Projet chef d'œuvre

CollectSpirit

Certification Simplon 2021

Développer data

Apprenant :
Mayel Pèllé
Années :
2020-2021

Sommaire

INTRODUCTION

Nous allons aborder dans ce rapport la thématique de la gestion d'une collection de spiritueux appliquée à un travail technique lié à la data. Le but est de concevoir une base de données permettant à un utilisateur de créer rapidement l'inventaire de sa collection.

L'idée du projet vient de mon histoire personnelle. J'ai travaillé pendant trois ans pour un producteur de rhum en Martinique appelé HSE. Cela m'a permis d'apprendre énormément sur l'univers des spiritueux. Et ce qui était d'abord un travail, s'est transformé petit à petit, de dégustation en dégustation, en passion. Au fil du temps, j'ai commencé à acquérir des bouteilles de plus en plus prestigieuses et ai ainsi pu constituer une petite collection d'une centaine de références. Lorsqu'il a fallu déménager en France, j'ai dû établir « à la main » un inventaire de ma collection à l'aide d'un fichier Excel. La tâche s'est révélée plus ardue que prévu. C'est à partir de cet événement qu'a germé l'idée de créer une application pour faciliter ce travail. Je profite de cette formation pour me lancer dans ce projet.

Dans un premier temps, rappelons que le rhum (anglais : rum, espagnol : ron) est une eau-de-vie originaire des Amériques. Elle est produite par distillation soit de sous-produits fermentés de l'industrie sucrière (aussi appelé mélasse : le rhum industriel ou traditionnel), soit à partir du jus de canne à sucre fermenté (le rhum agricole). [Source Wikipédia].

Le secteur du rhum a beaucoup évolué ces dernières années, tant en volume de vente qu'en production ? Considéré pendant longtemps comme un alcool bas de gamme, il est désormais jugé comme un spiritueux aussi noble que le whisky ou le cognac.

Ce changement n'a pas échappé aux collectionneurs qui sont de plus en plus nombreux à partir à la découverte des joyaux des tropiques. Mais le monde du rhum excède largement les frontières des caraïbes. Disposer d'une base de données sur le rhum sera à la fois un moyen de réaliser l'inventaire de sa collection mais permettra aussi de découvrir de nouvelles références.

Pour ce projet, nous allons essayer de répondre à des problématiques tournant autour de la création d'une base de données liée aux spiritueux et à la gestion d'une collection. Pour ce faire, nous allons analyser les envies et les attentes d'un collectionneur.

Ainsi, nous verrons dans ce rapport les différentes étapes de la construction de ce projet. Elles sont au nombre de trois :

- La préparation du projet : la phase de planification et de recherches
- La conception du projet : construction de la base de données
- L'exploitation des données : l'applicatif.

Nous allons ici nous focaliser essentiellement sur le rhum, car c'est mon domaine d'expertise. Mais je prévois, pour des versions ultérieures, d'intégrer d'autres alcools.

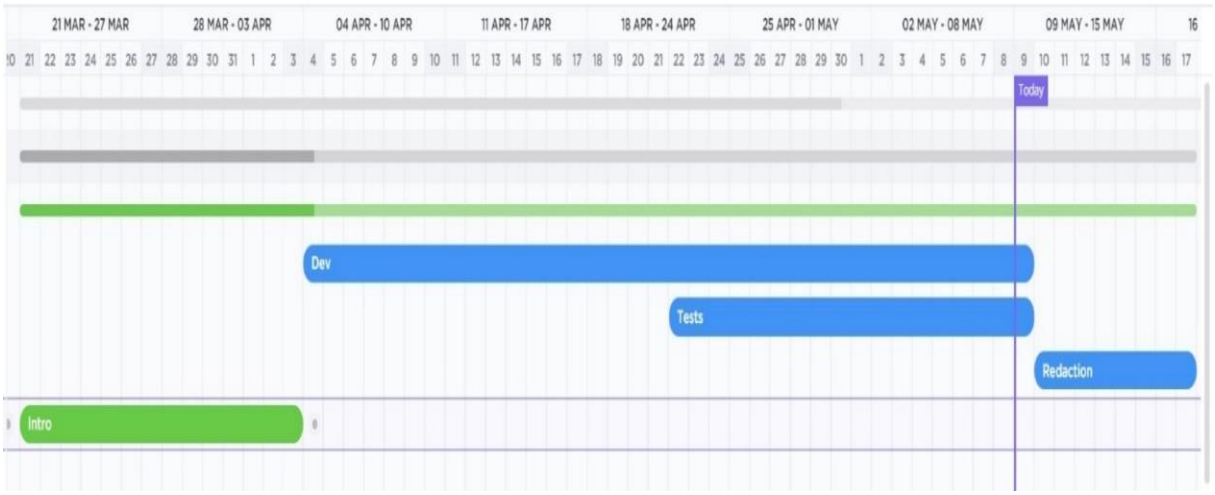
PREPARATION DU PROJET

Article I. La phase de planification

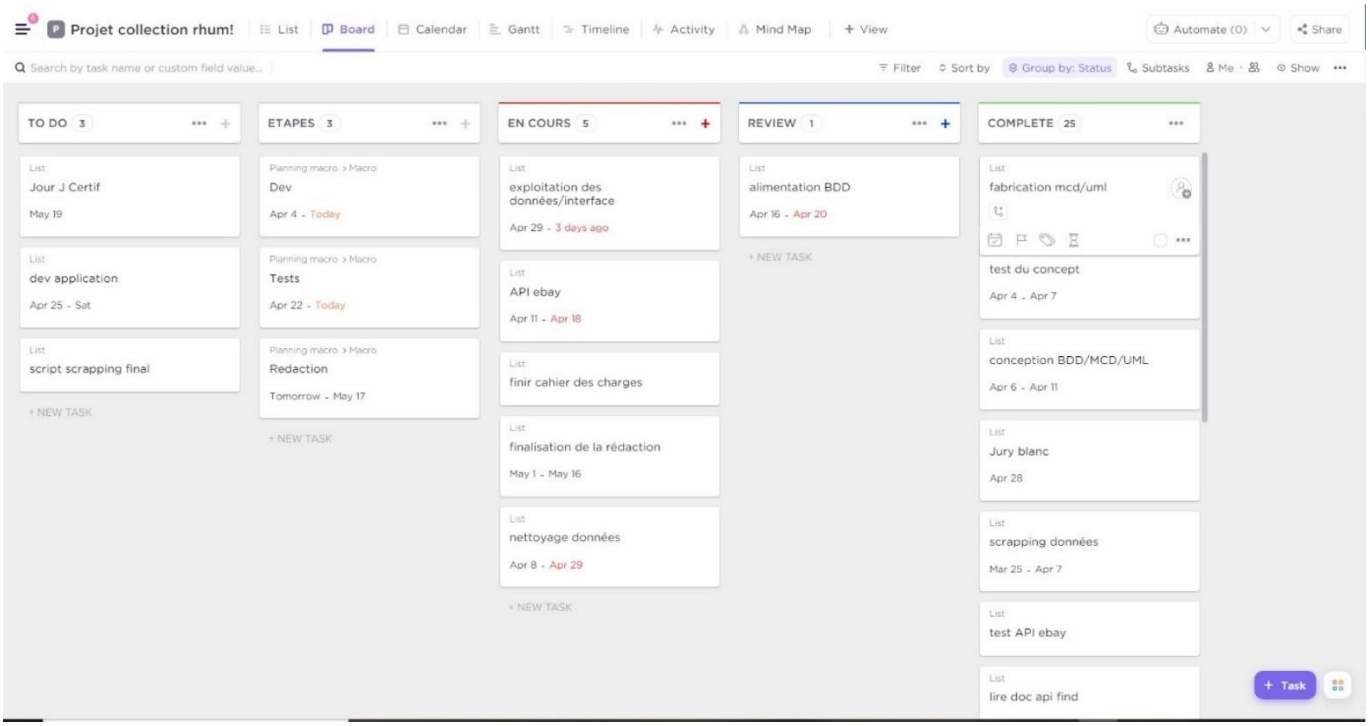
Notre projet de certification est un projet individuel dont le sujet est libre. La première étape a donc été de faire des recherches préliminaires pour déterminer le sujet et sa faisabilité.

Après la validation du sujet par nos formateurs, le premier challenge a été de déterminer le planning prévisionnel du projet. En tant que novice, il n'est pas évident d'évaluer le temps à consacrer à chaque tâche. Pour m'aider, j'ai utilisé l'application Clickup pour créer mon planning. Elle dispose de nombreuses fonctionnalités : calendrier, graphique de gant, todo-list… On a également implémenté des doses méthode Agile dans le suivi du projet (Kanban, review).

J'ai commencé par définir les grandes étapes du projet en trois grandes phases modélisées dans un graphique de Gant : la phase de recherche, celle de développement et de tests et enfin celle de rédaction.



Dans un second temps, j'ai essayé de définir plus finement les tâches à accomplir. Pour chacune d'entre elles j'ai estimé leurs priorités, leurs difficultés et le temps nécessaire à les accomplir. J'ai pu grâce à clickup, créer un tableau en m'inspirant de Kanban. Grâce à ce tableau je pouvais visualiser les tâches en cours ou celles qui restent à faire par exemple.



Cette approche, en deux temps, m'a permis d'avoir une vision globale du planning et une meilleure vision du temps à ma disposition pour réaliser ce projet.

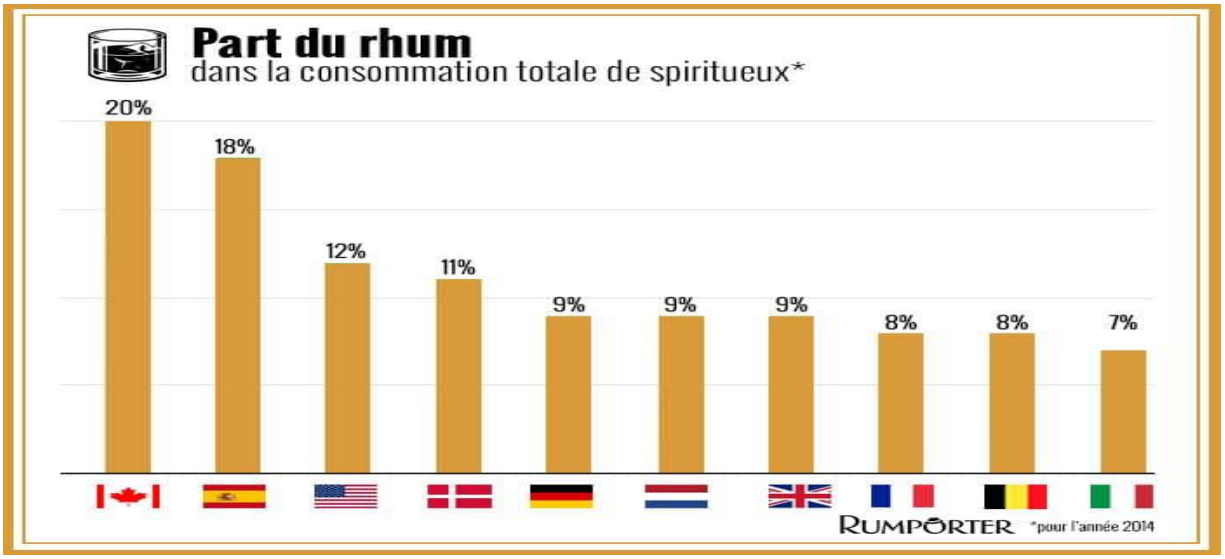
Pour la gestion quotidienne de mon projet, je notais simplement les tâches du jour sur un calepin.
Je contactais au moins deux fois par semaine mes formateurs, afin de leur faire un retour sur l'avancée de mon projet, les difficultés rencontrées et les solutions trouvées. Je les interpellais ponctuellement lorsque je rencontrais un problème. Ce contact régulier était essentiel, nous étions principalement en distanciel, pour s'assurer que l'on avance dans la bonne direction et aussi pour rester motivé.

Malgré ce travail, on s'aperçoit qu'on a toujours sous-estimé certaines tâches. Pour ma part, j'ai passé plus de temps sur la préparation de mon projet et sur la partie de nettoyage de données que prévu.

Article II. L'expression du besoin

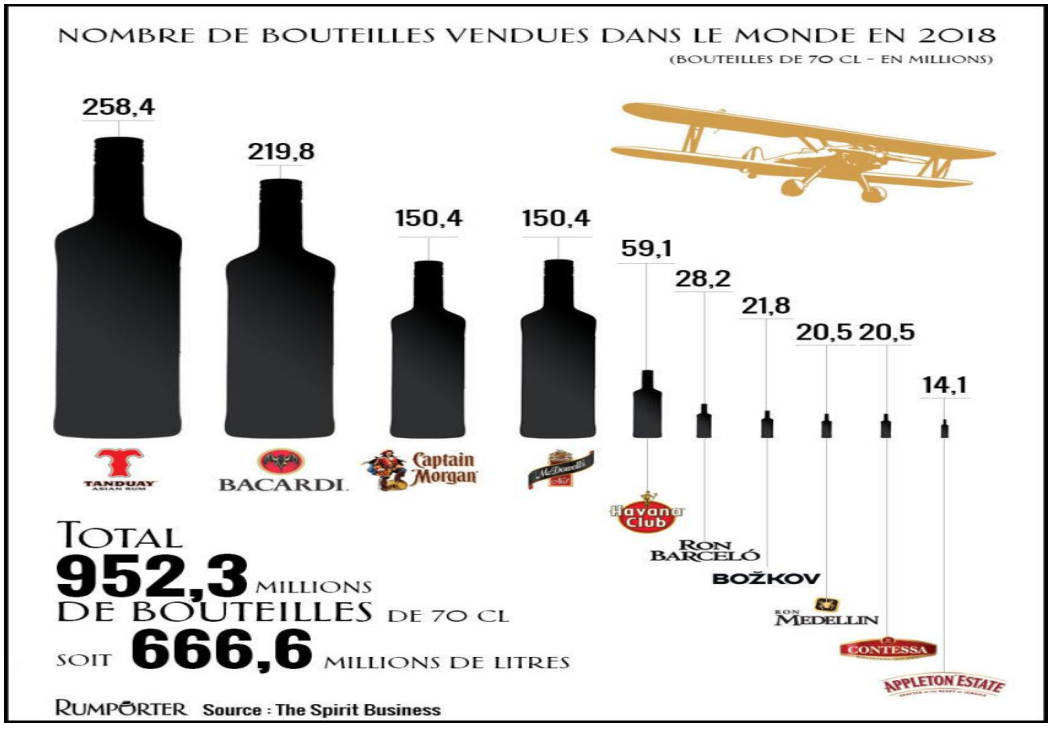
(a) La progression du marché du rhum

Le rhum est le troisième spiritueux le plus consommé au monde, derrière le whisky et le brandy. C'est le spiritueux ayant connu la plus forte croissance au cours de ces dix dernières années (+ 40 %). 536 000 tonnes de rhum sont bues chaque année dans le monde, soit 17 litres par seconde.



En France, le rhum représente 8 % de la consommation totale de spiritueux avec 34 millions de litres consommés et il reste un des principaux moteurs de croissance de la catégorie alcools. La France est d'ailleurs le deuxième plus grand consommateur de rhum en Europe derrière l'Espagne.
La production de rhum français, exclusivement dans les départements français d'Outre-Mer représente plus de 55 millions de litres par an :

- 20 % de cette production est consommée sur place (DOM)
- 30 % du rhum est exporté en Europe et dans le Monde
- 50 % du rhum est consommé en MétropoleLe rhum industriel représente aujourd'hui environ 90% de la production mondiale de rhum. Les principaux producteurs de rhum sont Tanduay (Philipine) et bacardi (Bermudes).



Les chiffres le montrent, le rhum séduit de plus en plus de consommateurs. Il n'est plus cantonné au ti-punch et aux mojitos. Il intéresse de plus en plus les mixologues ou les amateurs de spiritueux. Il fédère une communauté d'amateur de plus en plus grande. Par exemple, les groupes français sur Facebook de la Confrérie du Rhum et du Rhum Arrangé comptent respectivement 45 000 et 150 000 membres.

Quelles sont les raisons de ce succès ? Premièrement, le goût, on trouve des rhums à la qualité de fabrication et la complexité gustative qui rivalisent avec les plus grands cognacs ou scotchs. Deuxièmement, on trouve une grande diversité dans la fabrication de rhum, que ce soit au niveau de la production que du vieillissement. On utilise différents types d'alambics (colonne, pot still, double copper, colonne hybride, …), de levures (boulangères, œnologiques, indigènes, …), de types de fûts (style de bois, chauffe, …). Il y a beaucoup de dynamisme dans le secteur. La Martinique, productrice de rhum agricole, constitue un des fleurons actuels du secteur, étant la première au monde à posséder une AOC. Cette diversité permet de toucher un public plus large. Enfin, autre avantage de taille, le prix ! Le whisky est un marché mature qui subit une inflation galopante surtout les créneaux premium. Le rhum, même pour ses bouteilles les plus prestigieuses, reste plus abordable que les whiskys de même gamme.

(b) Les collectionneurs de spiritueux

Ce projet s'adresse aux collectionneurs de spiritueux. Mais qu'est-ce qu'un collectionneur de spiritueux ? Quelles sont ses envies ? Sur quels critères se base-t-il pour acheter ?

Il serait difficile de définir le collectionneur type de rhum. Il y a diverses raisons de commencer à acquérir des bouteilles. On peut être naturellement amateur de spiritueux ou spéculateur ou juste attiré par l'esthétique d'un flacon. Personnellement, j'ai commencé à acquérir des bouteilles par curiosité technique (production et gustative).

Essayons de déterminer les attentes d'un collectionneur moyen qui cherche à agrandir sa collection. De manières générale, un collectionneur fera attention à la réputation de la marque et/ou de la bouteille, au prix, à la qualité de conservation de la bouteille, à la rareté de la référence, à l'esthétique du flacon, …

J'ai pu rencontrer des collectionneurs plus exigeants qui vont faire attention au numéro de série, à la date de mise en fût et de mise en bouteille (surtout pour les références de rhum vieux millésimé car cela donne l'âge du spiritueux) ou qui vont collectionner toutes les références d'une seule marque.

Une particularité intéressante des collectionneurs d'alcool, c'est qu'ils vont souvent acquérir deux exemplaires d'une référence qui leurs plaisent. L'une pour la garde, l'autre pour l'ouvrir.

Le graal pour un collectionneur c'est d'acquérir une bouteille rare, ancienne, qui jouit d'une excellente réputation gustative, bien conservée et à un prix raisonnable.

Donc, en conclusion, le collectionneur est une personne qui est avide de renseignements. La question est de savoir, si toutes ces informations sont pertinentes à être collectées pour notre projet.

Il est important de préciser, que le collectionneur est souvent discret sur la teneur réelle de sa collection. Il y a plusieurs raisons à cela, par exemple pour éviter les vols ou les demandes incessantes d'autres collectionneurs.

La sécurité et la confidentialité des informations seront au cœur du développement de ce projet.

(c) Traduction du besoin

L'importance et la granularité des informations nécessaires pour satisfaire la curiosité d'un collectionneur peut être primordiale. Nous allons voir ici quels sont les types d'informations que nous sélectionnerons pour notre projet.

Comme nous l'avons dit, l'un de nos buts est de créer une base de données de références. Notre objectif est la gestion d'une collection et pas forcément d'être un Wikipédia du rhum.

Nous pourrions intégrer beaucoup d'informations dans notre base de données. Cependant, beaucoup d'entre elles seraient difficiles et fastidieuses à obtenir (comme par exemple le numéro de la bouteille, le numéro de l'embouteillage ou du batch pour une référence).
Par exemple, la base de données la plus complète sur le rhum est celle de wikirum. Ils ont passé plus de deux ans à la constituer.

Pour notre projet, nous allons nous concentrer sur les informations générales. Des informations pas forcément exhaustives, mais suffisantes pour permettre de constituer une première base de données.

Les informations seront :

- Le nom de la référence,
- La marque/le producteur,
- Le volume,
- Le degré,
- L'origine.

J'ai choisi ces informations car elles me permettent déjà de donner des renseignements précis sur une référence et elles sont plus faciles à obtenir.

Des informations comme la réputation, la conservation ou la qualité d'un produit sont trop subjectives. Des informations comme le numéro de la bouteille, la date d'embouteillage ou de mise en fût sont, elles, fastidieuses à trouver. C'est pour cela que je n'intégrerai pas directement ce type

d'informations très précises. Par contre, nous pouvons prévoir la possibilité à l'utilisateur d'ajouter des informations sur une référence.

(d) Les applications existantes

Il existe déjà plusieurs applications dans le domaine des vins et spiritueux. Celles-ci permettent de gérer une collection, d'obtenir des informations sur des références, d'écrire des notes de dégustations, … Pour les spiritueux, ce sont surtout des applications concernant le whisky que l'on trouve.

Il existe actuellement deux applications autour de la gestion de collection spécifique au rhum qui ont été lancées en 2020 et un site web américain plus ancien :

- Wikirum : c'est une web application construite avec Angular qui est la base de données la plus complète sur le rhum. C'est un projet porté par des nantais de Saint Herblain. Ils travaillent directement avec les distilleries et les marques pour obtenir des données. C'est avant tout un wikipédia du rhum. Il permet également de se constituer une collection et d'acquérir des bouteilles.

- RumX (anciennement rum tasting note) : c'était en premier lieu une application qui permettait de partager des notes de dégustation avec sa communauté. Dans sa nouvelle version, on peut également faire un inventaire de sa collection et acquérir de nouvelles bouteilles via un site d'enchères. C'est actuellement l'application qui offre le plus de fonctionnalités.

- Rumratings : c'est un site américain qui contient plus de 7000 références et qui permet de donner une note. Le site n'est actuellement plus activement maintenu. Ses deux créateurs cherchant à diversifier leurs activités vers d'autres alcools.

CONCEPTION DU PROJET

Article I. L'application

(a) Traduction du besoin client

Afin d'avoir une meilleure vision du projet, j'ai réalisé un schéma du projet. Celui-ci a beaucoup évolué au fur et à mesure de l'avancement du projet. Ce schéma représente l'architecture du projet. Nous avons à gauche les sources des données, au milieu la modélisation de la base de données et à droite la partie exploitation de celle-ci.

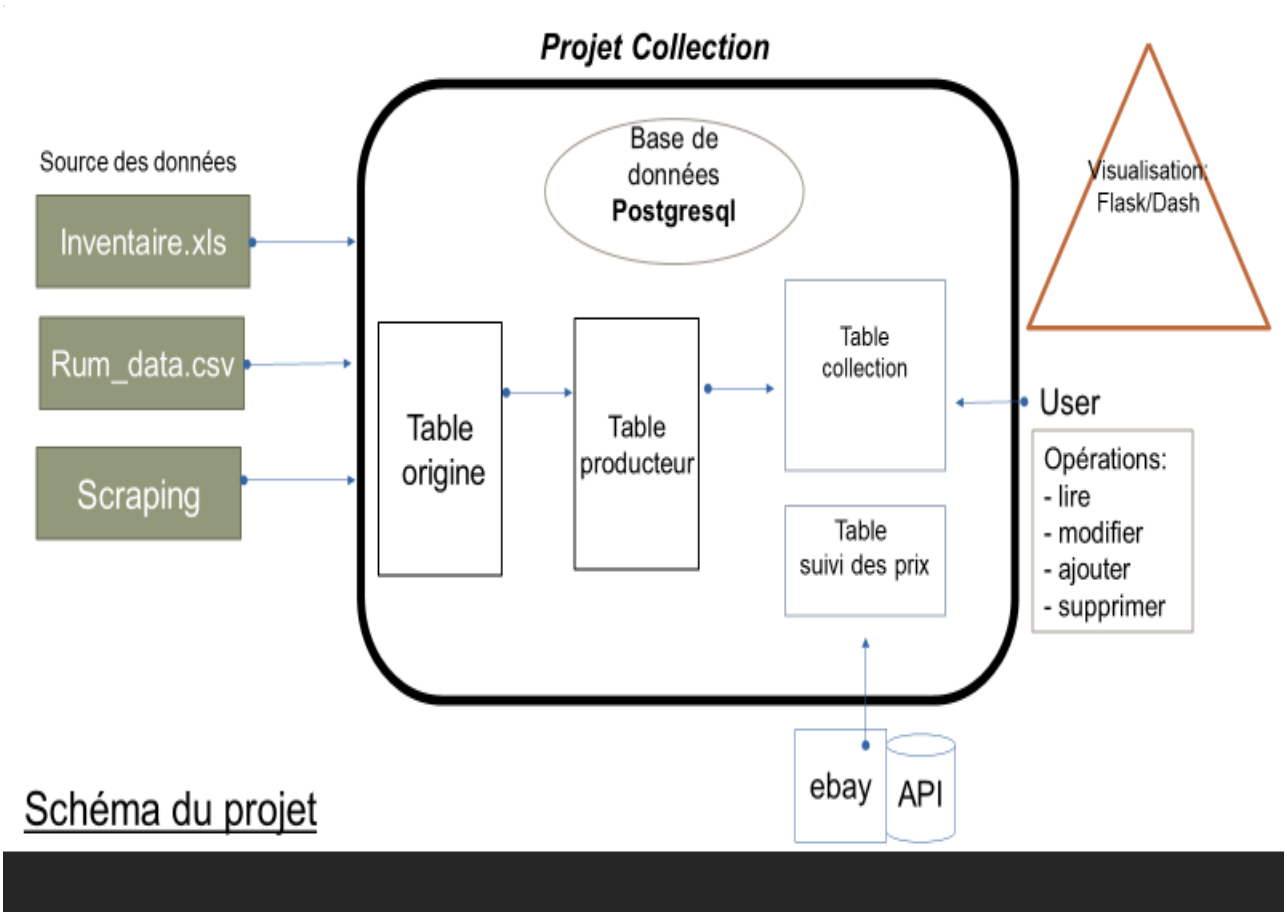
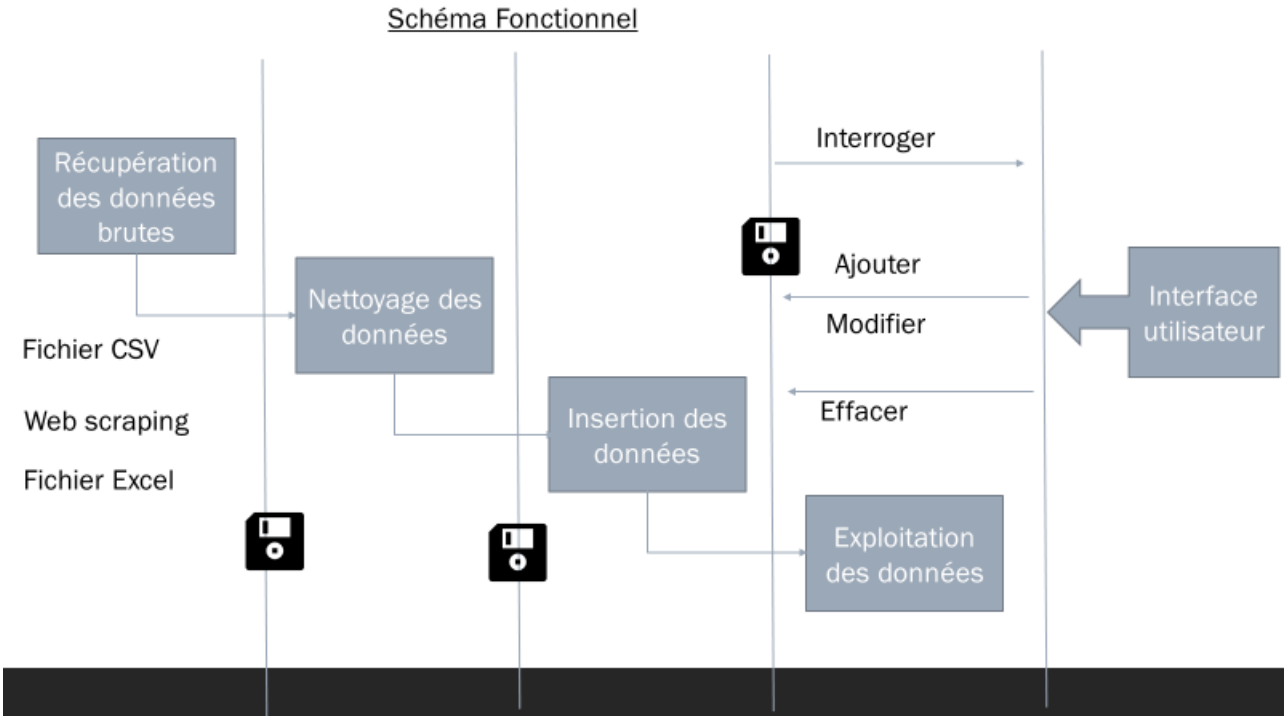


Schéma du projet

J'ai également réalisé un schéma fonctionnel qui, lui, modélise les étapes de réalisation du projet qui seront vu dans la suite de ce rapport.



(b) Choix des technologies

Voici un tableau représentant l'ensemble des technologies choisies pour la mise en œuvre du projet.

Langages	**Bibliothèques**	**Outils**
Python3.8	Pandas==1.2.3	Jupyter lab==3.0.12
SQL	Psycopg2==2.6.8	PgAdmin 4 V5.2
Nosql (JSON)	Plotly-express==0.4.1	Microsoft office 2019
	Folium==0.12.1	Exiftool==12.2.1.0
	Scrapy==2.4.1	VS Code==1.56.1
		Postman==8.3.1
		Looping.exe==3.0.1

		Bitwarden==1.26.23

Après analyse du besoin, j'ai décidé de proposer cette solution :

- Création d'une base de données relationnelle sous PostgreSQL.

- Création d'un dashboard pour visualiser les informations de la base de données avec Dash.

Article II. Les sources de données

(a) La collecte des données

Pendant la phase de recherche, j'avais prévu de récupèrer les données uniquement par scrapping. Je comptais scraper des sites spécialisés ou des sites d'e-commerce pour me constituer rapidement un inventaire exhaustif.

J'ai fait des essais avec les libraires python BeautifulSoup et Scrapy. Le scraping s'est révélé moins fructueux que prévu. Soit la donnée collectée était peu exploitable (informations incomplètes et parcellaires) soit l'automatisation du scrapping se révélait fastidieux (code html mal structuré). J'ai quand même pu obtenir des données exploitables grâce à la librairie Scrappy. Elle s'est révélée être plus difficile à appréhender que BeautifulSoup, bien que plus performante. Elle offre un véritable micro-framework pour le scrapping. J'ai suivi un tutoriel sur youtube qui m'a permis de scrapper le site e-commerce Christian de Montaguère. (https://www.youtube.com/watch?v=s4jtkzHhLzY&t=1216s)

J'ai pu récolter des jeux de données, de manière plus classique, auprès de différentes sources. En voici le descriptif :

- Test.json : fichier contenant 596 références issue du scraping du site Christian de Montaguère avec la librairie scrapy.

- Inventaire.xls : l'inventaire de ma propre collection établi lors de mon déménagement dans un fichier excel. Etant des données privées je peux librement les utiliser. Cet inventaire contient 89 références.

- Rum_data.csv : jeu de données en format csv trouvé sur kaggle (site qui contient des dataset opensource). Il contient une liste de 6000 rhums.

- Rhumfullinfo_wikirum.csv: fichier csv obtenu grâce à l'aimable participation du site wikirum. Il contient 1000 références.

- Api_ebay : je prévois de récolter l'évolution des prix grâce à l'API Find d'Ebay, cette fonctionnalité sera intégrée dans les futurs versions. J'ai déjà pu me procurer une clé d'identification pour développeur.

(b) RGPD

A l'exception des données obtenues par scrapping, tous les autres jeux de données sont obtenus sous licence open source ou par autorisation explicite des propriétaires.

Je me suis interrogé sur la légalité du scrapping. Le problème c'est que les données sont à la fois sur le domaine public (du moins l'accès) et peuvent également être protégé par la propriété intellectuelle.
Pour mon projet, je me contente de récupérer des données publiques, générales et non sensibles. L'utilisation de ces données à un but purement éducatif et non commercial. De plus, j'ai préféré au maximum obtenir l'aval des propriétaires avant de scraper leurs données. Bien que cette démarche m'ait fermé quelques portes. Cela m'a permis de monter une collaboration avec wikirum.

J'ai utilisé pour mon projet mes propres données, que j'ai anonymisé par nécessité et pour des raisons de discrétion.

Les données utilisées dans ce projet sont des données qui seront publiques. Elles pourront être utilisé par un tiers.

Article III. Création de la base de données

(a) Le modèle conceptuel des données (MCD)

J'ai choisi d'utiliser Looping pour la création de mon MCD. C'est un logiciel simple d'utilisation et complet.

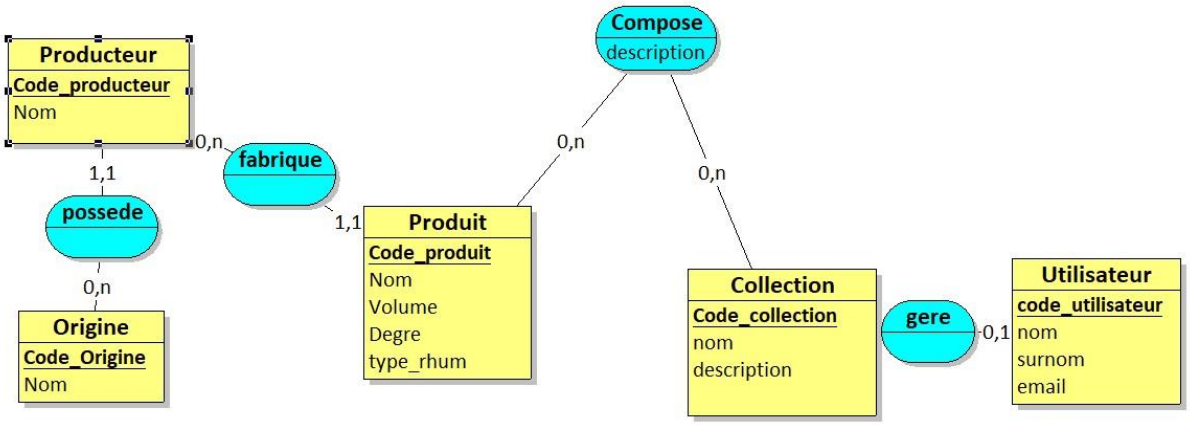
Pour construire notre MCD nous devons définir nos entités, leurs attributs et les relations entre entités.

Nos entités seront :

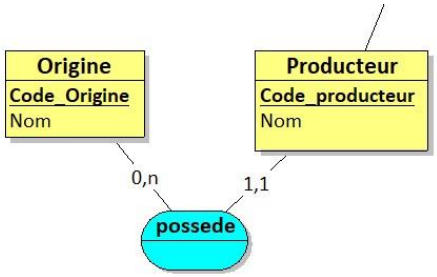
- Origine : telles que données dans les ressources.
    - Il peut être difficile de déterminer l'origine d'un rhum. Car il existe de nombreuses marques qui ne produisent pas de rhum. Elles achètent des lots de rhum de différents endroits du monde. Il se peut que ces lots d'origines différentes soit assemblés par la suite pour ne former qu'une seule référence. D'un point de vue marketing ou de réputation, il arrive aussi que la provenance du rhum soit plus importante que la marque qui le distribue (surtout pour les rhums Caribéens).
- Producteur : englobe les distillateurs, les sous-marques et les embouteilleurs indépendants.
- Produit : le nom du rhum.

- Collection : informations sur la collection d'un utilisateur.
- Utilisateur : les informations sur l'utilisateur.

Voici la modélisation de mon MCD.



Maintenant analysons plus en détails notre MCD :

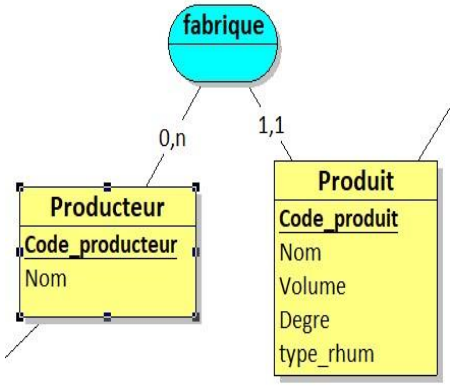


Les entités Origine et Producteur ont le même type d'attribut : un nom et un code spécifique qui fera office d'identifiant unique.

Elles sont associées par une relation « one to many » soit « un à plusieurs » en français.

Exemple de lecture :

- Un producteur possède une seule Origine.
- Une même origine peut être possédée par plusieurs producteurs

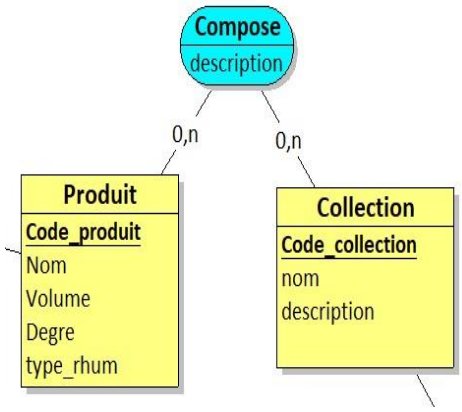


L'entité produit contient les informations nécessaires pour identifier une référence. C'est à partir de cette table que l'on pourra constituer les entités Collections des utilisateurs. Elle contient les attributs : code_produit (identifiant), nom (de la référence), volume, degré et type de rhum (vieux, blanc, ambré, …)

Les entités Produit et Producteur sont associées par une relation « one to many ».

Exemple de lecture :

- Un producteur fabrique un ou plusieurs produits.
- Un produit n'est fabriqué que par un seul producteur.



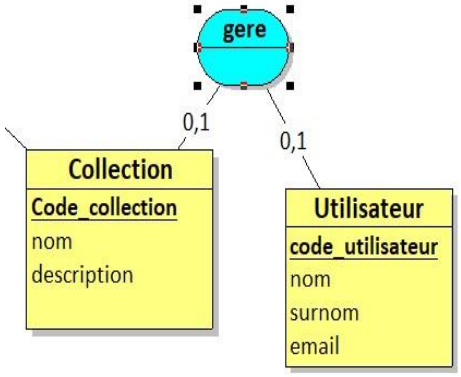
L'entité collection contient les informations nécessaires pour identifier une collection d'un utilisateur. Elle contient trois attributs : code_collection (identifiant), nom, description

Les entités Produit et Producteur sont associées par une relation « many to many » (plusieurs à plusieurs en français).

L'association « Compose » possède également un attribut « description » qui permettra d'ajouter des informations sur une référence.

Exemple de lecture :

- Un produit compose une ou plusieurs collections.
- Une collection est composée d'aucun ou plusieurs produits.



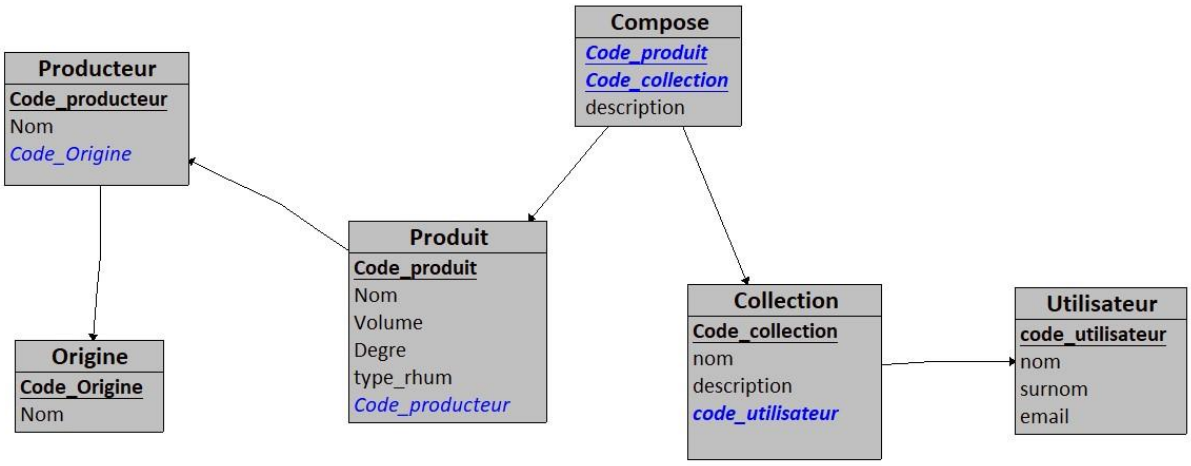
L'entité Utilisateur contient les informations d'identification d'un utilisateur. Elle contient les attributs suivants : code_utilisateur (clé primaire), nom, surnom, email.

Les entités collection et utilisateur sont associées par une relation « one to one » (un à un en français). C'est une relation forte entre deux entités.

Cela veut dire qu'un utilisateur ne peut gérer qu'une seule collection.

(b) Le Modèle Logique des Données (MLD)

Le MLD nous permet de passer du MCD à la modélisation effective de notre base de données que l'on pourra implémenter dans notre SGBD.



Légende : - en bleu clé étrangère

- souligné clé primaire

➔ La clé primaire de la table origine devient une clé étrangère de la table producteur.

➔ La clé primaire de la table producteur devient une clé étrangère de la table produit

➔ Ajout de la table compose pour réaliser la relation many to many entre les tables produit et collection. Cette table a une clé primaire composé des clé primaires code_produit et code_collection.

➔ La clé primaire de la table utilisateur devient une clé étrangère de la table collection.

Article IV. Nettoyage des données.

La préparation des données est une étape essentielle pour s'assurer de la qualité de la base de données.

Ce fut un processus fastidieux. Au niveau de la planification, c'est la tâche que j'ai le plus sous-estimé.

Le nettoyage a été réalisé de plusieurs façon, j'ai essentiellement utilisé Jupyter notebook pour écrire mes scripts, la bibliothèque Pandas et les expressions régulières.

(a) Méthodologie

Les sources de données étant assez hétérogènes, mon but a été de les rendre uniformes. Pour l'ensemble des fichiers, j'ai uniformisé le nom des colonnes et leurs nombres, ne gardant que les informations indispensables. J'ai également supprimé ou remplacé certains caractères spéciaux.

Pour chaque grande étape de nettoyage, je sauvegardais mon travail dans un nouveau fichier. Cela permet de revenir facilement en arrière en cas d'erreur ou de changement, sans devoir recommencer du début. La plus importante partie de nettoyage a été réalisé par l'intermédiaire des Jupiter

Notebook, mais certaines retouchent ont été directement fait dans les logiciels Excel 2019 et Libre Office (pour les fichiers csv).

Dans les parties suivantes, je vais m'attacher à détailler une particularité du nettoyage de chaque fichier.

(b) Fichier excel Inventaire_A.xlsx

La particularité de ce jeu de donnée, c'est qu'il contient des données personnelles. Comme il s'agit de mon propre inventaire, la granularité des informations y est poussée. J'y ai inclus les numéros de fûts, numéro de série, mise en bouteille... Ces informations étaient importantes. En cas de vols, il m'aurait été plus facile de repérer les bouteilles sur les sites e-commerce. Mais cela permet aussi pour une personne extérieure d'avoir des informations sur ma collection.

De plus, ce surplus d'information ajoute un surcroît de complexité. Ces informations complémentaires n'apportent rien de plus pour identifier une référence et elles pourront être rajouter plus tard dans le processus.

J'ai donc décider d'anonymiser le fichier. J'ai dans un premier temps, transférer dans une liste les informations de la colonne « nom ». Puis par une boucle, j'ai appliqué sur les chaînes de caractères la méthode split() sur le caractère '–'. Enfin je récupère la première partie (index [0]) de ma liste créée par la méthode qui sera réintégré dans la dataframe.

Pour ce fichier, j'ai rajouté une colonne « type ». Je l'ai implémenté directement dans la dataframe et je lui ai attribué la valeur par défaut « Vieux » car la plupart des références sont des rhums vieux. J'ai changé les autres valeurs directement dans le fichier Excel.

Capture d'écran de l'évolution du nettoyage :

INVENTAIRE RHUM

		DEGRE	NOMBRE	ACHAT		Marque
HSE						
C12 70	70	50	5	8,5		HSE
C16 70	70	50	1	11		HSE
C16 100	100	50	1	14		HSE
PARCELLAIRE 70	70	55	2	13		HSE
PARCELLAIRE 100	100	55	1	15		HSE
PARCELLAIRE 150L	150	55	1	40		HSE
RAGTIME ESB	70	40	1	11		HSE
Coffret XO ENO	70	43	1	40		HSE
Small Cask 2007-22/10/2015, n°848	50	46	1	26	cadeau	HSE
Small Cask 2007-20/08/2018, n°596	50	46	1	26	cadeau	HSE
Small Cask 2007-20/08/2018, n°1585	50	46	1	26	cadeau	HSE
Highland 2005-18/07/2016, n°1099	50	44	1	34	cadeau	HSE
Highland 2006-26/02/2018, n°179	50	44	1	34	cadeau	HSE
Pedro Ximenez 2007-26/02/18 n°2017/2800	50	46	1	34		HSE
Fino&Olorosso 2004-06/11/2013 , n°1676	50	45	1	34		HSE
Fino&Olorosso 2004-19/07/2016 , n°928	50	45	1	34		HSE
Fino&Olorosso 2004-19/07/2016 , n°1038	50	45	1	34		HSE

Version Brute

nom	volume	degre	producteur	type	origine
C12 70	70	50	Hse	Vieux	Martinique
C16 70	70	50	Hse	Vieux	Martinique
C16 100	100	50	Hse	Vieux	Martinique
PARCELLAIRE 70	70	55	Hse	Vieux	Martinique
PARCELLAIRE 100	100	55	Hse	Vieux	Martinique
PARCELLAIRE 150L	150	55	Hse	Vieux	Martinique
RAGTIME ESB	70	40	Hse	Vieux	Martinique
Coffret XO ENO	70	43	Hse	Vieux	Martinique
Small Cask 2007	50	46	Hse	Vieux	Martinique
Small Cask 2007	50	46	Hse	Vieux	Martinique
Small Cask 2007	50	46	Hse	Vieux	Martinique
Highland 2005	50	44	Hse	Vieux	Martinique
Highland 2006	50	44	Hse	Vieux	Martinique
Pedro Ximenez 2007	50	46	Hse	Vieux	Martinique
Fino&Olorosso 2004	50	45	Hse	Vieux	Martinique
Fino&Olorosso 2004	50	45	Hse	Vieux	Martinique
Fino&Olorosso 2004	50	45	Hse	Vieux	Martinique
Fino&Olorosso 2007	50	45	Hse	Vieux	Martinique

Version Intermédiaire

nom	volume	degre	producteur	type	origine
C12 70	70	50	Hse	Blanc	Martinique
C16 70	70	50	Hse	Blanc	Martinique
C16 100	100	50	Hse	Blanc	Martinique
PARCELLAIRE 70	70	55	Hse	Blanc	Martinique
PARCELLAIRE 100	100	55	Hse	Blanc	Martinique
PARCELLAIRE 150L	150	55	Hse	Blanc	Martinique
RAGTIME ESB	70	40	Hse	Ambre	Martinique
Coffret XO ENO	70	43	Hse	Vieux	Martinique
Small Cask 2007	50	46	Hse	Vieux	Martinique
Highland 2005	50	44	Hse	Vieux	Martinique
Highland 2006	50	44	Hse	Vieux	Martinique
Pedro Ximenez 2007	50	46	Hse	Vieux	Martinique
Fino&Olorosso 2004	50	45	Hse	Vieux	Martinique
Fino&Olorosso 2007	50	45	Hse	Vieux	Martinique
Marquis de Terme 2005	50	47	Hse	Vieux	Martinique
Marquis de Terme 2006	50	47	Hse	Vieux	Martinique

Version Finale.

(c) Fichier rum_data.csv

Dans la colonne « company », il y avait beaucoup de producteurs qui avaient la mention « Unknown ». Après analyse, il s'avérait que cela concernait beaucoup de petites marques, assez confidentiel (et très peu de référence). Très souvent dans le nom du produit se trouvait le nom du producteur. Il était impossible dans ce cas-là, d'utiliser des expressions régulières pour extraire le nom du producteur dans la colonne « nom ». J'ai donc décidé de remplacer directement la mention « Unknown » par le nom du produit (tel qu'il était dans la colonne « nom »). J'ai utilisé la fonction loc de pandas pour repérer dans la colonne « company » les mentions « Unknown » pour faire l'échange avec le nom du produit.

Pour certaines grandes producteurs (en volume et/ou réputation), j'ai fait les modifications directement dans le fichier csv.

Ce fichier était le seul a déjà avoir une colonne « type », j'ai juste uniformisé les labels, pour plus de simplicité, par exemple : « Aged, Dark » > « Vieux ».

name	company	country	price	ratings	score	type	rum_url	img_url	br_score
10000 Drops Silver	Unknown	United States	0	0	0	Light	/brands/10972-10000-drops-silver	https://d1jtwiy8m5zi8	4.77536295
10000 Drops Spiced	Unknown	United States	0	1	4	Spiced	/brands/7354-10000-drops-spiced	https://d19vk5i0q1xlsm	4.711954542
1000 Hills Gold	Unknown	Rwanda	0	1	4	Light	/brands/5285-1000-hills-gold	https://d19vk5i0q1xlsm	4.711954542
100 Fuegos Buckeye 2-Year	Unknown	Ecuador	0	0	0	Gold	/brands/9037-100-fuegos-buckeye-2-ye	https://d1jtwiy8m5zi8	4.77536295
100 Fuegos 8-Year	Unknown	Ecuador	0	0	0	Aged	/brands/1834-100-fuegos-8-year	https://d1jtwiy8m5zi8	4.77536295
105 Simonton	Unknown	United States	0	0	0	Gold	/brands/9303-105-simonton	https://d1jtwiy8m5zi8	4.77536295
10 Cane 1999 Guyana 15-Year	10 Cane	Trinidad and Tobago	0	0	0	Aged	/brands/3805-10-cane-1999-guyana-15	https://d1jtwiy8m5zi8	4.77536295
10 Cane Light	10 Cane	Trinidad and Tobago	0	88	6.4	Light	/brands/1-10-cane-light	https://d19vk5i0q1xlsm	6.216165428
10 Cane Signature Blend	10 Cane	Trinidad and Tobago	0	1	10	Gold	/brands/2312-10-cane-signature-blend	https://d1jtwiy8m5zi8	5.202628554
1423 Indian Rum	1423	Panama	0	1	5	Flavored	/brands/3998-1423-indian-rum	https://d1jtwiy8m5zi8	4.793733544
1423 Panama 12-Year	1423	Panama	0	1	7	Aged	/brands/1526-1423-panama-12-year	https://d19vk5i0q1xlsm	4.957291548
1423 Special Cask 25-Year	1423	Panama	0	4	9	Aged	/brands/1527-1423-special-cask-25-ye	https://d19vk5i0q1xlsm	5.885059683
1492151	1492	United States	0	0	0	Overproof	/brands/5997-1492-151	https://d1jtwiy8m5zi8	4.77536295
1492 Antique 15-Year	1492	United States	0	0	0	Aged	/brands/7074-1492-antique-15-year	https://d19vk5i0q1xlsm	4.77536295
1492 Blanco	1492	United States	0	0	0	Aged	/brands/7067-1492-blanco	https://d1jtwiy8m5zi8	4.77536295
1492 Blanco 7	1492	United States	0	0	0	Aged	/brands/7068-1492-blanco-7	https://d1jtwiy8m5zi8	4.77536295
1492 Cristobal Reserve	1492	United States	0	2	6.5	Aged	/brands/1215-1492-cristobal-reserve	https://d19vk5i0q1xlsm	5.036116942
1492 Golden Age	1492	United States	0	1	3	Gold	/brands/7070-1492-golden-age	https://d1jtwiy8m5zi8	4.63017554
1492 Reserva 7	1492	United States	0	0	0	Aged	/brands/7069-1492-reserva-7	https://d1jtwiy8m5zi8	4.77536295
1492 Spiced	1492	United States	0	1	2	Spiced	/brands/7071-1492-spiced	https://d1jtwiy8m5zi8	4.548396538
1492 21-Year	1492	United States	0	0	0	Aged	/brands/5998-1492-21-year	https://d19vk5i0q1xlsm	4.77536295
1492 18-Year	1492	United States	0	0	0	Aged	/brands/7073-1492-18-year	https://d1jtwiy8m5zi8	4.77536295
1492 10-Year	1492	United States	0	0	0	Aged	/brands/7072-1492-10-year	https://d19vk5i0q1xlsm	4.77536295
1653 Old Barrel	Unknown	Switzerland	0	67	7.2	Aged	/brands/5880-1653-old-barrel	https://d19vk5i0q1xlsm	5.619788106

Version brute

nom	producteur	origine	type	im_url
10000 Drops Silver	10000 Drops Silver	United States	Blanc	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
10000 Drops Spiced	10000 Drops Spiced	United States	Spiced	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/7354/small_10-000-drops-sp
1000 Hills Gold	1000 Hills Gold	Rwanda	Blanc	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/5285/small_1000-hills-gold.
100 Fuegos Buckeye 2-Year	100 Fuegos Buckeye	Ecuador	Ambre	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
100 Fuegos 8-Year	100 Fuegos 8-Year	Ecuador	Vieux	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
105 Simonton	105 Simonton	United States	Ambre	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
10 Cane 1999 Guyana 15-Y	10 Cane	Trinidad and Tobago	Vieux	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
10 Cane Light	10 Cane	Trinidad and Tobago	Blanc	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/1/small_10_cane_rum_400
10 Cane Signature Blend	10 Cane	Trinidad and Tobago	Ambre	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
1423 Indian Rum	1423	Panama	Spiced	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
1423 Panama 12-Year	1423	Panama	Vieux	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/1526/small_1423_panama_
1423 Special Cask 25-Year	1423	Panama	Vieux	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/1527/small_1423_special_c
1492151	1492	United States	Overproof	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
1492 Antique 15-Year	1492	United States	Vieux	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/7074/small_1492-antique-15
1492 Blanco	1492	United States	Vieux	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
1492 Blanco 7	1492	United States	Vieux	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
1492 Cristobal Reserve	1492	United States	Vieux	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/1215/small_1492_Cristobal_
1492 Golden Age	1492	United States	Ambre	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
1492 Reserva 7	1492	United States	Vieux	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
1492 Spiced	1492	United States	Spiced	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
1492 21-Year	1492	United States	Vieux	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/5998/small_1492-21-year.pr
1492 18-Year	1492	United States	Vieux	https://d1jtwiy8m5zi8a.cloudfront.net/assets/default_bottle_preview-c15f89a45f2ac217a
1492 10-Year	1492	United States	Vieux	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/7072/small_1492-10-year.pr
1653 Old Barrel	1653 Old Barrel	Switzerland	Vieux	https://d19vk5i0q1xlsm.cloudfront.net/uploads/brand/image/5880/small_1653-old-barrel.

Version finale

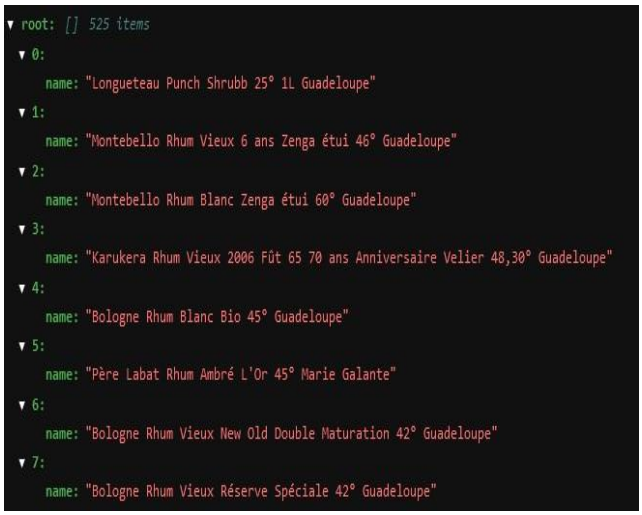
(d) Fichier test.json

J'ai utilisé des expressions régulières pour extraire les informations nécessaires. Je m'aidé du site Pytex (https://pythex.org/) qui permet de tester les expressions régulières.

Voici un exemple d'expression régulière qui m'a permis de trouver (quasiment) tous les producteurs :

```
    # Producteurs
    marque_pattern = re.compile(r"[a-zA-Zèé]+[&? [a-zA-Zèé0-9']+ Rhum|[a-zA-Zèé']+[ ]? Rhum|[a-zA-Zèé']+ [Liqueur|Punch|Coffret]+")
    marque = marque_pattern.findall(element)
#     marque = marque.replace("Rhum", "")
    if marque:
        marques.append(marque[0])
    else:
        marques.append("XXXXXXXX")
```

L'utilisation des expressions régulières n'a pas toujours fourni un résultat parfait. J'ai pu corriger certaines valeurs grâce à la fonction « replace » de pandas. Les anciennes et les nouvelles valeurs sont contenues dans un dictionnaire qui est passé en paramètre de la fonction.



Version brute

nom	producteur	origine	volume	degre	type
Montebello Rhum Vieux 6 ans Zenga etui	Montebello	Guadeloupe		46	Vieux
Karukera Rhum Vieux 2006 Fût 65 70 ans Anniversaire Velier	Karukera	Guadeloupe		48.30	Vieux
Bologne Rhum Vieux New Old Double Maturation	Bologne	Guadeloupe		42	Vieux
Bologne Rhum Vieux Reserve Speciale	Bologne	Guadeloupe		42	Vieux
Longueteau Rhum Vieux edition 120 ans	Longueteau	Guadeloupe		40	Vieux
Bologne Rhum Vieux Les Confidentiels 2014 Finish Sauternes Brut de Fût etui	Bologne	Guadeloupe	50	49.9	Vieux
Bologne Rhum Vieux Les Confidentiels Hors dAge 2009 Brut de Fût etui	Bologne	Guadeloupe	50	53.1	Vieux
Longueteau Rhum Vieux Concerto Batch 6 Harmonie Collection	Longueteau	Guadeloupe		48.9	Vieux
Longueteau Rhum Vieux Prelude Batch 8 Harmonie Collection	Longueteau	Guadeloupe		49.3	Vieux
Ferroni Rhum Vieux Gwada 2012 Brut de Fût	Ferroni	Guadeloupe		52.3	Vieux
Longueteau Rhum Vieux Cuvee Confrerie du Rhum XO Single Cask Fût 44	Confrerie du Rhum	Guadeloupe		48.6	Vieux
Damoiseau Rhum Vieux Concordia	Damoiseau	Guadeloupe		40	Vieux
Sixty Six Rum Rhum Vieux 12 ans Cask Strenght	Sixty Six Rum	Barbade		59	Vieux
Mount Gay Rhum Vieux 10 ans 1703 Master Select ed. 2020 carafe etui	Mount Gay	Barbade	70	43	Vieux
LEsprit Rhum Vieux Foursquare 2005	LEsprit	Barbade		60.5	Vieux
Mount Gay Rhum Vieux The Port Cask Expression Brut de Fût etui	Mount Gay	Barbade		55	Vieux
Doorlys Rhum Vieux 8 ans	Doorlys	Barbade		46	Vieux
Plantation Rhum Vieux 2011 etui	Plantation	Barbade		51.1	Vieux
La Maison du Rhum Rhum Rhum Vieux 5 ans	La Maison du Rhum	Barbade		40	Vieux

Version finale

(e) Fichier wikirum

Ce fichier m'a été amicalement fourni par l'équipe de wikirum. Donc c'était un fichier déjà bien formaté. Le nettoyage a été assez rapide. Les changements notables, ont été la transformation en « float » du type des colonnes « degre » et « volume » et l'ajout de la colonne « type ».

nom	marque	pays	volume	titrage_final
Cuba 5 Years	1731 Fine & Rare*	Pays-Bas	700	46
Cuba Dominican Republic (Spanish Caribbean XO)	1731 Fine & Rare*	Pays-Bas	700	46
Guatemala Panama Bélize (Central America XO)	1731 Fine & Rare*	Pays-Bas	700	46
Panama 6 Years	1731 Fine & Rare*	Pays-Bas	700	46
Trinidad Barbados Jamaica (British West Indies XO)	1731 Fine & Rare*	Pays-Bas	700	46
Caroni 12 Ans	Blackadder*	Angleterre	700	46
Caroni 18 Ans	Blackadder*	Angleterre	700	63.1
Fiji 2001-2012 - 11 Years	Blackadder*	Angleterre	700	63.9
Guadeloupe Belvédère 1998-2015- 17 Years	Blackadder*	Angleterre	700	57.5
Guadeloupe Belvédère 1998-2017- 19 Years	Blackadder*	Angleterre	700	56.6
Guyana Diamond 2008-2018- 10 Years	Blackadder*	Angleterre	700	57.6
Guyana Diamond 2003-2016- 12 Years	Blackadder*	Angleterre	700	64.3
Guyana Diamond 2003-2015- 14 Years	Blackadder*	Angleterre	700	63.1
Guyana Diamond 2003-2015- 15 Years	Blackadder*	Angleterre	700	48.6
Jamaica Hampden - 14 Years	Blackadder*	Angleterre	700	57.4
Jamaica Hampden 2000-2016- 15 Years	Blackadder*	Angleterre	700	57
Jamaica Hampden 2000-2017- 16 Years	Blackadder*	Angleterre	700	56.1
Nicaragua 1999	Blackadder*	Angleterre	700	46
Ste-Lucie 1999-2012	Blackadder*	Angleterre	700	68.2
Blanc VSOP	Bonpland*	Allemagne	700	40
Claire	Bonpland*	Allemagne	700	42
Forte	Bonpland*	Allemagne	700	55
Rouge VSOP	Bonpland*	Allemagne	700	40

Version brute

nom	producteur	origine	volume	degre	type
Cuba 5 Years	1731 Fine & Rare	Pays-Bas	70.0	46.0	Vieux
Cuba Dominican Republic (Spanish Caribbean XO)	1731 Fine & Rare	Pays-Bas	70.0	46.0	Vieux
Guatemala Panama Belize (Central America XO)	1731 Fine & Rare	Pays-Bas	70.0	46.0	Vieux
Panama 6 Years	1731 Fine & Rare	Pays-Bas	70.0	46.0	Vieux
Trinidad Barbados Jamaica (British West Indies XO)	1731 Fine & Rare	Pays-Bas	70.0	46.0	Vieux
Caroni 12 Ans	Blackadder	Angleterre	70.0	46.0	Vieux
Caroni 18 Ans	Blackadder	Angleterre	70.0	63.1	Vieux
Fiji 2001-2012 - 11 Years	Blackadder	Angleterre	70.0	63.9	Vieux
Guadeloupe Belvedere 1998-2015- 17 Years	Blackadder	Angleterre	70.0	57.5	Vieux
Guadeloupe Belvedere 1998-2017- 19 Years	Blackadder	Angleterre	70.0	56.6	Vieux
Guyana Diamond 2008-2018- 10 Years	Blackadder	Angleterre	70.0	57.6	Vieux
Guyana Diamond 2003-2016- 12 Years	Blackadder	Angleterre	70.0	64.3	Vieux
Guyana Diamond 2003-2015- 14 Years	Blackadder	Angleterre	70.0	63.1	Vieux
Guyana Diamond 2003-2015- 15 Years	Blackadder	Angleterre	70.0	48.6	Vieux
Jamaica Hampden - 14 Years	Blackadder	Angleterre	70.0	57.4	Vieux
Jamaica Hampden 2000-2016- 15 Years	Blackadder	Angleterre	70.0	57.0	Vieux
Jamaica Hampden 2000-2017- 16 Years	Blackadder	Angleterre	70.0	56.1	Vieux
Nicaragua 1999	Blackadder	Angleterre	70.0	46.0	Vieux
Ste-Lucie 1999-2012	Blackadder	Angleterre	70.0	68.2	Vieux
Blanc VSOP	Bonpland	Allemagne	70.0	40.0	Spiced
Claire	Bonpland	Allemagne	70.0	42.0	Spiced
Forte	Bonpland	Allemagne	70.0	55.0	Spiced
Rouge VSOP	Bonpland	Allemagne	70.0	40.0	Spiced

Version finale

Article V. Sauvegarde et stockage

J'ai choisi PostgreSQL pour sauvegarder ma base de données. J'ai décidé de sauvegarder en local ma base de données.

J'ai sauvegardé dans un fichier json, les métadonnées de mes fichiers. J'utilise le logiciel exiftool, qui me permet d'extraire plein d'informations. Pour l'intérêt technique, j'ai également sauvegardé mon répertoire de métadonnées sur MongoDB.

Pour sauvegarder mon projet j'ai adopté différentes stratégies :

- Un dépôt régulier sur le repositorie gitlab du projet : https://gitlab.com/simplon-dev-data-nantes-2020/projet-chef-d-oeuvre/-/tree/mayelpelle
- Une copie de l'ensemble du répertoire sur mon compte one-drive au moins une fois par semaine
- Un script (trouvé par notre collègue Joris Teyssier : https://nitratine.net/blog/post/python-file-backup-script/) qui me permet de réaliser un backup de ma base de données dans un répertoire défini.

L'EXPLOITATION DES DONNÉES

Dans cette partie nous allons aborder l'utilisation des données stockées dans la base de données PostgreSQL.

L'objectif de ce projet data est principalement de créer une base de données fonctionnel. Nous allons la tester via le biais de quelques requêtes.

J'ai réalisé quelques graphiques avec Plotly-express. C'est une librairie qui permet de faire des graphiques interactifs

(i) Première requête : Les 20 pays les plus représentés (en fonction du nombre de référence) :

```
SELECT origine.nom, count(produit.nom_produit)
FROM origine
INNER JOIN producteur
ON origine.code_origine = producteur.code_origine
INNER JOIN produit
ON producteur.code_producteur = produit.code_producteur
group by origine.nom
order by count(produit.nom_produit) DESC
Limit 20;
```

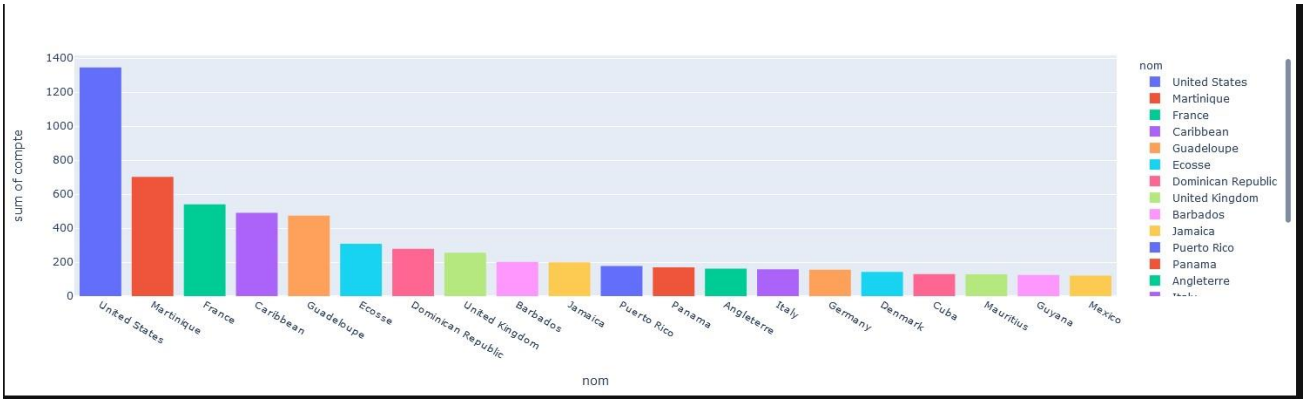
	nom character varying (100)	count bigint
1	United States	1346
2	Martinique	702
3	France	540
4	Caribbean	490
5	Guadeloupe	473
6	Ecosse	308
7	Dominican Republic	279
8	United Kingdom	255
9	Barbados	202
10	Jamaica	199
11	Puerto Rico	177
12	Panama	169
13	Angleterre	162
14	Italy	158
15	Germany	155
16	Denmark	143
17	Cuba	129
18	Mauritius	128
19	Guyana	125
20	Mexico	120

La requête exécutée dans pgadmin, fait ressortir une sur représentativité des rhums d'origine étasunienne. Ceux-ci occupent la première marche du podium. Ce bon résultat est dû au jeu de données rum_data.csv qui fait la part belle aux rhums nord-américain.

Les rhums caribéens et martiniquais figurent en bonne place. Ce ne sont pourtant pas les plus grands producteurs de rhum. Mais ce sont des rhums qui jouissent d'une excellente réputation. Dans la recherche des données sur le rhum, j'ai longtemps privilégié les caraïbes. Beaucoup des bouteilles qui intéressent les collectionneurs proviennent de cette région.

Pour le moment, les noms des pays n'ont pas été traduit. Le but étant ici de réaliser un premier POC fonctionnel. L'intégration dans une ou deux langues des noms des pays est une question importante. Elle sera aux centres des prochains développement.

Le graphique ci-dessous, réalisé avec plotly-express, nous confirme le résultat de notre requête.



(ii) Deuxième requête : Les 10 plus grands producteurs

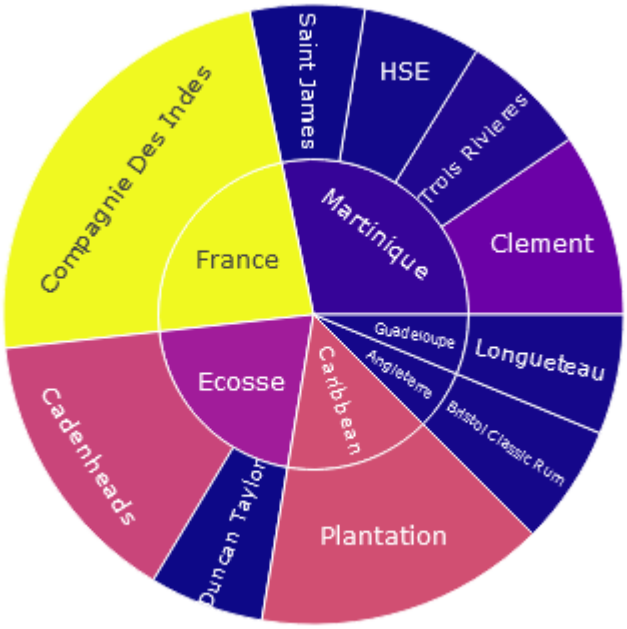
```
SELECT producteur.nom_producteur,
origine.nom,
count(produit.nom_produit)

FROM producteur

INNER JOIN origine

ON producteur.code_origine =
origine.code_origine

INNER JOIN produit

ON producteur.code_producteur =
produit.code_producteur

group by producteur.nom_producteur,
origine.nom

order by count(produit.nom_produit)
DESC

Limit 10;
```

	nom_producteur character varying (150)	nom character varying (100)	count bigint
1	Compagnie Des Indes	France	255
2	Plantation	Caribbean	165
3	Cadenheads	Ecosse	159
4	Clement	Martinique	104
5	Trois Rivieres	Martinique	72
6	Bristol Classic Rum	Angleterre	68
7	Longueteau	Guadeloupe	68
8	HSE	Martinique	67
9	Saint James	Martinique	65
10	Duncan Taylor	Ecosse	65

Ce classement est intéressant. Il fait la part belle aux embouteilleurs indépendants. Sur les 3 premiers, deux sont des embouteilleurs indépendants (Compagnie des Indes, pourtant fondé assez récemment et Cadenheads, une ancienne institution). Plantation est une marque française détenu par Maison Ferrand, grande maison de cognac. Elle est depuis quelques années propriétaire de distilleries à la Barbade et en Jamaïque. La marque plantation est très reconnu par les amateurs et bénéficie d'un beau succès commercial à l'export (surtout aux Etats Unis), ce qui explique sa haute

position. Il faut attendre la 8e place avec HSE, pour avoir un producteur qui détient et exploite sa propre distillerie.

Avec ce camembert, la domination de la Compagnie des Indes est encore plus flagrante.

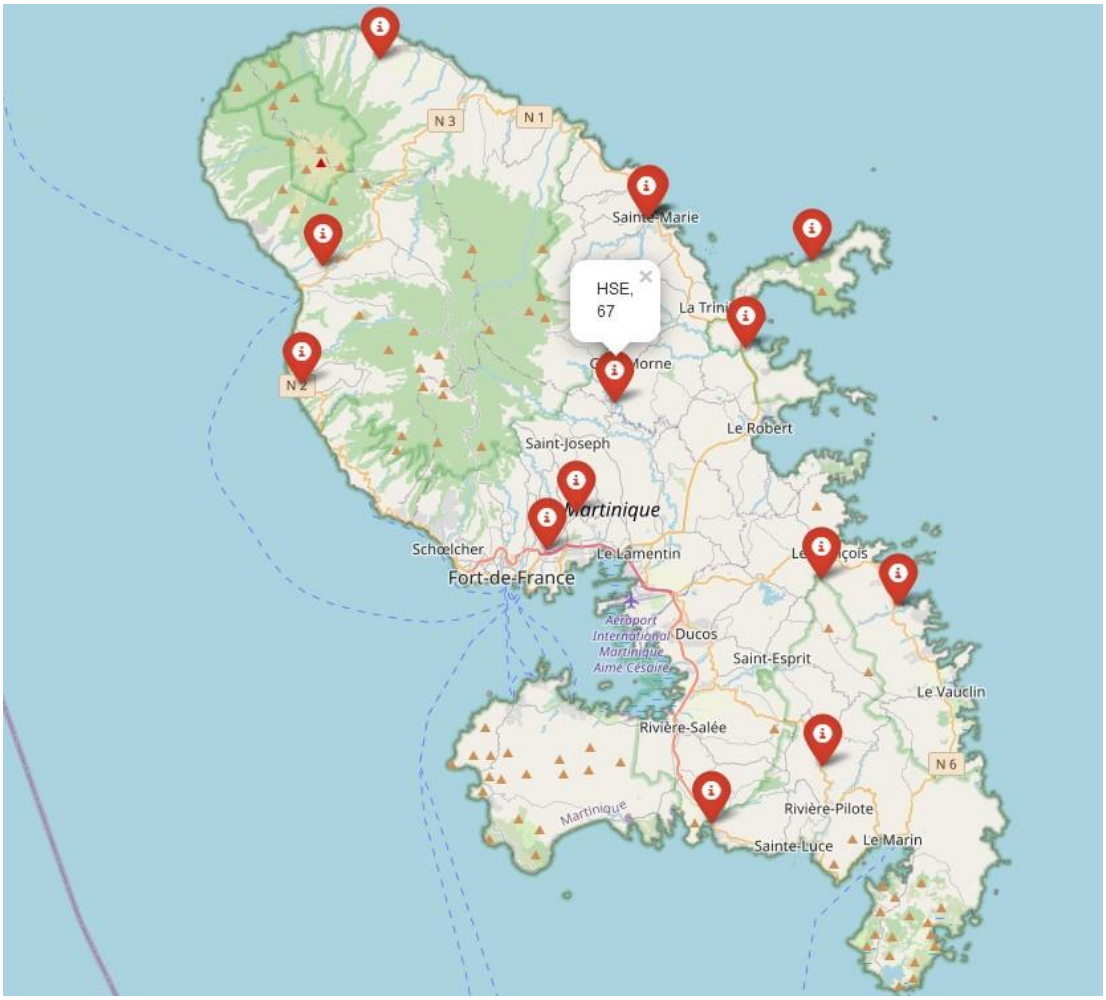


(iii) Troisième requête : Les références martiniquaises

Cette requête nous permet d'afficher le nombre de référence par marque martiniquaise. Grâce à folium, une librairie python, je peux modéliser ces résultats sur une carte de la Martinique.

```
Select producteur.nom_producteur,
origine.nom, count(produit.nom_produit)

from producteur

inner join origine

on producteur.code_origine =
origine.code_origine

inner join produit

on producteur.code_producteur =
produit.code_producteur
```

20	HSE	Martinique	67
21	J. Bally	Martinique	36
22	Jura	Martinique	1
23	La Favorite	Martinique	40
24	La Mauny	Martinique	50
25	LAmitie White	Martinique	1
26	Lauzea Chocolatier Le Shrubb	Martinique	1



N’ayant pas d’adresse dans mes jeux de données, j’ai dû chercher manuellement les coordonnées sur Openstreetmap. Grâce à un tooltip, je peux afficher sur la carte le nombre de références. On peut également zoomer sur la carte, pour voir où se situe précisément les ressources.

CONCLUSION

Nous arrivons à la fin de ce rapport. L'objectif initial qui était de créer une base de données fonctionnelle a été atteint. Après un mois passé sur ce projet, j'ai une idée plus précise des améliorations et problématiques futures.

Il y a la question de la traduction de certains termes, notamment les noms de pays.

Il y a la création de l'interface homme-machine pour permettre à l'utilisateur de gérer sa propre collection de façon effective.

Il y a des questions de sécurités. J'ai envie d'explorer la possibilité de chiffrer des données précises dans la base de données, par exemple les mots de passe ou les données qui seront enregistrées dans les champs de descriptions.

Ce sont quelques pistes de réflexion parmi les nombreuses questions qui restent encore en suspens.

Je profite de cette conclusion pour faire un point sur le déroulement de ce projet. Pour moi ce projet est vraiment positif. Il m'a permis de mieux cerner un projet de développement data. J'ai pu progresser dans toutes les étapes : la planification, la conception du mcd, le nettoyage des données, …

J'ai pu utiliser un large panel de technologie pour faire ce projet. Cela me donne vraiment envie d'aller plus loin dans mon apprentissage du développement.

Pourtant ces dernières semaines ont été intenses et laborieuses. Travailler en individuel a été un vrai challenge pour moi. Heureusement j'ai pu compter sur les formateurs et les collègues. Je les en remercie.

ANNEXE 1 : RESSOURCES

https://pandas.pydata.org/pandas-docs/stable/index.html

https://www.pytex.org/

https://www.youtube.com/

https://stackoverflow.com/

http://cours.pise.info/modelisation/index.htm

https://scrapy.org/

https://www.crummy.com/software/BeautifulSoup/

https://plotly.com/python/plotly-express/

https://developer.ebay.com/

https://www.delftstack.com/fr/howto/python/python-convert-list-into-dictionary/

https://www.delftstack.com/fr/howto/python-pandas/

https://moncoachdata.com/blog/nettoyage-de-donnees-python/

https://nitratine.net/blog/post/python-file-backup-script/

https://naysan.ca/2020/05/09/pandas-to-postgresql-using-psycopg2-bulk-insert-performance-benchmark/

https://github.com/CPingeon/exacdatamente/blob/main/04_Create_DB.ipynb

https://pynative.com/python-postgresql-insert-update-delete-table-data-to-perform-crud-operations/

https://github.com/python-visualization/folium

https://rumratings.com/

https://wikirum.fr/

https://rumporter.com/mentions-legales/

https://www.christiandemontaguere.com/

https://www.kaggle.com/mrpantherson/rum-data