Objetivo del proyecto:

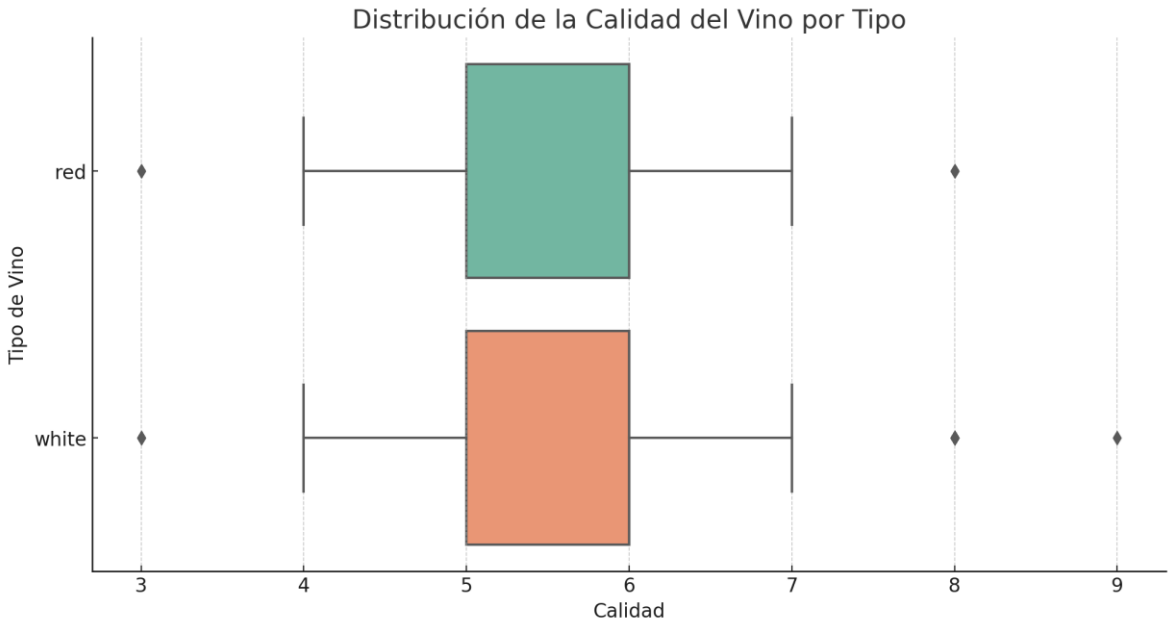
Dado un dataset de vinos blancos y vino tintos, queremos encontrar que variables pueden influir o ser candidatas para afectar la calidad de los vinos. El trabajo será realizado en Python y se documentará el proceso paso a paso en lo posible haciendo especial énfasis al uso de la estadística para acotar posibles insight:

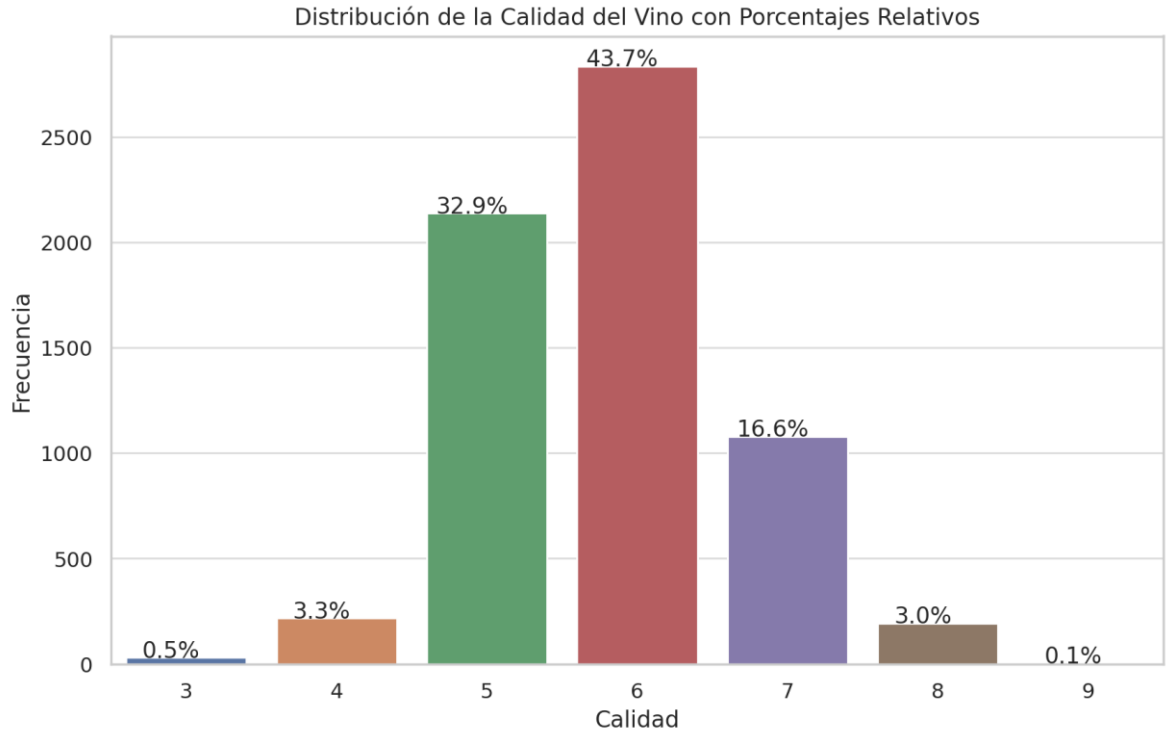
PREPARACION Y EVALUACION PRELIMINAR:

- Cargamos el dataset a partir del archivo: combined_wine_dataset.csv
- Revisamos la información básica del dataset para conocer sus propiedades y ver si tenemos registros nulos. (NO ENCONTRAMOS VALORES NULOS)
    a.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         6497 non-null   float64
 1   volatile acidity      6497 non-null   float64
 2   citric acid           6497 non-null   float64
 3   residual sugar        6497 non-null   float64
 4   chlorides             6497 non-null   float64
 5   free sulfur dioxide   6497 non-null   float64
 6   total sulfur dioxide  6497 non-null   float64
 7   density               6497 non-null   float64
 8   pH                    6497 non-null   float64
 9   sulphates             6497 non-null   float64
 10  alcohol               6497 non-null   float64
 11  quality               6497 non-null   int64
 12  tipo                  6497 non-null   object
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

- Vamos a realizar una función describe() para comprender los valores estadísticos relacionados con la calidad para comprender primero el desempeño de nuestros productos.
    a. Quality: Media: 5.82, Desviación estándar: 0.87, Mínimo: 3.00, Máximo: 9.00.
    b. Observamos la distribución de la calidad de los vinos separadas por tipo y también un gráfico de frecuencia sobre la calidad para entender mejor la calidad del catálogo.





de

- Valores de Tendencia Central y Desviación Estándar:
  - Para el Vino Tinto:
  - Media: 5.64
  - Desviación Estándar: 0.81
  - Mínimo: 3.00
  - Máximo: 8.00
- Para el Vino Blanco:

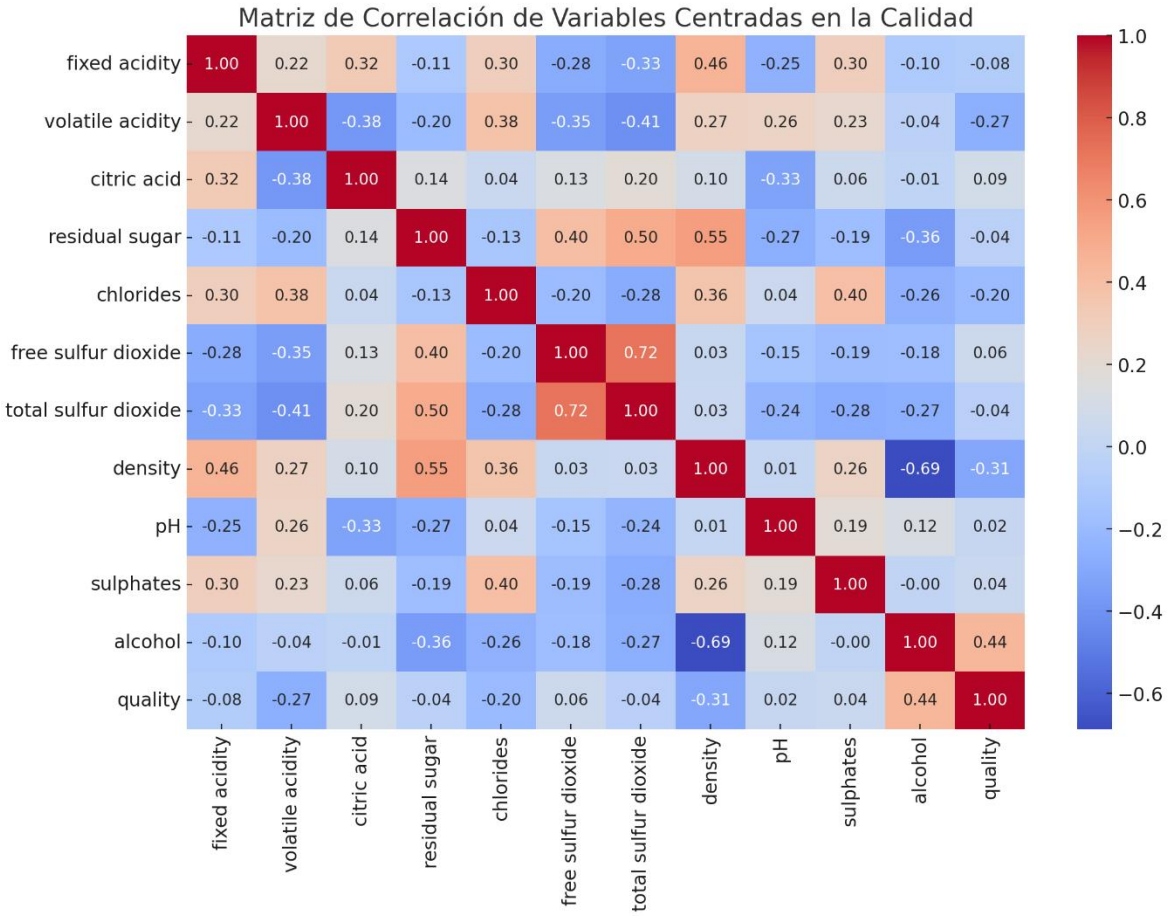
- Media: 5.88
- Desviación Estándar: 0.89
- Mínimo: 3.00
- Máximo: 9.00

Análisis de los Resultados:

- Los vinos blancos tienen, en promedio, una calidad ligeramente superior a los vinos tintos, como lo indica la media más alta.
- La desviación estándar es similar en ambos tipos de vino, lo que indica una variabilidad comparable en la calidad dentro de cada categoría.
- El rango de calidad es más amplio en los vinos blancos (3 a 9) en comparación con los vinos tintos (3 a 8), sugiriendo una mayor diversidad en la calidad de los vinos blancos.
- Ambos tipos de vino muestran una distribución de calidad centrada principalmente en torno a los valores medios (5 y 6), con pocos vinos alcanzando las calificaciones más bajas o altas.
- Casi el 75% de nuestros vinos se consideran de calidad promedio con tendencia a la baja.

Análisis de correlación:

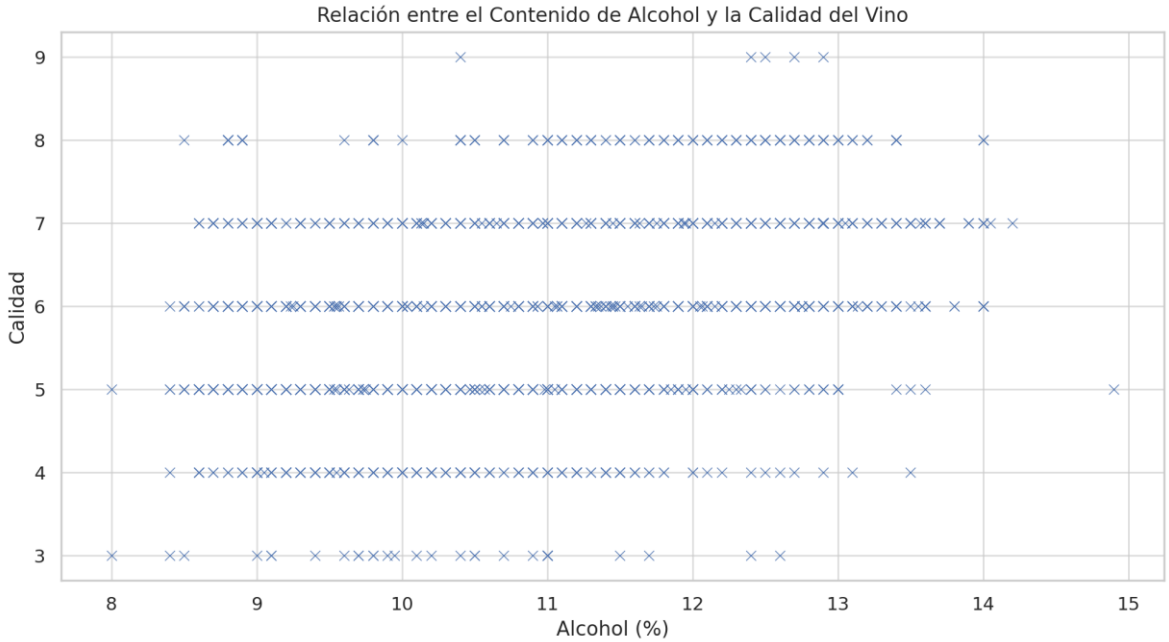
Entendiendo que la calidad de nuestros vinos es baja en su mayoría, vamos a centrarnos en la calidad del vino. Realizamos una matriz de correlación general para encontrar variables que afecten la calidad:



- Basándonos en esta matriz, las tres variables más influyentes en la calidad del vino son:
    a. Alcohol: Tiene la correlación más alta con la calidad (aproximadamente 0.444), lo que sugiere que a medida que aumenta el contenido de alcohol, generalmente mejora la calidad del vino.
    b. Ácido Cítrico: Tiene una correlación positiva moderada con la calidad (aproximadamente 0.086). Esto implica que una mayor concentración de ácido cítrico puede tener un impacto positivo en la calidad.
    c. Dióxido de Azufre Libre: Tiene una correlación positiva baja con la calidad (aproximadamente 0.055). Esto indica una influencia menos significativa en la calidad en comparación con el alcohol y el ácido cítrico.
    d.

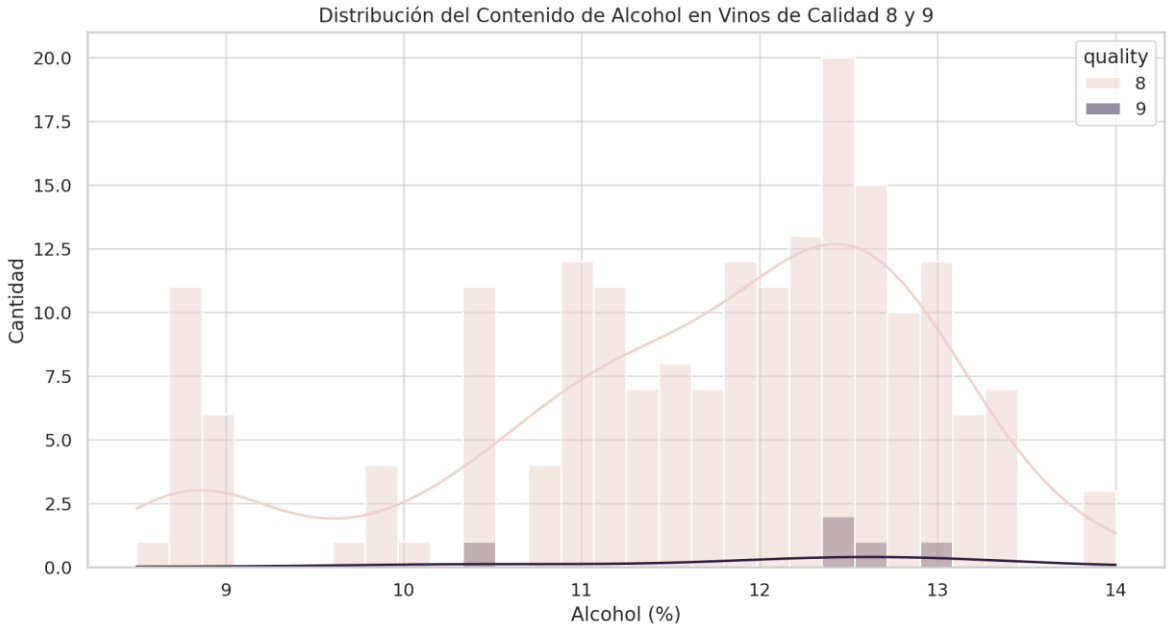
Evaluación del alcohol frente a la calidad de los vinos

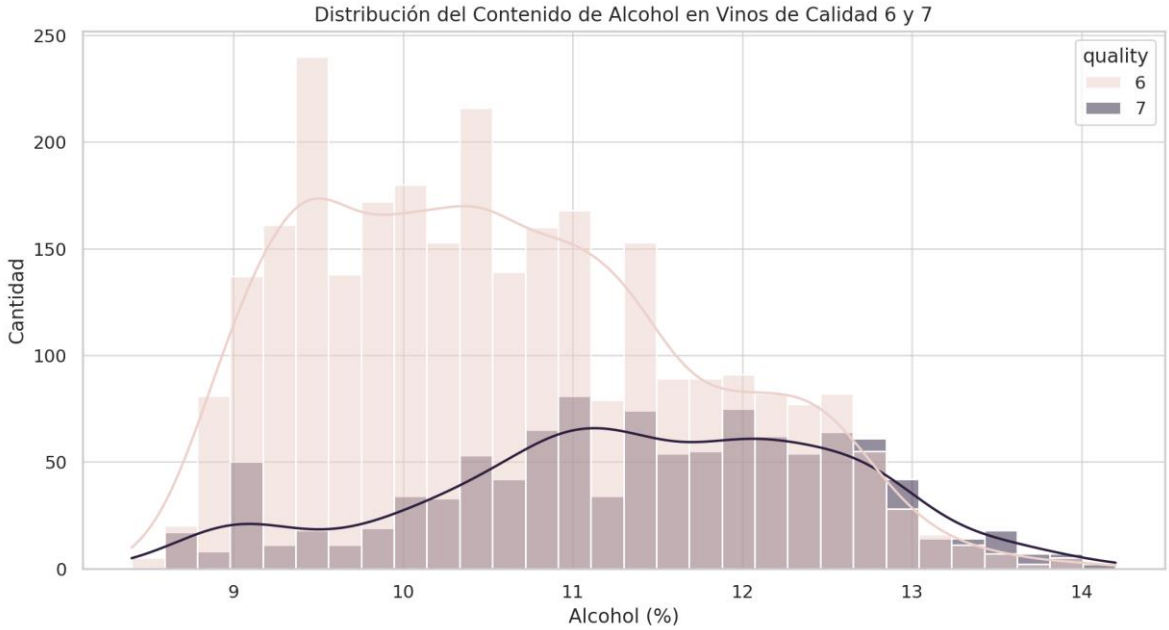
- Vamos a evaluar el nivel de alcohol si realmente tiene un comportamiento de causalidad.



a. En el gráfico de dispersión, observamos la relación entre el contenido de alcohol (%) y la calidad del vino. Parece haber una tendencia a que vinos con mayor contenido de alcohol tengan calificaciones de calidad más altas, aunque esta relación no es perfectamente clara.

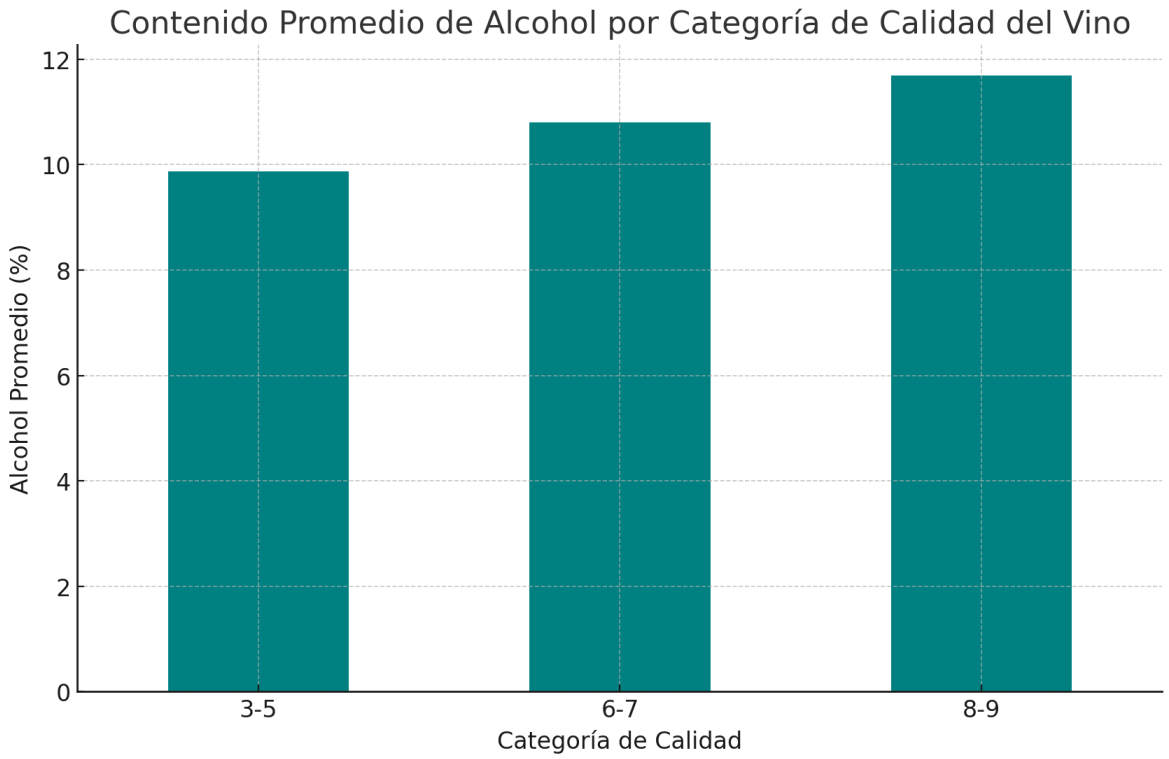
Análisis de Relación





- Aunque generalmente observamos una tendencia de que los vinos de mayor calidad (calificaciones 8 y 9) tienden a tener un mayor porcentaje de alcohol, esto no es una regla absoluta. En el dataset, es probable que haya algunos vinos de calidad 9 con un porcentaje de alcohol relativamente bajo.

Evaluemos la cantidad de alcohol por categorías de calidad.



a. Categoría 3-5: El contenido promedio de alcohol es aproximadamente 9.87%.
b. Categoría 6-7: El contenido promedio de alcohol es aproximadamente 10.81%.
c. Categoría 8-9: El contenido promedio de alcohol es aproximadamente 11.69%.

Usando el teorema de Bayes, veamos si la cantidad de alcohol es determinante en la calidad entendiendo lo construido hasta ahora:

Para calcular la probabilidad de que un vino de categoría 8 o 9 tenga un porcentaje de alcohol superior al 11% utilizando el teorema de Bayes, necesitamos definir y calcular ciertas probabilidades:

- P(Alcohol>11%|Calidad≥8)P(Alcohol>11%|Calidad≥8): La probabilidad de que un vino tenga un contenido de alcohol superior al 11% dado que su calidad es 8 o 9.
- P(Calidad≥8)P(Calidad≥8): La probabilidad de que un vino tenga una calidad de 8 o 9.
- P(Alcohol>11%)P(Alcohol>11%): La probabilidad de que un vino tenga un contenido de alcohol superior al 11%.

P(Calidad≥8|Alcohol>11%) P(Calidad≥8|Alcohol>11%), la probabilidad de que un vino sea de calidad 8 o 9 dado que su contenido de alcohol es superior al 11%.

Primero calcularemos las probabilidades:

- P(Calidad≥8)P(Calidad≥8): La probabilidad de que un vino tenga una calidad de 8 o 9 es aproximadamente 3.05%.

- P(Alcohol>11%)P(Alcohol>11%): La probabilidad de que un vino tenga un contenido de alcohol superior al 11% es aproximadamente 30.31%.
- P(Alcohol>11%|Calidad≥8)P(Alcohol>11%|Calidad≥8): La probabilidad de que un vino con calidad 8 o 9 tenga un contenido de alcohol superior al 11% es aproximadamente 73.74%.

Aplicando el Teorema de Bayes, calculamos:

P(Calidad≥8|Alcohol>11%): La probabilidad de que un vino sea de calidad 8 o 9 dado que su contenido de alcohol es superior al 11% es aproximadamente 7.41%.

Esto significa que dentro del conjunto de vinos con un contenido de alcohol superior al 11%, aproximadamente el 7.41% son vinos de calidad 8 o 9. Aunque una gran proporción de vinos de alta calidad tienen un alto contenido de alcohol, solo una pequeña fracción de todos los vinos con alto contenido de alcohol son de alta calidad.

CONCLUSION:

- Este fenómeno resalta la complejidad de la relación entre las características del vino y su calidad percibida. La calidad del vino no depende únicamente del contenido de alcohol, sino que está influenciada por una variedad de factores, incluyendo la acidez, los azúcares, los compuestos fenólicos, las prácticas de vinificación, y otras características sensoriales.