04 Traitement de données ouvertes

Durée
2h ou 3h30
(Hors éval. B)

Thématique

DONNEES STRUCTUREES

Activité Branchée

Description de l'activité

Dans cette activité, on manipule des données ouvertes provenant du site data.gouv.fr afin d'en extraire de l'information pertinente.

Objectifs pédagogiques ou compétences

Objectifs généraux	Objectifs intermédiaires	Compétences
Notions de cours	- Comprendre l'importance des données et de leur traitement automatique - Revoir / Découvrir quelques manipulations de base des données (Excel – trier / filtrer / croiser)	- Manipuler des données ouvertes - Traiter des données structurées avec un tableur - Traiter automatiquement des données avec un programme informatique

Matériel et outils

- Fiche élève à imprimer
- 1 poste par élève

Tags

#données structurées ; #données ouvertes ; #tableur ; #filtrer ; #trier ; #croiser

Déroulé de l'activité

Introduction : *(<20 minutes)*

- **Présenter les objectifs de la séance (contenu théorique et productions attendues)** *(2-3 minutes)*
- **Introduire la thématique des données :** *(~15 minutes)*

Pour lancer la thématique, on propose aux élèves une activité pour identifier les données qu'ils « créent » dans leur quotidien. En petits groupes, ils listent leurs activités en ligne (Recherche, utilisation d'outils, RS, jeux, …) et, pour chacune, ils cherchent qui y a accès et comment elles sont utilisées.

Les données sont créées chaque fois que des événements, des actions ou des mesures sont enregistrés. Elles permettent de comprendre, analyser et prendre des décisions dans de nombreux domaines, mais posent également des questions juridiques ou encore d'éthique, notamment pour ce qui est de la protection de la vie privée.

Pistes de réflexion :

Réseaux sociaux

- **Entités :** Facebook, Twitter/X, Instagram, Tiktok, Twitch, …
  - Personnalisation du contenu pour une expérience utilisateur améliorée.
  - Analyse des comportements pour un meilleur design (UX design)
  - Collecte de données personnelles pour le ciblage publicitaire.
  - Risque de manipulation de l'opinion publique à travers des bulles de filtre.
  - Accès étendu via la multi-connexion (Google Accounts, Twitter/X, …)

Recherche en ligne

- **Entités :** Google, Bing, DuckDuckGo, …
  - Fourniture de résultats de recherche pertinents en fonction des intérêts de l'utilisateur.
  - Amélioration des algorithmes de recherche grâce à l'apprentissage automatique.
  - Collecte de données de recherche pour la création de profils publicitaires.
  - Risque de biais dans les résultats de recherche en fonction des intérêts commerciaux.

Jeux vidéo en ligne

- **Entités :** Steam, Epic Games, Xbox / Playstation / Nintendo / …, …
  - Création d'expériences de jeu multijoueur fluides.
  - Analyse des comportements pour un meilleur design (UX design)
  - Adaptation graphique des avatars et des environnements de jeu en fonction des tendances.
  - Adaptation des designs (game design, ux design, …) pour créer des systèmes addictifs et encourageant les microtransactions.

Achats en ligne

- **Entités :** Amazon, eBay, Alibaba, Vinted, Aliexpress, …
    - Recommandation de produits basée sur les préférences de l'utilisateur.
    - Facilité d'achat et de livraison grâce aux données de localisation, données bancaires, etc.
    - Ciblage publicitaire.

Applications de messagerie

- **Entités :** WhatsApp, Facebook Messenger, Signal, …
    - Communication instantanée et gratuite avec des contacts dans le monde entier.
    - (Pour certains) Sécurisation des messages grâce au chiffrement de bout en bout.
    - Collecte de métadonnées de communication pour le ciblage publicitaire ou encore l'entraînement d'IA conversationnelles (article sur X, article sur Gmail et Instagram).
    - Risque de violation de la vie privée en cas de failles de sécurité.

Services de streaming

- **Entités :** Netflix, Spotify, YouTube, …
    - Recommandation de contenu personnalisé pour les utilisateurs, mais aussi publicités ciblées.
    - Risque de dépendance au contenu en streaming via l'UX design, le choix des séries à recommander, etc.
    - Surveillance des préférences de visionnage pour la publicité ciblée.

(Télé)travail

- **Entités :** Zoom, Microsoft Teams, Slack
    - Risque de surveillance excessive des employés en télétravail.
    - Possibilité de fuites de données sensibles.
- Employeurs (générique) :
    - Suivi de la productivité pour l'amélioration des performances, déceler des erreurs, etc. (article sur Amazon France)

Étape 1 – Traitement de données avec tableur *(~1h30)*

- **Présentation du site** *(5 minutes)*

En conclusion de l'activité précédente, l'enseignant.e explique qu'il existe également de nombreux sites qui mettent à disposition des données ouvertes (Open Data), accessibles gratuitement et à tous (liste des portails open data). On peut leur demander s'ils en connaissent un français (peut-être vu lors d'un autre cours) : data.gouv.fr.

L'enseignant.e explique ensuite le fonctionnement du site en présentant plusieurs types de données recueillies dans les thèmes à la une, puis en expliquant comment on peut faire des recherches plus poussées en cliquant sur

« Données » tout en montrant l'exemple avec « salles de cinéma », ce qui permettra de retrouver la feuille .csv qui servira pour l'exercice.

Feuille : https://www.data.gouv.fr/fr/datasets/les-salles-de-cinema-en-ile-de-france/ , mise à jour le 6 novembre 2021.

- **Importation des données ouvertes** *(10 minutes)*

Les élèves vont sur le site et tapent les mots-clés « véhicules CO^2 », puis accèdent à la page de la fiche intitulée « Emissions de CO2 et de polluants des véhicules commercialisés en France ». Ils vont récupérer la feuille CSV du même nom ainsi que le Dictionnaire des variables au format XLS et le fichier ZIP.

- **Filtrer et trier** *(~1h15)*

Les premières questions permettent de se familiariser avec les fichiers et la recherche rapide.

L'enseignant.e va montrer ensuite une série de manipulations à faire pour le filtrage et le tri des données :

- Avant de montrer la manipulation, on demande aux élèves comment ils procéderaient pour lister les voitures hybrides de 2 marques uniquement.
- S'ils n'ont pas la piste, on leur pose des questions pour les guider, le but ici étant d'introduire la notion de filtre et tri, voire traitement automatique, et on montre la fonction tri du logiciel.

Étape 2 – Données ouvertes : découpages administratifs *(~1h30)*

Cette partie est facultative, en fonction de la motivation des élèves. Dans cette activité, ils vont manipuler des données de l'INSEE en croisant les données, et manipuler une formule, ce qui permettra de les initier à l'automatisation des données.

Avant de passer à la conclusion, on peut notamment expliquer davantage en quoi consiste le traitement automatique des données, discuter des avantages, éventuellement des inconvénients ou réserves qu'on peut émettre.

Proposition :

Le traitement automatique des données (TAD), ou traitement automatisé de l'information, fait référence au processus par lequel des machines (ordinateurs et logiciels spécialisés), sont utilisées pour collecter, stocker, manipuler, analyser et interpréter des données numériques. Ce processus est devenu essentiel dans de nombreux domaines : affaires, recherche scientifique, médecine, industrie, entreprises, sport, loisirs, …

Avantages :

- **Vitesse et efficacité :** Les machines peuvent traiter d'énormes quantités de données beaucoup plus rapidement que les êtres humains, ce qui permet d'obtenir des résultats en temps réel ou dans des délais très courts.

- **Précision :** Les machines sont moins sujettes aux erreurs humaines. Une fois correctement programmées, elles effectuent des calculs et des analyses de manière cohérente et précise.
- **Stockage et récupération faciles :** Les données numériques peuvent être stockées de manière organisée et facilement récupérées lorsque nécessaire, ce qui permet de gagner du temps et d'éviter la perte de données.
- **Analyse avancée :** Les logiciels de traitement de données peuvent effectuer des analyses complexes qui servent de base aux Intelligences Artificielles, telles que l'apprentissage automatique, la modélisation statistique et l'exploration de données, pour extraire des informations précieuses et des tendances à partir des données.
- **Automatisation des tâches répétitives :** Le TAD peut automatiser des tâches routinières, et ainsi, nous évite d'avoir à le faire nous même (dans l'activité, les élèves auraient dû faire des allers-retours dans les 2 fichiers pour les 37 993 communes).
- **Accessibilité :** Les données peuvent être accessibles à partir de n'importe quel endroit avec une connexion Internet, ce qui facilite le travail à distance et la collaboration entre équipes dispersées géographiquement.

Inconvénients et réserves potentiels :

- **Dépendance technologique :** La dépendance à l'égard de la technologie peut entraîner des problèmes en cas de défaillance matérielle ou de pannes informatiques, et bloquer de plus en plus de secteurs. Avec l'émergence de machines autonomes et des objets connectés (transports en commun, maisons avec volets, portes et autres objets connectés par exemple), mais aussi l'usage des données dans des métiers tels que ceux de la santé et l'administration, il ne s'agit pas seulement de quelques entreprises bloquées pour 2 ou 3 heures, mais d'un réel impact sur nos vies.
- **Sécurité :** Les données numériques sont vulnérables aux menaces de sécurité (piratage informatique, logiciels malveillants, vol de données, …).
- **Protection de la vie privée :** La collecte et le traitement automatisés des données soulèvent des questions concernant la vie privée et la confidentialité. Les organisations doivent être conformes aux réglementations de protection des données, telles que le RGPD en Europe, pour éviter des problèmes juridiques. La question du réel respect des données se pose également (article sur Amazon France).
- **Biais algorithmique :** Les systèmes automatisés peuvent être biaisés si les données utilisées pour les former sont elles-mêmes biaisées. Cela peut entraîner des discriminations injustes, en particulier dans les domaines tels que les prêts bancaires, la justice pénale et l'embauche.
- **Perte de compétences manuelles :** L'automatisation excessive peut entraîner une perte de compétences manuelles, car les employés peuvent devenir dépendants des systèmes automatisés et perdre la capacité de résoudre manuellement certains problèmes.
- **Déshumanisation :** Dans certaines industries, l'automatisation peut entraîner la suppression de postes de travail, ce qui peut avoir des implications sociales, économiques et psychologiques.

Conclusion *(15 minutes)*

- **Bilan de la séance :** *(5 minutes)*

Pour clôturer la séance, on peut revenir sur les principales difficultés rencontrées pendant l'activité. Éventuellement, il est possible de finir sur un court échange autour :

- **L'importance de l'automatisation des données**

On peut, dans un premier temps, revenir sur le TAD si besoin.

- **Les métiers en lien** ***(10 minutes)***

On peut également évoquer les principaux métiers en lien avec l'analyse et le traitement des données. En fonction du temps, on peut demander aux élèves de faire des recherches, et éventuellement construire leur propre « top 3 » des métiers qui les intéressent dans ce domaine. Voici quelques exemples :

  - **Analyste de données :** Analyse les données pour identifier des tendances, des modèles et des informations utiles pour la prise de décision.
  - **Data Scientist :** Utilise des techniques avancées d'analyse de données, de machine learning et de statistiques pour extraire des prédictions à partir des données.
  - **Ingénieur en traitement des données :** Conçoit et développe des systèmes de gestion des données, des bases de données et des pipelines de traitement des données.
  - **Analyste de business intelligence :** Collecte, organise et analyse des données pour aider les entreprises à prendre des décisions stratégiques et à surveiller leurs performances.
  - **Ingénieur en apprentissage automatique :** Développe des modèles d'apprentissage automatique pour automatiser des tâches et créer des systèmes intelligents.
  - **Analyste de données marketing :** Analyse les données marketing pour évaluer l'efficacité des campagnes, identifier les segments de clientèle et améliorer les stratégies marketing.
  - **Analyste financier :** Utilise les données financières pour évaluer les performances de l'entreprise, prévoir les tendances économiques et recommander des stratégies d'investissement.
  - **Data Engineer :** Gère l'infrastructure de données, construit et optimise les pipelines de données et assure la qualité des données pour les analyses ultérieures.
  - **Architecte de données :** Conçoit l'architecture des systèmes de données, définit les flux de données et garantit la cohérence et l'intégrité des données.
  - **Analyste en cyber-sécurité :** Analyse les données de sécurité pour détecter et prévenir les menaces, les attaques et les vulnérabilités dans les systèmes informatiques.
  - **Data Analyst en santé :** Analyse les données médicales pour identifier des modèles de santé, améliorer les soins aux patients et prendre des décisions éclairées en matière de santé.
  - **Analyste en sciences sociales :** Utilise les données pour étudier les comportements humains, les tendances sociales et les modèles sociétaux.

Evaluation B :

On fournit aux élèves un nouveau fichier (par exemple : **Base des zones d'emploi**), et on pose des questions similaires (lecture de descripteurs, adaptation de la formule pour reporter les noms des régions).

Traitement des données ouvertes

Fiche activité - *Correction*

Étape 1 – Données ouvertes : véhicules et émission de CO^2

- **Importation des données ouvertes**
  - Allez sur le site data.gouv.fr, puis saisissez les mots-clés « véhicule CO2 » dans le champ de recherche.
  - Téléchargez le jeu de données sur les émissions de CO2 et de polluants des véhicules commercialisés en France en 2014. Quel est ce format ? Quel type de séparateur est utilisé ? (*Ouvrez-le avec un Notepad++ par exemple).*

[2014] Emissions de polluants, CO2 et caractéristiques des véhicules commercialisés en France

Mis à jour le 7 juillet 2014 — csv — 2339 téléchargements

CSV, le séparateur étant le « ; »

Les fichiers CSV sont des fichiers texte qui permettent de manipuler des données en colonnes, comme peut le faire un fichier de tableur type Excel. CSV signifie « Comma Separated Values », ce qui se traduit par « Valeurs séparées par des virgules. Les sauts de ligne correspondent, elles, aux lignes du tableau. »

  - Téléchargez également le dictionnaire des variables. Quel est son format ?

Dictionnaire des variables

Mis à jour le 4 juillet 2014 — xls — 1380 téléchargements

XLS, extension de nom de fichier pour tableur au format de Microsoft Excel.

La différence entre le format csv et xls est que, en principe, la quantité de données stockées est illimitée pour le premier, contrairement à Excel, qui ne peut stocker que 1048576 lignes×16384 colonnes de données au maximum.

  - Décompressez le fichier ZIP dans votre répertoire de travail (clic droit -> Extraire...), puis ouvrez dans un tableur les fichiers « carlab-annuaire-variable » et « fic_etiq_edition ».

Remarque : En général, l'importation de ce fichier se passe bien car il est reconnu par le tableur. Sinon, une boîte de dialogue s'ouvre et il faut procéder à certains réglages.

  - Pourquoi l'un des fichiers est-il « ouvert » et l'autre « importé » ?

L'extension XLSX est une extension tableur, alors que CSV non.

  - Combien de lignes et de colonnes remplies y a-t-il dans le fichier « fic_etiq_edition » ?

20881 lignes et 26 colonnes. En sélectionnant la première ligne ou la première colonne, l'information s'affiche en bas à droite (Excel).

- **Filtrer et trier**

Nous souhaitons à présent trier les véhicules suivant leur taux d'émission de CO2 et leur type de carburant.

- Quel est le descripteur de la collection des voitures commercialisées en France indiquant le type de carburant ?

« cod_cbr » (fichier carlab-annuaire-variable).

On peut demander aux élèves s'ils connaissent une méthode pour trouver une information plus rapidement : Ctrl+F permet de rechercher un mot-clé dans un fichier / sur une page web.

- Quelle est la donnée permettant de connaître la marque d'une voiture ?

« lib_mrq_utac ».

Méthode : Filtrer des données

Dans le menu Données du tableur, des outils de filtres permettent de n'afficher que certaines lignes d'une feuille de calcul suivant certains critères. La fonction **« Filtrer »** *(également appelée « AutoFiltre » dans certains tableurs)* insère, au niveau d'une ou de plusieurs colonnes de données, une zone combinée permettant de sélectionner les enregistrements (lignes) à afficher. Cette fonction s'utilise de la manière suivante :

- **Sélectionnez** (= mettez en surbrillance) **les colonnes** auxquelles vous souhaitez appliquer le filtrage (ne rien sélectionner revient à sélectionner toutes les colonnes).
- **Activez la fonction « Filtrer » :** des boîtes combinées sont visibles dans la première ligne de la plage sélectionnée.
- Cliquez sur la flèche de déroulement située dans l'en-tête de la colonne et **choisissez un ou plusieurs éléments**.

Une fois la fonction de filtrage exécutée, seules les lignes dont le contenu correspond aux critères de filtre sont affichées. La colonne utilisée pour le filtre est identifiée par une icône différente.

Si vous appliquez un filtre supplémentaire sur une autre colonne de la plage de données filtrées, alors les autres zones combinées listent seulement les données filtrées.

Pour afficher à nouveau tous les enregistrements, sélectionnez « Sélectionner tout » dans toutes les colonnes ayant servi à filtrer les données. Pour cesser d'utiliser la fonction « Filtrer », sélectionnez toutes les cellules sélectionnées initialement et désactivez la fonction « Filtrer » *(ou bien cliquez sur l'icône « Filtrer » 2 fois jusqu'à ce que tous les filtres soient désactivés).*

- Combien de modèles hybrides sont enregistrés dans ce fichier ?

Il suffit de filtrer la catégorie « Hybride » (colonne I) et sélectionner « Oui », puis de sélectionner la colonne une fois le filtre effectué. Actuellement, il y a 1024 entrées.

- En tout, combien de modèles hybrides des marques Audi, BMW et Peugot sont enregistrés ?

On ajoute au filtre hybride un filtre de marques : en tout, il y a actuellement 31 entrées.

- Une fois les données filtrées, les lignes (colonne tout à gauche) ne se suivent plus. Pourquoi ?

Comme les lignes non concernées par les filtres sont cachées, seuls les n° des lignes concernées par les filtres apparaissent.

Méthode : Trier des données

Dans le menu « Données », choisissez « Trier », puis sélectionnez les clés de tri primaires et éventuellement secondaires en spécifiant si le tri est croissant ou décroissant.

- Maintenant, sélectionnez les voitures hybrides et triez-les de manière à trouver la marque du modèle qui émet le plus de CO^2 et celui qui en émet le moins (CO^2 mixte).

Actuellement, c'est le modèle Laferrari (Ferari) qui émet le plus de CO^2, et le Série 1 (BMW) qui en émet le moins.

- Enquêtons davantage sur quelques marques françaises … Parmi les marques Peugeot, Renault et Citroën, identifiez les 3 modèles de voitures roulant uniquement au gazole et émettant le moins de CO2, puis ceux qui en émettent le plus. On donnera la marque, le modèle, le code national d'identification du type et le taux de CO2.

Modèles en émettant le moins :

Marque	Modèle	CNIT	Taux de CO^2
CITROEN	C3	M10CTRVP029B259	79
CITROEN	DS3	M10CTRVP029A258	79
CITROEN	C3	M10CTRVP0163616	82

Modèles en émettant le plus :

Marque	Modèle	CNIT	Taux de CO^2
CITROEN	JUMPY	M10CTRVP008R356	199
CITROEN	JUMPY	M10CTRVP008U360	199
PEUGEOT	EXPERT	M10PGTVP0083949	199

- Et vos modèles préférés ? Recherchez vos modèles préférés et remplissez le tableau suivant :

Marque	Modèle	CNIT	Taux de CO^2	Conso Mixte

Les élèves peuvent notamment s'aider du site type-mine pour retrouver les CNIT, et les retrouver dans le fichier grâce à la barre « Recherche » du filtre cnit.

Étape 2 – Données ouvertes : découpages administratifs

- **Récupération des données**

Allez sur le site https://www.insee.fr/fr/accueil et suivez le chemin suivant : Accueil > Définitions, méthodes et qualités > Géographie administrative et d'étude > Téléchargement > Code officiel géographique.

- Cliquez sur "Téléchargement des fichiers du millésime 2023".
- Puis allez à la section "Régions" *(n'oubliez pas l'astuce ctrl+F)*.
- Téléchargez le fichier csv, puis décompressez le fichier ".zip" dans votre espace de travail.
- Récupérez de la même façon les fichiers csv du millésime 2023 sur les communes et les départements.
    - De quel institut proviennent ces trois fichiers ? Quel est son rôle ?

Ces trois fichiers proviennent du site de l'INSEE (Institut National de la Statistique et des Etudes Economiques). Cet institut est chargé de la production, de l'analyse et de la publication des statistiques officielles en France : comptabilité nationale annuelle et trimestrielle, évaluation de la démographie nationale, du taux de chômage, etc. Le site web de l'INSEE fournit quantité de données ouvertes que tout citoyen peut récupérer et traiter suivant ses besoins. En revanche, il est parfois nécessaire de récupérer les données de plusieurs fichiers pour obtenir l'information désirée.

Importation des données dans un tableur

Nous allons voir comment importer ici les trois fichiers csv dans un même classeur sans utiliser de copier-coller.

- Ouvrez les trois fichiers csv avec Notepad++ et prenez connaissance de l'encodage et du type de séparateur utilisés.

- Lancez un tableur et ouvrez un classeur vide. Sauvegardez ce classeur avec le nom "Activite RegDepCom".

Vous désirez importer à présent les trois fichiers csv dans trois feuilles de calcul séparées, et renomer.

  - Recherchez sur internet comment on peut importer dans un tableur des données externes issues d'un fichier.

- **Tableur type Excel :** onglet Données -> Fichier texte.
  - **Étape 1 :** type délimité; Unicode utf-8.
  - **Étape 2 :** virgule.
  - **Étape 3 :** texte pour toutes les colonnes (sinon perte des 0 dans la première colonne).

- **Tableur type OpenOffice Calc :** cliquez à gauche des feuilles existantes en bas -> sélectionner "à partir d'un fichier". Unicode utf-8; séparateur : virgule; colonne en texte.

  - Importez les trois fichiers **puis renommez chacune des feuilles "Régions", "Départements" et "Communes"** suivant le type de données qu'elles contiennent.

  - Sauvegardez l'ensemble du classeur.

Croiser des données

On désire ajouter pour chaque commune le nom de la région à laquelle elle appartient. Pour cela, procédez de la façon suivante :

- Dans la feuille "Communes" du classeur "Activite RegDepCom", saisissez le mot "nreg" (pour Nom Région) dans la cellule L1.
- Dans la cellule L2 de cette même feuille, saisissez l'une des deux instructions suivantes :
  - **Tableur type Excel :** =RECHERCHEV(C2;Régions!$A$2:$D$19;4;FAUX)
  - **Tableur type OpenOffice Calc :** =RECHERCHEV(C2;$Régions:$A$2:$D$19;4;0)

  - La formule a-t-elle fonctionné ? Si ce n'est pas le cas, quelles peuvent être les erreurs commises ?

Pistes :
- Il n'y a pas de donnée dans la colonne reg (A) de la feuille « Régions » ou dans la colonne reg (C) de la feuille « Communes ».
- Les données ne correspondent pas.
- Le nom de la feuille « Régions » a été mal reporté dans la formule

- Reprenons les formules =RECHERCHEV(C2;Régions!$A$2:$D$19;4;FAUX) ou =RECHERCHEV(C2;$Régions:$A$2:$D$19;4;0), et traduisez en une ou plusieurs phrases ce qu'elles demandent au logiciel *(vous pouvez faire des recherches, notamment utiliser la fonction « Aide »)*.

L'instruction demande de chercher la valeur contenue dans la cellule C2 de la feuille courante (premier paramètre) dans la première colonne du tableau de la feuille "Régions" constituée des cellules allant de A2 (en haut à gauche) à D19 (en bas à droite) (deuxième paramètre), puis de renvoyer la valeur située sur la même ligne que la valeur trouvée et dans la colonne numéro 4 (troisième paramètre). Le dernier paramètre est un booléen (VRAI ou 1 / FAUX ou 0) qui indique si la première colonne est triée ou non. Si vous indiquez FAUX ou 0, il cherche la valeur exacte dans tous les cas. Les symboles $ servent à bloquer les cellules lors de la recopie de la formule dans les cellules suivantes.

- Pourquoi a-t-il été possible d'automatiser le remplissage de cette colonne ?

Les deux collections "Communes" et "Régions" ont la colonne correspondant au descripteur "reg" en commun, ce qui permet de faire une jointure entre les deux feuilles de calculs.

La colonne "typecom" indique le type de commune. Ses valeurs, fournies par l'INSEE, sont les suivantes :

- COM : commune
- COMA : commune associée
- COMD : commune déléguée
- ARM : arrondissement municipal

- Affichez sur la feuille "Communes" tous les arrondissements municipaux. Combien y en a-t-il et à quoi correspondent-ils ?

On effectue un filtre "ARM" sur "typecom". Il y en a 57, et ils correspondent aux différents arrondissements des villes de Marseille, Lyon et Paris.

La colonne "com" de la feuille "Communes" indique le code de chaque commune, et la colonne "cheflieu" de la feuille "Départements" indique le code de la commune chef-lieu du département.

- Affichez sur la feuille "Départements" le nom du chef-lieu et le nom de la région de chaque département.
- Écrivez "ncheflieu" dans H1, puis "=RECHERCHEV(C2;Communes!$B$2:$I$37933;8;FAUX)" dans H2 et étirez la formule. Écrivez "nreg" dans I1, puis "=RECHERCHEV(B2;Régions!$A$2:$F$19;6;FAUX)" dans I2 et étirez la formule.

- Quels sont les chefs-lieux des départements des Yvelines et de la Vendée ?

Yvelines : Versailles, Vendée : La Roche-sur-Yon.

- Quels sont les chefs-lieux de la région Île-de-France et de la région Auvergne-Rhône-Alpes ?

On filtre les données suivant les régions voulues et on obtient : Île-de-France : Paris, Melun, Versailles, Evry-Courcouronnes, Nanterre, Bobigny, Créteil, Pontoise. Auvergne-Rhône-Alpes : Bourg-en-Bresse, Moulins, Privas, Aurillac, Valence, Grenoble, Saint-Étienne, Le Puy-en-Velay, Clermont-Ferrand, Lyon, Chambéry, Annecy.

Traitement des données ouvertes

Fiche activité

Étape 1 – Données ouvertes : véhicules et émission de CO^2

- **Importation des données ouvertes**
  - Allez sur le site data.gouv.fr, puis saisissez les mots-clés « véhicule CO2 » dans le champ de recherche.
  - Téléchargez le jeu de données sur les émissions de CO2 et de polluants des véhicules commercialisés en France en 2014. Quel est ce format ? Quel type de séparateur est utilisé ? (*Ouvrez-le avec un Notepad++ par exemple).*

⊞ **[2014] Emissions de polluants, CO2 et caractéristiques des véhicules commercialisés en France**

Mis à jour le 7 juillet 2014 — csv — 2339 téléchargements

.....

.....

.....

.....

.....

.....

  - Téléchargez également le dictionnaire des variables. Quel est son format ?

⊞ **Dictionnaire des variables**

Mis à jour le 4 juillet 2014 — xls — 1380 téléchargements

.....

.....

.....

.....

.....

.....

- Décompressez le fichier ZIP dans votre répertoire de travail (clic droit -> Extraire...), puis ouvrez dans un tableur les fichiers « carlab-annuaire-variable » et « fic_etiq_edition ».

Remarque : En général, l'importation de ce fichier se passe bien car il est reconnu par le tableur. Sinon, une boîte de dialogue s'ouvre et il faut procéder à certains réglages.

- Pourquoi l'un des fichiers est-il « ouvert » et l'autre « importé » ?

.....

.....

.....

- Combien de lignes et de colonnes remplies y a-t-il dans le fichier « fic_etiq_edition » ?

.....

.....

.....

Filtrer et trier

Nous souhaitons à présent trier les véhicules suivant leur taux d'émission de CO2 et leur type de carburant.

- Quel est le descripteur de la collection des voitures commercialisées en France indiquant le type de carburant ?

.....

.....

.....

.....

- Quelle est la donnée permettant de connaître la marque d'une voiture ?

.....

Méthode : Filtrer des données

Dans le menu Données du tableur, des outils de filtres permettent de n'afficher que certaines lignes d'une feuille de calcul suivant certains critères. La fonction **« Filtrer »** *(également appelée « AutoFiltre » dans certains tableurs)* insère, au niveau d'une ou de plusieurs colonnes de données, une zone combinée permettant de sélectionner les enregistrements (lignes) à afficher. Cette fonction s'utilise de la manière suivante :

- **Sélectionnez** (= mettez en surbrillance) **les colonnes** auxquelles vous souhaitez appliquer le filtrage (ne rien sélectionner revient à sélectionner toutes les colonnes).
- **Activez la fonction « Filtrer » :** des boîtes combinées sont visibles dans la première ligne de la plage sélectionnée.
- Cliquez sur la flèche de déroulement située dans l'en-tête de la colonne et **choisissez un ou plusieurs éléments**.

Une fois la fonction de filtrage exécutée, seules les lignes dont le contenu correspond aux critères de filtre sont affichées. La colonne utilisée pour le filtre est identifiée par une icône différente.

Si vous appliquez un filtre supplémentaire sur une autre colonne de la plage de données filtrées, alors les autres zones combinées listent seulement les données filtrées.

Pour afficher à nouveau tous les enregistrements, sélectionnez « Sélectionner tout » dans toutes les colonnes ayant servi à filtrer les données. Pour cesser d'utiliser la fonction « Filtrer », sélectionnez toutes les cellules sélectionnées initialement et désactivez la fonction « Filtrer » *(ou bien cliquez sur l'icône « Filtrer » 2 fois jusqu'à ce que tous les filtres soient désactivés).*

- Combien de modèles hybrides sont enregistrés dans ce fichier ?

……

……

……

- En tout, combien de modèles hybrides des marques Audi, BMW et Peugot sont enregistrés ?

……

……

- Une fois les données filtrées, les lignes (colonne tout à gauche) ne se suivent plus. Pourquoi ?

……

……

……

Méthode : Trier des données

Dans le menu « Données », choisissez « Trier », puis sélectionnez les clés de tri primaires et éventuellement secondaires en spécifiant si le tri est croissant ou décroissant.

- Maintenant, sélectionnez les voitures hybrides et triez-les de manière à trouver la marque du modèle qui émet le plus de CO^2 et celui qui en émet le moins (CO^2 mixte).

.....

.....

- Enquêtons davantage sur quelques marques françaises … Parmi les marques Peugeot, Renault et Citroën, identifiez les 3 modèles de voitures roulant uniquement au gazole et émettant le moins de CO2, puis ceux qui en émettent le plus. On donnera la marque, le modèle, le code national d'identification du type et le taux de CO2.

Modèles en émettant le moins :

Marque	Modèle	CNIT	Taux de CO^2

Modèles en émettant le plus :

Marque	Modèle	CNIT	Taux de CO^2

- Et vos modèles préférés ? Recherchez vos modèles préférés et remplissez le tableau suivant :

Marque	Modèle	CNIT	Taux de CO^2	Conso Mixte

Étape 2 – Données ouvertes : découpages administratifs

- **Récupération des données**

Allez sur le site https://www.insee.fr/fr/accueil et suivez le chemin suivant : Accueil > Définitions, méthodes et qualités > Géographie administrative et d'étude > Téléchargement > Code officiel géographique.

- Cliquez sur "Téléchargement des fichiers du millésime 2023".
- Puis allez à la section "Régions" *(n'oubliez pas l'astuce ctrl+F)*.
- Téléchargez le fichier csv, puis décompressez le fichier ".zip" dans votre espace de travail.
- Récupérez de la même façon les fichiers csv du millésime 2023 sur les communes et les départements.
    - De quel institut proviennent ces trois fichiers ? Quel est son rôle ?

.....

.....

.....

.....

.....

.....

- **Importation des données dans un tableur**

Nous allons voir comment importer ici les trois fichiers csv dans un même classeur sans utiliser de copier-coller.

- Ouvrez les trois fichiers csv avec Notepad++ et prenez connaissance de l'encodage et du type de séparateur utilisés.

- Lancez un tableur et ouvrez un classeur vide. Sauvegardez ce classeur avec le nom "Activite RegDepCom".

Vous désirez importer à présent les trois fichiers csv dans trois feuilles de calcul séparées, et renomer.

- Recherchez sur internet comment on peut importer dans un tableur des données externes issues d'un fichier.

Méthode (pour rappel) :

.....

.....

……

……

……

……

……

……

……

……

……

……

- Importez les trois fichiers **puis renommez chacune des feuilles "Régions", "Départements" et "Communes"** suivant le type de données qu'elles contiennent.
- Sauvegardez l'ensemble du classeur.

Croiser des données

On désire ajouter pour chaque commune le nom de la région à laquelle elle appartient. Pour cela, procédez de la façon suivante :

- Dans la feuille "Communes" du classeur "Activite RegDepCom", saisissez le mot "nreg" (pour Nom Région) dans la cellule L1.
- Dans la cellule L2 de cette même feuille, saisissez l'une des deux instructions suivantes :
  - **Tableur type Excel :** =RECHERCHEV(C2;Régions!$A$2:$D$19;4;FAUX)
  - **Tableur type OpenOffice Calc :** =RECHERCHEV(C2;$Régions:$A$2:$D$19;4;0)
- La formule a-t-elle fonctionné ? Si ce n'est pas le cas, quelles peuvent être les erreurs commises ?

……

……

……

……

- Reprenons les formules **=RECHERCHEV(C2;Régions!$A$2:$D$19;4;FAUX)** ou **=RECHERCHEV(C2;$Régions:$A$2:$D$19;4;0)**, et traduisez en une ou plusieurs phrases ce qu'elles demandent au logiciel *(vous pouvez faire des recherches, notamment utiliser la fonction « Aide »)*.

.....

.....

.....

.....

.....

.....

- Pourquoi a-t-il été possible d'automatiser le remplissage de cette colonne ?

.....

.....

.....

.....

.....

.....

La colonne "typecom" indique le type de commune. Ses valeurs, fournies par l'INSEE, sont les suivantes :

- COM : commune
- COMA : commune associée
- COMD : commune déléguée
- ARM : arrondissement municipal

- Affichez sur la feuille "Communes" tous les arrondissements municipaux. Combien y en a-t-il et à quoi correspondent-ils ?

.....

.....

.....

La colonne "com" de la feuille "Communes" indique le code de chaque commune, et la colonne "cheflieu" de la feuille "Départements" indique le code de la commune chef-lieu du département.

- Affichez sur la feuille "Départements" le nom du chef-lieu et le nom de la région de chaque département.
- Écrivez "ncheflieu" dans H1, puis "=RECHERCHEV(C2;Communes!$B$2:$I$37933;8;FAUX)" dans H2 et étirez la formule. Écrivez "nreg" dans I1, puis "=RECHERCHEV(B2;Régions!$A$2:$F$19;6;FAUX)" dans I2 et étirez la formule.

    - Quels sont les chefs-lieux des départements des Yvelines et de la Vendée ?

.....

.....

    - Quels sont les chefs-lieux de la région Île-de-France et de la région Auvergne-Rhône-Alpes ?

.....

.....

.....

.....

.....

.....

.....

.....