```r
# Proyecto Final.
# Estudiantes:
# Miguel Colorado 14-10237
# Jose Barrera    15-10123
# Carlos Sivira   15-11377

########################################################################
# Carga el archivo del proyecto
sales_pre = read.table("datosproy.txt", header = TRUE)

########################################################################
# Limpieza de la data
sales = sales_pre[-c(23,67,122),]

########################################################################
# Pregunta 1
# Realice un analisis descriptivo y exploratorio de los datos. Incluya en
# este analisis la matriz de correlacion.

summary(sales_pre)
summary(sales)

# Ventas
ventas = sales$ventas
summary(ventas)
sd(ventas)
boxplot(ventas,main="Caja de Ventas", ylab= "Numero de productos vendidos")

# Facebook
facebook = sales$facebook
summary(facebook)
sd(facebook)
boxplot(facebook,main="Caja de Facebook", ylab= "Presupuesto")

# Periodico
periodico = sales$periodico
summary(periodico)
sd(periodico)
boxplot(periodico,main="Caja de Periodico", ylab= "Presupuesto")

# Instagram
instagram = sales$instagram
summary(instagram)
sd(instagram)
boxplot(instagram,main="Caja de Instagram", ylab= "Presupuesto")

# Tv
tv = sales$tv
summary(tv)
sd(tv)
boxplot(tv,main="Caja de Tv", ylab= "Presupuesto")

# Ebay
ebay = sales$ebay
summary(ebay)
sd(ebay)
boxplot(ebay,main="Caja de Ebay", ylab= "Presupuesto")

# Region
Region = sales$Region
```

```r
summary(Region)
sd(Region)
boxplot(Region,main="Caja de Region", ylab= "Distribucion")

# Desviaciones estandar
sd = c(sd(ventas), sd(facebook), sd(periodico), sd(instagram), sd(tv), sd(ebay),
sd(Region));
sd

# Medias
mn = c(mean(ventas), mean(facebook), mean(periodico), mean(instagram), mean(tv),
mean(ebay), mean(Region));
mn

# Coeficientes de variacion
cof = sd / mn
cof

# Generacion del factor para la grafica
fc = c(replicate(length(sales$ventas),"ventas"),
        replicate(length(sales$ventas),"facebook"),
        replicate(length(sales$ventas),"periodico"),
        replicate(length(sales$ventas),"instagram"),
        replicate(length(sales$ventas),"tv"),
        replicate(length(sales$ventas),"ebay"),
        replicate(length(sales$ventas),"region"))
fct = factor(fc)
dt = c(ventas, facebook, periodico, instagram, tv, ebay, Region)

# Es generada la grafica de todas las variables
boxplot(dt~fct,main="Caja de datos",ylab="Valores", xlab = "Variables")

# Histogramas
hist(ventas,main="Histograma de Ventas",ylab="Frecuencia",xlab="Ventas")
hist(facebook,main="Histograma de Facebook",ylab="Frecuencia",xlab="Facebook")
hist(periodico,main="Histograma de Periodico",ylab="Frecuencia",xlab="Periodico")
hist(instagram,main="Histograma de Instagram",ylab="Frecuencia",xlab="Instagram")
hist(tv,main="Histograma de Tv",ylab="Frecuencia",xlab="Tv")
hist(ebay,main="Histograma de Ebay",ylab="Frecuencia",xlab="Ebay")
hist(Region,main="Histograma de Region",ylab="Frecuencia",xlab="Region")

# Matriz de Correlacion
variables_cuant = sales[1:7]
sales.cor = cor(variables_cuant)
sales.cor

######################################################################
# Pregunta 2
# Calcular el intervalo de confianza del 95% para las medias de ventas por
# region. Discuta los resultados.

# Calculo del intervalo de confianza para la media de ventas en la region 1
sales_per_region = sales$ventas[sales$Region == 1]
t.test(sales_per_region, conf.level = 0.95)$conf.int

# Calculo del intervalo de confianza para la media de ventas en la region 2
sales_per_region = sales$ventas[sales$Region == 2]
t.test(sales_per_region, conf.level = 0.95)$conf.int

# Calculo del intervalo de confianza para la media de ventas en la region 3
```

```r
sales_per_region = sales$ventas[sales$Region == 3]
t.test(sales_per_region, conf.level = 0.95)$conf.int

# Calculo del intervalo de confianza para la media de ventas en la region 4
sales_per_region = sales$ventas[sales$Region == 4]
t.test(sales_per_region, conf.level = 0.95)$conf.int

#######################################################################
# Pregunta 3
# Encuentre el modelo de regresion simple que mejor se ajuste a
# los datos; realice las pruebas estadisticas que considere conveniente
# para justificar su respuesta, incluyendo un analisis de residuales.

# Facebook
ml1 = lm(sales$ventas ~ sales$facebook)
plot(sales$facebook, sales$ventas, main = "Ventas en funcion de la publicidad en Facebook", xlab = "Publicidad en Facebook", ylab = "Ventas")
abline(ml1)
summary(ml1)
plot(ml1, main = "Ventas ~ Facebook")

# Periodico
ml2 = lm(sales$ventas ~ sales$periodico)
plot(sales$periodico, sales$ventas, main = "Ventas en funcion de la publicidad en Periodico", xlab = "Publicidad en Periodico", ylab = "Ventas")
abline(ml2)
summary(ml2)
plot(ml2, main = "Ventas ~ Periodico")

# Instagram
ml3 = lm(sales$ventas ~ sales$instagram)
plot(sales$instagram, sales$ventas, main = "Ventas en funcion de la publicidad en Instagram", xlab = "Publicidad en Instagram", ylab = "Ventas")
abline(ml3)
summary(ml3)
plot(ml3, main = "Ventas ~ Instagram")

# Tv
ml4 = lm(sales$ventas ~ sales$tv)
plot(sales$tv, sales$ventas, main = "Ventas en funcion de la publicidad en TV", xlab = "Publicidad en Television", ylab = "Ventas")
abline(ml4)
summary(ml4)
plot(ml4, main = "Ventas ~ Tv")

# Ebay
ml5 = lm(sales$ventas ~ sales$ebay)
plot(sales$ebay, sales$ventas, main = "Ventas en funcion de la publicidad en Ebay", xlab = "Publicidad en Ebay", ylab = "Ventas")
abline(ml5)
summary(ml5)
plot(ml5, main = "Ventas ~ Ebay")

# Region
ml6 = lm(sales$ventas ~ sales$Region)
plot(sales$Region, sales$ventas, main = "Ventas en funcion de la Region", xlab = "Publicidad en FB", ylab = "Ventas")
abline(ml6)
summary(ml6)
plot(ml6, main = "Ventas ~ Region")
```

```r

#########################################################################
# Pregunta 4
# Consiga el modelo multiple mas apropiado. Realice, como en el inciso 3,
# todas las pruebas estadisticas que considere conveniente para justificar
# su respuesta, incluyendo un analisis de residuos. Considere un nivel del 5%

# Modelo multiple Y~facebook + periodico + instagram + tv + ebay + Region
mlm1 = lm(sales$ventas ~ sales$facebook + sales$periodico + sales$instagram +
sales$tv + sales$ebay + sales$Region)
summary(mlm1)
plot(mlm1, main = "Modelo multiple Y~facebook + periodico + instagram + tv + ebay +
Region")

# Modelo multiple Y~facebook + instagram + tv + ebay + Region
mlm2 = lm(sales$ventas ~ sales$facebook + sales$instagram + sales$tv + sales$ebay +
sales$Region)
summary(mlm2)
plot(mlm2, main = "Modelo multiple Y~facebook + instagram + tv + ebay + Region")

# Modelo multiple Y~facebook + instagram + tv + Region
mlm3 = lm(sales$ventas ~ sales$facebook + sales$instagram + sales$tv + sales$Region)
summary(mlm3)
plot(mlm3, main = "Modelo multiple Y~facebook + instagram + tv + Region")

# Modelo multiple Y~facebook + instagram + Region
mlm4 = lm(sales$ventas ~ sales$facebook + sales$instagram + sales$Region)
summary(mlm4)
plot(mlm4, main = "Modelo multiple Y~facebook + instagram + Region")

# Mejor modelo conseguido. Presenta problemas en la normalidad de los residuales con
datos atipicos
# Modelo multiple Y~facebook + instagram
mlm5 = lm(sales$ventas ~ sales$facebook + sales$instagram)
summary(mlm5)
predict(mlm5,sales,interval = "prediction")
plot(mlm5, main = "Modelo multiple Y~facebook + instagram")
boxplot(rstandard(mlm5), main = "Caja de residuales", ylab = "Residuales")
hist(rstandard(mlm5), main="Histograma de residuales",ylab="Frecuencia",xlab="Rango")

# Se eliminan los valores de la tabla que afectan la normalidad de los residuales
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),]

# Esta modificacion del modelo anterior mejora el modelo a costa de eliminar datos de
la tabla
# Se aprecia que los residuales se encuentran bien distribuidos de forma normal
# Modelo multiple Y~facebook + instagram modificado
mlm6 = lm(sales_mod$ventas ~ sales_mod$facebook + sales_mod$instagram)
summary(mlm6)
predict(mlm6,sales_mod,interval = "prediction")
plot(mlm6, main = "Modelo multiple Y~facebook + instagram modificado")
boxplot(rstandard(mlm6), main = "Caja de residuales", ylab = "Residuales")
hist(rstandard(mlm6), main="Histograma de residuales",ylab="Frecuencia",xlab="Rango")

#########################################################################
# Pregunta 5
# Estudios previos indican que las ventas en la region 1 muestran un
# precio de 150 (millones), aunque estudios suponen que dicha cantidad
# es superior a la mostrada por este analisis. Con un nivel de confianza
```

```r
# que usted considere necesario, realice un codigo en el software estadistico
# R que muestre el resultado de dicho analisis. Analice los resultados y concluya.

# Se obtiene las ventas de la primera region
sales_per_region = sales$ventas[sales$Region == 1]

# Se define la hipotesis nula como m0 igual a 150
# La hipotesis alternativa sera que m0 es mayor a 150
m0 = 150

# Se obtiene el tamano de la muestra (grande 49>30)
n = length(sales_per_region)

# Se obtiene la media de la muestra
sample_mean = mean(sales_per_region)

# Se obtiene la media de la muestra
standard_deviation = sd(sales_per_region)

# Se calcula el estadistico Z por tratarse de una muestra grande
z = ((sample_mean - m0) / (standard_deviation / sqrt(n)))
z

# Se obtiene el p-valor asociado a Z
p_value = pnorm(z, lower.tail=FALSE)
p_value

# Se realiza el estudio de las las hipotesis propuestas
t.test(sales_per_region, alternative = "greater", mu = 150, conf.level = 0.95)

# Seccion de observacion en el informe
# Estudio de la prueba de hipotesis para cola inferior
z = ((sample_mean - m0) / (standard_deviation / sqrt(n)))
z

# Se obtiene el p-valor asociado a Z
p_value = pnorm(z, lower.tail=TRUE)
p_value

# Se realiza el estudio de las las hipotesis propuestas
t.test(sales_per_region, alternative = "less", mu = 15, conf.level = 0.95)

#########################################################################
# Pregunta 6
# Para el modelo de regresion lineal simple obtenido en el
# inciso 3, realice la prediccion correspondiente para 5 ventas
# que se anexan a la muestra, los datos se presentan en el Cuadro 1.
# Grafique los intervalos de prediccion y de confianza
# respectivamente. Realice el anolisis respectivo.

# El mejor modelo fue modeloFacebook o ml1, entonces se utilizara este
# para predecir las ventas.
modeloFacebook = lm(ventas~facebook)

nuevosPresupuestos = data.frame(facebook=c(300, 320, 338, 350, 400))

# Se grafica el modelo y las bandas de confianza/prediccion
plot(facebook, ventas, main="Intervalos para el modelo ventas~facebook", xlab =
"Presupuesto asignado a Facebook")
abline(modeloFacebook)
```

```r

# Prediccion para el nuevo presupuesto asignado a publicidad
(predict(modeloFacebook, nuevosPresupuestos, interval = 'predict'))

# Se generan puntos para las bandas
sequence = data.frame(facebook = seq(min(facebook), max(facebook), 1))

# Intervalo de prediccion
predicFacebook = predict(modeloFacebook, sequence, interval = "prediction")
lines(sequence$facebook, predicFacebook[,2], lty = 2, col = "red")
lines(sequence$facebook, predicFacebook[,3], lty = 2, col = "red")

# Intervalo de confianza para el 95%
confFacebook = predict(modeloFacebook, sequence, interval = "confidence")
lines(sequence$facebook, confFacebook[,2], lty = 2, col = "blue")
lines(sequence$facebook, confFacebook[,3], lty = 2, col = "blue")

# Se agrega una leyenda
legend("topleft", legend=c("Interv. Prediccion", "Interv. Confianza para 95%"),
       col=c("red", "blue"), lty=2:2, cex=0.8)

######################################################################
# Pregunta 7
# Existe suficiente evidencia que permita concluir que las ventas media
# de las variables de estudio difieren con respecto a las regiones. Use
# el procedimiento de analisis de varianza para un diseno de un factor.
# Que concluiria usted con un nivel de significancia alpha = 0.03

# H0: Las medias de ventas entre las regiones son iguales
dat=sales$ventas
fact=factor(sales$Region)
tapply(dat,fact,mean)
boxplot(dat~fact, main = "Caja de regiones", xlab = "Regiones", ylab= "Ventas")

# Se observa que los valores de Q2 son cercanos entre si,
# de manera que las medias de cada region se encuentran dentro del
# rango intercuartil resto de las regiones, los cuales son similares entre si.
mod.lm=lm(dat~fact)
anova(mod.lm)

# Como pvalor>0.03 no podemos rechazar la hipotesis nula. Entonces no hay
# suficiente evidencia para concluir que que las ventas medias de las
# variables de estudio difieren con especto a las regiones.
```