# M2.851 - Tipología y ciclo de vida de los datos

# Práctica 2: Limpieza y validación de datos

## Autor

José Pérez Sánchez, jperezsanchez

## House Prices: Advanced Regression Techniques

# Table of Contents

# Análisis del precio de venta de inmuebles según distintas características

Se basa en la competición *kaggle* sobre técnicas de regresión avanzadas en precios de viviendas.

House Prices: Advanced Regression Techniques

## 1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

En este dataset se describen distintas características de casas así como el precio de éstas. Se trataría de crear métodos de regresión que nos permitan predecir el precio de una casa a partir de sus características. Esto es útil tanto para futuros compradores, que pueden estimar así si una vivienda está cara o barata según sus características y zona donde se encuentra y para los vendedores, que pueden estimar los precios adecuados para sus viviendas, al menos según el estado del mercado.

El dataset viene descrito en inglés en la propia web de *kaggle*, traducido, los campos son:

MSSubClass: Tipo de vivienda. Valor discreto, factor.

MSZoning: Calificacion urbanística del suelo. Valor discreto, factor.

```
A    Agrícola
C    Comercial
FV   Residencial flotante
I    Industrial
RH   Residencial alta densidad
RL   Residencial baja densidad
RP   Residencial baja densidad, parques
RM   Residencial densidad media
```

LotFrontage: Longitud de la parte de la propiedad que limita con la calle, en pies

LotArea: Tamaño de la propiedad en pies cuadrados

Street: Tipo de carretera de acceso a la propiedad

```
Grvl Grava
Pave Pavimentado
```

Alley: Tipo de acceso final a la propiedad

```
Grvl Grava
Pave Pavimentado
NA   Sin acceso
```

LotShape: Forma genérica de la propiedad

```
   Reg  Regular
   IR1  Ligeramente irregular
   IR2  Moderadamente irregular
   IR3  Irregular
```

LandContour: Nivelación de la propiedad

```
   Lvl  Casi plano.
   Bnk  Escalón - Subida rápida y significativa desde la calle al edificio
   HLS  Colina - Inclinación significativa de lado a lado
   Low  Depresión
```

Utilities: Servicios públicos disponibles

```
   AllPub   Todos (electricidad, gas, agua y alcantarillado)
   NoSewr   Electricidad, gas y agua (fosa séptica)
   NoSeWa   Sólo electricidad y gas
   ELO  Solamente electricidad
```

LotConfig: Configuración de la parcela

```
   Inside   Parcela interior
   Corner   Esquina
   CulDSac  Cul-de-sac, calle sin salida
   FR2  Frontal exterior por dos lados de la propiedad
   FR3  Frontal exterior por tres lados de la propiedad
```

LandSlope: Inclinación de la propiedad

```
   Gtl  Ligera
   Mod  Moderada
   Sev  Severa
```

Neighborhood: Ubicación de la propiedad, entre distintos vecindarios.

```
   Blmngtn  Bloomington Heights
   Blueste  Bluestem
   BrDale   Briardale
   BrkSide  Brookside
   ClearCr  Clear Creek
   CollgCr  College Creek
   Crawfor  Crawford
   Edwards  Edwards
   Gilbert  Gilbert
   IDOTRR   Iowa DOT and Rail Road
   MeadowV  Meadow Village
   Mitchel  Mitchell
   Names    North Ames
   NoRidge  Northridge
   NPkVill  Northpark Villa
```

```
   NridgHt  Northridge Heights
   NWAmes   Northwest Ames
   OldTown  Old Town
   SWISU    South & West of Iowa State University
   Sawyer   Sawyer
   SawyerW  Sawyer West
   Somerst  Somerset
   StoneBr  Stone Brook
   Timber   Timberland
   Veenker  Veenker
```

Condition1: Proximidad a comunicaciones e infrastructuras

```
   Artery   Adyacente a calle principal
   Feedr    Adyacente a calle secundaria
   Norm Normal
   RRNn A menos de 200 minutos del ferrocarril North-South
   RRAn Adyacente al ferrorarril North-South
   PosN cerca de parques, cinturón verde, etc.
   PosA Adyacente a otras atracciones urbanas (parques, etc..)
   RRNe A menos de 200 minutos del ferrocarril East-West
   RRAe Adyacente al ferrocarril East-West
```

Condition2: Proximidad a comunicaciones e infrastructuras (si se da más de una)

```
   Artery   Adyacente a calle principal
   Feedr    Adyacente a calle secundaria
   Norm Normal
   RRNn A menos de 200 minutos del ferrocarril North-South
   RRAn Adyacente al ferrorarril North-South
   PosN cerca de parques, cinturón verde, etc.
   PosA Adyacente a otras atracciones urbanas (parques, etc..)
   RRNe A menos de 200 minutos del ferrocarril East-West
   RRAe Adyacente al ferrocarril East-West
```

BldgType: tipo de vivienda

```
   1Fam Unifamiliar independiente
   2FmCon   Vivienda convertida en dos viviendas, inicialmente construida
como unifamiliar
   Duplx    Dúplex
   TwnhsE   Vivienda adosadada esquina
   TwnhsI   Vivienda adosada ambos lados
```

HouseStyle: Estilo de vivienda

```
   1Story   Una planta
   1.5Fin   Planta y media: Segundo nivel terminado
   1.5Unf   Planta y media: Segundo nivel no terminado
   2Story   Dos plantas
   2.5Fin   Dos plantas y media: último nivel terminado
```

```
   2.5Unf   Dos plantas y media: último nivel no terminado
   SFoyer   Casa con dos plantas y sótano.
   SLvl Vivienda en una planta de una casa.
```

OverallQual: Calidades del material y terminaciones de la casa

```
   10   Sobresaliente
   9    Excelente
   8    Muy bueno
   7    Bueno
   6    Encima de la media
   5    Medio
   4    Bajo la media
   3    Regular
   2    Malo
   1    Muy malo
```

OverallCond: Calificación del estado de la casa

```
   10   Sobresaliente
   9    Excelente
   8    Muy bueno
   7    Bueno
   6    Encima de la media
   5    Medio
   4    Bajo la media
   3    Regular
   2    Malo
   1    Muy malo
```

YearBuilt: Fecha inicial de construcción

YearRemodAdd: Fecha de renovación (la de construcción si no ha sido renovada)

RoofStyle: Tipo de tejado

```
   Flat Plano
   Gable    Tejado inclinado, 2 lados
   Gambrel  Granero
   Hip  Tejado inclinado, 4 lados
   Mansard  Buhardilla
   Shed Cobertizo
```

RoofMatl: Material del tejado

```
   ClyTile  Teja
   CompShg  Grava
   Membran  Membrana
   Metal    Metal
   Roll Rollo de alquitrán
```

```
Tar&Grv  Grava y alquitrán
WdShake  Tablones de madera
WdShngl  Tejas de madera
```

Exterior1st: Material de construcción exterior de la casa

```
AsbShng  Asbesto
AsphShn  Asfalto
BrkComm  Ladrillo normal
BrkFace  Ladrillo visto
CBlock   Bloque de hormigón
CemntBd  Cemento
HdBoard  Hard Board
ImStucc  Imitación estuco
MetalSd  Metalico
Other    Otro
Plywood  Plywood
PreCast  PreCast
Stone    Piedra
Stucco   Estuco
VinylSd  Vinilo
Wd Sdng  Tablones de madera
WdShing  Madera con forma
```

Exterior2nd: Cobertura exterior de la casa (si otro material del anterior)

```
AsbShng  Asbesto
AsphShn  Asfalto
BrkComm  Ladrillo normal
BrkFace  Ladrillo visto
CBlock     Bloque de cemento
CemntBd  Cemento
HdBoard  Hard Board
ImStucc  Imitación estuco
MetalSd  Metalico
Other      Otro
Plywood  Plywood
PreCast  PreCast
Stone      Piedra
Stucco   Estuco
VinylSd  Vinilo
Wd Sdng  Tablones de madera
WdShing  Madera con forma
```

MasVnrType: Tipo de muros interiores

```
BrkCmn   Ladrillo común
BrkFace  Ladrillo visto
CBlock     Bloque de cemento
```

```
   None     Ninguno
   Stone      Piedra
```

MasVnrArea: Área de construcción en pies cuadrados

ExterQual: Calidad del material exterior

```
   Ex   Excelente
   Gd   Bueno
   TA   Normal
   Fa   Regular
   Po   Malo
```

ExterCond: Estado actual del material del exterior

```
   Ex   Excelente
   Gd   Bueno
   TA   Normal
   Fa   Regular
   Po   Malo
```

Foundation: Tipo de cimientos

```
   BrkTil   Ladrillo y azulejo
   CBlock   Piedras, ladrillos, etc..
   PConc    Hormigón
   Slab Losas
   Stone    Piedra
   Wood Madera
```

BsmtQual: Altura de los cimientos

```
   Ex   Excelente (Más de 100 pulgadas)
   Gd   Bueno (Entre 90-99 pulgadas)
   TA   Normal (80-89 pulgadas)
   Fa   Regular (70-79 pulgadas)
   Po   Pobre (menos de 70 pulgadas)
   NA   Sin cimientos
```

BsmtCond: Estado general de los cimientos

```
   Ex   Excelente
   Gd   Bueno
   TA   Normal - Ligeras humedades permitidas
   Fa   Regular - Humedades, pequeñas roturas o asentamientos
   Po   Malo - Asentamientos, roturas o humedades graves
   NA   Sin cimientos
```

BsmtExposure: Estado de la salida o límites del jardin

```
Gd   Buena
Av   Media
Mn   Mínimo
No   Sin señalización del límite en la propiedad
NA   Sin especificar
```

BsmtFinType1: Clasificación de habitabilidad del sótano

```
GLQ  Buenas habitaciones de vivienda
ALQ  Habitaciones normales
BLQ  Habitabilidad inferior al resto de la vivienda
Rec  Habitabilidad media como trastero
LwQ  Baja calidad
Unf  No terminado
NA   Sin sótano
```

BsmtFinSF1: Tamaño del área sótano en pies cuadrados terminados

BsmtFinType2: Clasificación de habitabilidad del sótano (si varios tipos)

```
GLQ  Buenas habitaciones de vivienda
ALQ  Habitaciones normales
BLQ  Habitabilidad inferior al resto de la vivienda
Rec  Habitabilidad media como trastero
LwQ  Baja calidad
Unf  No terminado
NA   Sin sótano
```

BsmtFinSF2: Tamaño de la segundo área sótano en pies cuadrados terminados

BsmtUnfSF: Tamaño del área del sótano no terminada

TotalBsmtSF: Tamaño total del sótano

Heating: Tipo de calefacción

```
Floor    Suelo radiante
GasA Calefacción por aire, usando gas
GasW Calefacción y agua caliente con gas
Grav Horno de gravedad (caldera de aire)
OthW Agua caliente o vapor no usando gas
Wall Muro calefactor
```

HeatingQC: Calidad y estado de la calefacción

```
Ex   Excelente
Gd   Bueno
TA   Normal
Fa   Regular
```

```
   Po   Malo
```

CentralAir: Aire acondicionado

```
   N    No
   Y    Sí
```

Electrical: Instalación eléctrica

```
   SBrkr    Circuito estándar
   FuseA    Caja de fusibles de más de 60 amperios (Normal)
   FuseF    Caja de fusibles de 60 amperios o casi (regular)
   FuseP    Caja de fusibles de menos de 60, (pobre)
   Mix    Mixto
```

1stFlrSF: Tamaño primera planta

2ndFlrSF: Tamaño segunda planta

LowQualFinSF: Área de baja calidad, en todas las plantas

GrLivArea: Área habitable, pies cuadrados últiles.

BsmtFullBath: Baños en el sótano

BsmtHalfBath: Aseos en el sótano

FullBath: Baños sobre planta baja

HalfBath: Aseos encima de planta baja

Bedroom: Dormitorios encima de planta baja o sótano

Kitchen: Cocinas sobre planta baja

KitchenQual: Calidad cocina

```
   Ex   Excelente
   Gd   Buena
   TA   Normal
   Fa   Regular
   Po   Mala
```

TotRmsAbvGrd: Número de habitaciones encima del sótano (no incluye baños)

Functional: Funcionalidad

```
   Typ  Normal
   Min1 Deducciones mínimas 1
   Min2 Deducciones mínimas 2
```

```
   Mod  Deducciones moderadas
   Maj1 Deducciones mayores 1
   Maj2 Deducciones mayores 2
   Sev  Daños severos
   Sal  Sólo reconstrucción
```

Fireplaces: Número de chimeneas, estufas, hogares, etc…

FireplaceQu: Calidad de la hogar (chimenea)

```
   Ex   Excelente - Excepcional, en ladrillo.
   Gd   Bueno - En ladrillo, en la planta principal
   TA   Medio - Prefabricado o de ladrillo en el sótano
   Fa   Regular - Prefabricado en el sótano
   Po   Malo - Horno Ben Franklin (metálico)
   NA   Sin hogar para fuego.
```

GarageType: Ubicación del garaje

```
   2Types   Más de un tipo de garaje
   Attchd   Adosado a la casa
   Basment  Garaje de sótano
   BuiltIn  Empotrado (garaje parte de la casa, normalmente con una
habitación arriba)
   CarPort  Tejadillo para coche
   Detchd   Separado de la casa
   NA   Sin garaje
```

GarageYrBlt: Año de construcción del garaje

GarageFinish: Terminación interior del garaje

```
   Fin  Terminado
   RFn  Casi terminado
   Unf  No terminado
   NA   Sin garaje
```

GarageCars: Tamaño del garaje en capacidad de coches

GarageArea: Tamaño del garaje en pies cuadrados

GarageQual: Calidad del garaje

```
   Ex   Excelente
   Gd   Bueno
   TA   Normal
   Fa   Regular
   Po   Malo
```

```
   NA   Sin garaje
```

GarageCond: Estado del garaje

```
   Ex   Excelente
   Gd   Bueno
   TA   Normal
   Fa   Regular
   Po   Malo
   NA   Sin garaje
```

PavedDrive: Entrada de vehículos pavimentada

```
   Y    Pavimentada
   P    Parcialmente pavimentada
   N    Nada, gravilla
```

WoodDeckSF: Área de terraza de madera en pies cuadrados

OpenPorchSF: Área del porche descubierto en pies cuadrados

EnclosedPorch: Área del porche cubierto en pies cuadrados

3SsnPorch: Tamaño de la veranda de obra en pies cuadrado

ScreenPorch: Área de veranda ligera en pies cuadrados

PoolArea: Tamaño de la piscina, pies cuadrados

PoolQC: Calidad de la piscina

```
   Ex   Excelente
   Gd   Buena
   TA   Normal
   Fa   Regular
   NA   Sin piscina
```

Fence: Calidad de la valla

```
   GdPrv    Buena privacidad
   MnPrv    Privacidad mínima
   GdWo     Buena madera
   MnWw     Mínima madera/alambrada
   NA       Sin valla
```

MiscFeature: Otras características

```
   Elev Ascensor
   Gar2 Segundo garaje (si no incluido en la sección garaje)
```

```
   Othr Otro
   Shed Cobertizo (más de 100 pies cuadrados)
   TenC Pista de tenis
   NA     Nada
```

MiscVal: Coste de otras características

MoSold: Mes de venta (MM)

YrSold: Año de venta (YYYY)

SaleType: Condición de venta

```
   WD     Escritura convencional
   CWD    Escritura de garantía, efectivo
   VWD    Estritura con hipoteca
   New    Casa recién construida y vendida
   COD    Court Officer Deed/Estate
   Con    Contrato con 15% de depósito normal
   ConLw    Contrato con bajo depósito y bajos intereses
   ConLI    Contrato bajo interés
   ConLD    Contrato bajo depósito inicial
   Oth    Otro
```

SaleCondition: Tipo de venta

```
   Normal     Normal
   Abnorml  Anormal - embargo, subasta, etc...
   AdjLand  Compra del solar de la vivienda
   Alloca     Asignación: dos propiedades vinculadas con escrituras separadas
   Family     Venta entre familiares
   Partial  Vivienda no terminada en momento de venta (normalmente para casas
nuevas)
```

## 2. Integración y selección de los datos de interés a analizar.

Los datos proceden de la web de *kaggle*, de donde se descargan en formato CSV. Cada registro tiene un identificador de inmueble, y el resto de los datos de la vivienda.

El objetivo es el poder predecir el precio de una vivienda a partir de esos datos, por ello se comprobarán cuáles de los datos influyen más en el precio de una vivienda, si es la ubicación, el tamaño, el estado de la vivienda, etc…

La carga de datos se realizará descargando el fichero CSV de la url indicada, (en este caso desde la cuenta github) y ejecutando el siguiente código R:

```
# Lectura de datos

#inmuebles <- read.csv("Datos-Inmuebles.csv")
```

```
suppressWarnings(suppressMessages(library(RCurl)))
downF <- getURL("https://raw.githubusercontent.com/jperezsanchezU/house-
prices-advanced-regression-techniques/master/csv/Datos-Inmuebles.csv")
inmuebles <- read.csv(text = downF)

# Comprobación de datos cargados
head(inmuebles[,1:5])
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea
## 1  1         60       RL          65    8450
## 2  2         20       RL          80    9600
## 3  3         60       RL          68   11250
## 4  4         70       RL          60    9550
## 5  5         60       RL          84   14260
## 6  6         50       RL          85   14115
```

En los datos de inmuebles se prestará atención sobre todo a los datos relativos al tamaño de la vivienda y de la parcela, asi como la ubicacion, (vecindario, proximidad a medios de transporte, etc..) y secundariamente a otras caracteristicas, garaje, piscina, estado, etc...

Se analizarán valores extremos de los datos cuantitativos y se trataran si no fuese coherentes con los datos habituales en compra/venta de inmuebles.

Tras la limpieza de datos, se reducirá el conjunto de datos a analizar, en el punto 4, agrupando datos y eliminando aquellos más incompletos o con menor peso en el cálculo del precio de la vivienda, como es la calidad del porche, o el número de chimeneas.

## 3. Limpieza de los datos.

Los tipos de los datos inferidos en su carga son correctos y coherentes conforme a la descripción del conjunto de datos.

```
# Tipo de dato asignado a cada campo

sapply(inmuebles, function(x) class(x))
```

```
##           Id    MSSubClass      MSZoning   LotFrontage       LotArea
##    "integer"     "integer"      "factor"     "integer"     "integer"
##       Street         Alley      LotShape   LandContour     Utilities
##     "factor"      "factor"      "factor"      "factor"      "factor"
##    LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##     "factor"      "factor"      "factor"      "factor"      "factor"
##     BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##     "factor"      "factor"     "integer"     "integer"     "integer"
## YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##    "integer"      "factor"      "factor"      "factor"      "factor"
##   MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##     "factor"     "integer"      "factor"      "factor"      "factor"
##     BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##     "factor"      "factor"      "factor"      "factor"     "integer"
## BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
```

```
##      "factor"     "integer"     "integer"     "integer"      "factor"
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##      "factor"      "factor"      "factor"     "integer"     "integer"
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##     "integer"     "integer"     "integer"     "integer"     "integer"
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##     "integer"     "integer"     "integer"      "factor"     "integer"
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##      "factor"     "integer"      "factor"      "factor"     "integer"
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##      "factor"     "integer"     "integer"      "factor"      "factor"
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##      "factor"     "integer"     "integer"     "integer"     "integer"
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##     "integer"     "integer"      "factor"      "factor"      "factor"
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##     "integer"     "integer"     "integer"      "factor"      "factor"
##     SalePrice
##     "integer"
```

Sólo se detecta un error, en el campo MSSubClass, que determina el tipo del inmueble, pero se detecta como valor cuantitativo. Se convertirá en factor.

```
# Conversión de MSSubClass a factor

inmuebles$MSSubClass <- as.factor(inmuebles$MSSubClass)

class(inmuebles$MSSubClass)

## [1] "factor"
```

### 3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Varios datos cualitativos y cuantitativos vienen con datos NA, en el caso de que no disponga de ese atributo (garaje, etc..). A tratar con reemplazo por 0, o bien filtrando elementos.

Los datos provistos parecen bastante completos, aún asi hay varios casos donde alguno de los datos no esta disponible, *NA*.

En el caso de los datos cualitativos, tipo factor, el significado del indicador *NA* está explicado en la descripción del dataset y no precisa tratamiento. Suele corresponder con la ausencia del dato, por ejemplo, la ausencia de garaje.

Respecto a los datos cuantitativos está el indicador *NA* que indican datos no disponibles, y que habrá que tratar adecuadamente.

Se verifican en qué valores cuantitativos hay indicadores *NA*

```
# Tipo de dato asignado a cada campo
```

```
fun<-function(x){
ifelse(is.factor(x),0,{sum(is.na(x))})
}

sapply(inmuebles, function(x) fun(x))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0           259             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0             0             0             0             0
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             0             0             0             0             0
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##             0             8             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##             0             0             0             0             0
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##             0             0             0             0             0
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0             0             0            81
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##             0             0             0             0             0
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

Los datos que interesarian tratar son los cuantitativos que valores *NA*:

LotFrontage: 259 elementos. MasVnrArea: 8 elementos. GarageYrBlt: 81 elementos

El tratamiento será el reemplazo de los valores *NA* por el valor calculado a partir de los *k vecinos más próximos*, imputación kNN, empleado la libreria VIM.

```
# Tipo de dato asignado a cada campo

suppressWarnings(suppressMessages(library(VIM)))
```

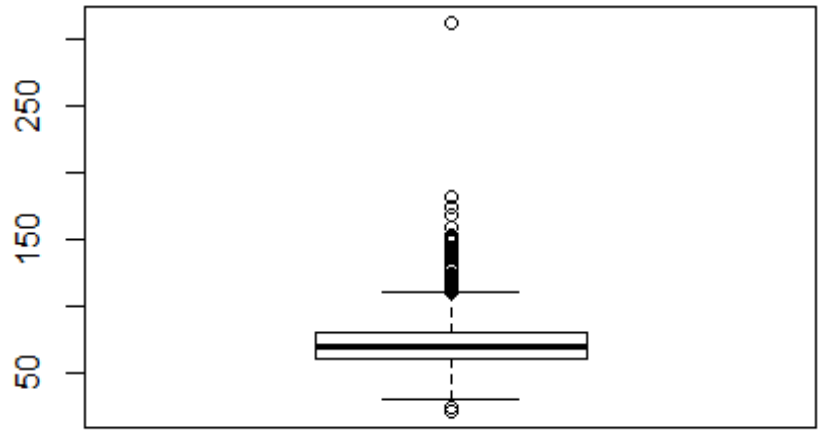
```
inmuebles$LotFrontage <- kNN(inmuebles)$LotFrontage

inmuebles$MasVnrArea <- kNN(inmuebles)$MasVnrArea

inmuebles$GarageYrBlt <- kNN(inmuebles)$GarageYrBlt

sapply(inmuebles, function(x) fun(x))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0             0             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0             0             0             0             0
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             0             0             0             0             0
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##             0             0             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##             0             0             0             0             0
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##             0             0             0             0             0
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0             0             0             0
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##             0             0             0             0             0
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

### 3.2. Identificación y tratamiento de valores extremos.

Para identificar los valores extremos se emplearán boxplots, con esa función se podrán representar graficamente estos valores y mostrarlos, aplicándolo a los datos cuantitaivos:

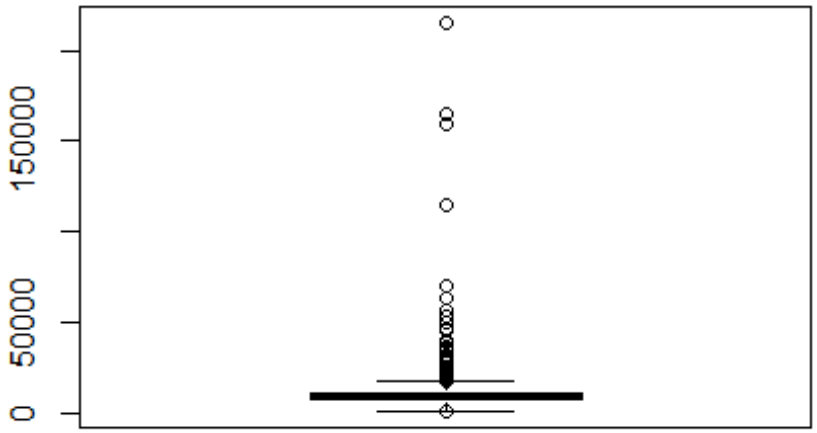
```r
# Ejecución de boxplot

boxplot(inmuebles$LotFrontage)
```




```r
boxplot.stats(inmuebles$LotFrontage)$out
```

```
##  [1] 112 115  24  21 121 122  24 120 134 141  24  24 174  21  21 174  21
## [18]  21 120 129 140 120 118 116 150 111  21 114 130  21  24  21 137  21
## [35]  21  24 130  24  21  21  21 120  24  24 144 114  24  21 128 116 149
## [52]  21 313  24  24  24 122 130 121  21 115  21  21  21 120  21  24  24
## [69]  24 114 168 182 134  24 120 118 138 160  24 152  21 124  21 313  24
## [86] 153 120 129 124  21  21
```

```r
boxplot(inmuebles$LotArea)
```




```
boxplot.stats(inmuebles$LotArea)$out
```

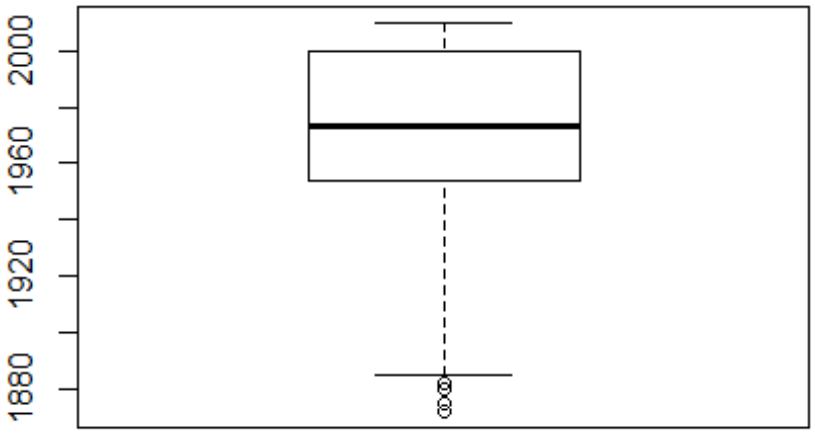
```
##  [1]  50271  19900  21000  21453  19378  31770  22950  25419 159000  19296
## [11]  39104  19138  18386 215245 164660  20431  18800  53107  34650  22420
## [21]  21750  70761  53227  40094  32668  21872  21780  25095  46589  20896
## [31]  18450  21535  26178 115149  21695  53504  21384  28698  45600  17920
## [41]  25286  27650  24090  25000   1300  21286  21750  29959  18000  23257
## [51]  17755  35760  18030  35133  32463  18890  24682  23595  17871  36500
## [61]  63887  20781  25339  57200  20544  19690  21930  26142
```

```
boxplot.stats(inmuebles$LandContour)$out
```

```
## Warning in Ops.factor(x[floor(d)], x[ceiling(d)]): '+' not meaningful for
## factors
```

```
## factor(0)
## Levels: Bnk HLS Low Lvl
```

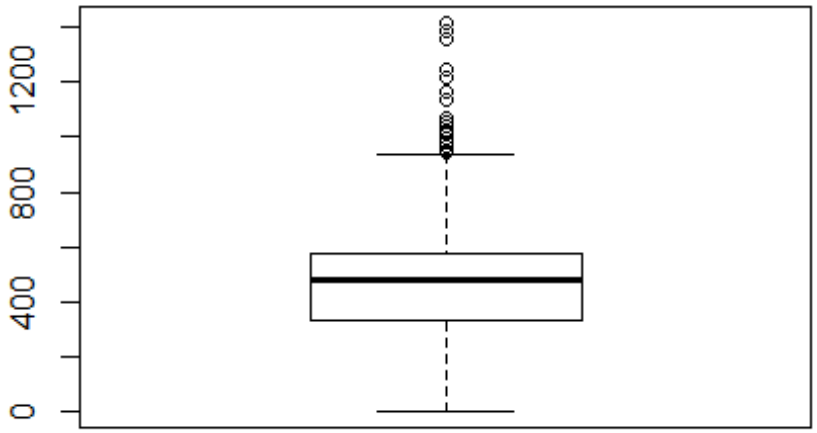
```
boxplot(inmuebles$YearBuilt)
```




```
boxplot.stats(inmuebles$YearBuilt)$out
```

```
## [1] 1880 1880 1880 1882 1880 1875 1872
```

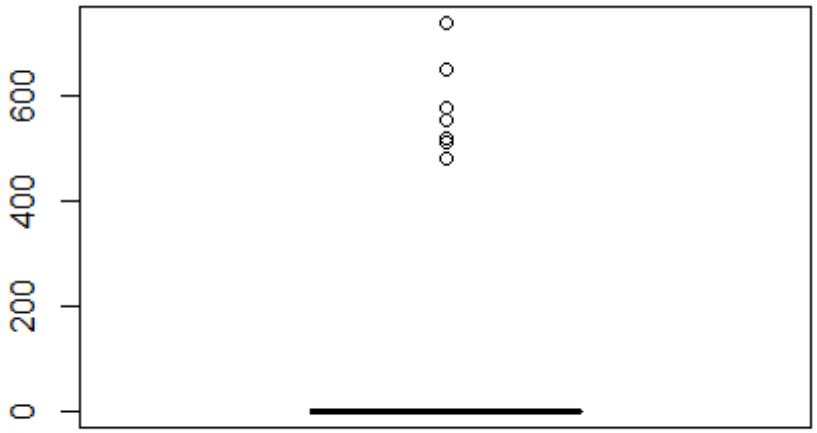
```
boxplot(inmuebles$GarageArea)
```




```
boxplot.stats(inmuebles$GarageArea)$out
```

```
##  [1] 1166  968 1053 1025  947 1390 1134  983 1020 1220 1248 1043 1052  995
## [15] 1356 1052  954 1014 1418  968 1069
```

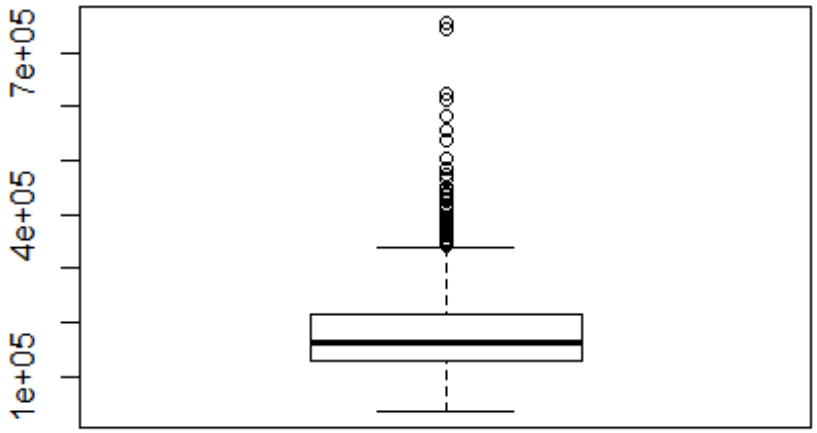
```
boxplot(inmuebles$PoolArea)
```




```
boxplot.stats(inmuebles$PoolArea)$out
```

```
## [1] 512 648 576 555 480 519 738
```

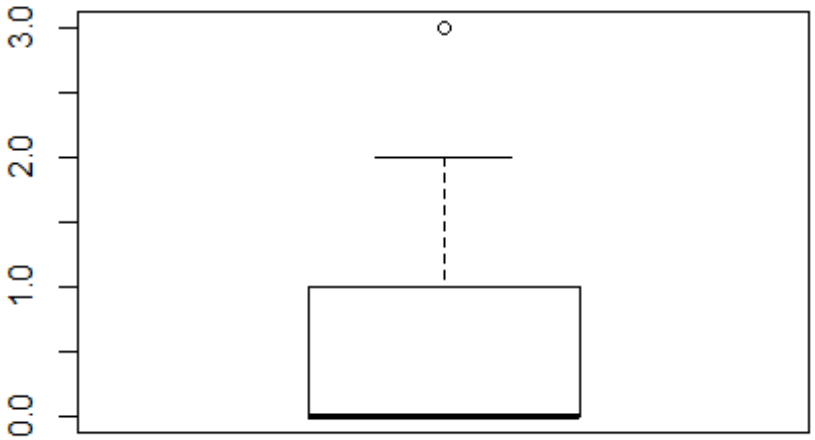
```
boxplot(inmuebles$SalePrice)
```




```
boxplot.stats(inmuebles$SalePrice)$out
```

```
##  [1] 345000 385000 438780 383970 372402 412500 501837 475000 386250 403000
## [11] 415298 360000 375000 342643 354000 377426 437154 394432 426000 555000
## [21] 440000 380000 374000 430000 402861 446261 369900 451950 359100 345000
## [31] 370878 350000 402000 423000 372500 392000 755000 361919 341000 538000
## [41] 395000 485000 582933 385000 350000 611657 395192 348000 556581 424870
## [51] 625000 392500 745000 367294 465000 378500 381000 410000 466500 377500
## [61] 394617
```

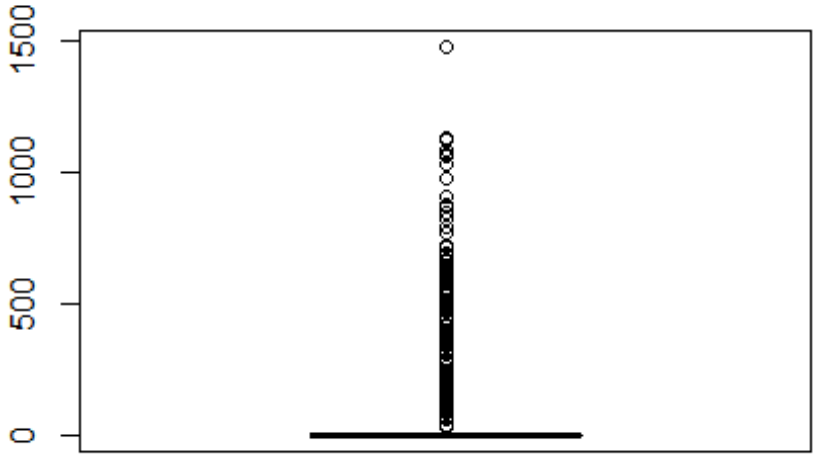
```
boxplot(inmuebles$BsmtFullBath)
```




```
boxplot.stats(inmuebles$BsmtFullBath)$out
```

```
## [1] 3
```

```
boxplot(inmuebles$BsmtFinSF2)
```




```
boxplot.stats(inmuebles$BsmtFinSF2)$out
```

```
##   [1]   32  668  486   93  491  506  712  362   41  169  869  150  670
28
##  [15] 1080  181  768  215  374  208  441  184  279  306  180  712  580
690
##  [29]  692  228  125 1063  620  175  820 1474  264  479  147  232  380
544
##  [43]  294  258  121  180  391  531  344  539  713  210  311 1120  165
532
##  [57]  279   96  495  180  174 1127  139  202  645  123  551  219  606
147
##  [71]  612  480  182  132  336  468  287   35  499  180  180  723  119
182
##  [85]   40  551  117  239   80  472   64 1057  127  630  480  128  377
764
##  [99]  345  539 1085  435  823  500  290  324  634  411  841 1061   93
466
## [113]  396  354  294  149  193  117  273  465  400  468   41  682   64
557
## [127]  230  106  791  240  287  547  391  469  177  108  374  600  492
211
## [141]  168   96 1031  438  375  144   81  906  608  276  661   68  173
972
## [155]  105  420  469  546  334  352  872  374  110  627  163 1029  290
```

Se han encontrado pocos valores *outliers* dentro de los datos, los mas relativos al area del garage o a la piscina, pues no muchas propiedades la tenían. Sin embargo estos datos no serán demasiado importante en el análisis de regresión sobre el precio, pues se centrara en el tamaño de la propiedad, vivienda y en la ubicacion. El año de construcción tampoco sera determinante. Hay que tener en cuenta que hay datos extremos en variables como la superficie de una tercera planta, o un segundo sótano, ya que la mayoría de las viviendas no tienen esos datos, o no tienen aseos en el sótano, etc.., con lo que una mayoría está a 0, mientras los valores extremos son las viviendas que sí tienen.

Para evitar este problema, algunos de esos campos, como el númeor de chimeneas, no entrará en el análisis, mientras que en el caso de las superficies, o número total de aseos o baños, se sumarán todos los de la vivienda, independientemente de dónde se encuentre.

Sin embargo, en general, no parecen datos erróneos, como podria ser valores multiplicados por 1000, etc.., por lo que no habra acción más acciones correctivas que una agrupación de datos para algunas variables.

## 4. Análisis de los datos.

### 4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Como, en principio, las casas o parcelas mayores serán más caras, se incluyen los datos de precio por pie cuadrado, para estandarizar mejor los precios. Se comprobara también si el tamaño total influye en el precio por pie cuadrado.

Según los principales fondos inmobiliarios, la característica clave que determina el valor de de una propiedad, además de su tamaño, es la ubicación, además, de todos los datos disponibles, lo que se tendrá en cuenta serán los distintos tamaños: parcela, tamaño útil de la vivienda, el tipo de la misma, garajes, materiales de construccion y calidades o estado, sobre todo de cimientos.

Para los restantes análisis emplearemos este subconjunto de los datos de inmuebles:

MSSubClass: Tipo de vivienda

MSZoning: Calificacion urbanística del suelo.

Neighborhood: Ubicación de la propiedad, entre distintos vecindarios.

Condition1: Proximidad a comunicaciones e infrastructuras.

OverallQual: Calidades del material y terminaciones de la casa.

OverallCond: Calificación del estado de la casa.

BsmtCond: Estado general de los cimientos.

YearBuilt: Año de construcción.

LotArea: Tamaño de la parcela.

GarageCars: Tamaño del garaje en coches

GarageArea: Área del garaje.

Bedroom: número de dormitorios

AllHalfBath: suma del total de aseos, BsmtHalfBath + HalfBath

AllBath: suma total de baños completos, BsmtFullBath + FullBath

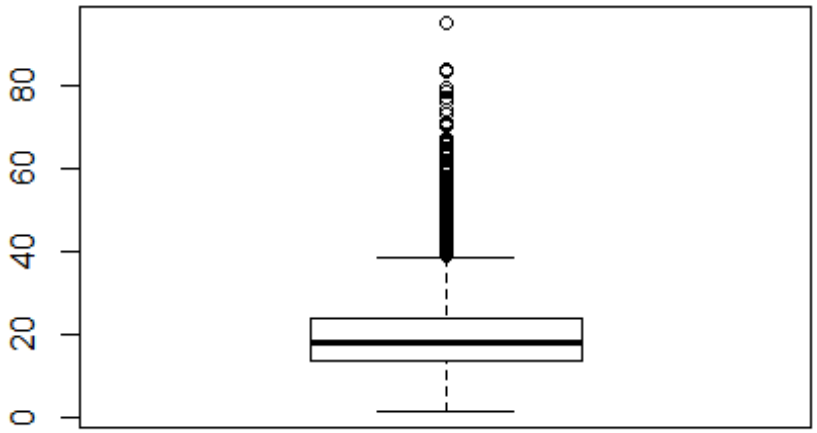
GrLivArea: Tamaño del área habitable.

```
# Selección de variables para el análisis y modelo

inmuebles["AllHalfBath"] <- NA

inmuebles$AllHalfBath <- inmuebles$BsmtHalfBath + inmuebles$HalfBath

inmuebles["AllBath"] <- NA

inmuebles["AllBath"] <- inmuebles$BsmtFullBath + inmuebles$FullBath
```

Se analizará la normalidad y la homogeneidad de los datos cuantitativos seleccionados.

Se añade el valor del pie cuadrado en base al tamaño de la parcela, *PriceLotSquareFoot* y se calcula sus valores *outliers*. Se utilizará tambien el precio del pie cuadrado del area construida, *GrLivArea*, dólares por pie cuadrado habitable, *PriceLotGrLivArea*.

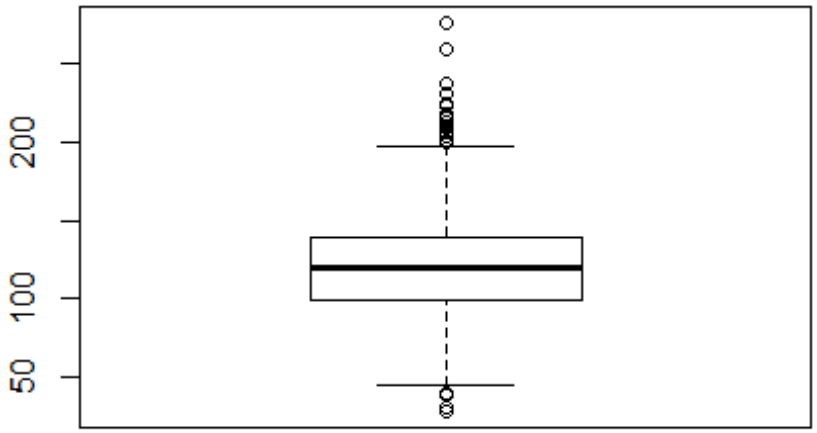
```
# Ejecución de boxplot


inmuebles["PriceLotSquareFoot"] <- NA
inmuebles$PriceLotSquareFoot <- (inmuebles$SalePrice / inmuebles$LotArea)

boxplot(inmuebles$PriceLotSquareFoot)
```




```
boxplot.stats(inmuebles$PriceLotSquareFoot)$out
```

```
##   [1] 41.77331 65.21739 57.01754 41.63503 54.48916 51.54639 45.04335
##   [8] 83.60888 58.09466 51.54639 65.13158 55.56146 66.66667 56.71482
##  [15] 60.49654 56.25000 53.27381 49.47368 40.98361 62.90850 40.67276
##  [22] 70.23810 60.33941 50.42017 39.96448 50.83333 63.80208 42.85714
##  [29] 43.98266 58.18359 40.27211 47.09193 62.00000 56.35649 43.15197
##  [36] 59.78836 63.69151 73.21652 40.51054 49.51581 77.43590 47.45443
##  [43] 50.63716 79.59376 45.64270 66.43923 40.36987 43.64669 52.38095
##  [50] 64.87549 59.92380 40.99170 48.46087 45.77300 40.70513 40.46243
##  [57] 41.90157 43.88569 53.40557 78.68421 50.34569 41.96480 57.94025
##  [64] 47.48428 59.52381 67.27159 39.61965 64.25083 47.34554 57.91100
##  [71] 95.38462 60.26439 67.15771 76.27866 62.24066 46.77117 50.24984
##  [78] 38.90068 56.88246 39.69328 70.85629 60.02514 70.23810 63.27462
##  [85] 54.16385 43.39431 42.29990 54.52022 59.40893 74.07407 45.99567
##  [92] 52.57937 47.23127 42.63623 45.99080 64.03509 42.79313 70.25237
##  [99] 47.68610 65.78450 54.46429 41.88551 66.78082 54.76722 44.58128
## [106] 49.23695 49.08192 71.13095 49.38534 52.78716 83.84506 45.33082
## [113] 47.65478 40.92105 42.49872 60.95871 71.57730 48.39065 60.01305
## [120] 39.45578
```

```
inmuebles["PriceLotGrLivArea"] <- NA
inmuebles$PriceLotGrLivArea <- (inmuebles$SalePrice / inmuebles$GrLivArea)

boxplot(inmuebles$PriceLotGrLivArea)
```




```
boxplot.stats(inmuebles$PriceLotGrLivArea)$out
```

```
##  [1]  30.37206 209.01194 217.77895 224.63608 212.51724 231.05745 222.67206
##  [8] 208.70536 199.43614  39.51027 210.01019 201.71674 276.25088 200.20274
## [15] 206.56733 258.73816  39.15289 203.70722  38.51091 237.59080 215.80141
## [22] 223.98844  28.35874 216.20848 204.25311
```

Se comprueba que el valor mas comun del pie cuadrado ronda los 20 dólares, mientras que el pie cuadrado construido/habitable, está entre los 100 y 140 dólares. Pero se verá que hay bastantes diferencias entre los distintos vecindarios. Curiosamente hay muchos menos valores *outliers* en el precio por superficie habitable/construida que por superficie total.

### 4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Comprobación de normalidad de las variables cuantitativas:

```
suppressWarnings(suppressMessages(library(nortest)))

alpha = 0.05
col.names = colnames(inmuebles)

for (i in 1:ncol(inmuebles)) {
  if (i == 1) cat("Variables que no siguen distribución normal:\n")
  if (is.integer(inmuebles[,i]) | is.numeric(inmuebles[,i])) {
    p_val = ad.test(inmuebles[,i])$p.value

    if (p_val < alpha) {
```

```
      cat(col.names[i])
      # Format output
      if (i < ncol(inmuebles)-1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no siguen distribución normal:
## Id, LotFrontage, LotArea, OverallQual,
## OverallCond, YearBuilt, YearRemodAdd,
## MasVnrArea,
## BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF,
## X1stFlrSF, X2ndFlrSF,
## LowQualFinSF, GrLivArea, BsmtFullBath,
## BsmtHalfBath, FullBath, HalfBath,
## BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces,
## GarageYrBlt,
## GarageCars, GarageArea,
## WoodDeckSF, OpenPorchSF, EnclosedPorch,
## X3SsnPorch, ScreenPorch, PoolArea,
## MiscVal, MoSold, YrSold,
## SalePrice,
## AllHalfBath, AllBath, PriceLotSquareFoot
## PriceLotGrLivArea
```

Análisis de la varianza.

Analizamos los precios por pie cuadrado respecto al vecindario, tipo de vivienda y cercanía a infraestructuras

```
fligner.test(inmuebles$PriceLotSquareFoot  ~ inmuebles$Neighborhood)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  inmuebles$PriceLotSquareFoot by inmuebles$Neighborhood
## Fligner-Killeen:med chi-squared = 196, df = 24, p-value < 2.2e-16
```

```
fligner.test(inmuebles$PriceLotGrLivArea  ~ inmuebles$Neighborhood)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  inmuebles$PriceLotGrLivArea by inmuebles$Neighborhood
## Fligner-Killeen:med chi-squared = 83.367, df = 24, p-value =
## 1.754e-08
```

```
fligner.test(inmuebles$PriceLotSquareFoot  ~ inmuebles$MSSubClass)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
```

```
##
## data:  inmuebles$PriceLotSquareFoot by inmuebles$MSSubClass
## Fligner-Killeen:med chi-squared = 204.4, df = 14, p-value <
## 2.2e-16
```

```
fligner.test(inmuebles$PriceLotGrLivArea  ~ inmuebles$MSSubClass)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  inmuebles$PriceLotGrLivArea by inmuebles$MSSubClass
## Fligner-Killeen:med chi-squared = 45.106, df = 14, p-value =
## 3.92e-05
```

Desafortunadamente los p-valor que obtenemos son muy pequeños, cercanos a 0, lo que nos indica que las varianzas son poco homogéneas. Es razonable ya que cada vivienda es, en cierto modo, muy distinta de la otra, en calidades, años de construcción, distribución, garajes, piscinas, y es más complicado en este caso la homogeneidad que en otros casos.

También hay que tener en cuenta que los precios de venta corresponden al año de la misma, con lo que actualizarlos a dólares constantes requeriría combinar los datos con la inflación, o incluso con la inflación en inmobiliario.

### 4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

Se trata de descubrir qué características de la vivienda afectan más al cálculo final del precio de la vivienda.

Se aplicará un modelo de regresion lineal múltiple tomando la variable dependiente precio, salePrice, y como variables explicativas la ubicacion, Neighborhood, tamaño de la parcela, el tipo de la misma y la cercanía a infrastructuras.

Aquí se puede ver la media de precios por pie cuadrado por cada vecindario, y cómo la ubicación sería la variable determinante.

```
## Classes 'rowwise_df', 'tbl_df', 'tbl' and 'data.frame':  25 obs. of  2
variables:
##  $ Neighborhood: Factor w/ 25 levels "Blmngtn","Blueste",..: 1 2 3 4 5 6 7
8 9 10 ...
##  $ mean        :List of 25
##   ..$ : num 194871
##   ..$ : num 137500
##   ..$ : num 104494
##   ..$ : num 124834
##   ..$ : num 212565
##   ..$ : num 197966
##   ..$ : num 210625
##   ..$ : num 128220
```

```
##   ..$ : num 192855
##   ..$ : num 1e+05
##   ..$ : num 98576
##   ..$ : num 156270
##   ..$ : num 145847
##   ..$ : num 335295
##   ..$ : num 142694
##   ..$ : num 316271
##   ..$ : num 189050
##   ..$ : num 128225
##   ..$ : num 136793
##   ..$ : num 186556
##   ..$ : num 225380
##   ..$ : num 310499
##   ..$ : num 142591
##   ..$ : num 242247
##   ..$ : num 238773
##  - attr(*, "vars")= chr "Neighborhood"
##  - attr(*, "drop")= logi TRUE

## Source: local data frame [25 x 2]
## Groups: <by row>
##
## # A tibble: 25 x 2
##    Neighborhood      mean
##  *       <fctr>    <list>
##  1      Blmngtn <dbl [1]>
##  2      Blueste <dbl [1]>
##  3       BrDale <dbl [1]>
##  4      BrkSide <dbl [1]>
##  5      ClearCr <dbl [1]>
##  6      CollgCr <dbl [1]>
##  7      Crawfor <dbl [1]>
##  8      Edwards <dbl [1]>
##  9      Gilbert <dbl [1]>
## 10       IDOTRR <dbl [1]>
## # ... with 15 more rows

## Classes 'rowwise_df', 'tbl_df', 'tbl' and 'data.frame':  25 obs. of  2
variables:
##  $ Neighborhood: Factor w/ 25 levels "Blmngtn","Blueste",..: 1 2 3 4 5 6 7
8 9 10 ...
##  $ mean        :List of 25
##   ..$ : num 57.6
##   ..$ : num 86.4
##   ..$ : num 58.3
##   ..$ : num 17.5
##   ..$ : num 13.8
##   ..$ : num 21.4
##   ..$ : num 21
```

```
##   ..$ : num 15.5
##   ..$ : num 18.6
##   ..$ : num 13.2
##   ..$ : num 45.4
##   ..$ : num 16.4
##   ..$ : num 15.1
##   ..$ : num 25.7
##   ..$ : num 49.6
##   ..$ : num 31.4
##   ..$ : num 16.5
##   ..$ : num 17
##   ..$ : num 14
##   ..$ : num 19
##   ..$ : num 35.5
##   ..$ : num 36.1
##   ..$ : num 18.6
##   ..$ : num 18.4
##   ..$ : num 18.1
##  - attr(*, "vars")= chr "Neighborhood"
##  - attr(*, "drop")= logi TRUE
```

```
## Source: local data frame [25 x 2]
## Groups: <by row>
##
## # A tibble: 25 x 2
##    Neighborhood      mean
##  *       <fctr>    <list>
##  1      Blmngtn <dbl [1]>
##  2      Blueste <dbl [1]>
##  3       BrDale <dbl [1]>
##  4      BrkSide <dbl [1]>
##  5      ClearCr <dbl [1]>
##  6      CollgCr <dbl [1]>
##  7      Crawfor <dbl [1]>
##  8      Edwards <dbl [1]>
##  9      Gilbert <dbl [1]>
## 10       IDOTRR <dbl [1]>
## # ... with 15 more rows
```

```
## Classes 'rowwise_df', 'tbl_df', 'tbl' and 'data.frame':  25 obs. of  2
variables:
##  $ Neighborhood: Factor w/ 25 levels "Blmngtn","Blueste",..: 1 2 3 4 5 6 7
8 9 10 ...
##  $ mean        :List of 25
##   ..$ : num 137
##   ..$ : num 99
##   ..$ : num 92
##   ..$ : num 106
##   ..$ : num 124
##   ..$ : num 137
```

```
##   ..$ : num 120
##   ..$ : num 102
##   ..$ : num 119
##   ..$ : num 89.1
##   ..$ : num 102
##   ..$ : num 126
##   ..$ : num 116
##   ..$ : num 132
##   ..$ : num 117
##   ..$ : num 165
##   ..$ : num 112
##   ..$ : num 91.4
##   ..$ : num 118
##   ..$ : num 120
##   ..$ : num 141
##   ..$ : num 165
##   ..$ : num 84.8
##   ..$ : num 140
##   ..$ : num 155
##  - attr(*, "vars")= chr "Neighborhood"
##  - attr(*, "drop")= logi TRUE

## Source: local data frame [25 x 2]
## Groups: <by row>
##
## # A tibble: 25 x 2
##    Neighborhood      mean
##  *       <fctr>    <list>
##  1      Blmngtn <dbl [1]>
##  2      Blueste <dbl [1]>
##  3       BrDale <dbl [1]>
##  4      BrkSide <dbl [1]>
##  5      ClearCr <dbl [1]>
##  6      CollgCr <dbl [1]>
##  7      Crawfor <dbl [1]>
##  8      Edwards <dbl [1]>
##  9      Gilbert <dbl [1]>
## 10       IDOTRR <dbl [1]>
## # ... with 15 more rows
```

Se va a comprobar como afectan las Variables cuantitativas al precio por pie cuadrado. Dado que las variables no corresponden a una distribucion normal, se empleara el test de Spearman.

Influencia en el precio por pie cuadrado:

```
##      variable      estimate               p-value
## [1,] "GarageCars"  "0.352251620555821"   "6.71741617692356e-44"
## [2,] "GarageArea"  "0.281497390818688"   "5.36551524720085e-28"
## [3,] "Bedroom"     "-0.0797632991518497" "0.00228840257717962"
## [4,] "AllHalfBath" "0.186394459508796"   "7.02104579230982e-13"
```

```
## [5,] "AllBath"     "0.347647184878147"   "9.89780369172683e-43"
## [6,] "GrLivArea"   "0.278158676876618"   "2.37982314488326e-27"
```

En este caso el número de plazas de garaje influye má que el número de dormitorios o aseos, curiosamente a nivel parecido al de los baños.

Influencia en el precio por pie cuadrado construido:

```
##      variable      estimate             p-value
## [1,] "LotArea"     "0.108161139786925"  "3.45122562945497e-05"
## [2,] "GarageCars"  "0.376229411689752"  "2.63945539874938e-50"
## [3,] "GarageArea"  "0.371980259242996"  "3.95146387007142e-49"
## [4,] "Bedroom"     "-0.323834006107162" "5.40357127039086e-37"
## [5,] "AllHalfBath" "-0.105760077411686" "5.14400946463704e-05"
## [6,] "AllBath"     "0.308815266476888"  "1.24725549546672e-33"
```

Por pie cuadrado construido, la influencia de las variables es parecida, de nuevo los garajes incluyen, así como los dormitorios y baños, y a más distancia los aseos.

Esto nos indica que un baño completo será más influyente en el precio que varios aseos (WCs), y que la distribución del garaje, cuántas plazas efectivas para vehículos, importa más que su tamaño total (que puede desperdiciarse por mala distribución, columnas, etc...)

### Modelado

A partir de los datos comprobados, se va a intentar crear un modelo de regresión donde la variable dependiente será el precio de venta, "SalesPrice" y las explicativas son el tamaño total de la parcela, LotArea, el tamaño construido/habitable, "GrLivArea", número de dormitorios, "Bedroom" y el número de garajes "GarageCars" y baños, "AllBath" y proximidad a infraestructuras "Condition1".

Se comprobará si hay alguno de los regresores que tiene influencia significativa (pvalor del contraste individual inferior al 5%).

```
#Regresion lineal

modeloRegLineal  <-  lm (formula = PriceLotSquareFoot ~ LotArea + GrLivArea +
Neighborhood +  GarageCars + AllBath + Condition1, family = binomial,
data=inmuebles)

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...)
:
##  extra argument 'family' will be disregarded

summary(modeloRegLineal)

##
## Call:
## lm(formula = PriceLotSquareFoot ~ LotArea + GrLivArea + Neighborhood +
##     GarageCars + AllBath + Condition1, data = inmuebles, family =
binomial)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.357  -3.745  -0.785   2.255  46.878
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.958e+01  2.349e+00  21.105  < 2e-16 ***
## LotArea             -3.522e-04  2.197e-05 -16.032  < 2e-16 ***
## GrLivArea            1.298e-03  5.034e-04   2.579  0.01000 *
## NeighborhoodBlueste  2.823e+01  5.524e+00   5.111 3.63e-07 ***
## NeighborhoodBrDale   2.376e+00  2.609e+00   0.910  0.36274
## NeighborhoodBrkSide -3.584e+01  2.098e+00 -17.082  < 2e-16 ***
## NeighborhoodClearCr -3.497e+01  2.366e+00 -14.783  < 2e-16 ***
## NeighborhoodCollgCr -3.470e+01  1.902e+00 -18.248  < 2e-16 ***
## NeighborhoodCrawfor -3.358e+01  2.107e+00 -15.936  < 2e-16 ***
## NeighborhoodEdwards -3.885e+01  1.989e+00 -19.534  < 2e-16 ***
## NeighborhoodGilbert -3.657e+01  1.995e+00 -18.332  < 2e-16 ***
## NeighborhoodIDOTRR  -3.982e+01  2.230e+00 -17.856  < 2e-16 ***
## NeighborhoodMeadowV -1.122e+01  2.580e+00  -4.346 1.48e-05 ***
## NeighborhoodMitchel -3.829e+01  2.099e+00 -18.246  < 2e-16 ***
## NeighborhoodNAmes   -3.878e+01  1.892e+00 -20.494  < 2e-16 ***
## NeighborhoodNoRidge -3.157e+01  2.190e+00 -14.413  < 2e-16 ***
## NeighborhoodNPkVill -8.118e+00  3.049e+00  -2.662  0.00785 **
## NeighborhoodNridgHt -2.576e+01  1.997e+00 -12.895  < 2e-16 ***
## NeighborhoodNWAmes  -3.837e+01  2.021e+00 -18.982  < 2e-16 ***
## NeighborhoodOldTown -3.698e+01  1.975e+00 -18.729  < 2e-16 ***
## NeighborhoodSawyer  -3.937e+01  2.035e+00 -19.346  < 2e-16 ***
## NeighborhoodSawyerW -3.685e+01  2.065e+00 -17.845  < 2e-16 ***
## NeighborhoodSomerst -2.124e+01  1.969e+00 -10.788  < 2e-16 ***
## NeighborhoodStoneBr -2.131e+01  2.339e+00  -9.109  < 2e-16 ***
## NeighborhoodSWISU   -3.702e+01  2.395e+00 -15.457  < 2e-16 ***
## NeighborhoodTimber  -3.393e+01  2.203e+00 -15.402  < 2e-16 ***
## NeighborhoodVeenker -3.547e+01  2.883e+00 -12.304  < 2e-16 ***
## GarageCars           1.323e-01  3.539e-01   0.374  0.70860
## AllBath              2.512e+00  3.453e-01   7.274 5.76e-13 ***
## Condition1Feedr     -6.911e-02  1.398e+00  -0.049  0.96058
## Condition1Norm       2.070e+00  1.148e+00   1.804  0.07147 .
## Condition1PosA      -7.084e-01  2.893e+00  -0.245  0.80661
## Condition1PosN       2.302e+00  2.067e+00   1.114  0.26559
## Condition1RRAe      -1.330e+00  2.580e+00  -0.516  0.60622
## Condition1RRAn      -2.158e+00  1.894e+00  -1.139  0.25470
## Condition1RRNe       3.395e+00  5.440e+00   0.624  0.53268
## Condition1RRNn      -9.730e-01  3.531e+00  -0.276  0.78293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.387 on 1423 degrees of freedom
## Multiple R-squared:  0.6531, Adjusted R-squared:  0.6444
## F-statistic: 74.43 on 36 and 1423 DF,  p-value: < 2.2e-16
```

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

```
## [1] 0.6531306
```

De nuevo se comprueba que la ubicación, por vecindario o por cercanía a infraestructuras tiene un peso mayor en el precio por pie cuadrado que el número de garajes o baños. El tamaño total de la parcela tampoco influye casi nada en el precio por pie cuadrado (aunque, obviamente, sí en el precio total de venta.) Aún así, el coeficiente r cuadrado no es muy alto, y cabría buscar cierta mejora en el modelo:

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...)
:
##  extra argument 'family' will be disregarded

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...)
:
##  extra argument 'family' will be disregarded

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...)
:
##  extra argument 'family' will be disregarded

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...)
:
##  extra argument 'family' will be disregarded

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...)
:
##  extra argument 'family' will be disregarded

## [1] 0.6531306

## [1] 0.5972421

## [1] 0.6056961

## [1] 0.620086

## [1] 0.6027803

## [1] 0.1798096
```

Comprobando otras combinaciones de modelo, aún la primera, basada en los cálculos anteriores, seguiría siendo la más fiable. Se comprueba cómo datos como el año de construcción tendría menos influencia que, por ejemplo, el estado de la vivienda, pero, en el último modelo se verifica cómo la ubicación, vecindario, es lo que determina el precio por pie cuadrado. El precio total vendría dado por éste y el tamaño de la propiedad.

Ejemplo de predicciones con su valor real:

```
##          Real estimado
## [1,] 24.67456 23.99145
```

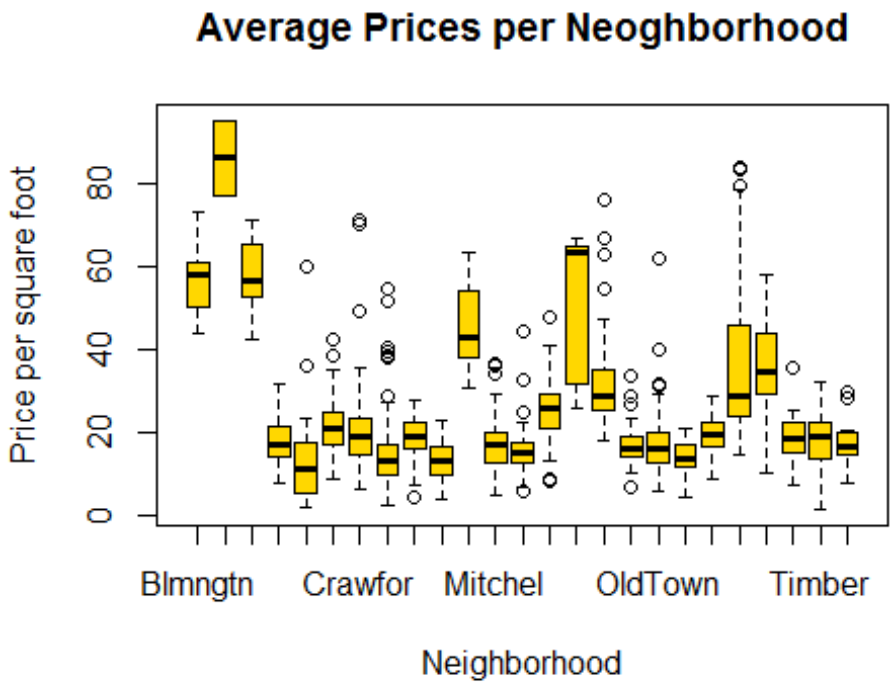
```
## [2,] 13.83584 16.20205
## [3,] 28.66229 32.67312
## [4,] 11.32143 17.28351
## [5,] 29.66154 37.36031
```

## 5. Representación de los resultados a partir de tablas y gráficas.

Las librerías gráficas de R nos permiten una gran variedad de gráficas, se mostrarán algunas relacionadas con el análisis de precio realizado.

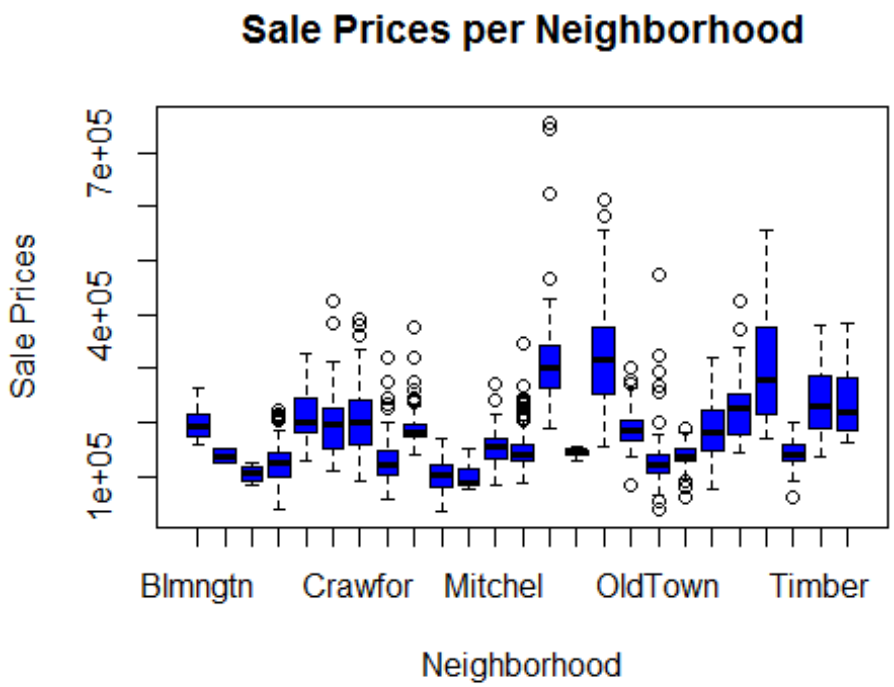
Dado que el vecindario es de los factores clave, es interesante ver la dispersión de precios por vecindario y entre ellos, se aprecia claramente con estos boxplot:

```
boxplot(inmuebles$PriceLotSquareFoot~inmuebles$Neighborhood,col="gold",
        xlab="Neighborhood",ylab="Price per square foot", main="Average
Prices per Neoghborhood")
```



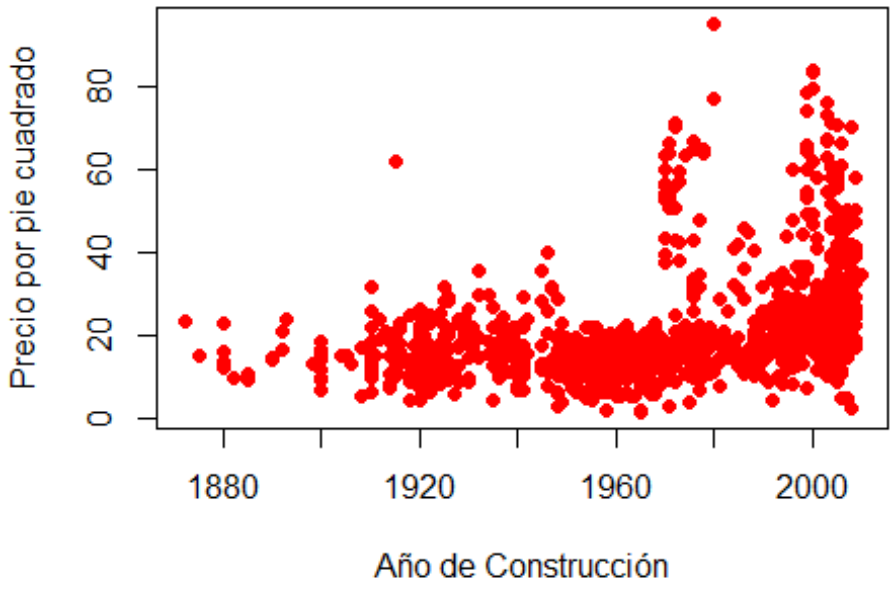

Si observamos los mismo datos por precios de venta se visualiza la diferencia, dada por el menor tamaño de las propiedades en vecindarios más caros. Cercanos a centros urbanos.

```
boxplot(inmuebles$SalePrice~inmuebles$Neighborhood,col="blue",
        xlab="Neighborhood",ylab="Sale Prices", main="Sale Prices per
Neighborhood")
```



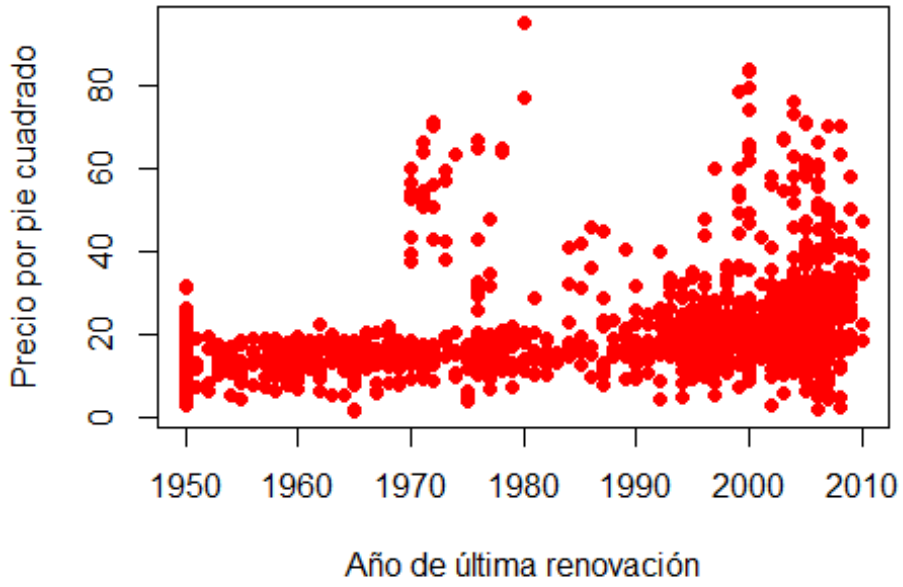

Comprobando también los precios por pie cuadrado respecto al año de construcción:

```
plot(inmuebles$YearBuilt, inmuebles$PriceLotSquareFoot,pch=19,col="red", xlab
= "Año de Construcción", ylab = "Precio por pie cuadrado")
```



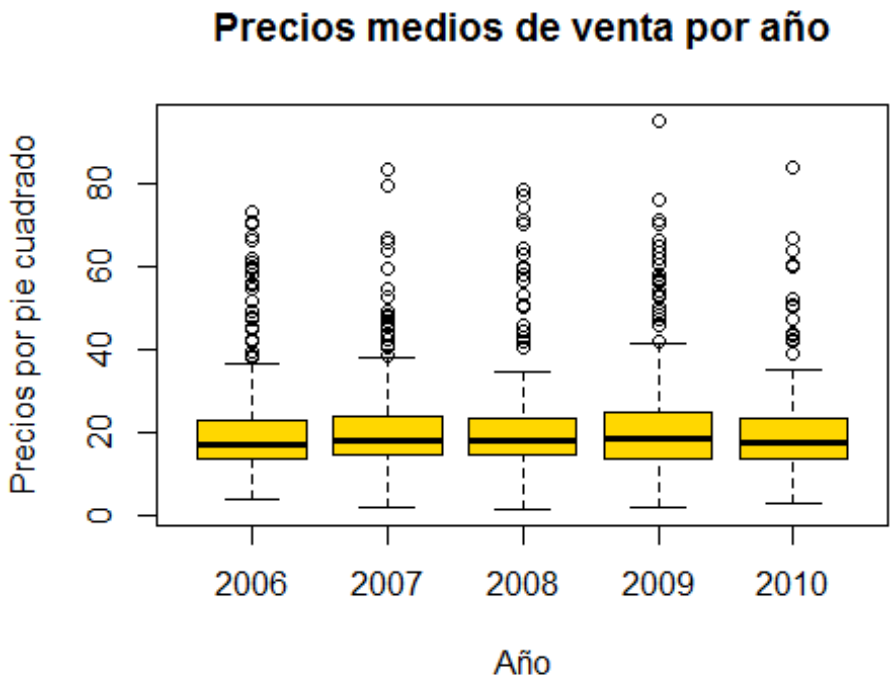
Los precios más altos se decantan, con excepciones, en las propiedades más recientes, esto nos indica cómo la antigüedad es importante en la determinación de precios, cabría analizar si irá a la par del año de renovación de la vivienda.

```
plot(inmuebles$YearRemodAdd, inmuebles$PriceLotSquareFoot,pch=19,col="red",
xlab = "Año de última renovación", ylab = "Precio por pie cuadrado")
```




Para completarlo, podemos analizar los rangos de precios según los años de venta:

```
boxplot(inmuebles$PriceLotSquareFoot~inmuebles$YrSold,col="gold",xlab="Año",y
lab="Precios por pie cuadrado",
        main="Precios medios de venta por año")
```




## 6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

El análisis de los datos de venta de viviendas nos ha permitido identificar los factores más influyentes en el precio, así como analizar por zonas y por equipamiento de los inmuebles.

Los distintos modelos probados nos han dado una aproximación a la tasación de viviendas, siendo útiles tanto para vendedores, que conocen cuáles son los precios medios demandados según la ubicación y características de la vivienda, como para compradores, para optimizar sus búsquedas según deseos y poder adquisitivo.

Los datos han permitido implementar unos modelos lineales básicos para ello, además de permitirnos visualizar la evolución de precios.

Se confirman hipótesis iniciales del mundo inmobiliario, como es la importancia determinante de la localización, vencindarios y proximidad a infraestructuras, por encima de otras consideraciones como materiales de construcción y, en algunos casos, hasta el estado de la vivienda.

## 7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos.

La práctica ha sido implementada en código R, empleado la herramienta RStudio, formateando los resultados finales con *rmarkdown*.

El codigo, fichero rmd, Practica2_topología_jperezsanchez.Rmd, se encuentra en el repositorio github de la practica. Con él se ha generado el documento pdf principal entregable de la práctica.

Se han empleado las librerías R: RCurl, VIM, nortest, magrittr y dplyr.

Algunos scripts más largos se han mutado en el documento generado con la opción ECHO = FALSE, pero es consultable en el fichero Rmd.

El script al ejecutarse descarga los datos directamente del fichero csv de github, para cambiar a lectura local hay que descomentar la línea:

– inmuebles <- read.csv("Datos-Inmuebles.csv")

y comentar las líneas:

downF <- getURL("https://raw.githubusercontent.com/jperezsanchezU/house-prices-advanced-regression-techniques/master/csv/Datos-Inmuebles.csv") inmuebles <- read.csv(text = downF)

## Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:


- Megan Squire (2015). Clean Data. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369 .
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science &Business Media.
- Wes McKinney (2012). Python for Data Analysis. Oâ□□Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world.
- Dataset de Precios de Casas en *kaggle* House Prices: Advanced Regression Techniques