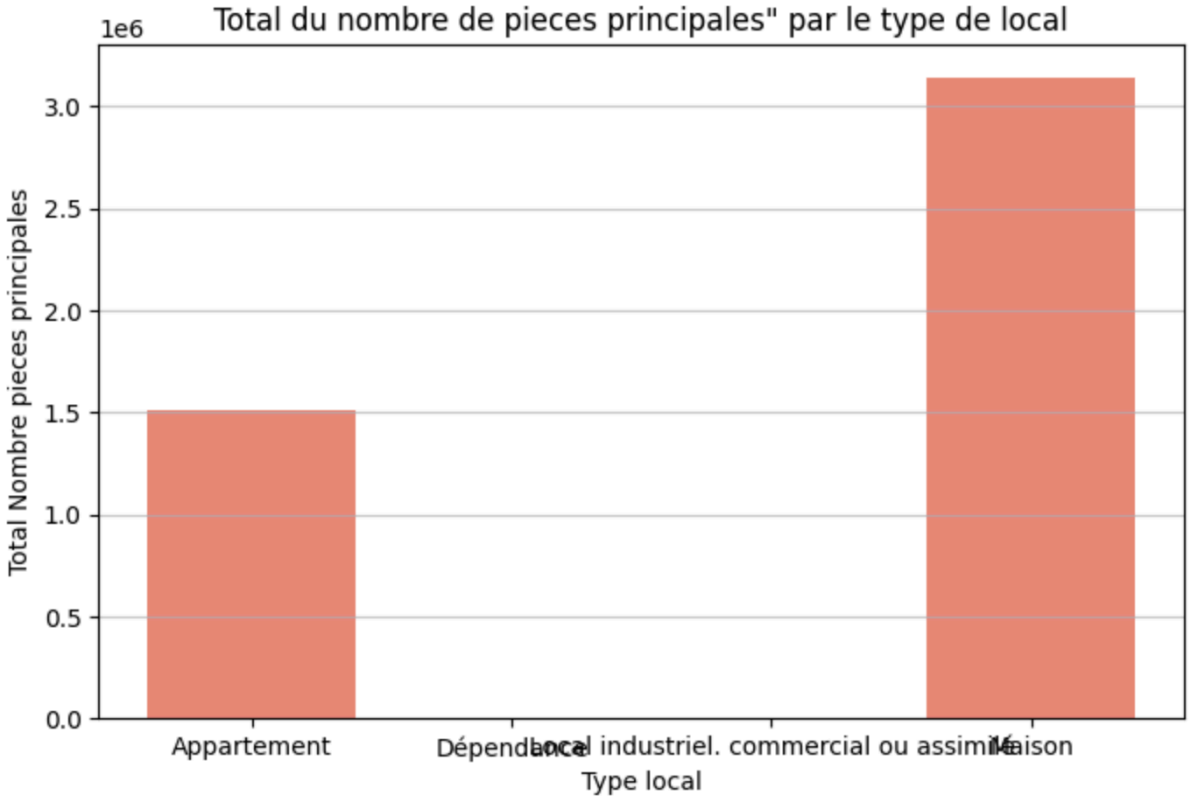
	Commune	pourcentage_augmenta
1	SAINT-MERD-LES-OUSSI	10200
2	SAUMEJAN	9800
3	MONTIERS SUR SAULX	8100
4	MERENS-LES-VALS	7800
5	THORAISE	7700
6	SENNELY	7550
7	LE PIN AU HARAS	7300
8	LOHITZUN-OYHERCQ	6700

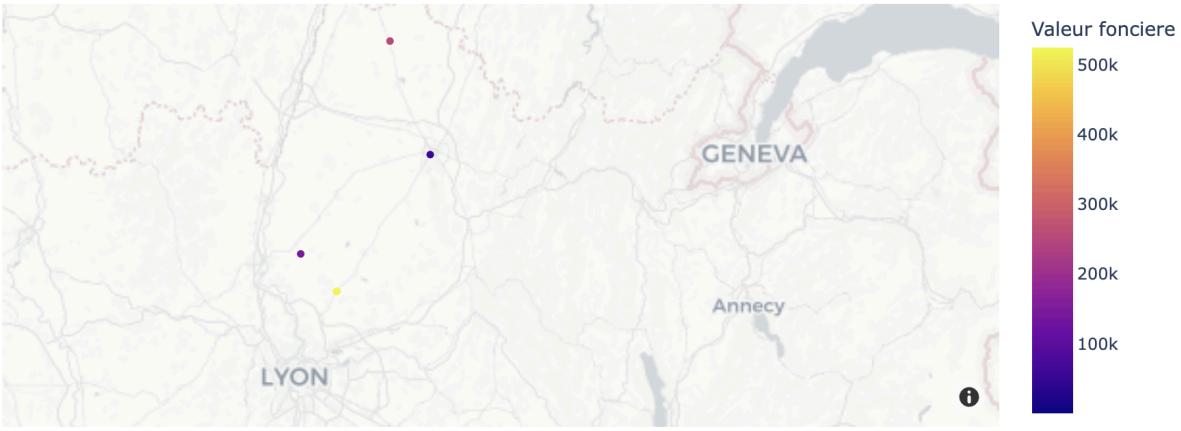
7. Graphique


Répartition des différents type de locaux vendues en 2022
585512
138364
24.3%
5.8%
38.8%
31.1%
933766
747946
Dépendance
Appartement
Maison
Local Industrielle/Commercial


Répartition des différentes natures de mutations en 2022
0.02%
0.20%
0.33%
6.56%
1.03%
91.86%
Ventes
Ventes en l"état futur d"achèvement
Echanges
Ventes terrain à bâtir
Adjudications
Expropriations


Top 10 des villes où il y a le plus de ventes
Quantité
35000
30000
25000
20000
15000
10000
5000
0
TOULOUSE
NICE
MONTPELLIER
NANTES
VANNES
BORDEAUX
SAINT-ETIENNE
LORIENT
RENNES
LILLE
Villes


Total du nombre de pieces principales" par le type de local
1e6
Total Nombre pieces principales
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Appartement
Dépendance
Local industriel. commercial ou assimilé
Maison
Type local



(problème technique qui nous a empêché de mettre davantage de valeurs)