



87.17 Análisis Multivariado

Comisión A

Primer Trabajo Práctico

Azul de los Ángeles Makk

Paula Ariana González

Cosatto Ammann, Pedro Camilo

Índice

1. Introducción	2
2. Base de datos	3
3. Análisis descriptivo de datos numéricos	3
3.1. fixed.acidity	4
3.2. density	7
3.3. alcohol	11
3.4. pH	15
3.5. Relaciones entre variables	18
4. Transformaciones previas	23
4.1. Ajuste Chi-cuadrado	23
4.2. Análisis de datos atípicos	25
4.3. Box y Cox	26
5. Componentes principales y biplot	28
5.1. Función de PCA	28
5.2. Biplot y coeficiente de determinación	30
6. Conclusiones	34
7. Bibliografía	35

1. Introducción

Este informe tiene como objetivo explorar y analizar un conjunto de datos utilizando técnicas de análisis multivariado. A lo largo del informe, se abordarán varios aspectos clave del análisis, con el objetivo de comprender mejor la información contenida en los datos y extraer conclusiones significativas.

Se comienza por presentar la base de datos que utilizaremos en nuestro análisis, destacando su importancia y contexto. A continuación, se centra en un análisis descriptivo de datos numéricos, prestando especial atención a cuatro variables específicas: 'fixed.acidity', 'density', 'alcohol' y 'pH'. Esta etapa nos proporciona una visión inicial de las características y distribuciones de estas variables.

Para preparar los datos para un análisis multivariado efectivo, se realizarán algunas transformaciones previas para que se garantice que los datos sean normales multivariados. Se utilizan las densidades de las distancias de Mahalanobis y la distribución chi-cuadrado para evaluar que

se cumpla la normalidad. Asimismo, se aborda la detección y análisis de datos atípicos.

El corazón del análisis residirá en la sección dedicada a los Componentes Principales (PCA) y el Biplot. Se utilizan estas técnicas para reducir la dimensionalidad de los datos y visualizar las relaciones entre las variables en un espacio multidimensional. El PCA permite identificar patrones subyacentes en los datos, mientras que el Biplot ayuda a visualizar y comprender estas relaciones de manera más clara.

2. Base de datos

Se decidió estudiar la base de datos que muestra la calidad de distintos tipos de vinos tintos llamada “winequality-red.csv”. 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. Para asegurar un comportamiento consistente entre las observaciones de cada categoría, se han seleccionado 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 -con reposición-.

Se ha identificado 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).

3. Análisis descriptivo de datos numéricos

En esta sección se realiza un análisis de cada una de las 4 variables seleccionadas para poder entender mejor cómo se comportan. Para llevar a cabo el análisis se utilizó R en RStudio y para producir los gráficos se instaló 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 prolífica y personalizada.

3.1. fixed.acidity

$$F = \text{Acidez fija de un vino Vinho Verde.}$$

La variable F mide la suma de todos aquellos ácidos que, al someter el vino al calor, no se evaporan. Primeramente, se observa 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, se grafica la función de distribución empírica que se puede observar en la *Figura I*.

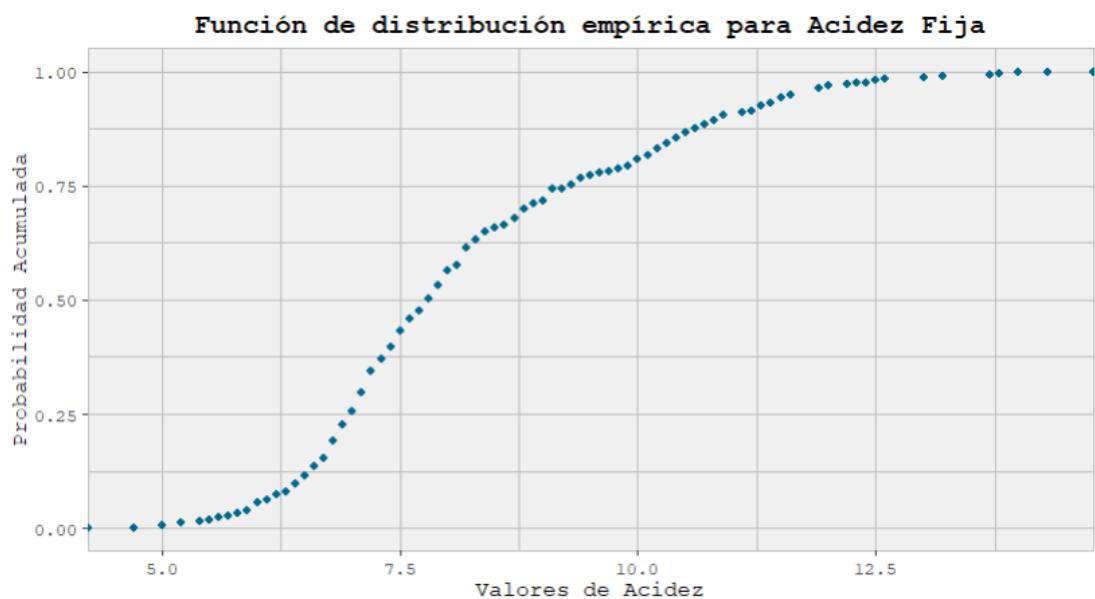


Figura I

Es posible observar en el gráfico que como los puntos son muy próximos unos de otros, esta variable es continua. Por esta razón, se grafica un histograma (véase *Figura II*) para poder distinguir la distribución de la variable F.

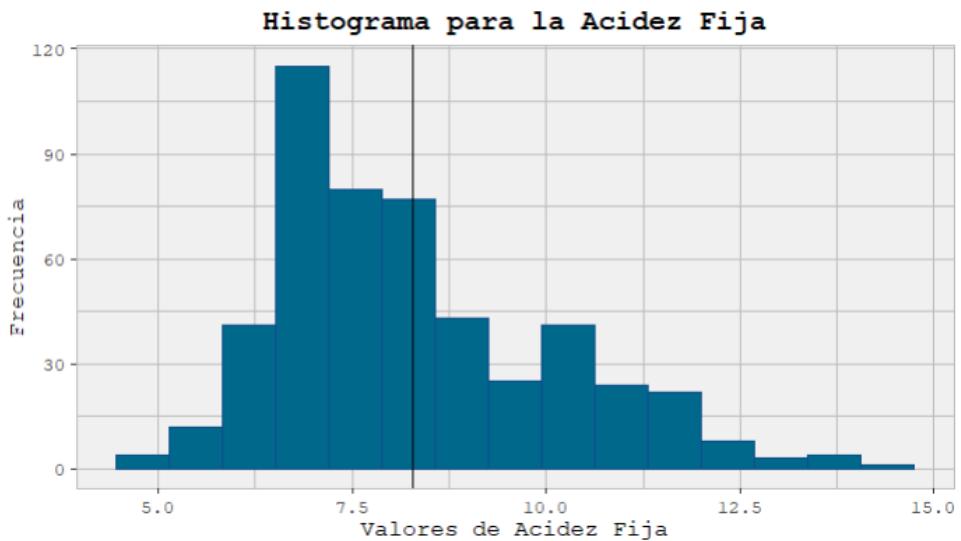


Figura II

A su vez, se grafica la función de densidad (véase *Figura III*), y es posible 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.

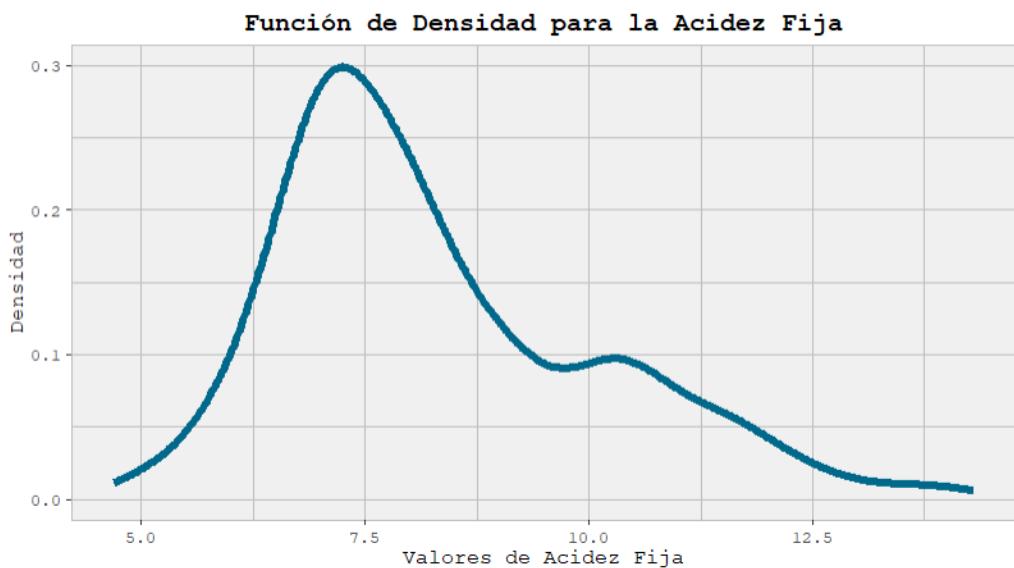


Figura III

Luego, se estima la media muestral de $\bar{f} = 8.27$ y la mediana de $f_{0.5} = 7.9$, de manera que al ser mayor la mediana que la media es posible afirmar que la distribución tiene asimetría positiva -luego respaldado en *Tabla II*-. Esto se puede confirmar en el boxplot de la *Figura IV*. En

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	$f_{0.25} = F^{-1}(0.25) = \min f_i \text{ tq } \{f_i: \hat{F}(f_i) \geq 0.25\}$	7.0
0.5	$f_{0.5} = F^{-1}(0.5) = \min f_i \text{ tq } \{f_i: \hat{F}(f_i) \geq 0.5\}$	7.9
0.75	$f_{0.75} = F^{-1}(0.75) = \min f_i \text{ tq } \{f_i: \hat{F}(f_i) \geq 0.75\}$	9.3

Tabla I

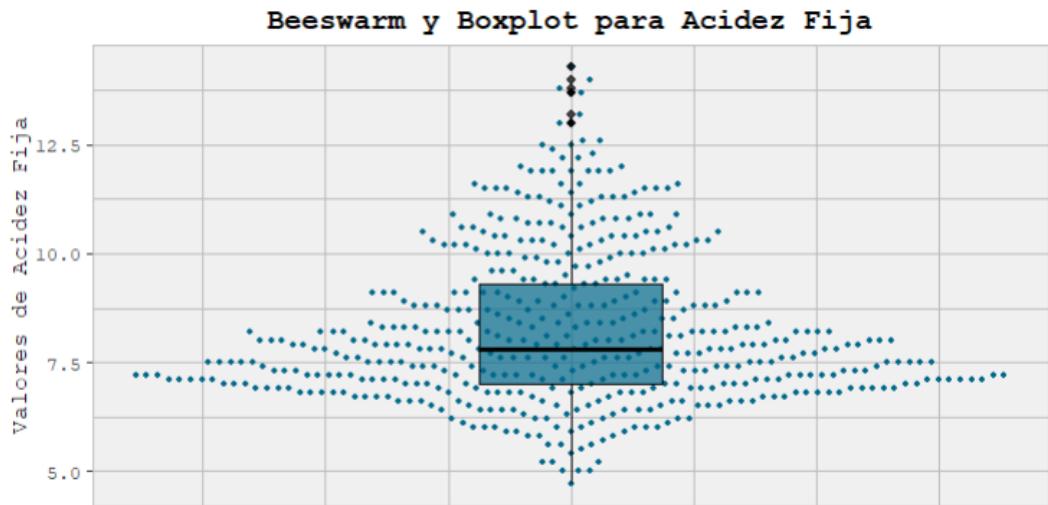


Figura IV

Como se puede observar 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 se optó no excluirlos ya que no son significativos ya que no alteran significativamente a las medidas no robustas con respecto a las robustas. Asimismo, se ha realizado 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 se considera 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, se ha confeccionado la siguiente *Tabla II* para visualizar un breve resumen estadístico de la variable de acidez del vino. De todos los valores, no resulta de principal interés 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 <i>fixed.acidity</i>		
Media muestral	$\bar{f} = \frac{\sum_{i=1}^n F_i}{N}$	8.27
Desvió estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (F_i - \bar{F})^2}{n}$	1.79
Coeficiente de variación	$r = \frac{\bar{V}(f)}{ f }$	0.21
Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\sum (\frac{f_j - \bar{f}}{\sqrt{V(f)}})^3}{n}$	0.81
Coeficiente de curtosis muestral	$\hat{k} = \frac{\sum (\frac{f_j - \bar{f}}{\sqrt{V(f)}})^4}{n}$	3.20

Tabla II

3.2. *density*

$$D = \text{Densidad de un vino Vinho Verde (gr/ml)}$$

La variable D mide la densidad del vino tinto en gramos por mililitro. El vino que registró menor densidad fue de 0.99 y el de mayor valor registrado de 1.003, presentando un rango total de 0.013. A continuación, en la *Figura V*, se observa su gráfico de la función de distribución empírica.

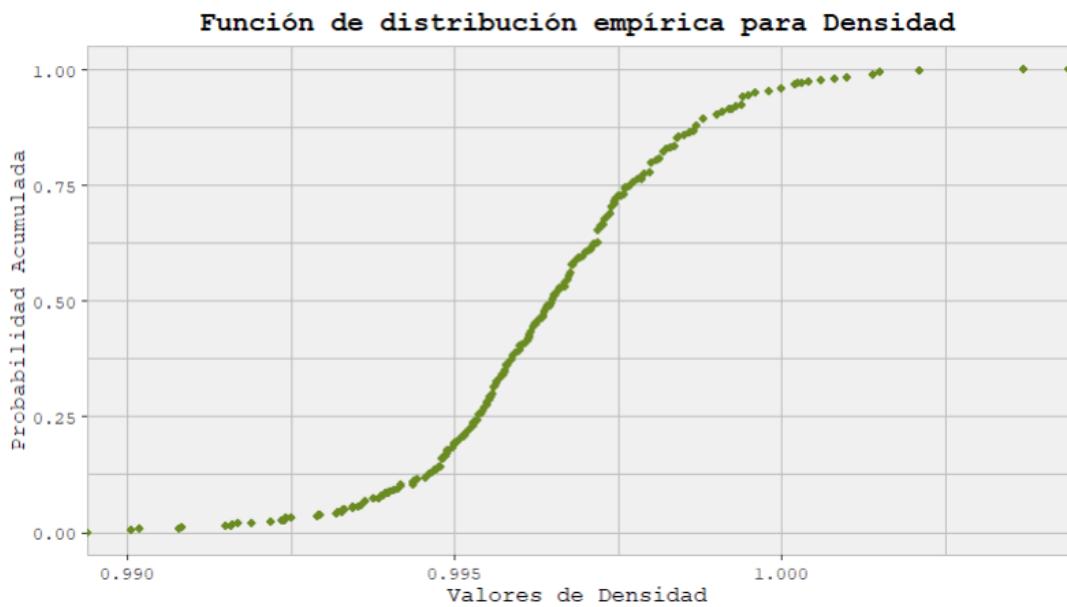


Figura V

Se podría decir que la variable D se comporta de manera aleatoria continua. También es posible visualizar que la gran mayoría de los puntos se encuentran concentrados en el centro de la distribución, cercano a la media. Se procede 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.

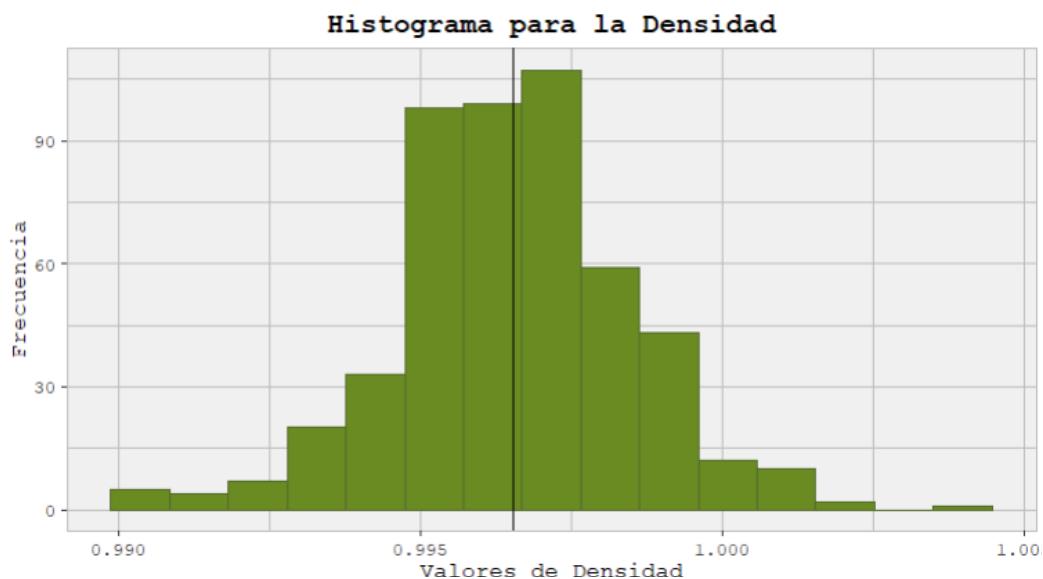


Figura VI

Función de Densidad para la Densidad del vino

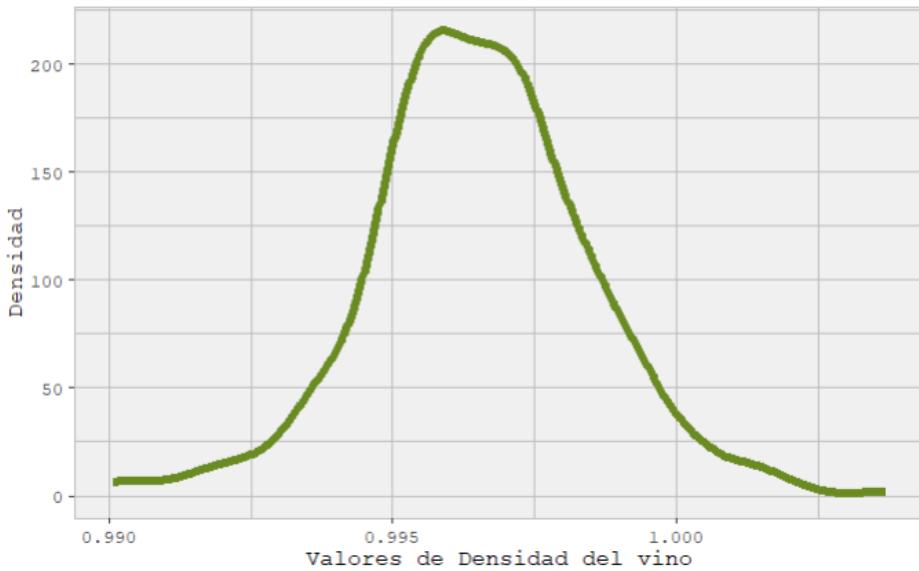


Figura VII

La *Figura VII* muestra que la curva se parece a una campana de Gauss (con algunas irregularidades). Luego, con su posterior cálculo se identifica que la media y la mediana están representadas por el mismo valor ($\bar{d} = d_{0.5} = 0.9$), 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	$d_{0.25} = F^{-1}(0.25) = \min d_i \text{ tq}\{d_i; \hat{F}(d_i) \geq 0.25\}$	0.99
0.5	$d_{0.5} = F^{-1}(0.5) = \min d_i \text{ tq}\{d_i; \hat{F}(d_i) \geq 0.5\}$	0.99
0.75	$d_{0.75} = F^{-1}(0.75) = \min d_i \text{ tq}\{d_i; \hat{F}(d_i) \geq 0.75\}$	0.99

Tabla III

Beeswarm y Boxplot para la Densidad

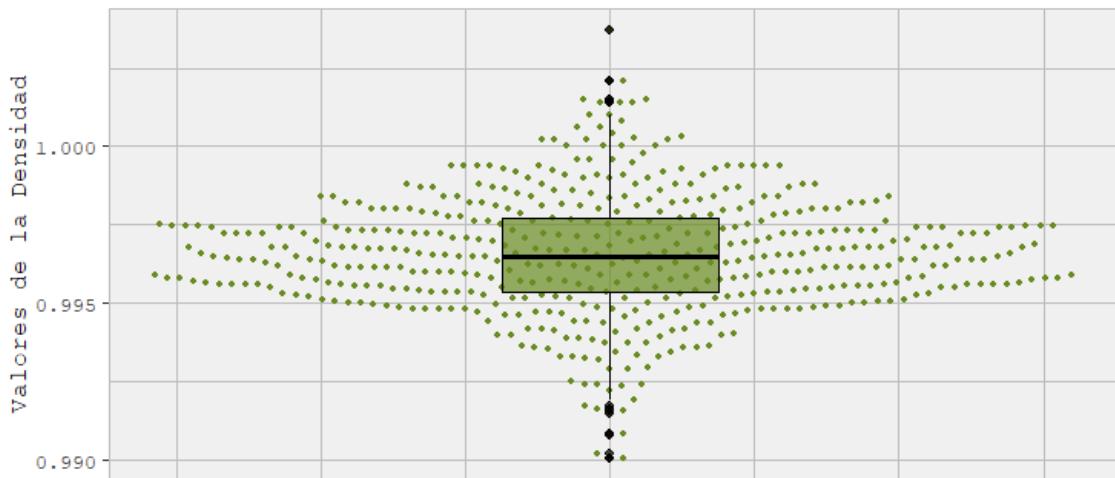


Figura VIII

Como se ha mencionado previamente, en la parte superior e inferior del gráfico es observable la presencia de outliers -marcados en color negro-. Se ha 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.

Función de Densidad para la Densidad del vino
(Verde oscuro = con outliers, Verde claro = sin outliers)

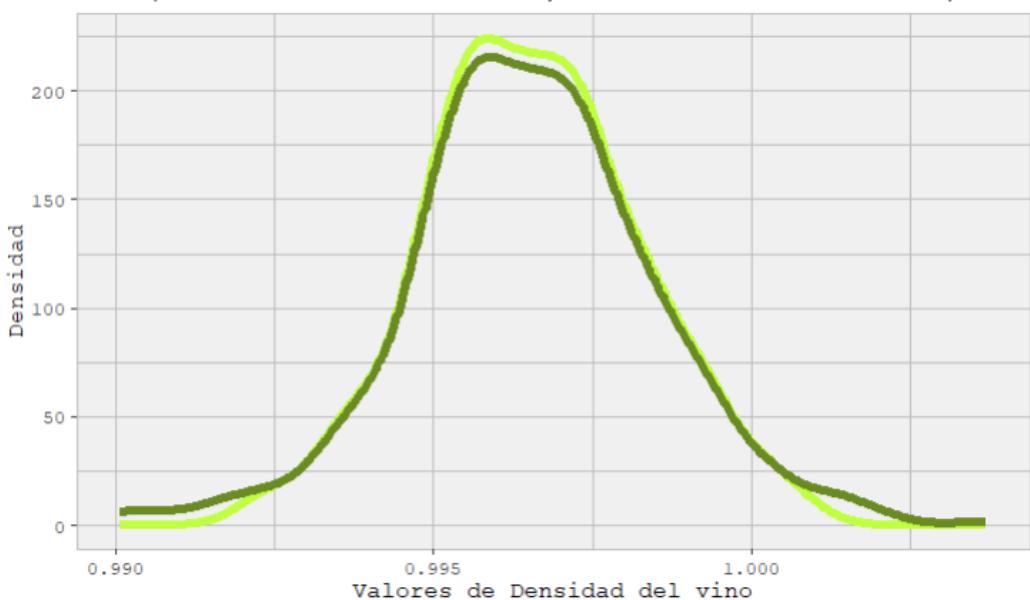


Figura IX

Para finalizar el análisis de esta variable, se confecciona 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 se ha supuesto anteriormente y presenta una asimetría levemente negativa.

Resumen estadístico para density		
Media muestral	$\bar{d} = \frac{\sum_{i=1}^n D_i}{N}$	0.99
Desvio estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n}$	0.001
Coeficiente de variación	$r = \frac{\bar{V}(d)}{ \bar{d} }$	0.001
Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\sum (\frac{d_i - \bar{d}}{\sqrt{V(d)}})^3}{n}$	-0.03
Coeficiente de curtosis muestral	$\hat{k} = \frac{\sum (\frac{d_i - \bar{d}}{\sqrt{V(d)}})^4}{n}$	5.85

Tabla IV

3.3. alcohol

$$A = \text{Porcentaje de alcohol de un vino Vinho Verde.}$$

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 continua.

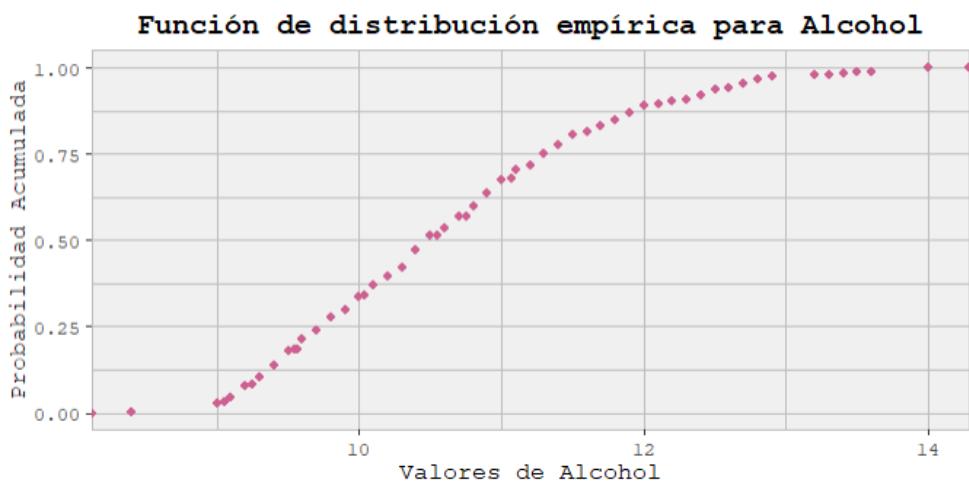


Figura X

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.

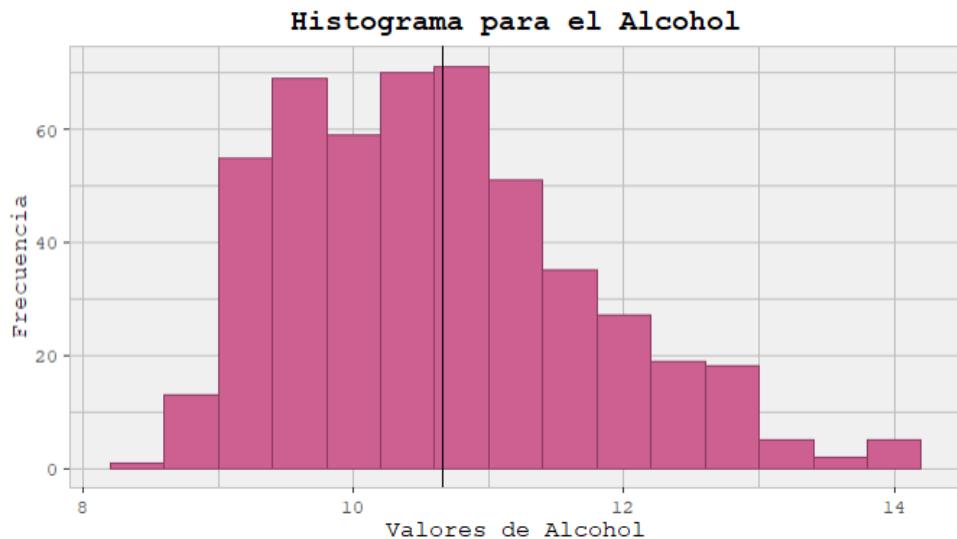


Figura XI

Es posible 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 $x=13$.

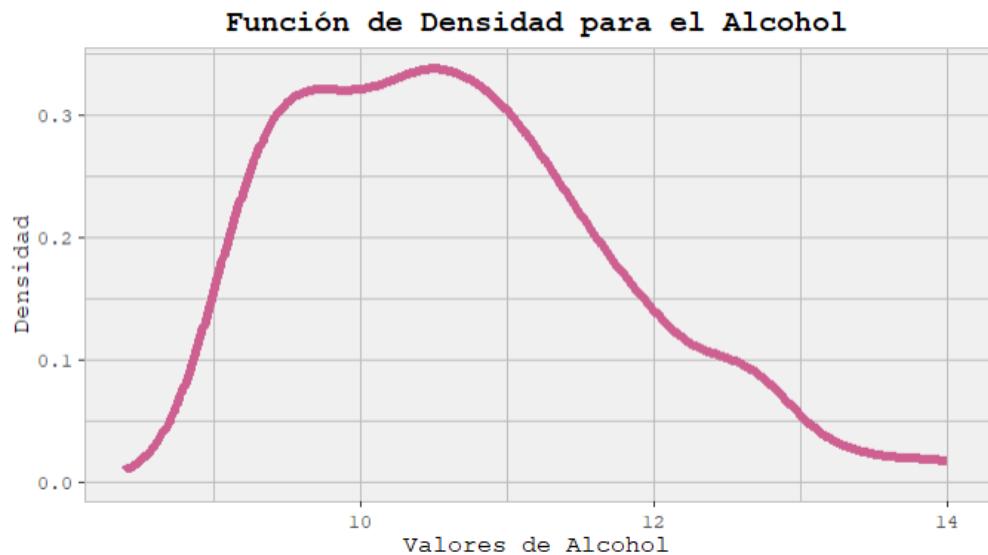


Figura XII

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	$a_{0.25} = F^{-1}(0.25) = \min a_i \text{ tq} \{a_i: \hat{F}(a_i) \geq 0.25\}$	9.80
0.5	$a_{0.5} = F^{-1}(0.5) = \min a_i \text{ tq} \{a_i: \hat{F}(a_i) \geq 0.5\}$	10.50
0.75	$a_{0.75} = F^{-1}(0.75) = \min a_i \text{ tq} \{a_i: \hat{F}(a_i) \geq 0.75\}$	11.32

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 se ha decidido no excluirlos de la muestra. Asimismo, se ha realizado 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 se considera que no es un gráfico que aporte tanto valor al informe y nos confirma que no es necesario eliminar los valores atípicos.

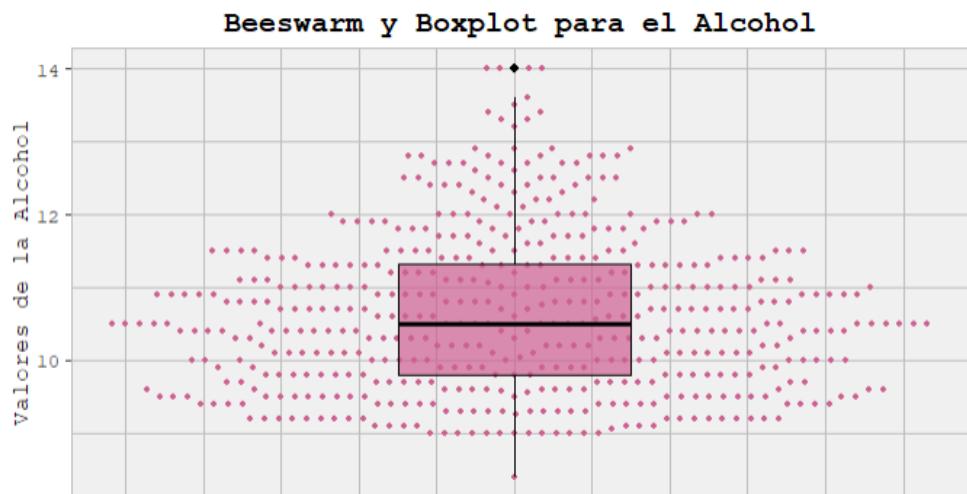


Figura XIII

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	$\bar{a} = \frac{\sum_{i=1}^n A_i}{N}$	10.65
Desvío estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (A_i - \bar{A})^2}{n}$	1.09
Coeficiente de variación	$r = \frac{\bar{V}(a)}{ \bar{a} }$	0.10

Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\sum(\frac{a_j - \bar{a}}{\sqrt{V(a)}})^3}{n}$	0.62
Coeficiente de curtosis muestral	$\hat{k} = \frac{\sum(\frac{a_j - \bar{a}}{\sqrt{V(a)}})^4}{n}$	2.94

Tabla VI

3.4. pH

P = Medida de pH 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, se supone que la forma de la distribución adquirirá la forma de una campana de Gauss. Asimismo, es posible afirmar que la variable P se comporta como una aleatoriedad continua.

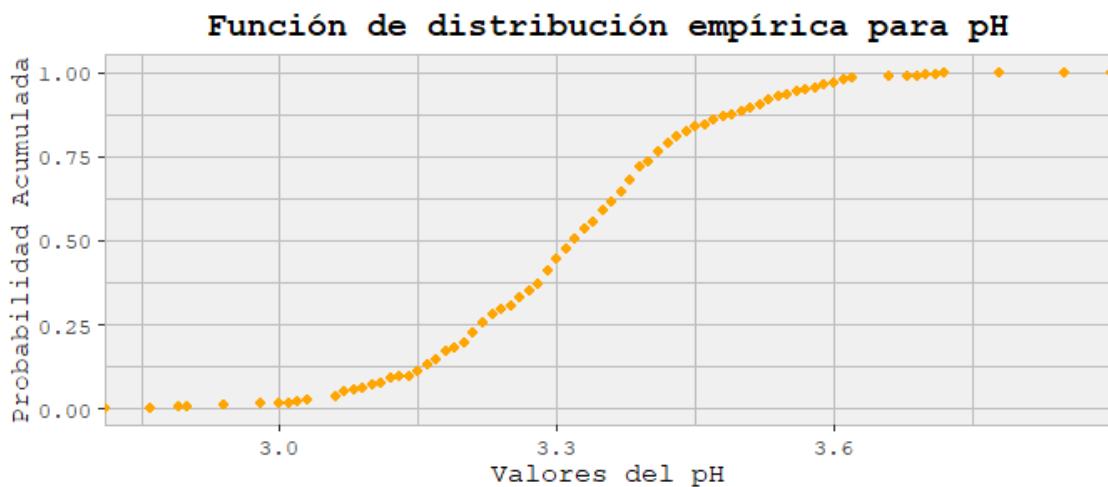


Figura XIV

En la muestra tomada es posible 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.

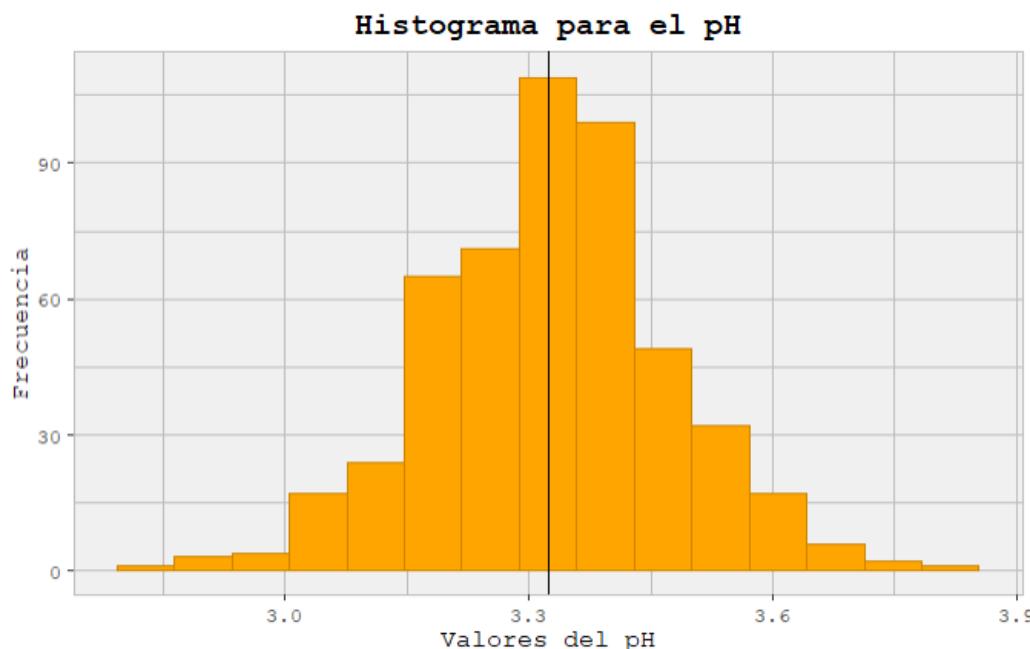


Figura XV

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.

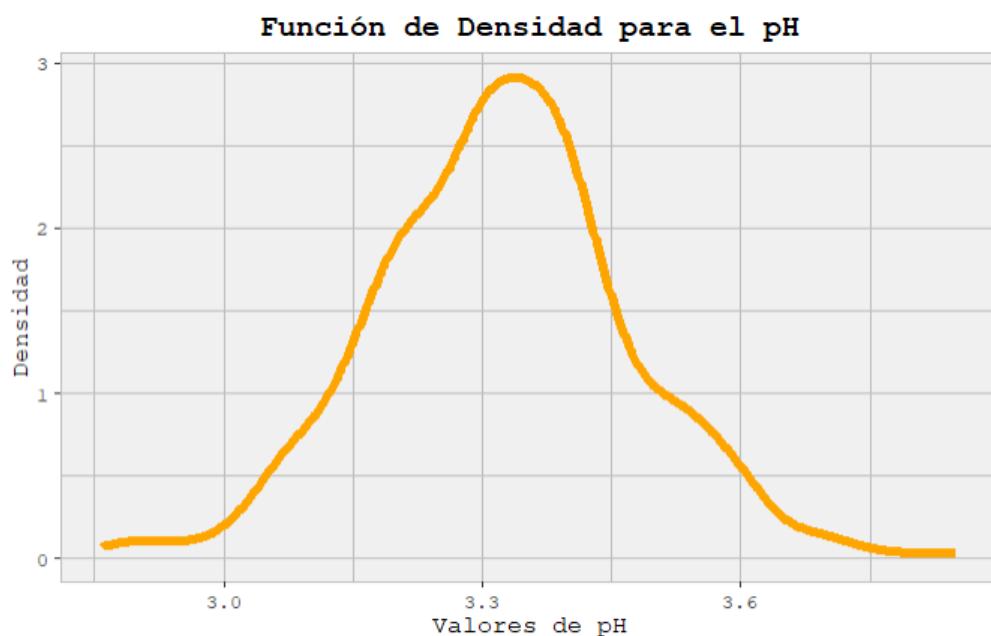


Figura XVI

Asimismo, se puede 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, es posible 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 numéricico
0.25	$p_{0.25} = F^{-1}(0.25) = \min p_i \text{ tq} \{p_i: \hat{F}(p_i) \geq 0.25\}$	3.22
0.5	$p_{0.5} = F^{-1}(0.5) = \min p_i \text{ tq} \{p_i: \hat{F}(p_i) \geq 0.5\}$	3.32
0.75	$p_{0.75} = F^{-1}(0.75) = \min p_i \text{ tq} \{p_i: \hat{F}(p_i) \geq 0.75\}$	3.41

Tabla VII

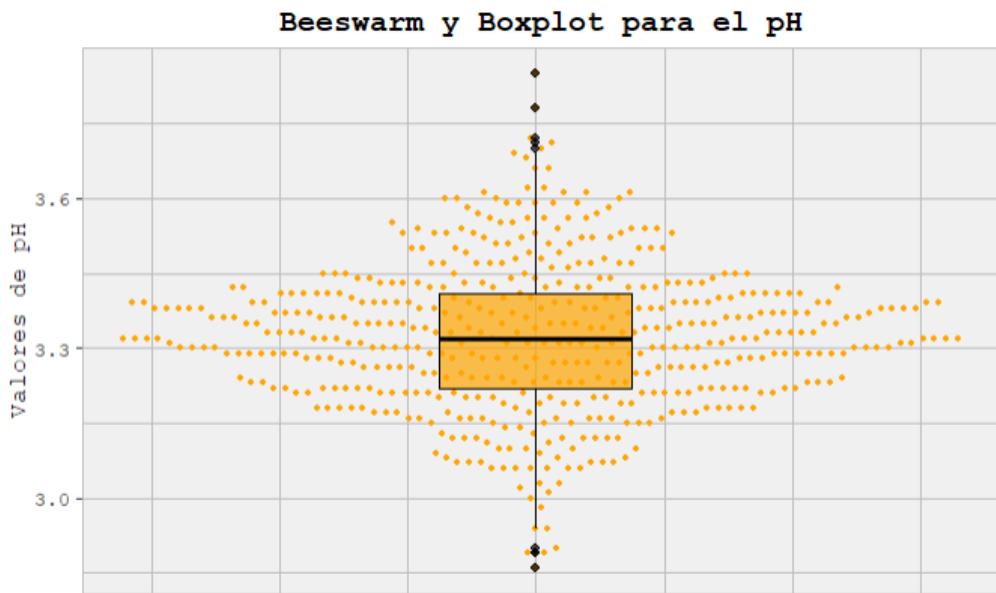


Figura XVII

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

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

Para finalizar, en la Tabla XVIII se puede apreciar 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, se ha identificado 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	$\bar{p} = \frac{\sum_{i=1}^n P_i}{N}$	3.32
Desvío estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (P_i - \bar{P})^2}{n}$	0.14
Coeficiente de variación	$r = \frac{\bar{V}(p)}{ p }$	0.04
Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\sum (\frac{p_i - \bar{p}}{\sqrt{V(p)}})^3}{n}$	0.05
Coeficiente de curtosis muestral	$\hat{k} = \frac{\sum (\frac{p_i - \bar{p}}{\sqrt{V(p)}})^4}{n}$	3.44

Tabla VIII

3.5. Relaciones entre variables

En esta sección se analizarán las relaciones entre las variables presentadas previamente. A fines de llevar adelante un análisis integral, se ha ejecutado 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	1.00	0.68	-0.95	-0.70
density	0.68	1.00	-0.51	-0.34
alcohol	-0.09	-0.51	1.00	0.17
pH	-0.70	-0.34	0.17	1.00

Tabla IX

Para poder facilitar el análisis de las correlaciones, se decidió ilustrar el siguiente gráfico (véase Figura XVIII), que visualizan los valores dados en la *Tabla IX*. El color rojo indica una fuerte correlación positiva entre dos variables mientras que un color azulado demuestra una correlación negativa.

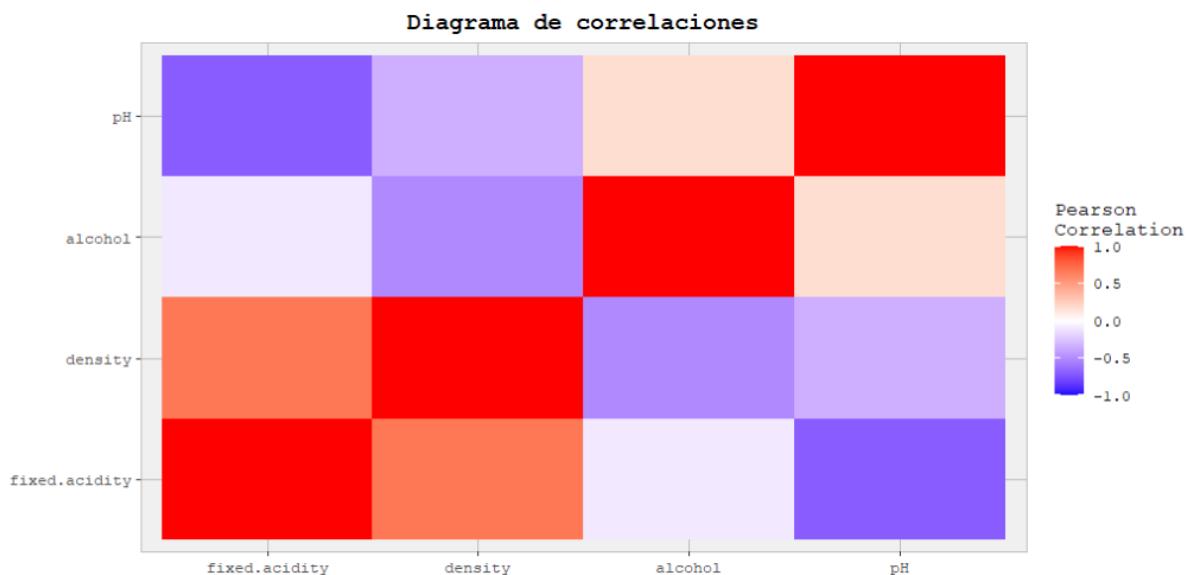


Figura XVIII

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

En primer lugar, se decidió relacionar la acidez del vino con el pH, lo cual había arrojado una correlación de -0.70, que señala que una estrecha relación entre ambas al estar cerca de -1. Tal información es

verificada visualmente mediante un scatter plot (*Figura XIX*), 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, se concluyó que a mayor acidez fija menor es el pH o, alternativamente, a mayor pH tiende a ser menor la acidez fija.

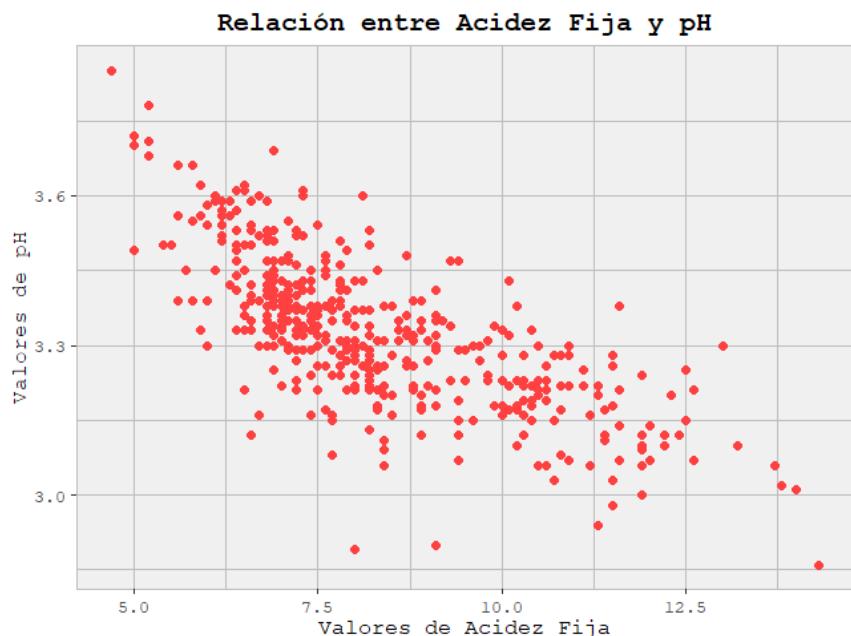


Figura XIX

Luego, se continuó 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 XX*) se puede ver una clara tendencia decreciente -ubicando la densidad en el eje horizontal y el alcohol en el eje vertical-. Por lo tanto, se infiere que a mayor densidad, menor es el nivel de alcohol y viceversa. En este caso, se puede 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 -0.51.

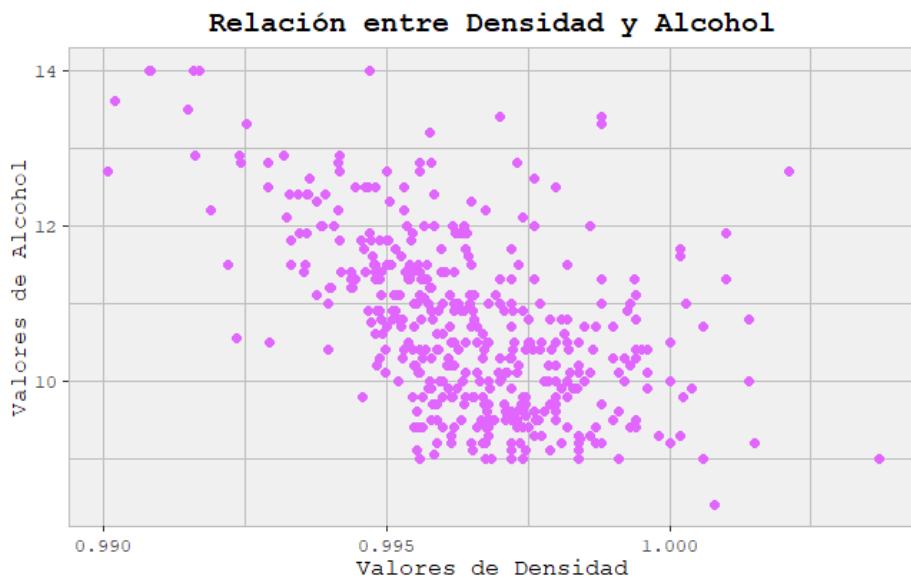


Figura XX

Para el tercer análisis sobre relaciones de variables se decidió observar la conexión entre el alcohol y el pH (*Figura XXI*). Previamente en la *Tabla IX* se arroja como resultado numérico una correlación de 0.17, 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.

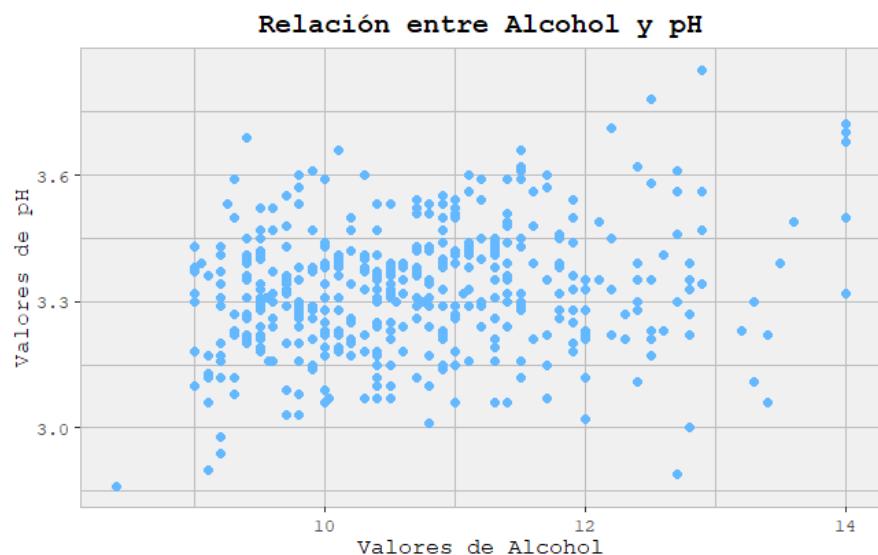


Figura XXI

Para finalizar, se eligió evaluar la relación entre la densidad y el pH (*Figura XXII*). Previamente en la *Tabla IV*, el resultado numérico nos muestra que su correlación es de -0.34, 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.

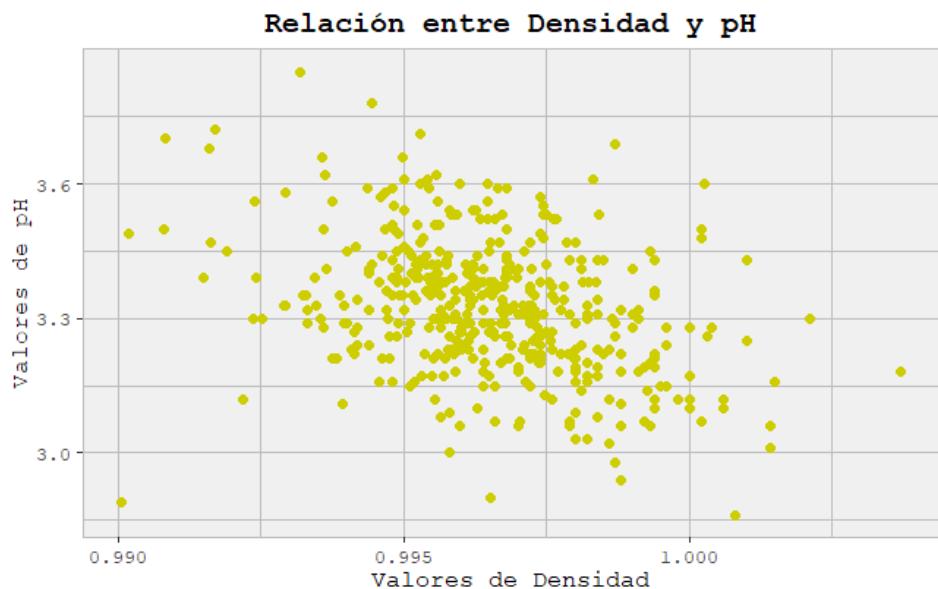


Figura XXII

Para que se pueda visualizar de forma más efectiva las relaciones entre las variables cuantitativas se grafica el siguiente gráfico de pairs (véase *Figura XXIII*), que permite ver en conjunto los distintos gráficos de dispersión de las distintas combinaciones, y asimismo, observa su respectivo coeficiente de correlación de la relación.

Una observación de relevancia que se desprende del presente gráfico es que se distinguen claramente dos estructuras lineales que exhiben una correlación más significativa. Concretamente, estas estructuras corresponden a las correlaciones entre las variables "fixed.acidity" y "pH", así como entre "fixed.acidity" y "density".

Asimismo, se puede observar que tres gráficos de dispersión muestran una forma similar a la de una elipse, siendo estos la relación entre "fixed.acidity" y "pH", "fixed.acidity" y "density" y "density" y "alcohol". Esto nos anticipa que esas variables se comportan como normales multivariadas.

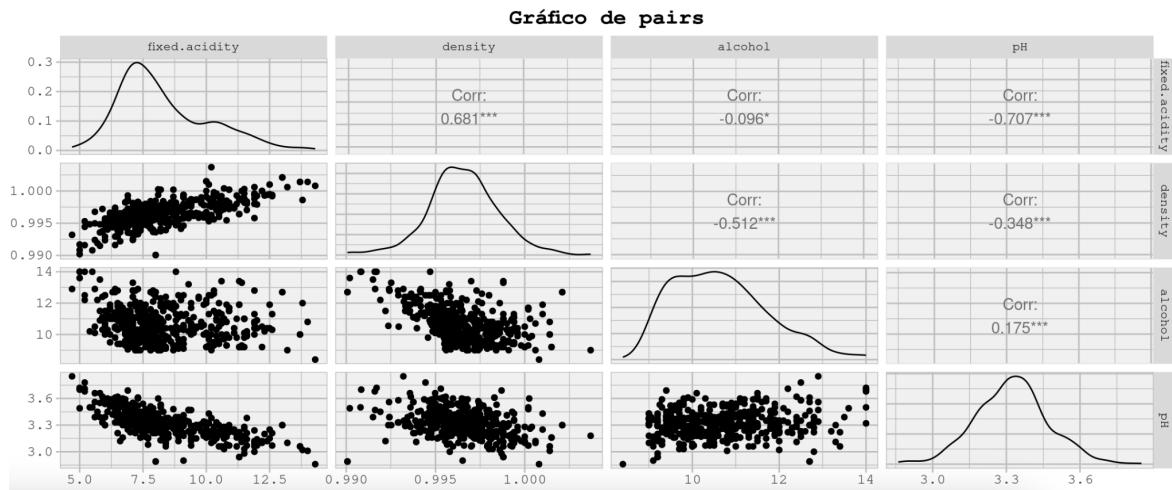


Figura XXIII

4. Transformaciones previas

4.1. Ajuste Chi-cuadrado

Las distancias de Mahalanobis es una métrica estadística empleada en el contexto del análisis multivariado. Su finalidad principal radica en cuantificar la distancia entre un punto de datos específico y un centro de distribución de datos multivariados. Lo que distingue a estas distancias es que toman en consideración las relaciones entre las diferentes variables en juego, teniendo en cuenta la covarianza que existe entre ellas.

$$dm_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^t S^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$

dm_{ii} = Representa la distancia de Mahalanobis.

x_i = Es el conjunto de datos para el cual se está evaluando la distancia.

x_j = Representa el vector de medias, es decir, los promedios de las variables en el centro de distribución.

S^{-1} = Es la matriz de covarianza que refleja cómo las variables se relacionan entre sí en el centro de distribución.

La relevancia de la distancia de Mahalanobis se encuentra en su capacidad para capturar la estructura de correlación existente entre múltiples variables. Este aspecto se torna especialmente significativo

cuando las variables en consideración presentan relaciones de correlación entre sí, dado que esta métrica refleja cómo estas variables varían conjuntamente.

Después de calcular las distancias de Mahalanobis, se procedió a crear un gráfico de densidad para estas distancias y posteriormente se han contrastado con datos generados de manera aleatoria que toman una distribución chi cuadrado. Este análisis se visualiza en la *Figura XXIV*, donde se representa la distribución de las distancias en color azul, mientras que en color rojo se muestra la distribución chi-cuadrado con un número de grados de libertad igual a 4, dado que se está trabajando con 4 variables numéricas independientes.

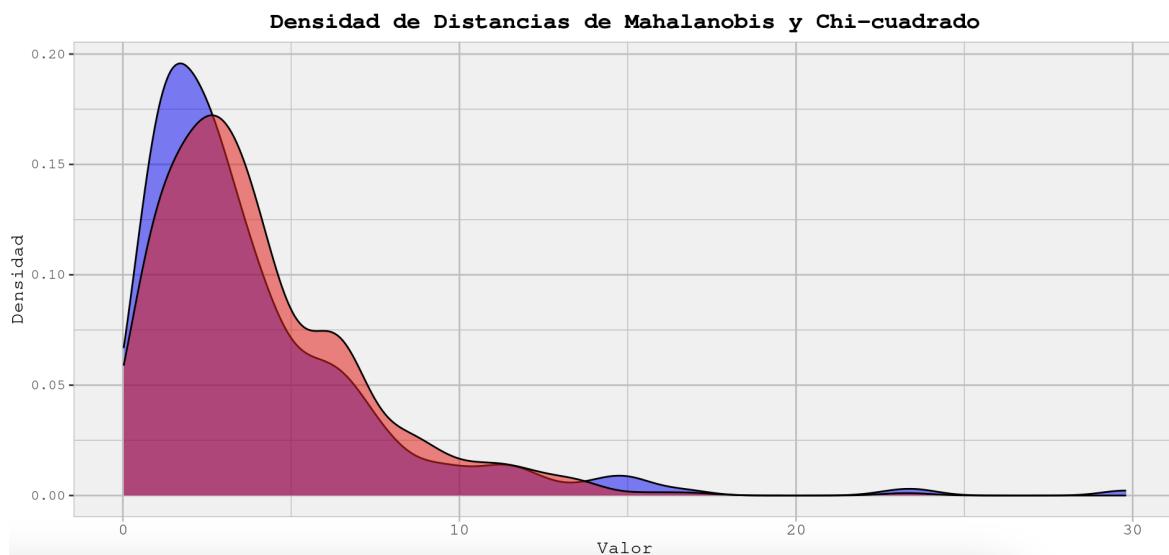


Figura XXIV

En el gráfico presentado, se advierte una notoria similitud entre la densidad de las distancias de Mahalanobis y una distribución chi-cuadrado, lo cual sugiere que los datos subyacentes que dieron origen a dichas distancias pueden considerarse como una muestra proveniente de una distribución normal multivariada. Esto se pudo anticipar previamente en nuestro análisis de la *Figura XXIII*. De todas formas, las curvas no se superponen del todo, lo cual tiene sentido ya que en el gráfico de pairs no todas las variables se comportan de manera normal multivariada, es decir que se observó que no todas tenían la forma de una que se asemeja a una elipse o una “pelota de rugby”.

4.2. Análisis de datos atípicos

Con respecto a los datos atípicos, delimitamos dos posibles clasificaciones:

- Outliers marginales: Interpretables en caso de que se presente un valor notablemente grande o pequeño. Es perceptible mediante las gráficas disponibles en la sección 3 del presente trabajo, utilizando métodos como por ejemplo boxplots o bee swarm plots.
- Outliers de correlación: Clasificados en función a las distancias de Mahalanobis (explicadas en detalle más adelante). Como tal hemos decidido elaborar la *Tabla X*, en función a enlistar las características de cada registro. Es importante tener en cuenta que si bien se muestran los datos con mayores distancias, aquellos que ocupan el primer y segundo puesto representan un dato que al realizar el muestreo fue tomado dos veces.

Puesto	Distancia	fixed.acidity	density	alcohol	pH
1	29,8	8,0	0,990	12,7	2,89
3	23,6	13,0	1,000	12,7	3,3
5	22,9	10,2	1,000	9,0	3,18
6	16,9	14,3	1,000	8,4	2,86
7	16,5	11,4	0,998	13,3	3,11

Tabla X

Con el objetivo de comprender la mecánica de comportamiento de los registros que poseen las mayores distancias, podemos identificar:

- Fixed acidity: Se observa que en los puestos 3, 5, 6 y 7 se presenta una acidez fija que excede el límite superior del rango intercuartílico. En otras palabras, estos valores superan el cuartil 0.75 al agregar el producto del rango entre el cuartil 0.25 y 0.75 multiplicado por 1.5. En consecuencia, se evidencia que estos registros exhiben valores notoriamente superiores en comparación con el conjunto de datos. Además, es relevante mencionar que

únicamente el primer registro se sitúa dentro del rango intercuartílico y se asemeja a la mediana.

- Density: El registro que se ubica por debajo del límite inferior del rango intercuartílico presenta la mayor discrepancia en términos de distancia con respecto al resto de los registros. Sin embargo, los demás registros se encuentran dentro del rango intercuartílico, lo que sugiere que no pueden ser considerados como valores atípicos marginales en relación a esta variable.
- Alcohol: Ninguno de los registros se clasifica como valor atípico marginal, ya que todos ellos se sitúan dentro del rango intercuartílico.
- pH: Se destaca que tanto el registro en el puesto 1 como el registro en el puesto 6 exhiben valores por debajo del límite inferior del rango intercuartílico.

En términos de mayor distancia con respecto al comportamiento general de la muestra, la variable "Density" muestra la discrepancia más notoria, seguida de cerca por "Fixed Acidity". Esto sugiere que estos dos aspectos (densidad y acidez fija) son los que más difieren entre los vinos de la muestra en función de los datos proporcionados.

4.3. Box y Cox

La transformación de Box-Cox se basa en una fórmula matemática que ajusta los valores de una variable en función de un parámetro llamado lambda (λ), de modo que los datos transformados se aproximan a una distribución normal. Esto se logra al aplicar una función que es una combinación de la exponenciación y una operación de desplazamiento. Esta transformación es especialmente útil cuando se trabaja con datos que presentan heterocedasticidad y no siguen una distribución normal.

Dado el gráfico de las densidades de la *Figura XXIV* y el gráfico de pairs demostrado en la *Figura XXIII*, se optó por aplicar la transformación de Box-Cox con el fin de investigar si es factible acercar aún más la densidad de las distancias de Mahalanobis a la densidad de una distribución chi-cuadrado.

La formulación de las variables transformadas con box y cox se basa en elevar los datos originales a los valores de lambda obtenidos con la

función bcPower. Se pueden ver las expresiones matemáticas de las transformaciones a continuación.

$$y_1 = x_1^{-0,5}$$

$$y_2 = x_2^{-27,12}$$

$$y_3 = x_3^{-1}$$

$$y_4 = x_4$$

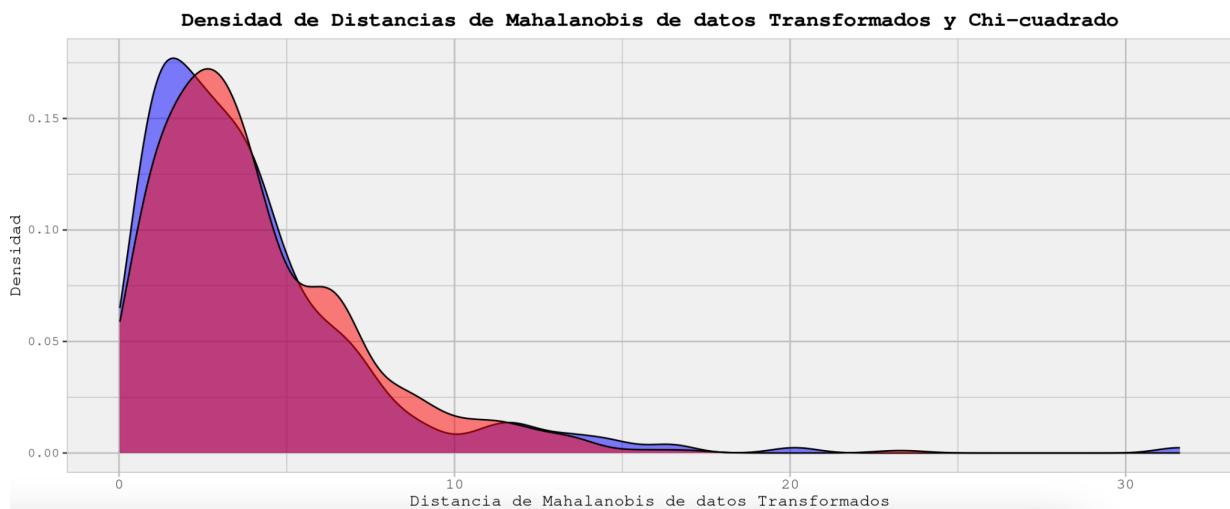


Figura XXV

En la *Figura XXV* se observa un gráfico de densidad, en rojo, de las distancias de Mahalanobis calculadas a partir de los datos sometidos a la transformación correspondiente, superpuesto a una densidad de la distribución chi-cuadrado que se luce en azul. Se aprecia una acentuación de la asimetría negativa en dichos datos y una clara mejoría en la superposición de las curvas de densidades, demostrando que la transformación de los datos ha resultado efectiva ya que se comportan normales multivariados.

Para verificar si la transformación de Box-Cox fue efectiva, se procedió a crear nuevamente un gráfico de pares para observar si los datos, que anteriormente no se asemejan a la forma de una elipse, ahora muestran esta característica de manera más clara y evidente.

Lo que se ve en la *Figura XXVI* es un ejemplo claro de éxito en la mejora de todas las variables, ya que todas ellas ahora muestran una forma que se asemeja a una elipse.

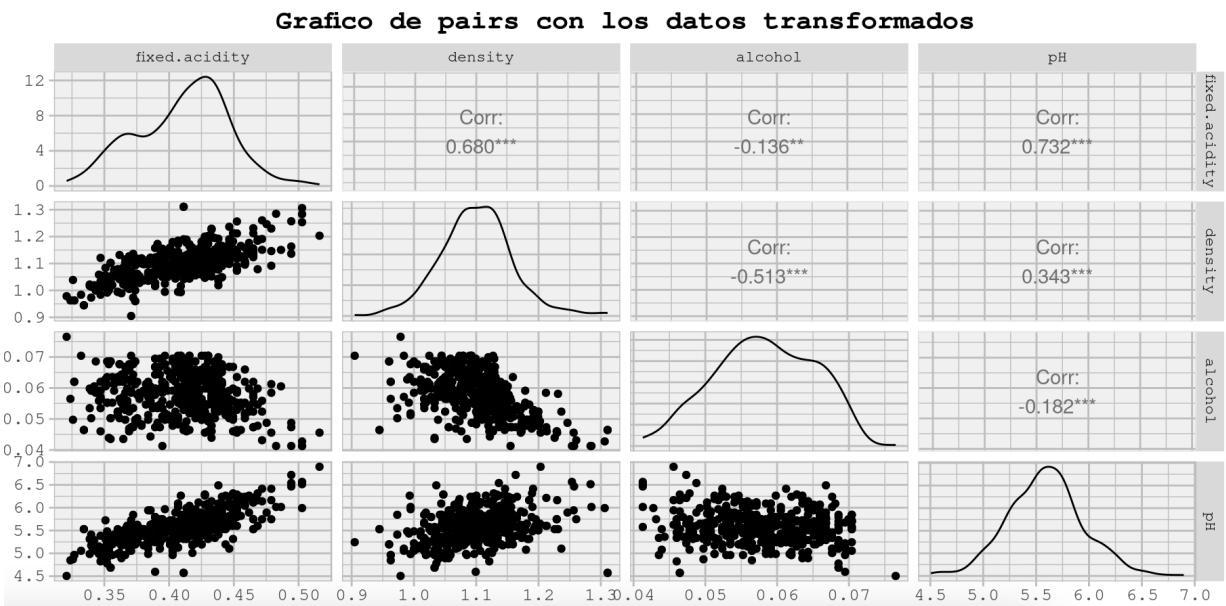


Figura XXVI

5. Componentes principales y biplot

5.1. Función de PCA

PCA (Análisis de Componentes Principales), es una técnica de análisis multivariado utilizada en estadísticas y análisis de datos para reducir la dimensionalidad de un conjunto de datos mientras conserva la mayor cantidad posible de información. PCA logra esto transformando las variables originales transformadas en un nuevo conjunto de variables, llamadas componentes principales, que son combinaciones lineales de las variables. Para este análisis se utilizaron los datos transformados con box y cox ya que, como se ha mencionado previamente, una vez aplicada la transformación los datos presentaron visualmente una similaridad con el comportamiento estándar de una distribución normal multivariada.

En este inciso se creó una función que ejecuta un PCA. Tal función obtiene los autovectores y autovalores de la matriz de covarianzas de las variables numéricas transformadas que permite reducir la dimensionalidad de los datos y seleccionar las dimensiones más importantes para transformar los datos en un nuevo espacio.

En nuestra función de PCA también se calculan las coordenadas de los “scores” que son las coordenadas de los datos transformados en un nuevo espacio de los componentes principales. A continuación, en la

Figura XXVI se visualiza un gráfico de los scores del PCA realizado eligiendo que sea de dos dimensiones. Los puntos en color rojo representan los datos de los vinos que pertenecen a calidad baja, los de color verde son los vinos de calidad media y los azules son los de calidad alta. El objetivo de la distinción de colores es identificar que los scores por categoría no se distribuyen en forma grupos aislados entre sí sino que se comportan de manera uniforme.

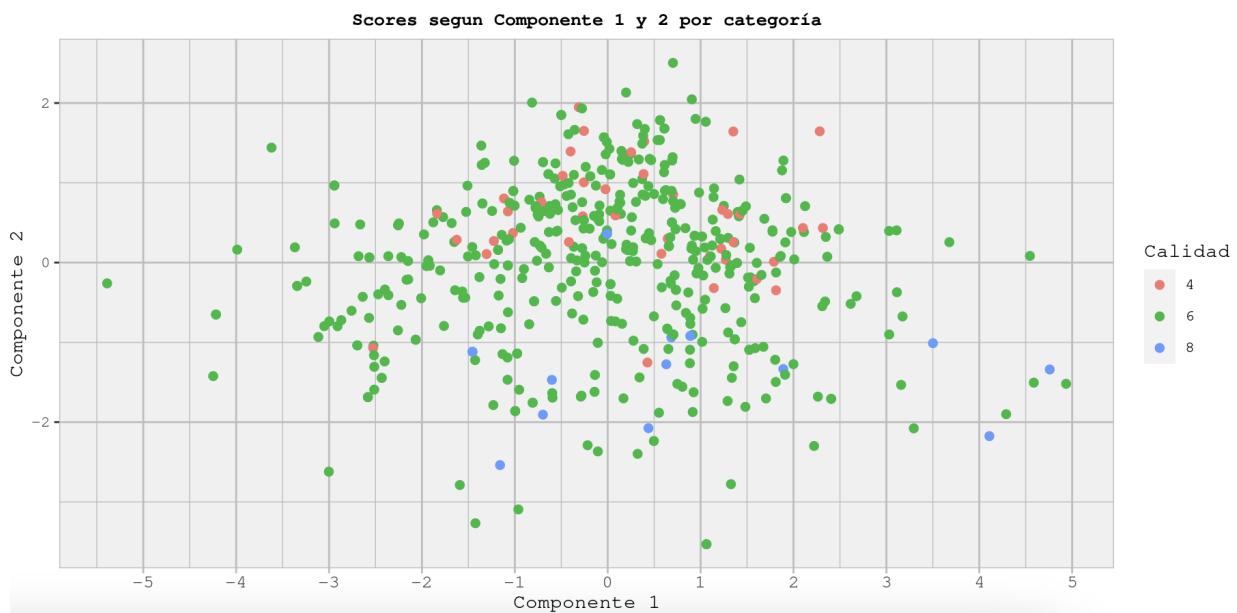


Figura XXVI

Como se puede observar, el primer componente principal “Componente 1” es más disperso que, lo que significa que ese componente principal captura mayor variabilidad en los datos en comparación con “Componente 2”. Esto significa que el primer componente principal tiene más varianza del segundo.

Asimismo, la función PCA calcula una matriz mediante la cual se transforman los datos transformados en los scores (llamada matriz de componentes principales). La matriz de componentes principales tiene dimensiones de $n \times k$. Esto significa que tiene n filas (una fila para cada variable) y k columnas (una columna para cada componente principal). Los elementos de la matriz de componentes principales representan los coeficientes de ponderación que se aplican a las variables transformadas para calcular los scores en el espacio de componentes principales. Esta matriz se puede observar a continuación en la Tabla XI.

	Componente Principal 1	Componente Principal 2
fixed.acidity	0,581	-0,348
density	0,547	0,298
alcohol	-0,335	-0,772
pH	0,499	-0,438

Tabla XI

Por último, la función devuelve un vector que tenga las varianzas de dichos componentes principales. Dichas varianzas son las siguientes que se muestran en la *Tabla XII*. Los valores tienen sentido ya que el primer componente principal debe tener mayor varianza que el segundo como se demuestra en la *Figura XXVI*.

	Componente Principal 1	Componente Principal 2
Varianza	2,348	1,033

Tabla XII

5.2. Biplot y coeficiente de determinación

A continuación, se grafica un biplot (véase *Figura XXVII*) con la intención de visualizar la función PCA creada para el inciso anterior. Tal se realiza en el contexto del PCA para comprender cómo las variables se relacionan entre sí y cómo se distribuyen las observaciones en función de las componentes principales. Su utilidad radica en que permite visualizar la estructura subyacente de datos multivariados de alta dimensionalidad de manera concisa y revelar patrones, relaciones y la importancia relativa de las variables en la variabilidad de los datos. En el ámbito académico, el biplot es esencial para explorar y comunicar resultados en análisis de datos, investigaciones científicas y toma de decisiones informadas en una amplia gama de disciplinas.

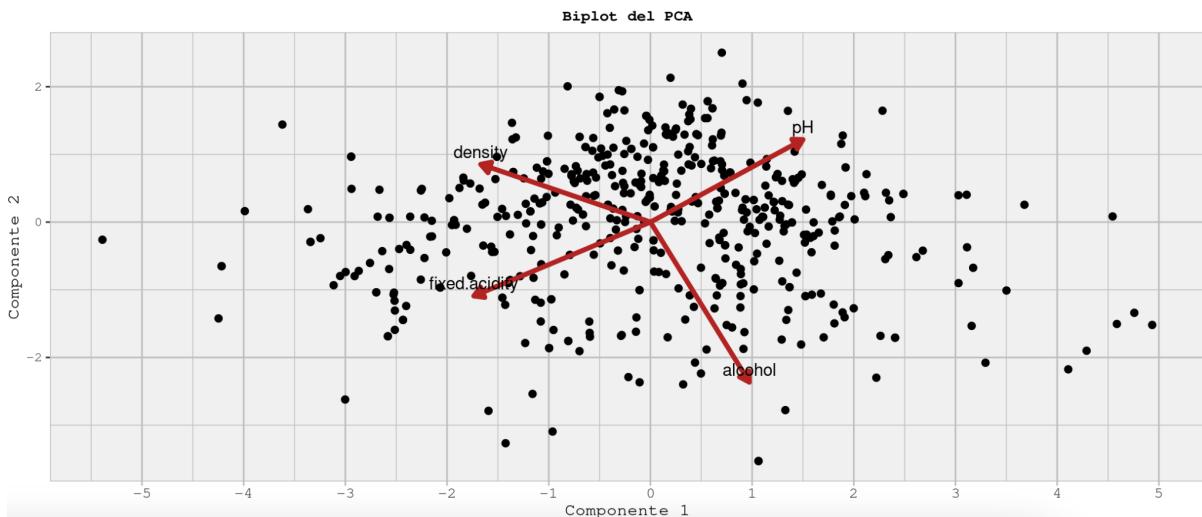


Figura XXVII

En primer lugar, observamos la flecha que representa la variable "Density." Esta flecha se extiende hacia la izquierda en la Componente 1, indicando que "Density" contribuye en gran medida a esa dirección en el espacio de las Componentes Principales. En consecuencia, las observaciones con valores más altos de densidad se encuentran en la parte izquierda del biplot en la Componente 1, mientras que aquellas con valores más bajos de densidad se sitúan en el lado derecho en este mismo componente. Además, la orientación hacia arriba en la Componente 2 sugiere una correlación positiva entre la densidad y otra variable en esa dirección vertical.

Por otro lado, la variable "pH" se representa mediante una flecha que apunta hacia la derecha en la Componente 1, lo que señala su influencia significativa en esa dirección horizontal del espacio de las Componentes Principales. Así, las observaciones con valores elevados de pH se ubican en el lado derecho del biplot en la Componente 1, mientras que las que presentan valores de pH más bajos se sitúan en el lado izquierdo en la misma componente. Esto sugiere una correlación positiva del pH con alguna otra variable en esa dirección horizontal.

Luego, la variable "Alcohol" se ilustra mediante una flecha que apunta hacia abajo en la Componente 2, indicando su contribución destacada en esa dirección vertical del espacio de las Componentes Principales. Por tanto, las observaciones con valores superiores de alcohol se encuentran en la parte superior del biplot en la Componente 2,

mientras que aquellas con valores más bajos de alcohol se hallan en la parte inferior en el mismo componente. Esto implica una correlación negativa del alcohol con alguna otra variable en esa dirección vertical.

Por último, la variable "Fixed Acidity" se representa con una flecha que apunta hacia abajo en la Componente 1, señalando su influencia relevante en esa dirección horizontal del espacio de las Componentes Principales. De esta manera, las observaciones con valores más altos de acidez fija se sitúan en el lado izquierdo del biplot en la Componente 1, mientras que aquellas con valores más bajos de acidez fija se ubican en el lado derecho en el mismo componente. Esto sugiere una correlación negativa de la acidez fija con alguna otra variable en esa dirección horizontal.

En el contexto del biplot, las observaciones en el centro pueden tener dos interpretaciones diferentes. En primer lugar, pueden representar valores intermedios en las variables representadas, lo que indica una neutralidad relativa en relación con esas variables. Sin embargo, en algunos casos, la presencia de observaciones en el centro puede sugerir que están afectando negativamente la estructura de correlación del modelo debido a valores atípicos o extremos en una o más variables, lo que distorsiona las relaciones generales entre las variables en el espacio de las Componentes Principales.

Para medir la eficacia del modelo, se calcula el coeficiente de determinación, el cual arroja un valor de 0,845. Siendo que el valor obtenido es mayor a 0.8, señala que los dos componentes representan de manera apropiada la variabilidad del modelo, según nuestro criterio, el cual se estructura de la siguiente manera:

$$R^2 = \frac{\hat{\lambda}_1 + \hat{\lambda}_2 + \dots + \hat{\lambda}_q}{tr(S)}$$

A continuación, se analiza la matriz de correlaciones entre las Y_i y las componentes principales que se observan en la siguiente *Tabla XIII*. La misma se utiliza para comprender la relación entre las variables transformadas y las nuevas dimensiones creadas por el PCA.

	Componente Principal 1	Componente Principal 2
fixed.acidity	0,891	-0,354
density	0,839	0,302
alcohol	-0,513	-0,785
pH	0,765	-0,446

Tabla XIII

Los valores altos indican una fuerte relación entre una variable transformada y una componente principal, mientras que las cargas cercanas a cero indican una relación débil o nula. En este caso, se visualiza que el “fixed.acidity”, “density” y “ph” tienen una fuerte relación con la Componente Principal 1, mientras que “alcohol” tiene una mayor relación con la Componente Principal 2. Complementariamente, en la Figura XXVIII se puede visualizar un heatmap de estos valores.

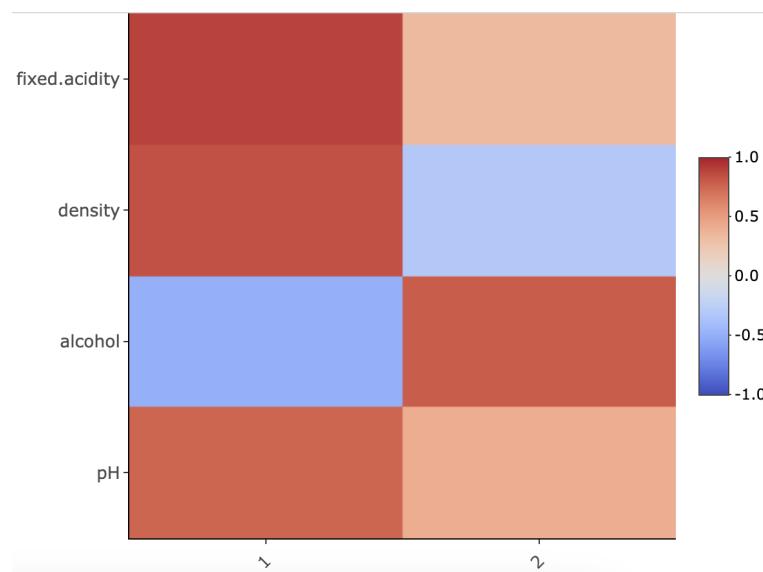


Figura XXVIII

La variable “fixed.acidity” tiene una correlación positiva significativa (0.895) con la primera componente principal. Esto significa que los valores más altos de “fixed.acidity” están asociados con valores más altos en la primera componente principal. Adicionalmente, la variable “density” también tiene una correlación positiva significativa (0.845) con la primera componente principal. Finalmente, la variable “pH” muestra una

correlación positiva (0.762) con la primera componente principal. Contrastando a ello, la variable "alcohol" tiene una correlación negativa significativa (-0.489) con la primera componente principal. Esto significa que los valores más bajos de "alcohol" están asociados con valores más altos en la primera componente principal.

La segunda componente principal muestra correlaciones significativas tanto positivas como negativas con algunas de las variables originales. Por ejemplo, la variable "fixed.acidity" tiene una correlación positiva (0.340), mientras que "density" tiene una correlación negativa (-0.307) con la segunda componente principal. Esto sugiere que la segunda componente principal captura una combinación de influencias de diferentes variables originales.

6. Conclusiones

El análisis multivariado revela dos estructuras lineales distintas en los datos, destacando las correlaciones notables entre "fixed.acidity" y "pH", así como entre "fixed.acidity" y "density". Además, se observa que tres gráficos de dispersión muestran una forma elíptica, sugiriendo que estas variables se comportan como normales multivariadas. El análisis de las distancias de Mahalanobis y su comparación con una distribución chi-cuadrado respalda la idea de que los datos siguen un patrón multivariado y normal. Se identifican registros atípicos con distancias de Mahalanobis significativamente altas.

La aplicación de la transformación de Box-Cox mejora la semejanza del modelo a una distribución normalidad multivariada de los datos, como se refleja en una mayor similitud con una distribución chi-cuadrado y una mejoría en la superposición de las curvas de densidad. El análisis de componentes principales (PCA) permite comprender cómo se relacionan las variables originales y cómo se distribuyen las observaciones en función de las componentes principales. El análisis de componentes principales (PCA) reveló que las variables "fixed.acidity", "density", y "pH" son las principales contribuyentes a la estructura de los datos en el primer componente principal (Componente 1), mientras que "alcohol" es la variable más influyente en el segundo componente principal (Componente 2), mostrando una correlación negativa con las otras variables. El coeficiente de determinación (R^2) de 0.845 indica que estos dos componentes principales son efectivos para representar la variabilidad en los datos.

7. 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 linea]. 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>