

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**

Junio 2024

Trabajo final de Modelos Lineales

Índice

Introducción	3
Los datos	5
Análisis exploratorio	10
Metodología	12
Selección de variables	12
Diagnóstico	14
Multicolinealidad	15
Linealidad #RESOLVER, EL SUPUESTO SE VERIFICA ASÍ??	16
Atípicos	21
Resultados	26
Conclusiones	26
Bibliografía	26

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'
```

```
head(df)
```

```
## # A tibble: 6 x 26
##   id host_id barrio      cod_postal latitud longitud tipo_habitacion personas
##   <dbl> <dbl> <chr>         <dbl>   <dbl>   <dbl> <chr>             <dbl>
## 1 18666  71615 Sant Marta      8026    41.4    2.19 Entire home/apt         6
## 2 18674  71615 La Sagrada~      8025    41.4    2.17 Entire home/apt         8
## 3 21605  82522 Sant Marta      8018    41.4    2.20 Private room           2
## 4 23197  90417 Sant Marta      8930    41.4    2.22 Entire home/apt         6
## 5 25786 108310 Vila de Gr~      8012    41.4    2.16 Private room           2
## 6 31377 134698 Horta-Guin~      8025    41.4    2.17 Private room           2
## # i 18 more variables: banios <dbl>, habitaciones <dbl>, camas <dbl>,
## #   precio_euros <dbl>, estancia_min <dbl>, puntuacion <dbl>, Internet <dbl>,
## #   TV <dbl>, Wifi <dbl>, Air_conditioning <dbl>, Elevator <dbl>,
## #   Breakfast <dbl>, Pets_allowed <dbl>, Cable_TV <dbl>, Pool <dbl>,
## #   Patio_or_balcony <dbl>, check_in_24_hs <dbl>, Smart_lock <dbl>
```

```
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.

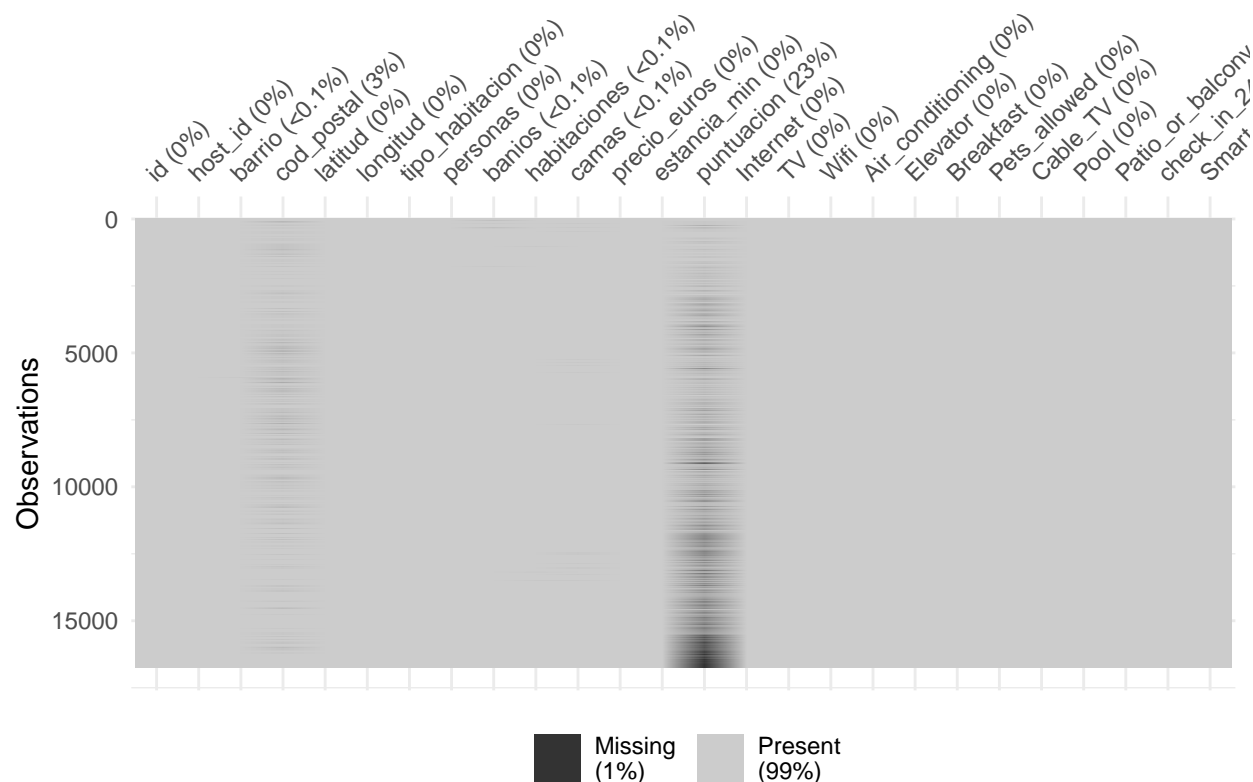
Se decidió prescindir de la latitud y longitud de cada apartamento como de algunas variables que se encontraban dentro de la variable amenities, tomando en cuenta finalmente las diez más #importantes.

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.

```
df %>% vis_miss()
```



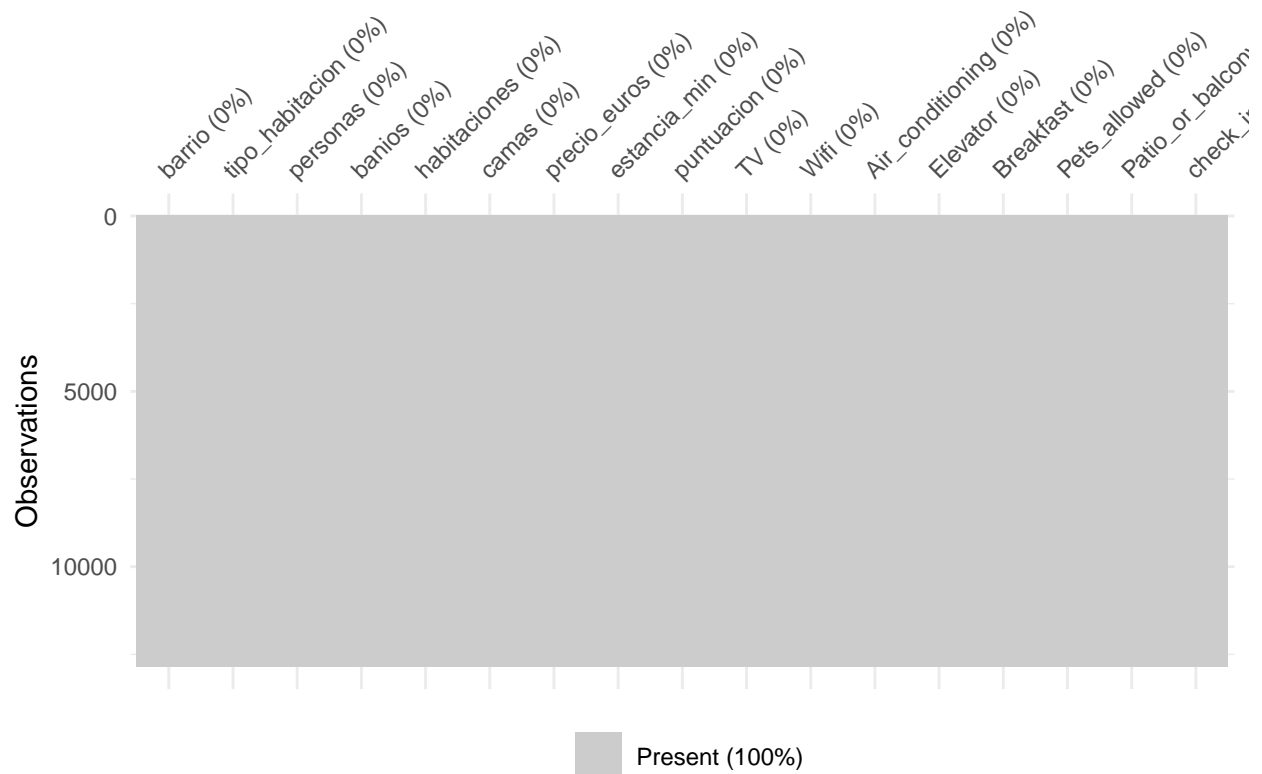
```
df = df %>% select(barrio,
                    tipo_habitacion,
                    personas,
```

```

    banios,
    habitaciones,
    camas,
    precio_euros,
    estancia_min,
    puntuacion,
    TV,
    Wifi,
    Air_conditioning,
    Elevator,
    Breakfast,
    Pets_allowed,
    Patio_or_balcony,
    check_in_24_hs
)%>% filter(is.na(puntuacion)==FALSE,
            is.na(barrio)==FALSE,
            is.na(banios)==FALSE,
            is.na(habitaciones)==FALSE,
            is.na(camas)==FALSE
)

```

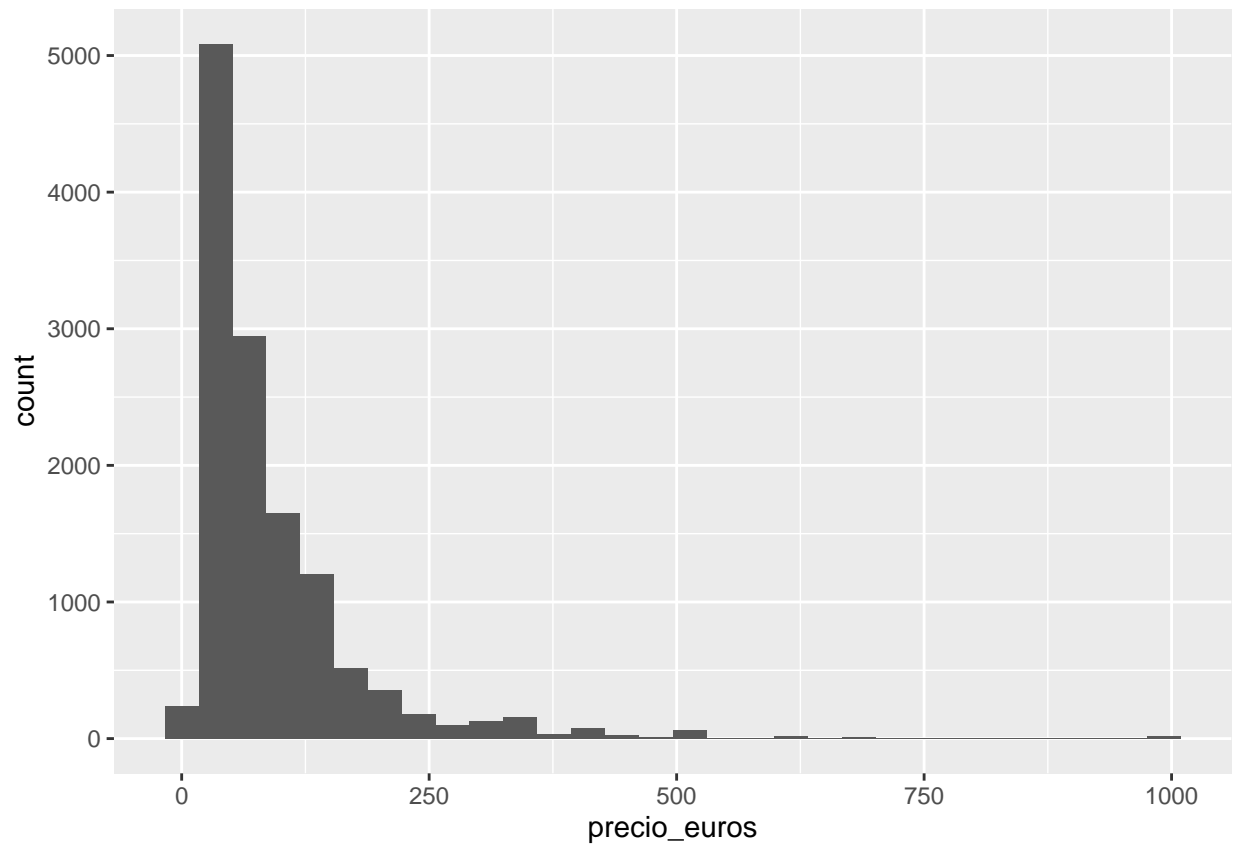
```
df%>%vis_miss()
```



```
#HAY DOS HISTOGRAMAS, CON CUAL NOS QUEDAMOS??? #RESOLVER
```

```
df%>%ggplot()+geom_histogram(aes(x=precio_euros))
```

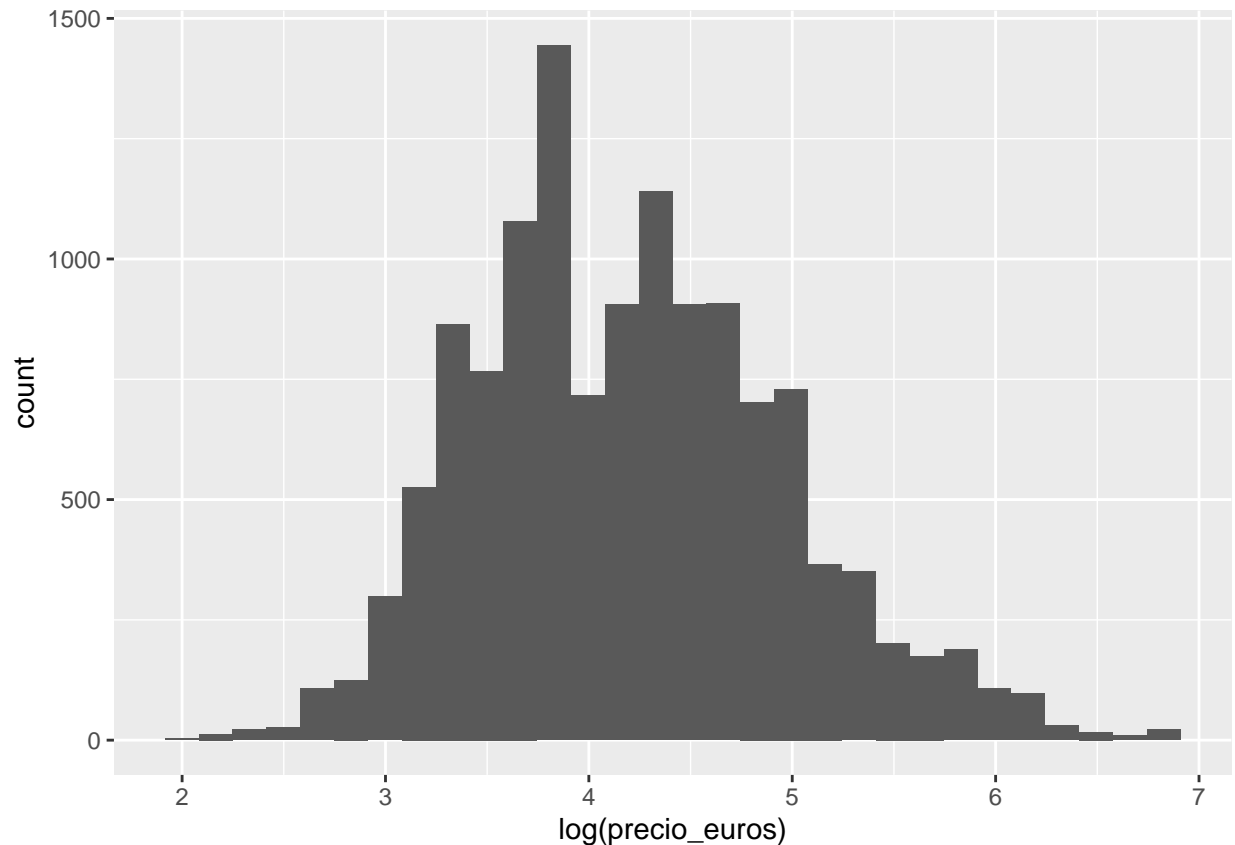
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# Se "normalizan" más los datos con Log(Precio_euros)
```

```
df%>%ggplot()+geom_histogram(aes(x=log(precio_euros)))
```

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



```
distributos<-character(length(nrow(df)))

for (i in seq_along(df$barrio)) {
  if (df$barrio[i] == "Sant Marta" ||
      df$barrio[i] == "La Guineueta - Canyelles" ||
      df$barrio[i] == "La Prosperitat" ||
      df$barrio[i] == "Nou Barris" ||
      df$barrio[i] == "Porta" ||
      df$barrio[i] == "Trinitat Nova" ||
      df$barrio[i] == "Turo de la Peira - Can Peguera" ||
      df$barrio[i] == "Verdum - Los Roquetes" ||
      df$barrio[i] == "Vilapicina i la Torre Llobeta") {
    distributos[i] <- "Nou Barris"
  } else if (df$barrio[i] == "La Sagrada Familia" ||
      df$barrio[i] == "Eixample" ||
      df$barrio[i] == "L'Antiga Esquerra de l'Eixample" ||
      df$barrio[i] == "Sant Antoni" ||
      df$barrio[i] == "Dreta de l'Eixample" ||
      df$barrio[i] == "La Nova Esquerra de l'Eixample" ||
      df$barrio[i] == "el Fort Pienc") {
    distributos[i] <- "L'Eixample"
  } else if (df$barrio[i] == "Vila de Gracia" ||
      df$barrio[i] == "Camp d'en Grassot i Gracia Nova" ||
      df$barrio[i] == "Gracia" ||
      df$barrio[i] == "El Coll" ||
      df$barrio[i] == "La Salut" ||
```



```

        df$barrio[i] == "Vallcarca i els Penitents") {
    distritos[i] <- "Gracia"
} else if (df$barrio[i] == "Horta-Guinarda" ||
df$barrio[i] == "Can Baro" ||
df$barrio[i] == "Carmel" ||
df$barrio[i] == "El Baix Guinardo" ||
df$barrio[i] == "Guinarda" ||
df$barrio[i] == "Horta" ||
df$barrio[i] == "La Font d'en Fargues" ||
df$barrio[i] == "La Teixonera" ||
df$barrio[i] == "La Vall d'Hebron" ||
df$barrio[i] == "Montbau" ||
df$barrio[i] == "Sant Genis dels Agudells") {
    distritos[i] <- "Horta"
} else if (df$barrio[i] == "Les Corts" ||
df$barrio[i] == "La Maternitat i Sant Ramon" ||
df$barrio[i] == "Pedralbes") {
    distritos[i] <- "Les Corts"
} else if (df$barrio[i] == "El Gotic" ||
df$barrio[i] == "La Barceloneta" ||
df$barrio[i] == "Ciutat Vella" ||
df$barrio[i] == "El Raval" ||
df$barrio[i] == "Sant Pere/Santa Caterina" ||
df$barrio[i] == "El Born") {
    distritos[i] <- "Ciutat Vella"
} else if (df$barrio[i] == "El Poble-sec" ||
df$barrio[i] == "Sants-Montjuic") {
    distritos[i] <- "Sants-Montjuic"
} else if (df$barrio[i] == "El Clot" ||
df$barrio[i] == "El Besos i el Maresme" ||
df$barrio[i] == "El Camp de l'Arpa del Clot" ||
df$barrio[i] == "El Poblenou" ||
df$barrio[i] == "La Vila Olimpica" ||
df$barrio[i] == "Diagonal Mar - La Mar Bella" ||
df$barrio[i] == "Glaries - El Parc" ||
df$barrio[i] == "La Verneda i La Pau" ||
df$barrio[i] == "Provincals del Poblenou" ||
df$barrio[i] == "Sant Marta de Provincals"
) {
    distritos[i] <- "Sant Martí"
} else if (df$barrio[i] == "Sant Gervasi - Galvany" ||
df$barrio[i] == "El Putget i Farro" ||
df$barrio[i] == "Les Tres Torres" ||
df$barrio[i] == "Sant Gervasi - la Bonanova" ||
df$barrio[i] == "Sarria" ||
df$barrio[i] == "Sarria-Sant Gervasi") {
    distritos[i] <- "Sarrià"
} else if (df$barrio[i] == "El Bon Pastor" ||
df$barrio[i] == "El Congres i els Indians" ||
df$barrio[i] == "La Sagrera" ||
df$barrio[i] == "La Trinitat Vella" ||
df$barrio[i] == "Navas" ||
df$barrio[i] == "Sant Andreu" ||

```

```

        df$barrio[i] == "Sant Andreu de Palomar") {
      distritos[i] <- "Sant Andreu"
    }
  }
df$distritos <- distritos

#Pasamos las variables a factor y definimos dos nuevas variables "grupo_habitacion" y "grupo_banios" qu

df_final= df%>% mutate(tipo_habitacion=as.factor(tipo_habitacion),
                      camas=as.factor(camas),
                      distritos=as.factor(distritos),
                      Wifi=as.factor(Wifi),
                      TV=as.factor(TV),
                      Air_conditioning=as.factor(Air_conditioning),
                      Elevator=as.factor(Elevator),
                      Breakfast=as.factor(Breakfast),
                      Pets_allowed=as.factor(Pets_allowed),
                      Patio_or_balcony=as.factor(Patio_or_balcony),
                      check_in_24_hs=as.factor(check_in_24_hs),
                      grupo_habitacion=as.factor(ifelse(habitaciones>=0 & habitaciones<3,'Chica',ifelse
                      grupo_banios=as.factor(ifelse(banios>=0 & banios<2,'Pocos',ifelse(banios>=2 & ban

)

df_final

```

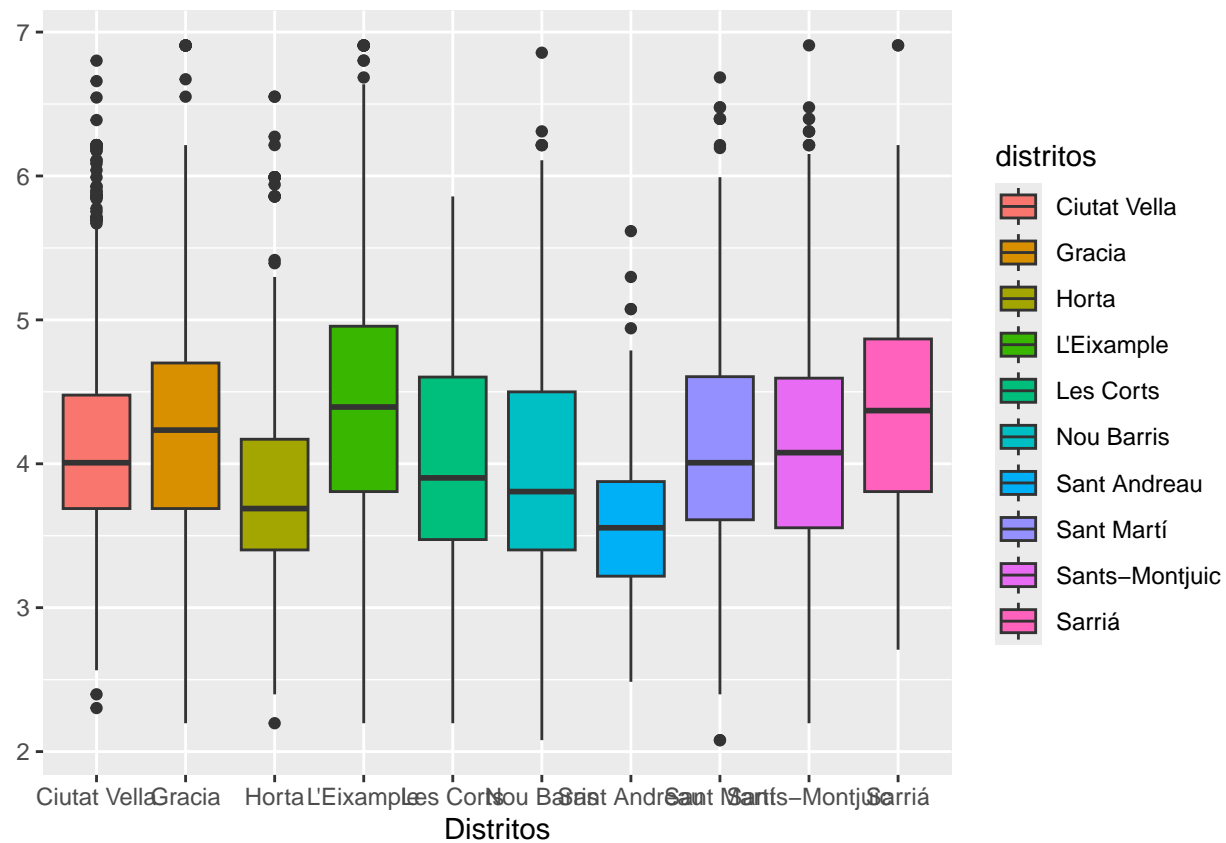
```

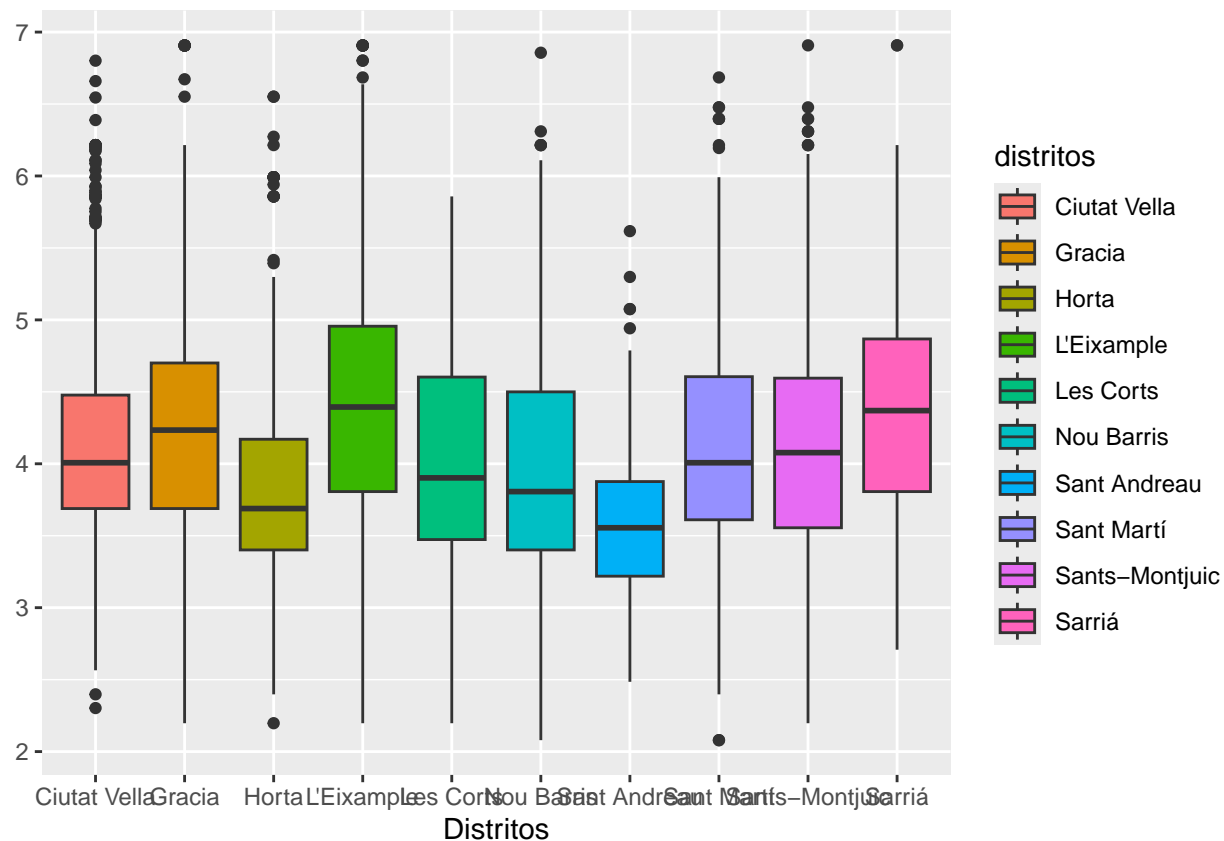
## # A tibble: 12,848 x 20
##   barrio      tipo_habitacion personas banios habitaciones camas precio_euros
##   <chr>      <fct>          <dbl>  <dbl>          <dbl> <fct>      <dbl>
## 1 Sant Marta Entire home/apt      6    1            2 4          130
## 2 La Sagrada F~ Entire home/apt      8    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.





#RESOLVER, QUE GRÁFICOS VAMOS A HACER???: De los anteriores gráficos lo primero que se puede apreciar es que hay muchos datos atípicos, situación en la que nos enfocaremos con profundidad más adelante. Además podríamos decir que existe diferencia entre el precio de los apartamentos dependiendo de si cuentan o no con aire acondicionado, pero adicional de lo ya mencionado no se puede interpretar con seguridad mucho más debido a que muchas de las cajas se solapan entre si. Esto se puede contemplar bien en el último gráfico, el cual toma como variable explicativa a los distritos, en este, todas las cajas tienen aproximadamente las mismas alturas además de las misma mediana, lo que si se puede observar que cambia distrito a distrito es la varianza, siendo la del distrito Sant Andreu la menor de todas.

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      88.20     9   42.9897 < 2.2e-16 ***
## tipo_habitacion 323.26     2  708.9869 < 2.2e-16 ***
## personas      247.62     1 1086.2028 < 2.2e-16 ***
## grupo_banios    25.61     2   56.1705 < 2.2e-16 ***
## grupo_habitacion  2.22     2    4.8657 0.0077204 **
## estancia_min   150.60     1  660.5900 < 2.2e-16 ***
## puntuacion      5.43     1   23.8201 1.070e-06 ***
## TV              4.74     1   20.7900 5.172e-06 ***
## Wifi           6.50     1   28.5037 9.512e-08 ***
## Air_conditioning 44.04     1  193.1834 < 2.2e-16 ***
## Elevator        3.24     1   14.1975 0.0001653 ***
## Breakfast       3.66     1   16.0574 6.180e-05 ***
## Pets_allowed    2.36     1   10.3720 0.0012826 **
## Patio_or_balcony 3.01     1   13.2017 0.0002808 ***
## check_in_24_hs  16.79     1   73.6708 < 2.2e-16 ***
## Residuals      2922.81 12821
## ---
## 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
## -2.8841 -0.3012 -0.0187  0.2780  3.3291
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.9633545   0.0478668   82.800 < 2e-16 ***
## distritosGracia -0.1000210   0.0173891  -5.752 9.02e-09 ***
## distritosHorta  -0.2150816   0.0258245  -8.329 < 2e-16 ***
## distritosL'Eixample -0.0180793   0.0124689  -1.450 0.147097
## distritosLes Corts -0.2547148   0.0319912  -7.962 1.83e-15 ***
## distritosNou Barris -0.1832534   0.0186913  -9.804 < 2e-16 ***
## distritosSant Andreu -0.3433516   0.0348821  -9.843 < 2e-16 ***
## distritosSant Martí -0.1052671   0.0211457  -4.978 6.50e-07 ***
## distritosSants-Montjuic -0.1785258   0.0153696 -11.616 < 2e-16 ***
## distritosSarrià    -0.0208862   0.0251468  -0.831 0.406232
## tipo_habitacionPrivate room -0.5242062   0.0143965 -36.412 < 2e-16 ***
## tipo_habitacionShared room -1.1557285   0.0758798 -15.231 < 2e-16 ***
```

```
## personas          0.1218599  0.0036975  32.958 < 2e-16 ***
## grupo_baniosMuchos 0.3818688  0.1262669   3.024 0.002497 **
## grupo_baniosPocos -0.1225746  0.0120703 -10.155 < 2e-16 ***
## grupo_habitacionGrande -0.1829563  0.0723115  -2.530 0.011414 *
## grupo_habitacionMediana 0.0151220  0.0156211   0.968 0.333038
## estancia_min      -0.0081306  0.0003163 -25.702 < 2e-16 ***
## puntuacion        0.0021954  0.0004498   4.881 1.07e-06 ***
## TV1               0.0841128  0.0184474   4.560 5.17e-06 ***
## Wifi1            -0.0996002  0.0186556  -5.339 9.51e-08 ***
## Air_conditioning1  0.1519067  0.0109293  13.899 < 2e-16 ***
## Elevator1         0.0360303  0.0095623   3.768 0.000165 ***
## Breakfast1        0.0743994  0.0185666   4.007 6.18e-05 ***
## Pets_allowed1      0.0436092  0.0135408   3.221 0.001283 **
## Patio_or_balcony1  0.0360932  0.0099337   3.633 0.000281 ***
## check_in_24_hs1    0.1144320  0.0133321   8.583 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4775 on 12821 degrees of freedom
## Multiple R-squared:  0.6013, Adjusted R-squared:  0.6004
## F-statistic: 743.5 on 26 and 12821 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. #RESPUESTA A LA PREGUNTA: VOLVER A CALCULAR 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 INTERPRETAR # (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 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. #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 líneas del script se pusieron a prueba cada uno de los supuestos antes mencionados.

Multicolinealidad

```
library(car)

X <- model.matrix(mod0)
XX <- t(X)%*%X
eigen_result <- eigen(XX) #valores y vectores propios de XX

# Obtener los valores propios
valores_propios <- eigen_result$values
print(valores_propios)

## [1] 1.081922e+08 2.487577e+06 7.043226e+04 4.945386e+03 3.874283e+03
## [6] 2.368994e+03 2.211754e+03 2.148727e+03 1.913668e+03 1.427351e+03
## [11] 1.340412e+03 1.233442e+03 1.122827e+03 9.732000e+02 9.376122e+02
## [16] 7.472145e+02 6.563968e+02 5.301557e+02 4.025563e+02 3.643128e+02
## [21] 3.025895e+02 2.230926e+02 1.578035e+02 1.009854e+02 4.728253e+01
## [26] 3.906690e+01 1.387576e+01

# numero de condicion
kA <- sqrt(max(valores_propios)/min(valores_propios))
print(kA)

## [1] 2792.348

vif(mod0)

##               distritosGracia               distritosHorta
##               1.292207               1.117259
## distritosL'Eixample               distritosLes Corts
##               2.011038               1.083230
## distritosNou Barris               distritosSant Andreu
##               1.264163               1.071535
## distritosSant Martí               distritosSants-Montjuic
##               1.201583               1.395021
## distritosSarrià tipo_habitacionPrivate room
##               1.147664               2.912569
## tipo_habitacionShared room               personas
##               1.032213               3.712360
## grupo_baniosMuchos               grupo_baniosPocos
##               1.117580               1.399095
## grupo_habitacionGrande               grupo_habitacionMediana
##               1.415227               1.914105
## estancia_min               puntuacion
##               1.093229               1.046659
## TV1               Wifi1
##               4.063336               3.710309
## Air_conditioning1               Elevator1
##               1.647713               1.205875
## Breakfast1               Pets_allowed1
##               1.047822               1.020872
## Patio_or_balcony1               check_in_24_hs1
##               1.042744               1.076293
```

```
#el número de condición es muy alto, lo que nos dice que hay problemas de multicolinealidad, mirando lo
#importante recordar que no estamos estimando el precio en euros, sino que el logaritmo del precio en e

## VER SCRIPTS DE CLASE
```

Linealidad #RESOLVER, EL SUPUESTO SE VERIFICA ASÍ??

```
res_ext<-rstudent(mod0) #residuos
res_int<-rstandard(mod0) #residuos

yhat<-fitted(mod0) #ygorro
#grafico

df_final$predichos <- yhat #agrego los residuos a la tabla
df_final$residuos <- res_ext #agrego los ygorro a la tabla
df_final$residuos_int <- res_int #agrego los ygorro a la tabla

ggplot(df_final, aes(x = predichos , y = residuos)) +
  geom_point() +
  geom_hline(yintercept = 0)
#en el gráfico se puede ver que la mayoría de los puntos se encuentran "agrupados", (Eso cuenta como un

#Homoscedasticidad
install.packages("skedastic")

breusch_pagan(mod0)
#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 HOMOSCEDASTICIDAD, p valor muy chico.

### Normalidad
# histogramas
hist(rstudent(mod0)) #OJO es sensible al tamaño de las barras (bins) por lo que se puede interpretar de
#nos gustaría ver que tenga la forma de una campana simetrica, en este caso tiene más o menos la forma,

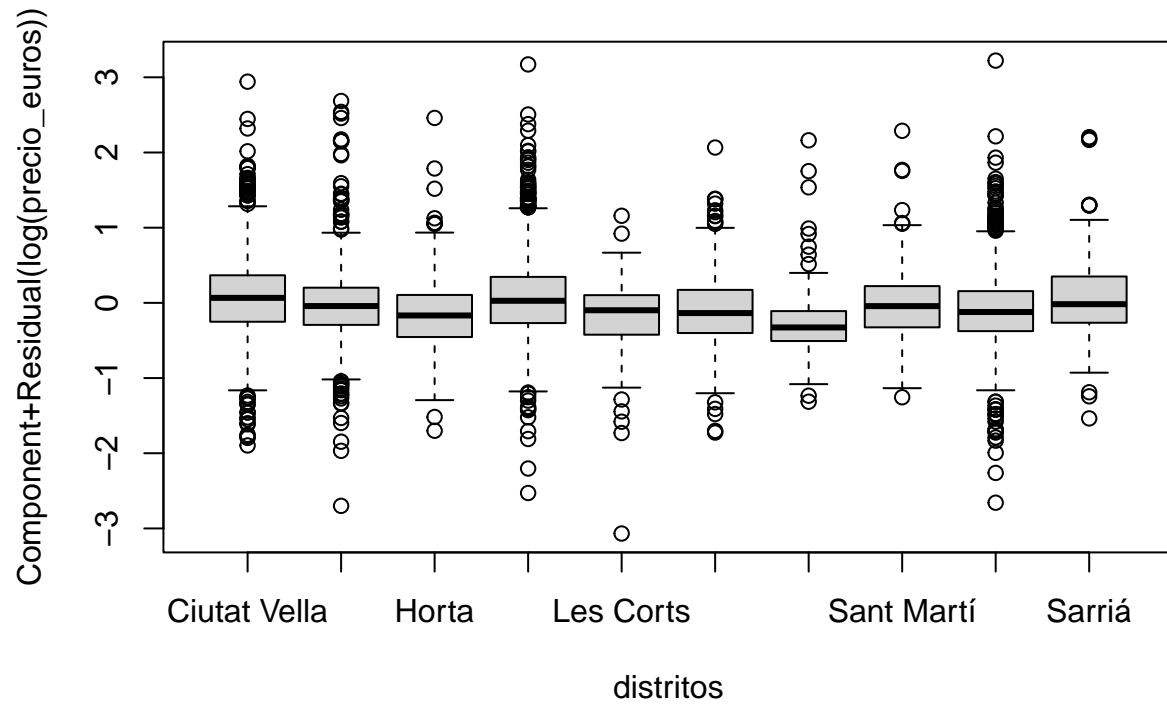
#Q-Q plot
qqPlot(res_ext)
plot(density(df_final$residuos))
#los puntos no están sobre la recta, nos hay normalidad

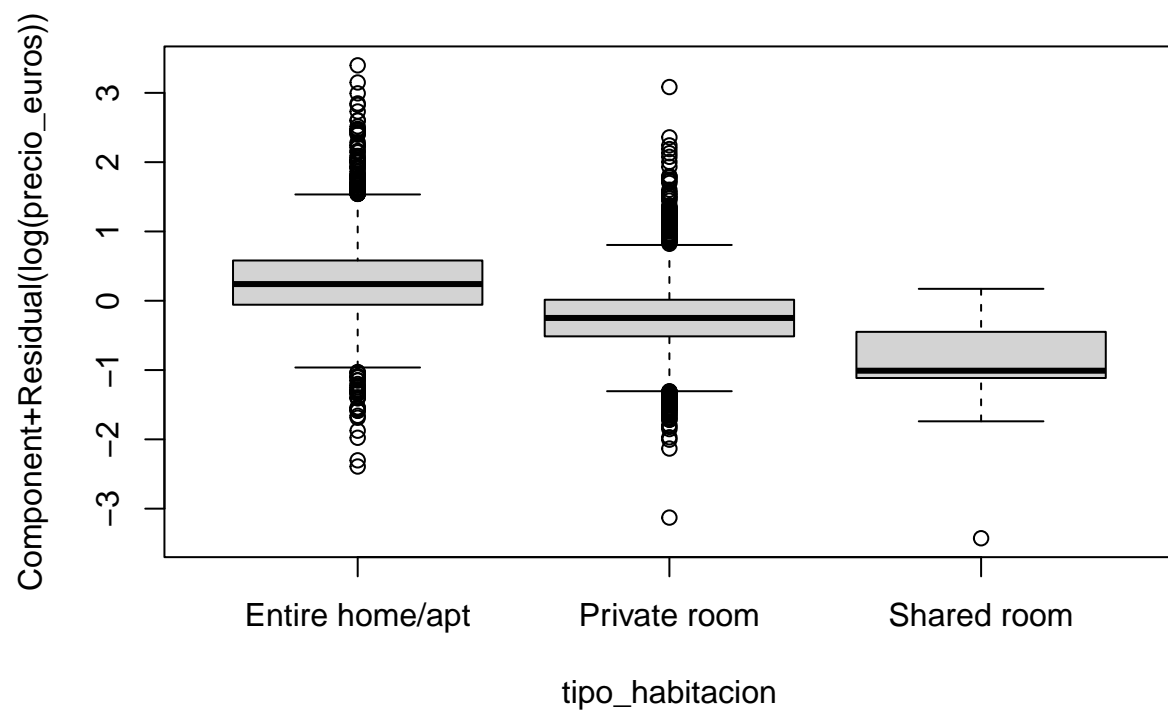
ggplot(data = df_final, aes(sample = residuos_int)) +
  stat_qq_band(fill = 2) +
  stat_qq_line(col = 2) +
  stat_qq_point() +
  xlab("Cuantiles teoricos")+
  ylab("Cuantiles empiricos")
```

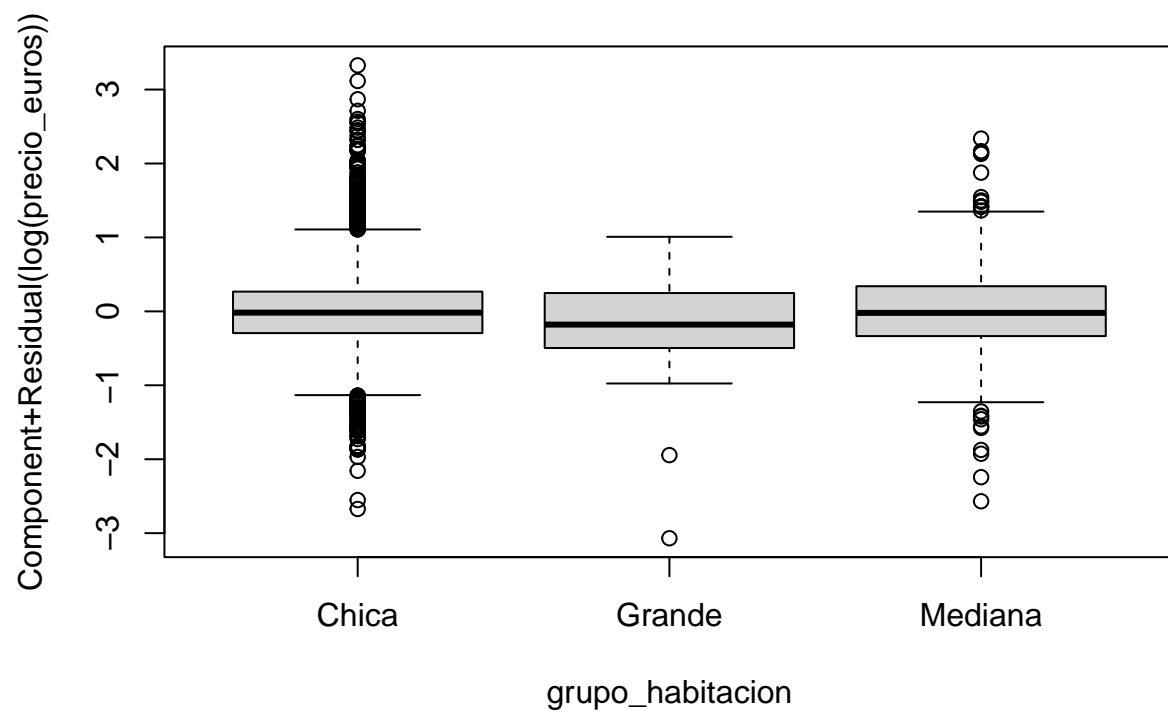


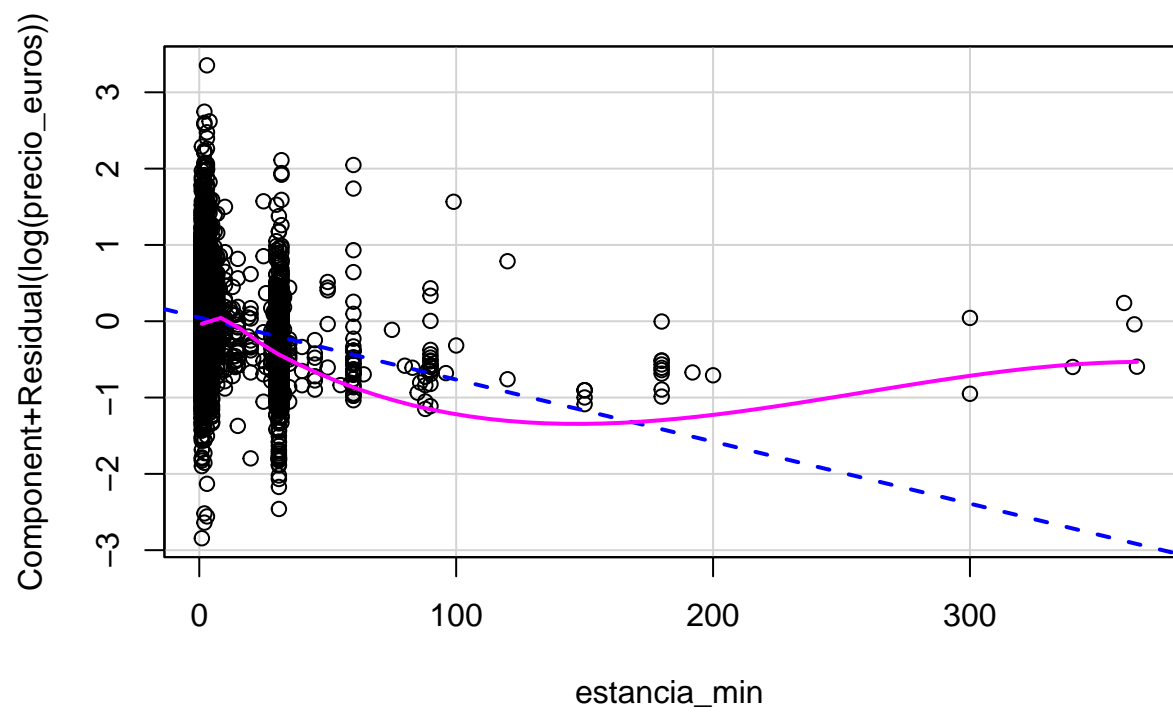
```
#DATOS FUERA DE LA BANDA. SE PUEDE VER QUE NO HAY NORMALIDAD.  
#Randomize
```

```
#NO FUNCIONA #RESOLVER  
ks.test(df_final$residuos_int)
```

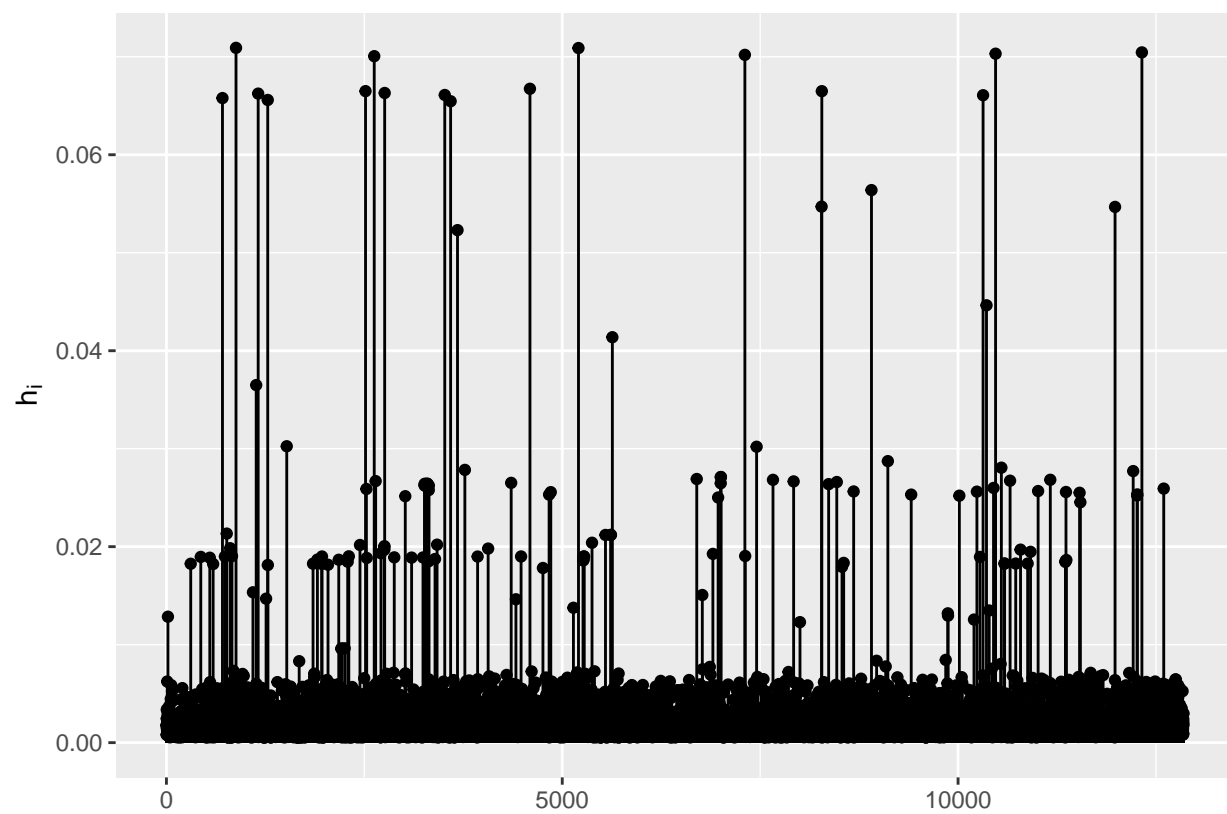
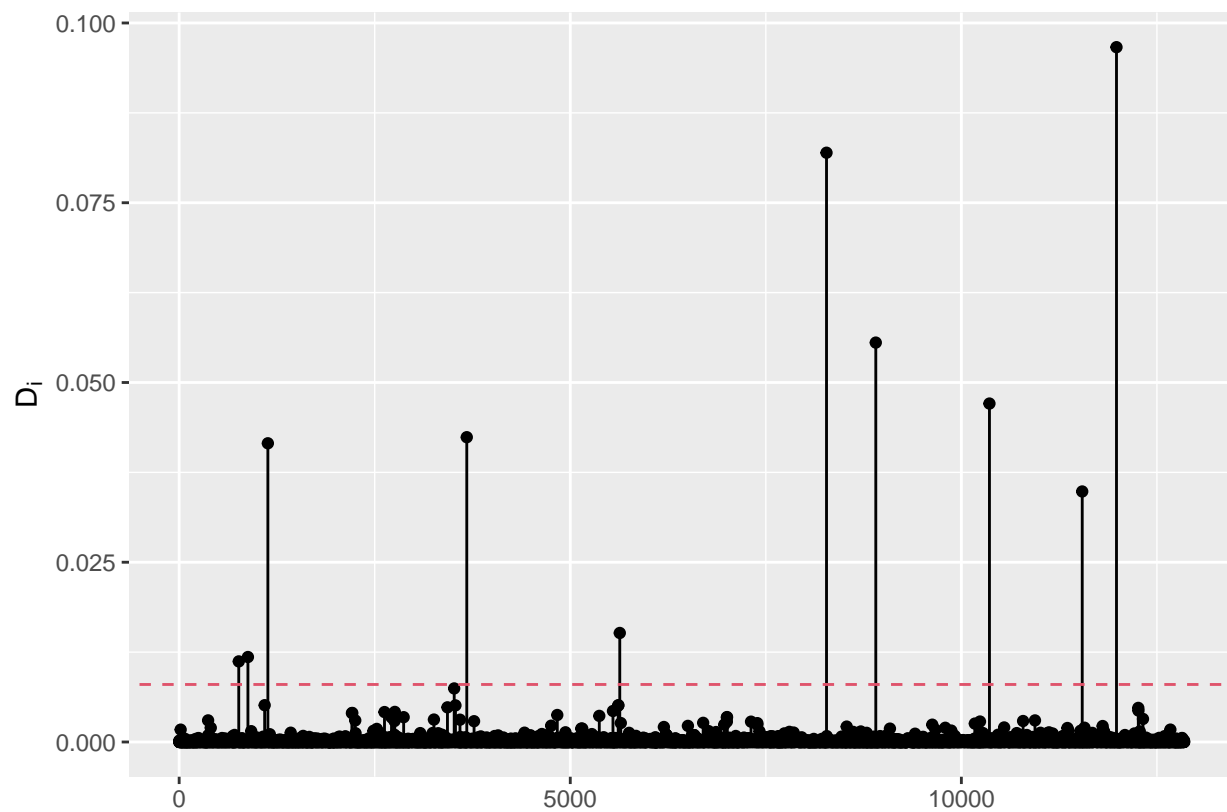








Atípicos



```
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	89.26	9	44.5209	< 2.2e-16 ***
tipo_habitacion	332.75	2	746.8677	< 2.2e-16 ***
personas	234.98	1	1054.8425	< 2.2e-16 ***
grupo_banios	26.59	2	59.6907	< 2.2e-16 ***
grupo_habitacion	0.80	2	1.7950	0.1661622
estancia_min	196.20	1	880.7531	< 2.2e-16 ***
puntuacion	4.98	1	22.3720	2.270e-06 ***
TV	4.55	1	20.4293	6.243e-06 ***
Wifi	6.07	1	27.2618	1.805e-07 ***
Air_conditioning	42.43	1	190.4628	< 2.2e-16 ***
Elevator	3.03	1	13.6001	0.0002271 ***
Breakfast	2.86	1	12.8529	0.0003382 ***
Pets_allowed	2.01	1	9.0240	0.0026697 **
Patio_or_balcony	2.23	1	9.9964	0.0015721 **
check_in_24_hs	16.34	1	73.3519	< 2.2e-16 ***
Residuals	2853.80	12811		

```
## ---
## 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
	-2.6807	-0.2990	-0.0191	0.2756	3.3284

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.0180495	0.0475550	84.493	< 2e-16 ***
distritosGracia	-0.1101168	0.0172051	-6.400	1.61e-10 ***
distritosHorta	-0.2224536	0.0255336	-8.712	< 2e-16 ***
distritosL'Eixample	-0.0307718	0.0123557	-2.491	0.012769 *
distritosLes Corts	-0.2445441	0.0316860	-7.718	1.27e-14 ***

```
## distritosNou Barris      -0.1911655  0.0184934 -10.337 < 2e-16 ***
## distritosSant Andreu    -0.3592169  0.0345652 -10.392 < 2e-16 ***
## distritosSant Martí     -0.1152925  0.0209133  -5.513 3.60e-08 ***
## distritosSants-Montjuic -0.1886021  0.0152250 -12.388 < 2e-16 ***
## distritosSarrià         -0.0334155  0.0248958  -1.342 0.179550
## tipo_habitacionPrivate room -0.5433593  0.0144086 -37.711 < 2e-16 ***
## tipo_habitacionShared room -1.1149013  0.0760785 -14.655 < 2e-16 ***
## personas                0.1200755  0.0036971  32.478 < 2e-16 ***
## grupo_baniosMuchos      0.3916356  0.1296464   3.021 0.002526 **
## grupo_baniosPocos      -0.1253912  0.0119362 -10.505 < 2e-16 ***
## grupo_habitacionGrande  -0.1052468  0.0725983  -1.450 0.147163
## grupo_habitacionMediana  0.0113039  0.0154732   0.731 0.465067
## estancia_min            -0.0110061  0.0003709 -29.677 < 2e-16 ***
## puntuacion              0.0021037  0.0004448   4.730 2.27e-06 ***
## TV1                     0.0824792  0.0182481   4.520 6.24e-06 ***
## Wifi1                   -0.0963425  0.0184519  -5.221 1.80e-07 ***
## Air_conditioning1       0.1491887  0.0108101  13.801 < 2e-16 ***
## Elevator1               0.0348727  0.0094562   3.688 0.000227 ***
## Breakfast1              0.0658846  0.0183774   3.585 0.000338 ***
## Pets_allowed1           0.0402363  0.0133943   3.004 0.002670 **
## Patio_or_balcony1       0.0310750  0.0098285   3.162 0.001572 **
## check_in_24_hs1        0.1129047  0.0131828   8.565 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.472 on 12811 degrees of freedom
## Multiple R-squared:  0.61, Adjusted R-squared:  0.6092
## F-statistic: 770.8 on 26 and 12811 DF, p-value: < 2.2e-16
```

```
mod2= lm(log(precio_euros) ~ distritos + tipo_habitacion+personas+ grupo_banios+estancia_min+puntuacion
Anova(mod2)
```

```
## 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	89.03	9	44.4023	< 2.2e-16 ***
## tipo_habitacion	346.01	2	776.5394	< 2.2e-16 ***
## personas	336.52	1	1510.4818	< 2.2e-16 ***
## grupo_banios	28.18	2	63.2455	< 2.2e-16 ***
## estancia_min	197.84	1	888.0313	< 2.2e-16 ***
## puntuacion	4.93	1	22.1237	2.583e-06 ***
## TV	4.55	1	20.4192	6.276e-06 ***
## Wifi	6.04	1	27.1237	1.938e-07 ***
## Air_conditioning	42.87	1	192.4429	< 2.2e-16 ***
## Elevator	3.04	1	13.6377	0.0002226 ***
## Breakfast	2.83	1	12.7205	0.0003630 ***
## Pets_allowed	2.00	1	8.9645	0.0027579 **
## Patio_or_balcony	2.33	1	10.4554	0.0012260 **

```
## check_in_24_hs      16.36      1    73.4397 < 2.2e-16 ***
## Residuals          2854.60 12813
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = log(precio_euros) ~ distritos + tipo_habitacion +
##      personas + grupo_banios + 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
## -2.6830 -0.2996 -0.0197  0.2759  3.3287
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.0262897   0.0471773   85.344 < 2e-16 ***
## distritosGracia  -0.1095743   0.0172026   -6.370 1.96e-10 ***
## distritosHorta    -0.2217640   0.0255237   -8.689 < 2e-16 ***
## distritosL'Eixample -0.0305357   0.0123486   -2.473 0.013419 *
## distritosLes Corts  -0.2434376   0.0316538   -7.691 1.57e-14 ***
## distritosNou Barris -0.1906434   0.0184919  -10.310 < 2e-16 ***
## distritosSant Andreu -0.3584802   0.0345500  -10.376 < 2e-16 ***
## distritosSant Martí  -0.1154214   0.0209098   -5.520 3.46e-08 ***
## distritosSants-Montjuic -0.1882363   0.0152200  -12.368 < 2e-16 ***
## distritosSarrià     -0.0321161   0.0248721   -1.291 0.196640
## tipo_habitacionPrivate room -0.5471541   0.0142209  -38.475 < 2e-16 ***
## tipo_habitacionShared room -1.1189726   0.0760345  -14.717 < 2e-16 ***
## personas           0.1194580   0.0030737   38.865 < 2e-16 ***
## grupo_baniosMuchos   0.3254496   0.1241874    2.621 0.008787 **
## grupo_baniosPocos    -0.1282330   0.0116710  -10.987 < 2e-16 ***
## estancia_min        -0.0110309   0.0003702  -29.800 < 2e-16 ***
## puntuacion           0.0020916   0.0004447    4.704 2.58e-06 ***
## TV1                  0.0824631   0.0182491    4.519 6.28e-06 ***
## Wifi1               -0.0960964   0.0184515   -5.208 1.94e-07 ***
## Air_conditioning1     0.1495044   0.0107771   13.872 < 2e-16 ***
## Elevator1            0.0349077   0.0094526    3.693 0.000223 ***
## Breakfast1           0.0655451   0.0183776    3.567 0.000363 ***
## Pets_allowed1         0.0400960   0.0133917    2.994 0.002758 **
## Patio_or_balcony1     0.0317594   0.0098221    3.233 0.001226 **
## check_in_24_hs1       0.1129034   0.0131747    8.570 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.472 on 12813 degrees of freedom
## Multiple R-squared:  0.6099, Adjusted R-squared:  0.6092
## F-statistic: 834.8 on 24 and 12813 DF, p-value: < 2.2e-16
```

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. 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.

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

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.

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 #RESOLVER

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 línea roja quedaba muy baja, haciendo referencia a que todas las osbervaciones eran atípicas, para poder interpretar mejor esto se decidió que la línea esté en el nivel de $4/300$, en ese punto las observaciones atípicas terminan siendo las más diferentes.

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.

```
coef(mod0)
summary(mod0)

#CUAL ES LA LIBRERIA??
#forward
library(mixlm)
modF <- forward(mod0, alpha = 0.05)
length(coef(modF)) # parametros
summary(modF)

# backward
modB <- backward(mod0, alpha = 0.01)
length(coef(modB)) # parametros
summary(modB)

# stepwise
modS <- stepWise(mod0, alpha.enter = 0.04, alpha.remove=0.05)
length(coef(modS)) # parametros
summary(modS)

#COMPARAMOS POR AIC, BIC O R2 AJUSTADO PARA VER CUAL ES EL MEJOR

#que paquete se usaba para ver la tabla con AIC BIC
install.packages("HH")
library(HH)
summaryHH(modF)
```

```
summaryHH(modB)
```

```
help(anova)
```

```
summaryHH(modS)
```

Resultados

Conclusiones

Bibliografía