Malik Jan
Mallet Oscar

Projet Big Data

Beer Reviews

1. **Choix de la problématique étudiée**

On s'intéresse ici à savoir si le taux d'alcool d'une bières influence son appréciation. Autrement dit, savoir quel est le taux d'alcool à privilégier lorsqu'on souhaite lancer une nouvelle bière ou améliorer les ventes selon les préférences des consommateurs.

Pour tenter de répondre à cette question (ou au moins en retirer des informations pertinentes le permettant), on a d'abord choisi de s'intéresser aux bières de type *American IPA* (style de bière le plus récurrents de notre base de données). S'intéresser à un style de bière se justifie par le fait que le taux d'alcool est susceptible de varier selon le type de bière (deux styles différents peuvent avoir une différence trop importante pour en faire une étude globale : mieux vaut étudier en préalable un cas : *FIGURE 1*).

*FIGURE 1 – **ABA** entre 4,8% et 5,5% / **AmIPA** entre 6% et 7,5% (en terme de **ABV**)*

Les *American IPA* le plus courantes sont celles ayant un taux d'alcool (qu'on notera dans la suite ***abv*** et exprimé en ***%***) compris entre *5,5%* et *7,5%*. Nous disposons de données pour lesquelles ces bières ont majoritairement un ***abv*** compris entre *6%* et *7,5%* *(FIGURES 2)*.

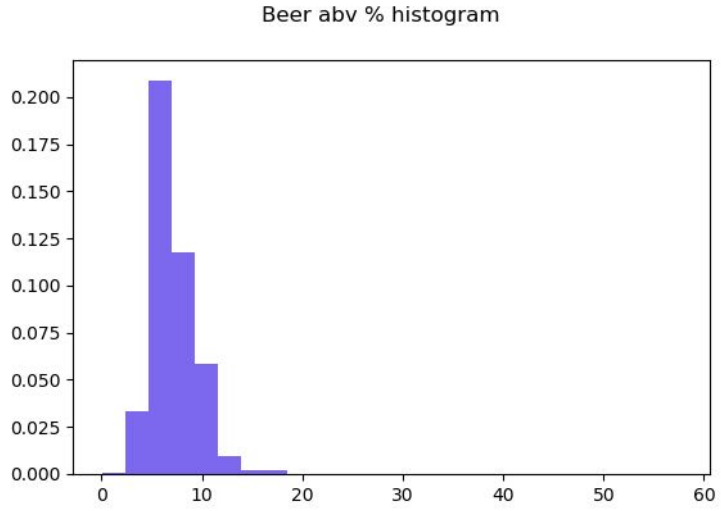

FIGURE 2.1 : Densité de abv de toutes les bières

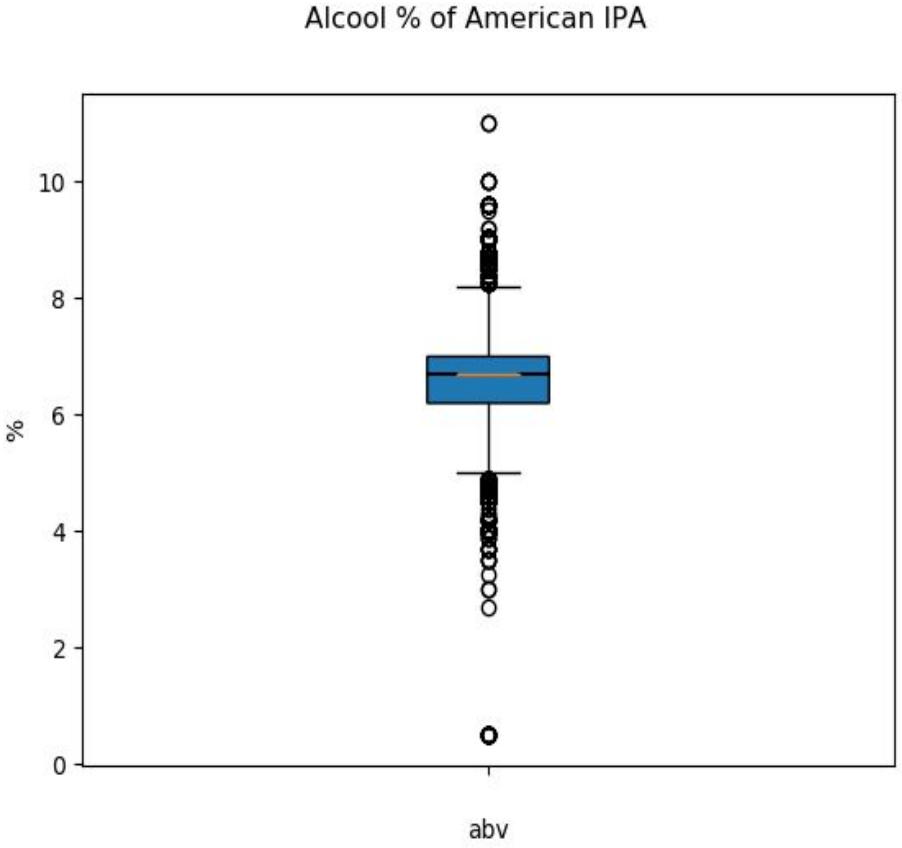

FIGURE 2.2 : Boxplot de abv pour American IPA

Afin de répondre à notre problématique, nous avons besoin de variables explicatives : des descripteur afin de mieux comprendre l'appréciation donnée à une bière. Nous disposons donc de 5 descripteurs (de grandeurs compris entre 0 et 5) : ***'overall'*** *(appréciation globale),* ***'aroma'*** *(appréciation des arômes),* ***'taste'*** *(appréciation du goût),* ***'palate'*** *(appréciation en bouche),* ***'appearance'*** *(appréciation donnée à l'apparence de la bière : packaging, couleur, ect...)*.

Nous allons donc déterminer, en fonction des ces descripteurs, quel est le taux d'alcool à privilégier pour les bières *American IPA* (qu'on notera, ***AmIPA***).

2. **Présentations des données**

Nous avons utilisé la base des données 'Beer Reviews' de Kaggle comprenant plus de 1.5 million de lignes et 13 colonnes (*https://www.kaggle.com/rdoume/beerreviews*). Ce fichier (*csv*) contient plus de 1.5 million de revues sur des bières (données provenant de ***BeerAdvocate*** : *FIGURE 3*).

Pour chaque revue nous avons les informations suivantes (*voir FIGURE 4 et 5*) :

- L'id de la brasserie (qui peut être considéré comme une colonne ***clé***),
- Nom de la brasserie (95 brasseries en total),
- L'id de la bière (qui peut aussi être considéré comme une colonne ***clé***),
- Nom de la bière (56875 bières en total),
- La date d'application de la revue,
- Appréciation générale du consommateur/juge (***note*** de 0 à 5),
- Appréciation de l'arôme du consommateur (***note*** de 0 à 5),
- Appréciation de l'apparence (***note*** allant de 0 à 5),
- Le non du consommateur/juge (33387 juges en total),
- Le style de bières (84 styles de bières en total),
- Appréciation en bouche (***note*** allant de 0 à 5),
- Appréciation du goût (***note*** allant de 0 à 5),
- Taux d'alcool de la bière (***abv*** exprimé en %).

	brewery_id	brewery_name	review_time	review_overall	review_aroma
	1 – 28.0k	Boston Beer Com... 2% Dogfish Head Bre... 2% Other (5741) 95%	841m – 1.33b	0 – 5	1 – 5
1	10325	Vecchio Birraio	1234817823	1.5	2
2	10325	Vecchio Birraio	1235915097	3	2.5
3	10325	Vecchio Birraio	1235916604	3	2.5
4	10325	Vecchio Birraio	1234725145	3	3
5	1075	Caldera Brewing Company	1293735206	4	4.5
6	1075	Caldera Brewing Company	1325524659	3	3.5
7	1075	Caldera Brewing Company	1318991115	3.5	3.5
8	1075	Caldera Brewing Company	1306276018	3	2.5
9	1075	Caldera Brewing Company	1290454503	4	3
10	1075	Caldera Brewing Company	1285632924	4.5	3.5

FIGURE 3 : 10 premières revues et 4 premières colonnes

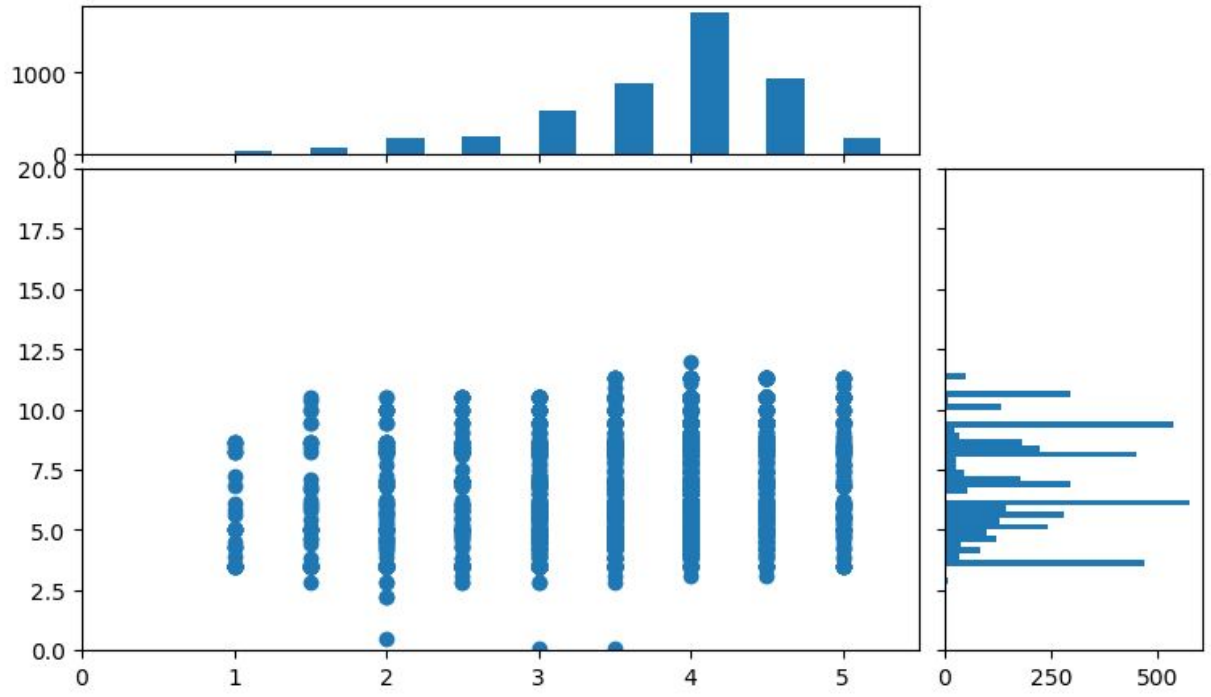

FIGURE 4 : 'review_overall' en abscisse et 'beer_abv' et ordonnée
(graphiques semblables sur les autres descripteurs).
Graphique effectué sur les 50.000 premières lignes.

FIGURE 5 : Boxplot des variables explicatives (descripteurs)
Il en va de même pour tout type de bières séparément.

Travailler sur un jeux de données de cette taille n'est pas évident si on garde le format csv (*format faiblement structuré*). Afin de trouver un compromis entre efficacité et flexibilité nous avons décidé de transformer ce fichier csv en un fichier de données semi-structuré JSON.
La forme choisie du document est la suivante :

doc = { 'beer_beerid' : ,
'beer_name' : ,
'beer_abv' : ,
'beer-style' : ,
'brewery_id' : ,
'beer_name' : ,
'brewery_name' : ,
'review' : [{ 'review_time' : ,
'review_overall' : ,
'review_aroma' : ,
'review_appearance' : ,
'review_profilname' : ,
'review_palate' : ,
'review_taste' : , }]}

Example :

{'_id': ObjectId('5bed9fb10b58051a98853040'),
'beer_abv': 57.7,
'beer_beerid': 73368,
'beer_name': 'Schorschbräu Schorschbock 57%',
'beer_style': 'Eisbock',
'brewery_id': 6513,
'brewery_name': 'Schorschbräu',
'review': [{'review_appearance': 4.0,
'review_aroma': 4.0,
'review_overall': 4.0,
'review_palate': 4.0,
'review_profilename': 'kappldav123',
'review_taste': 3.5,
'review_time': '2011-09-23'}]}

3. **Outils utilisés**

Pour traiter un si gros volume de données il est préférable d'utiliser des outils adaptés. Pour ce faire nous avons choisi un outil de gestion de base de données qui est MongoDB ainsi qu'un outil qui permet de faire du calcul distribué efficacement : Spark avec la méthode Map/Reduce. De plus nous travaillons en Python et sur un serveur distant. Nous avons donc eu recours aux bibliothèques Pymongo et Pyspark.

Enfin, l'écosystème choisi pour travailler en python est Jupyter Notebook qui est directement installé sur le serveur et permet de programmer directement en python sur un navigateur web avec la possibilité d'intercaler des MarkDown.

Accès au code permettant la résolution du problème se trouve dans le répertoire 'Binouze'. Le répertoire est totalement libre d'accès : lecture , écriture , exécution.
Dans ce répertoire se trouvent :

- fichier CSV : *'beer_reviews.csv'* ,
- fichier JSON : *'data.json'* ,
- script *'MongoDB'* (conversion CSV en JSON + étude avec PyMongo),
- script *'Spark'* (étude avec PySpark),
- document PDF

Chemin : [redacted]

4. **Analyse des données**

A) PyMongo

Dans un premier temps nous avons voulu faire des requêtes simples sur notre jeu de données pour en apprendre un peu plus sur ce dernier. Pour ce faire nous avons dû convertir notre jeux de donnée pour pouvoir l'exploiter en MongoDB (*'data.json'*). Initialement on disposait d'une ligne par review. Mais en réalité les objets d'intérêt de ce jeu de données sont bien les bières. Nous avons donc créé un petit programme pour convertir le jeu de données du format csv au format souhaité type Json (*voir script 'MongoDB'*).

Sur ce jeu de donnée nous avons donc commencé par faire de la prospection en faisant des requêtes avec PyMongo. Nous avons choisi de nous intéresser aux bières de style *American IPA* et nous avons donc regardé, dans un premier temps, le nombre de bières différentes de ce style là (36 110 bières *American IPA*).

Dans un deuxième temps nous nous sommes intéressé aux notes des bières sur les différents critères, on a donc fait des requêtes pour trouver le nombre de bières qui obtiennent la note maximale dans les différentes catégories. De même, on a regardé comment les bières se répartissaient selon leur taux d'alcool, on a vu notamment qu'il y a une majorité des bières dont le taux d'alcool se trouve entre 6.5 et 7 % (*cf. FIGURE 1*).

Comme on s'intéresse à l'effet du taux d'alcool sur l'appréciation de la bière nous avons regardé les

valeurs extrêmes, et on voit que la bière la plus forte du lot est à 11% d'alcool et elle a des notes plutôt dans la moyenne : ça n'est donc pas une bière particulièrement appréciée ni dépréciée.
De même, on a regardé la bière la moins forte (0.5% d'alcool) et elle a également une moyenne autour de 3.5. On conclut qu'on ne remarque pas de pattern directement en regardant les extrêmes.

Pour voir ce qui se passe entre les extrêmes on a choisi de découper le taux d'alcool en plusieurs intervalles et compter sur chaque intervalle le nombre de bières qui ont des des notes maximales sur toutes les reviews. Et pour comparer les intervalles de taux d'alcool on a créé un indice qui est le rapport entre :

- le pourcentage de bières parfaite sur l'intervalle ciblé
- le pourcentage de bière parfaite en tout

Donc un indice supérieur à 1 sur un intervalle indique que c'est un intervalle sur lequel les bières sont plus appréciées que la moyenne.
On note un pique d'appréciation de la bière pour un taux d'alcool autour de 7%, et un autre pique à 10 % mais ce dernier est probablement à prendre avec des pincettes car on a peu de données sur cet intervalle (54 bières dont 8 contient au moins 1 note maximale sur chaque catégorie).

Passons maintenant à l'analyse un peu plus complexe et en détail avec Spark, qui est un outil plus adapté et rapide sur des large dataset. en travaillant sur un RDD (*collection distribuée d'éléments*).

B) PySpark

Sur Spark on a pu regarder plus en détail les reviews pour chaque bière en faisant des moyennes. Les bières n'ont pas été noté un même nombre de fois et donc il est intéressant de faire des moyennes pour les comparer.

Nous avons effectué deux algorithme de type ***MapReduce*** :

- Moyenne des notes (de chaque revue) de chaque bière *American IPA* pour la sélection des meilleures moyenne.

- Moyenne du taux d'alcool selon les moyennes pour la sélection (décroissante) des bières *American IPA* les mieux notées.

L'algorithme ***MapReduce*** permettant d'afficher la moyenne du taux d'alcool en fonction de la moyenne des notes nous montre bien que ce premier ne permet pas de comprendre les préférences des consommateurs. En effet les *American IPA* ayant eu que des notes de 5 (sur chaque review et pour chaque descripteur) ont un taux d'alcool moyen de 6.75% alors que les bières *American IPA* ayant une note moyenne de 4.8 ont pour taux d'alcool moyen de 6.76% : une différence non significative. De même ceux avec une note moyenne de 1.5 ont pour taux d'alcool moyen 6.7%. Seul le taux d'alcool ne permet pas de définir l'appréciation d'une bière de type *American IPA.*

5. **Conclusion**

Lors de cette étude nous cherchions à savoir si l'appréciation d'une bière pouvait être liée au taux d'alcool dans celle-ci. Par soucis de cohérence on a ciblé l'étude sur un seul type de bière les *American IPA*. On a pu étudier la répartition des différentes notes attribuées aux bières en fonction des taux d'alcool. On a observé que les bières étaient en moyenne plus appréciées pour des taux d'alcool médians (entre 6 et 8 %). Donc globalement les gens ont plus de mal avec les bières dont les taux d'alcool sont "extrêmes" . On aurait aimé aller plus loin et eventuellement faire une ACP pour voir si on pouvait remarquer des liens entre les différents type de notes. On aurait aussi pu faire des groupes de juges pour découvrir si certaines catégories de personnes ont plus de chance d'aimer tel ou tel type de bière.

Nous avons choisi d'étudier ce jeu de donnée pour appliquer les technologies Big Data car il contenait un grand nombre de ligne et permet de répondre à une problématique intéressante. Bien évidemment sa taille ne justifie pas l'usage des technologies Big Data mais ça nous a permis de se faire la main.