

```

1 # Proyecto Final.
2 # Estudiantes:
3 # Miguel Colorado 14-10237
4 # Jose Barrera 15-10123
5 # Carlos Sivira 15-11377
6
7 #####
8 # Carga el archivo del proyecto
9 sales_pre = read.table("datosproy.txt", header = TRUE)
10
11 #####
12 # Limpieza la data
13 sales = sales_pre[-c(23,67,122),]
14
15 #####
16 # Pregunta 1
17 # Realice un analisis descriptivo y exploratorio de los datos. Incluya en
18 # este analisis la matriz de correlacion.
19
20 summary(sales_pre)
21 summary(sales)
22
23 # Ventas
24 ventas = sales$ventas
25 summary(ventas)
26 sd(ventas)
27 boxplot(ventas,main="Caja de Ventas", ylab= "Ganancias")
28
29 # Facebook
30 facebook = sales$facebook
31 summary(facebook)
32 sd(facebook)
33 boxplot(facebook,main="Caja de Facebook", ylab= "Presupuesto")
34
35 # Periodico
36 periodico = sales$periodico
37 summary(periodico)
38 sd(periodico)
39 boxplot(periodico,main="Caja de Periodico", ylab= "Presupuesto")
40
41 # Instagram
42 instagram = sales$instagram
43 summary(instagram)
44 sd(instagram)
45 boxplot(instagram,main="Caja de Instagram", ylab= "Presupuesto")
46
47 # Tv
48 tv = sales$tv
49 summary(tv)
50 sd(tv)
51 boxplot(tv,main="Caja de Tv", ylab= "Presupuesto")
52
53 # Ebay
54 ebay = sales$ebay
55 summary(ebay)
56 sd(ebay)
57 boxplot(ebay,main="Caja de Ebay", ylab= "Presupuesto")
58
59 # Region
60 Region = sales$Region

```

```

61 summary(Region)
62 sd(Region)
63 boxplot(Region,main="Caja de Region", ylab= "Distribucion")
64
65 # Desviaciones estandar
66 sd = c(sd(ventas), sd(facebook), sd(periodico), sd(instagram), sd(tv), sd(ebay),
67 sd(Region));
68
69 # Medias
70 mn = c(mean(ventas), mean(facebook), mean(periodico), mean(instagram), mean(tv),
71 mean(ebay), mean(Region));
72
73 # Coeficientes de variacion
74 cof = sd / mn
75 cof
76
77 # Generacion del factor para la grafica
78 fc = c(replicate(197,"ventas"),
79 replicate(197,"facebook"),
80 replicate(197,"periodico"),
81 replicate(197,"instagram"),
82 replicate(197,"tv"),
83 replicate(197,"ebay"),
84 replicate(197,"region"))
85 fct = factor(fc)
86 dt = c(ventas, facebook, periodico, instagram, tv, ebay, Region)
87
88 # Es generada la grafica de todas las variables
89 boxplot(dt~fct,main="Caja de datos",ylab="Valores", xlab = "Variables")
90
91 # Histogramas
92 hist(ventas,main="Histograma de Ventas",ylab="Frecuencia",xlab="Ventas")
93 hist(facebook,main="Histograma de Facebook",ylab="Frecuencia",xlab="Facebook")
94 hist(periodico,main="Histograma de Periodico",ylab="Frecuencia",xlab="Periodico")
95 hist(instagram,main="Histograma de Instagram",ylab="Frecuencia",xlab="Instagram")
96 hist(tv,main="Histograma de Tv",ylab="Frecuencia",xlab="Tv")
97 hist(ebay,main="Histograma de Ebay",ylab="Frecuencia",xlab="Ebay")
98 hist(Region,main="Histograma de Region",ylab="Frecuencia",xlab="Region")
99
100 # Matriz de Correlacion
101 variables_cuant = sales[1:7]
102 sales.cor = cor(variables_cuant)
103 sales.cor
104
105 #####
106 # Pregunta 2
107 # Calcular el intervalo de confianza del 95% para las medias de ventas por
108 # region. Discuta los resultados.
109
110 # Calculo del intervalo de confianza para la media de ventas en la region 1
111 sales_per_region = sales$ventas[sales$Region == 1]
112 t.test(sales_per_region, conf.level = 0.95)$conf.int
113
114 # Calculo del intervalo de confianza para la media de ventas en la region 2
115 sales_per_region = sales$ventas[sales$Region == 2]
116 t.test(sales_per_region, conf.level = 0.95)$conf.int
117
118 # Calculo del intervalo de confianza para la media de ventas en la region 3

```

```

119 sales_per_region = sales$ventas[sales$Region == 3]
120 t.test(sales_per_region, conf.level = 0.95)$conf.int
121
122 # Calculo del intervalo de confianza para la media de ventas en la region 4
123 sales_per_region = sales$ventas[sales$Region == 4]
124 t.test(sales_per_region, conf.level = 0.95)$conf.int
125
126 #####
127 # Pregunta 3
128 # Encuentre el modelo de regresion simple que mejor se ajuste a
129 # los datos; realice las pruebas estadisticas que considere conveniente
130 # para justificar su respuesta, incluyendo un analisis de residuales.
131
132 # Facebook
133 ml1 = lm(sales$ventas ~ sales$facebook)
134 plot(sales$facebook, sales$ventas, main = "Ventas en funcion de la publicidad en
Facebook", xlab = "Publicidad en Facebook", ylab = "Ventas")
135 abline(ml1)
136 summary(ml1)
137 plot(ml1, main = "Modelo 2.1")
138
139 # Periodico
140 ml2 = lm(sales$ventas ~ sales$periodico)
141 plot(sales$periodico, sales$ventas, main = "Ventas en funcion de la publicidad en
Periodico", xlab = "Publicidad en Periodico", ylab = "Ventas")
142 abline(ml2)
143 summary(ml2)
144 plot(ml2, main = "Modelo 2.2")
145
146 # Instagram
147 ml3 = lm(sales$ventas ~ sales$instagram)
148 plot(sales$instagram, sales$ventas, main = "Ventas en funcion de la publicidad en
Instagram", xlab = "Publicidad en Instagram", ylab = "Ventas")
149 abline(ml3)
150 summary(ml3)
151 plot(ml3, main = "Modelo 2.3")
152
153 # Tv
154 ml4 = lm(sales$ventas ~ sales$tv)
155 plot(sales$tv, sales$ventas, main = "Ventas en funcion de la publicidad en TV", xlab
= "Publicidad en Television", ylab = "Ventas")
156 abline(ml4)
157 summary(ml4)
158 plot(ml4, main = "Modelo 2.4")
159
160 # Ebay
161 ml5 = lm(sales$ventas ~ sales$ebay)
162 plot(sales$ebay, sales$ventas, main = "Ventas en funcion de la publicidad en Ebay",
xlab = "Publicidad en Ebay", ylab = "Ventas")
163 abline(ml5)
164 summary(ml5)
165 plot(ml5, main = "Modelo 2.5")
166
167 # Region
168 ml6 = lm(sales$ventas ~ sales$Region)
169 plot(sales$Region, sales$ventas, main = "Ventas en funcion de la Region", xlab =
"Publicidad en FB", ylab = "Ventas")
170 abline(ml6)
171 summary(ml6)
172 plot(ml6, main = "Modelo 2.6")

```

```

173
174 #####
175 # Pregunta 4
176 # Consiga el modelo multiple mas apropiado. Realice, como en el inciso 3,
177 # todas las pruebas estadisticas que considere conveniente para justificar
178 # su respuesta, incluyendo un analisis de residuos. Considere un nivel del 5%
179
180 # Modelo multiple Y~facebook + periodico + instagram + tv + ebay + Region
181 mlm1 = lm(sales$ventas ~ sales$facebook + sales$periodico + sales$instagram +
sales$tv + sales$ebay + sales$Region)
182 summary(mlm1)
183 plot(mlm1, main = "Modelo multiple Y~facebook + periodico + instagram + tv + ebay +
Region")
184
185 # Modelo multiple Y~facebook + instagram + tv + ebay + Region
186 mlm2 = lm(sales$ventas ~ sales$facebook + sales$instagram + sales$tv + sales$ebay +
sales$Region)
187 summary(mlm2)
188 plot(mlm2, main = "Modelo multiple Y~facebook + instagram + tv + ebay + Region")
189
190 # Modelo multiple Y~facebook + instagram + tv + Region
191 mlm3 = lm(sales$ventas ~ sales$facebook + sales$instagram + sales$tv + sales$Region)
192 summary(mlm3)
193 plot(mlm3, main = "Modelo multiple Y~facebook + instagram + tv + Region")
194
195 # Modelo multiple Y~facebook + instagram + Region
196 mlm4 = lm(sales$ventas ~ sales$facebook + sales$instagram + sales$Region)
197 summary(mlm4)
198 plot(mlm4, main = "Modelo multiple Y~facebook + instagram + Region")
199
200 # Mejor modelo conseguido. Presenta problemas en la normalidad de los residuales con
datos atipicos
201 # Modelo multiple Y~facebook + instagram
202 mlm5 = lm(sales$ventas ~ sales$facebook + sales$instagram)
203 summary(mlm5)
204 predict(mlm5,sales,interval = "prediction")
205 plot(mlm5, main = "Modelo multiple Y~facebook + instagram")
206 boxplot(rstandard(mlm5), main = "Caja de residuales", ylab = "Residuales")
207 hist(rstandard(mlm5), main="Histograma de residuales",ylab="Frecuencia",xlab="Rango")
208
209 # Se eliminan los valores de la tabla que afectan la normalidad de los residuales
210 sales_mod = sales[-c(128, 124, 56, 6, 77, 3, 74, 25, 186, 176, 171, 167, 164, 156,
129, 130, 101, 35, 80, 163, 133),]
211
212 # Esta modificacion del modelo anterior mejora el modelo a costa de eliminar datos de
la tabla
213 # Se aprecia que los residuales se encuentran bien distribuidos de forma normal
214 # Modelo multiple Y~facebook + instagram modificado
215 mlm6 = lm(sales_mod$ventas ~ sales_mod$facebook + sales_mod$instagram)
216 summary(mlm6)
217 predict(mlm6,sales_mod,interval = "prediction")
218 plot(mlm6, main = "Modelo multiple Y~facebook + instagram modificado")
219 boxplot(rstandard(mlm6), main = "Caja de residuales", ylab = "Residuales")
220 hist(rstandard(mlm6), main="Histograma de residuales",ylab="Frecuencia",xlab="Rango")
221
222 #####
223 # Pregunta 5
224 # Estudios previos indican que las ventas en la region 1 muestran un
225 # precio de 150 (millones), aunque estudios suponen que dicha cantidad
226 # es superior a la mostrada por este analisis. Con un nivel de confianza

```

```

227 # que usted considere necesario, realice un código en el software estadístico
228 # R que muestre el resultado de dicho análisis. Analice los resultados y concluya.
229
230 # Se obtiene las ventas de la primera región
231 sales_per_region = sales$ventas[sales$Region == 1]
232
233 # Se define la hipótesis nula como  $\mu_0$  igual a 150
234 # La hipótesis alternativa será que  $\mu_0$  es mayor a 150
235  $\mu_0 = 150$ 
236
237 # Se obtiene el tamaño de la muestra (grande  $49 > 30$ )
238 n = length(sales_per_region)
239
240 # Se obtiene la media de la muestra
241 sample_mean = mean(sales_per_region)
242
243 # Se obtiene la media de la muestra
244 standard_deviation = sd(sales_per_region)
245
246 # Se calcula el estadístico Z por tratarse de una muestra grande
247 z = ((sample_mean -  $\mu_0$ ) / (standard_deviation / sqrt(n)))
248 z
249
250 # Se obtiene el p-valor asociado a Z
251 p_value = pnorm(z, lower.tail=FALSE)
252 p_value
253
254 # Se realiza el estudio de las hipótesis propuestas
255 t.test(sales_per_region, alternative = "greater", mu = 150, conf.level = 0.95)
256
257 # Sección de observación en el informe
258 # Estudio de la prueba de hipótesis para cola inferior
259 z = ((sample_mean -  $\mu_0$ ) / (standard_deviation / sqrt(n)))
260 z
261
262 # Se obtiene el p-valor asociado a Z
263 p_value = pnorm(z, lower.tail=TRUE)
264 p_value
265
266 # Se realiza el estudio de las hipótesis propuestas
267 t.test(sales_per_region, alternative = "less", mu = 15, conf.level = 0.95)
268
269 #####
270 # Pregunta 6
271 # Para el modelo de regresión lineal simple obtenido en el
272 # inciso 3, realice la predicción correspondiente para 5 ventas
273 # que se anexan a la muestra, los datos se presentan en el Cuadro 1.
274 # Grafique los intervalos de predicción y de confianza
275 # respectivamente. Realice el análisis respectivo.
276
277 # El mejor modelo fue modeloFacebook o ml1, entonces se utilizará este
278 # para predecir las ventas.
279 modeloFacebook = lm(ventas~facebook)
280
281 nuevosPresupuestos = data.frame(facebook=c(300, 320, 338, 350, 400))
282
283 # Se grafica el modelo y las bandas de confianza/predicción
284 plot(facebook, ventas, main="Intervalos para el modelo ventas~facebook", xlab =
"Presupuesto asignado a Facebook")
285 abline(modeloFacebook)

```

```

286
287 # Predicción para el nuevo presupuesto asignado a publicidad
288 (predict(modeloFacebook, nuevosPresupuestos, interval = 'predict'))
289
290 # Se generan puntos para las bandas
291 sequence = data.frame(facebook = seq(0, 300, 1))
292
293 # Intervalo de predicción
294 predicFacebook = predict(modeloFacebook, sequence, interval = "prediction")
295 lines(sequence$facebook, predicFacebook[,2], lty = 2, col = "red")
296 lines(sequence$facebook, predicFacebook[,3], lty = 2, col = "red")
297
298 # Intervalo de confianza para el 95%
299 confFacebook = predict(modeloFacebook, sequence, interval = "confidence")
300 lines(sequence$facebook, confFacebook[,2], lty = 2, col = "blue")
301 lines(sequence$facebook, confFacebook[,3], lty = 2, col = "blue")
302
303 # Se agrega una leyenda
304 legend("topleft", legend=c("Interv. Predicción", "Interv. Confianza para 95%"),
305 col=c("red", "blue"), lty=2:2, cex=0.8)
306
307 #####
308 # Pregunta 7
309 # Existe suficiente evidencia que permita concluir que las ventas media
310 # de las variables de estudio difieren con respecto a las regiones. Use
311 # el procedimiento de análisis de varianza para un diseño de un factor.
312 # Que concluiría usted con un nivel de significancia  $\alpha = 0.03$ 
313
314 #  $H_0$ : Las medias de ventas entre las regiones son iguales
315 dat=sales$ventas
316 fact=factor(sales$Region)
317 tapply(dat,fact,mean)
318 boxplot(dat~fact, main = "Caja de regiones", xlab = "Regiones", ylab= "Ventas")
319
320 # Se observa que los valores de Q2 son cercanos entre sí,
321 # de manera que las medias de cada región se encuentran dentro del
322 # rango intercuartil resto de las regiones, los cuales son similares entre sí.
323 mod.lm=lm(dat~fact)
324 anova(mod.lm)
325
326 # Como  $p\text{-valor} > 0.03$  no podemos rechazar la hipótesis nula. Entonces no hay
327 # suficiente evidencia para concluir que las ventas medias de las
328 # variables de estudio difieren con respecto a las regiones.

```