

Projet d'étude - Analyse de la Supply Chain :

Analyse et prévision des avis laissés sur les plateformes TrustPilot et TrustedShop à partir des commentaires laissés par les clients/internautes

Cohorte de novembre 2022 - parcours Data Scientist

**Adnane MOUZAOU
François ROUXELIN**

Rapport d'étude n°1 :

**Rapport d'exploration, de data visualisation et de
pré-processing des données**



DataScientest.com

Agrément organisme de formation 11755665975

09 80 80 79 49

2 place de Barcelone, 75016 Paris

Sommaire

Introduction	2
Objectifs	3
Cadre	3
Webscraping et données personnelles	4
Disponibilité des données	4
Utilisation principale des données	5
Web scraping des avis clients	5
Pertinence	6
Pre-processing et feature engineering	7
Exploration du dataset issu du scraping de ShowRoomPrivé	7
Analyse et exploration rapide du dataframe	8
Choix de l'intervalle de dates	8
Retraitement des commentaires	9
Visualisations et Statistiques	10
Répartition de notre variable cible, le nombre d'étoiles	10
Répartition et distribution du nombre d'avis postés	11
Etude de la répartition pour l'année 2015 et 2016	13
Relation entre "star" et les caractéristiques des commentaires	16
Analyse des commentaires	19
Analyse des titres	22

Introduction

La satisfaction client est un élément essentiel pour toute entreprise souhaitant conserver sa clientèle et améliorer sa réputation. Dans ce contexte, l'analyse des commentaires et des avis des clients est devenue une pratique courante pour évaluer la qualité de la supply chain, la conformité des produits/services aux attentes du marché et pour identifier les points d'amélioration nécessaires.

Cependant, l'analyse manuelle de ces commentaires peut être fastidieuse et chronophage. Dans ce projet, nous proposons une approche automatisée pour extraire de l'information de commentaires afin de prédire la satisfaction d'un client. Plus précisément, nous visons à prédire le nombre d'étoiles associé à chaque commentaire à partir de données disponibles sur Trusted Shops et Trustpilot.

Cette étude présente un intérêt à la fois économique, technique et scientifique. Sur le plan économique, les résultats de cette étude peuvent aider les entreprises à mieux comprendre les attentes de leurs clients et à améliorer leur satisfaction. Ils peuvent également contribuer à la réduction des coûts liés à l'analyse manuelle des commentaires. D'un point de vue technique, l'utilisation de techniques de traitement du langage naturel et de machine learning permet de transformer des données non structurées en informations exploitables. Enfin, d'un point de vue scientifique, cette étude permet d'explorer les limites de la prédiction de la satisfaction client à partir de commentaires textuels.

En résumé, l'objectif de ce projet est de proposer une approche automatisée pour prédire la satisfaction client à partir de commentaires textuels, avec des implications importantes pour les entreprises, la technologie et la recherche scientifique.

Objectifs

L'objectif principal de notre étude est de développer un modèle prédictif performant pour évaluer la satisfaction des clients à partir de leurs commentaires. Pour cela, nous avons eu accès à des données de deux sources différentes : les avis vérifiés de Trusted Shops et les avis d'internautes de Trustpilot. Nous avons également enrichi notre base de données grâce au web scraping, en récupérant des données associées à l'entreprise Showroom Privé.

Dans le cadre de notre analyse, nous chercherons à comprendre les variables qui peuvent avoir un impact sur la satisfaction client, en examinant divers aspects des commentaires tels que la taille, la ponctuation, la présence de majuscules, etc. Nous effectuerons également une étape de nettoyage des données, en enlevant les mots sans signification ou redondants tels que les stopwords, et en utilisant la lemmatisation pour normaliser les mots.

En somme, nous chercherons à comprendre les facteurs clés qui contribuent à la satisfaction des clients et à développer un modèle prédictif robuste pour prédire leur note directement à partir de leurs commentaires. Cette étude présente des enjeux économiques, techniques et scientifiques importants, en permettant notamment aux entreprises de mieux comprendre les besoins de leurs clients et d'adapter leur offre en conséquence.

Cadre

Dans le cadre de notre projet, nous avons utilisé deux jeux de données pour atteindre nos objectifs. Le premier jeu de données a été fourni par notre organisme de formation et contenait des informations sur les entreprises ShowRoom et VeePee, récupérées à partir des données des sites TrustedShop et Trustpilot. Ce jeu de données comprenait environ 20 000 lignes.

Ce jeu de données a été enrichi en réalisant un scrapping de données sur le site Trustpilot, spécifiquement sur l'entreprise ShowRoom Privé. Ce procédé nous a permis d'ajouter près de 170 000 avis supplémentaires, et également de récupérer deux nouvelles variables, qui est le nombre d'avis laissés par un utilisateur, que nous avons nommé 'nb_avis', ainsi que le titre de l'avis, que nous avons nommé "Titre". Nous disposons ainsi d'une base de données conséquente pour entraîner notre modèle. Afin de limiter le nombre de doublons, nous avons choisi de ne conserver que la base de données scrappées.

Webscraping et données personnelles

Disponibilité des données

Les données ayant fait l'objet du scraping sont disponibles publiquement sur le site de trustpilot. Concernant le volet réglementation, de manière générale, le scraping de données est un domaine juridique complexe et il est important de prendre en compte les lois et les règles applicables à chaque site web et à chaque pays.

Le règlement général sur la protection des données, ou RGPD, s'applique aux données personnelles. Il s'agit de toute information personnelle identifiable (IPI) qui pourrait être utilisée pour identifier directement ou indirectement une personne physique.

Pour des personnes physiques, voici un aperçu de ce que relate **CNIL** (France) :

- nom, prénom, pseudonyme, date de naissance;
- photos, enregistrements sonores de voix;
- numéro de téléphone fixe ou portable, adresse postale, adresse email;
- adresse IP, identifiant de connexion informatique ou identifiant de cookie;
- empreinte digitale, empreinte rétinienne, etc..
- numéro de plaque d'immatriculation, de sécurité sociale ou de pièce d'identité.

L'identification n'est pas toujours possible à partir d'une seule de ces données personnelles mais peut être réalisée par un croisement de ces dernières.

Dans notre cas, du web scraping, une partie des informations récoltées contiennent le nom et prénom des clients et autres indications (liste ci-dessus). Il est utile et nécessaire de préciser qu'il ne nous est pas possible d'identifier les clients d'une manière directe ou indirecte avec ces seules informations, ni avec complément et croisement avec d'autres informations collectées. Aussi, ces données sont :

- Accessibles et publics sur le web
- Stockées de manière sécurisée et conformément aux meilleures pratiques.
- Ne sont pas vendues ou partagées avec des tiers.
- Non utilisées à des fins de démarchages commerciales
- Ne subissent pas de traitement à des fins de prospection.

Les avis sont affichés sur le site selon le format ci-dessus, nous avons extrait les informations suivantes :

- nom_client : Nom du client,
- note_avis : Note attribuée à l'achat
- date_avis : Date de publication de l'avis
- date_achat : Date d'achat
- nbr_avis : Nombre total des avis postés par le même client
- titre_avis : Titre du commentaire
- texte_avis : Texte du commentaire
- pays : Pays

Résultat du scraping

	ShowRoomPrivé	VeePee
Nombres d'avis récoltés	166 897	4074

Pertinence

Le jeu de données fourni de base comporte plusieurs variables dont le nombre d'étoiles ('star'), notre variable cible, et le commentaire laissé par l'internaute ('Commentaire'), commentaire qu'il conviendra d'étudier pour en déduire la note laissée par l'internaute. D'autres indicateurs, comme la date de publication de l'avis ('date'), le délai entre la date de commande et la date de l'avis ('ecart') ainsi que le nombre d'avis laissés par l'utilisateur ('nb_avis') pourraient nous aider à prédire à la note. Les autres indicateurs caractérisant les commentaires, qui seront créés par la suite, auront une importance primordiale dans les premiers modèles de prédictions que nous développerons. Notre jeu de données étant constitué majoritairement de données issues de ShowRoomPrivé, il pourra être intéressant de tester dans un temps futur la fiabilité du modèle proposé avec les données d'autres entreprises et services.

Pre-processing et feature engineering

Exploration du dataset issu du scraping de ShowRoomPrivé

Ce dataset comporte 168 897 avis récoltés.

il contient 8 colonnes :

- **nom_client** : Nom du client,
- **note_avis** : Note attribuée à l'achat
- **date_avis** : Date de publication de l'avis
- **date_achat** : Date d'achat
- **nbr_avis** : Nombre total des avis postés par le même client
- **titre_avis** : Titre du commentaire
- **texte_avis** : Texte du commentaire
- **pays** : Pays

nom_client	note_avis	date_achat	date_avis	nbr_avis	texte_avis	pays	titre_avis
Agani	1.0	: 26 août 2022	2023-02-27	2.0	Je n'ai jamais reçu ma commande. J'ai écrits à ...	FR	Je n'ai jamais reçu ma commande
CHANTAL SLATKINE	1.0	: 26 février 2023	2023-02-27	2.0	J'ai commandé 2 colliers. L'un est OK l'autre ...	FR	J'ai commandé 2 colliers
LDC	1.0	: 26 février 2023	2023-02-27	2.0	J ai commandé des airpods reconditionnés, dit ...	FR	Très déçue de la dernière commande.
anass jeffal	1.0	: 27 octobre 2022	2023-02-27	1.0	Produit acheté en Septembre 2022 retourné le m...	FR	Produit acheté en Septembre 2022...
aurélien	1.0	: 27 février 2023	2023-02-27	1.0	encore une commande partiellement annulée au d...	FR	encore une commande partiellement...
...

Analyse et exploration rapide du dataframe

Nous avons très peu de valeurs manquantes dans ce dataframe et peu de doublons en comparaison à la taille totale du fichier.

Nous allons procéder à leur suppression. Nous disposons au final pour ce dataset d'une bonne volumétrie et avec des informations complètes. ce qui est très utile et qualitatif pour la suite de notre analyse.

```
[3]: display(df.info())

display(df.isna().sum())

display(print('le nombre de doublons est:', df.duplicated().sum()))
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 168897 entries, 0 to 168896
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   nom_client  168896 non-null  object
1   note_avis   168896 non-null  float64
2   date_achat  168896 non-null  object
3   date_avis   168896 non-null  object
4   nbr_avis    168895 non-null  float64
5   text_avis   168895 non-null  object
6   pays        168895 non-null  object
7   titre_avis  168893 non-null  object
dtypes: float64(2), object(6)
memory usage: 10.3+ MB

None

nom_client    1
note_avis     1
date_achat    1
date_avis     1
nbr_avis      2
text_avis     2
pays          2
titre_avis    4
dtype: int64

le nombre de doublons est: 337
```

Choix de l'intervalle de dates

Les avis récoltés sont postés depuis fin décembre 2014 jusqu'à février 2023 nous allons nous intéresser aux années civiles complètes dans le but de traiter des intervalles de temps similaires.. Donc, nous prendrons les dates du 01 janvier 2015 jusqu'au 31 décembre 2022. Ceci nous facilitera les comparaisons, l'étude des tendances et tenter de repérer des saisonnalités, si existantes, en fonction des années, des saisons, des mois.

Le nouveau dataset comporte alors **165967** entrées × 8 colonnes

Retraitement des commentaires

Avant de procéder à une étape de traitement des commentaires, nous avons créée des nouvelles variable à partir des commentaires, et des titres :

- 'longueur' qui mesure la longueur de chaque commentaire, en nombre de caractères.
- 'nb_mots' qui mesure le nombre de mots dans chaque commentaire.
- 'majuscule' qui mesure le nombre de caractères en majuscules pour chaque commentaire.
- 'ponct' qui mesure le nombre de points d'exclamation et d'interrogation dans chaque commentaire.

Des équivalents à ces 4 nouvelles variables ont été créées pour le titre des commentaires.

Ensuite, nous avons procédé à une étape de prétraitement des données (Commentaire et Titre) en utilisant la librairie NLTK (Natural Language Toolkit) pour nettoyer les données textuelles dans la variable 'Commentaire'. Nous avons appliqué les étapes suivantes : suppression des caractères spéciaux, mise en minuscules, suppression des stopwords (mots très courants qui ne portent pas de sens), et enfin, la lemmatisation (réduction des mots à leur forme de base). Cette étape est importante pour améliorer la qualité de nos données textuelles et faciliter l'extraction des informations importantes pour la prédiction.

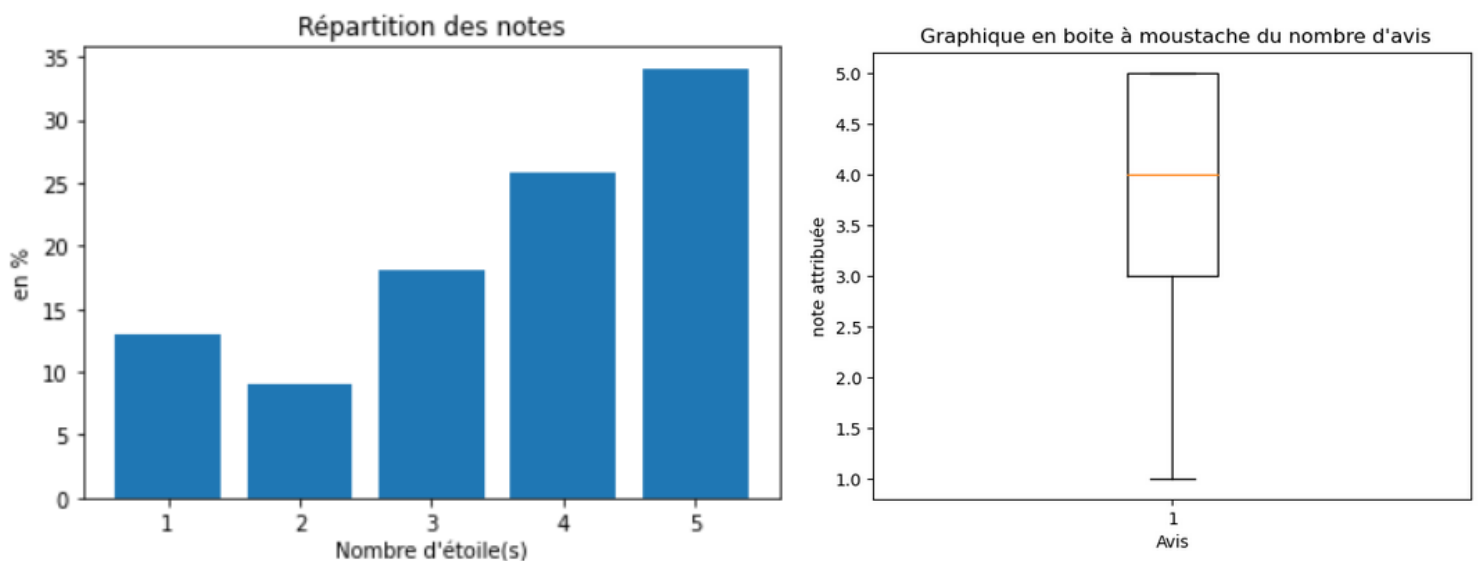
Nous avons également réfléchi à introduire, avant la lemmatisation, une étape de correction orthographique. Notre test s'effectua avec la méthode spellchecker, mais le temps d'exécution était très très long (plusieurs semaines d'exécution estimées sur le jeu de données scrappées), ce malgré l'exécution du code sur plusieurs cœurs de notre processeur en simultané. Bien que pouvant être intéressant, voire important, l'intégration de spellchecker à notre code de nettoyage a été déclinée pour le moment, d'autres solutions sont à l'étude, notamment des tests avec d'autres méthodes, ou l'amélioration de notre puissance de calcul afin de parvenir à réaliser ces corrections dans un temps raisonnable.

Concernant la normalisation ou la standardisation des données, nous n'avons pas jugé nécessaire de le faire pour la variable 'date', 'date' étant une variable temporelle. En fonction des performances des modèles à venir, nous ferons peut-être le choix de transformer la variable 'star' en variable binaire, avec un valeur pour les mauvaises notes, et une valeur pour les bonnes notes. Nous allons en revanche normaliser les variables 'ecart', 'longueur', 'longueur_titre' 'nb_mots', 'nb_mots_titre', 'majuscules', 'majuscules_titre', 'ponct', 'ponct_titre' et 'nb_avis' afin de réduire l'écart-type des variables.

Suite au nettoyage et à la visualisation des mots et ensemble de mots (ngrams) les plus fréquents parmi les commentaires et les titres, de nouvelles variables ont été créées, prenant les valeurs 0 ou 1 suivant la présence ou non des mots ou ensemble de mots sélectionnés, à l'intérieur des commentaires, ou des titres. Ces variables seront très utiles lors des premiers travaux de modélisation et de classification.

Visualisations et Statistiques

Observons les relations entre les différentes variables et la variable cible



Répartition de notre variable cible, le nombre d'étoiles

Selon ces deux graphiques, nous remarquons que les scores à 1* sont plus nombreux que les scores à 2*. ensuite de 2* à 5*, nous observons une certaine linéarité. Ceci est confirmé par le graphique en boîte à moustache et la répartition par quantile.

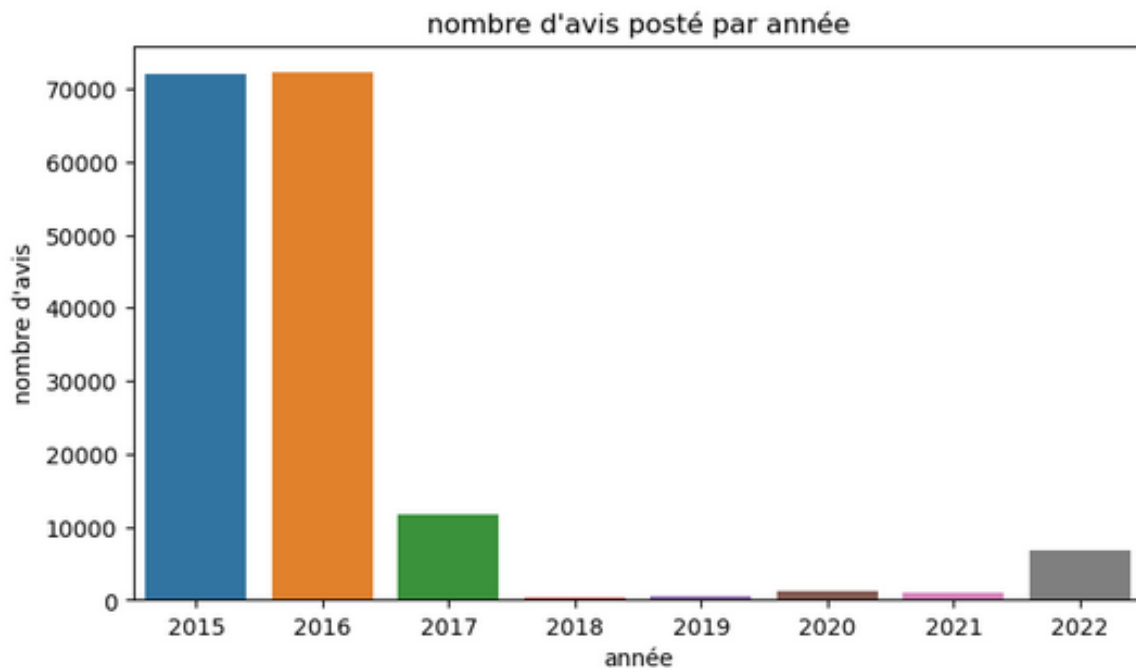
```
df['star'].value_counts(normalize=True)
```

```
5.0    0.341056
4.0    0.258348
3.0    0.180596
1.0    0.128997
2.0    0.091003
```

Répartition et distribution du nombre d'avis postés

Une première observation concerne le nombre d'avis publiés par an. Il est clairement visible que les années 2015 et 2016 ont connu de fortes publications de la part des clients. elles se distinguent des années suivantes.

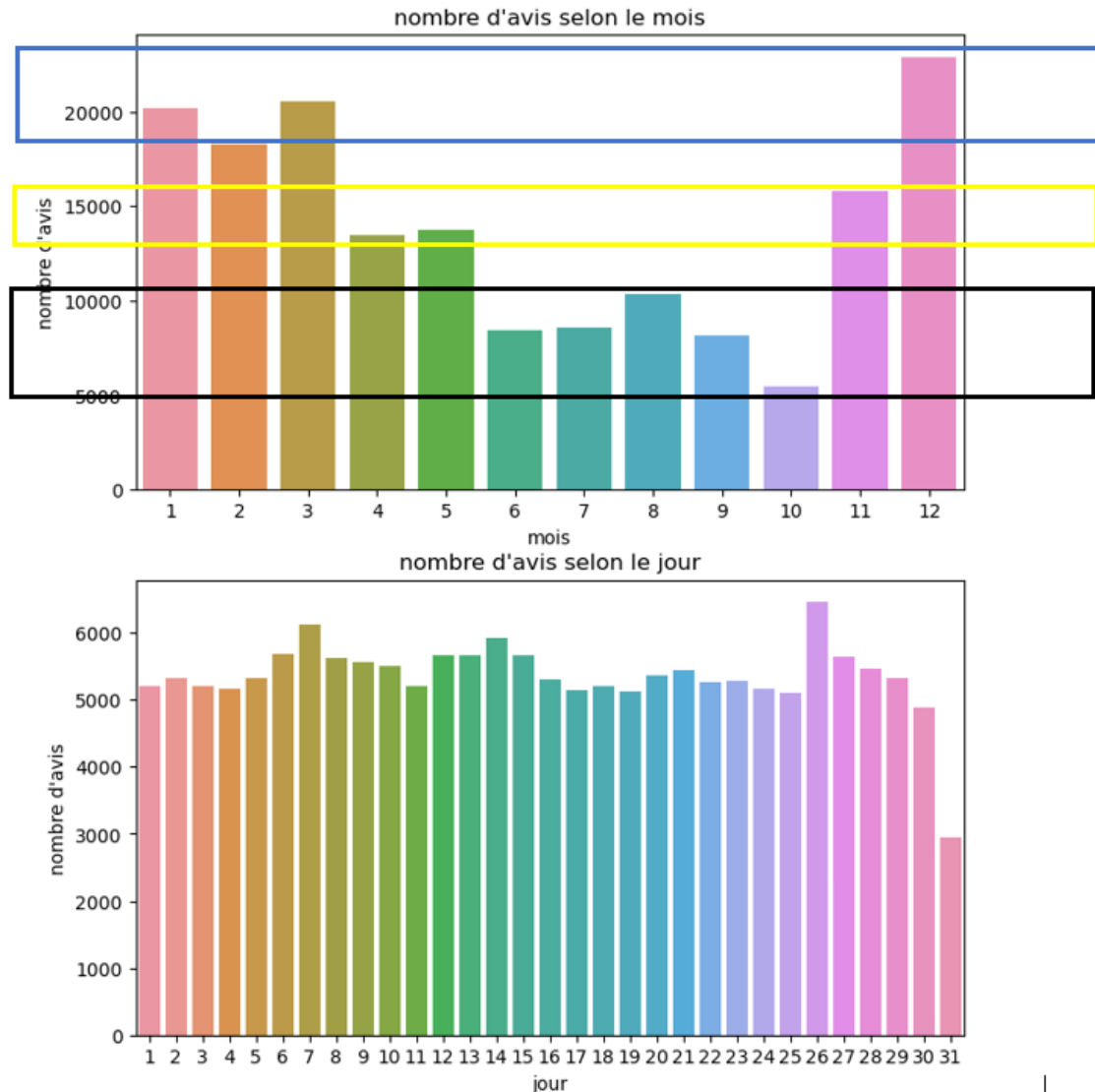
pour les autres années, le nombre d'avis ne semble pas suivre une courbe ou une tendance, il est assez dispersé.



```
df['date_avis'].dt.year.value_counts()
```

```
2016    72261
2015    71925
2017    11727
2022     6843
2020     1212
2021     1089
2019       573
2018       337
Name: date_avis, dtype: int64
```

Graphiques des répartitions du nombre d'avis



La répartition des avis selon les mois montre globalement 3 périodes :

- Période de forte activité : On compte le plus de publications par mois et concerne les mois de janvier, février, mars et décembre
- Période de moyenne activité : cela concerne les mois de avril, mai et novembre
- Période de moindre activité : en juin, juillet, août, septembre et octobre

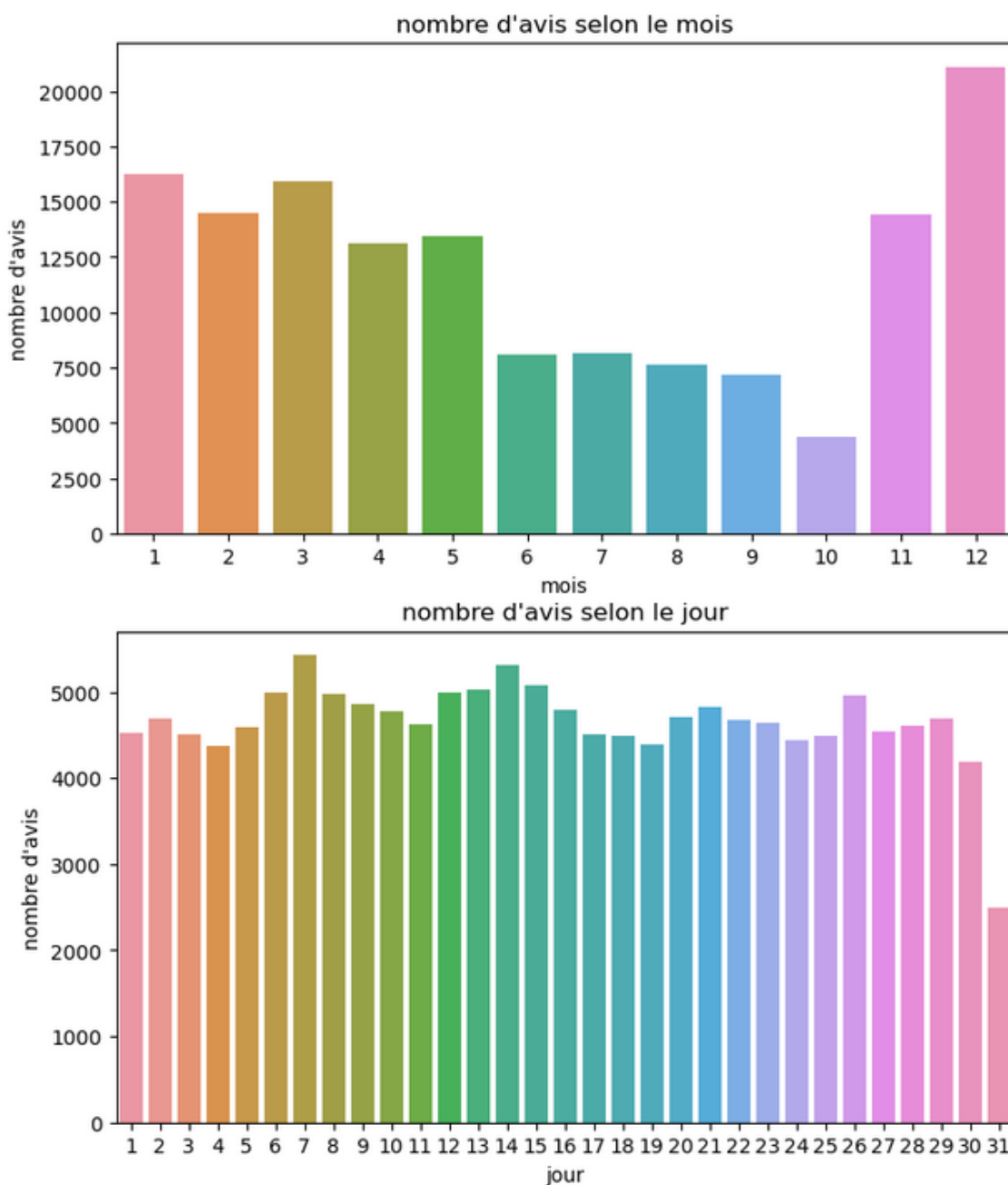
Au cours d'un mois, la distribution du nombre d'avis ne semble pas dégager une tendance particulière. Sur le graphique nous constatons que le jour numéro 31 est moins élevé que les autres, ceci est dû au fait que seuls 7 mois de l'année sont comptabilisés en jour 31.

Etude de la répartition pour l'année 2015 et 2016

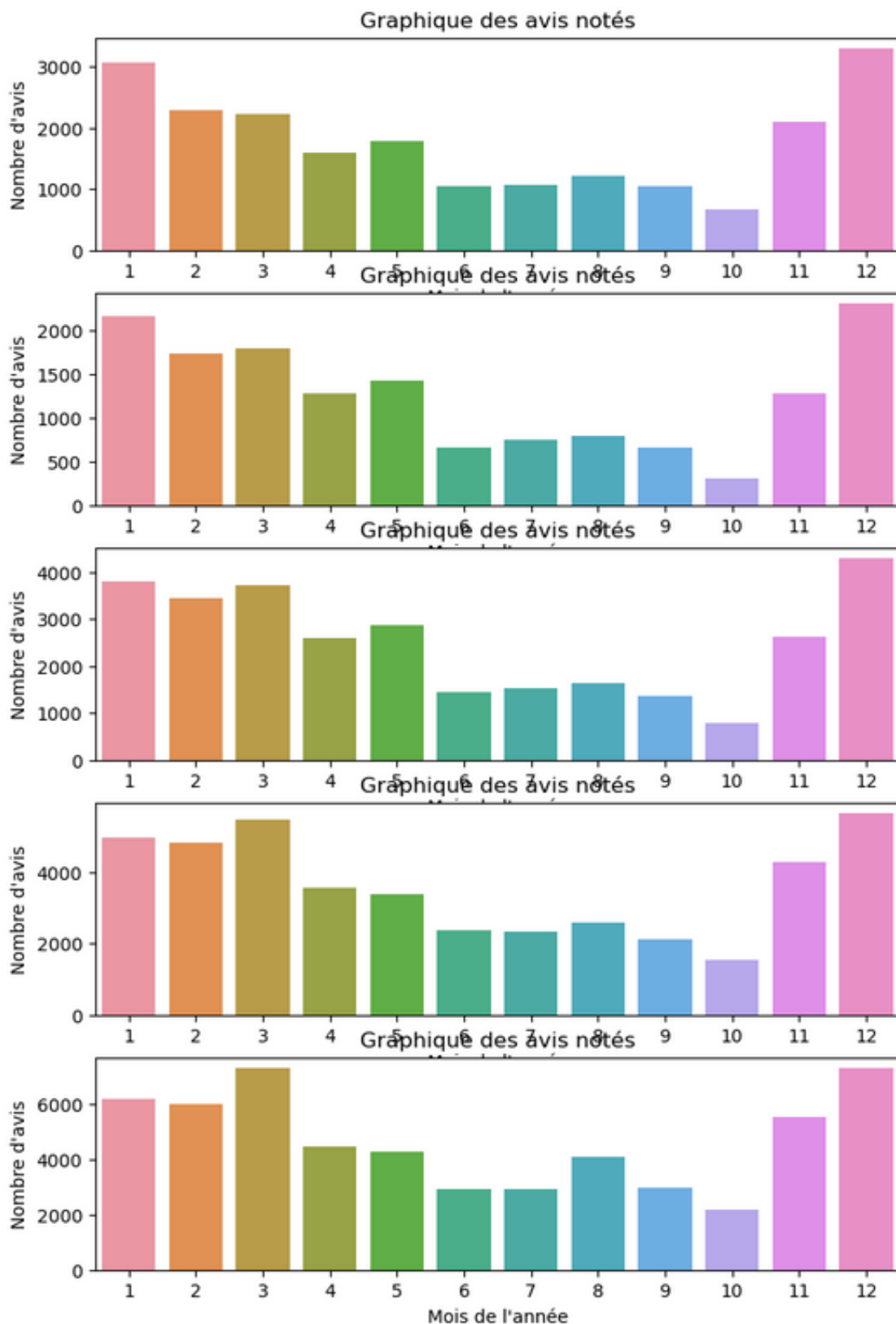
L'année 2015 et 2016 sont marquées par un nombre très élevé d'avis postés, 144 186 avis sur un total de **165967**, ce qui représente plus de 86% des entrées. Nous nous posons la question sur le poids de ces années dans notre analyse et future modélisation. Dans le graphique suivant, nous nous intéressons à la répartition des avis pour 2015, 2016.

Observation :

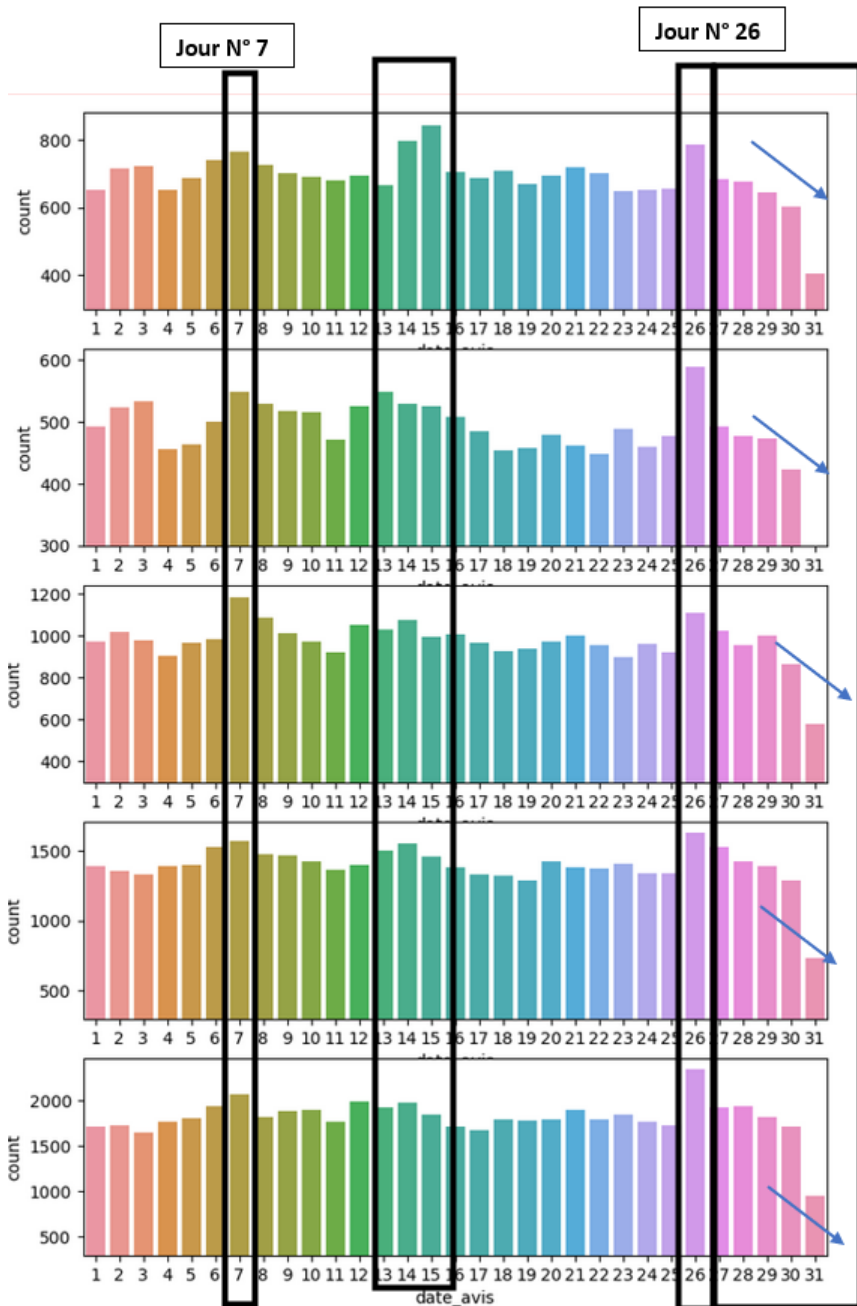
Nous remarquons à peu de chose près, que le nombre d'avis par mois et par jour suit le même graphique que précédemment.



Graphique de la répartition des avis de 1* à 5* selon les mois



Graphique de la répartition des avis de 1* à 5* selon les jours du mois



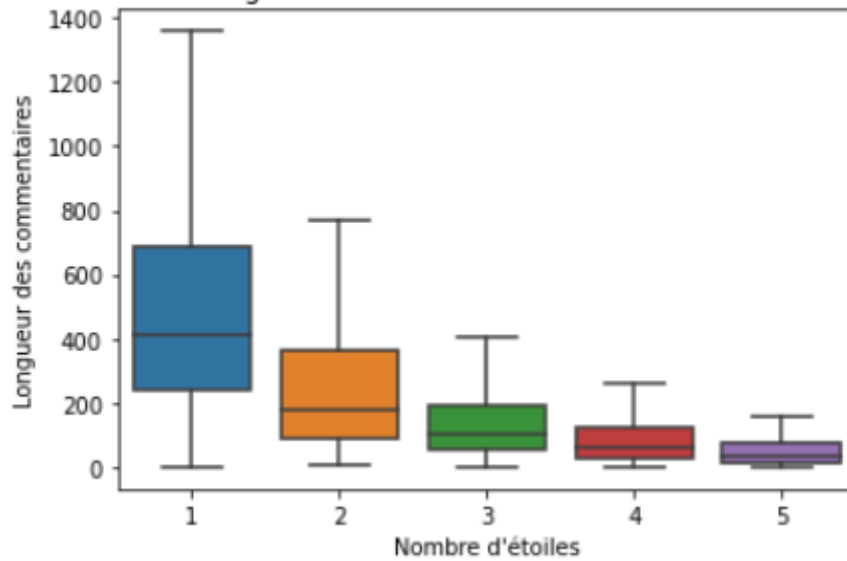
Visuellement, les avis allant de 1 à 5 suivent la même évolution tout au long des jours du mois. On observe :

- Une légère hausse des avis jusqu'au jour 7 qui détermine un pic (à explorer) suivi d'une légère baisse jusqu'à la mi de chaque mois.
- Une reprise des publication entre les 12-13-14-15 environ de chaque mois
- Un pic au jour 26 (à explorer) suivi d'une baisse du nombre d'avis

Remarque: Il est possible que ces pics apparaissent suite à des vagues de rappels de la part des entreprises et/ou de la plateforme à leurs clients de laisser des commentaires.

Relation entre “star” et les caractéristiques des commentaires

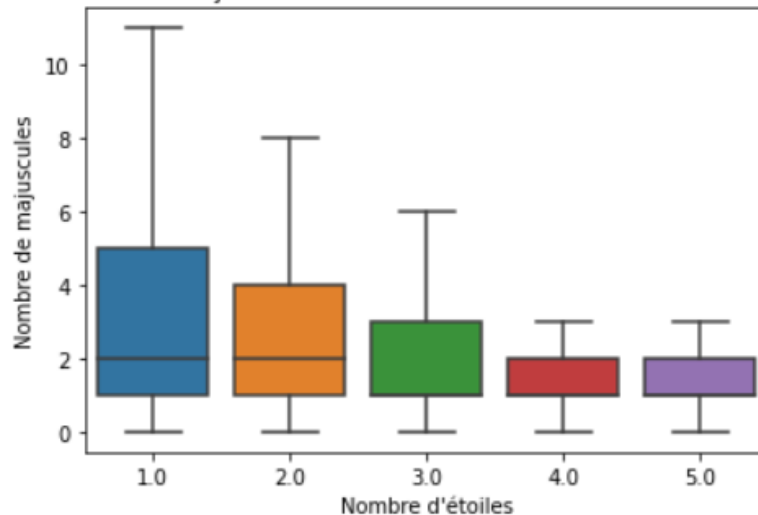
Distribution de la longueur des commentaires en fonction du nombre d'étoiles



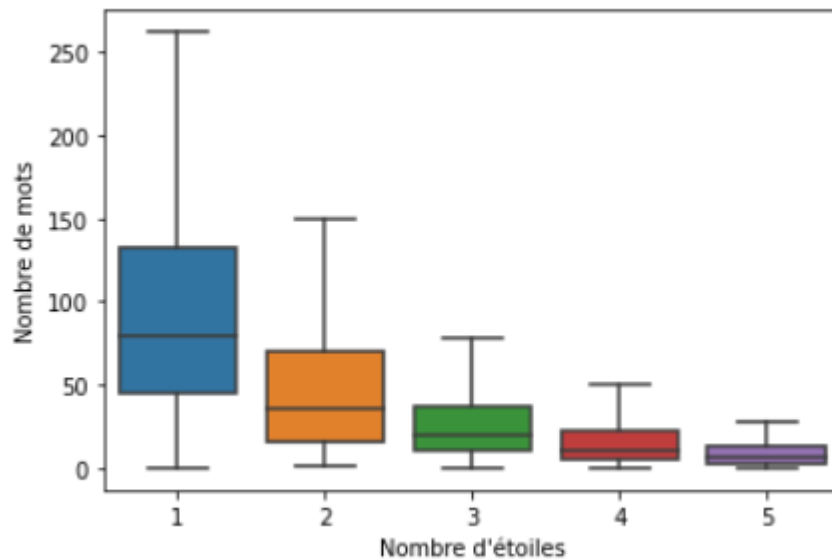
A l'analyse visuelle du dataset, nous avons identifié que plus le commentaire est long, plus il y a de chances que la note associée soit négative. Ce qui se vérifie par le coefficient de corrélation négatif plus bas.

Nous remarquons les mêmes phénomènes concernant le nombre de mots (variables très liées à la longueur des commentaires), les nombres de majuscules et les certains caractères de ponctuation.

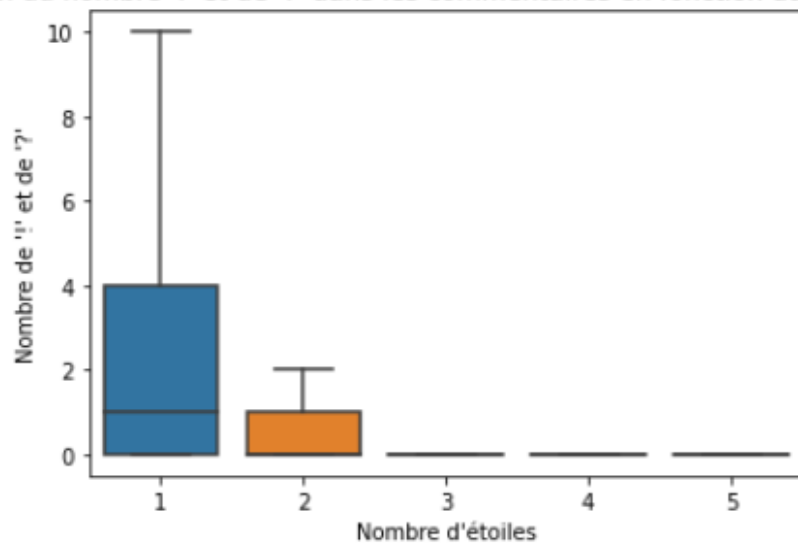
Distribution du nombre de majuscules dans les commentaires en fonction du nombre d'étoiles



Distribution du nombre de mots dans les commentaires en fonction du nombre d'étoiles



Distribution du nombre de '!' et de '?' dans les commentaires en fonction du nombre d'étoiles

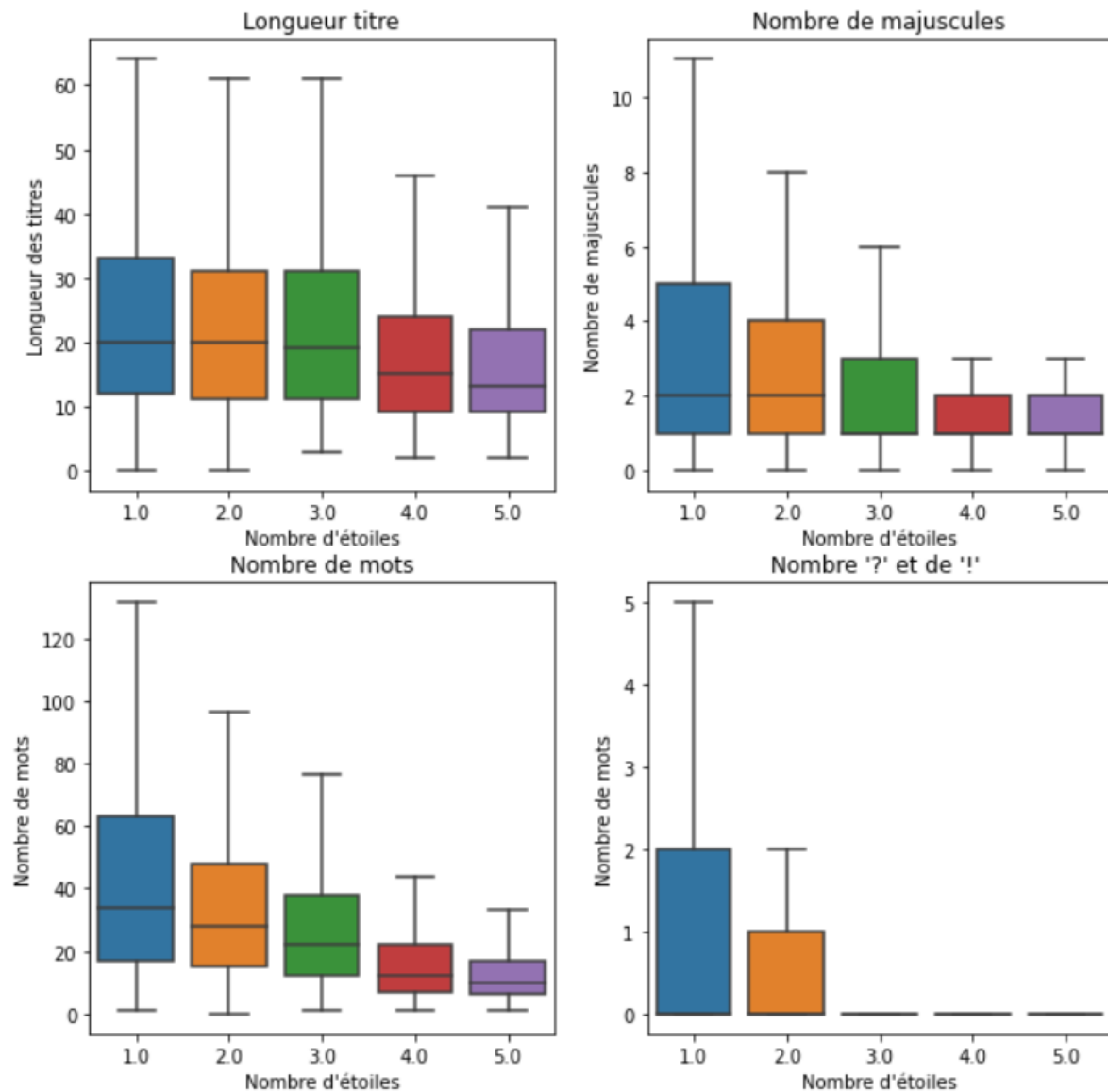


Matrice de corrélation :

```
1 df[["star", "longueur", "majuscule", "ponct", "nb_mots"]].corr()
```

	star	longueur	majuscule	ponct	nb_mots
star	1.000000	-0.405823	-0.103718	-0.208811	-0.415711

Des observations similaires peuvent être faites concernant les titres des commentaires, mais de manière moins marquée, sans doute en raison de la limitation du nombre de caractères dans les titres.



Matrice de corrélation :

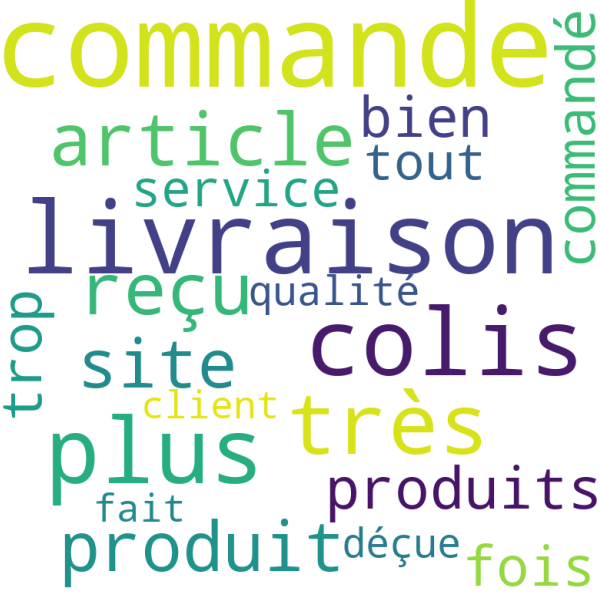

```
1 df[["star", "longueur_titre", "majuscule_titre", "ponct_titre", "nb_mots_titre"]].corr()
```

	star	longueur_titre	majuscule_titre	ponct_titre	nb_mots_titre
star	1.000000	-0.182322	-0.029035	-0.048556	-0.174929

Analyse des commentaires

A l'issue du nettoyage et de la lemmatisation des commentaires, des wordclouds ont été réalisés afin de représenter la fréquence des mots et des ensembles de mots (ngrams) les plus fréquents. Le plus intéressant étant de cibler les mots et ensemble de mots les plus fréquents pour les commentaires/titres négatifs (1, 2 et 3 étoiles), en parallèle avec les mots les plus fréquents des commentaires/titres positifs (4 et 5 étoiles).

WordCloud des commentaires, les mots seuls :

Mauvaises notes		Mot/litération commande 29600 livraison 21155 colis 18265 plus 17652 très 15066 article 11347 reçu 11055 produit 9801 site 9272 produits 9270 bien 8537 trop 8455 fois 8229 commandé 7920 tout 6953 service 6801 qualité 6520 déçue 6383 fait 5935 client 5825
Bonnes notes		Mot/litération livraison 27404 très 27332 commande 26621 bien 17122 site 14954 produits 12376 produit 11536 bon 11171 qualité 10236 tout 9521 délais 9500 prix 8931 satisfaite 8875 peu 7851 conforme 7752 temp 7417 long 7383 colis 7354 article 7238 délai 7108

Les mots seuls donnent une première information, mais on retrouve aussi des mots communs aux bonnes et mauvaises notes. Il est nécessaire d'observer les n-grams pour avoir plus d'informations sur le sens des mots.

WordCloud des commentaires, 2-grams :

Mauvaises notes		<p>Mots/litération</p> <p>('service', 'client') 3737 ('point', 'relais') 3218 ('trop', 'long') 2872 ('délai', 'livraison') 2193 ('très', 'déçue') 2189 ('frais', 'port') 2131 ('délais', 'livraison') 2078 ('livraison', 'trop') 1766 ('cette', 'commande') 1516 ('très', 'déçu') 1384 ('geste', 'commercial') 1344 ('première', 'fois') 1312 ('beaucoup', 'trop') 1237 ('nai', 'reçu') 1131 ('mondial', 'relay') 1113 ('j'avais', 'commandé') 1055 ('mauvaise', 'qualité') 1018 ('plus', 'tard') 951 ('retard', 'livraison') 950 ('reçu', 'colis') 944</p>
Bonnes notes		<p>Mots/litération</p> <p>('très', 'bien') 4793 ('très', 'bon') 3752 ('délai', 'livraison') 3706 ('délais', 'livraison') 3419 ('très', 'satisfaite') 3226 ('bonne', 'qualité') 3146 ('peu', 'long') 3039 ('rien', 'dire') 2776 ('livraison', 'peu') 2472 ('bien', 'passé') 2412 ('date', 'l'expérience') 2204 ('bon', 'site') 2167 ('produit', 'conforme') 2164 ('trop', 'long') 1894 ('bon', 'produit') 1872 ('rien', 'redire') 1858 ('livraison', 'rapide') 1738 ('très', 'contente') 1712 ('tout', 'bien') 1706 ('qualité', 'prix') 1705</p>

Avec les 2-grams, nous pouvons mieux identifier si le commentaire est positif ou non. Par exemple, les mots associés au mot “très” permettent de bien catégoriser les avis. On oppose bien “très satisfaite” à “très déçue” par exemple. Par contre, certains duos restent flous, comme “délai livraison”.

WordCloud des commentaires, 3-grams :



Mauvaises notes		Mots/litération ('livraison', 'trop', 'long') 951 ('beaucoup', 'trop', 'long') 619 ('plus', 'dun', 'mois') 523 ('livraison', 'beaucoup', 'trop') 505 ('délai', 'livraison', 'trop') 442 ('livraison', 'trop', 'longue') 419 ('délais', 'livraison', 'trop') 410 ('nest', 'première', 'fois') 359 ('très', 'mauvaise', 'qualité') 350 ('aucun', 'geste', 'commercial') 335 ('service', 'après', 'vente') 311 ('autre', 'point', 'relais') 274 ('entre', 'commande', 'livraison') 243 ('cette', 'fois', 'ci') 220 ('déçue', 'cette', 'commande') 217 ('colis', 'point', 'relais') 217 ('livraison', 'point', 'relais') 217 ('nai', 'toujours', 'reçu') 216 ('livraison', 'très', 'long') 212 ('jours', 'plus', 'tard') 210
Bonnes notes		Mots/litération ('très', 'bon', 'site') 1246 ('livraison', 'peu', 'long') 1200 ('tout', 'bien', 'passé') 1145 ('rapport', 'qualité', 'prix') 1009 ('très', 'bonne', 'qualité') 988 ('tout', 'très', 'bien') 898 ('très', 'bien', 'passé') 858 ('délai', 'livraison', 'peu') 695 ('délai', 'livraison', 'respecté') 691 ('très', 'bon', 'produit') 686 ('livraison', 'peu', 'longue') 672 ('bon', 'rapport', 'qualité') 636 ('livraison', 'trop', 'long') 636 ('très', 'satisfaite', 'commande') 494 ('délais', 'livraison', 'peu') 493 ('entre', 'commande', 'livraison') 447 ('délais', 'livraison', 'respectés') 413 ('produit', 'conforme', 'description') 390 ('date', 'l'expérience', 'août') 389 ('très', 'bon', 'état') 340

Avec les 3grams, on obtient les détails manquants sur certains termes, notamment “délai livraison”, qui prend la forme “délai livraison respecté” ou “délai livraison trop”...longue.

Analyse des titres



La même analyse peut être réalisée sur les titres des commentaires

WordCloud des titres, les mots seuls :

Mauvaises notes		<p>Mot/Itération</p> <ul style="list-style-type: none"> livraison 10078 commande 7297 très 3669 trop 3659 produit 3434 déçue 3326 colis 3244 non 3058 long 2458 déçu 2376 article 2324 qualité 2178 délai 1883 produits 1848 retard 1753 reçu 1747 bien 1645 mauvaise 1469 erreur 1357 délais 1350
Bonnes notes		<p>Mot/Itération</p> <ul style="list-style-type: none"> très 14117 bien 12814 commande 9688 satisfaite 7806 bon 7643 parfait 7426 site 5948 livraison 5847 super 4713 excellent 3836 tres 3578 produit 3529 rapide 2749 bonne 2697 conforme 2627 qualité 2240 produits 2204 rien 2184 tout 2183 satisfait 1762



Le titre doit être clair et concis, ce qui facilite le repérage de mots récurrents. Ainsi, à l'inverse des commentaires, on retrouve rapidement des mots positifs et négatifs, compréhensifs pour la plupart, tout seul.

WordCloud des titres, 2-grams :

Mauvaises notes		<p>Mots/litération</p> <ul style="list-style-type: none"> ('trop', 'long') 1670 ('livraison', 'trop') 1344 ('délai', 'livraison') 1132 ('délais', 'livraison') 748 ('non', 'conforme') 747 ('très', 'déçue') 713 ('trop', 'longue') 674 ('commande', 'incomplète') 590 ('mauvaise', 'qualité') 578 ('retard', 'livraison') 567 ('commande', 'non') 517 ('livraison', 'non') 469 ('très', 'déçu') 456 ('service', 'client') 453 ('mieux', 'faire') 388 ('frais', 'port') 346 ('point', 'relais') 334 ('problème', 'livraison') 333 ('non', 'respecté') 315 ('bon', 'produit') 299
Bonnes notes		<p>Mots/litération</p> <ul style="list-style-type: none"> ('très', 'bien') 5328 ('bon', 'site') 2543 ('très', 'bon') 2514 ('très', 'satisfaite') 2138 ('tres', 'bien') 1616 ('bon', 'produit') 1393 ('rien', 'dire') 919 ('bien', 'passé') 750 ('rien', 'redire') 683 ('délai', 'livraison') 658 ('très', 'contente') 620 ('tres', 'bon') 605 ('bonne', 'qualité') 582 ('bon', 'service') 573 ('aucun', 'problème') 560 ('qualité', 'prix') 552 ('très', 'bonne') 548 ('tout', 'bien') 507 ('peu', 'long') 504 ('trop', 'long') 499

Comme pour les commentaires, c'est principalement sur les termes "délais livraisons" qu'il est nécessaire d'avoir une analyse de niveau 3

WordCloud des commentaires, 3 grams :

Mauvaises notes		Mots/litération ('livraison', 'trop', 'long') 634 ('livraison', 'trop', 'longue') 458 ('délai', 'livraison', 'trop') 291 ('peut', 'mieux', 'faire') 220 ('délais', 'livraison', 'trop') 207 ('livraison', 'non', 'respecté') 189 ('produit', 'non', 'conforme') 174 ('délai', 'trop', 'long') 146 ('commande', 'non', 'conforme') 138 ('livraison', 'beaucoup', 'trop') 137 ('délai', 'livraison', 'non') 131 ('beaucoup', 'trop', 'long') 128 ('très', 'mauvaise', 'qualité') 125 ('non', 'conforme', 'commande') 109 ('peu', 'mieux', 'faire') 109 ('commande', 'non', 'reçu') 85 ('colis', 'non', 'reçu') 81 ('délais', 'livraison', 'non') 81 ('délai', 'livraison', 'trop') 78 ('livraison', 'non', 'conforme') 73
Bonnes notes		Mots/litération ('très', 'bon', 'site') 1206 ('rapport', 'qualité', 'prix') 417 ('très', 'bon', 'produit') 401 ('tout', 'bien', 'passé') 367 ('bon', 'rapport', 'qualité') 322 ('très', 'bien', 'passé') 268 ('très', 'bon', 'site') 262 ('bon', 'traitement', 'commande') 256 ('tout', 'très', 'bien') 254 ('livraison', 'trop', 'long') 230 ('très', 'bon', 'service') 215 ('livraison', 'peu', 'long') 162 ('très', 'bonne', 'qualité') 158 ('très', 'satisfaite', 'commande') 156 ('commande', 'sans', 'problème') 156 ('livraison', 'trop', 'longue') 144 ('livraison', 'peu', 'longue') 137 ('très', 'bon', 'produit') 116 ('délai', 'livraison', 'trop') 116 ('très', 'bon', 'rapport') 107

Les mots et ensemble de mots les plus présents, marquants, seront retenus pour créer des nouvelles variables qui prendront 0 ou 1 en fonction de leur présence ou non dans les titres et/ou commentaires.