Modelo de regresión lineal múltiple

Kevin Garcia - Alejandro Vargas

20 de febrero de 2018

Análisis exploratorio de datos

Para trabajar con la base de datos denominada 'cadata', generamos un número aleatorio con la ayuda del software R, el cuál nos arrojó el número 15529, por tanto nuestra base de datos final, quedo con las 9 variables (columnas) y con las filas desde la 15529 hasta la 16028.

El objetivo del estudio es ajustar un modelo de regresión para la variable 'Valor medio de la casa', tomando como variables explicativas las variables 'Ingreso medio','Edad media de la vivienda', 'Total de habitaciones','Total de dormitorios','Población','Hogares','Latitud' y 'Longitud'.

Análisis exploratorio de datos

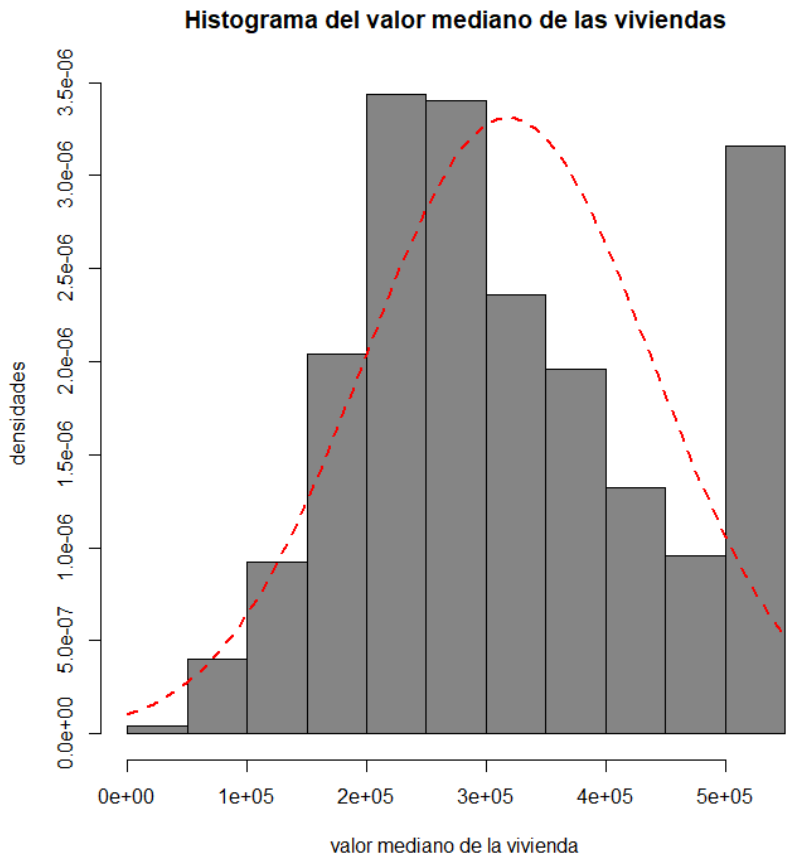
Previo al ajuste e interpretación del modelo, se llevo a cabo el respectivo análisis exploratorio de datos, para tener una idea de las descriptivas mas importantes de cada variable, su forma de distribución y su rango de valores.

Análisis exploratorio de datos

	Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.	Desviación Est.
Valor mediano de las viviendas	32500	229600	294400	318400	419100	500000	120509.8
Ingreso mediano	0.7054	2.87	3.697	3.905	4.554	15	1.735087
Edad mediana	2	34.75	52	42.2	52	52	14.5956
Total de habitaciones	52	1580	2218	2537	3166	12480	1539.323
Total de dormitorios	13	346	506	583.9	735.2	2747	360.0191
Población	55	836	1168	1303	1608	5640	779.5774
Hogares	13	327.80	469.5	538.8	664.5	2538	328.173

Figura: Resumen estadístico

Variable 'Valor mediano de las viviendas'



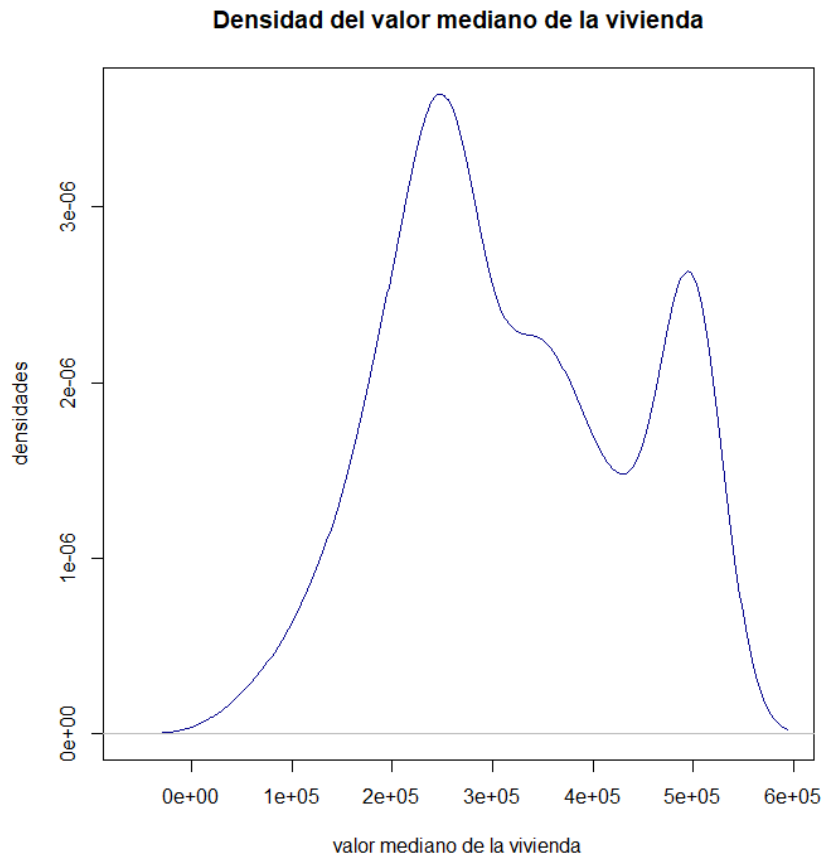


Figura: Histograma y densidad de la variable 'Valor mediano de las casas'

Variable 'Valor mediano de las viviendas'

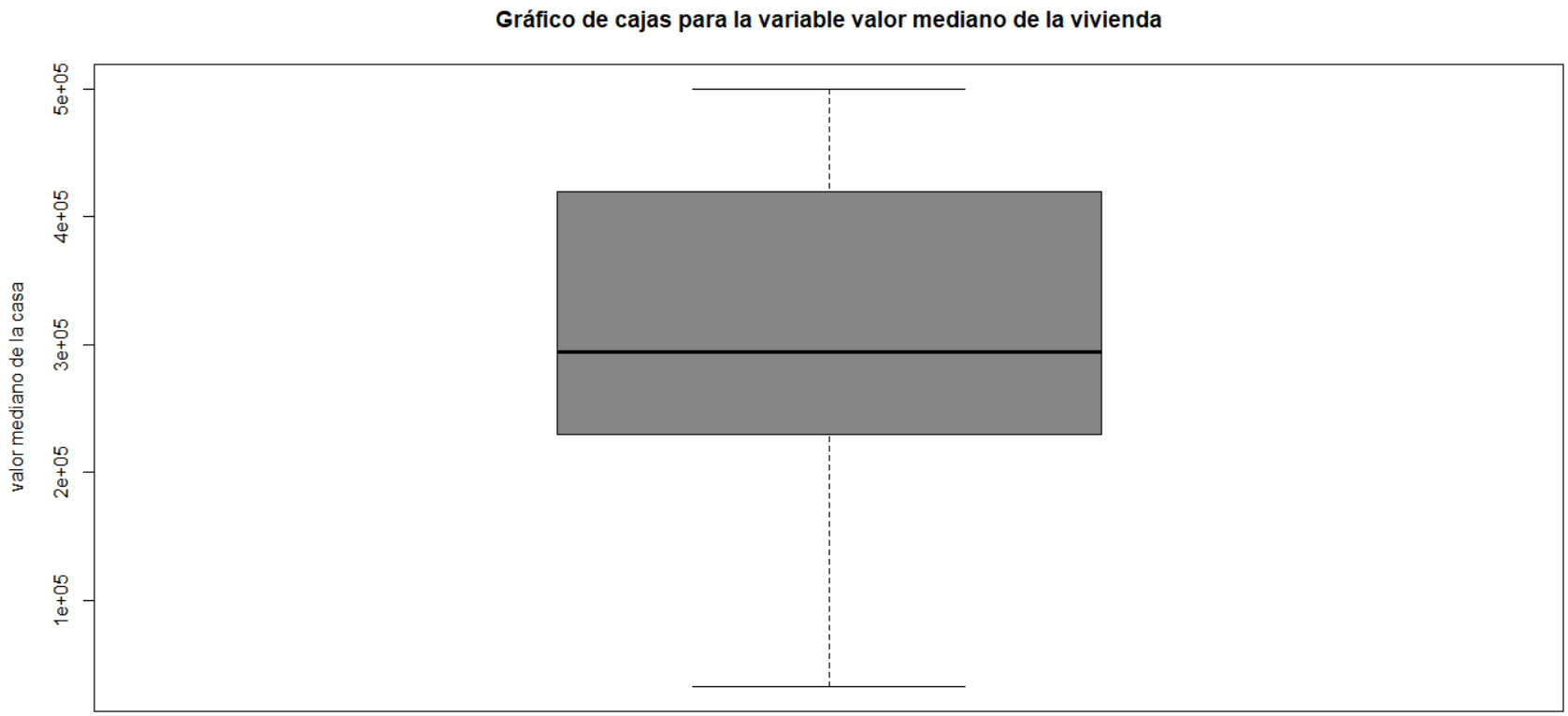


Figura: Gráfico de cajas para la variable 'Valor mediano de las casas'

Variable 'Ingreso mediano'

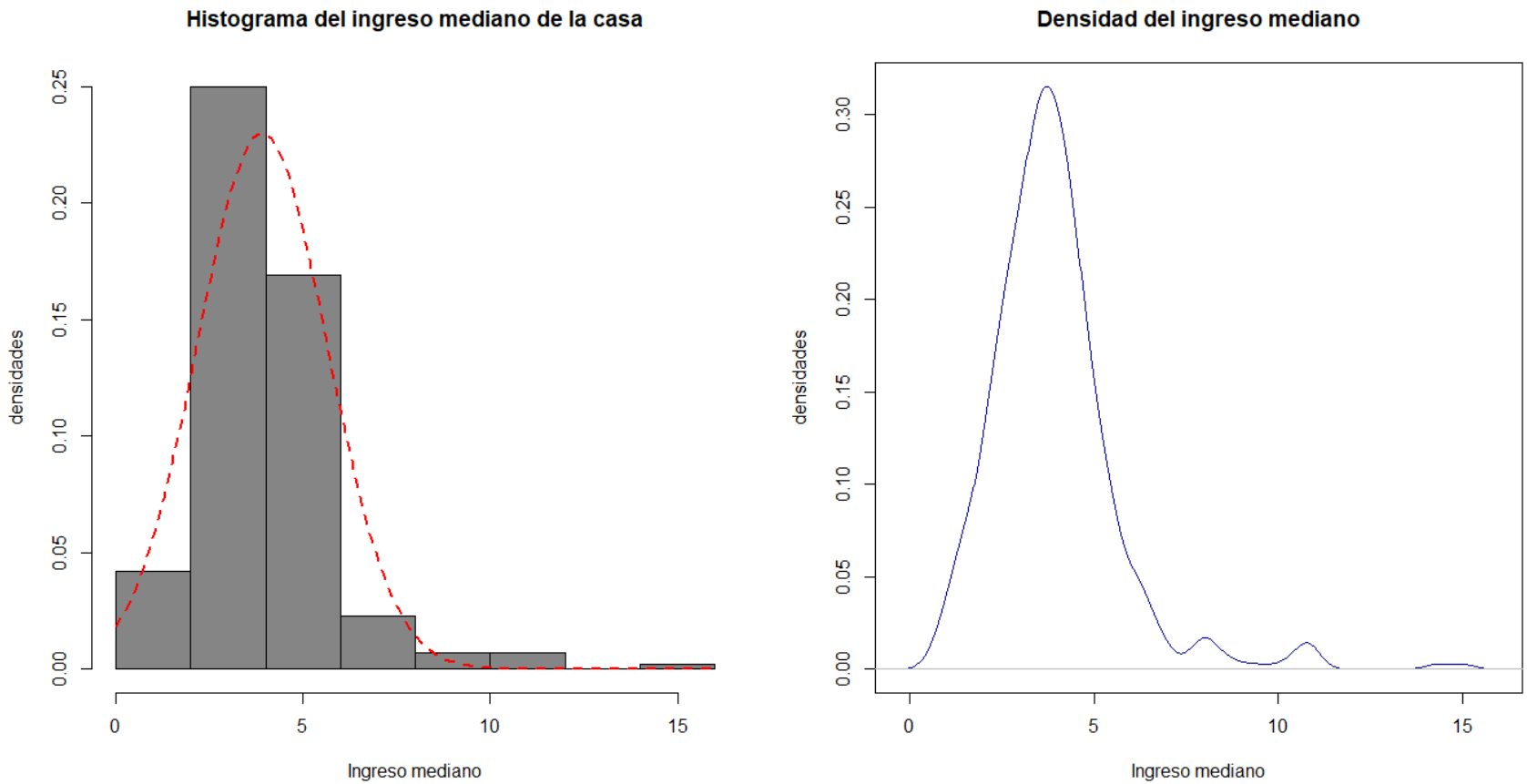


Figura: Histograma y densidad de la variable 'Ingreso mediano'

Variable 'Ingreso mediano'

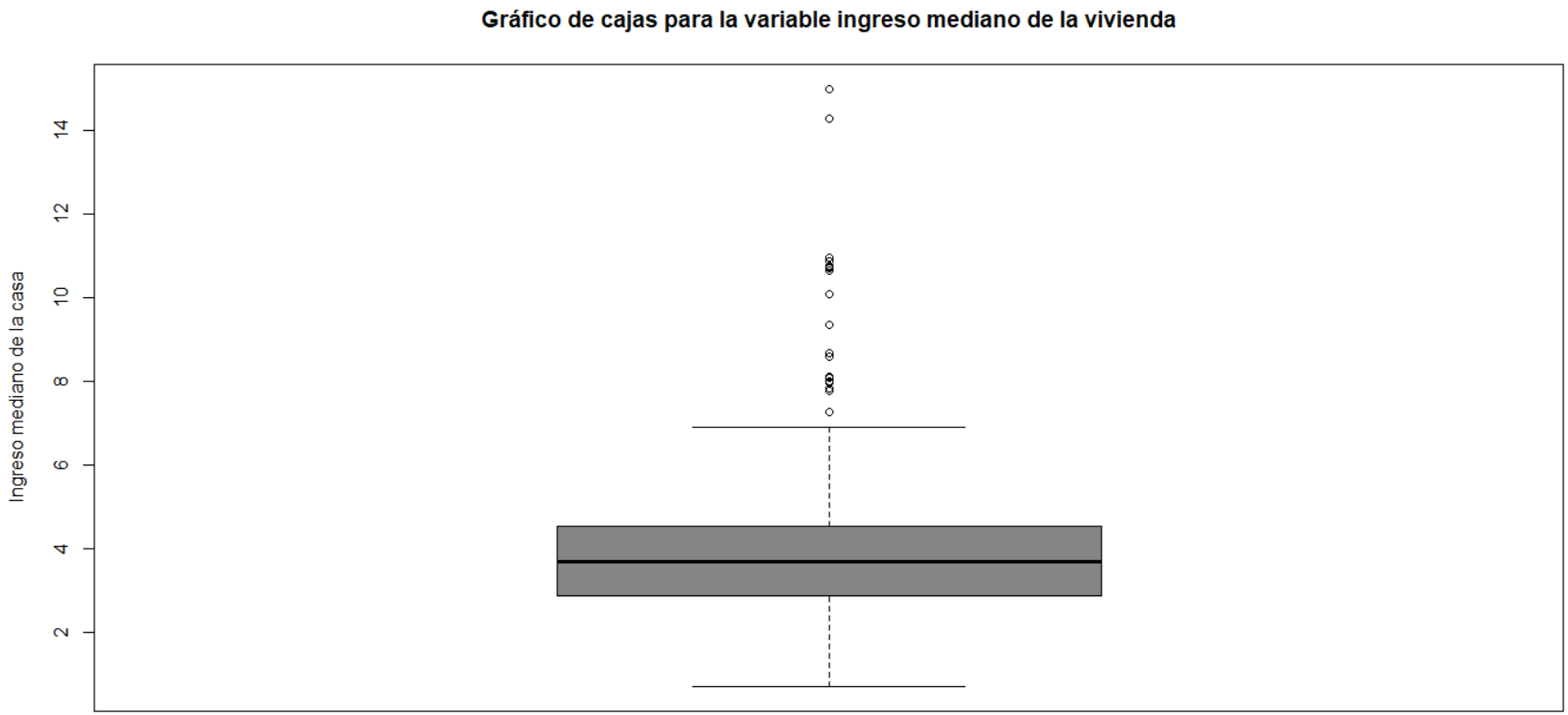


Figura: Gráfico de cajas para la variable 'Ingreso mediano de las casas'

Variable 'Edad mediana de las viviendas'

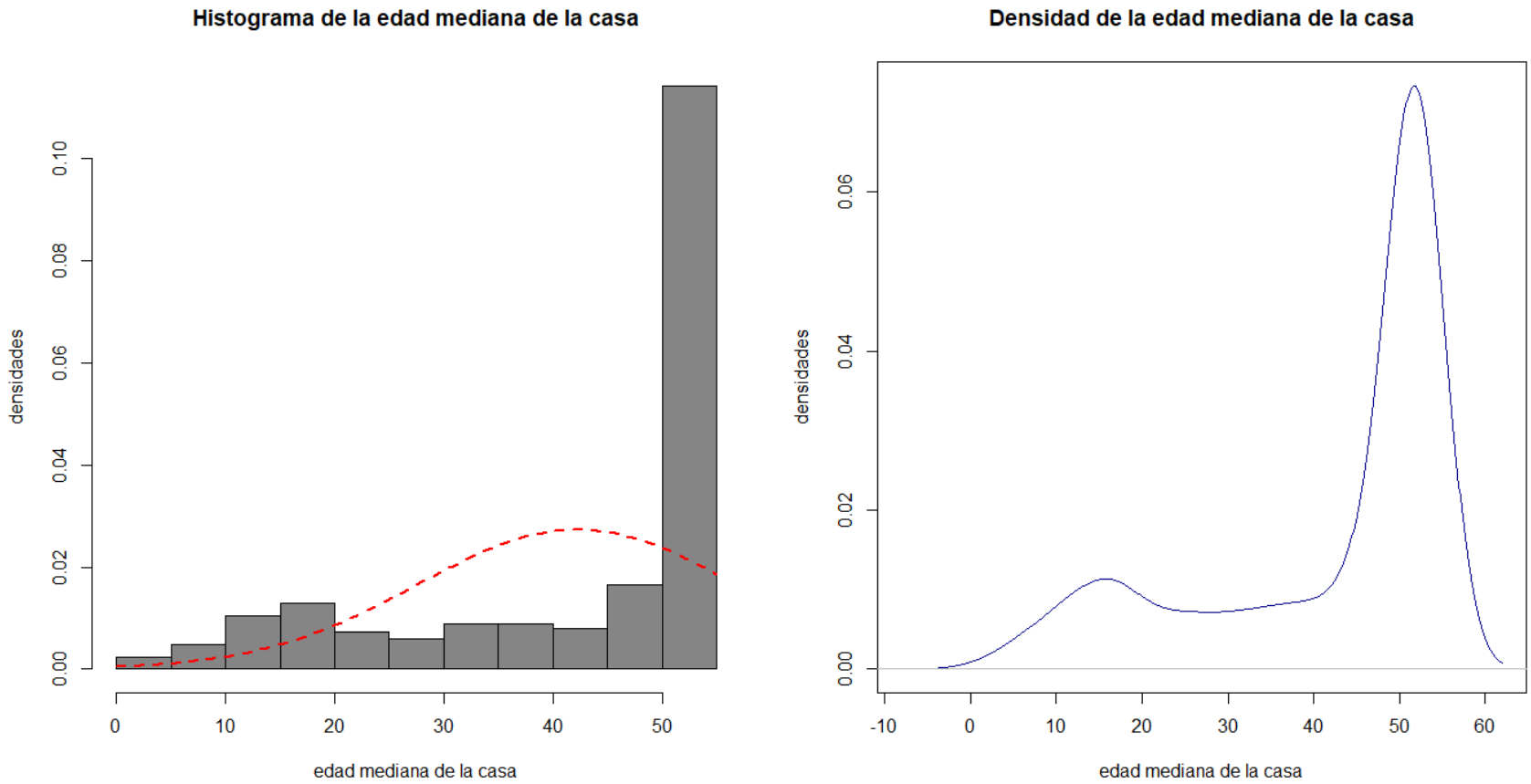


Figura: Histograma y densidad de la variable 'Edad mediana de las viviendas'

Variable 'Edad mediana de las viviendas'

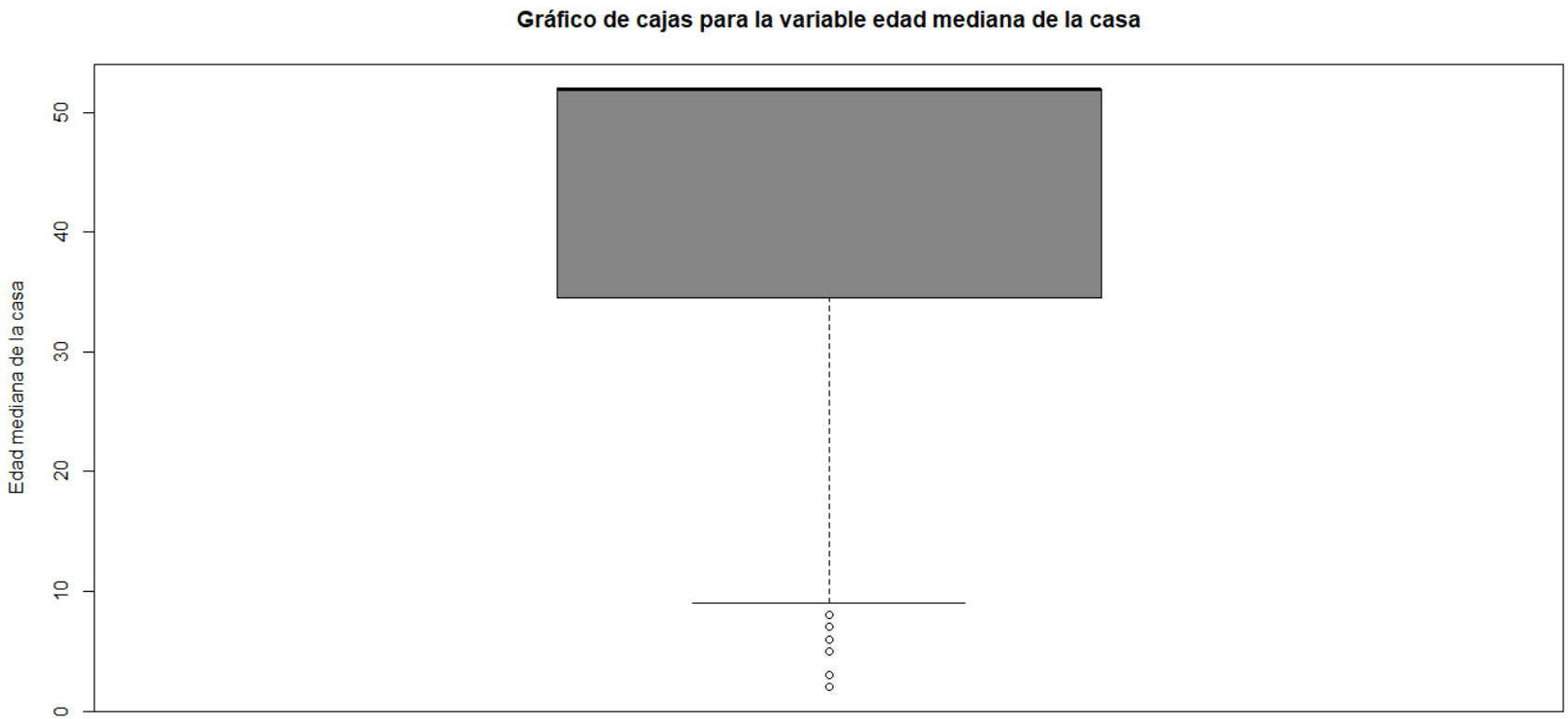


Figura: Gráfico de cajas para la variable 'Edad mediana de las viviendas'

Variable 'Total de habitaciones'

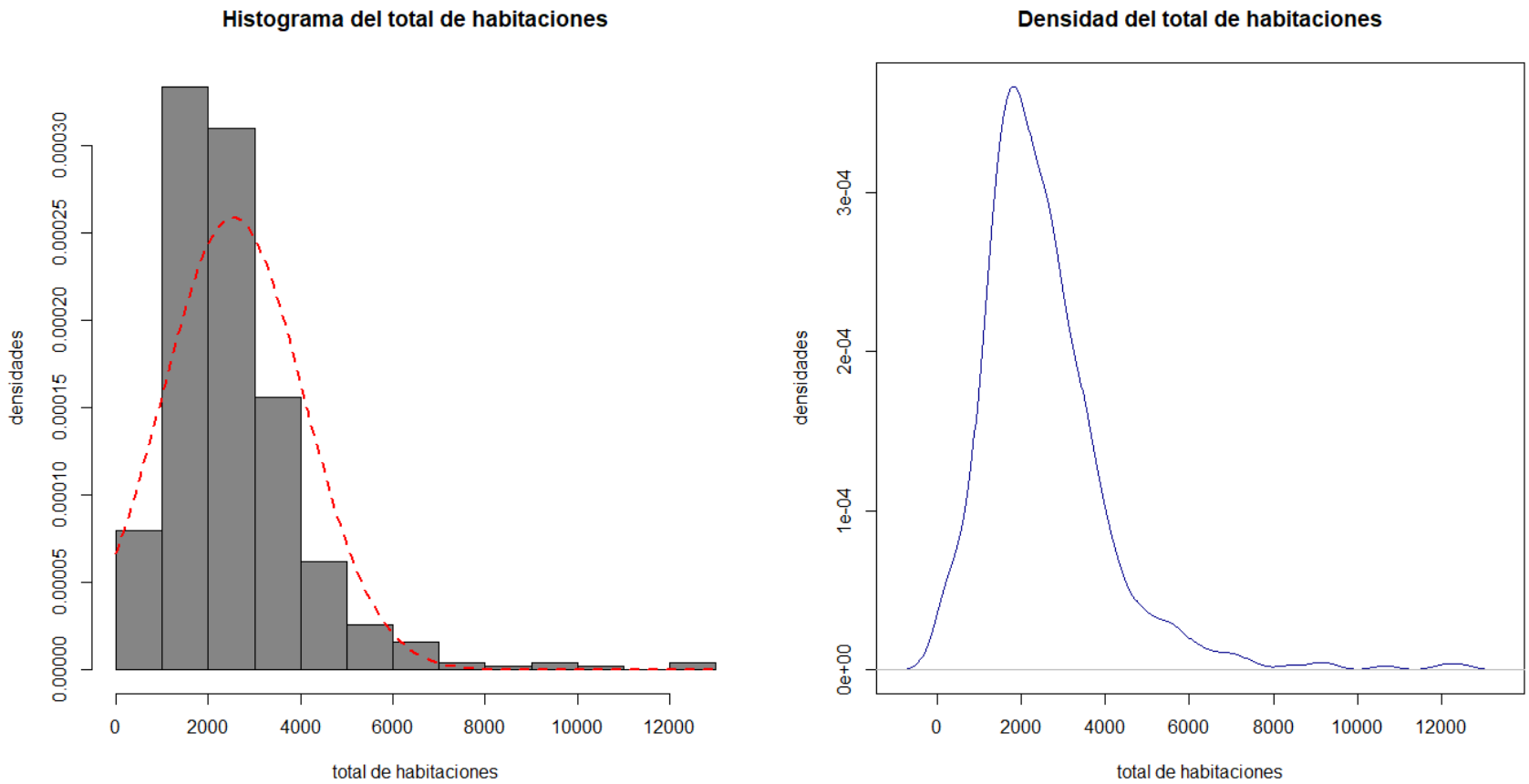


Figura: Histograma y densidad de la variable 'Total de habitaciones'

Variable 'Total de habitaciones'

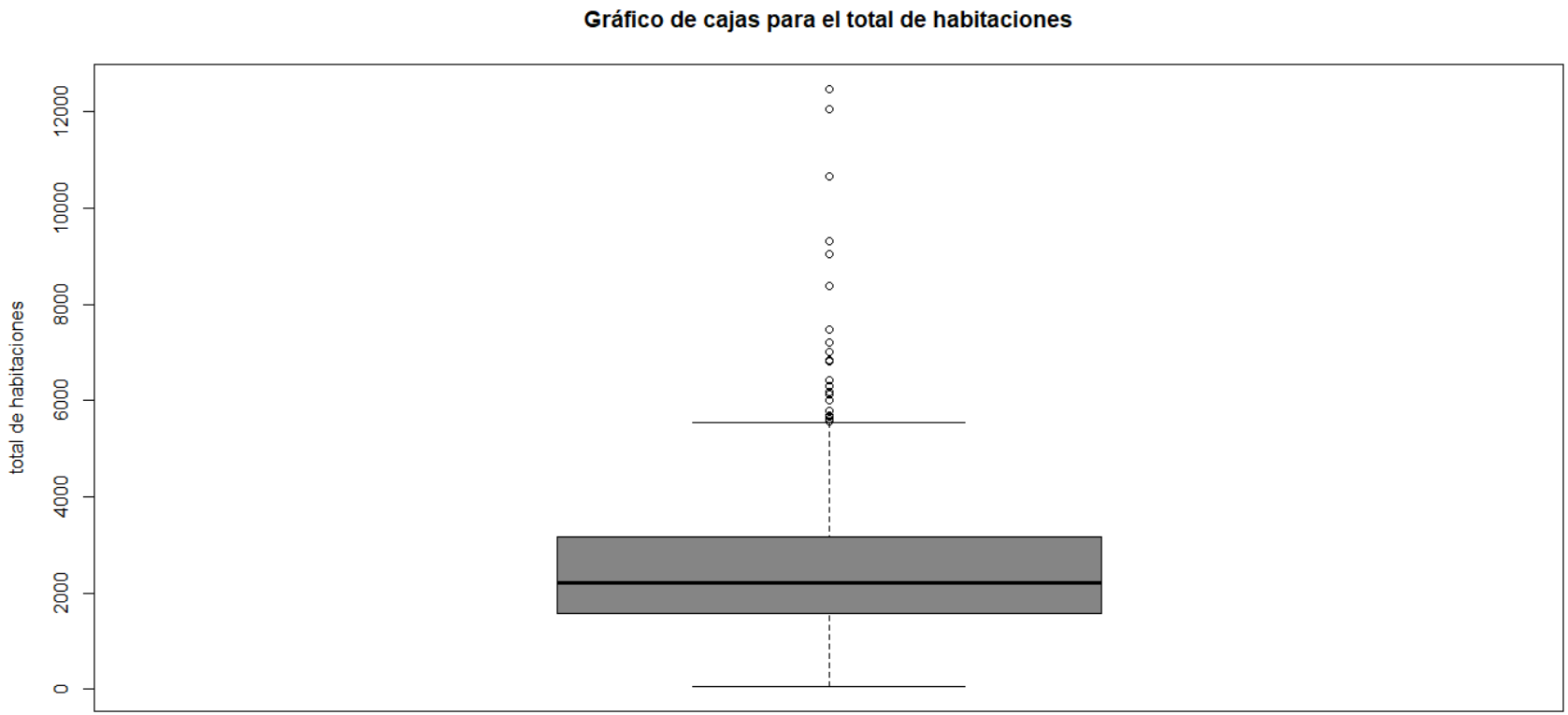


Figura: Gráfico de cajas para la variable 'Total de habitaciones'

Variable 'Total de dormitorios'

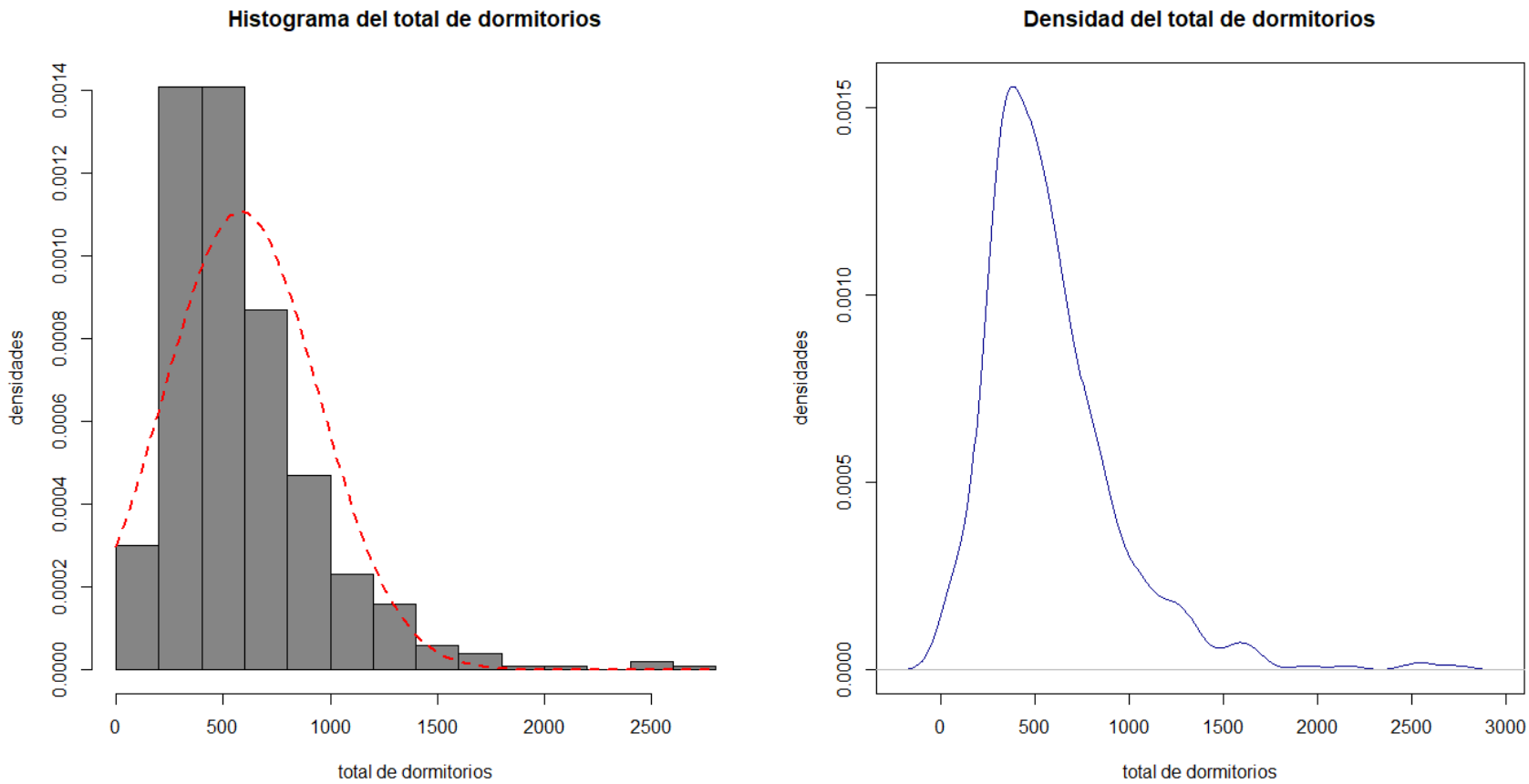


Figura: Histograma y densidad de la variable 'Total de dormitorios'

Variable 'Total de dormitorios'

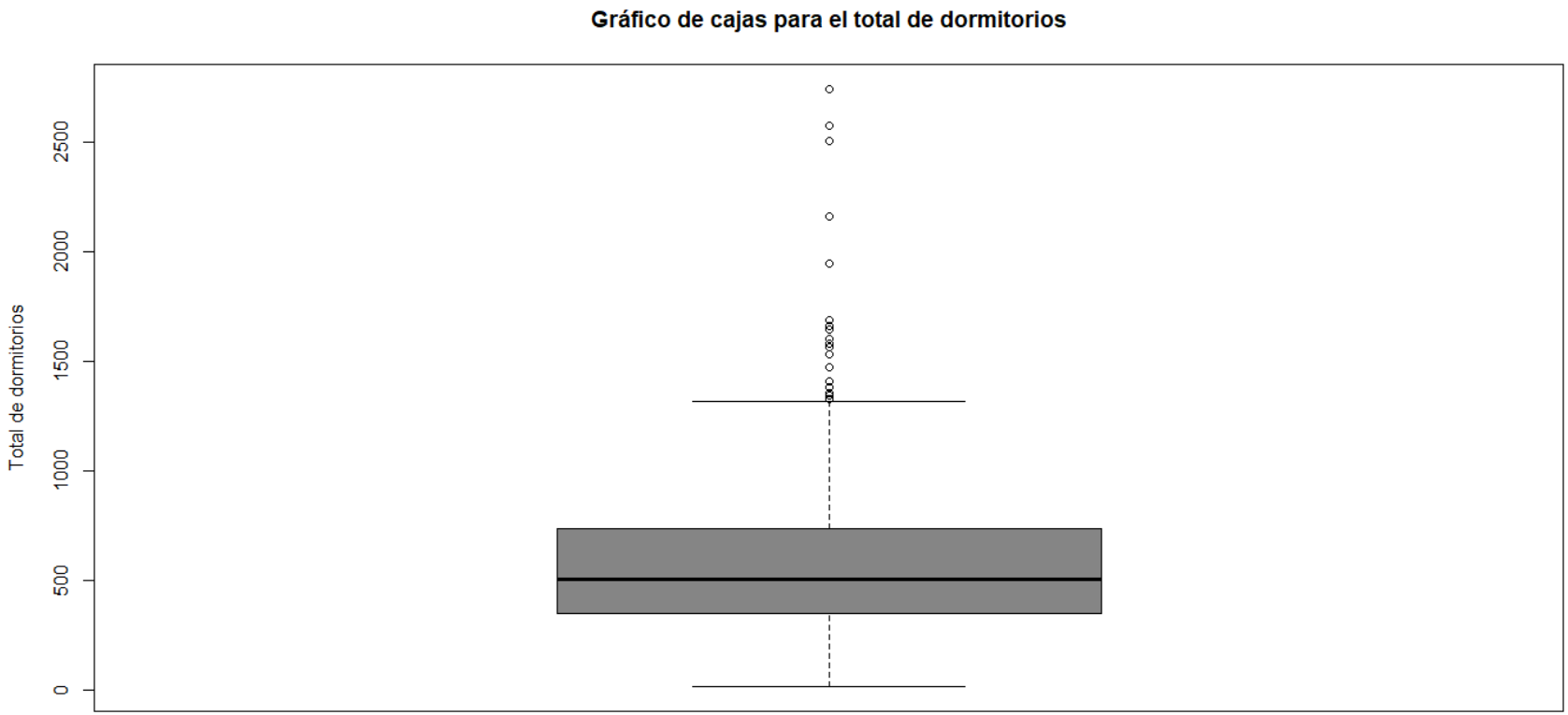


Figura: Gráfico de cajas para la variable 'Total de dormitorios'

Variable 'Población'

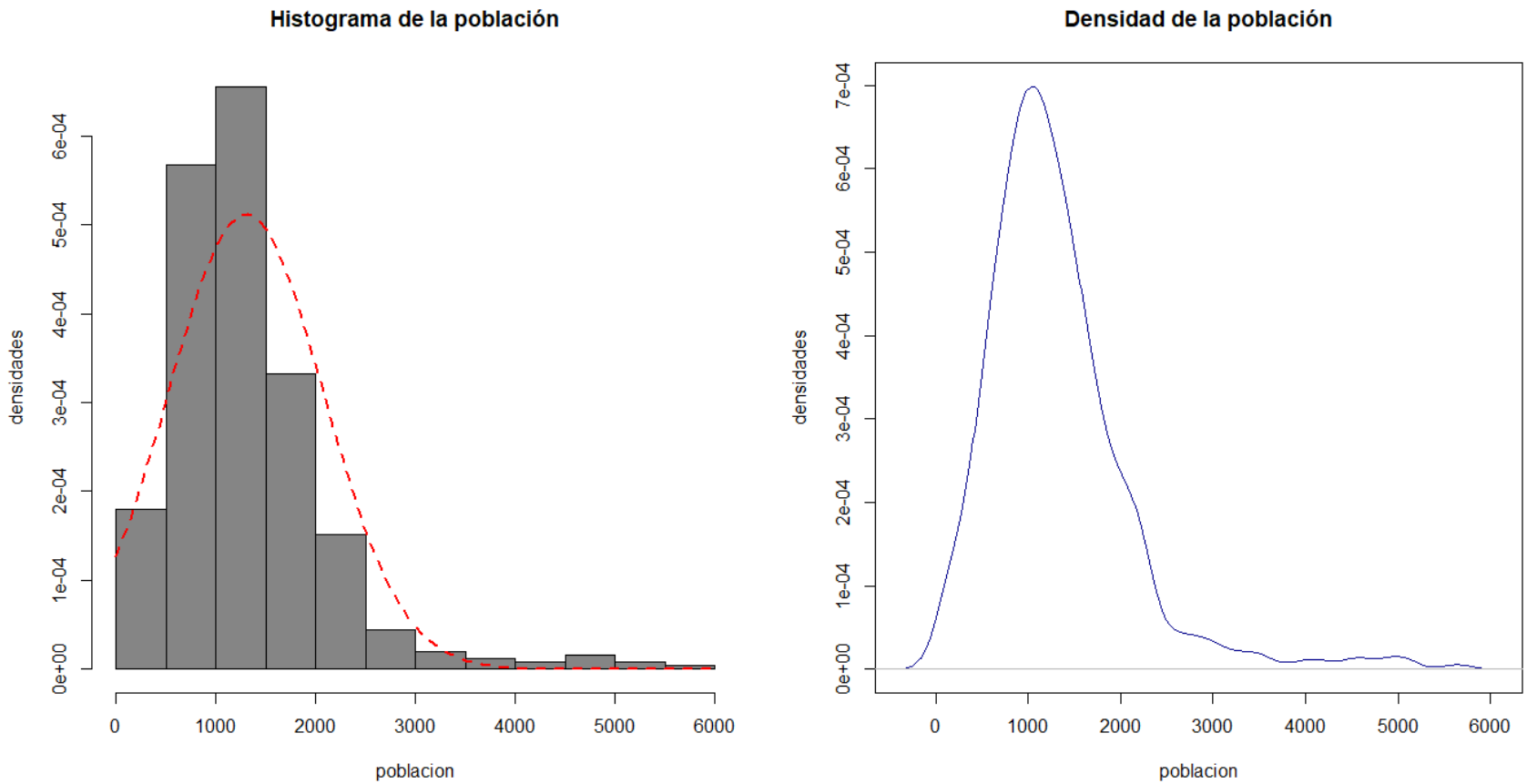


Figura: Histograma y densidad de la variable 'Población'

Variable 'Población'

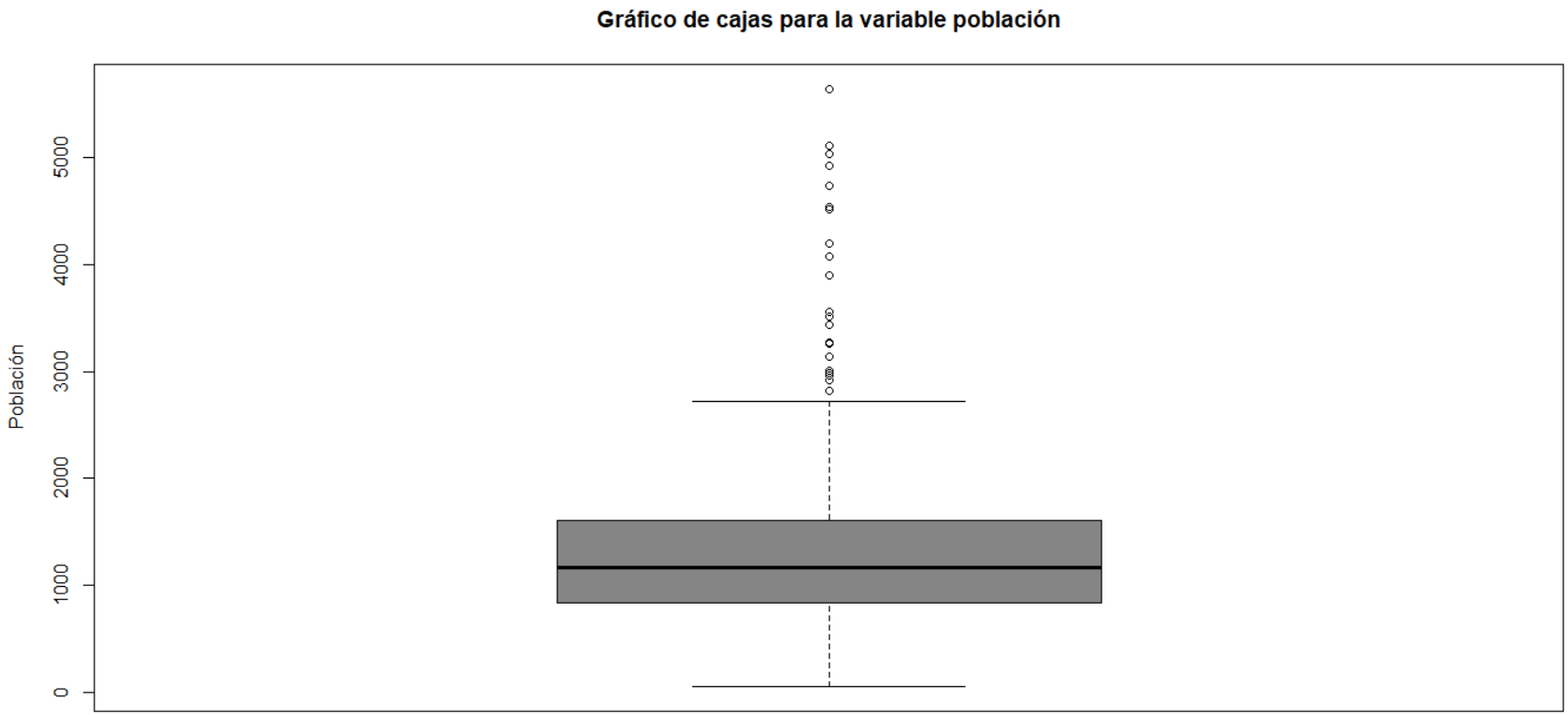


Figura: Gráfico de cajas para la variable 'Población'

Variable 'Hogares'

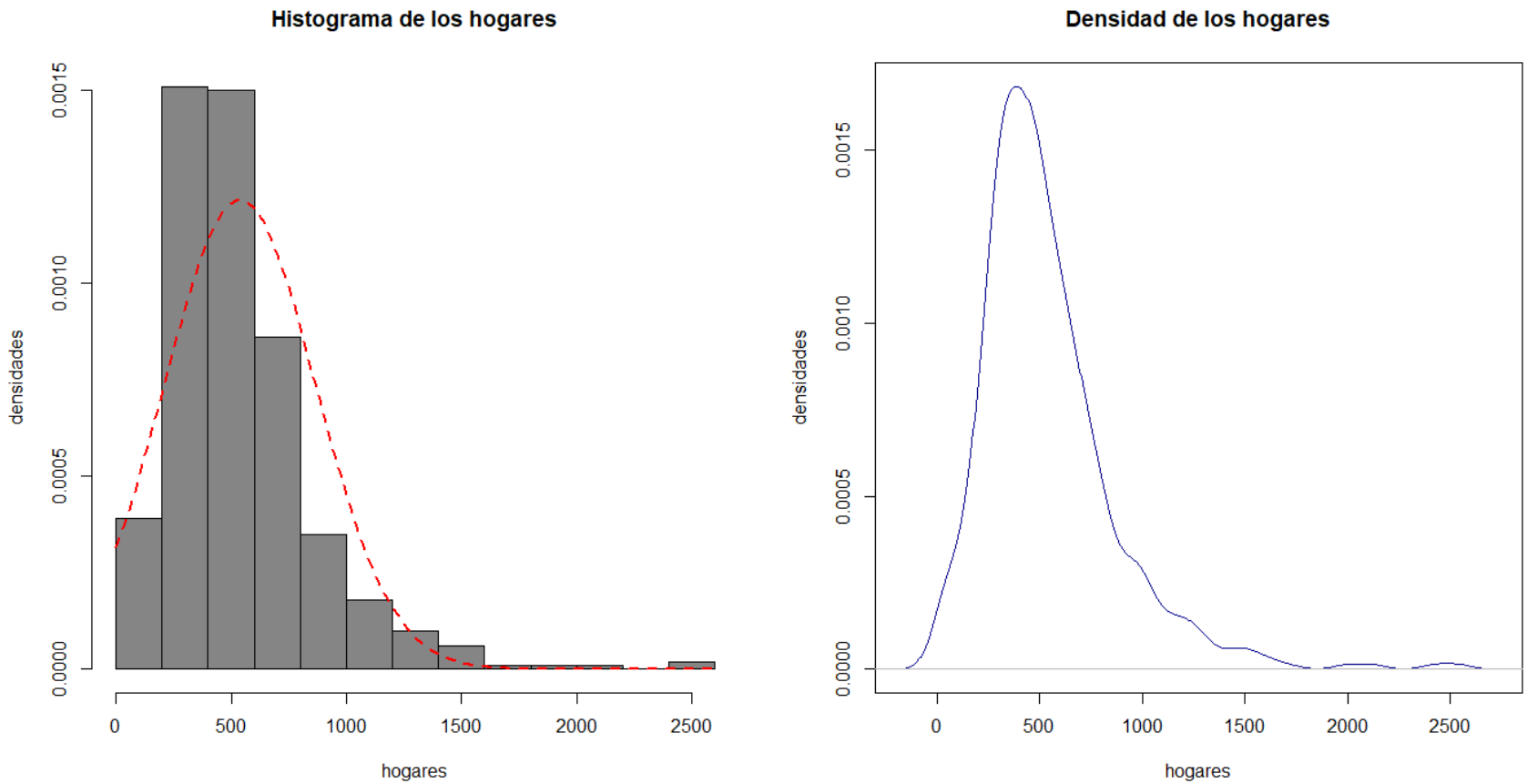


Figura: Histograma y densidad de la variable 'Hogares'

Variable 'Hogares'

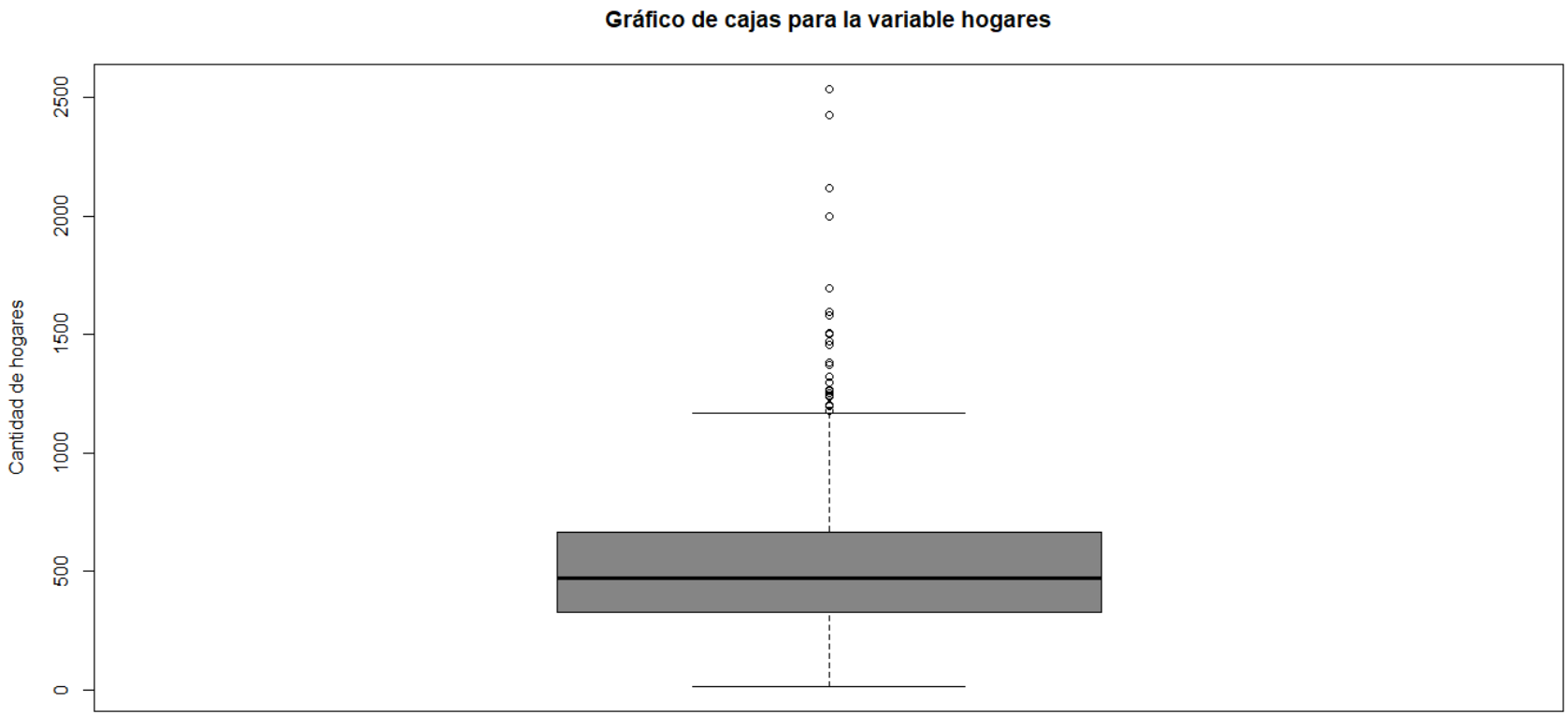


Figura: Gráfico de cajas para la variable 'Hogares'

California

Figura: Lugar de donde provienen los datos

San Francisco

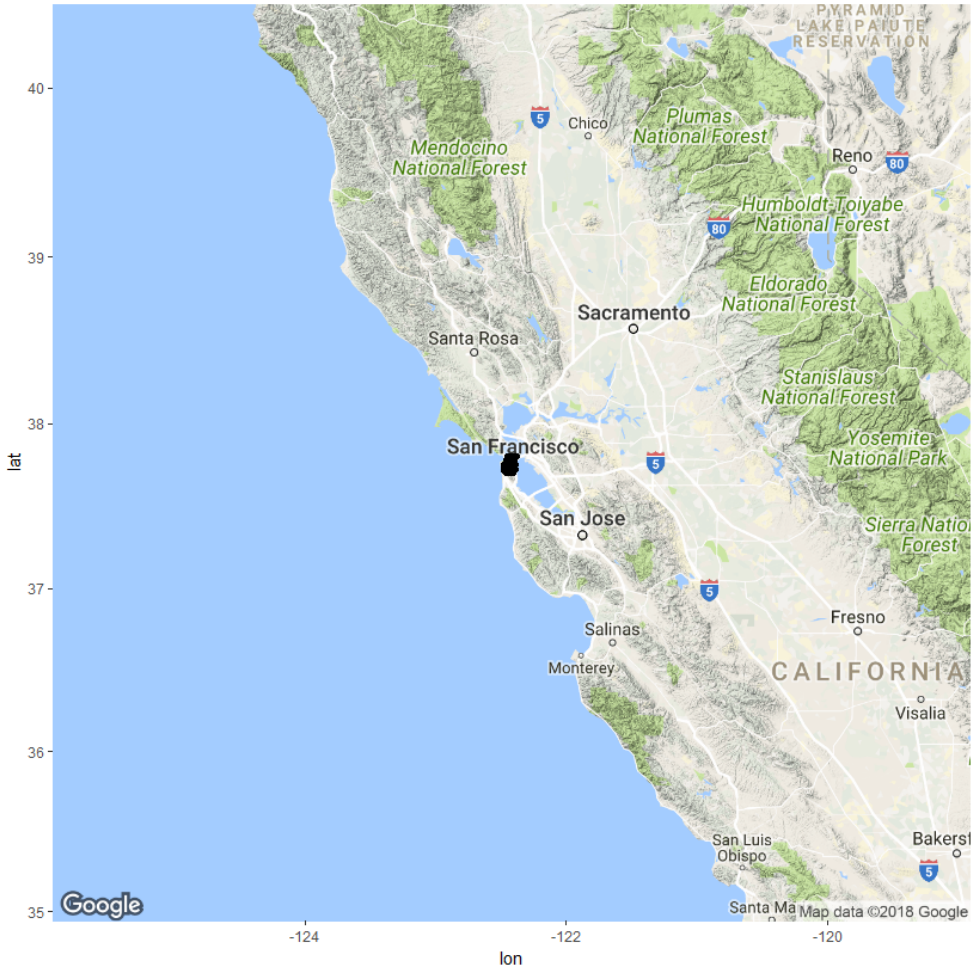


Figura: Mapa con los puntos dados de longitud y latitud

San Diego

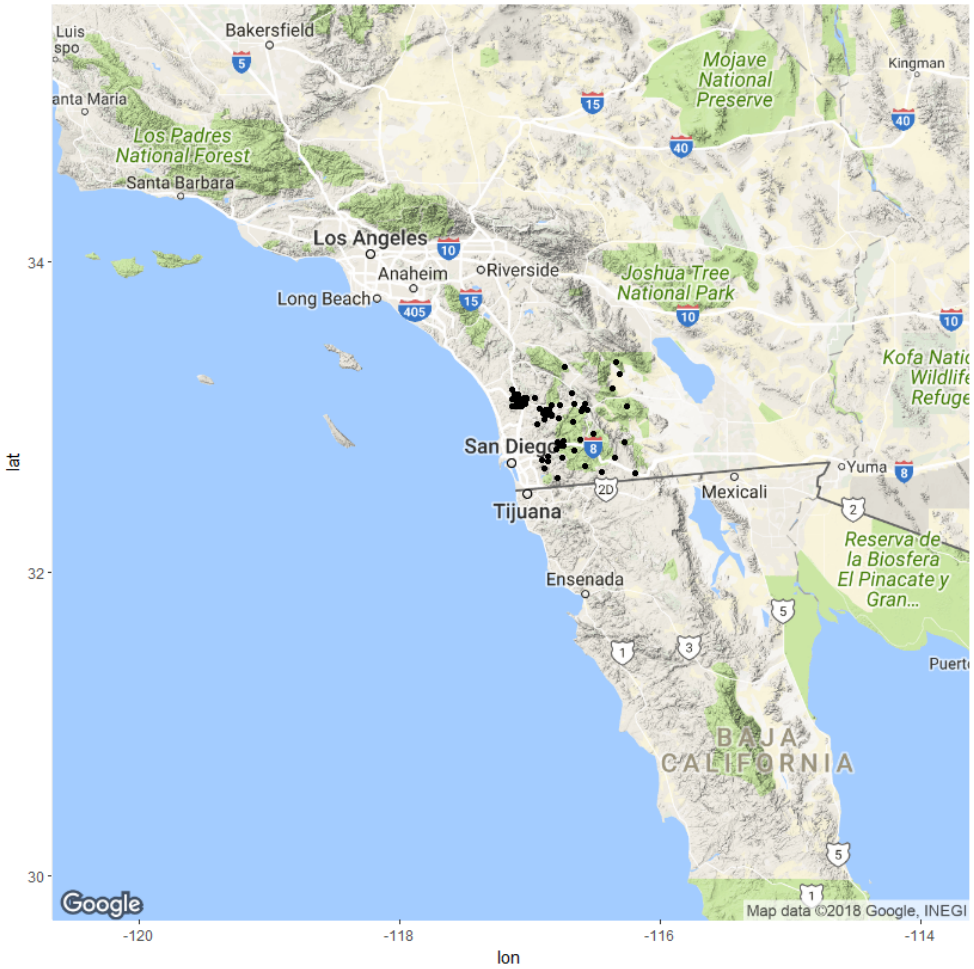


Figura: Mapa con los puntos dados de longitud y latitud

Posibles relaciones entre las variables explicativas

- Matriz de correlaciones:

	Ingreso	Edad	Habitaciones	Dormitorios	Población	Hogares
Ingreso	1.00000000	0.01725444	0.07902878	-0.1813679	-0.1625779	-0.1604979
Edad	0.01725444	1.00000000	-0.31169821	-0.1702448	-0.2651765	-0.1445663
Habitaciones	0.07902878	-0.31169821	1.00000000	0.8620727	0.8576863	0.8652765
Dormitorios	-0.18136789	-0.17024480	0.86207267	1.0000000	0.8340611	0.9887048
Población	-0.16257787	-0.26517653	0.85768626	0.8340611	1.0000000	0.8539180
Hogares	-0.16049785	-0.14456632	0.86527649	0.9887048	0.8539180	1.0000000

Figura: Matriz de correlaciones entre covariables

Posibles relaciones entre las variables explicativas

- Variables 'Total de dormitorios' y 'Hogares':

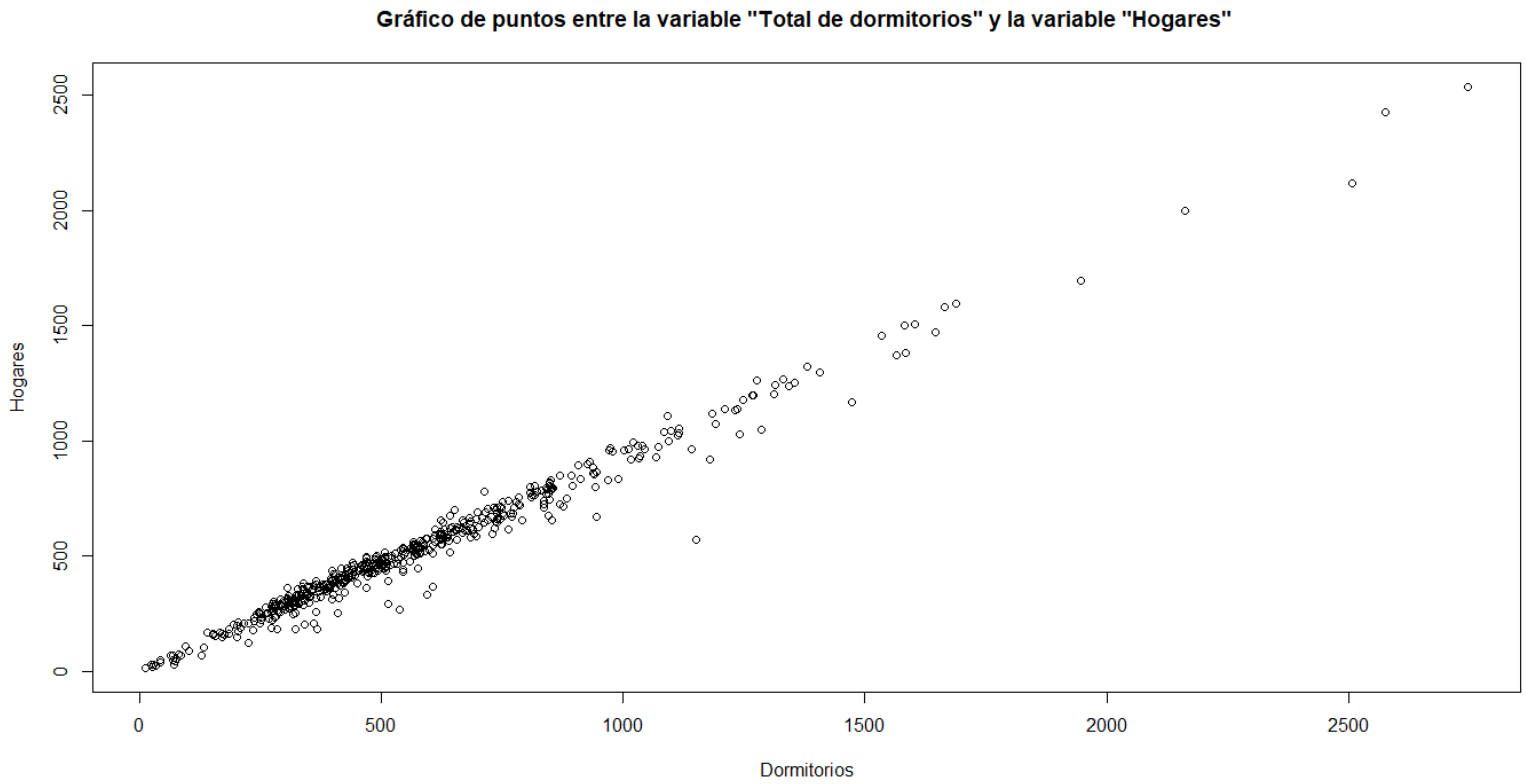


Figura: Gráfico de puntos entre las variables 'Total de dormitorios' y 'Hogares'

Posibles relaciones entre las variables explicativas

- correlación de Pearson: $r = 0{,}9887048$
- correlación de Spearman: $\rho = 0{,}9819133$

Posibles relaciones entre las variables explicativas

- Variables 'Población' y 'Hogares':

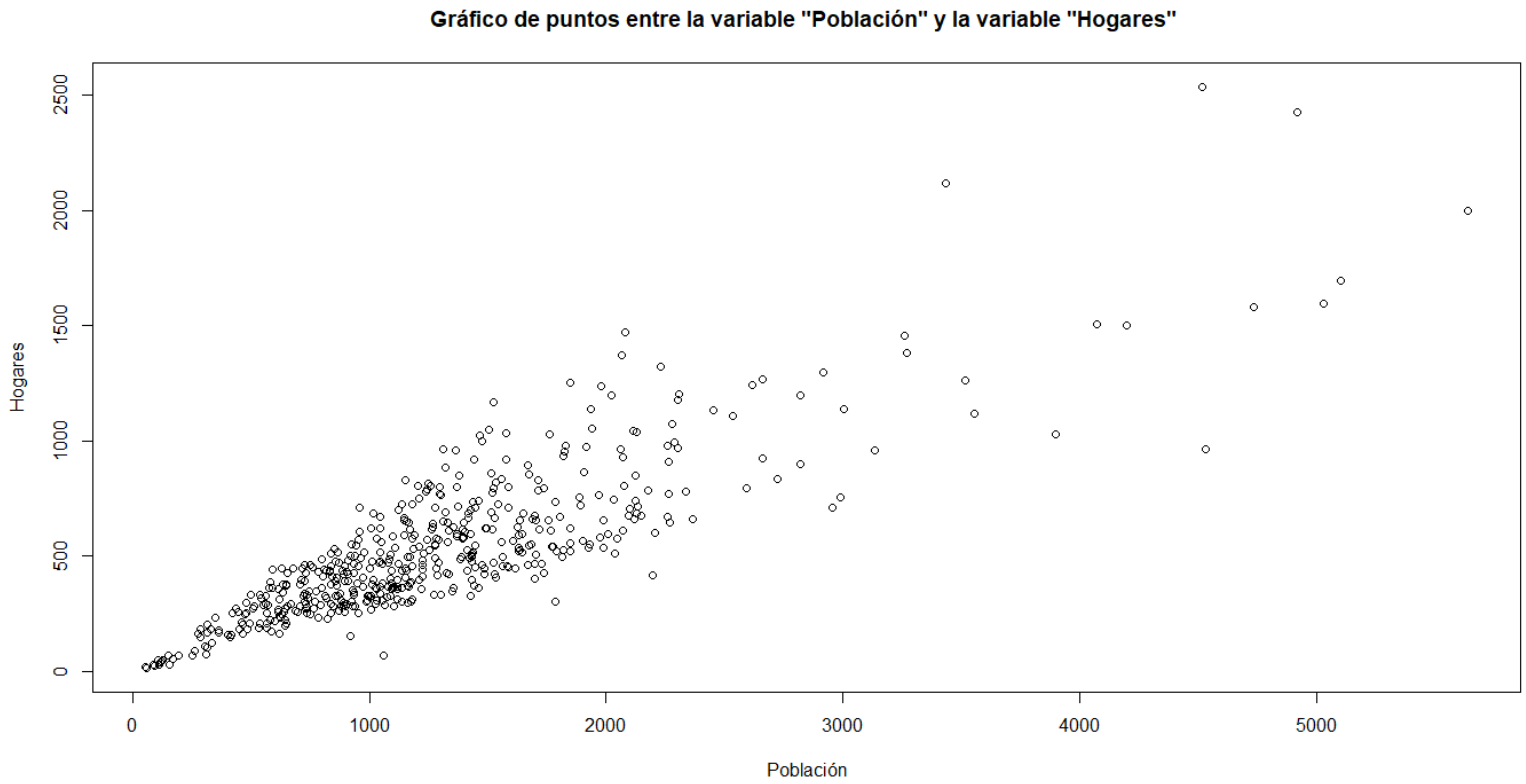


Figura: Gráfico de puntos entre las variables 'Población' y 'Hogares'

Posibles relaciones entre las variables explicativas

- correlación de Pearson: $r = 0{,}853918$
- correlación de Spearman: $\rho = 0{,}8395669$

Posibles relaciones entre las variables explicativas

- Variables 'total de habitaciones' y 'total de dormitorios':

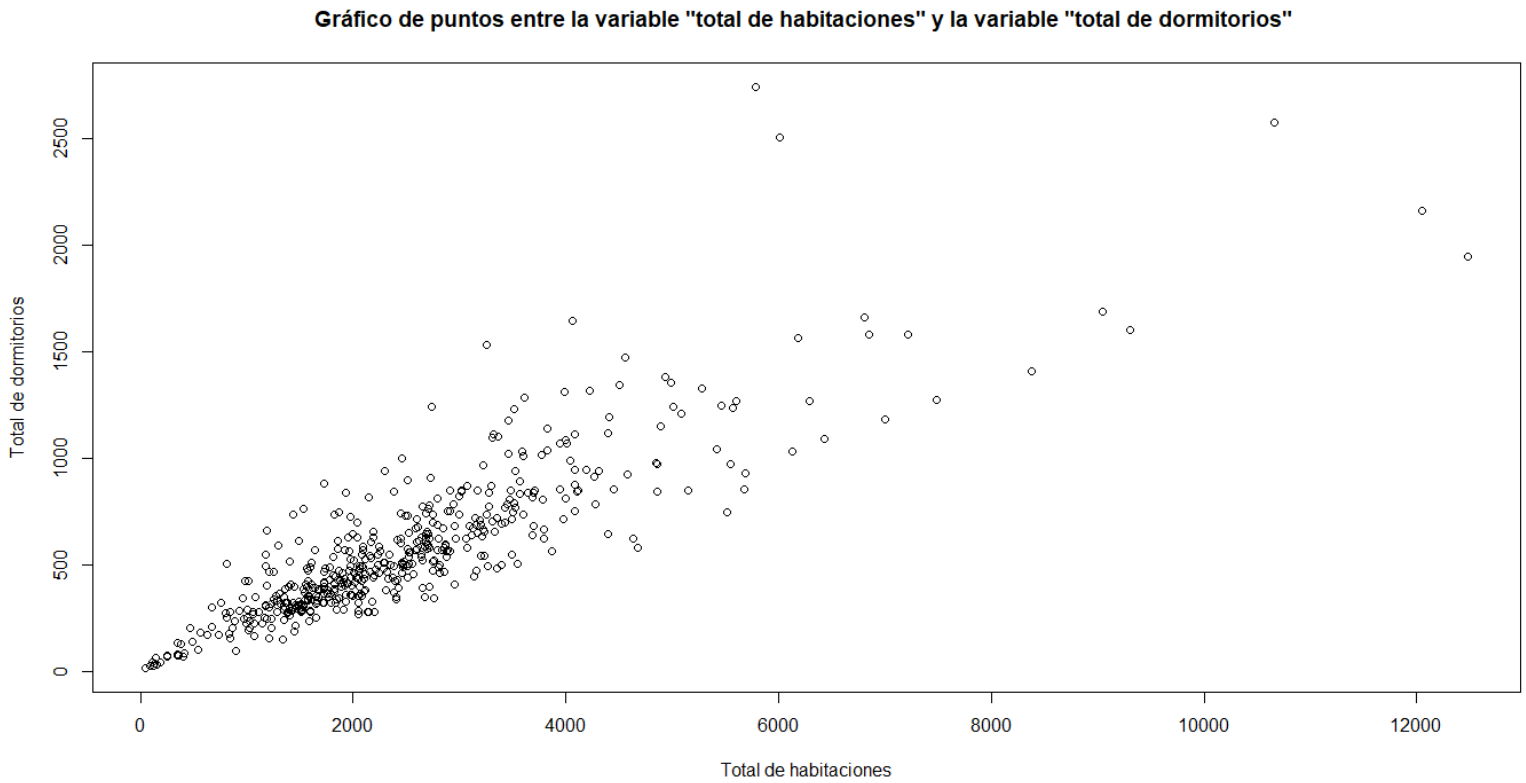


Figura: Gráfico de puntos entre las variables 'total de habitaciones' y 'total de dormitorios'

Posibles relaciones entre las variables explicativas

- correlación de Pearson: $r = 0{,}862072$
- correlación de Spearman: $\rho = 0{,}8716658$

Modelo ajustado e interpretación

El modelo ajustado incluyendo todas las variables sin transformación y sin selección de variables es:
$Y = 57720{,}52 + 24261{,}20X_1 + 3443{,}94X_2 + 19{,}09X_3 - 67{,}72X_4 - 121{,}66X_5 + 315{,}92X_6$

Donde: Y=Valor mediano de la casa, X_1 =Ingreso mediano, X_2 =Edad mediana de la vivienda, X_3 =Total de habitaciones, X_4 =Total de dormitorios, X_5 =Población, X_6 =Hogares.

- $R^2 = 0{,}5425$
- $R^2_{ajustado} = 0{,}537$
- $CME = \hat{\sigma^2} = 6724487923$

Modelo ajustado e interpretación

- β_0:Representa el valor del intercepto con el eje y, esto quiere decir que -4113000 siempre va a afectar mi variable dependiente sin importar en cuantas unidades aumente mis otras variables.
- β_1:por cada unidad que aumente de ingreso mediano se aumentan 22100 unidades del valor mediano de la casa.
- β_2:por cada unidad que aumente de edad mediana de la vivienda se aumentan 1835 unidades del valor mediano de la casa.
- β_3:por cada unidad que aumente de total de habitaciones se aumentan 30,35 unidades del valor mediano de la casa.
- β_4:por cada unidad que aumente de total de dormitorios se disminuye 14,43 unidades del valor mediano de la casa.

- β_5:por cada unidad que aumente de población se disminuye 124,5 unidades del valor mediano de la casa.
- β_6:por cada unidad que aumente de hogares se disminuye 210,2 unidades del valor mediano de la casa.
- β_7:por cada unidad que aumente de latitud se disminuye 33230 unidades del valor mediano de la casa.
- β_8:por cada unidad que aumente de longitud se disminuye 45080 unidades del valor mediano de la casa.

Modelo ajustado con selección de variables

El modelo ajustado, utilizando el método forward para seleccionar variables es:
$Y = -587755{,}25 + 22415{,}05X_1 + 1823{,}94X_2 + 30{,}59X_3 - 59{,}62X_4 - 123{,}27X_5 + 256{,}45X_6 - 19522{,}07X_7$

Donde: Y=Valor mediano de la casa, X_1 =Ingreso mediano, X_2 =Edad mediana de la vivienda, X_3 =Total de habitaciones, X_4 =Total de dormitorios, X_5 =Población, X_6 =Hogares, X_7 =Latitud.

- $R^2 = 0{,}5743$
- $R^2_{ajustado} = 0{,}5683$
- $CME = \hat{\sigma^2} = 6270012635$

Modelo ajustado con selección de variables

El modelo ajustado, utilizando el método backward para seleccionar variables es:
$Y = -1922000 + 22690X_1 + 1831X_2 + 29{,}72X_3 - 121{,}9X_5 + 192{,}6X_6 - 16910X_8$

Donde: Y=Valor mediano de la casa, X_1 =Ingreso mediano, X_2 =Edad mediana de la vivienda, X_3 =Total de habitaciones, X_5 =Población, X_6 =Hogares, X_8 =Longitud.

- $R^2 = 0{,}5747$
- $R^2_{ajustado} = 0{,}5695$
- $CME = \hat{\sigma^2} = 6251879617$

Modelo ajustado con selección de variables

El modelo ajustado, utilizando el método stepwise para seleccionar variables es:
$Y = -593652{,}6 + 22992{,}9X_1 + 1855{,}1X_2 + 29{,}3X_3 - 120{,}6X_5 + 192{,}4X_6 + 19572{,}7X_7$

Donde: Y=Valor mediano de la casa, X_1 =Ingreso mediano, X_2 =Edad mediana de la vivienda, X_3 =Total de habitaciones, X_5 =Población, X_6 =Hogares, X_7 =Latitud.

- $R^2 = 0{,}5737$
- $R^2_{ajustado} = 0{,}5685$
- $CME = \hat{\sigma^2} = 6266212125$

Conclusión

Al finalizar nuestra selección de variables, con el fin de ajustar el mejor modelo posible para la variable valor mediano de la casa, comparamos cada uno de los 4 modelos obtenidos (completo, forward, backward, stepwise) con respecto al $R^2_{ajustado}$ y el $CME = \hat{\sigma^2}$. Podemos concluir que el mejor modelo que logramos obtener para nuestros 500 datos sin hacer transformación de variables, fue el generado por el método de selección "Backward", el cuál nos elimino las variables predictoras X_4 :Total de dormitorios y X_7 :Latitud.