

Práctica 2 - Tipología ciclo de vida de los datos

Pau Ortí y Alexander Almendral

08/06/2021

Contents

Descripción del dataset	1
Explicación del dataset y sus variables	1
Carga de los datos	2
Integración y selección de los datos de interés a analizar	3
Limpieza de los datos	4
Valores nulos	4
Valores extremos	5
Análisis de los datos	6
Estudio de la distribución de <code>median_house_value</code>	6
Comparación del precio según <code>ocean_proximity</code>	7
Análisis de componentes principales (PCA)	11
Análisis de regresión	13
Análisis de correlaciones	15
Representación de los resultados y resolución del problema	17

Descripción del dataset

Explicación del dataset y sus variables

El dataset california housing prices nos da información sobre el habitatge de los distintos distritos que se encontraban en el estado de California (Estados Unidos) en 1990, los datos se han extraídos del censo oficial del estado. Cada fila del dataset representa un distrito concreto el cual consta de las siguientes variables:

- **longitude**: dato numérico. Nos da información de la ubicación del distrito.
- **latitude**: dato numérico. Nos da información de la ubicación del distrito.
- **housing_median_age**: dato numérico. Edad mediana de la población de dentro del distrito.
- **total_rooms**: dato numérico. Número total de habitaciones dentro del distrito.

- **total_bedrooms**: dato numérico. Número total de dormitorios dentro del distrito.
- **population**: dato numérico. Número total de individuos residiendo dentro del distrito.
- **households**: dato numérico. Número total de hogares dentro del distrito.
- **median_income**: dato numérico. Mediana de los ingresos por hogar dentro del distrito.
- **median_house_value**: dato numérico. Mediana del valor del habitaje dentro del distrito.
- **ocean_proximity**: dato categórico. Proximidad del distrito respecto al océano, con cuatro posibles valores: "NEAR BAY", "<1H OCEAN", "INLAND", "ISLAND" y "NEW OCEAN".

##Qué pregunta se pretende responder

Se pretende entender cuáles son las variables que más influyen en el precio del habitaje por distritos, y estudiar qué relación tienen estas respecto al precio final de las casas en la costa oeste de los estados unidos. Por lo tanto, escogeremos la variable **median_house_value**, que nos da información sobre el valor medio del habitatge por distrito, como variable explicada, y el resto como variables explicativas.

Carga de los datos

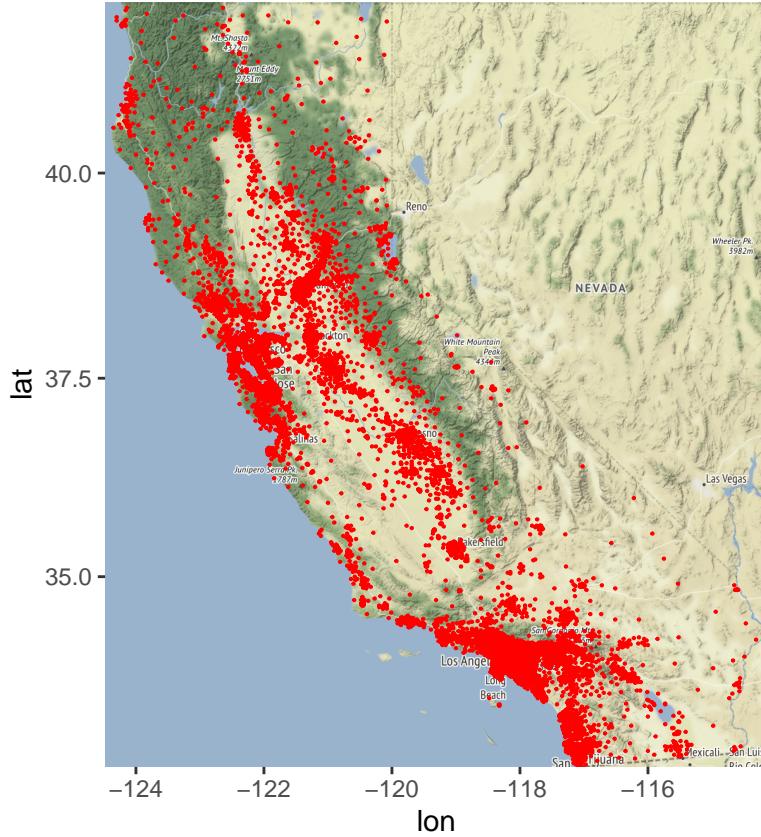
Empezamos inicializando dos variables donde guardaremos, en una la versión original de los datos (data_original), tal y como los hemos importado del csv, y en la otra (data) guardaremos los datos que iremos modificando a lo largo de la práctica.

```
# Cargamos los datos
data_original <- read.csv('Data/housing.csv')
data <- read.csv('Data/housing.csv')
head(data)

##   longitude latitude housing_median_age total_rooms total_bedrooms population
## 1    -122.23     37.88              41        880          129         322
## 2    -122.22     37.86              21       7099          1106        2401
## 3    -122.24     37.85              52       1467           190         496
## 4    -122.25     37.85              52       1274           235         558
## 5    -122.25     37.85              52       1627           280         565
## 6    -122.25     37.85              52       919            213         413
##   households median_income median_house_value ocean_proximity
## 1         126      8.3252          452600    NEAR BAY
## 2        1138      8.3014          358500    NEAR BAY
## 3         177      7.2574          352100    NEAR BAY
## 4         219      5.6431          341300    NEAR BAY
## 5         259      3.8462          342200    NEAR BAY
## 6         193      4.0368          269700    NEAR BAY

# Obtiene el mapa de California
map <- get_map(getbb("California"), maptype = "toner-background")
cali_map <- ggmap(map)
geo_data <- data[,c("longitude","latitude")]

# Muestra el mapa de California con las localizaciones de las viviendas
cali_map +
  geom_point(data = geo_data, mapping = aes(x = longitude, y = latitude),
             color = "red", size=0.1)
```



Integración y selección de los datos de interés a analizar

```
dim(data)
```

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

```
str(data)
```

```
## '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  : chr "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
```

Observamos que todas las variables del conjunto de datos corresponden a variables cuantitativas discretas excepto la variable `ocean_proximity`, la cual es cualitativa de tipo factor.

Limpieza de los datos

Valores nulos

Empezamos a analizar nuestra única variable categórica, la variable `ocean_proximity`.

```
# Tabla de frecuencias de ocean_proximity
table(data$ocean_proximity)
```

```
##
## <1H OCEAN      INLAND      ISLAND    NEAR BAY NEAR OCEAN
##      9136        6551         5       2290       2658
```

Vemos como esta está compuesta por 4 categorías diferentes. El resultado de la tabla de frecuencias no muestra ninguna categoría “vacía” que contenga valores, por lo que no tenemos valores nulos en esta variable. Finalmente la convertimos a tipo factor.

```
# Convertimos a factor
data$ocean_proximity <- factor(data$ocean_proximity)
```

A continuación comprobamos también la existencia de valores nulos en las otras variables:

```
# Comprobamos la existencia de datos perdidos
sapply(data, function(x) sum(is.na(x)))
```

```
##           longitude          latitude housing_median_age      total_rooms
##                 0                  0                  0                  0
##      total_bedrooms      population     households median_income
##                 207                  0                  0                  0
## median_house_value ocean_proximity
##                 0                  0
```

Observamos que sólamente hay 207 datos nulos, todos pertenecen a la variable `total_bedrooms`. La cantidad de datos nulos es mínima teniendo en cuenta que conjunto de datos contiene más de 20.000 registros, sin embargo, optamos por imputar los valores perdidos utilizando el algoritmo *knn-means*, para calcular la distancia Euclídea para la computación de los valores perdidos de la variable `total_bedrooms`, utilizaremos aquellas variables que presenten una correlación más alta con los valores de esta variable. Empezamos imprimiendo la correlación de los valores respecto a la variable `total_bedrooms`.

```
data_na_removed <- data[is.na(data$total_bedrooms) == FALSE, 1:9]
sort(cor(data_na_removed) [, 'total_bedrooms'], decreasing = TRUE)
```

```
##      total_bedrooms     households      total_rooms      population
##            1.000000000      0.97972827      0.93037950      0.87774674
##      longitude median_house_value median_income      latitude
##            0.06960802      0.04968618     -0.00772285     -0.06698283
## housing_median_age
##            -0.32045104
```

Observamos que las variables que presentan una correlación mayor son `households`, `total_rooms` and `population`. Por lo tanto, vamos a dejar que nuestro algoritmo **knn-means** calcule la distancia Euclídea con estas tres variables y después impute los valores perdidos de nuestra variable `total_rooms`.

```
data$total_bedrooms <- kNN(data[,c('total_bedrooms','households', 'total_rooms', 'population')])$total_
```

Comprobamos la existencia de datos perdidos

```
sapply(data, function(x) sum(is.na(x)))
```

```
##          longitude            latitude housing_median_age      total_rooms
##          0                  0                  0                  0
##      total_bedrooms      population     households median_income
##          0                  0                  0                  0
## median_house_value ocean_proximity
##          0                  0
```

Valores extremos

Para detectar outliers nos vamos a centrar en tres variables en concreto: median_house_value, median_income, housing_median_age y population.

```
par(mfrow=c(2,2))
hist(scale(data$median_house_value), main="median_house_value")
boxplot(scale(data$median_income), main="median_income")
hist(scale(data$housing_median_age), main="housing_median_age")
plot(scale(data$population), main="population")
```



Observamos que todas las variables representadas en el gráfico tienen valores extremos. Por un lado, la variable `median_house_value` consta de más de 1000 valores por encima de las 2 desviaciones estándares, las otras variables presentan características similares. En este caso decidimos dejar los valores extremos o *outliers* tal y como se nos presentan en el conjunto de datos, ya que corresponden a valores reales y normales de la población estudiada, son comunes y no son fruto de errores.

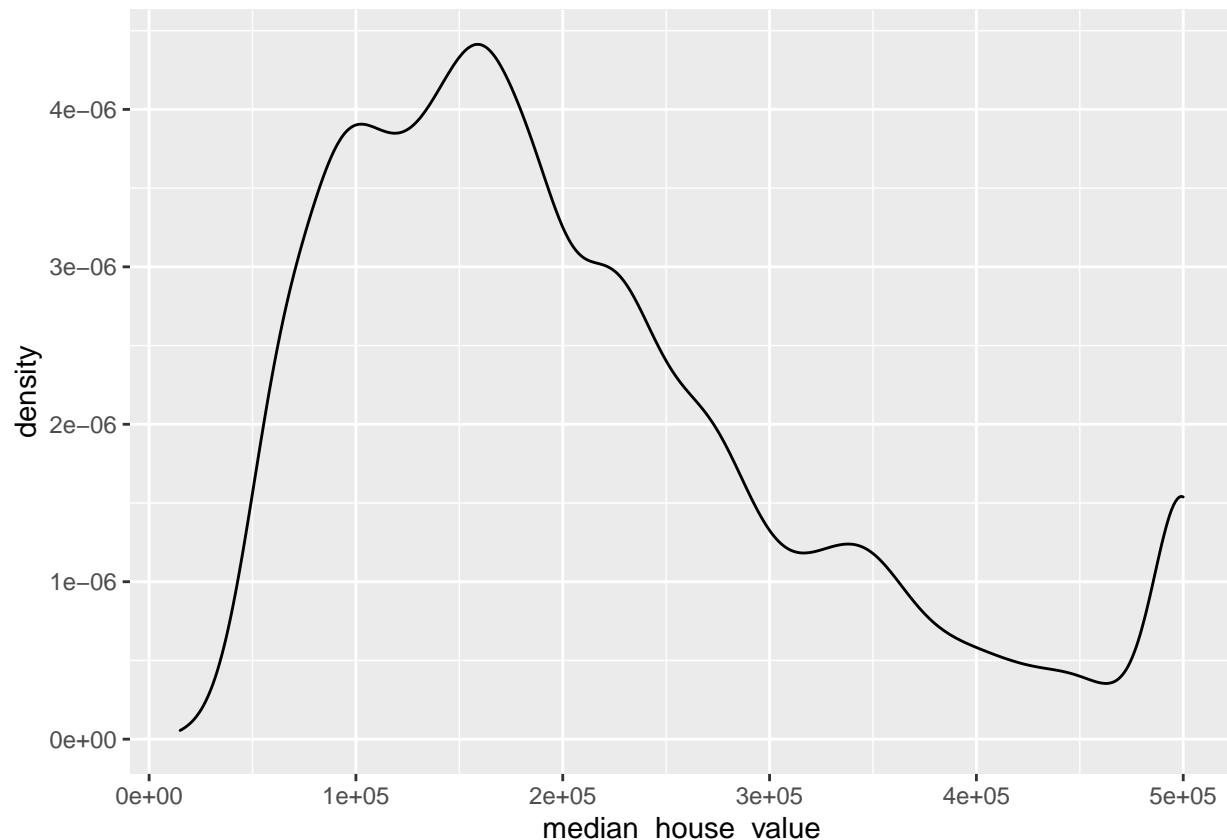
Análisis de los datos

A continuación realizaremos diversos análisis, tales como test de hipótesis, análisis de componentes principales, regresión cuantil y análisis de correlaciones.

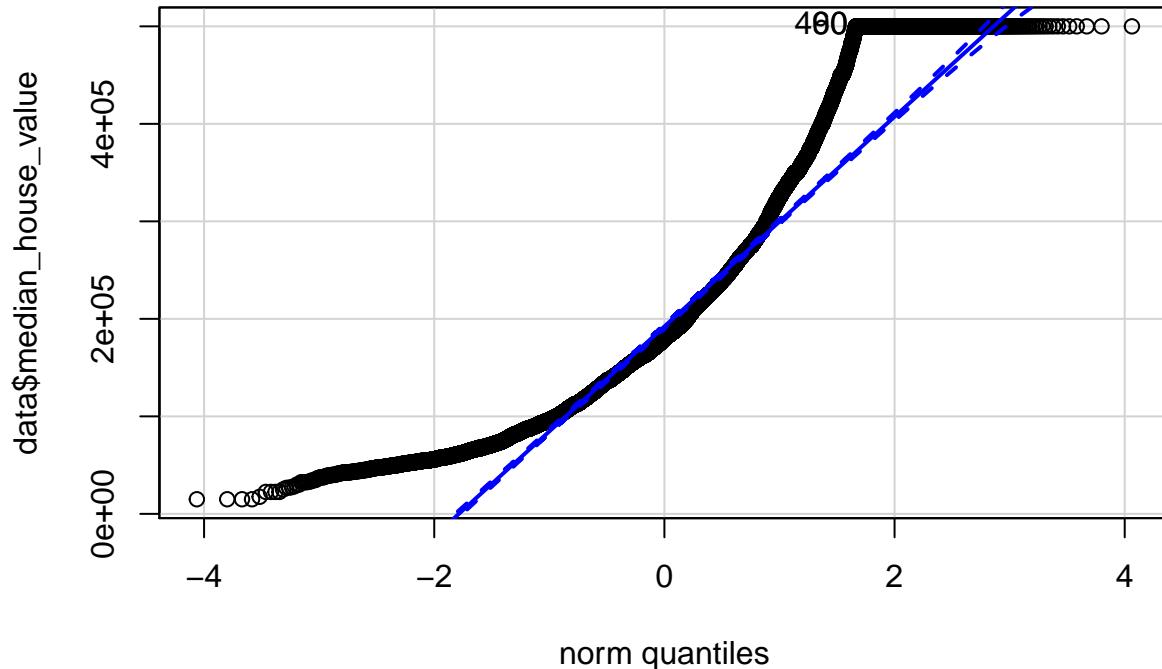
Estudio de la distribución de `median_house_value`

El primer paso en el análisis de los datos es estudiar la distribución de la variable de interés `median_house_value`, sobre la cual se comprobará la normalidad y homogeneidad de la varianza. Para ello visualizaremos la distribución de la variable mediante un gráfico de densidad y un Q-Q plot.

```
library(ggplot2)
ggplot(data, aes(x = median_house_value)) + geom_density()
```



```
qqPlot(data$median_house_value)
```



```
## [1] 90 460
```

Mediante el Q-Q plot ya podemos apreciar que la distribución de la variable `median_house_value` no sigue una distribución normal. Este gráfico nos sugiere que los residuos siguen una distribución normal en cuantiles cercanos a cero, no obstante, para cuantiles en las colas se observa una tendencia muy elevada que nos podría indicar que los datos presentan más valores extremos de los esperados en una distribución normal.

Finalmente, para comprobar la normalidad de los datos mediante un test de hipótesis, utilizaremos el test de normalidad de Lilliefors (Kolmogorov-Smirnov), el cual rechaza firmemente la hipótesis nula de normalidad indicandonos así que la variable `median_house_value` no es normal.

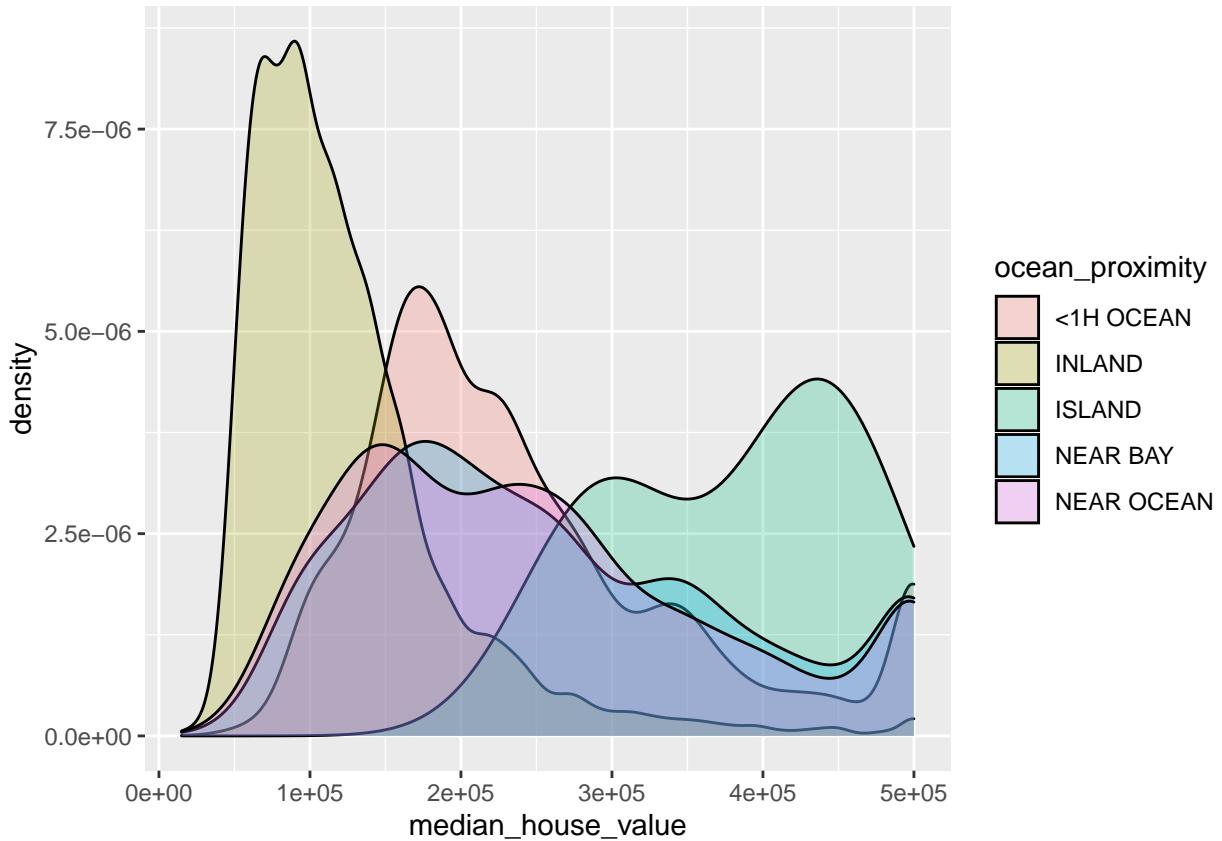
```
lillie.test(data$median_house_value)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: data$median_house_value
## D = 0.10299, p-value < 2.2e-16
```

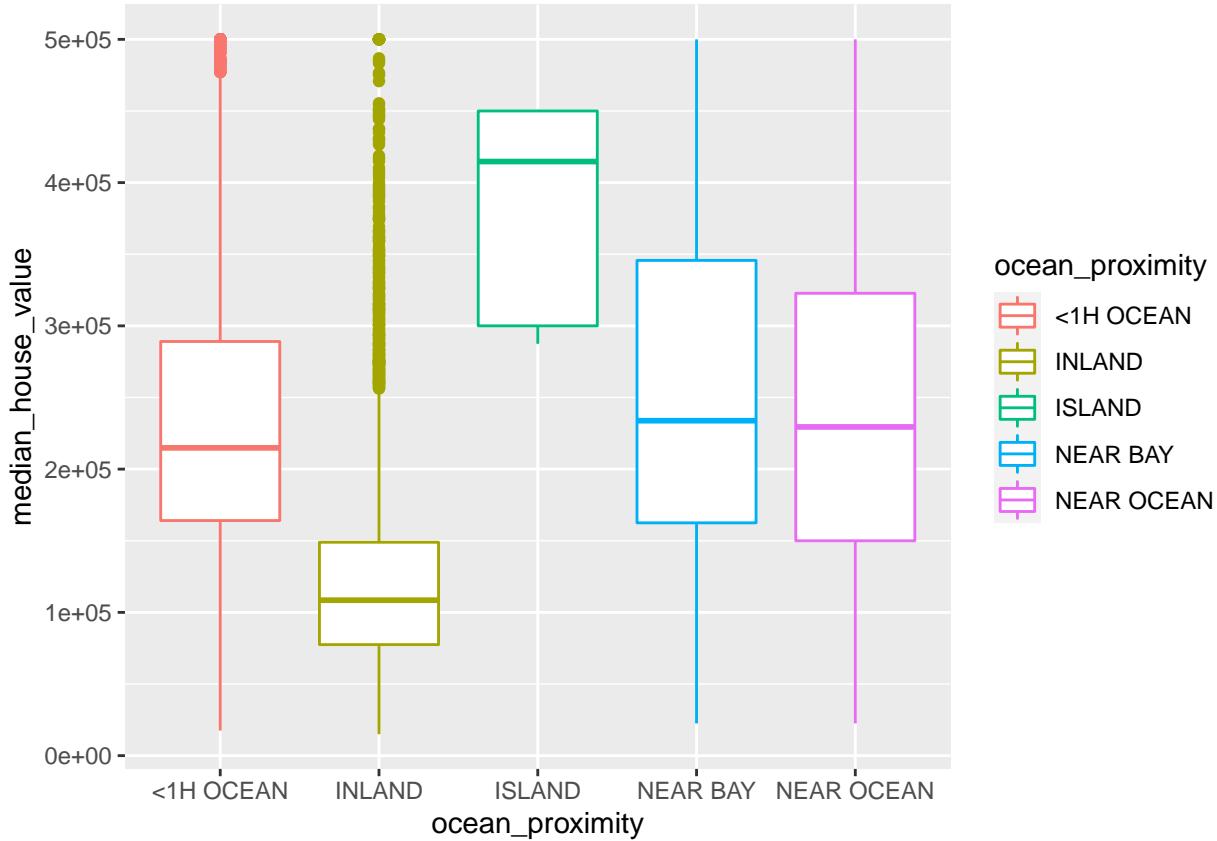
Comparación del precio según `ocean_proximity`

Otra variable que nos interesa estudiar es `ocean_proximity`, que nos servirá para comparar los precios de las viviendas según su proximidad al océano. Para ello se estudia primero el gráfico de densidad y el diagrama de caja `median_house_value` según las diferentes categorías de `ocean_proximity`.

```
ggplot(data, aes(x=median_house_value, fill=ocean_proximity)) + geom_density(alpha=0.25)
```



```
# Boxplot de
ggplot(data, aes(x=ocean_proximity, y=median_house_value, color=ocean_proximity)) +
  geom_boxplot()
```



Observamos en el gráfico de densidad de los distintos grupos dentro de la variable `ocean_proximity` respecto a la variable `median_house_value` que las categorías "<1H OCEAN", "NEAR BAY" y "NEAR OCEAN", no parecen mostrar diferencias significativas respeco a la distribución de precios en estas areas. Del mismo modo, en el diagrama de caja se no se aprecian diferencias entre las categorías ya observadas en el gráfico de densidad.

Ahora debemos decidir que test aplicar sobre los datos ya que, como se ha observado anteriormente, la distribución de `median_house_value` no sigue una distribución normal. Al aplicar un test de normalidad sobre cada una de las categorías observamos que únicamente sigue una distribución normal la categoría "ISLAND". También aplicamos un test de Levene para comprobar la homogeneidad de las variancias según las categorías de `ocean_proximity` que da como resultado una clara heterocedasticidad de las varianzas.

```
# Test de normalidad del precio por categorias
tapply(data$median_house_value,data$ocean_proximity,lillie.test)
```

```
## $`<1H OCEAN`
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: X[[i]]
## D = 0.11196, p-value < 2.2e-16
##
##
## $INLAND
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
```

```

## data: X[[i]]
## D = 0.12139, p-value < 2.2e-16
##
## $ISLAND
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: X[[i]]
## D = 0.26468, p-value = 0.2969
##
## $`NEAR BAY`
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: X[[i]]
## D = 0.086188, p-value < 2.2e-16
##
## $`NEAR OCEAN`
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: X[[i]]
## D = 0.07679, p-value < 2.2e-16

# Levene's test with multiple independent variables
leveneTest(median_house_value ~ ocean_proximity, data = data)

```

```

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group      4 439.23 < 2.2e-16 ***
##        20635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Posteriormente vamos a corroborar esta primera impresión con el test de suma de rangos de Kruskal-Wallis, el cual es un test no paramétrico y donde, sobre la hipótesis nula, se prueba si las diferentes categorías son iguales o pertenecen a la misma población.

Posteriormente calculamos las diferentes comparaciones, mediante el test de comparaciones múltiples de suma de rangos de Wilcoxon, entre la media del precio de la vivienda en cada una de las categorías. Aplicamos la corrección de Bonferroni para ajustar el error provocado por las múltiples comparaciones.

```

# Kruskal Wallis Test One Way Anova by Ranks
kruskal.test(data$median_house_value~data$ocean_proximity)

```

```

##
## Kruskal-Wallis rank sum test
##
## data: data$median_house_value by data$ocean_proximity
## Kruskal-Wallis chi-squared = 6634.6, df = 4, p-value < 2.2e-16

```

```

pairwise.wilcox.test(data$median_house_value, data$ocean_proximity, p.adjust.method = "bonferroni")

##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
##  data:  data$median_house_value and data$ocean_proximity
##
##          <1H OCEAN INLAND ISLAND NEAR BAY
## INLAND      < 2e-16   -     -     -
## ISLAND      0.0616   0.0019   -     -
## NEAR BAY    1.5e-07   < 2e-16  0.2531   -
## NEAR OCEAN  0.6653   < 2e-16  0.1656  0.0342
##
## P value adjustment method: bonferroni

```

Análisis de componentes principales (PCA)

Realizamos un análisis de componentes principales con tal de buscar las componentes que son combinación lineal unitaria de las variables seleccionadas, es decir, todas las variables numéricas excepto la variable de interés `median_house_price`.

Para ello, primeramente seleccionamos las variables necesarias y calculamos la matriz de covarianzas S y correlaciones R de las diferentes variables.

```

# Seleccionamos las variables numéricas
data_cp <- data[,1:8]
# Matriz de covarianzas
S <- cov(data_cp)
# Matriz de correlaciones
R <- cor(data_cp)

```

A continuación aplicamos el método de análisis de componentes principales partiendo de cada una de las matrices:

```

cpcov = princomp(covmat = S) # componentes principales saliendo de S
cpcor = princomp(covmat = R) # componentes principales saliendo de R
summary(cpcov, loading=T) # loading = TRUE añade los vectores propios

```

```

## Importance of components:
##                               Comp.1       Comp.2       Comp.3       Comp.4
## Standard deviation    2459.2979163 533.53551085 1.731674e+02 5.215970e+01
## Proportion of Variance 0.9501207  0.04471808 4.710733e-03 4.273923e-04
## Cumulative Proportion 0.9501207  0.99483876 9.995495e-01 9.999769e-01
##                               Comp.5       Comp.6       Comp.7       Comp.8
## Standard deviation    1.168138e+01 2.839246e+00 1.543609e+00 4.974131e-01
## Proportion of Variance 2.143605e-05 1.266376e-06 3.743095e-07 3.886789e-08
## Cumulative Proportion 9.999983e-01 9.999996e-01 1.000000e+00 1.000000e+00
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## longitude            0.675  0.154  0.722
## latitude             -0.733   0.680

```

```

## housing_median_age -1.000
## total_rooms 0.882 -0.447 0.150
## total_bedrooms 0.162 -0.743 0.647
## population 0.418 0.887 0.190
## households 0.147 0.102 -0.623 -0.761
## median_income -0.988 0.130

summary(cpcor, loading=T)

## Importance of components:
##                                         Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
## Standard deviation      1.9765454 1.3812850 1.0353331 0.9066495 0.38478514
## Proportion of Variance 0.4883414 0.2384935 0.1339893 0.1027517 0.01850745
## Cumulative Proportion  0.4883414 0.7268350 0.8608243 0.9635760 0.98208342
##                                         Comp.6   Comp.7   Comp.8
## Standard deviation      0.2857789 0.216585551 0.121464969
## Proportion of Variance 0.0102087 0.005863663 0.001844217
## Cumulative Proportion  0.9922921 0.998155783 1.000000000
##
## Loadings:
##             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## longitude          0.701           0.100  0.478  0.504
## latitude          -0.702           0.464  0.522
## housing_median_age -0.218       0.394 -0.886
## total_rooms         0.484       -0.115  0.317  0.558 -0.550  0.153
## total_bedrooms     0.491       0.117   0.377 -0.231  0.221 -0.702
## population         0.472       0.116   -0.849  0.131           -0.134
## households         0.492       0.110   0.139 -0.403  0.302  0.678
## median_income      -0.891 -0.408           0.169

```

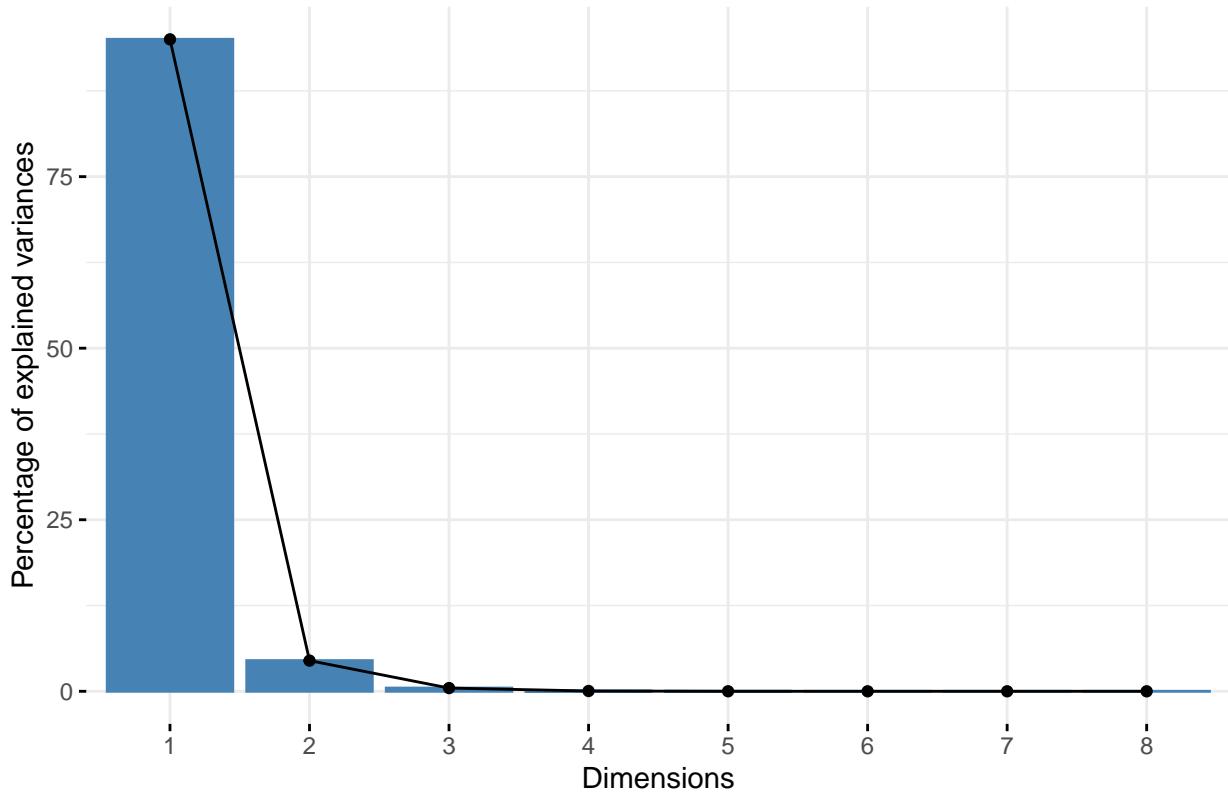
Podemos observar que las componentes principales partiendo de la matriz de covarianzas son muy diferentes a si partimos de la matriz de correlaciones, es decir, obtenemos desviaciones típicas y vectores propios muy diferentes. Para nuestro análisis partiremos de la matriz de covarianzas. En el gráfico siguiente se observa el porcentaje de variable explicado por cada componente principal, claramente, más de un 95% de la variable la explica un único componente principal.

```

data_cp = as.matrix(data_cp)
fviz_eig(cpcov)

```

Scree plot



```
cpdata = cpcor$loading[,1]
# Multiplicamos veps por los datos
cp = data_cp %*% cpdata
```

Análisis de regresión

Pasamos a hacer un análisis de regresión, específicamente un análisis de regresión cuantil. Debido a la gran cantidad de valores extremos que se recogen en la mayoría de variables, las suposiciones de normalidad y homocedasticidad del modelo lineal no se cumplían, por lo que un análisis de regresión por cuantiles nos permitirá estudiar el efecto de las diferentes variables regresoras sobre la distribución de la variable respuesta `median_house_price` a través de los cuantiles de ésta.

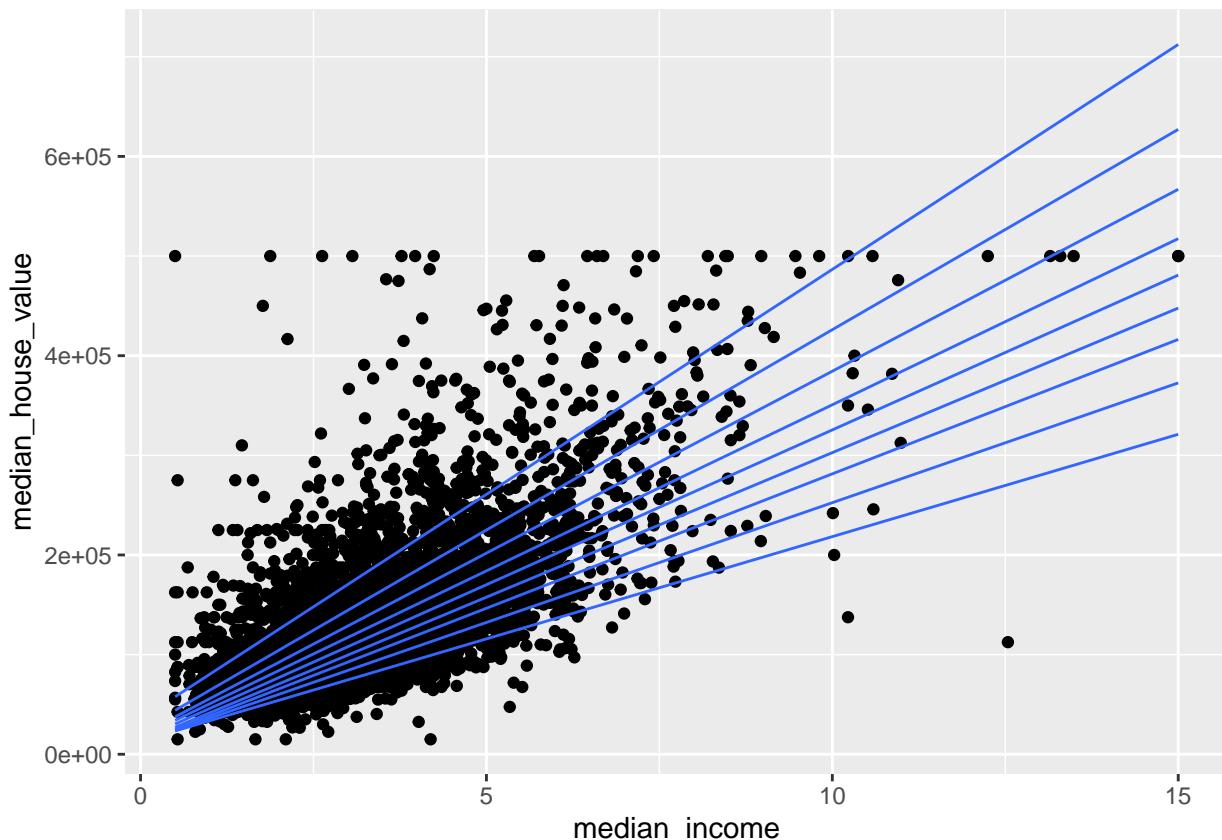
```
# Modelo reg cuantil con todas las variables y todos los cuantiles
model <- rq(median_house_value ~ housing_median_age + total_rooms + total_bedrooms +
             population + households + median_income + ocean_proximity,
             data = data, tau = 1:9/10)
# Modelo reg cuantil con todas las variables para la mediana
model_Med <- rq(median_house_value ~ housing_median_age + total_rooms + total_bedrooms +
                  population + households + median_income + ocean_proximity,
                  data = data, tau = 0.5)

# Pseudo-R squared de Nagelkerke
nagelkerke(model_Med) [[2]][3,1]
```

```
## [1] 0.689055
```

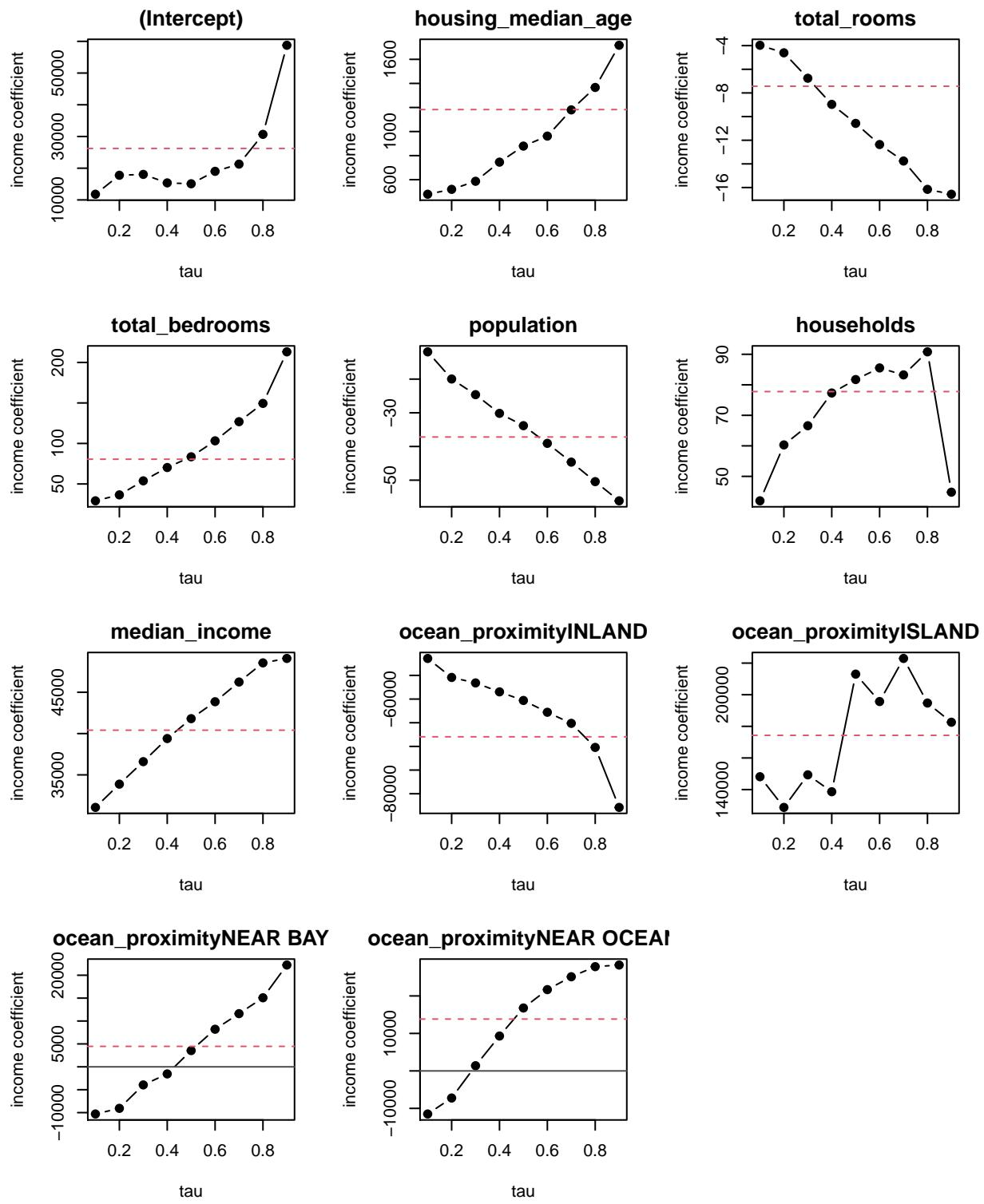
```
ggplot(data[data$ocean_proximity=="INLAND",], aes(median_income, median_house_value)) +
  geom_point() +
  geom_quantile(quantiles = 1:9/10)
```

Smoothing formula not specified. Using: $y \sim x$



Se observa un valor del coeficiente Pseudo- R^2 relativamente alto, con un valor de 0.68. En la gráfica siguiente observamos la variación de la coordenada en el origen y de los coeficientes de las variables del modelo según los diferentes cuantiles de la variable respuesta `median_house_price`.

```
plot(model, mar = c(5.1, 4.1, 2.1, 2.1), xlab = "tau",
      ylab = "income coefficient", cex = 1, pch = 19)
```



Análisis de correlaciones

Pasamos a realizar un análisis de correlaciones entre las variables numéricas del conjunto de datos. Primeramente mostramos en el siguiente panel el coeficiente de correlación entre las variables:

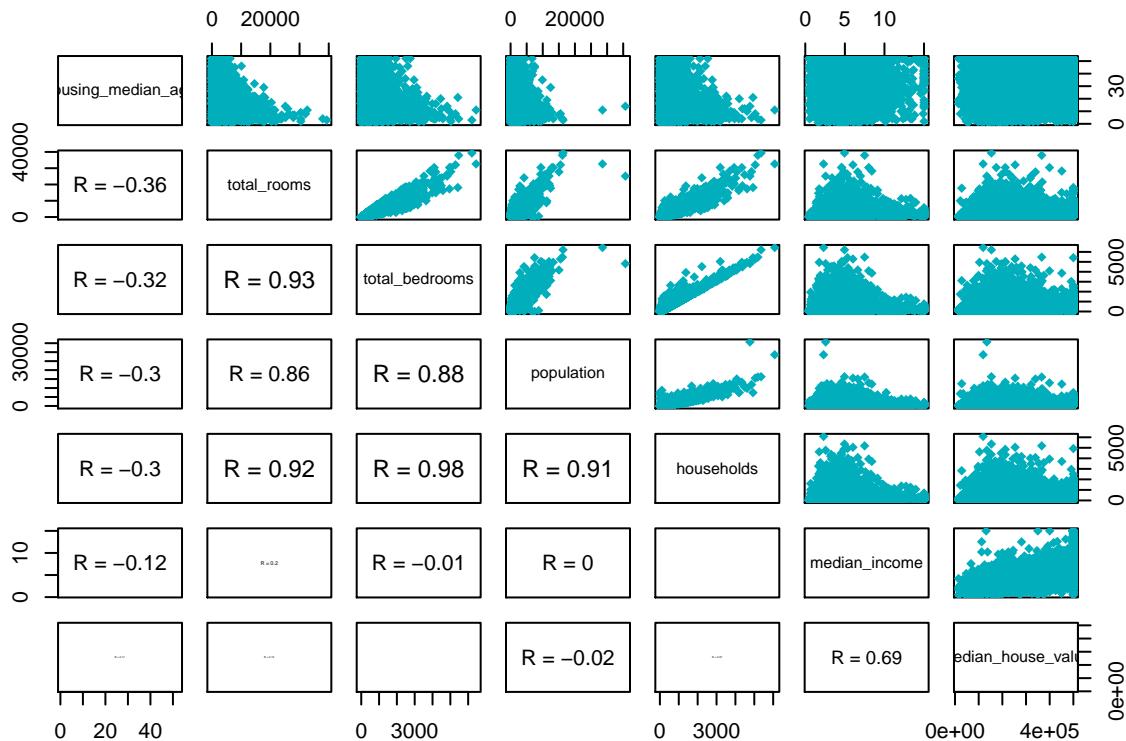
```

# Correlation panel
panel.cor <- function(x, y){
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y), digits=2)
  txt <- paste0("R = ", r)
  cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

# Customize upper panel
upper.panel<-function(x, y){
  #points(x,y, cex = 1, col = "#00AFBB")
  points(x,y,panel = panel.smooth, cex = 1.0,
  pch = 18, bg = "light blue", cex.labels = 2, font.labels = 2, col = "#00AFBB")
}

pairs(data[,c(3:9)],
      lower.panel = panel.cor,
      upper.panel = upper.panel)

```



Observamos en el panel inferior de correlaciones, donde observamos que las variables más correlacionadas entre si son: `total_rooms` y `total_bedrooms`, `total_bedrooms` y `households`, y `households` con `population`, todas ellas tiene un índice de correlación por encima de 0.90. Respecto a la variable que nos interesa, `median_house_value`, la variable más correlacionada es `median_income`, con una correlación de 0.69.

Mediante el paquete `correlation` podemos realizar un análisis de correlaciones directamente como se muestra a continuación

```
# tabla del análisis de correlaciones
results <- correlation(data[,3:9])
results

## # Correlation Matrix (pearson-method)
##
## Parameter1 | Parameter2 | r | 95% CI | t(20638) | p
## -----
## housing_median_age | total_rooms | -0.36 | [-0.37, -0.35] | -55.66 | < .001***  

## housing_median_age | total_bedrooms | -0.32 | [-0.33, -0.31] | -48.60 | < .001***  

## housing_median_age | population | -0.30 | [-0.31, -0.28] | -44.56 | < .001***  

## housing_median_age | households | -0.30 | [-0.32, -0.29] | -45.66 | < .001***  

## housing_median_age | median_income | -0.12 | [-0.13, -0.11] | -17.22 | < .001***  

## housing_median_age | median_house_value | 0.11 | [ 0.09, 0.12] | 15.26 | < .001***  

## total_rooms | total_bedrooms | 0.93 | [ 0.93, 0.93] | 363.31 | < .001***  

## total_rooms | population | 0.86 | [ 0.85, 0.86] | 239.05 | < .001***  

## total_rooms | households | 0.92 | [ 0.92, 0.92] | 333.66 | < .001***  

## total_rooms | median_income | 0.20 | [ 0.18, 0.21] | 29.03 | < .001***  

## total_rooms | median_house_value | 0.13 | [ 0.12, 0.15] | 19.45 | < .001***  

## total_bedrooms | population | 0.88 | [ 0.87, 0.88] | 263.55 | < .001***  

## total_bedrooms | households | 0.98 | [ 0.98, 0.98] | 705.19 | < .001***  

## total_bedrooms | median_income | -7.86e-03 | [-0.02, 0.01] | -1.13 | 0.517  

## total_bedrooms | median_house_value | 0.05 | [ 0.04, 0.06] | 7.29 | < .001***  

## population | households | 0.91 | [ 0.90, 0.91] | 309.83 | < .001***  

## population | median_income | 4.83e-03 | [-0.01, 0.02] | 0.69 | 0.517  

## population | median_house_value | -0.02 | [-0.04, -0.01] | -3.54 | < .01**  

## households | median_income | 0.01 | [ 0.00, 0.03] | 1.87 | 0.183  

## households | median_house_value | 0.07 | [ 0.05, 0.08] | 9.48 | < .001***  

## median_income | median_house_value | 0.69 | [ 0.68, 0.70] | 136.22 | < .001***  

##  

## p-value adjustment method: Holm (1979)  

## Observations: 20640
```

Representación de los resultados y resolución del problema

En primer lugar, en cuanto a las comparaciones múltiples sobre el precio del habitaje según las diferentes localizaciones de estos se observan diferencias significativas, con un nivel de confianza del 95%, entre la media del precio del habitaje de "INLAND" y "NEAR_BAY" con "<1H OCEAN"; también entre "NEAR_OCEAN" y "NEAR_BAY" con "INLAND"; aunque menos significativa, existe también diferencia entre "NEAR_OCEAN" y "NEAR_BAY".

En segundo lugar, en cuanto al análisis de componentes principales, se observa que una única componente principal podría explicar más del 95% de la variancia.

En tercer lugar, en cuanto a la regresión cuantil, se puede apreciar en el gráfico obtenido que los coeficientes β_0 y β_i , donde i son las diferentes variables estudiadas, varían mucho según en qué cuantil de la distribución de la variable respuesta se esté analizando. En el caso de `housing_median_age`, `total_bedrooms`, `median_income`, y las categorías "NEAR_OCEAN" y "NEAR_OCEAN" con respecto a la categoría de referencia "<1H OCEAN", se produce un incremento en el coeficiente a medida que aumentamos de cuantil; en caso contrario, los coeficientes de las variables `total_rooms`, `population`, y la categoría "ISLAND" con respecto

a la de referencia, presentan una disminución. Como destacable observamos el incremento del coeficiente de `household` hasta el cuantil del 80% y la gran disminución de este en el cuantil del 90%.

En cuarto y último lugar, en cuanto al análisis de correlaciones se observa que casi todos los tests de hipótesis para el coeficiente de correlación dan como resultado un valor $P < .001$, siendo significativos con un nivel de significación del 95%. Los coeficientes que no son significativos, es decir los que serían 0, son para las variables `median_income` con `tota_bedrooms`, `households` y `population`, con valores $P > .05$.

Enlace Github:

<https://github.com/AlexanderAlmendral/PRA2-California-House-Prices>

```
ft_firma = flextable()
data.frame(Contribuciones = c("Investigación previa",
                               "Redacción de las respuestas", "Desarrollo código"),
           Firma = c("Alexander Almendral, Pau Ortí",
                     "Alexander Almendral, Pau Ortí",
                     "Alexander Almendral, Pau Ortí")))
ft_firma = autofit(ft_firma)
ft_firma
```

Contribuciones	Firma
Investigación previa	Alexander Almendral, Pau Ortí
Redacción de las respuestas	Alexander Almendral, Pau Ortí
Desarrollo código	Alexander Almendral, Pau Ortí