Análisis de clustering para las propiedades de distintas marcas de cerveza en Colombia – Grupo 10

Santiago Pulido, Liliana Briceño, Cristhian Barbosa, Diego Peñaloza

Enlace Repositorio en Github: https://github.com/Cbarbo82/ANS-G10.git

1. RESUMEN

Se propone realizar un perfilamiento del mercado de cerveza nacional a partir de las características de las marcas que lo componen, estableciendo perfiles cerveceros que agrupen diferentes marcas para identificar posibles oportunidades de negocio.

Para ello, se cuenta con las características químicas, sensoriales y de producción global de 52 marcas de cerveza masivas y artesanales del mercado nacional para los años entre 2020 y 2023. Cada conjunto de características se compone de distintas variables que aportan información valiosa sobre las características de cada marca.

Con la información disponible se busca que, empleando técnicas de aprendizaje no supervisado, como el análisis de clústeres, permita segmentar el mercado de cervezas en Colombia a partir de resultados sensoriales, fisicoquímicos y cantidad de ventas. Esto podría ayudar a identificar oportunidades de crecimiento en sectores no explorados, preferencias del consumidor, desarrollo de marcas nuevas o potenciar las ya existentes.

2. INTRODUCCIÓN

El mercado de la cerveza en Colombia está experimentando un notable crecimiento, con un consumo anual estimado en 30.2 millones de hectolitros, representando ingresos de aproximadamente 22 billones de pesos en 2023. Según Euromonitor [2], el consumo per cápita ha alcanzado 60 litros anuales, posicionando a Colombia como el tercer consumidor de cerveza en América Latina, detrás de México y Brasil.

A pesar de este amplio mercado, la dinamización es limitada debido a la estructura monopolística del mercado. Esto ha restringido la diversificación, limitando la innovación y la introducción de nuevos tipos de cerveza. No obstante, dentro de las principales oportunidades de negocio se encuentra la inclusión de marcas y expansión de cervezas artesanales, un segmento que ha mostrado un crecimiento gradual en los últimos años [3]. Sin embargo, en comparación con otros países latinoamericanos como Argentina o Chile, donde la cerveza artesanal ha capturado una mayor participación de mercado, Colombia aún tiene un largo camino por recorrer.

Este vacío presenta una oportunidad única para cervecerías que deseen explorar y desarrollar el mercado, ofreciendo productos diferenciados que puedan captar la atención de consumidores que están comenzando a valorar la diversidad y la calidad en su consumo cervecero.

El cliente potencial para esta solución incluye cervecerías colombianas que buscan ingresar o ganar una mayor participación en el mercado colombiano. Mediante la aplicación de técnicas de aprendizaje no supervisado, como el análisis de clústeres, se espera encontrar patrones que ofrezcan información útil para detectar segmentos de mercado no explorados con potencial de crecimiento. Así, esperamos analizar cada segmento hallado para entender en qué aumento en ventas, está ganando tamaño de mercado, y que presentan un incremento reciente de ingresos, siendo las más exitosas. El resultado final permitiría recomendar los perfiles de cerveza en los que invertir puede resultar en grandes ganancias.

3. REVISIÓN PRELIMINAR DE ANTECEDENTES EN LA LITERATURA

La industria cervecera ha sido objeto de numerosos estudios a nivel nacional e internacional, enfocándose en aspectos como la caracterización sensorial, el análisis de mercado y la identificación de oportunidades de negocio. Esta revisión preliminar de la literatura se centra en investigaciones que abordan el perfilamiento del mercado cervecero, la segmentación de marcas y el uso de métodos de análisis multivariado, como la creación de modelos de clustering, para identificar oportunidades en el sector.

En el ámbito internacional, varios estudios han explorado la caracterización de la cerveza a través de métodos estadísticos avanzados. Por ejemplo, en “Sensory Characterization of Beer Flavor Using Cluster Analysis” [1] desarrollaron un método de análisis multivariado para evaluar las características sensoriales de diferentes marcas de cerveza en Japón, utilizando análisis de cluster y análisis de componentes principales. Este enfoque permitió identificar similitudes y diferencias en los perfiles de sabor de las cervezas, lo que resultó relevante para el desarrollo de estrategias de marketing y posicionamiento de productos en el mercado.

Por otra parte, para el contexto colombiano, se identificó que las cervezas artesanales tienen un potencial significativo para capturar una mayor cuota de mercado. Este estudio sugiere que la segmentación del mercado y la identificación de características únicas en las cervezas pueden llevar a estrategias efectivas para competir en un entorno dominado por grandes cerveceras [2], [3].

El enfoque propuesto en este proyecto se basa en la integración de características químicas, sensoriales y de producción para realizar un análisis de clusterización que permita identificar vacíos en el mercado colombiano y oportunidades de inversión. A diferencia de los estudios previos, que a menudo se centran en un solo aspecto, este proyecto busca una visión holística que combine múltiples variables para ofrecer un perfil más completo del mercado cervecero colombiano.

4. DESCRIPCIÓN DETALLADA DE LOS DATOS

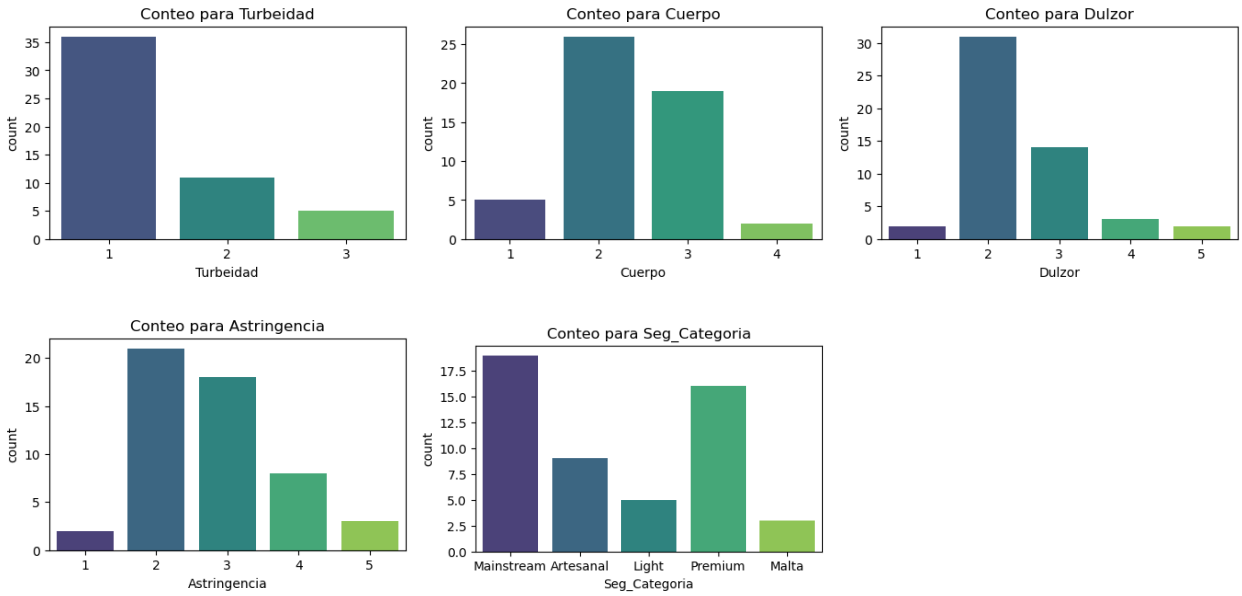
Para el estudio propuesto contamos con una base de datos con 52 marcas de cerveza y 15 variables. Algunos de los campos ofrecen información de las propiedades específicas de cada marca de cerveza respecto a sus condiciones fisicoquímicas, los parámetros sensoriales que ayudan a describir el perfil de cada bebida y el volumen de cerveza vendida a nivel nacional para cada marca, como se indica a continuación:

Variable	Unidad	Definición
Parámetro Físico químicos		
Extracto Original	°P	Concentración de azúcares en el mosto antes de la fermentación. Determina el potencial de alcohol que la cerveza puede alcanzar después de la fermentación
Extracto Aparente	°P	Concentración de azúcares que quedan después de la fermentación.
Alcohol	% v/v	Cantidad de alcohol en la cerveza
pH		Medida de acidez o alcalinidad de la cerveza. Asegura la estabilidad del sabor, claridad y conservación de la cerveza.
Color	EBC	Se refiere a la intensidad de color de la cerveza. Esto varía dependiendo del tipo de cerveza.
Amargo	IBU	Cantidad de compuestos amargos, sobre todo ácidos alfa de lúpulo en la cerveza. El valor de IBU varía dependiendo del tipo de cerveza.
Medición Sensorial		
Turbiedad	1. Claro / cristalino. 2. Ligeramente turbio. 3. Moderadamente turbio.	Se refiere a la clara u opacidad de la cerveza. Influye en la percepción visual de la bebida

	4. Turbio.	
Cuerpo	1. Cuerpo bajo. 2. Cuerpo ligero medio. 3. Cuerpo moderado. 4. Cuerpo alto.	Sensación de textura en boca de la cerveza. Se pueden llegar a percibir cervezas más “pesadas” o más “ligeras”.
Dulzor	1. Nada dulce. 2. Ligeramente dulce. 3. Moderadamente dulce. 4. Dulce alto.	Percepción de los azucares residuales en la cerveza.
Astringencia	1. Ninguna. 2. Leve. 3. Moderado. 4. Alto.	Sensación de sequedad o aspereza en la boca.
Ventas 2021	Hl (hectolitros)	Cantidad de hectolitro vendidos por año.
Ventas 2022		
Ventas 2023		

Del total de variables, 10 son numéricas y 15 son categóricas. De las observaciones solo hay valores faltantes en las ventas del año 2020, quizá explicado por la existencia de marcas nuevas que entraron en el mercado en ese año. Se observan también 2 valores faltantes en el volumen de ventas para el 2021 y el 2023.

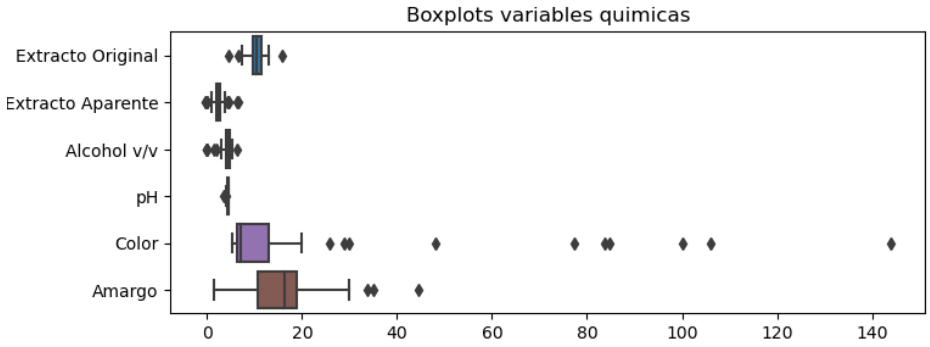
Para las variables numéricas están las estadísticas descriptivas (Anexo 7.1), que permiten observar que las variables tienen diferentes unidades de medida y escalas. Respecto a la variabilidad de los datos, destaca el ‘color’ con una desviación estándar de 30.78 y un coeficiente de variación del 151%. Las variables de ventas anuales entre 2020 y 2023 reflejan que existen marcas con diferencias significativas en la cantidad de hectolitros vendidos, lo que se observa claramente en la amplitud de los rangos intercuartílicos.
Las variables de tipo categórico y ordinal cuentan con los siguientes volúmenes por categorías:



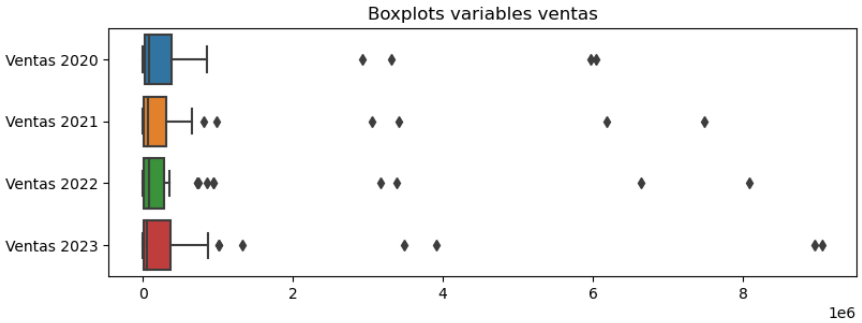
Los gráficos permiten ver que el 67 % de las cervezas tiene turbiedad clara/cristalina, el 86 % de las marcas tienen cuerpo ligero medio o moderado, alrededor del 73% de las revisadas tienen un grado de astringencia leve o moderada y el 84 % son leve o moderadamente dulces. Por último, la categoría revela que la mayor

parte de las cervezas son Mainstream o premium, mientras las malta y light se encuentran en menor medida.

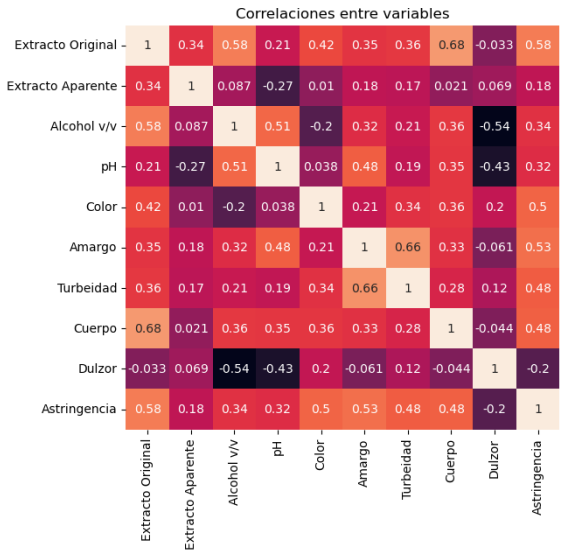
Las variables correspondientes a las características químicas cuentan con las siguientes distribuciones. Destaca la amplia presencia de valores atípicos en la variable 'color'.



Por otro lado, las variables de ventas anuales entre 2020 y 2023 cuentan con las siguientes distribuciones, señalando que existen algunas marcas con ventas muy por encima del promedio.



Al revisar posibles correlaciones entre las variables encontramos algunas correlaciones leves (entre 0.5 y 0.6) entre las variables turbiedad y amargura, la medición de cuerpo y extracto original y astringencia y extracto original. Esto nos lleva a reconsiderar la exclusión de algunas de estas variables que nos aportan información redundante.



Para adicionar información al modelo, se construyen medidas del incremento en ventas entre cada año que bien pueden emplearse al momento de construir el modelo de clúster o perfilar los resultados obtenidos.

	Incremento 2021	Incremento 2022	Incremento 2023
Promedio	-	-0.60	-0.135
Desviación Estándar	-	1.84	0.729
Min	-	-10.74	-2.783
25%	-	-0.48	-0.33
50%	-	-0.11	-0.09
75%	-	0.072	0.288
Max	-	0.938	0.89

5. PROPUESTA METODOLÓGICA

Con el propósito de abordar la pregunta de negocio planteada, se propone la siguiente metodología con una fase inicial de análisis y transformación de datos, seguido de la aplicación de diferentes modelos de clustering y la selección e interpretación del mejor modelo en atención al contexto del problema:

5.1. Análisis descriptivo

Se realiza un análisis de los datos proporcionados y validados, que permita identificar sus principales medidas de tendencia central y dispersión. A su vez se evaluará la distribución de cada una de las variables y las relaciones existentes entre ellas. Por último, se examinará la calidad de los datos respecto a la presencia de valores atípicos y valores faltantes. Los resultados obtenidos nos permiten obtener un resumen de las propiedades de los datos, como apoyo al análisis e interpretación de los modelos elaborados.

5.2. Transformación de datos

Se procede a aplicar One-Hot Encoding a las variables categóricas a fin de convertirlas en una serie de variables binarias las cuales sean compatibles con los modelos de clustering.

Acto seguido, se realiza la estandarización de las variables numéricas de modo que tengan media cero y una desviación estándar igual a uno. La normalización de los datos originales que incorporan diferentes escalas de medida evita que las variables con mayores magnitudes afecten de manera errada el cálculo de los modelos propuestos.

5.3. Clustering y validación de modelos

En esta etapa se propone implementar diferentes modelos a fin de identificar agrupaciones en las marcas de las cervezas. Para ello, se realizará el proceso con los diferentes algoritmos abordados en clase:

- K-medias: algoritmos basados en centroides que busca que “las observaciones dentro del clúster sean los más similares entre sí, y fuera de los clústeres lo más disimilares entre ellas”.
- K-medoides: algoritmo similar a k-medias, con una aproximación diferente para determinar los centros de los clústeres. Es posible aplicarlo considerando que usamos una muestra pequeña.
- Clustering Jerárquico aglomerativo: algoritmo para la construcción de clústeres de manera incremental, mediante dendogramas. Teniendo en cuenta no planteamos patrones iniciales sobre los datos, este modelo resulta de particular interés.
- DBSCAN: algoritmo de clustering basado en densidad, que destaca por la forma en que aborda datos atípicos, así como su robustez a la forma de los clústeres.

Los resultados de los modelos se validarán mediante el examen detallado de los clústeres obtenidos, para que cada agrupación tenga características diferenciables de los otros grupos, así como una alta cohesión

al interior de sus observaciones. El uso de cada modelo se acompañará de la aplicación de las diferentes métricas propuestas en clase, como el coeficiente de silueta, método del gráfico de codo, entre otros que apliquen según el caso.

5.4. Interpretación de resultados

Se selecciona el modelo que responde mejor a las preguntas de negocio propuestas, detallando las principales características de los clústeres generados y su impacto a las problemáticas propuestas.

6. BIBLIOGRAFÍA

[1] Mawatari, M., Nagashima, Y., Aoki, T., Hirota, T., & Yamada, M. (1991). Sensory Characterization of Beer Flavor Using Cluster Analysis. Journal of the American Society of Brewing Chemists, 49(2), 59–64. https://doi-org.ezproxy.uniandes.edu.co/10.1094/ASBCJ-49-0059

[2] El Tiempo. (2024). *Competir en un mercado que históricamente ha sido un monopolio no ha sido una tarea fácil: Central Cervecera*. https://www.eltiempo.com/economia/empresas/competir-en-un-mercado-que-historicamente-ha-sido-un-monopolio-no-ha-sido-una-tarea-facil-central-cervecera-3370142

[3] La República. (2024). *La industria cervecera artesanal tiene alrededor de 0,5% del mercado total de licores*. https://www.larepublica.co/consumo/la-industria-cervecera-artesanal-tiene-alrededor-de-0-5-del-mercado-total-de-licores-3444506

[4] Castellanos, J. y Sossa, C. (2022). Industria cervecera colombiana: un análisis desde su comercio internacional. Expresiones, Revista Estudiantil de Investigación, 9(17), 51-59.

[5] Mariño, Cynthia, (2023). Análisis de clustering para la segmentación de mercado: un caso de estudio de una aplicación de una bebida alcohólica en las principales ciudades de Colombia. Universidad del Bosque.

7. ANEXOS

- **Estadísticas descriptivas de las variables continuas**

	count	mean	std	min	25%	50%	75%	max
Extracto Original	52.0	10.31	1.80	4.43	9.74	10.50	11.34	15.75
Extracto Aparente	52.0	2.26	1.38	-0.48	1.92	2.20	2.66	6.64
Alcohol v/v	52.0	3.95	1.38	0.00	3.88	4.17	4.74	6.21
pH	52.0	4.32	0.24	3.50	4.27	4.39	4.47	4.61
Color	52.0	20.36	30.78	5.20	6.22	7.18	13.08	144.00
Amargo	52.0	16.28	7.97	1.50	10.62	16.25	18.85	44.40
Turbeidad	52.0	1.40	0.66	1.00	1.00	1.00	2.00	3.00
Cuerpo	52.0	2.35	0.71	1.00	2.00	2.00	3.00	4.00
Dulzor	52.0	2.46	0.83	1.00	2.00	2.00	3.00	5.00
Astringencia	52.0	2.79	0.96	1.00	2.00	3.00	3.00	5.00
Ventas 2020	39.0	616716.52	1448081.38	0.00	22258.28	77947.84	390283.35	6039650.02
Ventas 2021	50.0	548159.30	1457252.98	1.65	7968.62	68315.66	316828.05	7485885.84
Ventas 2022	49.0	602002.96	1567569.85	522.51	7457.17	86133.76	282336.14	8083773.64
Ventas 2023	50.0	698367.48	1874624.43	62.98	8429.30	55171.80	371626.87	9059599.28

- **Pairplots de variables continuas**