Projet chef d’œuvre

CollectSpirit

Certification Simplon 2021

Développer data

Apprenant :

Mayel Pèllé

Années :

2020-2021

# Sommaire

[Sommaire 2](#_Toc71884321)

[INTRODUCTION 3](#_Toc71884322)

[Préparation du projet 4](#_Toc71884323)

[Article I. La phase de planification 4](#_Toc71884324)

[Article II. L’expression du besoin 5](#_Toc71884325)

[(a) La progression du marché du rhum 5](#_Toc71884326)

[(b) Les collectionneurs de spiritueux 7](#_Toc71884327)

[(c) Traduction du besoin 8](#_Toc71884328)

[(d) Les applications existantes 9](#_Toc71884329)

[CONCEPTION DU PROJET 10](#_Toc71884330)

[Article I. L’application 10](#_Toc71884331)

[(a) Traduction du besoin client 10](#_Toc71884332)

[(b) Choix des technologies 11](#_Toc71884333)

[Article II. Les sources de données 12](#_Toc71884334)

[(a) La collecte des données 12](#_Toc71884335)

[(b) RGPD 13](#_Toc71884336)

[Article III. Création de la base de données 13](#_Toc71884337)

[(a) Le modèle conceptuel des données (MCD) 13](#_Toc71884338)

[(b) Le Modèle Logique des Données (MLD) 16](#_Toc71884339)

[Article IV. Nettoyage des données. 17](#_Toc71884340)

[(a) Méthodologie 17](#_Toc71884341)

[(b) Fichier excel Inventaire\_A.xlsx 17](#_Toc71884342)

[(c) Fichier rum\_data.csv 18](#_Toc71884343)

[(d) Fichier test.json 19](#_Toc71884344)

[(e) Fichier wikirum 20](#_Toc71884345)

[Article V. Sauvegarde et stockage 21](#_Toc71884346)

[L’EXPLOITATION DES DONNÉES 22](#_Toc71884347)

[CONCLUSION 26](#_Toc71884348)

[ANNEXE 1 : RESSOURCES 27](#_Toc71884349)

# 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 manuellement 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. 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. Créer cette base de données et l’exploiter, seront nos buts principaux de projet.

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.

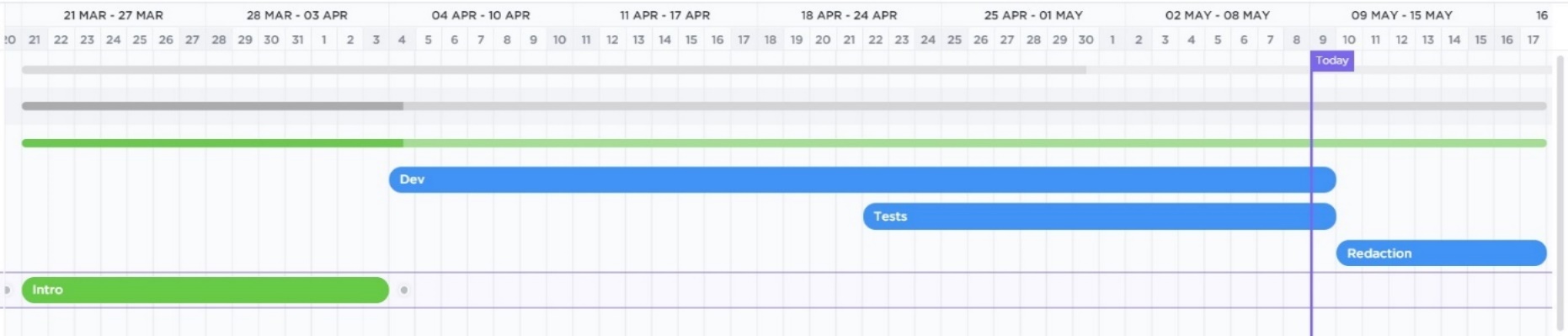
# Préparation du projet

# La phase de planification

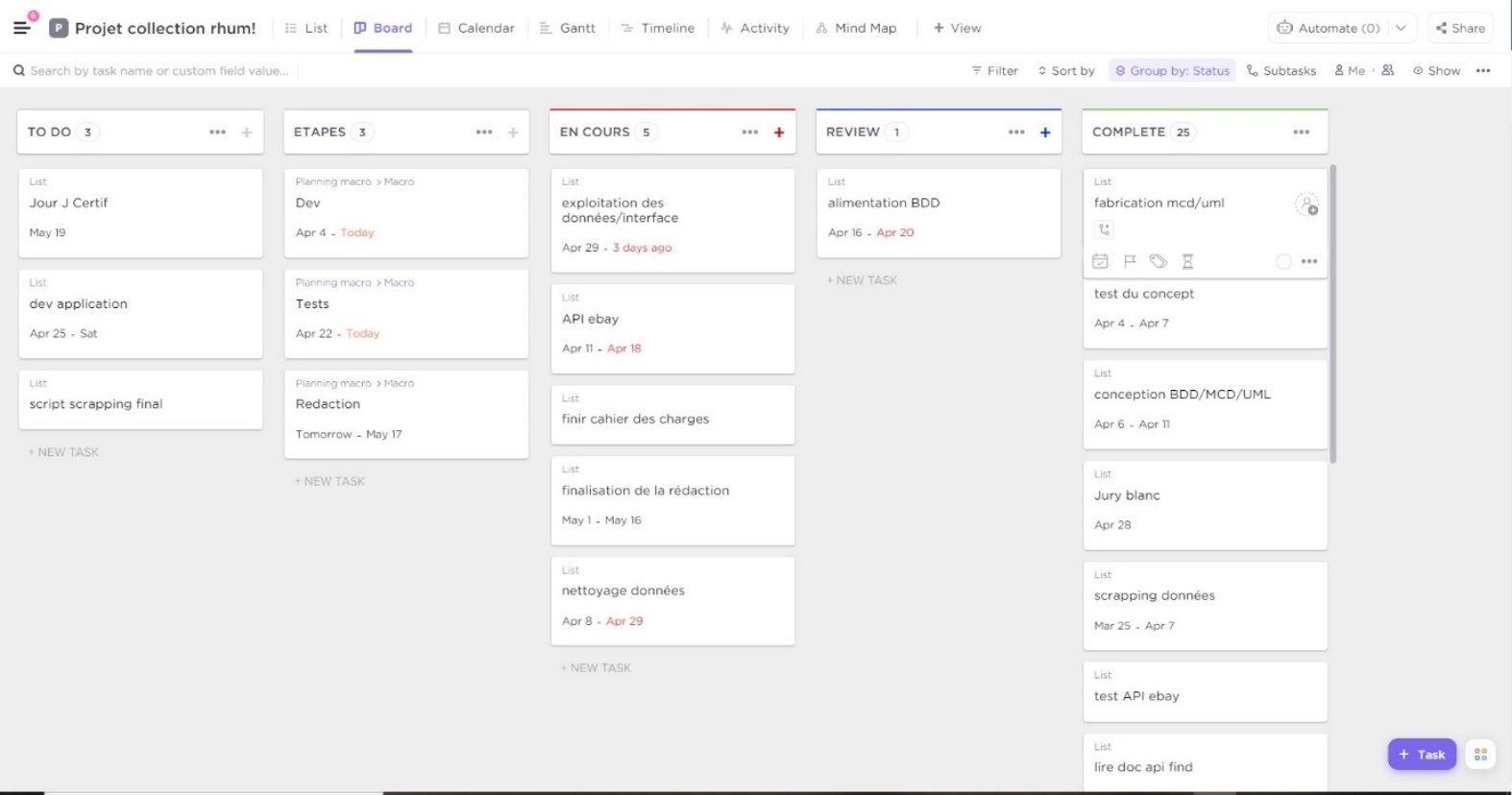
Notre projet de certification est un projet individuel dont le sujet est libre. La première étape a donc été de faire des recherches préliminaires pour déterminer le sujet et sa faisabilité.

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 d’agilité 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.

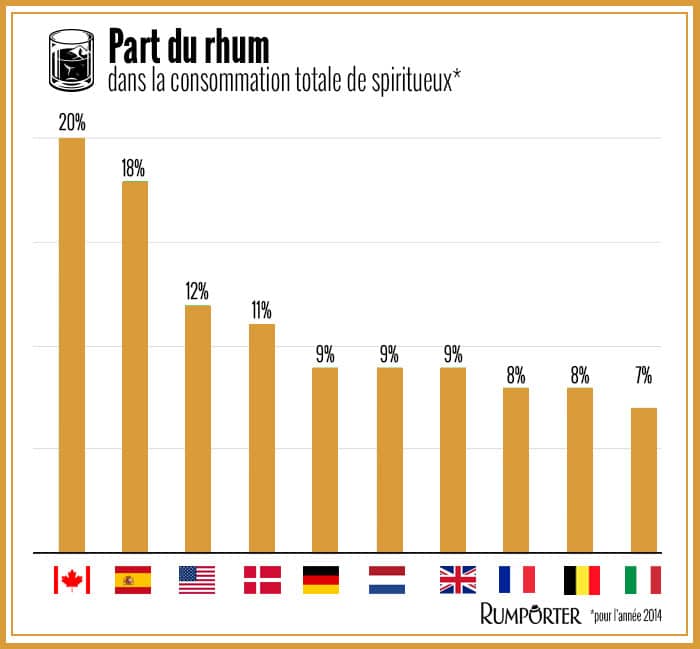
Je contactais au moins deux fois par semaine mes formateurs, afin de leur faire un retour sur l’avancée de mon projet, les difficultés rencontrées et les solutions trouvées. Je les interpellais ponctuellement lorsque je rencontrais un problème bloquant. Ce contact régulier était essentiel, nous étions 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 sous-estime le temps à passer sur 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.

# L’expression du besoin

### 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.

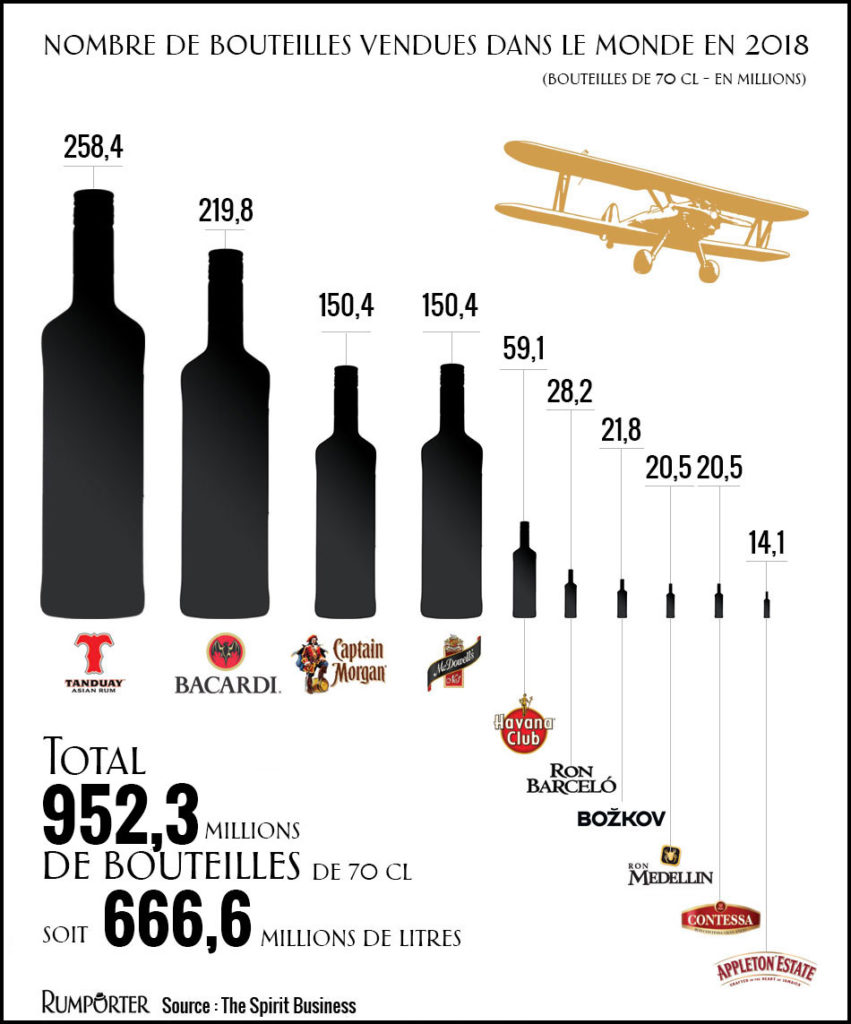


En France, le rhum représente 8 % de la consommation totale de spiritueux avec 34 millions de litres consommés et il reste un des principaux moteurs de croissance de la catégorie alcools. La France est d'ailleurs le deuxième plus grand consommateur de rhum en Europe derrière l'Espagne.

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).

(source : https://www.planetoscope.com/boisson/1510-consommation-mondiale-de-rhum.html)



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.

### Les collectionneurs de spiritueux

Ce projet s’adresse 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.

### 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 (lot) 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.

### 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 :

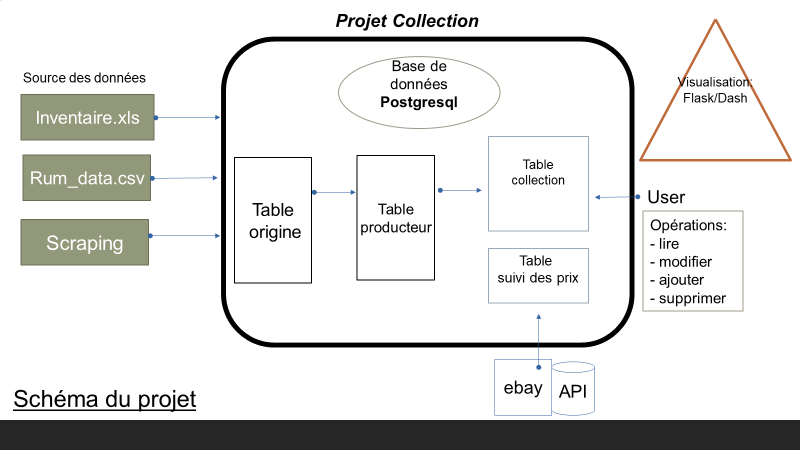
* Wikirum : c’est une web application construite avec Angular qui est la base de données la plus complète sur le rhum. C’est un projet porté par des nantais de Saint Herblain. Ils travaillent directement avec les distilleries et les marques pour obtenir des données. C’est avant tout un wikipédia du rhum. 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

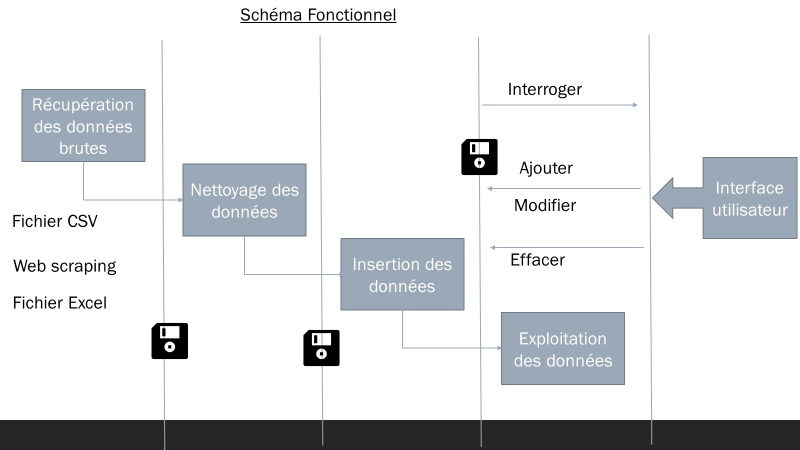
# L’application

### 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.



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.



Légende : les disquettes représentent les sauvegardes et backup.

### 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.

# Les sources de données

### La collecte des données

Pendant la phase de recherche, j’avais prévu de récupérer 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 Scrapy. 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.

### Conformité à la 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 protégées 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 utilisées par un tiers.

# Création de la base de données

### Le modèle conceptuel des données (MCD)

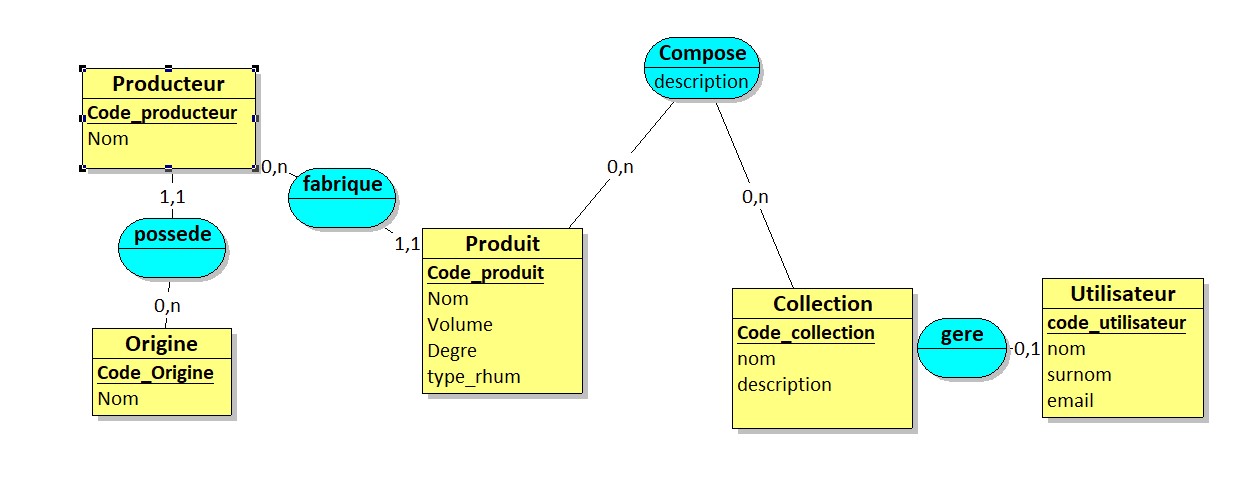
J’ai choisi d’utiliser Looping pour la création de mon MCD. C’est un logiciel simple d’utilisation et complet.

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

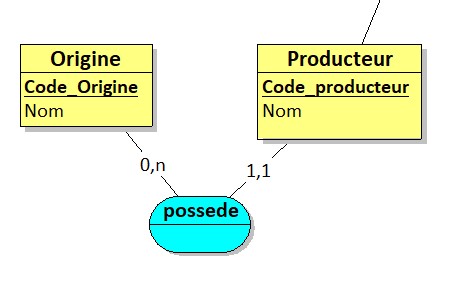
Nos entités seront :

* 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 :

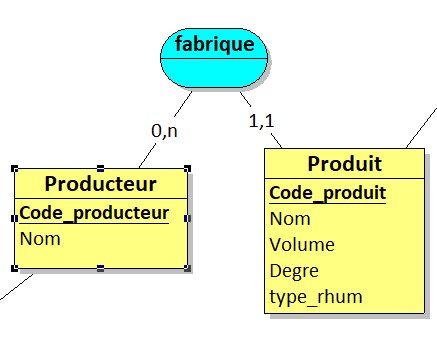


Les entités Origine et Producteur ont le même type d’attribut : un nom et un code spécifique qui fera office d’identifiant unique.

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

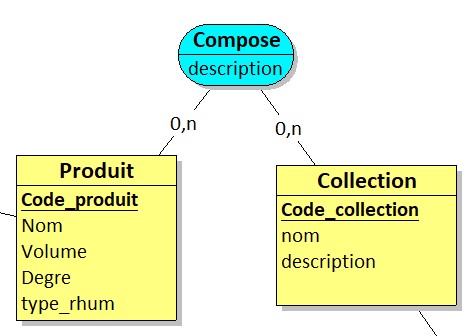


L’entité produit contient les informations nécessaires pour identifier une référence. C’est à partir de cette table que l’on pourra constituer les entités Collections des utilisateurs. Elle contient les attributs : code\_produit (identifiant), nom (de la référence), volume, degré et type de rhum (vieux,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



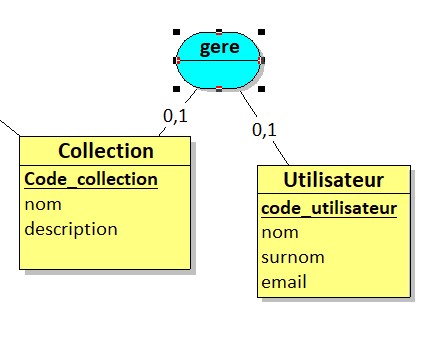
L’entité collection contient les informations nécessaires pour identifier une collection d’un utilisateur. Elle contient trois attributs : code\_collection (identifiant), nom, description

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.



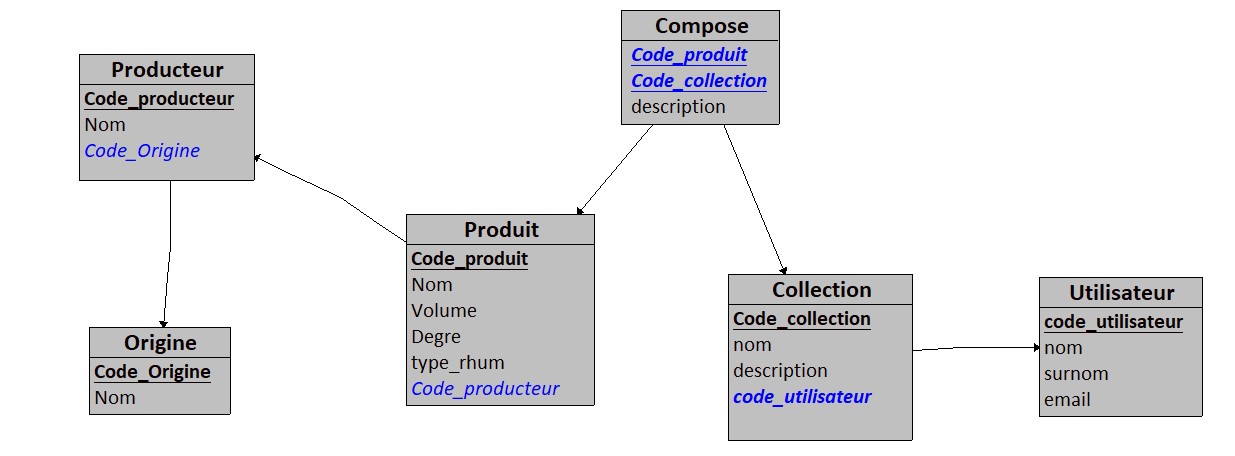
L’entité Utilisateur contient les informations d’identification d’un utilisateur. Elle contient les attributs suivants : code\_utilisateur (clé primaire), nom, surnom, email.

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.

### 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.

# 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.

### 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.

### Fichier excel Inventaire\_A.xlsx

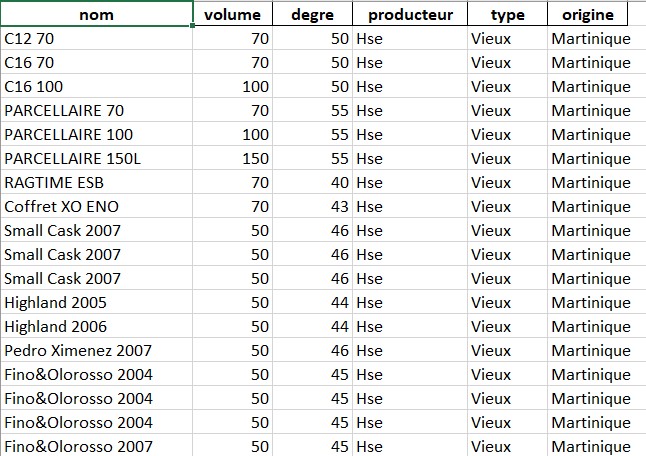
La particularité de ce jeu de données, c’est qu’il contient des données personnelles. Comme il s’agit de mon propre inventaire, la granularité des informations y est poussée. J’y ai inclus les numéros de fûts, numéro de série, mise en bouteille... Ces informations étaient importantes. En cas de vols, 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.

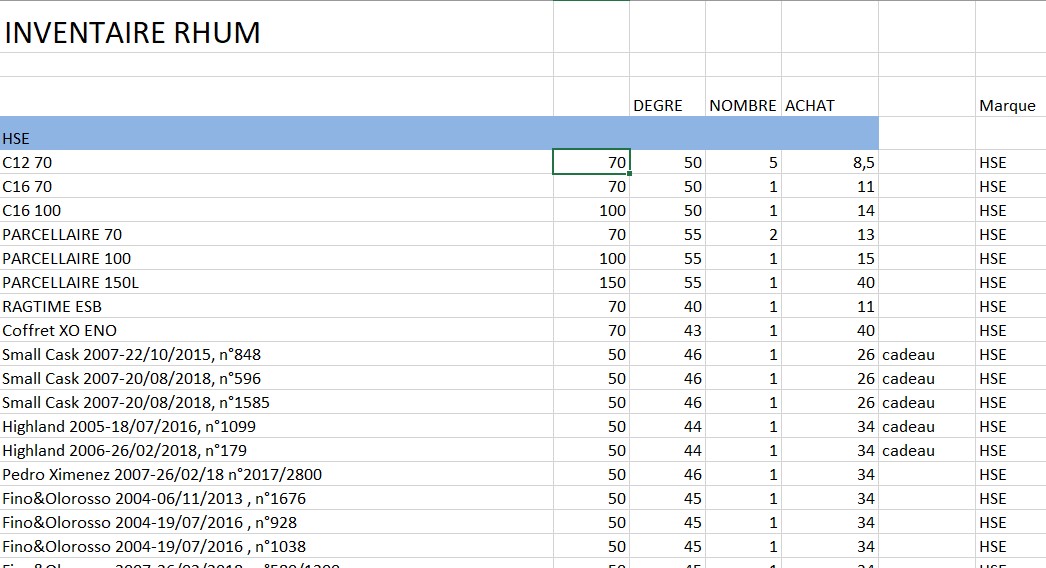
De plus, ce surplus d’information ajoute un surcroît de complexité. Ces informations complémentaires n’apportent rien de plus pour identifier une référence et elles pourront être rajouter plus tard dans le processus.

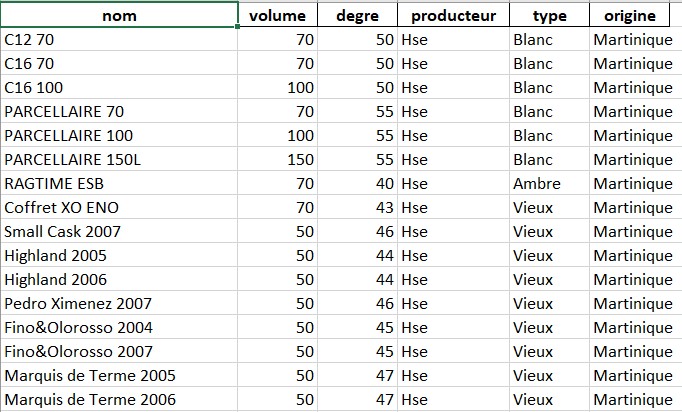
J’ai donc décider d’anonymiser le fichier. J’ai dans un premier temps, transférer dans une liste les informations de la colonne « nom ». Puis par une boucle, j’ai appliqué sur les chaînes de caractères la méthode split() sur le caractère ‘–‘. Enfin je récupère la première partie (index [0]) de ma liste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

Pour ce fichier, j’ai rajouté une colonne « type ». Je l’ai implémentée directement dans la dataframe et je lui ai attribuée 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 :

Version Intermédiaire

Version Brute

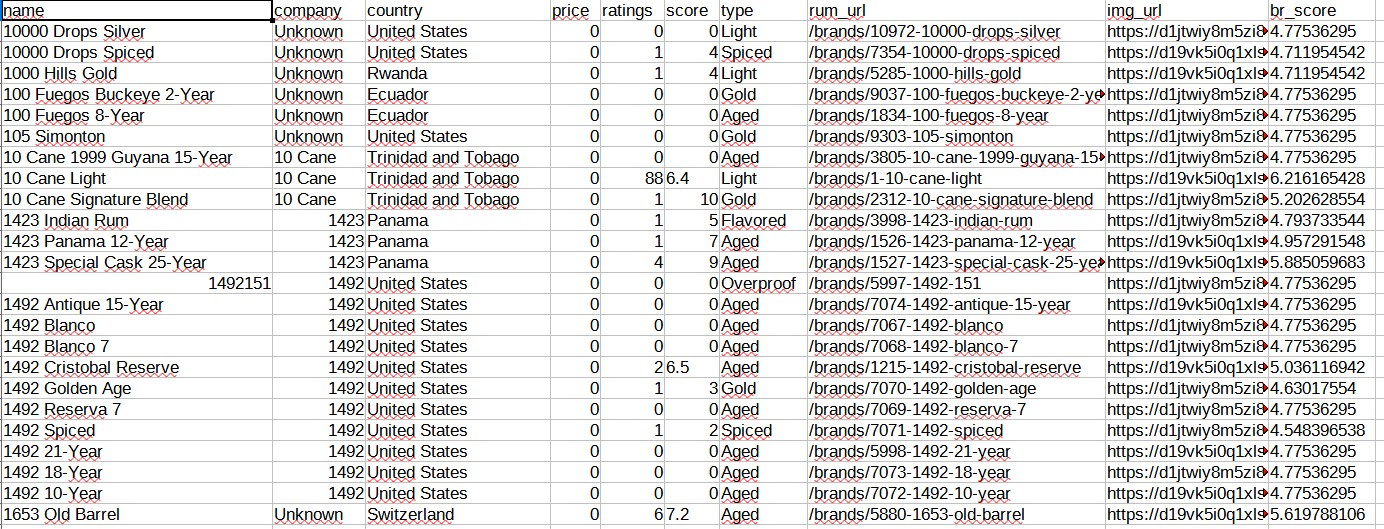
Version Finale.

### 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 grand 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 ».

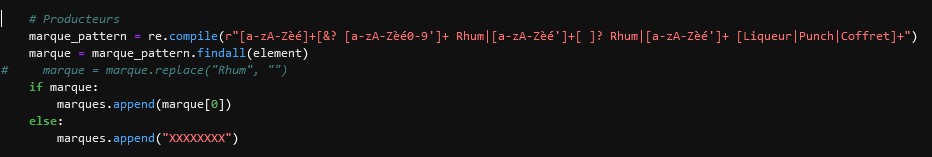
Version brute

Version finale

### Fichier test.json

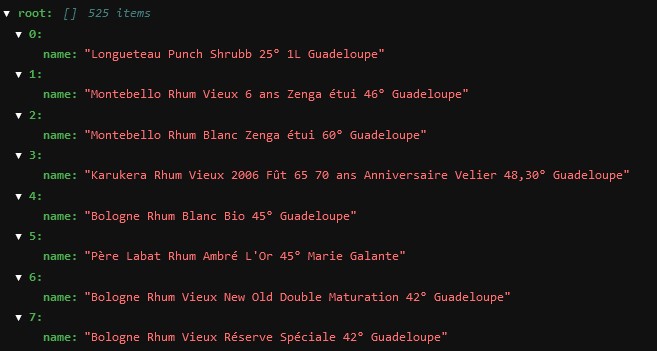
J’ai utilisé des expressions régulières pour extraire les informations nécessaires. Je me suis aidé du site Pytex (https://pythex.org/) qui permet de tester les expressions régulières.

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



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.

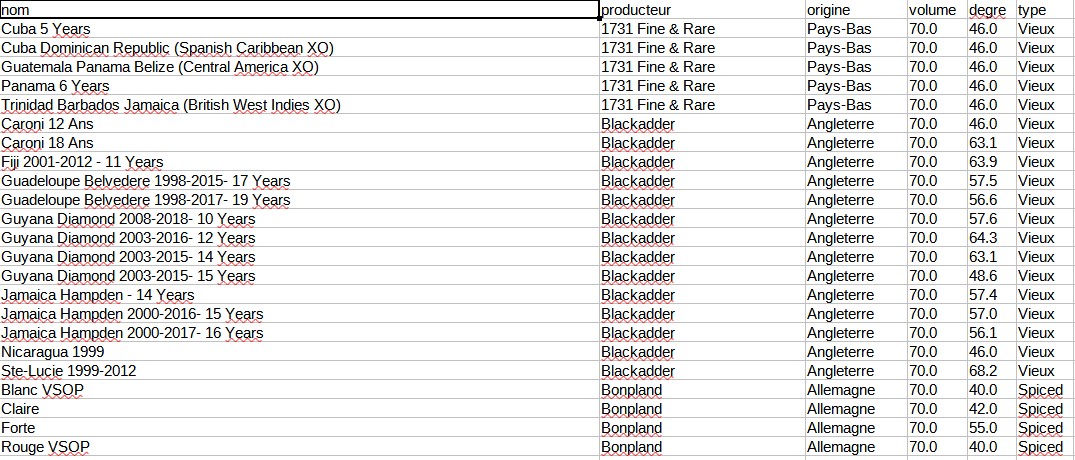
Version finale

Version brute

### 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 ».

Version brute

Version finale

# 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

#### 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

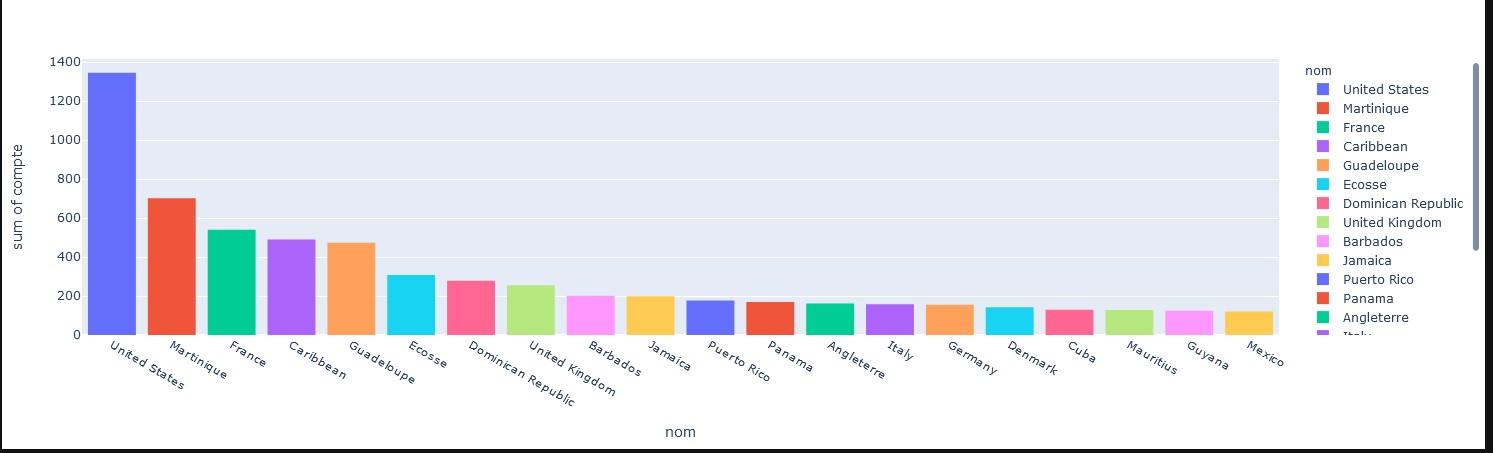
Limit 20;

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 traduits. 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 développements.

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



#### 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;

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 fondée assez récemment et Cadenheads, une ancienne institution). Plantation est une marque française détenue 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 reconnue 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.

#### 

#### Troisième requête : Les références martiniquaises

Cette requête nous permet d’afficher le nombre de références 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

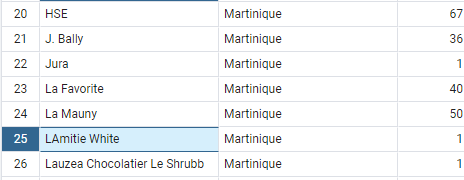
inner join produit

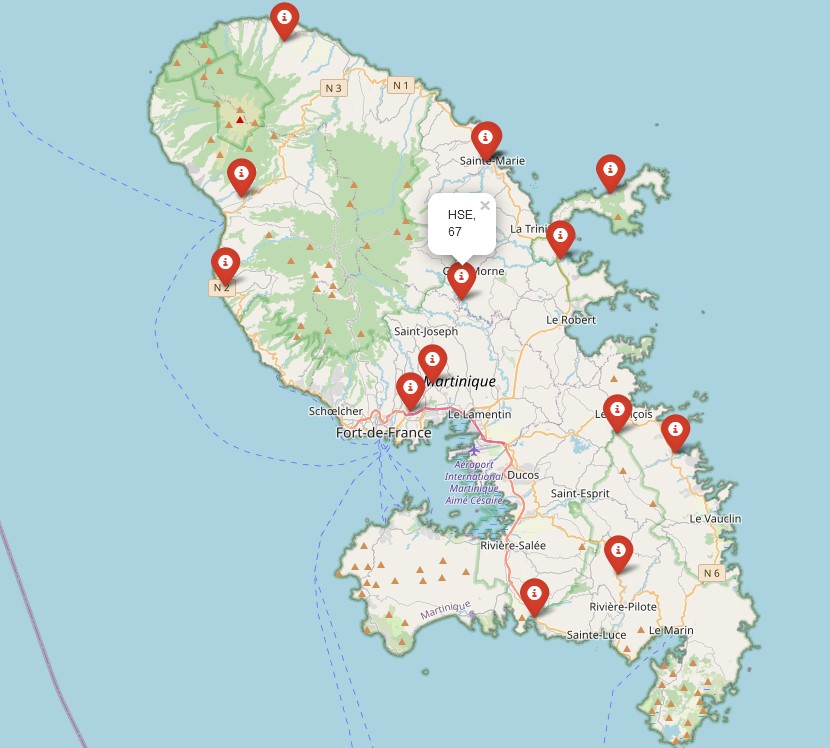
on producteur.code\_producteur = produit.code\_producteur

where origine.nom = 'Martinique'

group by producteur.nom\_producteur, origine.nom;

""")





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 situent 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 technologies 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.

Mayel PELLE

# 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>

<https://spirits-station.fr/spiritueux-chiffres-statistiques-france>