Facultad de Ciencias Exactas
Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN)

Fundamentos de Ciencia de Datos

Análisis Calidad de Vino Bodega del Sol

Un análisis exploratorio y estadístico de la calidad de vino en Bodega del Sol

Autores:
Federico Clutterbuck
Tomás Di Carlo

Fecha de Presentación:
14 de noviembre de 2024

Tandil, Buenos Aires, 2024

Índice

METODOLOGÍA

Descripción del conjunto de datos: Son 3231 muestras de vino obtenidas mediante pruebas fisicoquímicas en la bodega Del Sol, elaboradas a partir de dos tipos de uva.

Técnicas y herramientas utilizadas:

- Análisis univariado, bivariado y multivariado
- Métodos de clustering
- Test de hipótesis paramétricos y no paramétricos
- Reducción de Dimensionalidad
- Visualización de Datos
- Visual Studio Code
- Python

Análisis Exploratorio de Datos

Caracterización de Atributos: las muestras tenían asignadas 13 columnas de las cuales 1 era una variable cualitativa nominal, 11 son variables cuantitativas continuas que representan la prueba fisicoquímica y 1 variable cuantitativa discreta que asigna un valor de calidad al vino (se desconoce si está ligada a la prueba fisicoquímica o proviene de una cata de los vinos y fue un valor de puntuación que un grupo de personas le asignó).

Comportamiento de los Datos: Para analizar el comportamiento de los datos recurrimos a fuentes en internet que nos puedan ayudar a entender que valores de las variables fisicoquímicas tienen sentido. Vimos presencia de outliers en diferentes boxplots pero muchos de esos tenían valores razonables en el contexto de una prueba fisicoquímica de vinos.

Por otro lado investigamos acerca de los repetidos, los cuales encontramos muchos alrededor de un 18% del dataset. Decidimos eliminarlos ya que en una prueba tan exacta estas filas repetidas deben no ser de vinos sino de algún error en la carga y son ruidos para nuestro análisis.

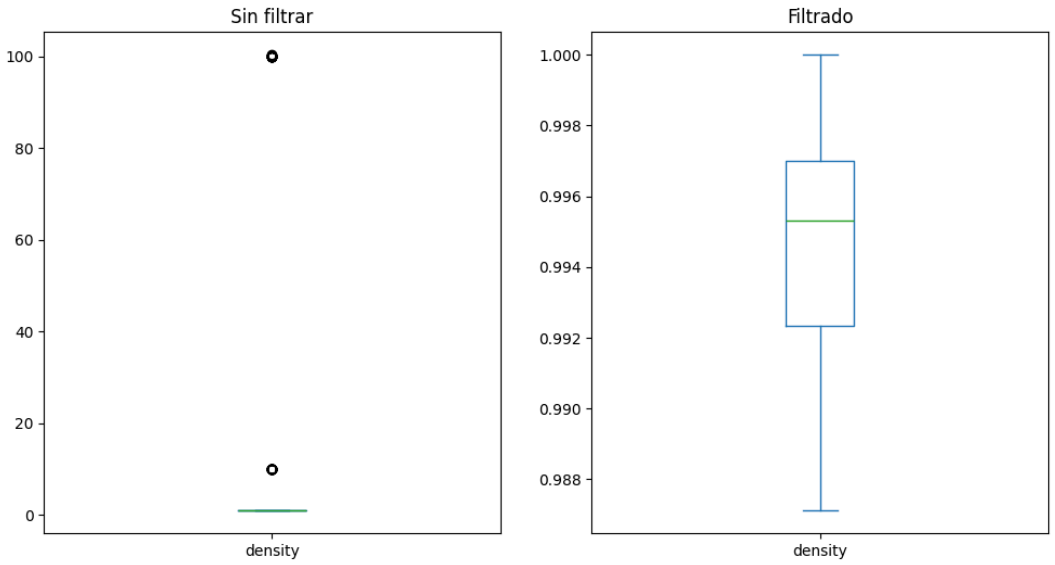
Limpieza de Datos

Acciones correctivas y justificación: Encontramos dos columnas que tuvimos que corregir dado que los valores que tomaban las muestras no cumplian con la condicion de la muestra o con un sentido físico químico

- "alcohol" medida de graduación alcohólica entre 0 a 100 en porcentaje. Tuvimos que eliminar las filas que superan este valor mayor a 100 ya que consideramos que era una muestra errónea.

```
object
alcohol
<class 'str'>    2579
Name: count, dtype: int64
['11.8' '10.2' '12.7' '9.4' '12' '11.3' '10.9' '9.8' '12.2' '11.1' '11.4'
 '12.8' '10.1' '12.9' '9.2' '12.3' '13.5' '11' '10.6' '12.4' '9.1' '11.2'
 '10.4' '12.5' '11.6' '12.6' '11.5' '10.5' '12.1' '9.9' '9.5' '10.8'
 '10.7' '13.2' '10.3' '8.8' '13.1' '13' '10' '13.9' '9.6' '13.4' '11.7'
 '13.6' '13.3' '13.7' '9.7' '9' '11.9' '9.3' '8.6' '8.4' '8.9' '14' '14.2'
 '8.7' '8.5' '11.94' '128.933.333.333.333' '114.666.666.666.667' '10.98'
 '100.333.333.333.333' '114.333.333.333.333' '105.333.333.333.333'
 '953.333.333.333.333' '109.333.333.333.333' '113.666.666.666.667'
 '113.333.333.333.333' '110.666.666.666.667' '973.333.333.333.333' '11.05'
 '9.75' '11.35' '9.55' '10.55' '11.45' '14.05' '123.333.333.333.333'
 '12.75' '13.8' '12.15' '13.05' '112.666.666.666.667'
 '105.666.666.666.667' '117.333.333.333.333' '11.75' '10.65'
 '109.666.666.666.667' '101.333.333.333.333' '10.15' '104.666.666.666.667'
 '116.333.333.333.333' '12.25' '11.85' '11.65' '13.55'
 '131.333.333.333.333' '11.95' '120.666.666.666.667' '11.55'
 '963.333.333.333.333' '12.05' '14.9' '956.666.666.666.667'
 '135.666.666.666.667' '9.95' '923.333.333.333.333' '9.25' '9.05' '10.75']
```

- “density” medida en miligramos por centímetro cúbico la cual tuvimos que eliminar algunas filas ya que tomaban valores que no son posibles de alcanzar nuestra medida de referencia fue el agua y en comparación con el vino sacado de esta fuente (https://www.larioja.com/opinion/densidad-vino-20210611214244-nt.html?ref=https%3A%2F%2Fwww.larioja.com%2Fopinion%2Fdensidad-vino-20210611214244-nt.html)



FORMULACION DE HIPOTESIS

Selección y Justificación de Hipótesis: Posterior al trabajo univariado de entender nuestras variables y pulir las muestras para quedarnos con un dataset lo más limpio posible, pensamos y elaboramos un camino de formulación de hipótesis.

La idea principal fue poder contar algo sobre la calidad del vino de estos datos.

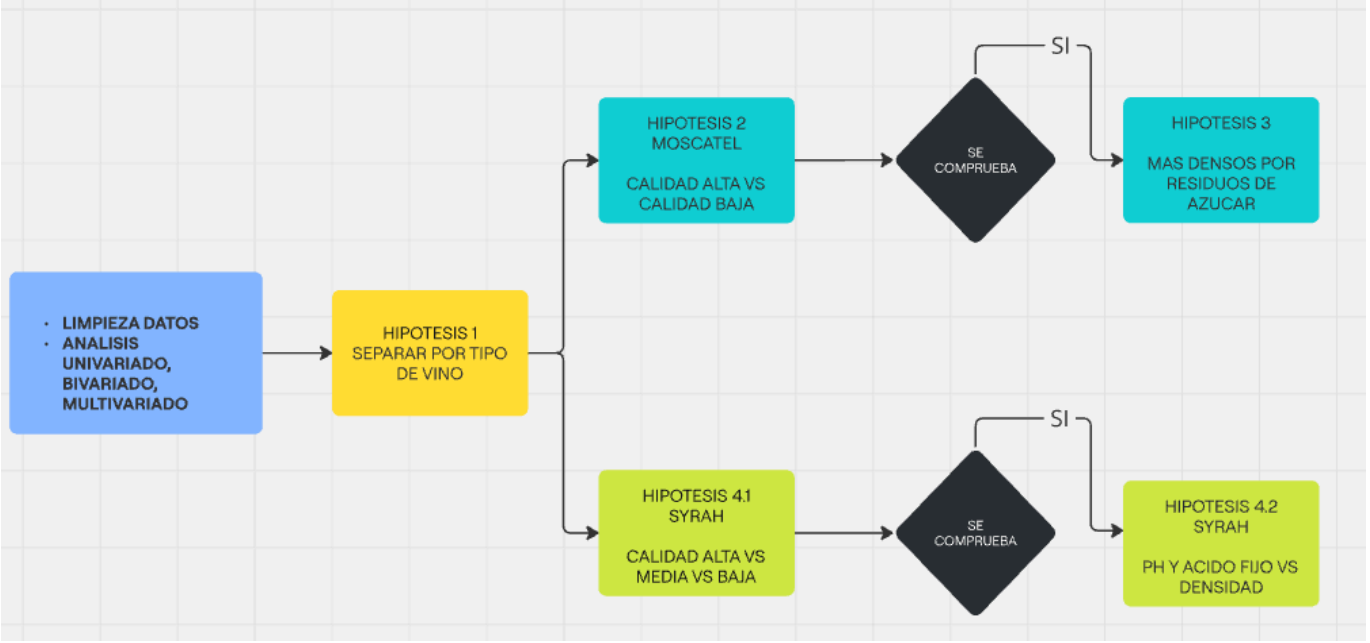
Previo a ello hicimos un análisis bivariado del dataset completo, viendo si existían correlaciones fuertes, estudiando si podíamos encontrar algún agrupamiento en los datos que nos pueda llevar a investigar y caracterizar de manera más clara el dataset ya que las muestras son muchas.

Encontramos que el dataset se podía dividir en 2 grupos por tipo vino y procedimos a comprobar que existían diferencias significativas para poder separar el dataset en un dataset con los vinos moscatel y otro con los vinos Syrah.

Luego nos enfocamos en la calidad del vino, creemos que es la variable que puede determinar acciones o medidas a tomar a futuro. Esta es la variable que nos puntúa el vino, de alguna manera podemos ver si un vino es mejor o no que el otro.

Tomamos la decisión de clasificar los dataset de los vinos por separado en 3 grupos en base al número de calidad del mismo. Los grupos son de calidad alta, media y baja. Siempre pensando en tener una división que sea significativa entre la cantidad de muestras de cada grupo.

Ruta de hipótesis



VALIDACIÓN DE HIPÓTESIS

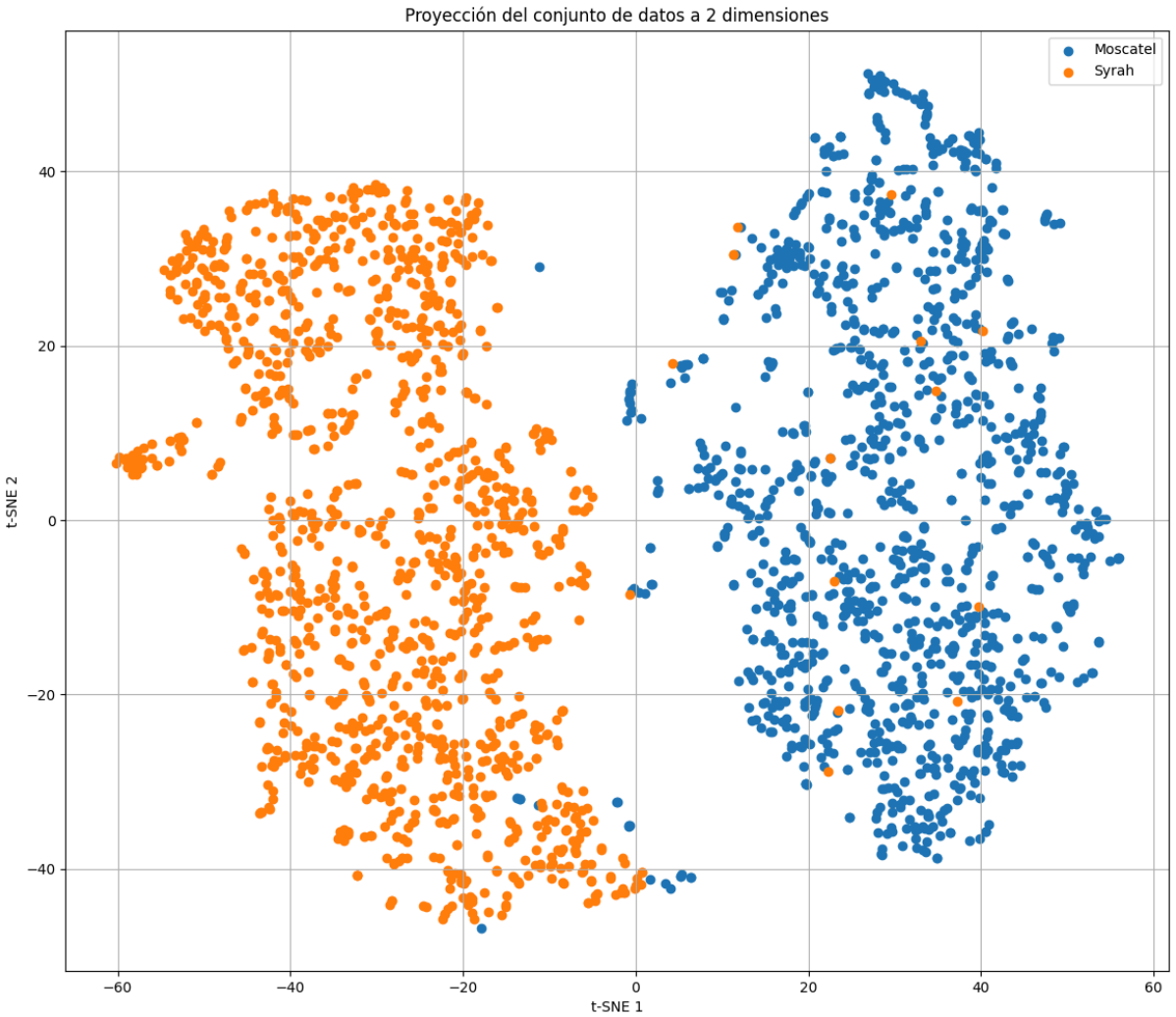
HIPÓTESIS 1

“ Los resultados químicos de las variables (todas menos calidad y tipo de vino) determinan el tipo de vino.”

Hipótesis nula: Las variables químicas no determinan el tipo de vino

Hipótesis alternativa: Las variables químicas determinan el tipo de vino

Previo a plantear esta hipótesis, pudimos observar utilizando la técnica de reducción de dimensionalidad T-SNE que las muestras se organizaban en dos grandes grupos, los cuales en un scatter plot pintamos de acuerdo a el tipo de uva.



Este gráfico parece decirnos que los vinos se están agrupando por el tipo de uva, por lo que decidimos utilizar el test de hipótesis de Kruskal Wallis para cada variable química. Kruskal Wallis nos dice en su hipótesis nula que las distribuciones de las variables químicas son iguales para los tipos de uva y en su hipótesis alternativa que las distribuciones de las variables químicas no son iguales para los tipos de uva.

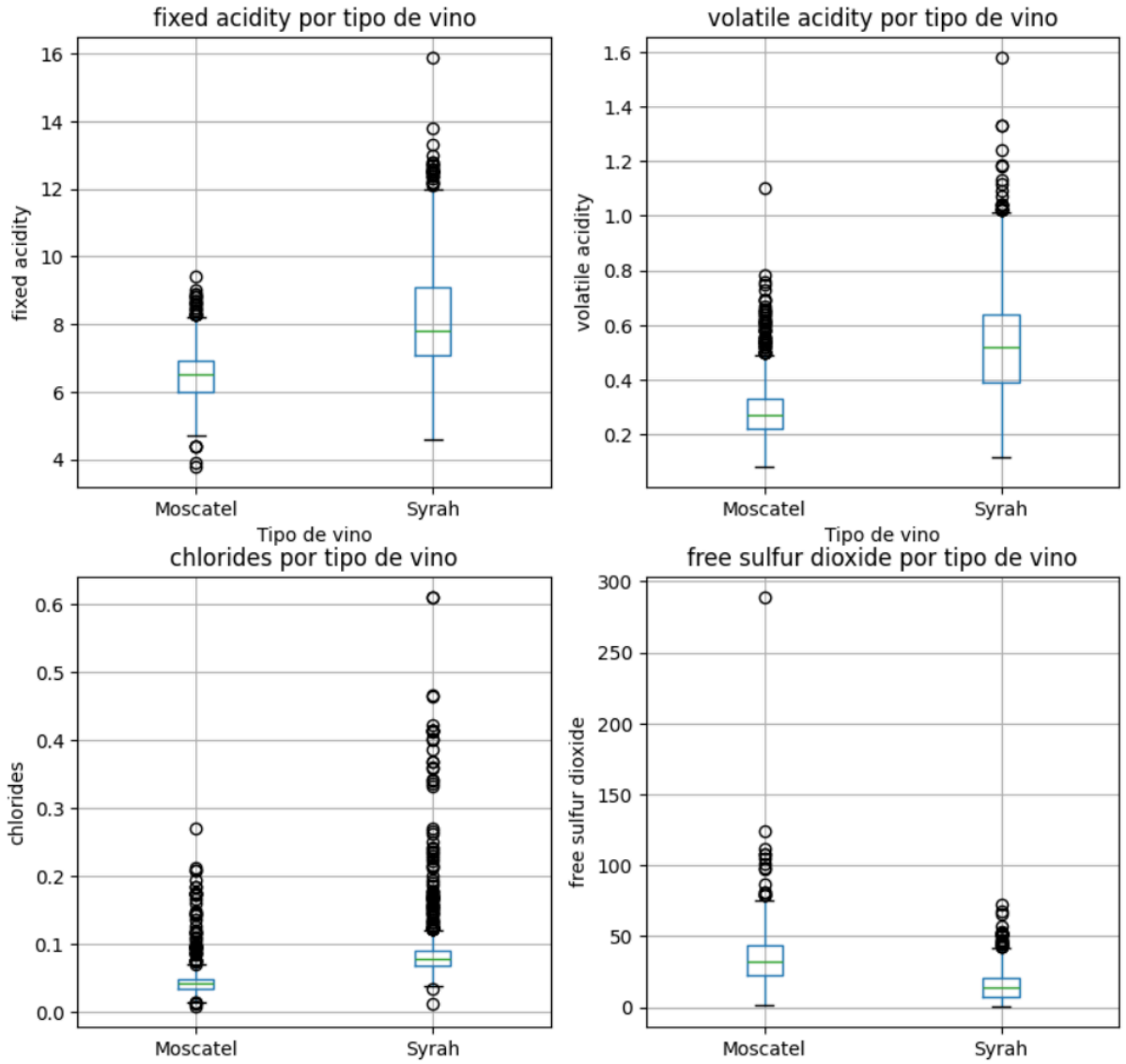
Si el test nos devuelve un P valor mayor a 0.05 quiere decir que las distribuciones de las variables químicas son iguales para los tipos de uva

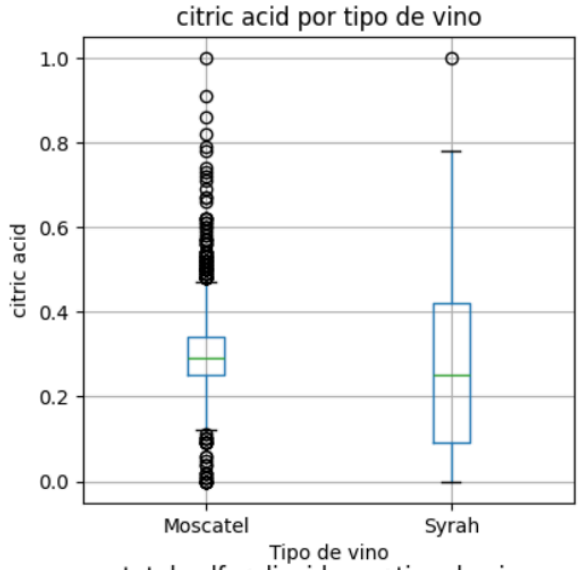
Si el test nos devuelve un P valor menor a 0,05 se rechaza la hipótesis nula y la distribución de las variables químicas son distintas por lo

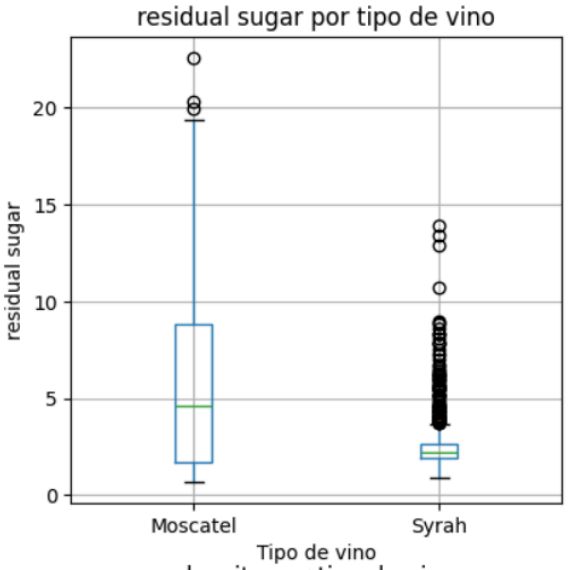
que estas estarían determinando el tipo de uva con el que se realizó el vino.

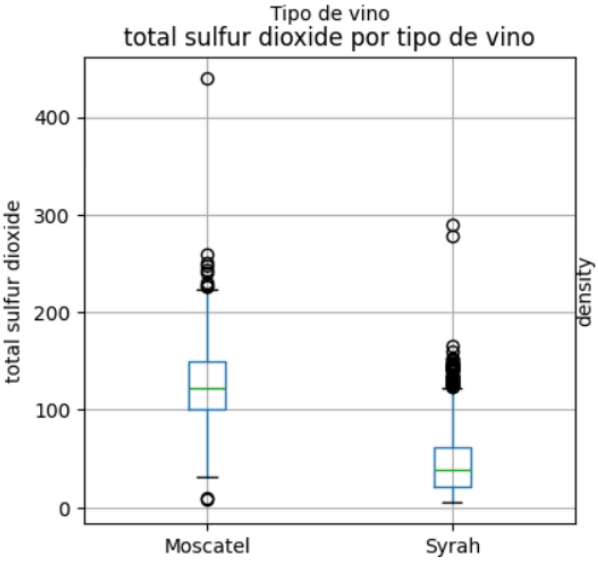
El test rechazó la hipótesis nula y las distribuciones de las variables químicas son distintas por tipo de uva.

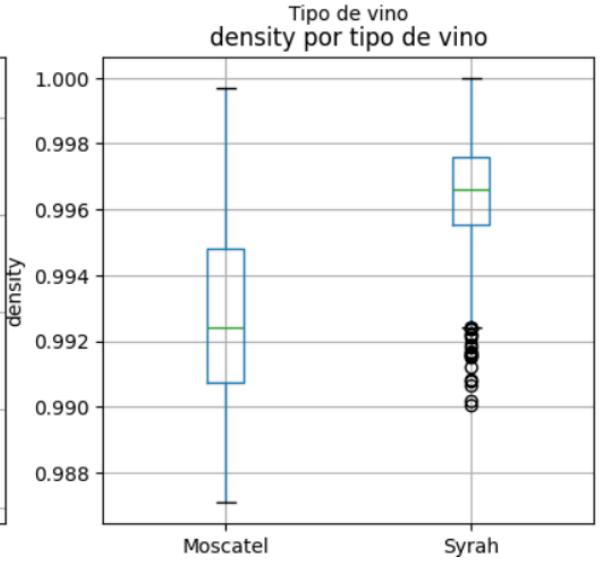
Mediante estos boxplot hechos durante el análisis univariado pudimos llegar a esta hipótesis.

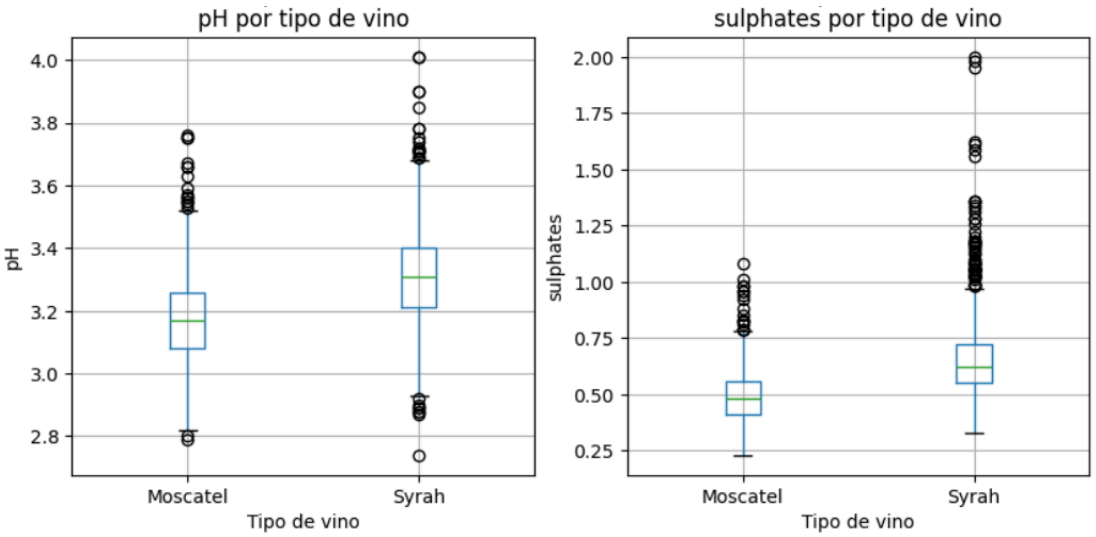


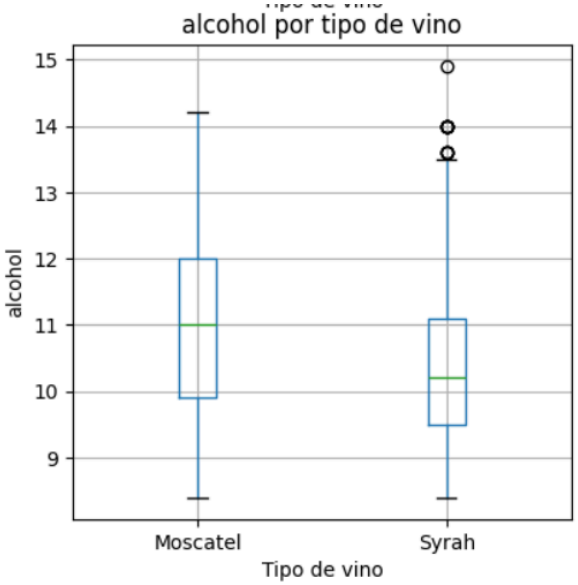

citric acid por tipo de vino
citric acid
1.0
0.8
0.6
0.4
0.2
0.0
Moscatel
Syrah
Tipo de vino


residual sugar por tipo de vino
residual sugar
20
15
10
5
0
Moscatel
Syrah
Tipo de vino


total sulfur dioxide por tipo de vino
total sulfur dioxide
400
300
200
100
0
Moscatel
Syrah


density por tipo de vino
density
1.000
0.998
0.996
0.994
0.992
0.990
0.988
Moscatel
Syrah





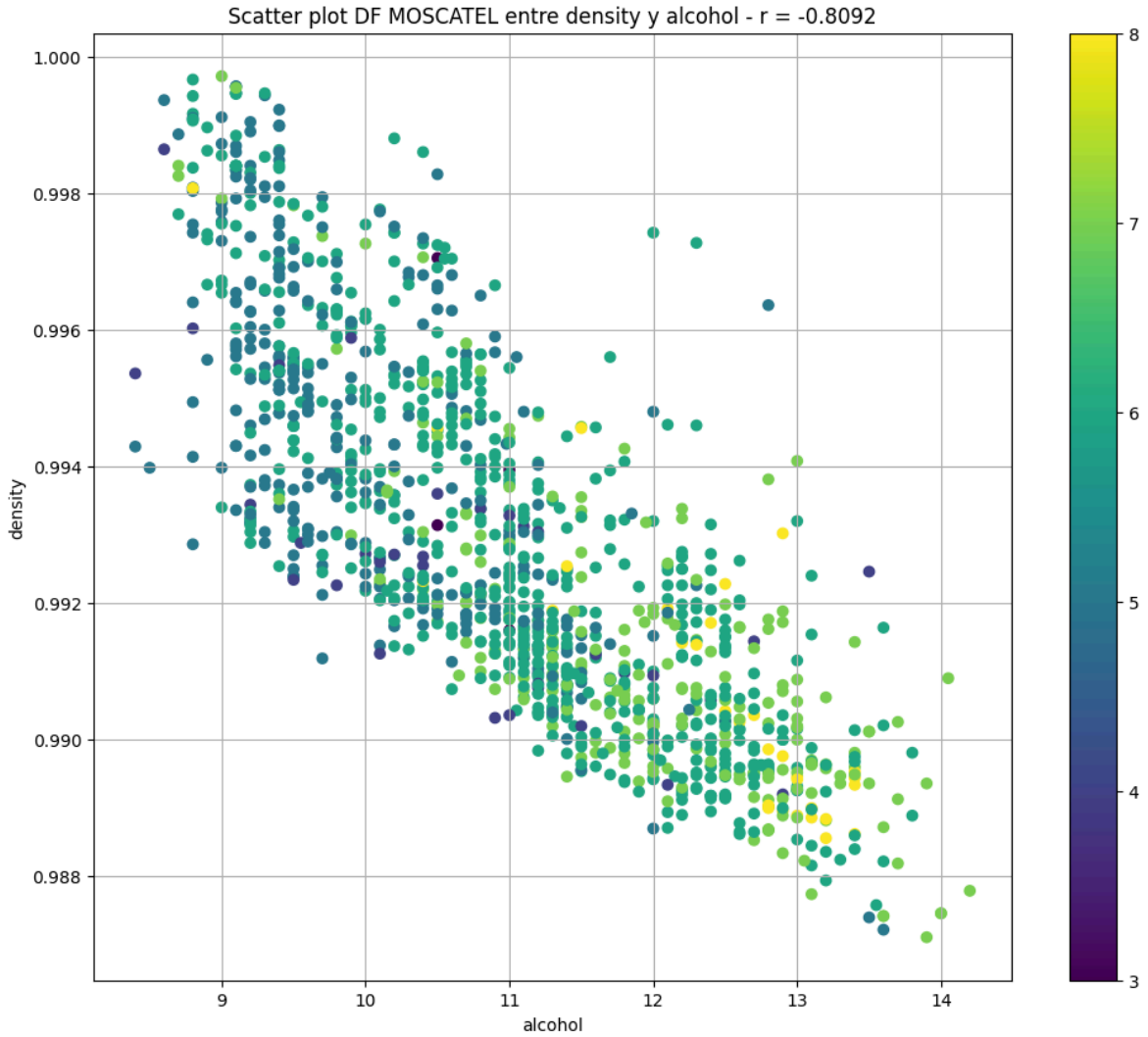
Conclusión hipótesis 1: Vimos que hay diferencias significativas entre los tipos uva con respecto a las variables químicas.

Esto nos abre las puertas a seguir analizando el dataset pero por tipo de uva. Por un lado analizaremos las muestras que correspondan a los tipos de uva Moscatel y por el otro a los Syrah. Esto también nos permite hacer los datos ya que el dataset nos queda con 1298 muestras para Syrah y 1249 para Moscatel.

HIPÓTESIS 2 - Moscatel

“Existen diferencias significativas entre los grupos de calidad de vino Alta vs Baja para los vinos M”

Contexto: Análisis sobre los vinos Moscatel. Comenzamos viendo las correlaciones más fuertes donde se pudo destacar la correlación entre la variable densidad y alcohol. La podemos observar en este ScatterPlot que está pintado con la calidad.



Notamos que la densidad y el alcohol tienen una correlación fuerte negativa y que la calidad pareciera querer seguir esa linealidad.

La variable calidad es una variable cuantitativa discreta por lo que se nos ocurrió utilizarla para clasificar los vinos Moscatel en grupos de calidad alta, media y baja. Aquí nace nuestra segunda hipótesis.

Hipótesis nula: No existen diferencias significativas en el alcohol y la densidad entre los grupos de calidad de vino Alta vs Baja para los vinos Moscatel.

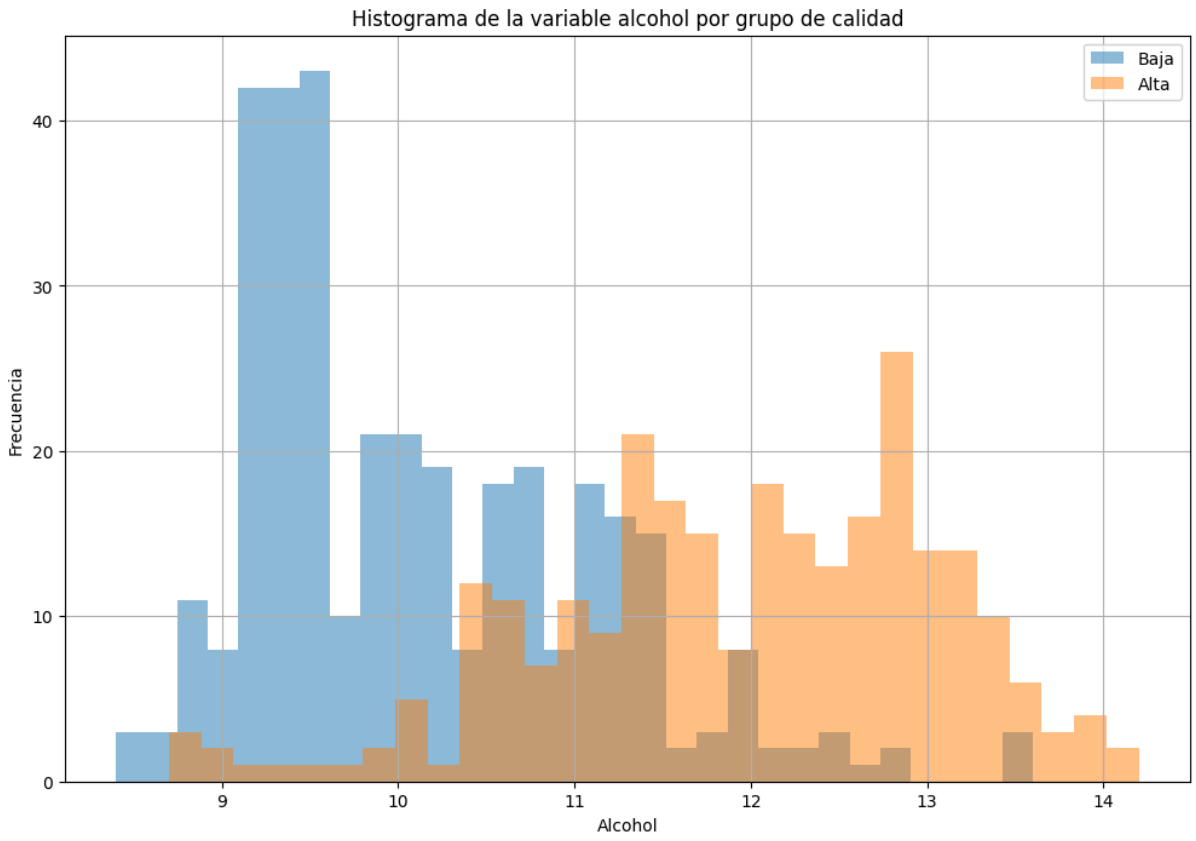
Hipótesis alternativa: Existen diferencias significativas en el alcohol y la densidad entre los grupos de calidad de vino Alta vs Baja para los vinos Moscatel.

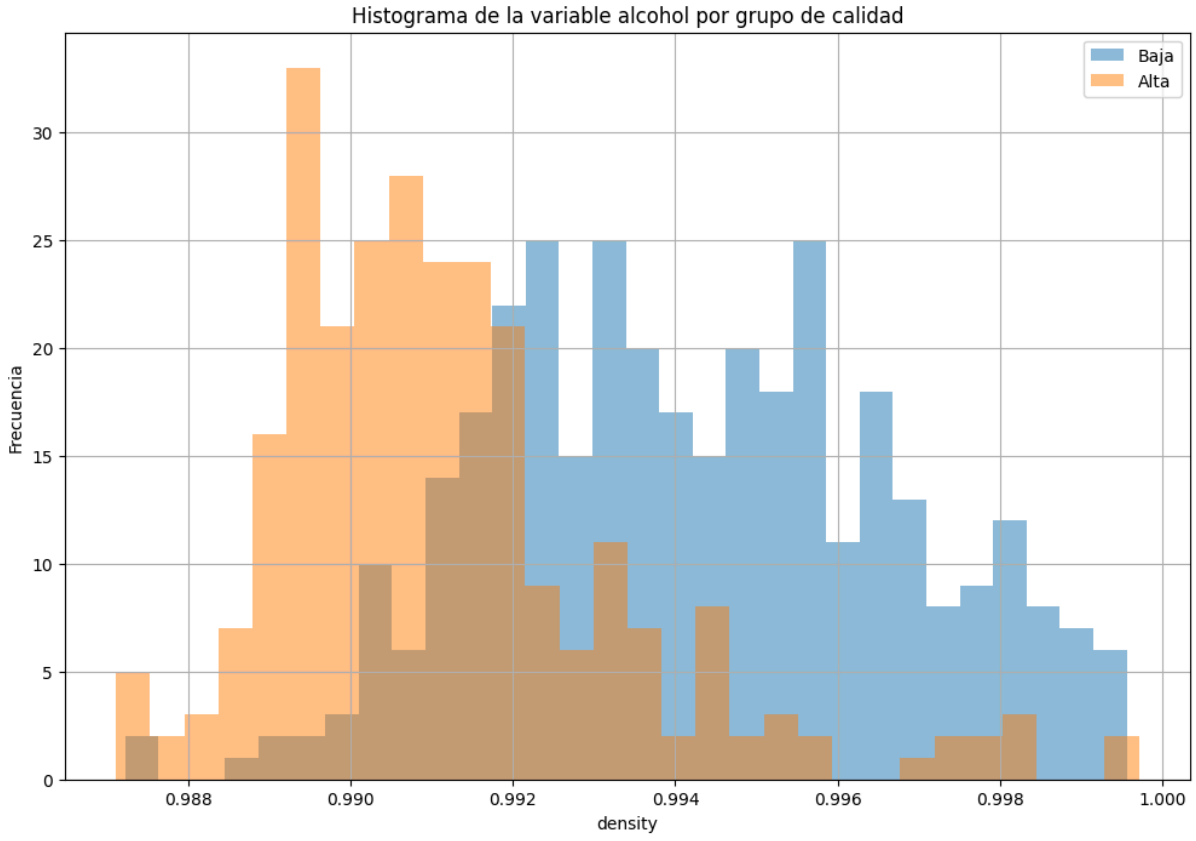
Primero procedimos a separar los vinos Moscatel de manera que nos quedarán equitativos los grupos. En base a esta decisión elegimos ver si los grupos de alta y baja calidad tenían diferencias significativas en sus valores de alcohol y densidad. Si las diferencias son significativas podemos ver que son dos variables que influyen en la calidad del vino.

Los grupos quedaron divididos de la siguiente manera, los límites de los grupos son calidad baja entre 0 y 5, los de calidad media 6 y los de calidad alta entre 7 y 10.

```
quality_group
Baja     351
Media    629
Alta     269
```

Veamos un poco gráficamente estos grupos, decidimos ver un histograma por separado de cada variable (alcohol y densidad) y ver por donde se están dando los valores dependiendo del grupo.


Histograma de la variable alcohol por grupo de calidad
Baja
Alta
Frecuencia
Alcohol
0
10
20
30
40
9
10
11
12
13
14


Histograma de la variable alcohol por grupo de calidad
Baja
Alta
Frecuencia
density
0
5
10
15
20
25
30
0.988
0.990
0.992
0.994
0.996
0.998
1.000

Pareciera que la distribución de las variables en cada grupo se están moviendo entre rangos acorde al ScatterPlot que vimos anteriormente donde los vinos de menor calidad eran los más densos y con menor graduación alcohólica y los de mayor calidad se movían en graduaciones alcohólicas altas y densidades bajas.

Pero serán estas diferencias grandes significativamente para poder decir que son determinantes a la hora de hablar de la calidad del vino Moscatel.

Para ellos probamos el test de Kruskal Wallis (dado que no se cumplieron los supuestos de normalidad y homocedasticidad) para las variables mencionadas.

Test de Kruskal-Wallis para la variable 'alcohol' en Moscatel: Estadístico=267.223, p-valor=0.000

Test de Kruskal-Wallis para la variable 'density' en Moscatel: Estadístico=208.715, p-valor=0.000

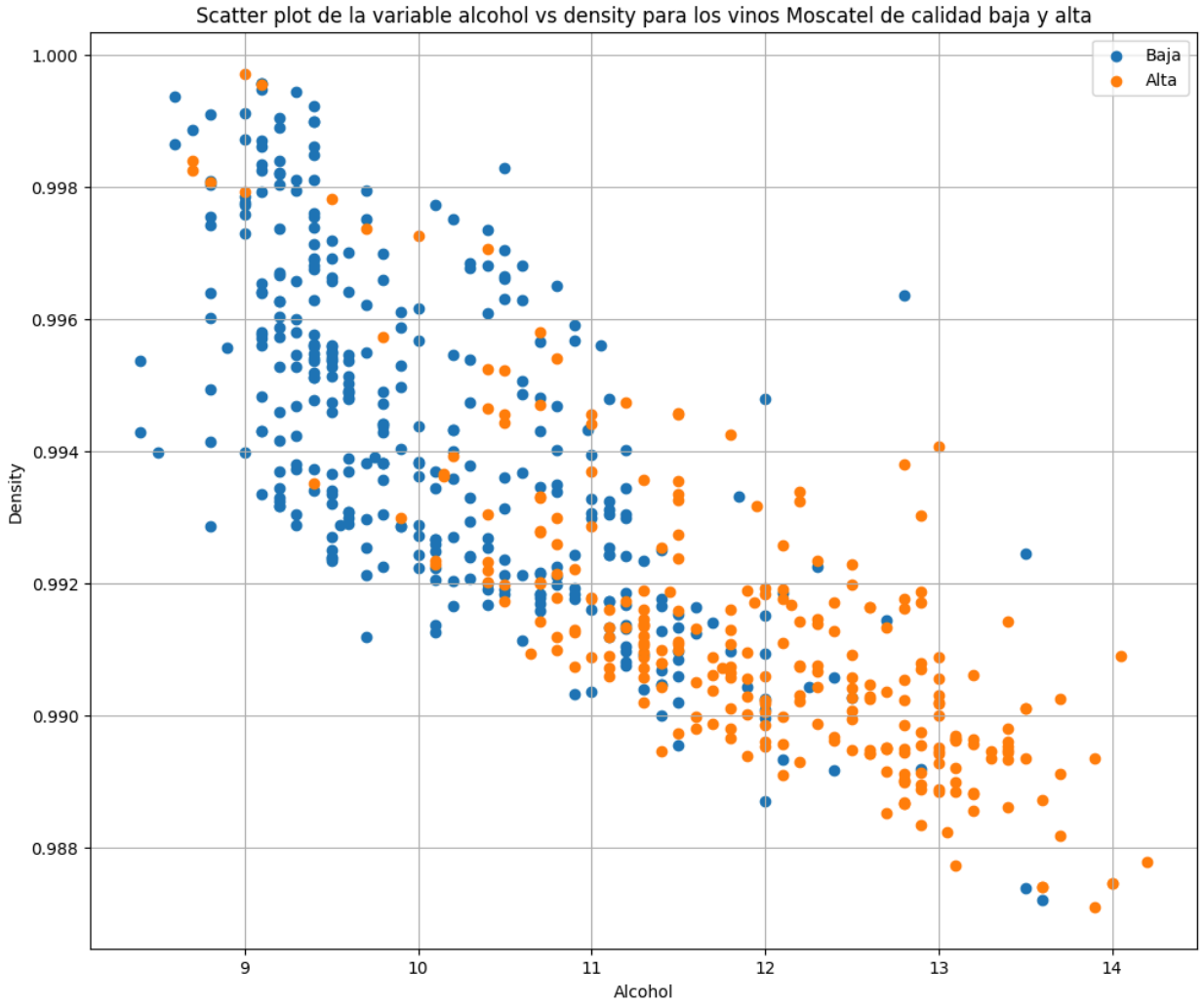
Hay suficiente evidencia para rechazar la hipótesis nula.

Las medias de las variables 'alcohol' y 'density' son diferentes en los grupos de calidad baja y alta.

Conclusión hipótesis 2

Existen diferencias significativas de las variables alcohol y densidad en los grupos de calidad alta y baja por lo que estas variables son determinantes cuando hablamos de la calidad de vino Moscatel

Veámoslo gráficamente con el siguiente scatter plot, donde en el eje y tenemos la densidad y en el eje x el alcohol y pintamos los puntos del grupo de calidad baja y lo del grupo de calidad alta.

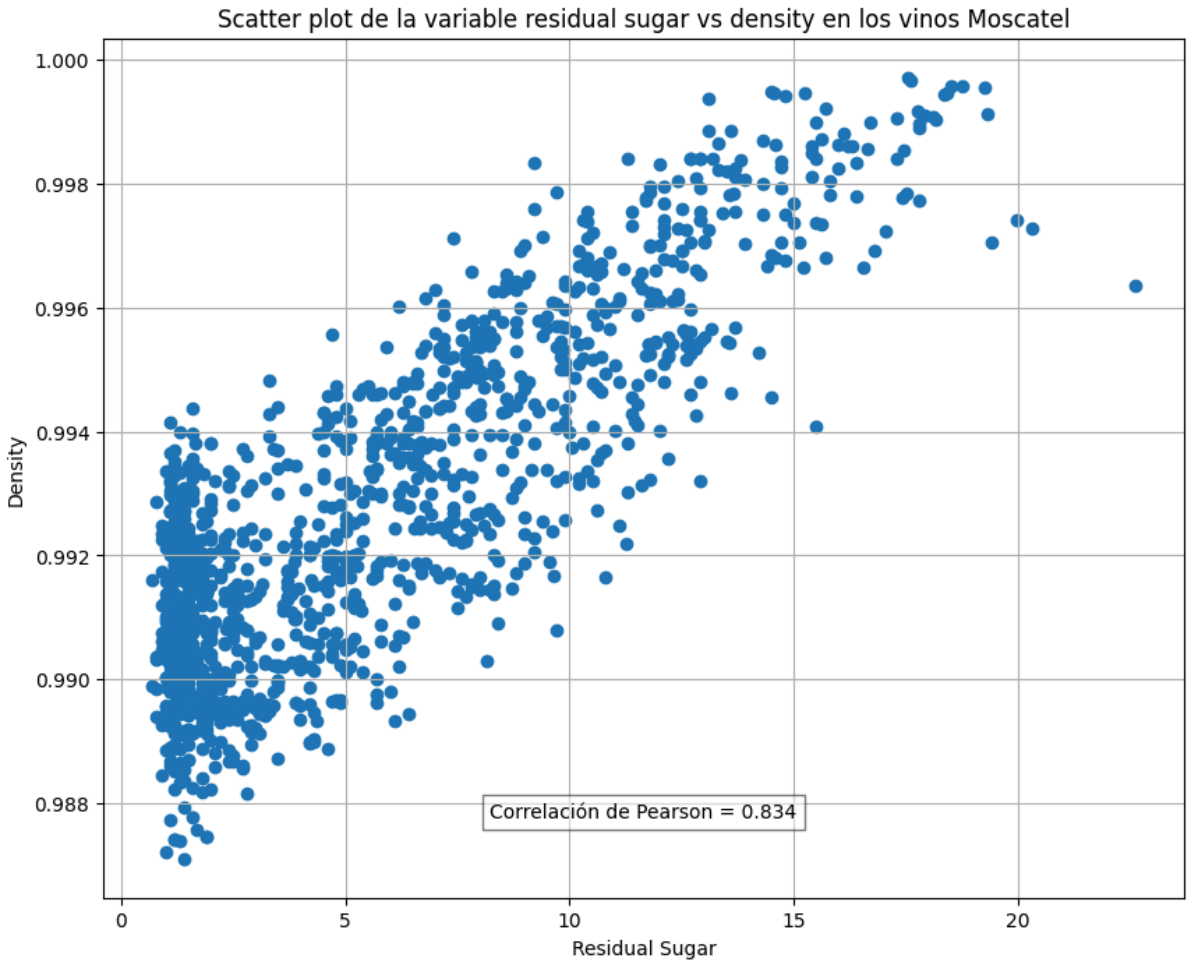


HIPÓTESIS 3

“ Los vinos Moscatel más densos tienen una diferencia significativa de residuos de azúcar que los menos densos.”

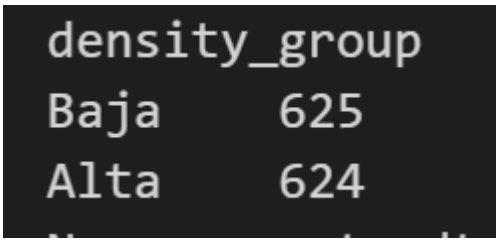
A partir de ver que la densidad es una variable importante en la calidad del vino, intentamos entender a qué se puede deber esta densidad.

Mediante la correlación de Pearson aplicada sobre un dataset de vinos Moscatel vimos que la densidad tiene una fuerte correlación positiva con los residuos de azúcar. Observemos mediante un ScatterPlot, a medida que la densidad aumenta los residuos de azúcar también, esto no nos habla de una causa y efecto, simplemente que son variables correlacionadas.

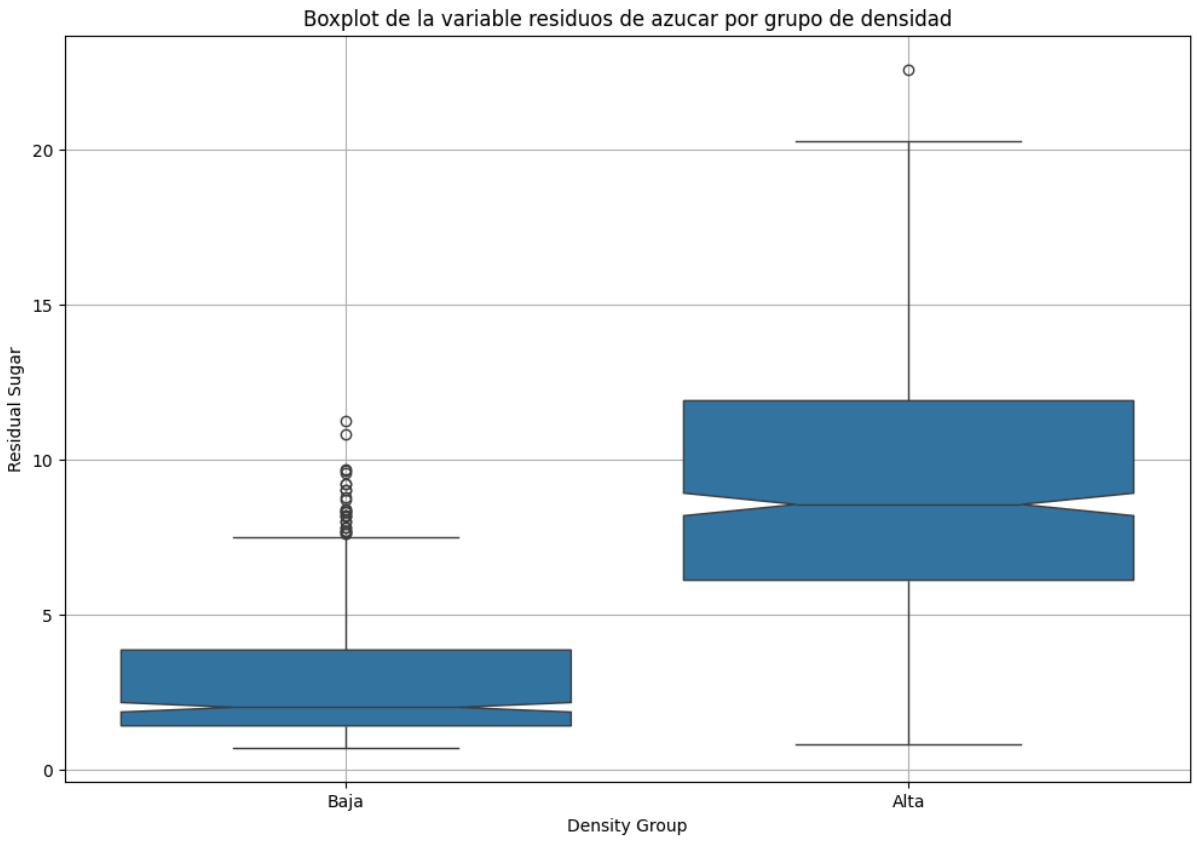


Creemos que la diferencia de residuos de azúcar en los vinos más densos es significativa pero debemos comprobarlo.

Separamos los vinos Moscatel en dos grupos pero ahora por los más densos y los menos densos. Ahora queremos saber si los vinos más densos tienen una diferencia significativa en residuos de azúcar que los vinos menos densos. Esto nos ayudará a entender más la densidad de los vinos Moscatel y poder a futuro trabajar sobre esto y tenerlo en cuenta a la hora de elaborar un vino porque vimos en la hipótesis anterior que los vinos de mayor calidad son menos densos que los vinos de menor calidad.



Separamos por la mediana así los datos nos quedan bien distribuidos y realizamos un boxplot de estos dos grupos por la variable residual sugar.



Bien, vemos que el grupo de densidad baja tiene ciertos outliers en cuanto a los residuos de azúcar pero no los quitamos porque son muestras válidas.

Planteamos las hipótesis y probamos mediante Kolmogorov-Smirnov (ya que los supuestos de normalidad y homocedasticidad no se cumplen). Este test comprueba que las distribuciones de residuos de azúcar en los grupos de densidad baja y alta son diferentes significativamente.

Hipótesis nula: No existen diferencias significativas en los residuos de azúcar de los vinos Moscatel de densidad alta y baja

Hipótesis alternativa: Existen diferencias significativas en los residuos de azúcar de los vinos Moscatel de densidad alta y baja

Test de Kolmogorov-Smirnov para la variable 'residual sugar' en los grupos de densidad baja y alta: Estadístico=0.692, p-valor=0.000

Hay suficiente evidencia para rechazar la hipótesis nula.

Las distribuciones de la variable 'residual sugar' son diferentes en los grupos de densidad baja y alta.

Conclusión hipótesis 3: Ahora podemos decir que además de tener una correlación positiva fuerte también hay una diferencia significativa en los residuos de azúcar entre los vinos Moscatel más densos y menos densos.

HIPÓTESIS 4

Introducción a la hipótesis 4: Esta hipótesis involucra dos hipótesis por separado. Una es continuación de la otra. Esto se debe a que al igual que con los vinos Moscatel nos interesó ver qué sucedía si dividimos por calidad los vinos Syrah y ver si la densidad y el alcohol son significativamente diferentes entre los grupos.

“ Los grupos de calidad de vino Syrah (alta, media y baja) tienen diferencias significativas de alcohol y densidad ”

Al haber probado esta hipótesis fuimos a buscar un patrón químico que pueda explicar la densidad y escribimos esta hipótesis.

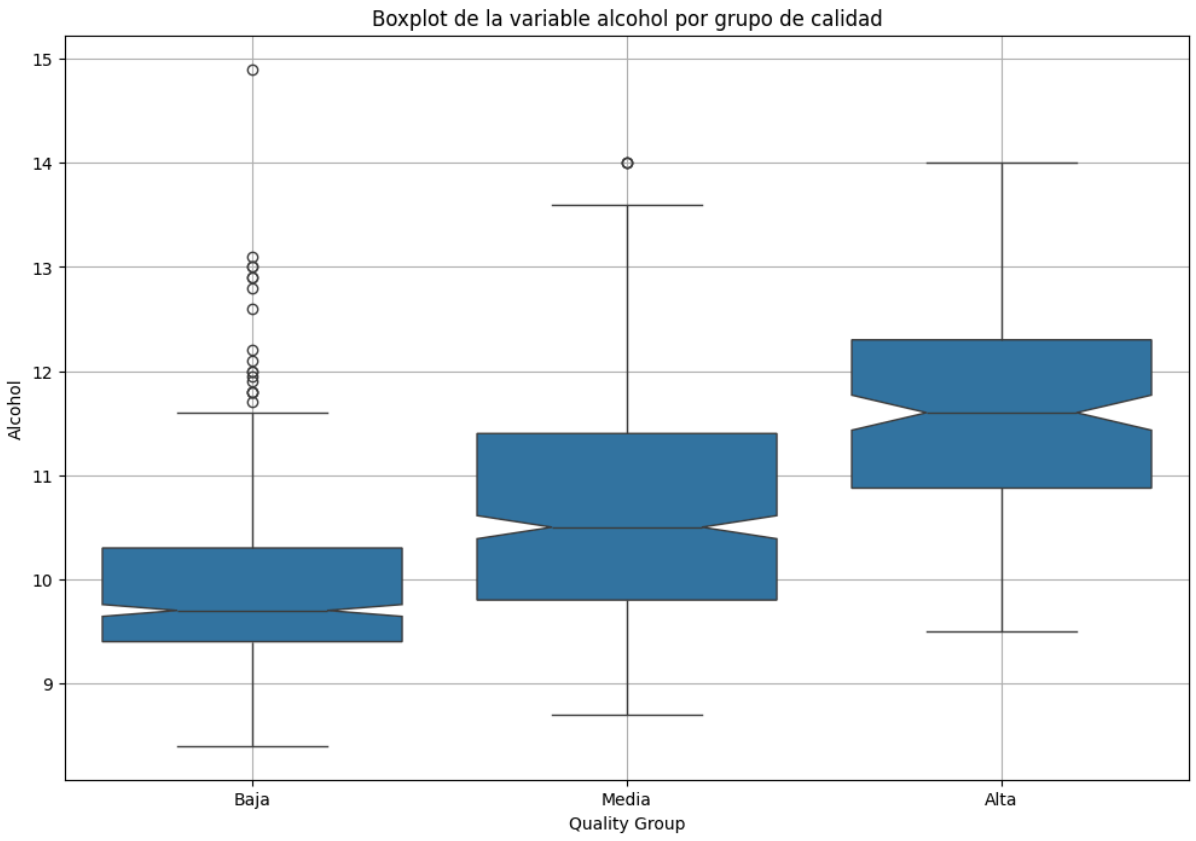
“ Los vinos Syrah de mayor ph y menor acido fijo son menos densos que los de menor ph y mayor acido fijo”

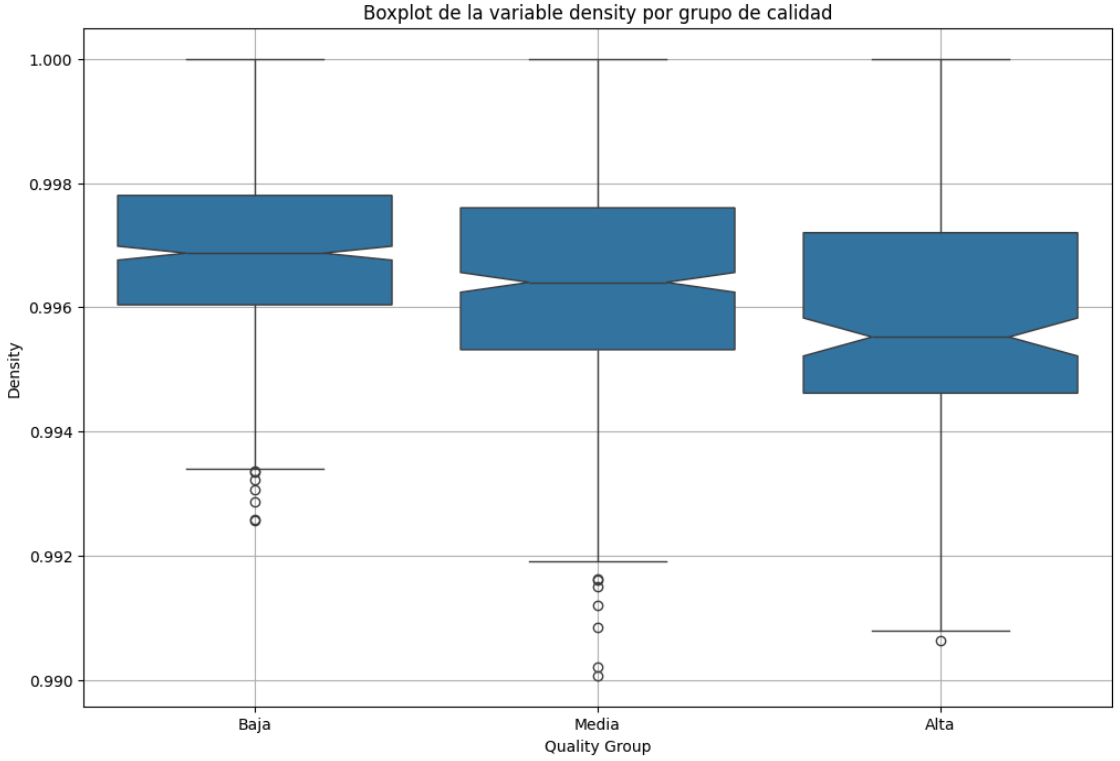
Desarrollo de la hipótesis

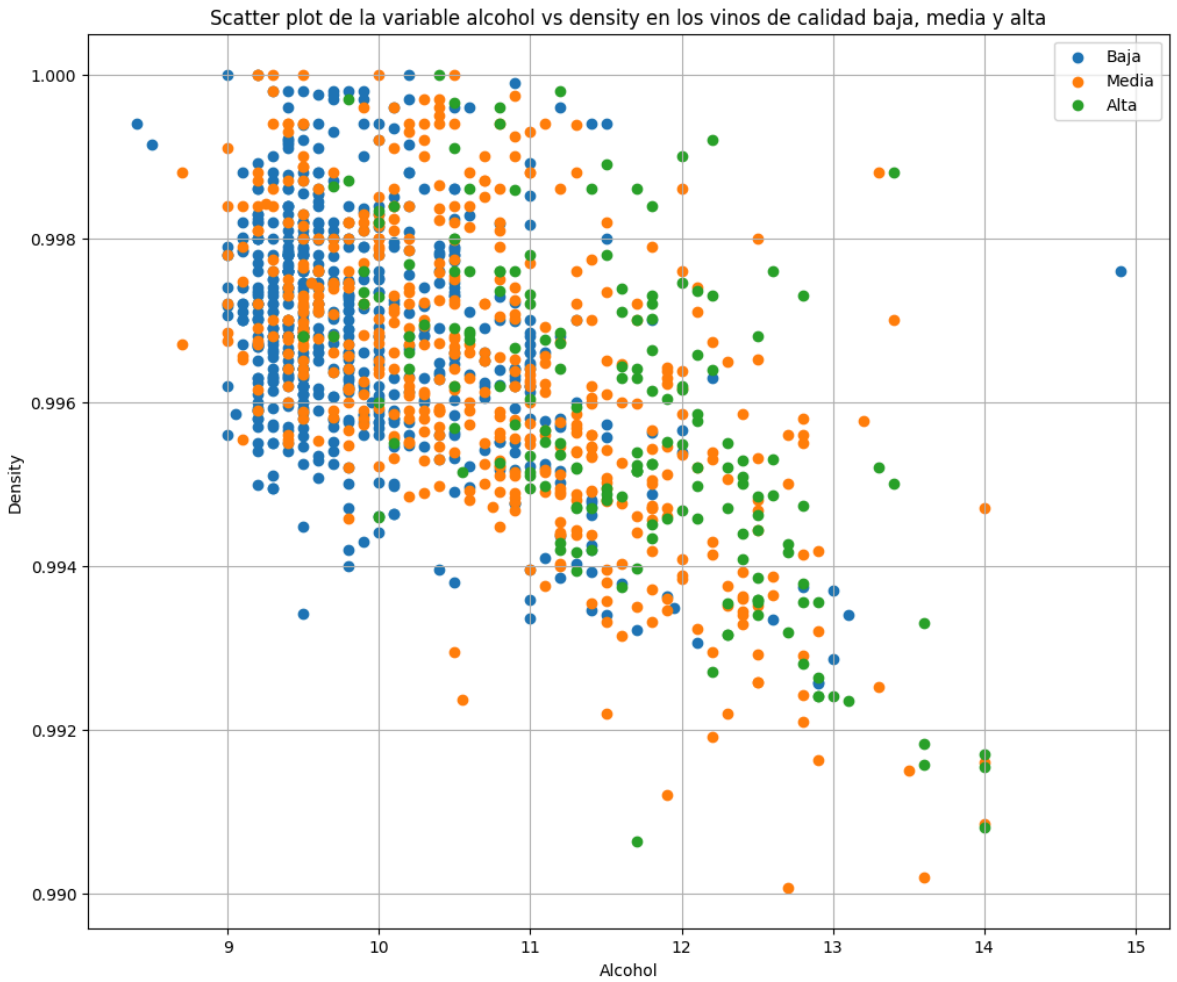
Para nuestra primera hipótesis bajo el mismo criterio que utilizamos en los vinos Moscatel dividimos los vinos Syrah en calidad alta (calidad con valores 7,8) calidad media (6) y calidad baja (entre 0 y 5).

Esta vez la división no quedó lo más equitativa posible, los de calidad baja quedaron con 612 muestras, los media con 510 y los de alta con 176. Aun así cada grupo tiene un número significativo para ser representado.

Veamos gráficamente cómo se distribuyen las variables alcohol y densidad en los grupos y hagamos un scatterplot pintando por grupo los puntos de densidad y alcohol


Boxplot de la variable alcohol por grupo de calidad
Alcohol
15
14
13
12
11
10
9
Baja
Media
Alta
Quality Group


Boxplot de la variable density por grupo de calidad
Density
1.000
0.998
0.996
0.994
0.992
0.990
Baja
Media
Alta
Quality Group



Si graficamos en un scatterplot los grupos pareciera que no vamos a encontrar una diferencia significativa porque los grupos conviven muy juntos y los valores que toman difieren en las centésimas pero con el boxplot que vimos anteriormente si se puede ver mejor esta diferencia, ambos gráficos nos ayudaron a entender esta hipótesis. Aunque en uno es más claro que en el otro.

Bien, las distribuciones parecen indicar que hay una diferencia en ambas variables y que además si existiese sigue el patrón de los vinos Moscatel, los vinos con mayor graduación alcohólica y menos densos son de más calidad.

Mediante el test estadístico de Kruskal-Wallis pudimos encontrar diferencias significativas entre los grupos.

Al test estadístico de Kruskal-Wallis le sumamos un test de Dunn ya que Kruskal-Wallis nos podrá confirmar que hay diferencias significativas en al menos un grupo pero nos interesa que sea en todos. Por ello miramos la matriz de resultado del Test de Dunn y si el p valor es menor a 0,05 se rechaza la hipótesis nula del test de Dunn (no hay diferencias), la matriz tendrá que darnos un p valor menor a 0,05 por cada comparación entre grupos.

Test de Kruskal-Wallis para la variable 'alcohol' en Syrah: Estadístico=354.949, p-valor=0.000

Hay suficiente evidencia para rechazar la hipótesis nula.

Las medias de la variable 'alcohol' son diferentes en los grupos de calidad baja, media y alta.

Resultado del test de Dunn para la variable 'alcohol' en los grupos de calidad baja, media y alta

	Media	Alta	Baja
Media	1.000000e+00	6.479124e-36	1.198340e-66
Alta	6.479124e-36	1.000000e+00	9.017514e-17
Baja	1.198340e-66	9.017514e-17	1.000000e+00

Test de Kruskal-Wallis para la variable 'density' en Syrah: Estadístico=58.094, p-valor=0.000

Hay suficiente evidencia para rechazar la hipótesis nula.

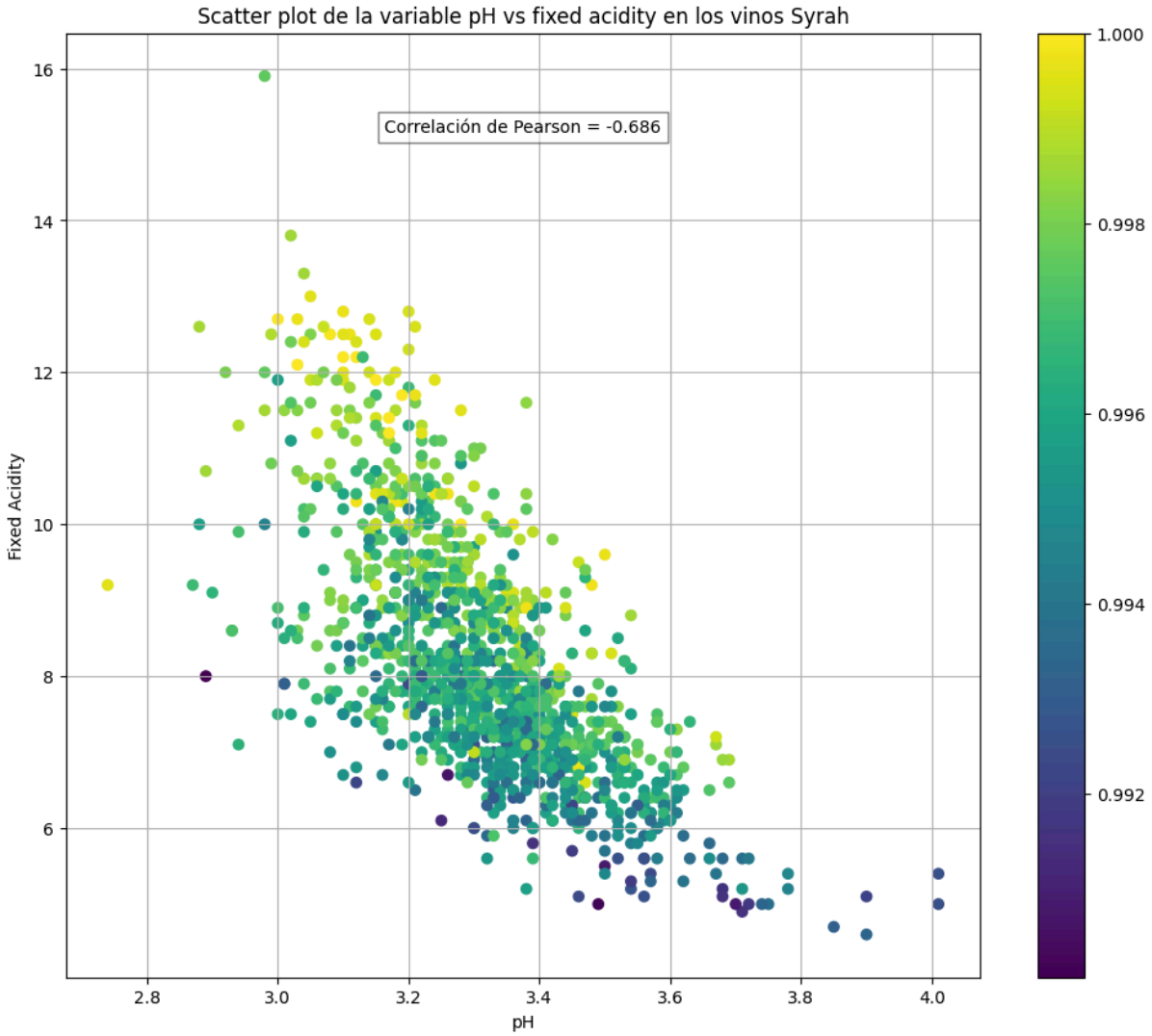
Las medias de la variable 'density' son diferentes en los grupos de calidad baja, media y alta.

Resultado del test de Dunn para la variable 'density' en los grupos de calidad baja, media y alta

	Media	Alta	3
Media	1.000000e+00	0.000004	2.339978e-12
Alta	4.113519e-06	1.000000	1.742964e-04
3	2.339978e-12	0.000174	1.000000e+00

Ahora bien encontramos diferencias significativas en los vinos agrupados por calidad. Podríamos buscar si en los vinos Syrah hay algún patrón químico que marque una diferencia en vinos más o menos densos.

Para ellos tomamos la matriz de correlación de los vinos Syrah y encontramos una correlación fuerte que parece estar ligada a la densidad.



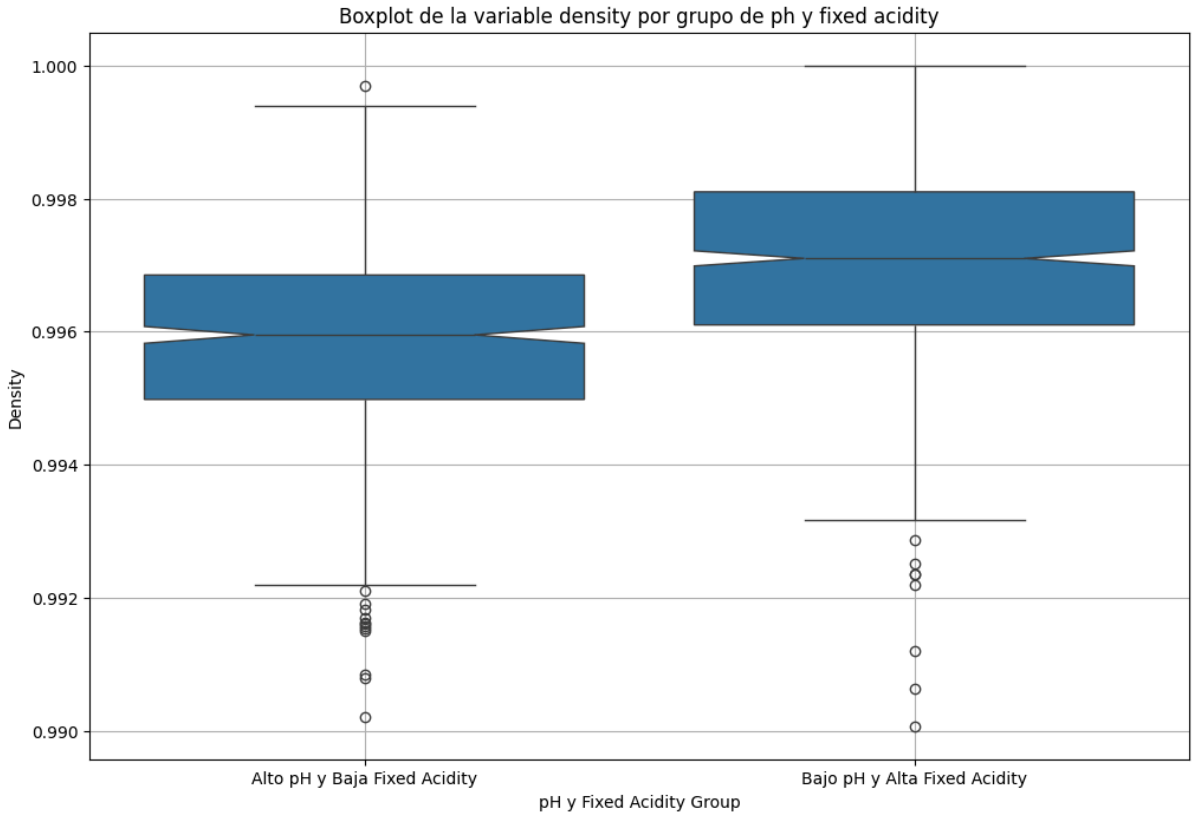
Mediante el scatter plot podemos ver que la correlación es negativa por lo que a mayor pH menor ácido fijo y si pintamos por densidad vemos que los colores menos opacos corresponden a una densidad más alta y viceversa.

Por ello se nos ocurrió separar el dataset de los vinos Syrah en 2, agrupamos por los vinos Syrah con mayor ph y menor acido fijo y con menor ph y mayor acido fijo a ver si esta diferencia que vemos en el scatter plot de la densidad tienen una diferencia que sea significativa.

Agrupamos los datos por la media y nos quedaron dos grupos con buenas cantidades de muestras.

```
ph_fixed_acidity_group
Alto pH y Baja Fixed Acidity    538
Bajo pH y Alta Fixed Acidity    760
```

Veamos mediante un boxplot como se distribuye la densidad en cada grupo.



Bien, parece que vamos a encontrar una diferencia significativa, probemoslo mediante un test.

Test de Mann-Whitney para la variable 'density' en los grupos de ph y fixed acidity: Estadístico=117050.000, p-valor=0.000

Hay suficiente evidencia para rechazar la hipótesis nula.

Las medias de la variable 'density' son diferentes en los grupos de ph y fixed acidity.

Conclusión hipótesis 4: el test nos permite rechazar la hipótesis nula de que la media de la densidad variable es igual en ambos grupos ya que da un p valor menor a 0,05 por lo que la media es distinta. Encontramos que la densidad en estos grupos es significativamente distinta y podemos ver mediante el boxplot que esta es mayor en bajo pH y alto ácido fijo.

CONCLUSIÓN

El trabajo desarrollado tuvo como fin poder analizar los datos crudos de los vinos. Determinar algún comportamiento en base a la calidad del mismo ya que es la variable a mejorar por la Bodega del Sol. Creemos que la calidad del vino es la variable más destacada porque es la que de alguna manera brinda un número asociado a que tan bien están haciendo los vinos. Poder mediante una muestra visualizar posibles problemáticas, comprobar ciertos patrones químicos que se dan y entender estos patrones desde el punto de vista analitico brindara una herramienta para la Bodega del Sol para decidir por qué camino seguir en la producción de vinos.