Mini-projet - Elections

A lire attentivement!

Ce travail est noté. Toute aide externe (telle que poser des questions sur des forums ou utiliser des assitants génératifs de type Copilot) est interdite.

Vous développerez vous-même votre code ! La copie d'une (petite) portion de code préexistant est tolérable si :

- Elle reste occasionnelle et largement minoritaire ;
- Elle est clairement signalée par un commentaire adéquat. Le non-respect de ces règles sera considéré comme de la tricherie et pourra occasionner des sanctions.

Organisation

- Le projet est réalisé par groupe de deux personnes
- Les semaines de cours du 21 et 28 décembre, ainsi que du 11 et 18 janvier seront consacrées à ce projet
- **Le projet est à rendre pour le dimanche 22 janvier 2023 à 23:59 sur Cyberlearn (pour chaque jour de retard, 1 point de note sera déduit)**
- **Les présentations sont obligatoires et auront lieu pendant le cours du 25 janvier**
- La présentation doit être courte et concise. Il faut expliquer les choix spécifiques pour votre pipeline de prétraitement (avec une représentation visuelle du pipeline), en particulier les étapes de nettoyage, et leur impact sur l'analyse, et les résultats d'analyse. La présentation doit durer au maximum 10 minutes, questions comprises. Chaque membre du groupe doit présenter avec une répartition uniforme.
- À rendre : `Nom1_Nom2.ipynb`
- Ajouter un fichier readme avec quelques commentaires sur ce qui fonctionne, ce qui ne fonctionne pas, et les éventuels autres points notables de votre implémentation.

Contexte

Nous allons explorer, prétraiter, et analyser des données liées aux élections présidentielles 2020 aux Etats-Unis. **Notre objectif final d'analyse sera de grouper les citoyens en fonction de leurs préférences électorales**. La démographie aux Etats-Unis est composée de 3006 comtés (counties), et la densité de chacun est représenté par couleur dans la carte ci-dessous.

Exploration des données

Les sources de données sont les suivantes (disponibles dans le dossier `data/raw`):

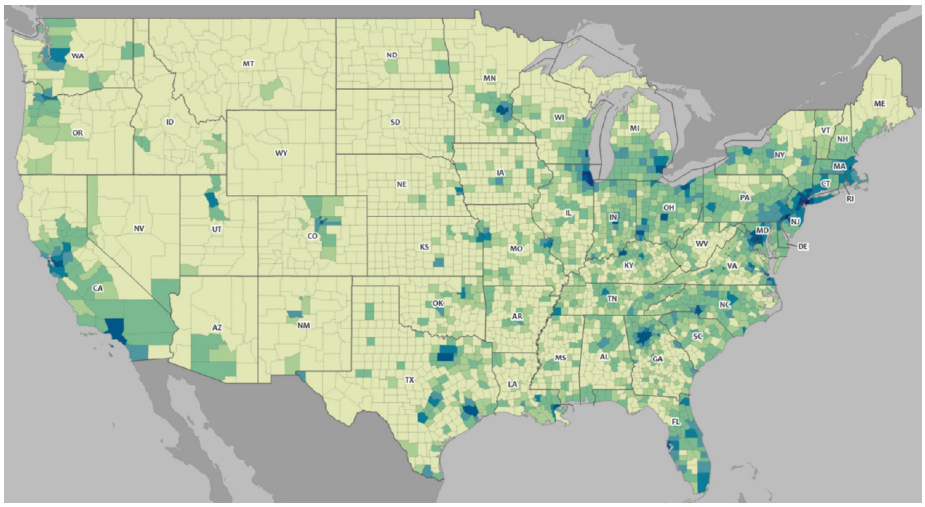

Figure 1: La démographie des Etats-Unis au niveaux des comtés

- `Education.xls` ==> à charger dans `edu_df`
- `PopulationEstimates.xls` ==> à charger dans `pop_df`
- `PovertyEstimates.xls` ==> à charger dans `pov_df`
- `Unemployment.xlsx` ==> à charger dans `employ_df`
- `countypres_2000-2020.csv` ==> à charger dans `election_df`

Ces fichiers proviennent du département américain de l'USDA (US Department of Agriculture Economic Research) et du US election results from MIT. L'objectif de l'exploration est de les importer (dans les noms de variables indiqués) et de les décrire pour bien les comprendre (graphiques, distributions, liens, valeurs manquantes, corrélations, …).

Prétraitement

Dans la partie de prétraitement, nous allons intégrer et nettoyer ces données avec les techniques vues au cours. L'objectif de prétraitement sera de les préprarer pour réaliser du clustering par comtés. Le résultat de cette étape sera donc d'avoir des données propres avec la structure suivante dans `county_df`:

- chaque ligne représente un comté
- les colonnes décrivent les comtés avec les attributs pertinents (les variables sélectionnées et les éventuelles features extraites selon vos choix d'analyse!)

Le résultat final du prétraitement devra être stocké dans `data/preprocessing/county_df` pour pouvoir être repris lors de l'analyse.

Analyse

Dans cette partie, nous allons effectuer un clustering basé sur le résultat du prétraitement (`county_df`). A vous de décider comment vous souhaitez réaliser le clustering!

Évaluation

Le projet donnera lieu à une note. Il sera évalué selon les critères suivants :

- Pertinence, maîtrise, clarté, et reproductibilité du pipeline de prétraitement (niveaux de nettoyage, pre-processing, etc.) (60%)
- Qualité générale du code Python (20%)
- Présentation et justifications des choix (20%)

De plus, il est judicieux de faire particulièrement attention au nettoyage de votre code avant de le rendre :

- Même si le code est sous forme de Notebook, utilisez des fonctions et structurez votre code avec des cellules de markdown
- Remettez votre code avec les outputs des cellules. De plus, le notebook doit pouvoir s'exécuter entièrerment sans erreur comme un script python!
- Pas de code inutile ! (Retirer les import et fonctions qui ne servent plus à rien, . . .)
- Évitez les blocs de 50 lignes de code mis en commentaire "pour l'instant" !
- Est-il encore besoin de le préciser : Commentez votre code !