Tipología y ciclo de vida de los datos

Ander Lopetegui

03/01/2022

Contents

* Detalles de la actividad

1. Descripción

En esta actividad se elabora un caso práctico consistente en el tratamiento de un conjunto de datos. Orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

2. Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

3. Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

* Desarrollo de la actividad

1. Descripción del dataset.

¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset se ha obtenido de la siguiente página: https://rdrr.io/cran/ISLR/man/Carseats.html#heading-0 El dataset contiene la informaión de ventas de asientos de niños para el coche de 400 tiendas diferentes. Entre los campos de este conjunto de datos, encontramos las siguientes 11 variables:

- **Sales**: Número de ventas (en miles) por cada localización
- **CompPrice**: Precio de de venta de las sillas de la competencia
- **Income**: Salario medio de la comunidad (en miles de dolares)
- **Population**: Tamaño de la población (en miles)
- **Price** - Precio de venta de la sillas
- **ShelveLoc** - Un factor con niveles para indicar la calidad de la localización de la estantería por cada sitio: (Bad,Medium, Good)
- **Age** - Edad media de la población local
- **Education** - Nivel de educación por cada localizaciónEducation.
- **Urban** - Un factor con niveles 'No' y 'Yes' para indicar si una tienda tiene una localización urbana o rural.
- **US** - Un factor con niveles 'No','Yes' para indicar si una tienda está en US o no.

Con este dataset queremos realizar una serie de análisis en relación a las variable Sales Price, para evaluar si hay diferencias entre las ventas dentro de USA y fuera de USA. Queremos obtener respuesta a las siguientes preguntas:

1. Tiene la variable Sales la misma media poblacional dentro y fiera de USA?
2. Cual es el modelo de regresión lineal de Ventas respecto a precio dentro y fuera de USA?
3. Son las ventas de producto superiores en las tiendas de USA que fuera de USA?
4. Tienen diferente estrategia de precios las tiendas dentro de USA y fuera de USA?

2. Integración y selección de los datos de interés a analizar

Para el análisis que vamos a realizar nos vamos a centrar en las variables Sales, CompPrice, Price, y US. Por lo tanto podemos eliminar el resto de variables del data set. Analizamos el dataset

```
#Carga del archivo de formato csv
CarSeatsAll <- read.csv('Carseats.csv',stringsAsFactors = FALSE)

#Analizamos los datos del dataset. Podemos ver las 11 variables de las que está
#compuesto el dataset, con una logitud de 400 registos.
#Comrobamos que las variables Sales y Advertising tienen valores mínimos igual a 0
summary(CarSeatsAll)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
```

```
##    Population        Price        ShelveLoc              Age
##  Min.   : 10.0   Min.   : 24.0   Length:400         Min.   :25.00
##  1st Qu.:139.0   1st Qu.:100.0   Class :character   1st Qu.:39.75
##  Median :272.0   Median :117.0   Mode  :character   Median :54.50
##  Mean   :264.8   Mean   :115.8                      Mean   :53.32
##  3rd Qu.:398.5   3rd Qu.:131.0                      3rd Qu.:66.00
##  Max.   :509.0   Max.   :191.0                      Max.   :80.00
##    Education       Urban               US
##  Min.   :10.0   Length:400         Length:400
##  1st Qu.:12.0   Class :character   Class :character
##  Median :14.0   Mode  :character   Mode  :character
##  Mean   :13.9
##  3rd Qu.:16.0
##  Max.   :18.0
```

Nos quedamos únicamente con las variables que necesitamos del datase.

```
CarSeats <- CarSeatsAll[c(1,2,6,11)]

#Con esta otra función podemos ver también algunos valores de cada variable,
str(CarSeats)
```

```
## 'data.frame':    400 obs. of  4 variables:
##  $ Sales    : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice: int  138 111 113 117 141 124 115 136 132 132 ...
##  $ Price    : int  120 83 80 97 128 72 108 120 124 124 ...
##  $ US       : chr  "Yes" "Yes" "Yes" "Yes" ...
```

3. Limpieza de los datos

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Primero comprobamos con al función sumari si los datos contienen valores con 0

```
#La función summary nos indican los valores máximos mínimos y media,
#así mpodemos comprobar si hay valores igual a 0
summary(CarSeats)
```

```
##      Sales          CompPrice       Price           US
##  Min.   : 0.000   Min.   : 77   Min.   : 24.0   Length:400
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.:100.0   Class :character
##  Median : 7.490   Median :125   Median :117.0   Mode  :character
##  Mean   : 7.496   Mean   :125   Mean   :115.8
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.:131.0
##  Max.   :16.270   Max.   :175   Max.   :191.0
```

```
#Observamos que únicamente la variable Sales tiene valores con 0
#Vamos a tratar de ahora comprobar si son válidos los datos con Sales igual a 0.
subset <- subset(CarSeats, Sales==0)
head(subset)
```

```
##     Sales CompPrice Price US
## 175     0       139   185 No
```

tenemos un único registro con este valor, y no parece razonable que no haya ni una única venta, además siendo únicamente un registro podemos considerar un valor perdido, por lo que procedemos a la eliminación de este registro del dataset.

```
CarSeats[CarSeats$Sales == 0, ]
```

```
##     Sales CompPrice Price US
## 175     0       139   185 No
```

```
#Eliminamos registros con Sales = 0
#Para ello hacemos un subset con Sales distinto a 0
CarSeats <- subset(CarSeats, Sales!=0)
```

Comprobamos si existen valores nulos en alguna columna.

```
#Observamos que no hay ningún valor nulo
summary(CarSeats)
```

```
##      Sales          CompPrice         Price           US
##  Min.   : 0.160   Min.   : 77.0   Min.   : 24.0   Length:399
##  1st Qu.: 5.410   1st Qu.:115.0   1st Qu.:100.0   Class :character
##  Median : 7.490   Median :125.0   Median :117.0   Mode  :character
##  Mean   : 7.515   Mean   :124.9   Mean   :115.6
##  3rd Qu.: 9.320   3rd Qu.:135.0   3rd Qu.:131.0
##  Max.   :16.270   Max.   :175.0   Max.   :191.0
```

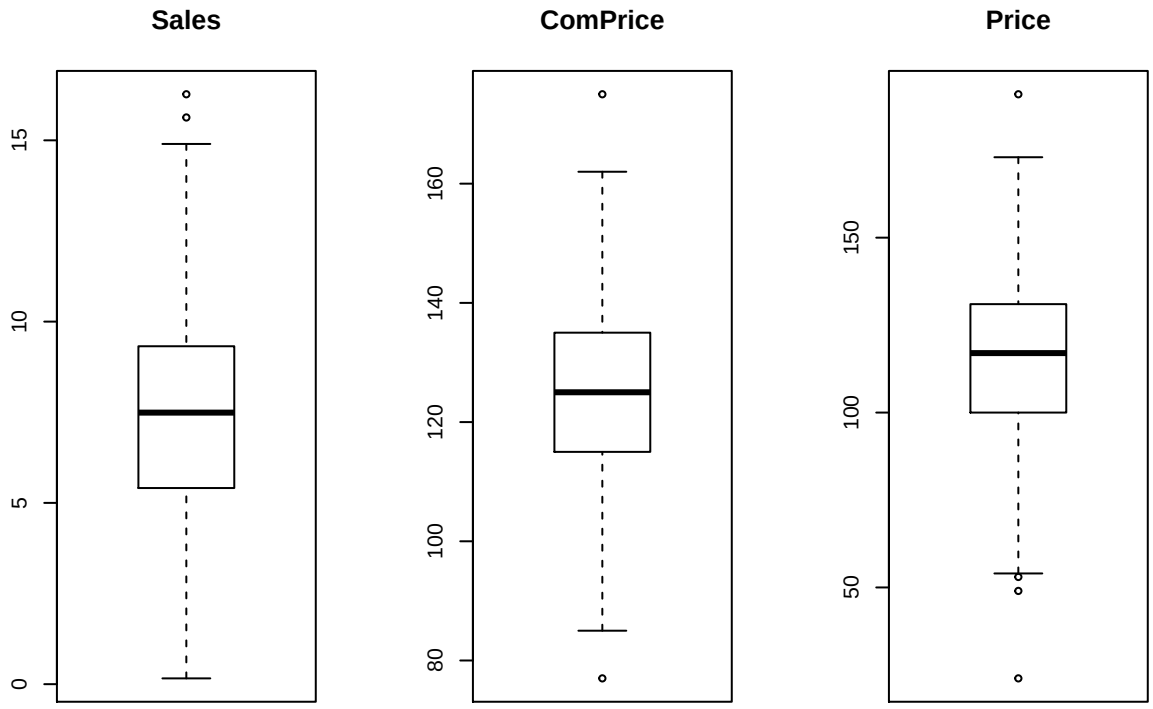
```
print(colSums(is.na(CarSeats)))
```

```
##     Sales CompPrice     Price        US
##         0         0         0         0
```

3.2. Identificación y tratamiento de valores extremos.

Podemos identificar los valores extrmos con las cajas boxplot para las variables numéricas La función boxplot() detecta outliers como todo valor que está más allá de los bigotes. Los bigotes son las líneas que se determinan como el tercer cuartil + 1.5 veces el rango intercuartílico (Tercer cuartil menos el primer cuartil) y el primer cuartil -1.5 veces el rango intercuartílico

```
#Analizamos cada variable numérica
par(mfrow=c(1,3))
g_caja1<-boxplot(CarSeats$Sales,main="Sales")
g_caja2<-boxplot(CarSeats$CompPrice,main="ComPrice")
g_caja3<-boxplot(CarSeats$Price,main="Price")
```



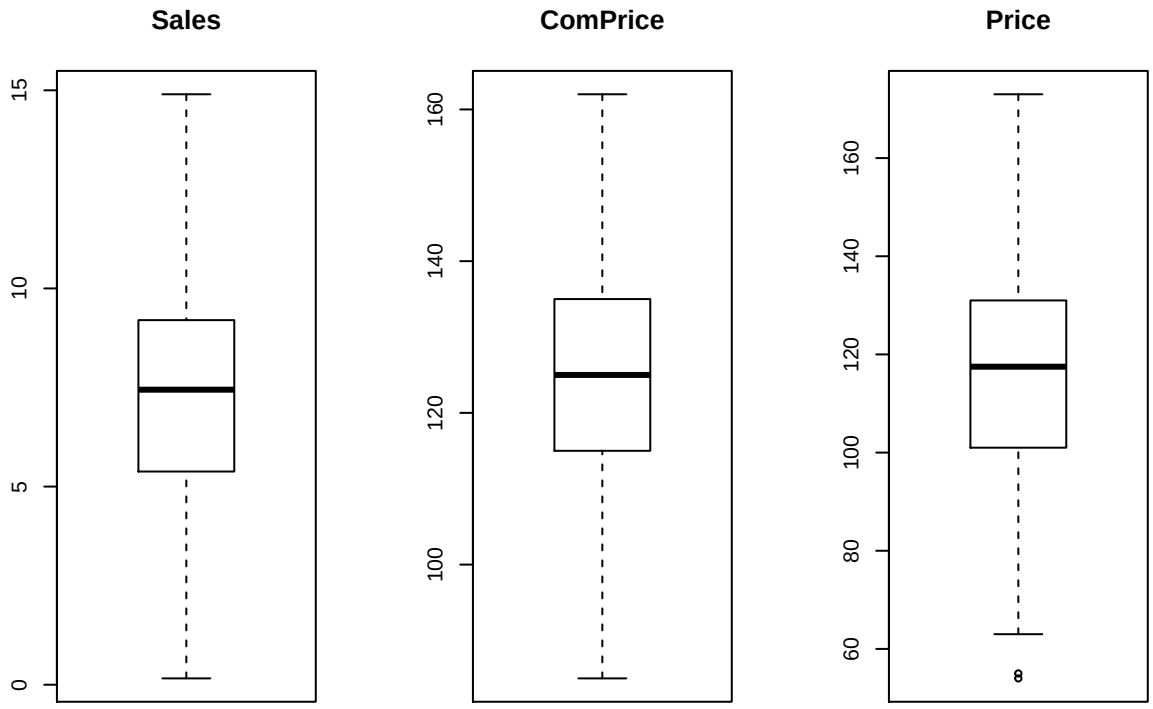

```
#Podemnos observar que hay puntos fuera de los vigotes en los tres gráficos
#Obtenemos los outliers de cada variable
outliers1<-g_caja1$out
outliers2<-g_caja2$out
outliers3<-g_caja3$out
```

Para el tratamiento de los outliers se puede hacer una sustitución de las variables, una discretización, o simplemente se pueden eliminar esos registros del dataset. En este caso vamos a optar por la eliminación, que se haría de la siguiente manera:

```
CarSeats<-CarSeats[-which(CarSeats$Sales %in% outliers1), ]
CarSeats<-CarSeats[-which(CarSeats$CompPrice %in% outliers2), ]
CarSeats<-CarSeats[-which(CarSeats$Price %in% outliers3), ]
```

Este serían los gráficos después de eliminar los ourliers

```
#Gráficos bloxplot después de eliminar outliers

par(mfrow=c(1,3))
boxplot(CarSeats$Sales,main="Sales")
boxplot(CarSeats$CompPrice,main="ComPrice")
boxplot(CarSeats$Price,main="Price")
```



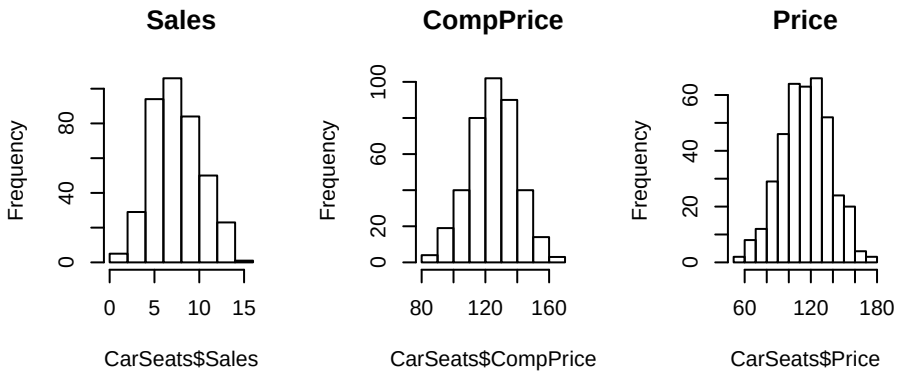
Una vez tenemos el dataset como queremos lo guardamos en otro csv de la siguiente manera:

```
write.csv(CarSeats,"CarSeats_clean.csv", row.names = FALSE)
```

4. Análisis general de los datos y planificación

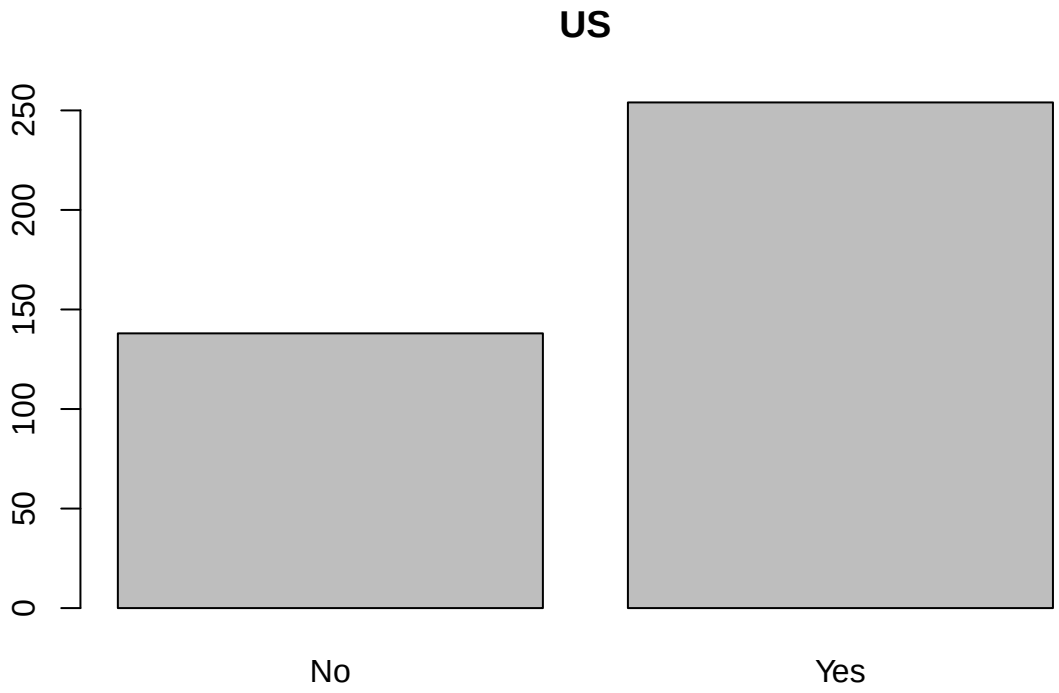
Analizamos las variables numéricas con gráficos de tipo histograma para hacernos una idea de la normalidad o no normalidad de cada variable. Las varibles normales se aproximan al dibujo de una campana de Gauss

```
#Histogramas
par(mfrow=c(2,4))
hist(CarSeats$Sales,main="Sales")
hist(CarSeats$CompPrice,main="CompPrice")
hist(CarSeats$Price,main="Price")

#Las tres varibles se aproximan a una campana de Gauss
```



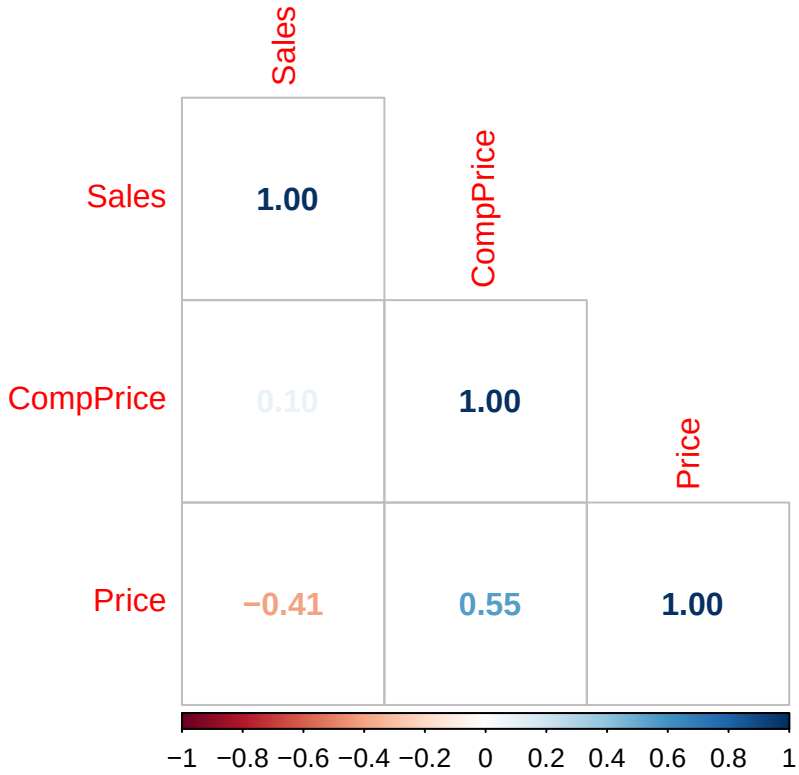
Factorizamos las variable discreta US de tipo character para posteriores análisis

```
#Factorizamos
CarSeats$US<-factor(CarSeats$US)

#Mostramos gráfico
plot(CarSeats$US,main="US")
```



Podemos realizar un matriz de correlación para analizar la relación que existe entre las diferentes variables numéricas

```
# Calculamos la matriz de correlación
M <- cor(CarSeats[, c(1,2,3)])

# Y Finalmente visualizamos en un gráfico la matriz de correlación,
#donde observamos una matriz que mustra la correlación a pares entre
#todas las variables numéricas del dataset
library("corrplot")
par( mfrow = c(1,1) )
corrplot(M,method='number',type = "lower")
```



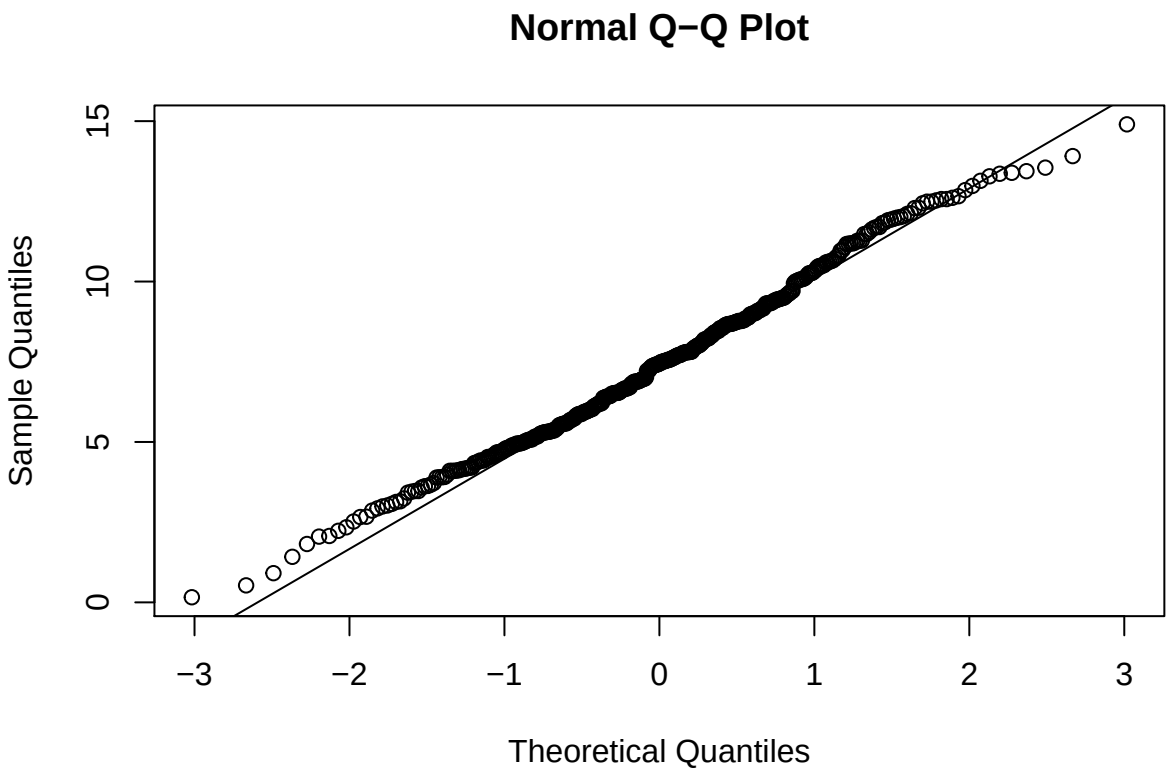
Este gráfico indica que los valores cuanto más cercanos a 1 o -1 mayor es la correlation. Vemos por ejemplo que hay una correlación negativa entre precio y venta, es decir que cuanto más alto es el precio menores son las ventas, Por otro lado se aprecio una correlación positiva entre precio y precio competencia, cuanto mayor es el precio, mayor es el precio de la competencia.

Comprobamos el intervalo de confianza de la media poblacional de la variable Sales.

```
#Test de normalidad para la variable Sales.
shapiro.test(CarSeats$Sales)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CarSeats$Sales
## W = 0.99343, p-value = 0.08591
```

```
#Valores de p del test inferiores al nivel de significancia alfa (0.05)
#permiten rechazar la hipótesis nula y, por lo tanto, llevarían a descartar la
#normalidad de los datos. En el ejemplo mostrado, no se puede rechazar la hipótesis nula,
#y se concluye que se puede asumir la normalidad de los datos.

#De manera cmplementaria podemos visualizar la desviación de los datos
#de la muestra en relación con una población normal, con el gráfico
# Q-Q, donde la Q denota cuantil
qqnorm(CarSeats$Sales)
qqline(CarSeats$Sales)
```




```
#En este caso los puntos están prácticamente sobre la línea y, por lo tanto,
# se puede asumir normalidad, confirmando el resultado del test Shapiro-Wilk.
```

La variable Sales tiene una distribución normal con una varianza desconocida por que se trata de una muestra poblacional, por lo tanto corresponde a una distribución t-student

```
#Comprobamos el intervalo de confianza con la función test y obtenemos que el
#intervalo de confianza están entre los valores 7.18-7.72
t.test(CarSeats$Sales)
```

```
##
##  One Sample t-test
##
## data:  CarSeats$Sales
## t = 54.502, df = 391, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  7.186643 7.724531
## sample estimates:
## mean of x
##  7.455587
```

El intervalo de confianza permie establecer los valores en los cuales se encuentra un determinado parámetro de la población, en esta caso indica con una confianza de 95% que el valor de las ventas unitarias en cada ubicación estará en un rango entre 7.18 y 7.72 miles. Podemos afirmar que si obtenemos infinitas muestras de las ventas realizadas en cada ubicación, el 95% de los itervalos de confianza calculados a partir de estas muestras contendrían el valor medio de las ventas unitarias.

5. Pruebas estadísticas y Resultados

5.1. Tiene la variable Sales la misma media poblacional dentro y fiera de USA?

Para ello debemos calcular el intervalo de confianza de los dos grupos

```
#Separamos la variable Sales del dataset entre los que son de US y los que no.
listaUS <- CarSeats$Sales[CarSeats$US=="Yes"]
listaNoUS <- CarSeats$Sales[CarSeats$US=="No"]

t.test(listaUS)
```

```
##
##  One Sample t-test
##
## data:  listaUS
## t = 45.076, df = 253, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  7.484292 8.168149
## sample estimates:
## mean of x
##   7.82622
```

```
t.test(listaNoUS)
```

```
##
##  One Sample t-test
##
## data:  listaNoUS
## t = 32.291, df = 137, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6.358617 7.188194
## sample estimates:
## mean of x
##  6.773406
```

Podemos observar que tienen un intervalo de confianza distinto, donde los rangos de un grupo no están dentro del otro. Por lo tanto podemos concluir que las dos variables tienen medias poblacionales diferentes

5.2. Cual es el modelo de regresión lineal de Ventas respecto a precio dentro y fuera de USA?

Estimamos por mínimos cuadrados ordinarios dos modelos lineales que expliquen la variable Sales en función de la variable Price, por un lado para ventas en US, y por otro para ventas fuera de US

```
#Utilizamos la función lm en ambos caso para realizar la regresión lineal
#Variable dependiente Sales(Y) en función de la variable Price (x)
listaUS <- subset(CarSeats, US=="Yes")
listaNoUS <- subset(CarSeats, US=="No")

head (listaUS)
```

```
##   Sales CompPrice Price  US
## 1  9.50       138   120 Yes
## 2 11.22       111    83 Yes
## 3 10.06       113    80 Yes
## 4  7.40       117    97 Yes
```

```
## 6 10.81       124    72 Yes
## 8 11.85       136   120 Yes
```

```
head (listaNoUS)
```

```
##    Sales CompPrice Price US
## 5   4.15       141   128 No
## 7   6.63       115   108 No
## 9   6.54       132   124 No
## 13  3.98       122   136 No
## 16  8.71       149   144 No
## 17  7.58       118   110 No
```

```
regresionUS <- lm(Sales ~ Price, data = listaUS)

regresionNoUS <- lm(Sales ~ Price, data = listaNoUS)
```

Describimos el modelo con la función summary

```
#Con la función summary podemos observar los dos modelos y así compararlos
summary(regresionUS)
```

```
##
## Call:
## lm(formula = Sales ~ Price, data = listaUS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8455 -1.7865  0.0611  1.7155  6.4529
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.783616   0.866515  15.907  < 2e-16 ***
## Price       -0.051093   0.007305  -6.994 2.39e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.537 on 252 degrees of freedom
## Multiple R-squared:  0.1626, Adjusted R-squared:  0.1592
## F-statistic: 48.92 on 1 and 252 DF,  p-value: 2.392e-11
```

```
summary(regresionNoUS)
```

```
##
## Call:
## lm(formula = Sales ~ Price, data = listaNoUS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7638 -1.4160 -0.2168  1.3896  6.4827
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.502252   0.970013  12.889  < 2e-16 ***
## Price       -0.049816   0.008277  -6.019 1.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.198 on 136 degrees of freedom
## Multiple R-squared:  0.2103, Adjusted R-squared:  0.2045
## F-statistic: 36.23 on 1 and 136 DF,  p-value: 1.54e-08
```

Podemos afirmar que la variable Price es significativa, porque tiene un p-valor menor que 0.05 en los dos casos. Por otro lado podemos observar que la precisión del modelo no es muy alta porque tienen valores R cuadrado de entre el 15 y 20 %.

Con el modelo podemos hacer la predicción de el número ventas que pueden haber en función de precio:

```
#Con Precio de 100, tenemos una predicción de 8.67 (miles) de ventas dentro de US
predict(regresionUS, data.frame(Price=100),
        interval = "prediction", level = 0.95)
```

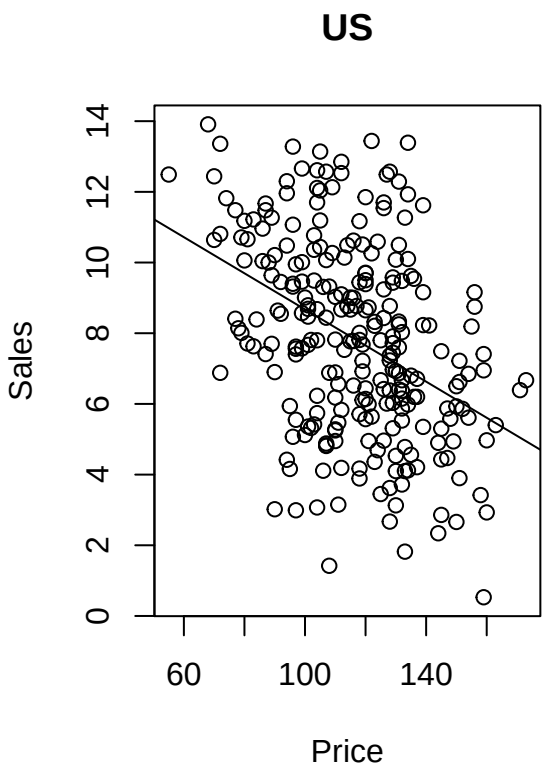
```
##        fit      lwr      upr
## 1 8.674288 3.661925 13.68665
```

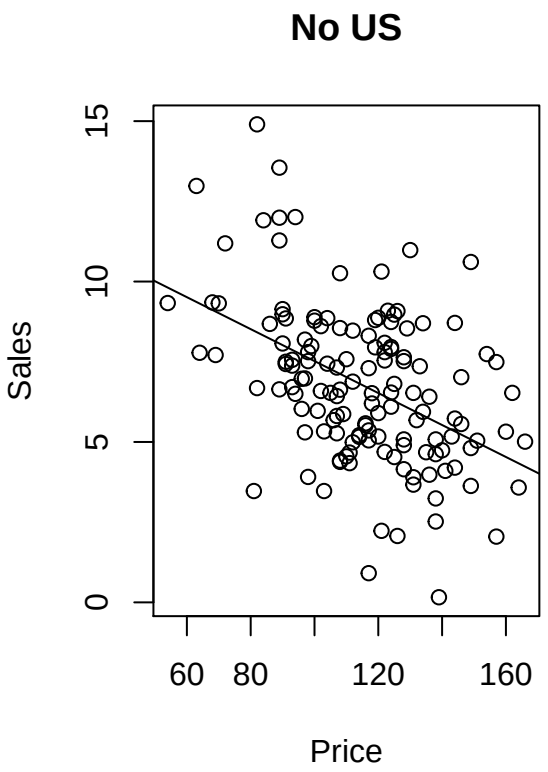
```
#Los mismo podemos hacer con el modelo fuera de US
predict(regresionNoUS, data.frame(Price=100),
        interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 7.520647 3.151886 11.88941
```

Representación gráfica del modelo:

```
#Variable dependiente Sales(Y) en función de la variable Price (x)

par(mfrow=c(1,2))
plot(listaUS$Price, listaUS$Sales, xlab='Price', ylab='Sales', main='US')
abline(regresionUS)
plot(listaNoUS$Price, listaNoUS$Sales, xlab='Price', ylab='Sales', main='No US')
abline(regresionNoUS)
```





Podemos observar en ambos la recta de regresión con pendiente descendiente, esto significa que cuanto el precio es mayor, menos ventas tenemos. En el gráfico se aprecia que la segunda recta se ajusta mejor a los puntos.

5.3. Son las ventas de producto superiores en las tiendas de USA que fuera de USA?

Hipótesis nula: La media de ventas en las tiendas de USA son igual a la media de ventas en las tiendas fuera de USA. Hipótesis alternativa: La media de ventas en las tiendas de USA son superiores a la media de ventas en las tiendas fuera de USA.

Debemos realizar el siguinete test -> Contrastes de dos muestras independientes sobre la media con varianzas desconocidas No conocemos la varianza de la poblacion por lo que debemos decidir si estamos en una de estas dos situaciones:

- las varianzas de las dos poblaciones son desconocidas pero iguales a un cierto parámetro
- o las dos varianzas son diferentes. Para decidir cuál de las dos situaciones se da, se puede realizar un test de igualdad de varianzas de dos muestras. Este tipo de test se denomina test de homoscedasticidad.

```
#Primero debemos hacer el test de homocentrisidad

#Test de igualdad de varianzas.
#Separamos el dataset en dos grupos:
H <- CarSeats$Sales[ CarSeats$US=="Yes" ]
D <- CarSeats$Sales[ CarSeats$US=="No" ]

#Calculamos con la función test
t.test(H,D)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  H and D
## t = 3.8665, df = 310.18, p-value = 0.0001345
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5170365 1.5885929
## sample estimates:
## mean of x mean of y
##  7.826220  6.773406
```

```
#El valor p obtenido es menor que el nivel de significancia (0.05),
#por lo tanto rechazamos la hipotesis nula que las varianzas son iguales
#Con lo cual debemos hacer los cálculos con la segunda opción.
```

Cálculos:

```
#Comprobamos la t de student y los grados de libertad
t.test(H,D,alternative = "greater",var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  H and D
## t = 3.8665, df = 310.18, p-value = 6.727e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.6035882       Inf
## sample estimates:
## mean of x mean of y
##  7.826220  6.773406
```

Conclusión: El valor p obtenido es inferior que el nivel de significancia, por lo tanto rechazamos la hipótesis nula de que las medias son iguales. Análogamente, debido a que la zona de aceptación es mayor que el valor observado, podemos aceptar la hipótesis nula que La media de ventas en las tiendas de USA son superiores a la media de ventas en las tiendas fuera de US

5.4. Diferencias en la estrategia de precios

¿la proporción de casos en los que el precio de la tienda es más bajo que la competencia (estrategia de precios bajos) es diferente en las tiendas de USA que en las tiendas fuera de USA?

Hipótesis nula: la proporción de casos en los que el precio de la tienda es más bajo que el de la competencia en las tiendas USA es igual que fuera de USA H0: p1=p2 Hipótesis alternativa: la proporción casos en los que el precio de la tienda es más bajo que el de la competencia en las tiendas USA es diferente que fuera de USA H1: P1 <> p2

Tipo de test: Contraste de dos muestras sobre la proporción, asumiendo que la muestra es grande

Cálculos:

```
library(dplyr)

CarSeatsUS<-CarSeats %>%  filter (US=="Yes")
CarSeatsNUS<-CarSeats %>%  filter (US=="No")

n1<-length(CarSeatsUS$CompPrice)
x1 <- CarSeatsUS %>% filter(Price < CompPrice)
```

```r
p1<-length(x1$Price)/length(CarSeatsUS$Price)

n2<-length(CarSeatsNUS$CompPrice)
x2 <- CarSeatsNUS %>% filter(Price < CompPrice)
p2<-length(x2$Price)/length(CarSeatsNUS$Price)

p<-(n1*p1 + n2*p2) / (n1+n2)
zobs <- (p1-p2)/( sqrt(p*(1-p)*(1/n1+1/n2)) )
zcrit <- qnorm(0.02, lower.tail=FALSE)
pvalue<- pnorm(zobs, lower.tail=FALSE)
c(zobs, zcrit, pvalue)
```

```
## [1] -0.6033117  2.0537489  0.7268493
```

```r
#Validación con prop.test: se construye un data frame que se le pasa
#a prop.test
success<-c( p1*n1, p2*n2)
nn<-c(n1,n2)
prop.test(success, nn, alternative="two.sided", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 0.36398, df = 1, p-value = 0.5463
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.12527520  0.06593479
## sample estimates:
##    prop 1    prop 2
## 0.6732283 0.7028986
```

Conclusión:

Con los resultados obtenidos, debemos rechazar la hipótesis nula puesto que el valor de p es 0.52, superior al nivel de confianza establecido 0.05. Dado que el valor observado es -0.64 y cae por fuera de la zona de aceptación, podemos concluir que la proporción casos en los que el precio de la tienda es más bajo que el de la competencia en las tiendas de USA es diferente que fuera de USA

6. Resolución del problema y conclusiones Finales

Con las pruebas realizadas hemos podido obtener las respuestas a las preguntas que nos hemos relizado inicialmente.

1. Tiene la variable Sales la misma media poblacional dentro y fuera de USA?

- No, no tienen las misma media poblacional, las ventas dentro de USA tiene una media poblaciona más alta.

2. Cual es el modelo de regresión lineal de Ventas respecto a precio dentro y fuera de USA?

- Hemos calculado el modello dee regresión de los dos grupos, ambos tienenuna recta descendente. La variable precio es significativa en las ventas. El modelo no tiene una gran precisión.

3. Son las ventas de producto superiores en las tiendas de USA que fuera de USA?

- La media de ventas en las tiendas de USA son superiores a la media de ventas en las tiendas fuera de US.

4. Tienen diferente estrategia de precios las tiendas dentro de USA y fuera de USA?

- Si, tienen una estrategia diferente. La proporción de casos en los que el precio de la tienda es más bajo que el de la competencia en las tiendas de USA es diferente que fuera de USA.

Como conclusión podemos decir que hemos comprobado significativas diferencias entre las ventas dentro de USA y fuera en cuanto al precio de las viviendas, las ventas realizadas o la estrategia de la competencia. Todo esto hemos podido comprobarlo mediante diferentes pruebas estadísticas después de analizar y limpiar el dataset de origen. Este dataset contiene otras variables que inicialmente hemos descartado porque quedan fuera del proósito de este análisis, pero sería interesante también poder analizarlas para estudiar otro tipo de relaciones.