EXAMEN DE POSTGRADO 2022 - (MAESTRIA EN CIENCIA DE DATOS)

APELLIDOS Y NOMBRES: …… CEL:…….

CORREO: …….

1.- (Regresión múltiple)(Se pide dar solución en R) Un estudio quiere generar un modelo que permita predecir la esperanza de vida media de los habitantes de una ciudad en función de diferentes variables. Se dispone de información sobre: habitantes, analfabetismo, ingresos, esperanza de vida, asesinatos, universitarios, heladas, área y densidad poblacional. Con la siguiente información

```
library(dplyr)
datos <- as.data.frame(state.x77)
datos <- rename(habitantes = Population, analfabetismo = Illiteracy,
          ingresos = Income, esp_vida = `Life Exp`, asesinatos = Murder,
           universitarios = `HS Grad`, heladas = Frost, area = Area,
          .data = datos)
datos <- mutate(.data = datos, densidad_pobl = habitantes * 1000/area)
datos
```

se pide: (todo con R)

a) Analizar la relación entre variables
b) Generar un modelo de regresión lineal con todas las variables (hacer un summary y analizar)
c) Selección de los mejores predictores (hacer un summary y analizar)
d) Realizar la validación de condiciones para la regresión múltiple lineal
e) Determinar si cumple con la distribución normal de los residuos
f) Realizar el test de normalidad
g) Realizar el test homocedasticidad
h) Realizar el Análisis de Inflación de Varianza (VIF)
i) Realizar ANOVA

2.- Considerando el estudio de rendimiento antes y después cuya información es

grupoAntes : 2, 4, 6, 1, 3
grupoDespues : 5, 2, 7, 1, 6

Se pide:

a) Realizar la prueba de rangos de signos de Wilcoxon en sus tres tipos de pruebas) (Se pide dar solución en R)
b) Verifique el resultado de R con el cálculo a mano. (en ambos incisos realizar la interpretación de todo el proceso)

3.- Un equipo de biólogos quiere generar un modelo estadístico que permita identificar a que especie (a o b) pertenece un determinado insecto. Para ello se han medido tres variables (longitud de las patas, diámetro del abdomen y diámetro del órgano sexual) en 10 individuos de cada una de las dos especies. *Los datos son los siguientes (aplique análisis discriminante con R). Con sus correspondientes interpretaciones.*

	especie	pata	abdomen	organo_sexual
1	a	191	131	53
2	a	185	134	50
3	a	200	137	52
4	a	173	127	50
5	a	171	128	49
6	a	160	118	47
7	a	188	134	54
8	a	186	129	51
9	a	174	131	52
10	a	163	115	47
11	b	186	107	49
12	b	211	122	49
13	b	201	144	47
14	b	242	131	54
15	b	184	108	43
16	b	211	118	51
17	b	217	122	49
18	b	223	127	51
19	b	208	125	50
20	b	199	124	46

4.- Considerando la siguiente información. La cuantificación del contenido en grasa de la carne pude hacerse mediante técnicas de analítica química, sin embargo, este proceso es costoso en tiempo y recursos. Una posible alternativa para reducir costes y optimizar tiempo es emplear un espectrofotómetro (instrumento capaz de detectar la absorbancia que tiene un material a diferentes tipos de luz en función de sus características). Para comprobar su efectividad se mide el espectro de absorbancia de 100 longitudes de onda en 215 muestras de carne, cuyo contenido en grasa se obtiene también por análisis químico para poder comparar los resultados. El set de datos meatspec del paquete faraway contiene toda la información.

```
library(faraway)

data(meatspec)
```

Se pide determinar: el modelo inicial, la suma de cuadrados del error, los predictores mediante stepwise, el número óptimo de componentes principales identificado por cross validation , el test-MSE y los gráficos correspondientes.(puede realizar otros aspectos no solicitados en esta pregunta que será tomado en cuenta en la evaluación).