Consommation d'électricité annuelle par secteur d'activité et par commune en France
Projet Data Mining

27/01/2023

ARMAND L'HUILLIER, CAMIL ZAHI

SUPERVISE PAR AMED COULIBALY

1. Description des données

Les bases de données

Séries historiques de population (1876 à 2019)

- 34 967 observations
- 37 variables
- 4 variables géographiques

La capacité des communes en hébergement touristique

- 34 983 observations
- 52 variables
- 1 variable géographique

Consommation annuelle d'électricité et de gaz par commune et par secteur d'activité

- 46 584 observations
- 36 variables
- 7 variables géographiques

Remarque : Les variables géographiques ont servies de clé de jointure entre ces trois tables.

2. Analyse exploratoire des données

Construction de la base d'analyse

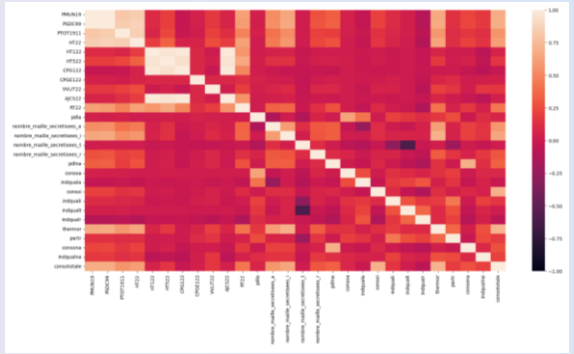
- Fusion en « innerjoin » entre la base consommation et la base population : on veut avoir sur l'ensemble des villes, les données de population et de consommation.
- Merge en « leftjoin » entre la nouvelle base et la base tourisme car on considère le tourisme comme bonus pour prédire la consommation d'électricité.
- Filtre sur la consommation d'électricité et non de gaz : 22 227 villes

Imputation des valeurs manquantes et valeurs extrêmes

- Valeurs manquantes :
  - Dans le coin inférieur gauche : La fusion a crée des valeurs manquantes : villes sans données touristiques : imputation par la régression.
  - Dans le centre droit vers le haut et tout en bas : peu de valeurs manquantes : quelques codes postaux manquants et quelques données de population manquantes : imputation par la médiane (populations) et par le code commune (code postal).
- Valeurs extrêmes : pas de traitement car on considère cela comme une perte d'information pour les villes les plus peuplées car moins nombreuses.

Corrélation des variables

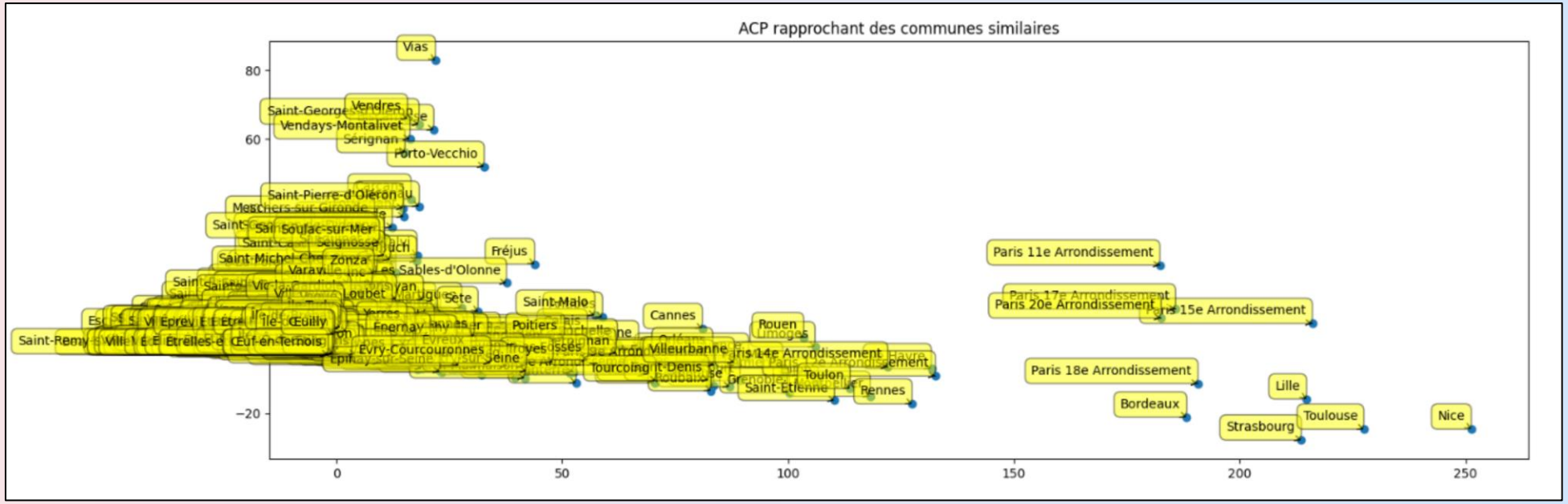
- Les variables semblent trop corrélées entre elles notamment celles de population où nous décidons de n'en garder que quelques unes. Pour le moment nous gardons les autres variables.



Traitement des doublons

- Suppression des doublons : observation faite sur des années différentes.
- 20 732 observations

3. Analyse en Composantes Principales



Interprétation :

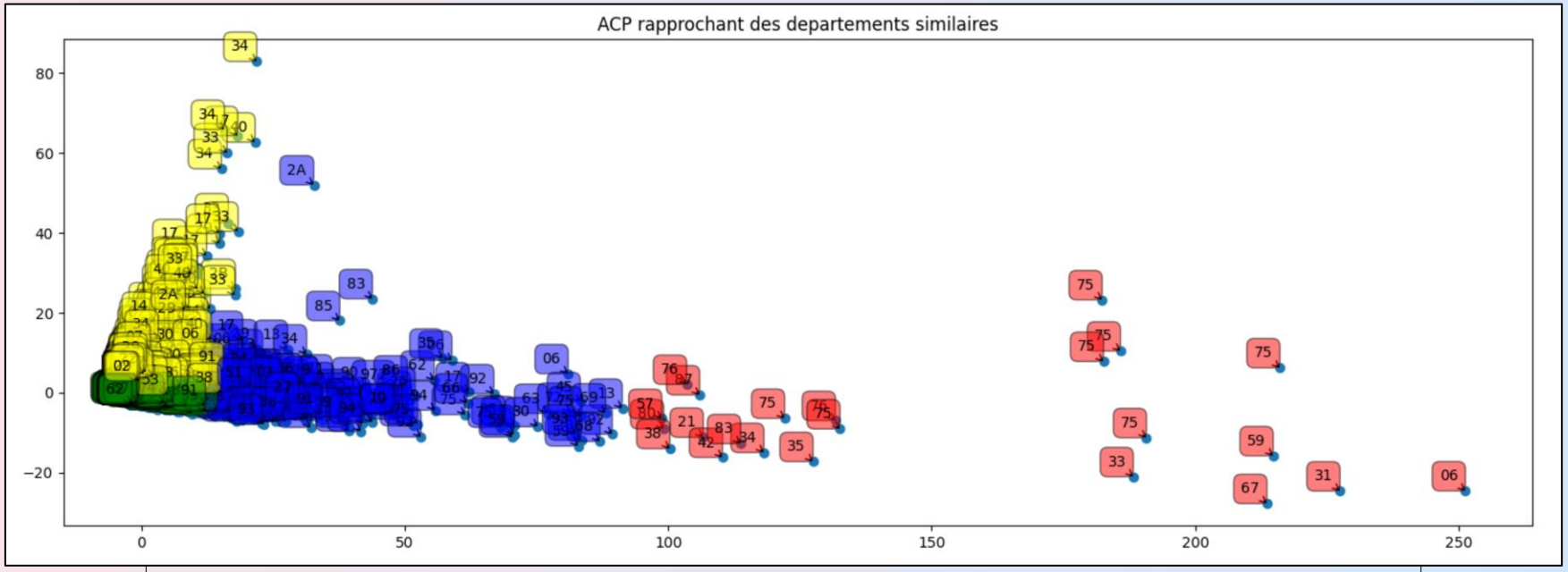
Les deux premiers axes représentent à peu près 55% de la variance expliquée.

- La composante principale 1 indique la grandeur en population de la ville (Est-Ouest).
- La composante principale 2 indique la capacité en tourisme de la ville relativement à sa population (Nord-Sud).

Plus la consommation est élevée, plus les communes se trouvent à droite dans le graphique.

PS : on a une grande masse de communes qui sont proches de 0 dans le graphique. C'est parce que on a une **grande quantité de communes qui ont une population faible et qui sont peu touristiques**.

4. Clustering



- Communes avec une population et capacité touristique très forte ;
- Communes avec une population et capacité touristique moyenne ;
- Communes avec une population faible mais avec une bonne capacité touristique ;
- Communes avec une population et capacité touristique faible.

Interprétation :

L'axe Ouest - Est s'explique par le niveau de la population de la commune : avec d'un côté les communes les plus peuplées et de l'autre les communes les moins peuplées.

L'axe Nord - Sud s'explique par le niveau de tourisme de la commune : avec d'un côté les communes les plus touristiques et de l'autre les communes les moins touristiques.

5. Machine Learning

Prédire la consommation électrique des communes par secteur agrégée

Random Forest

Dans l'agriculture	Dans l'industrie	Dans le résidentiel	Dans le tertiaire
Modèle peu performant R^2 Train = 0,54 et R^2 Test = 0,43	Modèle peu performant R^2 Train = 0,72 et R^2 Test = 0,21	Modèle très performant R^2 Train = 0,97 et R^2 Test = 0,93	Modèle très performant R^2 Train = 0,92 et R^2 Test = 0,87
Plus la population a augmenté rapidement et moins la commune consomme de l'électricité pour ce secteur	Pas d'interprétation fiable pour ce secteur qui n'a pas su donner de bon score	Plus il y a de points de branchements à l'électricité et plus la consommation d'électricité est élevée	En moyenne, plus la commune est peuplée, et plus son économie a de chance d'être tournée vers les services
Faible score car hétérogénéité du secteur	Non interprétable	Ce phénomène semble linéaire	Très bon score pour prédire la consommation d'électricité
✗	✗	✓	✓

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE
ÉCOLE D'ÉCONOMIE
DE LA SORBONNE

MoSEF
Data Science

Merci pour votre attention

27/01/2023

ARMAND L'HUILLIER, CAMIL ZAHI

SUPERVISE PAR AMED COULIBALY