

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 3**](#_ybo3mfsh670c)

[**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. 3**](#_jy30bt2p9iwd)

[**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. 3**](#_b3ft6usja70a)

[**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. 3**](#_xu4v09r5mxto)

[**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. 3**](#_llo8lkahs3yt)

[**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. 3**](#_ueljvy826n95)

[**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. 3**](#_hrv3uxlrcrci)

[**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 utilisateur de maîtriser l’exploitation des données. 4**](#_iouoshj55hut)

[**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. 4**](#_wu78rnjl1i7c)

[**Conclusion 4**](#_beorsspftaph)

# 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 ou 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 ou 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 enseignement.

En plus de cette mission, j’ai aussi eu l’opportunité de partir d’un projet de zéro pour le client Renault, ou 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 aider 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, ou 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 par un Data-Engineer Sénior durant la première année 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 Sénior 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 au yeux du client (ayant un Data-scientist en référent technique si besoin). J’ai réalisé des réunions avec le client une fois par semaine, ou 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 ou 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 sur ce mémoire a était 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 relevés lors de cette entretien afin de cadrer au mieu le travail d’analyse a produire:

**Problématique:**

Le pôle data du gouvernement français possède un jeu de données officiel sur l'évolution du prix des différentes essences vendu 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 celle-ci ont été calculées.

Ils se posent donc la question de la véracité et fiabilité des transformations des données anciennement récoltés, 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 vendus en France, afin d'en tirer des conclusions et avoir une meilleure visibilité en terme journalière. 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 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 issus 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.

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 anonymisé. 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 relevés durant la prise des besoins et des attentes, **les objectifs a réalisés permettant de répondre aux problématiques client** sont les suivants :

* Réaliser un audit sur la qualité des données officiels hebdomadaires en comparaison à un autre jeu de données journalier gouvernemental.
* Rechercher des décalages de prix sur les données historiques qui suggérerait un changement de méthode de calcul des moyennes hebdomadaires officielles.
* Analyser sur une granulométrie temporelle plus fine qu' actuellement sur 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 sauvegarder et restaurer 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** vendu 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 des **relevés de prix sur les 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 issus 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 des longues périodes identiques afin de pouvoir réaliser des comparaisons de données pertinentes sur le long terme.

Avec les jeux de données nécessaires, les problématiques clients peuvent se reformuler en problématique numérique permettant d’avoir une vision claire sur 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 suggérerait un changement de méthode de calcul des moyennes hebdomadaires officielles**

Ces objectifs semblent correspondre à un même besoin global qui pourrait nécessiter un plan d’analyse similaire.

Pour répondre à ces deux problématiques, les données nécessaires seront les deux jeux de données. Il faudra les données officielles ainsi que les données issus des stations essences.

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 traités 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 des prix de stations essences 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 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.**

Cette objectif ne demande pas de métriques spécifiques. En revanche, il indique clairement que la solution a déployé devra permettre de se connecter aux données de manière autonome et récurrente pour récupérer les dernières informations, puis 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 qui doivent pouvoir être sauvegarder et restaurer dans un environnement externe, afin d’assurer la facilité de déploiement du système dans d'autres environnements.**

La 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, zipper, 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.

En prenant en compte les métriques de données nécessaires que nous venons de voir et afin de répondre au mieux aux différentes problématiques, le travail de traitement de données a été réalisé par un serveur Flask avec des planificateurs de tâches permettant de récupérer les données les plus récentes chaque jour.

Ces données ont été extraites, transformées, puis stockées sur une base de données MongoDB afin d’être disponible pour la visualisation.

Tout ce travail de préparation de données a été décrit et analysé sur le Bloc 1 de ce titre RNCP.

# 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**

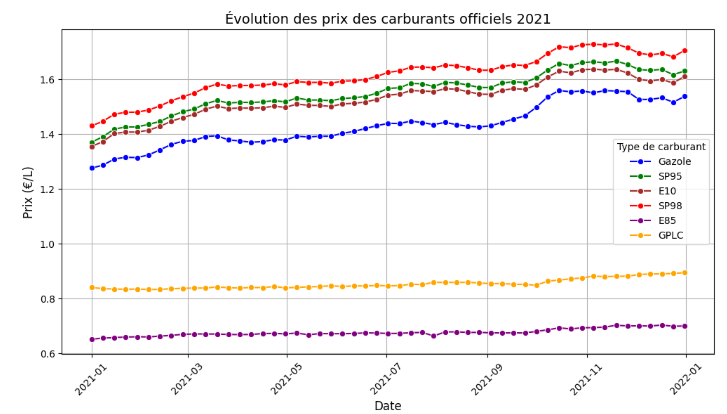
nécessité de créer des visualisations en graphique temporelles pour le jeu de données des prix des essences officielles, ainsi que pour le jeu de données des relevés de prix des stations essences.

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.

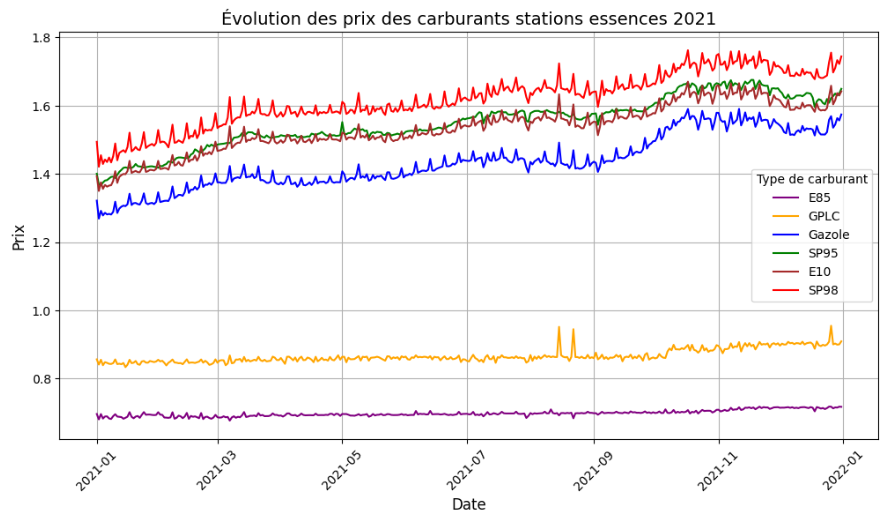
La première analyse se fait sur les données officielles année après année :



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 suivant les types d'essence.

Par exemple, le Gasoil commence en 1985, le SP95 et SP98 en 1990, etc..

La seconde analyse se fait sur les données des stations essences 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 suivant les types d'essence.

Par exemple, le Gasoil, SP95, GPLc et E85 commence en 2007, le SP95-10 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 18 ans pour le Gasoil ou le SP95 par exemple.

Nous pouvons aussi voir que sur les données journalières des stations essences, 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 essences en journalière.

Cette seconde visualisations des données des stations essences nous permet donc d’y voir un peu plus claire sur l’objectif de “**l’analyse sur une granulométrie temporelle plus fine qu'actuellement sur les prix des essences**” et sur les possibilités qu'offre ce jeux de données

Afin d’approfondir ces analyses, et de pouvoir plus facilement jouer avec les jeux de données, une implémentation de ces données sur un docker Métabase va permettre de pouvoir plus facilement faire des comparaisons, ainsi que d’avoir une navigation simplifiée sur les courbes.  
Pour réaliser cela, il faut connecter MongoDB a 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 essences ont été rassemblés sur la même collection, et des variables pertinentes permettant de faciliter la visualisation ont été ajoutées.

Il suffit donc simplement d’importer la collection qui nous intéresse sur Métabase.  
Voici l’architecture de cette collection appelé “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

**Day\_of\_week:** "Friday"

**Month:** "August"

**Year:** "2021"

**DayMonth:** "27august"

}

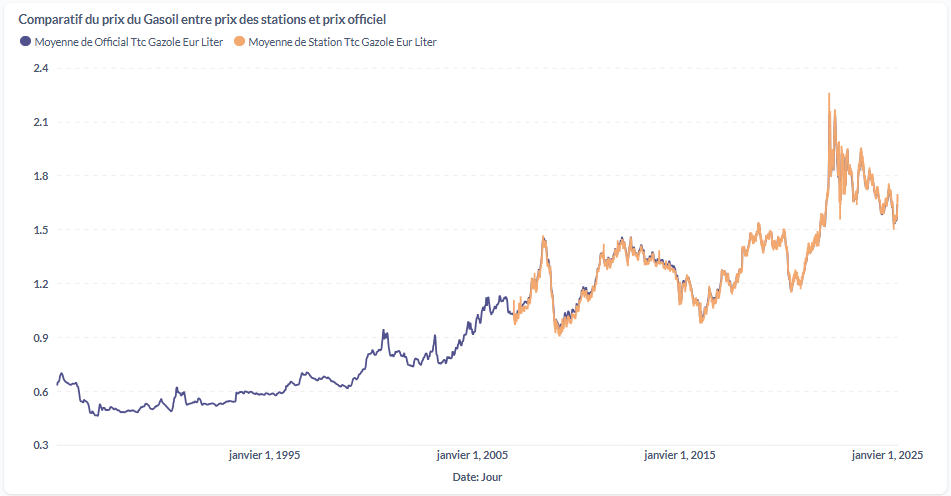
Une fois la table importée sur Métabase, on va créer un dashboard appelé “dataviz”. L’objectif de ce dashboard va être de réaliser tout types d’analyses et de juger de leurs pertinences.

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 essences.

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 qu’une valeur par jour sur chaque colonne)



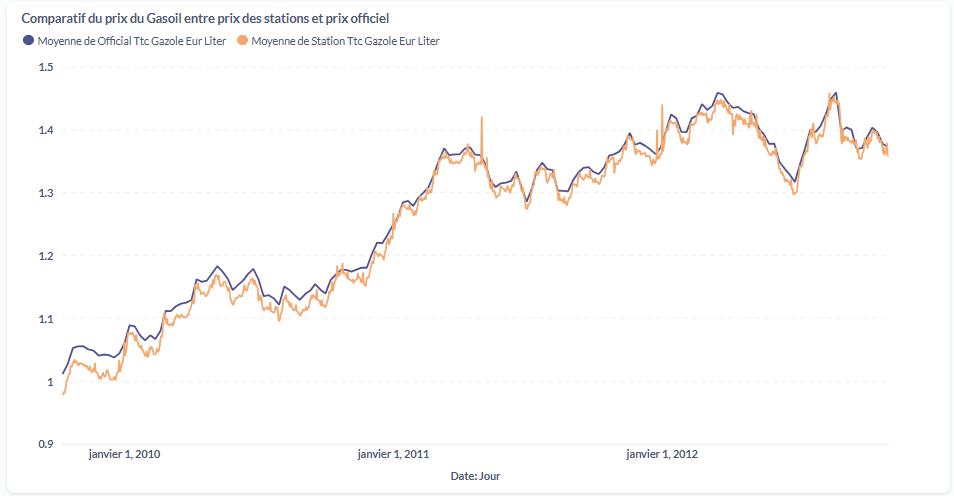
Voici le résultat de ce graphique temporel:



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 foncé), qui commence en 2007.

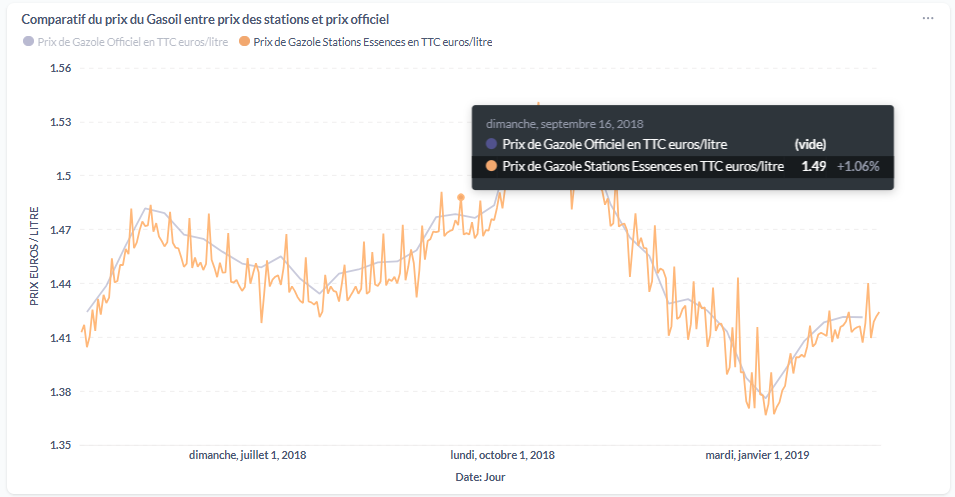
Le plus gros avantage de Métabase est que l’on peut zoomer dans les données.

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



C’est donc un point important a relevé 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 pattern de prix se dessiner constitué de pics et de creux à intervalle régulier.  
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 pattern est visible sur les données des stations essences, mais ne peut pas être visible sur les données des données officielles.

Des analyses similaires à celles ci-dessus sont ainsi effectués sur chacun des six types d’essences différents, afin de chercher des potentiels points 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.

test

# 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.

# 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.

# 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 utilisateur de maitriser l’exploitation des données.

# 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.

# Conclusion