Ejercicio 1:

1.1 Realizo clusterización jerárquica. El número de clusters adecuado es 2.

1.2 Selecciono el método average y la distancia euclídea por tener mayor coeficiente de Correlación cofenética. Probé varios métodos y elegí el de mayor coeficiente.

1.3 No he estandarizado las variables por tratarse de porcentajes en todos los casos

1.4 El cluster 1 está formado por animales que viven en la tierra. El cluster 2 está formado por animales que viven en el agua.

1.5 Calculo las medias de ambos clusters para visualizar los datos. Armé un data frame indicando en una columna el número de cluster.

Luego realizo test de hotelling:

H0: Todas las medias entre los subgrupos son iguales

H1: Hay al menos dos medias entre los subgrupos que difieren

El p valor del test de hotelling da muy chico, menor a 0.05. Por lo tanto rechazo la hipótesis nula de igualdad de medias. Hay diferencias entre las medias de ambos clusters.

Ejercicio 2:

2.1

Modelo de análisis:

HO: muA = muB = muC = muD = muE

Es decir,las medias del valor calórico de todas las marcas son iguales

H1: Existe al menos dos marcas cuyas medias no son iguales

Los supuestos del test son:

a) Independencia de las observaciones

b) Normalidad de los residuos

c) Homocedasticidad

2.2 Testear

Realizo primero Anova porque los supuestos se testean con los residuos del test.

Aplicando ANOVA el p-valor=1.03e-09 es <0.05 (pide confianza del 95%). Por lo tanto, si se verifican los supuestos de normalidad y homogeneidad de la varianza se rechaza que las medias de cada grupo sean todas iguales entre sí, es decir, hay al menos un par que difieren. Por lo tanto las distintas marcas influyen sobre la variable.

La independencia de las observaciones no se prueba, se consideran independientes por ser muestras tomadas al azar.

Verifico si se cumple el supuesto de normalidad de la distribución:

Aplico los tests de Shapiro, Agostino y Anderson y en todos los casos el pvalor es mayor que 0.05. Con lo cual no hay motivos para rechazar hipótesis nulas de los tests de normalidad.

Lo mismo en el gráfico, se puede ver que todas las observaciones están cercanas a la recta del medio.

Verifico si se cumple el supuesto de homocedasticidad:

Aplico los tests de Bartlett y Levene y en ambos casos el pvalor es mayor a 0.05.

Grafico los boxplot y se puede observar que el tamaño de las cajas es muy parecido para cada marca.

Calculo las varianzas por Marca.

En todos los casos, no hay motivo para rechazar hipótesis nula de homocedasticidad(igualdad de varianzas).

Dado que se cumplen los supuestos del test de ANOVA, rechazo HO. Es decir, hay por lo menos dos marcas de cervezas cuyas medias de calorias no son iguales.

2.3 Armo nuevo dataframe para testear por cervezas blancas y cervezas negras.

Hipótesis:

HO: Las medias de las cervezas blancas y de las cervezas negras son iguales

H1: Las medias de las cervezas blancas y de las cervezas negras son distintas

Si se cumplen los supuestos de independencia (muestras tomadas al azar), normalidad y homocedasticidad puedo aplicar ANOVA.

Calculo media y varianza de los subgrupos para ver los datos.

Verifico si se cumplen los supuestos de normalidad y homocedasticidad

Se cumple el supuesto de normalidad (Shapiro Test) pero no hay igualdad de varianzas (Ni con el test de Levene ni con Bartlett). Se ve la diferencia de varianzas en el temaño de las cajas en el gráfico de boxplot. Con lo cual realizo transformación de Box-Cox y valido los supuestos nuevamente.

El lambda que maximiza la función es 2.

Valido normalidad con Shapiro, Agostino y Ad test: No rechazo Hipótesis nula de los tests de normalidad.

Realizo test de Levene para validar homocedasticidad.

El pvalor del test de levene es prácticamente 0.05 (mejoró mucho con la transformación). Con lo cual no rechazo la hipótesis nula de este test y doy por validada la homocedasticidad.

Con lo cual, el anova con los datos transformados está dando un p valor menor a 0.05. Considerando que se cumplen los supuestos, rechazo la hipótesis nula de igualdad de las medias de calorías de cervezas por Tipo (Blancas y Negras)

2.4 Analizar diferencias entre los tipos de cervezas. Lo realizo con el dataframe original de las 5 marcas.

Hay diferencias entre todas las marcas, excepto entre la A y B (p valor alto y el intervalo contiene el 0) y entre las marcas EyD (p valor alto y el intervalo contiene el 0).

En todas las demás comparaciones el p valor es muy chico, menor al 0.05 y el intervalo no continee el 0.

2.5 Primero calculo los promedios de valores calóricos por marca.

La marca con mayor valor calórico es la B con 32.225 calorias promedio y la marca con menor valor calórico es la E con 24.925. El intervalo de confianza para B es 32.225 +/- z (0.95) y para E es 24.925 +/- z(0.95).

Ejercicio 3

1. El p valor de hotelling es muy chico, con lo cual hay diferencias entre las medias de cada grupo, tiene sentido realizar análisis discriminante
2. Analizo normalidad multivariada con mshapiro.test: p valor es muy chico (menor a 0.05), con lo cual rechazo la normalidad multivariada

Analizo igualdad de matrices de varianzas y covarianzas con BoxM: también es muy chico el p valor, menor a 0.05.

Con lo cual no se cumplen los supuestos para realizar análsis discriminante lineal.

1. Realizo análisis discriminante robusto porque no se cumplen los supuestos para análisis discriminante lineal o cuadrático. Para cuadrático se debería cumplir normalidad.
2. La tasa de buena clasificación ingenua (es ingenua porque evalúo el método con los mismos datos con los cuales lo genero) con el método robusto QdaCov da 85%.

Luego aplico otro método robusto (sde) y evalúo la clasificación anterior con este nuevo método y me da la tasa de buena clasificación 80%.

También se puede tomar el 70% de los datos para entrenamiento y testear con el 30% restante. Me está dando error este resultado. No pude calcular el % de buena clasificación.