



# Business Intelligence

## Rapport du projet

Kim-Anh Laura Nguyen
Arij Riabi
Promo DAC 2018-2019

*Enseignants :*
Laure Soulier
Benjamin Piwowarski
Clara Gainon de Forsan de Gabriac

# Table des matières

# 1 Présentation de la problématique

## 1.1 Introduction

Vous souhaitez investir dans l'immobilier en France mais l'instabilité du marché vous effraie ? Chez **Estatesy**, nous vous offrons une solution alliant outils de *business intelligence* et savoir-faire immobilier pour vous proposer **les meilleures villes où investir**.

## 1.2 Faits et dimensions

Afin de vous guider vers les meilleures décisions, nous proposons plusieurs analyses sur 40 grandes villes françaises. Notre problématique principale est basée sur le **rendement moyen de l'investissement dans un appartement et/ou une maison**, en fonction de la ville et du temps ; cet indicateur réflète l'attractivité globale du lieu.

De plus, nous rajoutons plusieurs autres faits, qui définissent nos sous-problématiques, toujours selon la ville et la date :

— **l'analyse de la structure de la population** (tranches d'âge, sexe, catégorie socio-professionnelle), qui décrit une grande part de l'environnement

— **l'analyse des services** (le taux d'infrastructures accordées à chaque type de service), qui reflètent la qualité de vie d'un territoire.

# 2 Description du datawarehouse

## 2.1 Données originales

Dans un premier temps, nous récupérons grâce à la transformation `pentaho/exportToCSV_cities_fr.ktr` les 40 villes françaises contenues dans la base `world`, sur lesquelles nous menons nos analyses.

Pour pouvoir décrire l'environnement de ces villes, nous récupérons sur le site de l'INSEE [1] une partie de la base permanente des équipements (BPE), qui fournit le niveau d'équipements et de services rendus à la population sur un territoire. Ces données sont répertoriées dans `data/data_csv`, et leur documentation se trouve dans `data/data_doc`. Nous utilisons :

- `commerce-com-.csv` : nombre de commerces d'un territoire
- `servi-sante-com-.csv` : le nombre d'infrastructures médicales par commune
- `serv-particuliers-com-.csv` : nombre de services aux particuliers par territoire
- `serv-ens-1er-degre-com-.csv`, `serv-ens-2e-degre-com-.csv`, `serv-ens-sup-form-serv-com-.csv` : données propres à l'enseignement pour chaque ville
- `sport-loisir-socio-com-.csv` : nombre d'infrastructures de sport et de divertissement par territoire
- `tour-transp-com-.csv` : nombre de bâtiments liés au transport et à l'accueil des touristes

Nous nous servons également des données de OpenDataSoft [2] relatives à la structure de la population (estimation de la population par tranches d'âge, professions, sexe), ainsi qu'à la densité d'habitants de chaque commune, contenues dans `structure-et-densite-de-la-population-2011.csv`. Ce fichier est légèrement modifié par notre script `scripts/struct_pop_to_int`, afin de convertir les estimations de populations en entier.

Pour récupérer les prix (au $m^2$) à la location et à la vente d'un appartement ou d'une maison dans chacune des 40 villes, nous effectuons un scrapping des données de MeilleursAgents [3] (décrit dans la section 2.2). Ce scrapping donne lieu aux fichiers suivants :

- `prix_locations.csv` contient les prix minimum, maximum, et moyen à la location d'un appartement ou d'une maison dans chaque ville
- `prix_vente.csv` contient les prix minimum, maximum, et moyen à la vente d'un appartement ou d'une maison dans chaque ville

## 2.2 Scrapping des données

Afin de scrapper les données sur l'estimation immobilière, nous définissons les fonctions suivantes, contenues dans `scrapping_prix` :

- `source_meilleursagents.py` : permet de collecter les données sur les prix immobiliers d'une ville ainsi que l'estimation des prix moyen à l'achat et à la location, et contient également des fonctions de formatage des donnes
- `fonction_headers.py` : choisit un user agent à utiliser pour le navigateur
- `fonction_get_session_proxy.py` : contient les fonctions pour ouvrir une nouvelle session de requête avec un nouveau proxy
- `my_controller.py` : choisit un user agent à utiliser pour le navigateur, change l'IP à la prochaine ouverture du navigateur, charge toutes les villes dans un tableau et créé tous les urls

1. https ://www.insee.fr/fr/statistiques/3568599 ?sommaire=3568656
2. https ://public.opendatasoft.com/explore/dataset/structure-et-densite-de-la-population-2011/export/
3. https ://www.meilleursagents.com/

## 2.3 Schémas du datawarehouse

Sur les figures 1, 2 et 3 apparaissent les schémas en étoile respectifs des faits **Analyse des services**, **Rendement de l'investissement**, et **Analyse des catégories socio-professionnelles**. Les dimensions sont représentées par des boîtes vertes, les faits, par des boîtes oranges.


FIGURE 1 – Fait **analyse des services**

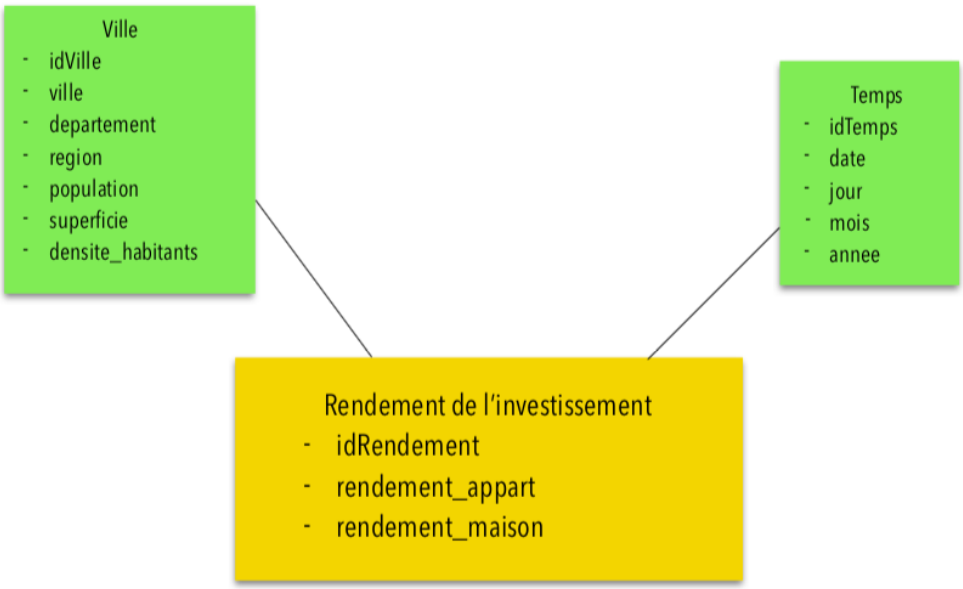



FIGURE 2 – Fait **rendement de l'investissement**


FIGURE 3 – Fait **analyse des catégories socio-professionnelles**

La figure 4 contient le schéma en constellation de notre datawarehouse. Par souci de lisibilité, les attributs de chaque fait et dimension sont retirés de cette représentation.


FIGURE 4 – Schéma en constellation du datawarehouse

## 2.4 Modélisation logique du datawarehouse

Ville(idVille, ville, departement, region, population, superficie, densite_habitants)

Temps(idTemps, date_extraction, jour, mois, annee)

CatAge(idCatAge, idVille*, enfants, jeunes_adultes, adultes, seniors)

Commerce(idCommerce, idVille*, magasin, commerce_detail, commerce_detai_frais, commerce_-divers, station_service)

Ens1(idEns1, idVille*, ecole_maternelle, ecole_elementaire)

Ens2(idEns1, idVille*, college, lycee_general_techno, lycee_pro)

Ens_sup(idEnsSup, idVille*, universite, ecole_ingenieurs, ecole_commerce, classe_preparatoire, restaurant_universitaire, residence_universitaire)

PrixLocation(idPrixLocation, idVille*, date_extraction, prix_appart_min, prix_appart_moyen, prix_appart_max, prix_appartement_min, prix_appartement_moyen, prix_appartement_max)

PrixVente(idPrixLocation, idVille*, date_extraction, prix_appart_min, prix_appart_moyen, prix_appart_max, prix_appartement_min, prix_appartement_moyen, prix_appartement_max)

Profession(idProfession, idVille*, agriculteurs, artisans_commercants, autres, cadres, employes, ouvriers, professions_inter, retraites)

Serv_particuliers(idServPart, idVille*, poste, reparation_travaux, ecole_conduite, pole_emploi, veterinaire, agence_immobiliere, esthetique, securite_justice, banques, restaurant)

Serv_sante(idServSante, idVille*, centre_medical, etablissement_sante, urgences, sante_divers)

Sexe(idSexe, idVille*, femmes, hommes)

Sports_loisirs(idSportsLoisirs, idVille*, cinema, conservatoire, musee, theatre, sports)

Tour_transport(idTourTransport, idVille*, agence_voyage, hotel, aeroport, taxi, camping, gare, information_touristique)

Fact_analyseServices(idAnalyseServices, idVille*, idCommerce*, idServPart*, idServSante*, idSportsLoisirs*, idTourTransport*,idEns1*, idEns2*, idEnsSup*, taux_commerce, taux_servPart, taux_servSante, taux_sportsLoisirs, taux_tourTransport, taux_ens)

Fact_rendement(idRendement, idVille*, idTemps*, rendement_appart, rendement_maison)

Fact_analyseServices(idAnalyseSP, idVille*, idCatAge*, idSexe*, idProfession*, pourcentage_femmes, pourcentage_hommes, pourcentage_enfants, pourcentage_jeunes_adultes, pourcentage_adultes, pourcentage_seniors, pourcentage_agriculteurs, pourcentage_artisans_commercants, pourcentage_autres, pourcentage_cadres, pourcentage_employes, pourcentage_ouvriers, pourcentage_prof_inter, pourcentage_retraites)

# 3 Intégration des données sur Pentaho

## 3.1 Pré-traitements

Les transformations mentionnées dans cette section se trouvent dans le dossier `pentaho`.

Nous récupérons tout d'abord les 40 villes françaises de la base `world`, que nous insérons dans un fichier csv, avec la transformation `exportToCSV_cities_fr`.

Avant d'insérer les données dans le datawarehouse, nous effectuons la jointure de ces villes avec les données de la section 2.1. Pour chaque type d'information, la transformation correspondante produit un fichier csv concernant uniquement les 40 villes récupérées précédemment. Les figures 5, 6, 7 montrent le nettoyage effectué, respectivement, sur les données `serv-particuliers-com-.csv`, `structure-et-densite-de-la-population-2011.csv`, et `prix_vente.csv`.

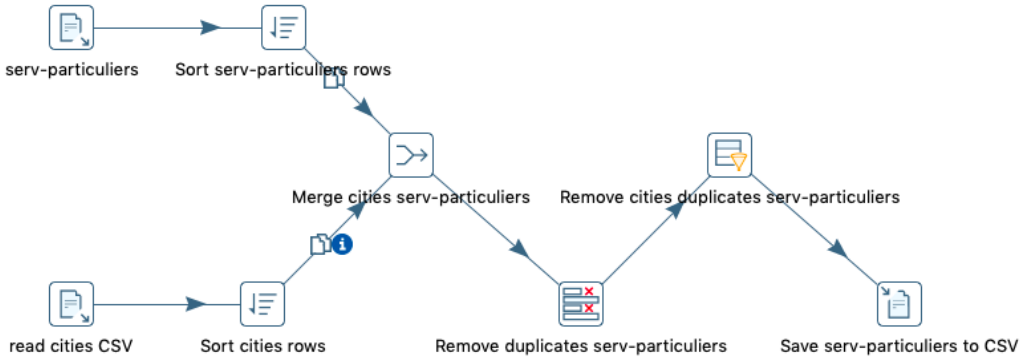



FIGURE 5 – Services aux particuliers

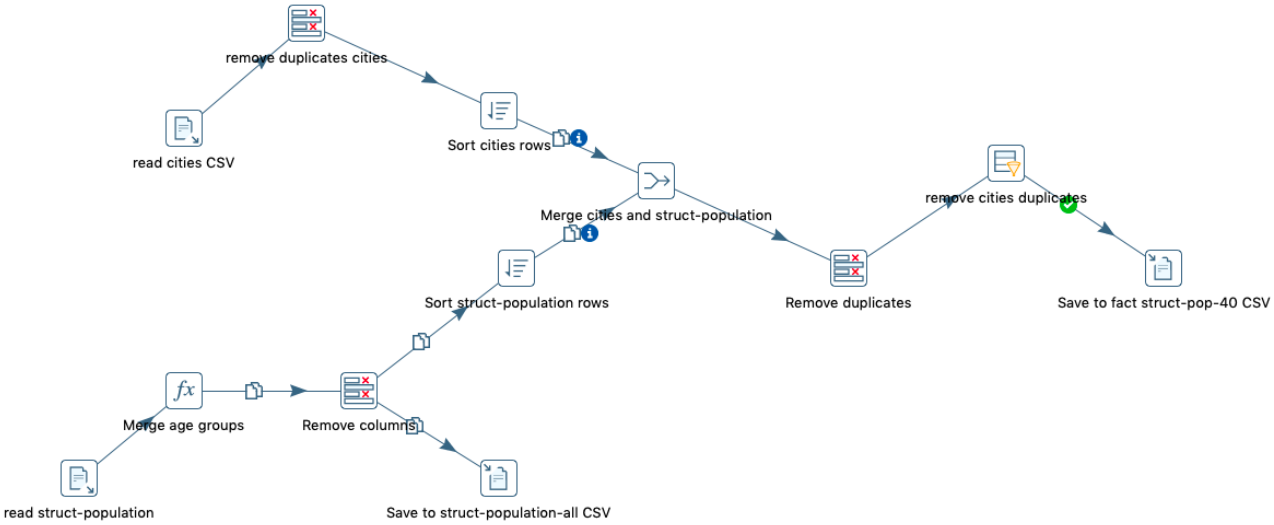



FIGURE 6 – Structure de la population

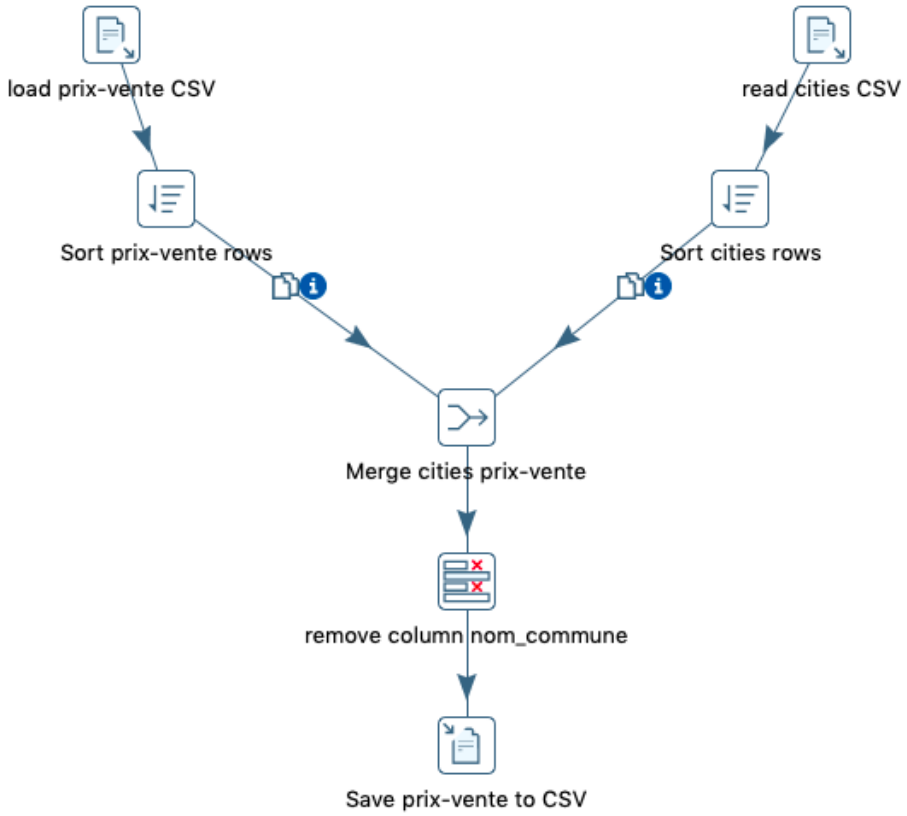



FIGURE 7 – Prix de vente

## 3.2 Insertion des données dans le datawarehouse

Nous remplissons ensuite les tables de dimensions et de faits avec les transformations `buildDim_*` et `buildFact_*`. Par exemple, les figures 8, 9, et 10 contiennent, respectivement, les étapes d'insertion des dates d'extraction des prix, des commerces par commune, et des rendements par ville.

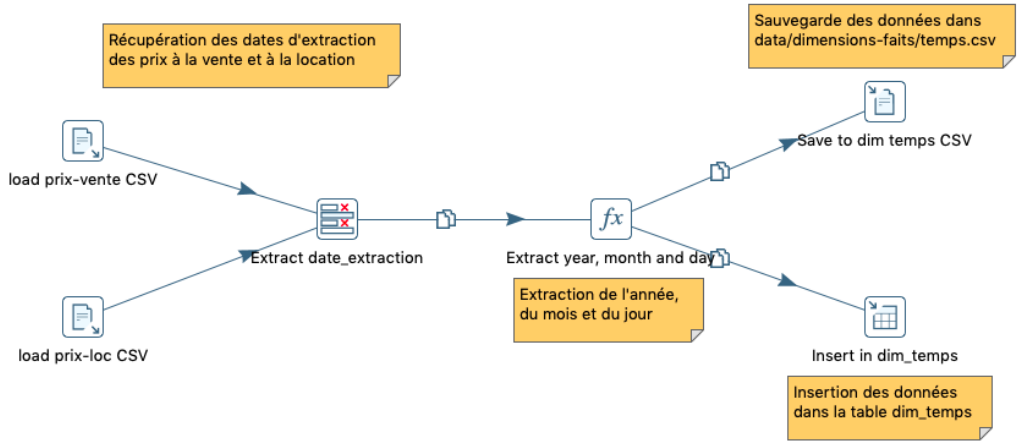



FIGURE 8 – Insertion des dates dans le datawarehouse

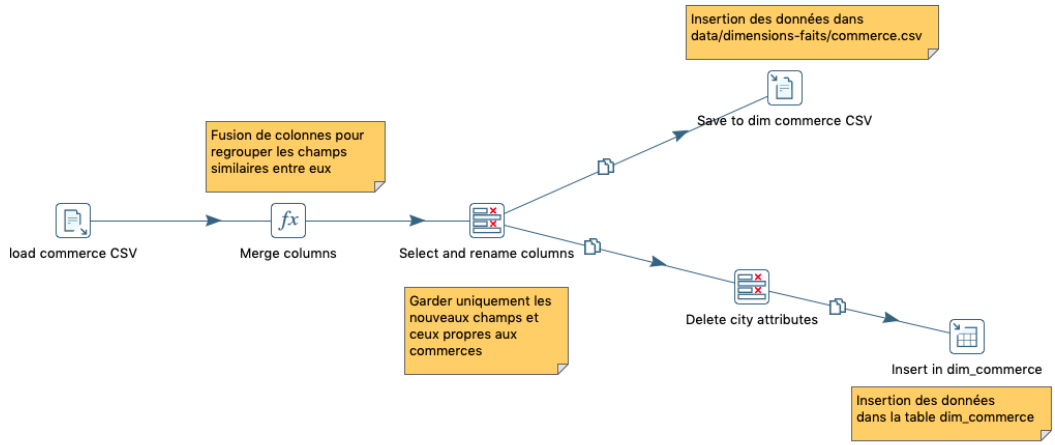



FIGURE 9 – Insertion des commerces dans le datawarehouse

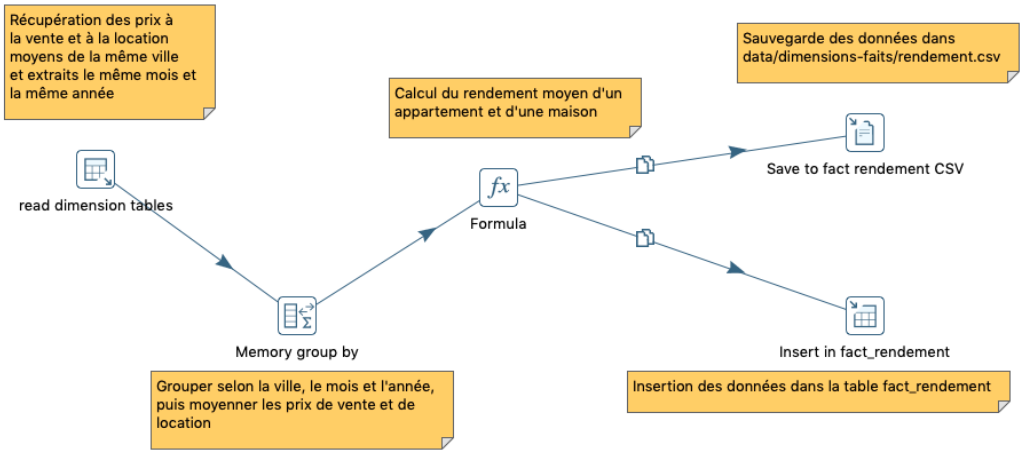



FIGURE 10 – Insertion des rendements dans le datawarehouse

Enfin, sur la figure 11 se trouve la tâche permettant de construire le datawarehouse. Chaque dimension est chargée par `load_dim_*`, chaque fait par `load_fact_*`.

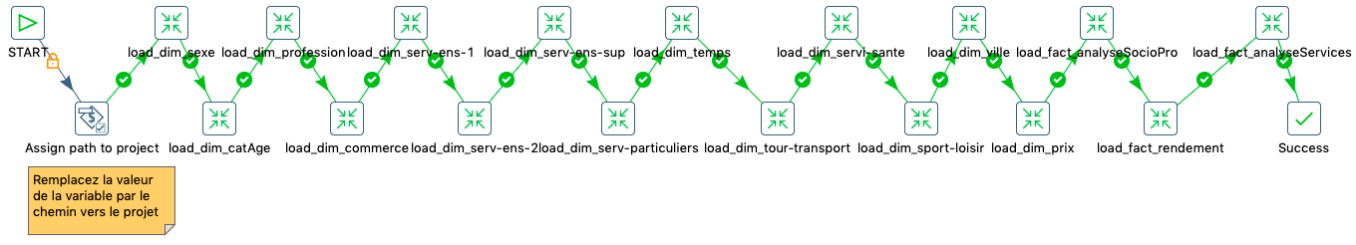



FIGURE 11 – Création du datawarehouse

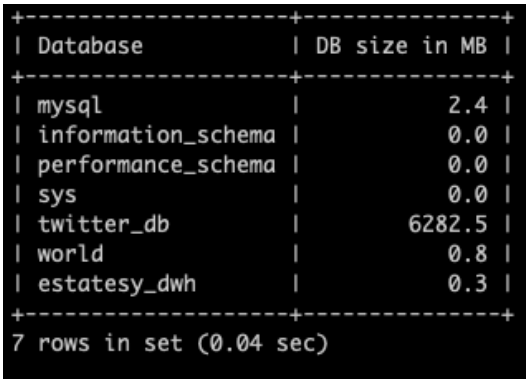
## 3.3 Résultats

Une fois les transformations effectuées, l'ensemble des faits et dimensions est stocké dans la base de données MySQL `estatesy_dwh`. Les résultats se trouvent sur la figure 12.

| Tables_in_estatesy_dwh |
|---|
| dim_catAge |
| dim_commerce |
| dim_ens1 |
| dim_ens2 |
| dim_ens_sup |
| dim_prixLocation |
| dim_prixVente |
| dim_profession |
| dim_serv_particuliers |
| dim_serv_sante |
| dim_sexe |
| dim_sports_loisirs |
| dim_temps |
| dim_tour_transport |
| dim_ville |
| fact_analyseServices |
| fact_analyseSocioPro |
| fact_rendement |

18 rows in set (0.01 sec)

(a) Tables de `estatesy_dwh`



| Database | DB size in MB |
|---|---|
| mysql | 2.4 |
| information_schema | 0.0 |
| performance_schema | 0.0 |
| sys | 0.0 |
| twitter_db | 6282.5 |
| world | 0.8 |
| estatesy_dwh | 0.3 |

7 rows in set (0.04 sec)

(b) Taille en MB de `estatesy_dwh`

FIGURE 12 – `estatesy_dwh`

# 4 Analyse des données

Dans cette section, nous procédons à l'analyse de nos données traitées afin de déterminer les villes où investir.

## 4.1 Analyse de la structure de la population

Nous souhaitons d'abord avoir une idée du nombre d'habitants de chacune de ces communes en affichant sur la figure 13 la population totale par ville. Nous observons que Paris, Marseille et Lyon constituent les trois villes les plus peuplées de France.

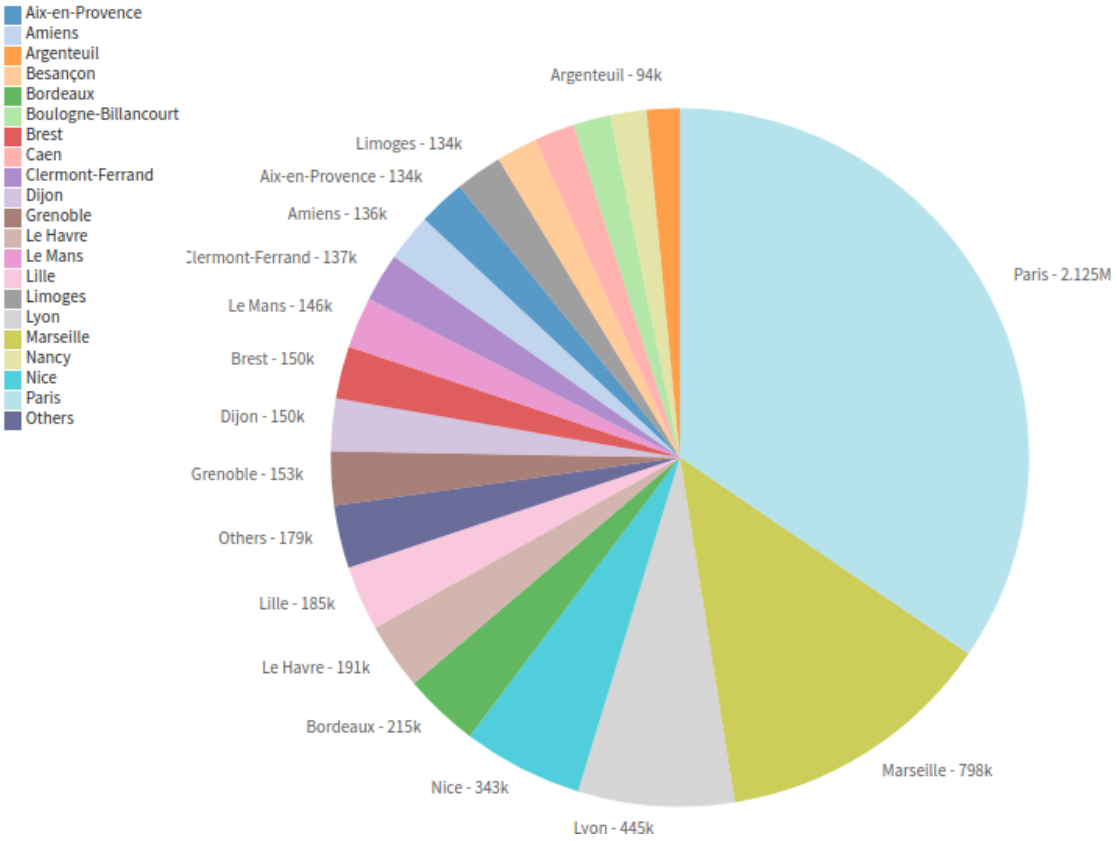



FIGURE 13 – Population totale par ville

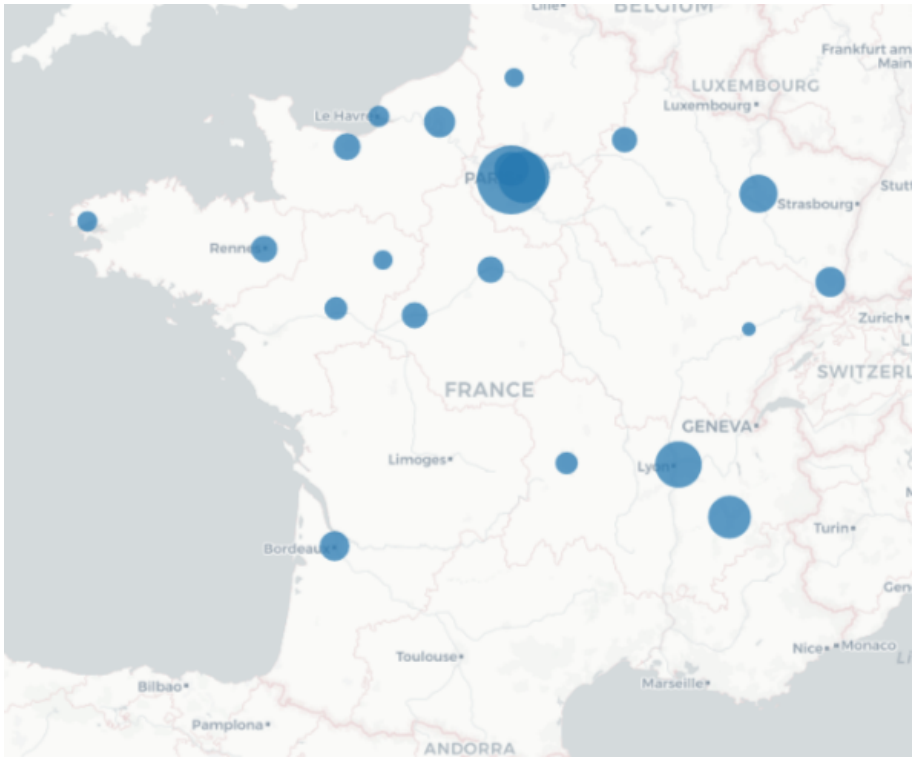



FIGURE 14 – Densité d'habitants par ville

La figure 14 contient la densité d'habitants par ville, représentée par la taille des cercles sur la carte de la France. Nous observons que Paris, Lyon et Grenoble ont de fortes densités de population ; ce qui reflète l'attractivité de ces villes.

NB : certaines villes ne sont pas représentées car le fichier des géo-localisations contient des données manquantes.

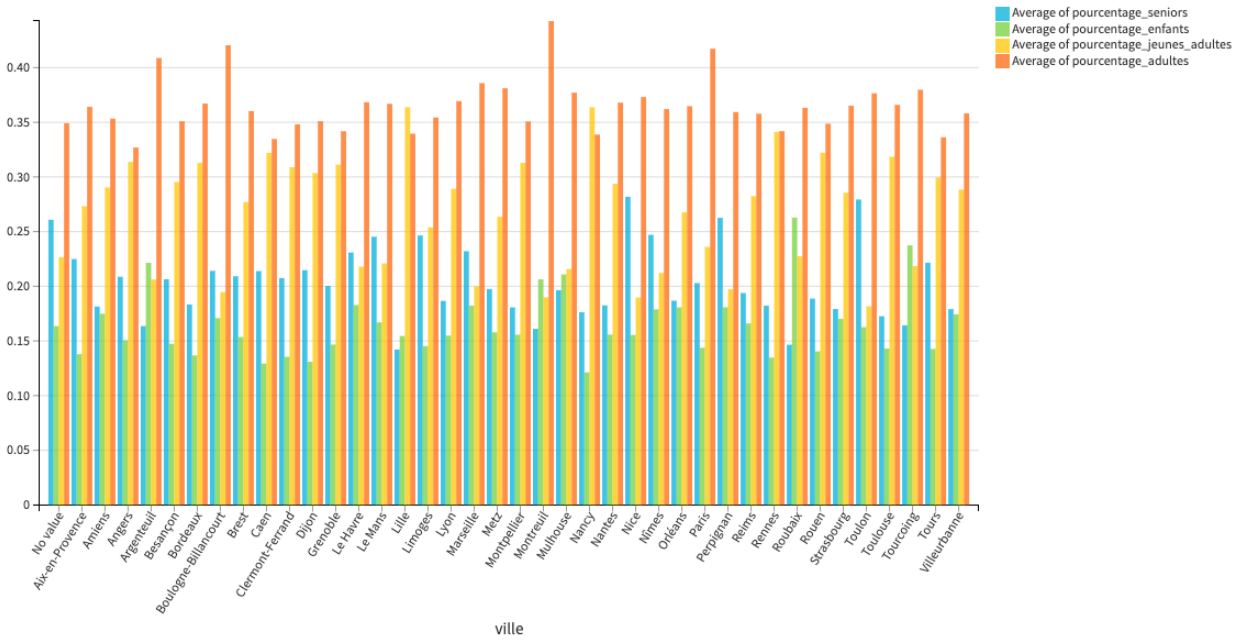



FIGURE 15 – Pourcentage de chaque tranche d'âge par ville

Il est important de prendre en compte la proportion des différentes tranches d'âge de la population par ville. En effet, le fort vieillissement de la population française est un facteur d'importance. Le parc immobilier doit donc s'adapter aux besoins d'une population vieillissante : les apparte-

ments au troisième étage ou plus sans ascenseurs auront encore plus de mal à trouver preneur qu'aujourd'hui. Sur la figure 15 sont représentés les pourcentages de chaque tranche d'âge dans la population de chacune des villes.

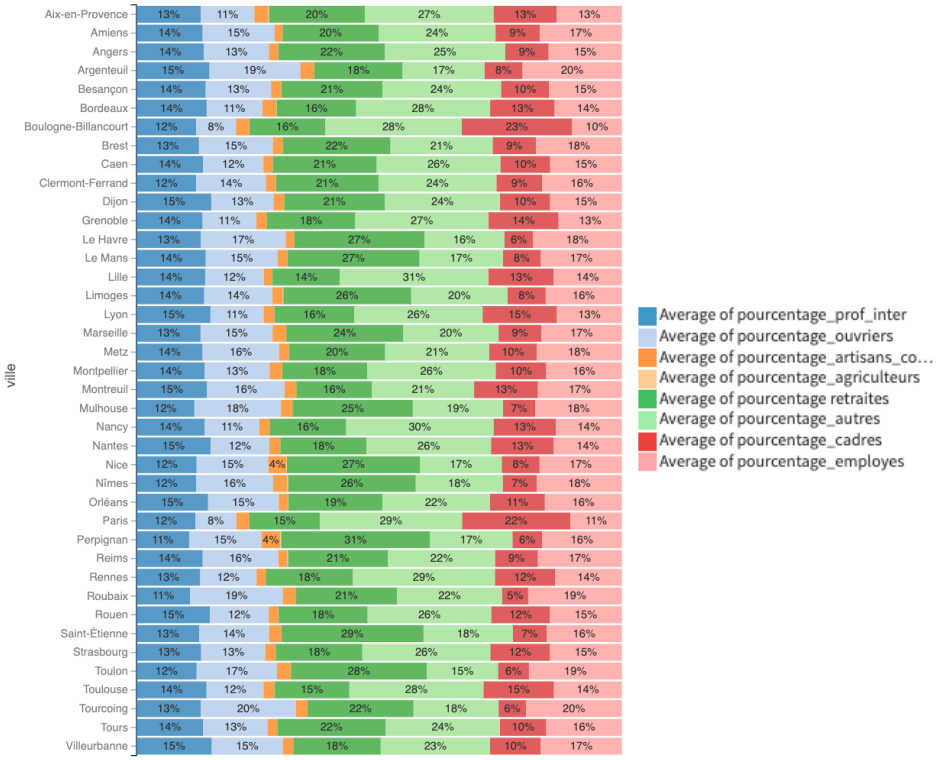



FIGURE 16 – Catégories socio-professionnelles par ville

Nous devons également prendre en compte les catégories socio-professionnelles de la population. À Paris, 64% des acheteurs sont des cadres ou des dirigeants d'entreprise. Ces catégories, qui représentent 21% de la population active, sont surreprésentées parmi les acheteurs franciliens. La figure 16 permet de confirmer que Paris et sa banlieue comportent un nombre importants de cadres.

## 4.2 Analyse des services rendus aux territoires

Nous pouvons constater, sur la figure 17, que la part de chaque moyen de divertissement diffère d'une ville à l'autre selon ses spécificités. Cela nous permet de choisir les villes selon la catégorie du client ciblé, puis de lier ces informations à la structure d'âge des habitants.

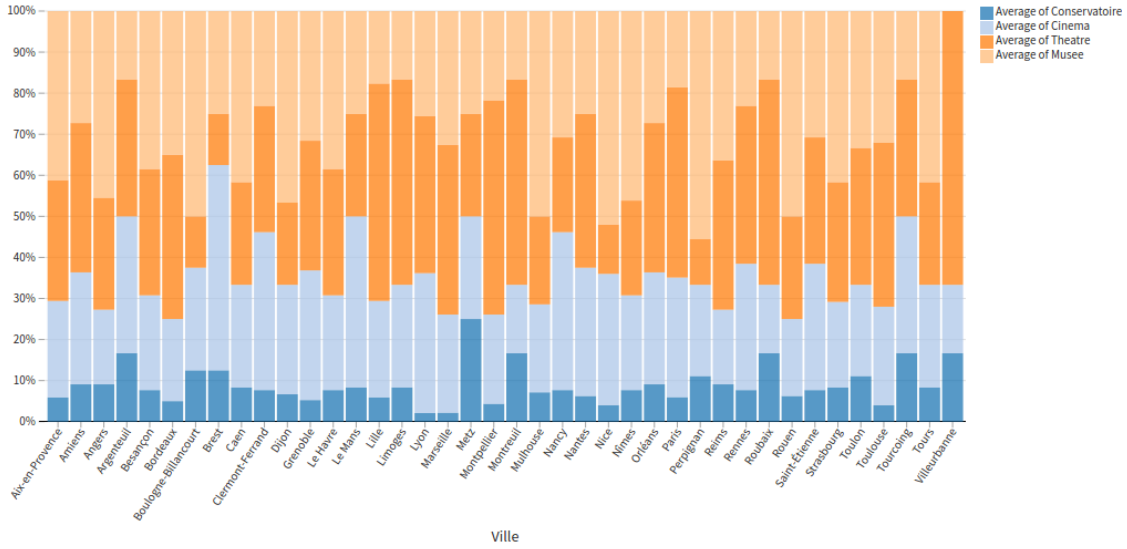



FIGURE 17 – Lieux de divertissement par ville

Nous considérons également que la part accordée à l'éducation est un facteur important de la qualité de vie d'un territoire. La figure 18 contient les 10 villes accordant le plus d'infrastructures liées à l'enseignement. Amiens, Orléans et Tourcoing sont en tête.

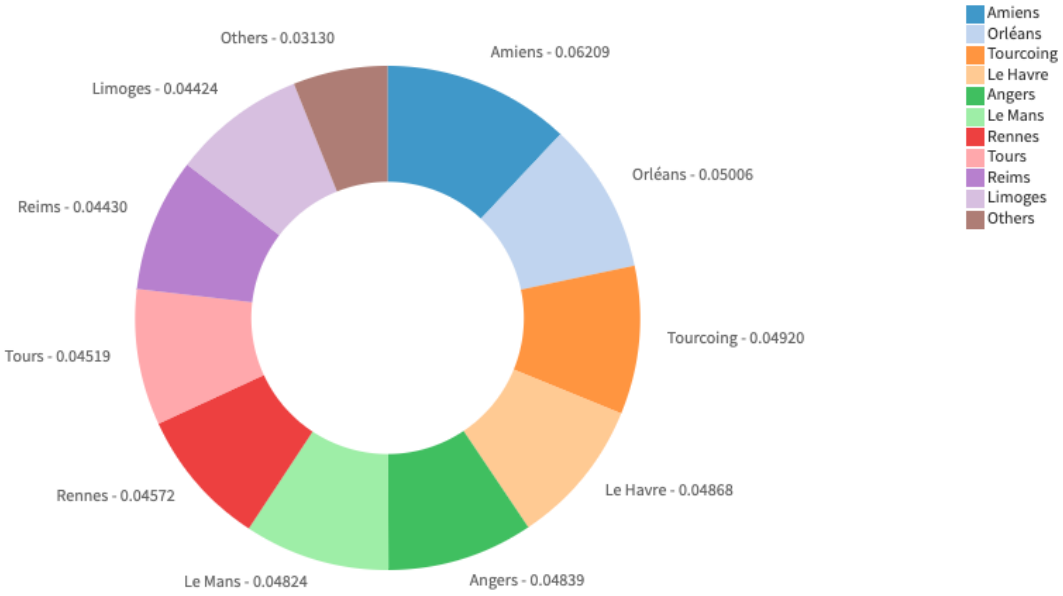



FIGURE 18 – Top 10 des villes ayant le plus haut taux d'infrastructures accordées à l'enseignement

Nous supposons de plus que le taux d'infrastructures accordées aux services de santé est important : d'après la figure 19, Caen, Amiens et Brest sont les villes qui en consacrent le plus.

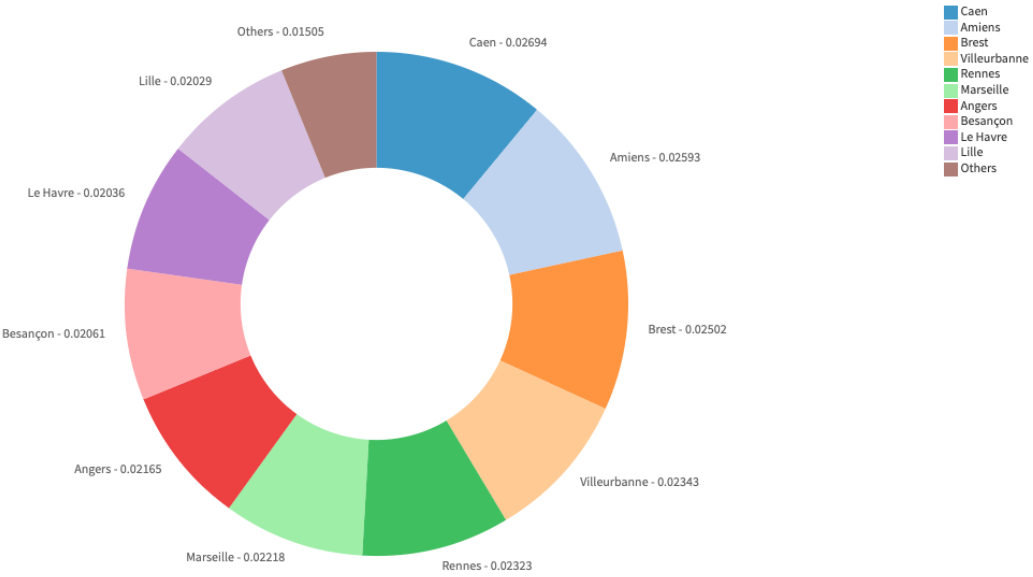



FIGURE 19 – Top 10 des villes ayant le plus haut taux d'infrastructures accordées aux services de santé


FIGURE 20 – Taux d'infrastructures accordées à chaque type de service par ville

D'après la figure 20, nous remarquons, sur 25 villes prises aléatoirement, que la part associée à chaque type de service varie peu d'une commune à l'autre.

## 4.3 Analyse des prix au m$^2$

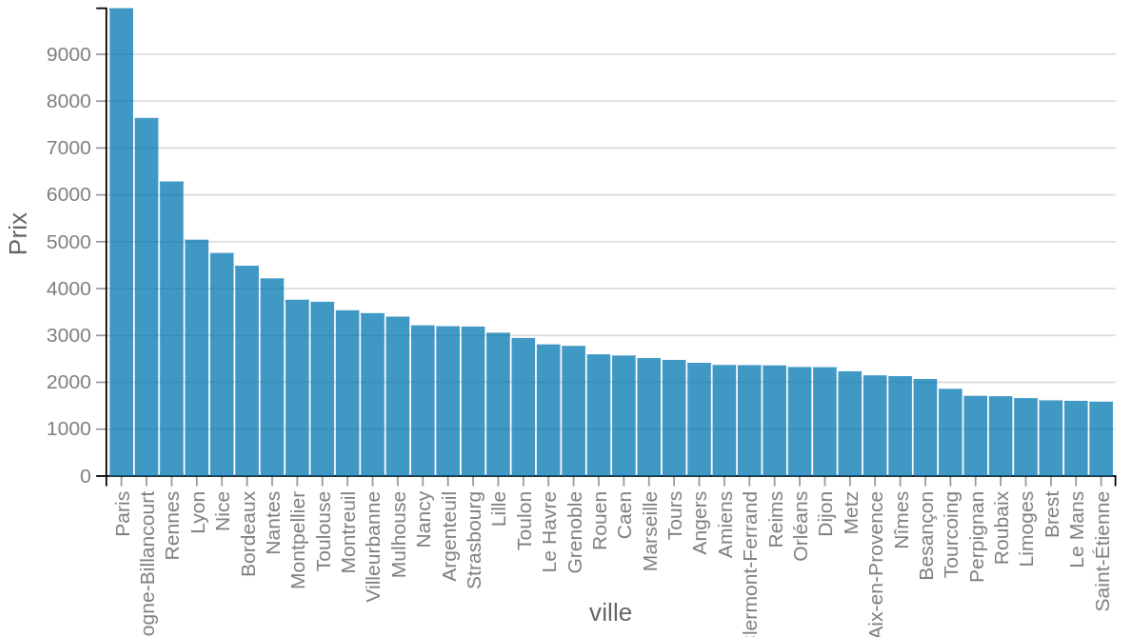



FIGURE 21 – Prix moyen à la vente d'un appartement par ville

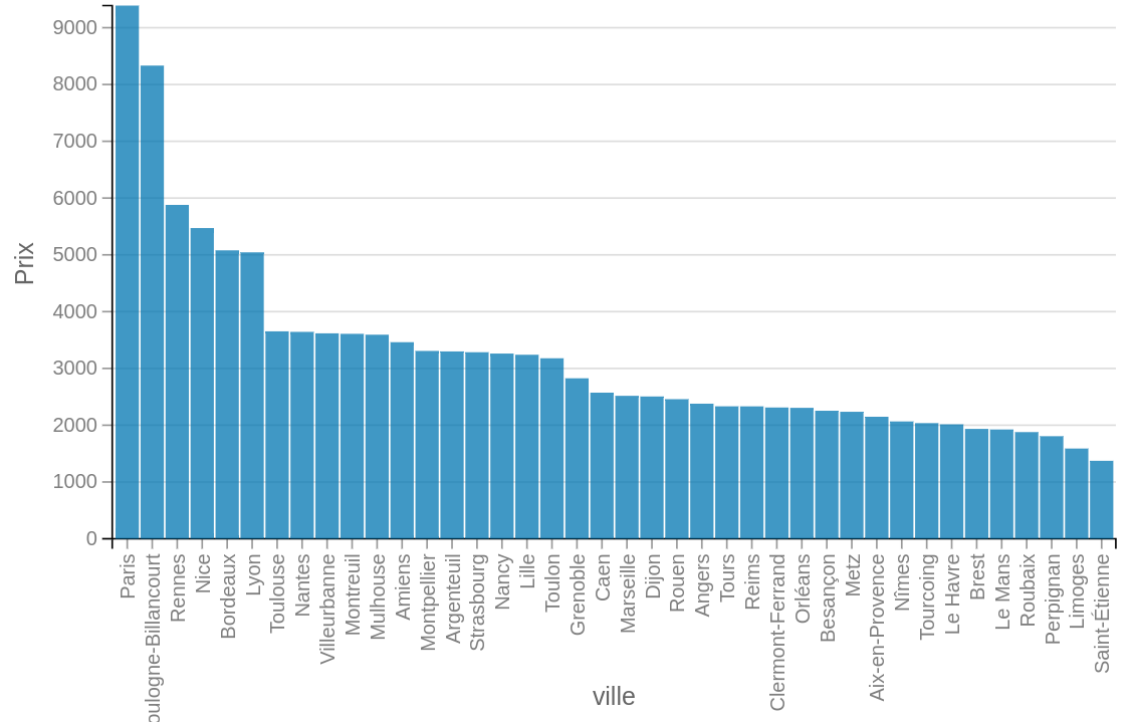



FIGURE 22 – Prix moyen à la vente d'une maison par ville

Sur les figures 21 et 21 se trouvent, respectivement, le prix moyen au m$^2$ à la vente d'un appartement par ville, et le prix moyen au m$^2$ à la vente d'une maison par ville. Nous observons que Paris, Boulogne-Billancourt, Rennes, Lyon et Nice font partie des villes où les biens immobiliers sont les plus chers. En revanche, Saint-Étienne, Le Mans, Brest, Limoges, Roubaix et Perpignan constituent les villes où il est le moins cher d'acheter un appartement ou une maison.

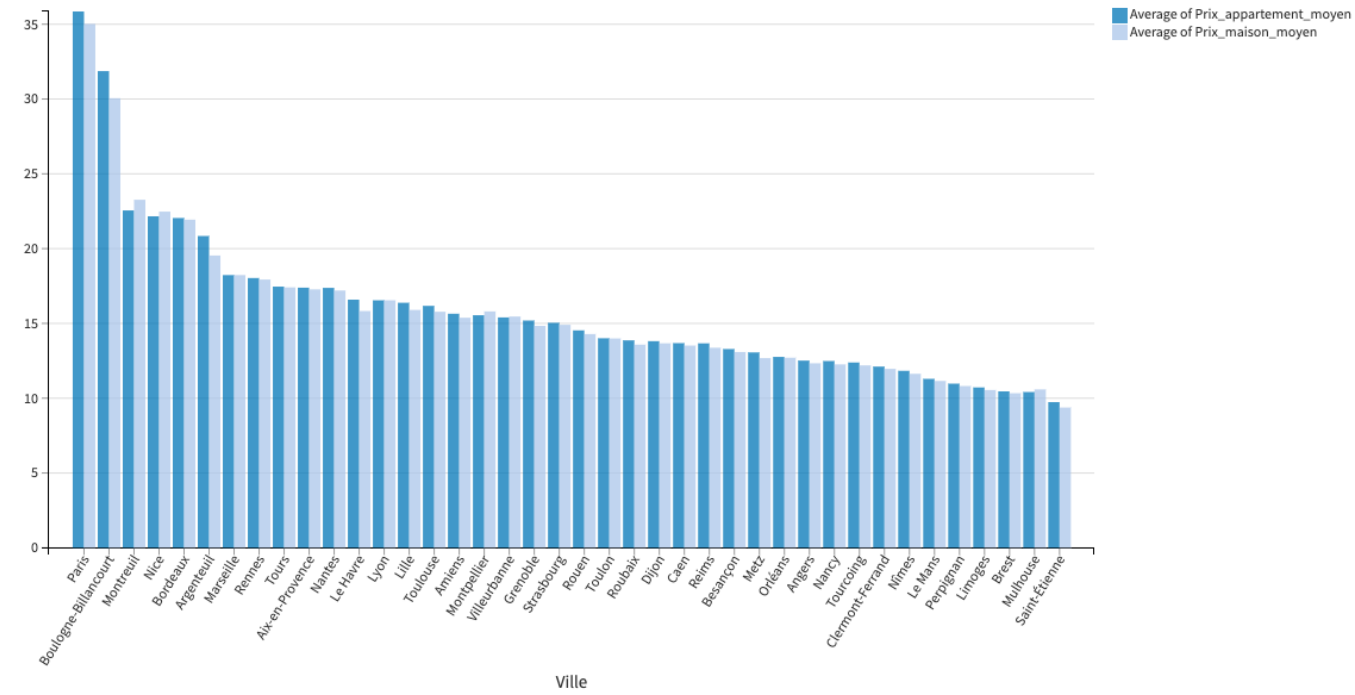



FIGURE 23 – Prix moyens à la location d'une maison et d'un appartement par ville

En ce qui concerne la location, la figure 25 montre que Paris et sa banlieue (Boulogne-Billancourt, Montreuil, Argenteuil) ainsi que Nice et Bordeaux sont les villes où louer un appartement ou une maison est le plus cher. Saint-Étienne, Mulhouse et Brest sont celles où les biens immobiliers sont les moins onéreux.

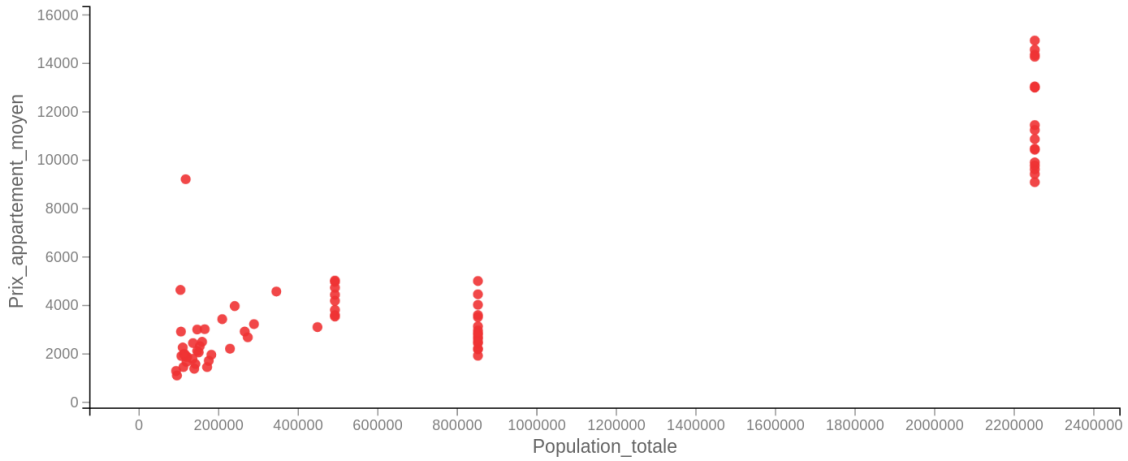



FIGURE 24 – Prix moyen à la vente d'un appartement en fonction du nombre d'habitants par ville

Nous traçons, sur la figure 24, l'évolution du prix moyen au $m^2$ à la vente d'un appartement en fonction du nombre d'habitants. Entre environ 100000 et 50000 habitants, il semble y avoir une légère corrélation entre la taille de la population et le prix moyen d'un appartement.

## 4.4 Analyse du rendement de l'investissement

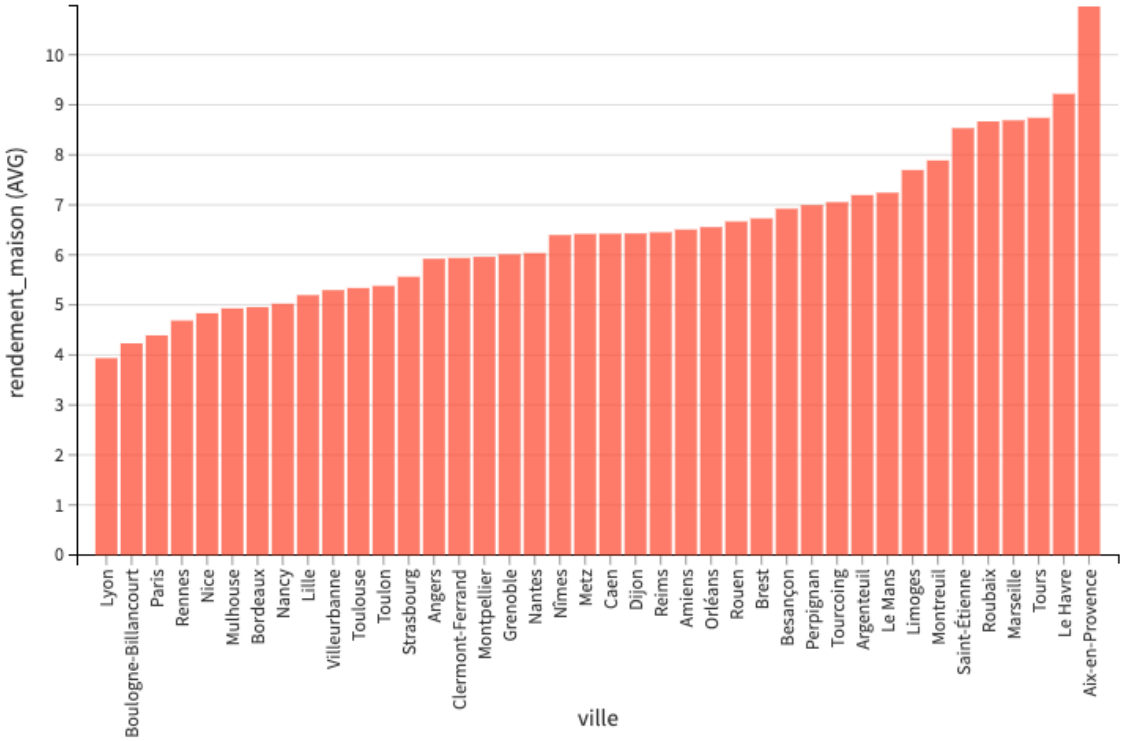



FIGURE 25 – Rendement moyen de l'investissement pour une maison par ville

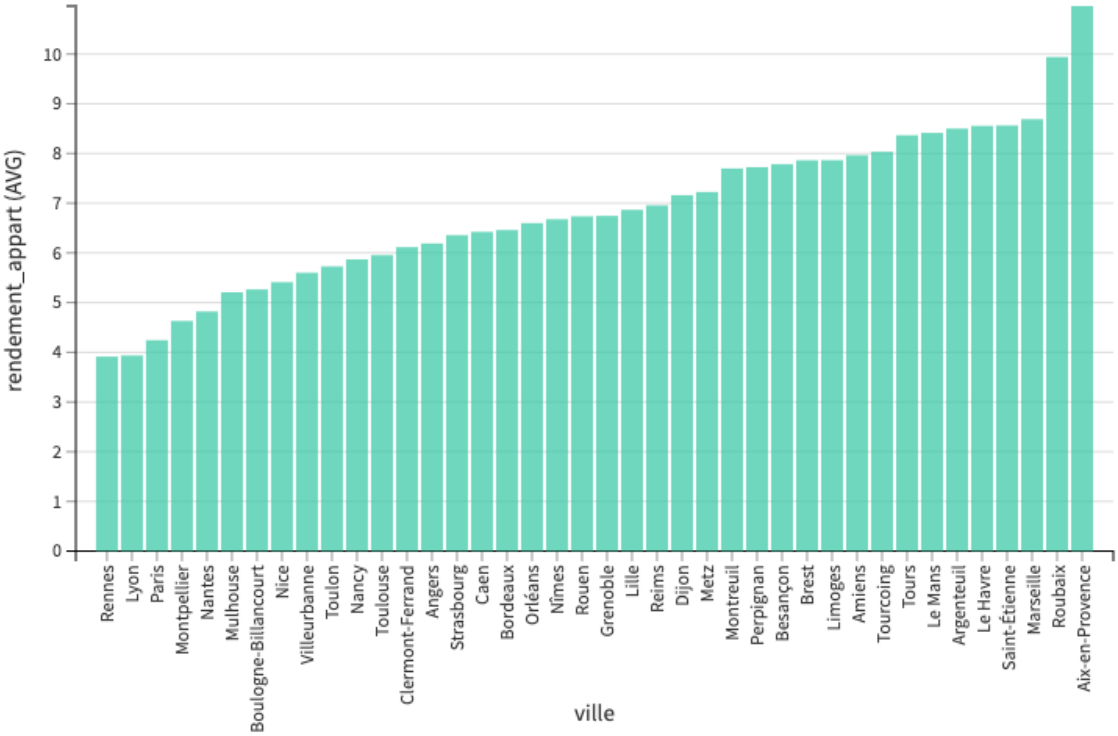



FIGURE 26 – Rendement moyen de l'investissement pour un appartement par ville

Enfin, nous affichons, sur les figures 25 et 26, le rendement moyen de l'investissement dans une maison et un appartement, respectivement. Aix-en-Provence, Le Havre, Tours, Marseille et

Roubaix constituent les cinq villes où l'investissement dans une maison est le plus profitable. En ce qui concerne un appartement, Aix-en-Provence, Roubaix, Marseille, Saint-Étienne et Le Havre arrivent en tête.