9 projets data science multi-sources pour une équipe de 7

Voici neuf sujets concrets, originaux et réalisables couvrant le commerce, la santé publique et l'énergie, chacun combinant au minimum deux sources de données ouvertes avec composante temporelle. Chaque proposition inclut des datasets vérifiés et téléchargeables, une stratégie de jointure multi-sources, et une estimation de faisabilité calibrée pour **945 heures-personne** (7 × 135h). Les sujets ont été sélectionnés pour leur richesse analytique (classification, séries temporelles, détection d'anomalies, NLP, géospatial) et leur potentiel de tableau de bord décisionnel percutant.

DOMAINE 1 — COMMERCE / E-COMMERCE

🛒 Projet 1 : Intelligence e-commerce brésilien — prédiction des retards de livraison et satisfaction client

Concept. Exploiter le plus grand jeu de données e-commerce public au monde (marketplace Olist, Brésil) enrichi d'indicateurs économiques nationaux et de tendances de recherche Google pour construire une plateforme d'aide à la décision logistique. Trois axes d'analyse interconnectés : prédire les retards de livraison avant qu'ils ne surviennent, modéliser les scores de satisfaction client, et prévoir les volumes de ventes régionaux.

Source 1 — Brazilian E-Commerce Public Dataset (Olist)

- **URL :** https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce
- **Volume :** 9 fichiers CSV reliés entre eux, **~45 Mo** total. Le fichier `olist_orders_dataset.csv` contient **~100 000 commandes**, (GitHub) `order_items` **~112 000 lignes**, et surtout `olist_geolocation_dataset.csv` atteint **~1 million de lignes** (source volumineuse et bruitée avec doublons géographiques).
- **Variables clés :** identifiants commande/client/vendeur/produit, timestamps d'achat/approbation/expédition/livraison, date estimée de livraison, prix, frais de port, score de review (1–5), commentaires texte en portugais, catégorie produit, dimensions/poids, coordonnées GPS. (Medium)
- **Temporalité :** commandes de **septembre 2016 à octobre 2018** (~2 ans), avec horodatage précis à chaque étape du parcours.

Source 2 — Indicateurs économiques brésiliens (IBGE SIDRA)

- **URL :** https://sidra.ibge.gov.br/home/pms/brasil (portail) — API : https://apisidra.ibge.gov.br/
- **Aussi sur Kaggle :** https://www.kaggle.com/datasets/unanimad/brazilian-economic-indicators
- **Volume :** milliers de lignes de séries temporelles mensuelles par indicateur.
- **Variables clés :** indice mensuel du commerce de détail (PMC) par État, indice des prix à la consommation (IPCA) par région, PIB trimestriel, taux de chômage par État.
- **Temporalité :** données mensuelles/trimestrielles disponibles depuis 2012.

Source 3 — Google Trends (via pytrends)

- **URL :** https://trends.google.com/trends/ — bibliothèque Python `pytrends` (https://github.com/GeneralMills/pytrends)
- **Volume :** scores d'intérêt hebdomadaires (0–100) par catégorie de produit au Brésil, par État.
- **Temporalité :** série hebdomadaire, alignable sur la période Olist 2016–2018.

Stratégie de jointure. Les 9 tables Olist se joignent par clés relationnelles (`order_id`, `customer_id`, `product_id`, `zip_code_prefix`) pour créer un dataset dénormalisé. Les ordres sont ensuite agrégés par (État, mois) et enrichis avec l'indice de commerce IBGE et l'IPC par jointure spatiotemporelle. Les catégories produit Olist sont mappées vers des requêtes Google Trends pour ajouter l'intérêt de recherche comme feature. Les coordonnées GPS permettent de calculer la distance vendeur-client (formule de Haversine).

Types d'analyse possibles :

- **Classification** : prédire le retard de livraison (binaire : livré après la date estimée) — problème de déséquilibre de classes
- **Régression** : prédire le score de review (1–5) à partir des features commande/produit/logistique/économiques
- **Prévision temporelle** : forecasting des ventes mensuelles par État via ARIMA/Prophet enrichi des indicateurs IBGE + Google Trends
- **NLP** : analyse de sentiment sur les commentaires en portugais pour extraire les motifs de plainte
- **Clustering** : segmentation client RFM enrichie de données géographiques et économiques

Pourquoi c'est motivant. Données commerciales réelles (pas synthétiques) d'un marketplace majeur. Architecture relationnelle riche (9 tables !) qui enseigne les jointures complexes. Combine données structurées, texte et géospatial. **Plus de 6 000 notebooks communautaires** sur Kaggle pour le benchmarking. Impact business direct : optimisation logistique, satisfaction client, stratégie régionale.

Faisabilité : ✅ **Très faisable.** Données bien documentées, jointures claires, volume gérable. Le Data Engineer se concentre sur le pipeline relationnel + API IBGE, le Data Analyst sur l'EDA géospatiale, les Data Scientists sur les modèles de prédiction et le NLP, le BI sur un dashboard Streamlit avec carte de suivi des livraisons. Charge estimée : ~200h collecte/préparation, ~150h EDA, ~250h modélisation, ~100h validation, ~150h dashboard, ~95h documentation.

🛍️ Projet 2 : Intelligence nutritionnelle du panier e-commerce — Instacart × Open Food Facts

Concept. Aller au-delà du market basket analysis classique en enrichissant les données d'achat massives d'Instacart (32 millions de lignes) avec la base nutritionnelle Open Food Facts et les indices de prix alimentaires américains. Questions de recherche : les clients qui achètent bio ont-ils des paniers différents ? Peut-on prédire le réachat à partir de profils nutritionnels ? Comment les prix alimentaires influencent-ils les comportements ?

Source 1 — Instacart Market Basket Analysis

- **URL :** https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis
- **Volume : ~1,1 Go** total. Le fichier order_products__prior.csv contient **~32 millions de lignes** (source volumineuse principale). orders.csv : ~3,4 M lignes. products.csv : ~50 000 produits, 134 rayons, 21 départements. GitHub Medium
- **Variables clés :** identifiants commande/utilisateur/produit, jour de la semaine, heure, jours depuis la commande précédente, indicateur de réachat, nom du produit, rayon, département.
- **Temporalité :** temporalité relative par utilisateur (jour de semaine, heure, séquence de commandes, intervalle entre commandes).

Source 2 — Open Food Facts

- **URL CSV :** https://static.openfoodfacts.org/data/en.openfoodfacts.org.products.csv.gz (~0,9 Go compressé) Open Food Facts
- **Aussi sur Kaggle :** https://www.kaggle.com/datasets/openfoodfacts/world-food-facts
- **Volume : ~4 millions de produits** mondiaux, **200+ colonnes** (utiliser le sous-ensemble américain : plusieurs centaines de milliers de produits). Emergent Mind
- **Variables clés :** nom produit, marque, catégories, Nutri-Score (A–E), groupe NOVA (1–4, niveau de transformation), énergie/lipides/glucides/protéines/fibres/sel pour 100g, ingrédients, allergènes, pays. Emergent Mind
- **Temporalité :** dates de création et de dernière modification des fiches produit.

Source 3 — FRED / USDA Food CPI

- **URL :** https://fred.stlouisfed.org/series/CPIUFDNS (IPC alimentaire mensuel) FRED
- **Aussi :** https://www.ers.usda.gov/data-products/food-at-home-monthly-area-prices (prix mensuels par catégorie alimentaire) Economic Research Service
- **Volume :** séries mensuelles (~1 000+ lignes sur plusieurs décennies), 90 catégories alimentaires × 15 zones géographiques. Economic Research Service
- **Temporalité :** mensuelle, forte composante temporelle pour le forecasting.

Stratégie de jointure. Le défi principal est le **fuzzy matching** entre les noms de produits Instacart et Open Food Facts (TF-IDF + similarité cosinus ou distance de Levenshtein). Aussi jointure par mapping catégoriel : rayon Instacart → catégories OFF. Les départements Instacart (dairy, produce, snacks) sont mappés vers les catégories USDA/FRED pour ajouter les tendances de prix. Le résultat : chaque produit Instacart est enrichi de son profil nutritionnel complet et du contexte de prix.

Types d'analyse :

- **Règles d'association** (Apriori/FP-Growth) enrichies par catégories nutritionnelles
- **Classification** : prédire le réachat d'un produit (binaire) en utilisant features produit + historique utilisateur + profil nutritionnel

- **Clustering** : segmentation client par régime alimentaire (health-conscious vs. convenience vs. prix)
- **Système de recommandation** hybride : filtrage collaboratif + content-based nutritionnel
- **Régression** : prédire la taille du panier et son score nutritionnel global

Pourquoi c'est motivant. Dataset massif de **32 millions de lignes** — un vrai défi big data. Le fuzzy matching NLP entre deux bases est un exercice avancé très formateur. L'angle santé/nutrition est personnellement relatable pour des étudiants. Le système de recommandation combine techniques avancées. Pertinence business directe dans un secteur en plein boom (Instacart, Amazon Fresh).

Faisabilité : ✅ **Faisable mais exigeant.** Le fuzzy matching produit-à-produit est le point le plus complexe. Solution de repli : utiliser des jointures au niveau catégorie si le matching produit s'avère trop difficile. ~250h collecte/préparation (dont le matching), ~150h EDA, ~250h modélisation, ~100h validation, ~125h dashboard, ~70h documentation.

📊 Projet 3 : Optimisation de la conversion e-commerce — clickstream massif + supply chain

Concept. Modéliser le funnel d'achat (vue → panier → achat) à partir d'un dataset clickstream de **285 millions d'événements** d'un magasin multi-catégories, (Kaggle) enrichi de données supply chain (DataCo) et de calendrier événementiel. Prédire la probabilité de conversion en temps réel, identifier les facteurs d'abandon, et analyser l'impact du Black Friday 2019.

Source 1 — eCommerce Behavior Data (REES46)

- **URL :** https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store
- **Volume : ~14 Go total**, **285 millions d'événements** répartis sur 2 fichiers (octobre : ~67M lignes, novembre : ~109M lignes incluant le Black Friday).
- **Variables clés (9 colonnes) :** event_time (timestamp à la seconde), event_type (view/cart/remove_from_cart/purchase), product_id, category_id, category_code (hiérarchique : "electronics.smartphone"), brand, price (USD), user_id, user_session.
- **Temporalité :** horodatage à la seconde sur 2 mois (oct–nov 2019), parfait pour l'analyse de sessions et l'impact Black Friday.

Source 2 — DataCo Smart Supply Chain

- **URL :** https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis
- **Volume : ~180 000 lignes**, 53 colonnes, ~80 Mo. (Medium)
- **Variables clés :** dates de commande/expédition, jours réels vs planifiés d'expédition, statut livraison, risque de retard, (Medium) catégorie produit, prix, segment client, mode d'expédition, région, profit par commande.
- **Temporalité :** commandes 2015–2018.

Source 3 — Calendrier jours fériés + Google Trends

- Bibliothèque Python holidays (100+ pays) + pytrends pour l'intérêt de recherche par catégorie produit.

Stratégie de jointure. Les événements REES46 sont agrégés en **features session-level** (durée, nombre de vues, ajouts panier, catégories parcourues, progression dans le funnel) et **features utilisateur** (fréquence d'achat, préférences catégorielles). DataCo fournit des patterns supply chain (relation mode d'expédition/délai → satisfaction) applicables comme features d'enrichissement catégoriel. Les données temporelles (jours fériés, Black Friday, tendances) enrichissent les agrégats journaliers/horaires.

Types d'analyse :

- **Classification binaire** : la session se terminera-t-elle par un achat ?
- **Analyse de funnel** : quantifier les taux de drop-off à chaque étape
- **Détection d'anomalies** : identifier les comportements utilisateurs inhabituels (bots, fraude) via Isolation Forest
- **Séries temporelles** : prévoir les taux de conversion horaires/journaliers, modéliser le surge du Black Friday
- **Modélisation séquentielle** : LSTM ou Transformer sur les séquences d'événements pour prédire l'action suivante

Pourquoi c'est motivant. Dataset véritablement massif (**285M lignes**) — un défi d'ingénierie data authentique nécessitant Dask ou PySpark. Contient le Black Friday 2019 comme expérience naturelle. L'analyse de sessions en temps réel est à la pointe de l'industrie. Directement applicable aux carrières en analytics e-commerce.

Faisabilité : ✅ **Faisable avec cadrage rigoureux.** Les étudiants **doivent** échantillonner les données (recommandé : échantillon de 5–10% des utilisateurs = ~28M événements). L'aspect big data engineering est une expérience d'apprentissage précieuse. Allouer **2 étudiants** spécifiquement au data engineering. ~300h ingénierie données, ~130h EDA, ~250h modélisation, ~100h validation, ~120h dashboard, ~45h documentation.

DOMAINE 2 — SANTÉ PUBLIQUE

🦠 Projet 4 : Prédiction de l'antibiorésistance en Europe à partir des patterns de consommation d'antibiotiques

Concept. Modéliser la relation entre consommation d'antibiotiques et évolution de la résistance antimicrobienne (RAM) à travers **29 pays EU/EEE**. L'OMS qualifie la RAM de « pandémie silencieuse » — elle cause **plus de 35 000 décès/an en Europe**. European Centre for Disease Pr… European Centre for Disease Pr… Le projet combine les données de surveillance ECDC avec les indicateurs socioéconomiques d'Eurostat pour prédire les niveaux futurs de résistance par pays et pathogène.

Source 1 — ECDC EARS-Net (résistance antimicrobienne)

- **URL :** https://atlas.ecdc.europa.eu/public/ (sélectionner "Antimicrobial resistance", export CSV)
- **Volume : ~15 000–50 000 lignes** selon la sélection (multiple pathogènes × pays × années × groupes d'antibiotiques). 30 pays, 8 espèces de pathogènes, ~14 phénotypes de résistance. Epi-net
- **Variables clés :** pays, année, espèce pathogène (E. coli, K. pneumoniae, MRSA, P. aeruginosa, Acinetobacter, S. pneumoniae, Enterococci), groupe antibiotique, % résistant (R), % intermédiaire (I), % sensible (S), nombre d'isolats testés.
- **Temporalité :** données annuelles de **2005 à 2024** (20 ans de séries temporelles).

Source 2 — ECDC ESAC-Net (consommation d'antibiotiques)

- **URL :** https://www.ecdc.europa.eu/en/antimicrobial-consumption/surveillance-and-disease-data/database
- **Volume : ~5 000–20 000 lignes** (29 pays × classes ATC × secteurs communautaire/hospitalier × années 1997–2024).
- **Variables clés :** pays, année, secteur, groupe ATC (pénicillines, céphalosporines, macrolides, quinolones...), consommation en DDD/1 000 habitants/jour, classification OMS AWaRe (Access/Watch/Reserve). PubMed Central
- **Temporalité :** annuelle, **1997–2024** (27 ans).

Source 3 — Eurostat (indicateurs socioéconomiques de santé)

- **URL :** https://ec.europa.eu/eurostat/databrowser/
- **Volume :** ~5 000–15 000 lignes pour les dépenses de santé + densité de population + lits d'hôpital par pays/année.
- **Variables clés :** dépenses de santé par habitant, lits d'hôpital pour 100k, densité de population, PIB par habitant, couverture vaccinale, causes de décès respiratoires/infectieuses.
- **Temporalité :** annuelle, 2000–2023.

Stratégie de jointure. Clé primaire : **code pays ISO + année**. EARS-Net joint à ESAC-Net pour corréler usage d'antibiotiques et résistance. Eurostat ajouté pour les facteurs confondants socioéconomiques. **Jointures décalées possibles** : consommation à l'année T prédisant la résistance à T+1, T+2.

Types d'analyse :

- **Prévision temporelle** : prédire les % de résistance futurs par pays via ARIMA/Prophet/LSTM avec consommation comme variable exogène
- **Régression multivariée** : lier les patterns de consommation (catégories AWaRe) à l'évolution de la résistance
- **Clustering** : grouper les pays par profils RAM/consommation (gradient nord-sud européen) European Centre for Disease Pr…
- **Classification** : classer les pays « en bonne voie » vs « hors cible » par rapport aux objectifs EU 2030
- **Détection d'anomalies** : détecter les pics de résistance inexpliqués par la consommation

Pourquoi c'est motivant. Enjeu critique de santé publique mondiale avec des **cibles politiques EU 2030** concrètes. Données multidimensionnelles avec des gradients géographiques nord-sud et est-ouest fascinants à explorer. Le dashboard pourrait réellement servir de tracker de progrès. Combine épidémiologie, pharmacologie et data science.

Faisabilité : ✅ **Très faisable.** Données bien structurées en CSV/XLSX, logique de jointure straightforward (pays + année). Documentation excellente de l'ECDC. ~100h collecte/nettoyage, ~150h EDA (cartes, heatmaps), ~250h modélisation, ~200h dashboard, ~150h documentation, ~95h buffer.

🌬️ Projet 5 : Impact de la pollution atmosphérique sur les maladies respiratoires — analyse spatiotemporelle européenne

Concept. Quantifier la relation entre mesures de qualité de l'air (PM2.5, PM10, NO2, O3) et outcomes respiratoires (mortalité, hospitalisations) à travers les régions européennes au niveau NUTS2. Le projet combine des **centaines de millions de mesures horaires** de stations de qualité de l'air avec les données de mortalité et d'hospitalisation d'Eurostat.

Source 1 — EEA Air Quality Station Data (via Zenodo)

- **URL :** https://zenodo.org/records/14513586
- **Alternative :** https://eeadmz1-downloads-webapp.azurewebsites.net/
- **Volume : 4,2 Go total**. Données horaires : 3,7 Go (partitionnées par pays, format Parquet), **centaines de millions de lignes**. Données pré-agrégées quotidiennes (204 Mo), mensuelles (~15 Mo), annuelles (~3 Mo).
- **Variables clés :** ID station, code pays, timestamp, NO2 (µg/m³), SO2, O3, PM10, PM2.5, flags de validité, métadonnées station (type : fond/industriel/trafic ; zone : urbain/périurbain/rural ; lat/lon).
- **Temporalité :** mesures horaires **2015–2023** (9 ans). zenodo

Source 2 — Eurostat, causes de décès par région NUTS2

- **URL :** https://ec.europa.eu/eurostat/databrowser/view/hlth_cd_asdr2/default/table?lang=en
- **Volume : ~50 000–200 000 lignes** (30+ pays × régions NUTS2 × causes × sexe × tranches d'âge × années).
- **Variables clés :** code NUTS2, année, cause de décès (CIM-10 : J00-J99 respiratoire, J09-J18 grippe/pneumonie, J40-J47 maladies respiratoires chroniques basses, J45-J46 asthme), sexe, tranche d'âge, taux standardisé de mortalité pour 100 000.
- **Temporalité :** annuelle 2011–2022 (niveau NUTS2), mensuelle 2019–2023 (niveau national). Europa

Source 3 — Eurostat, hospitalisations pour maladies respiratoires

- **URL :** https://ec.europa.eu/eurostat/databrowser/product/page/HLTH_CO_DISCH2

- **Volume :** ~10 000–30 000 lignes.
- **Variables clés :** pays, année, diagnostic CIM, sexe, taux de sortie d'hôpital pour 100 000, durée moyenne de séjour.
- **Temporalité :** annuelle, 2010–2021.

Stratégie de jointure. Agrégation spatiale des stations de qualité d'air vers les régions NUTS2 via géocodage des coordonnées GPS. Calcul des concentrations moyennes régionales annuelles/mensuelles. Jointure temporelle avec les données Eurostat sur (région NUTS2, année). Enrichissement avec les données de population et d'urbanisation par région NUTS2.

Types d'analyse :

- **Régression panel** : estimer l'effet de PM2.5/NO2 sur la mortalité respiratoire en contrôlant les facteurs confondants
- **Séries temporelles** : mortalité mensuelle vs pollution mensuelle avec effets de lag (courbes exposition-réponse)
- **Clustering spatial** : identifier les régions à haut risque combinant exposition à la pollution et vulnérabilité sanitaire
- **Détection d'anomalies** : détecter les épisodes de pollution (poussière saharienne, feux de forêt, canicules) et corréler avec les pics d'hospitalisation
- **Prévision** : prédire les tendances de mortalité respiratoire sous différents scénarios d'amélioration de la qualité de l'air

Pourquoi c'est motivant. Dataset véritablement massif (centaines de millions de mesures). Forte composante cartographique et de mapping (cartes européennes des hotspots). Pertinence politique directe : les normes EU de qualité de l'air sont activement débattues. Travailler avec le format **Parquet et des données géospatiales** enseigne des compétences modernes en data engineering.

Faisabilité : ✅ **Faisable mais techniquement ambitieux.** Les fichiers pré-agrégés quotidiens/mensuels/annuels sur Zenodo simplifient grandement le traitement. Les étudiants apprendront Parquet, les jointures géospatiales et les méthodes de données panel. ~200h prétraitement, ~150h EDA, ~250h modélisation, ~200h dashboard cartographique, ~100h documentation, ~45h buffer.

🥵 Projet 6 : Système de prévision épidémique de la grippe en France — multi-sources

Concept. Construire un système de prévision de la grippe pour la France en combinant le réseau Sentinelles (40+ ans de surveillance ILI hebdomadaire), les données des urgences hospitalières et SOS Médecins (data.gouv.fr), et les données météorologiques de Météo France. Prédire le début, le pic et l'intensité des épidémies au niveau régional pour la planification hospitalière.

Source 1 — Réseau Sentinelles (incidence grippe hebdomadaire)

- **URL :** https://www.sentiweb.fr/france/en/?page=database

- **Mirror data.gouv.fr :** https://www.data.gouv.fr/fr/datasets/estimation-dincidence-des-syndromes-grippaux/
- **Volume : ~30 000–50 000 lignes** pour l'extraction multi-maladie, multi-région (données hebdomadaires depuis 1984 pour la grippe, 13 régions métropolitaines).
- **Variables clés :** semaine ISO, année, région, taux d'incidence estimé pour 100 000, bornes IC 95%, seuil épidémique, surveillance virologique (% positifs influenza A/B, VRS, SARS-CoV-2), distribution des sous-types viraux.
- **Temporalité :** hebdomadaire, **1984–2026** (40+ ans au national ; régional depuis 2004). (Sentinelles)

Source 2 — Urgences hospitalières et SOS Médecins (data.gouv.fr)

- **URL :** https://www.data.gouv.fr/datasets/donnees-des-urgences-hospitalieres-et-de-sos-medecins-relatives-a-lepidemie-de-covid-19
- **Volume : ~116 000+ lignes** (données quotidiennes × départements × sexe × tranches d'âge, mars 2020 à 2024).
- **Variables clés :** date, code département/région, sexe, tranche d'âge (0-4, 5-14, 15-44, 45-64, 65-74, 75+), nombre de passages aux urgences pour suspicion respiratoire, hospitalisations depuis les urgences, consultations SOS Médecins.
- **Temporalité :** quotidienne, mars 2020–2024.

Source 3 — Météo France Open Data

- **URL :** https://donneespubliques.meteofrance.fr/ — aussi via https://www.data.gouv.fr (chercher "données climatologiques")
- **Alternative :** ECMWF ERA5 via Copernicus CDS (https://cds.climate.copernicus.eu/)
- **Volume : millions de lignes** de données horaires/quotidiennes de centaines de stations sur des décennies.
- **Variables clés :** température (min/max/moy), humidité relative, précipitations, vitesse du vent, rayonnement solaire, point de rosée.
- **Temporalité :** quotidienne/horaire, remontant à plusieurs décennies.

Stratégie de jointure. Agrégation des données quotidiennes SOS Médecins/Urgences en totaux hebdomadaires pour s'aligner avec les données Sentinelles. Clé spatiale : code région française (13 régions métropolitaines). Moyennes météo hebdomadaires régionales calculées à partir des stations locales. **Variables décalées** : features météo avec 1–3 semaines de lag (le froid/l'humidité précède les vagues grippales). Période de chevauchement principal : 2020–2024 pour les 3 sources ; extension possible avec Sentinelles + météo seuls sur des décennies.

Types d'analyse :

- **Prévision temporelle** : prédire l'incidence ILI 1–3 semaines à l'avance via ARIMA/Prophet/LSTM avec variables exogènes météo + urgences
- **Détection de début d'épidémie** : classification binaire — la semaine courante marque-t-elle le début d'une épidémie ? (signaux multi-sources)

- **Régression** : quantifier la contribution des facteurs météorologiques (température, humidité) à l'incidence grippale
- **Détection d'anomalies** : détecter les pics anormaux de passages aux urgences pouvant indiquer l'émergence d'un nouveau pathogène (comme COVID en mars 2020)
- **Analyse spatiale** : cartographie régionale du risque montrant la progression épidémique à travers la France

Pourquoi c'est motivant. Miroir direct des systèmes réels de surveillance sanitaire utilisés par Santé publique France. Contexte français avec des données ouvertes riches et bien documentées. Pertinence immédiate : la grippe reste une cause majeure de surcharge des urgences chaque hiver. La période COVID (2020–2024) fournit des **expériences naturelles** (confinements, changements comportementaux) ajoutant une richesse analytique unique.

Faisabilité : ✅ **Très faisable.** Les données Sentinelles sont propres et bien structurées. Le principal défi est l'intégration des 3 sources — excellente valeur pédagogique. Le projet reproduit fidèlement les systèmes de surveillance du monde réel. ~150h collecte/prétraitement, ~150h EDA, ~250h modélisation, ~200h dashboard, ~100h documentation, ~95h buffer.

DOMAINE 3 — ÉNERGIE / ENVIRONNEMENT

🔥 Projet 7 : Prédiction de la cascade feux de forêt → qualité de l'air

Concept. Prédire comment les feux de forêt actifs dégradent la qualité de l'air régionale (concentrations PM2.5) en combinant les détections satellite NASA, les capteurs de qualité d'air au sol EPA, et les observations météorologiques NOAA. Le système prévoit les événements dangereux 1–3 jours à l'avance en modélisant le transport de fumée via la proximité des feux, leur intensité et les patterns de vent.

Source 1 — NASA FIRMS Active Fire Data (MODIS/VIIRS)

- **URL :** https://firms.modaps.eosdis.nasa.gov/download/
- **Aussi sur Kaggle :** https://www.kaggle.com/datasets/vijayveersingh/nasa-firms-active-fire-dataset-modisviirs
- **Volume : millions de détections par an** globalement. Pour les USA seuls, centaines de milliers par an. **>5 millions de lignes** pour le sous-ensemble US sur 10+ ans.
- **Variables clés :** latitude, longitude, brightness (Kelvin), date/heure d'acquisition, satellite, instrument, confidence (%), FRP (Fire Radiative Power en MW), jour/nuit.
- **Temporalité :** détections quotidiennes avec heure exacte. MODIS depuis nov 2000, VIIRS depuis jan 2012.

Source 2 — EPA AQS Hourly Air Quality (PM2.5)

- **URL :** https://aqs.epa.gov/aqsweb/airdata/download_files.html
- **Exemple direct :** https://aqs.epa.gov/aqsweb/airdata/hourly_88101_2023.zip
- **Volume : ~5–10 millions de lignes par an** pour le PM2.5 horaire national. Source **très volumineuse et bruitée**.
- **Variables clés (29 colonnes) :** codes État/comté/site, coordonnées, date/heure locale et GMT, mesure (µg/m³), incertitude, qualificateurs, type de méthode, noms géographiques.
- **Temporalité :** mesures horaires de ~2 500+ moniteurs à travers les USA.

Source 3 — NOAA ISD-Lite Hourly Weather

- **URL :** https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database
- **Bulk :** https://registry.opendata.aws/noaa-isd/
- **Volume :** >35 000 stations globalement, ~14 000 actives. Millions de relevés horaires par an.
- **Variables clés :** température, point de rosée, pression, direction/vitesse du vent, couverture nuageuse, précipitations.
- **Temporalité :** horaire, couverture historique extensive.

Stratégie de jointure. Jointure **spatiotemporelle** : pour chaque moniteur EPA, calculer les distances à tous les feux actifs FIRMS dans un rayon configurable (50–300 km) et une fenêtre temporelle (0–72h avant). Features créées : nombre de feux dans le rayon, distance au feu le plus proche, FRP total, FRP pondéré par distance. Enrichissement météo : joindre NOAA de la station la plus proche à chaque moniteur AQ. Utiliser direction + vitesse du vent pour déterminer si la fumée est transportée vers le moniteur. Features avancées : inversions de température (point de rosée vs température), lags (feux 6h, 12h, 24h, 48h, 72h avant).

Types d'analyse :

- **Régression** : prédire l'augmentation de PM2.5 attribuable à la fumée de feux
- **Classification binaire** : le PM2.5 dépassera-t-il le seuil "Unhealthy for Sensitive Groups" (>35.5 µg/m³) dans les prochaines 24/48h ?
- **Prévision temporelle** : forecasting multi-pas du PM2.5
- **Détection d'anomalies** : distinguer les événements de fumée des pics de pollution urbaine
- **Analyse spatiale** : cartographier dynamiquement les zones d'impact de la fumée

Pourquoi c'est motivant. Impact direct sur la santé publique — la fumée des feux de forêt est une crise croissante. Utilise des données de **télédétection satellite** (excitant pour des étudiants en informatique). Implique des calculs spatiaux non triviaux (géo-distance, analyse directionnelle, indexation KD-tree). Pertinence changement climatique. Potentiel de dashboard impressionnant avec visualisations cartographiques en temps réel.

Faisabilité : ✅ **Faisable.** Pipeline clair. Data engineering : téléchargement et parsing des 3 sources, gestion des données manquantes (~200h). Moteur de jointure spatiale : indexation KD-tree/BallTree (~160h).

Modélisation : gradient boosting + LSTM (~200h). Dashboard : carte interactive Folium/Plotly (~160h). Validation + documentation (~225h).

⚡ Projet 8 : Prévision des prix négatifs de l'électricité liés à la surproduction renouvelable en Europe

Concept. Prédire quand les prix de l'électricité européens deviennent **négatifs** (sous 0 €/MWh) — un phénomène fascinant qui se produit quand la production éolienne et solaire dépasse la demande. On est littéralement **payé pour consommer** ! Le projet combine les données horaires du marché électrique avec la météo pour construire un système de forecasting pour les opérateurs de réseau et les traders d'énergie.

Source 1 — Open Power System Data (OPSD) — Time Series

- **URL :** https://data.open-power-system-data.org/time_series/2020-10-06/
- **ZIP complet :** https://data.open-power-system-data.org/time_series/opsd-time_series-2020-10-06.zip (277 Mo)
- **CSV horaire :** https://data.open-power-system-data.org/time_series/2020-10-06/time_series_60min_singleindex.csv
- **Volume : ~289 000 lignes** (timestamps horaires jan 2015 – mi-2020) × **500+ colonnes** (variables × 32 pays). La résolution 15 min contient ~1,16 M lignes.
- **Variables clés par pays :** charge réelle/prévue (MW), génération solaire/éolienne réelle (MW), capacités installées, **prix day-ahead (EUR/MWh)**, profils solaire/éolien.
- **Pays couverts :** 32 (AT, BE, BG, CH, CZ, **DE**, DK, EE, ES, FI, **FR**, GB, GR, HR, HU, IE, IT, LT, LU, LV, ME, NL, NO, PL, PT, RO, RS, SE, SI, SK).
- **Temporalité :** résolutions 15 min, 30 min, 60 min. **Période : 2015–2020.** Licence CC-BY 4.0.

Source 2 — OPSD Weather Data (ERA5)

- **URL :** https://data.open-power-system-data.org/weather_data/
- **Volume :** plusieurs Go. Données météo grillées au niveau NUTS-2 pour les pays européens.
- **Variables clés :** température (°C), radiation directe/diffuse/globale (W/m²), vitesse du vent (m/s), précipitations, chutes de neige, densité de l'air.
- **Temporalité :** horaire, correspondant à la période des séries temporelles. CC-BY 4.0.

Source 3 — ENTSO-E Transparency Platform (complémentaire)

- **URL :** https://transparency.entsoe.eu/
- **Variables clés :** génération réelle par type de production (nucléaire, gaz, charbon, éolien, solaire, hydro...), flux physiques transfrontaliers, prix day-ahead et intraday, données d'équilibrage.
- **Volume :** données horaires pour tous les États membres EU depuis 2015. Millions d'enregistrements.
- **Temporalité :** horaire/15 min, 2015–présent. Accès gratuit avec inscription.

Stratégie de jointure. Fusion temporelle directe : OPSD time series et weather data partagent les mêmes timestamps horaires et codes pays → merge direct sur (timestamp, pays). ENTSO-E ajoute la décomposition par type de combustible et les flux transfrontaliers. Feature engineering : **ratio de pénétration renouvelable** (gen_renouvelable / charge_totale), position nette d'export, heure/jour/mois, prix décalés, moyennes glissantes de production éolienne/solaire, erreurs de prévision météo.

Types d'analyse :

- **Classification binaire** : le prix day-ahead sera-t-il négatif demain ? (Très déséquilibré — excellent défi ML)
- **Régression** : prédire le prix day-ahead à partir des prévisions météo et de génération
- **Prévision temporelle multi-horizon** : 1h, 6h, 24h, 168h pour les prix et la production renouvelable
- **Clustering** : identifier les états typiques du réseau (fort-vent-faible-demande, canicule-forte-demande, etc.)
- **Analyse causale** : quantifier comment chaque GW supplémentaire de capacité renouvelable affecte la distribution des prix

Pourquoi c'est motivant. Les prix négatifs sont **contre-intuitifs et fascinants** — un hook parfait pour une présentation orale. Directement pertinent pour la transition énergétique et la politique climatique. Applications financières (stratégies de trading). Visualisations riches (graphiques de mix énergétique en aires empilées, heatmaps de prix, dashboards de comparaison par pays). **Données pré-nettoyées** et bien documentées par un consortium académique (TU Berlin, ETH Zürich).

Faisabilité : ✅ **Très faisable — le projet le plus "prêt à l'emploi".** Les données OPSD sont pré-nettoyées et structurées pour la recherche. Les étudiants peuvent se concentrer sur 2–3 pays (Allemagne — le plus de prix négatifs, Danemark — plus forte pénétration éolienne, France — dominé par le nucléaire). Split temporel clair (train 2015–2018, validation 2019, test 2020). Les étapes du pipeline mappent proprement sur les rôles : data engineering (2), EDA + feature engineering (2), modélisation (2), dashboard (1).

🎄 Projet 9 : Prédiction du risque de feux de forêt aux USA — données spatiotemporelles massives

Concept. Construire un système de prédiction spatiotemporelle des feux de forêt en combinant un dataset massif pré-jointé (grille GRIDMET + IRWIN, 9,5 millions de lignes), les archives historiques de 2,3 millions de feux américains (1992–2020), et les données météo NOAA. Prédire où et quand les feux sont les plus susceptibles de survenir et leur sévérité attendue.

Source 1 — US Wildfire Dataset (2014–2025) — GRIDMET+IRWIN

- **URL :** https://www.kaggle.com/datasets/firecastrl/us-wildfire-dataset
- **Volume : ~9,5 millions de lignes** (cellules de grille spatiotemporelles avec labels feu/pas feu).
- **Variables clés :** localisation de la cellule (lat/lon), date, features météo GRIDMET (température, précipitations, humidité, vent, radiation solaire, évapotranspiration, composante de libération d'énergie, indice de combustion), label d'occurrence de feu IRWIN.

- **Temporalité :** quotidien, **2014–2025**.

Source 2 — 2.3 Million US Wildfires (1992–2020)

- **URL Kaggle :** https://www.kaggle.com/datasets/behroozsohrabi/us-wildfire-records-6th-edition
- **Source originale USDA :** https://www.fs.usda.gov/rds/archive/catalog/RDS-2013-0009.6
- **Volume : ~2,3 millions d'enregistrements de feux**, base SQLite ~1,7 Go.
- **Variables clés :** nom du feu, année, date de découverte, cause (foudre, incendie criminel, feu de camp...), **taille en acres**, classe de taille (A–G), coordonnées, État, comté, date de maîtrise.
- **Temporalité :** chaque enregistrement a une date de découverte et de maîtrise (1992–2020). Dataset peer-reviewed publié dans *Scientific Data*.

Source 3 — NOAA ISD-Lite Hourly Weather

- **URL :** https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database
- Même source que le Projet 7 — données horaires de milliers de stations.

Stratégie de jointure. Le Dataset A est déjà un tableau ML spatiotemporel grillé — c'est la table d'analyse principale. Enrichissement historique depuis le Dataset B : agréger les 2,3M feux historiques en cellules de grille pour créer des features de fréquence et sévérité historique par localisation (feux par décennie, taille moyenne, cause dominante). Jointure spatiale. Matching station météo NOAA pour ajouter des observations temps réel complémentaires à GRIDMET. Feature engineering additionnel : jours consécutifs sans pluie (proxy sécheresse), degrés-jours cumulés, historique saisonnier des feux par localisation.

Types d'analyse :

- **Classification binaire** : un feu se produira-t-il dans cette cellule dans les 1/3/7 prochains jours ? (Dataset **très déséquilibré** — exercice ML formateur)
- **Régression** : prédire la taille attendue du feu (acres) selon les conditions d'ignition
- **Classification multi-classes** : prédire la classe de taille (A à G)
- **Clustering spatial** : identifier les zones à risque et leurs patterns météo caractéristiques
- **Analyse temporelle** : détecter les tendances dans la durée et l'intensité de la saison des feux sur les décennies
- **Analyse de survie** : modéliser le temps de maîtrise selon les caractéristiques du feu et de la météo

Pourquoi c'est motivant. Les feux de forêt sont l'une des catastrophes climatiques les plus dramatiques. Dataset principal de **9,5 millions de lignes** — vrai défi big data. Le problème de classification fortement déséquilibré enseigne des concepts ML importants (SMOTE, focal loss, métriques adaptées). Potentiel de **visualisations géospatiales spectaculaires** (heatmaps de risque, animations de propagation). Directement actionnable : les résultats pourraient informer le pré-positionnement des ressources de lutte contre les incendies.

Faisabilité : ✅ **Très faisable.** Le dataset A est **déjà pré-jointé et prêt pour l'analyse** — pas de jointure complexe nécessaire pour le modèle de base. Les datasets B et C offrent des opportunités d'enrichissement pour

les analyses avancées. Approche progressive possible : commencer avec A seul pour un baseline, puis ajouter B et C. Sous-ensemble géographique possible (ex : Californie, Pacific Northwest). ~200h data engineering, ~150h EDA, ~250h modélisation, ~200h dashboard, ~145h documentation/validation.

Tableau comparatif des 9 projets

#	Domaine	Sujet	Volume principal	Sources	Temporalité	Difficulté	Originalité
1	Commerce	Olist Brésil — livraison + satisfaction	1M lignes (géoloc) + 100k commandes	3	2 ans, timestamps	★★★☆☆	★★★★☆
2	Commerce	Instacart × Open Food Facts — nutrition	**32M lignes**	3	Relative + mensuelle	★★★★☆	★★★★★
3	Commerce	Clickstream REES46 + supply chain	**285M événements**	3	Secondes, 2 mois	★★★★★	★★★★★
4	Santé	Antibiorésistance Europe	~50k lignes	3	20 ans annuel	★★★☆☆	★★★★★
5	Santé	Pollution air × maladies respiratoires	**100M+ mesures**	3	9 ans horaire	★★★★★	★★★★☆
6	Santé	Grippe France multi-sources	~200k lignes	3	40+ ans hebdo	★★★☆☆	★★★★☆
7	Énergie	Feux de forêt → qualité de l'air	**5M+/an** (EPA)	3	Horaire, 10+ ans	★★★★☆	★★★★★
8	Énergie	Prix négatifs électricité renouvelable	289k × 500 cols	3	Horaire, 5 ans	★★★☆☆	★★★★★
9	Énergie	Risque feux de forêt USA	**9,5M lignes**	3	Quotidien, 11 ans	★★★★☆	★★★★☆

Recommandations finales pour le choix

Trois projets se distinguent comme **les meilleurs choix** selon les contraintes exprimées :

Pour maximiser la faisabilité et la répartition des rôles → Projet 8 (Prix négatifs de l'électricité). Les données OPSD sont pré-nettoyées, la documentation est excellente, et le sujet des prix négatifs est un hook narratif puissant pour la présentation orale. Chaque rôle de l'équipe a un périmètre clair et bien défini. C'est le projet le plus "prêt à l'emploi".

Pour maximiser l'originalité et l'impact → Projet 4 (Antibiorésistance en Europe). Sujet de santé publique critique avec des enjeux politiques EU 2030 concrets. Données bien structurées, jointures simples (pays + année), et le gradient géographique européen offre des visualisations cartographiques saisissantes. Excellent ratio originalité/difficulté.

Pour maximiser le défi technique et l'apprentissage → Projet 2 (Instacart × Open Food Facts). Le fuzzy matching entre deux bases massives est un exercice avancé hautement formateur. Le dataset de 32 millions de lignes impose une discipline d'ingénierie data sérieuse. L'angle nutritionnel rend le sujet personnellement engageant et le système de recommandation hybride est un livrable impressionnant.

Quel que soit le choix, chaque projet a été conçu pour permettre une **montée en charge progressive** : commencer avec un modèle baseline sur la source principale, puis enrichir avec les sources secondaires. Cette approche itérative protège l'équipe contre le risque d'échec tout en laissant de la marge pour aller plus loin.