README MSPR

bureau de vote et communes (44000)

communes -> revenues moyens des gens

définir les critères : on peut penser a l'immigration(pas important) , score gauche / droite

pertinence des jeux de données

sujets des médias

on part sur l'utilisation des données des communes

sur une communes juste

prédire un score sur une idée c'est compliqué

Est-ce qu'il y a des biais par rapport aux évolutions d'idées?

data gouv fr

on choisit ses jeux de données 🆗

sa zone géographique

qu'est-ce qu'on va prédire?

# Les choix

* On va prédire dans la zone géographique de Montpellier.
* On va prédire quel est le bord politique le plus susceptible de l’emporter aux prochaines élections municipales.
* Pour cela on va faire un un échelle qui servira de score pour déterminer le bord politique.
* Le modèle s'arrêtera à prédire si la droite ou la gauche à plus de chances de l’emporter aux prochaines élections. (Limitation des sorties du modèles et approximations sinon on s’en sort pas).
* Pour cela on pourra utiliser **dans un premier temps** les données des élections passées sur Montpellier. Puis **dans un second temps** les données suivantes, sans s’y limiter, pour tenter de trouver des tendances (liste non exhaustive et on est pas obligé de tous les prendre) :
  + Insécurité taux de criminalité
  + Revenu moyen
  + Revenu médian
  + Bonheur des habitants
  + Nombres de fonctionnaires dans les établissements publics
  + Taux d’éducation moyen (via diplômes)
  + Nombre d’entreprises sur le territoire
  + Nombres de demandeurs d’emplois
  + Densité de population
  + PIB/habitant de MTP
* ETL via spark ? -> à def.
* IA via scikit-learn ? -> à def.
* Une près et un rapport avec nos beaux graphiques et explications !

# Nouvelles notes - 22 mars 2024

Reporting ?

Dashboard via power bi?

Le rendu final? Qu'est ce qui rapporte des points ?? sur quoi on est noté ?

Les responsabilités ? Qui fait quoi dans l’équipe ?

Fabien : Documentation / Rapport

Mathis :

Fita :

Lelio : Dapo et structure de la présentation ? Méthode agile.

Sachant que tout le monde doit pouvoir parler aisément de l’ensemble du sujet, donc des parties des autres.

Les biais et limites :

* Les données ne sont pas significatives : cela fait à peine 16 élections entre 1958 et 2022, comment prétendre prédire de manière fiable avec un échantillon aussi restreint ?

On peut voir sur data.gouv.fr que le nombre d'élections législatives est supérieur au nombre des autres types d'élections disponible sur la plateforme. C’est donc un très bon choix.

Les élections municipales sont malheureusement trop peu nombreuses sur data.gouv.fr. Les élections présidentielles suivent de près les législatives en termes de quantité des données.

Les élections présidentielles ont l'avantage qu’on à pas à choisir de position géographiques.

Utiliser plusieurs algos et les comparer/critiquer.

Documenter la phase d’exploration des données.

Ne pas oublier de parler de méthode agile.

Minimum 20 pages de rapport.

# Premier approche du dataset, exploration des données

Le jeu de données au format CSV fait 12,9 Mo ou 13 549 861 octets. Le délimiteur utilisé est un point-virgule. Le fichier fait 66194 lignes. Le fichier est encodé au format Western European (Windows).

## Conversion en utf-8

Cela augmente légèrement la taille du dataset mais permet une plus large compatibilité.

En utilisant “head -n ./fichier” : on peut voir que le fichier contient les colonnes suivantes :

numéro de l’élection, nom de l’élection, Code Commune, Commune, Numéro du bureau de vote, Nom du bureau de vote, Circonscription, Canton, Inscrits, Votants, Nuls, Blancs, Exprimés, Procurations, Nombre de candidat, nom du Candidat et nombre de voix.

On peut donc directement constater qu’il manque une colonne pour l’année de l’élection.

Dans un premier lieu, on peut voir que les types d’élections sont mélangés dans le fichier. On y trouve :

* Des referundum
* Des élections du parlement européen

# Les sources de données

Dataset : <https://data.montpellier3m.fr/dataset/resultats-des-elections-montpellier>

# Les biais

On ne trouves pas de données pour le sexe et l’age des votants, logique, il ne sont renségné quand dans des sondages et on a fait le choix de ne prendre que des données factuelles.