

Nanterre p10
- Dev Data

SIMPLON
.CO

Chef d'oeuvre

d'Alain Juste NGABO

Certifications “Développer une base de données” (CNCP 3497)
et “Exploiter une base de données” (CNCP 3508)



02

AGENDA



Mon parcours La formation Dev Data Simplon

Le projet:

- Présentation
- Contexte
- Enjeux et objectifs
- Lexique des termes clés
- Planification et gestion
- Technologies utilisées
- Architecture générale
- Présentation technique
- Démo

Difficultés rencontrées

Quelle suite pour ce projet ?

Mon parcours

03



Master II en communication et Marketing – Lille, 2014

Master II Droit , Economie et Gestion, spécialité Action Publique et Relations Internationales – Paris XII, 2015



5 ans d'expérience professionnelle dans les métiers de la communication et du marketing. #secteur: Organisations internationales, entreprises et administrations publiques



Passionné du digital et de la data, choix d'une reconversion dans ce domaine en 2020, chez Simplon.co



Centre de formation:

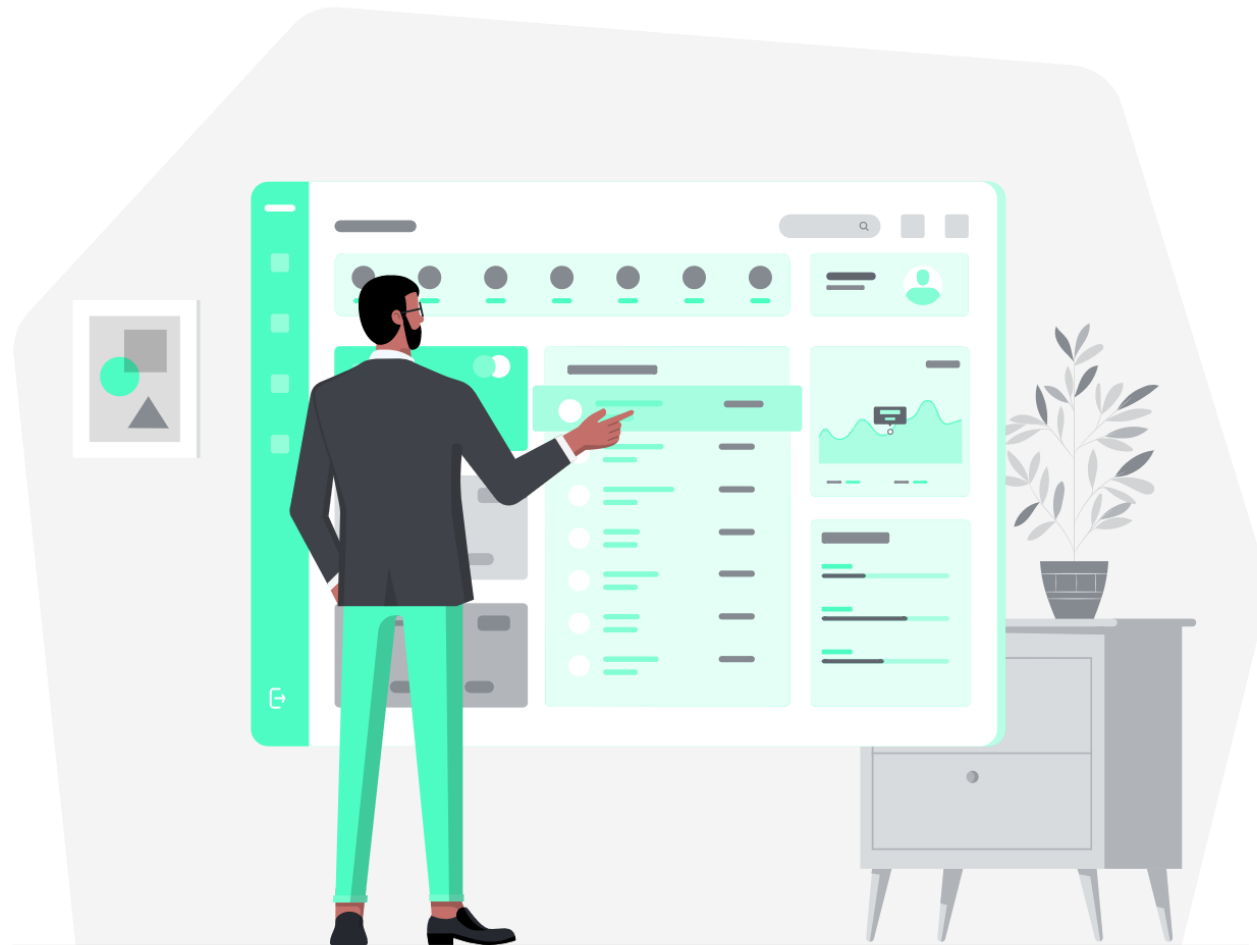
- ✓ Forme aux métiers du numérique
- ✓ Pédagogie active
- ✓ Labellisée Grande Ecole du Numérique

Formation : Développeur(se) data

- ✓ Développer une base de données
- ✓ Exploiter une base de données

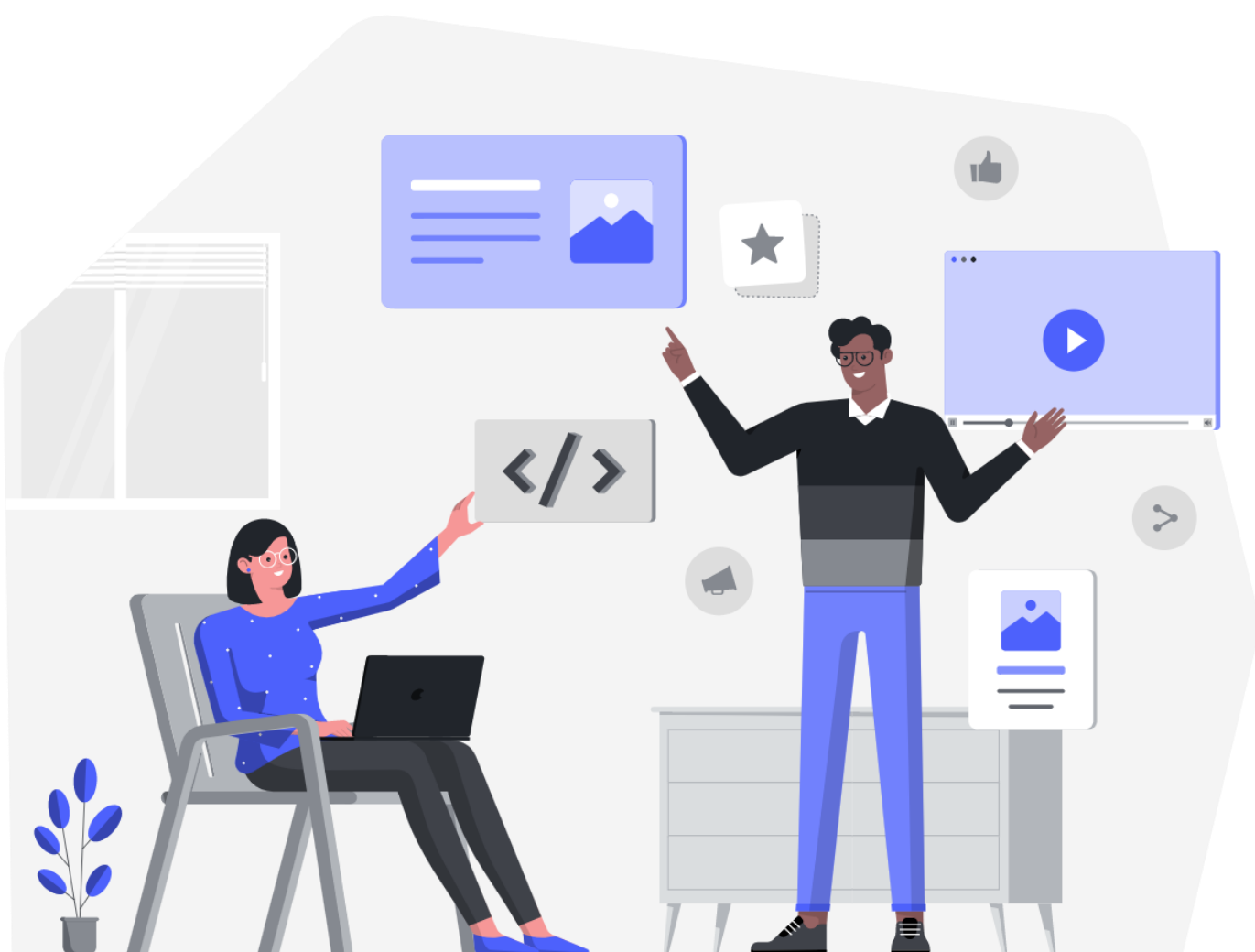
LE PROJET

05



Création d'une base de données et d'un dashboard sur la fréquentation touristique hôtelière en France de 2011 à 2020.

06



PRESENTATION DU PROJET

Ce projet comporte deux grandes parties :

**Une base de données relationnelle
conçue sous Mysql Workbench**

**Pour stocker de manière structurée l'ensemble des
informations.**

**Une datavisualisation réalisée via Microsoft
PowerBi**

**pour afficher et partager les résultats des analyses et autres
KPIs.**

07

ENJEUX ET OBJECTIFS DU PROJET



Enjeux

Mieux appréhender le tourisme en France centrée sur le comportement des clientèles touristiques à partir de l'analyse de la fréquentation des hotels.

Objectifs

L'objectif est d'observer en fonction du lieu et la périodicité:

- le volume de touristes accueillis,
- leur origine (clientèle Française ou étrangère)
- les pays de provenance
- la durée du séjour
- le taux d'occupation des hotels

Ceci sur la base des arrivées et nuitées recensées par l'INSEE auprès des hotels.

08

ENJEUX ET OBJECTIFS DU PROJET (suite)

Intérêt de ce travail

2011: Reflexion gouvernementale sur la mise en place d'une plateforme data sur le tourisme.

La France est le pays le plus visité au monde et la consommation touristique intérieure a atteint 7,4% du PIB en 2018.

En 2017, lancement de la plateforme datatourisme, mais les données ne sont pas toujours complètes. L'INSEE demeure la plateforme la plus fiable mais les données ne sont pas nettoyées et exploitables par tout le monde.

De manière aboutie, la base de données et le dashboard pourraient être une importante source d'informations pour :

- Les administrations en charge de l'attractivité territoriale,
- Les entreprises spécialisées,
- Les journalistes et toutes personnes intéressées par ce sujet





Lexique des termes clés du projet

09

Taux d'occupation : rapport entre le nombre de chambres occupés et le nombre de chambres offerts par les hotels.

$$TO \text{ (en \%)} = \frac{\text{Chambres vendues}}{\text{Capacité construite}} * 100$$

Arrivées / Nuitées

- On compte une arrivée dès qu'une nouvelle personne arrive dans un hôtel le mois concerné.
- On compte une nuitée par nuit et par personne passée dans un hôtel durant le mois concerné.

Exemple:

un couple arrive et séjourne 3 nuits dans le mois ; il faut compter 2 arrivées dans le mois et 6 nuitées (2 personnes x 3 nuits).

Une personne arrive le 25 janvier et séjourne 10 nuits ; il faut compter 1 arrivée et 7 nuitées en janvier et 0 arrivée et 3 nuitées au mois de février.

10

Planification et gestion

Le projet a été géré en mode agile avec une planification sur l'outil Trello. On y trouve les tâches, mais aussi les compétences correspondantes.

Lien d'accès du Trello : <https://trello.com/b/7U8yePl2/chef-doeuvredata-analyst>

The screenshot displays a Trello board for a project named "Chef d'oeuvre#DATA ANALYST". The board is organized into several columns representing different stages of the project workflow:

- Compétences**: Lists six tasks (C1 to C6) related to data analysis, each with a progress bar. C1: Concevoir et structurer physiquement une base de données relationnelle ou non, à partir des besoins, contraintes et données du commanditaire. C2: Acquérir des données, les combiner et les structurer en données propres en vue de leur intégration dans la structure de la base de données. C3: Intégrer des données propres et préparées dans la base de données finale, en utilisant des langages informatiques, logiciels ou outils. C4: Optimiser une base de données afin d'en maintenir la fiabilité et la qualité des données. Nettoyer et améliorer les performances. C5: Interroger la base de donnée afin de mettre à jour les données (brutes ou traitées) stockées, provisoirement ou durablement, en fonction du résultat recherché. C6: Concevoir et réaliser un rendu visuel des données issues du
- À faire**: Contains tasks like "Github", "Compilation recherches internet", "Flask ou html ?", and "diagramme de gantt et prévisionnel".
- En cours**: Contains tasks like "Refaire le power point présentation", "Publication en ligne", "Arborescence générale", and "indiquer la création des hierarchies".
- Terminé**: Contains tasks like "Terminé", "Création environnement", "[Tâche terminée]", "Création d'un compte sur datatourisme", "1er Point de suivi - Josselin - Manel", "Choix du sujet - Lecture données sur les statistiques du tourisme en France", "Télécharger CSV Frequentation touristique sur INSEE", "télécharger CSV villes et départements France", and "Créer MCD".
- Ressources**: Contains tasks like "Idéation", "Nuitées en hébergement touristique en France", "prise de notes", "explications insee", "scrap données meteo", "idees analyse chiffres notamment finances", "Aide problèmes rencontrées", "pib tourisme france", "idées statistiques", and "Présentation orale du projet".
- Abandon**: Contains tasks like "Test", "Télécharger CSV Latitude Longitude Départements France", and "PIB Tourisme France".
- Blocage**: Contains a task "Ajouter une carte".

Stack technique

11

Développer une BDD

Sourcing des données
Fichiers CSV

Traitement des données
Python
IDE: Jupyter notebook

Modélisation BDD
Looping

**Création et stockage
BDD**
SQL, My SQL Workbench

Exploiter une BDD

Traitement des données
Mysql,Python

IDE: Jupyter notebook

Connection
Pymysql,
Connector/Net,ODBC

Visualisation
Power BI Desktop, Power
BI SERVICE

Gestion de projet

Suivi de projet
Trello

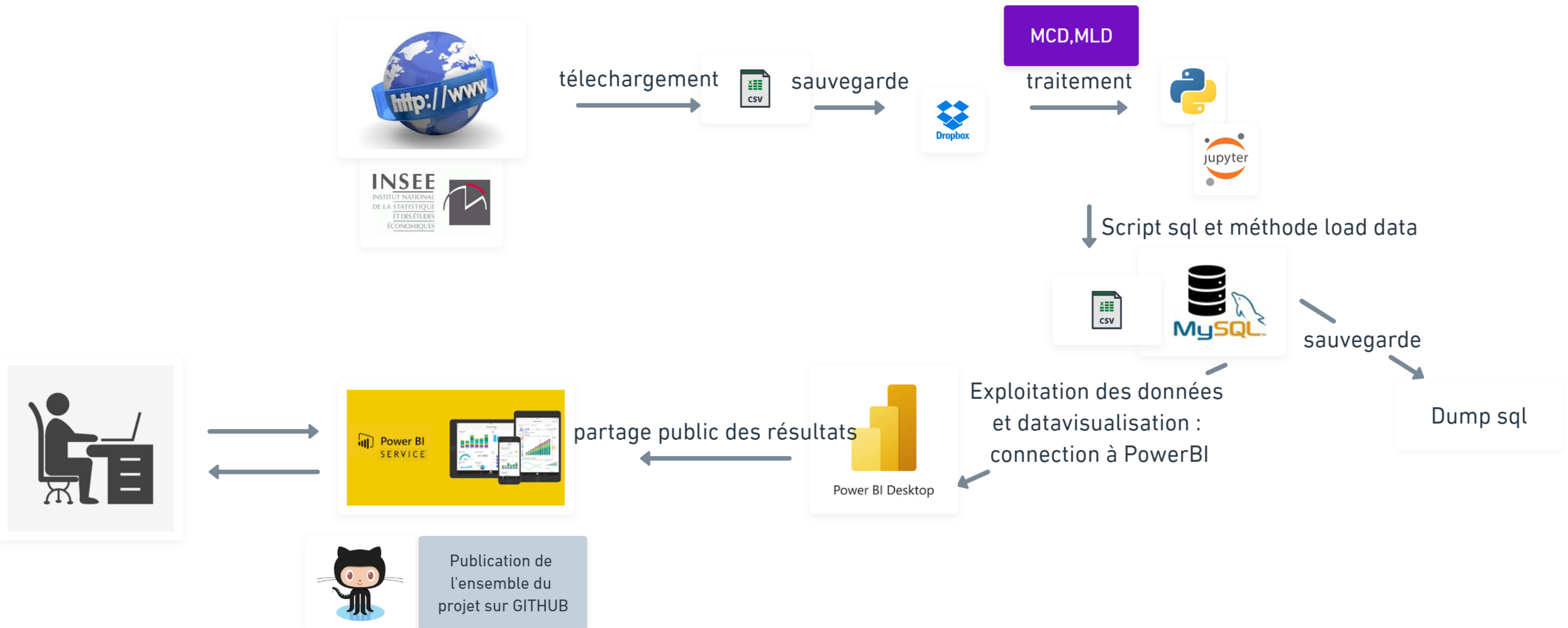
Idéation
Whimsical

Sauvegarde
GITHUB,Dropbox

Problématiques
Doc technique,
stackoverflow,Power BI community

Architecture générale

12

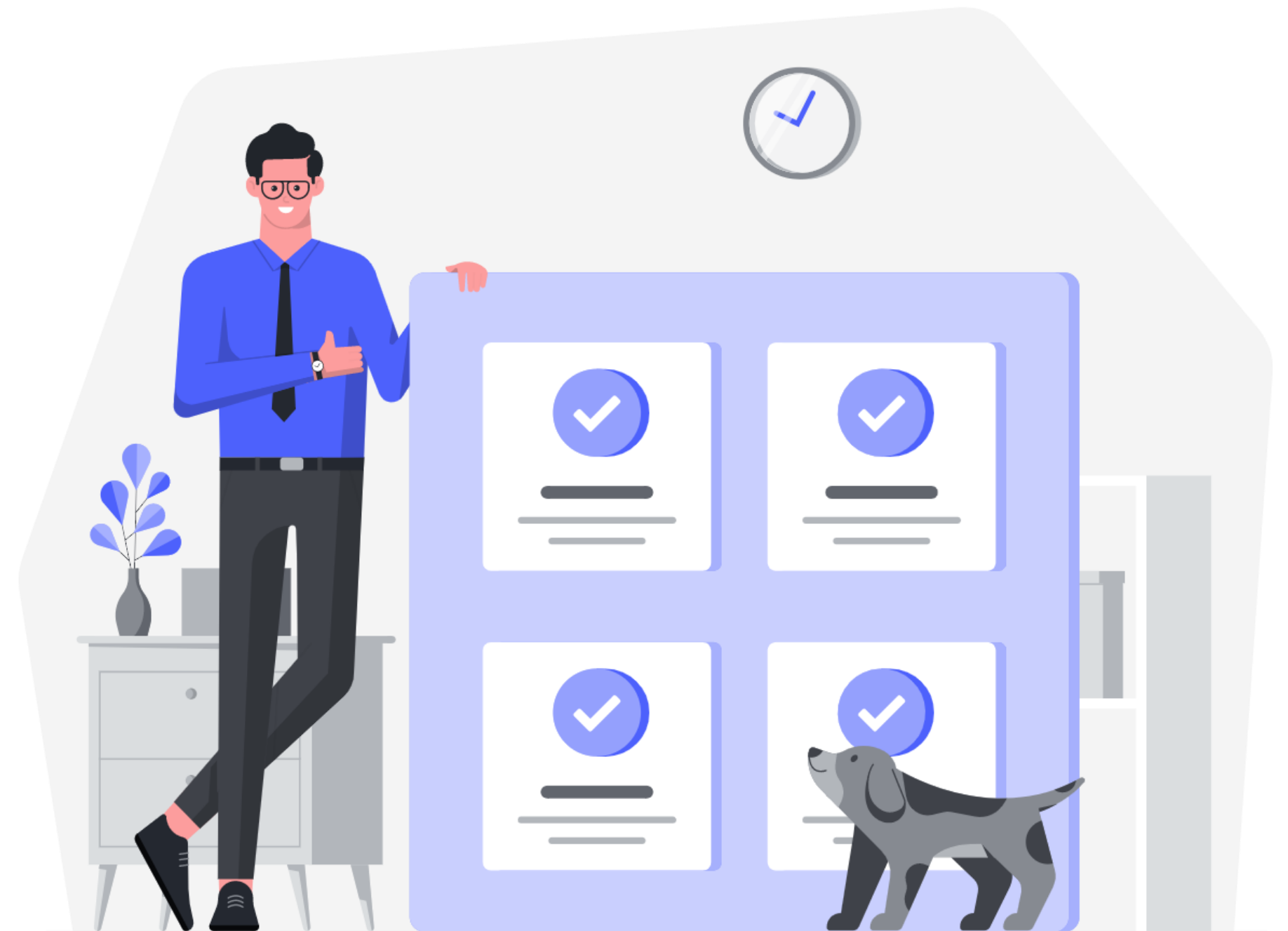


PRÉSENTATION TECHNIQUE

13

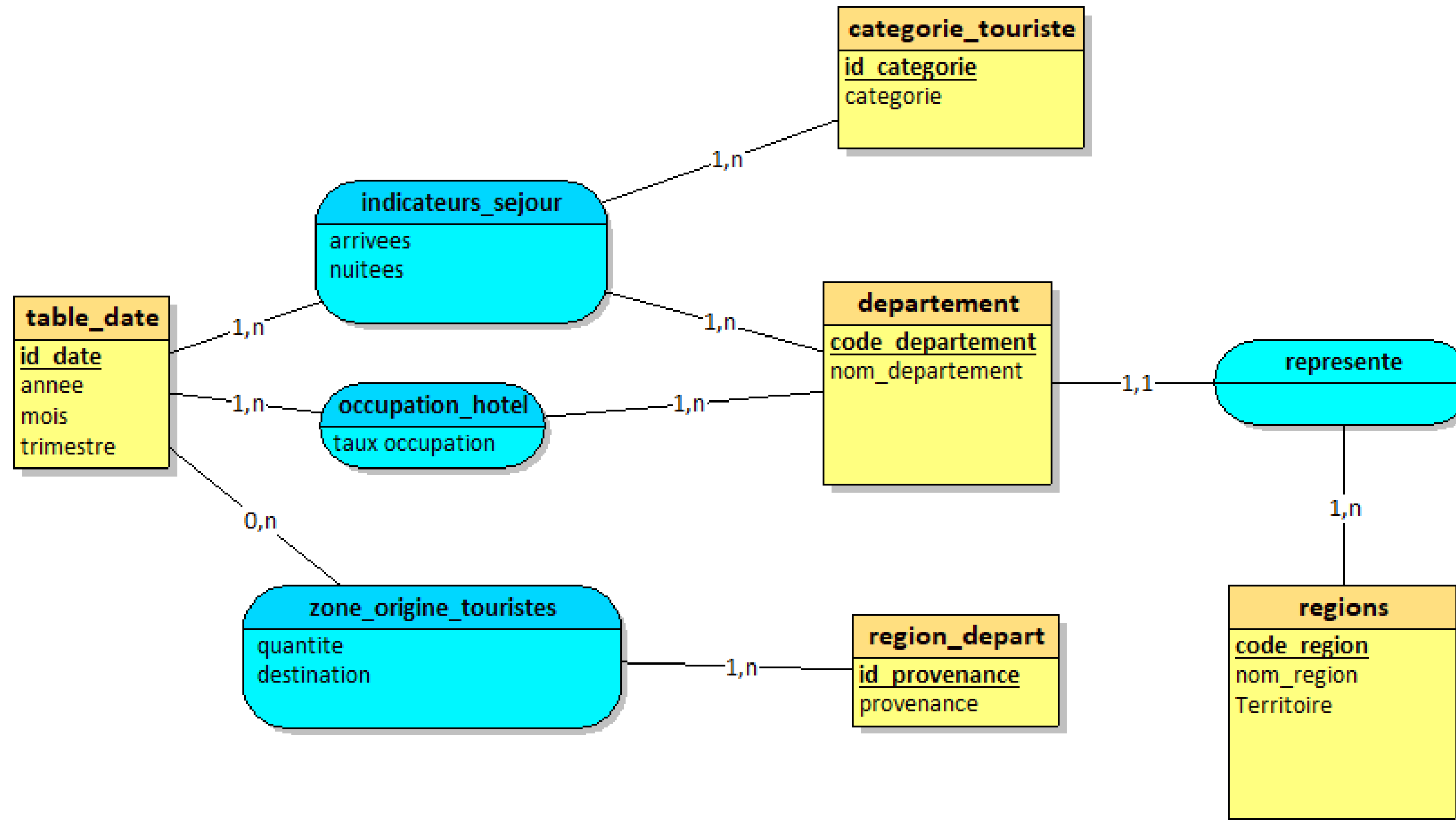
Dans cette sous-partie :

- ❑ Modèle Conceptuel des Données
- ❑ Modèle Physique des Données
- ❑ Détails des tables
- ❑ Exemple de requêtes sql et optimisation
- ❑ Sécurisation et sauvegarde des données
- ❑ Datavisualisation



MODÈLE CONCEPTUEL DES DONNÉES

14



Relations ManyToOne

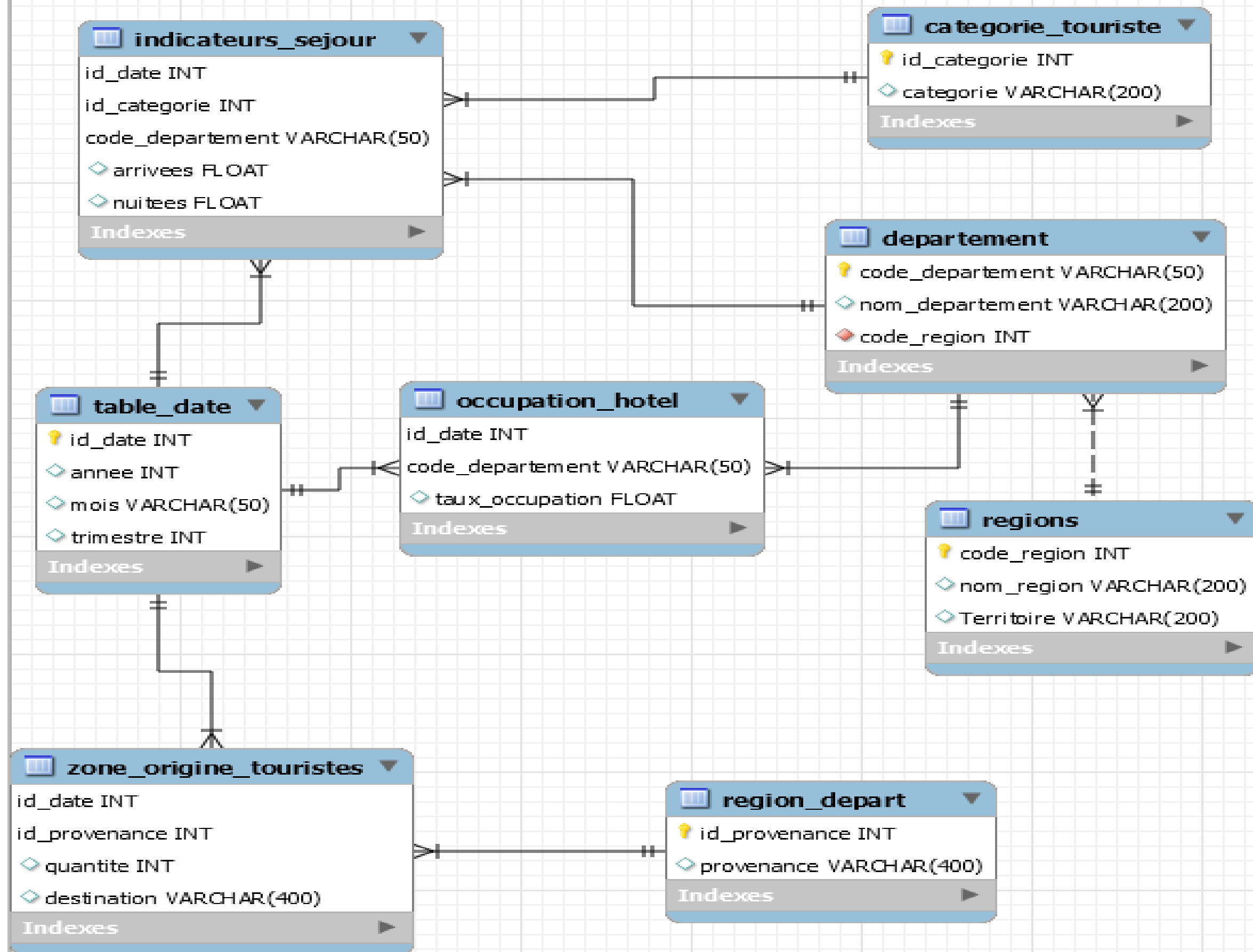
La base est liée à une seule target et la target peut être liée à plusieurs bases

Exemple de la relation **departement** et **region**

Relations ManyToMany

La base peut être liée à plusieurs targets et la target peut être liée à plusieurs bases

Exemple de la relation **table_date** et **indicateurs_sejour**



Permet de définir la mise en œuvre de structures physiques et de requêtes portant sur des données.

Ce sont les données telles qu'elles existent dans l'application informatique.

SCRIPT CREATION BDD

Démarche: création des tables, insertion des clés primaires, insertion des fichiers csv via load data, insertion des clés étrangères

16

```
#création de la base de donnée tourisme
```

```
drop database if exists tourisme;  
create database if not exists tourisme;  
use tourisme;
```

```
CREATE TABLE regions(  
    code_region INT,  
    nom_region VARCHAR(200),  
    Territoire VARCHAR(200)  
);
```

```
57  
58     #insertion des clés primaires  
59  
60 •   Alter table regions add PRIMARY KEY(code_region);  
61 •   alter table table_date add PRIMARY KEY(id_date);
```

SCRIPT CREATION BDD (suite)

17

```
130
131     #Insertion des clés étrangères
132
133 •   SET FOREIGN_KEY_CHECKS=0;
134 •   Alter table departement add FOREIGN KEY(code_region) REFERENCES regions(code_region);
135 •   alter table zone_origine_touristes add FOREIGN KEY(id_date) REFERENCES table_date(id_date);
136 •   alter table zone_origine_touristes add FOREIGN KEY(id_provenance) REFERENCES region_depart(id_provenance);
137 •   alter table indicateurs_sejour add FOREIGN KEY(id_date) REFERENCES table_date(id_date);
138 •   alter table indicateurs_sejour add FOREIGN KEY(id_categorie) REFERENCES categorie_touriste(id_categorie);
139 •   alter table indicateurs_sejour add FOREIGN KEY(code_departement) REFERENCES departement(code_departement);
140 •   alter table occupation_hotel add FOREIGN KEY(id_date) REFERENCES table_date(id_date);
141 •   alter table occupation_hotel add FOREIGN KEY(code_departement) REFERENCES departement(code_departement);
142 •   SET FOREIGN_KEY_CHECKS=1;
143
```

```
69
70     #connection avec les données contenues dans des fichiers csv et enregistrées en local
71     #afin de les insérer dans mes tables
72
73 •   SET GLOBAL local_infile=1;
74 •   LOAD DATA LOCAL INFILE 'C:/Users/Alain NGABO/Dropbox/SIMPLON/Chefdoeuvre/data_propre/regions.csv'
75     INTO TABLE regions
76     FIELDS TERMINATED BY ','
77     ENCLOSED BY ''
78     LINES TERMINATED BY '\n'
79     IGNORE 1 ROWS;
80
81 •   LOAD DATA LOCAL INFILE 'C:/Users/Alain NGABO/Dropbox/SIMPLON/Chefdoeuvre/data_propre/table_date.csv'
82     INTO TABLE table_date
83     FIELDS TERMINATED BY ','
84     ENCLOSED BY ''
85     LINES TERMINATED BY '\n'
86     IGNORE 1 ROWS;
87
```

Détails de quelques tables

	A	B	C	D	E	F	G	H
1	Libellé	idBank	Dernière mise à	Période	2010-0	2010-02	2010-03	2010-0
2	Arrivées dans l'hôtellerie - Total - Départements d'outre-mer - Série arrêtée	1717310	06/03/2019 15:20		98.29	97.11	94.35	97.0
3	Codes				A	A	A	A
4	Arrivées dans l'hôtellerie - Total - France métropolitaine et Départements d'outre-mer - Série arrêtée	1717313	06/03/2019 15:20		6098.8	6691.69	7929.4	9034.3
5	Codes				A	A	A	A
6	Arrivées dans l'hôtellerie - Total - France métropolitaine - Série arrêtée	1711123	06/03/2019 15:20		6000.6	6594.58	7835.05	8937.3
7	Codes				A	A	A	A
8	Arrivées dans l'hôtellerie - Total - Guadeloupe - Série arrêtée	1717210	06/03/2019 15:20		28.64	31.13	28.09	27.52
9	Codes				A	A	A	A
0	Arrivées dans l'hôtellerie - Total - Martinique - Série arrêtée	1717213	06/03/2019 15:20		27.65	30.41	27.26	27.6
1	Codes				A	A	A	A
2	Arrivées dans l'hôtellerie - Total - Guyane - Série arrêtée	1717216	06/03/2019 15:20		9.09	9.29	10.62	9.36
3	Codes				A	A	A	A
4	Arrivées dans l'hôtellerie - Total - La Réunion - Série arrêtée	1717219	06/03/2019 15:20		32.9	26.27	28.38	32.52
5	Codes				A	A	A	A
6	Arrivées dans l'hôtellerie - Total - Île-de-France - Série arrêtée	1711112	06/03/2019 15:20		2207.9	2201.94	2659.75	2617.5
7	Codes				A	A	A	A
8	Arrivées dans l'hôtellerie - Total - Centre-Val de Loire - Série arrêtée	1711107	06/03/2019 15:20		166.87	189.11	249.37	315.19
9	Codes				A	A	A	A
0	Arrivées dans l'hôtellerie - Total - Bourgogne-Franche-Comté - Série arrêtée	1739001	06/03/2019 15:20		241.91	308.87	333.78	414.1

Echantillon du csv d'origine

2 exemples de tables générées

	id_date	id_categorie	code_departement	arrivees	nuitées
	1	1	1	39.91	56.49
	1	1	10	30.37	39.68
	1	1	11	20.85	29.37
	1	1	12	15.9	23.07
	1	1	13	136.06	226.09
	1	1	14	72.25	114.29
	1	1	15	10.0	17.06

	id_date	code_departement	taux_occupation
	1	1	46.1
	1	10	49.7
	1	11	31.6
	1	12	25.3
	1	13	45.5
	1	14	36

Détails de quelques tables

	A	B	C	D	E	F	G	H
1	Libellé	idBank	Dernière mise à	Période	2010-0	2010-02	2010-03	2010-0
2	Arrivées dans l'hôtellerie - Total - Départements d'outre-mer - Série arrêtée	1717310	06/03/2019 15:20		98.29	97.11	94.35	97.0
3	Codes				A	A	A	A
4	Arrivées dans l'hôtellerie - Total - France métropolitaine et Départements d'outre-mer - Série arrêtée	1717313	06/03/2019 15:20		6098.8	6691.69	7929.4	9034.3
5	Codes				A	A	A	A
6	Arrivées dans l'hôtellerie - Total - France métropolitaine - Série arrêtée	1711123	06/03/2019 15:20		6000.6	6594.58	7835.05	8937.3
7	Codes				A	A	A	A
8	Arrivées dans l'hôtellerie - Total - Guadeloupe - Série arrêtée	1717210	06/03/2019 15:20		28.64	31.13	28.09	27.52
9	Codes				A	A	A	A
0	Arrivées dans l'hôtellerie - Total - Martinique - Série arrêtée	1717213	06/03/2019 15:20		27.65	30.41	27.26	27.6
1	Codes				A	A	A	A
2	Arrivées dans l'hôtellerie - Total - Guyane - Série arrêtée	1717216	06/03/2019 15:20		9.09	9.29	10.62	9.36
3	Codes				A	A	A	A
4	Arrivées dans l'hôtellerie - Total - La Réunion - Série arrêtée	1717219	06/03/2019 15:20		32.9	26.27	28.38	32.52
5	Codes				A	A	A	A
6	Arrivées dans l'hôtellerie - Total - Île-de-France - Série arrêtée	1711112	06/03/2019 15:20		2207.9	2201.94	2659.75	2617.5
7	Codes				A	A	A	A
8	Arrivées dans l'hôtellerie - Total - Centre-Val de Loire - Série arrêtée	1711107	06/03/2019 15:20		166.87	189.11	249.37	315.19
9	Codes				A	A	A	A
0	Arrivées dans l'hôtellerie - Total - Bourgogne-Franche-Comté - Série arrêtée	1739001	06/03/2019 15:20		241.91	308.87	333.78	414.1

Echantillon du csv d'origine

2 exemples de tables générées

	id_date	id_categorie	code_departement	arrivees	nuitées
	1	1	1	39.91	56.49
	1	1	10	30.37	39.68
	1	1	11	20.85	29.37
	1	1	12	15.9	23.07
	1	1	13	136.06	226.09
	1	1	14	72.25	114.29
	1	1	15	10.0	17.06

	id_date	code_departement	taux_occupation
	1	1	46.1
	1	10	49.7
	1	11	31.6
	1	12	25.3
	1	13	45.5
	1	14	36

Sécurisation et sauvegarde

Sécurisation : création d'un utilisateur qui n'a pas les privileges d'administrateur. Il ne peut que lire la BDD.

```
mysql> create user 'alain'@'localhost';
Query OK, 0 rows affected (0.13 sec)

mysql> alter user 'alain'@'localhost' identified by 'Portes1907*';
Query OK, 0 rows affected (0.01 sec)

mysql> grant select on *.* to 'alain'@'localhost';
Query OK, 0 rows affected (0.01 sec)

mysql> flush privileges;
Query OK, 0 rows affected (0.00 sec)
```

Sauvegarde : creation d'une sauvegarde via dump

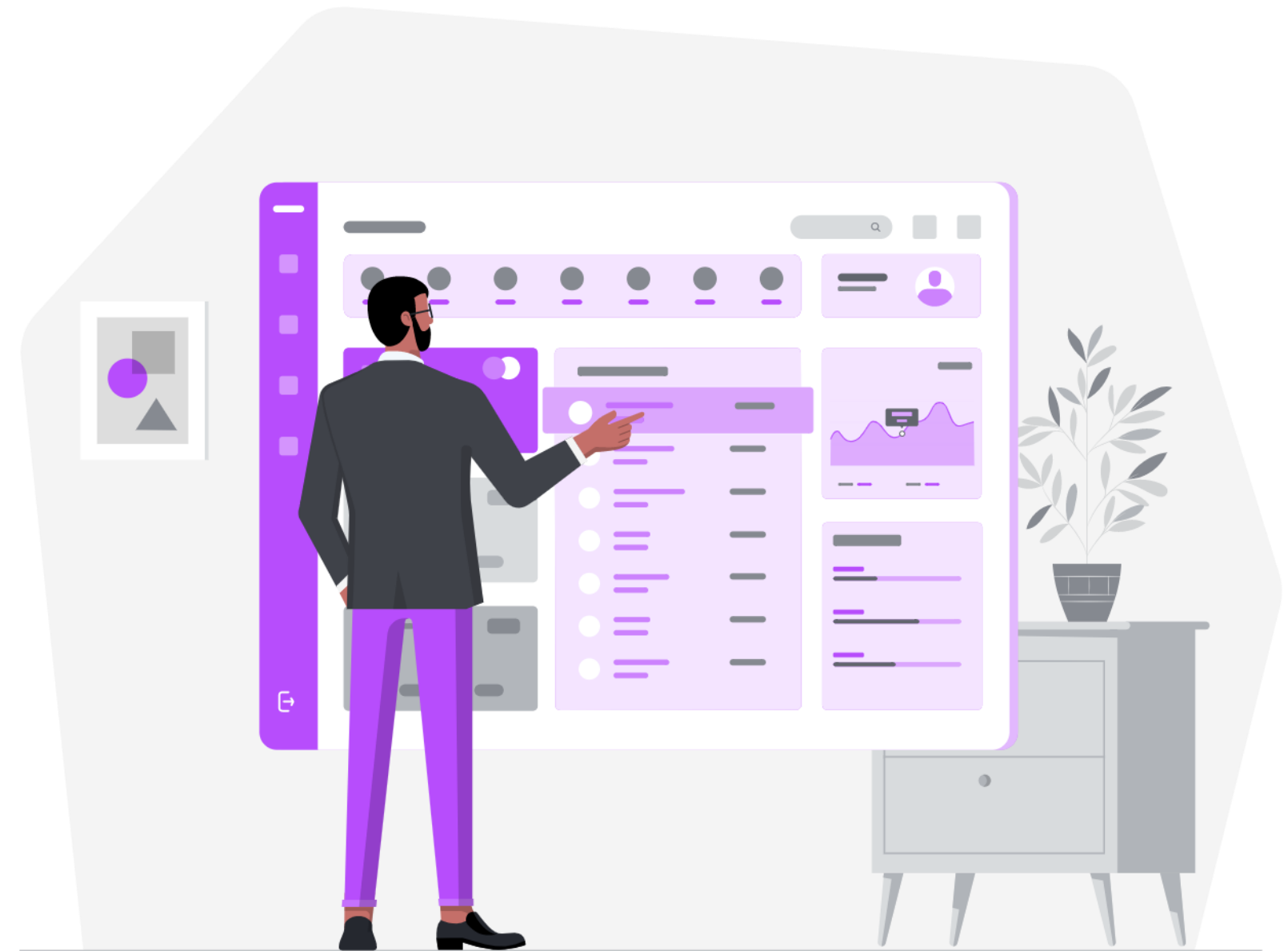
```
(base) C:\Users\Alain NGABO\Dropbox\SIMPLON\Chefdoeuvre>mysqldump.exe -u root -p tourisme > dumptourisme.sql
Enter password: *****

(base) C:\Users\Alain NGABO\Dropbox\SIMPLON\Chefdoeuvre>
```


Datavisualisation

Dans cette sous-partie :

- ❑ Développer un rapport power bi
- ❑ Connection à la base de données Mysql afin d'automatiser la mise à jour
- ❑ Partager le dashboard tout en veillant à la sécurité

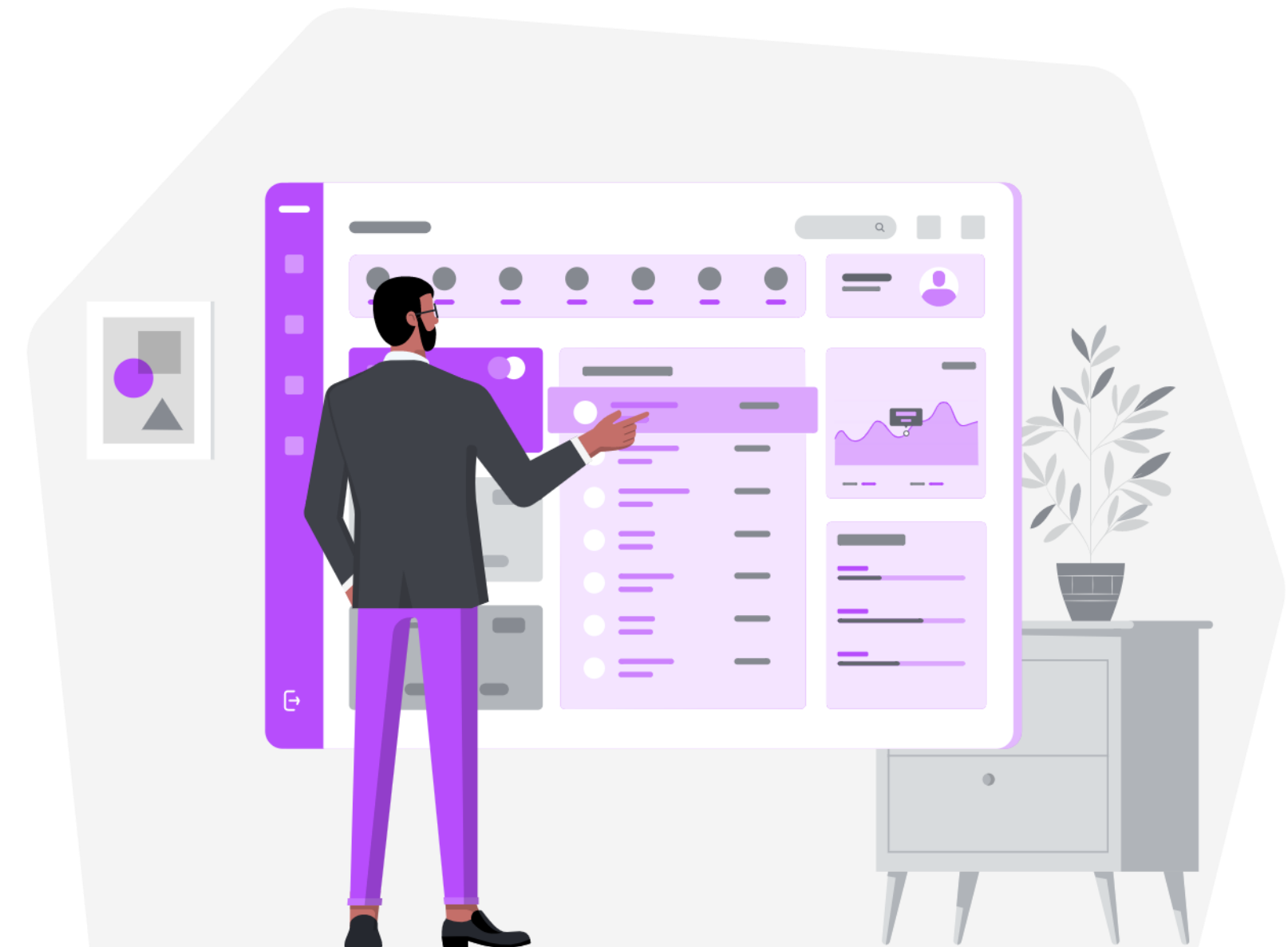


Datavisualisation

Rappel des objectifs pour la datavisualisation :

Observer en fonction du lieu et la périodicité:

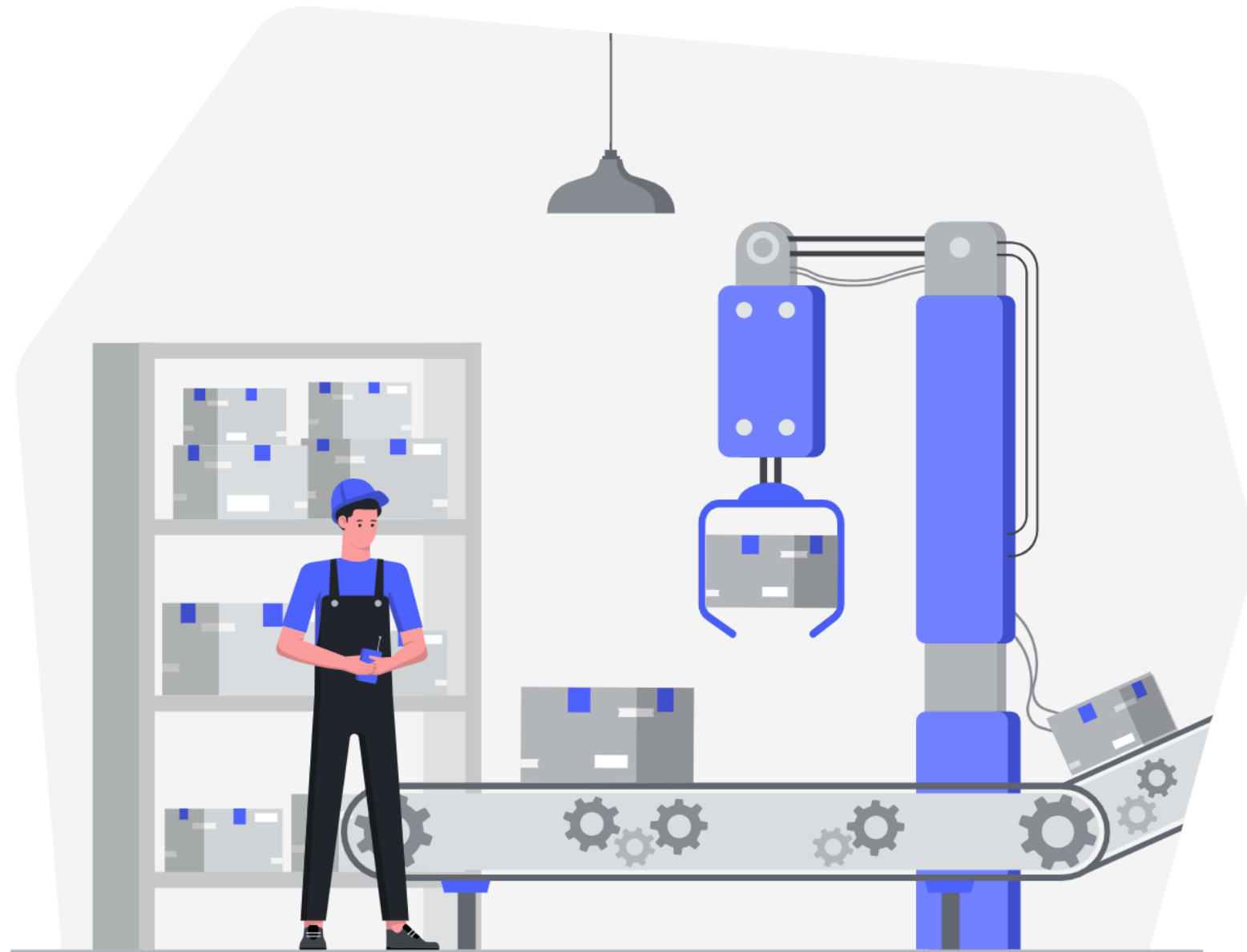
- le volume de touristes accueillis,
- leur origine (clientele Française ou étrangère)
- les pays de provenance
- la durée du séjour
- le taux d'occupation des hotels



[Lien accès dashboard](#)

ensuite email: alain.ngabo@agencengabo.onmicrosoft.com

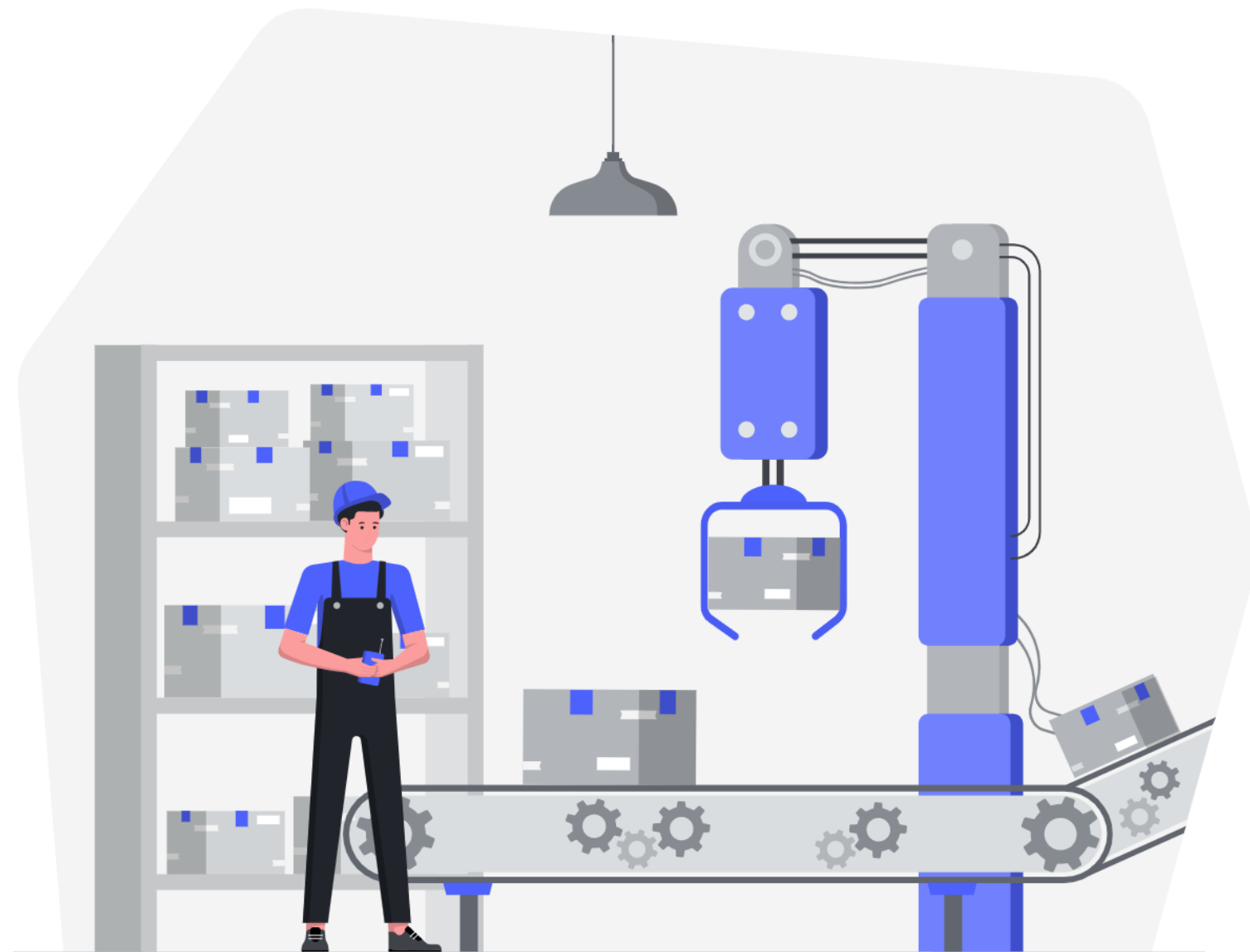
Mot de passe : Portes1907*



Dans cette sous-partie, focus sur quelques grandes étapes dans la preparation des données.

Langage : python

Panorama des méthodes et fonctions utilisées



Str.contains
str.replace
Split
Drop
Rename
set value
Apply
Merge

melt
drop_duplicates
groupby
pd.to_datetime
pd.get_dummies
Lambda
read_csv
to_csv

Librairies

Numpy; pandas; datetime, os

Aperçu du csv d'origine contenant les principaux indicateurs

25

```
Entrée [2]: #téléchargement du csv d'origine pour le mettre dans un dataframe pandas. Nous l'appelons df_mensuel
df_mensuel=pd.read_csv('data/data_avec_pays_origine/valeurs_mensuelles.csv', sep=',')
df_mensuel.sample(50)
```

	Libellé	idBank	Dernière mise à jour	Période	2010-01	2010-02	2010-03	2010-04	2010-05	2010-06	...	2020-03	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09	2020-10	2020-11	2020-12
375	Codes	NaN	NaN	NaN	O	O	O	A	A	A	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1069	Codes	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	O	O	O	O	A	A	A	A	O	O
1428	Arrivées dans l'hôtellerie - Résidents - Occit...	10598651.0	14/04/2021 18:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	234	NaN	NaN	NaN	728	975	629	443	NaN	NaN
1202	Arrivées dans l'hôtellerie - Résidents - Aveyron	10598936.0	14/04/2021 18:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	52.6	70.8	47.8	28.9	NaN	NaN
748	Arrivées dans l'hôtellerie - Total - Landes	10599176.0	14/04/2021 18:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	58.97	81.21	48.28	38.42	NaN	NaN
	Arrivées dans																				

Importation du csv et transformation en dataframe.

Les principales informations sont contenues dans la colonne Libellé.

Les dates ne sont pas dans une colonne.

Suppression des données avec le libellé “Total” et régions

26

```
Entrée [4]: #suppression des données calculées par "total" affichées plus haut
df_mensuel= df_mensuel[~df_mensuel['Libellé'].str.contains("Total")]
df_mensuel['Libellé'].unique()

Out[4]: array(['Codes',
               "Arrivées dans l'hôtellerie - Non-résidents - Départements d'outre-mer - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - France métropolitaine et Départements d'outre-mer - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - France métropolitaine - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Île-de-France - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Centre-Val de Loire - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Bourgogne-Franche-Comté - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Normandie - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Hauts-de-France - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Grand Est - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Pays de la Loire - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Bretagne - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Nouvelle-Aquitaine - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Occitanie - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Auvergne-Rhône-Alpes - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Provence-Alpes-Côte d'Azur - Série arrêtée",
               "Arrivées dans l'hôtellerie - Non-résidents - Corse - Série arrêtée",
               "Arrivées dans l'hôtellerie - Résidents - Départements d'outre-mer - Série arrêtée",
               "Arrivées dans l'hôtellerie - Résidents - France métropolitaine et Départements d'outre-mer - Série arrêtée",
               ...])
```

Dans notre jeu de données, on retrouve le volume de touristes au niveau national, par region et par département. Objectif: supprimer le total au niveau national ainsi que par region afin de garder la plus petite granularité, le département.

Une autre étape du nettoyage : regrouper toutes les dates en 2 colonnes (annee et mois)

```
#Pour créer la colonne variable qui contient toutes les années et mois, nous allons utiliser la méthode melt.
#Nous allons prendre toutes les colonnes qui se trouvent après le 4ème élément jusqu'au dernier.

df_mensuel=pd.melt(df_mensuel, id_vars = 'Libellé', value_vars =df_mensuel.columns[4:])
```

```
#nous allons renommer les colonnes suivantes pour ce que ce soit plus clair. Méthode rename

df_mensuel.rename(columns={'variable': 'annee',
                           'value': 'valeur'}, inplace=True)
```

```
: #Nous allons diviser en 2 la colonne annee afin d'avoir l'année et le mois.
#Pour cela nous allons convertir la colonne en datetime.
df_mensuel['annee']= pd.to_datetime(df_mensuel['annee'])
df_mensuel['Annee'] = df_mensuel['annee'].dt.year
df_mensuel['Mois'] = df_mensuel['annee'].dt.month
df_mensuel.drop(columns=["annee"],inplace=True) #suppression de la colonne annee d'origine puisque nous n'avons plus besoin.
df_mensuel.sample(50)
```

```
df_mensuel.head()
```

Out[42]:

	annee	mois	nom_departement	categorie	arrivees	nuitees	taux_occupation	id_categorie	code_departement
6060	2011	1	Ain	etrangers	7.41			2	1
6061	2011	1	Aisne	etrangers	2.4			2	2
6062	2011	1	Allier	etrangers	0.94			2	3
6063	2011	1	Alpes-de-Haute-Provence	etrangers	0.65			2	4
6064	2011	1	Hautes-Alpes	etrangers	3.01			2	5

Une autre étape du nettoyage : utilisation de la méthode dummies pour mettre les valeurs de la colonne Type dans des colonnes distinctes

	Annee	Mois	valeur	Departement	Type	typologie_touristes
21907	2013	8	62.4	Essonne	Arrivées dans l'hôtellerie	français
11954	2011	12	20.77	Indre	Nuitées dans l'hôtellerie	français
39846	2016	7	57.6	Haute-Marne	Taux d'occupation dans l'hôtellerie	Non applicable

Entrée [29]: `#utilisation de la méthode get dummies pour les elements de la colonne Type en colonne`
`dummies=pd.get_dummies(df_mensuel['Type'])`
`dummies.head()`

Out[29]:

	Arrivées dans l'hôtellerie	Nuitées dans l'hôtellerie	Taux d'occupation dans l'hôtellerie
6060	1	0	0
6061	1	0	0
6062	1	0	0

Entrée [30]: `#reaffectation des valeurs`
`for col in dummies.columns:`
`df_mensuel[col]=dummies[col]*df_mensuel['valeur']`
`df_mensuel.sample(50)`

Out[30]:

	Annee	Mois	valeur	Departement	Type	typologie_touristes	Arrivées dans l'hôtellerie	Nuitées dans l'hôtellerie	Taux d'occupation dans l'hôtellerie
63278	2020	6	NaN	Haute-Marne	Arrivées dans l'hôtellerie	français	NaN	NaN	NaN
66616	2020	12	NaN	Moselle	Taux d'occupation dans l'hôtellerie	Non applicable	NaN	NaN	NaN
64971	2020	9	227.8	Corse-du-Sud	Nuitées dans l'hôtellerie	français		227.8	
16812	2012	10	24.05	Lot	Arrivées dans l'hôtellerie	français	24.05		
31909	2015	4	47.17	Val-de-Marne	Arrivées dans l'hôtellerie	etrangers	47.17		

Entrée [68]: `#on fait un groupby afin d'avoir des indicateurs uniques`
`df_indicateurstourisme2=df_indicateurstourisme2.groupby(["id_date","id_categorie","code_departement"])[["arrivees","nuitées","taux_occupation"]].sum()`
`df_indicateurstourisme2.sample(100)`

Out[68]:

	id_date	id_categorie	code_departement	arrivees	nuitées	taux_occupation
63	3	5	0.00	0.00	51.0	
80	3	32	0.00	0.00	55.6	
32	2	90	5.21	7.21	0.0	
120	2	95	-600.00	-600.00	-600.0	

Nous avons pris soin de supprimer les nan avant

Difficultés rencontrées

29

Sourcing des données

Tous les départements ne déclarent pas le même niveau d'information et au même moment. Il a fallu identifier la bonne source INSEE et le rapport définitif car il existe beaucoup de rapports intermédiaires avec des chiffres différents.

Nettoyage des données

Utilisation de plusieurs fonctions et méthodes. Nous avons utilisé principalement stackoverflow, consulté la documentation technique et des comptes GitHub

Quelle suite pour ce projet ?

30

Ajouter des indicateurs économiques

Ajouter les autres types d'hébergement (camping, plateformes comme Airbnb...)

Pour les arrivées des étrangers, enrichir les pays de provenance et leur destination en France

Heberger la base de données dans un cloud

Merci pour votre attention!

AVEZ-VOUS DES QUESTIONS OU BESOIN
DE PRÉCISIONS ?

