Análisis Exploratorio de Datos

¿Cómo Elegir un Buen Vino sin ser un Experto?

Cristina Pérez Fernández

LOS DATOS

En un primer momento, se obtiene el dataset que sería objeto del posterior análisis en la web de "kaggle": se trataba de un conjunto de datos de más de 6000 vinos españoles procedentes de un usuario que había realizado un *webscrapping* de la página "Vivino". Sin embargo, nada más empezar a observar los datos, resultó evidente que se trataba de un dataset erróneo o, al menos, incompleto, pues el precio máximo de los vinos no superaba los 14 euros (a pesar de contar con más de 6.500 entradas).

Así pues, se recurre de nuevo a "kaggle" para la descarga del código de *webscrapping* del usuario; de este modo, alterando dicho código, nos permite conseguir los datos de vinos tintos de todos los rangos de precios, así como también, datos correspondientes con las otras clases de vinos: blancos, rosados y espumosos.

Una vez conseguidos los datos de los vinos según su clase, y convertidos en diferentes *dataframes*, se procede con la limpieza y transformación de los mismos:

- ⇒ Comprobación y eliminación de valores duplicados
- ⇒ Adición de una nueva columna en los *dataframes* donde se especifica la clase de vino: tinto, blanco, rosado o espumoso
- ⇒ Unión de los 4 *dataframes* en uno solo

A continuación, con todos los vinos en un mismo *dataframe*, homogeneizamos los datos de la columna "Año"; ya que vemos que existen varios problemas:

1º) Existen *np.nan*
2º) Hay algunos valores que presentan formato *float*
3º) El *string* "N.V." → Significa "Non vintage" que es un estilo de elaboración muy común en los vinos espumosos.

Con el fin de realizar un análisis exhaustivo del mercado de vinos españoles, nos proponemos añadir dos columnas más al *dataframe*:

- ⇒ Categoría
- ⇒ Denominación

Categorías del Vino

Los distintos vinos se clasifican en diferentes categorías en función de la edad, es decir, el tiempo que ha permanecido envejeciendo tanto en barrica como en botella. Esta clasificación variará según se trate de una clase de vino u otra, siendo especialmente importante en el vino tinto. Además, hay que tener presente que cada D.O. establece sus características particulares; sin embargo, a efectos de este análisis nos guiaremos por una clasificación general.

Categorías vino tinto

Atendiendo a lo anterior, distinguimos las siguientes categorías, que van de menor a mayor tiempo de envejecimiento:

⇒ Joven → no pasan tiempo en barrica; el de la cosecha
⇒ Roble → tiempo total entre barrica y botella 9 meses
⇒ Crianza → tiempo total entre barrica y botella 24 meses
⇒ Reserva → tiempo total entre barrica y botella 48 meses
⇒ Gran Reserva → tiempo total entre barrica y botella 60 meses

Categorías vino blanco y rosado

En esta ocasión, los tiempos de envejecimiento difieren con respecto a la clase anterior, por lo que se tienen las siguientes categorías, pues no es tan común en estos vinos la crianza en barrica:

⇒ Reserva → tiempo total 24 meses
⇒ Gran reserva → tiempo total 36 meses

Categorías vino espumoso

Por último, los vinos espumosos no atienden a las categorías mencionadas; sino que se clasifican por su método de elaboración en:

⇒ Non-vintage → elaboración con mezcla de añadas para conservar la calidad año tras año
⇒ Vintage

Denominación de Origen

Por último, vamos a determinar si la región de procedencia del vino tiene D.O. (denominación de origen), ya que así podremos analizar la influencia de este factor.

Además de la D.O, existen otras expresiones tradicionales como son los Vinos de Pago, los Vinos de Calidad, así como la Indicación Geográfica Protegida; sin embargo, tales denominaciones, a priori, no se tendrán en cuenta en este análisis dado que gozan de escasa popularidad entre el público general y, por tanto, su influencia en la valoración de un vino por parte de cualquier persona ajena al mundo vitivinícola será ínfima.

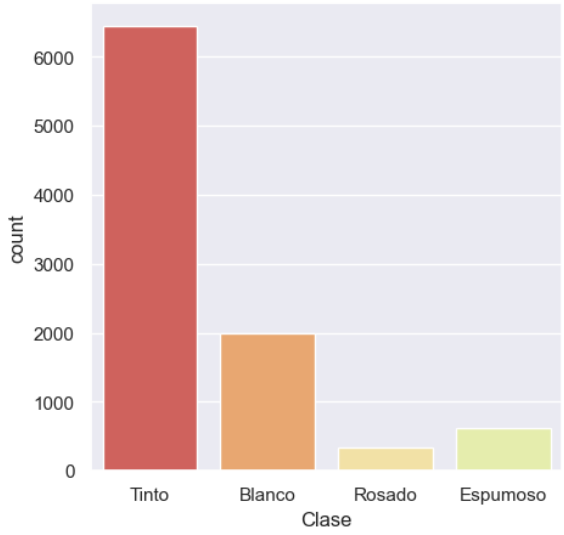
En España, actualmente, existen 69 D.O, 14 denominaciones de Vinos de Pago y 7 de Vinos de Calidad.

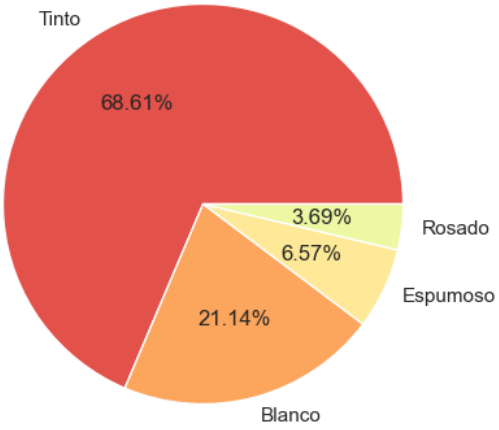
MERCADO VITIVINÍCOLA ESPAÑOL

Comenzamos realizando un breve análisis del mercado del vino español, asumiendo, en todo momento, las limitaciones debidas a que nuestra fuente de datos es única: la página web "Vivino".

Clases de Vino

Dadas las diferencias en los vinos según su clase, lo primero será ver cuál es la participación en el mercado de cada una de ellas partiendo del número de vinos que hay en nuestro *dataset* de cada clase:

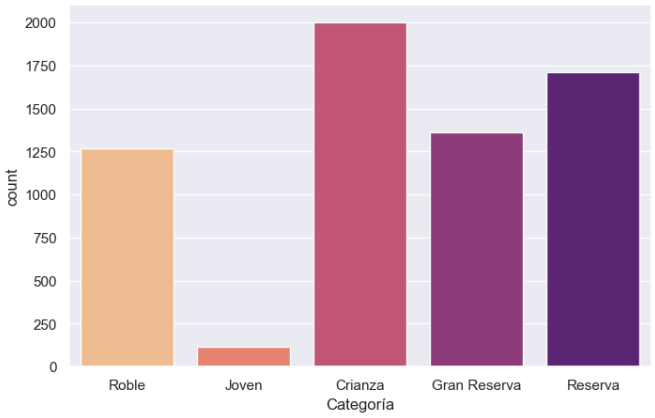




Categoría

Continuamos viendo cómo se reparten los vinos de cada clase en sus distintas categorías:

Vino tinto



Observamos como la totalidad de los vinos tintos se clasifican en las diferentes categorías existentes.

Se corrobora, por tanto, que en el vino tinto la categoría, de acuerdo con la edad del vino, es una característica de vital importancia.

Vino blanco y rosado

Dado lo poco común que es la crianza de estos vinos en barrica, la mayor parte de ellos aparecen con la categoría sin determinar en nuestro *dataset*, existiendo una gran desproporción, por lo que no es óptimo reflejar los resultados de la clasificación en un formato de gráfica.

Blancos

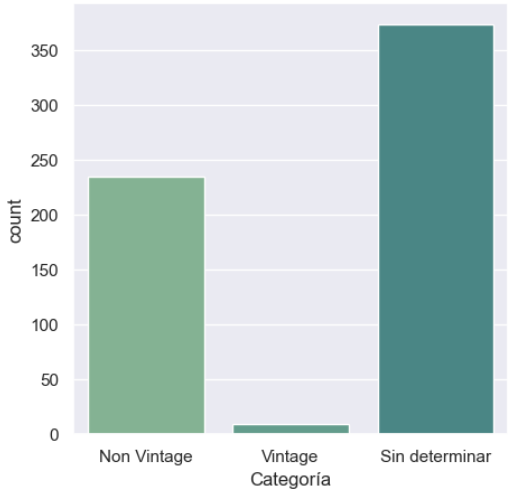
	Categoría
Sin determinar	1966
Reserva	15
Roble	3
Crianza	3
Gran Reserva	1

Rosados

	Categoría
Sin determinar	345
Gran Reserva	2

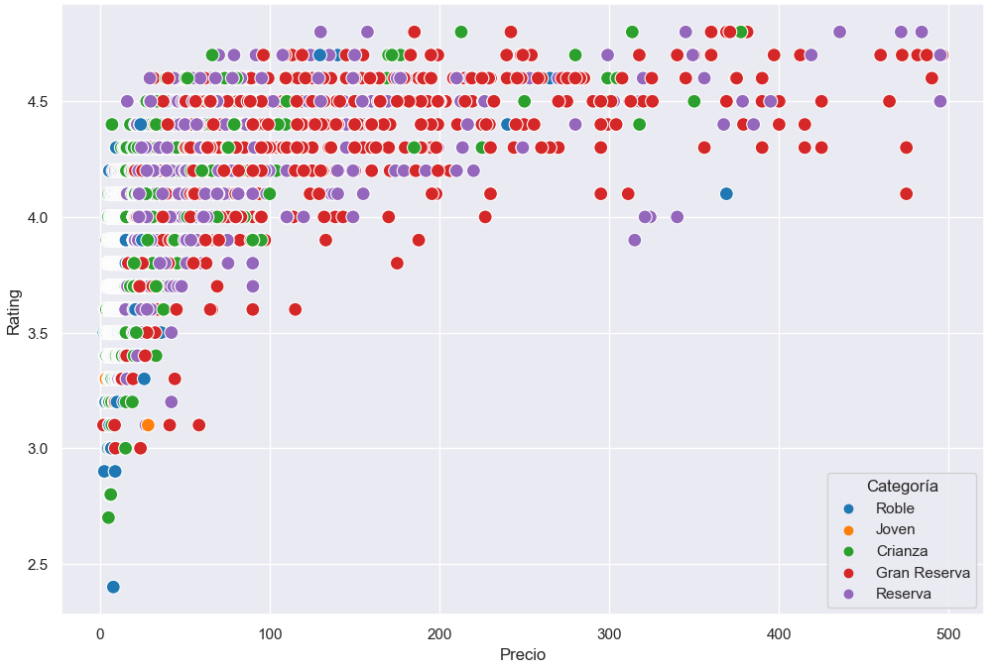
Vino espumoso

En cuanto al espumoso, ya vimos cómo las categorías varían en este caso:



De nuevo observamos, la gran cantidad de espumosos con la categoría sin determinar; aun así, la desproporción no es tan acusada como en los vinos blancos y rosados.

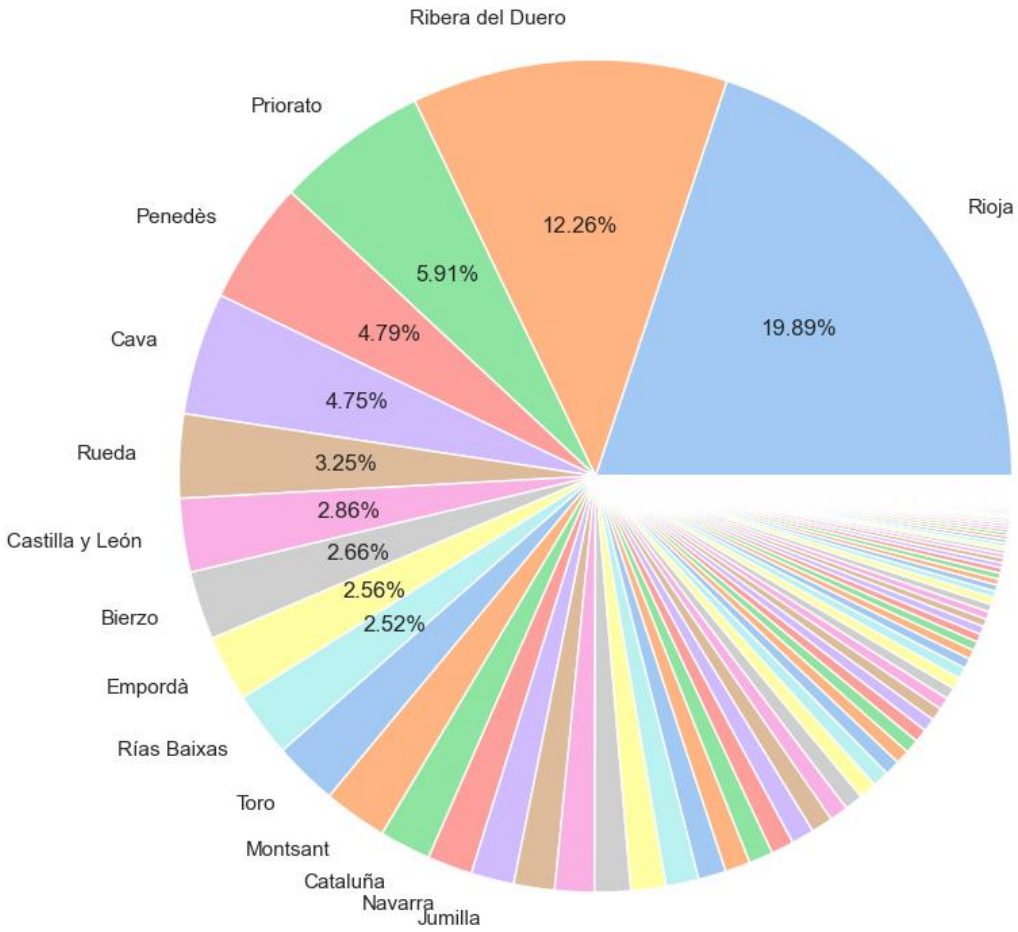
Concluimos, pues, la importancia de la categoría como característica del vino tinto. Ante esto, sería apropiado visualizar la relación de las categorías de tinto con el rating y el precio del vino:



Región

Ahora, vamos a estudiar las regiones vitivinícolas de España. Tendremos que ver cuáles son las regiones con mayor número de vinos en el mercado; y comprobar si dichas regiones varían en función de la clase del vino.

En primer lugar, vemos como en España existen más de 100 regiones productoras de vino:



Por otra parte, dentro del top 15 de las regiones con mayor número de vinos en el mercado, algunas ni siquiera llegan al 2'5% de participación en el mercado.

Seguidamente, se muestran el top 10 de las regiones según cada clase de vino:

Clase	Región	Vino
Tinto	Rioja	1604
	Ribera del Duero	1119
	Priorato	512
	Toro	232
	Bierzo	202
	Montsant	197
	Castilla y León	177
	Empordà	154
	Penedès	147
	Jumilla	143

Clase	Región	Vino
Blanco	Rueda	300
	Rías Baixas	216
	Penedès	194
	Rioja	184
	Castilla y León	73
	Empordà	66
	Cataluña	57
	Ribeiro	57
	Valdeorras	55
	Costers del Segre	50

Clase	Región	Vino
Rosado	Rioja	81
	Penedès	31
	Navarra	30
	Ribera del Duero	23
	Cataluña	20
	Empordà	17
	Castilla y León	11
	Cigales	11
	Montsant	11
	Somontano	11

Clase	Región	Vino
Espumoso	Cava	447
	Penedès	79
	Cataluña	33
	Valencia	10
	Castilla y León	8
	Castilla	6
	Vino de España	4
	Empordà	4
	Costers del Segre	4
	Rueda	3

Sello de Calidad / Denominación

Éste es un factor íntimamente ligado con la región.

Empezamos analizando qué regiones, de las incluidas en el top15, tienen sello de calidad:

		Vino
Región	**Sello_calidad**	
Rioja	Denominación de Origen	1871
Ribera del Duero	Denominación de Origen	1153
Priorato	Denominación de Origen	556
Penedès	Denominación de Origen	451
Cava	Denominación de Origen	447
Rueda	Denominación de Origen	306
Castilla y León	Sin denominación	269
Bierzo	Denominación de Origen	250
Empordà	Denominación de Origen	241
Rías Baixas	Denominación de Origen	237
Toro	Denominación de Origen	235
Montsant	Denominación de Origen	233
Cataluña	Denominación de Origen	188
Navarra	Denominación de Origen	164
Jumilla	Denominación de Origen	157

Todas las regiones menos una (CyL) tienen como sello "Denominación de Origen"; lo cual, no debería sorprendernos, primero, porque es el sello mayoritario y el de más antigüedad y, segundo, porque, por lo general, estamos hablando de las regiones con más tradición vitivinícola de España.

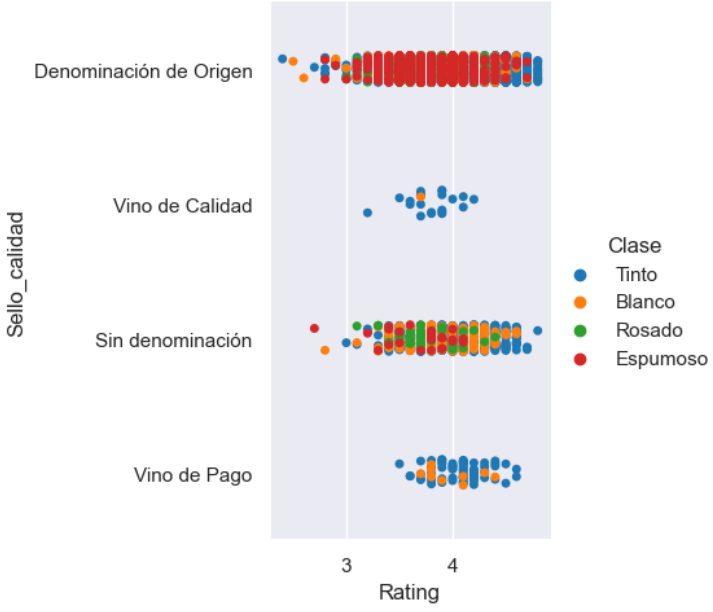
Empero, sí que es sorprendente la particularidad de la región de Castilla y León; ya que, algunas de las D.O más famosas se hallan dentro de esta comunidad y, por lo que vemos, aun así, existen más de 250 vinos que se quedan sin denominación por adscribirse a esta región que permanece sin sello (al contrario de lo que ocurre con Cataluña teniendo ésta, a su vez, otras D.O famosas en su territorio: Priorato, Cava, Penedés…)

Respecto a los otros sellos de calidad ("Vinos de Pago" y "Vinos de Calidad"), éstas son las regiones que los ostentan y la cantidad de vinos que cada una tiene dentro del *dataset*:

		Vino
Sello_calidad	**Región**	
Vino de Pago	Dominio de Valdepusa	28
	Dehesa del Carrizal	13
	El Terrerazo	12
	Arínzano	7
	Pago Calzadilla	6
	Otazu	4
	Campo de la Guardia	4
	Pago Florentino	3

		Vino
Sello_calidad	**Región**	
Vino de Calidad	Islas Canarias	13
	Sierra de Salamanca	6
	Cangas	2
	Granada Sur-Oeste	1

Por último, comprobamos que el distintivo de calidad nada tiene que ver con la calidad apreciada por los usuarios (rating); por lo que se puede deducir que el sello guarda una relación de índole administrativa/política.



Bodega

Para finalizar con el análisis del mercado del vino español, contaremos con otro dato: la bodega de procedencia de los vinos.

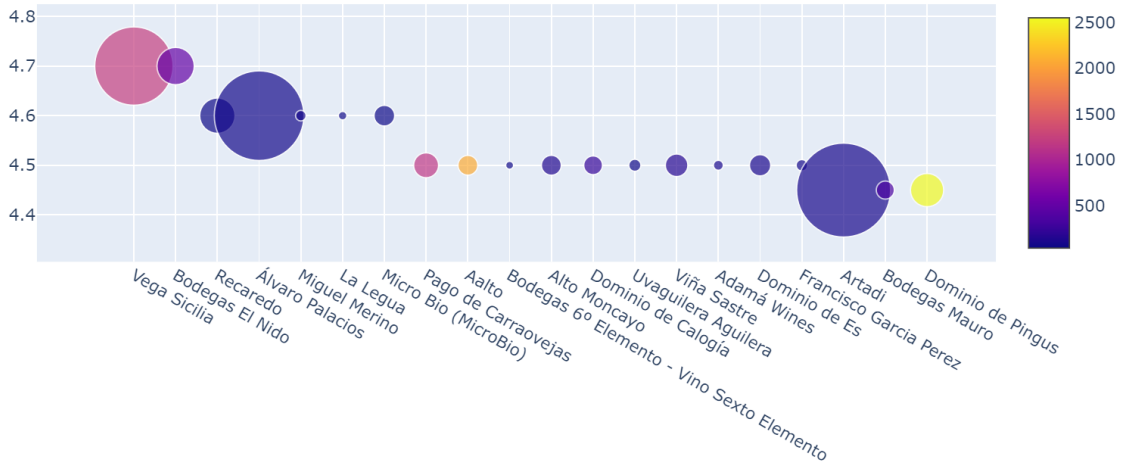
En el *dataset* aparecen 1.666 bodegas distintas, siendo la más prolífica la denominada "Familia Torres" con un total de 96 vinos diferentes. La última bodega en el ranking de las 20 con mayor número de vinos tiene un total de 35 vinos. Basándonos en esto, podemos decir que, generalmente, en España, priman las producciones vinícolas reducidas procedentes de bodegas familiares.

Sin embargo, existen excepciones a lo anterior, pues, tras comprobar las bodegas por regiones, se deduce que hay algunas implantadas en distintas regiones.

Por otra parte, se realiza otro ranking ordenando las bodegas según su *rating* (la mediana de todos sus vinos). En este caso, llama la atención la escasez de bodegas procedentes de la Rioja, dada su enorme penetración en el mercado; por el contrario, Ribera del Duero cuenta casi el 50% de las bodegas del mencionado ranking.

Las bodegas con un *rating* (mediana) alto nos da una noción sobre la fama de la misma; por lo que, a priori, un vino poco conocido pero amparado por un nombre de bodega de calidad, puede antojarse recomendable.

En el siguiente gráfico podemos observar los resultados del ranking de bodegas: aparte del nombre y su calificación, el tamaño de las burbujas nos indica el precio mediano de sus vinos, y el color, el número de valoraciones de usuarios recibidas la web de Vivino.



MEJORES VINOS

Una vez analizado el mercado, pasamos a valorar 3 factores de los vinos que nos servirán como filtros de nuestro *dataset* para poder así obtener un máximo de 10 vinos por cada clase, de modo que, de entre ellos, el usuario pueda elegir en base a su propio criterio/sesgo/gusto. Dichos factores son:

1. Calidad
2. Precio
3. Fama y popularidad

Calidad

Para analizar la calidad, nos apoyaremos en el *rating* que poseen los vinos en la página de Vivino.

Comenzamos viendo la distribución de densidad y comprobando la existencia de *outliers*. En esta ocasión, los *outliers* son por debajo del mínimo (además son muy pocos: 19); por tanto, debido a la importancia de esta variable, se desechan todos esos vinos.

A continuación, desagregamos los datos de *rating* y procedemos al primer filtrado basándonos en 2 referencias:

I. Rating → El rating de los vinos deberá ser mayor a la mediana de su clase
II. Consistencia de la valoración → Para considerar sólido el rating del vino, éste deberá fundamentarse en un mínimo de reseñas. Este mínimo será mayor al percentil25.

Precio

Lo primero que hay que destacar sobre esta variable en nuestro *dataset* es el gran rango que presenta, ya que los precios varían desde los 2€ hasta los casi 500€. Por ello, resulta más apropiado observar las distribuciones de precios por clases; lo que nos demuestra, que la mayor varianza en los precios de los tintos se corresponde con una base de la distribución de la densidad más amplia.

Por otra parte, se constata la existencia de muchísimos *outliers* por encima del máximo; hecho que puede explicarse atendiendo a la exclusividad del sector (y es que el mercado del vino puede llegar a ser muy *gourmet*, de ahí los precios de casi 500 euros una botella de vino).

El segundo filtrado consistirá en eliminar estos *outliers*. Así pues, estableceremos un precio máximo que dependerá de la distribución de precios por cada clase de vino:

$$máx.\ precio = q3 + 1.5*IQR$$

Popularidad y Fama

Para finalizar nuestro análisis y filtrado de datos, tratamos ahora la popularidad y la fama.

Evidentemente, tanto una como la otra son valoraciones subjetivas y aunque ninguno de estos 2 factores te asegura un buen producto, bien es sabido que, a la hora de elegir vino para un evento o para un regalo, tanto la región de procedencia como la bodega pueden predisponer un trato de favor por parte del público.

Por tanto, partiendo de datos objetivos, como el número de valoraciones o el rating de bodegas, intentaremos llegar a una valoración subjetiva acerca de la fama y popularidad de los vinos.

El tercer filtrado se realiza atendiendo a la popularidad de las regiones según las clases de vino; estimándose dicha popularidad en función del número de reseñas. Luego, la mediana de reseñas de las regiones de cada clase (tinto, blanco, etc.) deberá ser superior a la mediana de reseñas de los vinos de esa clase.

El cuarto y último filtro se fundamenta en la fama de las bodegas; para lo cual, nos fijaremos en el *ranking* de las 20 mejores bodegas para cada clase de vino según su *rating*.

Ahora que tenemos nuestro *dataset* separado en vinos tintos, blancos, rosados y espumosos, y filtrado de acuerdo con la calidad, precio, popularidad y fama; elaboramos 2 grupos:

I. Los 20 primeros vinos ordenados por precio (de menor a mayor)
II. Los 20 primeros vinos ordenados por *rating* (de mayor a menor)

Finalmente, con la intersección de estos 2 conjuntos, obtenemos los tops para cada clase de vino:

TOP vino tinto

	Bodega	Año	Vino_ID	Vino	Rating	Nº_reseñas	Precio	Región	Categoría	Sello_calidad
1	Pago de Carraovejas	2016	1139434	Tinto 2016	4.5	7623	29.80	Ribera del Duero	Reserva	Denominación de Origen
2	Bodegas El Nido	2016	1219218	Clio 2016	4.5	2259	31.75	Jumilla	Reserva	Denominación de Origen
3	Bodegas El Nido	2015	1219218	Clio 2015	4.5	2756	33.20	Jumilla	Reserva	Denominación de Origen
4	Alto Moncayo	2017	98782	Garnacha 2017	4.5	1567	33.50	Campo de Borja	Crianza	Denominación de Origen

TOP vino blanco

	Bodega	Año	Vino_ID	Vino	Rating	Nº_reseñas	Precio	Región	Categoría	Sello_calidad
1	Veigamoura	2013	2380009	Blanco 2013	4.5	45	14.96	Rías Baixas	Sin determinar	Denominación de Origen
2	Belondrade	2021	2929881	Quinta Apolonia 2021	4.0	72	15.70	Castilla y León	Sin determinar	Sin denominación
3	Rafael Palacios	2021	1620143	Louro Godello 2021	4.2	647	17.48	Valdeorras	Sin determinar	Denominación de Origen
4	Allende	2017	77131	Rioja Blanco 2017	4.0	249	18.90	Rioja	Sin determinar	Denominación de Origen
5	Tomás Postigo	2018	1553003	Fermentado En Barrica Blanco 2018	4.3	221	20.25	Castilla y León	Sin determinar	Sin denominación
6	Ossian	2019	6142915	Viñas Viejas Verdejo 2019	4.3	657	25.61	Castilla y León	Sin determinar	Sin denominación
7	Ossian	2018	6142915	Viñas Viejas Verdejo 2018	4.3	1002	25.61	Castilla y León	Sin determinar	Sin denominación
8	Remírez de Ganuza	2019	3183419	Rioja Blanco 2019	4.4	229	26.50	Rioja	Sin determinar	Denominación de Origen

TOP vino rosado

	Bodega	Año	Vino_ID	Vino	Rating	Nº_reseñas	Precio	Región	Categoría	Sello_calidad
1	Sinforiano	2021	1563481	Sinfo Rosado 2021	4.1	43	4.25	Cigales	Sin determinar	Denominación de Origen
2	Muga	2020	7490	Muga Rosado 2020	3.8	876	7.30	Rioja	Sin determinar	Denominación de Origen
3	Altanza	2020	6262054	Alma Bohemia Rosé 2020	4.1	43	7.95	Rioja	Sin determinar	Denominación de Origen
4	Sinforiano	2021	5610625	Quelías Rosé 2021	4.1	77	7.99	Cigales	Sin determinar	Denominación de Origen
5	Muga	2021	7490	Muga Rosado 2021	3.9	406	8.75	Rioja	Sin determinar	Denominación de Origen
6	Muga	2021	5513425	Flor de Muga Rosado 2021	4.1	109	17.82	Rioja	Sin determinar	Denominación de Origen
7	Cillar de Silos	2020	5959277	Dominio del Pidio Rosado 2020	4.1	64	19.95	Ribera del Duero	Sin determinar	Denominación de Origen
8	Muga	2019	5513425	Flor de Muga Rosado 2019	4.1	323	19.95	Rioja	Sin determinar	Denominación de Origen
9	Jose Luis Ripa	2017	8643220	Rosado 2017	4.1	56	20.90	Rioja	Sin determinar	Denominación de Origen

TOP vino espumoso

	Bodega	Año	Vino_ID	Vino	Rating	Nº_reseñas	Precio	Región	Categoría	Sello_calidad
1	Juvé & Camps	2013	1622570	Cava Gran Reserva Brut 2013	4.1	153	4.94	Cava	Sin determinar	Denominación de Origen
2	Miquel Pons	2015	1394624	Cava Gran Reserva Vintage Brut Nature 2015	4.1	63	10.90	Cava	Vintage	Denominación de Origen
3	Guilera	2010	1851083	Gran Reserva Brut Nature 2010	4.1	57	12.50	Cava	Sin determinar	Denominación de Origen
4	Pago de Tharsys	2018	1295891	Cava Millésime Brut Reserva 2018	4.2	54	15.50	Cava	Sin determinar	Denominación de Origen
5	Dominio de la Vega	2015	1454604	Cava Reserva Especial Brut 2015	4.1	171	18.11	Cava	Sin determinar	Denominación de Origen
6	Dominio de la Vega	2017	1454604	Cava Reserva Especial Brut 2017	4.1	104	18.11	Cava	Sin determinar	Denominación de Origen
7	Gramona	2016	1199954	Imperial Brut 2016	4.1	947	18.85	Cava	Sin determinar	Denominación de Origen
8	Mestres	2013	2181578	Visol Gran Reserva Brut Nature 2013	4.2	250	20.00	Cava	Sin determinar	Denominación de Origen
9	Juvé & Camps	2016	9125117	Cava Gran Reserva Brut Nature Singular Xarel-l...	4.1	94	21.50	Cava	Sin determinar	Denominación de Origen
10	Hispano Suizas	2016	1797114	Cava Tantum Ergo Pinot Noir Brut Nature 2016	4.1	97	22.00	Cava	Sin determinar	Denominación de Origen