Projet Datamining

Parc électrique automobile français

Thomas Barat | Mathéo Richer | Amaury Coutant

Table des matières

Introduction

Dans le cadre de l'Open Data University Challenge, nous avons été conviés à concevoir un tableau de bord sur le sujet de notre choix, en utilisant un jeu de données disponibles sur le site.
Notre groupe a choisi les données liées aux véhicules électriques. Suite au tableau de bord, une analyse du même jeu de données était demandée.
Cette analyse devait inclure une méthode d'analyse factorielle (ACP, AFC ou ACM), ainsi que du clustering afin d'identifier et d'interpréter des groupes d'individus.

Données

La première étape de notre travail a consisté à charger les données et à effectuer un nettoyage préliminaire afin de garantir leur qualité et leur cohérence.
Les jeux de données utilisés incluent:

- Un jeu de données sur les véhicules immatriculés en France.
- Une base de données sur la population française.
- Un jeu de données fournissant de nombreuses informations sur les bornes de recharge en France.
- Un dernier jeu contenant les noms des départements et des régions.

Ces différentes sources de données ont permis d'établir une analyse complète et croisée pour répondre aux objectifs fixés.

Une fois les données téléchargées, il a été nécessaire de les nettoyer afin de structurer une base de données par département intégrant plusieurs caractéristiques jugées pertinentes pour notre analyse.

Vehicules

Nous avons commencé par les données des véhicules. Initialement, ces données se présentaient de la manière suivante:

codgeo	libgeo	epci	libepci	date_arrete	nb_vp_rechargeables_el	nb_vp_rechargeables_gaz	nb_vp
42298	SAUVAIN	200065886	CA Loire Forez Agglomération	2024-03-31	0	0	445
42299	SAVIGNEUX	200065886	CA Loire Forez Agglomération	2020-12-31	28	0	3708

Le principal problème résidait dans la multiplicité des dates pour une même ville, ainsi que dans la présence de certaines données jugées non pertinentes, comme le code epci.
Pour y remédier, nous avons décidé de conserver uniquement la date la plus récente pour chaque ville. Ensuite, les données des différentes villes d'un même département ont été additionnées. Enfin, seules les variables nous intéressants ont été gardées pour poursuivre l'analyse.
Le résultat final se présente ainsi:

Département	nb_vp_rechargeable_el	nb_vp_rechargeables_gaz	nb_vp
02	12682	16	600407
03	5788	11	361228

Population

Le jeu de données suivant à nettoyer était celui lié la population, qui se présentait initialement sous la forme suivante:

AGE	CIVIL_STATUS	EMPSTA_ENQ	GEO	GEO_OBJECT	HAR	RP_MEASURE	SEX	TIME_PERIOD	OBS_VALUE
Y_GE15	_T	33	44107	UU2020	_T	POP	_T	2021	241,43284
Y_GE15	_T	33	41202	UU2020	_T	POP	_T	2021	272,34522
Y_GE15	_T	33	38102	UU2020	_T	POP	_T	2021	91,31346

Nous avons choisi d'appliquer les mêmes manipulations que pour le jeu de données précédent.
Nous avons conservé uniquement le nombre d'habitants par département ainsi que l'âge prédominant dans chaque département.
Le numéro de département a également été conservé, mais uniquement dans le but de concaténer les données ultérieurement.
Les données finales se présentent ainsi :

Dép	OBS_VALUE_REG	SEX	AGE
01	663202	F	Y55T64
02	527468	F	Y55T64
03	334872	F	Y55T64

Borne

Le jeu de données suivant à être nettoyé était celui concernant les bornes , présenté à l'origine sous la forme suivante:

nom_amenageur	siren_amenageur	contact_amenageur	nom_operateur	contact_operateur	telephone_operateur	nom_enseigne	id_station_itinerance	id_station_local	nom_station	im
Grupo Easycharger	NA	NA	Zunder	ES*ZUN	roaming@zunder.com	NA	Zunder	ESZUNP1249276401566619487	1130592	Zunder/143954	Sta rec

Ce jeu de données comportant un grand nombre de variables jugées inutiles pour notre analyse, celles-ci ont été supprimées.
Ensuite, nous avons recensé, pour chaque département, le nombre de bornes ainsi que certaines variables et leurs modalités prédominantes.
Par exemples, pour la variable de l'opérateur principal, nous avons conservé uniquement celui ayant le plus grand nombre de bornes implantées dans le département. Les données se présentent désormais ainsi:

NombreBorne	nom_operateur	implantation_station	puissance_nominale
738	Freshmile Services	FR*FR1	Voirie	22

Départements & Régions

Le dernier jeu de données modifié concernait les noms des départements et leurs régions.
Nous avons rencontré un problème avec cette base de données : elle incluait les multiples départements d'outre-mer, et la Corse y était divisée en deux départements distincts. Cette structuration était différente de celle de certaines bases de données utilisées précédemment:

num_dep	dep_name	region_name
2A	Corse-du-Sud	Corse
2B	Haute-Corse	Corse
971	Guadeloupe	Guadeloupe
972	Martinique	Martinique
973	Guyane	Guyane
974	La Réunion	La Réunion
976	Mayotte	Mayotte

Pour résoudre ce problème, nous avons choisi de regrouper tous les départements d'outre-mer en une seule entité et d'« unifier » la Corse en la considérant comme un unique département.

num_dep	dep_name	region_name
20	Corse	Corse
97	Outre-Mer	Outre-Mer

Concatenation

Pour terminer, nous avons concaténé toutes les tables de données en une seule. Cette table regroupe l'ensemble des variables que nous avons jugées utiles et pertinentes pour l'étude, couvrant les 96 départements français (94 départements métropolitains, la Corse "unifiée" et les départements d'outre-mer regroupés).
La table finale se présente ainsi :

	Région	Numéro département	Electrique	Gaz	Essence	Population	Genre prédominant	Tranche d'age prédominant	Nombre de borne	nom_operateur	implantation_station	Puissance prédominante
Ain	Auvergne-Rhône-Alpes	01	19746	48	733757	663202	F	Y55T64	738	Freshmile Services	FR*FR1	Voirie	22
Aisne	Hauts-de-France	02	12682	16	600407	527468	F	Y55T64	1129	L'Union des Secteurs d'Énergie du Département de l'Aisne (USEDA)	FR*S02	Parking privé à usage public	22
Allier	Auvergne-Rhône-Alpes	03	5788	11	361228	334872	F	Y55T64	776	SPBR1	Voirie	22
Alpes-de-Haute-Provence	Provence-Alpes-Côte d'Azur	04	3964	5	203365	166077	F	Y55T64	492	SPBR1	Voirie	22
Hautes-Alpes	Provence-Alpes-Côte d'Azur	05	3013	6	164875	140976	F	Y55T64	589	SPBR1	Voirie	22

ACP

Nous avons choisi de réaliser deux types d'analyses, en commençant par une ACP.

Initialement, l'ACP a été effectuée directement sur les données brutes. Cependant, cela a révélé un problème de proportionnalité entre les variables, ce qui réduisait considérablement la pertinence de l'analyse.

Pour résoudre ce problème, nous avons transformé les variables en proportions par habitant. Cette normalisation a permis de restaurer la pertinence de l'analyse.

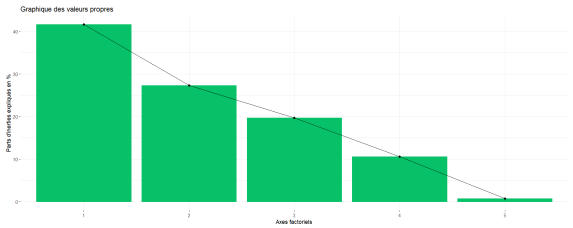
Nous avons procédé en deux étapes :

1.Une première ACP par département.
2.Une seconde ACP par région, afin de déterminer si les départements suivent les mêmes tendances à une échelle agrégée.

ACP par Départements

Ainsi comme dis precédement, une fois les variables changer par leur proportionalité par habitant , nous procédons à l'analyse à composantes principales.

Inertie

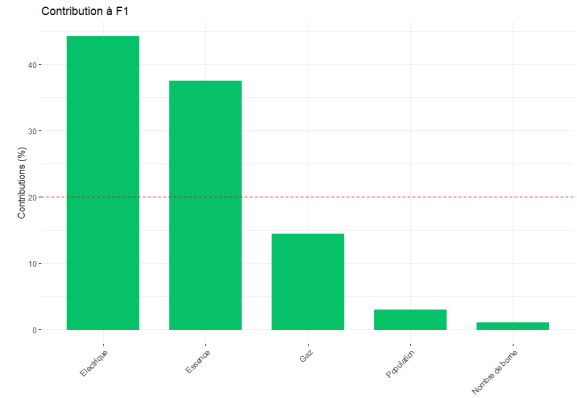


	F1	F2	F3	F4	F5
Valeur Propre	2.084	1.367	0.985	0.527	0.036
Parts d'inertie	41.690	27.331	19.701	10.549	0.729
Parts cumulées	41.690	69.021	88.722	99.271	100.000

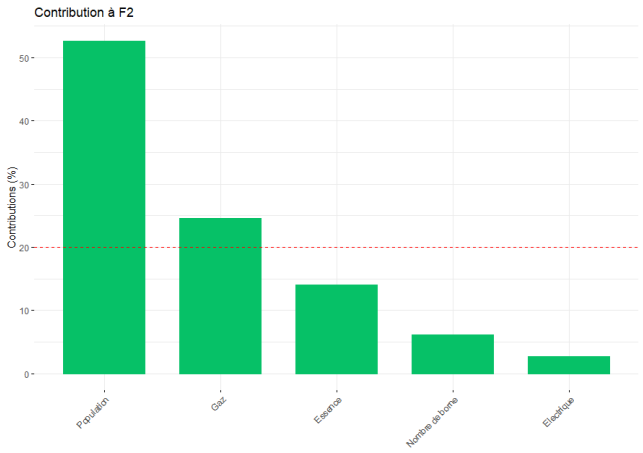
Sur le graphique des valeurs propres, on remarque que les deux premiers axes se distinguent clairement en capturant la majorité des informations importantes. Ensemble, ils expliquent près de 70 % de la dispersion totale (41,69 %pour le premier axe et 27,33 % pour le second).

Cela montre que ces deux axes résument une grande partie des variations dans les données. On a donc décidé de concentrer notre analyse dessus pour simplifier et mieux comprendre les tendances principales.

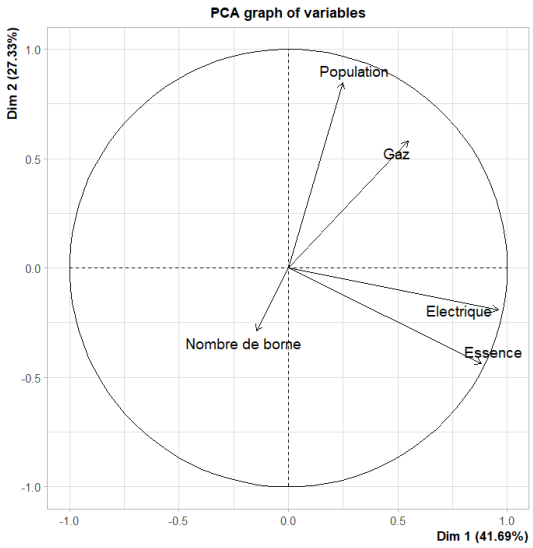
Contribution



Les deux variables les plus influentes sur l'axe 1 sont "Électrique", qui représente environ 44 %, et "Essence", à hauteur de 38 %.
En d'autres termes, l'axe 1 regroupe les départements où on trouve un fort taux de voitures, qu'elles soient électriques ou à essence, par habitant.
Cela nous permet de repérer les zones où la densité automobile est particulièrement élevée.



L'axe 2, quant à lui, est principalement influencé par la population, qui pèse pour plus de 52 %.
Il va donc opposer les départements en fonction de leur nombre d'habitants.
On peut aussi noter que la variable "Gaz" contribue également à l'axe 2, à hauteur de 25 %, ce qui suggère qu'il y a un lien entre la présence de véhicules au gaz et la taille des départements.



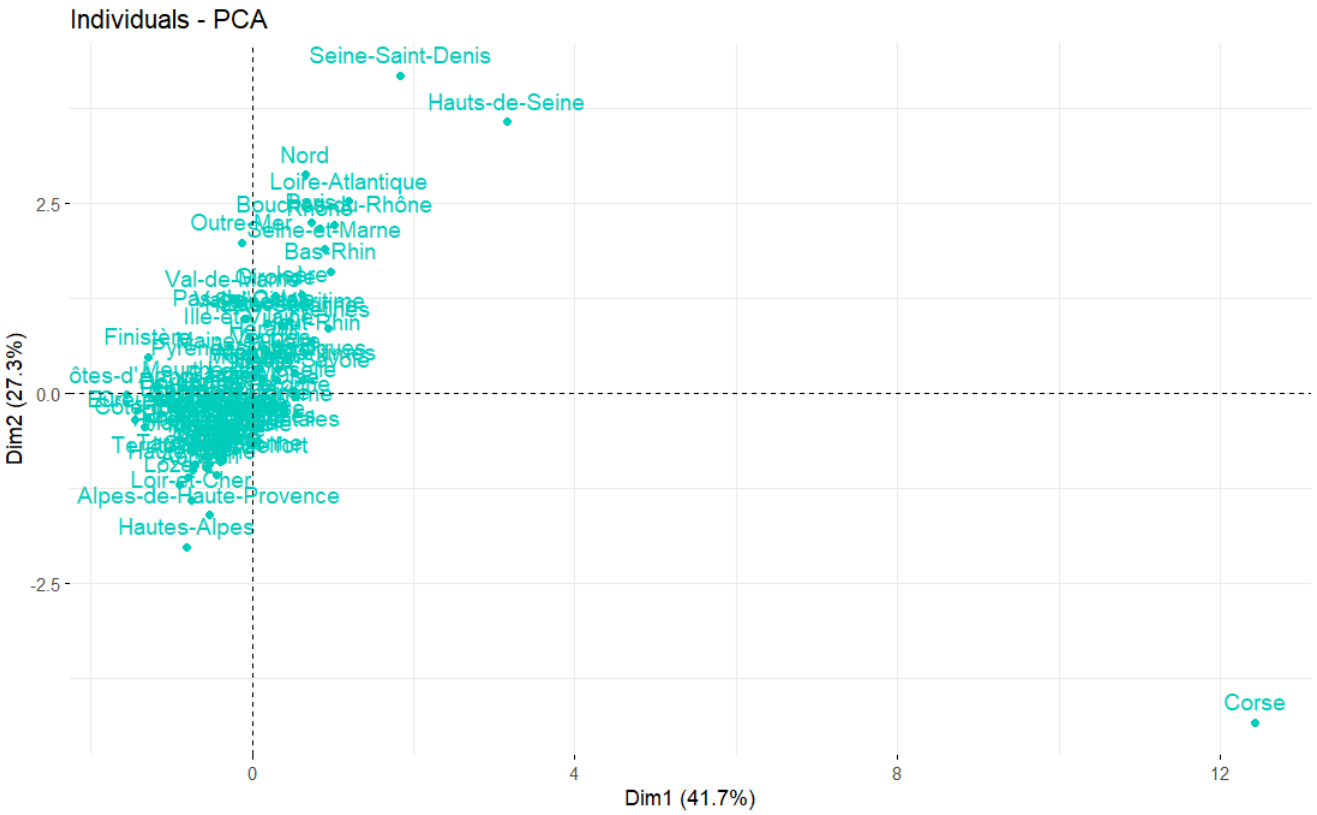
On voit clairement ici que les variables "Électrique" et "Essence" contribuent fortement et positivement à l'axe 1, tandis que la variable "Population" se démarque nettement sur l'axe 2, également de manière positive. On remarque aussi que la variable "Gaz" influence à la fois l'axe 1 et l'axe 2 de manière assez équilibrée.
Quant au nombre de bornes, il reste plutôt centré par rapport aux autres variables. Cela pourrait indiquer que les bornes de recharge sont réparties de façon relativement uniforme entre les départements, ou qu'elles ne sont pas directement corrélées ni à la population ni aux types de motorisation.

Nuage des individus

En regardant de plus près le nuage des individus, on observe une répartition diagonale. En bas à gauche, on trouve les départements les moins peuplés, avec un faible taux de voitures par habitant. À l'opposé, en haut à droite, se situent les départements fortement peuplés, avec un taux de voitures par habitant élevé.

Certains départements se démarquent particulièrement, comme la Seine-Saint-Denis et les Hauts-de-Seine. Ces départements, fortement urbanisés et densément peuplés, concentrent une grande quantité de véhicules.

Un point intéressant à regarder se situe en bas à droite, où la Corse apparaît isolée. Ce département atypique se distingue par une population relativement faible, mais un taux de voitures par habitant très élevé, ce qui reflète ses spécificités socio-économiques et géographiques.

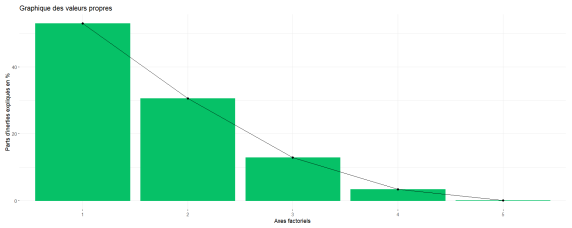


ACP par Régions

Nous passons maintenant à l'ACP par région, comme mentionné précédemment.

L'objectif est d'examiner si les départements suivent les mêmes tendances que leur région.
Si tel est le cas, cela pourrait suggérer que les politiques concernant le développement de l'électrique relèvent davantage d'une approche régionale que départementale.

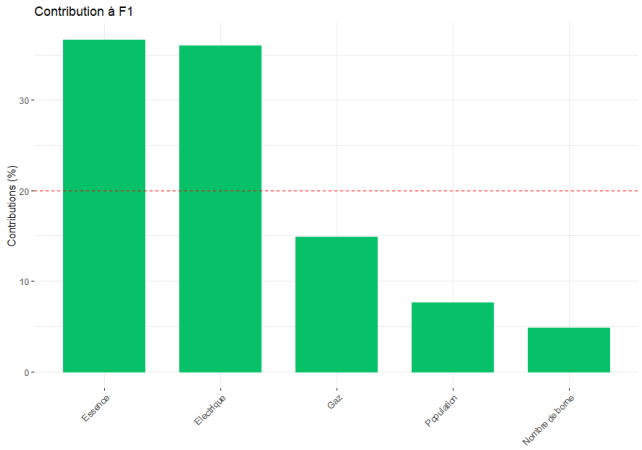
Inertie



	F1	F2	F3	F4	F5
Valeur Propre	2.654	1.530	0.645	0.170	0.001
Parts d'inertie	53.072	30.607	12.902	3.397	0.022
Parts cumulées	53.072	83.679	96.581	99.978	100.000

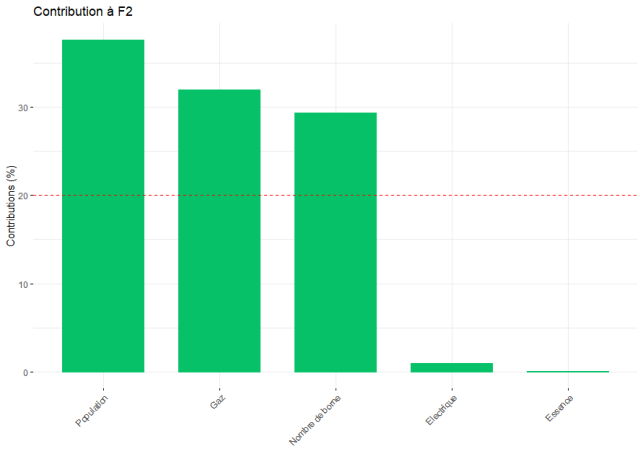
En observant notre graphique des valeurs propres, deux axes ressortent clairement.
Le premier explique à lui seul plus de 53,07 % de la dispersion, tandis que le second en capture un peu plus de 30.67 %.
Ensemble, ces deux axes totalisent plus de 83 % de la dispersion.
Cela les rend particulièrement pertinents pour poursuivre notre analyse et concentrer notre attention.

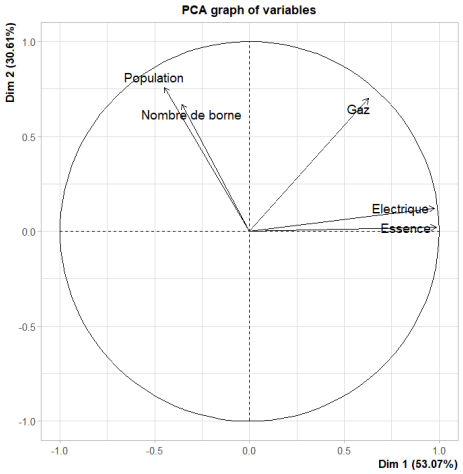
Contribution



Les deux variables qui contribuent le plus, et de manière assez équilibrée, sont "Essence" et "Électrique".
Ce résultat est similaire à ce que nous avions observé précédemment.
Ici aussi, l'axe 1 regroupe donc les régions présentant les plus forts taux de voitures, qu'elles soient électriques ou à essence, par habitant.

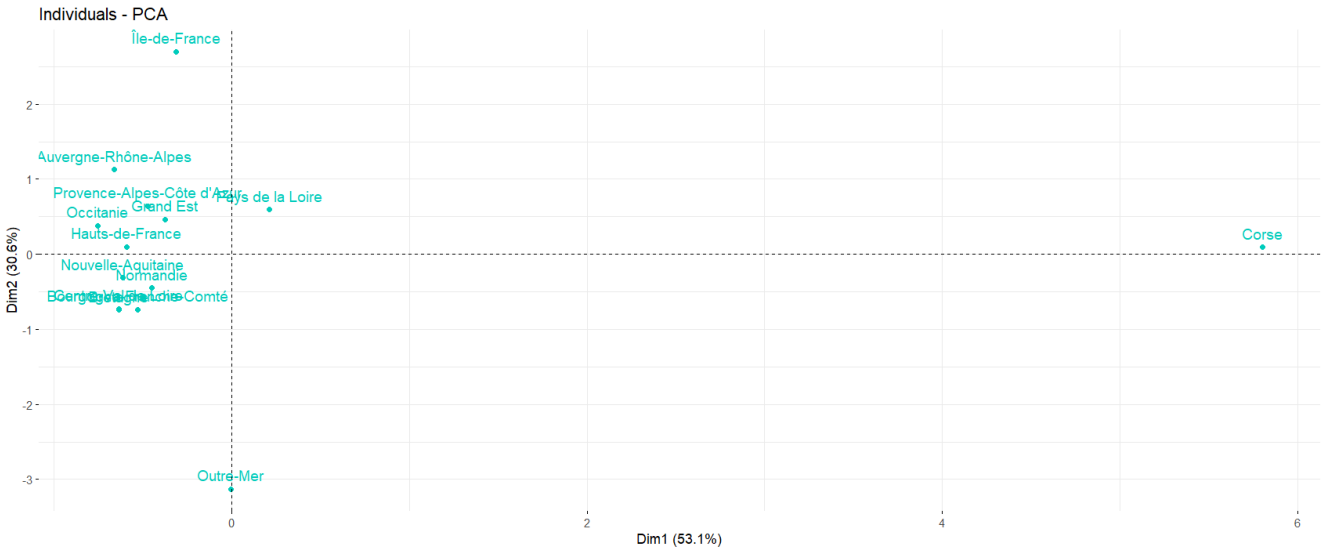
Pour l'axe 2, les résultats diffèrent davantage de ceux observés pour l'axe 1. Si la variable "Population" reste la plus contributive, son influence est moins marquée, tandis que la variable "Gaz" joue un rôle plus important. La principale différence réside dans le "Nombre de bornes", qui contribue désormais de manière significative, presque autant que "Gaz".
Cet axe 2 semble donc distinguer les régions fortement peuplées, qui investissent dans des infrastructures adaptées pour les bornes de recharge, des régions moins peuplées. Ce sont également des zones où le nombre de véhicules à gaz est relativement élevé.





Les résultats pour les deux premiers axes restent globalement similaires aux précédents. Cependant, on note que les variables "Électrique" et "Essence" sont désormais légèrement corrélées avec l'axe 2.
La variable "Gaz" continue de se répartir équitablement entre les axes 1 et 2.
Le principal changement vient du "Nombre de bornes", qui explique maintenant de manière significative l'axe 2, contrairement à l'ACP départementale. À l'échelle départementale, les bornes sont plus homogènes, tandis que des régions comme l'Île-de-France, avec un nombre plus élevé de bornes, montrent un impact plus marqué.

Nuage des individus



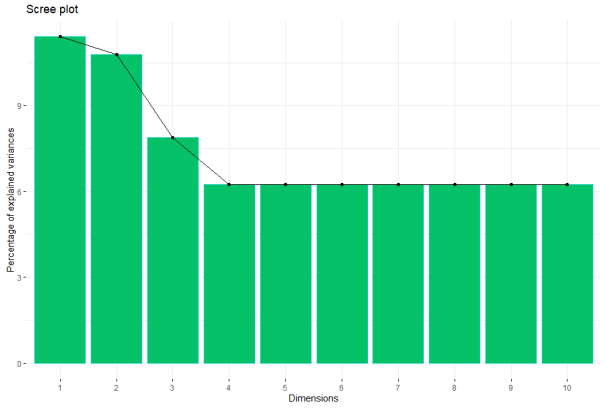
Ce que nous avons observé précédemment est clairement illustré par le nuage des individus.

Les régions se répartissent sur l'axe 2, avec à ses extrémités l'Île-de-France et l'Outre-Mer, très éloignées l'une de l'autre.
Cela montre une distinction entre les régions en fonction de leur population et de leur urbanisation. L'Île-de-France, plus urbanisée, est mieux équipée pour accueillir des infrastructures de bornes, contrairement à l'Outre-Mer.
Juste en dessous de l'Île-de-France, on trouve des régions comme Auvergne-Rhône-Alpes ou Provence-Alpes-Côte d'Azur, également urbanisées, tandis que la Bretagne et la Normandie sont moins urbanisées.
La Corse se distingue toujours sur l'extremité droite, en raison de son taux de voitures par habitant élevé malgré sa faible population.

ACM

Nous passons maintenant à l'Analyse des Correspondances Multiples (ACM), celle-ci est inévitable étant donné que de nombreuses variables de notre base de données sont qualitatives. Dans un premier temps, nous examinerons les inerties des axes. Ensuite, nous analyserons les contributions des axes, puis les nuages de points séparément.

Inertie

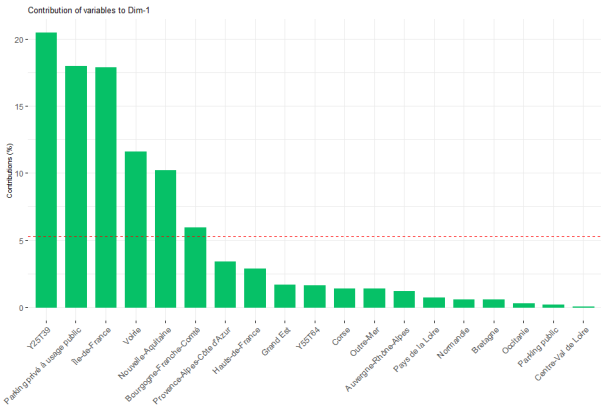


Le graphique des inertie nous indique qu'il est pertinent de concentrernotre étude sur les trois premières dimensions, qui expliquent à elles seules 33,74 % de la variance.

Au-delà de l'axe 4, la variance diminue fortement et se stabilise à des valeurs faibles, rendant ces axes peu intéressants pour l'analyse.

Contributions des axes

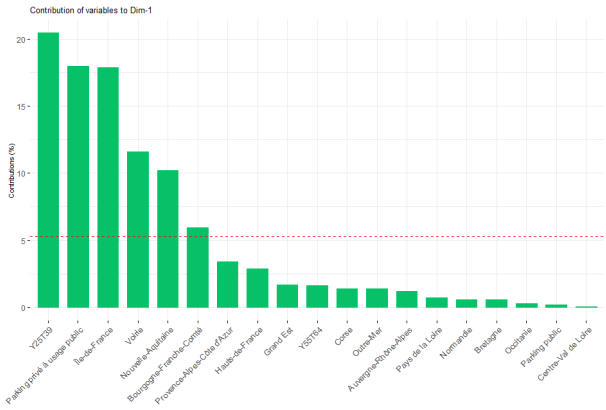
Axe 1



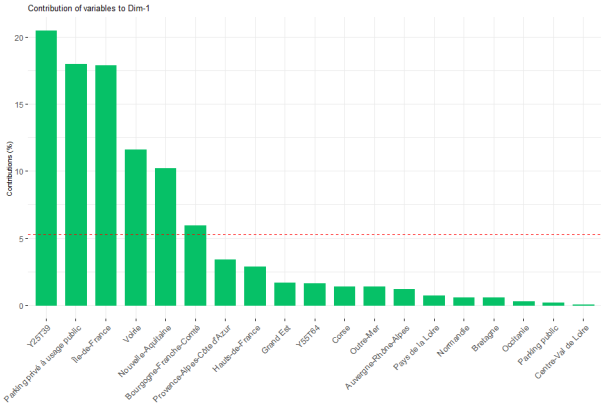
On observe que les variables les plus contributives à l'axe 1 sont les individus de sexe masculin et ceux âgés de 25 à 39 ans, qui totalisent à eux deux plus de 60 % de la contribution.
L'Île-de-France apporte également une légère contribution. Cet axe semble refléter une densité urbaine associée à un groupe socio-démographique jeune, davantage utilisateur de véhicules électriques.

Axe 2

Les variables les plus contributives à l'axe 2 sont "parking à usage public" et "voirie", accompagnées des régions Nouvelle-Aquitaine et Bourgogne-Franche-Comté.
Cet axe semble mettre en lumière une opposition entre des départements fortement urbanisés, disposant d'infrastructures de bornes accessibles sur la voirie et en usage public, et ceux moins équipés en la matière.



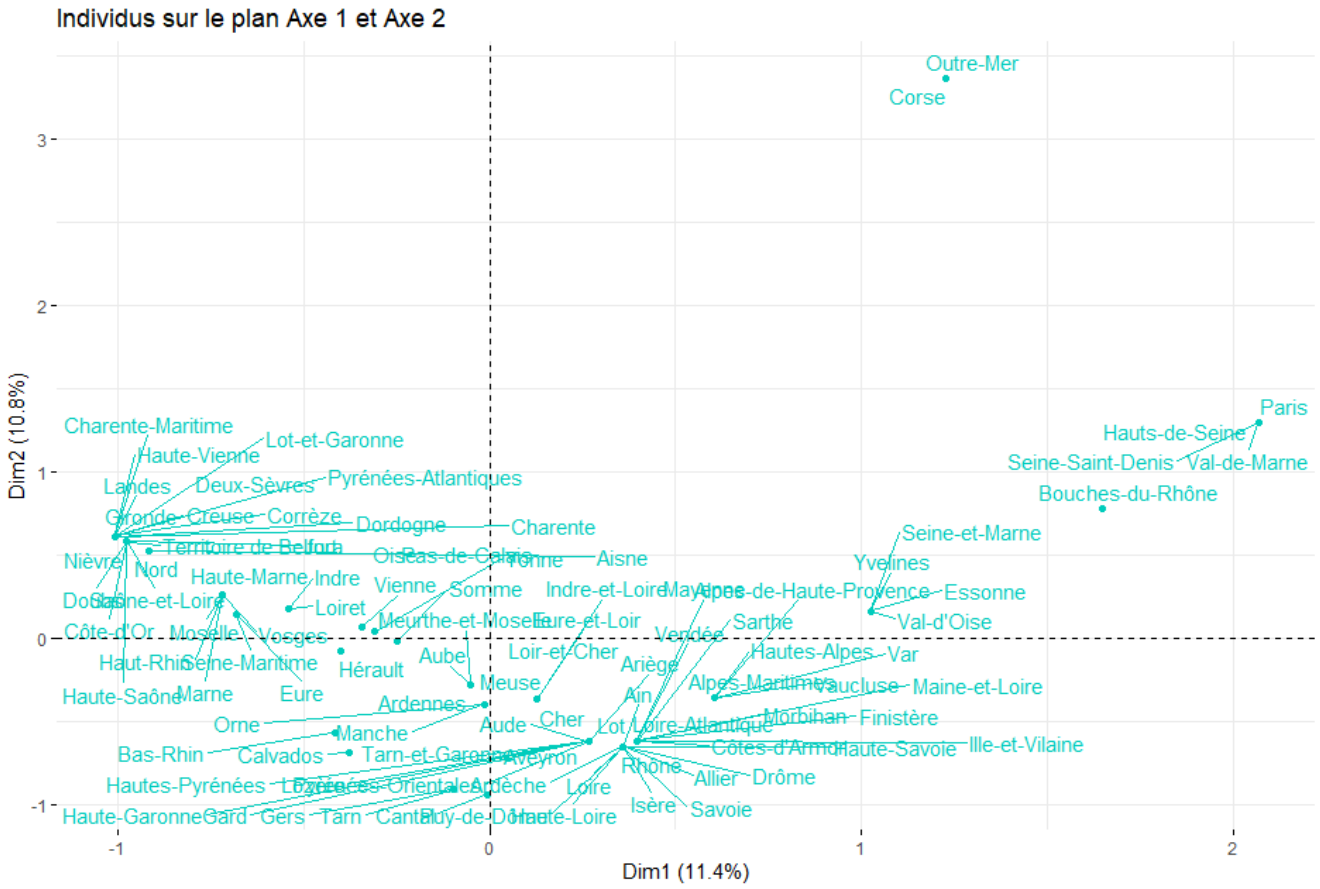
Axe 3



La variable qui domine le plus largement est parking public avec plus de 40%.
Les régions Normandie, Occitanie, et Pays de la Loire contribuent également significativement.
Cela pourrait indiquer des spécificités régionales liées à la répartition des infrastructures de recharge.

Nuages de points

Axe 1 & 2



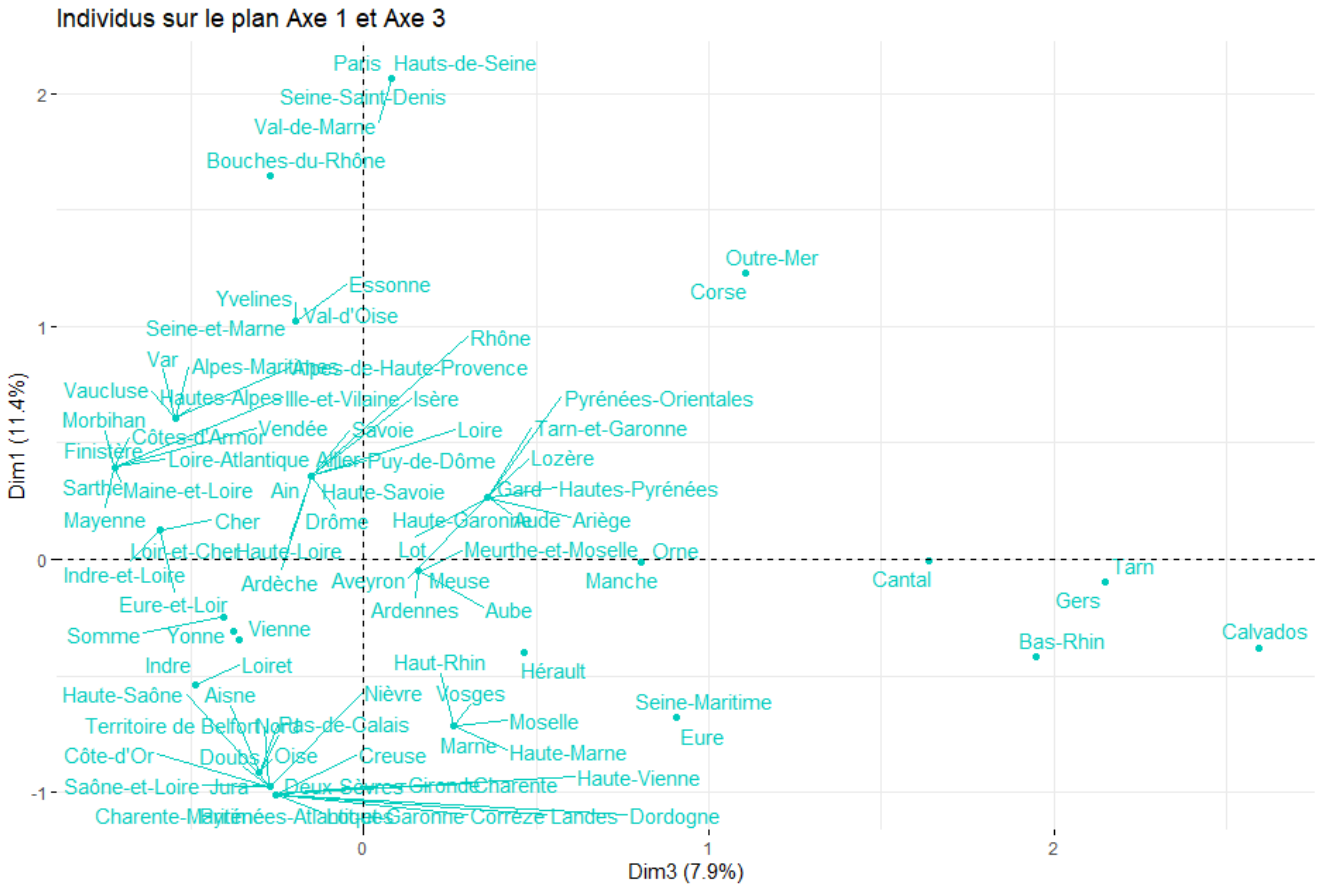
L'axe 1 reflète l'urbanisme et une population active plus jeune. Comme le montre le graphique, des départements comme Paris, les Hauts-de-Seine ou la Seine-Saint-Denis sont positivement corrélés à cet axe, ce qui est cohérent avec leur forte densité urbaine et leur population relativement jeune. Ce sont également les départements où l'utilisation de véhicules électriques est la plus élevée.

À l'opposé, sur la gauche de l'axe, on trouve des départements plus ruraux, où la moyenne d'âge est plus élevée et où les véhicules électriques ne sont pas une priorité. Les départements d'Outre-Mer et la Corse sont isolés, car ils présentent des modalités rares.

Pour l'axe 2, les départements qui contribuent le plus sont ceux de Nouvelle-Aquitaine, suivis de près par ceux de Bourgogne-Franche-Comté, des régions où les bornes sont majoritairement accessibles sur la voirie ou des parkings publics. En contraste, l'Occitanie et Auvergne-Rhône-Alpes se distinguent par une prédominance de bornes situées sur des parkings privés.

Ainsi, l'axe 2 illustre une opposition entre l'usage des bornes publiques et privées.

Axe 3 & 1



Avec ce graphique nous pouvons remarquer que les départements qui contribuent le plus à l'axe 3 sont le Calvados, le Tarn, le Bas-Rhin, le Gers et le Cantal qui se distinguent de par modalité parking public. On remarque que les départements plus ruraux comme ceux sités sont plus à meme d'avoir des parkings publics que des parking privé.

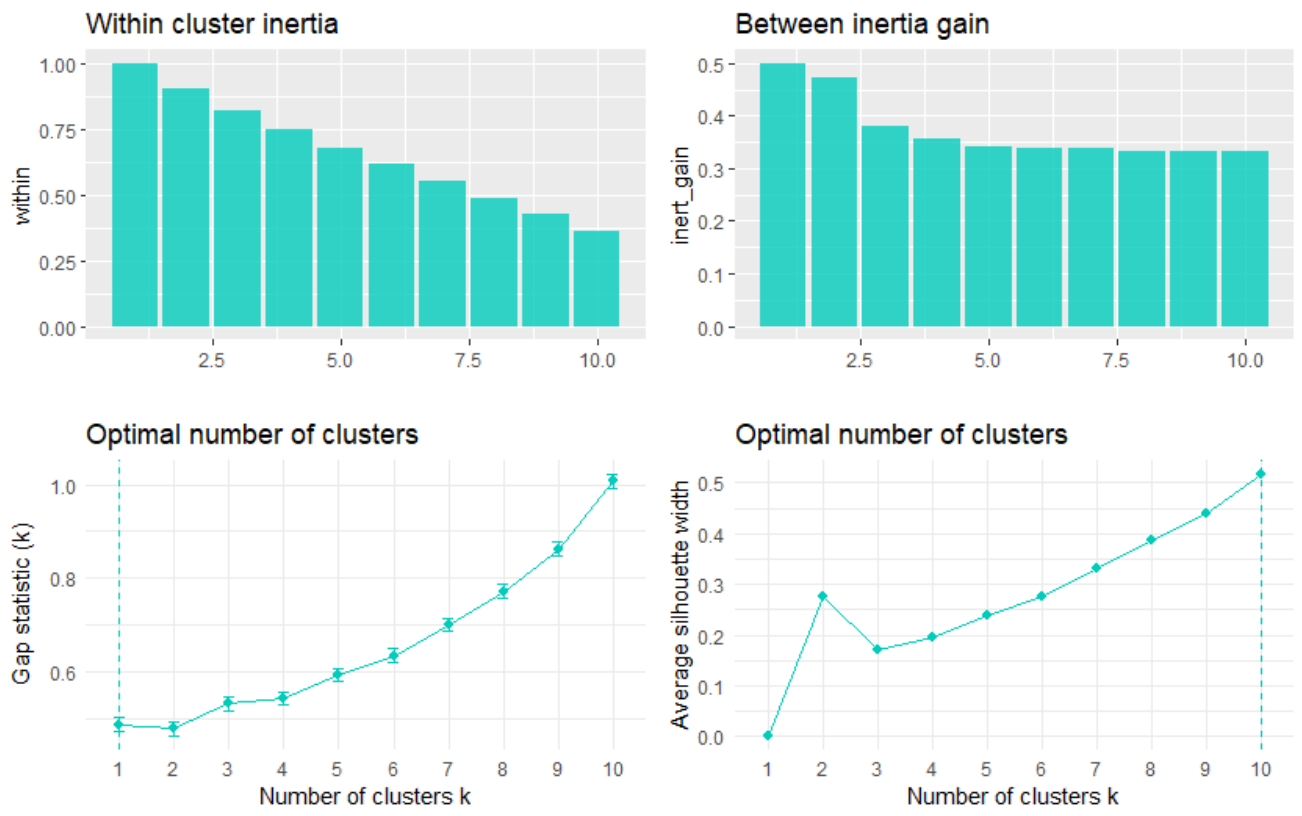
En effet les départements plus urbains si situent à leur opposé et les bornes sont plus souvent situés dans des parkings publics dans les zones ruales et inversement pour les zones urbaines.

Clustering

Nous allons maintenant réaliser un clustering afin d'identifier si des groupes de départements se forment en fonction des variables analysées précédemment.

Nombre de cluster

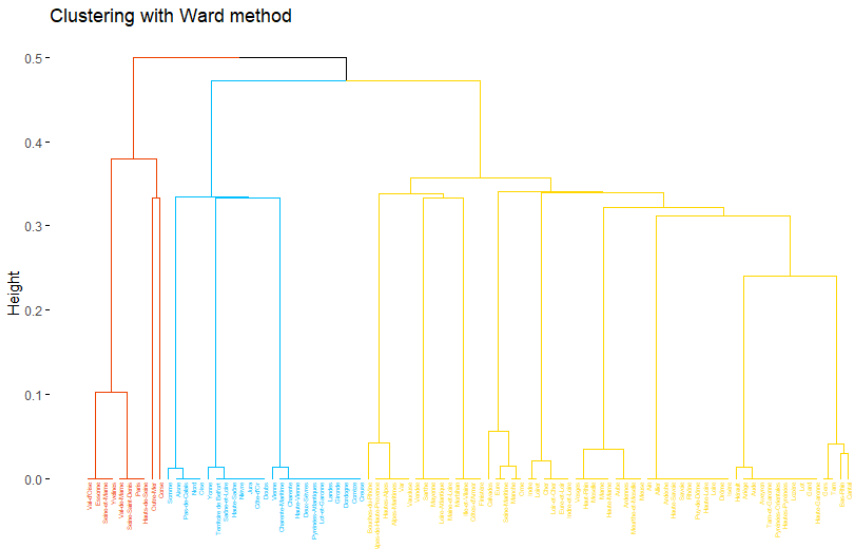
La première question qui se pose est de déterminer le nombre optimal de clusters. Pour cela, plusieurs méthodes différentes sont à notre disposition.



Le problème rapidement identifié est qu'aucune des techniques ne produit un résultat commun. Il nous revient donc de choisir une méthode ainsi qu'un nombre de clusters.

Nous avons opté pour trois clusters, comme suggéré par la méthode des gains d'inertie intra-cluster. Ce choix est également renforcé par l'évidence visuelle observée sur le dendrogramme.

Dendrogramme

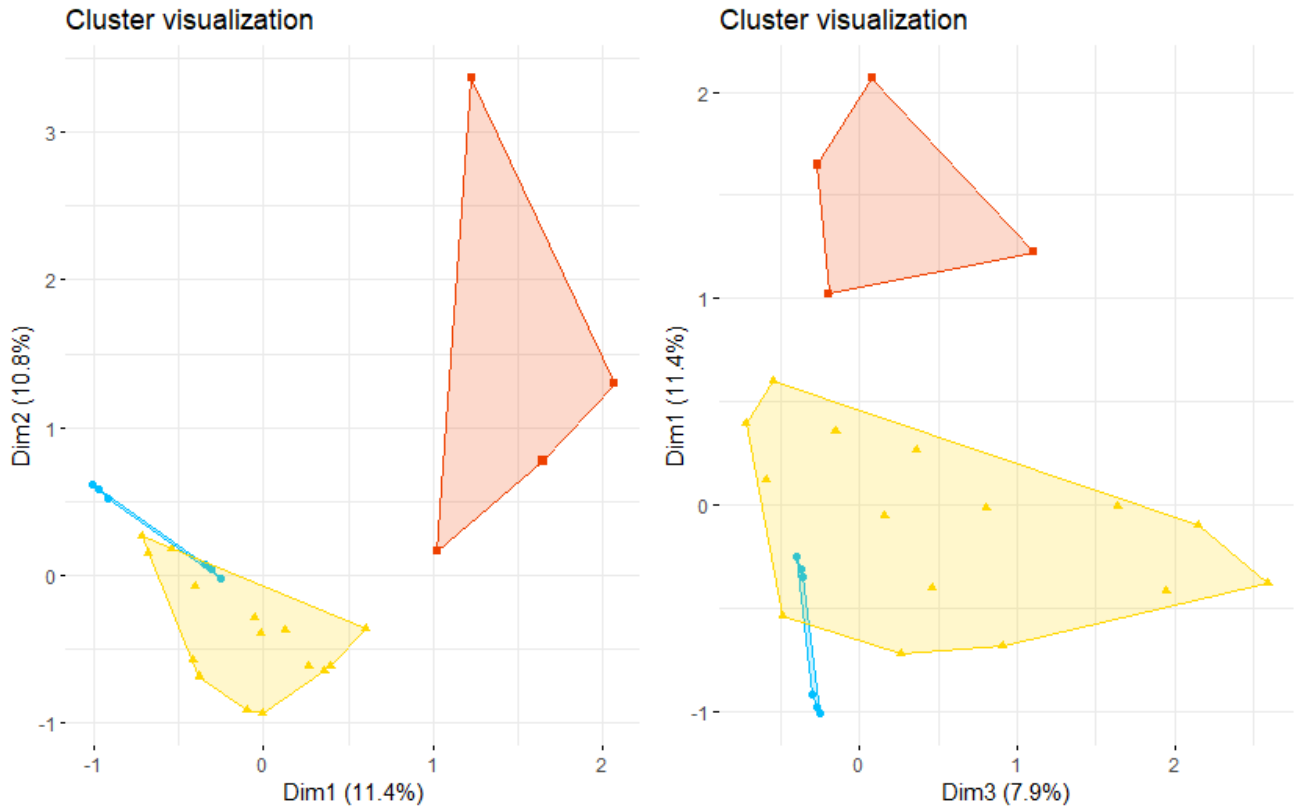


Trois groupes se dégagent immédiatement, ce qui est logique puisque l'on retrouve les mêmes tendances observées lors de l'analyse des axes.

Le premier cluster regroupe Paris, les Hauts-de-Seine et la Seine-Saint-Denis, des départements très urbains avec une population jeune et active, caractérisés par une forte présence de véhicules électriques et de bornes de recharge.

Le deuxième cluster rassemble les départements qui expliquent l'axe 2, et le troisième cluster suit une logique similaire.

Visualisation



La visualisation des résultats nous permet de constater que les clusters sont bien définis et qu'ils séparent les individus de manière pertinente en fonction des axes étudiés. Cette séparation logique et claire nous confirme que le nombre de clusters choisi est adapté et que ces derniers reflètent bien les groupes de départements en fonction des caractéristiques analysées. Cela nous rassure donc sur la validité et la pertinence de notre choix de clustering.

Conclusion

D'un point de vue global, les résultats reflètent des disparités importantes entre les zones urbaines et rurales. Les régions les plus urbanisées se distinguent par une forte densité de bornes, en lien avec une population jeune et active, tandis que les territoires plus ruraux, bien qu'équipés, montrent une dynamique différente, souvent marquée par des infrastructures publiques pour pallier une densité plus faible.

Cette ACM met en évidence trois grands groupes de départements. Le premier regroupe les zones fortement urbanisées, avec une densité élevée de bornes et des usagers jeunes et actifs. Le deuxième groupe est constitué de départements où les infrastructures sont majoritairement accessibles au public, souvent situées dans des régions intermédiaires en termes d'urbanisation. Enfin, le troisième groupe regroupe des territoires plus ruraux où les bornes se situent souvent dans des espaces publics, illustrant des besoins spécifiques et une autre dynamique d'aménagement.

En conclusion, cette ACM clôture l'étude en soulignant l'importance de prendre en compte les disparités territoriales dans le déploiement des bornes de recharge. Ces différences reflètent non seulement des caractéristiques socio-démographiques mais aussi des choix d'aménagement adaptés aux spécificités locales. Ces résultats apportent une meilleure compréhension des enjeux liés à l'électromobilité et permettent de dégager des pistes pour des politiques publiques et des stratégies d'aménagement mieux ciblées.

Sources

- Régions & départements : https://www.data.gouv.fr/fr/datasets/departements-et-leurs-regions/
- Population : https://catalogue-donnees.insee.fr/fr/catalogue/recherche/DS_RP_POPULATION_PRINC
- Véhicules : https://defis.data.gouv.fr/datasets/628310650b8550478a9ddd2d
- Bornes : https://defis.data.gouv.fr/datasets/5448d3e0c751df01f85d0572