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

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 publidcadas 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 agrupan 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.

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

```
## # A tibble: 12,848 x 20
##    barrio        tipo_habitacion personas banios habitaciones camas precio_euros
##    <chr>         <fct>              <dbl>  <dbl>        <dbl> <fct>        <dbl>
##  1 Sant Marta    Entire home/apt        6    1            2 4              130
##  2 La Sagrada F~ Entire home/apt        8    2            3 6               60
##  3 Sant Marta    Private room           2    1            1 1               33
##  4 Sant Marta    Entire home/apt        6    2            3 8              210
##  5 Vila de Grac~ Private room           2    1            1 1               45
##  6 Horta-Guinar~ Private room           2    1            1 2               42
##  7 Horta-Guinar~ Private room           3    1            1 2               53
##  8 Camp d'en Gr~ Entire home/apt        4    1            1 1               75
##  9 Gracia        Entire home/apt        5    1.5          3 3               85
## 10 Les Corts     Private room           1    1            1 1               30
## # i 12,838 more rows
## # i 13 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>
```

Análisis exploratorio

Como parte de la estadística descriptiva se crearon gráficos donde se relacionan cada una 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`.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

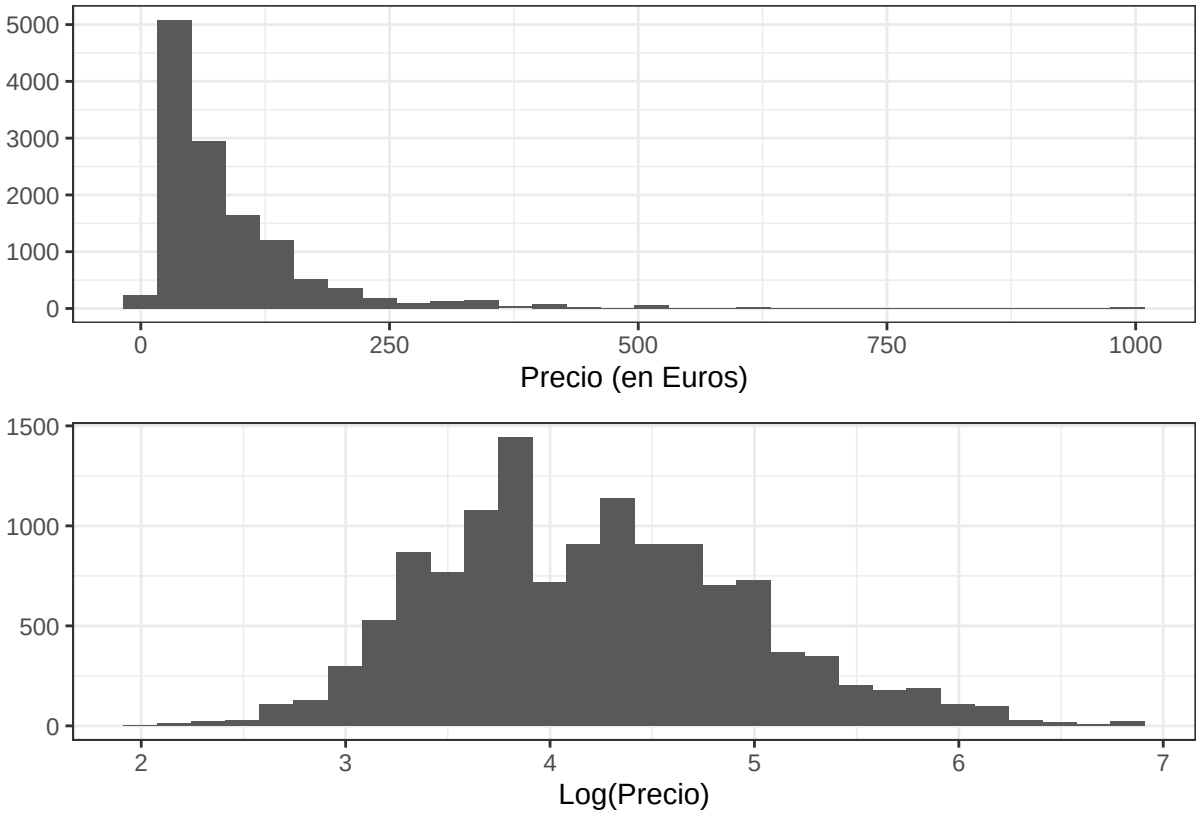


Figure 1: Histograma de la variable de respuesta Precio

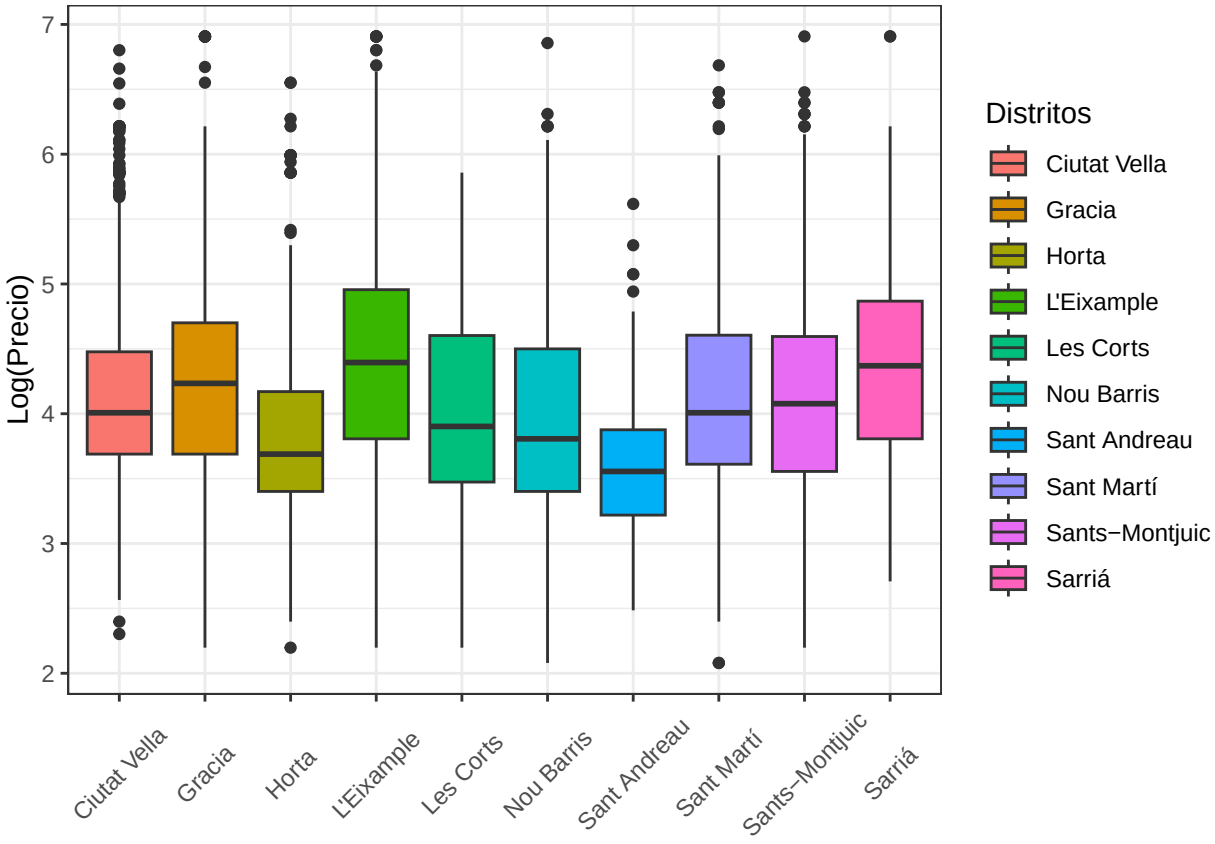


Figure 2: Dispersión de Precio (en euros) por Distrito

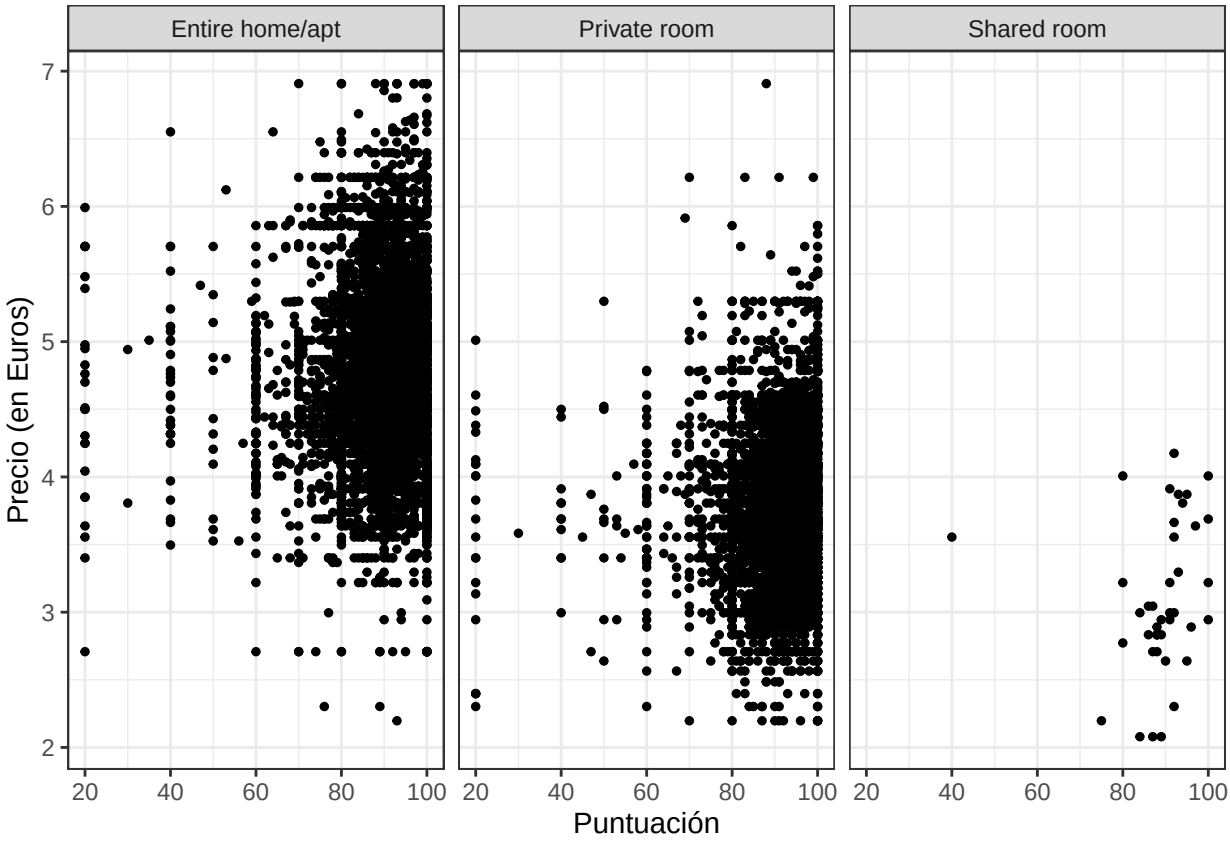
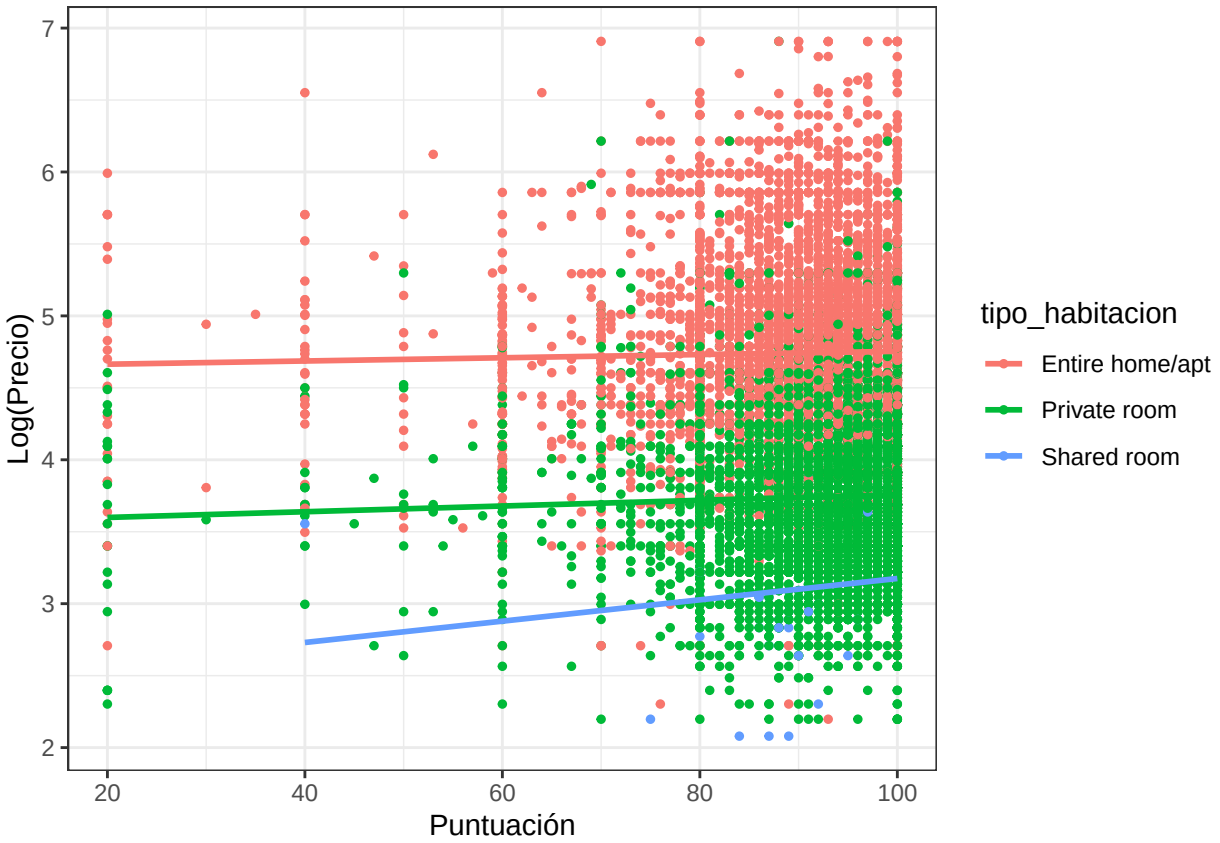


Figure 3: Dispersión de Precio (en euros) según la puntuación, por tipo de alojamiento



#TERMINAR No hay relación entre precio y puntuación. Los apartamentos enteros tienden a tener mayor precio.

Metodología

Selección de variables

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
#para "vichar"
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          94.81     9   45.8737 < 2.2e-16 ***
## tipo_habitacion   327.42     2  712.8747 < 2.2e-16 ***
## personas          294.53     1 1282.5275 < 2.2e-16 ***
## grupo_banios        5.08     1   22.1370 2.565e-06 ***
## grupo_habitacion    2.55     1   11.0997 0.0008659 ***
## estancia_min      151.32     1  658.9063 < 2.2e-16 ***
## puntuacion          5.66     1   24.6631 6.916e-07 ***
## TV                  4.97     1   21.6422 3.318e-06 ***
## Wifi                6.26     1   27.2780 1.790e-07 ***
## Air_conditioning   45.32     1  197.3564 < 2.2e-16 ***
## Elevator            4.47     1   19.4640 1.034e-05 ***
## Breakfast           3.76     1   16.3904 5.185e-05 ***
## Pets_allowed        2.18     1    9.5071 0.0020511 **
## Patio_or_balcony    3.52     1   15.3086 9.177e-05 ***
## check_in_24_hs     17.42     1   75.8530 < 2.2e-16 ***
## Residuals        2944.76 12823
## ---
## 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
## -3.0321 -0.3051 -0.0197  0.2773  3.3263
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.9895934  0.0581148  68.650  < 2e-16 ***
## distritosGracia              -0.1026450  0.0174491  -5.883 4.14e-09 ***
## distritosHorta               -0.2211466  0.0259112  -8.535  < 2e-16 ***
## distritosL'Eixample          -0.0126856  0.0124951  -1.015 0.310008
## distritosLes Corts           -0.2548561  0.0321179  -7.935 2.28e-15 ***
## distritosNou Barris          -0.1837821  0.0187587  -9.797  < 2e-16 ***
## distritosSant Andreau        -0.3489471  0.0350054  -9.968  < 2e-16 ***
## distritosSant Martí          -0.1076485  0.0212190  -5.073 3.97e-07 ***
## distritosSants-Montjuic      -0.1834576  0.0154162 -11.900  < 2e-16 ***
## distritosSarriá              -0.0159291  0.0252297  -0.631 0.527815
## tipo_habitacionPrivate room -0.5224468  0.0143080 -36.514  < 2e-16 ***
## tipo_habitacionShared room  -1.1575063  0.0761360 -15.203  < 2e-16 ***
## personas                     0.1238842  0.0034593  35.812  < 2e-16 ***
## grupo_baniosPocos           -0.1530209  0.0325231  -4.705 2.57e-06 ***
## grupo_habitacionGrande       0.0508267  0.0152559   3.332 0.000866 ***
## estancia_min                -0.0081616  0.0003180 -25.669  < 2e-16 ***
## puntuacion                   0.0022409  0.0004512   4.966 6.92e-07 ***
## TV1                          0.0861456  0.0185175   4.652 3.32e-06 ***
## Wifi1                       -0.0977870  0.0187230  -5.223 1.79e-07 ***
## Air_conditioning1            0.1539145  0.0109560  14.048  < 2e-16 ***
```

```
## Elevator1                    0.0422385  0.0095740   4.412 1.03e-05 ***
## Breakfast1                   0.0754272  0.0186309   4.049 5.18e-05 ***
## Pets_allowed1                0.0419002  0.0135891   3.083 0.002051 **
## Patio_or_balcony1            0.0389993  0.0099676   3.913 9.18e-05 ***
## check_in_24_hs1              0.1164991  0.0133763   8.709  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4792 on 12823 degrees of freedom
## Multiple R-squared:  0.5983, Adjusted R-squared:  0.5975
## F-statistic: 795.6 on 24 and 12823 DF,  p-value: < 2.2e-16
```

Diagnóstico

Luego de la limpieza de datos y estadística descriptiva se comenzó la etapa de diagnótico, 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 está 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. #RESPUESTA A LA PREGUNTA: VOLVER A CALCUAR VIF Y NÚMERO DE CONDICIÓN LUEGO DE HACERLE TODOS LOS ARREGLOS AL MODELO Y VER SI LOS RESULTADOS SE PUEDEN INTERPRETAR, EN EL CASO CONTRARIO EXPLICAR POR QUÉ NO TIENE SENTIDO INTERPETAR #(SI NO SE PUEDEN INTERPRETAR PREGUNTAR AL PROFE SI LA "RAZÓN" ESTÁ BIEN)

Linealidad: este supuesto se basa en la linealidad de la variable precio_euros y cada una de las variables explicativas, si en el modelo no hay linealidad se presentarán probelmas de correlación entre los residuos y variabilidad de los mismos. Para verificar el cumplimiento de este supuesto se realizó un análisis gráfico entre los Yˆ y ˆ, en el cual se busca no encontrar patrones. #NO NOS PREOCUPEMOS

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 verá mediante gráficos que relaciona cada variable explicativa con la de respuesta y a partir de una prueba de hipótesis donde buscamos 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 lineas del script se pusieron a prueba cada uno de los supuestos antes mencionados.

Multicolinealidad

[1] 1662.29 #TERMINAR

A partir de la etapa del diagnóstico se llegó a diferentes conclusiones sobre que hay que cambiar en el modelo para que este sea lo mejor posible, se comenzó evaluando el supuesto de multicolinealidad, para esto se calculó el número de condición, el cual dió 2792.348, es decir, hay problemas de multicolinealidad.

El número de condición es muy alto, lo que nos dice que hay problemas de multicolinealidad.

```r
vif(mod0)
```

```
##             distritosGracia              distritosHorta
##                    1.291644                    1.116569
##         distritosL'Eixample          distritosLes Corts
##                    2.004768                    1.083859
##         distritosNou Barris       distritosSant Andreau
##                    1.264004                    1.071245
##         distritosSant Martí     distritosSants-Montjuic
##                    1.201100                    1.393243
##             distritosSarriá tipo_habitacionPrivate room
##                    1.146819                    2.855889
##  tipo_habitacionShared room                     personas
##                    1.031610                    3.225694
##           grupo_baniosPocos      grupo_habitacionGrande
##                    1.142367                    1.853855
##                estancia_min                   puntuacion
##                    1.096326                    1.045461
##                         TV1                        Wifi1
##                    4.064418                    3.709875
##           Air_conditioning1                   Elevator1
##                    1.643705                    1.200002
##                  Breakfast1               Pets_allowed1
##                    1.047396                    1.020661
##           Patio_or_balcony1             check_in_24_hs1
##                    1.042212                    1.075528
```

```r
## VER SCRIPTS DE CLASE
```

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.

Linelidad

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

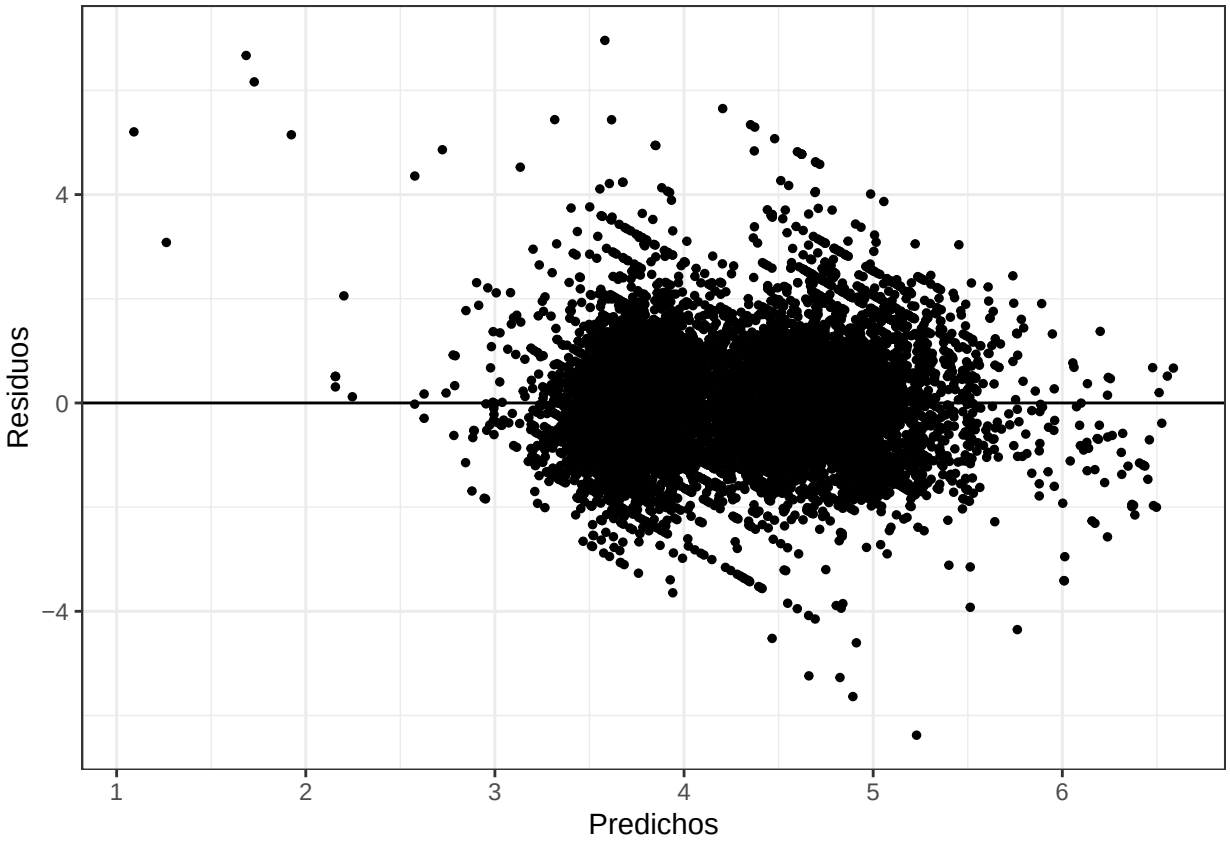
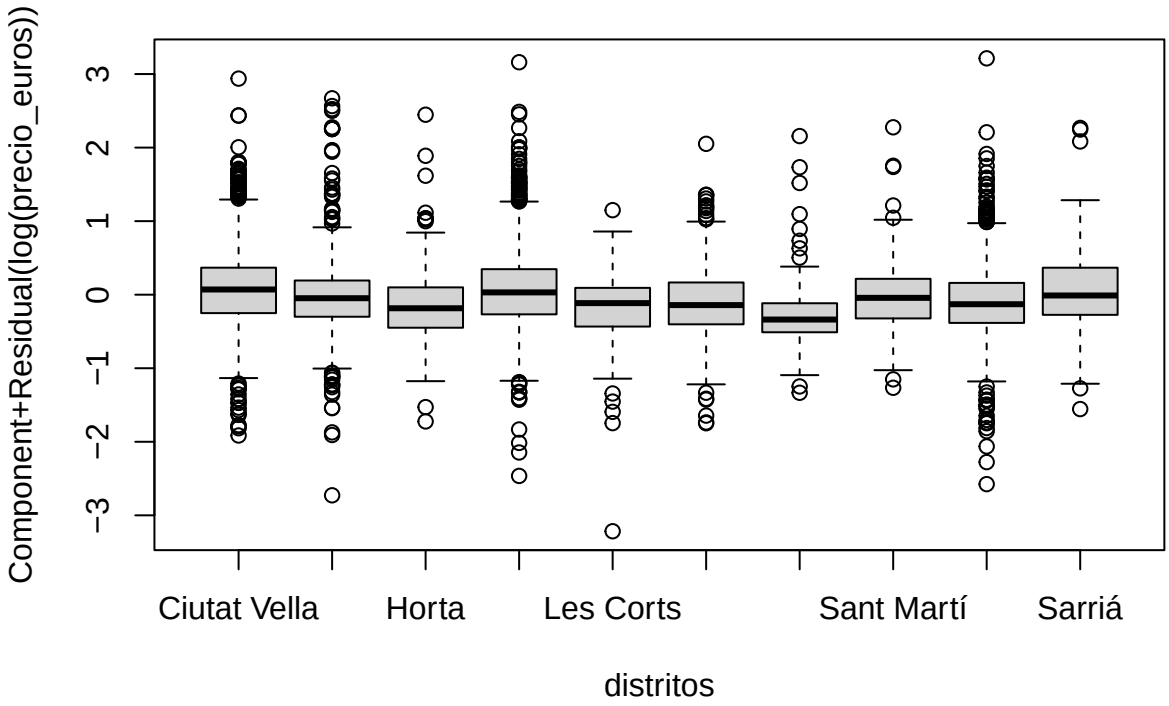
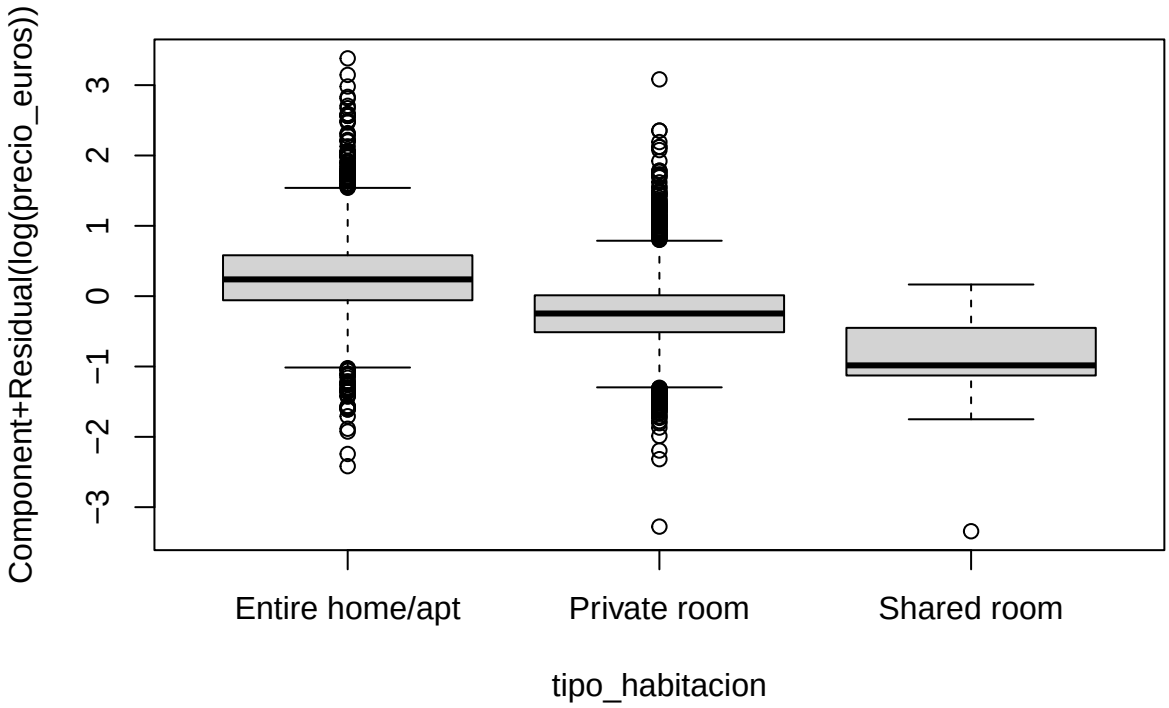
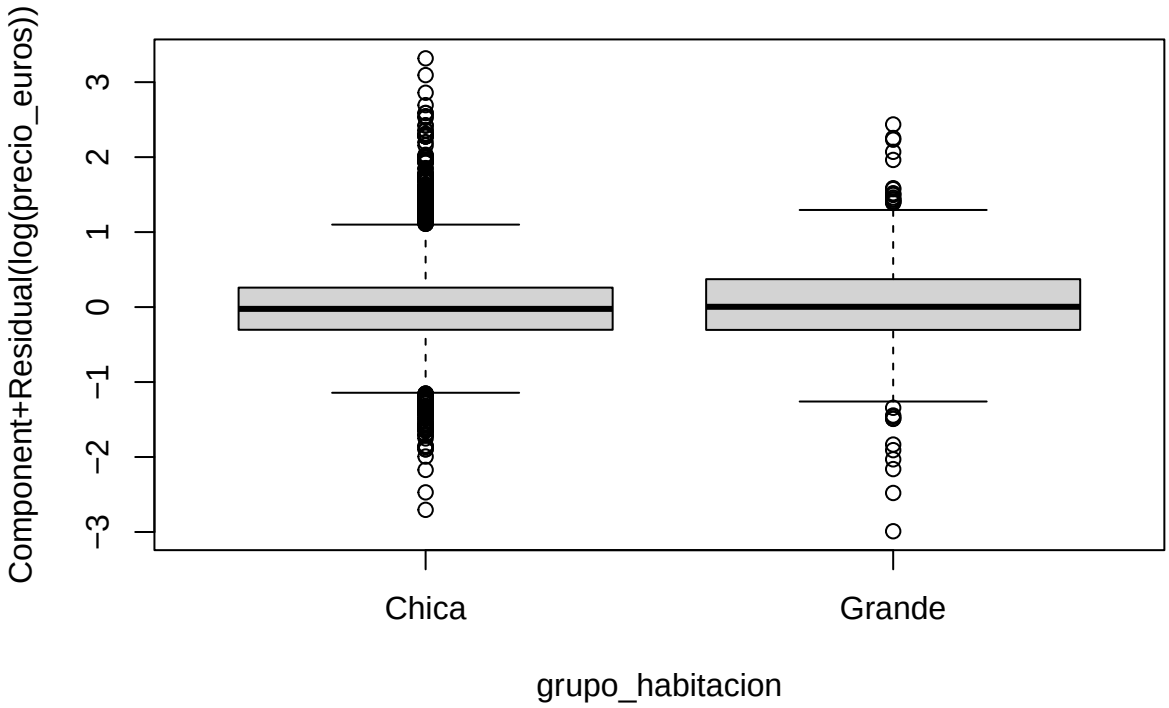
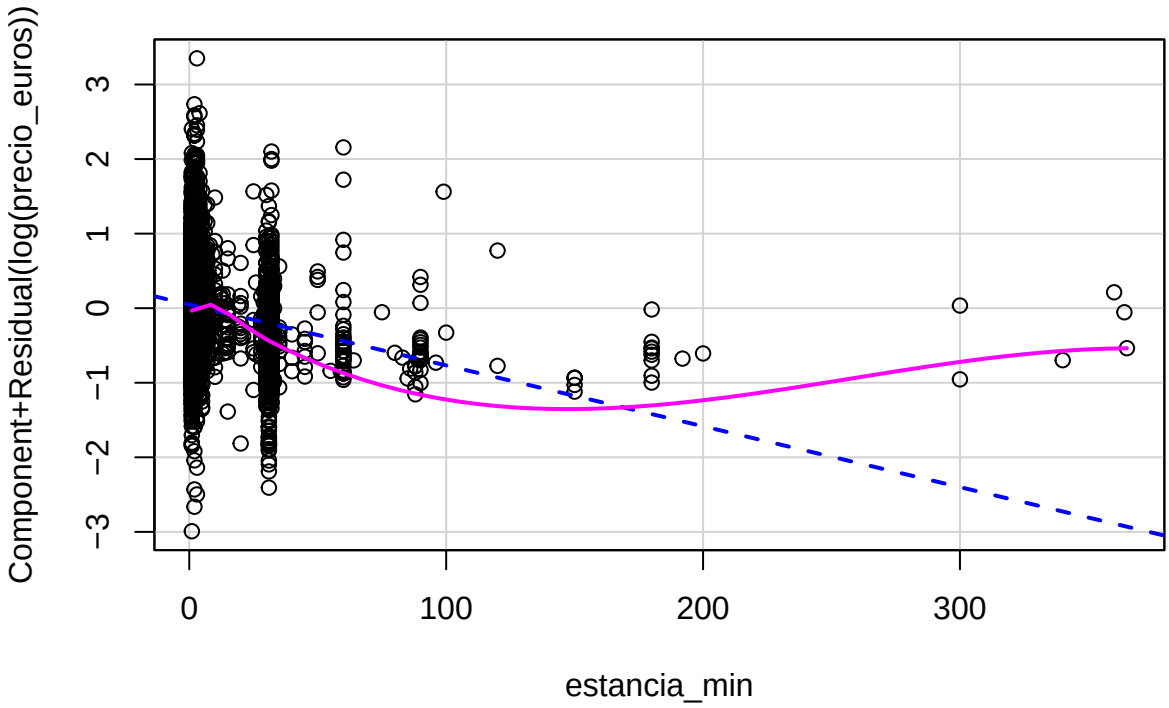


Figure 4: Gráfico de puntos de los residuos del modelo en función de los valores predichos


Component+Residual(log(precio_euros))
3
2
1
0
−1
−2
−3
Ciutat Vella
Horta
Les Corts
Sant Martí
Sarriá
distritos


Component+Residual(log(precio_euros))
3
2
1
0
−1
−2
−3
Entire home/apt
Private room
Shared room
tipo_habitacion


Component+Residual(log(precio_euros))
3
2
1
0
−1
−2
−3
Chica
Grande
grupo_habitacion



Homoscedasticidad

```
#install.packages("skedastic")
data.frame(breusch_pagan(mod0))
```

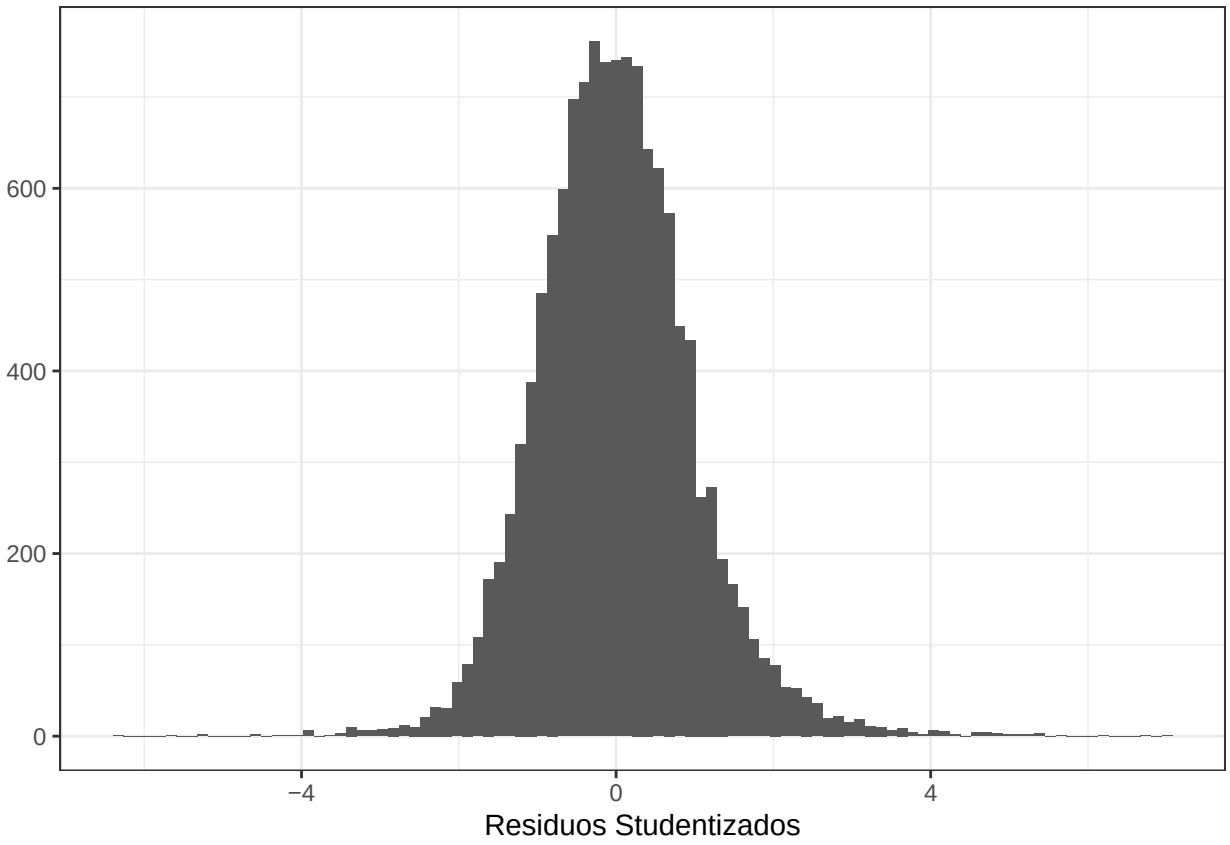
```
##   statistic       p.value parameter                method alternative
## 1  679.8462 4.280921e-128        24 Koenker (studentised)     greater
```

```
#knitr::kable(df_bp)%>%print()
#buscamos NO rechazar H0, el p-valor es 1.11e-129, muy chico, por lo cual se rechaza H0, debemos intent
#H0 nos dice que los errores son homoscedasticos

#hipotesis nula es que todos tienen la misma varianza, es decir, hay homoscedasticidad, la otra hipotes
#rechazo H0, NO HAY HOMOSEDASTICIDAD, p valor muy chico.

#RANDOMIZE o "Sanguche"
```

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 H0. Subir dicho p-valor fue unos de los cambios que se realizó en el modelo.

Normalidad

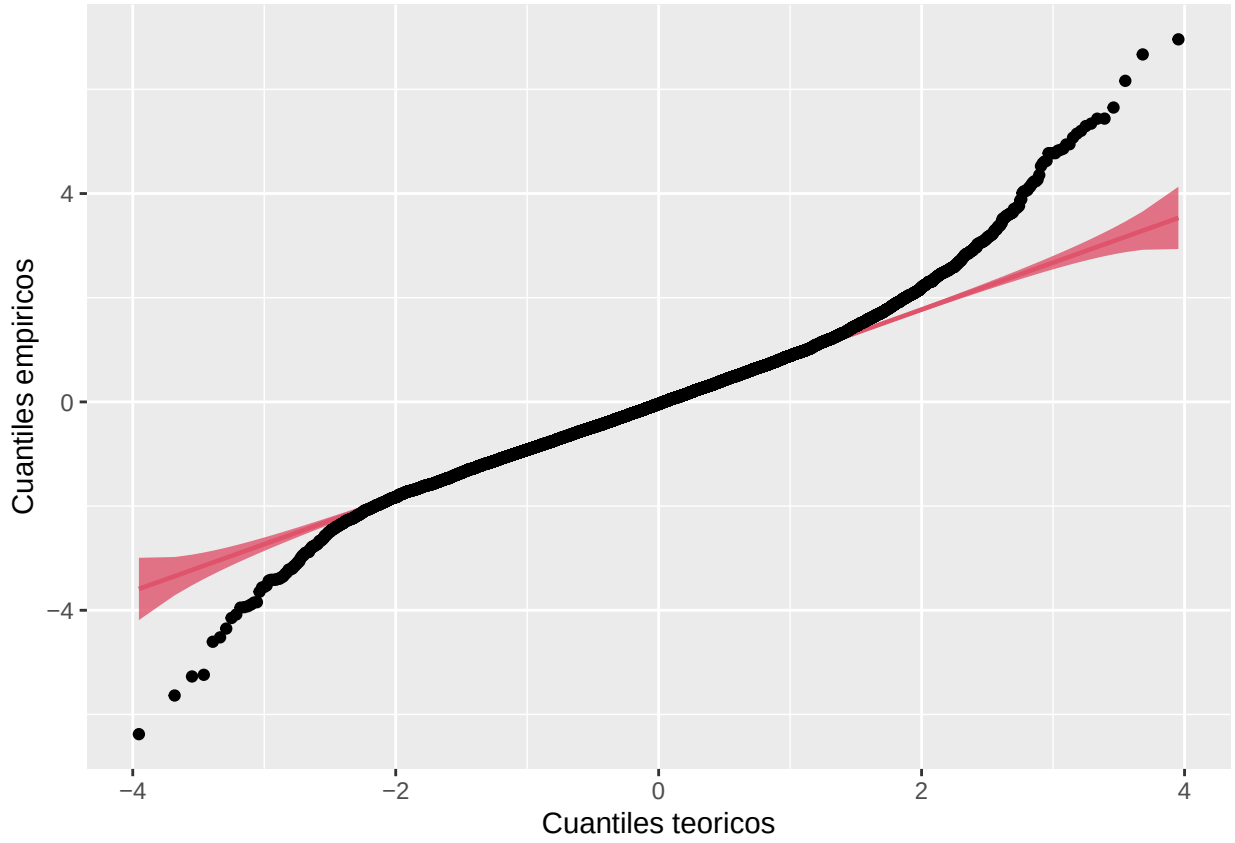
```
# histogramas
ggplot() + geom_histogram(aes(x=rstudent(mod0)),bins = 100) + xlab("Residuos Studentizados")+  theme_bw
```




```
#REVISAR: 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.

```
#REVISAR: PORNER 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 cumplia el supuesto de normalidad pero el segundo mostraba que no. Para definir el cumpliento o no del supuesto se realizó la Prueba de Kolmogorov-Smirnov

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

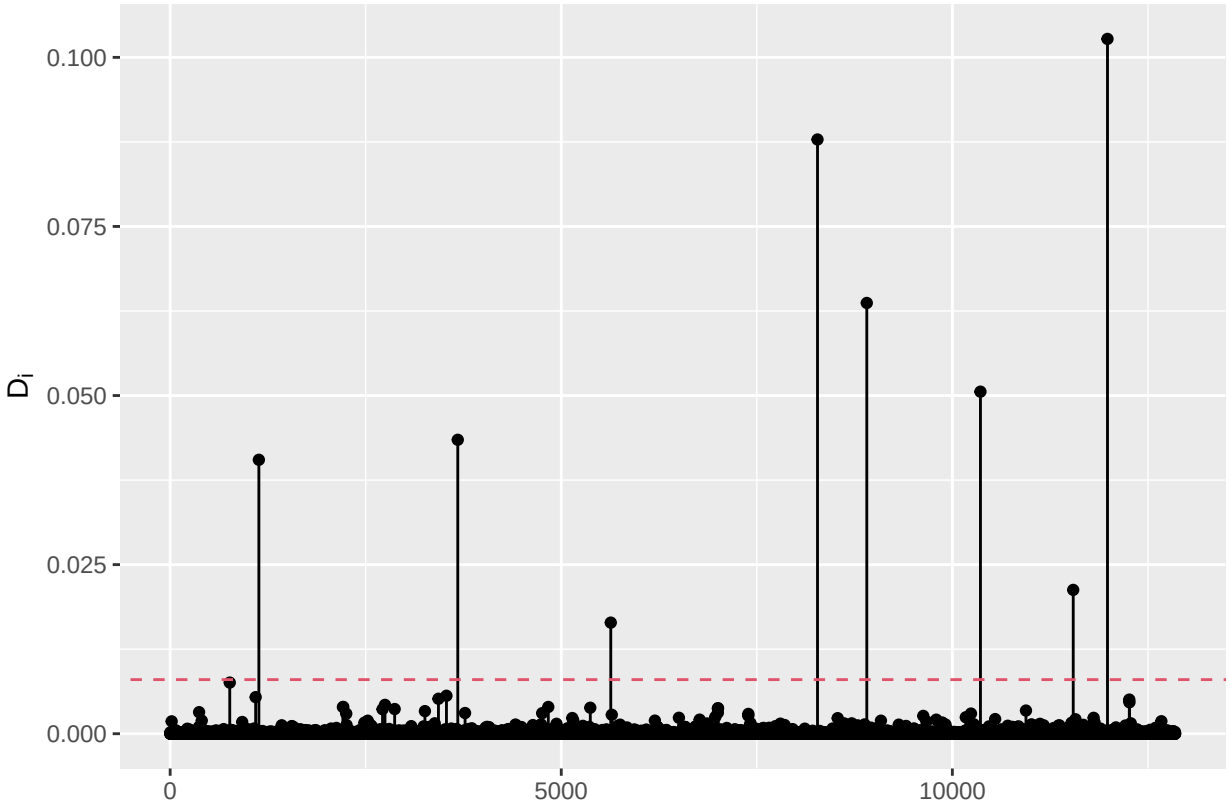
```
## Warning in ks.test.default(df_final$residuos_int, "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test

##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  df_final$residuos_int
## D = 0.033791, p-value = 3.617e-13
## 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: 8 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 Sants-Montjuic Shared room           16    2.5            1 14               9
## 7 La Maternitat~ Private room          16    2              6 18               9
## 8 Sants-Montjuic Private room           1    1              1 1               15
## # 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>
```

Correción 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.

```r
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           96.14     9   47.5377 < 2.2e-16 ***
## tipo_habitacion    336.52     2  748.7479 < 2.2e-16 ***
## personas           282.49     1 1257.0679 < 2.2e-16 ***
## grupo_banios         5.12     1   22.7616 1.854e-06 ***
## grupo_habitacion     2.15     1    9.5528 0.0020008 **
## estancia_min       196.58     1  874.7547 < 2.2e-16 ***
## puntuacion           5.28     1   23.5091 1.258e-06 ***
## TV                   5.07     1   22.5813 2.036e-06 ***
## Wifi                 6.16     1   27.4111 1.671e-07 ***
## Air_conditioning    43.01     1  191.3764 < 2.2e-16 ***
## Elevator             4.23     1   18.8165 1.450e-05 ***
## Breakfast            3.00     1   13.3656 0.0002573 ***
## Pets_allowed         1.86     1    8.2966 0.0039785 **
## Patio_or_balcony     2.61     1   11.6185 0.0006550 ***
## check_in_24_hs      17.23     1   76.6757 < 2.2e-16 ***
## Residuals         2879.81 12815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
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
## -2.7067 -0.3022 -0.0205  0.2756  3.3251
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.0391636  0.0577961  69.886  < 2e-16 ***
## distritosGracia            -0.1130884  0.0172771  -6.546 6.15e-11 ***
```

```
## distritosHorta              -0.2289074  0.0256379  -8.928  < 2e-16 ***
## distritosL'Eixample         -0.0252300  0.0123899  -2.036 0.041738 *
## distritosLes Corts          -0.2439854  0.0318380  -7.663 1.94e-14 ***
## distritosNou Barris         -0.1929163  0.0185666 -10.391  < 2e-16 ***
## distritosSant Andreau       -0.3649992  0.0347123 -10.515  < 2e-16 ***
## distritosSant Martí         -0.1172780  0.0210006  -5.585 2.39e-08 ***
## distritosSants-Montjuic     -0.1938114  0.0152824 -12.682  < 2e-16 ***
## distritosSarriá             -0.0284569  0.0249955  -1.138 0.254941
## tipo_habitacionPrivate room -0.5408822  0.0143281 -37.750  < 2e-16 ***
## tipo_habitacionShared room  -1.1176077  0.0763820 -14.632  < 2e-16 ***
## personas                     0.1229541  0.0034679  35.455  < 2e-16 ***
## grupo_baniosPocos           -0.1539298  0.0322642  -4.771 1.85e-06 ***
## grupo_habitacionGrande       0.0467433  0.0151236   3.091 0.002001 **
## estancia_min                -0.0110247  0.0003728 -29.576  < 2e-16 ***
## puntuacion                   0.0021647  0.0004465   4.849 1.26e-06 ***
## TV1                          0.0870508  0.0183188   4.752 2.04e-06 ***
## Wifi1                       -0.0969704  0.0185215  -5.236 1.67e-07 ***
## Air_conditioning1            0.1500005  0.0108430  13.834  < 2e-16 ***
## Elevator1                    0.0410973  0.0094742   4.338 1.45e-05 ***
## Breakfast1                   0.0674654  0.0184538   3.656 0.000257 ***
## Pets_allowed1                0.0387469  0.0134520   2.880 0.003978 **
## Patio_or_balcony1            0.0336376  0.0098685   3.409 0.000655 ***
## check_in_24_hs1              0.1158920  0.0132350   8.756  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.474 on 12815 degrees of freedom
## Multiple R-squared:  0.6065, Adjusted R-squared:  0.6058
## F-statistic:   823 on 24 and 12815 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

summary(mod0)$sigma
```

```
## [1] 0.4792147
```

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

```
## [1] 0.4740483
```

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

```
## [1] 0.5982583
```

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

```
## [1] 0.6065025
```

```
##  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? ¿Habrá que sacar la v
```

Resultados

ACÁ IRÍA LAS PREDICCIONES Y LAS INTERPRETACIONES DEL MODELO

Conclusiones

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