



CHEF D'ŒUVRE

D'Alain Juste NGABO

Nanterre p10
- Dev Data

SIMPLON
.CO

CERTIFICATIONS

Développer une base de données (CNCP 3497)

Exploiter une base de données (CNCP 3508)

Table des matières

INTRODUCTION	2
1- PRESENTATION ET ENJEU DU PROJET	2
2- CONTEXTE ET OBJECTIFS	2
Contexte	2
Objectifs	4
3- LEXIQUE DES TERMES CLES DU PROJET	4
4- STACK TECHNIQUE	5
5- MODELE ORGANISATIONNEL DES DONNEES	7
I- CONCEPTION DE LA BASE DE DONNEES	8
1. LES DIFFERENTES SOURCES DES DONNEES ET LEUR DESCRIPTION	8
2. MODELE CONCEPTUEL DE DONNEES (MCD)	10
3. MODELE LOGIQUE DE DONNEES (MLD)	12
4. MODELE PHYSIQUE DE DONNEES (MPD)	13
5. NETTOYAGE DES DONNEES ET CREATION DES TABLES	14
6. INSERTION DES DONNEES DANS LA BASE DE DONNEES MYSQL	16
7. SECURITE ET SAUVEGARDE DE LA BASE DE DONNEES	18
8. EXEMPLE DE REQUETE SQL OPTIMISEE	18
II – DATAVISUALISATION ET ANALYSE DES DONNEES	20
1. RAPPEL DES OBJECTIFS	20
2. OUTILS UTILISES (POWER BI DESKTOP ET POWER BI SERVICE)	20
3. DATAVISUALISATION	21
4. AUTOMATISATION DE LA RECUPERATION DES DONNEES	24
III – GESTION ET BILAN DU PROJET	24
1. TRELLO	24
2. DIFFICULTES RENCONTREES	25
3. AXES D’AMELIORATIONS DU PROJET	26
REMERCIEMENTS	27
BIBLIOGRAPHIE/WEBGRAPHIE	28

Introduction

L'objet de ce travail est une démonstration de l'ensemble des compétences acquises durant les sept mois de formation au métier de Développeur Data au sein de Simplon.

Débutée au mois de décembre 2020, cette formation intensive délivre principalement quatre certifications reconnues par l'Etat et par des organismes professionnels :

1. [Développer une base de données \(CNCP 3497\)](#)
2. [Exploiter une base de données \(CNCP 3508\)](#)
3. [Méthodes agiles de gestion et amorçage de projets \(Scrum, Kanban, Lean startup...\)](#)
4. [Maitrise de la qualité en gestion en projet web \(délivrée par OPQUAST\)](#)

A date, j'ai déjà obtenu les deux dernières certifications et ce rapport s'inscrit dans le processus d'obtention des deux premières certifications. Il décrit de manière conceptuelle et technique mon projet de fin de formation. Il accompagne l'ensemble des réalisations.

1- Présentation et enjeu du projet

Pour ce projet de fin de formation, j'ai choisi de m'intéresser à la data dans le tourisme en France, spécifiquement sur les données touristiques générées par l'hébergement touristique.

Il a pour enjeu de mieux appréhender le tourisme en France en se focalisant sur le comportement des clientèles touristiques à partir de l'analyse de la fréquentation des hôtels ces neuf dernières années (2011 – 2020). L'année 2010 n'est pas incluse car il n'y avait pas suffisamment de données disponibles.

Concrètement, ce projet est composé de deux parties principales :

- Une base de données pour stocker de manière structurée l'ensemble des données
- Une application de datavisualisation pour afficher et partager les résultats des analyses. Elle devra être dynamique et interactive afin que l'utilisateur puisse effectuer des requêtes et afficher les résultats.

2- Contexte et objectifs

Contexte

D'après les chiffres de l'Organisation Mondiale du Tourisme, la France est la première destination touristique mondiale depuis les années 1980 en termes d'arrivées de touristes internationaux (la 3ème en termes de recettes derrière l'Espagne et les Etats-Unis). La consommation

touristique intérieure atteint 7,4 % du PIB français en 2018. Le tourisme représente donc un atout important dans l'économie Française.

Afin de mieux connaître la typologie et les habitudes des touristes en France, et de mettre ces informations à la disposition des acteurs économiques et institutionnelles de ce secteur, le gouvernement a entrepris un travail dans ce sens qu'on peut résumer en deux grandes phases :

- **2011** : réflexion gouvernementale sur la mise en place d'une plateforme data sur le tourisme ;
- **2017** : lancement de la plateforme [datatourisme](https://datatourisme.gouv.fr/), plateforme open data du tourisme en France.

Cet espace est alimenté exclusivement par les organismes institutionnels chargés de collecter l'information touristique locale au sein d'un Système ou Réseau d'Information Touristique local. Il s'agit des Offices de Tourisme (OT), Agences et Comités Départementaux du Tourisme (ADT), et Comités Régionaux du Tourisme (CRT) mais également de tout autre producteur de données touristiques publiques œuvrant à l'échelle territoriale ou nationale.

En principe, en se connectant soit à une **API** (en informatique, une API est un ensemble de définitions et de protocoles qui facilite la création et l'intégration de logiciels d'applications) ou en téléchargeant des fichiers de type **Excel**, on a accès à un ensemble d'informations renseignées par les acteurs sus-cités.

Cependant, cette plateforme rencontre quelques problèmes qui ne facilitent pas l'analyse des données disponibles. Après avoir testé cet outil pendant quelques semaines, j'ai constaté que les données ne sont toujours pas correctement nettoyées, harmonisées et exploitables par tout le monde.

The screenshot shows the 'diffuseur.datatourisme.gouv.fr/' website. On the left, there's a sidebar with 'Dernières annonces' (Latest announcements) regarding a platform migration. The main content area displays a table of user-reported anomalies. A red line highlights the following entries:

Question	Nombre de votes	Nombre de réponses	Délai	Statut
Question sur "se déroule le " -> takesPlaceAt	3	30	il y a 3 jours	Qualité
Question qualité traduction	12	81	il y a 3 jours	Qualité
Qgis, réutilisation des données par catégories : style, catégorisation, icône	3	33	il y a 3 jours	Ontologie
L'api docker stack ne retourne que fr et en?	4	52	il y a 3 jours	API
Qualité de contenu de la description	16	108	il y a 8 jours	Qualité
Visiblement Obsolète mais ne possède pas le tag	0	25	il y a 11 jours	Qualité
Anomalie divers poi	7	59	il y a 15 jours	Qualité
Flux JSON zip invalides	9	55	il y a 16 jours	Application diffuseur

Capture d'écran page datatourisme avec des anomalies remontées par les utilisateurs. Version du 20/8/2020

A la quête de données intègres, structurées et validées sur le tourisme en France, j'ai remarqué que la plateforme open data du gouvernement (datagouv) indique que l'Institut National de la

Statistique et des Etudes Economiques (INSEE) met à disposition [des informations sur le tourisme en France](#).

Les informations qu'on y retrouve sont :

- Nombre de nuitées et d'arrivées par mois dans les hôtels par région
- Nombre de nuitées et d'arrivées par mois dans l'hôtellerie de plein-air par région
- Le taux d'occupation des hôtels

Objectifs

Dans le cadre de ce travail, j'ai choisi de m'intéresser au nombre de nuitées, d'arrivées et le taux d'occupation des hôtels par département et région.

L'objectif est d'observer en fonction du lieu (département, région) et la périodicité (mois, trimestre, année) :

- Le volume de touristes accueillis,
- Leur origine (clientèle Française ou étrangère),
- Les pays de provenance,
- La durée du séjour,
- Le taux d'occupation des hôtels.

De manière aboutie, la base de données et le dashboard (outil de datavisualisation) 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.

3- Lexique des termes clés du projet

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

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.

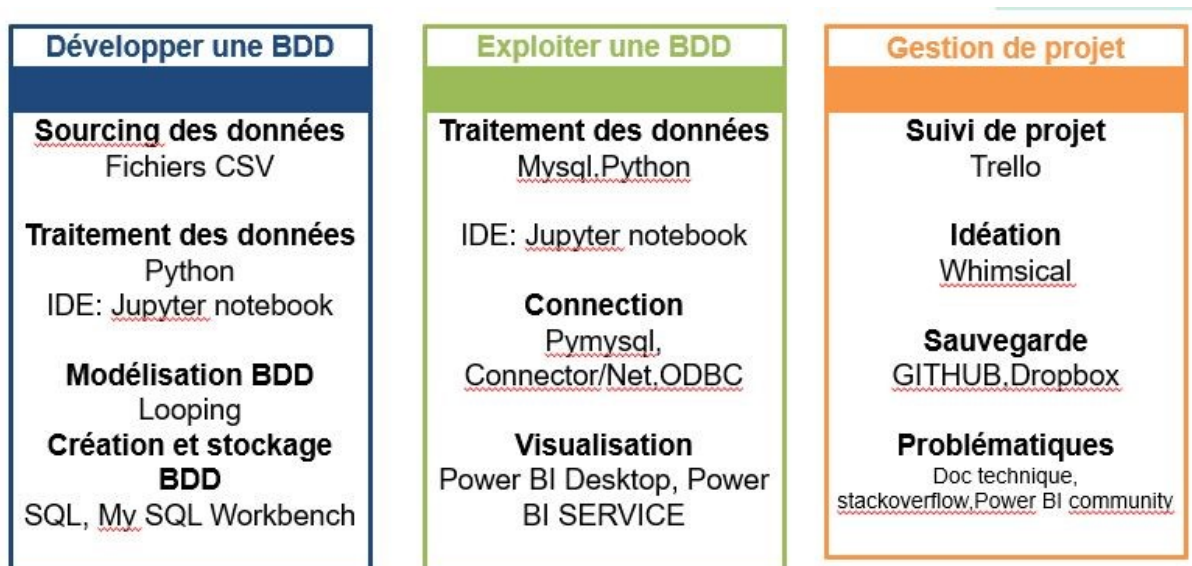
Exemple :

- 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.

4- Stack technique

Une stack technique, en anglais « technology stack », également appelée « tech stack », « pile de technologies » ou « écosystème de données », est une liste de tous les outils technologiques utilisés pour développer et faire fonctionner un programme.

Ci-après, voici la stack technique de notre projet :



Définitions de principaux outils et langages :

Fichier CSV :

Un fichier CSV (en anglais, comma separated values) est le fichier de base des données recueillies - sans formatage particulier. Chaque champ est séparé par une virgule ou autre ponctuation.

Python :

Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet.

Dans le domaine de la data, Python est le langage le plus utilisé. Sa flexibilité permet de prendre en charge le développement de modèles de Machine Learning, le forage et le traitement des données, ainsi que d'autres tâches plus rapidement que les autres langages.

Jupyter notebook :

Les notebooks Jupyter sont des environnements de développement ou cahiers électroniques qui, dans le même document, peuvent rassembler du texte, des images, des formules mathématiques et du code informatique exécutable. Ils sont manipulables interactivement dans un navigateur web.

Initialement développés pour les langages de programmation Julia, Python et R (d'où le nom Jupyter), les notebooks Jupyter supportent près de 40 langages différents.

La cellule est l'élément de base d'un notebook Jupyter. Elle peut contenir du texte formaté au format Markdown ou du code informatique qui pourra être exécuté.

Looping :

Looping est un logiciel de modélisation conceptuelle de données entièrement gratuit et libre d'utilisation.

SQL :

SQL ou " Structured Query Language " est un langage de programmation permettant de manipuler les données et les systèmes de bases de données relationnelles. Ce langage permet principalement de communiquer avec les bases de données afin de gérer les données qu'elles contiennent.

Il permet notamment de stocker, de manipuler et de retrouver ces données. Il est aussi possible d'effectuer des requêtes, de mettre à jour les données, de les réorganiser, ou encore de créer et de modifier le schéma et la structure d'un système de base de données et de contrôler l'accès à ses données.

MySql Workbench :

MySQL Workbench est un outil visuelle pour la conception de base de données. Il permet notamment de représenter la structure de la base de données et les liaisons entre les données. Les informations peuvent ensuite être exporter pour générer les requêtes SQL nécessaire à la création de la base.

Microsoft Power BI :

Power BI est une solution d'analyse de données de Microsoft. Il permet de créer des visualisations de données personnalisées et interactives avec une interface suffisamment simple pour que les utilisateurs finaux créent leurs propres rapports et tableaux de bord.

Trello :

Trello est un outil de gestion de projet en ligne, lancé en septembre 2011 et inspiré par la méthode Kanban de Toyota. Il repose sur une organisation des projets en planches listant des cartes, chacune représentant des tâches.

Whimsical :

Comme Trello, whimsical est un outil en ligne et gratuit, qui permet de représenter ses idées sous formes de cartes graphiques.

Github :

Lancé en 2008, GitHub est un site web conçu pour fédérer (héberger) et partager le code source d'un projet de développement d'application mis à œuvre par plusieurs programmeurs.

Dropbox :

Dropbox est un service de stockage et de partage de copies de fichiers locaux en ligne proposé par Dropbox, Inc., entreprise localisée à San Francisco, en Californie.

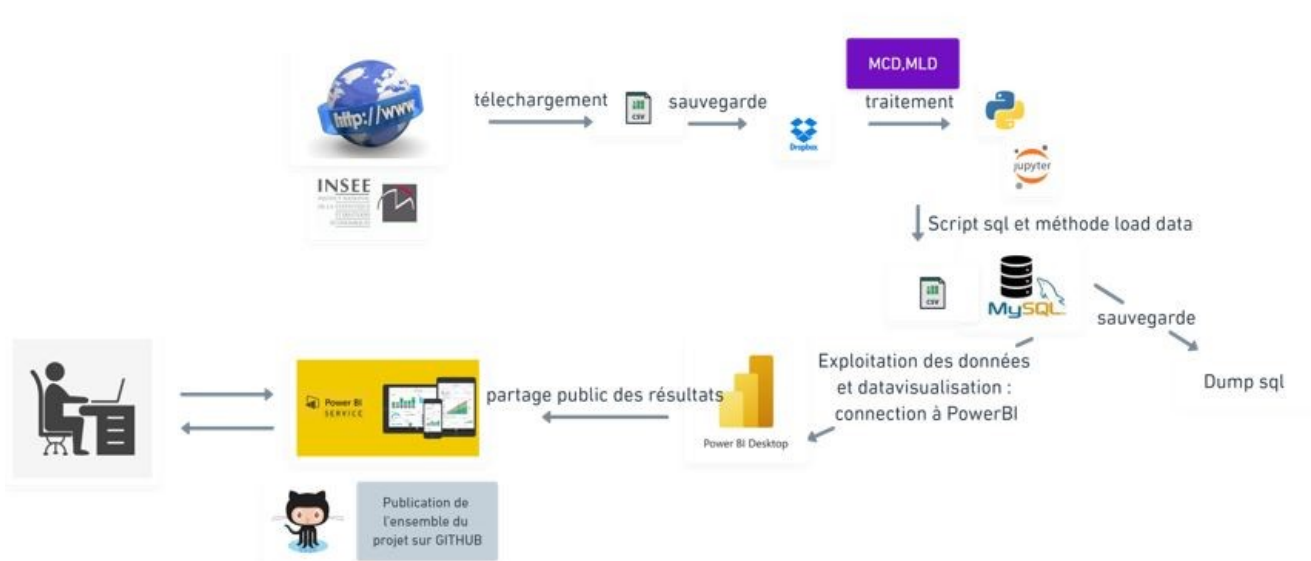
Stackoverflow :

Stack Overflow est un site web proposant des questions et réponses sur un large choix de thèmes concernant la programmation informatique.

5- Modèle organisationnel des données

Le modèle organisationnel des données ou architecture générale, décrit de manière schématique le fonctionnement de notre programme, application ou projet.

Ci-dessous, le modèle de notre chef d'œuvre :








I- Conception de la base de données

1. Les différentes sources des données et leur description

Les indicateurs arrivées, nuitées et taux d'occupation

Sur le site internet de l'INSEE <https://www.insee.fr/fr/statistiques/series/113990189#>, nous avons coché les différentes cases et téléchargé un dossier contenant 4 fichiers csv (image ci-dessous).

 caractéristiques.csv	10/05/2021 18:54	Classeur OpenOffi...	665 Ko
 famille_TOURISME-FRANCE-METHODE-R...	10/05/2021 18:54	Feuille de calcul ...	1 185 Ko
 valeurs_annuelles.csv	10/05/2021 18:54	Classeur OpenOffi...	393 Ko
 valeurs_mensuelles.csv	22/06/2021 16:21	Classeur OpenOffi...	1 232 Ko
 valeurs_trimestrielles.csv	10/05/2021 18:54	Classeur OpenOffi...	11 Ko

Nous nous intéresserons principalement à *Valeurs_mensuelles* et *valeurs_trimestrielles*.

- **Caractéristiques.csv et famille_Tourisme-France** >> sont une description des différentes données qu'on va retrouver dans les trois autres fichiers.
- **Valeurs_mensuelles et valeurs_annuelles** >> contiennent les informations d'arrivées et de nuitées par département, région et un total au niveau de la France dans la colonne *Libellé*.
 - o La colonne *IdBank* fournit des informations sur le code de stockage des données au sein de l'INSEE.
 - o La colonne *Dernière mise à jour* indique la date de mise à jour de chaque information.
 - o Enfin, la *colonne Période et toutes les suivantes* contiennent soit mensuellement ou annuellement les valeurs des arrivées et des nuitées. Nous avons choisi de travailler avec les données mensuelles afin d'avoir le maximum d'informations.

A	B	C	D	E	F	G	H	I
Libellé	IdBank	Dernière mise à jour	Période	2010-01	2010-02	2010-03	2010-04	2010-05
Arrivées dans l'hôtellerie - Total - Départements d'outre-mer - Série arrêtée	171731006/03/2019 15:20			98.29	97.11	94.35	97.0	96.31
Codes				A	A	A	A	A
Arrivées dans l'hôtellerie - Total - France métropolitaine et Départements d'outre-mer - Série arrêtée	171731306/03/2019 15:20			6098.89	6691.69	7929.4	9034.22	10261.47
Codes				A	A	A	A	A
Arrivées dans l'hôtellerie - Total - France métropolitaine - Série arrêtée	171112306/03/2019 15:20			6000.6	6594.58	7835.05	8937.22	10165.17
Codes				A	A	A	A	A
Arrivées dans l'hôtellerie - Total - Guadeloupe - Série arrêtée	171721006/03/2019 15:20			28.64	31.13	28.09	27.52	25.46
Codes				A	A	A	A	A
Arrivées dans l'hôtellerie - Total - Martinique - Série arrêtée	171721306/03/2019 15:20			27.65	30.41	27.26	27.6	28.89
Codes				A	A	A	A	A
Arrivées dans l'hôtellerie - Total - Guyane - Série arrêtée	171721606/03/2019 15:20			9.09	9.29	10.62	9.36	8.45
Codes				A	A	A	A	A
Arrivées dans l'hôtellerie - Total - La Réunion - Série arrêtée	171721906/03/2019 15:20			32.9	26.27	28.38	32.52	33.51
Codes				A	A	A	A	A
Arrivées dans l'hôtellerie - Total - Ile-de-France - Série arrêtée	171111206/03/2019 15:20			2207.96	2201.94	2659.75	2617.92	2938.34
Codes				A	A	A	A	A
Arrivées dans l'hôtellerie - Total - Centre-Val de Loire - Série arrêtée	171110706/03/2019 15:20			166.87	189.11	249.37	315.19	385.39
Codes				A	A	A	A	A
Arrivées dans l'hôtellerie - Total - Bourgogne-Franche-Comté - Série arrêtée	173900106/03/2019 15:20			241.91	308.87	333.78	414.1	474.67
Codes				A	A	A	A	A

Capture d'écran du fichier valeurs_mensuelles

- **Valeurs_trimestrielles** >> structuré comme le fichier précédent (valeurs mensuelles), il contient les taux d'occupation des hôtels. Les données reportées ici le sont de manière trimestrielle.

A	B	C	D	E	F	G	H	I
Libellé	Bank	Dernière mise à jour	Période	2011-T1	2011-T2	2011-T3	2011-T4	2012-T1
Evolution des nuitées en glissement trimestriel (%) dans l'hôtellerie de plein air - France métropolitaine	10569953	14/11/2019 12:00			8.2	1.5		
Codes			O	E	A	O	O	
Evolution des nuitées en glissement trimestriel (%) dans l'hôtellerie - France métropolitaine	10569951	07/02/2020 12:00		2.2	3.1	3.0	4.6	2.3
Codes			A	A	A	A		
Evolution des nuitées en glissement trimestriel (%) dans les Autres Hébergements Collectifs Touristiques - France métropolitaine	10569952	07/02/2020 12:00						3.8
Codes								A
Nuitées dans l'hôtellerie en provenance - Amérique centrale et du Sud - France métropolitaine	10609683	28/05/2020 16:31		385	689	815	521	435
Codes			A	A	A	A	A	
Nuitées dans l'hôtellerie en provenance - Afrique - France métropolitaine	10609689	28/05/2020 16:31		219	275	314	283	262
Codes			A	A	A	A	A	
Nuitées dans l'hôtellerie en provenance - Allemagne - France métropolitaine	10609674	28/05/2020 16:31		764	1995	2296	1071	781
Codes			A	A	A	A	A	

Capture d'écran du fichier valeurs_trimestrielles contenant les taux d'occupation des hôtels

Les informations géographiques

Toujours sur le site web de l'INSEE, nous avons téléchargé deux fichiers contenant les noms des départements et régions françaises, ainsi que leurs codes INSEE.

Lien : <https://www.insee.fr/fr/information/4316069>

	A	B	C
1	region code	code departement	nom departement
2	84	1	Ain
3	32	2	Aisne
4	84	3	Allier
5	93	4	Alpes-de-Haute-Provence
6	93	5	Hautes-Alpes
7	93	6	Alpes-Maritimes
8	84	7	Ardèche
9	44	8	Ardennes

Capture d'écran du fichier département

	A	B	C
	code region	nom region	Territoire
	1	guadeloupe	Outre mer
	2	martinique	Outre mer
	3	guyane	Outre mer
	4	la reunion	Outre mer
	6	mayotte	Outre mer
	11	ile de france	France metropolitaine
	24	centre val de loire	France metropolitaine
	27	bourgogne franche comte	France metropolitaine
	28	normandie	France metropolitaine
	32	hauts de france	France metropolitaine

Capture d'écran du fichier régions

Le fichier des dates

Grâce au logiciel excel, nous avons généré un fichier contenant les dates de janvier 2011 à Décembre 2020 (cette période couvre toutes nos données disponibles). Nous aurions aussi pu le créer avec un script python mais il était plus simple et rapide de le faire sous excel.

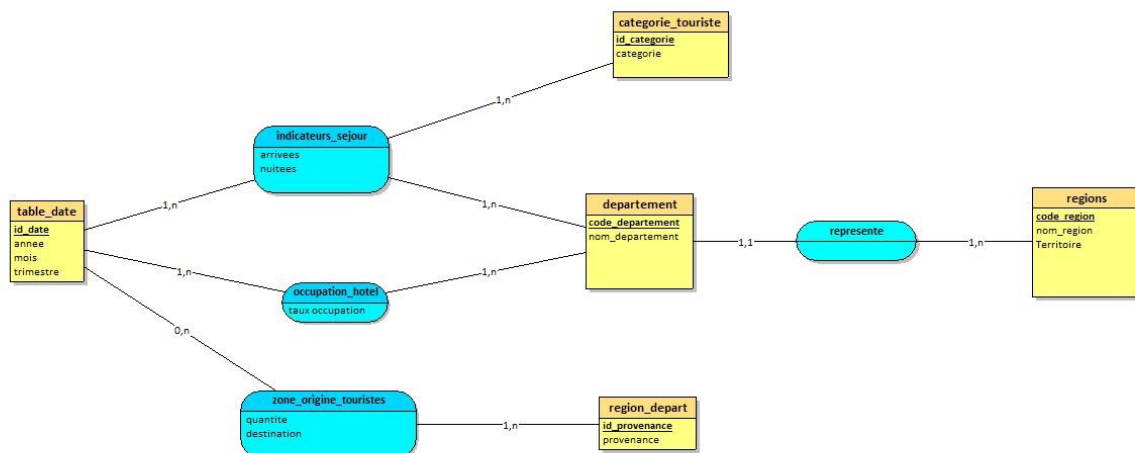
Dans ce fichier, la colonne id_date est un identifiant unique pour chaque date présente dans ce fichier.

id_date	annee	mois	trimestre
1	2011	janvier	1
2	2011	fevrier	1
3	2011	mars	1
4	2011	avril	2
5	2011	mai	2
6	2011	juin	2
7	2011	juillet	3

Capture d'écran de la table des dates

2. Modèle conceptuel de données (MCD)

Le modèle conceptuel des données (MCD) a pour but d'écrire de façon formelle les données qui seront utilisées par le système d'information. Il s'agit donc d'une représentation des données, facilement compréhensible, permettant de décrire le système d'information à l'aide d'entités (ainsi que les relations les cardinalités).



Notre MCD est composé de 5 entités :

- a) **Table_date** qui contient les attributs *année, mois et trimestre*. Les dates vont de janvier 2011 à décembre 2020. Chaque date est représentée par un identifiant unique >> *id_date*.
- b) **Categorie_touriste** qui permet d'identifier la typologie de touriste : français ou étranger. Chaque typologie a un identifiant unique qui est 1 ou 2.
- c) **departement** qui permet d'identifier le nom de chaque département et chacun ayant un code unique en France, nous l'avons choisi pour en faire l'identifiant unique.
- d) **regions** contient la liste de toutes les régions de France (métropole et départements d'outre-mer). L'attribut territoire renseigne s'il s'agit justement de la métropole ou non. Chaque région est représentée par un code unique (*code_region*).
- e) **region_depart** contient la liste des zones d'origines de touristes étrangers (pays, continent).

4 associations sont identifiables dans notre MCD et 3 d'entre-elles, étant des relations Many-to-many (plusieurs à plusieurs), vont générer des tables dans le modèle logique des données. Il s'agit de :

Indicateurs_sejour qui contient les attributs arrivées et nuitées. Dans le modèle logique, elle héritera d'une clé primaire composée par les identifiants uniques des tables ***categorie_touriste***, ***departement*** et ***table_date***.

Occupation_hotel qui contient le taux d'occupation des hôtels. Dans le modèle logique, elle héritera d'une clé primaire composée par les identifiants uniques des tables ***departement*** et ***table_date***.

Zone_origine_touristes dont l'attribut *quantité* indiquent le nombre de touristes arrivant en France selon la zone d'origine, ainsi que la destination (ici, c'est juste la France de manière globale. Il n'y a pas de granularités à plus petite échelle). Dans le modèle logique, elle héritera d'une clé primaire composée par les identifiants uniques des tables ***region_depart*** et ***table_date***.

L'association ***représente*** liant ***departement*** et ***regions*** étant comprise dans une relation many-to-one, ne va pas donner naissance à une table dans le modèle logique. Cependant la table département héritera d'une clé étrangère qui est la clé primaire de la table régions.

Explication des cardinalités de ce modèle conceptuel des données :

- **Relations Many-to-one ou plusieurs à un (ou vice versa)** : La base est liée à une seule target et la target peut être liée à plusieurs bases.

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

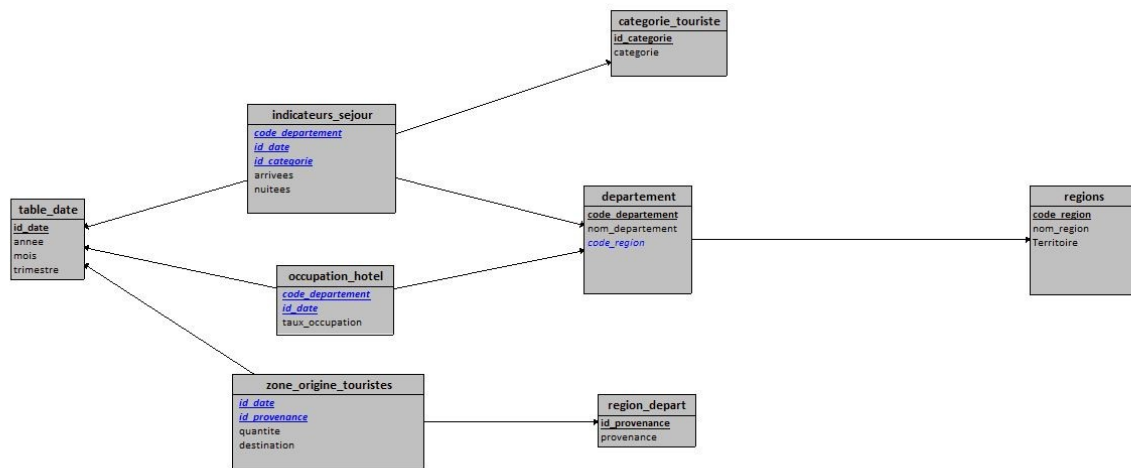
-

- **Relations Many-to-many ou plusieurs à plusieurs** : 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***

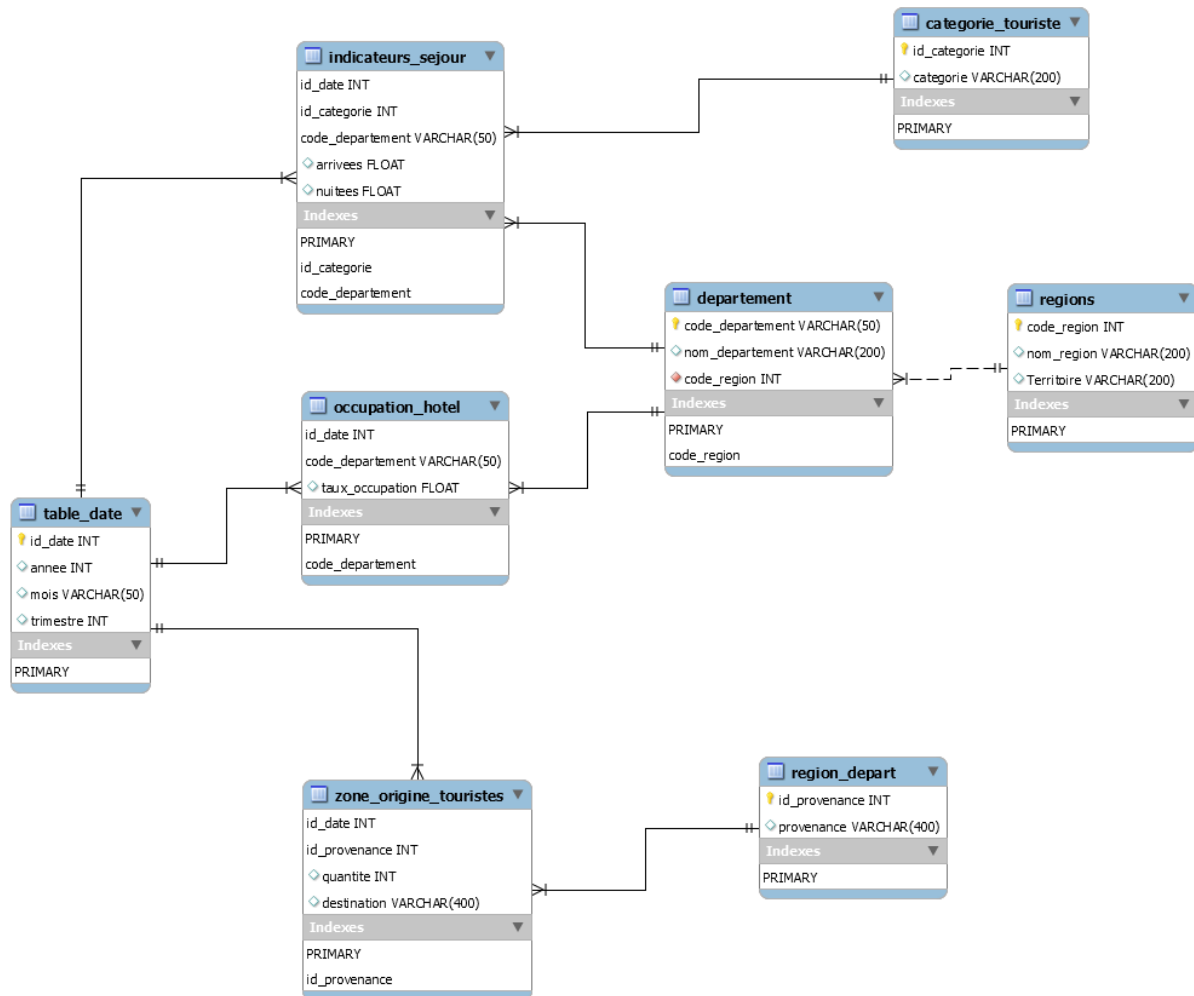
3. Modèle logique de données (MLD)

Dans le modèle logique des données, on représente chaque table avec les clés primaires et étrangères. C'est une vue plus concrète des tables telles qu'elles vont exister dans la base de données.



4. Modèle physique de données (MPD)

Généré après la création des tables dans la base de données, il 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.



5. Nettoyage des données et création des tables

Pour le traitement de nos données, nous avons travaillé avec le langage python grâce à Pandas et ses différentes librairies, dans l'environnement de développement Jupyter notebook.

Pour le nettoyage des données, voici ci-dessous les principales grandes étapes. Pour les détails, vous pouvez accéder au notebook (Notebook_Preprocessing) présent dans les documents annexes.

- Suppression des données avec le libellé Total, chiffres par région

Dans notre jeu de données, il était possible de retrouver le volume de touristes étrangers ou français soit au niveau national, soit par région et par département. Etant donné que dans notre modèle conceptuel de données la plus petite granularité est le département, et qu'il est essentiel que les données ne soient pas dupliquées, nous avons choisi de ne garder que les chiffres au niveau départemental.

Pour ce faire, nous avons utilisé la **méthode *str contains*** qui permet de rechercher une chaîne de caractères dans une autre.

```
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",
               ...])
```

Exemple : suppression des éléments avec le libellé Total

- Ensuite, à partir du dataframe ci-dessous, nous avons effectué les opérations suivantes (voir après image) :

```
Entrée [10]: 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
Taux d'occupation dans l'hôtellerie - Rhône	10598999.0	14/04/2021 18:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	47.4	48.0	46.9	37.7
Arrivées dans l'hôtellerie - Non-résidents - B...	10599339.0	14/04/2021 18:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	48.46	53.12	28.18	17.03
Nuitées dans l'hôtellerie - Résidents - Maine-...	10599235.0	14/04/2021 18:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	104.82	127.73	98.02	86.96

Dataframe avant d'autres manipulations.

- ✓ De la colonne Libellé :
 - Nous avons récupéré le nom du département pour le mettre dans une nouvelle colonne ;
 - Extraire les indicateurs (arrivées, taux d'occupation et nuitée pour les mettre dans une nouvelle colonne Type ;
 - Créer une nouvelle colonne avec la typologie de touristes (français ou étranger)
- ✓ Diviser la colonne date en 2 colonnes distinctes qui contiennent l'année et le mois. Cela nous évite de les avoir sur une ligne mais dans deux colonnes ;
- ✓ Afin de préparer la création des différentes tables, nous avons aussi rajouter :
 - Un id à chaque catégorie de touristes (2 : étrangers, 1 : français) ;
 - Le code Insee de chaque département.

Ci-dessous le dataframe intermédiaire obtenu :

```
df_mensuel.head()
```

Out[42]:

	annee	mois	nom_departement	categorie	arrivees	nuitées	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

Parmi les autres étapes, nous avons fait un « merge », qui signifie associer deux ou plusieurs tables en définissant les éléments communs. Dans le cas d'espèce, à partir de la table date, nous avons ajouté les id de chaque date.

Ensuite, enfin de garder des éléments uniques (date, département, catégorie), nous avons fait un group by (c'est une méthode qui permet de grouper des indicateurs).

```
Entrée [68]: >2= df_indicateurstourisme2.groupby(["id_date", "id_categorie", "code_departement"])[["arrivees", "nuitées", "taux_occupation"]].sum()  
>2.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
	78	3	62	0.00	0.00	66.8
	19	3	31	0.00	0.00	59.2
	81	2	5	6.84	9.35	0.0
	57	1	95	143.49	220.64	0.0
	94	2	52	3.92	5.22	0.0
	101	2	18	7.39	9.23	0.0

- Pour l'enregistrement des dataframe en csv, nous avons utilisé la méthode `pd.to_csv` en sélectionnant au préalable les colonnes nécessaires à la création de la table.

Exemple de la création du csv zone d'origine des touristes :

```
Entrée [100]: df_zoneoriginetouriste = df_fusionprovenance[['id_date', 'id_provenance', 'quantite', 'destination']]
df_zoneoriginetouriste.sample(10)
```

Out[100]:

	id_date	id_provenance	quantite	destination
	273	16	7	389 France métropolitaine
	1642	98	4	2089 France métropolitaine
	118	8	6	2225 France métropolitaine
	1136	69	5	3293 France métropolitaine
	1794	106	4	2715 France métropolitaine
	1614	94	12	390 France métropolitaine
	1512	88	12	350 France métropolitaine
	326	21	7	579 France métropolitaine
	1793	108	3	1306 France métropolitaine
	1630	95	17	703 France métropolitaine

```
Entrée [ ]: df_zoneoriginetouriste.to_csv('C:/Users/Alain NGABO/Dropbox/SIMPLON/Chefdoeuvre/data_propre/zone_origine_touristes.csv', index=False)
```

6. Insertion des données dans la base de données Mysql

Le script sql complet est disponible dans les annexes de ce document : Script_sql.sql

Pour la création et l'insertion des données, nous avons codé un script sql comme suit :

- Un « drop database if exists » au début de script afin de pouvoir le relancer si besoin sans qu'il n'y ait de problèmes de doublons ;
- La création de la base de données et des tables ;
- L'insertion des clés primaires ;
- L'insertion des clés étrangères ;
- Et enfin l'insertion des données.

Cette méthode permet d'optimiser la création de la base de données. La requête s'exécute automatiquement ligne par ligne, ce qui permet d'identifier rapidement une erreur et de la corriger si elle survient. Ci-dessous quelques captures d'écrans de ces différentes étapes :

```
#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)  
);
```

```
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);
```

```
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);  
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
```

7. Sécurité et sauvegarde de la base de données

- Sécurisation :

Pour la sécurité de notre base de données, nous avons créé un utilisateur qui n'a pas les privilèges d'administrateur. Cet utilisateur ne pourra que lire la base de données et ne pourra supprimer aucune donnée. Son accès est sécurisé par un mot de passe.

Voici le script en ligne de commande :

```
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 de notre base de données :

Pour la sauvegarde de notre base de données, nous avons effectué un dump (copie brute des données) en ligne de commande. Cela nous permet de restaurer à tout moment notre base de données.

```
(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>
```

A noter que notre script sql nous permet aussi de relancer si besoin la création de notre base de données.

8. Exemple de requête sql optimisée



Pour cet exemple, nous avons choisi une requête qui respecte ces conditions :

- ✓ Jointure sur 3 différentes tables
- ✓ Conditions et filtres : Where, group by, order by, limit, desc (décroissant)
- ✓ Opérateurs : round et sum
- ✓ Renommage des colonnes lors de l'affichage
- ✓ Utilisation des alias

```

#Quels sont les 5 départements qui ont accueillis le plus de touristes en 2020
use tourisme;
select nom_departement as "departement", round(sum(arrivees)) as "nombre de touristes" from departement as d
join indicateurs_sejour as i
on i.code_departement = d.code_departement
join table_date as t
on t.id_date = i.id_date
where annee = 2020
group by nom_departement
order by round(sum(arrivees)) desc
limit 5;

```

Result Grid   Filter Rows: <input type="text"/>		
	departement	nombre de touristes
▶	Paris	1834
	Bouches-du-Rhône	1091
	Alpes-Maritimes	1046
	Rhône	854
	Gironde	814

II – Datavisualisation et analyse des données

1. Rappel des objectifs

Dans cette partie, nous allons présenter comment a été développé notre rapport power bi, la connexion avec la base de données MySQL et la mise à disposition du Dashboard auprès des utilisateurs tout en veillant à la sécurité et la confidentialité des données.

Les objectifs de la datavisualisation à partir de nos données :

Observer en fonction du lieu (département, région) et de la périodicité (mois, trimestre, année), pour les touristes français et/ou étrangers :

- **Le volume de touristes accueillis ;**
- **Leur origine (clientèle française ou étrangère) ;**
- **Les pays de provenance ;**
- **La durée de séjour dans l'hôtel ;**
- **Le taux d'occupation des hôtels.** Outils utilisés (Power BI desktop et power bi service)

2. Outils utilisés (Power BI desktop et power bi service)

Comme détaillé en introduction dans la section « stack technique », nous avons choisi de travailler avec l'outil Power BI développé par l'entreprise Microsoft. C'est un outil assez complet et mis à jour assez souvent, qui intègre toute la chaîne de données (Extraction, transformation et mise à disposition des données).

La version pour ordinateur (Power BI Desktop) nous a permis de récupérer les données en se connectant à notre base de données et de transformer ces données afin de pouvoir faire les graphiques représentant les objectifs d'analyse listés précédemment.

Lors de la transformation des données dans Power BI desktop, nous avons dû effectuer plusieurs opérations. Par exemple, nous avons :

- Assigner la bonne catégorie aux données (géographiques, données calculées et non calculées...) ;
- Créer des mesures et colonnes calculées (ce sont des éléments qui n'étaient pas présents initialement dans notre base de données) ;
- Créer des cartes (exemple la carte représentant la France et ses régions outre-mer n'était pas présente initialement dans Power BI. Il a fallu rechercher sur des forums comment la créer).

Une fois le rapport power bi créé sur la version ordinateur, nous l'avons publié dans la version power bi service (en ligne) dans un espace dédié. Seules les personnes accréditées (ayant l'identifiant et le mot de passe peuvent y accéder).

3. Datavisualisation

Comment accéder au rapport ?

- Cliquer sur ce lien [ICI](#)
- Actualisez et entrer les codes d'accès ci-dessous :

email: alain.ngabo@agencengabo.onmicrosoft.com

Mot de passe : Portes1907*

Si vous disposez du logiciel POWER BI DESKTOP, vous pouvez aussi accéder au rapport annexé à de ce document.

Pour répondre aux objectifs assignés à la datavisualisation de notre jeu de données, notre Dashboard power bi est composé de 5 pages. Grâce aux icônes de navigation disponibles sur chaque page, il est possible de naviguer aisément dans ce rapport, d'une page à une autre.

Ci-dessous, un descriptif sommaire du contenu des pages :

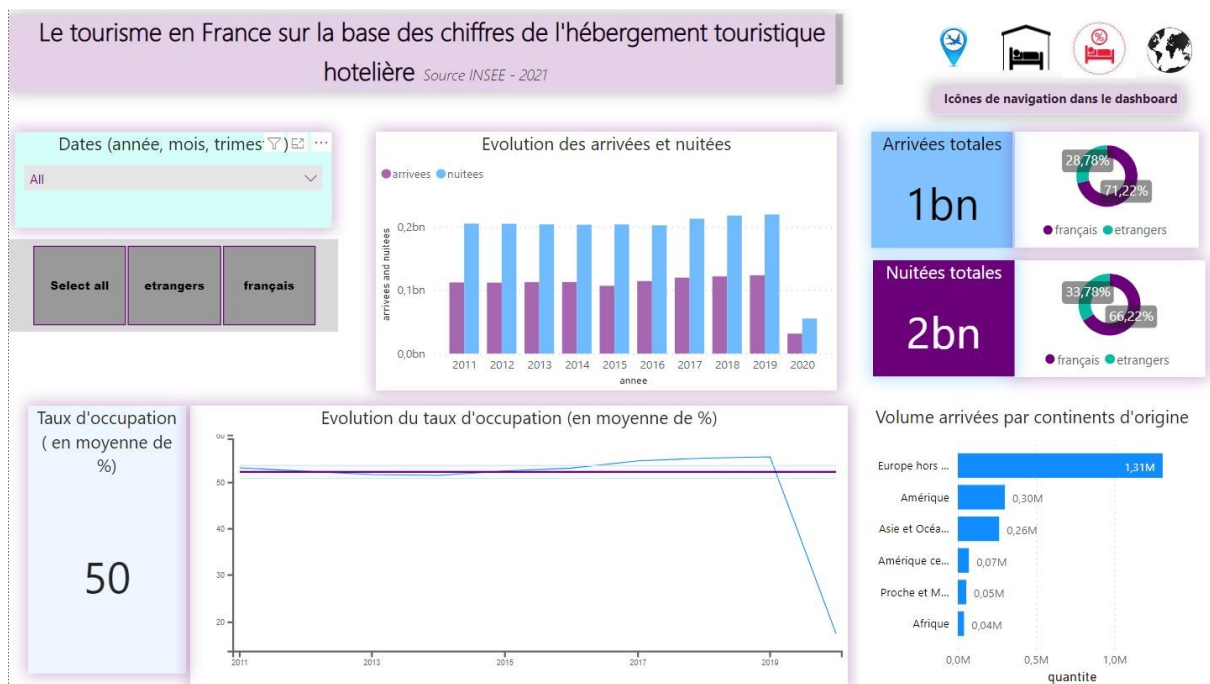
Page Accueil

Elément central de ce rapport, cette page permet à l'utilisateur d'avoir une vue globale sur tous les indicateurs.

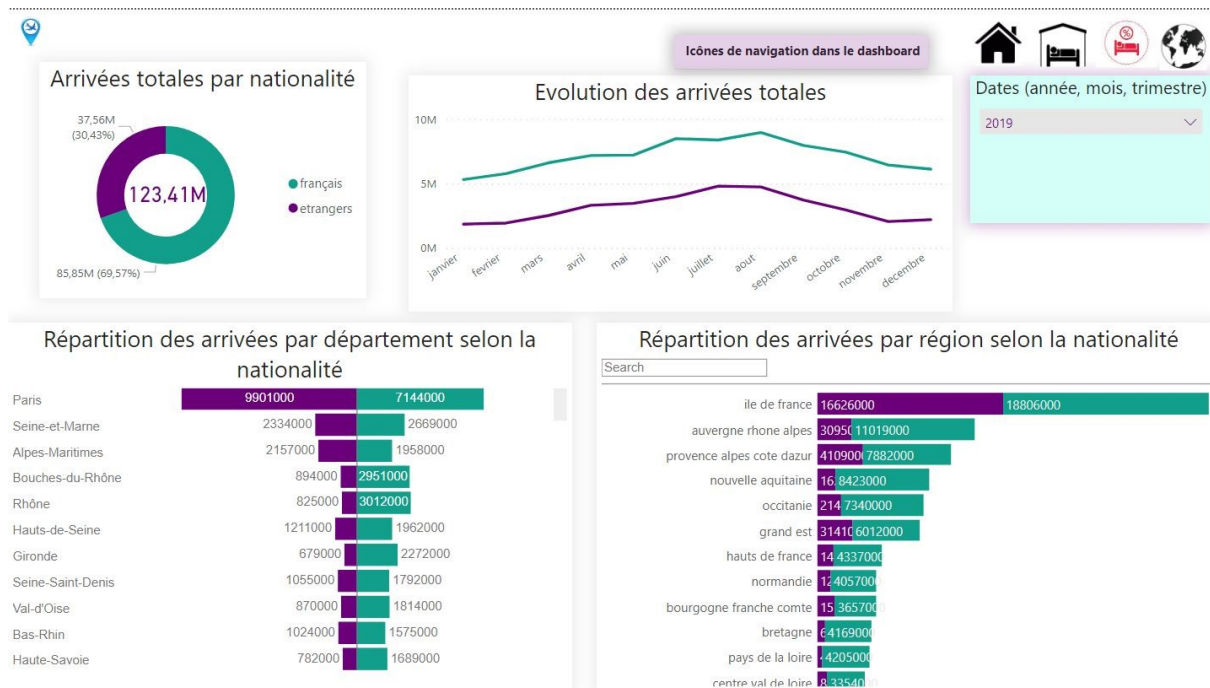
- ✓ Le graphique évolution des arrivées et nuitées lui indique l'évolution de ces 2 indicateurs de 2011 à 2020. Grâce à une navigation hiérarchique ou drill down, il peut aussi avoir une vue par mois ou trimestre ;
- ✓ Les graphiques « taux d'occupation » permettent de voir l'évolution de ce taux depuis 2011 ;
- ✓ Grâce au slicer « Dates », l'utilisateur peut isoler une période spécifique qui l'intéresse et les données des différents graphiques se mettent à jour ;
- ✓ Pareil que le slicer « Dates », celui avec SELECT ALL, ETRANGERS OU FRANÇAIS permet de sélectionner l'ensemble ou une typologie de touristes ;
- ✓ Les 2 graphiques à l'extrême droite en haut indiquent quant à eux les chiffres relatifs aux arrivées et nuitées totales. Il y a un détail en pourcentage par type de touristes. En cliquant sur l'un des 2 items (français ou étrangers), cette action aura pour effet d'isoler tous les chiffres sur la catégorie choisie ;
- ✓ Enfin, le graphique « volume arrivées par continents d'origine » indique la quantité de touristes étrangers selon les continents d'origine.

Les autres pages permettent d'explorer en détail les indicateurs

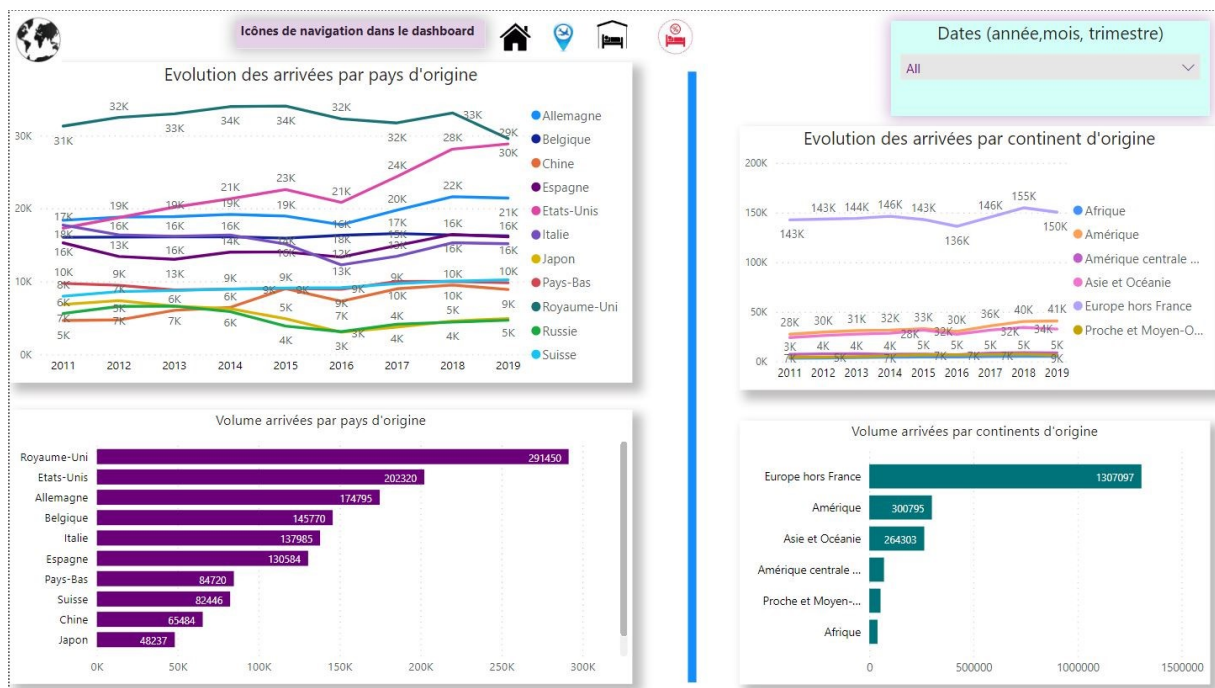
Capture d'écran de chaque page



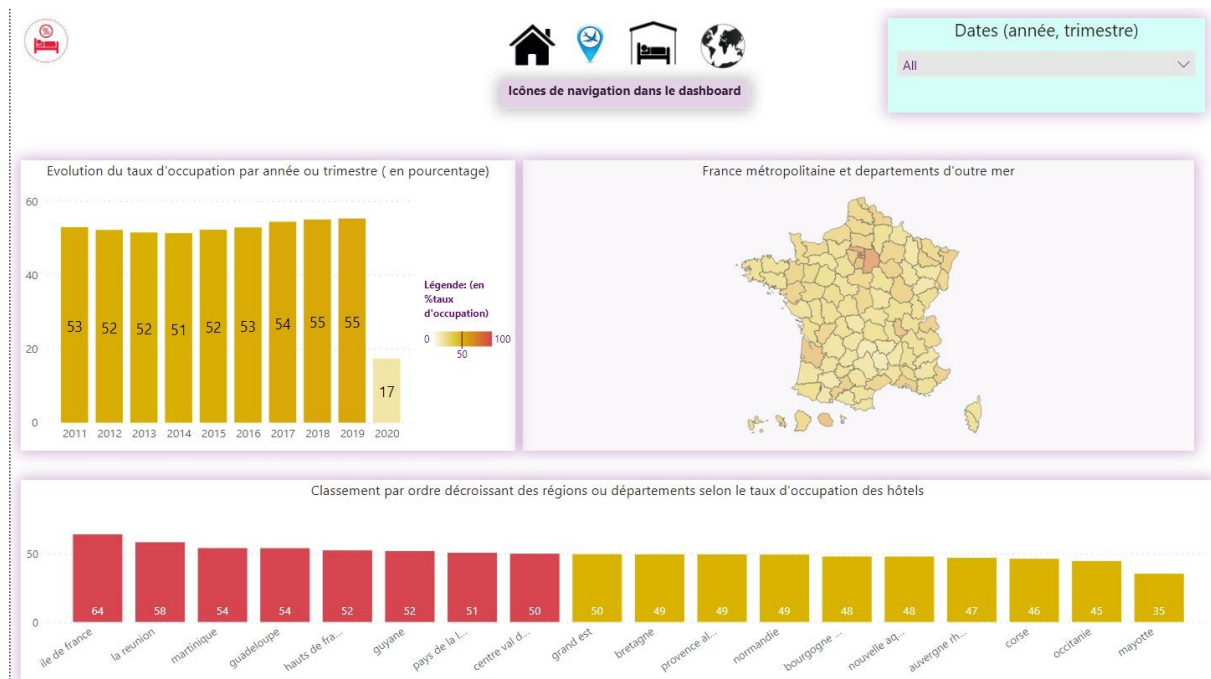
Page d'accueil



Arrivées



Pays et continents d'origine



Taux d'occupation

4. Automatisation de la récupération des données

Lors de la création de notre rapport Power BI Desktop, nous avons établi une connexion avec notre base de données (BDD). Lorsque nous faisons une mise à jour dans la BDD, la fonction refresh de Power BI desktop permet d'importer les nouvelles données et d'actualiser automatiquement notre rapport.

III – Gestion et bilan du projet

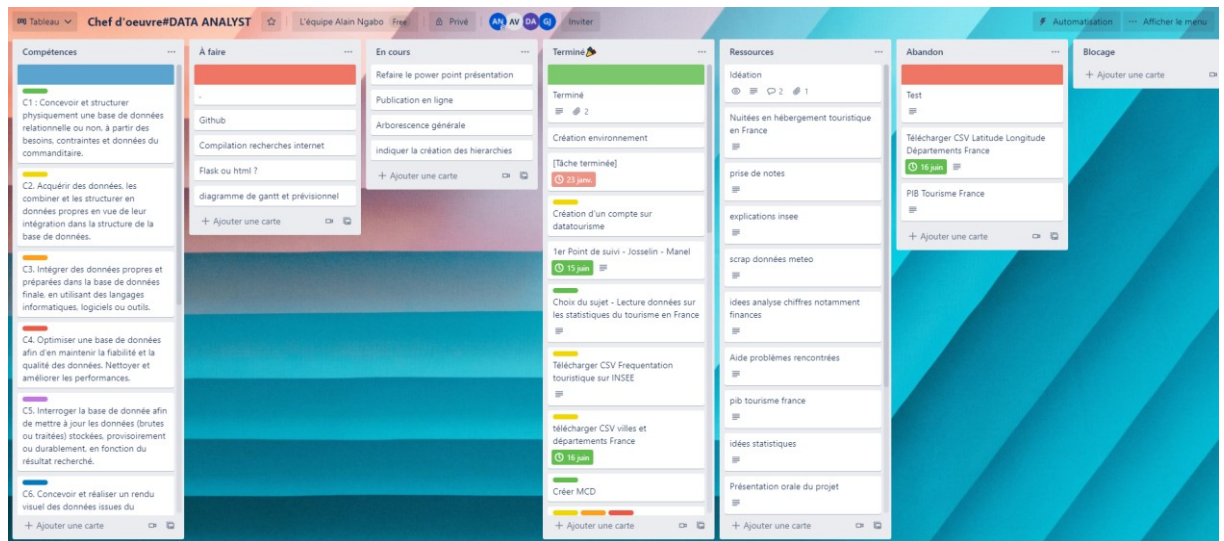
1. Trello

➤ **Lien d'accès :** <https://trello.com/b/7U8yePI2/chef-doeuvredata-analyst>

Le projet a été géré en mode agile (kanban) avec une planification sur l'outil Trello. Il y a 7 principales rubriques :

- **Compétences :** regroupe l'ensemble des compétences nécessaires pour valider les certifications en cours.
- **A faire :** regroupe de manière hiérarchique l'ensemble des tâches à faire. Celle qui est prioritaire est au-dessus. Chaque tâche correspond à une plusieurs compétences.
- **En cours :** tâches actuellement en cours. Tant qu'elles ne sont pas terminées ou déplacées (abandon, blocage), aucune autre tâche n'est tirée de la rubrique à faire.
- **Terminé :** regroupe l'ensemble des tâches terminées.
- **Ressources :** cette rubrique permet de retrouver pour chaque sujet ayant nécessité une recherche d'informations, de retrouver les liens ou références.
- **Abandon :** regroupe l'ensemble des tâches abandonnées.

- **Blocage** : regroupe l'ensemble des tâches nécessitant une aide extérieure afin de les résoudre (recherches internet, collègues). Ces tâches sont prioritaires et cette rubrique est limitée à 3 tâches maximales en même temps.



2. Difficultés rencontrées

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

Pour le preprocessing des données, nous avons fait appel à plusieurs méthodes et fonctions que nous n'avions pas auparavant utilisées. Pour trouver des solutions, nous avons consulté la documentation technique ainsi que des sites spécialisés parmi lesquels Stackoverflow et GITHUB.

Exemple ci-dessous : consultation de stackoverflow pour la méthode get dummies de python.

I have a dataframe like below. The column `Mfr Number` is a categorical data type. I'd like to preform `get_dummies` or one hot encoding on it, but instead of filling in the new column with a 1 if it's from that row, I want it to fill in the value from the `quantity` column. All the other new 'dummies' should remain a 0 on that row. Is this possible?

	Datetime	Mfr Number	quantity
0	2016-03-15 07:02:00	MWS0460MB	1
1	2016-03-15 07:03:00	TM-120-6X	3
2	2016-03-15 08:33:00	40.50699.0095	5
3	2016-03-15 08:42:00	40.50699.0100	1
4	2016-03-15 08:46:00	CXS-04T098-00-0703R-1025	10

python pandas one-hot-encoding

Share Improve this question Follow

asked Mar 20 '19 at 23:56
 Chris Macaluso
956 2 7 23

Add a comment

3 Answers

Active Oldest Votes

Do it in two steps:

```
dummies = pd.get_dummies(df['Mfr Number'])  
dummies.values[dummies != 0] = df['Quantity']
```

3. Axes d'améliorations du projet

Pour aller plus loin et mieux peaufiner ce projet, il serait intéressant :

- D'ajouter des indicateurs économiques du tourisme dans chaque département afin de voir son impact réel ;
- Pour les arrivées des étrangers, enrichir la liste des pays de provenance et leur département de destination en France ;
- Ajouter les autres types d'hébergements (camping, plateformes comme airbnb...) pour comparer l'évolution des types d'hébergements touristiques ;
- Enfin, héberger la base de données dans un cloud afin d'y donner l'accès à un plus grand nombre de personnes.

En outre, la présentation du dashboard pourrait être améliorée et certains axes d'analyses plus détaillés.

Remerciements

Au terme de ces sept mois de formation, j'ai vécu une belle aventure non seulement académique, mais aussi humaine. J'ai une profonde gratitude pour :

Nos formateurs (Manel, David, Josselin, Nicolas) et les différents intervenants professionnels ;

Les autres membres de Simplon (Ana, Alexia, Gilles et Julien) qui ont toujours été à nos côtés pour nous motiver, s'occuper de la partie pédagogique et insertion professionnelle ;

Mes camarades de promotion pour le soutien mutuel ;

Et enfin mon mentor, Jean-François, pour son aimable écoute, ses conseils avisés et sa très grande disponibilité.

Bibliographie/webographie

Données téléchargées

Fréquentation touristique (nuitées, arrivées) Insee, <https://www.insee.fr/fr/statistiques/series/113990189>, consulté le 20 Juin 2021

Contours des départements français, Data Gouv, <https://www.data.gouv.fr/fr/datasets/contours-des-departements-francais-issus-d-openstreetmap/>, consulté en juillet 2021

Carte des départements français, Data Gouv, <https://www.data.gouv.fr/en/datasets/carte-des-departements-2-1/>, consulté en juillet 2021

Sites consultés

<https://www.entreprises.gouv.fr/fr/tourisme/conseils-strategie/datatourisme-plateforme-open-data>

<https://www.economie.gouv.fr/cedef/statistiques-officielles-tourisme>

<https://atlasocio.com/classements/economie/tourisme/classement-etats-par-nombre-de-touristes-etrangers-monde.php>

<https://stackoverflow.com/questions/55271858/pandas-get-dummies-with-value-from-another-column>

<https://phoenixnap.com/kb/how-to-backup-restore-a-mysql-database>

<http://lamaisondeladonnee.bwabwanet.fr/2020/06/05/trier-les-mois-dans-un-graphique-ou-un-segment/>