MSPR Bloc 3 Big Data

Réaliser par

Salima BOUZERA, Marion MAGAUD, Roumaissa MOUNIR, Myriam ZALANI

Table des matières

I. Introduction

La start-up fictive de Jean-Edouard de la Motte Rouge, spécialisée dans le conseil en campagnes électorales, envisage d'utiliser l'intelligence artificielle pour prédire les tendances électorales futures. Pour ce faire, elle souhaite créer une preuve de concept (POC) basée sur des données géographiquement restreintes. L'objectif est de démontrer la capacité de l'IA à anticiper les résultats électoraux en se basant sur divers indicateurs tels que la sécurité, l'emploi, la vie associative, la population, et d'autres facteurs pertinents.

Dans le cadre de ce projet, notre équipe se charge de sélectionner un secteur géographique spécifique, de collecter des jeux de données pertinents, de créer des visualisations pour en faciliter la compréhension, de développer un modèle prédictif supervisé, et enfin, de présenter visuellement les prédictions sur 5 ans. Cette POC permettra à la start-up de prendre des décisions éclairées concernant d'éventuels investissements dans l'infrastructure et la recherche et développement pour développer cette solution innovante.

Ce rapport détaille notre démarche pour répondre aux besoins de la start-up, en mettant l'accent sur la collecte, l'analyse, et la présentation des données, ainsi que sur la création d'un modèle de prédiction fiable pour aider Jean-Edouard de la Motte Rouge à gagner un avantage concurrentiel significatif dans son domaine d'activité.

La POC que nous allons développer est conçue sur un secteur géographique restreint et unique (ville). Elle implique les étapes suivantes :

- **Sélection du Secteur Géographique** : le Choix d’une zone géographique spécifique où nous effectuons notre analyse et nos prédictions électorales. Cette zone servira de base à notre POC.
- **Collecte de Données** : la sélection des jeux de données pertinents. Ces données incluent des indicateurs tels que la sécurité, l'emploi, la population, et d'autres facteurs liés aux élections. Ces données alimenteront notre modèle de prédiction.
- **Visualisation des Données** : On a créé des visualisations graphiques pour rendre les données plus compréhensibles. Les graphiques et les tableaux aideront à mettre en évidence les tendances et les corrélations entre les indicateurs et les résultats électoraux passés.
- **Modèle Prédictif Supervisé** : Divisez les données en jeux d'apprentissage et de test, puis développez un modèle prédictif supervisé. Ce modèle utilisera les données historiques pour anticiper les résultats électoraux futurs.
- **Prédictions sur 5 ans** : Une fois le modèle construit, on propose des visualisations graphiques qui montrent les prédictions électorales pour les prochaines années (dans 5 ans). Ces prédictions aideront la start-up à prendre des décisions éclairées.

II. Secteur géographique

Afin d'avoir les meilleures prédictions possibles nous avons choisi de faire notre POC sur une sélection des neuf plus grandes villes de France (hors Paris). En ayant un échantillonnage de neuf grandes villes cela va lisser les orientations politiques et donc de permettre une meilleure prédiction. Nous avons choisi d'exclure Paris car il y a trop de divergences d'orientation politique / différence de salaire / de sécurité / … au sein des différents arrondissements et cela risquerait de fausser les résultats finaux.

III. Collecte des données

A. Sécurité

Le dataset comprend des données sur la délinquance enregistrée dans différentes communes et départements pour les années 2002, 2007, 2012, 2017, et 2022. Les principales variables incluent l'année de l'enregistrements, le type d'indicateur de crime ou de délit, le nombre de faits de délinquance enregistrés, ainsi que la ville. Les indicateurs de délinquance couvrent un large éventail, notamment les cambriolages, les coups et blessures volontaires, les destructions et dégradations, les violences sexuelles, les vols avec armes, les vols violent sans arme, les vols de véhicules, les vols dans les véhicules, les vols d'accessoires sur véhicules, et les vols sans violence contre des personnes. Ce dataset offre une vue complète des tendances de la délinquance au fil des années et dans différentes régions géographiques.

B. Population

Le dataset présente des données historiques de la population résidant en France métropolitain, pour les années 2002, 2007, 2012, 2017 et 2022. Il se focalise sur dix grandes villes françaises, notamment Marseille, Lyon, Toulouse, Nice, Nantes, Strasbourg, Bordeaux, Lille et Rennes. Les variables clés comprennent l'année de recensement, la ville de résidence. Ce dataset offre une précieuse perspective sur l'évolution démographique au fil du temps dans ces villes majeures.

C. PIB par Habitant

Le dataset présente des données sur le Produit Intérieur Brut (PIB) par habitant en euros pour différentes villes françaises, pour les années 2002, 2007, 2012, 2017 et 2022. Les trois principales variables incluses sont l'année de référence, les noms des villes (telles que Marseille, Lyon, Toulouse, Nice, Nantes, Strasbourg, Bordeaux, Lille et Rennes), et le PIB en euros par habitant.

D. Chômage

Le dataset présente des données brutes sur le nombre de demandeur d'emplois (nombre de personne inscrits à pôle emplois) dans les années 2002, 2007, 2012, 2017 et 2022 dans les villes suivantes Marseille, Lyon, Toulouse, Nice, Nantes, Strasbourg, Bordeaux, Lille et Rennes.

E. Revenus & impôts

Le dataset présente des données des revenus d'activité et des impôts en euro par habitant dans neuf grandes villes française Marseille, Lyon, Toulouse, Nice, Nantes, Strasbourg, Bordeaux, Lille et Rennes pour les années 2002, 2007, 2012, 2017 et 2022.

F. Election présidentielle

Le dataset présente des données sur différents candidats aux élections présidentielles. Il contient notamment les noms et prénoms des candidats ainsi que le nombre de voix qu'ils ont obtenu pour chaque commune. Ce dataset est construit à partir des dataset des élections 2002, 2007, 2012, 2017 et 2022 pour les 10 plus grandes villes de France (hors Paris). (voir Concatenation_vote.ipynb)

G. Parti Politique

Le dataset présente des données sur les différents candidats aux élections présidentielles et leurs orientations politiques (extrême droite/ gauche, centre droite/ gauche, centre/ droite/ gauche). Ce dataset contient la liste des candidats qui se sont présentés pour les élections des années 2002, 2007, 2012, 2017 et 2022. L'orientation politique est celle que les candidats avaient lorsqu'ils se sont présentés.

IV. Visualisation des données

A. Dictionnaire de données

a) Sécurité

Nom de la table	Nom de la vari	Definition	Propriété	Type	Unité	Source web	Exemple	Remarque
Sécurité.xlsx	classe	Indicateur des crimes et d	Indicateur des crimes et délits	string		https://www.data.gouv.fr/fr/datasets/bases-statistiques-commu		
Sécurité.xlsx	annee	Année	Année pendant laquelle la délinquance a été enregistrée	integer		https://www.data.gouv.fr/fr/datasets/bases-statistiques-commu		
Sécurité.xlsx	Villes	Villes	ville ou la délinquance a été enregistrée	string		https://www.data.gouv.fr/fr/datasets/bases-statistiques-commu		
Sécurité.xlsx	faits	faits	Nombre de faits de délinquance enregistrés établis en commune de commis	integer		https://www.data.gouv.fr/fr/datasets/bases-statistiques-commu		

b) Population

Nom de la table	Nom de la variable	Definition	Propriété	Type	Unité	Source web	Exemple	Remarque
population_totale_par_ville.xlsx	Année	Année		Integer		https://www.insee.fr/fr/statistiques/1893204#consulter	2002/2007/2012/2017/2022	
population_totale_par_ville.xlsx	Ville	Ville		String		https://www.insee.fr/fr/statistiques/1893204#consulter	Marseille','Lyon','Toulouse','Nice','Nantes','Stras bourg','Bordeaux','Lille','Rennes'	
population_totale_par_ville.xlsx	Population totale	Population totale		Integer		https://www.insee.fr/fr/statistiques/1893204#consulter		

c) PIB par Habitant

Nom de la table	Nom de la variable	Propriété	Type	Unité	Source web
valeur_pib_par_habitant .xlsx	Année	Année pendant laquelle le PIB a été enregistrée	int		https://www.insee.fr/fr/statist
valeur_pib_par_habitant .xlsx	Villes	nom de la ville	String		https://www.insee.fr/fr
valeur_pib_par_habitant .xlsx	PIB par habitant	PIB par habitant	int	euro par habitant	https://www.insee.fr/fr

d) Chômage

Nom de la table	Nom de la variable	Propriété	Type	Unité	Source web
nb_demandeurs_emploi.xlsx	Année	Année pendant laquelle le nb demandeurs d'emploi a été enregistré	int		https://statistiques.pole-emploi.org/stmt/defm?qp=1&fi
nb_demandeurs_emploi.xlsx	Villes	nom de la ville	String		https://statistiques.pole-emploi.org/stmt/defm?qp=1&fi
nb_demandeurs_emploi.xlsx	Nb_DE	nombre de demandeurs d'emploi	int	nb personnes	https://statistiques.pole-emploi.org/stmt/defm?qp=1&fi

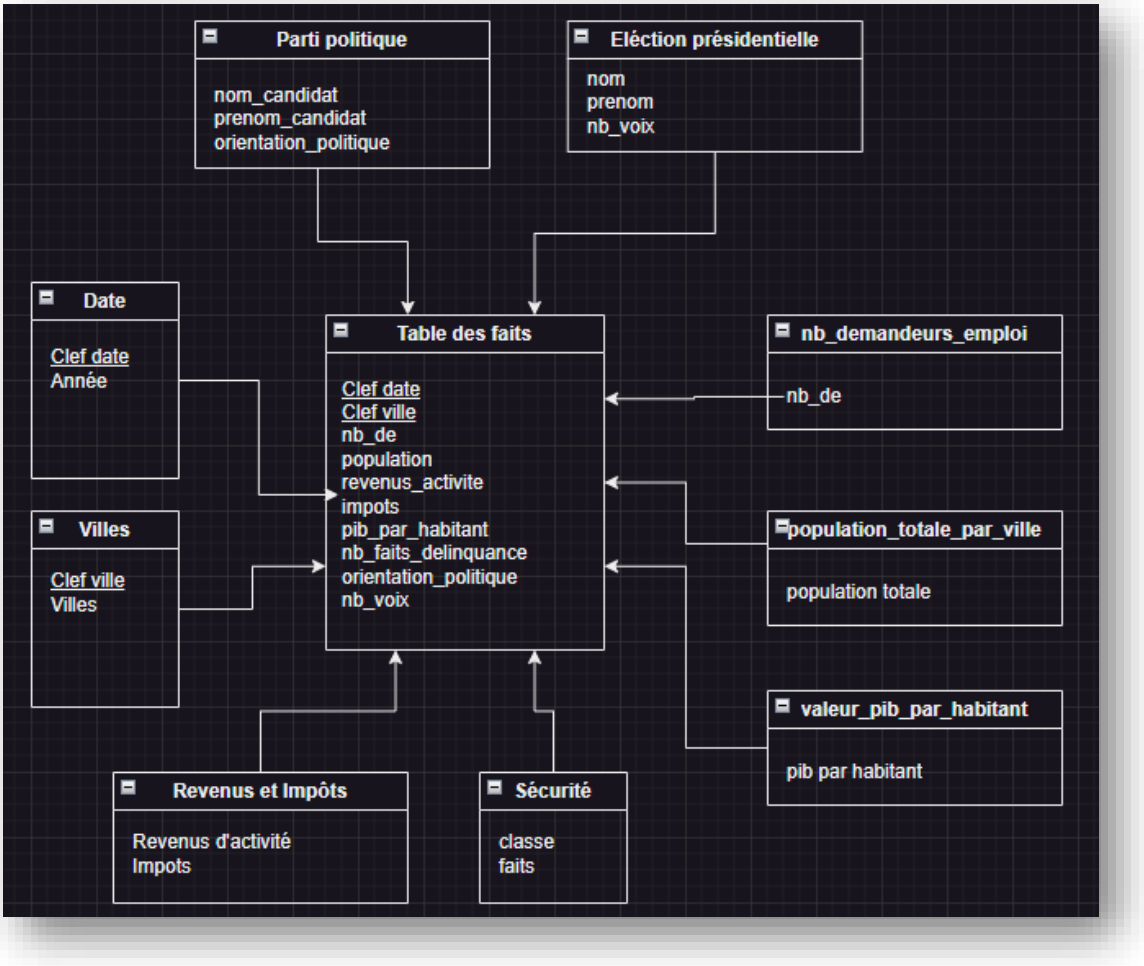
e) Revenus & impôts

Nom de la table	Nom de la variable	Propriété	Type	Unité	Source web
Revenus et Impôts .xlsx	Année	Année pendant laquelle le f	int		https://www.insee.fr/fr
Revenus et Impôts .xlsx	Villes	nom de la ville	String		https://www.insee.fr/fr
Revenus et Impôts .xlsx	Revenus d'activité	euro	int	euro par habitant	https://www.insee.fr/fr
Revenus et Impôts .xlsx	Impots	euro	int	euro par habitant	https://www.insee.fr/fr

f) Election présidentielle et parti politique

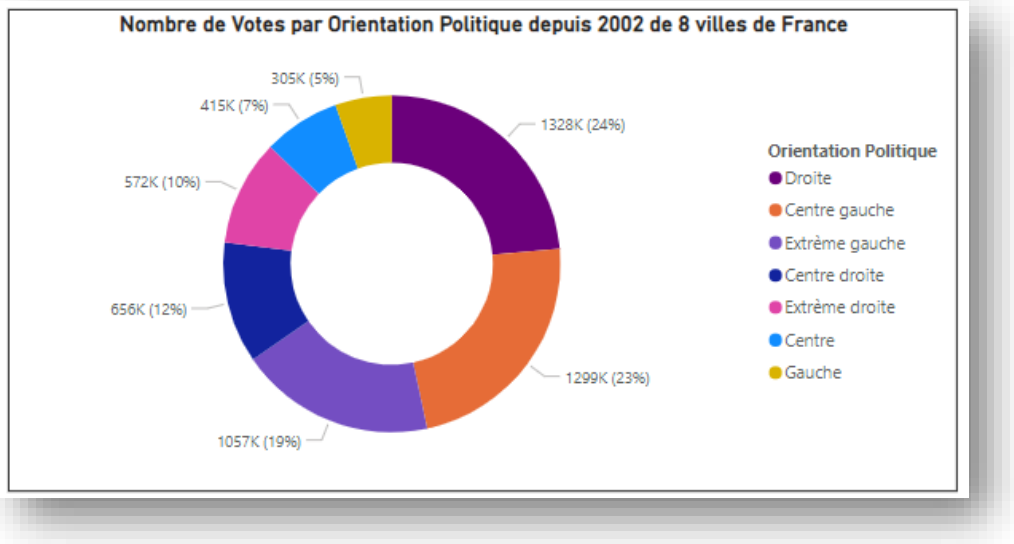
Nom de la table	Nom de la variable	Propriété	Type	Unité	Source web
Eléction présidentielle	ville	Top 10 villes France (Sans Paris)	String		https://www.data.gouv.fr/fr/pages/donnees-des-elections/
Eléction présidentielle	nom	Nom de famille du/de la candidat(e)	String		https://www.data.gouv.fr/fr/pages/donnees-des-elections/
Eléction présidentielle	prenom	Prénom du/de la candidat(e)	String		https://www.data.gouv.fr/fr/pages/donnees-des-elections/
Eléction présidentielle	nb_voix	Nombre de voix obtenues	int		https://www.data.gouv.fr/fr/pages/donnees-des-elections/
Eléction présidentielle	annee	Année	int		https://www.data.gouv.fr/fr/pages/donnees-des-elections/
Parti politique.xlsx	nom_candidat	Nom de famille du/de la candidat(e)	String		Création manuelle
Parti politique.xlsx	prenom_candidat	Prénom du/de la candidat(e)	String		Création manuelle
Parti politique.xlsx	orientation_politique	Orientation politique du/de la candidat(e	String		Création manuelle

Avec les différents jeux de données nous obtenons comme table des faits :

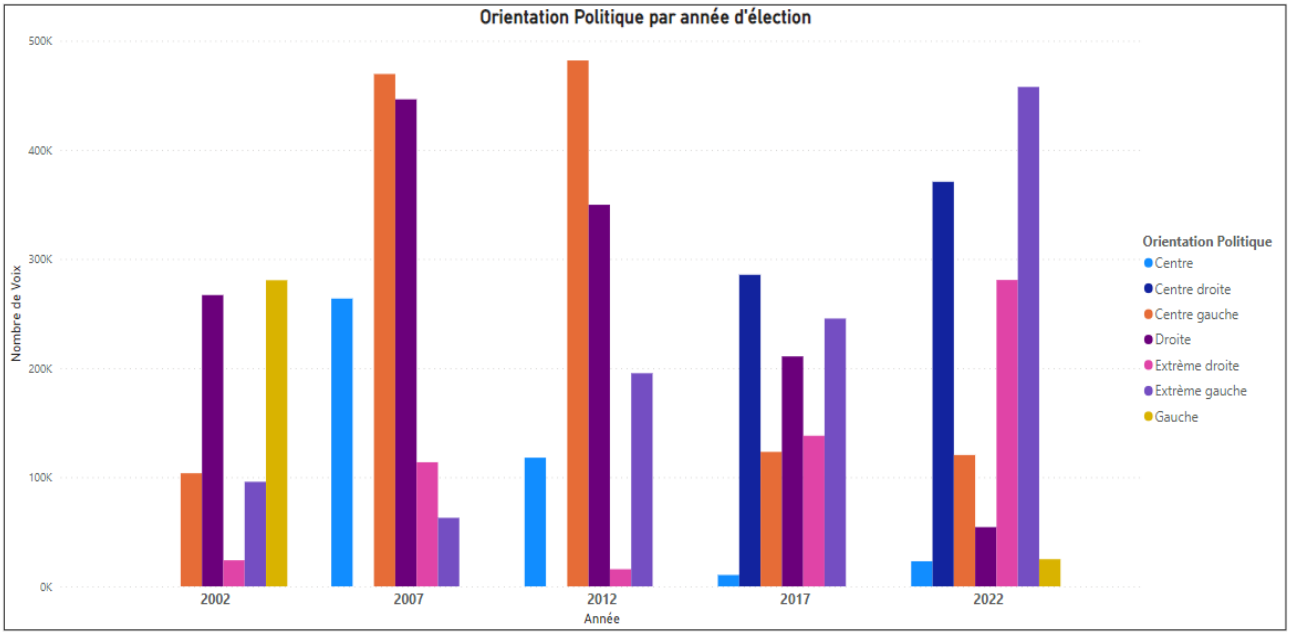


B. Analyse des données

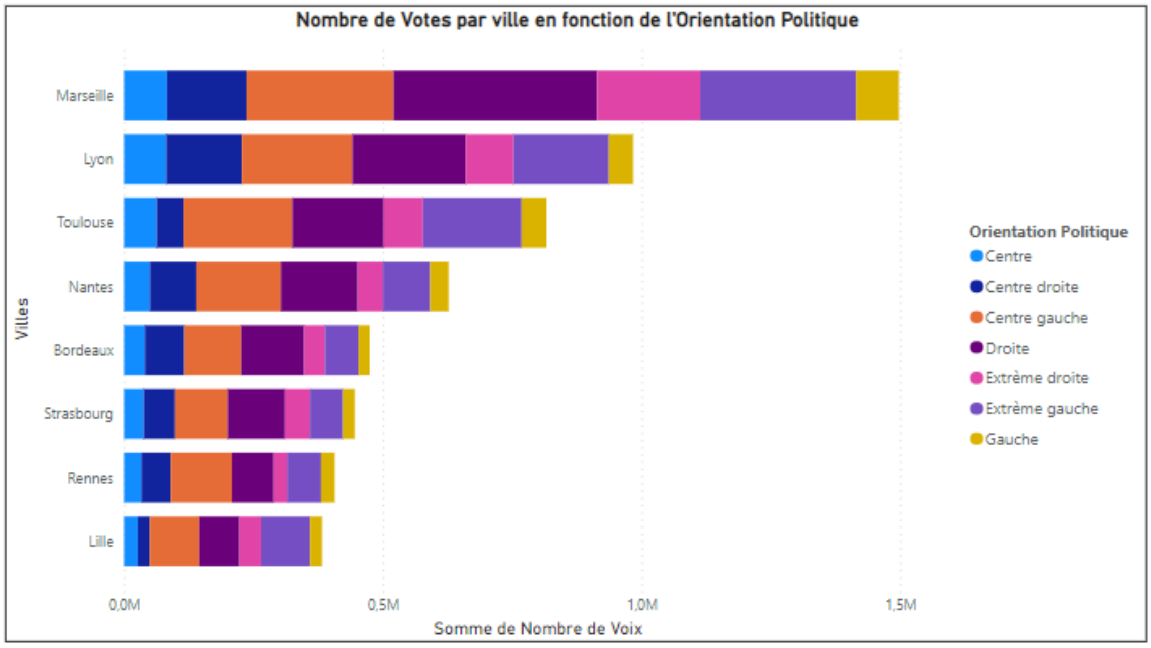
Nous avons utilisé l'outil Power BI pour réaliser différents visuels pour nous permettre de mieux analyser nos données.



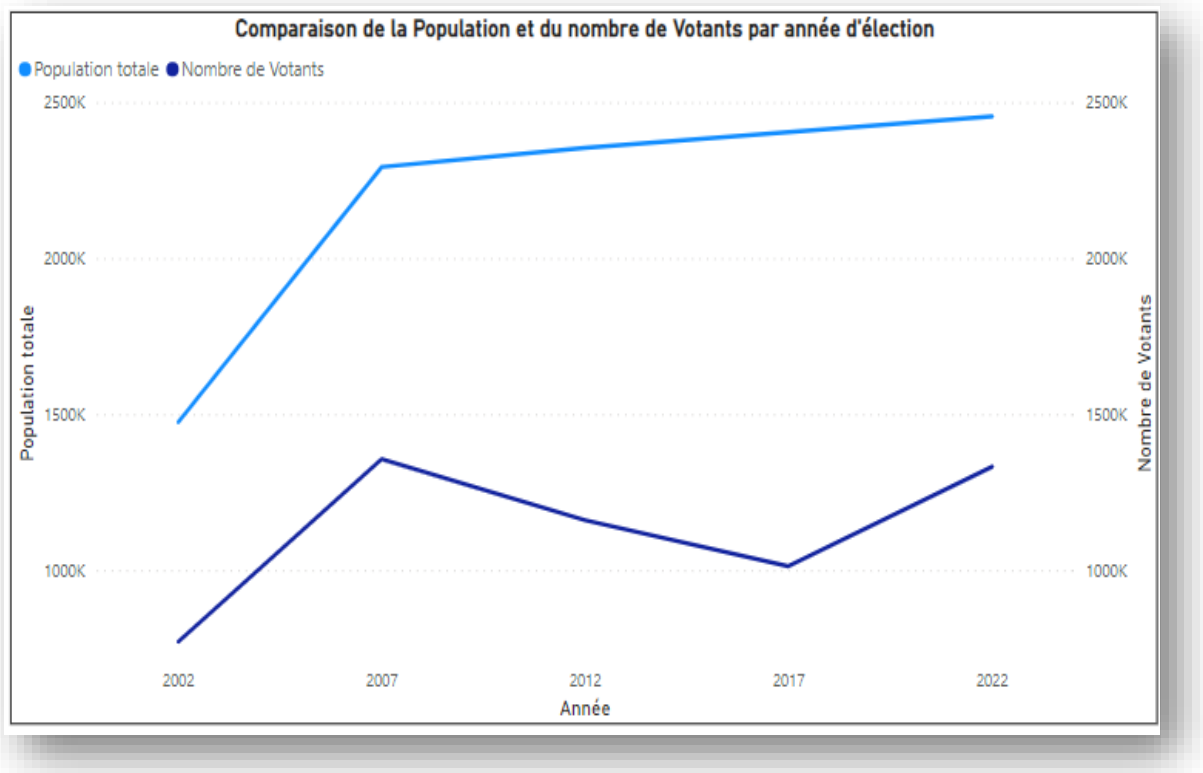
Le diagramme en anneau ci-dessus représente le nombre de votes par orientation politique dans 8 villes de France depuis 2002. Les sections sont colorées pour indiquer les différentes orientations politiques, telles que l'extrême droite, la gauche, etc. On peut remarquer une tendance à voter majoritairement pour la droite, le centre gauche et l'extrême gauche. Et une tendance à voter minoritairement pour la gauche et le centre.



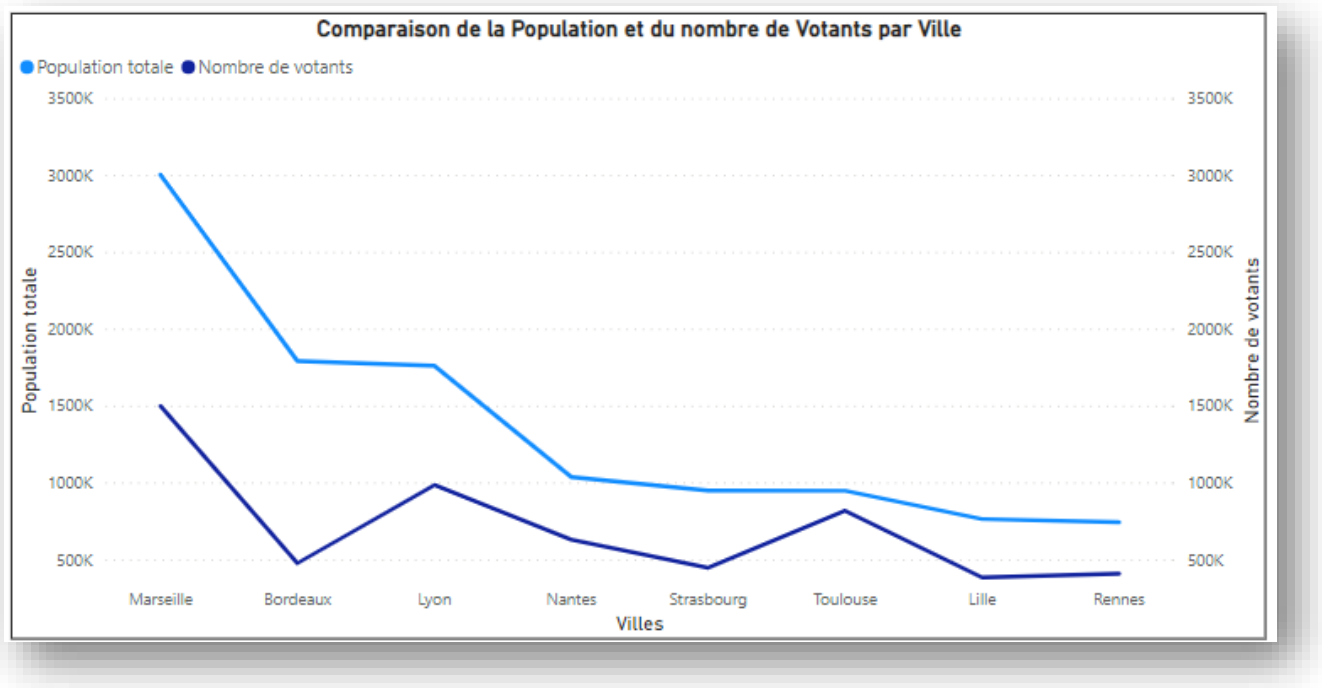
Le diagramme histogramme ci-dessus représente l'évolution du nombre de votes par orientation politique depuis 2002. On peut remarquer que le nombre de votes pour l'extrême gauche a augmenté régulièrement depuis 2002, tandis que d'autres parties tels que la droite ou la gauche ont connu une baisse constante.



Le diagramme à barres ci-dessus représente le nombre de votes par ville en fonction de l'orientation politique. On peut remarquer que par exemple la ville de Marseille vote majoritairement pour la droite ou l'extrême gauche et minoritairement pour le centre ou la gauche.



Le diagramme en courbes ci-dessus permet de comparer la population totale et le nombre de votants par année d'élection. On peut remarquer qu'il y a eu une hausse de la population qui elle est proportionnel à la hausse des votants en 2007, puis une forte baisse de votant jusqu'en 2017 et ensuite une forte hausse des votants en 2022.

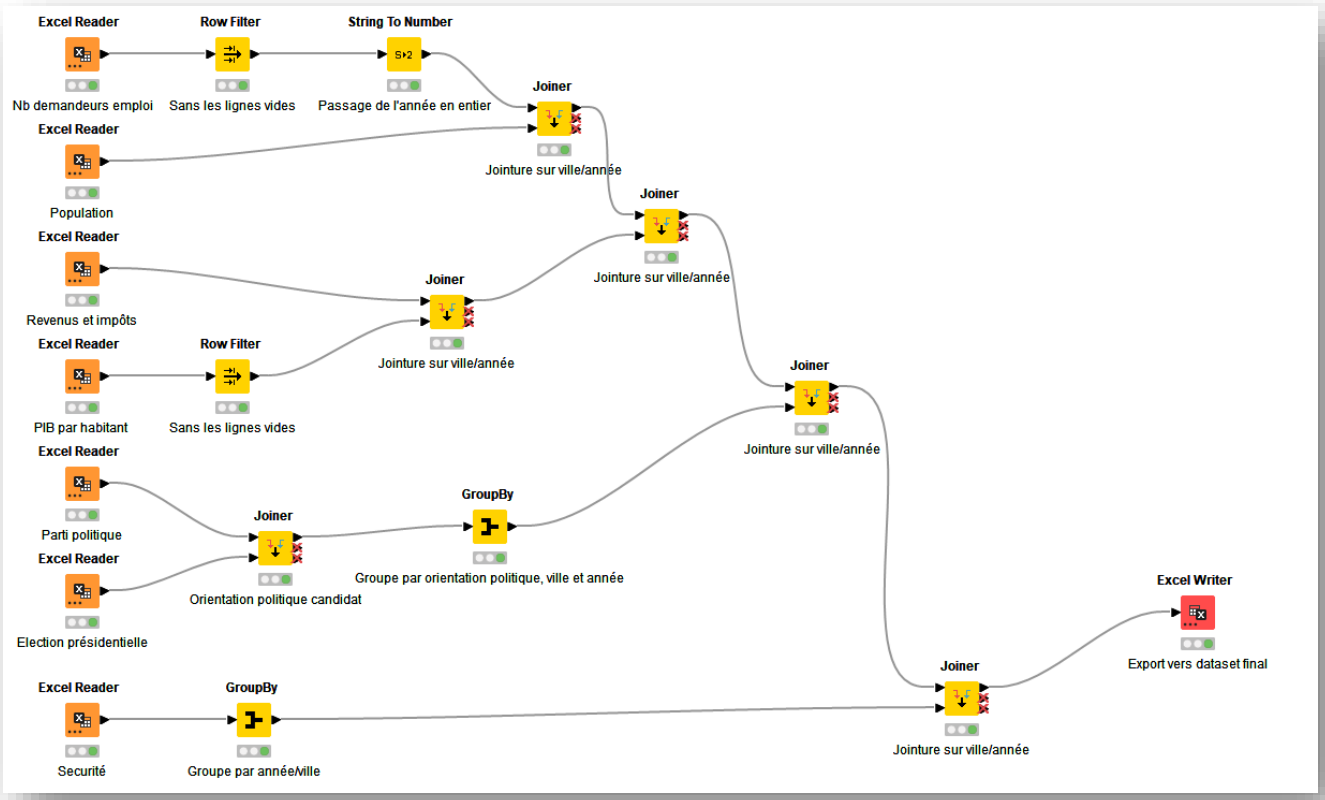


Le diagramme en courbes ci-dessus permet de comparer la population totale et le nombre de votants par ville. On peut remarquer que par exemple qu'il y a moins de 50% de la population de la ville de Bordeaux qui ne vote pas. Alors que pour la ville de Toulouse, le nombre de votants est quasiment égal au nombre d'habitants.

C. Flux ETL

Nous avons utilisé Knime afin de faire une jointure des différents dataset et de pouvoir faire un export en dataset_version_final.xlsx

(voir MSPR3_merge_all_dataset.knwf)



MSPR3_merge_all_dataset.knwf

D. Description du Dataset final

Cet ensemble de données est le résultat de la fusion de plusieurs sources de données distinctes, fournissant des informations pertinentes sur différentes années, villes, indicateurs socio-économiques, et variables politiques. Ces données sont recueillies sur une période de plusieurs années, offrant ainsi une vue globale et historique des caractéristiques de différentes régions.
Colonnes :

- Année : L'année à laquelle les données ont été enregistrées.
- Villes : Le nom de la ville concernée par les données.
- Nb_de_population : Le nombre total de population dans la ville.
- Population : La population totale de la ville.
- Revenus_activité : Les revenus issus de l'activité économique dans la ville.
- Impôts : Les impôts collectés dans la ville.
- PIB_par_habitant : Le produit intérieur brut (PIB) par habitant dans la ville.
- Nb_faits_delinquance : Le nombre total de faits de délinquance enregistrés.
- Orientation_politique : L'orientation politique dominante dans la ville.
- Nb_voix : Le nombre de voix ou votes enregistrés pour cette orientation politique.

Nombre de lignes : 222

Sources de données :

Ce dataset a été créé en combinant plusieurs ensembles de données provenant de sources diverses, afin de fournir une vue d'ensemble complète des données socio-économiques et politiques à travers le temps et l'espace.

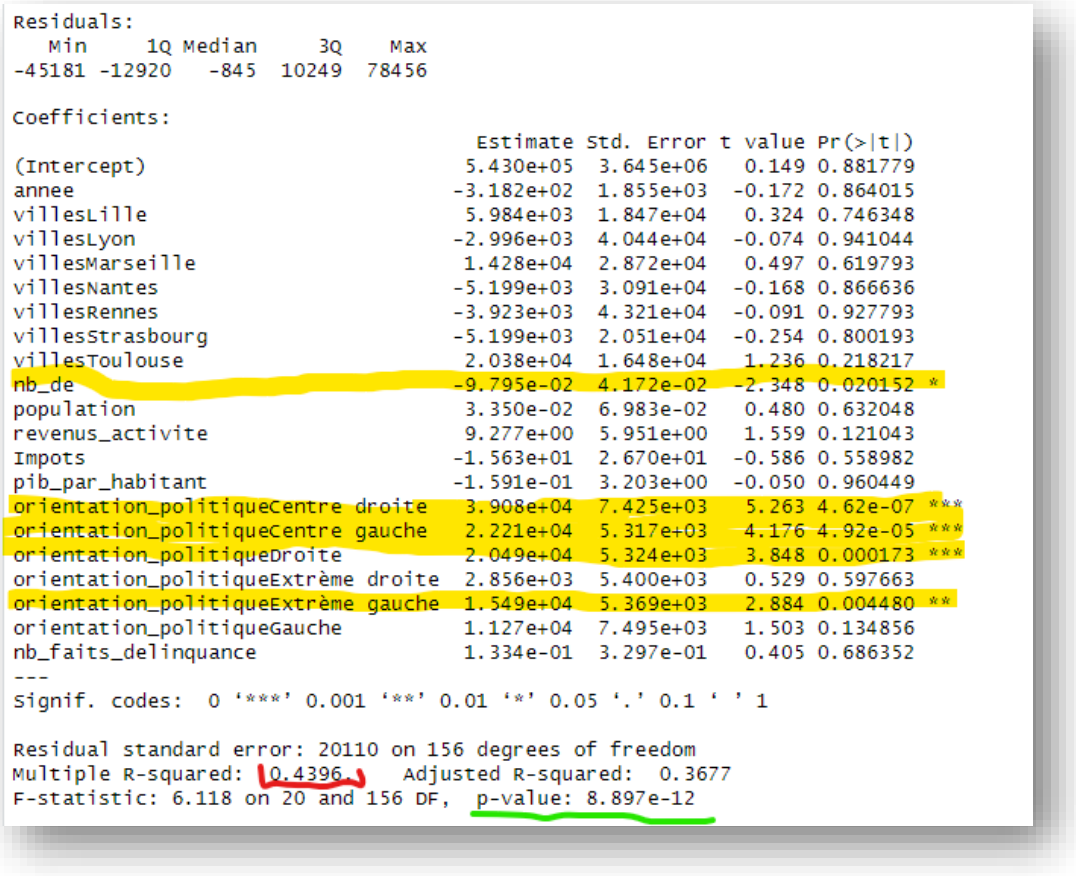
V. Modèle Prédictif Supervisé

A. Régression linéaire

(voir MSPR3_prédiction.Rmd)

Nous pouvons voir sur l'image ci-dessous que le nombre de demandeur d'emploi ainsi que certaines orientations politiques sont significatives par rapport à notre modèle.

La p-value est faible (<5%) ce qui veut dire que notre modèle est significatif. En revanche le modèle a un mauvais pouvoir prédictif environ 43%. Le mauvais pouvoir prédictif s'explique par le faible nombre de ligne de notre jeu de données. Un plus grand jeu de données permettrait de faire un meilleur modèle.

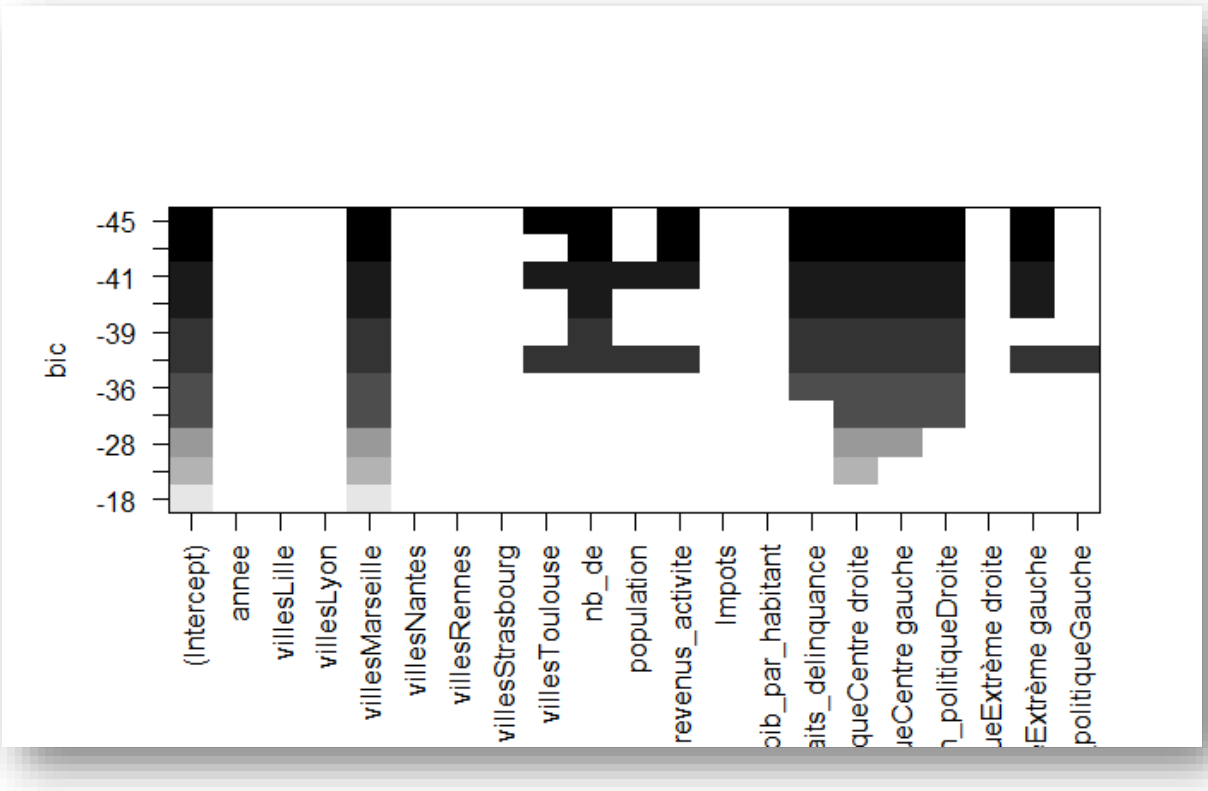


```
Residuals:
    Min      1Q  Median      3Q     Max
-45181 -12920   -845  10249  78456

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.430e+05  3.645e+06   0.149 0.881779
annee                           -3.182e+02  1.855e+03  -0.172 0.864015
villesLille                      5.984e+03  1.847e+04   0.324 0.746348
villesLyon                      -2.996e+03  4.044e+04  -0.074 0.941044
villesMarseille                  1.428e+04  2.872e+04   0.497 0.619793
villesNantes                    -5.199e+03  3.091e+04  -0.168 0.866636
villesRennes                    -3.923e+03  4.321e+04  -0.091 0.927793
villesStrasbourg                -5.199e+03  2.051e+04  -0.254 0.800193
villesToulouse                   2.038e+04  1.648e+04   1.236 0.218217
nb_de                           -9.795e-02  4.172e-02  -2.348 0.020152 *
population                       3.350e-02  6.983e-02   0.480 0.632048
revenus_activite                 9.277e+00  5.951e+00   1.559 0.121043
Impots                          -1.563e+01  2.670e+01  -0.586 0.558982
pib_par_habitant                -1.591e-01  3.203e+00  -0.050 0.960449
orientation_politiqueCentre droite  3.908e+04  7.425e+03   5.263 4.62e-07 ***
orientation_politiqueCentre gauche  2.221e+04  5.317e+03   4.176 4.92e-05 ***
orientation_politiqueDroite      2.049e+04  5.324e+03   3.848 0.000173 ***
orientation_politiqueExtrême droite  2.856e+03  5.400e+03   0.529 0.597663
orientation_politiqueExtrême gauche  1.549e+04  5.369e+03   2.884 0.004480 **
orientation_politiqueGauche      1.127e+04  7.495e+03   1.503 0.134856
nb_faits_delinquance             1.334e-01  3.297e-01   0.405 0.686352
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20110 on 156 degrees of freedom
Multiple R-squared:  0.4396,    Adjusted R-squared:  0.3677
F-statistic: 6.118 on 20 and 156 DF,  p-value: 8.897e-12
```

MSPR3_prédiction.Rmd



Les points où le BIC atteint son minimum représentent le modèle optimal en termes de compromis entre ajustement et complexité. En revanche il faut faire attention à ne pas être en sur-apprentissage et donc avoir un modèle à faible pourvoir prédictif.

B. SVM

(voir data_processing.ipynb)

Dans le cadre de notre étude visant à prédire l'orientation politique à partir de données électorales, nous avons utilisé un modèle SVM (Support Vector Machine) avec un noyau linéaire. Cependant, notre modèle a montré une précision de seulement 26,67%, ce qui est en deçà de nos attentes. Dans cette analyse, nous allons examiner la principale cause possible de ces résultats médiocres.

a) Analyse des résultats

La principale cause de la performance insatisfaisante de notre modèle réside dans le volume de données limité sur lequel le modèle a été formé. Avec un ensemble de données plus petit, le modèle peut avoir du mal à apprendre des relations complexes entre les caractéristiques et l'orientation politique. Le manque de données entraîne une capacité limitée à généraliser à partir de l'ensemble d'entraînement vers de nouvelles données, ce qui nuit à la précision du modèle.

b) Mesures pour améliorer la précision

L'acquisition d'un ensemble de données plus volumineux est impérative pour permettre au modèle d'apprendre plus efficacement les schémas sous-jacents. Un ensemble de données plus étendu offrira au modèle la possibilité de mieux comprendre les nuances de l'orientation politique.

VI. Conclusion

Afin de répondre au mieux à la demande, nous avons voulu nous focaliser sur les dix plus grandes villes de France (hors Paris). Les jointures entre les différents jeux de données nous ont permis d'avoir un dataset final regroupant plusieurs critères pour ces villes ainsi que les années 2002, 2007, 2012, 2017 et 2022. En revanche les jointures ont limité le nombre de ligne de notre dataset final ce qui ne permet pas de faire un modèle avec une bonne prédiction.

Nous avons tout de même constaté que le nombre de demandeurs d'emploi pour une ville ainsi que l'orientation politique du candidat influence le nombre de voix obtenue par le candidat.

Annexe

- https://www.data.gouv.fr/fr/pages/donnees-des-elections/
- https://www.data.gouv.fr/fr/pages/donnees-securite/
- https://www.data.gouv.fr/fr/pages/donnees-emploi/
- https://www.data.gouv.fr/fr/organizations/institut-national-de-la-statistique-et-des-etudes economiques-insee/?datasets_page=7#organization-datasets
- https://www.data.gouv.fr/fr/datasets/bases-statistiques-communale-et-departementale-de-la-delinquance-enregistree-par-la-police-et-la-gendarmerie-nationales/#/resources
- https://www.insee.fr/fr/statistiques/1893204#consulter
- https://www.insee.fr/fr/statistiques/5020211
- https://statistiques.pole-emploi.org/stmt/defm?qp=1&fi=84,53,44,32,75,76,52,93&li=0&mm=0&pp=200201-202201&ss=1
- https://www.data.gouv.fr/fr/pages/donnees-des-elections/