Modelo **Machine Learning** basado en **Regresión** para predicción en opiniones de cervezas

Institución: Coder House
Curso: Data Science

Integrante del equipo:
• Rodrigo Larroca

Link directo al informe web:
https://github.com/RodrigoLarroca/Beer/blob/0227ce585c32ad470deb54a567deb3ecdfe6add8/Prediccion_Cervezas_Larroca.ipynb

Lobby de contenidos:

Contexto analítico y comercial, problema comercial e hipótesis:

CONTEXTO ANALITICO: El equipo consiguió datos sobre una gran cantidad de cervezas, su composición, sus gustos y opiniones realizadas a 3200 personas. Los datos aquí no estarán etiquetados; es decir, no hay una variable que nos diga cuáles de estas opiniones son buenas o no. En su lugar, debemos utilizar modelos de regresión para abordar este problema de aprendizaje supervisado.

CONTEXTO COMERCIAL: Trabaja en una consultora privada de datos, donde un cliente del rubro de cervezas requiere sus servicios para encontrar el óptimo gusto en su nuevo proyecto de cerveza. El mercado actual presenta mucha demanda debido a las fechas festivas, por lo cual es imperante para el cliente obtener resultados en un tiempo determinado. Es función del equipo determinar qué gustos/aromas/estilo de cerveza es el de más gustado y por ende el posible más consumido.
PROBLEMA COMERCIAL: Se espera mejorar la elección de ingredientes en base a opiniones de personas. ¿Existen patrones particulares en los ingredientes de las cervezas que puedan ser indicativos de opiniones excelentes?

HIPÓTESIS: Utilizando un conjunto de datos que incluyen información sobre los ingredientes utilizados en diferentes cervezas, así como las opiniones de clientes, se puede entrenar un modelo de aprendizaje automático que sea capaz de predecir según los ingredientes que se utilizarán , cuál será la opinión del cliente. Esto podría ayudar a los cerveceros a mejorar la calidad de sus productos ya satisfacer mejor las necesidades y gustos de sus clientes. Así mismo aportar información sobre las mejores cervezas para poder dar excelentes recomendaciones a mi profesor 🍺🍻

Estructura de los datos:

El conjunto de datos principal (beer.csv) contiene las siguientes columnas:

Nombre : Nombre de la cerveza (etiqueta)

Estilo : Estilo de la cerveza

Cervecería : Nombre de la cervecería

Nombre de la cerveza (completo)

Descripción : Notas en la cerveza si está disponible

ABV : Contenido de alcohol de la cerveza (% por volumen)

Min IBU

Max IBU

Las siguientes once columnas representan las características del perfil de degustación de la cerveza.

(Sensación en boca)	(Sabor)	(Sabor Y Aroma)
Astringencia	Amargo	Frutas
Cuerpo	Dulce	Lúpulo
Alcohol	Agrio	Especias
	Salado	Malta

Las últimas seis columnas contienen información de reseñas

opinión _ aroma

opinión _ aspecto

opinión _ paladar

opinión _ sabor

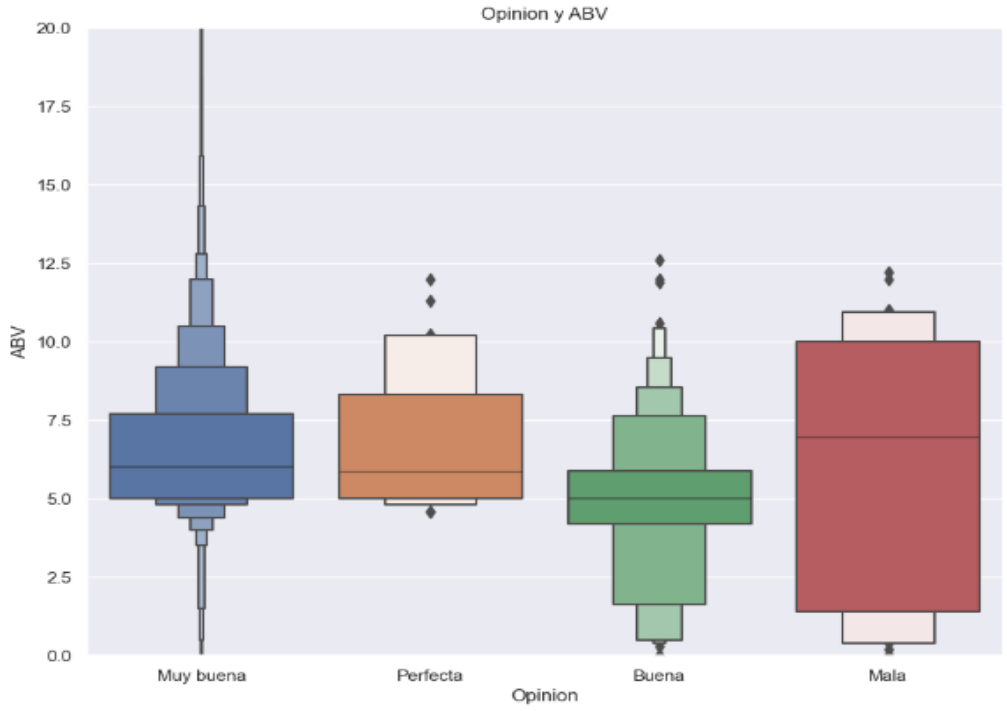
opinión _ general **#TARGET**

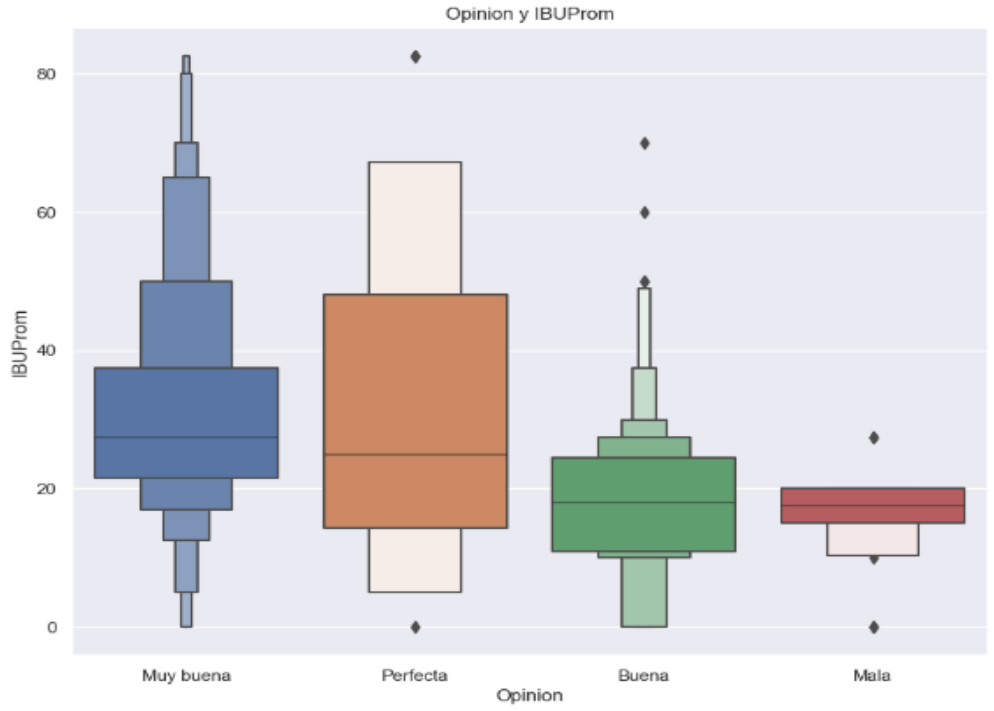
número _ de _ opiniones

Análisis exploratorio de los datos (EDA):

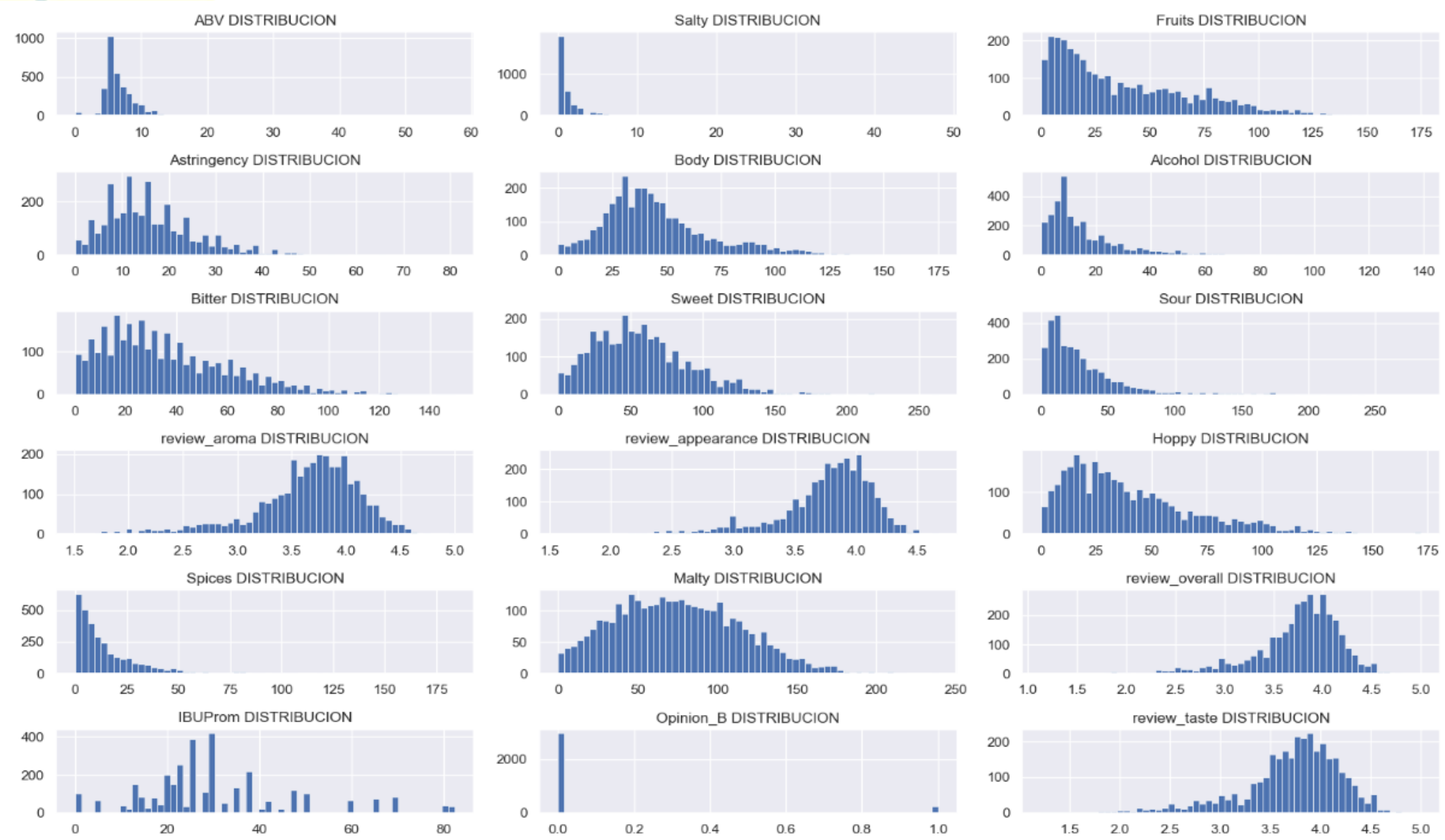
Insights generales y distribución:

El dataset cuenta con 3197 filas y 27 columnas.
No existen datos nulos.
Se creó una columna con el IBU promedio y review overall en letras. También una revisión overall tipo booleana.
Al existir tantas cervezas se analizará según estilos y se dividirá en grupos de opiniones.
Existen columnas innecesarias a los casos del análisis (Brewery, Beer Name(Full), Description).
Puede existir valores muy extremos, por ej valores 0 min y 60 max con 6 de media. Se buscará analizar por grupos entonces.





Es importante informar de la nueva columna “Opinión”. Siendo Perfecta la birra entre (5-4) puntos, Muy buena (4-3), buena entre (3-2), Mala (2-1) y Muy mala (1-0). Y es interesante ver la distribución de cada una con respecto a los ingredientes más importantes, como lo son el ABV (alcohol%) e IBU Promedio (medición del amargor).



De igual manera, una vista rápida a las distribuciones de las distintas columnas nos da una idea acerca de las relaciones entre las mismas, ya que a futuro se ve esto mismo en un mapa de calor y se debe seleccionar un método según la distribución, que a priori en algunos casos se asemeja a una normal pero sería muy generoso.

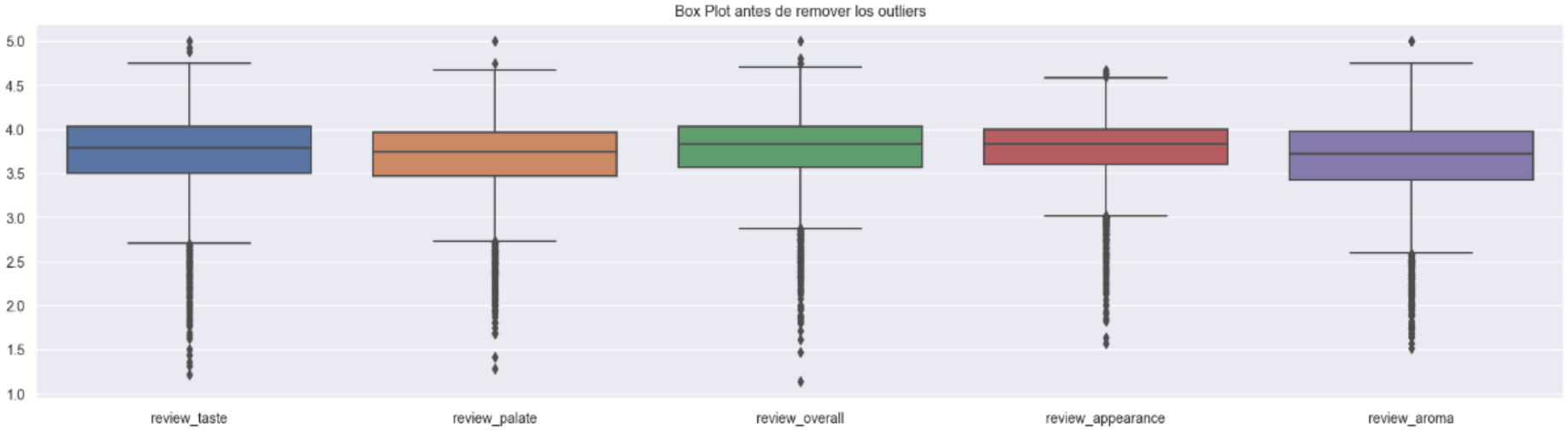
A continuación se tomó un análisis en los outliers para intentar resolver las distribuciones con esperanza de volverlas un poco más “normal”.

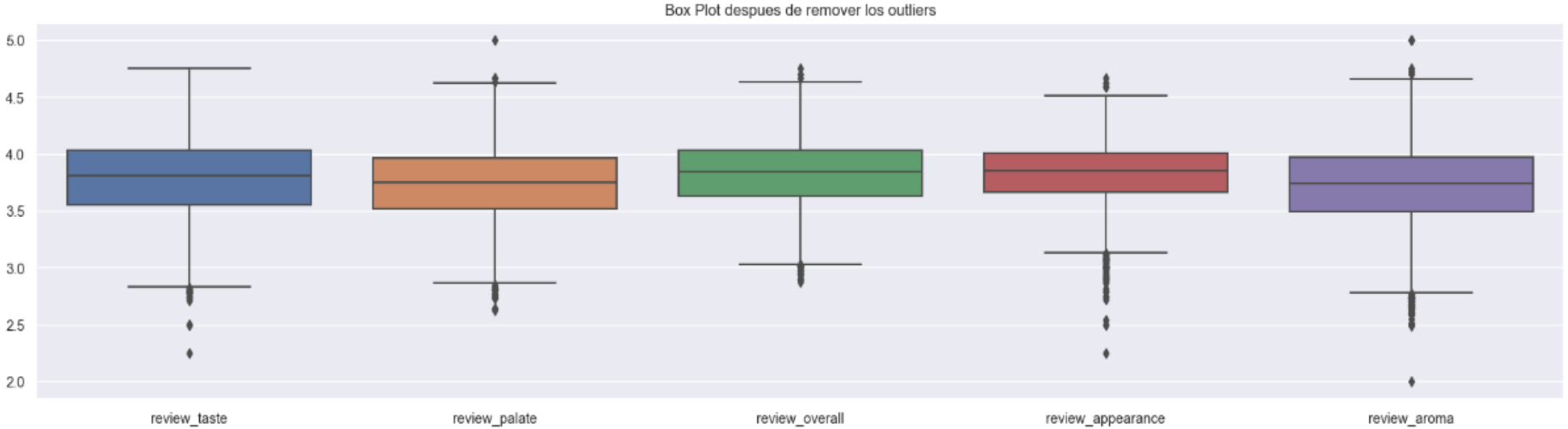
Outliers

Antes y después de la eliminación de outliers:

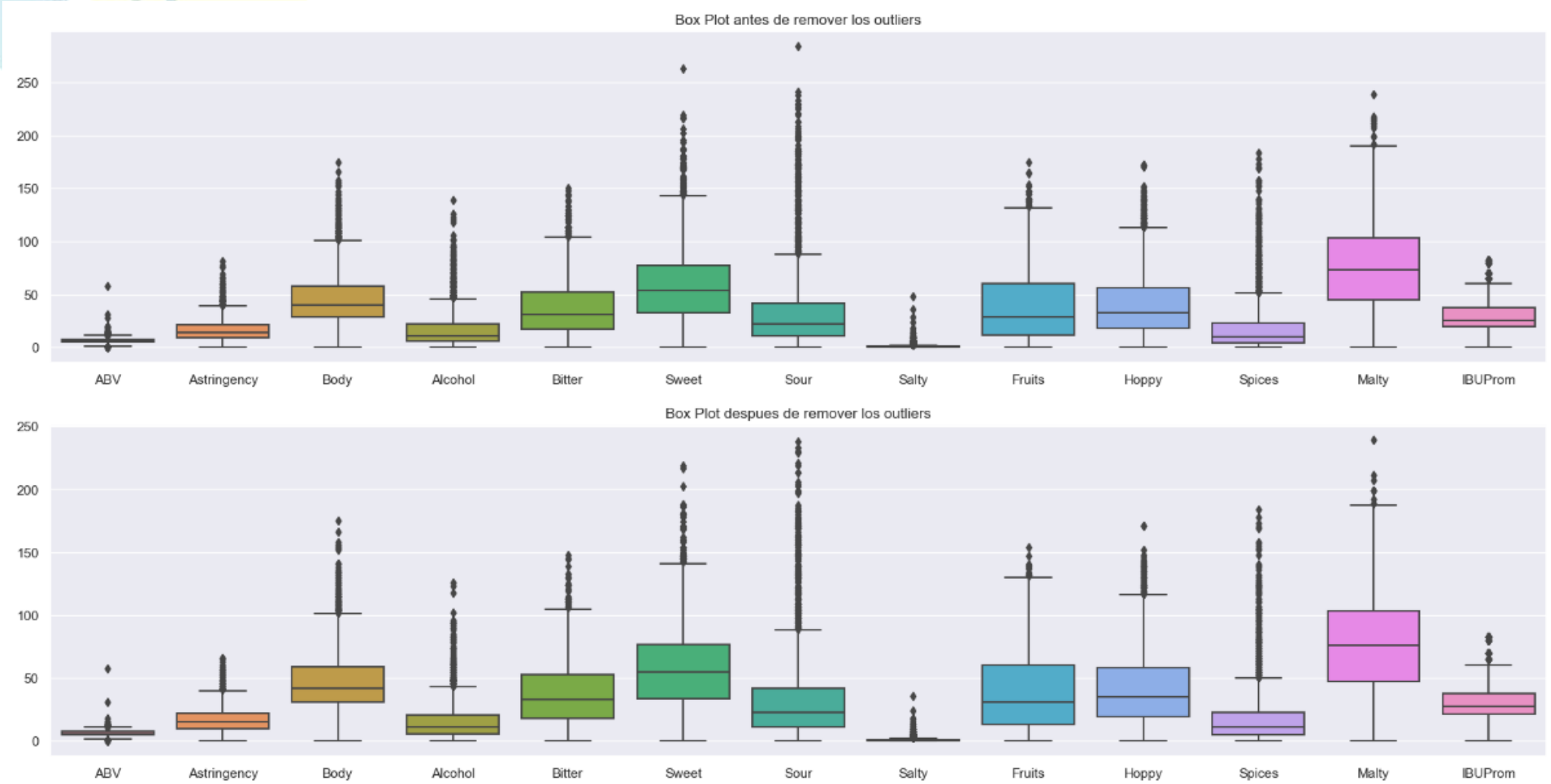
De 3179 cervezas se pasaron a 2833. Eliminando la mayoría de muy malas, a finales del análisis con malas y buenas es suficiente.

Primero una vista de los outliers en las opiniones:





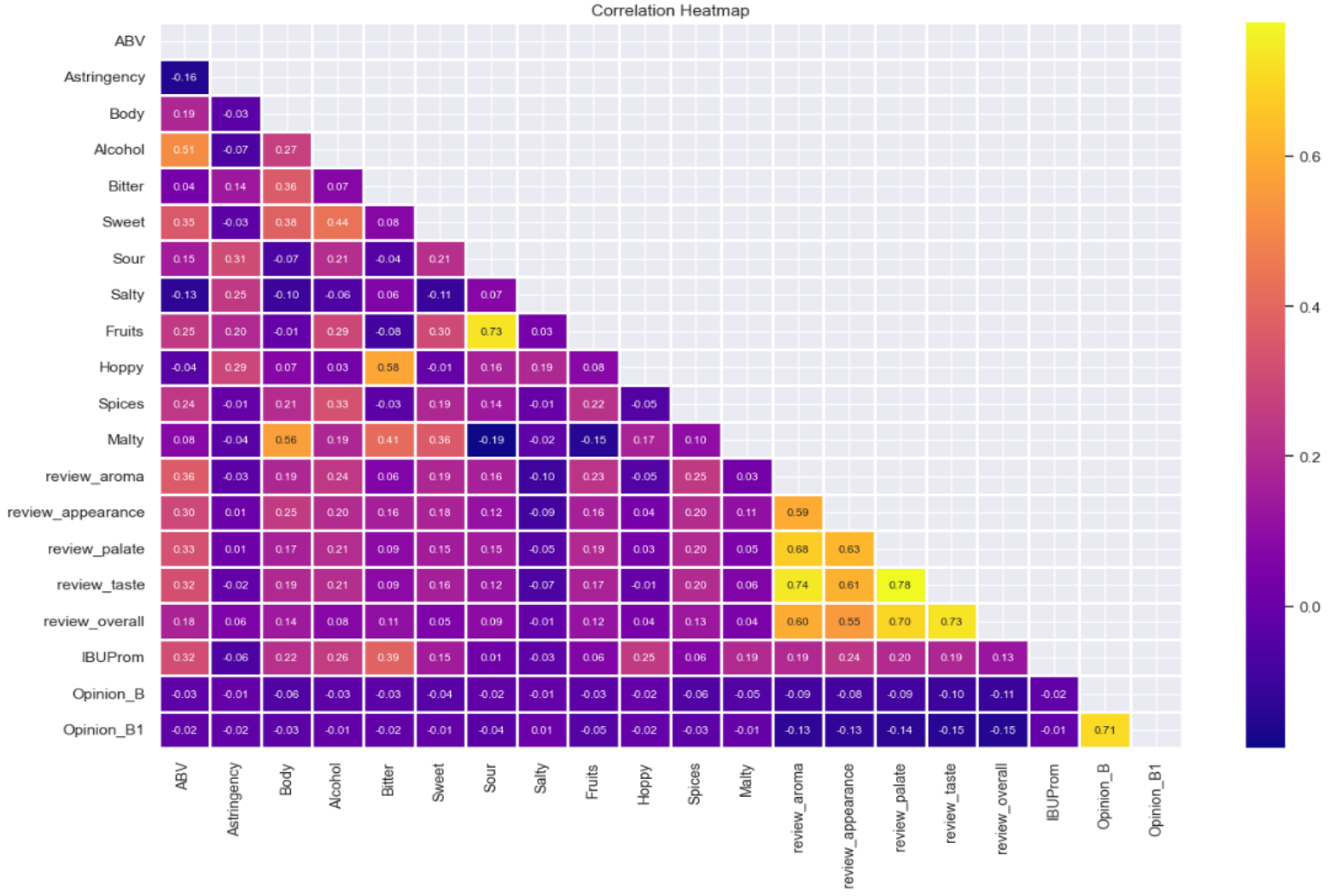
Y ahora una vista de los outliers en los gustos y sabores, también ingredientes:



En fin, a grandes rasgos no hubo muchos cambios. Pero si se nota una ligera mejora en el sentido de los datos. Posteriormente sirve para el tratamiento del modelo de machine learning =). Por ahora se verá un poco sobre elementos en especificos: Opiniones y ABV, sabores, composición y respuestas a interrogantes

Respuestas a interrogantes de relación y composición

Como comienzo, las relaciones de las variables entre sí es algo importante a saber. Si bien se analiza cada una con respecto al target, ya que es lo que se busca predecir, también puede ser factible saber sobre la relación entre todas en general. Además usar un mapa de calor a primera vista puede darnos idea de cómo responden las variables con el target.



En definitiva, se ven las relaciones de -1 a 1. Siendo relación directa o inversa, siendo un gráfico muy claro. Se remarca que se usó el método KENDALL debido a las distribuciones no tan normales.

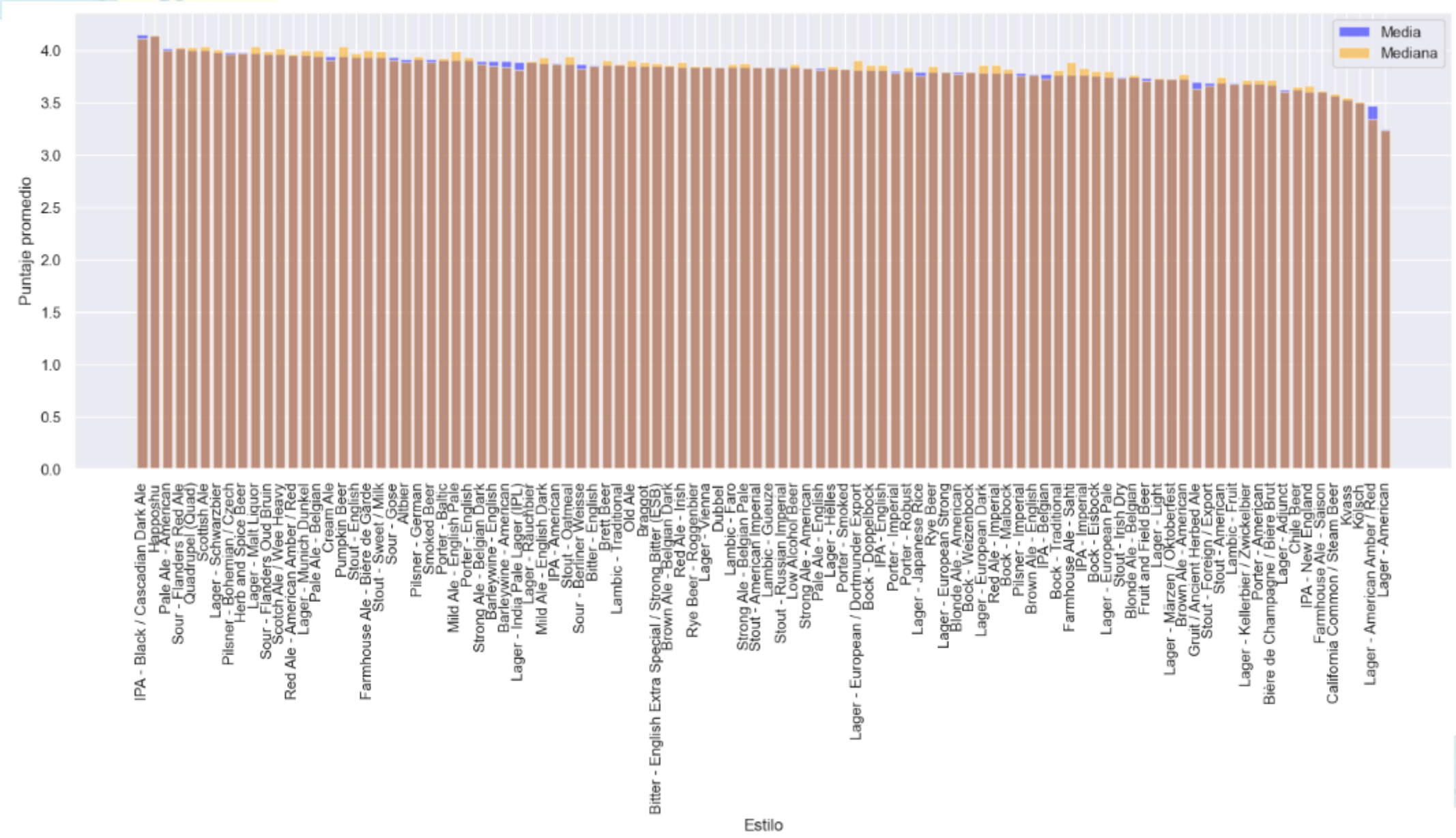
Ahora una pregunta muy importante ¿que afecta en mayor medida a las opiniones?
Se resumen los resultados en un tabla =)

R cuadrado de la regresión lineal simple						
review/variable	¿El ABV?	¿El gusto dulce?	¿El gusto amargo?	¿El gusto a frutas?	¿El gusto agrio?	¿Las especias?
review_aroma	0.9085	0.7813	0.6959	0.6411	0.5162	0.4081
review_appearance	0.9023	0.7782	0.7045	0.6271	0.5045	0.3966
review_palate	0.9053	0.7758	0.6971	0.6316	0.5092	0.3972
review_taste	0.9053	0.7772	0.6975	0.6318	0.5093	0.4004
review_overall	0.8952	0.7657	0.6979	0.6232	0.5042	0.3902

Entonces, en orden de influencia en las opiniones, queda:

1. ABV
2. Sweet
3. Bitter
4. Fruits
5. Sour
6. Spices

Comparaciones de estilos de cervezas y la opinión de las mismas ¿Cuáles estilos gustan más y cuáles menos?
La pregunta se refiere a estilo, ya que para seleccionar un ranking de cervezas más gustadas es un poco desordenado cualquier tipo de gráfico al tratarse de 2800 casos.
En cambio al estar asociadas por estilos, se vuelve una cantidad categórica mucho menor y permite comparaciones más sencillas.

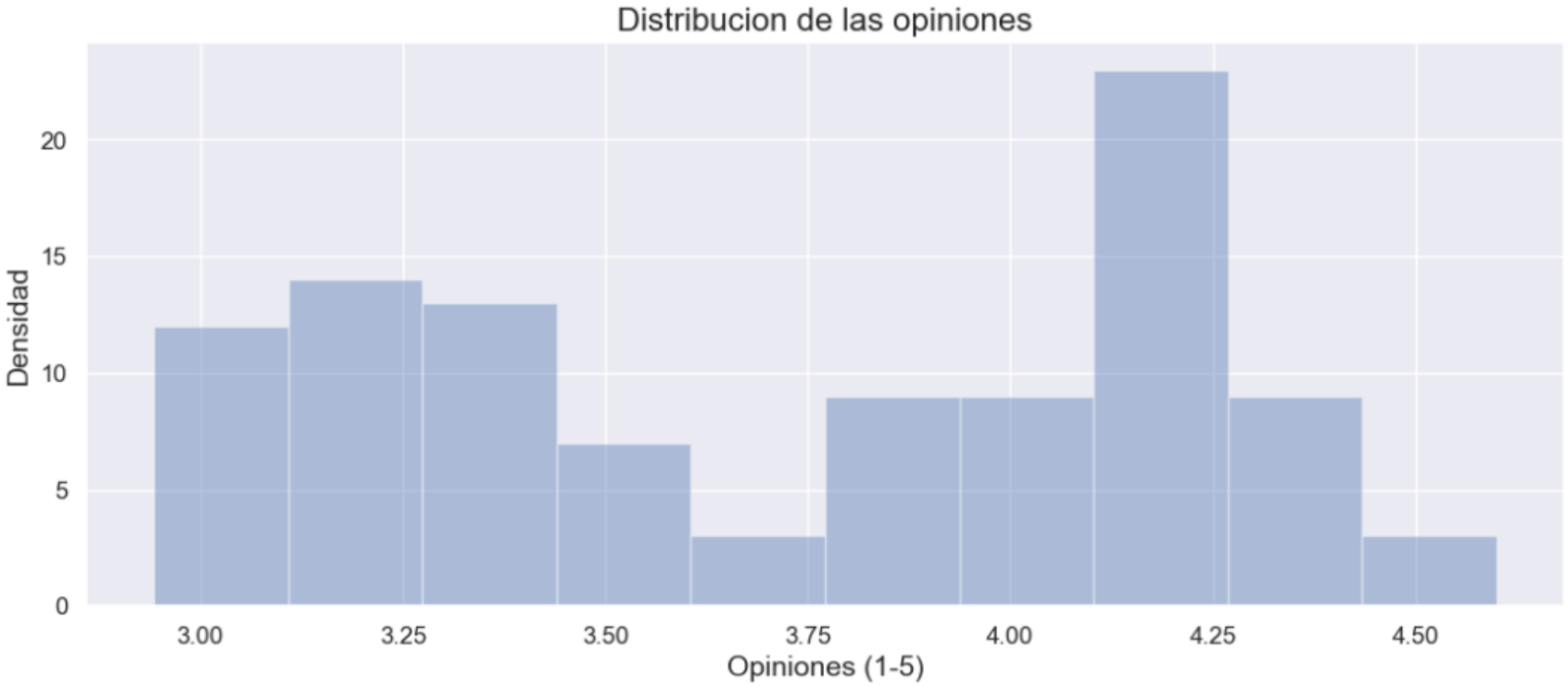


De igual forma y para no prestar a la confusión se seleccionan las 2 birras más puntuadas y las 2 menos puntuadas a fin de un análisis en particular.

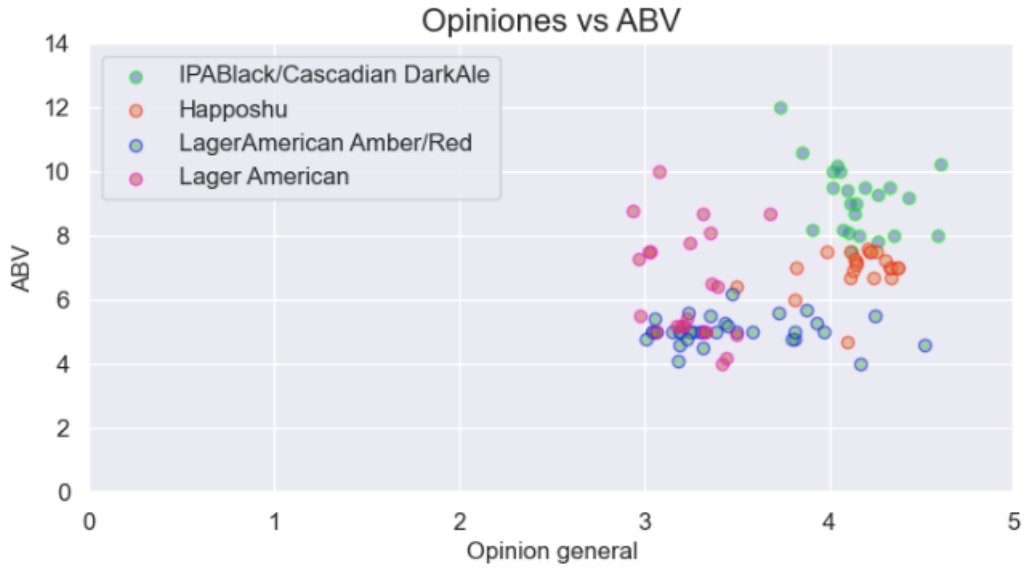
En orden de puntuación más alta:		En orden de puntuación más baja:	
IPA - Black / Cascadian Dark Ale	4.16	Lager - American	3.24
Happoshu	4.15	Lager - American Amber / Red	3.47

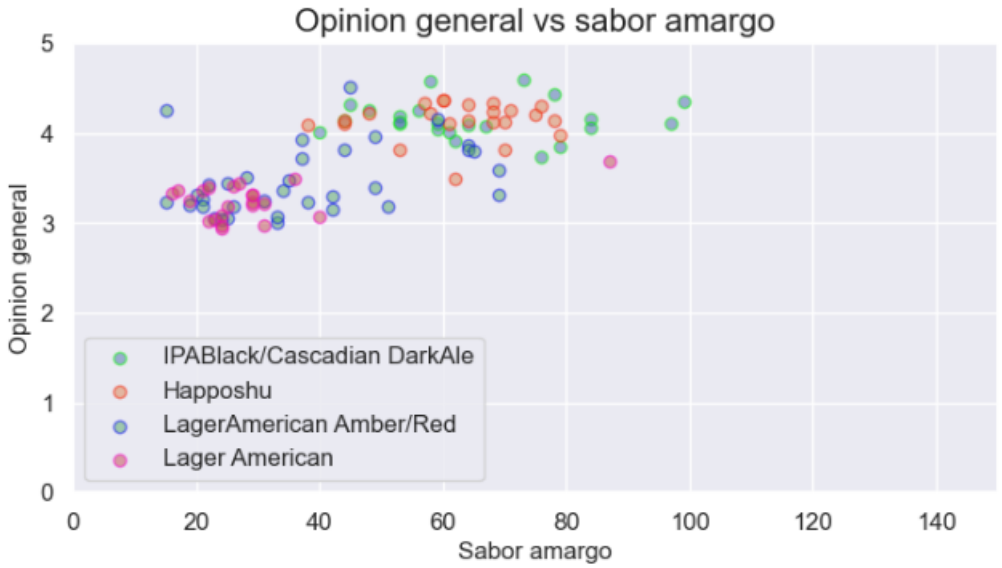
Este análisis particular, ¿cuantos datos posee?

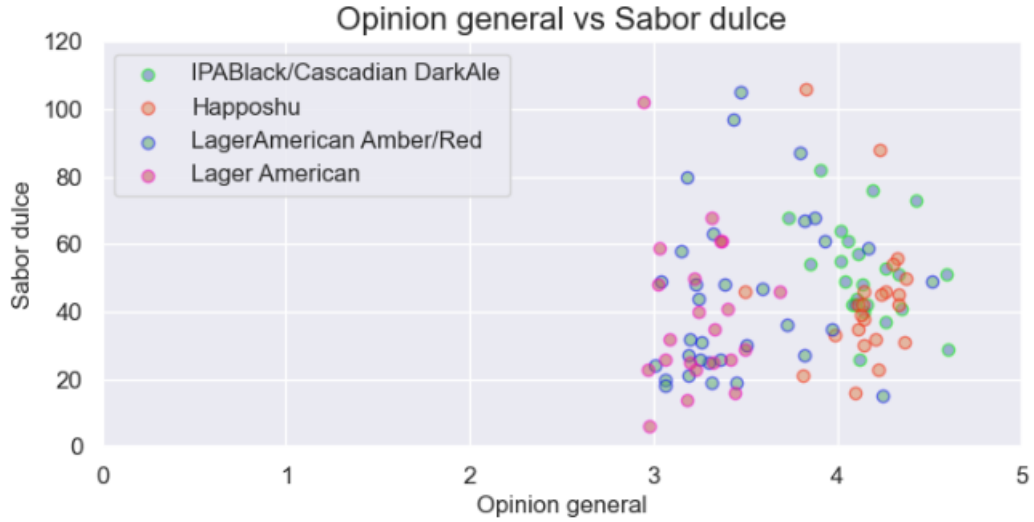
1. Cantidad de opiniones sobre IPA - Black / Cascadian Dark Ale: 23
2. Cantidad de opiniones sobre Happoshu: 23
3. Cantidad de opiniones sobre Lager - American Amber / Red: 34
4. Cantidad de opiniones sobre Lager - American: 22

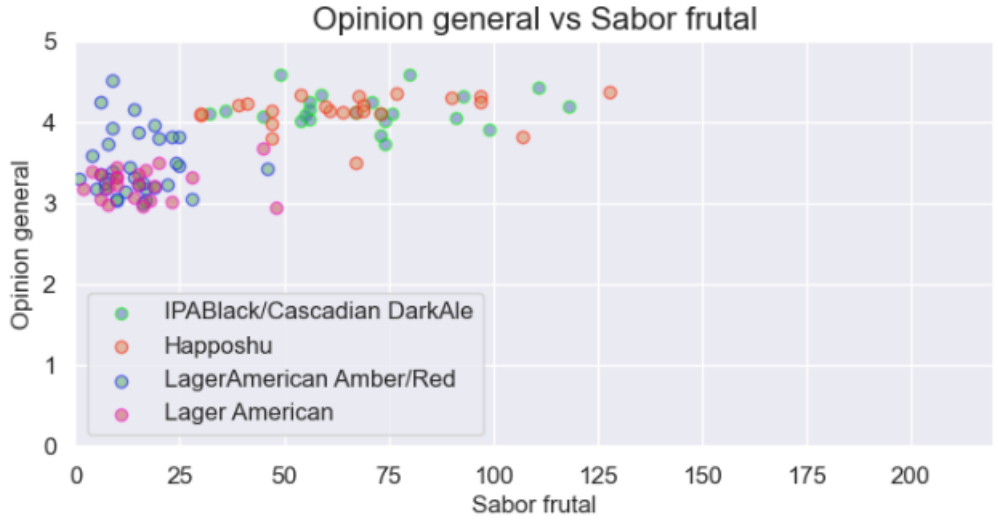


Realmente tiene la distribución esperada. Valores bajos en opiniones, y luego valores altos.
Lo que dará una comparación interesante al dividir los valores en grupos en un gráfico de dispersión.









Este gráfico tiene mucho que decir, y es uno de los que más información aportó. Como se puede ver a continuación.

Las gráficas muestran algunos grupos claros. Por ejemplo, hay un grupo de opiniones alrededor del rango de ABV (Alcohol By Volume) de 5-12 en la tabla de opiniones vs ABV, lo que podría deberse al hecho de que la gente prefiere esa cantidad de alcohol, porque es la cantidad que permite un control del efecto alcohol-consumicion.

También hay una franja de Amargor entre 40-100 en la tabla de Opinión frente a gusto amargo, que valen 4 puntos de opinión aproximadamente. No está claro exactamente cuál proporción de lúpulo es la dominante en gusto, pero podrían ser la conjunción con otros elementos como se ve luego (Ej. sabor dulce o agrio). También hay una franja de opiniones bajas donde existe menor número.

En el gráfico de Opinión versus sabor dulce, hay una franja de personas que opinan que existe un gusto dulce entre 20 a 80 en las cervezas más valoradas (opinion).
En el gráfico de Opinión versus sabor frutal,los estilos mejor puntuados tienen un alto nivel frutal entre 25 y 125.

EN CONCLUSIÓN:
IPABlack/Cascadian DarkAle: el ABV entre 8% y 12%, el sabor amargo entre 40 y 100, un sabor dulce entre 25 y 80, y un sabor frutal entre 30 y 120. Es la clave para la cerveza de mejor puntuacion.
DULCE, muy FRUTAL, AMARGA. ALCOHOL PROMEDIO-alto.

Happoshu: el ABV entre 4% y 8%, el sabor amargo entre 30 y 80, un sabor dulce entre 20 y 60, y un sabor frutal entre 30 y 120. Es la clave para la 2da cerveza. Semi-DULCE, semi-AMARGA, muy FRUTAL. ALCOHOL PROMEDIO.

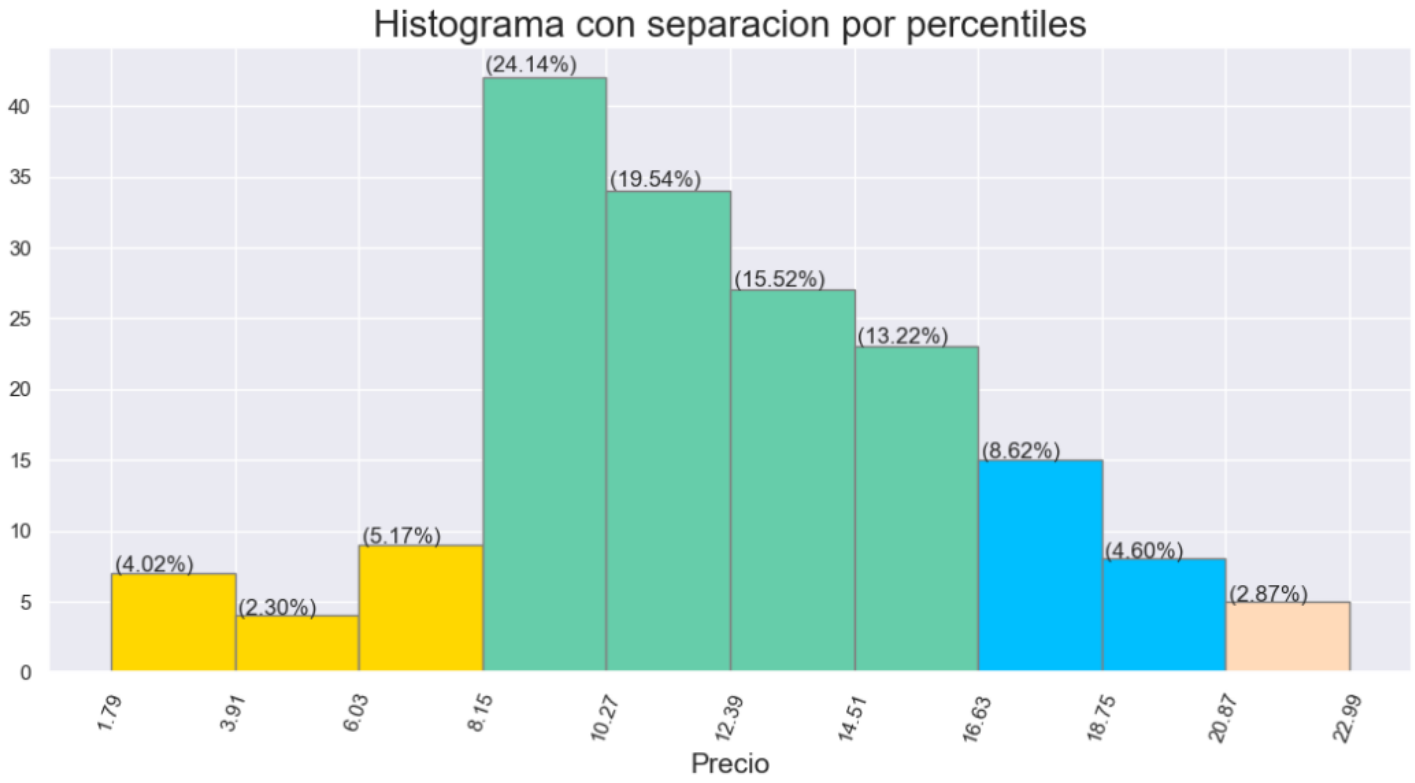
LagerAmerican Amber/Red: el ABV entre 4% y 6%, el sabor amargo entre 20 y 70, un sabor dulce entre 20 y 100, y un sabor frutal entre 0 y 20. Es la clave para una cerveza con opinión 'negativa'.
DULCE, POCO FRUTAL, POCO AMARGA. ALCOHOL PROMEDIO.

Lager American: el ABV entre 4% y 6%, el sabor amargo entre 20 y 40, un sabor dulce entre 20 y 50, y un sabor frutal entre 0 y 30. Es la clave para una cerveza con opinión muy negativa.
POCO DULCE, POCO FRUTAL, MEDIO AMARGO. POCO ALCOHOL.

PLUS - API de cervezas y sus precios

De manera resumida. Se analizó los precios de las cervezas a través del uso de los datos de una API.

El análisis con gráficas de distribución y el tratamiento de outliers se puede ver en el trabajo completo. A fines prácticos se verá el gráfico final y una pequeña conclusión =).



La muestra contiene información de 174 cervezas diferentes. La columna "price" es numérica, tiene un rango de valores entre 1.79 y 22.99, con una media de 12.23 y una desviación estándar de 4.24. Además, se observa un sesgo positivo en la distribución (0.01) y un nivel de apuntamiento normal (-0.20). Esto sugiere que la mayoría de las cervezas tienen precios similares, pero hay algunas cervezas con precios más altos.