DOSSIER DE VALIDATION

Ingénieur en science des données
spécialisé en infrastructure data
pour le Titre RNCP 39586

Bloc 2: Analyser, organiser et valoriser des données

Nom Prénom	CORIN Gaëtan
Nom Prénom du tuteur	MOULIN Alexis
Niveau visé	RNCP 7
Date de la soutenance	Septembre 2025
Lieu de la soutenance	Toulouse

Table des matières

Introduction

Je m'appelle Gaëtan Corin, j'ai 30 ans, et je suis en reconversion professionnelle depuis 4 ans. J'ai réalisé un CAP boulanger et un CAP pâtissier où j'ai exercé ces métiers pour une durée totale de 10 ans, puis j'ai décidé de me reconvertir.

Mon parcours de reconversion a commencé au centre de formation de l'Adrar Pôle Numérique où j'ai réalisé une année de formation en présentiel me donnant le titre RNCP 5 Développeur d'application Web avec option Devops.

J'ai ensuite réalisé une année en alternance avec l'école supérieure IPI Blagnac, ainsi que l'entreprise CELAD. Durant cette alternance, j'ai travaillé dans une équipe sur un projet interne en tant que Développeur Fullstack sur les langages Python et Angular, me donnant une première expérience professionnelle sur un projet ambitieux. J'ai eu l'opportunité d'avoir un chef de projet et un Scrum Master qui maîtrisaient parfaitement la méthode Scrum, ce qui a été riche en enseignements.
En plus de cette mission, j'ai aussi eu l'opportunité de partir d'un projet de zéro pour le client Renault, où j'ai réalisé l'intégralité de l'application de moi-même en 3 mois avec le support et les conseils de mon responsable d'alternance qui m'a aidé pour la conception, et qui faisait l'intégralité de mes revues de code.
Ce projet a pu me servir de sujet de mémoire pour passer mon titre RNCP 6 Concepteur développeur d'application numérique, ou j'ai obtenu les félicitations du Jury.

Je continue ainsi mon parcours avec l'école supérieure Ynov, où je me spécialise en Data-Engineer. C'est un métier qui m'intéresse depuis la fin de ma première année d'informatique et où j'ai eu la chance de concrétiser cette ambition pour mon mastère. Mon alternance est réalisée avec l'entreprise Menaps, une startup toulousaine. J'ai pu être encadré durant la première année par un Data-Engineer senior qui m'a appris les fondements de ce métier sur un grand projet orienté Data pour le client Stellantis. Par la suite, le Data-Engineer senior est parti à la fin de ma première année d'alternance, et j'ai pris la relève de son poste, étant le dernier Data-Engineer de l'entreprise. J'ai donc été promu au rang de Data-Engineer référent sur le projet aux yeux du client (ayant un Data-scientist en référent technique si besoin). J'ai réalisé des réunions avec le client une fois par semaine, où je devais lui expliquer l'état d'avancement du projet, prendre en compte les éventuelles remarques et demandes du client, et continuer les améliorations et réalisations en cours.
Le projet étant désormais terminé, je travaille avec mon chef de projet et l'équipe afin de rendre notre projet simple et fonctionnel pour un futur potentiel client.

J'ai énormément appris durant ces 4 années et j'ai comme ambition de continuer dans le métier de Data-Engineer où je me sens prêt à relever les défis qui me seront proposés.

Je tiens a préciser que l'ensemble de ce mémoire a été écrit entièrement à la main, sans aide de rédaction de LLM, et que le projet présenté dans ce mémoire a été entièrement réalisé par mes soins durant mon temps libre, disponible sur mon github:
https://github.com/gaetancorin/Datapipeline_comparaison_official_vs_gas_stations_reporting

C2.1.1 : Analyser les besoins métier et les enjeux exprimés par un commanditaire en réalisant des entretiens exploratoires et en récupérant les informations stratégiques nécessaires afin de cadrer le travail d'analyse des données à produire.

Un entretien fictif avec les personnes du pôle data du gouvernement français a été réalisé afin qu'ils partagent les différentes problématiques et enjeux métiers nécessitant notre intervention.

Voici les informations stratégiques qui ont été relevées lors de cet entretien afin de cadrer au mieux le travail d'analyse à produire:

Problématique:
Le pôle data du gouvernement français possède un jeu de données officiel sur l'évolution du prix des différentes essences vendues en France. Ces données sont rassemblées en moyenne hebdomadaire.

Malheureusement, ils ne connaissent pas la méthodologie qui a permis de calculer ces moyennes hebdomadaires à l'époque où celles-ci ont été calculées.

Ils se posent donc la question de la véracité et de la fiabilité des transformations des données anciennement recueillies, et souhaiteraient qu'un recalcul soit fait sur d'autres données journalières existantes afin de vérifier la crédibilité des transformations passées. Ils aimeraient aussi une vision plus fine des cours des différentes essences vendues en France, afin d'en tirer des conclusions et d'avoir une meilleure visibilité en termes journalièrs. Cet aperçu devra être mis à jour régulièrement.

Cadre réglementaire:
La "doctrine cloud au centre" de l'État incite les applications gouvernementales à être déployées sur des environnements cloud de confiance.
Il sera donc nécessaire de construire l'application de manière à pouvoir être entièrement déployé sur des clouds réputés.

Environnement:
L'environnement est laissé libre durant la réalisation du projet. Le projet devra ensuite être entièrement dockerisé afin d'être déployé sur les services cloud.
Les données pourront aussi être stockées dans des clouds, car il ne s'agit pas de données sensibles. En effet, nous ne travaillons qu'avec des données OpenData.

Contraintes:
Il est obligatoire de ne croiser que les données gouvernementales issues des différents sites internet de l'Etat, afin d'assurer une viabilité des données reconnue par celui-ci.

RSE:
Une attention particulière devra être faite sur le chargement de données.
En effet, il faut éviter le surchargement inutile sur l'ensemble de l'historique des données chaque jour, dans une optique de préservation écologique et économique.
Les problèmes de confidentialité des données suivant la loi RGPD sont faiblement impactant sur ce projet, car l'intégralité des données est déjà disponible en Open Data et anonymisés. Les données sources sont donc déjà considérées comme respectant la loi RGPD.

C2.1.2 : Définir les axes d'analyse et les métriques en identifiant les données à exploiter, celles disponibles et pertinentes pour traduire la problématique d'entreprise énoncée en problème numérique.

Au vu des informations relevées durant la prise des besoins et des attentes, **les objectifs permettant de répondre aux problématiques client** sont les suivants :

- Réaliser un audit sur la qualité des données officielles hebdomadaires en comparaison avec un autre jeu de données journalier gouvernemental.
- Rechercher des décalages de prix sur les données historiques qui suggéreraient un changement de méthode de calcul des moyennes hebdomadaires officielles.
- Analyser sur une granulométrie temporelle plus fine qu'actuellement les prix des essences afin d'en tirer des conclusions.
- Faire un rendu sous format de visualisations claires qui pourra être automatiquement rafraîchi lors de nouvelles données.
- Les visualisations doivent pouvoir être sauvegardées et restaurées dans un environnement externe, afin d'assurer la facilité de déploiement du système dans d'autres environnements.

Afin de réaliser des axes d'analyses et des métriques permettant de répondre à ces problématiques, deux jeux de données sont à exploiter

Le premier jeu de données va représenter les **données officielles des prix des essences** vendues en france en hebdomadaires. Il s'agit du jeu de données donné par le pôle data du gouvernement.
https://www.ecologie.gouv.fr/politiques-publiques/prix-produits-petroliers

Le second jeu de données représente les **relevés de prix des stations essences** de manière journalière. Ce jeu de données vient aussi d'une source gouvernementale, et respecte donc les contraintes imposées demandant de ne croiser que des données issues du gouvernement afin d'en assurer de leur viabilité.
https://www.prix-carburants.gouv.fr/rubrique/opendata/

Les deux jeux de données sont accessibles en OpenData. Il est nécessaire que ces deux jeux de données possèdent des prix de même type d'essence, ainsi que sur de longues périodes identiques afin de pouvoir réaliser des comparaisons de données pertinentes sur le long terme.

Avec les jeux de données nécessaires, les problématiques du client peuvent se reformuler en **problématiques numériques** permettant d'avoir une vision claire de comment traiter les données pour répondre aux différents besoins.

Les deux premiers objectifs qui seront traités seront:

- **l'audit entre les deux jeux de données**
- **la recherche du décalage de prix sur des temps précis qui suggéreraient un changement de méthode de calcul des moyennes hebdomadaires officielles**

Ces objectifs semblent correspondre à un même besoin global qui pourrait être résolu par un plan d'analyse similaire.
Pour répondre à ces deux problématiques, les données nécessaires seront les deux jeux de données. Il faudra donc les données officielles ainsi que les données issues des stations essence.
Les métriques nécessaires sur ces jeux de données sont pour chacun d'eux le prix des essences, les types d'essences concernés ainsi que la date. Cela permettra de réaliser une visualisation sur l'évolution des prix des essences en courbe temporelle qui pourra facilement être analysée, et permettra d'en faire des hypothèses permettant de continuer l'analyse.

Le troisième objectif qui devra être traité sera:

- **l'analyse sur une granulométrie temporelle plus fine qu'actuellement sur les prix des essences**

Pour répondre à cette problématique, le jeu de données des relevés de prix de stations essence sera suffisant.
Les métriques nécessaires sur ce jeu de données seront le prix des essences, les types d'essences concernés ainsi que la date. Cela permettra là aussi de réaliser une visualisation sur l'évolution des prix des essences en courbe temporelle journalière qui pourra facilement être analysée, et permettra d'en faire des hypothèses permettant de continuer l'analyse et en tirer des conclusions.

Le quatrième objectif est:

- **Le rendu sous format de visualisations claires qui pourra être automatiquement rafraîchi lors de nouvelles données.**

Cet objectif ne demande pas de métriques spécifiques. En revanche, il indique clairement que la solution à déployer devra permettre de se connecter aux données de manière autonome et récurrente pour récupérer les dernières informations, ainsi que d'avoir la capacité de les afficher de manière automatique.
L'outil de visualisation devra donc être adapté à cette tâche.
Afin de répondre à cette demande, le choix a été fait sur l'outil Métabase pour la visualisation car il permet de répondre aux objectifs demandés de mise à jour de données automatique sur les visualisations.

Le cinquième objectif est:

- **Les visualisations doivent pouvoir être sauvegardées et restaurées dans un environnement externe, afin d'assurer la facilité de déploiement du système dans d'autres environnements.**

Là non plus, cet objectif ne demande pas de métriques spécifiques.
Il faudra en revanche réaliser un système permettant de sauvegarder les visualisations dans un cloud, afin de pouvoir les déployer sur un autre environnement.
Pour réaliser cela, un serveur Flask sera connecté à un espace de stockage S3. À l'appel d'une API, le serveur Flask ira chercher les informations des visualisations dans Métabase sur le dossier natif “metabase.db ”. Ce dossier sera copié, zippé, daté, puis stocké sur le serveur S3. A l'appel d'une autre API, le dossier stocké pourra être restitué sur une autre machine contenant un autre Métabase afin d'accéder à ces mêmes visualisations sur un environnement différent.

Enfin, afin de joindre le travail réalisé dans le Bloc 1 de ce titre RNCP, le travail de traitement des données est réalisé par des ETLs en serveur Flask, avec des planificateurs de tâches permettant de récupérer les données les plus récentes chaque jour.
Ces données sont extraites, transformées, puis stockées sur une base de données MongoDB.
Elles sont donc directement disponibles pour répondre aux métriques nécessaires à la visualisation.

C2.1.3 : Réaliser des requêtes et des calculs en utilisant des outils de dashboarding, des tableurs, des requêtes SQL ou scripts Python afin de produire une analyse des données préalablement collectées.

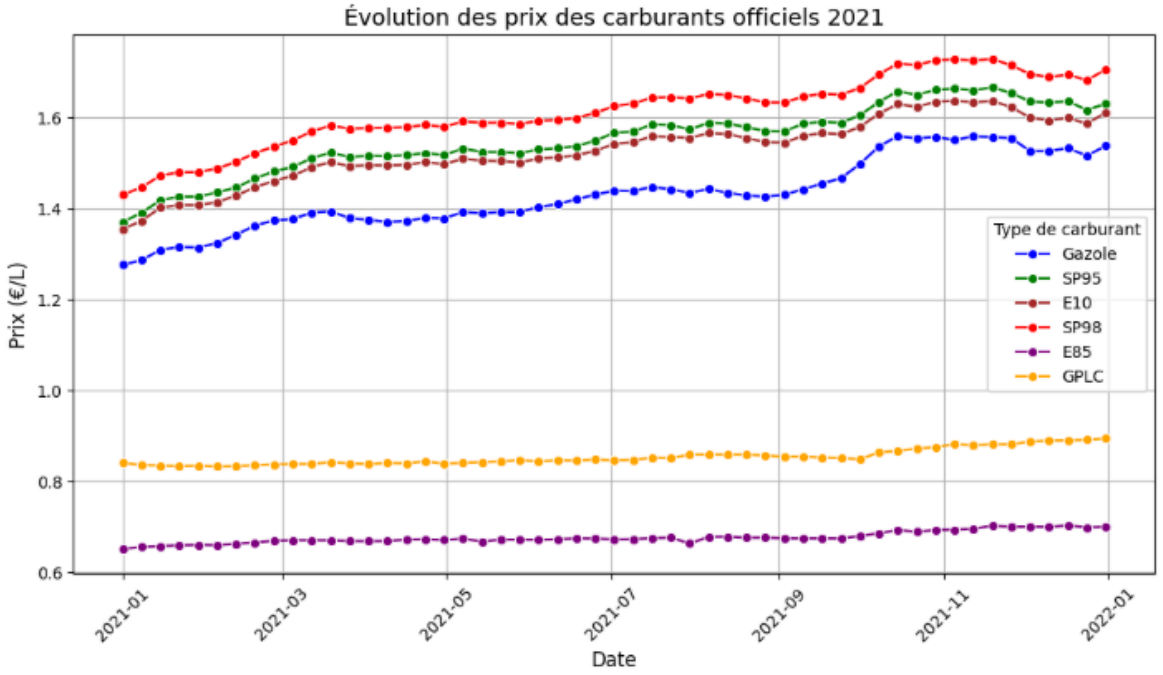
Comme dit précédemment, les deux premières problématiques:

- **l'audit entre les deux jeux de données**
- **la recherche du décalage de prix sur des temps précis qui suggérerait un changement de méthode de calcul des moyennes hebdomadaires officielles**

nécessitent de créer des visualisations en graphiques temporelles pour le jeu de données des prix des essences officielles, ainsi que pour le jeu de données des relevés de prix des stations essence.
Ayant 6 types de données à analyser par jeu de données (Gazole, SP95, SP95-E10, SP98, E85, GPLc), le choix de commencer par une analyse de chaque jeu de données de manière séparée semble plus pertinent pour éviter de surcharger les graphiques.

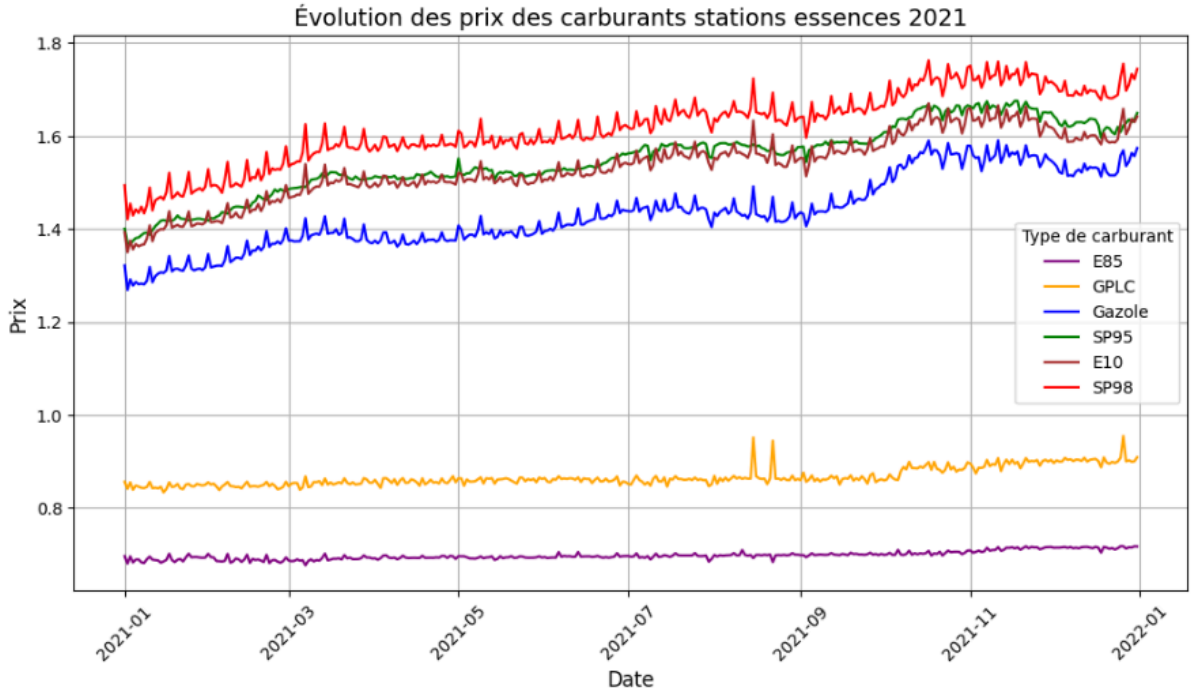
Pour réaliser cela, une première analyse en jupyter notebook est réalisée en utilisant la librairie de visualisation matplotlib.
Cela permet d'avoir une première vision globale des données séparées en offrant rapidement des résultats pertinents.

La première analyse se fait sur les données officielles en visualisation année après année :



D’après les premières analyses, il ne semble pas y avoir de trou de données dans les courbes. Les données sont donc complètes. Les données sont présentes en cycle hebdomadaire, et ce jusqu'à la semaine en cours. En revanche, les données commencent avec un historique à date différente selon les types d'essence.
Par exemple, le Gazole commence en 1985, le SP95 et SP98 en 1990, etc..

La seconde analyse se fait sur les données des stations essence année après année :



On peut se rendre compte, là aussi, que les données semblent complètes, car il n’y a pas de trou dans les courbes. Cette fois-ci, les données sont présentes en cycle journalier, et ce

jusqu’au jour de la veille. De la même manière, les données commencent avec un historique à date de début différente selon les types d'essence.
Par exemple, le Gazole, SP95, GPLc et E85 commencent en 2007, le SP95-E10 en 2009, etc..

Nous pouvons donc en conclure que les types d’essences sont les mêmes entre les deux jeux de données. De plus, de très nombreuses dates sont similaires entre les deux jeux de données, allant jusqu'à un range de similarité de 18 ans pour le Gazole ou le SP95 par exemple.
Nous pouvons aussi voir que sur les données journalières des stations essence, les informations sont bien plus pertinentes que sur les données officielles hebdomadaires, car de nombreuses variations non visibles sur les données hebdomadaires se révèlent sur les données des stations essence en journalier.

Cette seconde visualisation des données des stations essence nous permet donc d’y voir un peu plus clair sur l’objectif suivant: “**l’analyse sur une granulométrie temporelle plus fine qu'actuellement sur les prix des essences**”, ainsi que sur les possibilités qu'offre ce jeu de données

Afin d’approfondir ces analyses, et de pouvoir plus facilement jouer avec les jeux de données, une implémentation de ces données sur le logiciel Métabase va permettre de pouvoir plus facilement faire des comparaisons, ainsi que d’avoir une navigation simplifiée et plus détaillée sur les courbes.
Pour réaliser cela, il faut connecter MongoDB à Métabase grâce à son connecteur intégré.
Le travail de dénormalisation ayant déjà été réalisé lors du processus des ETLs sur le Bloc 1 de ce titre RNCP, l’ensemble des données officielles et des données des stations essence ont été rassemblés sur la même collection sur la même date journalière.
Étant donné que les données officielles sont uniquement hebdomadaires, elles seront donc en valeur “None” les 6 jours restants de la semaine.

Il suffit donc simplement d’importer la collection qui nous intéresse sur Métabase.
Voici l’architecture de cette collection appelée “denorm_station_vs_official_prices”:
{
_id: 685eaa1a984bc1281b199f01
Date: 2021-08-27T00:00:00.000+00:00
official_ttc_GAZOLE_eur_liter: 1.4252
official_ttc_SP95_eur_liter: 1.5692
official_ttc_E10_eur_liter: 1.5449
official_ttc_SP98_eur_liter: 1.6321
official_ttc_E85_eur_liter: 0.6765
official_ttc_GPLC_eur_liter: 0.8566
station_ttc_GAZOLE_eur_liter: 1.4192
station_ttc_SP95_eur_liter: 1.56399
station_ttc_E10_eur_liter: 1.5339
station_ttc_SP98_eur_liter: 1.62267
station_ttc_E85_eur_liter: 0.69809
station_ttc_GPLC_eur_liter: 0.86019

}

Une fois la table importée sur Métabase, on va créer un dashboard appelé “dataviz”. L’objectif de ce dashboard va être de réaliser tous types d’analyses et de juger de leur pertinence.
Dans ce dashboard, on va créer des graphiques temporels pour chaque type d’essence pour lequel on souhaite comparer le prix des données officielles avec le prix des données des stations essence.

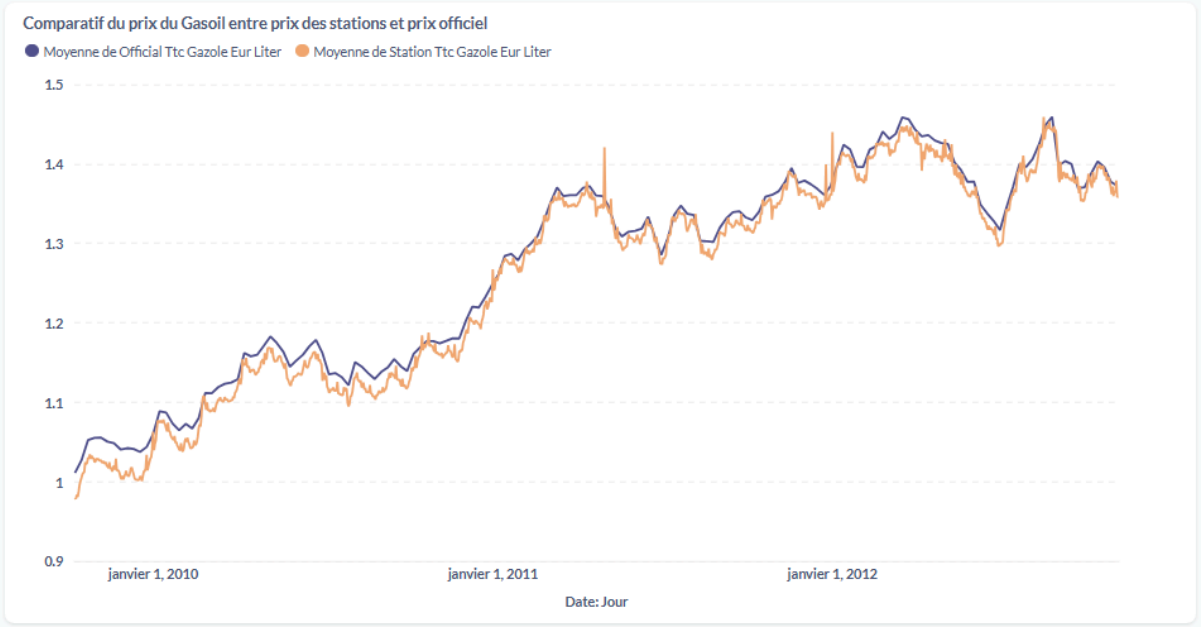
Par exemple, pour le gazole, on va prendre les colonnes “Official Ttc Gazole Eur Liter” ainsi que “Station Ttc Gazole Eur Liter”, et séparer ces valeurs par jour. (La moyenne étant obligatoire sur les colonnes, cela n’impacte pas notre jeu de données car il n’y a déjà au maximum qu’une valeur par jour sur chaque colonne)

Voici le résultat de ce graphique temporel:

On peut désormais bien voir la courbe des données officielles (en bleu foncé) qui commence en 1985, ainsi que la courbe des données des stations essence (en orange clair), qui commence en 2007.

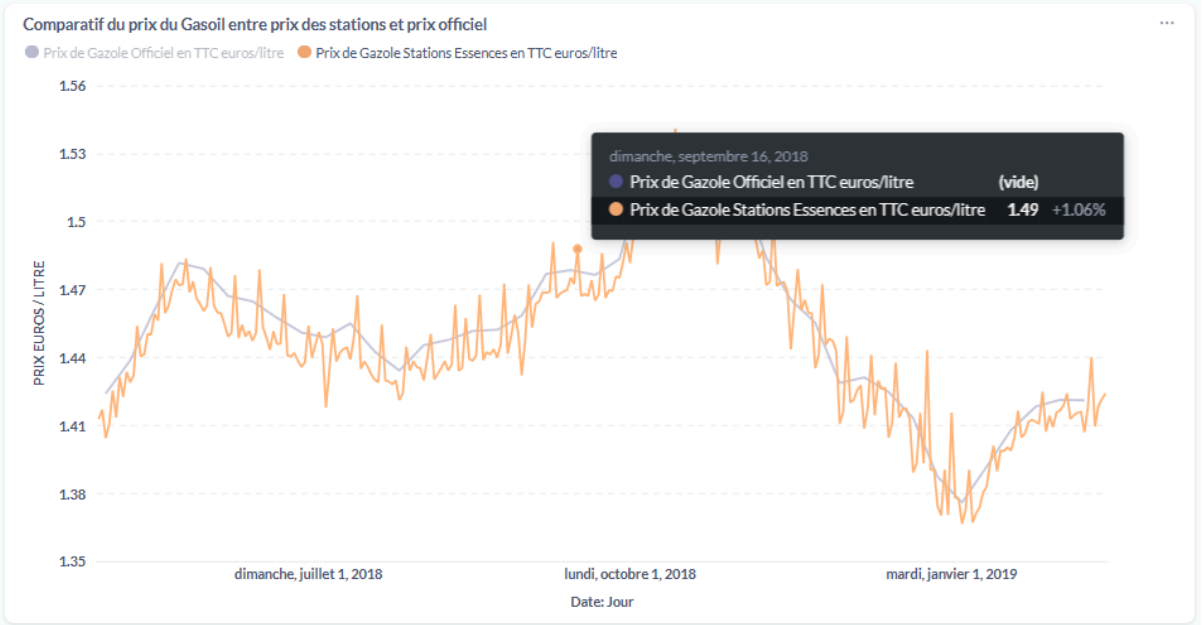
Dans notre cas d'usage, le plus gros avantage de Métabase en comparaison à Matplotlib est que l'on peut zoomer dans les données.

En enquêtant sur les valeurs de manière plus minutieuse, on peut se rendre compte que les données officielles sont légèrement supérieures aux données des stations essence sur certaines années, mais pas sur d'autres.



C'est donc un point important à relever pour la problématique de "**l'audit entre les deux jeux de données**" ainsi que pour la problématique de "**la recherche du décalage de prix sur des temps précis qui suggérerait un changement de méthode de calcul des moyennes hebdomadaires officielles**".

La deuxième remarque qui peut être faite est qu'en zoomant encore plus sur les données, on voit un schéma de prix se dessiner constitué de pics et de creux à intervalles réguliers. En effet, chaque dimanche semble être un pic sur la courbe.



Cela semble donc pertinent à analyser afin de répondre à la problématique de “**l’analyse sur une granulométrie temporelle plus fine qu'actuellement sur les prix des essences**”.
En effet, ce schéma est visible sur les données des stations essence, mais ne peut pas être visible sur les données officielles.

Des analyses similaires à celles présentées ci-dessus sont ainsi effectuées sur chacun des six types d’essences différents, et cela dans le but de chercher d’autres points potentiels à analyser plus en profondeur.

C2.1.4 : Élaborer des modèles statistiques et des tests d’hypothèses en modélisant des relations entre les variables, en évaluant la pertinence des résultats des simulations afin de valider ou réfuter des hypothèses.

D’après les premières analyses qui sont réalisées sur les données, on peut constater que les prix des données officielles hebdomadaires semblent légèrement supérieurs sur certains types d’essences en comparaison des prix des stations essence journalières, et cela sur certaines années uniquement.

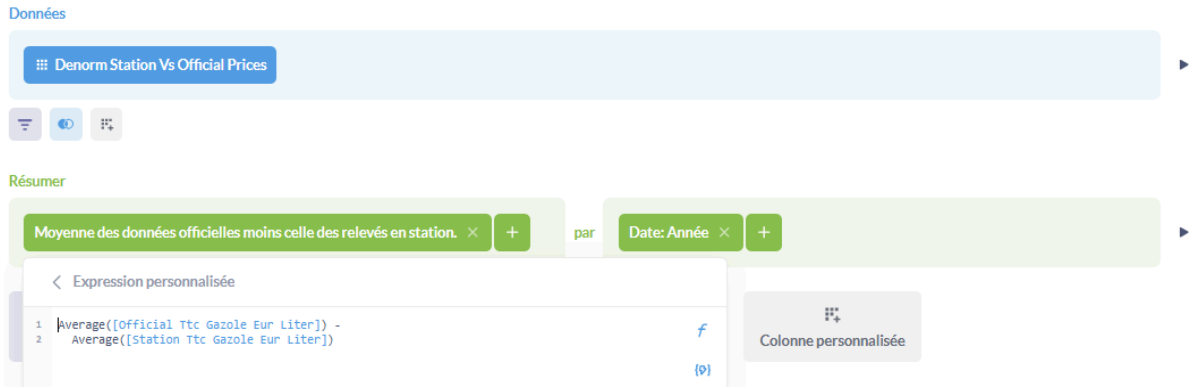
On peut donc faire l’**hypothèse** que cela n’est pas dû au hasard, et que probablement un schéma peut être trouvé parmi ces écarts de prix.

Les **tests associés** pour mieux comprendre ces différences de prix vont consister à réaliser des moyennes par année de ces écarts de prix, afin d’avoir une vision claire et globale de leur évolution, et de pouvoir répondre au mieux aux deux problématiques clients:

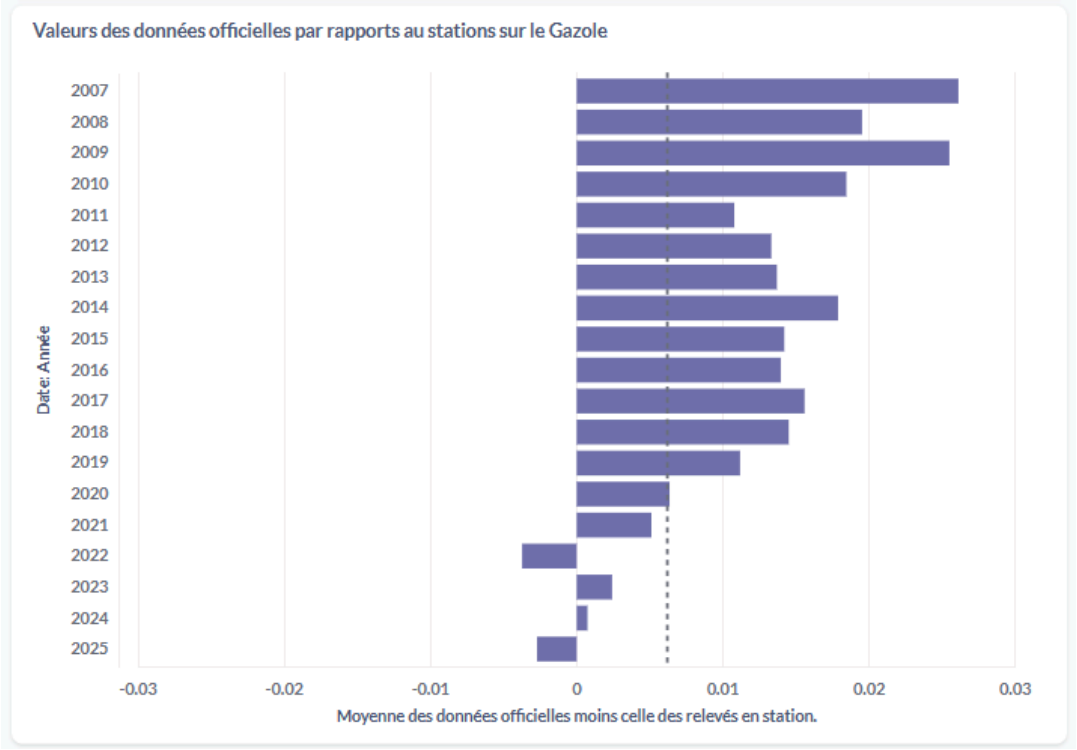
- “l’audit entre les deux jeux de données”
- “la recherche du décalage de prix sur des temps précis qui suggérerait un changement de méthode de calcul des moyennes hebdomadaires officielles”

Pour pouvoir réaliser ce nouveau test, on va devoir créer une nouvelle visualisation sur Métabase.
Afin de comparer les deux jeux de données sur le gazole par exemple, la colonne “Official Ttc Gazole Eur Liter” ainsi que la colonne “Station Ttc Gazole Eur Liter” doivent être comparées. Il va donc falloir créer une “expression personnalisée” permettant de réaliser le calcul suivant pour chaque date:
(moyenne de prix de gazole en données officielles) - (moyenne de prix de gazole en données de station)

Puis de mettre ce résultat par année.



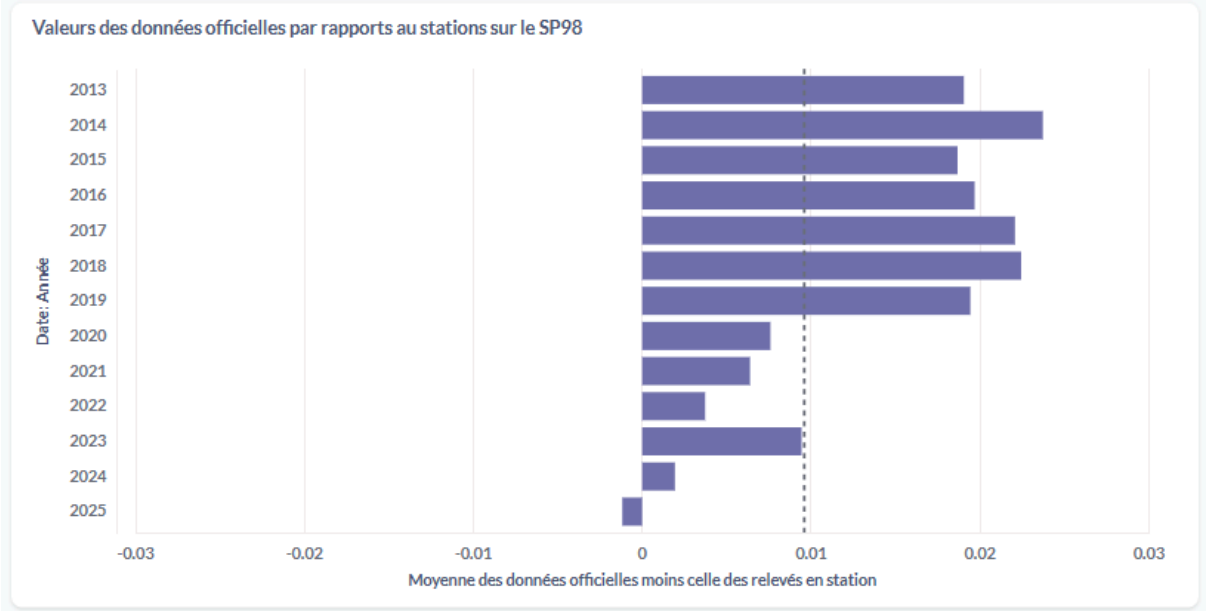
Le résultat permet de voir pour chaque année à combien de centimes supplémentaires ou inférieurs les prix des données officielles sont en comparaison des prix des données des stations essence.



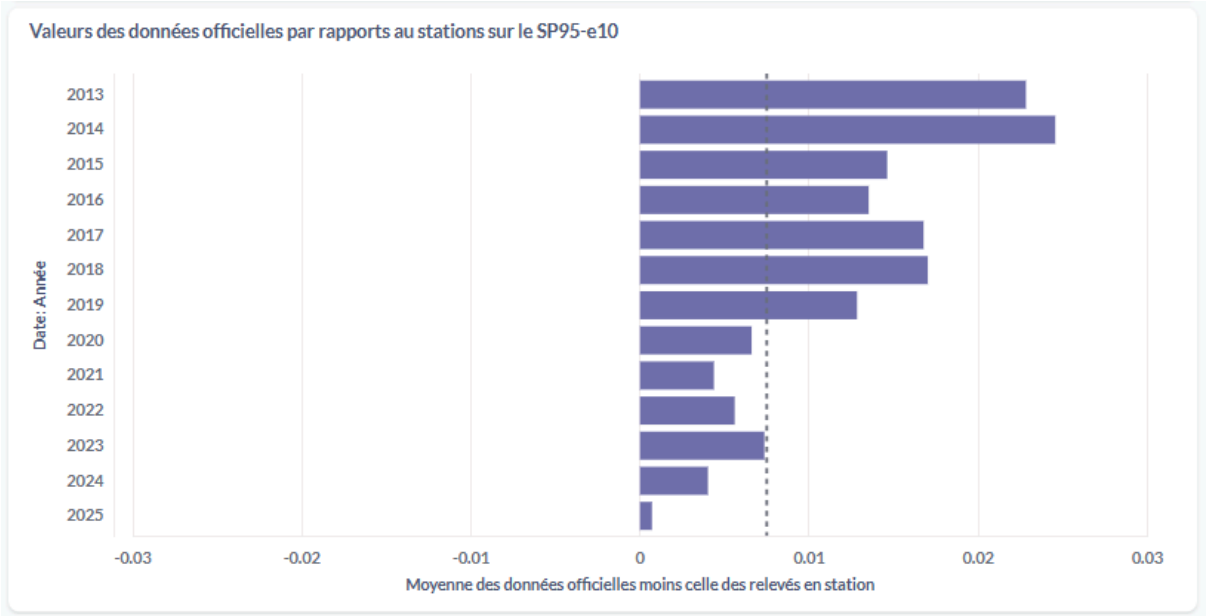
Sur les différences de prix du gazole, nous pouvons en conclure que pour les années de 2007 à 2019, les prix entre les deux jeux de données sont d'un écart d'environ 1.75 centime. Nous pouvons donc considérer que l'écart est vraiment important, et que probablement des erreurs de calcul ont été réalisées lors de l'agrégation des données en hebdomadaire sur le jeu de données officiel durant ces années plus anciennes.

En revanche, pour les années de 2020 à 2025, l'écart moyen est inférieur à 0.6 centime. On peut donc considérer que les résultats sont quasiment similaires pour cette seconde période, et que l'écart peut être expliqué par des erreurs d'arrondi par exemple.

En continuant ces comparaisons, nous pouvons trouver des écarts de prix similaires sur l' essence SP98:



ainsi que sur l'essence SP95-E10:



Les faits étonnants sont que pour chacun de ces types d'essence, l'écart est important du début de l'historique jusqu'à l'année 2019. Par la suite, l'écart de prix diminue fortement de l'année 2020 jusqu'à aujourd'hui.

En prenant en compte les deux problématiques client:

- "l'audit entre les deux jeux de données"
- "la recherche du décalage de prix sur des temps précis qui suggérerait un changement de méthode de calcul des moyennes hebdomadaires officielles"

L'interprétation des résultats permet donc de définir que les données sur les essences Gazole, SP98 et SP95-E10 ont potentiellement subi une modification des méthodes de

calcul en l'année 2020 qui pourrait expliquer des écarts de prix passés qui ne sont quasiment plus existants aujourd'hui. Ces interprétations restent encore à être confirmées par le métier et les professionnels du milieu.
Les résultats réalisés sur les autres types d'essences nous permettent aussi d'interpréter que les essences SP95, GPLc et E85 n'ont pas d'écart de prix significatif permettant d'imaginer un changement de méthode de calcul.

Lors des analyses préalables servant à rechercher des tests d'hypothèses, il a aussi été remarqué que sur les données journalières relevées par les stations essence, nous pouvons voir des pics et de creux à intervalle régulier, ainsi que des pics significatifs le dimanche.

La seconde **hypothèse** que l'on va vérifier est qu'un pattern de prix est présent de manière cyclique suivant les jours de la semaine.

Les **tests associés** pour pouvoir constater ce pattern consistent à réaliser des agrégations des prix par jour de la semaine, et cela pour chaque type d'essence du jeu de données des relevés de stations essence.
Cela permettra de répondre à la problématique:

- l'analyse sur une granulométrie temporelle plus fine qu'actuellement sur les prix des essences

Pour pouvoir réaliser ce nouveau test, on va devoir là aussi créer une nouvelle visualisation sur Métabase pour chaque type d'essence.
Pour rechercher un comportement cyclique de prix sur le gazole par exemple, on va réaliser une moyenne des prix des relevés de stations essence par jour de la semaine sur le gazole.
(Les données de la colonne "Date" étant formatées pour être reconnues par Métabase, l'outil permet directement de filtrer par jour de la semaine sans traitement supplémentaire à réaliser)

Voici le résultat des visualisations pour l'ensemble des essences

Nous pouvons constater que pour les essences Gazole, SP95-E10, SP98, et GPLc, la différence de prix est très significative le dimanche, avec une nette augmentation du prix pouvant aller jusqu'à plusieurs centimes.
Cette constatation est aussi visible sur l'essence E85, mais de manière un peu moins importante.
En revanche, pour l'essence SP95, le dimanche ne semble pas avoir d'augmentation de prix significative.

En prenant en compte la problématique client:

- l'analyse sur une granulométrie temporelle plus fine qu'actuellement sur les prix des essences

L'interprétation des résultats permet donc de définir que les essences Gazole, SP95-E10, SP98, et GPLc possèdent une augmentation très significative de leurs prix le dimanche.
Cette hausse de prix est aussi présente sur l'essence E85 mais de manière moins significative, et elle est quasiment inexistante sur l'essence SP95.
Enfin, les autres jours de la semaine ne semblent pas montrer de schéma particulier pour l'ensemble des essences.

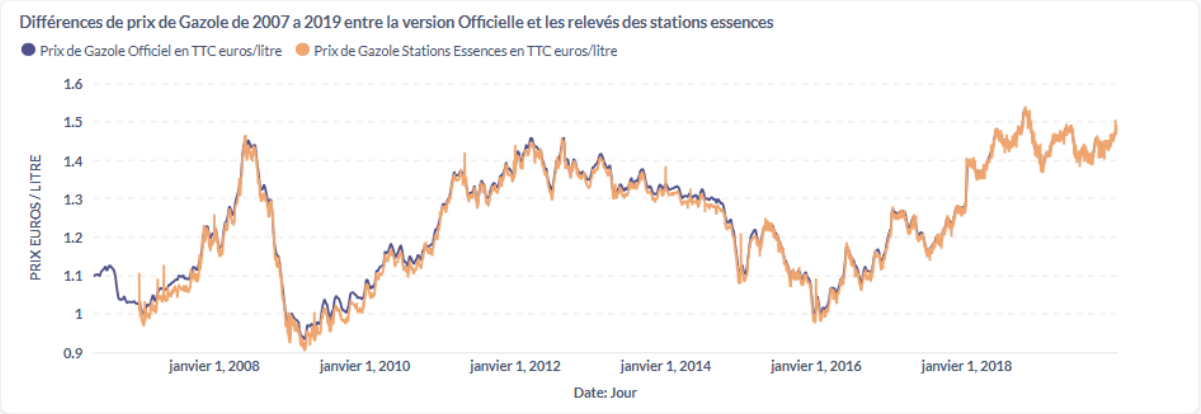
C2.2.1 : Représenter les données en choisissant les modèles de représentation les plus adaptés (ex : histogramme, Heat map, nuage de points) et en utilisant des outils de représentation adaptés (ex : Office, power BI) afin de permettre la compréhension et l'exploitation des données par le public visé.

Afin que les résultats des analyses réalisées soient impactants et compris par le client, il est important de définir comment les présenter pour que les informations soient correctement transmises.

La **première page** s'appelle "Différences de résultat sur le Gazole".
Elle a comme objectif de présenter au client le décalage de prix significatif du Gazole sur des dates précises entre les données officielles données par le client, et les données relevées par les stations essence.

Sur cette page, il y a trois représentations de données.

La première ainsi que la deuxième représentation de données sont des graphiques temporels similaires représentant simplement des dates différentes.
Le premier représente les dates 2007 à 2019 ou l'on voit les différences de prix entre les jeux de données, tandis que le second graphique représente les dates 2020 à 2025 où il n'y en a pas.



Sur ce graphique de 2007 à 2019, on souhaite montrer la différence des valeurs de prix sur une temporalité.
L'objectif ici étant de créer une visualisation de la comparaison des deux jeux de données sur leurs prix au sein de leurs évolutions temporelles.
Il est donc nécessaire de présenter les données en graphique temporel ayant un axe horizontal en temps et un axe vertical en prix.
Le découpage de l'axe horizontal uniquement de 2007 à 2019 représente les dates exactes où cette différence de prix est présente. Elle est un peu moins visible sur les dernières années de 2016 à 2019 à cause de la volatilité qui augmente et cela cache malheureusement un peu la courbe bleue foncé.

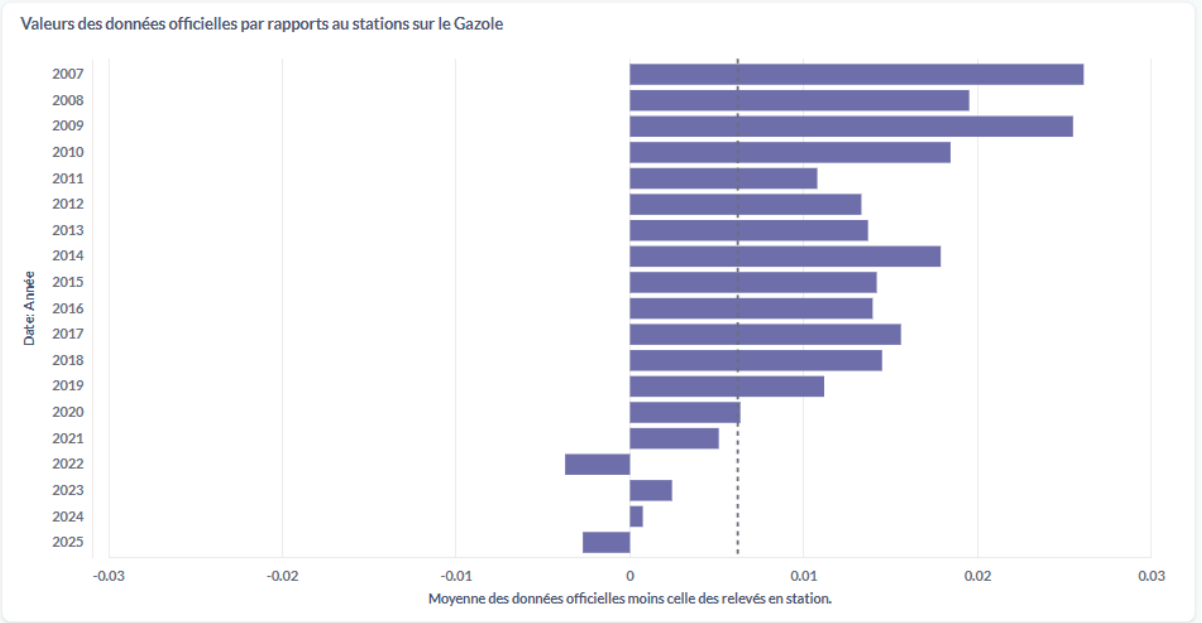
L'axe vertical, représentant les prix, ne commence pas au 0 absolu. En effet, les informations pertinentes ici ne sont pas les valeurs du prix de l'essence pour chacun des jeux de données, mais bien la différence de valeurs entre ces données. Le fait d'améliorer

l’amplitude de la courbe en ne commençant pas à 0 permet d’améliorer cette visibilité de comparaison, qui est le but recherché dans ce graphique.

Les légendes des deux courbes “Prix de Gazole Officiel en TTC euros/litre” et “Prix de Gazole Stations Essences en TTC euros/litre” permettent de fournir toutes les informations essentielles des données que l’on compare. Cela indique le type d’essence, s'il est hors taxe ou avec taxe, en quelle valeur monétaire et pour quelle quantité d’essence.

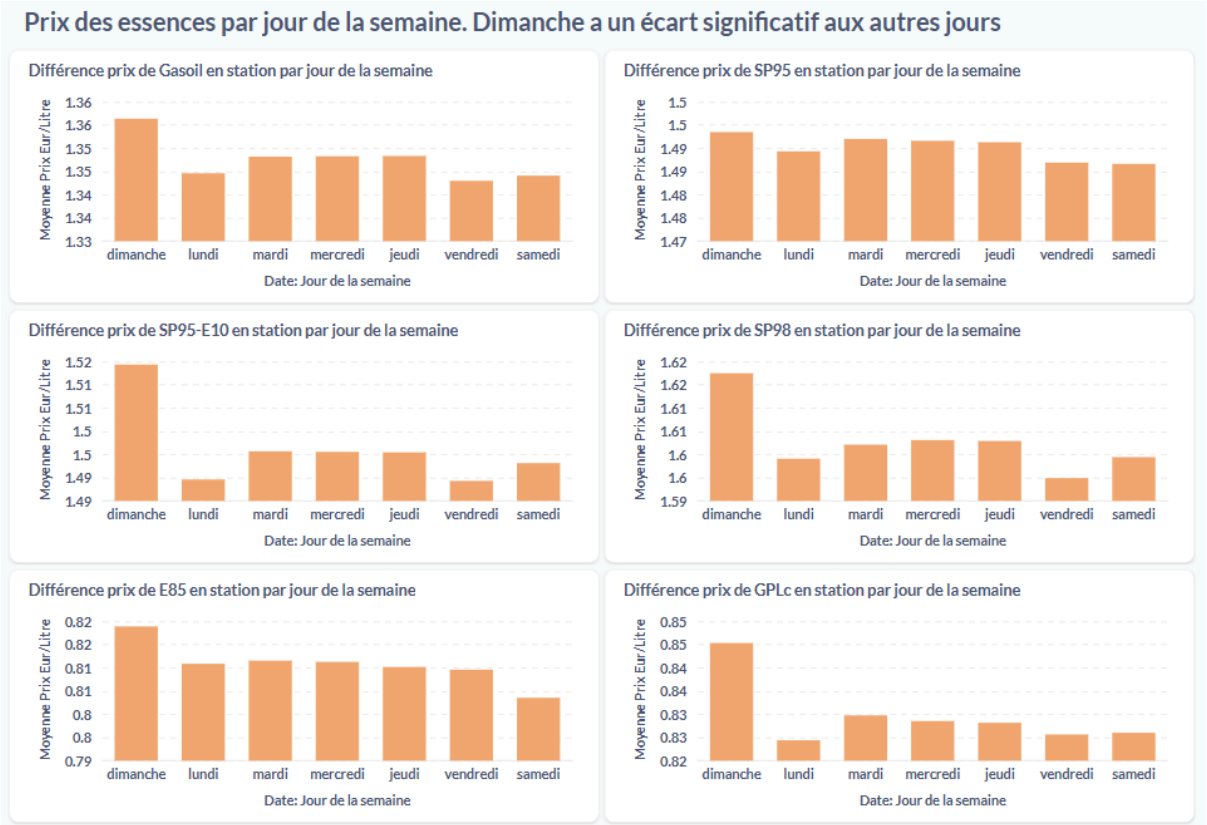
Enfin, les couleurs orange clair et bleue foncé ont été choisies car elles sont facilement distinguables pour les personnes atteintes de daltonisme. Cela permet d’améliorer l’inclusion des personnes souffrant de ce type de handicap, en leur offrant la possibilité d’accéder aux informations.

Le troisième graphique représente les différences de prix sur les données officielles par rapport aux données relevées par les stations essence, mais cette fois-ci présentées en moyenne annuelle.



Cette fois-ci, la temporalité est sur l’axe vertical et les différences en centimes sont sur l’axe horizontal.
Un trait en pointillé a été placé sur l’axe vertical. Il permet de délimiter l’année 2020 où un changement semble avoir eu lieu (en lien avec les deux autres graphiques temporels).
Cela permet de clairement distinguer les années à fortes différences de prix ainsi que les années à faibles différences.
La couleur violet foncé est choisie pour être en accord avec les données officielles sur les deux premiers graphiques, présentées aussi en violet foncé.
Le diagramme en barres horizontal est choisi sur ce graphique car il est bien adapté pour faciliter la comparaison visuelle sur notre cas d’usage de différence de prix, et il permet de facilement distinguer ces deux périodes temporelles.
De plus, il met bien en valeur la valeur 0 qui est le résultat le plus important puisque celui espéré lors d’une comparaison d’une même essence à une même date.

La deuxième page s’appelle “Prix par jour de la semaine”
Elle a comme objectif de présenter au client les réactions des prix de chaque type d’essence des relevés des stations essence, suivant les jours de la semaine.



Un graphique est réalisé pour chaque type d’essence sur une même page afin de pouvoir les comparer entre eux.

Pour chaque graphique, l’axe horizontal représente les jours de la semaine, et l’axe vertical représente les prix des essences.
Le diagramme en barres vertical est choisi sur ces graphiques car il est bien adapté pour présenter des données agrégées par type (comme ici pour les jours de la semaine). Cela permet de faciliter la comparaison et de voir immédiatement les données différentes, comme on peut le voir sur les dimanches.

L’axe horizontal représente les jours de la semaine présentés de manière chronologique, afin de faciliter la lecture et la logique temporelle.

L’axe vertical ne commence pas sur le zéro absolu, car là aussi, les informations pertinentes ici ne sont pas les valeurs du prix de l'essence pour chacun des jours de la semaine, mais bien la différence de valeurs entre ces jours. Le fait d'améliorer l’amplitude de la courbe en ne commençant pas à 0 permet d’améliorer et d’étendre cette visibilité de comparaison, qui est le but recherché dans ces diagrammes.
Afin de ne pas prendre une amplitude arbitraire différente sur l’axe vertical sur chaque visualisation de cette page, qui pourrait détériorer la comparaison entre les différents types d’essence, une fourchette de 0.03 centimes a été choisie entre la valeur la plus haute et la

valeur la plus basse de l'axe. Ainsi, les tailles de barres verticales sont cohérentes entre les différentes essences.

La couleur orange clair est choisie pour être en raccord avec les autres visualisations. En effet, c'est la couleur choisie pour représenter les données de relevés de stations essence.

Pour ce qui est des légendes, l'axe horizontal est assez évident avec les jours de la semaine. En revanche, la légende de l'axe vertical "Moyenne prix eur/liter" permet de donner l'information que nous sommes sur des moyennes de prix pour chaque jour de la semaine, de connaître la devise utilisée et la quantité que l'on mesure.

Enfin, le titre sur chaque graphique permet de bien définir quel est le type d'essence que nous sommes en train de visualiser.

Sur l'ensemble des visualisations, une attention est apportée à avoir une bonne visibilité typographique ainsi que des contrastes élevés afin d'être accessible aux personnes ayant des handicaps visuels.

C2.2.2 : Présenter des recommandations, en préparant son discours et des arguments, en structurant son analyse sur les données représentées afin d'aider les décideurs à établir leurs stratégies.

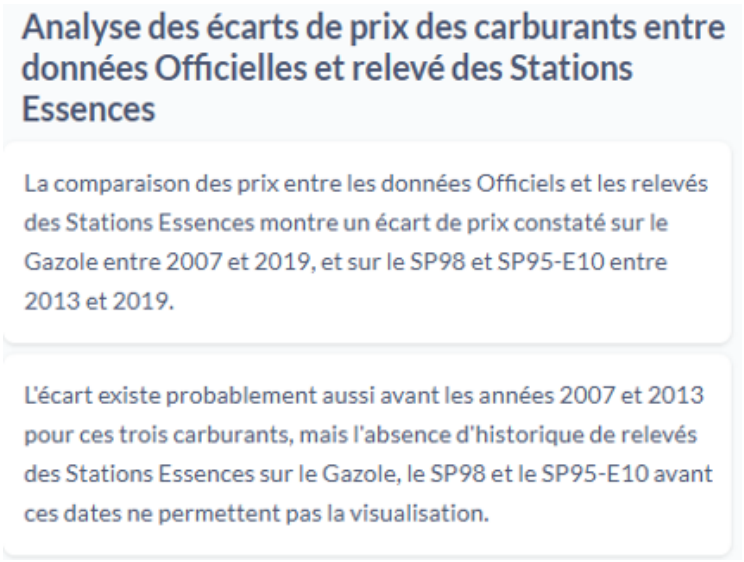
Afin de présenter les recommandations structurées, simples et impactantes, l'ensemble des recommandations a été rassemblées sur une seule page afin d'être très facile à lire et à comprendre pour le client.

La page se compose de trois grandes sections :
La première section est une explication du contexte global de ce qui a été analysé.
La seconde et troisième section sont uniques pour chaque type d'analyse.

Voici **la première section** rappelant le contexte global

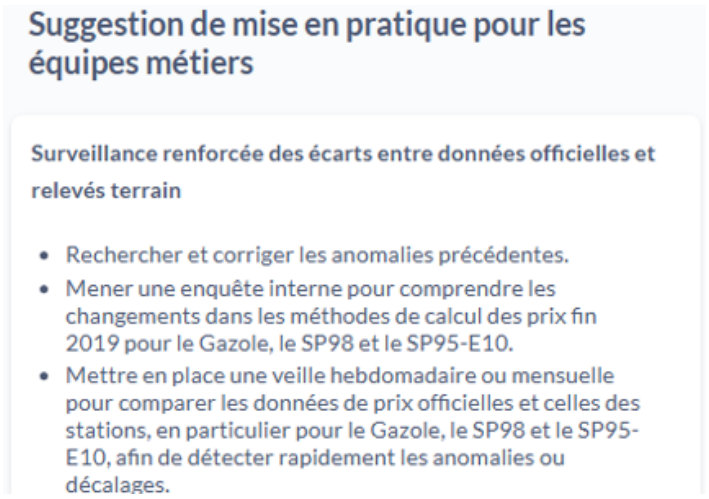
Il s'agit des différents jeux de données qui sont analysés ainsi que les types d'essence disponible et analysés dans ces données.

Pour **la seconde section**, on détaille les résultats d'analyses des écarts de prix des données officielles et celles relevées des stations essence, tout en se basant sur les visualisations des courbes de prix en graphiques temporelles journaliers que l'on a réalisées.



L'objectif est de présenter ici les analyses effectuées, les résultats trouvés, ainsi que l'interprétation de ces analyses, et cela de manière très synthétique.

La troisième section est directement à la suite de la deuxième section. Elle indique les suggestions de mise en pratiques à réaliser suivant les interprétations d'analyses de la seconde section.



On peut y voir les suggestions des différentes actions à réaliser, afin d'orienter les stratégies qui peuvent être déployées.

Les arguments qui justifient ces suggestions sont les informations fournies sur la section d'analyses, ainsi que les visualisations que l'on a conçues et qui sont fournies au client.

Pour présenter au client les résultats des analyses de l'impact temporelle des jours de la semaine sur le prix des relevés des stations essence, on réalise à nouveau une section 2 ainsi qu'une section 3

Dans **la section des analyses**, on peut y voir la constatation d'écart de prix sur les dimanches par rapport aux autres jours de la semaine, les essences fortement impactées ainsi que ceux faiblement impactés. Les diagrammes précédemment réalisés sont disponibles pour le client sur une autre page afin d'appuyer ces analyses.

Dans **la section des suggestions**, on développe plusieurs propositions en lien avec les résultats des analyses sur ce sujet, comme par exemple mener une enquête de terrain pour comprendre pourquoi les essences sont plus chères le dimanche, ou encore chercher à

identifier si cette différence de prix du dimanche vient de choix commerciaux ou de contraintes logistiques.

Enfin, sur l'architecture de rendu global qui sera fournie au client, le dashboard Métabase est constitué de plusieurs pages ayant pour chacune d'elles une vision d'analyse, avec des textes explicatifs permettant de mieux comprendre le pourquoi de ces analyses.
La page de bilan et recommandation est donc à la fin de ce dashboard, pour finir sur les recommandations une fois que le client a terminé de voir l'ensemble des analyses.
Le fait de tout rassembler sur un même dashboard Métabase permet de centraliser les informations, et de permettre qu'elles soient facilement partagées à travers les équipes du pôle data du gouvernement.

C2.3.1 : Former les utilisateurs à l'utilisation des données et des outils de visualisation en analysant le besoin de montée en compétences et en élaborant des supports de formation et de sensibilisation adaptés afin de permettre aux utilisateurs de maîtriser l'exploitation des données.

Il a été remonté l'information que de nombreuses personnes du pôle data du gouvernement n'ont jamais utilisé Métabase.
Un besoin de formation est donc nécessaire pour répondre à la situation.

Afin d'être disponible pour le plus grand nombre, un support de formation est donc créée et sera accessible sur un espace de stockage cloud S3 à cette adresse:
https://open-documentations.s3.eu-west-3.amazonaws.com/Support+de+formation+Metabase.pdf
Ce document est la suite logique du fichier "readme.md" du répertoire "project_master_METABASE" permettant de lancer l'application sur la partie Métabase.

Ce support de formation est destiné aux data-analystes, et éventuellement à certains data-engineers de l'équipe data du gouvernement.
Il s'agit d'un support de formation dédiés spécifiquement aux données mises en place, et permet d'apprendre comment utiliser le logiciel sans connaissances préalables sur ces données ou sur cet outil.

Cette formation commence en tout premier lieu par inviter les gens à lire le fichier "readme.md", afin d'être certain que leurs installations sont bien réalisées.

Ensuite, il leur est demandé d'**importer les appels APIs sur Postman** afin d'avoir directement accès à l'ensemble des APIs ainsi que leurs paramètres associés.

Il leur est ensuite demandé de lancé l'API '/utils/restore_metabase_db_from_S3' avec le paramètre de base de données d'exemple '"zipname": "metabase_db_example".
Cela va automatiquement modifier la base de données native de Métabase, en **installant une nouvelle base contenant l'ensemble des visualisations générées durant l'analyse des jeux de données** qui nous intéressent ainsi que la présentation finale.

Cela va aussi créer l'ensemble des connexions vers la base de données MongoDB cloud, afin d'accéder aux données.

Enfin, ils peuvent se connecter sur le port 3000 avec toutes les installations réalisées.

Sur la seconde partie du tutoriel qui est la découverte de l'outil de Métabase, **un parcours avec des screens et des flèches rouges** aux endroits où il faut cliquer permettent à l'utilisateur de monter en compétences petit à petit sur l'outil, sur les jeux de données et les graphiques des données qui nous concernent, tout en étant accompagné pas à pas.

Durant ce parcours, l'utilisateur va apprendre comment accéder à un dashboard, comment voir une visualisation et zoomer dedans et comment voir la méthode de calcul et les données utilisées lors de la création de cette visualisation.

Enfin, il va apprendre à créer ses propres visualisations.
Pour aider l'utilisateur à naviguer sur les jeux de données à utiliser, une explication détaillée des données ainsi que de leurs correspondances est fournie dans ce tutoriel.

Choisissez votre collection. Nous vous invitons à choisir la collection "Denorm_Station_VS Official_Prices" car il s'agit de la plus complète.
Voici toutes les bases et collections disponibles:
base datalake:

- Gas Stations Infos
- Gas Stations Price Logs Eur
- Official Oils Prices

bdd denormalization:

- Denorm Station Prices
- Denorm Station Vs Official Prices

explications:
Gas Stations Infos: contient les informations des stations essences (pas des prix). Les données sont mise a jour et non additionné (13 665 enregistrements)
Gas Stations Price Logs Eur: contient tous les prix enregistrés durant toutes la journée pour toutes les stations sur toutes les essences. (attention, 40 millions d'enregistrements)
Denorm Station Prices: Contient des moyennes de prix de tous les jours pour toutes les stations pour toutes les essences, afin de ne faire plus qu' un seul enregistrement par jour(6 700 enregistrements).

Official Oils Prices: Contient les prix officiels gouvernementaux de chaque types d'essence par cycle d' un enregistrement par moyenne hebdomadaire.(2 000 enregistrements).
Denorm Station Vs Official Prices: Rassemble la table Denorm Station Prices et Denorm Station Vs Official Prices sur la date pour en faire des comparaison. (7 900 enregistrements)

C2.3.2 : Rédiger la documentation technique d'utilisation du système d’analyse de données en identifiant le public concerné, en détaillant le fonctionnement du système d’analyse de données afin d'assurer la traçabilité et la transmission aux utilisateurs.

Il est important que les Data-Engineers et les Data-Analystes du pôle data du gouvernement puissent comprendre en détail le fonctionnement et les transformations des données de l’application développée.

Une documentation a donc été créée pour rassembler l'ensemble des informations techniques en lien avec cette application.

Afin d’être disponible pour toutes les personnes concernées, ce fichier sera stocké sur un espace de stockage cloud S3 et disponible à cette adresse:
https://open-documentations.s3.eu-west-3.amazonaws.com/Documentation+technique+Datapipeline_comparaison.pdf
Ce document est la suite logique du fichier “readme.md” du répertoire “project_master_ETL” permettant de lancer l’application sur la partie des ETLs.

Ce document est constitué de plusieurs sections:

- L’introduction

C’est dans cette section qu’il est expliqué quels sont les destinataires de cette documentation et l’information qu’il n’y a pas besoin d’être un expert dans le domaine pétrolier pour comprendre et pouvoir travailler sur cette application.

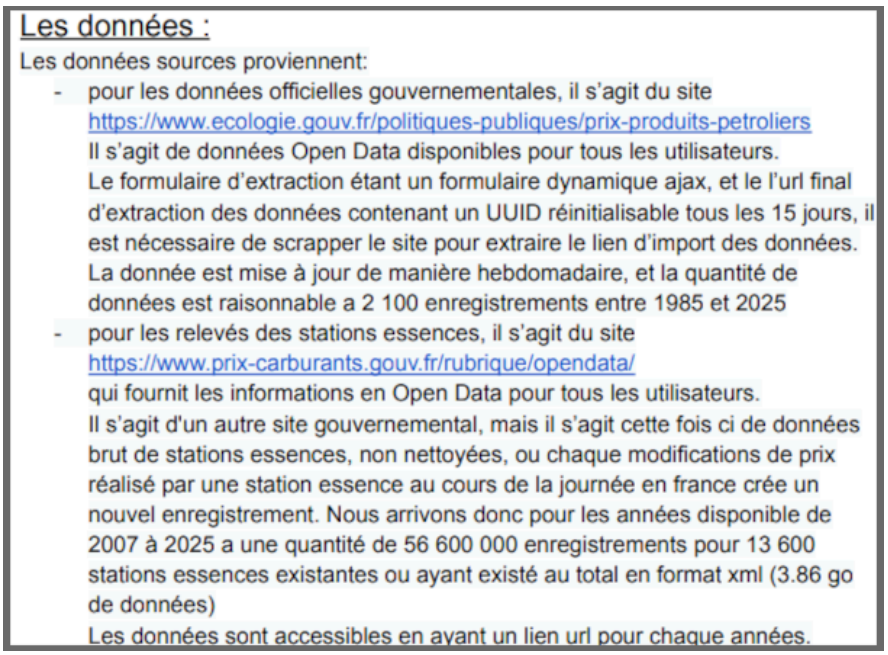
- Résumé du projet

Cette section consiste à expliquer de manière synthétique l’objectif du projet, c'est-à-dire de comparer les données officielles avec les données relevées par les stations essences sur six types d’essences différents, afin de comparer les résultats.

- Les données

Cette section est constituée de deux parties.

La première partie consiste à expliquer la provenance des jeux de données, avec des liens permettant d’accéder aux données sources et aux explications des sites fournissant ces données. Il est aussi expliqué comment les extraire ainsi qu’une petite explication du contenu des données.

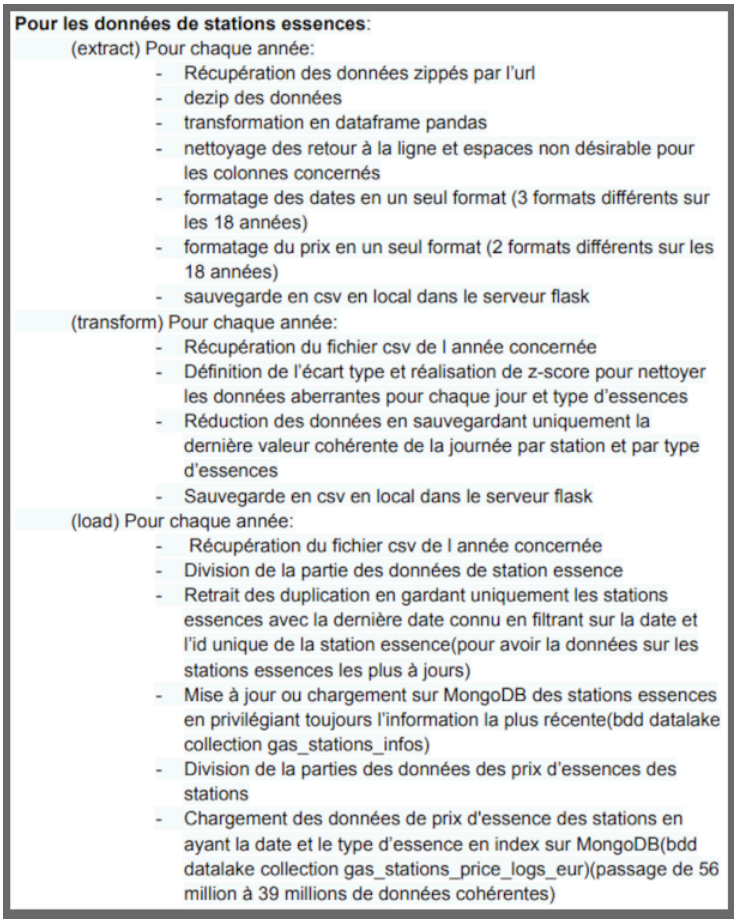


Les données :

Les données sources proviennent:

- pour les données officielles gouvernementales, il s'agit du site https://www.ecologie.gouv.fr/politiques-publiques/prix-produits-petroliers
  Il s'agit de données Open Data disponibles pour tous les utilisateurs.
  Le formulaire d'extraction étant un formulaire dynamique ajax, et le l'url final d'extraction des données contenant un UUID réinitialisable tous les 15 jours, il est nécessaire de scrapper le site pour extraire le lien d'import des données.
  La donnée est mise à jour de manière hebdomadaire, et la quantité de données est raisonnable a 2 100 enregistrements entre 1985 et 2025
- pour les relevés des stations essences, il s'agit du site https://www.prix-carburants.gouv.fr/rubrique/opendata/
  qui fournit les informations en Open Data pour tous les utilisateurs.
  Il s'agit d'un autre site gouvernemental, mais il s'agit cette fois ci de données brut de stations essences, non nettoyées, ou chaque modifications de prix réalisé par une station essence au cours de la journée en france crée un nouvel enregistrement. Nous arrivons donc pour les années disponible de 2007 à 2025 a une quantité de 56 600 000 enregistrements pour 13 600 stations essences existantes ou ayant existé au total en format xml (3.86 go de données)
  Les données sont accessibles en ayant un lien url pour chaque années.

La seconde partie consiste à expliquer les transformations réalisées sur les ETLs pour chacune de ces données sources.
On y retrouve les étapes d'extraction, de transformations et de chargement en base de données dans le datalake, avec pour chacune d'elles des explications détaillées.



Pour les données de stations essences:

(extract) Pour chaque année:

- Récupération des données zippés par l'url
- dezip des données
- transformation en dataframe pandas
- nettoyage des retour à la ligne et espaces non désirable pour les colonnes concernés
- formatage des dates en un seul format (3 formats différents sur les 18 années)
- formatage du prix en un seul format (2 formats différents sur les 18 années)
- sauvegarde en csv en local dans le serveur flask

(transform) Pour chaque année:

- Récupération du fichier csv de l année concernée
- Définition de l'écart type et réalisation de z-score pour nettoyer les données aberrantes pour chaque jour et type d'essences
- Réduction des données en sauvegardant uniquement la dernière valeur cohérente de la journée par station et par type d'essences
- Sauvegarde en csv en local dans le serveur flask

(load) Pour chaque année:

- Récupération du fichier csv de l année concernée
- Division de la partie des données de station essence
- Retrait des duplication en gardant uniquement les stations essences avec la dernière date connu en filtrant sur la date et l'id unique de la station essence(pour avoir la données sur les stations essences les plus à jours)
- Mise à jour ou chargement sur MongoDB des stations essences en privilégiant toujours l'information la plus récente(bdd datalake collection gas_stations_infos)
- Division de la parties des données des prix d'essences des stations
- Chargement des données de prix d'essence des stations en ayant la date et le type d'essence en index sur MongoDB(bdd datalake collection gas_stations_price_logs_eur)(passage de 56 million à 39 millions de données cohérentes)

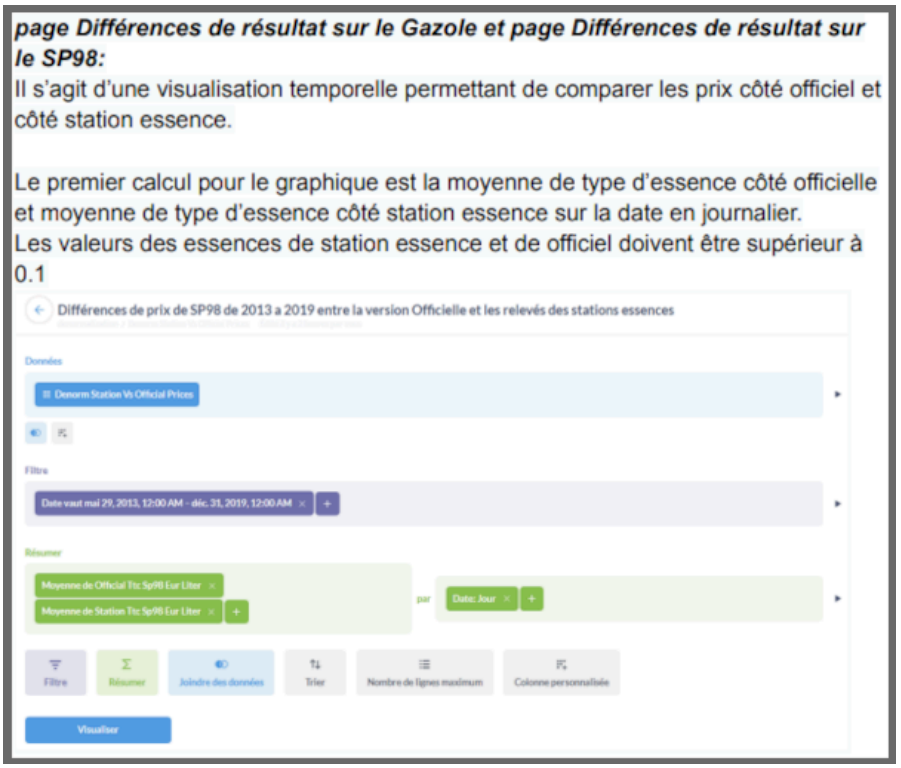
Ces explications détaillées sont aussi réalisées lors de l'étape de dénormalisation des données stockées du datalake vers la base de données de dénormalisation.

- Dataviz

La section de datavisualisation permet de fournir tous les éléments techniques permettant de comprendre et de recréer l'ensemble des visualisations par rapport aux données en bases de données.

page Différences de résultat sur le Gazole et page Différences de résultat sur le SP98:
Il s'agit d'une visualisation temporelle permettant de comparer les prix côté officiel et côté station essence.

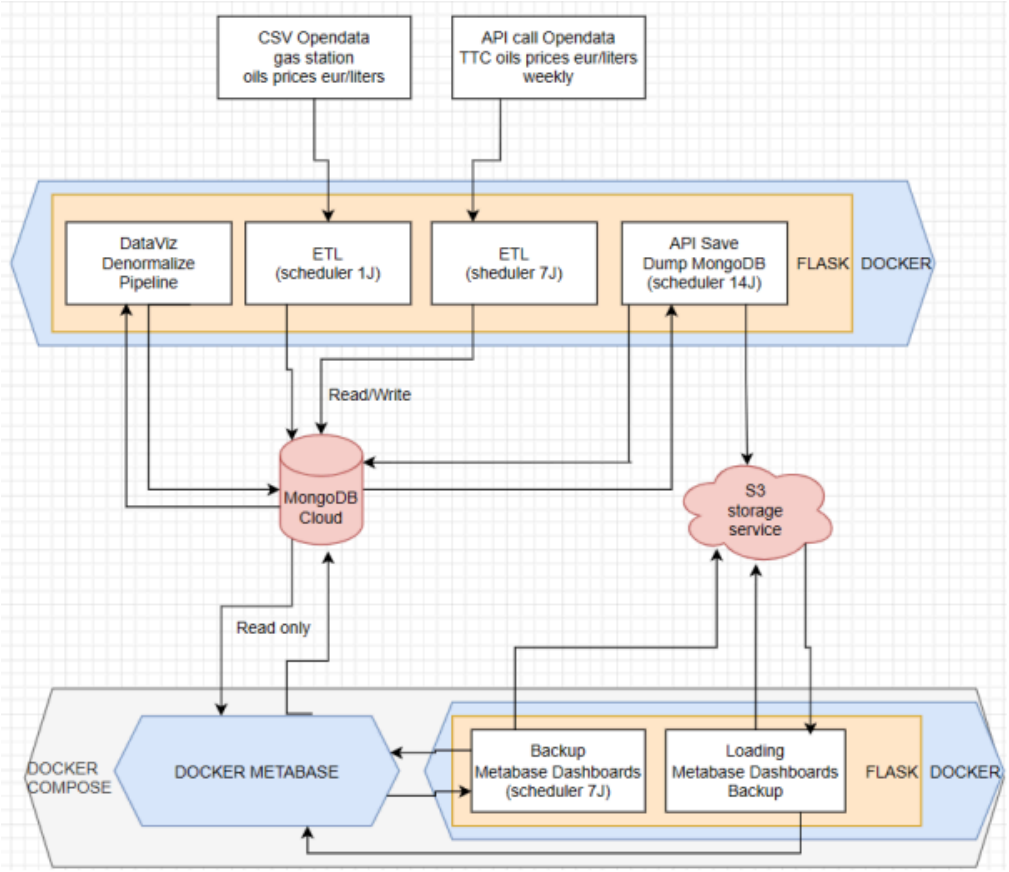
Le premier calcul pour le graphique est la moyenne de type d'essence côté officielle et moyenne de type d'essence côté station essence sur la date en journalier.
Les valeurs des essences de station essence et de officiel doivent être supérieur à 0.1



Pour chaque visualisation contenue dans le dashboard de présentation, une explication est fournie afin de comprendre les collections et les colonnes utilisées, ainsi que les méthodes de calcul et les différents filtres pour recréer cette même visualisation.
Un screen de la page des paramètres lors de la création de la visualisation permet de facilement mieux la comprendre et la recréer si besoin.

- Le schéma d'architecture

Cette section est dédiée à faire comprendre l'architecture globale de l'application.
Un schéma d'architecture est donc présentée avec les différents dockers, les serveurs Flask ainsi que la base de données mongoDB et l'espace de stockage cloud S3



Dans cette section, il est expliqué la partie du docker contenant toute la partie des ETLs, la partie du docker-compose contenant toute la partie de visualisation, ainsi que la partie centrale contenant MongoDB en cloud et l'espace de stockage cloud S3.

Conclusion

Durant la réalisation de ce projet, j'ai mis un effort supplémentaire sur la partie d'interconnection entre différents outils cloud, dans l'objectif de créer une vraie architecture semblable à un réel projet d'entreprise.

J'ai beaucoup appris sur ce projet, que cela soit sur la recherche de données OpenData, sur la partie de conception des ETLs, mais aussi sur toute la partie de visualisation, dont j'ai fait peu de réalisations durant mon alternance.

Cela m'a permis de découvrir l'outil Métabase, que je trouve révolutionnaire et que j'apprécie énormément.

L’aide de Large Language Model n'a pas été utilisée lors de l’écriture de ce mémoire, mais cela m’a aidé lors de difficultés d’implémentation sur le projet. Par exemple, lors de l’import de MongoDB-tools sur le dockerfile. Cela m’a fait gagner beaucoup de temps sur certains points, et cela me permet de mieux me concentrer sur les problèmes d’architectures ou de problèmes métiers.

Je vous remercie du temps que vous avez passé à lire mon mémoire.