Tipologia i cicle de vida de les dades: PRA2

Autors: Laura Guerra i Àlex Tort

Desembre 2023

Contents

Pràctica 2 - Tipologia i cicle de vida de les dades

Descripció del dataset

Durant els últims anys el mercat immobiliari andorrà ha observat un augment notable en la demanda d'habitatges, tant de compra com de lloguer, degut al creixent interès de residents estrangers pel país. La demanda és tal que s'ha superat l'oferta disponible, donat lloc a un mercat immobiliari molt competitiu amb preus constantment a l'alça; convertint l'accés a l'habitatge en un problema de dimensions cada vegada majors. Moltes persones, tant de nacionalitat andorrana com potencials nous residents, s'han vist obligats a buscar alternatives que sovint impliquen buscar un habitatge a localitats properes però fora del país, com per exemple La Seu d'Urgell.

Per aquest motiu hem considerat que disposar de dades detallades del mercat immobiliari andorrà, tant de venta com de lloguer, pot resultar molt útil per determinar, entre d'altres, els factors que influeixen en els preus de les propietats, la distribució geogràfica d'aquestes, quines parròquies experimenten un major creixement i quines han experimentat un major increment en els preus...

Un dels primers objectius de l'anàlisi de les dades obtingudes és determinar la situació actual de la oferta i la demanda, i per tant, de la disponibilitat d'immobles. A part d'un primer estudi generalitzat, també es pot determinar la distribució d'oferta per parròquies, per tal d'observar si la situació és similar a totes les parròquies o hi ha més oferta i/o demanda en funció de la

localització dels immobles. L'estudi de la distribució de la oferta també es pot realitzar en funció de la tipologia de la transacció, és a dir, si la propietat és de lloguer o està en venta. Aquest estudi pot complementar-se tenint en compte les estadístiques publicades pel Departament d'Estadística d'Andorra sobre el nombre d'edificis per parròquia i tipus o la població registrada per parròquia.

Una altra dimensió rellevant que volem explorar és la distribució de preus en funció de la ubicació de l'immoble, així com de les característiques d'aquest (nombre d'habitacions, superfície, garatge...), determinant quin pes pot arribar a tenir cada característica a l'hora d'establir els preus. Un cop s'hagi obtingut aquesta informació, es pot comparar amb els ingressos mitjans o medians per unitat de consum i determinar així quin percentatge d'aquests pot representar el lloguer d'un immoble.

Per dur a terme aquest estudi s'ha recopilat un conjunt de dades de propietats per zona geogràfica, tipus d'habitatge i altres característiques específiques del portal immobiliari pisos.ad a data de 9 de novembre. El conjunt conté 3442 registres, per als quals es recullen 11 variables:

- Price: preu establert a l'anunci. En cas de tractar-se d'una propietat de lloguer, el preu correspon al preu mensual d'aquesta. El format actual de la variable és el valor en format text, amb punts als milers i en €.

- Area: superfície de la propietat, em m2. De nou, actualment es tracta d'una variable textual composta pel valor de la superfície de la propietat i les unitats corresponents.

- Bedrooms: nombre d'habitacions de la propietat. En cas de no tractar-se d'un immoble, aquesta variable té com a valor "0 Habitacions".

- Parking: variable categòrica binària que indica si la propietat disposa de garatge. Els seus valors són "Inclòs" i "No Inclòs".

- Features: llista de característiques de la propietat.

- Agency: nom de l'agència immobiliària que ha publicat l'anunci.

- Id: identificador numèric únic de l'anunci, que es correspon amb la referència del portal.

- Type: tipologia de la transacció anunciada. Els possibles valors de la variable són "venda" i "lloguer".

- Zone: parròquia on es troba ubicada la propietat. Els possibles valors són "ordino", "canillo", "encamp", "andorra-la-vella", "sant-julia-de-loria", "escaldes-engordany" i "la-massana".

- URL: enllaç de l'anunci.

- Timestamp: data i hora de l'extracció. Degut a l'elevada demanda, és possible que part dels anuncis (sobretot els de lloguer) deixin d'estar publicats al portal al cap d'uns dies i per tant no es pugui accedir de nou a la informació. Per aquest motiu s'ha considerat rellevant indicar el període al qual pertanyen les dades.

El conjunt de dades pot trobar-se a https://doi.org/10.5281/zenodo.10112197

A part, es realitzarà un procés d'integració per addicionar les dades necessàries provinents del Departament d'Estadística d'Andorra.

Integració i selecció

Per tal de poder donar resposta a les qüestions plantejades anteriorment, i en concret, per tal de determinar la proporció entre l'oferta i el nombre total d'habitatges així com entre el preu dels lloguers i els ingressos medians de la població, es recopilen i s'integren les dades necessàries més recents publicades pel Departament d'Estadística d'Andorra. Els enllaços de dades utilitzats són:

- Habitatges parròquies
- Ingressos medians

```
hab_parroquia_taula <-jsonlite::fromJSON(url_habitatges_parroquies)
ingressos_medians_taula <-jsonlite::fromJSON(url_ingressos_medians)

hab_parroquia_df <- as.data.frame(hab_parroquia_taula$Data)
ingressos_medians_df <- as.data.frame(ingressos_medians_taula$Data)

hab_parroquia_df <- hab_parroquia_df[, c('A_Descripcio', 'A_2021')]
ingressos_medians_df <- ingressos_medians_df[, c('A_Descripcio', 'A_2020')]

hab_parroquia_df <- hab_parroquia_df %>% rename('Informacio' = 'A_Descripcio',
                                                'Valor' = 'A_2021')
ingressos_medians_df <- ingressos_medians_df %>% rename('Informacio' = 'A_Descripcio',
                                                        'Valor' = 'A_2020')
```

```
# Obtenim les dades del web scraping de zenodo
zenodo_url <- "https://zenodo.org/records/10112197/files/preus_propietats_andorra.csv?download=1"
preus_propietats_df <- read.csv2(file = url(zenodo_url))
```

Neteja de les dades

Tenint en compte que l'estudi es centra en vivendes, es descarten tots aquells registres de qualsevol altra classe, és a dir, que tinguin un total de zero habitacions. Alhora, també es descarten aquells registres que no tinguin valor per qualsevol variable, exceptuant les variables "Features", per la qual s'imputarà un valor més endavant, o "Agency".

```
# Es determinen els registers buits
num_rows_missing <- sum(apply(preus_propietats_df[
  , !(names(preus_propietats_df) %in% c("Features", "Agency"))], 1,
  function(row) any(is.na(row))))

print(paste("Registres amb valors buits :", num_rows_missing))
```

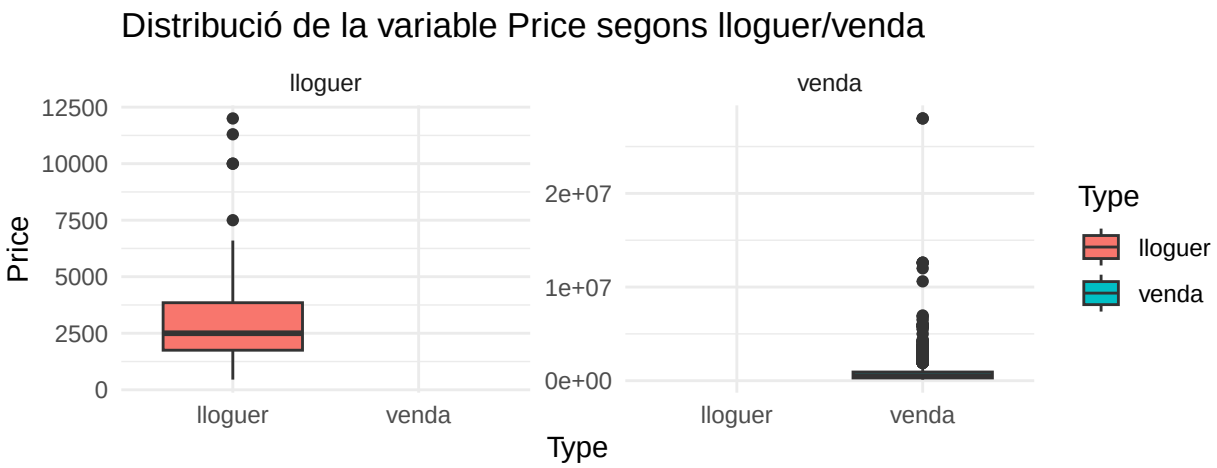
```
## [1] "Registres amb valors buits : 16"
```

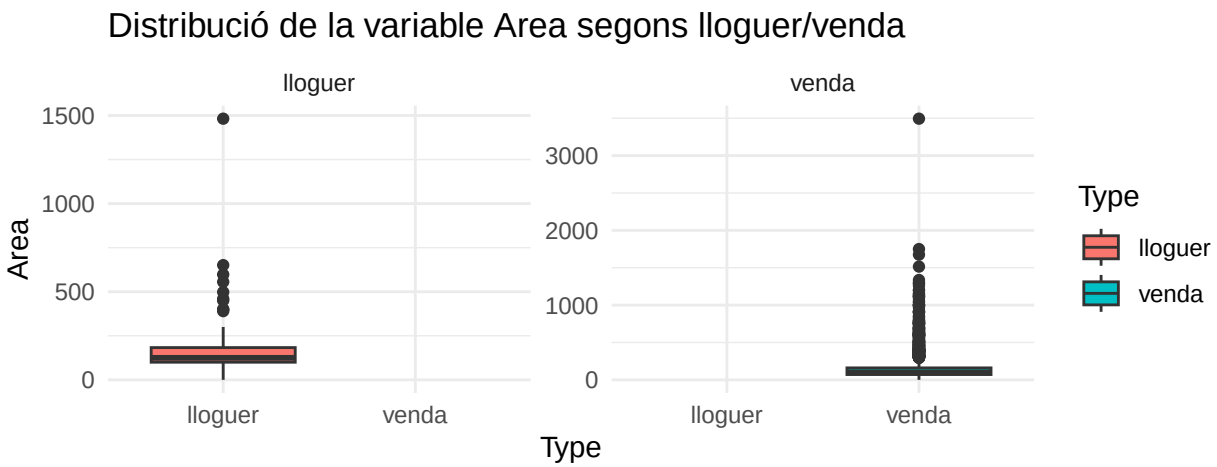
```
preus_propietats_df <- preus_propietats_df[
  complete.cases(preus_propietats_df[, !(names(preus_propietats_df) %in%
                                           c("Features", "Agency"))]), ]

# Es converteix Bedrooms en una variable numèrica i es descarten els valors igual a 0
preus_propietats_df$Bedrooms <- as.integer(gsub(" Habitacions",
                                                "", preus_propietats_df$Bedrooms))

preus_propietats_df <- preus_propietats_df[preus_propietats_df$Bedrooms > 0, ]

# Es converteixen les "Features" a variables dicotòmiques
process_data <- function(data) {
  data %>%
    mutate(Features = ifelse(is.na(Features) | Features == "", "NoFeature", Features)) %>%
    separate_rows(Features, sep = ", ") %>%
    mutate(value = 1) %>%
    pivot_wider(names_from = Features, values_from = value,
```

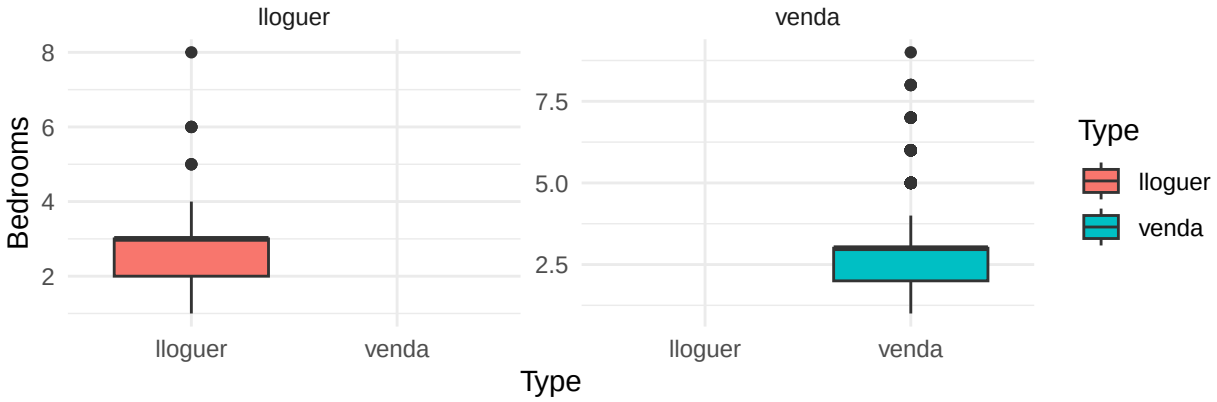
```
                values_fill = list(value = 0)) %>%
    select(-Parking)
}

processed_data <- process_data(preus_propietats_df)
```

A continuació, es tracten les variables i s'identifiquen i gestionen els valors extrems:

```
preus_propietats_df$Price <- gsub(" €", "", preus_propietats_df$Price)
preus_propietats_df$Price <- as.integer(gsub("[^0-9]", "",
                                             gsub("\\.", "",preus_propietats_df$Price)))
preus_propietats_df$Area <- as.integer(gsub(" m2", "",preus_propietats_df$Area))

# Eliminem valors nuls que s'hagin pogut generar:
preus_propietats_df <- na.omit(preus_propietats_df)

for (variable in c("Price", "Area", "Bedrooms")) {
  plt <- ggplot(preus_propietats_df, aes(x = Type, y = .data[[variable]], fill = Type)) +
    geom_boxplot() +
    labs(title = paste("Distribució de la variable", variable, "segons lloguer/venda"),
         y = variable) +
    theme_minimal() +
    facet_wrap(~Type, scales = "free_y")

  print(plt)
}
```







Pel que fa a la variable "Bedrooms", si bé sorprèn que algunes de les propietats arribin a tenir fins a vuit habitacions, no és un valor suficientment elevat com per considerar que es tracta d'un error, i per tant, s'ha optat per no descartar els registres. De la mateixa manera, tampoc s'han descartat els registres amb àrees superiors a la resta, exceptuant el valor màxim de les propietats en venda i lloguer, ja que es troben molt per sobre fins i tot dels valors més extrems. Pel que fa als preus, només es considera un valor extrem descartable el valor màxim de les propietats en venda. Més endavant, en cas que es consideri adient pels diferents anàlisi, es podran realitzar les seleccions corresponents, obviant aquells registres que es consideri que distorsionen excessivament els càlculs.

```
preus_propietats_df <- preus_propietats_df %>%
  filter(!(Type == "lloguer" & Area > 1000) & !(Type == "venda" & Area > 3000))

preus_propietats_df <- preus_propietats_df %>%
  filter(!(Type == "venda" & Price > 20000000))
```

Anàlisi de les dades

Tal com s'indicava a l'inici de la pràctica, l'objectiu és determinar a grans trets l'estat actual del mercat immobiliàri andorrà. Així doncs, s'ha considerat que per obtenir una visió global es poden realitzar els anàlisis següents:

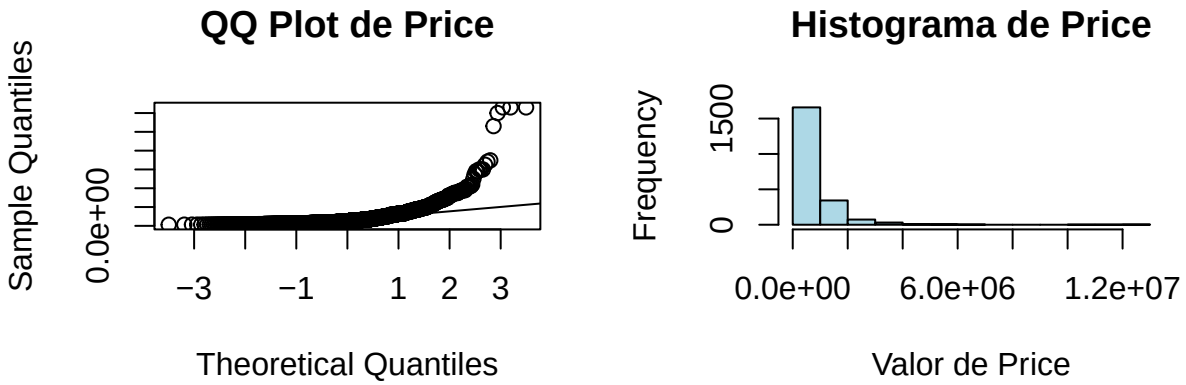
1. Determinar l'estat actual de l'oferta a través de la proporció entre el nombre d'habitatges en oferta i el nombre d'habitatges totals per parròquia. Per fer-ho, s'han delimitat dues zones, una considerada com la zona cèntrica, que inclou Andorra la Vella i Escaldees, i una zona no cèntrica, que inclou la resta de les parròquies.
2. Estudiar les distribucions de preus de lloguer i venda de les dues zones definides, i comparar els valors mitjans de les dos grups mitjançant un contrast d'hipòtesi per determinar si el preu a la zona cèntrica és superior.
3. Determinar la proporció dels ingressos mensuals medians nets que suposa el lloguer d'un pis en funció de la zona, i si aquesta supera el 35% recomanable.
4. Desenvolupar un model de regressió lineal per obtenir el preu òptim d'una vivenda a partir de les característiques d'aquesta (superfície, nombre d'habitacions, pàrquing, zona...).
5. Estudiar les possibles correlacions entre les diferents característiques dels habitatges.

Per a l'anàlisi del mercat immobiliari andorrà, seleccionem grups de dades basats en característiques com la ubicació (cèntric o no cèntric), tipus de transacció (venda o lloguer), i atributs de les propietats (nombre d'habitacions i superfície). Aquesta selecció permetrà comparar diferents paràmetres del mercat, com la diferència en preus i característiques entre de les diferents propietats en venda i lloguer. Abans de procedir als anàlisis, s'estudia la normalitat i homogeneïtat de la variància:

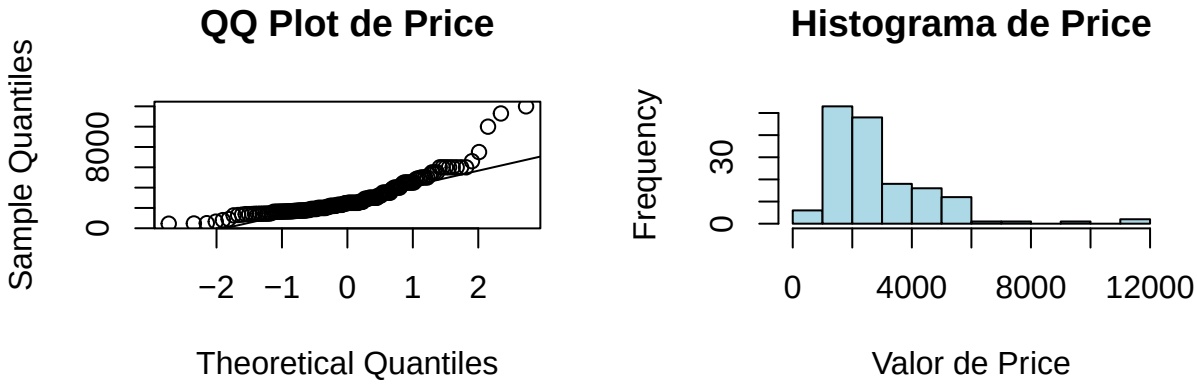
```
preus_propietats_df$Centric <- ifelse(preus_propietats_df$Zone %in%
                                      c("andorra-la-vella", "escaldes-engordany"), 1, 0)

venda_data <- preus_propietats_df[preus_propietats_df$Type == "venda", ]
lloguer_data <- preus_propietats_df[preus_propietats_df$Type == "lloguer", ]

normality_check <- function(data, variable) {
    shapiro_test <- shapiro.test(data[[variable]])
    print(shapiro_test)

    par(mfrow = c(1, 2))

    qqnorm(data[[variable]], main = paste("QQ Plot de", variable))
    qqline(data[[variable]])

    hist(data[[variable]], main = paste("Histograma de", variable),
         xlab = paste("Valor de", variable), col = "lightblue", border = "black")

    par(mfrow = c(1, 1))
}

# Aplicar la funció a les variables per a cada subconjunt de dades
for (variable in c("Price", "Area", "Bedrooms")) {
    cat("\nShapiro-Wilk Test per la variable", variable, "en propietats en venda:\n")
    normality_check(venda_data, variable)

    cat("\nShapiro-Wilk Test per a la variable", variable, "en propietats en lloguer:\n")
    normality_check(lloguer_data, variable)
}
```

```
##
## Shapiro-Wilk Test per la variable Price en propietats en venda:
```

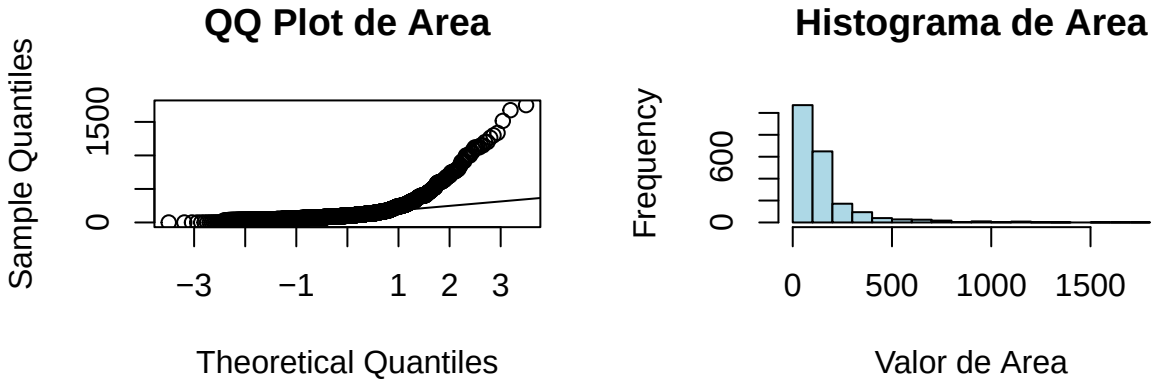
```
##
##  Shapiro-Wilk normality test
##
## data:  data[[variable]]
## W = 0.55637, p-value < 2.2e-16
```



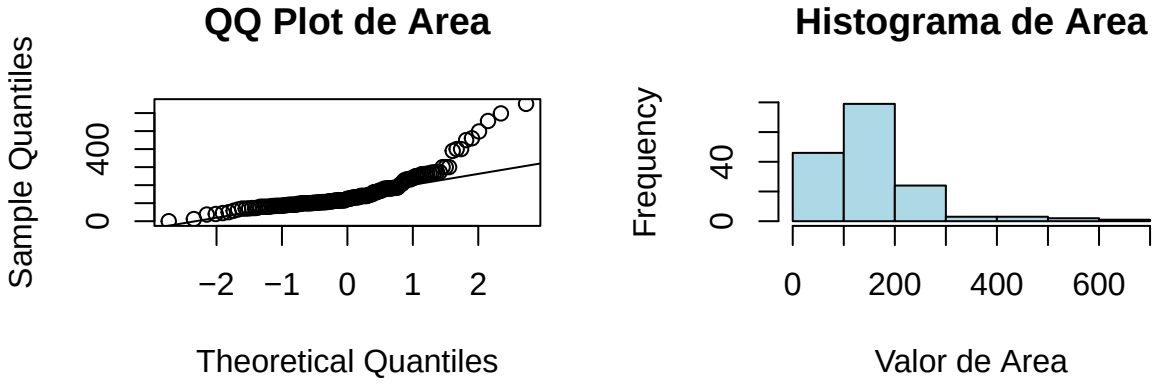

```
##
## Shapiro-Wilk Test per a la variable Price en propietats en lloguer:
##
##  Shapiro-Wilk normality test
##
## data:  data[[variable]]
## W = 0.81984, p-value = 1.124e-12
```



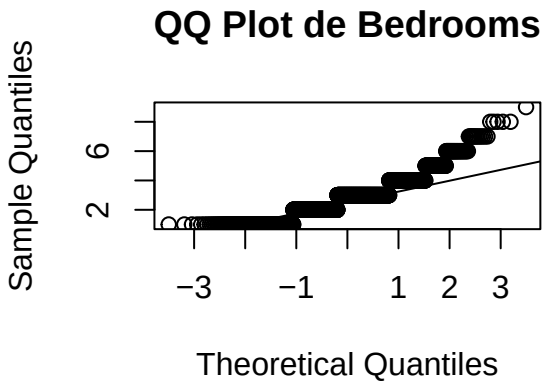

```
##
## Shapiro-Wilk Test per la variable Area en propietats en venda:
##
##  Shapiro-Wilk normality test
##
## data:  data[[variable]]
## W = 0.5989, p-value < 2.2e-16
```

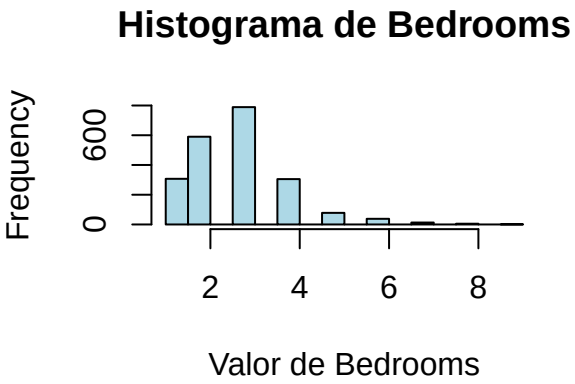



```
##
## Shapiro-Wilk Test per a la variable Area en propietats en lloguer:
##
##  Shapiro-Wilk normality test
##
## data:  data[[variable]]
## W = 0.77727, p-value = 3.183e-14
```

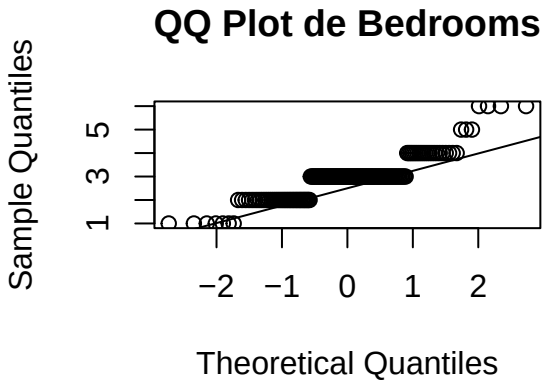


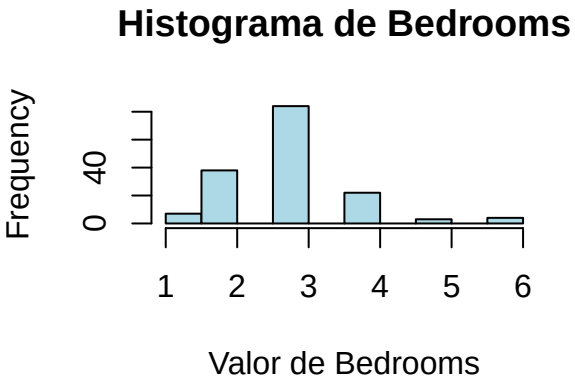

```
##
## Shapiro-Wilk Test per la variable Bedrooms en propietats en venda:
##
##  Shapiro-Wilk normality test
##
## data:  data[[variable]]
## W = 0.89565, p-value < 2.2e-16
```






```
##
## Shapiro-Wilk Test per a la variable Bedrooms en propietats en lloguer:
##
##  Shapiro-Wilk normality test
##
## data:  data[[variable]]
## W = 0.84823, p-value = 1.698e-11
```





Els resultats dels tests de Shapiro-Wilk indiquen clarament que les distribucions de les variables 'Price', 'Area' i 'Bedrooms', tant per a propietats en venda com en lloguer, no segueixen una distribució normal, ja que els valors p són extremadament baixos en tots els casos (menors de 0.05). Això implica que per aquestes variables, les dades es desvien significativament d'una distribució normal.

A continuació farem una comprovació de la igualtat de variàncies entre les propietats cèntriques i no cèntriques. En aquest cas, com les dades no compleixen la condició de normalitat, utilitzarem el test de Fligner-Killeen.

```
fligner_test_per_tipus <- function(variable) {
  cat("Fligner-Killeen Test per la variable", variable, "en propietats en venda:\n")
  print(fligner.test(venda_data[[variable]] ~ venda_data$Centric))
```

```r
  cat("Fligner-Killeen Test per la variable", variable, "en propietats en lloguer:\n")
  print(fligner.test(lloguer_data[[variable]] ~ lloguer_data$Centric))
}

for (variable in c("Price", "Area", "Bedrooms")) {
  fligner_test_per_tipus(variable)
}
```

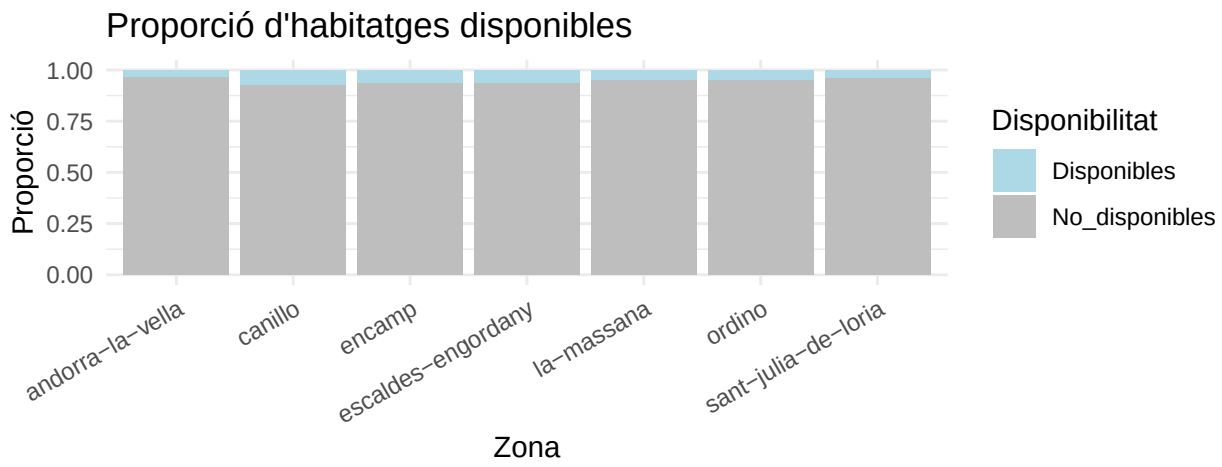
```
## Fligner-Killeen Test per la variable Price en propietats en venda:
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  venda_data[[variable]] by venda_data$Centric
## Fligner-Killeen:med chi-squared = 21.397, df = 1, p-value = 3.733e-06
##
## Fligner-Killeen Test per la variable Price en propietats en lloguer:
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  lloguer_data[[variable]] by lloguer_data$Centric
## Fligner-Killeen:med chi-squared = 2.4429, df = 1, p-value = 0.1181
##
## Fligner-Killeen Test per la variable Area en propietats en venda:
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  venda_data[[variable]] by venda_data$Centric
## Fligner-Killeen:med chi-squared = 2.2767, df = 1, p-value = 0.1313
##
## Fligner-Killeen Test per la variable Area en propietats en lloguer:
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  lloguer_data[[variable]] by lloguer_data$Centric
## Fligner-Killeen:med chi-squared = 0.035027, df = 1, p-value = 0.8515
##
## Fligner-Killeen Test per la variable Bedrooms en propietats en venda:
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  venda_data[[variable]] by venda_data$Centric
## Fligner-Killeen:med chi-squared = 46.022, df = 1, p-value = 1.169e-11
##
## Fligner-Killeen Test per la variable Bedrooms en propietats en lloguer:
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  lloguer_data[[variable]] by lloguer_data$Centric
## Fligner-Killeen:med chi-squared = 2.1518, df = 1, p-value = 0.1424
```

Els resultats del test de Fligner-Killeen mostren diferents patrons en la variabilitat de les dades entre les propietats en venda i en lloguer, així com entre zones cèntriques i no cèntriques:

- Preu (Price) en Venda: Hi ha una variància significativament diferent entre zones (p-value = 3.733e-06), indicant que el preu varia més d'una zona a l'altra.
- Preu (Price) en Lloguer: No es troben diferències significatives (p-value = 0.1181), suggerint una major uniformitat en els preus de lloguer independentment de la zona.
- Àrea (Area) en Venda i Lloguer: Les variàncies no mostren diferències significatives (p-values = 0.1313 i 0.8488, respectivament), indicant una certa consistència en la mida de les propietats independentment de la zona.
- Habitacions (Bedrooms) en Venda: Hi ha una diferència significativa en la variància (p-value = 1.169e-11), mostrant una major variabilitat en el nombre d'habitacions entre zones.
- Habitacions (Bedrooms) en Lloguer: Sense diferències significatives (p-value = 0.1834), indicant una uniformitat en la distribució del nombre d'habitacions en lloguer entre zones.

Tal com s'indicava a l'inici de l'apartat, un dels objectius de l'estudi és determinar l'estat actual de l'oferta a través de la proporció entre el nombre d'habitatges en oferta (tan de lloguer com de venda) i el nombre total d'habitatges per parròquia.

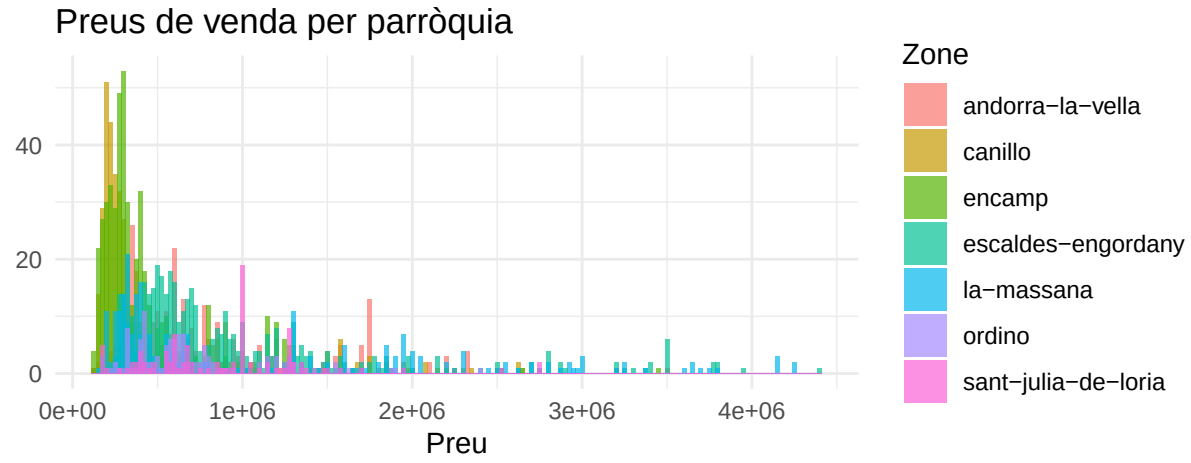
```
hab_parroquia_df$Informacio <- tolower(sub(".*?\\.\\s*(.*)\\.$", "\\1",
                                           hab_parroquia_df$Informacio))

zone_counts <- as.data.frame(table(preus_propietats_df$Zone))
colnames(zone_counts) <- c("Zone", "Registres")

zone_counts$Mod_zone <- sapply(strsplit(as.character(zone_counts$Zone), "-"),
                               function(x) x[1])

hab_parroquia_df$Mod_zone <- sapply(strsplit(as.character(hab_parroquia_df$Informacio),
                                             " "),function(x) x[1])

prop_by_zone <- merge(hab_parroquia_df, zone_counts, by.x = "Mod_zone",
                      by.y = "Mod_zone", all.x = TRUE)[, c("Zone", "Registres", "Valor")]
prop_by_zone <- prop_by_zone[!is.na(prop_by_zone$Zone), ]

prop_by_zone$Registres <- as.numeric(prop_by_zone$Registres)
prop_by_zone$Valor <- as.numeric(prop_by_zone$Valor)

# Calcul de la raó i el valor de cada zona
prop_by_zone$Disponibles <- prop_by_zone$Registres / prop_by_zone$Valor
# Calcula el complement fins a 1
prop_by_zone$No_disponibles <- 1 - prop_by_zone$Disponibles

aux_data <- prop_by_zone %>%
  gather(key = "Disponibilitat", value = "Valor", No_disponibles, Disponibles)

# Crea el gràfic de barres
ggplot(aux_data, aes(x = Zone, y = Valor, fill = Disponibilitat)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("lightblue", "grey")) +
  theme_minimal() +
  labs(title = "Proporció d'habitatges disponibles",
       x = "Zona",
       y = "Proporció") +
       theme(axis.text.x = element_text(angle = 30, vjust = 1, hjust = 1))
```



Tal i com es pot observar al gràfic anterior, la proporció d'habitatges disponibles, ja sigui mitjançant la seva venda o el seu lloguer, és molt reduïda, sobretot a les parròquies més cèntriques. Així doncs, es posa de manifest una de les afirmacions exposades a l'inici i que han donat lloc a l'estudi que es presenta: el mercat immobiliàri andorrà és molt reduït, i per tant també ho és l'accés a una vivenda, donant lloc a una problemàtica social patent.

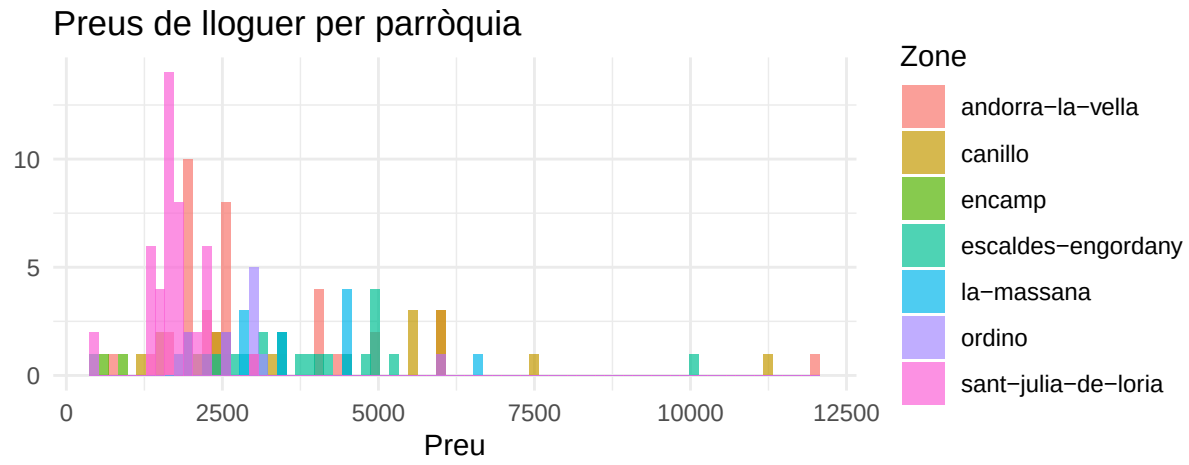
Alhora, es preten estudiar les distribucions de preus de lloguer i venda de les dues zones definides, i comparar els valors mitjans de les dos grups mitjançant un contrast d'hipòtesi per determinar si el preu a la zona cèntrica és superior.

```
ggplot(venda_data[venda_data$Price < 5000000, ], aes(x = Price, fill = Zone)) +
  geom_histogram(binwidth = 25000, position = "identity", alpha = 0.7) +
  labs(title = "Preus de venda per parròquia",
       x = "Preu",
       y = "") +
  theme_minimal()
```




```
ggplot(lloguer_data, aes(x = Price, fill = Zone)) +
  geom_histogram(binwidth = 150, position = "identity", alpha = 0.7) +
  labs(title = "Preus de lloguer per parròquia",
       x = "Preu",
```

```
       y = "") +
  theme_minimal()
```




```
habitatges_venda_centrics <- venda_data[venda_data$Centric == 1, ]
habitatges_venda_no_centrics <- venda_data[venda_data$Centric == 0, ]
t_test_result_venda <- t.test(habitatges_venda_centrics$Price,
                              habitatges_venda_no_centrics$Price,
                              alternative = "greater")

print("El resultat del contrast d'hipòtesis per les ofertes de venda és:")
```

```
## [1] "El resultat del contrast d'hipòtesis per les ofertes de venda és:"
```

```
print(t_test_result_venda)
```

```
##
##  Welch Two Sample t-test
##
## data:  habitatges_venda_centrics$Price and habitatges_venda_no_centrics$Price
## t = 5.49, df = 1920.2, p-value = 2.277e-08
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  151964.7      Inf
## sample estimates:
## mean of x mean of y
##  920329.7  703313.0
```

```
habitatges_lloguer_centrics <- lloguer_data[lloguer_data$Centric == 1, ]
habitatges_lloguer_no_centrics <- lloguer_data[lloguer_data$Centric == 0, ]
t_test_result_lloguer <- t.test(habitatges_lloguer_centrics$Price,
                                habitatges_lloguer_no_centrics$Price,
                                alternative = "greater")

print("El resultat del contrast d'hipòtesis per les ofertes de lloguer és:")
```

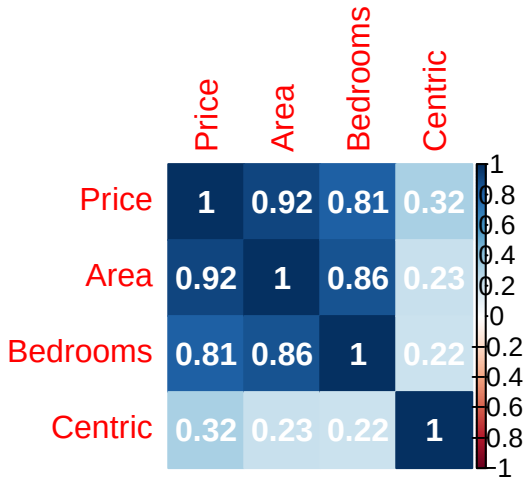
```
## [1] "El resultat del contrast d'hipòtesis per les ofertes de lloguer és:"
```

```
print(t_test_result_lloguer)
```

```
##
##  Welch Two Sample t-test
##
## data:  habitatges_lloguer_centrics$Price and habitatges_lloguer_no_centrics$Price
## t = 2.7073, df = 124.04, p-value = 0.00387
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  311.6675      Inf
## sample estimates:
## mean of x mean of y
##  3464.188  2660.638
```

Tenint en compte que els p-values obtinguts en ambdós casos són inferiors al nivell de significança, es pot rebutjar còmodament la hipòtesis nul · la i afirmar que, efectivament, els preus mitjans tant de lloguer com de venda són superiors al centre.

```
correlation_matrix <- cor(venda_data[, c("Price", "Area", "Bedrooms", "Centric")],
                          method = "spearman")
corrplot(correlation_matrix, method = "color", addCoef.col = "white")
```



Els resultats de la correlació de Spearman en el pisos en venda mostren una forta associació entre 'Price' i 'Area' (0.917), indicant que a major àrea de la propietat, major és el preu, en general. Una associació similar es troba entre 'Price' i 'Bedrooms' (0.812), suggerint que propietats amb més habitacions tendeixen a tenir preus més alts. La correlació entre 'Price' i 'Centric' (0.320) és positiva però més moderada, indicant que hi ha una tendència a preus més alts en zones cèntriques, però no és tan forta com les relacions amb l'àrea o el nombre d'habitacions.

Per altra banda, en els pisos de lloguer trobem aquesta casuística:

```
correlation_matrix <- cor(lloguer_data[, c("Price", "Area", "Bedrooms", "Centric")],
                          method = "spearman")
corrplot(correlation_matrix, method = "color", addCoef.col = "white")
```

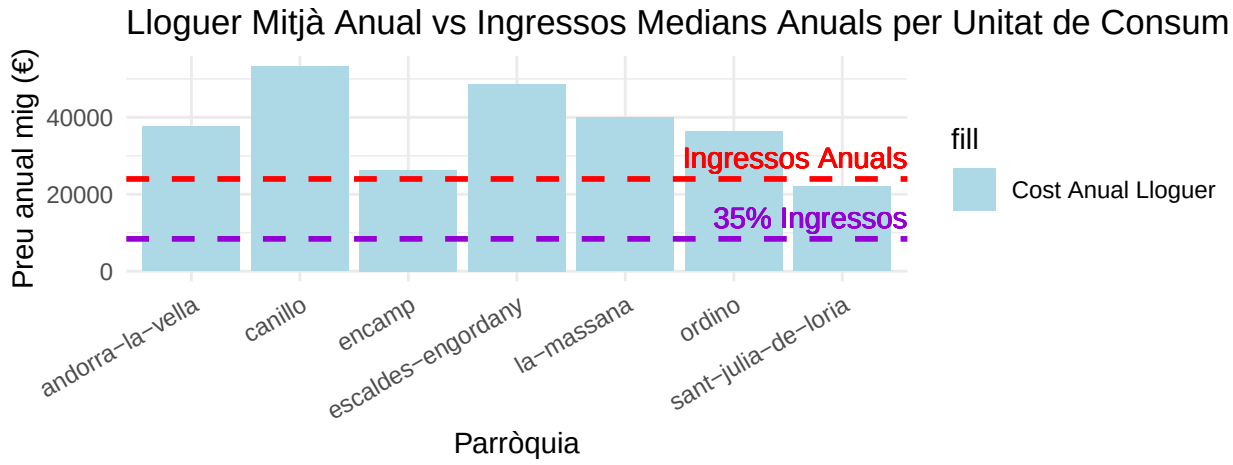
En el cas dels pisos en lloguer els resultats són similars als de pisos en venda. Els resultats de la correlació de Spearman mostren una relació forta entre 'Price' i 'Area' (0.755), indicant que els pisos més grans tendeixen a tenir lloguers més alts, cal remarcar però que aquesta correlació no és tan forta com en el cas dels pisos en venda. La correlació entre 'Price' i 'Bedrooms' (0.600) és també positiva, suggerint que propietats amb més habitacions generalment tenen lloguers més alts. La correlació entre 'Price' i 'Centric' (0.308) és positiva però més baixa, indicant una tendència a lloguers més alts en zones cèntriques, però aquesta tendència no és tan forta com la relació amb l'àrea o el nombre d'habitacions.

A continuació fem un anàlisi per a veure la relació entre el salari medià i el preu del lloguer. Segons les directrius estàndard d'habitatge i finances personals és recomanable que el cost de l'habitatge no superi el 35% dels ingressos totals. Aquesta anàlisi calcula la mediana i mitjana dels preus anuals de lloguer per parròquia i determina quin percentatge de lloguers supera aquest llindar del 35% en relació als ingressos anuals per unitat de consum. També es calcula quantes vegades els ingressos anuals cobreixen aquest cost del lloguer, proporcionant una visió de la càrrega financera associada al lloguer en diferents parròquies.

```r
# Preparar les dades
ingressos_anuals <- ingressos_medians_df$Valor
lloguer_data <- lloguer_data %>% mutate(Preu_Anual = Price * 12)

# Calcular estadístiques per parròquia
estadistiques_parroquia <- lloguer_data %>%
  group_by(Zone) %>%
  summarise(
    Lloguer_Mitja = mean(Preu_Anual, na.rm = TRUE),
    Mediana_Lloguer = median(Preu_Anual, na.rm = TRUE),
    Percentatge_Superior_35 = mean(Preu_Anual > ingressos_anuals * 0.35) * 100,
    Salaris_Necessaris = mean(Preu_Anual / (ingressos_anuals * 0.35))
  )

# Combinar les dades en una taula per visualització
dades_combinades <- estadistiques_parroquia %>%
  mutate(Parròquia = Zone) %>%
  select(Parròquia, everything())

# Mostrar la taula combinada
print(dades_combinades)
```

```
## # A tibble: 7 x 6
##   Parròquia           Zone  Lloguer_Mitja Mediana_Lloguer Percentatge_Superior~1
##   <chr>               <chr>         <dbl>           <dbl>                  <dbl>
## 1 andorra-la-vella    ando~        37568.           30000                  100
## 2 canillo             cani~        53273.           54000                  100
## 3 encamp              enca~        26314.           26400                   85.7
## 4 escaldes-engordany  esca~        48704.           46800                   95.7
## 5 la-massana          la-m~        39900            34800                  100
## 6 ordino              ordi~        36400            36000                  100
## 7 sant-julia-de-loria sant~        21970            20340                   95.8
## # i abbreviated name: 1: Percentatge_Superior_35
## # i 1 more variable: Salaris_Necessaris <dbl>
```

```r
# Crear el gràfic
ggplot(dades_combinades, aes(x = Parròquia, y = Lloguer_Mitja,
                             fill = "Cost Anual Lloguer")) +
  geom_col() +
  geom_hline(aes(yintercept = ingressos_anuals), color = "red", linetype = "dashed",
             size = 1) +
  geom_text(aes(x = Inf, y = ingressos_anuals, label = "Ingressos Anuals"),
            color = "red", vjust = -0.5, hjust = 1) +
  geom_hline(aes(yintercept = ingressos_anuals * 0.35), color = "darkviolet",
             linetype = "dashed", size = 1) +
  geom_text(aes(x = Inf, y = ingressos_anuals * 0.35, label = "35% Ingressos"),
            color = "darkviolet", vjust = -0.5, hjust = 1) +
  labs(title = "Lloguer Mitjà Anual vs Ingressos Medians Anuals per Unitat de Consum",
       x = "Parròquia",
       y = "Preu anual mig (€)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  scale_fill_manual(values = c("Cost Anual Lloguer" = "lightblue"))
```



L'anàlisi del mercat de lloguer a Andorra mostra una càrrega financera significativa per als inquilins en totes les parròquies. A Andorra-la-Vella, la mitjana de lloguer anual és de 37,568.20€, i es requereixen aproximadament 4.48 salaris per afrontar aquest cost. Canillo té la mitjana més alta amb 53,272.94€, exigint fins a 6.35 salaris, el que indica una gran pressió financera. Encamp presenta una situació més moderada amb una mitjana de 26,314.29€ i 3.14 salaris necessaris.

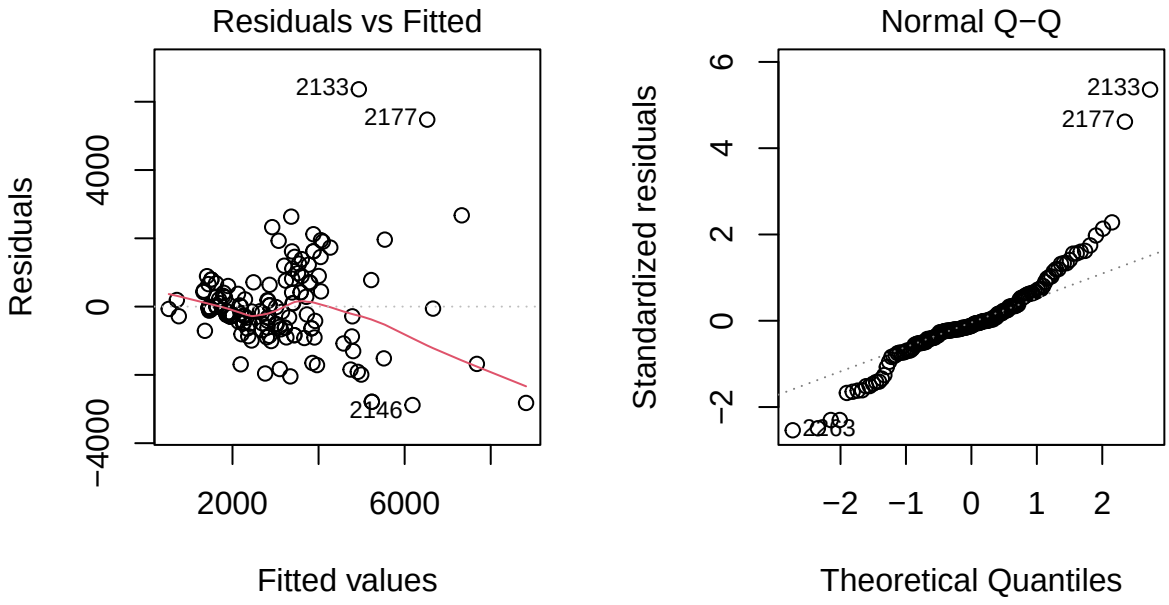
A Escaldes-Engordany, el lloguer mitjà és de 48,704.35€, necessitant gairebé 5.8 salaris. La Massana i Ordino mostren també una alta càrrega amb 4.75 i 4.34 salaris respectivament. Finalment, Sant Julià de Lòria té la mitjana més baixa amb 21,970€, però encara requereix 2.62 salaris.

Aquesta situació reflecteix una problemàtica d'accessibilitat als habitatges, sobretot a Canillo i Escaldes-Engordany. Per entendre millor com l'àrea, nombre d'habitacions, aparcament i ubicació afecten els preus, s'implementarà un model de regressió lineal per analitzar tant els pisos en venda com els de lloguer.

```
realitzar_regressio <- function(df, title="") {
    # Definir variables
    dependent_var <- "Price"
    independent_vars <- c("Area", "Bedrooms", "Parking", "Centric")

    # Convertir la variable 'Zone' a factor
    df$Zone <- factor(df$Zone)

    # Imprimir el título
    if (!missing(title)) {
        cat("Análisis per a:", title, "\n")
    }

    # Regressió
    model <- lm(Price ~ Area + Bedrooms + Parking + Centric + Zone, data=df)

    # Avaluació del model
    model_summary <- summary(model)
    print(model_summary)

    # Diagnòstic del model
    par(mfrow=c(1,2))
    plot(model, which=1) # Residuals vs Fitted
    plot(model, which=2) # Normal Q-Q

    return(list(model=model, summary=model_summary))
}

resultats_lloguer <- realitzar_regressio(lloguer_data, "Lloguer")
```

```
## Análisis per a: Lloguer
##
## Call:
## lm(formula = Price ~ Area + Bedrooms + Parking + Centric + Zone,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2882.1  -522.4  -141.2   423.1  6364.5
##
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              281.592    463.650   0.607 0.544559
## Area                       8.356      1.319   6.334 2.71e-09 ***
## Bedrooms                 321.697    157.308   2.045 0.042625 *
```

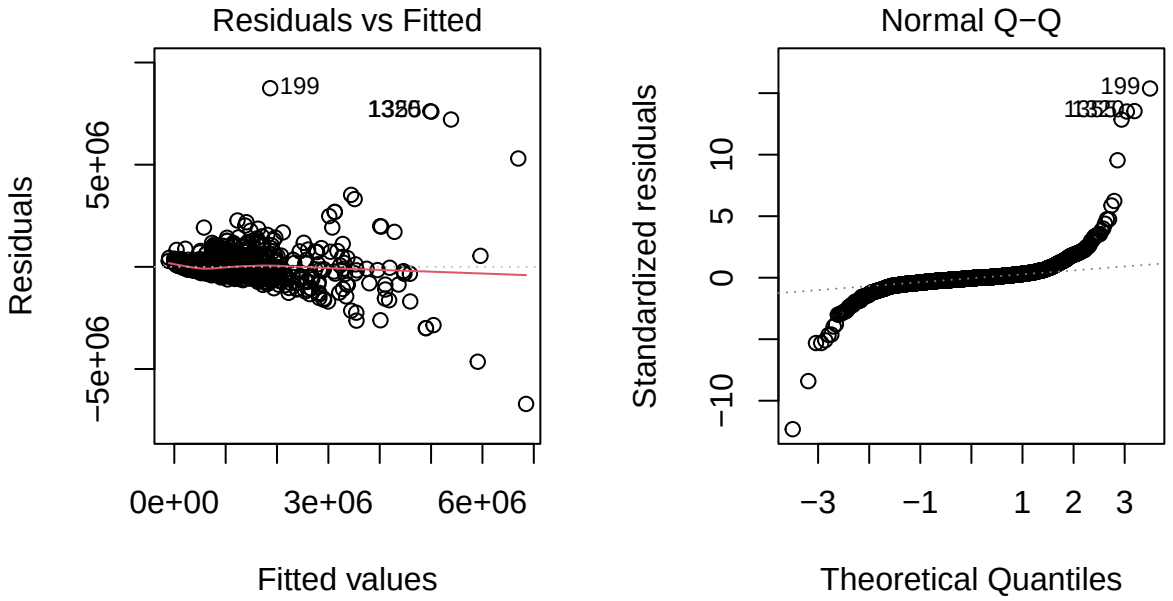
```
## ParkingNo Inclòs        -185.321    252.526  -0.734 0.464189
## Centric                  791.103    281.551   2.810 0.005628 **
## Zonecanillo             1463.382    417.606   3.504 0.000606 ***
## Zoneencamp               -20.158    526.550  -0.038 0.969514
## Zoneescaldes-engordany   563.629    329.475   1.711 0.089233 .
## Zonela-massana           279.230    386.296   0.723 0.470919
## Zoneordino              -492.477    595.888  -0.826 0.409873
## Zonesant-julia-de-loria       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1253 on 148 degrees of freedom
## Multiple R-squared:  0.5561, Adjusted R-squared:  0.5291
## F-statistic:  20.6 on 9 and 148 DF,  p-value: < 2.2e-16
```




```
resultats_venda <- realitzar_regressio(venda_data, "Venda")
```

```
## Análisis per a: Venda
##
## Call:
## lm(formula = Price ~ Area + Bedrooms + Parking + Centric + Zone,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6709415  -149697   -28756    95499  8742929
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)             -187705.64   70882.13  -2.648  0.00815 **
## Area                       3835.48      96.28  39.835  < 2e-16 ***
## Bedrooms                  98320.82   14290.88   6.880 7.86e-12 ***
## ParkingNo Inclòs         -19220.28   25152.59  -0.764  0.44486
## Centric                  147606.78   65029.93   2.270  0.02332 *
## Zonecanillo              108123.04   64936.71   1.665  0.09605 .
## Zoneencamp                16250.02   63835.91   0.255  0.79909
## Zoneescaldes-engordany    91004.06   41750.43   2.180  0.02939 *
## Zonela-massana           -11499.50   66321.74  -0.173  0.86236
## Zoneordino               178012.82   77082.05   2.309  0.02102 *
## Zonesant-julia-de-loria         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 569900 on 2116 degrees of freedom
## Multiple R-squared:  0.6344, Adjusted R-squared:  0.6329
## F-statistic:   408 on 9 and 2116 DF,  p-value: < 2.2e-16
```



De l'anàlisis sobre el lloguer s'observa que les variables "Area", "Bedrooms" i "Centric" mostren una significació estadística positiva en el model. Això significa que un augment en l'àrea, el nombre d'habitacions o la proximitat al centre estan associats amb increments en el preu de lloguer. La parròquia (Zone) també és significativa en el model, amb algunes zones com "Canillo" que tenen un impacte positiu significatiu en el preu. L'ajustament global del model és moderat, amb un R2 ajustat d'aproximadament 0.5291, el que indica que aproximadament el 52.91% de la variabilitat en el preu de lloguer es pot explicar pel model.

En l'anàlisi sobre la venda s'observa que un augment en l'àrea i el nombre d'habitacions es correlaciona amb increments en el preu de venda. Algunes zones també tenen un impacte significatiu en el preu de venda. Per exemple, la zona "Ordino" i "Canillo" mostren un impacte positiu significatiu en els preus de venda. L'ajustament global del model de venda és moderat, amb

un R2 ajustat d'aproximadament 0.6329, el que suggereix que aproximadament el 63.29% de la variabilitat en el preu de venda es pot explicar pel model.

En tots dos casos s'observa que la Zona Sant Julià de Lòria conté valors NA, això podria relacionarse amb una quantitat insuficient de dades.

Si s'observen els gràfics de residus, tan en el cas de venta com en el de lloguer s'observa un patró de dispersió irregular, sobretot en el cas dels habitatges en venda. Això indica que no es compleix el supòsit de variança constant en els errors del model, amb lo qual la regressió lineal pot no ser l'ajust més adequat per aquestes dades. Aquest patró també pot indicar que no s'estan considerant variables que podrien resultar rellevants per al model.

En canvi, els QQ plots mostren que les dades s'ajusten bastant bé excepte en les cues, això ens podria indicar que hi ha presencia de valors extrems.

Conclusions

L'estudi detallat del mercat immobiliari andorrà revela una complexa realitat en la disponibilitat d'habitatges i els preus de venda i lloguer. S'ha observat que els preus de venda varien significativament entre parròquies, a diferència dels lloguers que són més uniformes. No obstant això, la superfície dels habitatges no mostra diferències significatives entre parròquies cèntriques i no cèntriques, tot i que la distribució en nombre d'habitacions varia, especialment en la venda.

Les correlacions indiquen una relació forta entre 'Price' i 'Area', especialment en la venda. La ubicació cèntrica i el nombre d'habitacions també influeixen en els preus, però de manera més moderada. La proporció d'habitatges disponibles és baixa, particularment en les zones cèntriques, destacant una problemàtica social d'accés a l'habitatge. Els estudis de contrast confirmen que els preus mitjans de venda i lloguer són significativament més alts en aquestes zones, evidenciant una major demanda o preferència per aquestes localitzacions. A més, l'anàlisi de regressió subratlla la influència de factors com l'àrea, el nombre d'habitacions, i la ubicació en els preus, mentre que l'anàlisi de càrrega financera dels lloguers mostra que en algunes zones, el cost del lloguer representa una part significativa dels ingressos anuals, posant en evidencia una forta pressió financera sobre els residents. Aquestes troballes proporcionen una comprensió detallada de les dinàmiques del mercat immobiliari andorrà i les seves implicacions socioeconòmiques.

Taula de contribucions

Contribucions	Signatura
Investigació prèvia	Laura Guerra, Àlex Tort
Redacció de les respostes	Laura Guerra, Àlex Tort
Desenvolupament del codi	Laura Guerra, Àlex Tort
Participació al vídeo	Laura Guerra, Àlex Tort

Video i repositori Github

El repositori git associat a aquest projecte és https://github.com/Laura-Guerra/TCVD_PR2
L'enllaç del video és: