EPFL

Data Visualisation - Spring 2023

Course code : COM-480, credits : 4

MILESTONE 3

Students :
Nicolas TERMOTE
Michel MORALES

Process Book

-

Exploring the Relationship between Beer Styles, Flavor Profiles, Brewery, and Consumer Reviews

---

Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

1 Introduction

The main goal of our project is to uncover the key factors that contribute to the popularity of beer. In our exploration, we will look into the world of beer styles, flavour profiles, and how these parameters influence consumer preferences.

Our objective is to identify the most popular beer styles among consumers and examine how their flavour profiles differentiate from one another. Secondly, we wanted to visualize how specific flavour profiles create better reviews within each style. Additionally, we will analyse how consumer reviews vary based on the origin of the beer, questioning whether certain regions of the world have better quality for certain styles. Moreover, we will explore how review trends impact the popularity of different beer types, and investigate whether seasonal variations influence consumer preferences.

The driving force behind this project is our desire to provide valuable insights to both beer enthusiasts and industry professionals, enabling them to gain a profound understanding of the beer market and consumer preferences. We firmly believe that our findings will prove indispensable to brewery owners, beer distributors, beer critics, and curious consumers alike, all seeking to expand their knowledge on diverse beer styles and flavour profiles.

Within the pages of this process book, we will detail the different visualisation methods we used to provide information about this subject.

One important point to highlight is that, unfortunately, one of our team members had to withdraw from the course due to health-related issues, less than a week before the project deadline. This unexpected development posed a challenge as the remaining team members had to quickly take over his responsibilities. These specific aspects will be described in the peer assessment section of our report.

2 Research

In the first milestone, our initial challenge was to find a suitable dataset containing all the necessary information for our analysis, including reviews, aromas, and brewery locations. However, due to the specialization of available datasets, we faced difficulties in finding a single dataset that encompassed all the required information.

To overcome this problem, we devised a strategy to merge multiple datasets by matching breweries and beer names. Although this approach resulted in some loss due to non-existent matches, we were able to create two datasets that would serve as the foundation for our analysis.

The first dataset contains information about the reviews, encompassing a total of 9,072,914 entries. It includes columns such as 'beer_id', 'username', 'date', 'text', 'look', 'smell', 'taste', 'feel', 'overall', 'score', 'id', 'name', 'brewery_id', 'state', 'country', 'style', 'availability', 'abv', 'notes', 'retired', and 'meta_style'.

The second dataset focuses on aromas, comprising 4,368 entries. It includes columns such as 'beer_id', 'name_beer', 'brewery_id', 'style', 'availability', 'abv', 'notes_beer', 'retired', 'name_brewery', 'city', 'state', 'country', 'notes_brewery', 'types', 'key', 'Style', 'Style Key', 'Description', 'ABV', 'Ave Rating', 'Min IBU', 'Max IBU', 'Astringency', 'Body', 'Alcohol', 'Bitter', 'Sweet', 'Sour', 'Salty', 'Fruits', 'Hoppy', 'Spices', 'Malty', and 'meta_style'.

Notably, both datasets contain brewery information.

To construct these two datasets, we merged four separate datasets sourced from Kaggle. The first dataset provided 9,073,128 reviews, the second contained aroma information for 5,558 beers, the third encompassed details about 50,347 breweries, and the fourth dataset supplied general information about 358,873 beers.

By combining and curating these datasets, we have established a robust foundation for our analysis, enabling us to explore the relationships between reviews, aromas, and brewery information.

While examining the data, we discovered that there were 113 different beer styles within the review dataset and 111 within the aroma dataset. In order to consolidate and simplify the analysis, we decided to group these styles into meta-styles, which are more commonly recognized and representative of the substyles. To accomplish this, we created a dictionary mapping the meta-style names to the corresponding substyle keys.

Our aggregation choices were primarily based on the collective experience of our team members, who are involved in a beer association and have participated in numerous tastings. For certain specific styles, we conducted online research to determine the appropriate meta-style classification. As a result, we encountered a small number of beers that remained unclassified. These beers were subsequently grouped under the 'other' category.

The final outcome of our meta-style classification process yielded exceptional results, as evidenced by the radial tree visualization on our website, which will be presented later. The distribution of beers across the meta-styles is as follows :

- Ale : 2,369,825
- IPA : 2,115,085
- Stout : 1,643,013
- Lager : 902,957
- Sour : 700,603
- Belgian Blonde : 415,117
- Winter Beer : 218,710
- Other : 214,114
- Boozy : 197,172
- Wheat Beer : 178,940
- Ambree : 80,444
- Smoked Beer : 32,486
- Alcohol-free : 4,448

During our initial phase of analysis, we conducted preliminary examinations and generated graphs, unearthing captivating insights. The most impactful information we found was : the disparities in aromas between the highest and lowest rated beers, along with the variations influenced by seasons and regions. These findings served as the fundamental pillars upon which we constructed our subsequent analysis.

3 Concept Development

After conducting our initial analysis, we began thinking about how to implement and present our key findings in an engaging and interactive manner. While seeking inspiration online, we stumbled upon a fascinating website that analysed wine and adopted a similar approach to what we envisioned. The website's storytelling format, utilizing a scroll sidebar, caught our attention. To replicate this format, we explored various online programs and discovered "scrolly," which served as the core HTML structure for our website.

To quantify the influence of different aromas on beer preferences, we decided to incorporate regression analysis. This enabled us to determine which aromas had a positive or negative impact on each beer style.

In order to create a learning experience for our readers, we structured our storytelling into the following sections :

1. Understanding the optimal aromas for each beer style.
2. Identifying connections and similarities between different beer styles.
3. Exploring the regions that produce the best beers within each style.
4. Analyzing the preferred seasons for each beer style.

To visualize these sections effectively, we chose the following visualizations :

1. Violin plots showcasing aroma distributions for beers of the same rating within a specific beer type. This allowed us to identify aroma trends in the best and worst rated beers and determine if certain aromas were desirable traits. We also added a regression analysis presenting impact coefficients to quantify the effects observed in the violin plots.
2. Initially, we considered using a bubble plot to depict the similarity between styles based on their proximity, with colour representing the primary aroma profile and size representing popularity. However, due to limited information provided this way on similar aromas and the inadequacy of a "general" similarity representation, we opted for a chord diagram instead.
3. Initially, we planned to use a choropleth map to display the average ratings of each region for each beer style. However, as we also wanted to incorporate production information, we decided to employ a bivariate choropleth map. We also decide to only present the regions of the United States and of Europe, as our dataset contained information mainly in these regions.
4. For our radial point plot, we maintained the original structure, representing 52 weeks of the year with points varying in size based on the average ratings during each week. This visualization aimed to reveal any seasonal variations in how people rate different beer styles.

Finally, we decided to include a graphical tree representing beer styles and their corresponding meta-styles. This addition was intended to help readers better understand the relationships between different beers within the same meta-style, considering that many beer styles are unfamiliar to most.

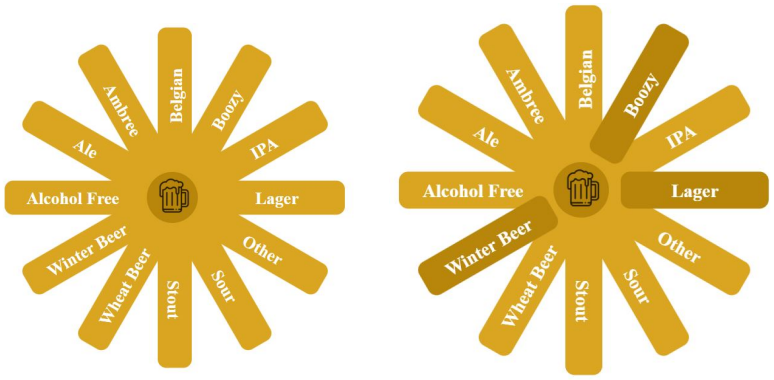
By utilizing these visualizations and storytelling elements, we aimed to provide an engaging and informative experience for our readers, facilitating their understanding of beer styles, aromas, and their associations.

4 Design Process

We wanted to give our website a beer related theme, and came up with a colour palette that reminded us of the common colours for beer, we ended up mainly using the following colours :

(a) Papayawhip (b) Goldenrod (c) Darkgoldenrod

For the style selection feature, we aimed to create an attractive and interactive interface. Thus, we implemented a radial list that appears when hovering over a beer button, as shown in Figure 2.a. The design of this feature was visually appealing and intuitive.



(a) Radial style chooser (b) Selected style chooser

(c) Warning message (d) Selected styles labels

FIGURE 2 – Design of the style chooser elements.

To visualize the chosen beer styles, we added labels to certain graphs where multiple styles needed to be selected. Multiple labels were added (Figure 2.d), and the colour of the list changed to indicate the chosen styles, as demonstrated in Figure 2.b. Additionally, to limit the number of beers that could be selected, we incorporated a warning message that appears when too many beers are chosen. An example of this warning message is presented in Figure 2.c.

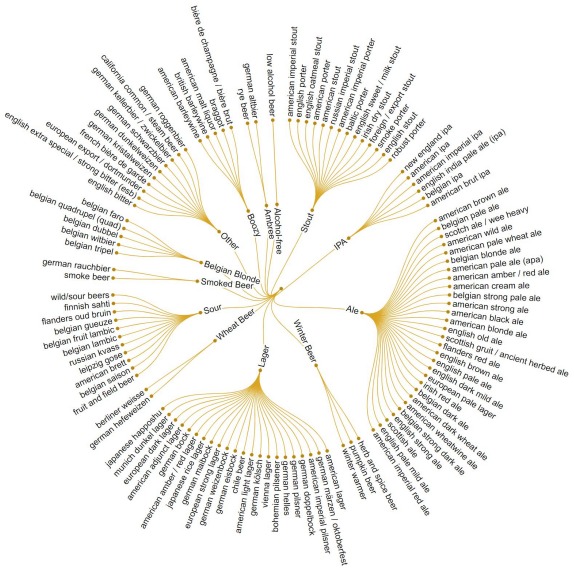


FIGURE 3 – Beer style tree

FIGURE 4 – Color palette of the chord diagram

For the beer style tree, we opted for a radial graph that effectively displayed the relationship between each style and its corresponding meta-style. The final design of the beer style tree can be observed in Figure 3. We aimed for an aesthetically pleasing presentation and found that the linked tree format surpassed our initial tabular design.

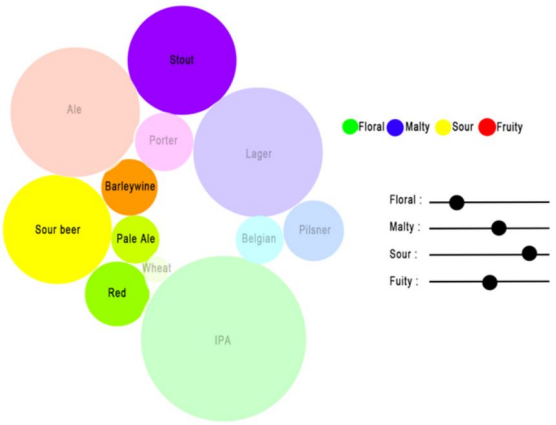


FIGURE 5 – Bubble diagram sketch

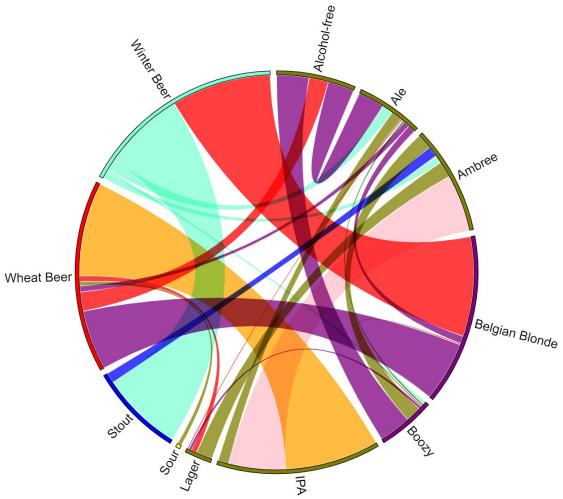


FIGURE 6 – Final chord diagram

To ensure optimal visibility and distinguishability of aromas, we carefully selected a colour palette, which can be seen in Figure 4. This palette was applied to the choropleth map to visualize aroma similarities between styles. Moreover, in order to provide information about the main aromas of each style, similar to the previous bubble graph sketch shown in Figure 5, we included the main aromatic colour of each style within the respective sections of the perimeter, as illustrated in Figure 6.

Regarding the violin plot, our initial intention was to present multiple plots side by side, as demonstrated in Figure 7. However, due to the clustered nature of the plot data, we ultimately opted for a single violin plot, as shown in Figure 8.

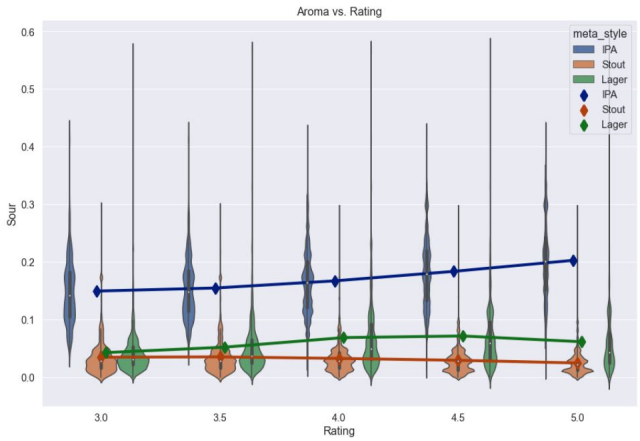


FIGURE 7 – Violin plot sketch

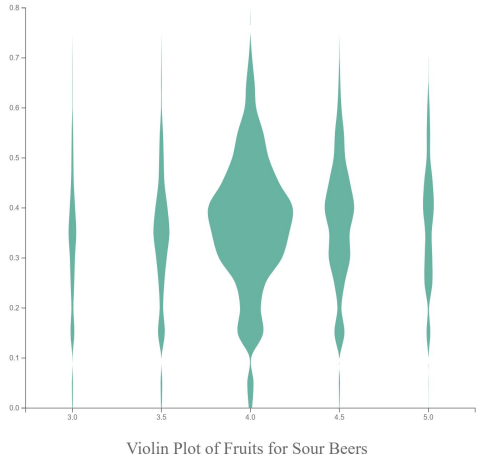
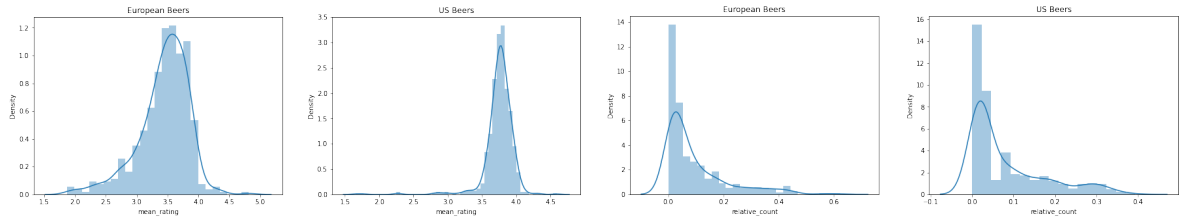


FIGURE 8 – Final violin plot

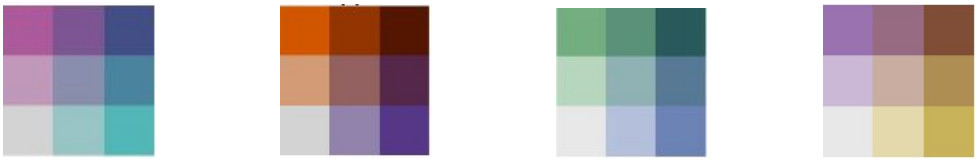
To enhance the presentation of regression analysis, we utilized a bar chart instead of relying solely on textual information. This allowed for a visual representation of the effect of each aroma on the average grade of each substyle.



(a) Average region rating distribution (b) Relative regional production distribution

FIGURE 9 – Value distribution for the bivariate map legend

For the bivariate choropleth, we discovered that the regional differences for certain styles were not particularly pronounced. To improve the visualization, we incorporated legends of different sizes, as exemplified in Figure 13. Additionally, we spent considerable time selecting the colour palettes, considering the options showcased in Figure 10. Ultimately, we found that option (a) provided the most readability, and we employed it for our choropleth map. The different ranges on the map were determined by examining the value distributions depicted in Figure 9. Our chosen distribution offered clear distinctions between the maps, meeting our requirements and expectations as seen in the final map depicted in Figure 12.



(a) purple-cyan (b) orange-purple (c) green-blue (d) purple-gold

FIGURE 10 – Different colour palettes considered

FIGURE 11 – Initial univariate map sketch

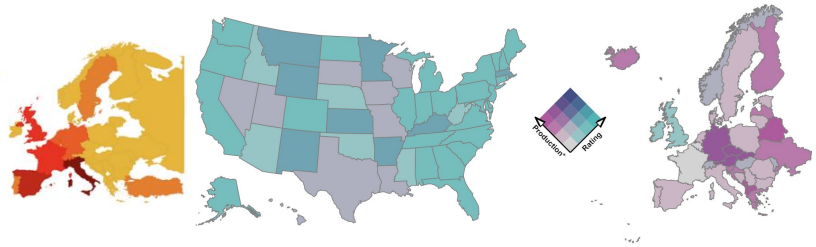


FIGURE 12 – Final bivariate Map design

(a) 3*3 (b) 4*4 (c) 5*5

FIGURE 13 – Different legend sizes

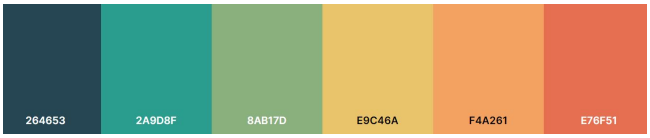


FIGURE 14 – Color palette of radial seasonal graph

For the final radial seasonal plot, we remained true to our initial sketch while incorporating a colour gradient in the outer edges to enhance visual impact and highlight seasonal variations. The initial sketch is presented in Figure 15, while the final implementation can be observed in Figure 16, with the categorical colour palette of Figure 14 that we found visually pleasing.

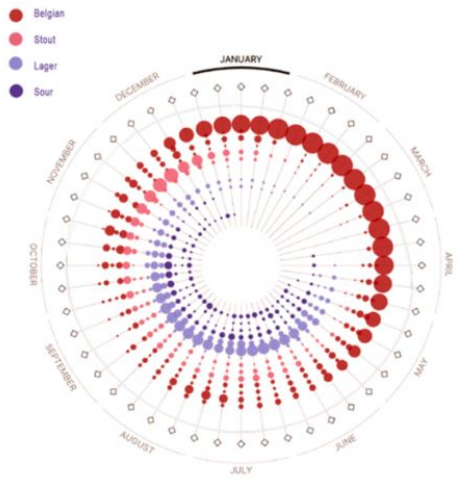


FIGURE 15 – Radial seasonal sketch

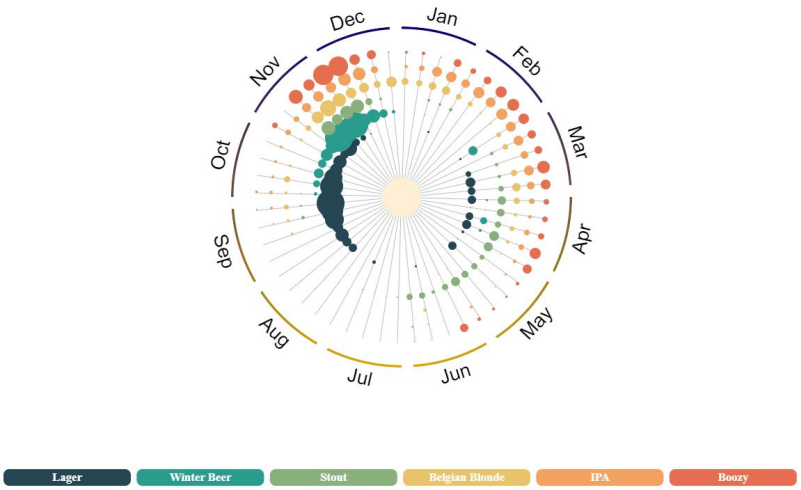


FIGURE 16 – Final radial seasonal plot

5 Conclusion

In summary, our project successfully delivers a captivating and interactive visual narrative that unveils the origins of beer and the interplay between aromas and styles. Our website offers a seamless and engaging user experience, allowing visitors to delve into the captivating world of beer. We invite you to explore our website and immerse yourself in the enthralling visual journey we have crafted.

Cheers to a better understanding of the factors that shape beer popularity !

6 Peer assessment

As previously mentioned, our group experienced an unfortunate setback with the unexpected departure of a member at the last minute. Consequently, the parts originally assigned to them are marked with a (*) symbol in our peer assessment. These sections were completed quickly, and we apologize for any limitations in interactivity or depth as a result.

Nicolas TERMOTE assignments :

- Data formatting for : choropleth, regression, chord diagram and style tree (docs/preprocessing/preprocessing.ipynb)
- Interactive bivariate choropleth (html, js, css)
- *Interactive regression coefficient plot (html, js, css)
- *Beer similarities chord diagram (html, js, css)
- Beer style radial tree (html, js, css)
- Visual design of radial time chart (html, css)
- Interactive and animated "radial beer style" picker and selected beer labels (html, js, css)
- Warning messages integration (html, js, css)
- *Creation, design and writing of the Process book
- *Website design, text (datastory), and graph scrolling with ".hidden" class method.
- Writing of the README
- Screencast

Michel MORALES assignments :

- Data formatting for : Violin plot, radial seasonal chart
- Interactive Violin plot (html, js, css)
- Data integration in radial seasonal chart (js)

Work done by our previous member :

- Website deployment with scrollama (milestone's 2 "initial page")