INGÉNIERIE DES DONNÉES

Projet 1 : Production d'une ressource en données ouvertes et FAIR

Professeur référent :

M. Aurélien MAX

Membres du groupe :

Issam CHAFIQ
Ilyes ARABET
Matheo BOUILLER
Marc CHARIGNON
Alexandre DEMONTAIGNE

POLYTECH PARIS-SACLAY

27 Janvier 2026

Table des matières

1 Description des caractéristiques du projet

L'objectif technique du projet était de produire un indicateur d'attractivité pour chaque commune, basé sur trois piliers : la densité de population, le niveau de vie médian et l'offre de transport.

1.1 Sources de données utilisées

Nous avons sélectionné quatre sources de données ouvertes, en veillant à éviter les problèmes de licence restrictive.

1. Transports (Point d'Accès National - PAN)
Jeu de données : `gtfs-stops-france-export`.
Mise à jour : 13 Janvier 2026.
Licence : Licence Ouverte / Open Licence 2.0.
Contenu : Données ponctuelles (POI) des arrêts de transport en commun sur toute la France.

2. Revenus (INSEE / Filosofi)
Jeu de données : Dispositif Fichier localisé social et fiscal (Filosofi).
Millésime : 2021 (Dernier disponible).
Avertissement source : Le millésime 2022 a été annulé par l'INSEE en raison d'une qualité statistique insuffisante liée à la suppression de la taxe d'habitation, empêchant le rattachement fiable des foyers fiscaux aux logements.

3. Population (INSEE)
Jeu de données : Recensement de la population.
Millésime : Population 2019 (limites territoriales 2021, entrée en vigueur 2022).
Contenu : Données officielles authentifiées par décret. Note importante : l'enquête annuelle 2021 ayant été reportée (Covid-19), les comparaisons temporelles doivent être faites avec prudence.

4. Référentiel Géographique (Opendatasoft / GeoRef)
Jeu de données : `georef-france-commune-arrondissement-municipal`.
Licence : Open License v2.0.
Spécificité : Ce référentiel fournit les délimitations administratives et les contours géographiques (polygones) précis de l'ensemble du territoire. Contrairement aux référentiels classiques, il inclut les arrondissements municipaux (Paris, Lyon, Marseille) en tant qu'entités distinctes, ce qui est crucial pour notre granularité d'analyse sur les métropoles.

1.2 Questions liées à l'ingestion des données

L'ingestion n'a pas été qu'un simple import technique, mais un travail de logique de données.

— **Tri et Filtrage :** La plupart des sources (INSEE, Filosofi) sont fournies sous forme de classeurs Excel multi-feuilles ou de fichiers CSV complexes contenant des métadonnées mélangées aux données. Il a fallu isoler spécifiquement les données communales et écarter les données régionales ou départementales.

— **Construction de la Clé Primaire :** Pour joindre ces tables hétérogènes, nous avons dû normaliser une clé de jointure commune : le Code Officiel Géographique (Code INSEE). Dans certaines tables, il a fallu le reconstruire en concaténant le code département et le code commune.

— **Compromis Temporel :** L'impossibilité d'avoir toutes les sources sur la même année (Revenus 2021 vs Transports 2026) est une contrainte forte. Nous avons fait le choix de combiner les versions les plus récentes disponibles de chaque producteur pour maximiser la pertinence actuelle, malgré le décalage temporel.

1.3 Format produit

Nous livrons la ressource sous deux formats : un fichier **.xlsx** et une visualisation **HTML interactive**(cf. Annexe A). Le choix du format Excel plutôt que SQL ou JSON se justifie par l'accessibilité : notre utilisateur cible (investisseur, décideur) doit pouvoir manipuler les données sans compétences en programmation. La carte HTML de son côté permet une exploration visuelle immédiate pour identifier les communes.

1.4 Données calculées et regroupées

— **Données regroupées :** La *population* et le *revenu médian* sont des données d'attributs directes, récupérées telles quelles après nettoyage.

— **Données calculées :**

  — `nb_arrets` : Cet attribut n'existait pas. Nous l'avons créé par intersection géométrique entre les points GPS du fichier PAN et les polygones des communes du GeoJSON.

  — `score_attractivite` : Une valeur calculée pondérant le revenu, la population et la densité de transport.
  Formule : $Score = 0.5 \times \text{Rev}_{norm} + 0.4 \times \text{Pop}_{norm} + 0.1 \times \text{Trans}_{norm}$

2 Discussion critique de la solution proposée

2.1 Pérennité de la stratégie automatique (Pipeline)

Notre pipeline actuel fonctionne mais présente des fragilités structurelles :

— **Pipeline "en dur" :** Seul le fichier GeoJSON est téléchargé dynamiquement par le code. Les fichiers Excel et CSV (INSEE, PAN) ont été téléchargés manuellement avant traitement.

— **Conséquence :** En cas de mise à jour des sources par les producteurs, notre code ne s'adaptera pas automatiquement. Il nécessitera une intervention manuelle pour télécharger les nouveaux fichiers et vérifier que les noms de colonnes n'ont pas changé. La solution est fonctionnelle mais pas entièrement pérenne sans maintenance.

2.2 Qualité et fiabilité de la ressource

Notre ressource ne représente pas une vérité absolue mais une estimation dépendante de la qualité des données d'entrée et de la richesse du modèle choisi.

— **Biais dans les données sources :** La fiabilité de notre indicateur est directement corrélée à la complétude des jeux de données ouverts utilisés.
Exemple : Nous avons identifié un manque significatif de données dans l'export national du PAN concernant les transports de la métropole lyonnaise. De nombreux arrêts étant manquants, l'attribut `nb_arrets` pour Lyon est factuellement sous-évalué, ce qui biaise son score d'attractivité final. Ce constat impose une transparence totale envers l'utilisateur via les métadonnées.

— **Limites de la modélisation :** Notre score actuel se fonde sur une sélection restreinte de trois attributs (Population, Revenu, Transport). Pour affiner la pertinence de l'indicateur et obtenir une représentation plus fidèle de la réalité territoriale, l'intégration de sources supplémentaires (ex : offre de soins, établissements scolaires, couverture numérique) aurait permis de construire un modèle multidimensionnel plus robuste.

2.3 Cycle de vie, mise à jour et contributions

Cette section aborde la gestion de l'évolution de la ressource.

— **Mise à jour (périodique) des données :** Actuellement, la mise à jour est manuelle : il faudrait retélécharger les fichiers Excel/CSV si une nouvelle version sortait. Pour améliorer la pérennité, l'idéal serait d'automatiser ce processus via des scripts interrogeant directement les API des fournisseurs plutôt que de manipuler des fichiers statiques.

— **Mise à jour du schéma :** Si une nouvelle source devait être ajoutée, notre code est modulaire : tant que la nouvelle table possède une clé `code_insee`, elle peut être jointe sans casser l'architecture existante.

— **Invalidation et correction de données :** Face aux données manquantes (ex : arrêts de bus absents pour Lyon), deux approches sont possibles. La correction pourrait se faire manuellement dans le fichier final Excel en comblant les trous par une recherche externe. Pour l'invalidation, nous avons choisi d'attribuer une valeur par défaut (0) pour permettre le calcul, mais une méthode plus rigoureuse serait de marquer ces cellules comme `NULL` ou d'ajouter une colonne "Indice de confiance" pour invalider la fiabilité du score sur ces communes spécifiques.

— **Contribution par des tiers :** Le caractère ouvert du code permettrait à un tiers disposant de meilleurs sources de données de remplacer notre fichier source pour corriger les biais, sans avoir à réécrire la logique de calcul.

3 Discussion de l'applicabilité des principes FAIR

3.1 Objectifs poursuivis

La mise à disposition de ces données vise plusieurs publics : des investisseurs cherchant le territoire idéal pour une implantation commerciale, ou des acteurs publics souhaitant visualiser les inégalités territoriales. L'objectif est de démocratiser l'accès à une information complexe qui nécessiterait sinon une expertise technique avancée.

3.2 Analyse selon les principes FAIR

En nous basant sur principes des données FAIR :

F - Findable
Actuellement, déposer le projet sur un GitHub ne suffit pas. Pour être réellement "Findable", la ressource devrait être indexée sur une plateforme de données (type data.gouv.fr) avec des métadonnées riches (mots-clés : *attractivité, commerce, communes*) et si possible un identifiant unique (DOI).

A - Accessible
Nous respectons ce principe par l'utilisation de protocoles standards. L'accès aux données ne nécessite pas d'authentification propriétaire ou de logiciel payant spécifique (le format .xlsx s'ouvre avec des outils libres comme LibreOffice).

I - Interoperable
En utilisant strictement le **Code Officiel Géographique (COG)** de l'INSEE comme pivot, nous garantissons que notre jeu de données est interopérable avec l'immense majorité des bases de données administratives françaises (Sirene, Éducation Nationale). Nous avons évité d'utiliser les noms de communes (sujets aux erreurs d'orthographe) comme clés.

R - Reusable
Pour garantir la réutilisabilité , nous devons fournir :

1. Une licence ouverte.
2. Un fichier de métadonnées ("Provenance") expliquant explicitement les calculs effectués (formule du score) et les biais connus.
3. Une solution pour mettre à jour les données (script Python fourni) afin que la ressource ne devienne pas obsolète.

4 Annexes

A Aperçu de la visualisation interactive

Ce rendu visuel permet d'explorer les données géographiques (arrêts, population, attractivité) de manière intuitive .

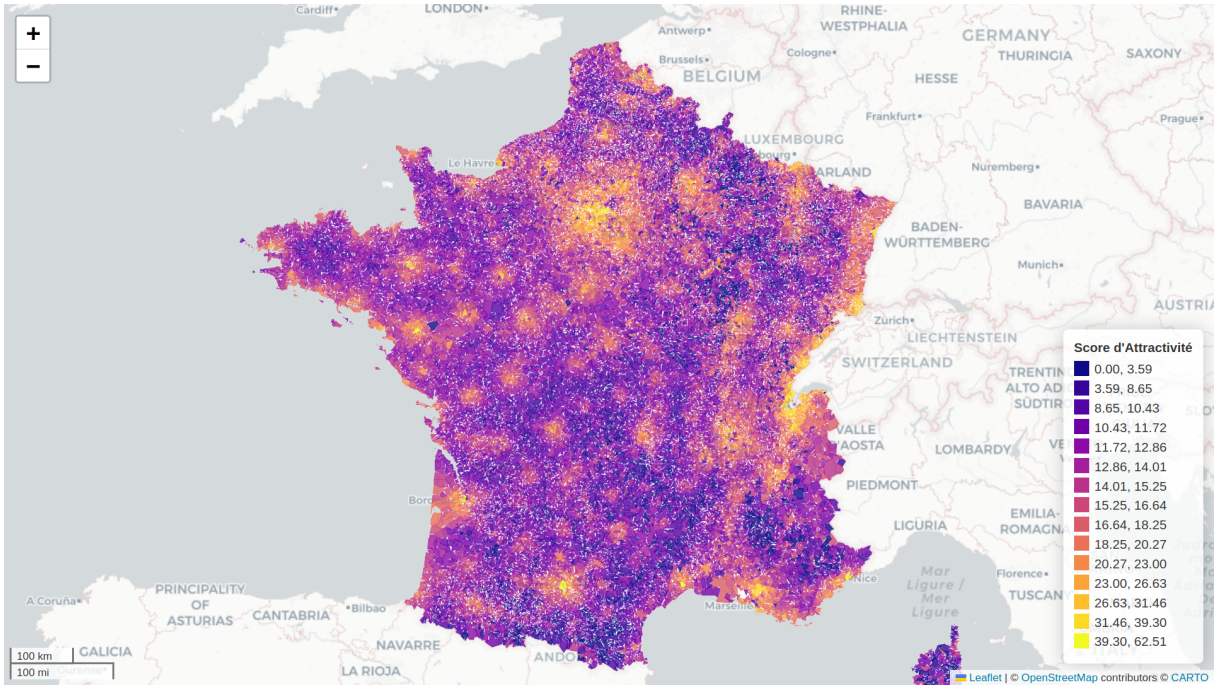

FIGURE 1 – Capture de la carte HTML