MEMOIRE

Développeur Data

Janvier 2021

Anthony JACQUEMIN

108 chemin de l'Olivet
Parc Auréa entrée D
06110 Le Cannet

06 63 78 65 03
anthonyjacquemin@hotmail.fr

SIMPLON
CANNES

Those who say it cannot be done

should not interrupt those doing it

Inconnu

Table des matières

Etat de l'art

Comme le précisait déjà le magazine The Economist sur une de ses couvertures en 2017, "*the world's most valuable resource is no longer oil, but data".* En effet, plusieurs éléments appuient ce constat :

- les ressources pétrolifères se tarissent alors que les données deviennent de plus en plus conséquentes : on estime le volume des données doublé tous les 2 ou 3 ans avec presque 50 zettaoctets (10^{21} octets) créés ces 10 dernières années
- les fortunes du début du 20ème siècle, composées de magnats du pétrole tel Rockefeller, ont laissé place aux patrons de la tech Elon Musk, Jeff Bezos et Bill Gates

Il faut toutefois prêter attention car l'analogie ne s'arrête pas là :

- cette ressource n'est détenue que par quelques gros acteurs qui s'en accaparent une exclusivité pressante
- des inquiétudes sont présentes quant à la mainmise des géants de la tech (et pas seulement de la part des GAFA) et surtout leur exploitation
- la dépendance à ces données et à leurs sociétés exploiteuses devient critique : de la même manière qu'au siècle dernier, la voiture à essence était le moyen de locomotion phare dont tout le monde dépendait (et dépend encore), qui pourrait se passer aujourd'hui du moteur de recherche de Google et il devient délicat pour beaucoup de se passer entre autres de plateformes de vente comme Amazon (acheteur comme vendeur) ou de certains réseaux sociaux (utilisateur comme publicitaire)
- tel le pétrole en plastique ou gaz, les données brutes n'apportent que trop peu d'information et doivent être « raffinées » pour apporter de la valeur

La donnée devient de plus en plus le cœur des enjeux au sein d'une entreprise comme élément d'analyse, de communication, de compréhension, d'aide à la décision. Les organisations essayent d'ailleurs de prendre le train en marche et de combler des déficiences marquées. D'après la société de conseil Gartner, 80% d'entre elles auraient amorcé le processus pour 2020 et 50% percevraient déjà ce manque dans l'achèvement de plus-value. Les employés ne sont pas en reste et s'estimeraient à hauteur de 75% mal à l'aise avec la manipulation des données selon l'étude de l'entreprise Accenture de l'année dernière également.

Les termes de culture des données ou *data literacy* sont d'ailleurs de plus en plus utilisés pour désigner la compétence associée et on considère désormais le cycle de de la donnée dans son intégralité : on cherche à l'identifier, la collecter, la vérifier, l'analyser, la rendre disponible et en décider l'accès, déterminer son utilisation.

Ce contexte laisse présager des opportunités d'emploi à différentes étapes :

- la collecte et le traitement avec des métiers de Data Architect ou Data Engineer dans le cadre de données clients ou d'historique des cours de prix (et respectivement Big Data Architect et Big Data Engineer dans le cadre du big data pour les réseaux sociaux ou le suivi sur marché financier)
- la gestion et le stockage des données avec le métier de Database Administrator
- l'analyse de données et l'extraction de facteurs pertinents ou *insights* avec établissement de tableaux de bord ou *dashboards* dans un cadre de Business Intelligence avec des métiers de BI Analyst, BI Consultant ou BI Developer
- l'établissement de prédictions par des méthodes statistiques dans le cadre de détermination d'expérience utilisateur ou de prévisions de vente avec des postes de Data Analyst ou Data Scientist, ce dernier étant amené aussi à utiliser l'intelligence artificielle, plus spécifique au Machine Learning Engineer, dans le cadre de la rétention de client ou la détection de fraude
- des métiers annexes concernant par exemple la protection et l'utilisation des données, vis-à-vis entre autres du RGPD, avec des postes de Data Protection Officer

Certaines technologies font figures de proue dans le domaine. On peut citer côté langages de programmation R, Python ou Matlab, même si certains comme Scala ou Java sont utilisés, voire même C ou C++ dans le cadre de l'IA. Par contre, le choix est partagé entre BDD SQL ou NoSQL en fonction des données. Des logiciels tels Power BI, Tableau, Qlik ou SAS se démarquent pour l'analyse des données quand Excel ne devient plus suffisant. Pour le déploiement sur le cloud, Amazon Web Service, Google Cloud et Microsoft Azure prédominent sur le marché.

Sous ces apparences parfois sombres et (trop) pragmatiques et mercantiles, de nombreux exemples illustrent les bénéfices que la *data* et son utilisation peuvent apporter :

- les données satellitaires (images visuelles, infrarouges…) amènent à la prédiction de la récolte des champs, la surveillance des pipelines pour éviter les incendies ou pour évacuer les citoyens lors de catastrophes naturelles (ce sujet a d'ailleurs été évoqué par l'équipe Simplon lors du dernier ActInSpace)
- l'analyse des images radiographiques et la détection automatique des métastases dans les ganglions ont assisté les oncologues dans la prévention des cancers du sein et permis une véritable synergie. Le système par apprentissage automatique du MIT était précis à 92% mais les humains l'étaient à 96% ; l'association des expertises des pathologistes à l'analyse automatique a permis d'atteindre 99,5% de précision et sauver 56 000 patients rien qu'aux Etats-Unis grâce à une meilleure prise en charge des patients et une prescription d'un meilleur traitement (toujours grâce à l'IA)
- les véhicules autonomes et toutes les données traitées y afférentes ont abouti à une diminution des accidents de circulation, une meilleure fluidité du trafic mais aussi indirectement à une diminution des gaz à effet de serre (de 40 à 60% selon l'université de Poznan)

Préambule

Seul un projet sera présenté dans ce mémoire par souci de simplicité. En effet, le projet Cliiink recoupe l'ensemble des compétences vues au cours de la formation, même s'il n'est pas aussi avancé que certains projets sur certains aspects (par exemple le projet de stage Perrenot pour Flask, le projet SDIS pour la modélisation ou encore le projet Netflix pour le traitement des données), et il aurait semblé peu judicieux de détailler tous ces projets pour évoquer juste quelques points spécifiques (qui peuvent être néanmoins évoqués si nécessaire lors d'entretiens ultérieurs).

L'ensemble du code source, des graphiques, de la base de données et des documents de suivi et de présentation du projet sont disponibles sur le dépôt GitHub suivant : https://github.com/antjacquemin/cliiink
Pour la partie relevant strictement de la période de stage, ce dépôt est également disponible : https://github.com/antjacquemin/cliiinkv0
Il se décompose de la manière suivante :

- bdd — Base de données et modélisation
- data — Jeux de données reçus ou créés
- doc — Documentation et gestion de projet
- src — Code source

La rédaction de ce mémoire s'appuie quasiment sur l'ordre chronologique suivi au fil du stage, à l'exception des compléments intégrés directement dans les sections concernées.

Projet de stage : Cliiink

Contexte

ENVIRONNEMENT ET ENJEUX

Le climat est en train de changer avec des conséquences inédites :

- sur l'accès à l'énergie, à l'eau
- sur les équilibres géostratégiques
- sur les mouvements de population
- sur les écosystèmes

Le changement climatique est malheureusement déjà enclenché jusqu'à 2050 malgré des initiatives prises. Sans initiative forte prise sur la décennie 2020-2030, le réchauffement climatique pourrait s'aggraver et dépasser les 4°C avec des conséquences alarmantes :

- accentuation des risques naturels comme les marées, crues, tempêtes et incendies
- diminution des ressources en eau avec principalement une limitation des nappes phréatiques et une réduction des cours d'eau de 30 à 60 % en été
- fonte continue des glaces avec diminution de la banquise, voire une disparition totale en été
- hausse du niveau de la mer de 52 à 98 cm d'ici 2100 avec disparition de zones côtières et d'îles
- perturbations sur différents secteurs économiques, notamment l'agriculture et le tourisme, avec des rendements affaiblis et des risques de déplacement des zones d'activité
- dangers sanitaires avec risque de surmortalité du fait, entre autres, des canicules, des contaminations et l'arrivée de nouvelles maladies

Des mesures sont à prendre à l'échelle globale avec en particulier la réduction des émissions de gaz à effet de serre, tel que prévu par l'accord de Paris 2015 lors de la COP21 pour contenir le réchauffement moyen en dessous de +2 °C, et la valorisation des économies d'énergie. Par contre, les mesures sont surtout à considérer à l'échelle locale, par exemple au sein des collectivités. En effet, même si cela ne paraît pas intuitif, le GIEC (Groupe d'experts Intergouvernemental sur l'Evolution du Climat) situe 50% à 70% des leviers d'action au niveau territorial.

CACPL

La Communauté d'Agglomération Cannes Pays de Lérins (CACPL) a été créée le 1er janvier 2014 suite à la loi sur la réforme territoriale et regroupe les 5 villes de Cannes, Le Cannet, Mandelieu-la-Napoule, Mougins et Théoule-sur-Mer. Ces dernières ont mutualisé au sein de la collectivité des services tels que les systèmes d'information et de télécommunications et l'aménagement du territoire. Elles exercent désormais en commun certaines de leurs compétences obligatoires selon la loi comme les transports, l'environnement, le développement économique, l'aménagement du territoire et la politique de la ville, notamment sur l'emploi, et certaines facultatives confiées par les communes dans le cadre d'un projet territorial comme la construction ou la gestion dans les domaines de la culture et des sports.
Elle se caractérise par une superficie d'environ 95 km² et une population, dénombrée en 2017, de 160 557 habitants, faisant d'elle la 2ème agglomération des Alpes-Maritimes derrière la Métropole Nice Côte d'Azur.

En 2019, elle a été lauréate des Trophées Climat Energie pour son engagement et plus particulièrement pour l'amélioration du réseau de points d'apport volontaire de verre par sa densification et sa modernisation avec de nouveaux compacteurs à déchets.

Parmi les pôles constitutifs de la CACPL, le Pôle Environnement Cadre de Vie et Transition Énergétique est consacré à la coordination, au pilotage et à la gestion des déchets ainsi qu'à la collecte. Il a supervisé, pour l'année 2019, 165 agents ayant effectué 14 900 tournées pour 10 679 tonnes d'emballages ménagers recyclables collectées, 77 136 tonnes d'ordures ménagères, 5 133 tonnes de verre et 540 tonnes de textile.

CLIIINK

Cliiink est une solution pour faciliter la gestion du tri, commercialisée depuis 2018 par la société Terradona. Il est composé d'un dispositif électronique qui s'adapte à tous les conteneurs de tri, de l'application mobile pour les usagers et et d'une plateforme de suivi en temps réel pour les collectivités. Pour l'utilisateur, cela se traduit de manière quotidienne par le repérage d'un conteneur à proximité et le déblocage du dispositif électronique via son téléphone portable afin qu'il puisse déposer ses déchets triés en échange de points qu'il pourra utiliser pour bénéficier de différentes promotions à valoir dans différents commerces locaux.

Cliiink s'est développé dans un premier temps dans la région PACA et reste encore un dispositif récent dans l'agglomération qui l'a mis en place en novembre 2018 sur les collecteurs de verre. Les retours sont toutefois déjà prometteurs avec 100 bornes équipées sur les 280, 145 commerçants partenaires, plus de 100 bons d'achat et remises en ligne, plus de 4 300 comptes utilisateurs au service et environ 3 400 000 dépôts atteints fin 2020. L'agglomération en a d'ailleurs fait la promotion encore récemment et a lancé plusieurs initiatives en fin d'année dernière, dont le 1er challenge inter-associations, finissant le 31 janvier 2021 et visant à récompenser les associations ayant le plus collecté, et la semaine du pouvoir d'achat en novembre dernier visant à redynamiser les commerces locaux sur les biens alimentaires.

OBJECTIFS DU STAGE

La CACPL souhaite être orientée dans sa gestion actuelle des déchets et mieux appréhender les facteurs de performance et les éventuelles optimisations à apporter à leur environnement actuel. Elle a initialement et volontairement laissé le sujet ouvert pour laisser libre court aux initiatives et suggestions personnelles et apporter tous les éléments d'interprétation que l'on juge pertinents. Après un premier entretien avec les responsables du service Environnement et du service Informatique, certains axes leur semblaient prédominants :

- la pertinence et l'efficacité du dispositif Cliiink pour le verre
- la répartition des collecteurs dans l'agglomération
- l'impact des professionnels dans la collecte, en particulier certains secteurs d'activité comme la restauration

Il y a été convenu des données supplémentaires nécessaires pour répondre au mieux aux demandes du client, dans la limite de la disponibilité des ressources par la CACPL.

TECHNOLOGIES UTILISEES

Pour des raisons de commodité, il a été choisi d'utiliser les technologies communes et connues par tous les apprenant du stage. De même, aucune contrainte n'était imposée par la CACPL.
Par conséquent, le langage de programmation Python et la base de données MySQL ont été choisis. Les deux présentent l'avantage à la fois d'être open source, et donc libres d'accès, et d'être très utilisés avec une forte communauté et documentation en ligne.
Des bibliothèques logicielles, méthodes de modélisation, technologies Web et méthodologies de gestion de projet ont été également exploitées et seront évoquées dans ce mémoire dans leur partie respective.

Planification et gestion

CAHIER DES CHARGES

Comme introduit auparavant dans les objectifs du stage, une note de cadrage a été rédigée dès la première semaine puis clarifiée la semaine suivante lors du premier entretien, amenant dès lors à la rédaction du cahier des charges.
L'objet du stage s'est orienté autour de ces axes :

- l'exploitation des données fournies
- leur nettoyage préalable et la détection d'anomalies
- leur transfert sur un SGBD (Système de Gestion de Base de Données) adéquat en données structurées
- la récupération de jeux de données complémentaires
- l'établissement de statistiques et graphiques pertinents
- leur exploitation via une interface simplifiée
- la mise en avant de facteurs significatifs de performance et d'optimisation

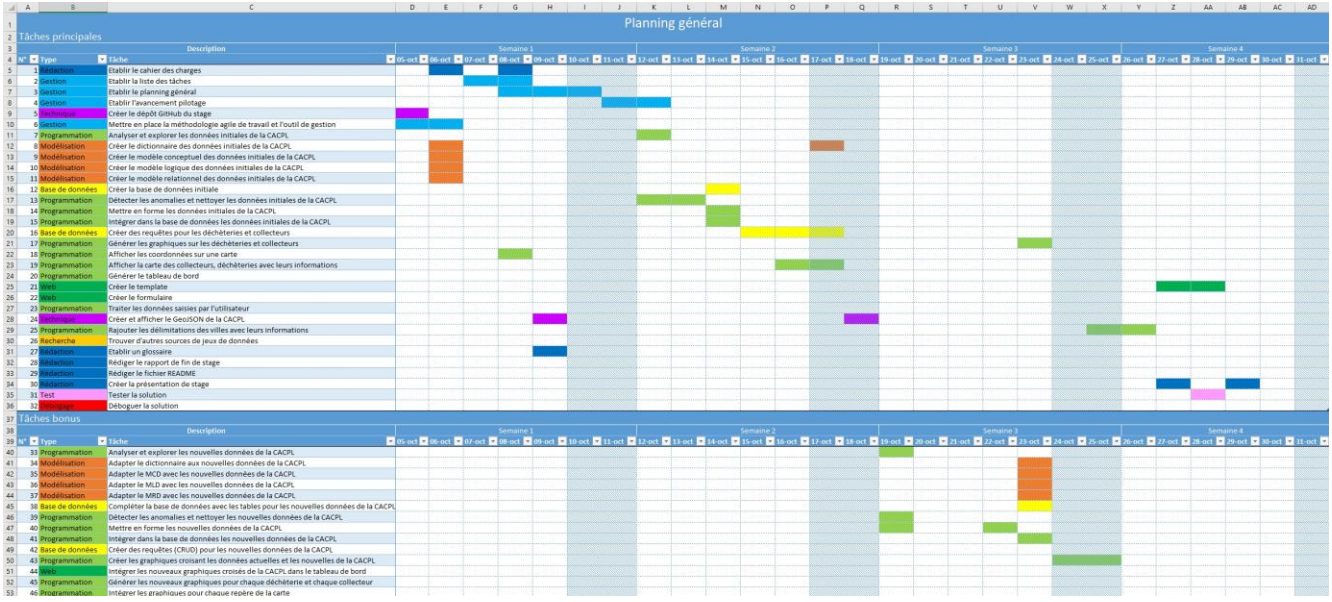
Les critères d'évaluation, les conditions opérationnelles et la méthodologie ont été spécifiés.
Un calendrier prévisionnel a réparti pour chacune des 4 semaines les différents segments :

- formalisation du contexte et des attentes, établissement du cahier des charges, de la planification et la gestion du projet, organisation des environnements et méthodologies de travail
- exploitation des jeux de données fournis et recherche de données complémentaires, transfert vers une base de données et création des premières statistiques
- création des graphiques et de l'interface de visualisation des données
- analyse et mesure des indicateurs de performance et d'optimisation

PLANNING ET AVANCEMENT

Une planification a été établie avec, dans un premier temps, la décomposition du travail en une liste de tâches principales pour fournir la solution attendue et une liste des tâches bonus permettant de compléter le produit avec des suggestions d'évolution et pouvant être effectuée par la suite lors de cette formation, dans le cadre de projet pédagogiques, ou lors d'une formation suivante plus en adéquation avec les connaissances requises. Plusieurs points y ont été apportés tels que l'estimation du temps de travail, l'objectif explicité avec ses critères d'achèvement et son livrable attendu pour chacune des tâches, ainsi que les dépendances entre elles. Ces tâches ont été également distinguées selon leur type (programmation, web, modélisation…) avec un jeu de couleurs défini pour en faciliter l'appréhension.

Dans un second temps, un planning général a été créé pour le mois de stage et complété au fur et à mesure de la complétion des tâches.



Le pilotage du projet a pu être surveillé via une feuille d'avancement disposant de quelques indicateurs permettant de mesurer l'avance ou le retard par rapport à la durée estimée.

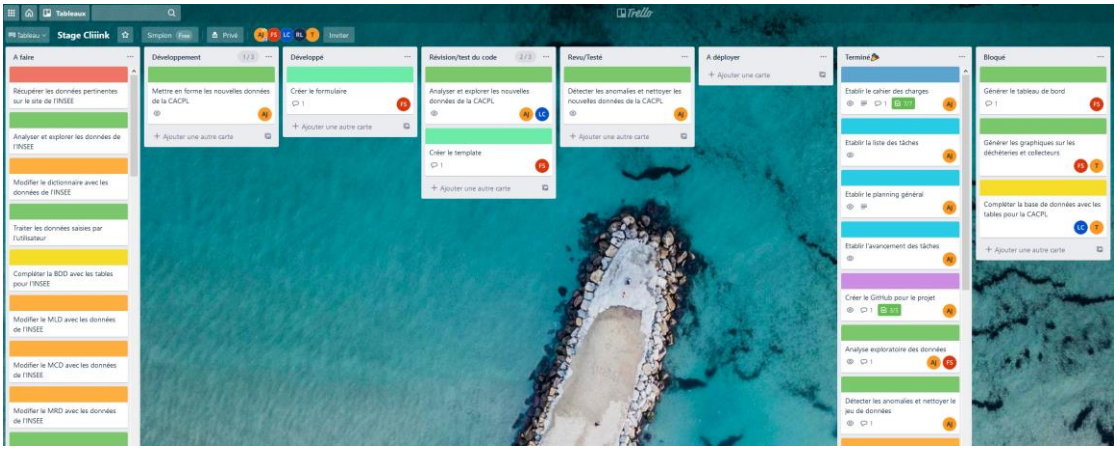
METHODE AGILE ET KANBAN

L'équipe étant réduite (4 membres) et autoorganisée, un cadre méthodologique léger et flexible était nécessaire pour répondre de manière pragmatique et adaptative aux besoins du client par une approche agile.
Le choix entre Kanban et Scrum s'est vite porté sur le premier pour plusieurs raisons :

- l'effectif de l'équipe ne permettait pas d'allouer des postes spécifiques à chacun et se serait alors centré sur une équipe « Développeurs » en auto-gestion sans Product Owner ni Scrum Master
- le sprint en Scrum, avec sa liste définie au préalable de fonctionnalités pendant l'itération, était trop contraignant en empêchant tout rajout ou toute modification en cours ; un flux continu de tâches en fonction des nouvelles demandes du client était à privilégier, surtout vu le temps court (4 semaines) disponible
- les objectifs n'étant pas forcément définis dès le départ et très évolutifs, une réactivité accrue aux changements était indispensable
- les sujets abordés étant très différents, Kanban semblait plus adapté pour leur traitement au sein d'une même équipe
- la répartition des tâches demandait une forte flexibilité

APPLICATION AVEC TRELLO

Kanban a pu être pratiqué, non pas sur un tableau ou *board*, mais en ligne via l'outil de gestion de projet Trello pour visualiser et harmoniser le travail d'équipe.



Pour faciliter la description du flux opérationnel ou *workflow*, le processus a été décomposé suivant ces colonnes : A faire, Développement, Développé, Révision/test du code, Revu/Testé, A déployer, Terminé. Une attention a été portée pour s'assurer que la personne qui teste le code soit différente de celle qui l'a développé. Une colonne Bloqué a été également rajoutée pour directement détecter d'éventuelles difficultés, rencontrées par le développeur ou dues à un manque d'information, alerter et mobiliser les autres membres, et parer ainsi potentiellement tout effet d'engorgement.

Chaque unité de travail a été organisée par *card* où l'on pouvait y retrouver un titre, une description, son responsable. Des compléments étaient parfois rajoutés, comme la subdivision des points à traiter ou des commentaires sur d'éventuels problèmes rencontrés ou tâches découvertes à rajouter.
Le jeu de couleurs du planning général a été repris pour distinguer facilement les types de tâches par *card* et permettre à chacun de choisir en fonction de ses préférences ou facilités.

Pour éviter la dispersion des développeurs sur un multi-tâches contre-productif, il a été décidé de limiter le travail en cours ou le *Work In Progress* sur plusieurs aspects. En particulier, les *cards* ont été limitées à 3 pour les colonnes Développement et Révision/Test du code afin d'avoir au moins 1 personne rattachée sur d'autres segments et éviter alors tout goulot d'étranglement.

Dans une moindre mesure, il a été essayé d'uniformiser les temps d'accomplissement des *cards*, ou *lead time*, en fonction des compétences de chacun.

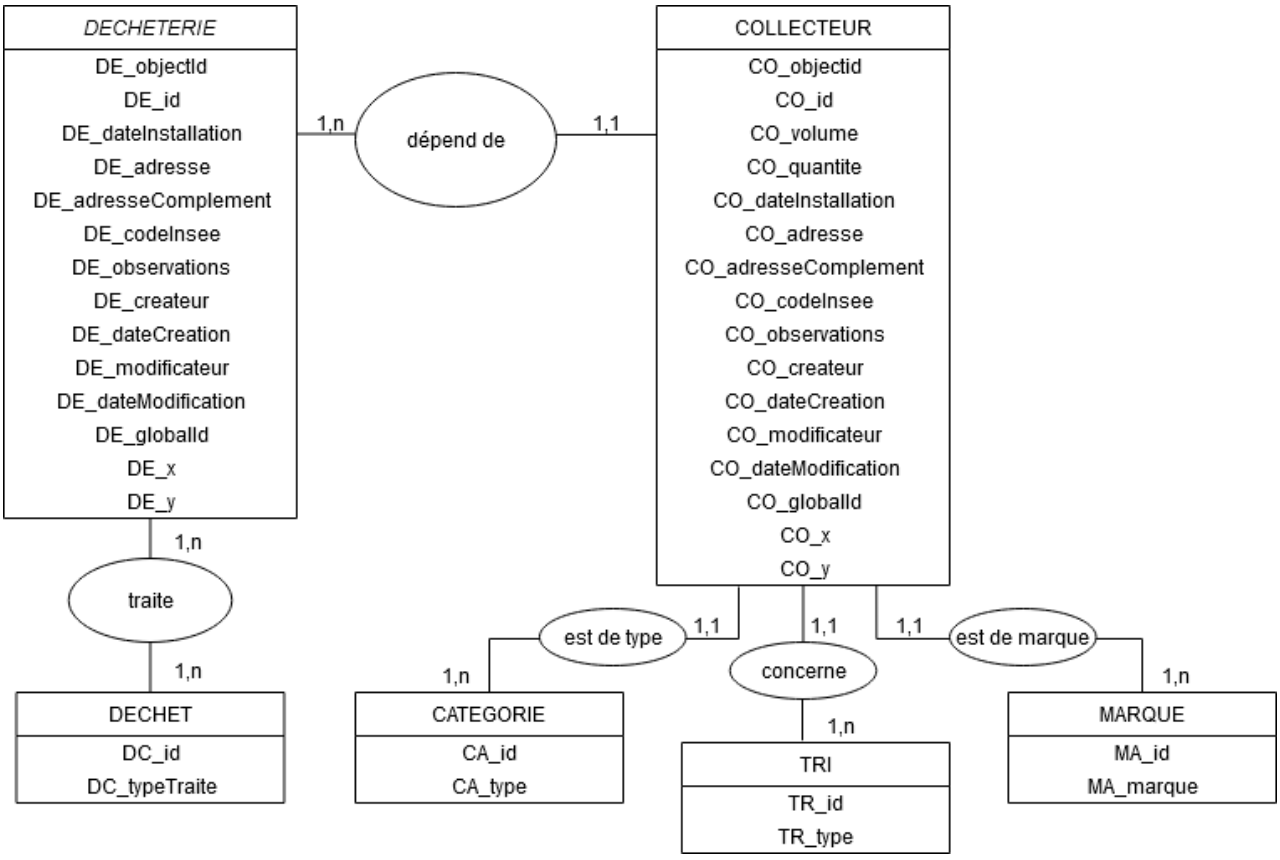
Analyse et modélisation

JEU DE DONNEES INITIAL

Le jeu de données initial correspond à l'inventaire des déchèteries et collecteurs sous forme respectivement d'1 jeu de données (3 éléments et 28 colonnes) et de 5 selon le type de tri (967 sites de collecteurs et entre 16 et 20 colonnes). Un lot complet de 942 photographies associées était fourni et s'est avéré bien plus utile que prévu.

Les données ont été complétées par l'ajout des codes INSEE manquants mais aussi des volumes et types de collecteur par comparaison directe avec les photographies de collecteurs avec les champs renseignés. S'est suivie une phase de mise en forme des données avec la standardisation des codes INSEE et des dates, la conversion des volumes dans la même unité et la conservation de colonnes pertinentes. Le transfert vers la base de données a été préparée avec l'isolation des tables pour les déchets, les traitements, les types de tri, les types et les marques de collecteurs.

La modélisation théorique se présente ainsi :



Quelques notes sont à souligner lors de la mise en pratique suite à l'analyse approfondie du jeu :

- la relation collecteur-déchèterie n'a pas pu être implémentée alors qu'elle aurait été cruciale pour déterminer et optimiser les tournées de collecte
- les cardinalités collecteur-catégorie et collecteur-marque ont été revues en 0,1 – 1,n pour pallier à l'absence de valeurs pour certaines lignes impossible à compléter

NOUVEAU JEU DE DONNEES

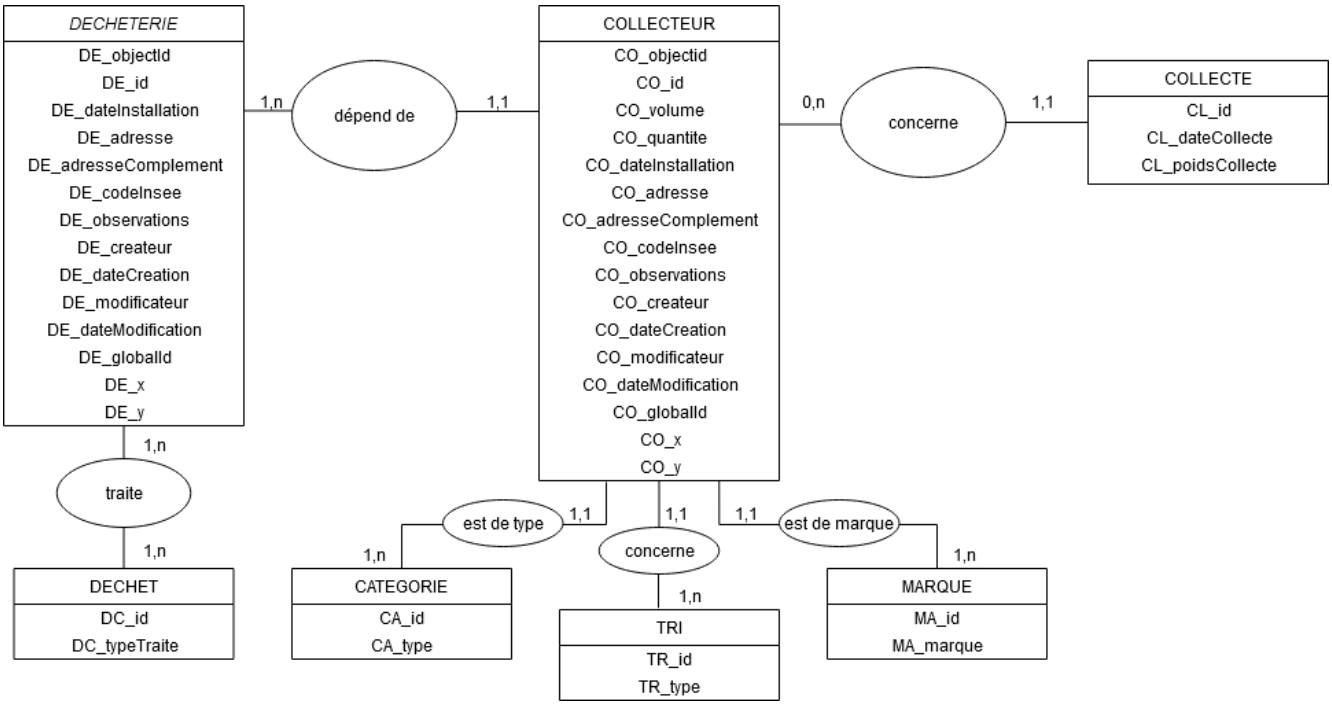
Le jeu de données fourni ensuite par la CACPL quant à Cliiink correspond au tonnage pour les collectes de verre par site sous forme de 2 jeux de données distincts (pour la période de 2017 à 2019 et pour 2020)

Un prétraitement a été accompli avec la suppression des doublons. De manière analogue au jeu initial, les données ont été mises en forme avec la conversion des poids dans la même unité et la conservation de colonnes pertinentes.

La grande difficulté s'est située dans l'identification partielle à la main des collecteurs inventoriés par rapport aux adresses postales, fournies dans le nouveau jeu, n'étant ni standardisées ni mises en association avec les collecteurs. Un géocodage n'a permis d'associer que très peu d'éléments et a empêché de raisonnablement structurer les données dans la base. De plus, il manquait des correspondances entre les secteurs d'une même ville d'une période considérée à la suivante et les cardinalités des collecteurs Cliiink, ou de manière générale de verre, n'étaient pas égales à celles du jeu de collecte.

Le choix a été pris de conserver les fichiers en l'état pour traitement plutôt que de transférer une sous-partie peu significative en quantité des données.

La modélisation ci-dessous n'est donc que la solution théorique qui aurait été appliquée :

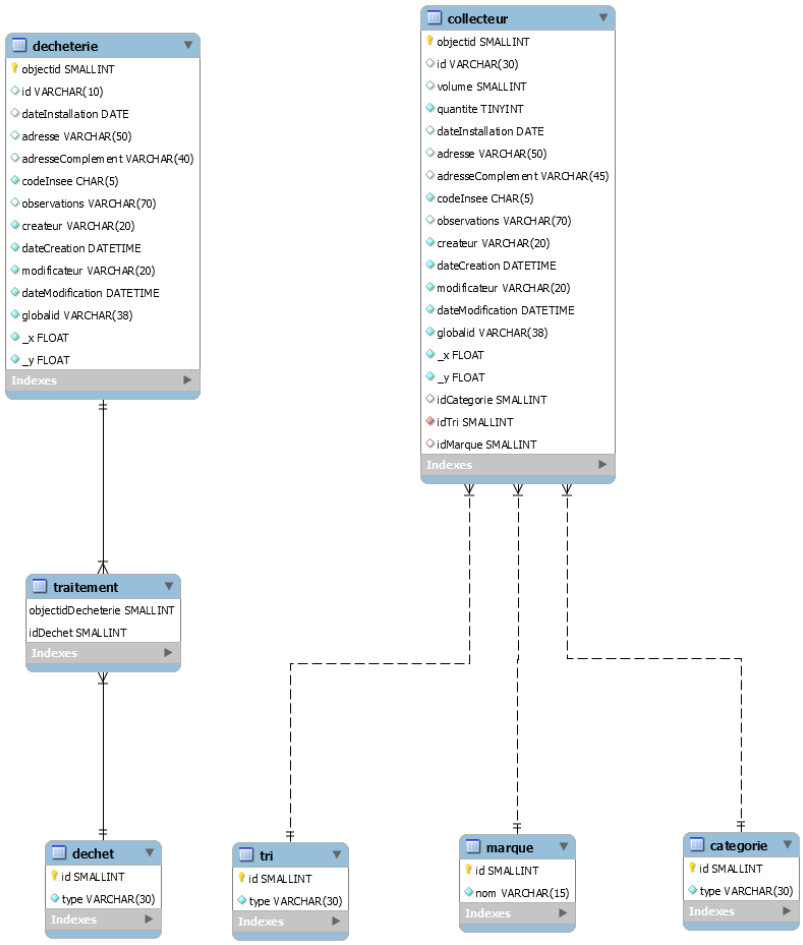


La conception de ces 2 modélisations a suivi l'approche Merise et respecte la 3ème forme normale. Par conséquent, chaque attribut possède une valeur atomique, chaque attribut non clé dépend de l'intégralité de la clé et il n'existe pas de dépendance entre attributs non clés. Ces règles permettent de limiter les redondances et diminuer le volume de données, ainsi que d'éviter toute incohérence.

Base de données

MODELE PHYSIQUE

Comme annoncé dans la section précédente, les tables ont été créées sur MySQL via l'outil Workbench et remplies pour l'inventaire des collecteurs et déchèteries, le transfert des données traitées côté Python ayant été réalisé grâce à la bibliothèque SQLAlchemy.



Toutes les fonctionnalités du CRUD (Create, Retrieve, Update, Delete) ont été apportées pour chaque table. Les noms ont été standardisés afin qu'elles puissent être utilisées librement par un développeur sans connaissance de la base de données exploitée. Des procédures stockées ont été apportées pour le repérage des collecteurs dans une zone à partir de coordonnées.
Un jeu partiel de tests unitaires a été conçu afin de s'assurer de la validité des procédures créées.

ELEMENTS D'AMELIORATION

Pour renforcer les contraintes d'intégrité, différents événements ont été proposés pour gérer les relations entre les clés :

- les traitements sont modifiés ou supprimés lors de la modification ou de la suppression des déchèteries ou des déchets correspondants
- la modification des collecteurs lors de la modification des types de collecteur, des marques ou des types de tri correspondants
- la modification des collecteurs lors de la suppression des types de collecteur ou des marques correspondants (qui peuvent être NULL)
- la suppression des collecteurs lors de la suppression des types de tri correspondants (qui sont indispensables)

Le dernier point peut être retiré (la contrainte redevenant RESTRICT) si la contrainte est considérée comme trop forte (la suppression d'un collecteur étant critique)

Les temps de traitement pour la datavisualisation peuvent être accélérés par la génération de tables de statistiques pour les collecteurs, plus enclins à évoluer en quantité que les déchèteries.
Une table de statistiques partielles avec les nombres de sites et de collecteurs est actualisée dès l'insertion ou la suppression dans la table Collecteur avec des triggers. Une table de statistiques complètes avec le nombre de sites de collecteurs et le nombre de collecteurs par ville, type de tri, marque, type de collecteur et volume, est planifiée annuellement, à un horaire de nuit pour éviter toute surcharge, avec un événement.

Un dispositif simple de sauvegarde mensuelle a été automatisé avec la sauvegarde logique de la base de données et sa reconstitution via son transfert sur une copie de la base de données.

Data visualisation

GENERATION DE GRAPHIQUES AVEC MATPLOTLIB ET SEABORN

Dans le cadre du stage, la bibliothèque Matplotlib sur Python et sa surcouche Seaborn ont été préférés de par leur facilité d'utilisation. Il a pu ainsi être généré de nombreux graphiques sur plusieurs critères :

- pour l'inventaire des collecteurs et des déchèteries
    - le nombre de collecteurs selon la ville, la marque, le volume, le type du collecteur ou le tri considéré
    - le nombre de sites de tri (pouvant regrouper plusieurs collecteurs) selon les mêmes critères
    - la répartition des collecteurs par ville et par type de tri
- pour la collecte du verre
    - le nombre de collectes de verre par année et par ville
    - leur répartition selon les mêmes critères
    - l'évolution du nombre de collectes de verre par secteur et par année pour chaque ville
    - l'évolution du nombre de collectes par mois et par ville
    - le poids collecté de verre par année et par ville
    - leur répartition
    - l'évolution du poids collecté de verre par secteur et par année pour chaque ville
    - l'évolution du poids collecté de verre par mois et par ville

GENERATION DE GRAPHIQUES DYNAMIQUES AVEC PLOTLY

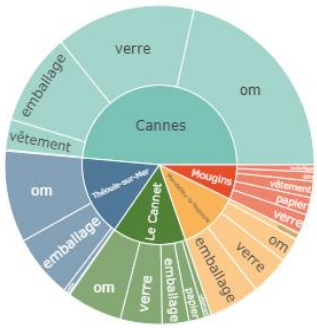
Après le stage, la bibliothèque Plotly a été choisie car elle permettait de rendre interactifs les graphiques (par exemple par sélection/désélection de certaines valeurs) et surtout car ceux-ci devenaient sérialisables et convertibles en JSON. Il n'était par conséquent plus nécessaire de générer les graphiques sous format d'images statiques, avec ses contraintes, que l'on exploitait côté web car on pouvait désormais les transférer en paramètres.
Plotly, en particulier Plotly Express, facilitait la génération de certains graphiques comparé à Seaborn comme certains diagrammes à barre, diagrammes circulaires (Sunburst) ou boîtes à moustaches.
Les graphiques précédents ont été revisités et complétés sur ces points :

- intégration du jeu de couleurs de la CACPL
- ajout d'histogrammes empilés ou par groupe sur le nombre de collecteurs et de sites selon plusieurs critères en même temps (par exemple la ville et un critère annexe)
- ajout d'histogrammes avec distribution de certaines variables quantitatives (comme le volume) sur un axe normé
- ajout de graphiques circulaires à plusieurs niveaux pour les répartitions selon les types de tri et les villes
- ajout de graphiques considérant la période complète de 2017 à 2020 au lieu des traitements séparés auparavant (2017 à 2019 et 2020)
- ajout de graphiques pour les comparaisons temporelles par année pour chaque ville
- ajout de diagrammes en boîtes pour la dispersion du poids collecté selon les secteurs de chaque ville

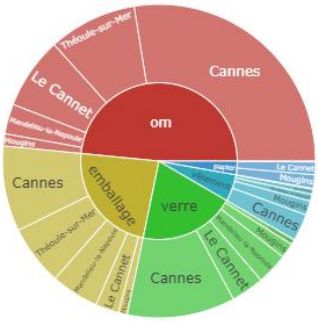
ELEMENTS D'INTERPRETATION

Sans en faire une liste exhaustive, voici quelques éléments d'interprétation intéressants.

Répartition des sites de collecteurs par ville et par type de tri

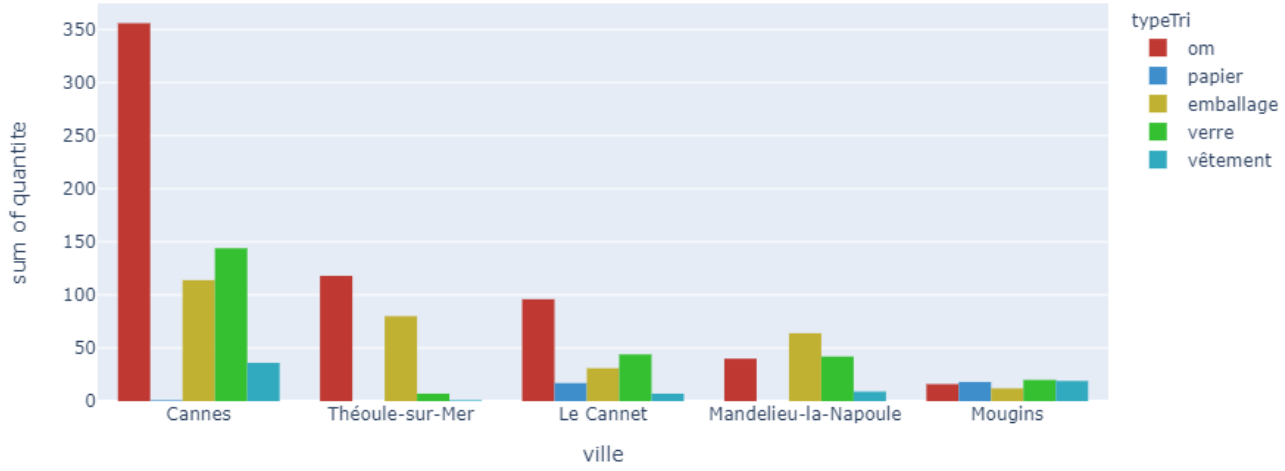


Répartition des collecteurs par type de tri et par ville



La répartition des collecteurs et de leurs sites confirme la prépondérance de Cannes qui regroupe à elle seule la moitié des sites de collectes. Les collecteurs classiques d'ordures ménagères restent logiquement quasi majoritaires avec néanmoins une part importante de collecteurs de verre et d'emballages.

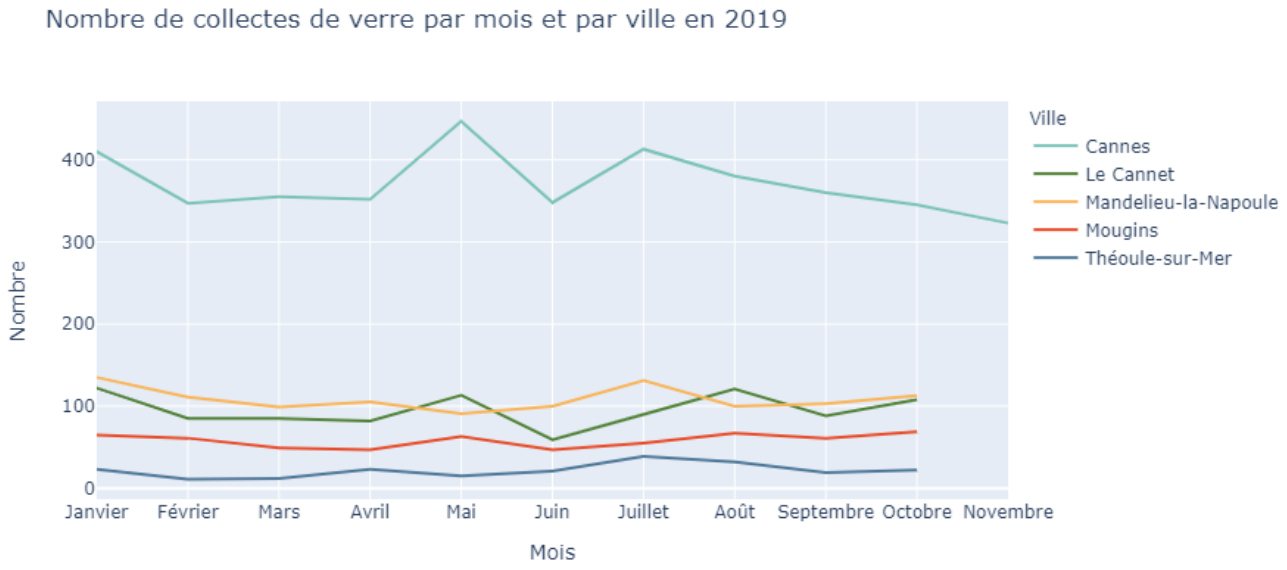
Nombre de collecteurs par ville et par type de tri



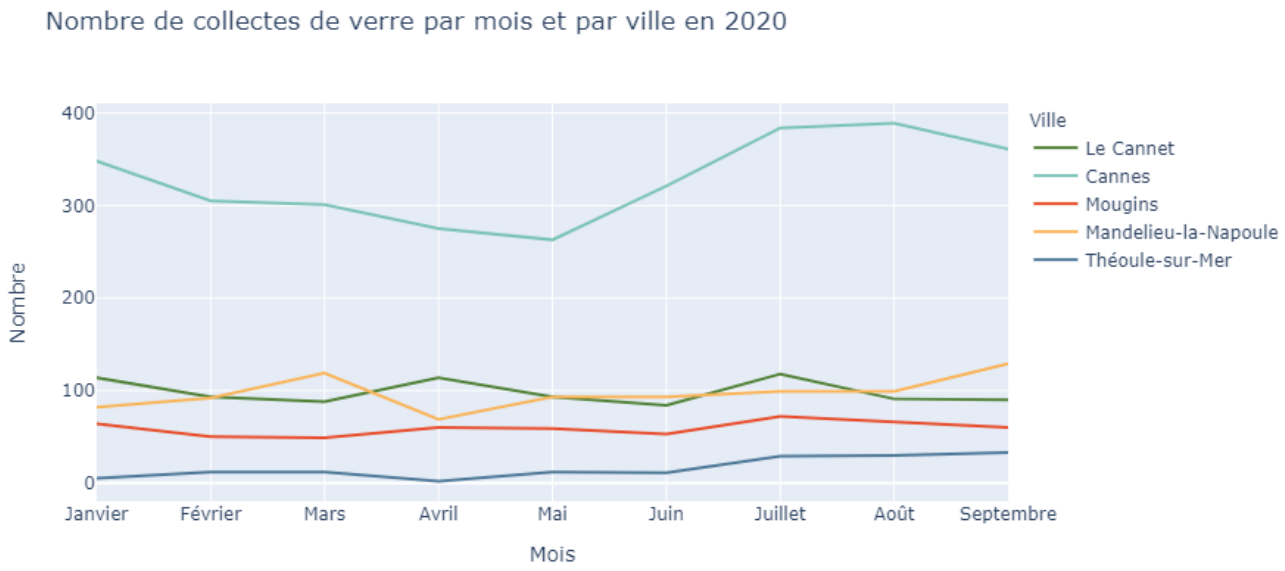
Quelques constats complètent les précédentes observations.

Mougins s'illustre par l'équilibre entre tous les types de tri et Mandelieu-la-Napoule dispose même de plus de collecteurs de verre ou d'emballage que d'ordures ménagères. Cela s'explique en grande partie par la configuration de ces villes où les collecteurs généraux sont centralisés au contraire de ceux de tri spécifique.
La disparition des collecteurs de papier s'explique par leur remplacement par les collecteurs d'emballage.

Les différences dans les collectes entre période normale et période atypique ont pu être mesurées par comparaison entre 2019 et 2020.

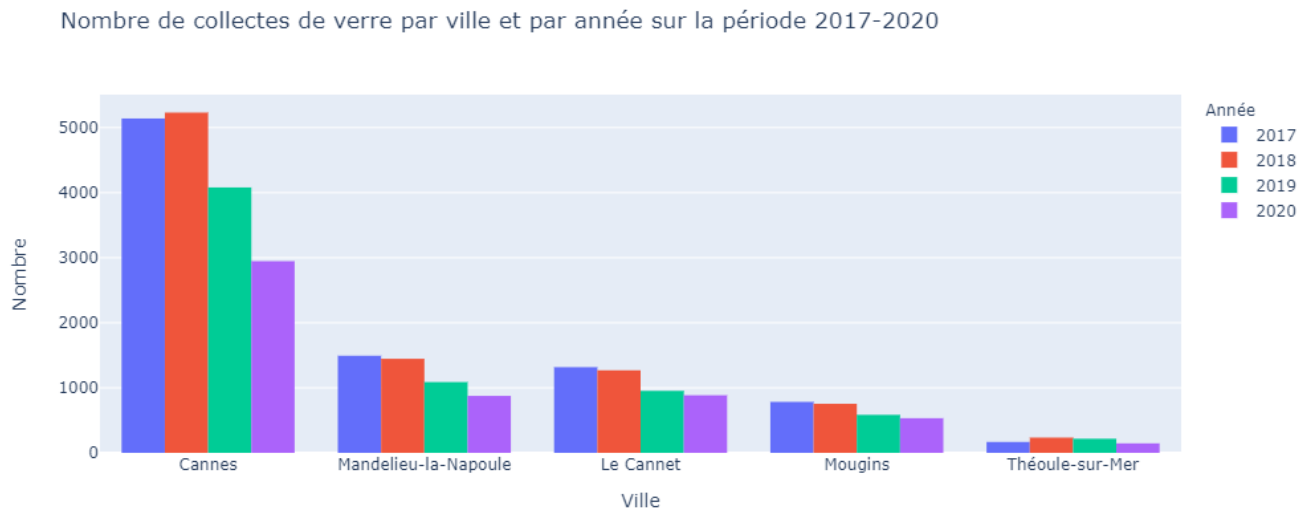


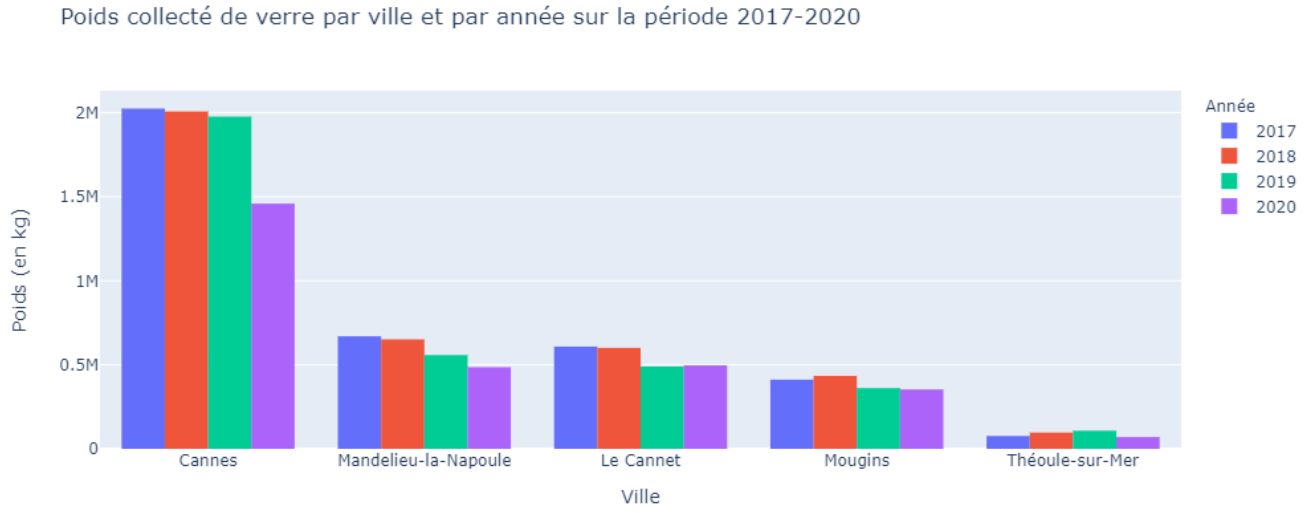
Sans surprise, mai est très significatif pour Cannes avec l'afflux de population lors du festival de Cannes avec même des répercussions pour Le Cannet et Mougins. Un regain lors de la période estivale est à souligner avec une incidence sur juillet ou août selon les villes.



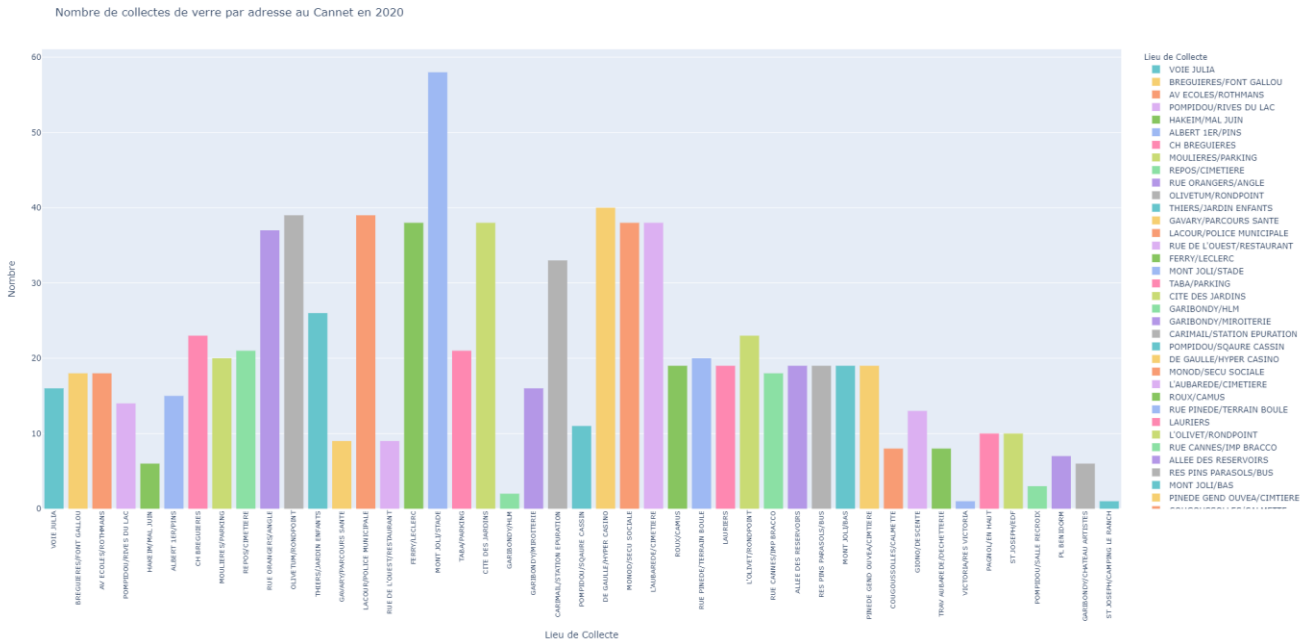
Suite au COVID-19, la chute des collectes a été marquée avec un déclin appuyé pendant la période de confinement. La baisse a été particulièrement marquée pour Cannes en mai 2020 avec l'annulation du festival. Par contre, l'arrivée significative de touristes a lancé un regain en été à Cannes.
Les courbes laissent présager un retour progressif des courbes vers les valeurs passées mais cela reste très sensible à l'évolution actuelle de la crise sanitaire.

De manière générale, les tendances restent identiques que l'on raisonne avec le nombre de collectes ou le poids collecté à l'exception de rares secteurs (la Croisette à Cannes par exemple).

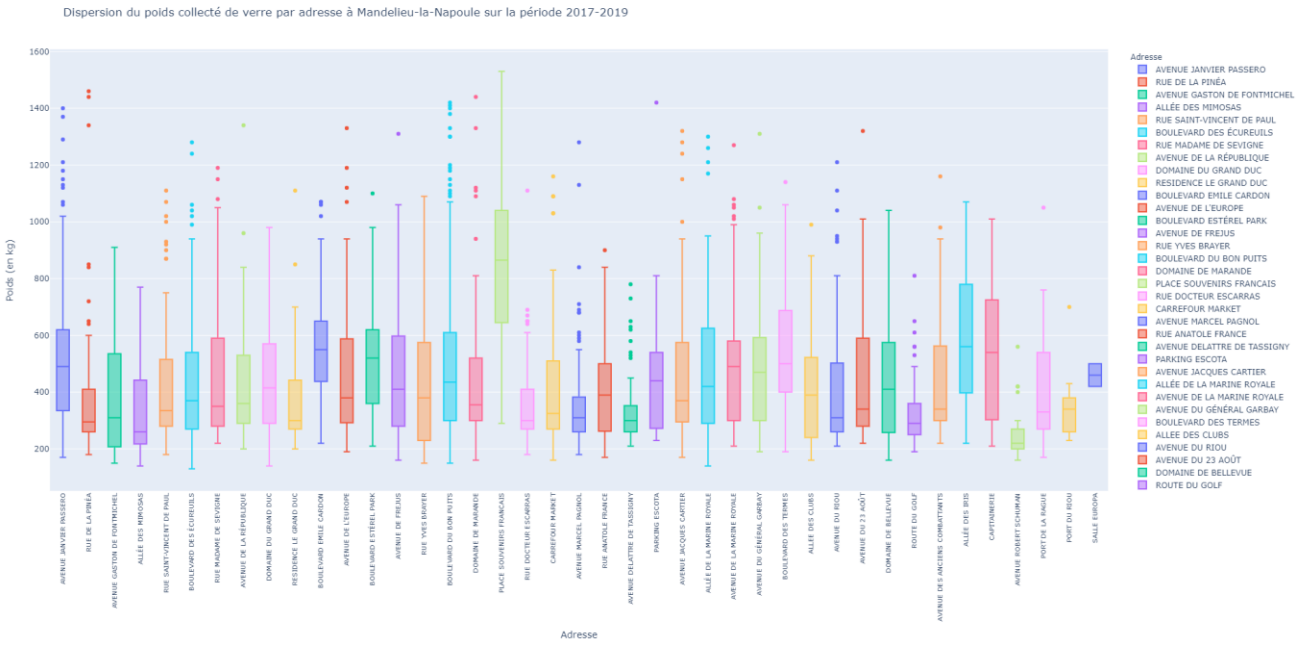




La comparaison entre ces deux paramètres, avec la baisse du nombre de collectes mais le maintien du poids collecté, suggère une meilleure organisation des collectes mais ne peut pas être affirmé car trop de paramètres extérieurs peuvent rentrer en jeu.



Les graphiques sur le nombre de collectes par secteurs dans une ville mettent clairement en exergue les quartiers à sensibiliser lorsque les valeurs sont trop faibles. Certains cas sont justifiés par la présence de travaux lors des périodes considérées.



De même, les boîtes à moustaches illustrent l'irrégularité des poids collectés de verre selon l'étirement des boîtes.

Géovisualisation

FOLIUM

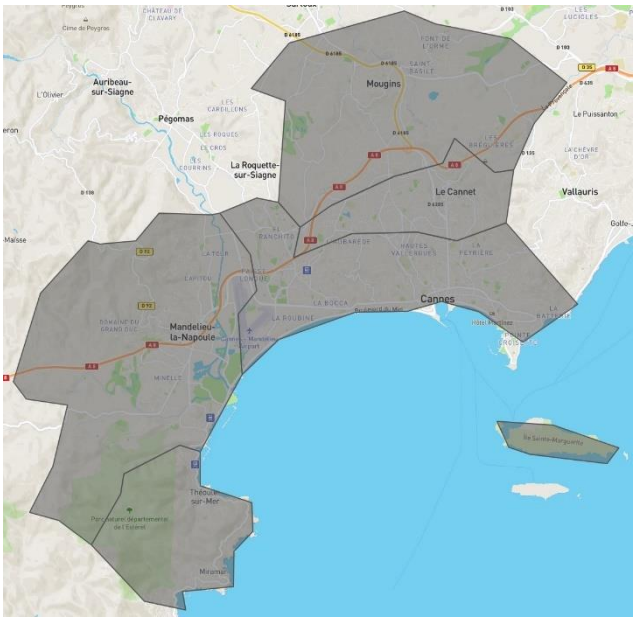
Folium est une bibliothèque sur Python pour de la cartographie interactive. Elle s'appuie sur la bibliothèque JavaScript connue Leaflet mais n'est encore qu'en phase de développement. Certaines difficultés ont été rencontrées :

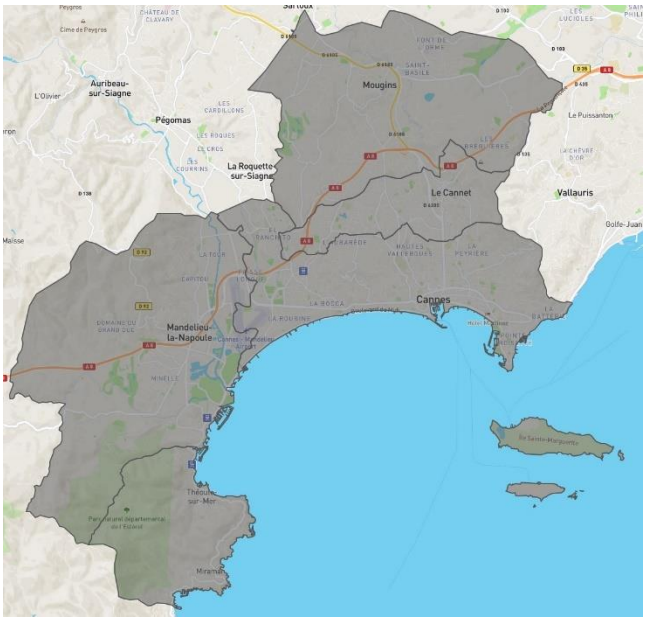
- il a fallu installer manuellement la version de développement, disponible dans la prochaine version à une date inconnue, depuis leur git pour pouvoir disposer de certaines fonctionnalités comme le géocodage
- une erreur d'affichage des caractères spéciaux, spécifique aux notebooks Jupyter (due à une différence entre l'encodage de l'HTML pour le notebook en UTF-8 via branca et le désencodage effectué en UTF-16 du côté du notebook pour son affichage), il a fallu également installer manuellement la version de développement réparée de la bibliothèque Branca
- un travail supplémentaire d'encodage/désencodage a été dans un premier temps requis pour afficher les images sur les pop-ups interactifs de la carte
- il est impossible d'afficher les cartes directement des notebooks en ligne sur GitHub car ce dernier produit un rendu statique et n'embarque pas le code JavaScript associé ; il faut rediriger vers Jupyter NBViewer.

GEOJSON

GeoJSON, pour Geographics JSON, désigne le format d'encodage pour des ensembles de données géospatiales. Il est utilisé entre autres par Leaflet et le projet libre de cartographie OpenStreetMap. Comme son nom l'indique, il utilise la norme JSON.
A titre de rappel, JSON ou JavaScript Object Notation est un format de données textuelles pour faciliter l'échange de contenus complexes en suivant un standard et se démarque de XML par sa facilité d'utilisation, la plupart des langages de programmation disposant de bibliothèques associées, et sa faible verbosité, JSON étant structuré autour d'ensembles de paires clé/valeur comme les dictionnaires en Python. Il présente néanmoins l'inconvénient de procurer des types restreints.

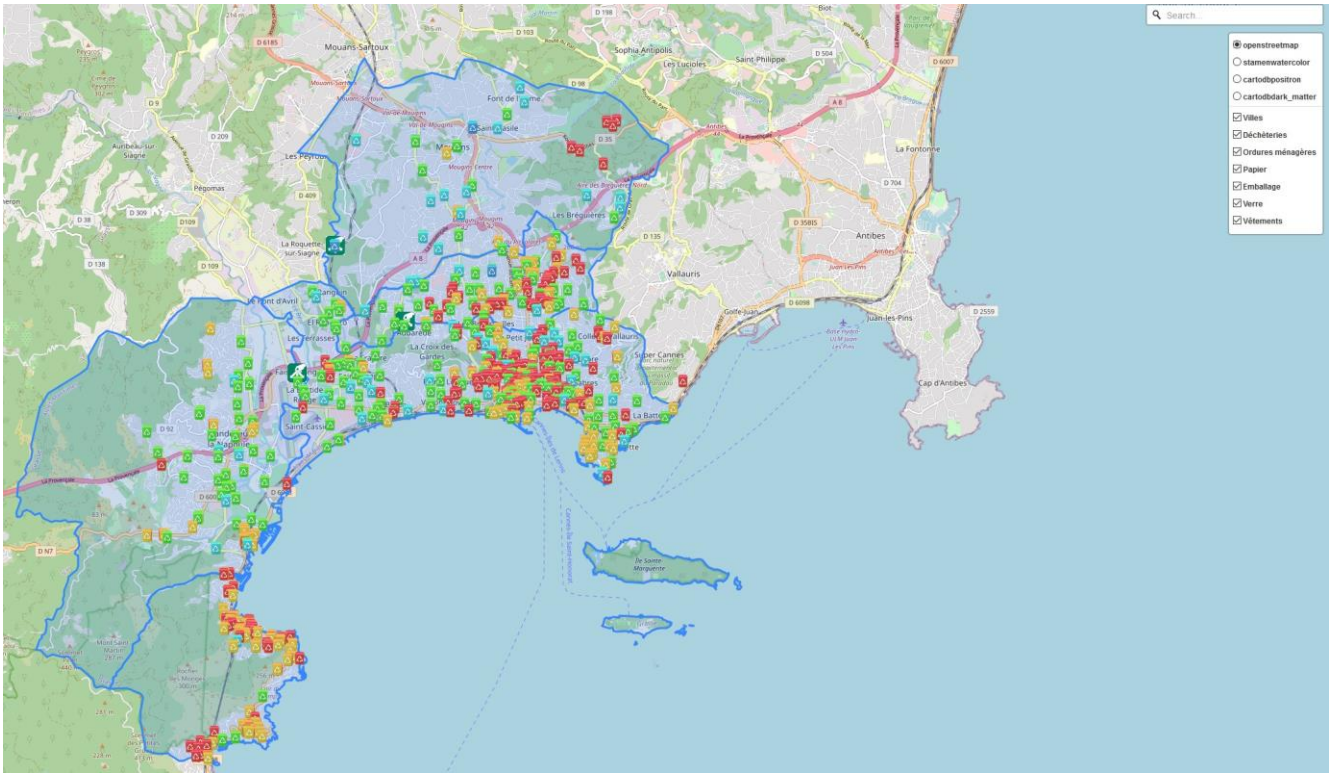
La définition originale des délimitations des villes était très limitée et mal définie en certaines zones, au point d'empêcher d'incorporer certains collecteurs dans les délimitations de la ville dont ils dépendaient, par exemple ceux de la pointe Croisette. Une édition complète à la main, point par point, a été faite en s'appuyant sur plusieurs sources, comme le plan cadastral, pour une résolution bien supérieure, comme l'atteste le comparatif avant/après ci-dessous :





APERÇU ET FONCTIONNALITES DE LA CARTE

La carte interactive se présente simplement

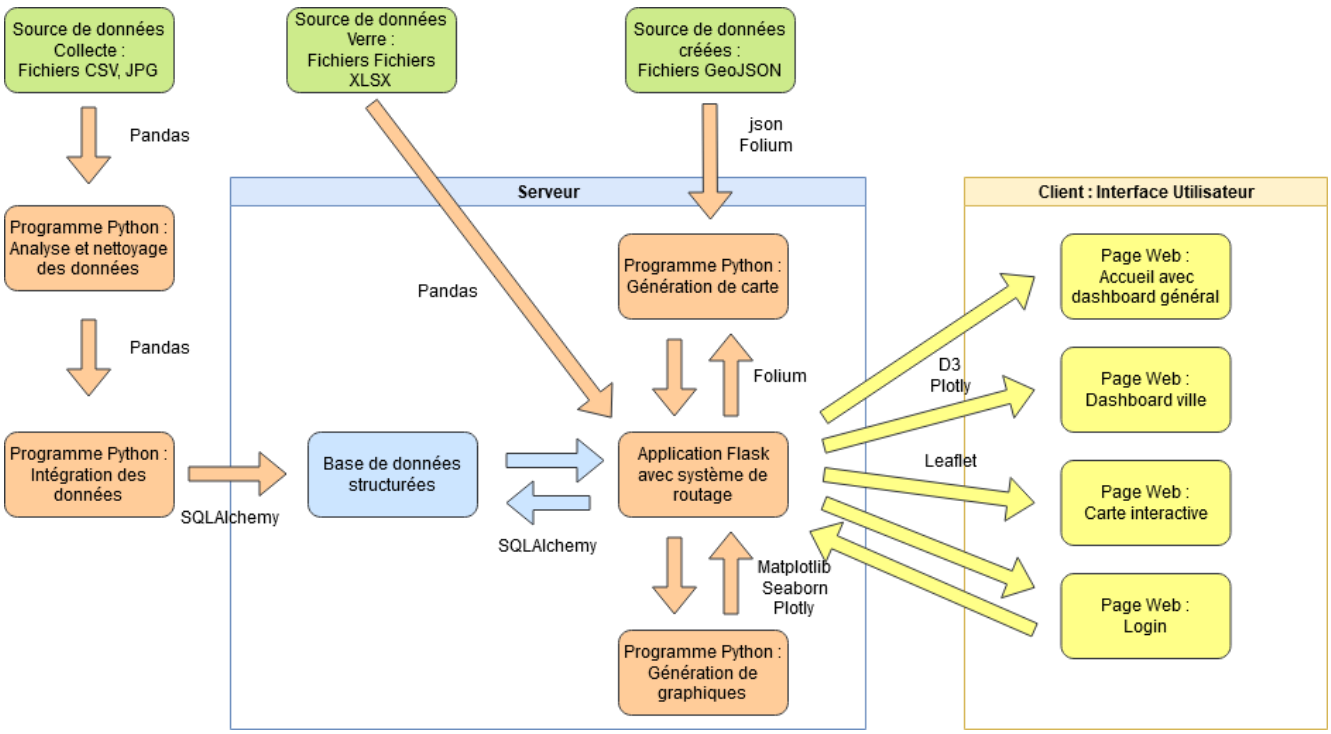


Différents types de cartes sont sélectionnables, de la vue schématique à celle plus artistique avec des effets d'aquarelle. Les collecteurs et déchèteries sont affichés selon leur pictogramme habituel et filtrables. A des fins de supervision rapide, le graphique des collectes de l'année actuelle par secteurs est affiché lors du clic sur la ville souhaitée.
Un moteur de recherche est proposé pour pouvoir afficher les points d'apport volontaire à proximité de l'adresse renseignée.

Tableau de bord

Un tableau de bord simplifié avec connexion est également disponible. Il propose les graphiques les plus pertinents pour l'agglomération dans son ensemble, la carte interactive complète ainsi que quelques graphiques spécifiques à chaque ville.

Les maquettes web ont été facilitées par l'utilisation de l'outil Bootstrap. L'application web a été construite facilement grâce au framework Flask, qui s'avère léger, et suit ce schéma fonctionnel :



Conclusion

EXTENSIONS POSSIBLES

Différents axes de développement sont à envisager pour prolonger ce projet :

- intégration de nouveaux jeu de données pour développer l'analyse
    - données générales de l'INSEE (en particulier celles carroyées) pour établir des dépendances avec les données démographiques (densité de population, familles nombreuses, niveaux de vie…)
    - données SIRENE pour mieux mesurer l'incidence de chaque secteur d'activité
    - données météorologiques pour mieux mesurer leur impact sur l'apport de déchets

Il est à noter que les 2 premiers points de cet axe ont été le centre d'intérêt majeur de l'équipe continuant ce projet

- amélioration de la carte en travaillant directement avec Leaflet.js qui semble plus judicieux en vue de la version provisoire de Folium
    - avec l'ajout de couches, vues et filtres supplémentaires
    - en renforçant son interactivité
- transformation du tableau de bord en service à part entière et création d'une API dédiée
- apport du caractère prédictif de l'analyse
    - définition des principaux facteurs d'utilisation des collecteurs (par une analyse PCA)
    - définition des types d'usager et des profils associés (par une partition K-Means)
    - défintion de groupes de collecteurs selon de nouveaux traits selon l'utilisation, le remplissage, la localisation… (par une partition K-Means)
    - détection et prévision d'utilisation des collecteurs (par des RNN, LSTM…)

RETOUR D'EXPERIENCE

Il a été intéressant d'appliquer les différentes connaissances vues le long de la formation dans une situation professionnelle concrète, telle une mission en tant que prestataire avec des résultats attendus.
Le défi était d'autant plus intéressant qu'on était en contact direct avec les responsables du service et qu'on devait être force de proposition, les éléments que l'on fournissait pouvant potentiellement avoir une incidence sur les décisions de l'agglomération. Il est particulièrement valorisant d'avoir contribué à un service public et entraperçu le fonctionnement et les responsabilités du service environnement dans l'agglomération.
Les circonstances de distanciel imposé par le contexte sanitaire ont originalement mis à l'épreuve les pratiques agiles.

La présentation finale auprès de la CACPL a vraiment permis de restituer le travail du stage comme la présentation d'une solution logicielle à un client et il a été particulièrement plaisant de recevoir des retours aussi positifs de la part du directeur de l'innovation et de la transformation numérique ainsi que de la directrice de la relation usagers et de la qualité avec qui j'ai pu échanger et comprendre certains résultats trouvés.

L'expérience est d'autant plus positive quand l'approche suggérée permet d'inspirer l'agglomération :
https://www.facebook.com/CannesPaysDeLerins/photos/a.330470310483454/1586682054862267/