Actividad 3: Diseño de la visualización

Caso de estudio: Brewer's Friend Recipes

1.- Busque en la web una base de datos sobre la cerveza. Puede tratarse de bases de datos sobre conocimiento general de la cerveza, consumo, comercio, distintos cereales, etc.

La base de datos elegida consiste en las recetas de Brewer's Friend (https://www.brewersfriend.com/homebrew-recipes) que contiene recetas publicadas por los usuarios del sitio y las propiedades de cada una.

2.- Defina 3 preguntas significativas sobre esos datos a resolver con una visualización.

Pregunta 1 ¿Cuáles son los estilos de cerveza más populares en el ámbito de la web de cervercería artesanal?
Pregunta 2 ¿Cuáles son la o las propiedades o características cuantitativas de la cerveza que mejor definen cada tipo o estilo?
Pregunta 3 ¿Es posible clasificar los estilos de acuerdo a estas propiedades?

3.- Investigue si existen soluciones existentes que puedan responder sus preguntas.

En el sitio https://www.kaggle.com/samlac79/beer-recipe-exploratory-analysis se muestra un EDA que presenta las características de los datos pero que no obtiene buenos resultados de clasificación.

En https://github.com/JinheonBaek/Beer-Classification se entrena un clasificador Random Forest que obtiene un 88% de precisión, pero no presenta visualizaciones útiles.

4.- Describa la audiencia para la cual estará destinada la visualización.

Fabricantes artensanales de cerveza, público general.
Ayudar al usuario a seleccionar el estilo adecuado en función de las propiedades del producto resultante que está publicando.

5.- Describa el conjunto de datos, qué atributos tiene, de que tipo son. Necesitará efectuar algún procesamiento adicional de los datos para responder alguna de sus preguntas?

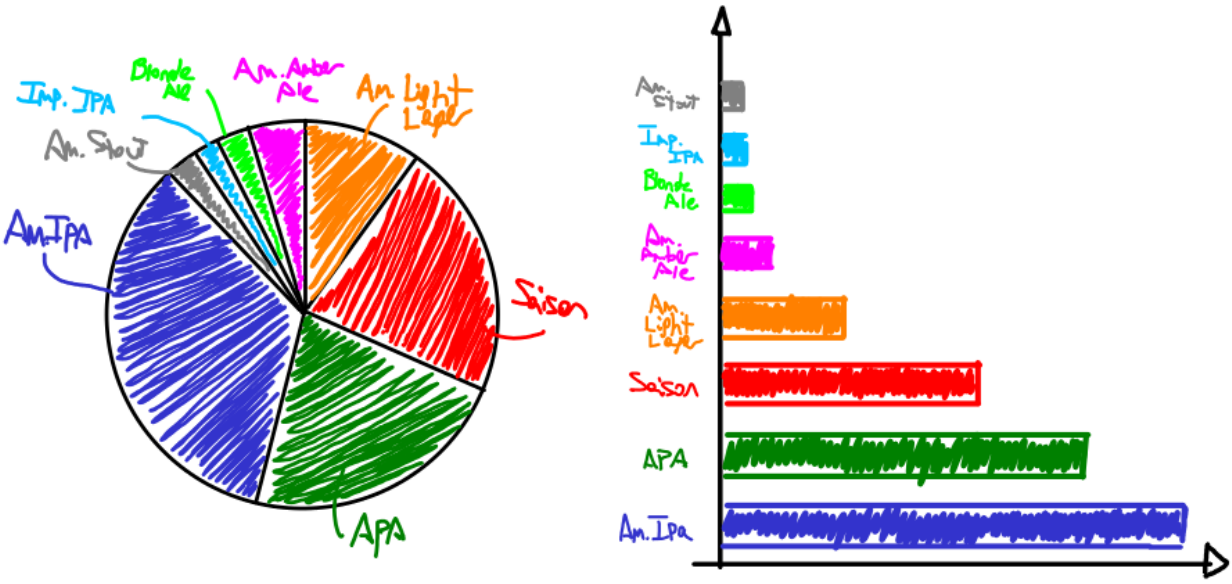
Atributos categóricos:
URL, Name, Style, SugarScale, BrewMethod, PrimingMethod.
Atributos numéricos:
BeerID, StyleID, UserID, Size, OG, FG, ABV, IBU, Color, BoilSize, BoilTime, BoilGravity, Efficiency, MashThickness, PitchRate, PrimingTemp, PrimingAmount.

Si los atributos por sí solos no permiten efectuar una clasificación directa, entonces es necesario definir componentes, por ejemplo, como combinación lineal de las variables originales, para encontrar un espacio bidimensional o tridimensional que sí permita separar en clases de acuerdo estilo correcto.

6.- Realice un sketch en papel describiendo el mapeo visual y las interacciones que necesitará para responder las preguntas planteadas.

Para responder la pregunta 1, se puede emplear un gráfico circular para mostrar las proporciones de los principales estilos, pero el problema es que si el número de clases es muy grande, se dificulta ver aquellas que tienen una pequeña proporción, aunque es posible agrupar las minorías en otra clase aparte y resaltar sólo las más importantes.

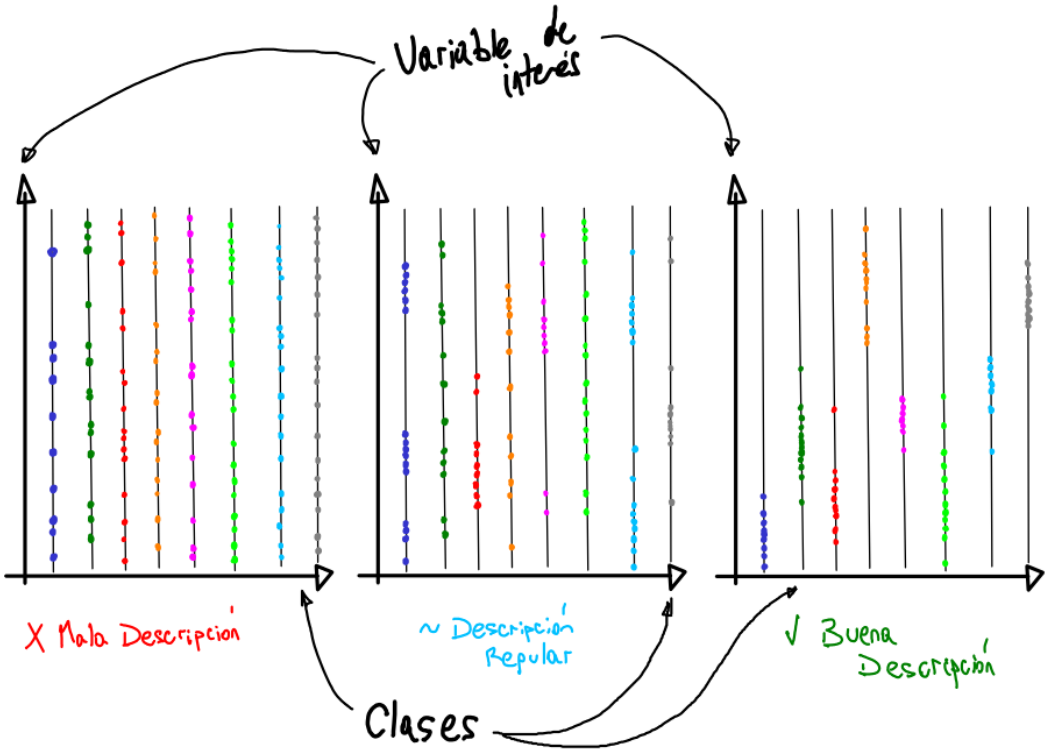
Una mejor opción es mediante un gráfico de barras horizontales o verticales. En este caso se podría agregar una interacción para ocultar la clase más numerosa y cambiar la escala de frecuencia de manera que se puede explorar las proporciones de todo el conjunto de clases.



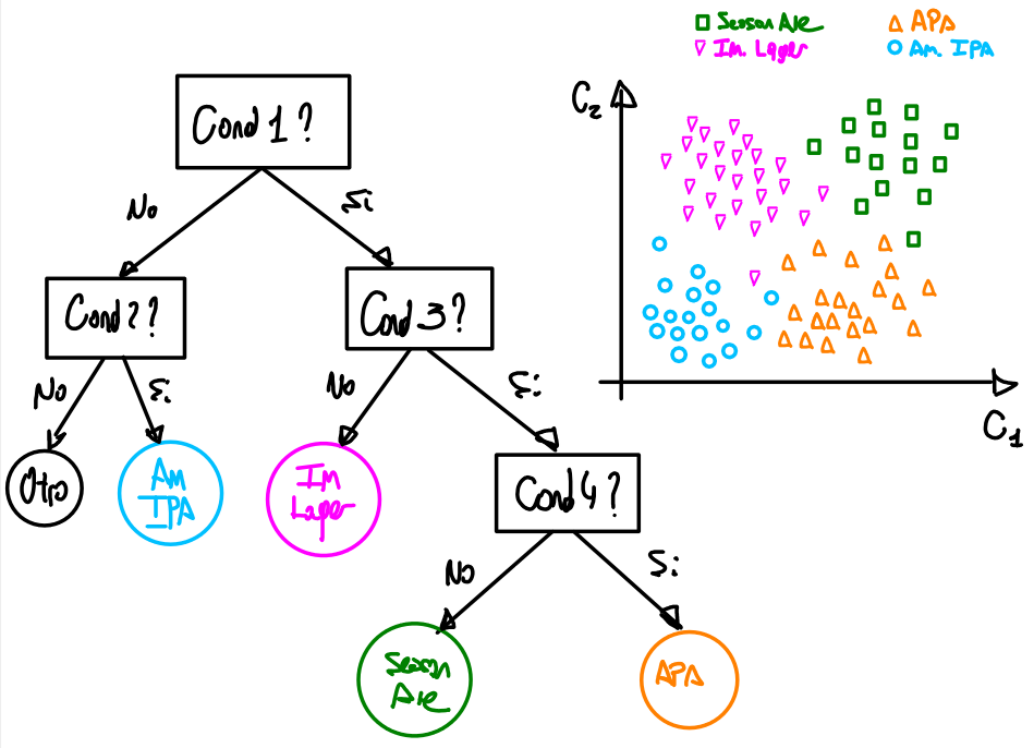
Este esquema no permite responder las preguntas 2 y 3 debido a que no aporta información sobre las propiedades de cada receta (OG, FG, ABV, Color, etc) y su relación con el estilo de cerveza.

Para lograr esto se podría graficar la distribución de una variable para cada clase o estilo y por lo tanto se requiere un gráfico por variable, aunque para este caso se pueden agrupar en una grilla de 3x3. Las distribuciones se pueden visualizar con diagramas de caja, de violín o simplemente de puntos dispersos. Por simplicidad se muestra un sketch del último caso.

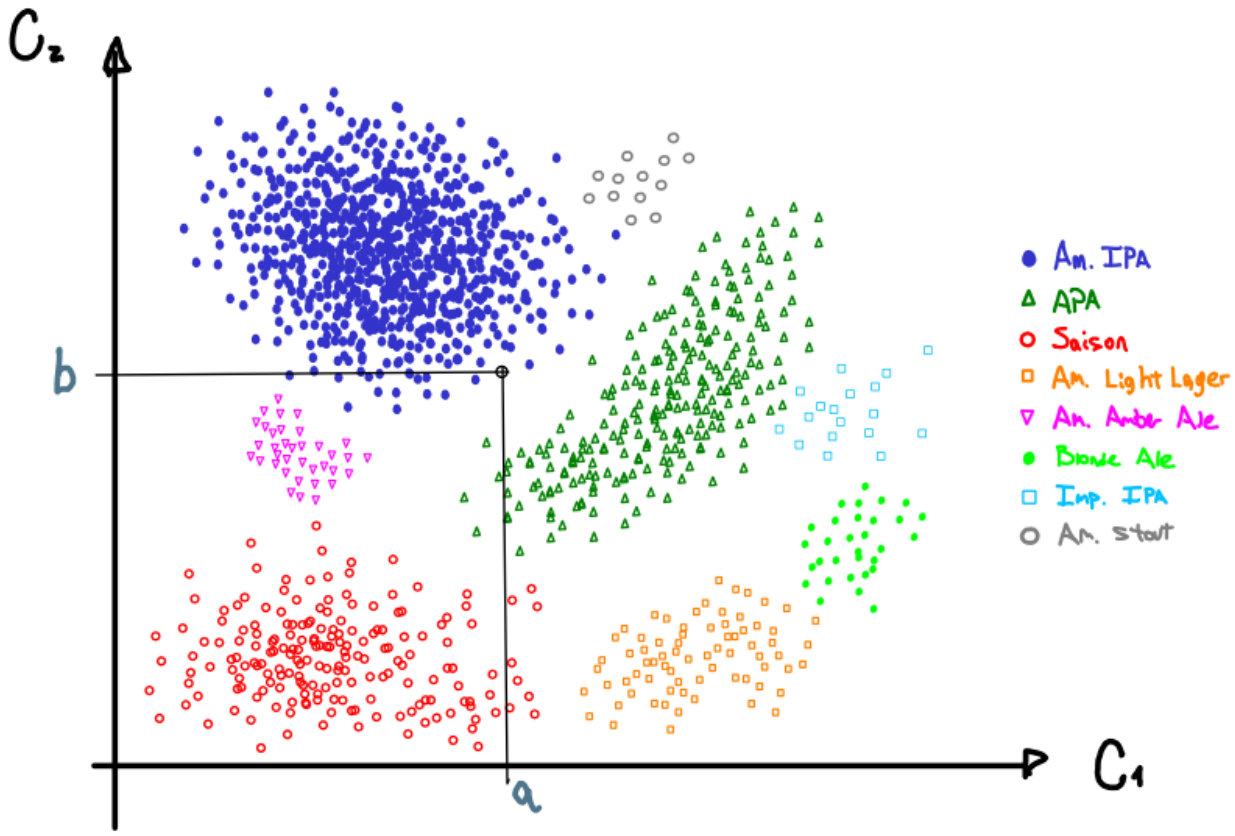
Este tipo de gráfica simplemente permite identificar las variables que potencialmente puedan emplearse para la clasificación de los estilos.



Para responder la pregunta 3, se proponen dos formas de visualizar la clasificación, una es por medio de un arbol de decisión o también dendrogramas y la otra es mediante un gráfico de datos dispersos del espacio de componentes (por ejemplo obtenido mediante LDA).



La opción más conveniente depende de la complejidad del modelo de clasificación, si las reglas son sencillas y la precisión del método es buena, entonces es conveniente emplear el árbol. En caso de que las reglas sean complejas o haya mucha dispersión de los datos, se recomienda definir dos o tres índices y graficar la posición del resultado en un gráfico de nube de puntos diferenciando las clases por color y marcador.
Finalmente se agrupa en un mismo mapeo visual el esquema que permite responder las tres preguntas planteadas, que en este caso podría ser los datos agrupados por clase en un espacio de índices apropiado.



En este gráfico se puede apreciar que la cantidad de muestras es proporcional a las recetas más populares y es posible visualizar las características de cada estilo dada la proximidad entre clases. Además, dada una nueva receta creada por un usuario, es posible saber a qué estilo se asemeja más y si se conocen las fórmulas de las variables del espacio, también sería posible determinar qué atributos de la receta se deben modificar para inclinar el resultado hacia otro estilo.