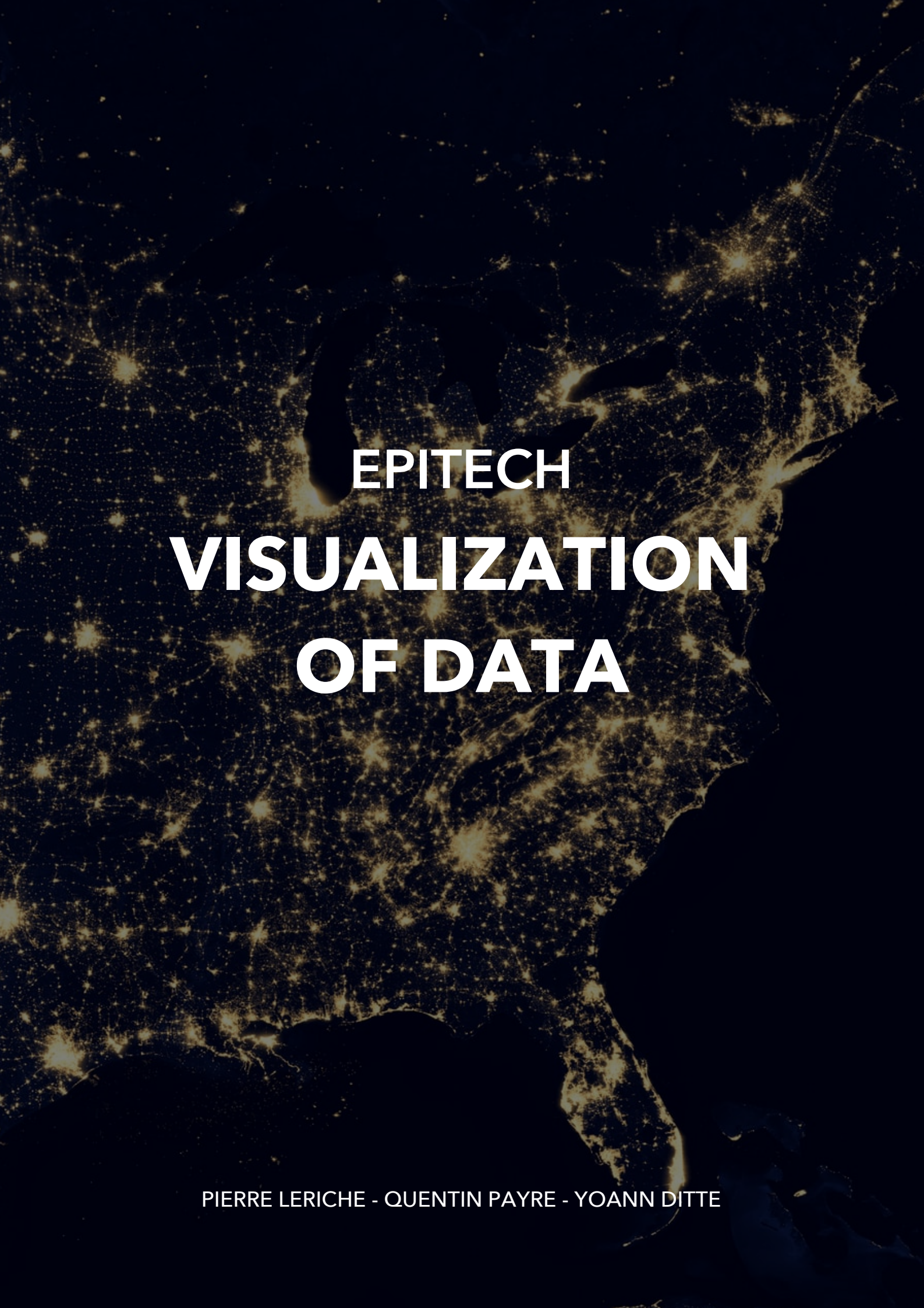

EPITECH

VISUALIZATION OF DATA

PIERRE LERICHE - QUENTIN PAYRE - YOANN DITTE

Jeu de données

Pour réaliser ce projet de visualisation de données, nous avons sélectionné le jeu de données « Boston Housing ». Ce dataset est constitué de 505 lignes et de 14 attributs, dont 1 catégorique. Les attributs de ce dataset constituent les situations géographiques des maisons.

Pour nous c'est un dataset intéressant et pratique à utiliser car nous pouvons facilement nous représenter les données incluses. Cela nous permettra lors des phases de visualisation et de prédiction de mieux interpréter les informations et de mieux comprendre les résultats que nous obtenons.

Notre objectif sur ce projet sera donc d'explorer le jeu de données, découvrir s'il existe des liens au sein des attributs, vérifier nos hypothèses et enfin prédire la valeur d'une maison en fonction des attributs les plus impactant.

Les attributs du dataset qualifient la population, les maisons et l'environnement de la maison. Cela va du ratio professeur-élèves jusqu'à la qualité de l'air en passant par le nombre de pièces dans la maison. L'attribut MEDV, qui correspond à la valeur de la maison en millier de dollars, est celui que l'on va chercher à prédire.

En regardant simplement les descriptions des attributs (disponibles dans le README) on peut d'ores et déjà émettre des hypothèses sur les attributs les plus importants pour notre prédiction. En effet, il serait logique que le nombre de pièces (RM) dans la maison influe sur la valeur de celle-ci. Aussi, le ratio de crime (CRIM) dans la ville va surement impacter la valeur du bien en négatif. On peut aussi penser que la pollution de l'air (NOX) par les oxides d'azote réduit la valeur de la maison, potentiellement à cause des industries polluantes proches (INDUS).

Visualisation

Dans cette étape du projet, nous allons créer 2 moyens de visualisation des données du dataset.

La première sera une matrice de graphiques de points. Elle aura notamment pour objectif de mettre en valeur les liens et les structures simples qu'il peut y avoir dans chaque pair d'attributs.

La seconde sera un graphique de coordonnées parallèles. Cette méthode de visualisation nous permettra entre autres de sélectionner plusieurs fourchettes de valeurs dans les attributs et d'éliminer les attributs qui sont présents partout.

Matrices de points

Les attributs les plus parlant sont RM (nombre de pièces) et LSTAT. (% de population au statut social faible). Plus RM est élevé, plus MEDV l'est aussi, à l'inverse un LSTAT haut montre un MEDV bas.

Voici les graphiques pour RM et LSTAT. Les points rose représentent un MEDV élevé :

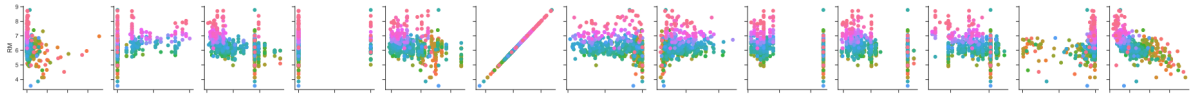

Figure 1: RM graphic line

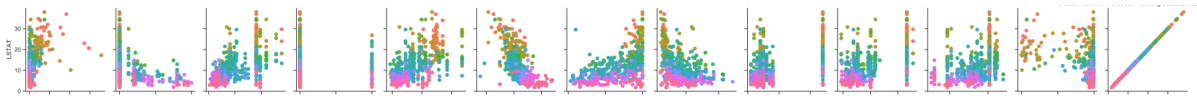

Figure 2: LSTAT ligne de graphiques

Sur le graphique LSTAT/RM nous voyons bien une majorité de points rose avec un RM élevé et un LSTAT faible.

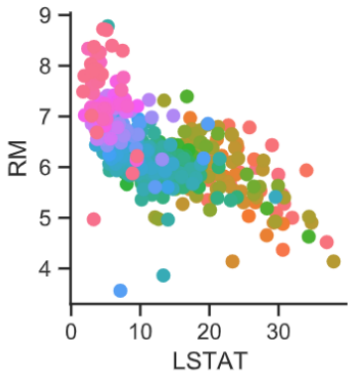

Figure 3: LSTAT/RM ligne de graphiques

Coordonnées parallèles

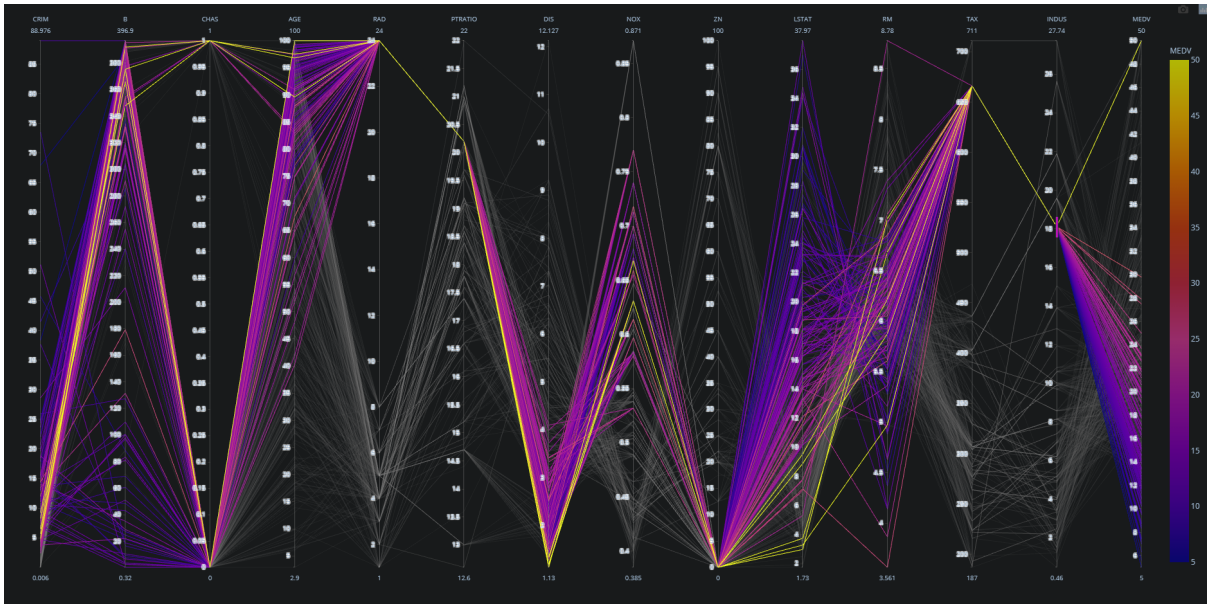

Figure 4: Coordonnées parallèles avec sélection

Grace au rendu web, nous pouvons manipuler les axes pour mettre en avant certaines valeurs sur certains axes. En cherchant des liens, nous avons remarqué que pour 2 valeurs, la valeur 18 d'INDUS (Proportion des activités hors commerce de détail) et 650 pour l'attribue TAX (taux de l'impôt foncier sur la valeur totale par 10 000$), les valeurs de MDEV restent très faibles.

Analyse quantitative

Cette dernière étape du projet a pour but de créer un modèle de prédiction de la valeur d'une maison en fonction de plusieurs attributs. Nous allons donc sélectionner les attributs puis appliquer un modèle de régression linéaire pour la prédiction.

Sélection des attributs

Pour sélectionner les attributs les plus pertinents, nous nous sommes basés sur nos précédentes observations et nous avons construit une grille de corrélation. L'objectif était de faire ressortir les attributs qui étaient le plus corrélé à MEDV, l'attribut à prédire.

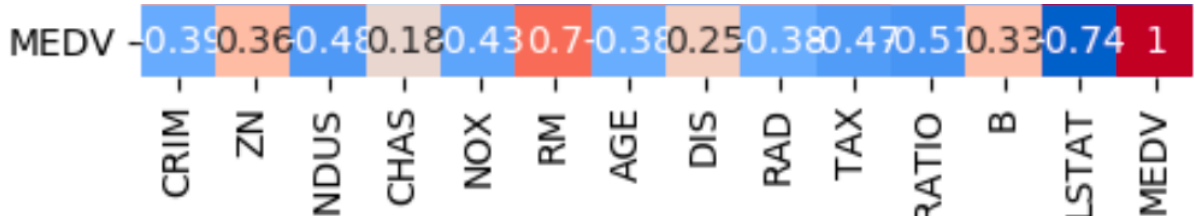

Figure 5: Corrélation vs MEDV

Sur l'image ci-dessus, on peut clairement voir les corrélations grâce aux couleurs. Il ressort que les attributs les plus corrélées (valeur tend vers 1, couleur rouge foncé) ou inversement corrélées (valeur tend vers -1, couleur bleu foncé) sont RM et LSTAT. Cela confirme les hypothèses émisent à l'exploration du dataset pour RM, et celles misent en avant lors des visualisations pour les 2. Sur la carte de corrélation, le 3^{e} attribut que nous sélectionnons est le PRATIO. Ensuite la corrélation s'estompe.

Régression linéaire

Enfin, nous avons utilisé un modèle de régression linéaire dans le but de prédire la valeur d'une maison. Nous avons uniquement sélectionné les attributs les plus pertinents pour nous : ceux qui sont ressorties lors des étapes précédentes. Il s'agit donc de RM, LSTAT, et PRATIO.

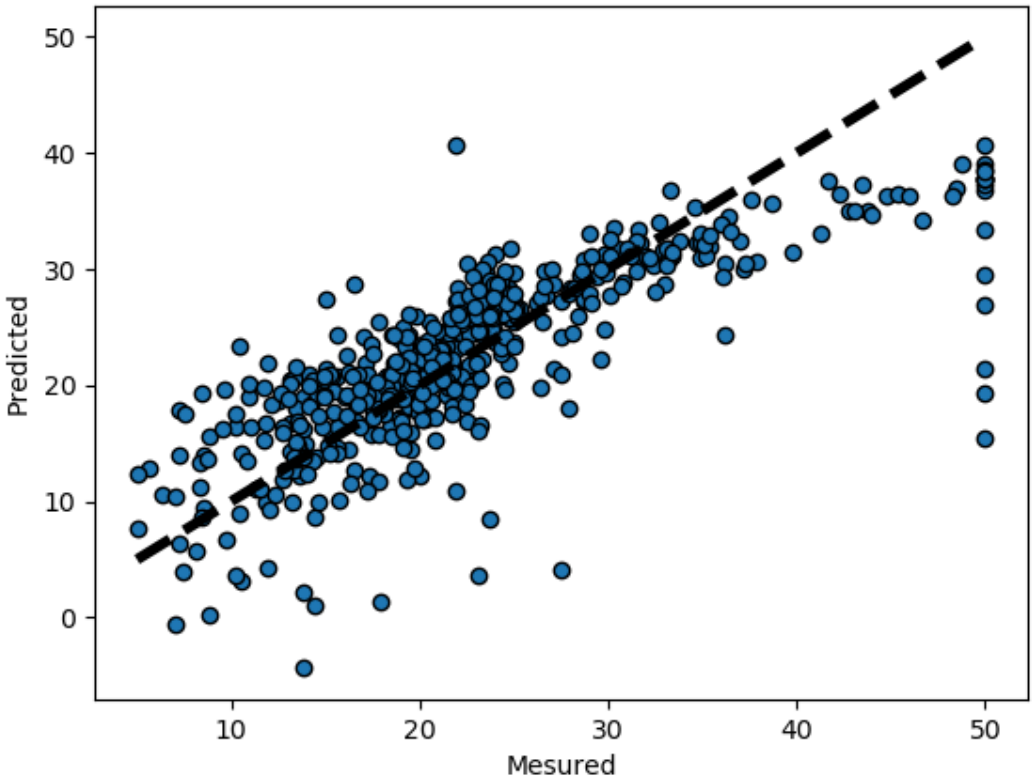

Figure 6: Linear Regression output

À propos des résultats de la prédiction, on peut voir sur la figure 6 qu'ils semblent être relativement bons, a ceci près que les hautes valeurs de MEDV (50) sont effectivement prédites comme plus basses.

```
The model performance for testing set
--------------------------------------
RMSE is 6.455849295190898
R2 score is 0.4881642015692508
```

Figure 7: Qualité du model avec 3 attributs

Le RMSE, qui permet d'évaluer la marge d'erreur moyenne en amplifiant les plus grosses nous donne 6.4. Cela donne une prédiction a notre avis de qualité malgré les valeurs hautes de MEDV qui sont mal gérés.

On peut d'ailleurs remarquer que ces résultats sont obtenus avec seulement 3 attributs, et lorsque nous utilisons la totalité des attributs, la marge d'erreur ne diminue pas énormément.

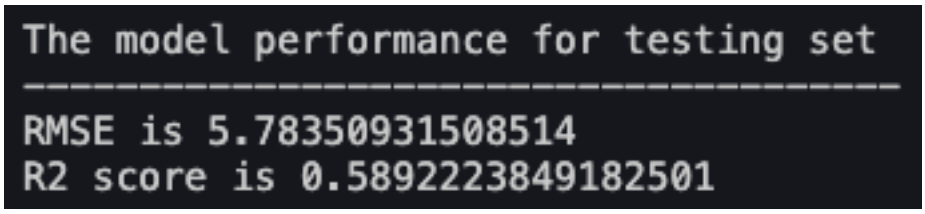

Figure 8: Qualité du model avec tous les attributs