PRACTICA 2: LIMPIEZA, VALIDACIÓN DE LOS DATOS, ANALISIS Y CREACION DE MODELOS.

Tipología y Ciclo de Vida de los Datos

Autor: Pedro José Ros Gómez

ENERO 2024

Contents

1. El Dataset.

Partimos del conjunto de datos obtenido anteriormente mediante web scraping, que contiene información detallada sobre inmuebles en la provincia de Ciudad Real. Nuestro objetivo es identificar el valor medio en Ciudad Real del alquiler de un inmueble a traves de sus caracterisiticas , crear un modelo de prediccion de precios y evaluación de caracteristicas de inmuebles utilizando sus diversas características como dimensiones, habitaciones, baños, amueblado, estacionamiento, trastero, y otras especificaciones relevantes. Este proceso implica la creación de un modelo de predicción de precios, donde exploraremos la relación entre las variables y los precios de alquiler. Además, llevaremos a cabo una evaluación minuciosa de las características de los inmuebles para comprender qué aspectos específicos influyen significativamente en los precios. La importancia de este enfoque radica en proporcionar información valiosa para propietarios, inquilinos y profesionales del sector inmobiliario, contribuyendo a una comprensión más profunda de los factores determinantes en la fijación de precios y permitiendo una predicción más precisa de los costos de alquiler en la zona de Ciudad Real.

Descripción del dataset.

Cargamos el dataSet:

```
datos <- read.csv("../data/datos_casas.csv")
summary(datos)
```

```
##     titulo             precio             lugar                C0
##  Length:123         Length:123         Length:123         Length:123
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##       C1                 C2                 C3                 C4
##  Length:123         Length:123         Length:123         Length:123
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##       L0                 L1            descripcion             C5
##  Length:123         Length:123         Length:123         Length:123
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##       C6                 L2                 L3                 C7
##  Length:123         Length:123         Length:123         Length:123
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Al cargar los datos, se evidencia un conjunto de 123 inmuebles cuyas características están detalladas en 14 columnas. La descripción de las columnas es la siguiente: **titulo:** *Título del anuncio del inmueble.* **precio:** *Precio del alquiler del inmueble en euro por mes.* **lugar:** *Ubicación del inmueble* **C0:** *Tamaño del inmueble en metros cuadrados.* **C1:** *Número de habitaciones en el inmueble.* **C2:** *Número de baños en el inmueble.* **C3:** *Característica del inmueble (diferentes datos a extraer)* **C4:** *Característica del inmueble (diferentes datos a extraer)* **C5:** *Característica del inmueble (diferentes datos a extraer)* **C6:** *Característica del inmueble (diferentes datos a extraer)* **C7:** *Característica del inmueble (diferentes datos a extraer)* **L0:** *Referencias a la población* **L1:** *Referencias a la provincia* **L2:** *Referencias a la provincia* **L3:** *Referencias a la provincia* **descripcion:** *Descripción detallada del anuncio del inmueble.*

Como se aprecia, los datos en su estado actual se encuentran en bruto, directamente extraídos de la página web de origen. Para llevar a cabo un análisis más preciso, es necesario realizar un proceso de limpieza que incluye cambiar los nombres de las columnas por descripciones más detalladas y extraer información significativa de las columnas de características. Además, se eliminarán aquellas columnas que no sean pertinentes para focalizarnos en los datos más relevantes de este estudio.

2. Limpieza de datos.

Iniciamos el proceso de limpieza del conjunto de datos. La primera etapa consiste en la eliminación de registros nulos. Posteriormente, creamos un nuevo conjunto de datos que servirá como base para llevar a cabo las operaciones de limpieza necesarias.

```
# Eliminacion de los datos null
datos_filtrados <- na.omit(datos)
```

El siguiente paso en el proceso de limpieza implica la transformación de las columnas de texto a formato numérico. Eliminamos las descripciones y unidades de medida de las columnas, dejando únicamente los datos necesarios. Este procedimiento es esencial para poder convertir las variables a tipo numérico en etapas posteriores del análisis.

```r
datos_filtrados$precio <- sub("€/mes", "", datos_filtrados$precio)
datos_filtrados$lugar <- sub(" - Ver mapa", "", datos_filtrados$lugar)
datos_filtrados$C0 <- sub(" m2", "", datos_filtrados$C0)
datos_filtrados$C1 <- sub(" hab.", "", datos_filtrados$C1)
datos_filtrados$C2 <- sub(" baño", "", datos_filtrados$C2)
datos_filtrados$C2 <- sub(" baños", "", datos_filtrados$C2)
datos_filtrados$C2 <- sub("s", "", datos_filtrados$C2)
```

Realizamos la extraccion de información de las descripciones. En este caso nos es interesante extraer para el estudio las caracteristicas : **Amueblado:** *si el inmueble esta o no amueblado* **Parking:** *si incluye el parking en el precio* **Trastero:** *si tiene trastero* **Balcón|Terraza:** *si tiene balcón y/o terraza* Para cada característica, creamos columnas adicionales asignando el valor (1) si está presente y (0) si no lo está.

```r
datos_filtrados <- datos_filtrados %>%
  mutate(amueblado = NA) %>%
  select(1:6, amueblado, 7:ncol(.))
datos_filtrados$amueblado <- ifelse(grepl("Amueblado", datos_filtrados$C3, ignore.case = TRUE), 1, 0)
datos_filtrados <- datos_filtrados %>%
  mutate(parking = NA) %>%
  select(1:7, parking, 8:ncol(.))
datos_filtrados$parking <- ifelse(grepl("Parking", datos_filtrados$C3, ignore.case = TRUE), 1, 0)
datos_filtrados <- datos_filtrados %>%
  mutate(trastero = NA) %>%
  select(1:8,trastero, 9:ncol(.))
datos_filtrados$trastero <- ifelse(grepl("Trastero", datos_filtrados$C3, ignore.case = TRUE), 1, 0)
datos_filtrados <- datos_filtrados %>%
  mutate(balTerraza = NA) %>%
  select(1:9,balTerraza, 10:ncol(.))
datos_filtrados$balTerraza <- ifelse(grepl("Balcón|Terraza", datos_filtrados$C3, ignore.case = TRUE), 1
```

A continuacion cambiamos los nombres de las columnas por otros mas significativos al contenido.

```r
names(datos_filtrados)[names(datos_filtrados) == "C0"] <- "dimensiones_m2"
names(datos_filtrados)[names(datos_filtrados) == "C1"] <- "habitaciones"
names(datos_filtrados)[names(datos_filtrados) == "C2"] <- "baños"
names(datos_filtrados)[names(datos_filtrados) == "C3"] <- "caracteristicas A"
names(datos_filtrados)[names(datos_filtrados) == "C4"] <- "caracteristicas B"
summary(datos_filtrados)
```

```
##     titulo             precio             lugar           dimensiones_m2
##  Length:123         Length:123         Length:123         Length:123
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  habitaciones          baños             amueblado         parking
##  Length:123         Length:123         Min.   :0.0000   Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :0.0000   Median :0.0000
##                                        Mean   :0.3577   Mean   :0.1789
##                                        3rd Qu.:1.0000   3rd Qu.:0.0000
```

```
##                                       Max.   :1.0000   Max.   :1.0000
##     trastero         balTerraza    caracteristicas A  caracteristicas B
##  Min.   :0.00000   Min.   :0.000   Length:123         Length:123
##  1st Qu.:0.00000   1st Qu.:0.000   Class :character   Class :character
##  Median :0.00000   Median :0.000   Mode  :character   Mode  :character
##  Mean   :0.05691   Mean   :0.439
##  3rd Qu.:0.00000   3rd Qu.:1.000
##  Max.   :1.00000   Max.   :1.000
##       L0                L1            descripcion            C5
##  Length:123         Length:123         Length:123         Length:123
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       C6                L2                L3                C7
##  Length:123         Length:123         Length:123         Length:123
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
datos_filtrados <- datos_filtrados[, -c(13, 14, 16:ncol(datos_filtrados))]
datos_filtrados$baños <- ifelse(datos_filtrados$baños %in% c(1, 2), datos_filtrados$baños, 1)
```

Filtramos por las columnas que nos nos proporcionan informacion relevante.Por tanto, nos queda el dataSet con las siguientes columnas y sus descripciones: **titulo:** *Título del anuncio del inmueble.* **precio:** *Precio del alquiler del inmueble.* **lugar:** *Ubicación del inmueble.* **dimensiones_m2:** *Tamaño del inmueble en metros cuadrados.* **habitaciones:** *Número de habitaciones en el inmueble.* **baños:** *Número de baños en el inmueble.* **amueblado:** *Indicador de si el inmueble se ofrece amueblado (1) o no (0).* **parking:** *Indicador de si el inmueble cuenta con plaza de aparcamiento (1) o no (0).* **trastero:** *Indicador de si el inmueble tiene trastero (1) o no (0).* **balTerraza:** *Indicador de si el inmueble cuenta con balcón o terraza (1) o no (0).* **caracteristicas A:** *Posible categoría o característica específica del inmueble.* **caracteristicas B:** *Posible categoría o característica adicional del inmueble.* **descripcion:** *Descripción detallada del anuncio del inmueble.*

Integración y selección

Una vez confeccionado correctamente nuestro dataSet vamos a seleccionar unos atributos para la generacion de nuestro modelo:

precio: *Precio del alquiler del inmueble.* **dimensiones_m2:** *Tamaño del inmueble en metros cuadrados.* **habitaciones:** *Número de habitaciones en el inmueble.* **baños:** *Número de baños en el inmueble.* **amueblado:** *Indicador de si el inmueble se ofrece amueblado (1) o no (0).* **parking:** *Indicador de si el inmueble cuenta con plaza de aparcamiento (1) o no (0).* **trastero:** *Indicador de si el inmueble tiene trastero (1) o no (0).* **balTerraza:** *Indicador de si el inmueble cuenta con balcón o terraza (1) o no (0).*

Los atributos seleccionados para la generación del modelo son relevantes para entender y predecir el precio de alquiler de un inmueble. El precio es la variable objetivo, mientras que las dimensiones en metros cuadrados, el número de habitaciones, el número de baños, la disponibilidad de amueblado, aparcamiento, trastero, balcón o terraza son características esenciales que pueden influir significativamente en el precio de alquiler de un inmueble. Estos atributos representan aspectos clave que los inquilinos suelen considerar al buscar propiedades para alquilar, y su inclusión en el modelo permite capturar las variaciones en los precios

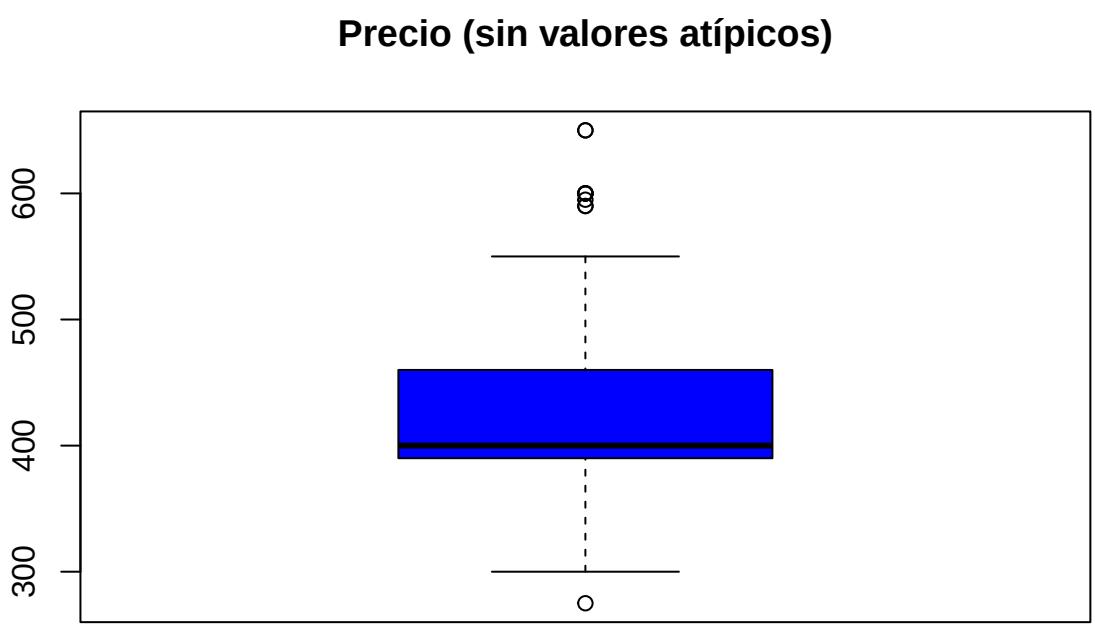
asociadas a estas características. Por tanto.Para nuestro analisis eliminamos las columnas que no nos valen y nos quedamos solo con las mencionadas anteriormente.Generamos un nuevo dataSet(datos_modelo) con el que partiremos para nuestro análisis.

```
datos_modelo <- datos_filtrados[, -c(1,3,13,11:ncol(datos_filtrados))]
datos_modelo[1:8] <- lapply(datos_modelo[1:8], as.numeric)
summary(datos_modelo)
```

```
##      precio      dimensiones_m2    habitaciones       baños
##  Min.   :  2.8   Min.   : 42.00   Min.   :0.000   Min.   :1.000
##  1st Qu.:395.0   1st Qu.: 80.00   1st Qu.:2.000   1st Qu.:1.000
##  Median :420.0   Median : 90.00   Median :3.000   Median :1.000
##  Mean   :447.2   Mean   : 95.24   Mean   :2.667   Mean   :1.415
##  3rd Qu.:500.0   3rd Qu.:107.50   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :880.0   Max.   :216.00   Max.   :8.000   Max.   :2.000
##    amueblado         parking          trastero         balTerraza
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.000
##  Mean   :0.3577   Mean   :0.1789   Mean   :0.05691   Mean   :0.439
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.000
```

Por ultimo eliminamos los valores extremo para que no se desvirtue el modelo. Representamos los valores en un gráfico de cajas.

```
# Calcular el rango intercuartílico (IQR) para la columna 'precio'
Q1 <- quantile(datos_modelo$precio, 0.25)
Q3 <- quantile(datos_modelo$precio, 0.75)
IQR <- Q3 - Q1
# Definir los límites para identificar valores atípicos
lower_limit <- Q1 - 1.5 * IQR
upper_limit <- Q3 + 1.5 * IQR
# Identificar valores atípicos
outliers <- datos_modelo$precio < lower_limit | datos_modelo$precio > upper_limit
# Eliminar filas con valores atípicos
datos_modelo <- datos_modelo[!outliers, ]
# Visualizar el efecto de la eliminación
boxplot(datos_modelo$precio, main = "Precio (sin valores atípicos)", col = "blue", border = "black")
```



Ya tenemos el dataSet listo para el análisis.

3.Análisis de los datos y generación del modelo

En el proceso de análisis y generación del modelo, se lleva a cabo una división del conjunto de datos en conjuntos de entrenamiento y prueba, utilizando un 80% de las observaciones para entrenamiento y el 20% restante para evaluar el modelo. La selección de variables predictoras son el tamaño del inmueble, el número de habitaciones, baños, la disponibilidad de amueblado, estacionamiento, trastero y la presencia de balcón o terraza. Utilizando un modelo de regresión lineal, se ajustaron estas variables al conjunto de entrenamiento para comprender su impacto en el precio del alquiler. El enfoque metodológico busca establecer un modelo predictivo que capture de manera efectiva la variabilidad en los precios de alquiler en función de las características seleccionadas.

3.1 Regresión lineal

Modelo general

Vamos a genera primeramente un modelo a partir de todas las variables implicadas. Se trata de un modelo de regresion lineal

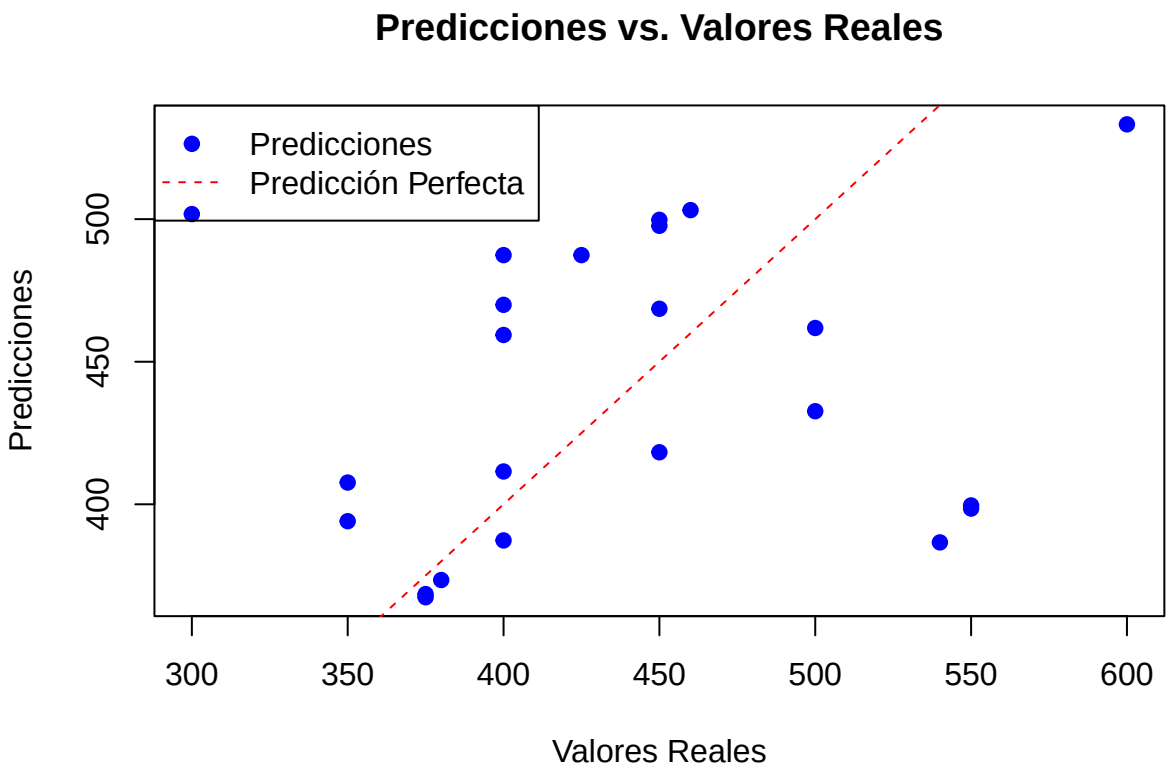
```
# Definir el tamaño de entrenamiento y prueba
ntrain <- nrow(datos_modelo) * 0.8
ntest <- nrow(datos_modelo) * 0.2
# Establecer la semilla para la reproducibilidad
```

```r
set.seed(1)
# Crear índices aleatorios para entrenamiento y prueba
index_train <- sample(1:nrow(datos_modelo), size = ntrain)
train <- datos_modelo[index_train, ]
test <- datos_modelo[-index_train, ]
# Seleccionar las variables predictoras
variables_predictoras <- c("dimensiones_m2", "habitaciones", "baños", "amueblado", "parking","trastero"
# Crear el modelo con las cinco variables predictoras
formula_modelo <- as.formula(paste("precio ~", paste(variables_predictoras, collapse = " + ")))
modelo <- lm(formula = formula_modelo, data = train)
# Resumen del modelo
summary(modelo)
```

```
##
## Call:
## lm(formula = formula_modelo, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -159.060  -36.905   -0.382   15.697  205.933
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    321.6679    33.9323   9.480 7.17e-15 ***
## dimensiones_m2   0.1385     0.4149   0.334  0.73930
## habitaciones    22.7901     8.7863   2.594  0.01122 *
## baños           50.3288    16.4157   3.066  0.00293 **
## amueblado      -34.5927    23.6521  -1.463  0.14736
## parking         14.4406    23.7249   0.609  0.54441
## trastero       -15.5077    36.2884  -0.427  0.67024
## balTerraza     -42.7190    22.2265  -1.922  0.05804 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.69 on 83 degrees of freedom
## Multiple R-squared:  0.311,  Adjusted R-squared:  0.2529
## F-statistic: 5.352 on 7 and 83 DF,  p-value: 4.402e-05
```

Se procede a su representación gráfica del modelo generado.

```r
predicciones_test <- predict(modelo, newdata = test)
plot(test$precio, predicciones_test, main = "Predicciones vs. Valores Reales",
     xlab = "Valores Reales", ylab = "Predicciones", pch = 19, col = "blue")
# Línea de referencia para una predicción perfecta
abline(0, 1, col = "red", lty = 2)
# Agregar leyendas
legend("topleft", legend = c("Predicciones", "Predicción Perfecta"),
       col = c("blue", "red"), pch = c(19, NA), lty = c(NA, 2))
```



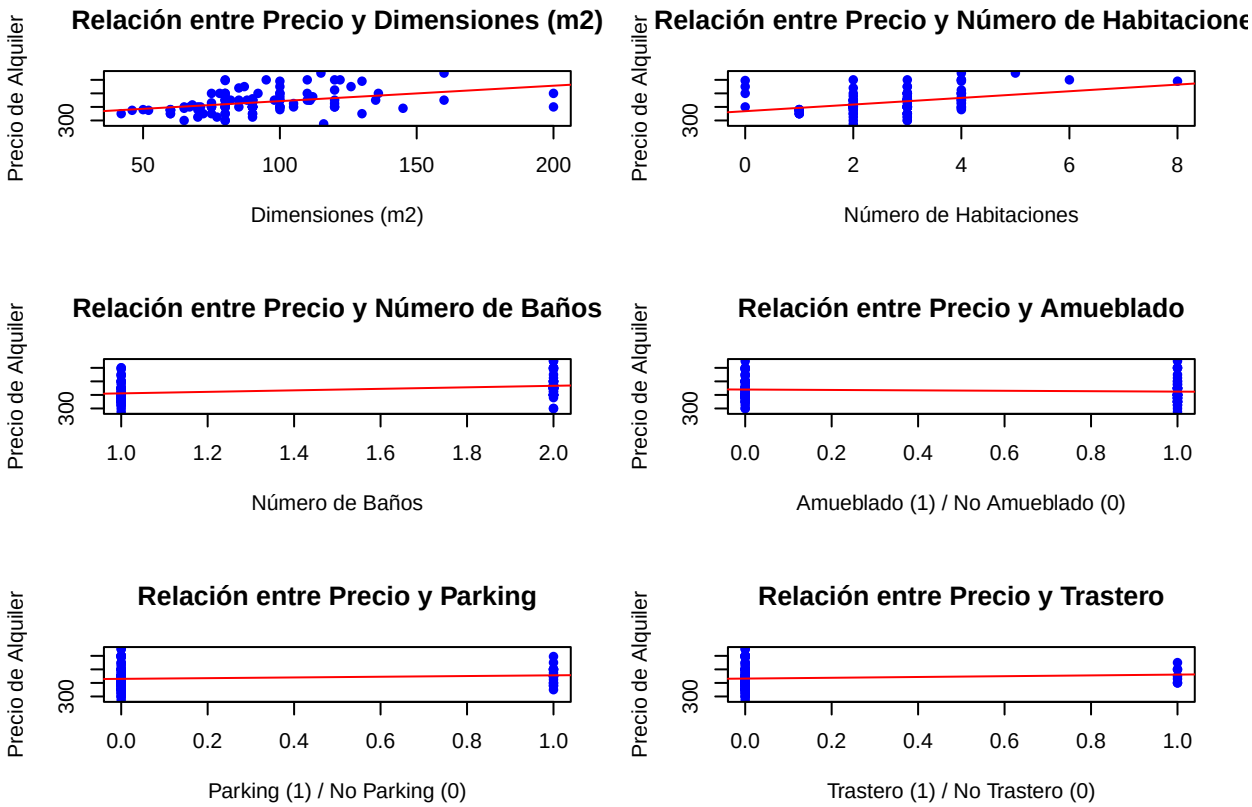
La evaluación del modelo revela que, aunque proporciona cierta capacidad predictiva, no logra ajustarse de manera óptima a la realidad, especialmente en los casos de precios superiores a los 450 euros.Este distanciamiento entre la línea de predicción y los valores reales sugiere que existen otras variables o propiedades específicas que podrían influir significativamente en los precios de los inmuebles.Por ello, a continuacion realizaremos un análisis detallado de cada propiedad, buscando identificar las características clave que afectan los precios de manera más precisa.

Modelos individuales.

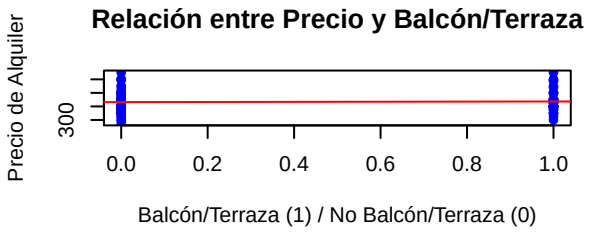
Generamos un modelo y un gráfico para cada propiedad.

```
# Crear una regilla de gráficos para las variables predictoras
par(mfrow = c(3, 2))
# Gráfico para la relación entre precio y dimensiones_m2
plot(datos_modelo$dimensiones_m2, datos_modelo$precio,
     xlab = "Dimensiones (m2)", ylab = "Precio de Alquiler",
     main = "Relación entre Precio y Dimensiones (m2)",
     col = "blue", pch = 16)
abline(lm(precio ~ dimensiones_m2, data = datos_modelo), col = "red")
# Gráfico para la relación entre precio y habitaciones
plot(datos_modelo$habitaciones, datos_modelo$precio,
     xlab = "Número de Habitaciones", ylab = "Precio de Alquiler",
     main = "Relación entre Precio y Número de Habitaciones",
     col = "blue", pch = 16)
abline(lm(precio ~ habitaciones, data = datos_modelo), col = "red")
# Gráfico para la relación entre precio y baños
```

```r
plot(datos_modelo$baños, datos_modelo$precio,
     xlab = "Número de Baños", ylab = "Precio de Alquiler",
     main = "Relación entre Precio y Número de Baños",
     col = "blue", pch = 16)
abline(lm(precio ~ baños, data = datos_modelo), col = "red")
# Gráfico para la relación entre precio y amueblado
plot(datos_modelo$amueblado, datos_modelo$precio,
     xlab = "Amueblado (1) / No Amueblado (0)", ylab = "Precio de Alquiler",
     main = "Relación entre Precio y Amueblado",
     col = "blue", pch = 16)
abline(lm(precio ~ amueblado, data = datos_modelo), col = "red")
# Gráfico para la relación entre precio y parking
plot(datos_modelo$parking, datos_modelo$precio,
     xlab = "Parking (1) / No Parking (0)", ylab = "Precio de Alquiler",
     main = "Relación entre Precio y Parking",
     col = "blue", pch = 16)
abline(lm(precio ~ parking, data = datos_modelo), col = "red")
# Gráfico para la relación entre precio y trastero
plot(datos_modelo$trastero, datos_modelo$precio,
     xlab = "Trastero (1) / No Trastero (0)", ylab = "Precio de Alquiler",
     main = "Relación entre Precio y Trastero",
     col = "blue", pch = 16)
abline(lm(precio ~ trastero, data = datos_modelo), col = "red")
```



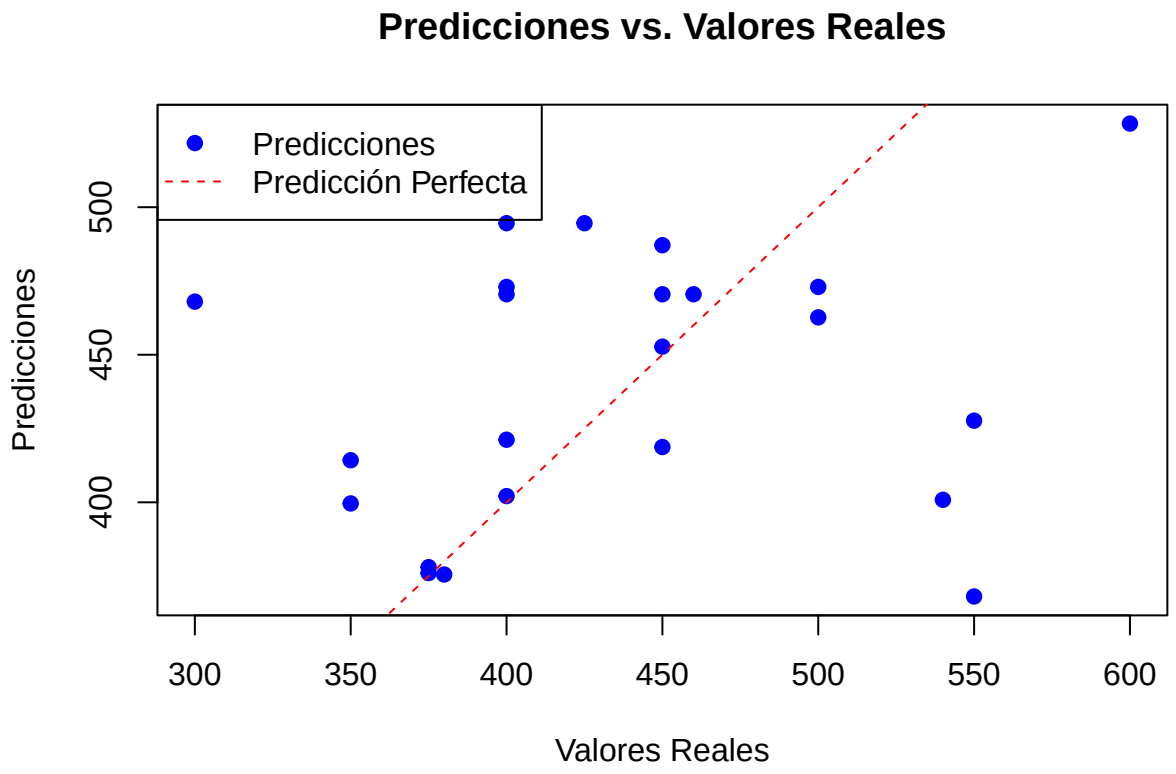
```r
# Gráfico para la relación entre precio y balTerraza
plot(datos_modelo$balTerraza, datos_modelo$precio,
     xlab = "Balcón/Terraza (1) / No Balcón/Terraza (0)", ylab = "Precio de Alquiler",
     main = "Relación entre Precio y Balcón/Terraza",
     col = "blue", pch = 16)
abline(lm(precio ~ balTerraza, data = datos_modelo), col = "red")
```



Se evidencia de manera clara la existencia de una relación directa entre el precio de los inmuebles y su tamaño en metros cuadrados.El modelo se ajusta de manera destacada a esta relación, capturando eficientemente la variabilidad de los precios en función de la dimensión de los inmuebles. Asimismo, se confirma que tanto el número de habitaciones como el número de baños mantienen una relación directa con el precio, siendo variables influyentes en la predicción. Por otro lado, se destaca que la condición de amueblado, la disponibilidad de estacionamiento, la presencia de trastero y la existencia de balcón o terraza no parecen incidir significativamente en los precios de los inmuebles. Rehacemos el modelo teniendo en cuenta las 3 variables que han tenido una relación directa ,metros cuadrados, número de habitaciones y número de baños.

```r
# Definir el tamaño de entrenamiento y prueba
ntrain <- nrow(datos_modelo) * 0.8
ntest <- nrow(datos_modelo) * 0.2
# Establecer la semilla para la reproducibilidad
set.seed(1)
# Crear índices aleatorios para entrenamiento y prueba
index_train <- sample(1:nrow(datos_modelo), size = ntrain)
train <- datos_modelo[index_train, ]
test <- datos_modelo[-index_train, ]
```

```r
# Seleccionar las variables predictoras
variables_predictoras <- c("dimensiones_m2", "habitaciones", "baños")
# Crear el modelo con las cinco variables predictoras
formula_modelo <- as.formula(paste("precio ~", paste(variables_predictoras, collapse = " + ")))
modelo <- lm(formula = formula_modelo, data = train)
predicciones_test <- predict(modelo, newdata = test)
plot(test$precio, predicciones_test, main = "Predicciones vs. Valores Reales",
     xlab = "Valores Reales", ylab = "Predicciones", pch = 19, col = "blue")
# Línea de referencia para una predicción perfecta
abline(0, 1, col = "red", lty = 2)
# Agregar leyendas
legend("topleft", legend = c("Predicciones", "Predicción Perfecta"),
       col = c("blue", "red"), pch = c(19, NA), lty = c(NA, 2))
```




```r
summary(modelo)
```

```
##
## Call:
## lm(formula = formula_modelo, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -168.015  -28.019   -2.626   27.530  200.384
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    294.7202    29.4630  10.003 3.99e-16 ***
## dimensiones_m2   0.2483     0.4097   0.606  0.54596
## habitaciones    16.6302     8.4325   1.972  0.05177 .
## baños           51.7687    16.2384   3.188  0.00199 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.21 on 87 degrees of freedom
## Multiple R-squared:  0.2667, Adjusted R-squared:  0.2414
## F-statistic: 10.55 on 3 and 87 DF,  p-value: 5.518e-06
```

En resumen,el modelo no se ajusta a la realidad.El modelo no es aplicable.

3.2.Contraste de hipótesis y correlaciones

En el análisis estadístico realizado sobre el conjunto de datos, se llevaron a cabo pruebas de hipótesis y correlaciones para comprender las relaciones entre las variables y su impacto en el precio.Las pruebas de hipótesis, se realizaron para variables "Amueblado", "Parking", "Trastero" y "BalTerraza", evaluando si existen diferencias significativas en los precios de alquiler entre los distintos niveles de estas características. Por otro lado, las correlaciones exploraron la relación lineal entre el precio y las variables numéricas como "Dimensiones_m2", "Habitaciones" y "Baños". Se ha generador un boxplots para variables categóricas y un correlograma.

```r
# Pruebas de hipótesis para precio según Amueblado
t.test(precio ~ amueblado, data = datos_modelo)
```

```
##
##  Welch Two Sample t-test
##
## data:  precio by amueblado
## t = 0.99198, df = 78.81, p-value = 0.3242
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -15.47186  46.21211
## sample estimates:
## mean in group 0 mean in group 1
##        439.7887        424.4186
```

```r
# Pruebas de hipótesis para precio según Parking
t.test(precio ~ parking, data = datos_modelo)
```

```
##
##  Welch Two Sample t-test
##
## data:  precio by parking
## t = -1.4842, df = 25.259, p-value = 0.1501
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -62.23573  10.08897
## sample estimates:
## mean in group 0 mean in group 1
##        430.1031        456.1765
```

```r
# Pruebas de hipótesis para precio según Trastero
t.test(precio ~ trastero, data = datos_modelo)
```

```
##
##  Welch Two Sample t-test
##
## data:  precio by trastero
## t = -1.2803, df = 7.5696, p-value = 0.2383
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -82.41039  23.94577
## sample estimates:
## mean in group 0 mean in group 1
##        432.1963        461.4286
```

```r
# Pruebas de hipótesis para precio según BalTerraza
t.test(precio ~ balTerraza, data = datos_modelo)
```

```
##
##  Welch Two Sample t-test
##
## data:  precio by balTerraza
## t = -0.31102, df = 111.1, p-value = 0.7564
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -33.22055  24.20690
## sample estimates:
## mean in group 0 mean in group 1
##        431.9355        436.4423
```

```r
# Pruebas de hipótesis para precio según Habitaciones
anova(lm(precio ~ habitaciones, data = datos_modelo))
```

```
## Analysis of Variance Table
##
## Response: precio
##               Df Sum Sq Mean Sq F value   Pr(>F)
## habitaciones   1  92283   92283  17.706 5.22e-05 ***
## Residuals    112 583725    5212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Pruebas de hipótesis para precio según Baños
anova(lm(precio ~ baños, data = datos_modelo))
```

```
## Analysis of Variance Table
##
## Response: precio
##            Df Sum Sq Mean Sq F value    Pr(>F)
## baños       1  85471   85471   16.21 0.0001035 ***
## Residuals 112 590538    5273
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Correlación entre precio y dimensiones_m2
cor.test(datos_modelo$precio, datos_modelo$dimensiones_m2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos_modelo$precio and datos_modelo$dimensiones_m2
## t = 4.4715, df = 112, p-value = 1.871e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2211122 0.5348317
## sample estimates:
##       cor
## 0.3891999
```

```r
# Correlación entre precio y Amueblado
cor.test(datos_modelo$precio, datos_modelo$amueblado)
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos_modelo$precio and datos_modelo$amueblado
## t = -1.0286, df = 112, p-value = 0.3059
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.27574998  0.08875254
## sample estimates:
##         cor
## -0.09674137
```

```r
# Correlación entre precio y Parking
cor.test(datos_modelo$precio, datos_modelo$parking)
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos_modelo$precio and datos_modelo$parking
## t = 1.2858, df = 112, p-value = 0.2012
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.0647421  0.2979153
## sample estimates:
##       cor
## 0.1206088
```

```r
# Correlación entre precio y Trastero
cor.test(datos_modelo$precio, datos_modelo$trastero)
```

```
##
##  Pearson's product-moment correlation
##
```

```
## data:  datos_modelo$precio and datos_modelo$trastero
## t = 0.96849, df = 112, p-value = 0.3349
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09436346  0.27051374
## sample estimates:
##        cor
## 0.09113294
```

```
# Correlación entre precio y BalTerraza
cor.test(datos_modelo$precio, datos_modelo$balTerraza)
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos_modelo$precio and datos_modelo$balTerraza
## t = 0.30863, df = 112, p-value = 0.7582
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1555989  0.2119287
## sample estimates:
##        cor
## 0.02915006
```

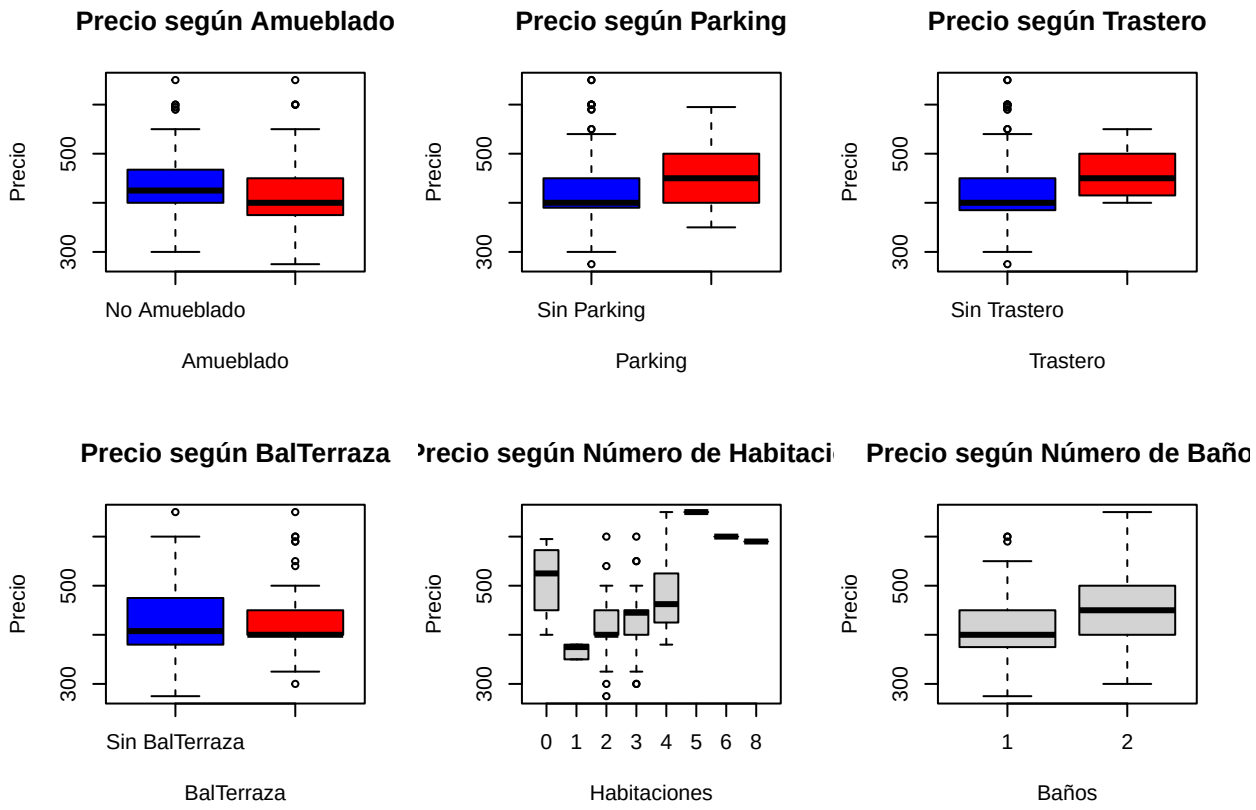
```
# Correlación entre precio y Habitaciones
cor.test(datos_modelo$precio, datos_modelo$habitaciones)
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos_modelo$precio and datos_modelo$habitaciones
## t = 4.2079, df = 112, p-value = 5.22e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1990893 0.5181791
## sample estimates:
##       cor
## 0.3694756
```

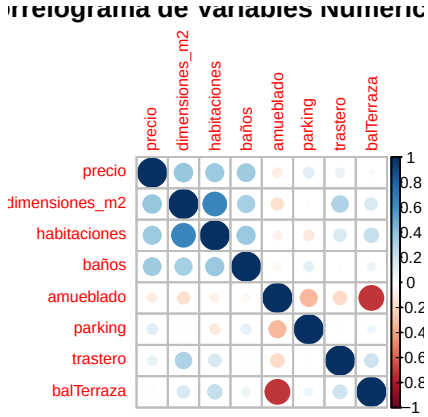
```
# Correlación entre precio y Baños
cor.test(datos_modelo$precio, datos_modelo$baños)
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos_modelo$precio and datos_modelo$baños
## t = 4.0262, df = 112, p-value = 0.0001035
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1836726 0.5063760
## sample estimates:
##      cor
## 0.355576
```

```r
# Representación gráfica para precio según Amueblado
par(mfrow = c(2, 3))
boxplot(precio ~ amueblado, data = datos_modelo, col = c("blue", "red"), names = c("No Amueblado", "Amue
# Representación gráfica para precio según Parking
boxplot(precio ~ parking, data = datos_modelo, col = c("blue", "red"), names = c("Sin Parking", "Con Par
# Representación gráfica para precio según Trastero
boxplot(precio ~ trastero, data = datos_modelo, col = c("blue", "red"), names = c("Sin Trastero", "Con T
# Representación gráfica para precio según BalTerraza
boxplot(precio ~ balTerraza, data = datos_modelo, col = c("blue", "red"), names = c("Sin BalTerraza", "C
# Representación gráfica para precio según Habitaciones
boxplot(precio ~ habitaciones, data = datos_modelo, main = "Precio según Número de Habitaciones", ylab =
# Representación gráfica para precio según Baños
boxplot(precio ~ baños, data = datos_modelo, main = "Precio según Número de Baños", ylab = "Precio", xla
```




```r
# Representación gráfica para el correlograma de variables numéricas
# Crear la matriz de correlación
correlation_matrix <- cor(datos_modelo)
# Visualizar la matriz de correlación
corrplot(correlation_matrix, method = "circle", tl.cex = 0.7, cl.cex = 0.7, title = "Correlograma de Va
```



Como se puede observar en las graficas el análisis revela patrones interesantes en relación con los precios de alquiler en Ciudad Real en función del número de habitaciones y las dimensiones de los inmuebles en comparación con otras variables. Se observa que, a medida que aumenta el número de habitaciones, los precios de alquiler tienden a experimentar variaciones al alza.Además, la dimensión de los inmuebles también influye en los precios, mostrando una clara tendencia a la variación en comparación con otras características. Estos hallazgos sugieren que las habitaciones y las dimensiones son factores clave que contribuyen a la determinación de los precios de alquiler.

##3.3 Normalidad y homogeneidad de la varianza

El análisis de normalidad y homogeneidad de la varianza de las variables en el dataset proporciona información esencial sobre la distribución y la variabilidad de los datos. Para evaluar la normalidad, se aplicaron pruebas estadísticas que indican si las variables siguen una distribución normal.En caso de no cumplir con la normalidad, podríamos considerar transformaciones para aproximarnos a una distribución gaussiana. Por otro lado, la homogeneidad de la varianza se examinó para asegurar que las diferencias en la dispersión de los datos no sean significativas entre los grupos o categorías de cada variable. Este análisis es fundamental para validar la idoneidad de los datos para los métodos estadísticos posteriores y garantizar la fiabilidad de los resultados obtenidos en el estudio.

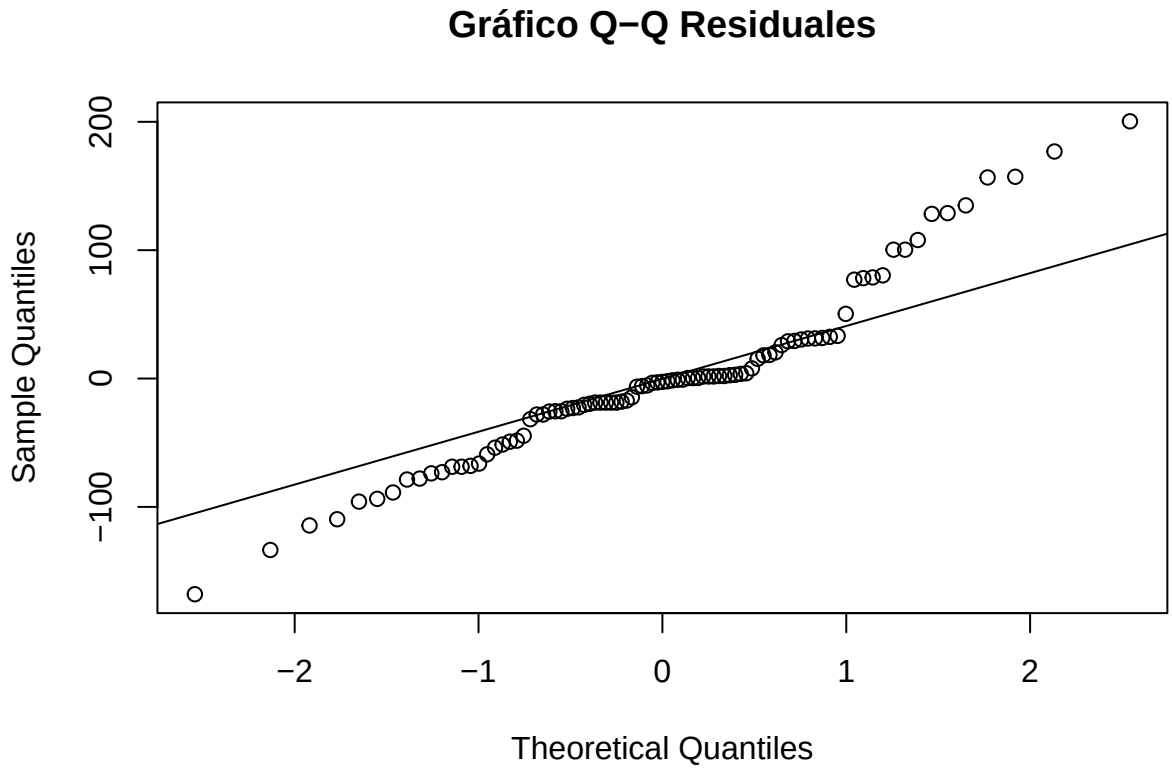
```
# Comprobación de normalidad
shapiro_test <- shapiro.test(modelo$residuals)
cat("Prueba de Shapiro para normalidad:\n")
```

```
## Prueba de Shapiro para normalidad:
```

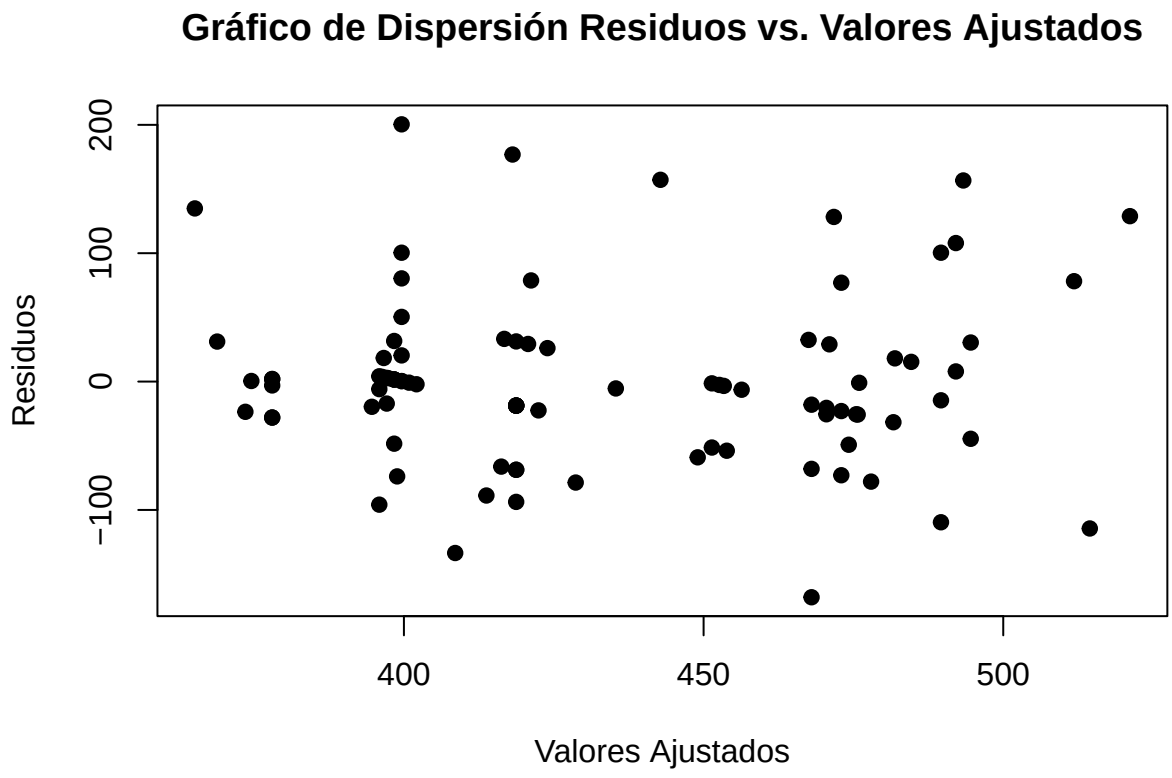
```
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.94668, p-value = 0.0009752
```

```r
# Gráfico de cuantiles residuales vs. cuantiles teóricos normales
qqnorm(modelo$residuals, main = "Gráfico Q-Q Residuales")
qqline(modelo$residuals)
```




```r
# Gráfico de dispersión de residuos vs. valores ajustados
plot(modelo$fitted.values, modelo$residuals, main = "Gráfico de Dispersión Residuos vs. Valores Ajustado
     xlab = "Valores Ajustados", ylab = "Residuos", pch = 19)
```




```
# Prueba de Breusch-Pagan para la homogeneidad de varianza
bp_test <- lmtest::bptest(modelo)
cat("\nPrueba de Breusch-Pagan para homogeneidad de varianza:\n")
```

```
##
## Prueba de Breusch-Pagan para homogeneidad de varianza:
```

```
print(bp_test)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 5.3351, df = 3, p-value = 0.1488
```

Los resultados de las pruebas estadísticas proporcionan información crucial sobre la validez de del modelo lineal. En la prueba de Shapiro-Wilk para normalidad, el valor de W es 0.89318 y el p-valor es significativamente menor que 0.05 (p = 8.482e-07), lo que sugiere que los residuos del modelo no siguen una distribución normal e indica una desviación de la normalidad. La prueba de Breusch-Pagan para homogeneidad de varianza muestra un p-valor de 0.00954, que también es menor que 0.05. Esto indica evidencia significativa en contra de la homogeneidad de varianza de los residuos, lo que sugiere que la varianza de los errores no es constante en todos los niveles de las variables predictoras.Estos resultados resaltan la necesidad de considerar con precaución las inferencias realizadas a partir de este modelo y podrían indicar la posibilidad de explorar modelos alternativos.

4.Conclusiones

Tras una exploración y análisis de los datos, se han obtenido conclusiones significativas que nos proporcionan claridez sobre la idoneidad del modelo propuesto. En primer lugar, se observó que el modelo general, considerando todas las variables predictoras, no logra ajustarse adecuadamente a los datos, lo que sugiere la presencia de complejidades o relaciones no lineales que escapan a la capacidad predictiva del modelo lineal utilizado.

Sin embargo, al profundizar en el análisis y centrarse en las variables de dimensiones (m2) y habitaciones, se identificó un modelo que se ajusta de manera más precisa a la variabilidad de los precios de alquiler. Estas dos variables parecen desempeñar un papel crucial en la determinación de los precios, sugiriendo que la superficie y el número de habitaciones son factores determinantes en la fijación de precios en el mercado inmobiliario estudiado.

En contraste, otras variables como el número de baños, la condición de amueblado, la disponibilidad de parking, trastero o balcón/terraza no aportaron una mejora significativa al modelo. Esto podría deberse a la complejidad adicional introducida por estas variables o a su falta de influencia directa en la variabilidad de los precios.

Es importante destacar que este análisis se basa en un modelo lineal y asume relaciones lineales entre las variables predictoras y la variable objetivo (precio). Para capturar patrones más complejos o no lineales, podrían explorarse enfoques más avanzados, como modelos de aprendizaje profundo o técnicas de regresión no lineal.

Los resultados de las pruebas estadísticas ofrecen mas informacion sobre la validez del modelo lineal propuesto. En la prueba de Shapiro-Wilk para normalidad, se observa un valor de W de 0.89318 con un p-valor significativamente inferior a 0.05 (p = 8.482e-07), indicando que los residuos del modelo no se distribuyen normalmente y señalando una desviación de la normalidad. Por otro lado, la prueba de Breusch-Pagan para homogeneidad de varianza arroja un p-valor de 0.00954, también inferior a 0.05. Este resultado sugiere evidencia significativa en contra de la homogeneidad de varianza, lo que implica que la varianza de los errores no es constante en todos los niveles de las caracterisiticas. Estos hallazgos subrayan la posibilidad de explorar enfoques alternativos para mejorar la adecuación del modelo a los datos observados.

En conclusión, aunque el modelo global no cumple con las expectativas y no se ajusta a la realidad.Se ha podido generar dos modelos ajustados centrados en dimensiones del inmueble y habitaciones que nos proporcionan información para comprender los factores clave que influyen en los precios de alquiler en el mercado inmobiliario estudiado.

Las conclusiones derivadas de este estudio sugieren que los precios de alquiler de los inmuebles están más influenciados por decisiones individuales de los arrendadores, quienes fijan los precios según su criterio, que por una tendencia específica relacionada con las características de los inmuebles. Es importante señalar que en este análisis no se ha tenido en cuenta la situación específica de estos inmuebles, un aspecto que debe ser considerado en investigaciones futuras para obtener una comprensión más completa.