Practica 2

Autors: Mireia Gutiérrez i Xavi Sellart

Assignatura: Tipologia i cicle de vida de les dades

Curs: 2021-2022

Índex

1. Descripció del dataset

El conjunt de dades conté dades de la majoria de les propietats importants dels pisos que es trobaven disponibles a la web d'Habitaclia, com ara la superfície del pis, el preu de compra, el barri on està ubicat el pis, nombre d'habitacions i lavabos, classificació energètica... També s'inclouen dos atributs més generals, que són "Característiques generals" i "Equipament comunitaris" en les que es llisten les característiques que no estan incloses en la resta d'atributs del joc de dades. S'ha realitzat un procés de neteja al joc de dades per tal de corregir incoherències i problemes de format de les dades.

L'objectiu del treball és analitzar el sector immobiliari a Barcelona. Aquest útils anys, a arrel de la pandèmia i del teletreball, han canviat les preferències immobiliàries de gran part de la població. A partir d'aquest estudi es vol veure quines són les característiques que mes influeixen en determinar el preu d'un pis i quines són les principals diferencies entre pisos de diferents districtes de la ciutat de Barcelona.

2. Integració i selecció de les dades d'interès a analitzar

S'ha fet una subselecció útil de les dades originals en base a l'objectiu que es vol aconseguir. Primer de tot, s'han eliminar aquells registres que estaven repetits en el joc de dades. També s'ha eliminat la columna de "Preu m2" de les dades originals, ja que la informació que contenia es podia obtenir a partir de la divisió de les columnes de "Preu" i "Àrea": També s'han eliminat les columnes "Title", "CaracteristicasGenerales", "EquipamientoComunitario": la primera perquè no aportava informació útil per l'anàlisi, i les altres perquè la informació que contenien s'ha transformat mitjançant one-hot-encoding.

Finalment, la columna de "Barri" s'ha substituït per la columna de "Districte", per tal de simplificar l'anàlisi posterior.

3. Preprocessat de les dades

3.1. Eliminació de zeros o elements buits

El joc de dades inicial contenia zeros o elements buits o incongruents. El procediment que s'ha seguit per tractar aquestes dades ha sigut eliminar les files que contenien algun d'aquest tipus d'elements. Això s'ha fet així perquè el joc de dades era gran i un cop eliminades les dades buides es seguia disposant d'un nombre de registres suficient com per realitzar l'estudi. Haver fet algun tractament de les dades buides podria haver introduït algun biaix en el resultat, per això s'ha decidit no fer-ho.

3.2. Identifica i gestiona els valors extrems

El tractament que s'ha fet dels valors extrems ha sigut calcular la mitjana i la desviació típica de cadascuna de les columnes, i s'han eliminat del joc de dades aquells registres que tenien algun atribut que quedava fora del rang de la mitjan més/menys 2 desviacions típiques.

En el cas de la variable Preu, com presenta diferencies significatives entre els valors en funció de molts paràmetres externs com poden ser el barri, l'àrea o si té ascensor o piscina, s'ha decidit establir com a límit inferior 5.000€ i com a superior 15.000.000€. Aquests valors s'han escollit a partir dels valors màxim i mínim de la pàgina d'Habitaclia en la ciutat de Barcelona. Això ens permetrà eliminar possibles errors en la recopilació de les dades, però no perdre registres reals de la base de dades generada.

3.3. Codi de Python

El script amb el codi de Python el trobem al fitxer *neteja_pisos_Barcelona.py* que es troba en el repositori de github.

3.4. Neteja per R Studio

Quan s'ha incorporat la base de dades a R Studio, s'han tingut que realitzar algunes transformacions de les dades per tal de que es treballes amb el tipus de dades adequats pels posteriors anàlisi.

En primer lloc, obrim el fitxer de dades i examinem el tipus de dades amb els que R ha interpretat cada variable.

```
pisos <- read.csv("~/MASTER/Tipologia i cicle de vida de les dades/Practica 2/pisos_barcelona.csv", dec=
str(pisos)
```

```
## 'data.frame':    8185 obs. of  27 variables:
##  $ X                       : int  1 4 6 8 9 10 11 12 13 16 ...
##  $ Barri                   : chr  "Esquerra Baixa de lÂ´Eixample" "Putget - FarrÃ³" "Esquerra Alta de
##  $ Area                    : int  101 114 122 63 94 124 85 65 105 85 ...
##  $ Habitacions             : int  3 3 3 2 3 4 3 3 3 2 ...
##  $ Lavabos                 : int  2 2 2 1 2 2 2 1 2 1 ...
##  $ Preu                    : chr  "410000.0" "550000.0" "650000.0" "289000.0" ...
##  $ Immobiliaria            : chr  "FINQUES GIRAMON" "Prontopiso" "Singular Properties" "Housfy Real Es
##  $ ConsumoEtiqueta         : chr  "G" "E" "E" "E" ...
##  $ ConsumokW               : chr  "999.0" "999.0" "145.0" "169.0" ...
##  $ EmisionesEtiqueta       : chr  "G" "E" "E" "E" ...
##  $ Emisioneskg             : chr  "999.0" "999.0" "30.0" "35.0" ...
##  $ Data.anunci             : chr  "07/04/2022" "21/03/2022" "21/03/2022" "09/04/2022" ...
##  $ Aire_acondicionat       : int  1 1 0 1 0 1 1 1 1 1 ...
##  $ Moblat                  : int  0 0 0 1 0 0 0 0 0 1 ...
##  $ Any_construccio         : int  1936 1979 1947 2002 1966 1969 1972 1996 1968 1920 ...
##  $ Calefaccio              : int  0 1 1 1 1 1 1 1 1 1 ...
##  $ Transport_public_proper: int  0 0 0 1 0 0 0 0 0 0 ...
##  $ Llar_de_foc             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Obra_nova               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Piscina_propia          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Plaza_parking           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Ascensor                : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Equipamient_esportiu    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Jardi_comunitari        : int  0 0 0 0 0 0 0 0 0 0 ...
```

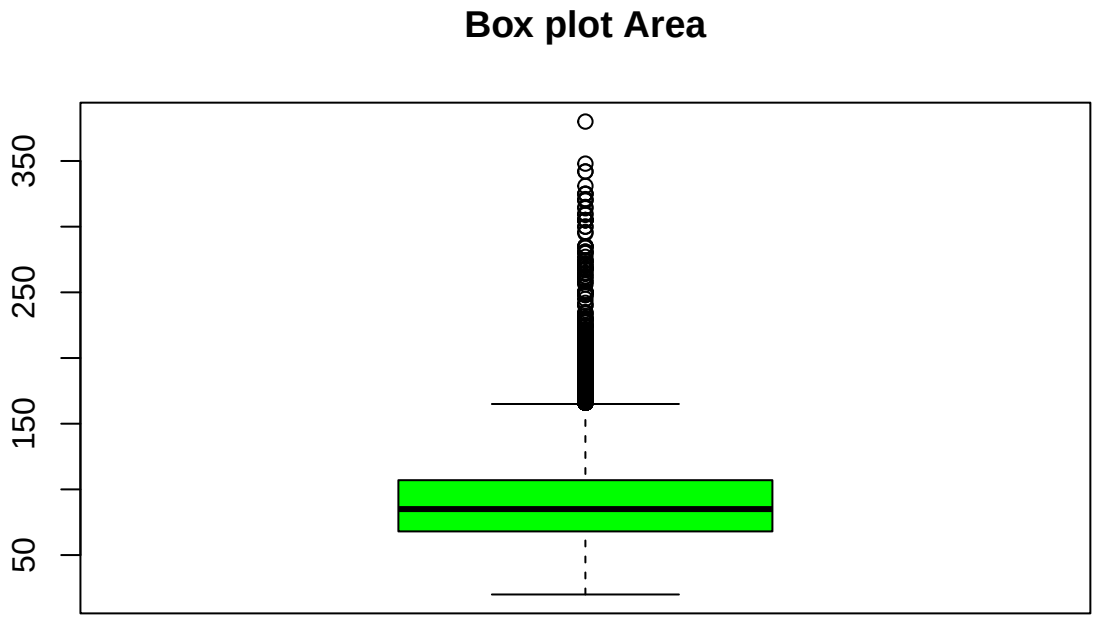
```
##  $ Piscina_comunitaria    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Vigilancia             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Districte              : chr  "Eixample" "Sarria Sant Gervasi" "Eixample" "Sants Montjuic" ...
```

Podem veure que les columnes “Area”, “Preu”, “ConsumokW” i “Emisioneskg” són dades de tipus text quan haurien de ser numèriques, així que realitzem les pertinents transformacions:

```
#Variable Area
pisos$Area <- as.numeric(pisos$Area)
summary(pisos$Area)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   68.00   85.00   93.32  107.00  380.00
```

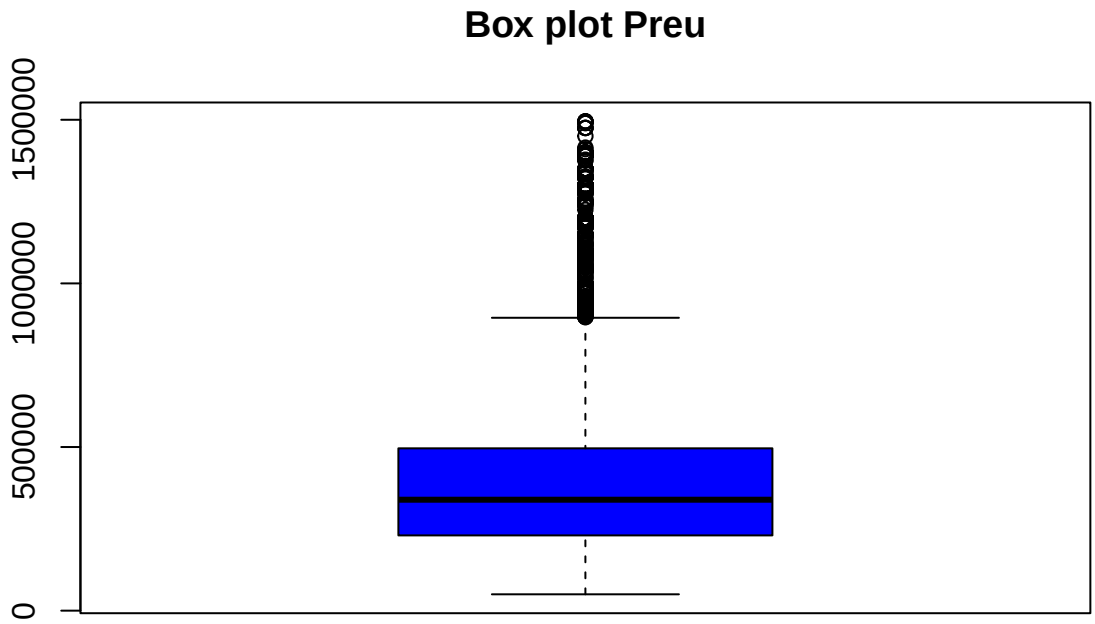
```
boxplot(pisos$Area,main="Box plot Area", col="green")
```




```
#Variable Preu
pisos$Preu <-as.numeric(pisos$Preu)
summary(pisos$Preu)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50000  230000  339000  411453  496000 1495000
```

```
boxplot(pisos$Preu,main="Box plot Preu", col="blue")
```




```
#Variable ConsumokW
pisos$ConsumokW <-as.numeric(pisos$ConsumokW)
summary(pisos$ConsumokW)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   132.0   191.0   395.7   999.0   999.0
```

```
#Variable Emisioneskg
pisos$Emisioneskg <-as.numeric(pisos$Emisioneskg)
summary(pisos$Emisioneskg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    30.0    45.0   289.4   999.0   999.0
```

A partir dels boxplots representats, condirem oportú eliminar els valors atípics en la base de dades:

```
Outliers<-boxplot.stats(pisos$Preu)$stats #Valors atipics
pisos$Preu[pisos$Preu<Outliers[1]]<- NA
pisos$Preu[pisos$Preu>Outliers[5]]<- NA
pisos <- pisos[!is.na(pisos$Preu),]
```

Creem el camp de preu/m2:

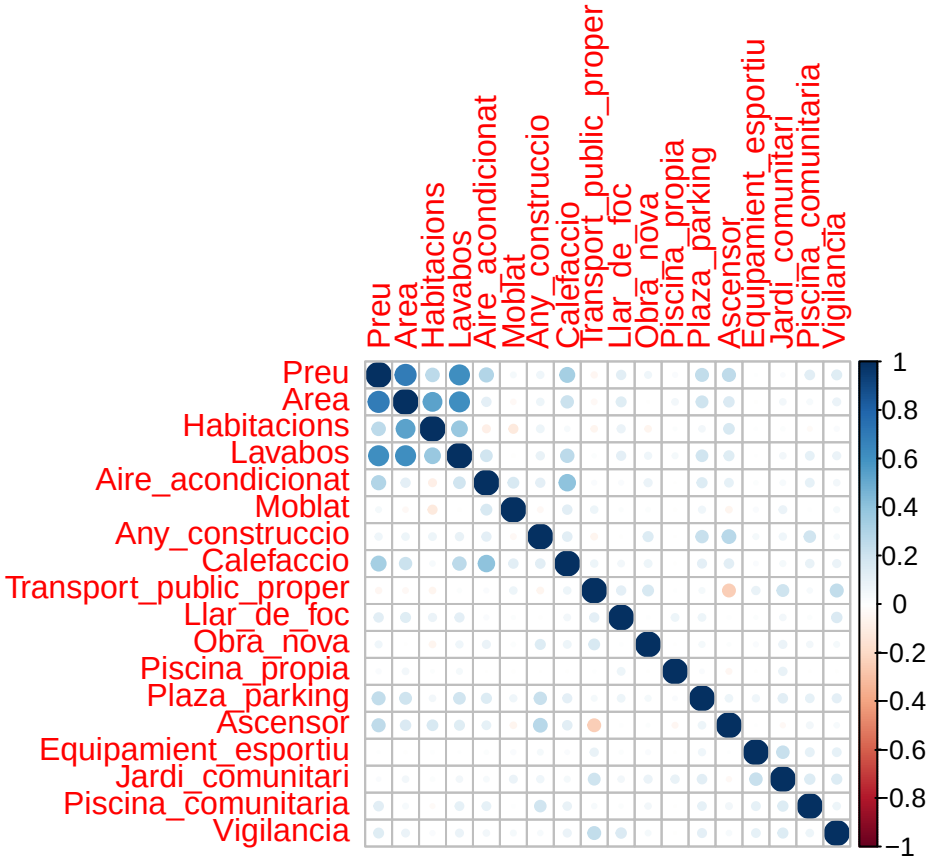
```
pisos$Preu_m2 <- pisos$Preu/pisos$Area
```

També serà interessant comprovar al correlació entre els diferents camps, així que la calculem entre els camps numèrics del data frame abans de convertir-los en factors:

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corr.res<-cor(pisos[,c(6,3,4,5,13,14,15,16,17,18,19,20,21,22,23,24,25,26)])
corrplot(corr.res,method="circle")
```



Com podem veure, el preu de la vivenda està altament correlacionada amb els camps Àrea i Lavabos. També podem veure una correlació menor amb els camps Aire acondicionat i calefacció.

Tenint en compte el tipus de dades i el que contenen, considerem convenient convertir en tipus "factor" els camps corresponents a les característiques generals dels pisos, les característiques de la comunitat i el Districte al que pertany l'habitatge:

```
pisos$ConsumoEtiqueta<-factor(pisos$ConsumoEtiqueta)
pisos$EmisionesEtiqueta<-factor(pisos$EmisionesEtiqueta)
pisos$Aire_acondicionat <- factor(pisos$Aire_acondicionat,c(0,1),labels=c("No","Yes"))
pisos$Moblat <- factor(pisos$Moblat,c(0,1),labels=c("No","Yes"))
pisos$Calefaccio <-factor(pisos$Calefaccio,c(0,1),labels=c("No","Yes"))
pisos$Transport_public_proper <- factor(pisos$Transport_public_proper,c(0,1),labels=c("No","Yes"))
```

```
pisos$Llar_de_foc <- factor(pisos$Llar_de_foc,c(0,1),labels=c("No","Yes"))
pisos$Obra_nova <-factor(pisos$Obra_nova,c(0,1),labels=c("No","Yes"))
pisos$Piscina_propia <- factor(pisos$Piscina_propia,c(0,1),labels=c("No","Yes"))
pisos$Plaza_parking <-factor(pisos$Plaza_parking,c(0,1),labels=c("No","Yes"))
pisos$Ascensor<-factor(pisos$Ascensor,c(0,1),labels=c("No","Yes"))
pisos$Equipamient_esportiu <- factor(pisos$Equipamient_esportiu,c(0,1),labels=c("No","Yes"))
pisos$Jardi_comunitari <-factor(pisos$Jardi_comunitari,c(0,1),labels=c("No","Yes"))
pisos$Piscina_comunitaria <-factor(pisos$Piscina_comunitaria,c(0,1),labels=c("No","Yes"))
pisos$Vigilancia <-factor(pisos$Vigilancia,c(0,1),labels=c("No","Yes"))
pisos$Districte <- factor(pisos$Districte)
```

Pel camp Districte, comprovem els factors que s'han generat i eliminem aquells registres que no tenen assignat cap Districte:

```
levels(pisos$Districte)
```

```
##  [1] ""                    "Ciutat Vella"        "Eixample"
##  [4] "Gracia"              "Horta Guinardo"      "Les Corts"
##  [7] "Nou Barris"          "Sant Andreu"         "Sant Marti"
## [10] "Sants Montjuic"      "Sarria Sant Gervasi"
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
index<-which(pisos$Districte == "")
pisos <- anti_join(pisos,pisos[index,])
```

```
## Joining, by = c("X", "Barri", "Area", "Habitacions", "Lavabos", "Preu",
## "Immobiliaria", "ConsumoEtiqueta", "ConsumokW", "EmisionesEtiqueta",
## "Emisioneskg", "Data.anunci", "Aire_acondicionat", "Moblat", "Any_construccio",
## "Calefaccio", "Transport_public_proper", "Llar_de_foc", "Obra_nova",
## "Piscina_propia", "Plaza_parking", "Ascensor", "Equipamient_esportiu",
## "Jardi_comunitari", "Piscina_comunitaria", "Vigilancia", "Districte",
## "Preu_m2")
```

Per últim, convertim la variable *Data.anunci* en tipus data, ja que s'ha incorporat com a caràcters:

```
pisos$Data.anunci<-as.Date(pisos$Data.anunci,"%d/%m/%Y")
```

Extraiem el data frame net en un fitxer .csv:

```
write.csv(pisos,"~/MASTER//Tipologia i cicle de vida de les dades/Practica 2/pisos_barcelona_final.csv"
```

---

4. Anàlisi de les dades

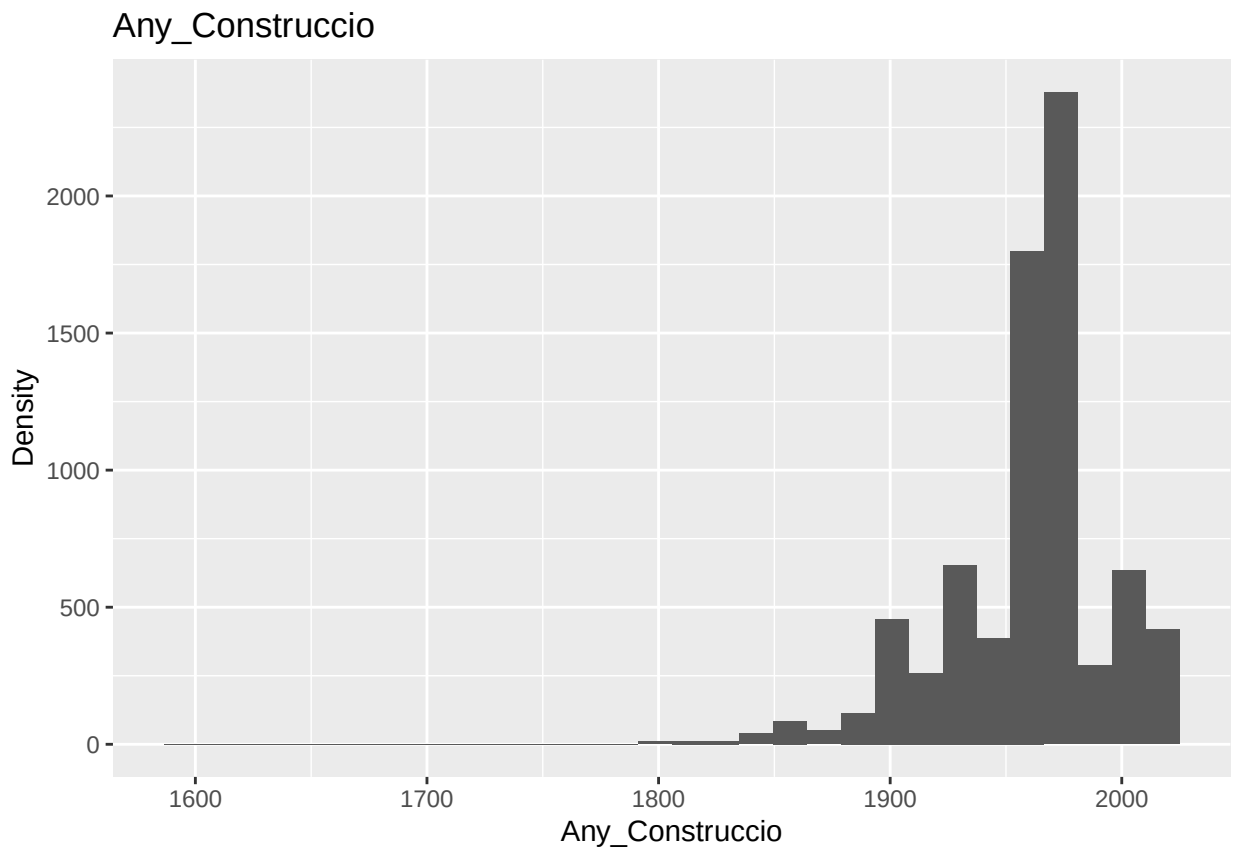
4.1. Contrast d'Hipòtesi

En primer lloc, volem comprovar si és cert que el valor dels pisos més nous és major que els que no ho són.

Per a això, representem l'any de construcció dels pisos en un histograma:

```
library(ggplot2)
#Representació de l'area en funció del districte
ggplot(mapping= aes(x=pisos$Any_construccio)) +
  geom_histogram() +
  ggtitle("Any_Construccio") +
  labs(x = "Any_Construccio",y = "Density")
```

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



Segons la representació anterior i el nostre criteri personal, decidim considerar pisos nous aquells que s'han construït després o a l'any 1990.

Les dues mostres no tindran la mateixa mida, però considerem adequat establir l'any 1990 com a separador, ja que suposa que els pisos tinguin com a molt uns 30 anys.

Per tant, creem una variable factorial per separar les dues mostres:

```
index_nous <- which(pisos$Any_construccio>=1990)
index_vells <- which(pisos$Any_construccio<1990)
pisos$Antiguitat <- 1
pisos$Antiguitat[index_vells] <- 0
pisos$Antiguitat <-factor(pisos$Antiguitat,c(0,1),labels=c("vell","nou"))
```

Comprovacions prèvies

Prèviament a realitzar l'anàlisi, hem de comprovar la normalitat i l'homoscedasticitat de la variable *Preu/m2*, ja que això ens permetrà determinar si podem utilitzar ètodes paramètrics o no paramètrics.

Per la comprovació de la normalitat utilitzem el test Kolmogorov-Smirnov:

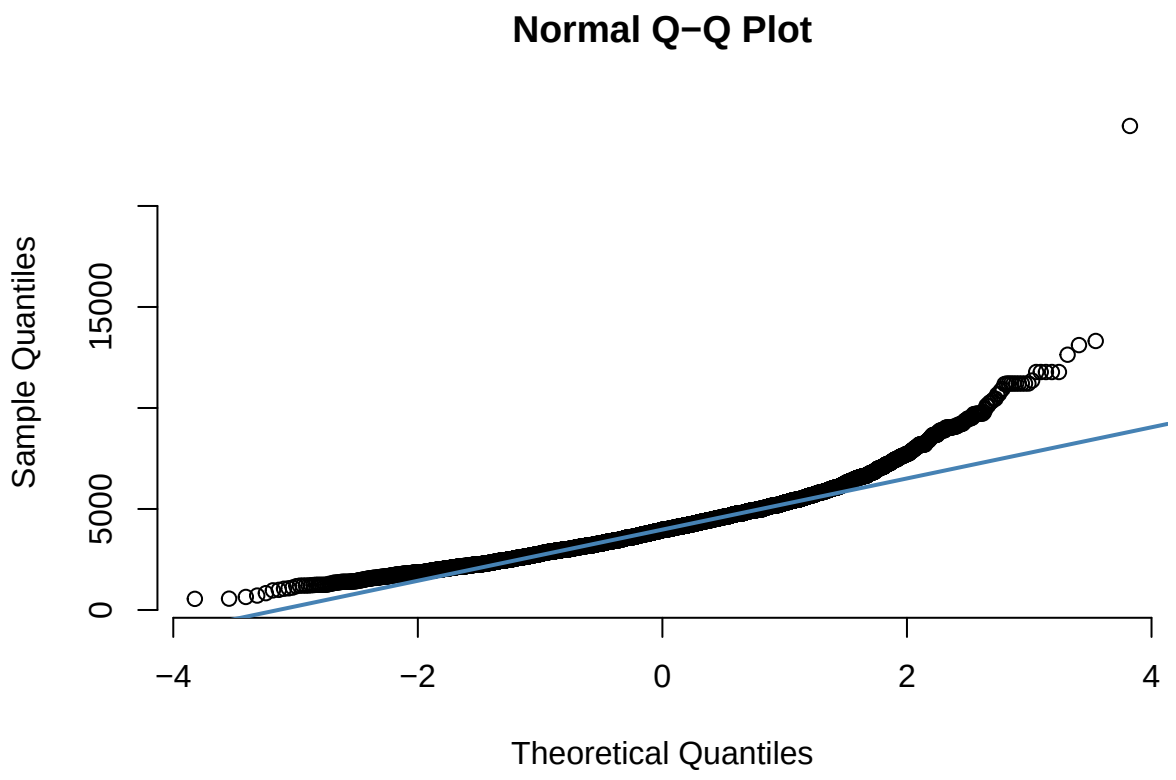
```
ks.test(pisos$Preu_m2, pnorm, mean(pisos$Preu_m2), sd(pisos$Preu_m2))
```

```
## Warning in ks.test(pisos$Preu_m2, pnorm, mean(pisos$Preu_m2),
## sd(pisos$Preu_m2)): ties should not be present for the Kolmogorov-Smirnov test
```

```
##
```

```
##  One-sample Kolmogorov-Smirnov test
##
## data:  pisos$Preu_m2
## D = 0.062641, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
qqnorm(pisos$Preu_m2, pch = 1, frame = FALSE)
qqline(pisos$Preu_m2, col = "steelblue", lwd = 2)
```



Tal com podem veure, obtenim un p-value per sota del nivell de significació del 0.05. Això implica que hem de rebutjar la hipòtesi que planteja el test i, per tant, les dades no presenten una distribució normal. Degut a la mida de la mostra, podríem aplicar el teorema del límit central, i considerar que la distribució és normal, però com tenim la opció d'utilitzar els mètodes no paramètrics, doncs procedim sense aplicar aquest teorema.

A continuació, comprovem l'homoscedasticitat a partir del test de Fligner-Killen, ja que es la opció més habitual quan les dades no compleixen amb la condició de normalitat:

```
fligner.test(Preu_m2 ~ Antiguitat, data = pisos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Preu_m2 by Antiguitat
## Fligner-Killeen:med chi-squared = 121.65, df = 1, p-value < 2.2e-16
```

Pel que hem determinat amb els tests anteriors, haurem d'utilitzar les proves no paramètriques, que pel contrast d'hipòtesi són les proves de Wilcoxon i Mann-Whitney.

Contrast d'Hipòtesi

En el nostra cas, la pregunta plantejada és si el valor del pisos construïts abans del 1990 és igual als construïts abans del 1990. Per tant, les hipòtesi son:

Ho: mean1 = mean2

Hi: mean1 != mean2

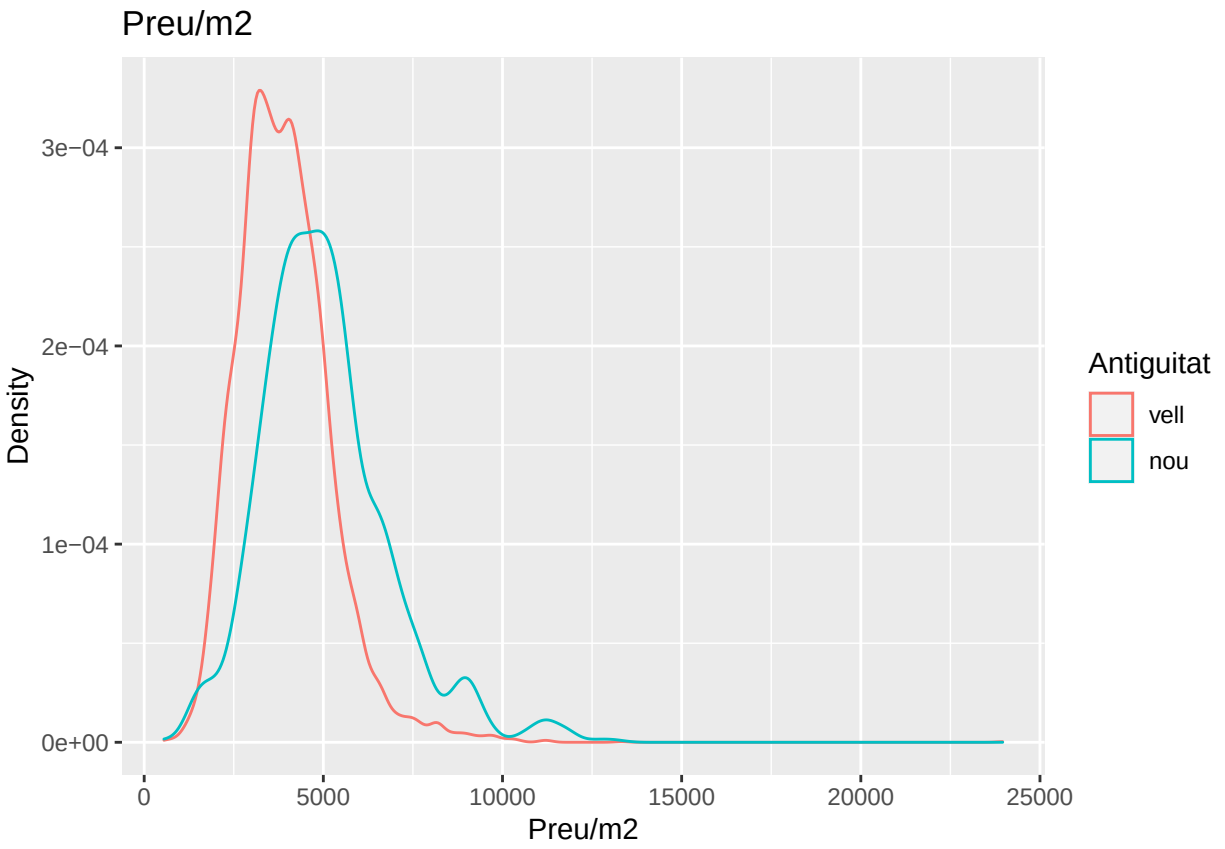
```
wilcox.test(Preu_m2 ~ Antiguitat, data = pisos)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Preu_m2 by Antiguitat
## W = 2261930, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

A partir del resultat del test, podem concloure que hi han diferències significatives en el valor dels pisos si aquests s'han construït abans o després del 1990.

Fem la representació de la variable de *Preu_m2* pels dos conjunts de dades per tal de visualitzar gràficament aquestes diferències entre les dues mostres:

```
#Representació del preu/m2 en funció de si es obra nova o no
ggplot(mapping= aes(x=pisos$Preu_m2, colour=pisos$Antiguitat)) +
  geom_density() +
  ggtitle("Preu/m2") +
  labs(x = "Preu/m2",y = "Density",colour = "Antiguitat")
```

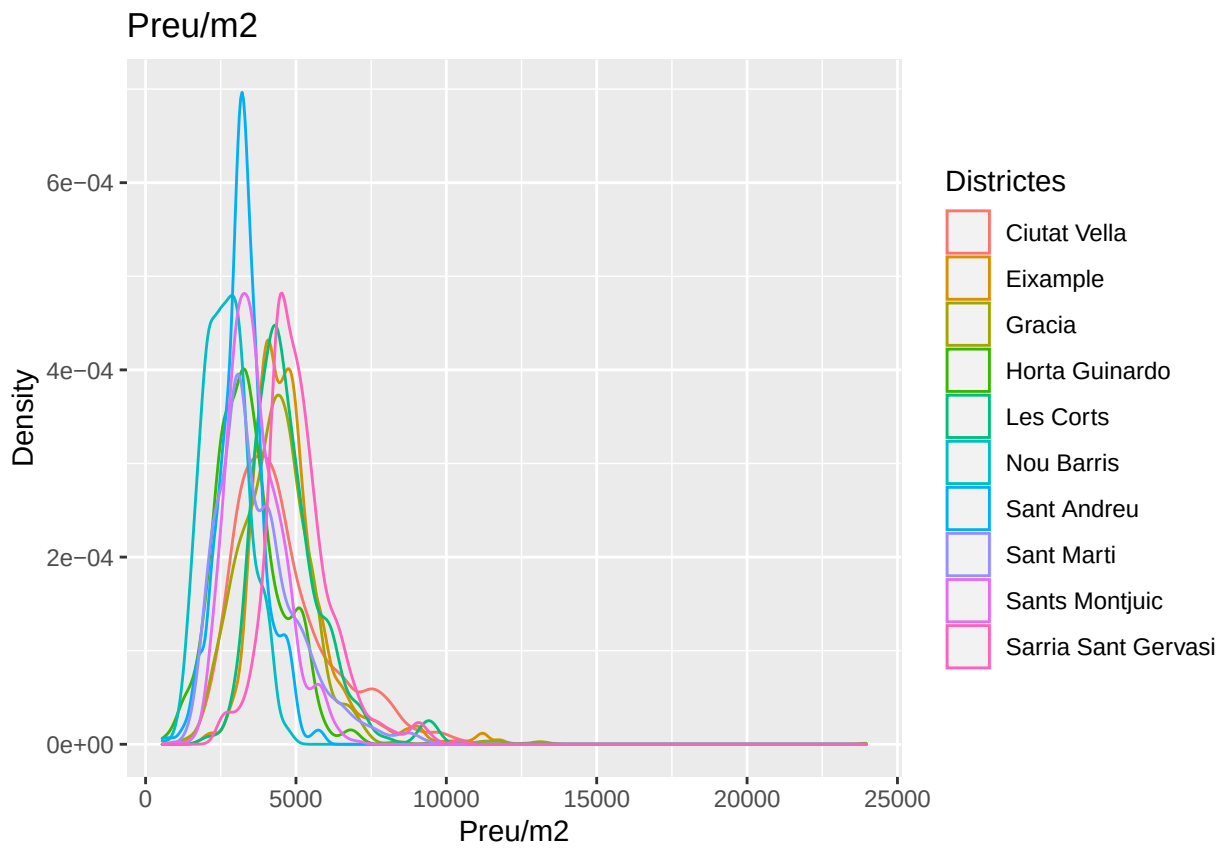


4.2 Comparació Districtes

El Districte al que pertany el pis també pot ser un factor important en el preu del pis, per tant, considerem convenient realitzar l'anàlisi de variància unidireccional (ANOVA) per aquest factor.

Per tal d'il · lustrar aquesta necessitat, representem el preu/m2 en funció del districte:

```
library(ggplot2)
#Representació del preu/m2 en funció del districte
ggplot(mapping= aes(x=pisos$Preu_m2, colour=pisos$Districte)) +
  geom_density() +
  ggtitle("Preu/m2") +
  labs(x = "Preu/m2",y = "Density",colour = "Districtes")
```



Com en l'apartat anterior ja hem comprovat que la variable *Preu/m2* no segueix una distribució normal, escollim l'alternativa no paramètrica als contrastos d'hipòtesis de més de 2 grups, que és el test de Kruskal-Wallis:

```
kruskal.test(Preu_m2 ~ Districte, data = pisos)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Preu_m2 by Districte
## Kruskal-Wallis chi-squared = 2438.4, df = 9, p-value < 2.2e-16
```

Al obtenir un p-value inferior al nivell de significació de 0.05 que esperem en els nostres anàlisis, podem concloure que el preu/m2 mostra diferències significatives en funció del Districte en el que es troba el pis.

4.3 Model de regressió lineal

També ens sembla interessant observar si existeix un model que ens permeti aproximar la variable del preu de l'habitatge en funció dels valors de la resta de les variables. Per a això, hem creat un model de regressió lineal que prediu el preu d'un pis en funció de totes les seves variables.

```r
m1 = lm(pisos$Preu~pisos$Area+pisos$Habitacions+pisos$Lavabos+pisos$Aire_acondicionat+pisos$Moblat+piso
summary(m1)
```

```
##
## Call:
## lm(formula = pisos$Preu ~ pisos$Area + pisos$Habitacions + pisos$Lavabos +
##     pisos$Aire_acondicionat + pisos$Moblat + pisos$Any_construccio +
##     pisos$Calefaccio + pisos$Transport_public_proper + pisos$Llar_de_foc +
##     pisos$Obra_nova + pisos$Piscina_propia + pisos$Plaza_parking +
##     pisos$Ascensor + pisos$Equipamient_esportiu + pisos$Jardi_comunitari +
##     pisos$Piscina_comunitaria + pisos$Vigilancia + pisos$Districte,
##     data = pisos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -639547  -57094   -6697   45818  476003
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -298327.65   73884.00  -4.038 5.45e-05 ***
## pisos$Area                         2638.23      50.62  52.123  < 2e-16 ***
## pisos$Habitacions                -14491.36    1518.03  -9.546  < 2e-16 ***
## pisos$Lavabos                     57398.92    2567.72  22.354  < 2e-16 ***
## pisos$Aire_acondicionatYes        36033.41    2555.68  14.099  < 2e-16 ***
## pisos$MoblatYes                    8625.57    3418.58   2.523  0.01165 *
## pisos$Any_construccio               182.94      38.85   4.709 2.53e-06 ***
## pisos$CalefaccioYes               17463.21    2608.81   6.694 2.33e-11 ***
## pisos$Transport_public_properYes  -13473.50    2625.24  -5.132 2.93e-07 ***
## pisos$Llar_de_focYes                6006.29   11455.82   0.524  0.60008
## pisos$Obra_novaYes                42351.40    9071.42   4.669 3.08e-06 ***
## pisos$Piscina_propiaYes          -32601.67   69763.85  -0.467  0.64029
## pisos$Plaza_parkingYes            37752.29    4155.02   9.086  < 2e-16 ***
## pisos$AscensorYes                 43364.77    3672.47  11.808  < 2e-16 ***
## pisos$Equipamient_esportiuYes     -1064.74   11428.48  -0.093  0.92577
## pisos$Jardi_comunitariYes        -20250.91    7096.86  -2.854  0.00434 **
## pisos$Piscina_comunitariaYes      72230.69    7381.11   9.786  < 2e-16 ***
## pisos$VigilanciaYes               28863.35    6296.79   4.584 4.64e-06 ***
## pisos$DistricteEixample           12548.19    4447.54   2.821  0.00479 **
## pisos$DistricteGracia            -31516.40    5824.12  -5.411 6.45e-08 ***
## pisos$DistricteHorta Guinardo   -115199.40    5843.07 -19.716  < 2e-16 ***
## pisos$DistricteLes Corts          -1317.45    6739.23  -0.195  0.84501
## pisos$DistricteNou Barris       -139022.92    5688.35 -24.440  < 2e-16 ***
## pisos$DistricteSant Andreu      -120556.08    6109.97 -19.731  < 2e-16 ***
## pisos$DistricteSant Marti        -76226.89    4990.39 -15.275  < 2e-16 ***
## pisos$DistricteSants Montjuic    -84554.47    5278.80 -16.018  < 2e-16 ***
## pisos$DistricteSarria Sant Gervasi  60704.21    6154.86   9.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97580 on 7576 degrees of freedom
## Multiple R-squared:  0.7018, Adjusted R-squared:  0.7008
## F-statistic: 685.9 on 26 and 7576 DF,  p-value: < 2.2e-16
```

Pel que podem veure del valor *R-squared* obtingut (Rˆ2=0.7042), el model permet fer una predicció més o menys precisa del preu amb la resta de les dades de la base de dades.

Pr(>|t|) és el nivell de significació de cada una de les variables. Tal i com es pot veure en el model anterior, hi ha algunes variables com Llar_de_foc, Piscina_propia i Equipamient_esportiu que tenen un coeficient realment baix, el que vol dir que la variable no afegeix res al model i es podria eliminar.

Per tal de millorar el model de regressió, provem a eliminar aquestes 3 variables en el model:

```
m2 = lm(pisos$Preu~pisos$Area+pisos$Habitacions+pisos$Lavabos+pisos$Aire_acondicionat+pisos$Moblat+piso
summary(m2)
```

```
##
## Call:
## lm(formula = pisos$Preu ~ pisos$Area + pisos$Habitacions + pisos$Lavabos +
##     pisos$Aire_acondicionat + pisos$Moblat + pisos$Any_construccio +
##     pisos$Calefaccio + pisos$Transport_public_proper + pisos$Obra_nova +
##     pisos$Plaza_parking + pisos$Ascensor + pisos$Jardi_comunitari +
##     pisos$Piscina_comunitaria + pisos$Vigilancia + pisos$Districte,
##     data = pisos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -639849  -57138   -6757   45681  476059
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -297474.60   73855.98  -4.028 5.69e-05 ***
## pisos$Area                         2638.87      50.49  52.268  < 2e-16 ***
## pisos$Habitacions                -14458.54    1516.58  -9.534  < 2e-16 ***
## pisos$Lavabos                     57408.31    2566.12  22.372  < 2e-16 ***
## pisos$Aire_acondicionatYes        36038.59    2555.23  14.104  < 2e-16 ***
## pisos$MoblatYes                    8656.49    3417.57   2.533  0.01133 *
## pisos$Any_construccio               182.37      38.83   4.697 2.69e-06 ***
## pisos$CalefaccioYes               17481.03    2608.14   6.703 2.20e-11 ***
## pisos$Transport_public_properYes  -13343.31    2612.86  -5.107 3.36e-07 ***
## pisos$Obra_novaYes                42334.54    9069.15   4.668 3.09e-06 ***
## pisos$Plaza_parkingYes            37804.83    4149.98   9.110  < 2e-16 ***
## pisos$AscensorYes                 43458.54    3667.48  11.850  < 2e-16 ***
## pisos$Jardi_comunitariYes        -20647.97    6934.11  -2.978  0.00291 **
## pisos$Piscina_comunitariaYes      72235.67    7360.83   9.814  < 2e-16 ***
## pisos$VigilanciaYes               29226.22    6246.95   4.678 2.94e-06 ***
## pisos$DistricteEixample           12539.14    4446.59   2.820  0.00482 **
## pisos$DistricteGracia            -31479.87    5822.13  -5.407 6.61e-08 ***
## pisos$DistricteHorta Guinardo   -115142.07    5840.78 -19.713  < 2e-16 ***
## pisos$DistricteLes Corts          -1298.90    6737.63  -0.193  0.84713
## pisos$DistricteNou Barris       -139028.62    5686.89 -24.447  < 2e-16 ***
## pisos$DistricteSant Andreu      -120526.89    6107.49 -19.734  < 2e-16 ***
## pisos$DistricteSant Marti        -76267.29    4988.72 -15.288  < 2e-16 ***
## pisos$DistricteSants Montjuic    -84592.00    5277.34 -16.029  < 2e-16 ***
## pisos$DistricteSarria Sant Gervasi  60819.55  6143.95   9.899  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97560 on 7579 degrees of freedom
```

```
## Multiple R-squared:  0.7018, Adjusted R-squared:  0.7009
## F-statistic: 775.6 on 23 and 7579 DF,  p-value: < 2.2e-16
```

Com podem veure a partir del valor *R-squared*, el model ha mantingut la seva qualitat predictiva, ja que el valor de *R-squared* s'ha mantingut constant, però es pot considerar que el model ha millorat perquè s'ha simplificat el nombre de variables predictives.

Per tal de millorar més el model de regressió, ara provem a eliminar aquestes les varaibles Moblat i Jardi_comunitari del model:

```
m3 = lm(pisos$Preu~pisos$Area+pisos$Habitacions+pisos$Lavabos+pisos$Aire_acondicionat+pisos$Any_construc
summary(m3)
```

```
##
## Call:
## lm(formula = pisos$Preu ~ pisos$Area + pisos$Habitacions + pisos$Lavabos +
##     pisos$Aire_acondicionat + pisos$Any_construccio + pisos$Calefaccio +
##     pisos$Transport_public_proper + pisos$Obra_nova + pisos$Plaza_parking +
##     pisos$Ascensor + pisos$Piscina_comunitaria + pisos$Vigilancia +
##     pisos$Districte, data = pisos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -636923  -57106   -7055   45693  474621
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -290115.15   73843.64  -3.929 8.61e-05 ***
## pisos$Area                         2633.45      50.50  52.144  < 2e-16 ***
## pisos$Habitacions                -14715.95    1514.64  -9.716  < 2e-16 ***
## pisos$Lavabos                     57275.07    2567.74  22.306  < 2e-16 ***
## pisos$Aire_acondicionatYes        36776.25    2539.16  14.484  < 2e-16 ***
## pisos$Any_construccio               179.98      38.84   4.634 3.64e-06 ***
## pisos$CalefaccioYes               17920.58    2603.65   6.883 6.33e-12 ***
## pisos$Transport_public_properYes  -14347.19    2581.04  -5.559 2.81e-08 ***
## pisos$Obra_novaYes                41030.54    9067.00   4.525 6.12e-06 ***
## pisos$Plaza_parkingYes            37390.87    4134.13   9.044  < 2e-16 ***
## pisos$AscensorYes                 43313.39    3668.33  11.807  < 2e-16 ***
## pisos$Piscina_comunitariaYes      69396.12    7310.27   9.493  < 2e-16 ***
## pisos$VigilanciaYes               27971.02    6229.18   4.490 7.22e-06 ***
## pisos$DistricteEixample           12104.08    4448.35   2.721  0.00652 **
## pisos$DistricteGracia            -32027.34    5824.33  -5.499 3.95e-08 ***
## pisos$DistricteHorta Guinardo    -116317.12    5837.25 -19.927  < 2e-16 ***
## pisos$DistricteLes Corts           -1932.29    6740.54  -0.287  0.77437
## pisos$DistricteNou Barris        -139950.51    5684.42 -24.620  < 2e-16 ***
## pisos$DistricteSant Andreu       -121703.96    6104.37 -19.937  < 2e-16 ***
## pisos$DistricteSant Marti         -77597.64    4979.19 -15.584  < 2e-16 ***
## pisos$DistricteSants Montjuic     -85315.97    5278.13 -16.164  < 2e-16 ***
## pisos$DistricteSarria Sant Gervasi  60144.14    6145.94   9.786  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97640 on 7581 degrees of freedom
## Multiple R-squared:  0.7013, Adjusted R-squared:  0.7004
```

```
## F-statistic: 847.4 on 21 and 7581 DF,  p-value: < 2.2e-16
```

Algunes de les conclusions que es poden treure del model són les següents:

- Com era d'esperar, els pisos més grans són més cars. Cada m² addicional d'un pis suposa uns 3.095€ d'increment en el preu.
- Tenir més lavabos incrementa el preu d'un pis però tenir més habitacions el disminueix. Això pot significar que es més valorat pels compradors que les habitacions siguin grans a que un pis en tingui moltes. Per a una àrea determinada, si el pis té menys habitacions, serà més car.
- A mesura que l'any de construcció augmenta, també ho fa el preu. Això vol dir que els pisos més nous són més cars.
- Característiques com l'aire condicionat, la calefacció, l'ascensor, aparcament, piscina comunitària i vigilància fan que el preu del pis augmenti. Per contra, la proximitat al transport públic decreix el valor d'un pis.
- Respecte els districtes: Els pisos que estan a l'Eixample, les Corts o a Sant Gervasi tenen un sobrecost pel fet d'estar en aquests districtes. Per contra, els pisos de Gràcia, Horta-Guinardó, Nou Barris, Sant Andreu, Sant Martí i Sants-Montjuïc són més barats pel fet de trobar-se en aquests districtes.

---

5. Resolució del problema

Un cop realitzat l'anàlisi, es poden extreure conclusions dels resultats obtinguts.

En el primer apartat s'ha vist la forta correlació que hi ha entre el preu d'un pis amb la seva superfície i el nombre de lavabos, i també una correlació menor amb els camps Aire acondicionat i calefacció. Posteriorment, en l'apartat de la regressió lineal, s'ha vist el pes que tenien aquests atributs en la predicció del preu. També s'ha vist una correlació entre ascensor i any de construcció, el que ens diu que els pisos més nous acostumen a tenir ascensor i els més vells no.

En l'apartat de contrast de hipòtesis s'ha pogut donar resposta a la pregunta si el valor dels preus dels pisos nous (construïts a partir del 1990) és diferent que el dels pisos antics. Per comprovar això s'ha realitzat un test de hipòtesis en el que s'ha rebutjat la hipòtesis que la mitjana dels pisos nous sigui igual a la dels vells, per tant es conclou que, de mitjana, els pisos nous tenen un valor més alt que els pisos vells.

També es volia comprovar si hi havia diferencies en el valors dels pisos segons el districte en el que estiguessin ubicats. Com Preu/m2 no segueix una distribució normal, s'ha hagut de realitzar el test de Kruskal-Wallis. A partir del resultat del test s'ha pogut concloure que hi ha diferències significatives en funció del Districte en el que es troba el pis. En el model de regressió lineal s'ha pogut quantificar aquestes diferencies.

Finalment, s'ha realitzat un model de regressió lineal per intentar predir el preu en funció de diversos atributs del pis. Durant la creació del model s'ha vist que algunes de les variables eren poc significatives i que el model millorava al no tenir-les en consideració: Llar_de_foc, Piscina_propia, Equipamient_esportiu, Moblat i Jardi_comunitari. Un cop construït el model s'ha pogut veure quins són els atributs amb més pes per definir el preu d'un pis. Tal i com s'ha comentat a l'apartat dels districtes, on s'ha vist que hi ha diferències significatives del preu en funció del Districte en el que es troba el pis, s'ha pogut quantificar aquestes diferencies: per exemple, un pis situat a Sant Gervasi amb les mateixes característiques que un pis Nou Barris seria 183.910,60€ més car.

---

6. Taula de contribucions

Contribuciones	**Firma**
Investigació prèvia	XSA, MGL
Redacció de les respostes	XSA, MGL
Desenvolupament codi	XSA, MGL