## horizontal line



GetItScraped - DRAFT

Octobre à Décembre 2023

Michel Douglas Piamou

Mike Boudhabhay

Edwige Fève

Introduction au projet

# Présentation de l'équipe projet:

## Pour chacun des membres du groupe, préciser le niveau d’expertise autour de la problématique adressée ?

* + Michel : Mon cœur de métier est le Conseil en transformation SI. sensibilisé aux enjeux de performance de la chaîne de production vis-à-vis de la satisfaction client.
  + Mike : Mon background est orienté sur l'entrepreneuriat ( Gérant TPE) avec un socle de connaissances financières, contrôle de gestion.Pas de spécialisation antérieures sur la Supply chain.
  + Edwige : mon cœur de métier est la Supply Chain, par contre le scraping de données et la catégorisation de commentaires complètement nouveaux pour moi.

## Êtes-vous entré en contact avec des experts métiers pour affiner la problématique et les modèles sous-jacents ? Si oui, détaillez l’apport de ces intéractions.

Pas pour le moment. A confirmer si l’on veut intégrer une approche plus commerciale/marketing à notre analyse.

## Avez-vous connaissance d’un projet similaire au sein de votre entreprise, ou bien dans votre entourage ? Quel est son état d’avancement ? En quoi vous a-t-il aidé dans la réalisation de votre projet ? En quoi votre projet contribue-t-il à l’améliorer ?

* + Michel: non
  + Mike: non
  + Edwige : non

# Contexte

**Contexte d’insertion du projet dans notre métier :**

La satisfaction du client, la réputation en ligne et l’ajustement aux besoins du marché sont au cœur de la problématique de toute entreprise qui souhaite rester compétitive et profitable sur le long terme. Préoccupation d'autant plus cruciale que nous sommes dans un monde dématérialisé où l'information est disponible et circule à grande vitesse.

Le développement de nouveaux produits ou la réponse à une demande plus importante que prévue peut engendrer - en interne - de nombreux surcoûts, des retards, des non-conformances qu’il est important de capturer mais aussi d’un point de vue externe - une insatisfaction du client final.

De la même façon ne pas être capable de s’adapter à une baisse de la demande sur un produit qui a été cannibalisé ou qui devient obsolète est critique pour l’entreprise et ses distributeurs afin d’éviter de générer des stocks de futurs invendus.

La latence d’une supply chain aujourd’hui est beaucoup plus longue que ce que l’on pourrait penser. Les flux sont globalisés avec des fournisseurs, des sites de production et des clients très dispersés géographiquement.

Réagir au plus vite et rester connecté à la demande est donc vital pour rationaliser les coûts opérationnels.

Enfin, il est à noter qu’au-delà de la qualité et de la disponibilité d’un produit, les attentes des clients sont aussi environnementales. Pouvoir capturer ces attentes permet aux entreprises de réfléchir à leur positionnement et de retravailler leurs produits.

**En résumé, être à l’écoute des ses clients au-delà d’une simple mesure de taux de service s’avère fondamental pour maintenir sa base de clients, mais en conquérir de nouveaux et ce afin de rester prospère mais aussi de se positionner avantageusement vis à vis de la concurrence.**

# Enjeux

Aujourd'hui, les données de satisfaction clients sous formes de commentaires, d’avis et de notation sont disponibles sur Internet.

Leur exploitation pertinente peut permettre aux entreprises d'identifier leurs domaines d'amélioration potentiels, d'évaluer la satisfaction globale de leurs clients, de communiquer les messages de la marque, d'identifier les produits ayant fait l'objet du plus grand nombre d'avis, et bien d'autres choses encore. De ce fait, l'exploitation de ces commentaires est centrale pour les entreprises d' e-commerce telles que Amazon, eBay, NetFlix, etc.

On note que le coût global de mauvaises expériences clients est estimé en trillion de dollars. (<https://blog.hubspot.com/service/customer-service-stats>)

Toutes ces informations permettent donc :

* d’évaluer de manière qualitative et quantitative, la performance de leur organisation supply chain et au-delà,
* De mesurer l’adéquation d’un service ou d’un produit par rapport aux attentes du consommateur final et de se réajuster si nécessaire,
* De synthétiser les retours clients pour travailler sur des axes d’ amélioration,
* D’automatiser - de manière performante - le prétraitement des retours des clients insatisfaits (réponse adéquate, redirection sur le service concerné).

**Notre projet : Répondre à ces enjeux avec les outils et techniques de la Data Science et être capable de fournir une classification des problématiques mentionnées par les clients.**

## Du point de vue technique :

La contrainte technique s'affine au fur et à mesure de l’avancée de notre projet mais d’ores et déjà nous pouvons mentionner les problématiques suivantes :

* + Aura-t-on des limites pour concaténer l’information ( i.e clés communes) ou de langue lorsqu’un site collecte des commentaires en plusieurs langues?
  + Sera-t-on capable d’intégrer la spécificité de chaque site ciblé qui demande d’adapter son programme d’extraction à une nouvelle source d’informations?
  + Devrons-nous mettre à jour régulièrement notre programme de scraping si le site évolue en parallèle?
  + Pourrons nous accéder aux données qui peuvent nécessiter des identifiants ou qui peuvent être bloquées car les sites se protègent des requêtes en masse que génèrent les programmes de scraping.

## Du point de vue économique :

Avec les médias sociaux et l’hyper connectivité aujourd’hui, les entreprises ne peuvent plus être auto-centrées sur leurs processus internes uniquement mais être à l'écoute du marché pour :

* + comprendre les nouvelles tendances,
  + anticiper des changements de comportement clients,
  + surveiller leur notoriété, la réputation ayant un impact direct et quasiment immédiat sur le chiffre d’affaires,
  + utiliser la disponibilité des retours clients comme un levier si l’information est exploitée correctement.

Ce travail de veille et d’analyse peut être extrêmement bénéfique si l’entreprise est assez agile pour y répondre ou catastrophique si elle ne sait pas capter l'évolution de son marché.

## Du point de vue scientifique :

Nous allons explorer des techniques de NLP ( Natural Language Processing), d’intelligence artificielle et de deep learning pour le traitement du texte des commentaires clients.

# Objectifs

## Quels sont les principaux objectifs à atteindre ?

* **Être capable de webscraper des données pour extraire des commentaires.**
* **Prédire la satisfaction client** ( note / nombre d’étoiles) à partir des commentaires.
* **De ces commentaires, identifier les catégories de sujets problématiques** : a-t on des problèmes de livraison et de retard, des produits qui arrivent abîmés, des produits potentiellement périmés, etc. )
* A partir des commentaires clients, **être capable d’automatiser une réponse**
* **Détection du sentiment client** : positif, neutre ou négatif - à confirmer
* Création d’une dashboard de suivi - à confirmer

# Cadre

Collecte des avis clients sur des plateformes spécialisées telles que :

* + TrustedShops
  + TrustPilot
  + Amazon

Nous avons considéré les critères suivants pour sélectionner notre jeu de données :

* + - Volumétrie des données disponibles
    - Volumétrie des réponses du fournisseur

Etape 0 - Collecte des données:

# Source des données :

Afin de collecter les données sur les avis clients, nous avons exploré 3 plateformes permettant de récupérer des avis clients avec différents outils ou scripts python :

* Amazon
* Trustpilot
* TrustedShop

Notre choix final s’est tourné vers la **plateforme TrustedShops** pour laquelle nous avons de la flexibilité pour extraire des données des avis clients sur les entreprises ciblées sans changer notre code de web scraping.

Pour des considérations de volumétrie des avis clients et de volumétrie des réponses des fournisseurs, nous avons décidé de réaliser l’étude avec les données d’avis clients sur La Redoute, leader français du e-commerce en mode et maison.

# Méthodes explorées:

Nous avons testé plusieurs méthodes/outils de web scraping pour collecter les avis clients:

* Réalisation de scripts python avec BeautifulSoup
* Usage d’ extensions de navigateur, par exemple le data instance scraper de google
* Applications spécialisées de web scraping telles que Scrapfly qui permettent de bypasser les captcha ou les anti bot.

Nous avons réalisé notre propre programme de web scraping en python avec Beautiful, afin de mettre en pratique les notions de web scraping vus en BootCamp.

Échantillons d’extractions disponibles :

* Rayon d’Or Bagage : <https://www.rayondor-bagages.fr/> : 11 colonnes x 1082 lignes
* Mechatechnic : <https://www.mecatechnic.com/fr-FR/> : 11 colonnes x 6245 lignes
* La Redoute: <https://www.trustedshops.fr/evaluation/info_XA3F5BBDE34B5AD3FB64015EB823AC7C6.html> : 11 col x 34760 lignes

# Web Scraping avec BeautifulSoup:

Le programme de web scraping réalisé se déroule en deux grandes étapes :

* **Étape 1** : **Scraper les données brutes des avis clients sur TrustedShops pour l’entreprise ciblée et les stocker dans un fichier csv.**

Cette étape est réalisée par la fonction *get\_reviews(url, n\_pages, output\_file)* avec :

* *url*: url de la 1ere page de résultats sur les avis clients pour l’entreprise ciblée
* *n\_pages* : nombre de pages de résultats que l’on veut récupérer. Ce nombre doit être inférieur au nombre total de pages web de résultats disponibles pour l’entreprise ciblé.
* *output\_file* : nom du fichier csv dans lequel sont stockés les les avis clients collectés.
* **Étape 2 : Réaliser un premier nettoyage et formatage des données**

Cette 2eme est réalisée à l’aide de la fonction *clean\_reviews(input\_file, output\_file)* avec :

* *input\_file* : fichier csv de données brut collecté dans l’étape 1
* *output\_file* : fichier csv final nettoyé et formaté

La fonction clean\_review réalise principalement les actions suivantes :

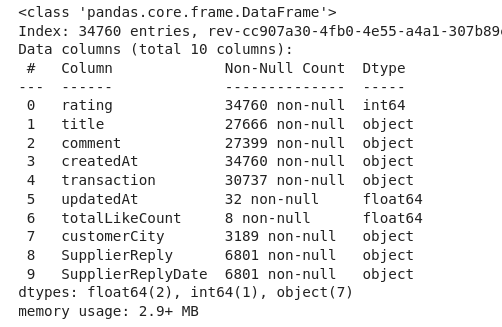
* Suppression des colonnes inutiles : "provider", "inModeration", "verificationStatus"
* Conversion des dates timestamps au format date
* Extraction des réponses des fournisseurs aux avis clients

Le détail du programme de web scraping des avis TrustedShops est disponible dans le notebook “[01\_mdp\_trustedshops\_scraping\_v1.0.ipynb](https://github.com/PMDOUGLAS23/DST_SC_CS/blob/main/notebooks/0_data_collection/01_mdp_trustedshops_scraping_v1.0.ipynb)” sur GitHub.

Étape 1 Exploration des données et Data Visualisation:

# Exploration des Données :

Voici un aperçu de ce que nous avons extrait pour le site de la Redoute, en termes de données brutes. Environ 35000 ratings sont disponibles pour environ 27400 commentaires.



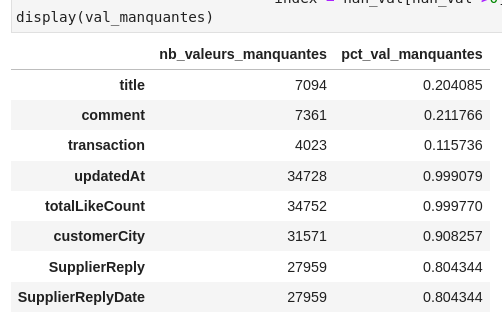
## 

## Préparation des données:

Après extraction des informations, nous avons effectué **un premier nettoyage** de données en supprimant les valeurs manquantes et les colonnes non pertinentes ou les pourcentages de Nan sont trop importants.

Nous avons aussi enlevé les lignes sans commentaires, les lignes ou les dates de transaction étaient manquantes.

Cf ci-dessous :



## Enrichissement du jeu de données :

Ensuite, nous avons souhaité a**pprofondir la dimension temporelle e**n extrayant les mois, jours ouvrés/week-end ainsi que les heures auxquelles les commentaires ont été postés.

* date\_of\_cmt
* year\_of\_cmt
* month\_of\_cmt
* cmt\_in\_weekend
* hour\_of\_cmt
* year\_of\_transaction
* month\_of\_transaction
* day\_of\_transaction
* weekday\_of\_transaction
* transaction\_in\_weekend
* transaction\_date

Nous avons aussi ajouté les variables suivantes qui nous permettent de calculer le nombre de mots des commentaires du texte ou du titre:

* comment\_length
* log\_comment\_length
* title\_length

Le log est une fonction logarithmique appliquée aux commentaires pour étudier la distribution des données en fonction du rating.

# Data Visualisation

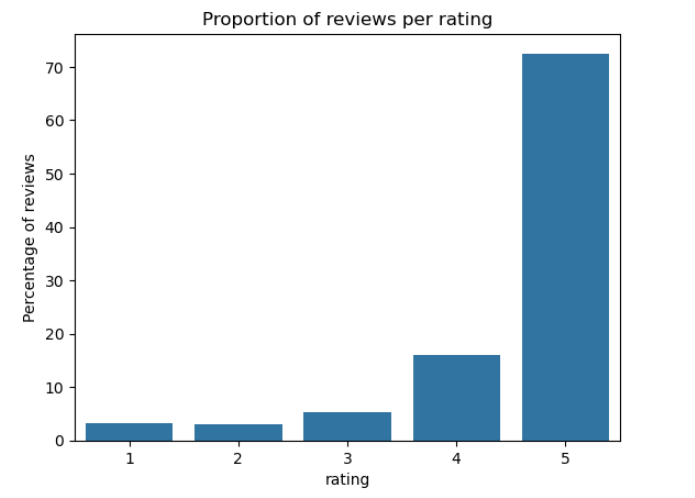
Nous avons ensuite grâce aux outils de visualisation connus cherché à comprendre comprendre la distribution de ces variables, leurs corrélations et plus particulièrement la relation entre le rating et les autres variables.

L’ensemble des graphiques se retrouve [ici](https://github.com/PMDOUGLAS23/DST_SC_CS/blob/main/notebooks/1_eda_dataviz/01_mdp_eda_redoute_reviews_1.0.ipynb) mais dans le cadre du compte-rendu écrit, nous avons décidé de choisir les plus pertinents.

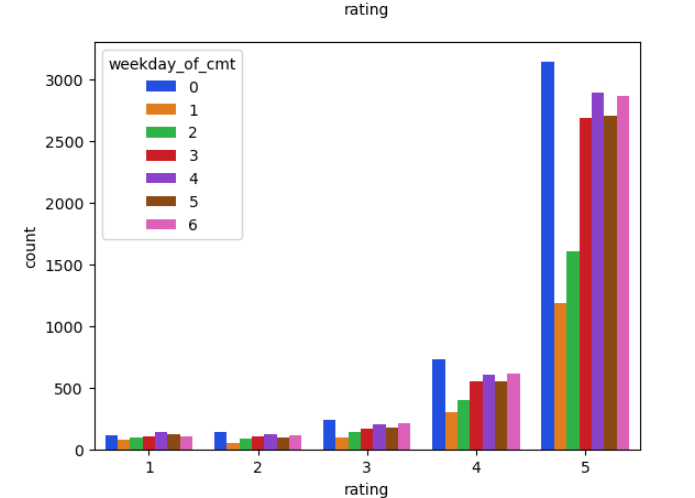
## Analyse de la distribution des variables:

### Graphique 1 : Proportion de reviews par rating:

On constate ici, avec surprise, que la **majorité des commentaires se fait lorsque le rating est extrêmement positif, càd à 5.** En d’autres termes, lorsque les clients sont satisfaits, ils sont plus enclins à commenter.



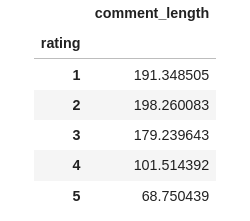
### Graphique 2: Nb de commentaires en fonction du jour de la semaine :

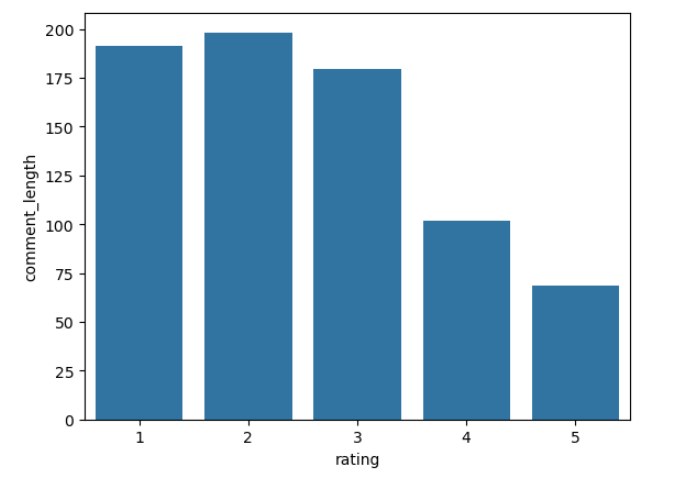


Complétés avec d’autres graphiques non retranscrits ici, on constate que :

* Les revues ont principalement été effectuées en 2023
* La distribution des ratings par mois est presque uniforme pour les notes allant de 1 à 3. Par contre, il y a plus de disparités sur les notes de 4 et 5.
* Il y a des **pics de commentaires en mars, avril, juillet, septembre et octobre** pour la plupart des ratings mais surtout pour ceux allant de 4 à 5.
* **La plupart des revues sont faites entre le vendredi et le lundi**
* Les revues sont principalement faites dans l’après-midi, entre midi et 14h, sinon le matin entre 6h et 10h

### Graphique 3: Longueur moyenne du commentaire =f(rating)

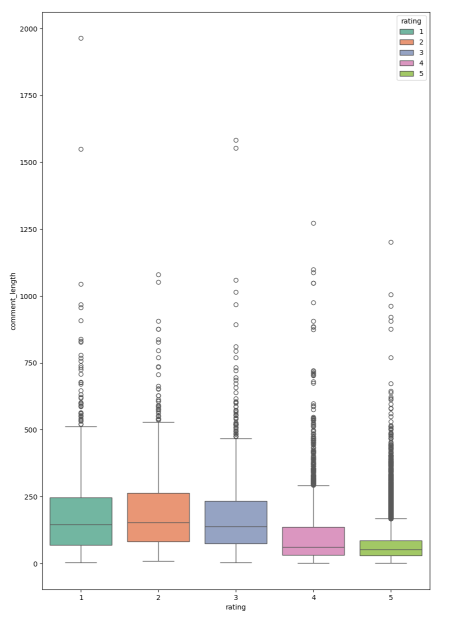




En moyenne, **on constate que les revues avec des notes inférieures ou égales à 3 ont tendance à avoir des commentaires plus longs**, ce qui paraît logique car le client manifeste son insatisfaction et utilise plus de mots pour la décrire.

**La moyenne de la longueur du commentaire diminue lorsque la note s’améliore.**

### Graphique 4: Boîte à moustaches pour vérifier les outliers et les médianes:



Avec ce type de graphique, on constate que **la médiane de la longueur de commentaires est nettement plus basse pour les notes de 4 et de 5** mais aussi qu’il y a une longue queue d’outliers avec asymétrie positive.

### Graphique 6 : Word cloud sur les commentaires



On constate qu’il y a d**u nettoyage de texte à faire pour avoir un rendu plus significatif**.

Il est aussi important de noter qu’un **mot pris hors de son contexte ne nous aide pas à prédire le sentiment du client.**

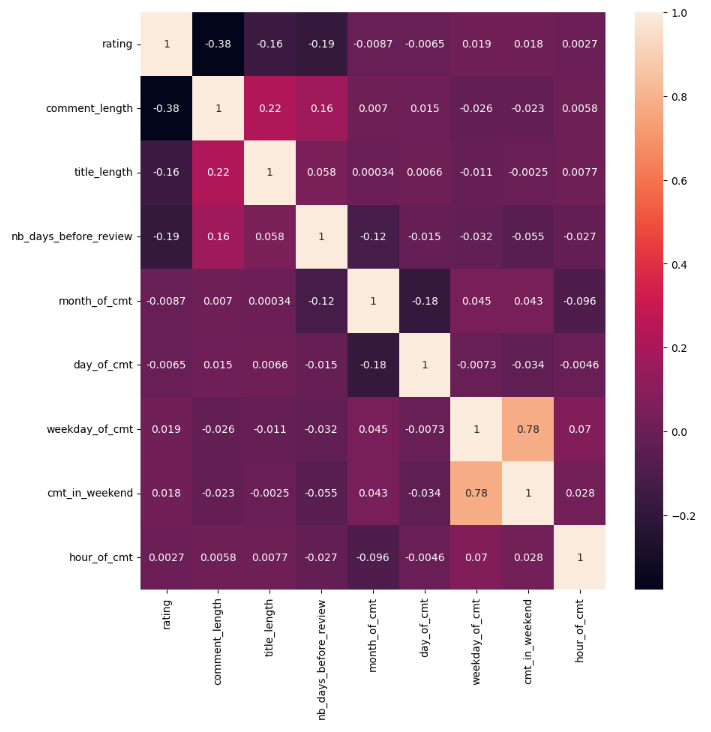
Par exemple, le mot “livraison” est mentionné très fréquemment mais on ne sait pas conclure si la livraison s’est bien passée ou si finalement elle a eu du retard ou si le colis a été endommagé.

## Analyse de corrélation :

L’objectif étant d’identifier les variables corrélées à **notre variable cible : le rating.**

### Graphique 5 : Analyse de corrélation par heatmap :

### 



Ici, on constate les **premières corrélations entre la longueur du commentaire (‘comment\_length’) et le rating, avec une valeur de -0.38.**

Il y aussi 2 autres corrélations moins marquées avec le rating et qui sont pour les variables : **‘nb\_days\_before\_review’ et ‘title\_length**’, respectivement avec les valeurs -0.19 et -0.16.

## Tests statistiques:

Notre premier objectif est de valider statistiquement la relation entre :

* le rating et les commentaires,
* Le rating et les dates de revue,
* Le rating et la longueur du titre

Le deuxième objectif est de vérifier la nature de la corrélation entre les variables explicatives.

Nous avons réalisé plusieurs tests pour identifier des corrélations entre les différentes variables.

* Test de Anova : nous n’avons pu appliquer Anova qu’entre la fonction logarithmique de la longueur du commentaire et le rating qui suivent une distribution normale confirmée par un diagramme QQ plot.

Pour les autres variables comme la longueur du titre et le nombre de jours avant revue, nous n’avons pas pu l’utiliser. A ce stade,  **nous pouvons confirmer que le rating a un effet significatif sur la longueur du commentaire.**

* Test de Kruska Willis : ce test nous permet de confirmer scientifiquement la relation statistique **entre le rating et la longueur des commentaires mais aussi entre les titres et le nombre de jours entre la transaction et la date du commentaire.**
* Test de Spearman : le test de Spearman nous fait conclure qu**’il y a une corrélation entre la longueur du commentaire et le rating, mais aussi entre la longueur du titre du commentaire et le nombre de jours entre la transaction et le commentaire ( nb\_days\_before\_review).**

A ce stade, nous avons donc **trois variables explicatives possibles pour prédire le rating** :

* La longueur du commentaire client
* La longueur du titre du commentaire
* Le nombre de jours entre la date du commentaire et la date de la transaction.

Afin de confirmer qu’il n ’y a pas de redondance entre ces variables explicatives candidates, nous avons analysé la colinéarité entre ces 3 variables à l’aide du VIF (Variance Inflation Factor qui est une mesure du degré de colinéarité entre les variables explicatives dans un modèle de régression multiple).

*Calcul du VIF avec variance\_inflation\_factor de statsmodels.stats.outliers\_influence***.**

|  |  |
| --- | --- |

**Le VIF calculé vaut pratiquement 1 pour chacune des variables. Donc, nos trois variables ne sont pas colinéaires.**

## Synthèse de l’analyse exploration et data visualization :

La première partie de nos analyses nous fait constater les choses suivantes :

* **Constat 1 - Analyse des ratings :** les notes excellentes (ratings 4 et 5) sont sur-représentées. On risque donc d’avoir un biais sur cette information car 70% des commentaires se font lorsque la note est excellente.

Il faudra prendre en compte ce déséquilibre des données dans la phase de modélisation et sûrement réfléchir à un ré-échantillonnage des données.

* **Constat 2 - Analyse des commentaires clients :**  l’analyse de mots hors contexte ne nous permet pas de conclure sur la satisfaction du client et il faudra par conséquent utiliser des outils plus sophistiqués capables de regrouper les mots ( word embedding?)

Il va falloir réfléchir à comment encoder du texte en numérique pour être capable d’utiliser les outils de NLP.

* **Constat 3 - Relation entre les variables et le rating :** L’analyse exploratoire et les tests statistiques nous ont permis de valider les points suivants :
  + La longueur du commentaire, la longueur du titre du commentaire, le nombre de jours entre la transaction et l’émission de l’avis sont statistiquement corrélés au rating
  + Le calcul du VIF montre que la longueur du commentaire, la longueur du titre du commentaire, le nombre de jours entre la transaction et l’émission de l’avis ne sont pas colinéaires.

Par conséquent, nous pourrons choisir ces 3 variables comme variables explicatives pour prédire le rating.

* **Constat 4 - Analyse des réponses fournisseurs :** on constate que les réponses semblent être génériques. Ce qui confirme l'intérêt de générer des réponses plus adaptées.
  + L'évaluation moyenne des fournisseurs et le nombre de réponses fournisseurs par mois varient mensuellement, avec un pic autour de mars 2023. Nous allons étudier la saisonnalité dans les prochaines semaines.
  + Le délai moyen de réponse s'allonge avec l'augmentation de l'évaluation.
  + Les fournisseurs avec une note de 5 ont un délai de réponse très long.
  + La majorité des délais de réponse se situent en dessous de 10 jours, avec un délai moyen de 6,62 jours.
  + Le nombre de réponses fournisseurs par jour présente une tendance à la hausse jusqu'en mai 2023, suivie d'une légère baisse.

Ces informations **suggèrent une corrélation entre le nombre de réponses et l'évaluation des fournisseurs. Elles indiquent aussi des variations dans les délais de réponse selon l'évaluation et également une corrélation entre le nombre de réponses et la saisonnalité.**

## Rapport d’exploration des données :

Le rapport d’exploration des données des disponible dans le GitHub projet via le lien suivant : “[02\_edf\_rapport\_’exploration\_données\_v1.0.xlsx](https://github.com/PMDOUGLAS23/DST_SC_CS/blob/main/reports/02_edf_rapport_exploration_donnees_v1.0.xlsx)”

# Etape 2: Pre Processing des données

Nous avons réalisé certaines étapes d’enrichissement de notre jeu de données lors de la première phase d’analyse.

Pour rappel, nous avons **exploré la dimension temporelle** afin d’isoler les jours, mois, années, semaines, week-end et heures des commentaires.

Aussi, nous avons créé de nouvelles variablespour **mesurer la longueur** - en nombre de mots - **du titre et des commentaires.**

Pour le moment nous n’avons pas croisé ce premier jeu de données de TrustedShops avec d’autres sources d’information.

—-----------------------------------

Rappel :

**Étape 1/** **Exploration des données + DataViz’**   
Votre première tâche consistera à définir le contexte et le périmètre du projet : j’attends que vous preniez vraiment le temps de bien comprendre le projet et de vous renseigner au mieux sur les notions que celui-ci va introduire.

* Il faudra ensuite prendre en main et découvrir votre jeu de données et faire une **analyse** presque exhaustive de celui-ci afin de mettre en lumière la **structure**, les **difficultés**et eventuels **biais**du dataset.

Vous pourrez utiliser ce **template**: [Template - Rapport exploration des données](https://docs.google.com/spreadsheets/d/1T7uEuAvwJZ9lVuPD0TBc1paAcSRIy_P-WLz4nCAjqUQ/edit?usp=sharing)  
J’attendrai également au moins **5 représentations graphiques** construites à partir de votre jeu de données, visuelles et surtout **pertinentes.**Pour chacune d’elle j’attendrai :

* Un commentaire **précis**, qui analyse la figure et apporte un avis “métier”.
* Une validation du constat par des manipulations de données, ou un test statistique.

**Etape 2/ Pre-processing et feature engineering : Deadline Vendredi 27/10**

* Suite aux constats établis lors de l'étape précédente, vous devrez **nettoyer**votre jeu de données et effectuer les étapes de **pre-processing**.
* Si besoin, il faudra **transformer**et/ou ajouter des features et **enrichir**votre jeu de données.
* L’objectif à la fin de cette étape est d’avoir un **dataset prêt** pour effectuer des analyses approfondies et/ou une modélisation à l’aide de Machine (et Deep) Learning.

**Rendu 1 : rapport d’exploration, de data visualisation et de pre-processing des données.**  
Après le rendu 1, si ce n’est pas déjà fait, votre mentor instanciera un**repo GitHub pour votre groupe**, celui-ci devra **respecter le template fourni et être transmis au jury avec le rapport final**.

# 