PEC 5

UOC

NOMBRE:

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)

```
## cargar datos
```

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).

Solución:

```
#Espacio para solución
```

b) Encontrad los siguientes dos parámetros del modelo de regresión lineal a estudiar: el intercepto (B_0) y la pendiente (B_1). Nota: Usa *citric.acid como variable dependiente* (1 punto).

Solución:

```
#Espacio para solución
```

c) ¿Qué porcentaje de la variación en la variable ácido cítrico no puede ser explicado por los azucares residuales? (1 punto)

Solución:

```
#Espacio para solución
```

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:

```
##                 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
```

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

Solución:

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:

```
#Espacio para solución
```

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

Solución:

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:

```
#Espacio para solución
```

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)

Solución:

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? (1 punto)

Solución:

```
#Espacio para solución
```

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: