

SOUTENANCE PROJET

BIG DATA

Maxime Feuillet - Paul Lemarquand - Pierre
Breganc -Aymeric Faure



SOMMAIRE

1. Présentation du sujet
2. Les différentes étapes du projet
3. Démonstration
4. Les axes d'amélioration

1. PRÉSENTATION DU SUJET

ANNONCE DE LOGEMENTS À LOUER SUR AIRBNB À BORDEAUX

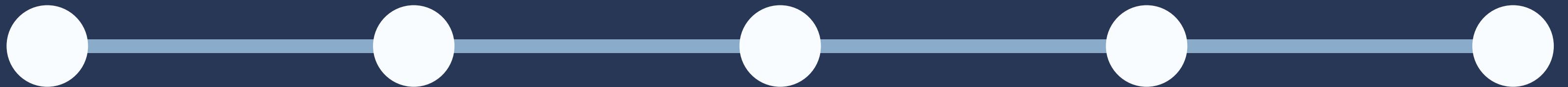


Proposer une analyse de données descriptive afin de les synthétiser

Proposer un modèle de prédiction du tarif nuitée



2. LES DIFFÉRENTES ÉTAPES DU PROJET



VISUALISER

Installer les données sur une VM et les visualiser
Rapatrier les données en local

PARTAGER

Pousser les données dans le cloud et les sécuriser

ENTRAÎNER UN MODÈLE

Sélectionner un modèle et l'entraîner
Avoir un modèle cohérent et efficace

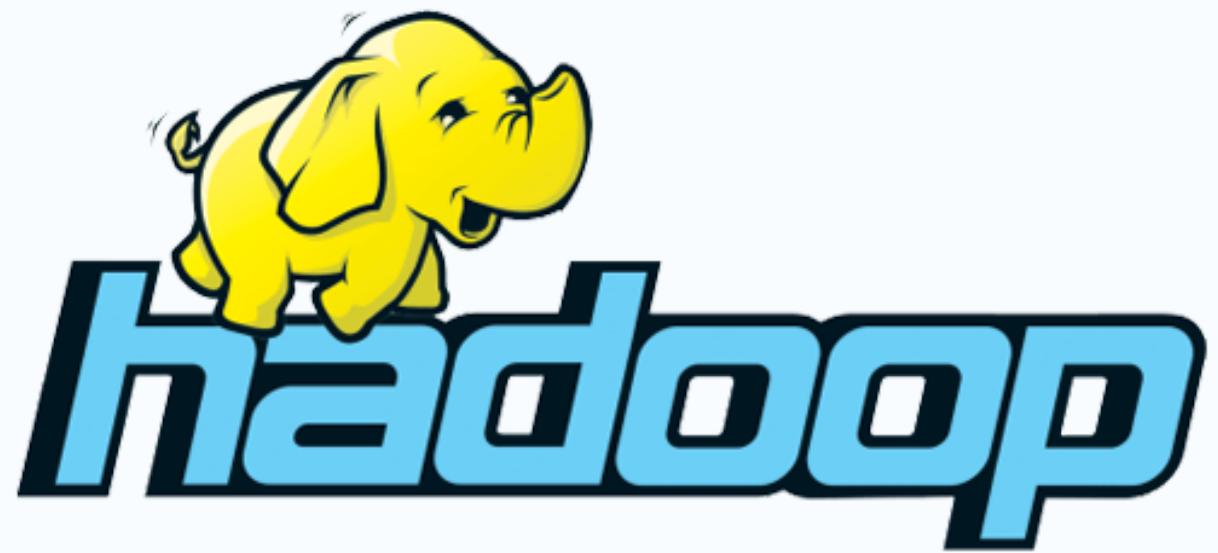
PRÉDIRE

Utiliser un modèle pour faire de prédictions du tarif nuitée

ANALYSER

Visualiser et analyser les données
Utiliser une base NoSQL et un outil de DataViz pour faire une analyse éclairée des données

ETAPES 0 & 1



VISUALISATION DES DONNÉES

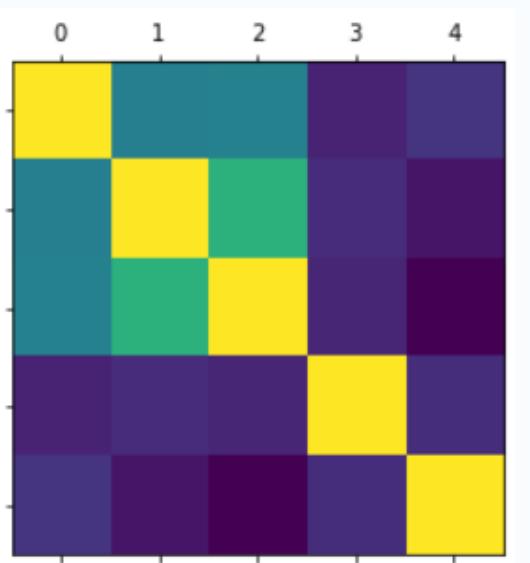
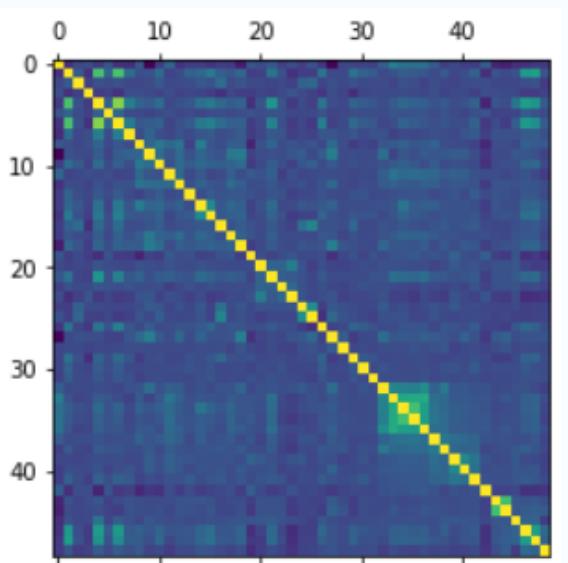
ETAPE 0

Installation des données sur HDFS, sur une VM
Hadoop téléchargée et installée en local et
première visualisation

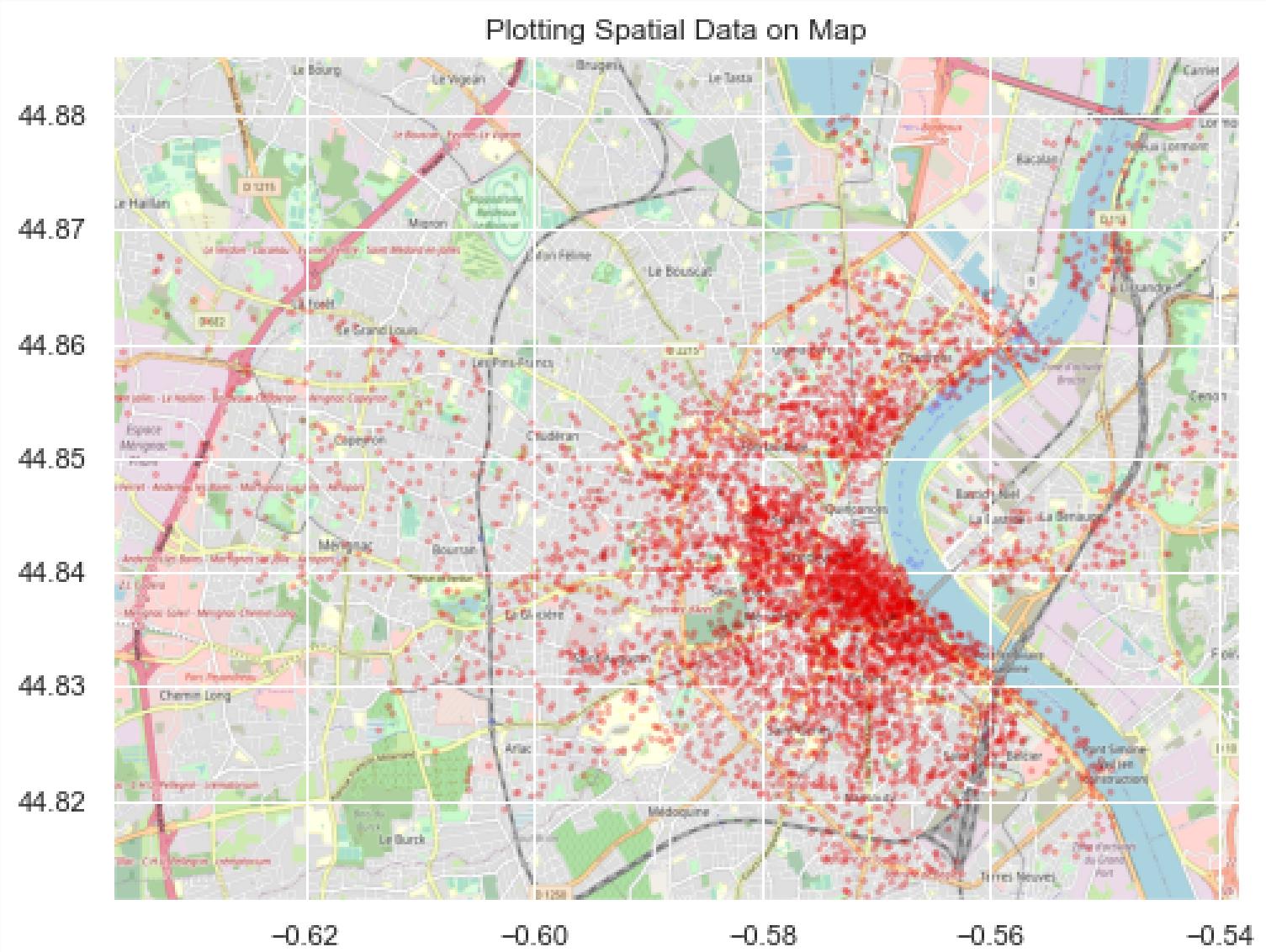
ETAPE 1

Script rapatriant les données en local, qui les
récupère depuis HDFS

Visualisation des données



	PrixNuitee	Capacite_accueil	NbChambres	frais_menage	Caution
PrixNuitee	1.000000	0.629278	0.632813	0.412444	0.444072
Capacite_accueil	0.629278	1.000000	0.766409	0.430599	0.383021
NbChambres	0.632813	0.766409	1.000000	0.417004	0.346687
frais_menage	0.412444	0.430599	0.417004	1.000000	0.430968
Caution	0.444072	0.383021	0.346687	0.430968	1.000000



ETAPE 2



PARTAGE DES DONNÉES ET SÉCURISATION

AMAZON WEB SERVICES

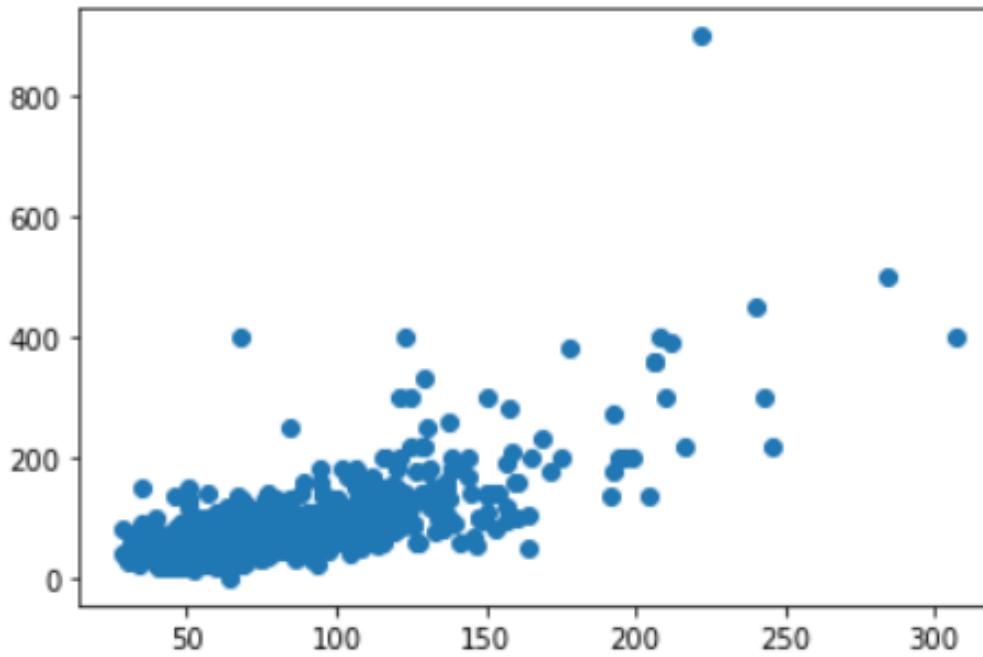
Données poussées sur une VM dans le cloud AWS
Script lancé avec un cron

SECURITÉ DES DONNÉES

Réalisation d'une connexion SSH entre la VM AWS et nos données en local
Scripts Python permettant un chiffrement symétrique

ETAPE 3

```
Erreur moyenne : 24.006438367346032  
Erreur medianne : 16.906689725355136  
Erreur min : [0.01147439]  
Erreur max : [678.27941942]
```



Paramètres utilisées : Capacité d'accueil, Nb de chambres, frais ménage, Caution et Position (latitude et longitude)

Essai de différentes méthodes (k-ppv, Linear regression, Ridge regression ...) avec différents hyper paramètres

Meilleurs modèle : Ridge regression avec alpha = 0.54
Résultats malgré tout moyennement concluant

Métrique d'erreur :
Erreur moyenne et médiane

ETAPE 4



PREDIRE

Séparation du set de données en 2 parties : 80% pour l'entraînement, 20% pour les tests

Exportation des lignes correspondants aux tests sous format CSV

Les colonnes du CSV sont celles du dataset moins les deux colonnes prix nuitée plus une colonne prix estimé



ETAPE 5

ANALYSE DES DONNÉES



PYMONGO

Les prédictions sont poussées dans une base NoSQL

DATA VISUALISATION

Représentation visuelle des résultats via MongoCharts

DÉMONSTRATION

4. AXES D'AMÉLIORATION

Setup l'environnement de la VM et création dynamique (utilisation de boto3)

Utilisation des bag of words

Le modèle pourrait être rendu plus performant, notamment en séparant les données extremes

Un modèle de Machine Learning serait sans doute plus efficace

Merci pour votre attention