

California

Laura

3 de diciembre de 2016

El conjunto de datos California

El conjunto de datos *California* contiene datos sobre viviendas de California y pretende estimar el precio de una nueva vivienda.

Las variables con las que se pretende estimar el precio de la vivienda son las siguientes:

- Longitud (*longitude*)
- Latitud (*latitude*)
- La edad media de las casas (*HousingMedianAge*)
- El número de habitaciones (*TotalRooms*)
- El número de dormitorios (*TotalBedrooms*)
- El número de habitantes (*Population*)
- El número de unidades familiares en el edificio (*Households*)
- La media de ingresos (*MedianIncome*)
- El valor medio de la casa (*MedianHouseValue*). El valor de esta variable es la que pretendemos obtener.

Hipótesis previas

Sin mirar el contenido del conjunto de datos se plantean las siguientes hipótesis:

1. El número de habitaciones incrementa el precio de forma lineal.
2. A mayor edad media más disminuye el precio.
3. A mayor población más incrementa el precio de la vivienda.
4. La distancia del centro disminuye el precio.
5. La media de ingresos hace aumenta el precio.
6. El número de unidades familiares en la vivienda hace disminuir el precio.

El conjunto de datos

La comprobación de las anteriores hipótesis la podemos realizar mirando como se comportan cada una de las variables con respecto a la salida.

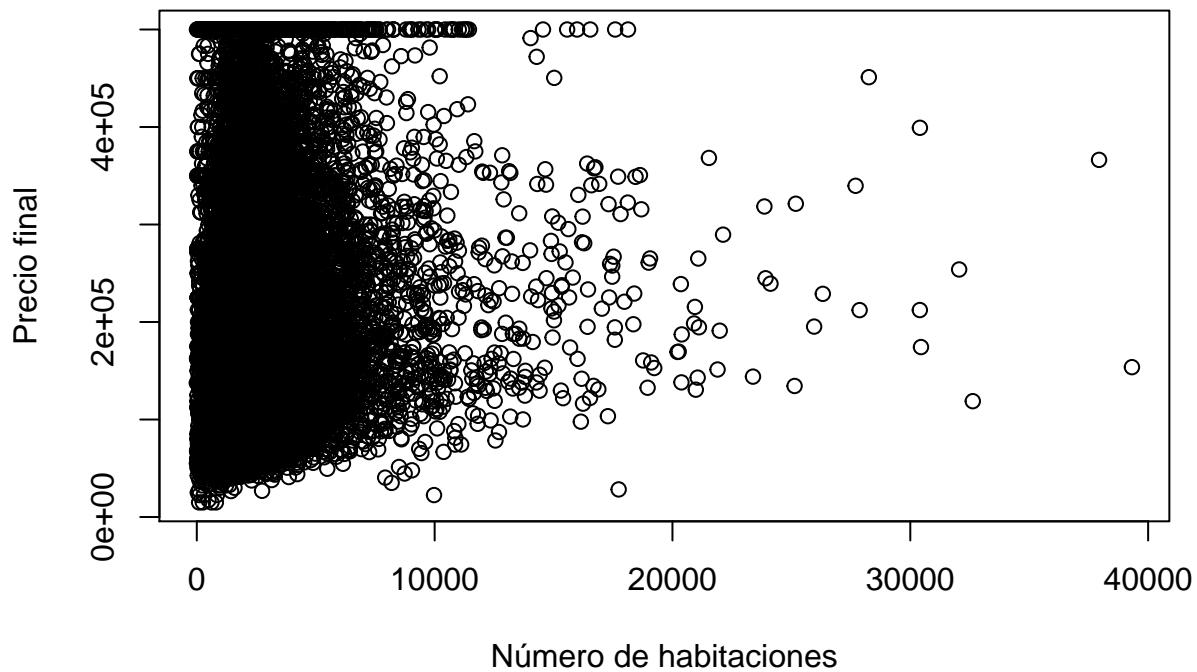
Para ello primero cargamos la base de datos

```
california_original <- read.csv("california.dat", header = FALSE,
  comment.char = "@")
names(california_original) <- c("Longitude", "Latitude", "HousingMedianAge",
  "TotalRooms", "TotalBedrooms", "Population", "Households",
  "MedianIncome", "MedianHouseValue")
```

Y mostramos las comparaciones con la salida de las variables anteriores

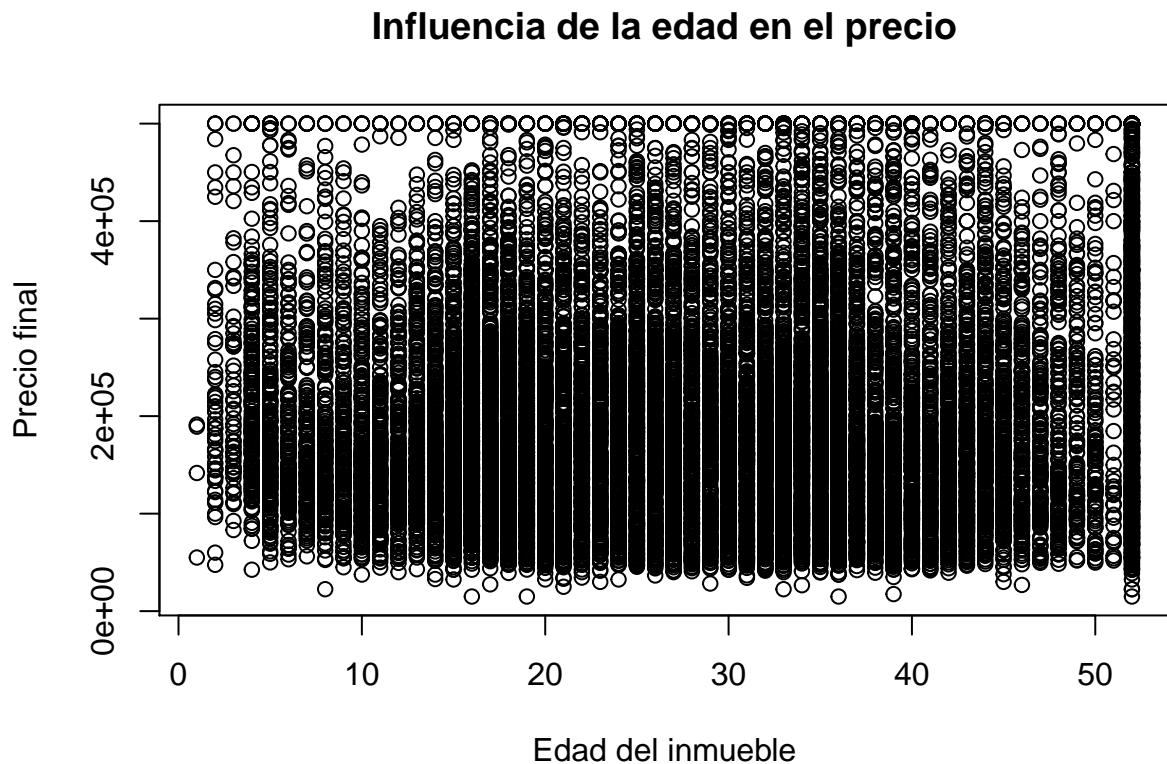
```
plot(california_original$TotalRooms, california_original$MedianHouseValue,  
      main = "Influencia del número de habitaciones en el precio",  
      xlab = "Número de habitaciones", ylab = "Precio final")
```

Influencia del número de habitaciones en el precio



Aparentemente no existe una relación lineal entre el número de habitaciones y el precio.

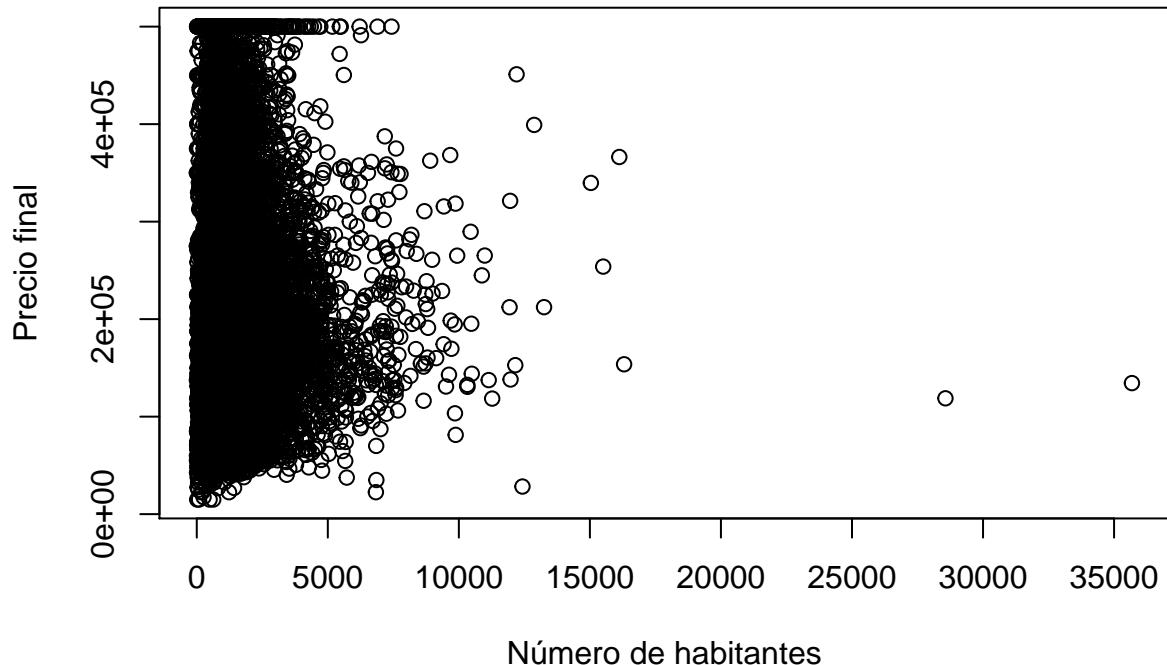
```
plot(california_original$HousingMedianAge, california_original$MedianHouseValue,  
     main = "Influencia de la edad en el precio", xlab = "Edad del inmueble",  
     ylab = "Precio final")
```



Esta variable tampoco tiene una relación lineal como habíamos deducido.

```
plot(california_original$Population, california_original$MedianHouseValue,  
      main = "Influencia del número de habitantes en el precio",  
      xlab = "Número de habitantes", ylab = "Precio final")
```

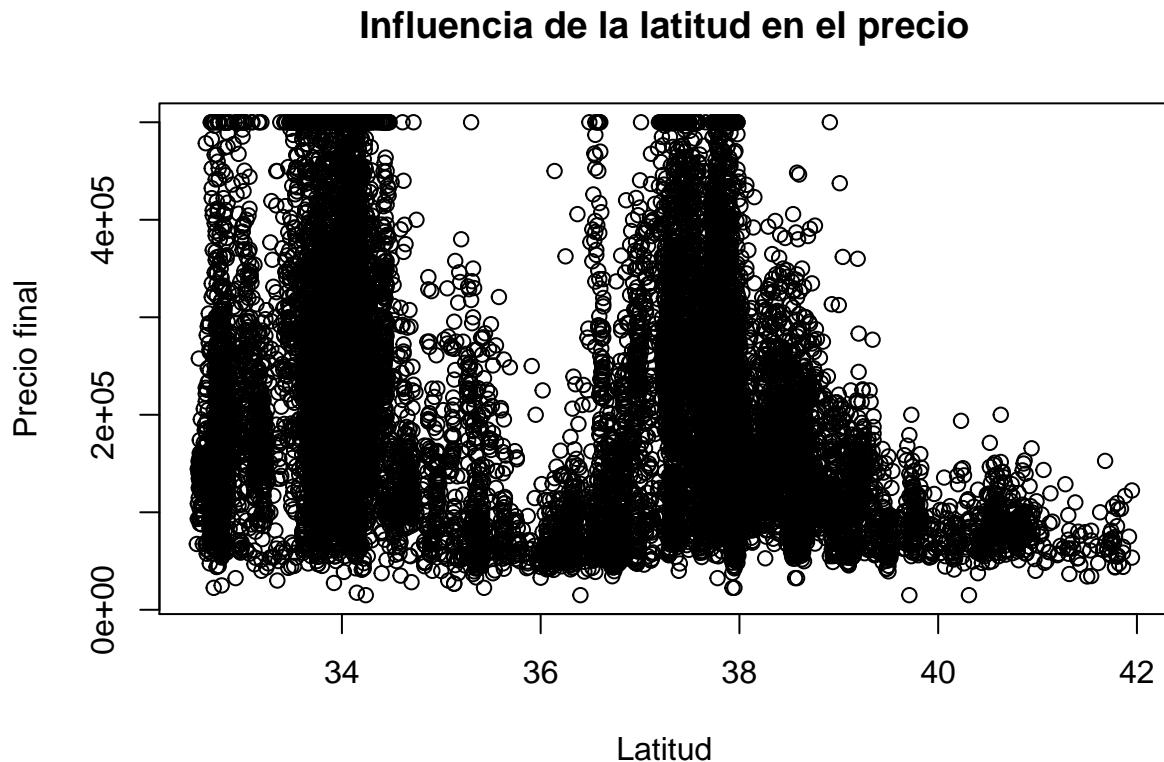
Influencia del número de habitantes en el precio



Esta variable tampoco tiene una relación lineal con la salida, pero tiene una forma similar al gráfico del número de habitaciones.

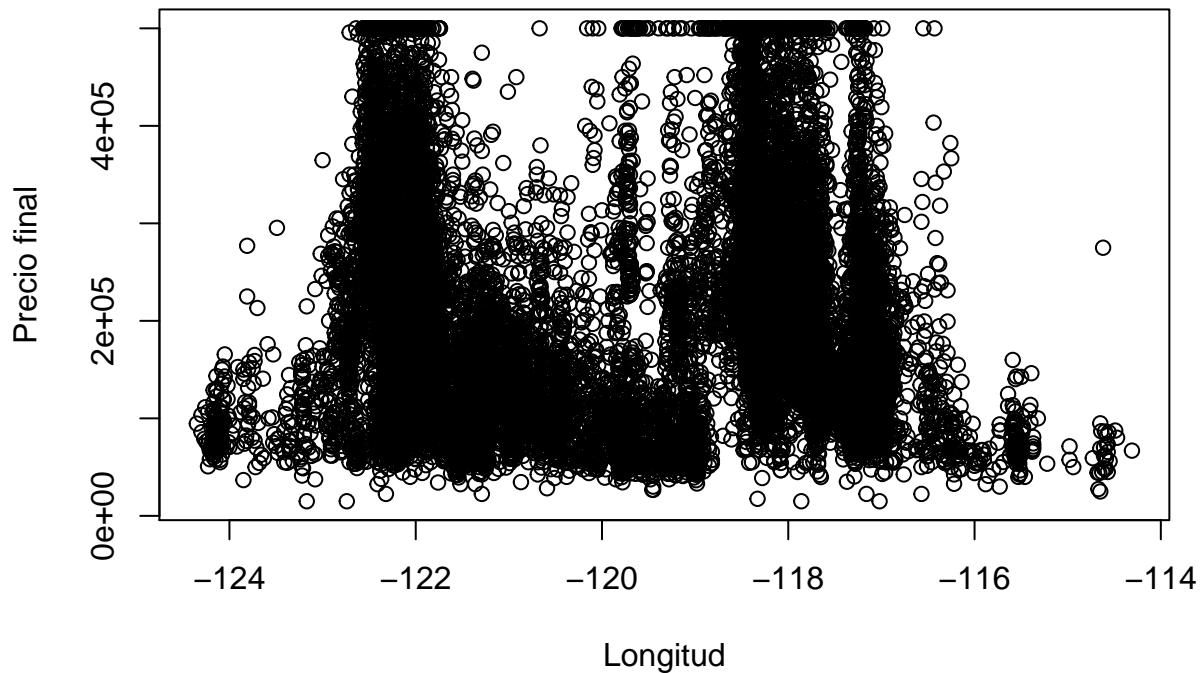
Para calcular la distancia al centro como hemos planteado en la hipótesis 4 tenemos que calcular un centro. Pero antes de eso miremos como se comportan las variables longitud y latitud por separado con respecto de la salida.

```
plot(california_original$Latitude, california_original$MedianHouseValue,  
      main = "Influencia de la latitud en el precio", xlab = "Latitud",  
      ylab = "Precio final")
```



```
plot(california_original$Longitude, california_original$MedianHouseValue,  
      main = "Influencia de la longitud en el precio", xlab = "Longitud",  
      ylab = "Precio final")
```

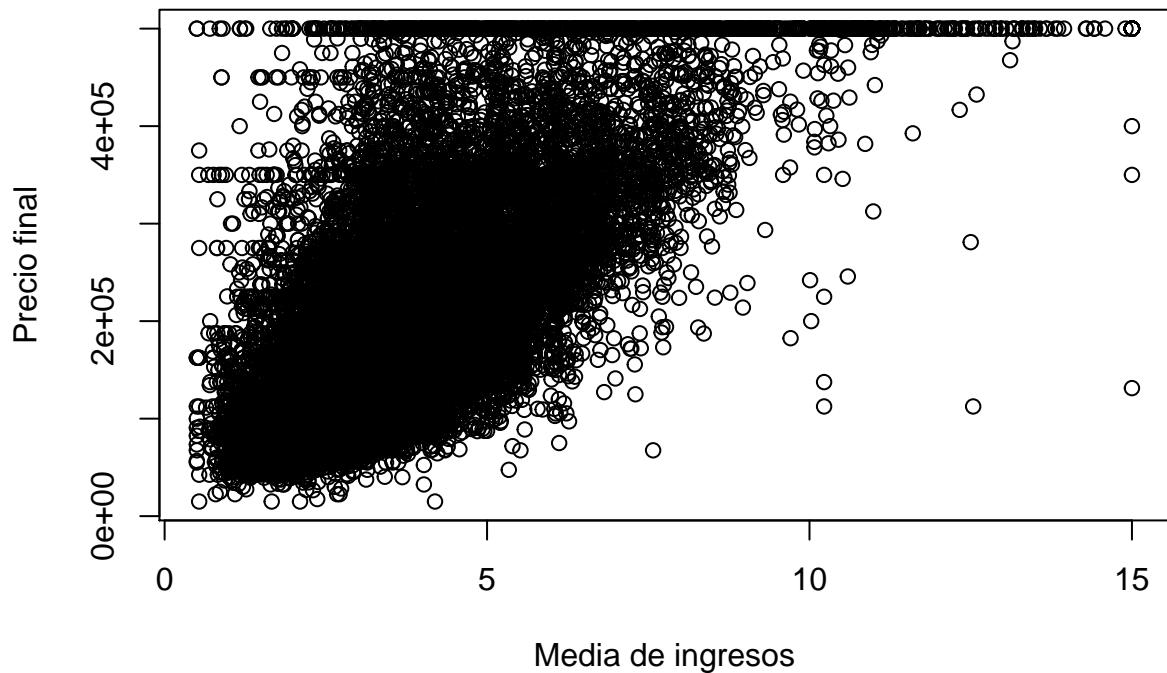
Influencia de la longitud en el precio



Claramente la relación latitud/longitud con el precio final no es lineal, pero el hecho de que existan varios picos me hace pensar que puede haber varias ciudades con lo que calcular la media para establecerlo como centro no es una buena opción.

```
plot(california_original$MedianIncome, california_original$MedianHouseValue,  
     main = "Influencia de la media de los ingresos en el precio",  
     xlab = "Media de ingresos", ylab = "Precio final")
```

Influencia de la media de los ingresos en el precio

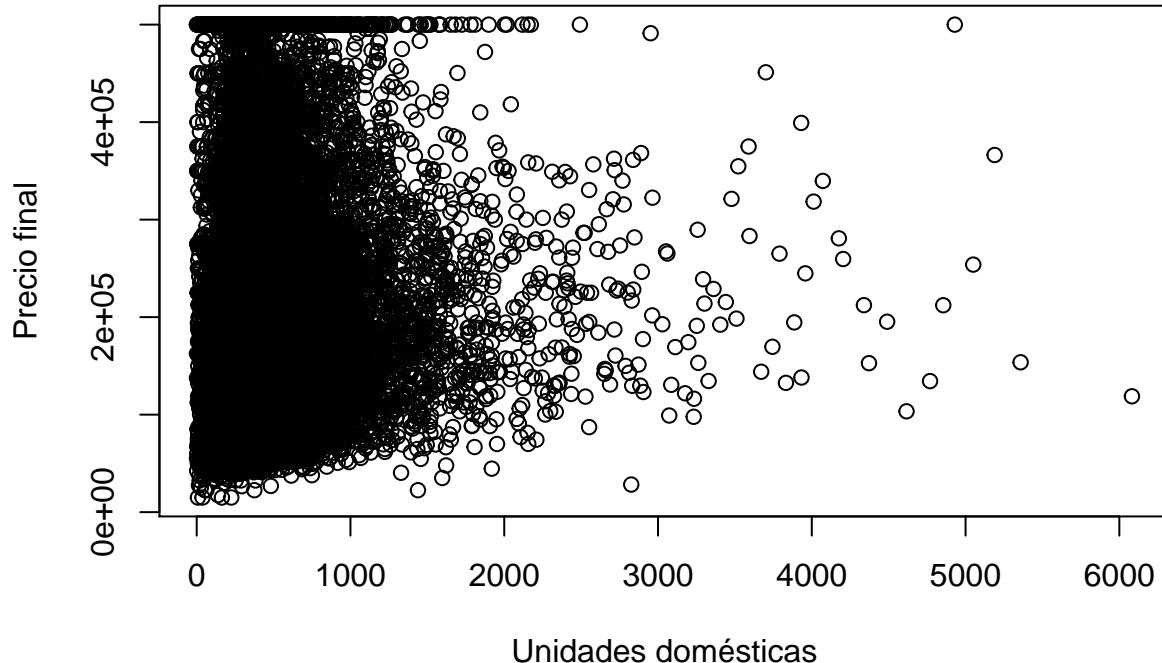


Puede ser una relación lineal con una dispersión muy alta, debido probablemente a la existencia de varias ciudades en el conjunto de datos.

Ya solo queda comprobar que al aumentar el número de unidades familiares disminuye el precio.

```
plot(california_original$Households, california_original$MedianHouseValue,
     main = "Influencia del número de unidades domésticas en el precio",
     xlab = "Unidades domésticas", ylab = "Precio final")
```

Influencia del número de unidades domésticas en el precio



No sigue una distribución lineal pero por la forma que tiene la hipótesis de que a mayor número de unidades domésticas disminuye también es falsa.

Conclusiones sobre las hipótesis previas.

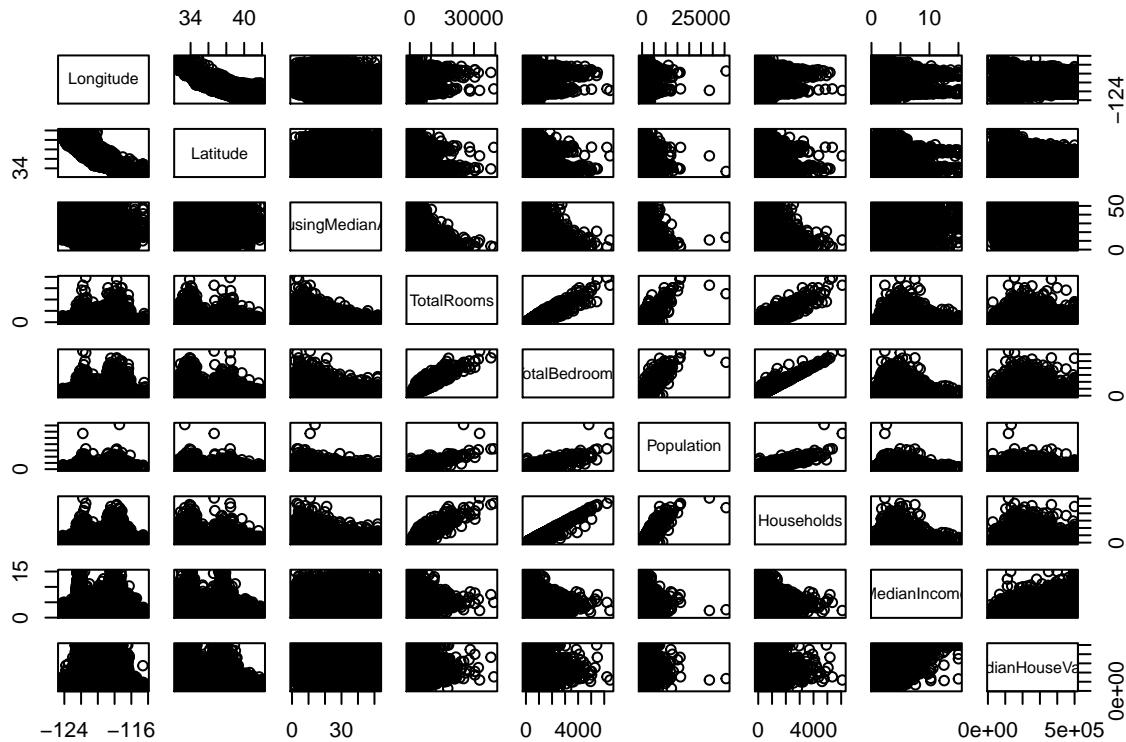
Tras ver la relación entre las variables de las hipótesis previas y la salida esperada, podemos llegar a las siguientes conclusiones:

- El conjunto de datos recoge varias ciudades.
- Al recoger varias ciudades las variables pueden presentar una tendencia lineal pero tener una amplia dispersión, relacionada con la ciudad en la que esté. ** Se debería clusterizar el conjunto de datos por ciudades. Pero el ejercicio es de modelos lineales por lo que no nos vamos a meter en la elaboración de clusters. Eso sí se espera que el modelo de regresión lineal sea bastante malo y que el modelo KNN presente mejores resultados.
- La única variable estudiada hasta ahora que parece independiente de la ciudad es la media de ingresos.

Comparación de todas las variables con la salida

Ahora compararemos todas las variables con todas para tener una idea de las interacciones que pueden existir entre ellas.

```
attach(california_original)
pairs(california_original)
```



Las variables *TotalRooms*, *TotalBedrooms*, *Population* y *Households* tienen una relación lineal entre ellas.

Construcción del modelo lineal

Modelo lineal simple

Una vez conocemos como se comportan las variables podemos plantear un modelo lineal que aproxime una solución a nuestro problema.

Podemos plantearlo como un modelo lineal de una sola variable, aunque sabemos que no va a ser un ajuste muy bueno pero de esta forma tendremos un parámetro de R cuadrada ajustada base para comparar más tarde.

La variable que se usará para este modelo es *MedianIncome* porque presentaba una tendencia lineal, más independiente que las demás estudiadas en el apartado anterior.

```
model1_singleLineal <- lm(MedianHouseValue ~ MedianIncome)
summary(model1_singleLineal)
```

```
Call:
lm(formula = MedianHouseValue ~ MedianIncome)
```

Residuals:

Min	1Q	Median	3Q	Max
-540697	-55950	-16979	36978	434023

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	45085.6	1322.9	34.08	<2e-16 ***
MedianIncome	41793.8	306.8	136.22	<2e-16 ***

Signif. codes:				
0	'***'	0.001	'**'	0.01
	*	0.05	.	0.1
	'	'	'	1

Residual standard error: 83740 on 20638 degrees of freedom
Multiple R-squared: 0.4734, Adjusted R-squared: 0.4734
F-statistic: 1.856e+04 on 1 and 20638 DF, p-value: < 2.2e-16

El valor del parámetro R cuadrada es de **0.4734**, es bastante malo pero nos sirve como base para ir aumentando el número de variables que intervienen en el modelo de regresión lineal múltiple.

Modelo lineal múltiple

Las variables que vamos a añadir al modelo serán HouseHold