11.67 Estadística Aplicada

Primer Entrega

Azul de los Ángeles Makk

Paula Ariana González

Fecha de entrega: 22.03.2023

Cosatto Ammann, Pedro Camilo

Índice

Selección de datos y limpieza

Para el presente trabajo práctico hemos decidido estudiar la base de datos que muestra la calidad de distintos tipos de vinos tintos llamada “winequality-red.cvs”. La misma contiene observaciones de vinos de la variedad Vinho Verde, elaborada en Portugal. La base de datos seleccionada posee 12 columnas: acidez fija, acidez volátil, acidez cítrica, azúcar residual, cloruro, sulfuro de dióxido neto, sulfuro de dióxido total, densidad, ph, sulfatos, alcohol y calidad.

Para el alcance de este trabajo se seleccionaron únicamente las columnas de acidez fija, densidad, alcohol y ph, siendo las cuatro variables numéricas. Adicionalmente, a modo de realizar un análisis categórico tomamos la variable ‘quality’, el cual le asigna un valor numérico a la calidad del vino. Para poder profundizar en el comportamiento de las observaciones de cada categoría, seleccionamos las que se le asignan los números 4, 6 y 8 a modo de reflejar una calidad baja, media y alta respectivamente. La muestra aleatoria tomada de un total de 709 observaciones, es de un total de 500 muestras -sin reposición-.

Identificamos que la base de datos se trata de una muestra transversal ya que se basa en la observación de los vinos al mismo tiempo. Su obtención fue mediante el sitio web Kaggle (véase bibliografía).

Análisis Descriptivo de datos numéricos

En esta sección analizaremos cada una de las 4 variables seleccionadas para poder entender mejor cómo se comportan. Para realizar el análisis utilizamos R en RStudio y para producir los gráficos instalamos el paquete *"tidyverse"* que viene con la librería de *GGPLOT2*. Dicha librería fue posteriormente empleada para generar todas las visualizaciones disponibles en las figuras adjuntas, de manera en la que se dispongan de manera prolija y personalizada.

1. fixed.acidity

F = Acidez fija de un vino Vinho Verde.

La variable F mide la suma de todos aquellos ácidos que, al someter el vino al calor, no se evaporan. Primeramente, pudimos observar que la mayor acidez que se hallada en la muestra fue de 14.3 gramos, mientras que la menor fue de 4.7 gramos, otorgando un rango de 9.6 gramos. A fines de poder observar mejor a la variable, graficamos la función de distribución empírica que se puede observar en la *Figura I*.

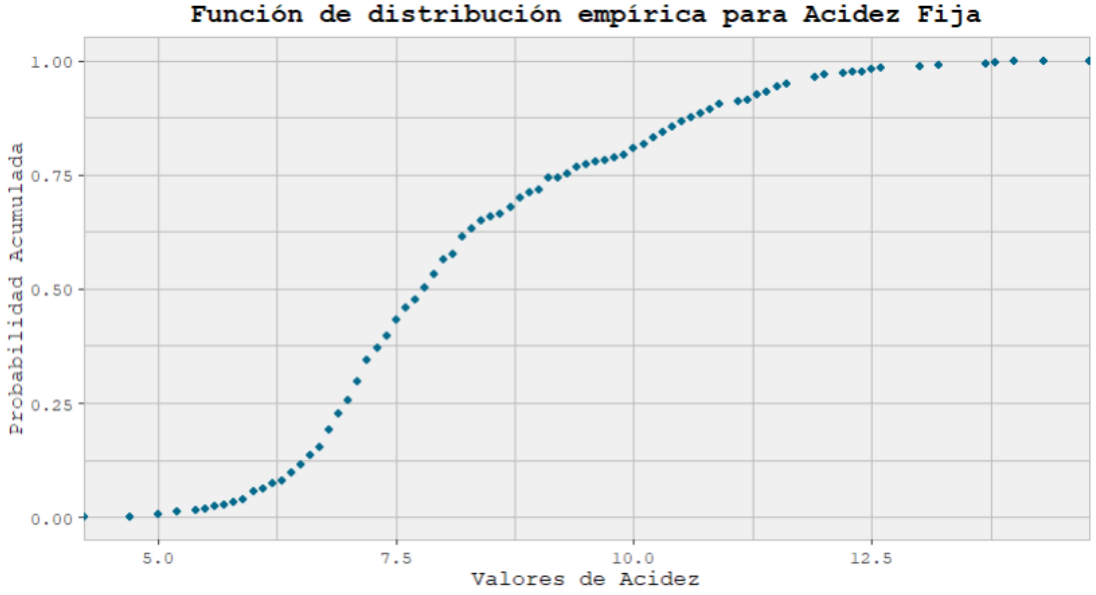

Figura I

Podemos observar en el gráfico que como los puntos son muy próximos unos de otros, esta variable es contínua. Por esta razón, graficamos un

histograma (véase *Figura II*) para poder distinguir la distribución de la variable F.

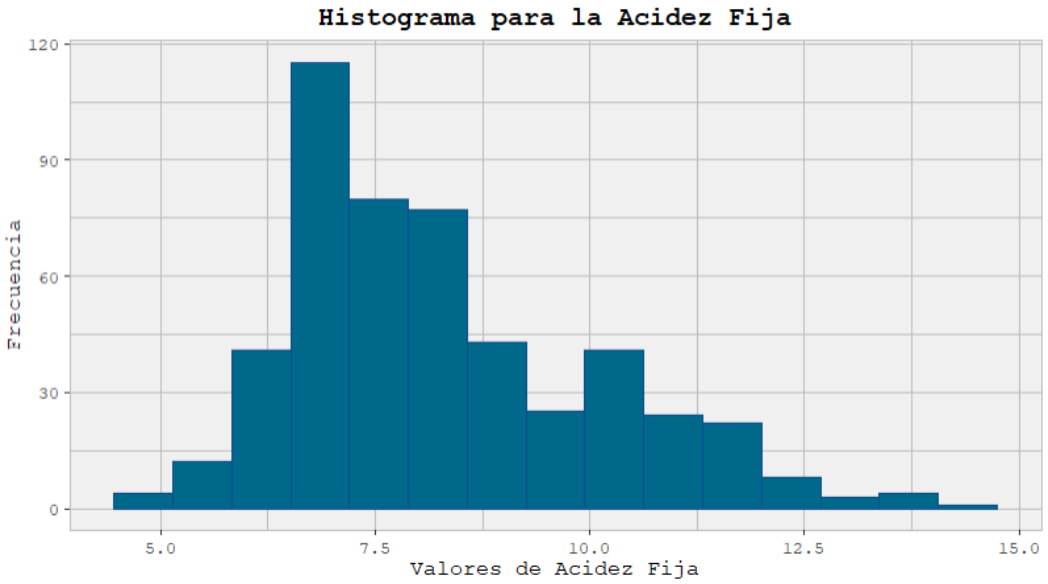

Figura II

A su vez, graficamos la función de densidad (véase *Figura III*), y pudimos observar mejor que en el histograma que hay un pico de densidad cercano a x = 7 y luego va disminuyendo progresivamente hasta aproximadamente x = 10.5, donde hay un nuevo pico y vuelve a bajar.

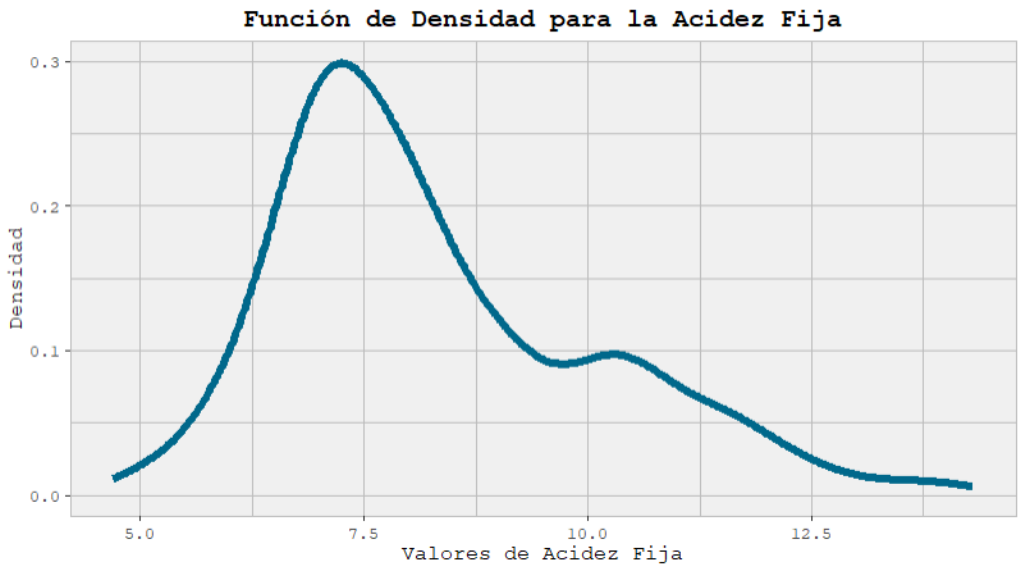

Figura III

Luego, estimamos la media muestral de $\bar{f}$=8.2718 y la mediana de $f_{0.5}$=7.9, de manera que al ser mayor la mediana que la media podemos decir que la distribución tiene asimetría positiva -luego respaldado en Tabla II-. Esto se puede confirmar en el boxplot de la *Figura IV*. Es este gráfico también se pueden observar un par de outliers -marcados en negro en la parte superior-, y los cuantiles son:

Cuantil	Expresión matemática	Resultado numérico
0.25	$f_{0.25} = F^{-1}(0.25) = min\, f_i\, tq\{f_i: \widehat{F}(f_i) \geq 0.25\}$	7.0
0.5	$f_{0.5} = F^{-1}(0.5) = min\, f_i\, tq\{f_i: \widehat{F}(f_i) \geq 0.5\}$	7.9
0.75	$f_{0.75} = F^{-1}(0.75) = min\, f_i\, tq\{f_i: \widehat{F}(f_i) \geq 0.75\}$	9.3

Tabla I

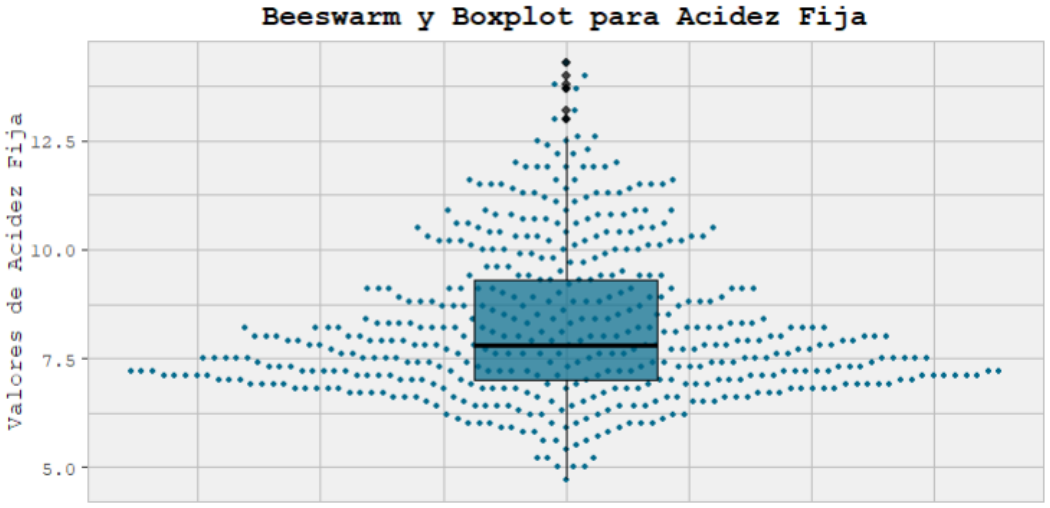

Figura IV

Cómo se puede ver en la *Figura IV*, hay una presencia de valores atípicos (los puntos de color negro, que muestran la cantidad de outliers) en la cantidad de acidez del vino pero decidimos no excluirlos ya que no son significativos ya que no alteran significativamente a las medidas no robustas con respecto a las robustas. Asimismo, realizamos un gráfico para visualizar si hay cambios en la distribución si se eliminan los outliers y el resultado fue que las curvas de las distribuciones se

superponen, por lo que consideramos que no es un gráfico que aporte tanto valor al informe y nos confirma que no es necesario eliminar los valores atípicos.

Para finalizar el análisis de esta variable, armamos la siguiente *Tabla II* para visualizar un breve resumen estadístico de la variable de acidez del vino. De todos los valores, podemos destacar que el coeficiente de curtosis es mayor a 3, eso quiere decir que la distribución de la variable es de colas pesadas.

Resumen estadístico para la *fixed.acidity*		
Media muestral	$\overline{f} = \frac{\sum_{i=1}^{n} F_i}{N}$	8.2718
Desvío estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (F_i - \overline{F})^2}{n}$	1.7923
Coeficiente de variación	$r = \frac{\overline{V}(f)}{\lvert\overline{f}\rvert}$	0.2166
Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\Sigma(\frac{f_j - \overline{f}}{\sqrt{\overline{V}(f)}})^3}{n}$	0.8181
Coeficiente de curtosis muestral	$\hat{k} = \frac{\Sigma(\frac{f_j - \overline{f}}{\sqrt{\overline{V}(f)}})^4}{n}$	3.2024

Tabla II

2. density

D = Densidad de un vino Vinho Verde (gr/ml)

La variable D mide la densidad del vino tinto en gramos por militro. El vino que registró menor densidad fue de 0.9901 y el de mayor valor registrado de 1.0037, presentando un rango total de 0.01361. A continuación, en la *Figura V*, se observa su gráfico de la función de distribución empírica.

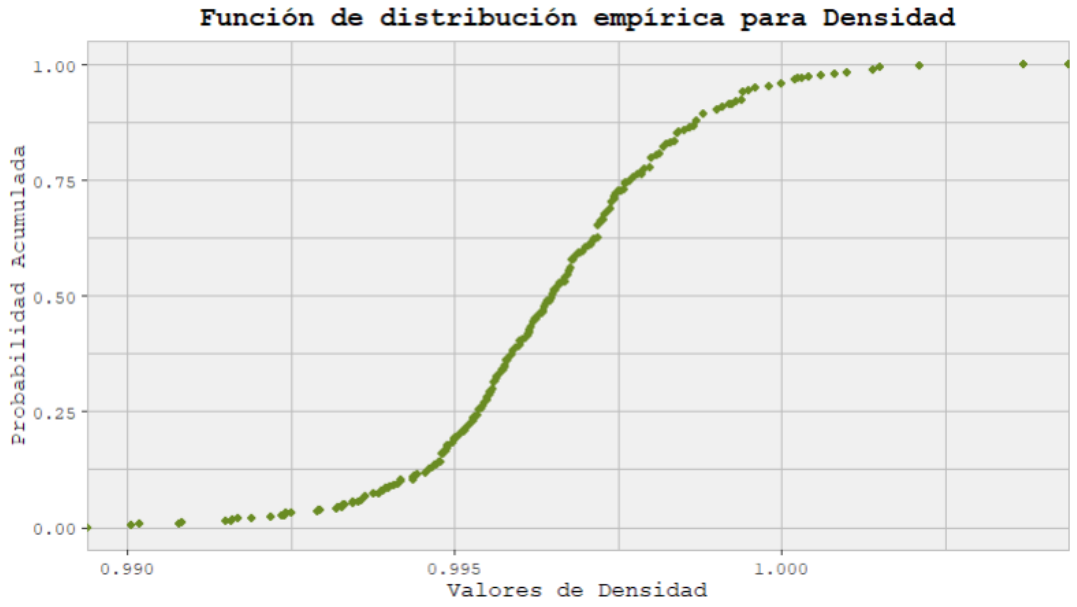

Figura V

Se podría decir que la variable D se comporta de manera aleatoria contínua. También podemos visualizar que la gran mayoría de los puntos se encuentran concentrados en el centro de la distribución, cercano a la media. Procedemos a graficar su histograma y la función de densidad correspondiente (véase *Figura VI* y *Figura VII*) para ver mejor cuál es la forma de la distribución.

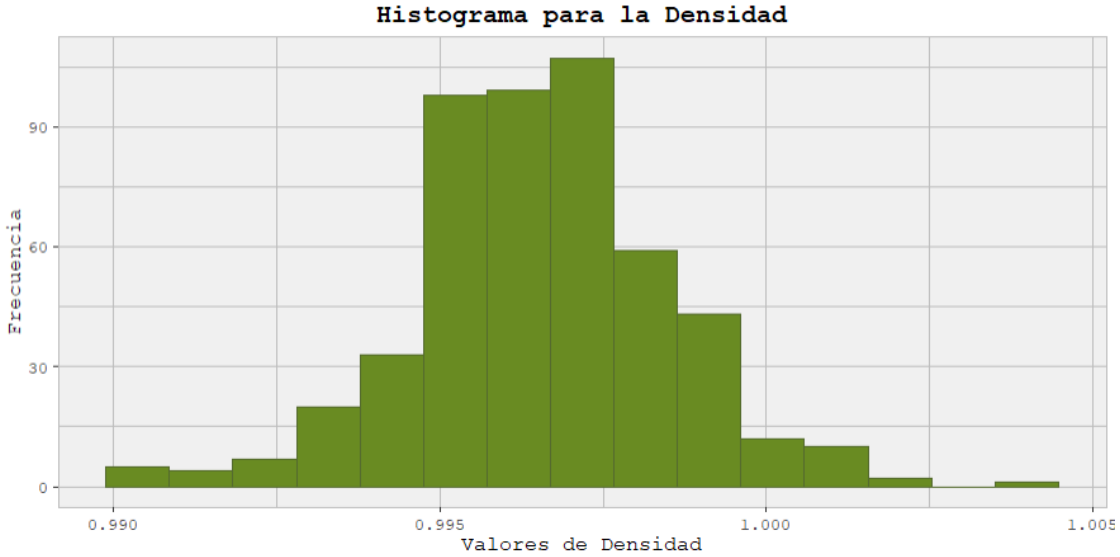

Figura VI

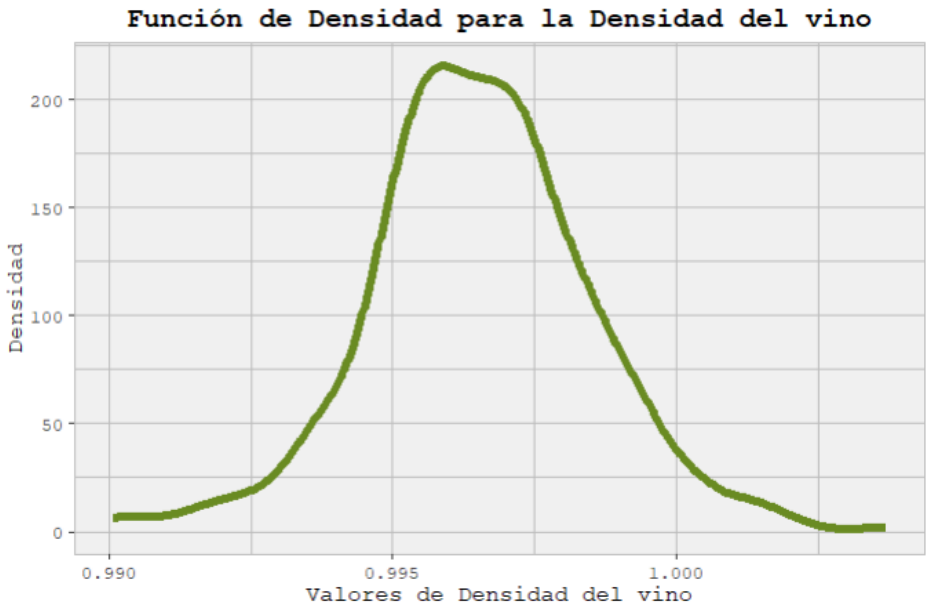

Figura VII

La *Figura VII* muestra que la curva se parece a una campana de Gauss (con algunas irregularidades). Luego, con su posterior cálculo identificamos que la media y la mediana están representadas por el mismo valor ($\bar{d} = d_{0.5}$ =0.9965), lo que nos muestra que la distribución es simétrica y posee normalidad. Esto se puede confirmar en el boxplot de la *Figura VIII*. En este gráfico también se pueden observar algunos outliers de ambos extremos y la delimitación de los cuantiles (representados también en *Tabla III*).

Cuantil	Expresión matemática	Resultado numérico
0.25	$d_{0.25} = F^{-1}(0.25) = min\, d_i\, tq\{d_i: \widehat{F}(d_i) \geq 0.25\}$	0.9953
0.5	$d_{0.5} = F^{-1}(0.5) = min\, d_i\, tq\{d_i: \widehat{F}(d_i) \geq 0.5\}$	0.9965
0.75	$d_{0.75} = F^{-1}(0.75) = min\, d_i\, tq\{d_i: \widehat{F}(d_i) \geq 0.75\}$	0.9977

Tabla III

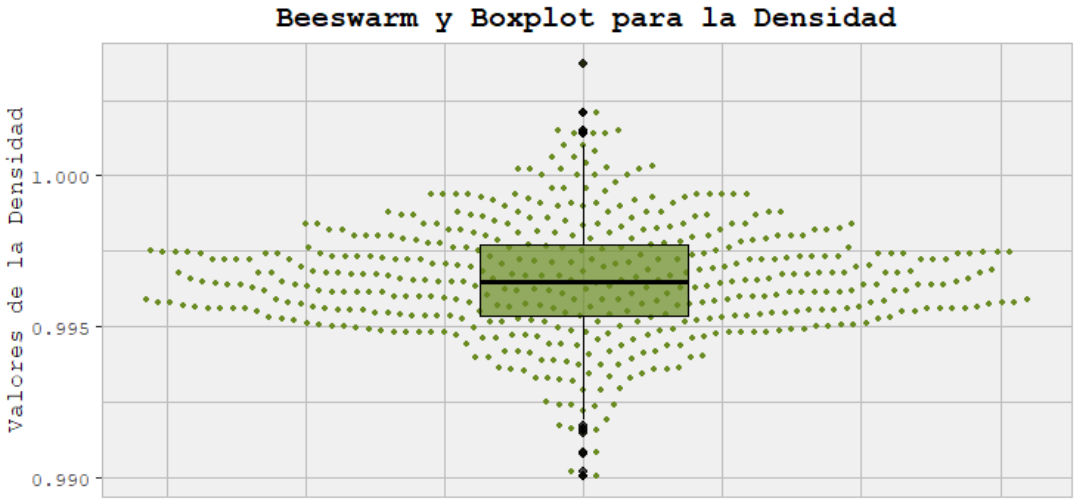

Figura VIII

Como mencionamos previamente, en la parte superior e inferior del gráfico es observable la presencia de outliers -marcados en color negro-. Hemos tomado la decisión de no separar tales valores de la muestra al no resguardar una significativa distancia con los límites del rango intercuartílico. De todas formas, a modo de evaluar el impacto de la inclusión de estos datos y su posible eliminación, en la *Figura IX*, se puede observar cómo se altera su distribución en ambos escenarios. De esta forma, es posible afirmar que no es sustancial su diferenciación y no altera significativamente la manera en la que esta se distribuye.

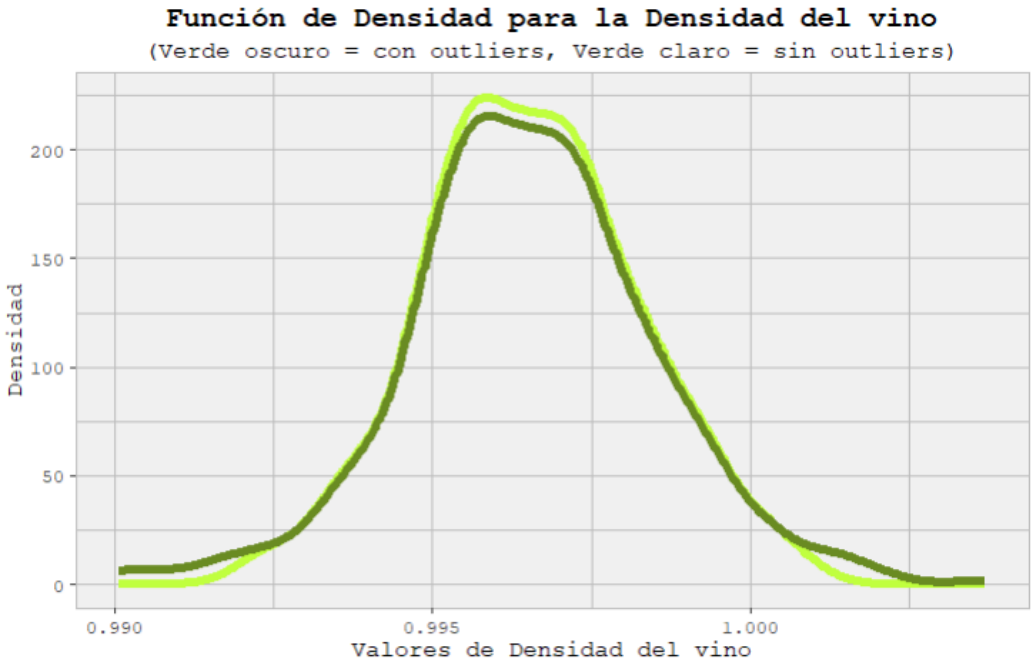

Figura IX

Para finalizar el análisis de esta variable, confeccionamos la *Tabla IV* a fines de visualizar un breve resumen estadístico de la variable de la densidad del vino. De esta tabla se puede destacar, nuevamente, que el coeficiente de curtosis nos dió mayor a cero, lo cual nos confirma que esta variable aleatoria tiene una distribución de colas pesadas como habíamos supuesto anteriormente y presenta una asimetría levemente negativa.

Resumen estadístico para *density*		
Media muestral	$\bar{d} = \frac{\sum_{i=1}^{n} D_i}{N}$	0.9965
Desvío estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n}$	0.0019
Coeficiente de variación	$r = \frac{\bar{V}(d)}{\lvert\bar{d}\rvert}$	0.0019
Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\sum(\frac{d_j - \bar{d}}{\sqrt{\bar{V}(d)}})^3}{n}$	-0.0383
Coeficiente de curtosis muestral	$\hat{k} = \frac{\sum(\frac{d_j - \bar{d}}{\sqrt{\bar{V}(d)}})^4}{n}$	5.8559

Tabla IV

3. alcohol

A = Porcentaje de alcohol de un vino Vinho Verde.

La variable alcohol mide la graduación alcohólica del vino. La misma se expresa en grados y mide el contenido de alcohol absoluto en, es decir, el porcentaje de alcohol que esta posee. A modo de representación, obsérvese en la *Figura X* su función empírica. Al ver cómo se comportan los puntos, se podría decir que A es una variable aleatoria contínua.

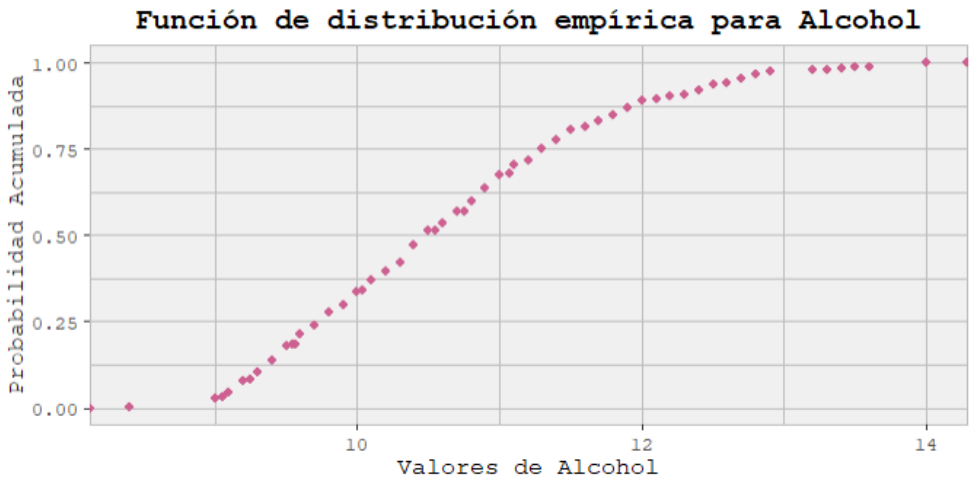

Figura X

Los valores registrados de tal variable son, como mínimo 8.4 y como máximo 14, presentando un rango de 5.6. Véase *Figura X* para observar en un histograma la manera en la que los datos se distribuyen.

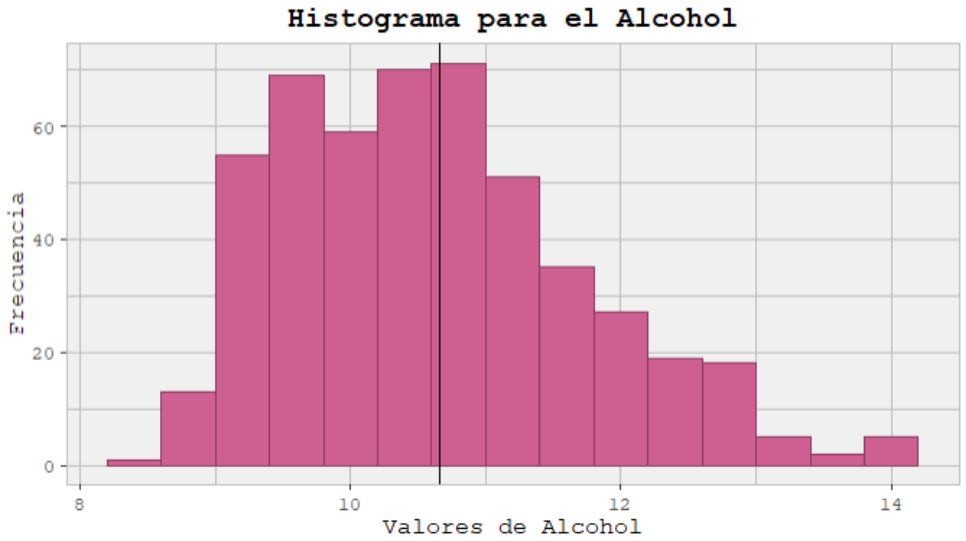

Figura XI

Podemos observar una amplia concentración de los datos cerca de la media (marcada con una línea vertical negra), prácticamente no teniendo agrupación o un significativo peso en sus colas, especialmente a la izquierda. Esto quiere decir que la distribución tiene asimetría positiva. A su vez, obsérvese en la *Figura XII* el diagrama de densidad de dicha variable. Nótese en aquel, un predominante aumento en la densidad,

a medida que se acerca a x=9.5 y una progresiva baja a partir de x=10.5 hasta x=13.

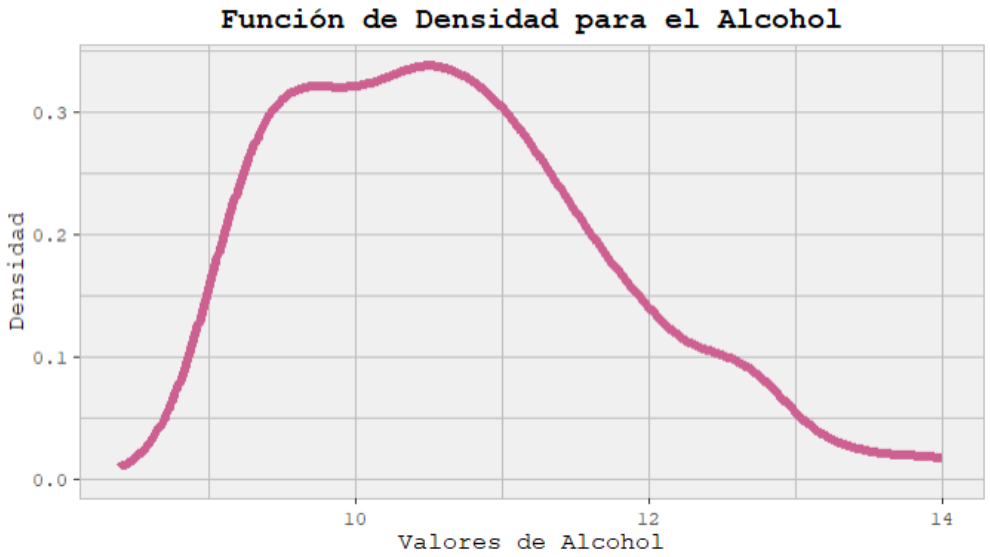

Figura XII

Continuamos luego con el cálculo de la media y la mediana. Ambas otorgan un valor cercano entre sí, siendo que la media es de 10.65 mientras que la mediana es de 10,5. Para mayor información de cuartiles, obsérvese la *Tabla V*, la cual muestra los valores numéricos para cuantiles 0.25, 0.5 y 0.75.

Cuantil	Expresión matemática	Resultado numérico
0.25	$a_{0.25} = F^{-1}(0.25) = min\, a_i\, tq\{a_i: \widehat{F}(a_i) \geq 0.25\}$	9.800
0.5	$a_{0.5} = F^{-1}(0.5) = min\, a_i\, tq\{a_i: \widehat{F}(a_i) \geq 0.5\}$	10.500
0.75	$a_{0.75} = F^{-1}(0.75) = min\, a_i\, tq\{a_i: \widehat{F}(a_i) \geq 0.75\}$	11.325

Tabla V

A continuación se puede visualizar el diagrama de caja y bigotes (véase *Figura XIII*) con un gráfico Bee Swarm. Es observable en tal caso que no existe una predominancia de datos fuera del rango intercuartílico, ubicándose sólo 4 valores atípicos que tienen el mismo valor fuera del rango superior estipulado. Al no ser una distancia muy predominante, no nos resultó de crucial importancia estudiarlo al no alterar

significativamente las medidas no robustas como la media. Por esa razón es que también decidimos no excluirlos de la muestra. Asimismo, realizamos un gráfico para visualizar si hay cambios en la distribución con la eliminación de outliers y el resultado fue que las curvas se superponen, por lo que consideramos que no es un gráfico que aporte tanto valor al informe y nos confirma que no es necesario eliminar los valores atípicos.

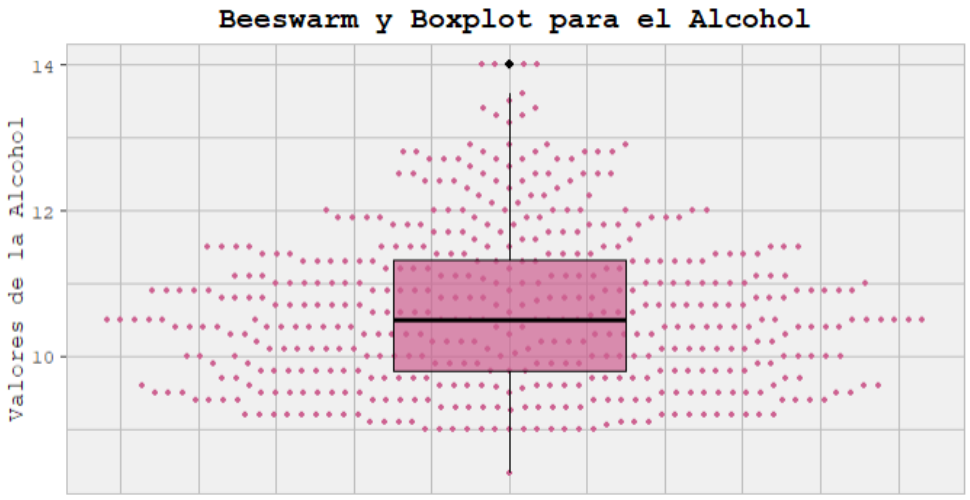

Figura XIII

Finalmente, a continuación en la *Tabla VI* se pueden observar junto a sus expresiones, los valores numéricos para la media muestral, el desvío estándar muestral, el coeficiente de variación, el coeficiente de asimetría muestral y el coeficiente de curtosis muestral. Podemos verificar que la muestra posee una leve asimetría positiva y una curtosis cercana a 3, lo cual sugiere la normalidad de la muestra.

Resumen estadístico para *alcohol*		
Media muestral	$\bar{a} = \frac{\sum_{i=1}^{n} A_i}{N}$	10.659
Desvío estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (A_i - \bar{A})^2}{n}$	1.097
Coeficiente de variación	$r = \frac{\bar{V}(a)}{\lvert\bar{a}\rvert}$	0.102

Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\Sigma(\frac{a_j - \bar{a}}{\sqrt{\bar{V}(a)}})^3}{n}$	0.620
Coeficiente de curtosis muestral	$\hat{k} = \frac{\Sigma(\frac{a_j - \bar{a}}{\sqrt{\bar{V}(a)}})^4}{n}$	2.948

Tabla VI

4. pH

P = Medida de acidez de un vino Vinho Verde.

La variable P describe qué tan *ácido* o *básico* es un vino en una escala del 0 -muy ácido- al 14 -muy básico-, la mayoría de los vinos suelen estar en un rango de 3-4. A modo de representación, obsérvese en la *Figura XIV* la función empírica. Al ver que la mayoría de los datos se encuentran concentrados en la mitad, y simultáneamente adquieren una forma de "s" suave, suponemos que la forma de la distribución adquirirá la forma de una campana de Gauss. Asimismo, podemos decir que la variable P se comporta como una aleatoria contínua.

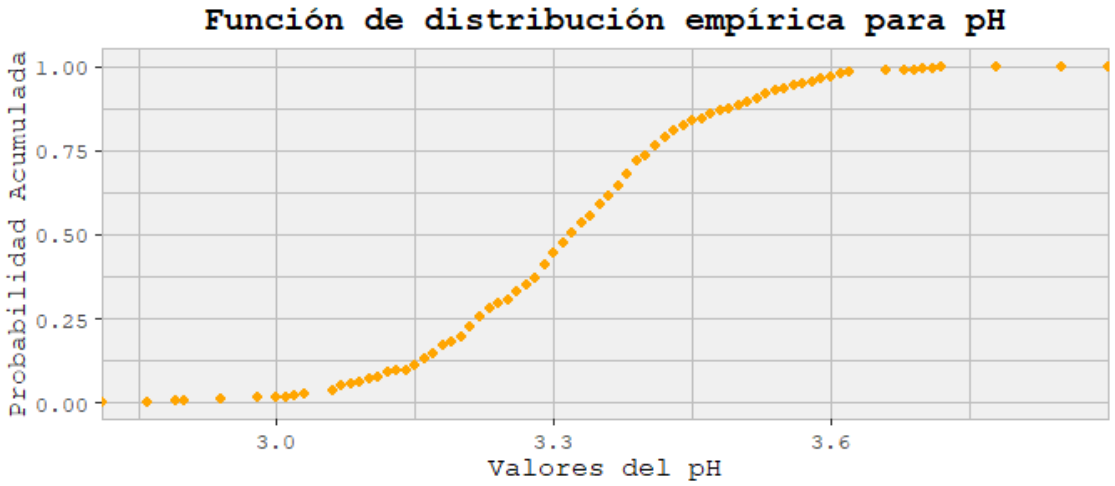

Figura XIV

En la muestra tomada podemos observar un valor mínimo de 2.86 y un máximo de 3.85, presentando un rango total de 0.99. Véase la *Figura XV* para el análisis de su distribución. En el mismo, se puede confirmar nuestra suposición de que la variable P morfológicamente se asemeja a grandes rasgos a una campana de Gauss, presentando una gran concentración

cerca de la media (marcada con una línea vertical negra) y con las colas similares a ambos lados, equitativamente presentando bajo peso.

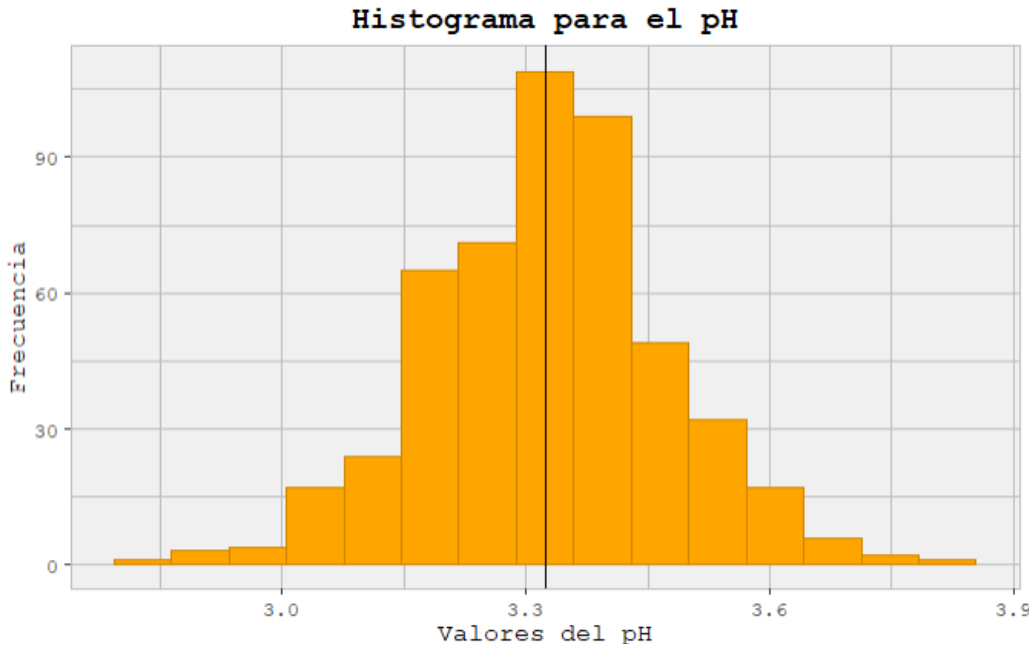

Figura XV

Adicionalmente, al confeccionar la curva de densidad es observable (*Figura XVI*) que la misma coincide con la descripción de normalidad previamente mencionada. Alrededor de la media, existe una notable predominancia de datos.

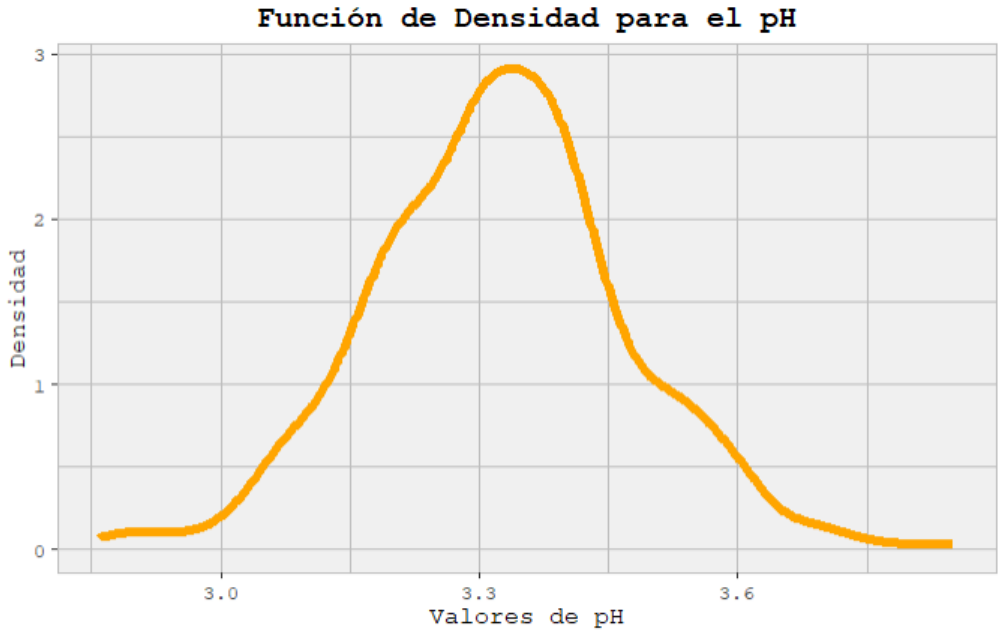

Figura XVI

Asimismo, podemos ver una notable similitud entre la mediana 3.32 y la media 3.3237. Al ser una medida robusta como la mediana tan similar a aquella no robusta, podemos afirmar que hay simetría en la distribución. A continuación en la *Tabla VII* se enlistan los cuantiles 0.25 0.5 y 0.75. También, obsérvese el diagrama de caja y bigotes con un gráfico Bee Swarm (*Figura XVII*).

Cuantil	Expresión matemática	Resultado numérico
0.25	$p_{0.25} = F^{-1}(0.25) = min\, p_i\, tq\{p_i: \widehat{F}(p_i) \geq 0.25\}$	3.22
0.5	$p_{0.5} = F^{-1}(0.5) = min\, p_i\, tq\{p_i: \widehat{F}(p_i) \geq 0.5\}$	3.32
0.75	$p_{0.75} = F^{-1}(0.75) = min\, p_i\, tq\{p_i: \widehat{F}(p_i) \geq 0.75\}$	3.41

Tabla VII

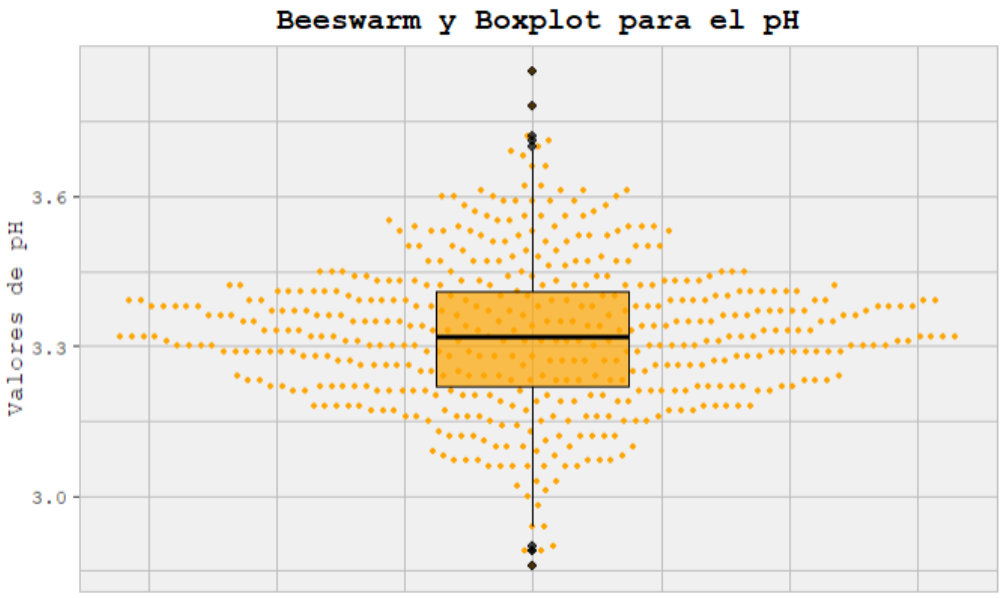

Figura XVII

Es observable en el boxplot una mayor presencia de outliers en relación con otras variables. Tanto en valores mínimos y máximos, podemos identificar valores atípicos. Sin embargo, en su gran mayoría ninguno de

ellos se aleja alarmantemente del rango intercuartílico, por esa razón decidimos no excluirlos de la muestra. Cuando realizamos el gráfico de las distribuciones con y sin los valores atípicos las curvas se superponen, confirmandonos que no es necesario eliminar los outliers ya que no cambia en nada la distribución.

Para finalizar, en la Tabla *XVIII* podemos observar enlistados la media muestral, el desvío estándar muestral, el coeficiente de variación, el coeficiente de asimetría muestral y el coeficiente de curtosis muestral. En ella, podemos identificar un coeficiente de asimetría positiva muy cercano al cero, lo que hace la muestra casi simétrica. A su vez, su curtosis es mayor a 3, lo cual nos confirma nuestra hipótesis de que la distribución es de colas pesadas.

Resumen estadístico para *pH*		
Media muestral	$\bar{p} = \frac{\sum_{i=1}^{n} P_i}{N}$	3.3237
Desvío estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (P_i - \bar{P})^2}{n}$	0.1481
Coeficiente de variación	$r = \frac{\bar{V}(p)}{\lvert\bar{p}\rvert}$	0.0445
Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\Sigma(\frac{p_j - \bar{p}}{\sqrt{\bar{V}(p)}})^3}{n}$	0.0565
Coeficiente de curtosis muestral	$\hat{k} = \frac{\Sigma(\frac{p_j - \bar{p}}{\sqrt{\bar{V}(p)}})^4}{n}$	3.4493

Tabla VIII

Relación entre variables

En esta sección analizaremos las relaciones entre las variables que hemos presentado previamente. A fines de llevar adelante un análisis integral, ejecutamos la matriz de correlación (*Tabla IX*), la cual detalla numéricamente la dependencia lineal entre dos variables. El determinante

de la matriz es de 0.12888 por lo que, sabiendo que a mayor semejanza con el 0 mayor correlación, las variables guardan cierta dependencia lineal.

	fixed.acidity	density	alcohol	pH
fixed.acidity	1.00000000	0.6811300	-0.9589555	-0.7074517
density	0.68113003	1.0000000	-0.51152948	-0.3475526
alcohol	-0.0958955	-0.51152948	1.00000000	0.1753842
pH	-0.70745173	-0.3475526	0.17538420	1.0000000

Tabla IX

Luego, elegimos hacer 4 relaciones entre las variables para ver cómo se ven gráficamente.

En primer lugar, decidimos relacionar la acidez del vino con el pH (*Figura XIX*), lo cual había arrojado una correlación de -0.70745173, lo cual señala que una estrecha relación entre ambas al estar cerca de -1. Tal información es verificada visualmente mediante un scatter plot (*Figura XVIII*), donde se puede ver una tendencia negativa posicionando a la acidez fija en el eje horizontal y el ph en el eje vertical. En consecuencia, concluimos que a mayor acidez fija menor es el pH o, alternativamente, a mayor pH tiende a ser menor la acidez fija.

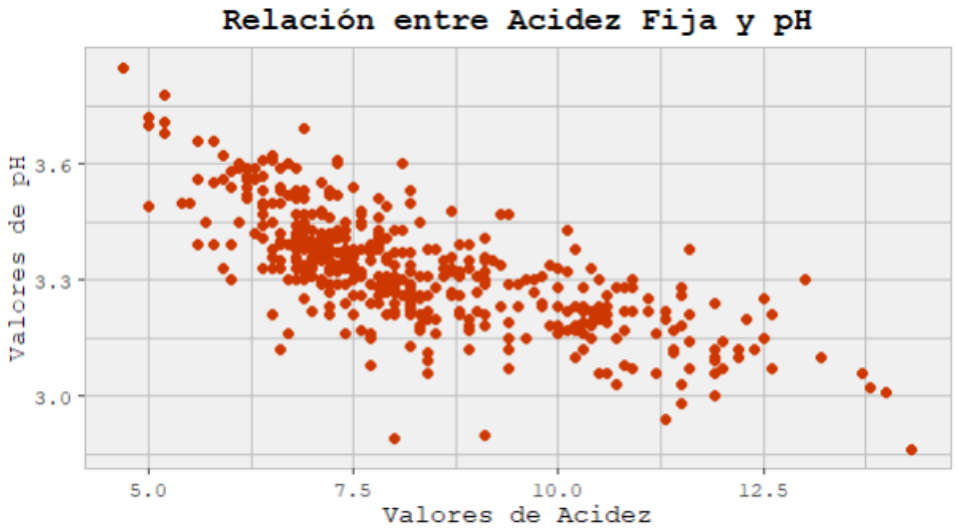

Figura XVIII

Luego, continuación con el mismo procedimiento para analizar la relación entre las variables de densidad de alcohol. Al igual que en la *Figura XIX*, en este nuevo gráfico (*Figura XIX*) podemos ver una clara tendencia decreciente -ubicando la densidad en el eje horizontal y el alcohol en el eje vertical-. Por lo tanto, inferimos que a mayor densidad, menor es el nivel de alcohol y viceversa. En este caso, podemos observar una menor concordancia o menos delimitada esta tendencia, lo cual es correspondido al presentar una concordancia más cercana al 0 que la variable relación anterior, siendo de -0.51152948.

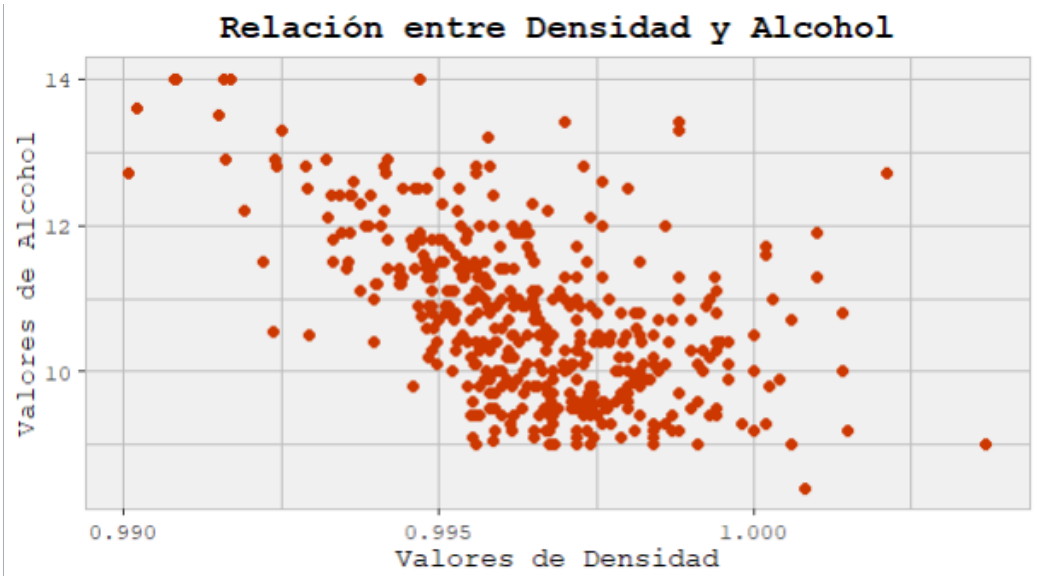

Figura XIX

Nuestro tercer análisis sobre relaciones de variables decidimos observar la conexión entre el alcohol y el ph (*Figura XX*). Previamente en la *Tabla IX* se arroja como resultado numérico una correlación de la *Tabla IX* de 0.1753842, notablemente cercano al 0. Si bien podría marcarse una tendencia ascendente (a mayor alcohol, mayor ph y viceversa), esto no es lo suficientemente diferenciable como para tomarlo como una afirmación válida.

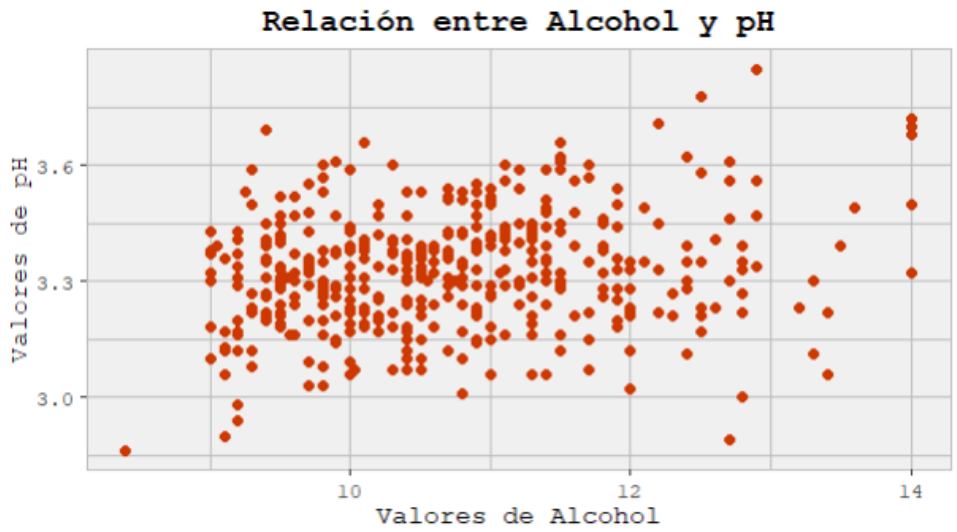

Figura XX

Para finalizar, decidimos evaluar la relación entre la densidad y el ph (*Figura XXI*). Previamente en la *Tabla IV*, el resultado numérico nos muestra que su correlación es de -0.3475526, notablemente cercano al 0. Es sensato entonces, que al observar su visualización, las mismas presentan una leve tendencia negativa (a mayor densidad, menor ph y viceversa), pero es la predominancia de dispersión lo que imposibilita determinarlo.

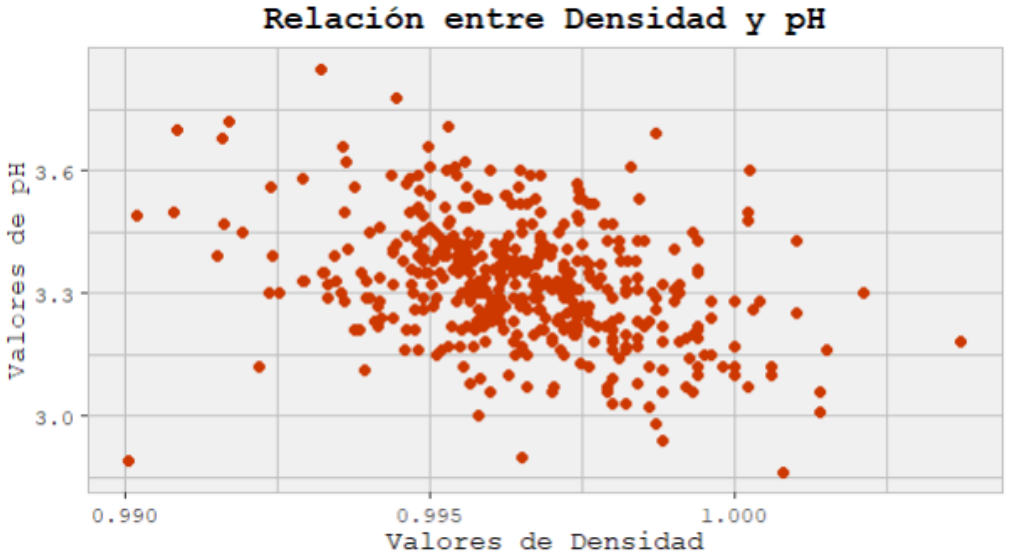

Figura XXI

Estudio de una variable a elección

Para esta sección se eligió la variable numérica del pH y se hizo un análisis gráfico de cómo cambia la distribución de los datos en los distintos niveles de la variable categórica *quality*. Como se ha mencionado anteriormente, la calidad en el dataset seleccionado está representada por un valor numérico. Dentro de ellos decidimos seleccionar aquellos de calidad 4 representando una calidad *baja*, los de calidad 6 para un nivel *medio* y los de calidad 8 para una categorización *alta*.

Primeramente, en la *Figura XXII* podemos observar el histograma de la variable P, filtrado únicamente para la calidad baja. Podemos observar que la misma se distribuye con ciertas irregularidades: en su cola izquierda presenta un inicio, una repentina ausencia de registros y luego un brusco ascenso cercano a la media.

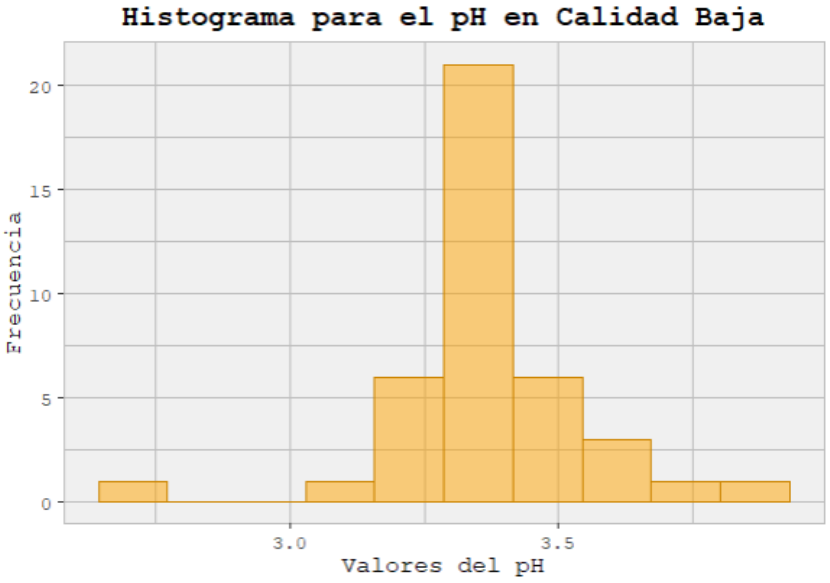

Figura XXII

Continuando con el análisis descriptivo de la variable P. pasamos a estudiar su calidad media (véase Figura XXIII). La misma se distribuye ciertamente con cierto rango de normalidad, a pesar de presentar una predominante curtosis. Posee colas tanto a izquierda como a derecha, siendo la derecha levemente más significativa. De los tres análisis realizados, se infiere que la calidad media es la que adquiere una mayor normalidad, pero a su vez más adelante estudiaremos que es la única que se alimenta de una numerosa cantidad de registros.

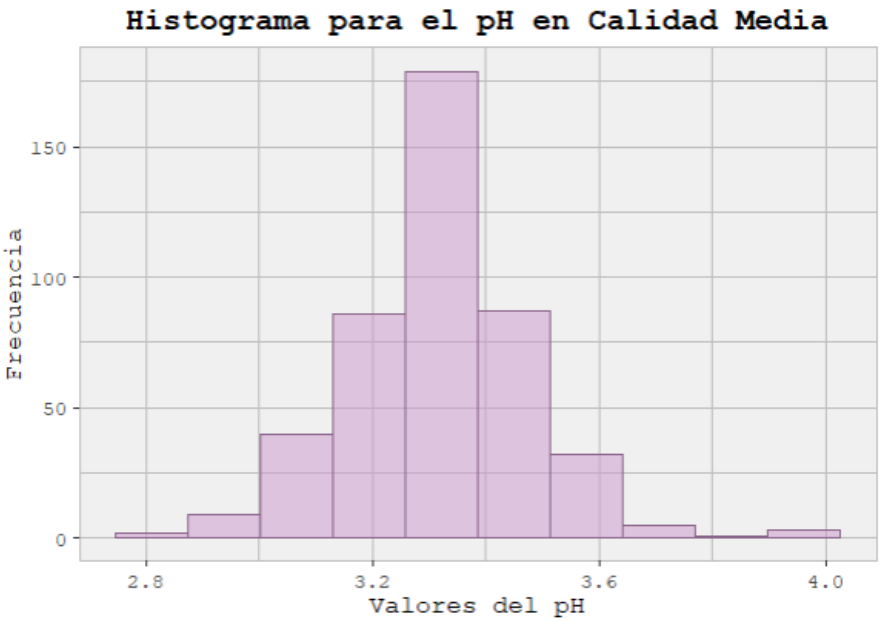

Figura XXIII

Procedemos luego para el análisis de la variable P para la calidad alta (véase Figura XXIV). La misma se distribuye de manera totalmente irregular, con repentinos picos y ausencia de registros en determinados rangos. Dicha inestabilidad, imposibilita el intento se señalar una posible distribución, lo cual no es novedad al tan solo tener un total de 14 observaciones.

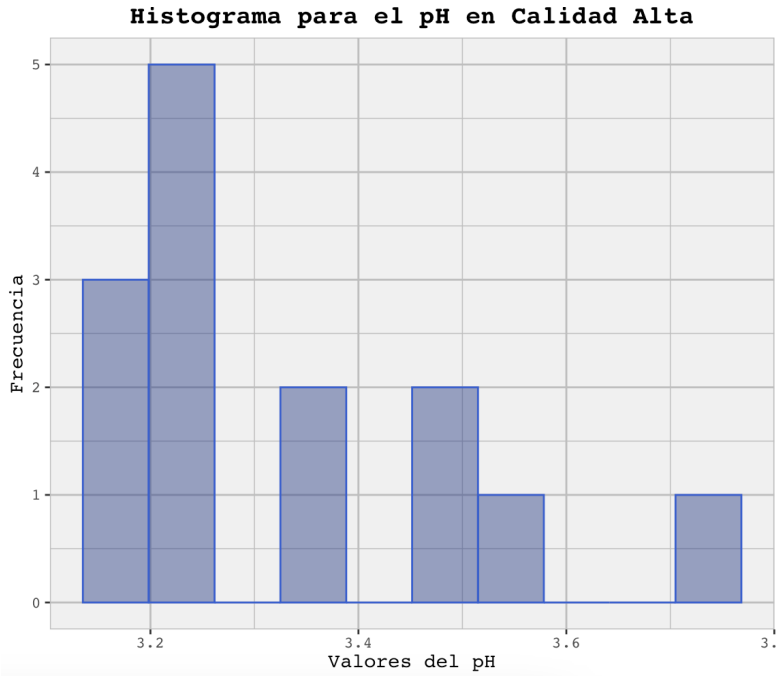

Figura XXIV

Finalmente, decidimos superponer los 3 histogramas mostrados previamente (véase *Figura XXV*) y se puede observar que existe una gran predominancia de observaciones de calidad media en relación a los de calidad baja y alta. A su vez, es posible afirmar que los menores valores de pH que tienen los vinos de calidad alta son mayores que los valores de las otras calidades.

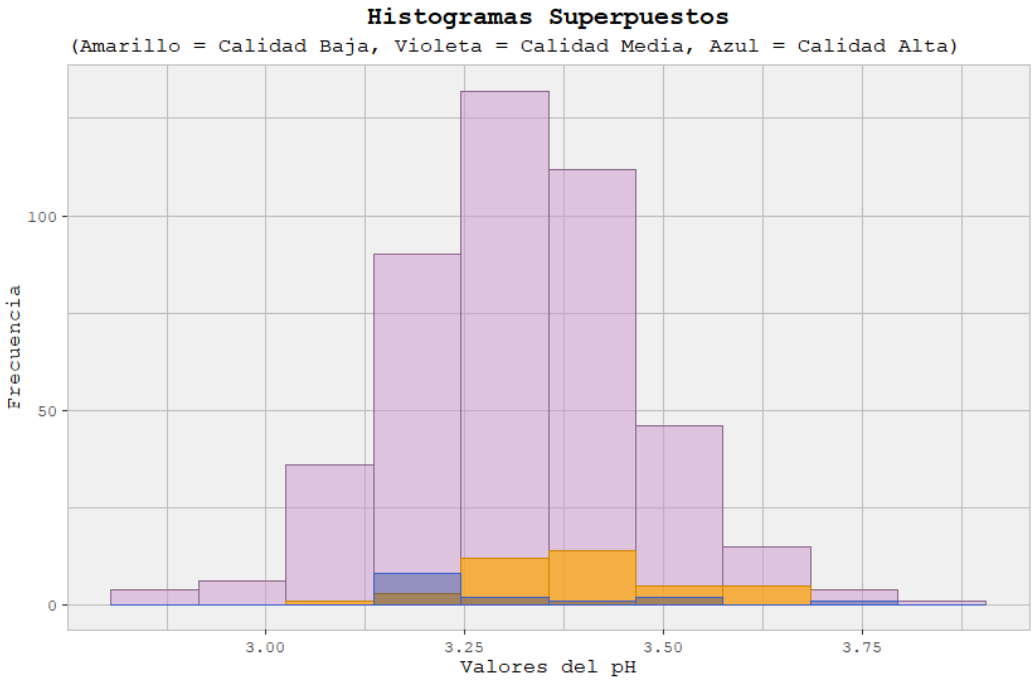

Figura XXV

Bibliografía

- Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009. [en línea]. Obtenido en: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

Recursos adicionales utilizados

- Documentacion de libreria GGPLOT [en linea] https://www.rdocumentation.org/packages/ggplot2/versions/3.4.1