California Housing Prices

Diego Alejandro Mernies

Primer cuatrimestre 2019

Introducción

Realizaremos una análisis de los datos de las casas que se encuentran en un distrito determinado de **California** y algunas estadísticas basadas en el censo de 1990.

Objetivo

Predecir el precio de las casa de la época con un modelo de regresión lineal.

En primer lugar cargamos las librerías requeridas. Si no las tiene en su sistema, puede instalarlas con `install.packages("librería")`.

```
library(readr)
library(dplyr)
library(corrplot)
library(ggplot2)
library(scales)
#library(rms)
```

Definición de contantes

A continuación se define las **constantes** que se utilizarán en el proyecto.

```
# URL donde reside el dataset a utilizar
dataurl <- "https://raw.githubusercontent.com/dmerniestic1987/tp_ciencia_datos_california_housing/maste
# Ubicación local en donde se guardará el dataset para su procesamiento
datadir <- "~/workspace/R/data"
```

Carga de datos

Set de datos

El set de datos es un archivo .csv (comma separated value) de exactamente 10 columnas y 20641 filas de las cuales la primera contiene los nombres. El archivo de input original se llama **housing.csv** y se tomó de [California Housing Price] (https://www.kaggle.com/camnugent/california-housing-prices), pero fue subido a un repositorio GIT para simplificar la descarga de los datos y controlar las versiones. El repositorio GIT se puede explorar ingresando a: https://github.com/dmerniestic1987/tp_ciencia_datos_california_housing.

Los datos pertenecen a las casas que se encuentran en un distrito de California y algunas estadísticas basadas en los datos del censo de 1990. Las variables son:

Variable	Descripción
longitude	Qué tan lejos al oeste este está una casa. Un valor más alto está más al oeste.
latitude	Qué tan lejos al norte está una casa. Un valor más alto está más al norte.
housing_median_age	Edad media de una casa dentro de un bloque de casas. Un número más bajo es un edificio más nuevo.
total_rooms	Número total de ambientes dentro de un bloque de casas.
total_bedrooms	Número total de habitaciones dentro de un bloque de casas.
population	Número total de personas que residen dentro de un bloque de casas.
households	Número total de hogares, un grupo de personas que residen dentro de una unidad de hogar, por un bloque.
median_income	Ingreso promedio para hogares dentro de un bloque de casas (medido en decenas de miles de dólares estadounidenses)
median_house_value	Valor medio de la vivienda para hogares dentro de un bloque (medido en dólares estadounidenses)

Variable	Descripción
ocean_proximity	Ubicación de la casa con relación al oceano o mar.

Descarga del set de datos

Descargamos los datos.

```
datafile <- paste(datadir, "housing.csv", sep = "/")
```

En primer lugar se verifica si es necesario crear un directorio para almacenar el archivo.

```
if (dir.exists(datadir)) {
  print(paste("El directorio ", datadir, " ya existe."))
} else {
  print(paste("Creando directorio de datos", datadir, "."))
  dir.create(datadir)
}
```

```
## [1] "El directorio  ~/workspace/R/data  ya existe."
```

El segundo lugar se descarga la última versión del archivo para poder utilizar la información actualizada.

```
if (file.exists(datafile)) {
  print(paste("El archivo ", datafile, " ya existe, lo elimino."))
  file.remove(datafile)
}
```

```
## [1] "El archivo  ~/workspace/R/data/housing.csv  ya existe, lo elimino."
```

```
## [1] TRUE
```

```
download.file(dataurl, datafile, method="auto")
```

Lectura de los datos

Leemos el archivo recientemente descargado convirtiendo los espacios vacíos en N/A. Las columnas sin información no necesitan ser eliminadas dado que el resto de la información del futbolista puede ser útil.

```
  dfhousing <- read.csv(datafile)
```

Control de datos

Verifcamos que nos nombres de las columnas sean correctos

```
colnames(dfhousing)
```

```
##  [1] "longitude"          "latitude"           "housing_median_age"
##  [4] "total_rooms"        "total_bedrooms"     "population"
##  [7] "households"         "median_income"      "median_house_value"
## [10] "ocean_proximity"
```

Verificamos los tipos de datos de las columnas

```r
dim(dfhousing)
```

```
## [1] 20640    10
```

```r
str(dfhousing)
```

```
## 'data.frame':    20640 obs. of  10 variables:
##  $ longitude         : num  -122 -122 -122 -122 -122 ...
##  $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
##  $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
##  $ total_rooms       : num  880 7099 1467 1274 1627 ...
##  $ total_bedrooms    : num  129 1106 190 235 280 ...
##  $ population        : num  322 2401 496 558 565 ...
##  $ households        : num  126 1138 177 219 259 ...
##  $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ median_house_value: num  452600 358500 352100 341300 342200 ...
##  $ ocean_proximity   : Factor w/ 5 levels "<1H OCEAN","INLAND",..: 4 4 4 4 4 4 4 4 4 4 ...
```

Verificamos los primeros registros del archivio para verificar los formatos

```r
head(dfhousing, give.attr=FALSE, 10)
```

```
##    longitude latitude housing_median_age total_rooms total_bedrooms
## 1    -122.23    37.88                41         880            129
## 2    -122.22    37.86                21        7099           1106
## 3    -122.24    37.85                52        1467            190
## 4    -122.25    37.85                52        1274            235
## 5    -122.25    37.85                52        1627            280
## 6    -122.25    37.85                52         919            213
## 7    -122.25    37.84                52        2535            489
## 8    -122.25    37.84                52        3104            687
## 9    -122.26    37.84                42        2555            665
## 10   -122.25    37.84                52        3549            707
##    population households median_income median_house_value ocean_proximity
## 1         322        126        8.3252             452600        NEAR BAY
## 2        2401       1138        8.3014             358500        NEAR BAY
## 3         496        177        7.2574             352100        NEAR BAY
## 4         558        219        5.6431             341300        NEAR BAY
## 5         565        259        3.8462             342200        NEAR BAY
## 6         413        193        4.0368             269700        NEAR BAY
## 7        1094        514        3.6591             299200        NEAR BAY
## 8        1157        647        3.1200             241400        NEAR BAY
## 9        1206        595        2.0804             226700        NEAR BAY
## 10       1551        714        3.6912             261100        NEAR BAY
```

Verificamos los últimos registros del archivio para verificar los formatos

```r
tail(dfhousing, give.attr=FALSE, 10)
```

```
##       longitude latitude housing_median_age total_rooms total_bedrooms
## 20631   -121.32    39.29                11        2640            505
## 20632   -121.40    39.33                15        2655            493
## 20633   -121.45    39.26                15        2319            416
## 20634   -121.53    39.19                27        2080            412
## 20635   -121.56    39.27                28        2332            395
## 20636   -121.09    39.48                25        1665            374
## 20637   -121.21    39.49                18         697            150
```

```
## 20638   -121.22    39.43                17        2254            485
## 20639   -121.32    39.43                18        1860            409
## 20640   -121.24    39.37                16        2785            616
##       population households median_income median_house_value
## 20631       1257        445        3.5673              112000
## 20632       1200        432        3.5179              107200
## 20633       1047        385        3.1250              115600
## 20634       1082        382        2.5495               98300
## 20635       1041        344        3.7125              116800
## 20636        845        330        1.5603               78100
## 20637        356        114        2.5568               77100
## 20638       1007        433        1.7000               92300
## 20639        741        349        1.8672               84700
## 20640       1387        530        2.3886               89400
##       ocean_proximity
## 20631          INLAND
## 20632          INLAND
## 20633          INLAND
## 20634          INLAND
## 20635          INLAND
## 20636          INLAND
## 20637          INLAND
## 20638          INLAND
## 20639          INLAND
## 20640          INLAND
```

Obtenemos un resumen de las variables para verificar los datos.

```
summary(dfhousing)
```

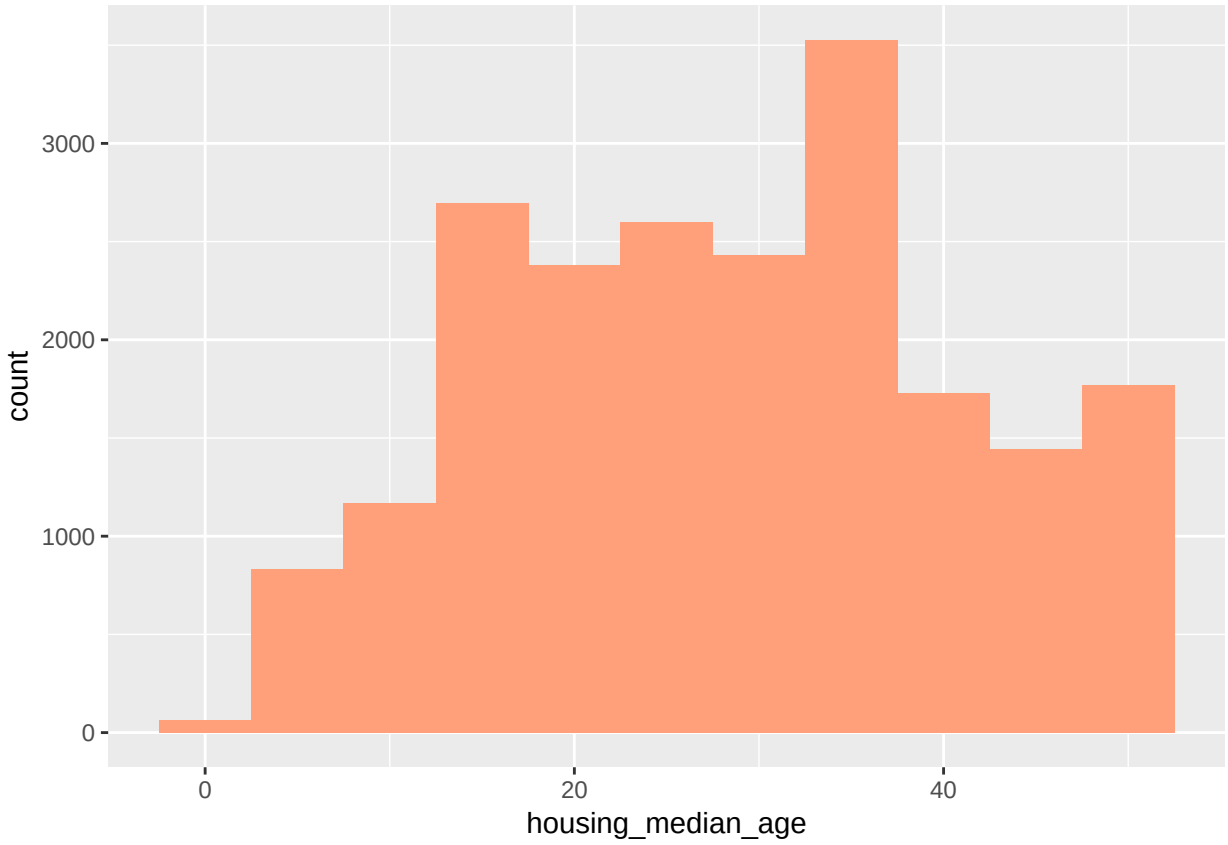
```
##    longitude         latitude     housing_median_age  total_rooms
##  Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2
##  1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448
##  Median :-118.5   Median :34.26   Median :29.00      Median : 2127
##  Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636
##  3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148
##  Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320
##
##  total_bedrooms     population      households     median_income
##  Min.   :   1.0   Min.   :    3   Min.   :   1.0   Min.   : 0.4999
##  1st Qu.: 296.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634
##  Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348
##  Mean   : 537.9   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707
##  3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432
##  Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001
##  NA's   :207
##  median_house_value   ocean_proximity
##  Min.   : 14999     <1H OCEAN :9136
##  1st Qu.:119600     INLAND    :6551
##  Median :179700     ISLAND    :   5
##  Mean   :206856     NEAR BAY  :2290
##  3rd Qu.:264725     NEAR OCEAN:2658
##  Max.   :500001
##
```

El resumen estadísticos se detectó: 1. Es necsario limpiar los NA'S de la columna total_bedrooms. 2. Existen

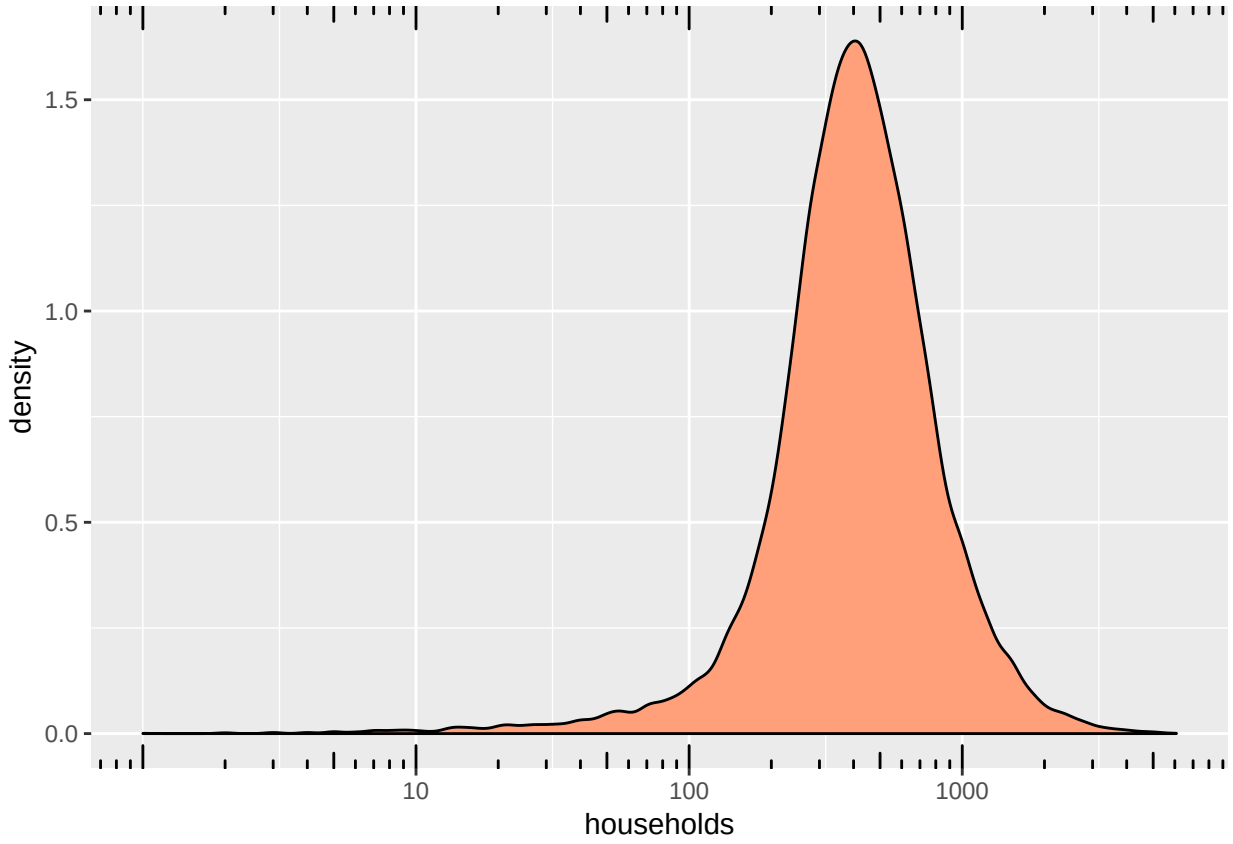
sólo 5 casas que están en una Isla. 3. Los valores máximos de total_rooms, total_bedrooms, population y households son muy altos en comparación a la media.

Se realizan algunos gráficos para observar la distribució de los datos.

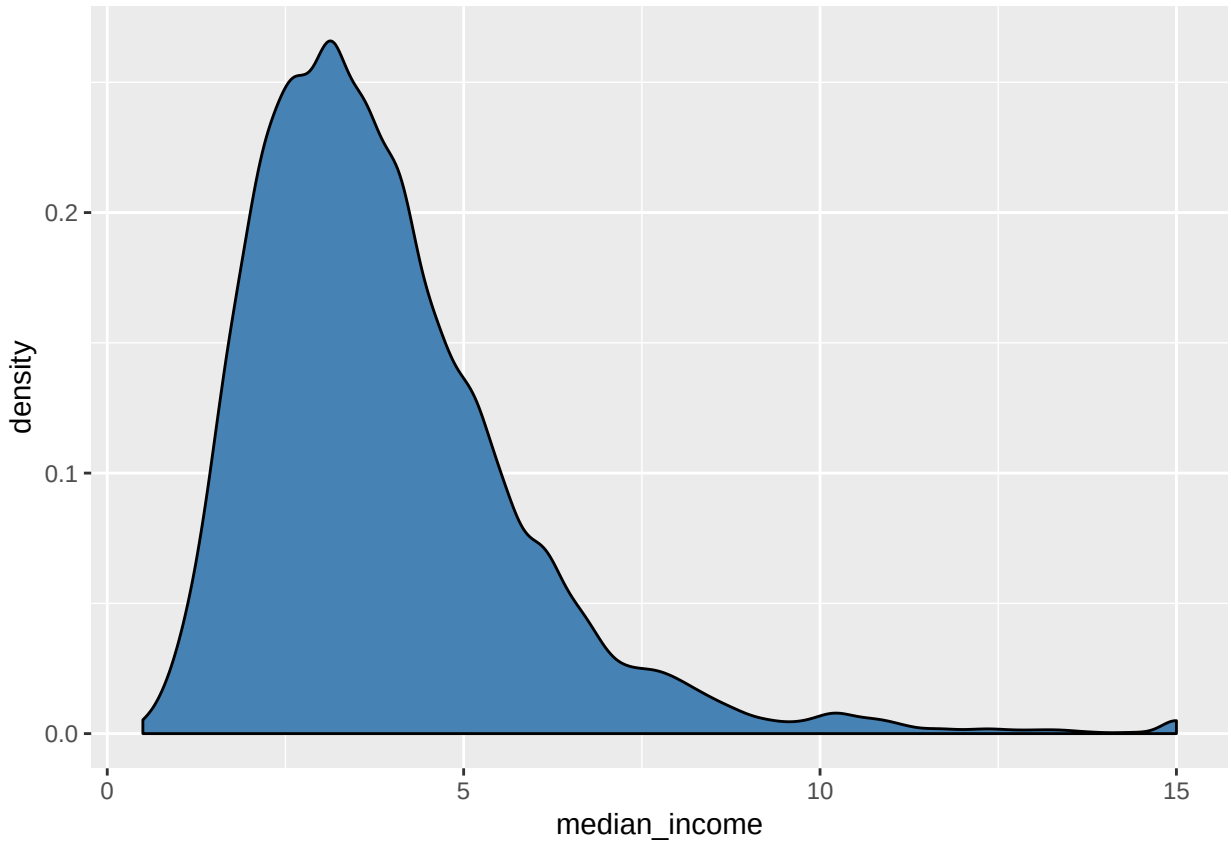
```
ggplot(dfhousing) +
  geom_histogram(aes(x=housing_median_age), binwidth=5, fill="lightsalmon", bins=30)
```



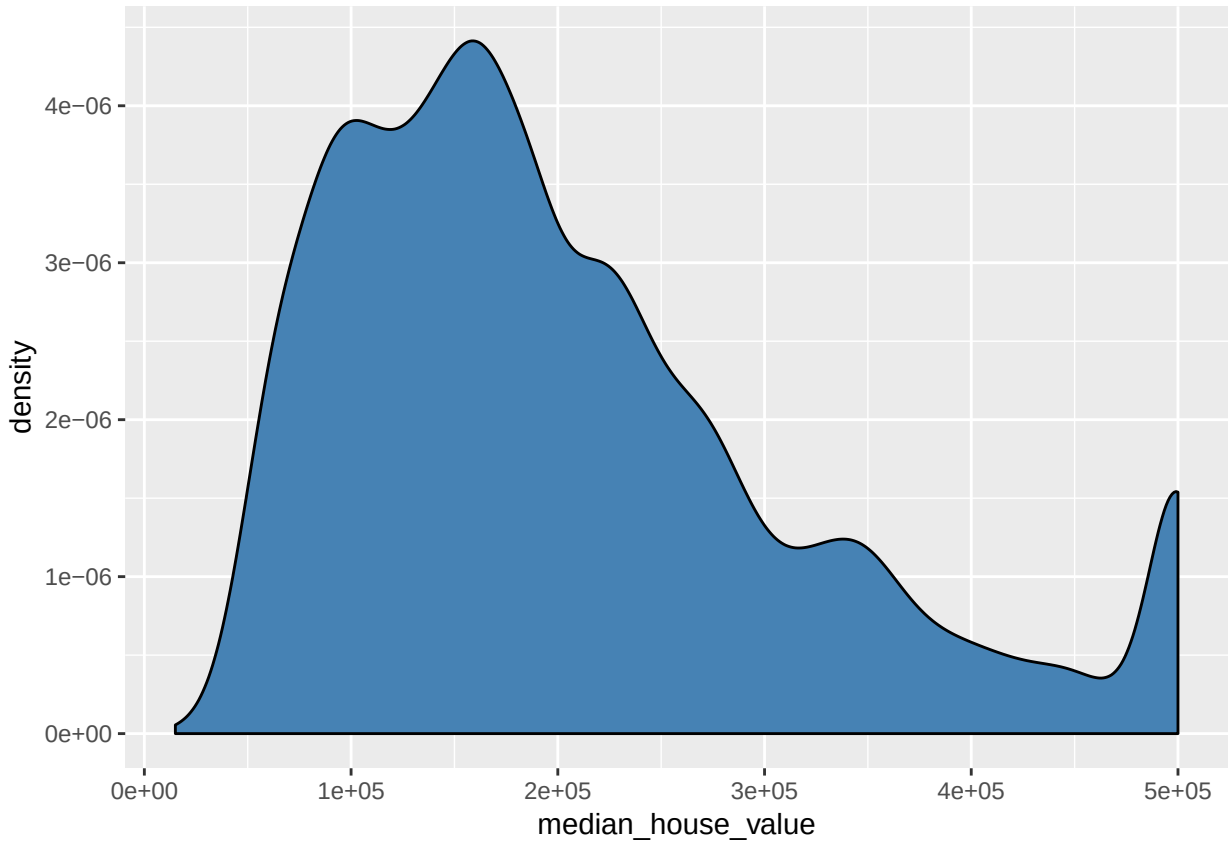

```
ggplot(dfhousing) +
  geom_density(aes(x=households), fill="lightsalmon") +
  scale_x_log10(breaks=c(10, 100, 1000, 100000)) +
  annotation_logticks(sides="bt")
```



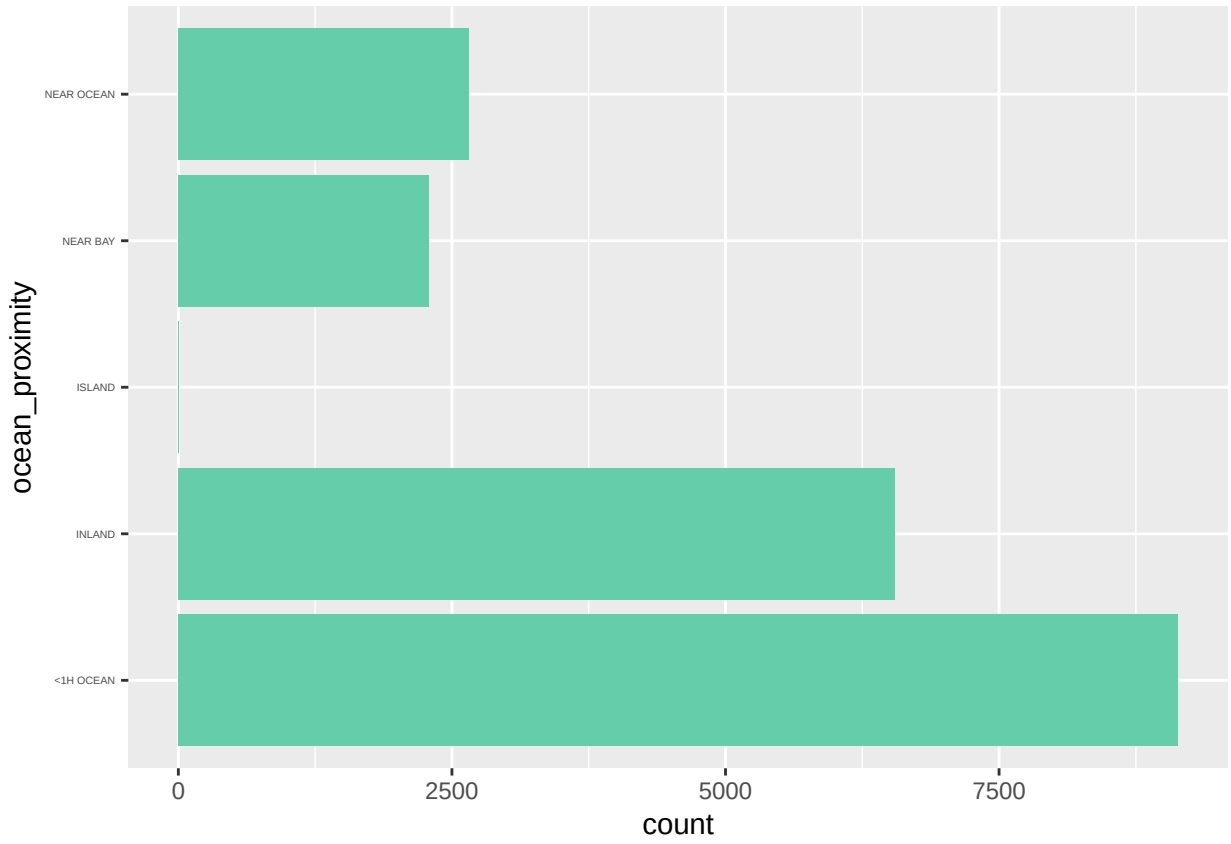

```
ggplot(dfhousing) +
  geom_density(aes(x=median_income), fill="steel blue")
```




```
ggplot(dfhousing) +
  geom_density(aes(x=median_house_value), fill="steel blue")
```




```
ggplot(dfhousing) +
  geom_bar(aes(x=ocean_proximity), fill="mediumaquamarine")+
  coord_flip() +
  theme(axis.text.y=element_text(size=rel(0.4)))
```



##Limpieza de datos Se eliminan los NA de **total_bedrooms**

```
dfhousing$total_bedrooms[is.na(dfhousing$total_bedrooms)]= median(dfhousing$total_bedrooms, na.rm=TRUE)
```

Para corregir los altos valores máximos se crean dos nuevas columnas: -mean_bedrooms: El cuociente entre habitaciones por hogares. -mean_rooms: El cuociente entre ambientes por hogares. Posteriormente eliminamos las columnas total_bedrroms y total_rooms

```
dfhousing$mean_bedrooms = dfhousing$total_bedrooms / dfhousing$households
dfhousing$mean_rooms = dfhousing$total_rooms / dfhousing$households
```

```
#Eliminamos las columnas total_bedrooms y total_rooms para usar el nuevo índice
drops = c('total_bedrooms', 'total_rooms')
dfhousing = dfhousing[ , !(names(dfhousing) %in% drops) ]
```

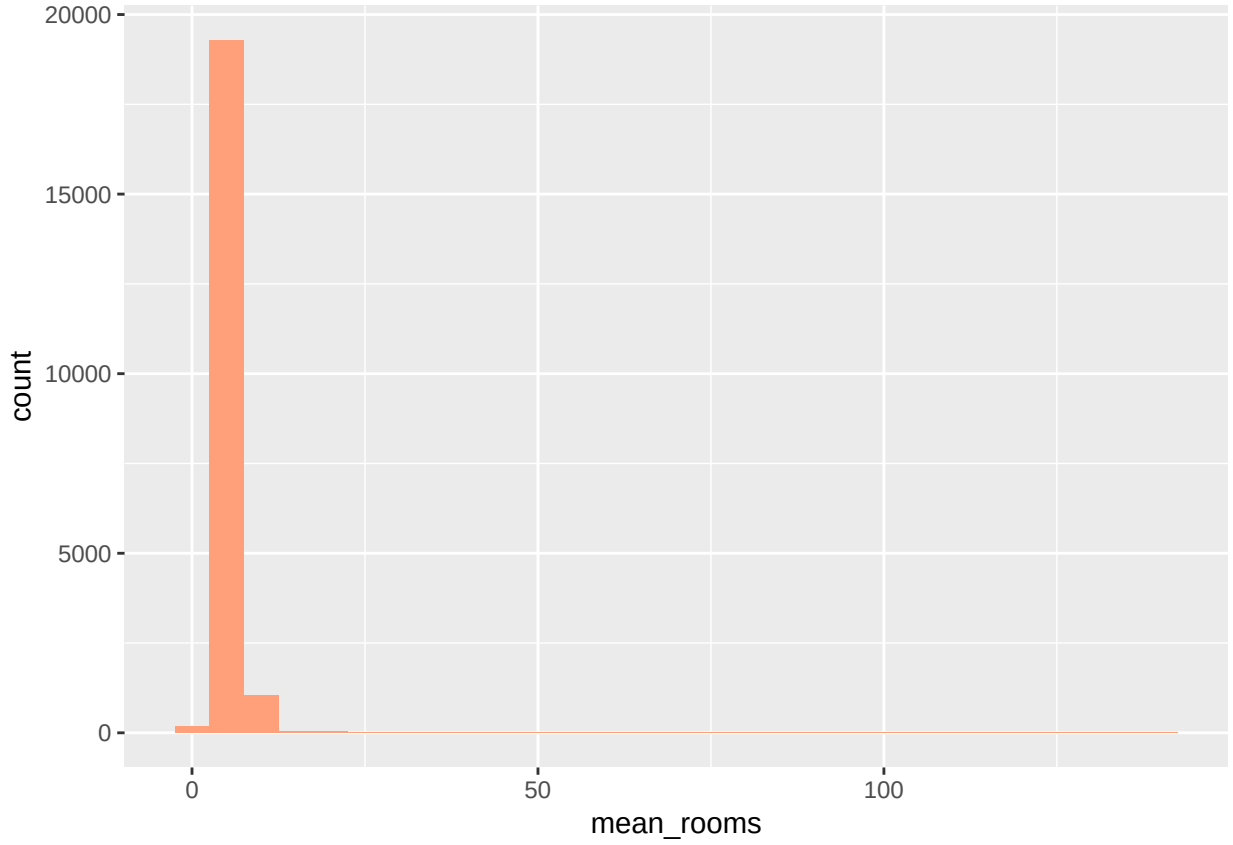
Controlamos la nueva estructura de la tabla

```
head(dfhousing)
```

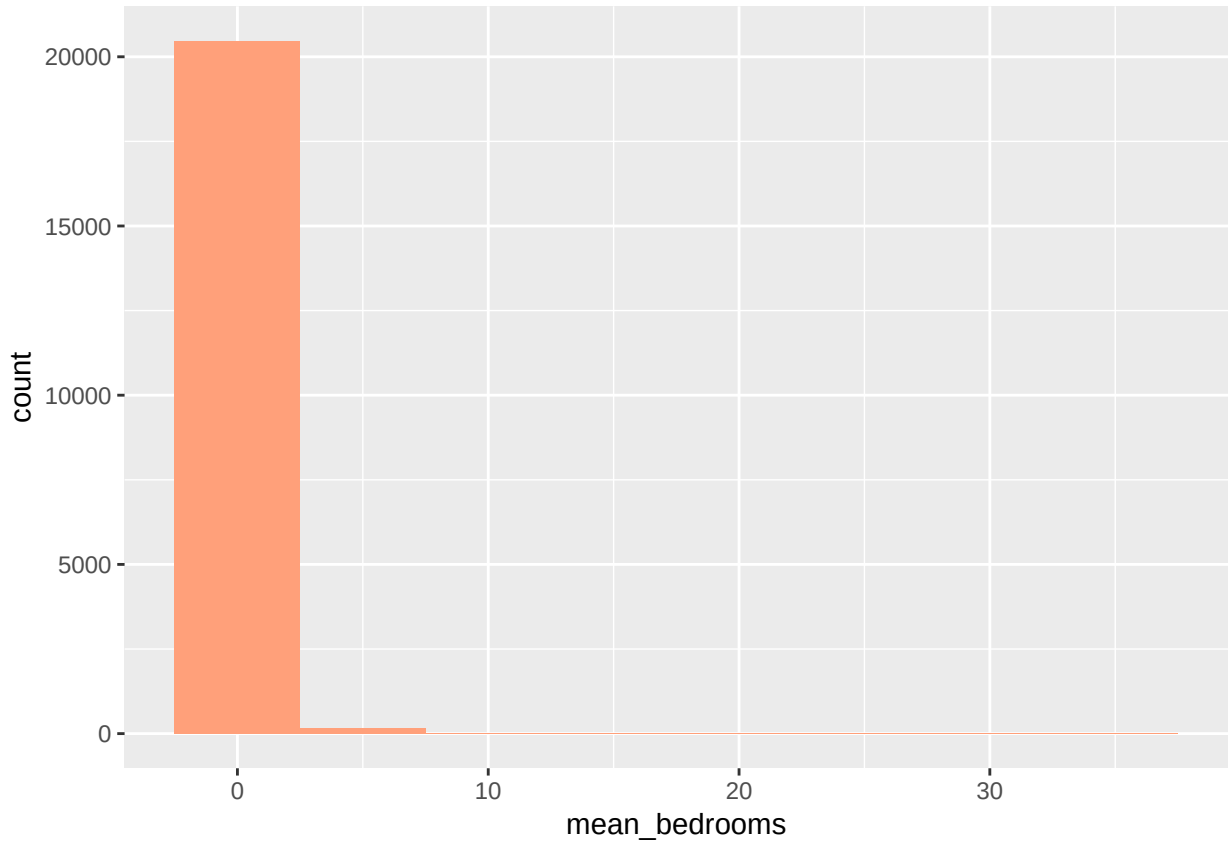
```
##   longitude latitude housing_median_age population households
## 1   -122.23    37.88                41        322        126
## 2   -122.22    37.86                21       2401       1138
## 3   -122.24    37.85                52        496        177
## 4   -122.25    37.85                52        558        219
## 5   -122.25    37.85                52        565        259
## 6   -122.25    37.85                52        413        193
##   median_income median_house_value ocean_proximity mean_bedrooms
## 1        8.3252             452600        NEAR BAY     1.0238095
## 2        8.3014             358500        NEAR BAY     0.9718805
## 3        7.2574             352100        NEAR BAY     1.0734463
```

```
## 4        5.6431             341300        NEAR BAY     1.0730594
## 5        3.8462             342200        NEAR BAY     1.0810811
## 6        4.0368             269700        NEAR BAY     1.1036269
##   mean_rooms
## 1   6.984127
## 2   6.238137
## 3   8.288136
## 4   5.817352
## 5   6.281853
## 6   4.761658
```

```
ggplot(dfhousing) +
  geom_histogram(aes(x=mean_rooms), binwidth=5, fill="lightsalmon", bins=30)
```




```
ggplot(dfhousing) +
  geom_histogram(aes(x=mean_bedrooms), binwidth=5, fill="lightsalmon", bins=30)
```




```
#Creamos un nuevo dataframe auxiliar sin ocean_proximity que es categórico, ni median_house_value
#que el dato que se intentará predecir. Luego se escalan los valores para trabajarlos con gráficos más
drops = c('ocean_proximity','median_house_value')
dfhousing_aux =  dfhousing[ , !(names(dfhousing) %in% drops)]
dfscaledhousing_aux = scale(dfhousing_aux)
```

```
#Creamos un nuevo dataframe que contenga solo la proximidad al mar. Luego limpiamos el resto de las col
dropsCategories = c('ocean_proximity', 'median_house_value')
dfcat_aux =  dfhousing[ , (names(dfhousing) %in% dropsCategories)]
```

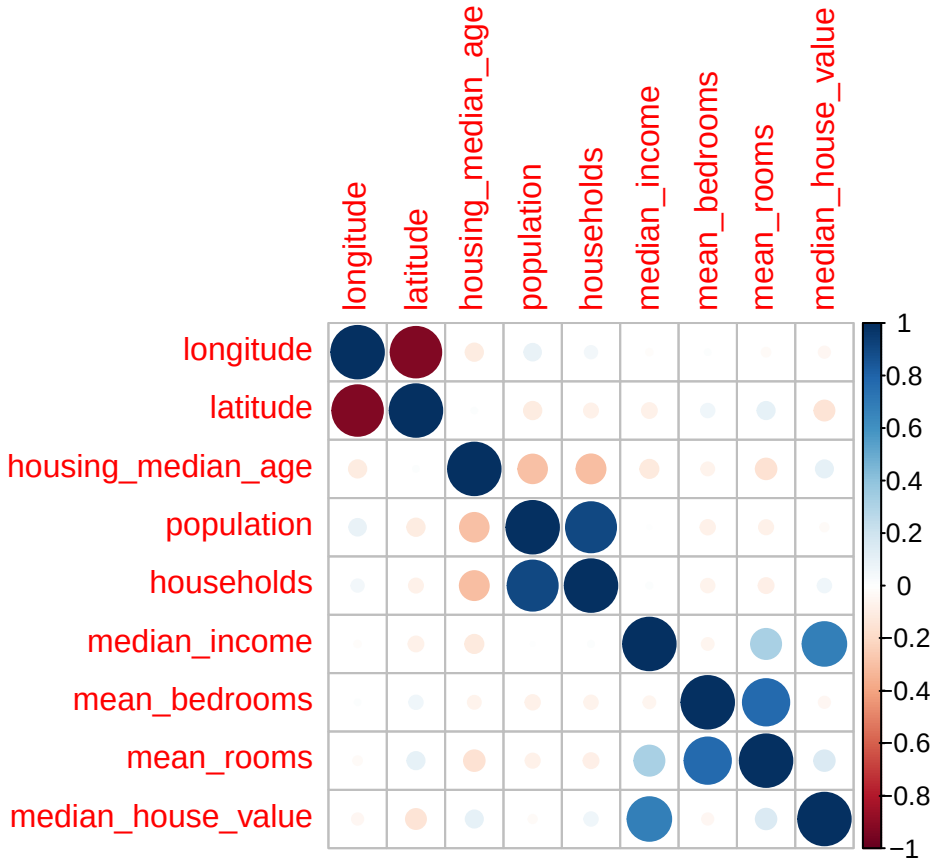
```
#Combinamos las dataframes y generamos un nuevo que contenga la combinación con los datos escalados y l
dfhousing_clean = cbind(DataSet1=dfcat_aux, DataSet2=dfscaledhousing_aux, median_house_value=dfhousing$

dropClean = c('DataSet1.median_house_value')
dfhousing_clean =  dfhousing_clean[ , !(names(dfhousing_clean) %in% dropClean)]

newNames = c('ocean_proximity', 'longitude', 'latitude', 'housing_median_age', 'population', 'household
colnames(dfhousing_clean) <- newNames
```

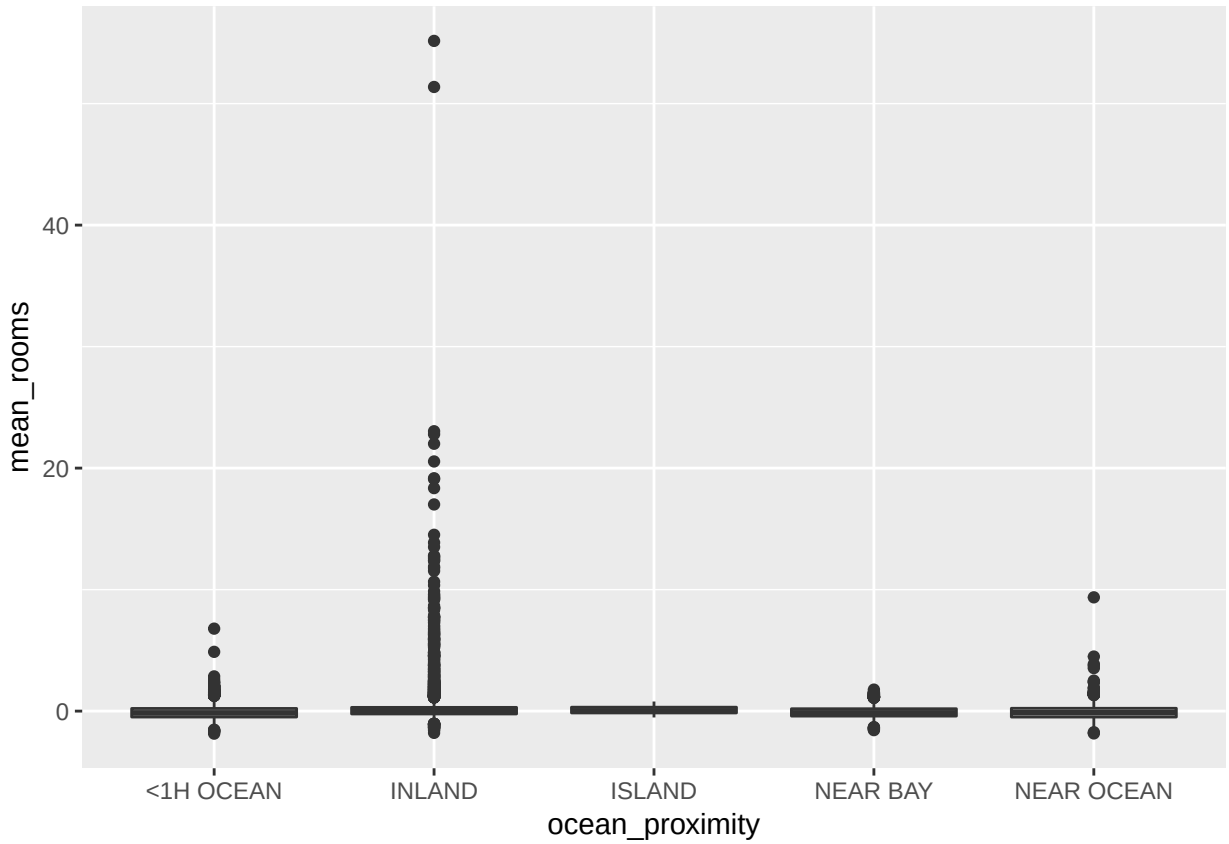
Visualización de correlaciones

```
dfhousing_clean %>%
    select_if(is.numeric) %>%
    cor() %>%
    corrplot()
```

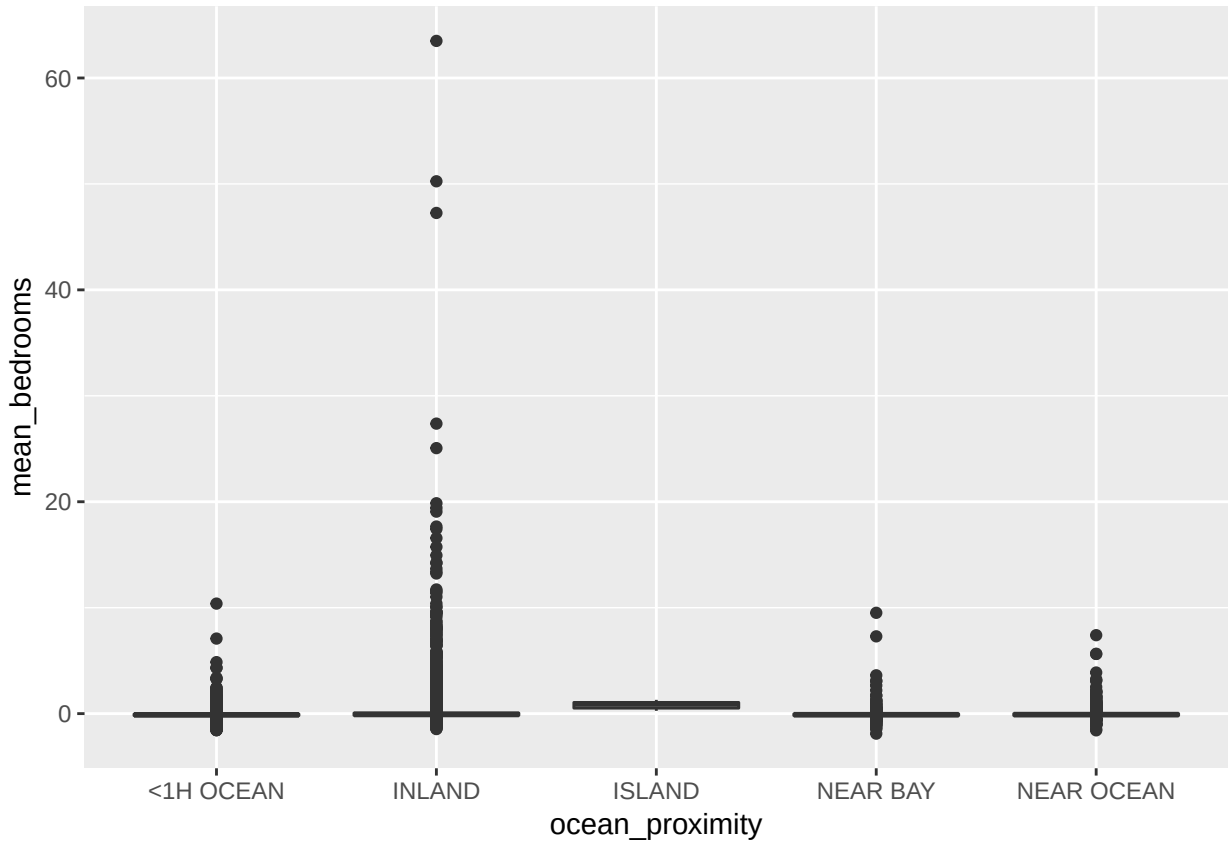


En base al gráfico de correlación se puede determinó que existe una fuerte relación entre la cantidad de la población y la cantidad de hogares. También que el precio de las casas está relacionado con el ingreso medio por lo que pueden haber barrios más caros.

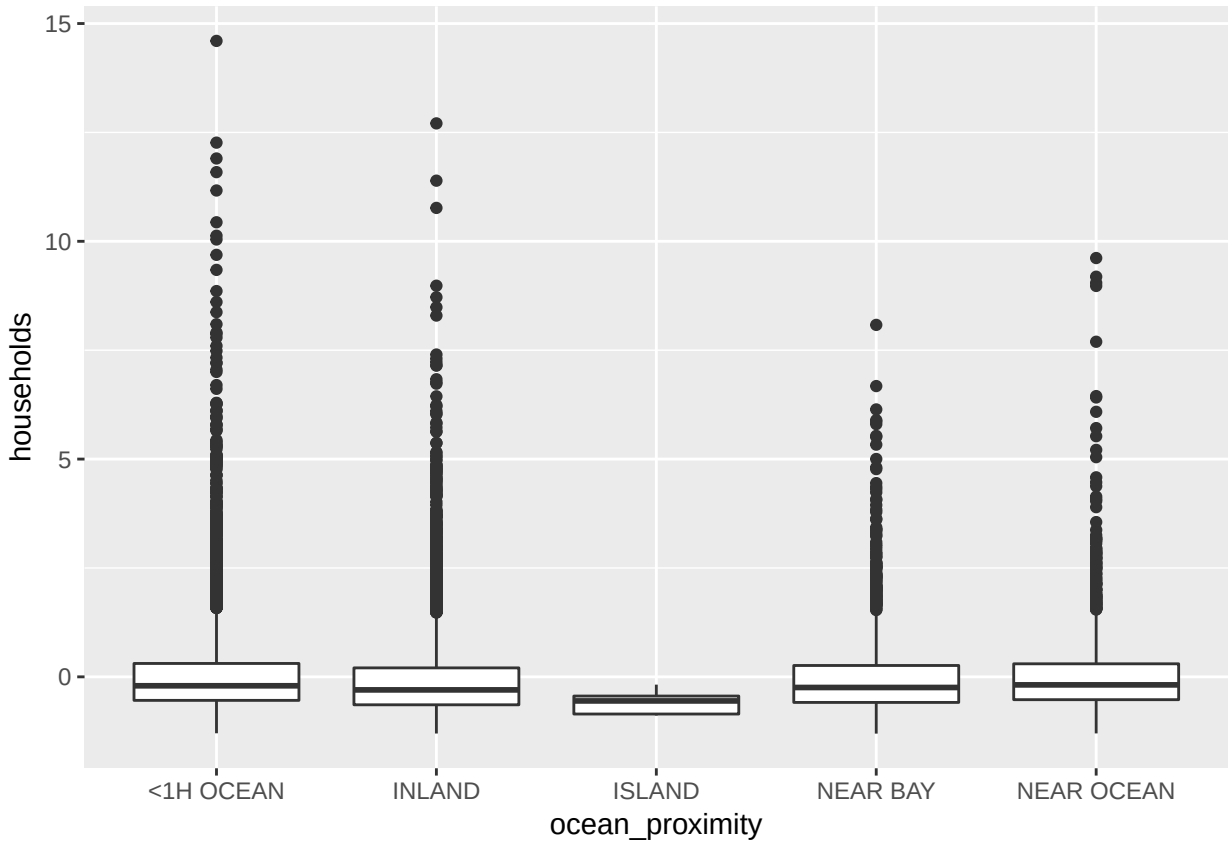
```
#Realizamos un gráfico de boxplot para analizar la relación que hay entre la cantidad de ploblación y l
#Variable categórica relacionada con la proximidad al Oceano.
ggplot(dfhousing_clean) +
    geom_boxplot(aes(x=ocean_proximity, y=mean_rooms))
```




```
ggplot(dfhousing_clean) +
    geom_boxplot(aes(x=ocean_proximity, y=mean_bedrooms))
```




```
ggplot(dfhousing_clean) +
    geom_boxplot(aes(x=ocean_proximity, households))
```


15
10
5
0
households
<1H OCEAN
INLAND
ISLAND
NEAR BAY
NEAR OCEAN
ocean_proximity