

UNIVERSIDAD DE LA REPÚBLICA  
Facultad de Ciencias Económicas y de Administración  
Licenciatura en Estadística

**Análisis explicativo del precio de apartamentos publicados en Airbnb de  
Barcelona**

**Viscailuz, Luciana Miranda, Germán  
Junio 2024**

**Trabajo final de Modelos Lineales**

# Índice

<b>Introducción</b>	<b>3</b>
Los datos . . . . .	4
<b>Análisis exploratorio</b>	<b>5</b>
<b>Metodología</b>	<b>7</b>
Selección de variables . . . . .	7
Diagnóstico . . . . .	9
Multicolinealidad . . . . .	10
Linealidad . . . . .	10
Homoscedasticidad . . . . .	14
Normalidad . . . . .	14
Atípicos . . . . .	16
Corrección del modelo . . . . .	17
<b>Resultados</b>	<b>20</b>
<b>Conclusiones</b>	<b>21</b>
<b>Bibliografía</b>	<b>21</b>

# Introducción

Airbnb es una compañía dedicada a la oferta de alojamientos de carácter vacacional en muchos de los países del mundo. Esta funciona a partir de un programa digital donde los anfitriones pueden publicar sus propiedades para que los clientes puedan verlas y elegir el alojamiento que más se adapte a sus necesidades.

En este proyecto se trabajó con algunas de las propiedades publicadas en esta plataforma en la ciudad Barcelona, España. Inicialmente, nuestra motivación del proyecto fue poder estimar el precio (en euros) de diferentes apartamentos según las características de cada uno. Para realizar lo mencionado se trabajó con los datos de Airbnb Barcelona, donde se tenía el registro de más de 16.000 apartamentos de dicha ciudad.

La información con la que se contaba era buena pero excesiva, lo que llevó a que algunos datos fueran redundantes, por lo cual una parte importante de este proyecto fue la inicial, donde se realizó una limpieza de datos para así disponer de información que permita realizar una buena estimación e interpretación de los datos.

Finalmente, el trabajo e interpretación de los datos, tanto sobre como actúan entre si y sus diversos efectos sobre la variable de interés fueron los que guiaron y generaron el interés en este proyecto provocando que el paso a paso sea tan importante como el resultado final.

```
df = read_excel(here("airbnb_barcelona_v2.xlsx"))
```

```
## Warning: Expecting numeric in D10008 / R10008C4: got '08001'
```

```
## Warning: Expecting numeric in D11270 / R11270C4: got 'barcelona'
```

```
## Warning: Expecting numeric in D11553 / R11553C4: got '13-08008'
```

```
## Warning: Expecting numeric in D11554 / R11554C4: got '13-08008'
```

```
summary(df)
```

```
##           id           host_id           barrio           cod_postal
## Min.      : 18666   Min.      : 10704   Length:16761   Min.      :    0
## 1st Qu.:11448633   1st Qu.: 7612142   Class :character   1st Qu.:  8004
## Median :22146039   Median : 45072553   Mode  :character   Median :  8012
## Mean      :20880757   Mean      : 86673374   Mean      :  8267
## 3rd Qu.:31623085   3rd Qu.:158838753   3rd Qu.:  8022
## Max.      :36582760   Max.      :274862556   Max.      :4008009
##                                     NA's      :506
##           latitud           longitud           tipo_habitacion           personas
## Min.      :41.35   Min.      :2.105   Length:16761   Min.      : 1.000
## 1st Qu.:41.38   1st Qu.:2.157   Class :character   1st Qu.:  2.000
## Median :41.39   Median :2.168   Mode  :character   Median :  2.000
## Mean      :41.39   Mean      :2.168   Mean      :  3.358
## 3rd Qu.:41.40   3rd Qu.:2.178   3rd Qu.:  4.000
## Max.      :41.46   Max.      :2.222   Max.      :18.000
##
##           banios           habitaciones           camas           precio_euros
## Min.      :0.000   Min.      : 0.000   Min.      : 0.000   Min.      :    7
## 1st Qu.:1.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:  40
## Median :1.000   Median : 1.000   Median : 2.000   Median :  63
## Mean      :1.288   Mean      : 1.586   Mean      : 2.239   Mean      :  92
```

```

## 3rd Qu.:1.500 3rd Qu.: 2.000 3rd Qu.: 3.000 3rd Qu.: 107
## Max. :8.000 Max. :12.000 Max. :30.000 Max. :1000
## NA's :9 NA's :3 NA's :16
## estancia_min puntuacion Internet TV
## Min. : 1.000 Min. : 20.00 Min. :0.0000 Min. :0.0000
## 1st Qu.: 1.000 1st Qu.: 88.00 1st Qu.:0.0000 1st Qu.:0.0000
## Median : 2.000 Median : 93.00 Median :0.0000 Median :1.0000
## Mean : 8.509 Mean : 90.98 Mean :0.2149 Mean :0.6973
## 3rd Qu.: 4.000 3rd Qu.: 97.00 3rd Qu.:0.0000 3rd Qu.:1.0000
## Max. :900.000 Max. :100.00 Max. :1.0000 Max. :1.0000
## NA's :3891
## Wifi Air_conditioning Elevator Breakfast
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.00000
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000
## Median :1.0000 Median :1.0000 Median :1.0000 Median :0.00000
## Mean :0.7383 Mean :0.5707 Mean :0.6167 Mean :0.05913
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.00000
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.00000
##
## Pets_allowed Cable_TV Pool Patio_or_balcony
## Min. :0.0000 Min. :0.00000 Min. :0.00000 Min. :0.0000
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.0000
## Median :0.0000 Median :0.00000 Median :0.00000 Median :0.0000
## Mean :0.1157 Mean :0.09898 Mean :0.01873 Mean :0.2261
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.0000
## Max. :1.0000 Max. :1.00000 Max. :1.00000 Max. :1.0000
##
## check_in_24_hs Smart_lock
## Min. :0.0000 Min. :0.000000
## 1st Qu.:0.0000 1st Qu.:0.000000
## Median :0.0000 Median :0.000000
## Mean :0.1107 Mean :0.007458
## 3rd Qu.:0.0000 3rd Qu.:0.000000
## Max. :1.0000 Max. :1.000000
##

```

## Los datos

Como se mencionó anteriormente, se disponía de la información de 16.761 apartamentos de Barcelona, donde se nombraban las características que los huéspedes toman en cuenta al momento de elegir su hospedaje y por lo tanto podrían llegar a incidir en su precio, entre estas se destacaba, ubicación, cantidad de camas y baños, cuantas personas se aceptaban, entre otras.

Habían variables cuantitativas pero la mayoría eran cualitativas, e incluso se pasaron a factor algunas de las cuantitativas para su mejor interpretación.

La información de código postal y barrio se decidió resumirla en una variable llamada distrito la cual agrupó los 73 barrios de Barcelona en 10 distritos.

Se decidió prescindir de la latitud y longitud de cada apartamento como de algunas variables que se encontraban dentro de la variable amenities, tomando en cuenta finalmente las diez más #importantes.

Se definieron dos nuevas variables, “grupo\_habitacion” y “grupo\_banios”, donde se agruparon la cantidad de habitaciones y baños respectivamente, para así disminuir la cantidad de categorías de cada variable.

Del total de observaciones se operó con 12.848 debido a que las restantes contaban con datos faltantes.

Por último se decidió trabajar con la transformación logarítmica de la variable respuesta, para así poder solucionar varios problemas relacionados con el diagnóstico que se verá posteriormente.

Finalizada la limpieza y organización de datos se pudo comenzar a trabajar con ellos.

## Análisis exploratorio

Como parte de la estadística descriptiva se crearon gráficos donde se relacionan algunas de las variables explicativas con la variable de respuesta (precio en euros), estos graficos permiten obtener interpretaciones de las diferentes relaciones, pero es muy importante destacar que las interpretaciones obtenidas son parciales, debido a que a diferencia del modelo, en cada uno de los gráficos se representa el efecto de una variable sin tomar en cuenta las demás.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

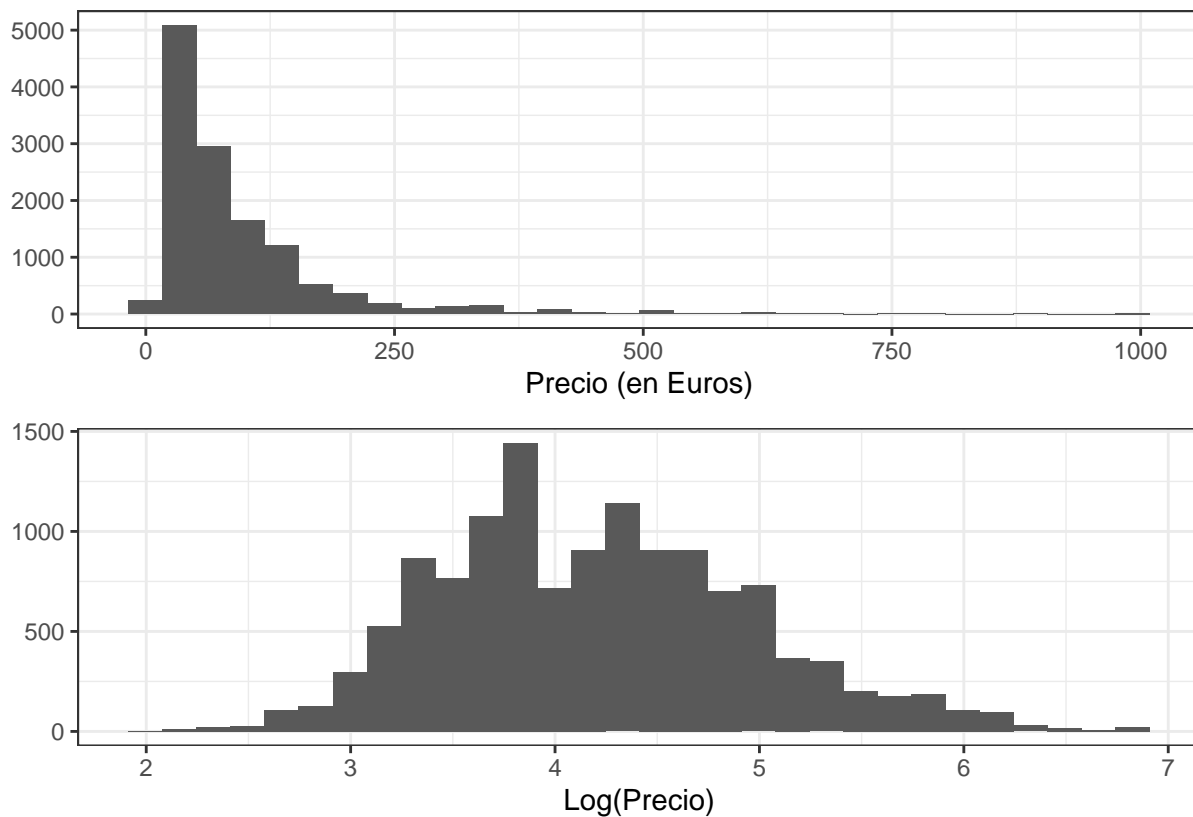


Figure 1: Histograma de la variable de respuesta Precio

En el presente gráfico se puede observar la distribución de los datos y la relación entre los distritos y  $\log(\text{Precio})$ . La mayoría de las cajas se encuentran superpuestas entre sí, por lo que no se podría afirmar cual es el distrito que lleva a apartamentos con un  $\log(\text{Precio})$  más alto, pero si se pueden interpretar algunas diferencias entre los distritos, como que los apartamentos del distrito L'Eixample tienen un  $\log(\text{Precio})$  mayor que el distrito de Nou Barris. También se resalta la presencia de observaciones atípicas, que tienden a tener un  $\log(\text{Precio})$  más alto que las mayoría.

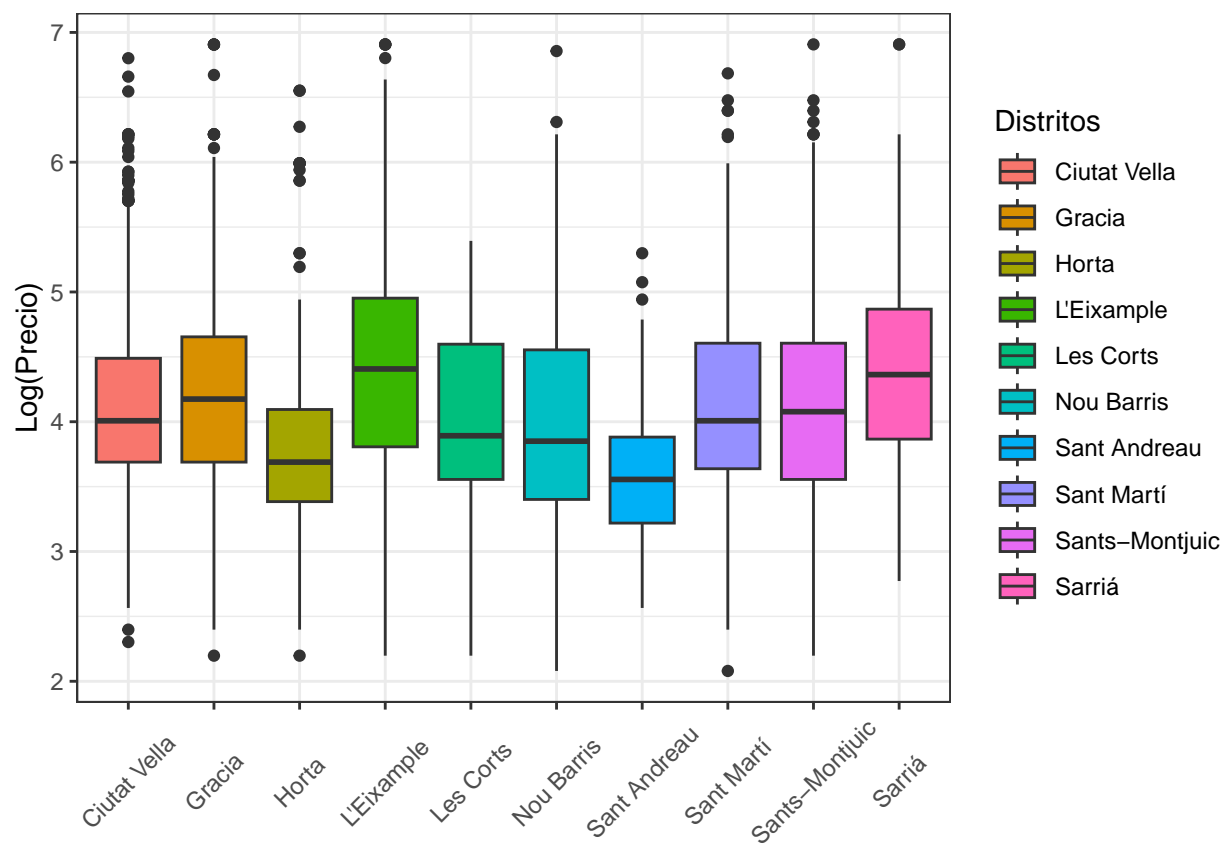
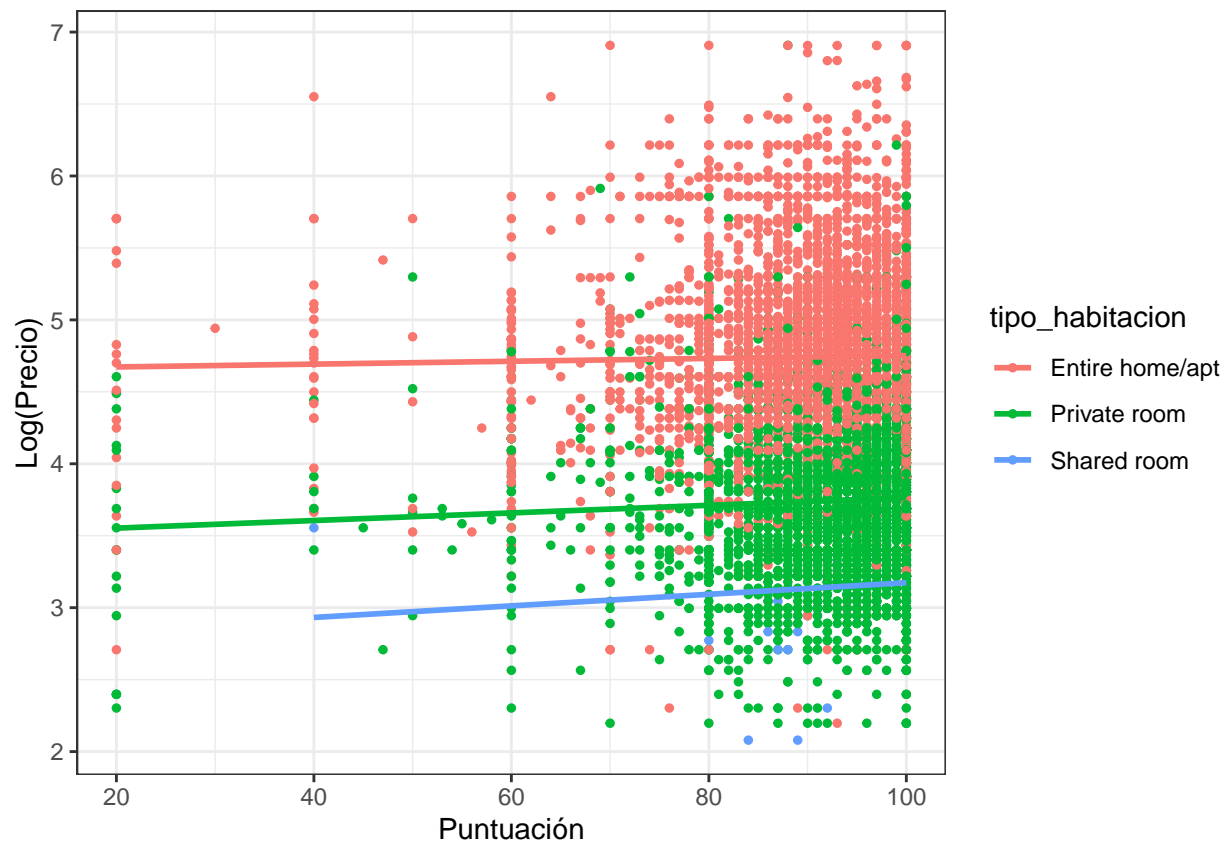


Figure 2: Dispersión de Precio (en euros) por Distrito



En el gráfico - se muestra a partir de un gráfico de dispersión la relación entre las diferentes combinaciones de las variables explicativas, puntuación y tipo de habitación, con la variable respuesta,  $\log(\text{Precio})$ . Lo primero a destacar es que los puntos tienden a estar acumulados en la parte derecha del gráfico, teniendo en su mayoría una puntuación mayor a 60, por lo que permite interpretar que no existe ninguna relación entre puntuación y  $\log(\text{precio})$ . Por otro lado también se puede observar que existe una relación muy marcada entre el tipo de habitación y la variable respuesta, los apartamentos enteros se encuentran en la parte derecha y superior del gráfico, es decir, que están relacionados con precios más altos, mientras que el resto se encuentran en el extremo inferior.

## Metodología

### Selección de variables

A partir de las diferentes combinaciones de las variables explicativas se puede llegar a una gran cantidad de modelos que busquen predecir el precio, cada uno de estos llevará a una predicción diferente de el mismo. Uno de los problemas centrales es encontrar cual de todos estos modelos cumple mejor su objetivo. En este proyecto finalmente se decidió que 15 de - variables explicativas formaran parte del modelo inicial.

```
((k=ncol(airbnb_barcelona)-1 modelos_posibles=2**k-1
```

hay 16383 modelos posibles, vamos a aplicar los procedimientos de hipotesis para llegar al mejor modelo)))

```
# Definimos la primer versión del modelo
```

```
mod0 <- lm(log(precio_euros) ~ distritos + tipo_habitacion+personas+ grupo_banios+ grupo_habitacion+es
```

```
Anova(mod0)
```

```
## Warning in printHypothesis(L, rhs, names(b)): one or more coefficients in the hypothesis include
## arithmetic operators in their names;
## the printed representation of the hypothesis will be omitted
```

```
## Anova Table (Type II tests)
```

```
##
```

```
## Response: log(precio_euros)
```

	Sum Sq	Df	F value	Pr(>F)
## distritos	66.88	9	31.9956	< 2.2e-16 ***
## tipo_habitacion	215.66	2	464.2744	< 2.2e-16 ***
## personas	212.28	1	914.0006	< 2.2e-16 ***
## grupo_banios	2.07	1	8.9077	0.0028473 **
## grupo_habitacion	1.98	1	8.5216	0.0035183 **
## estancia_min	95.97	1	413.2027	< 2.2e-16 ***
## puntuacion	4.22	1	18.1589	2.053e-05 ***
## TV	4.55	1	19.5917	9.702e-06 ***
## Wifi	5.58	1	24.0231	9.685e-07 ***
## Air_conditioning	35.26	1	151.8335	< 2.2e-16 ***
## Elevator	3.44	1	14.8165	0.0001193 ***
## Breakfast	4.31	1	18.5515	1.671e-05 ***
## Pets_allowed	1.96	1	8.4325	0.0036947 **
## Patio_or_balcony	2.23	1	9.6145	0.0019365 **
## check_in_24_hs	12.19	1	52.4982	4.662e-13 ***
## Residuals	2083.13	8969		

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod0)
```

```
##
```

```
## Call:
```

```
## lm(formula = log(precio_euros) ~ distritos + tipo_habitacion +
## personas + grupo_banios + grupo_habitacion + estancia_min +
## puntuacion + TV + Wifi + Air_conditioning + Elevator + Breakfast +
## Pets_allowed + Patio_or_balcony + check_in_24_hs, data = df_final)
##
```

```
## Residuals:
```

	Min	1Q	Median	3Q	Max
##	-3.0786	-0.3052	-0.0162	0.2809	3.3202

```
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	3.9217314	0.0700532	55.982	< 2e-16 ***
## distritosGracia	-0.1174924	0.0209170	-5.617	2.00e-08 ***
## distritosHorta	-0.2394162	0.0316167	-7.572	4.02e-14 ***
## distritosL'Eixample	-0.0106953	0.0149540	-0.715	0.474495
## distritosLes Corts	-0.2307730	0.0390481	-5.910	3.55e-09 ***
## distritosNou Barris	-0.1874278	0.0223157	-8.399	< 2e-16 ***
## distritosSant Andreu	-0.3569627	0.0426072	-8.378	< 2e-16 ***
## distritosSant Martí	-0.0921952	0.0256528	-3.594	0.000327 ***



```
## distritosSants-Montjuic      -0.1765240  0.0185342  -9.524 < 2e-16 ***
## distritosSarriá              -0.0120884  0.0305449  -0.396 0.692294
## tipo_habitacionPrivate room -0.5100940  0.0172527 -29.566 < 2e-16 ***
## tipo_habitacionShared room  -1.1024064  0.0942326 -11.699 < 2e-16 ***
## personas                     0.1259719  0.0041668  30.232 < 2e-16 ***
## grupo_baniosPocos            -0.1172766  0.0392941  -2.985 0.002847 **
## grupo_habitacionGrande        0.0535848  0.0183561   2.919 0.003518 **
## estancia_min                 -0.0074194  0.0003650 -20.327 < 2e-16 ***
## puntuacion                   0.0023214  0.0005448   4.261 2.05e-05 ***
## TV1                          0.0986626  0.0222903   4.426 9.70e-06 ***
## Wifi1                       -0.1106030  0.0225659  -4.901 9.69e-07 ***
## Air_conditioning1            0.1619744  0.0131451  12.322 < 2e-16 ***
## Elevator1                    0.0442603  0.0114985   3.849 0.000119 ***
## Breakfast1                   0.0959173  0.0222694   4.307 1.67e-05 ***
## Pets_allowed1                0.0477139  0.0164311   2.904 0.003695 **
## Patio_or_balcony1            0.0372403  0.0120102   3.101 0.001936 **
## check_in_24_hs1              0.1160455  0.0160161   7.246 4.66e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4819 on 8969 degrees of freedom
## Multiple R-squared:  0.5947, Adjusted R-squared:  0.5936
## F-statistic: 548.4 on 24 and 8969 DF, p-value: < 2.2e-16
```

## Diagnóstico

Luego de la limpieza de datos y estadística descriptiva se comenzó la etapa de diagnóstico, esta etapa es imprescindible debido a que el cumplimiento de todos los supuestos sobre el modelo es el que permite afirmar que las inferencias realizadas son validas. Entre estos supuestos se encuentran:

Multicolinealidad: en esta prueba se busca que ninguna de las columnas de la matriz X sea “casi” combinación lineal de las demás. Cuando esto si sucede el número de condición aumenta, lo que lleva finalmente a que la inversa de  $X'X$  sea inestable. Esta inestabilidad es la que finalmente se busca evitar.

Linealidad: este supuesto se basa en la linealidad de la variable  $\log(\text{Precio})$  y cada una de las variables explicativas, si en el modelo no hay linealidad se presentarán problemas de correlación entre los residuos y variabilidad de los mismos. Para verificar el cumplimiento de este supuesto se realizó un análisis gráfico entre los  $\hat{Y}$  y  $\hat{\epsilon}$ , en el cual se busca no encontrar patrones.

Homoscedasticidad: se busca que el modelo sea homoscedastico, es decir, que la varianza de todos los residuos sea constante, se va a entender esto como que la varianza no depende de ninguna de las variables explicativas. Esta prueba se suele ver mediante gráficos que relacionan cada variable explicativa con la de respuesta y a partir de una prueba de hipótesis donde se busca no rechazar la hipótesis nula. #EL PROFE DIJO QUE AL SER TANTAS VARIABLES (TODO) SIEMPRE VA A SER SIGNIFICATIVO, PODEMOS AGREGARLO POR ESCRITO

Normalidad: se refiere a que los residuos deben tener una distribución normal, este supuesto es muy importante debido a que es el que luego permite realizar inferencias. Sin embargo en los modelos donde el tamaño de muestra es grande, como en este caso, la falta de normalidad de los residuos no generan repercusiones. #AL SER TANTAS OBSERVACIONES SON “ROBUSTAS” A LA NORMALIDAD (ALGO ASÍ DIJO EL PROFE)

En las siguientes líneas del script se pusieron a prueba cada uno de los supuestos antes mencionados.

## Multicolinealidad

[1] 1711.793 distritosGracia distritosHorta 1.288725 1.115022 distritosL'Eixample distritosLes Corts 1.987565 1.082300 distritosNou Barris distritosSant Andreu 1.267226 1.069657 distritosSant Martí distritosSants-Montjuic 1.195899 1.386380 distritosSarrià tipo\_habitacionPrivate room 1.149912 2.873404 tipo\_habitacionShared room personas 1.029174 3.203203 grupo\_baniosPocos grupo\_habitacionGrande 1.134351 1.850647 estancia\_min puntuacion 1.094684 1.042523 TV1 Wifi1 4.069130 3.716428 Air\_conditioning1 Elevator1 1.635899 1.197727 Breakfast1 Pets\_allowed1 1.048010 1.023216 Patio\_or\_balcony1 check\_in\_24\_hs1 1.037473 1.079611

Se comenzó evaluando el supuesto de multicolinealidad, para esto se calculó el número de condición, el cual dió 1711.793, es decir, existían problemas de multicolinealidad. Luego se calculó el VIF para cada una de las variables explicativas, todas tenían un VIF mayor a 1 pero menor a 5, por lo cual, no se pudo determinar cual es la variable que generó el problema del cumplimiento de este supuesto. De igual forma, esto no provocó grandes problemas debido a que se trabajó con un modelo con muchas variables cualitativas, lo cual hace que la evaluación de este supuesto no tenga mucho sentido.

## Linellidad

Luego se siguió con el supuesto de linealidad, donde se realizó el grafico entre residuos y predichos.

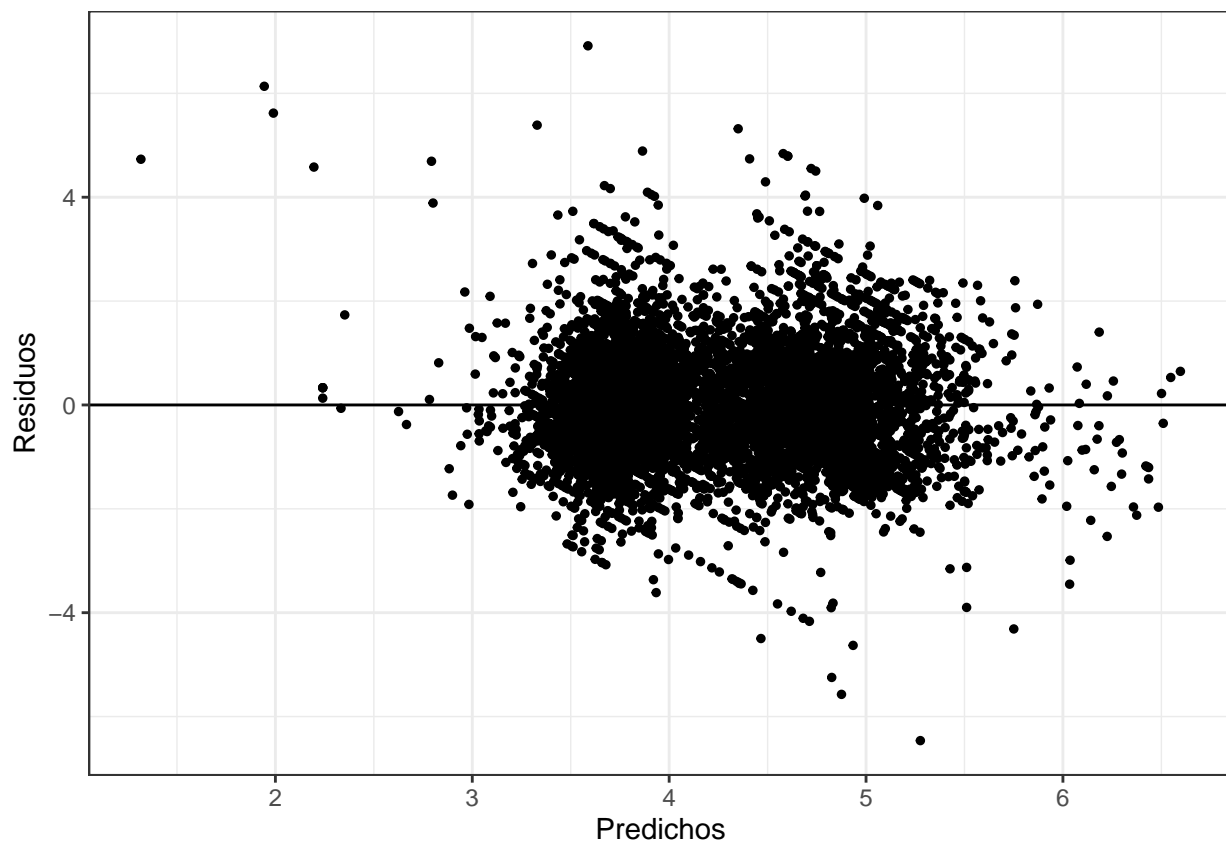
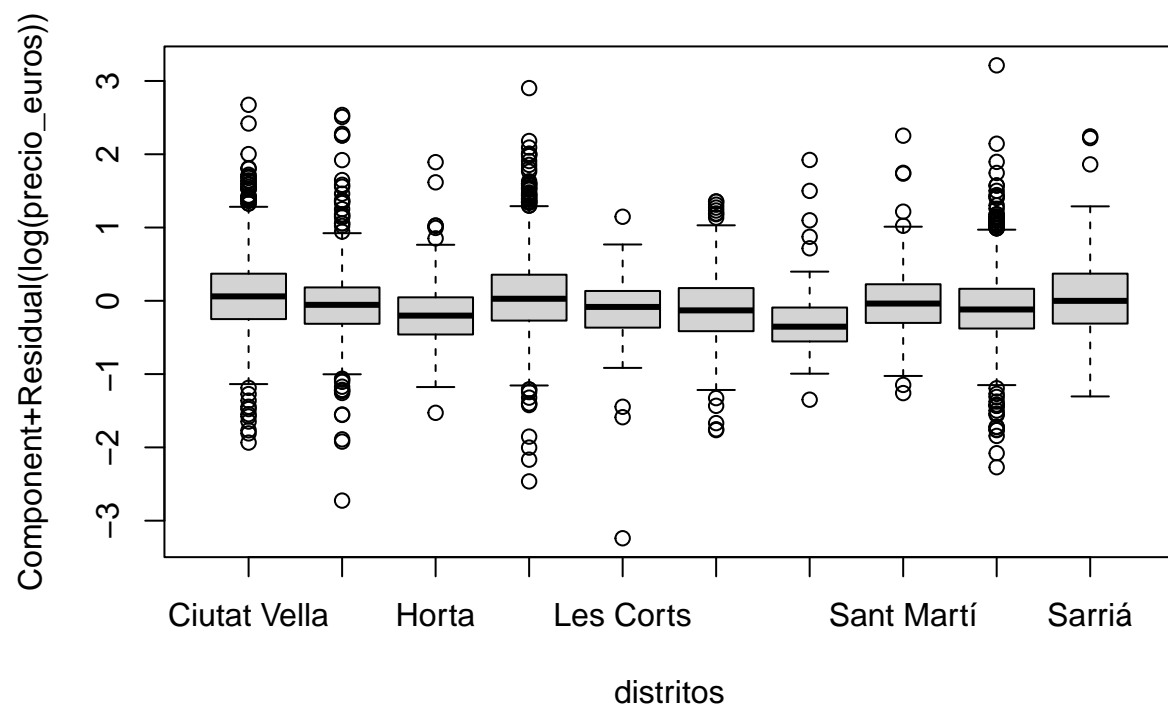
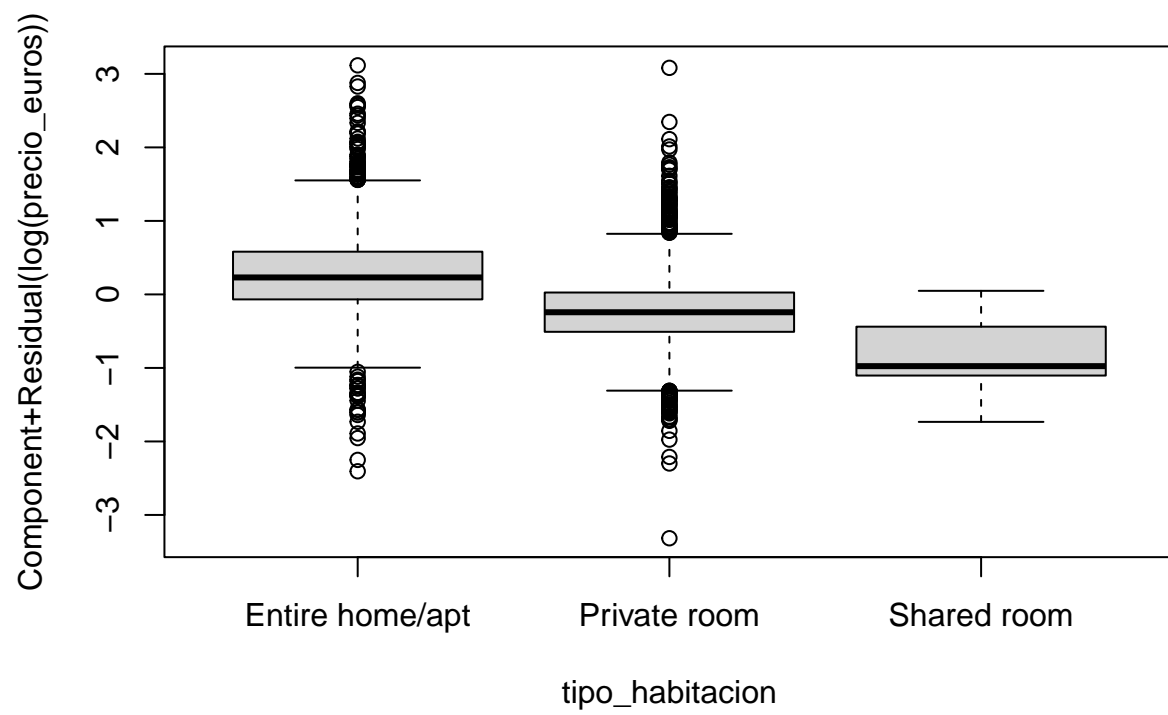
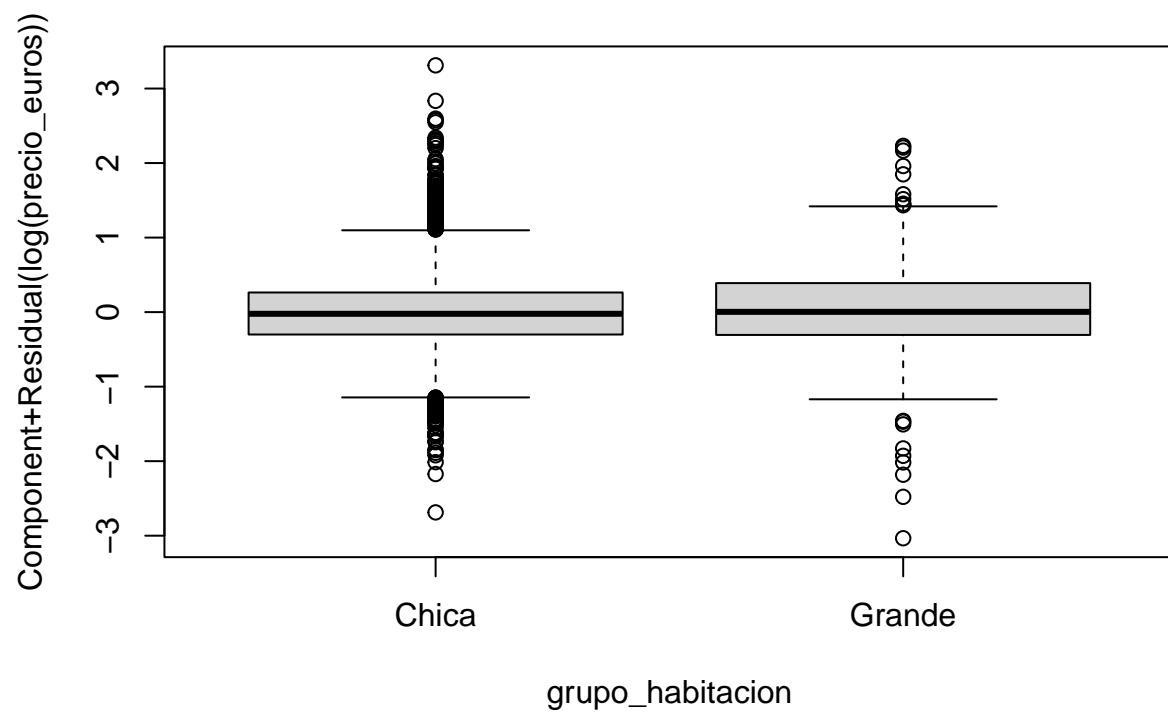
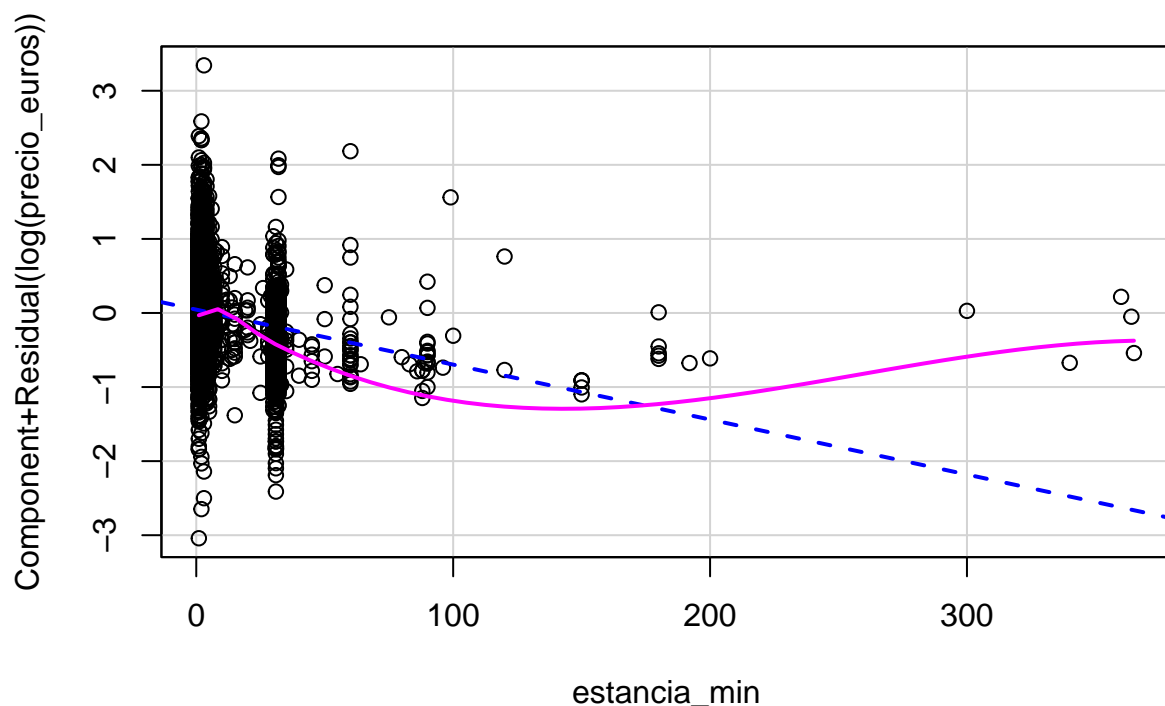


Figure 3: Gráfico de puntos de los residuos del modelo en función de los valores predichos









## Homoscedasticidad

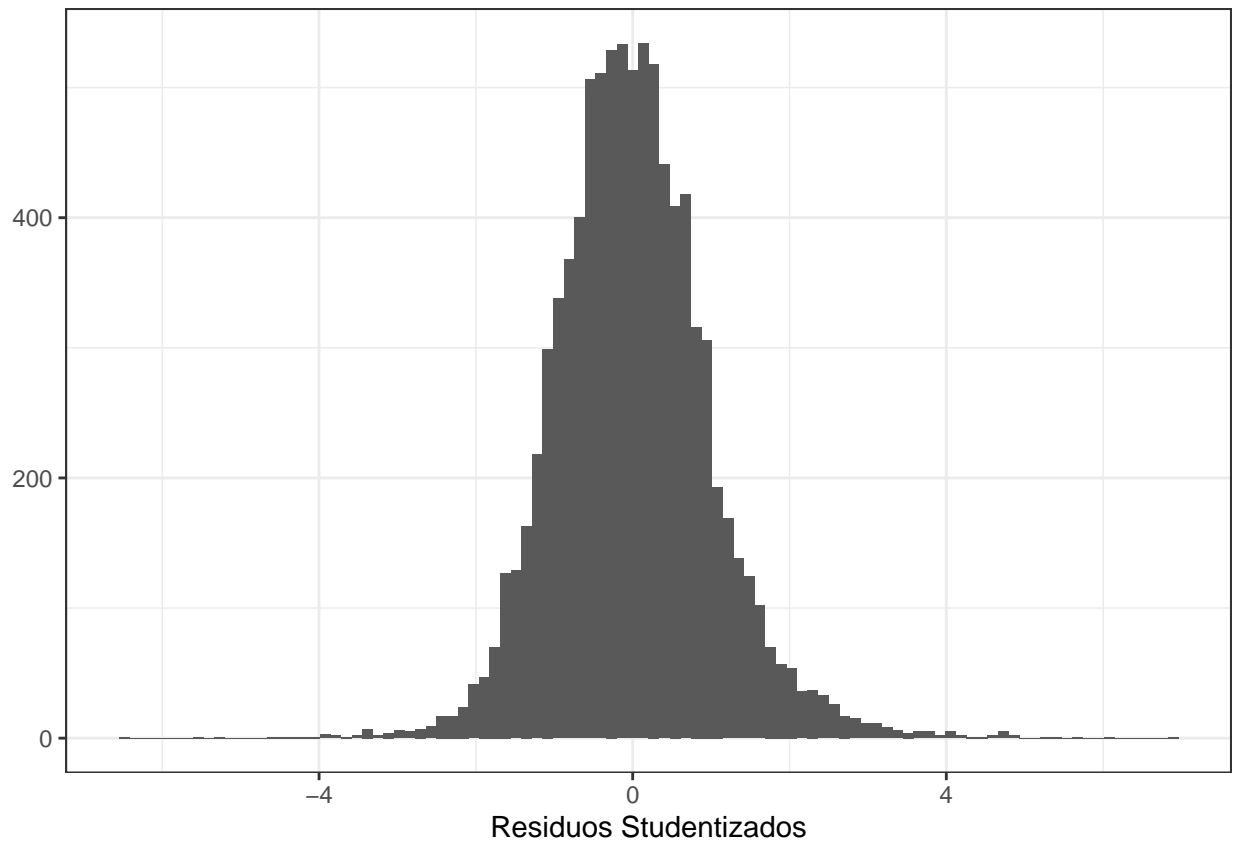
```
data.frame(breusch_pagan(mod0))
```

```
##      statistic      p.value parameter      method alternative
## 1  501.4026 8.526621e-91      24 Koenker (studentised)      greater
```

Como ya se mencionó, los dos primeros supuestos diagnosticados son importantes pero no tanto para un modelo como el que se presentó, por lo cual se centró la atención en diagnosticar los supuestos restantes, como la homoscedasticidad, donde a partir del test de BREUSCH-PAGAN se pudo observar que en el modelo inicial no se cumplía este supuesto. El p-valor dió muy bajo, lo que provocó que se rechazara  $H_0$ . Subir dicho p-valor fue uno de los cambios que se realizó en el modelo.

## Normalidad

```
ggplot() + geom_histogram(aes(x=rstudent(mod0)),bins = 100) + xlab("Residuos Studentizados")+ theme_bw
```

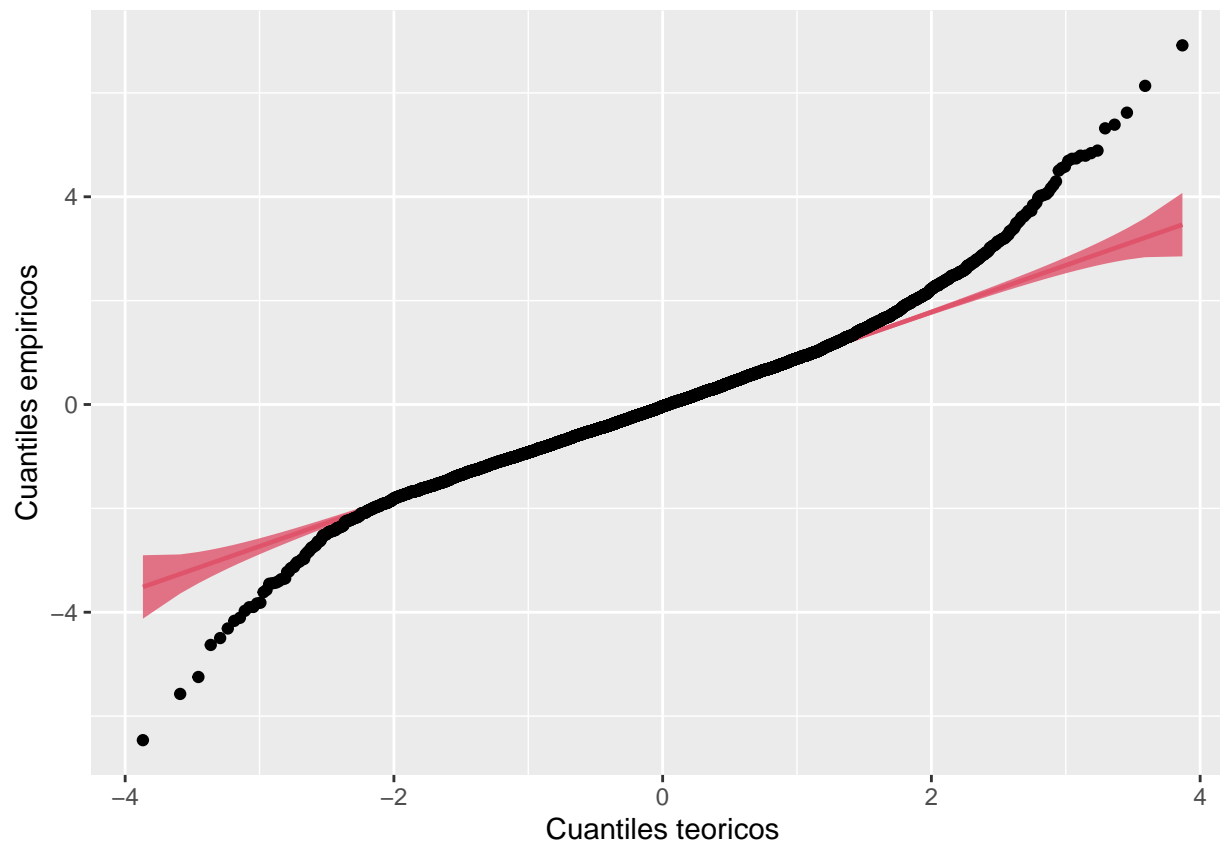


*#RESOLVER: VER SI SE PUEDE AGREGAR CAMPANA.*

OJO es sensible al tamaño de las barras (bins) por lo que se puede interpretar de diferentes formas nos gustaría ver que tenga la forma de una campana simetrica, en este caso tiene más o menos la forma, importante tomar en cuenta que el tamaño de los bins influye muchísimo en la forma del histograma.

*#RESOLVER: PONER EN FORMATO LINDO EL GRÁFICO*

```
ggplot(data = df_final, aes(sample = res_ext)) +
  stat_qq_band(fill = 2) +
  stat_qq_line(col = 2) +
  stat_qq_point() +
  xlab("Cuantiles teoricos")+
  ylab("Cuantiles empiricos")
```



Otro supuesto importante es el de la normalidad, para evaluarlo se realizó un histograma y un QQ-Plot, en el primer gráfico se pudo observar que tenía forma de una campana simétrica, en cambio en el segundo gráfico los puntos se encontraban fuera de la banda, es decir, el primer gráfico dio indicios de que si se cumplía el supuesto de normalidad pero el segundo mostraba que no. Para definir el cumplimiento o no del supuesto se realizó la Prueba de Kolmogorov-Smirnov

```
ks.test(df_final$residuos_int, 'pnorm')
```

```
## Warning in ks.test.default(df_final$residuos_int, "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: df_final$residuos_int
## D = 0.032889, p-value = 7.096e-09
## alternative hypothesis: two-sided
```

```
# QUEREMOS PVALOR ALTO PARA NO RECHAZAR
```

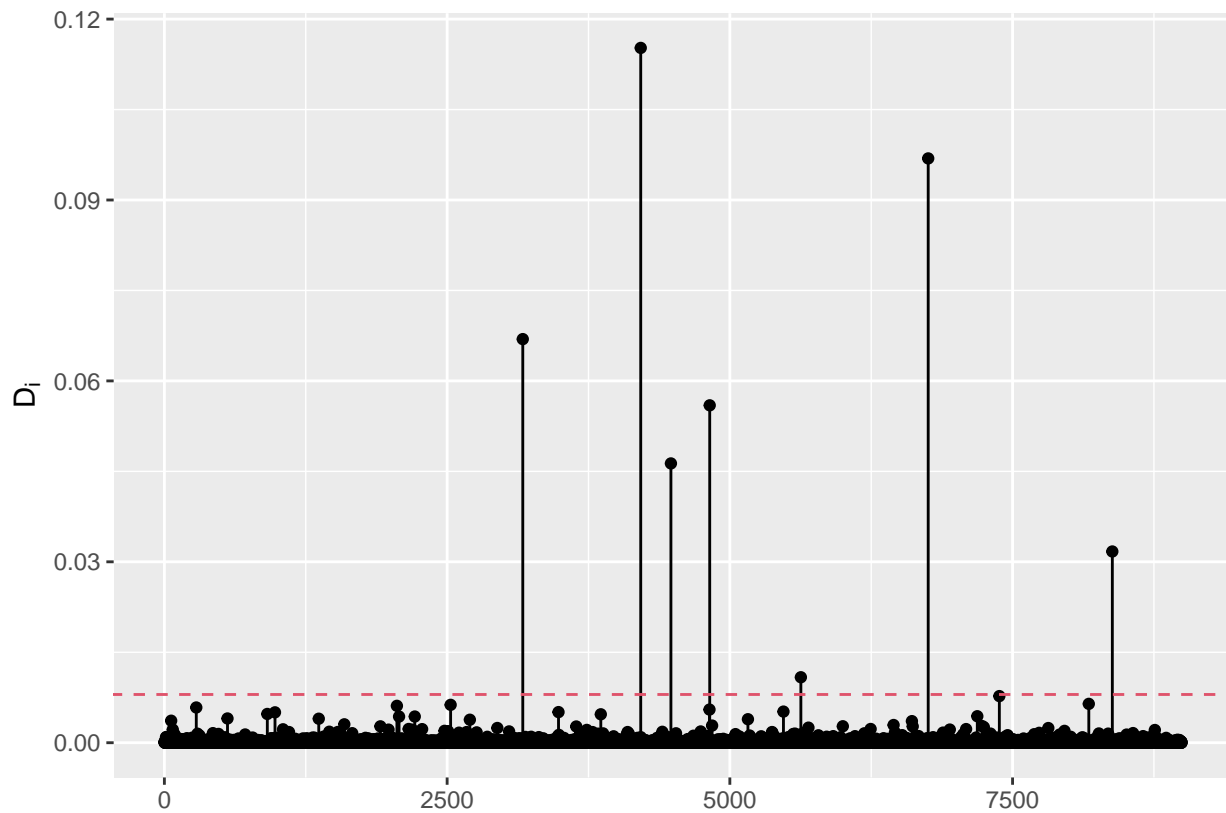
```
#RESOLVER: AGREGAR COMENTARIO DE ESTA PRUEBA.
```

## Atípicos

Por último se realizó el gráfico de la Distancia de Cook para observar si habían observaciones atípicas. Usualmente en este gráfico se hace la línea roja horizontal al nivel de  $4/n$ , donde  $n$  es el número de observaciones,



como en este modelo las observaciones son muchas esta linea roja quedaba muy baja, haciendo referencia a que todas las osbervaciones eran atípicas, para poder interpretar mejor esto se decidió que la linea esté en el nivel de 4/500, en ese punto las observaciones atípicas terminan siendo las más diferentes.



```
## # A tibble: 7 x 23
##   barrio      tipo_habitacion personas banios habitaciones camas precio_euros
##   <chr>      <fct>          <dbl>  <dbl>         <dbl> <fct>      <dbl>
## 1 Eixample    Entire home/apt      3      1             1 1         120
## 2 El Gotic    Entire home/apt      4      1             2 3          99
## 3 Eixample    Entire home/apt      5      2             3 3          75
## 4 Sant Andreu Private room         2      1             1 1          34
## 5 Sarria-Sant G~ Entire home/apt      6      3             4 6         100
## 6 La Maternitat~ Private room        16      2             6 18           9
## 7 Eixample    Private room        16      4             7 16          40
## # i 16 more variables: estancia_min <dbl>, puntuacion <dbl>, TV <fct>,
## #   Wifi <fct>, Air_conditioning <fct>, Elevator <fct>, Breakfast <fct>,
## #   Pets_allowed <fct>, Patio_or_balcony <fct>, check_in_24_hs <fct>,
## #   distritos <fct>, grupo_habitacion <fct>, grupo_banios <fct>,
## #   predichos <dbl>, residuos <dbl>, residuos_int <dbl>
```

## Corrección del modelo

A continuación se verán algunos de los cambios realizados en el modelo con respecto al cumplimiento de los principales supuestos junto con los nuevos resultados.

```
df_2=df_final[-id_atipico,]
```

```
mod1 <- lm(log(precio_euros) ~ distritos + tipo_habitacion+personas+ grupo_banios+ grupo_habitacion+es
```

```
#para "vichar"
```

```
Anova(mod1)
```

```
## Warning in printHypothesis(L, rhs, names(b)): one or more coefficients in the hypothesis include
## arithmetic operators in their names;
## the printed representation of the hypothesis will be omitted
```

```
## Anova Table (Type II tests)
```

```
##
```

```
## Response: log(precio_euros)
```

```
##          Sum Sq   Df F value    Pr(>F)
## distritos      67.77    9  33.2597 < 2.2e-16 ***
## tipo_habitacion 225.56    2 498.1404 < 2.2e-16 ***
## personas      202.31    1 893.6037 < 2.2e-16 ***
## grupo_banios     2.55    1  11.2424 0.0008028 ***
## grupo_habitacion  1.83    1   8.0832 0.0044778 **
## estancia_min   133.82    1 591.0658 < 2.2e-16 ***
## puntuacion      3.64    1  16.0812 6.118e-05 ***
## TV              4.36    1  19.2702 1.148e-05 ***
## Wifi           5.19    1  22.9104 1.724e-06 ***
## Air_conditioning 33.93    1 149.8781 < 2.2e-16 ***
## Elevator        3.15    1  13.9162 0.0001923 ***
## Breakfast       3.79    1  16.7617 4.275e-05 ***
## Pets_allowed    1.71    1   7.5489 0.0060165 **
## Patio_or_balcony 1.46    1   6.4337 0.0112143 *
## check_in_24_hs  11.90    1  52.5559 4.528e-13 ***
## Residuals      2028.99 8962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod1)
```

```
##
```

```
## Call:
```

```
## lm(formula = log(precio_euros) ~ distritos + tipo_habitacion +
##     personas + grupo_banios + grupo_habitacion + estancia_min +
##     puntuacion + TV + Wifi + Air_conditioning + Elevator + Breakfast +
##     Pets_allowed + Patio_or_balcony + check_in_24_hs, data = df_2)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.6878 -0.2995 -0.0176  0.2790  3.3179
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.9961080   0.0695977   57.417 < 2e-16 ***
## distritosGracia -0.1290656   0.0206743   -6.243 4.49e-10 ***
```

```
## distritosHorta -0.2455964 0.0312206 -7.866 4.07e-15 ***
## distritosL'Eixample -0.0242337 0.0148030 -1.637 0.101648
## distritosLes Corts -0.2098904 0.0386699 -5.428 5.86e-08 ***
## distritosNou Barris -0.1977588 0.0220471 -8.970 < 2e-16 ***
## distritosSant Andreu -0.3757242 0.0422130 -8.901 < 2e-16 ***
## distritosSant Martí -0.1038548 0.0253431 -4.098 4.20e-05 ***
## distritosSants-Montjuic -0.1881828 0.0183239 -10.270 < 2e-16 ***
## distritosSarrià -0.0275176 0.0302213 -0.911 0.362565
## tipo_habitacionPrivate room -0.5291639 0.0172606 -30.657 < 2e-16 ***
## tipo_habitacionShared room -1.1247260 0.0930823 -12.083 < 2e-16 ***
## personas 0.1246592 0.0041702 29.893 < 2e-16 ***
## grupo_baniosPocos -0.1306677 0.0389707 -3.353 0.000803 ***
## grupo_habitacionGrande 0.0515917 0.0181463 2.843 0.004478 **
## estancia_min -0.0106573 0.0004384 -24.312 < 2e-16 ***
## puntuacion 0.0021575 0.0005380 4.010 6.12e-05 ***
## TV1 0.0966987 0.0220281 4.390 1.15e-05 ***
## Wifi1 -0.1067233 0.0222968 -4.786 1.72e-06 ***
## Air_conditioning1 0.1589516 0.0129836 12.242 < 2e-16 ***
## Elevator1 0.0423650 0.0113566 3.730 0.000192 ***
## Breakfast1 0.0900894 0.0220047 4.094 4.28e-05 ***
## Pets_allowed1 0.0445963 0.0162314 2.748 0.006017 **
## Patio_or_balcony1 0.0301131 0.0118720 2.536 0.011214 *
## check_in_24_hs1 0.1146730 0.0158179 7.250 4.53e-13 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4758 on 8962 degrees of freedom
## Multiple R-squared: 0.6048, Adjusted R-squared: 0.6038
## F-statistic: 571.5 on 24 and 8962 DF, p-value: < 2.2e-16
```

```
#distritos(Gracia) 0.1495579
#exp(0.1495579)=1.161834
```

*#"Gracia es 16% más caro que el grupo de referencia DEJANDO TODO DEMAS CONSTANTE" TOMAR SOLO ALGUNAS VA*

```
#RESOLVER: PONER EN TABLA
summary(mod0)$sigma
```

```
## [1] 0.4819322
```

```
summary(mod1)$sigma
```

```
## [1] 0.4758147
```

```
summary(mod0)$r.squared
```

```
## [1] 0.5947203
```

```
summary(mod1)$r.squared
```

```
## [1] 0.6048337
```

```
## UNA VEZ DEFINIDO EL MODELO, PODEMOS HACER UN SPLIT DE DATOS 70/30 PARA ENTRENAR Y TESTEAR.
```

```
# p-valor de grupo_habitación alto, ¿está relacionado con el grupo de referencia? ¿Habría que sacar la v
```

## Resultados

```
df_test$predicciones = predict(mod1,newdata = df_test)

results=df_test %>% select(precio_euros,predicciones)%>%mutate(precio_predic=exp(predicciones),
                      delta_precio=round(precio_euros-exp(predicciones),2))

y_prom=mean(df_test$precio_euros)

num=sum((results$delta_precio)**2)

den=sum((df_test$precio_euros-y_prom)**2)

1-(num/den)
```

```
## [1] 0.3925068
```

```
summary(mod1)
```

```
##
## Call:
## lm(formula = log(precio_euros) ~ distritos + tipo_habitacion +
##     personas + grupo_banios + grupo_habitacion + estancia_min +
##     puntuacion + TV + Wifi + Air_conditioning + Elevator + Breakfast +
##     Pets_allowed + Patio_or_balcony + check_in_24_hs, data = df_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.6878 -0.2995 -0.0176  0.2790  3.3179
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.9961080   0.0695977   57.417 < 2e-16 ***
## distritosGracia -0.1290656   0.0206743   -6.243 4.49e-10 ***
## distritosHorta  -0.2455964   0.0312206   -7.866 4.07e-15 ***
## distritosL'Eixample -0.0242337   0.0148030   -1.637 0.101648
## distritosLes Corts -0.2098904   0.0386699   -5.428 5.86e-08 ***
## distritosNou Barris -0.1977588   0.0220471   -8.970 < 2e-16 ***
## distritosSant Andreu -0.3757242   0.0422130   -8.901 < 2e-16 ***
## distritosSant Martí -0.1038548   0.0253431   -4.098 4.20e-05 ***
## distritosSants-Montjuic -0.1881828   0.0183239  -10.270 < 2e-16 ***
## distritosSarrià    -0.0275176   0.0302213   -0.911 0.362565
## tipo_habitacionPrivate room -0.5291639   0.0172606  -30.657 < 2e-16 ***
## tipo_habitacionShared room -1.1247260   0.0930823  -12.083 < 2e-16 ***
```

```
## personas          0.1246592  0.0041702  29.893  < 2e-16 ***
## grupo_baniosPocos -0.1306677  0.0389707  -3.353  0.000803 ***
## grupo_habitacionGrande 0.0515917  0.0181463   2.843  0.004478 **
## estancia_min      -0.0106573  0.0004384 -24.312  < 2e-16 ***
## puntuacion        0.0021575  0.0005380   4.010  6.12e-05 ***
## TV1               0.0966987  0.0220281   4.390  1.15e-05 ***
## Wifi1            -0.1067233  0.0222968  -4.786  1.72e-06 ***
## Air_conditioning1  0.1589516  0.0129836  12.242  < 2e-16 ***
## Elevator1         0.0423650  0.0113566   3.730  0.000192 ***
## Breakfast1        0.0900894  0.0220047   4.094  4.28e-05 ***
## Pets_allowed1     0.0445963  0.0162314   2.748  0.006017 **
## Patio_or_balcony1  0.0301131  0.0118720   2.536  0.011214 *
## check_in_24_hs1   0.1146730  0.0158179   7.250  4.53e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4758 on 8962 degrees of freedom
## Multiple R-squared:  0.6048, Adjusted R-squared:  0.6038
## F-statistic: 571.5 on 24 and 8962 DF, p-value: < 2.2e-16
```

ACÁ IRÍA LAS PREDICCIONES Y LAS INTERPRETACIONES DEL MODELO

## Conclusiones

Luego de realizadas las predicciones, se logró concluir que el modelo es funcional según las pruebas realizadas aunque tal vez creemos que se podría seguir mejorando ya sea agregando nuevas variables que puedan ser relevantes y que no estén en los datos iniciales.

ACÁ IRÍA SI EL MODELO SIRVIÓ O NO, SI “RESPONDE” O NO LA CUESTIÓN PLANTEADA EN LA INTRO

## Bibliografía

CAPAZ PONER LAS REFERENCIAS DE LAS DIAPO DEL PROFE