****

NLP Projet 2 :

**Revues d’assurances**

**Authors:** Léo CICAL, Matthieu COLIN

**Instructor:** Christophe RODIGUES

# Etude du dataset

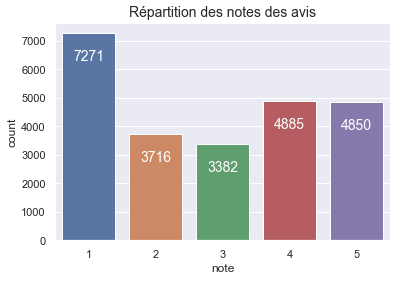
Nos données sont découpées en deux datasets :

* Un dataset train de 24106 lignes et 6 colonnes
* Un dataset test de 10332 lignes et 5 colonnes

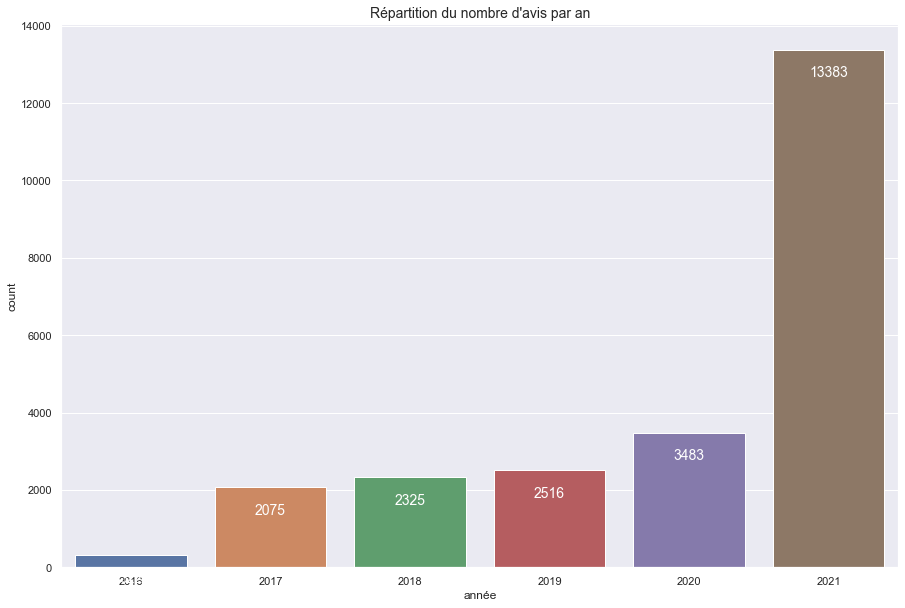
Nous avons des informations sur la date, l’auteur, l’avis, l’assureur et le produit. Pour le dataset de train s’ajoute aussi la note de l’avis.

Nous avons commencé par vérifier la qualité des données. Il nous manque peu d’informations dans chacun des datasets, nous avons donc peu de cleaning à faire sur celles-ci. Il faudra cependant surtout travailler sur la qualité des avis pour améliorer nos résultats des prochaines parties.

Les données en termes de notes ne sont pas parfaitement réparties. Cependant nous considérons en avoir assez dans chaque note pour ne pas à avoir recours à de la data augmentation.

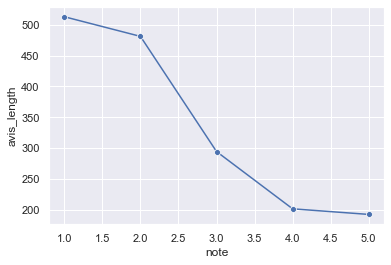


Nous avons reformaté la colonne date, donné en type string, afin de l’avoir en type date. Ainsi nous avons pu remarquer que nos avis sont principalement de 2021.

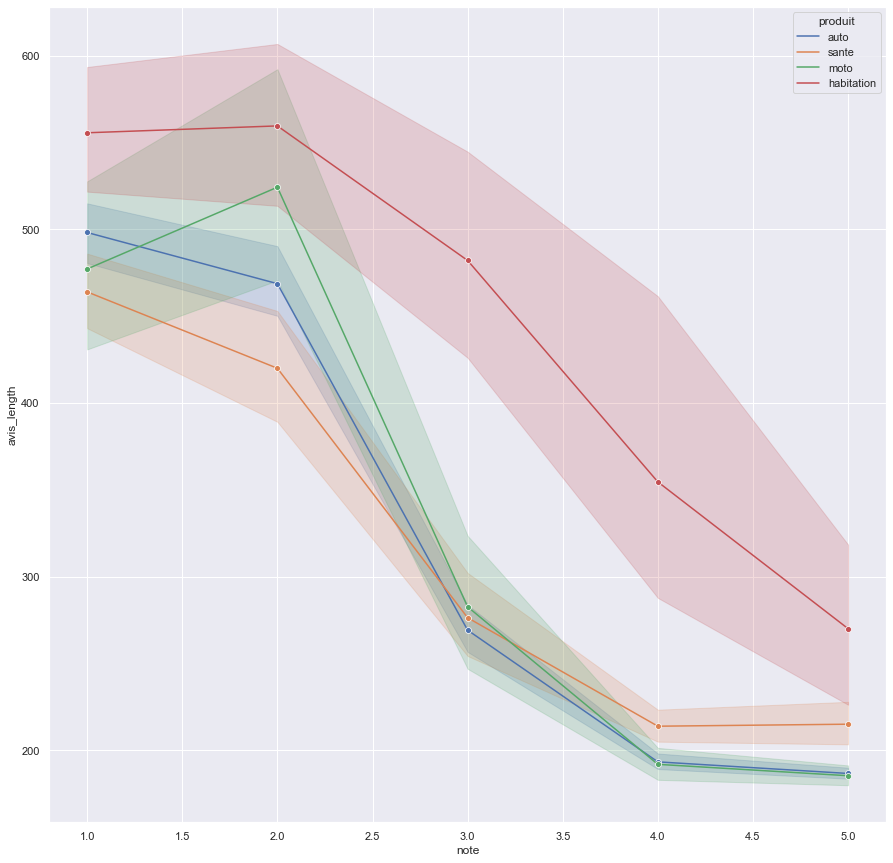


Pour améliorer l’efficacité de nos futurs modèles, nous avons ajouté quelques features à nos données de base :

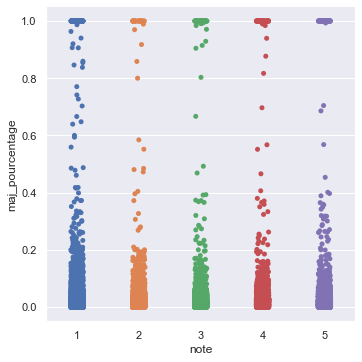
* La longueur de l’avis, après étude nous avons remarqué une grande différence sur la longueur des avis en fonction de la note.



En groupant cette donnée en fonction du produit, on remarque une légère différence pour les avis d’habitations.



* Le pourcentage de majuscules dans un avis. En parcourant les datasets nous avons noté certain avis qui pouvait être beaucoup plus agressif, et ainsi en majuscule. Pour mieux utiliser cette différence nous avons ajouté une colonne donnant le pourcentage de majuscule dans un avis. La différence entre les notes n’est pas flagrante mais le spectre semble cependant mieux représenté pour les avis ayant une note de 1. Cela pourrait s’expliquer par la plus grande représentation des avis ayant une note de 1. Nous décidons quand même de garder l’information.



* Nombre de majuscule maximum de suite dans un avis. Cette information s’ajoute à celle du pourcentage de majuscule. Certain avis mette des majuscules au début des phrases uniquement, cette information pour aider à équilibrer ces phénomènes.
* Nous avons fait le même processus avec les points d’exclamation, récupérant le pourcentage ainsi que le nombre maximum de points d’exclamations à la suite.

# Unsupervised Learning

# Supervised Learning

Notre objectif est de prédire le nombre d’étoile d’un avis à l’aide des autres colonnes. Comme vu précédemment nous avons complété notre dataset avec quelques données supplémentaires afin de guider notre modèle et d’obtenir de meilleurs résultats. Après avoir enlevé les informations qui ne nous sons pas utiles tel que le pseudonyme de l’auteur, voici les données que nous utiliserons :

* Avis : Avis écrit par l’auteur, nous expliquerons dans les prochaines étapes comment nous avons traités ces données
* avis\_length : Longueur de l’avis (en charactères)
* maj\_pourcentage : pourcentage de majuscules
* exclam\_pourcentage : pourcentage de points d’exclamations
* max\_maj\_in\_a\_row : nombre maximum de majuscules à la suite
* max\_exclam\_in\_a\_row : nombre maximum de point d’exclamation à la suite
* produit : information sur le type d’assurance (auto, assurances professionnelles, habitation, moto…)
* assureur : Nom de l’assureur (Allianz, AXA, L’olivier Assurance etc.)