ANÁLISIS MULTIVARIANTE UOC

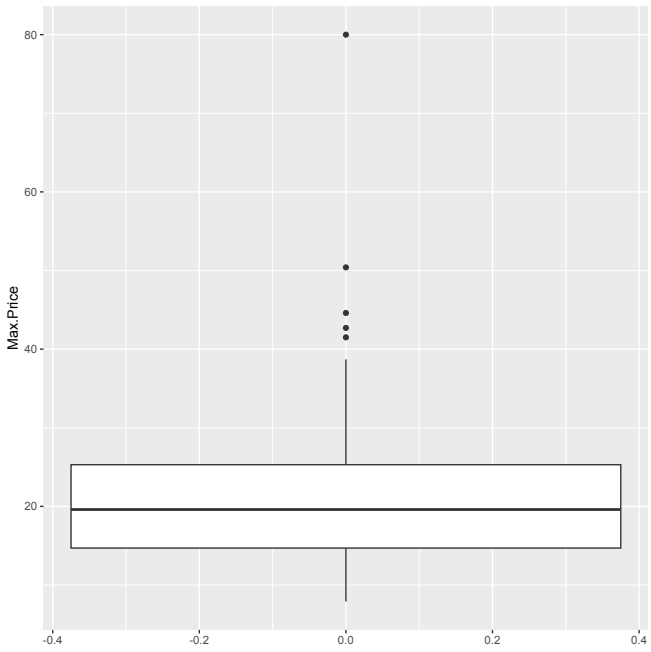
PEC 1 CURSO 24-25 Segundo Semetre

Pregunta 1:

Haced un diagrama de caja (box-plot) de la variable cuantitativa "Max.Price". Comentad los resultados obtenidos.

Obtenemos el diagrama de caja (box-plot) de la variable cuantitativa "Max.Price" con este código en R:

```
#Visualización de variables cuantitativas

g1<- ggplot( car, aes(y=Max.Price)) + geom_boxplot()

print(g1)
```



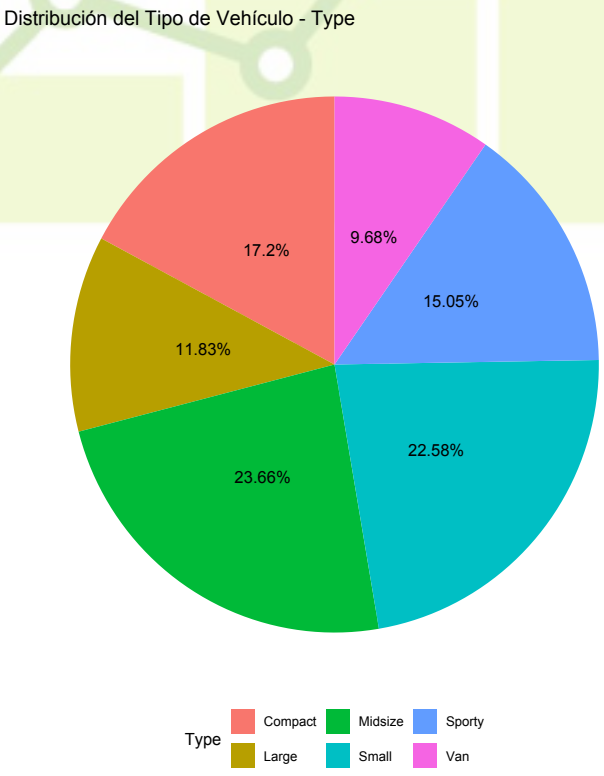
Como se puede apreciar en el diagrama de box- plot:

- la mediana de la variable precio es aproximadamente 20.000 dólares.
- El primer cuartil es aproximadamente 15.000 dólares.
- El tercer cuartil es aproximadamente 25.000 dólares.
- El recorrido intercuartílico (la diferencia entre el tercer cuartil y el primer cuartil) no es demasiado grande, aproximadamente 10.000 dólares.
- Existen 5 modelos de vehículo cuyo precio se puede considerar como atípico (superior a unos 40.000 dólares).

Pregunta 2: En esta pregunta consideraremos, además de variables cuantitativas, variables categóricas (cualitativas). En concreto, queremos analizar el tipo de vehículo. Para ello, haced en primer lugar, un gráfico de sectores según el tipo de vehículo ("Type"). Comentad los resultados obtenidos.

Obtenemos el gráfico de sectores de la variable cualitativa "Max.PriceType" con este código en R:

```
# Contar la frecuencia de cada tipo de vehículo

SumType <- summarize(group_by(car, Type), n = length(Type))

# Calcular los porcentajes

SumType$porcentaje <- (SumType$n / sum(SumType$n)) * 100

# Crear el gráfico de sectores

g2 <- ggplot(SumType, aes(x = "", y = n, fill = Type)) + coord_polar("y", start = 0) +

geom_bar(width = 1, stat = "identity") + geom_text(aes(label = paste0(round(porcentaje, 2), "%")),

          position = position_stack(vjust = 0.5)) +

theme_void() + theme(legend.position = "bottom") + ggtitle("Distribución del Tipo de Vehículo - Type")

print(g2)
```



El gráfico de sectores muestra la distribución de los tipos de vehículos. En este gráfico se puede ver qué tipo es más común y qué porcentaje representa cada uno.

Observamos:

- El modelo más frecuente en la muestra es “Midsize” que tiene un porcentaje de 23,66%.
- Le sigue muy de cerca “Small” con un porcentaje del 22,58%.
- Los modelos menos frecuentes son “Large” con un 11,83% y “Van” con un 9,68% que es el modelo menos frecuente.

Pregunta 3:

3a. Calculad las medias y los cuartiles de las variables cuantitativas Max.Price, MPG.city y MPG.highway. Comentad los resultados obtenidos.

Obtenemos el resumen de los cinco números en R de las variables “Max.Price”, “MPG.city” y “MPG.highway”:

```
> summary(car$Max.Price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    7.9    14.7    19.6    21.9    25.3    80.0
> summary(car$MPG.city)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   18.00   21.00   22.37   25.00   46.00
> summary(car$MPG.highway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   26.00   28.00   29.09   31.00   50.00
> IQR(car$Max.Price)
[1] 10.6
> IQR(car$MPG.city)
[1] 7
> IQR(car$MPG.highway)
[1] 5
```

- **Variable Max.Price**

La media de la variable Max.Price es 21.898,82 dólares, superior que la mediana que es 19.600 dólares, cosa que ya se apreciaba en el box-plot, ya que hay datos atípicos en la parte superior de la distribución. El recorrido intercuatílico es 10.600 dólares (25.300-14.700) que ya habíamos visto en el box-plot que no era demasiado grande. El vehículo más barato tiene un precio de 7.900 dólares y el más caro tiene un precio de 80.000 dólares.

- **Variable MPG.city**

La media de la variable MPG.city es 22,36 millas por galón en ciudad, ligeramente superior que la mediana que es 21 millas por galón en ciudad. El recorrido intercuatílico es 7 (25-18). El menor consumo de gasolina en ciudad es de 15 millas por galón y el mayor es de 46 millas por galón.

- **Variable MPG.highway**

La media de la variable MPG.highway es 29,08 millas por galón, la mediana que es 28 millas por galón. El recorrido intercuatílico es 5 (31-26). El menor consumo de gasolina es de 20 millas por galón y el mayor es de 50 millas por galón.

3b. Calculad el coeficiente de correlación entre las variables MPG.city y MPG.highway. Comentad el resultado obtenido.

```
> coef_corr <- cor(car$MPG.city, car$MPG.highway)
> print(coef_corr)
[1] 0.9439358
```

El coeficiente de correlación entre las variables MPG.city y MPG.highway es 0,943 que indica que existe una correlación positiva entre las variables (si una variable aumenta la otra también y viceversa) y también indica que esta relación es una relación fuerte ya que el valor del coeficiente se aproxima a 1.

Pregunta 4:

Obtened la tabla de frecuencias de la variable "Type". Comentad el resultado obtenido.

```
> tabla_frecuencias <- table(car$Type)
> print(tabla_frecuencias)

 Compact    Large Midsize    Small   Sporty      Van
      16       11       22       21       14        9
```

Las frecuencias absolutas son el número de veces que se observa cada categoría de la variable "Type" en la muestra. Como ya habíamos visto en el gráfico de sectores la categoría más frecuente es "Midsize" (22 veces), seguido de "Small" (21 veces) y que el que menos se repite es el "Van" (9 veces).

Pregunta 5:

Obtened la tabla de contingencia y realizad un contraste estadístico (prueba chi-cuadrado de Pearson) para analizar si existe asociación (dependencia) entre las variables cualitativas "Type" y "AirBags". ¿Las variables son dependientes o independientes?

```
> tabla_contingencia <- table(car$Type, car$AirBags)
> print(tabla_contingencia)

         Driver & Passenger Driver only None
  Compact                  2           9    5
  Large                    4           7    0
  Midsize                  7          11    4
  Small                    0           5   16
  Sporty                   3           8    3
  Van                      0           3    6
> prueba_chi <- chisq.test(tabla_contingencia)
Warning in chisq.test(tabla_contingencia) :
  Chi-squared approximation may be incorrect
> print(prueba_chi)

    Pearson's Chi-squared test

data:  tabla_contingencia
X-squared = 33.001, df = 10, p-value = 0.0002723
```

En la tabla de contingencia se puede observar el número de veces que se repite cada par de variables, vemos que la combinación más frecuente es "Type"=Small y "AirBags"=None que se repite 16 veces.

Para saber si estas dos variables son independientes hacemos el contraste de independencia de chi-cuadrado de Pearson.

El estadístico Chi-cuadrado es 33,001 con un p-valor asociado de 0,00027 (p-valor menor que 0,05) nos indica que podemos rechazar la hipótesis nula de independencia entre las dos variables "Type" y "AirBags". Podemos concluir que existe una relación de dependencia significativa entre las dos variables analizadas.

Pregunta 6:

Estimad un modelo de regresión lineal múltiple en el que el consumo en ciudad (MPG.city) venga explicado por las variables Horsepower, Origin y Type. A partir de los resultados obtenidos contestad las siguientes cuestiones:

```
> modelo <- lm(MPG.city ~ Horsepower + Origin + Type, data = car)
> summary(modelo)

Call:
lm(formula = MPG.city ~ Horsepower + Origin + Type, data = car)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9599 -1.6048  0.0632  1.3065 14.2087

Coefficients:
                   Estimate Std. Error t value   Pr(>|t|)
(Intercept)       27.292045   1.336380  20.422    < 2e-16 ***
Horsepower        -0.041035   0.008043  -5.102 0.00000202 ***
Origin[T.non-USA]  1.370773   0.711697   1.926   0.057439 .
Type[T.Large]     -1.564465   1.363502  -1.147   0.254441
Type[T.Midsize]   -1.391473   1.085532  -1.282   0.203386
Type[T.Small]      5.385448   1.094806   4.919 0.00000420 ***
Type[T.Sporty]     0.477681   1.177979   0.406   0.686122
Type[T.Van]       -4.768802   1.319450  -3.614   0.000509 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.137 on 85 degrees of freedom
Multiple R-squared:  0.712,   Adjusted R-squared:  0.6883
F-statistic: 30.03 on 7 and 85 DF,  p-value: < 2.2e-16
```

6a. Contrastad la significación individual del parámetro asociado a la variable Horsepower.

El parámetro asociado a la variable “Horsepower” es significativo puesto que el p-valor que tiene asociado el estadístico es 0,00000202, que es menor que el nivel de significación (0,05) por tanto rechazamos la hipótesis nula de no significación individual.

6b. Contrastad la significación global del modelo.

El modelo es globalmente significativo puesto que el p-valor que tiene asociado el estadístico F=30,03 es 2,2e-16 y este es menor que el nivel de significación (0,05) por tanto rechazamos la hipótesis nula de no significación global.

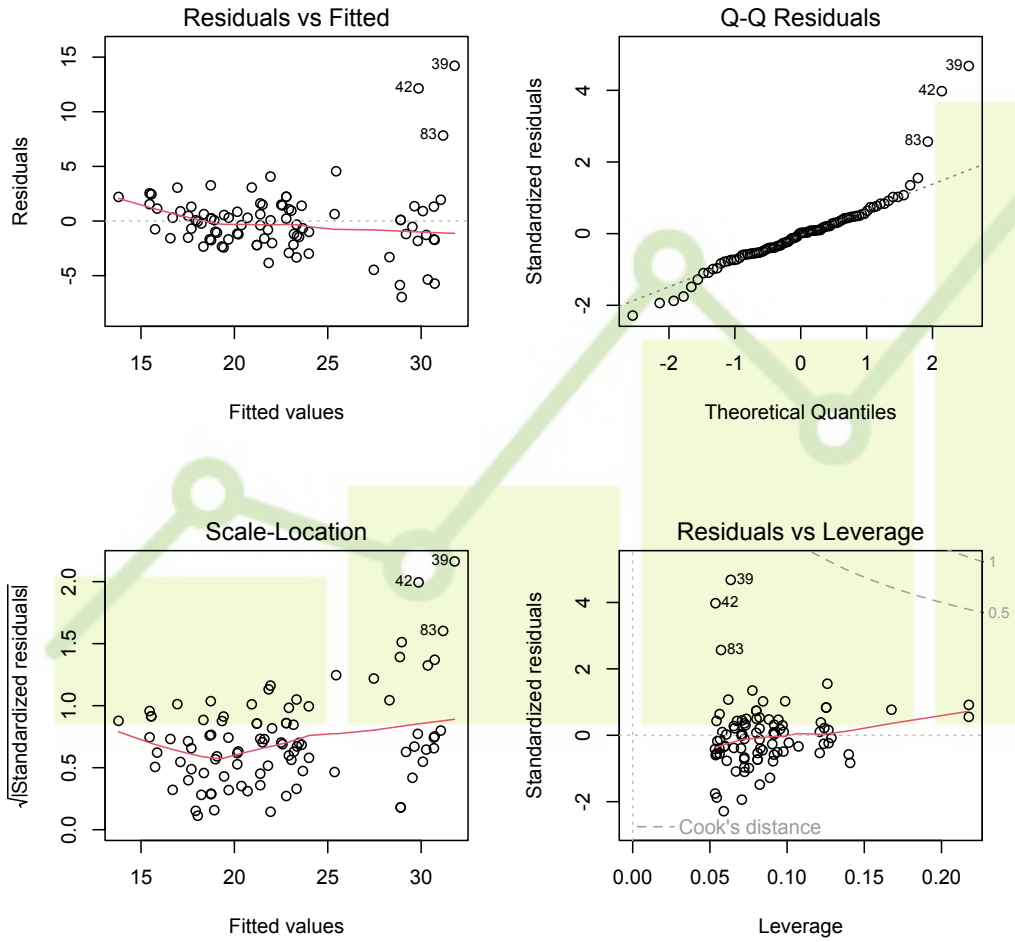
6c. Valorad la bondad del ajuste obtenido

El coeficiente de determinación R^2 es igual a 0,712, lo cual significa que con las variables exógenas (explicativas) consideradas conseguimos explicar el 71,2% de la variabilidad de la variable endógena MPG.city

Pregunta 7:

Con el fin de completar la valoración de la estimación obtenida del modelo de la pregunta anterior, queremos analizar los residuos de dicho modelo puesto que el hecho de que los residuos no sean esféricos (es decir, homoscedásticos y sin autocorrelación), implicará que las estimaciones que hemos obtenido no sean eficientes (es decir, de varianza mínima). ¿Hay indicios de problemas de heteroscedasticidad en la estimación? Para responder a esta pregunta realizad un diagnóstico gráfico de los residuos del modelo y comentad los resultados obtenidos.

```
par(mfrow=c(2,2))

plot(modelo)
```



- En el gráfico, “Residuals vs Fitted”, se puede determinar la linealidad del modelo. Cuanto más cercana sea la línea roja al 0 en abscisas, menos indicios hay de que el modelo siga un patrón y, por lo tanto, más aceptable es la hipótesis de linealidad. En nuestro caso, podemos concluir que no hay problemas de no linealidad en el modelo.

- En el gráfico “Normal Q-Q”, se puede comprobar la normalidad del modelo. Cuanto más cercanos estén los datos a la línea, mayor será el ajuste del modelo a la distribución normal. En este gráfico se observa que no hay problemas de no normalidad en el modelo aunque en las colas los puntos no se ajustan del todo. Los puntos que se observan en la cola derecha que no se ajustan bien a la línea probablemente son debido a a la presencia de valores extremos en la parte superior (lo hemos visto en los outliers del boxplot).

- En el gráfico “Scale-Location” se puede comprobar la homogeneidad del modelo o, más bien, la falta de heterocedasticidad. En nuestro caso, la no heterocedasticidad se aprecia con bastante claridad para las muestras más bajas puesto que se observa una línea prácticamente horizontal, pero existe una tendencia positiva para los valores ajustados más elevados.

- El gráfico “Residuals vs Leverage” muestra el estadístico leverage para cada observación. Tenemos algún caso de outliers, pero aparentemente no tenemos un problema importante de presencia de outliers.

Pregunta 8:

Con el fin de seguir completando la valoración de la estimación obtenida del modelo de la pregunta 6, ahora queremos analizar si hay o no hay problemas de multicolinealidad, puesto que si hay una fuerte correlación entre las variables independientes (esto es, multicolinealidad intensa), podemos tener un problema de sobre- información en el modelo, con información redundante. ¿Hay problemas de multicolinealidad en el modelo estimado? Para responder a esta pregunta calculad los factores de inflación de la varianza (VIF) y comentad los resultados obtenidos.

```
> vif(modelo)
                GVIF Df GVIF^(1/(2*Df))
Horsepower 1.658608  1        1.287869
Origin     1.195126  1        1.093218
Type       1.944036  5        1.068736
```

En esta salida de R se muestran los valores de cada uno de los VIF asociados a cada variable explicativa del modelo. Los tres VIF son inferiores a 5 por tanto no existen problemas de multicolinealidad en el modelo. Además todos los VIF tienen valores cercanos a 1 indicando que la correlación entre variables explicativas es muy baja.

CÓDIGO DE LA PRÁCTICA

```
# PREGUNTA 1: Diagrama de caja (box-plot) de Max.Price

g1 <- ggplot(car, aes(y=Max.Price)) + geom_boxplot()
print(g1)

# PREGUNTA 2: Gráfico de sectores por tipo de vehículo

# Contar la frecuencia de cada tipo de vehículo

SumType <- summarize(group_by(car, Type), n = length(Type))

# Calcular los porcentajes

SumType$porcentaje <- (SumType$n / sum(SumType$n)) * 100

# Crear el gráfico de sectores

g2 <- ggplot(SumType, aes(x = "", y = n, fill = Type)) +
  coord_polar("y", start = 0) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(aes(label = paste0(round(porcentaje, 2), "%")),
            position = position_stack(vjust = 0.5)) +
  theme_void() +
  theme(legend.position = "bottom") +
  ggtitle("Distribución del Tipo de Vehículo - Type")
print(g2)

# PREGUNTA 3a: Resumen de cinco números para variables cuantitativas

summary(car$Max.Price)
summary(car$MPG.city)
summary(car$MPG.highway)

# Rango intercuartílico (IQR)

IQR(car$Max.Price)
IQR(car$MPG.city)
IQR(car$MPG.highway)

# PREGUNTA 3b: Coeficiente de correlación

coef_corr <- cor(car$MPG.city, car$MPG.highway)
print(coef_corr)
```

```r
# PREGUNTA 4: Tabla de frecuencias de la variable Type

# Opción 1: Tabla de frecuencias

tabla_frecuencias <- table(car$Type)
print(tabla_frecuencias)

# Opción 2: Summary de variable cualitativa

varcual <- data.frame(car$Type)
summary(varcual)

# PREGUNTA 5: Tabla de contingencia y prueba chi-cuadrado

tabla_contingencia <- table(car$Type, car$AirBags)
print(tabla_contingencia)

# Opción 1: Chi-cuadrado con tabla de contingencia

prueba_chi <- chisq.test(tabla_contingencia)
print(prueba_chi)

# Opción 2: Chi-cuadrado directamente con las variables

chisq.test(car$AirBags, car$Type)

# PREGUNTA 6: Modelo de regresión lineal múltiple

modelo <- lm(MPG.city ~ Horsepower + Origin + Type, data = car)
summary(modelo)

# PREGUNTA 7: Diagnóstico gráfico de residuos

par(mfrow=c(2,2))
plot(modelo)

# PREGUNTA 8: Factores de inflación de la varianza (VIF)


vif(modelo)
```