PEC 5- MANUEL ROJAS GARCÍA

UOC

NOMBRE: MANUEL ROJAS GARCÍA

Introducción

En esta PEC utilizaremos el conjunto de datos 'winequality-red.csv' que contiene información técnica y gustativa de distintos tipos de vino tinto.

Se pueden consultar en https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Las variables que contiene son las siguientes:

- acidez fija
- acidez volátil
- ácido cítrico
- azúcar residual
- cloruros
- dióxido de azufre libre
- dióxido de azufre total
- densidad
- pH
- sulfatos
- alcohol
- calidad

Os puede ser útil consultar el siguiente material:

- Módulos teóricos de Regresión lineal simple, múltiple y ANOVA.
- Actividades resueltas del Reto 5 (regresión lineal simple, múltiple y ANOVA).

Hay que entregar la práctica en fichero pdf o html (exportando el resultado final a pdf o html por ejemplo). Se recomienda generar el informe con Rmarkdown que genera automáticamente el pdf/html a entregar.

NOTA 1: no es necesario ni limpiar ni preprocesar los datos para este ejercicio

NOTA 2: comprobar que el dataset ha cargado correctamente (vigilar con la separación que se utiliza en el csv)

```
dataset <- read.csv("C:/Users/Manuel/Desktop/UOC/SEMESTRE 3 (Sep 2023 - Feb 2024)/Estadistica/PEC 5/ENTREGA/data_pac5.csv", sep = ";", header = TRUE)
```

Pregunta 1. (resolver con R). (3 puntos)

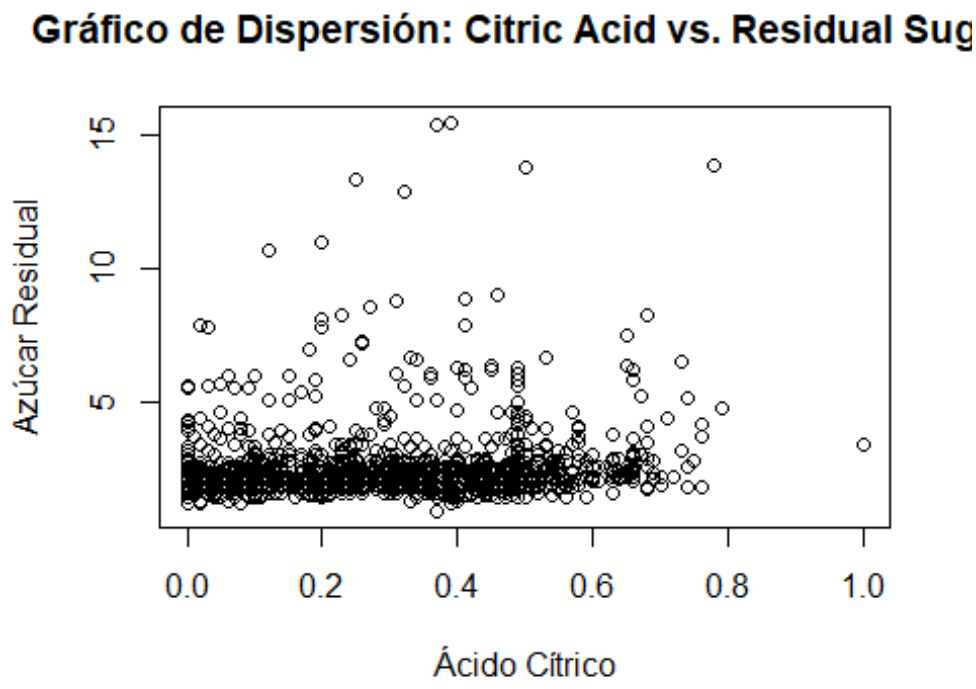
La empresa especializada en la creación de vinos de alta calidad está buscando comprender mejor las variables que influyen en la calidad del vino para optimizar sus estrategias de producción y marketing. Se realizará un análisis para identificar las características clave que contribuyen a la calidad del vino y determinar el enfoque para futuras campañas publicitarias.

a) Realiza un gráfico de dispersión entre la variable *citric.acid* y la variable *residual.sugar*. ¿Cuál es el coeficiente de correlación? Interpretad el resultado (1 punto).

https://rpubs.com/osoramirez/316691

Solución:

```
plot(dataset$citric.acid, dataset$residual.sugar,
     main = "Gráfico de Dispersión: Citric Acid vs. Residual Sugar",
     xlab = "Ácido Cítrico",
     ylab = "Azúcar Residual")
```




```
modelo <- lm(citric.acid ~ residual.sugar, data = dataset)

summary(modelo)

##
## Call:
```

```
## lm(formula = citric.acid ~ residual.sugar, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35733 -0.17624 -0.01235  0.15897  0.71194
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.220613   0.009935  22.205  < 2e-16 ***
## residual.sugar 0.019837   0.003422   5.798 8.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1928 on 1597 degrees of freedom
## Multiple R-squared:  0.02061,    Adjusted R-squared:   0.02
## F-statistic: 33.61 on 1 and 1597 DF,  p-value: 8.084e-09
```

La recta de regresión es: $\hat{y} = 0.220613 + 0.019837x$

La pendiente, de valor 0.019837 nos indica que por cada aumento de 1, el aumento del azucar residual es de 0.019837.

El coeficiente de determinación R2 es de 0.02061 y el de coeficiente de correlación lineal es: $\sqrt{0.2061}$

```
correlacion <- sqrt(0.02061)

cat ("El coeficiente de correlacion es de:", correlacion)

## El coeficiente de correlacion es de: 0.1435618
```

Hacemos los cálculos de forma automática.

```
correlacionauto <- cor(dataset$citric.acid, dataset$residual.sugar)
cat("Coeficiente de Correlación:", correlacionauto)

## Coeficiente de Correlación: 0.1435772
```

El coeficiente de determinación (R^2) es de 0.02061 lo que indica que el modelo de regresión lineal solo explica el 2.06% de la variabilidad en la variable de respuesta (citric.acid) en función de la variable predictora (residual.sugar). Un coeficiente de determinacion bajo indica que el modelo no explica una gran proporción de la variabilidad en la variable dependiente. Además, la correlación también es baja, con un valor de 0.1435 lo que determina que hay una relación débil entre las dos variables analizadas: citric.acid y residual.sugar, es decir, los cambios en residual.sugar no están asociados con cambios en citric.acid.

b) Encontrad los siguientes dos parámetros del modelo de regresión lineal a estudiar: el intercepto (B_0) y la pendiente (B_1) (1 punto).

Solución:

La recta de regresión es: $\hat{y} = 0.220613 + 0.019837x$

Como ya lo habiamos obtenido en el ejecicio anterior. Como conocemos que la recta de la regresión lineal es:

$\hat{y} = \beta 0 + \beta 1\ x$

$\beta 0 = 0.220613\ \beta 1 = 0.19837$

Podemos encontrar los resultados en la misma recta de regresion.

Realizamos el cálculo automático de los coeficientes

```
modelo <- lm(citric.acid ~ residual.sugar, data = dataset)

coefficients(modelo)
```

```
##    (Intercept) residual.sugar
##     0.22061287     0.01983718
```

c) ¿Qué porcentaje de la variación en la calidad del vino no puede ser explicado por los azucares residuales? (1 punto)

```
azucares <- lm(citric.acid ~ residual.sugar, data = dataset)
summary(azucares)
```

```
##
## Call:
## lm(formula = citric.acid ~ residual.sugar, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35733 -0.17624 -0.01235  0.15897  0.71194
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.220613   0.009935  22.205  < 2e-16 ***
## residual.sugar 0.019837   0.003422   5.798 8.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1928 on 1597 degrees of freedom
## Multiple R-squared:  0.02061,    Adjusted R-squared:   0.02
## F-statistic: 33.61 on 1 and 1597 DF,  p-value: 8.084e-09
```

Como hicimos en el ejercicio anterior, obtenemos el valor del modelo:

Multiple R-squared: 0.02061, Adjusted R-squared: 0.02

Para obtener el porcentaje de variación no explicada, restamos 1 al coeficiente de determinación y multiplicamos por 100 para obtener el porcentaje.

```
Porcentaje <- (1 - 0.02061)*100
cat ("Porcentaje de variación no explicada:", Porcentaje, "%")

## Porcentaje de variación no explicada: 97.939 %
```

El 97.939 de la variabilidad en solo enla calidad del vino no está siendo explicada por la variable residual.sugar.Deberiamos incluir otros valores u actores externos como podría ser el clima, sequia, etc.

Pregunta 2. (resolver con R). (3 puntos)

Para la creación de la próxima versión mejorada de vinos tintos, se han seleccionado distintos vinos y se han sometido a diversas catas.

Se procederá inicialmente a analizar los datos obtenidos de la evaluación de la influencia de la cantidad de sulfatos en la calidad del vino. Se buscará determinar si existen diferencias significativas entre las cantidades de la variable *sulphates* para distintos grupos definidos por la calidad (variable Quality_group **no disponible en el dataset**).

Si miramos la salida del modelo creado, contestad las preguntas siguientes:

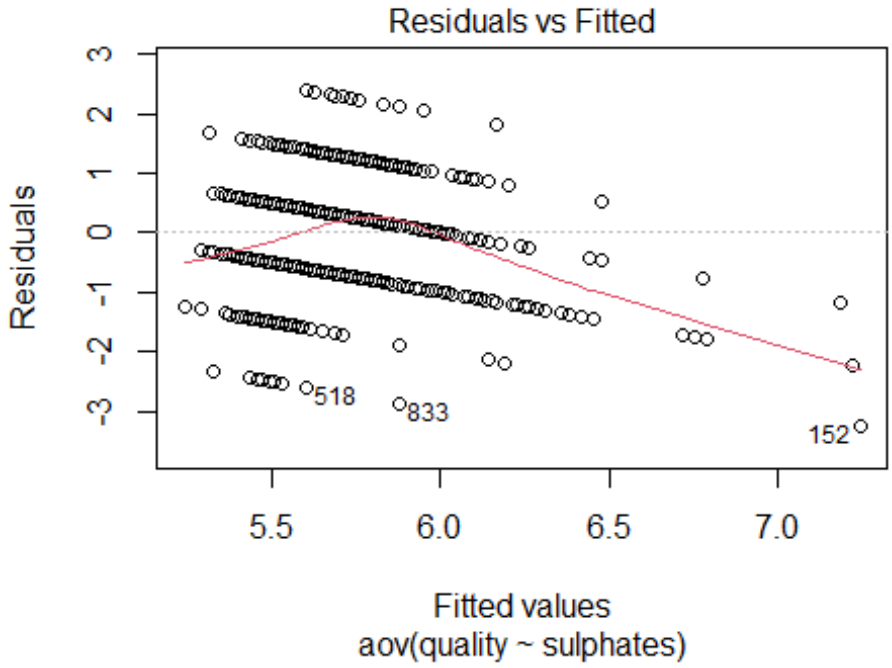
a) ¿Cuántos grupos y cuántas observaciones hay en el dataset? (1 punto)

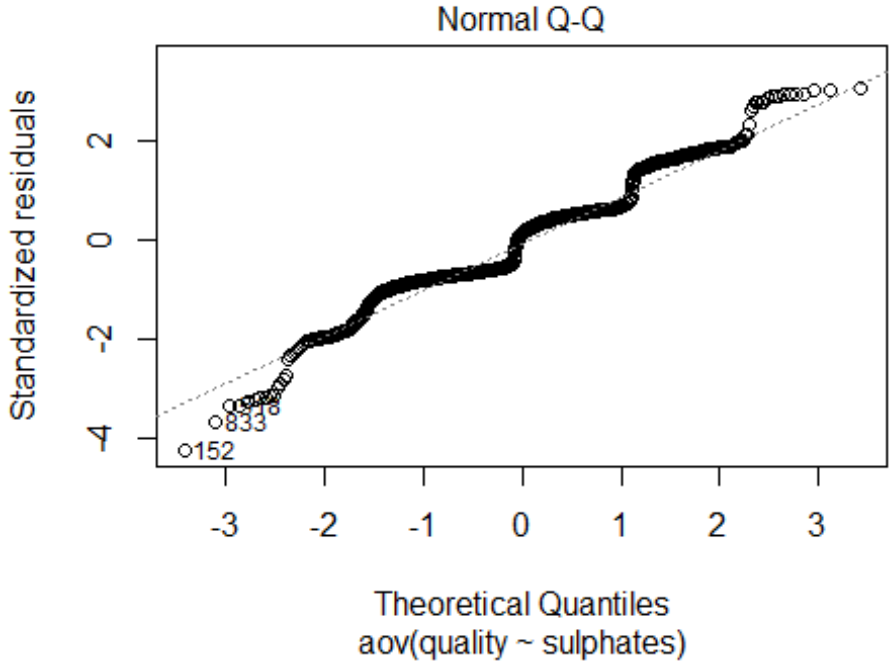
Solución:

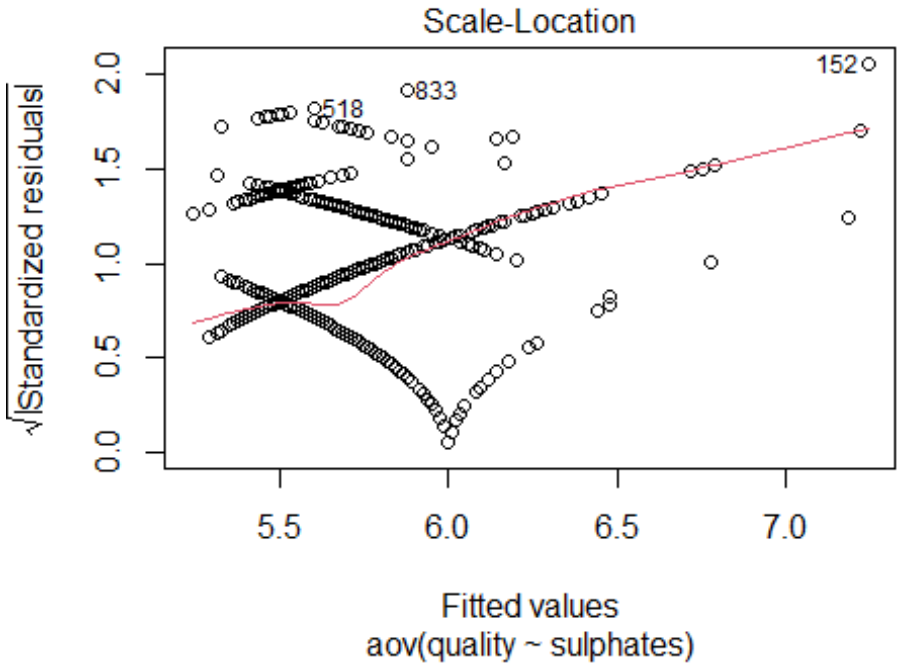
Hago un modelo parecido para mi propio dataset para analizar y comprender los datos con la siguiente web

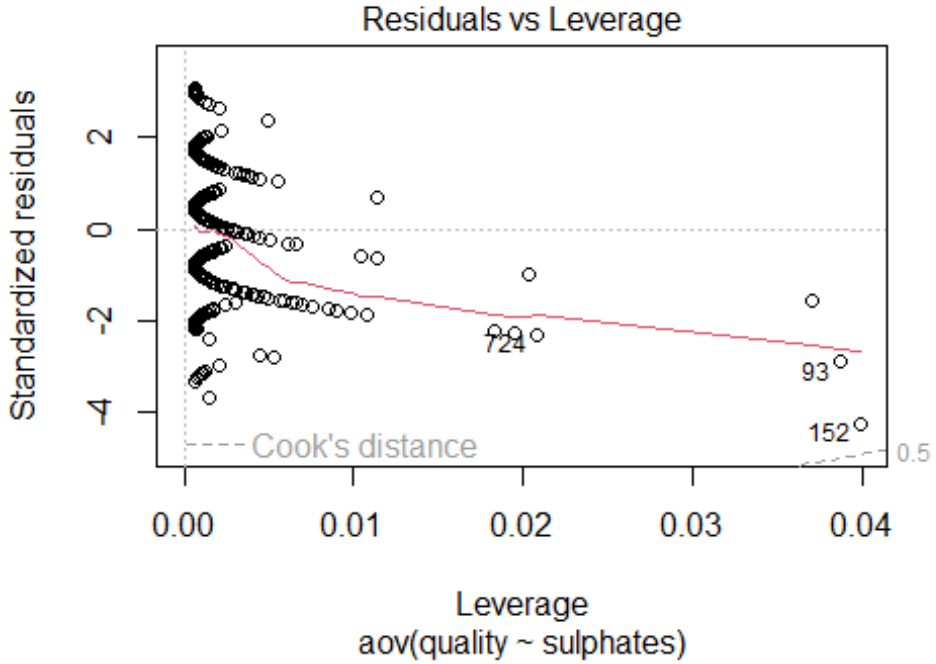
https://rpubs.com/Joaquin_AR/219148

```
modelo_anova <- aov(quality ~ sulphates , data = dataset)
summary(modelo_anova)

##               Df Sum Sq Mean Sq F value Pr(>F)
## sulphates      1   65.9   65.87   107.7 <2e-16 ***
## Residuals   1597  976.3    0.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(modelo_anova)
```


Residuals vs Fitted
Residuals
3
2
1
0
-1
-2
-3
518
833
152
5.5
6.0
6.5
7.0
Fitted values
aov(quality ~ sulphates)


Normal Q-Q
Standardized residuals
2
0
-2
-4
833
152
-3
-2
-1
0
1
2
3
Theoretical Quantiles
aov(quality ~ sulphates)





Una vez comprendidos los datos anova, usamos los que indica el ejercicio

Df Sum Sq Mean Sq F value Pr(>F)

Quality_group 3 2.98 0.9942 36.94 <2e-16 *

Residuals 1595 42.93 0.0269

—

Signif. codes: 0 ’’ *0.001* ’’ *0.01* ’’ 0.05 ’.’ 0.1 ’ ’ 1

Podemos determinar que tiene 3 grados de libertad (DF) para los grupos “Quality_group” y 1595 grados de libertad (DF) para “Residuals” Hay 4 grupos ya que en Anova se resta 1 a los grados de libertad (k -1) y el número total de observaciones en el dataset es la suma de los grados de libertad para los grupos y los residuos: 3 + 1595 =1598

b) Si se utiliza el nivel de significación $\alpha = 0.05$, ¿qué valor crítico se debe utilizar para realizar el análisis de la varianza? (1 punto)

Solución:

En primer lugar, determinamos que rechazamos la hipótesis nula ya que el pvalor (que es prácticamente 0) es menor que el valor de significación (Aunque en el siguiente apartado descubriremos la misma comparativa pero con el valor crítico)

En segundo lugar, para poder obtener el valor crítico necesitamos los grados de libertad ya obtenidos en el apartado anterior y usar qf para obtener la distribución F.

https://statologos.com/f-valor-critico-r/

```
alpha <- 0.05

valorcritico <- qf(1 - alpha, df1 = 3, df2 = 1595)

cat("Valor crítico de la distribución F:", valorcritico)

## Valor crítico de la distribución F: 2.610481
```

c) ¿Cuál es la conclusión del análisis de la varianza (con un nivel de significación del 5%) en función del valor crítico? (1 punto)

```
modeloanova <- aov(quality ~ sulphates, data = dataset)
summary(modeloanova)

##               Df Sum Sq Mean Sq F value Pr(>F)
## sulphates      1   65.9   65.87   107.7 <2e-16 ***
## Residuals   1597  976.3    0.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor F es de 107,7 si este valor es superior al valor crítico (2,61) debemos rechazar la hipótesis nula y afirmamos que hay diferencias significativas entre los grupos en términos de la variable “Quality_group”

Pregunta 3 (resolver con R). (4 puntos)

Exploraremos un modelo predictivo sobre la calidad del vino utilizando múltiples variables:

a) Escribe la ecuación que se obtiene del modelo de regresión múltiple para predecir la calidad del vino utilizando las variables de pH, contenido de azúcar residual y sulfatos. (1 punto)

Solución:

Quality = β0 + β1(pH) + β2(Residualsugar) + β3(sulphates) + error.

Sacamos el modelo como en regresion simple pero ahora multiple con las 3 variables que nos indica el ejercicio.

```
modelo2 <- lm(quality ~ pH + `residual.sugar` + sulphates, data =
dataset)

summary(modelo2)

##
## Call:
## lm(formula = quality ~ pH + residual.sugar + sulphates, data =
dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2617 -0.5460  0.1185  0.4541  2.3886
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.967242   0.456120  10.890   <2e-16 ***
## pH             -0.039750   0.129763  -0.306    0.759
## residual.sugar  0.006701   0.013932   0.481    0.631
## sulphates       1.190285   0.117755  10.108   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7823 on 1595 degrees of freedom
## Multiple R-squared:  0.06341,    Adjusted R-squared:  0.06165
## F-statistic: 35.99 on 3 and 1595 DF,  p-value: < 2.2e-16
```

Siendo la formula con los datos la siguiente:

Calidad del vino = 4.967242 – 0.039750(Ph) + 0.006701 (residual sugar) + 1.190285 (sulphates) + error.

Quality = β0 + β1(pH) + β2(Residualsugar) + β3(sulphates) + error.

b) ¿El modelo en su conjunto es significativo con un nivel del 5%? Además, ¿cuál es el coeficiente de determinación obtenido para este modelo? (1 punto)

F-statistic: 35.99 on 3 and 1595 DF, p-value: < 2.2e-16

Multiple R-squared: 0.06341, Adjusted R-squared: 0.06165

El modelo de determinación es 0.06341, y el de coeficiente de correlación lineal es: $\sqrt{0.06341} = 0.2518134$

Volvemos ha realizar el valor critico como un ejercicio anterior que ya obteniamos el valor critico

```
alpha <- 0.05

valorcritico <- qf(1 - alpha, df1 = 3, df2 = 1595)

cat("Valor crítico de la distribución F:", valorcritico)

## Valor crítico de la distribución F: 2.610481
```

Seguimos determinando que como el pvalor es tan pequeño que sigue significando que rechazamos la hipótesis nula

c) Dado un vino con un pH de 3.5, un contenido de azúcar residual de 2.5 y sulfatos de 0.6, ¿cuál sería su calidad según el modelo establecido?

Como ya tenemos la fórmula creada solo debemos sustituir los valores

Calidad del vino = 4.967242 – 0.039750(Ph) + 0.006701 (residual sugar) + 1.190285 (sulphates) + error.

Calidad del vino = 4.967242 – 0.039750(3.5) + 0.006701 (2.5) + 1.190285 (0.6)

```
Quality <- 4.967242 - 0.039750 * (3.5) + 0.006701 * (2.5) + 1.190285 *
(0.6)
Quality

## [1] 5.559041
```

d) Si tuvieras que eliminar alguna variable del modelo del apartado a), considerando un nivel de significación del 5%, ¿cuál eliminarías y por qué? (1 punto).

Solución:

```
matrizcorrelacion <- cor(dataset[c("quality", "pH", "residual.sugar",
"sulphates")])
```

```
print(matrizcorrelacion)
```

```
##                      quality          pH residual.sugar    sulphates
## quality          1.00000000 -0.05773139    0.013731637  0.251397079
## pH              -0.05773139  1.00000000   -0.085652422 -0.196647602
## residual.sugar   0.01373164 -0.08565242    1.000000000  0.005527121
## sulphates        0.25139708 -0.19664760    0.005527121  1.000000000
```

Quality vs. pH: La correlación entre “quality” y “pH” es -0.0577. Esta correlación es bastante baja y cercana a cero, lo que sugiere una relación débil o nula entre estas dos variables.

Quality vs. Residual Sugar: La correlación entre “quality” y “residual.sugar” es 0.0137. Similar a la correlación con pH, es cercana a cero, indicando una relación débil o nula.

Quality vs. Sulphates: La correlación entre “quality” y “sulphates” es 0.2514. Esta correlación es más fuerte en comparación con las anteriores, pero aún así no es extremadamente alta. Indica una relación positiva moderada entre la calidad del vino y la cantidad de sulfatos.

Basándonos en estos resultados, podríamos considerar eliminar la variable pH del modelo. Sin embargo, la decisión también dependerá del contexto del problema y de consideraciones teóricas sobre qué variables son importantes para tu análisis específico.