

# Projets 4: Data & Machine Learning

## ADN Tourisme

Narciso Alves, Bedi Ogur, Jin Xu



# CONTEXTE DU PROJET

## Problématique

- ADN tourisme regroupe les fédérations des acteurs institutionnels du tourisme.
- Leur base de données réunit les informations des points d'intérêt touristique en France.

⇒ Objectif: Améliorer la qualité de la base de données

⇒ Livrable: Tableau de bord, informations et cartographie des établissements, filtrées par catégorie.

# PLAN

1. Collecte de données (4 semaines)
2. Sélection des caractéristiques utiles, gérer le volume de données (1 semaine)
3. Analyse exploratoire des données (2 semaines)
4. Machine learning: classification d'établissements (1 semaine)
5. Tableau de bord (1 semaine)

# 1. Collecte de données

Source de data: <https://www.datatourisme.fr/>

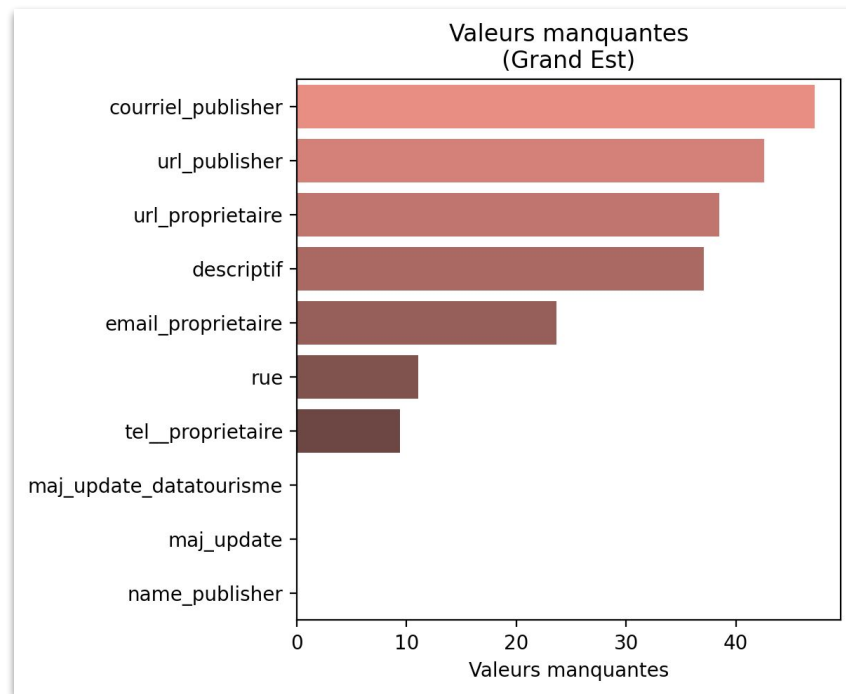
- 1 fichier flux-complet.zip qui contient 412621 fichiers Json
- 1 établissement ( 1 fichier Json par établissements)
- 412621 établissements (POI)
- Le fichier représente environ 10Go



## 2. Sélection des caractéristiques

21 caractéristiques sélectionnées dans le data frame pour chaque établissement. Le volume de données représente 230 Mo :

0	identifiant	11	region
1	nom	12	url_proprietaire
2	type_etablissement	13	email_proprietaire
3	lat	14	tel__proprietaire
4	lng	15	url_etablissement
5	code_insee	16	name_publisher
6	descriptif	17	courriel_publisher
7	rue	18	url_publisher
8	localite	19	maj_update
9	postal_code	20	maj_update_datatourisme
10	departement		



### 3. Analyse exploratoire des données

- **412 617** événements et points d'intérêt recensés
- **14** régions
- **98** départements

#### Statistiques

POI :

412617

Regions couvertes :

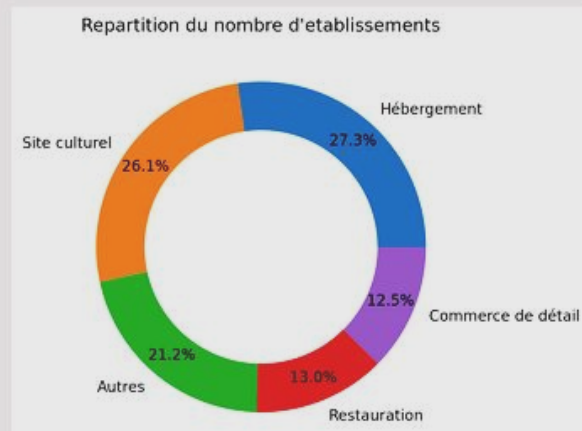
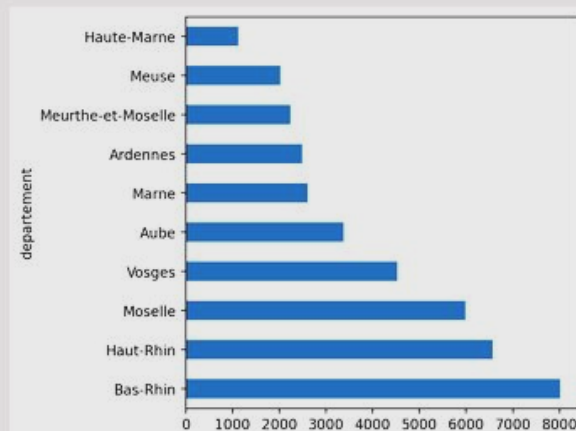
14

Département couverts :

98

POI dans le département :

6260



## 4. Machine Learning

- En fonction du descriptif de l'établissement, on prédit à quelle catégorie appartient l'établissement: restauration, hébergement, etc.
- Le modèle est basé sur une classification de régression logistique avec une précision de 97%
- Cela peut permettre de détecter si une catégorie est mal renseignée en fonction du descriptif de l'établissement.

```
modelLR.predict(tfidf.transform(pd.Series('Je vais manger chez lucette')))  
array(['Restauration'], dtype=object)
```



## Hébergement

## 5. Tableau de bord (Demo)



### Filtrer par Région et Département

Sélectionner une Région

Provence-Alpes-Côte d'Azur

Sélectionner un Département

Var

Sélectionner le type d'établissement

## Bienvenue chez ADN tourisme



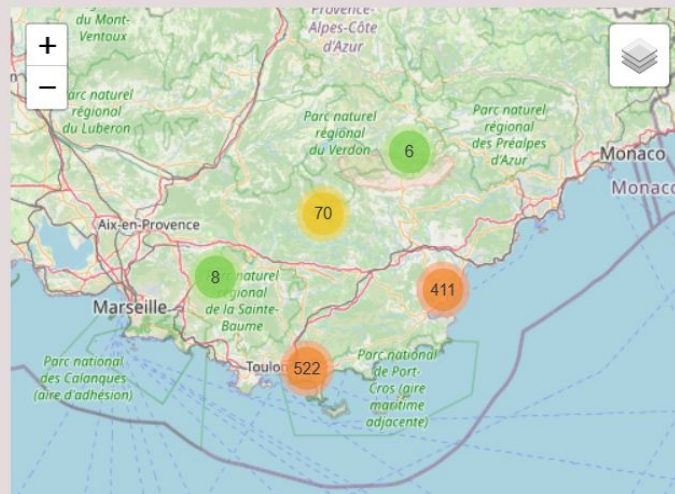
Accueil

Détail

Statistique

Robot ML


### Cartographie des POI





# SYNTHÈSE

⇒ Amélioration de la qualité des données.

- Tableau de bord:
  - informations pour chaque établissement
  - cartographie des POI
  - visualisation d'indicateurs: taux de valeurs non renseigné par exemple
- Proposition de prédiction grâce au machine learning 



# PERSPECTIVES

- Faciliter la collecte des données
- Indiquer aux propriétaires les champs essentiels à remplir
- Mise à jour automatique de la base de données
- Etendre le développement du tableau de bord à un format web (type HTML Java Script)

